Journal de la Société Française de Statistique
Vol. 160 No. 2 (2019)

Conditional inference in parametric models

Titre: Inférence conditionnelle dans les modèles paramétriques

Michel Broniatowski[1] and Virgile Caron[1]

Résumé : Cet article propose une nouvelle approche d'inférence statistique, fondée sur la simulation d'échantillons conditionnés par une statistique des données. L'approximation de la vraisemblance conditionnelle de longues séries d'échantillons sachant la statistique des données admet une forme explicite qui est présentée. Lorsque la statistique de conditionnement est exhaustive par rapport à un paramètre fixé, on montre que la densité approchée est également invariante par rapport à ce même paramètre. Une nouvelle procédure de Rao-Blackwell est proposée et les simulations réalisées montrent que le théorème de Lehmann Scheffé reste valide pour cette approximation. L'inférence conditionnelle sur les familles exponentielles avec paramètre de nuisance est également étudiée, menant à des tests de Monte Carlo, dont les performances sur échantillonnage conditionnel sont comparées à celles sur bootstrap paramétrique. Enfin, on s'intéresse à l'estimation du paramètre d'intérêt par la vraisemblance conditionnelle.

Abstract: This paper presents a new approach to conditional inference, based on the simulation of samples conditioned by a statistics of the data. Also an explicit expression for the approximation of the conditional likelihood of long runs of the sample given the observed statistics is provided. It is shown that when the conditioning statistics is sufficient for a given parameter, the approximating density is still invariant with respect to the parameter. A new Rao-Blackwellisation procedure is proposed and simulation shows that Lehmann Scheffé Theorem is valid for this approximation. Conditional inference for exponential families with nuisance parameter is also studied, leading to Monte Carlo tests; comparison with the parametric bootstrap method is discussed. Finally the estimation of the parameter of interest through conditional likelihood is considered.

Mots-clés : Inférence conditionnelle, Théorème de Rao Blackwell, Théorème de Lehmann Scheffé, Familles exponentielles, Paramètre de nuisance, Simulation
Keywords: Conditional inference, Rao Blackwell Theorem, Lehmann Scheffé Theorem, Exponential families, Nuisance parameter, Simulation
AMS 2000 subject classifications: 62B05, 62F03, 62F10

1. Introduction and context

This paper explores conditional inference in parametric models. A comprehensive overview on this area is the illuminating review paper by Reid (1995) [29]. Our starting point is as follows: given a model $\mathscr{P}$ defined as a collection of continuous distributions P_θ on $\mathbb{R}^d$ with density p_θ with respect to the Lebesgue measure where the parameter θ belongs to some subset Θ in $\mathbb{R}^s$ and given a sample of independent copies of a random variable with distribution P_{θ_T} for some unknown value θ_T of the parameter, we intend to provide some inference about θ_T conditioning on some observed statistics of the data. The situations which we have in mind are of two different kinds.

The first one is the Rao-Blackwellisation of estimators, which amounts to reduce the variance of an unbiased estimator by conditioning on any statistics; when the conditioning statistics is

[1] LPSM, Sorbonne-Université, Paris, France.

http://www.sfds.asso.fr/journal

complete and sufficient for the parameter then this procedure provides optimal reduction, as stated by Lehmann-Scheffé Theorem. These facts yield the following questions.

1. is it possible to provide good approximations for the density of a sample conditioned on a given statistics, and, when applied for a model where some sufficient statistics for the parameter is known, does sufficiency w.r.t. the parameter still holds for the approximating density?
2. in the case when the first question has positive answer, is it possible to simulate samples according to the approximating density, and to propose some Rao-Blackwellised version for a given preliminary estimator? Also we would hope that the proposed method would be feasible, that the programming burden would be light, that the run time for this method be short, and that the involved techniques would keep in the range of globally known ones by the community of statisticians.

The second application of conditional inference pertains to the role of conditioning in models with nuisance parameters. There is a huge bibliography on this topic, some of which will be considered in details in the sequel. The usual frame for this field of problems is the exponential families one, for reasons related both with the importance of these models in applications and on the role of the concept of sufficiency when dealing with the notion of nuisance parameter. Conditioning on a sufficient statistics for the nuisance parameter produces a new exponential family, which gets free of this parameter, and allows for simple inference on the parameter of interest, at least in simple cases. This will also be discussed, since the reality, as known, is not that simple, and since so many complementary approaches have been developed over decades in this area. Using the approximation of the conditional density in this context and performing simulations yields Monte Carlo tests for the parameter of interest, free from the nuisance parameter; comparison with the parametric bootstrap will also be discussed. Also conditional maximum likelihood estimators will be produced. The present paper relies on an approximation result for conditional distributions developed in [4], where some of the present statistical applications are merely sketched.

This paper is organized as follows. Section 2 describes a general approximation scheme for the conditional density of long runs of subsamples conditioned on a statistics, with explicit formulas. The proof of the main result of this section is presented in [4]. Discussion about implementation is provided. Section 3 presents two aspects of the approximating conditional scheme: we first show on examples that sufficiency is kept under the approximating scheme and, second, that this yields to an easy Rao-Blackwellisation procedure. An illustration of Lehmann-Scheffé Theorem is presented. Section 4 deals with models with nuisance parameters in the context of exponential families. We have found it useful to spend a few paragraphs on bibliographic issues. We address Monte Carlo tests based on the simulation scheme; in simple cases its performance is similar to that of parametric bootstrap; however conditional simulation based tests improve clearly over parametric bootstrap procedure when the test pertains to models for which the likelihood is multimodal with respect to the nuisance parameter; an example is provided. Finally we consider conditioned maximum likelihood based on the approximation of the conditional density; in simple cases its performance is similar to that of estimators defined through global likelihood optimization; however when the preliminary estimator of the nuisance is difficult to obtain, for example when it depends strongly on some initial point for a Newton-Raphson routine (this is indeed a very common situation), then, by the very nature of sufficiency, conditional inference based on the proxy of the conditional likelihood performs better; this is illustrated with examples.

2. The approximate conditional density of the sample

Most attempts which have been proposed for the approximation of conditional densities stem from arguments developed in [18] for inference on the parameter of interest in models with nuisance parameter; however the proposals in this direction hinge at the approximation of the distribution of the sufficient statistics for the parameter of interest given the observed value of the sufficient statistics of the nuisance parameter. We will present some of these proposals in the section devoted to exponential families. To our knowledge, no attempt has been made to approximate the conditional distribution of a sample (or of a long subsample) given some observed statistics.

However, generating samples from the conditional distribution itself (such samples are often called co-sufficient samples, following [22]) has been considered by many authors; see for example [14], [19] and references therein, and [20].

In [14], simulating exponential or normal samples under the given value of the empirical mean is proposed. For example under the exponential distribution $Exp(\theta)$, the minimal sufficient statistics for θ is the sum of the observations, say t_n; a co-sufficient sample x^* can be created by generating an $x^{'}$-sample from $Exp(1)$ and taking $x_i^* = x_i^{'} t_n / \overline{x'}$. However, this approach may be at odd in simple cases, as for the Gamma density in the non exponential case.

Lockhart et al. [22] proposed a different framework based on the Gibbs sampler, simulating the conditioned sample one at a time through a sequential procedure. The example which is presented is for the Gamma distribution under the empirical mean; in these examples it seems to perform well for location parameter, when the true parameter is in some range, therefore not uniformly on the model. Their paper contains a comparative study with the global maximum likelihood method. In a simple case, they argue favorably for both methods. We will turn back to global likelihood maximization in relation with conditional likelihood estimators, in the last section of this paper.

