A note on predictive densities based on composite likelihood methods

A note on predictive densities based on composite likelihood methods

Paolo Vidoni

Abstract Whenever the computation of the data distribution is not feasible or convenient, the classical predictive procedures are not useful since they rely on the conditional distribution of the future random variable given the observations, which is also not available. This paper aims at considering a notion of composite likelihood for specifying composite predictive distributions, viewed as surrogates for the true unknown predictive distribution. In particular, the focus is on the pairwise likelihood obtained as a weighted product of likelihood factors related to bivariate events associated to both the sample data and the future observation. The specification of the weights, and more generally the evaluation of the frequentist properties of alternative pairwise predictive distributions, is performed by considering the mean square prediction error of the associated predictors and the expected Kullback-Liebler loss of the related predictive densities. Finally, simple examples concerning autoregressive models are presented.

Keywords Kullback-Leibler divergence · Pairwise likelihood · Logarithmic prediction pool · Predictive distribution

1 Introduction

This paper concerns the prediction of the value of a future or not yet observed random variable, based on the available observations, in the challenging situation where the joint distribution of the data is not available. This happens whenever a full model specification is not reliable and the model turns out to be partially specified giving only low-dimensional marginal or conditional distributions. Moreover, even if the model may be potentially defined, a closed form expression for the joint distribution could not be computable in a closed form, or approximated using analytical or numerical procedures, due to the

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

complex interdependencies which are involved or to the presence of a huge amount of data.

In these situations, in order to perform likelihood-based inference, it may be useful to consider suitable pseudolikelihoods, such as composite likelihoods which are constructed by composing low-dimensional likelihood objects [10]. Composite likelihood inferential procedures have proved to be useful in a number of complex statistical models (see for example [14] and references therein) and they usually have good properties even if, compared to the full likelihood methods, they could be less efficient. A careful choice of the likelihood objects and the specification of a suitable system of weights may reduce this gap.

In this context the classical predictive procedures are not useful either, since the lack of an explicit expression for the joint distribution of data does not permit the computation of the conditional distribution of the future random variable given the observed ones. The aim of this paper is to consider a suitable notion of composite likelihood for specifying composite predictive distributions, as useful surrogates for the true unknown predictive distribution. Among various notions of composite likelihood, we focus on the pairwise likelihood, obtained as a weighted product of likelihood factors related to bivariate marginal or conditional events associated to both the observed sample and the future unknown observation.

The pairwise predictive distribution obtained in this way may be interpreted as a weighted pool of bivariate predictive distributions, which correspond to partially specified models for prediction. A further interesting interpretation involves the notion of exponential tilting and the information theoretical principle of maximum entropy. This new notion of predictive distribution can be considered for specifying point predictors and prediction intervals, whenever a genuine predictive distribution cannot be defined for computational or modelling problems. With particular regard to the construction of prediction intervals, a careful specification of the pairwise predictive distribution is required, in order to get a valid uncertainty assessment for the prediction statement. Under this respect, the specification of the weights, and more generally the evaluation of the properties of alternative pairwise predictive distributions is extremely important and, in this case, it is performed by considering the mean square prediction error of the associated predictors and, in particular, the expected Kullback-Leibler loss of the predictive distribution taken into account.

Finally, a simple example concerning autoregressive models with additive observation noise is presented with the aim of comparing the classical predictive procedures, available in this case, and those ones based on the pairwise predictive distribution. Furthermore, a more interesting application to autoregressive ordered probit models, where an exact predictive solution is not available, is also proposed.

2 Preliminaries on composite likelihood prediction

2.1 Composite likelihood inference

Let $(Y_1, \ldots, Y_n, Y_{n+1})$ be a random vector with joint density $f(y_1, \ldots, y_{n+1}; \theta)$, with respect to a suitable dominating measure, specified by the unknown d-dimensional parameter $\theta \in \Theta \subseteq \mathbf{R}^d$, $d \geq 1$; $Y = (Y_1, \ldots, Y_n)$, $n > 1$, is observable, while $Z = Y_{n+1}$ is a future or not yet available observation. Given the observed sample $y = (y_1, \ldots, y_n)$, there are two general aims to be considered: the first one is to make inference on the unknown parameter θ, while the second one, which is the main objective of this paper, is to predict the future observation z by means of suitable point predictors, predictive densities or prediction intervals. Although the predictive procedures presented in the paper can be considered for both continuous and discrete random vectors, in order to simplify the exposition we confine the presentation to the continuous case.

Whenever the computation of the joint density $f(y; \theta)$, and then of the full likelihood function, is cumbersome or infeasible, we may consider alternative inferential methods based on a suitable surrogate of the true likelihood, such as the composite likelihood. Following Lindsay [10], a composite likelihood is simply defined as the (weighted) product

$$L_C(\theta; y) = \prod_{k=1}^{K} L_k(\theta; y)^{w_k} \tag{1}$$

of likelihood components $L_k(\theta; y)$, $k = 1, \ldots, K$, with $K \geq 1$, generated from low-dimensional marginal or conditional densities associated to $f(y; \theta)$. Hereafter, $w = \{w_k\}$ indicates a set of non-negative weights. In particular, we shall consider the pairwise marginal (PM) and the pairwise conditional (PC) likelihoods, obtained by combining bivariate marginal or conditional densities. They are given, respectively, by

$$L_{PM}(\theta; y) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} f(y_i, y_j; \theta)^{w_{ij}}, \quad L_{PC}(\theta; y) = \prod_{i=1}^{n} \prod_{j \neq i} f(y_i | y_j; \theta)^{w_{ij}},$$

where functions $f(\cdot; \theta)$ specify the marginal or the conditional density of the random variables involved in the argument. Furthermore, the one-wise (independence) likelihood is constructed under the independence assumption and it corresponds to

$$L_O(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta)^{w_i}.$$

A generalization of the notion of one-wise and pairwise likelihoods corresponds to consider, in the specification of the likelihood objects, blocks of observations with dimension greater than one. These pseudolikelihoods are usually called block composite (in particular, block one-wise or pairwise) likelihoods. Further marginal or conditional composite likelihoods are reviewed in [14].

Compared to the usual likelihood methods, composite likelihood inferential procedures are usually less efficient. In regular problems, provided that the parameter is identifiable, the maximum pairwise (marginal or conditional) likelihood estimator $\widehat{\theta}_P = \widehat{\theta}_P(Y)$, and in general the maximum composite likelihood estimator $\widehat{\theta}_C$, is consistent and asymptotically normally distributed, with asymptotic mean θ and the inverse of the Godambe information as asymptotic variance matrix. Thus, there is usually a loss of efficiency, which may be reduced with a suitable choice for the weights w and for the likelihood components $L_k(\theta; y)$, $k = 1, \ldots, K$ (see [8], [10], [14]).

2.2 Composite likelihood prediction

Since the computation of the model function $f(y; \theta)$ is not feasible, the specification of the conditional density $f(z|y; \theta)$ of the future random variable Z given $Y = y$ turns out to be not available either. In order to overcome this drawback, we shall consider a suitable composite predictive likelihood with the aim of defining a surrogate for $f(z|y; \theta)$ to be used as predictive density for Z.

The notion of predictive likelihood stems from the fact that in prediction problems there are two unknown quantities to deal with, namely the future observation z and the parameter θ, and the evidence about these two quantities is contained in the joint likelihood function $L(\theta; y, z) = f(y, z; \theta)$. Since our primary aim is to get information about z, with θ viewed as nuisance parameter, we consider here the simplest version of a predictive likelihood, termed estimative or plug-in predictive likelihood, which is defined as $L(\widehat{\theta}; y, z) = f(y, z; \widehat{\theta})$, where $\widehat{\theta} = \widehat{\theta}(y)$ is the maximum likelihood estimate (or an alternative consistent estimate) for θ. Note that, after normalization with respect to z, we obtain the estimative or plug-in predictive density $f(z|y; \widehat{\theta})$. A complete review on a number of alternative approaches to finding predictive likelihoods, which differ by the procedure adopted to deal with the nuisance parameter θ, is given by Bjørnstad [2].

