Statistica Sinica **33** (2023), 1069-1091
doi:https://doi.org/10.5705/ss.202021.0027

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

Ruggero Bellio[1], Ioannis Kosmidis[2,3], Alessandra Salvan[4] and Nicola Sartori[4]

[1] *University of Udine,* [2] *University of Warwick,* [3] *The Alan Turing Institute, and* [4] *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 the 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, using the asymptotic normality of analytical modifications of the signed likelihood root is known to recover the inferential performance. Here, we prove that the parametric bootstrap of standard likelihood pivots results in inferences as accurate as those of analytical modifications of the signed likelihood root do in stratified models with stratum-specific nuisance parameters. We focus on the challenging case in which the number of strata increases as fast or faster than the size of the stratum samples. We further show that this equivalence holds regardless of whether we use the constrained or the unconstrained bootstrap. In contrast, when the number of strata is fixed or increases more slowly than the stratum sample size, we show that using the constrained bootstrap corrects inference to a higher order than when using the unconstrained bootstrap. Simulation experiments support the theoretical findings and demonstrate the excellent performance of the bootstrap in extreme scenarios.

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

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, the score and the Wald statistics. However, 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

Corresponding author: Ruggero Bellio, Department of Economics and Statistics, University of Udine, 33100 Udine, Italy. E-mail: ruggero.bellio@uniud.it.

the modified signed likelihood root (see, for instance, Severini (2000, Chap. 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 extreme scenarios with many nuisance parameters and very limited information (Sartori (2003)).

Parametric bootstrap methods provide an alternative assessment of tail probabilities of likelihood pivots. Furthermore, in standard asymptotic settings, where the number of nuisance parameters is fixed and regularity conditions are satisfied (Severini (2000, Sec. 3.4)), parametric bootstrap methods have been shown to guarantee a level of asymptotic accuracy equivalent to that of analytical modifications of the signed likelihood root (see Young and Smith (2005, Chap. 11)). Two main variants of parametric bootstrap are the constrained bootstrap and the unconstrained bootstrap (also known as the conventional bootstrap). In the latter variant, the sampling distribution of the statistic is computed at the full maximum likelihood estimate, and, in the former, is computed at the constrained maximum likelihood estimate for a given value of the parameter of interest. In standard asymptotic settings, the constrained bootstrap (DiCiccio, Martin and Stern (2001); Lee and Young (2005)) corrects inference about a scalar parameter in the presence of nuisance parameters to a higher order than the unconstrained bootstrap does. 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 nonregular cases in which the modified signed likelihood root is not computable.

We investigate the properties of the parametric bootstrap in models for 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 that $q = o(m)$, whereas 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 an error of order $O_p(m^{(\alpha-1)/2})$ (Sartori (2003, formula (8))), whereas the asymptotic normality fails in the highly stratified case with $\alpha \geq 1$. In this case, the aim of higher-order solutions is to recover the first-order validity of the inferential procedures.

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

order accuracy with the same order of error as that of higher-order analytical solutions.

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

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 examine the case in which 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 the maximum likelihood estimate of θ as $\hat\theta(y) = (\hat\psi(y), \hat\lambda(y)^\top)^\top$ and the constrained maximum likelihood estimate of θ, for fixed ψ, as $\hat\theta_\psi(y) = (\psi, \hat\lambda_\psi(y)^\top)^\top$. We let $U(\theta; y) = \nabla l(\theta; y)$ denote the score vector, and $j(\theta; y) = -\nabla\nabla^\top l(\theta; y)$ denote the observed information, with $i(\theta) = \mathrm{E}_\theta\{j(\theta; Y)\}$ denoting the expected information. The argument θ is dropped when no ambiguity arises, and the components of the vectors and blocks of matrices are denoted by subscripts. For instance, $U_\psi(\theta; y)$ denotes the component of the score vector corresponding to ψ. Furthermore, the argument y is dropped whenever the 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, respectively, of $i(\theta)$, and so on.

The signed likelihood root, the score and the Wald statistics for inference about ψ are

$$R(\psi; y) = \mathrm{sign}\left(\hat\psi(y) - \psi\right)\sqrt{2\left\{l(\hat\theta(y); y) - l(\hat\theta_\psi(y); y)\right\}}, \tag{2.1}$$

$$S(\psi; y) = \frac{U_p(\psi; y)}{\sqrt{i_{\psi\psi|\lambda}(\hat\theta_\psi(y))}}, \tag{2.2}$$

$$T(\psi; y) = (\hat\psi(y) - \psi)\sqrt{j_p(\hat\psi(y); y)}, \tag{2.3}$$

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. Although (2.1) and (2.2) are invariant under reparameterizations that preserve the parameter of interest, (2.3) is not.

