Information Geometry of Statistical Models

Xi Yang

A Thesis
in
The Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts (Mathematics) at
Concordia University
Montreal, Quebec, Canada

June 2018

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Xi Yang

Entitled: Information Geometry of Statistical Models

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Prof. Arusharka Sen Chair
Prof. Frédéric Godin Examiner
Prof. Arusharka Sen Examiner
Prof. Alina Stancu Supervisor

Approved by ____

Chair of Department or Graduate Program Director

____________ 2018 ____

Dean of Faculty of Arts and Science

ABSTRACT

Information Geometry of Statistical Models

Xi Yang

Information Geometry is a relatively young branch of Mathematics, which roots back to studies of invariant geometrical structure involved in statistical inference. It defines a Riemannian metric together with dually coupled affine connections in a manifold of probability distributions. These structures provide tools not only for studying statistical inference but also for research in wider areas of information sciences, such as machine learning, signal processing, optimization, and even neuroscience, not to mention mathematics and physics. The aim of this thesis is to give a brief introduction to Information Geometry with focus on the exponential family. In Chapter 1, we first introduce the notion and basic properties of statistical models. We then define some common notions in information geometry such as Fisher information, Christoffel symbols, connections, Skewness tensor, geodesic and Jeffreys Prior. We also introduce the geometry of entropy, including entropy, Kullback-Leibler divergence (or relative entropy) and information energy on statistical models. Chapter 2 focuses on the geometry of the exponential family of probability distributions. Examples and properties of exponential families are firstly discussed in this chapter. Fisher metric and geodesics are worked out explicitly for common exponential families. Chapter 3 contains important examples of exponential families for which the entropy, Kullback-Leibler relative entropy and information energy are worked out explicitly. This chapter deals also with the problem of finding the density of max-

imum entropy subject to the first N moment constraints, with unique solutions for the cases $N \leq 2$.

Acknowledgments

I would first like to thank my supervisor Professor Dr. Alina Stancu very much for all the guidance and help that she has provided throughout my study at Concordia University. Our meetings on a regular basis, her patience and knowledge helped very much in completing this thesis. I am particularly grateful to Dr. Stancu for her choosing the topic of this thesis with consideration to my interest.

I am grateful to the scholarship scheme of the Institut des sciences mathématiques, a consortium of nine Québec universities, and the department of mathematics and statistics at Concordia University for making it possible for me to study here.

I would also like to thank many professors and the staff at Concordia University for their help and encouragement on my study in mathematics and statistics. Among them, Professors Frédéric Godin, Arusharka Sen and Wei Sun have kindly given me access to audit their courses which are closely related to this thesis work.

I have been very fortunate to be surrounded by many great friends. I am thankful for their support on my study and the happy time that we spent together.

Finally, I must express my very profound gratitude to my family for providing me with unfailing support and continuous encouragement throughout my years of study.

Contents

Chapter 1

Introduction to Information Geometry of Statistical Models

Information geometry explores the world of information by means of modern geometry. It is a method to characterize the structure of statistical models from a viewpoint of differential geometry. By considering families of probability distributions as manifolds with coordinate charts determined by the parameters of each individual model, the tools of differential geometry such as divergences and metric tensors provide additional means to study statistical inference, information loss, and estimation.

Information geometry traces its roots back to the work of C. R. Rao in the mid-1940s [6]. Rao developed a way to measure the statistical distance between two populations through a Riemannian metric which was shown to be equivalent to Fisher's information matrix. Further contributions were made in the decades followed by H. Jeffreys, D. Cox, B. Efron, O. Barndorff-Nielsen, N. N. Chentsov, and

S. Amari among many others [1]. Information geometry reached maturity through the work of S. Amari and other Japanese mathematicians in the 1980s. Today, information geometry is a filed that is increasingly attracting the interest of researchers from many different areas of science, including mathematics, statistics, geometry, computer science, signal processing, physics and neuroscience [4]. It is an active area of research with international conferences held regularly. Springer is launching a new topical journal "Information Geometry" in 2018.

The thesis can be read as a brief introduction to information geometry. It is structured into three chapters.

In Chapter 1, we introduce the notion of statistical models, which is a space of density functions, and discuss the basic properties of statistical models. In this chapter, we also define some common notions in information geometry such as Fisher information, Christoffel symbols, connections, Skewness tensor, autoparallel curves and Jeffreys Prior. We also introduce the geometry of entropy, including entropy, Kullback-Leibler divergence (or relative entropy) and information energy on statistical models.

Chapter 2 focuses on the geometry of the exponential family of probability distributions. The exponential family is not only a typical statistical model, including many well-known families of probability distributions, but is associated with a convex function (used in the definition for each exponential family). Examples and properties of exponential families are discussed in this chapter. Fisher metric and geodesics are worked out explicitly for common exponential families.

Chapter 3 contains important examples of exponential families for which the entropy, Kullback-Leibler relative entropy and information energy are worked out explicitly. This chapter deals also with the problem of finding the density of maximum entropy subject to the first N moment constraints.

The thesis work is mainly based on the book of O. Calin and C. Udrişte (C&U) [4], published in 2014. Other important reference used for this work are the book of S. Amari [2], published in 2016, and the book of Ay et al. [3], published in 2017, which claiming the standard reference of the field. In this thesis, while many propositions and corollaries are cited directly or in a modified form from the book [4], many propositions and useful results for the exponential family of probability distributions have also been obtained by the author and supplemented. Some mistakes and typos in C&U's book [4] have also been corrected in this thesis.

1.1 Statistical Models

We first introduce the notion of statistical models by associating it with a family of probability distributions. We restrict our work on the statistical models given parametrically. When the family of distributions can be described smoothly by a set of parameters, it can be considered as a multidimensional hypersurface. Upon specifying the parameters of a distribution, we determine a unique element of the family or a unique point on the hypersurface.

1.1.1 Probability Spaces and Random Variables

Let $(\mathrm{S}, \mathcal{F}, P)$ be a *probability space*, where the finite or infinite set S is the sample space, $\mathcal{F}$ is a *σ-field* over S, and P is the *probability measure*. A *random variable* X is a measurable function used to measure the random outcomes contained in S.

A *discrete random variable* X takes finite or countably infinite values, $X : \mathrm{S} \rightarrow \mathcal{X} = \{x^1, x^2, x^3, \ldots\}$. The probability for $X(s) = x^k \in \mathcal{X}$ is described by a *probability mass function* $p : \mathcal{X} \rightarrow [0, 1]$

$$p_k = p\left(x^k\right) = P\left(X = x^k\right) = P\left(\left\{s \in \mathrm{S}; X(s) = x^k \in \mathcal{X}\right\}\right), \forall k \geq 1,$$

satisfying $\sum_{k \geq 1} p_k = 1$. A *discrete probability distribution* is characterized by a probability mass function. The function $p(x)$ defines a *probability distribution function* $F : \mathcal{X} \rightarrow [0, 1]$,

$$F(x) = \sum_{k=1}^{N} p(x),\ \forall x^N \leq x \leq x^{N+1}.$$

The Bernoulli distribution, binomial distribution, the geometric distribution and the Poisson distribution are among the most common discrete probability distributions.

A *continuous random variable* X takes a continuous range of values, $X : \mathrm{S} \rightarrow \mathcal{X} \subset \mathbb{R}^n$. The probability for $X \in \mathcal{D}$, an open set in $\mathcal{X}$, is described by a *probability density function* $p : \mathcal{X} \rightarrow [0, 1]$ satisfying $\int_{\mathcal{X}} p(x) = 1$,

$$P(X \in \mathcal{D}) = \int_{\mathcal{D}} p(x)\, dx.$$

A *continuous probability distribution* is characterized by a probability density function. When $\mathcal{X} = \mathbb{R}$, $p(x)$ defines a probability distribution function $F : \mathbb{R} \to [0, 1]$,

$$F(x) = \int_{-\infty}^{x} p(t)\, dt.$$

The normal distribution, the lognormal distribution, the exponential distribution, the gamma and the beta distributions are some most well-known examples of continuous probability distributions.

1.1.2 Parametric Models

Parameters are descriptive measures of the characteristics of a population that may be used as the inputs for a probability distribution function. This section deals with a family of probability density functions described by a set of parameters. Such a family can be organized as a parameterized hypersurface, each point on the hypersurface representing a probability density.

Let $\mathcal{S} = \{p_\xi = p(x; \xi) \,|\, \xi = (\xi^1, \ldots, \xi^n) \in \mathbb{E}\}$ be a family of probability distributions on $\mathcal{X}$, where each element p_ξ can be parameterized by n real-valued variables $\xi = (\xi^1, \ldots, \xi^n)$ and the set $\mathbb{E} \subset \mathbb{R}^n$ is called the *parameters space*. The set $\mathcal{S}$ is a subset of the infinite dimensional space of functions

$$\mathcal{P}(\mathcal{X}) = \left\{ f;\ f : \mathcal{X} \to \mathbb{R}, f \geq 0, \int_{\mathcal{X}} f\, dx = 1 \right\}.$$

Definition 1.1.1. *The set* $\mathcal{S} = \{p_\xi = p(x; \xi) \,|\, \xi = (\xi^1, \ldots, \xi^n) \in \mathbb{E}\}$ *is called a sta-*

tistical model or a parametric model of dimension n *if the mapping*

$$\iota : \mathbb{E} \to \mathcal{P}(\mathcal{X}),\ \iota(\xi) = p_\xi$$

is one-to-one and has rank $n = \dim \mathbb{E}$*.*

The rank n implies that $\{\partial_j p_\xi\}_{j=1}^{n}$ is a set of linearly independent functions, where $\partial_j = \frac{\partial}{\partial \xi^j}$. This condition defines the *regularity* of the statistical model.

The one-to-one condition of the mapping $\iota : \mathbb{E} \to \mathcal{P}(\mathcal{X}),\ \iota(\xi) = p_\xi$ for a statistical model implies that it is reasonable to consider the inverse function $\phi : \mathcal{S} \to \mathbb{E} \subset \mathbb{R}^n, \phi(p_\xi) = \xi$. Since ϕ assigns a parameter ξ to each p_ξ, we can take ϕ as a *coordinate system* for our statistical model.

Although a statistical model may change its parametrization, the geometric results obtained in one parametrization are valid for all parametrization. Thus, it is better to choose a convenient parametrization to work with.

1.1.3 Basic Properties of Statistical Models

We will also use the abbreviations $\mathcal{S} = \{p_\xi\}$ and $\mathcal{S} = \{p(x;\xi)\}$ when there is no doubt on the parameters space or sample space. The functions $\partial_j p_\xi(x)$, or denoted by $\varphi_j(x;\xi)$, are basic vector fields for the model $\mathcal{S} = \{p_\xi\}$. The vector field φ_j is a differentiation on smooth mapping $f : \mathcal{S} \to \mathcal{F}(\mathcal{X}, \mathbb{R})$

$$\varphi_j(f) = \frac{\partial(f(p_\xi))}{\partial \xi^j}.$$

A frequently-used mapping is the *log-likelihood function* $\ell : \mathcal{S} \to \mathcal{F}(\mathcal{X}, \mathbb{R})$ defined by

$$\ell(p_\xi)(x) = \ln p_\xi(x),$$

which is sometimes denoted by $\ell_x(\xi) = \ell(p_\xi(x))$. Its derivatives are

$$\varphi_j \ell_x(\xi) = \frac{\partial \ln p_\xi(x)}{\partial \xi^j} = \varphi_j(\ell_x(\xi)),\ 1 \leq j \leq n,$$

which play a core role in the information geometry of statistical models.

It is often easier to check the linear independence of $\{\partial_j \ell_x(\xi)\}_{j=1}^n$ than of $\{\partial_j p_\xi\}_{j=1}^n$.

Theorem 1.1.1. *[4] The regularity condition in the definition of the statistical model $\mathcal{S} = \{p_\xi\}$ holds if and only if for any $\xi \in \mathbb{E}$ the set $\{\partial_j \ell_x(\xi)\}_{j=1}^n$ is a system of n linearly independent functions of x.*

Proof. Since

$$\partial_j \ell_x(\xi) = \frac{\partial}{\partial \xi^j} \ln p(x;\xi) = \frac{1}{p(x;\xi)} \frac{\partial}{\partial \xi^j} p(x;\xi) = \frac{1}{p(x;\xi)} \partial_j p_\xi(x), \tag{1.1}$$

the two systems $\{\partial_j \ell_x(\xi)\}_{j=1}^n$ and $\{\partial_j p_\xi\}_{j=1}^n$ are proportional. Hence, their linear independence is equivalent.

□

We will often use the facts that,

$$\int_{\mathcal{X}} \varphi_j(x)\, dx = \partial_j \int_{\mathcal{X}} p_\xi(x)\, dx = \partial_j 1 = 0,\ \forall j \in \{1, \ldots, n\}; \tag{1.2}$$

for continuous distributions, and

$$\sum_{k\geq 1}\varphi_j\left(x_k\right)=\partial_j\sum_{k\geq 1}p_\xi\left(x_k\right)=\partial_j 1=0,\ \forall 1\leq j\leq n. \tag{1.3}$$

for discrete distributions. In equation (1.2), the interchangeability of the derivative with the integral holds if assuming the boundedness of $\mathcal{X}$ or the integrability of $p_\xi\left(x\right)$; while in equation (1.3), the derivative can be taken out of the sum by assuming $\mathcal{X}$ finite or the uniform convergence of the series.

Theorem 1.1.2. *[4]Assume that the equations (1.2) and (1.3) hold. The expectation of $\partial_j\ell_x\left(\xi\right)$ with respect to p_ξ is zero,*

$$E_\xi\left[\partial_j\ell_x\left(\xi\right)\right]=0. \tag{1.4}$$

Proof. According to equations (1.1) and (1.2), we have

$$\begin{aligned}E_\xi\left[\partial_j\ell_x\left(\xi\right)\right]&=E_\xi\left[\frac{1}{p\left(x;\xi\right)}\partial_j p_\xi\left(x\right)\right]=E_\xi\left[\frac{1}{p\left(x;\xi\right)}\varphi_j\left(x;\xi\right)\right]\\&=\int_{\mathcal{X}}\varphi_j\left(x;\xi\right)dx=0.\end{aligned}$$

Similarly, in the discrete case, we have

$$E_\xi\left[\partial_j\ell_x\left(\xi\right)\right]=\sum_{k\geq 1}p_\xi\left(x_k\right)\partial_j\ln p_\xi\left(x_k\right)=\sum_{k\geq 1}\partial_j p_\xi\left(x_k\right)=\partial_j\sum_{k\geq 1}p_\xi\left(x_k\right)=0.$$

□

1.2 Information Geometry of Statistical Models

1.2.1 Fisher Information

The Riemannian metric tensor g is a fundamental object in differential geometry. Similarly, we define a metric structure on a statistical model.

Definition 1.2.1. *The Fisher information matrix for a statistical model with the parameter* $\xi = (\xi^1, \dots, \xi^n) \in \mathbb{E}$ *is defined by*

$$g_{ij}(\xi) = E_\xi\left[\partial_i \ell(\xi)\, \partial_j \ell(\xi)\right], \quad \forall i, j \in \{1, \dots, n\}, \tag{1.5}$$

where $\ell(\xi)$ *and* ∂_i *denotes* $\ln p_\xi(x)$ *and* $\frac{\partial}{\partial \xi^i}$ *respectively.*

Proposition 1.2.1. *[4] The Fisher information matrix can be represented as*

$$g_{ij}(\xi) = 4 \int_{\mathcal{X}} \partial_i \sqrt{p_\xi(x)}\partial_j \sqrt{p_\xi(x)} dx \tag{1.6}$$

The discrete analogue is

$$g_{ij}(\xi) = 4 \sum_{\mathcal{X}} \partial_i \sqrt{p_\xi(x)}\partial_j \sqrt{p_\xi(x)}. \tag{1.7}$$

Proof. We prove the discrete case. The proof to the continuous case is similar.

$$\begin{aligned}
g_{ij}(\xi) &= E_\xi\left[\partial_i \ell(\xi)\,\partial_j \ell(\xi)\right] \\
&= \sum_{\mathcal{X}} p_\xi(x)\,\partial_i \ln p_\xi(x)\,\partial_j \ln p_\xi(x) \\
&= \sum_{\mathcal{X}} p_\xi(x)\,\frac{\partial_i p_\xi(x)}{p_\xi(x)}\frac{\partial_j p_\xi(x)}{p_\xi(x)} \\
&= 4\sum_{\mathcal{X}} \frac{\partial_i p_\xi(x)}{2\sqrt{p_\xi(x)}}\frac{\partial_j p_\xi(x)}{2\sqrt{p_\xi(x)}} \\
&= 4\sum_{\mathcal{X}} \partial_i \sqrt{p_\xi(x)}\partial_j \sqrt{p_\xi(x)}.
\end{aligned}$$

□

Proposition 1.2.2. *[4] The Fisher information matrix on any statistical model is symmetric, positive definite and non-degenerate.*

Proof. The symmetry follows from the definition (1.5).

For a statistical model $\mathcal{S}$ with the parameters ξ, $\forall v \in T_\xi\mathcal{S}$ and $v \neq 0$, we have

$$\begin{aligned}
v^t g v = \sum_{i,j} g_{ij} v^i v^j &= 4\sum_{i,j}\int_{\mathcal{X}}\left(v^i\partial_i\sqrt{p_\xi(x)}v^j\partial_j\sqrt{p_\xi(x)}\right)dx \\
&= 4\int_{\mathcal{X}}\left(\sum_i v^i\partial_i\sqrt{p_\xi(x)}\right)\left(\sum_j v^j\partial_j\sqrt{p_\xi(x)}\right)dx \\
&= 4\int_{\mathcal{X}}\left(\sum_i v^i\partial_i\sqrt{p_\xi(x)}\right)^2 dx \geq 0.
\end{aligned}$$

So, g is non-negative definite.

Since we have

$$v^t g v = 0 \Longleftrightarrow \int_{\mathcal{X}} \left(\sum_i v^i \partial_i \sqrt{p_\xi (x)} \right)^2 dx = 0 \Longleftrightarrow$$

$$\sum_i v^i \partial_i \sqrt{p_\xi (x)} = 0 \Longleftrightarrow \sum_i v^i \partial_i p_\xi (x) = 0,$$

by the linear independence of $\{\partial_i p_\xi (x)\}$ for a statistical model, $v^i = 0$ for all i, $i.e.\, v = 0$, which contradicts our assumption. Thus, g is non-degenerate and positive-definite.

□

The Fisher information matrix provides the coefficients of a Riemannian metric on the hypersurface $\mathcal{S}$. We can measure distances, angles, and define connections on statistical models by Fisher metric.

Theorem 1.2.1. *[4] The Fisher information matrix can be represented as*

$$g_{ij} (\xi) = -E_\xi \left[\partial_i \partial_j \ell (\xi) \right]. \tag{1.8}$$

Proof. We start from the normalization condition

$$\int_{\mathcal{X}} p_\xi (x) \, dx = 1.$$

Differentiating with respect to ξ^i yields

$$\int_{\mathcal{X}} \partial_i p_\xi (x) \, dx = 0.$$

So, we have

$$E_\xi\left[\partial_i \ell\left(\xi\right)\right] = \int_{\mathcal{X}} \partial_i \ln p_\xi\left(x\right) \cdot p_\xi\left(x\right) dx = \int_{\mathcal{X}} \partial_i p_\xi\left(x\right) dx = 0.$$

Differentiating in $\int_{\mathcal{X}} \partial_i \ln p_\xi\left(x\right) \cdot p_\xi\left(x\right) dx = 0$ again with respect to ξ^j yields

$$\begin{aligned}
&\int_{\mathcal{X}} \partial_j \partial_i \ln p_\xi\left(x\right) \cdot p_\xi\left(x\right) dx + \int_{\mathcal{X}} \partial_i \ln p_\xi\left(x\right) \cdot \partial_j p_\xi\left(x\right) dx = 0 \\
\Longleftrightarrow\ & E_\xi\left[\partial_j \partial_i \ln p_\xi\left(x\right)\right] + \int_{\mathcal{X}} \partial_i \ln p_\xi\left(x\right) \partial_j \ln p_\xi\left(x\right) \cdot p_\xi\left(x\right) dx = 0 \\
\Longleftrightarrow\ & -E_\xi\left[\partial_j \partial_i \ln p_\xi\left(x\right)\right] = E_\xi\left[\ln p_\xi\left(x\right) \partial_j \ln p_\xi\left(x\right)\right] = g_{ij}.
\end{aligned}$$

□

Proposition 1.2.3. *[4] The Fisher metric is invariant under reparametrizations of the sample space.*

Proof. Consider an invertible transform of sample spaces $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$, defined by $Y = f\left(X\right)$. Denote by $p_\xi\left(x\right)$ and $\tilde{p}_\xi\left(y\right)$ the density functions associated with the random variables X and Y respectively. The relation between $p_\xi\left(x\right)$ and $\tilde{p}_\xi\left(y\right)$ is given by

$$p_\xi\left(x\right) = \tilde{p}_\xi\left(y\right) \frac{\partial f\left(x\right)}{\partial x}.$$

Since the log-likelihood functions are given by

$$\ell\left(\xi\right) = \ln p_\xi\left(x\right) = \ln \tilde{p}_\xi\left(y\right) + \ln \frac{\partial f\left(x\right)}{\partial x},$$

we have

$$\partial_{\xi^i} \ln p_\xi\left(x\right) = \partial_{\xi^i} \ln \tilde{p}_\xi\left(y\right).$$

Hence,

$$\begin{aligned} g_{ij}\left(\xi\right) &= \int_{\mathcal{X}} \partial_i \ln p_\xi\left(x\right) \partial_j \ln p_\xi\left(x\right) \cdot p_\xi\left(x\right) dx \\ &= \int_{\mathcal{X}} \partial_i \ln \tilde{p}_\xi\left(y\right) \partial_j \ln \tilde{p}_\xi\left(y\right) \cdot \tilde{p}_\xi\left(y\right) \frac{\partial f\left(x\right)}{\partial x} dx \\ &= \int_{\mathcal{Y}} \partial_i \ln \tilde{p}_\xi\left(y\right) \partial_j \ln \tilde{p}_\xi\left(y\right) \tilde{p}_\xi\left(y\right) dy \\ &= \tilde{g}_{ij}\left(\xi\right). \end{aligned}$$

□

Theorem 1.2.2. *[4] The Fisher metric is covariant under reparametrizations of the parameters space.*

Proof. Let $\xi = \left(\xi^1, \ldots, \xi^n\right)$ and $\eta = \left(\eta^1, \ldots, \eta^n\right)$ be two sets of parameters related by the invertible relationship $\xi = \xi\left(\eta\right)$. Let $\tilde{p}_\eta\left(x\right) = p_{\xi(\eta)}\left(x\right)$.

By chain rule, we have

$$\partial_{\eta^i} \tilde{p}_\eta = \frac{\partial}{\partial \eta^i} \tilde{p}_\eta = \sum_k \frac{\partial \xi^k}{\partial \eta^i} \partial_{\xi^k} p_\xi, \ \partial_{\eta^j} \tilde{p}_\eta = \frac{\partial}{\partial \eta^j} \tilde{p}_\eta = \sum_r \frac{\partial \xi^r}{\partial \eta^j} \partial_{\xi^r} p_\xi.$$

So, we have the covariance relation between the components of g and $\tilde{g}$,

$$\begin{aligned}
\tilde{g}_{ij}(\eta) &= \int_{\mathcal{X}} \partial_{\eta^i} \ln \tilde{p}_\eta(x)\, \partial_{\eta^j} \ln \tilde{p}_\eta(x) \cdot \tilde{p}_\eta(x)\, dx \\
&= \int_{\mathcal{X}} \frac{1}{\tilde{p}_\eta(x)} \partial_{\eta^i} \tilde{p}_\eta(x)\, \partial_{\eta^j} \tilde{p}_\eta(x)\, dx \\
&= \sum_k \sum_r \left[\int_{\mathcal{X}} \frac{1}{p_{\xi(\eta)}(x)} \partial_{\xi^k} p_\xi \partial_{\xi^r} p_\xi dx \right] \frac{\partial \xi^k}{\partial \eta^i} \frac{\partial \xi^r}{\partial \eta^j} \\
&= \sum_{k,r} g_{kr}(\xi)\,|_{\xi=\xi(\eta)} \frac{\partial \xi^k}{\partial \eta^i} \frac{\partial \xi^r}{\partial \eta^j} \\
&= \sum_{k,r} g_{kr}(\xi)\,|_{\xi=\xi(\eta)} J_{ki} J_{rj},
\end{aligned}$$

where J is the Jacobian matrix $J(\xi, \eta)$. Writing in the matrix form, we have $\tilde{g}(\eta) = J^t g(\xi) J$, where J^t is the transpose of J. □

1.2.2 Christoffel Symbols

The Christoffel symbol is the most simple connection on the statistical model $\mathcal{S}$.

Definition 1.2.2. *Let g_{ij} denote a Riemannian metric, particularly the Fisher information matrix, then the Christoffel symbols of first kind are defined by*

$$\mathit{\Gamma}_{ij,k} = \frac{1}{2} \left(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij} \right), \tag{1.9}$$

where we used the notation $\partial_i = \partial_{\xi^i}$. The Christoffel symbols of second kind are *defined by*

$$\mathit{\Gamma}_{ij}^{p} = \frac{1}{2} g^{pk} \left(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij} \right), \tag{1.10}$$

Proposition 1.2.4. *[4] The Christoffel symbols of first kind can be represented in the following equivalent forms.*

(i) $\Gamma_{ij,k} = \frac{1}{2} \left(E_\xi \left[\left(\partial_i \partial_j \ell \right) \partial_k \ell \right] - E_\xi \left[\left(\partial_j \partial_k \ell \right) \partial_i \ell \right] - E_\xi \left[\left(\partial_k \partial_i \ell \right) \partial_j \ell \right] - E_\xi \left[\partial_i \partial_j \partial_k \ell \right] \right);$

(ii) $\Gamma_{ij,k} = E_\xi \left[\left(\partial_i \partial_j \ell + \frac{1}{2} \partial_i \ell \partial_j \ell \right) \partial_k \ell \right];$

(iii) $\Gamma_{ij,k} = 4 \int_{\mathcal{X}} \partial_i \partial_j \sqrt{p(x;\xi)} \partial_k \sqrt{p(x;\xi)} dx.$

Proof. By equation (1.8),

$$\begin{aligned}
\partial_k g_{ij}(\xi) &= -\partial_k E_\xi \left[\partial_i \partial_j \ell \right] \\
&= -\partial_k \int_{\mathcal{X}} \left(\partial_i \partial_j \ell \right) p(x;\xi) dx \\
&= -\int_{\mathcal{X}} \left(\partial_k \partial_i \partial_j \ell \right) p(x;\xi) dx - \int_{\mathcal{X}} \left(\partial_i \partial_j \ell \right) \partial_k p(x;\xi) dx \\
&= -\int_{\mathcal{X}} \left(\partial_k \partial_i \partial_j \ell \right) p(x;\xi) dx - \int_{\mathcal{X}} \left(\partial_i \partial_j \ell \right) \left(\partial_k \ell \right) p(x;\xi) dx \\
&= -E_\xi \left[\partial_k \partial_i \partial_j \ell \right] - E_\xi \left[\left(\partial_i \partial_j \ell \right) \left(\partial_k \ell \right) \right].
\end{aligned}$$

Similarly, we have

$$\partial_i g_{jk}(\xi) = -E_\xi \left[\partial_i \partial_j \partial_k \ell \right] - E_\xi \left[\left(\partial_j \partial_k \ell \right) \left(\partial_i \ell \right) \right],$$

$$\partial_j g_{ki}(\xi) = -E_\xi \left[\partial_i \partial_j \partial_k \ell\right] - E_\xi \left[\left(\partial_k \partial_i \ell\right)\left(\partial_j \ell\right)\right].$$

Substituting the above three results into (1.9) gives (i).

By the definition equation (1.5), we have

$$\begin{aligned}
\partial_k g_{ij}(\xi) &= \partial_k E_\xi \left[\partial_i \ell \partial_j \ell\right] \\
&= \partial_k \int_{\mathcal{X}} \left(\partial_i \ell \partial_j \ell\right) p(x;\xi) dx \\
&= \int \left(\partial_k \partial_i \ell\right)\left(\partial_j \ell\right) p(x;\xi) dx + \int \left(\partial_k \partial_j \ell\right)\left(\partial_i \ell\right) p(x;\xi) dx + \int_{\mathcal{X}} \left(\partial_i \ell \partial_j \ell\right) \partial_k p(x;\xi) dx \\
&= E_\xi \left[\partial_k \partial_i \ell \partial_j \ell\right] + E_\xi \left[\partial_k \partial_j \ell \partial_i \ell\right] + E_\xi \left[\partial_i \ell \partial_j \ell \partial_k \ell\right]. \qquad (1.11)
\end{aligned}$$

Similarly, we have

$$\partial_i g_{jk}(\xi) = E_\xi \left[\partial_i \partial_j \ell \partial_k \ell\right] + E_\xi \left[\partial_i \partial_k \ell \partial_j \ell\right] + E_\xi \left[\partial_j \ell \partial_k \ell \partial_i \ell\right],$$

$$\partial_j g_{ki}(\xi) = E_\xi \left[\partial_j \partial_k \ell \partial_i \ell\right] + E_\xi \left[\partial_j \partial_i \ell \partial_k \ell\right] + E_\xi \left[\partial_k \ell \partial_i \ell \partial_j \ell\right].$$

Substituting the above three results into (1.9) gives (ii).