In this framework, given a joint composite likelihood $L_C(\theta; y, z)$, obtained from (1) with (y, z) substituted for y, we define the estimative composite predictive likelihood by substituting the nuisance parameter θ with the estimate $\widehat{\theta}_C = \widehat{\theta}_C(y)$, and we get

$$L_C(\widehat{\theta}_C; y, z) \propto \prod_{h=1}^{H} L_h(\widehat{\theta}_C; y, z)^{w_h},$$

where only the likelihood components $L_h(\widehat{\theta}_C; y, z)$, $h = 1, \ldots, H$, $H \geq 1$, involving the future observation z are taken into account. The components not based on z are neglected, being not relevant for prediction purposes. Although function $L_C(\widehat{\theta}_C; y, z)$ depends on the weights w, this is not made explicit in order to simplify the notation. Alternative notions of composite predictive likelihood could be defined by mimicking those ones reviewed in [2]. However, these proposals, though quite complicated to compute, are usually first-order equivalent to the estimative one.

The predictive density obtained by considering the normalized version of $L_C(\widehat{\theta}_C; y, z)$, assuming that it is integrable, may be interpreted as a surrogate to the true density $f(z|y;\theta)$ and it can be used for prediction purposes instead of the true, unknown estimative predictive density $f(z|y;\widehat{\theta})$. In particular, we may obtain point predictors or prediction intervals for z. However, since the predictive conclusions are based on a misspecified or a partially specified model, their accuracy is affected by a potentially relevant misspecification error. For this reason, a suitable choice for the weights w is required, as we shall show in the following sections. Another potential difficulty is related to the additional uncertainty introduced by substituting θ with $\widehat{\theta}_C$, giving rise to the well-known plug-in error. Actually, this part of the error is not so relevant, provided that the sample size n is large enough and the estimator $\widehat{\theta}_C$ is proved to be consistent.

3 Pairwise predictive densities

3.1 Definition

This paper focuses on the estimative composite predictive likelihood based on both the one- and the two-dimensional marginal distributions, which is defined as

$$L_C(\widehat{\theta}_P; y, z) = L_O(\widehat{\theta}_P; y, z) L_{PM}(\widehat{\theta}_P; y, z) \propto f(z; \widehat{\theta}_P)^{w_0} \prod_{i=1}^{n} f(z, y_i; \widehat{\theta}_P)^{w_i}, \quad (2)$$

with $\widehat{\theta}_P = \widehat{\theta}_P(y)$ a suitable maximum pairwise likelihood estimate for θ. This function can be viewed as a simple surrogate to the full estimative predictive likelihood $L(\widehat{\theta}; y, z) = f(y, z; \widehat{\theta})$. Note that, if $L_{PC}(\widehat{\theta}_P; y, z)$ is considered instead of $L_{PM}(\widehat{\theta}_P; y, z)$, we obtain the same final expression as in (2). Thus, from a predictive view point, the marginal and the conditional pairwise likelihoods are in fact equivalent and, after normalization, they both define the estimative pairwise predictive density

$$f_P(z|y; \widehat{\theta}_P) = c(y, \widehat{\theta}_P, w) \prod_{i=0}^{n} f(z|y_i; \widehat{\theta}_P)^{w_i}, \quad (3)$$

with $c(y, \widehat{\theta}_P, w)$ the normalizing constant, assumed to be finite. Here, for simplifying the notation, we set $f(z|y_0; \widehat{\theta}_P) \equiv f(z; \widehat{\theta}_P)$. Indeed, we implicitly assume that the product of densities in (3) is integrable, so that it can be normalized to get a valid density function. Under this respect, a sufficient condition is that functions $f(z|y_i;\theta)$, $i = 0, \ldots, n$, are bounded probability densities for each $\theta \in \Theta$. This assumption holds, for example, for many probability densities belonging to the exponential family.

Expression (2) is rather general and it combines one-wise and pairwise composite likelihoods; when $w_0 = 0$, we get a predictive density based on the two-dimensional marginal distributions. Even if in most empirical applications we conveniently assume $w_0 = 0$, we maintain this formulation since $f(z; \widehat{\theta}_P)$ may indicate a preliminary or a prior guess on the distribution of the future random variable Z, specified under the independence assumption.

Function (3) can be interpreted as a misspecified version of the true density $f(z|y;\theta)$ and it determines the distribution which is in fact considered for prediction purposes, instead of the unknown estimative density $f(z|y;\widehat{\theta})$. It is almost immediate to see that

$$f(z|y_i;\theta) = f(z|y;\theta)\delta_i(y,z;\theta), \quad \delta_i(y,z;\theta) = \frac{f(y_{-i}|y_i;\theta)f(z|y_i;\theta)}{f(y_{-i},z|y_i;\theta)}, \tag{4}$$

$i = 1,\ldots,n$, with $y_{-i} = \{y_j, j \neq i\}$ and z such that $f(y_{-i},z|y_i;\theta) \neq 0$; for the case $i = 0$, $f(z|y_0;\theta) \equiv f(z;\theta)$ and $\delta_0(y,z;\theta) = f(y;\theta)f(z;\theta)/f(y,z;\theta)$. Note that, for $i = 1,\ldots,n$, $\delta_i(y,z;\theta) = 1$ if Z and Y_{-i} are conditionally independent given $Y_i = y_i$ and $\delta_0(y,z;\theta) = 1$ if Z and Y are independent. By substituting (4) in (3) we find that

$$f_P(z|y;\widehat{\theta}_P) = f(z|y;\widehat{\theta}_P)^{\sum_{i=0}^n w_i} c(y,\widehat{\theta}_P,w) \prod_{i=0}^n \delta_i(y,z;\widehat{\theta}_P)^{w_i},$$

so that the estimative pairwise predictive density is a function of the estimative predictive density $f(z|y;\widehat{\theta}_P)$ raised to the power $\sum_{i=0}^n w_i$ and, for this reason, it could be too much peaked or too much flat. Then a cautious assumption could be to consider weights normalized to sum up to one, even if this constraint can not be enough in order to have a density with a suitable shape and it may produce a sub-optimal predictive distributions. Relation (4) can be generalized to the case of block pairwise predictive densities, where blocks of observations are considered instead of single observations.

3.2 Point predictors and prediction intervals

The pairwise predictive density (3) can be considered to obtain point predictors or prediction intervals for the future observation z. In particular, as a point predictor for z we may consider the quantity $\widehat{z}_p = z_p(y;\widehat{\theta}_P,w)$ obtained by maximizing (3) or (2) with respect to z. The maximizer $\widehat{z}_p$ is called pairwise predictor and, for continuous random variables, it is usually specified as the solution with respect to z of the following score-type equation

$$\sum_{i=0}^n w_i \frac{d \log f(z|y_i;\widehat{\theta}_P)}{dz} = 0.$$

Alternative point predictors are the expected value, if it exists, or the median associated to the pairwise predictive density.

