Improved bootstrap simultaneous prediction limits

Limiti di previsione simultanei migliorati basati su procedure bootstrap

Paolo Vidoni

Abstract This paper concerns the problem of constructing joint prediction regions having coverage probability equal or close to the target nominal value. In particular, the focus is on prediction regions defined using a system of simultaneous prediction limits. These regions are not necessarily of rectangular form and each component prediction interval depends on the preceding future observations. The specification of prediction regions with well-calibrated coverage probability has been considered in [2] and [5]. In this paper we consider an asymptotically equivalent procedure, which extends to the multivariate setting the bootstrap-based approach proposed in [3]. A simple application to autoregressive time series models is presented.

Abstract *In questo lavoro si considera il problema della costruzione di regioni di previsione con probabilità di copertura uguale o prossima a quella nominale, con particolare attenzione a regioni basate su limiti di previsione simultanei. Tali regioni non hanno necessariamente una forma rettangolare e ogni intervallo componente dipende dalle osservazioni future precedenti. La specificazione di regioni ben calibrate è stata studiata in [2] e [5]. In questo contributo si presenta una procedura di calcolo asintoticamente equivalente, e di facile implementazione, che estende all'ambito multivariato la procedura bootstrap introdotta in [3]. Si propone, infine, una semplice applicazione al caso dei modelli autoregressivi per serie storiche.*

Key words: bootstrap calibration, coverage, simultaneous prediction, time series

1 Introduction and preliminaries

This paper concerns the problem of constructing multivariate prediction regions having coverage probability equal or close to the target nominal value. In partic-

Paolo Vidoni
Department of Economics and Statistics, University of Udine, via Tomadini, 30/A I-33100 Udine, Italy, e-mail: paolo.vidoni@uniud.it

ular, the focus here is on multivariate prediction regions defined using a system of simultaneous prediction limits. These regions are not necessarily of rectangular form and they can be usefully considered whenever there is a natural order in the observations, such as for time series data, since each component prediction interval turns out to be influenced by the preceding future observations. With regard to time series applications, a system of simultaneous prediction intervals could be viewed as an alternative to a sequence of marginal prediction intervals at different periods into the future, which do not properly account for the actual dynamic evolution of the interest phenomenon.

Let (Y,Z) be a continuous random vector having joint density function $f(y,z;\theta)$, with $\theta \in \Theta \subseteq \mathbf{R}^d$, $d \geq 1$, an unknown d-dimensional parameter; $Y = (Y_1,\ldots,Y_n)$, $n \geq 1$, is observable, while $Z = (Z_1,\ldots,Z_m)$, $m \geq 1$, denotes a future, or yet unobserved, random vector. Although prediction problems may be studied from different perspectives, the aim here is to define an α-prediction region for Z, that is a random set $R(Y,\alpha) \subset \mathbf{R}^m$, depending on the observable sample Y and on the nominal coverage probability α, such that

$$P_{Y,Z}\{Z \in R(Y,\alpha);\theta\} = E_Y[P_{Z|Y}\{Z \in R(Y,\alpha)|Y;\theta\};\theta] = \alpha, \tag{1}$$

for every $\theta \in \Theta$ and for any fixed $\alpha \in (0,1)$. The above probability is called coverage probability and it is calculated with respect to the joint distribution of (Z,Y); moreover, the expectation is with respect to Y and $P_{Z|Y}\{\cdot;\theta\}$ is the probability distribution of Z given Y.

When there exists a transitive statistics $U = g(Y)$, it is natural to consider the conditional coverage probability such that, exactly or approximately,

$$P_{Y,Z|U}\{Z \in R(Y,\alpha)|U=u;\theta\} = E_{Y|U}[P_{Z|U}\{Z \in R(Y,\alpha)|U;\theta\}|U=u;\theta] = \alpha, \tag{2}$$

where the probability and the expectation are conditioned on $U = u$. For example, if we consider an autoregressive (AR) model of order 1, the transitive statistics is $U = Y_n$. Obviously, conditional solutions satisfying (2) also satisfy (1) and, when we can not find a transitive statistic, the conditional approach is meaningless.

