

Approximating Markov Chains for Bootstrapping and Simulation

Roy Cerqueti, Paolo Falbo, Gianfranco Guastaroba and Cristian Pelizzari

Abstract In this work we develop a bootstrap method based on the theory of Markov chains. The method moves from the two competing objectives that a researcher pursues when performing a bootstrap procedure: (i) to preserve the structural similarity -in statistical sense- between the original and the bootstrapped sample; (ii) to assure a diversification of the latter with respect to the former. The original sample is assumed to be driven by a Markov chain. The approach we follow is to implement an optimization problem to estimate the memory of a Markov chain (i.e. its order) and to identify its relevant states. The basic ingredients of the model are the transition probabilities, whose distance is measured through a suitably defined functional. We apply the method to the series of electricity prices in Spain. A comparison with the Variable Length Markov Chain bootstrap, which is a well established bootstrap method, shows the superiority of our proposal in reproducing the dependence among data.

1 Introduction

The heart of the bootstrap -introduced by Efron (1979)- consists of resampling some given observations with the purpose of obtaining a good estimation of the statistical properties of the original population. Among the different bootstrap methods, a prominent role is played by those based on Markov chains (see e.g. Athreya and Fuh, 1992; Bühlmann, 1997; Horowitz, 2003; Paroditis and Politis, 2001a, 2001b;

Roy Cerqueti

University of Macerata, Department of Economics and Law, Via Crescimbeni, 20 - 62100 Macerata MC, Italy e-mail: roy.cerqueti@unimc.it

Paolo Falbo, Gianfranco Guastaroba and Cristian Pelizzari

University of Brescia, Department of Economics and Management, Contrada S. Chiara, 50 - 25122 Brescia BS, Italy e-mail: paolo.falbo@unibs.it, gianfranco.guastaroba@unibs.it and cristian.pelizzari@unibs.it

Anatolyev and Vasnev, 2002; Bertail and Cl emen on, 2007). The major advantage of this approach is that it is entirely data driven, so that it can smoothly capture the dependence structure of an observed time series, releasing a researcher from the risk of wrongly specifying the model, and from the difficulties of estimating its parameters.

In this paper we develop an original general approach to determine the relevant states and the memory (i.e. the order) of a Markov chain. The bootstrap procedure advanced here works similarly to that of Anatolyev and Vasnev (2002), who propose a *Markov chain bootstrap* where states correspond to the intervals resulting from a partition of the state space (of an observed time series) into I quantiles. However, differently from that work, our proposal places much greater care in identifying the states of the Markov chain. In particular, the approach we propose is based on the joint estimation of the relevant states and of the order of a Markov chain through an optimization problem. The solution identifies the partition which groups the states with the most similar transition probabilities. In this way the resulting groups emerge as the relevant states, that is the states which significantly influence the conditional distribution of the process. In this work we also extend theoretically the analysis in Cerqueti et al., (2010, 2012, 2013) by introducing L^p norm based distance measures. We also show that the minimization of the objective function represented by the distance measure of the partitions, which is based on the transition probabilities of the states, corresponds to the minimization of the information loss function in the sense of Kolmogorov (1965). The optimization problem includes also a ‘‘multiplicity’’ constraint, which controls for a sufficient diversification of the resampled trajectories. Our proposal exploits the powerful conditioning tool provided by the transition probability matrix of Markov chains to model correctly and efficiently random processes with arbitrary dependence structure. The results shown in the application to the electricity prices of the Spanish market confirm the better performances of the method proposed here with respect to a well established bootstrap approach, such as the Variable Length Markov Chain (VLMC) bootstrap of B uhlmann and Wyner (1999).

The paper is organized as follows. Section 2 introduces the settings of the model. Section 3 clarifies the theoretical foundation of the optimization problem we deal with. Section 4 formalizes the optimization problem. Section 5 provides a validation of the theoretical results through numerical experiments based on real data. Section 6 offers some conclusive remarks.

2 Model

We suppose that we observe N realizations homogeneously spaced in time of a phenomenon and we introduce the set of such time-ordered observations as $E = \{y_1, \dots, y_N\}$. There exist $J_N \geq 1$ distinct states $a_1, \dots, a_{J_N} \in E$. The corresponding subsets of E , denoted as E_1, \dots, E_{J_N} and defined as

$$E_z = \{y_i \in E \mid y_i = a_z\}, \quad z = 1, \dots, J_N, i = 1, \dots, N,$$

constitute a partition of E . Moreover, fixed $z = 1, \dots, J_N$, then the frequency of state a_z in the observed time series E is the cardinality of E_z . Let $A = \{a_1, \dots, a_{J_N}\}$ be the range of the observed time series.

