Aggregated Hold-Out

Guillaume Maillard

Université Paris Sud

February 13, 2020

Plan

- Presentation of Agghoo
- Theoretical performance
- Simulations in regression
- Classification

$\mathcal{X}, \mathcal{Y}$ are measurable spaces, P a distribution on $\mathcal{X} \times \mathcal{Y}$, and $\gamma : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a measurable *contrast function*. For ex:

- $\mathcal{Y} = \{-1; 1\}$ and $\gamma(u, y) = \mathbb{I}_{u \neq y}$ in classification
- $\mathcal{Y} = \mathbb{R}$ and $\gamma(u, y) = |y - u|$ in median regression.

Definition

- A *predictor* is a measurable function $t : \mathcal{X} \to \mathcal{Y}$.
- The *excess risk* of a predictor f is defined by

$$\ell(s, t) = \mathbb{E}\left[\gamma(f(X), Y)\right] - \inf_{g \text{ predictor}} \mathbb{E}\left[\gamma(g(X), Y)\right]$$

Fix an integer n. Denote

Definition

- $D_n = (X_i, Y_i)_{1 \leq i \leq n} \sim P^{\otimes n}$ a *sample* of size n.
- For a sample D_n and $T \subset \{1, \dots, n\}$, $D_n^T = (X_j, Y_j)_{j \in T}$.
- $\mathcal{A} : \cup_{k \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^k \to$ predictors a *learning rule*.

1.Presentation of Agghoo: hyperparameter selection

- Given learning rules $(\mathcal{A}_m)_{m\in\mathcal{M}}$ and a sample D_n, choose parameter m.
- Optimal choice m_* realizes $\inf_{m\in\mathcal{M}} \ell(s, \mathcal{A}_m(D_n))$.

Examples:

- $m = k$ in $k-$NN for classification.
- regularization parameter, $m = \lambda$ (Lasso, Ridge, SVM...)

Definition

Fix a sample $D_n = (X_i, Y_i)_{1 \leq i \leq n}$. Let $T \subset \{1, \dots, n\}$. For any learning rule $\mathcal{A}$,

$$\mathrm{HO}_T(\mathcal{A}) = \frac{1}{|T^c|} \sum_{j \notin T} \gamma\left(\mathcal{A}(D_n^T)(X_j), Y_j\right).$$

$1 \leq p \leq n$, $\mathcal{T} \subset \{T \subset \{1, ..., n\} : |T| = p\}$

$$CV_{\mathcal{T}}(\mathcal{A}) = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \mathrm{HO}_T(\mathcal{A})$$

Idea: estimate $\ell(s, \mathcal{A}_m(D))$ and optimize in m.

Definition

Let $\mathcal{M}$ be a set of hyperparameters.

- A *hold out predictor* is $\widehat{f}_T^{\mathrm{ho}} = \mathcal{A}_{\hat{m}_T^{ho}}(D_n^T)$ where $\hat{m}_T^{ho} = \mathrm{argmin}_{m \in \mathcal{M}} \mathrm{HO}_T(\mathcal{A}_m)$.
- A *CV predictor* is defined by $\widehat{f}_{\mathcal{T}}^{CV} = \mathcal{A}_{\hat{m}_{\mathcal{T}}^{CV}}(D_n)$ where $\hat{m}_{\mathcal{T}}^{CV} = \mathrm{argmin}_{m \in \mathcal{M}} CV_{\mathcal{T}}(\mathcal{A}_m)$

Idea: *aggregate* several hold out estimators.

Definition

Assume $\mathcal{Y}$ is convex (regression...).
$T_1, ..., T_V$ i.i.d $\sim \mathcal{U}(\{T \subset \{1, .., n\} : |T| = \lfloor \tau n \rfloor\})$.

$$\widehat{f}^{\mathrm{ag}}_{\tau,V} = \frac{1}{V} \sum_{i=1}^{V} \widehat{f}^{\mathrm{ho}}_{T_i}$$

