Title	A Geometrical Structure in the Statistical Information Loss under the Curved Exponential Family
Author(s)	熊谷, 悦生
Citation	大阪大学, 1997, 博士論文
Version Type	VoR
URL	https://doi.org/10.11501/3128788
rights	
Note	

A Geometrical Structure in the Statistical Information Loss under the Curved Exponential Family

Etsuo Kumagai

Department of Mathematical Science,
Division of Informatics and Mathematical Science,
Graduate School of Engineering Science,
Osaka University

November, 1996

Preface

R.A.Fisher and C.R.Rao tried to refine and compare more asymptotic efficiency of the asymptotic efficient estimators with the lower bound of the asymptotic variance that is equal to the inverse of Fisher information $I(\theta)$ (say). Fisher(1925) introduced the concept of the information loss by the difference between the total information and the information of the estimator in order to obtain more efficient estimator among the exactly or asymptotically efficient estimators. He calculated the asymptotic information loss in multinomial case, but there existed some confusions in it. Rao(1961) discussed the various concepts of the asymptotic second order efficiency in general and, particularly, completed the way of Fisher's evaluation of the difference between the likelihood score function and the estimator in multinomial case which is called the Rao's definition of the asymptotic second order efficiency. Then, he suggested that Rao's amount of second order efficiency could be equal to Fisher's information loss under some regularity conditions.

B.Efron(1975) showed that these amounts of second order efficiency derived from different motivations are equal in the curved exponential family with regularity conditions for continuity. He proved that this common amount are characterized by the statistical curvature $\Gamma_S(\theta)$ (say) of curved exponential family, and showed that the information loss defined by Fisher is asymptotically and geometrically specified to be $I(\theta)\Gamma_S(\theta)^2$ for the maximum likelihood estimator. Furthermore he showed by a counterexample that the above two amounts could be different in the multinomial case treated really by Fisher and Rao.

In this thesis, we shall mainly consider the exact information loss for the fixed sample number n in order to investigate the reason why Efron's statistical curvature is inevitable for the asymptotically information loss of maximum likelihood estimator. And we shall give relationships between the former classical likelihood theories and the later recent information geometry by the circular mechanism in the two dimensional curved exponential family.

We shall describe the construction in this thesis as follows: First of all, we shall prepare some definitions and notations (Chapter 1), and we shall briefly survey a history about the information loss by Fisher, Rao, and Efron (Chapter 2). Next, we shall explain the exponential family and the curved exponential family in general (Chapter 3). And, by restricting the curved exponential family with the two dimension, we shall demonstrate mainly

1. the contradictory demand on Efron's counterexample,
2. the exact information loss in Fisher's circle model,
3. the circular mechanism.

In Chapter 4, we shall investigate Efron's counterexample in detail, show some properties when the counterexample would hold, and prove that there exists a contradictory demand in the counterexample. Thus we shall obtain that Efron's counterexample does not hold.

In Chapter 5, we shall consider Fisher's circle model as the model for investigating the exact information loss and demonstrate the exact information loss in detail, so that we have one characteristic which the conditional variance of the likelihood score function given the maximum likelihood estimator reduces the conditional variance of the length given the angle. This implies the visualization of exact information loss. We shall also consider the asymptotic result based on the exact information loss.

In Chapter 6, we shall investigate the relationship between the mathematical curvature and the statistical curvature in the curved exponential family, so that we shall obtain the circular mechanism which is an algorithm to obtain the statistical curvature and the center of osculating circle with the radius of its inverse by using derivatives to second order of log-likelihood function. This implies the necessary of the statistical curvature as a signpost from the ordinary likelihood estimation theory to the information geometry.

In appendices, we shall add some explanations for the convex conjugate on the exponential family, for the information circle which was defined by Efron(1978), and for the fundamental of Amari's framework.

This thesis has been typeset by using LaTeX and the figures in the thesis have been drawn by using GNUPLOT3.5+3.1.2.

Etsuo Kumagai

November, 1996

Acknowledgments

The author would like to express his sincere gratitude to Professor N. Inagaki for his hearty proper instruction on this series of researches and for his constructive and detailed comments on a preliminary version of the manuscript.

In addition, the author is grateful to Professors S. Shirahata, M. Goto, H. Nagai for their encouragement and kindness.

Finally, the author acknowledges the other staffs of the Department of Mathematical Science, Osaka University for their various help in all respects.

Contents

Chapter 1

Preliminaries

Let $(\mathbf{R}^k, \mathbf{B}^k)$ be a Borel measurable space in the k dimensional Euclidean space and let a parameter space Θ be an open subset of $\mathbf{R}^r$, where $r < k$. Let $\Pi := \{P_\theta \,|\, \theta \in \Theta\}$ be a family of probability measure on $\mathbf{B}^k$. We assume that Π is dominated by a $\sigma-$finite measure μ. By the lemma of Halmos-Savage(1947), there exists a probability measure P_0 such that P_0 is equivalent to Π, that is, $\Pi \sim P_0$. P_0 is called the pivotal probability measure. By the Radon-Nikodym derivative of P_θ with respect to P_0, we shall describe its derivative as follows :

$$f(\boldsymbol{x} : \theta) \;=\; \frac{dP_\theta}{dP_0}, \quad \theta \in \Theta.$$

For an observation $\boldsymbol{X}$ with distribution P_θ, let $T = T(\boldsymbol{X})$ be an estimator of θ which is a measurable function from $\mathbf{R}^k$ to Θ. Then is holds that $P_\theta^T \ll P_0^T$ for the measure on $(\mathbf{R}^r, \mathbf{B}^r)$ induced by T. Thereby there exists

$$g(t : \theta) \;=\; \frac{dP_\theta^T}{dP_0^T}, \quad \theta \in \Theta.$$

Here we have the following lemma :

Lemma 1.0.1 (Inagaki(1983)) *For the above $f(\boldsymbol{x} : \theta)$ and $g(t : \theta)$, let*

$$h(\boldsymbol{x} : \theta \,|\, t) \equiv \begin{cases} f(\boldsymbol{x} : \theta)/g(t : \theta) & \textit{if } T(\boldsymbol{x}) = t \textit{ and } g(t : \theta) > 0, \\ 1 & \textit{if } T(\boldsymbol{x}) = t \textit{ and } g(t : \theta) = 0, \\ 0 & \textit{otherwise.} \end{cases}$$

Then it holds that

$$f(\boldsymbol{x} : \theta) \;=\; g(t : \theta)\, h(\boldsymbol{x} : \theta \,|\, t), \quad a.s.\ [P_\theta]. \qquad \Box$$

Let $\boldsymbol{X}$ be a random vector which is identically independent distributed (i.i.d.) with the probability (density) function $f(\boldsymbol{x} : \theta)$. For the joint probability (density) function

$$f_n(\boldsymbol{x} : \theta) := \prod_{j=1}^{n} f(\boldsymbol{x}_j : \theta),$$

since, in the statistical inference, we regard $f_n(\boldsymbol{x} : \theta)$ as the function of θ which represents the degree of likelihood of the parameter θ for the observation $\boldsymbol{x}$, we call it the likelihood function

$$L_n = L_n(\theta \,|\, \boldsymbol{x}) \equiv f_n(\boldsymbol{x} : \theta).$$

And the log-likelihood function is defined by

$$\ell_n(\theta \,|\, \boldsymbol{x}) \equiv \log L_n(\theta \,|\, \boldsymbol{x}) = \sum_{j=1}^{n} \ell(\theta \,|\, \boldsymbol{x}_j) = \sum_{j=1}^{n} \log f(\boldsymbol{x}_j : \theta).$$

Let a estimator $T_n = T_n(\boldsymbol{X})$ be a random variable. The (unknown) parameter θ of the distribution of $\boldsymbol{X}$ is estimated by the estimator T_n. We shall enumerate the definitions about the estimator T_n as follows :

(Unbiased Estimator) the expectation of T_n is equivalent to the parameter θ, that is, $E_\theta[T_n] = \theta$.

(Efficient Estimator) For the Fisher information $I(\theta)$, the variance of T_n is equivalent to $(nI(\theta))^{-1}$, that is, $n\, I(\theta)\, V_\theta(T_n) = 1$.

(Consistent Estimator) T_n convergences to θ in probability, that is, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr_\theta\{\, |T_n - \theta| \geq \varepsilon \,\} = 0 \quad (\forall \theta \in \Theta).$$

(Asymptotically Normal Estimator) $\sqrt{n}(T_n - \theta)$ converges to the normal distribution in law, that is, for any θ and any z,

$$\lim_{n\to\infty} \Pr_\theta \left(\frac{\sqrt{n}(T_n - \theta)}{\sigma(\theta)} \leq z \right) = \Phi(z),$$

where $\Phi(z)$ is the standard normal distribution and $\sigma(\theta)$ is the asymptotic variance.

(Asymptotically Efficient Estimator) T_n is asymptotically normal with the asymptotic variance $I(\theta)^{-1}$, that is,

$$\sqrt{n}(T_n - \theta) \to N(0, I(\theta)^{-1}) \quad \text{in law} \quad (n \to \infty).$$

(Maximum Likelihood Estimator) The maximum likelihood estimator $\hat{\theta}$ is a point of Θ such that

$$\ell_n(\hat{\theta} \,|\, \boldsymbol{x}) \;=\; \max\{\, \ell_n(\theta \,|\, \boldsymbol{x}) : \theta \in \Theta \,\}.$$

Chapter 2

History of information loss

We shall show a history of some authors with respect to the information loss.

2.1 Fisher's information loss

Fisher(1925) described about the loss of information as follows :

> *When the sets of samples which for one value of θ have the same value of $\partial \log L_n/\partial\theta$, have no longer the same value for other values of θ, there exists no sufficient statistic, and some loss of information will necessarily ensue upon the substitution of a single estimate for the original data upon which it was based. (in section 11)*

This may mean that if there exists a sufficient statistic, then, for any element $\boldsymbol{x}_1$ of the subset $\{\boldsymbol{x} : \partial \log L_n/\partial\theta_1 = C_1\}$ (C_1 is a constant), there exists a constant C_2 such that $\partial \log L_n/\partial\theta_2 = C_2$ for a different value θ_2 from θ_1, so that there shall exist no information loss by the substitution of a single estimate for the original data which it was based.

And Fisher denoted the total loss of information in the maximum likelihood estimator $\hat{\theta}$ as follows :

> *if now $\partial \log L_n/\partial\hat{\theta} = 0$, then to a first approximation*
>
> $$\frac{\partial \log L_n}{\partial\theta} \;=\; (\theta - \hat{\theta})\,\frac{\partial^2 \log L_n}{\partial\theta^2},$$
>
> *and the variance of $\partial \log L_n/\partial\theta$ in a set of samples for which $\hat{\theta}$ is constant, will be given by the variance of $\partial^2 \log L_n/\partial\theta^2$ within the set multiplied by $(\theta-\hat{\theta})^2$, or the total loss of information will be given by the general variance within such sets multiplied by $V(\hat{\theta})$. (in section 11)*

This may mean that the total loss of information for the maximum likelihood estimator $\hat{\theta}$ is approximately given by

$$V(\hat{\theta}) \cdot V\left[\frac{\partial^2 \log L_n}{\partial\theta^2}\,\middle|\,\hat{\theta}\right].$$

Fisher calculated this loss for the maximum likelihood estimator $\hat{\theta}$ in the large samples concretely under the multinomial distribution as follows : For the sample which consists of observed numbers $x_1, \ldots, x_s$ in categories in which the expectations are $m_1, \ldots, m_s$, if the expectations are functions of θ, then it holds that

$$\begin{aligned}
\log L_n &= \sum_{j=1}^{s} x_j \log m_j \;+\; c_0, \\
\frac{\partial \log L_n}{\partial \theta} &= \sum_{j=1}^{s} x_j \, \frac{\dot{m}_j}{m_j}, \\
\frac{\partial^2 \log L_n}{\partial \theta^2} &= \sum_{j=1}^{s} x_j \left(\frac{\ddot{m}_j}{m_j} - \frac{\dot{m}_j^2}{m_j^2} \right),
\end{aligned}$$

where the dot notation means the differentiation with respect to θ and c_0 is a constant which is independent of θ, so that the loss of information in large samples is represented by

$$\text{(2.1.1)} \qquad \frac{\sum_{j=1}^{s} \frac{1}{m_j} \left(\ddot{m}_j - \frac{\dot{m}_j^2}{m_j} \right)^2}{\sum_{j=1}^{s} \frac{\dot{m}_j^2}{m_j}} - \frac{1}{n} \sum_{j=1}^{s} \frac{\dot{m}_j^2}{m_j} - \left\{ \frac{\sum_{j=1}^{s} \frac{\dot{m}_j}{m_j} \left(\ddot{m}_j - \frac{\dot{m}_j^2}{m_j} \right)}{\sum_{j=1}^{s} \frac{\dot{m}_j^2}{m_j}} \right\}^2 ,$$

where $n = \sum_{j=1}^{s} x_j = \sum_{j=1}^{s} m_j$.

2.2 Rao's second order efficiency

Let I_{T_n} be the information for a statistic T_n. For Fisher's proposition about the information loss, first of all, Rao(1961) investigated a sufficient condition for the convergence of (I_{T_n}/n) to the Fisher information $I(\theta)$, as the sample size $n \to \infty$ as follows:

$$\left| \frac{1}{\sqrt{n}} \left(\frac{d \log L_n}{d\theta} \right) \;-\; a \;-\; b\,\sqrt{n}(T_n - \theta) \right| \;\to\; 0$$

in probability, where a, b are constants which may depend on θ. This proposition itself was first derived from Doob(1934). He defined this as the new definition of the asymptotic efficiency. This is called the first-order efficiency. Under the first-order efficiency, he proposed the second-order efficiency E_2 as the minimum asymptotic variance of

$$\text{(2.2.1)} \qquad \frac{d \log L_n}{d\theta} \;-\; \sqrt{n}\, a \;-\; b\, n(T_n - \theta) - \lambda\, n\, (T_n - \theta)^2$$

when minimized with respect to λ. Rao calculated the second-order efficiency E_2 for some various estimators based on the multinomial distribution. Let $\pi_j = m_j/n$ and $p_j = x_j/n$ for the previous notations.

Method of Estimation	Estimating equation	E_2
maximum likelihood	$\sum_{j=1}^s p_j \frac{\dot\pi_j}{\pi_j} = 0$	$E_2(m.l.)$
minimum χ^2	$\sum_{j=1}^s \dot\pi_j \frac{p_j^2}{\pi_j^2} = 0$	$E_2(m.l.) + \Delta$
minimum modified χ^2	$\sum_{j=1}^s \dot\pi_j \frac{\pi_j}{p_j} = 0$	$E_2(m.l.) + 4\Delta$
Haldane's minimum discrepancy D_k	$\sum_{j=1}^s \dot\pi_j \frac{\pi_j^k}{p_j^k} = 0$	$E_2(m.l.) + (k+1)^2\Delta$
minimum Hellinger distance	$\sum_{j=1}^s \dot\pi_j \frac{p_j^{\frac{1}{2}}}{\pi_j^{\frac{1}{2}}} = 0$	$E_2(m.l.) + \frac{1}{4}\Delta$
minimum KL separator	$\sum_{j=1}^s \dot\pi_j \log\left(\frac{\pi_j}{p_j}\right) = 0$	$E_2(m.l.) + \Delta$

In the above table, $E_2(m.l.)$ is

$$E_2(m.l.) = \frac{\mu_{02} - 2\mu_{21} + \mu_{40}}{\mu_{20}} - \mu_{20} - \frac{\mu_{11}^2 + \mu_{30}^2 - 2\mu_{11}\mu_{30}}{\mu_{20}^2}, \tag{2.2.2}$$

where

$$\mu_{ik} := \sum_{j=1}^s \pi_j \left(\frac{\dot\pi_j}{\pi_j}\right)^i \left(\frac{\ddot\pi_j}{\pi_j}\right)^k .$$

And Δ is

$$\Delta = \frac{1}{2}\sum_{j=1}^s \left(\frac{\dot\pi_j}{\pi_j}\right)^2 - \frac{\mu_{40}}{\mu_{20}} + \frac{1}{2}\frac{\mu_{30}^2}{\mu_{20}^2}.$$

Rao asserted that the minimum variance of (2.2.1) would be same as the limit of $(nI(\theta) - I_{T_n})$, that is,

$$\lim_{n\to\infty}(nI(\theta) - I_{T_n}) = E_2 \tag{2.2.3}$$

under some regularity conditions. Note that $E_2(m.l.)$ of maximum likelihood estimator is equivalent to Fisher's information loss (2.1.1).

2.3 Efron's information loss

Based on Fisher's information loss and Rao's second-order efficiency, Efron(1975) restarted the information loss by a geometric view point based on the curved exponential family. He defined the statistical curvature for the density $f(\boldsymbol{x} : \theta)$ at θ as follows : for the log-likelihood $\ell(\theta \,|\, \boldsymbol{x}) = \log f(\boldsymbol{x} : \theta)$,

$$\Gamma_S(\theta) \equiv \left(\frac{\det \begin{pmatrix} E\,\dot\ell(\theta\,|\,\boldsymbol{x})^2 & E\,\dot\ell(\theta\,|\,\boldsymbol{x})\,\ddot\ell(\theta\,|\,\boldsymbol{x}) \\ E\,\ddot\ell(\theta\,|\,\boldsymbol{x})\,\dot\ell(\theta\,|\,\boldsymbol{x}) & E\,\ddot\ell(\theta\,|\,\boldsymbol{x})^2 - I(\theta)^2 \end{pmatrix}}{I(\theta)^3} \right)^{\frac{1}{2}} .$$

Using the statistical curvature $\Gamma_S(\theta)$, Efron showed the asymptotic information loss of the maximum likelihood estimator $\hat{\theta}$ in the curved exponential family as follows :

$$\lim_{n\to\infty} (nI(\theta) - I_{\hat{\theta}}) \;=\; I(\theta)\,\Gamma_S(\theta)^2. \tag{2.3.1}$$

This representation is equivalent to Fisher's and Rao's representations (2.1.1), (2.2.2) in the multinomial distribution, but Efron's contribution is that he asserted a new view-point for the (asymptotic) information loss by bringing the geometrical curvature into the statistical problem. Also, by restricting the distribution to the curved exponential family, he made Fisher's argument for the information loss be simple as follows : for an estimator T_n of parameter θ, the exact information loss is

$$nI(\theta) - I_{T_n} \;=\; E_{T_n}\left[V\left[\dot{\ell}_n(\theta\,|\,\boldsymbol{x})\,\middle|\, T_n\right]\right], \tag{2.3.2}$$

where $\ell_n(\theta\,|\,\boldsymbol{x})$ means the log-likelihood for the n joint density and where the expectation in the right-hand side means the expectation by the marginal probability (density) function of T_n. Combining (2.3.1) and (2.3.2) implies that

$$\lim_{n\to\infty} E_{\hat{\theta}}\left[V\left[\dot{\ell}_n(\theta\,|\,\boldsymbol{x})\,\middle|\,\hat{\theta}\right]\right] \;=\; I(\theta)\,\Gamma_S(\theta)^2.$$

In this relation and (2.2.3), Efron attempted to show the counterexample with respect to the maximum likelihood estimator $\hat{\theta}$ of the trinomial distribution. We shall treat this counterexample in Chapter 4.

Chapter 3

Exponential family

We shall explain the definitions and properties of the exponential family and the curved exponential family.

3.1 Exponential family

Let $p_0(\boldsymbol{x})$ be a pivotal probability (density) function for the random vector $\boldsymbol{X}$ in $\mathbf{R}^k$. Let $\boldsymbol{\alpha}$ be a parameter on $\mathbf{R}^k$ and the parameter space $\mathcal{A}$ is defined as follows :

$$\mathcal{A} := \left\{ \boldsymbol{\alpha} \in \mathbf{R}^k : \int \exp\{\langle \boldsymbol{\alpha}, \, \boldsymbol{x} \rangle\} \, p_0(\boldsymbol{x}) \, d\,\boldsymbol{x} \; < \infty \right\}.$$

Note that if the pivotal is discrete then the above integration means the summation. We shall call $\mathcal{A}$ the parametric space. It is easy to check that the parametric space $\mathcal{A}$ is convex. For $\mathcal{A}$, we shall define the density of exponential family as follows :

$$f(\boldsymbol{x} : \boldsymbol{\alpha}) = \exp\{\langle \boldsymbol{\alpha}, \, \boldsymbol{x} \rangle \; - \; \psi(\boldsymbol{\alpha})\} \, p_0(\boldsymbol{x}),$$

where $\psi(\boldsymbol{\alpha})$ is the cumulant generating function, that is,

$$\psi(\boldsymbol{\alpha}) \;=\; \log \int \exp\{\langle \boldsymbol{\alpha}, \, \boldsymbol{x} \rangle\} \, p_0(\boldsymbol{x}) \, d\,\boldsymbol{x}.$$

Since the parametric $\mathcal{A}$ is determined by the existence of the moment generating function for $p_0(\boldsymbol{x})$, we have the interchangeability of integration and differentiation with respect to $\boldsymbol{\alpha}$, so that the derivatives of $\int f(\boldsymbol{x} : \boldsymbol{\alpha}) \, d\,\boldsymbol{x} \, = 1$ are

$$\begin{aligned} 0 &= \int (\boldsymbol{x} - \nabla\psi(\boldsymbol{\alpha})) \, f(\boldsymbol{x} : \boldsymbol{\alpha}) \, d\,\boldsymbol{x}, \\ 0 &= -\nabla' \, \nabla\psi(\boldsymbol{\alpha}) \; + \; \int (\boldsymbol{x} - \nabla\psi(\boldsymbol{\alpha})) \, (\boldsymbol{x} - \nabla\psi(\boldsymbol{\alpha}))' \, f(\boldsymbol{x} : \boldsymbol{\alpha}) \, d\,\boldsymbol{x}, \end{aligned}$$

where ∇ means the derivative with respect to $\boldsymbol{\alpha}$ and $'$ means the transpose. Thereby the expectation and variance are

$$E[\boldsymbol{X}] \;=\; \nabla\psi(\boldsymbol{\alpha}) \quad \text{and} \quad V[\boldsymbol{X}] \;=\; \nabla' \, \nabla\psi(\boldsymbol{\alpha}).$$

Let $\boldsymbol{\beta}$ be the expectation and let $\boldsymbol{\Sigma}$ the variance matrix, that is,

$$\boldsymbol{\beta}(\boldsymbol{\alpha}) \;=\; \nabla\psi(\boldsymbol{\alpha}) \quad \text{and} \quad \boldsymbol{\Sigma}(\boldsymbol{\alpha}) \;=\; \nabla'\,\nabla\psi(\boldsymbol{\alpha}).$$

When $\boldsymbol{\alpha} = \mathbf{0}$, they are equal to ones of $p_0(\boldsymbol{x})$, that is, $\boldsymbol{\beta}(\mathbf{0})$, $\boldsymbol{\Sigma}(\mathbf{0})$ are the expectation and the covariance of $p_0(\boldsymbol{x})$, respectively. Let $\mathcal{B}$ be the space of $\boldsymbol{\beta}(\boldsymbol{\alpha})$, that is,

$$\mathcal{B} \;=\; \{\boldsymbol{\beta}(\boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathcal{A}\}.$$

Though $\mathcal{A}$ is convex, $\mathcal{B}$ is not convex always. For this example, see Efron(1978). In order to escape a confusion, we shall not treat the trivial case which the variance $V[\boldsymbol{X}]$ is zero. Then, since the second derivative of $\psi(\boldsymbol{\alpha})$ is positive definite, it holds that $\psi(\boldsymbol{\alpha})$ is strictly convex, so that the correspondence between $\mathcal{A}$ and $\mathcal{B}$ is one-to-one.

