

Optimal prediction theory *

Jean-Marie Dufour [†]
McGill University

First version: May 1999

Revised: February 2021

This version: February 15, 2021

Compiled: February 15, 2021, 17:40

*This work was supported by the William Dow Chair in Political Economy (McGill University), the Bank of Canada (Research Fellowship), the Toulouse School of Economics (Pierre-de-Fermat Chair of excellence), the Universidad Carlos III de Madrid (Banco Santander de Madrid Chair of excellence), a Guggenheim Fellowship, a Konrad-Adenauer Fellowship (Alexander-von-Humboldt Foundation, Germany), the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Fonds de recherche sur la société et la culture (Québec).²

[†] William Dow Professor of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ). Mailing address: Department of Economics, McGill University, Leacock Building, Room 414, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada. TEL: (1) 514 398 6071; FAX: (1) 514 398 4800; e-mail: jean-marie.dufour@mcgill.ca. Web page: <http://www.jeanmariedufour.com>

Contents

List of Definitions, Assumptions, Propositions and Theorems	i
1. Optimal mean square prediction	1
2. Properties of conditional expectations	2
3. Linear regression	4
4. Properties of the projection operator	6
5. Prediction based on an infinite number of variables	8
6. Bibliographic notes	9

List of Definitions, Assumptions, Propositions and Theorems

2.1 Proposition : Linearity	2
2.2 Proposition : Positivity	2
2.3 Proposition : Monotonicity	2
2.4 Proposition : Invariance	2
2.5 Proposition : Orthogonality	2
2.6 Proposition : Iterated conditionings law	3
2.7 Proposition : Mean square optimality	3
2.8 Proposition : Characterization of optimality by orthogonality	3
2.9 Definition : Conditional covariance	3
2.10 Proposition : Variance decomposition	3
4.2 Proposition : Linearity	7
4.3 Proposition : Invariance	7
4.4 Proposition : Orthogonality	7
4.5 Proposition : Law of iterated projections	7
4.6 Proposition : Frisch-Waugh Theorem	8

1. Optimal mean square prediction

Let Y, X_1, \dots, X_k be real random variables in L^2 , and $X = (X_1, \dots, X_k)'$. We wish to find a function

$$g(X) = g(X_1, \dots, X_k)$$

such that

$$\mathbb{E}([Y - g(X)]^2) \text{ is minimal.}$$

Given the mean square criterion, we also restrict $g(X)$ to be in L^2 :

$$\mathbb{E}[g(X)^2] < \infty.$$

Then it is easy to see that the optimal solution to this problem is

$$g(X) = M(X)$$

where

$$M(X) = \mathbb{E}(Y | X) .$$

In general, $M(X)$ is a nonlinear function of X . The optimality of $M(X)$ can easily be shown on observing that :

$$\begin{aligned} \mathbb{E}\{[Y - g(X)]^2\} &= \mathbb{E}\{[Y - \mathbb{E}(Y | X) + \mathbb{E}(Y | X) - g(X)]^2\} \\ &= \mathbb{E}\{[Y - \mathbb{E}(Y | X)]^2 + [\mathbb{E}(Y | X) - g(X)]^2 \\ &\quad + 2[Y - \mathbb{E}(Y | X)][\mathbb{E}(Y | X) - g(X)]\} \\ &= \mathbb{E}\{[Y - \mathbb{E}(Y | X)]^2\} + \mathbb{E}\{[\mathbb{E}(Y | X) - g(X)]^2\} \\ &\quad + 2\mathbb{E}\{[\mathbb{E}(Y | X) - g(X)] \mathbb{E}[Y - \mathbb{E}(Y | X) | X]\} \\ &= \mathbb{E}\{[Y - \mathbb{E}(Y | X)]^2\} + \mathbb{E}\{[\mathbb{E}(Y | X) - g(X)]^2\} \end{aligned}$$

from which it follows that the optimal solution is

$$g(X) = \mathbb{E}(Y | X) .$$

The set of random variables

$$M_0 = \{Z : Z = g(X) \text{ is a random variable and } \mathbb{E}(Z^2) < \infty\}$$

is a closed subspace of L^2 . $M(X) = \mathbb{E}(Y | X)$ can be interpreted as the projection of Y on M_0 :

$$\mathbb{E}(Y | X) = P_{M_0} Y.$$

2. Properties of conditional expectations

Let

$$\begin{aligned} Y &= (Y_1, \dots, Y_q)' , \\ Z &= (Z_1, \dots, Z_q)' , \\ X &= (X_1, \dots, X_k) \end{aligned}$$

be random vectors whose components are all in L^2 . By definition,

$$E(Y | X) = \begin{bmatrix} E(Y_1 | X) \\ E(Y_2 | X) \\ \vdots \\ E(Y_q | X) \end{bmatrix}$$

and similarly for $E(Z | X)$.