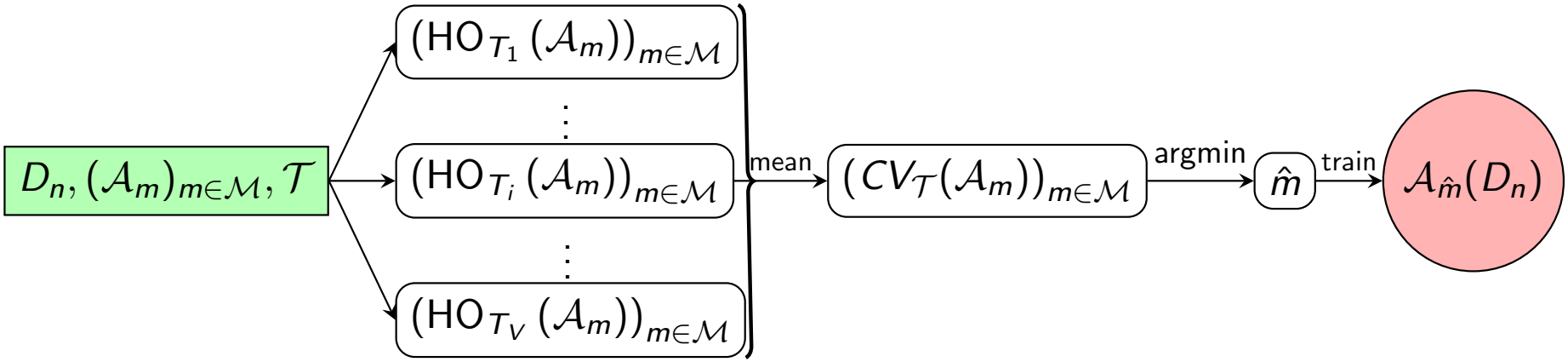
Parameters are V and τ.

V-fold CV $\Longrightarrow$ V-fold aggregation.

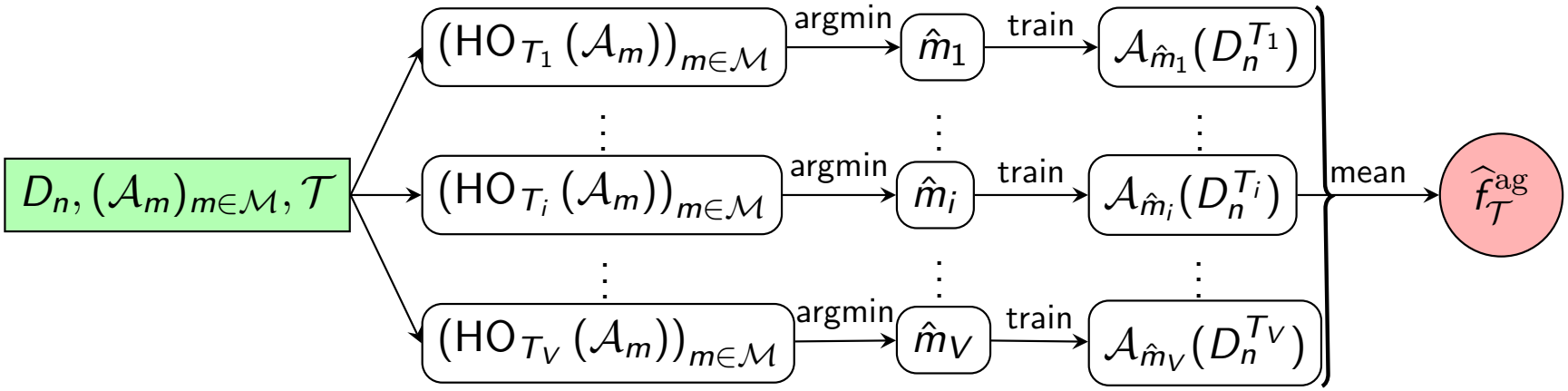
$CV_{\mathcal{T}} \Longrightarrow \widehat{f}^{\mathrm{ag}}_{\mathcal{T}}$.

1.Presentation of Agghoo: comparison with cross-validation

Cross-validation:



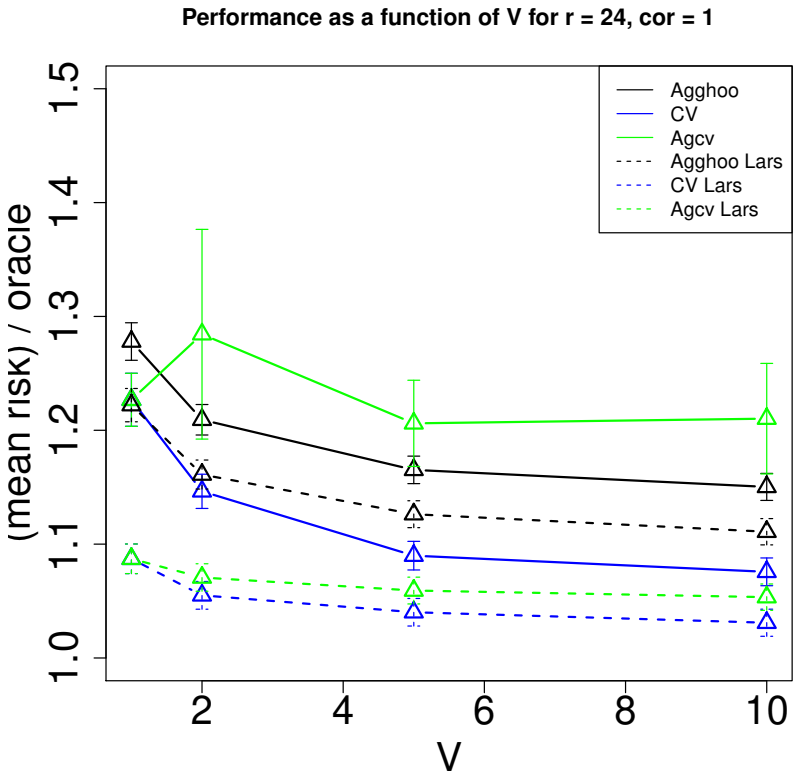
Agghoo:

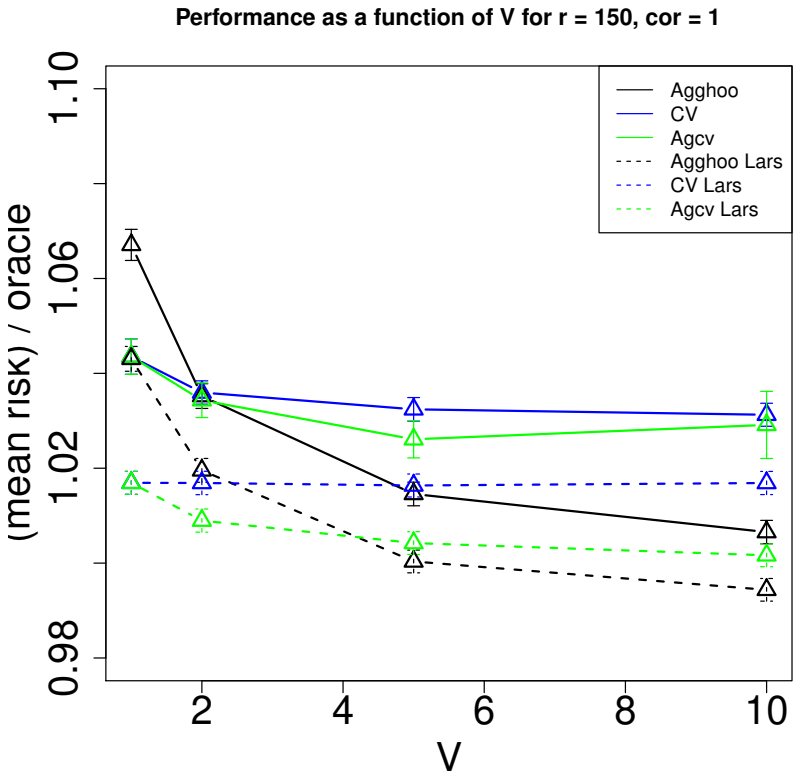


Simulation setting 1

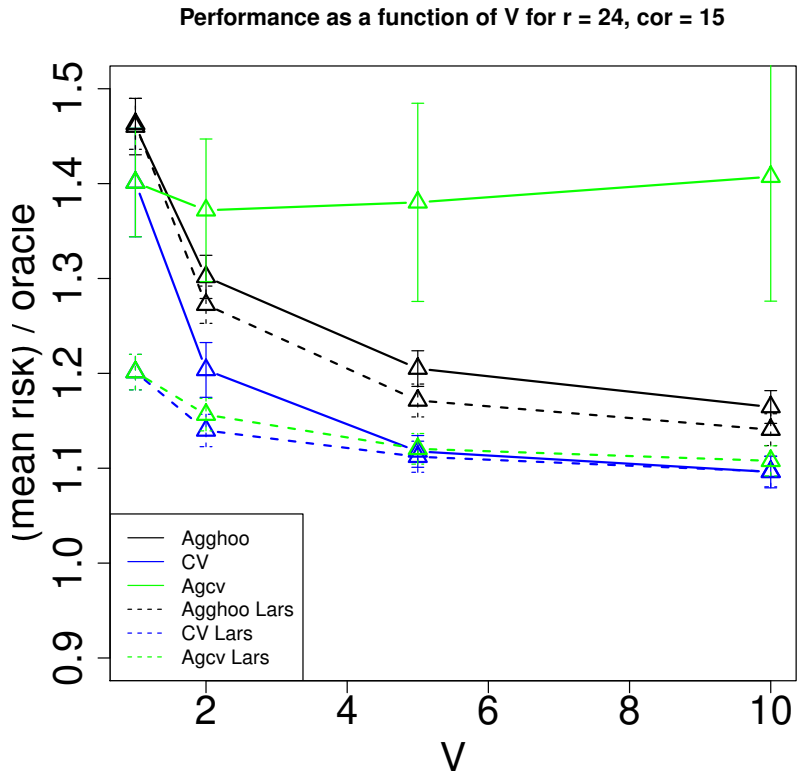
- $Y \sim \langle w_*, X \rangle + \varepsilon$ where $\varepsilon \sim \text{Cauchy}(0, 0.08)$
- $X_j = \sum_{i=1}^{1000} \mathbb{I}_{|i-j| \leq cor} Z_i e^{-2.33^2 \frac{(i-j)^2}{2cor^2}}$ where $Z =$ standard gaussian vector.
- $w_{*,j} = u_{*,g(j)}$ where $g =$ random permutation and $u_{*,j} = b$ for $1 \leq j \leq r$, $u_{*,j} = \frac{b}{4}$ for $r+1 \leq j \leq 3r$.
- b such that $\|\langle X, w_* \rangle\|_{L^2} = 1$.
- sample size $n = 100$

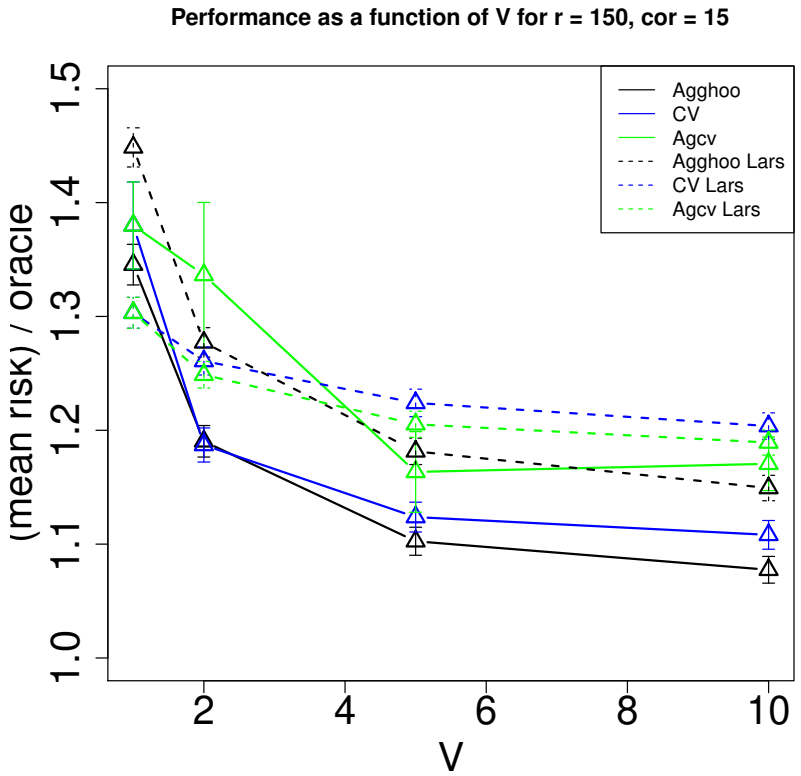
Agghoo vs CV for two parametrizations of the Lasso