Since the log-likelihood is

$$\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{x}) \;=\; \langle\boldsymbol{\alpha},\,\boldsymbol{x}\rangle \;-\; \psi(\boldsymbol{\alpha}) \;+\; \log p_0(\boldsymbol{x}),$$

the derivatives with respect to $\boldsymbol{\alpha}$ are

$$\begin{aligned} \nabla\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{x}) \;&=\; \boldsymbol{x} \;-\; \boldsymbol{\beta}(\boldsymbol{\alpha}), \\ \nabla'\,\nabla\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{x}) \;&=\; -\,\boldsymbol{\Sigma}(\boldsymbol{\alpha}). \end{aligned}$$

And the expectations and variances are

$$\begin{aligned} E[\,\nabla\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{X})\,] \;&=\; \mathbf{0}, & E[\,\nabla'\,\nabla\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{X})\,] \;&=\; -\boldsymbol{\Sigma}(\boldsymbol{\alpha}), \\ V[\,\nabla\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{X})\,] \;&=\; \boldsymbol{\Sigma}(\boldsymbol{\alpha}), & V[\,\nabla'\,\nabla\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{X})\,] \;&=\; \mathbf{0}, \end{aligned}$$

and the covariance is

$$Cov[\,\nabla\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{X}),\;\nabla'\,\nabla\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{X})\,] \;=\; \mathbf{0}.$$

Thereby the Fisher information is

$$I(\boldsymbol{\alpha}) \;\equiv\; V[\,\nabla\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{x})\,] \;=\; E[\,-\nabla'\,\nabla\ell(\boldsymbol{\alpha}\,|\,\boldsymbol{x})\,] \;=\; \boldsymbol{\Sigma}(\boldsymbol{\alpha}),$$

so that $I(\boldsymbol{\alpha})$ is positive definite.

We shall consider the maximum likelihood estimator(MLE) of exponential family for n i.i.d. observations. Since the MLE is determined by $\nabla\ell_n(\boldsymbol{\alpha}\,|\,\boldsymbol{x}) = \mathbf{0}$ given the observation $\boldsymbol{x}$, there exists $\hat{\boldsymbol{\alpha}}$ such that

$$n\,\overline{\boldsymbol{x}}_n \;-\; n\,\boldsymbol{\beta}(\hat{\boldsymbol{\alpha}}) \;=\; \mathbf{0},$$

where $\overline{\boldsymbol{x}}_n = \sum_{j=1}^{n} \boldsymbol{x}_j/n$. (See Theorem 7.1.3 in Appendices.) Thus the MLE is $\hat{\boldsymbol{\alpha}} = \boldsymbol{\beta}^{-1}(\overline{\boldsymbol{x}}_n)$. Since there exists only one $\boldsymbol{x}$ which satisfies $\nabla\ell_n(\hat{\boldsymbol{\alpha}}\,|\,\boldsymbol{x}) = \mathbf{0}$ given $\hat{\boldsymbol{\alpha}}$, there does not exist the information loss in this situation.

3.2 Curved Exponential family

We shall consider the curved exponential family which the exponential family is restricted as follows : the parametric $\boldsymbol{\alpha}$ is restricted by the parameter θ as

$$\{\boldsymbol{\alpha}(\theta) : \theta \in \Theta\} \ \subset \mathcal{A}.$$

Note that $\Theta \subset \mathbf{R}^r$. This probability density function is represented by

$$f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta)) \ = \ \exp\{\, \langle \boldsymbol{\alpha}(\theta), \boldsymbol{x} \rangle - \psi(\boldsymbol{\alpha}(\theta)) \,\} \cdot p_0(\boldsymbol{x}). \tag{3.2.1}$$

As the same way, we have the expectation and variance as follows :

$$\begin{aligned} \boldsymbol{\beta}(\theta) &= \boldsymbol{\beta}(\boldsymbol{\alpha}(\theta)) \ = \ E[\boldsymbol{X}] \ = \ \nabla\psi(\boldsymbol{\alpha}(\theta)), \\ \boldsymbol{\Sigma}(\theta) &= \boldsymbol{\Sigma}(\boldsymbol{\alpha}(\theta)) \ = \ V[\boldsymbol{X}] \ = \ \nabla'\,\nabla\psi(\boldsymbol{\alpha}(\theta)). \end{aligned}$$

The following relationship, which is one of the most characteristic property in the curved exponential family, is important which is derived by the differentiation of $\boldsymbol{\beta}(\theta)$ with respect to θ :

$$\dot{\boldsymbol{\beta}}(\theta) = \boldsymbol{\Sigma}(\theta)\, \dot{\boldsymbol{\alpha}}(\theta), \tag{3.2.2}$$

where the dot notation means the differentiation with respect to θ. And the relationship guarantees the local interchangeability between $\dot{\boldsymbol{\alpha}}(\theta)$ and $\dot{\boldsymbol{\beta}}(\theta)$ at θ. The differentiations of log-likelihood, that is, $\ell(\theta \,|\, \boldsymbol{x}) \ = \ \log f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta))$, with respect to θ are

$$\begin{aligned} \dot{\ell}(\theta \,|\, \boldsymbol{x}) &= \langle \dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{x} - \boldsymbol{\beta}(\theta) \rangle, \\ \ddot{\ell}(\theta \,|\, \boldsymbol{x}) &= \ddot{\boldsymbol{\alpha}}(\theta)'\, [\mathbf{I}_r \otimes \{\boldsymbol{x} - \boldsymbol{\beta}(\theta)\}\,] \ - \ \langle \dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\beta}}(\theta) \rangle, \end{aligned}$$

where the notation $\otimes$ means the Kronecker product. These differentiations imply the following relations;

$$\begin{aligned} E[\,\dot{\ell}(\theta \,|\, \boldsymbol{X})\,] &= \mathbf{0}, \\ E[\,\ddot{\ell}(\theta \,|\, \boldsymbol{X})\,] &= -\langle \dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\beta}}(\theta) \rangle \ = \ -\dot{\boldsymbol{\alpha}}(\theta)'\, \boldsymbol{\Sigma}(\theta)\, \dot{\boldsymbol{\alpha}}(\theta), \\ V[\,\dot{\ell}(\theta \,|\, \boldsymbol{X})\,] &= E[\,-\ddot{\ell}(\theta \,|\, \boldsymbol{X})\,] \ = \ \dot{\boldsymbol{\alpha}}(\theta)'\, \boldsymbol{\Sigma}(\theta)\, \dot{\boldsymbol{\alpha}}(\theta), \\ V[\,\ddot{\ell}(\theta \,|\, \boldsymbol{X})\,] &= \ddot{\boldsymbol{\alpha}}(\theta)'\, [\mathbf{I}_r \otimes \boldsymbol{\Sigma}(\theta)\,]\, \ddot{\boldsymbol{\alpha}}(\theta), \\ Cov[\,\ddot{\ell}(\theta \,|\, \boldsymbol{X}),\ \dot{\ell}(\theta \,|\, \boldsymbol{X})\,] &= \ddot{\boldsymbol{\alpha}}(\theta)'\, [\,\mathbf{I}_r \otimes \boldsymbol{\Sigma}(\theta)\,] \begin{bmatrix} \dot{\boldsymbol{\alpha}}^{(1)}(\theta) \\ \vdots \\ \dot{\boldsymbol{\alpha}}^{(r)}(\theta) \end{bmatrix}, \end{aligned}$$

where

$$\dot{\boldsymbol{\alpha}}^{(j)}(\theta) := \frac{\partial\, \boldsymbol{\alpha}(\theta)}{\partial \theta_j}, \qquad \text{that is,} \qquad \dot{\boldsymbol{\alpha}}(\theta) = \left[\, \dot{\boldsymbol{\alpha}}^{(1)}(\theta) \ \cdots \ \dot{\boldsymbol{\alpha}}^{(r)}(\theta) \,\right].$$

The Fisher information is

$$I(\theta) \ = \ \dot{\boldsymbol{\alpha}}(\theta)'\, \boldsymbol{\Sigma}(\theta)\, \dot{\boldsymbol{\alpha}}(\theta),$$

so that $I(\theta)$ is also positive definite.

We shall consider the maximum likelihood estimator of curved exponential family for n i.i.d. observations. Since the MLE is determined by the likelihood equation $\dot{\ell}_n(\theta \,|\, \boldsymbol{x}) = \mathbf{0}$ given the observation $\boldsymbol{x}$, there exists $\hat{\theta}$ such that

$$n\ \langle \dot{\boldsymbol{\alpha}}(\hat{\theta}),\, \overline{\boldsymbol{x}}_n - \boldsymbol{\beta}(\hat{\theta}) \rangle \ = \ \mathbf{0}.$$

For $\boldsymbol{x}$ which satisfies $\dot{\ell}_n(\hat{\theta}\,|\,\boldsymbol{x}) = \mathbf{0}$ given $\hat{\theta}$, we shall define a subset $L(\hat{\theta})$ as follows :

$$L(\hat{\theta}) \;=\; \{\, \overline{\boldsymbol{x}}_n : \dot{\ell}_n(\hat{\theta}\,|\,\boldsymbol{x}) = \mathbf{0}\,\}.$$

In the curved exponential family, $L(\hat{\theta})$ is orthogonal to $\boldsymbol{\alpha}(\hat{\theta})$, that is,

$$\dot{\boldsymbol{\alpha}}(\hat{\theta}) \quad \perp \quad \{\, \overline{\boldsymbol{x}}_n - \boldsymbol{\beta}(\hat{\theta})\,\} \quad \text{for any} \quad \overline{\boldsymbol{x}}_n \in L(\hat{\theta}).$$

The orthogonality is a characteristic in the curved exponential family.

We shall show the lemmas with respect to the MLE $\hat{\theta}$ in the curved exponential family. In general, we need to assume the following regularity conditions (For example, see Inagaki(1990)):

1. The parameter space Θ is an open (connected) set in $\mathbf{R}^r$.
2. The support $\{\boldsymbol{x} : f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta)) > 0\}$ of $f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta))$ does not depend on θ.
3. The joint density $f_n(\boldsymbol{x} : \boldsymbol{\alpha}(\theta))$ is three times continuously differentiable with respect to θ.
4. For a neighborhood $U(\theta)$ of any $\theta \in \Theta$, there exists a function $u(\boldsymbol{x} : \theta) \geq 0$ such that

   $$|\ell^{(3)}(\tau\,|\,\boldsymbol{x})| \leq u(\boldsymbol{x} : \theta) \quad \text{and} \quad E[\,u(\boldsymbol{X} : \theta)\,] < \infty,$$

   where $\tau \in U(\theta)$ and $\ell^{(3)}(\cdot)$ means the third differentiation.
5. For any $\theta \in \Theta$, there exists the Fisher information $I(\theta)$ which is positive definite and finite.

The following theorem is known with respect to the maximum likelihood estimator :

Theorem 3.2.1 *(Cramér) Under the regularity conditions, the maximum likelihood estimator (MLE) $\hat{\theta}$ is an asymptotically normal estimator with the asymptotic variance $I(\theta)^{-1}$, that is,*

$$\sqrt{n}(\hat{\theta} - \theta) \to N_r(\mathbf{0},\, I(\theta)^{-1}) \quad \textit{in law} \quad (n \to \infty).$$

In the curved exponential family, the above regularity conditions are satisfied. Thus, by the above Cramér Theorem, we have the following corollary :

Corollary 3.2.1 *The MLE $\hat{\theta}$ in the curved exponential family is always an asymptotically efficient estimator.* □

Chapter 4

Comment on Efron's Counterexample

4.1 Introduction

Efron(1975) showed by a counterexample that Fisher's information loss and Rao's second-order efficiency could be different in the multinomial case treated really by Fisher and Rao.

Our aim of the present chapter is to point out that we could not construct the valid form and explicit representation of Efron's counterexample because the contradictory demands are carried on the curved exponential family.

4.2 Efron's counterexample

We consider the trinomial distribution as the exponential family. Let $\beta_1,\ \beta_2$ be the probability of category $1, 2$ and let the domain of parameter $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ be

$$D = \left\{ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \ : \ 0 < \beta_1,\ 0 < \beta_2,\ \beta_1 + \beta_2 < 1 \right\},$$

where the probability of category 3 is $\beta_3 = 1 - \beta_1 - \beta_2$. For n independent trials, let n_1, n_2 be the number of occurrence of category $1, 2$, where the number of occurrence of category 3 is

$$n_3 = n - n_1 - n_2.$$

Then, we regard the probability function

$$\beta_1^{n_1}\ \beta_2^{n_2}\ \ (1 - \beta_1 - \beta_2)^{n - n_1 - n_2}$$

as the following two dimensional exponential family :

$$f_n(\boldsymbol{x} : \boldsymbol{\alpha}) = \exp[\ n\{\langle \boldsymbol{\alpha},\, \boldsymbol{x} \rangle - \psi(\boldsymbol{\alpha})\}\ ]$$

where $\boldsymbol{X} = (X_1,\, X_2)'$ is the vector of observation ratios of category $1, 2$:

$$x_1 = \frac{n_1}{n}, \quad x_2 = \frac{n_2}{n},$$

$\boldsymbol{\alpha} = (\alpha_1,\, \alpha_2)'$ is the vector of natural parameters :

$$\alpha_1 = \log \frac{\beta_1}{1-\beta_1-\beta_2}, \quad \alpha_2 = \log \frac{\beta_2}{1-\beta_1-\beta_2},$$

and

$$\langle \boldsymbol{\alpha},\, \boldsymbol{x} \rangle \;=\; \alpha_1\, x_1 \,+\, \alpha_2\, x_2$$

denotes the inner product of vectors $\boldsymbol{\alpha}$, $\boldsymbol{x}$. We note that

$$\psi(\boldsymbol{\alpha}) = \log(1 + e^{\alpha_1} + e^{\alpha_2})$$

is a convex function, the expectation parameter is

$$\text{(4.2.1)} \qquad \boldsymbol{\beta} = E\{\boldsymbol{X}\} = \nabla\psi(\boldsymbol{\alpha}) = \begin{pmatrix} \dfrac{\partial\psi(\boldsymbol{\alpha})}{\partial\alpha_1} \\[2ex] \dfrac{\partial\psi(\boldsymbol{\alpha})}{\partial\alpha_2} \end{pmatrix} = \begin{pmatrix} \dfrac{e^{\alpha_1}}{1+e^{\alpha_1}+e^{\alpha_2}} \\[2ex] \dfrac{e^{\alpha_2}}{1+e^{\alpha_1}+e^{\alpha_2}} \end{pmatrix},$$

and $n\,\boldsymbol{\Sigma}$ is the covariance matrix of $n\,(X_1,\, X_2)'$:

$$\begin{aligned} \boldsymbol{\Sigma} = \nabla\boldsymbol{\beta} = \nabla'\,\nabla\psi(\boldsymbol{\alpha}) \;&=\; \left[\frac{\partial^2\psi(\boldsymbol{\alpha})}{\partial\alpha_i\partial\alpha_j}\right] \qquad i,j = 1,2 \\ \text{(4.2.2)} \qquad &=\; \begin{pmatrix} \beta_1(1-\beta_1) & -\beta_1\beta_2 \\ -\beta_1\beta_2 & \beta_2(1-\beta_2) \end{pmatrix} = \boldsymbol{\Sigma}(\boldsymbol{\beta}), \end{aligned}$$

which is positive definite. Thus, we see that there is the one-to-one correspondence between the natural parameter $\boldsymbol{\alpha}$ and the expectation parameter $\boldsymbol{\beta}$.

Now, we define the curved exponential family as the exponential family with the expectation parameter vector indexified by one parameter :

$$\boldsymbol{\beta}(\theta) = \begin{pmatrix} \beta_1(\theta) \\ \beta_2(\theta) \end{pmatrix} \in D, \qquad \text{for } \theta \in \Theta,$$

which is assumed to belong to the C^1 class with respect to θ. According to the one-to-one correspondence between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the natural parameter vector $\boldsymbol{\alpha}$ is also indexified with the same parameter θ, which is denoted by $\{\boldsymbol{\alpha}(\theta) : \theta \in \Theta\}$. Then, we have the following fundamental equation :

$$\text{(4.2.3)} \qquad \dot{\boldsymbol{\beta}}(\theta) \;=\; \boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta),$$

where the dot mark means the differentiation with respect to θ, and $\boldsymbol{\Sigma}(\theta) = \boldsymbol{\Sigma}(\boldsymbol{\beta}(\theta))$ in (4.2.2) :

$$\text{(4.2.4)} \qquad \boldsymbol{\Sigma}(\theta) \;=\; \begin{pmatrix} \beta_1(\theta)(1-\beta_1(\theta)) & -\beta_1(\theta)\beta_2(\theta) \\ -\beta_1(\theta)\beta_2(\theta) & \beta_2(\theta)(1-\beta_2(\theta)) \end{pmatrix}.$$

In the sequel, we set up the Efron's counterexample which is a curved exponential family with the following special parameterization : We consider a union of half-lines emanating from the center point $\boldsymbol{c} = (-\sqrt{2}, -1)'$, and denote the one through a point $\boldsymbol{\beta}_0 = (1/3, 1/3)'$ by L_0

and the one with the angle θ from L_0 by L_θ . The angle between L_0 and the x_1 -axis is equal to A_0 :

$$A_0 = \arctan \frac{1 + \frac{1}{3}}{\sqrt{2} + \frac{1}{3}}.$$

We see that the parameter space Θ is located in such a way as :

$$\Theta \subseteq \{\theta \ : \ \theta_L < \theta < \theta_U\},$$

with

$$\theta_L := \arctan(1/(1+\sqrt{2})) - A_0, \quad \theta_U := \arctan(\sqrt{2}) - A_0.$$

Let us denote the unit directional vector of L_θ and the unit vector orthogonal to it by $\boldsymbol{\phi}_\theta,\ \boldsymbol{\varphi}_\theta$, respectively :

$$\boldsymbol{\phi}_\theta := \begin{pmatrix} \cos(\theta + A_0) \\ \sin(\theta + A_0) \end{pmatrix}, \quad \boldsymbol{\varphi}_\theta := \begin{pmatrix} -\sin(\theta + A_0) \\ \cos(\theta + A_0) \end{pmatrix},$$

where

$$\langle \boldsymbol{\phi}_\theta,\ \boldsymbol{\varphi}_\theta \rangle = 0 \quad \text{for any } \theta \in \Theta.$$

Let $\boldsymbol{\rho}_\theta$ be the unit directional vector of $\boldsymbol{\Sigma}(\theta)\boldsymbol{\varphi}_\theta$:

$$\boldsymbol{\rho}_\theta = \frac{\boldsymbol{\Sigma}(\theta)\boldsymbol{\varphi}_\theta}{|\boldsymbol{\Sigma}(\theta)\boldsymbol{\varphi}_\theta|}, \tag{4.2.5}$$

where the notation $|\cdot|$ means the Euclidean distance, and let B_θ be the angle between $\boldsymbol{\rho}_\theta$ and the line L_θ :

$$\cos B_\theta \ = \ \langle \boldsymbol{\rho}_\theta,\ \boldsymbol{\phi}_\theta \rangle. \tag{4.2.6}$$

Efron(1975) defined a curved exponential family by the following parameterization of the expectation parameter $\boldsymbol{\beta}(\theta)$:

$$\boldsymbol{\beta}(\theta) := \boldsymbol{\beta}(0) + \int_0^\theta h_\tau \ \boldsymbol{\rho}_\tau \ d\tau \tag{4.2.7}$$

where

$$h_\tau = \frac{|\boldsymbol{\beta}(\tau) - \boldsymbol{c}|}{\sin B_\tau}.$$

Note that h_θ is defined such that the infinitesimal variation of the angle of $\boldsymbol{\beta}(\theta)$ is equal to 1 :

$$\frac{\left|\dot{\boldsymbol{\beta}}(\theta)\right| \sin B_\theta}{|\boldsymbol{\beta}(\theta) - \boldsymbol{c}|} = 1, \tag{4.2.8}$$

which is the necessary condition for $\boldsymbol{\beta}(\theta)$ to be on the line L_θ and equivalently, to justify the above parameterization.