Other techniques have been developed in specific cases: for the inverse Gaussian distribution see [24], [8]; for the Weibull distribution see [23]. No unified technique exists in the literature which would work under general models.

2.1. Approximation of conditional densities

2.1.1. Notation and hypotheses

For sake of clearness we consider the case when the model $\mathscr{P}$ is a family of distributions on $\mathbb{R}$.

Denote $\mathbf{X}_1^n := (\mathbf{X}_1, .., \mathbf{X}_n)$ a set of n independent copies of a real random variable $\mathbf{X}$ with density $p_{\mathbf{X},\theta_T}$ on $\mathbb{R}$. Let $\mathbf{x}_1^n := (\mathbf{x}_1, ..., \mathbf{x}_n)$ denote the observed values of the data, each $\mathbf{x}_i$ resulting from the sampling of $\mathbf{X}_i$. Define the r.v. $\mathbf{U} := u(\mathbf{X})$ and $\mathbf{U}_{1,n} := u(\mathbf{X}_1) + ... + u(\mathbf{X}_n)$ where u is a real-valued measurable function on $\mathbb{R}$, and, accordingly, $u_{1,n} := u(\mathbf{x}_1) + ... + u(\mathbf{x}_n)$. Denote $p_{\mathbf{U},\theta_T}$ the density of the r.v. $\mathbf{U}$. We consider approximations of the density of the vector $\mathbf{X}_1^k = (\mathbf{X}_1, .., \mathbf{X}_k)$ on $\mathbb{R}^k$ when $\mathbf{U}_{1,n} = u_{1,n}$. It will be assumed that the observed value $u_{1,n}$ is "typical", in the sense that it satisfies the law of large numbers. Since the approximation scheme for the conditional density is validated through limit arguments, it will be assumed that the sequence $u_{1,n}$ satisfies

$$\lim_{n\to\infty} \frac{u_{1,n}}{n} = Eu(\mathbf{X}). \tag{1}$$

We propose an approximation for

$$p_{u_{1,n},\theta_T}\left(x_1^k\right) := p_{\theta_T}\left(x_1^k|\mathbf{U}_{1,n} = u_{1,n}\right)$$

where $x_1^k := (x_1,..,x_k)$ and $k := k_n$ is an integer sequence such that

$$\lim_{n\to\infty} n-k = \infty \tag{K}$$

which is to say that we may approximate $p_{u_{1,n},\theta_T}\left(x_1^k\right)$ on long runs. The rule which defines the value of k for a given accuracy of the approximation is stated in section 3.2 of [4]. Note that (K) is a very weak assumption in the context of approximation of conditional distributions; indeed it implies

$$0 \leq \limsup_{n\to\infty} k/n \leq 1.$$

Cases when $\limsup_{n\to\infty} k/n < 1$ have been considered in the literature (see e.g. [10] and [9]) but do not address approximations on long runs, whose application to statistics is the focus of the present paper.
The hypotheses pertaining to the function u and the r.v. $\mathbf{U} = u(\mathbf{X})$ are as follows.

1. u is real valued and the characteristic function of the random variable $\mathbf{U}$ is assumed to belong to $L^r(\lambda)$ where λ denotes the Lebesgue measure on $\mathbb{R}$ for some $r \geq 1$.
2. The r.v. $\mathbf{U}$ is supposed to fulfill the Cramer condition: the domain $\mathcal{N}$ of the moment generating function
$$\phi_{\mathbf{U}}(t) := E\exp t\mathbf{U}$$
contains a non void neighborhood of 0.

Define the functions $m(t), s^2(t)$ and $\mu_3(t)$ as the first, second and third derivatives of $\log \phi_{\mathbf{U}}(t)$.

Let α belong to the support of $P_{\mathbf{U},\theta_T}$, the distribution of $\mathbf{U}$. Assume that the mapping $t \to \phi_{\mathbf{U}}(t)$ is steep (see [1], p153 and followings). Under steepness the mapping m is a diffeomorphism from $\mathcal{N}$ onto the support of $\mathbf{U}$. It follows that the correspondence (α, t) defined through $m(t) = \alpha$ is one to one. For such a couple (α, t) denote

$$\pi_{u,\theta_T}^{\alpha}(x) := \frac{\exp tu(x)}{\phi_{\mathbf{U}}(t)} p_{\mathbf{X},\theta_T}(x).$$

We introduce a positive sequence ε_n which satisfies

$$\lim_{n\to\infty} \varepsilon_n\sqrt{n-k} = \infty \tag{E1}$$

$$\lim_{n\to\infty} \varepsilon_n (\log n)^2 = 0. \tag{E2}$$

2.2. *The proxy of the conditional density of the sample*

The density $g_{u_{1,n},\theta_T}(x_1^k)$ on $\mathbb{R}^k$, which approximates $p_{u_{1,n},\theta_T}\left(x_1^k\right)$ sharply with relative error smaller than $\varepsilon_n\,(\log n)^2$ is defined recursively as follows.

Set

$$m_0 := u_{1,n}/n$$

and

$$g_0(x_1) := \pi_{u,\theta_T}^{m_0}(x_1)$$

and for $1 \leq i \leq k-1$ define the density $g(x_{i+1}|x_1^i)$ in the following way.

Set t_i the unique solution of the equation

$$m_i := m(t_i) = \frac{u_{1,n} - u_{1,i}}{n-i} \tag{2}$$

where $u_{1,i} := u(x_1) + ... + u(x_i)$.

The tilted adaptive family of densities $\pi_{u,\theta_T}^{m_i}$ is the basic ingredient of the derivation of approximating scheme. Let us briefly recall two main properties of tilted distributions in order to motivate the notation, which may seem cumbersome. Firstly conditional distributions with respect to sums are invariant under any tilting, whenever defined. For any sequence of iid rv's $Z_1, Z_2, ..$ with common density p_Z wrt the Lebesgue measure,

$$p_{Z_1}(z|Z_1 + .. + Z_n = s) = \pi_Z^t(z|Z_1 + .. + Z_n = s).$$

Hence sampling under p_Z or under any π_Z^t leaves the conditional marginal distribution invariant, whatever t, where $\pi_Z^t(z) := e^{tz}\ p_Z(z)/\int e^{tz}\ p_Z(z)dz$. Secondly convolutions of tilted densities can be approximated sharply by Gaussian distributions through Edgeworth expansions, which involve moments of higher orders. In the current approximation, conditioning upon $u_{1,i}$, in order to get a proxy of the density of X_{i+1} given $u(X_1) + .. + u(X_i) = u_{1,i}$ we are led to introduce the tilted density

$$\pi_{u,\theta_T}^{m_i}(x) := \frac{e^{t_i u(x)}}{\phi_{\mathbf{U}}(t_i)} p_{X,\theta_T}(x)$$

with t_i defined in (2). The moment generating function of this tilted distribution is

$$\begin{aligned} \phi_{\mathbf{U},i}(t) \quad &: \quad = E_{\pi_{u,\theta_T}^{m_i}} \mathbf{e}^{tu(\mathbf{X})} \\ &= \quad \frac{\phi_{\mathbf{U}}(t_i + t)}{\phi_{\mathbf{U}}(t_i)} \end{aligned}$$

and its first three cumulants are

$$m_i = m(t_i) = \frac{d}{dt} \log \phi_{\mathbf{U},i}(0),$$

$$s_i^2 := \frac{d^2}{dt^2} \log \phi_{\mathbf{U},i}(0) = s^2(t_i)$$

and

$$\mu_3^i := \frac{d^3}{dt^3}\log\phi_{\mathbf{U},i}(0) = \mu_3(t_i).$$

. Let

$$g(x_{i+1}|x_1^i) = C_i p_{\mathbf{X},\theta_T}(x_{i+1})\mathfrak{n}\left(\alpha\beta + m_0, \beta, u\left(x_{i+1}\right)\right) \tag{3}$$

where $\mathfrak{n}\left(\mu,\tau,x\right)$ is the normal density with mean μ and variance τ at x. Here

$$\beta = s_i^2\left(n-i-1\right) \tag{4}$$

$$\alpha = t_i + \frac{\mu_3^i}{2s_i^4\left(n-i-1\right)} \tag{5}$$

and the C_i is a normalizing constant.

Define

$$g_{u_{1,n},\theta_T}(x_1^k) := g_0(x_1|x_0)\prod_{i=1}^{k-1} g(x_{i+1}|x_1^i). \tag{6}$$

It holds

Theorem 1. *Assume (K) together with (E1,E2). Then (i)*

$$p_{u_{1,n},\theta_T}(x_1^k) = g_{u_{1,n},\theta_T}(x_1^k)(1 + o_{P_{u_{1,n},\theta_T}}(\varepsilon_n\left(\log n\right)^2))$$

and (ii)

$$p_{u_{1,n},\theta_T}(x_1^k) = g_{u_{1,n},\theta_T}(x_1^k)(1 + o_{G_{u_{1,n},\theta_T}}(\varepsilon_n\left(\log n\right)^2)).$$

(iii) The total variation distance between $P_{u_{1,n},\theta_T}$ and $G_{u_{1,n},\theta_T}$ goes to 0 as n tends to infinity.

For the proof, see [4].

Statement (i) means that the conditional likelihood of any long sample path $\mathbf{X}_1^k$ given $(\mathbf{U}_{1,n} = u_{1,n})$ can be approximated by $G_{u_{1,n},\theta_T}(\mathbf{X}_1^k)$ with a small relative error on typical realizations of $\mathbf{X}_1^n$.

The second statement implies that typical samples $\mathbf{X}_1^k$ simulated under $g_{u_{1,n},\theta_T}$ are also typical under the conditional density $p_{u_{1,n},\theta_T}$.

2.3. Comments on implementation

The simulation of a sample X_1^k with density $g_{u_{1,n},\theta_T}$ is fast as easy. Indeed the r.v. X_{i+1} with density $g\left(x_{i+1}|x_1^i\right)$ is obtained through a standard acceptance -rejection algorithm. When $\mathbf{U}_{1,n}$ is sufficient for $p_{u_{1,n},\theta}$ it is nearly sufficient for its proxy $g_{u_{1,n},\theta}$ (see next section); indeed changing the value of this preliminary estimator does not alter the value of the likelihood of the sample; as shown in the simulations developed here after, any value of θ can be used; call θ^* the value of θ chosen as initial value , using henceforth $p_{\mathbf{X},\theta^*}$ instead of $p_{\mathbf{X},\theta_T}$ in (3). In exponential families the values of the parameters which appear in the Gaussian component of $g\left(x_{i+1}|x_1^i\right)$ in (3) are easily calculated; note also that due to (1) the parameters in $\mathfrak{n}\left(\alpha\beta,\beta,u\left(x_{i+1}\right)\right)$ are such that the dominating density can be chosen for all i as $p_{\mathbf{X},\theta^*}$. The constant in the acceptance rejection algorithm is then $C_i/\sqrt{2\pi\beta}$. The constant C_i need not be evaluated since it cancels in the ratio defining the acceptance-rejection rule.

In order to simulate X_{i+1} with density $g(x_{i+1}|x_1^i)$, the acceptance/rejection algorithm thus runs as follows: the proposal density is $p_{\mathbf{X},\theta^*}$.

Set $k=1$. Simulate Z_k with density $p_{\mathbf{X},\theta^*}$. Simulate U uniform$(0,1)$ independent on Z_j, $1 \leq j \leq k$

If

$$U/\sqrt{2\pi\beta} \leq \mathfrak{n}(\alpha\beta + m_0, \beta, u(Z_k)) \tag{7}$$

then set $X_{i+1} := Z_k$. Else increase k by 1 and repeat.

When (1) holds, β does not tend to 0 as n increases, and the runtime is short; indeed under (1) $\sup_{1\leq i \leq n} t_i$ does not go to infinity as $n \to \infty$ which implies that (7) is fulfilled for small k. .

This is in contrast with the case when the conditioning value is in the range of a large deviation with respect to $p_{\mathbf{X},\theta_T}$; in this case, which appears in a natural way in Importance sampling estimation for rare event probabilities, the simulation algorithm is more complex ; see [5].

3. Sufficient statistics and approximated conditional density

3.1. Keeping sufficiency under the proxy density

The density $g_{u_{1,n},\theta_T}(y_1^k)$ is used in order to handle Rao -Blackellisation of estimators or statistical inference for models with nuisance parameters. The basic property is sufficiency with respect to the nuisance parameter. We show on some examples that the family of densities $g_{u_{1,n},\theta}(y_1^k)$ defined in (6), when indexed by θ, inherits of the invariance with respect to the parameter θ when conditioning on a sufficient statistics.

Consider the Gamma density

$$f_{r,\theta}(x) := \frac{\theta^{-(r+1)}}{\Gamma(r+1)} x^r \exp -x/\theta \quad \text{for } x > 0. \tag{8}$$

As r varies in $(-1,\infty)$ and θ is positive, the density runs in an exponential family with parameters r and θ, and sufficient statistics $t(x) := \log x$ and $u(x) := x$ respectively for r and θ. Given a data set $\mathbf{x}_1,\ldots,\mathbf{x}_n$ obtained through sampling from i.i.d. r.v's $\mathbf{X}_1,\ldots\mathbf{X}_n$ with density f_{r_T,θ_T} the resulting sufficient statistics are respectively $t_{1,n} := \log \mathbf{x}_1 + \ldots + \log \mathbf{x}_n$ and $u_{1,n} := \mathbf{x}_1 + \ldots + \mathbf{x}_n$. We consider two parametric models $(f_{r_T,\theta}, \theta \geq 0)$ and $(f_{r,\theta_T}, r > -1)$ respectively assuming r_T or θ_T known.

We first consider sufficiency of $\mathbf{U}_{1,n} := \mathbf{X}_1 + \ldots + \mathbf{X}_n$ in the first model. The density $g_{u_{1,n},(r_T,\theta_T)}(y_1^k)$ should be free of the current value of the true parameter θ_T of the parameter under which the data are drawn. However as appears in (6) the unknown value θ_T should be used in its very definition. We show by simulation that whatever the value of θ inserted in place of θ_T in (6) the value of the likelihood of $\mathbf{x}_1^k$ under $g_{u_{1,n},(r_T,\theta)}$ does not depend upon θ. We thus observe that $\mathbf{U}_{1,n}$ is "sufficient" for θ in the conditional density approximating $p_{u_{1,n},(r_T,\theta)}$, as should hold as a consequence of Theorem 1 . Say that $\mathbf{U}_{1,n}$ is quasi sufficient for θ in $g_{u_{1,n},(r_T,\theta)}$ if this loose invariance holds.

Similarly the same fact occurs in the model $(f_{r,\theta_T}, r > -1)$.

In both cases whatever the value of the parameter θ (Figure 1) or r (Figure 2), the likelihood of $\mathbf{x}_1^k$ remains constant.

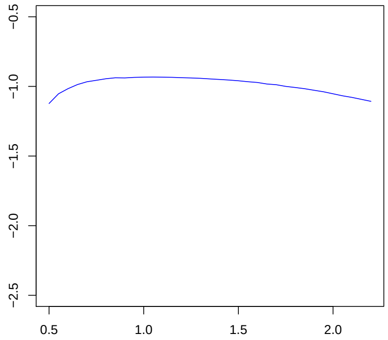

FIGURE 1. *Proxy of the conditional likelihood of X_1^k under $g_{u_{1,n},(r_T,\theta)}$ as a function of θ for $n = 100$ and $k = 80$ in the Gamma case.*

We also consider the Inverse Gaussian distribution with density

$$f_{\lambda,\mu}(x) := \left[\frac{\lambda}{2\pi}\right]^{1/2} \exp -\frac{\lambda\,(x-\mu)^2}{2\mu^2 x} \quad \text{for } x > 0$$

with both parameters λ and μ be positive. Given a data set $\mathbf{x}_1,...,\mathbf{x}_n$ generated from the i.i.d. r.v's $\mathbf{X}_1,...,\mathbf{X}_n$ with density $f_{\mu,\lambda}$, the resulting sufficient statistics are respectively $t_{1,n} := \mathbf{x}_1 + ... + \mathbf{x}_n$ and $u_{1,n} := \mathbf{x}_1^{-1} + ... + \mathbf{x}_n^{-1}$. Similarly as for the Gamma case we draw the likelihood of a subsample $\mathbf{x}_1^k$ under $g_{u_{1,n},(\lambda,\mu_T)}$ with $\mathbf{T}_{1,n} := \mathbf{X}_1 + ... + \mathbf{X}_n$,which is a sufficient statistics for μ (Figure 3), and upon $\mathbf{U}_{1,n} := \mathbf{X}_1^{-1} + ... + \mathbf{X}_n^{-1}$ which is sufficient for λ (Figure 4). In either cases the other coefficient is kept fixed at the true value of the parameter generating the sample. As for the Gamma case these curves show the invariance of the proxy of the conditional density with respect to the parameter for which the chosen statistics is sufficient.

3.2. Rao-Blackwellisation

Rao-Blackwell Theorem holds regardless of whether biased or unbiased estimators are used, since conditioning reduces the MSE. Although its statement is rather weak, in practice the improvement is often enormous. New interest in Rao-Blackwellisation procedures have risen in the recent years, conditioning on ancillary variables (see [15] for a survey on ancillaries in conditional inference); specific Rao-Blackwellisation schemes have been proposed in [6], [7], [28], [30] and [16], whose purpose is to improve the variance of a given statistics (for example a tail probability) under a *known* distribution, through a simulation scheme under this distribution; the ancillary variables used in the simulation process itself are used as conditioning ones for the Rao-Blackwellisation of the statistics. The present approach is more classical in this respect, since we do not assume that the parent distribution is known; conditioning on a sufficient statistics $\mathbf{U}_{1,n}$ with respect to the parameter θ and simulating samples according to the approximating density $g_{u_{1,n},\theta}$ will produce the improved estimator.

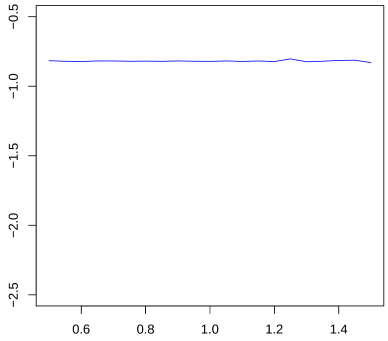

FIGURE 2. *Proxy of the conditional likelihood of X_1^k under $g_{t_{1,n},(r,\theta_T)}$ as a function of r for $n = 100$ and $k = 80$ in the Gamma case.*

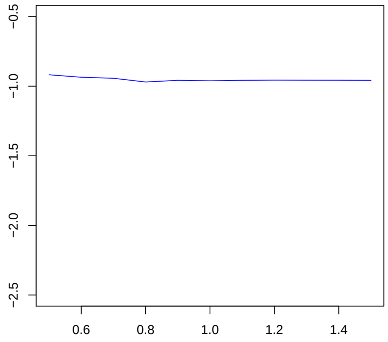

FIGURE 3. *Proxy of the conditional likelihood of X_1^k under $g_{t_{1,n},(\lambda_T,\mu)}$ as a function of μ for $n = 100$ and $k = 80$ in the Inverse Gaussian case.*

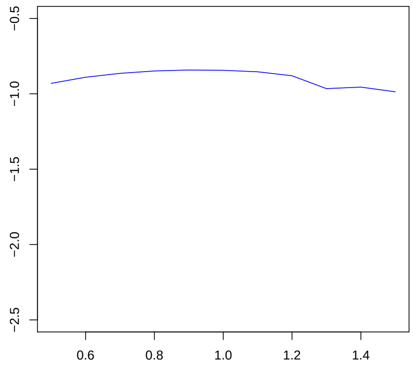

FIGURE 4. *Proxy of the conditional likelihood of X_1^k under $g_{u_{1,n},(\lambda,\mu_T)}$ as a function of λ for $n = 100$ and $k = 80$ in the Inverse Gaussian case.*

Since $\mathbf{U}_{1,n}$ is quasi sufficient for the parameter θ in $g_{u_{1,n},\theta}$ it can be used in order to obtain improved estimators of θ_T through Rao Blackwellization. We shortly illustrate the procedure and its results on some toy cases. Consider again the Gamma family defined here-above with canonical parameters r and θ.

First the parameter to be estimated is θ_T. A first unbiased estimator is chosen as

$$\widehat{\theta_2} := \frac{X_1 + X_2}{2r_T}.$$

Given an i.i.d. sample $\mathbf{X}_1^n$ with density f_{r_T,θ_T} the Rao-Blackwellised estimator of $\widehat{\theta_2}$ is defined through

$$\theta_{RB,2} := E\left(\widehat{\theta_2}\middle|\mathbf{U}_{1,n}\right)$$

whose variance is less than $Var\widehat{\theta_2}$. Given the data set $\mathbf{x}_1,...\mathbf{x}_n$ the estimate of $\theta_{RB,2}$ is produced through simulation of as many $\widehat{\theta_2}$'s as wished, under $g_{u_{1,n},(r_T,\theta_T)}$. Denote $\widehat{\theta}_{RB,2}$ the resulting Rao-Blawellised estimator of $\widehat{\theta_2}$.

