Efficient prediction in L2-differentiable families of distributions

Emmanuel Onzon

Pub. Inst. Stat. Univ. Paris
63, fasc. 2-3, 2019, 33-44
Numéro spécial en l'honneur des 80 ans de Denis Bosq /
Special issue in honour of Denis Bosq's 80th birthday

Efficient prediction in L^2-differentiable families of distributions

Emmanuel Onzon

Algolux Inc.

Abstract: A result about efficient predictors is presented. It is proved that the existence of an efficient predictor, *i.e.* which risk attains the Cramér-Rao bound for predictors, implies that the family of distributions is of a special form which can be seen as an extension of the notion of exponential family. The result is proved under L^2-differentiability conditions.

1. Unbiased statistical prediction

Statistical prediction relates to the inference of an unobserved random quantity from observations, it is considered here as an extension of point estimation, where the quantity to infer is not necessarily deterministic. We follow the framework posed by [18]. In full generality, the problem of statistical prediction is to estimate a quantity $f(X, Y, \theta)$, we shall say *predict* $f(X, Y, \theta)$, where X is an observed random variable representing the observations, Y an unobserved random variable and θ the parameter of the model $\{\mathrm{P}_\theta \,|\, \theta \in \Theta\}$ which the distribution of (X, Y) is supposed to belong to. The random variables X and Y may be of finite or infinite dimensions. For example in the finite dimensional case, X could be the value of a vector gaussian process $U = (U_t,\, t \in \mathbb{N})$ at time T with $U_t \in \mathbb{R}^k$ for all t, *i.e.* $X := U_T \in \mathbb{R}^k$. The random variable Y could be the value of the random process U at time $T+1$, $Y := U_{T+1} \in \mathbb{R}^k$. For example in the infinite dimensional case, X could be the path of a continuous time random process $(U_t,\, t \in [0, \infty))$ until time $T \in [0, \infty)$, with $U_t \in \mathbb{R}$ for all $t \in [0, \infty)$, *i.e.* $X := (U_t,\, t \in [0, T])$. The random variable Y could be the path of the random process U between times T and $T + h$, $Y := (U_t,\, T < t \leqslant T + h)$.

We shall assume that g takes its values in $\mathbb{R}^k$ and $\Theta \subset \mathbb{R}^d$. That framework encompasses a wide variety of statistical problems ranging from stochastic processes prediction and time series forecasting ([5], [1], [3], [16]) to latent variable models and random effects inference ([10], [12]). If $p(X)$ is used to predict $f(X, Y, \theta)$ we shall call it a predictor and measure its performance with its mean squared error of prediction which breaks down in the following sum

$$\mathrm{E}_\theta(p(X) - f(X, Y, \theta))^{\times 2} = \mathrm{E}_\theta(p(X) - g(X, \theta))^{\times 2} + \mathrm{E}_\theta(g(X, \theta) - f(X, Y, \theta))^{\times 2},$$

AMS 2000 subject classifications: 62M20, 62J02
Keywords and phrases: Cramér-Rao inequality, lower bound, prediction, L2 differentiable families

with $g(X,\theta) = E_\theta[f(X,Y,\theta)|X]$ and where we use the notation $A^{\times 2} = AA'$ the product of a matrix with its transpose. The second term of the right hand side is incompressible, it does not depend on the choice of the predictor. Hence in what follows we are interested in the first term which we call quadratic error of prediction (QEP) and denote by $R(\theta)$.

$$R(\theta) = \mathrm{E}_\theta(p(X) - g(X,\theta))^{\times 2}.$$

A lower bound of Cramér-Rao type has been proved for the QEP with conditions of point differentiability of the family of the densities of the distributions of the model with respect to the parameter and conditions of differentiability under the integral sign ([18], [11], [2]). The bound has also been proved for conditions of L^2-differentiability of the family of distributions of the model ([8], [15]). In the one-dimensional case ($k = d = 1$) and for unbiased predictors it reads,

$$\mathrm{E}_\theta(p(X) - g(X,\theta))^2 \geqslant \frac{\left(\mathrm{E}_\theta \partial_\theta g(X,\theta)\right)^2}{I(\theta)},$$

where $I(\theta)$ is the Fisher information. At the end of this section, we recall the statement of this inequality in the case of a multidimensional parameter and under conditions of L^2-differentiability of the family of distributions.

