Statistical Inference and Exact Saddle Point Approximations

Peter Harremoës
GSK Department
Copenhagen Business College
Copenhagen, Denmark
Email: harremoes@ieee.org

***Abstract*—Statistical inference may follow a frequentist approach or it may follow a Bayesian approach or it may use the minimum description length principle (MDL). Our goal is to identify situations in which these different approaches to statistical inference coincide. It is proved that for exponential families MDL and Bayesian inference coincide if and only if the renormalized saddle point approximation for the conjugated exponential family is exact. For 1-dimensional exponential families the only families with exact renormalized saddle point approximations are the Gaussian location family, the Gamma family and the inverse Gaussian family. They are conjugated families of the Gaussian location family, the Gamma family and the Poisson-exponential family. The first two families are self-conjugated implying that only for the two first families the Bayesian approach is consistent with the frequentist approach. In higher dimensions there are more examples.**

I. Introduction

In this paper we are interested in predition of the future given the past. We assume that a sequence $x_1^m = x_1, x_2, \ldots, x_m$ has been observed and the goal is to predict the next symbols $x_{m+1}^n = x_{m+1}, x_{m+2}, \ldots, x_n$ in the sense that we will assign a probability or a probability density to this sequence. The prediction is compared with iid models given by a parametrized family $(P_\theta)_{\theta\in\Theta}$ of probability distributions that assign probability $\prod_{i=1}^n P_\theta(x_i)$ (or the corresponding density) to the sequence x_1^n. One may think of the elements of the family $(P_\theta)_{\theta\in\Theta}$ as the models that some experts can choose among. For the techniques used in this paper the restriction to iid models is crusial, but some of the results may generalize to non-iid models.

All measures will be described by their density with respect to a dominating measure λ. Data are assumed to lie in $\mathcal{X} \subseteq \mathbb{R}^d$ and vectors will be marked with bold face. Assume that $(P_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$ is a natural exponential family with

$$\frac{\mathrm{d}P_{\boldsymbol{\theta}}}{\mathrm{d}\lambda}(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}\cdot\boldsymbol{x})}{Z(\boldsymbol{\theta})} = \exp(\boldsymbol{\theta}\cdot\boldsymbol{x} - A(\boldsymbol{\theta})).$$

Here $Z(\boldsymbol{\theta}) = \int \exp(\boldsymbol{\theta}\cdot\boldsymbol{x})\,\mathrm{d}\lambda\boldsymbol{x}$ is the *moment generating function* and $A(\boldsymbol{\theta}) = \ln(Z(\boldsymbol{\theta}))$ is the *cumulant generating function*. If the parameter has value $\boldsymbol{\theta}$ then the mean value is $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \nabla A(\boldsymbol{\theta})$. The density $\frac{\mathrm{d}P_{\boldsymbol{\theta}}}{\mathrm{d}\lambda}$ will be denoted $p_{\boldsymbol{\theta}}$, but sometimes we will also use $p_{\boldsymbol{\theta}}$ for iid sequences.

One approach is the frequentist approach where the sequence $\boldsymbol{x}_1^n$ is generated by the distribution $P_{\boldsymbol{\theta}}$ for some true but unknown value of $\boldsymbol{\theta}$. The sequence $\boldsymbol{x}_1^m$ is used to make inference about the value of $\boldsymbol{\theta}$ in terms of a confidence region. In a Bayesian approach one has a prior distribution π on the true parameter $\boldsymbol{\theta}$ and the sequence $\boldsymbol{x}_1^m$ is used to calculate a posterior distribution of $\boldsymbol{\theta}$ as

$$\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^m)\,\pi(\boldsymbol{\theta})}{\int_\Theta p_{\boldsymbol{\theta}}(\boldsymbol{x}^m)\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}.$$

Then the posterior distribution of $\boldsymbol{x}_{m+1}^n$ is given by

$$\begin{aligned} p_\pi\left(\boldsymbol{x}_{m+1}^n \mid \boldsymbol{x}^m\right) &= \int_\Theta p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{m+1}^n\right)\,\mathrm{d}\pi(\boldsymbol{\theta}\mid\boldsymbol{x}^m) \qquad (1)\\ &= \int_\Theta p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{m+1}^n\right)\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^m)\,\pi(\boldsymbol{\theta})}{\int_\Theta p_{\boldsymbol{\theta}}(\boldsymbol{x}^m)\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}\,\mathrm{d}\boldsymbol{\theta}\end{aligned}$$

One of the main problems is Bayesian statistics is the question of how to determine the prior distribution π.

