Can we trust $\mathbb{L}_2$-criteria and $\mathbb{L}_2$-losses?

Yannick BARAUD

Université Côte d'Azur, CNRS, LJAD, France

April 24, 2019

In this thorough survey, Sylvain Arlot adresses the important problem of estimator selection in Statistics. Given a n-sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ with distribution $P_{s^\star}$ belonging to a parametrized statistical model $\mathscr{P} = \{P_s,\ s \in S\}$ and a loss function ℓ on S^2, the problem of estimator selection is that of finding from the data an estimator which minimizes the mapping $t \mapsto \ell(s^\star, t)$ among a family $\{\widehat{s}_m,\ m \in \mathcal{M}\} \subset S$ of candidate ones. The slope heuristics as well as other selection procedures aim at designing a selection rule, i.e. a mapping $\boldsymbol{X} \mapsto \widehat{m}(\boldsymbol{X})$ with values in $\mathcal{M}$, for which one can assure with a probability p close to 1 that the selected estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}(\boldsymbol{X})$ satisfies an inequality of the form

$$\ell(s^\star, \widetilde{s}) \leqslant K_n \inf_{m \in \mathcal{M}} \ell(s^\star, \widehat{s}_m) + R_n \tag{1}$$

where K_n is a constant that we wish to be as close as possible to 1 and R_n is an additional term that we wish to be as small as possible compared to $\inf_{m \in \mathcal{M}} \ell(s^\star, \widehat{s}_m)$. Inequality (1) is already interesting when the family $\{\widehat{s}_m,\ m \in \mathcal{M}\}$ reduces to two distinct and deterministic points $\{s_0, s_1\} \subset S$ in which case the selection rule corresponds to a test between s_0 and s_1. When $s^\star$ belongs to $\{s_0, s_1\}$ and the right-hand side of (1) is smaller than $\ell(s_0, s_1)$, this inequality tells us that the procedure selects the true parameter with a probability at least p. Our aim is to discuss the property of this test and the pieces of information that an inequality like (1) brings on the true distribution $P_{s^\star}$ of the data.

When the parameter set S is a subset of a Hilbert space $\mathscr{H} = \mathbb{L}_2(\mu)$ for some suitable measure μ, a convenient loss (from the point of view of the mathematical analysis) is that given by the Hilbert norm $\|\cdot\|$. I shall refer to $\|\cdot\|$ as a $\mathbb{L}_2$-loss and more specifically focus on the following typical examples.

Example 1 (The regression setting)**.** We observe a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ with values in $(\mathbb{R}^d)^n$ and mean of the form $(s, \ldots, s)$ or equivalently a n-sample $X_1, \ldots, X_n$ with values in $\mathbb{R}^d$ and mean $s = (s_1, \ldots, s_d) \in S$ where S is a subset of the Hilbert space $\mathscr{H} = \mathbb{R}^d$. The space $\mathscr{H}$ can also be seen as $\mathbb{L}_2(\mu)$ for μ being the counting measure on $\{1, \ldots, d\}$. Given a subgaussian distribution Q on $\mathbb{R}$ and $s \in S$, P_s denotes here the distribution of the vector $X = s + \varepsilon \in \mathbb{R}^d$ with $\varepsilon \sim Q^{\otimes d}$ (I shall refer to the Gaussian framework when Q is Gaussian). In this regression setting, statisticians often base their selection rule on the least squares that I shall see as a $\mathbb{L}_2$-criterion and define for $t \in \mathbb{R}^d$ by

$$\widehat{\gamma}_{1,n}(t) = -\frac{2}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d} t_j X_{i,j}\right) + \|t\|^2 \quad \text{with} \quad \|t\|^2 = \sum_{j=1}^{d} t_j^2. \tag{2}$$

More precisely, given s_0 and s_1 in S, the test based on the least squares selects s_0 when $\widehat{\gamma}_{1,n}(s_0) < \widehat{\gamma}_{1,n}(s_1)$ and s_1 when $\widehat{\gamma}_{1,n}(s_0) > \widehat{\gamma}_{1,n}(s_1)$, the choice between s_0 and s_1 being unimportant in case of equality.

