The Canadian Journal of Statistics
Vol. 51, No. 3, 2023, Pages 877–896
La revue canadienne de statistique

Confidence sequences with composite likelihoods

Luigi PACE[1], Alessandra SALVAN[2], and Nicola SARTORI[2*]

[1]*Department of Economics and Statistics, University of Udine, Italy*
[2]*Department of Statistical Sciences, University of Padova, Italy*

Key words and phrases: Pseudo-likelihood; replicability; safe inference; statistical evidence; universal inference.

MSC 2020: Primary 62A99; secondary 62L99.

Abstract: In dominated parametric statistical models, confidence sequences provide conservatively valid frequentist inference directly from a likelihood ratio. They ensure a specific mode of replicability when inference is performed on accumulating data: inferential conclusions that are compatible with a guaranteed probability when the sample is enlarged, in the form of overlapping confidence regions. Here we consider both Robbins' mixture confidence sequences and running maximum likelihood confidence sequences recently considered by Wasserman, Ramdas, and Balakrishnan. We compare through simulation the replicability properties of the two kinds of confidence sequences, evaluating, along a prospected enlargement of the sample, the frequency of incompatible estimation intervals and the frequency of failure of simultaneous coverage of the true parameter value. Moreover, we propose a shortcut to extend the application of mixture confidence sequences to pseudo-likelihoods, in particular to composite likelihood. The main assumption required is that normal asymptotic theory offers a good approximation to the density of the maximizer of the pseudo-likelihood. When inference is about a scalar parameter of interest, the computation of the proposed sequence of confidence intervals is straightforward. The method is illustrated by an example with replicability properties evaluated through simulation.

Résumé: Dans les modèles statistiques paramétriques dominés, les séquences de confiance offrent une inférence fréquentiste conservatrice basée sur un rapport de vraisemblance. Elles garantissent un mode de reproductibilité spécifique en cas d'inférence à partir de données accumulées : il s'agit de conclusions inférentielles sous forme de chevauchements de régions de confiance possédant une probabilité garantie lorsque la taille de l'échantillon augmente. Les auteurs de ce travail se sont penchés sur des mélanges de séquences de confiance similaires à celles de Robbins (1970) et sur des séquences de confiance par maximum de vraisemblance comme celles de Wasserman, Ramdas et Balakrishnan (2020, Section 7). Afin de comparer les propriétés de reproductibilité des deux types de séquences de confiance considérées, ils explorent, selon un élargissement prospectif de la taille d'échantillon, la fréquence des intervalles d'estimation incompatibles et la fréquence d'échecs de couverture simultanée de la valeur du paramètre réel. En outre, ils montrent comment appliquer ces mélanges de séquences de confiance aux pseudo-vraisemblances, dont en particulier la vraisemblance composite. La principale hypothèse requise est que la normalité asymptotique offre une bonne approximation de la densité du maximum de la pseudo-vraisemblance. Lorsque l'inférence concerne un paramètre scalaire, l'évaluation de la séquence d'intervalles de confiance proposée est simple. Enfin, la méthode est illustrée au moyen d'un exemple avec des propriétés de reproductibilité évaluées par simulation.

* *Corresponding author: nicola.sartori@unipd.it*

1. INTRODUCTION

Ritualistic application of frequentist inferential tools such as *P*-values, even from likelihood ratio tests, is often pointed out as a source of the replicability crisis in science famously denounced by Ioannidis (2005). Indeed, calibration established on hypothetical repetitions of the experiment that produced the data at hand, considered in isolation, is too vague a guarantee of replicability. Such a calibration gives rise to episodic inferences that are vulnerable, *inter alia*, to selection bias and interim analyses. See Benjamini (2020) on selective inference as a killer of replicability.

Pace & Salvan (2020) underline that to request that compatible, i.e., noncontradictory, conclusions be reached when the sample is enlarged is a better basis than the repeated sampling principle to embed a concept of replicability into statistical theory. This view emphasizes statistical models as models for sequential environments.

With confidence regions for the same parameter, calculated at various sample sizes, inferential conclusions are compatible if these regions overlap, and incompatible if their intersection is empty. When inference is performed on the basis of accumulating data, hasty announcement of provisional conclusions that will turn out to be incompatible with the final conclusions may cause a large reputational damage.

A confidence sequence (Robbins, 1970; see also Darling & Robbins, 1967a,b) is a sequence of confidence regions constructed so as to be all compatible with a guaranteed probability. Research on confidence sequences seems to have been long neglected after the technical contributions in Lai (1976) and Csenki (1979). In recent years, however, there has been a renewed interest, with various aims. See, e.g., Wasserman, Ramdas & Balakrishnan (2020), Howard et al. (2021), Johari et al. (2021), Vovk & Wang (2021), and Howard & Ramdas (2022).

Outside proper sequential settings, use of confidence sequences enhances replicability of the conclusions of a stand-alone study, but, of course, only actual replication in follow-up studies may give the experimental demonstration of a finding. In the more recent literature, confidence sequences are often termed as "anytime-valid confidence regions" and the whole approach as "safe inference". The mixture device—used to obtain confidence sequences by Robbins (1970) and other early contributors—has been superseded by the device of data splitting in Wasserman, Ramdas & Balakrishnan (2020). While computationally more convenient (no integration is required), data splitting infringes the weak likelihood principle and may look dubious unless applied to intractable models when no other tool with frequentist guarantee is available.

With high-dimensional data, often the full likelihood is difficult to specify and inference may be based on a misspecified likelihood such as composite likelihood. A composite likelihood combines dependent likelihoods from small portions of the data using convenient weights. See Varin, Reid & Firth (2011) for a review. See also Pace, Salvan & Sartori (2019) and Fraser & Reid (2020) for results on optimal weights. When composite likelihood is the basis for inference, it is of interest to construct confidence sequences whose estimation regions are all compatible with an at least approximate frequentist guarantee. One proposal with exact frequentist validity is in Nguyen (2020), which generalizes the data splitting device of Wasserman, Ramdas & Balakrishnan (2020) to composite likelihoods.

In this work, we propose a shortcut to construct confidence sequences *à la* Robbins from composite likelihood. The main assumption required is that normal asymptotic theory offers a good approximation to the density of the maximizer of the pseudo-likelihood. When inference is about a scalar parameter of interest, the computation of the proposed sequence of confidence intervals is straightforward.

The outline of the article is as follows. Section 2 offers a brief review of the rationale behind confidence sequences. The main devices to obtain confidence sequences, namely mixture as in Robbins (1970) and splitting as in Wasserman, Ramdas & Balakrishnan (2020), are recalled in Section 3, where simple examples are examined, including simulations. In Section 4, attention is

devoted to confidence sequences associated with asymptotically normal estimators. The particular case of estimators from composite likelihoods is considered in detail in Section 5. Section 6 presents an example of confidence sequences with composite likelihoods, with simulations supporting the claim of approximate validity when the sample size is large enough. Section 7 concludes.

2. CONFIDENCE SEQUENCES

Let the potentially observable data be $y^{(n)} = (y_1, \ldots, y_n)$, a realization of the random vector $Y^{(n)} = (Y_1, \ldots, Y_n)$, $n = 1, 2, \ldots$. We denote by P_θ the joint probability distribution of the sequence $Y^{(\infty)} = (Y_1, Y_2, \ldots)$ and suppose that P_θ belongs to a statistical model with parameter space $\Theta \subseteq \mathbb{R}^p$. Moreover, we assume that $p_n\left(y^{(n)}; \theta\right) > 0$ is the density of $Y^{(n)}$ under P_θ, whose support is independent of θ. Ideally, by observing the sequence $y^{(n)}$ the statistician will eventually discover the truth, that is, the true value of θ in Θ, denoted by θ^*.

An estimation region based on $y^{(n)}$ is a subset of Θ, denoted by $\hat{\Theta}_n = \hat{\Theta}\left(y^{(n)}\right)$ or similar symbols. A confidence sequence is a sequence of estimation regions. A confidence sequence offers compatible inferential conclusions about θ if there are conclusions that are common to all confidence statements, which are thus noncontradictory. Consistency of an estimator that is always contained in $\hat{\Theta}_n$ often entails that the sequence $\hat{\Theta}_n$ shrinks towards the true value of θ.

A confidence sequence $\hat{\Theta}_n$ has persistence level $1 - \varepsilon$, where $0 < \varepsilon < 1$, if, for every $\theta \in \Theta$,

$$P_\theta\left(\theta \in \cap_{n \geq 1} \hat{\Theta}_n\right) \geq 1 - \varepsilon.$$

This implies the frequentist guarantee that, for every $\theta \in \Theta$,

$$P_\theta\left(\cap_{n \geq 1} \hat{\Theta}_n = \emptyset\right) \leq \varepsilon,$$

so that the probability of observing incompatible conclusions from a sequence with persistence level $1 - \varepsilon$ as evidence accumulates is as small as desired. Indeed,

$$P_\theta\left(\cap_{n \geq 1} \hat{\Theta}_n = \emptyset\right) \leq P_\theta\left(\theta \notin \cap_{n \geq 1} \hat{\Theta}_n\right) = 1 - P_\theta\left(\theta \in \hat{\Theta}_n \text{ for every } n \geq 1\right).$$

Confidence sequences $\hat{\Theta}_n$ with persistence level $1 - \varepsilon$ provide conservatively valid frequentist inference. For any given n, $\hat{\Theta}_n$ is an estimation region with confidence level at least $1 - \varepsilon$. As remarked in Wasserman, Ramdas & Balakrishnan (2020, page 16888), such regions are valid at arbitrary stopping times and at arbitrary data-dependent times that are chosen *post hoc*.