By Proposition 1.2.1, we have

$$\begin{aligned}
\partial_k g_{ij}\left(\xi\right) &= 4\partial_k \int_{\mathcal{X}} \partial_i \sqrt{p_\xi\left(x\right)} \partial_j \sqrt{p_\xi\left(x\right)} dx \\
&= 4 \int_{\mathcal{X}} \partial_k \partial_i \sqrt{p_\xi\left(x\right)} \partial_j \sqrt{p_\xi\left(x\right)} dx + 4 \int_{\mathcal{X}} \partial_i \sqrt{p_\xi\left(x\right)} \partial_k \partial_j \sqrt{p_\xi\left(x\right)} dx.
\end{aligned}$$

Similarly, we have

$$\partial_i g_{jk} = 4\int_{\mathcal{X}} \partial_i\partial_j\sqrt{p_\xi\left(x\right)}\partial_k\sqrt{p_\xi\left(x\right)}dx + 4\int_{\mathcal{X}} \partial_j\sqrt{p_\xi\left(x\right)}\partial_i\partial_k\sqrt{p_\xi\left(x\right)}dx,$$

$$\partial_j g_{ki} = 4\int_{\mathcal{X}} \partial_j\partial_k\sqrt{p_\xi\left(x\right)}\partial_i\sqrt{p_\xi\left(x\right)}dx + 4\int_{\mathcal{X}} \partial_k\sqrt{p_\xi\left(x\right)}\partial_j\partial_i\sqrt{p_\xi\left(x\right)}dx.$$

Substituting all three results into (1.9), we obtain (*iii*). □

1.2.3 Connections

Definition 1.2.3. *The coefficients given by* (1.5) *induce a Riemannian metric on $\mathcal{S}$, which is a 2-covariant tensor g defined locally by*

$$g\left(X_\xi,\, Y_\xi\right) = \sum_{i,j=1}^{n} g_{ij}\left(\xi\right)a^i\left(\xi\right)b^j\left(\xi\right),\ p_\xi \in \mathcal{S} \tag{1.12}$$

where $X_\xi = \sum_{i=1}^n a^i(\xi)\partial_i p_\xi$ and $Y_\xi = \sum_{j=1}^n b^j(\xi)\partial_j p_\xi$ are vector fields in the 0-representation on $\mathcal{S}$. Observe that $\{\partial_i p_\xi\}_{i=1}^n$, or simplified as $\{\partial_i\}_{i=1}^n$, forms a basis of the tangent space $T_\xi\mathcal{S}$. The tensor g is called the $Fisher$-Riemannian metric. Its associated Levi-Civita connection is denoted by $\nabla^{(0)}$ and is defined by

$$g\left(\nabla^{(0)}_{\partial_i}\partial_j, \partial_k\right) = \Gamma^{(0)}_{ij,k}, \tag{1.13}$$

where $\Gamma^{(0)}_{ij,k}$ is the Christoffel symbols of first kind defined in (1.9).

Definition 1.2.4. *Using the Fisher metric g, the $\nabla^{(1)}$-connection is defined by*

$$g\left(\nabla^{(1)}_{\partial_i}\partial_j,\, \partial_k\right) = \mathit{\Gamma}^{(1)}_{ij,k}(\xi) = E_\xi\left[\left(\partial_i\partial_j\ell\right)\partial_k\ell\right], \tag{1.14}$$

where ℓ is the log-likelihood function.

Definition 1.2.5. *Using the Fisher metric g, the $\nabla^{(-1)}$-connection on a statistical model $\mathcal{S}$ is defined by*

$$g\left(\nabla^{(-1)}_{\partial_i}\partial_j,\, \partial_k\right) = \mathit{\Gamma}^{(-1)}_{ij,k}(\xi) = E_\xi\left[\left(\partial_i\partial_j\ell + \partial_i\ell\partial_j\ell\right)\left(\partial_k\ell\right)\right], \tag{1.15}$$

where ℓ is the log-likelihood function.

Proposition 1.2.5. *[4] The relation among the foregoing three connections is given by*

$$\nabla^{(0)} = \frac{1}{2}\left(\nabla^{(1)} + \nabla^{(-1)}\right). \tag{1.16}$$

Proof. It suffices to show

$$\mathit{\Gamma}^{(0)}_{ij,k} = \tfrac{1}{2}\left(\mathit{\Gamma}^{(-1)}_{ij,k} + \mathit{\Gamma}^{(1)}_{ij,k}\right).$$

By equations (1.14), (1.15) and *(ii)* in Proposition 1.2.4, we have

$$\begin{aligned}
\mathit{\Gamma}^{(-1)}_{ij,k} + \mathit{\Gamma}^{(1)}_{ij,k} &= E_\xi\left[\left(\partial_i\partial_j\ell + \partial_i\ell\partial_j\ell\right)\left(\partial_k\ell\right)\right] + E_\xi\left[\left(\partial_i\partial_j\ell\right)\partial_k\ell\right] \\
&= 2E_\xi[\left(\partial_i\partial_j\ell + \frac{1}{2}\partial_i\ell\partial_j\ell\right)\partial_k\ell] \\
&= 2\mathit{\Gamma}^{(0)}_{ij,k}.
\end{aligned}$$

□

Proposition 1.2.6. *[4] For any vector fields X, Y, Z on the statistical model $\mathcal{S} = \{p_\xi\}$, we have*

$$Zg\left(X,Y\right) = g\left(\nabla_Z^{(1)} X, Y\right) + g\left(X, \nabla_Z^{(-1)} Y\right). \tag{1.17}$$

Proof. Choosing $X = \sum\limits_i a_i \partial_i p_\xi$, $Y = \sum\limits_j b_j \partial_j p_\xi$ and $Z = \sum\limits_k c_k \partial_k p_\xi$ with the basis $\{\partial_i p_\xi\}_{i=1}^n$ of the tangent space $T_\xi \mathcal{S}$, we have

$$Zg\left(X,Y\right) = Z\left(\sum_{i,j} g_{ij} a_i b_j\right) = \sum_k c_k \partial_k \left(\sum_{i,j} g_{ij} a_i b_j\right).$$

By equations (1.14) and (1.15), we have

$$g\left(\nabla_Z^{(1)} X, Y\right) = \sum_k c_k g\left(\nabla_{\partial_k}^{(1)} X, Y\right) = \sum_k c_k \sum_{i,j} a_i b_j \mathit{\Gamma}_{ki,j}^{(1)},$$

$$g\left(X, \nabla_Z^{(-1)} Y\right) = \sum_k c_k g\left(X, \nabla_{\partial_k}^{(-1)} Y\right) = \sum_k c_k \sum_{i,j} a_i b_j \mathit{\Gamma}_{kj,i}^{(-1)},$$

Thus, it suffices to prove

$$\partial_k g_{ij} = \mathit{\Gamma}_{kj,i}^{(-1)} + \mathit{\Gamma}_{ki,j}^{(1)}.$$

By equations (1.14), (1.15) and (1.11), we have

$$\begin{aligned} \mathit{\Gamma}_{kj,i}^{(-1)} + \mathit{\Gamma}_{ki,j}^{(1)} &= E_\xi\left[\left(\partial_k \partial_j \ell + \partial_k \ell \partial_j \ell\right)\left(\partial_i \ell\right)\right] + E_\xi\left[\left(\partial_k \partial_i \ell\right) \partial_j \ell\right] \\ &= \partial_k g_{ij}. \end{aligned}$$

□

Definition 1.2.6. *We define the 1-parameter family of connections*

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)}, \tag{1.18}$$

with real parameter α on the statistical model $\mathcal{S}$. Using the Fisher metric g, the connection components are denoted by

$$\Gamma^{(\alpha)}_{ij,k} = g\left(\nabla^{(\alpha)}_{\partial_i}\partial_j, \partial_k\right).$$

The connection components are given by the following proposition.

Proposition 1.2.7. *[4]*

$$\Gamma^{(\alpha)}_{ij,k} = E_\xi\left[\left(\partial_i\partial_j\ell + \frac{1-\alpha}{2}\partial_i\ell\partial_j\ell\right)\partial_k\ell\right]. \tag{1.19}$$

Proof. Writing definition (1.18) in terms of components and then use equations (1.14) and (1.15), we obtain

$$\begin{aligned}
\Gamma^{(\alpha)}_{ij,k} &= \frac{1+\alpha}{2}\Gamma^{(1)}_{ij,k} + \frac{1-\alpha}{2}\Gamma^{(-1)}_{ij,k} \\
&= \frac{1+\alpha}{2}E_\xi\left[\left(\partial_i\partial_j\ell\right)\partial_k\ell\right] + \frac{1-\alpha}{2}E_\xi\left[\left(\partial_i\partial_j\ell + \partial_i\ell\partial_j\ell\right)\left(\partial_k\ell\right)\right] \\
&= E_\xi[\left(\partial_i\partial_j\ell + \frac{1-\alpha}{2}\partial_i\ell\partial_j\ell\right)\partial_k\ell].
\end{aligned}$$

□

Proposition 1.2.8. *[4] For any vector fields X, Y, Z on the statistical model $\mathcal{S} =$*

$\{p_\xi\}$, *we have*

$$Zg\left(X,Y\right) = g\left(\nabla^{(\alpha)}_Z X, Y\right) + g\left(X, \nabla^{(-\alpha)}_Z Y\right). \tag{1.20}$$

Proof. Similar to the proof to Proposition (1.2.6), it suffices to show

$$\partial_k g_{ij} = \mathit{\Gamma}^{(-\alpha)}_{kj,i} + \mathit{\Gamma}^{(\alpha)}_{ki,j}.$$

By equations (1.19) and (1.11), we find

$$\begin{aligned}\mathit{\Gamma}^{(-\alpha)}_{kj,i} + \mathit{\Gamma}^{(\alpha)}_{ki,j} &= E_\xi[\left(\partial_k\partial_j\ell + \frac{1+\alpha}{2}\partial_k\ell\partial_j\ell\right)\partial_i\ell] + E_\xi[\left(\partial_k\partial_i\ell + \frac{1-\alpha}{2}\partial_k\ell\partial_i\ell\right)\partial_j\ell] \\ &= E_\xi\left[\partial_i\ell\partial_j\ell\partial_k\ell\right] + E_\xi\left[\left(\partial_k\partial_j\ell\right)\partial_i\ell\right] + E_\xi\left[\left(\partial_k\partial_i\ell\right)\partial_j\ell\right] \\ &= \partial_k g_{ij}\left(\xi\right).\end{aligned}$$

□

1.2.4 Skewness Tensor

The difference of two linear connections is a tensor field.

Definition 1.2.7. *We define the generalized difference tensor by*

$$K^{(\alpha,\beta)}\left(X,Y\right) = \nabla^{(\beta)}_X Y - \nabla^{(\alpha)}_X Y. \tag{1.21}$$

We define a 3-covariant, symmetric tensor T with components

$$T\left(\partial_i,\,\partial_j,\,\partial_k\right) = T_{ijk} = E_\xi\left[\partial_i\ell\,\partial_j\ell\,\partial_k\ell\right]. \tag{1.22}$$

T is called the skewness tensor.

Proposition 1.2.9. *[4] The skewness tensor T satisfies the following equation:*

$$g\left(K^{(\alpha,\beta)}\left(X,Y\right),Z\right)=\frac{\alpha-\beta}{2}T\left(X,Y,Z\right), \tag{1.23}$$

where g is the Fisher metric.

Proof. We prove the local coordinates version of the above equation by applying Proposition 1.2.7:

$$\begin{aligned}\Gamma^{(\beta)}_{ij,k}(\xi)-\Gamma^{(\alpha)}_{ij,k}(\xi)&=E_\xi[\left(\partial_i\partial_j\ell+\frac{1-\beta}{2}\partial_i\ell\partial_j\ell\right)\partial_k\ell]-E_\xi[\left(\partial_i\partial_j\ell+\frac{1-\alpha}{2}\partial_i\ell\partial_j\ell\right)\partial_k\ell]\\&=\frac{\alpha-\beta}{2}E_\xi\left[\partial_i\ell\,\partial_j\ell\,\partial_k\ell\right].\end{aligned}$$

□

Proposition 1.2.10. *The skewness tensor is covariant under reparametrizations.*

Proof. Let $T_{ijk}\left(\xi\right)=E_\xi\left[\partial_i\ell\partial_j\ell\partial_k\ell\right]$ and $\eta^i=\eta^i\left(\xi^1,\ldots,\xi^n\right)$ for $i\in\{1,\ldots,n\}$ be a reparametrization of ξ. Then, similar to the proof to Theorem1.2.2,

$$\begin{aligned}
\tilde{T}_{ij,k}(\eta) &= E_{\eta}\left[\partial_i \ell \partial_j \ell \partial_k \ell\right] \\
&= \int_{\mathcal{X}} \partial_{\eta^i} \ln \tilde{p}_{\eta}(x)\, \partial_{\eta^j} \ln \tilde{p}_{\eta}(x)\, \partial_{\eta^k} \ln \tilde{p}_{\eta}(x) \cdot \tilde{p}_{\eta}(x)\, dx \\
&= \int_{\mathcal{X}} \frac{\partial_{\eta^i} \tilde{p}_{\eta}(x)\, \partial_{\eta^j} \tilde{p}_{\eta}(x)\, \partial_{\eta^k} \tilde{p}_{\eta}(x)}{\left(\tilde{p}_{\eta}(x)\right)^2} dx \\
&= \sum_{a,b,c} \left[\int_{\mathcal{X}} \frac{\partial_{\xi^a} p_{\xi}(x)\, \partial_{\xi^b} p_{\xi}(x)\, \partial_{\xi^c} p_{\xi}(x)}{\left(p_{\xi(\eta)}(x)\right)^2} dx\right] \frac{\partial \xi^a}{\partial \eta^i} \frac{\partial \xi^b}{\partial \eta^j} \frac{\partial \xi^c}{\partial \eta^k} \\
&= \sum_{a,b,c} \left[\int_{\mathcal{X}} \partial_{\xi^a} \ln p_{\xi}(x)\, \partial_{\xi^b} \ln p_{\xi}(x)\, \partial_{\xi^c} \ln p_{\xi}(x) \cdot p_{\xi(\eta)}(x)\, dx\right] \frac{\partial \xi^a}{\partial \eta^i} \frac{\partial \xi^b}{\partial \eta^j} \frac{\partial \xi^c}{\partial \eta^k} \\
&= \sum_{a,b,c} T_{ab,c}(\xi)\, J_{ai} J_{bj} J_{ck}.
\end{aligned}$$

□

Proposition 1.2.11. *The skewness tensor is invariant under transformations of the random variable.*

Proof. Consider an invertible transform of sample spaces $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$, defined by $Y = f(X)$. Denote by $p_{\xi}(x)$ and $\tilde{p}_{\xi}(y)$ the density functions associated with the random variables X and Y respectively. The relation between $p_{\xi}(x)$ and $\tilde{p}_{\xi}(y)$ is given by

$$p_{\xi}(x) = \tilde{p}_{\xi}(y) \frac{\partial f(x)}{\partial x}.$$

Since the log-likelihood functions are given by

$$\ell(\xi) = \ln p_{\xi}(x) = \ln \tilde{p}_{\xi}(y) + \ln \frac{\partial f(x)}{\partial x},$$

we have

$$\partial_{\xi^i} \ln p_\xi(x) = \partial_{\xi^i} \ln \tilde{p}_\xi(y) \, .$$

Hence,

$$\begin{aligned} T_{ij,k}(\xi) &= \int_{\mathcal{X}} \partial_i \ln p_\xi(x) \, \partial_j \ln p_\xi(x) \, \partial_k \ln p_\xi(x) \cdot p_\xi(x) \, dx \\ &= \int_{\mathcal{X}} \partial_i \ln \tilde{p}_\xi(y) \, \partial_j \ln \tilde{p}_\xi(y) \, \partial_k \ln \tilde{p}_\xi(y) \cdot \tilde{p}_\xi(y) \, \frac{\partial f(x)}{\partial x} dx \\ &= \int_{\mathcal{Y}} \partial_i \ln \tilde{p}_\xi(y) \, \partial_j \ln \tilde{p}_\xi(y) \, \partial_k \ln \tilde{p}_\xi(y) \cdot \tilde{p}_\xi(y) \, dy \\ &= \tilde{T}_{ij,k}(\xi) \, . \end{aligned}$$

□

1.2.5 Autoparallel Curves

A curve on the statistical model $\mathcal{S} = \{p_\xi\}$ is defined by

$$\gamma(s) = \iota(\xi(s)) = p_{\xi}(s), \quad s \in [0, \, T] \, . \tag{1.24}$$

The velocity of the curve $\gamma(s)$ is given by

$$\dot{\gamma}(s) = \iota_* \left(\dot{\xi}(s) \right) = \frac{d}{ds} \iota \left(\dot{\xi}(s) \right) = \frac{d}{ds} p_\xi(s) \, . \tag{1.25}$$

Definition 1.2.8. *A curve $\gamma \, : \, [0, \, T] \to \mathcal{S}$ is called $\nabla^{(\alpha)}$-autoparallel if $\dot{\gamma}(s)$ is parallel transported along $\gamma(s)$. So, the acceleration with respect to the $\nabla^{(\alpha)}$-connection* *vanishes,*

$$\nabla^{(\alpha)}_{\dot{\gamma}(s)}\dot{\gamma}\left(s\right)=0,\quad \forall s\in\left[0,\,T\right]. \tag{1.26}$$

In local coordinates $\gamma\left(s\right)=\left(\gamma^{k}\left(s\right)\right)$, the autoparallelism means, for each fixed k,

$$\ddot{\gamma}^{k}\left(s\right)+\sum_{i,j}\mathit{\Gamma}^{(\alpha)k}_{ij}\dot{\gamma}^{i}\left(s\right)\dot{\gamma}^{j}\left(s\right)=0, \tag{1.27}$$

where, for fixed i, j, $\mathit{\Gamma}^{(\alpha)k}_{ij}=\sum_{l}\mathit{\Gamma}^{(\alpha)}_{ij,l}g^{lk}$ is evaluated along $\gamma\left(s\right)$, and g denotes the Fisher metric. This is a Riccati system of ordinary differential equations (ODEs). In particular, if $\alpha=0$, the autoparallel curves become geodesics with respect to the Fisher metric g.

Proposition 1.2.12. *[4] Let $\alpha\neq\beta$. If a curve $\gamma\left(s\right)$ is both $\nabla^{(\alpha)}$ and $\nabla^{(\beta)}$-autoparallel, then*

$$T\left(\dot{\gamma}\left(s\right),\dot{\gamma}\left(s\right),X\right)=0,$$

for any vector field X along the curve $\gamma\left(s\right)$.

Proof. By equation (1.26), we have $\nabla^{(\alpha)}_{\dot{\gamma}(s)}\dot{\gamma}\left(s\right)=0$ and $\nabla^{(\beta)}_{\dot{\gamma}(s)}\dot{\gamma}\left(s\right)=0$. Then, by equation (1.21), we have

$$K^{(\alpha,\beta)}\left(\dot{\gamma}\left(s\right),\dot{\gamma}\left(s\right)\right)=\nabla^{(\alpha)}_{\dot{\gamma}(s)}\dot{\gamma}\left(s\right)-\nabla^{(\beta)}_{\dot{\gamma}(s)}\dot{\gamma}\left(s\right)=0.$$

Hence, by Proposition 1.2.9, we obtain $T\left(\dot{\gamma}\left(s\right),\dot{\gamma}\left(s\right),X\right)=0$ for any vector field X along the curve $\gamma\left(s\right)$. □

1.2.6 Jeffreys Prior

Definition 1.2.9. *Let $\mathcal{S} = \{p_\xi ; \xi \in \mathbb{E}\}$ be a statistical model, and $G(\xi) = \det g(\xi)$ denote the determinant of the Fisher information matrix. Assume the volume*

$$Vol(\mathcal{S}) = \int_{\mathbb{E}} \sqrt{G(\xi)} d\xi < \infty. \tag{1.28}$$

Then, Jeffreys prior is a probability distribution on $\mathbb{E}$, defined by

$$Q(\xi) = \frac{1}{Vol(\mathcal{S})} \sqrt{G(\xi)}. \tag{1.29}$$

Proposition 1.2.13. *Jeffreys prior is invariant under reparametrizations of the parameters space $\mathbb{E}$.*

Proof. Let $\xi = (\xi^1, \ldots, \xi^n)$ and $\eta = (\eta^1, \ldots, \eta^n)$ be two sets of parameters of a statistical model $\mathcal{S}$. Suppose we have the invertible relationship $\xi = \xi(\eta)$.

By Theorem 1.2.2, the Fisher metric is covariant under reparametrizations of the parameters space, i.e.

$$\tilde{g}(\eta) = J(\xi, \eta)^t g(\xi(\eta)) J(\xi, \eta).$$

So, we have

$$\begin{aligned}\tilde{G}(\eta) = \det \tilde{g}(\eta) &= \det \left(J(\xi,\eta)^t \, g(\xi(\eta)) \, J(\xi,\eta) \right) \\ &= \det J(\xi,\eta)^t \det g(\xi(\eta)) \det J(\xi,\eta) \\ &= G(\xi(\eta)) \left[\det J(\xi,\eta) \right]^2 ,\end{aligned}$$

$$\begin{aligned}\tilde{Vol}(\mathcal{S}) &= \int_{\mathbb{E}} \sqrt{\tilde{G}(\eta)} d\eta \\ &= \int_{\mathbb{E}} \sqrt{G(\xi(\eta))} |\det J(\xi,\eta)| d\eta \\ &= \int_{\mathbb{E}} \sqrt{G(\xi)} d\xi \\ &= Vol(\mathcal{S}) .\end{aligned}$$

Hence, we have

$$\begin{aligned}\widetilde{Q}(\eta) d\eta &= \frac{\sqrt{\tilde{G}(\eta)}}{\tilde{Vol}(\mathcal{S})} d\eta \\ &= \frac{\sqrt{G(\xi(\eta))} |\det J(\xi,\eta)|}{Vol(\mathcal{S})} d\eta \\ &= Q(\xi) \, d\xi.\end{aligned}$$

□

1.3 The Geometry of Entropy on Statistical Models

1.3.1 Entropy

Definition 1.3.1. *The entropy is a function $H : \mathbb{E} \to \mathbb{R}$, which is defined by*

$$H(\xi) = \begin{cases} -\int_{\mathcal{X}} p(x,\xi) \ln p(x,\xi)\, dx, & \text{if } \mathcal{X} \text{ is continuous;} \\ -\sum_{x \in \mathcal{X}} p(x,\xi) \ln p(x,\xi), & \text{if } \mathcal{X} \text{ is discrete.} \end{cases} \tag{1.30}$$

Thus, the entropy is equal to the negative of the expectation of the log-likelihood function, $H(\xi) = -E_{p_\xi}[\ell_x(\xi)]$.

Proposition 1.3.1. *[4] The entropy is a concave function, i.e. for any densities $p_1, \ldots, p_n$ on $\mathcal{X}$ and $\lambda_i \in [0,1]$ with $\sum_{i=1}^{n} \lambda_i = 1$, we have*

$$H\left(\sum_{i=1}^{n} \lambda_i p_i\right) \geq \sum_{i=1}^{n} \lambda_i H(p_i). \tag{1.31}$$

Proof. We know that $f(u) = -u \ln u$ is concave for $u \in \mathbb{R}^+$. So, we have

$$f\left(\sum_{i=1}^{n} \lambda_i p_i\right) \geq \sum_{i=1}^{n} \lambda_i f(p_i).$$

Integrating (summing in the discrete case) over $\mathcal{X}$ leads to

$$H\left(\sum_{i=1}^{n} \lambda_i p_i\right) \geq \sum_{i=1}^{n} \lambda_i H(p_i).$$

□

Definition 1.3.2. *Let $\mathcal{S} = \{p_\xi(x); x \in \mathcal{X}, \xi \in \mathbb{E}\}$ be a statistical model. A point $q \in \mathcal{S}$ is a critical point for the entropy H if*

$$X(H) = 0,\ \forall X \in T_q\mathcal{S}. \tag{1.32}$$

Proposition 1.3.2. *[4] The probability distribution p_ξ is a critical point of the entropy H if and only if*

$$\int_{\mathcal{X}} \ln p(x,\xi)\,\partial_{\xi^i} p(x,\xi)\,dx = 0,\ \forall i \in \{1,\ldots,n\}. \tag{1.33}$$

In the discrete case, the above equation is replaced by

$$\sum_{x\in\mathcal{X}} \ln p(x,\xi)\,\partial_{\xi^i} p(x,\xi) = 0,\ \forall i \in \{1,\ldots,n\} \tag{1.34}$$

Proof. Choose $X = \partial_{\xi^i} = \partial_i$. Since $\{\partial_i\}$ form a basis of T_qS, by Definition 1.3.2, we obtain that the point $q = p_\xi \in \mathcal{S}$ is a critical point for H if and only if

$$\partial_i H(\xi) = 0,\ \forall i \in \{1,\ldots,n\},$$

i.e.

$$\begin{aligned}\partial_i H &= -\partial_i \int_{\mathcal{X}} p\left(x,\xi\right) \ln p\left(x,\xi\right) dx \\ &= -\int_{\mathcal{X}} \left(\partial_i p\left(x,\xi\right) \ln p\left(x,\xi\right) dx + \partial_i p\left(x,\xi\right)\right) dx \\ &= -\int_{\mathcal{X}} \ln p\left(x,\xi\right) \partial_i p\left(x,\xi\right) dx = 0,\end{aligned}$$

or, similarly in the discrete case,

$$\partial_i H = \sum_{x\in\mathcal{X}} \ln p\left(x,\xi\right) \partial_{\xi^i} p\left(x,\xi\right) = 0.$$

□

Proposition 1.3.3. *[4] The Hessian of the entropy is given by*

$$\partial_i \partial_j H\left(\xi\right) = -g_{ij}\left(\xi\right) - h_{ij}(\xi),$$

where $g_{ij}\left(\xi\right)$ is the Fisher-Riemann metric and

$$h_{ij}(\xi) = E_\xi\left[\left(\partial_j \ell\left(\xi\right) \partial_i \ell\left(\xi\right) + \partial_i \partial_j \ell\left(\xi\right)\right) \ell\left(\xi\right)\right].$$

Proof. The first derivative of the entropy can be expressed as

$$\begin{aligned}
\partial_i H\left(\xi\right) &= -\int_{\mathcal{X}} \ln p\left(x,\xi\right)\partial_i p\left(x,\xi\right)dx \\
&= -\int_{\mathcal{X}} p\left(x,\xi\right)\ln p\left(x,\xi\right)\partial_i \ln p\left(x,\xi\right)dx \\
&= -E_{\xi}\left[\ell\left(\xi\right)\partial_i \ell\left(\xi\right)\right].
\end{aligned}$$

Hence the Hessian of the entropy can be expressed as

$$\begin{aligned}
\partial_j\partial_i H\left(\xi\right) &= -\partial_j \int_{\mathcal{X}} p\left(x,\xi\right)\ell\left(\xi\right)\partial_i \ell\left(\xi\right)dx \\
&= -\int_{\mathcal{X}} \left(p\left(x,\xi\right)\partial_j \ell\left(\xi\right)\partial_i \ell\left(\xi\right) + \partial_j p\left(x,\xi\right)\ell\left(\xi\right)\partial_i \ell\left(\xi\right) + p\left(x,\xi\right)\ell\left(\xi\right)\partial_j\partial_i \ell\left(\xi\right)\right)dx \\
&= -\int_{\mathcal{X}} \left(p\left(x,\xi\right)\partial_j \ell\left(\xi\right)\partial_i \ell\left(\xi\right) + p\left(x,\xi\right)\partial_j \ell\left(\xi\right)\ell\left(\xi\right)\partial_i \ell\left(\xi\right) + p\left(x,\xi\right)\ell\left(\xi\right)\partial_j\partial_i \ell\left(\xi\right)\right)dx \\
&= -E_{\xi}\left[\partial_i \ell\left(\xi\right)\partial_j \ell\left(\xi\right)\right] - E_{\xi}\left[\left(\partial_j \ell\left(\xi\right)\partial_i \ell\left(\xi\right) + \partial_i\partial_j \ell\left(\xi\right)\right)\ell\left(\xi\right)\right] \\
&= -g_{ij}\left(\xi\right) - h_{ij}(\xi).
\end{aligned}$$

□

1.3.2 Kullback-Leibler Relative Entropy

Definition 1.3.3. *The Kullback-Leibler relative entropy is a non-commutative measure of the difference between two probability densities* p *and* q *on the same statistical* *hypersurface, and it is defined by*

$$D_{KL}\left(p||q\right) = E_p\left[\ln\frac{p}{q}\right] = \begin{cases} \int_{\mathcal{X}} p\left(x\right)\ln\dfrac{p\left(x\right)}{q\left(x\right)}dx, & \text{if } \mathcal{X} \text{ is continuous;} \\ \sum_{x^i\in\mathcal{X}} p\left(x^i\right)\ln\dfrac{p\left(x^i\right)}{q\left(x^i\right)}, & \text{if } \mathcal{X} \text{ is discrete.} \end{cases}$$

The Kullback-Leibler relative entropy can also be expressed as the expectation of the difference of two log-likelihood functions

$$D_{KL}\left(p||q\right) = E_p\left[\ell_p\right] - E_p\left[\ell_q\right].$$

Proposition 1.3.4. *[4] Let $\mathcal{S}$ be a statistical manifold.*

(i) The relative entropy $D_{KL}\left(p||q\right) \geq 0$ for any $p, q \in \mathcal{S}$ with $D_{KL}\left(p||q\right) = 0$ if and only if $p = q$.

