ARTICLE IN PRESS

Statistical Methodology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Jeffreys versus Shtarkov distributions associated with some natural exponential families

Shaul K. Bar-Lev [a,*], Daoud Bshouty [b], Peter Grünwald [c], Peter Harremoës [c]

[a] Department of Statistics, University of Haifa, Haifa 31905, Israel

[b] Department of Mathematics, Technion—Israel Institute of Technology, Haifa 32000, Israel

[c] Centrum Wiskunde & Informatica, Amsterdam, Netherlands

ARTICLE INFO

Article history:
Received 8 October 2009
Accepted 3 June 2010

Keywords:
Jeffreys prior
Natural exponential family
Regret
Shtarkov distribution
Variance function

ABSTRACT

Jeffreys and Shtarkov distributions play an important role in universal coding and minimum description length (MDL) inference, two central areas within the field of information theory. It was recently discovered that in some situations Shtarkov distributions exist while Jeffreys distributions do not. To demonstrate some of these situations we consider in this note the class of natural exponential families (NEF's) and present a general result which enables us to construct numerous classes of infinitely divisible NEF's for which Shtarkov distributions exist and Jeffreys distributions do not. The method used to obtain our general results is based on the variance functions of such NEF's. We first present two classes of parametric NEF's demonstrating our general results and then generalize them to obtain numerous multiparameter classes of the same type.

1. Introduction

Jeffreys and Shtarkov distributions play an important role in universal coding and *minimum description length* (MDL) inference, two central areas within the field of information theory [6,2]. Jeffreys and Shtarkov distributions are both defined relative to a given statistical model, i.e. a family of distributions $\mathcal{F}$. The Jeffreys distribution relative to a model $\mathcal{F}$ is just the distribution of data X obtained by equipping $\mathcal{F}$ with *Jeffreys prior*, which in turn is based on the *Jeffreys*

* Corresponding author. Tel.: +972 4 8240178.
E-mail addresses: barlev@stat.haifa.ac.il (S.K. Bar-Lev), daoud@techunix.technion.ac.il (D. Bshouty), Peter.Grunwald@cwi.nl (P. Grünwald), P.Harremoes@cwi.nl (P. Harremoës).

1572-3127/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2010.06.001

integral [9]. The Jeffreys distribution relative to $\mathcal{F}$ exists (i.e., is well-defined) if and only if the Jeffreys integral is finite. Jeffreys prior measure also plays an important role in Bayesian statistics. The Shtarkov distribution, also known as the *normalized maximum likelihood (NML)* distribution, has its roots in information theory [13,12,7]. It exists whenever a corresponding integral, the *Shtarkov integral*, is finite. We mention two important facts about Shtarkov and Jeffreys distributions: (a) For *natural exponential families* (NEF's) with canonical parameter space restricted to a compact subset of the interior of the full space, both integrals are finite and, for i.i.d. data $X_1, X_2, \ldots$, the Jeffreys and Shtarkov distributions converge asymptotically to one another in a very strong sense. (b) For all NEF's commonly encountered in the literature, the integrals are either both finite or both infinite.

Facts (a) and (b) have led some to conjecture that the Jeffreys distribution relative to $\mathcal{F}$ exists whenever the Shtarkov distribution relative to $\mathcal{F}$ exists [7]. However, [8] presents a single example of an NEF for which a Shtarkov distribution exists while a Jeffreys distribution does not. In this note, we formally demonstrate that there are, in fact, numerous such situations. We consider the class NEF's and present a general result which enables us to construct many classes of infinitely divisible NEF's for which Shtarkov distributions exist and Jeffreys distributions do not. The method used to obtain our general results is based on the variance functions of such NEF's. We first present two classes of parametric NEF's demonstrating our general results and then generalize them to obtain numerous multiparameter classes of the same type.

