LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Mathias Fuchs, Norbert Krautenbacher

A variance decomposition and a Central Limit Theorem for empirical losses associated with resampling designs

Technical Report Number 173, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

A VARIANCE DECOMPOSITION AND A CENTRAL LIMIT THEOREM FOR EMPIRICAL LOSSES ASSOCIATED WITH RESAMPLING DESIGNS

MATHIAS FUCHS
NORBERT KRAUTENBACHER

ABSTRACT. The mean prediction error of a classification or regression procedure can be estimated using resampling designs such as the cross-validation design. We decompose the variance of such an estimator associated with an arbitrary resampling procedure into a small linear combination of covariances between elementary estimators, each of which is a regular parameter as described in the theory of U-statistics. The enumerative combinatorics of the occurrence frequencies of these covariances govern the linear combination's coefficients and, therefore, the variance's large scale behavior. We study the variance of incomplete U-statistics associated with kernels which are partly but not entirely symmetric. This leads to asymptotic statements for the prediction error's estimator, under general non-empirical conditions on the resampling design. In particular, we show that the resampling based estimator of the average prediction error is asymptotically normally distributed under a general and easily verifiable condition. Likewise, we give a sufficient criterion for consistency. We thus develop a new approach to understanding small-variance designs as they have recently appeared in the literature. We exhibit the U-statistics which estimate these variances.

We present a case from linear regression where the covariances between the elementary estimators can be computed analytically. We illustrate our theory by computing estimators of the studied quantities in an artificial data example.

1. INTRODUCTION

This paper is concerned with the variance of resampling designs in machine learning and statistics. A resampling design — a collection of splits of the data into learning and test sets — yields an estimator of the expectation of a loss function of a model fitting procedure; see Section 2 for details of the set-up. An example of a resampling design is the leave-p-out estimator of the average prediction error; a recent preprint [5] exploits the fact that this estimator is a U-statistic to derive its properties. Consequently, it is asymptotically normally distributed under a very weak condition, namely that of existing and non-vanishing asymptotic variance.

In this work, we generalize from the leave-p-out estimator to general resampling designs such as cross-validation. We set up a very general condition on the resampling design that leads to consistency and a narrower one that leads to asymptotic normality.

In a similar framework, consistency of cross-validation was shown in Yang [13]; the principal difference between our work and Yang's is that in Theorems 4 and 5 we treat the case of a fixed learning set size; hence, we do not subsume cross-validation in these theorems.

A resampling design is an non-empirical datum in the sense that it can and should be

2010 *Mathematics Subject Classification.* 62G05, 62G09, 62G10, 62G20.

Key words and phrases. U-statistic, cross-validation, limit theorem, design, model selection.

supported by the German Science Foundation (DFG-Einzelförderung BO3139/2-2 to Anne-Laure Boulesteix).

specified by the experimenter before seeing the data. Moreover, a design is, by nature, algebraic: its definition involves no probability or analysis. On the other hand, a Central Limit Theorem is a probabilistic-analytical statement and thus of quite a different nature. Suppose we are given a sequence of resampling designs in the sense that for every sufficiently large sample size n a collection of learning sets is specified. Then, the Central Limit Theorem may either hold or not. By specifying a sufficient criterion in this work, we pose the question of whether there is any necessary criterion on the design. This question seems to be very challenging. Likewise, it seems to be difficult to determine sufficient or necessary conditions on the resampling design for the Strong Law of Large Numbers, the Berry-Esseen theorem, and the Law of the Iterated Logarithm to be valid. Therefore, it seems that the present work raises certain interesting and challenging problems for further research, at the boundary between the combinatorial world of designs and the probabilistic-analytical world of limit theorems.

Partial answers are given by the theory of incomplete U-statistics; however, the theory of incomplete U-statistics has only been developed thoroughly in the case of symmetric kernels. Here, in contrast, a resampling design is an incomplete U-statistic that is naturally associated with a non-symmetric kernel but usually not a symmetric one (note that only complete U-statistics are always associated with symmetric kernels).

Usually, in statistical design theory, the experimental units are allocated to blocks, and each experimental unit leads to a response [2, Section 3.1, Equation (2)]. Here, the picture is quite different: our independent observations correspond to what are called the treatments; thus, a response is measured for each “treatment”. It seems that this viewpoint appears less frequently in statistics.

Here, we point out the usefulness of statistical design theory to resampling. Although design theory has been used in resampling and variance estimation theory, previous papers seem to have focused on giving surveys [11], whereas we examine model fitting algorithms in general.

Design theory and U-statistics also seem to have been examined in the case where the blocks are the evaluation indices of symmetric kernels. Here, we look at a very different scenario: The blocks are the indices of the learning sets, and the kernel is non-symmetric since it involves a learning set together with a testing observation.

Likewise, the literature describing resampling procedures for model fitting in the language of U-statistics seems to be surprisingly sparse.

Let us now outline the main results in more detail.

The problem that cross-validation suffers high variance is well studied; further, approaches aiming at alleviating this are classical and treated in vast amounts of literature. Recently, in Zhang and Qian [14], cross-validation designs akin to Latin hyper-cube designs in experimental design theory were proposed, and it was shown that such designs, although of a computational cost similar to that of cross-validation, have clearly smaller variance and are therefore generally preferable. Zhang and Qian [14, after Formula 12] gives a variance decomposition of the average prediction error estimator associated with several particular designs; we will give the corresponding formula for any design. The same reference also contains an extensive overview of recent literature.

Moreover, Fuchs et al. [5] outlined that the leave-p-out prediction error estimator can be seen as a U-statistic and exploited this fact to deduce the existence of an approximately exact hypothesis test of the equality of the two prediction errors. Since Fuchs et al. [5] is a preprint, we give a synopsis of that paper in Section 2.4. Thus, we aim to exploit the fact that any resampling procedure is an incomplete U-statistic and to view the results of Zhang and Qian [14] in the light of the variance calculation framework of U-statistics.

There is a general theory of incomplete U-statistics designs such that the variance of such a incomplete U-statistics is as small as possible and, therefore, as close as possible to that of the leave-p-out classifier [9, Chapter 4]; let us recall the fact that any complete U-statistic associated with a possibly non-symmetric kernel is simultaneously a U-statistic associated with a symmetric kernel, namely the striation of the original kernel. Thus the theory of complete U-statistics is entirely covered by that for symmetric kernels. However, the picture is completely different for incomplete U-statistics. The reason is that if one defines an incomplete U-statistic just as an average taken over symmetric kernels of a collection of subsets, then one misses a good deal of interesting statistics. Here, we will investigate a more general definition that calls any average of non-symmetric kernels an incomplete U-statistic.

In contrast to a definition just containing symmetric kernels, we will have to perform optimization for non-symmetric kernels. Then, the kernel defining the U-statistic which is the leave-p-out error estimator, is genuinely non-symmetric. The associated symmetrization is the leave-one-out error estimator on a sample whose size is just one plus the original learning sample size. We are now faced with the difficulty that this kernel is computationally very unfortunate. Therefore, we set out to generalize the theory of incomplete U-statistics to that of non-symmetric kernels. However, we will do so just for the case of a mildly non-symmetric kernel such as ours – in fact, only a few summands are necessary in order to obtain a symmetric one.

Subsuming this point, it seems that the existing theories are restricted to the case of symmetric kernels. In contrast, a proper resampling procedure would not rely on a symmetric kernel, because there is no reason why small-variance procedures could be achieved with a symmetric kernel. Moreover, it seems very intuitive that the symmetrized formula of the kernel leads to a very high ratio of variance to computational cost.

In generalizing the theory of incomplete U-statistics to that of non-symmetric kernels, we give a conceptual approach to finding designs similar to the ad-hoc designs of Zhang and Qian [14] which were defined without any mention of U-statistics.

The main results of our paper concern a decomposition of the variance of any cross-validation-like procedure into a linear combination of four series of “core covariances”, generalizing the covariances appearing in Bengio and Grandvalet [3, Corollary 2]. Each of these, denoted by $\tau_d^{(i)}$ for $i = 1, \ldots, 4$, is a regular parameter and can therefore be estimated optimally by another U-statistic.

The variance estimation of U-statistics has already been considered in the literature [10, 12].

The coefficients of the linear combination are polynomials of degree at most two that only depend on the sample size and the learning set size. Thus, they are known in advance of seeing the data and easily calculable. The decomposition is a significant generalization of the classical decomposition of Hoeffding [7, Formula 5.18] for the variance of a U-statistic to the case of an incomplete U-statistic associated with a symmetric kernel. The difference of our variance formula to Lee’s is that ours extends over four series of covariances instead of just one.

It turns out that the variance expression thus attained is extremely difficult (or perhaps impossible) to minimize over all designs of a given size — uniformly over all underlying probability distributions P. Therefore, we will approximate an asymptotic case of large sample size.

Our main results are: proving the existence of unbiased variance estimators for the core covariances (Corollary 1), the variance structure of cross-validation (Theorem 16) and its estimation (Theorem 3), the analytical computation of the core covariances in a toy regression model in Section 4, the Central Limit Theorem 5 and the associated asymptotic test in (29) and the numerical computation of the estimators in a related regression model in Section 6.

The paper is structured as follows. In Section 2, we specify the set-up, Section 3 explains the variance decompositions, Section 4 presents an analytical computation of the core covariances, in Section 5, we define the variance estimators and show the Central Limit Theorem, and Section 6 illustrates our theory by means of a data example in which we compute the estimators numerically.

2. The set-up

2.1. **The loss estimator.** The general framework of the loss estimator is slightly more general than that underlying the largest part of statistical literature.

In the general framework, there is a univariate response variable Y ranging over a set $\mathscr{Y}$, and a multivariate predictor variable X ranging over $\mathscr{X}$ (both $\mathscr{X}$ and $\mathscr{Y}$ are assumed to be equipped with fixed σ-algebras). The joint distribution of (X,Y) is described by a probability measure P on $\mathscr{X}\times\mathscr{Y}$ equipped with the product σ-algebra. The quality of the prediction of Y is measured by a loss function $(y,y')\mapsto l(y,y')$. Typically, binary classification uses the misclassification loss $\mathbb{1}_{y\neq y'}$, but we can also use any other measurable loss. Other loss functions include, for instance, the usual regression mean-square loss $(y-y')^2$ or a survival analysis loss after extending the loss function's domain of definition to censored observations.