We now consider a time-homogeneous Markov chain of order $k \geq 1$, denoted as $X = \{X(t), t \geq 0\}$, with state space A . To ease the notation, in the following we will simply write Markov chain instead of time-homogeneous Markov chain. The k -lag memory of the Markov chain implies that the transition probability matrix should account for conditioning to trajectories of length k . Therefore, we refer hereafter to a k -path transition probability matrix.

Let us consider $a_z \in A$ and $\mathbf{a}_h = (a_{h,k}, \dots, a_{h,1}) \in A^k$. The row vector \mathbf{a}_h is the ordered set of k states $a_{h,w} \in A$, $w = 1, \dots, k$, listed, in a natural way, from the furthest to the closest realization of the chain. The row vector \mathbf{a}_h will be called k -path. This ordering of the realizations will be maintained throughout the paper. The Markov chain has transition probability from k -path \mathbf{a}_h to state a_z given by

$$P(a_z | \mathbf{a}_h) = P(X(t) = a_z \mid X(t-1) = a_{h,1}, \dots, X(t-k) = a_{h,k}). \quad (1)$$

According to Ching et al. (2008), we estimate $P(a_z | \mathbf{a}_h)$ with the empirical frequencies $f(a_z | \mathbf{a}_h)$ related to N realizations of the phenomenon. For the sake of simplicity, we avoid introducing throughout the paper a specific notation for the estimates of the probabilities, therefore we estimate $P(a_z | \mathbf{a}_h)$ by

$$P(a_z | \mathbf{a}_h) = \begin{cases} \frac{f(a_z | \mathbf{a}_h)}{\sum_{j: a_j \in A} f(a_j | \mathbf{a}_h)}, & \text{if } \sum_{j: a_j \in A} f(a_j | \mathbf{a}_h) \neq 0 \\ 0, & \text{otherwise} \end{cases}.$$

Let us now introduce the set Λ of the partitions of A . A generic element $\lambda \in \Lambda$ can be written as $\lambda = \{A_1, \dots, A_{|\lambda|}\}$, where $|\lambda|$ is the cardinality of λ , with $1 \leq |\lambda| \leq J_N$, and $\{A_q\}_{q=1, \dots, |\lambda|}$ is a partition of nonempty subsets of A . The cardinality of Λ is $B(J_N)$, i.e. the Bell number of the J_N elements in set A .

Extending our notation to a multidimensional context, we consider the set Λ_k of k -dimensional partitions. The set Λ_k contains the partitions we will focus on in the present paper. A k -dimensional partition of Λ_k is denoted as λ and is defined as

$$\lambda = \{A_{q_k, k} \times \dots \times A_{q_w, w} \times \dots \times A_{q_1, 1} \mid q_w \in \{1, \dots, |\lambda_w|\}, w = 1, \dots, k\},$$

where $A_{q_w, w}$ is a class of the partition λ_w and λ_w is a partition of A at time lag w . A k -dimensional partition of Λ_k can also be (more easily) represented by the k -tuple of the partitions λ_w , $w = 1, \dots, k$, which the classes $A_{q_w, w}$ belong to. So the partition λ can also be identified with the notation $\lambda = (\lambda_k, \dots, \lambda_w, \dots, \lambda_1)$. Such notation describes the fact that λ is a time-dependent partition of A , i.e. A is partitioned in different ways for each time lag w , $w = 1, \dots, k$. The cardinality of Λ_k is $[B(J_N)]^k$, the cardinality of the partition λ is $|\lambda| = \prod_{w=1}^k |\lambda_w|$.

