Journal of Applied Statistics

Taylor & Francis
Taylor & Francis Group

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Nonparametric bootstrapping for hierarchical data

Shiquan Ren , Hong Lai , Wenjing Tong , Mostafa Aminzadeh , Xuezhang Hou & Shenghan Lai

To cite this article: Shiquan Ren , Hong Lai , Wenjing Tong , Mostafa Aminzadeh , Xuezhang Hou & Shenghan Lai (2010) Nonparametric bootstrapping for hierarchical data, Journal of Applied Statistics, 37:9, 1487-1498, DOI: 10.1080/02664760903046102

To link to this article: http://dx.doi.org/10.1080/02664760903046102

Published online: 06 Sep 2010.

Journal of Applied Statistics
Vol. 37, No. 9, September 2010, 1487–1498

Taylor & Francis
Taylor & Francis Group

Nonparametric bootstrapping for hierarchical data

Shiquan Ren[a*], Hong Lai[b], Wenjing Tong[b], Mostafa Aminzadeh[c], Xuezhang Hou[c] and Shenghan Lai[b]

[a]*School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, Australia;* [b]*Departments of Radiology and Pathology, Johns Hopkins University School of Medicine, Baltimore, USA;* [c]*Mathematics Department of Towson University, Towson, USA*

(*Received 29 May 2008; final version received 13 May 2009*)

Nonparametric bootstrapping for hierarchical data is relatively underdeveloped and not straightforward: certainly it does not make sense to use simple nonparametric resampling, which treats all observations as independent. We have provided some resampling strategies of hierarchical data, proved that the strategy of nonparametric bootstrapping on the highest level (randomly sampling all other levels without replacement within the highest level selected by randomly sampling the highest levels with replacement) is better than that on lower levels, analyzed real data and performed simulation studies.

Keywords: random effects model; hierarchical data; nonparametric bootstrapping; resampling schemes; unbalanced data

1. Introduction

Hierarchical data, such as environmetric data, often include multiple sources of variation that can be described using a hierarchical or multilevel model [7,9]. When the model is sufficiently simple (e.g., a linear mixed model in which all random effects have a normal distribution with constant variance), the data can be expressed with a variance–covariance matrix that depends on the variance components. It may be important when setting up a resampling scheme to take careful account of the multiple sources of variation depending on the nature of the parameter being estimated.

Recent years have seen the widespread application of hierarchical and other random effect models. In the parametric setting, it is often most natural to take a Bayesian point of view and to fit the parameters using Markov chain Monte Carlo algorithms, and then uncertainty analysis for both parameters and random effects is straightforwardly tackled using simulation output, at least in principle. In practice this may be unsatisfactory, either because standard parametric models fit poorly or because of concerns about the impact of the assumed priors on inferences. The

[*] Corresponding author. Email: s.ren@unsw.edu.au

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664760903046102
http://www.informaworld.com

specification of prior distributions can be avoided by taking a frequentist approach, under which fitting is typically performed using an expectation–maximization algorithm, possibly stochastic, and it would then seem natural to try and use a bootstrap procedure for uncertainty analysis [2]. There are two useful bootstrap procedures for nonparametric and parametric inferences [3,4]. Ren *et al.* [11] found parametric bootstrap inference for the intraclass correlation coefficients of hierarchical binary outcomes is not good, while nonparametric bootstrapping for hierarchical data is relatively underdeveloped [2].

Nonparametric bootstrapping for hierarchical data is complicated by the need to estimate empirical distribution functions for two (or more) random variables. Davison and Hinkley [1] discussed a few of the basic points about nonparametric resampling in a relatively simple context – balanced one-way nested classification. Field and Welsh [5] described the asymptotic behavior of bootstrap moments that depends on the convergence of sample statistics to population quantities of the random effects model for the simple two-level data (balanced one-way nested classification). For hierarchical unbalanced data having more than two levels, there are more than two bootstrap resampling strategies. The resampling strategy that works best for three-level (or multilevel) data, such as the structure with regions, families, members [7,9] has not yet been determined. Because nonparametric bootstrapping for hierarchical data is not straightforward: certainly it does not make sense to use simple nonparametric resampling, which treats all observations as independent. Although procedures can be derived for specific cases [10,11], there is currently no general nonparametric method for bootstrapping hierarchical data.

Our aim in this paper is to address some nonparametric bootstrap resampling strategies of hierarchical or multilevel data. In Section 2, we provide some resampling strategies of hierarchical data and prove that the strategy of nonparametric bootstrapping of three-level data (or unbalanced two-way nested classification) on the highest level (randomly sampling without replacement within the highest level selected by randomly sampling the highest levels with replacement) is better than that on the lower levels. Section 3 contains analysis of a real example. A simulated study to the two-level data (unbalanced one-way nested classification) is performed in Section 4. In Section 5, we extended the study to four-level data (unbalanced three-way nested classification).

2. Nonparametric bootstrapping for three-level data

For observed values with a hierarchical structure, such as regions, families, members, etc., we assume that there are l regions, m_i families in the ith region and n_{ij} members in the jth family in the ith region. Response y_{ijk} is the observed value of the kth member, the jth family and the ith region, $N = \sum\sum n_{ij}$, such that

$$Y_{ijk} = \mu + a_i + b_{ij} + e_{ijk}, \quad i = 1, 2, \ldots, l;\ j = 1, 2, \ldots, m_i;\ k = 1, 2, \ldots, n_{ij}, \tag{1}$$

where $\mu = E(Y_{ijk})$, the highest level (region) effects $\{a_i\}$, the second level (family) effects $\{b_{ij}\}$ and the residual effects $\{e_{ijk}\}$ are independently normally distributed with mean 0 and variance σ_a^2, σ_b^2 and σ_e^2, respectively. The variance of Y_{ijk}, the covariance of Y_{ijk} and Y_{ist} and the covariance of Y_{ijk} and Y_{ijt} are, respectively:

$$\begin{aligned} \sigma_y^2 &= \sigma_a^2 + \sigma_b^2 + \sigma_e^2, \\ \mathrm{Cov}(Y_{ijk}, Y_{ist}) &= \sigma_a^2,\ j \neq s, \\ \mathrm{Cov}(Y_{ijk}, Y_{ijt}) &= \sigma_a^2 + \sigma_b^2,\ k \neq t. \end{aligned} \tag{2}$$

This random effects model without independent variables is the same as the variance components model which is a kind of hierarchical linear model.