We fix a learning sample size g and then consider a statistical model fitting procedure in the form of a function

$$\begin{aligned} s:(\mathscr{X}\times\mathscr{Y})^{\times g}\times\mathscr{X}&\to\mathscr{Y}\\ (x_1,y_1,\dots,x_g,y_g,x_{g+1})&\mapsto s(x_1,y_1,\dots,x_g,y_g;x_{g+1}) \end{aligned}\tag{1}$$

which maps the learning sample $(x_1,y_1,\dots,x_g,y_g)$ to the prediction rule applied to the test observation x_{g+1}. Equivalently, s can be seen as mapping the learning sample to a classification rule which is a map from predictors $\mathscr{X}$ to responses in $\mathscr{Y}$. (Sometimes, $s(x_1,y_1,\dots,x_g,y_g;x_{g+1})$ is denoted by $\widehat{f}(x_{g+1}|x_1,y_1,\dots,x_g,y_g)$ to describe a learned estimator $\widehat{f}$ for a true model $f:\mathscr{X}\to\mathscr{Y}$.) Throughout the paper, we will assume that s treats all learning arguments equally, so that it is invariant under permutation of the first g arguments, and we assume that s is measurable with respect to the product σ-algebra on $(\mathscr{X}\times\mathscr{Y})^{\times g}\times\mathscr{X}$.

The joint expectation of the loss function with respect to the $g+1$-fold product measure is

$$\mathbb{E}(l(s))=\int\cdots\int l(s(x_1,y_1,\dots,x_g,y_g;x_{g+1}),y_{g+1})dP(x_1,y_1),\dots,dP(x_{g+1},y_{g+1})\tag{2}$$

and is called the unconditional loss of the model fitting procedure, where the left-hand side uses a slightly sloppy but unambiguous notation. It is of practical interest to estimate it, together with the difference $\mathbb{E}(l_1(s_1))-\mathbb{E}(l_2(s_2))=\mathbb{E}(l_1(s_1)-l_2(s_2))$, for two model fitting procedures s_1 and s_2 and two loss functions l_1 and l_2.

Remark 1. $\mathbb{E}(l(s))$ generalizes the usual mean squared error in sense that the loss function is arbitrary instead of being the quadratic loss, the true model is arbitrary instead of being in the particular form $Y=f(X)+\varepsilon$, the predictors X are random, and the expectation is taken with respect to the learning data as well.

Even if the true model is of the form $Y = f(X) + \varepsilon$ and the loss is quadratic, one cannot immediately obtain a bias-variance decomposition as in Hastie et al. [6, Formula 2.47] because the joint testing and learning expectation instead of just the testing expectation leads to covariance between $f(X_{g+1})$ and $\widehat{f}(X_{g+1})$. The derivation of the bias-variance decomposition usually relies on ignoring this covariance by viewing the X_i as non-random.

2.2. **Estimators for the loss.** Let us define

$$\begin{aligned}\Gamma(i_1,\dots,i_g;i_{g+1}) := {} & l_1(s_1(x_{i_1},y_{i_1},\dots,x_{i_g},y_{i_g};x_{i_{g+1}}),y_{i_{g+1}}) \\ & - l_2(s_2(x_{i_1},y_{i_1},\dots,x_{i_g},y_{i_g};x_{i_{g+1}}),y_{i_{g+1}}),\end{aligned} \tag{3}$$

a function on a set of $g+1$ different indices $i_k \in 1,\dots,n$, for two model fitting procedures s_1, s_2 and two appropriate loss functions l_1, l_2. We allow for each model fitting procedure to have its own loss function because then the case $l_2 := 0$ yields the loss of a single procedure, which is of obvious practical interest.

We have: $\mathbb{E}(\Gamma) = \mathbb{E}(l_1(s_1) - l_2(s_2))$ and we define $\Theta = \mathbb{E}\Gamma$ as a slight generalization of (2). The expectations are taken with respect to the $(g+1)$-fold product space of $\mathscr{X} \times \mathscr{Y}$ and are assumed to exist.

A resampling procedure is a collection of disjoint learning and test sets. For every pair of learning set and test observation one obtains an "elementary" estimator of the error rate. Averaging these across all learning and test sets of the resampling procedure defines an unbiased estimator for Θ. Quite often, another convention is used where such an estimator is seen as an approximation for the prediction error on another learning set size such as the total sample size; then, unbiasedness is of course lost. It is now of interest to gain insight into the variance of such an estimator.

All expectations and variances are taken with respect to the $g+1$-fold product measure of P. The definition of Γ was such that the number $g+1$ of arguments is minimal under the restriction that $\Theta = \mathbb{E}\Gamma$ for all underlying probability distributions such that this expectation exists. This minimality would be lost if the definition of Γ involved a larger test set size than one.

Let $\mathscr{T}$ be a collection of pairs (S,a) where $S \subset \{1,\dots,n\}$ is an (unordered) set of disjoint learning indices, and $a \in \{1,\dots,n\} \setminus S$ is a test index. Then, each $\Gamma(S;a)$ is an "elementary" estimator of Θ, and we define

$$\widehat{\Theta}(\mathscr{T}) := \frac{1}{|\mathscr{T}|} \sum_{(S,a)\in\mathscr{T}} \Gamma(S;a).$$

In simple cases, it is possible to compute Θ analytically. For instance, we will do so in Section 4.

2.3. **Complete and incomplete U-statistics.** This section summarizes some definitions and ideas from [7]. Let n denote the sample size. A U-statistic is a statistic of the form $U = \binom{n}{k}^{-1} \sum h(z_{i_1},\dots,z_{i_k})$ for a symmetric function h of k vector arguments, where the summation extends over all possible subsets $(i_1,\dots,i_k)$. Since the number of such subsets is $\binom{n}{k}$, the expectation of U is equal to that of h with respect to the k-fold product measure of P, so U is an unbiased estimator of $\mathbb{E}(h)$. A regular parameter is a functional of the form $P \mapsto \int h d^k(P)$. The minimal k such that there exists a symmetric function h such that $\mathbb{E}(h) = \Theta$ holds for all probability distributions P is called the degree of the U-statistic. Any such minimal function is called a kernel of U. If a non-symmetric function with that property exists, then, by symmetrization, a symmetric function exists.

An important property of U-statistics is that they are the unique minimum variance estimator of the expected value Θ. Furthermore, the convergence of U towards Θ is controlled by precise theorems: the Laws of Large Numbers, the Law of the Iterated Logarithm, the Law of Berry-Esseen, and the Central Limit Theorem.

An incomplete U-statistic is often defined in the literature as one associated with a symmetric kernel, namely as a sum of the form $K^{-1}\sum_{S\in\mathscr{S}} h(z_{S_1},\dots,z_{S_k})$, where h is a symmetric function and $\mathscr{S}$ is a collection of k-subsets S. We write $|\mathscr{S}| =: K$ because it generalizes the corresponding nomenclature in K-fold cross-validation. Since h is symmetric, it suffices to extend the summation over collections of increasing subsets, and an evaluation of h is already determined by its evaluation on increasing indices: each subset S can be written as $S = (S_i)$ such that $1 \le S_1 < \cdots < S_k \le n$.

Here, we will consider statistics of the more general form $|\mathscr{R}|^{-1}\sum_{S\in\mathscr{R}} h(z_{R_1},\dots,z_{R_k})$ where h is not necessarily symmetric, and therefore $\mathscr{R}$ is a collection of arbitrary *ordered*, but not necessarily increasing, subsets $R = \{R_1,\dots,R_k\}$. Variance-minimizing designs have been set up for incomplete U-statistics with symmetric kernels but not yet for those with not necessarily symmetric kernels. We will do so in the special case of $h = \Gamma$. One could consider variance minimizing designs associated with the symmetrization Γ_0 (as defined below) but the variance can be reduced further in the general case.

2.4. A test for the comparison of two average prediction errors.

Here, we give a short, self-contained overview of the results of Fuchs et al. [5]. One defines

$$\Gamma_0(1,\dots,g+1) := (g+1)^{-1}\sum_{\pi}\Gamma(\pi(1),\dots,\pi(g);\pi(g+1))$$

where the sum is taken over all $g+1$ cyclic permutations π of $1,\dots,g+1$, namely all permutations of the form $(1,\dots,g+1)\mapsto(q,\dots,g+1,1,\dots,q-1)$, where $q\in\{1,\dots,g+1\}$. Then Γ_0 is the leave-one-out version of Γ, and Γ_0 is a symmetric function of $g+1$ vector arguments. Therefore, Γ_0 defines a U-statistic, and sorting out the terms shows that this U-statistic is the leave-p-out estimator of the error [1] where $p := n-g$ (this definition holds for the rest of the paper). Likewise, Γ_0 is obtained from Γ by symmetrizing over all $(g+1)!$ permutations; the sum then simplifies to the cyclic permutations because all learning observations are treated equally.

Let $\mathscr{T}_\star$ or, when the sample size is needed, $\mathscr{T}_{\star,n}$ denote the maximal design, consisting of all $\binom{n}{g}(n-g)$ possible pairs $(S;a)$. Then, the U-statistic associated with the symmetric kernel Γ_0 is $\widehat{\Theta}(\mathscr{T}_\star)$, the leave-$p$-out estimator.

An important consequence of identifying the leave-p-out estimator as a U-statistic is that it has minimal variance among all estimators of the error rate. Also, all of the many properties of U-statistics, such as asymptotic normality and so on, automatically apply to the leave-p-out estimator $\widehat{\Theta}(\mathscr{T}_\star)$.

We implicitly assume

Assumption 1. *The degree of* Θ *is exactly* $g+1$*. Similarly, the degree of* Θ^2 *is* $2g+2$*.*

Remark 2. It seems to be very hard to prove analytically the first part of the assumption, or to give numerical evidence. However, it seems to be very intuitive to assume that the true error can not be achieved by a smaller learning set size than g, across all distributions P. The second part of the assumption is violated, for instance, if $\sigma_1^2 = 0$ (defined in Definition 4), which corresponds to the case that the U-statistic is degenerate. It is unclear whether the second part of the assumption can be violated if the U-statistic is non-degenerate.

Furthermore, it turns out that the variance of a U-statistic, trivially given by $U^2-\Theta^2$, is another regular parameter and can therefore be estimated by a U-statistic. However, under

Assumption 1, the variance is a U-statistic of twice the degree of that of the underlying U-statistic, and therefore, there is no unbiased estimator of the variance of the leave-p-out error estimator unless $n \geq 2(g+1)$. Therefore, the learning set size must be less than half the total sample size.