Let $L^2(X)$ be the set of random variables W such that $W = g(X)$ and $E(W^2) < \infty$.

2.1 Proposition LINEARITY. *Let A an $m \times q$ fixed matrix and b an $m \times 1$ fixed vector. Then*

$$\begin{aligned} E(AY + b | X) &= AE(Y | X) + b, \\ E(Y + Z | X) &= E(Y | X) + E(Z | X). \end{aligned}$$

2.2 Proposition POSITIVITY. *If $Y_i \geq 0$, for $i = 1, \dots, q$, then*

$$E(Y_i | X) \geq 0, \quad \text{for } i = 1, \dots, q.$$

2.3 Proposition MONOTONICITY. *If $Y_i \geq Z_i$, for $i = 1, \dots, q$, then*

$$E(Y_i | X) \geq E(Z_i | X), \quad \text{for } i = 1, \dots, q.$$

2.4 Proposition INVARIANCE.

$$\begin{aligned} E(Y | X) = Y &\Leftrightarrow Y \text{ is a function of } X \\ &\Leftrightarrow \text{there is a function } g(x) \text{ such that } Y = g(X) \\ &\quad \text{with probability 1.} \end{aligned}$$

2.5 Proposition ORTHOGONALITY. *If $g_1(X) \in L^2$ and $g_2(Y) \in L^2$, then*

$$E\{g_1(X) [g_2(Y) - E(g_2(Y) | X)]\} = 0.$$

2.6 Proposition ITERATED CONDITIONINGS LAW. *If W is a random vector such that*

$$L^2(W) \subseteq L^2(X) ,$$

then

$$\begin{aligned} \mathbb{E}[\mathbb{E}(Y | X) | W] &= \mathbb{E}[\mathbb{E}(Y | W) | X] \\ &= \mathbb{E}(Y | W) . \end{aligned}$$

2.7 Proposition MEAN SQUARE OPTIMALITY.

$$\mathbb{E}[(Y_i - \mathbb{E}(Y_i | X))^2] = \min_{g_i(X) \in L^2(X)} \mathbb{E}[(Y_i - g_i(X))^2] , \quad i = 1, \dots, q .$$

2.8 Proposition CHARACTERIZATION OF OPTIMALITY BY ORTHOGONALITY. *For any $i = 1, \dots, q$,*

$$h_i(X) = \mathbb{E}(Y_i | X) \Leftrightarrow \mathbb{E}[g(X)(Y_i - h_i(X))] = 0 , \quad \forall g(X) \in L^2(X) .$$

2.9 Definition CONDITIONAL COVARIANCE. *The conditional covariance matrix of Y given X is the matrix*

$$\mathbf{V}(Y | X) = \mathbb{E}[(Y - \mathbb{E}(Y | X))(Y - \mathbb{E}(Y | X))' | X] .$$

If we define

$$\varepsilon(X) = Y - \mathbb{E}(Y | X) ,$$

we see easily that

$$\mathbf{V}[\varepsilon(X)] = \mathbb{E}[\mathbf{V}(Y | X)] .$$

We can then write

$$Y = \mathbb{E}(Y | X) + \varepsilon(X)$$

where $\mathbb{E}(Y | X)$ and $\varepsilon(X)$ are uncorrelated.

2.10 Proposition VARIANCE DECOMPOSITION.

$$\begin{aligned} \mathbf{V}(Y) &= \mathbf{V}[\mathbb{E}(Y | X)] + \mathbf{V}[\varepsilon(X)] \\ &= \mathbf{V}[\mathbb{E}(Y | X)] + \mathbb{E}[\mathbf{V}(Y | X)] . \end{aligned}$$

