九州大学学術情報リポジトリ
Kyushu University Institutional Repository

Time Series Analysis and Stochastic Prediction (III)

Ogawara, Masami
Tokyo Woman's Christian College

https://doi.org/10.5109/12988

出版情報：統計数理研究. 9 (1), pp.1-9, 1960-03. Research Association of Statistical Sciences
バージョン：
権利関係：

KYUSHU UNIVERSITY

TIME SERIES ANALYSIS AND STOCHASTIC PREDICTION (III)

By

Masami OGAWARA

(*Received September* 1, 1958)

Chapter VI. Efficiency of a Stochastic Prediction*

§ 6. 1. **Introduction.** In this chapter we shall include the fiducial prediction into the stochastic prediction in the wide sense. A characteristic point of our stochastic prediction consists in the expression of possibilities of future value in a form of distribution function in relation to the finite data observed in the past which becomes the conditional probability distribution in the limit case, while the usual theory of prediction seems to concern exclusively with the expected value or the linear least squares prediction value and the variance or the mean square error based on the data back to infinite past. On the other hand, as for the usual form in the practice of prediction, e. g. the weather forecasting, it seems that only the value which is considered to be most probable is predicted—let us call provisionally such forecasting the *decisional prediction.*

We shall be able to criticize various schemes of prediction from the stand point of an operations research. The operational efficiencies of a prediction have been discussed by several meteorologists, e.g. J. C. Thompson [37], [38], but they are confined to the case of categorical and decisional prediction. In the following paragraph, the efficiency of a prediction will be defined in a general and reasonable way.

Before proceed to the subject, we shall give some preliminary considerations on the precision of a stochastic prediction. In this chapter we denote the stochastic prediction of y by $F(y \mid x)$ where the x stands for an united set of fixed condition variates and other sample values of a time series.**

On the other hand, we shall denote an estimation of the distribution function of unconditional random variable y by $F(y)$, which should be defined, if possible, by an unconditional stochastic prediction of y such as (4.3.2). In general, however, the unconditional stochastic prediction is not always obtainable. In that case we use the ordinary estimation, the empirical distribution function of y instead. Any way, for a large sample, the $F(y \mid x)$

* Ogawara [10], [12], [17], [21].

** We shall denote by the x both a set of variates and a multidimensional variate.

and the $F(y)$ is approximately the conditional and the absolute distribution function of y respectively.

Definition. *Let y be a k dimensional predictand ($k \geq 1$), let the total variance of $F(y \mid x)$ and $F(y)$ be $\sigma^2(x)$ and σ^2 respectively, let $G(x)$ be the distribution function of x, $\sigma_s^2 = \int \sigma^2(x) dG(x)$, and suppose that*

$$F(y) = \int F(y \mid x) dG(x), \tag{6.1.1}$$

then

$$p_s = (\sigma^2 - \sigma_s^2)/\sigma^2 \tag{6.1.2}$$

*is said to be the precision of the stochastic prediction $F(y \mid x)$.**

Under the assumption (6.1.1), it is evident that $\sigma_s^2 \leqq \sigma^2$ and $0 \leqq p_s \leqq 1$.

Now, when a one dimensional time series $x_{t-\tau}$ ($\tau = 0, 1, \cdots, L$) is given, suppose that we can assume, for instance, a normal autoregression scheme of order h, then the stochastic prediction of $x(t+s)$ is given by the distribution of (4.2.2) and, roughly speaking, the variance σ_s^2 is proportional to $N/(N-h-1) \fallingdotseq (L-h)/(L-(h+1)(h+s+1))$, neglecting the other factors which do not remarkably decrease with N larger than certain value. Consequently, if h is too large the precision rather decreases. In general, the model faithful to the nature may be complex. However, if we adopt a model which involves too many unknown parameters, we need sufficiently long time series, otherwise the precision of the prediction will rather fall. So, for a given time series, we should assume preferably simple scheme, so far as it is accepted on a significance level.

