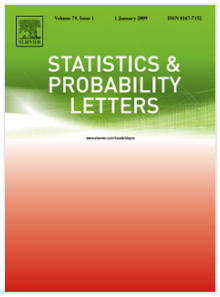
Multivariate Cramér-Rao inequality for prediction and efficient predictors

Emmanuel Onzon

Accepted Manuscript



Multivariate Cramér-Rao inequality for prediction and efficient predictors

Emmanuel Onzon

PII: S0167-7152(10)00347-0
DOI: 10.1016/j.spl.2010.12.007
Reference: STAPRO 5863

To appear in: *Statistics and Probability Letters*

Received date: 4 May 2010
Revised date: 13 December 2010
Accepted date: 15 December 2010

Please cite this article as: Onzon, E., Multivariate Cramér-Rao inequality for prediction and efficient predictors. *Statistics and Probability Letters* (2010), doi:10.1016/j.spl.2010.12.007

Multivariate Cramér-Rao inequality for prediction and efficient predictors

Emmanuel Onzon – UPMC LSTA

Abstract

We derive and discuss a matricial Cramér-Rao type inequality for the quadratic prediction error matrix. A study of the attainment of the bound follows. Then we introduce an unbiased predictor for a bivariate Poisson process and prove that it is efficient, i.e. its quadratic error attains the Cramér-Rao bound.

Keywords: Cramér-Rao inequality, Efficient predictor, Information inequality, Prediction, Fisher information, Parametric models

1. Introduction

We consider statistical *prediction* theory as an extension of statistical *estimation* theory as presented in Bosq and Blanke (2007). To put this extension in perspective we briefly recall the framework of statistical estimation before presenting the framework of statistical prediction.

In statistical estimation theory, the statistician observes a random variable $X : \Omega \to \mathcal{X}$ where $(\Omega, \mathcal{A})$ and $(\mathcal{X}, \mathcal{B})$ are measurable spaces. The distribution of X is unknown but is assumed to belong to some family of probability measures $(P_\theta)_{\theta\in\Theta}$ where $\theta \in \Theta$ is the parameter of the family (we have implicitly $\mathbb{P}_\theta$ a probability measure on $\mathcal{A}$ and $P_\theta = X(\mathbb{P}_\theta)$). In this framework, the problem is to estimate the unknown parameter θ or a function of this parameter, say $g(\theta)$, thanks to the observed variable X. To this end the statistician computes an *estimator* which is a measurable function of X. A classic reference for statistical estimation theory is the book Lehmann and Casella (1998).

Email address: emmanuel.onzon@upmc.fr (Emmanuel Onzon – UPMC LSTA)

In the framework of statistical *prediction* theory, we consider the unobserved random variable $Y : \Omega \to \mathcal{Y}$ in addition to the observed random variable X, with $(\mathcal{Y}, \mathcal{C})$ a measurable space. The probability measure P_θ is not the distribution of X anymore but the probability measure on the underlying probability space Ω, thus $(\Omega, \mathcal{A}, P_\theta, \theta \in \Theta)$ is a statistical model. The problem is to predict $g(X, Y, \theta)$ where $g : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathcal{Z}$ is a known function such that the function $(x, y) \mapsto g(x, y, \theta)$ is measurable for all $\theta \in \Theta$ and $g(X, Y, \theta) \in L^2(P_\theta)$ for all θ. To do this the statistician computes a *predictor* $p(X)$ which is a measurable function of X, such that $p(X) \in L^2(P_\theta)$ for all θ. If g only depends on Y the problem is said to be a *pure prediction* problem, if g only depends on X then this is an approximation problem, and if g only depends on θ then this is an estimation problem. So the framework of statistical prediction includes the framework of statistical estimation.

In this paper we present a few results about statistical predictors where we assume $\Theta \subset \mathbb{R}^d$ and $\mathcal{Z} = \mathbb{R}^k$.

The framework of statistical prediction theory can be used to pose the problem of prediction of a time series in the following way. Let $(Z_t)_{t \geqslant 0}$ be a stochastic process, and consider the problem of predicting Z_{t+h} assuming we know the process at time t (or until time t). If we put this problem in the previous setting, we have $X = Z_t$ (or $X = (Z_s)_{0 \leqslant s \leqslant t}$) and $Y = g(X, Y, \theta) = Z_{t+h}$. Since the conditional expectation $E_\theta[Z_{t+h}|Z_t]$ (or $E_\theta[Z_{t+h}|(Z_s)_{0 \leqslant s \leqslant t}]$) is the best predictor of Z_{t+h} for the quadratic error, it is also of interest to consider the function to be predicted $g(X, Y, \theta) = E_\theta[Z_{t+h}|Z_t]$ (or $g(X, Y, \theta) = E_\theta[Z_{t+h}|(Z_s)_{0 \leqslant s \leqslant t}]$) which is $E_\theta[Y|X]$.

A convenient way to evaluate the accuracy of a predictor is to use the quadratic prediction error (QPE). When $\mathcal{Z} = \mathbb{R}$ the QPE is

$$R_\theta(p, g) = E_\theta\big(p(X) - g(X, Y, \theta)\big)^2$$

This risk function induces the following preference relation between predictors. The predictor p_1 is said to be preferable to the predictor p_2 for predicting $g(X, Y, \theta)$ if $R_\theta(p_1, g) \leqslant R_\theta(p_2, g), \forall \theta \in \Theta$.

Many results of estimation theory regarding the accuracy of an estimator have been generalized to statistical prediction theory for the accuracy of a predictor when using

the QPE in the case $\mathcal{Z} = \mathbb{R}$. Those results are presented in details in Bosq and Blanke (2007). We review a few of them.