Background

Our result is of interest in two central areas within information theory: *universal coding*, also known as *universal data compression*, and the related field of *minimum description length* statistical inference [6,7]. The goal in universal coding is to achieve maximal lossless data compression of a data sequence $X = (X_1, \ldots, X_n)$ sampled from some unknown distribution P. P is usually called the "source". To explain the basic ideas, we assume in this section that X is discrete; but we emphasize that our results continue to be relevant in a continuous setting. Assume then that X has discrete support $\mathcal{X}$. The *Kraft inequality* [6] implies that for each probability mass function q on $\mathcal{X}$, there exists a lossless code that encodes each $x \in \mathcal{X}$ with length $L(x) = -\log_2 q(x)$, where, as is usual, we ignore rounding issues. The P-expected codelength of the code based on q is then given by $E_{X\sim P}[L(X)] = \sum_{x\in\mathcal{X}} -p(x)\log_2 q(x)$. By the information inequality [6], the minimum over q is achieved for the code with length $L(x) = -\log_2 p(x)$, resulting in a codelength which is equal to the entropy $H(P) = -\sum_x p(x)\log p(x)$. Thus, if the source P is known, in order to obtain the shortest mean codelength we should use a code with length $-\log p(x)$. In universal coding, one considers the case where P itself is unknown, but it is known that P belongs to some family of distributions $\mathcal{F}$. One is then interested in the code L which achieves the shortest lengths in a minimax redundancy or minimax regret sense; for precise definitions, we refer the reader to [7, Chapter 6]. If a code achieving minimax regret exists at all, it is given by the *Shtarkov distribution*; that is, it is the code with length $L(x) = -\log p_{\text{Shtarkov}}(x)$, where $p_{\text{Shtarkov}}(x)$ is the mass function corresponding to the Shtarkov distribution. In general, the Shtarkov distribution is hard to calculate. Instead, one may want to use the code based on the easier-to-calculate *Jeffreys distribution*, $L(x) = -\log p_{\text{Jeffreys}}(x)$. It can be shown [12,4,5] that, if parameters are restricted to a compact set, then asymptotically, the Jeffreys distribution assigns essentially the same codelength as the Shtarkov distribution, thus also achieving the minimax regret. The Jeffreys distribution is the distribution with mass function $p_{\text{Jeffreys}}(x) = \int p_\theta(x)w(\theta)\,\mathrm{d}\theta$, and $w(\theta)$ is the Jeffreys prior distribution [9]. For a one-parameter family, the Jeffreys prior is given by $w(\theta) = I(\theta)^{1/2}/J$ where $I(\theta)$ is the *Fisher information* at θ, and $J = \int I(\theta)^{1/2}\,\mathrm{d}\theta$ is called the Jeffreys integral. In a recent paper [8] we study under what conditions the minimax regret is finite. A natural question is then whether the Jeffreys distribution even exists if the Shtarkov distribution does. In this paper, we show that the answer is a definitive no.

In Section 2 we provide some basic facts regarding NEF's and we present our general result. In Section 3 we first present two classes of parametric NEF's demonstrating our general results and then generalize them to obtain numerous multiparameter classes of the same type.

Please cite this article in press as: S.K. Bar-Lev, et al., Jeffreys versus Shtarkov distributions associated with some natural exponential families, Statistical Methodology (2010), doi:10.1016/j.stamet.2010.06.001

2. The problem's formulation

Let ν be a nondegenerate positive Radon measure on $\mathbb{R}$ with *Laplace transform*

$$L(\theta) = \int_{\mathbb{R}} \mathrm{e}^{\theta x} \nu(\mathrm{d}x)$$

having an effective domain $D_\nu = \{\theta \in \mathbb{R} : L(\theta) < \infty\}$ such that $\Theta = int\, D_\nu \neq \varnothing$. Let S_ν and C_ν denote, respectively, the support and the convex support of ν. Then $\mathcal{F}$, the NEF generated by ν, is given by the probabilities

$$P(\theta, \nu(\mathrm{d}x)) = \exp\{\theta x - k(\theta)\}\, \nu(\mathrm{d}x), \quad \theta \in \Theta,$$

where $k(\theta) = \log L(\theta), \theta \in \Theta$, is the cumulant transform of ν. The *cumulant transform* k is strictly convex and real analytic on Θ and

$$\mu(\theta) = k'(\theta) = \int_{\mathbb{R}} x \exp\{\theta x - k(\theta)\}\, \nu(\mathrm{d}x)$$

is the mean function of $\mathcal{F}$. The open interval $M = k'(\Theta)$ is called the mean domain of $\mathcal{F}$. Since the map $\theta \longmapsto k'(\theta)$ is one-to-one, its inverse function $\theta : M \longrightarrow \Theta$ is well-defined. The variance function (VF) of $\mathcal{F}$ is the pair (V, M), where $V(\mu) = 1/\theta'(\mu), \mu \in M$. Such a VF defines the NEF $\mathcal{F}$ uniquely within the class of NEF's [11,10].