§ **6.2. Efficiency of a prediction.** Suppose we suffer a *loss* $L(c,y)$ by a realization of a stochastic quantity y, even when we do a *protection* at the *cost* c, where the function $L(c, y)$ may be negative when the y and the protection cost c bring a profit and where we assume that psychological damage and cost can be quantitatively measured. In such circumstances, we should do a protection c such that the *risk*

$$R(c, y) = c + L(c, y) \tag{6.2.1}$$

is minimum. However, we can not, in general, deterministically know the value of y that will be realized in future.

If we only know the (estimation of) absolute probability distribution of the y, $F(y)$, we may take the protection $c = c_a$ such that the expected risk

$$R_a(c) = \int R(c, y)\ dF(y) \tag{6.2.2}$$

is minimum. Thus, $R_a \equiv R_a(c_a)$ is the unavoidable mean minimum risk**

* $p_s(x) = (\sigma^2 - \sigma^2(x))/\sigma^2$ will be sometimes called the *individual precision* of the stochastic prediction.

** If we denote the mean value of $c = c(y)$ which minimize the $R(c, y)$ by $\bar{c} = \int c(y) dF(y)$, then $R_a \leqq \int R(\bar{c}, y) dF(y)$.

when we do not use any occasional prediction and constantly apply the protection c_a. For instance, in case of weather forecasting, $F(y)$ may be consistent to the climatic frequency distribution of a predictand y.

In order to make the best use of a stochastic prediction $F(y|x)$, we may choose the protection $c(x)$ such that $R_s(c,x) \equiv \int R(c,y)\, dF(y|x)$ is minimum on each occasion. The $c(x)$ may be called the *effective protection cost* under the condition x. If we denote the (multivariate) distribution function of x by $G(x)$, the mean value of $R_s(x) \equiv R_s(c(x), x)$,

$$R_s \equiv \int R_s(x)\, dG(x), \tag{6.2.3}$$

is the *mean minimum risk* in the case when we use the stochastic prediction on every occasion and, the *mean effective protection cost* in that case, is

$$c_s \equiv \int c(x)\, dG(x). \tag{6.2.4}$$

Suppose, as a special case, that we can deterministically foreknow the value of y as a function of the preceding condition x, $y=m(x)$—let us call this a *deterministic prediction*—, then we can choose the protection cost $c=c_0(x)$ such that $R_0(c,x) \equiv R(c, m(x))$ is minimum. Even in this case we can not avoid a risk $R_0 \equiv \int R_0(x) dG(x)$ in the mean, where $R_0(x) \equiv R_0(c_0(x), x)$, and the mean effective protection cost is $c_0 \equiv \int c_0(x)\, dG(x)$.

In our usual case, the case when we can not deterministically foreknow the y, suppose that we can only predict the expected value of y, $m(x) = \int y\, dF(y|x)$, on the knowledge of x and that we do not know the stochastic prediction $F(y|x)$ itself, then it is the only course open to us to proceed depending on the decisional prediction $m(x)$. Thus we choose the protection cost $c^*(x)$ such that $R^*(c,x) \equiv R(c, m(x))$ is minimum. However, various values of y are virtually possible according to the distribution $F(y|x)$, if we have no information other than the x. Therefore, the actually expected risk for that decisional prediction is given by $R^*(x) \equiv \int R(c^*(x), y)\, dF(y|x)$, and the mean minimum risk for the succession of this kind of prediction is given by

$$R^* \equiv \int R^*(x) dG(x) \tag{6.2.5}$$

and the mean effective protection cost is $c^* = \int c^*(x) dG(x)$.