Since h_θ and $\boldsymbol{\rho}_\theta$ include $\boldsymbol{\beta}(\theta)$, we have the simultaneous differential equations of $\boldsymbol{\beta}(\theta)$:

$$\dot{\boldsymbol{\beta}}(\theta) = h_\theta \ \boldsymbol{\rho}_\theta. \tag{4.2.9}$$

which, together with (4.2.3), implies

$$\dot{\boldsymbol{\beta}}(\theta) \propto \boldsymbol{\Sigma}(\theta)\boldsymbol{\varphi}_\theta, \quad \dot{\boldsymbol{\alpha}}(\theta) \propto \boldsymbol{\varphi}_\theta.$$

(See the figure 4.1.) By the last fact, it is easy to see the following theorem.

Theorem 4.2.1 *Under the Efron's parameterization, the following three conditions are equivalent : for any θ,*

$$\left\{\begin{array}{l} (C1)\ \boldsymbol{\beta}(\theta) \in L_\theta, \\ (C2)\ \langle \dot{\boldsymbol{\alpha}}(\theta),\ \boldsymbol{\beta}(\theta) - \boldsymbol{c} \rangle \ = \ 0, \\ (C3)\ \{\boldsymbol{\beta} \ : \ \langle \dot{\boldsymbol{\alpha}}(\theta),\ \boldsymbol{\beta} - \boldsymbol{\beta}(\theta) \rangle \ = \ 0\} \ \subseteq \ L_\theta. \end{array}\right.$$

Efron asserts that $\boldsymbol{\beta}(\theta)$ is on the line L_θ. Therefore, it follows from the definition of the maximum likelihood estimator (MLE) $\hat{\theta}$:

$$\langle \dot{\boldsymbol{\alpha}}(\hat{\theta}),\ \boldsymbol{x} - \boldsymbol{\beta}(\hat{\theta}) \rangle \ = \ 0,$$

that the observation ratio vector $\boldsymbol{x}$ is on the line $L_{\hat{\theta}}$. Then, the slope of the line

$$\frac{x_2 - c_2}{x_1 - c_1},$$

is irrational, which does not allow for the line $L_{\hat{\theta}}$ to have any other observation ratio vector on the same line because of the rationality. Hence, it follows that the MLE $\hat{\theta}$ corresponds to the observation vector $\boldsymbol{x}$ by one to one and that

$$E[\ \boldsymbol{X} \mid \hat{\theta}\ ] = \boldsymbol{X}, \quad V[\ \boldsymbol{X} \mid \hat{\theta}\ ] = 0.$$

The vanish of the conditional variance of the observation given the MLE means that the information loss is not equal to the statistical curvature. This is the result of the Efron's counterexample.

However, it may be difficult to obtain the explicit form of the solution $\boldsymbol{\beta}(\theta)$ of the differential equations (4.2.9), which is written down in detail as follows:

$$\begin{pmatrix} \dot{\beta}_1(\theta) \\ \dot{\beta}_2(\theta) \end{pmatrix} = \sqrt{\frac{(\beta_1(\theta) - c_1)^2 + (\beta_2(\theta) - c_2)^2}{1 - \left\{ \dfrac{\langle \boldsymbol{\phi}_\theta, \begin{pmatrix} \beta_1(\theta)(1-\beta_1(\theta)) & -\beta_1(\theta)\beta_2(\theta) \\ -\beta_1(\theta)\beta_2(\theta) & \beta_2(\theta)(1-\beta_2(\theta)) \end{pmatrix} \boldsymbol{\varphi}_\theta \rangle}{\left| \begin{pmatrix} \beta_1(\theta)(1-\beta_1(\theta)) & -\beta_1(\theta)\beta_2(\theta) \\ -\beta_1(\theta)\beta_2(\theta) & \beta_2(\theta)(1-\beta_2(\theta)) \end{pmatrix} \boldsymbol{\varphi}_\theta \right|} \right\}^2}}$$

$$\times \frac{\begin{pmatrix} \beta_1(\theta)(1-\beta_1(\theta)) & -\beta_1(\theta)\beta_2(\theta) \\ -\beta_1(\theta)\beta_2(\theta) & \beta_2(\theta)(1-\beta_2(\theta)) \end{pmatrix} \boldsymbol{\varphi}_\theta}{\left| \begin{pmatrix} \beta_1(\theta)(1-\beta_1(\theta)) & -\beta_1(\theta)\beta_2(\theta) \\ -\beta_1(\theta)\beta_2(\theta) & \beta_2(\theta)(1-\beta_2(\theta)) \end{pmatrix} \boldsymbol{\varphi}_\theta \right|}.$$

4.3 Contradictory demand

Our aim is to show that the definition (4.2.7), which is derived from the simultaneous differential equations (4.2.9), is necessary but not sufficient for the parameterization to be justified and thus, that the three conditions $(C1), (C2), (C3)$ are shown to be equivalent under the parameterization, but any one of them is not shown to hold.

Theorem 4.3.1 *The Efron's parameterization is contradictory to the demand that $\boldsymbol{\beta}(\theta)$ is on the line L_θ.*

Proof : Recall (4.2.5) and (4.2.9) :

$$\dot{\boldsymbol{\beta}}(\theta) = h_\theta \ \boldsymbol{\rho}_\theta, \qquad \text{with} \quad \boldsymbol{\rho}_\theta = \frac{\boldsymbol{\Sigma}(\theta)\boldsymbol{\varphi}_\theta}{|\boldsymbol{\Sigma}(\theta)\boldsymbol{\varphi}_\theta|}. \tag{4.3.1}$$

Suppose $\boldsymbol{\beta}(\theta)$ be on the line L_θ, we have

$$\boldsymbol{\beta}(\theta) := \boldsymbol{c} + r_\theta \begin{pmatrix} \cos(\theta + A_0) \\ \sin(\theta + A_0) \end{pmatrix} \ = \ \boldsymbol{c} + r_\theta \, \boldsymbol{\phi}_\theta,$$

where r_θ be a positive function which belongs to the C^1 class and $r_0 := |\boldsymbol{\beta}(0) - \boldsymbol{c}|$. Then, the tangent vector of $\boldsymbol{\beta}(\theta)$ is the first derivative of $\boldsymbol{\beta}(\theta)$ with respect to θ, that is,

$$\dot{\boldsymbol{\beta}}(\theta) = \dot{r}_\theta \ \boldsymbol{\phi}_\theta + r_\theta \ \boldsymbol{\varphi}_\theta. \tag{4.3.2}$$

Therefore, from the comparison between (4.3.1) and (4.3.2), we have that

$$\begin{aligned} h_\theta &= \sqrt{r_\theta^2 + \dot{r}_\theta^2}, \\ \boldsymbol{\rho}_\theta &= \frac{\dot{r}_\theta}{\sqrt{r_\theta^2 + \dot{r}_\theta^2}} \ \boldsymbol{\phi}_\theta + \frac{r_\theta}{\sqrt{r_\theta^2 + \dot{r}_\theta^2}} \ \boldsymbol{\varphi}_\theta, \end{aligned} \tag{4.3.3}$$

and from (4.2.6) that

$$\begin{aligned} \cos B_\theta &= \langle \boldsymbol{\rho}_\theta, \boldsymbol{\phi}_\theta \rangle = \frac{\dot{r}_\theta}{\sqrt{r_\theta^2 + \dot{r}_\theta^2}}, \\ \sin B_\theta &= \frac{r_\theta}{\sqrt{r_\theta^2 + \dot{r}_\theta^2}}. \end{aligned}$$

This leads to the alternative representation of $\boldsymbol{\rho}_\theta$ (4.3.3) as follows :

$$\begin{aligned} \boldsymbol{\rho}_\theta &= \cos B_\theta \ \boldsymbol{\phi}_\theta + \sin B_\theta \ \boldsymbol{\varphi}_\theta \\ &= \begin{pmatrix} \cos(B_\theta + \theta + A_0) \\ \sin(B_\theta + \theta + A_0) \end{pmatrix} = \boldsymbol{\phi}_{\theta + B_\theta}. \end{aligned} \tag{4.3.4}$$

Consequently, from $\boldsymbol{\rho}_\theta$ of (4.3.1) and (4.3.4), we have the following identity :

$$\boldsymbol{\rho}_\theta = \frac{\boldsymbol{\Sigma}(\theta)}{|\boldsymbol{\Sigma}(\theta)\boldsymbol{\varphi}_\theta|} \ \boldsymbol{\varphi}_\theta = \boldsymbol{\phi}_{\theta + B_\theta} \qquad \text{for any} \qquad \theta \in \Theta.$$

Since $\boldsymbol{\varphi}_\theta$, $\boldsymbol{\phi}_{\theta + B_\theta}$ belong to the unit circle $S^1 = \{\boldsymbol{u} \in \mathbf{R}^2 : \ |\boldsymbol{u}| = 1\}$, the type of the variance matrix $\boldsymbol{\Sigma}(\theta)$ must be as shown :

$$\frac{\boldsymbol{\Sigma}(\theta)}{|\boldsymbol{\Sigma}(\theta)\boldsymbol{\varphi}_\theta|} = \begin{pmatrix} \cos(\frac{\pi}{2} - B_\theta) & \sin(\frac{\pi}{2} - B_\theta) \\ -\sin(\frac{\pi}{2} - B_\theta) & \cos(\frac{\pi}{2} - B_\theta) \end{pmatrix} = \begin{pmatrix} \sin B_\theta & \cos B_\theta \\ -\cos B_\theta & \sin B_\theta \end{pmatrix}. \tag{4.3.5}$$

By the comparison of elements of $\mathbf{\Sigma}(\theta)$ in (4.2.4) and (4.3.5), we have

$$\begin{cases} \beta_1(\theta)(1-\beta_1(\theta)) = \beta_2(\theta)(1-\beta_2(\theta)), \\ \\ \beta_1(\theta)\beta_2(\theta) = -\beta_1(\theta)\beta_2(\theta). \end{cases} \tag{4.3.6}$$

This leads to the possible values of $\boldsymbol{\beta}(\theta)$:

$$\boldsymbol{\beta}(\theta) = \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}, \tag{4.3.7}$$

and then, to the contradiction : $\mathbf{\Sigma}(\theta) = \mathbf{0}$. Thus, we conclude that $\boldsymbol{\beta}(\theta)$ is not on the line L_θ for all $\theta \in \Theta$. □

Theorem 4.3.1 means that Efron's counterexample is invalid as the counterexample which attempts to demonstrate the gap between the information loss and the sufficiency with respect to the maximum likelihood estimator in discrete distributions.

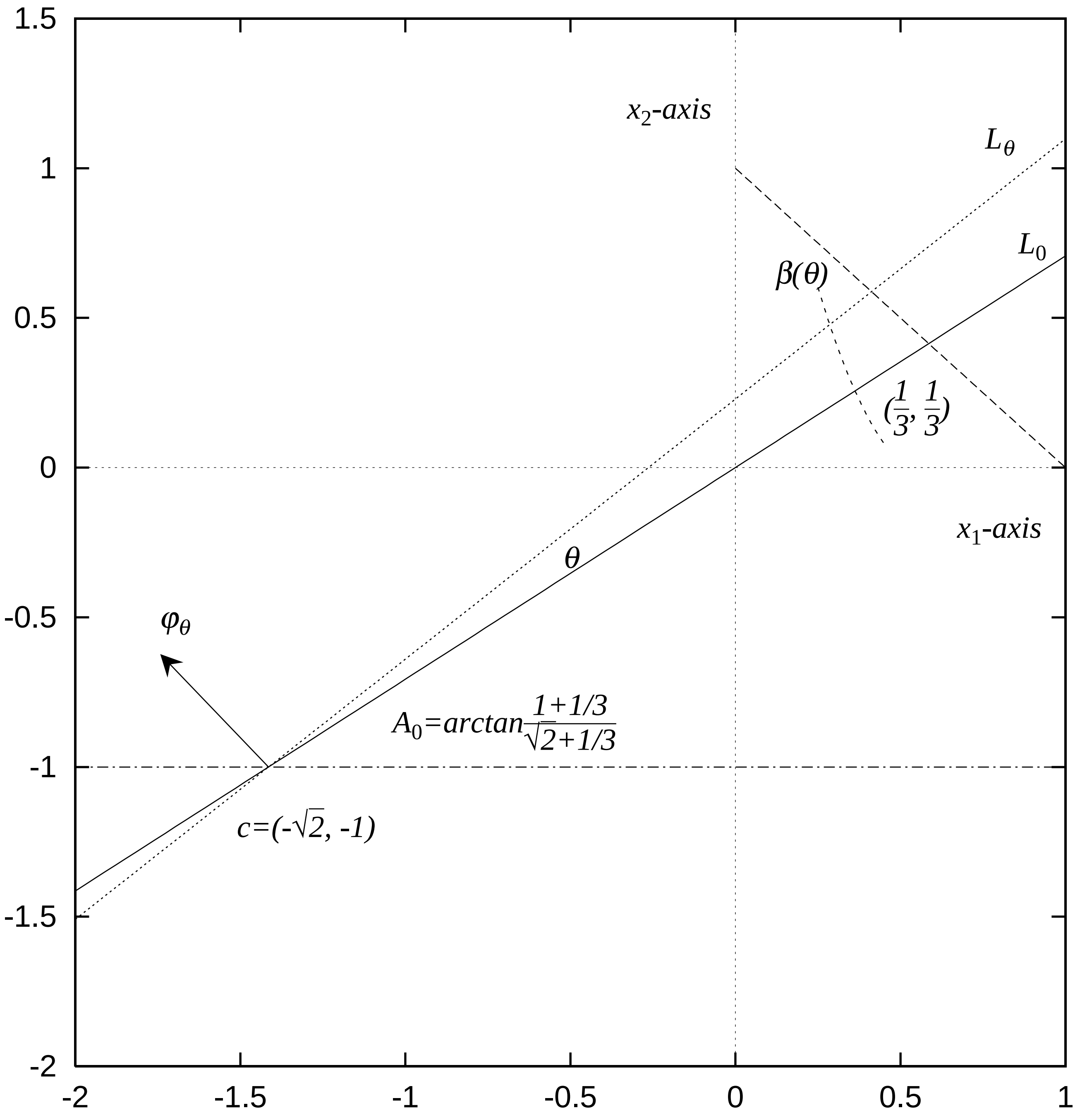

Figure 4.1: Figure of Counterexample

Chapter 5

Exact Information Loss in Fisher's Circle Model

5.1 Introduction

The circle model is the simplest and best one in order to illustrate the information loss, and thus, is often referred to in many papers. However, these amounts with respect to the distribution of the length given the angle are calculated asymptotically but not exactly. Also the distribution of the angle given the radius is well known in detail as the von Mises distribution. The main purpose of the present chapter is to calculate the distribution of the length given the angle in detail, which enable us to have the exact theory of the information loss by using the conditional variance, and to refine the geometric structure and asymptotical relation between information loss and statistical curvature.

5.2 Fisher's circle model

Two dimensional random vector $\boldsymbol{X} = (X_1, X_2)'$ is distributed to the normal distribution with mean vector $\boldsymbol{\alpha}$ and covariance matrix $\mathbf{I}$, the unit matrix, i.e., $N_2(\boldsymbol{\alpha}, \mathbf{I})$. Its density function is written in the exponential type :

$$\begin{aligned} f(\boldsymbol{x} : \boldsymbol{\alpha}) &= \exp\left\{\alpha_1 x_1 + \alpha_2 x_2 - \frac{1}{2}(\alpha_1^2 + \alpha_2^2)\right\} \; \varphi(x_1)\,\varphi(x_2) \\ &= \exp\left\{\boldsymbol{\alpha}'\boldsymbol{x} - \frac{1}{2}|\boldsymbol{\alpha}|^2\right\} \; p_0(\boldsymbol{x}), \qquad (5.2.1) \end{aligned}$$

where $\varphi(x)$ is the density function of the standard normal distribution and $p_0(\boldsymbol{x})$ is the pivotal density function :

$$\begin{aligned} \varphi(x) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \\ p_0(\boldsymbol{x}) &= f(\boldsymbol{x} : \mathbf{0}) \; = \; \varphi(x_1)\,\varphi(x_2). \end{aligned}$$

Let us consider the Fisher's circle model defined by the above normal family with mean vector $\boldsymbol{\alpha}$ on a circle :

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}(\theta) = \rho \boldsymbol{e}(\theta), \qquad \boldsymbol{e}(\theta) = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}.$$

where the radius ρ is known and the angle θ is unknown in $\Theta = [0, 2\pi)$, denoting a vector with length 1 and angle θ by $\boldsymbol{e}(\theta)$. Then, we have the simplest curved exponential type of density :

$$f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta)) = \exp\left\{ \rho \boldsymbol{e}(\theta)'\boldsymbol{x} - \frac{\rho^2}{2} \right\} \; p_0(\boldsymbol{x}).$$

Denote the differential with respect to θ by " $\dot{}$ ", for example, as follows :

$$\begin{aligned} \dot{\boldsymbol{e}}(\theta) = \frac{\partial}{\partial\theta}\boldsymbol{e}(\theta) &= \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix} = \begin{pmatrix} \cos(\frac{\pi}{2}+\theta) \\ \sin(\frac{\pi}{2}+\theta) \end{pmatrix} = \boldsymbol{e}(\frac{\pi}{2}+\theta) \\ \ddot{\boldsymbol{e}}(\theta) = \frac{\partial^2}{\partial\theta^2}\boldsymbol{e}(\theta) &= \begin{pmatrix} -\cos\theta \\ -\sin\theta \end{pmatrix} = -\boldsymbol{e}(\theta). \end{aligned}$$

The derivatives up to the second order of the log-likelihood with respect to θ are :

$$\begin{aligned} \frac{\partial}{\partial\theta}\log f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta)) &= \rho\, \dot{\boldsymbol{e}}(\theta)'\boldsymbol{x} = \rho\, \boldsymbol{e}(\frac{\pi}{2}+\theta)'\boldsymbol{x}, \\ \frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta)) &= \rho\, \ddot{\boldsymbol{e}}(\theta)'\boldsymbol{x} = -\rho\, \boldsymbol{e}(\theta)'\boldsymbol{x}. \end{aligned}$$

These expectations are :

$$\begin{aligned} E\left\{\frac{\partial}{\partial\theta}\log f(\boldsymbol{X} : \boldsymbol{\alpha}(\theta))\right\} &= \rho^2\, \boldsymbol{e}(\frac{\pi}{2}+\theta)'\boldsymbol{e}(\theta) = 0, \\ E\left\{\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X} : \boldsymbol{\alpha}(\theta))\right\} &= -\rho^2\, \boldsymbol{e}(\theta)'\boldsymbol{e}(\theta) = -\rho^2. \end{aligned}$$

Therefore, we have the Fisher information $I(\theta) = \rho^2$, which is independent of θ.