If the VF (V, M) is given then the canonical parameter θ and cumulant transform k of $\mathcal{F}$ are obtained by

$$\theta = \theta(\mu) = \int_{\mu_0}^{\mu} \frac{1}{V(t)}\, \mathrm{d}t \quad \text{and} \quad k(\mu) = \int_{\mu_0}^{\mu} \frac{t}{V(t)}\, \mathrm{d}t, \quad \text{for some } \mu_0 \in M. \tag{1}$$

Note that S_ν and C_ν are also the support and the convex support of any member of $\mathcal{F}$. Moreover, if ν is infinitely divisible (i.d.) then so are all members of $\mathcal{F}$, in which case we say that the respective VF is an i.d.

For an NEF, the Fisher information is given by $I(\mu) = V(\mu)^{-1}$. We pose the following question which relates to the difference between Jeffreys and Shtarkov distributions: Do there exist NEF's whose VF's satisfy the following condition: For arbitrary $s \in M$,

$$\int_s^{\infty} \frac{\mu}{V(\mu)}\, \mathrm{d}\mu < \infty \quad \text{whereas} \quad \int_s^{\infty} \frac{1}{V(\mu)^{1/2}}\, \mathrm{d}\mu = \infty? \tag{2}$$

We note that the leftmost integral is just the Jeffreys integral for the restricted family with a mean-value parameter ranging from s to ∞. Thus the rightmost term of (2) expresses that the Jeffreys prior (a distribution on M) does not exist, and hence neither does the Jeffreys distribution (a distribution on X). On the other hand, by a change of variable from μ to θ, it is seen that the leftmost integral in (2) is equal to $\int_{\theta_0}^{\theta_{\sup}} \mu(\theta)\, \mathrm{d}\theta$ for some $\theta_0 \in D_\nu$ and $\theta_{\sup} = \sup D_\nu$. Since $\mu(\theta) = k'(\theta)$, this integral is finite iff $k(\theta_{\sup}) < \infty$. In [8] it is shown that if $k(\theta_{\sup}) < \infty$, then the Shtarkov distribution exists. Thus, (2) expresses the situation that the Jeffreys distribution does not exist, whereas the Shtarkov distribution does.

Before presenting our results, we give some further intuition about (2). It is easily seen that, in order to satisfy (2), an NEF must satisfy a very specific (necessary but not sufficient) condition: If the limits $\lim_{\mu\to\infty} V(\mu)/\mu^2$ and, for $\varepsilon > 0$, $\lim_{\mu\to\infty} V(\mu)/\mu^{2+\varepsilon}$ exists at all, then they must satisfy

$$\lim_{\mu\to\infty} \frac{V(\mu)}{\mu^2} = \infty, \quad \text{yet for all } \varepsilon > 0, \quad \lim_{\mu\to\infty} \frac{V(\mu)}{\mu^{2+\varepsilon}} = 0. \tag{3}$$

To see this, note that if for some $\varepsilon > 0$, $\lim_{\mu\to\infty} V(\mu)/\mu^{2+\varepsilon} > 0$, then $V(\mu)^{-1/2} = O(\mu^{-1-\frac{\varepsilon}{2}})$, so the rightmost integral in (2) is finite, contradicting (2). Also, if $\lim_{\mu\to\infty} V(\mu)/\mu^2 < \infty$, then $\mu^2/V(\mu) > c;\ c > 0$, for large μ, and hence $\mu/V(\mu) > c\mu^{-1}$, so the leftmost integral in (2) is infinite, contradicting (2).

Please cite this article in press as: S.K. Bar-Lev, et al., Jeffreys versus Shtarkov distributions associated with some natural exponential families, Statistical Methodology (2010), doi:10.1016/j.stamet.2010.06.001

Condition (3) is not satisfied by any of the usual families encountered in the literature, and initially we wondered whether such families exist at all. Theorem 3 gives a positive answer to this question. Before introducing Theorem 3, we need the two following results which are taken from Bshouty [3] and Bar-Lev et al. [1], respectively.