When the mean squared error of an estimator attains the Cramér-Rao bound, we say that it is *efficient*. By analogy, an efficient predictor is a predictor which QEP attains the Cramér-Rao bound. In the case of estimation, it is proved that there exists an efficient estimator $\delta(X)$ of $\psi(\theta) \in \mathbb{R}^d$ if and only if the family of distributions of the model is exponential, *i.e.* of the form

$$\frac{\mathrm{dP}_\theta}{\mathrm{dP}_{\theta_0}}(x) = \exp\{A(\theta)'\delta(x) - B(\theta)\},$$

for some $\theta_0 \in \Theta$, and differentiable functions $A : \Theta \to \mathbb{R}^k$ and $B : \Theta \to \mathbb{R}$, with $(\mathrm{J}_\theta A(\theta))' = I(\theta)(\mathrm{J}_\theta \psi(\theta))^{-1}$ and $\nabla_\theta B(\theta) = (\mathrm{J}_\theta A(\theta))'\psi(\theta)$. The result has been proved under different conditions ([17], [4], [9]).

An analogous result for prediction appears in [2] in the one-dimensional case and in [14] in the multidimensional case. In both cases, the result is proved under conditions of point differentiability of the family of the densities of the distributions of the model and differentiability under the integral sign. For this result, the family is not necessarily exponential but has a form which may be seen as an extension of the notion of exponential family. There exists an efficient predictor $p(X)$ to predict $g(X,\theta) \in \mathbb{R}^k$, in the special case $k = d$, if and only if

$$\frac{\mathrm{dP}_\theta}{\mathrm{dP}_{\theta_0}}(x) = \exp\{A(\theta)'p(x) - B(x,\theta)\},$$

for some $\theta_0 \in \Theta$, and differentiable functions $A : \Theta \to \mathbb{R}^k$ and $B : \Theta \times E \to \mathbb{R}$, with $(\mathrm{J}_\theta A(\theta))' = I(\theta)(\mathrm{E}_\theta \mathrm{J}_\theta g(X,\theta))^{-1}$ and $\nabla_\theta B(x,\theta) = (\mathrm{J}_\theta A(\theta))'g(x,\theta)$. Section 2

presents a proof of this result under L^2-differentiability conditions. The proof is based on the proof of the result for estimation that appears in [9].

For convenience, the appendix gathers definitions and results on L^2-differentiability that are used in this paper. We now give a set of assumptions under which the Cramér-Rao bound for predictors holds. We use the following notations, let $\mathcal{X}$ be $\mathbb{R}^d$ (the space where the random variable X takes its values) and $\mathcal{B}$ its Borel σ-algebra.

Assumption 1.1. *Consider a model $(\mathcal{X}, \mathcal{B}, P_\theta, \theta \in \Theta)$, $\theta_0 \in \mathring{\Theta}$, a neighbourhood $U(\theta_0)$ of θ_0 and a function $g : \mathcal{X} \times \Theta \to \mathbb{R}^k$, with $g(\cdot, \theta)$ measurable for all $\theta \in \Theta$, such that the following conditions hold.*

1. *The family $(P_\theta, \theta \in \Theta)$ is L^2-differentiable at θ_0, with derivative $\dot{L}_{\theta_0}$.*
2. *Fisher matrix information $I(\theta_0)$ is invertible.*
3. *For all θ, $\theta' \in U(\theta_0)$, $g(X, \cdot)$ is P_θ-almost surely differentiable at θ' and*

$$\sup_{(\theta,\theta') \in U(\theta_0)^2} E_\theta \|J_\theta g(X, \theta')\|^2_{M_{k,d}} < \infty.$$

4. $\sup_{(\theta,\theta') \in U(\theta_0)^2} E_\theta L^2_{\theta,\theta'} < \infty$

Moreover consider a predictor $p(X)$ taking values in $\mathbb{R}^k$. There is $U(\theta_0)$, a neighbourhood of θ_0, such that

5. $\sup_{\theta \in U(\theta_0)} E_\theta \|p(X)\|^2_{\mathbb{R}^k} < \infty$.

We state below the inequality for unbiased predictors. A proof can be found in [15]. Here we say that $p(X)$ a predictor of $g(X, \theta)$ is an *unbiased predictor* if $E_\theta(p(X)) = E_\theta(g(X, \theta))$ for all $\theta \in \Theta$ (for other concepts of risk unbiasedness pertaining to prediction problems see [13]).

