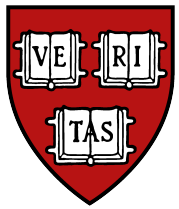
Cross-validation Confidence Intervals for Test Error

Alexandre Bayle

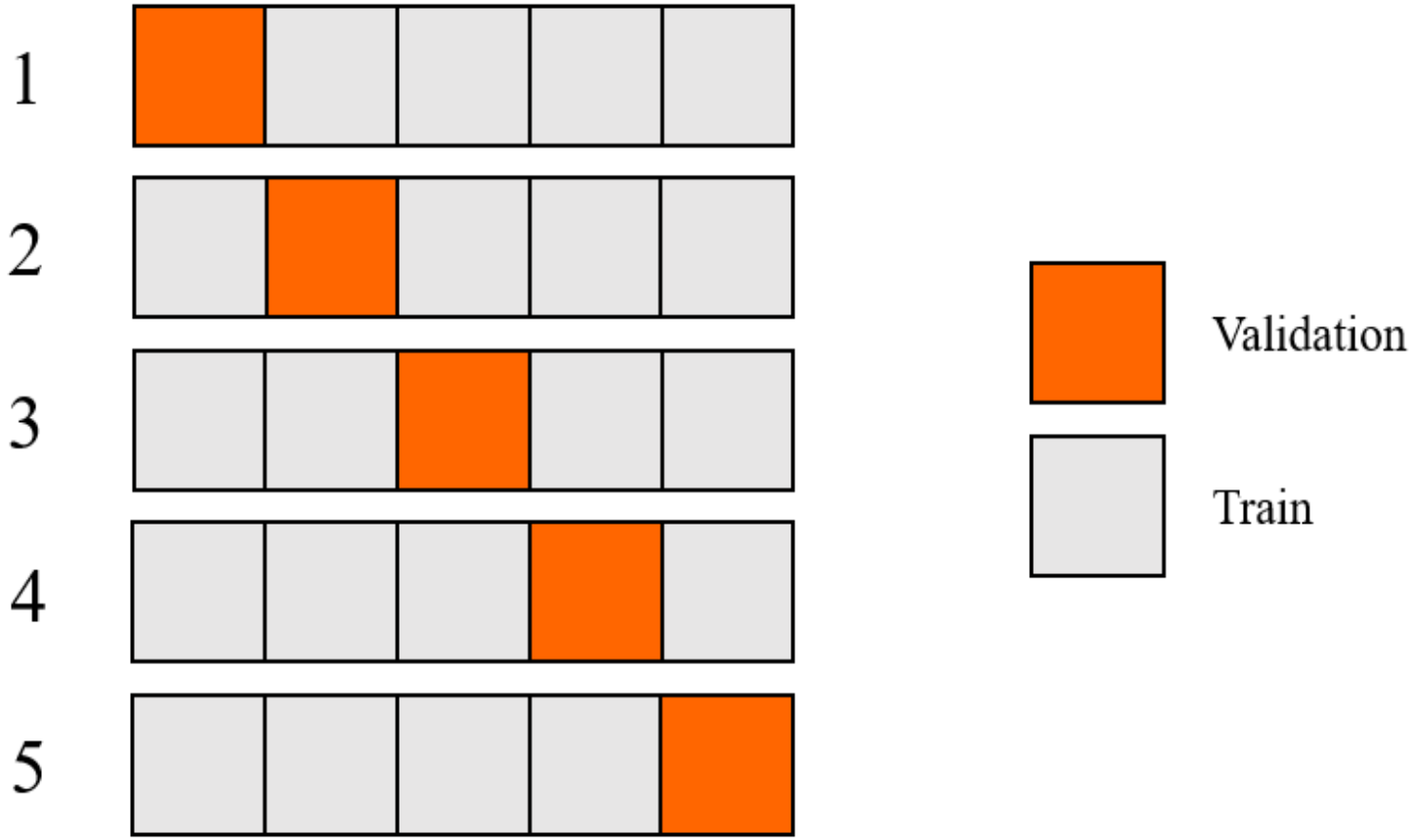
Harvard University



Joint work with **Pierre Bayle** (Princeton University),
Lucas Janson (Harvard University),
and **Lester Mackey** (Microsoft Research)