The best $\sigma(X)$-measurable quantity to predict $g(X, Y, \theta)$ for the quadratic error is its conditional expectation with respect to X. $E_\theta^X g(X, Y, \theta)$ usually depends on θ, then a common strategy is to compute an estimator $\hat{\theta}$ of θ and take $p(X) = E_{\hat{\theta}}^X g(X, Y, \hat{\theta})$ as a predictor. Such a predictor is called a *plug-in* predictor. The class of plug-in predictors is a useful and important one, nonetheless the results that follow are not limited to them.

The concept of sufficiency is generalized to prediction in the following way. A statistics $S(X)$ is said to be P-sufficient for predicting $g(X, Y, \theta)$ if the conditional distribution of X with respect to $S(X)$ does not depend on θ *and* for all θ, X and $g(X, Y, \theta)$ are conditionaly independent given $S(X)$. A Rao-Blackwell theorem for prediction states that if $S(X)$ is P-sufficient for predicting $g(X, Y, \theta)$ then $E^{S(X)}p(X)$ is preferable to $p(X)$ for predicting $g(X, Y, \theta)$. Unbiasedness is generalized to prediction too. A predictor $p(X)$ of $g(X, Y, \theta)$ is said to be unbiased if

$$E_\theta\big(p(X)\big) = E_\theta\big(g(X, Y, \theta)\big), \quad \theta \in \Theta$$

When the predictor $p(X)$ is not unbiased it is said to be biased and one calls the *bias* of $p(X)$ for predicting $g(X, Y, \theta)$ the quantity $b(\theta) = E_\theta\big(p(X) - g(X, Y, \theta)\big)$.

A Lehmann-Scheffé theorem states that, if the statistics $S(X)$ is complete and P-sufficient for predicting $g(X, Y, \theta)$ and $p(X)$ is unbiased, then $E^S\big(p(X)\big)$ is the unique optimal unbiased predictor of $g(X, Y, \theta)$ (an optimal predictor of $g(X, Y, \theta)$ is a predictor preferable to any predictor for predicting $g(X, Y, \theta)$).

A Cramér-Rao type inequality is obtained in Yatracos (1992). It is presented in Bosq and Blanke (2007) under the following form.

Assumptions 1. $\Theta \subset \mathbb{R}$ is an open set, the model associated with X is dominated by a σ-finite measure μ, the density $f(x, \theta)$ of X is such that $\{x : f(x, \theta) > 0\}$ does not depend on θ, $\partial f(x, \theta)/\partial\theta$ does exist. Finally the Fisher information $I_X(\theta) = E_\theta\Big(\frac{\partial}{\partial\theta} \ln f(X, \theta)\Big)^2$ satisfies $0 < I_X(\theta) < \infty$, $\theta \in \Theta$.

Theorem 1. *If assumptions 1 hold, and $p(X)$ is an unbiased predictor, and the equality*

$$\int p(x) f(x, \theta) d\mu(x) = E_\theta\big(g(X, Y, \theta)\big)$$

can be differentiated under the integral sign, then

$$E_\theta\big(p-g\big)^2 \geqslant E_\theta\big(g-E_\theta^X g\big)^2 + \frac{\Big(\frac{\partial}{\partial\theta}(E_\theta g) - E_\theta\big(E_\theta^X g\,\frac{\partial}{\partial\theta}\ln f(X,\theta)\big)\Big)^2}{I_X(\theta)}$$

where we noted p for $p(X)$ and g for $g(X,Y,\theta)$.

Since the quadratic prediction error can be decomposed the following way

$$E_\theta\big(p-g\big)^2 = E_\theta\big(p-E_\theta^X g\big)^2 + E_\theta\big(E_\theta^X g-g\big)^2$$

the inequality in theorem 1 can be written

$$E_\theta\big(p-E_\theta^X g\big)^2 \geqslant \frac{\Big(\frac{\partial}{\partial\theta}(E_\theta g) - E_\theta\big(E_\theta^X g\frac{\partial}{\partial\theta}\ln f(X,\theta)\big)\Big)^2}{I_X(\theta)}$$

Of course $E_\theta\big(E_\theta^X g(X,Y,\theta)\frac{\partial}{\partial\theta}\ln f(X,\theta)\big)$ can also be written $E_\theta\big(g(X,Y,\theta)\frac{\partial}{\partial\theta}\ln f(X,\theta)\big)$.

The following corollary is corollary 1.1 p.22 in Bosq and Blanke (2007)

Corollary 2. *If, in addition, the equality*

$$E_\theta g(X,Y,\theta) = \int E_\theta^{X=x}\big(g(X,Y,\theta)\big)f(x,\theta)d\mu(x)$$

is differentiable under the integral sign, then

$$E_\theta\Big(p(X)-E_\theta^X g(X,Y,\theta)\Big)^2 \geqslant \frac{\Big[E_\theta\Big(\frac{\partial E_\theta^X g(X,Y,\theta)}{\partial\theta}\Big)\Big]^2}{I_X(\theta)}$$

The Cramér-Rao bound for estimators has been generalized for *biased* estimators. In such a case, the bound not only depends on $g'(\theta)$ but also on the bias $b(\theta)$ and its derivative. It has been generalized for the multivariate case too, i.e. when $\theta\in\mathbb{R}^k$ or $g(\theta)\in\mathbb{R}^d$. Then, depending on the version of the inequality, the left-hand side of the inequality is either the covariance of the norm of the estimator or the covariance matrix of the estimator. In the later case the two matrices are compared with the Löwner semi-order (i.e. $A\leqslant B$ iff $B-A$ is positive semidefinite). We refer to Lehmann and Casella (1998) chapter 2 for the Cramér-Rao inequality for estimators (where it is called the information inequality).

Bosq and Blanke generalized the inequality of theorem 1 for $\Theta\subset\Theta_0$ and g taking its values in B, with Θ_0 and B separable Banach spaces. Under some regularity conditions

about the density function f they obtain the following inequality

$$E_\theta\big(x^*(p-g)\big)^2 \geqslant E_\theta\big(x^*(g-E_\theta^X g)\big)^2 + \frac{x^*\Big(\partial_u(E_\theta g) - E_\theta\big(E_\theta^X g\,\frac{\partial_u f(X,\theta)}{f(X,\theta)}\big)\Big)^2}{I_{X,u}(\theta)}$$

for any $x^* \in B^*$ the topological dual of B and any $u \in \Theta_0$ such that the derivatives exist, and where we noted $\partial_u h(\theta) = \frac{\partial}{\partial t} h(\theta + tu)\big|_{t=0}$, $I_{X,u}(\theta) = E_\theta\Big(\frac{\partial_u f(X,\theta)}{f(X,\theta)}\Big)^2$, $p = p(X)$ and $g = g(X, Y, \theta)$. We refer to Bosq and Blanke (2007) for more details.

Nayak introduced a matricial Cramér-Rao type inequality for prediction in the case of a multidimensional parameter θ and a random vector to predict, that generalizes theorem 1. In his paper, Nayak (2002), he also gives a Bhattacharyya type bound for predictors.

In the second section we give a proof of the matricial inequality given by Nayak and discuss it. Then in the third section we study the attainment of the bound. In the last section we consider an unbiased predictor for the bivariate Poisson process and prove its efficiency.

2. Cramér-Rao type inequality for prediction

We denote $J_\theta h(\theta_0)$ the jacobian matrix of a function h with respect to the variable θ and evaluated at the point θ_0, and $\dot{L}_\theta$ the gradient of $\ln f(x,\theta)$ with respect to θ a vector of $\mathbb{R}^d$, where $f(x,\theta)$ is the density of X.

$$\dot{L}_\theta = \nabla_\theta \ln f(x,\theta)$$

When an inequality involves matrices, then it always refers to the Löwner semiorder (i.e. $A \leqslant B$ iff $B - A$ is positive semidefinite).

Assumptions 2. $\Theta \subset \mathbb{R}^d$ is an open set, the model associated with X is dominated by a σ-finite measure μ, the density $f(x,\theta)$ of X is such that $\{x : f(x,\theta) > 0\}$ does not depend on θ, $\nabla_\theta f(x,\theta)$ does exist. Finally the Fisher information

$$I_X(\theta) = E_\theta\big(\dot{L}_\theta \dot{L}_\theta^T\big)$$

satisfies $\det\big(I_X(\theta)\big) \neq 0$ and $I_{X,i,j}(\theta) < \infty$, $\theta \in \Theta$, where $I_{X,i,j}(\theta)$ is the coefficient at line i and column j of the matrix $I_X(\theta)$.

Assumptions 3. The equality

$$\int E_\theta^{X=x}\big(g(X,Y,\theta)\big) f(x,\theta)\, d\mu(x) = E_\theta\big(g(X,Y,\theta)\big)$$

can be differentiated under the integral sign with respect to each component of θ (where $g : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}^k$ and the function $(x,y) \mapsto g(x,y,\theta)$ is measurable for all $\theta \in \Theta$).

Theorem 3. *Suppose assumptions 2 and 3 hold. Let $p : \mathcal{X} \to \mathbb{R}^k$ be an unbiased predictor of $g(X,Y,\theta)$ and $g : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}^k$ a function such that for all $\theta \in \Theta$ the function $(x,y) \mapsto g(x,y,\theta)$ is measurable and $g(X,Y,\theta) \in L^2(P_\theta)$ and $\Theta \subset \mathbb{R}^d$.*

If the equality

$$\int p(x) f(x,\theta)\, d\mu(x) = E_\theta\big(g(X,Y,\theta)\big) \tag{1}$$

can be differentiated under the integral sign with respect to each component of θ, then

$$E_\theta\big(p - E_\theta^X g\big)\big(p - E_\theta^X g\big)^T \;\geqslant\; G(\theta)\, I_X(\theta)^{-1}\, G(\theta)^T \tag{2}$$

where we denoted $p = p(X)$, $g = g(X,Y,\theta)$ and $G(\theta) = E_\theta\big(J_\theta E_\theta^X g\big)$.

Proof. Let $Z = \begin{pmatrix} p - E_\theta^X g \\ \dot{L}_\theta \end{pmatrix}$. Differentiability of (1) under the integral sign allows to show $J_\theta(E_\theta p) = E_\theta p \dot{L}_\theta^T$, hence $J_\theta(E_\theta g) = E_\theta p \dot{L}_\theta^T$. Using these relations one gets

$$E_\theta\big[ZZ^T\big] = \begin{pmatrix} E_\theta(p - E_\theta^X g)(p - E_\theta^X g)^T & J_\theta\big(E_\theta g\big) - E_\theta\big[\big(E_\theta^X g\big)\dot{L}_\theta^T\big] \\ \Big(J_\theta\big(E_\theta g\big) - E_\theta\big[\big(E_\theta^X g\big)\dot{L}_\theta^T\big]\Big)^T & I_X(\theta) \end{pmatrix} \geqslant 0$$

Using assumption 3 one gets

$$J_\theta\big(E_\theta g\big) - E_\theta\big[\big(E_\theta^X g\big)\dot{L}_\theta^T\big] = E_\theta\big(J_\theta E_\theta^X g\big)$$

which is $G(\theta)$. Hence

$$E_\theta\big[ZZ^T\big] = \begin{pmatrix} E_\theta(p - E_\theta^X g)(p - E_\theta^X g)^T & G(\theta) \\ G(\theta)^T & I_X(\theta) \end{pmatrix} \geqslant 0$$

$g(X,Y,\theta)$ and $\dot{L}_\theta$ are square integrable therefore $E_\theta\big[\big(E_\theta^X g\big)\dot{L}_\theta^T\big]$ does exist. For all $\alpha \in \mathbb{R}^{k+d}$ it holds $\alpha^T E_\theta\big[ZZ^T\big]\alpha \geqslant 0$ therefore, letting $\alpha = \begin{pmatrix} \beta \\ -I_X(\theta)^{-1} G(\theta)^T \beta \end{pmatrix}$ where β is any vector of $\mathbb{R}^k$, one gets

$$\alpha^T E_\theta\big[ZZ^T\big]\alpha = \beta^T\Big(E_\theta(p - E_\theta^X g)(p - E_\theta^X g)^T - bb^T - G(\theta) I_X(\theta)^{-1} G(\theta)^T\Big)\beta \geqslant 0$$

Which implies the matrix inequality. □

This theorem can be extended for the case of *biased* predictors

Theorem 4. *Suppose assumptions 2 and 3 hold. Let* $p : \mathcal{X} \to \mathbb{R}^k$ *be a predictor of* $g(X, Y, \theta)$ *with bias* $b(\theta)$ *a differentiable function and* $g : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}^k$ *a function such that for all* $\theta \in \Theta$ *the function* $(x, y) \mapsto g(x, y, \theta)$ *is measurable and* $g(X, Y, \theta) \in L^2(P_\theta)$ *and* $\Theta \subset \mathbb{R}^d$.

If the equalities

$$\int p(x) f(x, \theta)\, d\mu(x) = E_\theta\big(g(X, Y, \theta)\big) + b(\theta)$$

$$\int f(x, \theta)\, d\mu(x) = 1 \tag{3}$$

can be differentiated under the integral sign with respect to each component of θ*, then*

$$E_\theta\big(p - E_\theta^X g\big)\big(p - E_\theta^X g\big)^T \;\geqslant\; b(\theta)b(\theta)^T + \big(G(\theta) + J_\theta b(\theta)\big)\, I_X(\theta)^{-1}\, \big(G(\theta) + J_\theta b(\theta)\big)^T \tag{4}$$

where we denoted $p = p(X)$, $g = g(X, Y, \theta)$ *and* $G(\theta) = E_\theta\big(J_\theta E_\theta^X g\big)$.

Proof. The proof is the same except one takes $Z = \begin{pmatrix} p - E_\theta^X g - b(\theta) \\ \dot{L}_\theta \end{pmatrix}$ and $\alpha = \begin{pmatrix} \beta \\ -I_X(\theta)^{-1}\big(G(\theta) + J_\theta b(\theta)\big)^T \beta \end{pmatrix}$. Now one has $J_\theta(E_\theta g) = E_\theta p \dot{L}_\theta^T - J_\theta b(\theta)$. (the differentiability of (3) allows to have $E_\theta \dot{L}_\theta = 0$) □

Remark 1. A proof of this inequality in the case $k = 1$ (i.e. $g(X, Y, \theta) \in \mathbb{R}$) is given in Nayak (2002).

Remark 2. If we relieve the assumption 3 in theorem 3 and in theorem 4 then $G(\theta)$ in the bound for each theorem becomes

$$G(\theta) = J_\theta\big(E_\theta g(X, Y, \theta)\big) - E_\theta\big[\big(E_\theta^X g(X, Y, \theta)\big)\dot{L}_\theta^T\big] = J_\theta\big(E_\theta g(X, Y, \theta)\big) - E_\theta\big[g(X, Y, \theta)\dot{L}_\theta^T\big]$$

Remark 3. The bound is invariant by reparameterization $\theta = h(\xi)$ where h is differentiable.

Remark 4. In the previous theorems we obtain a matricial result. We can deduce a result about the norm of the error, taking the trace of the matrices. The matrix

$$E_\theta\big(p - E_\theta^X g\big)\big(p - E_\theta^X g\big)^T \; - \; G(\theta)\, I_X(\theta)^{-1}\, G(\theta)^T$$

in theorem 3 is positive semi-definite, thus its trace is positive (as the sum of its eigen values which are positive reals). Therefore

$$E_\theta\big|\big|p - E_\theta^X g\big|\big|^2 \; \geqslant \; \mathrm{trace}\big(G(\theta)\, I_X(\theta)^{-1}\, G(\theta)^T\big)$$

and in the case of a biased predictor

$$E_\theta\big|\big|p - E_\theta^X g\big|\big|^2 \; \geqslant \; \big|\big|b(\theta)\big|\big|^2 \; + \; \mathrm{trace}\Big(\big(G(\theta) + J_\theta b(\theta)\big)\, I_X(\theta)^{-1}\, \big(G(\theta) + J_\theta b(\theta)\big)^T\Big).$$

3. Attainment of the bound

When the bound is attained for some value θ of the parameter, $p(X) - E_\theta^X g(X, Y, \theta)$ takes a particular form.

Proposition 5. *Given the assumptions of theorem 3, the equality in (2) holds iff*

$$p(X) = E_\theta^X g(X, Y, \theta) + G(\theta) I_X(\theta)^{-1} \dot{L}_\theta, \quad P_\theta\text{-a.s.} \tag{5}$$

Proof. We denote $p = p(X)$ and $g = g(X, Y, \theta)$. One considers

$$Z = p - E_\theta^X g - G(\theta) I_X(\theta)^{-1} \dot{L}_\theta$$

$$E_\theta Z Z^T = E_\theta\big(p - E_\theta^X g\big)\big(p - E_\theta^X g\big)^T - G(\theta) I_X(\theta)^{-1} G(\theta)^T = 0$$

hence $p - E_\theta^X g = G(\theta) I_X(\theta)^{-1} \dot{L}_\theta$, P_θ-a.s. □

Proposition 6. *Given the assumptions of theorem 4, the equality in (4) holds iff*

$$p(X) = E_\theta^X g(X, Y, \theta) + b(\theta) + \big(G(\theta) + J_\theta b(\theta)\big) I_X(\theta)^{-1} \dot{L}_\theta, \quad P_\theta\text{-a.s.} \tag{6}$$

Proof. The proof is the same taking $Z = p - E_\theta^X g - b(\theta) - \big(G(\theta) + J_\theta b(\theta)\big) I_X(\theta)^{-1} \dot{L}_\theta$.