Theorem 1.1. *Let $(\mathcal{X}, \mathcal{B}, P_\theta, \theta \in \Theta)$ be a model, $\theta_0 \in \mathring{\Theta}$, and $p(X)$ an unbiased predictor of $g(X, \theta)$ taking values in $\mathbb{R}^k$, that satisfies Assumption 1.1.*

Then the QEP of $p(X)$ at θ_0 satisfies the following inequality.

$$E_{\theta_0}(p(X) - g(X, \theta_0))^{\times 2} \geqslant G(\theta_0) I(\theta_0)^{-1} G(\theta_0)', \tag{1.1}$$

with $G(\theta) = E_\theta J_\theta g(X, \theta)$. The equality holds in (1.1) iff

$$p(X) = g(X, \theta_0) + G(\theta_0) I(\theta_0)^{-1} \dot{L}_{\theta_0}, \quad P_{\theta_0}\text{-a.s.}$$

2. Efficient prediction

A predictor $p(X)$ is said *efficient* when its QEP attains the Cramér-Rao bound.

Theorem 2.1. *Suppose $k = d$. Let Θ be a connected open set of $\mathbb{R}^d$. Let $(\mathcal{X}, \mathcal{B}, P_\theta, \theta \in \Theta)$ be a model, $g : \mathcal{X} \times \Theta \to \mathbb{R}^k$ and $p(X)$ an unbiased predictor of $g(X, \theta)$, that satisfy Assumption 1.1 for all $\theta \in \Theta$.*

Suppose the following conditions hold.

1. *$p(X)$ is efficient.*
2. *For all $\theta \in \Theta$, $G(\theta) = \mathrm{E}_\theta \mathrm{J}_\theta g(X, \theta)$ is invertible.*
3. *There is $A : \Theta \to \mathbb{R}^k$ a differentiable function over Θ, such that $(\mathrm{J}_\theta A(\theta))' = I(\theta)G(\theta)^{-1}$, for all $\theta \in \Theta$.*
4. *$\mathcal{X}$ is a topological space and $(\mathcal{X}, \mathcal{B})$ is a σ-compact space.*
5. *For all compact sets $C \subset \mathcal{X}$, $\tilde{C} \subset \Theta$, $\sup_{x \in C, \theta \in \tilde{C}} \|\mathrm{J}_\theta g(x, \theta)\| < \infty$.*
6. *$\theta \mapsto I(\theta)$ and $\theta \mapsto G(\theta)$ are continuous.*

Then, for $\theta_0 \in \Theta$ fixed, there is a function $B : \mathcal{X} \times \Theta \to \mathbb{R}$, differentiable at $\theta \in \Theta$, such that for all $\theta \in \Theta$, for P_{θ_0}-almost all $x \in \mathcal{X}$,

$$\frac{d\mathrm{P}_\theta}{d\mathrm{P}_{\theta_0}}(x) = \exp\big(A(\theta)'p(x) - B(x, \theta)\big),$$

and $\nabla_\theta B(x, \theta) = (\mathrm{J}_\theta A(\theta))' g(x, \theta)$.

Proof. Let $\theta \in \Theta$. The predictor $p(X)$ is efficient hence P_θ-a.s.

$$\begin{aligned} p(X) &= g(X, \theta) + \big(\mathrm{E}_\theta \mathrm{J}_\theta g(X, \theta)\big)\, I(\theta)^{-1} \dot{L}_\theta \\ &= g(X, \theta) + G(\theta) I(\theta)^{-1} \dot{L}_\theta. \end{aligned}$$

Hence

$$\dot{L}_\theta = I(\theta)G(\theta)^{-1}\big(p(X) - g(X, \theta)\big).$$

Let $s \mapsto \theta_s$ be a continuously differentiable path from θ_0 to θ with $s \in [0, 1]$. This path exists because Θ is open and connected. We set

$$f(x) = \exp\Big(\int_0^1 \dot{\theta}_s' \dot{L}_{\theta_s}(x)\, ds\Big) = \exp\Big(\int_0^1 \big(\dot{\theta}_s' I(\theta_s) G(\theta_s)^{-1} p(x) - \phi(s, x)\big) ds\Big),$$

with

$$\phi(s, x) = \dot{\theta}_s' I(\theta_s) G(\theta_s)^{-1} g(x, \theta_s).$$

We prove that for all event $B \in \mathcal{B}$, the following equality holds

$$\int_B f(X)\, d\mathrm{P}_{\theta_0} = \mathrm{P}_\theta(B). \tag{2.1}$$

Since $\mathcal{B}$ is σ-compact, one may assume that B is a compact set. For P_θ-almost all $x \in \mathcal{X}$, $s \mapsto g(x, \theta_s)$ is differentiable over $[0, 1]$ (we remove from B the points x for which differentiability does not hold). We set

$$M = \sup_{x \in B, s \in [0,1]} \|\partial_s g(x, \theta_s)\| \leqslant \sup_{s \in [0,1]} \|\dot{\theta}_s\| \sup_{x \in B, t \in \{\theta_s, s \in [0,1]\}} \|\mathrm{J}_\theta g(x, t)\|.$$

The first supremum of the right hand side is finite because $(\theta_s, s \in [0, 1])$ is continuously differentiable. The second one is finite from condition 5. Hence $M < \infty$. Let $\varepsilon > 0$ and $(R_i)_{i \in \mathbb{N}}$ be a partition of $\mathbb{R}^k$ in rectangles of diameters at most ε, and let

$$n = \left\lceil \frac{M}{\varepsilon} \right\rceil.$$

For all $u \in \mathbb{N}^{n+1}$ we let

$$S_u = \left\{ x \in \mathcal{X} \,|\, \forall i \in \{0, \dots, n\},\, g(x, \theta_{i/n}) \in R_{u_i} \right\}.$$

We then define

$$B_{i,u} = B \cap p^{-1}(R_i) \cap S_u.$$

Let $x \in B_{i,u}$ and $s \in [0,1]$ then,

$$\begin{aligned} \|g(x,\theta_s)\| &\leqslant \|g(x,\theta_{\lfloor sn \rfloor/n})\| + \|g(x,\theta_{\lfloor sn \rfloor/n}) - g(x,\theta_s)\| \\ &\leqslant \sup_{y \in R_{u_{\lfloor sn \rfloor}}} \|y\| + M\,|\lfloor sn \rfloor/n - s| \\ &\leqslant \sup_{y \in R_{u_{\lfloor sn \rfloor}}} \|y\| + \frac{M}{n} \end{aligned}$$

$$\|g(x,\theta_s)\| \leqslant \sigma_u + \frac{M}{n} < \infty, \tag{2.2}$$

with

$$\sigma_u = \sup_{0 \leqslant i \leqslant n,\, y \in R_{u_i}} \|y\|.$$

We prove by contradiction that $\mathrm{P}_{\theta_0}(B_{i,u}) > 0$ iff $\mathrm{P}_{\theta}(B_{i,u}) > 0$. Without loss of generality, suppose that $\mathrm{P}_{\theta_s}(B_{i,u}) > 0$ for $s \in [0,1)$ and $\mathrm{P}_{\theta}(B_{i,u}) = 0$. We set $H(s) = \log \mathrm{P}_{\theta_s}(B_{i,u})$. From Proposition A.1, $s \mapsto \mathrm{P}_{\theta_s}(B_{i,u})$ is differentiable over $[0,1]$, hence it is continuous over $[0,1]$. Hence

$$\lim_{s \to 1^-} \mathrm{P}_{\theta_s}(B_{i,u}) = 0.$$

And therefore

$$\lim_{s \to 1^-} H(s) = -\infty. \tag{2.3}$$

Besides H is differentiable over $[0,1)$. Its derivative is

$$h(s) = \frac{\dot{\theta}_s' \nabla_\theta \mathrm{P}_{\theta_s}(B_{i,u})}{\mathrm{P}_{\theta_s}(B_{i,u})} = \frac{1}{\mathrm{P}_{\theta_s}(B_{i,u})} \dot{\theta}_s' \int_{B_{i,u}} \dot{L}_{\theta_s} d\mathrm{P}_{\theta_s} = m(s|B_{i,u}) - \phi(s|B_{i,u}),$$

where

$$\begin{aligned} m(s|B_{i,u}) &= \mathrm{P}_{\theta_s}(B_{i,u})^{-1} \int_{B_{i,u}} \dot{\theta}_s' I(\theta_s) G(\theta_s)^{-1} p(X)\, d\mathrm{P}_{\theta_s}, \\ \phi(s|B_{i,u}) &= \mathrm{P}_{\theta_s}(B_{i,u})^{-1} \int_{B_{i,u}} \phi(s,X)\, dP_{\theta_s}. \end{aligned} \tag{2.4}$$

