UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2019:10

On Curved Exponential Families

Emma Angetun

Examensarbete i matematik, 15 hp
Handledare: Silvelyn Zwanzig
Examinator: Örjan Stenflo
Mars 2019

Department of Mathematics
Uppsala University

On Curved Exponential Families

Emma Angetun

March 4, 2019

Abstract

This paper covers theory of inference statistical models that belongs to curved exponential families. Some of these models are the normal distribution, binomial distribution, bivariate normal distribution and the SSR model. The purpose was to evaluate the belonging properties such as sufficiency, completeness and strictly k-parametric. Where it was shown that sufficiency holds and therefore the Rao Blackwell Theorem but completeness does not so Lehmann Scheffé Theorem cannot be applied.

Contents

1 EXPONENTIAL FAMILIES

Statistical inference is concerned with looking for a way to use the information in observations $\mathbf{x}$ from the sample space $\mathcal{X}$ to get information about the partly unknown distribution of the random variable $\mathbf{X}$. Essentially one want to find a function called *statistic* that describe the data without loss of important information. The exponential family is in probability and statistics a class of probability measures which can be written in a certain form. The exponential family is a useful tool because if a conclusion can be made that the given statistical model or sample distribution belongs to the class then one can thereon apply the general framework that the class gives. If it can be stated that the distribution belongs to an exponential family one might be able to reduce the *statistic* to a lower dimension without the risk of losing information. [1]

Definition 1.1. A statistical model is a class of probability measures

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

where θ is a parameter and Θ is the parameter space and contains all possible parameterizations. The statistical model $\mathcal{P}$ is defined on the **sample space** $\mathcal{X}$ where the elements *(observations)* $\mathbf{x}$ of the set are realizations of the random variable $\mathbf{X}$.

Definition 1.2 (Statistic)**.** A **statistic** T is a function of the sample

$$T : \mathbf{x} \in \mathcal{X} \rightarrow T(\mathbf{x}) = t \in \mathcal{T}$$

where $\mathcal{T}$ is a suitable set. With the random variables $\mathbf{X}$ the function

$$T(\mathbf{X}) : \omega \in \Omega \rightarrow T(\mathbf{x}) = t \in \mathcal{T}$$

is a random variable. The distribution of T is given by

$$P_\theta^T(B) = P_\theta(\{\mathbf{x} : T(\mathbf{x}) \in B\}).$$

Consider a class of probability measures $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and assume that for each P_θ there exists a probability function $p(\cdot;\theta)$ then the definition of the exponential family can be presented.

Definition 1.3 (Exponential Family)**.** A class of probability measures $\mathcal{P} = \{ P_\theta : \theta \in \Theta \}$ is called an **exponential family** if there exist a number k $\in \mathbb{N}$, real-valued functions $\eta_1, ..., \eta_k$ on Θ, real-valued statistics $T_1, ..., T_k$ and a function h on $\mathcal{X}$ such that the probability function has the form

$$p(x;\theta) = A(\theta) \exp\Big(\sum_{i=1}^{k} T_i(x)\eta_i(\theta)\Big)h(x). \tag{1}$$

The exponential expression of the probability function will determine the statistical properties. The functions $\eta = (\eta_1, .., \eta_k)$ and $T = (T_1, ..., T_k)$ and k are not uniquely determined but we refer to equation (1) as a **k-parameter exponential family**.

Example 1.1 (Normal distribution)**.** Let the statistical model be the class of all normal distributions where μ and σ are unknown and $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. The probability function can be written in the form of an exponential family by some algebraic manipulations.

$$\begin{aligned} p(x;\theta) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-x^2 + 2x\mu - \mu^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\}. \end{aligned}$$

And with the resulting functions.

$$\begin{aligned} &A(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \\ &T_1(x) = x \qquad \eta_1(\theta) = \frac{\mu}{\sigma^2} \\ &T_2(x) = x^2 \qquad \eta_2(\theta) = -\frac{1}{2\sigma^2} \end{aligned}$$

The distribution with the chosen parameter can be written on the wanted form because it results in a two dimensional statistic $T = (T_1(x), T_2(x))$ with a corresponding $\eta(\theta)$ and also the normalizing factor $A(\theta)$.
In conclusion the normal distribution with unknown mean and variance belongs to a 2-parameter exponential family.

The dimension of the statistic will be determined by the dimension k of the family and statistical procedures are done with the statistic T. One therefore wants to choose the dimension of the exponential family as small as possible in order to also reduce the dimension of the statistic. When the dimension k is minimal we call it a **strictly k-parameter exponential family**.

Definition 1.4 (Strictly k-parameter exponential family)**.** A class of probability measures $\mathcal{P}$ belonging to an exponential family is said to be a strict k-parameter exponential family when k is minimal.

For exponential families we have that the set $\mathcal{A} = \{x : p(x;\theta) > 0\}$ is independent of θ.

Definition 1.5 ($\mathcal{P}$-affine independence)**.** The functions $T_1, .., T_k$ are called $\mathcal{P}$-affine independent if for $c_j \in \mathbb{R}$ and $c_0 \in \mathbb{R}$

$$\sum_{i=1}^{k} c_j T_j(x) = c_0 \quad \text{for all } x \in \mathcal{A} \text{ implies} \quad c_j = 0, \quad \text{for } j = 0, ..., k$$

The following theorem gives a technique in how to determine if the dimension k is minimal so that a strict k-parameter exponential family arises.