Consider $k = 2$ in $g_{u_{1,n},(r_T,\theta_T)}(y_1^k)$ and let (Y_1,Y_2) be distributed according to $g_{u_{1,n},(r_T,\theta_T)}(y_1^2)$; note that any value θ can be used in practice instead of the unknown value θ_T, by quasi sufficiency of $\mathbf{U}_{1,n}$. Replications of (Y_1,Y_2) produce an estimator $\widehat{\theta}_{RB,2}$ for fixed $u_{1,n}$; we have used 1000 replications (Y_1,Y_2). Iterating on 1000 simulations of the runs $\mathbf{X}_1^n$ produces, for $n = 100$ an i.i.d. sample with size 1000 of $\widehat{\theta}_{RB,2}$'s and $Var\theta_{RB,2}$ is estimated. The resulting variance shows a net improvement with respect to the estimated variance of $\widehat{\theta}_2$. It is of some interest to confront this gain in variance as the number of terms involved in $\widehat{\theta}_k$ increases together with k. As k approaches n the variance of $\widehat{\theta}_k$ approaches the Cramer-Rao bound. The graph below shows the decay in variance of $\widehat{\theta}_k$. We note that whatever the value of k the estimated value of the variance of $\theta_{RB,k}$ is constant. This is indeed an illustration of Lehmann-Scheffé's theorem.

Remark 2. *Lockhart and O'Reilly ([21]) establish, under certain conditions and for fixed k, the asymptotic equivalence of the plug-in estimate for the distribution $P_{\theta_{ML}}\left(\mathbf{X}_1^k \in B\right)$ and the*

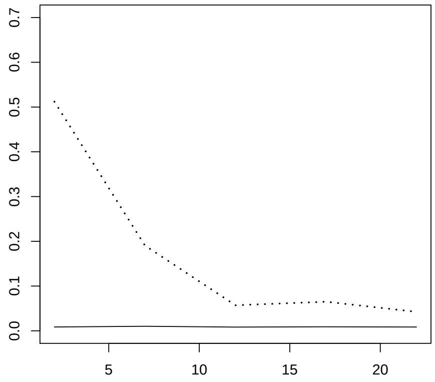

FIGURE 5. *Variance of $\widehat{\theta}_k$, the initial estimator (dotted line), along with the variance of $\theta_{RB,k}$, the Rao-Blackwellised estimator (solid line) with $n = 100$ as a function ok k.*

Rao-Blackwell estimate $P\left(\mathbf{X}_1^k \in B \middle| \mathbf{U}_{1,n}\right)$ where θ_{ML} is the maximum likelihood estimator of θ_T based on the whole sample $\mathbf{X}_1^n$ (this result is known as Moore's conjecture (see [25])). They also provide rates for this convergence.

4. Exponential models with nuisance parameters

4.1. Conditional inference in exponential families

We consider the case when the parameter consists in two distinct subparameters, one of interest denoted θ and a nuisance component denoted η. As is well known, conditioning on a sufficient statistics for the nuisance parameter produces a new exponential family which is free of it. Assuming the observed dataset $\mathbf{x}_1^n := (\mathbf{x}_1, ..., \mathbf{x}_n)$ resulting from sampling of a vector $\mathbf{X}_1^n := (\mathbf{X}_1, ..., \mathbf{X}_n)$ of i.i.d. random variables with distribution in the initial exponential model, and denoting $\mathbf{U}_{1,n}$ a sufficient statistics for η, simulation of samples under the conditional distribution of $\mathbf{X}_1^n$ given $\mathbf{U}_{1,n} = u_{1,n}$ and $\theta = \theta_0$ for some θ_0 produces the basic ingredient for Monte Carlo tests with $H0 : \theta_T = \theta_0$ where θ_T stands for the true value of the parameter of interest. Changing θ_0 for other values of the parameter of interest produces power curves as functions of the level of the test. This is the well known principle of Monte Carlo tests, which are considered hereunder. We consider a steep but not necessarily regular exponential family exponential family $\mathscr{P} := \{P_{\mathbf{X},(\theta,\eta)}, (\theta,\eta) \in \mathscr{N}\}$ defined on $\mathbb{R}$ with canonical parametrization (θ,η) and minimal sufficient statistics (t,u) defined through the density

$$p_{\mathbf{X},(\theta,\eta)}(x) := \frac{dP_{\mathbf{X},(\theta,\eta)}(x)}{dx} = \exp\left[\theta t(x) + \eta u(x) - K(\theta,\eta)\right] h(x). \tag{9}$$

For notational convenience and without loss of generality both θ and η belong to $\mathbb{R}$. Also the model can be defined on $\mathbb{R}^d$, $d > 1$, at the cost of similar but more involved tools. The natural parameter space is $\mathcal{N}$ (which is a convex set in $\mathbb{R}^2$) defined as the effective domain of

$$k(\theta,\eta) := \exp\left[K(\theta,\eta)\right] = \int \exp\left[\theta t(x) + \eta u(x)\right] h(x) dx. \tag{10}$$

As above denote $\mathbf{x}_1^n{:=}(\mathbf{x}_1,...,\mathbf{x}_n)$ be the observed values of n i.i.d. replications of a general random variable $\mathbf{X}$ with density (9). Denote

$$t_{1,n} := \sum_{i=1}^{n} t(\mathbf{x}_i) \text{ and } u_{1,n} := \sum_{i=1}^{n} u(\mathbf{x}_i). \tag{11}$$

Basu [3] discusses ten different ways for eliminating the nuisance parameters, among which conditioning on sufficient statistics and consider UMPU tests pertaining to the parameter of interest. In most cases, the density of $\mathbf{T}_{1,n}$ given $\mathbf{U}_{1,n} = u_{1,n}$ is unknown. Two main ways have been developed to deal with this issue: approximating this conditional density of a statistics or simulating samples from the conditional density. These two approaches are combined hereunder.

The classical technique is to approximate this conditional density using some expansion. Then integration produces critical values. For example, Pedersen [26] defines the mixed Edgeworth-saddlepoint approximation, or the single saddlepoint approximation. However, the main issue of this technique is that the approximated density still depends on the nuisance parameter. In order to obtain the expansion, some suitable values for the parameter of interest and for the nuisance parameter have to be chosen. In the method developed here, as seen before, the conditional approximated density inherits of the invariance with respect to the nuisance parameter when conditioning on a sufficient statistics pertaining to this parameter.

Rephrasing the notation of Section 2 in the present setting the MLE (θ_{ML},η_{ML}) satisfies

$$\left.\frac{\partial K(\theta,\eta)}{\partial\eta}\right|_{\theta_{ML},\eta_{ML}} = u_{1,n}/n$$

and therefore $u_{1,n}/n$ converges to $\left(\frac{\partial K(\theta_T,\eta)}{\partial\eta}\right)^{-1}(\eta_T)$.

For notational clearness denote μ the expectation of $u(\mathbf{X}_1)$ and σ^2 its variance under (θ_T,η_T) , hence

$$\mu := \mu_{(\theta_T,\eta_T)} := \partial K(\theta_T,\eta_T)/\partial\eta \qquad \sigma^2 := \sigma^2_{(\theta_T,\eta_T)} := \partial^2 K(\theta_T,\eta_T)/\partial r^2$$

Assume at present θ_T and η_T known. It holds

$$\phi(r) := E_{(\theta_T,\eta_T)} \exp[ru(\mathbf{X})] = \exp\left[K(\theta_T,\eta_T+r) - K(\theta_T,\eta_T)\right]$$

and

$$\begin{aligned} m(r) &= \mu_{(\theta_T,\eta_T+r)} \\ s^2(r) &= \sigma^2_{(\theta_T,\eta_T+r)} \\ \mu_3(r) &= \partial^3 K(\theta_T,\eta_T+r)/\partial\eta^3 \,. \end{aligned}$$

Further

$$\pi^{\alpha}_{u,\theta_T,\eta_T}(x) := \frac{\exp ru(x)}{\phi(r)} p_{\mathbf{X},(\theta_T,\eta_T)}(x) = p_{\mathbf{X},(\theta_T,\eta_T+r)}(x) \tag{12}$$

for any given α in the range of $P_{\mathbf{X},(\theta_T,\eta_T)}$. In the above formula (12) the parameter r denotes the only solution of the equation

$$m(r) = \alpha.$$

For large k depending on n, using Monte Carlo tests based on runs of length k instead of n does not affect the accuracy of the results.

