Fisher Information and Exponential Families Parametrized by a Segment of Means

Piotr Graczyk, Salha Mamane

Fisher Information and Exponential Families Parametrized by a Segment of Means

Piotr Graczyk
LAREMA, Université d'Angers
Salha Mamane
School of Statistics and Actuarial Science, University of the Witwatersrand
and LAREMA, Université d'Angers

Abstract

We consider natural and general exponential families $(Q_m)_{m\in M}$ on $\mathbb{R}^d$ parametrized by the means. We study the submodels $(Q_{\theta m_1+(1-\theta)m_2})_{\theta\in[0,1]}$ parametrized by a segment in the means domain, mainly from the point of view of the Fisher information. Such a parametrization allows for a parsimonious model and is particularly useful in practical situations when hesitating between two parameters m_1 and m_2. The most interesting examples are obtained when $\mathbb{R}^d$ is a linear space of matrices, in particular for Gaussian and Wishart models.

1 Introduction

Fisher information is a key concept in mathematical statistics. Its importance stems from the Cramér-Rao inequality which says that the variance of any unbiased estimator $T(X_1,\dots X_n)$ of an unknown parameter θ, is bounded by the inverse of the Fisher information: $\mathrm{Var}_\theta(T)-(I(\theta))^{-1}$ is semi-positive definite. Fisher information is therefore mainly used as a measure of how well a parameter can be a estimated. This justifies the use of Fisher information in experimental design for predicting the maximum precision an experiment can provide on model parameters. This also justifies the important role Fisher information plays in estimation theory where it provides bounds for confidence regions and also in Bayesian analysis where it provides a basis for noninformative priors. Fisher information can be used to investigate the trade-off between parsimony of parameters and precision of the estimation of the parameters [Andersson and Handel, 2006].

Besides its importance in statistical theory, Fisher information has different interpretations that lead to some practical applications. For example, the interpretation of Fisher information

Supported by l'Agence Nationale de la Recherche ANR-09-BLAN-0084-01.
AMS Subject Classification: 62H12, 62H10
Keywords: Fisher information, efficient estimator, exponential family, multivariate Gaussian distribution, Wishart distribution, parsimony.

as a measure of the state of disorder of a dynamic system leads to the use of Fisher information in stochastic optimal control as a tuning tool to stabilise the performance of a dynamic system [Ramirez et al., 2010]. Viewing Fisher information as a measure of information, leads to the statement of a "minimum information principle" akin to the well-known maximum entropy principle for determining the "maximally unpresumptive distribution" satisfying some predefined constraints [Bercher and Vignat, 2009]. Gupta and Kundu [2006] describe the use of Fisher information in model selection as a tool to discriminate between two models with otherwise very similar fit to some data. The use of Fisher information however goes far beyond statistics; Frieden [2004] shows that Fisher information is in fact a key concept in the unification of science in general, as it allows a systematic approach to deriving Lagrangians.

The objective of this work is the study of the Fisher information for exponential models $(P_m)_{m\in M}$ parametrized by a segment of means $[m_1, m_2]$. Exponential families of distributions have been extensively studied [Brown, 1986; Barndorff-Nielsen, 1978; Letac, 1992; Letac and Casalis, 2000]. A parametrization of the family by a segment instead of the whole means domain allows to obtain a parsimonious model when the mean domain is high-dimensional. The parametrization of the mean parameter by a segment is particularly useful in practical situations when hesitating between two equally convenient mean values m_1 and m_2. Such parametrization will also serve in sequential data collection, when an updated estimate of a parameter largely differs from the previous estimate. An important practical example is a Gaussian model $N(u, \theta C + D)$ in $\mathbb{R}^d$ with the mean vector u known and the covariance matrix in a segment $IC + D$, where $\theta \in I = [a, b] \subset \mathbb{R}$.

From the Fisher information point of view, exponential families constitute an interesting and important class of models. Their Fisher information coincides with the second derivative of the cumulant generating function of the measure generating the family and they are the only models for which the Cramér-Rao bound can always be attained [Brown, 1986; Letac, 1992].

The paper is organised as follows. In Section 2, basic definitions and results on Fisher information and exponential families are recalled and extended to matrix-parametrized models. Section 3 contains new results on the Fisher information of exponential families parametrized by the domain of the means and the sub-families parametrized by a segment of means $[m_1, m_2]$. In Section 4, these results are applied to Gaussian and Wishart families of distributions. When m_1 and m_2 are colinear, we construct efficient estimators for the segment parameter θ.

2 Preliminaries

In most expositions of the theory of exponential families and of the concept of Fisher information, the parameter is considered to be a vector whereas cases abound in multivariate analysis where the canonical parameter is a matrix. In this preliminary section we adapt the presentation of the usual objects of exponential families (mean function, variance function and Fisher information) to the case where the canonical parameter is a matrix.

We denote by $\mathbb{R}^{k\times m}$ the space of real matrices with k rows and m columns and by $A \otimes B$ the Kronecker product of two matrices. We use the usual notations A^T for the transpose matrix and $\langle A, B\rangle = \mathrm{Tr}(A^T B)$ for the scalar product of two matrices. The operator Vec converts a $k \times m$

matrix A into a vector $\mathrm{Vec}(A) \in \mathbb{R}^{km}$ by stacking the columns one underneath the other. The Vec operator is commonly used in applications of the matrix differential calculus in statistics, cf. [Magnus and Neudecker, 2007; Muirhead, 2005].

The following properties of the Kronecker product are used in this work [Magnus and Neudecker, 2007, p.32,35]. For non-singular squared matrices A, B we have $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$. For all matrices A, B and C such that the product ABC is well defined

$$\mathrm{Vec}(A\,B\,C) = (C^T \otimes A)\,\mathrm{Vec}(B). \tag{1}$$

In this paper we use the following convention of the matrix differential calculus: if a function $f : \mathbb{R}^{k\times p} \to \mathbb{R}^{n\times m}$ is differentiable then its derivative is a matrix $f'(x) \in \mathbb{R}^{nm\times kp}$ such that

$$\mathrm{Vec}(df(x)(u)) = f'(x)\,\mathrm{Vec}(u), \quad u \in \mathbb{R}^{k\times p}. \tag{2}$$

The only exception we will make is the derivative of a function $K : \mathbb{R}^{k\times m} \to \mathbb{R}$, for which the following convention is used: the derivative of K is not a row vector but the matrix $K'(x) \in \mathbb{R}^{k\times m}$, related to the differential of K by $dK(x)(u) = \langle K'(x), u\rangle = \mathrm{Tr}(K'(x)^T u)$, for all $u \in \mathbb{R}^{k\times m}$. This convention is needed to give sense to formula (5) for the mean of an exponential family.

In this section we consider probability models $(P_s(d\omega))_{s\in S}$, $S \subset \mathbb{R}^{k\times m}$, on a measurable space $(\Omega, \mathcal{A})$ such that there exists a σ-finite positive measure ν on $(\Omega, \mathcal{A})$ and a real function $(\omega, s) \mapsto l_\omega(s)$ such that

$$P_s(d\omega) = e^{l_\omega(s)}\nu(d\omega)$$

and $\int e^{l_\omega(s)}\nu(d\omega) = 1$. These models encompass but are not reduced to exponential families of distributions. We suppose that S is open and that the function $s \to l_\omega(s)$ is twice differentiable. We impose on $l_\omega(s)$ classical regularity conditions, allowing double differentiation under the integral sign in $\int e^{l_\omega(s)}\nu(d\omega)$.

The score function $l'_\omega(s)$ is the derivative with respect to s of the log-likelihood function $l_\omega(s)$. It is a $k \times m$ matrix with zero mean. The extension of the definition of Fisher information matrix from vector-parametrized models to matrix-parametrized models is straightforward.

