E-variables for Exponential Families

Long, L.

E-variables for Exponential Families

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE
in
MATHEMATICS

Author :	Long Long
Student ID :	s2583267
Supervisor :	Peter Grünwald
Second corrector :	Valentina Masarotto

Leiden, The Netherlands,

E-variables for Exponential Families

Long Long

Leiden, The Netherlands

Abstract

E-variables are a novel tool for constructing hypothesis tests that retain Type-I error guarantees when the sampling plan is not determined in advance, i.e. under optional stopping and optional continuation. We construct E-variables for null hypotheses that are univariate exponential families and point alternative hypotheses by calculating the *Reverse Information Projection*, abbreviated to RIPr, of the alternative on the set of mixtures over the null. We focus on RIPr's that are simple; this means that they coincide with a single element of the null hypothesis rather than a mixture of such elements. We find that there is no unique simple way to determine the RIPr for the whole class of exponential families. We give conditions under which the RIPr is simple (and then also easy to calculate), and conditions under which it is not (and then it is hard to calculate), and we give several examples of each case. For the case that an E-variable for a specific exponential family null is given, we establish E-variables for other exponential families by 1-to-1 transformations of random variables. We approximate a more complex RIPr (i.e. a mixture of exponential distributions) when the sample space consists of two outcomes of the exponential distribution in a specific setting by programming in R.

Contents

Chapter 1

Introduction

The *exponential families* (Altun et al., 2012, Banerjee, 2007, Brown, 1986) are a class of probability models that include the Bernoulli, binomial (with known number of trials n), Poisson, negative binomial (with known number of failures r), exponential, Weibull (with known shape k), normal, gamma, multinomial, and many other well-known sets of distributions. They are widely used because their general form, introduced in Definition 1, makes it easy to compute important quantities such the maximum likelihood estimator, the mean, variance, Fisher information, relative entropy, and so on. Brown (1986) and later Nielsen and Garcia (2009) listed properties (probability density function, maximum likelihood estimator, dual parameterizations: natural and expectation parameters, and so on) of some common exponential family distributions.

E-variables (Grünwald et al., 2020, Shafer, 2019, Vovk and Wang, 2021) have been proposed in recent years as an alternative to the p-value. Hypothesis testing using E-variables is 'safer' than traditional hypothesis testing using p-values: it guarantees Type-I error under optional continuation and optional stopping. It has by now been applied in several classic testing scenarios, such as contingency tables (Turner et al., 2021) and the logrank test (ter Schure et al., 2020).

In this thesis, we determine E-variables for exponential families in hypothesis testing. We consider a composite null hypothesis $\mathcal{H}_0$ that is a set of single parameter exponential families (i.e. $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$) together with a simple alternative hypothesis $\mathcal{H}_1$ (i.e. $\mathcal{H}_1 = \{P_1\}$). Grünwald et al. (2020) argued that there is a special distribution W_0^* on Θ_0 such that $p_1 / p_{W_0^*}$ is an E-variable where $p_{W_0^*}$ is the Bayes marginal distribution based on prior W_0^*. This W_0^* coincides with what has been called the *Reverse Information Projection (RIPr)* in the literature (Grünwald et al., 2020,

Li, 1999). We construct E-variables based on this RIPr.

In the remainder of this introduction, we formally introduce exponential families, the concept of E-variable and we provide examples of E-variables for some different types of hypotheses. We end by giving an overview of the thesis and the main results.

1.1 Exponential Family

In this section, we state the definition and some well-known properties of exponential families.

Definition 1 (Exponential Family)**.** *Let X be a random variable with sample space $\mathcal{X}$. Let $\mathcal{P}$ be a family of distributions for $\mathcal{X}$. We say that $\mathcal{P}$ is a d-dimensional exponential family if the probability density of every element of $\mathcal{P}$ can be written in the following* canonical *form:*

$$p_{\boldsymbol{\eta}}(x) = h(x) \exp(\boldsymbol{\eta}^T \boldsymbol{T}(x) - A(\boldsymbol{\eta}))$$

with $h(x)$ a function from $\mathcal{X}$ to $\mathbb{R}_0^+$, 'canonical' *parameter vector $\boldsymbol{\eta} \in \mathbb{R}^d$,* 'sufficient statistic' *vector $\boldsymbol{T} : \mathcal{X} \to \mathbb{R}^d$, and* 'log-partition function' *$A(\boldsymbol{\eta})$.*

$h(x)$ is the density of a measure (which may not be a probability measure, i.e. not integrate to 1) relative to Lebesgue measure (in continuous distributions) or counting measure (in discrete distributions). The statistic $\boldsymbol{T}(x)$ is called "sufficient" which intuitively means that we get the same (and no less) information about the unknown parameter $\boldsymbol{\eta}$ from observing the value of statistic $\boldsymbol{T}(x)$ than from observing the full x. If $\boldsymbol{T}$ is the identity and $\mathcal{X}$ is a subset of $\mathbb{R}^d$, we call the family *natural* and $\boldsymbol{\eta}$ a *natural* parameter. Since the integral or sum of $p_{\boldsymbol{\eta}}(x)$ must be equal to 1, we have that the log-partition function, also known as log-normalizer, is given by, in continuous distributions

$$A(\boldsymbol{\eta}) = \log \int_{x \in \mathcal{X}} h(x) \exp(\boldsymbol{\eta}^T \boldsymbol{T}(x)) dx,$$

or, in discrete distributions,

$$A(\boldsymbol{\eta}) = \log \sum_{x \in \mathcal{X}} h(x) \exp(\boldsymbol{\eta}^T \boldsymbol{T}(x)).$$

The parameter space in the canonical form is given by
$\boldsymbol{\Theta}_{\boldsymbol{\eta}} = \{\boldsymbol{\eta} : A(\boldsymbol{\eta}) < \infty\}$. This is the set of parameters for which the distribution of X is well-defined.

We classify exponential families into univariate and multivariate based on the dimension of sample space $\mathcal{X}$. In this thesis, we only consider univariate exponential families. The dimension of the natural parameter space $\mathbf{\Theta}_\eta$ determines the dimension of the exponential family.

The most commonly used parameterization of any specific exponential family is called its standard parameterization or standard *form*. It has a number of parameters with particular meanings, like mean, variance, location parameter, scale parameter, shape parameter and so on. For example, the probability density function of the Poisson distribution in the standard form is $\frac{\lambda^x \exp(-\lambda)}{x!}$ where parameter λ represents both the mean and the variance. In order to distinguish the canonical form (parameterization) from the standard form, we add a small circle to the standard form (eg: $p^\circ_\theta(x), A^\circ(\theta)$ in the standard form vs. $p_\eta(x), A(\eta)$ in the canonical form).

Example 1 (Bernoulli Distribution)**.** The Bernoulli distribution, also known as the $0-1$ distribution, is a univariate exponential family of dimension 1. It is a type of discrete distribution. Outcome 1 means the Bernoulli trial succeeded. 0 means the trial failed. In the standard parameterization, the success probability of the trial is $\theta \in [0,1]$. We have probability mass function

$$p^\circ_\theta(x) = \theta^x \cdot (1-\theta)^{(1-x)}, \qquad x = 0 \text{ or } 1.$$

We transform it to the canonical form. We obtain $h(x) = 1$. The natural parameter is $\eta = \log \frac{\theta}{1-\theta}$. The sufficient statistic is $T(x) = x$. The log-partition function is $A(\eta) = -\log(1-\theta)$.

Example 2 (Normal Distribution)**.** The normal distribution, also known as Gaussian, Gauss, or Laplace-Gauss distribution, is a univariate exponential family of dimension 2. It is a type of continuous distribution and is often used in natural and social sciences to represent real-valued random variables. In the standard parameterization, parameter $\mu \in \mathbb{R}$ is the mean of the distribution, which determines the location of the density curve, i.e. the curve is symmetric around $x = \mu$. Parameter $\sigma^2 > 0$ is the variance of the distribution, which determines the scale of the density curve, i.e. the smaller the variance, the more concentrated the curve. We have probability density function

$$p^\circ_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}.$$

We transform it to the canonical form. We have $h(x) = \frac{1}{\sqrt{2\pi}}$, natural pa-

rameter vector $\boldsymbol{\eta} = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}$, sufficient statistic vector $\boldsymbol{T}(\boldsymbol{x}) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$, and log-partition function $A(\boldsymbol{\eta}) = \frac{\mu^2}{2\sigma^2} + \log \sigma$.

1.2 E-variables, Test-martingale and Safety

This section introduces the central concepts related to E-variables, and provides an important lemma about them. Both the definitions and the lemma are taken from Grünwald et al. (2020), though the definition of 'test martingale' below is just the classical, standard definition of 'nonnegative supermartingale with starting value ≤ 1', which can be found in any advanced probability textbook, e.g. Williams (1991).

Let $\mathcal{H}_0$ be a set of distributions for random process $X_1, X_2, \ldots$. Throughout this thesis, $\mathcal{H}_0$ represents the *null hypothesis*.

Definition 2 (Conditional E-variable)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables defined on sample space Ω. Let $M_1, M_2, \ldots$ be a sequence of non-negative random variables where for all $n \in \mathbb{N}$, M_n is determined by $X_1, \ldots, X_n$, i.e. M_n is a function of $X_1, \ldots, X_n$. We say that M_n is an E-variable for X_n relative to the null hypothesis $\mathcal{H}_0$ conditional on $X_1, \ldots, X_{n-1}$ if*

$$\mathbb{E}_{P_0}[M_n | X_1, \ldots, X_{n-1}] \leq 1 \qquad \text{for all } P_0 \in \mathcal{H}_0.$$

We call a sequence $M_1, M_2, \ldots$ a conditional E-variable process.

Definition 3 (Test Martingale)**.** *If for all $n \in \mathbb{N}$, M_n is an E-variable for X_n conditional on $X_1, \ldots, X_{n-1}$, then the sequence $M^{(1)}, M^{(2)}, \ldots$ with $M^{(n)} = \prod_{i=1}^{n} M_i$ is called a test martingale relative to the null hypothesis $\mathcal{H}_0$.*

Definition 4 ((Unconditional) E-variable)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables defined on sample space Ω. Let $S^{(n)}$ be a non-negative random variable which is determined by $X_1, \ldots, X_n$, i.e. $S^{(n)}$ is a function of $X_1, \ldots, X_n$. We say that $S^{(n)}$ is an (unconditional) E-variable for X^n relative to the null hypothesis $\mathcal{H}_0$ if*

$$\mathbb{E}_{P_0}[S^{(n)}] \leq 1 \qquad \text{for } \forall\ P_0 \in \mathcal{H}_0.$$

An E-variable is called sharp if the above holds with equality for at least one $P_0 \in \mathcal{H}_0$.

We call the *value* that an E-variable takes on a given data sample the *E-value*; some authors use 'E-value' also for 'E-variable', similarly to what is customary for p-values.

Lemma 1 (E-variable Lemma)**.** *Suppose that $M_1, M_2, \ldots$ is a conditional E-variable process, then for all n, $M^{(n)}$ is an E-variable for X^n, i.e.*

$$\mathbb{E}_{P_0}[M^{(n)}] \leq 1 \qquad \text{for } \forall\, P_0 \in \mathcal{H}_0.$$

Safety

We are interested in E-variables and test martingales because type-I error probability bounds can be guaranteed irrespective of the stopping rule used: for any test martingale $\{S^{(i)}\}_{i\in\mathbb{N}}$, Ville's inequality (Grünwald et al., 2020, Shafer, 2019) shows that, for all $0 < \alpha < 1$, $P \in \mathcal{H}_0$,

$$P(\text{there exists } i \text{ such that } S^{(i)} \geq 1/\alpha) \leq \alpha.$$

Thus, if evidence against the null hypothesis $\mathcal{H}_0$ after observing i data units is measured by $S^{(i)}$, and we reject the null hypothesis if $S^{(i)} \geq 1/\alpha$, then our type-I error will be bounded by α, independently of the stopping rule used to determine i. We thus have type-I error control independently of the stopping rule that is used, even if it is externally imposed, or if it is chosen to be as aggressive as possible (keep sampling until $S^{(i)} \geq 1/\alpha$ or time runs out); in contrast, in classical testing based on p-values, the stopping rule must be determined in advance and must be adhered to to get Type-I error control.

Any test which is based on $\{S^{(i)}\}_{i\in\mathbb{N}}$ and a stopping time τ that, after stopping, rejects iff $S^{(\tau)} \geq 1/\alpha$ is called a level α-test that is a safe under optional stopping, or simply a safe test.

1.2.1 Hypotheses

The null hypothesis $\mathcal{H}_0$ and the alternative hypothesis $\mathcal{H}_1$ are both defined as sets of distributions of random process $X_1, X_2, \ldots$ and thus define marginal distributions for vector X^n. We now introduce some important E-variables for X^n for three different types of $\mathcal{H}_0, \mathcal{H}_1$. All members of any $\mathcal{H}_0$ or $\mathcal{H}_1$ mentioned below are thus probability distributions for X^n, and they are assumed to have densities or probability mass functions. For distribution P (or Q) we denote its corresponding density/mass function by p (or q, respectively).

Simple $\mathcal{H}_0$ & Simple $\mathcal{H}_1$

Example 3. Suppose that the null hypothesis $\mathcal{H}_0$ and the alternative hypothesis $\mathcal{H}_1$ are simple, meaning that each hypothesis contains just a sin-

gle distribution for X^n, i.e. $\mathcal{H}_0 = \{P_0\}$ and $\mathcal{H}_1 = \{P_1\}$. The likelihood ratio for n outcomes $\frac{p_1(X^n)}{p_0(X^n)}$ is a sharp E-variable since

$$\mathbb{E}_{p_0}\left[\frac{p_1(X^n)}{p_0(X^n)}\right] = \int p_0(x^n) \cdot \frac{p_1(x^n)}{p_0(x^n)} dx^n = 1,$$

with the integral replaced by a sum in case of probability mass functions.

Simple $\mathcal{H}_0$ & Composite $\mathcal{H}_1$

Example 4. Suppose that the null hypothesis $\mathcal{H}_0$ is simple, i.e. $\mathcal{H}_0 = \{P_0\}$, and the alternative hypothesis $\mathcal{H}_1$ is composite, containing many distributions for X^n, i.e. $\mathcal{H}_1 = \{P_\theta | \theta \in \Theta_1\}$ for some (nonsingleton) set Θ_1. Let W_1 be an arbitrary distribution on Θ_1, with density function $w_1(\theta)$. In Bayesian statistics, one interprets W_1 as a 'prior distribution' and measures the evidence in favor of $\mathcal{H}_1$ provided by the data X^n by the *Bayes factor* $\frac{p_{W_1}(X^n)}{p_0(X^n)}$ where we have

$$p_{W_1}(X^n) = \int_{\theta \in \Theta_1} p_\theta(X^n) \cdot w_1(\theta) d\theta = \prod_{i=1}^{n} p_{W_1}(X_i | X^{i-1})$$

with

$$p_{W_1}(X_i | X^{i-1}) = \int_{\theta \in \Theta_1} p_\theta(X_i) \cdot w_1(\theta | X^{i-1}) d\theta,$$

and $w_1(\theta | X^{i-1})$ the Bayesian posterior density of θ (Berger, 1985). No matter what distribution W_1 is chosen, the Bayes factor $\frac{p_{W_1}(X^n)}{p_0(X^n)}$ is an E-variable since

$$\mathbb{E}_{p_0}\left[\frac{p_{W_1}(X^n)}{p_0(X^n)}\right] = \int p_0(x^n) \cdot \frac{p_{W_1}(x^n)}{p_0(x^n)} dx^n = 1.$$

Composite $\mathcal{H}_0$ & Simple $\mathcal{H}_1$

In the case of a composite null hypothesis, E-variables are no longer as easy to construct as in the case of the simple null hypothesis $\mathcal{H}_0$. Grünwald et al. (2020) introduce a general way to find an E-variable nevertheless. To explain it, we first need to introduce KL divergence (Andersen, 1970, Kullback and Leibler, 1951) and reverse information projection (Li, 1999, Li and Barron, 1999).

Definition 5 (KL Divergence)**.** *Kullback-Leibler divergence, also known as relative entropy or expected log-likelihood ratio, is a measure of the difference between*

two probability distributions. Let P and Q be two probability distributions for a random variable X. The KL divergence of P from Q is denoted and defined as

$$D(P||Q) := \mathbb{E}_P\left[\log \frac{p(X)}{q(X)}\right].$$

Definition 6 (Reverse Information Projection (RIPr))**.** *Suppose that the null hypothesis $\mathcal{H}_0$ is a set of distributions for X^n with parameter θ, i.e. $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$. We enlarge it to a convex set $\{P_{W_0} | W_0 \in \mathcal{W}(\Theta_0)\}$ where $\mathcal{W}(\Theta_0)$ contains all distributions on Θ_0 and*

$$p_{W_0}(x^n) = \int_{\theta\in\Theta_0} p_\theta dW_0(\theta)$$

We call this 'the Bayes marginal distribution based on prior W_0'. We take the alternative hypothesis $\mathcal{H}_1$ to be simple, i.e. $\mathcal{H}_1 = \{P_1\}$. We call $P_{W_0^*}$ the Reverse Information Projection (RIPr) of P_1 *if*

$$W_0^* = \arg\min_{W_0\in\mathcal{W}(\Theta_0)} D(P_1||P_{W_0}).$$

In the remainder of this thesis we will also use the RIPr terminology for densities, i.e. we will also say '$p_{W_0^*}$ is the RIPr of p_1'.