Remark 5. This result improves on Nayak (2002) who remarked that, in the case $k = 1$ (i.e. $g : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}$), equality in (4) holds iff there exists $H : \Theta \to \mathbb{R}^d$ and $a : \Theta \to \mathbb{R}$ such that (in our notation) $E_\theta^X g(X, Y, \theta) = \langle H(\theta),\, \dot{L}_\theta \rangle + a(\theta) + p(X)$, P_θ-a.s.

We recall that an estimator which variance attains the Cramér-Rao bound is called *efficient*. Following this convention we call a predictor which quadratic prediction error attains the bound of inequality (4) an *efficient predictor*. When $k = d$ and the predictor is efficient, i.e when the bound is globally attained, the density family satisfies some special conditions.

Theorem 7. *Suppose assumptions 2 and 3 hold. Let $p : \mathcal{X} \to \mathbb{R}^k$ be an* efficient *unbiased predictor of $g(X, Y, \theta)$ and $g : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}^k$ a function such that for all $\theta \in \Theta$ the function $(x, y) \mapsto g(x, y, \theta)$ is measurable and $g(X, Y, \theta) \in L^2(P_\theta)$ and $\Theta \subset \mathbb{R}^d$. Denote $p = p(X)$, $g = g(X, Y, \theta)$ and $G(\theta) = E_\theta\big(J_\theta E_\theta^X g\big)$.*

If $G(\theta)$ is an invertible matrix and there exists a differentiable function $\theta \mapsto A(\theta)$, $\Theta \to \mathbb{R}^k$, such that $\big(J_\theta A(\theta)\big)^T = I_X(\theta)G(\theta)^{-1}$.

Then there exists a function $B : \mathcal{X} \times \Theta \to \mathbb{R}$ differentiable in $\theta \in \Theta$ such that

$$f(x, \theta) = \exp\big(\langle A(\theta), p(x)\rangle - B(x, \theta)\big), \quad \text{for all } \theta \in \Theta \text{ and } P_{X,\theta}\text{-a.a. } x \in \mathcal{X}$$

and $\nabla_\theta B(x, \theta) = \big(J_\theta A(\theta)\big)^T E_\theta^{X=x} g$ for all $\theta \in \Theta$ and $P_{X,\theta}$-a.a. $x \in \mathcal{X}$.

Remark 6. The family of densities $f_\theta(x) = \exp\{\langle A(\theta), p(x)\rangle - B(x, \theta)\}$ is not an exponential family since $B(x, \theta)$ might not be a sum $B_1(x) + B_2(\theta)$, we call such a family an *extended* exponential family. Nevertheless when g only depends on θ, the equality $\nabla_\theta B(x, \theta) = \big(J_\theta A(\theta)\big)^T E_\theta^{X=x} g$ implies that there exists B_1 and B_2 such that $B(x, \theta) = B_1(x) + B_2(\theta)$, for all $\theta \in \Theta$ and $x \in \mathcal{X}$. This case is precisely when the framework of prediction degenerates to the framework of estimation.

Remark 7. In the case of estimation (i.e. when g only depends on θ), the assumption that *there exists a differentiable function $\theta \mapsto A(\theta)$, $\Theta \to \mathbb{R}^k$, such that $\big(J_\theta A(\theta)\big)^T = I_X(\theta)G(\theta)^{-1}$* can be relieved, see Müller-Funk et al. (1989) or Liese and Miescke (2008).

Proof. We have $p(X) - E_\theta^X g(X, Y, \theta) = G(\theta) I_X(\theta)^{-1} \dot{L}_\theta$, P_θ-a.s. (proposition 5). Hence

$$\dot{L}_\theta = I_X(\theta)G(\theta)^{-1}\big(p(X) - E_\theta^X g(X, Y, \theta)\big), \quad P_\theta\text{-a.s.} \tag{7}$$

Since $\big(J_\theta A(\theta)\big)^T = I_X(\theta)G(\theta)^{-1}$, one has $I_X(\theta)G(\theta)^{-1} p(X) = \nabla_\theta \langle A(\theta), p(X)\rangle$. Therefore

$$I_X(\theta)G(\theta)^{-1} E_\theta^X g(X, Y, \theta) = \nabla_\theta\big(\log f(X, \theta) - \langle A(\theta), p(X)\rangle\big)$$

Thus there exists a function $B(x,\theta)$ differentiable in θ such that

$$\nabla_\theta B(x,\theta) = I_X(\theta)G(\theta)^{-1}E_\theta^{X=x}g(X,Y,\theta) \quad \text{for all } \theta \text{ and } P_{X,\theta}\text{-a.a. } x.$$

Therefore one can integrate (7) and get

$$f(x,\theta) = \exp\big(\langle A(\theta), p(x)\rangle - B(x,\theta)\big), \quad \text{for all } \theta \text{ and } P_{X,\theta}\text{-a.a. } x \quad \square$$

This proof can be extended to the case of a biased predictor.

Theorem 8. *Suppose assumptions 2 and 3 hold. Let $p : \mathcal{X} \to \mathbb{R}^k$ be a predictor of $g(X,Y,\theta)$ with differentiable bias $b(\theta)$ and $g : \mathcal{X}\times\mathcal{Y}\times\Theta \to \mathbb{R}^k$ a function such that for all $\theta\in\Theta$ the function $(x,y)\mapsto g(x,y,\theta)$ is measurable and $g(X,Y,\theta)\in L^2(P_\theta)$ and $\Theta\subset\mathbb{R}^d$, and the predictor $p(X)$ attains the bound in (4) for each θ. Denote $p=p(X)$, $g=g(X,Y,\theta)$ and $G(\theta)=E_\theta\big(J_\theta E_\theta^X g\big)$.*

If $G(\theta)$ is an invertible matrix and there exists a differentiable function $\theta\mapsto A(\theta)$, $\Theta\to\mathbb{R}^k$, such that $\big(J_\theta A(\theta)\big)^T = I_X(\theta)\big(G(\theta)+J_\theta b(\theta)\big)^{-1}$.

