V-fold cross-validation improved: V-fold penalization

Sylvain Arlot

[1]University Paris-Sud XI, Orsay

[2]Projet Select, Inria-Futurs

Cherry Bud Workshop, Keio University, Yokohama
March 26, 2008

arXiv:0802.0566

Statistical framework: regression on a random design

$$(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y} \quad \text{i.i.d.} \qquad (X_i, Y_i) \sim P \text{ unknown}$$

$$Y = s(X) + \epsilon \qquad X \in \mathcal{X} \subset \mathbb{R}^d, \quad Y \in \mathcal{Y} = [0; 1] \text{ or } \mathbb{R}$$

$$\text{noise } \epsilon : \qquad \mathbb{E}\left[\epsilon | X\right] = 0 \qquad \text{noise level} \quad \mathbb{E}[\epsilon^2 | X] = \sigma^2(X)$$

predictor $t : \mathcal{X} \mapsto \mathcal{Y}$?

Loss function, least-squares estimator

- Least-squares risk:

$$P\gamma(t,\cdot) = \mathbb{E}\gamma(t,(X,Y)) \quad \text{with} \quad \gamma(t,(x,y)) = (t(x)-y)^2 \ .$$

- Loss function:

$$\ell(s,t) = P\gamma(t,\cdot) - P\gamma(s,\cdot) = \mathbb{E}\left[(t(X)-s(X))^2\right]$$

- Empirical risk minimizer on S_m (= model):

$$\widehat{s}_m \in \arg\min_{t \in S_m} P_n\gamma(t,\cdot) = \arg\min_{t \in S_m} \frac{1}{n}\sum_{i=1}^{n}(t(X_i)-Y_i)^2 \ .$$

- *e.g.* histograms on a partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of $\mathcal{X}$.

$$\widehat{s}_m = \sum_{\lambda \in \Lambda_m} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \qquad \widehat{\beta}_\lambda = \frac{1}{\operatorname{Card}\{X_i \in I_\lambda\}} \sum_{X_i \in I_\lambda} Y_i \ .$$

Model selection

$$(S_m)_{m \in \mathcal{M}} \longrightarrow (\widehat{s}_m)_{m \in \mathcal{M}} \longrightarrow \widehat{s}_{\widehat{m}} \quad ???$$

- "Classical" oracle inequality:

$$\mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq C \inf_{m \in \mathcal{M}} \left\{\mathbb{E}\left[\ell\left(s, \widehat{s}_m\right)\right] + R(m,n)\right\}$$

- "Pathwise" (or "conditional") oracle inequality:

$$\mathbb{P}\left(\ell\left(s, \widehat{s}_{\widehat{m}}\right) \leq C \inf_{m \in \mathcal{M}} \left\{\ell\left(s, \widehat{s}_m\right) + R(m,n)\right\}\right) \geq 1 - Kn^{-2}$$

- Adaptivity (*e.g.*, α if s is α-hölder, $\sigma(X)$ in the heteroscedastic framework)

Cross-validation

$$\underbrace{(X_1, Y_1), \ldots, (X_q, Y_q)}_{\text{Training}}, \underbrace{(X_{q+1}, Y_{q+1}), \ldots, (X_n, Y_n)}_{\text{Validation}}$$

$$\widehat{s}_m^{(t)} \in \arg\min_{t \in S_m} \left\{ \frac{1}{q} \sum_{i=1}^{q} \gamma(t, (X_i, Y_i)) \right\}$$

$$P_n^{(v)} = \frac{1}{n-q} \sum_{i=q+1}^{n} \delta_{(X_i, Y_i)} \qquad \Rightarrow P_n^{(v)} \gamma \left(\widehat{s}_m^{(t)} \right)$$

V-fold cross-validation: $(B_j)_{1 \leq j \leq V}$ partition of $\{1, \ldots, n\}$

$$\Rightarrow \widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ \frac{1}{V} \sum_{j=1}^{V} P_n^{(j)} \gamma \left(\widehat{s}_m^{(-j)} \right) \right\} \qquad \widetilde{s} = \widehat{s}_{\widehat{m}}$$

Bias of cross-validation

Ideal criterion: $P\gamma(\widehat{s}_m)$

Regression on an histogram model of dimension D_m, when $\sigma(X) \equiv \sigma$:

$$\mathbb{E}\left[P\gamma(\widehat{s}_m)\right] \approx P\gamma(s_m) + \frac{D_m \sigma^2}{n}$$

$$\mathbb{E}\left[P_n^{(j)} \gamma\left(\widehat{s}_m^{(-j)}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-j)}\right)\right] \approx P\gamma(s_m) + \frac{V}{V-1}\frac{D_m \sigma^2}{n}$$

$\Rightarrow$ bias if V is fixed

Suboptimality of V-fold cross-validation

- $Y = X + \sigma\epsilon$ with ϵ bounded and $\sigma > 0$
- $\mathcal{M}_n$: family of regular histograms on $\mathcal{X} = [0,1]$
- V fixed

Theorem

With probability at least $1 - \Diamond n^{-2}$,

$$\ell\left(s, \widehat{s}_{\widehat{m}}\right) \geq \left(1 + \kappa(V) - \ln(n)^{-1/5}\right) \inf_{m \in \mathcal{M}} \left\{\ell\left(s, \widehat{s}_m\right)\right\}$$

with $\kappa(V) > 0$.

Choice of V

- Bias: decreases with V (can be corrected: Burman 1989)
- Variance: large if V is small ($V = 2$), or sometimes when V is very large ($V = n$, unstable algorithms)
- Computation time: complexity proportional to V

$\Rightarrow$ trade-off

$\Rightarrow$ classical conclusion: "$V = 10$ is fine"

Simulation framework

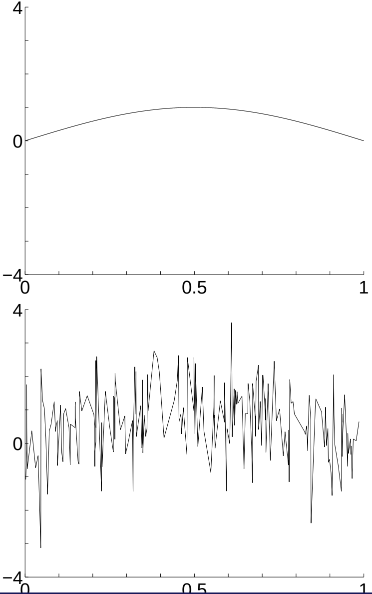
$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \qquad X_i \sim^{\text{i.i.d.}} \mathcal{U}([0;1]) \qquad \epsilon_i \sim^{\text{i.i.d.}} \mathcal{N}(0,1)$$

$$\mathcal{M}_n = \left\{ \text{regular histograms with } D \text{ pieces, } 1 \leq D \leq \frac{n}{\log(n)} \right. \\ \left. \text{and s.t. } \min_{\lambda \in \Lambda_m} \operatorname{Card}\{X_i \in I_\lambda\} \geq 2 \right\}$$

$\Rightarrow$ Benchmark:

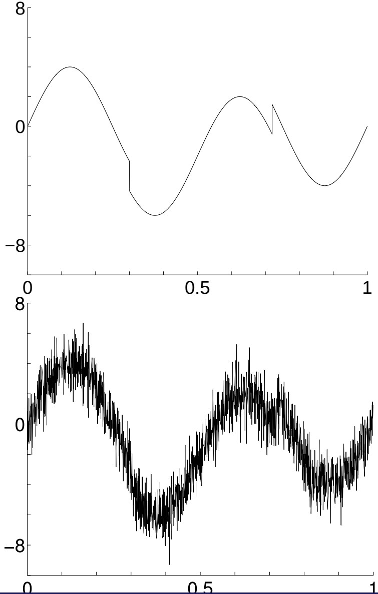
$$C_{\text{classical}} = \frac{\mathbb{E}[\ell(s, \widehat{s}_{\widehat{m}})]}{\mathbb{E}[\inf_{m \in \mathcal{M}} \ell(s, \widehat{s}_m)]} \qquad \text{computed with } N = 1000 \text{ samples}$$

Simulations: $s(x) = \sin(\pi x)$, $n = 200$, $\sigma \equiv 1$



2-fold	2.08 ± 0.04
5-fold	2.14 ± 0.04
10-fold	2.10 ± 0.05
20-fold	2.09 ± 0.04
leave-one-out	2.08 ± 0.04

Simulations: HeaviSine, $n = 2048$, $\sigma \equiv 1$



2-fold	1.002 ± 0.003
5-fold	1.014 ± 0.003
10-fold	1.021 ± 0.003
20-fold	1.029 ± 0.004
leave-one-out	1.034 ± 0.004

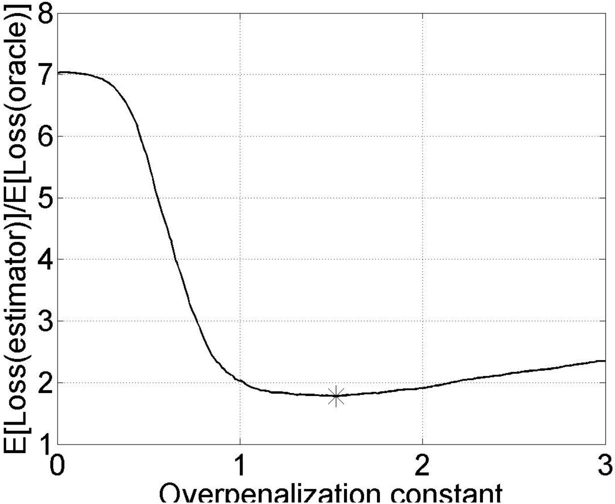
Overpenalization

- penalization: $\widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m) \right\}$
- ideal penalty: $\mathrm{pen}_{\mathrm{id}}(m) = P\gamma\left(\widehat{s}_m\right) - P_n\gamma\left(\widehat{s}_m\right)$
- V-fold cross-validation is overpenalizing:

$$\frac{\mathbb{E}\left[\frac{1}{V}\sum_{j=1}^{V} P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right) - P_n\gamma\left(\widehat{s}_m\right)\right]}{\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]} \approx 1 + \frac{1}{2(V-1)}$$

- non-asymptotic phenomenon:
  better to overpenalize when the signal-to-noise ratio n/σ^2 is small.

Overpenalization ($s = \sin$, $\sigma \equiv 1$, $n = 200$, Mallows' C_p)



Conclusions on V-fold cross-validation

- asymptotically suboptimal if V fixed
- optimal $V^\star$: trade-off variability–overpenalization
- $V^\star = 2$ can happen for prediction

- difficult to find $V^\star$ from the data (+ complexity issue)
- low signal-to-noise ratio $\Rightarrow$ $V^\star$ unsatisfactory (highly variable)
- large signal-to-noise ratio $\Rightarrow$ $V^\star$ too large (computation time)

Penalization

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) \right\}$$

Ideal penalty: $\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)(\gamma(\widehat{s}_m, \cdot))$

$$\mathrm{pen}(m) = \frac{2\sigma^2 D_m}{n} \qquad \text{(Mallows 1973)} \qquad \text{or} \qquad \frac{2\widehat{\sigma}^2 D_m}{n} \qquad \text{etc.}$$

Linear penalties may not work: heteroscedastic regression, classification, etc.

Resampling heuristics (bootstrap, Efron 1979)

Real world : $P \xrightarrow{\text{sampling}} P_n \Longrightarrow \widehat{s}_m$

$\rightsquigarrow$

Bootstrap world : $P_n \xrightarrow{\text{resampling}} P_n^W \Longrightarrow \widehat{s}_m^W$

$$(P - P_n)\gamma\left(\widehat{s}_m\right) = F(P, P_n) \rightsquigarrow F(P_n, P_n^W) = (P_n - P_n^W)\gamma\left(\widehat{s}_m^W\right)$$

V-fold: $\quad P_n^W = \dfrac{1}{n - \mathrm{Card}(B_J)} \sum_{i \notin B_J} \delta_{(X_i, Y_i)}$ with $J \sim \mathcal{U}(1, \ldots, V)$

Resampling heuristics (bootstrap, Efron 1979)

Real world : $P \xrightarrow{\text{sampling}} P_n \Longrightarrow \widehat{s}_m$

$\rightsquigarrow$

Bootstrap world : $P_n \xrightarrow{\text{subsampling}} P_n^W \Longrightarrow \widehat{s}_m^W$

$$(P - P_n)\gamma\left(\widehat{s}_m\right) = F(P, P_n) \rightsquigarrow F(P_n, P_n^W) = (P_n - P_n^W)\gamma\left(\widehat{s}_m^W\right)$$

V-fold: $$P_n^W = \frac{1}{n - \mathrm{Card}(B_J)} \sum_{i \notin B_J} \delta_{(X_i, Y_i)} \quad \text{with } J \sim \mathcal{U}(1, \ldots, V)$$

V-fold penalization

- Ideal penalty:

$$(P - P_n)(\gamma(\widehat{s}_m))$$

- V-fold penalty:

$$\operatorname{pen}(m) = \frac{C}{V} \sum_{j=1}^{V} \left[(P_n - P_n^{(-j)})(\gamma(\widehat{s}_m^{(-j)})) \right]$$

$$\widehat{s}_m^{(-j)} \in \arg\min_{t \in S_m} P_n^{(-j)} \gamma(t)$$

with $C \geq V - 1$ to be chosen
($C = V - 1 \Rightarrow$ we recover Burman's corrected V-fold, 1989)

- The final estimator is $\widehat{s}_{\widehat{m}}$ with

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \{ P_n \gamma(\widehat{s}_m) + \operatorname{pen}(m) \}$$

Model selection and resampling

- Hold-out, Cross-validation, Leave-one-out, V-fold cross-validation:
  $I \subset \{1, \ldots, n\}$ random sub-sample of size q (VFCV: $q = \frac{n(V-1)}{V}$).
- Efron's bootstrap penalties (Efron 1983, Shibata 1997):

$$\mathrm{pen}(m) = \mathbb{E}\left[(P_n - P_n^W)(\gamma(\widehat{s}_m^W)) \middle| (X_i, Y_i)_{1 \le i \le n} \right]$$

- Rademacher complexities (Koltchinskii 2001 ; Bartlett, Boucheron, Lugosi 2002): subsampling

$$\mathrm{pen}_{\mathrm{id}}(m) \le \mathrm{pen}_{\mathrm{id}}^{\mathrm{glo}}(m) = \sup_{t \in S_m} (P - P_n)\gamma(t, \cdot)$$

- idem with general exchangeable weights (Fromont 2004)
- Local Rademacher complexities (Bartlett, Bousquet, Mendelson 2004 ; Koltchinskii 2004)

Non-asymptotic pathwise oracle inequality

- $C \approx V - 1$
- Histogram regression on a random design
- Small number of models (at most polynomial in n)
- Model pre-selection: remove m when

$$\min_{\lambda \in \Lambda_m} \left\{ \mathrm{Card} \left\{ X_i \in I_\lambda \right\} \right\} \leq 1$$

- Fixed V or $V = n$

Theorem

Under a "reasonable" set of assumptions on P, with probability at least $1 - \Diamond n^{-2}$,

$$\ell \left(s, \widehat{s}_{\widehat{m}} \right) \leq \left(1 + \ln(n)^{-1/5} \right) \inf_{m \in \mathcal{M}} \left\{ \ell \left(s, \widehat{s}_m \right) \right\}$$

Sufficient assumptions

Reminder: the procedure *does not use* any of these assumptions.

- Bounded data: $\|Y\|_\infty \leq A < \infty$
- Minimal noise-level:

$$0 < \sigma_{\min} \leq \sigma(X)$$

- Smoothness of the regression function s: non-constant, belongs to some hölderian ball $\mathcal{H}_\alpha(R)$
- Regularity of the partition: $\min_\lambda \mathbb{P}(X \in I_\lambda) \geq \Diamond D_m^{-1}$

and they can be relaxed...

Corollaries

- Classical oracle inequality:

$$\mathbb{E}\left[\ell\left(s,\widehat{s}_{\widehat{m}}\right)\right] \leq \left(1+\ln(n)^{-1/5}\right)\mathbb{E}\left[\inf_{m\in\mathcal{M}}\left\{\ell\left(s,\widehat{s}_m\right)\right\}\right] + \Diamond n^{-2}$$

- Asymptotic optimality if $C \sim_{n\to+\infty} V-1$:

$$\frac{\ell\left(s,\widehat{s}_{\widehat{m}}\right)}{\inf_{m\in\mathcal{M}}\left\{\ell\left(s,\widehat{s}_m\right)\right\}} \xrightarrow[n\to+\infty]{a.s.} 1$$

- Adaptation to hölderian regularity in an heteroscedastic framework (regular histograms).

Simulation framework

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \qquad X_i \sim^{\text{i.i.d.}} \mathcal{U}([0;1]) \qquad \epsilon_i \sim^{\text{i.i.d.}} \mathcal{N}(0,1)$$

$$\mathcal{M}_n = \Big\{ \text{regular histograms with } D \text{ pieces, } 1 \leq D \leq \frac{n}{\log(n)} \text{ and s.t. } \min_{\lambda \in \Lambda_m} \operatorname{Card}\{X_i \in I_\lambda\} \geq 2 \Big\}$$

$\Rightarrow$ Benchmark:

$$C_{\text{classical}} = \frac{\mathbb{E}[\ell(s, \widehat{s}_{\widehat{m}})]}{\mathbb{E}[\inf_{m \in \mathcal{M}} \ell(s, \widehat{s}_m)]}$$

computed with $N = 1000$ samples

Model selection methods

- Mallows:

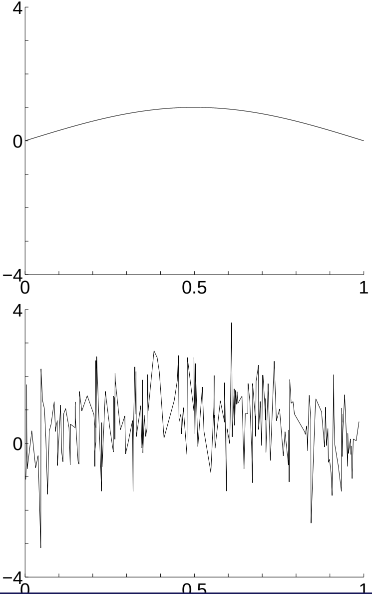
$$\operatorname{pen}(m) = 2 \widehat{\sigma}^2 D_m n^{-1}$$

- "Classical" V-fold cross-validation ($V \in \{2, 5, 10, 20, n\}$):

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ \frac{1}{V} \sum_{j=1}^{V} P_n^j \gamma \left(\widehat{s}_m^{(-j)}, \cdot \right) \right\} \qquad \widetilde{s} = \widehat{s}_{\widehat{m}}$$

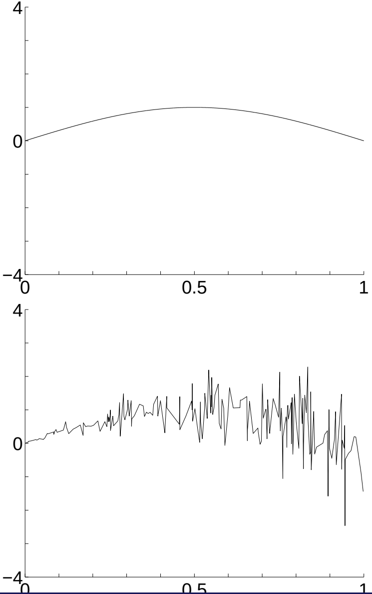
- V-fold penalties ($V \in \{2, 5, 10, n\}$), $C = V - 1$

Simulations: $s(x) = \sin(\pi x)$, $n = 200$, $\sigma \equiv 1$



Mallows	1.93 ± 0.04
2-fold	2.08 ± 0.04
5-fold	2.14 ± 0.04
10-fold	2.10 ± 0.05
20-fold	2.09 ± 0.04
leave-one-out	2.08 ± 0.04
pen 2-f	2.58 ± 0.06
pen 5-f	2.22 ± 0.05
pen 10-f	2.12 ± 0.05
pen Loo	2.08 ± 0.05
Mallows $\times 1.25$	1.80 ± 0.03
pen 2-f $\times 1.25$	2.17 ± 0.05
pen 5-f $\times 1.25$	1.91 ± 0.05
pen 10-f $\times 1.25$	1.87 ± 0.03
pen Loo $\times 1.25$	1.84 ± 0.03

Simulations: sin, $n = 200$, $\sigma(x) = x$, 2 bin sizes



Mallows	3.69 ± 0.07
2-fold	2.54 ± 0.05
5-fold	2.58 ± 0.06
10-fold	2.60 ± 0.06
20-fold	2.58 ± 0.06
leave-one-out	2.59 ± 0.06
pen 2-f	3.06 ± 0.07
pen 5-f	2.75 ± 0.06
pen 10-f	2.65 ± 0.06
pen Loo	2.59 ± 0.06
Mallows $\times 1.25$	3.17 ± 0.07
pen 2-f $\times 1.25$	2.75 ± 0.06
pen 5-f $\times 1.25$	2.38 ± 0.06
pen 10-f $\times 1.25$	2.28 ± 0.05
pen Loo $\times 1.25$	2.21 ± 0.05

Conclusions on V-fold penalization

- asymptotically optimal, even if V fixed
- optimal $V^{\star}$: the largest possible one
  $\Rightarrow$ easier to balance with the computational cost
- low signal-to-noise ratio $\Rightarrow$ easy to overpenalize and decrease variability (keep V large)
- large signal-to-noise ratio $\Rightarrow$ possible to stay unbiased with a small V (for computational reasons)

- flexibility improves V-fold cross-validation (according to both theoretical results and simulations)
- theory can be extended to exchangeable weighted bootstrap penalties (e.g. bootstrap, i.i.d. Rademacher, leave-one-out, leave-p-out with $p = \alpha n$).
- Open problems: consistency when $C \gg V - 1$, prediction in a general framework, etc.