LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Mathias Fuchs, Norbert Krautenbacher

Minimization and estimation of the variance of prediction errors for cross-validation designs

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

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

MINIMIZATION AND ESTIMATION OF THE VARIANCE OF PREDICTION ERRORS FOR CROSS-VALIDATION DESIGNS

MATHIAS FUCHS

LMU Munich
Institut für Medizinische Informationsverarbeitung Biometrie und Epidemiologie
Marchioninistr. 15, 81377 München, Germany

NORBERT KRAUTENBACHER

TU Munich
Institute of Computational Biology, Helmholtz-Zentrum Munich
Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

ABSTRACT. We consider the mean prediction error of a classification or regression procedure as well as its cross-validation estimates, and investigate the variance of this estimate as a function of an arbitrary cross-validation design. We decompose this variance into a scalar product of coefficients and certain covariance expressions, such that the coefficients depend solely on the resampling design, and the covariances depend solely on the data's probability distribution. We rewrite this scalar product in such a form that the initially large number of summands can gradually be decreased down to three under the validity of a quadratic approximation to the core covariances. We show an analytical example in which this quadratic approximation holds true exactly. Moreover, in this example, we show that the leave-p-out estimator of the error depends on p only by means of a constant and can, therefore, be written in a much simpler form. Furthermore, there is an unbiased estimator of the variance of K-fold cross-validation, in contrast to a claim in the literature. As a consequence, we can show that Balanced Incomplete Block Designs have smaller variance than K-fold cross-validation. In a real data example from the UCI machine learning repository, this property can be confirmed. We finally show how to find Balanced Incomplete Block Designs in practice.

1. INTRODUCTION

This paper is concerned with the variance of resampling designs in classification and regression. Our setup is the classical statistical setup: we suppose having observed n independent observations from an unknown probability distribution P, where each observations is comprised of a number of predictors, and a response variable. The latter might be binary, categorical, or on a metric scale. Furthermore, we suppose we have a particular learning procedure at hand: any rule that maps a new set of predictors to a pertaining "conjectured"

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

2010 *Mathematics Subject Classification.* 62G05, 62G09, 62G10, 62G20 (primary), and 62J05, 62K05, 05B05(secondary).

Key words and phrases. U-statistic; cross-validation; design; model selection.

response, given a collection of learning observations. For instance, this rule could consist of multiplying the new data by a vector of estimated regression coefficients, or it might be a very elaborate machine learning algorithm — just any procedure capable of predicting a new response, given the learning observations and test predictors, in a deterministic way. In this paper, we will not distinguish between regression and machine learning because we will be concerned with their prediction aspects which can be formally treated in the same way, no matter whether the predicted response was generated by a machine learning algorithm or a regression. The "success" of the learning algorithm is then measured by the value of a loss function — which might measure the observed misclassification rate, for example. There are a multitude of loss functions in use. Then, we will refer to the (unknown) expected value of this loss function as to the "true error", where both the learning observations and the test observations are random.

We recommend the reader to take a look at Section 2 for a closer introduction of the concept of "resampling design" as it is investigated here, and for understanding that here the notion of design is different from the usual statistical meaning of "experimental design". Any resampling design, such as cross-validation, defines an estimator of the true error, and estimation of this variance has proven to be an important but difficult and essentially unresolved issue [Nadeau and Bengio, 2003, Bengio and Grandvalet, 2003/04]. 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 [2013], 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 [2013, after Formula 12] give a variance decomposition of the average prediction error estimator associated with several particular designs; we will give the corresponding formula for any design.

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).

Here, we point out the usefulness of statistical design theory to resampling, using this interpretation of a design. Although design theory has been used in resampling and variance estimation theory, previous papers seem to have focused on giving surveys [Tang, 1999], 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.

Moreover, Fuchs et al. [2013] 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. [2013] is a preprint, we give a synopsis of that paper in Section 2.5. 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 [2013] 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 an incomplete U-statistic is as small as possible and, therefore, as close as possible to that of the leave-p-out classifier [Lee, 1990, 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 symmetrization 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 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 [2013] which were defined without any mention of U-statistics.

The variance estimation of U-statistics has already been considered in the literature [Maesono, 1998, Wang and Lindsay, 2014].