The predictive accuracy of a point predictor $\widehat{z}_p$ can be evaluated in terms of the associated (unconditional) mean square prediction error

$$MSPE(\widehat{z}_p, w) = E_{Y,Z}\left[\left\{Z - z_p(Y; \widehat{\theta}_P, w)\right\}^2\right], \tag{5}$$

where the expectation is with respect to the true joint distribution of (Y, Z). This quantity can be considered for choosing among competing point predictors and also for comparing different choices for the weights w. Although $\widehat{z}_p$ may be obtained quite easily using, if required, numerical optimization techniques, an explicit expression for $MSPE(\widehat{z}_p, w)$ is usually infeasible. An estimate can be obtained using a parametric bootstrap procedure, provided that simulated observations may be generated from the distribution of (Y, Z) with $\theta = \widehat{\theta}_P$. This happens, for example, within latent variable models, such as state space models and mixed models, where simulated samples can be easily obtained taking advantage of the hierarchical structure of the model. Then, if (y_b^*, z_b^*), $b = 1, \ldots, B$, are parametric bootstrap samples simulated from $f(y, z; \widehat{\theta}_P)$ and $\widehat{\theta}^*_{P,b}$, $b = 1, \ldots, B$, are the corresponding maximum pairwise likelihood estimates, the parametric bootstrap estimate for (5) is defined as

$$MSPE(\widehat{z}_p, w)_{boot} = \frac{1}{B}\sum_{b=1}^{B}\left\{z_b^* - z_p(y_b^*; \widehat{\theta}^*_{P,b}, w)\right\}^2. \tag{6}$$

Whenever the aim is to specify prediction intervals for z, we may consider suitable lower and upper prediction limits so that the coverage probability of the associated interval is equal or close to the required nominal value. Given the observed sample y, an α-prediction limit for Z is a quantity $c_\alpha(y)$ such that

$$\mathrm{pr}_{Y,Z}\{Z \leq c_\alpha(Y); \theta\} = \alpha,$$

for every $\theta \in \Theta$ and any fixed $\alpha \in (0, 1)$. This probability is called coverage probability and it is calculated with respect to the joint distribution of (Y, Z). In this framework, we may consider the estimative pairwise α-prediction limit $\widehat{z}_\alpha = z_\alpha(y; \widehat{\theta}_P)$ obtained as the α-quantile of the predictive density (3); namely, $\widehat{z}_\alpha$ is such that $F_P(\widehat{z}_\alpha | y; \widehat{\theta}_P) = \alpha$, with $F_P(\cdot | y; \widehat{\theta}_P)$ the pairwise predictive distribution function associated to (3).

As a consequence of both the model misspecification and the plug-in procedure, the actual coverage probability of $\widehat{z}_\alpha$ does not usually match the target nominal value α. More precisely, we have that

$$\begin{aligned}\mathrm{pr}_{Y,Z}\{Z \leq \widehat{z}_\alpha; \theta\} &= E_Y\{F(\widehat{z}_\alpha | Y; \theta)\} = E_Y\{F_P(\widehat{z}_\alpha | Y; \theta)\} \\ &\quad + E_Y\{F(\widehat{z}_\alpha | Y; \theta) - F_P(\widehat{z}_\alpha | Y; \theta)\} \\ &= \alpha + [E_Y\{F_P(\widehat{z}_\alpha | Y; \theta)\} - \alpha] + \\ &\quad E_Y\{F(\widehat{z}_\alpha | Y; \theta) - F_P(\widehat{z}_\alpha | Y; \theta)\}\end{aligned} \tag{7}$$

with $F(\cdot | Y; \theta)$ the true distribution function of Z given Y. Since an explicit expression for the coverage probability is not available, an estimate can be obtained using a simple parametric bootstrap simulation procedure similar to that one mentioned before, which can be applied whenever simulated observations may be generated from the distribution of (Y, Z) with $\theta = \widehat{\theta}_P$.

From equation (7) we may state that the coverage error term of $\widehat{z}_\alpha$ can be specified as the sum of two components. The first one, defined by the second term in (7), is related to the estimative procedure and it depends on the additional uncertainty introduced when θ is substituted by the estimator $\widehat{\theta}_P$. The second one, defined by the third term in (7), corresponds to the misspecification error due to the fact that the prediction limit $\widehat{z}_\alpha$ is calculated using the pairwise estimative predictive density rather than the true one. The plug-in error term, which is usually not relevant in case of large samples, can be substantially reduced by introducing analytical or simulation-based corrections (see [7], [13] and references therein). On the other hand, the misspecification error is usually remarkable, since it can produce a misleading assessment of the uncertainty of the predictive procedure, giving useless prediction intervals. A significant reduction of this source of error can be obtained by a suitable choice for the weights w, so that the discrepancy between the pairwise and the true predictive models is minimized, as described in the following subsection. Furthermore, an additional improvement could be achieved by taking into account further modifications for the pairwise predictive density, obtained by mimicking those ones introduced in the inferential framework (see, for example, [3] and [12]) for adjusting the magnitude and the curvature of a composite likelihood.

Finally, we emphasize that the two criteria introduced in this section for assessing, respectively, the accuracy of point predictors and prediction intervals are in fact unconditional. Although there is a strong motivation towards the fact that a prediction procedure should be judged conditionally on the observed value of the sample Y, or on the value of a suitable subset of Y, the conditional version of the mean square error and of the coverage probability are usually not computable in a closed form. Furthermore, parametric bootstrap estimates for these conditional quantities are hardly ever available, since to perform simulations from the distribution of Z given $Y = y$ is usually not possible. For this reason we focus only on non-conditional quantities so that the accuracy of prediction statements are evaluated under repeated random sampling from the unconditional distribution of (Y, Z).

3.3 Choice of the weights for improving prediction

In order to evaluate the closeness of the estimative pairwise predictive density $\widehat{f}_P = f_P(z|y; \widehat{\theta}_P)$ to the true density $f = f(z|y; \theta)$ we may consider the Kullback-Leibler divergence. For continuous random variables (analogous results may be obtained in the discrete case), it corresponds to

$$KL(f, \widehat{f}_P; w) = \int f(z|y; \theta) \log \frac{f(z|y; \theta)}{f_P(z|y; \widehat{\theta}_P)} \, dz$$

$$= E_{Z|Y=y}\left\{\log f(Z|y;\theta) - \log f_P(Z|y;\widehat{\theta}_P)\right\},$$

where the expectation is with respect to the true conditional distribution of Z given $Y = y$. It is easy to prove that

$$KL(f, \widehat{f}_P; w) = KL(f, \widehat{f}) + \int f(z|y;\theta) \log \frac{f(z|y;\widehat{\theta})}{f_P(z|y;\widehat{\theta}_P)}\, dz,$$

with $\widehat{f} = f(z|y, \widehat{\theta}_P)$, which points out the divergence contributions due to the plug-in procedure and to model misspecification, respectively.

In this context, we evaluate a predictive density $\widehat{f}_P$ by its Kullback-Leibler risk (expected loss) $R(f, \widehat{f}_P; w) = E_Y\{KL(f, \widehat{f}_P; w)\}$ or by the associated expected logarithmic score

$$EL(f, \widehat{f}_P; w) = E_{Y,Z}\left\{\log f_P(Z|Y;\widehat{\theta}_P)\right\}, \tag{8}$$

which may be estimated, using the bootstrap samples considered before, by

$$EL(f, \widehat{f}_P; w)_{boot} = \frac{1}{B}\sum_{b=1}^{B} \log f_P(z_b^*|y_b^*;\widehat{\theta}_{P,b}^*). \tag{9}$$

If we aim at comparing different predictive distributions, or in particular alternative choices for the weights, we look for the solution with higher expected logarithmic score and, consequently, lower Kullback-Leibler risk. Thus, the optimal weights with respect to the Kullback-Leibler loss will be defined as

$$w_{opt} = \arg\max_w \; EL(f, \widehat{f}_P; w),$$

subject to the following constraints: $w_i \geq 0$, $i = 0, \ldots, n$, and, possibly, $\sum_{i=0}^{n} w_i = 1$. Here, the bootstrap estimate (9) is usually considered as objective function instead of $EL(f, \widehat{f}_P; w)$.

Alternative (unconditional) criteria for weights selection can be defined. In particular, an optimality criterion based on the mean square prediction error may be considered. In this case, the optimal weights are such that

$$w_{opt} = \arg\min_w \; MSPE(\widehat{z}_p, w),$$

with the assumption that $w_i \geq 0$, $i = 0, \ldots, n$, and, possibly, $\sum_{i=0}^{n} w_i = 1$. The objective function $MSPE(\widehat{z}_p, w)$ is given by (5) and it may be substituted by its bootstrap estimate (6).