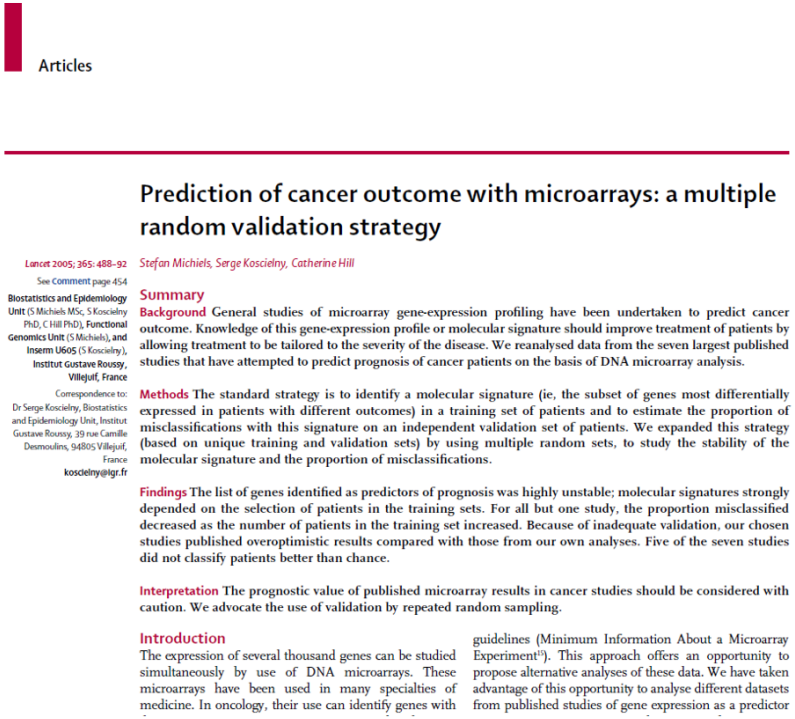
November 5, 2020

k-fold cross-validation



- **Unbiased**
- **Lower variance than a single train-test split**
- **Complex dependence structure**

High-stakes applications



Articles

Prediction of cancer outcome with microarrays: a multiple random validation strategy

Lancet 2005; 365: 488–92
See Comment page 454

Stefan Michiels, Serge Koscielny, Catherine Hill

Biostatistics and Epidemiology Unit (S Michiels MSc, S Koscielny PhD, C Hill PhD), **Functional Genomics Unit** (S Michiels), **and Inserm U605** (S Koscielny), **Institut Gustave Roussy, Villejuif, France**

Correspondence to:
Dr Serge Koscielny, Biostatistics and Epidemiology Unit, Institut Gustave Roussy, 39 rue Camille Desmoulins, 94805 Villejuif, France
koscielny@igr.fr

Summary

Background General studies of microarray gene-expression profiling have been undertaken to predict cancer outcome. Knowledge of this gene-expression profile or molecular signature should improve treatment of patients by allowing treatment to be tailored to the severity of the disease. We reanalysed data from the seven largest published studies that have attempted to predict prognosis of cancer patients on the basis of DNA microarray analysis.

Methods The standard strategy is to identify a molecular signature (ie, the subset of genes most differentially expressed in patients with different outcomes) in a training set of patients and to estimate the proportion of misclassifications with this signature on an independent validation set of patients. We expanded this strategy (based on unique training and validation sets) by using multiple random sets, to study the stability of the molecular signature and the proportion of misclassifications.

Findings The list of genes identified as predictors of prognosis was highly unstable; molecular signatures strongly depended on the selection of patients in the training sets. For all but one study, the proportion misclassified decreased as the number of patients in the training set increased. Because of inadequate validation, our chosen studies published overoptimistic results compared with those from our own analyses. Five of the seven studies did not classify patients better than chance.