Computing the p-values and confidence intervals for ψ requires the distribu-

tion of the 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, $\mathrm{pr}_\theta\{R(\psi) \leq R(\psi;y)\} = \Phi(R(\psi;y))\{1 + O(n^{-1/2})\}$, where $\Phi(\cdot)$ denotes the standard normal distribution function. The accuracy can be improved by using higher-order modifications $R^*(\psi;y)$ of $R(\psi;y)$, such that $\mathrm{pr}_\theta\{R(\psi) \leq R(\psi;y)\} = \Phi(R^*(\psi;y))\{1 + O(n^{-1})\}$. 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, several alternative versions of $R^*(\psi;y)$ have been proposed (see Pierce and Bellio (2017), for an overview).

An alternative to the asymptotic approximations of the distributions of (2.1), (2.2), and (2.3) is the 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) the unconstrained bootstrap, where samples are simulated from the model at $\hat\theta(y)$, and ii) the constrained bootstrap, where samples are simulated at $\hat\theta_\psi(y)$ (see DiCiccio, Martin and Stern (2001); Lee and Young (2005); Young and Smith (2005, Chap. 11)).

In standard asymptotic settings, the 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 we simulate samples y^k from the model with parameter $\hat\theta(y)$, for $k = 1, \ldots, K$, then the 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 an error of order $O(n^{-1})$, that is,

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

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

In contrast, the constrained bootstrap provides third-order accuracy. If we simulate samples y^k from the model with parameter $\hat\theta_\psi(y)$, for $k = 1, \ldots, K$, the p-values for (2.1), calculated as

$$\hat p_2^R(\psi) = \frac{1}{K}\sum_{k=1}^{K} I\{R(\psi; y^k) \leq R(\psi;y)\}\,, \tag{2.6}$$

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

is,

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

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, Sec. 11.4) note, the theoretical advantage of using the constrained over the unconstrained bootstrap is rarely supported by numerical evidence, because both types improve equally over the first-order results.

The advantage of the bootstrap p-values in (2.4) and (2.6) over using analytical modifications to common statistics is that the bootstrap does not require any additional, often tedious, algebraic derivation and implementation of the necessary modifications. Moreover, there are nonstandard modeling settings, in which $R(\psi; y)$ is computable, whereas $R^*(\psi; y)$ 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 the bootstrap is the additional computation typically required for the repeated model fits, which can be partly mitigated by parallel computing.

In some special cases, the distributions of (2.1), (2.2), and (2.3) depend only on ψ, so that the 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, Sec. 1.3)). For instance, if y is a realization of $Y = (Y_1, \ldots, Y_n)^\top$ with independent and identically distributed (i.i.d.) components from a shape and scale model with generic density

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

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, owing to the 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 $Y_1^0, \ldots, Y_n^0$. 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

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

which depends only on ψ.

An example with a stratified gamma model is provided in the Supplementary Material, where simulation results confirm the exactness of the constrained bootstrap.

3. Two-Index Asymptotic Theory for Stratified Models

We consider a stratified setting with q independent strata, each with m observations. 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 $y_i = (y_{i1}, \ldots, y_{im})^\top$, for $i = 1, \ldots, q$, 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 the 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 $q = O(m^\alpha)$, for $\alpha > 0$. The case $\alpha = 0$ corresponds to the standard asymptotic setting. Sartori (2003, Sec. 4) showed that $R(\psi)$, $S(\psi)$, and $T(\psi)$ are asymptotically equivalent to order $o_p(1)$ for $\alpha \geq 0$. Specifically, when $0 \leq \alpha < 1$, the equivalence of the three quantities holds with a 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 $\alpha \geq 1$, the asymptotic equivalence of $R(\psi)$, $S(\psi)$, and $T(\psi)$ holds with an error of order $O_p(m^{-1})$. 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 the strata profile scores. However, the same results hold for $R(\psi)$ and $T(\psi)$, because 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 main result of this study is that the asymptotic validity of both the constrained and the 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 that required for the validity of inference based on the modified signed likelihood root $R^*(\psi)$ (Sartori (2003)). In particular, we show that, when $0 < \alpha < 1$,

$$\mathrm{pr}_\theta\left(F_{\hat{\theta}_\psi}(S(\psi)) \le u\right) = u + O(m^{(\alpha-3)/2}) \tag{3.1}$$

and

$$\mathrm{pr}_\theta\left(F_{\hat{\theta}}(S(\psi)) \le u\right) = u + O(m^{-1})\,, \tag{3.2}$$

whereas, when $1 \le \alpha < 3$,

$$\mathrm{pr}_\theta\left(F_{\hat{\theta}_\psi}(S(\psi)) \le u\right) = u + O(m^{(\alpha-3)/2}) \tag{3.3}$$

and

$$\mathrm{pr}_\theta\left(F_{\hat{\theta}}(S(\psi)) \le u\right) = u + O(m^{(\alpha-3)/2})\,. \tag{3.4}$$