A fundamental martingale inequality for the likelihood ratio statistic is the basis for constructing confidence sequences with persistence level $1 - \varepsilon$. Let P and Q denote the joint probability distributions of the sequence $Y^{(\infty)}$ when $Y^{(n)}$, $n = 1, 2, \ldots$, have densities with the same supports $p_n\left(y^{(n)}\right)$ and $q_n\left(y^{(n)}\right)$, respectively. Then

$$P\left(\frac{q_n\left(Y^{(n)}\right)}{p_n\left(Y^{(n)}\right)} \geq k \text{ for some } n\right) \leq \frac{1}{k}, \tag{1}$$

for any $k > 0$.

For $k > 1$, inequality (1) gives a bound to the probability of reaching strongly misleading evidence from the likelihood ratio statistic (Royall, 1997, page 7). Note that, for every given n,

the inequality $P\left(q_n\left(Y^{(n)}\right)/p_n\left(Y^{(n)}\right) \geq k\right) \leq 1/k$ is an easy consequence of Markov's inequality. Result (1) follows from a well-known martingale inequality: see Jacod & Protter (2000, Theorem 26.1). In particular,

$$\frac{q_n\left(Y^{(n)}\right)}{p_n\left(Y^{(n)}\right)} = \frac{q_{n-1}\left(Y^{(n-1)}\right)}{p_{n-1}\left(Y^{(n-1)}\right)} \frac{q_n\left(Y_n|Y^{(n-1)}\right)}{p_n\left(Y_n|Y^{(n-1)}\right)}$$

is a nonnegative martingale with respect to the natural filtration $\mathcal{F}_n = \sigma(Y^{(n)})$. Indeed, under P, the ratio of conditional densities

$$\frac{q_n\left(Y_n|Y^{(n-1)}\right)}{p_n\left(Y_n|Y^{(n-1)}\right)}$$

has expectation, conditional on $Y^{(n-1)}$, equal to 1.

Inequality (1) is a universal basis for constructing a plethora of confidence sequences with persistence level $1-\varepsilon$, each corresponding to a particular choice of $q_n(\cdot)$. The general form is

$$\hat{\Theta}_{1-\varepsilon}\left(y^{(n)}\right) = \left\{\theta \in \Theta \ : \ \frac{q_n\left(y^{(n)}\right)}{p_n\left(y^{(n)};\theta\right)} < \frac{1}{\varepsilon}\right\} = \left\{\theta \in \Theta : \ p_n\left(y^{(n)};\theta\right) > \varepsilon q_n\left(y^{(n)}\right)\right\}. \quad (2)$$

The fact that the confidence sequence $\hat{\Theta}_{1-\varepsilon}\left(y^{(n)}\right)$ defined by (2) has persistence level $1-\varepsilon$ follows from inequality (1) with $p_n\left(y^{(n)}\right) = p_n(y^{(n)};\theta)$ and $k = 1/\varepsilon$. Confidence regions (2) are nested, that is, $\hat{\Theta}_{1-\varepsilon'}\left(y^{(n)}\right) \subseteq \hat{\Theta}_{1-\varepsilon}\left(y^{(n)}\right)$ when $1-\varepsilon' < 1-\varepsilon$. As underlined in Wasserman, Ramdas & Balakrishnan (2020, Section 1, after Remark 4), for a given n, regions of the form (2) are an instance of universal inference, meaning that the procedure has a valid frequentist guarantee with no regularity conditions.

3. MIXTURE AND SPLIT CONFIDENCE SEQUENCES

Robbins' (1970) confidence sequences, hereafter called mixture confidence sequences, have the form (2) with $q_n\left(y^{(n)}\right)$ given by the mixture device

$$q_n\left(y^{(n)}\right) = \int_\Theta p_n(y^{(n)};\theta)\pi(\theta)\, d\theta. \quad (3)$$

Therefore they have the form

$$\hat{\Theta}_{1-\varepsilon}\left(y^{(n)}\right) = \left\{\theta \in \Theta \ : \ p_n\left(y^{(n)};\theta\right) > \varepsilon \int_\Theta p_n\left(y^{(n)};\theta\right)\pi(\theta)\, d\theta\right\}. \quad (4)$$

The weight function $\pi(\theta)$ is a preset probability density over Θ with $\pi(\theta) > 0$ for every $\theta \in \Theta$ and invites Bayesian interpretation. Indeed, in the form (3), $q_n\left(y^{(n)}\right)$ can incorporate prior information about θ. One advantage of the choice (3) in definition (2) is that the maximum likelihood estimate $\hat{\theta}_n = \text{argmax}_{\theta\in\Theta}\, p_n\left(y^{(n)};\theta\right)$ is always a point in $\hat{\Theta}_{1-\varepsilon}\left(y^{(n)}\right)$, because $p_n\left(y^{(n)};\hat{\theta}_n\right) \geq \int_\Theta p_n\left(y^{(n)};\theta\right)\pi(\theta)\, d\theta$. More generally, mixture confidence sequences are likelihood-based. The set $\hat{\Theta}_{1-\varepsilon}\left(y^{(n)}\right)$ is the region of θ values whose likelihood $L\left(\theta; y^{(n)}\right) = p_n\left(y^{(n)};\theta\right)$ is larger than a fraction ε of the integrated likelihood (3).

The *ideal* mixture confidence sequence has $\pi(\theta) \sim D(\theta^*)$; that is, it degenerates at the true θ. This choice produces the sequence

$$\Theta^*_{1-\varepsilon}\left(y^{(n)}\right) = \left\{\theta \in \Theta \ : \ p_n\left(y^{(n)};\theta\right) > \varepsilon p_n\left(y^{(n)};\theta^*\right)\right\}. \tag{5}$$

Of course, sequence (5) has persistence level $1-\varepsilon$ but it is not a feasible confidence sequence. However, θ^* may be estimated using information besides $y^{(n)}$. This fact is at the basis of the data splitting device introduced in Wasserman, Ramdas & Balakrishnan (2020).

Considering for simplicity two groups with the same size n, $n = 1, 2, \ldots$, the data $y^{(2n)}$ are randomly split as $y^{(2n)} = \left(y_0^{(n)}, y_1^{(n)}\right)$ with $y_0^{(n)} = (y_1, \ldots, y_n)$ and $y_1^{(n)} = (y_{n+1}, \ldots, y_{2n})$. These observations are a realization of the random vector $Y^{(2n)} = \left(Y_0^{(n)}, Y_1^{(n)}\right)$, where $Y_0^{(n)} = (Y_1, \ldots, Y_n)$ and $Y_1^{(n)} = \left(Y_{n+1}, \ldots, Y_{2n}\right)$. The split estimation set, denoted by $\tilde{\Theta}_{1-\varepsilon}(Y^{(2n)})$ with realization $\tilde{\Theta}_{1-\varepsilon}\left(y^{(2n)}\right)$, is given by

$$\tilde{\Theta}_{1-\varepsilon}\left(y^{(2n)}\right) = \left\{\theta \in \Theta \ : \ p_n\left(y_0^{(n)};\theta\right) > \varepsilon p_n\left(y_0^{(n)};\hat{\theta}_1^{(n)}\right)\right\}, \tag{6}$$

where $\hat{\theta}_1^{(n)}$ is any consistent estimator of θ depending on $y_1^{(n)}$, typically the maximum likelihood estimator or a regularized form of it. The set (6) is still of the form (2), with n equal to $2n$, $p_{2n}\left(y^{(2n)}\right) = p_n\left(y_0^{(n)};\theta\right)$, and $q_{2n}\left(y^{(2n)}\right) = p_n\left(y_0^{(n)};\hat{\theta}_1^{(n)}\right)$. While mixture confidence sequences agree with the strong likelihood principle, data splitting infringes the weak likelihood principle and consequently it may lead to inferior inferences; see, however, Cox (1975). To somehow accommodate this drawback, Wasserman, Ramdas & Balakrishnan (2020, Section 4) have proposed various de-randomized variants, such as K-fold splitting and subsampling, which we will not discuss.

Actually, in Wasserman, Ramdas & Balakrishnan (2020), the set (6) is proposed as a universal confidence set having, for every fixed sample size $2n$, confidence level at least $1-\varepsilon$, as implied by Markov's inequality. It is not proposed as a confidence sequence with persistence level $1-\varepsilon$. The reason is, we guess, that it is not clear whether the martingale property holds for the sequence

$$\frac{p_n\left(Y_0^{(n)};\hat{\theta}_1^{(n)}\right)}{p_n\left(Y_0^{(n)};\theta\right)}.$$

We conjecture that the sequence (6) may be used as an approximate confidence sequence, with asymptotic persistence level $1-\varepsilon$, where we say that a sequence $\hat{\Theta}_n$ has asymptotic persistence level $1-\varepsilon$ if

$$\lim_{n\to\infty} P_\theta\left(\theta \in \hat{\Theta}_m \ \text{ for every } \ m \geq n\right) \geq 1-\varepsilon.$$

From this perspective, the sampling property of practical interest is that, for a sample size $n_{\min}$ large enough and a hypothetical observation horizon $n_{\max}$ much larger than $n_{\min}$,

$$P_\theta\left(\theta \in \hat{\Theta}_n \ \text{ for every } \ n \ \text{ such that } n_{\min} \leq n \leq n_{\max}\right) \geq 1-\varepsilon.$$

When, as for sequence (6), analytic tools do not provide any guidance, simulation may be used to ascertain whether the above simultaneous frequency guarantee holds. In the simulations in Examples 1 and 2 that follow, the confidence sequence is defined supposing that both datasets

$y_0^{(n)}$ and $y_1^{(n)}$ are increased by one independent observation at a time and do not mix. Thus, new observations come in pairs and each of the two observations is randomly assigned to one of the two groups once and for all.

With independent and identically distributed (i.i.d.) random variables $Y_1, Y_2, \ldots$, Wasserman, Ramdas & Balakrishnan (2020) introduce the confidence sequence with (exact) persistence level $1-\varepsilon$ defined as

$$\overline{\Theta}_{1-\varepsilon}(y^{(n)}) = \left\{ \theta \in \Theta \ : \ \prod_{i=1}^{n} p_1(y_i;\theta) > \varepsilon \prod_{i=1}^{n} p_1(y_i;\hat{\theta}(y^{(i-1)})) \right\}, \tag{7}$$

where, for $i \geq 2$, $\hat{\theta}(y^{(i-1)})$ is any estimate of θ depending on $y^{(i-1)}$, for instance the maximum likelihood estimate or a regularized form of it, while $\hat{\theta}(y^{(0)}) = \theta$, so that, on both sides of the inequality defining the sequence (7), the factor $p_1(y_1;\theta)$ cancels out. If the dimension of θ is $p > 1$, then $p_1(y_1;\theta)$ in (7) may be substituted by the density of a block of p or more observations. The sequence (7) is written supposing that the dataset is increased by one observation at a time. The definition is easily extended to cover cases when data are collected in groups. Wasserman, Ramdas & Balakrishnan (2020, Section 7) refer to the process giving rise to the sequence (7) as "running maximum likelihood ratio" and highlight that the idea originated in Wald (1947) and was further analyzed in Robbins & Siegmund (1972, 1974). In the following, a confidence sequence (6) will be referred to as split-naive while a confidence sequence (7) will be referred to as split-exact. At any given n, split-naive intervals are computationally much more convenient than split-exact intervals. The next two examples compare the average length of confidence intervals from mixture and from split confidence sequences, both naive and exact. The empirical percentage of incompatible inferences and of noncoverage of the true parameter value pertaining to confidence intervals from naive and exact split confidence sequences are also compared. The results for mixture intervals are overall satisfactory.

3.1. Example 1. Normal Population with Known Variance

Suppose that Y_i, $i = 1, 2, \ldots$, are i.i.d. $N(\theta, \sigma_0^2)$, with unknown mean θ and known variance σ_0^2. Sufficiency leads us to consider the sample mean $\overline{Y}_n = \sum_{i=1}^{n} Y_i/n$, with a $N(\theta, \sigma_0^2/n)$ density, $n = 1, 2, \ldots$. The mixture confidence sequence with persistence level $1-\varepsilon$ defined by the weight function

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left\{ -\frac{(\theta-\mu_0)^2}{2\tau_0^2} \right\},$$

corresponding to a $N(\mu_0, \tau_0^2)$ conjugate prior, is given in a closed form by the intervals

$$\overline{y}_n \pm \frac{\sigma_0}{\sqrt{n}} \sqrt{\log \frac{\tau_0^2 + \sigma_0^2/n}{\sigma_0^2/n} + \frac{(\overline{y}_n - \mu_0)^2}{\tau_0^2 + \sigma_0^2/n} - 2\log\varepsilon}\,. \tag{8}$$

See Pace & Salvan (2020, Example 1) for more details.

When the data are randomized into two streams, $y^{(2n)} = \left(y_0^{(n)}, y_1^{(n)}\right)$, with $y_0^{(n)} = (y_{01}, \ldots, y_{0n})$ and $y_1^{(n)} = (y_{11}, \ldots, y_{1n})$, the split-naive confidence sequence (6) is

$$\tilde{\Theta}_{1-\varepsilon}\left(y^{(2n)}\right) = \left\{ \theta \in \mathbb{R} \ : \ \frac{p_n\left(y_0^{(n)}; \hat{\theta}_1^{(n)}\right)}{p_n\left(y_0^{(n)}; \theta\right)} < \frac{1}{\varepsilon} \right\},$$

DOI: 10.1002/cjs.11749

where $p_n\left(y_0^{(n)};\theta\right) = c\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(y_{0i}-\theta)^2\right\}$ is the likelihood function from $y_0^{(n)}$, and $\hat{\theta}_1^{(n)} = \sum_{i=1}^n y_{1i}/n$ is the maximum likelihood estimate of θ from $y_1^{(n)}$. This sequence is given by closed-form intervals, namely

$$\hat{\theta}_0^{(n)} \pm \frac{\sigma_0}{\sqrt{n}}\sqrt{\frac{n\left\{\hat{\theta}_0^{(n)}-\hat{\theta}_1^{(n)}\right\}^2}{\sigma_0^2} - 2\log\varepsilon}\,, \tag{9}$$

which always contain $\hat{\theta}_0^{(n)} = \sum_{i=1}^n y_{0i}/n$, the maximum likelihood estimate of θ based on $y_0^{(n)}$. Intervals (9) may be formally obtained from intervals (8) by putting $\mu_0 = \hat{\theta}_1^{(n)}$ and considering the limit as $\tau_0^2 \to 0$. Even the split-exact confidence sequence (7) is given by closed-form intervals,

$$\hat{\theta}_{2:2n} \pm \frac{\sigma_0}{\sqrt{(2n-1)}}\sqrt{\frac{\sum_{i=2}^{2n}\left(y_i-\hat{\theta}_{i-1}\right)^2 - \sum_{i=2}^{2n}\left(y_i-\hat{\theta}_{2:2n}\right)^2}{\sigma_0^2} - 2\log\varepsilon}, \tag{10}$$

where $\hat{\theta}_{2:2n} = \sum_{j=2}^{2n} y_j/(2n-1)$ and $\hat{\theta}_{i-1} = \sum_{j=1}^{i-1} y_j/(i-1)$, for $i = 2,\ldots,2n$. Note that, unlike (8), intervals (9) and (10) are equivariant under location changes.

At any given sample size $2n$, mixture intervals depend on the data only through the minimal sufficient statistic, while split intervals, both naive and exact, do not. A small simulation experiment was conducted in order to compare the lengths of the three kinds of intervals. The persistence level selected is $1-\varepsilon = 0.80$. As observed in Pace & Salvan (2020, Example 2), the corresponding confidence sequence gives fixed-n intervals, for n in the range $[n_{\min}, n_{\max}]$ =[10, 4000], close to conventional confidence intervals with level 0.995. Various values of $2n$ from 20 to 1000 and several choices of the true θ in the range $[0, 2.5]$ were considered, with $\sigma_0^2 = 1$. The standard normal weight function $\pi(\theta) = e^{-\theta^2/2}/\sqrt{2\pi}$ was used for mixture intervals, such that $\mu_0 = 0$ and $\tau_0^2 = 1$ in (8). The results, based on 10,000 Monte Carlo replications, are displayed in Table 1. When the sample size is small or moderate, and the true θ is not far away from μ_0, mixture intervals are generally shorter. With $\theta > \mu_0 + \tau_0$, for larger sample sizes, the split intervals of the naive kind are the shortest, whereas the split intervals of the exact kind seem to be slightly inferior even when $2n$ =1000.

Next, we perform a simulation to investigate the conjecture that, for conveniently large sample sizes, the split-naive intervals have approximately persistence level $1-\varepsilon$. For 10,000 replications, sequences of samples with even size from $2n_{\min}$ to $2n_{\max}$ were generated with $n_{\min} = 100$ and $n_{\max} = 40{,}000$. The behaviour along the sequence of mixture intervals (8) with $\mu_0 = 0$ and $\tau_0^2 = 1$, split-naive intervals (9), and split-exact intervals (10) was observed, with the aim of detecting sequences that give incompatible conclusions (incompatibilities) and sequences that do not always cover the true parameter value (uncoverages), at various nominal persistence levels, $1-\varepsilon$. The results are displayed in Table 2. Analogous results are displayed in Table 3 for sequences of samples with even size from 2(400) to 2(80,000). The mixture and split-exact intervals look very conservative in the sense that the percentages of incompatibilities and uncoverages are both far below the bound $100\times\varepsilon$. The sequences of split-naive intervals with $1-\varepsilon = 0.80$ are anti-conservative. However, they improve their closeness to the nominal bound as $n_{\min}$ moves from 100 to 400.

Intervals (8) may provide a simple closed-form approximation for confidence sequences (4) for a scalar parameter θ when a normal weight function is used. Suppose that a normal approximation is available for the maximum likelihood estimator $\hat{\theta}_n$, that is, $\hat{\theta}_n \dot{\sim} N\left(\theta, v_n^2(\theta)\right) \dot{\sim} N\left(\theta, \hat{v}_n^2\right)$,

TABLE 1: Normal population with mean θ and variance 1: empirical averages of the length of intervals (8) with $\mu_0 = 0$ and $\tau_0^2 = 1$ (mixture), intervals (9) (split-naive), and intervals (10) (split-exact), all with nominal persistence level $1 - \varepsilon = 0.80$ in 10,000 samples with size $2n$, for various true values of θ.

$2n$	Method	$\theta = 0$	$\theta = 0.5$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$	$\theta = 2.5$
20	Mixture	1.123	1.144	1.204	1.299	1.421	1.564
	Split-exact	1.262	1.262	1.262	1.262	1.262	1.262
	Split-naive	1.407	1.407	1.407	1.407	1.407	1.407
40	Mixture	0.834	0.849	0.891	0.957	1.042	1.142
	Split-exact	0.921	0.921	0.921	0.921	0.921	0.921
	Split-naive	0.996	0.996	0.996	0.996	0.996	0.996
100	Mixture	0.560	0.569	0.595	0.635	0.687	0.749
	Split-exact	0.610	0.610	0.610	0.610	0.610	0.610
	Split-naive	0.631	0.631	0.631	0.631	0.631	0.631
200	Mixture	0.413	0.419	0.436	0.464	0.500	0.543
	Split-exact	0.447	0.447	0.447	0.447	0.447	0.447
	Split-naive	0.446	0.446	0.446	0.446	0.446	0.446
400	Mixture	0.304	0.308	0.320	0.339	0.363	0.393
	Split-exact	0.326	0.326	0.326	0.326	0.326	0.326
	Split-naive	0.315	0.315	0.315	0.315	0.315	0.315
1000	Mixture	0.201	0.204	0.211	0.222	0.238	0.256
	Split-exact	0.215	0.215	0.215	0.215	0.215	0.215
	Split-naive	0.200	0.200	0.200	0.200	0.200	0.200

Note: Due to equivariance, the results for split-exact and split-naive methods do not depend on θ.

with $\hat{v}_n^2$ an estimate of the asymptotic variance of $\hat{\theta}_n$ (that is, of $v_n^2(\theta) = \sigma^2(\theta)/n$). If a $N(\mu_0, \tau_0^2)$ density is used as a weight function where μ_0 is the conjectured central value for θ, then, in analogy with (8), a closed-form confidence sequence for θ is

$$\hat{\theta}_n \pm \hat{v}_n \sqrt{\log \frac{\tau_0^2 + \hat{v}_n^2}{\hat{v}_n^2} + \frac{\left(\hat{\theta}_n - \mu_0\right)^2}{\tau_0^2 + \hat{v}_n^2} - 2\log \varepsilon}, \tag{11}$$

for $n \geq n_{\min}$ with $n_{\min}$ sufficiently large and $\hat{v}_n = \sqrt{\hat{v}_n^2}$.

As Pace & Salvan (2020, Section 4) showed through simulation in some special models, for sequences starting from a moderate sample size $n_{\min}$, this proposal seems to maintain approximately the persistence level $1 - \varepsilon$ in the examples considered. Closed-form confidence intervals (11) have a Wald-type structure. Consequently, unlike intervals from a genuine likelihood, intervals (11) are not exactly equivariant under reparameterizations. On the other hand, intervals (11) rely only on the assumption that normal asymptotic theory offers a good approximation of the density of the estimator of θ. Therefore, any asymptotically normal estimator could be used in (11), like a robust estimator or the maximizer of a pseudo-likelihood (such as a composite likelihood).

TABLE 2: Normal population with mean θ and variance 1: empirical percentages of incompatibilities and uncoverages of intervals (8) with $\mu_0 = 0$ and $\tau_0^2 = 1$ (mixture), intervals (9) (split-naive), and intervals (10) (split-exact) at various nominal persistence levels $1 - \varepsilon$ in 10,000 sequences of samples with even size from 200 to 80,000. The true value of θ is zero.

		$100 \times \varepsilon$			
	Method	20	10	5	1
Incompatibilities	Mixture	0.99	0.43	0.20	0.04
	Split-exact	0.76	0.28	0.14	0.03
	Split-naive	7.82	2.09	0.54	0.03
Uncoverages	Mixture	3.22	1.79	0.94	0.21
	Split-exact	2.66	1.41	0.72	0.13
	Split-naive	23.66	9.15	3.35	0.37

TABLE 3: Normal population with mean θ and variance 1: empirical percentages of incompatibilities and uncoverages of intervals (8) with $\mu_0 = 0$ and $\tau_0^2 = 1$ (mixture), intervals (9) (split-naive), and intervals (10) (split-exact) at various nominal persistence levels $1 - \varepsilon$ in 10,000 sequences of samples with even size from 800 to 160,000. The true value of θ is zero.

		$100 \times \varepsilon$			
	Method	20	10	5	1
Incompatibilities	Mixture	0.40	0.14	0.06	0.01
	Split-exact	0.25	0.07	0.03	0.01
	Split-naive	5.77	1.50	0.41	0.02
Uncoverages	Mixture	1.74	1.02	0.44	0.08
	Split-exact	1.40	0.62	0.33	0.09
	Split-naive	21.67	8.30	3.36	0.28

3.2. Example 2. Poisson Population

For $i = 1, 2, \ldots$, let Y_i be a sequence of independent Poisson random variables with unknown mean $\lambda > 0$ and let y_i be their observations. The likelihood function based on $y^{(2n)}$ is

$$p_{2n}\left(y^{(2n)}; \lambda\right) = c\ e^{-2n\lambda} \lambda^{2n\hat{\lambda}^{(2n)}},$$

where $\hat{\lambda}^{(2n)} = \sum_{i=1}^{2n} y_i/(2n)$.

Using for $\lambda > 0$ the weight function $\pi(\lambda) = b^a \lambda^{a-1} e^{-b\lambda}/\Gamma(a)$ where $a, b > 0$, corresponding to a Gamma(a, b) distribution, Robbins' mixture is

$$q_{2n}\left(y^{(2n)}\right) = \int_0^{+\infty} p_{2n}\left(y^{(2n)}; \lambda\right)\pi(\lambda)\ d\lambda = c\frac{b^a\Gamma\left(a + 2n\hat{\lambda}^{(2n)}\right)}{(b + 2n)^{a+2n\hat{\lambda}^{(2n)}}\Gamma(a)},$$

so that the mixture confidence sequence with persistence level $1-\varepsilon$ is

$$\hat{\Theta}_{1-\varepsilon}\left(y^{(2n)}\right)=\left\{\lambda>0\ :\ \frac{q_{2n}\left(y^{(2n)}\right)}{p_{2n}\left(y^{(2n)};\lambda\right)}<\frac{1}{\varepsilon}\right\};$$

that is

$$\hat{\Theta}_{1-\varepsilon}\left(y^{(2n)}\right)=\left\{\lambda>0\ :\ \frac{b^a\Gamma\left(a+2n\hat{\lambda}^{(2n)}\right)e^{2n\lambda}}{(b+2n)^{a+2n\hat{\lambda}^{(2n)}}\Gamma(a)\lambda^{2n\hat{\lambda}^{(2n)}}}<\frac{1}{\varepsilon}\right\}. \tag{12}$$

Because $g(\lambda)=2n\lambda-2n\hat{\lambda}^{(2n)}\log\lambda$ is convex when $\lambda>0$, $\exp(g(\lambda))$ is convex as well, and therefore $\hat{\Theta}_{1-\varepsilon}\left(y^{(2n)}\right)$ is an interval containing $\hat{\lambda}^{(2n)}$. A closed-form approximation for (12) based on (11) can also be obtained, for instance from the approximation $\log\hat{\lambda}^{(2n)}\dot{\sim}N(\log\lambda,1/(2n\lambda))$ and using a normal weight function for $\log\lambda$ with mean μ_0 and variance τ_0^2 matching the mean and variance of $\log\lambda$ when $\lambda\sim\text{Gamma}(a,b)$.