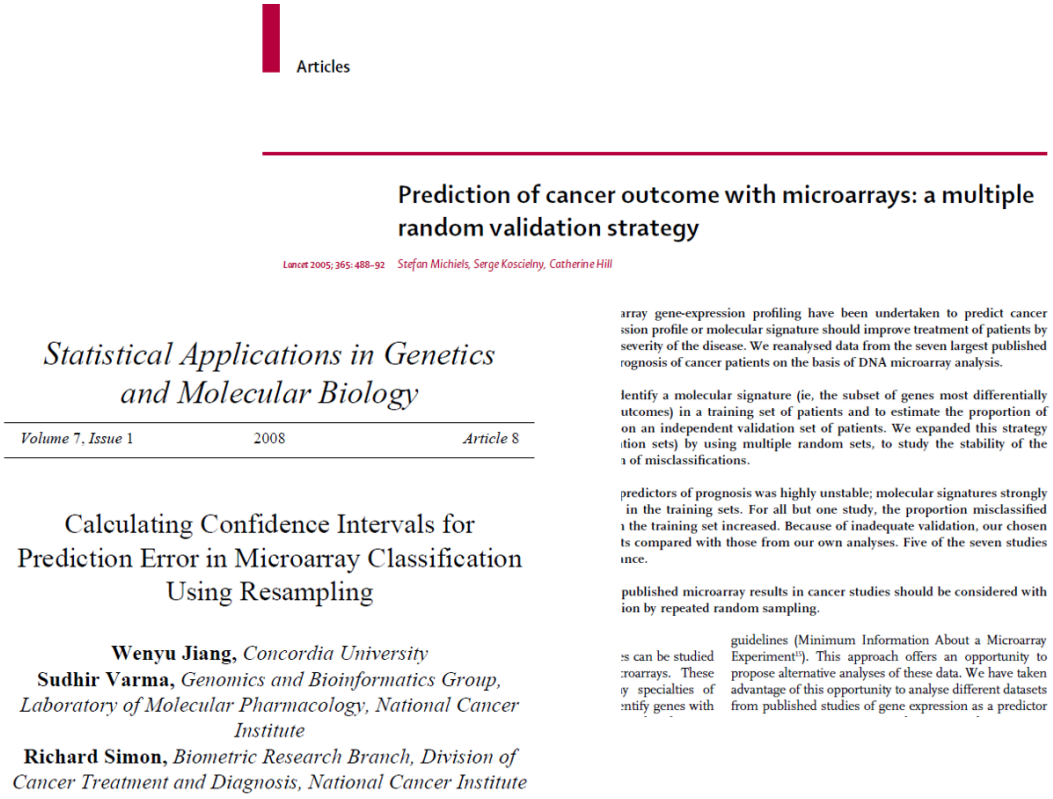
Interpretation The prognostic value of published microarray results in cancer studies should be considered with caution. We advocate the use of validation by repeated random sampling.

Introduction

The expression of several thousand genes can be studied simultaneously by use of DNA microarrays. These microarrays have been used in many specialties of medicine. In oncology, their use can identify genes with

guidelines (Minimum Information About a Microarray Experiment[1]). This approach offers an opportunity to propose alternative analyses of these data. We have taken advantage of this opportunity to analyse different datasets from published studies of gene expression as a predictor

High-stakes applications



Articles

Prediction of cancer outcome with microarrays: a multiple random validation strategy

Lancet 2005; 365: 488–92 Stefan Michiels, Serge Koscielny, Catherine Hill

array gene-expression profiling have been undertaken to predict cancer ssion profile or molecular signature should improve treatment of patients by severity of the disease. We reanalysed data from the seven largest published rognosis of cancer patients on the basis of DNA microarray analysis.

lentify a molecular signature (ie, the subset of genes most differentially utcomes) in a training set of patients and to estimate the proportion of on an independent validation set of patients. We expanded this strategy tion sets) by using multiple random sets, to study the stability of the a of misclassifications.

predictors of prognosis was highly unstable; molecular signatures strongly in the training sets. For all but one study, the proportion misclassified the training set increased. Because of inadequate validation, our chosen ts compared with those from our own analyses. Five of the seven studies nce.

published microarray results in cancer studies should be considered with ion by repeated random sampling.

s can be studied croarrays. These y specialties of ntify genes with guidelines (Minimum Information About a Microarray Experiment[15]). This approach offers an opportunity to propose alternative analyses of these data. We have taken advantage of this opportunity to analyse different datasets from published studies of gene expression as a predictor

Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 1 2008 *Article* 8

Calculating Confidence Intervals for Prediction Error in Microarray Classification Using Resampling

Wenyu Jiang, *Concordia University*
Sudhir Varma, *Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, National Cancer Institute*
Richard Simon, *Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute*

High-stakes applications

Articles

Prediction of cancer outcome with microarrays: a multiple random validation strategy

Lancet 2005; 365: 488–92 Stefan Michiels, Serge Koscielny, Catherine Hill

Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 1 2008 *Article* 8

Calculating Confidence Intervals for Prediction Error in Microarray Classification Using Resampling

Wenyu Jiang, *Concordia University*
Sudhir Varma, *Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, National Cancer Institute*
Richard Simon, *Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute*

array gene-expression profiling
ssion profile or molecular signatu
severity of the disease. We reanal
rognosis of cancer patients on the

lentify a molecular signature (ie
utcomes) in a training set of pa
on an independent validation se
tion sets) by using multiple ra
n of misclassifications.

predictors of prognosis was highl
in the training sets. For all but
n the training set increased. Beca
ts compared with those from ou
nce.

published microarray results in
ion by repeated random samplin

s can be studied
roarrays. These
y specialties of
ntify genes with

guidelines (M
Experiment[15]).
propose altern
advantage of t
from publishe

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH Public Access
Author Manuscript
Lancet Respir Med. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:
Lancet Respir Med. 2015 January ; 3(1): 42–52. doi:10.1016/S2213-2600(14)70239-5.