Let us transform the statistic vector $\boldsymbol{X}$ to length and angle (R, T), $R \geq 0$, $T \in \theta$ in the polar coordinates :

(5.2.2) $$\boldsymbol{X} = \left(\begin{array}{c} X_1 \\ X_2 \end{array} \right) = \left(\begin{array}{c} R \cos T \\ R \sin T \end{array} \right) = R\; \boldsymbol{e}(T).$$

Then, we have the joint density function of (R, T) :

(5.2.3) $$f_\theta(r,\, t) \equiv \frac{r}{2\pi}\; \exp\left\{ \rho\, r\; \cos(t-\theta) - \frac{r^2}{2} - \frac{\rho^2}{2} \right\}.$$

Since, for $m = 0, 1, 2, \ldots$,

$$\begin{aligned} \int_0^{2\pi} \cos^{2m} t\; dt &= \frac{1}{2^{2m}}\; {}_{2m}\mathrm{C}_m\; 2\pi, \\ \int_0^{2\pi} \cos^{2m+1} t\; dt &= 0, \end{aligned}$$

by the term-wise integral of the series of a exponential function :

$$e^{a\cos t} = \sum_{m=0}^{\infty} \frac{a^m}{m!} \cos^m t,$$

we have the modified Bessel function :

$$\begin{aligned} I_0(a) &= \frac{1}{2\pi}\int_0^{2\pi} e^{a\cos t}\, dt \\ &= \sum_{m=0}^{\infty} \frac{a^m}{m!}\frac{1}{2\pi}\int_0^{2\pi} \cos^m t\, dt = \sum_{m=0}^{\infty} \frac{a^{2m}}{(2m)!}\frac{1}{2^{2m}}\ {}_{2m}\mathrm{C}_m \\ &= \sum_{m=0}^{\infty} \left(\frac{a}{2}\right)^{2m} \frac{1}{m!\Gamma(m+1)}, \end{aligned}$$

where ${}_{2m}\mathrm{C}_m$ means the combination of $2m$ things m at a time and $\Gamma(\cdot)$ is the gamma function. In the similar way, the joint density $f_\theta(r,\, t)$ is decomposed into the sum of even terms and the one of odd terms in the expansion of the exponential function of its cross term :

$$\begin{aligned} f_\theta(r,\, t) &= \frac{r}{2\pi}\exp\left\{-\frac{r^2}{2}-\frac{\rho^2}{2}\right\}\sum_{m=0}^{\infty}\frac{(\rho r\ \cos(t-\theta))^{2m}}{(2m)!} \\ &\quad + \frac{r}{2\pi}\exp\left\{-\frac{r^2}{2}-\frac{\rho^2}{2}\right\}\sum_{m=0}^{\infty}\frac{(\rho r\ \cos(t-\theta))^{2m+1}}{(2m+1)!} \\ &\equiv f_\theta^e(r,\, t) + f_\theta^o(r,\, t) \qquad \text{(say)}. \end{aligned} \tag{5.2.4}$$

These are represented as Poisson mixture distributions of chi-distributions as follows :

$$\begin{aligned} f_\theta^e(r,\, t) &= \sum_{m=0}^{\infty} \exp\left(-\frac{\rho^2}{2}\right)\frac{\left(\frac{\rho^2}{2}\right)^m}{m!}\,\frac{1}{\Gamma(m+1)2^{m+1}}\,2\,r^{2m+1}\exp\left(-\frac{r^2}{2}\right) \\ &\quad \cdot 2^{2m}\ \frac{1}{{}_{2m}\mathrm{C}_m}\ \frac{1}{2\pi}\cos^{2m}(t-\theta) \\ &= \sum_{m=0}^{\infty} q_m\ \chi_{2m+2}\ 2^{2m}\ \frac{1}{{}_{2m}\mathrm{C}_m}\ \frac{1}{2\pi}\ \cos^{2m}(t-\theta), \end{aligned} \tag{5.2.5}$$

$$\begin{aligned} f_\theta^o(r,\, t) &= \frac{\rho}{\sqrt{2\pi}}\sum_{m=0}^{\infty} \exp\left(-\frac{\rho^2}{2}\right)\frac{\left(\frac{\rho^2}{2}\right)^m}{m!} \\ &\quad \cdot \frac{1}{\Gamma(\frac{2m+3}{2})\,2^{\frac{2m+3}{2}}}\,2\,r^{2m+2}\exp\left(-\frac{r^2}{2}\right)\cos^{2m+1}(t-\theta) \\ &= \frac{\rho}{\sqrt{2\pi}}\sum_{m=0}^{\infty} q_m\,\chi_{2m+3}\ \cos^{2m+1}(t-\theta), \end{aligned} \tag{5.2.6}$$

where $\{q_m\}$ is the Poisson probability function of $P_o(\lambda)$:

$$q_m = \exp(-\lambda)\frac{\lambda^m}{m!}, \qquad \lambda = \frac{\rho^2}{2},$$

and the density function of the chi-distribution of degree of freedom m, χ_m, is :

$$2\,\frac{1}{\Gamma(\frac{m}{2})2^{\frac{m}{2}}}\,r^{m-1}\,\exp\left\{-\frac{r^2}{2}\right\}.$$

Denote the log-likelihood function by

$$\ell(\theta) \equiv \log f_\theta(r,\,t),$$

then we have :

$$\begin{aligned}\dot{\ell}(\theta) &= \rho\, r\ \sin(t-\theta),\\ \ddot{\ell}(\theta) &= -\rho\, r\ \cos(t-\theta).\end{aligned}$$

Therefore, it follows from the likelihood equation $\dot{\ell}(\theta)=0$ that the angle T is the maximum likelihood estimator of θ. By the facts that

$$\begin{aligned}\int_0^{2\pi} \rho^2\, r^2\, \sin^2(t-\theta)\ f_\theta^o(r,\,t)\,dr\,dt &= 0\\ \text{and}\quad \int_0^{2\pi} \rho\, r\ \cos(t-\theta)\ f_\theta^e(r,\,t)\,dr\,dt &= 0,\end{aligned}$$

we have, again, the Fisher information in the following two ways :

$$\begin{aligned}I(\theta) = E\{(\dot{\ell}(\theta))^2\} &= \rho^2\int_0^\infty\int_0^{2\pi}\{r^2 - r^2\,\cos^2(t-\theta)\}\,f_\theta^e(r,\,t)\,dt\,dr\\
&= \rho^2\sum_{m=0}^\infty q_m\,\frac{1}{\Gamma(m+1)2^{m+1}}\int_0^\infty 2r^{2m+3}\,\exp(-\frac{r^2}{2})\,dr\\
&\quad -\rho^2\sum_{m=0}^\infty q_m\,\frac{1}{\Gamma(m+1)2^{m+1}}\int_0^\infty 2r^{2m+3}\,\exp(-\frac{r^2}{2})\,dr\\
&\quad \cdot 2^{2m}\,\frac{1}{{}_{2m}\mathrm{C}_m}\,\frac{1}{2\pi}\int_0^{2\pi}\cos^{2m+2}(t-\theta)\,dt\\
&= \rho^2\sum_{m=0}^\infty(2m+2)\,q_m - \rho^2\sum_{m=0}^\infty(2m+1)\,q_m = \rho^2.\end{aligned}$$

$$\begin{aligned}I(\theta) = -E\{\ddot{\ell}(\theta)\} &= \int_0^\infty\int_0^{2\pi}\rho\, r\ \cos(t-\theta)\,f_\theta^o(r,\,t)\,dt\,dr\\
&= \sum_{m=0}^\infty\int_0^\infty \exp\left\{-\frac{r^2}{2}-\frac{\rho^2}{2}\right\}\frac{(\rho r)^{2m+3}}{(2m+1)!}\,dr\,\frac{1}{2\pi}\int_0^{2\pi}\cos^{2m+2}(t-\theta)\,dt\\
&= \sum_{m=1}^\infty(2m)\,\exp\left(-\frac{\rho^2}{2}\right)\frac{\left(\frac{\rho^2}{2}\right)^m}{m!} = 2\sum_{m=0}^\infty m\,q_m\ =\ 2\lambda\ =\ \rho^2.\end{aligned}$$

The marginal density function of length R which is independent of θ is easily obtained by the integration of a periodic function :

$$\begin{aligned}\tilde{h}(r) &= \int_0^{2\pi} f_\theta(r,\,t)\,dt = r\,\exp\left\{-\frac{r^2}{2}-\frac{\rho^2}{2}\right\}\frac{1}{2\pi}\int_0^{2\pi}\exp\{\rho\, r\,\cos t\}dt\\
&= r\,\exp\left\{-\frac{r^2}{2}-\frac{\rho^2}{2}\right\}I_0(\rho\, r)\ =\ \sum_{m=0}^\infty q_m\,\chi_{2m+2}. \qquad (5.2.7)\end{aligned}$$

The last equation means that R is distributed with the non-central chi-distribution with degree of freedom 2 and non-central parameter ρ^2 : $R \sim \chi_2(\rho^2)$. Therefore the conditional density of angle T given length R is well known to be one of von Mises distribution $M(\theta, \rho r)$:

$$\tilde{g}_\theta(t\,|\,r) = \frac{f_\theta(r,\,t)}{\tilde{h}(r)} = \frac{\exp\{\rho\, r\, \cos(t-\theta)\}}{2\pi I_0(\rho\, r)}, \tag{5.2.8}$$

which implies the ancillarity of R to θ.

5.3 The conditional density of length R given angle T

Although we have the properties of the conditional distribution of angle T given length R from many studies of von Mises distribution, we have little about properties of the conditional distribution length R given angle T. Our main purpose in this chapter is to investigate exact and asymptotic properties about the Fisher informations of the marginal distribution of angle T and the conditional distribution of length R given it.

Lemma 5.3.1 *For any nonnegative integer k and real number u, let*

$$\mathcal{H}_k(u) = \int_0^\infty e^{ur}\, r^k\, e^{-\frac{r^2}{2}}\, dr.$$

Then, the following equations holds :

$$\begin{aligned}
&(5.3.1) && \mathcal{H}_k(u) &=&\ u\ \mathcal{H}_{k-1}(u) + (k-1)\mathcal{H}_{k-2}(u), \qquad \textit{for} \quad k \geq 2,\\
&(5.3.2) && \mathcal{H}_1(u) &=&\ 1 + u\ \mathcal{H}_0(u) = \varphi(u)^{-1}\{\varphi(u) + u\Phi(u)\},\\
&(5.3.3) && \mathcal{H}_0(u) &=&\ \varphi(u)^{-1}\Phi(u),
\end{aligned}$$

where $\varphi(u)$ and $\Phi(u)$ are the standard normal density and distribution functions, respectively, and $\mathcal{H}_1(u)$ is remarked to be the moment generating function of the chi-distribution with degree of freedom 2.

Proof : Equation (5.3.1) is easy to be seen by integral by part.
Equation (5.3.3) is shown as follows :

$$\begin{aligned}
\mathcal{H}_0(u) &= \int_0^\infty e^{ur}\, e^{-\frac{r^2}{2}}\, dr = e^{\frac{u^2}{2}} \int_0^\infty e^{-\frac{(r-u)^2}{2}}\, dr \\
&= \sqrt{2\pi} e^{\frac{u^2}{2}}\, \{1 - \Phi(-u)\} = \varphi(u)^{-1}\, \Phi(u).
\end{aligned}$$

Equation (5.3.2) follows from equations (5.3.1) and (5.3.3). □

Now, we have the marginal density of angle T and the conditional density of length R given T, as follows :

$$\begin{aligned}
g_\theta(t) &\equiv \int_0^\infty f_\theta(r,\,t)\, dr \\
&= \exp\left\{-\frac{\rho^2}{2}\right\} \frac{1}{2\pi} \int_0^\infty \exp\{\rho\cos(t-\theta)\, r\}\, r\, \exp\left\{-\frac{r^2}{2}\right\} dr
\end{aligned}$$

$$= \frac{1}{2\pi} \exp\left\{-\frac{\rho^2}{2}\right\} \mathcal{H}_1(\rho\cos(t-\theta)), \tag{5.3.4}$$

$$h_\theta(r \,|\, t) \equiv \frac{f_\theta(r,\, t)}{g_\theta(t)} = \frac{\exp\{\rho\, r\, \cos(t-\theta)\}\, r\, \exp\{-\frac{r^2}{2}\}}{\mathcal{H}_1(\rho\cos(t-\theta))}. \tag{5.3.5}$$

Theorem 5.3.1 *Let $a = \rho\, \cos(t-\theta)$. Then the following representations are obtained of the conditional mean and variance of length R given angle T :*

$$E[R \,|\, T=t] = \int_0^\infty r\, h_\theta(r \,|\, t)\, dr = \frac{\mathcal{H}_2(a)}{\mathcal{H}_1(a)}, \tag{5.3.6}$$

$$V[R \,|\, T=t] = 1 + \frac{1}{\mathcal{H}_1(a)} - \left(\frac{\mathcal{H}_0(a)}{\mathcal{H}_1(a)}\right)^2 \le 1. \tag{5.3.7}$$

Proof : It is easy to check the equation of conditional mean.
Since

$$E[R^2 \,|\, T] = \int_0^\infty r^2\, h_\theta(r|t)\, dr = \frac{\mathcal{H}_3(a)}{\mathcal{H}_1(a)},$$

we have, from (3) of Lemma 5.3.1,

$$\begin{aligned} V[R \,|\, T=t] &= E[R^2 \,|\, T=t] - (E[R \,|\, T=t])^2 = \frac{\mathcal{H}_3(a)}{\mathcal{H}_1(a)} - \left(\frac{\mathcal{H}_2(a)}{\mathcal{H}_1(a)}\right)^2 \\ &= \frac{a\{a\mathcal{H}_1(a) + \mathcal{H}_0(a)\} + 2\mathcal{H}_1(a)}{\mathcal{H}_1(a)} - \left(\frac{a\mathcal{H}_1(a) + \mathcal{H}_0(a)}{\mathcal{H}_1(a)}\right)^2 \\ &= \frac{\mathcal{H}_1(a) + 1 + 2a\mathcal{H}_0(a)}{\mathcal{H}_1(a)} - \frac{2a\mathcal{H}_1(a)\mathcal{H}_0(a) + \mathcal{H}_0(a)^2}{\mathcal{H}_1(a)^2} \\ &= 1 + \frac{1}{\mathcal{H}_1(a)} - \left(\frac{\mathcal{H}_0(a)}{\mathcal{H}_1(a)}\right)^2. \end{aligned}$$

On the other hand, we see, by Cauchy-Shwartz's inequality,

$$\begin{aligned} \mathcal{H}_1(a)^2 &= \left(\int_0^\infty e^{ar}\, r\, e^{-\frac{r^2}{2}}\, dr\right)^2 \\ &\le \int_0^\infty e^{ar}\, e^{-\frac{r^2}{2}}\, dr \int_0^\infty e^{ar}\, r^2\, e^{-\frac{r^2}{2}}\, dr \\ &= \mathcal{H}_0(a)\mathcal{H}_2(a) = \mathcal{H}_0(a)\{a\mathcal{H}_1(a) + \mathcal{H}_0(a)\} \\ &= \{\mathcal{H}_1(a) - 1\}\mathcal{H}_1(a) + \mathcal{H}_0(a)^2. \end{aligned}$$

That is,

$$\mathcal{H}_1(a) \le \mathcal{H}_0(a)^2.$$

This leads to the following inequality :

$$V[R \,|\, T=t] = 1 + \frac{1}{\mathcal{H}_1(a)} - \left(\frac{\mathcal{H}_0(a)}{\mathcal{H}_1(a)}\right)^2 \le 1. \qquad \square$$

Theorem 5.3.2 *According to the factorization of joint density $f_\theta(r,t)$ into the marginal $g_\theta(t)$ and the conditional $h_\theta(r|t)$:*

(5.3.8)
$$f_\theta(r,\,t) = g_\theta(t)\,h_\theta(r\,|\,t),$$

the Fisher information of joint density is decomposed into the sum of those of marginal and conditional densities :

(5.3.9)
$$I_f = I_g + I_h,$$

where each Fisher information is independent of parameter θ :

$$\begin{aligned} I_f &= \rho^2, \\ I_h &= \rho^2\, E_T\{\sin^2(T-\theta)V[R\,|\,T]\} \\ &= \rho^2\,\exp(-\frac{\rho^2}{2})\frac{1}{2\pi} \\ &\quad \cdot \int_0^{2\pi} \sin^2 t\left[1+\frac{1}{\mathcal{H}_1(\rho\,\cos t)} - \left(\frac{\mathcal{H}_0(\rho\,\cos t)}{\mathcal{H}_1(\rho\,\cos t)}\right)^2\right]\mathcal{H}_1(\rho\,\cos t)\,dt. \end{aligned}$$

Proof : The existence of the moment generating function $\mathcal{H}_1(u)$ for any real number u guarantees the exchangeability of the differential and integral in the following way :

$$\begin{aligned} \frac{\partial}{\partial\theta}\log g_\theta(t) &= \frac{\dot{\mathcal{H}}_1(\rho\,\cos(t-\theta))}{\mathcal{H}_1(\rho\,\cos(t-\theta))} \\ &= \frac{\int_0^\infty \exp\{\rho\, r\,\cos(t-\theta)\}\,\rho\, r\,\sin(t-\theta)\, r\,\exp\{-\frac{r^2}{2}\}\,dr}{\mathcal{H}_1(\rho\cos(t-\theta))} \\ &= \int_0^\infty \rho\, r\,\sin(t-\theta)\,h_\theta(r\,|\,t)\,dr \;=\; E[\dot{\ell}(\theta)\,|\,T=t]. \end{aligned}$$

The last equation shows that the differential of log-likelihood of the marginal distribution of angle T is equal to the conditional mean of that of the joint distribution given T. This leads to

$$\frac{\partial}{\partial\theta}\log h_\theta(r\,|\,t) = \frac{\partial}{\partial\theta}\log f_\theta(r,\,t) - \frac{\partial}{\partial\theta}\log g_\theta(t) = \dot{\ell}(\theta) - E[\dot{\ell}(\theta)\,|\,T].$$

Therefore, we have

$$I_h(\theta) = E_T\left\{\left(\frac{\partial}{\partial\theta}\log h_\theta(r\,|\,t)\right)^2\right\} = E_T\{V[\dot{\ell}(\theta)\,|\,T]\}$$

and

$$I_f(\theta) = I_g(\theta) + I_h(\theta).$$

Now, we already obtained the Fisher information of the joint density $I_f(\theta) = \rho^2$ in the previous section. By the equation

$$\dot{\ell}(\theta) = \rho\, r\,\sin(t-\theta),$$

we have the Fisher information of the conditional density h_θ :

$$I_h(\theta) = E_T\{\,\rho^2\,\sin^2(T-\theta)\,V[R\,|\,T]\,\}.$$

The last equation of I_h in this theorem follows from the formula (5.3.4) and the equation (5.3.7). At last, three informations $I_f(\theta)$, $I_g(\theta)$, and $I_h(\theta)$ are independent of θ, so that the proof is completed. □

Fisher proposes to utilize the difference between the Fisher informations of the original density of sample and the marginal density of an estimator in order to evaluate the efficiency of the estimator, and Fisher called it the information loss of the estimator. Theorem 5.3.2 means that the information loss of T is equal to the Fisher information of the conditional density of R given T. On the other hand, we have the corresponding expansion of the marginal density of T by the term-wise integral of the expansion of joint density (5.2.4),

$$g_\theta(t) = \int_0^{2\pi} f_\theta^e(r,\,t)\,dr + \int_0^{2\pi} f_\theta^o(r,\,t)\,dr \equiv g_\theta^e(t) + g_\theta^o(t), \tag{5.3.10}$$

where

$$\begin{aligned} g_\theta^e(t) &= \int_0^{2\pi} f_\theta^e(r,\,t)\,dr = \sum_{m=0}^{\infty} q_m\,2^{2m}\,\frac{1}{{}_{2m}\mathrm{C}_m}\,\frac{1}{2\pi}\,\cos^{2m}(t-\theta) \\ g_\theta^o(t) &= \int_0^{2\pi} f_\theta^o(r,\,t)\,dr = \frac{\rho}{\sqrt{2\pi}}\sum_{m=0}^{\infty} q_m\,\cos^{2m+1}(t-\theta). \end{aligned}$$

This formula (5.3.10) enables us to calculate the main part of the information loss in Theorem 5.3.2.

Lemma 5.3.2 *The main part of the information loss is calculated as follows :*

$$E\{\rho^2\sin^2(T-\theta)\} = 1 - \exp\left(-\frac{\rho^2}{2}\right).$$

Proof : Since

$$\int_0^{2\pi} \cos^2(t-\theta)\,g_\theta^o(t)\,dt \;=\; 0,$$

we have

$$\begin{aligned} E\{\cos^2(T-\theta)\} &= \int_0^{2\pi} \cos^2(t-\theta)\,g_\theta^e(t)\,dt \\ &= \sum_{m=0}^{\infty} q_m\,2^{2m}\,\frac{1}{{}_{2m}\mathrm{C}_m}\,\frac{1}{2\pi}\int_0^{2\pi}\cos^{2m+2}t\,dt \\ &= \sum_{m=0}^{\infty} q_m\,2^{2m}\,\frac{1}{{}_{2m}\mathrm{C}_m}\,2^{-2(m+1)}\;{}_{2(m+1)}\mathrm{C}_{m+1} \\ &= \sum_{m=0}^{\infty} q_m\,\frac{2m+1}{2(m+1)} = 1 - \frac{1}{2\lambda}\sum_{m=0}^{\infty} q_{m+1} = 1 - \frac{1-e^{-\lambda}}{2\lambda}, \end{aligned}$$

and thus,

$$E\{\sin^2(T-\theta)\} = \frac{1-e^{-\lambda}}{2\lambda} = \frac{1}{\rho^2}\left\{1 - \exp\left(-\frac{\rho^2}{2}\right)\right\}.$$

This leads to the conclusion of the lemma. □

Theorem 5.3.3 *The exact information loss is calculated as follows :*

$$\begin{aligned} I_h &= I_f - I_g \\ &= 1 - \exp\left(-\frac{\rho^2}{2}\right) + \frac{1}{2}\rho^2 \exp\left(-\frac{\rho^2}{2}\right) \\ &\quad - \rho^2 \int_0^{2\pi} \sin^2 t \left(\frac{\Phi(\rho\cos t)}{\varphi(\rho\cos t) + \rho\cos t\,\Phi(\rho\cos t)}\right)^2 g_0(t)\,dt, \end{aligned}$$

where

$$g_0(t) = g_\theta(t)\,|_{\theta=0} = \varphi(\rho\sin t)\,\{\varphi(\rho\cos t) + \rho\,\cos t\,\Phi(\rho\cos t)\}.$$

Proof
It is easy to see that the equation of $g_0(t)$ means the last equation about the marginal density at $\theta = 0$. Therefore, the equation

$$\frac{1}{2\pi}\int_0^{2\pi} \sin^2 t\,dt \;=\; \frac{1}{2},$$

and Lemma 5.3.2 lead to

$$\int_0^{2\pi} \rho^2 \sin^2 t \left[1 + \frac{1}{\mathcal{H}_1(\rho\cos t)}\right] g_0(t)\,dt = 1 - \exp\left(-\frac{\rho^2}{2}\right) + \frac{1}{2}\rho^2 \exp\left(-\frac{\rho^2}{2}\right).$$

The last term of the information loss follows from the fact that

$$\frac{\mathcal{H}_0(\rho\,\cos t)}{\mathcal{H}_1(\rho\,\cos t)} = \frac{\Phi(\rho\,\cos t)}{\varphi(\rho\,\cos t) + \rho\,\cos t\,\Phi(\rho\,\cos t)}.$$

Hence, we have the result of the theorem. □

This theorem is one of our main object that the exact information loss for one sample is shown, by which we obtain the asymptotic information loss of maximum likelihood estimator of θ on n i.i.d. samples in the next section.

5.4 Asymptotic information loss and statistical curvature

Let $\boldsymbol{X}_1 = (X_{11}, X_{21})', \ldots, \boldsymbol{X}_n = (X_{1n}, X_{2n})'$ be i.i.d. random vector samples from the Fisher's circle model. The joint density function is rewritten in the exponential type :

$$\prod_{i=1}^n f(\boldsymbol{x}_i : \boldsymbol{\alpha}(\theta)) = \exp\left\{n\,\rho\,\boldsymbol{e}(\theta)'\,\overline{\boldsymbol{x}}_n - n\,\frac{\rho^2}{2}\right\}\,\prod_{i=1}^n p_0(\boldsymbol{x}_i),$$

where

$$\overline{\boldsymbol{x}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i = \begin{pmatrix}\overline{x}_{1n}\\ \overline{x}_{2n}\end{pmatrix}.$$

Then, new random vector

$$\boldsymbol{Y}_n \equiv \sqrt{n}\;\overline{\boldsymbol{X}}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i$$

is normally distributed with mean vector

$$\boldsymbol{\alpha}_n(\theta) \equiv \rho_n \boldsymbol{e}(\theta), \qquad \rho_n \equiv \sqrt{n}\ \rho,$$

and covariance matrix $\mathbf{I}$. Let us transform $\boldsymbol{Y}_n$ to length and angle (R_n, T_n), $\quad R_n \geq 0, \quad T_n \in \theta$ in the polar coordinates :

$$\boldsymbol{Y}_n = \begin{pmatrix} Y_{1n} \\ Y_{2n} \end{pmatrix} = \begin{pmatrix} R_n \cos T_n \\ R_n \sin T_n \end{pmatrix} = R_n\ \boldsymbol{e}(T_n).$$

Then, we have the joint density function of (R_n, T_n), the marginal density of T_n, and the conditional density of R_n given T_n in the same way as in the previous sections by taking ρ_n in place of ρ :

$$f_{n\theta}(r,\, t) \;=\; \frac{r}{2\pi}\ \exp\left\{\rho_n\, r\ \cos(t-\theta) - \frac{r^2}{2} - \frac{\rho_n^2}{2}\right\}, \tag{5.4.1}$$

$$g_{n\theta}(t) \;=\; \frac{1}{2\pi}\ \exp\left\{-\frac{\rho_n^2}{2}\right\}\ \mathcal{H}_1(\rho_n\ \cos(t-\theta)), \tag{5.4.2}$$

$$h_{n\theta}(r|t) \;=\; \frac{\exp\{\rho_n r\ \cos(t-\theta)\}\, r\ \exp\{-\frac{r^2}{2}\}}{\mathcal{H}_1(\rho_n\ \cos(t-\theta))}. \tag{5.4.3}$$

It is easy to check that the likelihood equation is

$$\dot{\ell}_n(\theta) \;=\; \frac{\partial}{\partial\theta}\log f_{n\theta}(R_n,\, T_n) \;=\; \rho_n\, R_n\ \sin(T_n-\theta) \;=\; 0,$$

and thus T_n is the maximum likelihood estimator of θ.

Theorem 5.4.1 *Let $a_n = \rho_n\ \cos(T_n-\theta)$. Then the conditional variance of length R_n given angle T_n is*

$$V[R_n\,|\,T_n] = 1 + \frac{1}{\mathcal{H}_1(a_n)} - \left(\frac{\mathcal{H}_0(a_n)}{\mathcal{H}_1(a_n)}\right)^2 \to 1 \quad \textit{a.s.,} \quad \textit{as}\ \ n\to\infty. \tag{5.4.4}$$

Proof
Since

$$\frac{R_n}{\sqrt{n}}\,\boldsymbol{e}(T_n) = \frac{1}{\sqrt{n}}\boldsymbol{Y}_n = \overline{\boldsymbol{X}}_n \to \rho\boldsymbol{e}(\theta) \quad \text{a.s.,} \quad \text{as}\ \ n\to\infty,$$

we have

$$\frac{R_n}{\sqrt{n}} \to \rho, \quad T_n \to \theta, \qquad \text{a.s.,} \quad \text{as}\ \ n\to\infty.$$

and thus,

$$a_n \to \infty, \qquad \text{a.s.,} \quad \text{as}\ \ n\to\infty. \tag{5.4.5}$$

Lemma 5.3.1 and Theorem 5.3.1 follows that the conditional variance is rewritten as follows :

$$0 \leq 1 - V[R_n\,|\,T_n] \leq \left(\frac{\mathcal{H}_0(a_n)}{\mathcal{H}_1(a_n)}\right)^2 = \left(\frac{\Phi(a_n)}{\varphi(a_n) + a_n\,\Phi(a_n)}\right)^2,$$

and furthermore, that the last term is bounded by a_n^{-2} when $a_n > 0$:

$$\text{(5.4.6)} \qquad \left(\frac{\Phi(a_n)}{\varphi(a_n) + a_n\,\Phi(a_n)}\right)^2 \le a_n^{-2}, \qquad \text{for } a_n > 0,$$

where $a_n > 0$ a.s. (as $n \to \infty$) is guaranteed by (5.4.5). This proves the conclusion of the theorem. □

Theorem 5.4.2 *The asymptotic information loss is*

$$\text{(5.4.7)} \qquad I_{h_n} = E_{T_n}\{\,\rho_n^2\,\sin^2(T_n - \theta)\,V[R_n\,|\,T_n]\,\} \to 1, \qquad \textit{as} \quad n \to \infty.$$

Proof : By Lemma 5.3.2, we see that the main part of information loss converges to 1 as $n \to \infty$:

$$E\{\rho_n^2\,\sin^2(T_n - \theta)\} = 1 - \exp\left(-\frac{\rho_n^2}{2}\right) \quad \to \quad 1.$$

Let us show that the remain part

$$\int_0^{2\pi} \rho_n^2\,\sin^2(t-\theta)\,\{V[R_n\,|\,t] - 1\}\,g_{n\theta}(t)\,dt$$

converges to 0.

As $n \to \infty$, the marginal distribution of T_n converges to the distribution concentrated at θ. In fact, we see the marginal density of T_n :

$$\text{(5.4.8)} \quad \begin{aligned} g_{n\theta}(t) \quad &= \quad \varphi(\rho_n\,\sin(t-\theta))\,\{\varphi(\rho_n\,\cos(t-\theta)) + \rho_n\,\cos(t-\theta)\,\Phi(\rho_n\cos(t-\theta))\} \\ &\to \quad \begin{cases} 0, & \text{if} \quad t \ne \theta, \\ \infty, & \text{if} \quad t = \theta. \end{cases} \end{aligned}$$

At the same time, we see

$$\rho_n^2\,g_{n\theta}(t) \to 0,$$

outside the neighborhood of θ. Therefore, for any ε neighborhood $U = (-\varepsilon + \theta,\, \theta + \varepsilon)$, we have

$$\begin{aligned} 0 \quad &\le \quad \int_{U^c} \rho_n^2\,\sin^2(t-\theta)\,\{1 - V[R_n|t]\}\,g_{n\theta}(t)\,dt \\ &\le \quad \int_{U^c} \rho_n^2\,g_{n\theta}(t)\,dt \quad \to 0 \qquad \text{as} \quad n \to \infty, \end{aligned}$$

because of the finiteness of the integral range : $U^c \subset [0,\, 2\pi)$. On the neighborhood U, it holds that

$$a_n = \rho_n\,\cos(t-\theta) > 0,$$

and thus, from both the inequality (5.4.6) and the convergence (5.4.8) that

$$\begin{aligned} 0 \quad &\le \quad \int_U \rho_n^2\,\sin^2(t-\theta)\,\{1 - V[R_n\,|\,t]\}\,g_{n\theta}(t)\,dt \\ &\le \quad \int_U \tan^2(t-\theta)\,g_{n\theta}(t)\,dt \quad \to 0 \qquad \text{as} \quad n \to \infty, \end{aligned}$$

so that the proof is completed. □

It is well known that the maximum likelihood estimator T_n has the asymptotic consistency and normality, that is,

$$T_n \to \theta \quad \text{and} \quad \sqrt{n}\,(T_n - \theta) \to N(0,\, I(\theta)^{-1}) \quad \text{as } n \to \infty.$$

In fact, by using both the Fisher information

$$I(\theta) \;=\; \frac{1}{n}\, I_{f_n} \;=\; \rho^2$$

and Theorem 5.4.2, the information of T_n is equal to $I(\theta)$ asymptotically :

$$I_{T_n} = \frac{1}{n}\, I_{g_n}(\theta) = \frac{1}{n}\; (I_{f_n} - I_{h_n}) \quad \to \; \rho^2.$$

This is the first-order efficiency of maximum likelihood estimator.

5.5 Mathematical curvature and Statistical curvature

Let $\Gamma_M(\theta)$ be the mathematical curvature of curve $\boldsymbol{b}(\theta) = (b_1(\theta),\, b_2(\theta))'$, $\theta \in \Theta$:

$$\text{(5.5.1)} \qquad \Gamma_M(\theta) \equiv \frac{\dot{b}_1(\theta)\ddot{b}_2(\theta) - \dot{b}_2(\theta)\ddot{b}_1(\theta)}{\left(\sqrt{\dot{b}_1(\theta)^2 + \dot{b}_2(\theta)^2}\right)^3} = \frac{\det\left(\dot{\boldsymbol{b}}(\theta) \;:\; \ddot{\boldsymbol{b}}(\theta)\right)}{|\,\dot{\boldsymbol{b}}(\theta)\,|^3},$$

with

$$\left(\dot{\boldsymbol{b}}(\theta) \;:\; \ddot{\boldsymbol{b}}(\theta)\right) \equiv \begin{pmatrix} \dot{b}_1(\theta) & \ddot{b}_1(\theta) \\ \dot{b}_2(\theta) & \ddot{b}_2(\theta) \end{pmatrix}.$$

On the other hand, the statistical curvature in the curved exponential family

$$\text{(5.5.2)} \qquad f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta)) = \exp\left\{\boldsymbol{\alpha}(\theta)'\,\boldsymbol{x} - \psi(\boldsymbol{\alpha}(\theta))\right\}\; p_0(\boldsymbol{x})$$

is represented as follows :

$$\begin{aligned}\text{(5.5.3)} \qquad \Gamma_S(\theta) \;&\equiv\; \left(\frac{\dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta)\;\; \ddot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\ddot{\boldsymbol{\alpha}}(\theta) \;-\; (\dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\ddot{\boldsymbol{\alpha}}(\theta))^2}{|\,\dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta)\,|^3}\right)^{\frac{1}{2}}, \\ &=\; \frac{\left|\det\left(\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) \;:\; \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta)\right)\right|}{\left|\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\right|^3}.\end{aligned}$$

Efron showed that the information loss defined by Fisher is asymptotically and geometrically specified to be $I(\theta)\,\Gamma_S(\theta)^2$. In the Fisher circle model, the parametric $\boldsymbol{\alpha}(\theta)$ is equal to the mean vector $\boldsymbol{\mu}(\theta) = \rho\,\boldsymbol{e}(\theta)$ and the covariance $\boldsymbol{\Sigma}(\theta)$ is equal to $\mathbf{I}$, the unit matrix, by comparing two densities (5.2.1) and (5.5.2). The mathematical curvature (5.5.1) of mean curve $\boldsymbol{\mu}(\theta)$, letting $\boldsymbol{b}(\theta)$ be $\boldsymbol{\mu}(\theta)$ in (5.5.1), coincides with the statistical curvature (5.5.3) :

$$\Gamma_M(\theta) = \Gamma_S(\theta) = \frac{\det\left(\dot{\boldsymbol{\mu}}(\theta) \;:\; \ddot{\boldsymbol{\mu}}(\theta)\right)}{|\,\dot{\boldsymbol{\mu}}(\theta)\,|^3} = \frac{\det\left(\rho\,\dot{\boldsymbol{e}}(\theta) \;:\; \rho\,\ddot{\boldsymbol{e}}(\theta)\right)}{|\,\rho\,\dot{\boldsymbol{e}}(\theta)\,|^3} = \frac{1}{\rho}.$$

And the Fisher information loss is

$$I(\theta) = \dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta) = \dot{\boldsymbol{\mu}}(\theta)'\,\dot{\boldsymbol{\mu}}(\theta) = \rho^2.$$

Then, by using Efron's representation $I(\theta)\,\Gamma_S(\theta)^2$ and Theorem 5.4.2, the information loss is asymptotically represented as follows :

$$1 = I(\theta)\,\Gamma_S(\theta)^2 = I(\theta)\,\Gamma_M(\theta)^2 = \rho^2\left(\frac{1}{\rho}\right)^2. \tag{5.5.4}$$

This result is very particular, because the information loss is able to be asymptotically represented by either of curvatures. But, they are different in general. We shall demonstrate it by a simple example on Fisher circle model.

Let $\boldsymbol{X} = (X_1,\, X_2)'$ be distributed to the Fisher circle model with the covariance matrix $\sigma^2\mathbf{I}$, that is,

$$\boldsymbol{X} \;\sim\; N_2(\boldsymbol{\mu}(\theta),\, \sigma^2\mathbf{I}),$$

where $\boldsymbol{\mu}(\theta) = \rho\,\boldsymbol{e}(\theta)$ and σ is a positive constant. The density of $\boldsymbol{X}$ is rewritten in the exponential type :

$$f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta)) = \exp\left\{\left(\frac{1}{\sigma^2}\right)\boldsymbol{\mu}(\theta)'\boldsymbol{x} - \frac{1}{2\sigma^2}\,|\boldsymbol{\mu}(\theta)|^2\right\}\; p_0(\boldsymbol{x}). \tag{5.5.5}$$

The mathematical curvature of mean vector $\boldsymbol{\mu}(\theta)$ is

$$\Gamma_M(\theta) = \frac{\det\left(\dot{\boldsymbol{\mu}}(\theta)\ :\ \ddot{\boldsymbol{\mu}}(\theta)\right)}{|\,\dot{\boldsymbol{\mu}}(\theta)\,|^3} = \frac{1}{\rho}.$$

On the other hand, since the parametric in the density (5.5.5) is $\boldsymbol{\alpha}(\theta) = (1/\sigma^2)\,\boldsymbol{\mu}(\theta)$ and the covariance matrix is $\boldsymbol{\Sigma}(\theta) = \sigma^2\,\mathbf{I}$, the statistical curvature is

$$\Gamma_S(\theta) = \frac{\det\left(\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\,\dot{\boldsymbol{\alpha}}(\theta)\ :\ \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\,\ddot{\boldsymbol{\alpha}}(\theta)\right)}{|\,\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\,\dot{\boldsymbol{\alpha}}(\theta)\,|^3} = \frac{\det\left(\frac{1}{\sigma}\,\dot{\boldsymbol{\mu}}(\theta)\ :\ \frac{1}{\sigma}\,\ddot{\boldsymbol{\mu}}(\theta)\right)}{|\,\frac{1}{\sigma}\,\dot{\boldsymbol{\mu}}(\theta)\,|^3} = \frac{\sigma}{\rho}.$$

The Fisher information is

$$I(\theta) = \dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta) = \frac{1}{\sigma^2}\,\dot{\boldsymbol{\mu}}(\theta)'\,\dot{\boldsymbol{\mu}}(\theta) = \frac{\rho^2}{\sigma^2}.$$

Hence, we obtain the asymptotic representation of information loss as follows :

$$1 \;=\; I(\theta)\,\Gamma_S(\theta)^2 \;=\; \left(\frac{\rho^2}{\sigma^2}\right)\left(\frac{\sigma}{\rho}\right)^2.$$

Meanwhile

$$I(\theta)\,\Gamma_M(\theta)^2 \;=\; \left(\frac{\rho^2}{\sigma^2}\right)\left(\frac{1}{\rho}\right)^2 \;\;(\neq\; 1)$$

except for $\sigma\;=\;1$. That is, this example shows that the curvature used in the asymptotic representation of the information loss is the statistical curvature, but not the mathematical curvature.

Chapter 6

The Circular Mechanism

6.1 Introduction

Efron defined the statistical curvature $\Gamma_S(\theta)$ and showed that the asymptotic information loss of maximum likelihood estimator (MLE) $\hat{\theta}$ is represented by the product of the Fisher information and the statistical curvature square, that is,

$$\lim_{n\to\infty} E_{\hat{\theta}}\left[V\left[\dot{\ell}_n(\theta)\,|\,\hat{\theta} \right] \right] \;=\; I(\theta)\,\Gamma_S(\theta)^2,$$

where the expectation in the left-hand side means the expectation by the marginal probability (density) function of $\hat{\theta}$. Here we shall investigate the statistical curvature not in the information loss but itself in detail. In order to do it, we restrict the curved exponential family to one with the two dimensional.

In this chapter, we aim to grasp relationships between the former classical likelihood theories and the later recent information geometry by the "circular mechanism" in the two dimensional curved exponential family, where the circular mechanism is an algorithm to describe the osculating circle with the radius $|\,\Gamma_S(\theta)\,|^{-1}$ by using derivatives to second order of log-likelihood function.

We shall expose the mathematical curvature and the statistical curvature of the natural parameter vector indexed by the parameter θ in the curved exponential family. And we shall define the circular mechanism as the relationship between the frame of the traditional likelihood theory and the frame of the information geometry and prove some properties of the circular mechanism.

6.2 Mathematical and statistical curvatures in curved exponential family

For vectors $\boldsymbol{x}, \boldsymbol{\alpha}$ in the two dimensional Euclidean space $\mathcal{R}^2$:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix},$$

set the transpose and length of $\boldsymbol{x}$:

$$\boldsymbol{x}' = (x_1,\, x_2), \qquad\qquad |\,\boldsymbol{x}\,| = \sqrt{x_1^2 + x_2^2},$$

respectively, and the inner product of $\boldsymbol{x}, \boldsymbol{\alpha}$:

$$\langle \boldsymbol{\alpha},\, \boldsymbol{x} \rangle = \alpha_1 x_1 + \alpha_2 x_2.$$

Let two dimensional random vector $\boldsymbol{X} = (X_1, X_2)'$ be distributed to a curved exponential distribution with the density :

$$f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta)) \;=\; \exp\{\, \langle \boldsymbol{\alpha}(\theta), \boldsymbol{x} \rangle - \psi(\boldsymbol{\alpha}(\theta)) \,\}\; p_0(\boldsymbol{x}),$$

where $p_0(\boldsymbol{x})$ is the pivotal density function $\psi(\boldsymbol{\alpha}(\theta))$ is the cumulant generating function. Let us consider the natural parameter curve $\{\alpha(\theta) : \theta \in \Theta\}$ in the natural parameter space $\mathcal{A}$. Suppose the following conditions :

(C1) The parameter space Θ is a compact subspace of $\mathbf{R}^1$.

(C2) If $\theta_1 \neq \theta_2$ for $\theta_1, \theta_2 \in \Theta$, then $\boldsymbol{\alpha}(\theta_1) \neq \boldsymbol{\alpha}(\theta_2)$.

(C3) The curve $\boldsymbol{\alpha}(\theta)$ is twice continuous differentiable with respect to θ in the interior of Θ.

(C4) The second differentiation of $\boldsymbol{\alpha}(\theta)$ is not the zero vector $\mathbf{0}$ and not parallel to the first differentiation.

(C5) $\psi(\boldsymbol{\alpha})$ is strictly convex.

The following lemma is well-known and easy to see :

Lemma 6.2.1 *(1) The expectation of X is :*

$$\boldsymbol{\beta}(\theta) = \boldsymbol{\beta}(\boldsymbol{\alpha}(\theta)) = E[\boldsymbol{X}] = \nabla\psi(\boldsymbol{\alpha}(\theta)).$$

(2) The covariance matrix of X is positive definite :

$$\boldsymbol{\Sigma}(\theta) = \boldsymbol{\Sigma}(\boldsymbol{\alpha}(\theta)) = V[\boldsymbol{X}] = \nabla'\nabla\psi(\boldsymbol{\alpha}(\theta)).$$

(3) Let the dot notation " $\dot{}$ " mean the differentiation with respect to θ. Then,

$$\dot{\boldsymbol{\beta}}(\theta) = \boldsymbol{\Sigma}(\theta)\; \dot{\boldsymbol{\alpha}}(\theta). \tag{6.2.1}$$

Let the log-likelihood be denoted by :

$$\ell(\theta\,|\,\boldsymbol{x}) \;=\; \log f(\boldsymbol{x} : \boldsymbol{\alpha}(\theta)).$$

Then, the derivatives to the second order with respect to θ are

$$\dot{\ell}(\theta\,|\,\boldsymbol{x}) \;=\; <\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{x} - \boldsymbol{\beta}(\theta)>, \tag{6.2.2}$$

$$\ddot{\ell}(\theta\,|\,\boldsymbol{x}) \;=\; <\ddot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{x} - \boldsymbol{\beta}(\theta)> \;-\; <\dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\beta}}(\theta)>. \tag{6.2.3}$$

These differentiations imply the following relations;

Lemma 6.2.2 *The expectations and covariances of (6.2.2) and (6.2.3) are*

$$\begin{aligned}
E[\,\dot{\ell}(\theta\,|\,\boldsymbol{X})\,] &= \mathbf{0}, \\
E[\,\ddot{\ell}(\theta\,|\,\boldsymbol{X})\,] &= -\dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta), \\
V[\,\dot{\ell}(\theta\,|\,\boldsymbol{X})\,] &= \dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta) \;=\; E[\,-\ddot{\ell}(\theta\,|\,\boldsymbol{X})\,], \\
V[\,\ddot{\ell}(\theta\,|\,\boldsymbol{X})\,] &= \ddot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\ddot{\boldsymbol{\alpha}}(\theta), \\
Cov[\,\dot{\ell}(\theta\,|\,\boldsymbol{X}),\,\ddot{\ell}(\theta\,|\,\boldsymbol{X})\,] &= \dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\ddot{\boldsymbol{\alpha}}(\theta).
\end{aligned}$$

Thus the Fisher information is :

$$I(\theta) \;=\; \dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta),$$

and is positive definite and finite, because $\boldsymbol{\Sigma}(\theta)$ is positive definite.