However, under this constraint, studentization is possible because of the consistence of the variance estimator, the Laws of Large Numbers, and Slutsky's theorem. This leads to the fact that the standardized statistic $(U^2 - \widehat{\Theta^2})^{-1/2}U$ is approximately normal, implying that there is an approximately exact test for the comparison of the losses of two statistical procedures [5].

3. The core covariances and their theoretical properties

In the following, we will generalize the variance decomposition of Bengio and Grandvalet [3, Formula (7)] to arbitrary designs. Thus, we will derive the general formula for the variance of a resampling procedure. In particular, we will take advantage of the fact that the large number of covariance terms occurring in the variance of a resampling procedure reduces to a few core covariance terms which we will call $\tau_d^{(i)}$.

Our goal is the variance decompositions in formulas (14) and (19). These are variance decompositions of incomplete U-statistics associated with only partially symmetric kernels. In the particular case where the kernel is symmetric (which does not happen for kernels of the form (3)), we recover part of the variance decomposition of incomplete U-statistics as in Lee [9, Chapter 4].

However, it is quite important to note that our variance decomposition (19) is somewhat analogous to, but *does not* reduce to, the variance decomposition of Lee [9, Chapter 4, Formula (2)]. In fact, our quantities B_γ only refer to the learning sets and are therefore different from Lee's B_γ's.

Definition 1. Let $S = \{1,\dots,g\}$, $a = g+1$, $S' = \{g+1,\dots,2g+1\}$, $a' = 2g+2$. Then, the functional Θ^2 is defined by

$$\Theta^2(P) = \int \cdots \int \Gamma(S;a)\Gamma(S';a')d^{2g+2}P(Z_1,\dots,Z_{2g+2}).$$

This is a regular parameter of degree at most $2g+2$. In the case that Θ is degenerate (meaning that $\sigma_1 = 0$ for all P where σ_1 is defined in (4)), $\Theta^2 = \mathbb{E}(\Gamma_0(1,\dots,g+1)\Gamma_0(g+1,\dots,2g+1))$ and therefore it is of smaller degree, it seems reasonable to assume that this is the only way Θ^2 can have smaller degree.

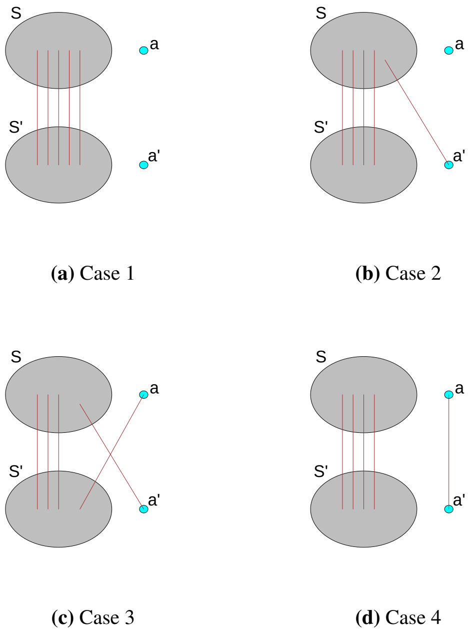
3.1. The four series - definition.

Let us now consider products of two evaluations of Γ where the index sets overlap in d indices, but there is either no overlap in the test indices, or one test observation occurs in the learning observation of the other, or both test observations occur in the other's learning set, respectively, or both test observations coincide. These four cases are illustrated by Figure 1 and describe all possible configurations.

Definition 2.

$$\tau_d^{(i)} := -\Theta^2 + \begin{cases} \mathbb{E}(\Gamma(1,\dots,g;g+1)\Gamma(1,\dots,d,g+2,\dots,2g+1-d;2g+2-d)), & i=1 \\ \mathbb{E}(\Gamma(1,\dots,g;g+1)\Gamma(1,\dots,d-1,g+1,\dots,2g+1-d;2g+2-d)), & i=2 \\ \mathbb{E}(\Gamma(1,\dots,g;g+1)\Gamma(1,\dots,d-2,g+1,\dots,2g+2-d;d-1)), & i=3 \\ \mathbb{E}(\Gamma(1,\dots,g;g+1)\Gamma(1,\dots,d-1,g+2,\dots,2g+2-d;g+1)), & i=4 \end{cases}$$

for $d = 1,\dots,g+1$, and the exceptional cases $\tau_0^{(i)} = 0$ for all i, and $\tau_1^{(3)} = \tau_{g+1}^{(1)} = \tau_{g+1}^{(2)} = 0$.



(a) Case 1 **(b)** Case 2

(c) Case 3 **(d)** Case 4

Figure 1. Let S, a and S', a' be any pair of g-subsets S and S', and $a \notin S$, $a' \notin S'$. Then $Cov(\Gamma(S;a), \Gamma(S';a'))$ only depends on which of the four cases describes the overlap pattern. Here: example for $d = 5$

Remark 3. Therefore, the quantity σ^2 from Bengio and Grandvalet [3] appears in this classification as $\tau^{(4)}_{n-n/K+1}$ where n is the total sample size and K is the number of blocks of cross-validation, their ω is our $\tau^{(1)}_{n-n/K}$ and their γ is our $\tau^{(3)}_{n-2n/K+1}$. The seemingly more complicated nomenclature, involving lower indices, allows for the treatment of any resampling procedure instead of only cross-validation.

Notational Convention 1. *Throughout this work, we denote the total overlap size*

$$\left|(S \cup \{a\}) \cap (S' \cup \{a'\})\right|$$

between two evaluation tuples by the letter d, and the overlap between two learning sets $|S \cap S'|$ by the letter c.

The interest in these quantities is that *any* occurring covariance between evaluations of Γ is equal to one of them. Note that there is an astronomical number of possible pairs of evaluations of Γ, but there are only $4g+1$ quantities $\tau^{(i)}_c$ unequal to zero.

Observation 1. Let S, a and S', a' be any pair of g-subsets S and S', and $a \notin S$, $a' \notin S'$. Then $Cov(\Gamma(S;a), \Gamma(S';a')) = \tau^{(i)}_{|(S \cup \{a\}) \cap (S' \cup \{a'\})|}$ for some $i = 1, \ldots, 4$ that describes the overlap pattern. This is obvious from the fact that Γ is symmetric in the learning indices, and the product measure $d^n P$ is permutation invariant.

3.2. σ^2_d as a linear combination of the core covariances.

Let us define

$$\sigma^2_d := \mathbb{E}(\Gamma_0(1, \ldots, g+1)\Gamma_0(g+2-d, \ldots, 2g+2-d)) - \Theta^2 \tag{4}$$

for $d = 1, \dots, g+1$. (σ_d^2 is called ζ_d in Hoeffding [7].) Thus, σ_d^2 measures the covariance between two symmetrized kernels whose overlap has size d. By a short computation, these numbers can be seen to be conditional variances, hence they are non-negative and it is justified to define them as squares. By plugging in the definition of Γ_0 and expanding the sum we arrive at the following expression in terms of the four series:

$$\sigma_d^2 = \frac{1}{(g+1)^2}\Big((g+1-d)^2 \cdot \tau_d^{(1)} + 2d(g+1-d)\cdot\tau_d^{(2)} + d(d-1)\cdot\tau_d^{(3)} + d\cdot\tau_d^{(4)}\Big). \tag{5}$$

In particular, we see that the right hand side must be non-negative.

The asymptotic variance of the complete U-statistic, the leave-p-out estimator, is $(g+1)^2\sigma_1^2/n$ [7, 5.23] (recall that $p = n-g$). So, the limiting variance is

$$\lim_{n\to\infty} n\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n})) = g^2\tau_1^{(1)} + 2g\tau_1^{(2)} + \tau_1^{(4)}, \tag{6}$$

where the limit is taken for g fixed.

3.3. Possible values for $\tau_d^{(i)}$.

Furthermore, plugging (5) into the inequality

$$\sigma_d^2/d \le \sigma_{d'}^2/d' \tag{7}$$

[7, 5.19] for $d \le d'$ puts constraints on the $\tau_d^{(i)}$, on top of those from Bengio and Grandvalet [3, Section 6].

Some of the $\tau_d^{(i)}$ can be identified as variances. Let us define the conditional expectations

$$\Gamma_{\flat,d}^{(i)}(z_1,\dots,z_d) := \begin{cases} \mathbb{E}(\Gamma(1,\dots,g;g+1)|Z_1 = z_1,\dots,Z_d = z_d) & \text{if } i = 1 \\ \mathbb{E}(\Gamma(1,\dots,g;g+1)|Z_1 = z_1,\dots,Z_{d-1} = z_{d-1}, Z_{g+1} = z_d) & \text{if } i = 4 \end{cases}$$

Under favorable regularity conditions on f and P, for instance those analogous to those stated in connection with Cramér [4, 21.1.5], these functions possess the more explicit form

$$\Gamma_{\flat,d}^{(i)}(z_1,\dots,z_d) := \begin{cases} \int\cdots\int G(z_1,\dots,z_d,Z_{d+1},\dots,Z_g;Z_{g+1})dP(Z_{d+1})\dots dP(Z_g)dP(Z_{g+1}), & i = 1 \\ \int\cdots\int G(z_1,\dots,z_{d-1},Z_d,\dots,Z_g;z_d)dP(Z_d)\dots dP(Z_g), & i = 4. \end{cases}$$

where we have to use the function G as a slightly different notation for Γ, namely

$$G(Z_1,\dots,Z_g;Z_{g+1}) := \Gamma(1,\dots,g+1)$$

Let us denote the $b\times b$-matrix implementing the binomial transform by P or $P(b)$, thus $P_{ij} = (-1)^{i+j}\binom{i}{j}$ for $1 \le j \le i \le b$ and $P_{ij} = 0$ for $j > i$. For any U-statistic U, let us denote the associated quantities σ_d and δ_d from [7, 5.27] by $\sigma_d(U)$ and $\delta_d(U)$, respectively. We then have $\delta(U) = P\sigma(U)$ as vectors with index $d = 1,\dots,\deg(U)$.

The following lemma puts strong constraints on the possible values of $\tau_d^{(1)}$ and $\tau_d^{(4)}$; these constraints do not appear in the treatment by Bengio and Grandvalet [3, Section 6].