4.2. Application of conditional sampling to MC tests

Consider a test defined through $H0 : \theta_T = \theta_0$ versus $H1 : \theta_T \neq \theta_0$. Monte Carlo (MC) tests aim at obtaining $p-$values through simulation when the distribution of the desired test statistics under $H0$ is either unknown or very cumbersome to obtain; a comprehensive reference is [17].

Recall the principle of those tests: denote t the observed value of the studied statistic based on the dataset and let $t_2,..,t_L$ the values of the resulting test statistics obtained through the simulation of $L-1$ samples $\mathbf{X}_1^n$ under $H0$. If t is the Mth largest value of the sample $(t,t_2,...,t_L)$, $H0$ will be rejected at the $\alpha = M/L$ significance level, since the rank of t is uniformly distributed on the integer $2,...,L$ when $H0$ holds. The present MC procedure uses simulated samples under the proxy of $p_{u_{1,n},(\theta_0,\eta_T)}$. Using quasi-sufficiency of $\mathbf{U}_{1,n}$ we may use any value in place of η_T; we have compared this simple choice with the common use, inserting the MLE $\hat{\eta}_{\theta_0}$ in place of η_T in $g_{u_{1,n},(\theta_0,\eta_T)}$. This estimate $\hat{\eta}_{\theta_0}$ is the MLE of η_T in the one parameter family $p_{\mathbf{X},(\theta_0,\eta)}$ defined through (9); this choice follows the commonly used one, as advocated for instance in [26] and [27]. Innumerous simulation studies support this choice in various contexts; we found no difference in the resulting procedures.

Consider the problem of testing the null hypothesis $H0 : \theta_T = \theta_0$ against the alternative $H1 : \theta_T > \theta_0$ in model (9) where η is the nuisance parameter.

When $p_{u_{1,n},(\theta_0,\eta_T)}$ is known, the classical conditional test $H0 : \theta_T = \theta_0$ versus $H1 : \theta_T > \theta_0$ with level α is UMPU.

Substituting $p_{u_{1,n},(\theta_0,\eta_T)}\,(\mathbf{X}_1^n = x_1^n | \mathbf{U}_{1,n} = u_{1,n})$ by $g_{u_{1,n},(\theta_0,\eta_T)}\left(x_1^k\right)$ defined in (6), i.e. substituting the test statistics $\mathbf{T}_1^n$ by $\mathbf{T}_1^k$ and $p_{\theta_0}\left(\mathbf{X}_1^k = x_1^k | \mathbf{U}_{1,n} = u_{1,n}\right)$ by $g_{u_{1,n},(\theta_0,\eta_T)}\left(x_1^k\right)$ i.e. changing the model for a proxy while keeping the same parameter of interest θ yields the conditional test with level α

$$\psi_\alpha(x_1^k) := \begin{cases} 1 \ \text{ if } \ t_{1,k} > t_\alpha \\ \gamma \ \text{ if } \ t_{1,k} = t_\alpha \\ 0 \ \text{ if } \ t_{1,k} < t_\alpha \end{cases}$$

and

$$E_{G_{u_{1,n},(\theta_0,\eta_T)}}[\psi_\alpha(X_1^k)] = \alpha$$

i.e. $\alpha := \int \mathbb{1}_{t_{1,k}>t_\alpha} g_{u_{1,n},(\theta_0,\eta_T)}\left(x_1^k\right) dx_1...dx_k$. Its power under a simple hypothesis $\theta_T = \theta$ is defined through

$$\beta_{\psi_\alpha}(\theta|u_n) = E_{G_{u_{1,n},(\theta_0,\eta_T)}}[\psi_\alpha(\mathbf{X}_1^k)].$$

By quasi-sufficiency of $\mathbf{U}_{1,n}$ with respect to η any value can be inserted in $g_{u_{1,n},(\theta_0,\eta_T)}$ in place of η_T.

Recall that the parametric bootstrap produces samples from a parametric model which is fitted to the data, often through maximum likelihood. In the present setting, the parameter θ is set to θ_0 and the nuisance parameter η is replaced by its estimator $\widehat{\eta}_{\theta_0}$ which is the MLE of η_T when the parameter θ is fixed at the value θ_0 defining $H0$. Comparing their exact conditional MC tests with parametric bootstrap ones for Gamma distributions, Lockhart et al [21] conclude that no significant difference can be noticed in terms of level or in terms of power. We proceed in the same vein, comparing conditional sampling MC tests with the parametric bootstrap ones, obtaining again similar results when the nuisance parameter is estimated accurately. However the results are somehow different when the nuisance parameter cannot be estimated accurately, which may occur in various cases.

4.3. Unimodal Likelihood: testing the coefficients of a Gamma distribution

Let $\mathbf{X}_1^n$ be an i.i.d. sample of random variables with Gamma distribution f_{r_T,θ_T} and $\mathbf{x}_1,...,\mathbf{x}_n$ the resulting data set. As r and θ vary this distribution is a two parameter exponential family. The statistics $\mathbf{T}_{1,n} := \log \mathbf{X}_1 + ... + \log \mathbf{X}_n$ is sufficient for r and $\mathbf{U}_{1,n} := \mathbf{X}_1 + ... + \mathbf{X}_n$ is sufficient for the parameter θ. Consider MC conditional test with $H0 : r_T = r_0$

Denote $u_{1,n} = \sum_{i=1}^n \mathbf{x}_i$ and $\widehat{\theta}_{r_0}$ the MLE of θ_T. Calculate for $l \in \{2, L\}$

$$t_l := \sum_{i=0}^k \log\left(Y_i(l)\right).$$

where the Y_i' are a sample from $g_{u_{1,n},\left(r_0,\widehat{\theta}_{r_0}\right)}$.

Consider the corresponding parametric bootstrap procedure for the same test, namely simulate $Z_i(l)$, $2 \le l \le L$ and $0 \le i \le k$ with distribution $f_{r_0,\widehat{\theta}_{r_0}}$; denote

$$s_l := \sum_{i=0}^k \log\left(Z_i(l)\right).$$

In this example simulation shows that for any α the Mth largest value of the sample $(t, t_2, ..., t_L)$ is very close to the corresponding empirical M/L-quantile of s_l's. Hence Monte Carlo tests through parametric bootstrap and conditional compete equally. Also in terms of power, irrespectively in terms of α and in terms of alternatives (close to $H0$), the two methods seem to be equivalent.

MC conditional test with $H0 : \theta_T = \theta_0$ Denote $t_{1,n} = \sum_{i=1}^n \log(\mathbf{x}_i)$ and $\widehat{r}_{\theta_0}$ the MLE of r_T. Calculate for $l \in \{2, L\}$

$$t_l := \sum_{i=0}^k Y_i(l)$$

where the Y_i' are a sample from $g_{u_{1,n},\left(\widehat{r}_{\theta_0},\theta_0\right)}$ and, as above define accordingly

$$s_l := \sum_{i=0}^k \log\left(Z_i(l)\right)$$

where the $Z_i(l)$'s are simulated under $f_{\widehat{r_{\theta_0}},\theta_0}$.