The moment generating function Z is related to the Laplace transform of the measure λ, so any of the functions Z and A can be used to reconstruct λ. The *Hesse matrix* of A with respect to $\boldsymbol{\theta}$ equals the *co-variance matrix* $Cov(\boldsymbol{\mu}_{\boldsymbol{\theta}})$. The Fisher information matrix with respect to the natural parameter is $Cov(\boldsymbol{\mu}_{\boldsymbol{\theta}})$ so that *Jeffreys' prior* is proportional to $|Cov(\boldsymbol{\mu}_{\boldsymbol{\theta}})|^{1/2}$. Therefore *Jeffreys' posterior* distribution of the parameter $\boldsymbol{\theta}$ after observing a sequence of length m with average $\bar{\boldsymbol{x}}$ is proportional to

$$\exp(m\cdot(\boldsymbol{\theta}\cdot\bar{\boldsymbol{x}} - A(\boldsymbol{\theta})))\cdot|Cov(\boldsymbol{\mu}_{\boldsymbol{\theta}})|^{1/2}.$$

One motivation for using Jeffreys' prior is that it is considered as an uninformative prior. Another motivation is that if one restricts to a bounded subset whose closure is in the interior of the full parameter space, then the use of Jeffrey's prior is asymptotically optimal in a MDL sense [1].

A co-variance matrix is positive semi-definite so the cumulant generating function is convex. The *convex conjugate* of the cumulant generating function A is $A^*(\boldsymbol{x}) = \sup_{\boldsymbol{\theta}}\{\boldsymbol{\theta}\cdot\boldsymbol{x} - A(\boldsymbol{\theta})\}$. The conjugate parameter $\boldsymbol{x}^*$ equals the value of $\boldsymbol{\theta}$ such that $P_{\boldsymbol{\theta}}$ has mean value $\boldsymbol{x}$, i.e. $\boldsymbol{x}^*$ is the solution to the equation $\nabla A(\boldsymbol{\theta}) = \boldsymbol{x}$. Usually the conjugate parameter $\boldsymbol{x}^*$ is denoted $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ and is called the maximum likelihood estimate of $\boldsymbol{\theta}$. We can define the *conjugated exponential family* (if it exists) as the exponential family with sufficient statistic $\boldsymbol{\theta}$ and with cumulant generating function $A^*(\boldsymbol{x})$.

Remark 1. For an exponential family the conjugated exponential family gives a set of "conjugated priors" as this concept is defined in the literature on Bayesian statistics (see [2] and [3, Sec. 12.2.6]), but a set of "conjugated priors" need not coincide with the conjugated exponential family as it is defined in this paper.

The Bregman divergence generated by the convex function A is defined by

$$D_A(\boldsymbol{\theta}_2,\boldsymbol{\theta}_1)=A(\boldsymbol{\theta}_2)-(A(\boldsymbol{\theta}_1)+(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1)\cdot\nabla A(\boldsymbol{\theta}_1))$$

Using convex conjugation the divergence can also be written as

$$D_A(\boldsymbol{\theta}_2,\boldsymbol{\theta}_1)=D_{A^*}(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2).$$

The information divergence can be calculated as

$$\begin{aligned}D(P_{\boldsymbol{\theta}_1}\|P_{\boldsymbol{\theta}_2})&=E_{\boldsymbol{\theta}_1}\left[\ln\left(\frac{\mathrm{d}P_{\boldsymbol{\theta}_1}}{\mathrm{d}P_{\boldsymbol{\theta}_2}}\right)\right]\\&=E_{\boldsymbol{\theta}_1}[(\boldsymbol{\theta}_1\cdot\boldsymbol{X}-A(\boldsymbol{\theta}_1))-(\boldsymbol{\theta}_2\cdot\boldsymbol{X}-A(\boldsymbol{\theta}_2))]\\&=D_A(\boldsymbol{\theta}_2,\boldsymbol{\theta}_1).\end{aligned}$$

The conjugated exponential family gives posterior distributions on the parameter $\boldsymbol{\theta}$, such that the maximum likelihood estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ is unbiased in the sense that it equals the mean value of $\boldsymbol{\theta}$ with respect to the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{x}$. Therefore the use of the conjugated exponential family implies that the maximum likelihood estimator equals the Bayes estimator with respect to the loss function D_A or any other Bregman divergence.