The main contributions of this paper are the following. The first key contribuition is a detailed variance decomposition of the variance of the error estimator of a particular resampling design. This variance decomposition will be optimal in the sense that it separates the contribution to the variance due to the probability distribution P from the contribution to the variance due to the resampling design. More precisely, we write the variance of the error estimator as a linear combination, where the coefficients only depend on the resampling design, and the "basis vectors" of the linear combination are what we call core covariances — namely, quantities that depend only on the underlying probability distribution P. The second key contribution is that we show how to exploit this variance decomposition for variance minimization. For this, we rewrite the linear combination just mentioned in a way which might shortly be described as a transposition operation: After an important preparatory step, namely the variance decomposition n Formula (3.4), we write the variance as (a constant multiple of) a scalar product $< f^{\ell}, \xi >$ (see Formula (3.6)), which will be rewritten in the variance decomposition in Formula (3.11). In short, the steps in between can be seen as the linear-algebraic transposition operation $< f^{\ell}, \xi >=< PB, \xi >=< B, P^T \xi >=< B, \alpha >$. The benefit of this is that the linear combination becomes much shorter under quadratic approximation to the vector B: instead of up to $g+1$ non-zero coefficients, only a term $B_1\alpha_1 + B_2\alpha_2$ has to be minimized. We then show that these observations imply that Balanced Incomplete Block Designs (BIBDs) are good candidates for variance minimizing designs. We then demonstrate in full mathematical detail in an analytical example in chapter 4, that BIBDs *are* in fact exactly the variance minimizing designs. We then support this numerically in a real data example (the concrete slump data from the well-known UCI machine learning repository), that the BIBDs indeed have notably smaller variance than the ordinary repeated cross-validations procedures of the same computational complexity.

Thus, this paper can, in total, be read as a plea for using BIBDs instead of cross-validation. We support this plea by showing how to find BIBDs in practice, using mathematical software. 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, Section 5 illustrates our theory by a real data set, and Section 6 helps readers to find Balanced Incomplete Block Designs in practice.

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 Y' 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 loss $(y-y')^2$ whose expectation is the mean-square error, 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

$$\begin{aligned} s : (\mathscr{X} \times \mathscr{Y})^{\times g} \times \mathscr{X} &\to \mathscr{Y} \\ (x_1, y_1, \ldots, x_g, y_g, x_{g+1}) &\mapsto s(x_1, y_1, \ldots, x_g, y_g; x_{g+1}) \end{aligned} \tag{2.1}$$

which maps the learning sample $(x_1, y_1, \ldots, x_g, y_g)$ to the prediction rule applied to the test observation x_{g+1}, and where the semicolon visually separates the learning observations from the predictor of the test observation. 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, $s(x_1, y_1, \ldots, x_g, y_g; x_{g+1})$ is denoted by $\widehat{f}(x_{g+1} | x_1, y_1, \ldots, x_g, y_g)$ 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

$$\mathbb{E}(l(s)) = \int \cdots \int l(s(x_1, y_1, \ldots, x_g, y_g; x_{g+1}), y_{g+1}) \, dP(x_1, y_1), \ldots, dP(x_{g+1}, y_{g+1}) \tag{2.2}$$

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. 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.

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. Therefore, we treat a broad class of frequently considered machine learning/statistics set-ups. In particular, our random design set-up, where the X_i are treated as random, is of most relevance in applications because it allows for out-of-sample prediction.

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. [2009, 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

$$\begin{aligned}\Gamma(i_1,\ldots,i_g;i_{g+1}) := l_1(s_1(x_{i_1},y_{i_1},\ldots,x_{i_g},y_{i_g};x_{i_{g+1}}),y_{i_{g+1}}) \\ - l_2(s_2(x_{i_1},y_{i_1},\ldots,x_{i_g},y_{i_g};x_{i_{g+1}}),y_{i_{g+1}}),\end{aligned} \tag{2.3}$$

a function on a set of $g+1$ different indices $i_k \in 1,\ldots,n$, for two model fitting procedures s_1, s_2 and two appropriate loss functions l_1, l_2.

We have: $\mathbb{E}(\Gamma) = \mathbb{E}(l_1(s_1) - l_2(s_2))$ and $\Theta := \mathbb{E}\Gamma$ as a slight generalization of (2.2). 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 mean difference of losses. 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 $S \subset \{1,\ldots,n\}$ is an (unordered) set of disjoint learning indices, and $a \in \{1,\ldots,n\} \setminus S$ is a test index. Then, each $\Gamma(S;a)$ is an "elementary" estimator of Θ, and

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

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

2.3. **Resampling.** We define a resampling design to be an arbitrary collection $\mathscr{S}$ of different size-g-subsets of $\{1,\ldots,n\}$, where n is a sample size and $1 \le g < n$ is a learning sample size. For each such g-subset, its complement — a subset of size $n-g$ — is interpreted as its corresponding test set. For instance, two-fold cross-validation, viewed as a resampling design, is described by the collection

$$\mathscr{S} = \big\{\{1,\ldots,n/2\},\{n/2+1,\ldots,n\}\big\}$$

containing the indices of the observations in the two learning sets. The leave-one-out resampling design is described by

$$\mathscr{S} = \big\{\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\}\big\}$$

for $n=4$, and so on. However, we will also consider less well-known examples of resampling designs $\mathscr{S}$.