Lemma 1 (*[3, Theorem 4 for a Special Case]*). *Let $\mu(\theta)$ be an absolutely monotone function on $(-\infty, r)$; then $V(\mu) = 1/\theta'(\mu)$ is a VF of an i.d. NEF.*

We denote by $\mathcal{A}$ the class of i.d. VF's (V, M) with $M = (0, \alpha)$ and $0 < \alpha \leq \infty$, which satisfy

$$\int_0^{\mu_1} \frac{1}{V(t)}\, \mathrm{d}t = \infty, \tag{4}$$

for some $\mu_1 \in M$.

Lemma 2 (*[1, Theorem 1]*). *For $i = 1, 2$, let $M_i = (0, \alpha_i)$ and $(V_i, M_i) \in \mathcal{A}$. Then $(V_1 V_2, M_1 \cap M_2) \in \mathcal{A}$.*

We are now ready to present our main theorem.

Theorem 3. *Let $\mathcal{F}$ be an i.d. NEF supported on $\mathbb{R}^+$ with VF $(V, M) \in \mathcal{A}$, where $M = \mathbb{R}^+$. Assume that (4) holds and that*

$$\lim_{\mu\to\infty} \frac{V(\mu)}{\mu} = \beta > 0. \tag{5}$$

Then $\mu \to (\mu+1)^2 V^2(\log(\mu+1))$ is also a VF of an i.d. NEF with mean parameter space $M = \mathbb{R}^+$ that satisfies condition (2).

Proof. We first show that $V_1(\mu) = (\mu+1)^2 V^2(\log(\mu+1))$ is also a VF of an i.d. NEF with mean parameter space $M = \mathbb{R}^+$. Indeed, since $\mathcal{F}$ is supported on $\mathbb{R}^+$, its mean function $\mu(\theta)$ is absolutely monotone on $\mathbb{R}^-$. Since once $\mu(\theta)$ is given its respective VF is well-defined, we have

$$V(\mu) = \frac{1}{\theta'(\mu)}.$$

Next, e^x is an absolutely monotone function on $\mathbb{R}$ and therefore so is the composition $\mu_0(\theta) \doteq \exp(\mu(\theta))$ on $\mathbb{R}^-$. Let θ_0 denote the inverse function of μ_0 and M_0 be the range of $\mu_0(\theta_0)$. Hence, by Lemma 1, $(V_0(\mu_0), M_0) \doteq (1/\theta_0'(\mu_0), M_0)$ is a VF of an i.d. NEF. To obtain the general form of $V_0(\mu)$ we use the fact that the inverse of $\mu_0(\theta)$ is $\theta_0(\mu) = \theta(\log\mu)$ implying that

$$V_0(\mu) = \frac{1}{\theta_0'(\mu)} = \frac{\mu}{\theta'(\log\mu)} = \mu V(\log\mu).$$

Using the translation $\mu \longmapsto \mu + 1$, we obtain that $(V_{01}, M_{01}) = ((\mu+1)V(\log(\mu+1)), M_{01})$ is an i.d. VF as well. Since M_{01} is the largest strip where V_{01} admits an analytic continuation or is zero and since by (5) $V(0) = 0$, we conclude that $M_{01} = \mathbb{R}^+$, implying that $(V_{01}, \mathbb{R}^+) \in \mathcal{A}$. Consequently, Lemma 2 yields that

$$(V_1, M_1) \doteq ((\mu+1)^2 V^2(\log(\mu+1)), \mathbb{R}^+) \in \mathcal{A}.$$

Next, we show that (2) holds. Since

$$\lim_{\mu\to\infty} \frac{V(\mu)^{-1/2}}{1/(\mu+1)\log(\mu+1)} = \beta^{\frac{1}{2}},$$

the two integrals $\int_s^\infty V(\mu)^{-1/2}\, \mathrm{d}\mu$ and $\int_s^\infty \frac{1}{(\mu+1)\log(\mu+1)}\, \mathrm{d}\mu$ (for arbitrary $s \in \mathbb{R}^+$) converge or diverge simultaneously. Since

$$\int_s^\infty \frac{1}{(\mu+1)\log(\mu+1)}\, \mathrm{d}\mu = \log(\log(\mu+1))\Big|_{\mu=s}^{\infty} = \infty,$$

we conclude that $\int_s^\infty V(\mu)^{-1/2}\, \mathrm{d}\mu$ diverges.