Example 2 (Density estimation)**.** The parameter set S is here a family of densities which are squared integrable with respect to some dominating measure μ and for $s \in S$, $P_s = s \cdot \mu$. Taking $\mathscr{H} = \mathbb{L}_2(\mu)$ and using the well-known $\mathbb{L}_2$-criterion $\widehat{\gamma}_{2,n}$ defined for $t \in \mathbb{L}_2(\mu)$ by

$$\widehat{\gamma}_{2,n}(t) = -\frac{2}{n}\sum_{i=1}^{n} t(X_i) + \|t\|^2 \quad \text{with} \quad \|t\|^2 = \int t^2 d\mu, \tag{3}$$

one may test $s^\star = s_0$ versus $s^\star = s_1$ in the same way as we did in Example 1 with $\widehat{\gamma}_{2,n}$ in place of $\widehat{\gamma}_{1,n}$.

The similarities between the $\mathbb{L}_2$-criteria (2) and (3) are striking and the common Hilbertian structure that underlines both frameworks allows additionally to use very similar mathematical technics for analyzing their properties. Although quite similar in their mathematical forms, there exist, from a more statistical point of view, major differences between these two criteria and their ways of testing between s_0 and s_1. First of all, unlike the least squares in the Gaussian framework, the $\mathbb{L}_2$-criterion in density estimation does not provide a reliable test in general. A very simple counter-example is the following one. Assume that one observes a n-sample

$X_1',\dots,X_n'$ with values in a very large interval $[0,2\sqrt{a}]$ for some parameter a satisfying $\sqrt{a}>n=100$. For convenience, we change the unit of the data by considering the random variables $X_i=X_i'/\sqrt{a}$ for all $i\in\{1,\dots,n\}$ and take as a reference measure $\mu=\sqrt{a}\lambda$ where λ denotes the usual Lebesgue measure on $\mathbb{R}$. By doing so, the rescaled data X_i take their values in the interval $[0,2]$. We assume that their true density (with respect to μ) is $s^\star=s_0=1\!\!1_{[0,1/a]}+a^{-1/2}(1-a^{-1/2})1\!\!1_{(1/a,1/a+1]}$ (which means that the data actually lie in the smaller interval $[0,1+1/a]$) and we introduce the candidate density $s_1=(4a)^{-1/2}1\!\!1_{[0,2]}$ which is supported on the interval $[0,2]$. Note that both densities are bounded by 1. With a probability at least $(1-1/\sqrt{a})^n\geqslant 0.36$, no observation falls into the interval $[0,1/a]$ hence

$$\widehat{\gamma}_{2,n}(s_0)=(1/\sqrt{a})\left[-2(1-1/\sqrt{a})+1+(1-1/\sqrt{a})^2\right]=1/a^{3/2}$$

while

$$\widehat{\gamma}_{2,n}(s_1)=(1/\sqrt{a})\left[-2/2+1/2\right]=-1/(2\sqrt{a})<\widehat{\gamma}_{2,n}(s_0).$$

This means that the $\mathbb{L}_2$-criterion $\widehat{\gamma}_{2,n}$ fails to select the true density s_0 among $\{s_0,s_1\}$ even though the distribution P_{s_0} is quite different from P_{s_1}. Using the classical likelihood ratio test between s_0 and s_1 and the fact that the Hellinger affinity between the distributions P_{s_0} and P_{s_1} is $\rho(s_0,s_1)=\int\sqrt{s_0s_1}d\mu=(\sqrt{2})^{-1}(a^{-3/4}+\sqrt{1-a^{-1/2}})<0.71$, we would make no mistake except on a set of probability not larger than

$$\mathbb{P}\left[\prod_{i=1}^n s_1(X_i)\geqslant\prod_{i=1}^n s_0(X_i)\right]=\mathbb{P}\left[\prod_{i=1}^n\sqrt{\frac{s_1(X_i)}{s_0(X_i)}}\geqslant 1\right]\leqslant\rho^n(s_0,s_1)<0.71^n,$$

which is smaller than 5% when $n\geqslant 9$ and becomes completely negligible when n passes 100. It is actually well-known that in density estimation the $\mathbb{L}_2$-criterion and the $\mathbb{L}_2$-loss suffer from many drawbacks and we refer the reader to Lugosi & Devroye (2001) Section 6.5 and Birgé (2014) for a related discussion.

