AGGREGATED HOLD-OUT FOR AGGREGATING SPARSE ESTIMATORS IN ROBUST REGRESSION

Guillaume Maillard
guillaume.maillard@u-psud.fr
Université Paris-Saclay

Hyperparameter selection for sparse estimators

The large dimensional setting gives rise to the additional need to consider "sparse" estimators, which only depend on a small number of covariates. How sparse should these estimators be? There is generally no a-priori answer to this question and the standard approach is to select this sparsity hyperparameter by cross-validation. However, in the Lasso case at least, existing theoretical results on cross-validation do not prove its optimality. Moreover, the idea of aggregating the Lasso has generated interest in the litterature [1, 4], with promising results. For this reason, aggregated hold-out, a general procedure that combines cross-validation with aggregation, seems like a good alternative. Theory and simulations show that it performs satisfactorily, and sometimes better than cross-validation. The following poster is based on my paper [3].

Problem

In robust regression, Lipschitz losses are preferred to the usual least-squares. A popular choice is Huber's loss.

Let $c > 0$. Huber's loss function is $\phi_c(u) = \frac{u^2}{2}\mathbb{I}_{|u|\leq c} + c\left(|u| - \frac{c}{2}\right)\mathbb{I}_{|u|>c}$.

From the statistical learning viewpoint, the goal of regression is to minimize the average error of a prediction $t(X)$ relative to Y, as measured by ϕ_c (in the Huber case). Therefore, the performance of a given linear predictor $x \mapsto q + \langle \theta, x\rangle$ is measured by its excess risk relative to the best possible prediction:

$$\ell_c(q,\theta) = \mathbb{E}[\phi_c(Y - q - \langle\theta, X\rangle)] - \inf_{t'} \mathbb{E}[\phi_c(Y - t'(X))].$$

Let now $x \mapsto \hat{q}_k + \langle \hat{\theta}_k, x\rangle$ denote a finite family of regression estimators. The goal is to construct a convex combination of these estimators that performs as well as the best element in the collection.

Agghoo

Hold-out, or simple validation, selects one of the estimators $(\hat{q}_k, \hat{\theta}_k)$ in the following manner.

Definition 0.2. *Let $D_n = (X_i, Y_i)_{1\leq i\leq n}$ be a dataset. For any $T \subset \{1, \dots, n\}$, denote $D_n^T = (X_i, Y_i)_{i\in T}$. Let then*

$$\hat{k}_T(D_n) = \min \operatorname*{argmin}_{1\leq k\leq K} \frac{1}{|T^c|}\sum_{i\notin T} \phi_c\left(Y_i - \hat{q}_k(D_n^T) - \langle \hat{\theta}_k(D_n^T), X_i\rangle\right).$$

Aggregated hold-out is, as its name suggests, an aggregate of several hold-out predictors T, obtained with different subsets T.

Definition 0.3. *Let $D_n = (X_i, Y_i)_{1\leq i\leq n}$ be a dataset. For any $V \in \mathbb{N}$, $n_t \in \{1 \dots n\}$ and $\tau = \frac{n_t}{n}$, let $T_1, \dots, T_V$ be uniform on $\{T \subset \{1 \dots n\}, |T| = (1-\tau)n.\}$. Let then:*

$$\hat{\theta}^{ag}_{\tau,V} = \frac{1}{V}\sum_{j=1}^{V} \hat{\theta}_{\hat{k}_{T_j}(D_n)}\big((X_i, Y_i)_{i\in T_j}\big)$$

$$\hat{q}^{ag}_{\tau,V} = \frac{1}{V}\sum_{j=1}^{V} \hat{q}_{\hat{k}_{T_j}(D_n)}\big((X_i, Y_i)_{i\in T_j}\big).$$

Assumptions

Sparsity played no role in the previous definitions: Agghoo is a generic method, just like cross-validation. However, the theoretical results deal specifically with the sparsity hyperparameter. We assume therefore that the parameter k controls the *sparsity* of $\hat{\theta}_k$ in the following sense:

$$\forall k \in \{1 \dots K\}, \left\|\hat{\theta}_k\right\|_0 = |\{i : \hat{\theta}_{k,i} \neq 0\}| \leq k \text{ a.s.}$$

Note that such a collection of estimators can be extracted from the LARS or Lasso regularization path, as suggested by Zou [6]. Two additional assumptions on $\hat{q}_k, \hat{\theta}_k$ are also needed.