A "resampling procedure", in the sense of this paper, is an estimator of a classification algorithm's true error rate, or of the expected value of a loss function, in general. This estimator is a random variable whose expectation depends only on the data generating process (the data distribution), the learning algorithm/regression procedure, the loss function, and

the learning sample size. We are interested in the variance of this random variable, viewed as a function of $\mathscr{S}$, for a fixed data generating process (distribution), and a fixed learning/regression algorithm. We are not going to distinguish between learning and regression, and will always speak of the error, defined by an appropriate loss function.

The resampling procedure estimates the prediction error of some algorithm or regression procedure in the following way: for each g-subset in the design $\mathscr{S}$, a prediction rule is fitted onto the corresponding observations, and the performance of this prediction rule is evaluated using the observations corresponding to the test set. The average of all these prediction errors (or loss functions, in general), taken across all learning sets in the design, is 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 collection of all $\binom{n}{g}$ subsets, leading to the leave-p-out estimator of the average prediction error; a recent preprint [Fuchs et al., 2013] 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.

We employ the notion of statistical design of experiments in the following way. The n independent observations — each comprising predictors and a response — correspond to the "treatments", and the K learning sets correspond to the "blocks".

The condition that all the blocks have the same size corresponds to our assumption that all learning sets have the same size g. Thus, a resampling design is a particular matrix of size $n \times K$ with binary entries such that the column sums are equal to g. The design is called equireplicate if each observation is contained in the same number of learning sets. The usual triple (b,k,v) of design theory (as in, for instance, [Bailey and Cameron, 2013]) becomes (K,g,n) in our notation. Thus, our point of view where the independent observations correspond to the treatments, differs from the usual interpretation of a design, where the independent observations are the experimental units that make up the blocks.

The true mean loss of the prediction rule learned on samples of size g can be interpreted as a statistical parameter Θ, depending only on the underlying distribution, the learning algorithm, and the number g, but not the sample size n. A resampling design in our sense yields, by construction, an unbiased estimator of Θ, as we will see below. In contrast, the bootstrap estimator of Θ is not unbiased; likewise, the jackknife variance estimators are not unbiased. In this paper, we focus our attention on unbiased estimators. For this reason, we shall not investigate any bootstrap or jackknife procedure.

2.4. **Complete and incomplete U-statistics.** This section summarizes some definitions and ideas from [Hoeffding, 1948]. Let n denote the sample size. A U-statistic is a statistic of the form $U = \binom{n}{k}^{-1} \sum h(z_{i_1},\ldots,z_{i_k})$ for a symmetric function h of k vector arguments, where the summation extends over all possible subsets $(i_1,\ldots,i_k)$. 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 $1 \le S_1 < \dots < S_k \le n$.

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 $\mathscr{R}$ is a collection of ordered size g-subsets of $\{1,\dots,n\}$.

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 $h = \Gamma$. This allows to treat the issue of finding small-variance designs — a problem from machine learning and statistics — in the powerful framework of U-statistics.

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.

2.5. **A test for the comparison of two mean losses.** Here, we give a short, self-contained overview of the results of Fuchs et al. [2013]. 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 [Arlot and Celisse, 2010] 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 mean difference of losses. 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 3.1), 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 U, trivially given by $\mathbb{E}U^2 - \Theta^2$, 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 [Fuchs et al., 2013].

3. The core covariances and their theoretical properties

In the following, we will generalize the variance decomposition of Bengio and Grandvalet [2003/04, Corollary 2, 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 $\tau_d^{(i)}$.

The important results are formulae (3.4), (3.6), (3.11), called "core covariances". The coefficients of these linear combinations only depend on the design $\mathscr{S}$, and the parameters only depend on the data distribution, the algorithm and the loss function.

These regular parameters $\tau_d^i, \xi_c, \alpha_\gamma$ are all closely related to the quantities ζ_d of Hoeffding [1948], which we will call σ_d^2 — in Section (3.2) we will explain Hoeffding's quantities as linear combinations of the more "fine-grained" $\tau_d^i, \xi_c, \alpha_\gamma$. We then go on to show that these decompositions are useful for deriving results on variance-minimization in the small sample case, as well as asymptotic results. For instance, we derive the variance structure of cross-validation in Theorem 3.8.

Our first goal is the variance decomposition in formulas (3.6) and (3.11). 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 (2.3)), we recover part of the variance decomposition of incomplete U-statistics as in Lee [1990, Chapter 4].

However, it is quite important to note that our variance decomposition (3.11) is somewhat analogous to, but *does not* reduce to, the variance decomposition of Lee [1990, 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.

