WARWICK
THE UNIVERSITY OF WARWICK

Manuscript version: Author's Accepted Manuscript
The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/161698

How to cite:
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:
Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

warwick.ac.uk/lib-publications

Statistica Sinica Preprint No: SS-2021-0027

Title	Parametric Bootstrap Inference for Stratified Models With High-Dimensional Nuisance Specifications
Manuscript ID	SS-2021-0027
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0027
Complete List of Authors	Ruggero Bellio, Ioannis Kosmidis, Alessandra Salvan and Nicola Sartori
Corresponding Author	Ruggero Bellio
E-mail	ruggero.bellio@uniud.it
Notice: Accepted version subject to English editing.	

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Statistica Sinica

PARAMETRIC BOOTSTRAP INFERENCE FOR STRATIFIED MODELS WITH HIGH-DIMENSIONAL NUISANCE SPECIFICATIONS

Ruggero Bellio, Ioannis Kosmidis, Alessandra Salvan and Nicola Sartori

University of Udine, University of Warwick and The Alan Turing Institute,

University of Padova and University of Padova

Abstract: Inference about a scalar parameter of interest typically relies on the asymptotic normality of common likelihood pivots, such as the signed likelihood root, the score and Wald statistics. Nevertheless, the resulting inferential procedures are known to perform poorly when the dimension of the nuisance parameter is large relative to the sample size and when the information about the parameters is limited. In many such cases, the use of asymptotic normality of analytical modifications of the signed likelihood root is known to recover inferential performance. It is proved here that parametric bootstrap of standard likelihood pivots results in as accurate inferences as analytical modifications of the signed likelihood root do in stratified models with stratum specific nuisance parameters. We focus on the challenging case where the number of strata increases as fast or faster than the stratum samples size. It is also shown that this equivalence holds regardless of whether constrained or unconstrained bootstrap is used. This is in

Statistica Sinica

contrast to when the number of strata is fixed or increases slower than the stratum sample size, where we show that constrained bootstrap corrects inference to a higher order than unconstrained bootstrap. Simulation experiments support the theoretical findings and demonstrate the excellent performance of bootstrap in extreme scenarios.

Key words and phrases: Incidental parameters, location and scale adjustment, modified profile likelihood, two-index asymptotics, profile score bias.

1. Introduction

Standard likelihood inference about a scalar parameter of interest is based on the asymptotic normality of likelihood pivots, such as the signed likelihood root, score and Wald statistics. This asymptotic approximation can be quite inaccurate in the presence of many nuisance parameters. An alternative, which guarantees higher accuracy, is based on the asymptotic normality of analytical modifications of the signed likelihood root, generally termed modified signed likelihood root (see for instance Severini, 2000, Chapter 7). In a two-index stratified asymptotic setting, in which both the dimension of the data and the number of nuisance parameters grow, the modified signed likelihood root has been proved to be highly accurate even in rather extreme scenarios with many nuisance parameters and very limited information (Sartori, 2003).

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Parametric bootstrap methods provide an alternative assessment of tail probabilities for likelihood pivots and, in standard asymptotic settings, where the number of nuisance parameters is fixed and regularity conditions are satisfied (Severini, 2000, Section 3.4), have been shown to guarantee an equivalent level of asymptotic accuracy as analytical modifications of the signed likelihood root (see Young and Smith, 2005, Chapter 11). In particular, the two main variants of parametric bootstrap are constrained and unconstrained bootstrap (also know as conventional bootstrap). In the latter, the sampling distribution of the statistic is computed at the full maximum likelihood estimate, and in the former at the constrained maximum likelihood estimate for a given value of the parameter of interest. In standard asymptotic settings, constrained bootstrap (DiCiccio et al., 2001; Lee and Young, 2005) corrects inference about a scalar parameter in the presence of nuisance parameters to a higher order than unconstrained bootstrap. On the other hand, numerical differences are rarely detectable. Although bootstrap methods are, typically, more computationally demanding than analytical approximations to the distribution of pivots, they are available in some non-regular cases in which the modified signed likelihood root is not computable.

We investigate the properties of parametric bootstrap in models for

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

stratified data in a two-index asymptotic setting, where both the number q of strata and the sample size m of each stratum grow. In this setting, the usual likelihood pivots are asymptotically standard normal provided $q = o(m)$, while the condition for the modified signed likelihood root is $q = o(m^3)$ (Sartori, 2003). If $q = O(m^\alpha)$, then for $0 \leq \alpha < 1$ the asymptotic normality of standard likelihood pivots still holds, with error of order $O_p(m^{(\alpha-1)/2})$ (Sartori, 2003, formula (8)), while asymptotic normality fails in the highly stratified case with $\alpha \geq 1$. In that case, the aim of higher-order solutions is to recover first-order validity of inferential procedures.

We show here that parametric bootstrap provides valid inference when $q = O(m^\alpha)$, provided that $\alpha < 3$. In particular, if $0 \leq \alpha < 1$, constrained bootstrap is theoretically more accurate than unconstrained bootstrap, and both improve over standard first-order asymptotic results. On the other hand, when $1 \leq \alpha < 3$ both variants of parametric bootstrap are equally accurate, recovering first-order accuracy with the same order of error as higher-order analytical solutions.

The theoretical results are supported by extensive simulation studies, which illustrate that parametric bootstrap is at least as accurate as use of the modified signed likelihood root, and provide evidence that constrained bootstrap can be even more accurate in some very extreme scenarios.

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

2. Background

Let $l(\theta; y)$ be the log-likelihood function for a parameter θ based on a sample y of size n, which is considered to be a realization of a random vector Y. We treat the case where the vector of parameters is partitioned as $\theta = (\psi, \lambda^\top)^\top$, where ψ is a scalar parameter of interest and λ is a vector of nuisance parameters, and denote by $\hat{\theta}(y) = (\hat{\psi}(y), \hat{\lambda}(y)^\top)^\top$ the maximum likelihood estimate of θ and by $\hat{\theta}_\psi(y) = (\psi, \hat{\lambda}_\psi(y)^\top)^\top$ the constrained maximum likelihood estimate of θ for fixed ψ. We let $U(\theta; y) = \nabla l(\theta; y)$ denote the score vector, and $j(\theta; y) = -\nabla\nabla^\top l(\theta; y)$ the observed information, with $i(\theta) = \mathrm{E}_\theta\{j(\theta; Y)\}$ denoting the expected information. The argument θ will be dropped when no ambiguity arises, and components of vectors and blocks of matrices will be denoted by subscripts, so that for instance $U_\psi(\theta; y)$ denotes the component of the score vector corresponding to ψ. Furthermore, the argument y will be dropped whenever evaluation is at the random variable Y instead of the sample y. For example, $U_\psi = U_\psi(\theta; Y)$, $U_\lambda = U_\lambda(\theta; Y)$, $i_{\psi\psi} = i_{\psi\psi}(\theta)$ and $i_{\psi\lambda} = i_{\psi\lambda}(\theta)$ are the (ψ, ψ) and (ψ, λ) blocks of $i(\theta)$, and so on.

The signed likelihood root, the score statistic and Wald statistic for

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

inference about ψ are

$$\begin{aligned} R(\psi;y) &= \text{sign}\left(\hat{\psi}(y)-\psi\right)\sqrt{2\left\{l(\hat{\theta}(y);y)-l(\hat{\theta}_\psi(y);y)\right\}} && (2.1) \\ S(\psi;y) &= \frac{U_p(\psi;y)}{\sqrt{i_{\psi\psi|\lambda}(\hat{\theta}_\psi(y))}}\,, && (2.2) \\ T(\psi;y) &= (\hat{\psi}(y)-\psi)\sqrt{j_p(\hat{\psi}(y);y)}\,, && (2.3) \end{aligned}$$

respectively, where $U_p(\psi;y)=U_\psi(\hat{\theta}_\psi(y);y)$ is the profile score, $j_p(\psi;y)=-dU_p(\psi;y)/d\psi$ is the profile observed information and $i_{\psi\psi|\lambda}=i_{\psi\psi}-i_{\psi\lambda}i_{\lambda\lambda}^{-1}i_{\lambda\psi}$ is the partial information about ψ. While (2.1) and (2.2) are invariant with respect to interest respecting reparameterizations, (2.3) is not.

Computation of p-values and confidence intervals for ψ requires the distribution of statistics (2.1), (2.2) and (2.3). In standard asymptotic settings, one possibility is to rely on the first-order asymptotic normal approximation to their distribution. For instance, $\text{pr}_\theta\{R(\psi)\leq R(\psi;y)\}=\Phi(R(\psi;y))\{1+O(n^{-1/2})\}$, where $\Phi(\cdot)$ denotes the standard normal distribution function. Improved accuracy can be obtained with higher-order modifications $R^*(\psi;y)$ of $R(\psi;y)$, such that $\text{pr}_\theta\{R(\psi)\leq R(\psi;y)\}=\Phi(R^*(\psi;y))\{1+O(n^{-1})\}$. Barndorff-Nielsen (1986) developed a modified signed likelihood root $R^*(\psi)$ which is standard normal with error of order $O(n^{-3/2})$. Following this seminal work, there have been various alternative versions of $R^*(\psi;y)$ (see Pierce and Bellio, 2017, for an accessible overview).