There are three simple strategies, for all of which the first stage is to randomly sample regions with replacement. The second stage is to randomly sample families within the regions selected at the first stage, either without replacement or with replacement. The third stage is to randomly sample members without replacement (Strategy 1) within the families selected without replacement at the second stage and the regions selected at the first stage; and to randomly sample members within the families selected with replacement at the second stage and the regions selected at the first stage, either without replacement (Strategy 2) or with replacement (Strategy 3). In short, there are three types of simple nonparametric samplings from hierarchical data with member–family–region structure, that is, randomly sampled data from members (Strategy 3), families (Strategy 2) or regions (Strategy 1), respectively. To see which strategy is likely to work best, one looks at the second moments of resampling data Y^*_{ijk} to see how well they match Equation (2).

Consider selecting Y^*_{ijk}. At the first stage, we select a random integer I^* from $\{1, 2, \dots, l\}$. At the second stage, we select m_{i^*} random integers J^* from $\{1, 2, \dots, m_{i^*}\}$, either without replacement or with replacement. At the third stage, we select $n_{i^*j^*}$ random integers K^* from $\{1, 2, \dots, n_{i^*j^*}\}$ without replacement within the families selected without replacement at the second stage and the regions selected at the first stage (Strategy 1); and either without replacement (Strategy 2) or with replacement (Strategy 3) within the families selected with replacement at the second stage and the regions selected at the first stage. Under the three strategies, $E(Y^*_{ijk}|I^* = i^*, J^* = j^*) = \bar{y}_{i^*j^*}$, $E(Y^{*2}_{ijk}|I^* = i^*, J^* = j^*) = \sum_{k=1}^{n_{i^*j^*}} y^2_{i^*j^*k}/n_{i^*j^*}$,

$$E(Y^*_{ijk}Y^*_{ijt}|I^* = i^*, J^* = j^*) = \begin{cases} \dfrac{\sum_{1\le k\ne t\le n_{i^*j^*}} y_{i^*j^*k}y_{i^*j^*t}}{(n_{i^*j^*}(n_{i^*j^*}-1))} & \text{Strategy 2,} \\[2ex] \dfrac{\sum_{k,t=1}^{n_{i^*j^*}} y_{i^*j^*k}y_{i^*j^*t}}{n^2_{i^*j^*}} & \text{Strategy 3.} \end{cases}$$

So, $E(Y^*_{ijk}|I^* = i^*) = \bar{y}_{i^*}$, $E(Y^{*2}_{ijk}|I^* = i^*) = \sum_{j=1}^{m_{i^*}} \sum_{k=1}^{n_{i^*j}} y^2_{i^*jk}/\sum_{j=1}^{m_{i^*}} n_{i^*j}$,

$$E(Y^*_{ijk}Y^*_{ist}|I^* = i^*) = \begin{cases} \dfrac{\sum_{1\le j\ne s\le m_{i^*}} \sum_{1\le k\le n_{i^*j},1\le t\le n_{i^*s},k\ne t} y_{i^*jk}y_{i^*st}}{\sum_{1\le j\ne s\le m_{i^*}} \sum_{1\le k\le n_{i^*j},1\le t\le n_{i^*s},k\ne t}} & \text{Strategy 1,} \\[2ex] \dfrac{\sum_{j,s=1}^{m_{i^*}} \sum_{1\le k\le n_{i^*j},1\le t\le n_{i^*s},k\ne t} y_{i^*jk}y_{i^*st}}{\sum_{j,s=1}^{m_{i^*}} \sum_{1\le k\le n_{i^*j},1\le t\le n_{i^*s},k\ne t}} & \text{Strategy 2,} \\[2ex] \dfrac{\sum_{j,s=1}^{m_{i^*}} \sum_{k=1}^{n_{i^*j}} \sum_{t=1}^{n_{i^*s}} y_{i^*jk}y_{i^*st}}{\sum_{j=1}^{m_{i^*}} n_{i^*j} \sum_{s=1}^{m_{i^*}} n_{i^*s}} & \text{Strategy 3.} \end{cases}$$

Therefore,

$$E(Y^*_{ijk}) = \bar{y}_{\cdot}, \quad \mathrm{Var}(Y^*_{ijk}) = \frac{(SS_1 + SS_2 + SS_3)}{N} \tag{3}$$

and

$$\mathrm{Cov}(Y^*_{ijk}, Y^*_{ist}) = \begin{cases} \dfrac{SS_1}{N} - \dfrac{SS_2}{\sum_{i=1}^{l} n_i(n_i - m_i)} - \dfrac{SS_3}{\sum_{i=1}^{l}\sum_{j=1}^{m_i} n_{ij}(n_i - n_{ij} - m_i + 1)} & \text{Strategy 1,} \\[2ex] \dfrac{SS_1}{N} - \dfrac{SS_2}{\left(\sum_{i=1}^{l} n_i(n_i - m_i)\right)} & \text{Strategy 2,} \\[2ex] \dfrac{SS_1}{N} & \text{Strategy 3,} \end{cases} \tag{4}$$

where

$$\bar{y}_{ij} = \frac{\sum_{k=1}^{n_{ij}} y_{ijk}}{n_{ij}}, \quad \bar{y}_i = \frac{\sum_{j=1}^{m_i} n_{ij}\bar{y}_{ij}}{n_i}, \quad \bar{y}_{.} = \frac{\sum_{i=1}^{l} n_i \bar{y}_i}{N}, \quad m = \sum_{l=1}^{l} m_i,$$

$$n_i = \sum_{j=1}^{m_i} n_{ij}, \quad N = \sum_{i=1}^{l}\sum_{j=1}^{m_i} n_{ij},$$

$$SS_1 = \sum_{i=1}^{l} n_i(\bar{y}_i - \bar{y}_{.})^2, \quad SS_2 = \sum_{i=1}^{l}\sum_{j=1}^{m_i} n_{ij}(\bar{y}_{ij} - \bar{y}_i)^2, \quad SS_3 = \sum_{i=1}^{l}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2.$$

Then

$$E[\mathrm{Var}(Y^*_{ijk})] = \frac{N^2 - \sum_{i=1}^{l} n_i^2}{N^2}\sigma_a^2 + \frac{N^2 - \sum_{i=1}^{l}\sum_{j=1}^{m_i} n_{ij}^2}{N^2}\sigma_b^2 + \frac{N-1}{N}\sigma_e^2 \tag{5}$$

and

$$E[\mathrm{Cov}(Y^*_{ijk}, Y^*_{ist})] = \begin{cases} \dfrac{E(SS_1)}{N} - \dfrac{E(SS_2)}{\sum_{i=1}^{l}\sum_{j=1}^{m_i} n_{ij}(n_i - m_i)} - \dfrac{E(SS_3)}{\sum_{i=1}^{l}\sum_{j=1}^{m_i} n_{ij}(n_i - n_{ij} - m_i + 1)} & \text{Strategy 1,} \\ \dfrac{E(SS_1)}{N} - \dfrac{E(SS_2)}{\left(\sum_{i=1}^{l}\sum_{j=1}^{m_i} n_{ij}(n_i - m_i)\right)} & \text{Strategy 2,} \\ \dfrac{E(SS_1)}{N} & \text{Strategy 3,} \end{cases} \tag{6}$$

where

$$E(SS_1) = \left(N - \frac{\sum_{i=1}^{l} n_i^2}{N}\right)\sigma_a^2 + \left(\sum_{i=1}^{l} \frac{\sum_{j=1}^{m_i} n_{ij}^2}{n_i} - \frac{\sum_{i=1}^{l}\sum_{j=1}^{m_i} n_{ij}^2}{N}\right)\sigma_b^2 + (l-1)\sigma_e^2,$$

$$E(SS_2) = \left(N - \sum_{i=1}^{l} \frac{\sum_{j=1}^{m_i} n_{ij}^2}{n_i}\right)\sigma_b^2 + (m-l)\sigma_e^2, \quad E(SS_3) = (N-m)\sigma_e^2.$$

Equation (6) includes $E[\mathrm{Cov}(Y^*_{ijk}, Y^*_{ijt})]$, that is,

$$E[\mathrm{Cov}(Y^*_{ijk}, Y^*_{ijt})] = \begin{cases} \dfrac{E(SS_1)}{N} - \dfrac{E(SS_2)}{\left(\sum_{i=1}^{l}\sum_{j=1}^{m_i} n_{ij}(n_i - m_i)\right)} & \text{Strategy 2,} \\ \dfrac{E(SS_1)}{N} & \text{Strategy 3.} \end{cases}$$

Thus, the nonparametric bootstrap (region) sampling strategy (Strategy 1) more closely mimics the variation properties of the data because its $E[\mathrm{Cov}(Y^*_{ijk}, Y^*_{ist})]$ is minimum based on Equation (6), and so is the preferable strategy.

3. Example

To gain a better understanding of the pathogenesis of human immunodeficiency virus (HIV)-related cardiac diseases at the early stages, harmonic phase magnetic resonance imaging

(HARP-MRI) was employed to study the differences in regional functions of the left ventricular myocardium between HIV-infected patients and normal controls. The regional systolic and diastolic functions were evaluated by measuring the average circumferential shortenings (Ecc) of the mid-wall myocardium during the systolic and diastolic phases of the left ventricular contraction. The recommendation for the number of myocardial segments for echocardiography had originally been 20, but was subsequently reduced to 16 segments [12]. The circumferential strains data was from 79 patients, 1260 segments, 13,427 time points in regional systolic functions; and from 79 patients, 1185 segments, 11,600 time points in regional diastolic functions.

Because of clustering within segments and within patients and because the circumferential strains data are subject to normal distribution, it is appropriate to analyze the data by using model (1). Standardized residuals (raw residuals divided by the corresponding standard errors) [9] are subject to standard normal distribution based on the test of normality (the ratio of the classical standard deviation to the robust standard deviation of the standardized residuals) [6]. The estimates of the fixed effects, random effects and 95% confidence intervals, as determined by using package nlme in R language [9], are listed in Table 1 for regional systolic functions and in Table 2 for regional diastolic functions, respectively.

There are three types of simple nonparametric samplings from hierarchical data with point–segment–patient structure, that is, randomly sampled data from time points (Strategy 3), segments (Strategy 2) and patients (Strategy 1).

3.1 *Nonparametric bootstrap sampling from time points*

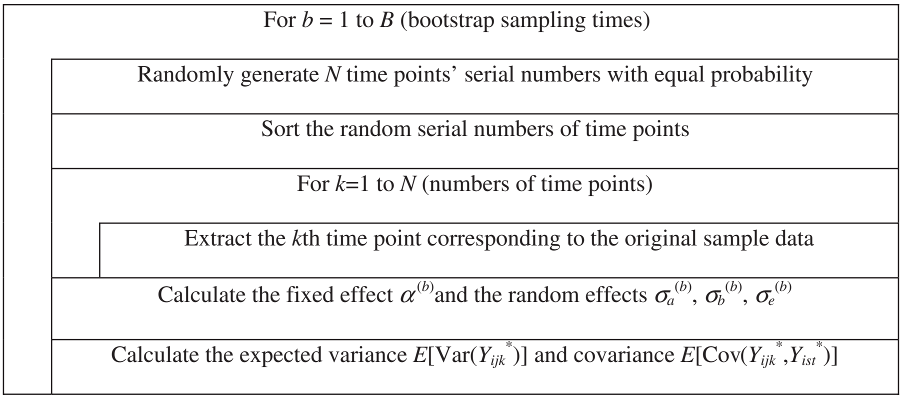
Nonparametric bootstrap sampling from time points is as follows (Figure 1): first, random generation of N serial numbers of time points with equal probability; second, from the original sample one-by-one extraction of those data that have the corresponding random serial numbers of time points; and third, organization of these data into a nonparametric bootstrap (point) sample.

Table 1. Comparison of the effects from original sample and the corresponding nonparametric bootstrap sampling for the regional systolic functions.

		Random effects		
	Fixed effects (α)	Patient (σ_a^2)	Segment (σ_b^2)	Residual (σ_e^2)
Original sample				
Estimation	−10.6064	2.7917	7.1795	53.7170
95% CI	(−11.02, −10.19)	(1.85, 4.21)	(6.27, 8.23)	(52.39, 55.08)
Bootstrap (point) sample				
Mean	−10.5839	2.7743	12.1307	48.8044
95% CI	(−10.7, −10.46)	(2.24, 3.31)	(10.78, 13.48)	(47.53, 50.08)
Median	−10.5874	2.7619	12.1205	48.7904
(2.5%, 97.5%)	(−10.71, −10.46)	(2.27, 3.35)	(10.89, 13.53)	(47.6, 50.11)
Bootstrap (segment) sample				
Mean	−10.6077	2.7752	7.1705	53.7597
95% CI	(−10.76, −10.46)	(2.08, 3.47)	(5.71, 8.63)	(51.96, 55.56)
Median	−10.6028	2.7815	7.1944	53.7491
(2.5%, 97.5%)	(−10.76, −10.46)	(2.08, 3.47)	(5.62, 8.52)	(51.92, 55.54)
Bootstrap (patient) sample				
Mean	−10.6031	2.7444	7.1306	53.6719
95% CI	(−11.01, −10.2)	(1.5, 3.99)	(4.54, 9.72)	(49.76, 57.59)
Median	−10.6014	2.6947	7.0667	53.5931
(2.5%, 97.5%)	(−11.02, −10.2)	(1.65, 4.11)	(4.68, 9.92)	(49.74, 57.69)

Table 2. Comparison of the effects from original sample and the corresponding nonparametric bootstrap sampling for the regional diastolic functions.

		Random effects		
	Fixed effects (α)	Patient (σ_a^2)	Segment (σ_b^2)	Residual (σ_e^2)
Original sample				
Estimation	−10.6451	4.549	14.8183	36.3792
95% CI	(−11.18, −10.11)	(3.04, 6.82)	(13.27, 16.55)	(35.4, 37.38)
Bootstrap (point) sample				
Mean	−10.6457	4.5421	19.1988	32.624
95% CI	(−10.77, −10.52)	(3.84, 5.25)	(17.78, 20.61)	(31.7, 33.55)
Median	−10.6485	4.5342	19.2197	32.6415
(2.5%, 97.5%)	(−10.76, −10.51)	(3.87, 5.28)	(17.79, 20.66)	(31.7, 33.51)
Bootstrap (segment) sample				
Mean	−10.639	4.6096	14.7155	36.3745
95% CI	(−10.86, −10.42)	(3.19, 6.03)	(11.57, 17.86)	(35.16, 37.59)
Median	−10.6359	4.5944	14.8158	36.3799
(2.5%, 97.5%)	(−10.84, −10.41)	(3.33, 6.18)	(11.45, 17.73)	(35.15, 37.6)
Bootstrap (patient) sample				
Mean	−10.6529	4.4895	14.9011	36.3061
95% CI	(−11.2, −10.11)	(2.28, 6.7)	(9.75, 20.05)	(33.4, 39.21)
Median	−10.652	4.398	14.7819	36.3081
(2.5%, 97.5%)	(−11.21, −10.12)	(2.62, 7.13)	(9.98, 20.56)	(33.45, 39.22)



```
For $b$ = 1 to $B$ (bootstrap sampling times)
    Randomly generate $N$ time points' serial numbers with equal probability
    Sort the random serial numbers of time points
    For $k$=1 to $N$ (numbers of time points)
        Extract the $k$th time point corresponding to the original sample data
    Calculate the fixed effect $\alpha^{(b)}$and the random effects $\sigma_a^{(b)}$, $\sigma_b^{(b)}$, $\sigma_e^{(b)}$
    Calculate the expected variance $E[\mathrm{Var}(Y_{ijk}^*)]$ and covariance $E[\mathrm{Cov}(Y_{ijk}^*, Y_{ist}^*)]$
```

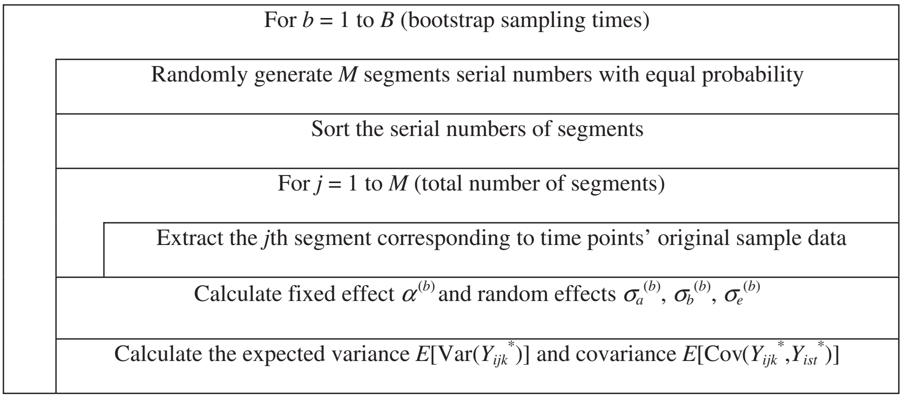
Figure 1. Nonparametric bootstrap sampling from time points.

3.2 *Nonparametric bootstrap sampling from segments*

Nonparametric bootstrap sampling from segments is as follows (Figure 2): first, random generation of M serial numbers of segments with equal probability; second, from the original sample one-by-one extraction of data that have the corresponding random serial numbers of segments; and third, organization of these data into a nonparametric bootstrap (segment) sample.

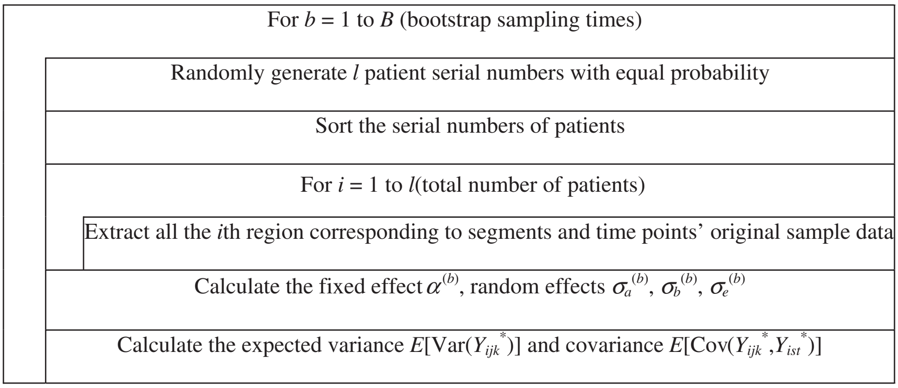
3.3 *Nonparametric bootstrap sampling from patients*

Nonparametric bootstrap sampling from patients is as follows (Figure 3): first, random generation of L serial numbers of patients; second, from the original sample one-by-one extraction of data that have the corresponding random serial numbers of patients; and third, organization of these data into a nonparametric bootstrap (patient) sample.



```
For $b = 1$ to $B$ (bootstrap sampling times)
    Randomly generate $M$ segments serial numbers with equal probability
    Sort the serial numbers of segments
    For $j = 1$ to $M$ (total number of segments)
        Extract the $j$th segment corresponding to time points' original sample data
    Calculate fixed effect $\alpha^{(b)}$ and random effects $\sigma_a^{(b)}$, $\sigma_b^{(b)}$, $\sigma_e^{(b)}$
    Calculate the expected variance $E[\mathrm{Var}(Y_{ijk}^*)]$ and covariance $E[\mathrm{Cov}(Y_{ijk}^*, Y_{ist}^*)]$
```

Figure 2. Nonparametric bootstrap sampling from segments.



```
For $b = 1$ to $B$ (bootstrap sampling times)
    Randomly generate $l$ patient serial numbers with equal probability
    Sort the serial numbers of patients
    For $i = 1$ to $l$(total number of patients)
        Extract all the $i$th region corresponding to segments and time points' original sample data
    Calculate the fixed effect $\alpha^{(b)}$, random effects $\sigma_a^{(b)}$, $\sigma_b^{(b)}$, $\sigma_e^{(b)}$
    Calculate the expected variance $E[\mathrm{Var}(Y_{ijk}^*)]$ and covariance $E[\mathrm{Cov}(Y_{ijk}^*, Y_{ist}^*)]$
```

Figure 3. Nonparametric bootstrap sampling from patients

We performed nonparametric bootstrap (point) sampling, nonparametric bootstrap (segment) sampling and nonparametric bootstrap (patient) sampling 1000 times. The results of those calculations corresponding to the estimations of the fixed effects and random effects are listed in Table 1 for regional systolic functions and in Table 2 for regional diastolic functions, respectively. The results corresponding to the estimations of the variance (5) and covariance (6) are given in Table 3.

Tables 1 and 2 show that the average values of random effects (segments) from the 1000-times nonparametric bootstrap (point) samples differ from the original samples. The average values of the fixed and random effects from the 1000-times nonparametric bootstrap (segment) samples or bootstrap (patient) samples are almost the same as the ones from the original samples, but the average values corresponding to the estimations of the covariance (6) from the 1000-times nonparametric bootstrap (patient) samples are less than the ones from the bootstrap (segment) samples or bootstrap (point) samples. With hierarchical data, the nonparametric bootstrap (patient) sampling method is closer to the original samples than the nonparametric bootstrap (point) or (segment) sampling method. The main reason may be that the nonparametric bootstrap (patient) sampling method can accurately reflect original sample information.

For each bootstrap, we gave exact bootstrap means and variances of the total, between and within sums of squares. As these bootstrap calculations are made conditionally on the observed data, they depend on the type of bootstrap but not on the underlying model [5]. To sum up, for hierarchical data with similar member–family–region structure, nonparametric bootstrap sampling

Table 3. Comparison of the expected results from the nonparametric bootstrap sampling.

	Systolic		Diastolic	
	$E[\mathrm{Var}(Y^*_{ijk})]$	$E[\mathrm{Cov}(Y^*_{ijk}, Y^*_{ist})]$	$E[\mathrm{Var}(Y^*_{ijk})]$	$E[\mathrm{Cov}(Y^*_{ijk}, Y^*_{ist})]$
Bootstrap (point) sample				
Mean	63.658	3.8776	56.2801	6.2268
95% CI	(62.01, 65.31)	(3.34, 4.41)	(54.64, 57.92)	(5.51, 6.94)
Median	63.666	3.8626	56.3028	6.2136
(2.5%, 97.5%)	(61.96, 65.23)	(3.37, 4.44)	(54.62, 57.91)	(5.54, 6.96)
Bootstrap (segment) sample				
Mean	63.6695	3.5777	55.6414	5.8878
95% CI	(60.69, 66.65)	(2.87, 4.28)	(51.81, 59.48)	(4.43, 7.35)
Median	63.7373	3.5866	55.7623	5.8715
(2.5%, 97.5%)	(60.67, 66.57)	(2.85, 4.3)	(51.59, 59.26)	(4.58, 7.46)
Bootstrap (patient) sample				
Mean	63.4373	3.1192	55.5623	5.4043
95% CI	(56.14, 70.73)	(1.85, 4.39)	(47.15, 63.98)	(3.03, 7.78)
Median	63.3056	3.0758	55.4087	5.3048
(2.5%, 97.5%)	(57.31, 69.94)	(2.03, 4.48)	(48.25, 63.99)	(3.45, 8.15)

from the highest (region) level is better than other nonparametric bootstrap sampling from the lower (family or member) levels.

4. Simulation study of the two-level data

If there are only two levels of data structure, such as families and members, we assume there are m families and n_i members in the ith family. Note that response y_{ij} is the observed value of the jth member, the ith family, $N = \sum n_i$, such that

$$Y_{ij} = \mu + a_i + e_{ij}, \quad i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n_i, \tag{7}$$

where $\mu = E(Y_{ij})$, the highest level (family) effects $\{a_i\}$ and the residual effects $\{e_{ij}\}$ are independently normally distributed with mean 0 and variances σ_a^2 and σ_e^2, respectively. The variance of Y_{ij} and the covariance of Y_{ij} and Y_{is} are $\sigma_y^2 = \sigma_a^2 + \sigma_e^2$, $\mathrm{Cov}(Y_{ij}, Y_{is}) = \sigma_a^2 (j \neq s)$, respectively.

There are two simple strategies, for both of which the first stage is to randomly sample families with replacement. At the second stage one randomly samples members within the families selected at the first stage, either without replacement (Strategy 1) or with replacement (Strategy 2). There are two types of simple nonparametric samplings from hierarchical data with member–family structure, that is, randomly sampled data from members (Strategy 2) or families (Strategy 1).

Consider selecting Y^*_{ij}, at the first stage a random integer I^* from $\{1, 2, \ldots, m\}$ is chosen. At the second stage, we select n_{i^*} random integer J^* from $\{1, 2, \ldots, n_{i^*}\}$, either without replacement (Strategy 1) or with replacement (Strategy 2). The two-level data are a special case of three-level data; therefore, we can obtain the following formulae based on Equations (5) and (6), respectively.

$$E[\mathrm{Var}(Y^*_{ij})] = \frac{N^2 - \sum_{i=1}^m n_i^2}{N^2}\sigma_a^2 + \frac{N-1}{N}\sigma_e^2 \tag{8}$$

and

$$E[\mathrm{Cov}(Y^*_{ij}, Y^*_{is})] = \begin{cases} \dfrac{E(SS_1)}{N} - \dfrac{E(SS_2)}{\left(\sum_{i=1}^m n_i(n_i - 1)\right)} & \text{Strategy 1,} \\[2ex] \dfrac{E(SS_1)}{N} & \text{Strategy 2,} \end{cases} \tag{9}$$

Table 4. Comparison of the results for two-level models from simulation data.

		Random effects			
	Fixed effects	Family (σ_a^2)	σ_e^2	$E[\mathrm{Var}(Y_{ij}^*)]$	$E[\mathrm{Cov}(Y_{ij}^*, Y_{is}^*)]$
Original sample	55.223	194.91	370.09	–	–
Bootstrap (family) sample					
Mean	55.226	190.29	370.32	538.29	234.5
95% CI	(50.63, 59.82)	(81.9, 298.6)	(282.5, 458.1)	(460.8, 615.8)	(146, 323)
Median	55.227	188.47	368.51	537.5	232.58
(2.5%, 97.5%)	(50.52, 59.82)	(87.3, 305.6)	(287, 463.7)	(463.6, 615.3)	(152.3, 327)
Bootstrap (member) sample					
Mean	55.039	189.58	371.54	553.84	304.45
95% CI	(52.8, 57.28)	(117, 262.2)	(302.1, 441)	(502.9, 604.8)	(247, 361.9)
Median	55.019	189.71	371.26	553.8	304.26
(2.5%, 97.5%)	(52.81, 57.29)	(117.2, 262.9)	(302.1, 441.1)	(502.2, 604)	(248.9, 364.3)

where $E(SS_1) = (N - \sum_{i=1}^{m} n_i^2/N)\sigma_a^2 + (m-1)\sigma_e^2$ and $E(SS_2) = (N-m)\sigma_e^2$. Thus, the nonparametric bootstrap (family) sampling strategy more closely mimics the variation properties of the data because its $E[\mathrm{Cov}(Y_{ij}^*, Y_{is}^*)]$ is minimum, and so it is the preferable strategy.

We generated simulation data from 50 families and 232 members, and each member was randomly assigned a value. After repetition of all the aforementioned steps, the results (Table 4) show the nonparametric bootstrap sampling from the family level to be better than the nonparametric bootstrap sampling from the member level for the hierarchical data with member–family structure. Furthermore, nonparametric bootstrap sampling can be extended to multilevel (i.e., more than three levels) data.

5. Extension to the four-level model

Assume there are w states (the first-stage class), v_i cities (the second-stage class) in the ith state, m_{ij} families (the third-stage class) in the jth city within the ith state and n_{ijk} members in the kth family in the jth city within the ith state. Note that response y_{ijkl} is the observed value of the lth member, the kth family, the jth city and the ith state, such that

$$Y_{ijkl} = \mu + a_i + b_{ij} + c_{ijk} + e_{ijkl}, \quad i = 1, 2, \ldots, w;\ j = 1, 2, \ldots, v_i;$$
$$k = 1, 2, \ldots, m_{ij};\ l = 1, 2, \ldots, n_{ijk},$$

where $\mu = E(Y_{ijk})$, the highest-level (state) effects $\{a_i\}$, and the second-level (city) effects $\{b_{ij}\}$, and the third-level (family) effects $\{c_{ijk}\}$ and the residual effects $\{e_{ijkl}\}$ are independently normally distributed with mean 0 and variances $\sigma_a^2, \sigma_b^2, \sigma_c^2$ and σ_e^2, respectively. These are the variance components that are to be estimated. The sampling variances of these estimates are found in [8].

There are four simple strategies, for all of which the first stage is to randomly sample states with replacement. The second stage is to randomly sample cities within the states selected at the first stage, either without replacement or with replacement. The third stage is to randomly sample families without replacement or with replacement within the cities selected at the second stage and the states selected at the first stage. The fourth stage is to randomly sample members without replacement within the families selected without replacement at the third stage, the cities selected without replacement at the second stage (Strategy 1) and the states selected at the first stage; and to randomly sample members without replacement within the families selected without replacement at the third stage (Strategy 2), the cities selected with replacement at the second stage and the states selected at the first stage; and to randomly sample members within the families selected with replacement at the third stage, the cities selected with replacement at the second stage and the states selected at the first stage, either without replacement (Strategy 3) or with replacement (Strategy 4). In short, there are four types of simple nonparametric samplings from hierarchical data with member–family–city–state structure, that is, randomly sampled data from members (Strategy 4), families (Strategy 3), cities (Strategy 2) or states (Strategy 1), respectively.

Consider selecting Y^*_{ijkl}. At the first stage, we select a random integer I^* from $\{1, 2, \ldots, w\}$. At the second stage, we select v_{i^*} random integers J^* from $\{1, 2, \ldots, v_{i^*}\}$, either without replacement or with replacement. At the third stage, we select $m_{i^*j^*}$ random integers K^* from $\{1, 2, \ldots, m_{i^*j^*}\}$ without replacement or with replacement within the cities selected at the second stage and the states selected at the first stage. At the fourth stage, we select $n_{i^*j^*k^*}$ random integers L^* from $\{1, 2, \ldots, n_{i^*j^*k^*}\}$, randomly sample members without replacement within the families selected without replacement at the third stage, the cities selected without replacement (Strategy 1) at the second stage and the states selected at the first stage; and we randomly sample members without replacement within the families selected without replacement at the third stage (Strategy 2), the cities selected with replacement at the second stage and the states selected at the first stage; and we randomly sample members within the families selected with replacement at the third stage, the cities selected with replacement at the second stage and the states selected at the first stage, either without replacement (Strategy 3) or with replacement (Strategy 4). Similarly, we can get the following formulae derived from the same concepts presented in Section 2.

$$
\begin{aligned}
E[\mathrm{Var}(Y^*_{ijkl})] = {} & \frac{N^2 - \sum_{i=1}^{w} n_i^2}{N^2}\sigma_a^2 + \frac{N^2 - \sum_{i=1}^{w}\sum_{j=1}^{v_i} n_i^2}{N^2}\sigma_b^2 \\
& + \frac{N^2 - \sum_{i=1}^{w}\sum_{j=1}^{v_i}\sum_{k=1}^{m_{ij}} n_{ijk}^2}{N^2}\sigma_c^2 + \frac{N-1}{N}\sigma_e^2
\end{aligned}
\tag{10}
$$

and

$$
\begin{aligned}
& E[\mathrm{Cov}(Y^*_{ijkl}, Y^*_{irst})] \\
& = \begin{cases}
\dfrac{E(SS_1)}{N} - \dfrac{E(SS_2)}{\sum_{i=1}^{w} n_i(n_i - m_i)} - \dfrac{E(SS_3)}{\sum_{i=1}^{w}\sum_{j=1}^{v_i}\sum_{k=1}^{m_{ij}} n_{ijk}(n_i - m_i - \sum_{r=1}^{v_i} n_{irk} + v_i)} \\
\quad - \dfrac{E(SS_4)}{\sum_{i=1}^{w}\sum_{j=1}^{v_i}\sum_{k=1}^{m_{ij}} n_{ijk}(n_i - m_i - \sum_{r=1}^{v_i} n_{irk} + v_i - n_{ij} + m_{ij} + n_{ijk} - 1)} & \text{Strategy 1,} \\
\dfrac{E(SS_1)}{N} - \dfrac{E(SS_2)}{\sum_{i=1}^{w} n_i(n_i - m_i)} - \dfrac{E(SS_3)}{\sum_{i=1}^{w}\sum_{j=1}^{v_i}\sum_{k=1}^{m_{ij}} n_{ijk}(n_i - m_i - \sum_{r=1}^{v_i} n_{irk} + v_i)} & \text{Strategy 2,} \\
\dfrac{E(SS_1)}{N} - \dfrac{E(SS_2)}{\left(\sum_{i=1}^{w} n_i(n_i - m_i)\right)} & \text{Strategy 3,} \\
\dfrac{E(SS_1)}{N}, & \text{Strategy 4,}
\end{cases}
\end{aligned}
\tag{11}
$$