The likelihood function can be written as

$$\begin{aligned}&p_{\boldsymbol{\theta}}(\boldsymbol{x})=\exp(\boldsymbol{\theta}\cdot\boldsymbol{x}-A(\boldsymbol{\theta}))=\\&\exp\left(-A(\boldsymbol{\theta})+\left(A\left(\hat{\boldsymbol{\theta}}(\boldsymbol{x})\right)+\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\boldsymbol{x})\right)\cdot\nabla A\left(\hat{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right)\right)\\&\qquad\cdot p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x})}(\boldsymbol{x})\\&=\exp\left(-D_A\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right)\cdot p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x})}(\boldsymbol{x}).\end{aligned}$$

As a consequence we have the following robustness property [1, Section 19.3, Eq. 19.12] of the exponential family

$$\frac{\mathrm{d}P_{\boldsymbol{\theta}}}{\mathrm{d}P_{\hat{\boldsymbol{\theta}}(\boldsymbol{x})}}(\boldsymbol{x})=\exp\left(-D_A\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right).\tag{2}$$

The likelihood function after observing the sequence $\boldsymbol{x}^m$ is

$$\begin{aligned}\prod_{i=1}^m p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)&=\prod_{i=1}^m\exp(\boldsymbol{\theta}\cdot\boldsymbol{x}_i-A(\boldsymbol{\theta}))\\&=\exp\left(\boldsymbol{\theta}\cdot\sum_{i=1}^m\boldsymbol{x}_i-m\cdot A(\boldsymbol{\theta})\right)\\&=\exp(m\cdot(\boldsymbol{\theta}\cdot\bar{\boldsymbol{x}}-A(\boldsymbol{\theta})))\\&=\exp\left(-m\cdot D_A\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\bar{\boldsymbol{x}})\right)\right)\\&\quad\cdot\exp\left(m\left(\hat{\boldsymbol{\theta}}(\bar{\boldsymbol{x}})\cdot\bar{\boldsymbol{x}}-A\left(\hat{\boldsymbol{\theta}}(\bar{\boldsymbol{x}})\right)\right)\right).\end{aligned}$$

In the minimum description length (MDL) approach to statistical inference there is no assumption about a true value of $\boldsymbol{\theta}$, and the quality of a prediction is compared with the maximum likelihood estimate of $\boldsymbol{\theta}$ in terms of a difference in code length. For a data sequence $\boldsymbol{x}^n$ the *regret* of predicting $p\left(\boldsymbol{x}_{m+1}^n\mid\boldsymbol{x}^m\right)$ is

$$-\ln\left(p\left(\boldsymbol{x}_{m+1}^n\mid\boldsymbol{x}^m\right)\right)-\left(-\ln\left(p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)}(\boldsymbol{x}^n)\right)\right).$$

Here the predictor $p(\cdot\mid\boldsymbol{x}^m)$ is used to code the future $\boldsymbol{x}_{m+1}^n$ while the expert is coding the whole sequence $\boldsymbol{x}^n$, but the expert is allowed to choose the model $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)$ that gives the best fit to data. We take the maximum over all possible data sequences and the predictor that minimizes the maximal regret is called the *conditional normalized maximum likelihood* predictor (CNML) [4] and is given by

$$\begin{aligned}&p_{cnml}^n\left(\boldsymbol{x}_{m+1}^n\mid\boldsymbol{x}^m\right)\\&=\frac{p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)}(\boldsymbol{x}^n)}{\int_{\mathcal{X}^{n-m}}p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^m\boldsymbol{y}^{n-m})}(\boldsymbol{x}^m\boldsymbol{y}^{n-m})\,\mathrm{d}\lambda^{n-m}(\boldsymbol{y}^{n-m})}.\end{aligned}\tag{3}$$

II. Main results

The essence of the following lemma was already present in [5, Lem. 3].

Lemma 2. *Assume that $(P_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$ is a natural exponential family. Assume that m is a number such that CNML and Bayesian prediction based on a prior π give equal prediction strategies for sequences $\boldsymbol{x}_{m+1}^n$ for all $n>m$. Then for any $n>m$ the integral*

$$\int_\Theta\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^n)}{p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)}(\boldsymbol{x}^n)}\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

is constant as a function of the data sequence $\boldsymbol{x}^n=\boldsymbol{x}_1\boldsymbol{x}_2\ldots\boldsymbol{x}_n$,

Remark 3. Prediction with CNML and prediction based of Jeffreys prior can only be equal if they are both defined. The values of m for which these prediction methods are defined, may in principle be different and may depend on the data sequence [6].