Now, let us consider the curvature of the natural parameter curve $\{\boldsymbol{\alpha}(\theta) : \theta \in \Theta\}$ in $\mathcal{A}$. The mathematical curvature $\Gamma_M(\theta)$ (say) of the curve $\boldsymbol{\alpha}(\theta)$ is defined by :

$$\Gamma_M(\theta) \;=\; \frac{|\det(\dot{\boldsymbol{\alpha}}(\theta) \,:\, \ddot{\boldsymbol{\alpha}}(\theta))\,|}{|\dot{\boldsymbol{\alpha}}(\theta)|^3} \;=\; \frac{|\dot{\alpha}_1(\theta)\ddot{\alpha}_2(\theta) - \dot{\alpha}_2(\theta)\ddot{\alpha}_1(\theta)|}{|\dot{\alpha}_1(\theta)^2 + \dot{\alpha}_2(\theta)^2|^{\frac{3}{2}}}, \tag{6.2.4}$$

where the notation (:) means the matrix making from two vectors, that is, for any two vectors $\boldsymbol{a} = (a_1,\, a_2)',\ \boldsymbol{b} = (b_1,\, b_2)'$,

$$(\boldsymbol{a} \,:\, \boldsymbol{b}) = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}.$$

It is easy to see :

$$\Gamma_M(\theta)^2 \;=\; \frac{\dot{\boldsymbol{\alpha}}(\theta)'\,\dot{\boldsymbol{\alpha}}(\theta)\ \ddot{\boldsymbol{\alpha}}(\theta)'\,\ddot{\boldsymbol{\alpha}}(\theta) \;-\; \{\,\dot{\boldsymbol{\alpha}}(\theta)'\,\ddot{\boldsymbol{\alpha}}(\theta)\,\}^2}{\{\,\dot{\boldsymbol{\alpha}}(\theta)'\,\dot{\boldsymbol{\alpha}}(\theta)\,\}^3}.$$

This is also represented by the inner product as follows :

$$\Gamma_M(\theta)^2 \;=\; \frac{\langle\,\dot{\boldsymbol{\alpha}}(\theta),\,\dot{\boldsymbol{\alpha}}(\theta)\,\rangle\ \langle\,\ddot{\boldsymbol{\alpha}}(\theta),\,\ddot{\boldsymbol{\alpha}}(\theta)\,\rangle \;-\; \langle\,\dot{\boldsymbol{\alpha}}(\theta),\,\ddot{\boldsymbol{\alpha}}(\theta)\,\rangle^2}{\langle\,\dot{\boldsymbol{\alpha}}(\theta),\,\dot{\boldsymbol{\alpha}}(\theta)\,\rangle^3}. \tag{6.2.5}$$

On the other hand, Efron(1975) defined the statistical curvature $\Gamma_S(\theta)$ (say) of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$ in general, as follows :

$$\Gamma_S(\theta)^2 = \frac{V[\,\dot{\ell}(\theta\,|\,\boldsymbol{X})\,]\,V[\,\ddot{\ell}(\theta\,|\,\boldsymbol{X})\,] \;-\; \Big\{Cov[\,\dot{\ell}(\theta\,|\,\boldsymbol{X}),\,\ddot{\ell}(\theta\,|\,\boldsymbol{X})\,]\Big\}^2}{\Big\{V[\,\dot{\ell}(\theta\,|\,\boldsymbol{X})\,]\Big\}^3}.$$

By Lemma 6.2.2, it is also represented in the curved exponential family, as follows :

$$\Gamma_S(\theta)^2 \;=\; \frac{\dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta)\ \ddot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\ddot{\boldsymbol{\alpha}}(\theta) \;-\; \{\dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\ddot{\boldsymbol{\alpha}}(\theta)\}^2}{\{\dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta)\}^3}. \tag{6.2.6}$$

Let a new inner product $\langle\langle\,,\,\rangle\rangle$ be

$$\langle\langle\,\boldsymbol{\alpha}(\theta),\,\boldsymbol{\alpha}(\theta)\,\rangle\rangle \;=\; \langle\,\boldsymbol{\alpha}(\theta),\,\boldsymbol{\alpha}(\theta)\,\rangle_{\boldsymbol{\Sigma}(\theta)} \;=\; \boldsymbol{\alpha}(\theta)'\,\boldsymbol{\Sigma}(\theta)\,\boldsymbol{\alpha}(\theta).$$

Then the statistical curvature (6.2.6) is also represented by the new inner product as follows :

$$\text{(6.2.7)} \qquad \Gamma_S(\theta)^2 \;=\; \frac{\langle\langle\, \dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\alpha}}(\theta)\,\rangle\rangle \;\langle\langle\, \ddot{\boldsymbol{\alpha}}(\theta),\, \ddot{\boldsymbol{\alpha}}(\theta)\,\rangle\rangle \;-\; \langle\langle\, \dot{\boldsymbol{\alpha}}(\theta),\, \ddot{\boldsymbol{\alpha}}(\theta)\,\rangle\rangle^2}{\langle\langle\, \dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\alpha}}(\theta)\,\rangle\rangle^3}.$$

By the comparison of (6.2.5) and (6.2.7), we can understand that the statistical curvature of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$ has formally a similar representation with the mathematical curvature of the curve $\boldsymbol{\alpha}(\theta)$. But there exists essentially a structural gap between the statistical curvature $\Gamma_S(\theta)$ and the mathematical curvature $\Gamma_M(\theta)$. We shall describe it in detail as the following section.

6.3 Circular mechanism

In the two dimensional curved exponential family, the log-likelihood function has the circular mechanism that derives the statistical curvature and the center of osculating circle. We shall describe the circular mechanism as follows :

Circular Mechanism *Let us consider the following two equations of $\boldsymbol{x} = (x_1, x_2)'$ under a fixed parameter θ :*

$$\begin{aligned} \dot{\ell}(\theta\,|\,\boldsymbol{x}) &= 0, \\ \ddot{\ell}(\theta\,|\,\boldsymbol{x}) &= 0, \end{aligned}$$

although the first one is, usually, well known as the likelihood equation of parameter θ under a given observation $\boldsymbol{x}$. The solution $\boldsymbol{c}(\theta)$ (say) implies the center of osculating circle and the length between the center and the expectation parameter leads to the statistical curvature $\Gamma_S(\theta)$.

We shall investigate relationships as the structure between the mathematical curvature and the statistical curvature in the circular mechanism by the following various cases. These case-studies shall elucidate the connection of two curvatures step by step.

(Case 1) $V[\boldsymbol{X}] = \mathbf{I}_2$

Theorem 6.3.1 *If the variance matrix is $V[\boldsymbol{X}] = \mathbf{I}_2$, then the two equations in the circular mechanism are, for a fixed parameter θ,*

$$\begin{aligned} \dot{\ell}(\theta\,|\,\boldsymbol{x}) &= \langle \dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{x} - \boldsymbol{b} - \boldsymbol{\alpha}(\theta)\rangle \;=\; 0, \\ \ddot{\ell}(\theta\,|\,\boldsymbol{x}) &= \langle \ddot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{x} - \boldsymbol{b} - \boldsymbol{\alpha}(\theta)\rangle \;-\; \langle \dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\alpha}}(\theta)\rangle \;=\; 0, \end{aligned}$$

where $\boldsymbol{b}$ is a constant vector. Thus, for the solution $\boldsymbol{c}(\theta)$ in the circular mechanism, the point $\boldsymbol{c}(\theta)$ is the center of osculating circle at $\boldsymbol{\alpha}(\theta)$ and the length between the center and the expectation parameter leads the statistical curvature $\Gamma_S(\theta)$ of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$.

Proof : This case is the simplest case. By Lemma 6.2.1, it holds that there exists a vector $\boldsymbol{b}$ such that

$$\psi(\boldsymbol{\alpha}(\theta)) \;=\; \frac{|\boldsymbol{\alpha}(\theta)|^2}{2} \;+\; \langle \boldsymbol{\alpha}(\theta),\, \boldsymbol{b}\rangle,$$

so that the expectation parameter is $\boldsymbol{\beta}(\theta) = \boldsymbol{\alpha}(\theta) + \boldsymbol{b}$. We shall consider the solution $\boldsymbol{c}(\theta)$ in the circular mechanism. Since the first equation is represented by

$$\langle \dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{x} - \boldsymbol{b} - \boldsymbol{\alpha}(\theta) \rangle \;=\; 0,$$

there exists r such that

$$\boldsymbol{x} - \boldsymbol{b} - \boldsymbol{\alpha}(\theta) \;=\; r\,\mathbf{S}(\frac{\pi}{2})\dot{\boldsymbol{\alpha}}(\theta),$$

where $\mathbf{S}(\cdot)$ is the rotation matrix in the two dimensional space, that is,

$$\mathbf{S}(\eta) = \begin{pmatrix} \cos\eta & -\sin\eta \\ \sin\eta & \cos\eta \end{pmatrix}, \qquad \forall\eta \in [0, 2\pi).$$

By substituting this into the second equation,

$$r\,\langle \ddot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\dot{\boldsymbol{\alpha}}(\theta) \rangle \;-\; \langle \dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\alpha}}(\theta) \rangle \;=\; 0,$$

so that we obtain the solution in the circular mechanism :

$$\begin{aligned} \boldsymbol{c}(\theta) \;&=\; \boldsymbol{b} + \boldsymbol{\alpha}(\theta) + \frac{\langle \dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\alpha}}(\theta) \rangle}{\langle \ddot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\dot{\boldsymbol{\alpha}}(\theta) \rangle}\;\mathbf{S}(\frac{\pi}{2})\dot{\boldsymbol{\alpha}}(\theta) \\ &=\; \boldsymbol{\beta}(\theta) + \frac{\langle \dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\alpha}}(\theta) \rangle}{\det\,(\dot{\boldsymbol{\alpha}}(\theta) : \ddot{\boldsymbol{\alpha}}(\theta))}\;\mathbf{S}(\frac{\pi}{2})\dot{\boldsymbol{\alpha}}(\theta). \end{aligned}$$

Thus the length between $\boldsymbol{c}(\theta)$ and $\boldsymbol{\beta}(\theta)$ leads the statistical curvature, that is,

$$(6.3.1) \qquad |\,\boldsymbol{c}(\theta) - \boldsymbol{\beta}(\theta)\,|^2 \;=\; \frac{\langle \dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\alpha}}(\theta) \rangle^2}{\{\det\,(\dot{\boldsymbol{\alpha}}(\theta) : \ddot{\boldsymbol{\alpha}}(\theta))\}^2}\,\langle \dot{\boldsymbol{\alpha}}(\theta),\, \dot{\boldsymbol{\alpha}}(\theta) \rangle \;=\; \frac{1}{\Gamma_S(\theta)^2}$$

by the definition (6.2.6). Since $\boldsymbol{\beta}(\theta) = \boldsymbol{\alpha}(\theta) + \boldsymbol{b}$, the mathematical curvature $\Gamma_M(\theta)$ of the curve $\boldsymbol{\alpha}(\theta)$ is equal to one of the curve $\boldsymbol{\beta}(\theta)$, that is,

$$\Gamma_M(\theta)^2 \;=\; \frac{\dot{\boldsymbol{\alpha}}(\theta)'\,\dot{\boldsymbol{\alpha}}(\theta)\;\ddot{\boldsymbol{\alpha}}(\theta)'\,\ddot{\boldsymbol{\alpha}}(\theta) \;-\; \{\,\dot{\boldsymbol{\alpha}}(\theta)'\,\ddot{\boldsymbol{\alpha}}(\theta)\,\}^2}{\{\,\dot{\boldsymbol{\alpha}}(\theta)'\,\dot{\boldsymbol{\alpha}}(\theta)\,\}^3},$$

so that the statistical curvature $\Gamma_S(\theta)$ of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$ is equivalent to the mathematical curvature $\Gamma_M(\theta)$ of the curve $\boldsymbol{\alpha}(\theta)$. Thus the point $\boldsymbol{c}(\theta)$ becomes the center of osculating circle at $\boldsymbol{\alpha}(\theta)$. □

(Case 2) $V[\boldsymbol{X}] = \sigma^2\mathbf{I}_2$

Theorem 6.3.2 *If the variance matrix is $V[\boldsymbol{X}] = \sigma^2\mathbf{I}_2$, then the two equations in the circular mechanism are, for a fixed parameter θ,*

$$\begin{aligned} \dot{\ell}(\theta\,|\,\boldsymbol{x}) \;&=\; \langle \dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{x} - \boldsymbol{b} - \sigma^2\,\boldsymbol{\alpha}(\theta) \rangle \;=\; 0, \\ \ddot{\ell}(\theta\,|\,\boldsymbol{x}) \;&=\; \langle \ddot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{x} - \boldsymbol{b} - \sigma^2\,\boldsymbol{\alpha}(\theta) \rangle \;-\; \langle \dot{\boldsymbol{\alpha}}(\theta),\, \sigma^2\,\dot{\boldsymbol{\alpha}}(\theta) \rangle \;=\; 0. \end{aligned}$$

Thus, for the solution $\boldsymbol{c}(\theta)$ in the circular mechanism, the point $\boldsymbol{c}(\theta)/\,\sigma$ is the center of osculating circle at $\sigma\,\boldsymbol{\alpha}(\theta)$ and the length between $\boldsymbol{c}(\theta)$ and $\boldsymbol{\beta}(\theta)$ leads the statistical curvature $\Gamma_S(\theta)$ of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$.

Proof : If $\sigma = 1$, then this is the same with Theorem 6.3.1, so we may assume that $\sigma \neq 1$. Since $V[\boldsymbol{X}] = \sigma^2 \mathbf{I}_2$,

$$\psi(\boldsymbol{\alpha}(\theta)) \;=\; \frac{\sigma^2\,|\boldsymbol{\alpha}(\theta)|^2}{2} \;+\; \langle \boldsymbol{\alpha}(\theta),\, \boldsymbol{b} \rangle,$$

so that the expectation parameter is $\boldsymbol{\beta}(\theta) = \sigma^2\,\boldsymbol{\alpha}(\theta) + \boldsymbol{b}$. We shall consider the solution $\boldsymbol{c}(\theta)$ in the circular mechanism. Here we shall convert the formulations of the above two equations as follows :

$$\begin{aligned}
\dot{\ell}(\theta\,|\,\boldsymbol{x}) \;&=\; \langle \sigma\,\dot{\boldsymbol{\alpha}}(\theta),\, \frac{1}{\sigma}\,(\boldsymbol{x} - \boldsymbol{b}) - \sigma\,\boldsymbol{\alpha}(\theta) \rangle \;=\; 0,\\
\ddot{\ell}(\theta\,|\,\boldsymbol{x}) \;&=\; \langle \sigma\,\ddot{\boldsymbol{\alpha}}(\theta),\, \frac{1}{\sigma}\,(\boldsymbol{x} - \boldsymbol{b}) - \sigma\,\boldsymbol{\alpha}(\theta) \rangle \;-\; \langle \sigma\,\dot{\boldsymbol{\alpha}}(\theta),\, \sigma\,\dot{\boldsymbol{\alpha}}(\theta) \rangle \;=\; 0.
\end{aligned}$$

The first equation implies that there exists r such that

$$\frac{1}{\sigma}\,(\boldsymbol{x} - \boldsymbol{b}) - \sigma\,\boldsymbol{\alpha}(\theta) \;=\; r\,\mathbf{S}(\frac{\pi}{2})\,\sigma\,\dot{\boldsymbol{\alpha}}(\theta).$$

By substituting this into the second equation,

$$r\,\langle \sigma\,\ddot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\,\sigma\,\dot{\boldsymbol{\alpha}}(\theta) \rangle \;-\; \langle \sigma\,\dot{\boldsymbol{\alpha}}(\theta),\, \sigma\,\dot{\boldsymbol{\alpha}}(\theta) \rangle \;=\; 0,$$

so that we obtain the solution in the circular mechanism :

$$\begin{aligned}
\boldsymbol{c}(\theta) \;&=\; \boldsymbol{b} + \sigma^2\,\boldsymbol{\alpha}(\theta) + \sigma\,\frac{\langle \sigma\,\dot{\boldsymbol{\alpha}}(\theta),\, \sigma\,\dot{\boldsymbol{\alpha}}(\theta) \rangle}{\langle \sigma\,\ddot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\sigma\,\dot{\boldsymbol{\alpha}}(\theta) \rangle}\,\mathbf{S}(\frac{\pi}{2})\sigma\,\dot{\boldsymbol{\alpha}}(\theta)\\
&=\; \boldsymbol{\beta}(\theta) + \sigma\,\frac{\langle \sigma\,\dot{\boldsymbol{\alpha}}(\theta),\, \sigma\,\dot{\boldsymbol{\alpha}}(\theta) \rangle}{\det\,(\sigma\,\dot{\boldsymbol{\alpha}}(\theta) : \sigma\,\ddot{\boldsymbol{\alpha}}(\theta))}\,\mathbf{S}(\frac{\pi}{2})\sigma\,\dot{\boldsymbol{\alpha}}(\theta).
\end{aligned}$$

Thus the length between $\boldsymbol{c}(\theta)$ and $\boldsymbol{\beta}(\theta)$ leads the statistical curvature, that is,

$$|\,\boldsymbol{c}(\theta) - \boldsymbol{\beta}(\theta)\,|^2 \;=\; \sigma^2\,\frac{\langle \sigma\,\dot{\boldsymbol{\alpha}}(\theta),\, \sigma\,\dot{\boldsymbol{\alpha}}(\theta) \rangle^2}{\{\det\,(\sigma\,\dot{\boldsymbol{\alpha}}(\theta) : \sigma\,\ddot{\boldsymbol{\alpha}}(\theta))\}^2}\,\langle \sigma\,\dot{\boldsymbol{\alpha}}(\theta),\, \sigma\,\dot{\boldsymbol{\alpha}}(\theta) \rangle \;=\; \frac{\sigma^2}{\Gamma_S(\theta)^2} \tag{6.3.2}$$

by the definition (6.2.6). (Compare this with the length (6.3.1) of Case 1.) In other words, the length square between $\boldsymbol{c}(\theta)/\,\sigma$ and $\boldsymbol{\beta}(\theta)/\,\sigma$ is exactly the inverse of $\Gamma_S(\theta)^2$. Since $\boldsymbol{\beta}(\theta) = \sigma^2\,\boldsymbol{\alpha}(\theta) + \boldsymbol{b}$, the mathematical curvature of the curve $\boldsymbol{\alpha}(\theta)$ is not equal to one of the curve $\boldsymbol{\beta}(\theta)$. But the mathematical curvature of the curve $\sigma\,\boldsymbol{\alpha}(\theta)$ is equal to one of the curve $\boldsymbol{\beta}(\theta)/\,\sigma$, that is,

$$\Gamma_M(\theta)^2 \;=\; \frac{1}{\sigma^2}\,\frac{\dot{\boldsymbol{\alpha}}(\theta)'\,\dot{\boldsymbol{\alpha}}(\theta)\;\ddot{\boldsymbol{\alpha}}(\theta)'\,\ddot{\boldsymbol{\alpha}}(\theta) \;-\; \{\,\dot{\boldsymbol{\alpha}}(\theta)'\,\ddot{\boldsymbol{\alpha}}(\theta)\,\}^2}{\{\,\dot{\boldsymbol{\alpha}}(\theta)'\,\dot{\boldsymbol{\alpha}}(\theta)\,\}^3},$$

so that the statistical curvature $\Gamma_S(\theta)$ of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$ is equivalent to the mathematical curvature $\Gamma_M(\theta)$ of the curve $\sigma\,\boldsymbol{\alpha}(\theta)$. Thus the point $\boldsymbol{c}(\theta)/\,\sigma$ becomes the center of osculating circle at $\sigma\,\boldsymbol{\alpha}(\theta)$. □

(Case 3) $V[\boldsymbol{X}] = \boldsymbol{\Sigma}$

Theorem 6.3.3 *If the variance matrix is $V[\boldsymbol{X}] = \boldsymbol{\Sigma}$, then the two equations in the circular mechanism are, for a fixed parameter θ,*

$$\begin{aligned} \dot\ell(\theta \,|\, \boldsymbol{x}) &= \langle \dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{x} - \boldsymbol{b} - \boldsymbol{\Sigma}\,\boldsymbol{\alpha}(\theta)\rangle \;=\; 0, \\ \ddot\ell(\theta \,|\, \boldsymbol{x}) &= \langle \ddot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{x} - \boldsymbol{b} - \boldsymbol{\Sigma}\,\boldsymbol{\alpha}(\theta)\rangle \;-\; \langle \dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}\,\dot{\boldsymbol{\alpha}}(\theta)\rangle \;=\; 0. \end{aligned}$$

Thus, for the solution $\boldsymbol{c}(\theta)$ in the circular mechanism, the point $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{c}(\theta)$ is the center of osculating circle at $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)$ and the length between $\boldsymbol{c}(\theta)$ and $\boldsymbol{\beta}(\theta)$ leads the statistical curvature $\Gamma_S(\theta)$ of the log-likelihood function $\ell(\theta \,|\, \boldsymbol{x})$.

