Abstracts

Data-driven penalties for linear estimators selection

Sylvain Arlot
(joint work with Francis Bach)

We consider the fixed-design regression framework, where one observes

$$Y = (Y_1, \dots, Y_n) = F + \varepsilon \in \mathbb{R}^n \quad ,$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d., with $\mathbb{E}[\varepsilon_1] = 0$ and $\mathbb{E}[\varepsilon_1^2] = \sigma^2$. The goal is to find from data some $t \in \mathbb{R}^n$ having a small least-squares loss

$$n^{-1} \| t - F \|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (t_i - F_i)^2 \quad .$$

We then tackle the problem of selecting among several linear estimators, *i.e.*, of the form

$$\widehat{F}_\lambda = A_\lambda Y \quad ,$$

where A_λ is a deterministic $n \times n$ matrix. This problem includes:

- model selection for linear regression,
- the choice of a regularization parameter in kernel ridge regression or spline smoothing,
- the choice of a kernel in multiple kernel learning,
- the choice of the number of neighbors (and of a distance in the feature space) for nearest-neighbor regression,
- the choice of a bandwidth (and of a kernel function) for Nadaraya-Watson estimators.

Given a family $(A_\lambda)_{\lambda \in \Lambda}$ of matrices, the goal is to choose some data-driven $\widehat{\lambda} \in \Lambda$ such that the corresponding estimator $\widehat{F}_{\widehat{\lambda}}$ has a quadratic risk $n^{-1} \mathbb{E} \| \widehat{F}_{\widehat{\lambda}} - F \|^2$ as small as possible. When $\operatorname{Card}(\Lambda) \leq K n^\alpha$ for some $K, \alpha \geq 0$, a well-known strategy is to follow the *unbiased risk estimation principle*, *i.e.*, to choose $\widehat{\lambda}$ by minimizing over $\lambda \in \Lambda$ an unbiased estimator of $n^{-1} \mathbb{E} \| \widehat{F}_\lambda - F \|_2^2$. In particular, penalization methods select

$$\widehat{\lambda} \in \arg\min_{\lambda \in \Lambda} \left\{ n^{-1} \| \widehat{F}_\lambda - Y \|_2^2 + \operatorname{pen}(\lambda) \right\} \quad , \tag{1}$$

where $\operatorname{pen} : \Lambda \to \mathbb{R}$ is called a penalty. Following the unbiased risk estimation principle, for every $\lambda \in \Lambda$, pen should be close to $n^{-1} \| \widehat{F}_\lambda - F \|_2^2 - n^{-1} \left\| \widehat{F}_\lambda - Y \right\|_2^2$.

Under mild conditions, concentration inequalities show that the risk $n^{-1}\|\widehat{F}_\lambda - F\|_2^2$ and the empirical risk $n^{-1}\|\widehat{F}_\lambda - Y\|_2^2$ both are close to their respective expectation. Therefore, the two key quantities in our problem are

$$\mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - F\|_2^2\right] = \frac{\|(A_\lambda - I_n)F\|_2^2}{n} + \frac{\mathrm{tr}(A_\lambda^\top A_\lambda)\sigma^2}{n} = \text{bias} + \text{variance} \ , \tag{2}$$

$$\mathbb{E}\left[n^{-1}\left\|\widehat{F}_\lambda - Y\right\|_2^2\right] = \frac{\|(A_\lambda - I_n)F\|_2^2}{n} - \frac{\left(2\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)\right)\sigma^2}{n} + \sigma^2 \ . \tag{3}$$

By (2), (3) and the unbiased risk estimation principle, an optimal penalty in (1) would be

$$\mathrm{pen}_{\mathrm{opt}}(\lambda) = \mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - F\|_2^2\right] - \mathbb{E}\left[n^{-1}\left\|\widehat{F}_\lambda - Y\right\|_2^2\right] - \sigma^2 = \frac{2\mathrm{tr}(A_\lambda)\sigma^2}{n} \ , \tag{4}$$

known as Mallows' C_L penalty [7]; its main drawback is its dependence on σ^2, usually unknown. Note that $\mathrm{tr}(A_\lambda)$ is often called *generalized degrees of freedom.*

We extend the notion of *minimal penalty* [4, 3] in order to define an estimator of σ^2 that could be plugged into (4) for designing a fully data-driven penalty. Indeed, let

$$\mathrm{pen}_{\mathrm{min}}(\lambda) = \frac{\left(2\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)\right)\sigma^2}{n}$$

$$\text{and} \quad \forall C > 0, \quad \widehat{\lambda}_{\min}(C) \in \arg\min_{\lambda\in\Lambda}\left\{n^{-1}\|\widehat{F}_\lambda - Y\|_2^2 + C\mathrm{pen}_{\mathrm{min}}(\lambda)\right\} \ .$$