(ii) The relative entropy is symmetric, i.e. $D_{KL}\left(p||q\right) = D_{KL}\left(q||p\right)$ if and only if

$$\int_{\mathcal{X}} \left(p\left(x\right) + q\left(x\right)\right)\ln\frac{p\left(x\right)}{q\left(x\right)}dx = 0.$$

(iii) The relative entropy satisfies the triangle inequality, i.e. $D_{KL}\left(p||q\right)+D_{KL}\left(q||r\right) \geq D_{KL}\left(p||r\right)$ if and only if

$$\int_{\mathcal{X}} \left(p\left(x\right) - q\left(x\right)\right)\ln\frac{q\left(x\right)}{r\left(x\right)}dx \leq 0.$$

Proof. (*i*) Since $\ln x \leq -1 + x$ for $x > 0$ with equality if and only if $x = 1$, we have

$$\begin{aligned}\int_{\mathcal{X}} p(x) \ln \frac{q(x)}{p(x)} dx &\leq \int_{\mathcal{X}} p(x) \left(-1 + \frac{q(x)}{p(x)} \right) dx \\ &= \int_{\mathcal{X}} q(x)\, dx - \int_{\mathcal{X}} p(x)\, dx \\ &= 0\end{aligned}$$

with equality if and only if $p = q$.

(*ii*) and (*iii*) are obtained by direct computations. □

Note that in the previous proposition the notations are only given for the continuous cases. This proposition shows that the Kullback-Leibler relative entropy does not satisfy all the axioms of a metric on the manifold $\mathcal{S}$. The non-symmetry can be removed by defining the symmetric *Kullback-Leibler quasimetric* $\mathcal{D}(p, q)$, $\mathcal{D}(p, q) = D_{KL}(p||q) + D_{KL}(q||p)$. However, in general, the problem of the triangle inequality cannot be fixed.

Definition 1.3.4. *The cross entropy of p with respect to q is defined as*

$$S(p, q) = -E_p[\ln q] = \begin{cases} -\int_{\mathcal{X}} p(x) \ln q(x) dx, & \textit{if } \mathcal{X} \textit{ is continuous;} \\ -\sum_{x \in \mathcal{X}} p(x) \ln q(x), & \textit{if } \mathcal{X} \textit{ is discrete.} \end{cases}$$

By direct computations, we have the following result.

Proposition 1.3.5. *[4] The relative entropy $D_{KL}(p||q)$, the entropy $H(p)$ and the* *cross entropy $S(p,q)$ are related by*

$$S(p,q) = D_{KL}(p||q) + H(p). \tag{1.35}$$

The following proposition gives the minimum of cross entropy, i.e. $\min_q S(p,q) = H(p)$.

Proposition 1.3.6. *[4] The entropy $H(p)$ and the cross entropy $S(p,q)$ satisfy the inequality*

$$S(p,q) \geq H(p) \tag{1.36}$$

with equality if and only if $p = q$.

Proof. This is the result of (i) in Proposition 1.3.4 and Proposition 1.3.5. □

Proposition 1.3.7. *[4] The diagonal part of the first variation of the Kullback-Leibler relative entropy is zero,*

$$\partial_{\xi^i} D_{KL}(p_{\xi_0}||p_\xi)|_{\xi=\xi_0} = \partial_i D_{KL}(p_{\xi_0}||p_\xi)|_{\xi=\xi_0} = 0. \tag{1.37}$$

Proof.

$$\begin{aligned}
\partial_i D_{KL}(p_{\xi_0}||p_\xi)|_{\xi=\xi_0} &= \partial_i \int_{\mathcal{X}} p_{\xi_0}(x) \ln \frac{p_{\xi_0}(x)}{p_\xi(x)} dx|_{\xi=\xi_0} \\
&= -\int_{\mathcal{X}} p_{\xi_0}(x)\, \partial_i \ln p_\xi(x)\, dx|_{\xi=\xi_0} \\
&= -\int_{\mathcal{X}} p_{\xi_0}(x)\, \partial_i \ell_x(\xi_0)\, dx \\
&= -E_{\xi_0}[\partial_i \ell_x(\xi_0)] = 0,
\end{aligned}$$

by Theorem 1.1.2. □

Proposition 1.3.8. *[4] The diagonal part of the Hessian of the Kullback-Leibler relative entropy is the Fisher metric,*

$$\partial_i \partial_j D_{KL} \left(p_{\xi_0} || p_\xi \right) |_{\xi = \xi_0} = g_{ij} \left(\xi_0 \right). \tag{1.38}$$

Proof. We have

$$\begin{aligned} \partial_i \partial_j D_{KL} \left(p_{\xi_0} || p_\xi \right) |_{\xi = \xi_0} &= \partial_i \partial_j \int_{\mathcal{X}} p_{\xi_0} \left(x \right) \ln \frac{p_{\xi_0} \left(x \right)}{p_\xi \left(x \right)} dx |_{\xi = \xi_0} \\ &= - \int_{\mathcal{X}} p_{\xi_0} \left(x \right) \partial_i \partial_j \ln p_\xi \left(x \right) dx |_{\xi = \xi_0} \\ &= - \int_{\mathcal{X}} p_{\xi_0} \left(x \right) \partial_i \partial_j \ell_x \left(\xi_0 \right) dx \\ &= - E_{\xi_0} \left[\partial_i \partial_j \ell_x \left(\xi_0 \right) \right] = g_{ij} \left(\xi_0 \right) \end{aligned}$$

by Theorem 1.2.1. □

Corollary 1.3.1. *[4] The density p_{ξ_0} is a minimum point for the functional $p_\xi \to D_{KL} \left(p_{\xi_0} || p_\xi \right)$.*

Proof. By Proposition 1.2.2, the Fisher information matrix g is positive definite, and by Proposition 1.3.8, the point p_{ξ_0} is a minimum. □

Definition 1.3.5. *Suppose p and q are two points on a statistical manifold $\mathcal{S}$. The Fisher distance, $d \left(p, q \right)$, represents the information distance between densities p and q. It is defined as the length of the shortest curve on $\mathcal{S}$ between p and q, i.e. the length of the geodesic curve joining p and q.*

Proposition 1.3.9. *[4] The Kullback-Leibler relative entropy and the Fisher distance are related as*

$$D_{KL}\left(p||q\right) = \frac{1}{2}d^2\left(p,q\right) + o\left(d^2\left(p,q\right)\right). \tag{1.39}$$

Proof. Consider a geodesic $\gamma\left(s\right)$ on the statistical model $\mathcal{S}$ joining densities p and q, satisfying $\gamma\left(s\right) = \iota\left(\xi\left(s\right)\right) = p_{\xi(s)}$ with $\gamma\left(0\right) = p_{\xi_0} = p$ and $\gamma\left(t\right) = p_{\xi} = q$. Since the arc length along the geodesic is the Riemannian distance, we have $d\left(p,q\right) = t$.

Consider the function $\varphi\left(s\right) = f\left(\xi\left(s\right)\right)$, with $f\left(\xi\right) = D_{KL}\left(p_{\xi_0}||p_{\xi}\right)$. Write the second order expansion of φ about $t = 0$,

$$\varphi\left(t\right) = \varphi\left(0\right) + t\varphi'\left(0\right) + \frac{t^2}{2}\varphi''\left(0\right) + o\left(t^2\right).$$

By Propositions 1.3.7 and 1.3.8, we have

$$\varphi\left(0\right) = f\left(\xi\left(0\right)\right) = D_{KL}\left(p_{\xi_0}||p_{\xi_0}\right) = 0,$$

$$\varphi'\left(0\right) = \sum_i \frac{\partial f}{\partial \xi^i}\left(\xi_0\right)\dot{\xi}\left(0\right) = 0,$$

and

$$\begin{aligned}\varphi''\left(0\right) &= \sum_{i,j} \frac{\partial^2 f}{\partial \xi^i \partial \xi^j}\left(\xi_0\right)\dot{\xi}^i\left(0\right)\dot{\xi}^j\left(0\right) + \sum_i \frac{\partial f}{\partial \xi^i}\left(\xi_0\right)\ddot{\xi}\left(0\right)\\ &= \sum_{i,j} g_{ij}\left(\xi_0\right)\dot{\xi}^i\left(0\right)\dot{\xi}^j\left(0\right).\end{aligned}$$

Hence, we have

$$\begin{aligned} D_{KL}\left(p||q\right) = f\left(\xi\left(t\right)\right) = \varphi\left(t\right) &= \frac{t^2}{2}\sum_{i,j} g_{ij}\left(\xi_0\right)\dot{\xi}^i\left(0\right)\dot{\xi}^j\left(0\right) + o\left(t^2\right) \\ &= \frac{t^2}{2} g\left(\dot{\gamma}\left(0\right), \dot{\gamma}\left(0\right)\right) + o\left(t^2\right) \\ &= \frac{t^2}{2} + o\left(t^2\right) \\ &= \frac{1}{2} d^2\left(p,q\right) + o\left(d^2\left(p,q\right)\right), \end{aligned}$$

where $g\left(\dot{\gamma}\left(0\right), \dot{\gamma}\left(0\right)\right) = 1$ since geodesics parametrized by the arc length are unit speed curves. □

Corollary 1.3.2. *[4] Let the Kullback-Leibler quasimetric $\mathcal{D}\left(p,q\right)$ be defined as $\mathcal{D}\left(p,q\right) = D_{KL}\left(p||q\right) + D_{KL}\left(q||p\right)$. Then*

$$\mathcal{D}\left(p,q\right) = d^2\left(p,q\right) + o\left(d^2\left(p,q\right)\right). \tag{1.40}$$

Proposition 1.3.10. *[4] The diagonal part of the third mixed derivatives of the Kullback-Leibler relative entropy is the negative of the Christoffel symbol, i.e.*

$$-\partial_{\xi_0^k}\partial_{\xi^i}\partial_{\xi^j} D_{KL}\left(p_{\xi_0}||p_\xi\right)|_{\xi=\xi_0} = \Gamma^{(1)}_{ij,k}\left(\xi_0\right). \tag{1.41}$$

Proof. We have

$$\begin{aligned}\partial_i\partial_j D_{KL}\left(p_{\xi_0}||p_\xi\right) &= \partial_i\partial_j \int_{\mathcal{X}} p_{\xi_0}\left(x\right)\ln\frac{p_{\xi_0}\left(x\right)}{p_\xi\left(x\right)}dx \\ &= -\int_{\mathcal{X}} p_{\xi_0}\left(x\right)\partial_i\partial_j\ln p_\xi\left(x\right)dx.\end{aligned}$$

Differentiating in ξ_0^k yields

$$\begin{aligned}-\partial_{\xi_0^k}\partial_{\xi^i}\partial_{\xi^j} D_{KL}\left(p_{\xi_0}||p_\xi\right) &= \partial_{\xi_0^k}\int_{\mathcal{X}} p_{\xi_0}\left(x\right)\partial_i\partial_j\ln p_\xi\left(x\right)dx \\ &= \int_{\mathcal{X}} \partial_{\xi_0^k}\ln p_{\xi_0}\left(x\right)\partial_i\partial_j\ln p_\xi\left(x\right)\ p_{\xi_0}\left(x\right)dx.\end{aligned}$$

Considering the diagonal part, we have

$$\begin{aligned}-\partial_{\xi_0^k}\partial_{\xi^i}\partial_{\xi^j} D_{KL}\left(p_{\xi_0}||p_\xi\right)|_{\xi=\xi_0} &= E_{\xi_0}\left[\partial_i\partial_j\ell\left(\xi\right)\partial_k\ell\left(\xi\right)\right] \\ &= \Gamma^{(1)}_{ij,k}\left(\xi_0\right).\end{aligned}$$

□

1.3.3 Informational Energy

Definition 1.3.6. *Let $\mathcal{S}=\{p_\xi=p\left(x;\xi\right)|\,\xi=\left(\xi^1,\ldots,\xi^n\right)\in\mathbb{E}\}$ be a statistical model. The information energy on $\mathcal{S}$ is a function $I:\mathbb{E}\to\mathbb{R}$ defined by*

$$I\left(\xi\right)=\begin{cases}\int_{\mathcal{X}} p^2\left(x,\xi\right)dx, & \text{if } \mathcal{X} \text{ is continuous;}\\ \sum_{x\in\mathcal{X}} p^2\left(x,\xi\right), & \text{if } \mathcal{X} \text{ is discrete.}\end{cases}. \tag{1.42}$$

Proposition 1.3.11. *[4] The informational energy of a system with n elementary outcomes is bounded above by 1 and below by $1/n$, i.e.,*

$$1/n \leq I \leq 1.$$

The minimum of the informational energy is realized for the uniform distribution.

Proof. The information energy is bounded above by 1 since

$$1 = \left(\sum_{i=1}^{n} p_i\right)^2 = \sum_{i=1}^{n} p_i^2 + \sum_{i\neq j}^{n} p_i p_j \geq \sum_{i=1}^{n} p_i^2 = I.$$

Let the distribution $q = \{q_i\}$ with $q_i = p_i + s_i$, $i = 1, \ldots, n$, be the perturbed distribution of $p = \{p_i\}$. We have $s_n = -\sum_{i=1}^{n-1} s_i$ since $\sum_{i=1}^{n} s_i = 0$. Therefore,

$$\begin{aligned} I(q) &= \sum_{i=1}^{n-1} q_i^2 + q_n^2 \\ &= \sum_{i=1}^{n-1} (p_i + s_i)^2 + (p_n + s_n)^2 \\ &= \sum_{i=1}^{n-1} (p_i + s_i)^2 + \left(p_n - \sum_{i=1}^{n-1} s_i\right)^2 . \end{aligned}$$

We obtain that the uniform distribution is a critical point since

$$0 = \frac{\partial I}{\partial s_i}|_{s_1=\ldots=s_{n-1}=0} = \left[2(p_i + s_i) - 2\left(p_n - \sum_{k=1}^{n-1} s_k\right)\right]|_{s_1=\ldots=s_{n-1}=0}, \ \forall i = 1, \ldots, n-1,$$

$\Longleftrightarrow p_i = p_n = \dfrac{1}{n}$, $\forall i = 1, \ldots, n-1$.

The $(n-1)\times(n-1)$ dimensional Hessian, $(H_{ij}) = \left(\frac{\partial^2 I}{\partial s_i \partial s_j}|_{s_1=\ldots=s_{n-1}=0}\right)$, is

$$H_n = \begin{pmatrix} 4 & 2 & \ldots & 2 \\ 2 & 4 & \ldots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 2 & 2 & \ldots & 4 \end{pmatrix}$$

since

$$H_{ij} = \frac{\partial^2 I}{\partial s_i \partial s_j}|_{s_1=\ldots=s_{n-1}=0} = 2\delta_{ij} + 2.$$

The Hessian is non-degenerate since we have

$$\det H_n = \det \begin{pmatrix} 4 & 2 & \ldots & 2 \\ 2 & 4 & \ldots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 2 & 2 & \ldots & 4 \end{pmatrix} = 2^{n-1} \det \begin{pmatrix} 2 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 2 \end{pmatrix}$$

$$=2^{n-1} \det \begin{pmatrix} n & n & \ldots & n \\ 1 & 2 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 2 \end{pmatrix} = n2^{n-1} \det \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 2 \end{pmatrix}$$

$$= n2^{n-1} \det \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} = n2^{n-1},\ \forall n \geq 2.$$

Furthermore, since all of the leading principal minors of H_n are positive, H_n is positive-definite.

It follows that the uniform distribution realizes the minimum for the informational energy with $\min I = I(p) = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 = \frac{1}{n}$. □

Proposition 1.3.12. *[4] The informational energy functional, defined on continuous distributions on $[a, b]$, satisfies the inequality*

$$\frac{1}{b-a} \leq I(p).$$

The minimum of the informational energy functional is realized by the uniform distribution $p(x) = 1/(b-a)$.

Proof. Since, by Cauchy-Schwarz inequality,

$$\int_a^b |p(x)\, q(x)| dx \leq \left(\int_a^b p^2(x)\, dx\right)^{1/2} \left(\int_a^b q^2(x)\, dx\right)^{1/2},$$

we have, by letting $q(x) = 1$,

$$1 = \int_a^b p(x)\, dx \leq \left(\int_a^b p^2(x)\, dx\right)^{1/2} (b-a)^{1/2} \Longleftrightarrow I = \int_a^b p^2(x)\, dx \geq \frac{1}{b-a}.$$

The equality is reached when $p(x)$ and $q(x) = 1$ are proportional, i.e., $p(x)$ is constant. Thus, $p(x) = 1/(b-a)$. □

Chapter 2

The Informational Geometry of Exponential Family

2.1 Exponential family

Definition 2.1.1. *The exponential family is a set of probability distributions whose probability density function (or probability mass function, for the case of a discrete distribution) can be expressed in the form*

$$p\left(x;\xi\right) = e^{C(x)+\xi^{i}F_{i}(x)-\psi(\xi)}, \tag{2.1}$$

where $C\left(x\right),\ F_{1}\left(x\right),\ldots,F_{n}\left(x\right)$ are real-valued smooth functions on $\mathcal{X}\subset\mathbb{R}^{k}$ such that $\{1\}\cup\{F_{i}\left(x\right)\}$ are linearly independent, and $\psi\left(\xi\right)$ is the normalization function such that $\int_{\mathcal{X}}p\left(x,\xi\right)dx=1$. The parameter space $\mathbb{E}$ is chosen to be a non-empty set of ξ with $\psi\left(\xi\right)<\infty$.

The exponential model $\mathcal{S} = \{p(x;\xi)\}$, with $p(x;\xi)$ given by equation (2.1), is also called an *exponential family* and ξ^j are its *natural parameters*. The parameter space $\mathbb{E}$ is an open subset of $\mathbb{R}^n$ and its dimension is called the *order* of the exponential family.

The set $\mathcal{S}$ is a statistical model according to the following facts. First, consider the mapping $\iota : \mathbb{E} \to \mathcal{S}$. Assume $\iota(\xi) = \iota(\theta)$. Then $\ln p(x;\xi) = \ln p(x;\theta)$, and $(\xi^i - \theta^i) F_i(x) - (\psi(\xi) - \psi(\theta)) = 0$. By the linear independence of $\{1\} \cup \{F_i(x)\}$, we have $\xi^i = \theta^i$ for all i and hence the infectivity of ι. Second, by differentiating the logarithm of equation (2.1), we have

$$\partial_j \ell_x(\xi) = \partial_j \big(C(x) + \xi^i F_i(x) - \psi(\xi)\big) = F_j(x) - \partial_j \psi(\xi). \qquad (2.2)$$

Because $\{F_i(x)\}$ are linearly independent, so are $\{\partial_j \ell_x(\xi)\}$. This implies that ι is also regular.

2.1.1 Examples of Exponential family

Here are some common examples of exponential family which are described in terms of functions $C(x), \{F_i(x)\}$ and $\psi(\xi)$. Note that reparametrizations of the usual parameters space are often necessary to obtain the form of (2.1).

- **Bernoulli Distribution**

The sample space is $\mathcal{X} = \{0, 1\}$ and the parameter space is $\mathbb{E} = [0, 1]$. The Bernoulli distribution is given by

$$p\left(x;\theta\right) = \theta^{x}\left(1-\theta\right)^{1-x},\ x \in \mathcal{X},\ \theta \in \mathbb{E}. \tag{2.3}$$

This forms a one-dimensional statistical model. Since

$$\ln p\left(x;\theta\right) = x \ln \frac{\theta}{1-\theta} + \ln\left(1-\theta\right),$$

we choose

$$\xi = \ln \frac{\theta}{1-\theta},\ C\left(x\right) = 0,\ F_1\left(x\right) = x,\ \psi\left(\xi\right) = -\ln\left(1-\theta\right) = \ln\left(1+e^{\xi}\right). \tag{2.4}$$

- **Binomial Distribution**

The sample space is $\mathcal{X} = \{0, 1, \ldots, n\}$, where n is the number of trials and is fixed, and the parameter space is $\mathbb{E} = [0, 1]$. The binomial distribution is given by

$$p\left(x;\theta\right) = \left(\begin{array}{c} n \\ x \end{array}\right)\theta^{x}\left(1-\theta\right)^{n-x},\ x \in \mathcal{X},\ \theta \in \mathbb{E}. \tag{2.5}$$

Since

$$\ln p\left(x;\theta\right) = \ln \left(\begin{array}{c} n \\ x \end{array}\right) + x \ln \frac{\theta}{1-\theta} + n \ln\left(1-\theta\right), \tag{2.6}$$

we choose

$$\xi = \ln\frac{\theta}{1-\theta},\ C\left(x\right) = \ln\begin{pmatrix} n \\ \\ x \end{pmatrix},$$

$$F_1\left(x\right) = x,\ \psi\left(\xi\right) = -n\ln\left(1-\theta\right) = n\ln\left(1+e^{\xi}\right) \tag{2.7}$$

- **Multinomial Distribution**

The sample space is $\mathcal{X} = \{0, 1, \ldots, n\}$, where n is the number of trials and is supposed to be fixed, and the parameter space is

$$\mathbb{E} = \left\{\left(\theta^1, \ldots, \theta^{k-1}\right); \theta^i > 0 \text{ for } i = 1, \ldots, k, \textstyle\sum_{i=1}^{k} \theta^i = 1\right\}.$$

The multinomial distribution is given by

$$p\left(x;\theta\right) = \ln(\begin{matrix} n \\ \\ x_1 \ \ldots \ x_k \end{matrix})\left(\theta^1\right)^{x_1}\ldots\left(\theta^k\right)^{x_k} = \ln(\frac{n!}{\prod\limits_{i=1}^{k} x_i!})\prod_{i=1}^{k}\left(\theta^i\right)^{x_i}, \tag{2.8}$$

where $x_i \in \mathcal{X},\ \left(\theta^1, \ldots, \theta^{k-1}\right) \in \mathbb{E}$ and $\sum_{i=1}^{k} x_i = n$.

This is a (k-1)-dimensional statistical model. Since

$$\ln p\left(x;\theta\right) = \ln n! - \ln\sum_{i=1}^{k} x_i! + \sum_{i=1}^{k} x_i \ln\theta^i, \tag{2.9}$$

we may choose

$$\xi^j = \ln \frac{\theta^j}{\theta^k} = \ln \frac{\theta^j}{1 - \sum\limits_{i=1}^{k-1} \theta^i} \text{ for } j \in \{1, \ldots, k-1\},$$

$$C(x) = \ln n! - \ln \sum_{i=1}^{k} x_i!,\ F_j(x) = x_j \text{ for } j \in \{1, \ldots, k-1\},$$

$$\psi(\xi) = -n \ln \theta^k = -n \ln \left(1 - \frac{\sum_{i=1}^{k-1} e^{\xi^i}}{1 + \sum_{i=1}^{k-1} e^{\xi^i}}\right) = n \ln(1 + \sum_{i=1}^{k-1} e^{\xi^i}). \tag{2.10}$$

Note that a multinomial distribution reduces to a categorial distribution if there is only 1 trial. In that case, let the sample space be $\mathcal{X} = \left\{x^1, x^2, \ldots, x^k\right\}$, a set of n individually identified items, and the parameter space

$$\mathbb{E} = \left\{\left(\theta^1, \ldots, \theta^{k-1}\right); \theta^i > 0 \text{ for } i = 1, \ldots, k,\ \textstyle\sum_{i=1}^{k} \theta^i = 1\right\},$$

the probability mass function of a categorial distribution is given by

$$p(x; \theta) = \prod_{i=1}^{k} \theta^{i^{[x=x^i]}}, \tag{2.11}$$

where $[x = x^i]$ evaluates to 1 if $x = x^i$, 0 otherwise.

This is a (k-1)-dimensional statistical model. Since

$$\ln p(x; \theta) = \sum_{i=1}^{k-1} \left[x = x^i\right] \ln \theta^i + [x = x^n] \ln \left(1 - \sum_{i=1}^{k-1} \theta^i\right), \tag{2.12}$$

if we choose the same parametrization as in (2.10), we have

$$C\left(x\right)=0,\ F_{j}\left(x\right)=\left[x=x^{j}\right]\text{for}\, j\in\{1,\ldots,k-1\},$$

$$\psi\left(\xi\right)=\ln(1+\sum_{i=1}^{k-1}e^{\xi^{i}}), \tag{2.13}$$

where we used the fact $\sum_{i=1}^{k}\left[x=x^{i}\right]=1$.

- **Geometric Distribution**

The sample space is $\mathcal{X}=\{0,1,2,...\}$ and the parameter space is $\mathbb{E}=[0,1]$. The geometric distribution is given by

$$p\left(x;\theta\right)=\theta\left(1-\theta\right)^{x-1},\ x\in\mathcal{X},\ \theta\in\mathbb{E}. \tag{2.14}$$

This is a one-dimensional statistical model. Since

$$\ln p\left(x;\theta\right)=\ln\frac{\theta}{1-\theta}+x\ln\left(1-\theta\right),$$

we choose

$$\xi=\ln\left(1-\theta\right),\ C\left(x\right)=0,\ F_{1}\left(x\right)=x,\ \psi\left(\xi\right)=\ln\frac{1-\theta}{\theta}=\ln\frac{e^{\xi}}{1-e^{\xi}}, \tag{2.15}$$

where $\xi\in(-\infty,0)$.

- **Poisson Distribution**

The sample space is $\mathcal{X}=\{0,1,2,...\}$ and the parameter space is $\mathbb{E}=(0,\infty)$. The

Poisson distribution is given by

$$p\left(x;\lambda\right)=e^{-\lambda}\frac{\lambda^{x}}{x!}. \tag{2.16}$$

This forms a one-dimensional statistical model. Since

$$\ln p\left(x;\lambda\right)=-\lambda+x\ln\lambda-\ln x!,$$

we choose

$$\xi=\ln\lambda,\ C\left(x\right)=-\ln x!,\ F_1\left(x\right)=x,\ \psi\left(\xi\right)=\lambda=e^{\xi}. \tag{2.17}$$

- **Joint Poisson Distribution**

Consider m independent Poisson distributions with parameters λ_i, $i=1,\ldots,m$. The joint probability mass function is given by

$$p\left(x;\lambda\right)=\prod_{i=1}^{m}p_{\lambda_i}\left(x_i\right)=\prod_{i=1}^{m}e^{-\lambda_i}\frac{\lambda_i^{x_i}}{x_i!},$$

where $\lambda=\left(\lambda_1,\ldots,\lambda_m\right)\in\mathbb{E}=\left(\mathbb{R}^+\right)^m$, and $x=\left(x_1,\ldots,x_m\right)\in\mathcal{X}=\left(\mathbb{N}\cup\{0\}\right)^m$.