Lemma 1. *For any $d = 1,\dots,g$, the quantity $\tau_d^{(1)}$ is the quantity $\sigma_d^2(U_\flat^{(1)})$ of the U-statistic $U_\flat^{(1)}$ associated with the kernel $\Gamma_{\flat,g}^{(1)}$ of degree g. Consequently,*

$$\tau_d^{(1)} = \mathbb{V}(\Gamma_{\flat,d}^{(1)}) \ge 0, \tag{8}$$

and

$$\tau_d^{(1)}/d \le \tau_e^{(1)}/e \tag{9}$$

for any $1 \leq d < e \leq g$*, and*

$$P\tau^{(1)} = P\sigma(U\flat^{(1)}) = \delta(U_\flat^{(1)}) \geq 0 \tag{10}$$

for the binomial matrix $P = P(g)$*.*
For type (4)*, one has only*

$$\tau_d^{(4)} = \mathbb{V}(\Gamma_{\flat,d}^{(4)}) \geq 0. \tag{11}$$

Proof. $\Gamma_{\flat,g}^{(1)}$ is a function of g arguments. Plugging in the random variables Z_i, we arrive at a random variable $\Gamma_{\flat,g}^{(1)}(Z_1,\ldots,Z_g)$. Using the fact that $\Gamma_{\flat,g}^{(1)}$ is symmetric, one obtains

$$\sigma_d(U_\flat^{(1)}) = Cov(\Gamma_{\flat,g}^{(1)}(Z_1,\ldots,Z_g),\Gamma_{\flat,g}^{(1)}(Z_1,\ldots,Z_d,Z_{g+1},\ldots,Z_{2g-d})).$$

Writing this covariance as the difference of the expectation of the product and the product of the expectations, the first term is seen to have overlap pattern of type $(d,(1))$, and the second is Θ^2, thus the first claim.
The claim on type (4) ensues analogously. □

In contrast, no such assertion seems to hold for the series (2) and (3), and there seems to be no positivity statement. Also, there seems to be no reason why $\tau^{(4)}$ could be identified with the quantities σ_d of some U-statistic, nor why $P\tau^{(4)}$ should be non-negative.
In contrast, Bengio and Grandvalet [3] also give constraints that are not covered by Lemma 1.

Lemma 2. (1) $\tau_d^{(4)} \geq \tau_{d-1}^{(1)}$ *for all* d*.*
(2) $-\tau_d^{(4)}/2 \leq \tau_2^{(3)} \leq \tau_d^{(4)}$ *for all* d*.*
(3) $\left|\tau_d^{(i)}\right| \leq \tau_g^{(4)}$ *for all* d *and* i*.*

Proof. The first two statements follow from plugging in all possible values for n and K into Bengio and Grandvalet [3, Lemma 8]. The third statement is the Cauchy-Schwarz inequality. □

3.4. Variance decomposition of incomplete U-statistics.

Let us turn our attention to the general incomplete U-statistic associated with a collection $\mathscr{T}$ of pairs (S,a) of a learning set S and a test observation $a \notin S$. We will briefly denote an overlap size and type of pattern by $\Psi((S,a),(S',a')) = (d,(i))$ when $|(S\cup\{a\})\cap(S'\cup\{a'\})| = d$ and the type is (i), and will then write $\tau(\Psi((S,a),(S',a'))$ instead indicating the type of the overlap pattern with lower and upper indices.
The variance of the cross-validation-like procedure associated with the collection $\mathscr{T}$ is

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T})) = |\mathscr{T}|^{-2}\sum_{i,j} Cov(\Gamma(S_i;a_i),\Gamma(S_j;a_j)) = |\mathscr{T}|^{-2}\sum_{i,j}\tau(\Psi((S,a),(S',a'))), \tag{12}$$

which is convenient because the last sum can be written as a linear combination with much fewer summands because many summands take the same value.

3.5. Variance decomposition of test-complete designs.

Definition 3. (1) Consider the following linear combination of the $\tau_d^{(i)}$:

$$\begin{aligned}\xi_c :=&(n-2g+c)(n-2g+c-1)\cdot\tau_c^{(1)} + 2(g-c)(n-2g+c)\cdot\tau_{c+1}^{(2)} + \\ &+ (g-c)^2\cdot\tau_{c+2}^{(3)} + (n-2g+c)\cdot\tau_{c+1}^{(4)}\end{aligned} \tag{13}$$

for all $c = 0,\ldots,g$, where we define $\tau_{g+2}^{(3)} = 0$.

(2) Furthermore, let us call a design $\mathscr{T}$ *test-complete* whenever the following holds: $(S,a) \in \mathscr{T} \implies (S,b) \in \mathscr{T}$ for any $b \notin S$. In words, a design is test-complete whenever it contains, together with a learning set S, the combinations of S with all possible test observations. Note that a test-complete design is uniquely specified by the learning sets it contains. Whenever a test-complete design $\mathscr{T}$ is specified by the collection of learning sets it contains, we will write $\mathscr{S}$ for the collection of learning sets, where each learning set S is counted only once even if it occurs in several pairs (S,a). Thus, $|\mathscr{T}| = K(n-g)$ (of course, we suppose $\mathscr{S}$ to contain each learning set only once).

(3) Let $\mathscr{T}$ be a test-complete design. For any $c = 0,\dots,g$, let $f_c^\ell \in \mathbb{N}_0$ be the number of ordered pairs of learning sets (S,S'), both occurring in $\mathscr{T}$, such that $|S \cap S'| = c$. Pairs (S,S) with the same learning set occurring twice are also allowed (where ℓ is a mere symbol instead of an index).

For instance, any cross-validation design is test-complete. The same holds for the complete design defining the leave-p-out estimator. For any test-complete design, the associated numbers f_c^ℓ are easily computable. For instance, they are given by the number of entries equal to c in $N^T N$, where N is the incidence matrix of the learning sets occurring in the design. Obviously, only test-complete designs seem to be relevant in practice because of the low computational cost of evaluating the loss function for a given model and given test observations.

Theorem 1. *Let $\mathscr{T}$ be a test-complete design and let $\mathscr{S}$ be the associated collection of learning sets. Then, the variance of the error estimator satisfies*

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T})) = |\mathscr{T}|^{-2} \sum_{c=0}^{g} f_c^\ell \xi_c \tag{14}$$

where ξ_c was defined in (13).

Proof. This follows from expanding the variance as in (12) into the form $|\mathscr{T}|^{-2}$ multiplied by the sum of all entries of the $|\mathscr{T}| \times |\mathscr{T}|$-covariance matrix between the non-rescaled summands of $\widehat{\Theta}(\mathscr{T})$ and counting the terms. Each entry of the covariance matrix is described by two pairs $(S,a),(S',a')$ and therefore defines a specific type $(1),\dots,(4)$ of the overlap pattern between (S,a) and (S',a'), and a particular overlap size $d = |(S \cup \{a\}) \cap (S' \cup \{a'\})|$. Any two summands of the same type (i) and the same overlap size d are equal, namely $\tau_d^{(i)}$. Now, counting and summing up all such terms with learning overlap size c, one obtains ξ_c. This implies the result. □

Minimization of the expression $\sum f_c^\ell \xi_c$ seems to be very hard in practice. However, we will outline below a few cases where this task is feasible.

Example 1 (Variance of cross-validation)**.** *Let us assume n is divisible by K, and that therefore the learning sets have size $g = n - n/K$. We then arrive at the following. For K-fold cross-validation, $K \geq 2$, we count*

$$f_c^\ell = \begin{cases} 0, & c \notin \{n-n/K, n-2n/K\} \\ K, & c = n-n/K \\ K^2 - K, & c = n-2n/K. \end{cases} \tag{15}$$

The variance of cross-validation is given by the formula

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T})) = (K^{-1} - n^{-1})\tau_{n-n/K}^{(1)} + (1-K^{-1})\tau_{n-2n/K+2}^{(3)} + n^{-1}\tau_{n-n/K+1}^{(4)} \tag{16}$$

In the case $K=2$, we obtain the expression

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T}))=\frac{1}{n}\left((n/2-1)\tau^{(1)}_{n/2}+n/2\cdot\tau^{(3)}_{2}+\tau^{(4)}_{n/2+1}\right) \tag{17}$$

Since it is unclear whether and how fast the $\tau^{(i)}_d$ converge to zero, one can not immediately deduce asymptotic statements from (17).

3.6. Non-asymptotic minimization of $f^\ell_c\xi_c$.

Definition 4. Let $\mathscr{T}$ be a test-complete design. For $\gamma=1,\dots,g$ and a subset $s\subset\{1,\dots,n\}$ such that $|s|=\gamma$, let $n(s)$ be the number of learning sets S in the design (where each single learning set is counted only once) such that $s\subset S$. Let $B^\ell_\gamma:=\sum_s n(s)^2$, where the sum is taken over all $\binom{n}{\gamma}$ subsets s. Analogously, let $B_0:=K^2=|\mathscr{T}|^2\,(n-g)^{-2}=\sum_{c=0}^g f^\ell_c$.

Lemma 3. *The quantities f^ℓ_c are uniquely determined by the B^ℓ_γ. In fact, $f^\ell_c=\sum_{\gamma=c}^g(-1)^{\gamma-c}\binom{\gamma}{c}B_\gamma$ for all $0\le c\le g$.*

Proof. For $1\le c\le g$, the proof proceeds in complete analogy to the proof of Lee [9, Chapter 4, Equation (7)], even though our f^ℓ_c and B_γ are quite different from Lee's f_c and B_γ. For $c=0$, one has $f^\ell_0=\sum_{c=0}^g f^\ell_c-\sum_{c=1}^g f^\ell_c=B_0-\sum_{c=1}^g\sum_{\gamma=c}^g B_\gamma=\dots=B_0+\sum_{\gamma=1}^g(-1)^\gamma B_\gamma=\sum_{\gamma=0}^g(-1)^\gamma B_\gamma$, using that $\sum_{c=1}^\gamma(-1)^c\binom{\gamma}{c}=-1$ for all $\gamma\ge c$. □

Let us write this result in the form $f^\ell=PB$ for the upper-triangular matrix P defined by $P_{c,\gamma}=(-1)^{\gamma-c}\binom{\gamma}{c}$ for all $0\le c\le\gamma\le g$ and $P_{c,\gamma}=0$ for $\gamma<c$, where $\binom{\gamma}{0}:=1$ for all $\gamma\ge0$ (The map described by the matrix P is often called the binomial transform.) Using (14), we can now write $\mathbb{V}(\widehat{\Theta}(\mathscr{T}))=|\mathscr{T}|^{-2}<f^\ell,\xi>=|\mathscr{T}|^{-2}<PB,\xi>=|\mathscr{T}|^{-2}<B^\ell,P^T\xi>$. For this reason, we consider the binomial transformation $P^T\xi$ of the vector ξ separately:

Definition 5.

$$\alpha_\gamma:=\sum_{c=0}^{\gamma}(-1)^{\gamma-c}\binom{\gamma}{c}\xi_c \text{ for all } 0\le\gamma\le g \tag{18}$$

Thus, we have shown that

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T}))=|\mathscr{T}|^{-2}\sum_{\gamma=0}^{g}B_\gamma\alpha_\gamma, \tag{19}$$

and in order to minimize this, we have to maximize those B_γ for which α_γ is negative, and minimize those for which it is positive. This stands in contrast to the classical case where all B_γ have to be minimized.

The usefulness of (19) lies in the fact that in the case that ξ_c is a polynomial of small degree in c, all α_γ vanish when γ is greater than the polynomial's degree because $\sum_{c=0}^\gamma(-1)^c\binom{\gamma}{c}c^d=0$ for any $d<\gamma$. In Section 4, we will exhibit a case where the ξ_c is a polynomial of degree two, and in Section 6 we will give numerical evidence that the ξ can more often be approximated well by a quadratic polynomial. Precisely, if ξ is of degree one, we have $\xi_c=b+Ac$ and then $\alpha_0=b,\alpha_1=A$ and $\alpha_\gamma=0$ for $\gamma\ge2$. If ξ is of degree two, we have $\xi_c=b+Ac+Cc^2$, and then it is easy to calculate that

$$\begin{aligned}\alpha_0&=b\\ \alpha_1&=A+C\\ \alpha_2&=2C\\ \alpha_\gamma&=0 \text{ for all } \gamma\ge3.\end{aligned} \tag{20}$$

4. Analytical computation of the core covariances in a toy intercept estimation model

Let us consider the following simple example. The random variable X is univariate and distributed according to some unknown distribution P_X, and the joint distribution of (Y,X) is given by the simple model $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon \sim \mathscr{N}(0,v)$ with β_0 and v unknown. This model is a close relative of that used in univariate ordinary regression where the slope coefficient β_1 is known and only the intercept is estimated.

We show the following facts in the supplement 6.4: There is an explicit formula for the kernel Γ, the $\tau_c^{(i)}$ are quadratic polynomials in c which we write down, consequently, ξ_c is a quadratic polynomial in c as well, and α_γ is non-zero if and only if $\gamma = 0,1,2$.

A first consequence is that by (19), two designs have the same variance already as soon as they have the same B_0, B_1 and B_2.

Example 2. *The calculations of 6.4 can be used to compare the variance of cross-validation with that of the leave-p-out estimator in closed form expressions. The full U-statistic associated with the kernel Γ, i.e., the leave-p-out estimator on a sample of size n, is equal to $\hat{v}(1+g^{-1})$. Since $\hat{v} \sim v(n-1)^{-1}\chi^2_{n-1}$, the leave-p-out estimator is distributed as the $v(1+g^{-1})(n-1)^{-1}$-fold of a chi-square of $n-1$ degrees of freedom.*

Therefore, the variance of the leave-p-out estimator is $\mathbb{V}(v(1+g^{-1})(n-1)^{-1}\chi^2_{n-1}) = v^2(1+g^{-1})^2(n-1)^{-2}\cdot 2(n-1) = 2v^2(1+g^{-1})^2(n-1)^{-1}$.

This is consistent with the fact that by (6) and (5), we have $\lim_{n\to\infty} n\mathbb{V}(\widehat{\Theta}(\mathscr{T}_\star)) = 2(g^{-2} + 2g^{-1}+1)v^2$ which could also be derived from the expression $\mathbb{V}(\widehat{\Theta}(\mathscr{T}_\star)) = |\mathscr{T}_\star|^{-2}\sum B_\gamma \alpha_\gamma$. So, we have derived the rescaled limiting variance of the leave-p-out estimator in three ways.

In contrast, for the design $\mathscr{T}_{2\text{-}CV}$ describing two-fold cross-validation ($g = n/2$), we obtain by (17):

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{CV})) = 2v^2\left[n^{-1} + 14n^{-2}\right]. \tag{21}$$

Thus, the ratio $\mathbb{V}_{2\text{-}CV} / \mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n}))$ is one for $n=2$ and tends to one for $n \to \infty$, and attains its maximum, 25/16, for $n = 6$. Note that here g and n both tend to infinity, in contrast to the rest of the paper:

$$\lim_{n\to\infty} n\mathbb{V}(\widehat{\Theta}(\mathscr{T}_\star)) = \lim_{n\to\infty} n\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{CV})) = 2v^2. \tag{22}$$

Also, one can check that $\mathbb{V}(\widehat{\Theta}(\mathscr{T}_\star)) < \mathbb{V}(\widehat{\Theta}(\mathscr{T}_{CV})$ for all n, as it should.

Let us go back to the usual scenario of fixed g, and let n tend to infinity.

Thus, in this example, the limiting variance $2v^2$ agrees in all three cases: where g is fixed, the case of cross-validation with $g = n/2$, and the leave-p-out case where $g = n/2$. Note also that minimizing $\sum B_\gamma \alpha_\gamma$ involves only three non-zero summands whereas $\sum f_c^\ell \xi_c$ involves $g+1$ summands. Therefore, the minimization problem's dimensionality is drastically reduced when passing from the ξ_c to the α_γ.

Let us now show how to apply our calculations to the variance minimization problem. Let us say we are given fixed values for n, g and K. The problem is to find a design that minimizes the expression $B_1\alpha_1 + B_2\alpha_2$, because the pre-factor as well as the summand corresponding to $c = 0$ of (19) can be ignored because they are determined by the pre-set quantity K.

Let us assume that each observation occurs in the same number of learning sets. This is analogous to the usual restriction to equireplicate designs as in Lee [9, Section 4.3.2]), and we also call such designs equireplicate, even though we are only referring to the learning sets. In such designs, the condition that $B_1 = K^2 g^2/n$ is imposed. Thus, only B_2 remains as a degree of freedom in the optimization, eliminating any trade-off between competing components. Since $\alpha_2 > 0$, B_2 has to be minimized. Subsuming the results of this section, we have shown the following:

Theorem 2. *In the intercept estimation model of this chapter, all equireplicate designs –for fixed n, g and K– that have the same B_2 have the same variance. Any equireplicate design with minimal B_2 among all equireplicate designs achieves the minimal variance among all equireplicate designs of the same n, g, and K. Assuming that the configuration of n, g, and K allows for the existence of a Balanced Incomplete Block Design (see Definition 7), any Balanced Incomplete Block Design of these n, g, and K is a design with minimal variance among all equireplicate designs of these n, g, and K.*

Proof. It only remains to show the last assertion. This is done in complete analogy to the proof of Lee [9, Chapter 4, Theorem 1]. □

For instance, for $g = 2$, B_2 is bound to be equal to K, and therefore all equireplicate designs have the same variance. Another simple example is the leave-one-out case $g = n-1, K = n$. Then, the minimality of the design's variance has been unveiled to be the minimality of a symmetric Balanced Incomplete Block Design's variance.

Since the α_γ, unlike those in the classical context, can happen to be negative, one might ask whether there exists a configuration (n, g, K) such that an equireplicate design exists but a non-equireplicate design has smaller variance than the best equireplicate one. Such a non-equireplicate design would then maximize B_1 instead of minimizing B_2. Thus, it would be, in some sense, the "opposite" of an equireplicate design.

It seems that whenever ξ_c is a polynomial in c of small degree, arguments similar to those in this chapter can be used to determine equireplicate minimal-variance designs in a non-empirical way.

5. A general Central Limit Theorem and a hypothesis test

5.1. **The core covariance as regular parameters and their estimation.** Let us recall that a linear combination of regular parameters is a regular parameter [7, middle of Page 295]. This allows us to split off the common regular parameter Θ^2 from an integral that is specific for the overlap pattern, in the following sense: each quantity $\tau_d^{(i)}$ can be written as

$$\tau_d^{(i)}(P) = \int \cdots \int \Gamma(S;a)\Gamma(S';a')d^{2g+2-d}P(Z_1,\ldots,Z_{2g+2-d}) - \Theta^2(P)$$

where the overlap pattern between (S,a) and (S',a') is of type (i) and $|(S\cup\{a\})\cap(S'\cup\{a'\})| = d$. Note that this implies that $|S\cup\{a\}\cup S'\cup\{a'\}| = 2g+2-d$, so the number of integrals in the first summand is correctly specified. Therefore, $\tau_d^{(i)}$ is a regular parameter of degree at most $2g+2$.

Lemma 4. *If Θ^2 has degree $2g+2$, $\tau_d^{(i)}$ is a regular parameter of degree exactly $2g+2$.*

Proof. Assume, there was a function f of only $2g+1$ arguments such that

$$\tau_d^{(i)}(P) = \int \cdots \int f d^{2g+2-1}P$$

for all P. This covers the general case because if there were a function with even fewer arguments then there would also be a function with $2g+1$ arguments, defined by ignoring the additional ones. Then, by linearity, $f - \Gamma(S;a)\Gamma(S';a')$ would be a kernel of degree $2g+1$ for Θ^2, in contradiction to Assumption 1. □

Definition 6. We define statistics $\widehat{\tau_d^{(i)}}$ as the U-statistics associated with these regular parameters.

As with U-statistics, they satisfy several optimality properties.

Corollary 1. *The estimators $\widehat{\tau_d^{(i)}}$ are the minimal variance unbiased estimators for $\tau_d^{(i)}$. They are consistent and satisfy the Weak and Strong Law of Large Numbers.*

Lemma 5. *Since Θ^2, ξ_c and α_γ are regular parameters, there exist U-statistics $\widehat{\Theta^2}, \widehat{\xi_c}$, and $\widehat{\alpha_\gamma}$. Let us abbreviate the U-statistic for the regular parameter $\mathbb{V}(\widehat{\Theta}(\mathscr{T}))$ as $[\mathbb{V}(\widehat{\Theta}(\mathscr{T}))]\widehat{}$. Then*

$$[\mathbb{V}(\widehat{\Theta}(\mathscr{T}))]\widehat{} = (\widehat{\Theta}(\mathscr{T}))^2 - \widehat{\Theta^2} = |\mathscr{T}|^{-2}\sum_{c=0}^{g} f_c^\ell \widehat{\xi_c} = |\mathscr{T}|^{-2}\sum_{\gamma=0}^{g} B_\gamma \widehat{\alpha_\gamma}.$$

Likewise, the estimators $\widehat{\xi_c}$ satisfy the empirical analog to (13).

This is in analogy to

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T})) = \mathbb{E}[(\widehat{\Theta}(\mathscr{T}))^2] - \Theta^2 = |\mathscr{T}|^{-2}\sum_{c=0}^{g} f_c^\ell \xi_c = |\mathscr{T}|^{-2}\sum_{\gamma=0}^{g} B_\gamma \alpha_\gamma.$$

Proof. This fact is not obvious but can be checked by straightforward computation. □

5.2. **Variance estimation of cross-validation.** In the following, we are referring to a situation where n observations are used to carry out cross-validation, and there exist n' so-called "extra" observations such that the total number of observations satisfies $n+n' \geq 2g+2$. Then, there exists a variance estimator for cross-validation, which may be contrasted with Bengio and Grandvalet [3]. Precisely, plugging (15) into Theorem 14, we obtain:

Theorem 3. *The empirical counterpart of the right hand-side of (16) is a U-statistic of degree $2g+2$ under Assumption 1.*
It defines the unique minimal variance unbiased estimator of the variance of cross-validation, if there are enough "extra observations". Otherwise, no unbiased estimator exists. This U-statistic $[\mathbb{V}(\widehat{\Theta}(\mathscr{T}))]\widehat{}$ is identical to the plug-in estimator

$$(K^{-1} - n^{-1})\widehat{\tau_{n-n/K}^{(1)}} + (1-K^{-1})\widehat{\tau_{n-2n/K+1}^{(3)}} + n^{-1}\widehat{\tau_{n-n/K+1}^{(4)}}$$

5.3. **The Weak Law of Large Numbers.** There is the following general criterion on the resampling design for the Weak Law of Large Numbers to hold.
Assume that for each sample size n a design $\mathscr{T}_n$ with learning sets $\mathscr{S}_n$ is given. We will only consider the case where the learning set sizes g are the same across all n, so $\lim_{n\to\infty} g/n = 0$. Let us write $K_n := |\mathscr{S}_n|$.

Theorem 4. *Assume that $U_\flat^{(1)}$ is non-degenerate in the sense that $\tau_1^{(1)} = \sigma_1^2(U_\flat^{(1)}) \neq 0$. Then, the following are equivalent:*

(1)

$$\lim_{n\to\infty} f^{\ell}_{0,n}/(\sum_{c=0}^{g} f^{\ell}_{c,n}) = 1. \tag{23}$$

(2) *$\widehat{\Theta}(\mathscr{T}_n)$ is weakly consistent in the sense that $\widehat{\Theta}(\mathscr{T}_n)$ converges in probability to Θ as $n \to \infty$.*

Proof. (1) $\implies$ (2): By (13), the quantity $\xi_{c,n}$ is $O(n)$ for $c=0$ because $\tau_0^{(1)} = 0$ and is $O(n^2)$ for $c \geq 1$. Therefore, the summand

$$|\mathscr{T}_n|^{-2} f^{\ell}_{0,n}\xi_{0,n} = (n-g)^{-2}(\sum_{c=0}^{g} f^{\ell}_{c,n})^{-1} f^{\ell}_{0,n}\xi_{0,n}$$

of (14) always vanishes, taking into account that $\sum_{c=0}^{g} f^{\ell}_{c,n} = (n-g)^{-2}|\mathscr{T}_n|^2$ for a test-complete design.
Similarly, the remaining summand of the decomposition (14) is

$$(n-g)^{-2}(\sum_{c=0}^{g} f^{\ell}_{c,n})^{-1}\sum_{c=1}^{g} f^{\ell}_{c,n}\xi_{c,n}.$$

Since $(n-g)^{-2}\sum_{c=1}^{g}\xi_{c,n}$ is a bounded sequence, it follows from the condition that $\lim_{n\to\infty}\mathbb{V}(\widehat{\Theta}(\mathscr{T}_n)) = 0$, and thus the assertion.
(2) $\implies$ (1): Convergence in probability implies convergence of the variance to zero. Since

$$\lim_{n\to\infty}\xi_0/(2gn\tau_1^{(2)} + n\tau_1^{(4)}) = 1$$

$$\lim_{n\to\infty}\xi_c/(n^2\tau_c^{(1)}) = 1, c \geq 1,$$

we have

$$\lim_{n\to\infty}\frac{1}{n^2K_n^2}(f^{\ell}_{0,n}(2gn\tau_1^{(2)} + n\tau_1^{(4)}) + n^2\sum_{c\geq 1} f^{\ell}_{c,n}\tau_c^{(1)}) = \lim_{n\to\infty}\frac{1}{K_n^2}\sum_{c\geq 1} f^{\ell}_{c,n}\tau_c^{(1)} = 0.$$

Since $\tau_c^1 \geq \tau_1^{(1)} \neq 0$, this implies that $\lim K_n^{-2} f^{\ell}_{c,n} = 0$ for all $c \geq 1$. Hence,

$$\lim_{n\to\infty}\frac{f^{\ell}_{0,n}}{K_n^2} = 1 - \sum_{c\geq 1}\lim_{n\to\infty}\frac{f^{\ell}_{c,n}}{K_n^2} = 1 - 0 = 1.$$

□

Example 3. *The condition appearing in the first equivalence of Theorem 4 is satisfied, for instance, for the complete design sequence. In contrast, it is violated for a design sequence such that there is an observation that is contained in every learning set for every n. One can also construct design sequences such that the limit (23) takes values between zero and one.*

5.4. **A Central Limit Theorem.** The following theorem generalizes Lee [9, Theorem 1, Chapter 4.3.1] to U-statistics with a non-symmetric kernel and thus to CV-like procedures.

Lemma 6. *Let $\widehat{\Theta}(\mathscr{T})$ be a CV-like procedure based on a fixed design $\mathscr{T} \subset \mathscr{T}_\star$ and $\widehat{\Theta}(\mathscr{T}_\star)$ be the leave-p-out estimator.*
Then

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T})) - \mathbb{V}(\widehat{\Theta}(\mathscr{T}_\star)) = \mathbb{V}\left(\widehat{\Theta}(\mathscr{T}) - \widehat{\Theta}(\mathscr{T}_\star)\right) \geq 0. \tag{24}$$

Proof. Since

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T})-\widehat{\Theta}(\mathscr{T}_\star))=\mathbb{V}(\widehat{\Theta}(\mathscr{T}))-2Cov(\widehat{\Theta}(\mathscr{T}),\widehat{\Theta}(\mathscr{T}_\star))+\mathbb{V}(\widehat{\Theta}(\mathscr{T}_\star)),$$

(24) holds if and only if $Cov(\widehat{\Theta}(\mathscr{T}),\widehat{\Theta}(\mathscr{T}_\star))=\mathbb{V}(\widehat{\Theta}(\mathscr{T}_\star))$. Since $Cov(\Gamma(S;a),\widehat{\Theta}(\mathscr{T}_\star))$ is the same for every $(S;a)\in\mathscr{T}_\star$, we have

$$\begin{aligned}\mathbb{V}(\widehat{\Theta}(\mathscr{T}_\star)) &= Cov(\widehat{\Theta}(\mathscr{T}_\star),\widehat{\Theta}(\mathscr{T}_\star))\\ &= Cov(|\mathscr{T}_\star|^{-1}\sum_{(S;a)\in\mathscr{T}_\star}\Gamma(S;a),\widehat{\Theta}(\mathscr{T}_\star))\\ &= |\mathscr{T}_\star|^{-1}\sum_{(S;a)\in\mathscr{T}_\star}Cov(\Gamma(S;a),\widehat{\Theta}(\mathscr{T}_\star))\\ &= |\mathscr{T}_\star|^{-1}\cdot|\mathscr{T}_\star|\cdot Cov(\Gamma(S;a),\widehat{\Theta}(\mathscr{T}_\star))\\ &= |\mathscr{T}|^{-1}\cdot|\mathscr{T}|\cdot Cov(\Gamma(S;a),\widehat{\Theta}(\mathscr{T}_\star))\\ &= |\mathscr{T}|^{-1}\sum_{(S;a)\in\mathscr{T}}Cov(\Gamma(S;a),\widehat{\Theta}(\mathscr{T}_\star))\\ &= Cov(\widehat{\Theta}(\mathscr{T}),\widehat{\Theta}(\mathscr{T}_\star)).\end{aligned}$$

□

The proof differs from that of [9] not only because we generalize his theorem, but also because there is a mistake in his proof. He assumes that the covariances of an incomplete U-statistic and a kernel are all equal, for each set of the design. This property, however, is not valid in general.

We are now going to investigate a situation where a design is pre-specified for each sample size, and will give a general sufficient criterion for a Central Limit Theorem.

Let us abbreviate the complete, leave-p-out design for n by $\mathscr{T}_{\star,n}$, and its learning sets by $\mathscr{S}_{\star,n}$ such that $K_{\star,n}:=|\mathscr{S}_{\star,n}|=\binom{n}{g}$.

Suppose, again, that for each $n\geq 2g+2$, a test-complete design $\mathscr{T}_n$ for sample size n is given such that the trivial condition $K_n\to\infty$ is satisfied. For each $n\geq 2g+2$, and any $c=0,\ldots,g$, let $f^\ell_{n,c}\in\mathbb{N}_0$ be the number of ordered pairs of learning sets (S,S'), both occurring in $\mathscr{S}_n$, such that $|S\cap S'|=c$.

Recall that $|\mathscr{T}_n|^2=(n-g)^2\sum_{c=0}^{g}f^\ell_{c,n}=(n-g)^2K_n^2$. Using $f^\ell_{\star,c,n}=\binom{n}{g}\binom{g}{c}\binom{n-g}{g-c}$, we see that the numbers $f^\ell_{\star,c,n}$ satisfy the following asymptotic properties:

$$\lim_{n\to\infty}n\frac{f^\ell_{\star,1,n}}{K^2_{\star,n}}=g^2 \tag{25}$$

for $c=1$, and

$$\lim_{n\to\infty}n\frac{f^\ell_{\star,c,n}}{K^2_{\star,n}}=0 \tag{26}$$

for all $2\leq c\leq g$.