We will see in the section devoted to applications and simulation experiments that the simple criterion based on the mean square error can produce untrustworthy solutions. The selection procedure based on the Kullback-Leibler divergence, which may require an additional computational effort, turns out to be more reliable since it considers the entire predictive distribution, instead of the first two moments of the prediction error associated to the particular point predictor taken into account. Furthermore, both the procedures for obtaining optimal weights require a substantial computational effort, since the dimension of the vector parameter $w = (w_0, \ldots, w_n)$ could be quite large. In the applications presented in Section 5, in order to mitigate this potential difficulty, we run several times the optimization procedure by considering only the pairs with a maximum lag k, ranging from 1 to 10; namely, we impose that $w_i = 0$ for $i = 1, \ldots, n-k$. This particular choice is motivated by the short range dependence structure of the models taken into account and it reduces the dimension of the parameter space. In these applications we use an optimization procedure based on the Nelder and Mead algorithm and we find that optimality is usually reached for $k = 2$ or $k = 3$. Different initial values for the parameters are considered in order to reach a global maximum (or minimum).

A final interesting issue should be pointed out with regard to the approach considered in this paper, which distinguishes between the estimation and the prediction phases. We have considered so far the situation where the component density functions $f(z|y_i; \theta)$, $i = 0, \ldots, n$, are estimated by assuming $\theta = \widehat{\theta}_P$. Nevertheless, in most practical situations, the pairwise likelihood estimate $\widehat{\theta}_P$ is only a preliminary or initial guess for the unknown model parameter θ. Thus, it is immediate to realize that conditioning on this parameter estimate may give inaccurate predictive statements.

In order to improve the pairwise likelihood estimator for θ, and consequently the accuracy of predictive inference on z, a two-step iterative procedure can be defined, where a weights calibration step is followed by a pairwise estimation step, until a suitable stopping criterion is satisfied. More precisely, given a set of weights $w = (w_0, \ldots, w_n)$, where $w_i = 0$ for $i = 0, \ldots, n-k$ for a fixed lag $k = 1, \ldots, n$, the pairwise predictive density $f_P(z|y; \theta) = f_P(z|y_{n-k+1}, \ldots, y_n; \theta)$, under stationarity assumptions, can be viewed as a surrogate to the true conditional densities $f(y_j|y_{j-k}, \ldots, y_{j-1}; \theta)$ with $j = k+1, \ldots, n$. In many applications the optimal weights are all zero except the ones assigned to observations close to the future observation z (in a suitable temporal or spatial sense). Thus, the lag k is usually a rather low quantity. In this context, the following conditional pairwise likelihood

$$
\begin{aligned}
L^{\dagger}_{PC}(\theta; y) &= \prod_{j=k+1}^{n} f_P(y_j|y_{j-k}, \ldots, y_{j-1}; \theta) \\
&= \prod_{j=k+1}^{n} c(y_{j-k}, \ldots, y_{j-1}; \theta, w) \prod_{i=j-k}^{j-1} f(y_j|y_i; \theta)^{w_i}
\end{aligned}
$$

defines a misspecified version of the true conditional likelihood

$$
L^{\dagger}_{C}(\theta; y) = f(y_{k+1}, \ldots, y_n|y_1, \ldots, y_k; \theta) = \prod_{j=k+1}^{n} f(y_j|y_{j-k}, \ldots, y_{j-1}; \theta).
$$

Whenever the aim is to define an estimator $\widehat{\theta}$ for θ, obtained as the maximizer of the conditional pairwise likelihood $L^{\dagger}_{PC}(\theta; y)$ based on a suitable choice for the weights w, an iterative optimization scheme can be specified. The results of a preliminary simulation study, not reported in this paper, show that the estimates for θ, obtained using a sequential algorithm where the weights are selected using predictive criteria, like those recalled above, are usually poor in terms of efficiency. Although these negative findings are not surprising, since inference and prediction do not usually share the same objectives, this issue is certainly interesting and it deserves further consideration in future research.

4 Alternative views for the pairwise predictive density

This section gives two alternative views of the pairwise predictive density, which can be useful for interpreting and studying this new predictive tool within some well-known theoretical frameworks. Firstly, the predictive density (3) may be defined as the logarithmic combination of the forecast densities $f(z;\widehat{\theta}_P)$, $f(z|y_i;\widehat{\theta}_P)$, $i=1,\ldots,n$, which correspond to partially specified, and usually misspecified, (estimative) predictive models for Z given $Y=y$. The problem of combining density forecasts, also termed prediction pooling, is considered quite often in the econometric and in the quantitative finance literature (see, for example, [6], [9]). In this context the objective is to combine alternative predictive distributions so that the combined distribution provides much more accurate prediction statements than selecting a single model.

Among the possible ways of aggregations, we can mention the linear prediction pool, which defines a mixture density, and the logarithmic prediction pool, which is in fact considered in the specification of the predictive densities based on composite likelihood presented in this paper. Compared to the linear pool, the logarithmic one gives predictive densities which are typically unimodal and less dispersed. Moreover, if the weights w are normalized to sum up to one, the logarithmic combination method is invariant under rescaling and it verifies the property of external Bayesianity ([5], [1]).

This last property characterizes the logarithmic pool and it essentially means that it does not matter whether new information arrives before or after the pooling, since the update of the pooled distribution corresponds to the distribution obtained by applying the pooling procedure to the formerly updated component distributions. Namely, the operation of updating the component distributions with a common likelihood commutes with the pooling operator.

Secondly, the logarithmic pooling formula, which is considered as aggregation method for specifying the estimative pairwise predictive density, can be defined as the solution of a well-known optimization problem in information theory (see [4], chapter 12). More precisely, according to the maximum entropy principle [17], we look for a predictive density $p=p(z|y)$ defined as the solution of the following constrained optimization problem

$$\widehat{p} = \arg\min_p \, KL\left\{p(z|y), f(z;\widehat{\theta}_P)\right\}, \tag{10}$$

subject to

$$E_p \left\{ \log \frac{f(Z|y_i;\widehat{\theta}_P)}{f(Z;\widehat{\theta}_P)} \right\} = E_{Z|Y=y} \left\{ \log \frac{f(Z|y_i;\widehat{\theta}_P)}{f(Z;\widehat{\theta}_P)} \right\}, \quad i = 1, \ldots, n, \tag{11}$$

and the normalization condition $E_p(1) = 1$, where $E_p(\cdot)$ is the expectation with respect to the density $p(z|y)$. Thus, we find out a predictive density which is closest, in the Kullback-Leibler sense, to $f(z;\widehat{\theta}_P)$, that is the preliminary guess, under the independence assumption, of the prediction model for Z, and it reproduces the same conditional expectation as the true density $f(z|y;\theta)$ on functions $\log\{f(Z|y_i;\widehat{\theta}_P)/f(Z;\widehat{\theta}_P)\}$, $i = 1, \ldots, n$. If we do not consider any preliminary predictive model for Z, the optimization problem simplifies to $\widehat{p} = \arg\max_p \, [-E_p\{\log p(z|y)\}]$, that is $\widehat{p}$ maximizes the entropy, subject to the constraints (11) where, in this case, function $f(Z;\widehat{\theta}_P)$ is substituted by the constant 1.

