# QUASI-LINEARLY INVARIANT PREDICTION

By J. Tiago de Oliveira

*Faculty of Sciences of Lisbon, Center of Applied Mathematics*

**1. Introduction, the geometry of the problem.** Consider a sequence of random variables $X_1, \cdots, X_n, X_{n+1}, \cdots, X_{m+n}$ from which we have observed the first $n$ and we want to predict the value of some known function of the next $m$ random variables.

Let us suppose that the distribution of the random sequence is known except for parameters of location and dispersion $(\lambda, \delta)$ $(-\infty < \lambda < +\infty, 0 < \delta < +\infty)$; this is a restricted but important case, which, in some applications, may be very useful. We will also suppose that the known function $\varphi(X_{n+1}, \cdots, X_{n+m})$ of the next $m$ random variables about which we want to predict is *quasi-linearly invariant*, that is,

$$\varphi(\lambda + \delta X_{n+1}, \cdots, \lambda + \delta X_{n+m}) = \lambda + \delta\varphi(X_{n+1}, \cdots, X_{n+m}).$$

Examples of such functions are the order statistics, and linear combinations of the order statistics.

Let us suppose that the process has a density and the function $\varphi$ is continuous. Denote by $(1/\delta^{n+1})\mathfrak{L}((x_1 - \lambda)/\delta, \cdots, (x_n - \lambda)/\delta; (z - \lambda)/\delta)$ the likelihood of the random vector $(X_1, \cdots, X_n; Z)$, where $Z = \varphi(X_{n+1}, \cdots, X_{n+m})$. Our purpose is to obtain mean-square predictors and prediction regions for $Z$, which will be quasi-linearly invariant.

Let $R^{n+1}$ denote $(n + 1)$-dimensional real space. As $Z$ must be quasi-linearly invariant the points $(x_1, \cdots, x_n; z)$ and $(\lambda + \delta x_1, \cdots, \lambda + \delta x_n; \lambda + \delta z)$ are in correspondence by the quasi-linear group of transformations $x \to \lambda + \delta x$ acting in $R^{n+1}$.

The equivalence relation introduced by this group splits $R^{n+1}$ into equivalence classes which are the half-planes of a bundle, whose axis is the line $x_1 = \cdots = x_n = z$. Each of the half-planes can be described by a system of quantities

$$\xi_3 = (x_3 - x_1)/(x_2 - x_1), \cdots, \xi_n = (x_n - x_1)/(x_2 - x_1), \quad \zeta = (z - x_1)/(x_2 - x_1)$$

invariant under the quasi-linear group acting on $R^{n+1}$.

**2. Best predictors.** We now consider the problem of finding a best quasi-linearly invariant predictor, $p(x_1, \cdots, x_n)$, where the loss function is given by $(z - p(x_1, \cdots, x_n))^2$.

If $p(x_1, \cdots, x_n)$ is an invariant predictor, write $h(\xi_3, \cdots, \xi_n) = p(0, 1, \xi_3, \cdots, \xi_n)$. Also, define $\mu(\xi_3, \cdots, \xi_n)$ by

$$\mu(\xi_3, \cdots, \xi_n) = \int_{-\infty}^{+\infty} d\zeta\zeta\mathfrak{L}^*(\xi_3, \cdots, \xi_n; \zeta)/\int_{-\infty}^{+\infty} d\zeta\mathfrak{L}^*(\xi_3, \cdots, \xi_n; \zeta)$$

where $\mathfrak{L}^*(\xi_3, \cdots, \xi_n; \zeta) = \int_0^{+\infty} dbb^{n+1} \int_{-\infty}^{+\infty} da\mathfrak{L}(a, a + b, a + b\xi_3, \cdots, a + b\xi_n;$

Received 7 March 1966; revised 21 May 1966.

$a + b\zeta$).