The easiest way for making prediction on Z is to define a prediction region by using the estimative (plug-in) predictive distribution $P_{Z|Y}\{\cdot;\hat{\theta}\}$, where the unknown parameter θ is substituted with an asymptotically efficient estimator $\hat{\theta}$ based on Y, such that $\hat{\theta} - \theta = O_p(n^{-1/2})$; we usually consider the maximum likelihood estimator or any asymptotically equivalent alternative estimator. However, estimative α-prediction regions $R_e(Y,\alpha)$ are not entirely adequate predictive solutions, since the additional uncertainty introduced by assuming $\theta = \hat{\theta}$ is underestimated and then the (conditional) coverage probability differs from α by a term usually of order $O(n^{-1})$. This lack of accuracy can be substantial for small n and/or large m.

Here we focus on a particular estimative prediction region $R_e(Y,\alpha)$ based on the system of simultaneous prediction limits defined as quantiles of the conditional distributions of the components of vector $Z = (Z_1,\ldots,Z_m)$. We assume, for simplifying the exposition, that (Y,Z) follows a first-order Markovian dependence structure (so

that $U = Y_n$) and then we set

$$R_e(Y,\alpha) = \{z \in \mathbf{R}^m : z_i \leq \hat{q}_i(\alpha_i),\, i = 1,\ldots,m\}, \tag{3}$$

where $\hat{q}_i(\alpha_i) = q_i(\alpha_i, z_{i-1}; \hat{\theta})$, $i = 1,\ldots,m$, is the α_i-quantile of the conditional distribution of Z_i given $Z_{i-1} = z_{i-1}$, evaluated at $\theta = \hat{\theta}$, with $Z_0 = Y_n$. Finally, we assume $\prod_{i=i}^{m} \alpha_i = \alpha$ in order to assure that $R_e(Y,\alpha)$ is an α-prediction region, namely that $P_{Z|Y_n}\{Z \in R_e(Y,\alpha)|Y_n = y_n; \hat{\theta}\} = \alpha$. Note that the conditional prediction limit $\hat{q}_i(\alpha_i)$, for each $i = 2,\ldots,m$, is obtained recursively as a function of the previous, unknown future observation z_{i-1}.

Corcuera and Giummolè [2] find that the (conditional) coverage probability of the estimative prediction region (3) is

$$\begin{aligned} P_{Y,Z|Y_n}\{Z \in R_e(Y,\alpha)|Y_n = y_n;\theta\} &= E_{Y|Y_n}\left\{\int_{-\infty}^{\hat{q}_1(\alpha_1)} \cdots \int_{-\infty}^{\hat{q}_m(\alpha_m)} f_{Z|Y_n}(z|Y_n;\theta)dz \Big| Y_n = y_n;\theta\right\} \\ &= C_m(\alpha_1,\ldots,\alpha_m;\theta,y_n) = \alpha + Q_m(\alpha_1,\ldots,\alpha_m;\theta,y_n) + O(n^{-3/2}), \end{aligned}$$

where $f_{Z|Y_n}(z|y_n;\theta)$ is the joint conditional density of Z given $Y_n = y_n$. Moreover, after tedious calculations, an explicit expression for the $O(n^{-1})$ coverage error term $Q_m(\alpha_1,\ldots,\alpha_m;\theta,y_n)$ is also derived.

2 Improved simultaneous prediction

In order to improve the estimative predictive approach a number of solutions have been proposed. One of these strategies (see, for example, [1] and [4]) is to define an explicit modification for the estimative prediction limits, so that the associated coverage probability turns out to be equal to the target α with a high degree of accuracy. With regard to the univariate case (namely, $m = 1$, $Z = Z_1$ and $\alpha = \alpha_1$), given the estimative α_1-prediction limit $\hat{q}_1(\alpha_1)$, it is easy to prove that the modified estimative prediction limit

$$\tilde{q}_1(\alpha_1) = \hat{q}_1(\alpha_1) - \frac{Q_1(\alpha_1;\hat{\theta},y_n)}{f_{Z_1|Y_n}(\hat{q}_1(\alpha_1)|y_n;\hat{\theta})}, \tag{4}$$