By an application of the Lagrange multipliers, we obtain that the unique solution to (10) has the following form

$$\widehat{p}(z|y) = p(z|y;\widehat{\theta}_P,\lambda) = f(z;\widehat{\theta}_P) \exp\left\{ \sum_{i=1}^n \lambda_i \log \frac{f(z|y_i;\widehat{\theta}_P)}{f(z;\widehat{\theta}_P)} - K(y,\widehat{\theta}_P,\lambda) \right\}, \tag{12}$$

provided that parameters $\lambda = (\lambda_1, \ldots, \lambda_n)$ satisfy conditions (11) and quantity $K(y,\widehat{\theta}_P,\lambda)$ assures normalization. Function (12) corresponds to a multiparameter exponential family obtained as the tilting of the carrier density $f(z;\widehat{\theta}_P)$ in the directions spanned by $\log\{f(z|y_i;\widehat{\theta}_P)/f(z;\widehat{\theta}_P)\}$, $i = 1, \ldots, n$, and $K(y,\widehat{\theta}_P,\lambda)$ specifies the associated cumulant generating function. Furthermore, if $\lambda_i \geq 0$, $i = 1, \ldots, n$, and $\sum_{i=1}^n \lambda_i \leq 1$, density $p(z|y;\widehat{\theta}_P,\lambda)$ equals the estimative pairwise predictive density (3) with $w_i = \lambda_i$, $i = 1, \ldots, n$, $w_0 = 1 - \sum_{i=1}^n \lambda_i$ and $c(y,\widehat{\theta}_P,w) = \exp\{-K(y,\widehat{\theta}_P,\lambda)\}$. Note that, whenever differentiation and integration may be interchanged, the weights w satisfy the score-type equations

$$\frac{\partial E_{Z|Y=y}\left\{\log f_P(Z|y;\widehat{\theta}_P)\right\}}{\partial w_i} = 0, \quad i = 1, \ldots, n,$$

since $E_p[\log\{f(Z|y_i;\widehat{\theta}_P)/f(Z;\widehat{\theta}_P)\}] = \partial K(y,\widehat{\theta}_P,\lambda)/\partial\lambda_i$, $i = 1, \ldots, n$. Thus, under the usual regularity conditions, w maximizes the conditional expected logarithmic score $E_{Z|Y=y}\{\log f_P(Z|y;\widehat{\theta}_P)\}$.

However, since the optimality criterion for weights selection presented in Section 3.2 is based on the Kullback-Leibler risk, or equivalently on the expected logarithmic score $E_{Y,Z}\{\log f_P(Z|Y;\widehat{\theta}_P)\}$, the information theoretical interpretation of the estimative pairwise predictive density (3) has to be properly adapted. More precisely, we have to compute the expectation, with respect to Y, of the quantities considered in the constrained optimization problem (10)-(11). In this case, the non-negative weights $w_1, \ldots, w_n$, such that

$\sum_{i=1}^{n} w_i \leq 1$, satisfy the score-type equations and, under regularity conditions, they in fact maximize the unconditional expected logarithmic score (8). Thus, the solution corresponds to the optimal values w_{opt} defined in Section 3.3 using the Kullback-Leibler loss.

5 Simulation studies

5.1 Autoregressive models with additive observation noise

Let us consider the following simple linear Gaussian state space model, called first-order autoregressive model with additive observation noise,

$$\begin{aligned} Y_r &= \beta + X_r + V_r, \quad r \geq 1, \\ X_r &= \gamma X_{r-1} + W_r, \quad r \geq 1, \end{aligned}$$

with $V_r \sim N(0, \sigma^2)$, $W_r \sim N(0, \tau^2)$, $r \geq 1$, mutually independent Gaussian random variables. Let us assume that $X_0 \sim N(0, \tau^2/(1-\gamma^2))$ and that the latent autoregressive process X_r, $r \geq 0$, is stationary, being $|\gamma| < 1$. We observe $Y = (Y_1, \ldots, Y_n)$, $n \geq 1$, and we aim at predicting the future random variable $Z = Y_{n+1}$. The parameter $\theta = (\beta, \sigma^2, \gamma, \tau^2)$ is unknown. In this elementary example the likelihood function and the conditional density of Z given $Y = y$ are available in a closed form and they can be efficiently computed by means of Kalman filter recursions. Thus, we can compare the performance of the classical predictive procedures and of those ones based on the notion of pairwise predictive density. With regard to the inferential issues, we refer to the procedure based on the marginal pairwise likelihood proposed by Varin and Vidoni [16] for general state space models.

In this case the estimative pairwise predictive density can be specified quite easily since $Z|Y_i = y_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$, where $\mu_i = \beta + \rho_i(y_i - \beta)$ and $\sigma_i^2 = (1-\rho_i^2)\{\sigma^2 + \tau^2/(1-\gamma^2)\}$, with $\rho_i = \{\tau^2\gamma^{n+1-i}/(1-\gamma^2)\}/\{\sigma^2 + \tau^2/(1-\gamma^2)\}$ the correlation coefficient between Z and Y_i. From (3) we obtain that $f_P(z|y; \widehat{\theta}_P) = \phi(z; \widehat{\mu}_P, \widehat{\sigma}_P^2)$, namely it is a Gaussian density with mean $\widehat{\mu}_P$ and variance $\widehat{\sigma}_P^2$ given by

$$\mu_P = \frac{\sum_{i=1}^{n} \mu_i w_i/(1-\rho_i^2)}{\sum_{i=1}^{n} w_i/(1-\rho_i^2)}, \quad \sigma_P^2 = \frac{\sigma^2 + \tau^2/(1-\gamma^2)}{\sum_{i=1}^{n} w_i/(1-\rho_i^2)},$$

evaluated at $\theta = \widehat{\theta}_P$. Hereafter we do not consider $f(z; \widehat{\theta}_P)$, since in the simulation experiments we always obtain $w_0 = 0$ as optimal choice for w_0.

It is immediate to conclude that a suitable point predictor based on the pairwise predictive density is $\widehat{z}_p = \widehat{\mu}_P$. However, since θ is unknown, the associated mean square prediction error $MSPE(\widehat{z}_p, w)$ is not explicitly available, but it can be estimated using a simple parametric bootstrap procedure. We shall consider weights w_i, $i = 1, \ldots, n$, normalized to sum up to one and this is motivated by the fact that, otherwise, the predictive density (3) could be too much peaked or flat, as emphasized at the end of Section 3.1. Furthermore, it is interesting to observe that, at least in this example, the point predictor $\widehat{z}_p$, unlike the variance $\widehat{\sigma}_P^2$ of the pairwise predictive density, is invariant under scale transformation of the weights. Moreover, it is easy to see that also $MSPE(\widehat{z}_p, w)$ is invariant and

$$MSPE(\widehat{z}_p, w) = \text{var}(Z - \widehat{z}_p) = \text{var}(Z) + \text{var}(\widehat{z}_p) - 2\text{cov}(Z, \widehat{z}_p),$$

with

$$\begin{aligned}
\text{var}(Z) &= \sigma^2 + \tau^2/(1-\gamma^2), \\
\text{var}(\widehat{z}_p) &= \frac{1}{\{\sum_{i=1}^n w_i/(1-\rho_i^2)\}^2} \left\{ \sum_{i=1}^n \frac{w_i^2 \rho_i^2}{(1-\rho_i^2)^2} \text{var}(Y_i) + \right. \\
&\qquad \left. 2\sum_{i=2}^n \sum_{j<i} \frac{w_i w_j \rho_i \rho_j}{(1-\rho_i^2)^2} \text{cov}(Y_i, Y_j) \right\}, \\
\text{cov}(Z, \widehat{z}_p) &= \frac{1}{\sum_{i=1}^n w_i/(1-\rho_i^2)} \sum_{i=1}^n \frac{w_i \rho_i}{1-\rho_i^2} \text{cov}(Y_i, Z),
\end{aligned}$$

where $\text{var}(Y_i) = \text{var}(Z)$ and $\text{cov}(Y_i, Y_j) = \tau^2 \gamma^{|i-j|}/(1-\gamma^2)$, $i = 1, \ldots, n+1$, $j \neq i$.