When $y^{(2n)}=\left(y_0^{(n)},y_1^{(n)}\right)$, with $y_0^{(n)}=(y_{01},\ldots,y_{0n})$ and $y_1^{(n)}=(y_{11},\ldots,y_{1n})$, the likelihood function based on $y_0^{(n)}$ is given by $p_n\left(y_0^{(n)};\lambda\right)=c\ e^{-n\lambda}\lambda^{n\hat{\lambda}_0^{(n)}}$, where $n\hat{\lambda}_0^{(n)}=\sum_{i=1}^n y_{0i}$. On the other hand, the estimate $\hat{\lambda}_1^{(n)}$ is the maximum likelihood estimate of λ from $y_1^{(n)}$, $\hat{\lambda}_1^{(n)}=\sum_{i=1}^n y_{1i}/n$, if $\sum_{i=1}^n y_{1i}>0$, and otherwise $\hat{\lambda}_1^{(n)}=0.5/n$. The split-naive confidence sequence (6) is then

$$\tilde{\Theta}_{1-\varepsilon}(y^{(2n)})=\left\{\lambda>0\ :\ \frac{p_n\left(y_0^{(n)};\hat{\lambda}_1^{(n)}\right)}{p_n\left(y_0^{(n)};\lambda\right)}<\frac{1}{\varepsilon}\right\},$$

that is,

$$\tilde{\Theta}_{1-\varepsilon}\left(y^{(2n)}\right)=\left\{\lambda>0\ :\ \exp\left(-n\left(\hat{\lambda}_1^{(n)}-\lambda\right)+n\hat{\lambda}_0^{(n)}\log\frac{\hat{\lambda}_1^{(n)}}{\lambda}\right)<\frac{1}{\varepsilon}\right\}. \tag{13}$$

Again, $\tilde{\Theta}_{1-\varepsilon}\left(y^{(2n)}\right)$ is an interval which always contains $\hat{\lambda}_0^{(n)}$.

The split-exact confidence sequence (7) is given by

$$\overline{\Theta}_{1-\varepsilon}(y^{(2n)})=\left\{\lambda>0:\ \exp\left(\sum_{i=2}^{2n}\hat{\lambda}_{i-1}-(2n-1)\lambda-\sum_{i=2}^{2n}y_i\log\frac{\hat{\lambda}_{i-1}}{\lambda}\right)>\varepsilon\right\}, \tag{14}$$

where $\hat{\lambda}_{i-1}=\sum_{j=1}^{i-1}y_j/(i-1)$ if $\sum_{j=1}^{i-1}y_j>0$ and $\hat{\lambda}_{i-1}=0.5/(i-1)$ otherwise.

At any given sample size $2n$, mixture intervals depend on the data only through the minimal sufficient statistic. Split intervals do not. A simulation experiment was conducted in order to compare the lengths of the four kinds of intervals. The persistence level selected is $1-\varepsilon=0.80$ and various values of $2n$ from 20 to 1000 and several choices of the true λ in the range $[0.3,3]$ were considered. The negative exponential weight function $\pi(\lambda)=e^{-\lambda}$ was used for mixture intervals, such that $a=b=1$ in (12), and $\mu_0=\Psi(1)=-\gamma$ and $\tau_0^2=\Psi'(1)=\pi^2/6$ in the approximate form (11) for $\log\lambda$, where $\Psi(\cdot)$ and $\Psi'(\cdot)$ denote the digamma and trigamma

TABLE 4: Poisson population with mean λ: empirical averages of the length of intervals (12) with $a = b = 1$ (mixture), (11) for $\log \lambda$ with $\mu_0 = \Psi(1)$ and $\tau_0^2 = \Psi'(1)$ (mixture appr.), (13) (split-naive) and (14) (split-exact) with persistence level $1 - \varepsilon = 0.80$ in 5000 samples with size $2n$ and various true values of λ

$2n$	Method	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 2.0$	$\lambda = 3.0$
20	Mixture	0.610	0.777	1.130	1.751	2.352
	Mixture appr.	0.697	0.860	1.242	1.891	2.448
	Split-exact	0.632	0.841	1.237	1.802	2.225
	Split-naive	0.790	1.020	1.413	1.995	2.443
40	Mixture	0.447	0.575	0.833	1.285	1.722
	Mixture appr.	0.479	0.611	0.897	1.370	1.772
	Split-exact	0.461	0.612	0.900	1.310	1.617
	Split-naive	0.549	0.710	0.995	1.409	1.725
100	Mixture	0.297	0.383	0.556	0.853	1.134
	Mixture appr.	0.309	0.400	0.592	0.901	1.162
	Split-exact	0.305	0.404	0.594	0.863	1.066
	Split-naive	0.336	0.439	0.622	0.883	1.083
200	Mixture	0.216	0.280	0.407	0.623	0.824
	Mixture appr.	0.224	0.293	0.432	0.657	0.845
	Split-exact	0.221	0.294	0.432	0.629	0.777
	Split-naive	0.233	0.306	0.434	0.619	0.761
400	Mixture	0.156	0.203	0.296	0.452	0.597
	Mixture appr.	0.164	0.214	0.316	0.478	0.614
	Split-exact	0.160	0.213	0.313	0.456	0.564
	Split-naive	0.163	0.213	0.305	0.436	0.536
1000	Mixture	0.100	0.131	0.193	0.294	0.388
	Mixture appr.	0.108	0.141	0.208	0.314	0.402
	Split-exact	0.102	0.137	0.203	0.297	0.368
	Split-naive	0.099	0.131	0.189	0.271	0.334

functions. With this choice of the weight function none of the selected values of λ is extreme. The results, based on 5000 Monte Carlo replications, are displayed in Table 4. When the sample size is small or moderate, mixture intervals (12) are generally preferred. However, for $2n = 1000$, split-naive intervals (13) are the shortest, although very close to the mixture intervals. Split-exact intervals (14) are, on average, shorter than split-naive intervals when $2n$ is small. The average length of approximate mixture intervals (11) is close to that of split-exact intervals.

As in Example 1, for 5000 replications, sequences of samples with even size from $2n_{\min}$ to $2n_{\max}$ were generated with $n_{\min} = 100$ and $n_{\max} = 10{,}000$. The behaviour along the sequence of intervals (12) with $a = b = 1$ (mixture), (11) for $\log \lambda$ with $\mu_0 = \Psi(1)$ and $\tau_0^2 = \Psi'(1)$ (mixture appr.), (13) (split-naive) and (14) (split-exact) has been observed, with the aim of evaluating incompatibilities and uncoverages at various nominal persistence levels $1 - \varepsilon$. The results are

TABLE 5: Poisson population with mean λ: empirical percentages of incompatibilities and uncoverages of intervals (12) with $a = b = 1$ (mixture), (11) for $\log \lambda$ with $\mu_0 = \Psi(1)$ and $\tau_0^2 = \Psi'(1)$ (mixture appr.), (13) (split-naive), and (14) (split-exact) at various nominal persistence levels $1 - \varepsilon$ in 5000 sequences of samples with even size from 100 to 10,000. The true value of λ is one.

		$100 \times \varepsilon$			
	Method	20	10	5	1
Incompatibilities	Mixture	0.72	0.32	0.08	0.00
	Mixture appr.	0.46	0.08	0.06	0.00
	Split-exact	0.62	0.26	0.08	0.04
	Split-naive	4.34	0.90	0.22	0.00
Uncoverages	Mixture	9.20	5.44	3.34	1.00
	Mixture appr.	2.42	1.54	0.78	0.16
	Split-exact	7.88	4.54	2.76	0.88
	Split-naive	43.00	20.34	8.76	1.34

displayed in Table 5. The mixture, approximate mixture, and split-exact intervals look very conservative. On the other hand, the split-naive intervals are always anti-conservative.

4. CONFIDENCE SEQUENCES ON A PARAMETER OF INTEREST

When $p > 1$ and the parameter is partitioned as $\theta = (\psi, \lambda)$, where $\psi \in \Psi$ is a p_0-dimensional component of interest and λ is nuisance, in special models, safe inference about ψ can be based on a statistic $t^{(n)} = t(y^{(n)})$ inducing a marginal or conditional model free of λ. In full generality, anytime-valid inference about ψ may be obtained from the profile likelihood using the mixture or the data splitting device.

The projection on the subspace Ψ of a confidence sequence $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$ for θ with persistence level $1 - \varepsilon$,

$$\hat{\Psi}_{1-\varepsilon}(y^{(n)}) = \left\{ \psi \in \Psi \ : \ (\psi, \lambda) \in \hat{\Theta}_{1-\varepsilon}\left(y^{(n)}\right) \ \text{ for some } \ \lambda \right\}, \tag{15}$$

is clearly a confidence sequence for ψ with persistence level $1 - \varepsilon$. The confidence sequence (15) turns out to be based on the profile likelihood. For instance, with mixture confidence sequences we obtain

$$\hat{\Psi}_{1-\varepsilon}\left(y^{(n)}\right) = \left\{ \psi \in \Psi \ : \ p_n\left(y^{(n)}; \psi, \hat{\lambda}_\psi\right) > \varepsilon q_n\left(y^{(n)}\right) \right\}, \tag{16}$$

where $\hat{\lambda}_\psi$ is the maximum likelihood estimate of λ in the model for $y^{(n)}$ with ψ fixed and

$$q_n(y^{(n)}) = \int_\Theta p_n\left(y^{(n)}; \theta\right) \pi(\theta) \, d\theta.$$

If a normal approximation is available for the maximum likelihood estimator $\hat{\psi}_n$ of a scalar ψ, that is, $\hat{\psi}_n \sim N\left(\psi, v_n^2(\theta)\right) \dot{\sim} N\left(\psi, \hat{v}_n^2\right)$, with $\hat{v}_n^2$ an estimate of the asymptotic variance of $\hat{\psi}_n$, and an $N\left(\psi_0, \tau_0^2\right)$ density is used as a weight function, where ψ_0 is the conjectured central value for ψ, a closed-form confidence sequence for ψ, analogous to sequence (11) is

$$\hat{\psi}_n \pm \hat{v}_n \sqrt{\log \frac{\tau_0^2 + \hat{v}_n^2}{\hat{v}_n^2} + \frac{(\hat{\psi}_n - \psi_0)^2}{\tau_0^2 + \hat{v}_n^2} - 2 \log \varepsilon}. \tag{17}$$

Like sequence (11), closed-form intervals (17) have a Wald-type structure and consequently they are not exactly equivariant under interest-respecting reparameterizations, that is, reparameterizations $\omega = \omega(\theta) = (\phi, \chi)$ where $\phi = \phi(\psi)$ and $\chi = \chi(\psi, \lambda)$.

Also, the splitting device may be applied to the profile likelihood, as we see substituting $\hat{\Theta}_{1-\varepsilon}\left(y^{(n)}\right)$ in (15) with the split-exact set $\overline{\Theta}_{1-\varepsilon}\left(y^{(2n)}\right)$ or even the split-naive set $\tilde{\Theta}_{1-\varepsilon}\left(y^{(2n)}\right)$, obtaining the confidence sequences $\overline{\Psi}_{1-\varepsilon}\left(y^{(n)}\right)$ and $\tilde{\Psi}_{1-\varepsilon}\left(y^{(n)}\right)$, respectively. See Wasserman, Ramdas & Balakrishnan (2020, Section 5) for details on $\overline{\Psi}_{1-\varepsilon}\left(y^{(n)}\right)$. Admittedly, the persistence level of $\tilde{\Psi}_{1-\varepsilon}\left(y^{(2n)}\right)$ will be $1-\varepsilon$ asymptotically at best. However, confidence sequences $\hat{\Psi}_{1-\varepsilon}\left(y^{(n)}\right)$ are likely to be much more conservative than confidence sequences $\hat{\Theta}_{1-\varepsilon}\left(y^{(n)}\right)$. The same remark holds for the analogous confidence sequences from the splitting device, $\overline{\Psi}_{1-\varepsilon}\left(y^{(n)}\right)$ and $\tilde{\Psi}_{1-\varepsilon}\left(y^{(n)}\right)$. Thus, conservativeness should ease the task of respecting the nominal persistence level for $\tilde{\Psi}_{1-\varepsilon}\left(y^{(n)}\right)$.

Unfortunately, the computational burden imposed by definition (16) may be very heavy. In this respect, things are a little bit better for the confidence sequence $\tilde{\Psi}_{1-\varepsilon}\left(y^{(n)}\right)$. In regular models, asymptotic sufficiency and asymptotic normality of the maximum likelihood estimator of θ offer a comfortable way out of the predicament.

5. CONFIDENCE SEQUENCES FROM COMPOSITE LIKELIHOODS: NO NUISANCE PARAMETERS

In this section, after a brief review of composite likelihood theory, we focus on confidence sequences from composite likelihoods when the parameter of the model is scalar and the maximizer of the composite likelihood is asymptotically normal. The more general case of a scalar parameter of interest in the presence of nuisance parameters will be treated in the next section.

In complex models, complexity often entails that the full likelihood $L(\theta) = p_Y(y;\theta)$ is computationally intractable, or even difficult to specify. It is then convenient to trade a certain loss of efficiency for material relief of the computational burden. In their simplest form, composite likelihoods are pseudo-likelihoods for θ composed by multiplying elemental contributions $L_j(\theta)$, $j = 1, \ldots, q$. Contributions $L_j(\theta)$ are in turn genuine likelihoods for θ based on low-dimensional but dependent parts of the data. Dependence occurs when Y has dependent components or when the same block of the data appears in more than one of the factors $L_j(\theta)$. For spatial data, Besag (1974) proposed to use elemental likelihoods $L_j(\theta)$ from conditional densities. More generally, likelihoods $L_j(\theta)$ from conditional or marginal densities were considered in Lindsay (1988). See Varin, Reid & Firth (2011) for a comprehensive review of composite likelihoods.

A composite log likelihood has the general form

$$\ell_{CP}(\theta) = \sum_{j=1}^{q} w_j \ell_j(\theta), \tag{18}$$

where $\ell_j(\theta) = \log L_j(\theta)$ and w_j are convenient weights, often chosen all equal to 1 (Sang & Genton, 2014). When the contributions $\ell_j(\theta)$ are independent random variables, unitary weights are optimal and make $\ell_{CP}(\theta)$ a genuine log likelihood. Under regularity conditions, the score function and the Fisher information matrix of $\ell_j(\theta)$ are $u_j(\theta) = \partial/(\partial\theta)\ell_j(\theta)$ and $I_{jj}(\theta) = Var_\theta\left\{u_j(\theta)\right\} = E_\theta\left\{-\partial/\left(\partial\theta^\top\right)u_j(\theta)\right\}$. The score function from the composite likelihood $\ell_{CP}(\theta)$ is then

$$u_{CP}(\theta) = \partial/(\partial\theta)\ell_{CP}(\theta) = \sum_{j=1}^{q} w_j u_j(\theta).$$

The estimating equation $u_{CP}(\theta) = 0$ is unbiased, meaning that $E_\theta\big(u_{CP}(\theta)\big) = 0$. Under the usual regularity conditions, the resulting estimator $\tilde{\theta}_{CP}$ is asymptotically normal with mean θ and variance equal to the inverse of the Godambe information matrix, that is,

$$\tilde{\theta}_{CP} \dot{\sim} N_p\big(\theta, G_{CP}^{-1}(\theta)\big). \tag{19}$$

The Godambe information matrix is the $p \times p$ matrix

$$G_{CP}(\theta) = H_{CP}(\theta) J_{CP}^{-1}(\theta) H_{CP}(\theta),$$

where $J_{CP}(\theta)$ and $H_{CP}(\theta)$ are, respectively, the variability and sensitivity matrices of the estimating function $u_{CP}(\theta)$:

$$J_{CP}(\theta) = Var_\theta\left\{u_{CP}(\theta)\right\}$$

and

$$H_{CP}(\theta) = E_\theta\left\{-\partial/(\partial\theta^\top) u_{CP}(\theta)\right\}.$$

Both matrices $J_{CP}(\theta)$ and $H_{CP}(\theta)$ are symmetric and are assumed to be invertible. When $J_{CP}(\theta) = H_{CP}(\theta)$, the Godambe information simplifies and the composite likelihood behaves more like a genuine likelihood. The estimating equation $u_{CP}(\theta) = 0$ is then called information unbiased (Lindsay, 1982). As an illustrative example, we consider below inference about a scalar θ, with true value θ^*, using a composite likelihood and even the optimal composite likelihood.

Concatenating, for $j = 1, \ldots, q$, the quantities $\ell_j(\theta)$, $u_j(\theta)$, and $I_{jj}(\theta) = Var_\theta\left\{u_j(\theta)\right\}$, we obtain three $q \times 1$ vectors, $\ell_V(\theta) = [l_j(\theta)]$, $u_V(\theta) = [u_j(\theta)]$, and $i_V(\theta) = [I_{jj}(\theta)]$. Consider also the $q \times q$ matrix $\Sigma_V(\theta) = [I_{jk}(\theta)]$, where $I_{jk}(\theta) = Cov_\theta(u_j(\theta), u_k(\theta))$, $j, k = 1, \ldots, q$. The vector $i_V(\theta)$ has the same entries as the main diagonal of $\Sigma_V(\theta)$, and we will write in short $i_V(\theta) = \text{diag}\left(\Sigma_V(\theta)\right)$.

The composite log likelihood (18) may be written as

$$\ell_{CP}(\theta) = w^\top \ell_V(\theta),$$

where $w = [w_i]$ are weights that may depend on θ^*. The score function from $\ell_{CP}(\theta)$ is

$$u_{CP}(\theta) = w^\top u_V(\theta).$$

Differentiating $E_\theta(u_{CP}(\theta)) = 0$ with respect to θ gives

$$Cov_\theta\big(u_{CP}(\theta), u(\theta)\big) = \sum_{j=1}^{q} w_j I_{jj}(\theta) = w^\top i_V(\theta),$$

where $u(\theta) = \partial/(\partial\theta) \log L(\theta)$ is the score from the full likelihood. Variability and sensitivity of $u_{CP}(\theta)$ are $J_{CP}(\theta) = w^\top \Sigma_V(\theta) w$ and $H_{CP}(\theta) = w^\top i_V(\theta)$, whence we see that to have meaningful weights we have to restrict w so that that $H_{CP}(\theta) > 0$.

The Godambe information of the estimating function $u_{CP}(\theta)$ is

$$G_{CP}(\theta) = \frac{\{w^\top i_V(\theta)\}^2}{w^\top \Sigma_V(\theta) w},$$

such that, under θ,

$$\tilde{\theta}_{CP} \dot{\sim} N(\theta, w^\top \Sigma_V(\theta) w / \{w^\top i_V(\theta)\}^2). \tag{20}$$

In the asymptotic variance, θ may be substituted by a consistent estimator, such as $\tilde{\theta}_I$, the maximizer of the so-called independence likelihood corresponding to the widely used pseudo-log likelihood $\tilde{\ell}_I(\theta) = \sum_{j=1}^{q} \ell_j(\theta) = 1_q^\top \ell_V(\theta)$, where $1_q^\top = (1, \ldots, 1)$.

Following Fraser & Reid (2020) and Pace, Salvan & Sartori (2019), the asymptotically most efficient estimator from a composite log likelihood of the form (18), $\tilde{\theta}^*$, is obtained when the weights are

$$w^* = w(\theta^*) = \Sigma_V(\theta^*)^{-1} i_V(\theta^*);$$

that is, w^* are the regression coefficients of the multiple linear regression of $u(\theta^*)$ on $u_V(\theta^*)$, because $i_V(\theta^*) = Cov_{\theta^*}(u_V(\theta^*), u(\theta^*))$. To use the optimal composite log likelihood in practice, the unknown θ^* has, of course, to be replaced by a consistent estimator such as $\tilde{\theta}_I$.

When the maximizer of the composite likelihood is asymptotically normal and we use a $N(\theta_0, \tau_0^2)$ density as a weight function for θ, a confidence sequence consisting of closed-form intervals is obtained. From approximation (20) we get an interval of the form (11), where $\hat{\theta}_n = \tilde{\theta}_n$ and $\hat{v}_n = \tilde{v} = \sqrt{w^\top \Sigma_V(\tilde{\theta}_I) w} / \{w^\top i_V(\tilde{\theta}_I)\}$ for a given vector of weights w. Refining the above formula, a sequence obtained from the composite likelihood that uses the optimal weights w^* has v^* instead of $\tilde{v}$, with $v^* = 1/\sqrt{i_V(\tilde{\theta}_I)^\top \Sigma_V(\tilde{\theta}_I)^{-1} i_V(\tilde{\theta}_I)}$ or $v^* = 1/\sqrt{i_V(\tilde{\theta}^*)^\top \Sigma_V(\tilde{\theta}^*)^{-1} i_V(\tilde{\theta}^*)}$.

6. CONFIDENCE SEQUENCES FROM COMPOSITE LIKELIHOODS WITH NUISANCE PARAMETERS: AN EXAMPLE

When the p-dimensional parameter of the model is $\theta = (\psi, \lambda)$, where ψ is a scalar parameter of interest and λ is a nuisance parameter, suppose that a composite log likelihood $\ell_{CP}(\theta)$ provides the estimate $\tilde{\theta}^\top = (\tilde{\psi}, \tilde{\lambda}^\top)$. If approximation (19) holds, the asymptotic sampling distribution of $\tilde{\psi}$ under θ is

$$\tilde{\psi} \dot{\sim} N\left(\psi, v^2(\tilde{\theta})\right),$$

where $v^2(\theta) = G_{CP}^{-1}(\theta)_{11}$ is the entry of the inverse of the Godambe information matrix at the first row and first column. Obtaining optimal weights for profile inference about ψ is not straightforward: see Pace, Salvan & Sartori (2019, Section 3). Here, we suppose therefore that the composite likelihood is defined using a given vector of weights w, in general, unrelated to optimality.

With the weight function $\pi(\psi)$ corresponding to the $N\left(\psi_0, \tau_0^2\right)$ density, if we put $\tilde{v} = \sqrt{v^2(\tilde{\theta})}$, a confidence sequence for ψ is then given by formula (17) with $\hat{\psi}_n$ replaced by $\tilde{\psi}$ and $\hat{v}_n^2$ replaced by $\tilde{v}^2$. We conjecture that this method provides a confidence sequence with asymptotic persistence level $1 - \epsilon$. To support this conjecture, a simulation study has been performed, considering the following example.

6.1. Example 3. Symmetric Multivariate Normal Population, Scalar μ of Interest

Inspired by an example in Section 1 of Cox & Reid (2004), let $Y_i^\top = (Y_{i1}, \ldots, Y_{iq})$, for $i = 1, \ldots, n$, be i.i.d. with

$$Y_1 \sim N_q\left(\mu 1_q, \sigma^2(1-\rho) I_q + \sigma^2 \rho 1_q 1_q^\top\right),$$

where $1_q^\top = (1, \ldots, 1)$ is the vector in $\mathbb{R}^q$ having all components 1, I_q denotes the identity matrix of order q, and $\mu \in \mathbb{R}$ is unknown, while parameters $\sigma^2 > 0$ and $\rho \geq 0$ are provisionally supposed to be known.

The process generating the data y_{ij}, $i = 1, \dots, n$, $j = 1, \dots, q$, can be analyzed as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad i = 1, \dots, n, \quad j = 1, \dots, q,$$

where α_i, $i = 1, \dots, n$, and ϵ_{ij}, $i = 1, \dots, n$, $j = 1, \dots, q$, are independent random variables having marginal distributions $\alpha_i \sim N\left(0, \sigma_\alpha^2\right)$ and $\epsilon_{ij} \sim N\left(0, \sigma_\epsilon^2\right)$. Then

$$\sigma^2 = \sigma_\alpha^2 + \sigma_\epsilon^2 , \qquad \rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}.$$

A collection of q genuine log likelihoods for μ provided by the independent marginals $Y_{1j}, \dots, Y_{nj}$ is given by

$$\ell_j(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{ij} - \mu)^2, \qquad j = 1, \dots, q,$$

with corresponding score functions

$$u_j(\mu) = \sum_{i=1}^{n}(y_{ij} - \mu)/\sigma^2, \qquad j = 1, \dots, q.$$

The estimating equation that is most efficient among the scores from the combined log likelihoods of the form (18) is obtained with $w^* = 1_q$: that is, from

$$\ell_I(\mu) = \sum_{j=1}^{q}\ell_j(\mu) = -\frac{1}{2\sigma^2}\sum_{j=1}^{q}\sum_{i=1}^{n}(y_{ij} - \mu)^2 \tag{21}$$

with corresponding estimating function

$$u_I(\mu) = \sum_{j=1}^{q}u_j(\mu) = \frac{1}{\sigma^2}\sum_{j=1}^{q}\sum_{i=1}^{n}(y_{ij} - \mu).$$

Indeed, the maximizer of $\ell_I(\mu)$ is $\overline{y} = \sum_{i=1}^{n}\sum_{j=1}^{q} y_{ij}/(nq)$, which is also the maximizer of the full likelihood.

We have $I_{jj}(\mu) = Var(u_j(\mu)) = n/\sigma^2$, and therefore

$$i_V(\mu) = \frac{n}{\sigma^2}1_q,$$

and from $I_{jk}(\mu) = Cov(u_j(\mu), u_k(\mu)) = n\rho/\sigma^2$ when $j \neq k$ we obtain

$$\Sigma_V(\mu) = \frac{n}{\sigma^2}\left\{(1-\rho)I_q + \rho 1_q 1_q^\top\right\}.$$

It is easily checked from the above expressions that

$$\left[\Sigma_V(\mu)\right]^{-1} i_V(\mu) = 1_q,$$

that is, $w^* = 1_q$.

The Godambe information of the estimating function $u_I(\mu)$ is

$$G_I(\mu) = \frac{\{u^\top i_V(\mu)\}^2}{u^\top \Sigma_V(\mu) u} = \frac{n}{\sigma^2} \frac{q}{1+(q-1)\rho} .$$

It follows that

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n} \frac{1+(q-1)\rho}{q}\right) .$$

The sampling distribution of $\overline{Y}$ is exact and not merely asymptotic. Therefore, when ρ and σ^2 are known, using the weight function $\pi(\mu)$ corresponding to the $N(\mu_0, \tau_0^2)$ density, the mixture confidence sequence for μ with exact persistence level $1-\varepsilon$ is

$$\overline{y} \pm v \sqrt{\log \frac{\tau_0^2 + v^2}{v^2} + \frac{(\overline{y}-\mu_0)^2}{\tau_0^2+v^2} - 2\log\varepsilon} , \tag{22}$$

where $v = \sqrt{\sigma^2\{1+(q-1)\rho\}/(qn)}$. The sequence is obtained by increasing n for a fixed q.