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

An alternative to the asymptotic approximations to the distribution of (2.1), (2.2) and (2.3) is parametric bootstrap, which provides higher-order approximations for p-values, such as $\mathrm{pr}_\theta\{R(\psi) \leq R(\psi; y)\}$. There are two main variants of parametric bootstrap: i) unconstrained bootstrap where samples are simulated from the model at $\hat{\theta}(y)$, and ii) constrained bootstrap where samples are simulated at $\hat{\theta}_\psi(y)$ (see DiCiccio et al., 2001; Lee and Young, 2005; Young and Smith, 2005, Chapter 11).

In standard asymptotic settings, unconstrained bootstrap provides second-order accuracy. Let $G_\theta(\cdot)$ denote the distribution function of $R(\psi)$ at θ, so that $G_\theta(R(\psi))$ is exactly uniform. If data y^k are simulated from the model with parameter $\hat{\theta}(y)$, $k = 1, \ldots, K$, then p-values for (2.1) calculated as

$$\hat{p}_1^R(\psi) = \frac{1}{K} \sum_{k=1}^{K} I\{R(\hat{\psi}(y); y^k) \leq R(\psi; y)\} \tag{2.4}$$

are Monte Carlo estimates of $G_{\hat{\theta}}(R(\psi))$, which is uniform on $(0, 1)$ under repeated sampling with error of order $O(n^{-1})$, i.e.

$$\mathrm{pr}_\theta \left(G_{\hat{\theta}}(R(\psi)) \leq u\right) = u + O(n^{-1})\,. \tag{2.5}$$

In (2.4), $I\{\cdot\}$ is the indicator function.

In contrast, constrained bootstrap provides third-order accuracy; if data y^k are simulated at $\hat{\theta}_\psi(y)$, $k = 1, \ldots, K$, p-values for (2.1) calculated as

$$\hat{p}_2^R(\psi) = \frac{1}{K} \sum_{k=1}^{K} I\{R(\psi; y^k) \leq R(\psi; y)\} \tag{2.6}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

are Monte Carlo estimates of $G_{\hat{\theta}_\psi}(R(\psi))$, which is uniform on $(0,1)$ under repeated sampling with error of order $O(n^{-3/2})$ (Lee and Young, 2005), i.e.

$$\mathrm{pr}_\theta\left(G_{\hat{\theta}_\psi}(R(\psi)) \leq u\right) = u + O(n^{-3/2})\,. \tag{2.7}$$

Similar results hold for $S(\psi)$ and $T(\psi)$ (Lee and Young, 2005; Young, 2009) with p-values $\hat{p}_1^S$ and $\hat{p}_2^S$, and $\hat{p}_1^T$ and $\hat{p}_2^T$, respectively.

As Young and Smith (2005, Section 11.4) note, the theoretical advantage of constrained over unconstrained bootstrap is rarely supported by numerical evidence, because both types of bootstrap are able to equally improve over first-order results.

The advantage of bootstrap p-values in (2.4) and (2.6) over the use of analytical modifications to common statistics is that bootstrap does not require any additional, often tedious, algebraic derivation and implementation of the necessary modifications. Moreover, there are non-standard modelling settings, where $R(\psi; y)$ is computable while $R^*(\psi; y)$ is not. One instance is when one or more components of $\hat{\theta}(y)$ are on the boundary of the parameter space. The main disadvantage of bootstrap is the additional computation that is typically required for the repeated model fits, which can be partly mitigated by parallel computing.

In some special cases, the distribution of (2.1), (2.2) and (2.3) depends

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

only on ψ, so that constrained bootstrap, as well as simulating data at $(\psi, \hat{\lambda}(y)^\top)^\top$ or even at $(\psi, \lambda_0^\top)^\top$ for arbitrary nuisance vectors λ_0, produces samples from the hypothesized model. This is the case when the model for fixed ψ is a transformation model (see Severini, 2000, Section 1.3). For instance, if y is a realization of $Y = (Y_1, \dots, Y_n)^\top$ with independent and identically distributed components from a shape and scale model with generic density

$$g(y_i; \psi, \lambda) = \frac{1}{\lambda} g^0(y_i/\lambda; \psi)\,,$$

we may write $Y_i = \lambda Y_i^0$, with $Y_i^0 \sim g^0(y_i; \psi) = g(y_i; \psi, 1)$. Hence, due to equivariance of the maximum likelihood estimator, $\hat{\lambda}$ and $\lambda\hat{\lambda}^0$ have the same distribution, where $\hat{\lambda}^0$ is the maximum likelihood estimator of λ based on Y_i^0's. The same representation holds for $\hat{\lambda}_\psi$, so that the profile likelihood ratio

$$\exp\{l(\hat{\psi}, \hat{\lambda}) - l(\psi, \hat{\lambda}_\psi)\} = \prod_{i=1}^{n} \frac{\hat{\lambda}_\psi g^0(Y_i/\hat{\lambda}; \hat{\psi})}{\hat{\lambda} g^0(Y_j/\hat{\lambda}_\psi; \psi)}$$

has the same distribution as

$$\prod_{i=1}^{n} \frac{\lambda\hat{\lambda}^0_\psi g^0\left(\frac{\lambda Y_i^0}{\lambda\hat{\lambda}^0}; \hat{\psi}\right)}{\lambda\hat{\lambda}^0 g^0\left(\frac{\lambda Y_i^0}{\lambda\hat{\lambda}_\psi}; \psi\right)} = \prod_{i=1}^{n} \frac{\hat{\lambda}^0_\psi g^0(Y_i^0/\hat{\lambda}^0; \hat{\psi})}{\hat{\lambda}^0 g^0(Y_i^0/\hat{\lambda}_\psi; \psi)}\,,$$

which depends on ψ only.

An example with a stratified gamma model is provided in the Supplementary Materials where simulation results confirm the exactness of the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

constrained bootstrap.

3. Two-index asymptotic theory for stratified models

We consider a stratified setting with q independent strata with m observations each. Therefore, the total number of observations is $n = mq$. The models considered here have $\lambda = (\lambda_1, \ldots, \lambda_q)^\top$ as nuisance parameter, where λ_i is a stratum-specific parameter. Let $y_i = (y_{i1}, \ldots, y_{im})^\top$, $i = 1, \ldots, q$, denote the vector of observations in the ith stratum and let $y = (y_1^\top, \ldots, y_q^\top)^\top$. The vectors $y_1, \ldots, y_q$ are assumed to be realizations of independent random variables $Y_1, \ldots, Y_q$ from a parametric model with densities $g_1(y_1; \psi, \lambda_1), \ldots, g_q(y_q; \psi, \lambda_q)$, respectively. The observations within strata are also assumed to be realizations of independent random variables, so that $g_i(y_i; \psi, \lambda_i) = \prod_{j=1}^m g_{ij}(y_{ij}; \psi, \lambda_i)$, where $g_{ij}(\cdot)$ may be conditional on a covariate vector x_{ij}. Under this specification, for fixed ψ, the likelihood has separable parameters $\lambda_1, \ldots, \lambda_q$, so that $U_p(\psi) = \sum_{i=1}^q U_\psi^i(\psi, \hat{\lambda}_{i\psi})$, where U_ψ^i is the contribution to U_ψ from the ith stratum.

We work in a two-index asymptotic setting where q increases with m, as $q = O(m^\alpha)$, $\alpha > 0$. The case $\alpha = 0$ corresponds to the standard asymptotic setting. Sartori (2003, Section 4) showed that $R(\psi)$, $S(\psi)$ and $T(\psi)$ are asymptotically equivalent to order $o_p(1)$ for $\alpha \geq 0$. Specifically,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

when $0 \leq \alpha < 1$, the equivalence of the three quantities holds with relative error of order $O_p(n^{-1/2}) = O_p(m^{-(\alpha+1)/2})$, and these are asymptotically standard normal. On the other hand, when $\alpha \geq 1$, asymptotic equivalence of $R(\psi)$, $S(\psi)$ and $T(\psi)$ holds with error of order $O_p(m^{-1})$ and, more critically, the three statistics are not asymptotically standard normal, so that, for instance, $\Phi\{R(\psi)\}$ is not asymptotically uniform.

The derivation of the results is more straightforward for $S(\psi)$ because the profile score is the sum of strata profile scores. However, the same results hold also for $R(\psi)$ and $T(\psi)$, since, as recalled above, they are both asymptotically equivalent to $S(\psi)$. Let $F_\theta(\cdot)$ denote the distribution function of $S(\psi)$ under θ, so that $F_\theta(S(\psi))$ is exactly uniform.