Definition. *Let $L(c,y)$ be a loss function integrable with respect to any distribution function of y and let $R(c,y) = c + L(c,y)$. Let $H(y|x)$ be any prediction, based on a set of preceding condition variates x, in the form of a distribution function, let $F(y|x)$ be a stochastic prediction of y or an (estimated) conditional distribution of y given x, let $G(x)$ be the (estimated) distribution function of x, and let*

$$\inf_c \int R(c,y)\, dH(y|x) \equiv \int R(c(x), y) dH(y|x), \tag{6.2.6}$$

$$R(x) \equiv \int R(c(x), y) dF(y|x), \tag{6.2.7}$$

$$R \equiv \int R(x)dG(x). \tag{6.2.8}$$

Then a number

$$e = (R_a - R)/|R_a| \tag{6.2.9}$$

is said to be the efficiency of the prediction $H(y|x)$ *with respect to the loss function* $L(c, y)$, *where* $R_a = \inf_c \int R(c,y)\, dF(y)$, $F(y)$ *being the estimated absolute distribution function of* y.

For a deterministic prediction, $H(y|x)$ and $F(y|x)$ are the same unit distribution, and, for a decisional prediction, $H(y|x)$ is only an unit distribution.

The efficiency of a prediction defined above by a sample indicates, when the prediction represented in the form of a distribution function is most effectively and repeatedly used, how much the mean minimum risk R decreases relatively to the mean minimum risk R_a in the case where only the prediction by the estimated absolute distribution is constantly used, provided that the same stochastic model is applicable on each occasion. According to our definition, the efficiency of a prediction depends on the loss function $L(c, y)$ as well as the prediction scheme and it may assume values less than zero or larger than one, but if we assume without loss of generality that $L(c, y) \geqq 0$ for all c and y, then we have $0 \leqq e \leqq 1$ for any stochastic prediction, as we see in the next paragraph.

We did not take the cost required for the preparation of a prediction, say c', into our consideration. If we introduce it in our theory, the prediction and the risk function may be written as $H(y|x, c')$ and $R(c, c', y) \equiv c + c' + L(c, y)$ respectively, and we should minimize the function $\int R(c, c', y)\, dF(y|x, c')$ with respect to c and c', and the generalized effective protection cost and the minimum risk may be written as $C(x) = c(x) + c'(x)$ and $R(x) = \int R(c(x), c'(x); y)dF(y|x)$ respectively.

§ 6.3. **Comparison of efficiencies.** Assume that a loss function $L(c, y, t)$ which depends on the time t has a finite expectation for all $c \geqq 0$ and t, let $x_1 = x_1(t)$ and $x_2 = x_2(t)$ be two subsequences, corresponding to time t, of a stationary time series, and let the minimum risk functions of two predictions (which may not be necessarily stochastic) $F_1(y|x_1)$ and $F_2(y|x_2)$ $(y = y(t+s), s > 0)$ be $R_1(x_1, t)$ and $R_2(x_2, t)$, and let us define the mean minimum risks by

$$R_i \equiv \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} dt \int R_i(x_i, t)dG(x_i) \quad (i = 1, 2)$$

respectively, where $G(x_i)$ is the (estimated) distribution function of x_i. If then $R_1(x_1, t) \leqq R_2(x_2, t)$ for all t, we say that $F_1(y|x_1)$ is *individually more effective* than $F_2(y|x_2)$ with respect to the loss function $L(c, y, t)$, and if $R_1 < R_2$, $F_1(y|x_1)$ is said to be *more effective in the mean* than $F_2(y|x_2)$

with respect to the loss function. If $F_1(y \mid x_1)$ is individually more effective than $F_2(y \mid x_2)$, the former is more effective in the mean than the latter (with respect to $L(c, y, t)$).

In the following, we assume that the loss function is independent of t as was in the last paragraph, and we shall consider the relations in magnitude between efficiencies which hold for any loss function. The following lemma will be useful for the comparison of efficiencies.

Lemma 6. *If a function $\varphi(c, y)$ $(0 \leqq c < \infty)$ and $\inf_c \varphi(c, y)$ are integrable with respect to a distribution function $\Phi(y)$, then*

$$\int_D \{\inf_c \varphi(c, y)\} d\Phi(y) \leqq \inf_c \int_D \varphi(c, y) d\Phi(y), \tag{6.3.1}$$

where D is the entire space of y.