reduces the coverage error to order $o(n^{-1})$. Here, $f_{Z_1|Y_n}(z_1|y_n;\theta)$ is the conditional density function of Z_1 given $Y_n = y_n$. A potential drawback of this strategy is that the evaluation of the fundamental term $Q_1(\alpha_1;\theta,y_n)$ may require complicated asymptotic calculations. To overcome this difficulty, Ueki and Fueda [3] show that the modifying term of the improved prediction limit (4) can be equivalently expressed as $\hat{q}_1(C_1(\alpha_1;\theta,y_n)) - \hat{q}_1(\alpha_1)$ and then, to the relevant order of approximation, the modified estimative prediction limit (4) corresponds to

$$\tilde{q}_1(\alpha_1) = 2\hat{q}_1(\alpha_1) - \hat{q}_1(C_1(\alpha_1;\theta,y_n)). \tag{5}$$

Therefore, the computation is greatly simplified, since we need only the value of the coverage probability $C_1(\alpha_1;\theta,y_n) = E_{Y|Y_n}\{F_{Z_1|Y_n}(\hat{q}_1(\alpha_1)|Y_n;\theta)|Y_n = y_n;\theta\}$, with $F_{Z_1|Y_n}(z_1|y_n;\theta)$ the conditional distribution function of Z_1 given $Y_n = y_n$, which can be usually estimated using a simple parametric bootstrap procedure.

The approach based on high-order analytical corrections can be extended to the multivariate case. In particular, Corcuera and Giummolè [2] specify a system of improved prediction limits $\tilde{q}_1(\alpha_1),\ldots,\tilde{q}_m(\alpha_m)$, where $\tilde{q}_1(\alpha_1)$ is defined as in (4), whereas each $\tilde{q}_i(\alpha_i)$, $i = 2,\ldots,m$, requires a further correction term in order to account for the additional dependence introduced, among the limits, by substituting θ with the same $\hat{\theta}$. This second correction term is far more complex that the first one.

In order to simplify the calculation, using a general result presented in [5], we prove that it is possible to extend the Ueki and Fueda's procedure to the multivariate setting. More precisely, an asymptotically equivalent expression for the improved simultaneous prediction limits corresponds to

$$\tilde{q}_i(\alpha_i) = 2\hat{q}_i(\alpha_i) - \hat{q}_i(C_i(\alpha_i;\theta,y_n,z_{(i-1)})),\ i = 1,\ldots,m, \tag{6}$$

with $z_{(i-1)} = (z_1,\ldots,z_{i-1})$. For $i = 1$, $C_1(\alpha_1;\theta,y_n)$ is the coverage probability of $\hat{q}_1(\alpha_1)$ as given in (5) and, for $i = 2,\ldots,m$, we consider the conditional coverage probability of $\hat{q}_i(\alpha_i)$ given $Z_{(i-1)} = z_{(i-1)}$ defined as

$$C_i(\alpha_i;\theta,y_n,z_{(i-1)}) = \frac{E_{Y|Y_n}\left\{\frac{f_{Z_{(i-1)}|Y_n}(z_{(i-1)}|Y_n;\theta)}{f_{Z_{(i-1)}|Y_n}(z_{(i-1)}|Y_n;\hat{\theta})}F_{Z_i|Z_{i-1}}(\hat{q}_i(\alpha_i)|z_{i-1};\theta)\Big|Y_n = y_n,\theta\right\}}{E_{Y|Y_n}\left\{\frac{f_{Z_{(i-1)}|Y_n}(z_{(i-1)}|Y_n;\theta)}{f_{Z_{(i-1)}|Y_n}(z_{(i-1)}|Y_n;\hat{\theta})}\Big|Y_n = y_n,\theta\right\}}, \tag{7}$$

where $F_{Z_i|Z_{i-1}}(z_i|z_{i-1};\theta)$ is the conditional distribution function of Z_i given $Z_{i-1} = z_{i-1}$ and $f_{Z_{(i-1)}|Y_n}(z_{(i-1)}|y_n;\theta)$ is the joint conditional density of $Z_{(i-1)}$ given $Y_n = y_n$. Also the conditional coverage probability (7) can be estimated using a fairly simple bootstrap parametric approach and, since the explicit expression for the correction terms is not required, this greatly simplifies the computation of the improved limits.