For $x_1 < x_2$, the mean square error can be written

$$\delta^2 \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} [\zeta - h(\xi_3, \cdots, \xi_n)]^2 d\xi_3 \cdots d\xi_n d\zeta \mathfrak{L}^*(\xi_3, \cdots, \xi_n; \zeta).$$

Using variational methods or writing $\zeta - h(\xi_3, \cdots, \xi_n)$ as $\zeta - \mu(\xi_3, \cdots, \xi_n) + \mu(\xi_3, \cdots, \xi_n) - h(\xi_3, \cdots, \xi_n)$, we see that the mean square error is minimized (for $x_1 < x_2$) by $h(\xi_3, \cdots, \xi_n) = \mu(\xi_3, \cdots, \xi_n)$. Thus the least squares predictor (for $x_1 < x_2$) is given by

$$p^*(x_1, \cdots, x_n) = x_1 + (x_2 - x_1)\mu((x_3 - x_1)/(x_2 - x_1), \cdots, (x_n - x_1)/(x_2 - x_1)).$$

The same result is obtained for $x_1 > x_2$ so that the minimum mean square error predictor is $p^*(x_1, \cdots, x_n)$. Another form for $p^*(x_1, \cdots, x_n)$ is

$$p^*(x_1, \cdots, x_n) = [\int_0^{+\infty} d\beta\beta^{n+1} \int_{-\infty}^{+\infty} d\alpha \int_{-\infty}^{+\infty} dz z\mathfrak{L}(\alpha + \beta x_1, \cdots, \alpha + \beta x_n; \alpha + \beta z)] \cdot [\int_0^{+\infty} d\beta\beta^{n+1} \int_{-\infty}^{+\infty} d\alpha \int_{-\infty}^{+\infty} dz\mathfrak{L}(\alpha + \beta x_1, \cdots, \alpha + \beta x_n; \alpha + \beta z)]^{-1}$$

Note that this predictor has a constant mean-square error although $\delta$ is in evidence when we pass to standard ($\delta = 1$) units.

We remark that the methods developed here are similar to the ones used by Pitman (1939) for the search of estimators and tests with location and dispersion parameters.

In the case where the likelihood function can be written $\mathfrak{L}(x_1, \cdots, x_n; z) = \bar{\mathfrak{L}}(x_1, \cdots, x_n)g(z)$, the formula for $p^*(x_1, \cdots, x_n)$ becomes

$$p^*(x_1, \cdots, x_n) = [\int_0^{+\infty} d\beta\beta^{n-1} \int_{-\infty}^{+\infty} d\alpha\bar{\mathfrak{L}}(\alpha + \beta x_1, \cdots, \alpha + \beta x_n)(\mu - \alpha)] \cdot [\int_0^{+\infty} d\beta\beta^{n} \int_{-\infty}^{+\infty} d\alpha\bar{\mathfrak{L}}(\alpha + \beta x_1, \cdots, \alpha + \beta x_n)]^{-1}$$

where $\mu = \int_{-\infty}^{+\infty} zg(z)\, dz$.

**3. Prediction regions.** We now want to find a quasi-linearly invariant prediction region for $Z$. Recall that a prediction region is determined by a function $\varphi(x_1, \cdots, x_n; z)$ taking the values 0 or 1 (randomization is not necessary for the problem under consideration). For fixed $x_1, \cdots, x_n$, if $\varphi(x_1, \cdots x_n; z) = 1$, then $z$ is in the prediction region determined by $\varphi$, otherwise, $z$ is not in the prediction region. Attention is restricted to quasi-linearly invariant functions $\varphi$, that is, functions $\varphi$ which satisfy $\varphi(a + bx_1, \cdots, a + bx_n; a + bz) = \varphi(x_1, \cdots, x_n; z)$.

Let $\omega$ denote the desired prediction level. We set ourselves the task of determining an invariant function $\varphi$ whose prediction level is $\omega$ such that the following quantity is minimized:

$$W(\varphi) = \delta^{-1} \int_{-\infty}^{+\infty} [\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \delta^{-n}\bar{\mathfrak{L}}((x_1 - \lambda)/\delta, \cdots, (x_n - \lambda)/\delta) \cdot\varphi(x_1, \cdots, x_n; z)\, dx_1 \cdots dx_n]\, dz$$

where $\bar{\mathfrak{L}}(x_1, \cdots, x_n) = \int_{-\infty}^{+\infty} dz\mathfrak{L}(x_1, \cdots, x_n; z)$.

