A survey of cross-validation procedures for model selection

Sylvain Arlot, Alain Celisse

A survey of cross-validation procedures for model selection

July 27, 2009

Sylvain Arlot,
CNRS ; Willow Project-Team,
Laboratoire d'Informatique de l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
45, rue d'Ulm, 75 230 Paris, France
Sylvain.Arlot@ens.fr

Alain Celisse,
Laboratoire Paul Painlevé, UMR CNRS 8524,
Université des Sciences et Technologies de Lille 1
F-59 655 Villeneuve dAscq Cedex, France
Alain.Celisse@math.univ-lille1.fr

Abstract

Used to estimate the risk of an estimator or to perform model selection, cross-validation is a widespread strategy because of its simplicity and its apparent universality. Many results exist on the model selection performances of cross-validation procedures. This survey intends to relate these results to the most recent advances of model selection theory, with a particular emphasis on distinguishing empirical statements from rigorous theoretical results. As a conclusion, guidelines are provided for choosing the best cross-validation procedure according to the particular features of the problem in hand.

Contents

1 Introduction

Many statistical algorithms, such as likelihood maximization, least squares and empirical contrast minimization, rely on the preliminary choice of a model, that is of a set of parameters from which an estimate will be returned. When several candidate models (thus algorithms) are available, choosing one of them is called the model selection problem.

Cross-validation (CV) is a popular strategy for model selection, and more generally algorithm selection. The main idea behind CV is to split the data (once or several times) for estimating the risk of each algorithm: Part of the data (the training sample) is used for training each algorithm, and the remaining part (the validation sample) is used for estimating the risk of the algorithm. Then, CV selects the algorithm with the smallest estimated risk.

Compared to the resubstitution error, CV avoids overfitting because the training sample is independent from the validation sample (at least when data are *i.i.d.*). The popularity of CV mostly comes from the generality of the data splitting heuristics, which only assumes that data are *i.i.d.*. Nevertheless, theoretical and empirical studies of CV procedures do not entirely confirm this "universality". Some CV procedures have been proved to fail for some model selection problems, depending on the goal of model selection: estimation or identification (see Section 2). Furthermore, many theoretical questions about CV remain widely open.

The aim of the present survey is to provide a clear picture of what is known about CV, from both theoretical and empirical points of view. More precisely, the aim is to answer the following questions: What is CV doing? When does CV work for model selection, keeping in mind that model selection can target different goals? Which CV procedure should be used for each model selection problem?

The paper is organized as follows. First, the rest of Section 1 presents the statistical framework. Although non exhaustive, the present setting has been chosen general enough for sketching the complexity of CV for model selection. The model selection problem is introduced in Section 2. A brief overview of some model selection procedures that are important to keep in mind for understanding CV is given in Section 3. The most classical CV procedures are defined in Section 4. Since they are the keystone of the behaviour of CV for model selection, the main properties of CV estimators of the risk for a fixed model are detailed in Section 5. Then, the general performances of CV for model selection are described, when the goal is either estimation (Section 6) or identification (Section 7). Specific properties of CV in some particular frameworks are discussed in Section 8. Finally, Section 9 focuses on the algorithmic complexity of CV procedures, and Section 10 concludes the survey by tackling several practical questions about CV.

1.1 Statistical framework

Assume that some data $\xi_1, \ldots, \xi_n \in \Xi$ with common distribution P are observed. Throughout the paper—except in Section 8.3—the ξ_i are assumed to be independent. The purpose of statistical inference is to estimate from the data $(\xi_i)_{1 \leq i \leq n}$ some target feature s of the unknown distribution P, such as the mean or the variance of P. Let $\mathbb{S}$ denote the set of possible values for s.

The quality of $t \in \mathbb{S}$, as an approximation of s, is measured by its loss $\mathcal{L}(t)$, where $\mathcal{L} : \mathbb{S} \mapsto \mathbb{R}$ is called the *loss function*, and is assumed to be minimal for $t = s$. Many loss functions can be chosen for a given statistical problem.

Several classical loss functions are defined by

$$\mathcal{L}(t) = \mathcal{L}_P(t) := \mathbb{E}_{\xi \sim P}[\gamma(t;\xi)] \ , \tag{1}$$

where $\gamma : \mathbb{S} \times \Xi \mapsto [0, \infty)$ is called a *contrast function*. Basically, for $t \in \mathbb{S}$ and $\xi \in \Xi$, $\gamma(t;\xi)$ measures how well t is in accordance with observation of ξ, so that the loss of t, defined by (1), measures the average accordance between t and new observations ξ with distribution P. Therefore, several frameworks such as transductive learning do not fit definition (1). Nevertheless, as detailed in Section 1.2, definition (1) includes most classical statistical frameworks.

Another useful quantity is the *excess loss*

$$\ell(s,t) := \mathcal{L}_P(t) - \mathcal{L}_P(s) \geq 0 \ ,$$

which is related to the risk of an estimator $\widehat{s}$ of the target s by

$$R(\widehat{s}) = \mathbb{E}_{\xi_1,\ldots,\xi_n \sim P}[\ell(s, \widehat{s})] \ .$$

1.2 Examples

The purpose of this subsection is to show that the framework of Section 1.1 includes several important statistical frameworks. This list of examples does not pretend to be exhaustive.

Density estimation aims at estimating the density s of P with respect to some given measure μ on Ξ. Then, $\mathbb{S}$ is the set of densities on Ξ with respect to μ. For instance, taking $\gamma(t;x) = -\ln(t(x))$ in (1), the loss is minimal when $t = s$ and the excess loss

$$\ell(s,t) = \mathcal{L}_P(t) - \mathcal{L}_P(s) = \mathbb{E}_{\xi \sim P}\left[\ln\left(\frac{s(\xi)}{t(\xi)}\right)\right] = \int s \ln\left(\frac{s}{t}\right) \mathrm{d}\mu$$

is the Kullback-Leibler divergence between distributions $t\mu$ and $s\mu$.

Prediction aims at predicting a quantity of interest $Y \in \mathcal{Y}$ given an explanatory variable $X \in \mathcal{X}$ and a sample of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. In other words, $\Xi = \mathcal{X} \times \mathcal{Y}$, $\mathbb{S}$ is the set of measurable mappings $\mathcal{X} \mapsto \mathcal{Y}$ and the contrast $\gamma(t;(x,y))$ measures the discrepancy between the observed y and its predicted value $t(x)$. Two classical prediction frameworks are regression and classification, which are detailed below.

Regression corresponds to continuous $\mathcal{Y}$, that is $\mathcal{Y} \subset \mathbb{R}$ (or $\mathbb{R}^k$ for multivariate regression), the feature space $\mathcal{X}$ being typically a subset of $\mathbb{R}^\ell$. Let s denote the regression function, that is $s(x) = \mathbb{E}_{(X,Y)\sim P}[Y \mid X = x]$, so that

$$\forall i, \quad Y_i = s(X_i) + \epsilon_i \qquad \text{with} \quad \mathbb{E}[\epsilon_i \mid X_i] = 0 \ .$$

A popular contrast in regression is the *least-squares contrast* $\gamma(t;(x,y)) = (t(x) - y)^2$, which is minimal over $\mathbb{S}$ for $t = s$, and the excess loss is

$$\ell(s,t) = \mathbb{E}_{(X,Y)\sim P}\left[(s(X) - t(X))^2\right] \ .$$

Note that the excess loss of t is the square of the L^2 distance between t and s, so that prediction and estimation are equivalent goals.

Classification corresponds to finite $\mathcal{Y}$ (at least discrete). In particular, when $\mathcal{Y} = \{0, 1\}$, the prediction problem is called *binary (supervised) classification*. With the 0-1 contrast function $\gamma(t;(x,y)) = \mathbb{1}_{t(x)\neq y}$, the minimizer of the loss is the so-called Bayes classifier s defined by

$$s(x) = \mathbb{1}_{\eta(x)\geq 1/2} \ ,$$

where η denotes the regression function $\eta(x) = \mathbb{P}_{(X,Y)\sim P}(Y = 1 \mid X = x)$.

Remark that a slightly different framework is often considered in binary classification. Instead of looking only for a classifier, the goal is to estimate also the confidence in the classification made at each point: $\mathbb{S}$ is the set of measurable mappings $\mathcal{X} \mapsto \mathbb{R}$, the classifier $x \mapsto \mathbb{1}_{t(x)\geq 0}$ being associated to any $t \in \mathbb{S}$. Basically, the larger $|t(x)|$, the more confident we are in the classification made from $t(x)$. A classical family of losses associated with this problem is defined by (1) with the contrast $\gamma_\phi(t;(x,y)) = \phi(-(2y-1)t(x))$ where $\phi : \mathbb{R} \mapsto [0,\infty)$ is some function. The 0-1 contrast corresponds to $\phi(u) = \mathbb{1}_{u\geq 0}$. The convex loss functions correspond to the case where ϕ is convex, nondecreasing with $\lim_{-\infty}\phi = 0$ and $\phi(0) = 1$. Classical examples are $\phi(u) = \max\{1+u, 0\}$ (hinge), $\phi(u) = \exp(u)$, and $\phi(u) = \log_2(1+\exp(u))$ (logit). The corresponding losses are used as objective functions by several classical learning algorithms such as support vector machines (hinge) and boosting (exponential and logit).

Many references on classification theory, including model selection, can be found in the survey by Boucheron et al. (2005).

1.3 Statistical algorithms

In this survey, a *statistical algorithm* $\mathcal{A}$ is any (measurable) mapping $\mathcal{A} : \bigcup_{n\in\mathbb{N}} \Xi^n \mapsto S$. The idea is that data $D_n = (\xi_i)_{1\leq i\leq n} \in \Xi^n$ will be used as an input of $\mathcal{A}$, and that the output of $\mathcal{A}$, $\mathcal{A}(D_n) = \widehat{s}^{\mathcal{A}}(D_n) \in \mathbb{S}$, is an estimator of s. The quality of $\mathcal{A}$ is then measured by $\mathcal{L}_P\left(\widehat{s}^{\mathcal{A}}(D_n)\right)$, which should be as small as possible. In the sequel, the algorithm $\mathcal{A}$ and the estimator $\widehat{s}^{\mathcal{A}}(D_n)$ are often identified when no confusion is possible.

Minimum contrast estimators form a classical family of statistical algorithms, defined as follows. Given some subset S of $\mathbb{S}$ that we call a *model*, a minimum contrast estimator of s is any minimizer of the empirical contrast

$$t \mapsto \mathcal{L}_{P_n}(t) = \frac{1}{n}\sum_{i=1}^{n} \gamma(t;\xi_i), \qquad \text{where} \quad P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i} \ ,$$

over S. The idea is that the empirical contrast $\mathcal{L}_{P_n}(t)$ has an expectation $\mathcal{L}_P(t)$ which is minimal over $\mathbb{S}$ at s. Hence, minimizing $\mathcal{L}_{P_n}(t)$ over a set S of candidate values for s hopefully leads to a good estimator of s. Let us now give three popular examples of empirical contrast minimizers:

- *Maximum likelihood estimators*: take $\gamma(t;x) = -\ln(t(x))$ in the density estimation setting. A classical choice for S is the set of piecewise constant functions on a regular partition of Ξ with K pieces.

- *Least-squares estimators*: take $\gamma(t;(x,y)) = (t(x)-y)^2$ the least-squares contrast in the regression setting. For instance, S can be the set of piecewise constant functions on some fixed partition of $\mathcal{X}$ (leading to regressograms), or a vector space spanned by the first vectors of wavelets or Fourier basis, among many others. Note that regularized least-squares algorithms such as the Lasso, ridge regression and spline smoothing also are least-squares estimators, the model S being some ball of a (data-dependent) radius for the L^1 (resp. L^2) norm in some high-dimensional space. Hence, tuning the regularization parameter for the LASSO or SVM, for instance, amounts to perform model selection from a collection of models.

- *Empirical risk minimizers*, following the terminology of Vapnik (1982): take any contrast function γ in the prediction setting. When γ is the 0-1 contrast, popular choices for S lead to linear classifiers, partitioning rules, and neural networks. Boosting and Support Vector Machines classifiers also are empirical contrast minimizers over some data-dependent model S, with contrast $\gamma = \gamma_\phi$ for some convex functions ϕ.

Let us finally mention that many other classical statistical algorithms can be considered with CV, for instance local average estimators in the prediction framework such as k-Nearest Neighbours and Nadaraya-Watson kernel estimators. The focus will be mainly kept on minimum contrast estimators to keep the length of the survey reasonable.

2 Model selection

Usually, several statistical algorithms can be used for solving a given statistical problem. Let $(\widehat{s}_\lambda)_{\lambda\in\Lambda}$ denote such a family of candidate statistical algorithms. The *algorithm selection problem* aims at choosing from data one of these algorithms, that is, choosing some $\widehat{\lambda}(D_n) \in \Lambda$. Then, the final estimator of s is given by $\widehat{s}_{\widehat{\lambda}(D_n)}(D_n)$. The main difficulty is that the same data are used for training the algorithms, that is, for computing $(\widehat{s}_\lambda(D_n))_{\lambda\in\Lambda}$, and for choosing $\widehat{\lambda}(D_n)$.