Theorem 21. *Using the notations in the preceding paragraph, the following propositions hold for the same predictand and for any loss function.*

i) $R_s \leqq R_a$,

consequently, for the efficiency e_s of an arbitrary stochastic prediction, we have $0 \leqq e_s$, and, if $L(c, y) \geqq 0$ for all c and y,

$$0 \leqq e_s \leqq 1. \tag{6.3.2}$$

ii) *Stochastic prediction is individually most efficient among the predictions based on the same set of preceding condition variates.*

iii) *The case of $R^* > R_s$ is possible; that is, the efficiency of a decisional prediction may be negative in certain circumstances.*

iv) *The deterministic prediction has the largest efficiency in the mean.*

Proof. i) follows from (6.3.1) by putting

$$\varphi(c, x) \equiv \int R(c, y)\ dF(y \mid x) \text{ and } \Phi(x) \equiv G(x).$$

ii) Let $F(y \mid x)$ be a stochastic prediction and $H(y \mid x)$ be any prediction. From (6.2.6) and (6.2.7) we have

$$\begin{aligned} R_s(x) &= \inf_c \int R(c, y)\ dF(y \mid x) \\ &\leqq \int R(c(x), y)\ dF(y \mid x) = R(x). \end{aligned}$$

iii) is proved by an example which will be given in the later part of this paragraph.

iv) It will be sufficient to compare with an arbitrary stochastic prediction. For a deterministic prediction $y = m(x)$, we have

$$R_0(x) = \inf_c R(c, m(x)) = R(c_0(x), y).$$

On the other hand, for a stochastic prediction $F(y \mid x_1)$, we get

$$R_s(x_1) = \inf_c \int R(c, y)\ dF(y \mid x_1)$$
$$= \int R(c(x_1), y)\ dF(y \mid x_1),$$

whereas

$$R_0(x) \equiv R(c_0(x), y) \leqq R(c(x_1),\ y) \equiv R(c(x_1),\ m(x))$$

for every value of x and x_1. Taking the mean of both sides of the last inequality with respect to x and x_1, we have $R_0 \leqq R_s$.

In the last part of the theorem, it may be natural to assume that $x_1 \subset x$. Then, from $R_0(x) \leqq R(c(x_1), y)$ we get $\bar{R}_0(x_1) \leqq R_s(x_1)$, where $\bar{R}_0(x_1)$ is the mean value of $R_0(x)$ with respect to the x such that $x \supset x_1$; in this meaning, the deterministic prediction is individually more effective than any other predictions.

Theorem 22. *Let $x_1 \subset x_2$ and suppose that two stochastic predictions $F(y \mid x_1)$ and $F(y \mid x_2)$ are convergent to the corresponding conditional distribution functions with probability* 1 *at each point of y when the size of sample series tends to infinity. Then the precision of $F(y \mid x_2)$ is not smaller than that of $F(y \mid x_1)$ asymptotically and the asymptotic efficency of $F(y \mid x_2)$ is not smaller than that of $F(y \mid x_1)$.*

Proof. The first part of this theorem is obvious. The second part is proved as follows. If we denote the conditional distribution function of x_2 given x_1 by $G(x_2 \mid x_1)$, then the following relation holds asymptotically:

$$F(y \mid x_1) = \int F(y \mid x_2)\ dG(x_2 \mid x_1).$$

Therefore, if we denote the minimum risks for the predictions $F(y \mid x_1)$ and $F(y \mid x_2)$ by $R_1(x_1)$ and $R_2(x_2)$ respectively and if we put, in (6.3.1),

$$\varphi(c, x_2) \equiv \int R(c, y)\ dF(y \mid x_2)$$
$$\Phi(x_2) \equiv G(x_2 \mid x_1) \quad \text{(for a fixed } x_1),$$

then we get

$$R_2(x_1) = \int R_2(x_2)\ dG(x_2 \mid x_1)$$
$$= \int \{\inf_c \varphi(c, x_2)\}\ dG(x_2 \mid x_1)$$
$$\leqq \inf_c \{\int \varphi(c, x_2)\ dG(x_2 \mid x_1)\}$$
$$= \inf_c \{\int\int R(c, y)\ dF(y \mid x_2)\ dG(x_2 \mid x_1)\}$$
$$= \inf_c \{\int R(c, y)\ dF(y \mid x_1)\} = R_1(x_1) \quad \text{(asymptotically)}.$$

