Model-free prediction

Politis, Dimitris
Department of Mathematics
University of California at San Diego
La Jolla, CA 92093-0112, USA
E-mail: dpolitis@ucsd.edu

Introduction

In the classical setting of an i.i.d. (independent and identically distributed) sample, the problem of prediction is not very interesting. Consequently, practitioners have focused on estimation and hypothesis testing in this case. However, when the i.i.d. assumption breaks down, the prediction problem is both important and intriguing; see Geisser (1993) for an introduction. Typical examples include regression problems and/or dependent data.
Some key models are given below. The data are $\{Y_t$, for $t = 1, \ldots, n\}$, the errors ε_t are assumed i.i.d. $(0, 1)$ throughout, and $\underline{X}_t$ is a fixed-length vector of explanatory (predictor) variables. Letters σ, a, a_i, b_i, etc. represent model parameters, $\underline{b}$ is a parameter vector, and $\mu(\cdot)$ and $s(\cdot)$ are functions.

- **Regression: linear/homoskedastic**

$$Y_t = \underline{X}_t' \, \underline{b} + \sigma\varepsilon_t, \quad t = 1, \ldots, n. \tag{1}$$

- **Regression: nonparametric/heteroskedastic**

$$Y_t = \mu(\underline{X}_t) + s(\underline{X}_t) \, \varepsilon_t, \quad t = 1, \ldots, n. \tag{2}$$

- **Time series: parametric (ARMA/ARCH)**

$$Y_t = b + \sum_{i=1}^{p} b_i Y_{t-i} + (a + \sum_{i=1}^{p} a_i Y_{t-i}^2)^{1/2} \cdot \varepsilon_t, \quad t = 1, \ldots, n. \tag{3}$$

- **Time series: nonparametric/heteroskedastic**

$$Y_t = \mu(Y_{t-1}, \cdots, Y_{t-p}; \underline{X}_t) + s(Y_{t-1}, \cdots, Y_{t-p}; \underline{X}_t) \, \varepsilon_t, \quad t = 1, \ldots, n. \tag{4}$$

The above examples represent some popular models for regression and time series data. The models are general enough to include possible heteroskedasticities in the error variance in addition to their potential nonparametric components. Given any one of these models, the optimal *model-based* predictors can be formed. Nevertheless, in what follows we show how the prediction problem can be addressed in a model-free setting.

A general model-free prediction principle

In models such as (1)—(4), the predictive distribution of Y_{n+1} given the data $\underline{Y}_n = (Y_1, \ldots, Y_n)'$ in general may depend on $\underline{Y}_n$ and on $\mathbf{X}_{n+1}$ which is a matrix of observable, explanatory (predictor) variables; for concreteness, assume the predictors are deterministic although provisions for random regressors can be made. The notation $\mathbf{X}_n$ here is cumulative, i.e., $\mathbf{X}_n$ is the collection of all predictor variables associated with the data $\underline{Y}_t$ for $t = 1, \ldots, n$; for instance, in the linear regression example of eq. (1), the matrix $\mathbf{X}_n$ would be formed by concatenating together all the fixed-length predictor vectors $\underline{X}_t, t = 1, \ldots, n$.

Let Y_t take values in the linear space $\mathbf{B}$ which typically will be $\mathbf{R}^d$ for some integer d. The goal is to predict $g(Y_{n+1})$ based on $\underline{Y}_n$ and $\mathbf{X}_{n+1}$ *without invoking any particular model*; here g is some real-valued (measurable) function. The key to successful model-free prediction is the following *model-free prediction principle*. In a nutshell, the basic idea is to transform the non-i.i.d. set-up to an i.i.d. dataset for which prediction is easy—even trivial—, and then transform back to the original setting to obtain the model-free prediction.

Model-Free Prediction Principle.

(a) For any natural number m, suppose that a transformation H_m is found that maps the data $\underline{Y}_m = (Y_1, \ldots, Y_m)'$ and the explanatory variables $\mathbf{X}_m$ onto the i.i.d. *sequence $\underline{\epsilon}_m^{(m)} = (\epsilon_1^{(m)}, \ldots, \epsilon_m^{(m)})'$ where each $\epsilon_i^{(m)}, i = 1, \ldots, m$ has distribution F_m, and F_m is such that $F_m \stackrel{\mathcal{L}}{\Longrightarrow}$ some F as $m \to \infty$.*

(b) Suppose that the transformation H_m is invertible for all m (possibly modulo some initial conditions denoted by IC), and—in particular—that one can solve for Y_m in terms of $\underline{Y}_{m-1}, \mathbf{X}_m$, and $\epsilon_m^{(m)}$ alone, i.e., that

$$Y_m = g_m(\underline{Y}_{m-1}, \mathbf{X}_m, \epsilon_m^{(m)}) \tag{5}$$

and

$$\underline{Y}_{m-1} = f_m(\mathbf{X}_m;\ \epsilon_1^{(m)}, \ldots, \epsilon_{m-1}^{(m)};\ IC) \tag{6}$$

for some functions g_m and f_m and for all $m = 1, 2, \ldots$.

