Journal of Multivariate Analysis 142 (2015) 16–25

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Calibrated multivariate distributions for improved conditional prediction

Paolo Vidoni

Department of Economics and Statistics, University of Udine, via Tomadini 30/a, I-33100 Udine, Italy

ARTICLE INFO

Article history:
Received 6 March 2015
Available online 12 August 2015

AMS 2000 subject classifications:
primary 62M20
secondary 62E20

Keywords:
Bootstrap calibration
Coverage probability
Prediction region
Simultaneous prediction
Time series

ABSTRACT

The specification of multivariate prediction regions, having coverage probability closed to the target nominal value, is a challenging problem both from the theoretical and the practical point of view. In this paper we define a well-calibrated multivariate predictive distribution giving suitable conditional prediction intervals with the desired overall coverage accuracy. This distribution is the extension in the multivariate setting of a calibrated predictive distribution defined for the univariate case and it is found on the idea of calibrating prediction regions for improving the coverage probability. This solution is asymptotically equivalent to that one based on asymptotic calculations and, whenever its explicit computation is not feasible, an approximation based on a simple bootstrap simulation procedure is readily available. Moreover, we state a simple, simulation-based, procedure for computing the associated improved conditional prediction limits.

1. Introduction

Predictive inference for an unobserved multivariate random variable may be of considerable interest in a number of application, such as the specification of simultaneous prediction intervals for future time series observations and the construction of prediction regions for observables from multivariate models. In this paper, prediction is considered from the frequentist perspective and the aim is to define a well-calibrated multivariate predictive distribution giving conditional prediction intervals, and in particular conditional prediction limits, with overall coverage probability closed to the target nominal value. The associated prediction regions are not necessarily of rectangular form, in the two-sided case, or of semi-infinite rectangular form, in the one-sided case. This is an important point and it will be appropriately discussed throughout the paper.

Let (Y, Z) be a continuous random vector having joint density function $p(y, z; \theta)$, 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. This is a fairly general formulation which includes both the simple case with $Y_1, \ldots, Y_n$ and Z independent, identically distributed, multivariate random variables and the more general situation with dependent Y and Z and, in particular, with Y and Z defined within a stochastic process model. For ease of exposition we consider $Y = (Y_1, \ldots, Y_n)$ as a random vector with density $f(\cdot; \theta)$ and Z as an independent future random vector with density $g(\cdot; \theta)$, possibly different from $f(\cdot; \theta)$, with θ the same d-dimensional parameter as before; $G(\cdot; \theta)$ indicates the distribution function of Z. We also assume that $f(\cdot; \theta)$, $g(\cdot; \theta)$ and $G(\cdot; \theta)$ are sufficiently smooth functions of the parameter θ. The extension of the results to the case with dependent Y and Z is considered in the final part of the paper.

E-mail address: paolo.vidoni@uniud.it.

http://dx.doi.org/10.1016/j.jmva.2015.08.001

Although prediction problems may be tackled with different objectives, 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

$$P_{Y,Z}\{Z \in R(Y, \alpha); \theta\} = \alpha, \tag{1.1}$$

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, it can be rewritten as $E_Y[P_Z\{Z \in R(Y, \alpha); \theta\}; \theta]$, where the expectation is with respect to Y and $P_Z\{\cdot; \theta\}$ is the probability distribution for Z.

Given a suitable predictive probability distribution for Z, namely $\tilde{P}_Z\{\cdot; Y\}$, defined as an estimator for the true $P_Z\{\cdot; \theta\}$ based on the sample Y, we may define an α-prediction region for Z as a set $\tilde{R}(Y, \alpha) \subset \mathbf{R}^m$ such that

$$\tilde{P}_Z\{Z \in \tilde{R}(Y, \alpha); Y\} = \alpha.$$

Under this respect, we aim at introducing a predictive distribution function such that the corresponding α-prediction region $\tilde{R}(Y, \alpha)$ fulfils (1.1) exactly or with a high degree of accuracy, for each $\alpha \in (0, 1)$.

Although there are some special cases where there is an exact solution to (1.1), these situations are extremely rare. Thus, in general, we look for approximate solutions satisfying (1.1) almost exactly, for each $\alpha \in (0, 1)$. The easiest way for making prediction on Z is by using the estimative predictive distribution $P_Z\{\cdot; \hat{\theta}\}$, 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, it is well-known that the estimative α-prediction regions $R_e(Y, \alpha)$ are not entirely adequate predictive solutions, since their coverage probability differs from α by a term usually of order $O(n^{-1})$ and prediction statements may be rather inaccurate for small n. In fact, this naive solution underestimates the additional uncertainty introduced by assuming $\theta = \hat{\theta}$.

Concerning the univariate case, Barndorff-Nielsen and Cox [1,2], Ueki and Fueda [9] and Vidoni [10,11] suggest a way to correct, by means of asymptotic calculations, the quantiles of the estimative predictive distribution, thus obtaining prediction intervals with a coverage error of order $o(n^{-1})$. A calibrating approach, useful in the multivariate case as well, has been suggested by Beran [3] and applied, for example, by Hall et al. [6], using a bootstrap procedure for improving the estimative prediction intervals. The key idea, behind this approach, is to determine a suitable value $\bar{\alpha}$ such that the coverage probability of the estimative, recalibrated, prediction region $R_e(Y, \bar{\alpha})$ is equal or close to the target value α. The effect of (bootstrap) calibration is that of reducing the magnitude of the coverage error but it is valid for a specific α and it does not provide a general solution to the problem, such as those based on the notion of predictive distribution.

Recently, Fonseca et al. [5] extend the calibrating procedure to predictive distribution functions in such a way that the associated prediction intervals have coverage probability equal or close to the target value. This solution has similarities with that one specified by Lawless and Fredette [8], involving (approximate) pivotal quantities, but it has the advantage that, whenever its computation is not feasible, it can be approximated using a suitable bootstrap simulation procedure or considering high-order asymptotic expansions, giving the same improved predictive distributions already known in the literature.

In the present paper, this result is properly extended to deal with multivariate prediction problems. In particular, a well-calibrated multivariate predictive distribution is derived. We prove that the associated joint predictive density is asymptotically equivalent to that one proposed by Corcuera and Giummolè [4], which gives improved conditional prediction limits. This new solution, contrary to the Corcuera and Giummolè's one, has a simple and intuitive form and, when computations are hard to perform, it is readily available an approximation based on bootstrap simulation methods. Furthermore, generalizing a result presented in [9], we state a simple, simulation-based, procedure for computing the associated improved conditional prediction limits. The paper is organized as follows. Section 2 reviews some known results on improved predictive procedures. Section 3 introduces the new approach based on the calibrated multivariate predictive distribution. Section 4 presents the procedure for calculating the improved conditional prediction limits, following the Ueki and Fueda's approach, and Section 5 briefly considers the extension to the case of dependent observations. Finally, Section 6 is dedicated to a simple example concerning autoregressive time series models.

2. Review on improved multivariate prediction

Let us review the calibrated predictive distribution function proposed by Fonseca et al. [5] for the univariate case, that is for $m = 1$. Let us consider the estimative α-prediction limit $\hat{q}(\alpha) = q(\alpha; \hat{\theta})$, defined as the α-quantile of the estimative distribution function $G(z; \hat{\theta})$, such that $G\{\hat{q}(\alpha); \hat{\theta}\} = \alpha$. The associated coverage probability is

$$P_{Y,Z}\{Z \leq \hat{q}(\alpha); \theta\} = E_Y[G\{\hat{q}(\alpha); \theta\}; \theta] = C(\alpha; \theta).$$

By substituting α with $G(z; \hat{\theta})$ in $C(\alpha; \theta)$, we obtain

$$G_c(z; \hat{\theta}, \theta) = C\{G(z; \hat{\theta}); \theta\},$$

which is a predictive distribution function if $C(\cdot; \theta)$ is a sufficiently smooth function. The associated density function is a suitable modification of the estimative predictive density $g(z; \hat{\theta})$ given by

$$g_c(z; \hat{\theta}, \theta) = g(z; \hat{\theta})C'\{G(z; \hat{\theta}); \theta\},$$