Lemma 7. *Assume $f^\ell_{c,n}$ satisfies equations (25), (26) with $f^\ell_{c,n}$ in place of $f^\ell_{\star,c,n}$ and K_n in place of $K_{\star,n}$. Then*

$$\lim_{n\to\infty}n(\mathbb{V}(\widehat{\Theta}(\mathscr{T}_n)-\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n}))=0. \tag{27}$$

Proof. Let us express the fact that ξ_c depends on n by writing $\xi_{c,n}$. One considers the order of magnitude of the $\xi_{c,n}$ as given in the proof of Theorem 4.
One substitutes (14) for $\mathbb{V}(\widehat{\Theta}(\mathscr{T}_n))$ and for $\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n}))$ and considers each summand of the left-hand side of (27) separately (the finite sum over c can be pulled outside the limit). For instance, the summand belonging to $c=0$ is zero because, by assumption,

$$\begin{aligned}\lim_{n\to\infty}((nf^{\ell}_{0,n}\xi_{0,n})/|\mathscr{T}_n|^2) &= \lim_{n\to\infty}((n^2 f^{\ell}_{0,n}/|\mathscr{T}_n|^2)\cdot\xi_{0,n}/n) = \\ \lim_{n\to\infty}(n^2 f^{\ell}_{0,n}/|\mathscr{T}_n|^2)\cdot\lim_{n\to\infty}(\xi_{0,n}/n) &= \lim_{n\to\infty}(\xi_{0,n}/n)\end{aligned}$$

and, similarly

$$\lim_{n\to\infty}((nf^{\ell}_{\star,0,n}\xi_{0,n})/|\mathscr{T}_{\star,n}|^2) = \lim_{n\to\infty}(\xi_{0,n}/n).$$

For $c=1$, we have

$$\begin{aligned}\lim_{n\to\infty}((nf^{\ell}_{1,n}\xi_{1,n})/|\mathscr{T}_n|^2) &= \lim_{n\to\infty}((n^3 f^{\ell}_{1,n}/|\mathscr{T}_n|^2)\cdot\xi_{1,n}/n^2) = \\ \lim_{n\to\infty}(n^3 f^{\ell}_{1,n}/|\mathscr{T}_n|^2)\cdot\lim_{n\to\infty}(\xi_{1,n}/n^2) &= g^2\lim_{n\to\infty}(\xi_{1,n}/n^2)\end{aligned}$$

and

$$\lim_{n\to\infty}((nf^{\ell}_{\star,1,n}\xi_{1,n})/|\mathscr{T}_{\star,n}|^2) = g^2\lim_{n\to\infty}(\xi_{1,n}/n^2).$$

Analogously, one shows that every summand belonging to any other c vanishes. □

Theorem 5. *The following are equivalent:*

(1) *Equation* (27)

(2)
$$(\widehat{\Theta}(\mathscr{T}_n)-\Theta)(\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n})))^{-1/2}\to\mathscr{N}(0,1)$$
in distribution as g remains fixed, $n\to\infty$.

The condition of Lemma 7 implies the condition (23) of Theorem 4; so we pass to a more specific case. Likewise, the situation of Theorem 4 is that a relaxation of (27) holds, namely that the left-hand side of (27) lacks the factor n.
It is easy to construct examples of design sequences $\mathscr{T}_n$ such that (23) holds but (27) is violated.

Proof. (1) $\Longrightarrow$ (2): We have

$$\begin{aligned}(\widehat{\Theta}(\mathscr{T})-\Theta)(\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n})))^{-1/2} &= (\widehat{\Theta}(\mathscr{T})-\widehat{\Theta}(\mathscr{T}_{\star,n}))(\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n})))^{-1/2} \\ &\quad+(\widehat{\Theta}(\mathscr{T}_{\star,n})-\Theta)(\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n})))^{-1/2}.\end{aligned}\tag{28}$$

We will show that the first summand converges in probability to zero while the second converges in distribution to $\mathscr{N}(0,1)$. The claim then follows from the standard fact that distribution in convergence is invariant under perturbation with a term that converges to zero in probability.
Using Lemma 6, the variance of the first summand of the right-hand side of (28) is

$$\mathbb{V}\big[(\widehat{\Theta}(\mathscr{T})-\widehat{\Theta}(\mathscr{T}_{\star,n}))(\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n})))^{-1/2}\big] = \mathbb{V}(\widehat{\Theta}(\mathscr{T}_n))/\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n}))-1,$$

which converges to zero as we have just shown.
The second summand of the right-hand side of (28) is the standardized U-statistic which satisfies the Central Limit Theorem [7, Theorem 7.1], multiplied by the square root of the ratio of the variances which converges to one. This completes the proof.
(2) $\Longrightarrow$ (1): Taking the variance of the left-hand side of (27), we see that

$\lim_{n\to\infty}\mathbb{V}(\widehat{\Theta}(\mathscr{T}_n))\,\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n}))^{-1}=1$. Together with the fact that $\lim_{n\to\infty} n\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n}))$ exists and is non-zero by (6), this implies (27). □

Corollary 2. *Further, under the conditions of the preceding theorem*

$$(\widehat{\Theta}(\mathscr{T}_n)-\Theta)(\mathbb{V}(\widehat{\Theta}(\mathscr{T}_n)))^{-1/2}\to\mathscr{N}(0,1).$$

Corollary 3. *The two-sided test of $H^0:\Theta=0$ with the rejection region*

$$\left\{\left|\widehat{\Theta}(\mathscr{T}_n)\right|\geq([\mathbb{V}(\widehat{\Theta}(\mathscr{T}_n))]\widehat{\;})^{1/2}\phi^{-1}(1-\alpha/2)\right\} \tag{29}$$

has asymptotic alpha level α, where ϕ is the standard normal cumulative distribution function. An asymptotically exact confidence interval for Θ at level $1-\alpha$ is

$$\left[\widehat{\Theta}(\mathscr{T}_n)-([\mathbb{V}(\widehat{\Theta}(\mathscr{T}_n))]\widehat{\;})^{1/2}\phi^{-1}(1-\alpha/2),\widehat{\Theta}(\mathscr{T}_n)+([\mathbb{V}(\widehat{\Theta}(\mathscr{T}_n))]\widehat{\;})^{1/2}\phi^{-1}(1-\alpha/2)\right]. \tag{30}$$

Proof. The strong consistency of U-statistics together with repeated applications of Slutsky's theorem shows that Theorem 5 remains valid when the denominator $(\mathbb{V}(\widehat{\Theta}(\mathscr{T})))^{1/2}$ is replaced by its empirical analog, the plug-in estimator $([\mathbb{V}(\widehat{\Theta}(\mathscr{T}))]\widehat{\;})^{1/2}$. □

6. A NUMERICAL ILLUSTRATION

In this section, we will present a data set together with the results of the computation of the optimal estimators of the quantities $\tau_d^{(i)}$.

We are going to illustrate how to find small-variance designs empirically in an example of a simple pre-specified and numerically convenient distribution P, and univariate regression with a non-quadratic loss. In this case, all quantities $\tau_d^{(i)}$ could be estimated.

6.1. **Layout.** We chose the following univariate scenario. Predictors X are uniformly distributed on $[0,2]$, and the response variable is given by $Y=X^2$. This specifies the joint distribution P of (X,Y). To a good degree of approximation, the predictors in the data could be taken to be equi-spaced on $[0,2]$ instead of being drawn independently.

Thus, for $i\in\{1,\dots,n\}$ we have

$$x_i=2\cdot i/n$$
$$y_i=x_i^2.$$

The number of observations was set to $n=80$. Learning set size was chosen to be $g=10$ so that $n\geq 2g+2$. (Note, however, that the quantities $\tau_d^{(i)}$ only depend on P.) Coefficients were computed using the fast **R**-function `fastLmPure` from the **RcppArmadillo**-package. We chose the loss function

$$L(y_i,\hat{y}_i)=\arctan\{(y_i-\hat{y}_i)^2\}\cdot\frac{2}{\pi},$$

which is the squared error loss mapped to $[0;1)$, so that the convergence speed was controlled. Let β_i^S, $i=0,1$ be estimated coefficients, fitted on a learning set of size g with indices $S\subset 1,\dots,n,|S|=g$. Then, the kernel is given by

$$\Gamma(S;a)=L(y_a,\widehat{\beta}_0^S+\widehat{\beta}_1^S\cdot x_a). \tag{31}$$

In this set-up, it was possible to perform the estimations.

6.2. **Estimation of the regular parameter components of the variance.** It is convenient to estimate $\lambda_d^{(i)} := \tau_d^{(i)} + \Theta^2$ rather than $\tau_d^{(i)}$, so that Θ^2 needs to be estimated only once. Even in this simple example the computation of the estimators $\widehat{\lambda_d^{(i)}}$, $\widehat{\Theta}$ and $\widehat{\Theta^2}$ by a U-statistic was computationally too expensive. Therefore, we estimated them by drawing between $N = 10^5$ and $N = 10^7$ random subsets (of size $g+1$, $2g+2$ or $2g+2-d$, according to the degree of the corresponding kernel) from the data set instead of using all possible evaluation tuples. This procedure is theoretically justified by the inequality of Hoeffding [8, Theorem 2] which implies that the difference between the true estimator and the estimator obtained after N random iterations is greater than or equal to δ with a probability of at most $2\exp\left(-\delta N^2/2\right)$. As a rule of thumb, l digits are fixed after at most 10^{2l+1} iterations.

The parameter of interest $\widehat{\Theta}(\mathscr{T}_\star)$ took a value of 0.0746. The quantity $\widehat{\Theta^2}$ was estimated to be 0.0055.

After estimating Θ^2 and all $\lambda_d^{(i)}$ for $d = 1, \ldots, g+1, i = (1), \ldots, (4)$ by directly resampling the expectation value appearing in the right-hand side of Definition 2, $\widehat{\tau_d^{(i)}}$ was given by $\widehat{\lambda_d^{(i)}} - \widehat{\Theta^2}$. Table 1 in 6.5 shows the results and Figure 2 depicts them using scatter plots.