Mortality prediction in the ICU: can we do better? Results from the Super ICU Learner Algorithm (SICULA) project, a population-based study

Romain Pirracchio, MD[*,$,£], **Maya L. Petersen, MD**[*], **Marco Carone, PhD**[#], **Matthieu Resche Rigon, MD**[$], **Prof. Sylvie Chevret, MD**[$], and **Prof. Mark J. van der LAAN, PhD**[*]

[*]Division of Biostatistics, School of Public Health, University of Califonia, Berkeley, Berkeley, USA

[$]Service de Biostatistique et Information Médicale, Unité INSERM 1153, Equipe ECSTRA, Hôpital Saint Louis, Paris, France

[£]Service d'Anesthésie-Réanimation, Hôpital Européen Georges Pompidou, Université Paris 5 Descartes, Sorbonne Paris Cité, Paris, France

[#]Department of Biostatistics, School of Public Health, University of Washington, Seattle, USA

Abstract

Background—Improved mortality prediction for patients in intensive care units (ICU) remains an important challenge. Many severity scores have been proposed but validation studies have concluded that they are not adequately calibrated. Many flexible algorithms are available, yet none of these individually outperform all others regardless of context. In contrast, the *Super Learner* (SL), an ensemble machine learning technique that leverages on multiple learning algorithms to obtain better prediction performance, has been shown to perform at least as well as the optimal member of its library. It might provide an ideal opportunity to construct a novel severity score with an improved performance profile. The aim of the present study was to provide a new

High-stakes applications

Is algorithm A actually better than algorithm B?

Notations and setting

- $(Z_i)_{i\geq 1} = (X_i, Y_i)_{i\geq 1}$, X_i vector of covariates and Y_i target variable

Notations and setting

- $(Z_i)_{i\geq 1} = (X_i, Y_i)_{i\geq 1}$, X_i vector of covariates and Y_i target variable
- For any vector B of indices in $\{1, \ldots, n\}$, let Z_B denote the subvector of $Z_{1:n}$ corresponding to ordered indices in B.

Notations and setting

- $(Z_i)_{i\geq 1} = (X_i, Y_i)_{i\geq 1}$, X_i vector of covariates and Y_i target variable
- For any vector B of indices in $\{1, \ldots, n\}$, let Z_B denote the subvector of $Z_{1:n}$ corresponding to ordered indices in B.
- Consider a set of k train-validation splits $\{(B_j, B'_j)\}_{j=1}^k$ with validation indices $\{B'_j\}_{j=1}^k$ partitioning $\{1, \ldots, n\}$ into k folds.

Notations and setting

- $(Z_i)_{i\geq 1} = (X_i, Y_i)_{i\geq 1}$, X_i vector of covariates and Y_i target variable
- For any vector B of indices in $\{1, \ldots, n\}$, let Z_B denote the subvector of $Z_{1:n}$ corresponding to ordered indices in B.
- Consider a set of k train-validation splits $\{(B_j, B'_j)\}_{j=1}^k$ with validation indices $\{B'_j\}_{j=1}^k$ partitioning $\{1, \ldots, n\}$ into k folds.
- $h_n(Z_i, Z_B)$: scalar loss function, evaluating the loss on the test point $Z_i = (X_i, Y_i)$ of the prediction rule learned on the training data Z_B. Examples of h_n:

$$h_n(Z_i, Z_B) = \begin{cases} (Y_i - \hat{f}(X_i; Z_B))^2 \\ \mathbb{1}[Y_i \neq \hat{f}(X_i; Z_B)] \end{cases}$$

k-fold cross-validation error and k-fold test error

Definition (k-fold cross-validation error)

$$\hat{R}_n \triangleq \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in B'_j} h_n(Z_i, Z_{B_j})$$

- k either fixed or dependent on n
- The terms are not independent. What is the asymptotic behavior?

k-fold cross-validation error and k-fold test error

Definition (k-fold cross-validation error)

$$\hat{R}_n \triangleq \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in B'_j} h_n(Z_i, Z_{B_j})$$

- k either fixed or dependent on n
- The terms are not independent. What is the asymptotic behavior?

Definition (k-fold test error)

$$R_n \triangleq \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in B'_j} \mathbb{E}[h_n(Z_i, Z_{B_j}) \mid Z_{B_j}]$$

It is a standard inferential target and represents the average test error of the k prediction rules $\hat{f}(\cdot; Z_{B_j})$ for $j = 1, \ldots, k$.

k-fold cross-validation error and k-fold test error

Definition (k-fold cross-validation error)

$$\hat{R}_n \triangleq \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in B'_j} h_n(Z_i, Z_{B_j})$$

- k either fixed or dependent on n
- The terms are not independent. What is the asymptotic behavior?