Hence, when $1 \le \alpha < 3$, the same order of error is obtained with both the constrained bootstrap and the unconstrained bootstrap, in contrast to what happens when $0 \le \alpha < 1$. The case $\alpha = 0$ corresponds to the standard asymptotic setting in which $n = O(m)$, and (3.1) and (3.2) reduce to (2.7) and (2.5), respectively. Intuitively, the reason why the two types of bootstrap have the same accuracy when $\alpha \ge 1$ is because 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 present when $0 \le \alpha < 1$. A formal development of the result is given below.

In the following, we concentrate on the more extreme case, $\alpha \ge 1$. The proof of (3.1) and (3.2) for the case $0 < \alpha < 1$ is given in the Supplementary Material. In order to prove both (3.3) and (3.4), 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)), $U_p = U_p(\psi)$ can be expanded as

$$U_p = U_{\psi|\lambda} + B + Re\,, \tag{3.5}$$

where $U_{\psi|\lambda} = U_\psi - i_{\psi\lambda} i_{\lambda\lambda}^{-1} U_\lambda = O_p(\sqrt{n}) = O_p(m^{(\alpha+1)/2})$, with zero mean and variance $i_{\psi\psi|\lambda}$, $B = B(\theta) = O_p(m^\alpha)$, and, when $\alpha > 1$, $Re = O_p(m^{\alpha-1})$. Details about the orders in (3.5) are provided in the Appendix. When $0 \le \alpha < 1$, the terms in (3.5) are in descending order. When $1 \le \alpha < 3$, B becomes the leading term, followed by $U_{\psi|\lambda}$. Finally, when $\alpha \ge 3$, $U_{\psi|\lambda}$ is dominated both by B and by Re. In practice, when $1 \le \alpha < 3$, 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 the expansion in (3.5).

Let $M(\theta) = \mathrm{E}_\theta(S(\psi))$ and $\mathrm{Var}_\theta(S(\psi))$ be the expectation and variance, re-

spectively, of $S(\psi)$. The asymptotic expansions in the Appendix can be used to show that

$$M(\theta) = \frac{b(\theta)}{i_{\psi\psi|\lambda}(\theta)^{1/2}} + M_1(\theta) + O(m^{(\alpha-5)/2}) \tag{3.6}$$

$$\mathrm{Var}_\theta(S(\psi)) = 1 + v(\theta) + O\left(\frac{1}{m^2}\right), \tag{3.7}$$

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 $S(\psi)$ admits a valid Edgeworth expansion. Severini (2000, Sec. 5.1–5.4) gives the conditions and details for the extension of the Edgeworth expansions for i.i.d. random variables to likelihood pivots, such as $R(\psi)$, $S(\psi)$, and $T(\psi)$. 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

$$F_\theta(x) = \mathrm{pr}_\theta\left(S(\psi) \le x\right) = \Phi\left(\frac{x - M(\theta)}{\sqrt{\mathrm{Var}_\theta(S(\psi))}}\right) + O(m^{-(\alpha+1)/2}), \tag{3.8}$$

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,

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

and

$$F_\theta(x) = \Phi\left(x^*(\theta)\right) + O\left(m^{-\min(2,(\alpha+1)/2)}\right). \tag{3.9}$$

We first focus on the constrained bootstrap. From (3.9),

$$F_{\hat\theta_\psi}(x) = \Phi\left(x^*(\hat\theta_\psi)\right) + O_p\left(m^{-\min(2,(\alpha+1)/2)}\right). \tag{3.10}$$

The Taylor expansions in the Appendix give

$$M(\hat\theta_\psi) = M(\theta) + \Delta + O_p\left(m^{-\min(1,(5-\alpha)/2)}\right) \tag{3.11}$$

and

$$v(\hat\theta_\psi) = v(\theta) + O_p(m^{-2}), \tag{3.12}$$

where $\Delta = O_p(m^{(\alpha-3)/2})$ is given in expression (A.9) of the Appendix. Using

(3.11) and (3.12), 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 (3.10) holds:

$$F_{\hat{\theta}_\psi}(x) = F_\theta(x) - \phi(x^*(\theta))\Delta + O_p(m^{-1}), \tag{3.13}$$

where the error is of order $O_p(m^{-1})$, because, for $\alpha < 3$, $\min(1, (5-\alpha)/2) = 1$, whereas the error term in (3.10) is $o_p(m^{-1})$ whenever $\alpha > 1$.

In order to prove (3.3), note that $F_{\hat{\theta}_\psi}(S(\psi)) \le u$ is equivalent to $S(\psi) \le s_u$, 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 (3.13),

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

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

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

we get

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

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

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

with $\bar{S}(\psi) = S(\psi) - \Delta + O_p(m^{\alpha-3}) + O_p(m^{-1})$, and where Δ is given by (A.9) and is such that $\mathrm{E}_\theta(\Delta) = O(m^{(\alpha-3)/2})$. Moreover, we have

$$\begin{aligned}
\mathrm{E}_\theta(\bar{S}(\psi)) &= \mathrm{E}_\theta(S(\psi)) + O(m^{(\alpha-3)/2}), &(3.14)\\
\mathrm{Var}_\theta(\bar{S}(\psi)) &= \mathrm{Var}_\theta(S(\psi) - \Delta) + O(m^{-2})\\
&= \mathrm{Var}_\theta(S(\psi)) + \mathrm{Var}_\theta(\Delta) - 2\mathrm{Cov}_\theta(S(\psi), \Delta) + O(m^{-2})\\
&= \mathrm{Var}_\theta(S(\psi)) + O(m^{-2}), &(3.15)
\end{aligned}$$

because $\mathrm{Var}_\theta(\Delta) = O(m^{-2})$ and $\mathrm{Cov}_\theta(S(\psi), \Delta) = O(m^{-2})$, where the order of the latter is determined by the orthogonality between $U_{\psi|\lambda}$ and the leading term

of $b_1(\theta)$ in (A.7). Finally, (3.3) holds because

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

The proof of (3.4) for the unconstrained bootstrap is obtained using the same steps as above. In particular, the expansion (A.12) holds for $F_{\hat{\theta}}(x)$, having the same form as (3.13), with Δ replaced by Δ_1, which is still of order $O_p(m^{(\alpha-3)/2})$. Details are given in the Appendix. However, although (3.14) is still true, (3.15) 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 (A.10). Therefore, for the unconstrained bootstrap, we have

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

Hence, when $\alpha \ge 1$, the errors in (3.3) and (3.4) 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 (3.3) and (3.4) suggest that the location and scale adjustments to the statistic, as done for $R(\psi)$ in a standard asymptotic setting by DiCiccio, Martin and Stern (2001) and Stern (2006), are the key requirement to recovering the approximate uniformity of the p-values. In this respect, a bootstrap location and scale adjustment of $R(\psi)$, $S(\psi)$, or $T(\psi)$ is expected to be as effective as bootstrapping the distribution of the statistic. This conjecture is confirmed by the numerical results in the following section and in the Supplementary Material.

4. Simulation Studies

The finite-sample properties of the unconstrained and constrained parametric bootstraps are assessed using 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

Table 1. Statistics compared in the simulation experiments. The mean $\tilde{\mu}^R$ and the standard deviation $\tilde{\sigma}^R$ of $R(\psi)$ are estimated using the 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_l^c	Location adjusted $R(\psi)$
$(R(\psi) - \tilde{\mu}^R)/\tilde{\sigma}^R$	R_{ls}^c	Location-and-scale adjusted $R(\psi)$

for further models reported in the Supplementary Material. For each model, we conduct nine simulation experiments, one for each combination of the number of strata $q \in \{10, 100, 1000\}$ and the stratum sample size $m \in \{4, 8, 16\}$.

Each simulation experiment involves 10,000 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 six bootstrap-based p-values are computed to test $\psi = \psi_0$. In particular, we compute the following statistics: i) $R(\psi)$, $S(\psi)$, and $T(\psi)$; ii) the location- and location-and-scale-adjusted versions of $R(\psi)$, $S(\psi)$, and $T(\psi)$, where the mean and variance of each statistic are estimated using the unconstrained bootstrap (at $\hat{\theta}$) and the constrained boostrap (at $\hat{\theta}_\psi$); and iii) $R^*(\phi)$ and the signed likelihood root computed from the modified profile likelihood (see, for instance, Severini (2000, Chap. 8)). The higher-order adjustment required for the latter two statistics is obtained using the expected moments of likelihood quantities, as in Severini (2000, Sec. 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 (2.6), respectively.