The core result of the paper is that asymptotic validity of both constrained and unconstrained bootstrap is guaranteed even in a two-index asymptotic setting, provided that $\alpha < 3$, that is $q = o(m^3)$. The latter condition is the same as the one required for validity of inference based on the modified signed likelihood root $R^*(\psi)$ (Sartori, 2003). In particular, we show that, when $0 < \alpha < 1$,

$$\mathrm{pr}_\theta \left(F_{\hat{\theta}_\psi}(S(\psi)) \leq u \right) = u + O(m^{(\alpha-3)/2}) \tag{3.8}$$

and

$$\mathrm{pr}_\theta \left(F_{\hat{\theta}}(S(\psi)) \leq u \right) = u + O(m^{-1}) \,, \tag{3.9}$$

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

while, when $1 \leq \alpha < 3$,

$$\mathrm{pr}_\theta \left(F_{\hat{\theta}_\psi}(S(\psi)) \leq u \right) = u + O(m^{(\alpha-3)/2}) \tag{3.10}$$

and

$$\mathrm{pr}_\theta \left(F_{\hat{\theta}}(S(\psi)) \leq u \right) = u + O(m^{(\alpha-3)/2}) \,. \tag{3.11}$$

Hence, when $1 \leq \alpha < 3$, the same order of error is obtained both with constrained and unconstrained bootstrap, unlike what happens with $0 \leq \alpha < 1$. The case $\alpha = 0$ corresponds to the standard asymptotic setting in which $n = O(m)$, and (3.8) and (3.9) reduce to (2.7) and (2.5), respectively. A first intuition about why the two types of bootstrap have the same accuracy when $\alpha \geq 1$ is that the major effect of both bootstrap procedures is to remove the diverging bias term of the statistic, which overshadows any minor differences in theoretical performance that are found when $0 \leq \alpha < 1$. A formal development of the result is given below.

In the following we will concentrate on the more extreme case, i.e. $\alpha \geq 1$, while the proof of (3.8) and (3.9) for the case $0 < \alpha < 1$ is given in the Supplementary Materials. In order to prove both (3.10) and (3.11) we need some preliminary results about the distribution function $F_\theta(x)$ of $S(\psi)$ in the two-index asymptotic setting. From Sartori (2003, formula (6)),

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$U_p = U_p(\psi)$ can be expanded as

$$U_p = U_{\psi|\lambda} + B + Re\,, \tag{3.12}$$

where $U_{\psi|\lambda} = U_\psi - i_{\psi\lambda} i_{\lambda\lambda}^{-1} U_\lambda = O_p(\sqrt{n}) = O_p(m^{(\alpha+1)/2})$, having zero mean and variance $i_{\psi\psi|\lambda}$, $B = B(\theta) = O_p(m^\alpha)$ and, with $\alpha > 1$, $Re = O_p(m^{\alpha-1})$. Details about the orders in (3.12) are provided in the Appendix. When $0 \leq \alpha < 1$ the terms in (3.12) are in descending order. Instead, when $1 \leq \alpha < 3$, B becomes the leading term, followed by $U_{\psi|\lambda}$. Finally, $U_{\psi|\lambda}$ is dominated by Re as well when $\alpha \geq 3$. In practice, when $1 \leq \alpha < 3$, bootstrap procedures, as well as higher-order analytical solutions, are able to correct for B, so that $U_{\psi|\lambda}$ is again the leading term in expansion (3.12).

Let $M(\theta) = \mathrm{E}_\theta(S(\psi))$ and $\mathrm{Var}_\theta(S(\psi))$ be the expectation and variance of $S(\psi)$. Asymptotic expansions detailed in the Appendix can be used to show that

$$M(\theta) = \frac{b(\theta)}{i_{\psi\psi|\lambda}(\theta)^{1/2}} + M_1(\theta) + O(m^{(\alpha-5)/2}) \tag{3.13}$$

$$\mathrm{Var}_\theta(S(\psi)) = 1 + v(\theta) + O(1/m^2)\,, \tag{3.14}$$

where $b(\theta) = \mathrm{E}_\theta(B) = O(m^\alpha)$, $M_1(\theta) = O(m^{(\alpha-3)/2})$ and $v(\theta) = (\mathrm{Var}_\theta(B) + 2\,\mathrm{E}_\theta(U_{\psi|\lambda} B))/i_{\psi\psi|\lambda} = O(m^{-1})$. The cumulants of $S(\psi)$ of order $r \in \{3, 4, \ldots\}$ are $O\left(m^{(\alpha+1)(1-r/2)}\right) = O(n^{1-r/2})$, as in standard asymptotics.

For the development here, we assume that the distribution function of

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$S(\psi)$ admits a valid Edgeworth expansion. Severini (2000, Sections 5.1-5.4) gives conditions and details for the extension of Edgeworth expansions for independent and identically distributed random variables to likelihood pivots, such as $R(\psi)$, $S(\psi)$, $T(\psi)$. The basic requirement, in the continuous case, is that an Edgeworth expansion exists for the joint distribution of log-likelihood derivatives up to the third order, implying

$$F_\theta(x) = \text{pr}_\theta\left(S(\psi) \le x\right) = \Phi\left(\frac{x - M(\theta)}{\sqrt{\text{Var}_\theta(S(\psi))}}\right) + O(m^{-(\alpha+1)/2})\,, \qquad (3.15)$$

where the order of the remainder term is that of the third cumulant of $S(\psi)$. Let $x^*(\theta) = (x - M(\theta))/\sqrt{1+v(\theta)}$. Then

$$\frac{x - M(\theta)}{\sqrt{\text{Var}_\theta(S(\psi))}} = \frac{x - M(\theta)}{\sqrt{1 + v(\theta) + O(m^{-2})}} = x^*(\theta) + O(m^{-2})$$

and

$$F_\theta(x) = \Phi\left(x^*(\theta)\right) + O\left(m^{-\min\left(2, \frac{\alpha+1}{2}\right)}\right)\,. \qquad (3.16)$$

We first focus on constrained bootstrap. From (3.16),

$$F_{\hat{\theta}_\psi}(x) = \Phi\left(x^*(\hat{\theta}_\psi)\right) + O_p\left(m^{-\min\left(2, \frac{\alpha+1}{2}\right)}\right)\,. \qquad (3.17)$$

The Taylor expansions in the Appendix give

$$M(\hat{\theta}_\psi) = M(\theta) + \Delta + O_p\left(m^{-\min\left(1, \frac{5-\alpha}{2}\right)}\right) \qquad (3.18)$$

and

$$v(\hat{\theta}_\psi) = v(\theta) + O_p(m^{-2})\,, \qquad (3.19)$$

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

where $\Delta = O_p(m^{(\alpha-3)/2})$ is given in expression (A9) of the Appendix. Using (3.18) and (3.19), we can write $x^*(\hat{\theta}_\psi) = x^*(\theta) - \Delta + O_p(m^{-\min(1,(5-\alpha)/2})$. As a result, if $\alpha < 3$, then the following Taylor expansion of (3.17) holds

$$F_{\hat{\theta}_\psi}(x) = F_\theta(x) - \phi(x^*(\theta))\Delta + O_p(m^{-1})\,, \tag{3.20}$$

where the error is of order $O_p(m^{-1})$ because, for $\alpha < 3$, $\min(1, (5-\alpha)/2) = 1$, while the error term in (3.17) is $o_p(m^{-1})$ whenever $\alpha > 1$.

In order to prove (3.10), note that $F_{\hat{\theta}_\psi}(S(\psi)) \leq u$ is equivalent to $S(\psi) \leq s_u$, with s_u the u-quantile of $F_{\hat{\theta}_\psi}(\cdot)$, such that $F_{\hat{\theta}_\psi}(s_u) = u$. Let s_u^0 be the u-quantile of $F_\theta(\cdot)$. It is useful to express s_u in terms of s_u^0. Using (3.20),

$$u = F_\theta(s_u^0) = F_{\hat{\theta}_\psi}(s_u) = F_\theta(s_u) - \phi(s_u^*(\theta))\Delta + O_p(m^{-1})\,,$$

where $s_u^*(\theta) = (s_u - M(\theta))/\sqrt{1+v(\theta)}$. Hence, $F_\theta(s_u) - F_\theta(s_u^0) = \phi(s_u^*(\theta))\Delta + O_p(m^{-1})$. On the other hand, letting $F'_\theta(x) = dF_\theta(x)/dx$, from

$$F_\theta(s_u^0) = F_\theta(s_u) + (s_u^0 - s_u)F'_\theta(s_u) + O_p((s_u^0 - s_u)^2)$$

and

$$F'_\theta(x) = \phi(x^*(\theta))/\sqrt{1+v(\theta)} + O(m^{-(\alpha+1)/2}) = \phi(x^*(\theta)) + O(m^{-1})$$

we get

$$s_u = s_u^0 + \Delta + O_p(m^{-1}) + O_p(m^{\alpha-3})\,,$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