Theorem 1.1. *Let $\mathcal{P}$ be an exponential family. Then*

1. *The family $\mathcal{P}$ is strictly k-dimensional if in (1) the functions $\eta_1,...,\eta_k$ are linearly independent and the statistics $T_1,...,T_k$ are $\mathcal{P}$-affine independent.*

2. *The functions are $T_1,...,T_k$ are $\mathcal{P}$-affine independent, if the covariance matrix $Cov_\theta T$ is positive definite for all $\theta \in \Theta$.*

Proof of (1) is given by H. Witting. Mathematische Statistik.

Proof. (2) We assume that the $Cov_\theta T$ is a $k \times k$ positive definite. We set the functions in a $k \times 1$ matrix T and assume that the functions in T are linearly dependent. This means that we can write $T_{k\times 1} = B_{k\times p} D_{p\times 1}$ where $p < k$ so that T is of rank p. Now $\Sigma_D = \mathrm{Cov} D > 0$. Now the covariance matrix of T can be written on the form $Cov_\theta T = B_{k\times p} \Sigma_D B^T_{p\times k}$. Then it's known by the fact that Σ_D is symmetric it has a Single Value Decomposition $\Sigma_D = \Gamma \Lambda \Gamma^T$. Where Λ and Γ are $p \times p$ matrices and Λ is the diagonal matrix with positive eigenvalues. We then conclude that the equation is of rank p and that $Cov_\theta T$ is a singular matrix which is a contradiction to the assumption that $Cov_\theta T$ is positive definite. □

Example 1.2 (Normal distribution)**.** Consider **Example 1.1** again and show that it is a strictly 2-parameter exponential family by applying *Theorem 1.1* and thereon calculate the covariance matrix. To calculate $Cov_\theta \mathbf{T}$ is the same as to calculate the variances of respective statistic, and the covariances between the statistics. The known moments for the normal distribution are used for these calculations.

$$EX^2 = M_X^{(2)}(0) = \mu^2 + \sigma^2$$
$$EX^4 = M_X^{(4)}(0) = 3\sigma^4 + 6\sigma^2\mu^2 + \mu^4$$
$$Var(T_2) = Var(X^2) = EX^4 - (EX^2)^2 = 2\sigma^4 + 4\mu^2\sigma^2$$

$$\begin{aligned} Cov[T_1, T_2] &= Cov[X, X^2] = EX^3 - EXEX^2 \\ &= \mu^3 + 3\mu\sigma^2 - \mu(\mu^2 + \sigma^2) \\ &= 2\mu\sigma^2. \end{aligned}$$

$$Cov_\theta \mathbf{T} = \begin{pmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 2\sigma^4 + 4\sigma^2\mu^2 \end{pmatrix}.$$

Which is positive definite by the fact that the determinant is not equal to zero, hence we can conclude that assumption (2) in **Theorem 1.1** is fulfilled. In conclusion the normal distribution with unknown mean and variance is a **strictly 2-parameter exponential family**.

1.1 NATURAL PARAMETER SPACE

The real valued functions $\eta_1, ..., \eta_k$ were introduced in the definition of the exponential family and shown in the example of the normal distribution. These functions are defined on the parameter space Θ and can be set to η which will be referred to as the *natural parameter*. The natural parameter space is the set of points η where the probability measure on the form of the exponential family is a probability function. The function A(θ) is just a normalizing factor which depends on the parameter θ from $\eta(\theta)$. We consider the class of probabilities $\mathcal{P} = \{P_\eta : \eta \in Z\}$ where Z := $\eta(\Theta)$. The probability function can be expressed in the new parametrization,

$$p(x;\eta) = A(\eta)\exp\left(\sum_{i=1}^{k} \eta_i T_i(x)\right) h(x) \quad A(\eta) = \left(\int \exp\left[\sum_{i=1}^{k} \eta_i T_i(x)\right] h(x) dx\right)^{-1}.$$

It's necessary to consider a set where the parametrization gives a defined probability function. From the statement above we can derive the following definition.

Definition 1.6. Let $\mathcal{P}$ be a class of probabilities which belongs to an exponential family with the parametrization $\eta := \eta(\theta)$ called the **natural parameter**. The set of the natural parameters where the probability function is finite is called the **natural parameter space**;

$$\mathcal{N} = \{\eta : 0 < \int \exp\left(\sum_{i=1}^{k} \eta_i T_i(x)\right) h(x) dx < \infty, \ \eta \in \mathbb{R}^k\}.$$

Generally we have that $\eta(\Theta) \subset \mathcal{N}$.

Theorem 1.2. *The natural parameter space of a k-parameter exponential family is convex.*

Proof. We assume that η_0, $\eta_1 \in \mathcal{N}$ and that $\alpha \in (0, 1)$.
$\mathcal{N}$ is a convex set if $\alpha\eta_0$ + (1 - α)$\eta_1 \in \mathcal{N}$.

$$
\begin{aligned}
0 &< \int \exp(\alpha \sum_{i=1}^{k} \eta_{0i} T_i(x)) \exp((1-\alpha) \sum_{i=1}^{k} \eta_{1i} T_i(x)) dx \\
&= \int \exp\left(\sum_{i=1}^{k} \eta_{0i} T_i(x)\right)^{\alpha} \exp\left(\sum_{i=1}^{k} \eta_{1i} T_i(x)\right)^{1-\alpha} dx \\
&\leq \left(\int \exp \sum_{i=1}^{k} \eta_{0i} T_i(x) dx\right)^{\alpha} \left(\int \exp \sum_{i=1}^{k} \eta_{1i} T_i(x) dx\right)^{1-\alpha} < \infty
\end{aligned}
$$

With Hölders Inequality it can be concluded that the integral is bounded from above and the convexity of the set is proved. □

Theorem 1.3. *The natural parameter space $\mathcal{N}$ of a strictly k-parameter exponential family contains a nonempty k-dimensional interval.*

[1]

2 CURVED EXPONENTIAL FAMILIES

Curved exponential families are a subset of the class of exponential families where the dimension of the parameter space does not match the dimension of the exponential family. These special cases are interesting to discuss because they are more likely to violate assumptions and statistical properties belonging to the exponential families. We'll bring up some useful models such as the multivariate normal distribution and so on. Because of vague formulations and no definition on the curved exponential family in books and articles that were considered, we'll define it.