3. Linear regression

Consider again the setup of Section 1. We now study the problem of finding a function of the form

$$\begin{aligned} L(X) &= b_0 + b_1 X_1 + \cdots + b_k X_k \\ &= \sum_{i=0}^k b_i X_i = b'x \end{aligned}$$

where

$$X_0 = 1, \quad b = (b_0, b_1, \dots, b_k)' \quad (3.1)$$

$$x = (X_0, X_1, \dots, X_k)', \quad (3.2)$$

such that the mean square prediction error

$$\mathbb{E} \{ [Y - L(X)]^2 \} = \mathbb{E} [(Y - b'x)^2]$$

is minimal. In other words, we wish to minimize (with respect to b) the function

$$\begin{aligned} S(b) &= \mathbb{E} \{ [Y - b'x]^2 \} \\ &= \mathbb{E} (Y^2) - 2b' \mathbb{E} (xY) + b' \mathbb{E} (xx') b. \end{aligned}$$

It is easy to see that the optimal value of b must satisfy the equation

$$\mathbb{E} [x(Y - b'x)] = 0$$

or

$$\mathbb{E} (xx') b = \mathbb{E} (xY).$$

If we write

$$b = \begin{pmatrix} \beta_0 \\ \gamma \end{pmatrix}, \quad \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_k \end{pmatrix}, \quad X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix},$$

we see that

$$\begin{bmatrix} 1 & \mathbb{E}(X)' \\ \mathbb{E}(X) & \mathbb{E}(XX') \end{bmatrix} \begin{bmatrix} \beta_0 \\ \gamma \end{bmatrix} = \begin{bmatrix} \mathbb{E}(Y) \\ \mathbb{E}(XY) \end{bmatrix},$$

hence

$$\beta_0 + E(X)' \gamma = E(Y) \quad (3.3)$$

$$E(Y) \beta_0 + E(XX') \gamma = E(XY) \quad (3.4)$$

and

$$\beta_0 = E(Y) - E(X)' \gamma .$$

Further, by the basic properties of the expectation operator,

$$E(XX') = V(X) + E(X) E(X)' ,$$

$$E(XY) = C(X, Y) + E(X) E(Y)$$

where

$$V(X) = E \{ E[X - E(X)] [X - E(X)]' \} , \quad (3.5)$$

$$C(X, Y) = E \{ [X - E(X)] [Y - E(Y)]' \} . \quad (3.6)$$

By the equations (3.3)-(3.6), we then see easily that

$$E(X) \beta_0 + E(X) E(X)' \gamma = E(X) E(Y) ,$$

$$E(X) \beta_0 + V(X) \gamma + E(X) E(X)' \gamma = C(X, Y) + E(X) E(Y)$$

hence

$$V(X) \gamma = C(X, Y) .$$

Thus,

$$\beta_0 = E(Y) - E(X)' \gamma , \quad (3.7)$$

$$V(X) \gamma = C(X, Y) . \quad (3.8)$$

The function

$$L(X) = \beta_0 + X' \gamma$$

is called the

linear regression of X on Y

or the

affine projection of Y on X. (3.9)

We write

$$L(X) = P_L(Y | X) = \beta_0 + X' \gamma$$

where β_0 and γ are any solution of the normal equations:

$$\begin{aligned} V(X) \gamma &= C(X, Y) , \\ \beta_0 &= E(Y) - E(X)' \gamma . \end{aligned}$$

If we denote by

$$\varepsilon = Y - P_L(Y | X)$$

the prediction error, we see easily that:

$$\begin{aligned} E(\varepsilon) &= 0 , \\ C(X, \varepsilon) &= 0 . \end{aligned}$$

In the language of Hilbert space theory, we can also write

$$L(X) = P_M Y = P_L(Y | X)$$

where

$$M = \overline{\text{sp}} \{1, X\} = \overline{\text{sp}} \{1, X_1, \dots, X_k\} .$$

If

$$\det [V(X)] \neq 0 ,$$

the optimal coefficients β_0 and γ are uniquely defined :

$$\gamma = V(X)^{-1} C(X, Y) , \quad \beta_0 = E(Y) - E(X)' \gamma .$$

4. Properties of the projection operator

Let

$$\begin{aligned} Y &= (Y_1, \dots, Y_q)' , \\ Z &= (Z_1, \dots, Z_q)' , \\ X &= (X_1, \dots, X_k) \end{aligned}$$

be random vectors whose components are all in L^2 . By definition,

$$P_L(Y | X) = \begin{bmatrix} P_L(Y_1 | X) \\ P_L(Y_2 | X) \\ \vdots \\ P_L(Y_q | X) \end{bmatrix}$$

We call $\mathcal{L}(X)$ the set of all linear transformations of X .