As above, parametric bootstrap and conditional sampling yield equivalent Monte Carlo tests in terms of power function under alternatives close to $H0$.

In the two cases studied above the value of k has been obtained through the rule exposed in section 3.2 of [4].

4.3.1. Bimodal likelihood: testing the mean of a normal distribution in dimension 2

In contrast with the above mentioned examples, the following case study shows that estimation through the unconditional likelihood may fail to provide consistent estimators when the likelihood surface has multiple critical points.

Sundberg [31] proposes four examples that allow likelihood multimodality. Two of them can also be found in [11] and [12], and in [2], Ch 2. We consider the "Normal parabola" model which is a curved (2, 1) family (see Example 2.35 in [2], Ch 2). Two independent Gaussian variates have unknown means and known variances; their means are related by a parabolic relationship.

Let $\mathbf{X}$ and $\mathbf{Y}$ be two independent Gaussian r.v.'s with same variance σ_T^2 with expectation ψ_T and ψ_T^2. In the present example $\sigma_T^2 = 1$ and $\psi_T = 2$.

Let $(\mathbf{x}_i, \mathbf{y}_i)$, $1 \leq i \leq n$ be i.i.d. realizations of $(\mathbf{X}_i, \mathbf{Y}_i)$.

The parameter of interest is σ^2 whilst the nuisance parameters is ψ. Derivation of the likelihood function of the observed sample with respect to ψ yields the following equation

$$(u_{1,n} - \psi) + 2\psi \left(v_{1,n} - \psi^2\right) = 0$$

with $u_{1,n} := \mathbf{x}_1 + \ldots + \mathbf{x}_n$ and $v_{1,n} := \mathbf{y}_1 + \ldots + \mathbf{y}_n$. Define accordingly $\mathbf{U}_{1,n}$ and $\mathbf{V}_{1,n}$.The following table shows that the likelihood function is bimodal in ψ.

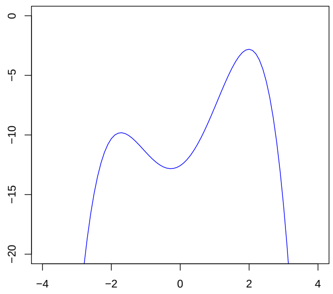

FIGURE 6. *Bimodal likelihood in ψ.*

Estimation of the nuisance parameter ψ is performed through the standard Newton Raphson method. The Newton-Raphson optimizer of the likelihood function converges to the true value when the initial value is larger than 1 and fails to converge to $\psi_T = 2$ otherwise. Henceforth the ML estimation based on this preliminary estimate of the nuisance parameter may lead to erroneous estimates of the parameter of interest. Indeed according to the initial value we obtained estimators of ψ_T close to 2 or to -2. When the estimator of the nuisance parameter is close to its true value 2 then parametric bootstrap yields Monte Carlo tests with power close to 1 for any α and any alternative close to $H0$. At the contrary when this estimate is close to the second maximizer of the likelihood (i.e. close to -2) then the resulting Monte Carlo test based on parametric bootstrap has power close to 0 irrespectively of the value of α and of the alternative, when close to $H0$. In contrast with these results, Monte Carlo tests based on conditional sampling provide powers close to 1 for any α; we have considered alternatives close to $H0$. This result is of course a consequence of the quasi sufficiency of the statistics $(\mathbf{U}_{1,n}, \mathbf{V}_{1,n})$ for the parameter $\left(\psi, \psi^2\right)$ of the distribution of the sample $(\mathbf{x}_i, \mathbf{y}_i)_{i=1,\dots,n}$; see next paragraph for a discussion of this point.

4.4. Estimation through conditional likelihood

Theorem 1 states that the density $g_{u_{1,n},\theta_T}$ on $\mathbb{R}^k$ approximates $p_{u_{1,n},\theta_T}$ on the sample $\mathbf{X}_1^n$ generated under $P_{u_{1,n},\theta_T}$. However, in some cases, the r.v.'s $\mathbf{X}_i$'s in Theorem 1 may at time be generated under some other parameters, say under $P_{u_{1,n},\theta_0}$. This is indeed required here, where a procedure somehow similar to parametric bootstrap will be achieved. Theorem 11 in [4] states that the approximation scheme holds true in this case. The reason for this lies in the fact that approximation schemes hold on typical paths generated by conditional distributions stemming from any basic light tailed distributions P (i.e. with finite moment generating function in a non void neighborhood of 0), and not only under those based on P_{θ_T}; this result is stronger than approximation in total variation norm as stated in Theorem 1.

Theorem 3. *With the same hypotheses and notation as in Theorem 1,*

$$p_{\theta_T}\left(\mathbf{X}_1^k = Y_1^k \,|\mathbf{U}_{1,n} = u_{1,n}\right) = g_{u_{1,n},\theta_T}(Y_1^k)(1 + o_{P_{u_{1,n},\theta_0}}(\varepsilon_n (\log n)^2)).$$

An easy extension of the above result allows to change θ_0 by any θ_n converging to θ_0.

Considering model (9) we intend to perform an estimation of θ_T irrespectively upon the value of η_T. Denote $\widehat{\eta}_\theta$ the MLE of η_T when θ holds; the model $p_{\mathbf{X},(\theta,\widehat{\eta}_\theta)}(x)$ is a one parameter model which is fitted to the data for any peculiar choice of θ.The optimizer in θ of the resulting likelihood function is the global MLE. Properties of the resulting estimators strongly rely on the consistency properties of $\widehat{\eta}_\theta$ at any given θ.

Consider the consequence of Theorem 3. Condition on the value of the sufficient statistics $\mathbf{U}_{1,n}$, and consider the conditional likelihood of the observed subsample $\mathbf{x}_1^k$ under parameter $(\theta, \widehat{\eta}_\theta)$; recall that $\mathbf{x}_1^k$ is generated under (θ_T, η_T) . By Theorem 3 this likelihood is approximated by $g_{u_{1,n},(\theta,\widehat{\eta}_\theta)}\left(\mathbf{x}_1^k\right)$ with a small relative error. Conditioned likelihood estimation is performed optimizing $g_{u_{1,n},(\theta,\widehat{\eta}_\theta)}\left(\mathbf{x}_1^k\right)$ upon θ. Any value of the nuisance parameter η can be used in place of $\widehat{\eta}_\theta$ as seen in Section 3.1.

In most cases, as the normal, gamma or inverse-gaussian, estimations through the unconditional likelihood or through conditional likelihood give a consistent estimator.

We consider the example of the bimodal likelihood from the above subsection, inheriting of the notation and explore the behavior of the proxy of the conditional likelihood of the sample

$(\mathbf{x}_i, \mathbf{y}_i)$, $1 \leq i \leq n$ when conditioning on $u_{1,n}$ and $v_{1,n}$, as a function of σ^2. This likelihood writes

$$\begin{aligned} & L\left(\sigma^2 \middle| u_{1,n}, v_{1,n}\right) \\ & = p_{u_{1,n}\sigma^2}\left(\mathbf{x}_1^n\right) p_{v_{1,n}\sigma^2}\left(\mathbf{y}_1^n\right) \end{aligned}$$

where we have used the independence of the r.v.'s $\mathbf{X}_i$'s and $\mathbf{Y}_i$'s.