We specify the following point predictors: $\widehat{z}_p^1$, obtained by considering all the pairs with $w_i = 1/n$, $i = 1, \ldots, n$, and $\widehat{z}_p^2(k)$, specified using only the pairs with a maximum lag k with $w_i = 1/k$, $i = n-k+1, \ldots, n$ and $w_i = 0$, $i = 1, \ldots, n-k$. We also take into account the true estimative predictor $\widehat{z}_{KF}$, based on Kalman filter recursions, and the point predictor $\widehat{z}_p^{opt}$ based on the optimal normalized weights minimizing the mean square prediction error $MSPE(\widehat{z}_p, w)$.

Samples of dimension $n = 200, 500$ are generated from a first-order autoregressive model plus additive observation noise, with $\beta = 0.2$, $\sigma = 1$, $\tau = 1$, (a) $\gamma = 0.5$ and (b) $\gamma = 0.95$ as true parameter values. Similar results are obtained with alternative choices for θ. In order to estimate the mean square prediction errors, we consider $5,000$ parametric bootstrap samples and the unknown parameter θ is estimated using the maximum (marginal) pairwise likelihood estimator of order six, as proposed in [16]. With regard to the optimal weights, we obtain, for the case (a), $w_i = 0$, $i = 1, \ldots, 497$, $w_{498} = 0.015$, $w_{499} = 0$, $w_{500} = 0.985$ for $n = 200$ and $w_i = 0$, $i = 1, \ldots, 498$, $w_{499} = 0.327$, $w_{500} = 0.673$ for $n = 500$ and, for the case (b), $w_i = 0$, $i = 1, \ldots, 498$, $w_{499} = 0.321$, $w_{500} = 0.688$ for $n = 200$ and $w_i = 0$, $i = 1, \ldots, 497$, $w_{498} = 0.089$, $w_{499} = 0.330$, $w_{500} = 0.581$ for $n = 500$. With samples having a smaller dimension n, we obtain unstable results, since in the present application the size of the bootstrap simulation procedure is conveniently reduced in order to lighten the computational burden.

The results are outlined in Table 1 and show that $\widehat{z}_p^{opt}$ presents values for the mean square prediction error which are very close to those ones related to the benchmark $\widehat{z}_{KF}$. Indeed, $\widehat{z}_p^2(3)$ has a better predictive performance than $\widehat{z}_p^2(6)$, which has the same order as the maximum pairwise likelihood estimator $\widehat{\theta}_P$. Notice that, in the more challenging situation related to case (b), the mean square prediction error of $\widehat{z}_p^1$, the predictor based on all the pairs, is very high. Moreover, since the model under consideration presents a short memory structure, we compute also the mean square prediction error of the simple predictor $\widehat{z}_p^2(1)$ based only on one lag pairs. For both the cases (a) and (b), the predictive performance is rather good and, in one circumstance, it equals that one of the optimal predictor.

			(a)			
n	$\widehat{z}_{KF}$	$\widehat{z}_p^{opt}$	$\widehat{z}_p^1$	$\widehat{z}_p^2(1)$	$\widehat{z}_p^2(3)$	$\widehat{z}_p^2(6)$
200	2.201	2.236	2.420	2.235	2.334	2.371
500	2.198	2.220	2.420	2.236	2.316	2.356
			(b)			
n	$\widehat{z}_{KF}$	$\widehat{z}_p^{opt}$	$\widehat{z}_p^1$	$\widehat{z}_p^2(1)$	$\widehat{z}_p^2(3)$	$\widehat{z}_p^2(6)$
200	2.737	2.832	7.636	2.949	3.072	3.624
500	2.695	2.698	9.698	2.902	2.733	3.193

Table 1 Mean square prediction error of the predictors $\widehat{z}_{KF}$, $\widehat{z}_p^{opt}$, $\widehat{z}_p^1$, $\widehat{z}_p^2(k)$, with $k = 1, 3, 6$, estimated using 5,000 bootstrap replications. Predictors based on samples of dimension $n = 200, 500$, from a first-order autoregressive model with additive observation noise with $\beta = 0.2$, $\sigma = 1$, $\tau = 1$, (a) $\gamma = 0.5$ and (b) $\gamma = 0.95$.

Although the mean square prediction error defines a valuable criterion for selecting point predictors, it could not be useful for choosing among alternative pairwise predictive densities, due to its invariance under rescaling of the weights. The point is that a set of weights, obtained as a scale transformation of the optimal ones, gives the same mean square prediction error but, on the contrary, the associated predictive density may have a different variance and it can produce misleading prediction intervals. For this reason we shall consider the Kullback-Leibler risk in order to evaluate pairwise predictive densities based on alternative choices for w and, in particular, to make a comparison with the true estimative predictive density, available in this case using Kalman filter recursions.

We consider the same simulated samples and the same $5,000$ parametric bootstrap replications used before and, in particular, we focus on the case $n = 500$ since we obtain more stable results. The optimal weights are now defined by maximizing the bootstrap estimate (9) of the expected logarithmic score $EL(f, \widehat{f}_P; w)$, which is equivalent to minimize the Kullback-Leibler risk. For the case (a), with $n = 500$, we have $w_i = 0$, $i = 1, \dots, 498$, $w_{499} = 0.334$, $w_{500} = 0.669$, corresponding to a $EL(f, \widehat{f}_P; w) = -1.819$, and for the case (b), with $n = 500$, $w_i = 0$, $i = 1, \dots, 497$, $w_{498} = 0.111$, $w_{499} = 0.365$, $w_{500} =$

0.626, corresponding to a $EL(f, \widehat{f}_P; w) = -1.919$. Here we do not impose the normalization constraint, otherwise we would have found suboptimal solutions.

It can be interesting to compare the values thus obtained, for the unconditional expected logarithmic score, with those corresponding to the optimal weights based on the mean square prediction error criterion. While for the case (a) the values nearly coincide, for case (b) we obtain the value -1.921, which is lower than the previous one, indicating a higher Kullback-Leibler loss. On the other hand the values for the mean square prediction error, when we use the optimal weights based on the Kullback-Leibler risk, are almost exactly the same as those presented in Table 1 for the optimal predictor.

In Figure 1 we compare the behaviour of the estimative pairwise predictive densities, based on the two alternative sets of optimal weights, with respect to the true estimative predictive density. The pairwise predictive densities based on all the pairs and on the pairs with a maximum lag 3 and 6 are also taken into account. We consider only the sample of dimension $n = 500$ simulated from a first-order autoregressive model with additive observation noise where (b) $\beta = 0.2$, $\sigma = 1$, $\tau = 1$, $\gamma = 0.95$, since in this case the differences are more pronounced. We find that, as expected, the pairwise predictive density based on the maximization of the expected logarithmic score turns out to be closer to the true estimative predictive density, obtained using the Kalman filter updating formula. Thus, at least in this case, it improves the predictive density with weights based on the mean square prediction error. Moreover, the predictive densities based on all the pairs or on the pairs with a maximum lag 3 and 6 present a poor behaviour.

Finally, at least in this preliminary analysis, we may state that the goodness of a pairwise predictive density, and then the usefulness of the associated prediction intervals, strongly depend on the specification of w. Under this respect, the definition of the weights according to the Kullack-Leibler risk seems to provide satisfactory results.

5.2 Autoregressive ordered probit models

The class of autoregressive ordered probit models (see, for example, [11]) specifies a flexible device for describing ordinal categorical time series. These models constitute a dynamic extension of the ordered probit models, which maintain the regression part and introduce a latent autoregressive time evolution. In this case, the likelihood function and the predictive distribution are not available in a closed form and the use of numerical procedures is problematic in terms of accuracy and computational effort. Thus, for inference and prediction we may conveniently exploit the procedures based on the notion of pairwise likelihood.