Definition 2.1. Curved Exponential Families
A class of probability measures $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is called a **curved exponential family** if there exists two numbers $q < k \in \mathbb{N}$, real valued functions $\eta_1, ..., \eta_k$ on $\Theta \subseteq \mathbb{R}^q$, real valued statistics $T_1, ..., T_k$ and a function h on $\mathcal{X}$. If there exists a θ where $Cov_\theta T$ is positive definite and that the probability measure has the form of the exponential family

$$p(x;\theta) = A(\theta) \exp\left(\sum_{i=1}^{k} T_i(x)\eta_i(\theta)\right) h(x).$$

In order to assure that k is not an arbitrary large number we state that the covariance matrix for the statistics must be positive definite for some θ. To set $\forall\theta$ would be a too strict requirement.

Example 2.1 (Normal Distribution)**.** Let the statistical model be the class of all normal distributions with $N(\mu, \mu^2)$ where μ is unknown and $\mu \neq 0$ with parameter $\theta = \mu$ and $\theta \in (-\infty, 0) \cup (0, \infty)$.

$$\begin{aligned} p(x;\theta) &= \frac{1}{\sqrt{2\pi}\mu} \exp\left\{\frac{-(x-\mu)^2}{2\mu^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\mu} \exp\left\{\frac{-x^2 + 2x\mu - \mu^2}{2\mu^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\mu} \exp\left\{-\frac{x^2}{2\mu^2} + \frac{x}{\mu} - \frac{1}{2}\right\}. \end{aligned}$$

By the same argument as the previous example it results in a two dimensional statistic.

$$\begin{aligned} T_1(x) &= x & \eta_1(\theta) &= \frac{1}{\mu} \\ T_2(x) &= x^2 & \eta_2(\theta) &= -\frac{1}{2\mu^2}. \end{aligned}$$

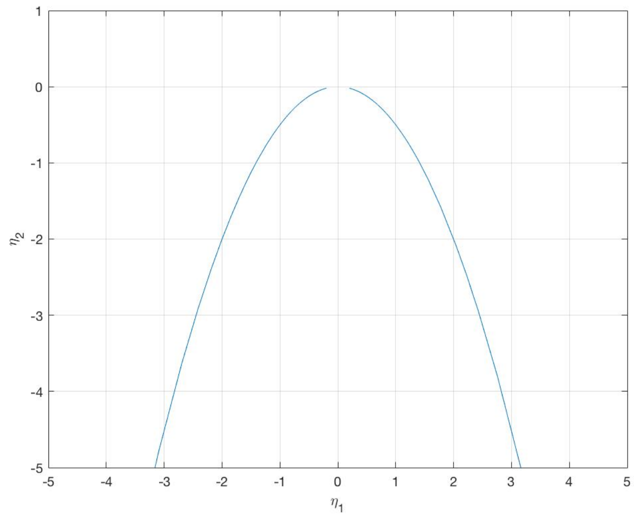

Figure 1: plot of the function $\eta(\theta)$

Consider the natural parameters;

$$\eta(\theta) = \left(\frac{1}{\theta}, \ -\frac{1}{2\theta^2}\right)$$

which creates a curve that can be seen in the figure below. This examples satisfies the definition of Curved Exponential families because the parameter θ is one dimensional but results in a 2-parameter exponential family. It can also be called a $(2, 1)$-parameter exponential family to emphasize that it's curved. The covariance matrix can be calculated with the known moments like in the example before.

$$Cov_\theta T = \begin{pmatrix} \mu^2 & 2\mu^3 \\ 2\mu^3 & 6\mu^4 \end{pmatrix}.$$

The covariance matrix is shown to be positive definite for all $\theta \in \Theta$ through the determinant.

$$\begin{vmatrix} \mu^2 & 2\mu^3 \\ 2\mu^3 & 6\mu^4 \end{vmatrix} = 6\mu^6 - 4\mu^6 = 2\mu^2 > 0.$$

By that the example satisfies condition of statement (2) in **Theorem 1.1** implies that the functions T_1, T_2 are $\mathcal{P}$-affine independent. This concludes that this statistical model fulfills the definition of the curved exponential family. One can evaluate if there exists any dependencies between the functions η and it can be seen in Figure 1 that a nonlinear dependence exists.

This gives us the results that this curved exponential family is not a **strictly 2-parametric family**. [2]

Example 2.2 (Multivariate Normal distribution)**.** Let the statistical model be the class of all bivariate normal distribution where $(X_1, X_2)^T$ are jointly bivariate. Where $\boldsymbol{\mu} = (0,0)^T$ and Σ is a 2×2 matrix with variances 1 and correlation ρ with parameter $\theta = \rho$.

$$\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}, \qquad -1 < \rho < 1 \; and \; \rho \neq 0.$$

$$f_X(x_1, x_2; \theta) = \frac{1}{\sqrt{(2\pi)^2}\sqrt{det(\Sigma)}} \exp\{-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\}$$
$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\{-\frac{1}{2}\frac{1}{(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)\}.$$

$$T_1(\mathbf{x}) = x_1^2 + x_2^2 \qquad T_2(\mathbf{x}) = x_1 x_2$$
$$\eta_1(\theta) = \frac{-1}{2(1-\rho^2)} \qquad \eta_2(\theta) = \frac{\rho}{1-\rho^2}$$

Hence we get a two dimensional statistic with θ as one dimensional. If there exists a θ such that the covariance matrix is positive definite it means that the model results in a curved 2-parameter exponential family. One can use the Wishart distribution to find the variances of T_1, T_2.