(c) Then, the L_2-optimal model-free predictor of $g(Y_{n+1})$ on the basis of the data $\underline{Y}_n$ and the predictors $\mathbf{X}_{n+1}$ is given by the (conditional) expectation $\int G_{n+1}(\underline{Y}_n, \mathbf{X}_{n+1}, \epsilon) dF_{n+1}(\epsilon)$ where $G_{n+1} = g \circ g_{n+1}$ denotes composition of functions.

(d)The whole predictive distribution of $g(Y_{n+1})$ is given by the distribution of the random variable $G_{n+1}(\underline{Y}_n, \mathbf{X}_{n+1}, \epsilon_{n+1})$ where ϵ_{n+1} is drawn from distribution F_{n+1} and is independent to $\underline{Y}_n$. The median of this predictive distribution yields the L_1-optimal model-free predictor of $g(Y_{n+1})$ given $\underline{Y}_n$ and $\mathbf{X}_{n+1}$.

Typically, the distribution F_{n+1} will be unknown but it can be consistently estimated by $\hat{F}_n$, the empirical distribution of $\epsilon_1^{(n)}, \ldots, \epsilon_n^{(n)}$, under the assumed convergence in part (a). The estimator $\hat{F}_n$ can then be plugged-in to compute estimates of the aforementioned (conditional) mean, median, and predictive distribution.

The abovementioned predictive distribution in part (d), and the expectation in part (c) are both conditional on the value of $\underline{Y}_n$ (and the value of $\mathbf{X}_{n+1}$ when the latter is random). Note also the tacit understanding that the 'future' ϵ_{n+1} is independent to the conditioning variable $\underline{Y}_n$; this assumption

is directly implied by eq. (6) which itself—under some assumptions on the function g_m—could be obtained by iterating (back-solving) eq. (5). The presence of initial conditions such as IC in eq. (6) is familiar in time series problems of autoregressive nature where IC would typically represent values $Y_0, Y_{-1}, \ldots, Y_{-p}$ for a finite value p; the effect of the initial conditions is negligible for large n. Note that in regression problems the presence of initial conditions would only be required if the regression errors are not independent.

Remark 1 The above empirical estimates of the (conditional) mean and median would typically be quite accurate but the empirical estimate of the predictive distribution may be a bit too narrow, i.e., possessing a smaller variance and/or inter-quartile range than ideal. The reason is that a true predictive distribution should incorporate the variability of $\hat{F}_n$; in other words, the predictive distribution's width/scale should be an increasing function of the degree of uncertainty regarding the shape of F, i.e., the variance of $\hat{F}_n$, and the same is true concerning estimation/fitting of any parameters in the 'model-like' equation (5). The only general way to practically capture such a widening of the predictive distribution is given by *resampling* and/or *subsampling* methods should these be applicable in the setting at hand; see e.g. Efron and Tibshirani (1993), or Politis, Romano and Wolf (1999).

Remark 2 Eq. (5) with $\epsilon_i^{(m)}$ being i.i.d. from distribution F_m looks like a model equation but it is more general than a typical model. For one thing, the functions g_m and F_m may change with m, and so does $\epsilon_i^{(m)}$ which, in essense, is a triangular array of i.i.d. random variables. Furthermore, no assumptions are made *a priori* on the form of g_m. However, the process of starting without a model, and—by this transformation technique—arriving at a model-like equation deserves the name *model-free model-fitting* (MF^2, for short).