Definition 2.1. The Fisher information matrix of the model $(P_s(d\omega))_{s\in S}$, $S \subset \mathbb{R}^{k\times m}$ on a measurable space $(\Omega, \mathcal{A})$ is the $km \times km$ symmetric matrix

$$I(s) = \mathrm{Cov}(l'_\omega(s)) = \int_\Omega \mathrm{Vec}(l'_\omega(s))\,\mathrm{Vec}(l'_\omega(s))^T P_s(d\omega).$$

Similarly as for vector parametrized models, the Fisher information can be written as the negative of the mean of the second derivative of the log-likelihood function.

Proposition 2.1. *The Fisher information of the model $(P_s(d\omega))_{s\in S}$, $S \subset \mathbb{R}^{k\times m}$ equals*

$$I(s) = -\int_\Omega l''_\omega(s) P_s(d\omega).$$

Proof. As a probability distribution, $P_s(d\omega)$ satifies

$$\begin{aligned}
\int_\Omega P_s(d\omega) = 1 \quad &\Longrightarrow \quad \frac{d}{ds}\int_\Omega P_s(d\omega) = \int_\Omega \frac{d}{ds}P_s(d\omega) = \int_\Omega e^{l_\omega(s)} l'_\omega(s)\nu(d\omega) = 0_{\mathbb{R}^{k\times m}} \\
&\Longrightarrow \quad \frac{d}{ds}\int_\Omega e^{l_\omega(s)} l'_\omega(s)\nu(d\omega) = \int_\Omega \frac{d}{ds}\left(e^{l_\omega(s)} l'_\omega(s)\right)\nu(d\omega) = 0_{\mathbb{R}^{km\times km}} \\
&\Longrightarrow \quad \int_\Omega \left(e^{l_\omega(s)}\operatorname{Vec}(l'_\omega(s))\operatorname{Vec}(l'_\omega(s))^T + e^{l_\omega(s)} l''_\omega(s)\right)\nu(d\omega) = 0_{\mathbb{R}^{km\times km}}
\end{aligned}$$

and the Proposition follows. □

Consider a σ-finite measure space $(\Omega, \mathcal{A}, \nu)$, a finite-dimensional normed vector space E and a measurable map $T : \Omega \to E$. Let μ be the image of the measure ν by T and let S_μ be the interior of the domain of the moment generating function of μ, i.e. the set $\{s \in E : \int_E e^{\langle s,x\rangle} d\mu(x) < \infty\}$. We suppose that μ is a σ-finite measure on E and that $S_\mu \neq \emptyset$. The cumulant generating function of μ is defined as the logarithm of the moment generating function of μ:

$$k_\mu(s) = \log \int_E e^{\langle s,x\rangle}\mu(dx) = \log \int_\Omega e^{\langle s,T(\omega)\rangle}\nu(d\omega), \quad s \in S_\mu.$$

Definition 2.2. The general exponential family generated by the measure ν and the map T is the family of probability distributions

$$\{P(s,T,\nu)(d\omega) = e^{\langle s\,,T(\omega)\rangle - k_\mu(s)}\nu(d\omega) : \quad s \in S_\mu\}. \tag{3}$$

The natural exponential family associated with the above general exponential family is the family of probability distributions defined on the space E by

$$P(s,\mu)(dx) = e^{\langle s\,,x\rangle - k_\mu(s)}\mu(dx), \quad s \in S_\mu. \tag{4}$$

Natural exponential families may be viewed as a special case of general exponential families with $\Omega = E$, $T(\omega) = \omega$ and $\nu = \mu$. The following result is well-known for vector-valued and matrix-valued exponential families [Letac and Casalis, 2000].

Proposition 2.2.
1. *The set S_μ is convex. If μ is not concentrated on some affine hyperplane of E, then k_μ is a strictly convex function on S_μ.*
2. *The map $s \mapsto k'_\mu(s)$ is an analytic diffeomorphism from S_μ to its image $M = k'_\mu(S_\mu) \subset \mathbb{R}^{k\times m}$ called the domain of the means of the family. In particular M is open.*

The name "domain of the means" for the set M is justified by formula (5) of the following Proposition, well-known in the vector case.

Proposition 2.3. *The mean and covariance of a random matrix X following the distribution $P(s,\mu)$ belonging to the natural exponential family generated by a measure μ are given by*

$$m(s) = \mathbb{E}_s(X) = k'_\mu(s) \tag{5}$$
$$v(s) = \operatorname{Cov}_s(\operatorname{Vec}(X)) = k''_\mu(s). \tag{6}$$

Proof. Formula (5) follows from

$$k'_\mu(s) = \frac{\int_{S_\mu} x e^{\langle s,x\rangle}\mu(dx)}{\int_{S_\mu} e^{\langle s,x\rangle}\mu(dx)} = \int_{S_\mu} x e^{\langle s,x\rangle - k_\mu(s)}\mu(dx) = \mathbb{E}_s(X).$$

Next, using (5), we obtain

$$\begin{aligned} k''_\mu(s) &= \int_{S_\mu} \frac{d}{ds}\left(x e^{\langle s,x\rangle - k_\mu(s)}\right)\mu(dx) = \int_{S_\mu} \mathrm{Vec}(x)\,\mathrm{Vec}(x - k'_\mu(s))^T e^{\langle s,x\rangle - k_\mu(s)}\mu(dx) \\ &= \int_{S_\mu} \mathrm{Vec}(x - m(s))\,\mathrm{Vec}(x - m(s))^T e^{\langle s,x\rangle - k_\mu(s)}\mu(dx) = \mathrm{Cov}_s(\mathrm{Vec}(X)). \end{aligned}$$

□

Remark 2.1. *If W is a random matrix with a law $P(s, T, \nu)$ from the general exponential family, then $T(W) = X$ in law and $m(s)$ and $v(s)$ are the mean and the covariance of $T(W)$.*

Now we compute the Fisher information of the exponential families parametrized by the canonical parameter $s \in S_\mu$.

Proposition 2.4. *The Fisher information for the parameter s of exponential families (3) and (4) is given by*

$$I(s) = k''_\mu(s) = v(s). \tag{7}$$

Proof. The log-likelihood is equal to $l_\omega(s) = \langle s\,,\, T(\omega)\rangle - k_\mu(s)$, so $l''_\omega(s) = -k''_\mu(s)$ does not depend on ω. Formula (7) follows by Proposition 2.1. □

Definition 2.3. Denote by $\psi : M \to S_\mu$, $m \mapsto \psi(m) = {k'_\mu}^{-1}(m)$ the inverse of the diffeomorphism k'_μ. The general exponential family, parametrized by the domain of the means M is given by the family of distributions

$$Q(m, T, \nu)(d\omega) = e^{\langle \psi(m)\,,\, T(\omega)\rangle - k_\mu(\psi(m))}\nu(d\omega), \quad m \in M. \tag{8}$$

The natural exponential family, parametrized by the domain of the means M is the family of probability distributions defined on the space E by

$$Q(m, \mu)(dx) = e^{\langle \psi(m)\,,\, x\rangle - k_\mu(\psi(m))}\mu(dx), \quad m \in M. \tag{9}$$

The mean of the families (8) and (9) is equal to m. We denote the covariance of the families (8) and (9) by $V(m)$ and we have by (6)

$$V(m) = v(\psi(m)) = k''(\psi(m)). \tag{10}$$

The function $V : m \in M \to V(m)$ is called the *variance function* of the exponential family.

We will compute the Fisher information of the exponential families (8) and (9) parametrized by the mean $m \in M$ in the next section. We will need the following formula giving the Fisher information for a reparametrized model.