Let us consider an ordinal categorical time series Y_r, $r \geq 1$, described as a simple first-order autoregressive ordered probit model defined as follows. The random variable Y_r, $r \geq 1$, takes values in the ordered set $\{1, \ldots, D\}$, with $D > 1$ the number of categories, according to a censoring mechanism obtained

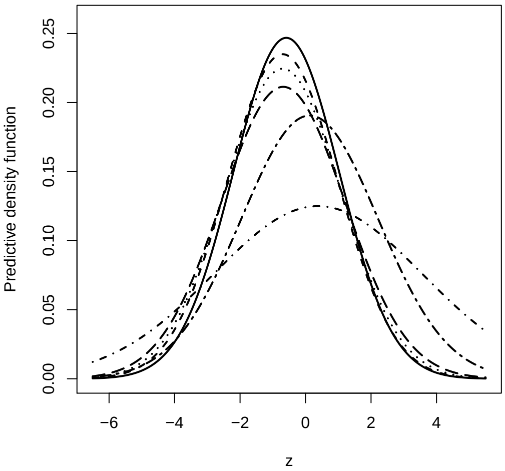

Fig. 1 Predictive densities based on a sample of dimension $n = 500$ from a first-order autoregressive model with additive observation noise with $\beta = 0.2$, $\sigma = 1$, $\tau = 1$, $\gamma = 0.95$: true estimative (solid line), estimative pairwise based on the optimal (expected logarithmic score criterion) weights (dashed line), based on the optimal (mean square prediction error criterion) weights (dotted line), based on all the pairs (dot-dashed line), based on the pairs with a maximum lag 3 (longdashed line) and with a maximum lag 6 (two-dashed line).

by partitioning the sample space of the latent random variable X_r, $r \geq 1$, into non-overlapping intervals. More precisely, we specify

$$\begin{aligned} &Y_r = 1, \quad \text{if and only if } -\infty < X_r \leq c_1, \\ &Y_r = d, \quad \text{if and only if } c_{d-1} < X_r \leq c_d, \quad d = 2, \ldots, D-1, \\ &Y_r = D, \quad \text{if and only if } c_{D-1} < X_r < +\infty, \end{aligned}$$

where the integer value $d = 1, \ldots, D$ is assigned to the d-th category and $\{c_1, \ldots, c_{D-1}\}$ is a set of ordered threshold parameters, with $c_0 = -\infty$ and $c_D = +\infty$. The latent first-order autoregressive process X_r, $r \geq 1$, is such that

$$X_r = \beta_0 + \beta_1 x_{r,1} + \ldots + \beta_p x_{r,p} + \gamma X_{r-1} + W_r, \quad r \geq 1,$$

with $W_r \sim N(0, \sigma^2)$, $r \geq 1$, independent Gaussian distributed random variables; $\beta = (\beta_0, \ldots, \beta_p)^T$ is the column vector of the regression coefficients and $x_r = (1, x_{r,1}, \ldots, x_{r,p})^T$ is the vector of the observed covariates at time r. We assume that $X_0 \sim N(0, \tau^2/(1-\gamma^2))$ and that the latent autoregressive process is stationary, being $|\gamma| < 1$. For overcoming identifiability problems we set $\sigma^2 = 1$ and $c_1 = 0$. Thus, $\theta = (c_2, \ldots, c_{D-1}, \beta_0, \ldots, \beta_p, \gamma)$ is the unknown parameter vector.

We observe $Y = (Y_1, \ldots, Y_n)$, $n \geq 1$, and we aim at predicting the future discrete random variable $Z = Y_{n+1}$. As mentioned before, in this context the classical predictive solutions are not available since the calculation of joint probabilities, such as $\mathrm{pr}(Y = y; \theta)$, requires the computation of intractable high-dimensional Gaussian integrals. For instance,

$$\mathrm{pr}(Y = y; \theta) = \int_{R(y)} p(x; \theta)\, dx_1 \cdots dx_n,$$

where $R(y) = \{x = (x_1, \ldots, x_n) : c_{y_i - 1} < x_i \leq c_{y_i}, i = 1, \ldots, n\}$ and $p(x; \theta)$ is the joint Gaussian density of $X = (X_1, \ldots, X_n)$. Thus, the unknown parameter θ is estimated using the procedure based on the marginal pairwise likelihood proposed by Varin and Vidoni [15] and, as predictive probability function for Z, we consider $f_P(z|y; \widehat{\theta}_P)$, where the weights are determined minimizing the Kullback-Leibler loss. Hereafter we do not consider $f(z; \widehat{\theta}_P)$, since in the simulation experiments we always obtain $w_0 = 0$ as optimal choice for w_0.

We shall present the results of a simple simulation study where we compare the unconditional expected logarithmic score $EL(f, \widehat{f}_P; w)$ of alternative pairwise predictive probability functions. We consider pairwise predictive distributions defined by considering all the pairs with equal normalized weights, namely $w_i = 1/n$, $i = 1, \ldots, n$, and the pairs with a maximum lag k, $k = 1, 2, 3, 5, 10, 50, 100$ with equal normalized weights, namely $w_i = 1/k$, $i = n - k + 1, \ldots, n$ and $w_i = 0$, $i = 1, \ldots, n - k$. Moreover, we consider also the pairwise predictive distribution based on the optimal weights obtained by maximizing the unconditional expected logarithmic score $EL(f, \widehat{f}_P; w)$.

Samples of dimension $n = 500$ are generated from a first-order autoregressive ordered probit model, also considered by Müller and Czado [11], where $D = 7$, $p = 2$ and the covariates x_1 and x_2 are obtained by simulating from a $N(-1, 1)$ and a $N(-0.25, 0.0324)$ distribution, respectively. The thresholds are $c_2 = 1.2$, $c_3 = 2.2$, $c_4 = 3.1$, $c_5 = 4.1$, $c_6 = 5.3$ and the regression coefficients are $\beta_0 = 2$, $\beta_1 = -0.6$, $\beta_2 = 9$; for the autoregressive parameter we consider the cases (a) $\gamma = 0.5$ and (b) $\gamma = 0.8$. Similar results are obtained with alternative parameter values. In order to estimate the unconditional expected logarithmic scores, we use $1,000$ parametric bootstrap samples and the unknown parameter θ is estimated using the maximum (marginal) pairwise likelihood estimator of order one, as proposed in [15]. With regard to the optimal weights, we obtain $w_i = 0$, $i = 1, \ldots, 498$, $w_{499} = 0.136$, $w_{500} = 0.871$, for the case (a), and $w_i = 0$, $i = 1, \ldots, 498$, $w_{499} = 0.058$, $w_{500} = 0.865$, for the case (b). Notice that these weights are not normalized to sum up to one. As in the simulation study presented in Section 5.1, in order to obtain more stable results, we consider large samples, since the size of the bootstrap simulation procedure is conveniently reduced in order to lighten the computational burden.

The results are outlined in Table 2 and show that the pairwise predictive distributions defined by considering the pairs with a lag $k = 1, 2$ present a predictive performance which is quite close to the that of the optimal (with respect to the Kullback-Leibler risk) solution. This conclusion is not unexpected, as the autoregressive ordered probit model under consideration presents a short range dependence structure. In the more challenging situation related to case (b), the differences in the estimated values are, obviously, more pronounced.

	k								
γ	opt	1	2	3	5	10	50	100	500
0.5	-0.477	-0.478	-0.482	-0.492	-0.502	-0.513	-0.522	-0.524	-0.525
0.8	-1.184	-1.187	-1.213	-1.253	-1.313	-1.398	-1.525	-1.547	-1.565

Table 2 Unconditional expected logarithmic scores of the pairwise predictive distributions based on the optimal weights and with equal normalized weights, based on pairs with a maximum lag $k = 1, 2, 3, 5, 10, 50, 100$ and on all the pairs ($k = 500$), estimated using 1,000 bootstrap replications. Predictive distributions computed on samples of dimension $n = 500$ from a first-order autoregressive ordered probit model with $D = 7$, $p = 2$, $c_2 = 1.2$, $c_3 = 2.2$, $c_4 = 3.1$, $c_5 = 4.1$, $c_6 = 5.3$, $\beta = (2, -0.6, 9)$ and (a) $\gamma = 0.5$, (b) $\gamma = 0.8$.