We refer to the probability law P introduced in (1) and define

$$P(a_z | \mathbf{A}_q) = P(X(t) = a_z \mid X(t-1) \in A_{q_1, 1}, \dots, X(t-k) \in A_{q_k, k}), \quad (2)$$

where $\mathbf{A}_q = A_{q_k,k} \times \dots \times A_{q_w,w} \times \dots \times A_{q_1,1} \subseteq A^k$, and $a_z \in A$. The quantity in (2) is the transition probability to reach state a_z at time t after the process has been in the classes $A_{q_k,k}, \dots, A_{q_1,1}$ in the previous k times. The transition probabilities $P(a_z|\mathbf{A}_q)$ in (2) are estimated, as usual, through the empirical frequencies:

$$P(a_z|\mathbf{A}_q) = \begin{cases} \frac{\sum_{i:\mathbf{a}_i \in \mathbf{A}_q} f(a_z|\mathbf{a}_i)}{\sum_{i:\mathbf{a}_i \in \mathbf{A}_q} \sum_{j:a_j \in A} f(a_j|\mathbf{a}_i)}, & \text{if } \sum_{i:\mathbf{a}_i \in \mathbf{A}_q} \sum_{j:a_j \in A} f(a_j|\mathbf{a}_i) \neq 0 \\ 0, & \text{otherwise} \end{cases}.$$

The quantities $P(a_z|\mathbf{A}_q)$ estimate a new transition probability matrix. To keep the notation as simple as possible, we continue to refer to this matrix as to the k -path transition probability matrix.

We deal in our paper with a couple of questions related to finding the Markov chain which best describes the observed time series E :

- Which is the optimal k ?
- Which is the optimal clustering of A for each time lag w , with $w = 1, \dots, k$?

3 Theoretical foundation of the optimization problem

In the context of bootstrapping, the conflicting scopes of a resampling procedure are two: on the one side, to maintain the statistical properties of the original sample (similarity); on the other side, to allow for a sufficient level of diversification between the original and the bootstrapped sample (multiplicity). In our model an optimal clustering procedure of the state space of a Markov chain -based on the fulfillment of similarity and multiplicity requirements- is implemented, to gain mathematical tractability when resampling.

The theoretical framework closer to our proposal is the field of information theory, with specific reference to information loss. We can, in general, define a functional space \mathcal{G} whose elements g act on the Markov chain X by defining a new Markov chain \tilde{X} . The states of \tilde{X} are the elements of a partition of A^k . There is a clear bijection between the g of \mathcal{G} and the partitions λ_g of A^k , and they are associated to a specific amount of information. Let us write $\tilde{X} = X|\lambda_g$ to formalize that the new Markov chain is X conditioned to the information generated by the partition λ_g .

Since merging two k -paths cancels part of the information on the transition probabilities available letting them distinct, we can say that the partitions λ_g imply in general information loss. This argument is in line with Kolmogorov (1965).

In order to measure such an information loss, a nonnegative functional $\eta_X \in [0, \bar{\eta}]$ can be introduced, such that $\eta_X(\tilde{X})$ represents a *distance measure* between X and \tilde{X} , which increases as the loss of information does. If $\eta(\tilde{X}) = 0$, no information is lost, while $\eta_X(\tilde{X}) = \bar{\eta}$ means that \tilde{X} is generated by a partition providing the maximum level of information loss (no information left in passing from X to \tilde{X}). The consistency requirements in the boundary situations of 0 and $\bar{\eta}$ lead to specific situations in our setting. We list them below, along with a brief explanation.

1. $\eta_X(X) = 0$. This is the case of full information and occurs when $\tilde{X} = X$. In this case the partition λ_g of the state space is the finest one: λ_g is the partition keeping separate all the states of the Markov chain (*singleton partition*).
2. If $\lambda_g = \{\emptyset, \Omega\}$ -where Ω represents the sample space of the probability space where the Markov chain is defined-, then $\eta_X(\tilde{X})$ attains its maximum. In this case the maximum level of information is lost. In fact, the corresponding partition λ_g collects all the elements of the state space in a unique set (*all-comprehensive partition*).

In the following section we will introduce a distance indicator, d_λ , and a multiplicity measure, m_λ , which will be used to measure similarity and multiplicity, respectively. They are two specific (information loss) distance measures η , which indeed satisfy conditions 1. and 2..

It can be expected that a partition of states of a Markov chain minimizing, in a controlled way, an information loss distance measure, will condition the evolution of the bootstrapped samples more consistently than it would occur if that partition had been organized otherwise.

4 Optimization problem: A formalization

The concept of optimality must be intended as satisfying the requirements of the bootstrap procedures of statistical closeness between the original and the bootstrapped sample -minimization of a distance indicator- and a certain degree of diversification -constraint on the level of a multiplicity measure-. A constrained minimization problem can be defined following this line. We enter into its details.