Theorem 2.1. *Consider a model $(P_s(d\omega))_{s\in S}$ and a reparametrization $f : \tilde{S} \subset \mathbb{R}^{k\times p} \to S \subset \mathbb{R}^{n\times m}$, where f is a differentiable map. Let $I(s)$ be the information matrix of $(P_s(d\omega))_{s\in S}$. The Fisher information matrix of the model $(Q_t(d\omega))_{t\in\tilde{S}} = \left(P_{f(t)}(d\omega)\right)_{t\in\tilde{S}}$ is*

$$\tilde{I}(t) = f'(t)^T I(f(t)) f'(t). \tag{11}$$

Proof. Let us denote $h_\omega(t) = l_\omega(f(t))$. We have $Q_t(d\omega) = e^{h_\omega(t)}\nu(d\omega)$. For all $t \in \tilde{S}$ and $u \in \mathbb{R}^{k\times p}$,

$$\begin{aligned} dh_\omega(t)(u) &= dl_\omega(f(t))(df(t)(u)) = \langle l'_\omega(f(t)), df(t)(u)\rangle = \\ &= \operatorname{Vec}(l'_\omega(f(t)))^T \operatorname{Vec}(df(t)(u)) = \operatorname{Vec}(l'_\omega(f(t)))^T f'(t) \operatorname{Vec}(u). \end{aligned}$$

Thus, using the convention introduced after (2), $\operatorname{Vec}(h'_\omega(t))^T = \operatorname{Vec}(l'_\omega(f(t)))^T f'(t)$ and

$$\operatorname{Vec}(h'_\omega(t)) \operatorname{Vec}(h'_\omega(t))^T = f'(t)^T \operatorname{Vec}(l'_\omega(f(t))) \operatorname{Vec}(l'_\omega(f(t)))^T f'(t).$$

Therefore, by Definition 2.1 we get $\tilde{I}(t) = f'(t)^T I(f(t)) f'(t)$. ☐

3 Fisher information of exponential families parametrized by the mean

In this section we first compute the Fisher information of the exponential families (8) and (9) parametrized by the mean. Next we consider the same problem for a submodel parametrized by a segment of means. In order to avoid confusion, when the parameter of an exponential family is the mean m we will denote the Fisher information by $J(m)$.

Theorem 3.1. *The Fisher information of the exponential families (8) and (9) parametrized by the mean $m \in M$ equals*

$$J(m) = V(m)^{-1} = \psi'(m), \tag{12}$$

where $V(m)$ is the variance function of the exponential family, given by (10).

Proof. We use Theorem 2.1 with $f = \psi : M \to S_\mu$. Since $\psi(m) = {k'_\mu}^{-1}(m)$, we have $\psi'(m) = [k''_\mu(\psi(m))]^{-1}$. Thus $J(m) = [k''_\mu(\psi(m))]^{-1} k''_\mu(\psi(m)) [k''_\mu(\psi(m))]^{-1} = [k''_\mu(\psi(m))]^{-1} = V(m)^{-1}$. ☐

Remark 3.1. *Note a striking contrast in the formulas (7) and (12) for the Fisher information of an exponential family parametrized either by the canonical parameter $s \in S_\mu$ or by the mean $m \in M$; in the first case we have $I(\psi(m)) = V(m)$, in the second $J(m) = V(m)^{-1}$.*

3.1 Fisher information of exponential families parametrized by a segment of means

Consider a general exponential family $\{Q(m, T, \nu)(d\omega) \ : \ m \in M\}$ parametrized by the domain of the means M. Let $A \neq 0, B \in \mathbb{R}^{k\times m}$ be two matrices. Define $\Theta = \{\theta \in \mathbb{R} : \theta A + B \in M\}$. The set $\Theta \subset \mathbb{R}$ is open because M is open. Suppose that $\Theta \neq \emptyset$. The parametrization by a segment of means $I \subset \Theta$ consists in considering the submodel

$$\{Q(\theta A + B, T, \nu) : \theta \in I\}. \tag{13}$$

In statistical practice, the following situation will be concerned by such models. Let $m_1 \in M$ and $m_2 \in M$ be two different estimations of the true mean m of an exponential family (8) or (9) . When one hesitates between them as estimators, and when M is convex, it is natural to consider the model

$$\{Q(\theta m_1 + (1-\theta)m_2, T, \nu) : \theta \in [0,1]\}.$$

Writing $\theta m_1 + (1-\theta)m_2 = \theta(m_1 - m_2) + m_2$ we see that this is a special case of the model (13).

The following theorem gives the Fisher information of a general exponential family parametrized by a segment of means. By analogy to the notation $J(m)$, we denote this information by $J(\theta)$.

Theorem 3.2. *The Fisher information of the model* $\{Q(\theta A + B, T, \nu) : \theta \in I\}$ *equals*

$$J(\theta) = \mathrm{Vec}(A)^T\, V(\theta A + B)^{-1}\, \mathrm{Vec}(A). \tag{14}$$

Remark 3.2. *This and the following results are also true for submodels* $\{Q(\theta A + B, \mu) : \theta \in I\}$ *of natural exponential families.*

Proof. By Theorem 3.1, the Fisher information of the model $\{Q(m, T, \nu)\ : m \in M\}$ is $J(m) = V^{-1}(m)$. We apply Theorem 2.1 to the reparametrization $f : I \to M, f(\theta) = \theta A + B$. We have $f'(\theta) = \mathrm{Vec}(A)$. Then formula (11) gives (14). □

The following Lemma is useful for the derivation of an alternative formula for the Fisher information of an exponential family parametrized by a segment of means and verifying an additional condition (15). We will see in Section 4 that this condition holds for Gaussian and Wishart models.

Lemma 3.1. *Assume that for all* $m \in M$,

$$\langle m\,,\, \psi(m)\rangle = C, \tag{15}$$

for some constant $C \in \mathbb{R}$. *Then, for all* $u \in M$,

$$\langle m\,,\, d\psi(m)(u)\rangle = -\langle u\,,\, \psi(m)\rangle. \tag{16}$$

Proof. By (15) the differential of the function $g : M \to \mathbb{R}$, $m \mapsto \langle m\,,\, \psi(m)\rangle$ is zero. Therefore, for all $m, u \in M$

$$dg(m)(u) = \langle m\,,\, d\psi(m)(u)\rangle + \langle u\,,\, \psi(m)\rangle = 0$$

and (16) follows. □

Corollary 3.1. *Let* $\{Q(\theta A + B, T, \nu)(d\omega) : \theta \in I\}$ *be an exponential model parametrized by a segment of means. If the condition (15) holds then the Fisher information of the model equals*

$$J(\theta) = -\frac{d^2}{d\theta^2}\left[k_\mu(\psi(\theta A + B))\right]. \tag{17}$$

Proof. Let $h(\theta) = k_\mu(\psi(\theta A + B))$ and $f(\theta) = \theta A + B$. We want to compute $h''(\theta)$. If $\theta, u \in \mathbb{R}$,

$$\begin{aligned} dh(\theta)(u) &= dk_\mu(\psi(f(\theta)))\big(d\psi(f(\theta))(df(\theta)(u))\big) \\ &= \langle k'_\mu(\psi(f(\theta)))\,,\, d\psi(f(\theta))(df(\theta)(u))\rangle \\ &= \langle f(\theta)\,,\, d\psi(f(\theta))(df(\theta)(u))\rangle \\ &= -\langle df(\theta)(u)\,,\, \psi(f(\theta))\rangle \\ &= -u\langle A\,,\, \psi(f(\theta))\rangle, \end{aligned}$$

where we used successively: the convention on k'_μ introduced after (2), the equality $k'_\mu \circ \psi(m) = m$, Lemma 3.1 and the formula $df(\theta)(u) = uA$. Thus we have $h'(\theta) = -\langle A\,,\, \psi(f(\theta))\rangle$. Now, starting as in the computation of $h'(\theta)$ and using (2), we get

$$h''(\theta) = -\langle A\,,\, d\psi(f(\theta))(A)\rangle = -\operatorname{Vec}(A)^T \operatorname{Vec}(d\psi(f(\theta))(A)) = -\operatorname{Vec}(A)^T \psi'(\theta A + B)\operatorname{Vec}(A).$$

We conclude using (12) and Theorem 3.2. □

4 Applications

In this section, we apply the results from preceding sections to the study of some important exponential families parametrized by a segment of means.