The variance decomposition (3.11) will turn out to be particularly useful for minimization: we will show that in the case where the quantities ξ_c are polynomial, the decomposition is particularly short and, thus, lends itself well for optimization. We show in Section 4 that this strategy works in the case of a simple regression toy model. In it, 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 [1948, 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.

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.

Definition 1. Let $S=\{1,\dots,g\}$, $a=g+1$, $S'=\{g+2,\dots,2g+1\}$, $a'=2g+2$. 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 (3.1)), $\Theta^2=\mathbb{E}(\Gamma_0(1,\dots,g+1)\Gamma_0(g+2,\dots,2g+2))$ 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.

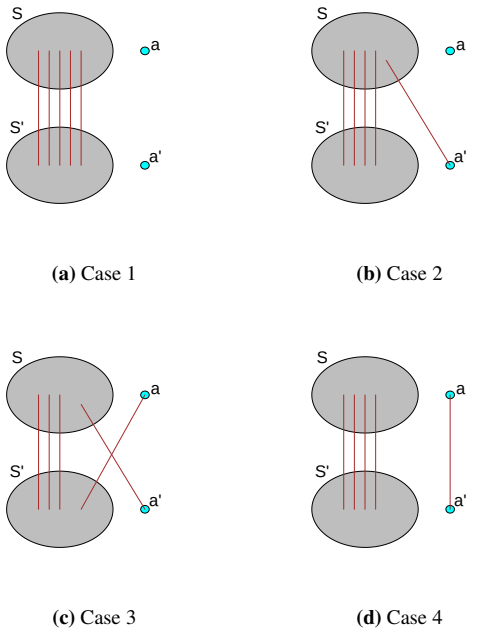
$$\tau_d^{(i)}:=-\Theta^2+\begin{cases}\mathbb{E}(\Gamma(1,\dots,g;g+1)\Gamma(1,\dots,d,g+2,\dots,2g+1-d;2g+2-d)), & i=1\\ \mathbb{E}(\Gamma(1,\dots,g;g+1)\Gamma(1,\dots,d-1,g+1,\dots,2g+1-d;2g+2-d)), & i=2\\ \mathbb{E}(\Gamma(1,\dots,g;g+1)\Gamma(1,\dots,d-2,g+1,\dots,2g+2-d;d-1)), & i=3\\ \mathbb{E}(\Gamma(1,\dots,g;g+1)\Gamma(1,\dots,d-1,g+2,\dots,2g+2-d;g+1)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

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

Remark 3. Therefore, the quantity σ^2 from Bengio and Grandvalet [2003/04] 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 between two evaluation tuples* $(S\cup\{a\})$ *and* $(S'\cup\{a'\})$ *by*

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



(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 $d = 5$

The overlap between two learning sets S and S' is denoted by

$$c := \left|S \cap S'\right|$$

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 $\tau_c^{(i)}$ 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 $i = 1,\dots,4$ that describes the overlap pattern. This is obvious from the fact that Γ is symmetric in the learning indices, and the product measure d^nP is permutation invariant.

3.2. σ_d^2 **as a linear combination of the core covariances.** Let us define

$$\sigma_d^2 := \mathbb{E}(\Gamma_0(1,\dots,g+1)\Gamma_0(g+2-d,\dots,2g+2-d)) - \Theta^2 \tag{3.1}$$

for $d = 1,\dots,g+1$. (σ_d^2 is called ζ_d in Hoeffding [1948].) 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:

(3.2)
$$\sigma_d^2 = \frac{1}{(g+1)^2}\Big((g+1-d)^2\cdot\tau_d^{(1)} + 2d(g+1-d)\cdot\tau_d^{(2)} + d(d-1)\cdot\tau_d^{(3)} + d\cdot\tau_d^{(4)}\Big).$$

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 $(g+1)^2\sigma_1^2/n$ [Hoeffding, 1948, 5.23] (recall that $p = n-g$). So, the limiting variance is

(3.3)
$$\lim_{n\to\infty} n\,\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{\star,n})) = g^2\tau_1^{(1)} + 2g\tau_1^{(2)} + \tau_1^{(4)},$$

where the limit is taken for g fixed.

3.3. **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

(3.4)
$$\begin{aligned}\mathbb{V}(\widehat{\Theta}(\mathscr{T})) &= |\mathscr{T}|^{-2}\sum_{k,l} Cov(\Gamma(S_k;a_k),\Gamma(S_l;a_l))\\ &= |\mathscr{T}|^{-2}\sum_{k,l}\tau(\Psi((S_k,a_k),(S_l,a_l)))\\ &= |\mathscr{T}|^{-2}\sum_{d,(i)}\left(\sum_{k,l} 1_{\Psi((S_k,a_k),(S_l,a_l))=(d,(i))}\right)\tau_d^{(i)}.\end{aligned}$$

The expression in brackets "counts" the number of occurences of the overlap pattern $d,(i)$ among all pairs of (S,a). These expressions are known to the researcher and need not be estimated. Hence, they can be viewed as scalar coefficient; by which the huge sum over all pairs (k,l) is abbreviated to a much shorter one over the numbers $(d,(i))$. Also, this sum can be read as a sort of scalar product between the vector of these numbers, indexed by the $(d,(i))$, and the vector $\tau_d^{(i)}$. This point of view will become very useful below.