2.1 The model selection paradigm

Following Section 1.3, let us focus on the *model selection problem*, where candidate algorithms are minimum contrast estimators and the goal is to choose a model S. Let $(S_m)_{m\in\mathcal{M}_n}$ be a family of models, that is, $S_m \subset \mathbb{S}$. Let γ be a fixed contrast function, and for every $m \in \mathcal{M}_n$, let $\widehat{s}_m$ be a minimum contrast estimator over model S_m with contrast γ. The goal is to choose $\widehat{m}(D_n) \in \mathcal{M}_n$ from data only.

The choice of a model S_m has to be done carefully. Indeed, when S_m is a "small" model, $\widehat{s}_m$ is a poor statistical algorithm except when s is very close to S_m, since

$$\ell(s,\widehat{s}_m) \geq \inf_{t\in S_m} \{\ell(s,t)\} := \ell(s,S_m) \quad .$$

The lower bound $\ell(s,S_m)$ is called the *bias* of model S_m, or *approximation error*. The bias is a nonincreasing function of S_m.

On the contrary, when S_m is "huge", its bias $\ell(s, S_m)$ is small for most targets s, but $\widehat{s}_m$ clearly overfits. Think for instance of S_m as the set of all continuous functions on $[0,1]$ in the regression framework. More generally, if S_m is a vector space of dimension D_m, in several classical frameworks,

$$\mathbb{E}[\ell(s, \widehat{s}_m(D_n))] \approx \ell(s, S_m) + \lambda D_m \tag{2}$$

where $\lambda > 0$ does not depend on m. For instance, $\lambda = 1/(2n)$ in density estimation using the likelihood contrast, and $\lambda = \sigma^2/n$ in regression using the least-squares contrast and assuming $\mathrm{var}(Y \mid X) = \sigma^2$ does not depend on X. The meaning of (2) is that a good model choice should balance the bias term $\ell(s, S_m)$ and the *variance* term λD_m, that is solve the so-called *bias-variance trade-off*. By extension, the variance term, also called *estimation error*, can be defined by

$$\mathbb{E}[\ell(s, \widehat{s}_m(D_n))] - \ell(s, S_m) = \mathbb{E}[\mathcal{L}_P(\widehat{s}_m)] - \inf_{t \in S_m} \mathcal{L}_P(t) \quad ,$$

even when (2) does not hold.

The interested reader can find a much deeper insight into model selection in the Saint-Flour lecture notes by Massart (2007).

Before giving examples of classical model selection procedures, let us mention the two main different goals that model selection can target: estimation and identification.

2.2 Model selection for estimation

On the one hand, the goal of model selection is *estimation* when $\widehat{s}_{\widehat{m}(D_n)}(D_n)$ is used as an approximation of the target s, and the goal is to minimize its loss. For instance, AIC and Mallows' C_p model selection procedures are built for estimation (see Section 3.1).

The quality of a model selection procedure $D_n \mapsto \widehat{m}(D_n)$, designed for estimation, is measured by the excess loss of $\widehat{s}_{\widehat{m}(D_n)}(D_n)$. Hence, the best possible model choice for estimation is the so-called *oracle* model $S_{m^\star}$, defined by

$$m^\star = m^\star(D_n) \in \arg\min_{m \in \mathcal{M}_n} \{\ell(s, \widehat{s}_m(D_n))\} \quad . \tag{3}$$

Since $m^\star(D_n)$ depends on the unknown distribution P of data, one cannot expect to select $\widehat{m}(D_n) = m^\star(D_n)$ almost surely. Nevertheless, we can hope to select $\widehat{m}(D_n)$ such that $\widehat{s}_{\widehat{m}(D_n)}$ is almost as close to s as $\widehat{s}_{m^\star(D_n)}$. Note that there is no requirement for s to belong to $\bigcup_{m \in \mathcal{M}_n} S_m$.

Depending on the framework, the optimality of a model selection procedure for estimation is assessed in at least two different ways.

First, in the asymptotic framework, a model selection procedure $\widehat{m}$ is called *efficient* (or asymptotically optimal) when it leads to $\widehat{m}$ such that

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

Sometimes, a weaker result is proved, the convergence holding only in probability.

Second, in the non-asymptotic framework, a model selection procedure satisfies an *oracle inequality* with constant $C_n \geq 1$ and remainder term $R_n \geq 0$ when

$$\ell\left(s, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right) \leq C_n \inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s, \widehat{s}_m(D_n)\right) \right\} + R_n \tag{4}$$

holds either in expectation or with large probability (that is, a probability larger than $1 - C'/n^2$, for some positive constant C'). Note that if (4) holds on a large probability event with C_n tending to 1 when n tends to infinity and $R_n \ll \ell\left(s, \widehat{s}_{m^\star}(D_n)\right)$, then the model selection procedure $\widehat{m}$ is efficient.

In the estimation setting, model selection is often used for building *adaptive estimators*, assuming that s belongs to some function space $\mathcal{T}_\alpha$ (Barron et al., 1999).Then, a model selection procedure $\widehat{m}$ is optimal when it leads to an estimator $\widehat{s}_{\widehat{m}(D_n)}(D_n)$ (approximately) minimax with respect to $\mathcal{T}_\alpha$ without knowing α, provided the family $(S_m)_{m \in \mathcal{M}_n}$ has been well-chosen.

2.3 Model selection for identification

On the other hand, model selection can aim at identifying the "true model" S_{m_0}, defined as the "smallest" model among $(S_m)_{m \in \mathcal{M}_n}$ to which s belongs. In particular, $s \in \bigcup_{m \in \mathcal{M}_n} S_m$ is assumed in this setting. A typical example of model selection procedure built for identification is BIC (see Section 3.3).

The quality of a model selection procedure designed for identification is measured by its probability of recovering the true model m_0. Then, a model selection procedure is called *(model) consistent* when

$$\mathbb{P}\left(\widehat{m}(D_n) = m_0\right) \xrightarrow[n \to \infty]{} 1 \ .$$

Note that identification can naturally be extended to the general algorithm selection problem, the "true model" being replaced by the statistical algorithm whose risk converges at the fastest rate (see for instance Yang, 2007).

2.4 Estimation *vs.* identification

When a true model exists, model consistency is clearly a stronger property than efficiency defined in Section 2.2. However, in many frameworks, no true model does exist so that efficiency is the only well-defined property.

Could a model selection procedure be model consistent in the former case (like BIC) and efficient in the latter case (like AIC)? The general answer to this question, often called the AIC-BIC dilemma, is negative: Yang (2005) proved in the regression framework that no model selection procedure can be simultaneously model consistent and minimax rate optimal. Nevertheless, the strengths of AIC and BIC can sometimes be shared; see for instance the introduction of a paper by Yang (2005) and a recent paper by van Erven et al. (2008).

3 Overview of some model selection procedures

Several approaches can be used for model selection. Let us briefly sketch here some of them, which are particularly helpful for understanding how CV works.

Like CV, all the procedures considered in this section select

$$\widehat{m}(D_n) \in \arg\min_{m \in \mathcal{M}_n} \{ \operatorname{crit}(m; D_n) \} \quad , \tag{5}$$

where $\forall m \in \mathcal{M}_n$, $\operatorname{crit}(m; D_n) = \operatorname{crit}(m) \in \mathbb{R}$ is some data-dependent criterion.

A particular case of (5) is *penalization*, which consists in choosing the model minimizing the sum of empirical contrast and some measure of complexity of the model (called penalty) which can depend on the data, that is,

$$\widehat{m}(D_n) \in \arg\min_{m \in \mathcal{M}_n} \{ \mathcal{L}_{P_n}(\widehat{s}_m) + \operatorname{pen}(m; D_n) \} \quad . \tag{6}$$

This section does not pretend to be exhaustive. Completely different approaches exist for model selection, such as the Minimum Description Length (MDL) (Rissanen, 1983), and the Bayesian approaches. The interested reader will find more details and references on model selection procedures in the books by Burnham and Anderson (2002) or Massart (2007) for instance.

Let us focus here on five main categories of model selection procedures, the first three ones coming from a classification made by Shao (1997) in the linear regression framework.

3.1 The unbiased risk estimation principle

When the goal of model selection is estimation, many model selection procedures are of the form (5) where $\operatorname{crit}(m; D_n)$ unbiasedly estimates (at least, asymptotically) the loss $\mathcal{L}_P(\widehat{s}_m)$. This general idea is often called unbiased risk estimation principle, or Mallows' or Akaike's heuristics.

In order to explain why this strategy can perform well, let us write the starting point of most theoretical analysis of procedures defined by (5): By definition (5), for every $m \in \mathcal{M}_n$,

$$\ell(s, \widehat{s}_{\widehat{m}}) + \operatorname{crit}(\widehat{m}) - \mathcal{L}_P(\widehat{s}_{\widehat{m}}) \leq \ell(s, \widehat{s}_m) + \operatorname{crit}(m) - \mathcal{L}_P(\widehat{s}_m) \quad . \tag{7}$$

If $\mathbb{E}[\operatorname{crit}(m) - \mathcal{L}_P(\widehat{s}_m)] = 0$ for every $m \in \mathcal{M}_n$, then concentration inequalities are likely to prove that $\varepsilon_n^-, \varepsilon_n^+ > 0$ exist such that

$$\forall m \in \mathcal{M}_n, \quad \varepsilon_n^+ \geq \frac{\operatorname{crit}(m) - \mathcal{L}_P(\widehat{s}_m)}{\ell(s, \widehat{s}_m)} \geq -\varepsilon_n^- > -1$$

with high probability, at least when $\operatorname{Card}(\mathcal{M}_n) \leq Cn^\alpha$ for some $C, \alpha \geq 0$. Then, (7) directly implies an oracle inequality like (4) with $C_n = (1+\varepsilon_n^+)/(1-\varepsilon_n^-)$. If $\varepsilon_n^+, \varepsilon_n^- \to 0$ when $n \to \infty$, this proves the procedure defined by (5) is efficient.

Examples of model selection procedures following the unbiased risk estimation principle are FPE (Final Prediction Error, Akaike, 1970), several cross-validation procedures including the Leave-one-out (see Section 4), and GCV (Generalized Cross-Validation, Craven and Wahba, 1979, see Section 4.3.3). With the penalization approach (6), the unbiased risk estimation principle is that $\mathbb{E}[\operatorname{pen}(m)]$ should be close to the "ideal penalty"

$$\operatorname{pen}_{\mathrm{id}}(m) := \mathcal{L}_P(\widehat{s}_m) - \mathcal{L}_{P_n}(\widehat{s}_m) \quad .$$

Several classical penalization procedures follow this principle, for instance:

- With the log-likelihood contrast, AIC (Akaike Information Criterion, Akaike, 1973) and its corrected versions (Sugiura, 1978; Hurvich and Tsai, 1989).
- With the least-squares contrast, Mallows' C_p (Mallows, 1973) and several refined versions of C_p (see for instance Baraud, 2002).
- With a general contrast, covariance penalties (Efron, 2004).

AIC, Mallows' C_p and related procedures have been proved to be optimal for estimation in several frameworks, provided $\mathrm{Card}(\mathcal{M}_n) \leq Cn^\alpha$ for some constants $C, \alpha \geq 0$ (see the paper by Birgé and Massart, 2007, and references therein).

The main drawback of penalties such as AIC or Mallows' C_p is their dependence on some assumptions on the distribution of data. For instance, Mallows' C_p assumes the variance of Y does not depend on X. Otherwise, it has a suboptimal performance (Arlot, 2008b).

Several resampling-based penalties have been proposed to overcome this problem, at the price of a larger computational complexity, and possibly slightly worse performance in simpler frameworks; see a paper by Efron (1983) for bootstrap, and a paper by Arlot (2008a) and references therein for generalization to exchangeable weights.

Finally, note that all these penalties depend on multiplying factors which are not always known (for instance, the noise-level, for Mallows' C_p). Birgé and Massart (2007) proposed a general data-driven procedure for estimating such multiplying factors, which satisfies an oracle inequality with $C_n \to 1$ in regression (see also Arlot and Massart, 2009).

3.2 Biased estimation of the risk

Several model selection procedures are of the form (5) where $\mathrm{crit}(m)$ does not unbiasedly estimate the loss $\mathcal{L}_P(\widehat{s}_m)$: The weight of the variance term compared to the bias in $\mathbb{E}[\mathrm{crit}(m)]$ is slightly larger than in the decomposition (2) of $\mathcal{L}_P(\widehat{s}_m)$. From the penalization point of view, such procedures are *overpenalizing*.

Examples of such procedures are FPE_α (Bhansali and Downham, 1977) and GIC_λ (Generalized Information Criterion, Nishii, 1984; Shao, 1997) with $\alpha, \lambda > 2$, which are closely related. Some cross-validation procedures, such as Leave-p-out with $p/n \in (0,1)$ fixed, also belong to this category (see Section 4.3.1). Note that FPE_α with $\alpha = 2$ is FPE, and GIC_λ with $\lambda = 2$ is close to FPE and Mallows' C_p.

