Prediction in a multidimensional setting

Giovanni Fonseca, Federica Giummolè and Paolo Vidoni

Abstract This paper concerns the problem of prediction in a multidimensional setting. Generalizing a result presented in Ueki and Fueda (2007), we propose a method for correcting estimative predictive regions to reduce their coverage error to third-order accuracy. The improved prediction regions are easy to calculate using a suitable bootstrap procedure. Furthermore, the associated predictive distribution function is explicitly derived. Finally, an example concerning the exponential distribution shows the good performance of the proposed method.

Key words: coverage probability, estimative prediction region, parametric bootstrap.

1 Introduction

Let us assume that $Y = (Y_1, \ldots, Y_n), n \geq 1$, is an observable continuous random vector. The problem of prediction, in a multidimensional setting, consists in defining a suitable prediction region, that is a subset of $\mathbb{R}^m, m \geq 1$, with a fixed probability of including a further continuous random vector $Z = (Z_1, \ldots, Z_m)$. The joint distribution of $Z$ and $Y$ is assumed to be known, up to a $k$-dimensional parameter $\omega \in \Omega \subseteq \mathbb{R}^k, k \geq 1$; $\hat{\omega} = \hat{\omega}(Y)$ denotes an asymptotically efficient estimator for $\omega$. 

Giovanni Fonseca
Università di Udine, Dipartimento di Scienze Economiche e Statistiche, via Treppo 18, I-33100 Udine, Italy, e-mail: giovanni.fonseca@uniud.it

Federica Giummolè
Università Ca’ Foscari di Venezia, Dipartimento di Scienze Ambientali, Informatica e Statistica, San Giobbe, Cannaregio 873, I-30121 Venezia, Italy, e-mail: giummole@unive.it

Paolo Vidoni
Università di Udine, Dipartimento di Scienze Economiche e Statistiche, via Treppo 18, I-33100 Udine, Italy, e-mail: paolo.vidoni@uniud.it
usually the maximum likelihood estimator. For simplicity, \( Y \) and \( Z \) are considered independent and we denote by \( f(z; \omega) \) the joint density function of \( Z \).

The simplest predictive solution is the estimative or plug-in one. An estimative prediction region, with nominal probability \( \alpha \in (0, 1) \), is a suitable subset of \( \mathbb{R}^m \) derived from the estimative predictive density \( f(z; \hat{\omega}) \), which is obtained by substituting the unknown parameter \( \omega \) by \( \hat{\omega} \) in \( f(z; \omega) \). Unfortunately the associated coverage probability is not equal to the target value \( \alpha \). The error term has order \( O(n^{-1}) \) and it is often considerable. For scalar \( Z \), improved predictive solutions have been proposed in Barndorff-Nielsen and Cox (1996) and Vidoni (1998), involving complicated asymptotic calculations with the aim of reducing the coverage error to order \( o(n^{-1}) \). Recently, Ueki and Fueda (2007) suggested a simple simulation-based procedure, useful to easily compute improved \( \alpha \)-prediction limits. In this work we extend the Ueki and Fueda’s procedure to the case of \( Z \) being a multidimensional random variable. Furthermore, we specify a predictive distribution function associated to improved prediction regions. An application, concerning exponential distribution, shows the good performance of the proposed method.

2 Improved prediction region

As suggested in Beran (1990) and Ueki and Fueda (2007), we consider estimative prediction regions of the form \( D(r, \hat{\omega}) = \{z \in \mathbb{R}^m : R(z, \hat{\omega}) \leq r\} \), for some real value \( r \) and some smooth real function \( R(z, \omega) \). Notice that the so-called highest prediction density region is a special case with \( R(z, \omega) = -f(z; \omega) \). Prediction regions of this form are identified by the value of \( r \), which we refer to as the limit of the region. From now on, our aim is to find a prediction limit \( \tilde{r}_\alpha(y) \) such that

\[
P_{Y,Z}[R\{Z, \hat{\omega}(Y)\} \leq \tilde{r}_\alpha(Y) \} = E_Y \left[ \int_{D[\tilde{r}_\alpha(Y), \hat{\omega}]} f(z; \omega) dz \right] = \alpha,
\]

for all \( \alpha \in (0, 1) \), at least to a high-order of approximation. The above probability is the coverage probability of the prediction region and it is intended with respect to the joint distribution of \( Y, Z \) with parameter \( \omega \). When \( Z \) is unidimensional and \( R(Z, \omega) = Z, \tilde{r}_\alpha(Y) \) is the \( \alpha \)-prediction limit for \( Z \).