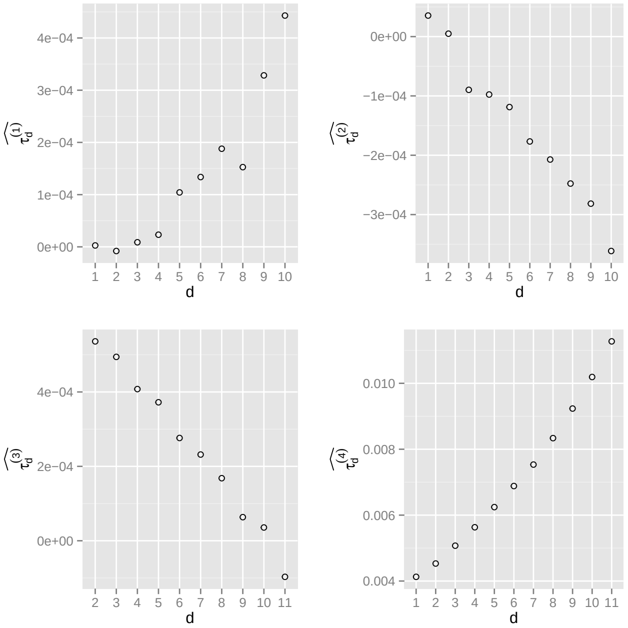
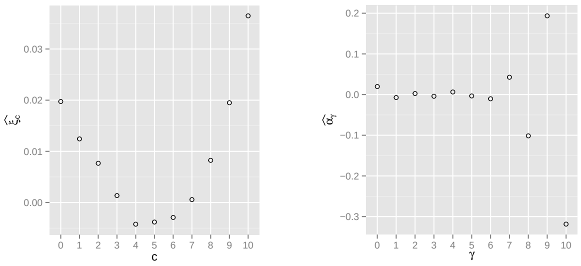


Figure 2. Scatter plots for the estimators $\widehat{\tau_d^{(i)}} \neq 0$. The plots show that the estimators $\widehat{\tau_d^{(1)}}$ and $\widehat{\tau_d^{(4)}}$ are positive and grow with d, whereas $\widehat{\tau_d^{(2)}}$ and $\widehat{\tau_d^{(3)}}$ decrease and can take negative values. Thus, the quantites $\tau_d^{(2)}, \tau_d^{(3)}$ cannot be variances. The fact that here $\widehat{\tau_2^{(1)}} < 0$ may be explained by the inaccuracy of the estimations. Therefore, almost all constraints on the $\tau_d^{(i)}$ described in Section 3.3 are satisfied by the estimators.



(a) The estimated quantities ξ_c

(b) The estimated quantities α_γ

Figure 3. Scatter plots for the estimators $\widehat{\xi_c}$ and $\widehat{\alpha_\gamma}$.

Thus, the computation suggests that the quantities $\tau_d^{(1)}$ divided by d indeed grow, in accordance with inequality (9). The remaining three estimated quantities decrease after dividing by d.

Having estimated the quantities $\tau_d^{(i)}$ – which do not depend on the sample size – on a large sample and thus with high precision, we will, from now on, suppose a hypothetical sample size of $n_{CV} = 13$. The computed estimators $\widehat{\xi_c}$ (see Table 2 of 6.5 and Figure 3a) of the quantities ξ_c – which *do* depend on the sample size – took both negative and positive values and did not decrease nor increase monotonically in d.

Figure 3a shows, in accordance with 6.4, that the ξ_c are roughly quadratic in c. Furthermore, the plot gives some indications on how to choose a variance-minimizing design . Since for $c = 4,5,6$, the $\widehat{\xi_c}$ are negative, the corresponding coefficients f_c^ℓ in the variance expression (14) have to be maximized while the remaining ones have to be minimized. Thus, a small-variance design should favor medium-sized learning overlaps over large and small sizes. Figure 4 illustrates (7): all $\widehat{\sigma_d^2}$ are positive and σ_d^2/d is monotonous. We estimated these quantities using the empirical analogue of (5).

The estimators $\widehat{\alpha_\gamma}$ took positive and negative values (see Table 3 in 6.5 and Figure 3b). Hence, it was not possible to find a design with minimum variance by simply minimizing or maximizing the terms B_γ. However, we considered the following designs and evaluated their variances empirically to obtain hints as to which CV-like procedures have small variance.

6.3. **Design with smaller variance than l-fold-K-fold CV.** We now selected several designs that typically have small variance. Since it is well known that K-fold cross-validation has high variance, one may consider cross-validating for fixed K several times by partitioning the sets differently (usually at random), in order to reduce the variance. We will refer to such a procedure as to *l-fold-K-fold CV*, where $K/l \in \mathbb{N}$. The number of iterations or learning sets of such a design is $K' = l \cdot K$.

We chose $n = 13$ so that a Balanced Incomplete Block Design (BIBD) exists. For the convenience of the reader, we repeat the definition of a BIBD adapted to our framework.

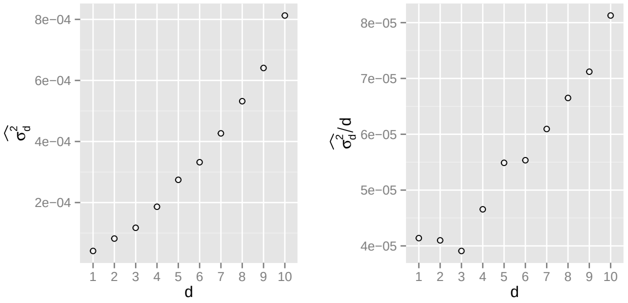


Figure 4. Scatter plots for the estimators $\widehat{\sigma_d^2}$ (left figure) and $\widehat{\sigma_d^2}$ divided by d (right figure). The plots show that σ_d^2 increases in both cases as it should. Thus, the estimators confirm the theoretical structure.

	$[\mathbb{V}(\widehat{\Theta}(\mathscr{T}))]\widehat{}$
MCCV	0.000798
5-F-6-F CV	0.000762
BIBD	0.000709
leave-p-out	0.000701

Table 1. Variance estimators $[\mathbb{V}(\widehat{\Theta}(\mathscr{T}))]\widehat{}$ for different designs in case of $g = 10$ and $n_{CV} = 13$

Definition 7. A Balanced Incomplete Block Design in the case of a learning set design $\mathscr{S}$ is a design in which each learning observation is contained in r learning sets and any pair of learning observations is contained in the same number λ of learning sets of size n_{CV}.

Thus, what is commonly called a (v,k,λ)-design corresponds, in our context, to a (n,g,λ)-design. More precisely, a (v,b,r,k,λ)-design corresponds to a (n,K,r,g,λ)-design.

Such a design exists for $n = 13$ and $K' = 26$, where $r = 20$ and $\lambda = 15$. This fact can be checked using the incidence matrix of this design (cf. attached R-code).

In order to compare this design appropriately to K-fold cross-validation, we chose 5-fold-6-fold CV for comparison, with $n_{CV} = 13$. Specifically, we used the design given by Table 4 in 6.5. Thus the number of learning sets (here $l \cdot K = 30$) even exceeded the one of the BIBD-design of 26. We considered the design given in Table 5 in 6.5.

For comparison we generated a further design for this set-up: we randomly chose $g = 10$ indices from $\{1,\dots,n_{CV}\}$ for every learning set. We also generated 30 learning sets. Such a CV-like procedure is commonly known as “Monte Carlo Cross-Validation” (MCCV).

We again computed the variance of the leave-p-outestimator for $n_{CV} = 13$. Table 1 presents the corresponding results.

The results show that the variance of the MCCV-estimator we estimated to be higher than the variance of l-fold-K cross-validation, as one would expect.

However, we have an interesting result for the BIBD variance estimator. It is clearly smaller than that for cross-validation and is close to the variance of the unique minimum-variance unbiased estimator, the leave-p-out.

Acknowledgements

We would like to thank Prof. Volker Betz for a very friendly and helpful discussion on the subject. Furthermore, we thank Prof. Anne-Laure Boulesteix for the support and the opportunity to carry out the present work in her group. Last, but not least we would like to thank Rory Wilson very much for a very careful, constructive and helpful proofreading. The first author was supported by the German Science Foundation (DFG-Einzelförderung BO3139/2-2 to Anne-Laure Boulesteix).

Supplements

6.4. **Supplement A.** Analytical computation of the quantities $\tau_d^{(i)}, \xi_c$, and α_c in an example related to linear regression.

6.5. **Supplement B.** More tables.

References

[1] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010. ISSN 1935-7516. URL http://dx.doi.org/10.1214/09-SS054.

[2] R. A. Bailey and Peter J. Cameron. Using graphs to find the best block designs. In *Topics in structural graph theory*, volume 147 of *Encyclopedia Math. Appl.*, pages 282–317. Cambridge Univ. Press, Cambridge, 2013.

[3] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.*, 5:1089–1105, 2003/04. ISSN 1532-4435.

[4] Harald Cramér. *Mathematical Methods of Statistics*. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946.

[5] Mathias Fuchs, Roman Hornung, Riccardo De Bin, and Anne-Laure Boulesteix. A u-statistic estimator for the variance of resampling-based error estimators. *LMU Department of Statistics: Technical Reports148*, 2013.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. ISBN 978-0-387-84857-0. URL http://dx.doi.org/10.1007/978-0-387-84858-7. Data mining, inference, and prediction.

[7] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, 19:293–325, 1948. ISSN 0003-4851.

[8] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963. ISSN 0162-1459.

[9] A. J. Lee. *U-statistics*, volume 110 of *Statistics: Textbooks and Monographs*. Marcel Dekker, Inc., New York, 1990. ISBN 0-8247-8253-4. Theory and practice.

[10] Yoshihiko Maesono. Asymptotic comparisons of several variance estimators and their effects for Studentizations. *Ann. Inst. Statist. Math.*, 50(3):451–470, 1998. ISSN 0020-3157. doi: 10.1023/A:1003521327411. URL http://dx.doi.org/10.1023/A:1003521327411.

[11] Boxin Tang. Balanced bootstrap in sample surveys and its relationship with balanced repeated replication. *J. Statist. Plann. Inference*, 81(1):121–127, 1999. ISSN 0378-3758. URL http://dx.doi.org/10.1016/S0378-3758(99)00013-0.

[12] Q. Wang and B. Lindsay. Variance estimation of a general *u*-statistic with application to cross-validation. *Statistica Sinica*, 24:1117–1141, 2014.

[13] Yuhong Yang. Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, 35(6):2450–2473, 2007. ISSN 0090-5364. URL http://dx.doi.org/10.1214/009053607000000514.

[14] Qiong Zhang and Peter Z. G. Qian. Designs for crossvalidating approximation models. *Biometrika*, 100(4):997–1004, 2013. ISSN 0006-3444. URL http://dx.doi.org/10.1093/biomet/ast034.

E-mail address: fuchs@ibe.med.uni-muenchen.de, krautenbacher@helmholtz-muenchen.de

INSTITUT FÜR MEDIZINISCHE INFORMATIONSVERARBEITUNG BIOMETRIE UND EPIDEMIOLOGIE, LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN,, MARCHIONINISTR. 15, 81377 MÜNCHEN, GERMANY