We denote by $\mathcal{S}_d$ the vector space of $d \times d$ symmetric matrices and by $\mathcal{S}_d^+$ the open cone of positive definite matrices.

4.1 Exponential families of Gaussian distributions

Let us recall the construction of the multivariate Gaussian model $\{N(u, \Sigma);\, \Sigma \in \mathcal{S}_d^+\}$ as a general exponential family. We consider $\Omega = \mathbb{R}^d$ equiped with a normalised Lebesgue measure $\nu(d\omega) = d\omega/(2\pi)^{d/2}$, the space $E = \mathcal{S}_d$ and the map

$$T : \mathbb{R}^d \to \mathcal{S}_d, \quad T(\omega) = -\frac{1}{2}(\omega - u)(\omega - u)^T.$$

The image measure μ on E is concentrated on the opposite of the cone of semi-positive definite matrices of rank one. For $s \in \mathcal{S}_d$, the moment generating function of μ equals

$$\int_\Omega e^{\langle s, T(\omega)\rangle}\nu(d\omega) = \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{-\frac{1}{2}\operatorname{Tr}\left(s(\omega-u)(\omega-u)^T\right)}d\omega = (\det s)^{-1/2}$$

when $s \in \mathcal{S}_d^+$ and it is infinite otherwise. Thus $S_\mu = \mathcal{S}_d^+$ and the cumulant function is

$$k_\mu(s) = -\frac{1}{2}\log\det(s), \quad s \in S_\mu = \mathcal{S}_d^+.$$

The general exponential family is therefore

$$P(s, T, \nu)(d\omega) = \frac{1}{(2\pi)^{d/2}} e^{\langle s\,,\, -\frac{1}{2}(\omega-u)(\omega-u)^T\rangle + \frac{1}{2}\log\det(s)}d\omega = \frac{(\det s)^{1/2}}{(2\pi)^{d/2}} e^{-\frac{1}{2}(\omega-u)^T s(\omega-u)}d\omega, \tag{18}$$

which is the family of Gaussian distributions $N(u, s^{-1})$ on $\mathbb{R}^d$ with a fixed mean $u \in \mathbb{R}^d$, parametrized by $s = \Sigma^{-1}$, the inverse of the covariance matrix Σ supposed to be invertible.

The derivative of the function $X \in \mathbb{R}^{d\times d} \to \det X$ is the cofactor matrix $X^\sharp$ which equals $(\det X)(X^{-1})^T$ when X is inversible. It follows that

$$m(s) = k'_\mu(s) = -\frac{1}{2}s^{-1}, \quad s \in \mathcal{S}_d^+.$$

This can be also deduced from Remark 2.1; if W is a random vector with law $N(u, s^{-1})$, then

$$m(s) = k'_\mu(s) = \mathbb{E}\,T(W) = \mathbb{E}(-\frac{1}{2}(W-u)(W-u)^T) = -\frac{1}{2}\operatorname{Cov} W \in -\mathcal{S}_d^+.$$

The means domain is $M = -\mathcal{S}_d^+$ and the inverse mean map is $\psi(m) = -\frac{1}{2}m^{-1}$. The Gaussian general exponential family parametrized by m in the means domain $M = -\mathcal{S}_d^+$ is therefore the family

$$Q(m, T, \nu) = N(u, -2m). \tag{19}$$

Up to a trivial affine change of parameter $\Sigma = -2m$, this parametrization by the covariance parameter is more natural than the parametrization of the family $(N(u, s^{-1}))_{s\in\mathcal{S}_d^+}$ by the canonical parameter s.

In order to compute the variance function, recall that $XX^{-1} = I_d$ implies that $dX^{-1} = -X^{-1}dX\,X^{-1}$ and $(X^{-1})' = -X^{-1} \otimes X^{-1}$. Thus $k''_\mu(s) = \frac{1}{2}s^{-1} \otimes s^{-1}$ and formula (10) implies that

$$V(m) = 2m \otimes m. \tag{20}$$

By Proposition 2.4, the Fisher information of the family $(N(u, s^{-1}))_{s\in\mathcal{S}_d^+}$ is $I(s) = \frac{1}{2}s^{-1} \otimes s^{-1}$.

By Theorem 3.1 and formula (20), the Fisher information of the model $(N(u, -2m))_{m\in-\mathcal{S}_d^+}$ equals $J(m) = \frac{1}{2}m^{-1} \otimes m^{-1}$.

Corollary 4.1. *The Fisher information matrix of the Gaussian model $(N(u, \Sigma))_{\Sigma\in\mathcal{S}_d^+}$ is*

$$J(\Sigma) = \frac{1}{2}\Sigma^{-1} \otimes \Sigma^{-1}.$$

Proof. Using Theorem 2.1 and a reparametrization $\Sigma = -2m$ we obtain $\tilde{J}(\Sigma) = \frac{1}{2}\Sigma^{-1} \otimes \Sigma^{-1} = J(\Sigma)$. □

Let us now consider Gaussian models parametrized by a segment of covariances.

Corollary 4.2. *Let C and D be two symmetric matrices and let $I \subset \mathbb{R}$ be a non-empty segment such that $I \subset \Theta = \{\theta \in \mathbb{R} : \theta C + D \in \mathcal{S}_d^+\}$. The Fisher information of the Gaussian model $\{N(u, \theta C + D), \theta \in I\}$ is*

$$J(\theta) \;=\; \frac{1}{2}\operatorname{Tr}\left(C(\theta C + D)^{-1}C(\theta C + D)^{-1}\right).$$

Proof. We use Corollary 4.1 and Theorem 2.1 with $f(\theta) = \theta C + D$. It follows that

$$J(\theta) = \mathrm{Vec}(C)^T J(\theta C + D)\,\mathrm{Vec}(C) = \frac{1}{2}\,\mathrm{Vec}(C)^T \left((\theta C + D)^{-1} \otimes (\theta C + D)^{-1}\right) \mathrm{Vec}(C).$$

Applying (1) we get

$$J(\theta) = \frac{1}{2}\,\mathrm{Vec}(C)^T\,\mathrm{Vec}\left((\theta C + D)^{-1} C (\theta C + D)^{-1}\right) = \frac{1}{2}\,\mathrm{Tr}\left(C(\theta C + D)^{-1} C (\theta C + D)^{-1}\right).$$

□

On the other hand we have the following alternative formula for the information $J(\theta)$.

Corollary 4.3. *The Fisher information of the Gaussian model* $\{N(u, \theta C + D), \theta \in I\}$ *is*

$$J(\theta) = -\frac{1}{2}\frac{d^2}{d\theta^2}(\log\det(\theta C + D)). \tag{21}$$

Proof. Observe that the condition (15) holds for the Gaussian exponential families $Q(m, t, \nu)$:

$$\langle m\,,\,\psi(m)\rangle = -\frac{1}{2}\,\mathrm{Tr}(mm^{-1}) = -\frac{d}{2}.$$

The model $N(u, \theta C + D) = N(u, -2m) = Q(m, T, \nu)$ with $m = \theta A + B \in M = -\mathcal{S}_d^+$ where $A = -\frac{C}{2}$ and $B = -\frac{D}{2}$. We apply Corollary 3.1 and the fact that

$$k_\mu(\psi(\theta A + B)) = -\frac{1}{2}\log\det(\theta C + D).$$

Formula (21) follows. □

Now we characterize the information $J(\theta)$ in terms of the eigenvalues of the matrix $D^{-1/2}CD^{-1/2}$.