This weakness of the $\mathbb{L}_2$-criterion completely disappears in the Gaussian framework since choosing there between the parameters s_0 and s_1 by means of the least squares is equivalent to using the likelihood ratio test between P_{s_0} and P_{s_1} (which is of course optimal). The good property of the least squares criterion actually extends to the subgaussian case and there is no need for $s^\star$ to be an element of $\{s_0,s_1\}$ to make the correct choice since it can be proven that this criterion is likely to select the closest point to $s^\star$ (for the Euclidean distance) even in the case of a slight misspecification of this model.

However, the superiority of the $\mathbb{L}_2$-criterion in the Gaussian framework compared to the density one and our enthusiasm that may celebrate an inequality like (1) in the regression setting should be nuanced. The closeness of $s^\star$ and $\widetilde{s}$ with respect to the Euclidean distance may say very little on the closeness of $P_{s^\star}$ and $P_{\widetilde{s}}$. To see this, consider Example 1 with $S = \mathbb{R}^d$, $s^\star = (1/2, \ldots, 1/2)$, Q the uniform distribution in $[-1/2, 1/2]$ (so that the coordinates of the X_i are uniformly distributed on $[0,1]$) and the problem of selecting between $s_0 = s^\star + (a, \ldots, a)$ and $s_1 = s^\star + (0, \ldots, 0, b)$ with $a = 3/d$ and $b = 10^{-1}$. I shall assume here that both the dimension d and the number n of observations are large, say larger than 15 000 for the sake of convenience. Then the parameter s_0 is much closer to the truth than s_1 since

$$\|s^\star - s_0\|^2 = da^2 < 6.10^{-4} < 10^{-2} = b^2 = \|s^\star - s_1\|^2\,,$$

and applying Proposition 5 of Birgé (2006) (with $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathbb{R}^{nd}$, $y = 4\|s^\star - s_0\|^2 - \|s_0 - s_1\|^2/4$ and $z = 0$), we obtain that the least squares criterion selects s_0 (as expected) with a probability at least

$$1 - \exp\left[-\frac{3n}{100}\left(\|s_0 - s_1\|^2 - \frac{98}{25}y\right)\right] > 0.99.$$

However, the distribution P_{s_0} is quite far away from the true one since

$$\rho(P_{s^\star}, P_{s_0}) = (1-a)^d < e^{-ad} = e^{-3} \quad \text{hence} \quad h^2(P_{s^\star}, P_{s_0}) > 1 - e^{-3}$$

while P_{s_1} satisfies $h^2(P_{s^\star}, P_{s_1}) = b = 10^{-1}$ and is therefore much closer! If the reader does not like the Hellinger distance, he can alternatively consider the total variation distance $\|\cdot\|_{\mathrm{TV}}$ and check that $\|P_{s^\star} - P_{s_1}\|_{\mathrm{TV}} = b = 0.01$ while $1 \geqslant \|P_{s^\star} - P_{s_0}\|_{\mathrm{TV}} \geqslant 0.95$.

This discussion shows that the use of the $\mathbb{L}_2$-criterion (and the $\mathbb{L}_2$-distance) for estimating a density should be avoided in general and that, in the regression setting, an inequality like (1) (for the Euclidean loss) should be interpreted with caution. The closeness of the parameters does not translate in general to the closeness of the corresponding distributions. Comparing the behaviours of candidate estimators of $s^\star$ by comparing their Euclidean distances to it may lead to quite erroneous conclusions, at least when one is interested in the estimation of the true distribution of the data.

References

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325.

Birgé, L. (2014). Model selection for density estimation with L_2-loss. *Probab. Theory Related Fields*, 158(3-4):533–574.

Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation.* Springer Series in Statistics. Springer-Verlag, New York.