We prove that $h(s)$ is bounded. The function $s \mapsto \dot{\theta}_s' I(\theta_s) G(\theta_s)^{-1}$ is continuous over $[0,1]$, from condition 6, hence

$$c = \sup_{s \in [0,1]} \|\dot{\theta}_s' I(\theta_s) G(\theta_s)^{-1}\| < \infty.$$

Let $x \in B_{i,u}$, then $p(x) \in R_i \cup \{0\}$ hence

$$|\dot\theta_s' I(\theta_s) G(\theta_s)^{-1} p(x)| \leqslant c \sup_{y \in R_i} \|y\| = c\rho_i.$$

Hence $|m(s|B_{i,u})| \leqslant c\rho_i$. From the continuity argument above and the bound (2.2) we deduce

$$|\phi(s,x)| \leqslant c\,\|g(x,\theta_s)\| \leqslant c(\sigma_u + M/n). \tag{2.5}$$

Hence $\phi(s|B_{i,u}) \leqslant c(\sigma_u + M/n)$. We deduce that h is bounded over $[0,1)$, which contradicts (2.3). Hence $\mathrm{P}_{\theta_0}(B_{i,u}) > 0$ iff $\mathrm{P}_\theta(B_{i,u}) > 0$, which implies that the distributions P_θ and P_{θ_0} are absolutely continuous with respect to each other.

We now prove (2.1) when $\mathrm{P}_\theta(B) > 0$. One may write

$$\begin{aligned}
\int_{B_{i,u}} f(X)\, d\mathrm{P}_{\theta_0} &= \int_{B_{i,u}} \exp\Bigg(\int_0^1 \Big(\dot\theta_s' I(\theta_s) G(\theta_s)^{-1} p(X) - m(s|B_{i,u}) + h(s) \\
&\qquad + \phi(s|B_{i,u}) - \phi(s,X) \Big) ds \Bigg)\, d\mathrm{P}_{\theta_0} \\
&= \int_{B_{i,u}} \exp\Bigg(\int_0^1 \Big(\dot\theta_s' I(\theta_s) G(\theta_s)^{-1} p(X) - m(s|B_{i,u}) \Big)\, ds \\
&\qquad + \int_0^1 \big(\phi(s|B_{i,u}) - \phi(s,X) \big)\, ds \Bigg)\, d\mathrm{P}_{\theta_0} \frac{\mathrm{P}_\theta(B_{i,u})}{\mathrm{P}_{\theta_0}(B_{i,u})}.
\end{aligned}$$

For all $x \in B_{i,u}$, $\dot\theta_s' I(\theta_s) G(\theta_s)^{-1} p(x)$ lies in the image of R_i by the map

$$\nu : \, y \mapsto \dot\theta_s' I(\theta_s) G(\theta_s)^{-1} y.$$

The quantity $m(s|B_{i,u})$ also lies in the image of R_i by the map ν, since it is the mean of $\dot\theta_s' I(\theta_s) G(\theta_s)^{-1} p(x)$ over $B_{i,u}$. Hence

$$\Big| \dot\theta_s' I(\theta_s) G(\theta_s)^{-1} p(X) - m(s|B_{i,u}) \Big| \leqslant \sup_{s \in [0,1]} \|\dot\theta_s' I(\theta_s) G(\theta_s)^{-1}\| \operatorname{diam}(R_i) \leqslant c\varepsilon.$$

Hence for all $x \in B_{i,u}$,

$$\Big| \int_0^1 \big(\dot\theta_s' I(\theta_s) G(\theta_s)^{-1} p(x) - m(s|B_{i,u}) \big)\, ds \Big| \mathbb{1}_{B_{i,u}} \leqslant c\varepsilon. \tag{2.6}$$

Moreover

$$\phi(s|B_{i,u}) - \phi(s,x) = \frac{\dot\theta_s' I(\theta_s) G(\theta_s)^{-1}}{\mathrm{P}_{\theta_s}(B_{i,u})} \int_{B_{i,u}} \big(g(X,\theta_s) - g(x,\theta_s) \big)\, d\mathrm{P}_{\theta_s}.$$