Proof: For all $\boldsymbol{x}^{\boldsymbol{n}}\in\mathcal{X}^n$ we must have

$$p_\pi\left(\boldsymbol{x}_{m+1}^n\mid\boldsymbol{x}_m\right)=p_{cnml}^n\left(\boldsymbol{x}_{m+1}^n\mid\boldsymbol{x}_m\right).$$

Using (1) and (3) we get

$$\begin{aligned}&\int_\Theta p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{m+1}^n\right)\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^m)\pi(\boldsymbol{\theta})}{\int_\Theta p_{\boldsymbol{\theta}}(\boldsymbol{x}^m)\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}\mathrm{d}\boldsymbol{\theta}\\&=\frac{p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)}(\boldsymbol{x}^n)}{\int_{\mathcal{X}^{n-m}}p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^m\boldsymbol{y}^{n-m})}(\boldsymbol{x}^m\boldsymbol{y}^{n-m})\,\mathrm{d}\lambda^{n-m}(\boldsymbol{y}^{n-m})}\end{aligned}$$

and

$$\begin{aligned}&\frac{\int_\Theta p_{\boldsymbol{\theta}}(\boldsymbol{x}^n)\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}{p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)}(\boldsymbol{x}^n)}\\&=\frac{\int_\Theta p_{\boldsymbol{\theta}}(\boldsymbol{x}^m)\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}{\int_{\mathcal{X}^{n-m}}p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^m\boldsymbol{y}^{n-m})}(\boldsymbol{x}^m\boldsymbol{y}^{n-m})\,\mathrm{d}\lambda^{n-m}(\boldsymbol{y}^{n-m})}.\end{aligned}$$

The quantity on the left side is a function of $\boldsymbol{x}^n$ while the quantity on the right side is a function of the sub-string

$\boldsymbol{x}^m$. Since the model is invariant under permutations of the elements in the string $\boldsymbol{x}^n$ both sides must equal a constant. Finally we note that

$$\frac{\int_\Theta p_{\boldsymbol{\theta}}(\boldsymbol{x}^n)\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}{p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)}(\boldsymbol{x}^n)} = \int_\Theta \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^n)}{p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)}(\boldsymbol{x}^n)}\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,,$$

which proves the lemma. ■

Note that we have not really used that the parametrized family is an exponential family, so a similar result holds as long as the parametrization is sufficiently smooth. If the parametrization is sufficiently smooth one can also prove that the prior must be proportional to Jeffrey's prior. We conjecture that if conditional MDL is a Bayesian prediction for some smoothly parametrized family where the parameter space is finitely dimensional, then the family must be exponential. Recall that the saddle point approximation [7] for the exponential family is

$$\exp\left(-nD_A\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)\right)\right)\frac{|Cov(\boldsymbol{\mu}_{\boldsymbol{\theta}})|^{1/2}}{\tau^{d/2}}\,,$$

where τ is short for 2π.

Theorem 4. *Assume that $(P_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta}$ is a natural exponential family. Then the following conditions are equivalent:*

- *CNML is a Bayesian prediction strategy.*
- *Jeffreys' posterior distributions are elements of the conjugated exponential family.*
- *The renormalized saddle-point approximation is exact for the conjugated exponential family.*

Proof: According to expression (2) we may define a constant C_n by

$$C_n = \int_\Theta \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^n)}{p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)}(\boldsymbol{x}^n)}\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}.$$

Then

$$\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^n)}{p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)}(\boldsymbol{x}^n)}\cdot\frac{\pi(\boldsymbol{\theta})}{C_n} \tag{4}$$

is a probability density function for $\boldsymbol{\theta}$. We will demonstrate that the family of probability measures (4) parametrized by $\boldsymbol{x}^n$ is the conjugated exponential family with $\boldsymbol{\theta}$ as sufficient statistic. We have

$$\begin{aligned}&\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}^n)}{p_{\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)}(\boldsymbol{x}^n)}\cdot\frac{\pi(\boldsymbol{\theta})}{C_n}\\ &\quad= \frac{\exp(n(\boldsymbol{\theta}\cdot\bar{\boldsymbol{x}} - A(\boldsymbol{\theta})))}{\exp\left(n\left(\hat{\theta}(\boldsymbol{x}^n)\cdot\bar{\boldsymbol{x}} - A\left(\hat{\theta}(\boldsymbol{x}^n)\right)\right)\right)}\cdot\frac{\pi(\boldsymbol{\theta})}{C_n}\\ &\quad= \exp(n(\boldsymbol{\theta}\cdot\bar{\boldsymbol{x}} - A^*(\bar{\boldsymbol{x}})))\cdot\frac{\pi(\boldsymbol{\theta})}{\exp(nA(\boldsymbol{\theta}))\,C_n}.\end{aligned}$$