This forms an m-dimensional statistical model. Since

$$\ln p\left(x;\lambda\right)=-\sum_{i=1}^{m}\left(\lambda_i+x_i\ln\lambda_i-\ln x_i!\right),$$

we choose

$$\xi_i=\ln\lambda_i,\ C\left(x\right)=-\sum_{i=1}^{m}\ln x_i!,\ F_i\left(x\right)=x_i,\ \psi\left(\xi\right)=\sum_{i=1}^{m}\lambda_i=\sum_{i=1}^{m}e^{\xi_i}. \tag{2.18}$$

- **Normal Distribution**

Let $\mathcal{X} = \mathbb{R}$ and $\mathbb{E} = \mathbb{R} \times (0, \infty)$. The normal distribution is defined by the formula

$$p\,(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x \in \mathcal{X},\ (\mu, \sigma) \in \mathbb{E}. \tag{2.19}$$

Since

$$\ln p\,(x; \mu, \sigma) = -\ln \sigma - \ln \sqrt{2\pi} - \frac{(x-\mu)^2}{2\sigma^2}, \tag{2.20}$$

we choose

$$\xi^1 = \frac{\mu}{\sigma^2},\ \xi^2 = -\frac{1}{2\sigma^2},$$

$$\begin{aligned} &C\,(x) = 0,\ F_1\,(x) = x,\ F_2\,(x) = x^2, \\ &\psi\,(\xi) = \frac{\mu^2}{2\sigma^2} + \ln\left(\sigma\sqrt{2\pi}\right) = \frac{-\left(\xi^1\right)^2}{4\xi^2} + \frac{1}{2}\ln\left(\frac{-\pi}{\xi^2}\right), \end{aligned} \tag{2.21}$$

where $(\xi^1, \xi^2) \in \mathbb{R} \times (-\infty, 0)$.

- **Multivariate Normal Distribution**

The multivariate normal distribution with a mean vector μ and a covariance matrix A is defined by

$$p\left(x; \mu, A^{-1}\right) = \frac{1}{(2\pi)^{k/2}\,(\det A)^{1/2}} e^{-\frac{1}{2}(x-\mu)^t A^{-1}(x-\mu)},\ \mathcal{X} = \mathbb{R}^k \tag{2.22}$$

with the parameter space

$$\mathbb{E} = \left\{\left(\mu, A^{-1}\right); \mu \in \mathbb{R}^k, \text{positive definite}\, A^{-1} \in \mathbb{R}^{k\times k}\right\}.$$

Since A^{-1} is symmetric, there are $k+\dfrac{k\,(k+1)}{2}$ independent entries in each element of $\mathbb{E}$.

The log-likelihood function for this statistical model is

$$\begin{aligned}\ell_x\left(\mu, A^{-1}\right) &= \ln p\left(x;\mu, A^{-1}\right)\\ &= -\frac{1}{2}\left(x-\mu\right)^t A^{-1}\left(x-\mu\right)+\frac{1}{2}\ln\left(\det A^{-1}\right)-\frac{k}{2}\ln\left(2\pi\right). \end{aligned} \tag{2.23}$$

A multivariate normal distribution is an exponential family, which can be verified by choosing

$$\xi^1 = A^{-1}\mu,\ \xi^2 = -\frac{1}{2}A^{-1}, \tag{2.24}$$

with

$$\mu = -\frac{1}{2}\left(\xi^2\right)^{-1}\xi^1,\ A = -\frac{1}{2}\left(\xi^2\right)^{-1},$$

and

$$\begin{aligned} C\left(x\right) &= -\frac{k}{2}\ln\left(2\pi\right),\ F_1\left(x\right) = x,\ F_2\left(x\right) = xx^t,\\ \psi\left(\xi\right) &= -\frac{1}{4}\left(\xi^1\right)^t\left(\xi^2\right)^{-1}\xi^1-\frac{1}{2}\ln\det(-2\xi^2). \end{aligned} \tag{2.25}$$

- **Lognormal Distribution**

Let $\mathcal{X}=\mathbb{R}^+$ and $\mathbb{E}=\mathbb{R}\times\mathbb{R}^+$. The lognormal distribution is defined by the formula

$$p\left(x;\mu,\sigma\right) = \frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}},\ x\in\mathcal{X},\ (\mu,\sigma)\in\mathbb{E}. \tag{2.26}$$

Since

$$\begin{aligned}\ln p\left(x;\mu,\sigma\right) &= -\ln x - \ln\left(\sigma\sqrt{2\pi}\right) - \frac{\left(\ln x-\mu\right)^2}{2\sigma^2},\\ &= -\ln x - \ln\left(\sigma\sqrt{2\pi}\right) + \frac{\mu}{\sigma^2}\ln x - \frac{1}{2\sigma^2}\left(\ln x\right)^2 - \frac{\mu}{2\sigma^2} \qquad (2.27)\end{aligned}$$

we choose

$$\xi^1 = \frac{\mu}{\sigma^2},\ \xi^2 = -\frac{1}{2\sigma^2},$$

$$\begin{aligned}&C\left(x\right) = -\ln x,\ F_1\left(x\right) = \ln x,\ F_2\left(x\right) = \left(\ln x\right)^2,\\ &\psi\left(\xi\right) = \frac{\mu^2}{2\sigma^2} + \ln\left(\sigma\sqrt{2\pi}\right) = \frac{-\left(\xi^1\right)^2}{4\xi^2} + \frac{1}{2}\ln\left(\frac{-\pi}{\xi^2}\right), \qquad (2.28)\end{aligned}$$

where $(\xi^1, \xi^2) \in \mathbb{R} \times (-\infty, 0)$.

- **Exponential Distribution**

Let $\mathcal{X} = [0,\infty)$ and consider the one-dimensional parameter space $\mathbb{E} = (0,\infty)$, which is an open interval in $\mathbb{R}$. The exponential distribution with parameter ξ is given by the formula

$$p\left(x;\xi\right) = \xi e^{-\xi x}. \qquad (2.29)$$

This is corresponding to the case, in (2.1), choosing $n = 1$ and

$$C\left(x\right) = 0,\ F_1\left(x\right) = -x,\ \xi^1 = \xi,\ \psi\left(\xi\right) = -\ln\xi. \qquad (2.30)$$

- **Gamma Distribution**

The sample space is $\mathcal{X} = [0, \infty)$ and the parameter space is $\mathbb{E} = (0, \infty) \times (0, \infty)$. The Gamma distribution is given by

$$p(x; \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta},\ x \in \mathcal{X},\ (\alpha, \beta) \in \mathbb{E}, \tag{2.31}$$

where $\Gamma(\alpha)$ is defined by the *Gamma function*

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt,\ \alpha > 0.$$

Since

$$\ln p(x; \lambda) = -\ln(\beta^{\alpha}\Gamma(\alpha)) - \ln x + \alpha \ln x - \frac{x}{\beta},$$

we choose

$$\begin{aligned} \xi^1 = \alpha,\ \xi^2 = \frac{-1}{\beta},\ C(x) = -\ln x,\ F_1(x) = \ln x,\ F_2(x) = x, \\ \psi(\xi) = \ln(\beta^{\alpha}\Gamma(\alpha)) = \ln\left(\left(\frac{-1}{\xi^2}\right)^{\xi^1} \Gamma(\xi^1)\right). \end{aligned} \tag{2.32}$$

Since the **Chi-squared distribution** is a special case of Gamma distribution, *i.e.* $\chi^2(k) = \Gamma(k/2, 2)$, so it is among the exponential family.

- **Beta Distribution**

The sample space is $\mathcal{X} = [0, 1]$ and the parameters $(a, b) \in \mathbb{E} = (0, \infty) \times (0, \infty)$. The Beta distribution is given by

$$p(x; a, b) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}, \tag{2.33}$$

where $B(a,b)$ is defined by the *Beta function*

$$B(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}\,dt,\ a,b>0.$$

Note that

$$B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Since

$$\ln p(x;\lambda)=-\ln(B(a,b))-\ln(x(1-x))+a\ln x+b\ln(1-x),$$

we choose

$$\begin{aligned}&\xi^1=a,\ \xi^2=b,\ C(x)=-\ln(x(1-x)),\ F_1(x)=\ln x,\ F_2(x)=\ln(1-x),\\ &\psi(\xi)=\ln(B(a,b)).\end{aligned}\tag{2.34}$$

2.1.2 Properties of Exponential Family

The following results show that, for distributions in exponential family, the expectations of F_j with respect to p_ξ are the corresponding derivatives of $\psi(\xi)$ with respect to ξ^j; furthermore, their covariance matrix is exactly the Fisher information matrix.

Proposition 2.1.1. *Suppose the sample space $\mathcal{X}$ of an exponential family is bounded and $E_\xi[F_i]<\infty$ for all $F_i(x)$. Then*

$$E_\xi[F_j]=\partial_j\psi(\xi),\ 1\le j\le n,\tag{2.35}$$

$$g_{ij}(\xi) = E_\xi\left[\partial_i \ell_x(\xi)\, \partial_j \ell_x(\xi)\right] = \mathrm{Cov}\left(F_i, F_j\right), \tag{2.36}$$

In particular,

$$\mathrm{Var}\left(F_j\right) = E_\xi\left[\partial_j^2 \ell_x(\xi)\right]. \tag{2.37}$$

Proof. Differentiating $\int_{\mathcal{X}} p(x,\xi)\, dx = 1$ and using the notation $\partial_j = \frac{\partial}{\partial \xi^j}$, we have

$$\int_{\mathcal{X}} \partial_j p(x,\xi)\, dx = 0 \Leftrightarrow$$

$$\int_{\mathcal{X}} p(x,\xi)\, \partial_j \ell(x,\xi)\, dx = 0 \Leftrightarrow$$

by (2.2),

$$\int_{\mathcal{X}} p(x,\xi)\left[F_j(x) - \partial_j \psi(\xi)\right] dx = 0 \Leftrightarrow$$

$$\int_{\mathcal{X}} p(x,\xi)\, F_j(x) = \partial_j \psi(\xi) \int_{\mathcal{X}} p(x,\xi)\, dx \Leftrightarrow$$

$$E_\xi\left[F_j\right] = \partial_j \psi(\xi).$$

By equations (2.2) and (2.35), we have $\partial_j \ell_x(\xi) = F_j(x) - E_\xi\left[F_j\right]$, which leads to

$$\begin{aligned} g_{ij}(\xi) = E_\xi\left[\partial_i \ell_x(\xi)\, \partial_j \ell_x(\xi)\right] &= E_\xi\left[\left(F_i(x) - E_\xi\left[F_i\right]\right)\left(F_j(x) - E_\xi\left[F_j\right]\right)\right] \\ &= \mathrm{Cov}\left(F_i, F_j\right). \end{aligned}$$

□

This is another way to verify that the Fisher information matrix on statistical models in exponential family is symmetric and positive definite.

There are several other important properties of exponential family as described in the following results.

Proposition 2.1.2.

$$\partial_i \partial_j \ell_x(\xi) = -\partial_i \partial_j \psi(\xi). \tag{2.38}$$

Proof. This is the immediate result of differentiating (2.2) with respect to ξ^j. □

Theorem 2.1.1. *The Fisher information matrix for an exponential family is given by*

$$g_{ij} = \partial_i \partial_j \psi(\xi), \tag{2.39}$$

where $\psi(\xi)$ is the normalization function.

Proof. By equations (1.8) and (2.38), we obtain

$$g_{ij} = -E_\xi \left[\partial_i \partial_j \ell_x(\xi)\right] = -E_\xi \left[-\partial_i \partial_j \psi(\xi)\right] = \partial_i \partial_j \psi(\xi).$$

□

Proposition 2.1.3. *In an exponential family model, the normalization function $\psi(\xi)$ is convex.*

Proof. By differentiating equation (2.35) with respect to ξ^k, we have

$$\begin{aligned}
\partial_k \partial_j \psi(\xi) &= \partial_k \int_{\mathcal{X}} e^{C(x)+\xi^i F_i(x)-\psi(\xi)} F_j(x)\, dx \\
&= \int_{\mathcal{X}} \partial_k \left(e^{C(x)+\xi^i F_i(x)-\psi(\xi)} F_j(x) \right) dx \\
&= \int_{\mathcal{X}} e^{C(x)+\xi^i F_i(x)-\psi(\xi)} F_j(x) \left(F_k(x) - \partial_k \psi(\xi) \right) dx \\
&= E\left[F_j(x) F_k(x)\right] - E\left[F_j(x)\right] \partial_k \psi(\xi) \\
&= E\left[F_j(x) F_k(x)\right] - E\left[F_j(x)\right] E\left[F_k(x)\right] \\
&= \operatorname{Cov}\left[F_j(x), F_k(x)\right] \\
&= \begin{cases} \operatorname{Var}(F_j), & \text{if } j = k; \\ 0, & \text{if } j \neq k. \end{cases}
\end{aligned}$$

This implies that the Hessian matrix of second partial derivatives of $\psi(\xi)$ is positive definite. □

Proposition 2.1.4. *The Christoffel symbols of first kind for an exponential family can be expressed as*

$$\Gamma_{ij,k} = \frac{1}{2} \partial_i \partial_j \partial_k \psi(\xi), \tag{2.40}$$

where $\psi(\xi)$ is the normalization function.

Proof. By equations (2.2), (2.38) and (2.35), we have

$$\begin{aligned}
E_\xi\left[(\partial_i \partial_j \ell)\, \partial_k \ell\right] &= E_\xi\left[(-\partial_i \partial_j \psi)\left(F_k(x) - \partial_k \psi\right)\right] \\
&= \partial_i \partial_j \psi\left(-E_\xi\left[F_k(x)\right] + \partial_k \psi\right) = 0.
\end{aligned} \tag{2.41}$$

Similarly, we have

$$E_\xi\left[\left(\partial_j\partial_k\ell\right)\partial_i\ell\right]=E_\xi\left[\left(\partial_k\partial_i\ell\right)\partial_j\ell\right]=0.$$

Substituting into equation (i) in Proposition 1.2.4, we obtain

$$\begin{aligned}\Gamma_{ij,k}&=\frac{1}{2}\left(E_\xi\left[\left(\partial_i\partial_j\ell\right)\partial_k\ell\right]-E_\xi\left[\left(\partial_j\partial_k\ell\right)\partial_i\ell\right]-E_\xi\left[\left(\partial_k\partial_i\ell\right)\partial_j\ell\right]-E_\xi\left[\partial_i\partial_j\partial_k\ell\right]\right)\\&=-\frac{1}{2}E_\xi\left[\partial_i\left(\partial_j\partial_k\ell\right)\right]=\frac{1}{2}E_\xi\left[\partial_i\left(\partial_j\partial_k\psi\right)\right]=\frac{1}{2}\partial_i\partial_j\partial_k\psi\left(\xi\right).\end{aligned}$$

□

Proposition 2.1.5. *The Christoffel symbols of first kind for an exponential family is covariant under reparametrization.*

Proof. By equation (ii) in Proposition (1.2.4) and equation (2.41), we have

$$\Gamma_{ij,k}\left(\xi\right)=\frac{1}{2}E_\xi\left[\partial_i\ell\partial_j\ell\partial_k\ell\right]=\frac{1}{2}T_{ij,k}\left(\xi\right).$$

From Proposition 1.2.10 we know that the skewness tensor is covariant under reparametrization, and so is each Christoffel symbol. □

2.2 The Fisher Metric

In this section, we will work on some common exponential families to derive their Fisher information matrices. Both the method using the definition of Fisher information matrix, equation (1.5), and that based on equation (2.39) will be applied and, when it is necessary, be compared.

- **Binomial Distribution**

Under the parametrization $\xi = \theta$, the log-likelihood function for a binomial distribution model is

$$\ln p\left(x;\xi\right) = \ln \begin{pmatrix} n \\ x \end{pmatrix} + x \ln \xi + (n-x) \ln\left(1-\xi\right). \tag{2.42}$$

Hence, the Fisher information is given by

$$\begin{aligned} g_{11}\left(\xi\right) &= -E_{\xi}\left[\partial_{\xi}^{2} \ell_{x}\left(\xi\right)\right] = -E_{\xi}\left[-\frac{x}{\xi^{2}} - \frac{n-x}{\left(1-\xi\right)^{2}}\right] \\ &= \frac{E_{\xi}\left[x\right]}{\xi^{2}} + \frac{n-E_{\xi}\left[x\right]}{\left(1-\xi\right)^{2}} = \frac{n\xi}{\xi^{2}} + \frac{n-n\xi}{\left(1-\xi\right)^{2}} = \frac{n}{\xi\left(1-\xi\right)}. \end{aligned} \tag{2.43}$$

Under the parametrization,

$$\eta = \ln\frac{\theta}{1-\theta},\ C\left(x\right) = \ln(\begin{matrix} n \\ x \end{matrix}),\ F_{1}\left(x\right) = x,\ \psi\left(\eta\right) = -n\ln\left(1-\theta\right) = n\ln\left(1+e^{\eta}\right),$$

it is easy to obtain the Fisher information,

$$\tilde{g}_{11}\left(\eta\right) = \partial_{\eta}^{2}\psi\left(\eta\right) = \partial_{\eta}\left[\frac{ne^{\eta}}{1+e^{\eta}}\right] = \frac{ne^{\eta}}{\left(1+e^{\eta}\right)^{2}}. \tag{2.44}$$

By Theorem 2.1.1, we can verify that Fisher matrix is covariant under reparametrizations of the parameter space,

$$\begin{aligned} g_{11}(\xi) = \tilde{g}_{11}(\eta) J_{11} J_{11} &= \frac{ne^{\eta}}{(1+e^{\eta})^2} \left(\frac{1}{\theta(1-\theta)} \right)^2 \\ &= n\theta(1-\theta) \left(\frac{1}{\theta(1-\theta)} \right)^2 = \frac{n}{\xi(1-\xi)}, \end{aligned}$$

where $J_{11} = \frac{\partial}{\partial \xi}\eta = \frac{\partial}{\partial \theta} \ln \frac{\theta}{1-\theta} = \frac{1}{\theta(1-\theta)}$.

- **Multinomial Distribution**

By equation (2.9), we have

$$\ell_x(\theta) = \ln p(x;\theta) = \ln n! - \ln \sum_{i=1}^{k} x_i! + \sum_{i=1}^{k} x_i \ln \theta^i,$$

$$\begin{aligned} \partial_{\theta_i} \ell_x(\theta) = \partial_i \ell_x(\theta) &= \partial_i \left[\ln n! - \ln \sum_{i=1}^{k} x_i! + \sum_{i=1}^{k-1} x_i \ln \theta^i + x_k \ln \left(1 - \sum_{i=1}^{k-1} \theta^i \right) \right] \\ &= \frac{x_i}{\theta^i} - \frac{x_k}{1 - \sum_{i=1}^{k-1} \theta^i}, \end{aligned}$$

$$\partial_j \partial_i \ell_x(\theta) = -\frac{x_i \delta_{ij}}{(\theta^i)^2} - \frac{x_k}{\left(1 - \sum_{i=1}^{k-1} \theta^i \right)^2}.$$

Hence, the Fisher information under natural parametrization is given by

$$\begin{aligned} g_{rs}(\theta) &= -E_\theta\left[\partial_s\partial_r\ell_x(\theta)\right] = E_\theta\left[\frac{x_r\delta_{rs}}{(\theta^r)^2} + \frac{x_k}{\left(1-\sum_{i=1}^{k-1}\theta^i\right)^2}\right] \\ &= \frac{n\theta^r\delta_{rs}}{(\theta^r)^2} + \frac{n\theta^k}{\left(1-\sum_{i=1}^{k-1}\theta^i\right)^2} \\ &= n\left(\frac{\delta_{rs}}{\theta^r} + \frac{1}{1-\sum_{i=1}^{k-1}\theta^i}\right),\ r,s\in\{1,\ldots,k-1\}. \end{aligned} \tag{2.45}$$

A multinomial distribution reduces to a categorial distribution when there is only 1 trial, i.e. $n = 1$. For a categorial distribution with the probability mass function given by equation (2.11), we can similarly obtain the Fisher information matrix as

$$g_{rs}(\theta) = \frac{\delta_{rs}}{\theta^r} + \frac{1}{1-\sum_{i=1}^{k-1}\theta^i},\ r,s\in\{1,\ldots,k-1\},\ r,s\in\{1,\ldots,k-1\}. \tag{2.46}$$

Under the reparametrization,

$$\xi^j = \ln\frac{\theta^j}{\theta^k} = \ln\frac{\theta^j}{1-\sum_{i=1}^{k-1}\theta^i}\ \text{for}\ j\in\{1,\ldots,k-1\},$$

the Jacobian matrix $J\left(\xi,\theta\right)$ is given by $\left(\frac{\delta_{rs}}{\theta^r}+\frac{1}{1-\sum\limits_{i=1}^{k-1}\theta^i}\right)_{r,s=1}^{k-1}$.

By equation (2.10) for a multinomial distribution, we have

$$\partial_r\psi\left(\xi\right)=n\partial_r\ln(1+\sum_{i=1}^{k-1}e^{\xi^i})=n\frac{e^{\xi^r}}{1+\sum_{i=1}^{k-1}e^{\xi^i}}.$$

So, the Fisher information matrix is given by

$$\begin{aligned}\tilde{g}_{rs}\left(\xi\right)&=\partial_s\partial_r\psi\left(\xi\right)=n\partial_s(\frac{e^{\xi^r}}{1+\sum_{i=1}^{k-1}e^{\xi^i}})\\&=\frac{-ne^{\xi^r}e^{\xi^s}}{\left(1+\sum_{i=1}^{k-1}e^{\xi^i}\right)^2}+\frac{n\delta_{rs}e^{\xi^r}}{1+\sum_{i=1}^{k-1}e^{\xi^i}},\ \forall r,s\in\{1,\ldots,k-1\}.\end{aligned}$$

Similarly, for a categorial distribution, the Fisher information matrix is given by

$$\tilde{g}_{rs}\left(\xi\right)=\frac{-e^{\xi^r}e^{\xi^s}}{\left(1+\sum_{i=1}^{k-1}e^{\xi^i}\right)^2}+\frac{\delta_{rs}e^{\xi^r}}{1+\sum_{i=1}^{k-1}e^{\xi^i}},\ \forall r,s\in\{1,\ldots,k-1\}.\tag{2.47}$$

By Theorem 2.1.1, we confirm the same result for the parameters space $\mathbb{E}=$

$\{\theta_i, i \in \{1, \ldots k-1\}\}$,

$$\begin{aligned}
g_{rs}(\theta) &= \sum_{p,q=1}^{k-1} \tilde{g}_{pq}(\xi)\,|_{\xi=\xi(\theta)} J_{pr} J_{qs} \\
&= n \sum_{p,q=1}^{k-1} \left(\frac{-e^{\xi^p} e^{\xi^q}}{\left(1 + \sum_{i=1}^{k-1} e^{\xi^i}\right)^2} + \frac{n\delta_{pq} e^{\xi^p}}{1 + \sum_{i=1}^{k-1} e^{\xi^i}} \right) \left(\frac{\delta_{pr}}{\theta^p} + \frac{1}{\theta^k} \right) \left(\frac{\delta_{qs}}{\theta^q} + \frac{1}{\theta^k} \right) \\
&= n \sum_{p,q=1}^{k-1} \left(-\theta^p \theta^q + \delta_{pq} \theta^p\right) \left(\frac{\delta_{pr}}{\theta^p} + \frac{1}{\theta^k} \right) \left(\frac{\delta_{qs}}{\theta^q} + \frac{1}{\theta^k} \right) \\
&= n \sum_{q=1}^{k-1} \left[\sum_{p=1}^{k-1} \left(-\delta_{pr}\theta^q - \frac{\theta^p\theta^q}{\theta^k} + \delta_{pq}\delta_{pr} + \delta_{pq}\frac{\theta^q}{\theta^k} \right) \right] \left(\frac{\delta_{qs}}{\theta^q} + \frac{1}{\theta^k} \right) \\
&= n \sum_{q=1}^{k-1} \left[-\theta^q - \frac{\sum_{p=1}^{k-1} \theta^p\theta^q}{\theta^k} + \delta_{qr} + \frac{\theta^q}{\theta^k} \right] \left(\frac{\delta_{qs}}{\theta^q} + \frac{1}{\theta^k} \right) \\
&= n \sum_{q=1}^{k-1} \delta_{qr} \left(\frac{\delta_{qs}}{\theta^q} + \frac{1}{\theta^k} \right) \\
&= n \left(\frac{\delta_{rs}}{\theta^r} + \frac{1}{\theta^k} \right), \; r, s \in \{1, \ldots, k-1\}.
\end{aligned}$$

A third way to compute the Fisher information matrix is using the proposition

1.2.1,

$$\begin{aligned}
g_{rs}(\theta) &= 4\sum_{\mathcal{X}} \partial_r \sqrt{p_\theta(x)} \partial_s \sqrt{p_\theta(x)} \\
&= 4n\sum_{j=1}^{k} \partial_r \sqrt{p_\theta(x_j)} \partial_s \sqrt{p_\theta(x_j)} \\
&= 4n\left(\sum_{j=1}^{k-1} \partial_r \sqrt{\theta^j} \partial_s \sqrt{\theta^j} + \partial_r \sqrt{\theta^k} \partial_s \sqrt{\theta^k}\right) \\
&= 4n\left(\sum_{j=1}^{k-1} \frac{\delta_{rj}}{2\sqrt{\theta^j}} \frac{\delta_{sj}}{2\sqrt{\theta^j}} + \frac{-1}{2\sqrt{1-\sum_{j=1}^{k-1}\theta^j}} \frac{-1}{2\sqrt{1-\sum_{j=1}^{k-1}\theta^j}}\right) \\
&= n\left(\frac{\delta_{rs}}{\theta^r} + \frac{1}{1-\sum_{j=1}^{k-1}\theta^j}\right). \qquad (2.48)
\end{aligned}$$

where we used the fact that, for each trial among n trials, we have $p_\theta(x_j) = \theta^j$ for $j = 1, \ldots, k$. When $n = 1$, the result is the Fisher information matrix of a categorial distribution.

- **Poisson Distribution**

By equation (2.17), we have

$$\partial_\lambda \ell_x(\xi) = -1 + \frac{x}{\lambda},\ \partial_\lambda^2 \ell_x(\xi) = -\frac{x}{\lambda^2}.$$

For the natural parameter λ, the Fisher metric is given by

$$g = g_{11}\left(\lambda\right) = -E_{\lambda}\left[\partial_{\lambda}\ell_{x}\left(\xi\right)\right] = -E_{\lambda}\left[-\frac{x}{\lambda^{2}}\right] = \frac{E_{\lambda}\left[x\right]}{\lambda^{2}} = \frac{1}{\lambda}. \tag{2.49}$$

For parameters $\eta = \ln\lambda$, by equations (2.17) and (2.39), the Fisher metric is given by

$$\tilde{g} = \tilde{g}_{11}\left(\eta\right) = \partial_{1}^{2}\psi\left(\eta\right) = \partial_{1}^{2}\left(e^{\eta}\right) = e^{\eta}. \tag{2.50}$$

By Theorem 2.1.1, we can verify that Fisher matrix is covariant under reparametrizations of the parameter space,

$$g_{11}\left(\lambda\right) = \tilde{g}_{11}\left(\eta\right)J_{11}J_{11} = e^{\eta}\left(\frac{1}{\lambda}\right)^{2} = \lambda\left(\frac{1}{\lambda}\right)^{2} = \frac{1}{\lambda},$$

where $J_{11} = \frac{\partial}{\partial\lambda}\eta = \frac{\partial}{\partial\lambda}\ln\lambda = \frac{1}{\lambda}$.

- **Joint Poisson Distribution**

By applying equation (2.35) to the joint distribution of m independent Poisson distributions (see equation (2.18)), we have

$$E_{\eta}\left[x_{j}\right] = E_{\eta}\left[F_{j}\right] = \partial_{j}\psi\left(\eta\right) = e^{\eta_{j}},\ \forall j \in \{1,\ldots,m\}. \tag{2.51}$$

According to equation (2.18), we have

$$\partial_{j}\ell_{x}\left(\lambda\right) = -1 + \frac{x_{j}}{\lambda_{j}},\ \partial_{k}\partial_{j}\ell_{x}\left(\lambda\right) = -\frac{x_{j}}{\lambda_{j}^{2}}\delta_{jk}.$$

Then, using the natural parameters, the Fisher information for the joint distribution of m independent Poisson distributions is obtained as

$$\begin{aligned} g_{jk}\left(\lambda\right) &= -E_{\lambda}\left[\partial_k\partial_j\ell_x\left(\lambda\right)\right] = -E_{\lambda}\left[-\frac{x_j}{\lambda_j^2}\delta_{jk}\right] \\ &= \frac{E_{\eta}\left[x_j\right]}{\lambda_j^2}\delta_{jk} = \frac{1}{\lambda_j^2}\delta_{jk}\lambda_j = \frac{1}{\lambda_j}\delta_{jk}, \end{aligned} \tag{2.52}$$

where we used the fact expressed in equation (2.51).