Definition (k-fold test error)

$$R_n \triangleq \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in B'_j} \mathbb{E}[h_n(Z_i, Z_{B_j}) \mid Z_{B_j}]$$

It is a standard inferential target and represents the average test error of the k prediction rules $\hat{f}(\cdot; Z_{B_j})$ for $j = 1, \ldots, k$.

Goal

Central Limit Theorem on $\hat{R}_n$ under mild assumptions

Algorithmic stability

Stability

How much does the performance of a learned prediction rule change when one point in the training set is changed? Different kinds of stability, for example:

- Uniform stability
- Mean-square stability γ_{ms}
- Loss stability γ_{loss}

Note: $\gamma_{loss} \leq \gamma_{ms}$

[Bousquet and Elisseeff, 2002, Kale et al., 2011, Kumar et al., 2013, Celisse and Guedj, 2016, . . .]

Asymptotic normality of k-fold CV

Theorem

Suppose $(Z_i)_{i\geq 1}$ are i.i.d. copies of a random element Z_0.
Let $\bar{h}_n(Z_0) = \mathbb{E}[h_n(Z_0, Z_{1:n(1-1/k)}) \mid Z_0]$ and $\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0))$.
If the following conditions hold:

- $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$,
- the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is uniformly integrable,

then

$$\frac{\sqrt{n}}{\sigma_n}\left(\hat{R}_n - R_n\right) \xrightarrow{d} \mathcal{N}(0,1).$$

Improvement upon previous literature

Dudoit and van der Laan, Theorem 3 (2005)

- Assumes a bounded loss function
- Excludes leave-one-out CV
- Requires the prediction rule to be loss-consistent for a risk-minimizing prediction rule

Improvement upon previous literature

Dudoit and van der Laan, Theorem 3 (2005)

- Assumes a bounded loss function
- Excludes leave-one-out CV
- Requires the prediction rule to be loss-consistent for a risk-minimizing prediction rule

Austern and Zhou, Theorem 1 (2020)

- Assumes variance parameter converging to a non-zero limit
- Requires $o(1/n)$ mean-square stability
- Requires $o(1/n^2)$ second-order mean-square stability
- Assumes learning algorithms symmetric in the training points

k-fold test error confidence intervals

Goal

Construct an asymptotically-exact $(1-\alpha)$-confidence interval for the unknown k-fold test error R_n

k-fold test error confidence intervals

Goal

Construct an asymptotically-exact $(1-\alpha)$-confidence interval for the unknown k-fold test error R_n

Confidence interval

Consider $\hat{\sigma}_n^2$ a variance estimator satisfying relative error consistency, $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$. With the CLT,

$$C_\alpha \triangleq \hat{R}_n \pm q_{1-\alpha/2}\,\hat{\sigma}_n/\sqrt{n}$$

satisfies

$$\lim_{n\to\infty} \mathbb{P}(R_n \in C_\alpha) = 1-\alpha$$

where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution

Testing for algorithm improvement

Goal

Given a dataset $Z_{1:n}$, a k-fold partition $\{B'_j\}_{j=1}^k$, and two algorithms $\mathcal{A}_1$, $\mathcal{A}_2$ for fitting prediction rules, test whether $\mathcal{A}_1$ improves upon $\mathcal{A}_2$ on the fold partition

Testing for algorithm improvement

Goal

Given a dataset $Z_{1:n}$, a k-fold partition $\{B'_j\}_{j=1}^k$, and two algorithms $\mathcal{A}_1$, $\mathcal{A}_2$ for fitting prediction rules, test whether $\mathcal{A}_1$ improves upon $\mathcal{A}_2$ on the fold partition

Test

Define

$$h_n(Z_0, Z_B) = \ell(Y_0, \hat{f}_1(X_0; Z_B)) - \ell(Y_0, \hat{f}_2(X_0; Z_B)).$$

Consider $\hat{\sigma}_n^2$ a variance estimator satisfying relative error consistency, $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$.
Test the null $H_0 : R_n \geq 0$ against the alternative hypothesis $H_1 : R_n < 0$. Asymptotically-exact level-α test is given by

$$\text{REJECT } H_0 \Leftrightarrow \hat{R}_n < q_\alpha \hat{\sigma}_n / \sqrt{n}$$

where q_α is the α-quantile of a standard normal distribution

Consistent variance estimation

Want to find an estimator $\hat{\sigma}_n^2$ such that $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$ under weak conditions.