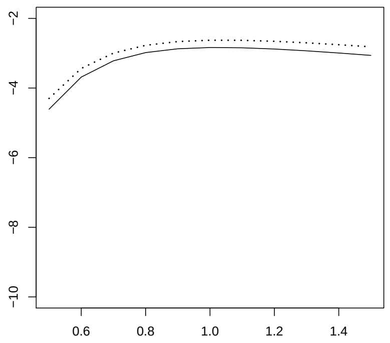

FIGURE 7. *Proxy of the conditional likelihood (solid line) along with the classical likelihood (dotted line) as function of σ^2 for $n = 100$ and $k = 99$ in the case where a good initial point in Newton-Raphson procedure is chosen.*

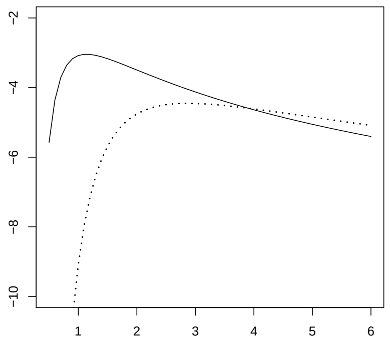

FIGURE 8. *Proxy of the conditional likelihood (solid line) along with the classical likelihood (dotted line) as function of σ^2 for $n = 100$ and $k = 99$ in the case where a bad initial point in Newton-Raphson procedure is chosen.*

Applying Theorem 1 to the above expression it appears that ψ cancels in the resulting densities $g_{u_{1,n,\sigma^2}}$ and $g_{v_{1,n,\sigma^2}}$. This proves that the proxy of the conditional likelihood provides consistent estimation of σ_T^2 as shown on Figures 7 and 8 (see the solid lines).

On Figure 7, the dot line is the likelihood function

$$L\left(\sigma^2\right) := \sum_{i=1}^{n} \log p_{\mathbf{X},\left(\sigma^2,\widehat{\psi}_{\sigma^2}\right)}(\mathbf{x}_i)$$

where $\widehat{\psi}_{\sigma^2}$ is a consistent estimator of the nuisance parameter; the resulting maximizer in the variable σ^2 is close to $\sigma_T^2 = 1$. At the opposite in Figure 8 an inconsistent preliminary estimator of ψ_T obtained through a bad tuning of the initial point in the Newton-Raphson procedure leads to inconsistency in the estimation of σ_T^2, the resulting likelihood function being unbounded.

References

[1] Barndorff-Nielsen, O.E. (1978). Information and exponential families in statistical theory. Wiley Series in Probability and Mathematical Statistics. Chichester: John Wiley & Sons.

[2] Bardnörff-Nielsen, O.E. and Cox, R.R. (1994). Inference and Asymptotics. Chapman & Hall, London.

[3] Basu, D. (1977). On the elimination of nuisance parameters. J. Amer. Statist. Assoc. 72 (1977), no. 358, 355–366.

[4] Broniatowski M. and Caron, V. (2011). Long runs under a conditional limit distribution. Ann. Appl. Probab. 24 (2014), no. 6, 2246–2296.

[5] Broniatowski M. and Caron, V. (2011). Small variance estimators for rare event probabilities. ACM Trans. Model. Comput. Simul. 23 (2013), no. 1, Art. 7, 23 pp.

[6] Casella, G. and Robert, C. P. (1996). Rao-Blackwellisation of sampling schemes. Biometrika 83 , no. 1, 81–94.

[7] Casella, G. and R. C. P. (1998). Post-processing accept-reject samples: recycling and rescaling. J. Comput. Graph. Statist. 7 , no. 2, 139–157

[8] Cheng, R.C.H. (1984). Generation of inverse Gaussian variates with given sample mean and dispersion. Appl. Statist. 33, 309–16

[9] Dembo, A. and Zeitouni, O. (1996) Refinements of the Gibbs conditioning principle. *Probab. Theory Related Fields* **104** 1–14.

[10] Diaconis, P. and Freedman, D.A. (1988) Conditional limit theorems for exponential families and finite versions of de Finetti's theorem. *J. Theoret. Probab.* **1** 381–410.

[11] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). Ann. Statist. 3, 1189-1242.

[12] Efron, B. (1978). The geometry of exponential families. Ann. Statist. 6, 362-376.

[13] Efron, B. (1979), Bootstrap methods: another look at the jackknife. Ann. Statist. 7 , no. 1, 1–26.

[14] Engen, S. and Lillegard, M. (1997). Stochastic simulations conditioned on sufficient statistics. Biometrika, 84, 235–240.

[15] Fraser, D. A. S. (2004). Ancillaries and conditional inference. With comments by Ronald W. Butler, Ib M. Skovgaard, Rudolf Beran and a rejoinder by the author. Statist. Sci. 19 , no. 2, 333–369

[16] Iacobucci, A., Marin, J.-M., Robert, C. (2010). On variance stabilisation in population Monte Carlo by double Rao-Blackwellisation. Comput. Statist. Data Anal. 54 , no. 3,

[17] Jöckel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. Ann. of Stat., 14, 336–347.

[18] Lehmann, E.L. (1986). Testing Statistical Hypotheses. Springer.

[19] Lindqvist, B.H., Taraldsen, G., Lillegärd, M. and Engen, S. (2003). A counterexample to a claim about stochastic simulations. Biometrika 90 , no. 2, 489–490.

[20] Lindqvist, B.H. and Taraldsen, G. (2005). Monte Carlo conditioning on a sufficient statistic. Biometrika 92 , no. 2, 451–464.

[21] Lockhart, R. and O'Reilly, F. (2005) A note on Moore's conjecture. Statist. Probab. Lett. 74 , no. 2, 212–220.

[22] LOCKHART, R A., O REILLY, F J. AND STEPHENS, A. (2007) Use of the Gibbs sampler to obtain conditional tests, with applications. Biometrika 94 , no. 4, 992–998.

[23] LOCKHART, R.A. AND STEPHENS, M. A. (1994). Estimation and tests of fit for the three–parameter Weibull distribution. J. Roy. Statist. Soc.. B. 56:491–500.

[24] O'REILLY, F. AND GRAVIA-MEDRANO, L. (2006). On the conditional distribution of goodness-of-fit tests. Commun. Statist. A, 35:541–9.

[25] MOORE, DAVID S. (1973). A note on Srinivasan's goodness-of-fit test. Biometrika 60 , 209–211.

[26] PEDERSEN, B.V., (1979). Approximating conditional distributions by the mixed Edgeworth-saddlepoint expansion. Biometrika, 66(3), 597–604.

[27] PACE L. AND SALVAN A., (1992). A note on conditional cumulants in canonical exponential families. Scand. J. Statist., 19, 185–191.

[28] PERRON, F. (1999). Beyond accept-reject sampling. Biometrika 86 , no. 4, 803–813

[29] REID, N. (1995). The roles of conditioning in inference. With comments by V. P. Godambe, Bruce G. Lindsay and Bing Li, Peter McCullagh, George Casella, Thomas J. DiCiccio and Martin T. Wells, A. P. Dawid and C. Goutis and Thomas Severini. With a rejoinder by the author. Statist. Sci. 10 , no. 2, 138–157, 173–189, 193–196.

[30] DOUC, R. AND ROBERT, C.P.(2011). A vanilla Rao-Blackwellization of Metropolis-Hastings algorithms. Ann. Statist. 39 , no. 1, 261–277

[31] SUNDBERG, R. (2009). Flat and multimodal likelihoods and model lack of fit in curved exponential families. Research Report 2009:1, http://www.math.su.se/matstat.