To conserve space, we report only the results for the six 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 presented only 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 for all of the simulation experiments are provided in the Supplementary Material.

The above experiments involve high-dimensional parameter spaces with as many as 1,000 nuisance parameters. As a result, the assessment of the statistics

requiring bootstrapping is demanding in terms of computational time and cost, even when using parallel computing with a large number of cores. Therefore, the number of bootstrap samples is limited to 1,000 in all simulation experiments.

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

Table 2. Empirical tail probabilities $\times 100$ for the statistics in Table 1 and all models considered in the simulation studies of Section 4. The figures shown are rounded to one decimal place, and are for $q = 1000$ and $m = 8$.

Model	Statistic	Nominal					
		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

4.1. Beta model

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

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

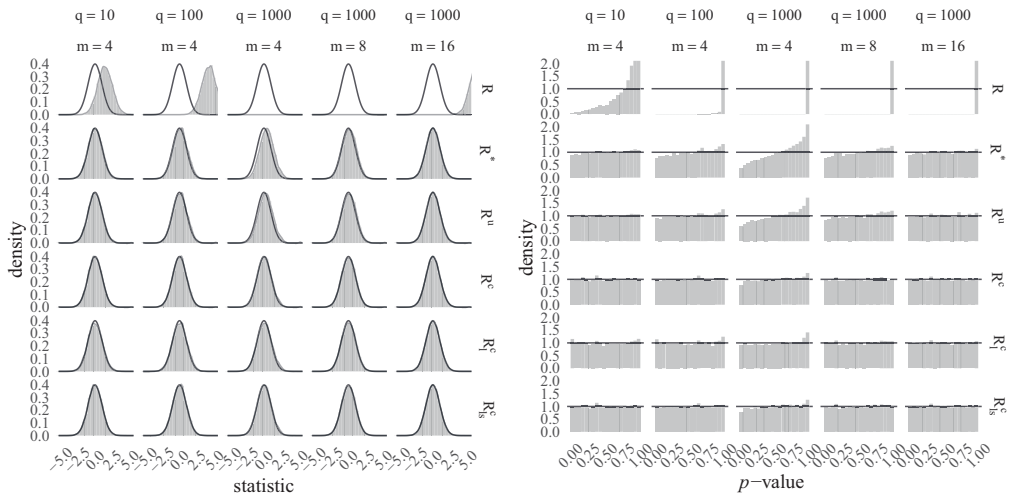

Figure 1. Beta model. Estimated null distribution of the statistics (left) and estimated distribution of the p-values (right) for the statistics in Table 1 for various combinations of q and m. The $\mathrm{N}(0,1)$ and $\mathrm{Uniform}(0,1)$ density functions are superimposed for the statistics (left)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

where $B(\cdot)$ is the beta function. The parameter of interest is $\psi = \log \phi$, and 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 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 also valid for the constrained and unconstrained bootstrap p-values, because they have been mapped onto the standard normal scale using 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 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 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.

The right panel of Figure 1 shows 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 $R(\psi)$ close to the standard normal appears to be mainly a location adjustment.

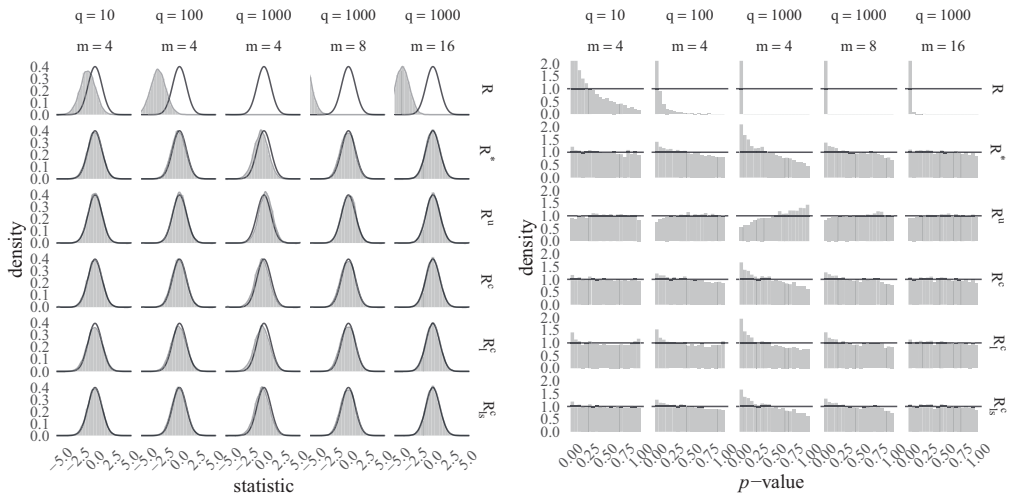

Figure 2. Curved exponential family model. Estimated null distribution of the statistics (left) and estimated distribution of the p-values (right) for the statistics in Table 1 for various combinations of q and m. The $\mathrm{N}(0,1)$ and $\mathrm{Uniform}(0,1)$ density functions are superimposed for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

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), who point 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 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. As in the previous example, the empirical, finite-sample distributions of $R(\psi)$ are far from standard normal, whereas all the 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 that no statistic appears to perform well for $(q, m) = (1000, 4)$; see also the empirical tail probabilities in Table 2.