3.4. **Variance decomposition of test-complete designs.**

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

(3.5)
$$\begin{aligned}\xi_c :=& (n-2g+c)(n-2g+c-1)\cdot\tau_c^{(1)} + 2(g-c)(n-2g+c)\cdot\tau_{c+1}^{(2)} +\\ &+ (g-c)^2\cdot\tau_{c+2}^{(3)} + (n-2g+c)\cdot\tau_{c+1}^{(4)}\end{aligned}$$

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

(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^TN, 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*

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T})) = |\mathscr{T}|^{-2}\sum_{c=0}^{g} f_c^\ell \xi_c \tag{3.6}$$

where ξ_c was defined in (3.5).

Proof. This follows from expanding the variance as in (3.4) 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*

$$f_c^\ell = \begin{cases} 0, & c \notin \{n-n/K, n-2n/K\} \\ K, & c = n-n/K \\ K^2-K, & c = n-2n/K. \end{cases} \tag{3.7}$$

The variance of cross-validation is given by the formula

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T})) = (K^{-1}-n^{-1})\tau^{(1)}_{n-n/K} + (1-K^{-1})\tau^{(3)}_{n-2n/K+2} + n^{-1}\tau^{(4)}_{n-n/K+1} \tag{3.8}$$

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

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T})) = \frac{1}{n}\big((n/2-1)\tau^{(1)}_{n/2} + n/2\cdot\tau^{(3)}_2 + \tau^{(4)}_{n/2+1}\big) \tag{3.9}$$

Since it is unclear whether and how fast the $\tau_d^{(i)}$ converge to zero, one can not immediately deduce asymptotic statements from (3.9).

3.5. **Non-asymptotic minimization of $f_c^\ell \xi_c$.**

Definition 4. Let $\mathscr{T}$ be a test-complete design. For $\gamma = 1,\ldots,g$ and a subset $s \subset \{1,\ldots,n\}$ 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 $B_\gamma^\ell := \sum_s n(s)^2$, where the sum is taken over all $\binom{n}{\gamma}$ subsets s. Analogously, let $B_0 := K^2 = |\mathscr{T}|^2 (n-g)^{-2} = \sum_{c=0}^g f_c^\ell$.

Lemma 1. *The quantities f_c^ℓ are uniquely determined by the B_γ^ℓ. In fact,*

$$f_c^\ell = \sum_{\gamma=c}^g (-1)^{\gamma-c} \binom{\gamma}{c} B_\gamma$$

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 [1990, Chapter 4, Equation (7)], even though our f_c^ℓ and B_γ are quite different from Lee's f_c and B_γ.

For $c = 0$, one has

$$f_0^\ell = \sum_{c=0}^g f_c^\ell - \sum_{c=1}^g f_c^\ell = B_0 - \sum_{c=1}^g \sum_{\gamma=c}^g B_\gamma = \cdots = B_0 + \sum_{\gamma=1}^g (-1)^\gamma B_\gamma = \sum_{\gamma=0}^g (-1)^\gamma B_\gamma,$$

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 $\gamma \ge 0$ (The map described by the matrix P is often called the binomial transform.) Using (3.6), 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.

$$\alpha_\gamma := \sum_{c=0}^\gamma (-1)^{\gamma-c} \binom{\gamma}{c} \xi_c \text{ for all } 0 \le \gamma \le g \tag{3.10}$$

Thus, we have shown that

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T})) = |\mathscr{T}|^{-2} \sum_{\gamma=0}^g B_\gamma \alpha_\gamma, \tag{3.11}$$

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 of [Lee, 1990, Chapter 4] where all B_γ have to be minimized.

The usefulness of (3.11) 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. 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 $\gamma \ge 2$. If ξ is of degree two, we have $\xi_c = b + Ac + Cc^2$, and then it is easy to calculate that

$$\begin{aligned} \alpha_0 &= b \\ \alpha_1 &= A + C \\ \alpha_2 &= 2C \\ \alpha_\gamma &= 0 \text{ for all } \gamma \ge 3. \end{aligned} \tag{3.12}$$

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

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{4.1}$$

where $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ with β_0 and σ^2 unknown. This model is a close relative of that used in univariate ordinary regression except that the slope coefficient β_1 is known and only the intercept is estimated.