Definition (Within-fold variance estimator)

$\hat{\sigma}_{n,in}^2$ is the average of the k within-fold empirical variances

Definition (All-pairs variance estimator)

$$\hat{\sigma}_{n,out}^2 \triangleq \frac{1}{n}\sum_{j=1}^{k}\sum_{i\in B'_j}(h_n(Z_i, Z_{B_j}) - \hat{R}_n)^2$$

Advantage: can also be used for leave-one-out cross-validation

Low computational cost

$\hat{\sigma}_{n,in}^2$ and $\hat{\sigma}_{n,out}^2$ can be computed in $O(n)$ time

Consistent variance estimation

Theorem

Suppose $(Z_i)_{i\geq 1}$ are i.i.d. copies of a random element Z_0.
Let $\bar{h}_n(Z_0) = \mathbb{E}[h_n(Z_0, Z_{1:n(1-1/k)}) \mid Z_0]$ and $\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0))$.
If the following conditions hold:

1. $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$,
2. the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is uniformly integrable,

then

$$\hat{\sigma}^2_{n,in}/\sigma_n^2 \xrightarrow{L^1} 1.$$

Consistent variance estimation

Theorem

Suppose $(Z_i)_{i\geq 1}$ are i.i.d. copies of a random element Z_0.
Let $\bar{h}_n(Z_0) = \mathbb{E}[h_n(Z_0, Z_{1:n(1-1/k)}) \mid Z_0]$ and $\sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0))$.
If the following conditions hold:

1. $\gamma_{loss}(h_n) = o(\sigma_n^2/n)$,
2. the sequence of $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is uniformly integrable,

then

$$\hat{\sigma}^2_{n,in}/\sigma_n^2 \xrightarrow{L^1} 1.$$

If additionally:

3. $\gamma_{ms}(h_n) = o(k\sigma_n^2/n)$,

then

$$\hat{\sigma}^2_{n,out}/\sigma_n^2 \xrightarrow{L^1} 1.$$

Experiments – k-fold test error confidence intervals

$$C_\alpha = \hat{R}_n \pm q_{1-\alpha/2}\,\hat{\sigma}_n/\sqrt{n} \quad \text{with} \quad \alpha = 0.05$$