Consequently, taking the mean with respect to x_1, we get $R_2 \leqq R_1$ which was

to be proved.

According to the above theorems, we see that, so far as based on the same predictors (the set of condition variates) x, the stochastic prediction has the largest efficiency and that, among the stochastic predictions for the same predictand, the prediction with broader set of condition variates has a larger precision and a larger efficiency at least for a large sample under some weak conditions.

It should be noticed that, for a time series with small size, even if $x_1 \subset x_2$ the precision of $F(y|x_2)$ is not always larger than that of $F(y|x_1)$, as we have mentioned in § 6.1. (Ogawara [15], [20])

Example. Suppose that the weather (y) is classified into two categories no rain (y_1) and rain (y_2) and that we have got the following stochastic prediction scheme.

Preceding condition x	Probability of the x	Stochastic prediction y_1	y_2	Decisional prediction
x_1	p_1	p_{11}	p_{12}	y_2 (Rain)
x_2	p_2	p_{21}	p_{22}	y_2 (Rain)
x_3	p_3	p_{31}	p_{32}	y_1 (No Rain)
Absolute probability		q_1	q_2	

In this table $\sum p_i=1$, $p_{i1}+p_{i2}=1$ $(i=1, 2, 3)$, $q_j=\sum_{i=1}^{3} p_i p_{ij}$ $(j=1, 2)$ and we may assume that $p_{32}<q_2<p_{22}<p_{12}$, $p_{11}<p_{12}$ and $p_{31}>p_{32}$; the other cases will be similar to this case.

Next, we assume that

$$L(c, y_1)=0 \qquad \text{for all } c$$

$$L(c, y_2)=\begin{cases} L-cr & \text{for } 0\leqq c\leqq rL \\ 0 & \text{for } rL\leqq c, \end{cases}$$

where L and r are positive constants.

Then, according to the definitions, we get the following table after some calculations.

Case	Stochastic prediction e_s	c	Decisional prediction e^*	c^*
$1 \leqq r$	0	0	$p_1(p_{12}-r)/rq_2 + p_2(p_{22}-r)/rq_2$	$(1-p_3)rL$
$p_{12} \leqq r<1$	0	0		
$p_{22} \leqq r<p_{12}$	$p_1(p_{12}-r)/q_2$	p_1rL	$p_1(p_{12}-r)/q_2 + p_2(p_{22}-r)/q_2$	$(1-p_3)L$
$q_2 \leqq r<p_{22}$	$p_1(p_{12}-r)/q_2 + p_2(p_{22}-r)/q_2$	p_1rL+p_2rL		
$p_{32} \leqq r<q_2$	$p_3(r-p_{32})/r$	p_1rL+p_2rL	$p_3(r-p_{32})/r$	$(1-p_3)L$
$0 < r\leqq p_{32}$	0	rL		

e=efficiency, c=mean effective protection cost.

We see, in this table, that in the case where r is too large, that is, too much protection cost is necessary in order to lighten the damage to be caused by a rain, it is preferable to do nothing; on the contrary, in the case where r is very small relatively to the probability of rain and we can diminish the damage with a small protection cost, it is desirable to provide a protection constantly, and in both cases the efficiency of our stochastic prediction is zero. We see also that, in the case of $1 \leqq r$ or $r < p_{32}$, the efficiency of the decisional prediction is negative.