1. There exist L, α such that $\forall n \in \mathbb{N}, \mathbb{E}\left[\sup_{1\leq k\leq n}\left\|\hat{\theta}_k(D_n)\right\|_1\right] \leq Ln^\alpha$.
2. Given $\hat{\theta}_k$, $\hat{q}_k$ is obtained through (unpenalized) empirical risk minimization. As ϕ_c isn't strongly convex, a special tie-breaking rule is needed (see [3] for more details).

The first assumption states that $\hat{\theta}_k$ is not too large compared to n, while the second corresponds with practice (with the possible exception of the tie-breaking rule).

Result

The main result can now be stated.

Theorem 0.4. *Let $D_n = (X_i, Y_i)_{1\leq i\leq n}$ be an i.i.d dataset and let $(X, Y) \sim (X_1, Y_1)$ be independent from D_n. Let $n_t \in [\![1; n-3]\!]$, $n_v = n - n_t$ and $\tau = \frac{n_t}{n}$.*
Assume that there exists $s \in \operatorname{argmin}_t \mathbb{E}\left[\phi_c(Y - t(X))\right]$ and $\eta > 0$ such that

$$\forall x, \mathbb{P}\left[|Y - s(x)| \leq \frac{c}{2}\,\middle|\, X = x\right] \geq \eta. \tag{1}$$

Let $\bar{X} = X - \mathbb{E}[X]$. Assume that $K \in \{1 \dots n_t\}$ and that for some $b_0 > 1$,

$$\sup_{\theta\neq 0, \|\theta\|_0\leq 2K} \frac{\left\|\langle \bar{X}, \theta\rangle\right\|_{L^\infty}}{\left\|\langle \bar{X}, \theta\rangle\right\|_{L^2}} \leq \frac{\eta}{8}\sqrt{\frac{n_v}{8b_0 \log n_t}}. \tag{2}$$

For any $T \subset \{1 \dots n\}$, let $\hat{q}_k^T = \hat{q}_k\big((X_i, Y_i)_{i\in T}\big), \hat{\theta}_k^T = \hat{\theta}_k\big((X_i, Y_i)_{i\in T}\big)$. Then for any $\theta \in \left[\frac{1}{\sqrt{b_0}}; 1\right]$,

$$(1-\theta)\mathbb{E}\left[\ell_c\big(\hat{q}^{ag}_{\tau,V}, \hat{\theta}^{ag}_{\tau,V}\big)\right] \leq (1+\theta)\mathbb{E}\left[\min_{1\leq k\leq K} \ell_c(\hat{q}_k^{T_1}, \hat{q}_k^{T_1})\right] + 24\theta b_0 \frac{c\log n_t}{\eta n_v}\left[c + \frac{cK}{n_t^{\theta^2 b_0}} + 16K\frac{L\left\|\bar{X}\right\|_{L^\infty}}{n_t^{\theta^2 b_0 - \alpha}}\right]. \tag{3}$$

Consider a sequence $X(n)$ of design matrices. If for any $b_0 > 1$, hypothesis (2) holds for all n large enough and if $\left\|\bar{X}(n)\right\|_{L^\infty}$ is at most polynomial in n, then the remainder term in equation (3) is always of order $\frac{\log n}{n}$. If this is negligible relative to the oracle (as in non-parametric minimax rates), equation (3) yields an asymptotically optimal oracle inequality. Note however that the oracle is based on estimators calculated on a restricted data-set of size $n_t = \tau n$. In most cases, this loses only a constant factor relative to the full dataset, and taking $\tau \to 1$ suffices to gain asymptotic equivalency to the full data oracle.