We show the following facts in supplement A: 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 (3.11), 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 supplement A can be used to compare the variance of cross-validation with that of the leave-p-out estimator in closed form expressions. Since $\Theta = \mathbb{E}(\widehat{\Theta}(\mathscr{T}_\star)) = \sigma^2(1+1/g)$, one may suspect and check by direct computation that the leave-p-out estimator on a sample of size n has the short form*

$$\widehat{\Theta}(\mathscr{T}_\star) = s^2(1+g^{-1})$$

with the usual unbiased residual variance estimator s^2 for σ^2, defined as

$$s^2(x_1,y_1,\ldots,x_n,y_n) = (n-1)^{-1}\sum_{i=1}^{n}(z_i-\overline{z})^2,$$

where $z_i = y_i - \beta_1 x_i = \beta_0 + \varepsilon_i$. Since $s^2 \sim \sigma^2(n-1)^{-1}\chi^2_{n-1}$, the leave-p-out estimator is distributed as the $\sigma^2(1+g^{-1})(n-1)^{-1}$-fold of a chi-square of $n-1$ degrees of freedom.

Therefore, the variance of the leave-p-out estimator is $\mathbb{V}(\sigma^2(1+g^{-1})(n-1)^{-1}\chi^2_{n-1}) = \sigma^4(1+g^{-1})^2(n-1)^{-2}\cdot 2(n-1) = 2\sigma^4(1+g^{-1})^2(n-1)^{-1}$.

This is consistent with the fact that by (3.3) and (3.2), we have $\lim_{n\to\infty} n\,\mathbb{V}(\widehat{\Theta}(\mathscr{T}_\star)) = 2(g^{-2}+2g^{-1}+1)\sigma^4$ 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 (3.9):

$$\mathbb{V}(\widehat{\Theta}(\mathscr{T}_{CV})) = 2\sigma^4\left[n^{-1}+14n^{-2}\right]. \tag{4.2}$$

One can check that $\mathbb{V}(\widehat{\Theta}(\mathscr{T}_\star)) < \mathbb{V}(\widehat{\Theta}(\mathscr{T}_{CV}))$ for all n, as it should.
Denote by s_2^2 the unbiased estimator s^2 calculated on a sample of size $n/2$. Then,

$$(z_1,\ldots,z_n) \mapsto 2\left[n^{-1}+14n^{-2}\right]\cdot s_2^2(z_1,\ldots,z_{n/2})s_2^2(z_{n/2+1},\ldots,z_n)$$

is an unbiased estimator of (4.2), which exists as soon as $n \geq 4$, in contrast to Bengio and Grandvalet [2003/04]. Note that this unbiased variance estimator is targeted at the particular situation, and is unrelated to the general variance estimator mentioned in 2.5 which only exists if $n \geq 2g+2$ and therefore excludes two-fold cross-validation where $n = 2g$.

Note 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 (3.11) 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 [1990, 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 (4.1), 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 6), 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 [1990, 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. Real data example: the concrete slump dataset

5.1. **Preparation: Explanation of the reported table.** In the preceding chapters, we have collected theoretical evidence that balanced incomplete block designs provide a better approximation to the optimal, namely the leave-p-out estimator of the error rate, than ordinary cross-validation.

In order to properly explain how we go about estimating these variances in the results table 5.3, we need some preparation.

We have the following goals:

(1) First, we will show how to find a balanced incomplete block design suitable for a given real dataset.

(2) Second, we want to illustrate the main point of the present paper: namely, to show that the BIBD does indeed have smaller variance than ordinary cross-validation with comparable values of the resampling design size K.

Let us explain the latter point. We want to use the dataset of the limited size n itself, and we want to take an intuitive approach to doing so, instead of having to estimate any of the quantities $\tau_d^{(i)}$, etc.

Let us assume we have fixed a learning set size $g < n$.

We know theoretically that the best (i.e., minimal variance) estimator of Θ is the leave-$(n-g)$-out estimator for a given g. However, the leave-$(n-g)$-out is computationally inaccessible.

Both cross-validation and BIBD have higher variance than the leave-$(n-g)$-out. In order to demonstrate that a BIBD is superior to ordinary cross-validation, we would like to estimate the variance of cross-validation, the variance of the BIBD, and the variance of the leave-$(n-g)$-out estimator, to show how one is successively smaller than the other, in that order.

Variance estimation is possible by estimating all the quantities $\tau_d^{(i)}$, because the variance of any design is a linear combination of these $\tau_d^{(i)}$, as we have abundantly discussed in the preceding chapters, and one can use a simple plug-in estimator.

However, we have seen that unbiased estimation of $\tau_d^{(i)}$ requires unbiased estimation of Θ^2 which is impossible unless $n \geq 2g+2$. This excludes ordinary cross-validation in which $g \geq n/2$.