where the $O_p(m^{\alpha-3})$ term on the right hand side comes from $O_p((s_u^0 - s_u)^2)$. Hence, $S(\psi) \leq s_u$ is equivalent to $S(\psi) \leq s_u^0 + \Delta + O_p(m^{-1}) + O_p(m^{\alpha-3})$, and

$$\text{pr}_\theta \left(F_{\hat{\theta}_\psi}(S(\psi)) \leq u \right) = \text{pr}_\theta \left(\bar{S}(\psi) \leq F_\theta^{-1}(u) \right) ,$$

where $\bar{S}(\psi) = S(\psi) - \Delta + O_p(m^{\alpha-3}) + O_p(m^{-1})$, with Δ given by (A9), and such that $\text{E}_\theta(\Delta) = O(m^{(\alpha-3)/2})$. Moreover, we have

$$\begin{aligned} \text{E}_\theta(\bar{S}(\psi)) &= \text{E}_\theta(S(\psi)) + O(m^{(\alpha-3)/2}) \,, && (3.21) \\ \text{Var}_\theta(\bar{S}(\psi)) &= \text{Var}_\theta(S(\psi) - \Delta) + O(m^{-2}) \\ &= \text{Var}_\theta(S(\psi)) + \text{Var}_\theta(\Delta) - 2\text{Cov}_\theta(S(\psi), \Delta) + O(m^{-2}) \\ &= \text{Var}_\theta(S(\psi)) + O(m^{-2}) \,, && (3.22) \end{aligned}$$

since $\text{Var}_\theta(\Delta) = O(m^{-2})$ and $\text{Cov}_\theta(S(\psi), \Delta) = O(m^{-2})$, where the order of the latter is determined by the orthogonality between $U_{\psi|\lambda}$ and the leading term of $b_1(\theta)$ in (A7). Finally, (3.10) holds because

$$\begin{aligned} \text{pr}_\theta \left(\bar{S}(\psi) \leq F_\theta^{-1}(u) \right) &= \text{pr}_\theta \left(S(\psi) \leq F_\theta^{-1}(u) \right) + O(m^{(\alpha-3)/2}) + O(m^{-2}) \\ &= \text{pr}_\theta \left(S(\psi) \leq F_\theta^{-1}(u) \right) + O(m^{(\alpha-3)/2}) \\ &= u + O(m^{(\alpha-3)/2}) \,. \end{aligned}$$

The proof of (3.11) for unconstrained bootstrap is obtained along the same steps as above. In particular, expansion (A12) holds for $F_{\hat{\theta}}(x)$, having

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

the same form as (3.20), with Δ replaced by Δ_1, which is still of order $O_p(m^{(\alpha-3)/2})$. Details are given in the Appendix. However, while (3.21) is still true, (3.22) holds with an error of order $O(m^{-1})$, because there is no orthogonality between $U_{\psi|\lambda}$ and the leading terms of $b_2(\theta)$, given in (A10). Therefore, for unconstrained bootstrap we have

$$\begin{aligned}
\mathrm{pr}_\theta\left(\bar{S}(\psi) \leq F_\theta^{-1}(u)\right) &= \mathrm{pr}_\theta\left(S(\psi) \leq F_\theta^{-1}(u)\right) + O(m^{(\alpha-3)/2}) + O(m^{-1}) \\
&= \mathrm{pr}_\theta\left(S(\psi) \leq F_\theta^{-1}(u)\right) + O(m^{(\alpha-3)/2}) \\
&= u + O(m^{(\alpha-3)/2})\,.
\end{aligned}$$

Hence, when $\alpha \geq 1$, errors in (3.10) and (3.11) are of the same order because the $O(m^{(\alpha-3)/2})$ error in the mean of $\bar{S}(\psi)$ dominates the $O(m^{-2})$ and $O(m^{-1})$ errors in the variance of $\bar{S}(\psi)$ in the constrained and unconstrained cases, respectively. However, the different errors in the variance of $\bar{S}(\psi)$ may have some effects and explain why the constrained bootstrap is sometimes numerically more accurate in extreme settings.

The arguments used in the proofs of (3.10) and (3.11) suggest that a location and scale adjustment to the statistic, as done for $R(\psi)$ in a standard asymptotic setting by DiCiccio et al. (2001) and Stern (2006), is the key requirement to recover approximate uniformity of p-values. In this respect, a bootstrap location and scale adjustment of $R(\psi)$, $S(\psi)$ or $T(\psi)$ is

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

expected to be as effective as bootstrapping the distribution of the statistic. This conjecture is confirmed by the numerical results, both in the following section and in the Supplementary Materials.

4. Simulation studies

The finite-sample properties of unconstrained and constrained parametric bootstrap are assessed through extensive simulation studies, for three statistical models for stratified data. In particular, we consider a beta model, a curved exponential family model and a truncated regression model, with the results for further models reported in the Supplementary Materials. For each model, we conduct 9 simulation experiments, one for each combination of number of strata $q \in \{10, 100, 1000\}$ and stratum sample size $m \in \{4, 8, 16\}$.

Each simulation experiment involves 10000 simulated samples under the model at a fixed parameter vector $\theta_0 = (\psi_0, \lambda_0^\top)^\top$. For each simulated sample, 17 statistics and 6 bootstrap-based p-values are computed for testing $\psi = \psi_0$. In particular, the statistics that are computed are i) $R(\psi)$, $S(\psi)$, $T(\psi)$, ii) the location and location-and-scale adjusted versions of $R(\psi)$, $S(\psi)$, $T(\psi)$, where the mean and variance of each statistic are estimated using unconstrained bootstrap (at $\hat{\theta}$) and constrained boostrap (at $\hat{\theta}_\psi$), and

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

iii) $R^{*}(\phi)$ and the signed likelihood root computed from the modified profile likelihood (see, for instance Severini, 2000, Chapter 8). The higher-order adjustment required for the latter two statistics is obtained using expected moments of likelihood quantities as in Severini (2000, Section 7.5). Finally, for each of $R(\psi)$, $S(\psi)$, and $T(\psi)$, we compute the unconstrained and constrained bootstrap p-values in (2.4) and in (2.6), respectively.

In the interest of space, in what follows, we only report results for the 6 statistics based on $R(\psi)$ shown in Table 1. The conclusions for the remaining statistics and p-values are qualitatively the same. Results are also only presented for $(q, m) = (10, 4)$, $(q, m) = (100, 4)$, $(q, m) = (1000, 4)$, $(q, m) = (1000, 8)$, and $(q, m) = (1000, 16)$, because these combinations of q and m are sufficient for assessing the performance of the statistics as q and m grow. The results from all simulation experiments are provided in the Supplementary Materials.

The above experiments involve high-dimensional parameter spaces with as many as 1000 nuisance parameters. As a result, the assessment of the statistics requiring bootstrapping is demanding in terms of computational time and cost, even when parallel computing with a large number of cores is used. For this reason, the number of bootstrap samples is limited to 1000 in all simulation experiments.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 1: Statistics considered for the results of the simulation experiments. The mean $\tilde{\mu}^R$ and the standard deviation $\tilde{\sigma}^R$ of $R(\psi)$ are estimated through constrained bootstrap, by simulating from the model at $\theta = \hat{\theta}_\psi$.

Statistic	Plotting Symbol	Description
$R(\psi)$	R	Signed likelihood root
$R^*(\psi)$	R^*	Modified signed likelihood root
$\Phi^{-1}\{\hat{p}_1^R(\psi)\}$	R^u	Transformed p-value from unconstrained bootstrap of $R(\psi)$
$\Phi^{-1}\{\hat{p}_2^R(\psi)\}$	R^c	Transformed p-value from constrained bootstrap of $R(\psi)$
$R(\psi) - \tilde{\mu}^R$	R^c_l	Location adjusted $R(\psi)$
$(R(\psi) - \tilde{\mu}^R)/\tilde{\sigma}^R$	R^c_{ls}	Location-and-scale adjusted $R(\psi)$

The three blocks of rows in Table 2 give the estimated tail probabilities of the statistics of interest for the case $q = 1000$ and $m = 8$ for all three models considered. This combination of q and m was selected because it is the least extreme setting (compared to the most extreme $q = 1000$, $m = 4$) where departures from the expected behaviour in terms of the

Statistica Sinica

distribution of the statistics starts becoming apparent; the results for all the other combinations of q and m are provided in the Supplementary Materials. The following sections give a more detailed discussion on the figures shown in Table 2.

Table 2: Empirical tail probabilities ×100 for the statistics in Table 1 and all models considered in the simulation studies of Section 4. The figures shown have been rounded to 1 decimal and are for $q = 1000$ and $m = 8$.