Remarks on model-free model-fitting

The prediction principle sounds deceptively simple but its application is not. The task of finding a set of candidate transformations H_n for any given particular set-up is challenging, and demands expertise and ingenuity. Once, however, a set of candidate transformations is identified (and denoted by $\mathcal{H}$), the procedure is easy to delineate: *Choose the transformation $H_n \in \mathcal{H}$ that minimizes the (pseudo)distance $d(\mathcal{L}(H_n(\underline{Y}_n)), \mathcal{F}_{iid,n})$ over all $H_n \in \mathcal{H}$;* here $\mathcal{L}(H_n(\underline{Y}_n))$ is the probability law of $H_n(\underline{Y}_n)$, and $\mathcal{F}_{iid,n}$ is the space of all distributions associated with an n-dimensional random vector whose $\mathbf{B}$–valued coordinates are i.i.d., i.e., the space of all distributions of the type $F \times F \times \cdots \times F$ where F is an arbitrary distribution on space $\mathbf{B}$. There are many choices of distance or pseudo-distance for d; see e.g. Hong and White (2005) and the references therein.

The application of the prediction principle appears similar in spirit to the Minimum Distance Method (MDM) of Wolfowitz (1957). Nevertheless, their objectives are quite different since MDM is typically employed for parameter estimation and hypothesis testing whereas in the prediction paradigm there is no interest in parameters. A typical MDM searches for the parameter $\hat{\theta}$ that minimizes the distance $d(\hat{F}_n, \mathcal{F}_\theta)$, i.e., the distance of the empirical distribution $\hat{F}_n$ to a parametric family $\mathcal{F}_\theta$. In this sense, it is apparent that MDM sets an ambitious target (the parametric family $\mathcal{F}_\theta$) but there is no necessity of actually 'hitting' this target. By contrast, the prediction principle sets the minimal target of independence but its successful application requires that this minimal target is more or less achieved.

Remark 3 If a model such as (1)—(4) is available, then the model itself suggests the form of the transformation H_n, and the residuals from model-fitting would serve as the 'transformed' values $\epsilon_t^{(n)}$. Of course, the goodness of the model should now be assessed in terms of achieved "i.i.d."—ness of these residuals. It is relatively straightforward—via the usual graphical methods—to check that the residuals have identical distributions but gauging their independence is trickier. Nevertheless, if the residuals happened to be (jointly) Gaussian, then checking their independence would be easier since in this case it would be equivalent to checking for correlation; for example, in the time series case a standard correlation test is the Ljung-Box.

The above ideas motivate the following variation of the prediction principle that may be of particular usefulness in the case of dependent data.

Transformation into Gaussianity as a Prediction Stepping Stone.
(a) For any natural number m, suppose that a transformation H_m on $\mathbf{B}^m$ is found that maps the data $\underline{Y}_m = (Y_1, \ldots, Y_m)'$ into the jointly Gaussian *vector $\underline{W}_m^{(m)} = (W_1^{(m)}, \ldots, W_m^{(m)})'$ with covariance matrix V_m whose eigenvalues—viewed as sequences in m—are bounded above and below by positive constants.*
(b) Also suppose that the transformation H_m is invertible (possibly modulo some initial conditions denoted by IC), and—in particular—that one can solve for Y_m in terms of $\underline{Y}_{m-1}, \mathbf{X}_m$, and $W_m^{(m)}$ alone, i.e., that

$$Y_m = \tilde{g}_m(\underline{Y}_{m-1}, \mathbf{X}_m, W_m^{(m)}) \tag{7}$$

and

$$\underline{Y}_{m-1} = \tilde{f}_m(\mathbf{X}_m;\ W_1^{(m)}, \ldots, W_{m-1}^{(m)};\ IC) \tag{8}$$

for some functions $\tilde{g}_m$ and $\tilde{f}_m$ for $m = 1, 2, \ldots$. Finally, define the vector $\underline{\epsilon}_m^{(m)} = (\epsilon_1^{(m)}, \ldots, \epsilon_m^{(m)})'$ to equal $V_m^{-1/2}\underline{W}_m^{(m)}$ where $V_m^{1/2}$ is a square root of matrix V_m. Note that $Y_m = \tilde{g}_m(\underline{Y}_{m-1}, \mathbf{X}_m, W_m^{(m)}) = \tilde{g}_m(\underline{Y}_{m-1}, \mathbf{X}_m, V_m^{1/2}\underline{\epsilon}_m^{(m)})$ which we can rename as $g_m(\underline{Y}_{m-1}, \mathbf{X}_m, \epsilon_m^{(m)})$ since the random vector $(\epsilon_1^{(m)}, \ldots, \epsilon_{m-1}^{(m)})'$ is related in a one-to-one fashion to $\underline{Y}_{m-1}$ (by induction on m).
Let F_n denote the common normal distribution of $\epsilon_1^{(n)}, \ldots, \epsilon_n^{(n)}$ that are i.i.d. by construction. Then, the L_1 and L_2–optimal model-free predictors and the predictive distribution of $g(Y_{n+1})$ given $\underline{Y}_n$ and $\mathbf{X}_{n+1}$ are given verbatim *by parts (c) and (d) of the Prediction Principle.*