Theorem 1 (Theorem 1 from (Grünwald et al., 2020))**.** *Based on the conditions in Definition 6, if $P_{W_0^*}$ is a RIPr of P_1, then $\frac{p_1}{p_{W_0^*}}$ is an E-variable, i.e.*

$$\mathbb{E}_{p_\theta}\left[\frac{p_1(X^n)}{p_{W_0^*}(X^n)}\right] \leq 1 \qquad \text{for } \forall\theta \in \Theta_0.$$

In this thesis, we focus on this case of composite null hypothesis $\mathcal{H}_0$ and the simple alternative hypothesis $\mathcal{H}_1$, since obtaining E-variables for simple null is easy, as indicated above, and, once we have an E-variable for simple alternative vs. composite null, it is easy to extend this to composite vs. composite. Various methods for doing this are described by Grünwald et al. (2020) (the 'GROW' and 'REGROW' e-variables). Thus, the only inherently difficult case is that of composite null and simple alternative, and this the only case we will consider from now on.

1.3 Overview of thesis and main results

Calculating the RIPr is a general method for finding an E-variable. There are three cases: the RIPr may be a single distribution in the null model, or

a mixture of such distributions, or it may be arbitrarily well-approximated (in the sense described by Li (1999)) by such mixtures, while not being itself such a mixture. When the RIPr is a single distribution, it is easy to calculate. We find that there is no general rule to find the RIPr for the whole exponential families. We aim to find some cases of exponential families in which the RIPr is simple (i.e. a single distribution in the null model). For three types of problems in which the null is an exponential family and the alternative is an element of some other exponential family (Chapter 2,3,4), we can obtain a simple RIPr when the simple conditions of Lemma (2,4,6) and Theorem (2,3,4) are satisfied. Lemma (3,5,7) are used to exclude the existence of a simple RIPr. Chapter 2 discusses E-variables where the alternative hypothesis is an element of a two-parameter exponential family. Chapter 3 discusses E-variables in which the alternative hypothesis is taken from a set of one-parameter exponential families indexed by an additional integer parameter k. Chapter 4 considers E-variables in the context of n outcomes of one-parameter exponential families. In Chapter 5, we extend the cases in which there exists a simple RIPr by transforming random variables and introduce some conditions (Theorem 5) of transformations that make it hold. We give several examples in all these chapters. We especially mention Example 4.2.2 which shows that there does not exist a simple RIPr for two outcomes of the exponential distribution. In Chapter 6, we approximate this non-simple RIPr anyway, as a finite mixture of distributions. We calculate it in a specific setting by programming in R and verify that it has the right properties.

Chapter 2

Two-parameter Exponential Families

In this chapter, we discuss E-values where the alternative is an element of an exponential family of dimension 2. In many cases, a particular value of one parameter can simplify this two-parameter exponential family to a well-known one-parameter exponential family which we can then take to be the null. For example, when the shape parameter α of a gamma distribution $Gamma(\alpha, \beta)$ is equal to 1, the gamma distribution is an exponential distribution with parameter β.

Assuming that RIPr is a single distribution, Lemma 2 simplifies the process of computing the parameter of this possible simple RIPr. Theorem 2 describes when this possible simple RIPr holds. Lemma 3 gives a simple sufficient condition to exclude this possible simple RIPr.

We consider the canonical form of exponential families with two parameters, i.e. with density

$$p_{\boldsymbol{\eta}}(x) = h(x) \exp(\boldsymbol{\eta}^T \boldsymbol{T}(x) - A(\boldsymbol{\eta})),$$

natural parameter vector

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_a \\ \eta_b \end{pmatrix},$$

sufficient statistic vector

$$\boldsymbol{T}(\boldsymbol{x}) = \begin{pmatrix} T_a(x) \\ T_b(x) \end{pmatrix},$$

and log-partition function

$$\begin{aligned} A(\boldsymbol{\eta}) &= \log \int_{x\in\mathcal{X}} h(x)\exp(\boldsymbol{\eta}^T\boldsymbol{T}(x))dx \qquad \text{in continuous distributions} \\ &= \log \sum_{x\in\mathcal{X}} h(x)\exp(\boldsymbol{\eta}^T\boldsymbol{T}(x)) \qquad \text{in discrete distributions,} \end{aligned}$$

where $h : \mathcal{X} \to \mathbb{R}_0^+$ is a non-negative function.

Let $\boldsymbol{\Theta}_{\boldsymbol{\eta}}$ be the set of values $\boldsymbol{\eta}$ for which this distribution is well-defined, i.e. $\boldsymbol{\Theta}_{\boldsymbol{\eta}} = \{\boldsymbol{\eta} : A(\boldsymbol{\eta}) < \infty\}$.

Now, we calculate some useful partial derivatives. The partial derivative for $A(\boldsymbol{\eta})$ in continuous distribution with respect to η_b is

$$\begin{aligned} \frac{\partial A(\boldsymbol{\eta})}{\partial \eta_b} &= \frac{\int T_b(x)h(x)\exp(\boldsymbol{\eta}^T\boldsymbol{T}(x))dx}{\int h(x)\exp(\boldsymbol{\eta}^T\boldsymbol{T}(x))dx} \\ &= \int T_b(x)h(x)\exp(\boldsymbol{\eta}^T\boldsymbol{T}(x) - A(\boldsymbol{\eta}))dx = \mathbb{E}_{p_{\boldsymbol{\eta}}}[T_b(X)]. \end{aligned}$$

The second partial derivative for $A(\boldsymbol{\eta})$ with respect to η_b is

$$\begin{aligned} \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta_b^2} &= \int T_b(x)h(x)\exp(\boldsymbol{\eta}^T\boldsymbol{T}(x) - A(\boldsymbol{\eta}))(T_b(x) - \frac{\partial}{\partial\eta_b}A(\boldsymbol{\eta}))dx \\ &= \int T_b(x)h(x)\exp(\boldsymbol{\eta}^T\boldsymbol{T}(x) - A(\boldsymbol{\eta}))(T_b(x) - \mathbb{E}_{p_{\boldsymbol{\eta}}}[T_b(X)])dx \\ &= \mathbb{E}_{p_{\boldsymbol{\eta}}}[T_b^2(X)] - \mathbb{E}_{p_{\boldsymbol{\eta}}}[T_b(X)]\mathbb{E}_{p_{\boldsymbol{\eta}}}[T_b(X)] = \mathrm{var}_{p_{\boldsymbol{\eta}}}[T_b(X)] \end{aligned}$$

We have the same results in discrete distributions.

2.1 Simple RIPr for Two-parameter Exponential Family

Let $\mathcal{P} = \{p_{\boldsymbol{\eta}} = p_{\eta_a,\eta_b} : \eta_a \in \Theta_{\eta_a},\ \eta_b \in \Theta_{\eta_b}\}$ with parameter space $\Theta_{\eta_a}, \Theta_{\eta_b} \subset \mathbb{R}$ denote a two-parameter exponential family. We suppose the null hypothesis $\mathcal{H}_0$ is composite and is given by

$$\mathcal{H}_0 : X \sim P_{\boldsymbol{\eta}_0} = p_{\eta_{0a},\eta_{0b}} \quad \text{for a fixed } \eta_{0a} \in \Theta_{\eta_a} \text{ and varying } \eta_{0b} \in \Theta_{\eta_b}.$$

The alternative hypothesis $\mathcal{H}_1$ simple and is given by

$$\mathcal{H}_1 : X \sim P_{\boldsymbol{\eta}_1} = p_{\eta_{1a},\eta_{1b}} \quad \text{for a fixed } \eta_{1a} \neq \eta_{0a}\,, \eta_{1a} \in \Theta_{\eta_a} \text{ and a fixed } \eta_{1b} \in \Theta_{\eta_b}.$$

In this chapter, we try to identify cases in which the RIPr $p_{W_0^*}$ is simple, which means that RIPr is a single distribution, i.e. $W_0^*(\boldsymbol{\eta_0^*}) = 1$ for some $\eta_0^* \in \Theta_{\eta_a} \times \Theta_{\eta_b}$ and then $p_{W_0^*} = p_{\boldsymbol{\eta_0^*}}$. By definition of RIPr, if such an $\boldsymbol{\eta}_0^*$ exists at all, it must minimize $D(p_{\boldsymbol{\eta_1}}||p_{\boldsymbol{\eta_0}})$ over $\boldsymbol{\eta_0} = (\eta_{0a}, \eta_{0b}), \eta_{0b} \in \Theta_{\eta_b}$. So our strategy will be to determine the $\boldsymbol{\eta_0^*}$ minimizing $D(p_{\boldsymbol{\eta_1}}||p_{\boldsymbol{\eta_0}})$ and then check whether it is a RIPr by checking whether $\frac{p_1}{p_{W_0^*}} = \frac{p_1}{p_{\boldsymbol{\eta_0^*}}}$ is an E-variable later.

Lemma 2. *Assuming there exists $\boldsymbol{\eta_0^*} = (\eta_{0a}, \eta_{0b}^*)^T$ satisfying $\mathbb{E}_{p_{\boldsymbol{\eta_0^*}}}[T_b(X)] = \mathbb{E}_{p_{\boldsymbol{\eta_1}}}[T_b(X)]$, then this $\boldsymbol{\eta_0^*}$ minimizes $D(p_{\boldsymbol{\eta_1}}||p_{\boldsymbol{\eta_0}})$ over $\boldsymbol{\eta_0} = (\eta_{0a}, \eta_{0b}), \eta_{0b} \in \Theta_{\eta_b}$.*

Proof. For all $\eta_{0b} \in \Theta_{\eta_b}$, the KL divergence satisfies:

$$\begin{aligned} D(p_{\boldsymbol{\eta_1}}||p_{\boldsymbol{\eta_0}}) &= \mathbb{E}_{p_{\boldsymbol{\eta_1}}}\left[\log \frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0}}}\right] = \mathbb{E}_{p_{\boldsymbol{\eta_1}}}\left[\log \frac{h(X)\exp(\boldsymbol{\eta_1}^T\boldsymbol{T}(X) - A(\boldsymbol{\eta_1}))}{h(X)\exp(\boldsymbol{\eta_0}^T\boldsymbol{T}(X) - A(\boldsymbol{\eta_0}))}\right] \\ &= \mathbb{E}_{p_{\boldsymbol{\eta_1}}}[\boldsymbol{\eta_1}^T\boldsymbol{T}(X) - \boldsymbol{\eta_0}^T\boldsymbol{T}(X) - A(\boldsymbol{\eta_1}) + A(\boldsymbol{\eta_0})] \\ &= \mathbb{E}_{p_{\boldsymbol{\eta_1}}}[\boldsymbol{\eta_1}^T\boldsymbol{T}(X) - \eta_{0a}T_a(X) - \eta_{0b}T_b(X) - A(\boldsymbol{\eta_1}) + A(\boldsymbol{\eta_0})] \end{aligned}$$

Taking the derivative with respect to η_{0b}, we have

$$\begin{aligned} \frac{dD(p_{\boldsymbol{\eta_1}}||p_{\boldsymbol{\eta_0}})}{d\eta_{0b}} &= -\mathbb{E}_{p_{\boldsymbol{\eta_1}}}[T_b(X)] + \frac{d}{d\eta_{0b}}A(\boldsymbol{\eta_0}) \\ &= -\mathbb{E}_{p_{\boldsymbol{\eta_1}}}[T_b(X)] + \mathbb{E}_{p_{\boldsymbol{\eta_0}}}[T_b(X)] \end{aligned}$$

and when $\boldsymbol{\eta_0^*} = (\eta_{0a}, \eta_{0b}^*)^T$ satisfies $\mathbb{E}_{p_{\boldsymbol{\eta_0^*}}}[T_b(X)] = \mathbb{E}_{p_{\boldsymbol{\eta_1}}}[T_b(X)]$, $\frac{dD(p_{\boldsymbol{\eta_1}}||p_{\boldsymbol{\eta_0}})}{d\eta_{0b}} = 0$. Taking the second derivative with respect to η_{0b}, we get

$$\frac{d^2D(p_{\boldsymbol{\eta_1}}||p_{\boldsymbol{\eta_0}})}{d\eta_{0b}^2} = \frac{d^2}{d\eta_{0b}^2}A(\boldsymbol{\eta_0}) = \mathrm{var}_{p_{\boldsymbol{\eta_0}}}[T_b(X)] > 0,$$

so this KL divergence is a convex function in η_{0b}, and is minimal when $\eta_{0b} = \eta_{0b^*}$. □

Theorem 2. *If $\boldsymbol{\eta_2} = \boldsymbol{\eta_1} + \boldsymbol{\eta_0} - \boldsymbol{\eta_0^*}$ is included in the parameter space $\Theta_{\boldsymbol{\eta}}$, i.e. $\eta_{2a} = \eta_{1a} \in \Theta_{\eta_a}$ and $\eta_{2b} = \eta_{1b} + \eta_{0b} - \eta_{0b}^* \in \Theta_{\eta_b}$, then*

$$\mathbb{E}_{p_{\boldsymbol{\eta_0}}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right] = \exp(A(\boldsymbol{\eta_0^*}) - A(\boldsymbol{\eta_1}) - A(\boldsymbol{\eta_0}) + A(\boldsymbol{\eta_2})).$$

Moreover, define $f(\eta_{0b}) := \log \mathbb{E}_{p_{\boldsymbol{\eta_0}}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right]$ *as a function of* η_{0b} *(with* η_{1a}*,* η_{1b}*,* η_{0a} *fixed). Then* $f(\eta_{0b}^*)$ *is a local maximum or minimum value.*
If $f(\eta_{0b})$ *takes the global maximum value at* η_{0b}^**, in particular, if* $f(\eta_{0b})$ *is increasing for* $\eta_{0b} < \eta_{0b}^*$ *and decreasing for* $\eta_{0b} > \eta_{0b}^*$*, then* $\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}$ *is an E-variable, i.e.*

$$\mathbb{E}_{p_{\boldsymbol{\eta_0}}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right] \leq 1, \qquad \forall\, \eta_{0b} \in \Theta_{\eta_b}.$$

Otherwise, if $f(\eta_{0b})$ *does not take the global maximum value at* η_{0b}^**, in particular, if* $f(\eta_{0b}^*)$ *is a local minimum value, then* $\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}$ *is not an E-variable and there does not exist a simple RIPr.*

Proof.

$$\begin{aligned}
\mathbb{E}_{p_{\boldsymbol{\eta_0}}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right] &= \int \frac{h(x)\exp(\boldsymbol{\eta_1}^T\boldsymbol{T}(x) - A(\boldsymbol{\eta_1}))}{h(x)\exp(\boldsymbol{\eta_0^*}^T\boldsymbol{T}(x) - A(\boldsymbol{\eta_0^*}))} \cdot h(x)\exp(\boldsymbol{\eta_0}^T\boldsymbol{T}(x) - A(\boldsymbol{\eta_0}))dx \\
&= \int h(x)\exp((\boldsymbol{\eta_1}^T + \boldsymbol{\eta_0}^T - \boldsymbol{\eta_0^*}^T)\boldsymbol{T}(x) - A(\boldsymbol{\eta_1}) - A(\boldsymbol{\eta_0}) + A(\boldsymbol{\eta_0^*}))dx \\
&= \int h(x)\exp(\boldsymbol{\eta_2}^T\boldsymbol{T}(x) - A(\boldsymbol{\eta_1}) - A(\boldsymbol{\eta_0}) + A(\boldsymbol{\eta_0^*}))dx \\
&= \exp(A(\boldsymbol{\eta_0^*}) - A(\boldsymbol{\eta_1}) - A(\boldsymbol{\eta_0}) + A(\boldsymbol{\eta_2})) \\
&\quad \cdot \int h(x)\exp(\boldsymbol{\eta_2}^T\boldsymbol{T}(x) - A(\boldsymbol{\eta_2}))dx \\
&= \exp(A(\boldsymbol{\eta_0^*}) - A(\boldsymbol{\eta_1}) - A(\boldsymbol{\eta_0}) + A(\boldsymbol{\eta_2})) \cdot 1 \qquad (2.1)
\end{aligned}$$

In (2.1), since $\boldsymbol{\eta_2}$ is included in the parameter space, $h(x)\exp(\boldsymbol{\eta_2}^T\boldsymbol{T}(x) - A(\boldsymbol{\eta_2}))$ is a density of the two-parameter exponential family with parameter $\boldsymbol{\eta_2}$, and the integral of a probability is equal to 1.

Recall that η_{1a}, η_{1b}, η_{0a} are fixed to particular values. We take the logarithm of the expectation as a function $f(\eta_{0b})$:

$$f(\eta_{0b}) = \log \mathbb{E}_{p_{\boldsymbol{\eta_0}}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right] = A(\boldsymbol{\eta_0^*}) - A(\boldsymbol{\eta_1}) - A(\boldsymbol{\eta_0}) + A(\boldsymbol{\eta_2}).$$

When $\eta_{0b} = \eta_{0b}^*$, we get $f(\eta_{0b}^*) = 0$ and $\mathbb{E}_{p_{\boldsymbol{\eta_0}^*}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right] = 1$.

Taking the derivative with respect to η_{0b}, we get

$$\begin{aligned}
\frac{df(\eta_{0b})}{d\eta_{0b}} &= \frac{d}{d\eta_{0b}}(A(\boldsymbol{\eta_2}) - A(\boldsymbol{\eta_0})) \\
&= \mathbb{E}_{p_{\boldsymbol{\eta_2}}}[T_b(X)] - \mathbb{E}_{p_{\boldsymbol{\eta_0}}}[T_b(X)].
\end{aligned}$$

When $\eta_{0b} = \eta_{0b}^*$, $\frac{df(\eta_{0b})}{d\eta_{0b}} = 0$, so $f(\eta_{0b})$ is a local maximum or minimum value.

If $f(\eta_{0b})$ takes the global maximum value at η_{0b}^*, then for $\forall\ \eta_{0b} \in \Theta_{\eta_b}$, $f(\eta_{0b}) \leq 0$ and $\mathbb{E}_{p_{\boldsymbol{\eta_0}}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right] \leq 1$.