Proof : If $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_2$, then this is the same with Theorem 6.3.2, so that we may assume that $\boldsymbol{\Sigma} \neq \sigma^2\mathbf{I}_2$. Since $V[\boldsymbol{X}] = \boldsymbol{\Sigma}$,

$$\psi(\boldsymbol{\alpha}(\theta)) \;=\; \frac{\boldsymbol{\alpha}(\theta)'\,\boldsymbol{\Sigma}\,\boldsymbol{\alpha}(\theta)}{2} \;+\; \langle \boldsymbol{\alpha}(\theta),\, \boldsymbol{b}\rangle,$$

so that the expectation parameter is $\boldsymbol{\beta}(\theta) = \boldsymbol{\Sigma}\,\boldsymbol{\alpha}(\theta) + \boldsymbol{b}$. We shall consider the solution $\boldsymbol{c}(\theta)$ in the circular mechanism. Here we shall convert the formulations of the above two equations as follows :

$$\begin{aligned} \dot\ell(\theta \,|\, \boldsymbol{x}) &= \langle \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{b}) - \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)\rangle \;=\; 0, \\ \ddot\ell(\theta \,|\, \boldsymbol{x}) &= \langle \boldsymbol{\Sigma}^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{b}) - \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)\rangle \;-\; \langle \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle \;=\; 0. \end{aligned}$$

The first equation implies that there exists r such that

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{b}) - \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\alpha}(\theta) \;=\; r\,\mathbf{S}(\frac{\pi}{2})\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta).$$

By substituting this into the second equation,

$$r\,\langle \boldsymbol{\Sigma}^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle \;-\; \langle \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle \;=\; 0,$$

so that we obtain the solution in the circular mechanism :

$$\begin{aligned} \boldsymbol{c}(\theta) &= \boldsymbol{b} + \boldsymbol{\Sigma}\,\boldsymbol{\alpha}(\theta) + \boldsymbol{\Sigma}^{\frac{1}{2}}\,\frac{\langle \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle}{\langle \boldsymbol{\Sigma}^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle}\mathbf{S}(\frac{\pi}{2})\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) \\ &= \boldsymbol{\beta}(\theta) + \boldsymbol{\Sigma}^{\frac{1}{2}}\,\frac{\langle \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle}{\det\Big(\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) : \boldsymbol{\Sigma}^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta)\Big)}\mathbf{S}(\frac{\pi}{2})\,\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta). \end{aligned}$$

Thus the length between $\boldsymbol{c}(\theta)$ and $\boldsymbol{\beta}(\theta)$ leads the statistical curvature, that is,

$$\begin{aligned} |\,\boldsymbol{c}(\theta) - \boldsymbol{\beta}(\theta)\,|^2 &= \frac{\langle \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle^2}{\Big\{\det\Big(\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) : \boldsymbol{\Sigma}^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta)\Big)\Big\}^2}\;\langle\langle \mathbf{S}(\frac{\pi}{2})\,\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\,\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle\rangle \\ &= \frac{1}{\Gamma_S(\theta)^2}\,\frac{\langle\langle \mathbf{S}(\frac{\pi}{2})\,\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\,\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle\rangle}{\langle \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle} \qquad (6.3.3) \end{aligned}$$

by the definition (6.2.6). (Compare this with the length (6.3.2) of Case 2.) In other words, the length square between $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{c}(\theta)$ and $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\beta}(\theta)$ is exactly the inverse of $\Gamma_S(\theta)^2$. Since $\boldsymbol{\beta}(\theta) =$

$\boldsymbol{\Sigma}\,\boldsymbol{\alpha}(\theta)+\boldsymbol{b}$, the mathematical curvature of the curve $\boldsymbol{\alpha}(\theta)$ is not equal to one of the curve $\boldsymbol{\beta}(\theta)$. But the mathematical curvature of the curve $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)$ is equal to one of the curve $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\beta}(\theta)$, that is,

$$\Gamma_M(\theta)^2 \;=\; \frac{\dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}\,\dot{\boldsymbol{\alpha}}(\theta)\;\ddot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}\,\ddot{\boldsymbol{\alpha}}(\theta)\;-\;\{\dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}\,\ddot{\boldsymbol{\alpha}}(\theta)\}^2}{\{\dot{\boldsymbol{\alpha}}(\theta)'\,\boldsymbol{\Sigma}\,\dot{\boldsymbol{\alpha}}(\theta)\}^3},$$

so that the statistical curvature $\Gamma_S(\theta)$ of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$ is equivalent to the mathematical curvature $\Gamma_M(\theta)$ of the curve $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)$. Thus the point $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{c}(\theta)$ becomes the center of osculating circle at $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)$. □

From the above three cases we shall obtain a relation between the statistical curvature and the mathematical curvature as follows :

Theorem 6.3.4 *If the variance matrix $V[\boldsymbol{X}]$ does not depend on the parameter θ, that is, $V[\boldsymbol{X}]=\boldsymbol{\Sigma}$, then the statistical curvature $\Gamma_S(\theta)$ of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$ is equal to the mathematical curvature $\Gamma_M(\theta)$ of the curve $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)$.* □

(Case 4) $V[\boldsymbol{X}]=\boldsymbol{\Sigma}(\theta)$

Theorem 6.3.5 *If the variance matrix is $V[\boldsymbol{X}]=\boldsymbol{\Sigma}(\theta)$, then the two equations in the circular mechanism are, for a fixed parameter θ,*

$$\begin{aligned}\dot{\ell}(\theta\,|\,\boldsymbol{x}) &= \langle\dot{\boldsymbol{\alpha}}(\theta),\,\boldsymbol{x}-\boldsymbol{\beta}(\theta)\rangle \;=\; 0,\\ \ddot{\ell}(\theta\,|\,\boldsymbol{x}) &= \langle\ddot{\boldsymbol{\alpha}}(\theta),\,\boldsymbol{x}-\boldsymbol{\beta}(\theta)\rangle \;-\; \langle\dot{\boldsymbol{\alpha}}(\theta),\,\boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta)\rangle \;=\; 0.\end{aligned}$$

Thus, for the solution $\boldsymbol{c}(\theta)$ in the circular mechanism, the length between $\boldsymbol{c}(\theta)$ and $\boldsymbol{\beta}(\theta)$ leads the statistical curvature $\Gamma_S(\theta)$ of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$.

Proof : If the variance $\boldsymbol{\Sigma}(\theta)$ does not depend on the parameter θ, this is the same with Theorem 6.3.3, so we may assume that the variance depends on the parameter θ. Then the cumulant generating function $\psi(\boldsymbol{\alpha}(\theta))$ may be not expressed by $\boldsymbol{\alpha}(\theta)$ explicitly, but we can use the differential relation (6.2.1) :

$$\dot{\boldsymbol{\beta}}(\theta) \;=\; \boldsymbol{\Sigma}(\theta)\,\dot{\boldsymbol{\alpha}}(\theta).$$

We shall consider the solution $\boldsymbol{c}(\theta)$ in the circular mechanism. Here we shall convert the formulations of the above two equations as follows :

$$\begin{aligned}\dot{\ell}(\theta\,|\,\boldsymbol{x}) &= \langle\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\,\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\beta}(\theta))\rangle \;=\; 0,\\ \ddot{\ell}(\theta\,|\,\boldsymbol{x}) &= \langle\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta),\,\boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\beta}(\theta))\rangle \;-\; \langle\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\,\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle \;=\; 0.\end{aligned}$$

The first equation implies that there exists r such that

$$\boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\beta}(\theta)) \;=\; r\,\mathbf{S}(\frac{\pi}{2})\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta).$$

By substituting this into the second equation,

$$r\,\langle\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta),\,\mathbf{S}(\frac{\pi}{2})\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle \;-\; \langle\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\,\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle \;=\; 0,$$

so that we obtain the solution in the circular mechanism :

$$\begin{aligned}\boldsymbol{c}(\theta) &= \boldsymbol{\beta}(\theta) + \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\, \frac{\langle \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle}{\langle \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle} \mathbf{S}(\frac{\pi}{2})\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) \\ &= \boldsymbol{\beta}(\theta) + \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\, \frac{\langle \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle}{\det\left(\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) : \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta)\right)} \mathbf{S}(\frac{\pi}{2})\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta).\end{aligned}$$

Thus the length between $\boldsymbol{c}(\theta)$ and $\boldsymbol{\beta}(\theta)$ leads the statistical curvature, that is,

$$\begin{aligned}|\,\boldsymbol{c}(\theta) - \boldsymbol{\beta}(\theta)\,|^2 &= \frac{\langle \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle^2}{\left\{\det\left(\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) : \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta)\right)\right\}^2} \langle\langle \mathbf{S}(\frac{\pi}{2})\,\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\,\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle\rangle \\ \text{(6.3.4)} \qquad &= \frac{1}{\Gamma_S(\theta)^2}\, \frac{\langle\langle \mathbf{S}(\frac{\pi}{2})\,\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \mathbf{S}(\frac{\pi}{2})\,\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle\rangle}{\langle \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),\, \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\rangle}\end{aligned}$$

by the definition (6.2.6). (Compare this with the length (6.3.3) of Case 3.) In other words, the length square between $\boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}}\boldsymbol{c}(\theta)$ and $\boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}}\boldsymbol{\beta}(\theta)$ is exactly the inverse of $\Gamma_S(\theta)^2$. □

In Theorem 6.3.5, we do not insist that the solution $\boldsymbol{c}(\theta)$ implies the center of osculating circle, because there exists the fatal difference as the structure between the statistical curvature of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$ and the mathematical curvature.

Theorem 6.3.6 *If the variance $V[\boldsymbol{X}]$ depends on the parameter θ, the statistical curvature of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$ is different from the mathematical curvatures of the curves $\boldsymbol{\alpha}(\theta)$, $\boldsymbol{\beta}(\theta)$, and $\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)$.*

Proof : Under this condition, the statistical curvature of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$ is also represented by

$$\Gamma_S(\theta) = \frac{\det\left(\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) : \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta)\right)}{\left|\,\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\,\right|^3}$$

by easy calculations. Thus it is clear that this curvature is different from the mathematical curvature of the curves $\boldsymbol{\alpha}(\theta)$ and $\boldsymbol{\beta}(\theta)$. Then what we must consider is the case of curve $\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)$. Since the variance depends on the parameter θ, the derivatives of the curve are

$$\text{(6.3.5)} \qquad \frac{\partial}{\partial\theta}\left\{\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)\right\} = \dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta) + \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta),$$

$$\text{(6.3.6)} \qquad \frac{\partial^2}{\partial\theta^2}\left\{\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)\right\} = \ddot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta) + 2\,\dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) + \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta).$$

We shall investigate whether the above first equation (6.3.5) is equal to $\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)$. If the first equation (6.3.5) is not equal to $\boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)$, then it is obvious that $\Gamma_M(\theta) \neq \Gamma_S(\theta)$.

Assume that $\partial/\,\partial\theta\ \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta) = \boldsymbol{\Sigma}^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)$, that is,

$$\dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta) = \mathbf{0}.$$

In this case the determinant of $\dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}$ must be zero. Since

$$\frac{\partial}{\partial\theta}\left\{\dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)\right\} \;=\; \ddot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)+\dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) \;=\; \mathbf{0},$$

the second equation (6.3.6) is

$$\frac{\partial^2}{\partial\theta^2}\left\{\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)\right\} \;=\; \dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) \;+\; \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta).$$

Here suppose that $\dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)=\mathbf{0}$. Then it holds that

$$\dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\;\left(\boldsymbol{\alpha}(\theta):\dot{\boldsymbol{\alpha}}(\theta)\right) \;=\; \mathbf{0} \quad\text{that is,}\quad \det\left(\boldsymbol{\alpha}(\theta):\dot{\boldsymbol{\alpha}}(\theta)\right) \;=\; 0,$$

so that $\boldsymbol{\alpha}(\theta)$ is parallel to $\dot{\boldsymbol{\alpha}}(\theta)$. Thus there exists k such that $\dot{\boldsymbol{\alpha}}(\theta)=k\,\boldsymbol{\alpha}(\theta)$, so that

$$\boldsymbol{\alpha}(\theta) \;=\; \begin{pmatrix}\alpha_1(\theta)\\ \alpha_2(\theta)\end{pmatrix} \;=\; \begin{pmatrix}\exp\{k\,\theta+a_1\}\\ \exp\{k\,\theta+a_2\}\end{pmatrix},$$

where a_1, a_2 are constants. This implies that the curve $\boldsymbol{\alpha}(\theta)$ is a straight line, so that the condition $\dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)=\mathbf{0}$ contradicts the assumption (C4) of curve $\boldsymbol{\alpha}(\theta)$. Thus

$$\dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta)\neq\mathbf{0},$$

so that the second equation (6.3.6) is

$$\frac{\partial^2}{\partial\theta^2}\left\{\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)\right\} \;=\; \dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\dot{\boldsymbol{\alpha}}(\theta) \;+\; \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta) \;\neq\; \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\ddot{\boldsymbol{\alpha}}(\theta).$$

Therefore the mathematical curvature $\Gamma_M(\theta)$ of the curve $\boldsymbol{\Sigma}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)$ is not equal to the statistical curvature $\Gamma_S(\theta)$ of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$ even if $\dot{\boldsymbol{\Sigma}}(\theta)^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)=\mathbf{0}$. The proof is completed. □

A key in Theorem 6.3.6 is that the variance structure changes locally by the point of parametric $\boldsymbol{\alpha}(\theta)$. As we see in Theorem 6.3.4, if the variance does not depend on the parameter θ, then the variance structure does not change by the point of parametric $\boldsymbol{\alpha}(\theta)$, that is, the variance structure is constant globally.

In the frame of the traditional likelihood estimation theories, since the variance changes by the point of $\boldsymbol{\alpha}(\theta)$, the *curve* $\boldsymbol{\alpha}(\theta)$ on the likelihood can not be regarded continuous ordinarily against the curve on the mathematical curvature. This is the reason why we purposely describe the statistical curvature as the statistical curvature of the log-likelihood function $\ell(\theta\,|\,\boldsymbol{x})$. This localization implies the information geometry like Amari's frame(1985) based on the Fisher information $I(\theta)$ as the local metric in the differential geometry. Thus we obtain that the circular mechanism is also a connection between the frame of the traditional likelihood estimation theories and one of the information geometry.

6.4 Some properties of circular mechanism

We shall show some properties with respect to the circular mechanism. Let

$$L(\theta) \ = \ \{\, \boldsymbol{x} : \dot{\ell}(\theta \,|\, \boldsymbol{x}) = 0 \,\}$$

be a subset for any fixed θ. As a geometrical interpretation of the solution $\boldsymbol{c}(\theta)$ in the circular mechanism, we have the following lemma :

Lemma 6.4.1 *If the point $\boldsymbol{x}$ satisfies that*

$$\dot{\ell}(\theta \,|\, \boldsymbol{x}) = 0 \quad \textit{and} \quad \ddot{\ell}(\theta \,|\, \boldsymbol{x}) = 0,$$

then it approximately holds that

(6.4.1)
$$\boldsymbol{x} \in L(\theta) \cap L(\theta + \delta)$$

where $\theta + \delta$ belongs to a neighborhood of θ.

Proof : From the first equation $\dot{\ell}(\theta \,|\, \boldsymbol{x}) = 0$ and (6.2.2), it holds that $\boldsymbol{x} \in L(\theta)$, that is,

$$\boldsymbol{x} - \boldsymbol{\beta}(\theta) \quad \perp \quad \dot{\boldsymbol{\alpha}}(\theta).$$

For any $\theta + \delta$ in a neighborhood of θ, we consider $\dot{\ell}(\theta \,|\, \boldsymbol{x})$. Under the second equation $\ddot{\ell}(\theta \,|\, \boldsymbol{x}) = 0$, we have

$$\dot{\ell}(\theta + \delta \,|\, \boldsymbol{x}) \ = \ \dot{\ell}(\theta \,|\, \boldsymbol{x}) + \delta\, \ddot{\ell}(\theta \,|\, \boldsymbol{x}) + O(\delta^2) \ = \ O(\delta^2),$$

so that it approximately holds that

$$\boldsymbol{x} - \boldsymbol{\beta}(\theta + \delta) \quad \perp \quad \dot{\boldsymbol{\alpha}}(\theta + \delta).$$

That is, (6.4.1) holds by ignoring the second and higher terms of δ. □

This lemma means that the point $\boldsymbol{c}(\theta)$ in the circular mechanism is less variation in the neighborhood with respect to $L(\theta)$.

We shall define a pseudo-length, circle, and semi-line in the circular mechanism as follows : Let $r(\boldsymbol{x},\, \boldsymbol{\beta}(\theta))$ be a pseudo-length such that

$$\boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}}\, (\boldsymbol{x} - \boldsymbol{\beta}(\theta)) \ = \ r(\boldsymbol{x},\, \boldsymbol{\beta}(\theta))\, \frac{\mathbf{S}(\frac{\pi}{2})\, \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}} \dot{\boldsymbol{\alpha}}(\theta)}{\left| \boldsymbol{\Sigma}(\theta)^{\frac{1}{2}} \dot{\boldsymbol{\alpha}}(\theta) \right|}$$

for any $\boldsymbol{x} \in L(\theta)$. It holds that $r(\boldsymbol{c}(\theta),\, \boldsymbol{\beta}(\theta)) = 1/\Gamma_S(\theta)$, and we have the angle $\xi(\theta)$ such that

$$\boldsymbol{e}(\xi(\theta)) \ = \ \begin{pmatrix} \cos \xi(\theta) \\ \sin \xi(\theta) \end{pmatrix} \ = \ \frac{\boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}} \boldsymbol{\beta}(\theta) - \boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}} \boldsymbol{c}(\theta)}{\left| \boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}} \boldsymbol{\beta}(\theta) - \boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}} \boldsymbol{c}(\theta) \right|},$$

where $\xi(\theta) \in [0, 2\pi)$. Thereby the following circle and semi-line are defined by regarding the point $\boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}} \boldsymbol{c}(\theta)$ as the center;

(6.4.2)
$$B(\boldsymbol{c}(\theta)) \ = \ \left\{ \boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}} \boldsymbol{c}(\theta) + \frac{1}{|\Gamma_S(\theta)|}\, \mathbf{S}(t)\, \boldsymbol{e}(\xi(\theta)) \ : \ t \in [0, 2\pi) \right\},$$

(6.4.3)
$$L(\boldsymbol{c}(\theta))^{+} \ = \ \left\{ \boldsymbol{\Sigma}(\theta)^{-\frac{1}{2}} \boldsymbol{c}(\theta) + r\, \boldsymbol{e}(\xi(\theta)) \ : \ r \in [0, \infty) \right\}.$$

We shall prepare the following lemma in order to obtain the next theorem:

Lemma 6.4.2 *In the curved exponential family, the variance matrix $\boldsymbol{\Sigma}(\theta)$ does not depend on the parameter θ if and only if the pivotal probability (density) function $p_0(\boldsymbol{x})$ is the normal density.*

Proof : The sufficiency is trivial. We shall demonstrate the necessity. By Lemma 6.2.1, $\psi(\boldsymbol{\alpha}(\theta))$ satisfies the following differential equation :

$$\nabla' \nabla \psi(\boldsymbol{\alpha}(\theta)) \; = \; \boldsymbol{\Sigma}.$$

Thereby we have

$$\psi(\boldsymbol{\alpha}(\theta)) \; = \; \frac{1}{2}\boldsymbol{\alpha}(\theta)' \,\boldsymbol{\Sigma}\, \boldsymbol{\alpha}(\theta) \; + \; \langle \boldsymbol{\alpha}(\theta), \boldsymbol{b} \rangle \; + \; C,$$

where $\boldsymbol{b}$ is a constant vector and C is a constant. Thus the moment generating function by the parameter $\boldsymbol{\alpha}(\theta)$ of $p_0(\boldsymbol{x})$ is

$$\exp\left\{\frac{1}{2}\boldsymbol{\alpha}(\theta)' \,\boldsymbol{\Sigma}\, \boldsymbol{\alpha}(\theta) \; + \; \langle \boldsymbol{\alpha}(\theta), \boldsymbol{b} \rangle \; + \; C\right\}.$$

By substituting $\boldsymbol{\alpha}(\theta) = \boldsymbol{0}$, it holds that the constant term C is zero, so that, by the relation between the moment generating function and the distribution, $p_0(\boldsymbol{x})$ is the density of the normal distribution with the expectation $\boldsymbol{b}$ and the variance $\boldsymbol{\Sigma}$. The proof is completed. □

The following theorem is important for giving some significances to the quantities with respect to the circular mechanism.

Theorem 6.4.1 *If the pivotal probability (density) function in the two dimensional curved exponential family is the normal density, then, for the curve $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)$,*

1. *The mathematical curvature $\Gamma_M(\theta)$ is equivalent to the statistical curvature $\Gamma_S(\theta)$.*
2. *The point $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{c}(\theta)$ is the center of curvature.*
3. *$B(\boldsymbol{c}(\theta))$ is the osculating circle at $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\alpha}(\theta)$.*

Proof : From Lemma 6.4.2, the variance matrix $\boldsymbol{\Sigma}$ does not depend on θ. Thus, by Theorem 6.3.4, the first result holds, so that it is easy to check the second and third results. The proof is completed. □

For the quantities (6.4.2) and (6.4.3) of a simple circular mechanism, see Figure 6.1.