Theorem 4.1. *Let* C *and* D *be two symmetric matrices and let* $I \subset \mathbb{R}$ *be a segment such that* $IC + D \subset \mathcal{S}_d^+$. *Let* $a_1, \ldots, a_d$ *be the eigenvalues of the matrix* $D^{-1/2}CD^{-1/2}$.

The Fisher information of the Gaussian model $\{N(u, \theta C + D), \theta \in I\}$ *equals*

$$J(\theta) = \frac{1}{2}\sum_{j=1}^{d}\left(\frac{a_j}{1 + a_j\theta}\right)^2. \tag{22}$$

Proof. The idea of the proof is to use formula (21). Let $P(\lambda)$ be the characteristic polynomial of the matrix $D^{-1/2}CD^{-1/2}$. We have

$$P(\lambda) = \det(D^{-1/2}CD^{-1/2} - \lambda I_n) = \det(D^{-1}C - \lambda I_n) = (\det D)^{-1}\det(C - \lambda D).$$

On the other hand $P(\lambda) = \prod_{j=1}^{n}(a_j - \lambda)$. It follows that

$$\det(\theta C + D) = \det D \times \theta^d P(-1/\theta) = \det D \prod_{j=1}^{d}(\theta a_j + 1).$$

The last formula allows to compute easily $\frac{d^2}{d\theta^2}(\log\det(\theta C + D))$. First we see that

$$\frac{d}{d\theta}(\log\det(\theta C + D)) = \frac{\frac{d}{d\theta}\det(\theta C + D)}{\det(\theta C + D)} = \sum_{j=1}^{d} \frac{a_j}{\theta a_j + 1}.$$

One more derivation and formula (21) lead to (22). $\square$

We finish by computing the Fisher information of two Gaussian models in $\mathbb{R}^d$, parametrized by an explicitely given segment of covariances. First, let A be a circulant matrix with the first row $e_2 + e_d = (0, 1, 0, \dots, 0, 1)$. Then for a segment $I \subset \mathbb{R}$ containing 0 and $\theta \in I$

$$\theta A + I_d = \begin{pmatrix} 1 & \theta & 0 & \dots & 0 & \theta \\ \theta & 1 & \theta & 0 & \dots & 0 \\ 0 & \theta & 1 & \theta & 0 & \dots \\ & & \ddots & \ddots & \ddots & \\ 0 & \dots & 0 & \theta & 1 & \theta \\ \theta & 0 & \dots & 0 & \theta & 1 \end{pmatrix} \in \mathcal{S}_d^+. \tag{23}$$

Corollary 4.4. *The Fisher information of the model $(N(0, \theta A + I_d))_{\theta \in I}$ is given by*

$$J(\theta) \quad = \quad \frac{1}{2}\sum_{j=0}^{d-1}\left(\frac{2\cos(\frac{2\pi j}{d})}{1 + 2\theta\cos(\frac{2\pi j}{d})}\right)^2. \tag{24}$$

Proof. Let $\mathcal{A}$ be a circulant matrix with the first row $(r_0, r_1, \dots, r_{d-1})$. It is well known (see e.g. [Gray, 2006]) and easy to check that if ϵ is a d-th root of unity, $\epsilon^d = 1$, then $a = \sum_{l=0}^{d-1} r_l \epsilon^l$ is an eigenvalue of $\mathcal{A}$ with an eigenvector $(1, \epsilon, \epsilon^2, \dots, \epsilon^{d-1})$.

Therefore if $\epsilon_j = e^{\frac{2\pi j i}{d}}$, $j = 0, \dots, d-1$ are the d distinct d-th roots of unity, then the matrix $\mathcal{A}$ has d distinct eigenvalues $a_j = \sum_{l=0}^{d-1} r_l \epsilon_j^l$. In our particular case,

$$a_j = e^{\frac{2\pi j i}{d}} + e^{\frac{2(d-1)\pi j i}{d}} = 2\cos\left(\frac{2\pi j}{d}\right).$$

Formula (24) follows from Theorem 4.1. $\square$

Now, let us consider a tridiagonal matrix C such that

$$\theta C + I_d = \begin{pmatrix} 1 & \theta & 0 & 0 & 0 & \dots \\ \theta & 1 & \theta & 0 & 0 & \dots \\ 0 & \theta & 1 & \theta & 0 & \dots \\ & \ddots & \ddots & \ddots & \ddots & \\ 0 & \dots & 0 & \theta & 1 & \theta \\ 0 & \dots & 0 & 0 & \theta & 1 \end{pmatrix}. \tag{25}$$

As in the preceding case, there exists a segment $I \subset \mathbb{R}$ such that $\theta C + I_d \in \mathcal{S}_d^+$ for $\theta \in I$.

Corollary 4.5. *The Fisher information of the model* $(N(0, \theta C + I_d))_{\theta \in I}$ *is given by*

$$J(\theta) = \frac{1}{2} \sum_{j=1}^{d} \left(\frac{2\cos\left(\frac{j}{d+1}\pi\right)}{1 + 2\theta \cos\left(\frac{j}{d+1}\pi\right)} \right)^2. \tag{26}$$

Proof. We will apply Theorem 4.1 with C and $D = I_d$. Expanding $\psi_d(\lambda) = \det(C - \lambda I_d)$ along the first row, we get $\psi_d(\lambda) = -\lambda \psi_{d-1}(\lambda) - M^{1,2}$. Expanding the minor $M^{1,2}$ along its first column gives $M^{1,2} = \psi_{d-2}(\lambda)$ and

$$\psi_d(\lambda) = -\lambda \psi_{d-1}(\lambda) - \psi_{d-2}(\lambda), \ d \geq 3.$$

We set $\varphi_d(\lambda) = (-1)^d \psi_d(2\lambda)$ and we obtain

$$\varphi_d(\lambda) = 2\lambda \varphi_{d-1}(\lambda) - \varphi_{d-2}(\lambda), \ d \geq 3$$

with initial conditions $\varphi_1(\lambda) = 2\lambda$, $\varphi_2(\lambda) = 4\lambda^2 - 1$. Therefore φ_d is a Tchebyshev polynomial of the second kind [Mason and Handscomb, 2003] and it satisfies $\varphi_d(\cos x) = \frac{\sin(d+1)x}{\sin x}$, $d \geq 1$.

We have, for all $\lambda \in [-2, 2]$,

$$\psi_d(\lambda) = 0 \iff \varphi_d\left(\frac{\lambda}{2}\right) = 0 \implies \frac{\sin(d+1)x}{\sin x} = 0, \quad x = \arccos\frac{\lambda}{2}.$$

Therefore $\lambda_j = 2\cos\left(\frac{j}{d+1}\pi\right)$, $1 \leq j \leq d$, are d distinct eigenvalues of the matrix C. □

4.2 Exponential families of Wishart distributions

Let $E = \mathcal{S}_d$ be the space of symmetric real matrices of order d. The Riesz measures μ_p on the cone $\overline{\mathcal{S}_d^+}$ are unbounded positive measures such that their Laplace transform equals for $t \in \mathcal{S}_d^+$

$$\mathcal{L}\,\mu_p(t) = \int_{\overline{\mathcal{S}_d^+}} e^{-\langle t, x \rangle} d\mu_p(x) = (\det t)^{-p}.$$

By the celebrated Gindikin theorem, such measures exist if and only if p belongs to the Gindikin set $\Lambda_d = \{\frac{1}{2}, \ldots, \frac{d-1}{2}\} \cup \left(\frac{d-1}{2}, \infty\right)$. They are supported by the cone $\overline{\mathcal{S}_d^+}$ if and only if $p > \frac{d-1}{2}$ and they are absolutely continuous in that case, with a density $\Gamma_d(p)^{-1}(\det x)^{p - \frac{d+1}{2}}$, $x \in \mathcal{S}_d^+$, $\Gamma_d(p) = \Gamma(p)\Gamma(p - \frac{1}{2}) \ldots \Gamma(p - \frac{d-1}{2})$. Otherwise, when $p \in \{\frac{1}{2}, \ldots, \frac{d-1}{2}\}$, the measures μ_p are singular and concentrated on semipositive symmetric matrices of rank $2p$.