According to the robustness property (2) the density can be rewritten as

$$\exp\left(-nD_A\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\boldsymbol{x}^n)\right)\right)\cdot\frac{\pi(\boldsymbol{\theta})}{C_n}.$$

Since this should hold for n tending to infinity the saddle point approximation implies that $\pi(\boldsymbol{\theta})$ is proportional to $|Cov(\boldsymbol{\mu}_{\boldsymbol{\theta}})|^{1/2}$. Therefore the density in the exponential family is proportional to the saddle point approximation. ■

Corollary 5. *If any of the equivalent conditions of Theorem 4 are fulfilled the exponential family is steep and the parameter space is maximal.*

The goal is now to identify exponential families where Jeffreys' posterior distributions form exponential families with exact renormalized saddle point approximations. In [8] it was proved that under certain regularity conditions the renormalized saddle point approximation is exact for *reproductive exponential families*. The reproductive exponential families were defined and described in detail in [9] where it was proved in 1 dimension the following families were reproductive: the Gaussian location families, the Gamma exponential families and the Inverse Gaussian families. The idea of reproductive exponential families can be used to construct reproductive exponential families in higher dimension by combining reproductive exponential families in lower dimensions. Five non-trivial examples of 2-dimensional (strongly) reproducible exponential families obtained by combining reproductive 1 dimensional families were listed in [9]. For each reproductive exponential family the conjugate exponential family (if it exists) will satisfy the conditions of Theorem 4. We will illustrate how this works for 1-dimensional reproductive exponential families.

The only 1-dimensional natural exponential families where the renormalized saddle point approximation is exact, are the three reproductive exponential families mentioned above [10], and it can be proved by solving ordinary differential equations [8]. A complete classification of exponential families with exact renormalized saddle point approximation in dimension 2 or higher would require solving some complicated partial differential equations. Therefore a complete catalog of families for which the equivalent conditions of Theorem 4 are fulfilled, seems inaccessable.

For the 1-dimensional reproductive exponential families the functions A^* is exactly the ones used in [9] to prove that the exponential family is reproductive. Exploration of this fact in higher dimensions will be covered in a future paper.

III. The Gamma family

A Gamma distribution can be parametrized by the shape parameter α and the rate parameter β. With these parameters the Gamma distribution $\Gamma(\alpha,\beta)$ has density

$$\frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)}\exp(-\beta x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)}\exp(-\beta x + \alpha\ln(\beta))$$

for $x > 0$. For a fixed value of α this is a natural exponential family with natural parameter $\theta = -\beta < 0$. Therefore $A(\theta) = -\alpha\ln(-\theta)$. The mean value is $\mu = -\alpha/\theta$ so that $\theta = -\alpha/\mu$. The variance is $Var = \alpha\theta^{-2}$, so that the variance function is $V(\mu) = \frac{\mu^2}{\alpha}$. In terms of the parameter β the mean value is $\mu = \alpha/\beta$ and the variance is $Var = \alpha\cdot\beta^{-2}$. Jeffreys' prior has density proportional to $\frac{\alpha^{1/2}}{\beta}$, which cannot be normalized.

The Bregman divergence is

$$D_A(\theta_1,\theta_2) = \alpha\ln\left(-\frac{1}{\theta_1}\right) - \left(\alpha\ln\left(-\frac{1}{\theta_2}\right) + (\theta_1-\theta_2)\cdot\frac{-\alpha}{\theta_2}\right) = \alpha\left(\frac{\theta_1}{\theta_2} - 1 - \ln\left(\frac{\theta_1}{\theta_2}\right)\right).$$

For $\alpha = 1$ this Bregman divergence is called the *Itakura-Saito divergence*.