Agghoo vs CV for two parametrizations of the Lasso

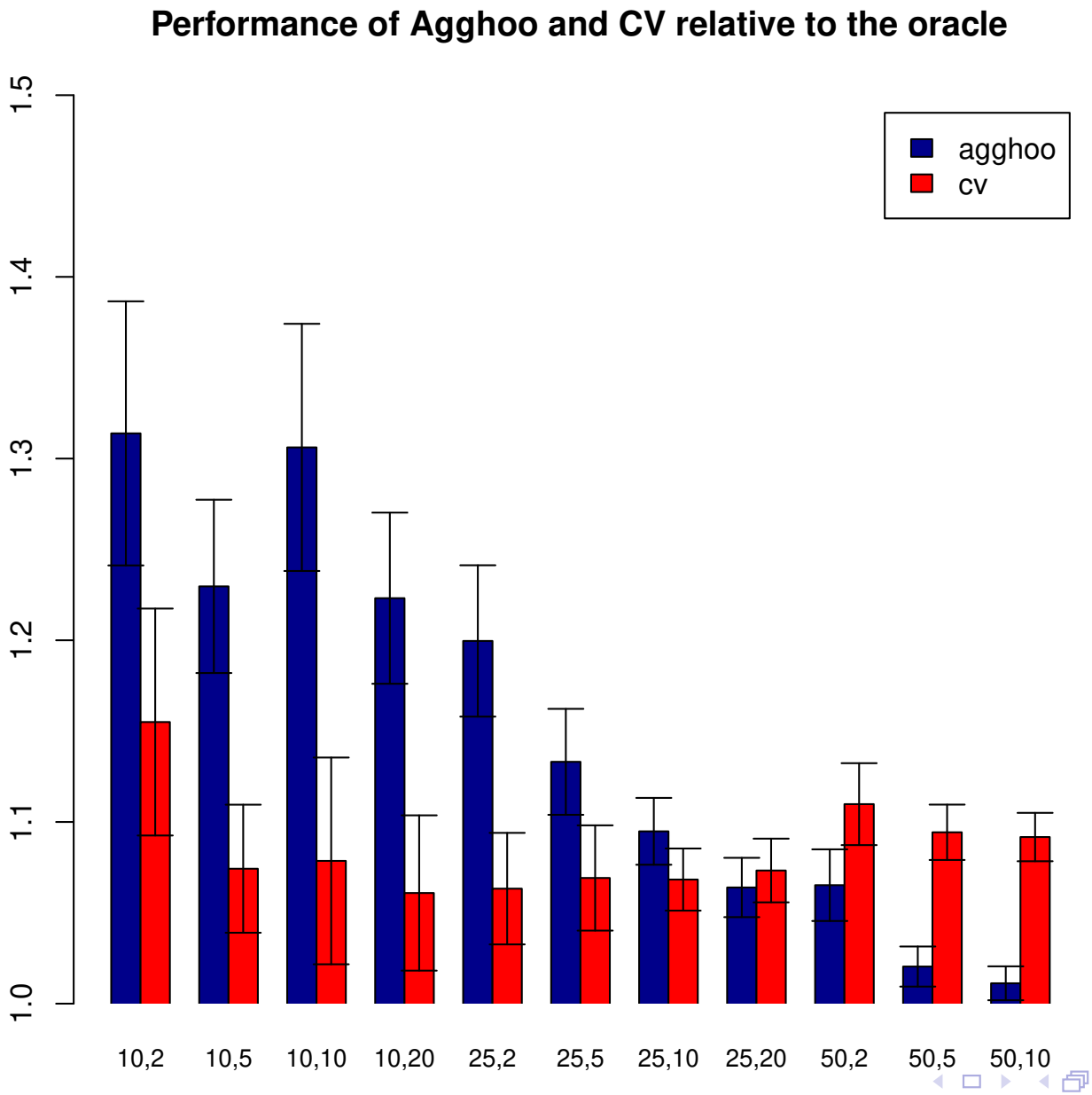




Simulation setting 2

- $Y \sim \langle w_*, X \rangle + \varepsilon$ where $\varepsilon \sim \text{Cauchy}(0, 0.3)$
- $X_{js+i} = \sqrt{0.8} Z^0_{j+1} + \sqrt{0.2} Z_{j,i}$ for $0 \leq j \leq r-1$, $1 \leq i \leq s$, $X_i = W_{i-r}$ for $rs < i \leq 1000$, where Z^0, Z, W independent standard gaussian.
- $w_{*,js} = b$ for $0 \leq j \leq r$, 0 otherwise.
- b such that $\|\langle X, w_* \rangle\|_{L^2} = 3$.
- sample size $n = 100$

Agghoo vs CV: impact of confounders



Simulation setting 3

- $Y \sim \langle w_*, X \rangle + \varepsilon$ where $\varepsilon \sim \text{Cauchy}(0, 0.3)$
- $X_j = \sqrt{\rho} Z^0 + \sqrt{1-\rho} Z_j$ for $1 \leq j \leq r$, $X_i = W_{i-r}$ for $r < i \leq 1000$, where Z^0, Z, W independent standard gaussian.
- $w_{*,j} = b$ for $0 \leq j \leq r$, 0 otherwise.
- b such that $\|\langle X, w_* \rangle\|_{L^2} = 3$.
- sample size $n = 100$

Agghoo vs CV: impact of correlations between predictive covariates