We have to determine if the covariance matrix of the statistics is positive definite so it can be concluded that the model belongs to the curved exponential family. We have that

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}^T = \begin{pmatrix} X_1^2 & X_1X_2 \\ X_1X_2 & X_2^2 \end{pmatrix} \sim \mathcal{W}_2(\Sigma, 1),$$

from this distribution we can find the variance and the expected value for $T_2 = X_1X_2$.

$$Var(X_1X_2) = n(\sigma_{12}^2 + \sigma_{11}\sigma_{22}) = \rho^2 + 1$$
$$EX_1X_2 = \rho.$$

The σ_{ii} is the element of the covariance matrix in position ii.
To find the variance and the expected value for T_1 one can use the Wishart distribution once again.

$$\begin{pmatrix} X_1 & X_2 \\ X_2 & X_1 \end{pmatrix}\begin{pmatrix} X_1 & X_2 \\ X_2 & X_1 \end{pmatrix} \sim \mathcal{W}_2(\Sigma, 2).$$

In the same way as above we can now get the variance and the expected value.

$$Var(X_1^2 + X_2^2) = 4\rho^2$$
$$E(X_1^2 + X_2^2) = E(X_1^2) + E(X_2^2) = 2.$$

The expected value can also easily be found by the fact that both X_1 and X_2 separately are standard normal. The covariance matrix can now be calculated with help of some properties of the multivariate normal distribution.

$$\begin{aligned} Cov(X_1^2 + X_2^2, X_1X_2) &= E((X_1^2 + X_2^2)X_1X_2) - E(X_1^2 + X_2^2)E(X_1X_2) \\ &= E(X_1^3X_2) + E(X_2^3X_1) - 2\rho \\ &= E_{x_1}(E(X_1^3X_2|X_1)) + E_{x_2}(E(X_2^3X_1|X_2)) - 2\rho \\ &= E_{x_1}(X_1^3E(X_2|X_1)) + E_{x_2}(X_2^3E(X_1|X_2)) - 2\rho \\ &= E_{x_1}(X_1^3\rho X_2) + E_{x_2}(X_2^3\rho X_1) - 2\rho \\ &= \rho X_2E_{x_1}(X_1^3) + \rho X_1E_{x_2}(X_2^3) - 2\rho \\ &= 0 + 0 - 2\rho \\ &= -2\rho. \end{aligned}$$

Now the resulting Covariance matrix of T_1, T_2

$$\begin{pmatrix} Var(X_1^2 + X_2^2) & Cov(X_1^2 + X_2^2, X_1X_2) \\ Cov(X_1^2 + X_2^2, X_1X_2) & Var(X_1X_2) \end{pmatrix} = \begin{pmatrix} 4\rho^2 & -2\rho \\ -2\rho & \rho^2 + 1 \end{pmatrix}.$$

The determinant is calculated

$$\begin{vmatrix} 4\rho^2 & -2\rho \\ -2\rho & \rho^2 + 1 \end{vmatrix} = 8\rho^2 + 4\rho^4 > 0.$$

The covariance matrix of T_1, T_2 is shown to be positive definite which concludes that the Multivariate Normal distribution with inference on ρ belongs to a 2-parameter curved exponential family.

Example 2.3 (Joint Distribution of Two Binomial r.v.s)**.** Consider the class of all joint distributions of $\mathbf{Z} = (Z_1, Z_2)^T$ where $Z_1 \sim \text{Bin}(n, p)$ and $Z_2 \sim \text{Bin}(m, p^2)$ where Z_1 and Z_2 are independent and with parameter $\theta = p$. We assume that $0 < p < 1$ and $m, n > 0$. Since Z_1 and Z_2 are independent we have the joint distribution,

$$P(\mathbf{Z} = \mathbf{z}) = P(Z_1 = z_1 \textit{ and } Z_2 = z_2) = P(Z_1 = z_1)P(Z_2 = z_2).$$

$$
\begin{aligned}
p(\mathbf{z};\theta) &= \binom{n}{z_1} p^{z_1}(1-p)^{n-z_1}\binom{m}{z_2} p^{2z_2}(1-p^2)^{m-z_2} \\
&= \exp\{z_1 \ln\frac{p}{1-p} + \ln(1-p)^n + 2z_2 \ln p + z_2 \ln\frac{1}{1-p^2} + \ln\{1-p^2\}^m\}\binom{n}{z_1}\binom{m}{z_2} \\
&= (1-p)^n(1-p^2)^m \exp\{z_1 \ln\frac{p}{1-p} + z_2(2\ln p + \ln\frac{1}{1-p^2})\}\binom{n}{z_1}\binom{m}{z_2}.
\end{aligned}
$$

$$
\begin{aligned}
T_1(\mathbf{z}) &= z_1 & \eta_1(\theta) &= \ln\frac{p}{1-p} \\
T_2(\mathbf{z}) &= z_2 & \eta_2(\theta) &= 2\ln p + \ln\frac{1}{1-p^2}
\end{aligned}
$$

$$
Cov_\theta T = \begin{pmatrix} np(1-p) & 0 \\ 0 & mp^2(1-p^2) \end{pmatrix}
$$

$$
\begin{vmatrix} np(1-p) & 0 \\ 0 & mp^2(1-p^2) \end{vmatrix} = mnp^3(1-p)(1-p^2) > 0
$$