The convex conjugate of A is

$$A^*(x) = \sup_\theta \{x\cdot\theta - A(\theta)\} = x\cdot\left(-\frac{\alpha}{x}\right) - A\left(-\frac{\alpha}{x}\right) = -\alpha + \alpha\ln\left(\frac{\alpha}{x}\right) = -\alpha + \alpha\ln(\alpha) - \alpha\ln(x).$$

We see that the conjugated exponential family of $\beta = -\theta$ is again a Gamma exponential family with shape parameter α, i.e. the Gamma exponential family is *self-conjugated*. If x is observed the posterior distribution of β has rate parameter x. If a sequence of length m has been observed then the posterior distribution is a Gamma distribution with shape parameters $m\alpha$ and rate parameter $m\bar{x}$.

Since the density of a Gamma distribution equals the re-normalized saddle point approximation we have that the conditions in Theorem 4 are fulfilled and the CNML predictor equals Bayesian prediction based on Jeffreys' prior. This also holds for exponential families like the inverse Gamma family, the Pareto family, the Nakagima family, and the Weibull family where the sufficient statistic is a smooth 1-to-1 function of the sufficient statistic in a Gamma family.

We will now look at the consequences of self-conjugation for calculations of one-sided credible intervals and one-sided confidence intervals.

Let G denote the distribution function of $\Gamma(m\alpha, m\bar{x})$, i.e. the posterior distribution of β if the average is observed to be $\bar{x}$. Then $\left[0, G^{-1}(1-\tilde{\alpha})\right]$ is a $1-\tilde{\alpha}$ *credible interval* for β. We can write

$$G^{-1}(1-\tilde{\alpha}) = \frac{F^{-1}(1-\tilde{\alpha})}{\bar{x}}$$

where F is the distribution function of $\Gamma(m\alpha, m)$. If $X_i \sim \Gamma(\alpha,\beta)$ then $\sum_{i=1}^m X_i \sim \Gamma(m\alpha,\beta)$ and $\frac{1}{m}\sum_{i=1}^m X_i \sim \Gamma(m\alpha, m\beta)$ so that $\beta\bar{X} \sim \Gamma(m\alpha, m)$. Therefore

$$P\left(\beta \in \left[0, \frac{F^{-1}(1-\tilde{\alpha})}{\bar{X}}\right]\right) = P\left(\bar{X} \in \left[0, \frac{F^{-1}(1-\tilde{\alpha})}{\beta}\right]\right) = 1-\tilde{\alpha}$$

so that the $1-\tilde{\alpha}$ credible interval $\left[0, \frac{F^{-1}(1-\tilde{\alpha})}{\bar{x}}\right]$ is also a $1-\tilde{\alpha}$ *confidence interval* for β as defined in the frequentist approach to statistics.

IV. The Gaussian location family

If the parameter space equals $\mathbb{R}^d$ the notion of self-conjugation becomes very simple. The proof of the following lemma is an easy exercise.

Lemma 6. *Let $B : \mathbb{R}^d \to \mathbb{R}^d$ denote a linear invertible self-adjoint mapping. If G is a convex function and $F = G \circ B$ then $F^* = G^* \circ B^{-1}$.*

The Gaussian location model has density

$$\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\cdot B^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\tau^{d/2}\cdot|B|^{1/2}}$$

where $\boldsymbol{\mu}$ is the mean and B denotes the co-variance matrix.

Theorem 7. *If an exponential family has a cumulant generating function $A : R^d \to R$ that satisfies $A^* = A \circ B$ for some positiv definite linear function $B : \mathbb{R}^d \to \mathbb{R}^d$ then the exponential family is a Gaussian location model where B can be identified with the co-variance matrix.*

Proof: Define $F = A \circ B^{1/2}$. Then

$$F^* = A^* \circ \left(B^{1/2}\right)^{-1} = A \circ B \circ B^{-1/2} = A \circ B^{1/2} = F.$$

Since F is self-conjugated and defined on $\mathbb{R}^d$ we can apply [11, Prop. 29a] to get $F(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$. Therefore

$$A(\boldsymbol{x}) = F\left(B^{-1/2}(\boldsymbol{x})\right) = \frac{1}{2}B^{-1/2}(\boldsymbol{x})\cdot B^{-1/2}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}\cdot B^{-1}(\boldsymbol{x}).$$

It is easy to prove that the Gaussian location model also has cumulant generating function $\frac{1}{2}\boldsymbol{x}\cdot B^{-1}(\boldsymbol{x})$. ■

Since the saddle point approximation is exact for the Gaussian location family the conditions of Threorem 4 are fulfilled. For the Gaussian location family the Bregman divergence is symmetric in its arguments and inference reduces to the principle of least squares.