Our CV CLT procedure: valid coverage, smallest width

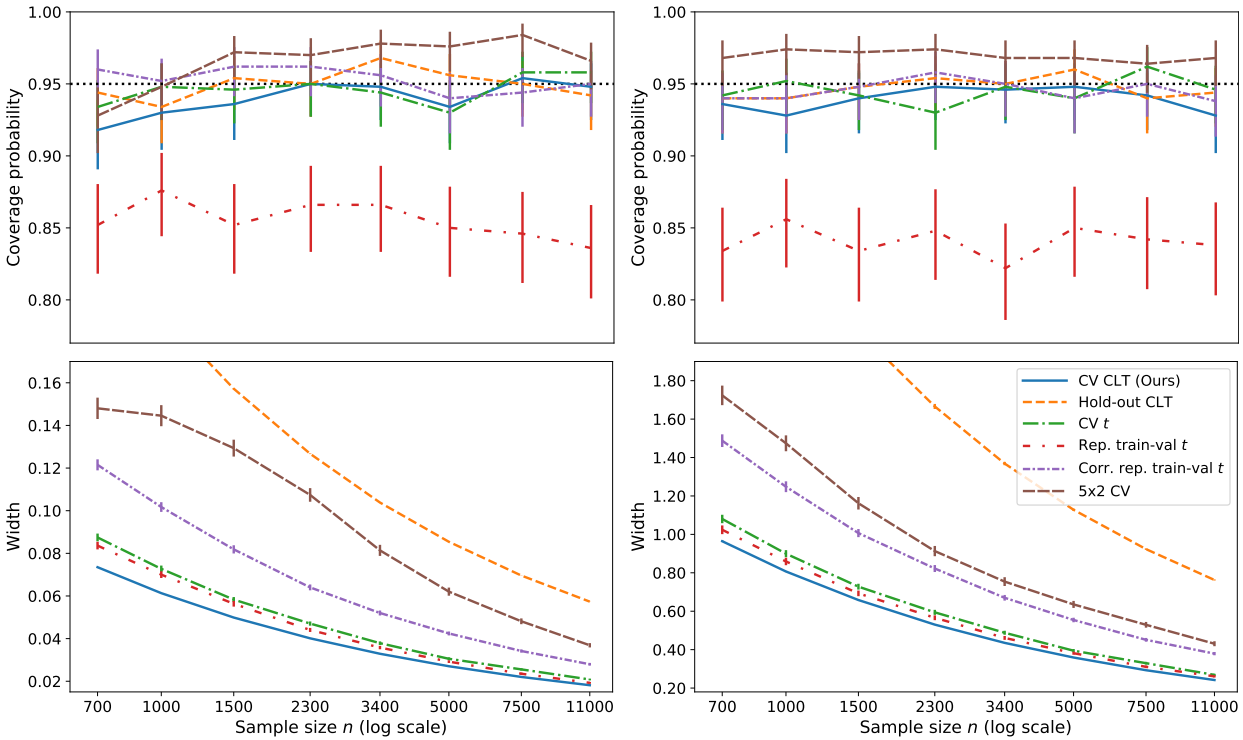


Figure: Test error coverage (top) and width (bottom) of 95% confidence intervals. **Left:** ℓ^2-regularized logistic regression classifier. **Right:** Random forest regression.

Experiments – Testing for algorithm improvement

$$\text{REJECT } H_0 \Leftrightarrow \hat{R}_n < q_\alpha \hat{\sigma}_n / \sqrt{n} \quad \text{with} \quad \alpha = 0.05$$

Our CV CLT procedure: valid size, most powerful

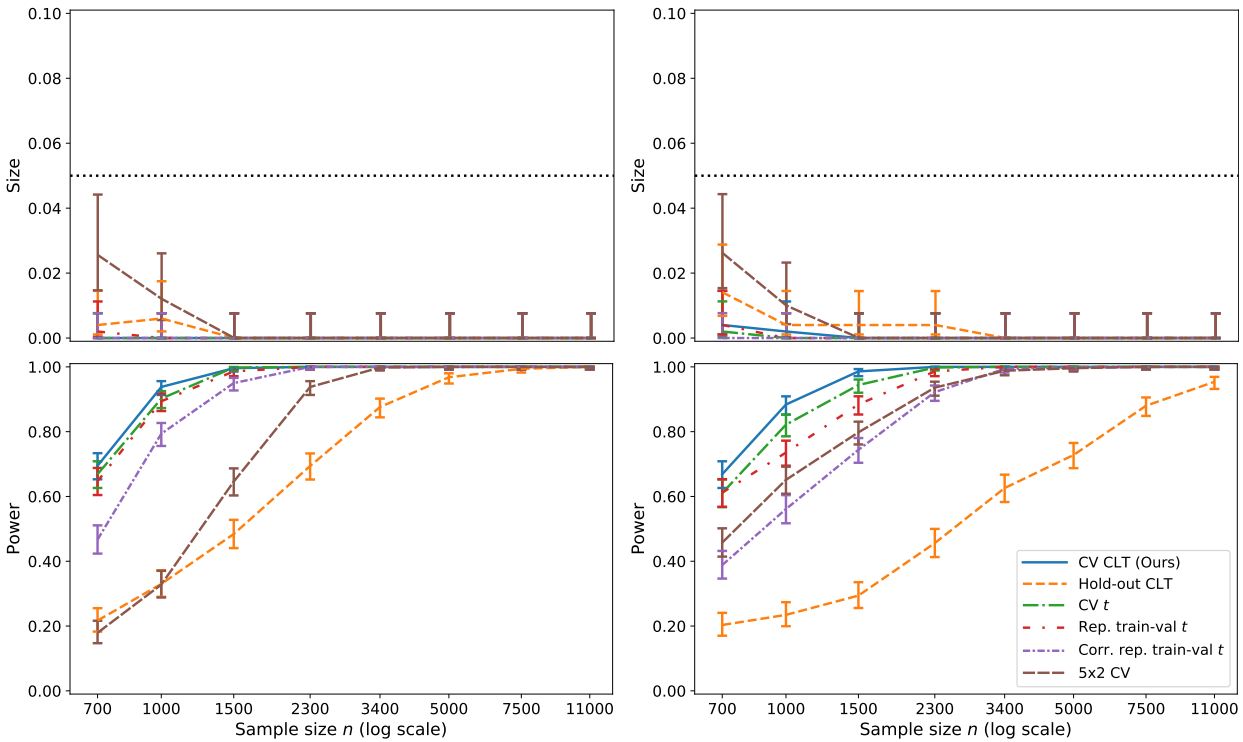
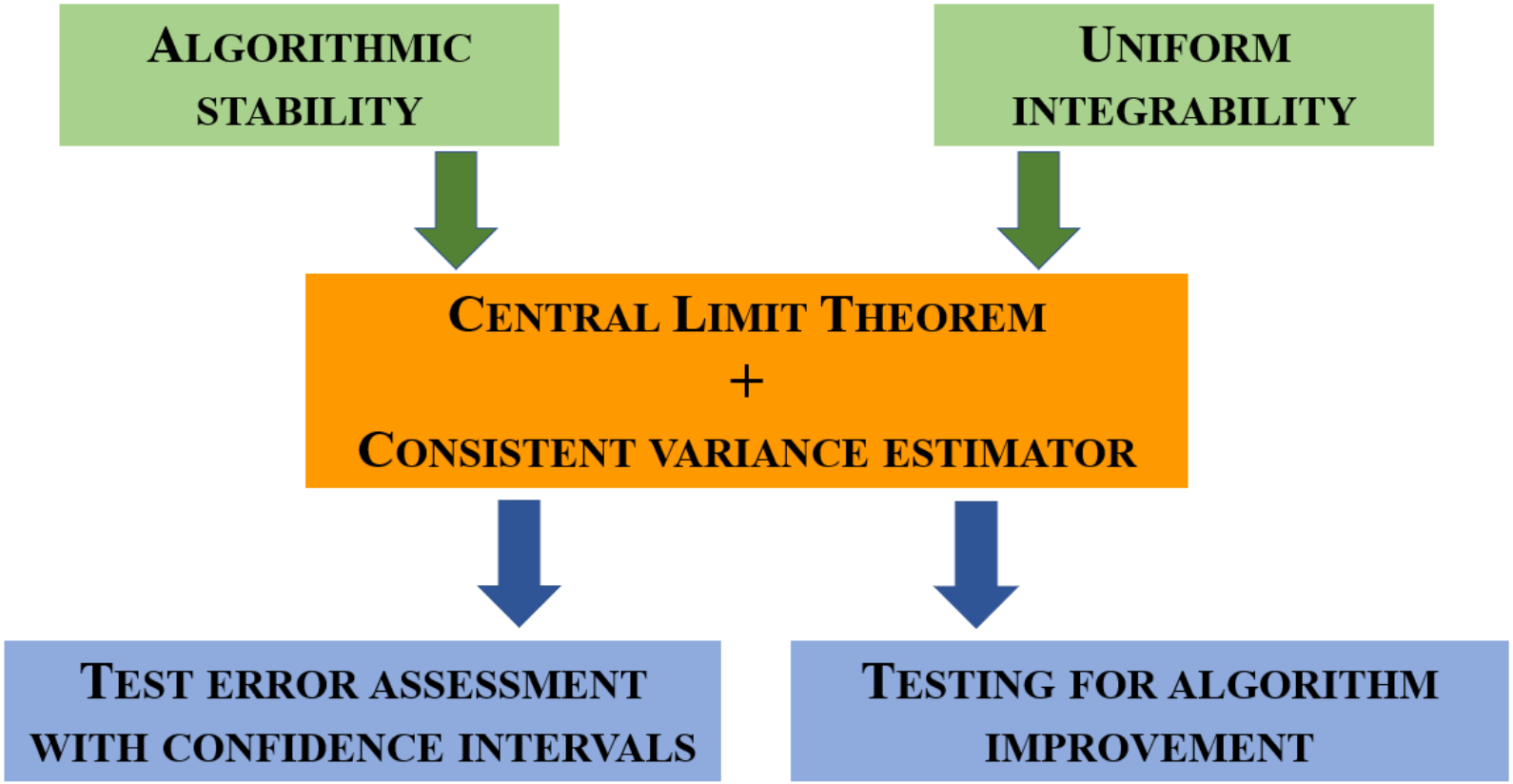


Figure: Size when testing $H_1 : \text{Err}(\mathcal{A}_1) < \text{Err}(\mathcal{A}_2)$ (top) and power when testing $H_1 : \text{Err}(\mathcal{A}_2) < \text{Err}(\mathcal{A}_1)$ (bottom) of level-0.05 tests for improved test error. **Left**: $\mathcal{A}_1 = \ell^2$-regularized logistic regression, $\mathcal{A}_2 =$ neural network classification. **Right**: $\mathcal{A}_1 =$ random forest, $\mathcal{A}_2 =$ ridge regression.

Conclusion


ALGORITHMIC STABILITY
UNIFORM INTEGRABILITY
CENTRAL LIMIT THEOREM
+
CONSISTENT VARIANCE ESTIMATOR
TEST ERROR ASSESSMENT WITH CONFIDENCE INTERVALS
TESTING FOR ALGORITHM IMPROVEMENT

Thank you!

Cross-validation Confidence Intervals for Test Error

Paper: https://arxiv.org/abs/2007.12671

Code: https://github.com/alexandre-bayle/cvci

Also in the paper:

- additional theoretical results
- experiments in the leave-one-out setting
- experiments illustrating the importance of stability