The family of Wishart distributions $W(p; s)$ on $\overline{\mathcal{S}_d^+}$ is defined as the natural exponential family generated by the Riesz measure μ_p. According to (4), it means that $p \in \Lambda_d$, $s \in S_{\mu_p} = -\mathcal{S}_d^+$ and

$$W(p; s)(dx) = \frac{e^{\langle s\,,\, x \rangle}}{\mathcal{L}\,\mu_p(-s)} \mu_p(dx) = e^{\langle s\,,\, x \rangle} (\det(-s))^p \mu_p(dx) = e^{\langle s\,,\, x \rangle - k_{\mu_p}(s)} \mu_p(dx),$$

with $k_{\mu_p}(s) = -p \log\det(-s)$. It follows that $\mathcal{L}\,W(p; s)(t) = \det(I_d + (-s)^{-1} t)^{-p}$ and that $\mu_p(dx) = e^{\operatorname{Tr} x} W(p; -I_d)$.

Wishart distributions are multivariate analogs of the gamma distributions $\lambda^p\Gamma(p)^{-1}e^{-\lambda x}x^{p-1}dx$ on $\mathbb{R}^+$($p>0,\lambda>0$), considered with a canonical parameter $s=-\lambda<0$. Similarly as in dimension 1, the Wishart distributions are often parametrized by a scale parameter $\sigma=(-s)^{-1}\in\mathcal{S}_d^+$ and then the notation $\gamma(p;\sigma)=W(p;(-\sigma)^{-1})$ is used, cf. [Letac and Massam, 2008]. The study of Wishart distributions is motivated by their importance as estimators of the covariance matrix of a Gaussian model in $\mathbb{R}^d$.

Let us apply our results on the Fisher information to a natural exponential family of Wishart distributions $\{W(p;s):\ s\in-\mathcal{S}_d^+\}$. The mean equals $m(s)=k'_{\mu_p}(s)=p(-s)^{-1}\in M=\mathcal{S}_d^+$ and the inverse mean map $\psi:\mathcal{S}_d^+\to-\mathcal{S}_d^+$ is $\psi(m)=-pm^{-1}$.

Thus the Wishart family $Q(m,\mu_p)$ parametrized by the domain of means is, up to a trivial reparametrization $m\to\frac{1}{p}m$, the family parametrized by its scale parameter:

$$Q(m,\mu_p)=W(p;-pm^{-1})=\gamma(p;\frac{1}{p}m),\quad m\in\mathcal{S}_d^+. \tag{27}$$

As $v(s)=k''_{\mu_p}(s)=p(s^{-1}\otimes s^{-1})$, it follows that the variance function is

$$V(m)=\frac{1}{p}(m\otimes m). \tag{28}$$

By Proposition 2.4, the Fisher information of the model $\{W(p;s):\ s\in-\mathcal{S}_d^+\}$ is $I(s)=ps^{-1}\otimes s^{-1}$. By Theorem 3.1 the Fisher information of the model $\{Q(m,\mu_p),\ m\in M\}$ is $J(m)=pm^{-1}\otimes m^{-1}$.

Consequently, using Theorem 2.1 and a reparametrization $m\to\frac{1}{p}m=\sigma$ we see that the the Fisher information matrix of the Wishart model $\{\gamma(p;\sigma):\ \sigma\in\mathcal{S}_d^+\}$ parametrized by a scale parameter σ equals $J(\sigma)=p\sigma^{-1}\otimes\sigma^{-1}$.

Theorem 4.2. *Let $I=(a,b)\subset\mathbb{R}$ and $C,D\in\mathcal{S}_d$ such that $IC+D\subset\mathcal{S}_d^+$. The Fisher information $J(\theta)$ of the Wishart model $\{\gamma(p;\theta C+D):\ \theta\in I\}$ verifies the formulas*

$$\begin{aligned} J(\theta) &= p\,\mathrm{Tr}\left(C(\theta C+D)^{-1}\right)^2 && (29)\\ J(\theta) &= -p\frac{d^2}{d\theta^2}(\log\det(\theta C+D)) \\ J(\theta) &= p\sum_{j=1}^{d}\left(\frac{a_j}{1+a_j\theta}\right)^2 && (30)\end{aligned}$$

where $a_1,\dots,a_d$ are the eigenvalues of the matrix $D^{-1/2}CD^{-1/2}$.

Proof. The proofs are similar to the proofs of the analogous results for exponential Gaussian families in the previous subsection. The condition (15) holds true: $\langle m\,,\,\psi(m)\rangle=-pd$, the model $\{\gamma(p;\theta C+D):\ \theta\in I\}$ is equal to the model $\{Q(\theta pC+pD,\mu_p):\ \theta\in I\}$ parametrized by the means and we have $k_\mu(\psi(\theta pC+pD))=p\log\det(\theta C+D)$. □

Corollary 4.6. *Let $\sigma_1,\sigma_2\in\mathcal{S}_d^+$ and let I be the open interval containing θ such that $\sigma_\theta=\theta\sigma_1+(1-\theta)\sigma_2\in S_d^+$. The Fisher information of the model $\{\gamma(p;\sigma_\theta):\ \theta\in I\}$ is equal to $J(\theta)=p\,\mathrm{Tr}\left(\left((\sigma_1-\sigma_2)\,\sigma_t^{-1}\right)^2\right)$.*

Proof. We write $\theta\sigma_1 + (1-\theta)\sigma_2 = \theta(\sigma_1 - \sigma_2) + \sigma_2$ and we apply formula (29). $\square$

Using (30) we obtain the following corollary, analogous to Corollaries 4.4 and 4.5.

Corollary 4.7. *1. Consider the model $\{\gamma(p;\theta A + I_d) : \theta \in I\}$ with $\theta A + I_d$ as in (23). Then its Fisher information equals $J(\theta) = p\sum_{j=0}^{d-1}\left(\frac{2\cos(\frac{2\pi j}{d})}{1+2\theta\cos(\frac{2\pi j}{d})}\right)^2$.*
2. Consider the model $\{\gamma(p;\theta C + I_d) : \theta \in I\}$ with $\theta C + I_d$ as in (25). Then its Fisher information equals $J(\theta) = p\sum_{j=1}^{d}\left(\frac{2\cos\left(\frac{j}{d+1}\pi\right)}{1+2\theta\cos\left(\frac{j}{d+1}\pi\right)}\right)^2$.

Remark 4.1. *Let $P(s,\mu)$ be the natural exponential family corresponding to the Gaussian general exponential family (18). If W has the law $N(u, s^{-1})$ given by (18), then $T(W)$ has the law $P(s,\mu)$. On the other hand it is well known that $-T(W) = \frac{1}{2}(W-u)(W-u)^T$ has the Wishart law $\gamma(\frac{1}{2}; 2s^{-1})$. This explains why the formulas for the Fisher information are the same for the Gaussian family and for the Wishart family with $p = \frac{1}{2}$.*

Exponential families of noncentral Wishart distributions. Let us finish the section on the Wishart models by considering the non-central case. The main reference is [Letac and Massam, 2008].