Then there exists a function $B:\mathcal{X}\times\Theta\to\mathbb{R}$ differentiable in $\theta\in\Theta$ such that

$$f(x,\theta) = \exp\big(\langle A(\theta), p(x)\rangle - B(x,\theta)\big), \quad \text{for all } \theta\in\Theta \text{ and } P_{X,\theta}\text{-a.a. } x\in\mathcal{X}$$

and $\nabla_\theta B(x,\theta) = \big(J_\theta A(\theta)\big)^T\big(E_\theta^{X=x}g + b(\theta)\big)$ for all $\theta\in\Theta$ and $P_{X,\theta}$-a.a. $x\in\mathcal{X}$.

The following theorem is a converse.

Theorem 9. *Suppose assumption 2 holds. Let $g:\mathcal{X}\times\mathcal{Y}\times\Theta\to\mathbb{R}^k$ be a function such that for all $\theta\in\Theta$ the function $(x,y)\mapsto g(x,y,\theta)$ is measurable and $g(X,Y,\theta)\in L^2(P_\theta)$ and $\Theta\subset\mathbb{R}^d$. Suppose the observed variable X has density*

$$f(x,\theta) = \exp\big(\langle A(\theta), p(x)\rangle - B(x,\theta)\big), \quad \theta\in\Theta \tag{8}$$

with $p:\mathcal{X}\to\mathbb{R}^k$ a measurable function and $A:\Theta\to\mathbb{R}^d$ differentiable with invertible Jacobian matrix and $B:\mathcal{X}\times\Theta\to\mathbb{R}$ twice differentiable with respect to θ, where A and B satisfy $E_\theta^X g(X,Y,\theta) = \big(J_\theta A(\theta)\big)^{-1\,T}\nabla_\theta B(X,\theta)$.

If $\int f(x,\theta)\,d\mu(x)$ is two times differentiable under the integral sign, then $p(X)$ is an efficient unbiased predictor of $g(X,Y,\theta)$.

Remark 8. Any density $f(x,\theta)$ that does not vanish can be written like (8), by choosing $A(\theta)=\theta$, $p(x)=0$, and $B(x,\theta)=-\log(f(x,\theta))$. But under this form, the quantity to predict given by the theorem is not necessarily interesting. The theorem is useful when it is possible to write the density in such a way that the quantity $\big(J_\theta A(\theta)\big)^{-1T}\nabla_\theta B(X,\theta)$ is an interesting quantity to predict.

Proof. It holds $\int f(x,\theta)\,d\mu(x)=1$ hence for all $i=1,\dots,d$, $\int \frac{\partial}{\partial\theta_i}f(x,\theta)\,d\mu(x)=0$ therefore

$$E_\theta\Big(\frac{\partial}{\partial\theta_i}A(\theta)^T p(X)-\frac{\partial}{\partial\theta_i}B(X,\theta)\Big)=0 \tag{9}$$

Thus $E_\theta\big(\big(J_\theta A(\theta)\big)^T p(X)-\nabla_\theta B(X,\theta)\big)=0$, hence

$$E_\theta p(X)=E_\theta\big((J_\theta A(\theta))^{-1T}\nabla_\theta B(X,\theta)\big)=E_\theta g(X,Y,\theta).$$

Therefore $p(X)$ is unbiased for predicting $g(X,Y,\theta)$. We now compute the Cramér-Rao bound. We denote $p=p(X)$ and $g=g(X,Y,\theta)$. Here we did not assume assumption (3) thus $G(\theta)=J_\theta\big(E_\theta g\big)-E_\theta\big[\big(E^X_\theta g\big)\dot L^T_\theta\big]$.

$$\begin{aligned}
G(\theta) &= J_\theta(E_\theta p)-E_\theta\big((J_\theta A(\theta))^{-1\,T}\nabla_\theta B(X,\theta)\dot L^T_\theta\big)\\
&= E_\theta p(X)\dot L^T_\theta-E_\theta\big((J_\theta A(\theta))^{-1\,T}\nabla_\theta B(X,\theta)\dot L^T_\theta\big)\\
&= E_\theta\big[\,\big(p(X)-(J_\theta A(\theta))^{-1\,T}\nabla_\theta B(X,\theta)\big)\,\dot L^T_\theta\,\big]
\end{aligned}$$

Differentiating (8) we get $\dot L_\theta=\nabla_\theta\big(A(\theta)^T p(X)-B(X,\theta)\big)$ hence

$$\begin{aligned}
G(\theta) &= E_\theta\big[\,\big(p(X)-(J_\theta A(\theta))^{-1\,T}\nabla_\theta B(X,\theta)\big)\,\big(\nabla_\theta(A(\theta)^T p(X)-B(X,\theta))\big)^T\,\big]\\
&= \big(J_\theta A(\theta)\big)^{-1\,T}E_\theta\big[\,\big((J_\theta A(\theta))^T p(X)-\nabla_\theta B(X,\theta)\big)\,\big(\nabla_\theta(A(\theta)^T p(X)-B(X,\theta))\big)^T\,\big]\\
&= \big(J_\theta A(\theta)\big)^{-1\,T}E_\theta\big[\,\nabla_\theta\big(A(\theta)^T p(X)-B(X,\theta)\big)\,\big(\nabla_\theta(A(\theta)^T p(X)-B(X,\theta))\big)^T\,\big]\\
&= \big(J_\theta A(\theta)\big)^{-1\,T}E_\theta\big[\dot L_\theta\dot L^T_\theta\big]\\
&= \big(J_\theta A(\theta)\big)^{-1\,T}I_X(\theta)
\end{aligned}$$