In applications, the covariance matrix V_n must be estimated from $W_1^{(n)}, \ldots, W_n^{(n)}$ using some extra assumption on its structure (e.g., a Toeplitz structure in stationary time series), or an appropriate shrinkage and/or regularization technique—see e.g. Bickel and Li (2006) and the references therein; then, the estimate $\hat{V}_n$ must be extrapolated to give an estimate of V_{n+1}. As before, the distribution F_{n+1} can be consistently estimated by $\hat{F}_n$, the empirical distribution of $\epsilon_1^{(n)}, \ldots, \epsilon_n^{(n)}$, or by a Gaussian distribution with unit variance and estimated mean; the former option may be more robust in practice. Applying the Gaussian 'stepping stone' can be formalized in much the same way as before. To elaborate, once $\mathcal{H}$, the set of candidate transformations is identified, the procedure is to: *choose the transformation $H_n \in \mathcal{H}$ that minimizes the distance $d(\mathcal{L}(H_n(\underline{Y}_n)), \Phi_n)$ over all $H_n \in \mathcal{H}$* where now Φ_n is the space of all n-dimensional Gaussian distributions on $\mathbf{B}$. Many choices for the distance d are

again available, including usual goodness-of-fit favorites such as the Kolmogorov-Smirnov or χ^2 test; a pseudo-distance based on the Shapiro-Wilk statistic is also a valid alternative.
However, now that H_n is essentially a *normalizing* transformation, a collection of graphical and exploratory data analysis (EDA) tools are also available to facilitate this search. Some of these tools include: (a) Q-Q plots of the $W_1^{(n)}, \ldots, W_n^{(n)}$ data to test for Gaussianity; (b) Q-Q plots of linear combinations of $W_1^{(n)}, \ldots, W_n^{(n)}$ to test for *joint* Gaussianity; and (c) autocorrelation plots of $\epsilon_1^{(n)}, \ldots, \epsilon_n^{(n)}$ to test for independence—since in the (jointly) Gaussian case, independence is tantamount to zero correlation.

Application to financial time series

We now consider data $Y_1, \ldots, Y_n$ arising as an observed stretch from a financial returns time series $\{Y_t, t \in \mathbf{Z}\}$ such as the percentage returns of a stock price, stock index or foreign exchange rate. The series $\{Y_t\}$ will be assumed stationary with mean zero which—from a practical point of view—implies that trends and other nonstationarities have been successfully removed.
The modeling work-horse in such a context is given by the well-known ARCH/GARCH models. The simplest ARCH(p) model of Engle (1982) is described by eq. (3) with $b = b_1 = b_2 = \cdots = 0$ and errors ε_t that are i.i.d. $N(0,1)$. Observe that, under such an ARCH(p) model, the quantity $\varepsilon_t = Y_t(a + \sum_{i=1}^p a_i Y_{t-i}^2)^{-1/2}$ is thought of as perfectly normalized and variance–stabilized as it is assumed to be i.i.d. $N(0,1)$. Thus, as in Remark 3, the ARCH model seems to suggest the form of a normalizing transformation.
However, as practitioners realized early-on, the residuals from ARCH and GARCH fitting do *not* look normal; see e.g. Shephard (1996). Rather than discarding the ARCH model altogether, we may instead attempt to tweak it in order to obtain a proper normalizing transformation for the Gaussian 'stepping stone' prediction approach. Note that the ARCH model residuals appear to be studentized returns, i.e., the return divided by a (time-localized) estimate of its standard deviation. But there is no reason—other than coming up with a neat model—to exclude the value of Y_n from an empirical estimate of the standard deviation of the same Y_n based on the data $\{Y_s, s \leq n\}$.
So, we may define a new studentized quantity

$$W_t := \frac{Y_t}{\sqrt{\alpha s_{t-1}^2 + a_0 Y_t^2 + \sum_{i=1}^p a_i Y_{t-i}^2}} \quad \text{for} \quad t = p+1, p+2, \ldots, n \tag{9}$$

with $W_t = Y_t$ for $t = 1, \ldots, p$; here $s_{t-1}^2 = (t-1)^{-1} \sum_{k=1}^{t-1} Y_k^2$ is an estimator of $Var(Y_1)$. The order p and the vector of nonnegative parameters $(\alpha, a_0, \ldots, a_p)$ are chosen by the practitioner with the normalization of W_t as target. As shown in Politis (2003a,b), it is always possible to find data-based configurations of the above parameters so that the normalization goal is indeed achieved. Note that eq. (9) can be uniquely solved for Y_t to give:

$$Y_t = \frac{W_t}{\sqrt{1 - a_0 W_t^2}} \sqrt{\alpha s_{t-1}^2 + \sum_{i=1}^p a_i Y_{t-i}^2} \quad \text{for} \quad t = p+1, p+2, \ldots, n. \tag{10}$$

