Approximations of empirical processes

Abstracts of Ph.D. Thesis

By Gábor Szűcs

Supervisor: Professor Sándor Csörgő

Doctoral School in Mathematics and Computer Science
Bolyai Institute, University of Szeged
Szeged, 2011

1 Introduction

In the thesis we investigate the asymptotic behavior of some empirical processes based on independent and identically distributed random variables. In most cases we apply the approximation technique, that is, on a suitable probability space we construct a representation of the underlying process and copies of an appropriate Gaussian process such that the distance between the empirical process and the Gaussian processes converges to zero in almost sure or stochastic sense as the sample size goes to infinity.

We study two types of empirical processes. In Chapter 3 we investigate the parametric and the non-parametric bootstrap versions of the parameter estimated empirical process defined on a parametric family of distributions. The main goal of the chapter is to show the convergence of the processes by proving weak approximation theorems for them. We present a bootstrap algorithm for testing goodness of fit, and we demonstrate the bootstrap method with a simulation study.

In Chapter 4 we investigate empirical processes based on the probability generating function of non-negative valued random variables, and we work out an effective and flexible background for the study of the subject. Using this framework we prove a strong approximation result and a law of the iterated logarithm for the generating process and its derivatives. Also, we define the bootstrapped and the parameter estimated version of the probability generating process, and we apply them to construct confidence bands for the probability generating function and to test goodness of fit.

Chapter 2 is a technical one, we introduce some basic concepts which will be applied in our research. We present the main results of the Hungarian construction method, we prove a background theorem for the bootstrap technique, and we extend the definition of stochastic integration on a finite interval to stochastic integral on the real line.

The author has written three papers on the subject of the thesis. The convergence of the parametric bootstrapped version of the estimated empirical process is published in Szűcs (2008). The related theorem for the non-parametric bootstrap process and the simulation study in Chapter 3 are the subjects of an accepted paper, see Szűcs (20??) for reference. Finally, Szűcs (2005) contains the statements on the probability generating process for non-negative integer valued variables. The generalization of the results on generating processes for arbitrary non-negative valued variables is new and unpublished.

2 Some basic concepts

In the chapter we introduce three concepts. The first one is the Hungarian construction or so-called KMT approximation for the uniform empirical process $\beta_n(u)$, $0 \le u \le 1$, based on independent variables $U_1, \ldots, U_n$ distributed uniformly on the interval $[0,1]$. By the construction one can define the variables on a suitable probability space carrying a sequence of Brownian bridges $B_1, B_2, \ldots$ such that we have

$$\sup_{0\le u\le 1} \big|\beta_n(u) - B_n(u)\big| = \mathcal{O}\big(n^{-1/2}\log n\big)\,, \qquad n\to\infty\,, \qquad \text{a.s.} \tag{1}$$

This construction of Komlós, Major and Tusnády will be essential in our research.

The second tool is Efron's bootstrap method for estimating the distribution of some statistics τ_n based on a given sample having n elements. Note that such an estimation can not be obtained by using only the standard statistical techniques, because having only one set of sample variables we have only one observation for the variable τ_n. By the bootstrap heuristics if we estimate the unknown distribution function $F(x)$ of the sample variables with a function $\hat{F}_n(x)$, $x \in \mathbb{R}$, and consider conditionally independent variables $X^*_{1,n}, \ldots, X^*_{m_n,n}$ having conditional distribution function $\hat{F}_n$ with respect to the original observations, then the corresponding statistics $\tau^*_{m_n,n} = \tau_{m_n}(X^*_{1,n}, \ldots, X^*_{m_n,n})$ has "similar" distribution as τ_n has. Since the distribution of $\tau^*_{m_n,n}$ can be obtained in arbitrary precision by direct calculations or by applying Monte Carlo simulation, we can get a better or worse estimation for the distribution of τ_n.

In our research we apply two versions of the bootstrap technique, the parametric and the non-parametric bootstrap. In the non-parametric case the estimator $\hat{F}_n$ is the empirical distribution function of the observations $X_1, \ldots, X_n$, that is, the bootstrap sample is obtained by replacement from the original variables. The parametric bootstrap method can be applied only in the case when the distribution function of the X_i's is a member $F(x, \theta_0)$ of a parametric family of distribution functions $F(x, \theta)$,$x \in \mathbb{R}$, $\theta \in \Theta$. By considering an estimator $\hat{\theta}_n$ of θ_0 based on the original observations we define the parametric bootstrap sample as conditionally independent variables having conditional distribution function $F(x, \hat{\theta}_n)$.

Finally, we need to extend the standard theory of stochastic integration on a finite interval with respect to a locally square integrable martingale to stochastic integration on the whole real line. We provide a condition for the existence of the integral, and we examine the distribution of processes defined as the integrals of bivariate functions with respect to the one-dimensional standard Wiener process.

3 Bootstrap parameter estimated processes

Introduction and preliminary results

Consider a parametric collection of distributions $\mathcal{F} = \{F(x, \theta) : x \in \mathbb{R}, \theta \in \Theta \subseteq \mathbb{R}^d\}$, and independent variables $X_1, X_2, \ldots$ having common distribution function $F(x, \theta_0)$, $x \in \mathbb{R}$, with a fixed $\theta_0 \in \Theta$. If $F_n(x)$, $x \in \mathbb{R}$, stands for the empirical distribution function of the first n elements of the sequence, and $\hat{\theta}_n$ is an estimator of θ_0 based on this sample, then the corresponding parameter estimated empirical process is

$$\hat{\alpha}_n(x) = n^{1/2}\big[F_n(x) - F(x, \hat{\theta}_n)\big]\,, \qquad x \in \mathbb{R}\,. \tag{2}$$

Since Durbin (1973) proved the convergence of $\hat{\alpha}_n$ in distribution to a Gaussian process $G(x)$, $x \in \mathbb{R}$, as the sample size n goes to infinity, the parameter estimated empirical process became a widely used tool to test goodness of fit to parametric distribution families. In general, statistical methods based on the process are not distribution free, and the critical values can not be calculated in theoretical way. However, one can avoid these difficulties by applying the parametric or the non-parametric bootstrap technique.

Consider bootstrapped sample variables $X^*_{1,n},\dots,X^*_{m_n,n}$ based on $X_1,\dots,X_n$, let $F^*_{m_n,n}(x)$, $x\in\mathbb{R}$, denote the empirical distribution function of the bootstrapped sample variables, and let θ^*_n be a parameter estimator based on the bootstrapped sample. The bootstrapped parameter estimated empirical process can be defined as the parameter estimated empirical process based on the bootstrapped sample, that is, by the form

$$\bar{\alpha}^*_{m_n,n}(x)=n^{1/2}\big[F^*_{m_n,n}(x)-F(x,\theta^*_n)\big]\,,\qquad x\in\mathbb{R}\,.$$

The process is denoted by $\hat{\alpha}^*_{m_n,n}$ in the parametric and by $\tilde{\alpha}^*_{m_n,n}$ in the non-parametric bootstrap case. The heuristics of the bootstrap method is that if $\hat{\alpha}^*_{m_n,n}$ and/or $\tilde{\alpha}^*_{m_n,n}$ converges in distribution to the same weak limit as $\hat{\alpha}_n$ does, then the critical values of a test statistic $\psi(\hat{\alpha}_n)$ can be estimated by the empirical quantiles of the corresponding functionals $\psi(\hat{\alpha}^*_{m_n,n})$ and/or $\psi(\tilde{\alpha}^*_{m_n,n})$. Stute et al. (1993), and independently, Babu and Rao (2004) proved the weak convergence of $\hat{\alpha}^*_{m_n,n}$ in the case $m_n=n$ for continuous distribution families, and hence, the parametric bootstrap works in this setup. On the other hand, Babu and Rao pointed out that the process $\tilde{\alpha}^*_{m_n,n}$ is not convergent because it requires bias correction.

Assumptions and results

Consider the distribution family $\mathcal{F}=\{F(x,\theta):x\in\mathbb{R},\theta\in\Theta\subseteq\mathbb{R}^d\}$, and let θ_0 be a fixed and let $\theta=(\theta^{(1)},\dots,\theta^{(d)})$ be an arbitrary vector in the set Θ. Let $X_1,X_2,\dots$ and $Y_1,Y_2,\dots$ be sequences of independent and identically distributed random variables having distribution function $F(x,\theta_0)$ and $F(x,\theta)$, $x\in\mathbb{R}$, respectively. Also, consider a statistical function

$$\theta_n=\theta_n(Y_1,\dots,Y_n)$$

as an estimator of the general parameter θ, and let $\hat{\theta}_n=\theta_n(X_1,\dots,X_n)$ be the estimation of θ_0 based on $X_1,\dots,X_n$. Let m_n be the bootstrap sample size, and based on the bootstrapped sample let $\hat{\theta}^*_n$ and $\tilde{\theta}^*_n$ denote the estimator of $\hat{\theta}_n$ in the parametric and of θ_0 in the non-parametric case, respectively. The primary notation of vectors refers to row vectors, and V^T and $V^{(k)}$ stand for the transpose and the k-th component of a row vector V.

Assumption. We will use the following assumptions in our main results.

(a1) The vector

$$\nabla_\theta F(x,\theta)=\left(\frac{\partial}{\partial\theta^{(1)}}F(x,\theta),\dots,\frac{\partial}{\partial\theta^{(d)}}F(x,\theta)\right)$$

of partial derivatives exists for all $(x,\theta)\in\mathbb{R}\times\Lambda$, where the set $\Lambda\subseteq\Theta$ is a proper neighborhood of θ_0.

(a2) $\nabla_\theta F(x,\theta)$, $x\in\mathbb{R}$, converges uniformly to $\nabla_\theta F(x,\theta_0)$, $x\in\mathbb{R}$, as $\theta\to\theta_0$.

(a3) $\nabla_\theta F(x, \theta_0)$, $x \in \mathbb{R}$, is bounded.

(a4) There exist Borel measurable functions $l(\theta_0) : \mathbb{R} \to \mathbb{R}^d$ and $\varepsilon_n(\theta_0) : \mathbb{R}^n \to \mathbb{R}^d$ such that

$$\hat{\theta}_n - \theta_0 = \frac{1}{n} \sum_{i=1}^{n} l(X_i, \theta_0) + n^{-1/2} \varepsilon_n(\theta_0) \qquad \text{a.s.}$$

holds with $\varepsilon_n(\theta_0) = \varepsilon_n(X_1, \ldots, X_n, \theta_0)$.

(a5) There exist Borel measurable functions $l : \mathbb{R} \times \Lambda \to \mathbb{R}^d$ and $\varepsilon_n : \mathbb{R}^n \times \Lambda \to \mathbb{R}^d$ such that

$$\theta_n - \theta = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, \theta) + n^{-1/2} \varepsilon_n(\theta) \qquad \text{a.s.}$$

holds with $\varepsilon_n(\theta) = \varepsilon_n(Y_1, \ldots, Y_n, \theta)$ for every $\theta \in \Lambda$.

(a6) There exist a Borel measurable function $\varepsilon_{m,n}(\theta_0) : \mathbb{R}^{n+m} \to \mathbb{R}^d$ such that

$$\tilde{\theta}_n^* - \hat{\theta}_n = \frac{1}{m} \sum_{i=1}^{m} l(X_{i,n}^*, \theta_0) - \frac{1}{n} \sum_{i=1}^{n} l(X_i, \theta_0) + m^{-1/2} \varepsilon_{m,n}(\theta_0) \qquad \text{a.s.}$$

holds with the function l defined in (a4), and with the non-parametric bootstrap sample $X_{1,n}^*, \ldots, X_{m,n}^*$, and with $\varepsilon_{m,n}(\theta_0) = \varepsilon_{m,n}(X_1, \ldots, X_n, X_{1,n}^*, \ldots, X_{m,n}^*, \theta_0)$.

(a7) $E\, l(X_i, \theta_0) = 0$.

(a8) $E\, l(Y_i, \theta) = 0$ for every $\theta \in \Lambda$.

(a9) $M(\theta_0) = E\, l(X_i, \theta_0)^T l(X_i, \theta_0)$ is a finite non-negative definite matrix.

(a10) $M(\theta) = E\, l(Y_i, \theta)^T l(Y_i, \theta)$ is a finite non-negative definite matrix for every $\theta \in \Lambda$.

(a11) The function $M(\theta)$, $\theta \in \Lambda$, is continuous at θ_0.

(a12) Each component of $l(x, \theta_0)$, $x \in \mathbb{R}$, is of bounded variation on every finite interval.

(a13) $l(x, \theta)$ converges uniformly to $l(x, \theta_0)$, $x \in \mathbb{R}$, on every finite interval as $\theta \to \theta_0$.

(a14) $\varepsilon_n(\theta_0) \xrightarrow{P} 0$.

(a15) $\varepsilon_{m_n}(\hat{\theta}_n) \xrightarrow{P} 0$.

(a16) $\varepsilon_{m_n,n}(\theta_0) \xrightarrow{P} 0$.

Note that Burke, Csörgő, Csörgő and Révész (1979) proved a weak approximation theorem for the non-bootstrapped process $\hat{\alpha}_n$. They showed that under assumptions (a1)–(a4), (a7), (a9), (a12) and (a14) one can define a representation of the variables $X_1, X_2, \ldots$ and copies $G_1, G_2, \ldots$ of the process G such that

$$\sup_{x \in \mathbb{R}} \left| \hat{\alpha}_n(x) - G_n(x) \right| \xrightarrow{P} 0, \qquad n \to \infty. \tag{3}$$

This result immediately implies the weak convergence of $\hat{\alpha}_n$ to G. The limit process is Gaussian and can be written as

$$G(x) = B\big(F(x,\theta_0)\big) - \left[\int_{\mathbb{R}} l(x,\theta_0)\, dB\big(F(x,\theta_0)\big)\right] \nabla_\theta F(x,\theta_0)^T\,, \qquad x \in \mathbb{R}\,, \tag{4}$$

where $B(u)$, $0 \leq u \leq 1$, is a Brownian bridge. Our main results in the subject of bootstrapped empirical processes are the following theorems. Theorem 3.2 deals with the parametric and Theorem 3.3 is about the non-parametric case.

Theorem 3.2. *Assume that the bootstrap sample size $m_n \to \infty$, and assume that the distribution family $\mathcal{F}$, the fixed parameter θ_0 and the applied estimation method satisfy conditions (a1)–(a3), (a5), (a8), (a10)–(a15). Then, on a suitable probability space, one can construct random variables X_i and $X_{i,\theta}$, $\theta \in \Theta$, $i = 1,2,\ldots$, having distribution function $F(x,\theta_0)$ and $F(x,\theta)$, respectively, and a sequence of Brownian bridges $B_1, B_2, \ldots$, such that the random variables $X_1, X_{1,\theta}, X_2, X_{2,\theta}, \ldots$ are independent for every θ, and the parametric bootstrapped estimated empirical process $\hat{\alpha}^*_{m_n,n}(x)$, $x \in \mathbb{R}$, based on the variables $X_1, \ldots, X_n$ and on the parametric bootstrap sample*

$$\big(X^*_{1,n}, \ldots, X^*_{m_n,n}\big) = \big(X_{1,\hat{\theta}_n}, \ldots, X_{m_n,\hat{\theta}_n}\big)\,, \qquad n = 1,2,\ldots\,,$$

satisfies

$$\sup_{x\in\mathbb{R}} \big|\hat{\alpha}^*_{m_n,n}(x) - G_{m_n}(x)\big| \xrightarrow{P} 0\,, \qquad n \to \infty\,,$$

where $G_1, G_2, \ldots$ are defined by (4) based on $B_1, B_2, \ldots$

Theorem 3.3. *Assume that the bootstrap sample size $m_n \to \infty$, and assume that the distribution family $\mathcal{F}$, the parameter θ_0 and the estimation method satisfy conditions (a1)–(a4), (a6), (a7), (a9), (a12), (a14) and (a16). Then, on a suitable probability space, one can construct independent random variables $X_1, X_2, \ldots$ with distribution function $F(x,\theta_0)$, and non-parametric bootstrap sample variables $X^*_{1,n}, \ldots, X^*_{m_n,n}$, for $n = 1,2,\ldots$, and a sequence of Brownian bridges $B_1, B_2, \ldots$, such that the non-parametric bootstrapped parameter estimated empirical process $\tilde{\alpha}^*_{m_n,n}$ based on the variables $X_1, \ldots, X_n$ and the bootstrap sample $X^*_{1,n}, \ldots, X^*_{m_n,n}$ satisfies*

$$\sup_{x\in\mathbb{R}} \left|\tilde{\alpha}^*_{m_n,n}(x) - \Big(\frac{m_n}{n}\Big)^{1/2} \hat{\alpha}_n(x) - G_{m_n}(x)\right| \xrightarrow{P} 0\,, \qquad n \to \infty\,,$$

where $\hat{\alpha}_n$ is the estimated empirical process of (2) and $G_1, G_2, \ldots$ are defined by (4) based on $B_1, B_2, \ldots$

Since the process G_{m_n} has the same distribution as G for every n, we obtain the following consequence of Theorems 3.2 and 3.3.

Corollary 3.4. *Under the assumptions of Theorems 3.2 and 3.3 the processes*

$$\hat{\alpha}^*_{m_n,n}(x) \qquad \text{and} \qquad \tilde{\alpha}^*_{m_n,n}(x) - \Big(\frac{m_n}{n}\Big)^{1/2} \hat{\alpha}_n(x)\,, \qquad x \in \mathbb{R}\,,$$

converges weakly to $G(x)$, $x \in \mathbb{R}$, in the space $D[-\infty,\infty]$.

For the proofs of the theorems we drew inspiration from two papers. The construction of the random elements is based on the method of Csörgő and Mason (1989) who introduced the idea of representing the original and the bootstrapped sample variables on the product of two KMT spaces. Fortunately, this technique does not require any regularity condition on the model. Also, we adopt the approximation method which was applied for the non-bootstrapped process $\hat{\alpha}_n$ by Burke, Csörgő, Csörgő and Révész (1979), from where we inherit the conditions of the result (3) along with the technique. That is, we require the existence and the smoothness of the function $\nabla_\theta F(x,\theta)$ in a neighborhood of θ_0, and we must also take some assumptions on the regularity of the applied estimation method at the single point θ_0. In the non-parametric bootstrap case we do not need much more additional assumptions, we only need a similar sum representation for the bootstrapped parameter estimator $\tilde{\theta}_n^*$ as we already have in the non-bootstrapped model. Contrary, in the parametric bootstrap case the bootstrapped sample comes from the distribution $F(x,\hat{\theta}_n)$, $x \in \mathbb{R}$, and hence, we must extend the assumptions on the estimation method for a neighborhood of θ_0, and we need uniformity for the functions $M(\theta)$ and $l(x,\theta)$. Surveying the earlier proofs for the convergence of the bootstrapped processes by Stute et al. (1993) and Babu and Rao (2004) we meet similar conditions.

The bootstrap algorithm

Consider independent and identically distributed observations $X_1,\dots,X_n$ having an unknown distribution function $F(x)$, $x \in \mathbb{R}$, and consider a distribution family

$$\mathcal{F} = \big\{ F(x,\theta) : x \in \mathbb{R}, \theta \in \Theta \subseteq \mathbb{R}^d \big\}$$

endowed with a parameter estimation $\theta_n : \mathbb{R}^n \to \Theta$. In this setup one can test the fit of the sample to the family $\mathcal{F}$, that is, the null-hypotheses $\mathcal{H}_0 : F \in \mathcal{F}$ by applying the test statistics $\psi_n = \psi(\hat{\alpha}_n)$, where $\hat{\alpha}_n$ is the estimated empirical process defined in (2) and ψ is a real valued functional on the space $D[-\infty,\infty]$ satisfying certain conditions. Since $\hat{\alpha}_n$ converges weakly to the process G, the theoretical quantiles of the variable $\varphi = \psi(G)$ serves as asymptotically correct critical values for the statistics ψ_n.

As we have specified in the introduction of this chapter, the main difficulty with this method is that the quantiles of φ can not be determined in theoretical way. However, if the distribution family $\mathcal{F}$ and the estimation statistics θ_n satisfy the assumptions of Theorems 3.2 and/or 3.3, and the functional ψ satisfies some additional conditions which are not detailed here, then the parametric and nonparametric functionals

$$\psi_{m_n,n}^{*p} = \psi\big(\hat{\alpha}_{m_n,n}^*\big) \qquad \text{and} \qquad \psi_{m_n,n}^{*np} = \psi\Big(\tilde{\alpha}_{m_n,n}^* - (m_n/n)^{1/2}\hat{\alpha}_n\Big)$$

converge in distribution to the variable φ, and the $(1-\alpha)$ quantile of φ can be estimated by

$$c_n^{*p}(\alpha) = \inf\Big\{ x \in \mathbb{R} : P\big(\psi_{m_n,n}^{*p} \le x \mid X_1,\dots,X_n\big) \ge 1-\alpha \Big\}$$

in the parametric bootstrap, and by

$$c_n^{*np}(\alpha) = \inf\Big\{x \in \mathbb{R} : P\big(\psi_{m_n,n}^{*np} \le x \mid X_1, \dots, X_n\big) \ge 1 - \alpha\Big\}$$

in the non-parametric bootstrap case. As a result we obtain the following bootstrap algorithm to test the null-hypotheses $\mathcal{H}_0$..

1. Calculate the estimator $\hat{\theta}_n$ based on the observations $X_1, \dots, X_n$.

2. Calculate the test statistics ψ_n.

3. Generate independent parametric or non-parametric bootstrapped observations $X_{1,n}^*, \dots, X_{m_n,n}^*$ having distribution function $F(x, \hat{\theta}_n)$ or $F_n(x)$, respectively.

4. Calculate the estimator $\hat{\theta}_n^*$ or $\tilde{\theta}_n^*$ based on the bootstrapped sample.

5. Calculate the bootstrapped statistics $\psi_{m_n,n}^{*p}$ or $\psi_{m_n,n}^{*np}$.

6. Repeat the steps 3–5 R times, and let $\psi_{n,1}^* \le \dots \le \psi_{n,R}^*$ be the order statistics of the resulting R values of $\psi_{m_n,n}^{*p}$ or $\psi_{m_n,n}^{*np}$.

7. Let $c_{n,\alpha}^*$ be the $(1-\alpha)$ empirical quantile of $\psi_{m_n,n}^{*p}$ or $\psi_{m_n,n}^{*np}$, that is, the $\lceil R(1-\alpha)\rceil$-th largest order statistic, where $\lceil y \rceil = \min\{j \in \mathbb{Z} : y \le j\}$ for $y \in \mathbb{R}$.

8. Reject $\mathcal{H}_0$ if ψ_n is greater than $c_{n,\alpha}^*$.

In Section 3.4 we show that the method can be used with the Kolmogorov–Smirnov supremum functionals of the empirical processes. In Section 3.5 we discuss on the regularity conditions of the results, and we check the validity of the assumptions for the Poisson and the normal distribution family endowed with the maximum likelihood parameter estimation method. To demonstrate the bootstrap technique in an application in the last section we report on simulation studies. Using the parametric and the non-parametric bootstrap method we test the fit of negative binomial variables having various parameters to the Poisson distribution, and also, the fit of location and scale contaminated normal samples to the normal family.

4 Empirical probability generating processes

Introduction and preliminary results

Let $X, X_1, X_2, \dots$ be a sequence of independent and identically distributed nonnegative valued random variables having distribution function $F(x)$, $x \in \mathbb{R}$. Let

$$g(t) = Et^X = \int_{\mathbb{R}} t^x dF(x) \qquad \text{and} \qquad g_n(t) = \frac{1}{n}\sum_{j=1}^{n} t^{X_j}\,, \qquad 0 \le t \le 1\,,$$

be the common probability generating function and its empirical counterpart based on the first n observations. Throughout this chapter the symbol 0^0 is interpreted as 1, because we will need the continuity of the function t^x in variable x. Then the empirical probability generating process can be defined by

$$\gamma_n(t) = n^{1/2}\big[g_n(t) - g(t)\big]\,, \qquad 0 \le t \le 1\,.$$

The idea of the application of generating functions to solve various statistical problems is not unusual, similar transformed processes based on the empirical characteristic and moment generating functions are well-known. (See Csörgő (1981) and Csörgő, Csörgő, Horváth and Mason (1986).) In each case the theoretical basis of the method is the fact, that under appropriate conditions the transformed processes converge in distribution in some function space. In the case of the empirical probability generating process, Rémillard and Theodorescu (2000) state that γ_n converges in distribution in $C[0,1]$ to the process

$$Y(t) = \int_{\mathbb{R}} t^x \, dB\big(F(x)\big)\,, \qquad 0 \le t \le 1\,,$$

for every non-negative integer valued variable X. Unfortunately, there is an oversight in their proof, but we show that their basic idea is good, and the proof can be corrected.

The aim of the chapter is to present a general approach to convergence problems for probability generating functions and processes, for their derivatives, and for the bootstrapped and/or parameter estimated versions of the empirical probability generating process. Our results are general in the other sense, as well, that they hold not only for an integer valued variable, but for an arbitrary non-negative valued X.

Generalized probability generating processes

In Section 4.2 we investigate processes defined by the integral

$$I_r(t) = \int_{\mathbb{R}} x(x-1)\cdots(x-r+1)t^{x-r} \, dK(x)\,,$$

where the function $K(x)$ can be represented by the sum of a locally square integrable martingale $M(x)$ and a process $A(x)$ being of bounded variation on finite intervals, $x \in \mathbb{R}$. Also, we assume that M and A vanish on the negative half-line $(-\infty, 0)$ and have càdlàg trajectories. In Propositions 4.2, 4.3 and 4.8 we provide conditions under which the process I_r exists on certain subsets $[a,b]$ of the interval $(-1,1]$. The main results of the section are Theorems 4.7 and 4.9 where we prove inequalities in the form

$$\sup_{a \le t \le b} |Y_r(t)| \le C \sup_{x \in \mathbb{R}} |K(x)| \tag{5}$$

with a constant $C = C(r,a,b)$ not depending on the process K.

In the applications of these general results the process K will stand for an empirical type process (empirical or theoretical distribution function, empirical bootstrapped and/or parameter estimated process) based on the sample variables $X_1,\dots,X_n$. Since in our case the sample comes from a distribution having only non-negative values, the related empirical type processes satisfy the assumptions on K. Furthermore, using (5) we can prove convergence and approximation results for the process I_r by transferring similar results which are already obtained for the corresponding empirical process K.

Elementary properties of the empirical generating process

In Section 4.3 we prove that the empirical probability generating process γ_n has r-th ($r=0,1,\dots$) derivative

$$\gamma_n^{(r)}(t)=n^{1/2}\big[g_n^{(r)}(t)-g^{(r)}(t)\big]=\int_{\mathbb{R}} x(x-1)\cdots(x-r+1)t^{x-r}\,d\alpha_n(x)\,,$$

where $g_n^{(r)}$ and $g^{(r)}$ are the r-th derivatives of the empirical and the theoretical probability generating function of the variables $X_1,\dots,X_n$, and $\alpha_n(x)$, $x\in\mathbb{R}$, is the empirical process corresponding to the sample. Also, we investigate the process

$$Y_r(t)=\int_{\mathbb{R}} x(x-1)\cdots(x-r+1)t^{x-r}\,dB\big(F(x)\big)\,,$$

with the Brownian bridge B. In Proposition 4.10 we find that $\gamma_n^{(r)}$ is continuous and Y_r has a sample-continuous modification on $[a,b]$, where $[a,b]$ can be chosen as

- $[\varepsilon,1-\varepsilon]$ with any $0<\varepsilon<1/2$, if X is an arbitrary non-negative valued variable;
- $[-\tau,\tau]$ with any $0<\tau<1$, if X is a non-negative integer valued variable;
- $[0,1]$, if X is a non-negative integer valued variable and $r=0$.

In Section 4.11 we show that $Y_r(t)$, $a\le t\le b$, is a Gaussian process having pointwise mean zero and a continuous covariance function. Also, using a theorem of Tsirel'son for the supremum of normal variables we prove that the distributions functions of the Kolmogorov–Smirnov type functionals

$$S_r=\sup_{a\le x\le b}|Y_r(t)|\,,\qquad S_r^+=\sup_{a\le x\le b}Y_r(t)\qquad\text{and}\qquad S_r^-=-\inf_{a\le x\le b}Y_r(t)\tag{6}$$

are absolute continuous on $(0,\infty)$ and have bounded density functions on the interval $[s,\infty)$ for any $s>0$. Furthermore, by applying the Karhunen–Loève expansion of Y_r we show that the Cramér–von Mises type integral functional

$$\int_a^b Y_r^2(t)\,dt\tag{7}$$

is absolute continuous and has a bounded density function on the real line. These results will be essential when we investigate the convergence of the distribution functions of the corresponding functionals of the process $\gamma_n^{(r)}$.

Asymptotic results for the empirical generating process

As we specified earlier, Rémillard and Theodorescu (2000) showed the convergence of γ_n to Y in distribution for non-negative integer valued random variables, but there is an oversight in their proof. However, the justification method is interesting, and we correct it in Section 4.5. The basic idea of the proof is that we consider a function Ψ being defined on a subspace D_0 of $D[0,1]$ and having values in $C[0,1]$ such that $\gamma_n = \Psi(\beta_n)$ and $Y = \Psi(B)$, where β_n is a uniform empirical process and B is a Brownian bridge. We show that Ψ is measurable with respect to the Skorohod topology and continuous on $D_0 \cap C[0,1]$ with respect to the supremum metric. Since β_n converges in distribution to B and the Brownian bridge lies in $D_0 \cap C[0,1]$ almost surely, we obtain the weak convergence of $\Psi(\beta_n)$ to $\Psi(B)$ as n goes to infinity.

Our main result on the empirical probability generating process is the following uniform strong approximation. This result and the following theorems hold not only for non-negative integer valued variables, but for arbitrary non-negative valued X's.

Theorem 4.18. *Consider the distribution function $F(x)$, $x \in \mathbb{R}$, of an arbitrary non-negative valued random variable X. On a suitable probability space, one can construct independent variables $X_1, X_2, \dots$ with common distribution function F and copies $Y_{r,1}, Y_{r,2}, \dots$ of the process Y_r such that we have*

$$\sup_{a \le t \le b} \left| \gamma_n^{(r)}(t) - Y_{r,n}(t) \right| = \mathcal{O}\big(n^{-1/2} \log n\big)$$

almost surely as $n \to \infty$.

In the proof we define the random elements of the theorem based on the uniform variables and the Brownian bridges of the KMT space introduced in Chapter 1, and we obtain the approximation by transferring the convergence in (1) by inequality (5). Also, we can achieve the following law of the iterated logarithm similarly by applying Chung's law of the iterated logarithm.

Theorem 4.23. *For any non-negative valued random variable X we have*

$$\limsup_{n \to \infty} \frac{\sup_{a \le t \le b} |\gamma_n^{(r)}(t)|}{(\log\log n)^{1/2}} \le \frac{C}{2^{1/2}} \qquad a.s.$$

with a positive constant $C = C(a,b,r)$ not depending on the distribution of X.

As a consequence of Theorem 4.23 we obtain the uniform convergence of the r-th derivative of the empirical probability generating function g_n to the r-th derivative of its theoretical counterpart, and we can state a rate of convergence, too. We have

$$\sup_{a \le t \le b} \left| g_n^{(r)}(t) - g^{(r)}(t) \right| = \mathcal{O}\Big(n^{-1/2} (\log\log n)^{1/2}\Big)$$

almost surely as $n \to \infty$.

The KMT approximation in (1) has an important consequence for certain functionals of the uniform empirical process β_n. Komlós, Major and Tusnády (1975) showed that if ψ is a Lipschitzian functional on $D[0,1]$, and $\psi(B)$ has a bounded density function, then the distribution function of $\psi(\beta_n)$ converges uniformly to the distribution function of $\psi(B)$ with the rate $\mathcal{O}(n^{-1/2}\log n)$. By applying the ideas of their proof and using inequality (5) we can state a similar theorem for the functionals of the empirical probability generating process.

Theorem 4.20. *Consider an arbitrary nonnegative valued variable X, and let ψ stand for a functional on the space $C[a,b]$ satisfying the Lipschitz condition*

$$\big|\psi(h_1)-\psi(h_2)\big| \leq M \sup_{a\leq u\leq b}\big|h_1(u)-h_2(u)\big| \,, \qquad h_1,h_2\in C[a,b]\,,$$

with some finite positive constant M. If $\psi(Y_r)$ has bounded density function on $[s,\infty)$ with some $s\in\mathbb{R}$, then we have

$$\sup_{x\geq s}\Big|P\big(\psi(\gamma_n^{(r)})\leq x\big)-P\big(\psi(Y_r)\leq x\big)\Big| = \mathcal{O}\big(n^{-1/2}\log n\big)\,, \qquad n\to\infty\,.$$

Note that by (6) this statement can be applied for the supremum functionals of the generating process. Also, by making some changes in the proof, we can show that we have similar convergence for the distribution function of the square integral of $\gamma_n^{(r)}(t)$, $a\leq t\leq b$, to the distribution function of (7).

It is a known and nice result due to Finkelstein (1971) that the empirical process β_n converges at most with probability 0 in $D[0,1]$ as n goes to infinity. Instead, it is relative compact with limit set C_β, where C_β is the set of those functions $h\in D[0,1]$, which are absolute continuous, vanish at the points 0 and 1, and have Radon–Nikodym derivative $h'(u)$, $0\leq u\leq 1$, with respect to the Lebesgue measure such that

$$\int_0^1 \big(h'(u)\big)^2\,du \leq 1\,.$$

Our next result states that the empirical generating process has a similar nature.

Theorem 4.25. *For an arbitrary non-negative valued random variable X and integer $r=0,1,\ldots$ the process $\gamma_n^{(r)}(t)$, $a\leq t\leq b$, is relative compact in the space $C[a,b]$, and the set of limit points is*

$$C_r[a,b] = \bigg\{\int_{\mathbb{R}} x(x-1)\cdots(x-r+1)t^{x-r}\,dh\big(F(x)\big)\,, a\leq t\leq b : h\in C_\beta\bigg\}\,.$$

The bootstrapped generating process and confidence bands

In Section 4.8 we investigate the non-parametric bootstrapped version of the empirical probability generating process. Consider independent non-negative valued variables $X_1,X_2,\ldots$ with the same distribution function $F(x)$, $x\in\mathbb{R}$, and let $X_{1,n}^*,\ldots,X_{m_n,n}^*$ be an Efron type, that is, non-parametric bootstrap sample based on the observations

$X_1, \ldots, X_n$. Consider the bootstrapped version of the empirical probability generating function by the form

$$g^*_{m_n,n}(t) = \frac{1}{n} \sum_{i=1}^{m_n} t^{X^*_{i,n}}\,, \qquad a \le t \le b\,,$$

and define the bootstrap empirical probability generating process with

$$\gamma^*_{m_n,n}(t) = n^{1/2}\big[g^*_{m_n,n}(t) - g_n(t)\big]\,, \qquad a \le t \le b\,.$$

Based on the approximation of Csörgő and Mason (1989) for the bootstrap empirical process we obtain the following result.

Theorem 4.26. *Consider the distribution function $F(x)$, $x \in \mathbb{R}$, of an arbitrary non-negative valued random variable, and assume that there exist positive constants C_1 and C_2 such that*

$$C_1 < m_n/n < C_2\,, \qquad n = 1, 2, \ldots$$

On a sufficiently rich probability space one can define independent random variables $X_1, X_2, \ldots$ having common distribution function $F(x)$, $x \in \mathbb{R}$, and bootstrapped sample variables $X^*_{1,n}, \ldots, X^*_{m_n,n}$, $n = 1, 2, \ldots$ based on the X_i's, and copies $Y^*_{r,1}, Y^*_{r,2}, \ldots$ of the process Y_r, such that

$$\sup_{a \le t \le b} \left|\gamma^{*(r)}_{m_n,n}(t) - Y^*_{r,m_n}(t)\right| = \mathcal{O}\big(\max\{l(m_n), l(n)\}\big)\,,$$

with the function $l(n) = n^{-1/4}(\log n)^{1/2}(\log\log n)^{1/4}$.

An interesting application of the statement is that based on a sample $X_1, \ldots, X_n$ we can construct an asymptotically correct confidence band for the unknown probability generating function $g(t)$, $a \le t \le b$, of the variables. That is, for a given $0 < \alpha < 1$ we can define a sequence $c_n(\alpha)$, $n = 1, 2, \ldots$, such that

$$P\Big(g_n(t) - c_n(\alpha) \le g(t) \le g_n(t) + c_n(\alpha), a \le t \le b\Big) \to 1 - \alpha\,,$$

as the sample size n goes to infinity.

Parameter estimated probability generating processes

In Chapter 2 we present a bootstrap algorithm to test the fit of independent and identically distributed sample variables to a parametric family of univariate distributions $\mathcal{F} = \{F(x,\theta), x \in \mathbb{R}, \theta \in \Theta \subseteq \mathbb{R}^d\}$. Unfortunately, there are families whose parametric distribution functions $F(x,\theta)$ are not provided in simple forms, and by this reason the application of the parameter estimated empirical process $\hat{\alpha}_n$ and the corresponding bootstrapped processes can be very difficult. However, in many cases the probability generating process can be written in friendly formulas. By this reason we define the parameter estimated version of the probability generating process.

Suppose that $\mathcal{F}$ contains only non-negative valued distributions, and consider the parametric probability generating function of the family by the form

$$g(t,\theta) = \int_{\mathbb{R}} t^x \, dF(x,\theta)\,, \qquad a' \le t \le b'\,, \quad \theta \in \Theta\,,$$

with some a' and b'. Also, consider a sequence of independent variables $X_1, X_2, \ldots$ having distribution function $F(x,\theta_0)$, $x \in \mathbb{R}$, with a fixed $\theta_0 \in \Theta$, and let $\hat{\theta}_n$ be an estimator of θ_0 based on the first n observations. If g_n denotes the empirical probability generating function, then the parameter estimated empirical probability generating process can be defined as

$$\hat{\gamma}_n(t) = n^{1/2}\big[g_n(t) - g(t,\hat{\theta}_n)\big] = \int_{\mathbb{R}} t^x \, d\hat{\alpha}_n(x)\,, \qquad a' \le t \le b'\,.$$

Also, with the Gaussian process $G(x)$, $x \in \mathbb{R}$, of (4) let

$$\hat{Y}(t) = \int_{\mathbb{R}} t^x \, dG(x)\,, \qquad a' \le t \le b'\,.$$

Using the approximation of Burke, Csörgő, Csörgő and Révész (1979) in (3) for the parameter estimated empirical process $\hat{\alpha}_n$, we can construct a representation of the X_i's and copies $\hat{Y}_1, \hat{Y}_2, \ldots$ of $\hat{Y}$ such that we have the weak approximation

$$\sup_{0 \le t < 1} \big|\hat{\gamma}_n(t) - \hat{Y}_n(t)\big| \xrightarrow{P} 0\,, \qquad n \to \infty\,.$$

This is our Theorem 4.29 which holds under the assumptions of approximation (3) and under some additional conditions on the asymptotic behavior of $\nabla_\theta F(x,\theta)$ as $x \to \infty$.

It is important to observe that the application of the process $\hat{\gamma}_n$ can lead to an other difficulty. As we have detailed earlier, in most cases the critical values corresponding to a test statistics $\psi(\hat{\alpha}_n)$ can not be determined by theoretical calculations, and these statistics usually are not distribution free, either. Since the limit process $\hat{Y}$ is defined as an integral transformation of G, the same problems can arise for a functional $\psi(\hat{\gamma}_n)$. As a possible solution, we introduce the bootstrapped versions of the parameter estimated probability generating process, and using our approximations in Theorems 3.2 and 3.3 we prove the weak convergence of the processes. As a consequence of this result, we can test the fit of a non-negative valued sample $X_1, \ldots, X_n$ to the family $\mathcal{F}$ by using the bootstrap algorithm presented earlier with the change that the empirical process $\hat{\alpha}_n$ and its bootstrapped versions are replaced by $\hat{\gamma}_n$ and the bootstrapped variants.

References

Babu, G.J. and Rao, C.R. (2004) Goodness-of-fit tests when parameters are estimated. *Sankhyā: The Indian Journal of Statistics* **66** 63–74.

Burke, M.D., Csörgő, M., Csörgő, S. and Révész, P. (1979) Approximations of the empirical process when parameters are estimated. *The Annals of Probability* **7** 790–810.

Csörgő, M., Csörgő, S., Horváth, L. and Mason, D.M. (1986) Supnorm convergence of the empirical process indexed by functions and applications. *Probability and Mathematical Statistics* **7** 13–26.

Csörgő, S. (1981) Limit behaviour of the empirical characteristic function. *The Annals of Probability* **9** 130–144.

Csörgő, S. and Mason, D.M. (1989) Bootstraping empirical functions. *The Annals of Statistics* **17** 1447–1471.

Durbin, J. (1973) Weak convergence of the sample distribution, when parameters are estimated. *The Annals of Statistics* **1** 279–290.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7** 1–26.

Finkelstein, H. (1971) The law of the iterated logarithm for empirical processes. *The Annals of Mathematical Statistics* **42** 607–615.

Komlós, J., Major, P. and Tusnády, G. (1975) An approximation of partial sums of independent R.V.'s, and the sample D.F.I. *Z. Wahrscheinlichkeitstheorie und verw. Gebiete* **32** 111–131.

Rémillard, B. and Theodorescu, R. (2000) Inference based on the empirical probability generating function for mixtures of Poisson distributions. *Statistics & Decisions* **18** 349–366.

Szűcs, G. (2005) Approximations of empirical probability generating processes. *Statistics & Decisions* **23** 67–80.

Szűcs, G. (2008) Parametric bootstrap tests for discrete distributions. *Metrika* **67** 63–81.

Szűcs, G. (20??) Non-parametric bootstrap tests for parametric distribution families. *Acta Scientarium Mathematicarum*, accepted.

Stute, W., Manteiga, W.G. and Quindimil, M.P. (1993) Bootstrap based goodness-of-fit-tests. *Metrica* **40** 243–256.