Hence the Cramér-Rao bound is $G(\theta)I^{-1}(\theta)G(\theta)^T=\big(J_\theta A(\theta)\big)^{-1\,T}I_X(\theta)\big(J_\theta A(\theta)\big)^{-1}$.
We now compute $E_\theta(p-E^X_\theta g)(p-E^X_\theta g)^T$, we differentiate (9) with respect to θ_j

$$E_\theta\Big(\frac{\partial}{\partial\theta_i}A(\theta)^T p(X)-\frac{\partial}{\partial\theta_i}B(X,\theta)\Big)^2=\Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}B(X,\theta)-\frac{\partial^2}{\partial\theta_i\partial\theta_j}A(\theta)^T p(X)\Big)$$

$$E_\theta\Big(\big(J_\theta A(\theta)\big)^T p(X)-\nabla_\theta B(X,\theta)\Big)\Big(\big(J_\theta A(\theta)\big)^T p(X)-\nabla_\theta B(X,\theta)\Big)^T$$

$$= \Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}B(X,\theta) - \frac{\partial^2}{\partial\theta_i\partial\theta_j}A(\theta)^T p(X)\Big)_{1\leqslant i,j\leqslant d}$$

Now Fisher's information is also equal to

$$I_X(\theta) = -E_\theta\Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(f(X,\theta))\Big)_{1\leqslant i,j\leqslant d} = \Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}B(X,\theta) - \frac{\partial^2}{\partial\theta_i\partial\theta_j}A(\theta)^T p(X)\Big)_{1\leqslant i,j\leqslant d}$$

Hence

$$E_\theta\Big(\big(J_\theta A(\theta)\big)^T p(X) - \nabla_\theta B(X,\theta)\Big)\Big(\big(J_\theta A(\theta)\big)^T p(X) - \nabla_\theta B(X,\theta)\Big)^T = I_X(\theta)$$

i.e.

$$E_\theta\Big(p(X) - \big(J_\theta A(\theta)\big)^{-1\,T}\nabla_\theta B(X,\theta)\Big)\Big(p(X) - \big(J_\theta A(\theta)\big)^{-1\,T}\nabla_\theta B(X,\theta)\Big)^T = \big(J_\theta A(\theta)\big)^{-1\,T} I_X(\theta)\big(J_\theta A(\theta)\big)^{-1}.$$

Therefore

$$E_\theta\big(p - E_\theta^X g\big)\big(p - E_\theta^X g\big)^T = \big(J_\theta A(\theta)\big)^{-1\,T} I_X(\theta)\big(J_\theta A(\theta)\big)^{-1} = G(\theta)I_X(\theta)^{-1}G(\theta)^T.$$

The Cramér-Rao bound is attained, p is an efficient unbiased predictor for g. □

The following is a multivariate version of theorem 1.8 from Bosq and Blanke (2007).

Theorem 10. *Suppose assumption 2 holds. Let $g : \mathcal{X}\times\mathcal{Y}\times\Theta \to \mathbb{R}^k$ be a function such that for all $\theta\in\Theta$ the function $(x,y)\mapsto g(x,y,\theta)$ is measurable and $g(X,Y,\theta)\in L^2(P_\theta)$ and $\Theta\subset\mathbb{R}^d$. Let $s(X)$ be an efficient unbiased estimator of $\psi(\theta):\Theta\to\mathbb{R}^k$ a differentiable function and suppose g is such that $E_\theta^X g(X,Y,\theta) = \phi(X) + \psi(\theta)$, $\theta\in\Theta$ with ϕ mesurable, ϕ and ψ known functions. Then $p(X) = \phi(X) + s(X)$ is an efficient unbiased predictor of $g(X,Y,\theta)$.*

Proof. The efficiency of $s(X)$ for estimating $\psi(\theta)$ implies

$$E_\theta\big(s(X) - \psi(\theta)\big)\big(s(X) - \psi(\theta)\big)^T = \big(J_\theta\psi(\theta)\big)I_X(\theta)^{-1}\big(J_\theta\psi(\theta)\big)^T$$

Let us denote $p = p(X)$ and $g = g(X,Y,\theta)$. The Cramér-Rao bound of p predictor of g is $G(\theta)I_X(\theta)^{-1}G(\theta)$ with

$$\begin{aligned} G(\theta) &= J_\theta\big(E_\theta g\big) - E_\theta\big[\big(E_\theta^X g\big)\dot{L}_\theta^T\big] \\ &= J_\theta\big(E_\theta\phi(X)\big) + J_\theta\psi(\theta) - \big[E_\theta(\phi(X)\dot{L}_\theta^T) + E_\theta\big(\psi(\theta)\dot{L}_\theta^T\big)\big] \\ &= J_\theta\big(E_\theta\phi(X)\big) + J_\theta\psi(\theta) - J_\theta\big(E_\theta\phi(X)\big) - \psi(\theta)E_\theta\big(\dot{L}_\theta^T\big) \end{aligned}$$

Now $E_\theta(\dot{L}_\theta) = 0$, hence $G(\theta) = J_\theta\psi(\theta)$ and the Cramér-Rao bound is $\big(J_\theta\psi(\theta)\big)I_X(\theta)^{-1}\big(J_\theta\psi(\theta)\big)^T$. The quadratic error of the predictor with respect to the conditional expectation is

$$E_\theta\big(p-E_\theta^X g\big)\big(p-E_\theta^X g\big)^T = E_\theta\big(s(X)-\psi(\theta)\big)\big(s(X)-\psi(\theta)\big)^T = \big(J_\theta\psi(\theta)\big)I_X(\theta)^{-1}\big(J_\theta\psi(\theta)\big)^T$$