By equation (2.18), we have

$$\partial_j\psi\left(\eta\right) = e^{\eta_j},\ \partial_k\partial_j\psi\left(\eta\right) = e^{\eta_j}\delta_{jk}.$$

For parameters $\eta = \{\ln\lambda_i\}_{i=1}^{m}$, by equations (2.18) and (2.39), the Fisher metric is given by

$$\tilde{g}_{jk}\left(\eta\right) = \partial_k\partial_j\psi\left(\eta\right) = e^{\eta_j}\delta_{jk}. \tag{2.53}$$

Since the Jacobian of (η, λ) is

$$J\left(\eta,\lambda\right) = \begin{pmatrix} \frac{1}{\lambda_1} & & & 0 \\ & \frac{1}{\lambda_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\lambda_m} \end{pmatrix},$$

we can verify the covariant relation between the Fisher information under reparametrizations of the parameters space,

$$
\begin{aligned}
g_{jk}(\lambda) &= \sum_{l,m} \tilde{g}_{lm}(\eta) J_{lj} J_{mk} = \tilde{g}_{jk} \frac{1}{\lambda_j} \frac{1}{\lambda_k} \\
&= e^{\eta_k} \delta_{jk} \frac{1}{\lambda_j} \frac{1}{\lambda_k} = \frac{1}{\lambda_j} \delta_{jk}.
\end{aligned}
$$

- **Normal Distribution**

By Equation (2.20), we have

$$
\partial_\sigma \ell_x(\xi) = -\frac{1}{\sigma} - \frac{(x-\mu)^2}{2}\left(-\frac{2}{\sigma^3}\right) = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3},\ \partial_\sigma^2 \ell_x(\xi) = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4},
$$

$$
\partial_\mu \ell_x(\xi) = \frac{x-\mu}{\sigma^2},\ \partial_\mu^2 \ell_x(\xi) = -\frac{1}{\sigma^2},\ \partial_\sigma \partial_\mu \ell_x(\xi) = \partial_\mu \partial_\sigma \ell_x(\xi) = -\frac{2(x-\mu)}{\sigma^3}. \tag{2.54}
$$

For parameters $(\xi^1, \xi^2) = (\mu, \sigma)$, the Fisher-Riemann metric components are given by

$$
g_{11}(\xi) = -E_\xi\left[\partial_\mu^2 \ell_x(\xi)\right] = -E_\xi\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2},
$$

$$
g_{12}(\xi) = g_{21}(\xi) = -E_\xi\left[\partial_\mu \partial_\sigma \ell_x(\xi)\right] = -E_\xi\left[-\frac{2(x-\mu)}{\sigma^3}\right] = \frac{2(E_\xi[x]-\mu)}{\sigma^3} = 0,
$$

$$
\begin{aligned}
g_{22}(\xi) &= -E_\xi\left[\partial_\sigma^2 \ell_x(\xi)\right] = -E_\xi\left[\frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}\right] = -\frac{1}{\sigma^2} + \frac{3}{\sigma^4} E_\xi\left[(x-\mu)^2\right] \\
&= -\frac{1}{\sigma^2} + \frac{3}{\sigma^4}\mathrm{Var}[x-\mu] = -\frac{1}{\sigma^2} + \frac{3}{\sigma^4}\mathrm{Var}[x] = \frac{2}{\sigma^2}.
\end{aligned} \tag{2.55}
$$

For parameters $(\eta^1, \eta^2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$, by equations (2.21) and (2.39), the Fisher-

Riemann metric components are given by

$$\tilde{g}_{11}(\eta) = \partial_1^2 \psi(\eta) = \partial_1^2 \left(\frac{-\left(\eta^1\right)^2}{4\eta^2} + \frac{1}{2} \ln\left(\frac{-\pi}{\eta^2} \right) \right) = -\frac{1}{2\eta^2},$$

$$\tilde{g}_{12}(\eta) = \tilde{g}_{21}(\eta) = \partial_2 \partial_1 \psi(\eta) = \partial_2 \left(\frac{-\eta^1}{2\eta^2} \right) = \frac{\eta^1}{2\left(\eta^2\right)^2},$$

$$\tilde{g}_{22}(\eta) = \partial_2^2 \psi(\eta) = \partial_2^2 \left(\frac{-\left(\eta^1\right)^2}{4\eta^2} + \frac{1}{2} \ln\left(\frac{-\pi}{\eta^2} \right) \right) = -\frac{\left(\eta^1\right)^2}{2\left(\eta^2\right)^3} + \frac{1}{2\left(\eta^2\right)^2}. \qquad (2.56)$$

By Theorem 2.1.1, we can verify that Fisher matrix is covariant under reparametrizations of the parameter space,

$$\begin{aligned} g_{11}(\xi) &= \tilde{g}_{11}(\eta) J_{11} J_{11} + \tilde{g}_{12}(\eta) J_{11} J_{21} + \tilde{g}_{21}(\eta) J_{21} J_{11} + \tilde{g}_{22}(\eta) J_{21} J_{21} \\ &= \tilde{g}_{11}(\eta) J_{11} J_{11} = -\frac{1}{2\eta^2} \frac{1}{\sigma^2} \frac{1}{\sigma^2} = \sigma^2 \frac{1}{\sigma^2} \frac{1}{\sigma^2} = \frac{1}{\sigma^2}, \end{aligned}$$

$$\begin{aligned} g_{12}(\xi) &= \tilde{g}_{11}(\eta) J_{11} J_{12} + \tilde{g}_{12}(\eta) J_{11} J_{22} + \tilde{g}_{21}(\eta) J_{21} J_{12} + \tilde{g}_{22}(\eta) J_{21} J_{22} \\ &= \sigma^2 \frac{1}{\sigma^2} \left(-\frac{2\mu}{\sigma^3} \right) + 2\mu\sigma^2 \frac{1}{\sigma^2} \frac{1}{\sigma^3} + 0 + 0 = 0, \end{aligned}$$

$$\begin{aligned} g_{21}(\xi) &= \tilde{g}_{11}(\eta) J_{12} J_{11} + \tilde{g}_{12}(\eta) J_{12} J_{21} + \tilde{g}_{21}(\eta) J_{22} J_{11} + \tilde{g}_{22}(\eta) J_{22} J_{21} \\ &= g_{12}(\xi) = 0, \end{aligned}$$

$$\begin{aligned} g_{22}(\xi) &= \tilde{g}_{11}(\eta) J_{12} J_{12} + \tilde{g}_{12}(\eta) J_{12} J_{22} + \tilde{g}_{21}(\eta) J_{22} J_{12} + \tilde{g}_{22}(\eta) J_{22} J_{22} \\ &= \sigma^2 \left(-\frac{2\mu}{\sigma^3}\right)^2 + 2\mu\sigma^2 \left(-\frac{2\mu}{\sigma^3}\right) \frac{1}{\sigma^3} 2 + 2\sigma^2 \left(2\mu^2 + \sigma^2\right) \left(\frac{1}{\sigma^3}\right)^2 = \frac{2}{\sigma^2}, \end{aligned}$$

where the Jacobian is

$$J(\eta, \xi) = \begin{pmatrix} \frac{1}{\sigma^2} & -\frac{2\mu}{\sigma^3} \\ 0 & \frac{1}{\sigma^3} \end{pmatrix}.$$

- **Multivariate Normal Distribution**

With the notation

$$\mu = (\mu_1, \ldots, \mu_k), \; A = (A_{ij}), \; A^{-1} = \left(A^{ij}\right),$$

we have

$$\begin{aligned} \partial_{\mu_r} \ell_x \left(\mu, A^{-1}\right) &= \partial_{\mu_r} \left[-\frac{1}{2} (x-\mu)^t A^{-1} (x-\mu) \right] \\ &= -\frac{1}{2} \partial_{\mu_r} \left[\sum_{i,j=1}^{k} A^{ij} (x-\mu)_i (x-\mu)_j \right] \\ &= -\frac{1}{2} \left[\sum_{j=1}^{k} A^{rj} \left(\partial_{\mu_r} (x-\mu)_r\right) (x-\mu)_j + \sum_{i=1}^{k} A^{ir} (x-\mu)_i \left(\partial_{\mu_r} (x-\mu)_r\right) \right] \\ &= \frac{1}{2} \left[\sum_{j=1}^{k} A^{rj} (x-\mu)_j + \sum_{i=1}^{k} A^{ir} (x-\mu)_i \right] \\ &= \frac{1}{2} \sum_{j=1}^{k} \left(A^{rj} + A^{jr}\right) (x_j - \mu_j) \\ &= \sum_{j=1}^{k} A^{rj} (x_j - \mu_j), \; \forall r \in \{1, \ldots, k\}. \end{aligned} \tag{2.57}$$

By Jacobi's formula [5] for the differential of a determinant,

$$\frac{\partial}{\partial t}\det M\left(t\right)=\det M\left(t\right)\operatorname{tr}(M\left(t\right)^{-1}\frac{\partial}{\partial t}M\left(t\right)), \tag{2.58}$$

we have

$$\begin{aligned}\frac{\partial}{\partial A^{is}}\det A^{-1}&=\det A^{-1}\operatorname{tr}\left(A\frac{\partial}{\partial A^{is}}A^{-1}\right)\\&=\det A^{-1}\operatorname{tr}\left(AB\right),\end{aligned}$$

where $B=\left(B_{pq}\right)$, and $B_{is}=1$, $B_{pq}=0$ for $p\neq i$ or $q\neq s$. Denote $C=AB$,

$$\frac{\partial}{\partial A^{is}}\det A^{-1}=\det A^{-1}\operatorname{tr}\left(C\right),$$

where $C_{pq}=\sum_{t=1}^{k}A_{pt}B_{tq}$ for $p,q\in\{1,\ldots,k\}$. We have $C_{ps}=A_{pi}$, and $C_{tq}=0$ for $q\neq s$ and $t\in\{1,\ldots,k\}$.

Since on the diagonal of C the only possible nonzero entry is $C_{ss}=A_{si}=A_{is}$, we obtain

$$\frac{\partial}{\partial A^{is}}\det A^{-1}=\det A^{-1}A_{si}=\frac{A_{is}}{\det A}, \tag{2.59}$$

where we applied $\left(\det A\right)\left(\det A^{-1}\right)=1$.

Hence, we have

$$\frac{\partial}{\partial A^{\alpha\beta}}\ln\left(\det A\right)=-\frac{\partial}{\partial A^{\alpha\beta}}\ln\left(\det A^{-1}\right)=-\frac{1}{\det A^{-1}}\frac{\partial}{\partial A^{\alpha\beta}}\det A^{-1}=-A_{\alpha\beta},$$

and, by equation (**??**),

$$\begin{aligned}\partial_{A^{\alpha\beta}}\ell_x\left(\mu, A^{-1}\right) &= \frac{1}{2}A_{\alpha\beta} - \frac{1}{2}\partial_{A^{\alpha\beta}}\left[\sum_{i,j=1}^{k} A^{ij}\left(x-\mu\right)_i\left(x-\mu\right)_j\right] \\ &= \frac{1}{2}A_{\alpha\beta} - \frac{1}{2}\partial_{A^{\alpha\beta}}\left[A^{\alpha\beta}\left(x-\mu\right)_\alpha\left(x-\mu\right)_\beta\right] \\ &= \frac{1}{2}A_{\alpha\beta} - \frac{1}{2}(x_\alpha-\mu_\alpha)(x_\beta-\mu_\beta). \qquad (2.60)\end{aligned}$$

We choose the components of μ and the entries of the upper triangle part of A^{-1} as parameters, denoted by

$$\begin{aligned}\xi &= \left(\xi^1, \ldots, \xi^{k+\frac{k(k+1)}{2}}\right) \\ &= \left(\mu_1, \ldots, \mu_k, A^{11}, \ldots, A^{1k}, A^{22}, \ldots, A^{2k}, A^{33}, \ldots, A^{3k}, \ldots A^{kk}\right).\end{aligned}$$

For $1 \leq r, s \leq k$, by equation (2.57), we have

$$\begin{aligned}g_{rs}\left(\xi\right) &= -E_\xi\left[\partial_s\partial_r\ell\left(\xi\right)\right] \\ &= -E_\xi\left[\partial_s\left(\sum_{j=1}^{k} A^{rj}(x_j-\mu_j)\right)\right] \\ &= -E_\xi\left[-A^{rs}\right] = A^{rs}. \qquad (2.61)\end{aligned}$$

For $1 \leq r \leq k$ and $k+1 \leq s \leq k + \frac{k(k+1)}{2}$, by equations (2.57) and (2.60), we have

$$\begin{aligned} g_{rs}(\xi) = g_{sr}(\xi) &= -E_{\xi}\left[\partial_s \partial_r \ell(\xi)\right] \\ &= -E_{\xi}\left[\partial_s \left(\sum_{j=1}^{k} A^{rj}(x_j - \mu_j)\right)\right] \\ &= -E_{\xi}\left[x_{\hat{j}} - \mu_{\hat{j}}\right] = 0, \end{aligned} \tag{2.62}$$

where $\hat{j}$ is an index such that $A^{r\hat{j}} = \xi^s$.

For $k+1 \le r \le s \le k + \frac{k(k+1)}{2}$, let $A^{ab} = \xi^r, A^{cd} = \xi^s$. By equation (2.60), we have

$$\begin{aligned} g_{rs}(\xi) &= -E_{\xi}\left[\partial_s \partial_r \ell(\xi)\right] = -E_{\xi}\left[\partial_{A^{cd}} \partial_{A^{ab}} \ell(\xi)\right] \\ &= -E_{\xi}\left[\partial_{A^{cd}}\left(\frac{1}{2}A_{ab} - \frac{1}{2}(x_a - \mu_a)(x_b - \mu_b)\right)\right] \\ &= -\frac{1}{2}E_{\xi}\left[\partial_{A^{cd}} A_{ab}\right] = -\frac{1}{2}\partial_s A_{ab}(\xi). \end{aligned} \tag{2.63}$$

Thus, we obtain the $k + \frac{k(k+1)}{2}$-dimensional Fisher information matrix of a multivariate normal distribution:

$$g = \begin{pmatrix} A^{-1}_{k\times k} & 0_{k\times \frac{k(k+1)}{2}} \\ 0_{\frac{k(k+1)}{2}\times k} & B_{\frac{k(k+1)}{2}\times \frac{k(k+1)}{2}} \end{pmatrix}, \tag{2.64}$$

where

$$B_{\frac{k(k+1)}{2}\times\frac{k(k+1)}{2}} = -\frac{1}{2}\begin{pmatrix} \partial_{k+1}A_{11}(\xi) & \dots & \partial_{k+\frac{k(k+1)}{2}}A_{11}(\xi) \\ \dots & \dots & \dots \\ \partial_{k+1}A_{1k}(\xi) & \dots & \partial_{k+\frac{k(k+1)}{2}}A_{1k}(\xi) \\ \partial_{k+1}A_{22}(\xi) & \dots & \partial_{k+\frac{k(k+1)}{2}}A_{22}(\xi) \\ \dots & \dots & \dots \\ \partial_{k+1}A_{kk}(\xi) & \dots & \partial_{k+\frac{k(k+1)}{2}}A_{kk}(\xi) \end{pmatrix}.$$

We can also use the vector/matrix parameters $\eta = (\mu, A^{-1})$ directly. Then, we have

$$\begin{aligned} \partial_\mu \ell_x\left(\mu, A^{-1}\right) &= \partial_\mu\left[-\frac{1}{2}(x-\mu)^t A^{-1}(x-\mu) + \frac{1}{2}\ln\left(\det A^{-1}\right) - \frac{k}{2}\ln(2\pi)\right] \\ &= -\frac{1}{2}\left(A^{-1} + \left(A^{-1}\right)^t\right)(x-\mu) \\ &= A^{-1}(x-\mu), \end{aligned} \tag{2.65}$$

$$\begin{aligned} \partial_{A^{-1}} \ell_x\left(\mu, A^{-1}\right) &= \partial_{A^{-1}}\left[-\frac{1}{2}(x-\mu)^t A^{-1}(x-\mu) + \frac{1}{2}\ln\left(\det A^{-1}\right) - \frac{k}{2}\ln(2\pi)\right] \\ &= -\frac{1}{2}(x-\mu)^t(x-\mu) + \frac{1}{2}A. \end{aligned} \tag{2.66}$$

By equation (1.8), we have

$$\begin{aligned} g_{11}\left(\mu, A^{-1}\right) = -E_\eta\left[\partial_\mu^2 \ell_x\left(\mu, A^{-1}\right)\right] &= -E_\eta\left[\partial_\mu\left(A^{-1}(x-\mu)\right)\right] \\ &= -E_\eta\left[-A^{-1}\right] \\ &= A^{-1}, \end{aligned}$$

$$\begin{aligned} g_{12}\left(\mu, A^{-1}\right) = g_{21}\left(\mu, A^{-1}\right) &= -E_\eta\left[\partial_\mu \partial_{A^{-1}} \ell_x\left(\mu, A^{-1}\right)\right] \\ &= -E_\eta\left[\partial_\mu\left(-\frac{1}{2}(x-\mu)^t(x-\mu)+\frac{1}{2}A\right)\right] \\ &= E_\eta[x-\mu] = 0, \end{aligned}$$

$$\begin{aligned} g_{22}\left(\mu, A^{-1}\right) &= -E_\eta\left[\partial^2_{A^{-1}} \ell_x\left(\mu, A^{-1}\right)\right] \\ &= -E_\eta\left[\partial_{A^{-1}}\left(-\frac{1}{2}(x-\mu)^t(x-\mu)+\frac{1}{2}A\right)\right] \\ &= \frac{1}{2}A^2. \end{aligned}$$

Hence, we obtain the Fisher information matrix for the vector/matrix parameters $\eta = (\mu, A^{-1})$:

$$g\left(\mu, A^{-1}\right) = \begin{pmatrix} A^{-1} & 0 \\ 0 & \frac{1}{2}A^2 \end{pmatrix}. \tag{2.67}$$

When $k = 1$, the above result reduces to the Fisher information matrix for a univariate normal distribution:

$$g\left(\mu, \sigma^{-2}\right) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{\sigma^4}{2} \end{pmatrix}.$$

For this case, we confirm the covariance of the Fisher information matrix (see equation (2.55)):

$$\tilde{g}_{11}(\mu, \sigma) = \sum_{r,s} g_{rs}\left(\mu, \sigma^{-2}\right) J_{r1} J_{s1} = g_{11}\left(\mu, \sigma^{-2}\right) = \frac{1}{\sigma^2},$$

$$\tilde{g}_{12}(\mu,\sigma)=\tilde{g}_{21}(\mu,\sigma)=0,$$

$$\tilde{g}_{22}(\mu,\sigma)=\sum_{r,s} g_{rs}\left(\mu,\sigma^{-2}\right)J_{r2}J_{s2}=g_{22}\left(\mu,\sigma^{-2}\right)\left(\frac{-2}{\sigma^3}\right)^2=\frac{2}{\sigma^2}.$$

- **Lognormal Distribution**

By equation (2.27), the log-likelihood function of a lognormal distribution is

$$\ln p(x;\mu,\sigma)=-\ln x-\ln\left(\sigma\sqrt{2\pi}\right)+\frac{\mu}{\sigma^2}\ln x-\frac{1}{2\sigma^2}(\ln x)^2-\frac{\mu}{2\sigma^2}.$$

For parameters $(\xi^1,\xi^2)=(\mu,\sigma)$,

$$\partial_\mu^2\ell_x(\xi)=\partial_\mu\left[-\frac{1}{\sigma^2}(\ln x-\mu)(-1)\right]=-\frac{1}{\sigma^2},$$

$$\partial_\sigma^2\ell_x(\xi)=\partial_\sigma\left[-\frac{1}{\sigma}+\frac{1}{\sigma^3}(\ln x-\mu)^2\right]=\frac{1}{\sigma^2}-\frac{3}{\sigma^4}(\ln x-\mu)^2,$$

$$\partial_\mu\partial_\sigma\ell_x(\xi)=\partial_\mu\left[-\frac{1}{\sigma}+\frac{1}{\sigma^3}(\ln x-\mu)^2\right]=-\frac{2}{\sigma^3}(\ln x-\mu).$$

By equations (2.28), (2.35) and (2.37), we have

$$E_\xi[\ln x]=E_\xi[F_1(x)]=\partial_1\psi(\xi)=\partial_1\left[\frac{-\left(\xi^1\right)^2}{4\xi^2}+\frac{1}{2}\ln\left(\frac{-\pi}{\xi^2}\right)\right]=-\frac{\xi^1}{2\xi^2}=\mu,\quad(2.68)$$

$$\begin{aligned}\mathrm{Var}\left[\ln x\right] &= E_\xi\left[(\ln x)^2\right] - \left(E_\xi\left[\ln x\right]\right)^2\\ &= E_\xi\left[F_2(x)\right] - \left(E_\xi\left[F_1(x)\right]\right)^2\\ &= \partial_2\psi(\xi) - \left(\partial_1\psi(\xi)\right)^2\\ &= \partial_2\left[\frac{-\left(\xi^1\right)^2}{4\xi^2} + \frac{1}{2}\ln\left(\frac{-\pi}{\xi^2}\right)\right] - \mu^2\\ &= \frac{\left(\xi^1\right)^2}{4\left(\xi^2\right)^2} - \frac{1}{2\xi^2} - \mu^2\\ &= \mu^2 + \sigma^2 - \mu^2 = \sigma^2. \end{aligned} \tag{2.69}$$

The Fisher-Riemann metric of a lognormal distribution coincides with that of a normal distribution model, as shown by

$$g_{11}(\xi) = -E_\xi\left[\partial_\mu^2 \ell_x(\xi)\right] = -E_\xi\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2},$$

$$g_{12}(\xi) = g_{21}(\xi) = -E_\xi\left[\partial_\mu\partial_\sigma \ell_x(\xi)\right] = -E_\xi\left[-\frac{2(\ln x - \mu)}{\sigma^3}\right] = \frac{2\left(E_\xi\left[\ln x\right] - \mu\right)}{\sigma^3} = 0,$$

$$\begin{aligned} g_{22}(\xi) = -E_\xi\left[\partial_\sigma^2 \ell_x(\xi)\right] &= -E_\xi\left[\frac{1}{\sigma^2} - \frac{3(\ln x - \mu)^2}{\sigma^4}\right] = -\frac{1}{\sigma^2} + \frac{3}{\sigma^4}E_\xi\left[(\ln x - \mu)^2\right]\\ = -\frac{1}{\sigma^2} + \frac{3}{\sigma^4}\mathrm{Var}\left[\ln x - \mu\right] &= -\frac{1}{\sigma^2} + \frac{3}{\sigma^4}\mathrm{Var}\left[\ln x\right] = \frac{2}{\sigma^2}, \end{aligned} \tag{2.70}$$

where we applied Equations (2.68) and (2.69).

From equations (2.21) and (2.28), we see that the normalization functions of the normal and lognormal distributions are the same. By (2.39), the Fisher metric components are given by equation (2.55), which agree well with the result in equation

(2.70).

- **Exponential Distribution**

For the exponential distribution, starting from the log-likelihood function

$$\ln p\left(x;\lambda\right) = -\xi x + \ln \xi,$$

we have

$$\partial_\xi \ell_x\left(\xi\right) = -x + \frac{1}{\xi},$$

$$\partial_\xi^2 \ell_x\left(\xi\right) = -\frac{1}{\xi^2}.$$

Then, the Fisher information is given by

$$g(\xi) = g_{11}\left(\xi\right) = -E_\xi\left[\partial_\xi^2 \ell_x\left(\xi\right)\right] = \frac{1}{\xi^2}. \tag{2.71}$$

- **Gamma Distribution**

For the gamma distribution, starting from the log-likelihood function

$$\ln p\left(x;\lambda\right) = -\alpha \ln \beta - \ln \Gamma\left(\alpha\right) - \ln x + \alpha \ln x - \frac{x}{\beta},$$

we have

$$\partial_\beta \ell_x\left(\xi\right) = -\frac{\alpha}{\beta} + \frac{x}{\beta^2},$$

where $\xi = (\xi^1, \xi^2) = (\alpha, \beta)$,

$$\partial_\alpha \partial_\beta \ell_x\left(\xi\right) = -\frac{1}{\beta},$$

$$\partial_\beta^2 \ell_x(\xi) = \frac{\alpha}{\beta^2} - \frac{2x}{\beta^3},$$

$$\partial_\alpha \ell_x(\xi) = -\ln\beta - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \ln x = -\ln\beta - \psi(\alpha) + \ln x,$$

where $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the *digamma function* of α,

$$\partial_\alpha^2 \ell_x(\xi) = -\psi'(\alpha) = -\psi_1(\alpha),$$

where $\psi_1(\alpha) = \psi'(\alpha)$ is the *trigamma function* of α.

Then, the Fisher information is given by

$$g_{11}(\xi) = -E_\xi\left[\partial_\alpha^2 \ell_x(\xi)\right] = E_\xi\left[\psi_1(\alpha)\right] = \psi_1(\alpha),$$

$$g_{21}(\xi) = g_{12}(\xi) = -E_\xi\left[\partial_\alpha \partial_\beta \ell_x(\xi)\right] = -E_\xi\left[-\frac{1}{\beta}\right] = \frac{1}{\beta},$$

$$g_{22}(\xi) = -E_\xi\left[\frac{\alpha}{\beta^2} - \frac{2x}{\beta^3}\right] = -\frac{\alpha}{\beta^2} + \frac{2E_\xi[x]}{\beta^3} = -\frac{\alpha}{\beta^2} + \frac{2\alpha\beta}{\beta^3} = \frac{\alpha}{\beta^2}. \qquad (2.72)$$

If we choose $\eta = (\eta^1, \eta^2) = \left(\alpha, -\frac{1}{\beta}\right)$, by equation (2.32), we have

$$\psi(\eta) = \ln\left(\beta^\alpha \Gamma(\alpha)\right) = \ln\left(\left(\frac{-1}{\eta^2}\right)^{\eta^1} \Gamma\left(\eta^1\right)\right).$$

Hence, the Fisher information under this reparametrization is given by

$$\tilde{g}_{11}(\eta) = \partial_1^2 \psi(\eta) = \partial_1^2 \left[\ln \left(\left(\frac{-1}{\eta^2} \right)^{\eta^1} \Gamma\left(\eta^1\right) \right) \right] = \partial_1^2 \left[-\eta^1 \ln\left(-\eta^2\right) + \ln \Gamma\left(\eta^1\right) \right]$$
$$= \partial_1 \left[-\ln\left(-\eta^2\right) + \frac{\Gamma'\left(\eta^1\right)}{\Gamma\left(\eta^1\right)} \right] = \psi_1\left(\eta^1\right),$$

$$\tilde{g}_{12}(\eta) = \tilde{g}_{21}(\eta) = \partial_2 \partial_1 \psi(\eta) = \partial_2 \left[-\ln\left(-\eta^2\right) + \frac{\Gamma'\left(\eta^1\right)}{\Gamma\left(\eta^1\right)} \right] = -\frac{1}{\eta^2},$$

$$\tilde{g}_{22}(\eta) = \partial_2^2 \psi(\eta) = \partial_2^2 \left[-\eta^1 \ln\left(-\eta^2\right) + \ln \Gamma\left(\eta^1\right) \right]$$
$$= \partial_2 \left[-\frac{\eta^1}{\eta^2} \right] = \frac{\eta^1}{\left(\eta^2\right)^2}. \tag{2.73}$$

With the Jacobian,

$$J(\eta, \xi) = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\beta^2} \end{pmatrix},$$

we have

$$g_{11}(\xi) = \tilde{g}_{11}(\eta) J_{11} J_{11} + \tilde{g}_{12}(\eta) J_{11} J_{21} + \tilde{g}_{21}(\eta) J_{21} J_{11} + \tilde{g}_{22}(\eta) J_{21} J_{21}$$
$$= \tilde{g}_{11}(\eta) J_{11} J_{11} = \psi_1\left(\eta^1\right) = \psi_1(\alpha),$$

$$g_{12}(\xi) = \tilde{g}_{11}(\eta) J_{11} J_{12} + \tilde{g}_{12}(\eta) J_{11} J_{22} + \tilde{g}_{21}(\eta) J_{21} J_{12} + \tilde{g}_{22}(\eta) J_{21} J_{22}$$
$$= \tilde{g}_{12}(\eta) J_{11} J_{22} = \left(-\frac{1}{\eta^2} \right) \left(\frac{1}{\beta^2} \right) = \frac{\beta}{\beta^2} = \frac{1}{\beta},$$

$$\begin{aligned} g_{21}(\xi) &= \tilde{g}_{11}(\eta) J_{12} J_{11} + \tilde{g}_{12}(\eta) J_{12} J_{21} + \tilde{g}_{21}(\eta) J_{22} J_{11} + \tilde{g}_{22}(\eta) J_{22} J_{21} \\ &= g_{12}(\xi) = 0, \end{aligned}$$

$$\begin{aligned} g_{22}(\xi) &= \tilde{g}_{11}(\eta) J_{12} J_{12} + \tilde{g}_{12}(\eta) J_{12} J_{22} + \tilde{g}_{21}(\eta) J_{22} J_{12} + \tilde{g}_{22}(\eta) J_{22} J_{22} \\ &= \tilde{g}_{22}(\eta) J_{22} J_{22} = \frac{\eta^1}{(\eta^2)^2} \left(\frac{1}{\beta^2} \right)^2 = \frac{\alpha}{\beta^2}. \end{aligned}$$

- **Beta Distribution**

For the beta distribution, starting from the log-likelihood function

$$\begin{aligned} \ln p(x; \lambda) &= -\ln(B(a,b)) - \ln(x(1-x)) + a \ln x + b \ln(1-x) \\ &= -\ln\left(\frac{\Gamma(a)\Gamma(b)}{\Gamma(a,b)} \right) - \ln(x(1-x)) + a \ln x + b \ln(1-x), \end{aligned}$$

we have

$$\partial_b \ell_x(\xi) = -\frac{\Gamma'(b)}{\Gamma(b)} + \frac{\Gamma'(a+b)}{\Gamma(a+b)} + \ln(1-x),$$

where $\xi = (\xi^1, \xi^2) = (a, b)$. Therefore,

$$\partial_a \partial_b \ell_x(\xi) = \psi_1(a+b),$$

$$\partial_b^2 \ell_x(\xi) = -\psi_1(b) + \psi_1(a+b),$$

$$\partial_a \ell_x(\xi) = -\frac{\Gamma'(a)}{\Gamma(a)} + \frac{\Gamma'(a+b)}{\Gamma(a+b)} + \ln x,$$

$$\partial_\alpha^2 \ell_x(\xi) = -\psi_1(a) + \psi_1(a+b),$$

where ψ_1 is the *trigamma function*.