4.1 L^p -type distance indicator

We define an L^p -type measure of the multidimensional class \mathbf{A}_q as follows:

$$d_{\mathbf{A}_q} = \max_{i,j:\mathbf{a}_i,\mathbf{a}_j \in \mathbf{A}_q} d_{i,j}, \quad (3)$$

where

$$d_{i,j} = \sum_{z=1}^{J_N} |P(a_z|\mathbf{a}_i) - P(a_z|\mathbf{a}_j)|^p, \quad p > 0.$$

In this case, we preserve the similarity by imposing that the classes of a suitable partition have a low value of the indicator defined in (3). We can finally characterize the distance d_λ of partition λ with the average value of its classes distances:

$$d_\lambda = \frac{1}{C} \cdot \sum_{q=1}^{|\lambda|} d_{\mathbf{A}_q} \cdot |\mathbf{A}_q|,$$

where $|\mathbf{A}_q|$ is the cardinality of partition class \mathbf{A}_q and $C = \sum_{q=1}^{|\lambda|} |\mathbf{A}_q|$.

4.2 L^r -type multiplicity measure

The multiplicity measures we propose are based on the size of the partition classes.

Let us define l_λ an *absolute multiplicity measure* of the partition λ :

$$l_\lambda = \sum_{q=1}^{|\lambda|} |\mathbf{A}_q|^r, \quad r > 0.$$

We define the *relative multiplicity measure* m_λ , related to the partition λ , by normalizing l_λ as follows:

$$m_\lambda = \frac{\sqrt[r]{l_\lambda} - \sqrt[r]{C}}{C - \sqrt[r]{C}}.$$

4.3 Optimization problem

We now present the optimization problem based on the similarity and multiplicity criteria developed so far.

Definition 1. Consider $\gamma \in [0, 1]$, $k^* \in \{1, \dots, N\}$, and $\lambda^* = (\lambda_{k^*}^*, \dots, \lambda_1^*) \in \Lambda_{k^*}$. The couple (k^*, λ^*) is said to be d - γ -optimal when it is the solution of the following minimization problem:

$$\min_{(k, \lambda) \in \{1, \dots, N\} \times \Lambda_k} d_\lambda \quad \text{s.t.} \quad m_\lambda \geq \gamma. \quad (4)$$

In Definition 1 we have that k^* is the optimal order of a Markov chain describing the evolutive phenomenon. Moreover, λ^* provides the optimal time-dependent clustering of the state space A , in order to have an approximation of the k^* -path transition probability matrix.

According to the definitions of d_λ and m_λ , we can briefly discuss the optimization problem. Letting the multiplicity measure reach its minimum ($\gamma = 0$) is equivalent to allow for the singleton partition, which ensures the minimum distance ($d_\lambda = 0$). Letting $\gamma = 1$ corresponds to forcing the maximum level of multiplicity. This boundary in our case is satisfied only by the all-comprehensive partition, when the distance indicator takes its maximum value.

5 Empirical validation of the model

This section aims at comparing the proposed method with another well established bootstrap procedure, namely the Variable Length Markov Chain (VLMC) bootstrap (Bühlmann and Wyner, 1999). Only the case of $p = 1$ and $r = 2$ is discussed, to further support the strength of the method. Moreover, our method is bounded to order $k = 7$, while VLMC self-calibrates k . The original sample is the daily Mibel Spanish Electric System Arithmetic Average Price (euros per MWh) from January 2nd, 1998 to December 31st, 2003.

To assess the quality of the method, we analyze the statistical properties of the bootstrapped samples and compare them with the ones of the original sample. To this goal, we calculate the following statistics: average, standard deviation, skewness, kurtosis, minimum, maximum, and autocorrelations at lag k (where $k = 1, \dots, 8$).

The comparison focuses on the percentile rank that the original sample takes with respect to the bootstrap distribution. A package written in R named “VLMC” (available at the web page <http://cran.r-project.org/>) was used to generate the bootstrapped samples for VLMC. For what concerns our method, the optimization problem (4) was solved heuristically by means of a Tabu Search algorithm in order to control for its computational complexity (see Cerqueti et al., 2013). The number of bootstrapped samples is 5000.

The results of the comparison in Table 1 show that VLMC regularly generates narrower ranges between the 5th and the 95th percentiles than our method. However, they are only seldom consistent. In particular, autocorrelations at all lags are severely under-replicated in the VLMC bootstrapped samples. Such results confirm the expectation that the method proposed here, thanks to the minimum information loss pursued in our optimization problem, generates bootstrapped samples reproducing more carefully the original dependence among the data of an original sample.