Thus, as desired, the transformation from $\underline{Y}_n$ to $\underline{W}_n$ is invertible (given the initial conditions $Y_1, \ldots, Y_p$) with an explicit formula relating Y_t to $W_1, \ldots, W_t$, i.e., a 'model-like' equation of type (5).

Eq. (10) looks like a regular ARCH(p) model with the non-normal errors $U_t = W_t/\sqrt{1 - a_0 W_t^2}$. Now if the W_ts are deemed to be independent, then this is indeed true and leads to a particular suggestion for a heavy-tail distibution for the errors; see Politis (2004). However, if the W_ts are not independent, then the prediction principle associated with eq. (10) will yield quite different predictions than an ARCH model with heavy-tailed errors. Furhermore, if the objective is just prediction, say of $g(Y_{n+1})$, then the predictor follows immediately from the prediction principle, and modelling issues are superfluous. For example, in predicting squared returns, i.e., $g(x) = x^2$, Politis (2007) shows empirically that the model-free prediction principle outperforms predictors arising from ARCH/GARCH models with normal and/or heavy-tailed errors. In addition, it is conceivable that the prediction principle would be more robust in practice since it is totally data-based and, in the end, an empirical distribution is the driving force. This robustness is being corroborated by simulations and real data examples in Politis and Thomakos (2007) who show the superior performance of model-free predictors in many problematic settings such as time series with structural breaks, regime switching, local—as opposed to global—stationarity, etc.

REFERENCES (RÉFERENCES)

Bickel, P. and Li, B. (2006). Regularization in Statistics, *Test*, vol. 15, no. 2, 271-344.

Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation, *Econometrica*, 50, 987-1008.

Geisser, S. (1993). *Predictive Inference: An Introduction*, Chapman and Hall, New York.

Hong, Y. and White, H. (2005). Asymptotic distribution theory for nonparametric entropy measures of serial dependence, *Econometrica*, Vol. 73, No. 3, 837-901.

Politis, D.N. (2003a). Model-free volatility prediction. UCSD Dept. of Economics Discussion Paper 2003-16. [http://repositories.cdlib.org/ ucsdecon/2003-16]

Politis, D.N. (2003b). A normalizing and variance-stabilizing transformation for financial time series, in *Recent Advances and Trends in Nonparametric Statistics*, (M.G. Akritas and D.N. Politis, Eds.), Elsevier (North Holland), pp. 335-347.

Politis, D.N. (2004). A heavy-tailed distribution for ARCH residuals with application to volatility prediction, *Annals of Economics and Finance*, vol. 5, pp. 283-298.

Politis, D.N. (2007). Model-free vs. model-based volatility prediction. *J. Financial Econometrics*, vol. 5, no. 3, pp. 1-31.

Politis, D.N., Romano, J.P. and Wolf, M. (1999), *Subsampling,* Springer Verlag, New York, 1999.

Politis, D.N., and Thomakos, D. (2007). NoVaS transformations: flexible inference for volatility forecasting, working paper.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In *Time Series Models in Econometrics, Finance and Other Fields*, D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen (eds.), London: Chapman & Hall, pp. 1-67.

Wolfowitz, J. (1957). The minimum distance method, *Ann. Math. Statist.*, 28, 75–88.

ABSTRACT

Some principles of model-free prediction are laid out based on the notion of transforming a given set-up into one that is easier to work with, e.g., i.i.d. or Gaussian. An application to financial time series is discussed in detail, namely the problem of prediction of squared returns. As it turns out, the transformation technique outperforms the well-known ARCH/GARCH models in terms of predictive accuracy.

RÉSUMÉ

Quelques principes de prévision sans-modèle sont présentés ont basé sur la notion de transformer une installation donnée en une avec laquelle est plus facile de travailler, par exemple, indépendance ou Normalité. Une application à la série chronologique financière est discutée en détail, à savoir le problème de la prévision des retours carrés. Pendant qu'elle s'avère, la technique de transformation surpasse les modèles bien connus ARCH/GARCH en termes d'exactitude prédictive.