The function $W(\varphi)$ is easily interpreted as the average linear measure (in standard units) of the prediction region determined by $\varphi$. Thus we want to minimize $W(\varphi)$ over invariant functions $\varphi$ subject to the condition that

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \delta^{-(n+1)} \mathcal{L}((x_1 - \lambda)/\delta, \cdots, (x_n - \lambda)/\delta; (z - \lambda)/\delta) \cdot \varphi(x_1, \cdots, x_n; z)\, dx_1 \cdots dx_n\, dz = \omega.$$

The substitutions used previously yield

$$W(\varphi) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} d\xi_3 \cdots d\xi_n\, d\zeta \varphi(0, 1, \xi_3, \cdots, \xi_n; \zeta) \cdot \left( \int_0^{+\infty} dbb^{n-1} \int_{-\infty}^{+\infty} da \bar{\mathcal{L}}(a, a + b, a + b\xi_3, \cdots, a + b\zeta_n) \right)$$

and

$$\omega = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} d\xi_3 \cdots d\xi_n\, d\zeta \varphi(0, 1, \xi_3, \cdots, \xi_n; \zeta) \cdot \left( \int_0^{+\infty} dbb^{n-1} \int_{-\infty}^{+\infty} da \mathcal{L}(a, a + b, a + b\xi_3, \cdots, a + b\xi_n; a + b\zeta) \right).$$

Using the Neyman-Pearson lemma, the optimum $\varphi$, say $\varphi^*$, takes the value 1 on the set of $(\xi_3, \cdots, \xi_n; \zeta)$ where

$$\int_0^{+\infty} dbb^{n+1} \int_{-\infty}^{+\infty} da \mathcal{L}(a, a + b, a + b\xi_3, \cdots, a + b\xi_n; a + b\zeta) \geqq k \int_0^{+\infty} dbb^{n+1} \int_{-\infty}^{+\infty} da \bar{\mathcal{L}}(a, a + b, a + b\xi_3, \cdots, a + b\xi_n)$$

where $k$ is computed so that the prediction level is $\omega$. In terms of the original variables, $\varphi^*(x_1, \cdots, x_n, z) = 1$ on the set where

$$\int_0^{+\infty} d\beta\beta^{n-1} \int_{-\infty}^{+\infty} d\alpha \mathcal{L}(\alpha + \beta x_1, \cdots, \alpha + \beta x_n; \alpha + \beta z) \geqq k \int_0^{+\infty} d\beta\beta^{n-1} \int_{-\infty}^{+\infty} d\alpha \bar{\mathcal{L}}(\alpha + \beta x_1, \cdots, \alpha + \beta x_n)$$

Let us now compute some examples: The one-step predictors in the normal, exponential and uniform independent random sequences are, respectively

$p^*(x_1, \cdots, x_n) = \bar{x}$ ($\bar{x}$ having the usual meaning)

$p^*(x_1, \cdots, x_n) = l + [(n - 1)/n](l - \bar{x})$ ($l$ being the observed minimum)

$p^*(x_1, \cdots, x_n) = (l + u)/2$ ($u$ being the observed maximum).

The prediction region for the normal independent sequences is, as obtained before by Hickman (1965) ($s$ being the sample standard deviation), $|(z - \bar{x})/s| \leqq c$.

In the case of existence of sufficient statistics for $\lambda$ and $\delta$ ($\hat{\lambda}$ and $\hat{\delta}$) the mean-square invariant predictor is of the form $\hat{\lambda} + \tau\hat{\delta}$ ($\tau$ a constant) and the best prediction regions are easily expressed in terms of $\hat{\lambda}$ and $\hat{\delta}$.

As a final remark we note that those techniques can also be used for the case of only translation invariance (only a location parameter) or homothetic invariance (only a dispersion parameter)

**4. Acknowledgment.** The author wishes to thank the referee in particular for his improvement of the English text.

## REFERENCES

[1] Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.
[2] Hickman, James C. (1963). Preliminary regional forecasts of the outcome of an estimation problem. *J. Amer. Statist. Assoc.* **58** 1104–1112.
[3] Pitman, E. J. G. (1939). The estimation of location and scale parameters of a continuous population of any given form. *Biometrika* **31** 391.
[4] Rosenblatt, Murray (1962). *Random Processes.* Oxford Univ. Press.