Please cite this article in press as: S.K. Bar-Lev, et al., Jeffreys versus Shtarkov distributions associated with some natural exponential families, Statistical Methodology (2010), doi:10.1016/j.stamet.2010.06.001

Similarly,

$$\lim_{\mu\to\infty}\frac{\mu/V(\mu)}{1/(\mu+1)\log^2(\mu+1)}=\beta$$

so $\int_s^\infty \mu/V(\mu)\,\mathrm{d}\mu$ and $\int_s^\infty \frac{1}{(\mu+1)\log^2(\mu+1)}\,\mathrm{d}\mu$ converge or diverge simultaneously. Since

$$\int_s^\infty \frac{1}{(\mu+1)\log^2(\mu+1)}\,\mathrm{d}\mu=\frac{-1}{\log(\mu+1)}\Big|_{\mu=s}^{\infty}<\infty,$$

it follows that V_1 satisfies condition (2) and the result follows. □

Remark 4. Assuming that $(V,\mathbb{R}^+)\in\mathcal{A}$, then it is easy to show that also $(V_1,\mathbb{R}^+)\doteq(aV,\mathbb{R}^+)\in\mathcal{A}$ for any positive a. Consequently, the class $(aV,\mathbb{R}^+)$ generates a two-parameter class of i.d. VF's (or, equivalently, of i.d. NEF's), with parameters $a>0$ and $\mu\in M$, that satisfy (2).

3. Constructions of parametric classes of VF's satisfying condition (2)

In this section we apply Theorem 3 to construct three-parameter classes of i.d. VF's that satisfy condition (2). We then generalize the two classes in Theorem 5 to a huge class containing these two classes as special cases.

Class 1:

The pair $(V,M)=(\mu,\mathbb{R}^+)$ is the VF of the Poisson family which belongs to $\mathcal{A}$ and corresponds to the mean function $\mu(\theta)=\mathrm{e}^\theta$. The choice $\mu_1(\theta)=\mathrm{e}^{b\theta}$, $b>0$, gives rise to $(V_1,M_1)=(b\mu,\mathbb{R}^+)$. The procedure stated in Theorem 3 applies here and yields the VF $(V_1,M_1)=(b^2(\mu+1)^2\log^2(\frac{\mu}{b}+1),\mathbb{R}^+)$. Hence, by Remark 4,

$$\mathcal{A}_1=\left\{(V_2,M_2)=\left(a(\mu+1)^2\log^2\left(\frac{\mu}{b}+1\right),\mathbb{R}^+\right),\ a>0,b>0,\mu>0\right\}\tag{6}$$

is a three-parameter class of i.d. VF's which satisfies condition (2).

Class 2:

Let $\mu(\theta)=a\exp\left\{\exp(\mathrm{e}^{b\theta})\right\}$, $a>0$, $b>0$, which is evidently an absolutely monotone function on $\mathbb{R}^-$. Then

$$V(\mu)=\frac{1}{\theta'(\mu)}=b\mu\,(\log\mu/a)\,(\log\log(\mu/a)),\quad \mu>ae,$$

is an i.d. VF. By using the translation $\mu\longmapsto\mu+ae$, we obtain an i.d. VF

$$\left(V_1,\mathbb{R}^+\right)=\left(b(\mu+ae)\log((\mu+ae)/a)\log\log((\mu+ae)/a),\mathbb{R}^+\right),\tag{7}$$

which vanishes at zero. Then, by squaring V_1 in (7) and employing Lemma 2, we get that

$$\mathcal{A}_2=\left\{(V_2,M_2)=\left([b(\mu+ae)\log((\mu+ae)/a)\log\log((\mu+ae)/a)]^2,\mathbb{R}^+\right),\ a>0,b>0,\mu>0\right\}$$

is a three-parameter class of i.d. VF's which can be easily verified to satisfy condition (2).