Then, the Fisher information is given by

$$g_{11}(\xi) = -E_\xi\left[\partial_a^2 \ell_x(\xi)\right] = \psi_1(a) - \psi_1(a+b),$$

$$g_{21}(\xi) = g_{12}(\xi) = -E_\xi\left[\partial_a \partial_b \ell_x(\xi)\right] = -\psi_1(a+b),$$

$$g_{22}(\xi) = -E_\xi\left[\partial_b^2 \ell_x(\xi)\right] = \psi_1(b) - \psi_1(a+b). \tag{2.74}$$

By equation (2.34), the normalization function for the beta distribution is given by

$$\psi(\xi) = \ln\left(\frac{\Gamma(a)\Gamma(b)}{\Gamma(a,b)}\right).$$

We can directly compute the Fisher information under the same parametrization by equation (2.39) and thus obtain the same results.

$$g_{11}(\xi) = \partial_1^2 \psi(\xi) = \partial_1^2\left[\ln\left(\frac{\Gamma(a)\Gamma(b)}{\Gamma(a,b)}\right)\right] = \psi_1(a) - \psi_1(a+b),$$

$$g_{21}(\xi) = g_{12}(\xi) = \partial_2\partial_1\psi(\xi) = \partial_2\partial_1\left[\ln\left(\frac{\Gamma(a)\Gamma(b)}{\Gamma(a,b)}\right)\right] = -\psi_1(a+b),$$

$$g_{22}(\xi) = \partial_2^2\psi(\xi) = 2\left[\ln\left(\frac{\Gamma(a)\Gamma(b)}{\Gamma(a,b)}\right)\right] = \psi_1(b) - \psi_1(a+b).$$

2.3 Christoffel Symbols

In this section, we compute the Christoffel symbols for some common statistical models based on Definition 1.2.2 and Proposition (2.1.4).

- **Binomial Distribution**

We obtained the Fisher information for the binomial distribution in equation (2.43), $g\left(\xi\right)=\frac{n}{\xi(1-\xi)}$. The inverse $g^{-1}\left(\xi\right)$ is $\frac{\xi(1-\xi)}{n}$.

The Christoffel symbols of first and second kind are given by

$$\Gamma_{11,1}\left(\xi\right)=\frac{1}{2}\left(\partial_1 g_{11}+\partial_1 g_{11}-\partial_1 g_{11}\right)=\frac{1}{2}\partial_\xi\left(\frac{n}{\xi\left(1-\xi\right)}\right)=\frac{n\left(-1+2\xi\right)}{2\xi^2\left(1-\xi\right)^2}, \qquad (2.75)$$

$$\Gamma_{11}^{1}\left(\xi\right)=g^{11}\left(\xi\right)\Gamma_{11,1}\left(\xi\right)=\frac{\xi\left(1-\xi\right)}{n}\frac{n\left(-1+2\xi\right)}{2\xi^2\left(1-\xi\right)^2}=\frac{-1+2\xi}{2\xi\left(1-\xi\right)}. \qquad (2.76)$$

Under the parametrization $\eta=\ln\frac{\theta}{1-\theta}$, by Proposition (2.1.4) and equation (2.7), the Christoffel symbol of first kind is given by

$$\begin{aligned}\Gamma_{11,1}\left(\eta\right)&=\frac{1}{2}\partial_1^3\psi\left(\eta\right)=\frac{1}{2}\partial_\eta^3\left(n\ln\left(1+e^\eta\right)\right)\\&=\frac{ne^\eta\left(1-e^\eta\right)}{2\left(1+e^\eta\right)^3}=\frac{1}{2}n\xi\left(1-\xi\right)\left(1-2\xi\right).\end{aligned}$$

By equations (1.9) and (2.44), we have the same result:

$$\Gamma_{11,1}\left(\eta\right)=\frac{1}{2}\left(\partial_1 g_{11}+\partial_1 g_{11}-\partial_1 g_{11}\right)=\frac{1}{2}\partial_\eta\left(\frac{ne^\eta}{\left(1+e^\eta\right)^2}\right)=\frac{ne^\eta\left(1-e^\eta\right)}{2\left(1+e^\eta\right)^3}. \qquad (2.77)$$

The covariance of the Christoffel symbols for an exponential family (see Proposition 2.1.5) can be verified by the results in equations (2.75), (2.77) and the Jacobian $J_{11}(\eta, \xi) = \frac{1}{\xi(1-\xi)}$.

By equations (1.10), the Christoffel symbol of second kind is given by

$$\Gamma_{11}^{1}(\eta) = g^{11}(\eta)\Gamma_{11,1}(\eta) = \frac{(1+e^{\eta})^{2}}{ne^{\eta}} \frac{ne^{\eta}(1-e^{\eta})}{2(1+e^{\eta})^{3}} = \frac{1-e^{\eta}}{2(1+e^{\eta})}. \tag{2.78}$$

- **Poisson Distribution**

We obtained the Fisher information for the Poisson distribution in equation (2.49), $g(\lambda) = \frac{1}{\lambda}$. The inverse $g^{-1}(\lambda)$ is λ.

The Christoffel symbols of first and second kind are given by

$$\Gamma_{11,1} = \frac{1}{2}(\partial_1 g_{11} + \partial_1 g_{11} - \partial_1 g_{11}) = \frac{1}{2}\partial_{\lambda}\left(\frac{1}{\lambda}\right) = -\frac{1}{2\lambda^{2}},$$

$$\Gamma_{11}^{1} = g^{11}\Gamma_{11,1} = \lambda\left(-\frac{1}{2\lambda^{2}}\right) = -\frac{1}{2\lambda}. \tag{2.79}$$

- **Normal Distribution**

We obtained the Fisher information matrix for the normal distribution in equation (2.55),

$$g(\mu, \sigma) = \begin{pmatrix} \frac{1}{\sigma^{2}} & 0 \\ 0 & \frac{2}{\sigma^{2}} \end{pmatrix}.$$

The inverse matrix is given by

$$g^{pk}(\mu, \sigma) = \begin{pmatrix} \sigma^{2} & 0 \\ 0 & \frac{\sigma^{2}}{2} \end{pmatrix}.$$

By a straightforward computation, we obtain the nonzero Christoffel symbols of first kind with parameters $(\xi^1, \xi^2) = (\mu, \sigma)$,

$$\Gamma_{11,2} = \frac{1}{2}\left(\partial_1 g_{12} + \partial_1 g_{21} - \partial_2 g_{11}\right) = -\frac{1}{2}\partial_\sigma\left(\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^3},$$

$$\Gamma_{12,1} = \frac{1}{2}\left(\partial_1 g_{21} + \partial_2 g_{11} - \partial_1 g_{12}\right) = \frac{1}{2}\partial_\sigma\left(\frac{1}{\sigma^2}\right) = -\frac{1}{\sigma^3},$$

$$\Gamma_{22,2} = \frac{1}{2}\left(\partial_2 g_{22} + \partial_2 g_{22} - \partial_2 g_{22}\right) = \frac{1}{2}\partial_\sigma\left(\frac{2}{\sigma^2}\right) = -\frac{2}{\sigma^3}. \tag{2.80}$$

Accordingly, the Christoffel symbols of second kind are:

$$\Gamma^1_{11} = g^{11}\Gamma_{11,1} + g^{12}\Gamma_{11,2} = 0,$$

$$\Gamma^1_{21} = \Gamma^1_{12} = g^{11}\Gamma_{12,1} + g^{12}\Gamma_{12,2} = \sigma^2\left(-\frac{1}{\sigma^3}\right) = -\frac{1}{\sigma},$$

$$\Gamma^1_{22} = g^{11}\Gamma_{22,1} + g^{12}\Gamma_{22,2} = 0,$$

$$\Gamma^2_{11} = g^{21}\Gamma_{11,1} + g^{22}\Gamma_{11,2} = \frac{\sigma^2}{2}\frac{1}{\sigma^3} = \frac{1}{2\sigma},$$

$$\Gamma^2_{21} = \Gamma^2_{12} = g^{21}\Gamma_{12,1} + g^{22}\Gamma_{12,2} = 0,$$

$$\Gamma^2_{22} = g^{21}\Gamma_{22,1} + g^{22}\Gamma_{22,2} = \frac{\sigma^2}{2}\left(-\frac{2}{\sigma^3}\right) = -\frac{1}{\sigma}. \tag{2.81}$$

- **Exponential Distribution**

We obtained the Fisher information matrix for the exponential distribution in equation (2.71), $g(\xi) = \frac{1}{\xi^2}$. The inverse $g^{-1}(\xi)$ is ξ^2.

The Christoffel symbols of first and second kind are given by

$$\Gamma_{11,1} = \frac{1}{2}\left(\partial_1 g_{11} + \partial_1 g_{11} - \partial_1 g_{11}\right) = \frac{1}{2}\partial_\xi \left(\frac{1}{\xi^2}\right) = -\frac{1}{\xi^3},$$

$$\Gamma^1_{11} = g^{11}\Gamma_{11,1} = \xi^2\left(-\frac{1}{\xi^3}\right) = -\frac{1}{\xi}. \tag{2.82}$$

- **Gamma Distribution**

We obtained the Fisher information matrix for the gamma distribution in equation (2.72),

$$g\left(\alpha,\beta\right) = \begin{pmatrix} \psi_1\left(\alpha\right) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}.$$

The inverse matrix is given by

$$g^{pk}\left(\alpha,\beta\right) = \frac{1}{\alpha\psi_1\left(\alpha\right) - 1}\begin{pmatrix} \alpha & -\beta \\ -\beta & \psi_1\left(\alpha\right)\beta^2 \end{pmatrix}.$$

By a straightforward computation, we obtain the Christoffel symbols of first kind with parameters $(\xi^1, \xi^2) = (\alpha, \beta)$,

$$\Gamma_{11,1} = \frac{1}{2}\left(\partial_1 g_{11} + \partial_1 g_{11} - \partial_1 g_{11}\right) = \frac{1}{2}\psi_1'\left(\alpha\right),$$

$$\Gamma_{11,2} = \frac{1}{2}\left(\partial_1 g_{12} + \partial_1 g_{21} - \partial_2 g_{11}\right) = 0,$$

$$\Gamma_{21,1} = \Gamma_{12,1} = \frac{1}{2}\left(\partial_1 g_{21} + \partial_2 g_{11} - \partial_1 g_{12}\right) = 0,$$

$$\Gamma_{21,2} = \Gamma_{12,2} = \frac{1}{2}\left(\partial_1 g_{22} + \partial_2 g_{21} - \partial_2 g_{12}\right) = \frac{1}{2\beta^2},$$

$$\Gamma_{22,1} = \frac{1}{2}\left(\partial_2 g_{21} + \partial_2 g_{12} - \partial_1 g_{22}\right) = -\frac{3}{2\beta^2},$$

$$\Gamma_{22,2} = \frac{1}{2}\left(\partial_2 g_{22} + \partial_2 g_{22} - \partial_2 g_{22}\right) = \frac{1}{2}\partial_\beta\left(\frac{\alpha}{\beta^2}\right) = -\frac{\alpha}{\beta^3}. \tag{2.83}$$

Accordingly, the Christoffel symbols of second kind are:

$$\Gamma^1_{11} = g^{11}\Gamma_{11,1} + g^{12}\Gamma_{11,2} = \frac{\alpha\psi_1'(\alpha)}{2\left(\alpha\psi_1(\alpha) - 1\right)},$$

$$\Gamma^1_{21} = \Gamma^1_{12} = g^{11}\Gamma_{12,1} + g^{12}\Gamma_{12,2} = -\frac{1}{2\beta\left(\alpha\psi_1(\alpha) - 1\right)},$$

$$\Gamma^1_{22} = g^{11}\Gamma_{22,1} + g^{12}\Gamma_{22,2} = -\frac{\alpha}{2\beta^2\left(\alpha\psi_1(\alpha) - 1\right)},$$

$$\Gamma^2_{11} = g^{21}\Gamma_{11,1} + g^{22}\Gamma_{11,2} = -\frac{\beta\psi_1'(\alpha)}{2\left(\alpha\psi_1(\alpha) - 1\right)},$$

$$\Gamma^2_{21} = \Gamma^2_{12} = g^{21}\Gamma_{12,1} + g^{22}\Gamma_{12,2} = \frac{\psi_1(\alpha)}{2\left(\alpha\psi_1(\alpha) - 1\right)},$$

$$\Gamma^2_{22} = g^{21}\Gamma_{22,1} + g^{22}\Gamma_{22,2} = \frac{\sigma^2}{2}\left(-\frac{2}{\sigma^3}\right) = \frac{3 - 2\alpha\psi_1(\alpha)}{2\beta\left(\alpha\psi_1(\alpha) - 1\right)}. \tag{2.84}$$

- **Beta Distribution**

We obtained the Fisher information matrix for the beta distribution in equation (2.74),

$$g(a,b) = \begin{pmatrix} \psi_1(a) - \psi_1(a+b) & -\psi_1(a+b) \\ -\psi_1(a+b) & \psi_1(b) - \psi_1(a+b) \end{pmatrix}$$
$$= \sum_{n=0}^{\infty} \begin{pmatrix} \frac{1}{(a+n)^2} - \frac{1}{(a+b+n)^2} & -\frac{1}{(a+b+n)^2} \\ -\frac{1}{(a+b+n)^2} & \frac{1}{(b+n)^2} - \frac{1}{(a+b+n)^2} \end{pmatrix}.$$

The inverse matrix is given by

$$g^{pk}\left(\alpha,\beta\right)=\frac{1}{\det\left(g\left(a,b\right)\right)}\left(\begin{array}{cc}\psi_{1}\left(b\right)-\psi_{1}\left(a+b\right) & \psi_{1}\left(a+b\right)\\ \psi_{1}\left(a+b\right) & \psi_{1}\left(a\right)-\psi_{1}\left(a+b\right)\end{array}\right),$$

where $\det\left(g\left(a,b\right)\right)=\psi_{1}\left(a\right)\psi_{1}\left(b\right)-\left[\psi_{1}\left(a\right)+\psi_{1}\left(b\right)\right]\psi_{1}\left(a+b\right).$

By a straightforward computation, we obtain the Christoffel symbols of first kind with parameters $(\xi^1,\xi^2)=(a,b)$,

$$\Gamma_{11,1}=\frac{1}{2}\left(\partial_1 g_{11}+\partial_1 g_{11}-\partial_1 g_{11}\right)=\sum_{n=0}^{\infty}\left(-\frac{1}{\left(a+n\right)^3}+\frac{1}{\left(a+b+n\right)^3}\right),$$

$$\Gamma_{11,2}=\frac{1}{2}\left(\partial_1 g_{12}+\partial_1 g_{21}-\partial_2 g_{11}\right)=\sum_{n=0}^{\infty}\frac{1}{\left(a+b+n\right)^3},$$

$$\Gamma_{21,1}=\Gamma_{12,1}=\frac{1}{2}\left(\partial_1 g_{21}+\partial_2 g_{11}-\partial_1 g_{12}\right)=\sum_{n=0}^{\infty}\frac{1}{\left(a+b+n\right)^3},$$

$$\Gamma_{21,2}=\Gamma_{12,2}=\frac{1}{2}\left(\partial_1 g_{22}+\partial_2 g_{21}-\partial_2 g_{12}\right)=\sum_{n=0}^{\infty}\frac{1}{\left(a+b+n\right)^3},$$

$$\Gamma_{22,1}=\frac{1}{2}\left(\partial_2 g_{21}+\partial_2 g_{12}-\partial_1 g_{22}\right)=\sum_{n=0}^{\infty}\frac{1}{\left(a+b+n\right)^3},$$

$$\Gamma_{22,2}=\frac{1}{2}\left(\partial_2 g_{22}+\partial_2 g_{22}-\partial_2 g_{22}\right)=\sum_{n=0}^{\infty}\left(-\frac{1}{\left(b+n\right)^3}+\frac{1}{\left(a+b+n\right)^3}\right). \qquad (2.85)$$

Accordingly, the Christoffel symbols of second kind are:

$$\begin{aligned}\Gamma_{11}^{1} &= g^{11}\Gamma_{11,1} + g^{12}\Gamma_{11,2} \\ &= \frac{\psi_1(b) - \psi_1(a+b)}{\psi_1(a)\psi_1(b) - [\psi_1(a) + \psi_1(b)]\psi_1(a+b)} \sum_{n=0}^{\infty} \frac{-1}{(a+n)^3} \\ &+ \frac{\psi_1(b)}{\psi_1(a)\psi_1(b) - [\psi_1(a) + \psi_1(b)]\psi_1(a+b)} \sum_{n=0}^{\infty} \frac{1}{(a+b+n)^3},\end{aligned}$$

$$\begin{aligned}\Gamma_{21}^{1} = \Gamma_{12}^{1} &= g^{11}\Gamma_{12,1} + g^{12}\Gamma_{12,2} \\ &= \frac{\psi_1(b)}{\psi_1(a)\psi_1(b) - [\psi_1(a) + \psi_1(b)]\psi_1(a+b)} \sum_{n=0}^{\infty} \frac{1}{(a+b+n)^3},\end{aligned}$$

$$\begin{aligned}\Gamma_{22}^{1} &= g^{11}\Gamma_{22,1} + g^{12}\Gamma_{22,2} \\ &= \frac{\psi_1(b)}{\psi_1(a)\psi_1(b) - [\psi_1(a) + \psi_1(b)]\psi_1(a+b)} \sum_{n=0}^{\infty} \frac{1}{(a+b+n)^3} \\ &+ \frac{\psi_1(a+b)}{\psi_1(a)\psi_1(b) - [\psi_1(a) + \psi_1(b)]\psi_1(a+b)} \sum_{n=0}^{\infty} \frac{-1}{(b+n)^3},\end{aligned}$$

$$\begin{aligned}\Gamma_{11}^{2} &= g^{21}\Gamma_{11,1} + g^{22}\Gamma_{11,2} \\ &= \frac{\psi_1(a+b)}{\psi_1(a)\psi_1(b) - [\psi_1(a) + \psi_1(b)]\psi_1(a+b)} \sum_{n=0}^{\infty} \frac{-1}{(a+n)^3} \\ &+ \frac{\psi_1(a)}{\psi_1(a)\psi_1(b) - [\psi_1(a) + \psi_1(b)]\psi_1(a+b)} \sum_{n=0}^{\infty} \frac{1}{(a+b+n)^3},\end{aligned}$$

$$
\begin{aligned}
\Gamma_{21}^{2} &= \Gamma_{12}^{2} = g^{21}\Gamma_{12,1} + g^{22}\Gamma_{12,2} \\
&= \frac{\psi_1(a)}{\psi_1(a)\psi_1(b) - [\psi_1(a)+\psi_1(b)]\psi_1(a+b)} \sum_{n=0}^{\infty} \frac{1}{(a+b+n)^3},
\end{aligned}
$$

$$
\begin{aligned}
\Gamma_{22}^{2} &= g^{21}\Gamma_{22,1} + g^{22}\Gamma_{22,2}. \\
&= \frac{\psi_1(a) - \psi_1(a+b)}{\psi_1(a)\psi_1(b) - [\psi_1(a)+\psi_1(b)]\psi_1(a+b)} \sum_{n=0}^{\infty} \frac{-1}{(b+n)^3} \\
&+ \frac{\psi_1(a)}{\psi_1(a)\psi_1(b) - [\psi_1(a)+\psi_1(b)]\psi_1(a+b)} \sum_{n=0}^{\infty} \frac{1}{(a+b+n)^3}.
\end{aligned}
\tag{2.86}
$$

2.4 Geodesics

The geodesic equations (1.27) are solutions of a Riccati ODE system. In this section, we work on some examples of common statistical models.

- **Binomial Distribution**

Based on the results obtained in equation (2.78), we have

$$
\ddot{\xi} - \frac{-1+2\xi}{2\xi(1-\xi)}\left(\dot{\xi}\right)^2 = 0. \tag{2.87}
$$

Let $\dot{\xi} = u$, then $\ddot{\xi} = \frac{du}{ds} = \frac{du}{d\xi}\dot{\xi} = \frac{du}{d\xi}u$. So, we have

$$
\frac{du}{d\xi}u = \frac{-1+2\xi}{2\xi(1-\xi)}u^2.
$$

Rewrite it as $d(\ln u) = \frac{-1+2\xi}{2\xi(1-\xi)}d\xi = \frac{1}{2}\left(\frac{1}{1-\xi} - \frac{1}{\xi}\right)d\xi$. Integrating gives

$$\ln|u| = \frac{1}{2}\ln\frac{1-\xi}{\xi} + C_1.$$

So, we have

$$\dot{\xi} = u = C_2\sqrt{\frac{1-\xi}{\xi}} \Longleftrightarrow \sqrt{\frac{\xi}{1-\xi}}d\xi = C_2 ds \Longleftrightarrow \int\sqrt{\frac{\xi}{1-\xi}}d\xi = \int C_2 ds.$$

Let $\xi = \sin^2 t$, then $d\xi = 2\sin t\cos t\,dt$, and

$$\int 2\sin^2 t\,dt = \int (1-\cos 2t)\,dt = C_2 s + C_3 \Longleftrightarrow \sin 2t = As + B$$

with constants A and B. So, the geodesic equations are

$$\xi\left(s\right) = \sin^2 t = \frac{1-\cos 2t}{2} = \frac{1-\sqrt{1-\left(As+B\right)^2}}{2}, \tag{2.88}$$

where the parameter s is at the value such that $1-\left(As+B\right)^2 \geq 0$.

- **Poisson Distribution**

Based on the results obtained in equation (2.79), we have

$$\ddot{\lambda} - \frac{1}{2\lambda}\left(\dot{\lambda}\right)^2 = 0. \tag{2.89}$$

Writing the equation as

$$\frac{\ddot{\lambda}}{\dot{\lambda}} = \frac{\dot{\lambda}}{2\lambda}$$

and integrating yields $\ln|\dot{\lambda}| = \ln\sqrt{C'\lambda}$ with $C' > 0$ constant. Rewrite it as $\frac{\dot{\lambda}}{\sqrt{\lambda}} = C$

and integrating gives the geodesic equations

$$\lambda\left(s\right) = \left(K_1 s + K_2\right)^2 \tag{2.90}$$

with K_1, K_2 constants.

- **Normal Distribution**

Based on the results obtained in equations (2.80) and (2.81), we have

$$\ddot{\mu} - 2\frac{1}{\sigma}\dot{\mu}\dot{\sigma} = 0, \tag{2.91}$$

$$\ddot{\sigma} + \frac{1}{2\sigma}\left(\dot{\mu}\right)^2 - \frac{1}{\sigma}\left(\dot{\sigma}\right)^2 = 0. \tag{2.92}$$

By equation (2.91), we have

$$\frac{\dot{\mu}}{\mu} = \frac{2\dot{\sigma}}{\sigma} \Longleftrightarrow d\ln\dot{\mu} = 2d\ln\sigma \Longleftrightarrow \dot{\mu} = c\sigma^2 \tag{2.93}$$

with c constant.

If $c = 0$, then $\dot{\mu} = 0$. and thus μ is a constant, which corresponds to a vertical half line since $\sigma > 0$. By equation (2.92), we have

$$\ddot{\sigma} = \frac{1}{\sigma}\left(\dot{\sigma}\right)^2 \Longleftrightarrow \frac{\ddot{\sigma}}{\dot{\sigma}} = \frac{\dot{\sigma}}{\sigma} \Longleftrightarrow d\ln\dot{\sigma} = d\ln\sigma.$$

By integrating, we have

$$\sigma\left(s\right) = K_1 e^{C_1 s},\ s \in \left[0, T\right]. \tag{2.94}$$

with C_1, K_1 positive constants.

Hence, the geodesics in this case have the following equations:

$$\mu(s) = c_1, \tag{2.95}$$

$$\sigma(s) = \sigma(0)\, e^{C_1 s},\ s \in [0, T] \tag{2.96}$$

with constants $c_1 \in \mathbb{R}, C_1 \in \mathbb{R}^+$.

If $c \neq 0$ in equation (2.93), substituting $\dot{\mu} = c\sigma^2$ into equation (2.92) gives

$$\ddot{\sigma} + \frac{1}{2\sigma}\left(c\sigma^2\right)^2 - \frac{1}{\sigma}(\dot{\sigma})^2 = 0. \tag{2.97}$$

Let $\dot{\sigma} = u$, so we have $\ddot{\sigma} = \frac{du}{d\sigma}u$. The above equation becomes

$$\frac{du}{d\sigma}u\sigma + \frac{c^2\sigma^4}{2} - u^2 = 0.$$

Multiplying by the integrant factor $\frac{1}{\sigma^3}$ leads to the exact equation

$$Mdu + Nd\sigma = 0.$$

with

$$M = \frac{u}{\sigma^2},\ N = \frac{c^2\sigma}{2} - \frac{u^2}{\sigma^3}, \tag{2.98}$$

since

$$\frac{\partial M}{\partial \sigma} = -\frac{2u}{\sigma^3} = \frac{\partial N}{\partial u}.$$

Now we look for the function $f(\sigma, u)$ such that $df(\sigma, u) = Mdu + Nd\sigma = 0$. We start by

$$\frac{df\left(\sigma,u\right)}{du}=M=\frac{u}{\sigma^2}.$$

Integrating yields

$$f\left(\sigma,u\right)=\frac{u^2}{2\sigma^2}+h\left(\sigma\right) \tag{2.99}$$

By differentiating result with respect to σ and comparing the result to N,

$$\frac{df\left(\sigma,u\right)}{d\sigma}=-\frac{u^2}{\sigma^3}+h'\left(\sigma\right)=\frac{c^2\sigma}{2}-\frac{u^2}{\sigma^3},$$

we obtain

$$h'\left(\sigma\right)=\frac{c^2\sigma}{2},$$

so

$$h\left(\sigma\right)=\frac{c^2\sigma^2}{4}+c_2,$$

with c_2 constant. Thus

$$f\left(\sigma,u\right)=\frac{u^2}{2\sigma^2}+\frac{c^2\sigma^2}{4}=K_2 \tag{2.100}$$

with K_2 positive constant. Solving for u, we have

$$u^2=2\sigma^2\left(K_2-\frac{c^2\sigma^2}{4}\right)\Longleftrightarrow\frac{\dot{\sigma}}{\sigma}=\sqrt{2K_2-\frac{c^2\sigma^2}{2}}=\frac{c}{\sqrt{2}}\sqrt{\frac{4K_2}{c^2}-\sigma^2}=\frac{c}{\sqrt{2}}\sqrt{C_2^2-\sigma^2}$$

where $C_2^2=\frac{4K_2}{c^2}$. Integrating the above equation, we have

$$\int_{\sigma(s_0)}^{\sigma(s)}\frac{d\sigma}{\sigma\sqrt{C_2^2-\sigma^2}}=\int_{s_0}^{s}\frac{c}{\sqrt{2}}dt. \tag{2.101}$$

By the integral formulae

$$\int \frac{dx}{x\sqrt{a^2-x^2}} = -\frac{1}{a}\mathrm{sech}^{-1}\left(\frac{x}{a}\right)+C,$$

we obtain

$$\frac{c}{\sqrt{2}}\left(s-s_0\right) = -\frac{1}{C_2}\mathrm{sech}^{-1}\left(\frac{\sigma\left(s\right)}{C_2}\right)+K_3,$$

where

$$K_3 = \frac{1}{C_2}\mathrm{sech}^{-1}\left(\frac{\sigma\left(s_0\right)}{C_2}\right).$$

Solving for σ, we have

$$\sigma\left(s\right) = C_2\mathrm{sech}\left[C_2\left(K_3-\frac{c}{\sqrt{2}}\left(s-s_0\right)\right)\right]. \tag{2.102}$$

Since $\dot{\mu} = c\sigma^2$, we have

$$\begin{aligned}\mu\left(s\right) &= \int_{s_0}^{s} cC_2^2\mathrm{sech}^2\left[C_2\left(K_3-\frac{c}{\sqrt{2}}\left(t-s_0\right)\right)\right]dt\\ &= \int_{s_0}^{s} cC_2^2\mathrm{sech}^2\left[C_2K_3-\frac{cC_2}{\sqrt{2}}\left(t-s_0\right)\right]dt\\ &= cC_2^2\frac{\sqrt{2}}{-cC_2}\left\{\tanh\left[C_2K_3-\frac{cC_2}{\sqrt{2}}\left(s-s_0\right)\right]-\tanh\left(C_2K_3\right)\right\}\\ &= -\sqrt{2}C_2\tanh\left[C_2K_3-\frac{cC_2}{\sqrt{2}}\left(s-s_0\right)\right]+K_4, \end{aligned} \tag{2.103}$$

where $K_4 = \sqrt{2}C_2\tanh\left(C_2K_3\right)$.

Since we have

$$\sigma\left(s\right)^2+\frac{1}{2}\left(\mu\left(s\right)-K_4\right)^2 = \frac{4K_2}{c^2},$$

the geodesics are half-ellipses with $\sigma > 0$.

- **Exponential Distribution**

Based on the results obtained in equation (2.82), we have

$$\ddot{\xi} - \frac{1}{\xi}\left(\dot{\xi}\right)^2 = 0. \tag{2.104}$$

Writing the equation as

$$\frac{\ddot{\xi}}{\dot{\xi}} = \frac{\dot{\xi}}{\xi}$$

and integrating yields $\ln \dot{\xi} = \ln (C\xi)$ with $C > 0$ constant. Rewrite it as $\frac{\dot{\xi}}{\xi} = C$ and Integrating gives the geodesics equation

$$\xi(s) = Ke^{Cs},\ s \in [0, T] \tag{2.105}$$

with $C, K > 0$ constants.

Gamma Distribution

Based on the results obtained in equation (2.84), the geodesic equations are the solutions of the system of ODEs:

$$\ddot{\alpha} - \frac{\alpha\psi_1'(\alpha)}{2(\alpha\psi_1(\alpha) - 1)}(\dot{\alpha})^2 - \frac{1}{\beta(\alpha\psi_1(\alpha) - 1)}\dot{\alpha}\dot{\beta} - \frac{\alpha}{2\beta^2(\alpha\psi_1(\alpha) - 1)}\left(\dot{\beta}\right)^2 = 0, \tag{2.106}$$

$$\ddot{\beta} - \frac{\beta\psi_1'(\alpha)}{2(\alpha\psi_1(\alpha) - 1)}(\dot{\alpha})^2 + \frac{\psi_1(\alpha)}{2(\alpha\psi_1(\alpha) - 1)}\dot{\alpha}\dot{\beta} + \frac{3 - 2\alpha\psi_1(\alpha)}{2\beta(\alpha\psi_1(\alpha) - 1)}\left(\dot{\beta}\right)^2 = 0. \tag{2.107}$$

2.5 α-Autotoparallel Curves

- **Normal Distribution**

We obtained the following results in equation (2.54):

$$\partial_\sigma \ell_x(\xi) = -\frac{1}{\sigma} - \frac{(x-\mu)^2}{2}\left(-\frac{2}{\sigma^3}\right) = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3},\ \partial_\sigma^2 \ell_x(\xi) = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4},$$

$$\partial_\mu \ell_x(\xi) = \frac{x-\mu}{\sigma^2},\ \partial_\mu^2 \ell_x(\xi) = -\frac{1}{\sigma^2},\ \partial_\sigma\partial_\mu \ell_x(\xi) = \partial_\mu\partial_\sigma \ell_x(\xi) = -\frac{2(x-\mu)}{\sigma^3}.$$

By Proposition 1.2.7, we compute the components of $\nabla^{(\alpha)}$-connection for the normal distribution. We have

$$\begin{aligned}
\Gamma_{11,1}^{(\alpha)} &= E_\xi\left[\left(\partial_\mu^2 \ell + \frac{1-\alpha}{2}(\partial_\mu \ell)^2\right)\partial_\mu \ell\right] \\
&= E_\xi\left[\left(-\frac{1}{\sigma^2} + \frac{1-\alpha}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right)\left(\frac{x-\mu}{\sigma^2}\right)\right] \\
&= 0,
\end{aligned}$$

where we used the fact, $E_\xi\left[(x-\mu)^{2k+1}\right] = 0$ for $k \in \mathbb{Z}$. Similarly, we obtain the other zero components:

$$\Gamma_{21,2}^{(\alpha)} = \Gamma_{12,2}^{(\alpha)} = \Gamma_{22,1}^{(\alpha)} = 0.$$

We also have

$$\begin{aligned}
\Gamma^{(\alpha)}_{11,2} &= E_\xi\left[\left(\partial^2_\mu \ell + \frac{1-\alpha}{2}\left(\partial_\mu \ell\right)^2\right)\partial_\sigma \ell\right] \\
&= E_\xi\left[\left(-\frac{1}{\sigma^2} + \frac{1-\alpha}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right)\left(-\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}\right)\right] \\
&= \frac{1}{\sigma^3} + E_\xi\left[\frac{\alpha-3}{2}\frac{(x-\mu)^2}{\sigma^5} + \frac{1-\alpha}{2}\frac{(x-\mu)^4}{\sigma^7}\right] \\
&= \frac{1}{\sigma^3} + \frac{\alpha-3}{2}\frac{\sigma^2}{\sigma^5} + \frac{1-\alpha}{2}\frac{3\sigma^4}{\sigma^7} \\
&= \frac{1-\alpha}{\sigma^3}.
\end{aligned}$$

where we used the fact, $E_\xi\left[(x-\mu)^2\right] = \sigma^2$ and $E_\xi\left[(x-\mu)^4\right] = 3\sigma^4$. Similarly, we obtain the other nonzero components:

$$\Gamma^{(\alpha)}_{12,1} = \Gamma^{(\alpha)}_{21,1} = -\frac{1+\alpha}{\sigma^3},\ \Gamma^{(\alpha)}_{22,2} = -\frac{2\left(1+2\alpha\right)}{\sigma^3}.$$

Based on the Christoffel symbols of first kind obtained as above, we compute the Christoffel symbols of second kind:

$$\begin{aligned}
\Gamma^{(\alpha)1}_{ij} &= g^{11}\Gamma^{\alpha}_{ij,1} + g^{12}\Gamma^{\alpha}_{ij,2} = \sigma^2\Gamma^{\alpha}_{ij,1} \\
&= \sigma^2\begin{pmatrix} 0 & -\frac{1+\alpha}{\sigma^3} \\ -\frac{1+\alpha}{\sigma^3} & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1+\alpha}{\sigma} \\ -\frac{1+\alpha}{\sigma} & 0 \end{pmatrix},
\end{aligned}$$

$$\begin{aligned}\Gamma^{(\alpha)2}_{ij} &= g^{21}\Gamma^{\alpha}_{ij,1} + g^{22}\Gamma^{\alpha}_{ij,2} = \frac{\sigma^2}{2}\Gamma^{\alpha}_{ij,2} \\ &= \frac{\sigma^2}{2}\begin{pmatrix} \frac{1-\alpha}{\sigma^3} & 0 \\ 0 & -\frac{2(1+2\alpha)}{\sigma^3} \end{pmatrix} = \begin{pmatrix} \frac{1-\alpha}{2\sigma} & 0 \\ 0 & -\frac{1+2\alpha}{\sigma} \end{pmatrix}.\end{aligned}$$

Hence, the Riccati equations (1.27) for the α-autoparallel curves are given by

$$\ddot{\mu} - 2\frac{(1+\alpha)}{\sigma}\dot{\mu}\dot{\sigma} = 0, \tag{2.108}$$

$$\ddot{\sigma} + \frac{1-\sigma}{2\sigma}\left(\dot{\mu}\right)^2 - \frac{1+2\alpha}{\sigma}\left(\dot{\sigma}\right)^2 = 0. \tag{2.109}$$

By equation (2.108), we have

$$\frac{\dot{\mu}}{\mu} = \frac{2(1\dot{+}\alpha)\,\sigma}{\sigma} \Longleftrightarrow d\ln\dot{\mu} = 2\left(1+\alpha\right)d\ln\sigma \Longleftrightarrow \dot{\mu} = c\sigma^{2(1+\alpha)} \tag{2.110}$$

with c constant. Substituting into equation (2.109) gives

$$\ddot{\sigma} + \frac{1-\alpha}{2\sigma}c^2\sigma^{4(1+\alpha)} - \frac{1+2\alpha}{\sigma}\left(\dot{\sigma}\right)^2 = 0. \tag{2.111}$$

Let $\dot{\sigma} = u$, so we have $\ddot{\sigma} = \frac{du}{d\sigma}u$. The above equation becomes

$$\frac{du}{d\sigma}u + \frac{\left(1-\alpha\right)c^2\sigma^{4(1+\alpha)}}{2\sigma} - \frac{1+2\alpha}{\sigma}u^2 = 0.$$

Multiplying by the integrant factor $\frac{1}{\sigma^{4\alpha+2}}$ leads to the exact equation

$$Mdu + Nd\sigma = 0.$$

with

$$M = \frac{u}{\sigma^{4\alpha+2}},\ N = \frac{(1-\alpha)\,c^2\sigma}{2} - \frac{(1+2\alpha)\,u^2}{\sigma^{4\alpha+3}}, \tag{2.112}$$

since

$$\frac{\partial M}{\partial \sigma} = -\frac{(4\alpha+2)\,u}{\sigma^{4\alpha+3}} = \frac{\partial N}{\partial u}.$$

Now we look for the function $f(\sigma, u)$ such that $df(\sigma, u) = Mdu + Nd\sigma = 0$. We start by

$$\frac{df(\sigma, u)}{du} = M = \frac{u}{\sigma^{4\alpha+2}}.$$

Integrating yields

$$f(\sigma, u) = \frac{u^2}{2\sigma^{4\alpha+2}} + h(\sigma) \tag{2.113}$$

By differentiating result with respect to σ and comparing the result to N,

$$\frac{df(\sigma, u)}{d\sigma} = -\frac{(2\alpha+1)\,u^2}{\sigma^{4\alpha+3}} + h'(\sigma) = \frac{(1-\alpha)\,c^2\sigma}{2} - \frac{(1+2\alpha)\,u^2}{\sigma^{4\alpha+3}},$$

we obtain

$$h'(\sigma) = \frac{(1-\alpha)\,c^2\sigma}{2},$$

so

$$h(\sigma) = \frac{(1-\alpha)\,c^2\sigma^2}{4} + c_2,$$

with c_2 constant. Thus

$$f(\sigma, u) = \frac{u^2}{2\sigma^{4\alpha+2}} + \frac{(1-\alpha)\,c^2\sigma^2}{4} = K_2 \tag{2.114}$$

with K_2 positive constant. Solving for u, we have

$$u^2 = 2\sigma^{4\alpha+2}\left(K_2 - \frac{(1-\alpha)\,c^2\sigma^2}{4}\right)$$

$$\Longleftrightarrow \frac{\dot{\sigma}}{\sigma^{2\alpha+1}} = \sqrt{2K_2 - \frac{(1-\alpha)\,c^2\sigma^2}{2}} = c\sqrt{\frac{1-\alpha}{2}}\sqrt{\frac{4K_2}{(1-\alpha)\,c^2} - \sigma^2} = c\sqrt{\frac{1-\alpha}{2}}\sqrt{C_2^2 - \sigma^2}$$

where $C_2^2 = \frac{4K_2}{(1-\alpha)c^2}$. Integrating the above equation, we have

$$\int_{\sigma(s_0)}^{\sigma(s)} \frac{d\sigma}{\sigma^{2\alpha+1}\sqrt{C_2^2-\sigma^2}} = \int_{s_0}^{s} c\sqrt{\frac{1-\alpha}{2}}dt. \tag{2.115}$$

By equation (2.110), the μ-component is given by

$$\mu = c\int \sigma\,(s)^{2(1+\alpha)}\,ds. \tag{2.116}$$

If $\alpha = -1$, the above equation becomes

$$\int_{\sigma(s_0)}^{\sigma(s)} \frac{\sigma d\sigma}{\sqrt{C_2^2-\sigma^2}} = \int_{s_0}^{s} c\sqrt{\frac{1-\alpha}{2}}dt.$$

The integrating result is

$$-\sqrt{C_2^2 - \sigma\,(s)^2} + \sqrt{C_2^2 - \sigma\,(s_0)^2} = c\sqrt{\frac{1-\alpha}{2}}\,(s-s_0)$$

$$\Rightarrow\ \sigma\,(s) = \sqrt{C_2^2 - \left[\sqrt{C_2^2 - \sigma\,(s_0)^2} - c\sqrt{\frac{1-\alpha}{2}}\,(s-s_0)\right]^2} \tag{2.117}$$

with $C_2^2 = \frac{2K_2}{c^2}$. By equation (2.116), the μ-component is given by

$$\mu = cs + \mu(0). \tag{2.118}$$

- **Exponential Distribution**

For the exponential distribution, we have the following derivatives:

$$\partial_\xi \ell_x(\xi) = -x + \frac{1}{\xi},$$

$$\partial_\xi^2 \ell_x(\xi) = -\frac{1}{\xi^2}.$$

By Proposition 1.2.7, the component of $\nabla^{(\alpha)}$-connection for the exponential distribution is

$$\begin{aligned}
\Gamma_{11,1}^{(\alpha)} &= E_\xi\left[\left(\partial_\xi^2 \ell + \frac{1-\alpha}{2}\left(\partial_\xi \ell\right)^2\right)\partial_\xi \ell\right] \\
&= E_\xi\left[\left(-\frac{1}{\xi^2} + \frac{1-\alpha}{2}\left(-x+\frac{1}{\xi}\right)^2\right)\left(-x+\frac{1}{\xi}\right)\right] \\
&= \frac{\alpha-1}{\xi^3},
\end{aligned}$$

where we used the fact, $E_\xi[x^n] = \frac{n!}{\xi^n}$ for $n \in \mathbb{Z}$.

Based on equation (2.71) and the Christoffel symbol of first kind obtained as above, we obtain the Christoffel symbol of second kind:

$$\Gamma_{11}^{(\alpha)1} = g^{11}\Gamma_{11,1}^{(\alpha)} = \xi^2 \frac{\alpha-1}{\xi^3} = \frac{\alpha-1}{\xi}.$$

Hence, the equation (1.27) for the α-autoparallel curves is given by

$$\ddot{\xi} + \frac{\alpha - 1}{\xi}\left(\dot{\xi}\right)^2 = 0. \tag{2.119}$$

Therefore,

$$\frac{\ddot{\xi}}{\dot{\xi}} = (1-\alpha)\frac{\dot{\xi}}{\xi} \Longrightarrow d\ln\dot{\xi} = (1-\alpha)\, d\ln\xi \Longrightarrow \ln\dot{\xi} = (1-\alpha)\ln\left(C_1\xi\right)$$

with $C_1 > 0$ constant. Rewrite it as $\frac{\dot{\xi}}{\xi^{1-\alpha}} = C_1^{1-\alpha} = C_2$ and integrating, we have

$$\int \xi^{\alpha-1} d\xi = C_2 s + C_{3.}$$

If $\alpha \neq 0$, we obtain the equation of α-autoparallel curves:

$$\xi\left(s\right) = \left(Cs + D\right)^{1/\alpha},\ s \in \left[0, T\right]. \tag{2.120}$$

with C, D constants.

2.6 Jeffreys Prior

- **Normal Distribution**

We consider the prior on a normal distribution model with the mean fixed,

$$\mathcal{S}_\mu = \left\{p_\xi; E_\xi\left[x\right] = \mu, Var\left[x\right] > 1\right\} = \left\{p_{(\mu,\sigma)}; \sigma > 1\right\}.$$

This is a vertical half line in the upper-half plane. The determinant of Fisher information matrix is

$$G(\xi) = \det g(\xi) = \det \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix} = \frac{2}{\sigma^4}.$$

Then the volume is

$$Vol(\mathcal{S}_\mu) = \int_1^\infty \sqrt{G(\xi)}d\sigma = \int_1^\infty \sqrt{\frac{2}{\sigma^4}}d\sigma = \sqrt{2} < \infty.$$

Hence, the Jeffreys prior on $\mathcal{S}_\mu$ is given by

$$Q(\xi) = \frac{\sqrt{G(\xi)}}{Vol(\mathcal{S}_\mu)} = \frac{1}{\sigma^2}.$$

- **Exponential Distribution**

We consider a statistical model of the exponential distribution as

$$\mathcal{S}_\lambda = \left\{p_\lambda; E_\lambda[x] = \frac{1}{\lambda}, \lambda \in [1, e]\right\} = \{p_\lambda; \lambda \in [1, e]\}.$$

The determinant of Fisher information matrix is

$$G(\lambda) = \det g(\lambda) = \frac{1}{\lambda^2}.$$

Then the volume is

$$Vol\left(\mathcal{S}_\lambda\right) = \int_1^e \sqrt{G\left(\lambda\right)} d\lambda = \int_1^e \frac{1}{\lambda} d\lambda = 1 < \infty.$$

Hence, the Jeffreys prior on $\mathcal{S}_\lambda$ is given by

$$Q\left(\lambda\right) = \frac{\sqrt{G\left(\lambda\right)}}{Vol\left(\mathcal{S}_\lambda\right)} = \frac{1}{\lambda}.$$

Chapter 3

The Geometry of Entropy of Exponential Families

3.1 Entropy

In Section 1.3.1, we gave the definition of entropy on a statistical model. The entropy for some distributions of the exponential family is computed in the following examples.

- **Poisson Distribution**

The probability mass function of a Poisson distribution is

$$p\left(x;\xi\right)=e^{-\xi}\frac{\xi^{x}}{x!},\ x\in\mathbb{N},\xi\in\mathbb{R}.$$

The entropy is

$$\begin{aligned} H(\xi) &= -\sum_{x\in\mathbb{N}} p(x,\xi)\ln p(x,\xi) \\ &= -\sum_{x\in\mathbb{N}} p(x,\xi)\left(-\xi + x\ln\xi - \ln x!\right) \\ &= \xi - \ln\xi E_{\xi}[x] + e^{-\xi}\sum_{x\in\mathbb{N}} \frac{\xi^x}{x!}\ln x! \\ &= \xi(1-\ln\xi) + e^{-\xi}\sum_{x\in\mathbb{N}} \frac{\xi^x \ln x!}{x!}. \end{aligned}$$

Note that $\lim_{\xi\to 0+} H = \lim_{\xi\to 0+} \dfrac{1-\ln\xi}{1/\xi} = \lim_{\xi\to 0+} \dfrac{-1/\xi}{-1/\xi^2} = 0$. Since

$$\lim_{x\to\infty} \left| \frac{\frac{\xi^{x+1}\ln(x+1)!}{(x+1)!}}{\frac{\xi^x \ln x!}{x!}} \right| = \xi \lim_{x\to\infty} \frac{\ln(x+1)!}{(x+1)\ln x!} = 0,$$

the series $\sum_{x\in\mathbb{N}} \frac{\xi^x \ln x!}{x!}$ has an infinite radius of convergence. Hence, $H(\xi) < \infty$.

- **Normal Distribution**

The density of a normal distribution is

$$p(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x \in \mathcal{X} = \mathbb{R},\ (\mu,\sigma) \in \mathbb{R}\times(0,\infty).$$

The entropy is

$$\begin{aligned} H(\mu,\sigma) &= -\int_{\mathcal{X}} p(x;\mu,\sigma)\ln p(x;\mu,\sigma)\,dx \\ &= \int_{\mathcal{X}} p(x;\mu,\sigma)\left(\ln\sigma + \ln\sqrt{2\pi} + \frac{(x-\mu)^2}{2\sigma^2}\right)dx \\ &= \ln\sigma + \ln\sqrt{2\pi} + \int_{\mathcal{X}} p(x;\mu,\sigma)\frac{(x-\mu)^2}{2\sigma^2}dx \\ &= \ln\sigma + \ln\sqrt{2\pi} + \frac{1}{2\sigma^2}\cdot\sigma^2 \\ &= \ln\left(\sigma\sqrt{2\pi e}\right). \end{aligned}$$

It follows that the entropy is independent of μ. The change of coordinates $\varphi : \mathbb{E} \to \mathbb{E}$ under which the entropy is invariant are only the translations $\varphi(\mu,\sigma) = \varphi(\mu + c,\sigma), c \in \mathbb{R}$. Also, the entropy is increasing logarithmically as a function of σ, with $\lim\limits_{\sigma\to 0+} H = -\infty$ and $\lim\limits_{\sigma\to\infty} H = \infty$.

- **Lognormal Distribution**

The density of a lognormal distribution is

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}},\ x > 0,\ (\mu,\sigma) \in \mathbb{R}\times(0,\infty).$$

By equation (2.35) in Proposition 2.1.1 and equation (2.28),

$$\begin{aligned} E\left[\ln x\right] = E\left[F_1\left(x\right)\right] &= \partial_1 \psi\left(\xi\right) \\ &= \frac{\partial}{\partial \xi^1}\left[\frac{-\left(\xi^1\right)^2}{4\xi^2} + \frac{1}{2}\ln\left(\frac{-\pi}{\xi^2}\right)\right] \\ &= \frac{-\xi^1}{2\xi^2} = \mu, \end{aligned}$$

$$\begin{aligned} E\left[\left(\ln x\right)^2\right] = E\left[F_2\left(x\right)\right] &= \partial_2 \psi\left(\xi\right) \\ &= \frac{\partial}{\partial \xi^2}\left[\frac{-\left(\xi^1\right)^2}{4\xi^2} + \frac{1}{2}\ln\left(\frac{-\pi}{\xi^2}\right)\right] \\ &= \frac{\left(\xi^1\right)^2}{4\left(\xi^2\right)^2} - \frac{1}{2\xi^2} = \mu^2 + \sigma^2. \end{aligned}$$

The entropy is

$$\begin{aligned} H\left(\mu,\sigma\right) &= -\int_{\mathcal{X}} p\left(x;\mu,\sigma\right)\ln p\left(x;\mu,\sigma\right)dx \\ &= -\int_{\mathcal{X}} p\left(x;\mu,\sigma\right)\left(-\ln x - \ln\left(\sigma\sqrt{2\pi}\right) + \frac{\mu}{\sigma^2}\ln x - \frac{1}{2\sigma^2}\left(\ln x\right)^2 - \frac{\mu}{2\sigma^2}\right)dx \\ &= \left(1 - \frac{\mu}{\sigma^2}\right)E\left[\ln x\right] + \ln\left(\sigma\sqrt{2\pi}\right) + \frac{1}{2\sigma^2}E\left[\left(\ln x\right)^2\right] + \frac{\mu}{2\sigma^2} \\ &= \left(1 - \frac{\mu}{\sigma^2}\right)\mu + \ln\left(\sigma\sqrt{2\pi}\right) + \frac{1}{2\sigma^2}\left(\mu^2 + \sigma^2\right) + \frac{\mu}{2\sigma^2} \\ &= \mu + \ln\sigma + \frac{1}{2} + \ln\sqrt{2\pi}. \end{aligned}$$

It follows that, being different from the case for a normal distribution, the entropy of a lognormal distribution is linearly dependent on μ.

- **Exponential Distribution**

The density of an exponential distribution is

$$p\left(x;\xi\right)=\xi e^{-\xi x},\ x>0,\ \xi>0.$$

The entropy is

$$\begin{aligned} H\left(\xi\right) &= -\int_0^\infty p\left(x\right)\ln p\left(x\right)dx \\ &= -\int_0^\infty p\left(x\right)\left(\ln\xi-\xi x\right)dx \\ &= -\ln\xi+\xi\int_0^\infty xp\left(x\right)dx \\ &= 1-\ln\xi. \end{aligned}$$

It follows that the entropy is a decreasing function of ξ. If we choose the parameter $\lambda=\frac{1}{\xi}$, $H\left(\lambda\right)=1+\ln\lambda$.

- **Gamma Distribution**

The density of a gamma distribution is

$$p\left(x;\alpha,\beta\right)=\frac{1}{\beta^{\alpha}\Gamma\left(\alpha\right)}x^{\alpha-1}e^{-x/\beta},\ x>0,\alpha>0,\beta>0.$$

By equation (2.35) in Proposition 2.1.1 and equation (2.32),

$$\begin{aligned} E\left[\ln x\right] = E\left[F_1\left(x\right)\right] &= \partial_\alpha \psi\left(\xi\right) \\ &= \frac{\partial}{\partial \alpha}\left[\alpha \ln \beta + \ln \Gamma\left(\alpha\right)\right] \\ &= \ln \beta + \frac{\Gamma'\left(\alpha\right)}{\Gamma\left(\alpha\right)} \\ &= \ln \beta + \psi\left(\alpha\right), \end{aligned} \tag{3.1}$$

where $\psi\left(\alpha\right)$ is the *digamma function*. Hence, the entropy is

$$\begin{aligned} H\left(\alpha, \beta\right) &= -\int_0^\infty p\left(x\right) \ln p\left(x\right) dx \\ &= -\int_0^\infty p\left(x\right) \left(-\ln\left(\beta^\alpha \Gamma\left(\alpha\right)\right) - \ln x + \alpha \ln x - \frac{x}{\beta}\right) dx \\ &= \ln\left(\beta^\alpha \Gamma\left(\alpha\right)\right) - \left(\alpha - 1\right) \int_0^\infty \ln x \, p\left(x\right) dx + \frac{1}{\beta} \int_0^\infty x \, p\left(x\right) dx \\ &= \alpha \ln \beta + \ln \Gamma\left(\alpha\right) - \left(\alpha - 1\right)\left[\ln \beta + \psi\left(\alpha\right)\right] + \frac{1}{\beta} \alpha \beta \\ &= \alpha + \ln \Gamma\left(\alpha\right) + \left(1 - \alpha\right) \psi\left(\alpha\right) + \ln \beta. \end{aligned}$$

- **Beta Distribution**

The density of a beta distribution is

$$p\left(x; a, b\right) = \frac{1}{B\left(a, b\right)} x^{a-1} \left(1 - x\right)^{b-1}, \; 0 < x < 1, a > 0, b > 0.$$

where $B(a,b)$ is defined by the *Beta function*

$$B(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}\,dt,\ a,b>0.$$

Note that

$$B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

By equation (2.35) in Proposition 2.1.1 and equation (2.34),

$$\begin{aligned}E[\ln x]=E[F_1(x)]&=\partial_a\psi(\xi)\\&=\frac{\partial}{\partial a}\ln B(a,b)\\&=\frac{\partial}{\partial a}[\ln\Gamma(a)+\ln\Gamma(b)-\ln\Gamma(a+b)]\\&=\psi(a)-\psi(a+b),\end{aligned}\tag{3.2}$$

$$\begin{aligned}E[\ln(1-x)]=E[F_2(x)]&=\partial_b\psi(\xi)\\&=\frac{\partial}{\partial b}\ln B(a,b)\\&=\frac{\partial}{\partial b}[\ln\Gamma(a)+\ln\Gamma(b)-\ln\Gamma(a+b)]\\&=\psi(b)-\psi(a+b).\end{aligned}\tag{3.3}$$

The entropy is

$$\begin{aligned}
H\left(a,b\right) &= -\int_0^\infty p\left(x\right)\ln p\left(x\right)dx \\
&= -\int_0^\infty p\left(x\right)\left[-\ln B\left(a,b\right) - \ln x - \ln\left(1-x\right) + a\ln x + b\ln\left(1-x\right)\right]dx \\
&= \ln B\left(a,b\right) + \left(1-a\right)\int_0^\infty p\left(x\right)\ln x\,dx + \left(1-b\right)\int_0^\infty p\left(x\right)\ln\left(1-x\right)dx \\
&= \ln B\left(a,b\right) + \left(1-a\right)\left[\psi\left(a\right) - \psi\left(a+b\right)\right] + \left(1-b\right)\left[\psi\left(b\right) - \psi\left(a+b\right)\right] \\
&= \ln B\left(a,b\right) + \left(1-a\right)\psi\left(a\right) + \left(1-b\right)\psi\left(b\right) + \left(a+b-2\right)\psi\left(a+b\right). \qquad (3.4)
\end{aligned}$$

3.2 Maximum Distributions

In this section, we deal with the problem of finding the density p of maximum entropy subject to the first N moment constraints. Given the numbers $m_1, m_2, \ldots, m_N$, we are interested in finding the distribution p that maximizes the following entropy functional with Lagrange multipliers

$$J\left(p\right) = -\int_{\mathcal{X}} p\left(x\right)\ln p\left(x\right)dx + \sum_{j=0}^{N}\lambda_j\left(\int_{\mathcal{X}} x^j p\left(x\right)dx - m_j\right), \qquad (3.5)$$

where we choose $m_0 = 1$ for convenience. By

$$\frac{\partial J\left(p\right)}{\partial p} = -\ln p\left(x\right) - 1 + \sum_{j=0}^{N}\lambda_j x^j = 0,$$

we obtain that the maximum entropy distribution belongs to the following exponential family

$$p\left(x\right)=e^{-1+\lambda_0+\lambda_1 x+\lambda_2 x^2+\ldots+\lambda_N x^N}=Ce^{\lambda_1 x+\lambda_2 x^2+\ldots+\lambda_N x^N} \tag{3.6}$$

with $C=e^{-1+\lambda_0}$ and Lagrange multipliers λ_j determined by the moment constraints.