For x, $x' \in B_{i,u}$,

$$\begin{aligned}\|g(x,\theta_s) - g(x',\theta_s)\| &\leqslant \|g(x,\theta_{\lfloor sn\rfloor/n}) - g(x',\theta_{\lfloor sn\rfloor/n})\| \\ &\quad + \|g(x,\theta_{\lfloor sn\rfloor/n}) - g(x,\theta_s)\| \\ &\quad + \|g(x',\theta_{\lfloor sn\rfloor/n}) - g(x',\theta_s)\| \\ &\leqslant \operatorname{diam}(R_{u_{\lfloor sn\rfloor}}) + \frac{2M}{n} \leqslant 3\varepsilon.\end{aligned}$$

Hence

$$|\phi(s|B_{i,u}) - \phi(s,x)| \leqslant \sup_{s\in[0,1]} \|\dot\theta_s' I(\theta_s) G(\theta_s)^{-1}\| \times 3\varepsilon = 3c\varepsilon. \tag{2.7}$$

Hence

$$e^{-4c\varepsilon} \mathrm{P}_\theta(B) \leqslant \int_B f(X)\, d\mathrm{P}_{\theta_0} \leqslant e^{4c\varepsilon} \mathrm{P}_\theta(B),$$

for all $\varepsilon > 0$. And therefore $\int_B f(X)\, d\mathrm{P}_{\theta_0} = \mathrm{P}_\theta(B)$. Hence, for P_{θ_0}-almost all $x \in \mathcal{X}$,

$$\frac{d\mathrm{P}_\theta}{d\mathrm{P}_{\theta_0}}(x) = \exp\left(A(\theta)' p(x) - B(x,\theta)\right),$$

with

$$\begin{aligned}A(\theta)' &= \int_0^1 \dot\theta_s' I(\theta_s) G(\theta_s)^{-1}\, ds, \\ B(x,\theta) &= \int_0^1 \dot\theta_s' I(\theta_s) G(\theta_s)^{-1} g(x,\theta_s)\, ds.\end{aligned}$$

From condition 3 and the gradient theorem, $A(\theta)$ does not depend on $(\theta_s, s \in [0,1])$, the chosen path. Yet

$$\frac{d\mathrm{P}_\theta}{d\mathrm{P}_{\theta_0}}(x) = \log f(x) = \int_0^1 \left(\dot\theta_s' I(\theta_s) G(\theta_s)^{-1} p(x) - \phi(s,x)\right) ds,$$

does not depend on it either, hence $B(x,\theta)$ does not depend on it. Therefore

$$\nabla_\theta B(x,\theta) = I(\theta) G(\theta)^{-1} g(x,\theta) = (\mathrm{J}_\theta A(\theta))' g(x,\theta).$$

□

Remark 2.1. In Theorem 2.1 we did not assumed *continuous* L^2-differentiability as [9] did for their analogous result in the case of estimation. If we add a condition of continuous L^2-differentiability in Theorem 2.1, this makes possible to save some assumptions. More precisely, the result of Theorem 2.1 also holds under the following conditions.

1. The family $(\mathrm{P}_\theta, \theta \in \Theta)$ is continuously L^2-differentiable and Θ is a connected open set of $\mathbb{R}^d$.

2. The matrix $I(\theta)$ is invertible for all $\theta \in \Theta$.
3. $p(X)$ is an unbiased efficient predictor of $g(X, \theta)$.
4. For all θ, $\mathrm{E}_\theta \|p(X)\|^2 < \infty$.
5. For all $\theta \in \theta$, $G(\theta) = \mathrm{J}_\theta \mathrm{E}_\theta g(X, \theta) - \mathrm{E}_\theta g(X, \theta) \dot{L}'_\theta$ is invertible, or equivalently, $\mathrm{E}_\theta \left(p(X) - g(X, \theta)\right)^{\times 2}$ is invertible.
6. There exists $A : \Theta \to \mathbb{R}^k$ a differentiable function over Θ, such that $(\mathrm{J}_\theta A(\theta))' = I(\theta) G(\theta)^{-1}$, for all $\theta \in \Theta$.
7. $\mathcal{X}$ is a topological space and $(\mathcal{X}, \mathcal{B})$ is a σ-compact space.
8. For all compact set $C \subset \mathcal{X}$, $\tilde{C} \subset \Theta$, $\sup_{x \in C, \theta \in \tilde{C}} \|\mathrm{J}_\theta g(x, \theta)\| < \infty$.