The determinant concludes that the covariance matrix is positive definite for some θ and the requirements for the curved exponential family holds. So that the statistical model belongs to a *curved 2-parameter exponential family.*

Example 2.4 (Simple Structural Relation)**.** Let's consider a regression model that has the structure

$$
\begin{aligned}
Y_i &= \beta\xi_i + \varepsilon_i \\
X_i &= \xi_i + \delta_i,
\end{aligned}
$$

where $\beta \in \mathbb{R}$ and ζ_i, ξ_i and δ_i are all standard normal and independent.

$$
\begin{aligned}
&EX_i = 0 & Var(X_i) &= Var(\xi_i) + Var(\delta_i) = 2 \\
&E[Y_i] = 0 & Var(Y_i) &= \beta^2 Var(\xi_i) + Var(\varepsilon_i) = \beta^2 + 1 \\
&Cov[Y_i, X_i] = \beta
\end{aligned}
$$

$$
\begin{aligned}
\Sigma &= \begin{pmatrix} 2 & \beta \\ \beta & \beta^2+1 \end{pmatrix} \\
\Sigma^{-1} &= \begin{pmatrix} \frac{1+\beta^2}{2+\beta^2} & -\frac{\beta}{2+\beta^2} \\ -\frac{\beta}{2+\beta^2} & \frac{2}{2+\beta^2} \end{pmatrix}
\end{aligned}
$$

$\mathbf{Z}_i = (X_i, Y_i)^T$, $Z_i \sim N_2(0, \Sigma)$. Now consider the class $\mathcal{P}$ of all these distributions with parameter β. From the structure of the bivariate normal distribution the exponential in the distribution will be determined by

$$z^T \Sigma^{-1} z = x^2 \frac{1+\beta^2}{2+\beta^2} - 2xy \frac{\beta}{2+\beta^2} + y^2 \frac{2}{2+\beta^2}.$$

The statistics

$$T(\mathbf{z}) = \left(x^2, xy, y^2\right) \qquad \eta(\beta) = \left(\frac{1+\beta^2}{2+\beta^2}, -\frac{\beta}{2+\beta^2}, \frac{2}{2+\beta^2}\right).$$

This seems to be a curved exponential family since β is one-dimensional but it results in a 3-parameter exponential family. It is not strictly 3-parametric by **Theorem 1.1** where it breaks the condition of linearly independence in $\eta(\beta)$.

$$\sum_{j=1}^{3} c_j \eta(\beta) = c_0$$

$$\text{when} \quad c_0 = c_1 = c, \qquad c_2 = 0, \qquad c_3 = \frac{c}{2}$$

$$\sum_{j=1}^{3} c_j \eta(\beta) = \frac{c((1+\beta^2)+1)}{\beta^2+2} = c$$

Hence all three functions are not linearly independent We know how Y and X are distributed and with the model of X and Y and also the moments of the normal distribution. Odd moments for ξ_i, δ_i and ε_i are equal to 0.

$$
\begin{aligned}
Var(XY) &= EX^2Y^2 - EXYEXY \\
&= E[(\xi_i + \delta_i)^2(\beta\xi_i + \varepsilon_i)^2] - (E[(\xi_i + \delta_i)(\beta\xi_i + \varepsilon_i)])^2 \\
&= 3\beta^2 + 1 \\
Var(X^2) &= 8 \\
Var(Y^2) &= Var(\beta\xi_i + \varepsilon_i)^2 = 2\beta^4 + 4\beta + 2 \\
Cov[Y^2, XY] &= E[Y^3X] - E[Y^2]E[XY] \\
&= E[(\beta\xi_i + \varepsilon_i)^3(\xi_i + \delta_i)] - E[(\beta\xi_i + \varepsilon_i)^2]E[(\xi_i + \delta_i)(\beta\xi_i + \varepsilon_i)] \\
&= \beta^3 E[\xi_i^4] + 3\beta E[\varepsilon_i^2]E[\xi_i^2] - (\beta^2 E[\xi_i^2] + E[\varepsilon_i^2])\beta E[\xi_i^2] \\
&= 3\beta^3 + 3\beta - (\beta^3 + \beta) \\
&= 2\beta^3 + 2\beta \\
Cov[X^2, XY] &= E[X^3Y] - E[X^2]E[XY] \\
&= \beta E[\xi_i^4] + 3\beta E[\delta_i^2]E[\xi_i^2] - 2\beta E[\xi_i^2] \\
&= 4\beta \\
Cov[X^2, Y^2] &= E[X^2Y^2] - E[X^2]E[Y^2] \\
&= E[(\xi_i + \delta_i)^2(\beta\xi_i + \varepsilon_i)^2] - 2(\beta + 1) \\
&= 4\beta^2 - 2\beta - 1
\end{aligned}
$$

$$
Cov_\theta \mathbf{T} = \begin{pmatrix} 8 & 4\beta & 4\beta^2 - 2\beta - 1 \\ 4\beta & 3\beta^2 + 1 & 2\beta^3 + 2\beta \\ 4\beta^2 - 2\beta - 1 & 2\beta^3 + 2\beta & 2(\beta^2 + 1)^2 \end{pmatrix}
$$

The long calculations of the determinant are left out and results in

$$
\begin{vmatrix} 8 & 4\beta & 4\beta^2 - 2\beta - 1 \\ 4\beta & 3\beta^2 + 1 & 2\beta^3 + 2\beta \\ 4\beta^2 - 2\beta - 1 & 2\beta^3 + 2\beta & 2(\beta^2 + 1)^2 \end{vmatrix} = 16\beta^5 - 20\beta^4 - 28\beta^3 + \beta^2 - 4\beta + 15.
$$

The equation of the determinant can be shown to have roots in $\mathbb{R}$ which leads to the conclusion that there exists a θ where the covariance matrix is positive definite. Therefore by the definition of the curved exponential family the statistical model belongs to the curved exponential family. The SSR-model does not fulfill any part of **Theorem 1.1** and is therefore not a strictly 3-parameter exponential family.