In Bayesian statistics a $1-\tilde{\alpha}$ *credible region* for the mean value parameter can be calculated as a divergence ball

$$\left\{\theta \in R^d \mid D_A\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\boldsymbol{x})\right) \leq r\right\} \tag{5}$$

where the radius r is chosen so that the ball has probability $1-\tilde{\alpha}$. Using that the exponential family is self-conjugated we see that the ball (5) is also a $1-\tilde{\alpha}$ *confidence region* as defined in frequentist statistics.

V. The Poisson-exponential family

The saddle point approximation is exact for the inverse Gaussian family with density

$$\left(\frac{\kappa}{\tau\beta^3}\right)^{1/2}\exp\left(-\kappa\frac{(\beta-\beta_0)^2}{2\beta_0^2\beta}\right),$$

where β is the sufficient statistic and β_0 denotes the mean value of the distribution and κ denotes the *shape parameter*. We are going to identify the conjugated exponential family. First we rewrite

$$\left(\frac{\kappa}{\tau\beta^3}\right)^{1/2}\exp\left(-\kappa\frac{(\beta-\beta_0)^2}{2\beta_0^2\beta}\right) = \left(\frac{\kappa}{\tau\beta^3}\right)^{1/2}\exp\left(-\frac{\kappa}{2\beta}\right)\exp\left(\frac{-\kappa}{2\beta_0^2}\cdot\beta + \frac{\kappa}{\beta_0}\right).$$

The natural parameter is $\theta = \frac{-\kappa}{2\beta_0^2}$ and the cumulant generating function is $A(\theta) = (-2\kappa\theta)^{1/2}$.

The convex conjugate is

$$A^*(\beta) = \sup\{\beta\cdot\theta - A(\theta)\} = \beta\cdot\frac{-\kappa}{2\beta^2} - \left(-2\kappa\cdot\frac{-\kappa}{2\beta^2}\right)^{1/2} = \frac{\kappa}{2\beta}.$$

One can identify an exponential family with this function as cumulant generating function by taking the inverse Laplace transform, but it is more instructive to identify it by calculating the variance function. We have

$$(A^*)'(\beta) = -\frac{\kappa}{2\beta^2} \text{ and } (A^*)''(\beta) = \frac{\kappa}{\beta^3}.$$

Thus $\hat{\theta}(\beta) = -\frac{\kappa}{2}\beta^{-2}$ so that $\beta(\theta) = \left(-\frac{\kappa}{2\theta}\right)^{1/2}$ and $V(\theta) = \kappa(\beta(\theta))^{-3} = 2^{3/2}\kappa^{-1/2}(-\theta)^{3/2} = \phi\cdot(-\theta)^{3/2}$ where $\phi = 2^{3/2}\kappa^{-1/2}$. Since the variance function is a power function of order $3/2$ one says that the corresponding exponential family is a *Tweedie family* of order $p = 3/2$. Jeffreys' prior for this family is proportional to

$$\left((A^*)''(\beta)\right)^{1/2} = \kappa^{1/2}\cdot\beta^{-3/2},$$

which cannot be normalized. Credible intervals and confidence intervals can be calculated using `tweedie` and the `statmod` package in the R program, but the $1-\tilde{\alpha}$ credible intervals do not coincide with the $1-\tilde{\alpha}$ confidence intervals reflecting that the Poisson-exponential family is not self-conjugated.

One cannot calculate the density of elements of the Tweedie family of order $p = 3/2$ exactly, but they can be obtained by the following construction. Let N denote a random variable with a Poisson distribution $Po(\lambda)$. Let $X_1, X_2, \ldots$ denote a sequence of iid random variables each exponentially distributed $Exp(\beta)$. Then we may define

$$Y = \sum_{n=1}^{N} X_n\,.$$

Then the distribution of Y is a compound Poisson distribution. Distributions where X_i are Gamma distributions were called Poisson-gamma distributions in [12], so we will call the distribution of Y a *Poisson-exponential distribution* when X_i are exponential. The density of $\sum_{n=1}^{\alpha} X_n$ is

$$\frac{\tilde{\beta}^{\alpha}x^{\alpha-1}\exp\left(-\tilde{\beta}x\right)}{\Gamma(\alpha)}.$$