If $f(\eta_{0b})$ does not take the maximum value at η_{0b}^*, then there is a η_{0b} such that $f(\eta_{0b}) > f(\eta_{0b}^*) = 0$ and $\mathbb{E}_{p_{\boldsymbol{\eta_0}}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right] > \mathbb{E}_{p_{\boldsymbol{\eta_0^*}}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right] = 1$. □

Lemma 3. *$f(\eta_{0b}) := \log \mathbb{E}_{p_{\boldsymbol{\eta_0}}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right]$ takes a local minimum value at η_{0b}^* iff*

$$\mathrm{var}_{p_{\boldsymbol{\eta_1}}}[T_b(X)] - \mathrm{var}_{p_{\boldsymbol{\eta_0^*}}}[T_b(X)] > 0.$$

Proof. Recall that $\frac{df(\eta_{0b})}{\eta_{0b}}|_{\eta_{0b}=\eta_{0b}^*} = 0$ in the proof of Theorem 2. Taking the second derivative of $f(\eta_{0b})$ with respect to η_{0b}, we find

$$\begin{aligned}\frac{d^2 f(\eta_{0b})}{d\eta_{0b}^2} &= \frac{d^2}{d\eta_{0b}^2}(A(\boldsymbol{\eta_2}) - A(\boldsymbol{\eta_0})) \\ &= \mathrm{var}_{p_{\boldsymbol{\eta_2}}}[T_b(X)] - \mathrm{var}_{p_{\boldsymbol{\eta_0}}}[T_b(X)].\end{aligned}$$

From the condition of the lemma, we have

$$\frac{d^2 f(\eta_{0b})}{d\eta_{0b}^2}|_{\eta_{0b}=\eta_{0b}^*} = \mathrm{var}_{p_{\boldsymbol{\eta_1}}}[T_b(X)] - \mathrm{var}_{p_{\boldsymbol{\eta_0^*}}}[T_b(X)] > 0,$$

then $f(\eta_{0b}^*)$ is a local minimum value. □

2.2 Examples

This section presents two examples.

2.2.1 Example: The Normal Distribution

We consider the set of normal distributions $\mathcal{N}(\mu, \sigma^2)$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ which is a common exponential family of dimension 2. The density function with parameters μ, σ^2, i.e. in the standard parameterization, is

$$p^{\circ}_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp(-\frac{(x-\mu)^2}{2\sigma^2}), \qquad x \in \mathbb{R}.$$

Straightforward calculation shows that the The canonical form is as below:

$$p^{\circ}_{\mu,\sigma^2}(x) = p_{\boldsymbol{\eta}}(x) = h(x)\exp(\boldsymbol{\eta}^T\boldsymbol{T}(x) - A(\boldsymbol{\eta})),$$

with $h(x) = \frac{1}{\sqrt{2\pi}}$, natural parameter vector

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_a \\ \eta_b \end{pmatrix} = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix},$$

sufficient statistic vector

$$\boldsymbol{T}(\boldsymbol{x}) = \begin{pmatrix} T_a(x) \\ T_b(x) \end{pmatrix} = \begin{pmatrix} x \\ x^2 \end{pmatrix},$$

and log-partition function

$$A(\boldsymbol{\eta}) = A^{\circ}(\mu,\sigma^2) = \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{\eta_a^2}{4\eta_b} - \frac{1}{2}\log(-2\eta_b).$$

The set of natural parameters $\Theta_{\boldsymbol{\eta}}$ is $\Theta_{\boldsymbol{\eta}} = \{\boldsymbol{\eta} = (\eta_a, \eta_b) : \eta_a \in \mathbb{R}, \eta_b < 0\}$.

Hypotheses: The null hypothesis $\mathcal{H}_0$ is a set of normal distributions with density p°_{0,σ_0^2}, mean 0 and variance σ_0^2 ($\sigma_0 \in \mathbb{R}_+$) (i.e. the normal scale family), and the corresponding canonical form with density $p_{0,\eta_{0b}}$ ($\eta_{0b} \in \mathbb{R}_-$). The alternative hypothesis $\mathcal{H}_1$ is a fixed normal distribution with density $p^{\circ}_{\mu_1,\sigma_1^2}$ with mean $\mu_1 \neq 0$, $\mu_1 \in \mathbb{R}$ and variance σ_1^2 ($\sigma_1 \in \mathbb{R}_+$). The corresponding canonical form with density $p_{\eta_{1a},\eta_{1b}}$ has $\eta_{1a} \neq 0$, $\eta_{1a} \in \mathbb{R}$ and $\eta_{1b} \in \mathbb{R}_-$.

Result: There is a simple RIPr $p^{\circ}_{0,\sigma_0^{*2}}$ for these hypotheses where $\sigma_0^{*2} = \sigma_1^2 + \mu_1^2$, so that $\frac{p^{\circ}_{\mu_1,\sigma_1^2}}{p^{\circ}_{0,\sigma_0^{*2}}}$ is an E-variable, i.e. for all $\mu_1 \in \mathbb{R}$,

$$\mathbb{E}_{p^{\circ}_{0,\sigma_0^2}}\left[\frac{p^{\circ}_{\mu_1,\sigma_1^2}}{p^{\circ}_{0,\sigma_0^{*2}}}\right] \leq 1, \qquad \text{for } \forall\ \sigma_0 \in \mathbb{R}_+.$$

Proof. By the parameter transformation, $D(p_{\boldsymbol{\eta_1}}||p_{\boldsymbol{\eta_0}})$ is the same as $D(p^{\circ}_{\mu_1,\sigma_1^2}||p^{\circ}_{0,\sigma_0^2})$, and getting η^*_{0b} from $\mathbb{E}_{p_{\boldsymbol{\eta_0^*}}}[X^2] = \mathbb{E}_{p_{\boldsymbol{\eta_1}}}[X^2]$ is equivalent to getting σ_0^{*2} from $\mathbb{E}_{p^{\circ}_{0,\sigma_0^{*2}}}[X^2] = \mathbb{E}_{p^{\circ}_{\mu_1,\sigma_1^2}}[X^2]$, with

$$\mathbb{E}_{p^{\circ}_{\mu,\sigma^2}}[X^2] = \mathrm{var}_{p^{\circ}_{\mu,\sigma^2}}[X] + (\mathbb{E}_{p^{\circ}_{\mu,\sigma^2}}[X])^2 = \sigma^2 + \mu^2.$$

We have

$$\sigma_0^{*2} = \sigma_1^2 + \mu_1^2 \text{ satisfying } \mathbb{E}_{p^\circ_{0,\sigma_0^{*2}}}[X^2] = \mathbb{E}_{p^\circ_{\mu_1,\sigma_1^2}}[X^2].$$

According to Lemma 2, this σ_0^{*2} minimizes KL divergence $D(p^\circ_{\mu_1,\sigma_1^2}||p^\circ_{0,\sigma_0^2})$. We have

$$\eta_{2a} = \eta_{1a} + \eta_{0a} - \eta^*_{0a} = \eta_{1a} = \frac{\mu_1}{\sigma_1^2}$$

and

$$\eta_{2b} = \eta_{1b} + \eta_{0b} - \eta^*_{0b} = -\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2} + \frac{1}{2\sigma_0^{*2}} = -\frac{\sigma_0^2\mu_1^2 + \sigma_1^2\sigma_0^{*2}}{2\sigma_0^2\sigma_1^2\sigma_0^{*2}} < 0.$$

We get

$$\begin{aligned} A(\boldsymbol{\eta_2}) &= -\frac{\eta_{2a}^2}{4\eta_{2b}} - \frac{1}{2}\log(-2\eta_{2b}) \\ &= \frac{\sigma_0^{*2}\mu_1^2}{2\sigma_1^2} \cdot \frac{\sigma_0^2}{\sigma_0^2\mu_1^2 + \sigma_1^2\sigma_0^{*2}} - \frac{1}{2}\log(\sigma_0^2\mu_1^2 + \sigma_1^2\sigma_0^{*2}) + \log\sigma_1 + \log\sigma_0 + \log\sigma_0^*. \end{aligned}$$

From Theorem 2, we have

$$f^\circ(\sigma_0) = f(\eta_{0b}) = \log \mathbb{E}_{p_{\boldsymbol{\eta}_0}}\left[\frac{p_{\boldsymbol{\eta}_1}}{p_{\boldsymbol{\eta}_0^*}}\right] = A(\boldsymbol{\eta_0^*}) - A(\boldsymbol{\eta_1}) - A(\boldsymbol{\eta_0}) + A(\boldsymbol{\eta_2}).$$

Taking the first derivative of $f^\circ(\sigma_0)$ with respect to σ_0, we get

$$\begin{aligned} \frac{df^\circ(\sigma_0)}{d\sigma_0} &= \frac{dA(\boldsymbol{\eta_2})}{d\sigma_0} - \frac{dA(\boldsymbol{\eta_0})}{d\sigma_0} \\ &= \frac{d}{d\sigma_0}\left[\frac{\sigma_0^{*2}\mu_1^2}{2\sigma_1^2} \cdot \frac{\sigma_0^2}{\sigma_0^2\mu_1^2 + \sigma_1^2\sigma_0^{*2}} - \frac{1}{2}\log(\sigma_0^2\mu_1^2 + \sigma_1^2\sigma_0^{*2}) + \log\sigma_0\right] \\ &\quad - \frac{d}{d\sigma_0}\left[\log\sigma_0\right] \\ &= \frac{\sigma_0\mu_1^4(\sigma_0^{*2} - \sigma_0^2)}{(\sigma_0^2\mu_1^2 + \sigma_1^2\sigma_0^{*2})^2}. \end{aligned}$$

If $\sigma_0 < \sigma_0^*$, then $\frac{df^\circ(\sigma_0)}{d\sigma_0} > 0$. If $\sigma_0 > \sigma_0^*$, then $\frac{df^\circ(\sigma_0)}{d\sigma_0} < 0$. So $f^\circ(\sigma_0)$ first increases and then decreases, taking the maximum value at σ_0^*. Moreover, $f^\circ(\sigma_0^*) = f(\eta_{0b}^*) = 0$, so for all $\sigma_0 \in \mathbb{R}_+$, $f^\circ(\sigma_0) \leq f^\circ(\sigma_0^*) = 0$ and $\mathbb{E}_{p^\circ_{0,\sigma_0^2}}\left[\frac{p^\circ_{\mu_1,\sigma_1^2}}{p^\circ_{0,\sigma_0^{*2}}}\right] \leq \mathbb{E}_{p^\circ_{0,\sigma_0^{*2}}}\left[\frac{p^\circ_{\mu_1,\sigma_1^2}}{p^\circ_{0,\sigma_0^{*2}}}\right] = 1.$ □

2.2.2 Example: The Gamma Distribution

Another well-known two-dimensional exponential family distribution is the gamma distribution $Gamma(\alpha, \beta)$ ($\alpha > 0$, $\beta > 0$). It has density in the standard parameterization, with parameters α, β:

$$p^{\circ}_{\alpha,\beta}(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \quad x > 0,$$

where $\Gamma(\alpha)$ is the gamma function: when $n \in \mathbb{N}_+$,

$$\Gamma(n) = (n-1)!$$

and when z is a real number,

$$\Gamma(z) = \int_0^{\infty} x^{z-1} \exp(-x) dx.$$

In particular, $\Gamma(1) = 1$. Moreover, the mean and variance of this distribution are, respectively,

$$\mathbb{E}_{p^{\circ}_{\alpha,\beta}}[X] = \frac{\alpha}{\beta}, \qquad \mathrm{var}_{p^{\circ}_{\alpha,\beta}}[X] = \frac{\alpha}{\beta^2}.$$

When $\alpha = 1$, it can be simplified to an exponential distribution with parameter β as below:

$$p^{\circ}_{1,\beta}(x) = \frac{\beta^{1}}{\Gamma(1)} x^{1-1} \exp(-\beta x) = \beta \exp(-\beta x).$$

We require $\eta_a = 0$ in the canonical form when $\alpha = 1$. The canonical form is

$$p^{\circ}_{\alpha,\beta}(x) = p_{\boldsymbol{\eta}}(x) = h(x) \exp(\boldsymbol{\eta}^T \boldsymbol{T}(x) - A(\boldsymbol{\eta})),$$

with $h(x) = 1$, natural parameter vector

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_a \\ \eta_b \end{pmatrix} = \begin{pmatrix} \alpha - 1 \\ -\beta \end{pmatrix},$$

sufficient statistic vector

$$\boldsymbol{T}(\boldsymbol{x}) = \begin{pmatrix} T_a(x) \\ T_b(x) \end{pmatrix} = \begin{pmatrix} \log x \\ x \end{pmatrix},$$

and log-partition function

$$A(\boldsymbol{\eta}) = A^{\circ}(\alpha, \beta) = \log \Gamma(\alpha) - \alpha \log \beta = \log \Gamma(\eta_a + 1) - (\eta_a + 1) \log(-\eta_b).$$

The natural parameter space $\Theta_{\boldsymbol{\eta}}$ is $\Theta_{\boldsymbol{\eta}} = \{\boldsymbol{\eta} = (\eta_a, \eta_b) : \eta_a > 1, \eta_b < 0\}$.

Hypotheses: Assume that the null hypothesis $\mathcal{H}_0$ consists of the gamma distributions with density p°_{1,β_0} ($\beta_0 > 0$). These gamma distributions are also exponential distributions with parameter β_0. If we convert them to canonical form, then the density becomes $p_{0,\eta_{0b}}$ with $\eta_{0b} < 0$. For the alternative hypothesis $\mathcal{H}_1$, we take a fixed gamma distribution with density $p^{\circ}_{\alpha_1,\beta_1}$, parameters in the standard parameterization $\alpha_1 \neq 1$, $\alpha_1 > 0$ and $\beta_1 > 0$. The density in the canonical form is $p_{\eta_{1a},\eta_{1b}}$ with natural parameters $\eta_{1a} \neq 0$, $\eta_{1a} > -1$ and $\eta_{1b} < 0$.

Result: Based on the hypotheses above, if $\alpha_1 > 1$, then there is a simple RIPr p°_{1,β_0^*} where $\beta_0^* = \frac{\beta_1}{\alpha_1}$, and $\frac{p^{\circ}_{\alpha_1,\beta_1}}{p^{\circ}_{1,\beta_0^*}}$ is an E-variable, i.e.

$$\mathbb{E}_{p^{\circ}_{1,\beta_0}}\left[\frac{p^{\circ}_{\alpha_1,\beta_1}}{p^{\circ}_{1,\beta_0^*}}\right] \leq 1, \qquad \text{for } \forall\ \beta_0 \in \mathbb{R}_+.$$

If $0 < \alpha_1 < 1$ and $\beta_0 > (\frac{1}{\alpha_1} - 1)\beta_1$, then $\frac{p^{\circ}_{\alpha_1,\beta_1}}{p^{\circ}_{1,\beta_0^*}}$ is not an E-variable and there does not exist a simple RIPr.
If $0 < \alpha_1 < 1$ and $\beta_0 < (\frac{1}{\alpha_1} - 1)\beta_1$, then $\mathbb{E}_{p^{\circ}_{1,\beta_0}}\left[\frac{p^{\circ}_{\alpha_1,\beta_1}}{p^{\circ}_{1,\beta_0^*}}\right]$ is not well-defined.

Proof. By the parameter transformation, $p_{0,\eta_{0b}} = p^{\circ}_{1,\beta_0}$ and $p_{\eta_{1a},\eta_{1b}} = p^{\circ}_{\alpha_1,\beta_1}$. The mean of a gamma distribution is

$$\mathbb{E}_{p_{\eta_a,\eta_b}}[X] = \mathbb{E}_{p^{\circ}_{\alpha,\beta}}[X] = \frac{\alpha}{\beta}.$$

We know $\beta_0^* = \frac{\beta_1}{\alpha_1}$, so it holds that

$$\mathbb{E}_{p_{0,\eta_{0b}^*}}[X] = \mathbb{E}_{p^{\circ}_{1,\beta_0^*}}[X] = \frac{1}{\beta_0^*} = \frac{\alpha_1}{\beta_1} = \mathbb{E}_{p^{\circ}_{\alpha_1,\beta_1}}[X] = \mathbb{E}_{p_{\eta_{1a},\eta_{1b}}}[X].$$

According to Lemma 2, this β_0^* makes the KL divergence $D(p_{\eta_{1a},\eta_{1b}}||p_{\eta_{0a},\eta_{0b}})$ (in other words, $D(p^{\circ}_{\alpha_1,\beta_1}||p^{\circ}_{\alpha_0,\beta_0})$) take its minimum value.