6 Conclusive remarks

This paper proposes an optimization problem to the goal of estimating the dimensions of the transition probability matrix of a Markov chain for simulation and bootstrap purposes. The optimization problem here formalized extends that presented in Cerqueti et al. (2010), in that it introduces distance measures of L^p (L^r) type. The satisfactorily results obtained in the above-mentioned paper are theoretically further improved. The model is grounded on information theory, and it has been numerically validated through an experiment based on real data.

Table 1 Percentiles and percentile ranks of the original daily Spanish electricity prices. Comparison between the VLMC bootstrap and the bootstrap method proposed here (case $p = 1$ and $r = 2$).

Statistics	Value ^a	VLMC bootstrap			Our bootstrap		
		5 th pctl ^b	95 th pctl ^b	Pctl rank ^c	5 th pctl ^b	95 th pctl ^b	Pctl rank ^c
Average	29.692	28.707	30.704	53	26.574	32.074	57
Standard dev.	9.570	8.157	10.581	72	7.113	10.979	68
Skewness	1.381	0.414	1.953	68	0.114	2.033	66
Kurtosis	5.081	0.293	9.352	59	-0.571	9.327	63
Minimum	5.469	5.726	9.311	0	4.546	11.754	58
Maximum	103.758	66.604	110.968	73	50.971	111.382	71
Aut. at lag 1	0.818	0.737	0.817	95	0.737	0.859	62
Aut. at lag 2	0.706	0.579	0.702	95	0.579	0.772	61
Aut. at lag 3	0.706	0.463	0.615	99	0.547	0.745	63
Aut. at lag 4	0.667	0.371	0.540	99	0.529	0.733	63
Aut. at lag 5	0.661	0.297	0.476	99	0.52	0.73	62
Aut. at lag 6	0.721	0.236	0.423	99	0.614	0.764	65
Aut. at lag 7	0.802	0.187	0.378	99	0.728	0.829	68
Aut. at lag 8	0.683	0.148	0.338	99	0.581	0.727	64

^a: value is the actual value of the statistic observed in the original sample

^b: pctl stands for percentile

^c: pctl rank stands for the percentile rank of the original sample value

References

1. Anatolyev S and Vasnev A (2002) Markov chain approximation in bootstrapping autoregressions. *Econ. Bull.* 3(19):1–8
2. Athreya KB and Fuh CD (1992) Bootstrapping Markov chains: Countable case. *J. Stat. Plan. Inference* 33(3):311–331
3. Bertail P and Cléménçon S (2007) Second-order properties of regeneration-based bootstrap for atomic Markov chains. *TEST* 16:109–122
4. Bühlmann P (1997) Sieve bootstrap for time series. *Bernoulli* 3(2):123–148
5. Bühlmann P and Wyner AJ (1999) Variable length Markov chains. *Ann. Stat.* 27(2):480–513
6. Cerqueti R, Falbo P, Guastaroba G and Pelizzari C (2013) A Tabu Search heuristic procedure in Markov chain bootstrapping. *European J. Oper. Res.* 227(2):367–384
7. Cerqueti R, Falbo P and Pelizzari C (2010) Relevant states and memory in Markov chain bootstrapping and simulation. Munich Personal RePEc Archive Paper 46254, University Library of Munich, Germany
8. Cerqueti R, Falbo P, Pelizzari C, Ricca F and Scozzari A (2012) A mixed integer linear programming approach to Markov chain bootstrapping. *Quaderni del Dipartimento di Economia e Diritto dell'Università degli Studi di Macerata* 67, Università degli Studi di Macerata, Macerata, Italy
9. Ching W-K, Ng MK and Fung ES (2008) Higher-order multivariate Markov chains and their applications. *Linear Algebra Appl.* 428(2-3):492–507
10. Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann. Stat.* 7(1):1–26
11. Horowitz JL (2003) Bootstrap methods for Markov processes. *Econometrica* 71(4):1049–1082
12. Kolmogorov AN (1965) Three approaches to the quantitative definition of information. *Probl. Inf. Transm.* 1(1):3–11
13. Paparoditis E and Politis DN (2001a) Tapered block bootstrap. *Biometrika* 88(4):1105–1119
14. Paparoditis E and Politis DN (2001b) A Markovian local resampling scheme for nonparametric estimators in time series analysis. *Econometric Theory* 17(3):540–566