where

$$
\begin{aligned}
E(SS_1) &= \left(N - \frac{\sum_{i=1}^{w} n_i^2}{N}\right)\sigma_a^2 + \left(\sum_{i=1}^{w} \frac{\sum_{j=1}^{v_i} n_{ij}^2}{n_i} - \frac{\sum_{i=1}^{w}\sum_{j=1}^{v_i} n_{ij}^2}{N}\right)\sigma_b^2 \\
&\quad + \left(\sum_{i=1}^{w} \frac{\sum_{j=1}^{v_i}\sum_{k=1}^{m_{ij}} n_{ijk}^2}{n_i} - \frac{\sum_{i=1}^{w}\sum_{j=1}^{v_i}\sum_{k=1}^{m_{ij}} n_{ijk}^2}{N}\right)\sigma_c^2 + (w-1)\sigma_e^2, \\
E(SS_2) &= \left(N - \sum_{i=1}^{w} \frac{\sum_{j=1}^{v_i} n_{ij}^2}{n_i}\right)\sigma_b^2 + \left(\sum_{i=1}^{w}\sum_{j=1}^{v_i} \frac{\sum_{k=1}^{m_{ij}} n_{ijk}^2}{n_{ij}} - \sum_{i=1}^{w} \frac{\sum_{j=1}^{v_i}\sum_{k=1}^{m_{ij}} n_{ijk}^2}{n_i}\right)\sigma_c^2 \\
&\quad + \left(\sum_{i=1}^{w} v_i - w\right)\sigma_e^2,\ N = \sum_{i=1}^{w}\sum_{j=1}^{v_i}\sum_{k=1}^{m_{ij}} n_{ijk}, \\
E(SS_3) &= \left(N - \sum_{i=1}^{w}\sum_{j=1}^{v_i} \frac{\sum_{k=1}^{m_{ij}} n_{ijk}^2}{n_{ij.}}\right)\sigma_c^2 + \left(\sum_{i=1}^{w}\sum_{j=1}^{v_i} m_{ij} - \sum_{i=1}^{w} v_i\right)\sigma_e^2, \\
E(SS_4) &= \left(N - \sum_{i=1}^{w}\sum_{j=1}^{v_i} m_{ij}\right)\sigma_e^2,\ n_{ij} = \sum_{k=1}^{m_{ij}} n_{ijk},\ m_i = \sum_{j=1}^{v_i} m_{ij},\ n_i = \sum_{j=1}^{v_i}\sum_{k=1}^{m_{ij}} n_{ijk}.
\end{aligned}
$$

Thus, the nonparametric bootstrap (state) sampling strategy (Strategy 1) more closely mimics the variation properties of the data because its $E[\mathrm{Cov}(Y_{ijkl}^{*}, Y_{irst}^{*})]$ is minimum according to Equation (11), and so is the preferable strategy.

6. Conclusions

For hierarchical unbalanced data having more than two levels, there are more than two bootstrap resampling strategies. By comparing the results of the nonparametric bootstrap methods with the hierarchical data, we proved that the strategy of nonparametric bootstrapping on the highest level (randomly sampling all other levels without replacement within the highest level selected by randomly sampling the highest levels with replacement) is better than that on the lower levels. The main reason may be that the nonparametric bootstrap (the highest level) sampling method can accurately reflect original sample information.

Acknowledgements

This research was supported by grants from the National Institute on Drug Abuse (DA12777) of United States of America. The authors are grateful to the editor and two referees for comments which improved this paper.

References

[1] A.C. Davison and D.V. Hinkley, *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge, 1997.

[2] A.C. Davison, D.V. Hinkley, and G.A. Young, *Recent developments in bootstrap methodology*. Statist. Sci. 18 (2003), pp. 141–157.

[3] B. Efron, *Bootstrap. Another look at jackknife*. Ann. Statist. 7 (1979), pp. 1–26.

[4] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[5] C.A. Field and A.H. Welsh, *Bootstrapping clustered data*. J. R. Statist. Soc. B 69 (2007), pp. 369–390.

[6] Y.R. Gel, W. Miao, and J.L. Gastwirth, *Robust directed tests of normality against heavy tailed alternatives*. Comput. Statist. Data Anal. 51 (2007), pp. 2734–2746.

[7] H. Goldstein, *Multilevel Statistical Models*, Edward Arnold, London, 1995.

[8] D.M. Mahamunulu, *Sampling variances of the estimates of variance components in the unbalanced three-way nested classification.* Ann. Math. Statist. 34 (1963), pp. 521–527.

[9] J.C. Pinheiro and D.M. Bates, *Mixed-Effects Models in S and S-PLUS*, Springer, New York, 2000.

[10] S. Ren, *Statistical methods for dependent data and their application in medicine*, doctorial diss., West China University of Medical Sciences, 1999.

[11] S. Ren, S. Yang, and S. Lai, *Intraclass correlation coefficients and bootstrap methods of hierarchical binary outcomes*. Statist. Med. 25 (2006), pp. 3576–3588.

[12] N.B. Schiller, *et al*. *Recommendations for quantitation of the left ventricle by two-dimensional echocardiography*. J. Am. Soc. Echocardiogr. 2 (1989), pp. 358–367.