The Cramér-Rao bound is attained, p is an efficient unbiased predictor of g. □

4. Prediction of a bivariate Poisson process

Let us consider the bivariate Poisson process $\big(N_t\big)_{t\geqslant 0} = \big(N_1(t), N_2(t)\big)_{t\geqslant 0}$ following the definition of Marshall and Olkin (1967). It is markovian and its increments are independent and stationary. The parameter of the model is $\theta = (\lambda_1,\, \lambda_2,\, \lambda_3)$ where $(\lambda_1 - \lambda_3,\, \lambda_2 - \lambda_3,\, \lambda_3) \in (\mathbb{R}_+^*)^3$. The distribution is

$$f(x,\theta) = P_\theta\Big(N_t = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\Big) = e^{-(\lambda_1+\lambda_2-\lambda_3)t} \sum_{k=0}^{\min(x_1,\,x_2)} \frac{\lambda_3^k(\lambda_1-\lambda_3)^{x_1-k}(\lambda_2-\lambda_3)^{x_2-k}t^{x_1+x_2-k}}{k!(x_1-k)!(x_2-k)!}$$

with $x = \begin{pmatrix} x_1 & x_2 \end{pmatrix}^T \in \mathbb{N}^2$. Kocherlakota and Kocherlakota (1992) p. 106 and the reparameterization $(\lambda_1,\, \lambda_2,\, \lambda_3) \mapsto (\lambda_1 t,\, \lambda_2 t,\, \lambda_3 t)$ (see formula (6.16) in Lehmann and Casella (1998) p.125) allow to show that the inverse of the information matrix is

$$I(\theta)^{-1} = \frac{1}{t}\begin{pmatrix} \lambda_1 & \lambda_3 & \lambda_3 \\ \lambda_3 & \lambda_2 & \lambda_3 \\ \lambda_3 & \lambda_3 & \delta \end{pmatrix}$$

(δ is given in Kocherlakota and Kocherlakota (1992))

In what follows we study the problem of predicting N_{t+h} assuming we know the process at time t (we do not assume that we know the process before time t). Thus in the framework of statistical prediction, presented in the introduction, we have $X = N_t$ and $Y = N_{t+h}$. The assumption 2 is fulfilled.

The following equality in distribution holds

$$\begin{pmatrix} N_1(t) \\ N_2(t) \end{pmatrix} \overset{d}{=} \begin{pmatrix} Z_1 + Z_3 \\ Z_2 + Z_3 \end{pmatrix}$$

where Z_1, Z_2 and Z_3 are independent random variables with Poisson distribution and respective parameters $(\lambda_1 - \lambda_3)t$, $(\lambda_2 - \lambda_3)t$ and $\lambda_3 t$. Using this property we compute

the conditional expectation of N_{t+h}

$$E_\theta^X Y = E_\theta^{N_t} N_{t+h} = E_\theta[N_{t+h} - N_t] + N_t = E_\theta N_h + N_t = h \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} + N_t$$

Using this expression of the conditional expectation of $Y = N_{t+h}$ given $X = N_t$ and Lebesgue's theorem one can prove that assumption 3 is fulfilled.

The predictor $p(N_t) = \frac{t+h}{t} N_t$ is unbiased. The equality (1) can be differentiated under the integral sign with respect to λ_1, λ_2 and λ_3, using Lebesgue's theorem.

We know that $p(N_t)$ is efficient in the univariate case (see Bosq and Blanke (2007) example 1.13 p. 26). We are going to see that it is also true in the bivariate case. The quadratic error of the predictor with respect to the conditional expectation is

$$E_\theta \big(p(N_t) - E_\theta^{N_t} N_{t+h}\big) \big(p(N_t) - E_\theta^{N_t} N_{t+h}\big)^T = \frac{h^2}{t} \begin{pmatrix} \lambda_1 & \lambda_3 \\ \lambda_3 & \lambda_2 \end{pmatrix}$$

We now compute the Cramér-Rao bound $G(\theta) I(\theta)^{-1} G(\theta)^T$.

$$G(\theta) = E_\theta \big(J_\theta E_\theta^{N_t} N_{t+h}\big) = \begin{pmatrix} h & 0 & 0 \\ 0 & h & 0 \end{pmatrix}$$

$$E_\theta \big(p(N_t) - E_\theta^{N_t} N_{t+h}\big) \big(p(N_t) - E_\theta^{N_t} N_{t+h}\big)^T = G(\theta) I(\theta)^{-1} G(\theta)^T$$

The Cramér-Rao bound is attained.

Acknowledgement

I am thankful to my PhD advisor Professor Denis Bosq for introducing the topic of statistical prediction to me and for his guidance while working on this paper. I thank my PhD co-advisor Professor Delphine Blanke for her advice for writing this second version of the paper. I thank the two referees of Statistics and Probability Letters for their time reading the first version of this paper and for their thoughtful and detailed comments which allowed to improve this paper.

References

Bosq, D., Blanke, D., 2007. Inference and prediction in large dimensions. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester.

Kocherlakota, S., Kocherlakota, K., 1992. Bivariate discrete distributions. Vol. 132 of Statistics: Textbooks and Monographs. Marcel Dekker Inc., New York.

Lehmann, E. L., Casella, G., 1998. Theory of point estimation, 2nd Edition. Springer Texts in Statistics. Springer-Verlag, New York.

Liese, F., Miescke, K.-J., 2008. Statistical decision theory. Springer Series in Statistics. Springer, New York, estimation, testing, and selection.

Marshall, A. W., Olkin, I., 1967. A generalized bivariate exponential distribution. J. Appl. Probability 4, 291–302.

Müller-Funk, U., Pukelsheim, F., Witting, H., 1989. On the attainment of the Cramér-Rao bound in $\mathbf{L}_r$-differentiable families of distributions. Ann. Statist. 17 (4), 1742–1748.

Nayak, T. K., 2002. Rao-Cramer type inequalities for mean squared error of prediction. Amer. Statist. 56 (2), 102–106.
URL http://dx.doi.org/10.1198/000313002317572763

Yatracos, Y. G., 1992. On prediction and mean squared error. Canad. J. Statist. 20 (2), 187–200.
URL http://dx.doi.org/10.2307/3315467