When the goal is estimation, there are two main reasons for using "biased" model selection procedures. First, experimental evidence show that overpenalizing often yields better performance when the signal-to-noise ratio is small (see for instance Arlot, 2007, Chapter 11).

Second, when the number of models $\mathrm{Card}(\mathcal{M}_n)$ grows faster than any power of n, as in the complete variable selection problem with n variables, then the unbiased risk estimation principle fails. From the penalization point of view, Birgé and Massart (2007) proved that when $\mathrm{Card}(\mathcal{M}_n) = e^{\kappa n}$ for some $\kappa > 0$,

the minimal amount of penalty required so that an oracle inequality holds with $C_n = \mathcal{O}(1)$ is much larger than $\mathrm{pen}_{\mathrm{id}}(m)$. In addition to the FPE_α and GIC_λ with suitably chosen α, λ, several penalization procedures have been proposed for taking into account the size of $\mathcal{M}_n$ (Barron et al., 1999; Baraud, 2002; Birgé and Massart, 2001; Sauvé, 2009). In the same papers, these procedures are proved to satisfy oracle inequalities with C_n as small as possible, typically of order $\ln(n)$ when $\mathrm{Card}(\mathcal{M}_n) = e^{\kappa n}$.

3.3 Procedures built for identification

Some specific model selection procedures are used for identification. A typical example is BIC (Bayesian Information Criterion, Schwarz, 1978).

More generally, Shao (1997) showed that several procedures identify consistently the correct model in the linear regression framework as soon as they overpenalize within a factor tending to infinity with n, for instance, GIC_{λ_n} with $\lambda_n \to +\infty$, FPE_{α_n} with $\alpha_n \to +\infty$ (Shibata, 1984), and several CV procedures such as Leave-p-out with $p = p_n \sim n$. BIC is also part of this picture, since it coincides with $\mathrm{GIC}_{\ln(n)}$.

In another paper, Shao (1996) showed that m_n-out-of-n bootstrap penalization is also model consistent as soon as $m_n \sim n$. Compared to Efron's bootstrap penalties, the idea is to estimate $\mathrm{pen}_{\mathrm{id}}$ with the m_n-out-of-n bootstrap instead of the usual bootstrap, which results in overpenalization within a factor tending to infinity with n (Arlot, 2008a).

Most MDL-based procedures can also be put into this category of model selection procedures (see Grünwald, 2007). Let us finally mention the Lasso (Tibshirani, 1996) and other ℓ^1 penalization procedures, which have recently attracted much attention (see for instance Hesterberg et al., 2008). They are a computationally efficient way of identifying the true model in the context of variable selection with many variables.

3.4 Structural risk minimization

In the context of statistical learning, Vapnik and Chervonenkis (1974) proposed the structural risk minimization approach (see also Vapnik, 1982, 1998). Roughly, the idea is to penalize the empirical contrast with a penalty (over)-estimating

$$\mathrm{pen}_{\mathrm{id},g}(m) := \sup_{t \in S_m} \left\{ \mathcal{L}_P(t) - \mathcal{L}_{P_n}(t) \right\} \geq \mathrm{pen}_{\mathrm{id}}(m) \ .$$

Such penalties have been built using the Vapnik-Chervonenkis dimension, the combinatorial entropy, (global) Rademacher complexities (Koltchinskii, 2001; Bartlett et al., 2002), (global) bootstrap penalties (Fromont, 2007), Gaussian complexities or the maximal discrepancy (Bartlett and Mendelson, 2002). These penalties are often called *global* because $\mathrm{pen}_{\mathrm{id},g}(m)$ is a supremum over S_m.

The localization approach (see Boucheron et al., 2005) has been introduced in order to obtain penalties closer to $\mathrm{pen}_{\mathrm{id}}$ (such as local Rademacher complexities), hence smaller prediction errors when possible (Bartlett et al., 2005; Koltchinskii, 2006). Nevertheless, these penalties are still larger than $\mathrm{pen}_{\mathrm{id}}(m)$

and can be difficult to compute in practice because of several unknown constants.

A non-asymptotic analysis of several global and local penalties can be found in the book by Massart (2007) for instance; see also Koltchinskii (2006) for recent results on local penalties.

3.5 *Ad hoc* penalization

Let us finally mention that penalties can also be built according to particular features of the problem. For instance, penalties can be proportional to the ℓ^p norm of $\widehat{s}_m$ (similarly to ℓ^p-regularized learning algorithms) when having an estimator with a controlled ℓ^p norm seems better. The penalty can also be proportional to the squared norm of $\widehat{s}_m$ in some reproducing kernel Hilbert space (similarly to kernel ridge regression or spline smoothing), with a kernel adapted to the specific framework. More generally, any penalty can be used, as soon as $\mathrm{pen}(m)$ is larger than the estimation error (to avoid overfitting) and the best model for the final user is not the oracle $m^\star$, but more like

$$\arg\min_{m \in \mathcal{M}_n} \{ \ell(s, S_m) + \kappa \, \mathrm{pen}(m) \}$$

for some $\kappa > 0$.

3.6 Where are cross-validation procedures in this picture?

The family of CV procedures, which will be described and deeply investigated in the next sections, contains procedures in the first three categories. CV procedures are all of the form (5), where $\mathrm{crit}(m)$ either estimates (almost) unbiasedly the loss $\mathcal{L}_P(\widehat{s}_m)$, or overestimates the variance term (see Section 2.1). In the latter case, CV procedures either belong to the second or the third category, depending on the overestimation level.

This fact has two major implications. First, CV itself does not take into account prior information for selecting a model. To do so, one can either add to the CV estimate of the risk a penalty term (such as $\|\widehat{s}_m\|_p$), or use prior information to pre-select a subset of models $\widetilde{\mathcal{M}}(D_n) \subset \mathcal{M}_n$ before letting CV select a model among $(S_m)_{m \in \widetilde{\mathcal{M}}(D_n)}$.

Second, in statistical learning, CV and resampling-based procedures are the most widely used model selection procedures. Structural risk minimization is often too pessimistic, and other alternatives rely on unrealistic assumptions. But if CV and resampling-based procedures are the most likely to yield good prediction performances, their theoretical grounds are not that firm, and too few CV users are careful enough when choosing a CV procedure to perform model selection. Among the aims of this survey is to point out both positive and negative results about the model selection performance of CV.

4 Cross-validation procedures

The purpose of this section is to describe the rationale behind CV and to define the different CV procedures. Since all CV procedures are of the form (5), defining a CV procedure amounts to define the corresponding CV estimator of the risk of an algorithm $\mathcal{A}$, which will be $\mathrm{crit}(\cdot)$ in (5).

4.1 Cross-validation philosophy

As noticed in the early 30s by Larson (1931), training an algorithm and evaluating its statistical performance on the same data yields an overoptimistic result. CV was raised to fix this issue (Mosteller and Tukey, 1968; Stone, 1974; Geisser, 1975), starting from the remark that testing the output of the algorithm on new data would yield a good estimate of its performance (Breiman, 1998).

In most real applications, only a limited amount of data is available, which led to the idea of *splitting the data*: Part of the data (the training sample) is used for training the algorithm, and the remaining data (the validation sample) is used for evaluating its performance. The validation sample can play the role of new data as soon as data are *i.i.d.*.

Data splitting yields the *validation* estimate of the risk, and averaging over several splits yields a *cross-validation* estimate of the risk. As will be shown in Sections 4.2 and 4.3, various splitting strategies lead to various CV estimates of the risk.

The major interest of CV lies in the universality of the data splitting heuristics, which only assumes that data are identically distributed and the training and validation samples are independent, two assumptions which can even be relaxed (see Section 8.3). Therefore, CV can be applied to (almost) any algorithm in (almost) any framework, for instance regression (Stone, 1974; Geisser, 1975), density estimation (Rudemo, 1982; Stone, 1984) and classification (Devroye and Wagner, 1979; Bartlett et al., 2002), among many others. On the contrary, most other model selection procedures (see Section 3) are specific to a framework: For instance, C_p (Mallows, 1973) is specific to least-squares regression.

4.2 From validation to cross-validation

In this section, the hold-out (or validation) estimator of the risk is defined, leading to a general definition of CV.

4.2.1 Hold-out

The *hold-out* (Devroye and Wagner, 1979) or (simple) *validation* relies on a single split of data. Formally, let $I^{(t)}$ be a non-empty proper subset of $\{1,\dots,n\}$, that is, such that both $I^{(t)}$ and its complement $I^{(v)} = \left(I^{(t)}\right)^c = \{1,\dots,n\} \setminus I^{(t)}$ are non-empty. The *hold-out* estimator of the risk of $\mathcal{A}(D_n)$ with training set $I^{(t)}$ is defined by

$$\widehat{\mathcal{L}}^{\mathrm{H-O}}\left(\mathcal{A}; D_n; I^{(t)}\right) := \frac{1}{n_v} \sum_{i \in D_n^{(v)}} \gamma\left(\mathcal{A}(D_n^{(t)}); (X_i, Y_i)\right) \quad , \tag{8}$$

where $D_n^{(t)} := (\xi_i)_{i \in I^{(t)}}$ is the *training sample*, of size $n_t = \mathrm{Card}(I^{(t)})$, and $D_n^{(v)} := (\xi_i)_{i \in I^{(v)}}$ is the *validation sample*, of size $n_v = n - n_t$; $I^{(v)}$ is called the validation set. The question of choosing n_t, and $I^{(t)}$ given its cardinality n_t, is discussed in the rest of this survey.

4.2.2 General definition of cross-validation

A general description of the CV strategy has been given by Geisser (1975): In brief, CV consists in averaging several hold-out estimators of the risk corresponding to different splits of the data. Formally, let $B \geq 1$ be an integer and $I_1^{(t)}, \ldots, I_B^{(t)}$ be a sequence of non-empty proper subsets of $\{1, \ldots, n\}$. The CV estimator of the risk of $\mathcal{A}(D_n)$ with training sets $\left(I_j^{(t)}\right)_{1 \leq j \leq B}$ is defined by

$$\widehat{\mathcal{L}}^{\mathrm{CV}}\left(\mathcal{A}; D_n; \left(I_j^{(t)}\right)_{1 \leq j \leq B}\right) := \frac{1}{B} \sum_{j=1}^{B} \widehat{\mathcal{L}}^{\mathrm{H-O}}\left(\mathcal{A}; D_n; I_j^{(t)}\right) \quad . \tag{9}$$

All existing CV estimators of the risk are of the form (9), each one being uniquely determined by the way the sequence $\left(I_j^{(t)}\right)_{1 \leq j \leq B}$ is chosen, that is, the choice of the splitting scheme.

Note that when CV is used in model selection for identification, an alternative definition of CV was proposed by Yang (2006, 2007) and called *CV with voting* (CV-v). When two algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ are compared, $\mathcal{A}_1$ is selected by CV-v if and only if $\widehat{\mathcal{L}}^{\mathrm{H-O}}(\mathcal{A}_1; D_n; I_j^{(t)}) < \widehat{\mathcal{L}}^{\mathrm{H-O}}(\mathcal{A}_2; D_n; I_j^{(t)})$ for a majority of the splits $j = 1, \ldots, B$. By contrast, CV procedures of the form (9) can be called "CV with averaging" (CV-a), since the estimates of the risk of the algorithms are averaged before their comparison.

4.3 Classical examples

Most classical CV estimators split the data with a fixed size n_t of the training set, that is, $\mathrm{Card}(I_j^{(t)}) \approx n_t$ for every j. The question of choosing n_t is discussed extensively in the rest of this survey. In this subsection, several CV estimators are defined. Two main categories of splitting schemes can be distinguished, given n_t: exhaustive data splitting, that is considering all training sets $I^{(t)}$ of size n_t, and partial data splitting.

4.3.1 Exhaustive data splitting

Leave-one-out (LOO, Stone, 1974; Allen, 1974; Geisser, 1975) is the most classical exhaustive CV procedure, corresponding to the choice $n_t = n-1$: Each data point is successively "left out" from the sample and used for validation. Formally, LOO is defined by (9) with $B = n$ and $I_j^{(t)} = \{j\}^c$ for $j = 1, \ldots, n$:

$$\widehat{\mathcal{L}}^{\mathrm{LOO}}\left(\mathcal{A}; D_n\right) = \frac{1}{n} \sum_{j=1}^{n} \gamma\left(\mathcal{A}\left(D_n^{(-j)}\right); \xi_j\right) \tag{10}$$

where $D_n^{(-j)} = (\xi_i)_{i \neq j}$. The name LOO can be traced back to papers by Picard and Cook (1984) and by Breiman and Spector (1992), but LOO has several other names in the literature, such as *delete-one CV* (see Li, 1987), *ordinary CV* (Stone, 1974; Burman, 1989), or even only *CV* (Efron, 1983; Li, 1987).