Therefore the Poisson-exponential distribution has a point mass in 0 of weight $\exp(-\lambda)$ and it has density

$$\sum_{\alpha=0}^{\infty}\frac{\lambda^{\alpha}\exp(-\lambda)}{\alpha!}\cdot\frac{\beta^{\alpha}x^{\alpha-1}\exp(-\beta x)}{\Gamma(\alpha)}$$

for $x > 0$. We introduce $\kappa = \frac{\tilde{\beta}\cdot\lambda}{2}$ so that the density can be written as

$$\sum_{\alpha=0}^{\infty}\frac{\left(\frac{\kappa}{2}\right)^{\alpha}x^{\alpha-1}}{\alpha!\Gamma(\alpha)}\cdot\exp\left(-\beta\cdot x - \frac{\kappa}{2\beta}\right).$$

This is a natural exponential family with with natural parameter $-\beta$ and cumulant generating function $\kappa/(2\beta)$. Except for a change of sign it is the conjugated exponential family of the inverse Gaussian family.

Since the saddle point approximation is exact for the inverse Gaussian family, prediction for the Poisson-exponential family based on CNML equals prediction based on Jeffreys prior, and Jeffreys posterior equals an inverse Gaussian distribution.

The Poisson-exponential families have been used to model the accumulated amount of rain in rainfalls, where the amount of rain in each rainfall is modeled by an exponential distribution and the number of rainfalls is modeled by a Poisson distribution [13], [14]. This application dates back to Cornish and Fisher. Reference to other applications as well as a derivation of the basic properties of Poisson-gamma distributions can be found in [15]. Note that the Poisson-exponential family is a Tweedie family of order $p = 3/2$ and that some of the literature on applications of the Poisson-exponential family treat the order p as a free parameter that should be estimated in order to give a good fit with data. According to our results the value $p = 3/2$ is special with respect to statistical inference, so that p cannot be considered as a free parameter if we want to have the properties developed here.

ACKNOWLEDGEMENT

I would like to thank Wojciech Kotłowski for useful comments to this paper.

REFERENCES

[1] P. Grünwald, *The Minimum Description Length principle*. MIT Press, 2007.

[2] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory, Division of Research*. Grad. School of Busimess Adm. Harvard, 1961.

[3] H. Liu and L. Wasserman, *Statistical Machine Learning*. 2014.

[4] J. Rissanen and T. Roos, “Conditional NML universal models,” in *Information Theory and Applications Workshop (ITA-07)*, pp. 37–341, 2007.

[5] P. Bartlett, P. Grünwald, P. Harremoës, F. Hedayati, and W. Kotłowski, “Horizon-independent optimal prediction with log-loss in exponential families,” in *Conference on Learning Theory (COLT 2013)*, p. 23, 12-14 June 2013.

[6] P. Harremoës, “Extendable MDL,” in *International Symposium on Information Theory*, (Boston), pp. 1516–1520, IEEE, July 2013.

[7] H. E. Daniels, “Saddlepoint approximations in statistics,” *Ann. Math. Statist.*, vol. 25, no. 4, pp. 631–650, 1954.

[8] P. Blæsild and J. L. Jensen, “Saddlepoint formulas for reproductive exponential models,” *Scand. J. Statist.*, 1985.

[9] O. Barndorff-Nielsen and P. Blæsild, “Reproductive exponential families,” *The Annals of Statistics*, vol. 11, no. 3, pp. 770–782, 1983.

[10] H. E. Daniels, “Exact saddlepoint approximations,” *Biometrika*, vol. 67, no. 1, pp. 59–63, 1980.

[11] J. J. Moreau, “Proximité et dualité dans un espace hilbertien,” *Bulletin de la S. M. F.*, vol. 93, pp. 273–299, 1965.

[12] G. K. Smyth, “Regression analysis of quantity data with exact zeros,” in *Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 572–580, 1996.

[13] C. S. Thompson, “Homogeneity analysis of rainfall series: An application of the use of a realistic rainfall model,” *Journal of Climatology*, vol. 4, no. 6, pp. 609–619, 1984.

[14] K. J. A. Revfeim, “An initial model of the relationship between rainfall events and daily rainfalls,” *Journal of Hydrology*, vol. 75, pp. 357–364, Dec. 1984.

[15] C. S. Withers and S. Nadarajah, “On the compound Poisson-gamma distribution,” *Kybernetika*, vol. 47, no. 1, pp. 15–37, 2011.