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 vari-

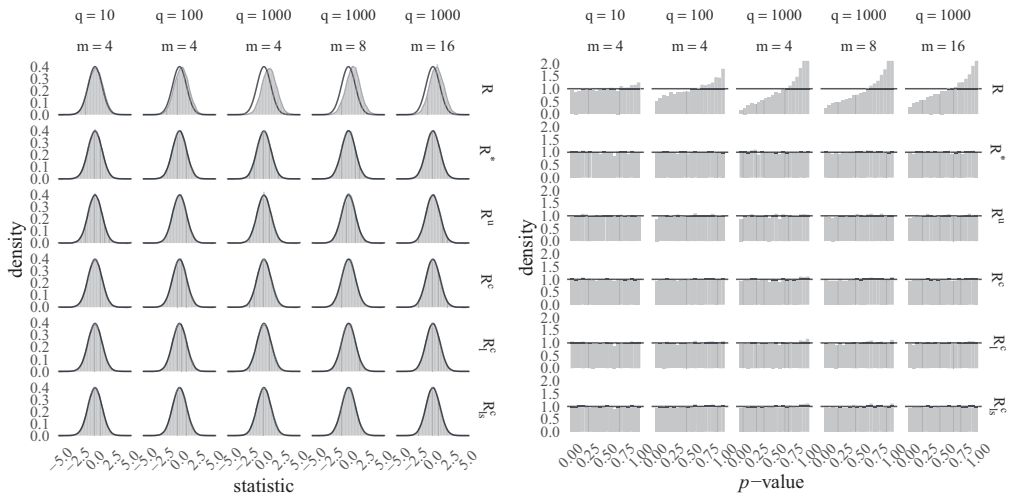

Figure 3. Truncated linear regression model. Estimated null distribution of the statistics (left) and estimated distribution of the p-values (right) for the statistics in Table 1 for various combinations of q and m. The $\mathrm{N}(0,1)$ and $\mathrm{Uniform}(0,1)$ density functions are superimposed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

able Y_{ij} to be distributed as Y_{ij}^* conditionally on $Y_{ij}^* > 0$, with

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

where the error term ε_{ij} follows a standard normal distribution. For the simulation study, we use $\psi = 1$, and the elements of λ_0 and x_{ij} are both independently generated from a standard normal distribution and held fixed over all 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. In contrast to the other examples, the distribution of the first-order statistics requires only a moderate adjustment, even in the most extreme settings. Furthermore, both bootstrap-based statistics and the $R^*(\psi)$ statistic perform rather well, providing results very close to the target distributions.

5. Conclusion

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

Unconstrained and constrained bootstrap for the signed likelihood ratio root, the score and the Wald statistics can both recover inferential performance in stratified settings when $q = O(m^\alpha)$, for $0 < \alpha < 3$. As in the case of $\alpha = 0$ (Lee and Young (2005)), when $0 < \alpha < 1$, the constrained bootstrap has a higher degree of asymptotic accuracy than that of the unconstrained bootstrap. On the other hand, the two bootstraps are asymptotically equivalent when $1 \leq \alpha < 3$. The condition $q = O(m^\alpha)$, for $0 < \alpha < 3$, is the same as that in Sartori (2003) for the validity of inference based on R^* and on the signed likelihood root computed from the modified profile likelihood.

The results in Section 4 from the simulation studies for the finite-sample assessment of the performance of the constrained bootstrap and the unconstrained bootstrap are in line with those expected from the theory. In extreme settings, such as the beta model with $(q, m) = (1000, 4)$, the constrained bootstrap appears to perform slightly better than the unconstrained bootstrap. Furthermore, in all simulation experiments, as q/m diverges, the inferential performance of the constrained bootstrap and the unconstrained bootstrap of the first-order statistics seems to deteriorate more slowly than that of R^* and the signed likelihood root computed from the modified profile likelihood (see also the Supplementary Material). As a result, the evidence from the simulation studies indicates that inference from the parametric bootstrap is more resilient to increasing q/m than that of the analytically available higher-order statistics, with the constrained bootstrap being the most accurate in extreme scenarios.