Leave-p-out (LPO, Shao, 1993) with $p \in \{1, \ldots, n\}$ is the exhaustive CV with $n_t = n - p$: every possible set of p data points are successively "left out" from the sample and used for validation. Therefore, LPO is defined by (9) with $B = \binom{n}{p}$ and $(I_j^{(t)})_{1 \leq j \leq B}$ are all the subsets of $\{1, \ldots, n\}$ of size p. LPO is also called *delete-p CV* or *delete-p multifold CV* (Zhang, 1993). Note that LPO with $p = 1$ is LOO.

4.3.2 Partial data splitting

Considering $\binom{n}{p}$ training sets can be computationally intractable, even for small p, so that partial data splitting methods have been proposed.

V-fold CV (VFCV) with $V \in \{1, \ldots, n\}$ was introduced by Geisser (1975) as an alternative to the computationally expensive LOO (see also Breiman et al., 1984, for instance). VFCV relies on a preliminary partitioning of the data into V subsamples of approximately equal cardinality n/V; each of these subsamples successively plays the role of validation sample. Formally, let $A_1, \ldots, A_V$ be some partition of $\{1, \ldots, n\}$ with $\mathrm{Card}(A_j) \approx n/V$. Then, the VFCV estimator of the risk of $\mathcal{A}$ is defined by (9) with $B = V$ and $I_j^{(t)} = A_j^c$ for $j = 1, \ldots, B$, that is,

$$\widehat{\mathcal{L}}^{\mathrm{VF}}\left(\widehat{s}; D_n; (A_j)_{1 \leq j \leq V}\right) = \frac{1}{V} \sum_{j=1}^{V} \left[\frac{1}{\mathrm{Card}(A_j)} \sum_{i \in A_j} \gamma\left(\widehat{s}\left(D_n^{(-A_j)}\right); \xi_i\right) \right] \quad (11)$$

where $D_n^{(-A_j)} = (\xi_i)_{i \in A_j^c}$. By construction, the algorithmic complexity of VFCV is only V times that of training $\mathcal{A}$ with $n - n/V$ data points, which is much less than LOO or LPO if $V \ll n$. Note that VFCV with $V = n$ is LOO.

Balanced Incomplete CV (BICV, Shao, 1993) can be seen as an alternative to VFCV well-suited for small training sample sizes n_t. Indeed, BICV is defined by (9) with training sets $(A^c)_{A \in \mathcal{T}}$, where $\mathcal{T}$ is a balanced incomplete block designs (BIBD, John, 1971), that is, a collection of $B > 0$ subsets of $\{1, \ldots, n\}$ of size $n_v = n - n_t$ such that:

1. $\mathrm{Card}\{A \in \mathcal{T} \text{ s.t. } k \in \mathcal{A}\}$ does not depend on $k \in \{1, \ldots, n\}$.
2. $\mathrm{Card}\{A \in \mathcal{T} \text{ s.t. } k, \ell \in \mathcal{A}\}$ does not depend on $k \neq \ell \in \{1, \ldots, n\}$.

The idea of BICV is to give to each data point (and each pair of data points) the same role in the training and validation tasks. Note that VFCV relies on a similar idea, since the set of training sample indices used by VFCV satisfy the first property and almost the second one: Pairs (k, ℓ) belonging to the same A_j appear in one validation set more than other pairs.

Repeated learning-testing (RLT) was introduced by Breiman et al. (1984) and further studied by Burman (1989) and by Zhang (1993) for instance. The RLT estimator of the risk of $\mathcal{A}$ is defined by (9) with any $B > 0$ and $(I_j^{(t)})_{1 \leq j \leq B}$ are B different subsets of $\{1, \ldots, n\}$, chosen randomly and independently from the data. RLT can be seen as an approximation to LPO with $p = n - n_t$, with which it coincides when $B = \binom{n}{p}$.

Monte-Carlo CV (MCCV, Picard and Cook, 1984) is very close to RLT: B independent subsets of $\{1, \dots, n\}$ are randomly drawn, with uniform distribution among subsets of size n_t. The only difference with RLT is that MCCV allows the same split to be chosen several times.

4.3.3 Other cross-validation-like risk estimators

Several procedures have been introduced which are close to, or based on CV. Most of them aim at fixing an observed drawback of CV.

Bias-corrected versions of VFCV and RLT risk estimators have been proposed by Burman (1989, 1990), and a closely related penalization procedure called V-fold penalization has been defined by Arlot (2008c), see Section 5.1.2 for details.

Generalized CV (GCV, Craven and Wahba, 1979) was introduced as a rotation-invariant version of LOO in least-squares regression, for estimating the risk of a linear estimator $\widehat{s} = M\mathbf{Y}$ where $\mathbf{Y} = (Y_i)_{1 \leq i \leq n} \in \mathbb{R}^n$ and M is an $n \times n$ matrix independent from $\mathbf{Y}$:

$$\mathrm{crit}_{\mathrm{GCV}}(M, \mathbf{Y}) := \frac{n^{-1} \left\| \mathbf{Y} - M\mathbf{Y} \right\|^2}{\left(1 - n^{-1} \operatorname{tr}(M)\right)^2} \quad \text{where} \quad \forall t \in \mathbb{R}^n, \, \|t\|^2 = \sum_{i=1}^{n} t_i^2 \ .$$

GCV is actually closer to C_L (Mallows, 1973) than to CV, since GCV can be seen as an approximation to C_L with a particular estimator of the variance (Efron, 1986). The efficiency of GCV has been proved in various frameworks, in particular by Li (1985, 1987) and by Cao and Golubev (2006).

Analytic Approximation When CV is used for selecting among linear models, Shao (1993) proposed an analytic approximation to LPO with $p \sim n$, which is called APCV.

LOO bootstrap and .632 bootstrap The bootstrap is often used for stabilizing an estimator or an algorithm, replacing $\mathcal{A}(D_n)$ by the average of $\mathcal{A}(D_n^\star)$ over several bootstrap resamples $D_n^\star$. This idea was applied by Efron (1983) to the LOO estimator of the risk, leading to the *LOO bootstrap*. Noting that the LOO bootstrap was biased, Efron (1983) gave a heuristic argument leading to the .632 *bootstrap* estimator of the risk, later modified into the .632+ *bootstrap* by Efron and Tibshirani (1997). The main drawback of these procedures is the weakness of their theoretical justifications. Only empirical studies have supported the good behaviour of .632+ bootstrap (Efron and Tibshirani, 1997; Molinaro et al., 2005).

4.4 Historical remarks

Simple validation or hold-out was the first CV-like procedure. It was introduced in the psychology area (Larson, 1931) from the need for a reliable alternative to the *resubstitution error*, as illustrated by Anderson et al. (1972). The hold-out was used by Herzberg (1969) for assessing the quality of predictors. The problem of choosing the training set was first considered by Stone (1974), where

"controllable" and "uncontrollable" data splits were distinguished; an instance of uncontrollable division can be found in the book by Simon (1971).

A primitive LOO procedure was used by Hills (1966) and by Lachenbruch and Mickey (1968) for evaluating the error rate of a prediction rule, and a primitive formulation of LOO can be found in a paper by Mosteller and Tukey (1968). Nevertheless, LOO was actually introduced independently by Stone (1974), by Allen (1974) and by Geisser (1975). The relationship between LOO and the jackknife (Quenouille, 1949), which both rely on the idea of removing one observation from the sample, has been discussed by Stone (1974) for instance.

The hold-out and CV were originally used only for estimating the risk of an algorithm. The idea of using CV for model selection arose in the discussion of a paper by Efron and Morris (1973) and in a paper by Geisser (1974). The first author to study LOO as a model selection procedure was Stone (1974), who proposed to use LOO again for estimating the risk of the selected model.

5 Statistical properties of cross-validation estimators of the risk

Understanding the behaviour of CV for model selection, which is the purpose of this survey, requires first to analyze the performances of CV as an estimator of the risk of a single algorithm. Two main properties of CV estimators of the risk are of particular interest: their bias, and their variance.

5.1 Bias

Dealing with the bias incurred by CV estimates can be made by two strategies: evaluating the amount of bias in order to choose the least biased CV procedure, or correcting for this bias.

5.1.1 Theoretical assessment of the bias

The independence of the training and the validation samples imply that for every algorithm $\mathcal{A}$ and any $I^{(t)} \subset \{1, \ldots, n\}$ with cardinality n_t,

$$\mathbb{E}\left[\widehat{\mathcal{L}}^{\mathrm{H-O}}\left(\mathcal{A}; D_n; I^{(t)}\right)\right] = \mathbb{E}\left[\gamma\left(\mathcal{A}\left(D_n^{(t)}\right); \xi\right)\right] = \mathbb{E}\left[\mathcal{L}_P\left(\mathcal{A}\left(D_{n_t}\right)\right)\right] \ .$$

Therefore, assuming that $\mathrm{Card}(I_j^{(t)}) = n_t$ for $j = 1, \ldots, B$, the expectation of the CV estimator of the risk only depends on n_t :

$$\mathbb{E}\left[\widehat{\mathcal{L}}^{\mathrm{CV}}\left(\mathcal{A}; D_n; \left(I_j^{(t)}\right)_{1 \le j \le B}\right)\right] = \mathbb{E}\left[\mathcal{L}_P\left(\mathcal{A}\left(D_{n_t}\right)\right)\right] \ . \tag{12}$$

In particular (12) shows that the bias of the CV estimator of the risk of $\mathcal{A}$ is the difference between the risks of $\mathcal{A}$, computed respectively with n_t and n data points. Since $n_t < n$, the bias of CV is usually nonnegative, which can be proved rigorously when the risk of $\mathcal{A}$ is a decreasing function of n, that is, when $\mathcal{A}$ is a smart rule; note however that a classical algorithm such as 1-nearest-neighbour in classification is not smart (Devroye et al., 1996, Section 6.8). Similarly, the bias of CV tends to decrease with n_t, which is rigorously true if $\mathcal{A}$ is smart.

More precisely, (12) has led to several results on the bias of CV, which can be split into three main categories: asymptotic results ($\mathcal{A}$ is fixed and the sample size n tends to infinity), non-asymptotic results (where $\mathcal{A}$ is allowed to make use of a number of parameters growing with n, say $n^{1/2}$, as often in model selection), and empirical results. They are listed below by statistical framework.

Regression The general behaviour of the bias of CV (positive, decreasing with n_t) is confirmed by several papers and for several CV estimators. For LPO, non-asymptotic expressions of its bias were proved by Celisse (2008b) for projection estimators, and by Arlot and Celisse (2009) for regressograms and kernels estimators when the design is fixed. For VFCV and RLT, an asymptotic expansion of their bias was yielded by Burman (1989) for least-squares estimators in linear regression, and extended to spline smoothing (Burman, 1990). Note finally that Efron (1986) proved non-asymptotic analytic expressions of the expectations of the LOO and GCV estimators of the risk in regression with binary data (see also Efron, 1983, for some explicit calculations).

Density estimation shows a similar picture. Non-asymptotic expressions for the bias of LPO estimators for kernel and projection estimators with the quadratic risk were proved by Celisse and Robin (2008) and by Celisse (2008a). Asymptotic expansions of the bias of the LOO estimator for histograms and kernel estimators were previously proved by Rudemo (1982); see Bowman (1984) for simulations. Hall (1987) derived similar results with the log-likelihood contrast for kernel estimators, and related the performance of LOO to the interaction between the kernel and the tails of the target density s.

Classification For the simple problem of discriminating between two populations with shifted distributions, Davison and Hall (1992) compared the asymptotical bias of LOO and bootstrap, showing the superiority of the LOO when the shift size is $n^{-1/2}$: As n tends to infinity, the bias of LOO stays of order n^{-1}, whereas that of bootstrap worsens to the order $n^{-1/2}$. On realistic synthetic and real biological data, Molinaro et al. (2005) compared the bias of LOO, VFCV and .632+ bootstrap: The bias decreases with n_t, and is generally minimal for LOO. Nevertheless, the 10-fold CV bias is nearly minimal uniformly over their experiments. In the same experiments, .632+ bootstrap exhibits the smallest bias for moderate sample sizes and small signal-to-noise ratios, but a much larger bias otherwise.

CV-calibrated algorithms When a family of algorithm $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ is given, and $\widehat{\lambda}$ is chosen by minimizing $\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A}_\lambda; D_n)$ over λ, $\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A}_{\widehat{\lambda}}; D_n)$ is biased for estimating the risk of $\mathcal{A}_{\widehat{\lambda}}(D_n)$, as reported from simulation experiments by Stone (1974) for the LOO, and by Jonathan et al. (2000) for VFCV in the variable selection setting. This bias is of different nature compared to the previous frameworks. Indeed, $\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A}_{\widehat{\lambda}}, D_n)$ is biased simply because $\widehat{\lambda}$ was chosen using the same data as $\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A}_\lambda, D_n)$. This phenomenon is similar to the optimism of $\mathcal{L}_{P_n}(\widehat{s}(D_n))$ as an estimator of the loss of $\widehat{s}(D_n)$. The correct way of estimating the risk of $\mathcal{A}_{\widehat{\lambda}}(D_n)$ with CV is to consider the full algorithm

$\mathcal{A}' : D_n \mapsto \mathcal{A}_{\widehat{\lambda}(D_n)}(D_n)$, and then to compute $\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A}'; D_n)$. The resulting procedure is called "double cross" by Stone (1974).