Fortunately, there is another approach suitable for demonstration of the fact that the BIBD has smaller variance than cross-validation. In fact, let us consider the set S_n of all permutations of the sample n (S_n is commonly called the symmetric group on n elements).

Let us consider the probability distribution generated by drawing the data, as usual, from the underlying probabilty distribution P, followed by randomly permuting the data.

Let us abbreviate the dataset before permutation, consisting of the matrix X of predictors and the response vector Y by (X,Y). Then, by the law of total variance, applied to conditioning on the observed data before the permutation procedure, we have

$$\begin{aligned} \mathbb{V}\big(\widehat{\Theta}(\mathscr{S})\big) &= \mathbb{E}\big(\mathbb{V}\big(\widehat{\Theta}(\mathscr{S})|(X,Y)\big)\big) + \mathbb{V}\big(\mathbb{E}\big(\widehat{\Theta}(\mathscr{S})|(X,Y)\big)\big) \\ &= \mathbb{E}\big(\mathbb{V}\big(\widehat{\Theta}(\mathscr{S})|(X,Y)\big)\big) + \mathbb{V}(\star) \end{aligned} \tag{5.1}$$

because the expectation $\mathbb{E}(\widehat{\Theta}(\mathscr{S})|(X,Y))$ of $U(\mathscr{S})$ under permutation of the data, given the data, can easily seen to be exactly equal to the leave-$(n-g)$-out $\star$.

Furthermore, the left-hand side of the equation, $\mathbb{V}\big(U(\mathscr{S})\big)$ is equal to the variance we are interested in, investigated in the rest of this paper, because the distribution of the permuted data is equal to the distribution of the data itself, so, in particular, the variances are the same.

Now, let $\mathscr{S}_1$ and $\mathscr{S}_2$ be two different designs which we want to compare. Then, plugging these into (5.1) in place of $\mathscr{S}$ results in two expressions for $\mathbb{V}\big(\widehat{\Theta}(\mathscr{S}_1)\big)$ and $\mathbb{V}\big(\widehat{\Theta}(\mathscr{S}_2)\big)$. Subtracting these two expressions from each other, the term $\mathbb{V}(\star)$ cancels out because it does not depend on $\mathscr{S}$ anymore, so we obtain

$$\mathbb{V}\big(U(\mathscr{S}_1)\big) - \mathbb{V}\big(U(\mathscr{S}_2)\big) = \mathbb{E}\big(\mathbb{V}\big(U(\mathscr{S}_1)|(X,Y)\big)\big) - \mathbb{E}\big(\mathbb{V}\big(U(\mathscr{S}_2)|(X,Y)\big)\big) \tag{5.2}$$

Thus, we arrive at a very important observation, which will help us to report the results on a real dataset: Even though it became clear in the preceding chapters that the variance of a single design can not be estimated without bias unless $n \geq 2g+2$, the difference between

two such variances can very easily be estimated without bias, namely by permuting the data and applying the usual naive variance estimator across the permutations.

In a nutshell: Even though one can not estimate the variance of a design globally, i.e. by itself (unless g is small: $g \leq (n-2)/2$, as we have discussed in the preceding chapters), we can actually estimate the difference of two such variances, belonging to two different designs with arbitrary but the same g.

As another byproduct of this discussion, we retain a simple fact: The naive variance estimator applied to some random permutations of the data does not estimate the variance of the error itself, but estimates the variance of the error *minus the variance of the corresponding leave-$(n-g)$-out estimator*.

Therefore, we can illustrate the main message of this paper by reporting the permuted variance estimator, and showing that BIBD has a smaller permutation variance than cross-validation. It then follows from this subsection that the total variance of the BIBD is estimated to be smaller than that of cross-validation, which is what we show in this paper.

5.2. **Description of the dataset.** A canonical source of datasets for demonstration of practical applicability is the UCI Machine Learning Repository. Since this paper is concerned with any regression/prediction method, and applies to, but is not peculiar to high-dimensional problems, there was no need to choose a high-dimensional dataset ($p \gg n$). Instead, we looked for a medium-dimensional dataset, where $p < n$, but not $p \ll n$.

The concrete slump dataset by I-Cheng [2007] consists of $n = 103$ observations of concrete specimens. On each of them, the content of seven components were measured in kilograms per cubic meter: slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate. The response variable is slump, a measure of the concrete's fluidity and workability. Usually, slump is to be optimized by balancing the seven ingredients under certain constraints. This explains the practical interest in predicting slump from the seven covariables.

5.3. **Numerical results.** This dataset lended itself for analysis because it was amenable to an ordinary least squares fitting procedure, using the squared difference between measured and observed slump as outcome variable. A preliminary analysis showed that the residuals were sufficiently normally distributed.