Conditions to have $G(\theta) = \mathrm{E}_\theta \mathrm{J}_\theta g(X, \theta)$ are not fulfilled anymore, hence we only get the expression $G(\theta) = \mathrm{J}_\theta \mathrm{E}_\theta g(X, \theta) - \mathrm{E}_\theta g(X, \theta) \dot{L}'_\theta$. In the list of conditions above one saves conditions 3, 4 and 5 of Assumption 1.1 and condition 6 of Theorem 2.1.

Remark 2.2. The essential idea in the proof of Theorem 2.1 is to cut the set B with the family of subsets with the following form

$$B_{i,u} = B \cap p^{-1}(R_i) \cap S_u,$$

while for the result in the case of estimation, Müller-Funk *et al.* [9] took the family of subsets with the simpler form $B_i = B \cap p^{-1}(R_i)$. More specifically, we can see why our more precise partition of B is useful, in the case of prediction, in two instances. First when we prove that h is bounded, and then when we prove (2.1) in the case $\mathrm{P}_{\theta_0}(B) > 0$. In the first instance, for proving that h is bounded, we need to prove that $\phi(s|B_{i,u})$ (2.4) is bounded, which is done in (2.5) thanks to the bound on $\|g(x, \theta)\|$ established in (2.2). The derivation of the bound (2.2) crucially takes advantage of the set S_u. Contrast this with the special case of estimation, in which g does not depend on x but only on θ. A consequence, in that special case, is that $\phi(x, \theta)$ also only depends on θ and therefore $\phi(s|B_{i,u})$ reduces to $\phi(s) = \dot{\theta}'_s I(\theta_s) G(\theta_s)^{-1} g(s)$, which can be shown to be bounded on $[0, 1)$, without using the set S_u, by continuity arguments. In the second instance, for proving (2.1) in the case $\mathrm{P}_{\theta_0}(B) > 0$, we need to prove the bounds (2.6) and (2.7). For the bound (2.6) we do not use the property of the set S_u and the derivation of the bound is identical as in the case of estimation. For proving the bound (2.7) we rely crucially on the property of the set S_u. We remark that this bound becomes trivial in the special case of estimation since $\phi(s|B_{i,u}) = \phi(s, x) = \phi(s)$.

Remark 2.3. In the particular case where g does not depend on X, $g(X, \theta) = g(\theta)$, Theorem 2.1 gives the well-known result that the existence of an efficient unbiased estimator implies the family is exponential.

Acknowledgement. The author thanks the anonymous referee for many constructive remarks that lead to a great improvement of the original manuscript.

References

[1] Adke, S. R., Ramanathan, T. V., 1997. On optimal prediction for stochastic processes. J. Statist. Plann. Inference 63 (1), 1–7.

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

[3] Bosq, D., Onzon, E., 2012. Asymptotically efficient statistical predictors. Journal de la Société Francaise de Statistique 153 (1), 22–43.

[4] Fabian, V., Hannan, J., 1977. On the Cramér-Rao inequality. Ann. Statist. 5 (1), 197–205.

[5] Johansson, B., 1990. Unbiased prediction in the Poisson and Yule processes. Scand. J. Statist. 17 (2), 135–145.

[6] Lehmann, E. L., Romano, J. P. (2006). Testing statistical hypotheses. Springer Science & Business Media.

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

[8] Miyata, Y., 2001. The lower bound for MSE in statistical prediction theory. J. Japan Statist. Soc. 31 (1), 111–127.

[9] 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.

[10] Nayak, T. K., 2000. On best unbiased prediction and its relationships to unbiased estimation. J. Statist. Plann. Inference 84 (1-2), 171–189.

[11] Nayak, T. K., 2002. Rao-Cramer type inequalities for mean squared error of prediction. Amer. Statist. 56 (2), 102–106.

[12] Nayak, T. K., 2003. Finding optimal estimators in survey sampling using unbiased estimators of zero. J. Statist. Plann. Inference 114 (1-2), 21–30, c. R. Rao 80th birthday felicitation volume, Part IV.

[13] Nayak, T. K., Qin, M., 2010. The concept of risk unbiasedness in statistical prediction. J. Statist. Plann. Inference 140 (7), 1923–1938.

[14] Onzon, E., 2011. Multivariate Cramér-Rao inequality for prediction and efficient predictors. Statistics & Probability Letters 81 (3), 429–437.

[15] Onzon, E., 2012. Prédiction efficace et asymptotiquement efficace. Ph.D. thesis, Université Paris 6.

[16] Onzon, E., 2014. Asymptotically efficient prediction for lan families. Ann. ISUP 58 (1–2), 3–36.