In addition, we have

$$\eta_{2a} = \eta_{1a} + \eta_{0a} - \eta_{0a^*} = \eta_{1a} = \alpha_1 - 1$$

and

$$\eta_{2b} = \eta_{1b} + \eta_{0b} - \eta_{0b^*} = -\beta_1 - \beta_0 + \beta_0^* = -\beta_1(1 - \frac{1}{\alpha_1}) - \beta_0.$$

For $\alpha_1 > 1$, we have $\eta_{2b} < 0$, which is in this canonical parameter space. We have

$$\begin{aligned} A(\boldsymbol{\eta_2}) &= \log\Gamma(\eta_{2a}+1) - (\eta_{2a}+1)\log(-\eta_{2b}) \\ &= \log\Gamma(\alpha_1) - \alpha_1\log(\beta_1+\beta_0-\beta_0^*). \end{aligned}$$

From Theorem 2, we obtain

$$f^\circ(\beta_0) = f(\eta_{0b}) = \log \mathbb{E}_{p_{\boldsymbol{\eta_0}}}\left[\frac{p_{\boldsymbol{\eta_1}}}{p_{\boldsymbol{\eta_0^*}}}\right] = A(\boldsymbol{\eta_0^*}) - A(\boldsymbol{\eta_1}) - A(\boldsymbol{\eta_0}) + A(\boldsymbol{\eta_2}).$$

Taking the first derivative of $f^\circ(\beta_0)$ with respect to β_0, we get

$$\begin{aligned} \frac{df^\circ(\beta_0)}{d\beta_0} &= \frac{dA(\boldsymbol{\eta_2})}{d\beta_0} - \frac{dA(\boldsymbol{\eta_0})}{d\beta_0} \\ &= \frac{d}{d\beta_0}\left[-\alpha_1\log(\beta_1+\beta_0-\beta_0^*)\right] - \frac{d}{d\beta_0}\left[-\log\beta_0\right] \\ &= \frac{1}{\beta_0} - \frac{\alpha_1}{\beta_1+\beta_0-\beta_0^*} \\ &= \frac{(\alpha_1-1)(\beta_0^*-\beta_0)}{\beta_0[\beta_0+\beta_1(1-\frac{1}{\alpha_1})]}. \end{aligned}$$

Since $\alpha_1 > 1$, if $\beta_0 < \beta_0^*$, then $\frac{df^\circ(\beta_0)}{d\beta_0} > 0$. If $\beta_0 > \beta_0^*$, then $\frac{df^\circ(\beta_0)}{d\beta_0} < 0$. So $f^\circ(\beta_0)$ first increases and then decreases, taking the maximum value at β_0^*. Moreover, $f^\circ(\beta_0^*) = f^\circ(\eta_{0b}^*) = 0$, so for all $\beta_0 \in \mathbb{R}_+$, $f^\circ(\beta_0) \le f^\circ(\beta_0^*) = 0$ and $\mathbb{E}_{p^\circ_{1,\beta_0}}\left[\frac{p^\circ_{\alpha_1,\beta_1}}{p^\circ_{1,\beta_0^*}}\right] \le \mathbb{E}_{p^\circ_{1,\beta_0^*}}\left[\frac{p^\circ_{\alpha_1,\beta_1}}{p^\circ_{1,\beta_0^*}}\right] = 1$. Then $\frac{p^\circ_{\alpha_1,\beta_1}}{p^\circ_{1,\beta_0^*}}$ is an E-variable.

For $0 < \alpha_1 < 1$, only when $\beta_0 > (\frac{1}{\alpha_1}-1)\beta_1$, $\eta_{2b} < 0$ holds.
We know that the variance of gamma distribution is

$$\mathrm{var}_{p_{\eta_a,\eta_b}}[X] = \mathrm{var}_{p^\circ_{\alpha,\beta}}[X] = \frac{\alpha}{\beta^2},$$

so

$$\begin{aligned} \mathrm{var}_{p_{\eta_{1a},\eta_{1b}}}[X] - \mathrm{var}_{p_{0,\eta_{0b}^*}}[X] &= \mathrm{var}_{p^\circ_{\alpha_1,\beta_1}}[X] - \mathrm{var}_{p^\circ_{1,\beta_0^*}}[X] \\ &= \frac{\alpha_1}{\beta_1^2} - \frac{1}{\beta_0^{*2}} = \frac{\alpha_1}{\beta_1^2}(1-\alpha_1) > 0. \end{aligned}$$

From Lemma 3, $\log \mathbb{E}_{p^\circ_{1,\beta_0}}\left[\frac{p^\circ_{\alpha_1,\beta_1}}{p^\circ_{1,\beta_0^*}}\right]$ takes a local minimum value at β_0^*. According to Theorem 2, $\frac{p_{\alpha_1,\beta_1}}{p_{1,\beta_0^*}}$ is not an E-variable.

For $0 < \alpha_1 < 1$, when $\beta_0 < (\frac{1}{\alpha_1} - 1)\beta_1$, $\eta_{2b} > 0$ and then it is not included in the parameter space Θ_{η_b}. □

Chapter 3

One-parameter Exponential Families with Fixed Parameter k

This chapter discusses E-values in which the alternative is taken from a set of exponential families of dimension 1 indexed by an additional integer parameter k. The sufficient statistic $T(x)$ and natural parameter η are independent of parameter k in the canonical form and the probability density (mass) function of this distribution is given by

$$p_{k,\eta}(x) = h_k(x) \exp(\eta T(x) - A_k(\eta)), \qquad (k \text{ fixed})$$

where natural parameter η is a function of original parameter (i.e. in the standard parameterization) θ

$$\eta = \eta(\theta),$$

$h_k(x)$ is a non-negative function of x with fixed k

$$h_k(x) = h(x,k),$$

and log-partition function is

$$\begin{aligned} A_k(\eta) &= \log \int h_k(x) \exp(\eta T(x)) dx \qquad \text{in continuous distributions} \\ &= \log \sum_x h_k(x) \exp(\eta T(x)) \qquad \text{in discrete distributions.} \end{aligned}$$

The treatment in this chapter is analogous to the treatment in chapter 2. Lemma 4 gives us a candidate for the RIPr with a simple form (i.e. a single distribution). Theorem 3 indicates when this candidate is indeed the RIPr and Lemma 5 gives an easy characterization when it is not.

We compute some useful derivatives. The first derivative for $A_k(\eta)$ in continuous distributions is

$$\begin{aligned}\frac{dA_k(\eta)}{d\eta} &= \frac{\int T(x)h_k(x)\exp(\eta T(x))dx}{\int h_k(x)\exp(\eta T(x))dx} \\ &= \int T(x)h_k(x)\exp(\eta T(x) - A_k(\eta))dx = \mathbb{E}_{p_{k,\eta}}[T(X)].\end{aligned}$$

The second derivative for $A_k(\eta)$ is

$$\begin{aligned}\frac{d^2A_k(\eta)}{d\eta^2} &= \int T(x)h_k(x)\exp(\eta T(x) - A_k(\eta))(T(x) - \frac{dA_k(\eta)}{d\eta})dx \\ &= \int T(x)h_k(x)\exp(\eta T(x) - A_k(\eta))(T(x) - \mathbb{E}_{p_{k,\eta}}[T(X)])dx \\ &= \mathbb{E}_{p_{k,\eta}}[T^2(X)] - \mathbb{E}_{p_{k,\eta}}[T(X)]\mathbb{E}_{p_{k,\eta}}[T(X)] = \text{var}_{p_{k,\eta}}[T(X)].\end{aligned}$$

We have the same results for discrete distributions.

3.1 Simple RIPr for One-parameter Exponential Families with Fixed k

We use $\mathcal{P} = \{p_{k,\eta} : \text{fixed } k \in \Theta_k, \eta \in \Theta_\eta\}$ with $\Theta_k \subset \mathbb{N}$, $\Theta_\eta \subset \mathbb{R}$ to denote the one-parameter families we mentioned above. The null hypothesis is composite and is given by

$$\mathcal{H}_0 : X \sim P_{k_0,\eta_0} \qquad \text{for a fixed } k_0 \in \Theta_k \text{ and varying } \eta_0 \in \Theta_\eta$$

The alternative hypothesis is simple and is given by

$$\mathcal{H}_1 : X \sim P_{k_1,\eta_1} \qquad \text{for a fixed } k_1 \neq k_0, k_1 \in \Theta_k \text{ and a fixed } \eta_1 \in \Theta_\eta$$

In this chapter, just like in the previous chapter, we aim to find conditions under which the RIPr $p_{W_0^*}$ is simple, that is, RIPr is a single distribution, i.e. $W_0^*(\eta_0^*) = 1$ and then $p_{W_0^*} = p_{\eta_0^*}$. Again, we first find the η_0^* minimizing KL divergence $D(p_{k_1,\eta_1}||p_{k_0,\eta_0})$ over $\eta_0 \in \Theta_\eta$; if the RIPr is simple it must be given by $p_{\eta_0^*}$. After finding $p_{\eta_0^*}$, we must check whether it is a RIPr by checking whether $\frac{p_1}{p_{\eta_0^*}}$ is an E-variable.

Lemma 4. *The η_0^* satisfying $\mathbb{E}_{p_{k_1,\eta_1}}[T(X)] = \mathbb{E}_{p_{k_0,\eta_0^*}}[T(X)]$, if it exists, minimizes $D(p_{k_1,\eta_1}||p_{k_0,\eta_0})$ over $\eta_0 \in \Theta_\eta$.*

Proof. For $\forall\ \eta_0 \in \Theta_\eta$, KL divergence

$$\begin{aligned} D(p_{k_1,\eta_1}||p_{k_0,\eta_0}) &= \mathbb{E}_{p_{k_1,\eta_1}}\left[\log \frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0}}\right] \\ &= \mathbb{E}_{p_{k_1,\eta_1}}\left[\log \frac{h_{k_1}(X)\exp(\eta_1 T(X) - A_{k_1}(\eta_1))}{h_{k_0}(X)\exp(\eta_0 T(X) - A_{k_0}(\eta_0))}\right] \\ &= \mathbb{E}_{p_{k_1,\eta_1}}\left[\log h_{k_1}(X) - \log h_{k_0}(X) + \eta_1 T(X) - \eta_0 T(X)\right. \\ &\quad \left. - A_{k_1}(\eta_1) + A_{k_0}(\eta_0)\right] \end{aligned}$$

Taking the derivative with respect to η_0,we get

$$\begin{aligned} \frac{dD(p_{k_1,\eta_1}||p_{k_0,\eta_0})}{d\eta_0} &= -\mathbb{E}_{p_{k_1,\eta_1}}[T(X)] + \frac{d}{d\eta_0}A_{k_0}(\eta_0) \\ &= -\mathbb{E}_{p_{k_1,\eta_1}}[T(X)] + \mathbb{E}_{p_{k_0,\eta_0}}[T(X)] \end{aligned}$$

When η_0^* satisfies $\mathbb{E}_{p_{k_1,\eta_1}}[T(X)] = \mathbb{E}_{p_{k_0,\eta_0}}[T(X)]$, $\frac{dD(p_{k_1,\eta_1}||p_{k_0,\eta_0})}{d\eta_0} = 0$. Taking the second derivative with respect to η_0, we have

$$\frac{d^2 D(p_{k_1,\eta_1}||p_{k_0,\eta_0})}{d\eta_0^2} = \frac{d^2}{d\eta_0^2}A_{k_0}(\eta_0) = \mathrm{var}_{p_{k_0,\eta_0}}[T(x)] > 0,$$

so this KL divergence is a convex function, and is minimal when $\eta_0 = \eta_0^*$. □

Theorem 3. *If $\eta_2 = \eta_1 + \eta_0 - \eta_0^*$ as a function of η_0 is included in parameter space Θ_η, then*

$$\mathbb{E}_{p_{k_0,\eta_0}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right] = \exp(A_{k_0}(\eta_0^*) - A_{k_1}(\eta_1) - A_{k_0}(\eta_0) + A_{k_1}(\eta_2)).$$

Moreover, define $f(\eta_0) := \log \mathbb{E}_{p_{k_0,\eta_0}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right]$ as a function of η_0 (with k_1, η_1, k_0 fixed). Then $f(\eta_0)$ is a local maximum or minimum value.
If $f(\eta_0)$ takes the maximum value at η_0^*, in particular, if $f(\eta_0)$is increasing for $\eta_0 < \eta_0^*$ and decreasing for $\eta_0 > \eta_0^*$, then $\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}$ is an E-variable, i.e.

$$\mathbb{E}_{p_{k_0,\eta_0}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right] \leq 1, \qquad \forall \eta_0 \in \Theta_\eta,$$

Otherwise, if $f(\eta_0)$ does not take the maximum value at η_0^*, in particular, if $f(\eta_0^*)$ is a local minimum value, then $\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}$ is not an E-variable, and there does not exist a simple RIPr.

Proof.

$$\begin{aligned}
\mathbb{E}_{p_{k_0,\eta_0}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right] &= \int \frac{h_{k_1}(x)\exp(\eta_1 T(x) - A_{k_1}(\eta_1))}{h_{k_0}(x)\exp(\eta_0^* T(x) - A_{k_0}(\eta_0^*))} h_{k_0}(x) \\
&\quad \cdot \exp(\eta_0 T(x) - A_{k_0}(\eta_0))dx \\
&= \int h_{k_1}(x)\exp((\eta_1 + \eta_0 - \eta_0^*)T(x) - A_{k_1}(\eta_1) \\
&\quad - A_{k_0}(\eta_0) + A_{k_0}(\eta_0^*))dx \\
&= \int h_{k_1}(x)\exp(\eta_2 T(x) - A_{k_1}(\eta_1) - A_{k_0}(\eta_0) + A_{k_0}(\eta_0^*))dx \\
&= \exp(A_{k_0}(\eta_0^*) - A_{k_1}(\eta_1) - A_{k_0}(\eta_0) + A_{k_1}(\eta_2)) \\
&\quad \cdot \int h_{k_1}(x)\exp(\eta_2 T(x) - A_{k_1}(\eta_2))dx \\
&= \exp(A_{k_0}(\eta_0^*) - A_{k_1}(\eta_1) - A_{k_0}(\eta_0) + A_{k_1}(\eta_2)) \cdot 1 \qquad (3.1)
\end{aligned}$$

In (3.1), since η_2 is included in the parameter space, $h_{k_1}(x)\exp(\eta_2 T(x) - A_{k_1}(\eta_2))$ is density of the one-parameter exponential family with parameter η_2, and the integral of a probability is equal to 1.

For fixed k_1, η_1 and k_0, set the logarithm of the expectation to the function $f(\eta_0)$:

$$f(\eta_0) = \log \mathbb{E}_{p_{k_0,\eta_0}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right] = A_{k_0}(\eta_0^*) - A_{k_1}(\eta_1) - A_{k_0}(\eta_0) + A_{k_1}(\eta_2)$$

When $\eta_0 = \eta_0^*$, we get $f(\eta_0^*) = 0$ and $\mathbb{E}_{p_{k_0,\eta_0^*}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right] = 1$.

Taking the derivative with respect to η_0, we have

$$\begin{aligned}
\frac{df(\eta_0)}{d\eta_0} &= \frac{d}{d\eta_0}(A_{k_1}(\eta_2) - A_{k_0}(\eta_0)) \\
&= \mathbb{E}_{p_{k_1,\eta_2}}[T(X)] - \mathbb{E}_{p_{k_0,\eta_0}}[T(X)]
\end{aligned}$$

When $\eta_0 = \eta_0^*$, $\frac{df(\eta_0)}{d\eta_0} = 0$, so $f(\eta_0)$ is a local maximum or minimum value.

If $f(\eta_0)$ takes the maximum value at η_0^*, then for $\forall\ \eta_0 \in \Theta_\eta$, $f(\eta_0) \leq 0$ and $\mathbb{E}_{p_{k_0,\eta_0}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right] \leq 1$.

If $f(\eta_0)$ does not take the maximum value at η_0^*, then there is a η_0 such that $f(\eta_0) > f(\eta_0^*) = 0$ and $\mathbb{E}_{p_{k_0,\eta_0}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right] > \mathbb{E}_{p_{k_0,\eta_0^*}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right] = 1$. □

Lemma 5. *$f(\eta_0) := \log \mathbb{E}_{p_{k_0,\eta_0}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right]$ takes a local minimum value at η_0^* iff*

$$var_{p_{k_1,\eta_1}}[T(X)] - var_{p_{k_0,\eta_0^*}}[T(X)] > 0$$

Proof. We have $\frac{df(\eta_0)}{d\eta_0} = 0$ when $\eta_0 = \eta_0^*$ in the proof of Theorem 3. Taking the second derivative of $f(\eta_0)$ with respect to η_0, we get

$$\begin{aligned}\frac{d^2 f(\eta_0)}{d\eta_0^2} &= \frac{d^2}{d\eta_0^2}(A_{k_1}(\eta_2) - A_{k_0}(\eta_0)) \\ &= \mathrm{var}_{p_{k_1,\eta_2}}[T(X)] - \mathrm{var}_{p_{k_0,\eta_0}}[T(X)]\end{aligned}$$

From the condition of the lemma, we have

$$\frac{d^2 f(\eta_0)}{d\eta_0^2}|_{\eta_0=\eta_0^*} = \mathrm{var}_{p_{k_1,\eta_1}}[T(X)] - \mathrm{var}_{p_{k_0,\eta_0^*}}[T(X)] > 0$$

then $f(\eta_0^*)$ is a local minimum value. □

3.2 Example

3.2.1 Example: Negative Binomial Distributions

The negative binomial distribution $NB(r,\theta)$ describes the number of success before r failures occurs, with stopping parameter $r \in \mathbb{N}_+$ and success probability $\theta \in [0,1)$. It is a one-parameter exponential family distribution for every fixed value of r. The density of this distribution is

$$p_{r,\theta}(x) = \binom{x+r-1}{r-1}\theta^x(1-\theta)^r, \qquad x \in \mathbb{N}$$

The mean and variance of this distribution are

$$\mathbb{E}_{p_{r,\theta}}[X] = \frac{\theta r}{1-\theta} \qquad \text{and} \qquad \mathrm{var}_{p_{r,\theta}}[X] = \frac{\theta r}{(1-\theta)^2}.$$

The canonical form is

$$p_{r,\theta}(x) = p_{r,\eta}(x) = h_r(x)\exp(\eta T(x) - A_r(\eta)) \qquad (\text{for fixed } r)$$

with

$$h_r(x) = \binom{x+r-1}{r-1},$$

natural parameter $\eta = \log\theta$, sufficient statistic $T(x) = x$ and log-partition function

$$A_r(\eta) = A_r^\circ(\theta) = -r\log(1-\theta) = -r\log(1-\exp(\eta)).$$

The natural parameter space is $\Theta_\eta = \{\eta : \eta < 0\}$

Hypotheses: The null hypothesis $\mathcal{H}_0$ is the set of negative binomial distributions $NB(r_0, \theta_0)$ with fixed $r_0 \in \mathbb{N}_+$ and varying $\theta_0 \in [0,1)$. The corresponding canonical form p_{r_0,η_0} has parameter $\eta_0 < 0$. The alternative hypothesis $\mathcal{H}_1$ is a single negative binomial distribution $NB(r_1, \theta_1)$ with a fixed $r_1 \in \mathbb{N}_+$, $r_0 \neq r_1$ and a fixed $\theta_1 \in [0,1)$. The corresponding canonical form p_{r_1,η_1} has parameter $\eta_1 < 0$.

Result: If $0 < r_0 < r_1$, there is a simple RIPr p_{r_0,θ_0^*} based on the above hypotheses where $\theta_0^* = \frac{\theta_1 r_1}{\theta_1 r_1 + (1-\theta_1)r_0}$, and $\frac{p^\circ_{r_1,\theta_1}}{p^\circ_{r_0,\theta_0^*}}$ is an E-variable, i.e.

$$\mathbb{E}_{p^\circ_{r_0,\theta_0}}\left[\frac{p^\circ_{r_1,\theta_1}}{p^\circ_{r_0,\theta_0^*}}\right] \leq 1, \qquad \text{for } \forall\theta_0 \in [0,1).$$

If $0 < r_1 < r_0$ and $0 \leq \theta_0 < \frac{1}{(\frac{r_0}{r_1}-1)(1-\theta_1)+1} < 1$, then $\frac{p^\circ_{r_1,\theta_1}}{p^\circ_{r_0,\theta_0^*}}$ is not an E-variable and there does not exist a simple RIPr.
If $0 < r_1 < r_0$ and $0 \leq \frac{1}{(\frac{r_0}{r_1}-1)(1-\theta_1)+1} \leq \theta_0 < 1$, then $\mathbb{E}_{p^\circ_{r_0,\theta_0}}\left[\frac{p^\circ_{r_1,\theta_1}}{p^\circ_{r_0,\theta_0^*}}\right]$ is not well-defined.

Proof. We know $p^\circ_{r_0,\theta_0} = p_{r_0,\eta_0}$ and $p^\circ_{r_1,\theta_1} = p_{r_1,\eta_1}$ by the parameter transformation. The mean of a negative binomial distribution is

$$\mathbb{E}_{p_{r,\eta}}[X] = \mathbb{E}_{p^\circ_{r,\theta}}[X] = \frac{\theta r}{1-\theta}.$$

From the condition $\theta_0^* = \frac{\theta_1 r_1}{\theta_1 r_1 + (1-\theta_1)r_0}$, we get

$$\mathbb{E}_{p_{r_1,\eta_1}}[X] = \mathbb{E}_{p^\circ_{r_1,\theta_1}}[X] = \frac{\theta_1 r_1}{1-\theta_1} = \frac{\theta_0^* r_0}{1-\theta_0^*} = \mathbb{E}_{p^\circ_{r_0,\theta_0^*}}[X] = \mathbb{E}_{p_{r_0,\eta_0^*}}[X]$$

According to Lemma 4, this θ_0^* minimizes KL divergence $D(p_{r_1,\eta_1}||p_{r_0,\eta_0})$. Moreover,

$$\begin{aligned}\eta_2 &= \eta_1 + \eta_0 - \eta_0^* = \log\theta_1 + \log\theta_0 - \log\theta_0^* \\ &= \log\frac{\theta_1\theta_0}{\theta_0^*} \\ &= \log\left[(\frac{r_0}{r_1}-1)(1-\theta_1)+1\right] + \log\theta_0.\end{aligned}$$

If $0 < r_0 < r_1$ and $\theta_0, \theta_1 \in [0,1)$, then $\eta_2 < \log(0+1) + \log\theta_0 < 0$ is included in the natural parameter space.
We have

$$\begin{aligned}A_{r_1}(\eta_2) &= -r_1\log(1-\exp(\eta_2)) = -r_1\log(1-\exp(\eta_2)) \\ &= r_1\log\theta_0^* - r_1\log(\theta_0^* - \theta_1\theta_0).\end{aligned}$$

From Theorem 3, we obtain

$$f^\circ(\theta_0) = f(\eta_0) = \log\mathbb{E}_{p_{k_0,\eta_0}}\left[\frac{p_{k_1,\eta_1}}{p_{k_0,\eta_0^*}}\right] = A_{k_0}(\eta_0^*) - A_{k_1}(\eta_1) - A_{k_0}(\eta_0) + A_{k_1}(\eta_2).$$

Taking the first derivative of $f^\circ(\theta_0)$ with respect to θ_0, we get

$$\begin{aligned}\frac{df^\circ(\theta_0)}{d\theta_0} &= \frac{dA_{k_1}(\eta_2)}{d\theta_0} - \frac{dA_{k_0}(\eta_0)}{d\theta_0} \\ &= \frac{d}{d\theta_0}\left[-r_1\log(\theta_0^* - \theta_1 p_0)\right] - \frac{d}{d\theta_0}\left[-r_0\log(1-\theta_0)\right] \\ &= \frac{\theta_1(r_1-r_0)(\theta_0^*-\theta_0)}{(\theta_0^*-\theta_1\theta_0)(1-\theta_0)}\end{aligned}$$

Since $\eta_2 = \log\frac{\theta_1\theta_0}{\theta_0^*} < 0$, then $\frac{\theta_1\theta_0}{\theta_0^*} < 1$ and $\theta_0^* - \theta_1\theta_0 > 0$. If $\theta_0 < \theta_0^*$, then $\frac{df^\circ(\theta_0)}{d\theta_0} > 0$. If $\theta_0 > \theta_0^*$, then $\frac{df^\circ(\theta_0)}{d\theta_0} < 0$. So $f^\circ(\theta_0)$ first increases and then decreases, taking the maximum value at θ_0^*. Moreover, $f^\circ(\theta_0^*) = f^\circ(\eta_0^*) = 0$, so for all $\theta_0 \in [0,1)$, $f^\circ(\theta_0) \le f^\circ(\theta_0^*) = 0$ and $\mathbb{E}_{p^\circ_{r_0,\theta_0}}[\frac{p^\circ_{r_1,\theta_1}}{p^\circ_{r_0,\theta_0^*}}] \le \mathbb{E}_{p^\circ_{r_0,\theta_0^*}}[\frac{p^\circ_{r_1,\theta_1}}{p^\circ_{r_0,\theta_0^*}}] = 1$. Then $\frac{p^\circ_{r_1,\theta_1}}{p^\circ_{r_0,\theta_0^*}}$ is an E-variable.

If $0 < r_1 < r_0$ and $0 \le \theta_0 < \frac{1}{(\frac{r_0}{r_1}-1)(1-\theta_1)+1} < 1$, then $\eta_2 < \log 1 = 0$.
The variance of the negative binomial distribution is

$$\mathrm{var}_{p_{r,\eta}}[X] = \mathrm{var}_{p^\circ_{r,\theta}}[X] = \frac{\theta r}{(1-\theta)^2}$$

so

$$\begin{aligned}\mathrm{var}_{p_{r_1,\eta_1}}[X] - \mathrm{var}_{p_{r_0,\eta_0^*}}[X] &= \mathrm{var}_{p^\circ_{r_1,\theta_1}}[X] - \mathrm{var}_{p^\circ_{r_0,\theta_0^*}}[X] \\ &= \frac{\theta_1 r_1}{(1-\theta_1)^2} - \frac{\theta_0^* r_0}{(1-\theta_0^*)^2} \\ &= \frac{\theta_1^2 r_1 (r_0 - r_1)}{(1-\theta_1)^2 r_0} > 0\end{aligned}$$

According to Lemma 5, $\log \mathbb{E}_{p^\circ_{r_0,\theta_0}}[\frac{p^\circ_{r_1,\theta_1}}{p^\circ_{r_0,\theta_0^*}}]$ takes a local minimum at θ_0. Then according to Theorem 3, $\frac{p^\circ_{r_1,\theta_1}}{p^\circ_{r_0,\theta_0^*}}$ is not an E-variable.

If $0 < r_1 < r_0$ and $0 \leq \frac{1}{(\frac{r_0}{r_1}-1)(1-\theta_1)+1} \leq \theta_0 < 1$, then $\eta_2 \geq \log 1 = 0$ and it is not included in the parameter space Θ_η. □

Chapter 4

N Outcomes of One-parameter Exponential Families

This chapter considers E-values in the context of a sample of n independent outcomes of a one-dimensional exponential family. We take as the alternative a special parameter vector $(\eta_1, \ldots, \eta_n)$. The density function of the corresponding distribution is the product of the densities of n individual one-parameter exponential families, and their canonical form is

$$p_{\eta_1,\ldots,\eta_n}(x^n) = \prod_{i=1}^{n} p_{\eta_i}(x_i) = \prod_{i=1}^{n} h(x_i) \exp(\eta_i T(x_i) - A(\eta_i))$$

with log-partition function ($i = 1, \ldots, n$)

$$\begin{aligned} A(\eta_i) &= \log \int h(x) \exp(\eta_i T(x)) dx && \text{in continuous distributions} \\ &= \log \sum h(x) \exp(\eta_i T(x)) && \text{in discrete distributions.} \end{aligned}$$

We take as the null hypothesis that $X_1, \ldots, X_n \sim$ *i.i.d.* according to P_η, where P_η is any number of the corresponding family.

The treatment in this chapter is also analogous to the treatment in Chapter 2. Lemma 6 gives us a possible simple RIPr. Theorem 4 states the conditions that make this possible simple RIPr true. Lemma 7 provides a simple condition to reject this possible simple RIPr.

Let us compute some useful derivatives. The first derivative for $A(\eta_i)$

in continuous distributions is

$$\begin{aligned}\frac{dA(\eta_i)}{d\eta_i} &= \frac{\int T(x)h(x)\exp(\eta_i T(x))dx}{\int h(x)\exp(\eta_i T(x))dx} \\ &= \int T(x)h(x)\exp(\eta_i T(x) - A(\eta_i))dx = \mathbb{E}_{p_{\eta_i}}[T(X)].\end{aligned}$$

The second derivative for $A(\eta_i)$ is

$$\begin{aligned}\frac{d^2A(\eta_i)}{d\eta_i^2} &= \int T(x)h(x)\exp(\eta_i T(x) - A(\eta_i))(T(x) - \frac{dA(\eta_i)}{d\eta_i})dx \\ &= \int T(x)h(x)\exp(\eta_i T(x) - A(\eta_i))(T(x) - \mathbb{E}_{p_{\eta_i}}[T(X)])dx \\ &= \mathbb{E}_{p_{\eta_i}}[T^2(X)] - \mathbb{E}_{p_{\eta_i}}[T(X)]\mathbb{E}_{p_{\eta_i}}[T(X)] = \text{var}_{p_{\eta_i}}[T(X)].\end{aligned}$$

We have the same results for discrete distributions.

4.1 Simple RIPr for N Outcomes of One-parameter Exponential Family

$\mathcal{P} = \{p_{\eta_1,\ldots,\eta_n} = \prod_{i=1}^n p_{\eta_i} : \eta_1,\ldots,\eta_n \in \Theta_\eta\}$ with $\Theta_\eta \subset \mathbb{R}$ denotes a sample of n n independent outcomes of a one-parameter exponential family. The null hypothesis is composite and is given by

$$\mathcal{H}_0 : X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} P_{\eta_0} \qquad \text{for varying } \eta_0 \in \Theta_\eta.$$

The alternative hypothesis is simple and is given by

$$\mathcal{H}_1 : X_1 \sim P_{\eta_{11}},\ldots,X_n \sim P_{\eta_{1n}} \text{ for fixed, and not all identical } \eta_{11},\ldots,\eta_{1n} \in \Theta_\eta.$$

As in the previous two chapters, we investigate whether the RIPr $p_{W_0^*,\ldots,W_0^*}$ is simple, i.e. $p_{W_0^*,\ldots,W_0^*} = p_{\eta_0^*,\ldots,\eta_0^*}$. Just like in the previous chapters, after finding the KL minimizing $p_{\eta_0^*,\ldots,\eta_0^*}$, we must check if it is a RIPr by checking if $\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}}$ is an E-variable.

Lemma 6. *The $\eta_0^* \in \Theta_\eta$ satisfying $\sum_{i=1}^n \mathbb{E}_{p_{\eta_{1i}}}[T(X)] = n\mathbb{E}_{p_{\eta_0^*}}[T(X)]$, if it exists, minimizes $D(p_{\eta_{11},\ldots,\eta_{1n}}||p_{\eta_0,\ldots,\eta_0})$ over Θ_η.*

Proof. For $\forall \eta_0 \in \Theta_\eta$, the KL divergence satisfies

$$\begin{aligned}
D(p_{\eta_{11},\dots,\eta_{1n}} || p_{\eta_0,\dots,\eta_0}) &= \mathbb{E}_{p_{\eta_{11},\dots,\eta_{1n}}} \left[\log \frac{p_{\eta_{11},\dots,\eta_{1n}}}{p_{\eta_0,\dots,\eta_0}} \right] \\
&= \mathbb{E}_{p_{\eta_{11},\dots,\eta_{1n}}} \left[\log \frac{\prod_{i=1}^n p_{\eta_{1i}}(X_i)}{\prod_{i=1}^n p_{\eta_0}(X_i)} \right] \\
&= \sum_{i=1}^n \mathbb{E}_{p_{\eta_{1i}}} \left[\log \frac{p_{\eta_{1i}}(X)}{p_{\eta_0}(X)} \right] \\
&= \sum_{i=1}^n \mathbb{E}_{p_{\eta_{1i}}} \left[\log \frac{h(X) \exp(\eta_{1i} T(X) - A(\eta_{1i}))}{h(X) \exp(\eta_0 T(X) - A(\eta_0))} \right] \\
&= \sum_{i=1}^n \mathbb{E}_{p_{\eta_{1i}}} [\eta_{1i} T(X) - \eta_0 T(X)] - nA(\eta_{1i}) + nA(\eta_0)
\end{aligned}$$

Taking the derivative with respect to η_0, we have

$$\begin{aligned}
\frac{dD(p_{\eta_{11},\dots,\eta_{1n}} || p_{\eta_0,\dots,\eta_0})}{d\eta_0} &= -\sum_{i=1}^n \mathbb{E}_{p_{\eta_{1i}}}[T(X)] + n \cdot \frac{dA(\eta_0)}{d\eta_0} \\
&= -\sum_{i=1}^n \mathbb{E}_{p_{\eta_{1i}}}[T(X)] + n\mathbb{E}_{p_{\eta_0}}[T(X)]
\end{aligned}$$

According to the condition, we get $\frac{dD(p_{\eta_{11},\dots,\eta_{1n}} || p_{\eta_0,\dots,\eta_0})}{d\eta_0} = 0$. Taking the second derivative with respect to η_0, we have

$$\frac{d^2 D(p_{\eta_{11},\dots,\eta_{1n}} || p_{\eta_0,\dots,\eta_0})}{d\eta_0^2} = n \cdot \frac{d^2 A(\eta_0)}{d\eta_0^2} = n \operatorname{var}_{p_{\eta_0}}[T(X)] > 0,$$

so this KL divergence is a convex function, and takes the minimum value at η_0^*. □

Theorem 4. *Consider $\eta_{(1i)} = \eta_{1i} + \eta_0 - \eta_0^*$ as a function of η_0. If $\eta_{(1i)} \in \Theta_\eta$, then*

$$\mathbb{E}_{p_{\eta_0,\dots,\eta_0}} \left[\frac{p_{\eta_{11},\dots,\eta_{1n}}}{p_{\eta_0^*,\dots,\eta_0^*}} \right] = \exp(nA(\eta_0^*) - \sum_{i=1}^n A(\eta_{1i}) - nA(\eta_0) + \sum_{i=1}^n A(\eta_{(1i)}))$$

Moreover, define $f(\eta_0) := \log \mathbb{E}_{p_{\eta_0,\dots,\eta_0}} \left[\frac{p_{\eta_{11},\dots,\eta_{1n}}}{p_{\eta_0^*,\dots,\eta_0^*}} \right]$ as a function of η_0 (with fixed $\eta_{11}, \dots, \eta_{1n}$). Then $f(\eta_0^*)$ is a local maximum or minimum value.

If $f(\eta_0)$ takes the maximum value at η_0^*, then $\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}}$ is an E-variable, i.e.

$$\mathbb{E}_{p_{\eta_0,\ldots,\eta_0}}\left[\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}}\right] \leq 1, \qquad \forall \eta_0 \in \Theta_\eta.$$

Otherwise, if $f(\eta_0)$ does not take the maximum value at η_0^*, in particular, if $f(\eta_0^*)$ is a local minimum value, then $\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}}$ is not an E-variable and there does not exist a simple RIPr.

Proof.

$$\begin{aligned}
\mathbb{E}_{p_{\eta_0,\ldots,\eta_0}}\left[\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}}\right] &= \mathbb{E}_{p_{\eta_0,\ldots,\eta_0}}\left[\prod_{i=1}^{n}\frac{p_{\eta_{1i}}(X_i)}{p_{\eta_0^*}(X_i)}\right] \\
&= \prod_{i=1}^{n}\mathbb{E}_{p_{\eta_{1i}}}\left[\frac{p_{\eta_{1i}}(X)}{p_{\eta_0^*}(X)}\right] \quad \text{since } x_1,\ldots,x_n \text{ are independent} \\
&= \prod_{i=1}^{n}\int \frac{h(x)\exp(\eta_{1i}T(x)-A(\eta_{1i}))}{h(x)\exp(\eta_0^*T(x)-A(\eta_0^*))}h(x) \\
&\quad \cdot \exp(\eta_0 T(x)-A(\eta_0))dx \\
&= \prod_{i=1}^{n}\int h(x)\exp((\eta_{1i}+\eta_0-\eta_0^*)T(x)-A(\eta_{1i}) \\
&\quad -A(\eta_0)+A(\eta_0^*))dx \\
&= \prod_{i=1}^{n}\int h(x)\exp(\eta_{(i)}T(x)-A(\eta_{1i})-A(\eta_0)+A(\eta_0^*))dx \\
&= \prod_{i=1}^{n}[\exp(A(\eta_0^*)-A(\eta_{1i})-A(\eta_0)+A(\eta_{(1i)})) \\
&\quad \cdot \int h(x)\exp(\eta_{(1i)}T(x)-A(\eta_{(1i)}))dx] \\
&= \prod_{i=1}^{n}\exp(A(\eta_0^*)-A(\eta_{1i})-A(\eta_0)+A(\eta_{(1i)})\cdot 1 \qquad (4.1) \\
&= \exp(nA(\eta_0^*)-\sum_{i=1}^{n}A(\eta_{1i})-nA(\eta_0)+\sum_{i=1}^{n}(A(\eta_{(1i)}))
\end{aligned}$$

In (4.1), $h(x)\exp(\eta_{(1i)}T(x)-A(\eta_{(1i)}))$ is the density of the one-parameter exponential family with parameter $\eta_{(1i)}$, and the integral of a probability is equal to 1.

$\eta_{11}, \ldots, \eta_{1n}$ are fixed and denote the logarithm of the expectation by $f(\eta_0)$

$$f(\eta_0) = \log \mathbb{E}_{p_{\eta_0,\ldots,\eta_0}} \left[\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}} \right] = nA(\eta_0^*) - \sum_{i=1}^{n} A(\eta_{1i}) - nA(\eta_0) + \sum_{i=1}^{n} A(\eta_{(1i)})$$

When $\eta_0 = \eta_0^*$, we have $f(\eta_0^*) = 0$ and $\mathbb{E}_{p_{\eta_0^*,\ldots,\eta_0^*}} \left[\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}} \right] = 1$.

Taking the derivative with respect to η_0, we get

$$\begin{aligned} \frac{df(\eta_0)}{d\eta_0} &= \frac{d}{d\eta_0}(\sum_{i=1}^{n}(A(\eta_{(1i)}) - nA(\eta_0)) \\ &= \sum_{i=1}^{n} \mathbb{E}_{p_{\eta_{(1i)}}}[T(X)] - n\mathbb{E}_{p_{\eta_0}}[T(X)] \end{aligned}$$

When $\eta_0 = \eta_0^*$, $\frac{df(\eta_0)}{d\eta_0} = 0$, so $f(\eta_0)$ is a local maximum or minimum value.

If $f(\eta_0)$ takes the maximum value at η_0^*, then for $\forall\ \eta_0 \in \Theta_\eta$, $f(\eta_0) \leq 0$ and $\mathbb{E}_{p_{\eta_0,\ldots,\eta_0}} \left[\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}} \right] \leq 1$.

If $f(\eta_0)$ does not take the maximum value at η_0^*, then there is a η_0 such that $f(\eta_0) > f(\eta_0^*) = 0$ and $\mathbb{E}_{p_{\eta_0,\ldots,\eta_0}} \left[\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}} \right] > \mathbb{E}_{p_{\eta_0^*,\ldots,\eta_0^*}} \left[\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}} \right] = 1$. □

Lemma 7. *$f(\eta_0) := \log \mathbb{E}_{p_{\eta_0,\ldots,\eta_0}} \left[\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}} \right]$ takes a local minimum value at η_0^* iff*

$$\sum_{i=1}^{n} \mathrm{var}_{p_{\eta_{1i}}}[T(X)] - n\mathrm{var}_{p_{\eta_0^*}}[T(X)] > 0$$

Proof. We know $\frac{f(\eta_0)}{d\eta_0}|_{\eta_0=\eta_0^*} = 0$ in the proof of Theorem 4. Taking the second derivative of $f(\eta_0)$ with respect to η_0, we have

$$\begin{aligned} \frac{d^2 f(\eta_0)}{d\eta_0^2} &= \frac{d^2}{d\eta_0^2}(\sum_{i=1}^{n}(A(\eta_{(1i)}) - nA(\eta_0)) \\ &= \sum_{i=1}^{n} \mathrm{var}_{p_{\eta_{(1i)}}}[T(X)] - n\mathrm{var}_{p_{\eta_0}}[T(X)]. \end{aligned}$$

According to the condition of the lemma, we obtain

$$\frac{d^2 f(\eta_0)}{d\eta_0^2}|_{\eta_0=\eta_0^*} = \sum_{i=1}^{n} \mathrm{var}_{p_{\eta_{1i}}}[T(X)] - n\mathrm{var}_{p_{\eta_0^*}}[T(X)] > 0,$$

then $f(\eta_0^*)$ is a local minimum value. □

4.2 Examples

This section discusses two examples.

4.2.1 Example: The Poisson Distribution

Poisson distributions form a one-parameter exponential family. The probability of a sample of n outcomes from n Poisson distributions is equal to the product of the probabilities of n independent Poisson distributions Pois(λ_i) ($\lambda_i > 0$, $i = 1, \ldots, n$). The density function, with λ_i denoting the parameter in the standard parameterization, is

$$p^{\circ}_{\lambda_1,\ldots,\lambda_n}(x^n) = \prod_{i=1}^{n} p^{\circ}_{\lambda_i}(x_i) = \prod_{i=1}^{n} \frac{\lambda_i^{x_i} \exp(-\lambda_i)}{x_i!}, \quad x_i \in \mathbb{N} \ (i = 1, \ldots, n).$$

The mean and variance of a Poisson distribution are

$$\mathbb{E}_{p^{\circ}_{\lambda_i}}[X] = \mathrm{var}_{p^{\circ}_{\lambda_i}}[X] = \lambda_i.$$

Transforming this n-sample density into the canonical form, we get

$$p^{\circ}_{\lambda_1,\ldots,\lambda_n}(x^n) = p_{\eta_1,\ldots,\eta_n}(x^n) = \prod_{i=1}^{n} p_{\eta_i}(x_i) = \prod_{i=1}^{n} h(x_i) \exp(\eta_i T(x_i) - A(\eta_i))$$

with $h(x_i) = \frac{1}{x_i!}$, natural parameter $\eta_i = \log \lambda_i$, sufficient statistic $T(x_i) = x_i$, and log-partition function

$$A(\eta_i) = A^{\circ}(\lambda_i) = \lambda_i = \exp(\eta_i).$$

The space of natural parameter η_i is $\Theta_{\eta_i} = \{\eta_i : \eta_i \in \mathbb{R}\}$.

Hypotheses: The null hypothesis $\mathcal{H}_0$ says that $X_1, \ldots, X_n$ are i.i.d. samples from a Poisson distribution with unknown parameter $\lambda_0 > 0$, hence the null hypothesis is composite. The density in the canonical form is $p_{\eta_0,\ldots,\eta_0}$ with $\eta_0 \in \mathbb{R}$. The alternative hypothesis is a sample of n outcomes from n Poisson distributions whose parameters $\lambda_{11}, \ldots, \lambda_{1n} (> 0)$ are fixed and not all identical. The density in the canonical form is $p_{\eta_{11},\ldots,\eta_{1n}}$ with not all identical natural parameters $\eta_{11}, \ldots, \eta_{1n} < 0$.

Result: Under these hypotheses, there is a simple RIPr $p^{\circ}_{\lambda_0^*,\ldots,\lambda_0^*}$ where $\lambda_0^* = \frac{1}{n} \sum_{i=1}^{n} \lambda_{1i}$ and $\frac{p^{\circ}_{\lambda_{11},\ldots,\lambda_{1n}}}{p^{\circ}_{\lambda_0^*,\ldots,\lambda_0^*}}$ is an E-variable, i.e.

$$\mathbb{E}_{p^{\circ}_{\lambda_0,\ldots,\lambda_0}}\left[\frac{p^{\circ}_{\lambda_{11},\ldots,\lambda_{1n}}}{p^{\circ}_{\lambda_0^*,\ldots,\lambda_0^*}}\right] = 1, \qquad \text{for } \forall\ \lambda_0 > 0.$$

Proof. We know the mean of an Poisson distribution is

$$\mathbb{E}_{p_{\eta_i}}[X] = \mathbb{E}_{p^\circ_{\lambda_i}}[X] = \lambda_i.$$

From the condition $\lambda_0^* = \frac{1}{n}\sum_{i=1}^n \lambda_{1i}$, we obtain

$$\sum_{i=1}^n \mathbb{E}_{p_{\eta_{1i}}}[X] = \sum_{i=1}^n \mathbb{E}_{p^\circ_{\lambda_{1i}}}[X] = \sum_{i=1}^n \lambda_{1i} = n\lambda_0^* = n\mathbb{E}_{p^\circ_{\lambda_0^*}}[X] = n\mathbb{E}_{p_{\eta_0^*}}[X].$$

According to Lemma 6, KL divergence $D(p^\circ_{\lambda_{11},\ldots,\lambda_{1n}} || p^\circ_{\lambda_0,\ldots,\lambda_0})$ takes the minimum value when $\lambda_0 = \lambda_0^*$.

Moreover,

$$\eta_{(1i)} = \eta_{1i} + \eta_0 - \eta_0^* = \log\lambda_{1i} + \log\lambda_0 - \log\lambda_0^* \ \in \Theta_{\eta_i}.$$

We have

$$A(\eta_{(1i)}) = \exp(\eta_{(1i)}) = \frac{\lambda_{1i}\lambda_0}{\lambda_0^*}.$$

According to Theorem 4, since $A(\eta_i) = A(\lambda_i) = \lambda_i$, then we obtain

$$\begin{aligned}
\mathbb{E}_{p_{\eta_0,\ldots,\eta_0}}\left[\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}}\right] &= \exp(nA(\eta_0^*) - \sum_{i=1}^n A(\eta_{1i}) - nA(\eta_0) + \sum_{i=1}^n A(\eta_{(1i)})) \\
&= \exp(n\lambda_0^* - \sum_{i=1}^n \lambda_{1i} - n\lambda_0 + \sum_{i=1}^n \frac{\lambda_{1i}\lambda_0}{\lambda_0^*}) \\
&= \exp(0) = 1,
\end{aligned}$$

so $\frac{p_{\eta_{11},\ldots,\eta_{1n}}}{p_{\eta_0^*,\ldots,\eta_0^*}}$ is an E-variable. □

4.2.2 Example: The Exponential Distribution

Exponential distributions form a one-parameter exponential family. For simplicity, we consider a sample of two independent outcomes of the exponential distribution, whose density is the product of densities of 2 independent exponential distributions with parameters λ_1 and λ_2. In the standard parameterization, the density of this distribution is

$$p^\circ_{\lambda_1,\lambda_2}(x^2) = p^\circ_{\lambda_1}(x_1)p^\circ_{\lambda_2}(x_2) = \lambda_1\exp(-\lambda_1 x_1)\cdot\lambda_2\exp(-\lambda_2 x_2),$$

where $x_1, x_2 > 0$.
The mean and variance of an exponential distribution with parameter λ_j are

$$\mathbb{E}_{p^\circ_{\lambda_j}}[X] = \frac{1}{\lambda_j} \qquad \text{and} \qquad \text{var}_{p^\circ_{\lambda_j}}[X] = \frac{1}{\lambda_j^2}.$$

The canonical form is as below:

$$p^{\circ}_{\lambda_1,\lambda_2}(x^2) = p_{\eta_1,\eta_2}(x^2) = p_{\eta_1}(x_1)p_{\eta_2}(x_2) = \prod_{i=1}^{2} h(x_i)\exp(\eta_i T(x_i) - A(\eta_i))$$

with $h(x_i) = 1$, natural parameter $\eta_i = -\lambda_i$, sufficient statistic $T(x_i) = x_i$ and log-partition function

$$A(\eta_i) = A^{\circ}(\lambda_i) = -\log\lambda_i = -\log(-\eta_i).$$

The natural parameter space is $\Theta_{\eta_i} = \{\eta_i : \eta_i < 0\}$.

Hypotheses: The null hypothesis $\mathcal{H}_0$ says that X_1, X_2 are i.i.d. samples from an exponential distribution with parameter $\lambda_0 > 0$. In the canonical form, the density is p_{η_0,η_0} with $\eta_0 < 0$. The alternative hypothesis says that X_1, X_2 is a sample of two outcomes from two exponential distributions with fixed parameters $\lambda_{11} > \lambda_{12} > 0$. In the canonical form, the density is $p_{\eta_{11},\eta_{12}}$ with fixed $\eta_{11} < \eta_{12} < 0$.

Result: If $\lambda_0 > \lambda_0^* - \lambda_{12}$ and $\lambda_0^* = \frac{2\lambda_{11}\lambda_{12}}{\lambda_{11}+\lambda_{12}}$, then $\frac{p^{\circ}_{\lambda_{11},\lambda_{12}}}{p^{\circ}_{\lambda_0^*,\lambda_0^*}}$ is not an E-variable and there does not exist a simple RIPr.
If $0 < \lambda_0 < \lambda_0^* - \lambda_{12}$, then $\mathbb{E}_{p^{\circ}_{\lambda_0,\lambda_0}}\left[\frac{p^{\circ}_{\lambda_{11},\lambda_{12}}}{p^{\circ}_{\lambda_0^*,\lambda_0^*}}\right]$ is not well-defined.

Proof. The mean of an exponential distribution is

$$\mathbb{E}_{p_{\eta_j}}[X] = \mathbb{E}_{p^{\circ}_{\lambda_j}}[X] = \frac{1}{\lambda_j}.$$

According to the condition of λ_0^*,

$$\begin{aligned}\mathbb{E}_{p_{\eta_{11}}}[X] + \mathbb{E}_{p_{\eta_{12}}}[X] = \mathbb{E}_{p^{\circ}_{\lambda_{11}}}[X] + \mathbb{E}_{p^{\circ}_{\lambda_{12}}}[X] = \frac{1}{\lambda_{11}} + \frac{1}{\lambda_{12}} = \frac{2}{\lambda_0^*} \\ = 2\mathbb{E}_{p^{\circ}_{\lambda_0^*}}[X] = 2\mathbb{E}_{p_{\eta_0^*}}[X].\end{aligned}$$

By Lemma 6, this λ_0^* minimizes KL divergence $D(p^{\circ}_{\lambda_{11},\lambda_{12}}||p^{\circ}_{\lambda_0,\lambda_0})$.
In addition, we have

$$\lambda_{11} = \frac{\lambda_{11}^2 + \lambda_{11}\lambda_{12}}{\lambda_{11}+\lambda_{12}} > \frac{2\lambda_{11}\lambda_{12}}{\lambda_{11}+\lambda_{12}} = \lambda_0^* > \frac{\lambda_{12}^2 + \lambda_{11}\lambda_{12}}{\lambda_{11}+\lambda_{12}} = \lambda_{12},$$

and

$$\eta_{(1i)} = \eta_{1i} + \eta_0 - \eta_0^* = -\lambda_{1i} - \lambda_0 + \lambda_0^* \qquad (i = 1, 2).$$

If $\lambda_0 > \lambda_0^* - \lambda_{12}$, then $\eta_{(11)} < \eta_{(12)} < 0$. The variance of an exponential distribution is

$$\text{var}_{p_{\eta_j}}[X] = \text{var}_{p^\circ_{\lambda_j}}[X] = \frac{1}{\lambda_j^2}.$$

so

$$\begin{aligned}\text{var}_{p_{\eta_{11}}}[X] + \text{var}_{p_{\eta_{12}}}[X] - 2\text{var}_{p_{\eta_0^*}}[X] &= \text{var}_{p^\circ_{\lambda_{11}}}[X] + \text{var}_{p^\circ_{\lambda_{12}}}[X] - 2\text{var}_{p^\circ_{\lambda_0^*}}[X] \\ &= \frac{1}{\lambda_{11}^2} + \frac{1}{\lambda_{12}^2} - \frac{2}{\lambda_0^{*2}} \\ &= \frac{(\lambda_{11} - \lambda_{12})^2}{2\lambda_{11}^2\lambda_{12}^2} > 0\end{aligned}$$

According to Lemma 7, $\mathbb{E}_{p^\circ_{\lambda_0,\lambda_0}}\left[\frac{p^\circ_{\lambda_{11},\lambda_{12}}}{p^\circ_{\lambda_0^*,\lambda_0^*}}\right]$ takes a local minimum value at λ_0. According to Theorem 4, $\frac{p^\circ_{\lambda_{11},\lambda_{12}}}{p^\circ_{\lambda_0^*,\lambda_0^*}}$ is not an E-variable.

If $0 < \lambda_0 < \lambda_0^* - \lambda_{12}$, then $\eta_{(11)} < 0 < \eta_{(12)}$ and $\eta_{(12)}$ is not included in the parameter space $\Theta_{\eta_{12}}$. □

Chapter 5

Transformation of Random Variables

In this chapter, we discuss the situation when the random variable X is transformed into a new random variable Y, where X follows an exponential family distribution. We denote a d-dimensional exponential family by $\mathcal{P} = \{p_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \Theta_{\boldsymbol{\eta}}^d\}$ with $\Theta_{\boldsymbol{\eta}}^d \subset \mathbb{R}^d$. The definition of $p_{\boldsymbol{\eta}}$ was shown in Definition 1. We consider two exponential families, for $k = 0, 1$: $\mathcal{P}_k = \{p_{k,\boldsymbol{\eta}_k} : \boldsymbol{\eta}_k \in \Theta_k^d\}$ with $\Theta_k^d \in \mathbb{R}^d$. The null hypothesis $\mathcal{H}_0$ for X is composite and is given as

$$\mathcal{H}_0 : X \sim P_{0,\boldsymbol{\eta}_0} \qquad \text{for varying } \boldsymbol{\eta}_0 \in \Theta_0^d.$$

The alternative $\mathcal{H}_1$ for X is simple and is given that

$$\mathcal{H}_1 : X \sim P_{1,\boldsymbol{\eta}_1} \qquad \text{for a fixed } \boldsymbol{\eta}_1 \in \Theta_1^d.$$

We suppose the new random variable Y satisfies $Y = h(X)$ and their distribution is $Q_{k,\boldsymbol{\eta}_k}$ with density $q_{k,\boldsymbol{\eta}_k}$ $(k = 0, 1)$. The new null hypothesis $\mathcal{H}_0'$ and new alternative hypothesis $\mathcal{H}_1'$ are given by

$$\mathcal{H}_0' : Y \sim Q_{0,\boldsymbol{\eta}_0} \qquad \text{for varying } \boldsymbol{\eta}_0 \in \Theta_0^d,$$
$$\mathcal{H}_1' : Y \sim Q_{1,\boldsymbol{\eta}_1} \qquad \text{for a fixed } \boldsymbol{\eta}_1 \in \Theta_1^d.$$

Theorem 5. *If $y = h(x)$ is a continuous function that piecewise has a differentiable inverse (i.e, in each piece, it is strictly monotonic and differentiable), and $\frac{p_{1,\boldsymbol{\eta}_1}(x)}{p_{0,\boldsymbol{\eta}_0^*}(x)}$ is an E-variable, that is*

$$\mathbb{E}_{p_{0,\boldsymbol{\eta}_0}}\left[\frac{p_{1,\boldsymbol{\eta}_1}(X)}{p_{0,\boldsymbol{\eta}_0^*}(X)}\right] \leq 1, \qquad \forall \boldsymbol{\eta}_0 \in \Theta_0^d.$$

Then $\frac{q_{1,\eta_1}(y)}{q_{0,\eta_0^*}(y)}$ *is also an E-variable, i.e.*

$$\mathbb{E}_{q_{0,\eta_0}}\left[\frac{q_{1,\eta_1}(Y)}{q_{0,\eta_0^*}(Y)}\right] \leq 1, \qquad \forall \boldsymbol{\eta}_0 \in \Theta_0^d.$$

Proof. Firstly, we consider $y = h(x)$ to be strictly monotonically increasing and differentiable, then their unique inverse function $x = g(y)$ is also strictly monotonically increasing and differentiable, i.e, $g'(x) > 0$. The cumulative distribution function is

$$F_Y(y) = Pr(Y \leq y) = Pr(h(X) \leq y) = Pr(X \leq g(y)) = F_X(g(y)).$$

Then the density function is

$$q_{k,\eta_k}(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(g(y)) = F_X'(g(y)) \cdot g'(y) = p_{k,\eta_k}(g(y)) \cdot g'(y).$$

So

$$\begin{aligned}
\mathbb{E}_{q_{0,\eta_0}}\left[\frac{q_{1,\eta_1}(Y)}{q_{0,\eta_0^*}(Y)}\right] &= \int \frac{q_{1,\eta_1}(y)}{q_{0,\eta_0^*}(y)} \cdot q_{0,\eta_0}(y)dy \\
&= \int \frac{p_{1,\eta_1}(g(y)) \cdot g'(y)}{p_{0,\eta_0^*}(g(y)) \cdot g'(y)} \cdot p_{0,\eta_0}(g(y)) \cdot g'(y)dy \\
&= \int \frac{p_{1,\eta_1}(g(y))}{p_{0,\eta_0^*}(g(y))} \cdot p_{0,\eta_0}(g(y))d(g(y)) \\
&= \int \frac{p_{1,\eta_1}(x)}{p_{0,\eta_0^*}(x)} \cdot p_{0,\eta_0}(x)dx \\
&= \mathbb{E}_{p_{0,\eta_0}}\left[\frac{p_{1,\eta_1}(X)}{p_{0,\eta_0^*}(X)}\right] \leq 1
\end{aligned}$$

Secondly, we consider $y = h(x)$ to be strictly monotonically decreasing and differentiable. Set $x \in (a_0, a_1)$, then $y \in (h(a_1), h(a_0))$. The unique inverse function $x = g(y)$ is also strictly monotonically decreasing and differentiable, i.e, $g'(x) < 0$. And $g(h(a_1)) = a_1$, $g(h(a_0)) = a_0$. The cumulative distribution function is

$$\begin{aligned}
F_Y(y) &= Pr(Y \leq y) = Pr(h(X) \leq y) = Pr(X \geq g(y)) \\
&= 1 - Pr(X \leq g(y)) = 1 - F_X(g(y)).
\end{aligned}$$

Then the density function is

$$\begin{aligned} q_{k,\eta_k}(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy}\left[1 - F_X(g(y))\right] \\ &= -F_X'(g(y)) \cdot g'(y) = -p_{k,\eta_k}(g(y)) \cdot g'(y). \end{aligned}$$

So

$$\begin{aligned} \mathbb{E}_{q_{0,\eta_0}}\left[\frac{q_{1,\eta_1}(Y)}{q_{0,\eta_0^*}(Y)}\right] &= \int_{h(a_1)}^{h(a_0)} \frac{q_{1,\eta_1}(y)}{q_{0,\eta_0^*}(y)} \cdot q_{0,\eta_0}(y)dy \\ &= \int_{h(a_1)}^{h(a_0)} \frac{-p_{1,\eta_1}(g(y)) \cdot g'(y)}{-p_{0,\eta_0^*}(g(y)) \cdot g'(y)} \cdot \left[-p_{0,\eta_0}(g(y)) \cdot g'(y)dy\right] \\ &= \int_{a_1}^{a_0} -\frac{p_{1,\eta_1}(g(y))}{p_{0,\eta_0^*}(g(y))} \cdot p_{0,\eta_0}(g(y))d(g(y)) \\ &= \int_{a_0}^{a_1} \frac{p_{1,\eta_1}(x)}{p_{0,\eta_0^*}(x)} \cdot p_{0,\eta_0}(x)dx \\ &= \mathbb{E}_{p_{0,\eta_0}}\left[\frac{p_{1,\eta_1}(X)}{p_{0,\eta_0^*}(X)}\right] \leq 1. \end{aligned}$$

Taking together the two conditions above, if $y = h(x)$ is a strictly monotone continuous function and is differentiable, then it has a unique inverse function $x = g(x)$, and the density function of Y is

$$q_{k,\eta_k}(y) = p_{k,\eta_k}(g(y)) \cdot |g'(y)|. \tag{5.1}$$

Moreover,

$$\mathbb{E}_{q_{0,\eta_0}}\left[\frac{q_{1,\eta_1}(Y)}{q_{0,\eta_0^*}(Y)}\right] = \mathbb{E}_{p_{0,\eta_0}}\left[\frac{p_{1,\eta_1}(X)}{p_{0,\eta_0^*}(X)}\right] \leq 1.$$

Finally, we consider $y = h(x)$ as a continuous function that piecewise has a differentiable inverse (i.e, in each piece, it is strictly monotonic and differentiable). Set $x \in (a_0, a_n)$. If $h'(x) = 0$ at some positions $x = a_j \ \ (j = 1, \ldots, n-1)$, then the definition space of x (i.e. set (a_0, a_n)) can be split into n disjoint intervals $I_j = (a_{j-1}, a_j) \ \ (j = 1, \ldots, n)$ and $\bigcup_{j=1}^{n} I_j = (a_0, a_n)$. So when $x_j \in I_j$, $y_j = h_j(x_j)$ is strictly monotonic and differentiable, then the corresponding inverse function $x_j = g_j(y_j) \ \ (x_j \in I_j)$ is also strictly monotonic and differentiable. The cumulative distribution function is

$$F_Y(y) = Pr(Y \leq y) = \sum_{j=1}^{n} Pr(h_j(X_j) \leq y).$$

From (5.1), we get the density

$$q_{k,\eta_k}(y) = f_Y(y) = \sum_{j=1}^{n} p_{k,\eta_k}(g_j(y_j))|g_j'(y_j)|.$$

So

$$\begin{aligned}
\mathbb{E}_{q_{0,\eta_0}}\left[\frac{q_{1,\eta_1}(Y)}{q_{0,\eta_0^*}(Y)}\right] &= \sum_{j=1}^{n} \mathbb{E}_{q_{0,\eta_0}}\left[\frac{q_{1,\eta_1}(Y_j)}{q_{0,\eta_0^*}(Y_j)}\right] \\
&= \sum_{j=1}^{n} \mathbb{E}_{p_{0,\eta_0}}\left[\frac{p_{1,\eta_1}(X_j)}{p_{0,\eta_0^*}(X_j)}\right] \\
&= \mathbb{E}_{p_{0,\eta_0}}\left[\frac{p_{1,\eta_1}(X)}{p_{0,\eta_0^*}(X)}\right] \leq 1.
\end{aligned}$$

□

5.1 Example

5.1.1 Example: The Exponential Distributions and the Pareto Distributions

As in the setting in Example 4.2.2, random variables X_1, X_2 are two independent outcomes of the exponential distribution. The composite null hypothesis $\mathcal{H}_0$ and the alternative hypothesis $\mathcal{H}_1$ for X_1, X_2 are given by

$$\begin{aligned}
&\mathcal{H}_0 : X_1, X_2 \overset{\text{i.i.d.}}{\sim} P^{\circ}_{\lambda_0} \qquad \text{for varying } \lambda_0 > 0, \\
&\mathcal{H}_1 : X_1 \sim P^{\circ}_{\lambda_{11}}, X_2 \sim P^{\circ}_{\lambda_{12}} \qquad \text{for fixed } \lambda_{11} > \lambda_{12} > 0.
\end{aligned}$$

Transform X_1, X_2 to new random variables Y_1, Y_2 by function $y = h(x) = y_{min} \cdot \exp x$ where y_{min} is fixed. Then Y_j follows a Pareto distribution with the same parameter as the corresponding X_j. The density with parameter λ_i of Y is

$$q^{\circ}_{\lambda_i} = \frac{\lambda_i y_{min}^{\lambda_i}}{y^{\lambda_i+1}} \qquad \text{where } y_{min} \text{ is fixed.}$$

So Y follows the Pareto distribution with parameter λ_i and fixed y_{min}. The new composite null hypothesis $\mathcal{H}_0'$ and the new alternative hypothesis $\mathcal{H}_1'$

for Y_1, Y_2 are given by

$$\begin{aligned}
&\mathcal{H}_0' : Y_1, Y_2 \overset{\text{i.i.d.}}{\sim} Q^\circ_{\lambda_0} \qquad \text{for varying } \lambda_0 > 0, \\
&\mathcal{H}_1' : Y_1 \sim Q_{\lambda_{11}}, Y_2 \sim Q^\circ_{\lambda_{12}} \qquad \text{for fixed } \lambda_{11} > \lambda_{12} > 0.
\end{aligned}$$

Since $y = h(x) = y_{min} \cdot \exp x$ is a monotonic continuous function and is differentiable, according to Theorem 5, if $\frac{p^\circ_{\lambda_{11},\lambda_{12}}}{p^\circ_{\lambda_0^*,\lambda_0^*}}$ is an E-variable, then $\frac{q^\circ_{\lambda_{11},\lambda_{12}}}{q^\circ_{\lambda_0^*,\lambda_0^*}}$ is also an E-variable. However, it is not an E-variable from the results of Example 4.2.2, so it can not show that there exists a simple RIPr for the Pareto distribution with two independent outcomes each with a different parameter‘.

Chapter 6

Two Outcomes of Exponential Distributions in a Specific Case

This chapter uses programming in R to approximate the RIPr E-variable for Example 4.2.2: a sample of two independent outcomes from exponential distributions each with a specific parameter value. As was seen, there does not exist a simple RIPr, i.e., single point probability distribution $p_{\lambda_0^*,\lambda_0^*}$. To approximate the RIPr anyhow, we set the probability p°_{0n} to be the weighted sum of $n(\geq 2)$ distinct probabilities $p^\circ_{\lambda_{0i},\lambda_{0i}}$ ($\lambda_{0i} \in \Theta_\lambda = \mathbb{R}_+, i = 1..n$) where the weights $w_i \in \Theta_w = \{w_i : 0 < w_i < 1\}$ corresponding to probability $p^\circ_{\lambda_{0i},\lambda_{0i}}$ sum to 1 ($i = 1, \ldots, n$), i.e.

$$p^\circ_{0n} = \sum_{i=1}^{n} w_i \cdot p^\circ_{\lambda_{0i},\lambda_{0i}} \qquad \text{and } \sum_{i=1}^{n} w_i = 1,$$

where $p^\circ_{\lambda_{0i},\lambda_{0i}}$ is the product of the probabilities of exponential distributions with parameter λ_{0i}, that is

$$p^\circ_{\lambda_{0i},\lambda_{0i}} = p^\circ_{\lambda_{0i}}(x_1) \cdot p^\circ_{\lambda_{0i}}(x_2) = \lambda_{0i} \exp(-\lambda_{0i} x_1) \cdot \lambda_{0i} \exp(-\lambda_{0i} x_2).$$

and $\boldsymbol{p^\circ_{0n}}$ is the set of all p°_{0n} that can be written in this manner. If there exists

$$p_0^{\circ *} := \arg \min_{p^\circ_{0n} \in \boldsymbol{p^\circ_{0n}}, n \geq 1} D(p^\circ_{\lambda_{11},\lambda_{12}} || p^\circ_{0n}),$$

then $p_0^{\circ *}$ must be the RIPr and $\frac{p^\circ_{\lambda_{11},\lambda_{12}}}{p_0^{\circ *}}$ must be an E-variable by Theorem 1 from (Grünwald et al., 2020), i.e.

$$\mathbb{E}_{p^\circ_{\lambda_0,\lambda_0}} \left[\frac{p^\circ_{\lambda_{11},\lambda_{12}}}{p_0^{\circ *}} \right] \leq 1 \qquad \forall \lambda_0 \in \Theta_\lambda. \tag{6.1}$$

While we do not succeed in finding such $p_0^{\circ *}$, we will approximate it with the minimum above restricted to some large but finite n. We may then expect that this satisfies (6.1) up to a small additive constant ϵ that goes to 0 as n gets larger.

For convenience, we simplify the function of KL divergence as follows:

$$\begin{aligned} D(p^{\circ}_{\lambda_{11},\lambda_{12}}||p^{\circ}_{0n}) &= \mathbb{E}_{p^{\circ}_{\lambda_{11},\lambda_{12}}}\left[\log \frac{p^{\circ}_{\lambda_{11},\lambda_{12}}}{p^{\circ}_{0n}}\right] \\ &= \mathbb{E}_{p^{\circ}_{\lambda_{11},\lambda_{12}}}[\log p^{\circ}_{\lambda_{11},\lambda_{12}}] - \mathbb{E}_{p^{\circ}_{\lambda_{11},\lambda_{12}}}[\log p^{\circ}_{0n}] \\ &= \mathbb{E}_{p^{\circ}_{\lambda_{11}}}[\log p^{\circ}_{\lambda_{11}}] + \mathbb{E}_{p^{\circ}_{\lambda_{12}}}[\log p^{\circ}_{\lambda_{12}}] - \mathbb{E}_{p^{\circ}_{\lambda_{11},\lambda_{12}}}[\log p^{\circ}_{0n}] \\ &= \log\lambda_{11} + \log\lambda_{12} - 2 - \mathbb{E}_{p^{\circ}_{\lambda_{11},\lambda_{12}}}[\log p^{\circ}_{0n}]. \end{aligned}$$

Since λ_{11} and λ_{12} are given, we find that the minimum value of KL divergence depends on the minimum value of $-\mathbb{E}_{p^{\circ}_{\lambda_{11},\lambda_{12}}}[\log p^{\circ}_{0n}]$.

We use an iterative method to approximate $p_0^{\circ *}$ to avoid too much computation. We assume $n = h$ in iteration h. $p^{\circ *}_{0h}$ denotes p°_{0h} that minimizes the KL divergence in iteration h and has parameters $\boldsymbol{w}^* = (w_1^*, \ldots, w_h^*)$ and $\boldsymbol{\lambda}_0^* = (\lambda_{01}^*, \ldots, \lambda_{0h}^*)$. The probability $p^{\circ *}_{0h}$ is based on $p^{\circ *}_{0(h-1)}$ of the last iteration and we only need to find a λ^* and a w^* in each iteration. We have the general form of probability p°_{0h} of iteration $h(\geq 2)$ which is easy to prove by mathematical induction.

$$\begin{aligned} p^{\circ *}_{0h} &= w_h \cdot p^{\circ *}_{0(h-1)} + (1 - w_h) \cdot p^{\circ}_{\lambda_{0h},\lambda_{0h}} \\ &= w_h \left[p^{\circ}_{\lambda_{01}^*,\lambda_{01}^*} \cdot \prod_{i=1}^{h-1} w_i^* + \sum_{i=2}^{h-2} \left((1 - w_i^*) p^{\circ}_{\lambda_{0i}^*,\lambda_{0i}^*} \cdot \prod_{j=i+1}^{h-1} w_j^* \right) \right. \\ &\quad \left. + (1 - w^*_{(h-1)}) p^{\circ}_{\lambda^*_{0(h-1)},\lambda^*_{0(h-1)}} \right] + (1 - w_h) \cdot p^{\circ}_{\lambda_{0h},\lambda_{0h}}. \end{aligned}$$

The iteration stops when the following conditions are satisfied:

$$\mathbb{E}_{p^{\circ}_{\lambda_0,\lambda_0}}\left[\frac{p^{\circ}_{\lambda_{11},\lambda_{12}}}{p^{\circ *}_{0h}}\right] \leq 1, \forall \lambda_0 \in \Theta_\lambda.$$

In practice, the optimal $p^{\circ *}$ may not be expressible as a mixture with a finite number of components; or the number of components may be too large to be found in the available time. For that reason, in practice, we will stop if the above holds up to some small $\epsilon > 0$.

6.1 Specific Steps in R

Now, we introduce the specific steps in programming R.

Step 1: Compute exact λ_{01}^*

In iteration 1, we have $p_{01}^\circ = w_1 \cdot p_{\lambda_{01},\lambda_{01}}^\circ$, where w_1 is equal to 1. We get $w_1^* = w_1 = 1$ and exact $\lambda_{01}^* = \frac{2\lambda_{11}\lambda_{12}}{\lambda_{11}+\lambda_{12}}$ easily.

Step 2: Narrow the value range of w_i and λ_{0i}

We choose n_w evenly spaced values of possible w_i from 0 to 1 and n_λ evenly spaced values of possible λ_{0i} from 0 to $4\lambda_{11}\lambda_{12}$. The more subdivided w_i and λ_{0i} are, the closer the approximate value of w_i^* and λ_{0i}^* is to the exact value. We do 10 iterations to get rough $w_2^*, \ldots, w_{10}^*$ and $\lambda_{02}^*, \ldots, \lambda_{010}^*$.

In iteration 2, we add new parameters w_2 and λ_{02}. We have

$$p_{02}^\circ = w_2 \cdot {p_{01}^\circ}^* + (1 - w_2) \cdot p_{\lambda_{02},\lambda_{02}}^\circ = w_2 \cdot w_1^* \cdot p_{\lambda_{01}^*,\lambda_{01}^*}^\circ + (1 - w_2) \cdot p_{\lambda_{02},\lambda_{02}}^\circ.$$

Substitute these possible values of w_2 and λ_{02} into this probability, compute $-\mathbb{E}_{p_{\lambda_{11},\lambda_{12}}^\circ}[\log p_{02}^\circ]$ respectively by programming in R. We get approximate values of w_2^* andλ_{02}^* by minimizing this set of $-\mathbb{E}_{p_{\lambda_{11},\lambda_{12}}^\circ}[\log p_{02}^\circ]$.

In iteration 3, we add new parameters w_3 and λ_{03}. Substituting into approximate w_2^* and λ_{02}^* above, We have

$$\begin{aligned} p_{03}^\circ &= w_3 \cdot {p_{02}^\circ}^* + (1 - w_3) \cdot p_{\lambda_{03},\lambda_{03}}^\circ \\ &= w_3 \cdot (w_2^* \cdot w_1^* \cdot p_{\lambda_{01}^*,\lambda_{01}^*}^\circ + (1 - w_2^*) \cdot p_{\lambda_{02}^*,\lambda_{02}^*}^\circ) + (1 - w_3) \cdot p_{\lambda_{03},\lambda_{03}}^\circ. \end{aligned}$$

By similar steps as in iteration 3, we get the approximate w_3^* and λ_{03}^*.

...

From these approximate $w_2^*, \ldots, w_{10}^*$ and $\lambda_{02}^*, \ldots, \lambda_{010}^*$, we obtain a smaller range of w_i: I_w and a smaller range of λ_{0i}^*: I_λ.

Step 3: Find n, $w_1^*, \ldots, w_n^*$ and $\lambda_{02}^*, \ldots, \lambda_{0n}^*$

We choose n_w' evenly spaced values of possible w_i in I_w and n_λ' evenly spaced values of possible λ_{0i} in I_λ. Since these $\boldsymbol{w}^*$ and $\boldsymbol{\lambda}_0^*$ are not exact, then we allow a very small error $\epsilon > 0$. Iterate and stop these iterations when

$$\mathbb{E}_{p_{\lambda_0,\lambda_0}^\circ}\left[\frac{p_{\lambda_{11},\lambda_{12}}^\circ}{p_{0h}^{\circ *}}\right] \leq 1 + \epsilon.$$

We pick $n_\lambda{}''$ evenly spaced values of λ_0 in I_λ'. Substitute these possible values of λ_0 and $p_{0h}^{\circ *}$ into $\mathbb{E}_{p_{\lambda_0,\lambda_0}^\circ}\left[\frac{p_{\lambda_{11},\lambda_{12}}^\circ}{p_{0h}^{\circ *}}\right]$ and judge whether the maximum

value of $\mathbb{E}_{p^{\circ}_{\lambda_0,\lambda_0}}\left[\frac{p^{\circ}_{\lambda_{11},\lambda_{12}}}{p^{\circ *}_{0h}}\right]$ is smaller than $1+\epsilon$. For convenience, we make a judgement every five iterations. We stop when the condition is reached. So we get n, $w^*_1,\ldots,w^*_n$ and $\lambda^*_{02},\ldots,\lambda^*_{0n}$.

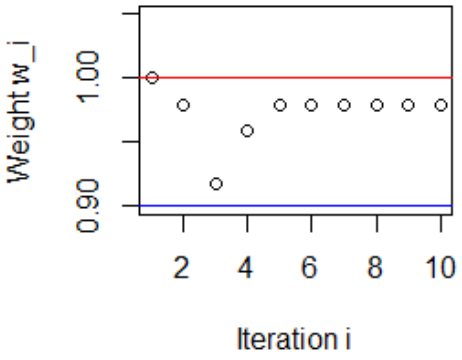
6.2 Examples

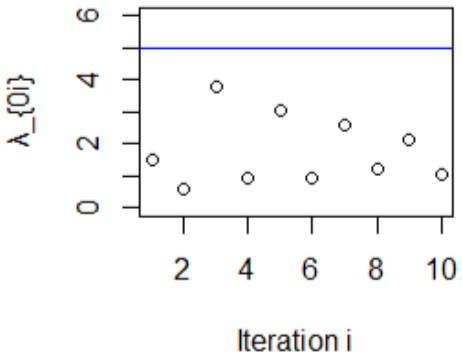
We consider three examples. We set $n_w = 50$, $n_\lambda = 80$, $n'_w = 60$, $n'_\lambda = 250$, $n_\lambda{}'' = 20000$ and $I'_\lambda = [0,5]$.

Example 5 ($\lambda_{11} = 3$,$\lambda_{12} = 1$)**.** Set $\epsilon = 0.0007$.

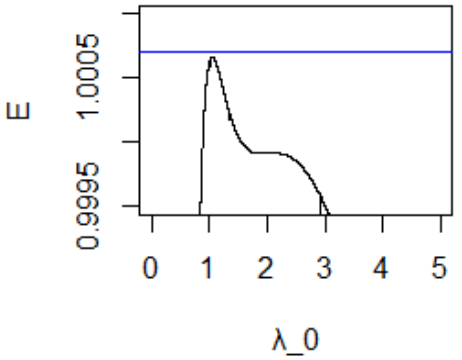
Step 1: $\lambda_{01} = \frac{2\lambda_{11}\lambda_{12}}{\lambda_{11}+\lambda_{12}} = \frac{3}{2}$.

Step 2: Do 10 iterations. We get $I_w = [0.9,1]$ and $I_\lambda = [0,5]$.

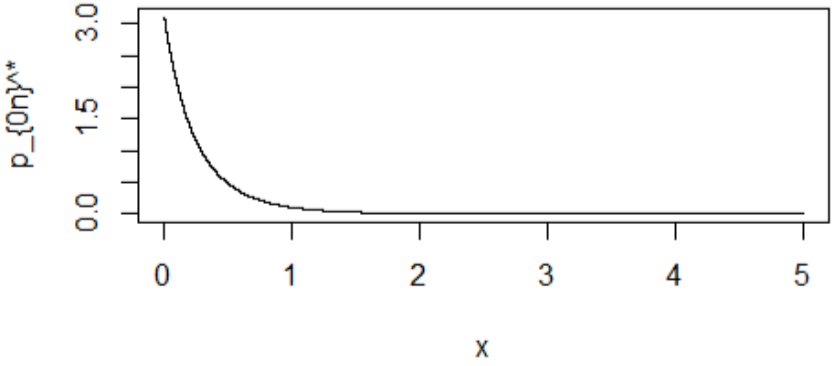




Step 3: We get $n = 125$. The graph of $\mathbb{E}_{p^{\circ}_{\lambda_0,\lambda_0}}\left[\frac{p^{\circ}_{\lambda_{11},\lambda_{12}}}{p^{\circ *}_{0125}}\right]$ is as follows:



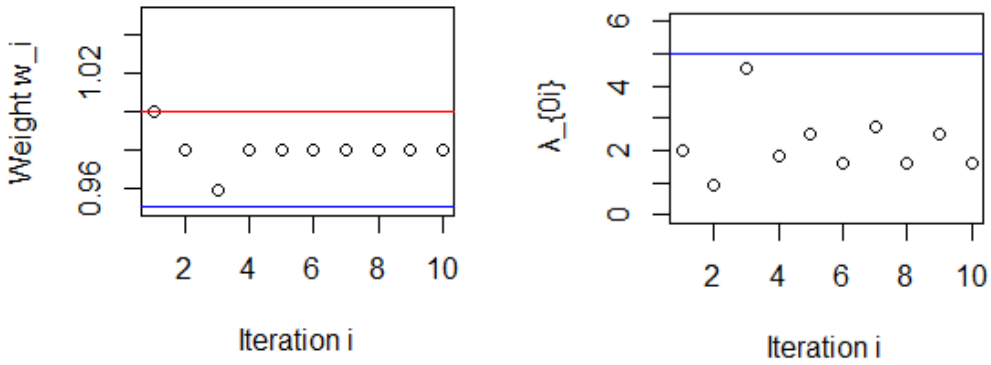
The graph of $p^{\circ *}_{0125}$ is



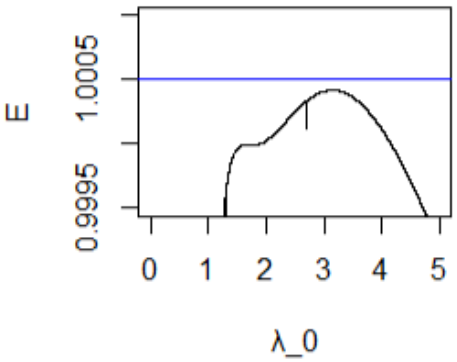
Example 6 ($\lambda_{11} = 3, \lambda_{12} = \frac{3}{2}$)**.** Set $\epsilon = 0.0005$.

Step 1: $\lambda_{01} = \frac{2\lambda_{11}\lambda_{12}}{\lambda_{11}+\lambda_{12}} = 2$.

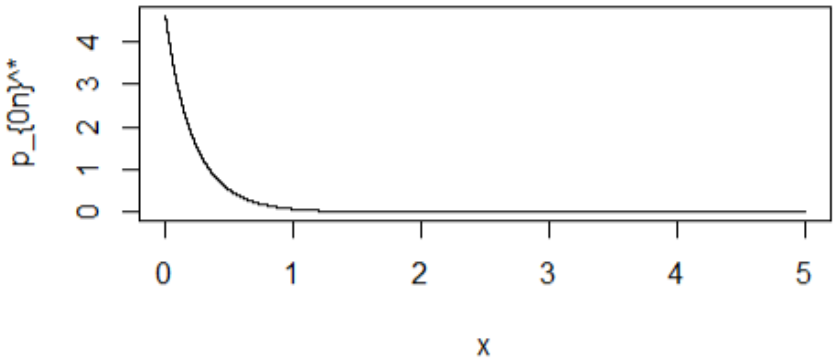
Step 2: Do 10 iterations. We get $I_w = [0.95, 1]$ and $I_\lambda = [0, 5]$.



Step 3: We get $n = 110$. The graph of $\mathbb{E}_{p^\circ_{\lambda_0,\lambda_0}}\left[\frac{p^\circ_{\lambda_{11},\lambda_{12}}}{{p^\circ_{0110}}^*}\right]$ is as follows:



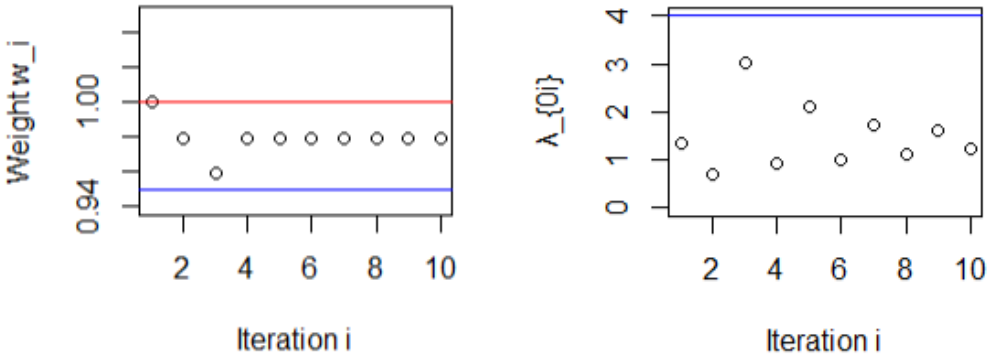
The graph of ${p^\circ_{0110}}^*$ is



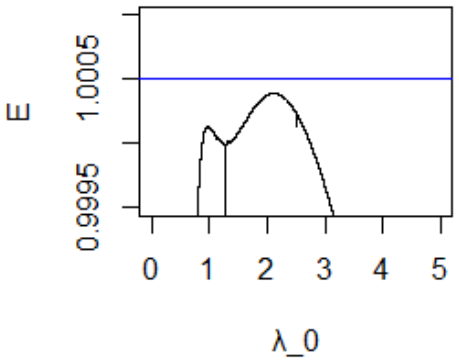
Example 7 ($\lambda_{11} = 2, \lambda_{12} = 1$)**.** Set $\epsilon = 0.0005$.

Step 1: $\lambda_{01} = \frac{2\lambda_{11}\lambda_{12}}{\lambda_{11}+\lambda_{12}} = \frac{4}{3}$.

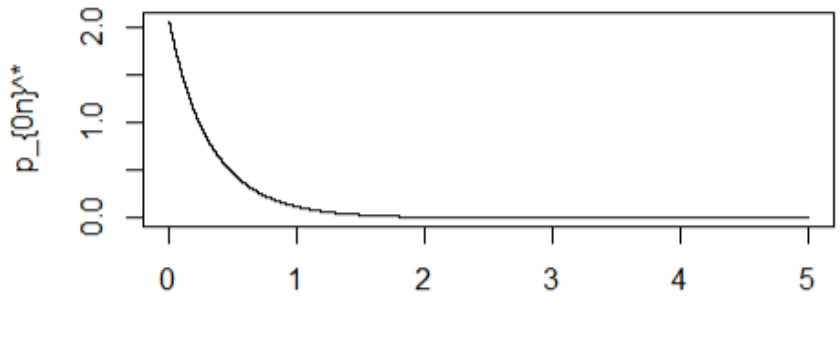
Step 2: Do 10 iterations. We get $I_w = [0.95, 1]$ and $I_\lambda = [0, 4]$.



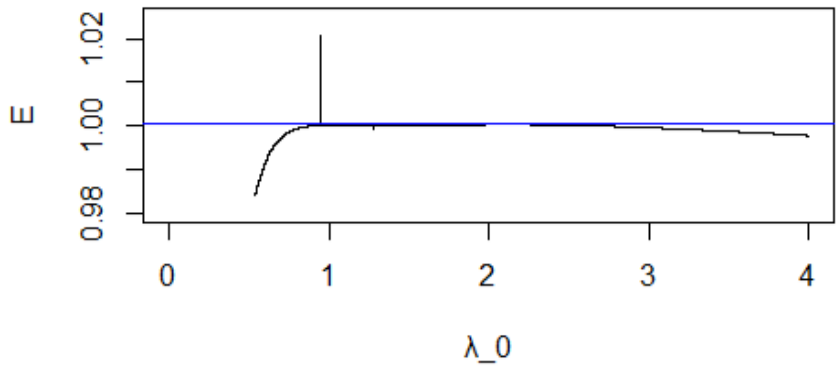
Step 3: We get $n = 90$.The graph of $\mathbb{E}_{p^\circ_{\lambda_0,\lambda_0}}\left[\frac{p^\circ_{\lambda_{11},\lambda_{12}}}{p^{\circ *}_{090}}\right]$ is as follows:



The graph of $p^{\circ *}_{090}$ is



However, the graph of $\mathbb{E}_{p^{\circ}_{\lambda_0,\lambda_0}}\left[\frac{p^{\circ}_{\lambda_{11},\lambda_{12}}}{p^{*}_{090}}\right]$ is different when n_λ" is 50000 instead of 20000. Only one of the 50000 points is greater than 1.0005, which seems to be a numerical issue.



Chapter 7

Conclusion

In this thesis, we establish E-variables by calculating RIPrs, especially simple RIPrs (i.e. a single distribution in the null hypothesis, rather than a mixture of such distributions). We find that there is no general rule to easily determine the RIPr for the whole of exponential family null models. We have obtained a simple RIPr and established E-variables for a simple alternative taken from three types of exponential families: (1) two-parameter exponential families; (2) one-parameter exponential families indexed by an additional integer parameter k; (3) a single alternative representing n independent outcomes of one-parameter exponential families. The corresponding RIPr is simple if some easily checkable conditions in the theorems are satisfied. We have shown how E-variables used for one exponential family can be used for some other exponential families by transformations of random variables. We have also established E-variables for two outcomes of the exponential distribution in a specific setting, by assuming that the RIPr is a mixture of exponential distributions. This thesis raises two interesting open questions for future work:

1. Is there a general rule for some exponential families that describes when exactly the RIPr is not simple? (we only have a sufficient condition). And, if it is not simple, is there a general formula or rule that allows us to calculate it easily?

2. We only consider the case when the null hypothesis is a set of distribution with one parameter in this thesis. What about distributions with two or more parameters?

Bibliography

Yasemin Altun, Alex Smola, and Thomas Hofmann. Exponential families for conditional random fields. *arXiv preprint arXiv:1207.4131*, 2012.

Arindam Banerjee. An analysis of logistic models: Exponential family connections and online performance. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 204–215. SIAM, 2007.

Lawrence D Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.

Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE, 2020.

Glenn Shafer. The language of betting as a strategy for statistical and scientific communication. *arXiv preprint arXiv:1903.06991*, 2019.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Rosanne Turner, Alexander Ly, and Peter Grünwald. Safe tests and always-valid confidence intervals for contingency tables and beyond. *arXiv preprint arXiv:2106.02693*, 2021.

Judith ter Schure, Muriel F Pérez-Ortiz, Alexander Ly, and P Grunwald. The safe logrank test: Error control under continuous monitoring with unlimited horizon. *arXiv preprint arXiv:2011.06931*, 2020.

Qiang Jonathan Li. *Estimation of mixture models*. Yale University, 1999.

D. Williams. *Probability with Martingales*. Cambridge Mathematical Textbooks, 1991.

J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, New York, revised and expanded 2nd edition, 1985.

Erling Bernhard Andersen. Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65(331): 1248–1255, 1970.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Jonathan Li and Andrew Barron. Mixture density estimation. *Advances in neural information processing systems*, 12, 1999.