The theoretical developments in this study do not cover situations in which the random variables have discrete support, because the Edgeworth expansion in (3.8) is only valid for models with continuous support. The impact of discreteness on the performance of the parametric bootstrap is examined in the Supplementary Material using 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 Material indicate that the equivalence between the unconstrained bootstrap and the constrained bootstrap of the first-order statistics in continuous models may not hold for discrete settings. In those cases, despite the unconstrained bootstrap appearing to deliver a marked inferential improvement to using first-order statistics, the constrained bootstrap, similarly to R^*, is found to perform considerably better for most combinations of q and m.

The simulation experiments reported here were carried out with 1,000 boot-

strap replications. This value is smaller than the recommendations of millions of replications in the literature for standard asymptotics settings (Young (2009); DiCiccio, Kuffner and Young (2017)). For stratified settings with $\alpha > 1$, the bootstrap adjustments recover the asymptotic uniformity of the p-values, instead of providing a small-sample refinement of p-values that are already asymptotically uniform. As a result, using a huge number of bootstrap replications is less essential. This is supported by the few experiments we carried out with more than 1,000 bootstrap replications. More comprehensive simulation studies to support this statement are unfortunately not feasible with current computing capabilities.

Supplementary Material

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

Acknowledgments

We thank two anonymous referees and the associate editor for their constructive comments. The work of Ioannis Kosmidis was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. Part of this work was completed during a visit by Ruggero Bellio, Alessandra Salvan, and Nicola Sartori to The Alan Turing Institute. The work of Alessandra Salvan and Nicola Sartori was supported by the University of Padova under grants BIRD185955 and BIRD203991.

Appendix

Asymptotic orders in (3.5)

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

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

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 (3.5) 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 (A.1), 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 $B = \sum_{i=1}^q B^i(\psi, \lambda_i) = O_p(m^\alpha)$, 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 (3.6) **and** (3.7)

As a first step, consider the expansion

$$i_{\psi\psi|\lambda}(\hat\theta_\psi) = i_{\psi\psi|\lambda} + C + O_p(m^{\alpha-1}), \tag{A.2}$$

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 (A.1). Hence,

$$\{i_{\psi\psi|\lambda}(\hat\theta_\psi)\}^{-1/2} = i_{\psi\psi|\lambda}^{-1/2}\left\{1 - \frac{1}{2}\frac{C}{i_{\psi\psi|\lambda}} + O_p(m^{-2})\right\}, \tag{A.3}$$

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

Using (3.5) and (A.3),

$$\begin{aligned} S(\psi) &= i_{\psi\psi|\lambda}^{-1/2}\left\{U_{\psi|\lambda} + B + Re\right\}\left\{1 - \frac{1}{2}\frac{C}{i_{\psi\psi|\lambda}} + O_p(m^{-2})\right\} \\ &= \frac{U_{\psi|\lambda}}{i_{\psi\psi|\lambda}^{1/2}} + \frac{B}{i_{\psi\psi|\lambda}^{1/2}} + \frac{Re}{i_{\psi\psi|\lambda}^{1/2}} - \frac{1}{2}\frac{U_{\psi|\lambda}C}{i_{\psi\psi|\lambda}^{3/2}} - \frac{1}{2}\frac{B\,C}{i_{\psi\psi|\lambda}^{3/2}} - \frac{1}{2}\frac{Re\,C}{i_{\psi\psi|\lambda}^{3/2}} \\ &\quad + O_p(m^{-2}) + O_p(m^{(\alpha-5)/2}) + O_p(m^{(\alpha-7)/2}), \end{aligned} \tag{A.4}$$

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 (A.3). Its expectation is of order $O(m^{(\alpha-5)/2})$. The orders of terms in (A.4) are as follows:

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

Expansion (3.6) for $\mathrm{E}_\theta(S(\psi))$ is obtained using (A.4) and recalling that $b(\theta) = O(m^\alpha)$. We have

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

giving (3.6) with

$$M_1(\theta) = \mathrm{E}_\theta\left(\frac{Re}{i_{\psi\psi|\lambda}^{1/2}}\right) + \mathrm{E}_\theta\left(\frac{1}{2}\frac{B\,C}{i_{\psi\psi|\lambda}^{3/2}}\right) = O(m^{(\alpha-3)/2})\,. \tag{A.5}$$