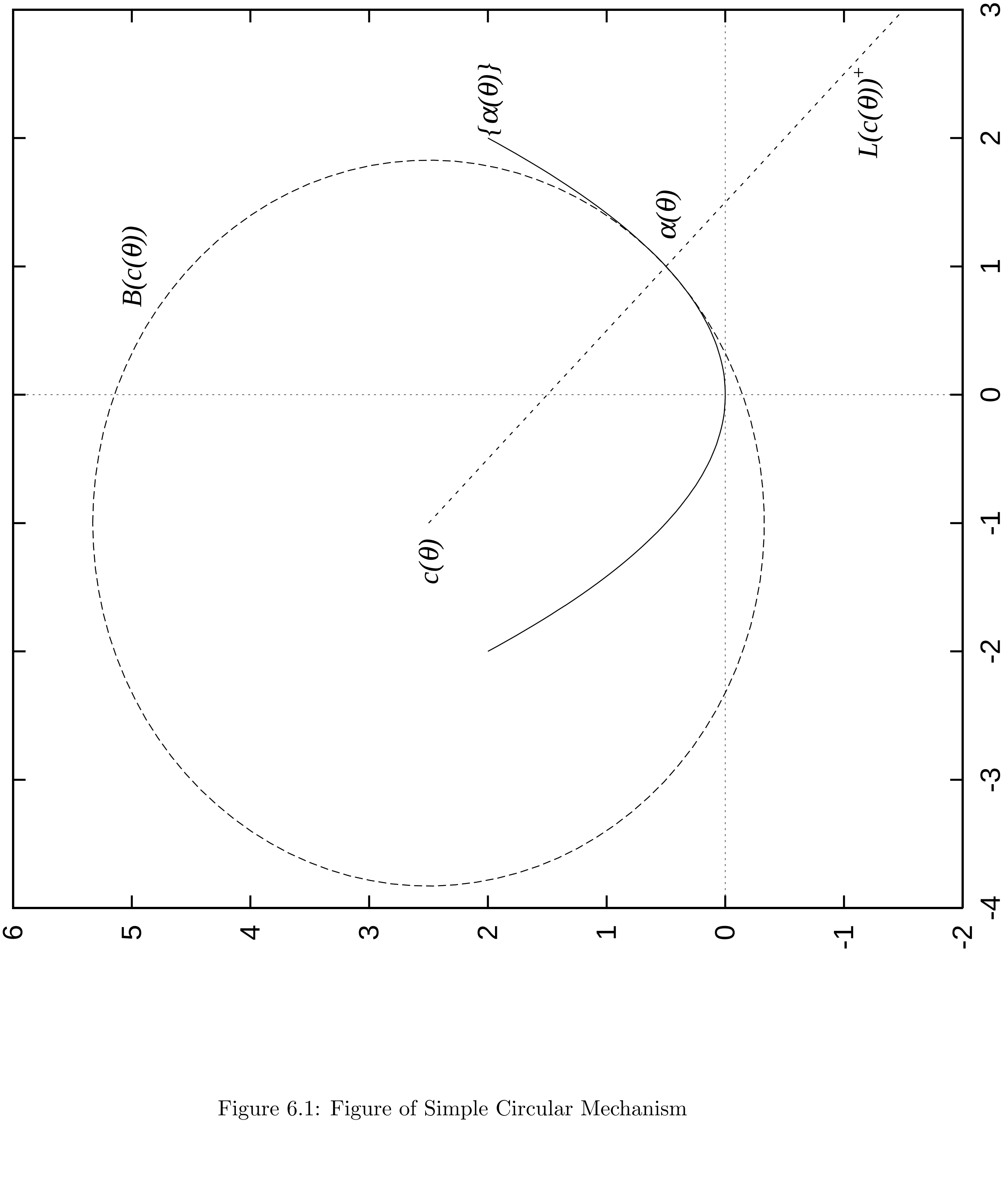

Figure 6.1: Figure of Simple Circular Mechanism

Chapter 7

Appendices

7.1 Convex Conjugate

We shall describe some basic properties about the convex conjugate and Legendre transformation based on Rockafellar(1970).

Let ψ be a function such that

$$\psi \ : \ \mathbf{R}^k \ \longrightarrow \mathbf{R} \cup \{\pm\infty\},$$

where k is an integer. The epigraph of ψ, denoted by $\mathrm{epi}(\psi)$, is defined as follows;

$$\mathrm{epi}(\psi) = \{\, (\boldsymbol{\alpha}, a) \mid \boldsymbol{\alpha} \in \mathbf{R}^k, \ a \in \mathbf{R}, \ a \geq \psi(\boldsymbol{\alpha}) \,\}.$$

The function ψ is defined to be convex on $\mathbf{R}^k$ if the epigraph $\mathrm{epi}(\psi)$ is convex as a subset of $\mathbf{R}^{k+1}$. Note that a function ψ is concave if $\mathrm{epi}(-\psi)$ is convex. And the effective domain of a convex function ψ on $\mathbf{R}^k$, denoted by $\mathrm{dom}(\psi)$, is defined by

$$\mathrm{dom}(\psi) = \{\, \boldsymbol{\alpha} \mid \exists a \text{ such that } (\boldsymbol{\alpha}, a) \in \mathrm{epi}(\psi) \,\} = \{\, \boldsymbol{\alpha} \mid \psi(\boldsymbol{\alpha}) < +\infty \,\}.$$

Note that this domain is convex. A convex function ψ is proper if $\mathrm{dom}(\psi)$ is non-empty and the restriction of ψ to $\mathrm{dom}(\psi)$ is finite. The closure $\mathrm{cl}(\psi)$ of a convex function ψ is defined by

$$\mathrm{cl}(\psi)(\boldsymbol{\alpha}_0) = \liminf_{\boldsymbol{\alpha} \to \boldsymbol{\alpha}_0} \psi(\boldsymbol{\alpha}).$$

A convex function ψ is said to be closed if $\mathrm{cl}(\psi) = \psi$. Therefore the following lemma is known (see Rockafellar, page 51);

Lemma 7.1.1 *The following three properties are equivalent;*

(a) ψ is a closed proper convex function.

(b) $\{\, \boldsymbol{\alpha} \mid \psi(\boldsymbol{\alpha}) \leq a \,\}$ is closed for any $a \in \mathbf{R}$.

(c) $\mathrm{epi}(\psi)$ is a closed set in $\mathbf{R}^{k+1}$. □

Any affine function η on $\mathbf{R}^k$ is represented by

$$\eta(\boldsymbol{\alpha}) = \ \langle \boldsymbol{\alpha},\, \boldsymbol{z} \rangle \ - \ b, \qquad \text{for } \boldsymbol{z} \in \mathbf{R}^k, \ b \in \mathbf{R},$$

where the notation $\langle\ ,\ \rangle$ means the usual inner product in the Euclidean space. Then we have the theorem (See Rockafellar, page 102);

Theorem 7.1.1 *A closed proper convex function ψ is the pointwise supremum of the collection of all affine functions η such that $\eta \leq \psi$.* □

This theorem implies the following corollary (See Rockafellar, page 103);

Corollary 7.1.1 *Given any proper convex function ψ on $\mathbf{R}^k$, there exists some $\boldsymbol{z} \in \mathbf{R}^k$ and $b \in \mathbf{R}$ such that $\psi(\boldsymbol{\alpha}) \geq \ \langle \boldsymbol{\alpha},\, \boldsymbol{z} \rangle - b$ for every $\boldsymbol{\alpha}$.* □

We shall consider the set $\{(\boldsymbol{z}, b)\}$ such that

$$\psi(\boldsymbol{\alpha}) \geq \eta(\boldsymbol{\alpha}) = \ \langle \boldsymbol{\alpha},\, \boldsymbol{z} \rangle - b \quad \text{for epi}(\psi).$$

Since that $\psi(\boldsymbol{\alpha}) \geq \eta(\boldsymbol{\alpha})$ for any $\boldsymbol{\alpha}$ is equal to that

$$b \geq \sup\{\langle \boldsymbol{\alpha},\, \boldsymbol{z} \rangle - \psi(\boldsymbol{\alpha}) \mid \boldsymbol{\alpha} \in \mathbf{R}^k\},$$

by defining the function ψ^* as follows

$$\psi^*(\boldsymbol{z}) = \sup_{\boldsymbol{\alpha}}\{\langle \boldsymbol{\alpha},\, \boldsymbol{z} \rangle - \psi(\boldsymbol{\alpha})\},$$

the set $\{(\boldsymbol{z}, b)\}$ we desire consists with the epigraph of ψ^*. Actually, the function ψ^* is the pointwise supremum of

$$\eta^*(\boldsymbol{z}) = \ \langle \boldsymbol{\alpha},\, \boldsymbol{z} \rangle - a \quad \text{such that } (\boldsymbol{\alpha}, a) \in \text{epi}(\psi).$$

The function ψ^* is called the conjugate of ψ and it is easy to check that ψ^* is a closed proper convex function. Since the function ψ is the pointwise supremum of

$$\eta(\boldsymbol{\alpha}) = \ \langle \boldsymbol{\alpha},\, \boldsymbol{z} \rangle - b \quad \text{such that } (\boldsymbol{z}, b) \in \text{epi}(\psi^*),$$

it holds that

$$\psi(\boldsymbol{\alpha}) = \sup_{\boldsymbol{z}}\{\langle \boldsymbol{\alpha},\, \boldsymbol{z} \rangle - \psi^*(\boldsymbol{z})\},$$

so that the conjugate ψ^{**} of ψ^* is ψ. It is known that the inequality

$$\langle \boldsymbol{\alpha},\, \boldsymbol{z} \rangle \ \leq \psi(\boldsymbol{\alpha}) + \psi^*(\boldsymbol{z}), \quad \text{for any } \boldsymbol{\alpha}, \boldsymbol{z},$$

holds for any proper convex function ψ and its conjugate ψ^*. This inequality is called Fenchel's inequality.

Let a proper convex function ψ be essentially smooth if it satisfies the following three conditions for the interior $\boldsymbol{S}$ of dom(ψ);

(a) $\boldsymbol{S}$ is not empty.

(b) ψ is differentiable throughout $\boldsymbol{S}$.

(c) $\lim_{j\to\infty} |\nabla\psi(\boldsymbol{\alpha})| = +\infty$ whenever $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots$ is a sequence in $\boldsymbol{S}$ converging to a boundary point $\boldsymbol{\alpha}$ of $\boldsymbol{S}$.

Hereafter, we assume that the above three conditions are satisfied when we consider the differentiability of ψ. Note that the notation $\nabla\psi(\boldsymbol{\alpha})$ means the gradient of ψ at $\boldsymbol{\alpha}$, that is,

$$\nabla\psi(\boldsymbol{\alpha}) = \frac{\partial}{\partial\boldsymbol{\alpha}}\,\psi(\boldsymbol{\alpha}).$$

The following theorem is known (See Rockafellar, page 242);

Theorem 7.1.2 *Let ψ be a convex function, and let $\boldsymbol{\alpha}$ be a point where ψ is finite. If ψ is differentiable at $\boldsymbol{\alpha}$, then it holds that*

$$\psi(\boldsymbol{\alpha}_1) \;\geq\; \psi(\boldsymbol{\alpha}) + \;\langle\nabla\psi(\boldsymbol{\alpha}),\, \boldsymbol{\alpha}_1 - \boldsymbol{\alpha}\rangle, \quad \textit{for any } \boldsymbol{\alpha}_1.$$

□

Also we have that the gradient mapping $\nabla\psi : \boldsymbol{\alpha} \to \nabla\psi(\boldsymbol{\alpha})$ is continuous on the set of points where ψ is differentiable.

Now we shall consider the Legendre transformation. Let ψ be a differentiable real-valued function on an open subset $\boldsymbol{S}$ of $\mathbf{R}^k$. The Legendre conjugate of $(\boldsymbol{S}, \psi)$ is defined to be $(\boldsymbol{S}^L, \psi^L)$ where

$$\boldsymbol{S}^L = \nabla\psi(\boldsymbol{S}), \quad \psi^L(\boldsymbol{\alpha}^L) = \;\langle(\nabla\psi)^{-1}(\boldsymbol{\alpha}^L),\, \boldsymbol{\alpha}^L\rangle \;-\; \psi((\nabla\psi)^{-1}(\boldsymbol{\alpha}^L)),$$

and where $(\nabla\psi)^{-1}(\boldsymbol{\alpha}^L) = \{\boldsymbol{\alpha} \mid \nabla\psi(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^L\}$. The Legendre transformation is defined by the transformation from $(\boldsymbol{S}, \psi)$ to the Legendre conjugate $(\boldsymbol{S}^L, \psi^L)$ when the latter is well-defined (that is, single-valued). For a convex case, the following theorem is known (See Rockafellar, page 256);

Theorem 7.1.3 *Let ψ be any closed proper convex function such that the interior $\boldsymbol{S}$ of* dom(ψ) *is non-empty and ψ is differentiable on $\boldsymbol{S}$. The Legendre conjugate $(\boldsymbol{S}^L, \psi^L)$ of $(\boldsymbol{S}, \psi)$ is then well-defined. Moreover, $\boldsymbol{S}^L$ is a subset of* dom(ψ^*)*, and ψ^L is the restriction of ψ^* to $\boldsymbol{S}^L$.* □

Since the gradient mapping $\nabla\psi$ is continuous, under the change of variables $\boldsymbol{z} = \nabla\psi(\boldsymbol{\alpha})$, by Theorem 7.1.3, it holds that

$$\psi^*(\nabla\psi(\boldsymbol{\alpha})) = \;\langle\boldsymbol{\alpha},\, \nabla\psi(\boldsymbol{\alpha})\rangle \;-\; \psi(\boldsymbol{\alpha}).$$

If the gradient mapping $\nabla\psi$ is one-to-one, then it holds that

$$\psi^*(\boldsymbol{z}) = \;\langle(\nabla\psi)^{-1}(\boldsymbol{z}),\, \boldsymbol{z}\rangle \;-\; \psi((\nabla\psi)^{-1}(\boldsymbol{z})).$$

7.2 Information circle

We shall point out that the osculating circle $B(\boldsymbol{c}(\theta))$ in the circular mechanism is different from the information circle which Efron(1978) showed. The definition of information circle is that, for any fixed $\boldsymbol{\alpha}_0 \in A$ and a constant $d \geq 0$,

$$\mathcal{C}_A \;=\; \{\,\boldsymbol{\alpha} \in A \;:\; I(\boldsymbol{\alpha}_0 \,\|\, \boldsymbol{\alpha}) = d\,\},$$

where

$$I(\boldsymbol{\alpha}_0 \,\|\, \boldsymbol{\alpha}) = \int \log \frac{f(\boldsymbol{x} : \boldsymbol{\alpha}_0)}{f(\boldsymbol{x} : \boldsymbol{\alpha})} \; f(\boldsymbol{x} : \boldsymbol{\alpha}_0) \; d\boldsymbol{x}$$

is the Kullback-Leibler information. We shall show its counterexample by Fisher's circle model. Fisher's circle model is the two dimensional normal distribution with the expectation $\boldsymbol{\beta}(\theta) = \rho\, \boldsymbol{e}(\theta)$ and the variance $\mathbf{I}_2$ where $\boldsymbol{e}(\theta) = (\cos\theta,\, \sin\theta)'$ and ρ is a positive constant. Then the density is represented by

$$f(\boldsymbol{x} : \theta) = \exp\left\{ \langle \boldsymbol{\beta}(\theta),\, \boldsymbol{x} \rangle - \frac{|\,\boldsymbol{\beta}(\theta)\,|^2}{2} \right\} \cdot \frac{1}{2\pi} \exp\left\{ -\frac{|\,\boldsymbol{x}\,|^2}{2} \right\}.$$

Note that $\boldsymbol{\alpha}(\theta) = \boldsymbol{\beta}(\theta)$. Since the center of curvature is, by Theorem 6.3.1 in the circular mechanism, $\boldsymbol{c}(\theta) = \mathbf{0}$ for any fixed θ, the osculating circle $B(\boldsymbol{c}(\theta))$ is equivalent to the original expectation circle, that is, the radius is ρ. On the other hand, the information circle of $\boldsymbol{\alpha}_0 = \mathbf{0}$ for $\boldsymbol{\alpha}(\theta)$ is

$$I(\mathbf{0} \,\|\, \boldsymbol{\alpha}(\theta)) \;=\; \frac{|\,\boldsymbol{\alpha}(\theta)\,|^2}{2} \;=\; \frac{\rho^2}{2} \quad (\neq\; \rho).$$

thus the information circle is not equivalent to the osculating circle except for $\rho = 2$.

7.3 Fundamental of Amari's frame

We shall explain the fundamental of Amari's frame(1985). In general, we assume that the parametric space $\mathcal{A}$ is a subset of $\mathbf{R}^k$, that is, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$. As a set of density $f(\boldsymbol{x} : \boldsymbol{\alpha})$, let $S = \{\, f(\boldsymbol{x} : \boldsymbol{\alpha}) \,\}$ be a statistical model parameterized by $\boldsymbol{\alpha}$. Then Amari said

> *When the density $f(\boldsymbol{x} : \boldsymbol{\alpha})$ is sufficiently smooth in $\boldsymbol{\alpha}$, it is natural to introduce in a statistical model S the structure of an $k-$ dimensional manifold, where $\boldsymbol{\alpha}$ plays the role of a coordinate system. (page 12)*

And he assumed the following regularity conditions :

1. All the $f(\boldsymbol{x} : \boldsymbol{\alpha})$'s have a common support so that $f(\boldsymbol{x} : \boldsymbol{\alpha}) > 0$ for all $\boldsymbol{x}$.
2. Let $\ell(\boldsymbol{\alpha} \,|\, \boldsymbol{x}) = \log\, f(\boldsymbol{x} : \boldsymbol{\alpha})$. For every fixed $\boldsymbol{\alpha}$, k functions in $\boldsymbol{x}$

   $$\frac{\partial}{\partial \alpha_j}\, \ell(\boldsymbol{\alpha} \,|\, \boldsymbol{x}), \qquad j = 1, \ldots, k$$

   are linearly independent.
3. The moments of random variables $(\partial/\partial\alpha_j)\, \ell(\boldsymbol{\alpha} \,|\, \boldsymbol{x})$ exist up to necessary orders.
4. For any measurable function $a(\boldsymbol{x},\, \boldsymbol{\alpha})$,

   $$\frac{\partial}{\partial \alpha_j} \int a(\boldsymbol{x},\, \boldsymbol{\alpha})\, f(\boldsymbol{x} : \boldsymbol{\alpha})\, d\,\boldsymbol{x} \;=\; \int \frac{\partial}{\partial \alpha_j}\, a(\boldsymbol{x},\, \boldsymbol{\alpha})\, f(\boldsymbol{x} : \boldsymbol{\alpha})\, d\,\boldsymbol{x}.$$

Let $T_{\boldsymbol{\alpha}}(S)$ be the tangent vector space at $\boldsymbol{\alpha} \in S$. By the second regularity condition, the bases of $T_{\boldsymbol{\alpha}}(S)$ are represented as the following vectors

$$\partial_j \;=\; \partial_j \, \ell(\boldsymbol{\alpha} \,|\, \boldsymbol{x}) \;=\; \frac{\partial}{\partial \alpha_j} \, \ell(\boldsymbol{\alpha} \,|\, \boldsymbol{x}), \qquad j = 1, \ldots, k \ ,$$

that is, any element A of $T_{\boldsymbol{\alpha}}(S)$ is

$$A \;=\; \sum_{j=1}^{k} A_j \, \partial_j \;=\; \sum_{j=1}^{k} A_j \, \partial_j \, \ell(\boldsymbol{\alpha} \,|\, \boldsymbol{x}).$$

The manifold S is called a Riemannian space if the inner product $\langle A, \, B \rangle$ of two tangent vectors $A, B \in T_{\boldsymbol{\alpha}}(S)$ is defined. Then their inner product is defined by the following :

$$\langle A, \, B \rangle \;=\; E \left[\sum_{i,j=1}^{k} A_i \, B_j \, \partial_i \, \ell(\boldsymbol{\alpha} \,|\, \boldsymbol{x}) \, \partial_j \, \ell(\boldsymbol{\alpha} \,|\, \boldsymbol{x}) \right] .$$

Hence the inner product of the two basis vectors ∂_i and ∂_j is

$$g_{ij}(\boldsymbol{\alpha}) \;=\; \langle \partial_i, \, \partial_j \rangle \;=\; E[\, \partial_i \, \ell(\boldsymbol{\alpha} \,|\, \boldsymbol{x}) \; \partial_j \, \ell(\boldsymbol{\alpha} \,|\, \boldsymbol{x}) \,].$$

k^2 quantities $g_{ij}(\boldsymbol{\alpha})$ are called the metric tensor, so that the manifold S with the metric $g = \{g_{ij}\}$ becomes the Riemannian manifold $(S, \, g)$. The matrix $(g_{ij}(\boldsymbol{\alpha}))$ is known in statistics as Fisher information matrix. Thus the statistical model S is regarded as Euclidean space locally and as Riemannian manifold globally. This is the fundamental of Amari's framework.

Thereby, in Amari's framework, we may need to consider the local behavior of $f(\boldsymbol{x} : \boldsymbol{\alpha})$ at $\boldsymbol{\alpha}$ under the metric g, since the metric g changes with the point $\boldsymbol{\alpha}$. In the exponential family, Fisher information matrix is represented by $I(\boldsymbol{\alpha}) = \boldsymbol{\Sigma}(\boldsymbol{\alpha})$. If the variance matrix is constant as we considered some cases in the circular mechanism, then the metric g is constant in Amari's framework, so that the statistical model S is regard as Euclidean space globally.

References

[1] S. Amari(1985), *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics, **28**, Springer-Verlag

[2] J. Doob(1934), Probability and Statistics, *Trans. Amer. Math. Soc.*, **36**, 759–775

[3] B. Efron(1975), Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency), *Ann. Statist.*, **3**, 1189–1242

[4] B. Efron(1978), The Geometry of Exponential Families, *Annals of Statistics*, **6**, 362–376

[5] R. A. Fisher(1925), Theory of Statistical Estimation, *Proc. Camb. Phil. Soc.*, **22**, 700–725

[6] R. A. Fisher(1956), *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh

[7] P. R. Halmos and L. J. Savage(1949), Application of the Radon-Nikodym theorem to the theory of sufficient statistics, *Ann. Math. Statist.*, **20**, 225–241

[8] N. Inagaki(1983), The Decomposition of the Fisher Information, *Ann. Inst. Statist. Math.*, **35**, 151–165

[9] N. Inagaki(1990), *Statistical Mathematics*, Shokabo, Tokyo (in Japanese)

[10] N. Inagaki and E. Kumagai(1996), Exact Information Loss in Fisher's Circle Model, *Mathematica Japonica*, **44**, 455–467

[11] E. Kumagai and N. Inagaki(1996), Comment on Efron's Counterexample, *Mathematica Japonica*, **44**, 449–454

[12] E. Kumagai and N. Inagaki(1996), The Circular Mechanism in the Curved Exponential Family, (submitted)

[13] C. R. Rao(1961), Asymptotic Efficiency and Limiting Information, *Proc. Forth Berkeley Symposium on Math. Statist. and Probab.*, **1**, 531–546

[14] R. T. Rockafellar(1970), *Convex Analysis*, Princeton University Press