3 An application to a simple autoregressive model

Let us consider a stationary AR(1) process $\{Y_j\}_{j\geq 1}$ defined as

$$Y_j = \mu + \rho\,(Y_{j-1} - \mu) + \varepsilon_j, \quad j \geq 1,$$

where $\mu \in \mathbf{R}$, $|\rho| < 1$ and $\{\varepsilon_j\}_{j\geq 1}$ is a sequence of independent normal distributed random variables with zero mean and variance $\sigma^2 > 0$. Then, using the notation introduced in Section 1, $Y = (Y_1,\ldots,Y_n)$, $Z = (Z_1,\ldots,Z_m) = (Y_{n+1},\ldots,Y_{n+m})$ and $\theta = (\theta_1,\theta_2,\theta_3) = (\mu,\rho,\sigma^2)$ is the unknown model parameter. Furthermore $\hat{\theta} =$

$(\hat{\mu}, \hat{\rho}, \hat{\sigma}^2)$ is the vector of the corresponding maximum likelihood estimators which, in this case, are explicitly known.

Since Z_i given $Z_{i-1} = z_{i-1}$, $i = 1, \ldots, m$, follows a normal distribution with mean $\mu + \rho\,(z_{i-1} - \mu)$ and variance σ^2, it is immediate to define the estimative prediction region $R_e(Y, \alpha)$ as specified by (3), with simultaneous prediction limits $\hat{q}_i(\alpha_i) = \hat{\mu} + \hat{\rho}\,(z_{i-1} - \hat{\mu}) + u_{\alpha_i}\hat{\sigma}$, $i = 1, \ldots, m$. Here, u_{α_i} is such that $\Phi(u_{\alpha_i}) = \alpha_i$, where $\Phi(\cdot)$ is the distribution function of a standard normal random variable, and $\prod_{i=1}^{m} \alpha_i = \alpha$. Using the bootstrap-based procedure outlined in Section 2, we obtain the modified simultaneous prediction limits (6), which are supposed to improve the coverage accuracy of the estimative solution.

We also consider a sequence of m marginal prediction limits, which correspond the the plug-in estimates of the α_i-quantile, for $i = 1, \ldots, m$, of the conditional distribution of Z_i given $Y_n = y_n$. Notice that the first marginal prediction limit corresponds to $\hat{q}_1(\alpha_1)$. These prediction limits are computed repeatedly one period at a time and they define a rectangular-shaped prediction region. In this case, the nominal coverage probability is not equal to $\prod_{i=1}^{m} \alpha_i = \alpha$, since the component prediction limits are not independent of each other. Furthermore, by applying a bootstrap-calibrated procedure to these marginal prediction limits, as supposed to be independent, we try to improve, also in this different situation, the coverage accuracy of the corresponding prediction region.

Table 1 presents the results of a preliminary simulation study for comparing the coverage accuracy of prediction regions based on estimative and bootstrap-calibrated simultaneous prediction limits and on estimative and bootstrap-calibrated marginal prediction limits. Conditional coverage probabilities, with nominal level $\alpha = 0.9, 0.95$, are estimated using 1,000 samples of dimension $n = 50, 100$ simulated from an AR(1) model with the last observation fixed to $y_n = 1$ and assuming $y_0 = 0$; indeed, we consider $\mu = 1$, $\sigma^2 = 1$ and (a) $\rho = 0.5$, (b) $\rho = 0.8$. The prediction regions have dimension $m = 5, 10$ and $\alpha_i = \alpha^{1/m}$, $i = 1, \ldots, m$. The bootstrap procedure is based on 1,000 conditional bootstrap samples. The results are in accordance with the theoretical findings and show that the improved bootstrap-based procedures remarkably improve on the estimative ones. The improvement is more pronounced when the dimension m of the future random vector is high with respect to n. Moreover, the bootstrap-calibrated technique seems to improve the coverage accuracy of the marginal estimative prediction limits as-well, accounting also for the dependence among the component prediction limits. This is an important point which require further attention.

Finally, we conclude this section by presenting the following Figure 1, which describes a simulated path of dimension $n = 80$ from an AR(1) Gaussian model with $y_0 = 1$, $\mu = 1$, $\sigma^2 = 1$ and $\rho = 0.5$, and a sequence of $m = 50$ future simulated observations generated from the same model. Moreover, we draw the sequence of estimative and improved simultaneous prediction intervals with level $\alpha_i = 0.9$, together with the estimative and improved marginal prediction intervals with the same nominal level. The simultaneous prediction intervals account for the actual evolution of the interest time series. Note that, using the bootstrap-based approach, the

Table 1 AR(1) Gaussian model with $\mu = 1$, $\sigma^2 = 1$ and (a) $\rho = 0.5$, (b) $\rho = 0.8$. Conditional coverage probabilities for the simultaneous (estimative and improved) and marginal (estimative and improved) prediction limits of level $\alpha = 0.9, 0.95$, with $m = 5, 10$. Estimation is based on 1,000 Monte Carlo conditional (on $y_n = 1$) samples of dimension $n = 50, 100$, with $y_0 = 0$. The bootstrap procedure is based on 1,000 conditional bootstrap samples.

			(a)				(b)			
α	n	m	Simultaneous		Marginal		Simultaneous		Marginal	
			Estimative	Improved	Estimative	Improved	Estimative	Improved	Estimative	Improved
0.9	50	5	0.860	0.905	0.885	0.910	0.860	0.908	0.878	0.897
		10	0.838	0.898	0.838	0.881	0.824	0.869	0.832	0.866
	100	5	0.884	0.907	0.888	0.896	0.886	0.890	0.886	0.890
		10	0.882	0.913	0.889	0.910	0.875	0.914	0.904	0.910
0.95	50	5	0.921	0.953	0.936	0.955	0.920	0.956	0.924	0.948
		10	0.905	0.946	0.900	0.931	0.888	0.929	0.884	0.915
	100	5	0.938	0.954	0.932	0.946	0.937	0.957	0.933	0.940
		10	0.937	0.962	0.940	0.955	0.934	0.958	0.951	0.958

estimative prediction limits are suitably calibrated in order to improve the coverage accuracy.

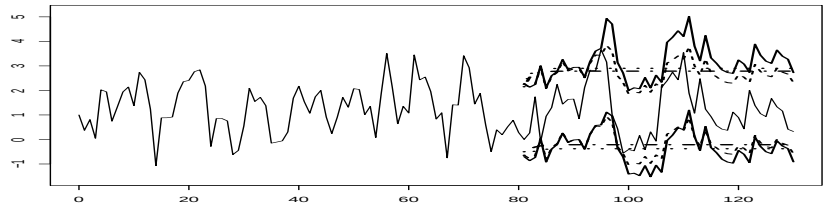

Fig. 1 Simulated observations from an AR(1) Gaussian model with $y_0 = 1$, $\mu = 1$, $\sigma^2 = 1$ and $\rho = 0.5$. Simultaneous estimative (dashed) and improved (solid) prediction intervals and marginal estimative (dotted) and improved (dot-dashed) prediction intervals with coverage probability 0.9.

Acknowledgements This research was partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003 and by the University of Udine under the PRID 2017 research grants.

References

1. Barndorff-Nielsen, O.E., Cox, D.R.: Prediction and asymptotics. Bernoulli **2**, 319–340 (1996)
2. Corcuera, J.M., Giummolè, F.: Multivariate prediction. Bernoulli **12**, 157–168 (2006)
3. Ueki, M., Fueda, K.: Adjusting estimative prediction limits. Biometrika **94**, 509–511 (2007)
4. Vidoni, P.: A note on modified estimative prediction limits and distributions. Biometrika **85**, 949–953 (1998)
5. Vidoni, P.: Calibrated multivariate distributions for improved conditional prediction. Journal of Multivariate Analysis **142**, 16–25 (2015)