4.1 Proposition If $\det[V(X)] \neq 0$,

$$\begin{aligned} P_L(Y | X) &= E(Y) + C(Y, X)V(X)^{-1}(X - E(X)) \\ &= [E(Y) - C(Y, X)V(X)^{-1}E(X)] + C(Y, X)V(X)^{-1}X, \end{aligned} \quad (4.1)$$

and

$$\begin{aligned} \varepsilon_L(X) &: = Y - P_L(Y | X) \\ &= [Y - E(Y)] - C(Y, X)V(X)^{-1}[X - E(X)]. \end{aligned} \quad (4.2)$$

4.2 Proposition LINEARITY. Let A and B be two fixed matrices of dimensions $n \times q$ and $1 \times n$ respectively. Then

$$P_L(AY | X) = A P_L(Y | X), \quad (4.3)$$

$$P_L(YB | X) = P_L(Y | X) B, \quad (4.4)$$

$$P_L(Y + Z | X) = P_L(Y | X) + P_L(Z | X). \quad (4.5)$$

4.3 Proposition INVARIANCE.

$$\begin{aligned} P_L(Y | X) = Y &\Leftrightarrow Y \text{ is a linear function of } X \\ &\Leftrightarrow Y = AX + b \text{ with probability } 1 \end{aligned}$$

where A and b are fixed matrices.

Note that

4.4 Proposition ORTHOGONALITY. If $\varepsilon_L(X) = Y - P_L(Y | X)$,

$$C(\varepsilon_L(X), X) = 0. \quad (4.6)$$

4.5 Proposition LAW OF ITERATED PROJECTIONS. If W is a random vector such that

$$\mathcal{L}(W) \subseteq \mathcal{L}(X),$$

then

$$\begin{aligned} P_L[P_L(Y | X) | W] &= P_L[P_L(Y | W) | X] \\ &= P_L(Y | W). \end{aligned}$$

In particular, if $X = W$,

$$P_L [P_L (Y | X) | X] = P_L (Y | X) \quad (4.7)$$

4.6 Proposition FRISCH-WAUGH THEOREM.

$$\begin{aligned} P_L (Y | X, W) &= P_L (Y | X) + P_L (Y - P_L (Y | X) | W - P_L (W | X)) \\ &= P_L (Y | X) + P_L (Y | W - P_L (W | X)) . \end{aligned} \quad (4.8)$$

5. Prediction based on an infinite number of variables

It is possible to generalize the concept of projection to the case where X contains an infinite number of variables

$$X \equiv \overline{X}_{t-1} = (X_{t-1}, X_{t-2}, \dots) = (X_{t-k} : k \geq 1) . \quad (5.1)$$

Let Y a scalar random variable. If we consider a potentially infinite set I of random variables such that the variables in I have finite second order moments, we can define the set $\mathcal{L}^2(I)$ of linear transformations of a finite set of variables from I . Then we can define $\mathcal{H}(I)$ the smallest set of random variables in L^2 such that $\mathcal{H}(I)$ is closed, i.e. $\mathcal{H}(I)$ satisfies the following condition: if

$$\{Y_n : n \in \mathbb{Z}\} \subseteq \mathcal{H}(I) \quad (5.2)$$

then

$$E [(Y_m - Y_n)^2] \longrightarrow 0 \text{ when } m, n \longrightarrow \infty \quad (5.3)$$

entails

$$\text{there exists } Y \in \mathcal{H}(I) \text{ such that } E [(Y_n - Y)^2] \xrightarrow{n \rightarrow \infty} 0 . \quad (5.4)$$

We call $\mathcal{H}(I)$ the “Hilbert space” generated by I .

5.1 Theorem *There exists a unique random variable $\widehat{Y}_{|t-1} \equiv P_L(Y | I)$ such that*

$$E[(Y - \widehat{Y}_{|t-1})^2] = \inf_{Z \in \mathcal{H}(I)} E[(Y - Z)^2] . \quad (5.5)$$

The operator $P_L(Y | I)$ enjoys properties sated in Propositions 4.2 to 4.6.

6. Bibliographic notes

On the properties of conditional expectations, see Gouriéroux and Monfort (1995, Appendix B) and Williams (1991).

References

- GOURIÉROUX, C., AND A. MONFORT (1995): *Statistics and Econometric Models, Volumes One and Two*. Cambridge University Press, Cambridge, U.K., Translated by Quang Vuong.
- WILLIAMS, D. (1991): *Probability with Martingales*. Cambridge University Press, Cambridge, U.K.