		Nominal					
Model	Statistic	1.0	2.5	5.0	95.0	97.5	99.0
Beta	R	0.0	0.0	0.0	0.0	0.0	0.0
	R^*	0.7	1.8	3.8	93.7	96.8	98.8
	R^u	0.8	1.9	4.1	94.0	97.0	98.7
	R^c	1.0	2.3	4.8	95.0	97.4	99.1
	R^c_l	1.1	2.5	5.1	94.7	97.3	98.9
	R^c_{ls}	0.9	2.3	4.8	95.1	97.5	99.0
Curved exponential family	R	100.0	100.0	100.0	100.0	100.0	100.0
	R^*	1.4	3.5	6.9	96.6	98.3	99.4
	R^u	0.6	1.8	4.0	95.0	97.7	99.2
	R^c	1.2	3.3	6.4	96.2	98.2	99.4
	R^c_l	1.5	3.6	7.1	95.8	98.0	99.2
	R^c_{ls}	1.3	3.2	6.5	96.3	98.2	99.4
Truncated regression	R	0.2	0.5	1.1	84.2	90.4	95.1
	R^*	1.0	2.5	5.2	94.8	97.3	98.9
	R^u	0.9	2.3	4.8	94.9	97.2	98.9
	R^c	0.9	2.4	4.9	94.5	97.2	98.7
	R^c_l	1.0	2.4	5.0	94.4	97.0	98.8
	R^c_{ls}	0.9	2.4	5.0	94.4	97.0	98.8

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

4.1 Beta model

As a first example, we suppose that Y_{ij} has a beta distribution, with density function

$$g(y_{ij};\mu_i,\phi)=\frac{1}{B\{\mu_i\phi,(1-\mu_i)\phi\}}y_{ij}^{\mu_i\phi-1}(1-y_{ij})^{(1-\mu_i)\phi-1}\quad(0<y_{ij}<1)\,,$$

where $B(\cdot)$ is the beta function. The parameter of interest is $\psi=\log\phi$, whereas the stratum-specific nuisance parameters are given by $\lambda_i=\log\{\mu_i/(1-\mu_i)\}$. The simulation experiments are carried out for $\psi_0=\log(2)$ and the elements of λ_0 are generated from a standard normal distribution and held fixed over all the replications.

The left panel of Figure 1 shows the empirical densities for the statistics in Table 1. The performance of the statistics is evaluated in terms of the closeness of their empirical density to the standard normal density. This assessment is valid also for the constrained and unconstrained bootstrap p-values, since they have been mapped into the standard normal scale by the $\Phi^{-1}(\cdot)$ transformation.

The large location bias of the distribution of $R(\psi)$ is apparent for all shown combinations of q and m, and it becomes huge for $q=1000$ and $m\in\{4,8\}$. All higher-order accurate statistics result in a marked finite-sample correction, with $R^*(\psi)$ and the unconstrained bootstrap illustrating some

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

discrepancy from the standard normal distribution for large q/m ratios, such as $q = 1000$ and $m \in \{4, 8\}$. This is also apparent from the entries in Table 2.

From the right panel of Figure 1, it is noticeable that the p-values based on R^c, the location adjusted version R^c_l and the location-and-scale adjusted version R^c_{ls} are all close to one another. Hence, the necessary adjustment for making the distribution of $R(\psi)$ to be close to standard normal appears to be mainly a location adjustment.

4.2 Curved exponential family

This example involves normally distributed random variables Y_{ij}, each with mean $\exp(\lambda_i)$ and variance $\exp(\psi + \lambda_i/2)$. This model was studied in Sartori et al. (1999), where it is pointed out that a marginal likelihood for ψ is not available. The simulation experiments are carried out for $\psi_0 = \log(1/2)$ and the elements of λ_0 are generated from a standard normal distribution and held fixed over all the replications.

The left panel in Figure 2 shows the empirical density functions of the statistics in Table 1, and the right panel shows the corresponding p-value distributions. Like in the previous example, the empirical, finite-sample distributions of $R(\psi)$ are far from standard normal, while all the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

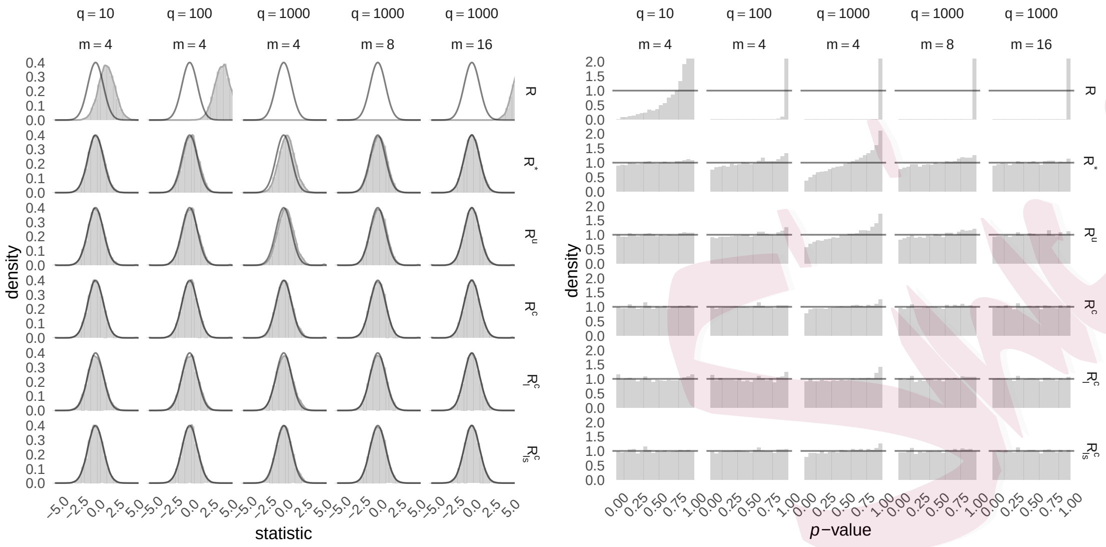

Figure 1: Beta model. Estimated null distribution of statistics (left) and estimated distribution of p-values (right) for the statistics in Table 1 for various combinations of q and m. The N$(0, 1)$ and Uniform$(0, 1)$ density functions are superimposed for statistics (left) and p-values (right).

higher-order statistics perform considerably better. The conclusions are similar to those from the simulation experiments for the beta model, in that the required adjustment to $R(\psi)$ seems to be a location correction. The main difference is the fact that no statistic appears to perform well for $(q, m) = (1000, 4)$; see also the empirical tail probabilities in Table 2.

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

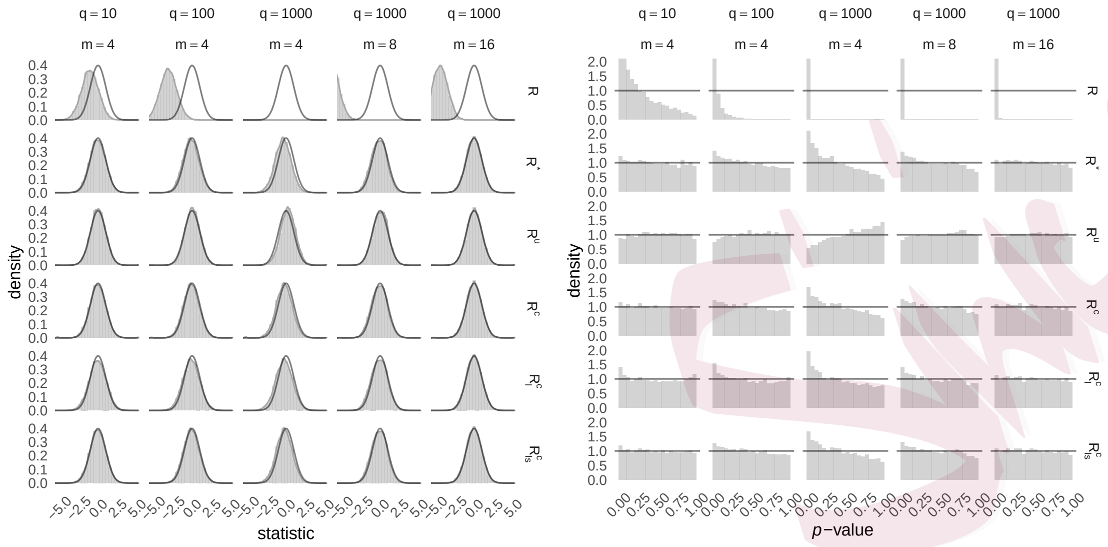

Figure 2: Curved exponential family model. Estimated null distribution of statistics (left) and estimated distribution of p-values (right) for the statistics in Table 1 for various combinations of q and m. The $\mathrm{N}(0,1)$ and Uniform$(0,1)$ density functions are superimposed for statistics (left) and p-values (right).

4.3 Truncated linear regression model

The last example is taken from the econometric literature; see Greene (2004), Bartolucci et al. (2016) and the references therein. We define the response variable Y_{ij} to be distributed as Y_{ij}^* conditionally on $y_{ij}^* > 0$, with

$$y_{ij}^* = \lambda_i + x_{ij}\,\psi + \varepsilon_{ij}\,, \quad i = 1, \ldots, q, \quad j = 1, \ldots, m\,,$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

where the error term ε_{ij} is standard normally distributed. For the simulation study, ψ is set to 1 and the elements of λ_0 are generated from a standard normal distribution and held fixed over all the replications. Likewise, the values x_{ij} are generated from a standard normal distribution and held fixed over all the replications.

The left panel in Figure 3 shows the empirical density functions of the statistics in Table 1, and the right panel shows the corresponding p-value distributions. Differently from the other examples, here the distribution of the first-order statistics requires only a moderate adjustment even in the most extreme settings, and both the bootstrap-based statistics as well the $R^*(\psi)$ statistic perform rather well, providing results very close to the target distributions.

5. Concluding remarks

The contribution of this paper is to formally show that, in stratified settings, inference based on either unconstrained or constrained parametric bootstrap of usual likelihood pivots is effective in recovering their inferential performance, even in rather extreme settings, where the bias of the profile score renders vanilla first-order inference invalid.

Unconstrained and constrained bootstrap for the signed likelihood ratio

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

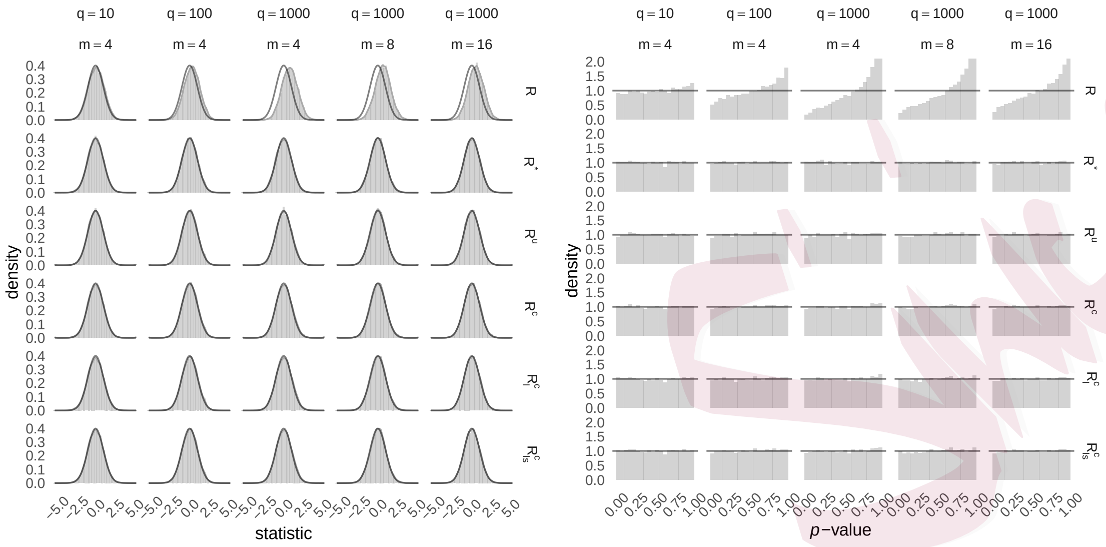

Figure 3: Truncated linear regression model. Estimated null distribution of statistics (left) and estimated distribution of p-values (right) for the statistics in Table 1 for various combinations of q and m. The N$(0,1)$ and Uniform$(0,1)$ density functions are superimposed for statistics (left) and p-values (right).

root, the score statistic and the Wald statistic can both recover inferential performance in stratified settings when $q = O(m^{\alpha})$, for $0 < \alpha < 3$. As in the case for $\alpha = 0$ (Lee and Young, 2005), when $0 < \alpha < 1$, constrained bootstrap is seen to have a higher degree of asymptotic accuracy than unconstrained bootstrap. On the other hand, the two bootstraps are asymptotically equivalent when $1 \leq \alpha < 3$. The condition $q = O(m^{\alpha})$, for

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$0 < \alpha < 3$, is the same as the one found in Sartori (2003) for validity of inference based on R^* and on the signed likelihood root computed from the modified profile likelihood.

The results of Section 4 from the extensive simulation studies for the finite-sample assessment of the performance of constrained and unconstrained bootstrap are in par to what is expected from theory. In extreme settings, like the beta model with $(q, m) = (1000, 4)$, constrained bootstrap appears to perform slightly better than unconstrained bootstrap. Furthermore, in all simulation experiments we carried out and as q/m diverges, the inferential performance from constrained and unconstrained bootstrap of first-order statistics seems to be deteriorating much slower than that of R^* and the signed likelihood root computed from the modified profile likelihood (see also the Supplementary Materials). As a result, the evidence from the simulation studies points out that inference from parametric bootstrap is more resilient to increasing q/m than inference from well-used, analytically available higher-order statistics, with constrained bootstrap being the most accurate in extreme scenarios.

The theoretical developments in this paper do not immediately cover situations where the random variables have discrete support, because the Edgeworth expansion in (3.15) can only be valid for models with continu-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

ous support. The impact of discreteness on the performance of parametric bootstrap is examined in the Supplementary Materials through a binomial matched pairs model. In particular, the experimental setup of Section 4 is used for a stratified logistic regression model, where Y_{ij} has a Bernoulli distribution with probability $\exp(\lambda_i + \psi x_j)/\{1+\exp(\lambda_i+\psi x_j)\}$, with $x_j = 1$ for $j \in \{1, \ldots, m/2\}$ and $x_j = 0$ for $j \in \{m/2+1, \ldots, m\}$. The results in Figures S21-S24 and Tables S3-S11 in the Supplementary Materials indicate that the equivalence between unconstrained and constrained bootstrap of the first-order statistics in continuous models may not hold for discrete settings. In those cases, despite that unconstrained bootstrap appears to deliver a marked inferential improvement to first-order statistics, constrained bootstrap, similarly to R^*, is found to perform considerably better for most combinations of q and m.

The simulation experiments in this paper have been carried out with 1000 bootstrap replications. This value is smaller than some of the recommendations of millions of replications that have appeared in the literature for standard asymptotics settings (Young, 2009; DiCiccio et al., 2017). For stratified settings with $\alpha > 1$ the bootstrap adjustments have the role of recovering asymptotic uniformity of p-values, rather that providing a small-sample refinement of p-values that are asymptotically uniform. As a result,

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

use of a huge number of bootstrap replications is less essential, and the few experiments we carried out with more than 1000 bootstrap replications are in support of that statement. More comprehensive simulation studies to support that statement are unfortunately not feasible with current computing capabilities.

Supplementary Materials

The Supplementary Materials provide the outputs from the simulation experiments described in Section 4, for all models and all combinations of statistics, q and m. Outputs are also provided for other models, given by a gamma model, a Behrens-Fisher model and by the logistic regression model described in Section 5. The outputs include null distributions of the various statistics and distributions of p-values, through extended versions of the Figures 1-3, and empirical tail probabilities, through extended versions of Table 2.

Acknowledgements

We thank two anonymous Referees and the Associate Editor for their constructive comments which enabled us to significantly improve our work. Ioannis Kosmidis has been supported by The Alan Turing Institute under

Statistica Sinica

the EPSRC grant EP/N510129/1. Part of this work was completed during a visit of Ruggero Bellio, Alessandra Salvan and Nicola Sartori to The Alan Turing Institute. Alessandra Salvan and Nicola Sartori have been supported by the University of Padova under grants BIRD185955 and BIRD203991.

Appendix

Asymptotic orders in (3.12)

The following representation from Sartori (2003, Appendix) will be used to determine the order of quantities in a stratified setting. Let μ_i and σ_i^2 denote mean and variance of independent the random variables $X_1, \ldots, X_q$. Then

$$\sum_{i=1}^{q} X_i = O_p\left(\sum_{i=1}^{q} \mu_i\right) + O_p\left(\sqrt{\sum_{i=1}^{q} \sigma_i^2}\right) . \qquad \text{(A1)}$$

We have $U_\psi = \sum_{i=1}^q U_\psi^i$, where U_ψ^i is the contribution to U_ψ from the ith stratum, and $U_\lambda = (U_{\lambda_1}, \ldots, U_{\lambda_q})^\top$. Here and in the following, when the argument is omitted, evaluation at θ is understood.