By the way, if we put $p_1 = q_2$, $p_2 = 0$, $p_3 = q_1$, $p_{12} = p_{31} = 1$ in the above tables, we then get the results for the case of a deterministic prediction.*

In conclusion, the stochastic prédiction informs us of an effective protection cost which minimizes the expected risk, and it gives a guide to human behaviour and social activities on each occasion of prediction, and when it is repeatedly used the effect will be displayed in the mean. Although, as mentioned by Thompson [38], there may be several practical difficulties for its general public use in the weather forecasting, they will and should be overcome in the near future.

In order to apply our theory to practice, however, it is the first and the most important problem to estimate the form of loss function by means of a damage survey or some experiments. Of course, the loss function depends upon the protection techniques. If the function is concretely given, the comparison of efficiencies will be done in detail further, as was shown in the above example.

§ 6.4. Amount of information of a stochastic prediction.**

Let us use the notations $F(y \mid x)$, $F(y)$ and $G(x)$ in the same meaning as in § 6.1.

Definition. *For an r dimensional predictand* $y = (y_1, y_2, \cdots, y_r)$,

$$H_s(x) = -\int (\log \delta_y F(y \mid x)) dF(y \mid x) \tag{6.4.1}$$

is called the entropy of an individual stochastic prediction $F(y \mid x)$, *where* $\delta_y F(y \mid x)$ *denotes the r th order finite difference of* $F(y \mid x)$ *corresponding to the increment* $\delta_y = (\delta y_1, \delta y_2, \cdots, \delta y_r)$ *of the y.* (6.4.1) *depends on the* δ_y, *butas we are interested in the difference between two such quantities, it will be out of the question.*

Definition. *Let*

$$H_a = -\int (\log \delta_y F(y))\ dF(y), \tag{6.4.2}$$

then

$$I_s(x) = H_a - H_s(x) \tag{6.4.3}$$

* Ogawara [12]; Some other examples are seen in Ogawara [10], [12], [21].

** Ogawara [17]

is called the amount of information brought by the stochastic prediction $F(y|x)$, *and*

$$I_s = \int I_s(x)\ dG(x) \tag{6.4.4}$$

is said to be the mean amount of information of the stochastic prediction.

The $I_s(x)$ and the I_s are related to the precision of the stochastic prediction. If, for instance, $F(y|x)$ and $F(y)$ are normal distribution, we have

$$I_s(x) = \log(\sigma/\sigma(x)), \tag{6.4.5}$$

where σ^2 is the variance of $F(y)$ and $\sigma^2(x)$ is the conditional variance of $F(y|x)$. In general, so far as the individual precision is positive, $I_s(x) > 0$; otherwise the prediction is worthless as a rule. However, we sometimes need the prediction with negative amount of information. For instance, the absolute probability p of great earthquake for 24 hours at a city is very small, say $p = 0.0001$. If then the probability increased to $p(x) = 0.1$ by some premonitory symptoms x, a warning may be necessary, according to the loss function. In such case, however, the amount of information of the stochastic prediction is negative. On the other hand, we have

Theorem 23. *If* $F(y) = \int F(y|x)\ dG(x)$, *the mean amount of information of stochastic prediction is always non-negative.*

Proof. Since $z \log z$ is a convex function of z, we have

$$\begin{aligned}
&\int (\log \delta_y\ F(y))\ dF(y) \\
&= \int \Big[\log \delta_y \{\int F(y|x) dG(x)\}\Big] d\{\int F(y|x) dG(x)\} \\
&\leqq \int \Big[\int \log \delta_y F(y|x)\ dF(y|x)\Big]\ dG(x).
\end{aligned}$$

Consequently, $I_s \geqq 0$.

TOKYO WOMAN'S CHRISTIAN COLLEGE

References (*continued*)

[37] J. C. THOMPSON: *On the operational deficiencies in categorical weather forecasts*, Bull. Amer. Met. Soc., **33** (1952), 223–226.

[38] J. C. THOMPSON and G. W. BRIER: *The economic utility of weather forecasts*, Month. Weath. Rev., **83** (1955), 249–254.