The parameter μ is orthogonal to the block of parameters (σ^2, ρ). This entails that $\ell_I(\mu)$, cf. (21), with σ^2 and ρ substituted by a consistent estimate, has the same asymptotic properties as a profile likelihood. In particular the optimality of the estimating function $u_I(\mu)$ with estimated σ^2 and ρ is preserved. Moment estimates of σ^2 and ρ are based on the sum of squares statistics

$$SS_E = \sum_{i=1}^{n}\sum_{j=1}^{q} (y_{ij} - \overline{y}_i)^2 , \qquad SS_B = q\sum_{i=1}^{n} (\overline{y}_i - \overline{y})^2 ,$$

where $\overline{y}_i = \sum_{j=1}^{q} y_{ij}/q$. They have the following expressions (see Searle, Casella & McCulloch, 1992, Section 3.5):

$$\tilde{\sigma}^2 = \frac{SS_B}{(n-1)q} + \frac{SS_E}{nq} , \qquad \tilde{\rho} = \max\left(\frac{SS_B/(n-1) - SS_E/(n(q-1))}{SS_E/n + SS_B/(n-1)}, 0\right) .$$

Inserting these estimates in Equation (22), we obtain the confidence sequence for μ with (σ^2, ρ) nuisance:

$$\overline{y} \pm \tilde{v} \sqrt{\log \frac{\tau_0^2 + \tilde{v}^2}{\tilde{v}^2} + \frac{(\overline{y}-\mu_0)^2}{\tau_0^2+\tilde{v}^2} - 2\log\varepsilon} , \tag{23}$$

where $\tilde{v} = \sqrt{\tilde{\sigma}^2\{1+(q-1)\tilde{\rho}\}/(qn)}$.

Finally, a simulation to investigate the conjecture that, for conveniently large sample sizes, the confidence sequence (23) has approximately persistence level $1-\varepsilon$ is in order. For 10,000 replications, sequences of samples with n from $n_{\min}$ to $n_{\max}$ have been generated with $n_{\min} = 25$ and $n_{\max} = 10{,}000$. The behaviour along the sequence of intervals (22) and (23) has been observed, with the aim of detecting incompatibilities and uncoverages, at various nominal persistence levels $1-\varepsilon$ and $q = 5$. The results are displayed in Table 6. The exact intervals look very conservative. The intervals with estimated v show some more incompatibilities and uncoverages.

TABLE 6: Symmetric multinormal population with mean $\mu 1_q$, variances $\sigma^2 1_q$, equicorrelation ρ: empirical percentages of incompatibilities and uncoverages of mixture intervals (22) (exact) and (23) (estimated) with $\mu_0 = 0$ and $\tau_0^2 = 1$, at various nominal persistence levels $1 - \varepsilon$, in 10,000 sequences of samples from $n_{\min} = 25$ to $n_{\max} = 10{,}000$. The true values of μ and σ^2 are zero and one, respectively, $q = 5$.

			$100 \times \varepsilon$			
ρ	Property	Method	20	10	5	1
0.3	Incompatibilities	Exact	1.85	0.84	0.34	0.06
		Estimated	3.42	1.95	1.14	0.25
	Uncoverages	Exact	5.23	2.87	1.53	0.27
		Estimated	7.57	4.66	2.98	0.95
0.5	Incompatibilities	Exact	2.20	0.94	0.41	0.08
		Estimated	3.80	2.14	1.29	0.27
	Uncoverages	Exact	5.94	3.24	1.68	0.27
		Estimated	8.45	5.12	3.24	1.02
0.8	Incompatibilities	Exact	2.50	1.16	0.47	0.09
		Estimated	4.28	2.46	1.41	0.30
	Uncoverages	Exact	6.84	3.68	1.93	0.35
		Estimated	9.54	5.73	3.52	1.12

7. CONCLUSIONS

In this article, we have dealt with a concept of replicability according to which, under the assumed statistical model, the current region and regions from arbitrarily enlarged samples have a large enough probability of overlapping. The definition of the persistence level $1 - \varepsilon$ makes the idea precise. We have emphasized, as a means to reach this end, the mixture confidence sequences described in Robbins (1970). An advantage of mixture confidence sequences is their justification under various views of inference, as stressed in Pace & Salvan (2020). The price to pay for controlling for the probability of sequence-wise overlapping of confidence regions is that wider regions are needed in comparison with the usual confidence regions with the same confidence level. These results are exact and refer to using the full likelihood as the grounds of parametric inference. Using an asymptotic normal approximation for the estimator of a scalar parameter of interest, approximate closed-form confidence sequences are easily calculated. This article has explored such a shortcut to extend the application of approximate confidence sequences to composite likelihoods. Simulation results support the conjecture that confidence sequences obtained in this way, with nominal persistence level $1 - \varepsilon$, have guaranteed sequence-wise compatibility.

Although all the examples in the article consider a scalar parameter of interest, the general construction based on (2) applies naturally to a multiparameter setting. Moreover, the approximate expression based on asymptotic normality of the estimator of the parameter extends easily to a vector parameter of interest using a Wald-type statistic.

Directions of future investigation include the efficient computation of confidence sequences when the estimator has no closed-form expression and is the solution of an estimating equation. Focusing on estimating equations might also overcome the lack of parameterization equivariance of the approximate solution considered here.

ACKNOWLEDGEMENT

It is a great pleasure to contribute to this special issue in honour of Nancy Reid, whose deep and far-reaching contributions to statistical theory have been highly influential for our work. We gratefully acknowledge an Associate Editor and a Reviewer for their useful suggestions. We also thank the organizers and participants of the workshop "Safe, Anytime-Valid Inference (SAVI) and Game-theoretic Statistics", Eindhoven (NL), 2022, for enlightening contributions and discussions on confidence sequences. Nicola Sartori's work was supported by a grant from the University of Padova (BIRD203991). Open Access Funding was provided by Università degli Studi di Padova within the CRUI-CARE Agreement.

REFERENCES

Benjamini, Y. (2020). Selective inference: The silent killer of replicability. *Harvard Data Science Review*, 2(4). https://doi.org/10.1162/99608f92.fc62b261

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B*, 36, 192–236.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62, 441–444.

Cox, D. R. & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91, 729–737.

Csenki, A. (1979). A note on confidence sequences in multiparameter exponential families. *Journal of Multivariate Analysis*, 9, 337–340.

Darling, D. A. & Robbins, H. (1967a). Iterated logarithm inequalities. *Proceedings of the National Academy of Sciences USA*, 57, 1188–1192.

Darling, D. A. & Robbins, H. (1967b). Confidence sequences for mean, variance and median. *Proceedings of the National Academy of Sciences USA*, 58, 66–68.

Fraser, D. A. S. & Reid, N. (2020). Combining likelihoods and significance functions. *Statistica Sinica*, 30, 1–15.

Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49, 1055–1080.

Howard, S. R. & Ramdas, A. (2022). Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28, 1704–1728.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, e124.

Jacod, J. & Protter, P. (2000). *Probability Essentials*, Springer, Berlin.

Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2021). Always valid inference: Continuous monitoring of A/B tests. *Operations Research*, 70, 1806–1821.

Lai, T. L. (1976). On confidence sequences. *Annals of Statistics*, 4, 265–280.

Lindsay, B. (1982). Conditional score functions: Some optimality results. *Biometrika*, 69, 503–512.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80, 221–240.

Nguyen, H. D. (2020). Universal inference with composite likelihoods. arXiv preprint, arXiv:2009.00848v4.

Pace, L., Salvan, A., & Sartori, N. (2019). Efficient composite likelihood for a scalar parameter of interest. *Stat*, 8, e222.

Pace, L. & Salvan, A. (2020). Likelihood, replicability and Robbins' confidence sequences. *International Statistical Review*, 88, 599–615.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41, 1397–1409.

Robbins, H. & Siegmund, D. (1972). A class of stopping rules for testing parametric hypotheses. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, University of California, Vol. 4, 37–41.

Robbins, H. & Siegmund, D. (1974). The expected sample size of some tests of power one. *Annals of Statistics*, 2, 415–436.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*, Chapman and Hall, London.

Sang, H. & Genton, M. G. (2014). Tapered composite likelihood for spatial max–stable models. *Spatial Statistics*, 8, 86–103.

Searle, S., Casella, G., & McCulloch, C. (1992). *Variance Components*, Wiley, New York.
Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5–42.
Vovk, V. & Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49, 1736–1754.
Wald, A. (1947). *Sequential Analysis*, Wiley, New York.
Wasserman, L., Ramdas, A., & Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences USA*, 117, 16880–16890.

Received 4 April 2022
Accepted 12 June 2022