Expansion (3.7) for $\mathrm{Var}_\theta(S(\psi))$ is also obtained using (A.4). 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 (A.4) 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 (3.11)**) and** (3.12)

Let $\overline{Re} = \mathrm{E}_\theta(Re)$ and $\overline{BC} = \mathrm{E}_\theta(B\,C)$. Then, from (3.6) and (A.5),

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

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,

$$b(\hat{\theta}_\psi) = b(\theta) + b_1(\theta) + O_p(m^{\alpha-2}) \,, \tag{A.6}$$

where

$$b_1(\theta) = \sum_{i=1}^{q} b^i_{\lambda_i}(\psi, \lambda_i)(\hat{\lambda}_{i\psi} - \lambda_i) + \frac{1}{2} \sum_{i=1}^{q} b^i_{\lambda_i\lambda_i}(\psi, \lambda_i)(\hat{\lambda}_{i\psi} - \lambda_i)^2 \,, \tag{A.7}$$

and $b^i_{\lambda_i}(\psi, \lambda_i) = \partial b^i(\psi, \lambda_i)/\partial\lambda_i$, and so on. Using (A.1), 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 (A.6) 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 (A.3), we get

$$M(\hat{\theta}_\psi) = i^{-1/2}_{\psi\psi|\lambda} b(\theta) + \tilde{M}_1 + O_p\left(m^{-\min\{1,(5-\alpha)/2\}} \right) \,, \tag{A.8}$$

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, (A.8), (3.6) and (A.5) give (3.11) with

$$\Delta = \tilde{M}_1 - M_1(\theta) = \frac{b_1(\theta)}{i^{1/2}_{\psi\psi|\lambda}} - \frac{C\, b(\theta)}{2\, i^{3/2}_{\psi\psi|\lambda}} \tag{A.9}$$

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

To obtain expansion (3.12) recall that in (3.7)

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

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 (A.2),

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

which gives (3.12).

Derivation of (3.4)

First, from (3.9), we have

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

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 (3.13) 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 (A.6)–(A.9), giving (3.11). In particular, we have

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

where

$$\begin{aligned} b_2(\theta) = b_2(\psi,\lambda) = &\sum_{i=1}^{q} b^i_{\psi}\,(\hat{\psi}-\psi) + \sum_{i=1}^{q} b^i_{\lambda_i}(\hat{\lambda}_i - \lambda_i) + \frac{1}{2}\sum_{i=1}^{q} b^i_{\lambda_i\lambda_i}(\hat{\lambda}_i-\lambda_i)^2 \\ &+\frac{1}{2}\sum_{i=1}^{q} b^i_{\psi\psi}(\hat{\psi}-\psi)^2 + \sum_{i=1}^{q} b^i_{\psi\lambda_i}(\hat{\lambda}_i-\lambda_i)(\hat{\psi}-\psi)\,. \end{aligned} \tag{A.10}$$

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

$$M(\hat{\theta}) = M(\theta) + \Delta_1 + O_p\left(m^{-\min(1,(5-\alpha)/2)}\right)\,, \tag{A.11}$$

where the term Δ_1 is of order $O_p(m^{(\alpha-3)/2})$, as its expected value, because the leading terms in (A.10) are of the same order as $b_1(\theta)$ in (A.6).

Using (A.11) and an expansion similar to (3.12) we obtain

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

so that the same error as in (3.13) holds also for unconstrained bootstrap, i.e.

$$F_{\hat{\theta}}(x) = F_\theta(x) - \phi(x^*(\theta))\Delta_1 + O_p(m^{-1}). \tag{A.12}$$

The steps leading from (A.12) to (3.4) are the same as those from (3.13) to (3.3).

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

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

DiCiccio, T. J., Martin, M. A. and Stern, S. E. (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 Young, G. A. (2005). Parametric bootstrapping with nuisance parameters. *Statistics and Probability Letters* **71**, 143–153.

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

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

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90**, 533–549.

Sartori, N., Bellio, R., Salvan, A. and Pace, L. (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 University Press, Oxford.

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 Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge University Press, Cambridge.

Ruggero Bellio

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

E-mail: ruggero.bellio@uniud.it

Ioannis Kosmidis
Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK.
The Alan Turing Institute, London, NW1 2DB, UK.
E-mail: ioannis.kosmidis@warwick.ac.uk
Alessandra Salvan
Department of Statistical Sciences, University of Padova, 35121 Padova, Italy.
E-mail: alessandra.salvan@unipd.it
Nicola Sartori
Department of Statistical Sciences, University of Padova, 35121 Padova, Italy.
E-mail: nicola.sartori@unipd.it

(Received January 2021; accepted September 2021)