5.1.2 Correction of the bias

An alternative to choosing the CV estimator with the smallest bias is to correct for the bias of the CV estimator of the risk. Burman (1989, 1990) proposed a corrected VFCV estimator, defined by

$$\widehat{\mathcal{L}}^{\mathrm{corrVF}}(\mathcal{A}; D_n) = \widehat{\mathcal{L}}^{\mathrm{VF}}(\widehat{s}; D_n) + \mathcal{L}_{P_n}(\mathcal{A}(D_n)) - \frac{1}{V}\sum_{j=1}^{V} \mathcal{L}_{P_n}\left(\mathcal{A}(D_n^{(-A_j)})\right) \quad ,$$

and a corrected RLT estimator was defined similarly. Both estimators have been proved to be asymptotically unbiased for least-squares estimators in linear regression.

When the A_js have exactly the same size n/V, the corrected VFCV criterion is equal to the sum of the empirical risk and the V-fold penalty (Arlot, 2008c), defined by

$$\mathrm{pen}_{\mathrm{VF}}(\mathcal{A}; D_n) = \frac{V-1}{V}\sum_{j=1}^{V}\left[\mathcal{L}_{P_n}\left(\mathcal{A}(D_n^{(-A_j)})\right) - \mathcal{L}_{P_n^{(-A_j)}}\left(\mathcal{A}(D_n^{(-A_j)})\right)\right] \quad .$$

The V-fold penalized criterion was proved to be (almost) unbiased in the non-asymptotic framework for regressogram estimators.

5.2 Variance

CV estimators of the risk using training sets of the same size n_t have the same bias, but they still behave quite differently; their variance $\mathrm{var}(\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A}; D_n; (I_j^{(t)})_{1\leq j\leq B}))$ captures most of the information to explain these differences.

5.2.1 Variability factors

Assume that $\mathrm{Card}(I_j^{(t)}) = n_t$ for every j. The variance of CV results from the combination of several factors, in particular (n_t, n_v) and B.

Influence of (n_t, n_v) Let us consider the hold-out estimator of the risk. Following in particular Nadeau and Bengio (2003),

$$\begin{aligned}
&\mathrm{var}\left[\widehat{\mathcal{L}}^{\mathrm{H-O}}\left(\mathcal{A}; D_n; I^{(t)}\right)\right] \\
&= \mathbb{E}\left[\mathrm{var}\left(\mathcal{L}_{P_n^{(v)}}\left(\mathcal{A}(D_n^{(t)})\right) \,\middle|\, D_n^{(t)}\right)\right] + \mathrm{var}\left[\mathcal{L}_P\left(\mathcal{A}(D_{n_t})\right)\right] \\
&= \frac{1}{n_v}\mathbb{E}\left[\mathrm{var}\left(\gamma(\widehat{s}, \xi) \,|\, \widehat{s} = \mathcal{A}(D_n^{(t)})\right)\right] + \mathrm{var}\left[\mathcal{L}_P\left(\mathcal{A}(D_{n_t})\right)\right] \quad . \qquad (13)
\end{aligned}$$

The first term, proportional to $1/n_v$, shows that more data for validation decreases the variance of $\widehat{\mathcal{L}}^{\mathrm{H-O}}$, because it yields a better estimator of $\mathcal{L}_P\left(\mathcal{A}(D_n^{(t)})\right)$. The second term shows that the variance of $\widehat{\mathcal{L}}^{\mathrm{H-O}}$ also depends on the distribution of $\mathcal{L}_P\left(\mathcal{A}(D_n^{(t)})\right)$ around its expectation; in particular, it strongly depends on the *stability* of $\mathcal{A}$.

Stability and variance When $\mathcal{A}$ is unstable, $\widehat{\mathcal{L}}^{\mathrm{LOO}}(\mathcal{A})$ has often been pointed out as a variable estimator (Section 7.10, Hastie et al., 2001; Breiman, 1996). Conversely, this trend disappears when $\mathcal{A}$ is stable, as noticed by Molinaro et al. (2005) from a simulation experiment.

The relation between the stability of $\mathcal{A}$ and the variance of $\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A})$ was pointed out by Devroye and Wagner (1979) in classification, through upper bounds on the variance of $\widehat{\mathcal{L}}^{\mathrm{LOO}}(\mathcal{A})$. Bousquet and Elisseff (2002) extended these results to the regression setting, and proved upper bounds on the maximal upward deviation of $\widehat{\mathcal{L}}^{\mathrm{LOO}}(\mathcal{A})$.

Note finally that several approaches based on the bootstrap have been proposed for reducing the variance of $\widehat{\mathcal{L}}^{\mathrm{LOO}}(\mathcal{A})$, such as LOO bootstrap, .632 bootstrap and .632+ bootstrap (Efron, 1983); see also Section 4.3.3.

Partial splitting and variance When (n_t, n_v) is fixed, the variability of CV tends to be larger for partial data splitting methods than for LPO. Indeed, having to choose $B < \binom{n}{n_t}$ subsets $(I_j^{(t)})_{1 \leq j \leq B}$ of $\{1, \ldots, n\}$, usually randomly, induces an additional variability compared to $\widehat{\mathcal{L}}^{\mathrm{LPO}}$ with $p = n - n_t$. In the case of MCCV, this variability decreases like B^{-1} since the $I_j^{(t)}$ are chosen independently. The dependence on B is slightly different for other CV estimators such as RLT or VFCV, because the $I_j^{(t)}$ are not independent. In particular, it is maximal for the hold-out, and minimal (null) for LOO (if $n_t = n - 1$) and LPO (with $p = n - n_t$).

Note that the dependence on V for VFCV is more complex to evaluate, since B, n_t, and n_v simultaneously vary with V. Nevertheless, a non-asymptotic theoretical quantification of this additional variability of VFCV has been obtained by Celisse and Robin (2008) in the density estimation framework (see also empirical considerations by Jonathan et al., 2000).

5.2.2 Theoretical assessment of the variance

Understanding precisely how $\mathrm{var}(\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A}))$ depends on the splitting scheme is complex in general, since n_t and n_v have a fixed sum n, and the number of splits B is generally linked with n_t (for instance, for LPO and VFCV). Furthermore, the variance of CV behaves quite differently in different frameworks, depending in particular on the stability of $\mathcal{A}$. The consequence is that contradictory results have been obtained in different frameworks, in particular on the value of V for which the VFCV estimator of the risk has a minimal variance (Burman, 1989; Hastie et al., 2001, Section 7.10). Despite the difficulty of the problem, the variance of several CV estimators of the risk has been assessed in several frameworks, as detailed below.

Regression In the linear regression setting, Burman (1989) yielded asymptotic expansions of the variance of the VFCV and RLT estimators of the risk with homoscedastic data. The variance of RLT decreases with B, and in the case of VFCV, in a particular setting,

$$\mathrm{var}\left(\widehat{\mathcal{L}}^{\mathrm{VF}}(\mathcal{A})\right) = \frac{2\sigma^2}{n} + \frac{4\sigma^4}{n^2}\left[4 + \frac{4}{V-1} + \frac{2}{(V-1)^2} + \frac{1}{(V-1)^3}\right] + o\left(n^{-2}\right) \quad .$$

The asymptotical variance of the VFCV estimator of the risk decreases with V, implying that LOO asymptotically has the minimal variance.

Non-asymptotic closed-form formulas of the variance of the LPO estimator of the risk have been proved by Celisse (2008b) in regression, for projection and kernel estimators for instance. On the variance of RLT in the regression setting, see the asymptotic results of Girard (1998) for Nadaraya-Watson kernel estimators, as well as the non-asymptotic computations and simulation experiments by Nadeau and Bengio (2003) with several learning algorithms.

Density estimation Non-asymptotic closed-form formulas of the variance of the LPO estimator of the risk have been proved by Celisse and Robin (2008) and by Celisse (2008a) for projection and kernel estimators. In particular, the dependence of the variance of $\widehat{\mathcal{L}}^{\mathrm{LPO}}$ on p has been quantified explicitly for histogram and kernel estimators by Celisse and Robin (2008).

Classification For the simple problem of discriminating between two populations with shifted distributions, Davison and Hall (1992) showed that the gap between asymptotic variances of LOO and bootstrap becomes larger when data are noisier. Nadeau and Bengio (2003) made non-asymptotic computations and simulation experiments with several learning algorithms. Hastie et al. (2001) empirically showed that VFCV has a minimal variance for some $2 < V < n$, whereas LOO usually has a large variance; this fact certainly depends on the stability of the algorithm considered, as showed by simulation experiments by Molinaro et al. (2005).

5.2.3 Estimation of the variance

There is no universal—valid under all distributions—unbiased estimator of the variance of RLT (Nadeau and Bengio, 2003) and VFCV estimators (Bengio and Grandvalet, 2004). In particular, Bengio and Grandvalet (2004) recommend the use of variance estimators taking into account the correlation structure between test errors; otherwise, the variance of CV can be strongly underestimated.

Despite these negative results, (biased) estimators of the variance of $\widehat{\mathcal{L}}^{\mathrm{CV}}$ have been proposed by Nadeau and Bengio (2003), by Bengio and Grandvalet (2004) and by Markatou et al. (2005), and tested in simulation experiments in regression and classification. Furthermore, in the framework of density estimation with histograms, Celisse and Robin (2008) proposed an estimator of the variance of the LPO risk estimator. Its accuracy is assessed by a concentration inequality. These results have recently been extended to projection estimators by Celisse (2008a).

6 Cross-validation for efficient model selection

This section tackles the properties of CV procedures for model selection when the goal is estimation (see Section 2.2).

6.1 Relationship between risk estimation and model selection

As shown in Section 3.1, minimizing an unbiased estimator of the risk leads to an efficient model selection procedure. One could conclude here that the best CV procedure for estimation is the one with the smallest bias and variance (at least asymptotically), for instance, LOO in the least-squares regression framework (Burman, 1989).

Nevertheless, the best CV estimator of the risk is not necessarily the best model selection procedure. For instance, Breiman and Spector (1992) observed that uniformly over the models, the best risk estimator is LOO, whereas 10-fold CV is more accurate for model selection. Three main reasons for such a difference can be invoked. First, the asymptotic framework ($\mathcal{A}$ fixed, $n \to \infty$) may not apply to models close to the oracle, which typically has a dimension growing with n when s does not belong to any model. Second, as explained in Section 3.2, estimating the risk of each model with some bias can be beneficial and compensate the effect of a large variance, in particular when the signal-to-noise ratio is small. Third, for model selection, what matters is not that every estimate of the risk has small bias and variance, but more that

$$\operatorname{sign}\left(\operatorname{crit}(m_1) - \operatorname{crit}(m_2)\right) = \operatorname{sign}\left(\mathcal{L}_P\left(\widehat{s}_{m_1}\right) - \mathcal{L}_P\left(\widehat{s}_{m_2}\right)\right)$$

with the largest probability for models m_1, m_2 near the oracle.

Therefore, specific studies are required to evaluate the performances of the various CV procedures in terms of model selection efficiency. In most frameworks, the model selection performance directly follows from the properties of CV as an estimator of the risk, but not always.

6.2 The global picture

Let us start with the classification of model selection procedures made by Shao (1997) in the linear regression framework, since it gives a good idea of the performance of CV procedures for model selection in general. Typically, the efficiency of CV only depends on the asymptotics of n_t/n:

- When $n_t \sim n$, CV is asymptotically equivalent to Mallows' C_p, hence asymptotically optimal.
- When $n_t \sim \lambda n$ with $\lambda \in (0,1)$, CV is asymptotically equivalent to GIC_κ with $\kappa = 1+\lambda^{-1}$, which is defined as AIC with a penalty multiplied by $\kappa/2$. Hence, such CV procedures are overpenalizing by a factor $(1+\lambda)/(2\lambda) > 1$.

The above results have been proved by Shao (1997) for LPO (see also Li, 1987, for the LOO); they also hold for RLT when $B \gg n^2$ since RLT is then equivalent to LPO (Zhang, 1993).

In a general statistical framework, the model selection performance of MCCV, VFCV, LOO, LOO Bootstrap, and .632 bootstrap for selection among minimum contrast estimators was studied in a series of papers (van der Laan and Dudoit, 2003; van der Laan et al., 2004, 2006; van der Vaart et al., 2006); these results apply in particular to least-squares regression and density estimation. It turns out that under mild conditions, an oracle-type inequality is proved, showing that up to a multiplying factor $C_n \to 1$,

the risk of CV is smaller than the minimum of the risks of the models with a sample size n_t. In particular, in most frameworks, this implies the asymptotic optimality of CV as soon as $n_t \sim n$. When $n_t \sim \lambda n$ with $\lambda \in (0,1)$, this naturally generalizes Shao's results.

6.3 Results in various frameworks

This section gathers results about model selection performances of CV when the goal is estimation, in various frameworks. Note that model selection is considered here with a general meaning, including in particular bandwidth choice for kernel estimators.

Regression First, the results of Section 6.2 suggest that CV is suboptimal when n_t is not asymptotically equivalent to n. This fact has been proved rigorously for VFCV when $V = \mathcal{O}(1)$ with regressograms (Arlot, 2008c): with large probability, the risk of the model selected by VFCV is larger than $1 + \kappa(V)$ times the risk of the oracle, with $\kappa(V) > 0$ for every fixed V. Note however that the best V for VFCV is not the largest one in every regression framework, as shown empirically in linear regression (Breiman and Spector, 1992; Herzberg and Tsukanov, 1986); Breiman (1996) proposed to explain this phenomenon by relating the stability of the candidate algorithms and the model selection performance of LOO in various regression frameworks.

Second, the "universality" of CV has been confirmed by showing that it naturally adapts to heteroscedasticity of data when selecting among regressograms. Despite its suboptimality, VFCV with $V = \mathcal{O}(1)$ satisfies a non-asymptotic oracle inequality with constant $C > 1$ (Arlot, 2008c). Furthermore, V-fold penalization (which often coincides with corrected VFCV, see Section 5.1.2) satisfies a non-asymptotic oracle inequality with $C_n \to 1$ as $n \to +\infty$, both when $V = \mathcal{O}(1)$ (Arlot, 2008c) and when $V = n$ (Arlot, 2008a). Note that n-fold penalization is very close to LOO, suggesting that it is also asymptotically optimal with heteroscedastic data. Simulation experiments in the context of change-point detection confirmed that CV adapts well to heteroscedasticity, contrary to usual model selection procedures in the same framework (Arlot and Celisse, 2009).

The performances of CV have also been assessed for other kinds of estimators in regression. For choosing the number of knots in spline smoothing, Burman (1990) proved that corrected versions of VFCV and RLT are asymptotically optimal provided $n/(Bn_v) = \mathcal{O}(1)$. Furthermore, in kernel regression, several CV methods have been compared to GCV in kernel regression by Härdle et al. (1988) and by Girard (1998); the conclusion is that GCV and related criteria are computationally more efficient than MCCV or RLT, for a similar statistical performance.

Finally, note that asymptotic results about CV in regression have been proved by Györfi et al. (2002), and an oracle inequality with constant $C > 1$ has been proved by Wegkamp (2003) for the hold-out, with least-squares estimators.

Density estimation CV performs similarly than in regression for selecting among least-squares estimators (van der Laan et al., 2004): It yields a risk smaller than the minimum of the risk with a sample size n_t. In particular,

non-asymptotic oracle inequalities with constant $C > 1$ have been proved by Celisse (2008b) for the LPO when $p/n \in [a, b]$, for some $0 < a < b < 1$.

The performance of CV for selecting the bandwidth of kernel density estimators has been studied in several papers. With the least-squares contrast, the efficiency of LOO was proved by Hall (1983) and generalized to the multivariate framework by Stone (1984); an oracle inequality asymptotically leading to efficiency was recently proved by Dalelane (2005). With the Kullback-Leibler divergence, CV can suffer from troubles in performing model selection (see also Schuster and Gregory, 1981; Chow et al., 1987). The influence of the tails of the target s was studied by Hall (1987), who gave conditions under which CV is efficient and the chosen bandwidth is optimal at first-order.

Classification In the framework of binary classification by intervals (that is, with $\mathcal{X} = [0, 1]$ and piecewise constant classifiers), Kearns et al. (1997) proved an oracle inequality for the hold-out. Furthermore, empirical experiments show that CV yields (almost) always the best performance, compared to deterministic penalties (Kearns et al., 1997). On the contrary, simulation experiments by Bartlett et al. (2002) in the same setting showed that random penalties such as Rademacher complexity and maximal discrepancy usually perform much better than hold-out, which is shown to be more variable.

Nevertheless, the hold-out still enjoys quite good theoretical properties: It was proved to adapt to the margin condition by Blanchard and Massart (2006), a property nearly unachievable with usual model selection procedures (see also Massart, 2007, Section 8.5). This suggests that CV procedures are naturally adaptive to several unknown properties of data in the statistical learning framework.

The performance of the LOO in binary classification was related to the stability of the candidate algorithms by Kearns and Ron (1999); they proved oracle-type inequalities called "sanity-check bounds", describing the worst-case performance of LOO (see also Bousquet and Elisseff, 2002).

An experimental comparison of several CV methods and bootstrap-based CV (in particular .632+ bootstrap) in classification can also be found in papers by Efron (1986) and Efron and Tibshirani (1997).

7 Cross-validation for identification

Let us now focus on model selection when the goal is to identify the "true model" S_{m_0}, as described in Section 2.3. In this framework, asymptotic optimality is replaced by (model) consistency, that is,

$$\mathbb{P}\left(\widehat{m}(D_n) = m_0\right) \xrightarrow[n\to\infty]{} 1 \ .$$

Classical model selection procedures built for identification, such as BIC, are described in Section 3.3.

7.1 General conditions towards model consistency

At first sight, it may seem strange to use CV for identification: LOO, which is the pioneering CV procedure, is actually closely related to the unbiased risk

estimation principle, which is only efficient when the goal is estimation. Furthermore, estimation and identification are somehow contradictory goals, as explained in Section 2.4.

This intuition about inconsistency of some CV procedures is confirmed by several theoretical results. Shao (1993) proved that several CV methods are inconsistent for variable selection in linear regression: LOO, LPO, and BICV when $\liminf_{n\to\infty}(n_t/n) > 0$. Even if these CV methods asymptotically select all the true variables with probability 1, the probability that they select too much variables does not tend to zero. More generally, Shao (1997) proved that CV procedures behave asymptotically like GIC_{λ_n} with $\lambda_n = 1 + n/n_t$, which leads to inconsistency as soon as $n/n_t = \mathcal{O}(1)$.

In the context of ordered variable selection in linear regression, Zhang (1993) computed the asymptotic value of the probability of selecting the true model for several CV procedures. He also numerically compared the values of this probability for the same CV procedures in a specific example. For LPO with $p/n \to \lambda \in (0,1)$ as n tends to $+\infty$, $\mathbb{P}(\widehat{m} = m_0)$ increases with λ. The result is slightly different for VFCV: $\mathbb{P}(\widehat{m} = m_0)$ increases with V (hence, it is maximal for the LOO, which is the worst case of LPO). The variability induced by the number V of splits seems to be more important here than the bias of VFCV. Nevertheless, $\mathbb{P}(\widehat{m} = m_0)$ is almost constant between $V = 10$ and $V = n$, so that taking $V > 10$ is not advised for computational reasons.

These results suggest that if the training sample size n_t is negligible in front of n, then model consistency could be obtained. This has been confirmed theoretically by Shao (1993, 1997) for the variable selection problem in linear regression: CV is consistent when $n \gg n_t \to \infty$, in particular RLT, BICV (defined in Section 4.3.2) and LPO with $p = p_n \sim n$ and $n - p_n \to \infty$.

Therefore, when the goal is to identify the true model, a larger proportion of the data should be put in the validation set in order to improve the performance. This phenomenon is somewhat related to the *cross-validation paradox* (Yang, 2006).

7.2 Refined analysis for the algorithm selection problem

The behaviour of CV for identification is better understood by considering a more general framework, where the goal is to select among statistical algorithms the one with the fastest convergence rate. Yang (2006, 2007) considered this problem for two candidate algorithms (or more generally any finite number of algorithms). Let us mention here that Stone (1977) considered a few specific examples of this problem, and showed that LOO can be inconsistent for choosing the best among two "good" estimators.

The conclusion of Yang's papers is that the sufficient condition on n_t for the consistency in selection of CV strongly depends on the convergence rates $(r_{n,i})_{i=1,2}$ of the candidate algorithms. Let us assume that $r_{n,1}$ and $r_{n,2}$ differ at least by a multiplicative constant $C > 1$. Then, in the regression framework, if the risk of $\widehat{s}_i$ is measured by $\mathbb{E}\,\|\widehat{s}_i - s\|_2$, Yang (2007) proved that the hold-out, VFCV, RLT and LPO with voting (CV-v, see Section 4.2.2) are consistent in selection if

$$n_v, n_t \to \infty \quad \text{and} \quad \sqrt{n_v}\max_i r_{n_t,i} \to \infty \ , \tag{14}$$

under some conditions on $\|\widehat{s}_i - s\|_p$ for $p = 2, 4, \infty$. In the classification framework, if the risk of $\widehat{s}_i$ is measured by $\mathbb{P}(\widehat{s}_i \neq s)$, Yang (2006) proved the same consistency result for CV-v under the condition

$$n_v, n_t \to \infty \quad \text{and} \quad \frac{n_v \max_i r_{n_t,i}^2}{s_{n_t}} \to \infty \ , \tag{15}$$

where s_n is the convergence rate of $\mathbb{P}(\widehat{s}_1(D_n) \neq \widehat{s}_2(D_n))$.

Intuitively, consistency holds as soon as the uncertainty of each estimate of the risk (roughly proportional to $n_v^{-1/2}$) is negligible in front of the risk gap $|r_{n_t,1} - r_{n_t,2}|$ (which is of the same order as $\max_i r_{n_t,i}$). This condition holds either when at least one of the algorithms converges at a non-parametric rate, or when $n_t \ll n$, which artificially widens the risk gap.

Empirical results in the same direction were proved by Dietterich (1998) and by Alpaydin (1999), leading to the advice that $V = 2$ is the best choice when VFCV is used for comparing two learning procedures. See also the results by Nadeau and Bengio (2003) about CV considered as a testing procedure comparing two candidate algorithms.

The sufficient conditions (14) and (15) can be simplified depending on $\max_i r_{n,i}$, so that the ability of CV to distinguish between two algorithms depends on their convergence rates. On the one hand, if $\max_i r_{n,i} \propto n^{-1/2}$, then (14) or (15) only hold when $n_v \gg n_t \to \infty$ (under some conditions on s_n in classification). Therefore, the cross-validation paradox holds for comparing algorithms converging at the parametric rate (model selection when a true model exists being only a particular case). Note that possibly stronger conditions can be required in classification where algorithms can converge at fast rates, between n^{-1} and $n^{-1/2}$.

On the other hand, (14) and (15) are milder conditions when $\max_i r_{n,i} \gg n^{-1/2}$: They are implied by $n_t/n_v = \mathcal{O}(1)$, and they even allow $n_t \sim n$ (under some conditions on s_n in classification). Therefore, non-parametric algorithms can be compared by more usual CV procedures ($n_t > n/2$), even if LOO is still excluded by conditions (14) and (15).

Note that according to a simulation experiments, CV with averaging (that is, CV as usual) and CV with voting are equivalent at first but not at second order, so that they can differ when n is small (Yang, 2007).

8 Specificities of some frameworks

Originally, the CV principle has been proposed for *i.i.d.* observations and usual contrasts such as least-squares and log-likelihood. Therefore, CV procedures may have to be modified in other specific frameworks, such as estimation in presence of outliers or with dependent data.

8.1 Density estimation

In the density estimation framework, some specific modifications of CV have been proposed.

First, Hall et al. (1992) defined the "smoothed CV", which consists in presmoothing the data before using CV, an idea related to the smoothed bootstrap.

They proved that smoothed CV yields an excellent asymptotical model selection performance under various smoothness conditions on the density.

Second, when the goal is to estimate the density at one point (and not globally), Hall and Schucany (1989) proposed a local version of CV and proved its asymptotic optimality.

8.2 Robustness to outliers

In presence of outliers in regression, Leung (2005) studied how CV must be modified to get both asymptotic efficiency and a consistent bandwidth estimator (see also Leung et al., 1993).

Two changes are possible to achieve robustness: Choosing a "robust" regressor, or choosing a robust loss-function. In presence of outliers, classical CV with a non-robust loss function has been shown to fail by Härdle (1984).

Leung (2005) described a CV procedure based on robust losses like L^1 and Huber's (Huber, 1964) ones. The same strategy remains applicable to other setups like linear models in Ronchetti et al. (1997).

8.3 Time series and dependent observations

As explained in Section 4.1, CV is built upon the heuristics that part of the sample (the validation set) can play the role of *new data* with respect to the rest of the sample (the training set). "New" means that the validation set is independent from the training set with the same distribution.

Therefore, when data $\xi_1, \ldots, \xi_n$ are not independent, CV must be modified, like other model selection procedures (in non-parametric regression with dependent data, see the review by Opsomer et al., 2001).

Let us first consider the statistical framework of Section 1 with $\xi_1, \ldots, \xi_n$ identically distributed but not independent. Then, when for instance data are positively correlated, Hart and Wehrly (1986) proved that CV overfits for choosing the bandwidth of a kernel estimator in regression (see also Chu and Marron, 1991; Opsomer et al., 2001).

The main approach used in the literature for solving this issue is to choose $I^{(t)}$ and $I^{(v)}$ such that $\min_{i \in I^{(t)},\, j \in I^{(v)}} |i - j| > h > 0$, where h controls the distance from which observations i and j are independent. For instance, the LOO can be changed into: $I^{(v)} = \{ J \}$ where J is uniformly chosen in $\{ 1, \ldots, n \}$, and $I^{(t)} = \{ 1, \ldots, J - h - 1, J + h + 1, \ldots, n \}$, a method called "modified CV" by Chu and Marron (1991) in the context of bandwidth selection. Then, for short range dependences, ξ_i is almost independent from ξ_j when $|i - j| > h$ is large enough, so that $(\xi_j)_{j \in I^{(t)}}$ is almost independent from $(\xi_j)_{j \in I^{(v)}}$. Several asymptotic optimality results have been proved on modified CV, for instance by Hart and Vieu (1990) for bandwidth choice in kernel density estimation, when data are α-mixing (hence, with a short range dependence structure) and $h = h_n \to \infty$ "not too fast". Note that modified CV also enjoys some asymptotic optimality results with long-range dependences, as proved by Hall et al. (1995), even if an alternative block bootstrap method seems more appropriate in such a framework.

Several alternatives to modified CV have also been proposed. The "h-block CV" (Burman et al., 1994) is modified CV plus a corrective term, similarly to

the bias-corrected CV by Burman (1989) (see Section 5.1). Simulation experiments in several (short range) dependent frameworks show that this corrective term matters when h/n is not small, in particular when n is small.

The "partitioned CV" has been proposed by Chu and Marron (1991) for bandwidth selection: An integer $g > 0$ is chosen, a bandwidth $\widehat{\lambda}_k$ is chosen by CV based upon the subsample $(\xi_{k+gj})_{j\geq 0}$ for each $k = 1, \ldots, g$, and the selected bandwidth is a combination of $(\widehat{\lambda}_k)$.

When a parametric model is available for the dependency structure, Hart (1994) proposed the "time series CV".

An important framework where data often are dependent is time-series analysis, in particular when the goal is to predict the next observation ξ_{n+1} from the past $\xi_1, \ldots, \xi_n$. When data are stationary, h-block CV and similar approaches can be used to deal with (short range) dependences. Nevertheless, Burman and Nolan (1992) proved in some specific framework that unaltered CV is asymptotic optimal when $\xi_1, \ldots, \xi_n$ is a stationary Markov process.

On the contrary, using CV for non-stationary time-series is a quite difficult problem. The only reasonable approach in general is the hold-out, that is, $I^{(t)} = \{1, \ldots, m\}$ and $I^{(v)} = \{m+1, \ldots, n\}$ for some deterministic m. Each model is first trained with $(\xi_j)_{j\in I^{(t)}}$. Then, it is used for predicting successively ξ_{m+1} from $(\xi_j)_{j\leq m}$, ξ_{m+2} from $(\xi_j)_{j\leq m+1}$, and so on. The model with the smallest average error for predicting $(\xi_j)_{j\in I^{(v)}}$ from the past is chosen.

8.4 Large number of models

As mentioned in Section 3, model selection procedures estimating unbiasedly the risk of each model fail when, in particular, the number of models grows exponentially with n (Birgé and Massart, 2007). Therefore, CV cannot be used directly, except maybe with $n_t \ll n$, provided n_t is well chosen (see Section 6 and Celisse, 2008b, Chapter 6).

For least-squares regression with homoscedastic data, Wegkamp (2003) proposed to add to the hold-out estimator of the risk a penalty term depending on the number of models. This method is proved to satisfy a non-asymptotic oracle inequality with leading constant $C > 1$.

Another general approach was proposed by Arlot and Celisse (2009) in the context of multiple change-point detection. The idea is to perform model selection in two steps: First, gather the models $(S_m)_{m\in\mathcal{M}_n}$ into meta-models $(\widetilde{S}_D)_{D\in\mathcal{D}_n}$, where $\mathcal{D}_n$ denotes a set of indices such that $\mathrm{Card}(\mathcal{D}_n)$ grows at most polynomially with n. Inside each meta-model $\widetilde{S}_D = \bigcup_{m\in\mathcal{M}_n(D)} S_m$, $\widehat{s}_D$ is chosen from data by optimizing a given criterion, for instance the empirical contrast $\mathcal{L}_{P_n}(t)$, but other criteria can be used. Second, CV is used for choosing among $(\widehat{s}_D)_{D\in\mathcal{D}_n}$. Simulation experiments show this simple trick automatically takes into account the cardinality of $\mathcal{M}_n$, even when data are heteroscedastic, contrary to other model selection procedures built for exponential collection of models which all assume homoscedasticity of data.

9 Closed-form formulas and fast computation

Resampling strategies, like CV, are known to be time consuming. The naive implementation of CV has a computational complexity of B times the complexity of training each algorithm $\mathcal{A}$, which is usually intractable for LPO, even with $p = 1$. The computational cost of VFCV or RLT can still be quite costly when $B > 10$ in many practical problems. Nevertheless, closed-form formulas for CV estimators of the risk can be obtained in several frameworks, which greatly decreases the computational cost of CV.

In density estimation, closed-form formulas have been originally derived by Rudemo (1982) and by Bowman (1984) for the LOO risk estimator of histograms and kernel estimators. These results have been recently extended by Celisse and Robin (2008) to the LPO risk estimator with the quadratic loss. Similar results are more generally available for projection estimators as settled by Celisse (2008a). Intuitively, such formulas can be obtained provided the number N of values taken by the $B = \binom{n}{n_v}$ hold-out estimators of the risk, corresponding to different data splittings, is at most polynomial in the sample size.

For least-squares estimators in linear regression, Zhang (1993) proved a closed-form formula for the LOO estimator of the risk. Similar results have been obtained by Wahba (1975, 1977), and by Craven and Wahba (1979) in the spline smoothing context as well. These papers led in particular to the definition of GCV (see Section 4.3.3) and related procedures, which are often used instead of CV (with a naive implementation) because of their small computational cost, as emphasized by Girard (1998).

Closed-form formulas for the LPO estimator of the risk were also obtained by Celisse (2008b) in regression for kernel and projection estimators, in particular for regressograms. An important property of these closed-form formulas is their additivity: For a regressogram associated to a partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of $\mathcal{X}$, the LPO estimator of the risk can be written as a sum over $\lambda \in \Lambda_m$ of terms which only depend on observations (X_i, Y_i) such that $X_i \in I_\lambda$. Therefore, dynamic programming (Bellman and Dreyfus, 1962) can be used for minimizing the LPO estimator of the risk over the set of partitions of $\mathcal{X}$ in D pieces. As an illustration, Arlot and Celisse (2009) successfully applied this strategy in the change-point detection framework. Note that the same idea can be used with VFCV or RLT, but for a larger computational cost since no closed-form formulas are available for these CV methods.

Finally, in frameworks where no closed-form formula can be proved, some efficient algorithms exist for avoiding to recompute $\widehat{\mathcal{L}}^{\mathrm{H-O}}(\mathcal{A}; D_n; I_j^{(t)})$ from scratch for each data splitting $I_j^{(t)}$. These algorithms rely on updating formulas such as the ones by Ripley (1996) for LOO in linear and quadratic discriminant analysis; this approach makes LOO as expensive to compute as the empirical risk.

Very similar formulas are also available for LOO and the k-nearest neighbours estimator in classification (Daudin and Mary-Huard, 2008).

10 Conclusion: which cross-validation method for which problem?

This conclusion collects a few guidelines aiming at helping CV users, first interpreting the results of CV, second appropriately using CV in each specific problem.

10.1 The general picture

Drawing a general conclusion on CV methods is an impossible task because of the variety of frameworks where CV can be used, which induces a variety of behaviors of CV. Nevertheless, we can still point out the three main criteria to take into account for choosing a CV method for a particular model selection problem:

- *Bias*: CV roughly estimates the risk of a model with a sample size $n_t < n$ (see Section 5.1). Usually, this implies that CV overestimates the variance term compared to the bias term in the bias-variance decomposition (2) with sample size n.
  When the goal is estimation and the signal-to-noise ratio (SNR) is large, the smaller bias usually is the better, which is obtained by taking $n_t \sim n$. Otherwise, CV can be asymptotically suboptimal. Nevertheless, when the goal is estimation and the SNR is small, keeping a small upward bias for the variance term often improves the performance, which is obtained by taking $n_t \sim \kappa n$ with $\kappa \in (0, 1)$. See Section 6.
  When the goal is identification, a large bias is often needed, which is obtained by taking $n_t \ll n$; depending on the framework, larger values of n_t can also lead to model consistency, see Section 7.

- *Variability*: The variance of the CV estimator of the risk is usually a decreasing function of the number B of splits, for a fixed training size. When the number of splits is fixed, the variability of CV also depends on the training sample size n_t. Usually, CV is more variable when n_t is closer to n. However, when B is linked with n_t (as for VFCV or LPO), the variability of CV must be quantified precisely, which has been done in few frameworks. The only general conclusion on this point is that the CV method with minimal variability seems strongly framework-dependent, see Section 5.2 for details.

- *Computational complexity*: Unless closed-form formulas or analytic approximations are available (see Section 9), the complexity of CV is roughly proportional to the number of data splits: 1 for the hold-out, V for VFCV, B for RLT or MCCV, n for LOO, and $\binom{n}{p}$ for LPO.

The optimal trade-off between these three factors can be different for each problem, depending for instance on the computational complexity of each estimator, on specificities of the framework considered, and on the final user's trade-off between statistical performance and computational cost. Therefore, no "optimal CV method" can be pointed out before having taken into account the final user's preferences.

Nevertheless, in density estimation, closed-form expressions of the LPO estimator have been derived by Celisse and Robin (2008) with histograms and kernel estimators, and by Celisse (2008a) for projection estimators. These expressions allow to perform LPO without additional computational cost, which reduces the aforementioned trade-off to the easier bias-variability trade-off. In particular, Celisse and Robin (2008) proposed to choose p for LPO by minimizing a criterion defined as the sum of a squared bias and a variance terms (see also Politis et al., 1999, Chapter 9).

10.2 How the splits should be chosen?

For hold-out, VFCV, and RLT, an important question is to choose a particular sequence of data splits.

First, should this step be random and independent from D_n, or take into account some features of the problem or of the data? It is often recommended to take into account the structure of data when choosing the splits. If data are stratified, the proportions of the different strata should (approximately) be the same in the sample and in each training and validation sample. Besides, the training samples should be chosen so that $\widehat{s}_m(D_n^{(t)})$ is well defined for every training set; in the regressogram case, this led Arlot (2008c) and Arlot and Celisse (2009) to choose carefully the splitting scheme. In supervised classification, practitioners usually choose the splits so that the proportion of each class is the same in every validation sample as in the sample. Nevertheless, Breiman and Spector (1992) made simulation experiments in regression for comparing several splitting strategies. No significant improvement was reported from taking into account the stratification of data for choosing the splits.

Another question related to the choice of $(I_j^{(t)})_{1\leq j\leq B}$ is whether the $I_j^{(t)}$ should be independent (like MCCV), slighly dependent (like RLT), or strongly dependent (like VFCV). It seems intuitive that giving similar roles to all data points in the B "training and validation tasks" should yield more reliable results as other methods. This intuition may explain why VFCV is much more used than RLT or MCCV. Similarly, Shao (1993) proposed a CV method called BICV, where every point and pair of points appear in the same number of splits, see Section 4.3.2. Nevertheless, most recent theoretical results on the various CV procedures are not accurate enough to distinguish which one may be the best splitting strategy: This remains a widely open theoretical question.

Note finally that the additional variability due to the choice of a sequence of data splits was quantified empirically by Jonathan et al. (2000) and theoretically by Celisse and Robin (2008) for VFCV.

10.3 V-fold cross-validation

VFCV is certainly the most popular CV procedure, in particular because of its mild computational cost. Nevertheless, the question of choosing V remains widely open, even if indications can be given towards an appropriate choice.

A specific feature of VFCV—as well as exhaustive strategies—is that choosing V uniquely determines the size of the training set $n_t = n(V-1)/V$ and the number of splits $B = V$, hence the computational cost. Contradictory phenomena then occur.

On the one hand, the bias of VFCV decreases with V since $n_t = n(1-1/V)$ observations are used in the training set. On the other hand, the variance of VFCV decreases with V for small values of V, whereas the LOO ($V = n$) is known to suffer from a high variance in several frameworks such as classification or density estimation. Note however that the variance of VFCV is minimal for $V = n$ in some frameworks like linear regression (see Section 5.2). Furthermore, estimating the variance of VFCV from data is a difficult problem in general, see Section 5.2.3.

When the goal of model selection is estimation, it is often reported in the literature that the optimal V is between 5 and 10, because the statistical performance does not increase much for larger values of V, and averaging over 5 or 10 splits remains computationally feasible (Hastie et al., 2001, Section 7.10). Even if this claim is clearly true for many problems, the conclusion of this survey is that better statistical performance can sometimes be obtained with other values of V, for instance depending on the SNR value.

When the SNR is large, the asymptotic comparison of CV procedures recalled in Section 6.2 can be trusted: LOO performs (nearly) unbiased risk estimation hence is asymptotically optimal, whereas VFCV with $V = \mathcal{O}(1)$ is suboptimal. On the contrary, when the SNR is small, overpenalization can improve the performance. Therefore, VFCV with $V < n$ can yield a smaller risk than LOO thanks to its bias and despite its variance when V is small (see simulation experiments by Arlot, 2008c). Furthermore, other CV procedures like RLT can be interesting alternatives to VFCV, since they allow to choose the bias (through n_t) independently from B, which mainly governs the variance. Another possible alternative is V-fold penalization, which is related to corrected VFCV (see Section 4.3.3).

When the goal of model selection is identification, the main drawback of VFCV is that $n_t \ll n$ is often required for choosing consistently the true model (see Section 7), whereas VFCV does not allow $n_t < n/2$. Depending on the frameworks, different (empirical) recommandations for choosing V can be found in the literature. In ordered variable selection, the largest V seems to be the better, $V = 10$ providing results close to the optimal ones (Zhang, 1993). On the contrary, Dietterich (1998) and Alpaydin (1999) recommend $V = 2$ for choosing the best learning procedures among two candidates.

10.4 Future research

Perhaps the most important direction for future research would be to provide, in each specific framework, precise quantitative measures of the variance of CV estimators of the risk, depending on n_t, the number of splits, and how the splits are chosen. Up to now, only a few precise results have been obtained in this direction, for some specific CV methods in linear regression or density estimation (see Section 5.2). Proving similar results in other frameworks and for more general CV methods would greatly help to choose a CV method for any given model selection problem.

More generally, most theoretical results are not precise enough to make any distinction between the hold-out and CV methods having the same training sample size n_t, because they are equivalent at first order. Second order terms do matter for realistic values of n, which shows the dramatic need for theory

that takes into account the variance of CV when comparing CV methods such as VFCV and RLT with $n_t = n(V-1)/V$ but $B \neq V$.

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127.

Alpaydin, E. (1999). Combined 5 x 2 cv F test for comparing supervised classification learning algorithms. *Neur. Comp.*, 11(8):1885–1892.

Anderson, R. L., Allen, D. M., and B., C. F. (1972). Selectioon of predictor variables in linear multiple regression. In bancroft, T. A., editor, *In Statistical papers in Honor of George W. Snedecor*. Iowa: iowa State University Press.

Arlot, S. (2007). *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11. oai:tel.archives-ouvertes.fr:tel-00198803_v1.

Arlot, S. (2008a). Model selection by resampling penalization. *Electronic Journal of Statistics*. To appear. oai:hal.archives-ouvertes.fr:hal-00262478_v1.

Arlot, S. (2008b). Suboptimality of penalties proportional to the dimension for model selection in heteroscedastic regression. arXiv:0812.3141.

Arlot, S. (2008c). V-fold cross-validation improved: V-fold penalization. arXiv:0802.0566v2.

Arlot, S. and Celisse, A. (2009). Segmentation in the mean of heteroscedastic data via cross-validation. arXiv:0902.3977v2.

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic).

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic).

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.

Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, 48:85–113.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3(Spec. Issue Comput. Learn. Theory):463–482.

Bellman, R. E. and Dreyfus, S. E. (1962). *Applied Dynamic Programming*. Princeton.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.*, 5:1089–1105 (electronic).

Bhansali, R. J. and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika*, 64(3):547–551.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73.

Blanchard, G. and Massart, P. (2006). Discussion: "Local Rademacher complexities and oracle inequalities in risk minimization" [Ann. Statist. **34** (2006), no. 6, 2593–2656] by V. Koltchinskii. *Ann. Statist.*, 34(6):2664–2671.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375 (electronic).

Bousquet, O. and Elisseff, A. (2002). Stability and Generalization. *J. Machine Learning Research*, 2:499–526.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24(6):2350–2383.

Breiman, L. (1998). Arcing classifiers. *Ann. Statist.*, 26(3):801–849. With discussion and a rejoinder by the author.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. *International Statistical Review*, 60(3):291–319.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514.

Burman, P. (1990). Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā Ser. A*, 52(3):314–345.

Burman, P., Chow, E., and Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, 81(2):351–358.

Burman, P. and Nolan, D. (1992). Data-dependent estimation of prediction functions. *J. Time Ser. Anal.*, 13(3):189–207.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference*. Springer-Verlag, New York, second edition. A practical information-theoretic approach.

Cao, Y. and Golubev, Y. (2006). On oracle inequalities related to smoothing splines. *Math. Methods Statist.*, 15(4):398–414.

Celisse, A. (2008a). Density estimation via cross-validation: Model selection point of view. Technical report, arXiv: 0811.0802.

Celisse, A. (2008b). *Model Selection Via Cross-Validation in Density Estimation, Regression and Change-Points Detection*. PhD thesis, University Paris-Sud 11, `http://tel.archives-ouvertes.fr/tel-00346320/en/`.

Celisse, A. and Robin, S. (2008). Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368.

Chow, Y. S., Geman, S., and Wu, L. D. (1987). Consistent cross-validated density estimation. *Ann. Statist.*, 11:25–38.

Chu, C.-K. and Marron, J. S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, 19(4):1906–1918.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4):377–403.

Dalelane, C. (2005). Exact oracle inequality for sharp adaptive kernel density estimator. Technical report, arXiv.

Daudin, J.-J. and Mary-Huard, T. (2008). Estimation of the conditional risk in classification: The swapping method. *Comput. Stat. Data Anal.*, 52(6):3220–3232.

Davison, A. C. and Hall, P. (1992). On the bias and variability of bootstrap and cross-validation estimates of error rate in discrimination problems. *Biometrika*, 79(2):279–284.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York.

Devroye, L. and Wagner, T. J. (1979). Distribution-Free performance Bounds for Potential Function Rules. *IEEE Transaction in Information Theory*, 25(5):601–604.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neur. Comp.*, 10(7):1895–1924.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.*, 81(394):461–470.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, 99(467):619–642. With comments and a rejoinder by the author.

Efron, B. and Morris, C. (1973). Combining possibly related estimation problems (with discussion). *J. R. Statist. Soc. B*, 35:379.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *J. Amer. Statist. Assoc.*, 92(438):548–560.

Fromont, M. (2007). Model selection by bootstrap penalization for classification. *Mach. Learn.*, 66(2–3):165–207.

Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika*, 61(1):101–107.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328.

Girard, D. A. (1998). Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression. *Ann. Statist.*, 26(1):315–334.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, USA.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.*, 11(4):1156–1174.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15(4):1491–1519.

Hall, P., Lahiri, S. N., and Polzehl, J. (1995). On bandwidth choice in nonparametric regression with both short- and long-range dependent errors. *Ann. Statist.*, 23(6):1921–1936.

Hall, P., Marron, J. S., and Park, B. U. (1992). Smoothed cross-validation. *Probab. Theory Related Fields*, 92(1):1–20.

Hall, P. and Schucany, W. R. (1989). A local cross-validation algorithm. *Statist. Probab. Lett.*, 8(2):109–117.

Härdle, W. (1984). How to determine the bandwidth of some nonlinear smoothers in practice. In *Robust and nonlinear time series analysis (Heidelberg, 1983)*, volume 26 of *Lecture Notes in Statist.*, pages 163–184. Springer, New York.

Härdle, W., Hall, P., and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.*, 83(401):86–101. With comments by David W. Scott and Iain Johnstone and a reply by the authors.

Hart, J. D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *J. Roy. Statist. Soc. Ser. B*, 56(3):529–542.

Hart, J. D. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.*, 18(2):873–890.

Hart, J. D. and Wehrly, T. E. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.*, 81(396):1080–1088.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York. Data mining, inference, and prediction.

Herzberg, A. M. and Tsukanov, A. V. (1986). A note on modifications of jackknife criterion for model selection. *Utilitas Math.*, 29:209–216.

Herzberg, P. A. (1969). The parameters of cross-validation. *Psychometrika*, 34:Monograph Supplement.

Hesterberg, T. C., Choi, N. H., Meier, L., and Fraley, C. (2008). Least angle and l1 penalized regression: A review. *Statistics Surveys*, 2:61–93 (electronic).

Hills, M. (1966). Allocation Rules and their Error Rates. *J. Royal Statist. Soc. Series B*, 28(1):1–31.

Huber, P. (1964). Robust estimation of a local parameter. *Ann. Math. Statist.*, 35:73–101.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

John, P. W. M. (1971). *Statistical design and analysis of experiments*. The Macmillan Co., New York.

Jonathan, P., Krzanowki, W. J., and McCarthy, W. V. (2000). On the use of cross-validation to assess performance in multivariate prediction. *Stat. and Comput.*, 10:209–229.

Kearns, M., Mansour, Y., Ng, A. Y., and Ron, D. (1997). An Experimental and Theoretical Comparison of Model Selection Methods. *Machine Learning*, 27:7–50.

Kearns, M. and Ron, D. (1999). Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. *Neural Computation*, 11:1427–1453.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5):1902–1914.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656.

Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of Error Rates in Discriminant Analysis. *Technometrics*, 10(1):1–11.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *J. Edic. Psychol.*, 22:45–55.

Leung, D., Marriott, F., and Wu, E. (1993). Bandwidth selection in robust smoothing. *J. Nonparametr. Statist.*, 2:333–339.

Leung, D. H.-Y. (2005). Cross-validation in nonparametric regression with outliers. *Ann. Statist.*, 33(5):2291–2310.

Li, K.-C. (1985). From stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.*, 13(4):1352–1377.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975.

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

Markatou, M., Tian, H., Biswas, S., and Hripcsak, G. (2005). Analysis of variance of cross-validation estimators of the generalization error. *J. Mach. Learn. Res.*, 6:1127–1168 (electronic).

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307.

Mosteller, F. and Tukey, J. W. (1968). Data analysis, including statistics. In Lindzey, G. and Aronson, E., editors, *Handbook of Social Psychology, Vol. 2.* Addison-Wesley.

Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, 52:239–281.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, 12(2):758–765.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.*, 16(2):134–153.

Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *J. Amer. Statist. Assoc.*, 79(387):575–583.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *J. Roy. Statist. Soc. Ser. B.*, 11:68–84.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press.

Rissanen, J. (1983). Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416–431.

Ronchetti, E., Field, C., and Blanchard, W. (1997). Robust linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 92:1017–1023.

Rudemo, M. (1982). Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9:65–78.

Sauvé, M. (2009). Histogram selection in non gaussian regression. *ESAIM: Probability and Statistics*, 13:70–86.

Schuster, E. F. and Gregory, G. G. (1981). On the consistency of maximum likelihood nonparametric density estimators. In Eddy, W. F., editor, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 295–298. Springer-Verlag, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422):486–494.

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.*, 91(434):655–665.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica*, 7(2):221–264. With comments and a rejoinder by the author.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, 71(1):43–49.

Simon, F. (1971). Prediction methods in criminology. volume 7.

Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–35.

Sugiura, N. (1978). Further analysis of the data by akaike's information criterion and the finite corrections. *Comm. Statist. A—Theory Methods*, 7(1):13–26.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. Royal Statist. Soc. Series B*, 58(1):267–288.

van der Laan, M. J. and Dudoit, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Working Paper Series Working Paper 130, U.C. Berkeley Division of Biostatistics. available at http://www.bepress.com/ucbbiostat/paper130.

van der Laan, M. J., Dudoit, S., and Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Stat. Appl. Genet. Mol. Biol.*, 3:Art. 4, 27 pp. (electronic).

van der Laan, M. J., Dudoit, S., and van der Vaart, A. W. (2006). The cross-validated adaptive epsilon-net estimator. *Statist. Decisions*, 24(3):373–395.

van der Vaart, A. W., Dudoit, S., and van der Laan, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statist. Decisions*, 24(3):351–371.

van Erven, T., Grünwald, P. D., and de Rooij, S. (2008). Catching up faster by switching sooner: A prequential solution to the aic-bic dilemma. arXiv:0807.1005.

Vapnik, V. (1982). *Estimation of dependences based on empirical data*. Springer Series in Statistics. Springer-Verlag, New York. Translated from the Russian by Samuel Kotz.

Vapnik, V. N. (1998). *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Vapnik, V. N. and Chervonenkis, A. Y. (1974). *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya*. Izdat. "Nauka", Moscow. Theory of Pattern Recognition (In Russian).

Wahba, G. (1975). Periodic splines for spectral density estimation: The use of cross validation for determining the degree of smoothing. *Communications in Statistics*, 4:125–142.

Wahba, G. (1977). Practical Approximate Solutions to Linear Operator Equations When the Data are Noisy. *SIAM Journal on Numerical Analysis*, 14(4):651–667.

Wegkamp, M. (2003). Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

Yang, Y. (2006). Comparing learning methods for classification. *Statist. Sinica*, 16(2):635–657.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, 35(6):2450–2473.

Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Statist.*, 21(1):299–313.