In conclusion we can see by our several examples that they either fail the linear independence of η, T or both which means that k is not minimal in these cases. This gives a hint that curved families might not fulfill that k is minimal but a generalization of this claim is left out.

2.1 Natural parameter space of curved families

We have shown that the natural parameter space of a *k-parameter exponential family* is a convex set which also applies to the curved exponential family. The natural parameter space of a *Curved Exponential Family* does not contain a non-empty k-dimensional interval since it relies on the $\mathcal{P}$-affine independence of η and as seen in **Example 2.5** that is not always the case. This will lead to complications further on in the section of completeness.

3 Sufficiency

A statistic is called sufficient if one cannot find another estimator calculated from the sample space that provides additional information as to the value of the parameter. For a family of distributions a statistic is sufficient if the sample space from which it's deduced from gives no additional information. The sufficiency for statistics assures us that all information about the parameter θ included in $\mathbf{X}$ is also contained in the statistic. This makes it a strong property since it can be viewed as a way of data reduction, where all the important information in the sample is condensed into the statistic. [1]

Definition 3.1. (Sufficient Statistic) A statistic T is said to be **sufficient** for the statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of $\mathbf{X}$ if the conditional distribution of $\mathbf{X}$ given $\mathrm{T} = \mathrm{t}$ is independent of θ for all t.

Theorem 3.1. *(Factorization criterion) Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model with probability function $p(\cdot;\theta)$. A statistic T is sufficient for $\mathcal{P}$ if and only if there exists nonnegative functions $g(\cdot;\theta)$ and h such that the probability functions satisfy*

$$p(\boldsymbol{x};\theta) = g(T(\boldsymbol{x});\theta)h(\boldsymbol{x}).$$

Example 3.1. (Sufficiency for Normal Distribution) Let $\mathbf{X} \sim N(\mu, \sigma^2)$ where both μ and σ are unknown, $\theta = (\mu, \sigma^2)$.

$$\begin{aligned} p(x;\theta) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2)\right) \end{aligned}$$

thus $T(\mathbf{X}) = (\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i)$ are sufficient statistics from the Factorization criterion **(3.1)** where the function $\mathrm{h(x)} = 1$ and g is the rest.

Example 3.2. (Sufficiency for Curved Normal) Let $\mathbf{X} \sim N(\mu, \mu^2)$ where $\mu = \theta$ is unknown.

$$p(x;\theta) = \left(\frac{1}{2\pi\mu^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\mu^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$
$$= \left(\frac{1}{2\pi\mu^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\mu^2}(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2)\right),$$

thus $T(\mathbf{X}) = (\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i)$ are again sufficient statistics from the Factorization criterion **(3.2)** and so on h(x) = 1 and g(x) is the rest.

When the statistic generates the coarsest sufficient partition of the sample space we define this property as *minimal sufficient*. Instead of evaluating the idea of partition of the sample space the following definition can be used.

Definition 3.2 (Minimal sufficiency). A statistic is **minimal sufficient** iff T is a function of any other sufficient statistic.

Definition 3.3. The set $\mathcal{K}$ is the set of all pairs $(\mathbf{x},\mathbf{y})$ for which there is a $k(\mathbf{x},\mathbf{y}) > 0$ such that

$$L(\theta;\mathbf{x}) = k(\mathbf{x},\mathbf{y})L(\theta;\mathbf{y}) \qquad \text{for all } \theta \in \Theta.$$

Note that the function $L(\theta;\mathbf{x})$ is the likelihood function.

Theorem 3.2. *Let T be a sufficient statistic for* $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. *If for all* $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{K}$ *the statistic T satisfies T(**x**) = T(**y**) then T is minimal sufficient.*

Example 3.3 (Minimal Sufficiency for Normal). $\mathbf{X} \sim N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$ by theorem above we construct the likelihood ratio.

$$\frac{L(\theta;\mathbf{x})}{L(\theta;\mathbf{y})} = \frac{(\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)}{(\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)}$$
$$= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$
$$= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}y_i^2\right) - 2\mu(\sum_{i=1}^{n}x_i - y_i)\right).$$

By the same reasoning we have that the ratio is independent of θ if $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2$ Thus the statistics

$$T(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x})) = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right)$$

are minimal sufficient.

These examples gives us an idea that sufficiency generally should hold for k-parameter exponential families. Minimal sufficiency requires that the sample of i.i.d random variables comes from a strictly k-parameter exponential family in order to assure that k is minimal.

Theorem 3.3. *For a sample of i.i.d random variables from a strictly k-parameter exponential family it holds:*

1. The statistic

$$T_{(n)}(\boldsymbol{x}) = \left(\sum_{i=1}^{n} T_1(x_i), ..., \sum_{i=1}^{n} T_n(x_n) \right) \tag{2}$$

is minimal sufficient.