First of all, we show that the existence of the maximum entropy distribution for any number of constraints, i.e. among all distributions $q\left(x\right)$ that satisfy the moment constraints

$$\int_{\mathcal{X}} x^j p\left(x\right)dx=m_j,\ \forall j=0,1,\ldots,N$$

where $m_0=1$, the distribution given by equation (3.6) reaches the maximum entropy. By the non-negativity and non-degeneracy of the Kullback-Leibler relative entropy, we have that the arbitrary entropy $H\left(q\right)$ is less than or equal to $H\left(p\right)$,

$$\begin{aligned}
H\left(q\right)&=-\int_{\mathcal{X}} q\ln q=-\int_{\mathcal{X}} q\ln\left(\frac{q}{p}p\right)=-\int_{\mathcal{X}} q\ln\frac{q}{p}-\int_{\mathcal{X}} q\ln p\\
&=-D_{KL}\left(q||p\right)-\int_{\mathcal{X}} q\ln p\leq-\int_{\mathcal{X}} q\ln p\\
&=-\int_{\mathcal{X}} q\left(x\right)\left(-1+\lambda_0+\lambda_1 x+\lambda_2 x^2+\ldots+\lambda_N x^N\right)dx\\
&=-\left(-1+\lambda_0+\lambda_1 m_1+\lambda_2 m_2+\ldots+\lambda_N m_N\right)\\
&=-\int_{\mathcal{X}} p\left(x\right)\left(-1+\lambda_0+\lambda_1 x+\lambda_2 x^2+\ldots+\lambda_N x^N\right)dx\\
&=-\int_{\mathcal{X}} p\ln p=H\left(p\right),
\end{aligned}$$

with equality when $p=q$.

However, the uniqueness of the maximum entropy distribution is a complicated problem for general N. For the cases $N \leq 2$, we will show the uniqueness by showing the uniqueness of the Lagrange multipliers λ_j satisfying the given constraints.

- **Case $N = 0$: Constraint-Free Distribution**

Proposition 3.2.1. *[4] Among all distributions defined on the finite interval (a, b), the one with maximum entropy is the uniform distribution.*

Proof. In this case, the distribution given by equation (3.6) is

$$p(x) = e^{-1+\lambda_0} = C.$$

By the only constraint

$$\int_a^b p(x)\, dx = m_0 = 1,$$

we have

$$\int_a^b p(x)\, dx = e^{-1+\lambda_0}(b-a) = 1,$$

$$\lambda_0 = 1 - \ln(b-a).$$

Hence, the uniform distribution is the unique constraint-free distribution with maximum entropy. □

- **Case $N = 1$: Matching the Mean**

Proposition 3.2.2. *[4] Among all distributions defined on $(0, \infty)$, with given positive mean μ, the one with maximum entropy is the exponential distribution $p(x) = \frac{1}{\mu}e^{-\frac{x}{\mu}}$.*

Proof. In this case, equation (3.5) becomes

$$J\left(p\right)=-\int_{0}^{\infty}p\left(x\right)\ln p\left(x\right)dx+\lambda_{1}\left(\int_{0}^{\infty}xp\left(x\right)dx-\mu\right)+\lambda_{0}\left(\int_{0}^{\infty}p\left(x\right)dx-1\right). \tag{3.7}$$

The maximum entropy density given by equation (3.6) is

$$p\left(x\right)=Ce^{\lambda_{1}x},\ C=e^{\lambda_{0}-1},$$

with the constants C and λ_1 to be determined from the constraints

$$\int_{0}^{\infty}p\left(x\right)dx=\int_{0}^{\infty}Ce^{\lambda_{1}x}dx=1,\ \int_{0}^{\infty}xp\left(x\right)dx=\int_{0}^{\infty}xCe^{\lambda_{1}x}dx=\mu.$$

We obtain $C=\frac{1}{\mu}$, $\lambda_1=-\frac{1}{\mu}$. Hence, the maximum entropy distribution is the exponential distribution with the parameter $\frac{1}{\mu}$. □

- **Case $N=2$: Matching Mean and Variance**

Proposition 3.2.3. *[4] Among all distributions defined on $\mathbb{R}$, with given positive mean μ and variance σ^2, the one with maximum entropy is the normal distribution $p\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.*

Proof. In this case, the constraints are

$$\int_{-\infty}^{\infty}p\left(x\right)dx=1,\ \int_{-\infty}^{\infty}xp\left(x\right)dx=\mu,\ \int_{-\infty}^{\infty}\left(x-\mu\right)^{2}p\left(x\right)dx=\sigma^{2}. \tag{3.8}$$

We write

$$\begin{aligned} J(p) = &-\int_{-\infty}^{\infty} p(x)\ln p(x)\,dx - \gamma\left(\int_{-\infty}^{\infty}(x-\mu)^2 p(x)\,dx - \sigma^2\right) \\ &+\beta\left(\int_{-\infty}^{\infty} xp(x)\,dx - \mu\right) + \alpha\left(\int_{-\infty}^{\infty} p(x)\,dx - 1\right) \\ = &\int_{-\infty}^{\infty}[-p\ln p - \gamma(x-\mu)^2 p + \beta(x-\mu)p + \alpha p + \gamma\sigma^2 - \alpha]dx. \end{aligned} \tag{3.9}$$

The Euler-Lagrange equation can be written as

$$\begin{aligned} &\frac{\partial}{\partial p}\left[-p\ln p - \gamma(x-\mu)^2 p + \beta(x-\mu)p + \alpha p + \gamma\sigma^2 - \alpha\right] = 0 \\ &\Leftrightarrow -\ln p - 1 - \gamma(x-\mu)^2 + \beta(x-\mu) + \alpha = 0 \\ &\Leftrightarrow -\gamma(x-\mu)^2 + \beta(x-\mu) + \alpha - 1 = \ln p. \end{aligned}$$

Hence, the distribution takes the form

$$p(x) = Ce^{-\gamma(x-\mu)^2+\beta(x-\mu)}, \tag{3.10}$$

where $C = e^{\alpha-1}$.

Hence, by the constraints given in equation (3.8), we have

$$\begin{aligned} 1 = \int_{-\infty}^{\infty} p(x)\,dx &= \int_{-\infty}^{\infty} e^{\alpha-1}e^{-\gamma(x-\mu)^2+\beta(x-\mu)}dx \\ &= e^{\alpha-1}\int_{-\infty}^{\infty} e^{-\gamma x^2+\beta x}dx \\ &= \sqrt{\frac{\pi}{\gamma}}e^{\frac{\beta^2}{4\gamma}+\alpha-1}, \end{aligned} \tag{3.11}$$

$$\begin{aligned}0 = \int_{-\infty}^{\infty} xp\left(x\right)dx - \mu = e^{\alpha-1}\int_{-\infty}^{\infty}\left(x-\mu\right)e^{-\gamma(x-\mu)^2+\beta(x-\mu)}dx\\ = e^{\alpha-1}\int_{-\infty}^{\infty} xe^{-\gamma x^2+\beta x}dx\\ = \sqrt{\frac{\pi}{\gamma}}\left(\frac{\beta}{2\gamma}\right)e^{\frac{\beta^2}{4\gamma}+\alpha-1},\end{aligned} \tag{3.12}$$

$$\begin{aligned}\sigma^2 = \int_{-\infty}^{\infty}\left(x-\mu\right)^2 p\left(x\right)dx = e^{\alpha-1}\int_{-\infty}^{\infty}\left(x-\mu\right)^2 e^{-\gamma(x-\mu)^2+\beta(x-\mu)}dx\\ = e^{\alpha-1}\int_{-\infty}^{\infty} x^2e^{-\gamma x^2+\beta x}dx\\ = \sqrt{\frac{\pi}{\gamma}}\frac{1}{2\gamma}\left(1+\frac{\beta^2}{2\gamma}\right)e^{\frac{\beta^2}{4\gamma}+\alpha-1}.\end{aligned} \tag{3.13}$$

By equation (3.12), $\beta = 0$. Substituting $\beta = 0$ into equations (3.11) and (3.13) yields $\sqrt{\frac{\pi}{\gamma}} = e^{1-\alpha}$, $\sigma^2 = \frac{1}{2\gamma}$. Therefore, by equation (3.10), we have

$$\begin{aligned}p\left(x\right) &= Ce^{-\gamma(x-\mu)^2+\beta(x-\mu)}\\ &= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}.\end{aligned}$$

□

3.3 Kullback-Leibler Relative Entropy

In Section 1.3.2, we gave the definition of Kullback-Leibler relative entropy on a statistical model. We shall compute Kullback-Leibler relative entropy for pairs of densities in the same class of the exponential family.

- **Poisson Distribution**

Consider Poisson distributions

$$p\left(x;\lambda\right)=e^{-\lambda}\frac{\lambda^{x}}{x!},\ q\left(x;\xi\right)=e^{-\xi}\frac{\xi^{x}}{x!}.$$

We have

$$\begin{aligned}D_{KL}\left(p||q\right)&=\sum_{x=0}^{\infty}p\left(x\right)\ln\frac{p\left(x\right)}{q\left(x\right)}\\&=\sum_{x=0}^{\infty}p\left(x\right)\left[\left(\xi-\lambda\right)+x\ln\frac{\lambda}{\xi}\right]\\&=\xi-\lambda+\ln\frac{\lambda}{\xi}\sum_{x=0}^{\infty}xp\left(x\right)\\&=\xi-\lambda-\lambda\ln\frac{\xi}{\lambda}\\&=\lambda\left(-\ln\frac{\xi}{\lambda}+\frac{\xi}{\lambda}-1\right)\end{aligned}$$

Hence

$$D_{KL}\left(p||q\right)=\lambda f\left(\frac{\xi}{\lambda}\right),$$

with $f\left(x\right)=-\ln x+x-1\geq 0$. This verifies $D_{KL}\left(p||q\right)\geq 0$, the equality being reached if and only if $\frac{\xi}{\lambda}=1$, i.e. if $\xi=\lambda$.

- **Normal Distribution**

For two normal densities

$$p_1\left(x\right)=\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{\left(x-\mu_1\right)^2}{2\sigma_1^2}},\ p_2\left(x\right)=\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{\left(x-\mu_2\right)^2}{2\sigma_2^2}},$$

we have

$$\begin{aligned}
D_{KL}\left(p_1||p_2\right) &= \int_{-\infty}^{\infty} p_1\left(x\right)\ln\frac{p_1\left(x\right)}{p_2\left(x\right)}dx \\
&= \int_{-\infty}^{\infty} p_1\left(x\right)\left[\ln\frac{\sigma_2}{\sigma_1} - \frac{\left(x-\mu_1\right)^2}{2\sigma_1^2} + \frac{\left(x-\mu_2\right)^2}{2\sigma_2^2}\right]dx \\
&= \ln\frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{1}{2\sigma_2^2}\int_{-\infty}^{\infty}\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{\left(x-\mu_1\right)^2}{2\sigma_1^2}}\left[\left(x-\mu_1\right)+\left(\mu_1-\mu_2\right)\right]^2 dx \\
&= \ln\frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{1}{2\sigma_2^2}\left[\sigma_1^2 + 0 + \left(\mu_1-\mu_2\right)^2\right] \\
&= \frac{1}{2}\left[-\ln\left(\frac{\sigma_1}{\sigma_2}\right)^2 + \left(\frac{\sigma_1}{\sigma_2}\right)^2 - 1\right] + \frac{\left(\mu_1-\mu_2\right)^2}{2\sigma_2^2} \qquad (3.14) \\
&\geq \frac{\left(\mu_1-\mu_2\right)^2}{2\sigma_2^2}
\end{aligned}$$

with equality being reached for $\frac{\sigma_1}{\sigma_2} = 1$, i.e. $\sigma_1 = \sigma_2$.

- **Multivariate Normal Distribution**

Consider two multivariate normal distributions

$$p_1\left(x;\mu_1,A_1\right) = \frac{1}{\left(2\pi\right)^{n/2}\left(\det A_1\right)^{1/2}}e^{-\frac{1}{2}\left(x-\mu_1\right)^t A_1^{-1}\left(x-\mu_1\right)},$$

$$p_2\left(x;\mu_2,A_2\right) = \frac{1}{\left(2\pi\right)^{n/2}\left(\det A_2\right)^{1/2}}e^{-\frac{1}{2}\left(x-\mu_2\right)^t A_2^{-1}\left(x-\mu_2\right)}.$$

Applying the following equalities,

$$
\begin{aligned}
E\left[\mathrm{tr} A\right] &= \mathrm{tr} E\left[A\right], \\
\mathrm{tr}\left(ABC\right) &= \mathrm{tr}\left(BCA\right) = \mathrm{tr}\left(CAB\right), \\
xx^t &= \left(x-\mu\right)\left(x-\mu\right)^t + \mu\left(x-\mu\right)^t + \left(x-\mu\right)\mu^t + \mu\mu^t,
\end{aligned}
$$

where A, B, C are matrices, we have

$$
\begin{aligned}
D_{KL}\left(p_1||p_2\right) &= E_{p_1}\left[\ln\frac{p_1\left(x\right)}{p_2\left(x\right)}\right] \\
&= \frac{1}{2}E_{p_1}\left[\ln\frac{\det A_2}{\det A_1} - \left(x-\mu_1\right)^t A_1^{-1}\left(x-\mu_1\right) + \left(x-\mu_2\right)^t A_2^{-1}\left(x-\mu_2\right)\right] \\
&= \frac{1}{2}\ln\frac{\det A_2}{\det A_1} + \frac{1}{2}E_{p_1}\left[-\mathrm{tr}\left(A_1^{-1}\left(x-\mu_1\right)\left(x-\mu_1\right)^t\right) + \mathrm{tr}\left(A_2^{-1}\left(x-\mu_2\right)\left(x-\mu_2\right)^t\right)\right] \\
&= \frac{1}{2}\ln\frac{\det A_2}{\det A_1} + \frac{1}{2}E_{p_1}\left[-\mathrm{tr}\left(A_1^{-1}A_1\right) + \mathrm{tr}\left(A_2^{-1}\left(xx^t - 2x\mu_2^t + \mu_2\mu_2^t\right)\right)\right] \\
&= \frac{1}{2}\ln\frac{\det A_2}{\det A_1} - \frac{1}{2}n + \frac{1}{2}\mathrm{tr}\left(A_2^{-1}\left(A_1 + \mu_1\mu_1^t - 2\mu_2\mu_1^t + \mu_2\mu_2^t\right)\right) = \\
&= \frac{1}{2}\left[\ln\frac{\det A_2}{\det A_1} - n + \mathrm{tr}\left(A_2^{-1}A_1\right) + \mathrm{tr}\left(\mu_1^t A_2^{-1}\mu_1 - 2\mu_1^t A_2^{-1}\mu_2 + \mu_2^t A_2^{-1}\mu_2\right)\right] \\
&= \frac{1}{2}\left[\ln\frac{\det A_2}{\det A_1} - n + \mathrm{tr}\left(A_2^{-1}A_1\right) + \left(\mu_2-\mu_1\right)^t A_2^{-1}\left(\mu_2-\mu_1\right)\right], \quad (3.15)
\end{aligned}
$$

which, when $n = 1$, reduces to the univariate case in equation (3.14).

- **Exponential Distribution**

For two exponential densities

$$p\left(x\right) = \xi e^{-\xi x},\ q\left(x\right) = \eta e^{-\eta x},$$

we have

$$\begin{aligned} D_{KL}(p||q) &= \int_0^\infty p(x)\ln\frac{p(x)}{q(x)}dx \\ &= \int_0^\infty \xi e^{-\xi x}\ln\frac{\xi e^{-\xi x}}{\eta e^{-\eta x}}dx \\ &= \int_0^\infty \xi e^{-\xi x}\left[\ln\frac{\xi}{\eta}+(\eta-\xi)x\right]dx \\ &= \ln\frac{\xi}{\eta}+\frac{\eta}{\xi}-1. \end{aligned}$$

Hence

$$D_{KL}(p||q) = f\left(\frac{\eta}{\xi}\right),$$

with $f(x) = -\ln x + x - 1 \geq 0$. This verifies $D_{KL}(p||q) \geq 0$, the equality being reached if and only if $\frac{\eta}{\xi} = 1$, i.e. if $p = q$.

- **Gamma Distribution**

Consider two gamma distributions

$$p(x;a,b) = \frac{1}{b^a\Gamma(a)}x^{a-1}e^{-x/b},$$

$$q(x;c,d) = \frac{1}{d^c\Gamma(c)}x^{c-1}e^{-x/d}.$$

By applying the result in equation (3.1), we have

$$\begin{aligned}
D_{KL}\left(p||q\right) &= \int_0^\infty p\left(x\right)\ln\frac{p\left(x\right)}{q\left(x\right)}dx \\
&= \int_0^\infty p\left(x\right)\left[\ln\frac{d^c\Gamma\left(c\right)}{b^a\Gamma\left(a\right)} + \left(a-c\right)\ln x + \left(\frac{1}{d}-\frac{1}{b}\right)x\right]dx \\
&= \ln\frac{d^c\Gamma\left(c\right)}{b^a\Gamma\left(a\right)} + \left(a-c\right)\left(\ln b + \psi\left(a\right)\right) + \left(\frac{1}{d}-\frac{1}{b}\right)ab \\
&= \left(a-c\right)\psi\left(a\right) - \ln\Gamma\left(a\right) + \ln\Gamma\left(c\right) + c\ln\frac{d}{b} + \frac{a\left(b-d\right)}{d}, \qquad (3.16)
\end{aligned}$$

where $\psi\left(a\right)$ denotes the digamma function.

- **Beta Distribution**

Consider two gamma distributions

$$\begin{aligned}
p\left(x;a,b\right) &= \frac{1}{B\left(a,b\right)}x^{a-1}\left(1-x\right)^{b-1}, \\
q\left(x;c,d\right) &= \frac{1}{B\left(c,d\right)}x^{c-1}\left(1-x\right)^{d-1}.
\end{aligned}$$

By applying the results in equations (3.2) and (3.3), we have

$$\begin{aligned}
D_{KL}\left(p||q\right) &= \int_0^\infty p\left(x\right)\ln\frac{p\left(x\right)}{q\left(x\right)}dx \\
&= \int_0^\infty p\left(x\right)\left[\ln\frac{B\left(c,d\right)}{B\left(a,b\right)} + \left(a-c\right)\ln x + \left(b-d\right)\ln\left(1-x\right)\right]dx \\
&= \ln\frac{B\left(c,d\right)}{B\left(a,b\right)} + \left(a-c\right)\left(\psi\left(a\right)-\psi\left(a+b\right)\right) + \left(b-d\right)\left(\psi\left(b\right)-\psi\left(a+b\right)\right) \\
&= \ln\frac{B\left(c,d\right)}{B\left(a,b\right)} + \left(a-c\right)\psi\left(a\right) + \left(b-d\right)\psi\left(b\right) \\
&\quad + \left(-a-b+c+d\right)\psi\left(a+b\right). \qquad (3.17)
\end{aligned}$$

3.4 Informational Energy

In Section 1.3.3, we gave the definition of information energy on a statistical model. Here, we shall compute the information energy for a few distributions in the exponential family.

- **Poisson Distribution**

Consider the Poisson distribution given by

$$p\left(n;\xi\right)=e^{-\xi}\frac{\xi^{n}}{n!}.$$

We have

$$\begin{aligned} I\left(\xi\right) &= \sum_{n=0}^{\infty} p^{2}\left(n,\xi\right) \\ &= e^{-2\xi}\sum_{n=0}^{\infty}\frac{\xi^{2n}}{\left(n!\right)^{2}} \\ &= e^{-2\xi} I_{0}\left(2\xi\right), \end{aligned} \tag{3.18}$$

where

$$I_{0}\left(z\right)=\sum_{n=0}^{\infty}\frac{\left(z/2\right)^{2n}}{\left(n!\right)^{2}}$$

is the modified Bessel function of order 0. The informational energy decreases to 0 as $\xi \to \infty$. Hence, $I\left(\xi\right) < I\left(0\right) = 1$ for any $\xi > 0$.

- **Normal Distribution**

Consider the normal density

$$p\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(x-\mu\right)^{2}}{2\sigma^{2}}}.$$

We have

$$\begin{aligned}I\left(\mu,\sigma\right)&=\int_{\mathbb{R}}p^{2}\left(x;\mu,\sigma\right)dx\\&=\frac{1}{2\pi\sigma^{2}}\int_{\mathbb{R}}e^{-\frac{\left(x-\mu\right)^{2}}{\sigma^{2}}}dx\\&=\frac{1}{2\pi\sigma^{2}}\sigma\sqrt{\pi}\\&=\frac{1}{2\sqrt{\pi}\sigma}.\end{aligned}\tag{3.19}$$

Note that the information energy does not depend on the mean μ and decreases by increasing the standard deviation σ.

- **Lognormal Distribution**

Consider the lognormal distribution given by

$$p\left(x;\mu,\sigma\right)=\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{\left(\ln x-\mu\right)^{2}}{2\sigma^{2}}}.$$

Using the substitution $y = \ln x - \mu$, we have

$$\begin{aligned} I(\mu, \sigma) &= \int_0^\infty p^2(x; \mu, \sigma)\, dx \\ &= \frac{1}{2\pi\sigma^2} \int_0^\infty \frac{1}{x^2} e^{-\frac{(\ln x - \mu)^2}{\sigma^2}} dx \\ &= \frac{1}{2\pi\sigma^2} \int_{-\infty}^\infty e^{-\frac{y^2}{\sigma^2} - 2y - 2\mu} dy \\ &= \frac{1}{2\pi\sigma^2} \sqrt{\pi\sigma^2} e^{\sigma^2 - 2\mu} \\ &= \frac{1}{2\sqrt{\pi}\sigma} e^{\sigma^2 - 2\mu}. \end{aligned} \tag{3.20}$$

It follows that, being different from the case for a normal distribution, the information energy of a lognormal distribution is dependent on μ.

- **Exponential Distribution**

Consider the exponential density

$$p(x, \xi) = \xi e^{-\xi x}.$$

We have

$$\begin{aligned} I(\xi) &= \int_0^\infty p^2(x, \xi)\, dx \\ &= \int_0^\infty \xi^2 e^{-2\xi x} dx \\ &= \frac{\xi}{2}. \end{aligned} \tag{3.21}$$

Hence, the information energy of an exponential distribution is linear in ξ.

- **Gamma Distribution**

Consider the gamma distribution given by

$$p\left(x;\alpha,\beta\right)=\frac{1}{\beta^{\alpha}\Gamma\left(\alpha\right)}x^{\alpha-1}e^{-x/\beta}.$$

Assuming $\alpha>1/2$ and using the substitution $a=2\alpha-1$ and $b=\beta/2$, the information energy becomes

$$\begin{aligned}I\left(\alpha,\beta\right)&=\int_0^{\infty}p^2\left(x;\mu,\sigma\right)dx\\&=\frac{1}{\beta^{2\alpha}\Gamma\left(\alpha\right)^2}\int_0^{\infty}x^{2\alpha-2}e^{-2x/\beta}dx\\&=\frac{1}{\beta^{2\alpha}\Gamma\left(\alpha\right)^2}b^{a}\Gamma\left(a\right)\int_0^{\infty}\frac{1}{b^{a}\Gamma\left(a\right)}x^{a-1}e^{-x/b}dx\\&=\frac{1}{\beta^{2\alpha}\Gamma\left(\alpha\right)^2}\left(\frac{\beta}{2}\right)^{2\alpha-1}\Gamma\left(2\alpha-1\right)\\&=\frac{\Gamma\left(2\alpha-1\right)}{2^{2\alpha-1}\beta\Gamma\left(\alpha\right)^2}.\end{aligned}\tag{3.22}$$

The case $\alpha\leq 1/2$ is eliminated by the divergence of the improper integral.

By the Legendre's duplication formula

$$\Gamma\left(2\alpha\right)=\frac{2^{2\alpha-1}}{\sqrt{\pi}}\Gamma\left(\alpha\right)\Gamma\left(\alpha+\frac{1}{2}\right),\tag{3.23}$$

equation (3.22) becomes

$$
\begin{aligned}
I(\alpha, \beta) &= \frac{\Gamma(2\alpha - 1)}{2^{2\alpha-1}\beta\Gamma(\alpha)^2} \\
&= \frac{\Gamma\left(\alpha + \frac{1}{2}\right)}{\sqrt{\pi}(2\alpha - 1)\beta\Gamma(\alpha)} \\
&= \frac{\Gamma\left(\alpha + \frac{1}{2}\right)}{\Gamma(1/2)(2\alpha - 1)\beta\Gamma(\alpha)} \\
&= \frac{1}{\beta(2\alpha - 1)B(\alpha, 1/2)}. \qquad (3.24)
\end{aligned}
$$

- **Beta Distribution**

Consider the gamma distribution given by

$$
p(x; a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1}.
$$

Assuming $a > 1/2$, $b > 1/2$ and using the substitution $\alpha = 2a - 1$ and $\beta = 2b - 1$, we have

$$
\begin{aligned}
I(a, b) &= \int_0^1 p^2(x; \mu, \sigma)\, dx \\
&= \frac{1}{B^2(a, b)} \int_0^1 x^{2a-2}(1-x)^{2b-2}\, dx \\
&= \frac{1}{B^2(a, b)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx \\
&= \frac{B(\alpha, \beta)}{B^2(a, b)} = \frac{B(2a-1, 2b-1)}{B^2(a, b)}. \qquad (3.25)
\end{aligned}
$$

The case $a \leq 1/2$ or $b \leq 1/2$ is eliminated by the divergence of the improper integral.

By equation (3.23), we have

$$\begin{aligned}
I\left(a,b\right) &= \frac{B\left(2a-1,2b-1\right)}{B^{2}\left(a,b\right)} \\
&= \frac{\Gamma\left(2a\right)\Gamma\left(2b\right)\Gamma\left(a+b\right)^{2}}{\left(2a-1\right)\left(2b-1\right)\Gamma\left(2a+2b-2\right)\Gamma\left(a\right)^{2}\Gamma\left(b\right)^{2}} \\
&= \frac{\left(2a+2b-1\right)\left(2a+2b-2\right)}{\left(2a-1\right)\left(2b-1\right)}\frac{\Gamma\left(2a\right)\Gamma\left(2b\right)\Gamma\left(a+b\right)^{2}}{\Gamma\left(2a+2b\right)\Gamma\left(a\right)^{2}\Gamma\left(b\right)^{2}} \\
&= \frac{\left(a+b-1/2\right)\left(a+b-1\right)}{\left(a-1/2\right)\left(b-1/2\right)}.\frac{\Gamma\left(a+1/2\right)\Gamma\left(b+1/2\right)\Gamma\left(a+b\right)}{2\sqrt{\pi}\Gamma\left(a+b+1/2\right)\Gamma\left(a\right)\Gamma\left(b\right)}. \qquad (3.26)
\end{aligned}$$

Bibliography

[1] Shun-ichi Amari, *Differential-geometrical methods in statistics. Vol. 28*, Springer Science & Business Media, 1985.

[2] Shun-ichi Amari, *Information Geometry and Its Applications*, Springer, 2016.

[3] Nihat Ay, Jürgen Jost, Hông Vân Lê, Lorenz Schwachhöfer, *Information Geometry*, Springer, 2017.

[4] Ovidiu Calin; Constantin Udrişte, *Geometric Modeling in Probability and Statistics*, Springer, 2014.

[5] Jan R. Magnus; Heinz Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, 1999.

[6] C. Radhakrishna Rao, *Information and the accuracy attainable in the estimation of statistical parameters.*, Bulletin of the Calcutta Mathematical Society, 37 (3), 1945, 81-91.