Furthermore, Figure 2 describes the behaviour of estimative pairwise predictive probability functions based on alternative choices for the weights, including the optimal one. We consider the same simulated sample as before and we focus, in particular, on the case (b), where the differences in the probabilities associated to the D categories may be substantial. Thus, at least in this situation, we may state that the predictive conclusions turn out to be significantly affected by the system of weights which is adopted.

In order to further emphasize how the choice of the weights may influence the predictive performance, we shall present an additional simple simulation study related to case (b), where we compare the mean square prediction error and the mean absolute prediction error of alternative point predictors based on different choices for the weights. As stated in Section 3.2, a point predictor for Z can be defined, also for a discrete predictand, as the category $\widehat{z}_p \in \{1, \ldots, D\}$ which maximizes the estimative pairwise predictive probability function $f_P(z|y; \widehat{\theta}_P)$. We consider the pairwise predictive distributions based on the optimal weights and on the normalized weights for pairs with maximum lag $k = 1, 2, 3, 5, 10, 50, 100$. The mean square prediction error $MSPE(\widehat{z}_p, w) = E_{Y,Z}\{(Z - \widehat{z}_p)^2\}$ and the mean absolute prediction error $MAPE(\widehat{z}_p, w) = E_{Y,Z}(|Z - \widehat{z}_p|)$ are estimated using the 1,000 bootstrap samples of dimension $n = 500$.

The results, outlined in Table 3, show that the pairwise predictive distributions based on the optimal weights (with respect to the Kullback-Leibler risk) and on the the pairs with a lag $k = 1$ present a similar predictive ability and they outperform the other predictive distributions in terms of both the mean square prediction error and the mean absolute prediction error. Namely, with a suitable choice for the weights, the mean prediction errors of the predictor $\widehat{z}_p$ may be considerably reduced.

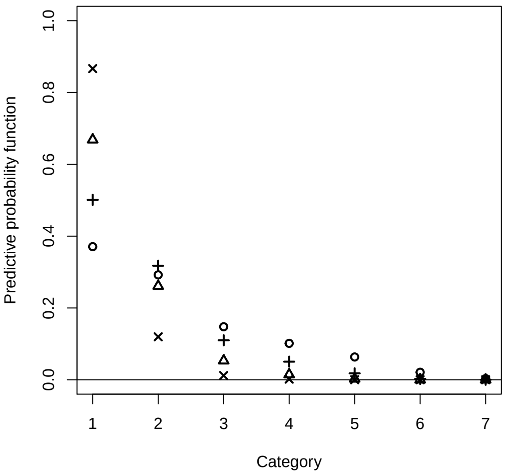

Fig. 2 Predictive probability functions based on the optimal weights (×) and on pairs with a maximum lag $k = 2$ (△), $k = 5$ (+) and $k = 50$ (∘), with normalized weights. Sample of dimension $n = 500$ from a first-order autoregressive ordered probit model with $D = 7$, $p = 2$, $c_2 = 1.2$, $c_3 = 2.2$, $c_4 = 3.1$, $c_5 = 4.1$, $c_6 = 5.3$, $\beta = (2, -0.6, 9)$ and $\gamma = 0.8$.

	opt	$k = 1$	$k = 2$	$k = 3$	$k = 5$	$k = 10$	$k = 50$	$k = 100$
$MSPE$	1.192	1.176	1.287	1.436	1.749	2.351	3.036	3.145
$MAPE$	0.730	0.728	0.755	0.798	0.881	1.015	1.186	1.215

Table 3 Mean square prediction error and mean absolute prediction error of point predictors associated to $f_P(z|y; \widehat{\theta}_P)$ with optimal weights (opt) and with equal normalized weights, based on pairs with maximum lag $k = 1, 2, 3, 5, 10, 50, 100$. Estimates obtained using 1,000 bootstrap samples of dimension $n = 500$ from a first-order autoregressive ordered probit model with $D = 7$, $p = 2$, $c_2 = 1.2$, $c_3 = 2.2$, $c_4 = 3.1$, $c_5 = 4.1$, $c_6 = 5.3$, $\beta = (2, -0.6, 9)$ and $\gamma = 0.8$.

6 Conclusions

Although the inferential procedures based on the notion of composite likelihood draw an increasing attention in the scientific literature, the problem of prediction, in the situation where the complete specification of the model is not available, is not considered with the same interest. In this paper, using the notion of weighted composite likelihood, and in particular the notion of weighted pairwise likelihood, we introduce a useful surrogate for the true unknown predictive density function, which can be considered for specifying predictors and prediction intervals.

Since this new predictive density is obtained as a weighted combination of partially specified predictive models, we emphasize that the specification of the weights and the selection of the component models is crucial for making reliable predictive inference. Under this respect, we propose criteria related to the mean square error of the associated predictor and to the divergence, in the Kullback-Leibler sense, between the true unknown predictive distribution and that one based on composite likelihood methods. This procedure, specifying optimal weights for composite likelihood prediction, could be possibly considered as a part of a sequential algorithm aiming at improving composite likelihood inference as well. However, the preliminary results on this iterative inferential procedure are not positive in terms of efficiency and they may constitute a matter for future research.

A further interesting issue, to be considered in future work, concerns the specification of an optimality criterion for weights selection, which aims at reducing the coverage error of the prediction intervals given by the pairwise predictive distribution. Moreover, an additional improvement could be achieved by suitably modifying the shape of the pairwise predictive density, as done in the inferential framework by [3] and [12] for adjusting the uncertainty assessment of the maximum composite likelihood estimators.

Acknowledgements The author thanks the anonymous referees for suggestions and comments which served to improve the paper. The research was partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003.

References

1. Allard, D., Comunian, A., Renard, P.: Probability aggregation methods in geoscience. Math. Geosci. 44, 545-581 (2012)
2. Bjørnstad, J.F.: Predictive likelihood: a review (with discussion). Statist. Sci. 5, 242-265 (1990)
3. Chandler, R.E., Bate, S.: Inference for clustered data using the independence loglikelihood. Biometrika 94, 167-183 (2007)
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory 2nd edn. Wiley, New York (2006)
5. Genest, C.: A characterization theorem for externally Bayesian groups. Ann. Statist. 12, 1100-1105 (1984)
6. Geweke, J., Amisano, G.: Optimal prediction pools. J. Econometrics. 164, 130-141 (2011)
7. Hall, P., Peng, L., Tajvidi, N.: On prediction intervals based on predictive likelihood or bootstrap methods. Biometrika 86, 871-880 (1999)
8. Joe, H., Lee, Y.: On weighting of bivariate margins in pairwise likelihood. J. Multivariate. Anal. 100, 670-685 (2009)
9. Kascha, C., Ravazzolo, F.: Combining inflation density forecasts. J. Forecast. 29, 231-250 (2010)
10. Lindsay, B.G.: Composite likelihood methods. Contemp. Math. 80, 221-239 (1988)
11. Müller, G., Czado, C.: An autoregressive ordered probit model with application to high frequency financial data. J. Comput. Graph. Statist. 14, 320-338 (2005)
12. Ribatet, M., Cooley, D., Davison, A.C.: Bayesian inference from composite likelihoods, with and application to spatial extremes. Statist. Sinica 22, 813-845 (2012)
13. Ueki, M., Fueda, K.: Adjusting estimative prediction limits. Biometrika 94, 509-511 (2007)

14. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. Statist. Sinica 21; 5-42 (2011)
15. Varin, C., Vidoni, P.: Pairwise likelihood inference for ordinal categorical time series. Comput. Statist. Data. Anal. 51, 2365-2373 (2006)
16. Varin, C., Vidoni, P.: Pairwise likelihood inference for general state space models. Econometric. Rev. 28, 170-185 (2009)
17. Zhu, S.C., Wu, Y.N., Mumford, D.: Minimax entropy principle and its application to texture modeling. Neural. Comput. 9, 1627-1660 (1997)