Proof. If we use **Theorem 3.1** then we have that the ratio of the likelihood function at points $\mathbf{x}$ and $\mathbf{y}$

$$\frac{L(\theta; \mathbf{x})}{L(\theta; \mathbf{y})} = \frac{\prod_{i=1}^{n} h(x_i)}{\prod_{i=1}^{n} h(y_i)} \exp \left[\sum_{j=1}^{k} \eta_j(\theta) \left(\sum_{i=1}^{n} T_j(x_i) - \sum_{i=1}^{n} T_j(y_i) \right) \right]$$

We can conclude that the ratio is independent of θ i.e. $(\mathbf{x},\mathbf{y}) \in \mathcal{K}$ iff

$$\sum_{i=1}^{n} T_j(x_i) = \sum_{i=1}^{n} T_j(y_i) \qquad \text{for all } j = 1, ..., k.$$

And therefore the statistic $T_{(n)}(\mathbf{x})$ is minimal sufficient. □

[1]

When dealing with strictly k-parameter exponential families one can broaden some theorems but an important part is that the strong property of sufficiency still holds for non-strictly families although minimal sufficiency does not. This means that when we have a curved exponential family we only have to state that it's a k-parameter exponential family to know that it has a sufficient statistic.

4 Completeness

Completeness describes the ranges of the parameter space related to the range of the sample space. Generalized one could say that a statistical model is large enough if it is complete.

Definition 4.1 (Completeness)**.** A statistical model $\mathcal{P} = \{ P_\theta : \theta \in \Theta \}$ is called **complete** if for any function h: $\mathcal{X} \to \mathbb{R}$:

$$\begin{aligned} &E_\theta h(\mathbf{X}) = 0 \qquad for\ all\ \theta \in \Theta \\ &\Longrightarrow \\ &P_\theta(h(\mathbf{X}) = 0) = 1 \qquad for\ all\ \theta \in \Theta \end{aligned}$$

A statistic T $\sim P_\theta^T$ is called complete *iff* the statistical model $\{ P_\theta^T : \theta \in \Theta \}$ is complete.

Theorem 4.1. *Assume that $\mathcal{P}$ is a k-parameter exponential family with natural parameter $\eta = (\eta_1, ..., \eta_k)$ and the natural parameter space $\mathcal{N}$ contains a non-empty k-dimensional interval. Then the statistic T(**X**) is sufficient and complete.*

[1]

Corollary 4.1.1. *Let us assume that P_θ belongs to a strictly k-parameter exponential family, then the statistic T(**X**) is sufficient and complete.*

Example 4.1 (Normal distribution)**.** Let **X** be a sample with distribution $N(\mu, \sigma^2)$ with statistics $\mathbf{T}_{(n)}(x) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$, then the image of the mapping

$$(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ \mapsto (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$$

contains an open subset of $\mathbb{R}^2$ hence by **Theorem 4.1** we can conclude that the statistics $\mathbf{T}_{(n)}(x)$ are sufficient and complete. By previous calculations in **Example 1.1** we also know that it belongs to an strict 2-parameter exponential family and therefore we could have used **Corollary 4.1.1** instead.

Example 4.2 (Normal Distribution)**.** We have a joint distribution $\mathbf{X} \sim N(\mu, \mu^2)$. With $Cov_\theta \mathbf{T}$ and statistic $\mathbf{T}_{(n)}(x) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$

$$Var(T_1) = ET_1^2 - (ET_1)^2$$

$$E_\theta T_1^2 = Var(T_1) + (ET_1)^2 = Var(\sum_{i=1}^n X_i) + (\sum_{i=1}^n EX_i)^2 = n\theta^2 + n^2\theta^2$$

$$E_\theta T_2 = nE_\theta X_i^2 = n((E_\theta X_i)^2 + Var(X_i)) = 2n\theta^2.$$

A relationship can be found between these and be described as a function; $g(\mathbf{T}) = 2T_1^2 - (n+1)T_2 \neq 0$ but this gives

$$E_\theta(g(\mathbf{T})) = E_\theta(2T_1^2 - (n+1)T_2) = 0.$$

The function $g(\mathbf{T}) \neq 0$ except for n = 1 and it gives that the statistic $\mathbf{T}$ is not complete and neither is the model. [2]

Example 4.3 (Simple structural Relation)**.** Consider a joint distribution of i.i.d variables distributed as the simple structural relation in **Example 2.4** with statistics $\mathbf{T}_{(n)}(z) = \Big(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i y_i, \sum_{i=1}^n y_i^2 \Big)$

$$E(T_1) = E(\sum_{i=1}^n x_i^2) = \sum_{i=1}^n E(X_i^2) = n(E([\xi_i + \delta_i]^2)) = 2n$$
$$E(T_2) = E(\sum_{i=1}^n x_i y_i) = n(E([\xi_i + \delta_i][\beta\xi_i + \varepsilon_i])) = n\beta$$
$$E(T_3) = E(\sum_{i=1}^n y_i^2) = n(E([\beta\xi_i + \varepsilon_i]^2)) = n(\beta + 1)$$

Consider the function $g(\mathbf{T}) = T_3 - T_2 - \frac{1}{2}T_1 \neq 0$, although it gives

$$E_\theta(g(\mathbf{T})) = E_\theta(T_3 - T_2 - \frac{1}{2}T_1) = n(\beta+1) - n\beta - n = 0.$$

The function $g(\mathbf{T}) \neq 0$ and results in that the statistic T is not complete and neither is the model.