By (3), up to concentration inequalities that are detailed in [1, 2], $\widehat{\lambda}_{\min}(C)$ behaves like a minimizer of

$$g_C(\lambda) = \mathbb{E}\left[\frac{\|\widehat{F}_\lambda - Y\|_2^2}{n} + C\mathrm{pen}_{\mathrm{min}}(\lambda)\right] - \sigma^2 = \frac{\|(A_\lambda - I_n)F\|_2^2}{n} + (C-1)\mathrm{pen}_{\mathrm{min}}(\lambda) \ .$$

Therefore, two main cases can be distinguished:

- if $C < 1$, then $g_C(\lambda)$ decreases with $\mathrm{tr}(A_\lambda)$ so that $\mathrm{tr}(A_{\widehat{\lambda}_{\min}(C)})$ is huge: $\widehat{\lambda}_{\min}(C)$ overfits.
- if $C > 1$, then $g_C(\lambda)$ increases with $\mathrm{tr}(A_\lambda)$ when $\mathrm{tr}(A_\lambda)$ is large enough, so that $\mathrm{tr}(A_{\widehat{\lambda}_{\min}(C)})$ is much smaller than when $C < 1$.

As a conclusion, $\mathrm{pen}_{\mathrm{min}}(\lambda)$ is the minimal amount of penalization needed so that a minimizer $\widehat{\lambda}$ of a penalized criterion is not clearly overfitting.

Since $\sigma^{-2}\mathrm{pen}_{\mathrm{min}}(\lambda)$ is known, we deduce the following algorithm:

Input: Λ a finite set with $\mathrm{Card}(\Lambda) \leq Kn^\alpha$ for some $K, \alpha \geq 0$, and matrices A_λ.

- $\forall C > 0$, compute $\widehat{\lambda}_0(C) = \widehat{\lambda}_{\min}(C\sigma^{-2}) \in \arg\min_{\lambda\in\Lambda}\{\|\widehat{F}_\lambda - Y\|_2^2 + C(2\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda))\}$.
- Find $\widehat{C}$ corresponding to the largest jump of $C \to \mathrm{tr}(A_{\widehat{\lambda}_0(C)})$.

Output: $\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda}\{\|\widehat{F}_\lambda - Y\|_2^2 + 2\widehat{C}\mathrm{tr}(A_\lambda)\}$.

We prove in [1, 2] that if the ε_i are Gaussian, under mild assumptions on the bias term $\|(A_\lambda - I_n)F\|_2^2$, then $|\sigma^{-2}\widehat{C} - 1| \leq \kappa\sqrt{\ln(n)}n^{-1/4}$ with large probability, for some constant $\kappa > 0$. Furthermore, we deduce that $\widehat{\lambda}$ satisfies an oracle inequality with leading constant $1 + \epsilon_n$ on an event of probability at least $1 - n^{-2}$.

Previous results on minimal penalties [4, 3, 6] considered the case of projection estimators, for which $\operatorname{tr}(A_\lambda^\top A_\lambda) = \operatorname{tr}(A_\lambda)$, so that the minimal penalty is exactly half the optimal penalty. Our result shows that for general linear estimators, the optimal and minimal penalties have different shapes, and their ratio

$$\frac{\operatorname{pen}_{\mathrm{opt}}(\lambda)}{\operatorname{pen}_{\min}(\lambda)} = \frac{2\operatorname{tr}(A_\lambda)}{2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)}$$

can take any value in $(1; 2]$.

Simulation experiments with kernel ridge regression and multiple kernel learning show that the proposed algorithm often improves significantly existing calibration procedures such as 10-fold cross-validation or generalized cross-validation [5], for moderate values of the sample size [1, 2].

References

[1] S. Arlot and F. Bach, *Data-driven calibration of linear estimators with minimal penalties*, In Advances in Neural Information Processing Systems **22** (2009).

[2] S. Arlot and F. Bach, *Data-driven calibration of linear estimators with minimal penalties*, arXiv:0909.1884v1 (2009).

[3] S. Arlot and P. Massart, *Data-driven calibration of penalties for least-squares regression*, Journal of Machine Learning Research **10** (2009), 245–279.

[4] L. Birgé and P. Massart, *Minimal penalties for Gaussian model selection*, Probability Theory and Related Fields, **138** (2007), 33–73.

[5] P. Craven and G. Wahba, *Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation*, Numerische Mathematik **31** (1978/79), 377–403.

[6] M. Lerasle, *Optimal model selection in density estimation*, http://hal.archives-ouvertes.fr/hal-00422655/en/ (2009).

[7] C. L. Mallows, *Some comments on C_p*, Technometrics **15** (1973), 661–675.

Reporter: