Statistics Surveys
Vol. 4 (2010) 40–79
ISSN: 1935-7516
DOI: 10.1214/09-SS054

A survey of cross-validation procedures for model selection*

Sylvain Arlot†

CNRS; Willow Project-Team,
Laboratoire d'Informatique de l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
23 avenue d'Italie, F-75214 Paris Cedex 13, France
e-mail: sylvain.arlot@ens.fr

and

Alain Celisse†

Laboratoire de Mathématique Paul Painlevé
UMR 8524 CNRS - Université Lille 1,
59 655 Villeneuve d'Ascq Cedex, France
e-mail: 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 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.

AMS 2000 subject classifications: Primary 62G08; secondary 62G05, 62G09.
Keywords and phrases: Model selection, cross-validation, leave-one-out.

Received July 2009.

Contents

*This paper was accepted by Yuhong Yang, the Associate Editor for the IMS.
†The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-09-JCJC-0027-01.

1. Introduction

Likelihood maximization, least squares and empirical contrast minimization require to choose some model, that is, a set from which an estimator will be returned. Let us call *statistical algorithm* any function that returns an estimator from data—for instance, likelihood maximization on some given model. Then, *model selection* can be seen as a particular *(statistical) algorithm selection* problem.

Cross-validation (CV) is a popular strategy for algorithm selection. The main idea behind CV is to split data, once or several times, for estimating the risk of each algorithm: Part of 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 "universality" of the data splitting heuristics. Nevertheless, 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: 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 a given model selection problem?

The paper is organized as follows. First, the rest of Section 1 presents the statistical framework. Although non exhaustive, the 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 is given in Section 3; these are important for better understanding CV. The most classical CV procedures are defined in Section 4. Section 5 details the main properties of CV estimators of the risk for a fixed model; they are the keystone of any analysis of the model selection behaviour of CV. 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 or modifications of CV in several 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

Throughout the paper, $\xi_1, \dots, \xi_n \in \Xi$ denote some random variables with common distribution P (the observations). Except in Section 8.1, the ξ_is 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 density of P w.r.t. some measure μ, or the regression function. Let $\mathbb{S}$ denote the set of possible values for s. The quality of $t \in \mathbb{S}$, as an approximation to s, is measured by its loss $\mathcal{L}(t)$, where $\mathcal{L} : \mathbb{S} \mapsto \mathbb{R}$ is called the *loss function*; the loss is assumed to be minimal for $t = s$. Several loss functions can be chosen for a given statistical problem. Many of them are defined by

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

where $\gamma : \mathbb{S} \times \Xi \mapsto [0, \infty)$ is called a *contrast function*. For $t \in \mathbb{S}$, $\mathbb{E}_{\xi \sim P}[\gamma(t; \xi)]$ measures the average discrepancy between t and a new observation ξ with distribution P. 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. Given a loss function $\mathcal{L}_P(\cdot)$, two useful quantities are the *excess loss*

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

and the *risk of an estimator* $\widehat{s}(\xi_1, \dots, \xi_n)$ of the target s

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

1.2. Statistical problems

The following examples illustrate how general the framework of Section 1.1 is.

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

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

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 $(X_1, Y_1), \dots, (X_n, Y_n)$. 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 y and its predicted value $t(x)$. Two classical prediction frameworks are regression and classification, which are detailed below.

Regression corresponds to continuous 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

$$\forall i \in \{1,\dots,n\} \quad , \quad Y_i = s(X_i) + \varepsilon_i \qquad \text{with} \quad \mathbb{E}[\varepsilon_i \mid X_i] = 0 \ .$$

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

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

Note that the excess loss of t is the square of the $L^2(P)$ distance between t and s, so that *prediction* and *estimation* here 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 $\gamma(t;(x,y)) = \mathbf{1}_{t(x)\neq y}$, the minimizer of the loss is the so-called Bayes classifier s defined by

$$\forall x \in \mathcal{X} \quad , \quad s(x) = \mathbf{1}_{\eta(x)\geq 1/2} \ ,$$

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

Note that classification with convex losses—such as the hinge, exponential and logit losses—can also be put into the framework of Section 1.1. Many references on classification theory, including model selection, can be found in the survey by Boucheron et al. (2005).

1.3. Statistical algorithms and estimators

In this survey, a *statistical algorithm* $\mathcal{A}$ is any measurable mapping $\mathcal{A} : \bigcup_{n\in\mathbb{N}} \Xi^n \mapsto \mathbb{S}$. Given a sample $D_n = (\xi_i)_{1\leq i\leq n} \in \Xi^n$, 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.

Minimum contrast estimators refer to a classical family of statistical algorithms. Given some subset S of $\mathbb{S}$ called a *model*, a minimum contrast estimator over S is any $\widehat{s}(D_n) \in \mathbb{S}$ that minimizes over S the empirical contrast

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

and δ_x denotes the Dirac measure at point x.

The corresponding *minimum contrast algorithm*, associated with the model S, is the mapping $D_n \mapsto \widehat{s}(D_n)$. The idea is that the empirical contrast $\mathcal{L}_{P_n}(t)$ has an expectation $\mathcal{L}_P(t)$ which is minimal for $t = s$. 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 minimum contrast estimators:

- *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 the regular partition of $\Xi = [0,1]$ with K intervals.
- *Least-squares estimators*: take $\gamma(t;(x,y)) = (t(x)-y)^2$ 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 Fourier basis. Note that regularized least-squares algorithms such as the Lasso, ridge regression and spline smoothing are also least-squares estimators; S is some ball—with a radius depending on data and on the regularization parameter—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 increasing balls.
- *Empirical risk minimizers*: Following Vapnik (1982), this terminology applies to any contrast function γ and model S 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 are also empirical contrast minimizers over some data-dependent model S, with different "convex" contrasts γ.

Let us finally mention that many other estimators can be considered with CV. Classical ones are local averaging estimators such as k-Nearest Neighbours and Nadaraya-Watson kernel estimators, which are not minimum contrast estimators. The following mainly focuses on minimum contrast estimators for length reasons.

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 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 the output of each candidate algorithm is a minimum contrast estimator and the choice of an algorithm amounts to choose a model. 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 $\widehat{s}_m(D_n)$ be a minimum contrast estimator over S_m. The problem 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. When S_m is "too small", any $t \in S_m$ is a poor approximation to s, so that

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

is large for most $s \in \mathbb{S}$. The lower bound $\ell(s, S_m)$ is called the *approximation error*, or *bias* of model S_m. Thinking of nested models, the bias is a nonincreasing function of S_m.

Conversely, when S_m is "too large", $\ell(s, S_m)$ is small, but $\widehat{s}_m(D_n)$ is likely to overfit: This results from the *estimation error*. Think for instance of S_m as the set of all continuous functions on $\mathcal{X} = [0,1]$ in the regression framework. If S_m is a vector space of dimension D_m, it can be proved in several classical frameworks that,

$$\mathbb{E}[\ell(s, \widehat{s}_m(D_n))] \approx \ell(s, S_m) + \alpha_n D_m = \text{ Approx. error} + \text{Estimation error} \quad (2)$$

where $\alpha_n > 0$ does not depend on m. For instance, $\alpha_n = 1/(2n)$ in density estimation using the log-likelihood contrast, and $\alpha_n = \sigma^2/n$ in regression using the least-squares contrast and assuming $\text{var}(Y \mid X) = \sigma^2$ does not depend on X. More generally according to (2), a good model choice reaches the best trade-off between the *approximation error* $\ell(s, S_m)$ and the *estimation error* $\alpha_n D_m$, which is often called the *bias-variance trade-off* ($\alpha_n D_m$ is often called "variance"). By extension, the *estimation error*, can be defined by

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

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 distinguish the two main different goals that model selection can target: *estimation* and *identification*.

2.2. Model selection for estimation

The goal of model selection is *estimation* when $\widehat{s}_{\widehat{m}(D_n)}(D_n)$ is used for estimating s, and the goal is to minimize its loss. In particular, s is not required to belong to $\bigcup_{m \in \mathcal{M}_n} S_m$. For instance, AIC and Mallows' C_p are built for estimation (see Section 3.1).

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

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

Since $m^\star(D_n)$ depends on the unknown distribution P of data, one can only hope to select $\widehat{m}(D_n)$ such that $\widehat{s}_{\widehat{m}(D_n)}$ is almost as close to s as $\widehat{s}_{m^\star(D_n)}$.

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 is called *efficient* (or asymptotically optimal) when

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

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 a constant $C' > 0$). 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. Note that the oracle is often defined as $\arg\min_{m \in \mathcal{M}_n} \left\{ \mathbb{E}\left[\ell\left(s, \widehat{s}_m(D_n)\right) \right] \right\}$, leading to a weaker form of oracle inequality

$$\mathbb{E}\left[\ell\left(s, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right) \right] \leq C_n \inf_{m \in \mathcal{M}_n} \left\{ \mathbb{E}\left[\ell\left(s, \widehat{s}_m(D_n)\right) \right] \right\} + R_n \ .$$

Model selection procedures designed for estimation are often used for building *minimax adaptive estimators* provided the collection $(S_m)_{m \in \mathcal{M}_n}$ is well-chosen (Barron et al., 1999). This notion is closely related to efficiency.

2.3. Model selection for identification

Model selection can also 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 3.2.2).

The quality of a model selection procedure designed for identification is measured by its probability of recovering the true model S_{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 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 when $s \in \bigcup_{m \in \mathcal{M}_n} S_m$ (like BIC) and efficient when $s \notin \bigcup_{m \in \mathcal{M}_n} S_m$ (like AIC)? This problem is often called the *AIC-BIC dilemma*. Some strengths of AIC (efficiency) and BIC (model consistency) can sometimes be shared (see the introduction of Yang (2005), and van Erven et al. (2008)). However, Yang (2005) proved in the regression framework that no model selection procedure can be simultaneously *model consistent* (like BIC) and *adaptive in the minimax sense* (like AIC).

2.5. *Model selection* vs. *model averaging*

When the goal is estimation (Section 2.2), model selection can suffer some troubles due to *instability* in the choice of the algorithm (Yang, 2001): Any perturbation of original data entails the selection of a completely different algorithm. *Model averaging*, also called *aggregation* (Nemirovski, 2000), enables to remedy this deficiency by combining the outputs of several algorithms rather than selecting one of them. Note that the purpose of aggregation cannot be identification.

Oracle inequalities with leading constant $C_n = 1$ have been derived for instance for aggregation with exponential weights (Lecué, 2006). In some specific frameworks, Lecué (2007) has even shown the suboptimality of model selection with respect to aggregation. See also Hoeting et al. (1999) on Bayesian model averaging.

3. Overview of some model selection procedures

Let us sketch the properties of several model selection procedures, which will help understanding how CV works. Like CV, all procedures in this section select

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

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

A particular case of (5) is *penalization*. It consists in choosing the model minimizing the sum of the empirical contrast and some measure of complexity of the model (called penalty):

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

Following a classification made by Shao (1997) in the linear regression framework, we first focus on procedures of the form (5) such that

$$\operatorname{crit}(m; D_n) \approx \text{ Approx. error } + \kappa_n \, \text{Estim. error} \tag{7}$$

for some $\kappa_n \geq 1$.

3.1. The unbiased risk estimation principle ($\kappa_n \approx 1$)

The unbiased risk estimation principle—also called Mallows' or Akaike's heuristics—states that $\mathrm{crit}(m; D_n)$ in (5) should unbiasedly estimate $\mathbb{E}[\mathcal{L}_P(\widehat{s}_m(D_n))]$, that is, should satisfy (7) with $\kappa_n \approx 1$ (up to a term independent of m). Let us explain why this strategy can perform well when the goal of model selection is estimation. By (5), for every $m \in \mathcal{M}_n$,

$$\ell(s, \widehat{s}_{\widehat{m}}(D_n)) + \mathrm{crit}(\widehat{m}) - \mathcal{L}_P(\widehat{s}_{\widehat{m}}(D_n)) \leq \ell(s, \widehat{s}_m(D_n)) + \mathrm{crit}(m) - \mathcal{L}_P(\widehat{s}_m(D_n)) . \tag{8}$$

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

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

with high probability, at least when $\mathrm{Card}(\mathcal{M}_n) \leq Cn^\alpha$ for some $C, \alpha \geq 0$. Then, (8) implies an oracle inequality like (4) with $C_n = (1 + \varepsilon_n^+)/(1 - \varepsilon_n^-)$. If $\varepsilon_n^+, \varepsilon_n^- \to 0$ when $n \to \infty$, the procedure defined by (5) is efficient. Let us remark that $\varepsilon_n^+, \varepsilon_n^-$ mostly depend on the variance of $\mathrm{crit}(m) - \mathcal{L}_P(\widehat{s}_m(D_n))$, which is therefore important for precisely understanding the performance of $\widehat{m}$.

Examples of such model selection procedures 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 $\mathbb{E}[\mathrm{pen}(m)]$ should be close to the expectation of the "ideal penalty"

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

The main instances of penalization procedures following this principle are: AIC (Akaike's Information Criterion, Akaike, 1973), with the log-likelihood contrast; C_p and C_L (Mallows, 1973), and several refined versions of C_p (e.g., by Baraud, 2002), with the least-squares contrast; covariance penalties (Efron, 2004), with a general contrast. AIC, C_p and related procedures have been proved to be efficient or to satisfy oracle inequalities in several frameworks, provided $\mathrm{Card}(\mathcal{M}_n) \leq Cn^\alpha$ for some constants $C, \alpha \geq 0$ (see Birgé and Massart, 2007, and references therein).

The main drawback of penalties such as AIC or C_p is their asymptotic nature (for AIC) or their dependence on some assumptions on the distribution of data: C_p assumes the variance of Y does not depend on X. Otherwise, it has a suboptimal performance (Arlot, 2008a). In addition to CV, resampling-based penalties have been proposed to overcome this problem (Efron, 1983; Arlot, 2009); see Section 10.4 on this question.

3.2. Biased estimation of the risk ($\kappa_n > 1$)

Several model selection procedures do not follow the unbiased risk estimation principle. In Sections 3.2.1 and 3.2.2, we review procedures that nevertheless satisfy (7) with $\liminf_{n\to\infty} \kappa_n > 1$. From the penalization point of view, such procedures are *overpenalizing*.

Examples of such procedures are FPE_α (Bhansali and Downham, 1977) and the closely related GIC_λ (Generalized Information Criterion, Nishii, 1984; Shao, 1997) with $\alpha, \lambda > 2$ (the unbiased risk estimation principle suggests taking $\alpha = \lambda = 2$), and some CV procedures, such as Leave-p-out with $p/n \geq C > 0$ (see Sections 4.3.1 and 5.1).

3.2.1. Estimation

When the goal is estimation, there are two main reasons for using "biased" model selection procedures.

First, experimental evidence show that having $1 < \kappa_n = \mathcal{O}(1)$—that is, overpenalizing—often yields better performance when the signal-to-noise ratio is small (see for instance Arlot, 2007, Chapter 11).

Second, unbiased risk estimation fails when the number of models $\mathrm{Card}(\mathcal{M}_n)$ grows faster than any power of n, as in complete variable selection with p variables if $p \gg \ln(n)$. From the penalization point of view, Birgé and Massart (2007) proved that the minimal amount of penalization required so that an oracle inequality holds is much larger than $\mathrm{pen}_{\mathrm{id}}(m)$ when $\mathrm{Card}(\mathcal{M}_n) = e^{\lambda n}$, $\lambda > 0$: κ_n must at least be of order $\ln(n)$. In addition to FPE_α and GIC_λ with suitably chosen α, λ, procedures taking the size of $\mathcal{M}_n$ into account were proposed in several papers (for instance, Barron et al., 1999; Baraud, 2002; Birgé and Massart, 2001; Sauvé, 2009).

3.2.2. Identification ($\kappa_n \to +\infty$)

Some specific model selection procedures are used for identification, like BIC (Bayesian Information Criterion, Schwarz, 1978).

Shao (1997) showed that several procedures consistently identify the true model in the linear regression framework, as long as they satisfy (7) with $\kappa_n \to \infty$ when $n \to +\infty$. Instances of such procedures are 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) proved that m_n-out-of-n bootstrap penalization is also model consistent as long as $m_n \ll n$, which precisely means that $\kappa_n \to +\infty$ (Arlot, 2009).

Most MDL-based procedures can also be put into the category of model selection procedures built for identification (see Grünwald, 2007). Let us finally mention the Lasso (Tibshirani, 1996) and other ℓ^1 penalization procedures (see

for instance Hesterberg et al., 2008) for an appropriate choice of the regularization parameter. They are a computationally efficient way of identifying the true model in the context of variable selection with p variables, $p \gg n$.

3.2.3. Other approaches

Structural risk minimization was introduced by Vapnik and Chervonenkis (1974) in the context of statistical learning (see also Vapnik, 1982, 1998). Roughly, the idea is to penalize the empirical contrast with a penalty—often distribution-free—(over)-estimating

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

for instance using the Vapnik-Chervonenkis dimension, (global) Rademacher complexities (Koltchinskii, 2001; Bartlett et al., 2002) or global bootstrap penalties (Fromont, 2007). The recent *localization* approach leads to smaller upper bounds on $\mathrm{pen}_{\mathrm{id},g}$ (for references, see the survey by Boucheron et al., 2005).

***Ad hoc* penalization** consists in using features of the problem for building a penalty. For instance, the penalty can be proportional to some particular norm of $\widehat{s}_m$, when it is known this norm should be small. This strategy is similar to regularized learning algorithms such as the Lasso, kernel ridge regression or spline smoothing. More generally, any penalty can be used, as long as $\mathrm{pen}(m)$ is large enough to avoid overfitting, and the best choice for the final user is not the oracle $m^\star$, but more like

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

Note finally that completely different approaches exist for model selection, such as the Minimum Description Length (MDL, Rissanen, 1983), and the Bayesian approaches (Raftery, 1995). Interested readers will find more details on model selection procedures in the books by Burnham and Anderson (2002) or by Massart (2007) for instance.

3.3. 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 categories of Sections 3.1, 3.2.1 and 3.2.2: they all satisfy (7) with different values of κ_n. Therefore, studying the bias of CV as an estimator of the risk (as we do in Section 5.1) is crucial to deduce the model selection performances of CV, which are detailed in Sections 6 and 7.

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 specify the corresponding CV estimator of the risk of $\mathcal{A}(D_n) = \widehat{s}^{\mathcal{A}}(D_n)$.

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, starting from the remark that testing the output of the algorithm on new data would yield a good estimate of its performance (Mosteller and Tukey, 1968; Stone, 1974; Geisser, 1975).

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

A single data split yields a *validation* estimate of the risk, and averaging over several splits yields a *cross-validation* estimate. Various splitting strategies lead to various CV estimates of the risk as explained in Sections 4.2 and 4.3.

The major interest of CV lies in the universality of the data splitting heuristics. It only assumes that data are identically distributed, and training and validation samples are independent, which can even be relaxed (see Section 8.1). Therefore, CV can be applied to (almost) any algorithm in (almost) any framework, such as regression (Stone, 1974; Geisser, 1975), density estimation (Rudemo, 1982; Stone, 1984), and classification (Devroye and Wagner, 1979; Bartlett et al., 2002) among many others. This universality is not shared by most other model selection procedures (see Section 3), which often are specific of a framework and can be completely misleading in another one. For instance, C_p (Mallows, 1973) is specific of 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

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 given by

$$\widehat{\mathcal{L}}^{\mathrm{HO}}\left(\mathcal{A}; D_n; I^{(t)}\right) := \frac{1}{n_v} \sum_{i \in D_n^{(v)}} \gamma\left(\mathcal{A}(D_n^{(t)}); \xi_i\right) \quad , \tag{9}$$

where $D_n^{(t)} := (\xi_i)_{i \in I^{(t)}}$ is the *training sample*, of size $n_t = \mathrm{Card}(I^{(t)})$, $D_n^{(v)} := (\xi_i)_{i \in I^{(v)}}$ is the *validation sample*, of size $n_v = n - n_t$, and $I^{(v)}$ is called the *validation set*.

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 data splits. Let $B \geq 1$ be an integer and $I_1^{(t)}, \ldots, I_B^{(t)}$ 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 $(I_j^{(t)})_{1 \leq j \leq B}$, is defined by

$$\widehat{\mathcal{L}}^{\mathrm{CV}}\left(\mathcal{A}; D_n; \left(I_j^{(t)}\right)_{1 \leq j \leq B}\right) := \frac{1}{B} \sum_{j=1}^{B} \widehat{\mathcal{L}}^{\mathrm{HO}}\left(\mathcal{A}; D_n; I_j^{(t)}\right) \quad . \tag{10}$$

All usual CV estimators of the risk are of the form (10). Each one is uniquely determined by $(I_j^{(t)})_{1 \leq j \leq B}$, that is, the choice of the splitting scheme.

Note that in model selection for identification, an alternative definition of CV was proposed by Yang (2006, 2007), 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 $\widehat{\mathcal{L}}^{\mathrm{HO}}(\mathcal{A}_1; D_n; I_j^{(t)}) < \widehat{\mathcal{L}}^{\mathrm{HO}}(\mathcal{A}_2; D_n; I_j^{(t)})$ for a majority of the splits $j = 1, \ldots, B$. By contrast, CV procedures of the form (10) can be called "CV with averaging" (CV-a), since the estimates of the risk 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 $I^{(t)}$ —in particular its cardinality n_t —is discussed in the rest of this survey. In this subsection, two main categories of splitting schemes are distinguished, given n_t: *exhaustive data splitting*, that is considering all training sets 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. It corresponds 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 (10) with $B = n$ and $I_j^{(t)} = \{j\}^c$ for $j = 1, \dots, n$:

$$\widehat{\mathcal{L}}^{\mathrm{LOO}}(\mathcal{A}; D_n) = \frac{1}{n} \sum_{j=1}^{n} \gamma\left(\mathcal{A}\left(D_n^{(-j)}\right); \xi_j\right) \tag{11}$$

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); LOO has several other names in the literature, such as *delete-one CV* (Li, 1987), *ordinary CV* (Stone, 1974; Burman, 1989), or simply *CV* (Efron, 1983; Li, 1987).

Leave-p-out (LPO, Shao, 1993) with $p \in \{1, \dots, n-1\}$ is the exhaustive CV with $n_t = n - p$: Every possible subset of p data is successively "left out" of the sample and used for validation. Therefore, LPO is defined by (10) with $B = \binom{n}{p}$, and $(I_j^{(t)})_{1 \leq j \leq B}$ are all subsets of $\{1, \dots, n\}$ of size $n - p$. 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 when p is small. Partial data splitting schemes have been proposed as alternatives.

V-fold CV (VFCV) with $V \in \{1, \dots, n\}$ 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 data into V subsamples of approximately equal cardinality n/V. Each subsample successively plays the role of validation sample. Formally, let $A_1, \dots, A_V$ be some partition of $\{1, \dots, n\}$ with $\forall j$, $\mathrm{Card}(A_j) \approx n/V$. Then, the VFCV estimator of the risk of $\mathcal{A}(D_n)$ is given by (10) with $B = V$ and $I_j^{(t)} = A_j^c$ for $j = 1, \dots, B$:

$$\widehat{\mathcal{L}}^{\mathrm{VF}}\left(\mathcal{A}; D_n; (A_j)_{1 \leq j \leq V}\right) = \frac{1}{V} \sum_{j=1}^{V} \left[\frac{1}{\mathrm{Card}(A_j)} \sum_{i \in A_j} \gamma\left(\widehat{s}\left(D_n^{(-A_j)}\right); \xi_i\right) \right] \tag{12}$$

where $D_n^{(-A_j)} = (\xi_i)_{i \in A_j^c}$. The computational cost of VFCV is only V times that of training $\mathcal{A}$ with $n - n/V$ points; it 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 when the training sample size n_t is small. Indeed, BICV is defined by (10) with training sets $(A^c)_{A \in \mathcal{T}}$, where $\mathcal{T}$ is a balanced incomplete block design (BIBD, John, 1971), that is, a collection of $B > 0$ subsets of $\{1, \dots, n\}$ of size $n_v = n - n_t$ such that:

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

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 satisfies the first property and almost satisfies the second one.

Repeated learning-testing (RLT) was introduced by Breiman et al. (1984) and further studied by Burman (1989) and Zhang (1993). The RLT estimator of the risk of $\mathcal{A}$ is defined by (10) with any $B > 0$, and any sequence $(I_j^{(t)})_{1 \leq j \leq B}$ of different subsets of $\{1, \dots, n\}$, chosen randomly, without replacement, and independently of 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. Unlike RLT, MCCV allows the same split to be chosen several times.

4.3.3. Other cross-validation-like risk estimators

Several procedures have been designed to fix possible drawbacks of CV.

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

Generalized CV (GCV, Craven and Wahba, 1979) was introduced in least-squares regression, as a rotation-invariant version of LOO, for estimating the risk of a linear estimator $\widehat{s}(D_n) = M\mathbf{Y}$, 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 (*i.e.*, C_p generalized to linear estimators; Mallows, 1973) than to CV, since GCV can be seen as an approximation to C_L with a particular estimator of variance (Efron, 1986). The efficiency of GCV has been proved in various frameworks (Li, 1985, 1987; Cao and Golubev, 2006).

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

LOO bootstrap and .632 bootstrap. The bootstrap is often used for stabilizing 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 *LOO bootstrap*. Noticing the bias of LOO bootstrap, Efron (1983) proposed a heuristic argument leading to the .632 *bootstrap* estimator, later modified into .632+ *bootstrap* by Efron and Tibshirani (1997). However, these procedures have nearly no theoretical justification and 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 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). It 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 are distinguished.

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

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). LOO, as a model selection procedure, was first studied by 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

As noticed in Section 3, understanding the behaviour of CV for model selection—which is the purpose of this survey—requires to analyze the performances of CV as an estimator of the risk of a single algorithm, at least in terms of bias and variance.

5.1. Bias

Analyzing the bias of CV enables to minimize or to correct this bias (following Section 3.1); alternatively, when some bias is needed (see Section 3.2), such an analysis allows to tune the bias of CV as desired.

5.1.1. Theoretical assessment of bias

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

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

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

$$\mathbb{E}\left[\widehat{\mathcal{L}}^{\mathrm{CV}}\left(\mathcal{A}; D_n; \left(I_j^{(t)}\right)_{1 \le j \le B}\right)\right] = \mathbb{E}\left[\mathcal{L}_P\left(\mathcal{A}\left(D_{n_t}\right)\right)\right] \ . \tag{13}$$

According to (13), the bias of the CV estimator of the risk of $\mathcal{A}(D_n)$ is the difference between the risks of $\mathcal{A}$ respectively trained with n_t and with n observations. Since $n_t < n$, the bias of CV is usually nonnegative and tends to decrease when n_t increases. This holds true when the risk of $\mathcal{A}(D_n)$ 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).

More precisely, (13) 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), and empirical results. They are organized below by statistical framework.

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

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 derived by Rudemo (1982); see Bowman (1984) for simulations. Hall (1987) provided similar results with the log-likelihood contrast for kernel estimators; he related the performance of LOO to the interaction between the kernel and the tails of the target density s.

Classification. For discriminating between two populations with shifted distributions, Davison and Hall (1992) compared the asymptotical bias of LOO and bootstrap. LOO is less biased 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 synthetic and real 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. Furthermore, .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 algorithms $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ 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 $\mathcal{A}_{\widehat{\lambda}}(D_n)$ (see Stone (1974) for the LOO, and Jonathan et al. (2000) for VFCV). 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 for the same reason as the empirical contrast $\mathcal{L}_{P_n}(\widehat{s}(D_n))$ suffers some *optimism* as an estimator of the loss of $\widehat{s}(D_n)$. Estimating the risk of $\mathcal{A}_{\widehat{\lambda}}(D_n)$ with CV can be done by considering the full algorithm $\mathcal{A}' : D_n \mapsto \mathcal{A}_{\widehat{\lambda}(D_n)}(D_n)$, and then computing $\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A}'; D_n)$. This procedure is called "double cross" by Stone (1974).

5.1.2. Bias correction

An alternative to choosing the CV estimator with the smallest bias is to correct this bias. Burman (1989, 1990) proposed a corrected VFCV estimator

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

A similar correction holds for RLT. Both estimators have been proved to be asymptotically unbiased for least squares in linear regression.

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

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

The V-fold penalized criterion was proved by Arlot (2008b) to be (almost) unbiased in the non-asymptotic framework for regressograms.

5.2. Variance

With training sets of the same size n_t, CV estimators have the same bias, but still behave 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 Nadeau and Bengio (2003),

$$\begin{aligned}
&\mathrm{var}\left[\widehat{\mathcal{L}}^{\mathrm{HO}}\left(\mathcal{A}; D_n; I^{(t)}\right)\right] \\
&= \mathbb{E}\left[\mathrm{var}\left(\mathcal{L}_{P_n^{(v)}}\left(\mathcal{A}(D_n^{(t)})\right) \,\middle|\, D_n^{(t)}\right)\right] + \mathrm{var}\left[\mathcal{L}_P\left(\mathcal{A}(D_{n_t})\right)\right] \\
&= \frac{1}{n_v}\mathbb{E}\left[\mathrm{var}\left(\gamma\left(\widehat{s}, \xi\right) \,\middle|\, \widehat{s} = \mathcal{A}(D_n^{(t)})\right)\right] + \mathrm{var}\left[\mathcal{L}_P\left(\mathcal{A}(D_{n_t})\right)\right] \quad . \qquad (14)
\end{aligned}$$

Assuming n_t is fixed, the first term is proportional to $1/n_v$. Therefore, more data for validation decreases the variance of $\widehat{\mathcal{L}}^{\mathrm{HO}}$, because it yields a better estimator of $\mathcal{L}_P(\mathcal{A}(D_n^{(t)}))$. Both terms show that the variance of $\widehat{\mathcal{L}}^{\mathrm{HO}}$ also depends on the distribution of $\mathcal{L}_P(\mathcal{A}(D_n^{(t)}))$ 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., 2009; Breiman, 1996). Conversely, Molinaro et al. (2005) noticed, from a simulation experiment, that this trend disappears when $\mathcal{A}$ is *stable*. The relation between the stability of $\mathcal{A}$ and the variance of $\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A})$ was stressed by Devroye and Wagner (1979) in classification, through upper bounds on the variance of $\widehat{\mathcal{L}}^{\mathrm{LOO}}(\mathcal{A})$; see also Bousquet and Elisseff (2002) for extended results in the regression setting. Note that various techniques have been proposed for reducing the variance of $\widehat{\mathcal{L}}^{\mathrm{LOO}}(\mathcal{A})$, see Section 4.3.3.

Partial splitting and variance. When (n_t, n_v) is fixed, the variance of CV tends to be larger for partial data splitting methods. Choosing $B < \binom{n}{n_t}$ 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 $n_t = n - p$. The variability due to the choice of the data splits is maximal for hold-out, and minimal (null) for exhaustive splitting schemes like LOO (if $n_t = n - 1$) and LPO (with $p = n - n_t$). With MCCV, this variability decreases like B^{-1} since the $I_j^{(t)}$ are chosen independently. The dependence on B is different for other CV estimators, such as RLT or VFCV, because the $I_j^{(t)}$s are not independent.

Note that the dependence of $\mathrm{var}(\widehat{\mathcal{L}}^{\mathrm{VF}}(\mathcal{A}))$ on V 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 variance

Precisely understanding how $\mathrm{var}(\widehat{\mathcal{L}}^{\mathrm{CV}}(\mathcal{A}))$ depends on the splitting scheme is complex in general, since $n_t + n_v = n$, and the number of splits B is generally linked with n_t (for instance, for LPO and VFCV). Furthermore, the variance of CV strongly depends on the framework and on the stability of $\mathcal{A}$. Therefore, radically different results have been obtained in different frameworks, in particular on the value of V for which the VFCV estimator has a minimal variance (Burman, 1989; Hastie et al., 2009, Section 7.10). Despite these difficulties, the variance of several CV estimators has been assessed in various frameworks, as detailed below.

Regression. In a simple linear regression setting with homoscedastic data, Burman (1989) proved an asymptotic expansion of the variance of VFCV

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

Asymptotically, the variance decreases with V, implying that LOO asymptotically has the minimal variance among VFCV estimators. Similar results have been derived for RLT as well.

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

Another argument for the small variance of LOO in regression was provided by Davies et al. (2005), with the log-likelihood contrast: assuming a well-specified parametric model is available, the LOO estimator of the risk is the minimum variance unbiased estimator of its expectation.

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

Classification. For 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. (2009) empirically showed that VFCV has a minimal variance for some $2 < V < n$, whereas LOO usually has a large variance. Simulation experiments by Molinaro et al. (2005) suggest this fact mostly depends on the stability of the considered algorithm.

5.2.3. Variance estimation

There is no universal—valid under all distributions—unbiased estimator of the variance of RLT (Nadeau and Bengio, 2003) and VFCV (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.

Despite these negative results, (biased) estimators of the variance of $\widehat{\mathcal{L}}^{\mathrm{CV}}$ in regression and classification have been proposed and assessed by Nadeau and Bengio (2003), Bengio and Grandvalet (2004), and Markatou et al. (2005). In the density estimation framework, Celisse and Robin (2008) proposed an estimator of the variance of the LPO risk estimator based on closed-form formulas (see also Celisse (2008a) for extended results to projection estimators).

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. Risk estimation and model selection

As shown in Section 3, the model selection performances of CV mostly depend on two factors. The first one is the bias of CV as an estimator of the risk; in particular, when the collection of models is not too large, minimizing an unbiased estimator of the risk leads to an efficient model selection procedure. The second factor, usually less important—at least asymptotically—, is the variance of CV as an estimator of the risk. One could conclude 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. According to Breiman and Spector (1992) the best risk estimator is LOO, whereas 10-fold CV is more accurate for model selection. Such a difference mostly comes from three reasons. First, the asymptotic framework ($\mathcal{A}$ fixed, $n \to \infty$) may not apply to models close to the oracle. Second, as explained in Section 3.2, estimating the risk of each model with some bias can compensate the effect of a large variance, for instance when the signal-to-noise ratio of data is small. Third, what really matters in model selection is that

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

with the largest probability, for all m_1, m_2 "close to" $m^\star(D_n)$.

Therefore, specific studies are required to evaluate the performances of CV procedures in terms of model selection efficiency.

6.2. The big picture

In several frameworks such as least-squares regression on linear models, the estimation error depends on the sample size only through a factor n^{-1}. Then, Section 5.1 shows that CV satisfies (7) with $\kappa_n = n/n_t$. Thus, according to Section 3, the efficiency of CV mostly 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 C_p with a penalty multiplied by $\kappa/2$. Such CV procedures are overpenalizing by a factor $(1+\lambda)/(2\lambda) > 1$.

The above results have been proved in linear regression by Shao (1997) for LPO (see also Li (1987) for LOO, and Zhang (1993) for RLT when $B \gg n^2$.

In a general statistical framework, the model selection performance of several CV-based procedures applied to minimum contrast estimation algorithms 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). An oracle-type inequality is proved, showing that up to a multiplying factor $C_n \to 1$, the risk of the algorithm selected by CV is smaller than the risk of the oracle with n_t observations $m^\star(D_{n_t})$. In most frameworks, this implies the asymptotic optimality of CV as long as $n_t/n = \mathcal{O}(1)$. When $n_t \sim \lambda n$ with $\lambda \in (0,1)$, this 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, including the problem of bandwidth choice for kernel estimators.

Regression. First, Section 6.2 suggests the suboptimality of CV when n_t is not asymptotically equivalent to n, which has been settled with regressograms by Arlot (2008b) for VFCV when $V = \mathcal{O}(1)$. Note however that the best V for VFCV is not always the largest one (see Breiman and Spector, 1992; Herzberg and Tsukanov, 1986). Breiman (1996) proposed to explain this phenomenon by relating the stability of the candidate algorithms to 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 (Arlot, 2008b). Despite its suboptimality, VFCV with $V = \mathcal{O}(1)$ satisfies an oracle inequality with constant $C > 1$. V-fold penalization (which often coincides with corrected VFCV, see Section 5.1.2) satisfies an oracle inequality with $C_n \to 1$ as $n \to +\infty$, both when $V = \mathcal{O}(1)$ (Arlot, 2008b) and when $V = n$ (Arlot, 2009). Note that n-fold penalization is very close to LOO, suggesting that LOO is also asymptotically optimal with heteroscedastic data. Simulation

experiments in the context of change-point detection confirmed that CV adapts to heteroscedasticity, unlike usual model selection procedures in the same framework (Arlot and Celisse, 2009).

The performances of CV have also been assessed for other kinds of statistical algorithms in regression. For choosing the number of knots in spline smoothing, corrected versions of VFCV and RLT are asymptotically optimal provided $n/(Bn_v) = \mathcal{O}(1)$ (Burman, 1990). Härdle et al. (1988) and Girard (1998) compared several CV methods to GCV for choosing the bandwidth of kernel estimators; 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). An oracle inequality, with constant $C > 1$, has been derived by Wegkamp (2003) for hold-out with least squares.

Density estimation. CV performs as in regression for selecting among least-squares density estimators (van der Laan et al., 2004). In particular, non-asymptotic oracle inequalities with constant $C > 1$ have been proved by Celisse (2008a,b) 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 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 binary classification with piecewise constant classifiers, Kearns et al. (1997) proved an oracle inequality for hold-out. Empirical experiments show that hold-out yields (almost) always the best performance compared to deterministic penalties (Kearns et al., 1997), but is less accurate than random penalties, such as Rademacher complexity or maximal discrepancy, due to large variability (Bartlett et al., 2002).

Nevertheless, hold-out still enjoys *adaptivity* 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).

The performance of 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", which describe the worst-case performance of LOO (see also Bousquet and Elisseff, 2002).

For comparisons of CV and bootstrap-based CV in classification, see papers by Efron (1986) and by Efron and Tibshirani (1997).

7. Cross-validation for identification

This section tackles the properties of CV when the goal is identification, as defined in Section 2.3.

7.1. General conditions towards model consistency

The use of CV for *identification* may seem strange, since the pioneering LOO procedure is closely related to unbiased risk estimation, which is only efficient when the goal is *estimation*. Furthermore, estimation and identification are somehow contradictory goals, see Section 2.4.

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

With ordered variable selection in linear regression, Zhang (1993) computed the asymptotic value of $\mathbb{P}\left(\widehat{m}(D_n) = m_0\right)$, and numerically compared several CV procedures in a specific example. For LPO with $p/n \to \lambda \in (0,1)$ as n tends to $+\infty$, $\mathbb{P}\left(\widehat{m}(D_n) = m_0\right)$ increases with λ. However for VFCV, $\mathbb{P}\left(\widehat{m}(D_n) = m_0\right)$ increases with V and is therefore maximal for the LOO, which is the worst case of LPO: The variability induced by the V splits seems more important here than the bias of VFCV. Besides, $\mathbb{P}\left(\widehat{m}(D_n) = m_0\right)$ is almost constant between $V = 10$ and $V = n$, so that taking $V > 10$ is not advised for computational reasons.

As in Section 6.2, let us assume that CV satisfies (7) with $\kappa_n = n/n_t$, as in several frameworks. Then, the results of Section 3.2.2 suggest that model consistency can be obtained for CV if $n_t \ll n$. This was theoretically confirmed by Shao (1993, 1997) for the variable selection problem in linear regression: RLT, BICV (defined in Section 4.3.2), and LPO with $p = p_n \sim n$ and $n - p_n \to +\infty$ lead to *model consistency*.

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

7.2. Refined analysis for the algorithm selection problem

Let us consider the algorithm selection problem: Identifying the true model is then replaced by identifying the algorithm with the fastest convergence rate. Yang (2006, 2007) considered this problem for two (or more generally any finite number of) candidate algorithms. Note that Stone (1977) also 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 model consistency of CV strongly depends on the convergence rates $(r_{n,i})_{i=1,2}$ of the two candidate algorithms. Intuitively, consistency holds as long as the uncertainty of each estimate of the risk (roughly proportional to $n_v^{-1/2}$) is negligible relative to the risk gap $|r_{n_t,1} - r_{n_t,2}|$. 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. For instance, in the regression framework, if the risk of $\widehat{s}_i$ is measured by $\mathbb{E}\left\|\widehat{s}_i - s\right\|_2$, Yang (2007) proved that hold-out, VFCV, RLT, and LPO with voting (CV-v, see Section 4.2.2) are consistent in selection if

$$n_v, n_t \to +\infty \quad \text{and} \quad \sqrt{n_v} \max_{i=1,2} r_{n_t,i} \to +\infty \ , \tag{15}$$

under some conditions on $\left\|\widehat{s}_i - s\right\|_p$ for $p = 2, 4, \infty$ (see also Yang (2006) for similar results in classification).

The sufficient condition (15) can be simplified depending on $\max_i r_{n,i}$. On the one hand, if $\max_i r_{n,i} \propto n^{-1/2}$, then (15) holds when $n_v \gg n_t \to \infty$. 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, (15) is implied by $n_t/n_v = \mathcal{O}(1)$, when $\max_i r_{n,i} \gg n^{-1/2}$. It even allows $n_t \sim n$ (under some conditions). Therefore, non-parametric algorithms can be compared by more usual CV procedures ($n_t > n/2$), even if LOO is still excluded by conditions (15).

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; note however that $n_t = n - n/V \geq n/2$ for VFCV, so that (15) does not hold for any V if $\max_i r_{n,i} = \mathcal{O}(n^{-1/2})$. See also the results by Nadeau and Bengio (2003) about CV considered as a testing procedure comparing two candidate algorithms.

Note that according to 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

This section tackles frameworks where modifying CV can be necessary, in particular because the main assumptions of the CV heuristics are not satisfied.

8.1. Time series and dependent observations

When data are dependent, all previous analyses break down since the validation and training samples are no longer independent. Therefore, CV must be

modified. We refer to the review by Opsomer et al. (2001) on model selection in non-parametric regression with dependent data.

Let us consider the statistical framework of Section 1 with $\xi_1, \dots, \xi_n$ identically distributed but not independent. When 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 ξ_i and ξ_j are independent. For instance, LOO can be changed into taking $I^{(v)} = \{ J \}$ and $I^{(t)} = \{ 1, \dots, J-h-1, J+h+1, \dots, n \}$, where J is uniformly chosen in $\{ 1, \dots, n \}$. This method is called "modified CV" by Chu and Marron (1991) in the context of bandwidth selection. 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 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). Alternatives to modified CV have been proposed in various frameworks by Burman et al. (1994), by Burman and Nolan (1992), by Chu and Marron (1991) and by Hart (1994). Nevertheless, CV without modification was proved to be asymptotically optimal when $\xi_1, \dots, \xi_n$ is a stationary Markov process in a specific framework (Burman and Nolan, 1992).

8.2. Large number of models

The unbiased risk estimation principle (see Section 3.1) is known to fail when the number of models grows exponentially with n (Birgé and Massart, 2007). Therefore, the analysis of Section 6 is no longer valid, and n_t must be carefully chosen for avoiding overfitting (see Celisse, 2008b, Chapter 6).

For least-squares regression with homoscedastic data, Wegkamp (2003) proposed to add a penalty term to the hold-out estimator to penalize for the number of models. The resulting procedure satisfies an 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 $\mathcal{L}_{P_n}(t)$. 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, unlike other model selection procedures built for exponential collections of models.

8.3. 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 a robust loss function. In presence of outliers, 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 (Ronchetti et al., 1997).

8.4. Density estimation

Hall et al. (1992) defined the "smoothed CV", which consists in pre-smoothing the data before using CV, an idea related to the smoothed bootstrap. Under various smoothness conditions on the density, this procedure yields excellent asymptotic model selection performances.

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.

9. Closed-form formulas and fast computation

Resampling strategies, like CV, are known to be time consuming. The naive implementation of CV with B data splits has a computational complexity of B times that of training each algorithm $\mathcal{A}$. This can be prohibitive, even for rather small B (say, $B = 10$), depending on the problem. 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 extended by Celisse and Robin (2008) to the LPO risk with the quadratic loss. Similar results are available for projection estimators as settled by Celisse (2008a).

For least squares in linear regression, Zhang (1993) proved a closed-form formula for the LOO estimator of the risk. See Wahba (1975, 1977) and Craven and Wahba (1979) for similar results in the spline smoothing context. These papers led to the definition of GCV (Section 4.3.3) and related procedures, which are often used because of their small computational cost, as emphasized by Girard (1998).

Closed-form formulas for the LPO estimator are also given by Celisse (2008b) in regression for kernel and projection estimators.

When no closed-form formula can be proved, some efficient algorithms avoid recomputing $\widehat{\mathcal{L}}^{\mathrm{HO}}(\mathcal{A}; D_n; I_j^{(t)})$ from scratch for each data split $I_j^{(t)}$. Such algorithms rely on updating formulas like the ones by Ripley (1996) for LOO in

linear and quadratic discriminant analysis. Note that very similar formulas are also available for LOO and the k-nearest neighbours algorithm in classification (Daudin and Mary-Huard, 2008).

When CV is used for selecting among regressograms, an important property of the CV estimators of the risk (computed thanks to a closed-form formula or not) is their "additivity": For a regressogram associated with a partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$, any CV estimator defined by (10) is 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 $\widehat{\mathcal{L}}^{\mathrm{CV}}$ over the set of all partitions of $\mathcal{X}$ with a given number of pieces, a strategy successfully applied by Arlot and Celisse (2009) for change-point detection.

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

This conclusion collects a few guidelines to help CV users interpreting the results of CV, and appropriately using CV in each specific setting.

10.1. The big picture

Drawing a general conclusion on CV is nearly an impossible task because of the variety of frameworks. Nevertheless, we can still point out the three main criteria to take into account for choosing a CV procedure for a particular model selection problem:

- *Bias*: CV roughly estimates the risk of an algorithm trained with a sample size $n_t < n$ (see Section 5.1). Usually, this implies that CV overestimates the *estimation error* compared to the *approximation error* in decomposition (2) with sample size n.
  When the goal is estimation and the signal-to-noise ratio (SNR) is large, the smallest bias usually is the best, 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 estimation error often improves the performance, which is obtained by taking $n_t \sim \kappa n$ with $\kappa \in (0,1)$ (Section 6).
  When the goal is identification, a large bias is often needed, which is obtained by taking $n_t \ll n$. Larger values of n_t can also lead to model consistency (see Section 7).
- *Variance*: Model selection performances of CV usually are optimal when $\mathrm{var}(\widehat{\mathcal{L}}^{\mathrm{CV}})$ is as small as possible. This variance usually decreases when the number B of splits increases, with a fixed training sample size n_t. When B is fixed, the variance of CV also depends on 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 variance 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 variance seems strongly framework-dependent (see Section 5.2 for details).

- *Computational complexity*: Closed-form formulas or analytic approximations are available in several frameworks (see Section 9). Otherwise, the computational complexity of CV is roughly proportional to the number of data splits: 1 for 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 on the computational complexity of each algorithm, on specificities of the framework, 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 balancing bias and variance. 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 should the splits 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? 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. With regressograms, this led Arlot (2008b) 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 in every validation sample is the same as in the sample (which should be done carefully since it strongly breaks the CV heuristics). Nevertheless, Breiman and Spector (1992) compared several splitting strategies by simulations in regression. No significant improvement was reported from taking into account stratification of data.

Another question related to the choice of $(I_j^{(t)})_{1\leq j\leq B}$ is whether the $I_j^{(t)}$s 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. 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 splitting strategy is the best: 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). Moreover, estimating the variance of VFCV is a difficult task in general (Section 5.2.3).

When the goal of model selection is estimation, it is often reported that the optimal V is between 5 and 10, because the statistical performance does not increase a lot for larger values of V, and averaging over less than 10 splits remains computationally feasible (Hastie et al., 2009, Section 7.10). Even if this claim is true for many problems, this survey concludes 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 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. Conversely, when the SNR is small, overpenalization can improve the performance. Therefore, VFCV with $V < n$ can select an algorithm with a smaller risk than LOO does (see simulation experiments by Arlot, 2008b). 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 *model consistency* (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 ordered variable selection, the largest V seems to be the better, $V = 10$ providing results close to the optimal ones (Zhang, 1993). However, Dietterich (1998) and Alpaydin (1999) recommend $V = 2$ for choosing the best among two learning procedures.

10.4. Cross-validation or penalized criteria?

A natural question is whether penalized criteria—in particular the most classical ones, such as AIC and C_p—should be preferred to CV for a given model selection problem. The strongest argument for CV is its quasi-universality: Provided data are i.i.d., CV yields good model selection performances in (almost) any framework. Nevertheless, universality has a price: Compared to procedures designed to be optimal in a specific framework (like AIC), the model selection performances of CV can be less accurate, while its computational cost is higher. In least-squares regression (with $\mathrm{Card}(\mathcal{M}_n)$ not growing too fast with n), C_p often outperforms CV when data are homoscedastic. Otherwise, C_p is no longer efficient and can strongly overfit (Arlot, 2008a). Therefore, when homoscedasticity is questionable, CV should be preferred to C_p. More generally, because of its versatility, CV should be prefered to any model selection procedure relying on assumptions which are likely to be wrong.

Conversely, penalization procedures may seem more convenient than CV, because they allow to choose more easily the value of κ_n in (7), which is crucial for optimizing model selection performances (see Section 3). Then, resampling-based penalties (Efron, 1983; Arlot, 2009)—in particular V-fold penalties (Arlot, 2008b)—are natural alternatives to CV. The resampling heuristics on which they rely is almost as universal as the CV heuristics. The main drawback of resampling-based penalties, compared to CV, may only be that fewer theoretical studies exist about their model selection performances.

10.5. 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 with respect to n_t, the number B of splits, and the way 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 real distinction between hold-out and CV procedures having the same training sample size n_t: They are all equivalent at first order. However, second order terms do matter for realistic values of n, which shows the dramatic need for theory to take into account the variance of CV when comparing CV procedures 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. MR0286233

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. MR0483125

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127. MR0343481

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 Cady, F. B. (1972). Selection 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. MR0418296

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

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

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

Arlot, S. (2009). Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624 (electronic). MR2519533

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

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic). MR1918295

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413. MR1679028

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

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). MR2248010

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. MR0494751

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268. MR1848946

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73. MR2288064

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. MR2329460

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375 (electronic). MR2182250

Bousquet, O. and Elisseff, A. (2002). Stability and Generalization. *J. Machine Learning Research*, 2:499–526. MR1929416

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360. MR0767163

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24(6):2350–2383. MR1425957

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. MR0726392

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. MR1040644

Burman, P. (1990). Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā Ser. A*, 52(3):314–345. MR1178041

Burman, P., Chow, E., and Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, 81(2):351–358. MR1294896

Burman, P. and Nolan, D. (1992). Data-dependent estimation of prediction functions. *J. Time Ser. Anal.*, 13(3):189–207. MR1168164

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference*. Springer-Verlag, New York, second edition. A practical information-theoretic approach. MR1919620

Cao, Y. and Golubev, Y. (2006). On oracle inequalities related to smoothing splines. *Math. Methods Statist.*, 15(4):398–414. MR2301659

Celisse, A. (2008a). Model selection in density estimation via cross-validation. 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. MR2411944

Chow, Y. S., Geman, S., and Wu, L. D. (1987). Consistent cross-validated density estimation. *Ann. Statist.*, 11:25–38. MR0684860

Chu, C.-K. and Marron, J. S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, 19(4):1906–1918. MR1135155

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. MR0516581

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. MR2424787

Davies, S. L., Neath, A. A., and Cavanaugh, J. E. (2005). Cross validation model selection criteria for linear regression based on the Kullback-Leibler discrepancy. *Stat. Methodol.*, 2(4):249–266. MR2205599

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. MR1185130

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. MR1383093

Devroye, L. and Wagner, T. J. (1979). Distribution-Free performance Bounds for Potential Function Rules. *IEEE Transaction in Information Theory*, 25(5):601–604. MR0545015

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. MR0711106

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.*, 81(394):461–470. MR0845884

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. MR2090899

Efron, B. and Morris, C. (1973). Combining possibly related estimation problems (with discussion). *J. R. Statist. Soc. B*, 35:379. MR0381112

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *J. Amer. Statist. Assoc.*, 92(438):548–560. MR1467848

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. MR0418322

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. MR1608164

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. MR0720261

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15(4):1491–1519. MR0913570

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. MR1389858

Hall, P., Marron, J. S., and Park, B. U. (1992). Smoothed cross-validation. *Probab. Theory Related Fields*, 92(1):1–20. MR1156447

Hall, P. and Schucany, W. R. (1989). A local cross-validation algorithm. *Statist. Probab. Lett.*, 8(2):109–117. MR1017876

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. MR0786307

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. MR0941001

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. MR1278225

Hart, J. D. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.*, 18(2):873–890. MR1056341

Hart, J. D. and Wehrly, T. E. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.*, 81(396):1080–1088. MR0867635

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York. Data mining, inference, and prediction. 2nd edition. MR1851606

Herzberg, A. M. and Tsukanov, A. V. (1986). A note on modifications of jackknife criterion for model selection. *Utilitas Math.*, 29:209–216. MR0846203

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). MR2520981

Hills, M. (1966). Allocation Rules and their Error Rates. *J. Royal Statist. Soc. Series B*, 28(1):1–31. MR0196879

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: A tutorial. *Statistical Science*, 14(4):382–417. MR1765176

Huber, P. (1964). Robust estimation of a local parameter. *Ann. Math. Statist.*, 35:73–101. MR0161415

John, P. W. M. (1971). *Statistical design and analysis of experiments*. The Macmillan Co., New York. MR0273748

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. MR1842526

Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of Error Rates in Discriminant Analysis. *Technometrics*, 10(1):1–11. MR0223016

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

Lecué, G. (2006). Optimal oracle inequality for aggregation of classifiers under low noise condition. In Gabor Lugosi, H. U. S., editor, *19th Annual Conference On Learning Theory, COLT06.*, pages 364–378. Springer. MR2280618

Lecué, G. (2007). Suboptimality of penalized empirical risk minimization in classification. In *COLT 2007*, volume 4539 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin. MR2397584

Leung, D., Marriott, F., and Wu, E. (1993). Bandwidth selection in robust smoothing. *J. Nonparametr. Statist.*, 2:333–339. MR1256384

Leung, D. H.-Y. (2005). Cross-validation in nonparametric regression with outliers. *Ann. Statist.*, 33(5):2291–2310. MR2211087

Li, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.*, 13(4):1352–1377. MR0811497

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. MR0902239

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). MR2249851

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. MR2319879

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.

Nemirovski, A. (2000). Topics in Non-Parametric Statistics. In Bernard, P., editor, *Lecture Notes in Mathematics*, Lectures on Probability Theory and Statistics, Ecole d'ete de Probabilities de Saint-Flour XXVIII - 1998. M. Emery, A. Nemirovski , D. Voiculescu. MR1775638

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, 12(2):758–765. MR0740928

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.*, 16(2):134–153. MR1861070

Picard, R. R. and Cook, R. D. (1984). Cross-validation of regression models. *J. Amer. Statist. Assoc.*, 79(387):575–583. MR0763576

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York. MR1707286

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *J. Roy. Statist. Soc. Ser. B.*, 11:68–84. MR0032176

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, 25:111–163.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press. MR1438788

Rissanen, J. (1983). Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416–431. MR0696056

Ronchetti, E., Field, C., and Blanchard, W. (1997). Robust linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 92:1017–1023. MR1482132

Rudemo, M. (1982). Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9:65–78. MR0668683

Sauvé, M. (2009). Histogram selection in non gaussian regression. *ESAIM: Probability and Statistics*, 13:70–86. MR2502024

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. MR0650809

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464. MR0468014

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422):486–494. MR1224373

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.*, 91(434):655–665. MR1395733

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica*, 7(2):221–264. With comments and a rejoinder by the author. MR1466682

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, 71(1):43–49. MR0738324

Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297. MR0760688

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147. With discussion and a reply by the authors. MR0356377

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–35. MR0474601

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. Royal Statist. Soc. Series B*, 58(1):267–288. MR1379242

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). MR2101455

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. MR2305113

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. MR2305112

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. MR0672244

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. MR1641250

Vapnik, V. N. and Chervonenkis, A. Y. (1974). *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya*. Izdat. "Nauka", Moscow. Theory of Pattern Recognition (In Russian). MR0474638

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. MR0471299

Wegkamp, M. (2003). Model selection in nonparametric regression. *The Annals of Statistics*, 31(1):252–273. MR1962506

Yang, Y. (2001). Adaptive Regression by Mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588. MR1946426

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. MR2234196

Yang, Y. (2006). Comparing learning methods for classification. *Statist. Sinica*, 16(2):635–657. MR2267253

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, 35(6):2450–2473. MR2382654

Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Statist.*, 21(1):299–313. MR1212178