The terms on the right-hand side of (3.12) are seen to be of order $O_p(m^{(\alpha+1)/2})$, $O_p(m^\alpha)$ and $O_p(m^{\alpha-1})$, respectively. Indeed, using (A1), we have $U_{\psi|\lambda} = \sum_{i=1}^q U_{\psi|\lambda_i} = O_p(m^{(\alpha+1)/2})$, with $U_{\psi|\lambda_i} = U_\psi^i - i_{\psi\lambda_i} i_{\lambda_i\lambda_i}^{-1} U_{\lambda_i}$ being $\mathrm{E}_\theta(U_{\psi|\lambda_i}) = 0$ and $\mathrm{Var}_\theta(U_{\psi|\lambda_i}) = i_{\psi\psi|\lambda_i} = O(m)$. Note that $i_{\psi\psi|\lambda} = \mathrm{Var}_\theta(U_{\psi|\lambda}) = \sum_{i=1}^q i_{\psi\psi|\lambda_i}$. Similarly, we have $B = \sum_{i=1}^q B^i(\psi, \lambda_i) = O_p(m^\alpha)$,

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

where $B^i(\psi, \lambda_i)$ is the term of order $O_p(1)$ of the expansion of the profile score in the ith stratum, having both mean and variance of order $O(1)$. The same additivity property holds for $b(\theta)$, so that $b(\theta) = \sum_{i=1}^q b^i(\psi, \lambda_i) = O(m^\alpha)$. Finally, the remainder term is $Re = \sum_{i=1}^q Re^i(\psi, \lambda_i)$, with $Re^i(\psi, \lambda_i)$ having mean and variance of order $O(m^{-1})$, so that $Re = O_p(m^{\max\{\alpha-1,(\alpha-1)/2\}}) = O_p(m^{\alpha-1})$ when $\alpha > 1$.

Derivation of (3.13) and (3.14)

As a first step, consider the expansion

$$i_{\psi\psi|\lambda}(\hat{\theta}_\psi) = i_{\psi\psi|\lambda} + C + O_p(m^{\alpha-1}) \, , \tag{A2}$$

with

$$C = \sum_{i=1}^q \frac{d}{d\lambda_i} i_{\psi\psi|\lambda_i}(\hat{\lambda}_{i\psi} - \lambda_i) + \frac{1}{2}\sum_{i=1}^q \frac{d^2}{d\lambda_i^2} i_{\psi\psi|\lambda_i}(\hat{\lambda}_{i\psi} - \lambda_i)^2 = O_p(m^\alpha) \, ,$$

where when $\alpha > 1$ both terms in C are of the same order, which is again determined using (A1). Hence,

$$\{i_{\psi\psi|\lambda}(\hat{\theta}_\psi)\}^{-1/2} = i_{\psi\psi|\lambda}^{-1/2} \left\{ 1 - \frac{1}{2}\frac{C}{i_{\psi\psi|\lambda}} + O_p(m^{-2}) \right\} \, , \tag{A3}$$

with $C/i_{\psi\psi|\lambda} = O_p(m^{-1})$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Using (3.12) and (A3),

$$\begin{aligned} S(\psi) &= i_{\psi\psi|\lambda}^{-1/2}\left\{U_{\psi|\lambda}+B+Re\right\}\left\{1-\frac{1}{2}\frac{C}{i_{\psi\psi|\lambda}}+O_p(m^{-2})\right\} \\ &= \frac{U_{\psi|\lambda}}{i_{\psi\psi|\lambda}^{1/2}}+\frac{B}{i_{\psi\psi|\lambda}^{1/2}}+\frac{Re}{i_{\psi\psi|\lambda}^{1/2}}-\frac{1}{2}\frac{U_{\psi|\lambda}C}{i_{\psi\psi|\lambda}^{3/2}}-\frac{1}{2}\frac{B\,C}{i_{\psi\psi|\lambda}^{3/2}}-\frac{1}{2}\frac{Re\,C}{i_{\psi\psi|\lambda}^{3/2}} \\ &\quad +O_p(m^{-2})+O_p(m^{(\alpha-5)/2})+O_p(m^{(\alpha-7)/2})\,, \qquad \text{(A4)} \end{aligned}$$

where $O_p(m^{-2})+O_p(m^{(\alpha-5)/2})+O_p(m^{(\alpha-7)/2}) = O_p(m^{(\alpha-5)/2})$ as long as $\alpha > 1$. The term of order $O_p(m^{(\alpha-5)/2})$ is given by $i_{\psi\psi|\lambda}^{-1/2}B$ times the term of order $O_p(m^{-2})$ in (A3). Its expectation is of order $O(m^{(\alpha-5)/2})$. The orders of terms in (A4) are as follows:

$$\begin{aligned} &\frac{U_{\psi|\lambda}}{i_{\psi\psi|\lambda}^{1/2}} = O_p(1)\,, \quad \frac{B}{i_{\psi\psi|\lambda}^{1/2}} = O_p(m^{(\alpha-1)/2})\,, \quad \frac{Re}{i_{\psi\psi|\lambda}^{1/2}} = O_p(m^{(\alpha-3)/2})\,, \\ &\frac{1}{2}\frac{U_{\psi|\lambda}C}{i_{\psi\psi|\lambda}^{3/2}} = O_p(m^{-1}) = o_p(1)\,, \quad \frac{1}{2}\frac{B\,C}{i_{\psi\psi|\lambda}^{3/2}} = O_p(m^{(\alpha-3)/2})\,, \\ &-\frac{1}{2}\frac{Re\,C}{i_{\psi\psi|\lambda}^{3/2}} = O_p(m^{(\alpha-5)/2})\,. \end{aligned}$$

Expansion (3.13) for $\mathrm{E}_\theta(S(\psi))$ is obtained using (A4) and recalling that $b(\theta) = O(m^\alpha)$. We have

$$\begin{aligned} &\mathrm{E}_\theta\left(\frac{U_{\psi|\lambda}}{i_{\psi\psi|\lambda}^{1/2}}\right) = 0\,, \quad \mathrm{E}_\theta\left(\frac{B}{i_{\psi\psi|\lambda}^{1/2}}\right) = \frac{b(\theta)}{i_{\psi\psi|\lambda}^{1/2}} = O(m^{(\alpha-1)/2})\,, \\ &\mathrm{E}_\theta\left(\frac{Re}{i_{\psi\psi|\lambda}^{1/2}}\right) = O(m^{(\alpha-3)/2})\,, \quad \mathrm{E}_\theta\left(\frac{1}{2}\frac{U_{\psi|\lambda}C}{i_{\psi\psi|\lambda}^{3/2}}\right) = O(m^{-(\alpha+3)/2}) = o(1)\,, \\ &\mathrm{E}_\theta\left(\frac{1}{2}\frac{B\,C}{i_{\psi\psi|\lambda}^{3/2}}\right) = O_p(m^{(\alpha-3)/2})\,, \quad \mathrm{E}_\theta\left(-\frac{1}{2}\frac{Re\,C}{i_{\psi\psi|\lambda}^{3/2}}\right) = O(m^{(\alpha-5)/2})\,, \end{aligned}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

giving (3.13) with

$$M_1(\theta) = \mathrm{E}_\theta\left(\frac{Re}{i_{\psi\psi|\lambda}^{1/2}}\right) + \mathrm{E}_\theta\left(\frac{1}{2}\frac{B\,C}{i_{\psi\psi|\lambda}^{3/2}}\right) = O(m^{(\alpha-3)/2})\,. \qquad \text{(A5)}$$

Expansion (3.14) for $\mathrm{Var}_\theta(S(\psi))$ is also obtained using (A4). In particular, the leading term has variance equal to 1, and, using a standard expansion for the stratum profile score $U_\psi^i(\psi, \hat\lambda_{i\psi})$ (see e.g Pace and Salvan, 1997, formula (8.88)), $\mathrm{Cov}_\theta(U_{\psi|\lambda}, B)$ and $\mathrm{Var}_\theta(B)$ are easily seen to be of order $O(m^\alpha)$. Further terms of (A4) give contributions to the variance of order $O(m^{-2})$.

Higher order cumulants of $S(\psi)$, $r = 3, 4, \ldots$, have the form

$$\kappa_r(S(\psi)) = \frac{O(m^{\alpha+1})}{O(m^{r(\alpha+1)/2})} = O(m^{(\alpha+1)(1-r/2)}) = O(n^{1-r/2})$$

as in standard asymptotics.