[17] Wijsman, R. A., 1973. On the attainment of the Cramér-Rao lower bound. Ann. Statist. 1, 538–542.

[18] Yatracos, Y. G., 1992. On prediction and mean squared error. Canad. J. Statist. 20 (2), 187–200.

A. L^2-differentiable families

We remind some definitions and results about L^2-differentiable families of distributions, we refer to [7] p. 58 and next. For θ, θ_0 in Θ, any random variable $L_{\theta_0,\theta}$ taking values in $[0,+\infty]$ is called likelihood ratio of P_θ with respect to P_{θ_0} if, for all $A \in \mathcal{A}$,

$$P_\theta(A) = \int_A L_{\theta_0,\theta} dP_{\theta_0} + P_\theta\left(A \cap \{L_{\theta_0,\theta} = +\infty\}\right).$$

$L_{\theta_0,\theta}$ is a probability density of P_θ with respect to P_{θ_0} if and only if $P_\theta \ll P_{\theta_0}$. If ν is a measure over $\mathcal{A}$ that dominates $\{P_\theta, P_{\theta_0}\}$ with $\{f_\theta, f_{\theta_0}\}$ the corresponding densities then

$$L_{\theta_0,\theta} = \frac{f_\theta}{f_{\theta_0}} \mathbb{1}_{\{f_{\theta_0}>0\}} + \infty \mathbb{1}_{\{f_{\theta_0}=0, f_\theta>0\}}, \quad \{P_\theta, P_{\theta_0}\}\text{-a.s.}$$

For all $\theta \in \Theta$, for all $u \in \mathbb{R}^d$ such that $u + \theta \in \Theta$, we set

$$L_\theta(u) = L_{\theta,\theta+u}.$$

Definition A.1. The family $(P_\theta, \theta \in \Theta)$ is said L^2-*differentiable* at $\theta_0 \in \mathring{\Theta}$, if there is $U(\theta_0)$ a neighbourhood of θ_0, such that for all $\theta \in U(\theta_0)$, $P_\theta \ll P_{\theta_0}$, and if there is $\dot{L}_{\theta_0} \in L^2_{P_{\theta_0}}(\mathbb{R}^d)$, called the L^2-derivative of the model at θ_0, such that as $u \to 0$,

$$E_{\theta_0}\left(L_{\theta_0}^{1/2}(u) - 1 - \frac{1}{2} u' \dot{L}_{\theta_0}\right)^2 = o(\|u\|_{\mathbb{R}^d}).$$

The matrix $I(\theta_0) = E_{\theta_0} \dot{L}_{\theta_0} \dot{L}'_{\theta_0}$ is called the *Fisher information matrix* of the model at θ_0.

The following result is a recasting of Propositions 1.110 and 1.111 of Liese and Miescke (2008) [7].

Proposition A.1. *Let $(P_\theta, \theta \in \Theta)$ be a L^2-differentiable family at $\theta_0 \in \mathring{\Theta}$ with $\dot{L}_{\theta_0}$ the L^2-derivative and let δ a r.v. taking values in $\mathbb{R}^k$ such that there is a neighbourhood $U(\theta_0)$ of θ_0 with*

$$\sup_{\theta \in U(\theta_0)} E_\theta \|\delta\|^2_{\mathbb{R}^k} < \infty.$$

Then $\psi : \theta \mapsto E_\theta \delta$ is differentiable at θ_0, and the jacobian matrix of ψ is

$$J_\theta \psi(\theta_0) = E_{\theta_0}\left(\delta \dot{L}'_{\theta_0}\right).$$

In particular, $\theta \in \Theta$, $E_\theta \dot{L}_\theta = 0$.

We give the definition of *continuous* L^2-differentiability.

Definition A.2. Let $(P_\theta, \theta \in \Theta)$ be an L^2-differentiable family over Θ, with $\dot{L}_\theta$ as L^2-derivative. We say that $(P_\theta, \theta \in \Theta)$ is a *continuously* L^2-differentiable family over Θ if for all $\theta_0 \in \Theta$,

$$\lim_{\theta \to \theta_0} \|L_{\theta,\theta_0}^{1/2} \dot{L}_\theta - \dot{L}_{\theta_0}\|^2 = 0.$$

Algolux Inc.
Putzbrunner Strasse 71
81739 Munich
Allemagne
e-mail: emmanuel.onzon@algolux.com