Let $p \in \Lambda_d$, $a \in \overline{\mathcal{S}_d^+}$ and $\sigma \in \mathcal{S}_d^+$. The noncentral Wishart distribution $\gamma(p,a;\sigma)$ is defined by its Laplace transform

$$\mathcal{L}\,\gamma(p,a;\sigma)(t) = \int_{\overline{\mathcal{S}_d^+}} e^{-\operatorname{Tr}(tx)}\gamma(p,a;\sigma)(dx) = \det(I_d + \sigma t)^{-p} e^{-\operatorname{Tr}\left(t(I_d+\sigma t)^{-1}\sigma a \sigma\right)},$$

for all $t \in \mathcal{S}_d^+$. When $p \geq \frac{d-1}{2}$, then non-central Wishart laws exist for all $a \in \overline{\mathcal{S}_d^+}$; when $p \in \{\frac{1}{2}, \ldots, \frac{d-2}{2}\}$ then a must be of rank at most $2p$ [Letac and Massam, 2011]. When $p = \frac{n}{2}$, $n \in \mathbb{N}$, the non-central Wishart distributions are constructed in the following way from n independent d-dimensional Gaussian vectors $Y_1, \ldots, Y_n$. Let $Y_j \sim N(m_j\,, \Sigma)$ and let M be the $d \times n$ matrix $[m_1, \ldots, m_n]$. Then, the $d \times d$ matrix $W = Y_1 Y_1^T + \ldots + Y_n Y_n^T$ has the noncentral Wishart distribution $\gamma(p,a;\sigma)$ with $p = \frac{n}{2}$, $\sigma = 2\Sigma$ and $\sigma a \sigma = MM^T$. Such Wishart distributions are studied in [Muirhead, 2005].

The non-central Wishart distributions may be constructed as a natural exponential family $\{W(p,a;s) : s \in -\mathcal{S}_d^+\}$ generated by the positive measure $\mu = \mu_{a,p}(dx) = e^{\operatorname{Tr}(a+x)}\gamma(p,a;I_d)(dx)$. Its moment generating function is given for $s \in -\mathcal{S}_d^+$ by

$$\int_{\overline{\mathcal{S}_d^+}} e^{\operatorname{Tr}(sx)}\mu_{a,p}(dx) = \det(-s)^{-p} e^{\operatorname{Tr}\left(a(-s)^{-1}\right)}.$$

We have $W(p,a;s) = \gamma(p,a;(-s)^{-1})$. Like for central Wishart families, $S_\mu = -\mathcal{S}_d^+$. The cumulant generating function is

$$k_\mu(s) = -p\log\det(-s) + \operatorname{Tr}(a(-s)^{-1}).$$

As before, we denote $\sigma = (-s)^{-1}$. We see that the mean equals

$$m(s) = k'_\mu(s) = p(-s)^{-1} + (-s)^{-1}a(-s)^{-1} = p\sigma + \sigma a \sigma \tag{31}$$

and the covariance

$$v(s) = k''_\mu(s) = p\sigma \otimes \sigma + (\sigma a \sigma) \otimes \sigma + \sigma \otimes (\sigma a \sigma) = -p\sigma \otimes \sigma + m \otimes \sigma + \sigma \otimes m. \tag{32}$$

When the matrix a is non-singular, the inverse mean map $\psi(m) = s$ is such that

$$(-s)^{-1} = \sigma = -\frac{p}{2}a^{-1} + a^{-1/2}\left(a^{1/2}ma^{1/2} + \frac{p^2}{4}I_d\right)^{1/2} a^{-1/2}. \tag{33}$$

For other cases see [Letac and Massam, 2008, Prop.4.5]. In order to write the variance function $V(m) = v(\psi(m))$ we compose the last expression from (32) and the formula (33).

For a model $\{W(p, a; \psi(\theta A + B)) : \theta \in I\}$ parametrized by a segment of means, the Fisher information $J(\theta)$ is obtained from the expression of $V(m)$ and Theorem 3.2.

Example. Suppose that $a = I_d$, $A = \alpha I_d$ and $B = \beta I_d$, $\alpha, \beta > 0$. The Fisher information on θ is $J(\theta) = \alpha^2 d\left((p^2 + 2\theta\alpha + 2\beta)(\theta\alpha + \beta + \frac{p^2}{4})^{1/2} - 2p(\theta\alpha + \beta) - \frac{p^3}{2}\right)^{-1}$.

4.3 Estimation of the mean in exponential families parametrized by a segment of means

Consider a sample $X_1, \ldots, X_n$ of a random variable X from a natural exponential family $Q(m, \mu)$ parametrized by the domain of means M, where the parameter $m = \mathbb{E}\, X$ is unknown and M is open. The following qualities of the sample mean $\bar{X}_n$ as an estimator of m seem to be known; for the sake of completeness we provide a short proof of properties which are less evident.

Proposition 4.1. *The sample mean $\bar{X}_n$ is an unbiased, consistent and efficient estimator of the parameter m. It is also a maximum likelihood estimator of m.*

Proof. By Theorem 3.1 we have $\operatorname{Cov} X = V(m) = J(m)^{-1}$, so the Cramér-Rao bound is attained by X. Consequently, the sample mean $\bar{X}_n$ is an efficient estimator of m. It follows from (9) that the sample mean $\bar{X}_n$ is a maximum likelihood estimator of m. One can also first show by (4) that the maximum likelihood estimator of s is $\hat{s} = k'^{-1}_\mu(X) = \psi(X)$ and next use the functional invariance of the maximum likelihood estimator [Casella and Berger, 2002, Theorem 7.2.10]. ☐

Remark 4.2. *For general exponential families $Q(m, T, \nu)$ parametrized by an open domain of means M, all these properties remain valid for $\hat{m} = \overline{T(X)}_n$ as an estimator of $m = \mathbb{E}\, T(X)$.*

Consider an exponential family $Q(\theta A + B, \mu)$ parametrized by a segment of means $IA + B \subset M$ with $A \neq 0, B \in E$ and $\theta \in I$, a segment in $\mathbb{R}$. We will now discuss estimators of the real parameter θ when we know that the mean $\mathbb{E}\, X = m \in IA + B$.

The segment $IA + B \subset M$ is of dimension one and has an empty interior. That's why the efficiency and maximum likelihood properties of the estimator $\hat{m} = \bar{X}_n$ are not automatically inherited by natural estimators of the real parameter θ. Determining a maximum likelihood estimator for θ seems impossible explicitly. This is the "price to pay" for the parsimony of the segment model parametrized by $m \in IA + B$. On the other hand, the efficiency of estimators of θ may be studied thanks to Theorem 3.2 and its corollaries.