Hypothesis (2) is the key to the application of Theorem 3. We give two examples where it can be simplified. First, if the components of $X = X(n)$ are uniformly bounded, uncorrelated and of variance 1, then (assuming $\tau = 0.8$) hypothesis (2) amounts to $K \leq \delta\frac{n}{\log n}$ for some constant $\delta > 0$. This means that predictors should depend on less than $\frac{\delta n}{\log n}$ variables, which is reasonable in sparse regression.

By [5, Lemma 7], hypothesis (2) also holds if X is a dictionary of piecewise polynomial functions on a "regular" partition with less than $\frac{\delta \log n}{n}$ pieces. In this case, there are no constraints on K.

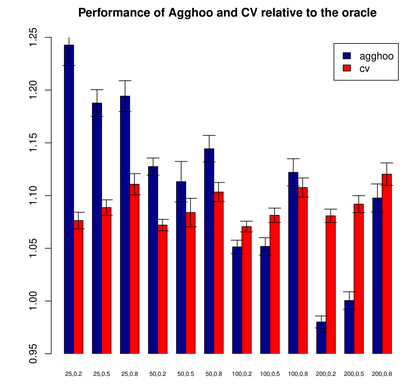
Simulation setup

A natural benchmark for Agghoo is the CV procedure with splits generated independently in the same manner as for Agghoo, known as monte-carlo CV: it has the same parameters τ, V as Agghoo. Previous work showed that $\tau = 0.8$, $V = 10$ were reasonable choices of Agghoo's parameters. On the other hand, the litterature on (v-fold) cross-validation suggests that $V = 5$ or $V = 10$ with $\tau = 0.8$ or $\tau = 0.9$ are good choices for model selection in practice [2]. Thus, we compared Agghoo and CV with the same parameters $\tau = 0.8$ and $V = 10$.
Both were applied to the huberized Lasso with Huber loss ϕ_2. Given independent standard gaussians $Z_0, (Z_i)_{1\leq i\leq r}, (W_i)_{1\leq i\leq 1000-r}$, the predictive covariates were generated as $X_i = \sqrt{\rho}Z_0 + \sqrt{1-\rho}Z_i$ for $1 \leq i \leq r$, together with noise covariates $X_i = W_{i-r}$ for $i > r$. The variable of interest was generated as

$$Y = 3\frac{\sum_{i=1}^r X_i}{\left\|\sum_{i=1}^r X_i\right\|_{L^2}} + \varepsilon$$

with $\varepsilon \sim$ Cauchy$(0, 0.3)$.

Simulation results



Shown above is the mean excess risk $\mathbb{E}[\ell_c(\cdot)]$ of Agghoo and CV, as a fraction of the mean excess risk of the oracle. For small values of r, CV performs better than Agghoo, while for large values of r, the situation is reversed. This seems due to the behaviour of Agghoo, since CV's performance seems to be unaffected by the increase in r. Lower correlations ρ also seem to favor Agghoo over CV. In conclusion, Agghoo seems to be the superior alternative in problems that aren't *too* sparse ($r \geq 100$ predictive covariates out of 1000).

References

[1] Francis Bach. "Bolasso: Model Consistent Lasso Estimation through the Bootstrap". In: *Proceedings of the 25th international conference on Machine learning* 33-40 (May 2008). DOI: 10.1145/1390156.1390161.

[2] Leo Breiman and Philip Spector. "Submodel Selection and Evaluation in Regression. The X-Random Case". In: *International Statistical Review / Revue Internationale de Statistique* 60.3 (1992), pp. 291–319. ISSN: 03067734, 17515823. URL: http://www.jstor.org/stable/1403680.

[3] Guillaume Maillard. "Aggregated hold out for sparse linear regression with a robust loss function". working paper or preprint. Feb. 2020. URL: https://hal.archives-ouvertes.fr/hal-02485694.

[4] Nicolai Meinshausen and Peter Bühlmann. "Stability Selection". In: *Journal of the Royal*