The estimative solution is based on the estimative prediction limit \( r_\alpha(\hat{\omega}) \), such that

\[
\int_{D[r_\alpha(\hat{\omega}), \hat{\omega}]} f(z; \hat{\omega}) dz = \alpha.
\]

The coverage probability of the estimative prediction region \( D[r_\alpha(\hat{\omega}), \hat{\omega}] \) is \( \hat{\alpha}(\omega) = \alpha + O(n^{-1}) \) and, in order to eliminate the \( O(n^{-1}) \) coverage error term, we modify \( r_\alpha(\hat{\omega}) \) as done by Ueki and Fueda (2007) in the unidimensional case. More precisely, the adjusted prediction limit, achieving coverage probability \( \alpha + o(n^{-1}) \), is

\[
\tilde{r}_\alpha(\hat{\omega}) = 2r_\alpha(\hat{\omega}) - r_{\hat{\alpha}(\omega)}(\hat{\omega}),
\]

(1)
Prediction in a multidimensional setting

where \( r_{\hat{\omega}}(\hat{\omega}) \) is the \( \hat{\alpha}(\omega) \)-estimative prediction limit. The improved estimative prediction region is \( D(\hat{r}_{\alpha}(\hat{\omega}), \hat{\omega}) = \{ z \in \mathbb{R}^m : R(z, \hat{\omega}) \leq \hat{r}_{\alpha}(\hat{\omega}) \} \). In order to explicitely calculate \( \hat{r}_{\alpha}(\hat{\omega}) \), we only need to evaluate the estimative coverage probability \( \hat{\alpha}(\omega) \). This can be easily computed in practice, using a suitable parametric bootstrap procedure.

Finally, as proved in Fonseca et al. (2011), we may obtain an explicit expression for the distribution function which gives, up to terms of order \( O(n^{-1}) \), the improved limit \( \hat{r}_{\alpha}(\hat{\omega}) \) as \( \alpha \)-quantile, for all \( \alpha \in (0, 1) \). Let \( F_R(r; \omega) \) be the distribution function of \( R(Z, \omega) \); thus, \( \hat{r}_{\alpha}(\hat{\omega}) \) is such that \( F_R(\hat{r}_{\alpha}(\hat{\omega}); \hat{\omega}) = \alpha \). The improved predictive distribution function corresponds to

\[
\hat{F}_R(r; Y) = F_R(r; \hat{\omega}) + f_R(r; \hat{\omega}) \left[ F_R^{-1}\{ \hat{\alpha}(\omega); \hat{\omega} \} \big|_{\alpha=F_R(r;\hat{\omega})} - r \right],
\]

with \( f_R(:; \omega) \) the density function of \( R(Z, \omega) \) and \( F_R^{-1}(,:; \omega) \) the inverse of function \( f_R(:, \omega) \). When the distribution function \( F_R(r; \omega) \) is not available, it may be approximated by means of a further bootstrap procedure.

3 Example

Let \( Y_1, \ldots, Y_n, Z_1, \ldots, Z_m, n, m \geq 1 \), be independent exponential random variables with unknown scale parameter \( \omega > 0 \). The maximum likelihood estimator for \( \omega \) is \( \hat{\omega} = \bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i \). A highest prediction density region is \( D(r, \hat{\omega}) = \{ z \in [0, +\infty)^m : z/\hat{\omega} \leq r \} \), with \( z = n^{-1} \sum_{j=1}^{m} z_j \). Notice that \( \bar{Z}/\hat{\omega} \) is a pivotal quantity, having a Fisher \( F \) distribution, \( F(2m, 2n) \). Thus, a prediction region with exact coverage probability \( \alpha \) can be obtained by choosing as limit of the region \( f_{\hat{\alpha}, 2m, 2n} \), the \( \alpha \)-quantile of a \( F(2m, 2n) \) distribution. Nonetheless, the aim of this example is to test the performance of the improved prediction region. In order to do this, note that \( R(Z, \omega) = \bar{Z}/\omega \) has a Gamma distribution with shape parameter \( m \) and scale parameter \( 1/m \), so that the estimative limit \( r_{\alpha}(\hat{\omega}) \) coincides with the \( \alpha \)-quantile of a Gamma\((m, 1/m) \) distribution. The corresponding coverage probability, \( \hat{\alpha}(\omega) \), can be evaluated using a suitable parametric bootstrap procedure. The improved prediction limit can thus be calculated by means of expression (1).

Table 1 shows the results of a simulation study for comparing coverage probabilities for estimative and improved prediction regions of level \( \alpha = 0.9, 0.95 \). The scale parameter of the true distribution is \( \omega = 10 \). It can be noticed that the coverage probability associated to improved prediction limits is closer to the nominal value \( \alpha \) than that one corresponding to the estimative solution, especially as the number of future variables \( m \) increases.

Finally, Figure 1 considers the case where \( \omega = 1 \) and it shows the upper tail of the exact predictive distribution function, which is based on the pivotal quantity \( \bar{Z}/\hat{\omega} \), together with those ones of the estimative and the improved predictive distribution. The exact solution turns out to be better approximated by the improved predictive distribution.
Table 1 Independent exponential random variables with scale parameter $\omega = 10$, $n = 10, 20$ and $m = 1, 5, 10$. Coverage probabilities for estimative and improved prediction regions of level $\alpha = 0.9, 0.95$. Estimation based on 10,000 Monte Carlo replications and bootstrap procedure based on 5,000 bootstrap samples. Estimated standard errors are smaller than 0.005.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.9$</th>
<th>$\alpha = 0.95$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$m$</td>
<td>Estimative</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.878</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.818</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.784</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0.884</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.855</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.830</td>
</tr>
</tbody>
</table>

Fig. 1 Independent exponential random variables with scale parameter $\omega = 1$. Plots of upper-tail of estimative (dashed), improved (dotted) and exact (solid) predictive distribution functions, for different values of the sample size $n = 10, 20$ and dimension of the future vector $m = 1, 5, 10$.

References