Derivation of (3.18) and (3.19)

Let $\overline{Re} = \mathrm{E}_\theta(Re)$ and $\overline{BC} = \mathrm{E}_\theta(B\,C)$. Then, from (3.13) and (A5),

$$M(\hat\theta_\psi) = \left\{i_{\psi\psi|\lambda}(\hat\theta_\psi)\right\}^{-1/2}\left\{b(\hat\theta_\psi) + \overline{Re}(\hat\theta_\psi) + \frac{1}{2}\frac{1}{i_{\psi\psi|\lambda}(\hat\theta_\psi)}\overline{BC}(\hat\theta_\psi)\right\} + O_p(m^{(\alpha-5)/2})\,,$$

where $\overline{Re}(\hat\theta_\psi)$ and $i_{\psi\psi|\lambda}(\hat\theta_\psi)^{-1}\overline{BC}(\hat\theta_\psi)$ are of order $O(m^{\alpha-1})$. Now,

$$b(\hat\theta_\psi) = b(\theta) + b_1(\theta) + O_p(m^{\alpha-2})\,, \qquad \text{(A6)}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

where

$$b_1(\theta) = \sum_{i=1}^{q} b^i_{\lambda_i}(\psi,\lambda_i)(\hat{\lambda}_{i\psi} - \lambda_i) + \frac{1}{2}\sum_{i=1}^{q} b^i_{\lambda_i\lambda_i}(\psi,\lambda_i)(\hat{\lambda}_{i\psi} - \lambda_i)^2 \,, \qquad \text{(A7)}$$

and $b^i_{\lambda_i}(\psi,\lambda_i) = \partial b^i(\psi,\lambda_i)/\partial\lambda_i$, and so on. Using (A1), and being $b^i_{\lambda_i}(\psi,\lambda_i)$ and $b^i_{\lambda_i\lambda_i}(\psi,\lambda_i)$ both of order $O(1)$,

$$\sum_{i=1}^{q} b^i_{\lambda_i}(\psi,\lambda_i)(\hat{\lambda}_{i\psi} - \lambda_i) = O_p(m^{\alpha-1}) + O(m^{(\alpha-1)/2})$$

and

$$\sum_{i=1}^{q} b^i_{\lambda_i\lambda_i}(\psi,\lambda_i)(\hat{\lambda}_{i\psi} - \lambda_i)^2 = O_p(m^{\alpha-1}) + O_p(m^{(\alpha-2)/2}) \,.$$

The remainder in (A6) is of order $O_p(m^{\alpha-2}) + O_p(m^{(\alpha-3)/2}) = O_p(m^{\alpha-2})$, when $\alpha > 1$. Moreover, $\overline{Re}(\hat{\theta}_\psi) = \overline{Re} + O_p(m^{\alpha-2})$ and $i_{\psi\psi|\lambda}(\hat{\theta}_\psi)^{-1}\overline{BC}(\hat{\theta}_\psi) = i^{-1}_{\psi\psi|\lambda}\overline{BC} + O_p(m^{\alpha-2})$.

Using (A3), we get

$$M(\hat{\theta}_\psi) = i^{-1/2}_{\psi\psi|\lambda} b(\theta) + \tilde{M}_1 + O_p\left(m^{-\min\{1,(5-\alpha)/2\}}\right) \,, \qquad \text{(A8)}$$

with

$$\tilde{M}_1 = i^{-1/2}_{\psi\psi|\lambda}\left\{ b_1(\theta) - \frac{C\,b(\theta)}{2 i_{\psi\psi|\lambda}} + \overline{Re} + \frac{\overline{BC}}{2 i_{\psi\psi|\lambda}} \right\} \,,$$

which is of order $O_p(m^{(\alpha-3)/2})$ because all terms are of the same order.

Therefore, (A8), (3.13) and (A5) give (3.18) with

$$\Delta = \tilde{M}_1 - M_1(\theta) = \frac{b_1(\theta)}{i^{1/2}_{\psi\psi|\lambda}} - \frac{C\,b(\theta)}{2\,i^{3/2}_{\psi\psi|\lambda}} \qquad \text{(A9)}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

that is of order $O_p(m^{(\alpha-3)/2})$.

To obtain expansion (3.19) recall that in (3.14)

$$v(\theta) = (\mathrm{Var}_\theta(B) + 2\,\mathrm{E}_\theta(U_{\psi|\lambda}B))/i_{\psi\psi|\lambda}\,.$$

The numerator of $v(\hat\theta_\psi)$ is equal to $\mathrm{Var}_\theta(B) + 2\,\mathrm{E}_\theta(U_{\psi|\lambda}B)$ plus a term of order $O_p(m^{\alpha-1})$. From (A2),

$$\frac{1}{i_{\psi\psi|\lambda}(\hat\theta_\psi)} = \frac{1}{i_{\psi\psi|\lambda}}\left\{1 + O_p(m^{-1})\right\}\,.$$

which gives (3.19).

Derivation of (3.11)

First, from (3.16), we have

$$F_{\hat\theta}(x) = \Phi\left(x^*(\hat\theta)\right) + O_p\left(m^{-\min\left(2,\frac{\alpha+1}{2}\right)}\right)\,.$$

In order to obtain expansions of $M(\hat\theta)$ and $v(\hat\theta)$ around θ, we use the fact that, when $\alpha > 1$, $\hat\psi - \psi = O_p(m^{-1})$ (Sartori, 2003). This implies that an expansion for $F_{\hat\theta}(x)$ of the form (3.20) holds with a different Δ term, which is still of order $O_p(m^{(\alpha-3)/2})$.

In order to obtain an expansion for $M(\hat\theta)$ we follow the same steps as in (A6)–(A9), giving (3.18). In particular, we have

$$b(\hat\theta) = b(\theta) + b_2(\theta) + O_p(m^{\alpha-2})\,,$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

where

$$b_2(\theta) = b_2(\psi, \lambda) \;=\; \sum_{i=1}^{q} b^i_{\psi}\,(\hat{\psi} - \psi) + \sum_{i=1}^{q} b^i_{\lambda_i}(\hat{\lambda}_i - \lambda_i) + \frac{1}{2}\sum_{i=1}^{q} b^i_{\lambda_i\lambda_i}(\hat{\lambda}_i - \lambda_i)^2 + \frac{1}{2}\sum_{i=1}^{q} b^i_{\psi\psi}(\hat{\psi} - \psi)^2 + \sum_{i=1}^{q} b^i_{\psi\lambda_i}(\hat{\lambda}_i - \lambda_i)(\hat{\psi} - \psi)\,. \quad (A10)$$

From Sartori (2003, below formula (9)), with $\alpha > 1$, $\hat{\psi} - \psi = O_p(m^{-1})$, so that the first three summands on the right hand side of the last formula are of order $O_p(m^{\alpha-1})$, while the remaining two are of order $O_p(m^{\alpha-2})$. This leads to

$$M(\hat{\theta}) = M(\theta) + \Delta_1 + O_p\left(m^{-\min\left(1, \frac{5-\alpha}{2}\right)}\right)\,, \quad (A11)$$

where the term Δ_1 is of order $O_p(m^{(\alpha-3)/2})$, as its expected value, because the leading terms in (A10) are of the same order as $b_1(\theta)$ in (A6).

Using (A11) and an expansion similar to (3.19) we obtain

$$x^*(\hat{\theta}) = x^*(\theta) + O_p(m^{(\alpha-3)/2})\,,$$

so that the same error as in (3.20) holds also for unconstrained bootstrap, i.e.

$$F_{\hat{\theta}}(x) = F_{\theta}(x) - \phi(x^*(\theta))\Delta_1 + O_p(m^{-1})\,. \quad (A12)$$

The steps leading from (A12) to (3.11) are the same as those from (3.20) to (3.10).

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

References

Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. Biometrika 73, 307–322.

Bartolucci, F., R. Bellio, A. Salvan, and N. Sartori (2016). Modified profile likelihood for fixed-effects panel data models. Econometric Reviews 35, 1271–1289.

DiCiccio, T. J., T. A. Kuffner, and G. A. Young (2017). The formal relationship between analytic and bootstrap approaches to parametric inference. Journal of Statistical Planning and Inference 191, 81–87.

DiCiccio, T. J., M. A. Martin, and S. E. Stern (2001). Simple and accurate one-sided inference from signed roots of likelihood ratios. The Canadian Journal of Statistics 29, 67–76.

Greene, W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. The Econometrics Journal 7, 98–119.

Lee, S. M. S. and G. A. Young (2005). Parametric bootstrapping with nuisance parameters. Statistics and Probability Letters 71, 143–153.

Pace, L. and A. Salvan (1997). Principles of Statistical Inference from a Neo-Fisherian Perspective. Singapore: World Scientific.

Pierce, D. A. and R. Bellio (2017). Modern likelihood-frequentist inference. International Statistical Review 85, 519–541.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters.

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Biometrika 90, 533–549.

Sartori, N., R. Bellio, A. Salvan, and L. Pace (1999). The directed modified profile likelihood in models with many nuisance parameters. Biometrika 86, 735–742.

Severini, T. A. (2000). Likelihood Methods in Statistics. Oxford: Oxford University Press.

Stern, S. E. (2006). Simple and accurate one-sided inference based on a class of M-estimators. Biometrika 93, 973–987.

Young, G. A. (2009). Routes to higher-order accuracy in parametric inference. Australian & New Zealand Journal of Statistics 51, 115–126.

Young, G. A. and R. L. Smith (2005). Essentials of Statistical Inference. Cambridge: Cambridge University Press.

Department of Economics and Statistics, University of Udine, Udine, 33100, Italy

E-mail: ruggero.bellio@uniud.it

Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK, and

The Alan Turing Institute, London, NW1 2DB, UK

E-mail: ioannis.kosmidis@warwick.ac.uk

Department of Statistical Sciences, University of Padova, 35121, Padova, Italy

E-mail: alessandra.salvan@unipd.it

Department of Statistical Sciences, University of Padova, 35121, Padova, Italy

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

E-mail: nicola.sartori@unipd.it

Statistica Sinica