Knowing that

$$m = \theta A + B \tag{34}$$

for a value $\theta \in I$, we have many possibilities of writing down a solution θ of equation (34). If $A \neq 0$ then the solution θ is unique ($A\theta + B = A\theta' + B$ implies $\theta = \theta'$ when $A \neq 0$.) For any C such that $\langle A\,,\,C\rangle \neq 0$ we have

$$\theta \;=\; \frac{\langle m - B\,,\,C\rangle}{\langle A\,,\,C\rangle}.$$

We define an estimator $\hat\theta_C$ of the parameter θ by

$$\hat\theta_C \;=\; \frac{\langle \bar X_n - B\,,\,C\rangle}{\langle A\,,\,C\rangle}$$

All the estimators $\hat\theta_C$ are unbiased and consistent. The natural question is whether they are efficient. The variance of $\hat\theta_C$ may be computed using the variance function $V(m)$ of the exponential family:

$$\operatorname{Var}\hat\theta_C = \frac{1}{\langle A, C\rangle^2}\operatorname{Var}\langle \bar X_n, C\rangle = \frac{\operatorname{Vec}(C)^T V(\theta A + B)\operatorname{Vec}(C)}{n\langle A, C\rangle^2}. \tag{35}$$

On the other hand, the Cramér-Rao bound is equal by Theorem 3.2 to

$$\frac{1}{nJ(\theta)} = \frac{1}{n\operatorname{Vec}(A)^T\, V(\theta A + B)^{-1}\, \operatorname{Vec}(A)}. \tag{36}$$

When the space E is a squared matrix space $\mathbb{R}^{d\times d}$ and the matrix A is invertible, we can take $C = A^{-1}$ and consider the estimator

$$\hat\theta_{A^{-1}} = \frac{\langle \bar X_n - B, A^{-1}\rangle}{d}.$$

The following theorem shows that for Gaussian and central Wishart exponential families and for linearly dependent A and B the estimator $\hat\theta_{A^{-1}}$ is efficient as an estimator of the mean m (with X_i replaced by $T(X_i) = -\frac{1}{2}(X_i - u)(X_i - u)^T$ in the Gaussian case). In conclusion, we obtain efficient estimators for Gaussian models parametrized by a covariance segment parameter and for Wishart models parametrized by a scale segment parameter.

Theorem 4.3. *1. Let $I \subset \mathbb{R}^+$ be a non-empty segment. Let $c \geq 0$, $A \in \mathcal{S}_d^+$ and $B = cA$.*
(1a) Consider an n-sample $(X_1, \ldots, X_n)$ from a Gaussian family $Q(m, T, \nu)$ defined by (19), where $m = \theta A + B$, $\theta \in I$. Then

$$\hat\theta_{A^{-1}} = \frac{\langle \overline{T(X)}_n - B, A^{-1}\rangle}{d}$$

is an unbiased efficient estimator of the parameter θ.
(1b) Consider an n-sample $(X_1, \ldots, X_n)$ from a Wishart model $Q(m, \mu_p)$ defined by (27), where $m = \theta A + B$, $\theta \in I$. Then

$$\hat\theta_{A^{-1}} = \frac{\langle \bar X_n - B, A^{-1}\rangle}{d}$$

is an unbiased efficient estimator of the parameter θ.
2. Let $c \geq 0$, $C \in \mathcal{S}_d^+$ *and* $D = cC$.
(2a) Consider an n*-sample* $(X_1, \ldots, X_n)$ *from a Gaussian model* $\{N(u, \theta C + D), \theta \in I\}$ *parametrized by a segment of covariances. An unbiased efficient estimator of* θ *is given by*

$$\hat{\theta} = \frac{1}{d}\langle\frac{1}{n}\sum_{i=1}^{n}(X_i - u)(X_i - u)^T - D, C^{-1}\rangle.$$

(2b) Consider an n*-sample* $(X_1, \ldots, X_n)$ *from a Wishart model* $\{\gamma(p, \theta C + D), \theta \in I\}$ *parametrized by a segment of scale parameters. Un unbiased efficient estimator of* θ *is given by*

$$\hat{\theta} = \frac{\langle\frac{1}{p}\bar{X}_n - D, C^{-1}\rangle}{d}.$$

Proof. For the first part of the Theorem, we give the proof in the Wishart case. The proof in the Gaussian case is identical, with $p = \frac{1}{2}$, cf. Remark 4.1. By formulas (35) and (28)

$$\mathrm{Var}\,\hat{\theta}_{A^{-1}} = \frac{1}{pd^2n}\,\mathrm{Tr}((A\theta + B)A^{-1}(A\theta + B)A^{-1}) = \frac{(\theta + c)^2}{pdn}$$

On the other hand, by (36) and (28)

$$\frac{1}{nJ(\theta)} = \frac{1}{np\,\mathrm{Tr}(A(A\theta + B)^{-1}A(A\theta + B)^{-1})} = \frac{1}{np(\theta + c)^{-2}d}.$$

Thus $\mathrm{Var}\,\hat{\theta} = \frac{1}{nJ(\theta)}$ and the estimator $\hat{\theta}_{A^{-1}}$ is efficient.

The second part of the Theorem follows by necessary reparametrizations. For (2a), using (19), we write $\theta C + D = -2m$ with $m = \theta A + B$, where $A = -\frac{C}{2}$ and $B = -\frac{D}{2}$. The part (2b) follows similarly from (27). □

Remark 4.3. *It is an open question whether* $\hat{\theta}_{A^{-1}}$ *may be efficient for independent* A *and* B. *Let* $n = 1$. *The equality* $\mathrm{Var}\,\hat{\theta} = \frac{1}{J(\theta)}$ *holds if and only if, writing* $D_\theta = (A\theta + B)A^{-1}(A\theta + B)A^{-1}$, *the equality* $\frac{1}{d^2}\,\mathrm{Tr}(D_\theta) = \frac{1}{\mathrm{Tr}(D_\theta^{-1})}$ *holds for all* $\theta \in I$.

Acknowledgements. The authors would like to thank Gérard Letac for his insightful suggestions. The second author acknowledges financial support(co-tutelle fellowship) from the French Ministry of Foreign Affairs and the Embassy of France in South Africa. The authors are greatly indebted to l'Agence Nationale de la Recherche for the research grant ANR-09-Blan-0084-01.

References

Andersson, T. and Handel, P. [2006]. Cramer-Rao bound and the parsimony principle. *Instrumentation and Measurement, IEEE Transactions*, **55** (1), 44–53.

Barndorff-Nielsen, O. E. [1978]. *Information And Exponential Families In Statistical Theory.* John Wiley and Sons.

Bercher, J.-F. and Vignat, C. [2009]. On minimum Fisher information distributions with restricted support and fixed variance. *Information Sciences*, **179**, 3832–3842.

Brown, L. D. [1986]. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, volume 9. IMS Lecture Notes-Monograph Ser.: Vol.9.

Casella, G. and Berger, R. [2002]. *Statistical Inference.* Duxbury Advances Series.

Frieden, B. R. [2004]. *Science from Fisher information: a unification.* Cambridge University Press.

Gray, R. M. [2006]. *Toeplitz and Circulant Matrices: A review.* Foundations and Trends in Communications and Information Theory, Vol. 2 Issue 3, NOW Publishers.

Gupta, R. D. and Kundu, D. [2006]. On the comparison of Fisher information of the Weibull and GE distributions. *Journal of Statistical Planning and Inference*, **136**, 3130–3144.

Letac, G. [1992]. *Lectures on natural exponential families and their variance functions.* CNDCT, Instituto de Matematica Pura e Aplicada, Rio de Janeiro.

Letac, G. and Casalis, M. [2000]. *Natural exponential families.* In: Kotz, S.; Balakrishnan, N.; Johnson, N. L. Continuous multivariate distributions. Vol. 1. Models and applications. Second edition. Wiley Series in Probability and Statistics.

Letac, G. and Massam, H. [2008]. The noncentral Wishart as an exponential family, and its moments. *Journal of Multivariate Analysis*, **99** (7).

Letac, G. and Massam, H. [2011]. Existence and non-existence of the noncentral Wishart distributions. *preprint.*

Magnus, J. R. and Neudecker, H. [2007]. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons.

Mason, J. C. and Handscomb, D. C. [2003]. *Chebyshev polynomials.* Chapman & Hall/CRC.

Muirhead, R. J. [2005]. *Aspects of multivariate statistical theory.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Ramirez, V. R., Hernandez, P. Q., Castro, S. H., and Diwekar, U. M. [2010]. Fisher information as a novel tool for process control applications. *20th European Symposium on Computer Aided Process Engineering - ESCAPE20 S. Pierucci and G. Buzzi Ferraris (Editors), 2010, Elsevier B.V.*