The methods used to construct classes 1 an 2 can be simply extended to establish a general form of classes of the same type. Indeed, let $(a_i)_{i=1}^\infty$ be an arbitrary sequence of positive real numbers and define

$$\mathrm{e}^{[a_1]}(\theta)\doteq a_1\mathrm{e}^\theta\quad\text{and}\quad \mathrm{e}^{[a_1,a_2,\ldots,a_n]}(\theta)\doteq a_n\mathrm{e}^{\mathrm{e}^{[a_1,a_2,\ldots,a_{n-1}](\theta)}},\quad n=2,3,\ldots,\tag{8}$$

i.e.,

$$\mathrm{e}^{[a_1,a_2]}(\theta)=a_2\mathrm{e}^{[a_1]}(\theta)=a_2\mathrm{e}^{a_1\mathrm{e}^\theta},\qquad \mathrm{e}^{[a_1,a_2,a_3]}(\theta)=a_3\mathrm{e}^{[a_1,a_2]}(\theta)=a_3\mathrm{e}^{a_2\mathrm{e}^{a_1\theta}}$$

and so on. Consequently, we have the following theorem.

Please cite this article in press as: S.K. Bar-Lev, et al., Jeffreys versus Shtarkov distributions associated with some natural exponential families, Statistical Methodology (2010), doi:10.1016/j.stamet.2010.06.001

Theorem 5. *Let* $\mu_n(\theta) \doteq e^{[a_1,a_2,\ldots,a_n]}(\theta)$ *be defined as in* (8)*; then by appropriate translations,* $\mu_n(\theta)$, $n \in \mathbb{N}$, *generates VF's* $\left(V_n, \mathbb{R}^+\right)$ *of i.d. NEF's that satisfy* (2)*.*

Proof. Clearly, the $\mu_n(\theta)$'s are absolutely monotone functions on $\mathbb{R}^-$. Then, in a method analogous to the steps conducted for class 2 (while using an appropriate translation and then by employing Lemma 2), one obtains that

$$\mathcal{B}_n = \left\{\left(V_n, \mathbb{R}^+\right), a_i > 0, i = 1, \ldots, n\right\}, \quad n \in \mathbb{N}$$

are such classes. We omit the details for the sake of brevity. □

References

[1] S.K Bar-Lev, D. Bshouty, P. Enis, A.Y. Ohayon, Compositions and products of infinitely divisible variance functions, Scandinavian Journal of Statistics 19 (1992) 83–89.
[2] A. Barron, J. Rissanen, B. Yu, The minimum description length principle in coding and modeling, IEEE Transactions on Information Theory 44 (6) (1998) 2743–2760. (Special commemorative issue: Information Theory, 1948–1998.)
[3] D. Bshouty, On a characterization of variance functions of natural exponential families, Mathematical Methods of Statistics 4 (1995) 92–98.
[4] B. Clarke, A. Barron, Information-theoretic asymptotics of Bayes methods, IEEE Transactions on Information Theory IT-36 (3) (1990) 453–471.
[5] B. Clarke, A. Barron, Jeffreys' prior is asymptotically least favorable under entropy risk, Journal of Statistical Planning and Inference 41 (1994) 37–60.
[6] T. Cover, J.A. Thomas, Elements of Information Theory, Wiley, 1991.
[7] P. Grünwald, The Minimum Description Length principle, MIT Press, 2007.
[8] P. Grünwald, P. Harremoës, Finiteness of redundancy, regret, Shtarkov Sums and Jeffreys integrals in exponential families, in: Proceedings of the International Symposium for Information Theory, ISIT 2009, IEEE, 2009, pp. 714–718.
[9] H. Jeffreys, An invariant form for the prior probability in estimation problems, Proceedings of the Royal Statistical Society (London) Series A 186 (1946) 453–461.
[10] G. Letac, M. Mora, Natural real exponential families with cubic variance functions, The Annals of Statistics 18 (1990) 1–37.
[11] C.N. Morris, Natural exponential families with quadratic variance functions, The Annals of Statistics 10 (1982) 65–80.
[12] J. Rissanen, Fisher information and stochastic complexity, IEEE Transactions on Information Theory 42 (1) (1996) 40–47.
[13] Y.M. Shtarkov, Universal sequential coding of single messages, Problems of Information Transmission 23 (3) (1987) 3–17.

Please cite this article in press as: S.K. Bar-Lev, et al., Jeffreys versus Shtarkov distributions associated with some natural exponential families, Statistical Methodology (2010), doi:10.1016/j.stamet.2010.06.001