For the curved examples we cannot use theorems for strictly k-parameter exponential families because in previous sections we concluded that k is not minimal. The examples for curved gives an hint that curved exponential families cannot be complete either since there is an linear dependence in the functions η and T which creates a problem with $\mathcal{N}$ containing a non-empty k-dimensional interval.

5 THE RAO-BLACKWELL AND LEHMANN-SCHEFFÉ THEOREMS

When one has evaluated the statistical model one wants to find a good estimator and also to state that it is the best one. The previous sections leads up to these theorems. [1]

Theorem 5.1 (Rao-Blackwell). *Let T be a sufficient statistic for the statistical model $\mathcal{P}$, and let $\tilde{\gamma}$ be an unbiased estimator for the parameter $\gamma = g(\theta) \in \mathbb{R}^k$. Define*

$$\hat{\gamma}(T) = E_\theta(\tilde{\gamma}|T) \tag{3}$$

The conditional expectation $\hat{\gamma}$ is independent of θ, i.e., $\hat{\gamma}(T) = E(\tilde{\gamma}|T)$. Furthermore, for all $\theta \in \Theta$

$$E_\theta \hat{\gamma} = g(\theta)$$

and

$$Cov_\theta \hat{\gamma} \preceq Cov_\theta \tilde{\gamma}$$

If trace$(Cov_\theta \tilde{\gamma}) < \infty$, *then* $Cov_\theta \hat{\gamma} = Cov_\theta \tilde{\gamma}$ *iff* $P_\theta(\hat{\gamma} = \tilde{\gamma}) = 1$.

Example 5.1 (Rao Blackwell Theorem and Curved Normal). Let $X_1, .., X_n$ be an i.i.d sample from $N(\mu, \mu^2)$. The sufficient statistic are $T = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ consider the unbiased estimator $\tilde{\gamma} = (\frac{1}{n}\sum_{i=1}^n x_i, \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2) = g(T)$.

$$\hat{\gamma}(T) = E(\tilde{\gamma}|T_1, T_2) = E(g(T)|T_1, T_2) = g(T) = (\frac{1}{n}\sum_{i=1}^n x_i, \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2)$$

Hence no improvement came from applying the Rao Blackwell theorem on the estimator for this statistical model.

The theorem of Rao-Blackwell says that you often can improve your estimator by taking the conditional expectation with respect to the sufficient statistic **T**. It can be difficult to find a crude estimator or to compute conditional expectations needed to apply the theorem. The theorem might seem weak but the application of it can lead to big improvements for the estimator and can give insights for the construction of estimators.

Theorem 5.2 (Lemann-Scheffé). *Let T be a sufficient and complete statistic for the statistical model $\mathcal{P}$, and let $\tilde{\gamma}_1$ be an unbiased estimator for the parameter $\gamma = g(\theta) \in \mathbb{R}^k$. Then the estimator*

$$\hat{\gamma} = \hat{\gamma}(T) = E(\tilde{\gamma}_1|T)$$

has the smallest covariance matrix among all unbiased estimators for the parameter $\gamma = g(\theta)$. *That is, for all estimators* $\tilde{\gamma}$ *with* $E_\theta \tilde{\gamma} = g(\theta)$ *we have*

$$Cov_\theta \hat{\gamma} \preceq Cov_\theta \tilde{\gamma} \qquad for\ all \quad \theta \in \Theta.$$

Example 5.2 (Normal distribution)**.** Consider the statistical model in **Example 1.1**, in the previous sections we concluded that the statistic T is sufficient and complete. We know that the sample mean and the uncorrected sample variance,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad S^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

are unbiased estimators for this model. Since they are functions of the sufficient statistics the theorem says that these are the best minimum variance unbiased estimator (UMVUE). That the meanbased estimator is best follows from the Cramér Rao Bound and the optimality of S^2 follows from the theorem of Lehmann Scheffé.

As we have seen for curved exponential families, the statistics will not be complete and therefore the Lemann Scheffé theorem fails to apply to them. The Lehmann Scheffé theorem gives the conclusion that if an estimator is complete, unbiased, and sufficient, then it's the best possible unbiased estimator and Rao Blackwell does not give us that conclusion. Therefore with the curved exponential families we have lost an important way to claim that the chosen estimator is the best. Curved exponential families will have sufficient statistics which means that one can apply the Rao Blackwell theorem. It is often a useful theorem because it only demands sufficiency and not completeness and it says that a crude estimator can be found to start with and then improve it by applying the theorem over and over again. Although with the curved exponential family we reach the conclusion that the Rao Blackwell theorem can be applied but does not improve the estimator.
We finish this statement with a corollary and a proof.

Corollary 5.2.1. *If you have a family of distributions which belongs to the curved exponential family then applying the Rao-Blackwell theorem will not improve the estimator.*

Proof. Consider the case $k = 2$ and assume that the Rao Blackwellization will improve the estimator. If the family belongs to a curved exponential family then there exists a sufficient statistic

$$\mathbf{T} = (T_1, T_2)$$

. Set the estimator to a function of the sufficient statistic which is unbiased $\tilde{\gamma} = g(T)$, and apply the Rao Blackwell theorem.

$$\hat{\gamma}(T) = E(\tilde{\gamma}|T_1, T_2) = E(g(T)|T_1, T_2) = g(T) = \tilde{\gamma}$$

Hence the Rao Blackwellization leads to no improvement of the estimator which contradicts the assumption.

$\square$

References

[1] H. Liero, S. Zwanzig, *Introduction to the Theory of Statistical Inference*. CRC press, 2012.

[2] R. W. Keener, *Theoretical Statistics*. Springer, 2010.

[3] H. Witting *Mathematische Statistik* Vieweg + Teubner Verlag, 1985.