Since we rely on speed and computational efficiency, we applied a principal component analysis dimension reduction to the dataset retaining the first three principal components, reducing the dimensionality from seven to $p = 3$. The reason was the following: In order to be able to report the leave-$(n-g)$-out estimator which is computationally very costly, we implemented a very fast version of the OLS fitting procedure in the C programming language, benefitting from matrix inversion for symmetric three-by-three matrices by explicit formulae.

We investigated three values of the learning set size: $g = 50, 76, 96$, because for $g = 76$ and $g = 96$ a Balanced Incomplete Block Design was readily available, albeit for a slightly differing value of n, namely for $n = 101$. We are going to explain how we adapted the design to the larger sample size. Finally, the learning set size $g = 50$ was chosen so that the condition $n \geq 2g + 2$ was satisfied. Thus, there was an unbiased variance estimator for $g = 50$.

In each of the steps described in the following, the slight mismatch of sample size (either between 103 and 101 for the BIBD, and between 103 and 100 for two-fold cross-validation), was accounted for by randomly omitting two or three observations in each fold, as usual.

Table 1. Estimation of several quantities discussed in this paper on the concrete slump dataset. The first column is the learning set size g. The values in the last column being positive confirm the main finding of this paper: All values are positive, which corresponds to the BIBD having smaller variance than ordinary cross-validation.
Some values are not indicated, for the following reasons: For $g = 50$, there was no BIBD available, so that the cells that required a BIBD had to remain empty. For $g = 76$ and $g = 96$, the condition $n \geq 2g+2$ was not satisfied, so that there was no unbiased variance estimator for $\mathbb{V}(\star)$, and the cells requiring that variance had to remain empty.

g	K_{BIBD}	K_{CV}	$\star$	$\mathbb{V}(\star)$	$\mathbb{V}(\text{BIBD}\|(X,Y))$	$\mathbb{V}(\text{CV})$	$\mathbb{V}(\text{CV}\|(X,Y))$	$\mathbb{V}(\text{CV})$ - $\mathbb{V}(\text{BIBD})$ $= \mathbb{V}(\text{CV}\|(X,Y))$ $-\mathbb{V}(\text{BIBD}\|(X,Y))$
50	-	$1 \cdot 2 = 2$	36.09	19.5	-	21.77	2.27	-
76	101	$26 \cdot 4 = 104$	35.4	-	0.35	-	0.52	0.17
96	505	$25 \cdot 20 = 500$	35.01	-	0.33	-	0.48	0.15

The first row of Table 5.3 displays the results for $g = 50$. The number of permutations for estimation of the conditional variances, given the data, was 5000. The first column is the leave-$(n-g)$-out estimator. Since for $g = 50$ there exists no Balanced Incomplete Block Design, columns involving BIBDs had to remain empty in the first row. However, it was possible to perform two-fold cross-validation, and estimating its permutation variance, as explained in the preceding subsection. Finally, the estimator for the total variance of cross-validation, was obtained by applying Formula (5.1), and presented in the sixth column.

Next, we investigated the sample size $g = 76$. The BIBD with parameters $n = 101, g = 76, K = 101$ was found by taking the complement of the symmetrical BIBD with $n = 101, g = 25, K = 101$, as explained in Section 6. Here, it was possible to estimate the permutation variances, and plug them into Formula (5.2). The results confirms that the BIBD in fact has lower variance than two-fold cross-validation.

Finally, for $g = 96$ a BIBD was obtained by executing the command
`sage.combinat.designs.bibd.balanced_incomplete_block_design(101, 5).blocks()`
in the SAGE numerical software (version 6.9). This yielded a BIBD with parameters $(n, g, \lambda) = (101, 5, 1)$. The final BIBD was obtained by taking, again, the complement, which had, therefore, parameters $(101, 96, 456)$ (see chapter 6).

The design has a size of $K = 505$. We compared with 25-fold repetition of 20-fold cross-validation. Thus, the cross-validation had size $K = 25 \cdot 20 = 500$ which is sufficiently close.

We also report the leave-$(n-g)$-out estimator, the value of the permuation variance of the BIBD and of cross-validation. The difference, by Formula (5.2) is again the difference of the total variances. Again, the value was positive, confirming again the main finding of this paper, namely that the BIBD has smaller variance than cross-validation.

6. Finding Balanced Incomplete Block Designs for cross-validation in practice

In the preceding chapters, we have collected both theoretical and practical evidence that BIBD designs have small variance. Here, we want to show approaches to finding BIBDs in practice.

For the convenience of the reader interested in the practical aspects of cross-validation, let us re-state a definition of a cross-validation design:

Definition 6.
- A *design* for the sample size n is a collection (or multiset) $\mathscr{S}$ of K subsets of the size $1 \leq g < n$ of the set $\{1,\ldots,n\}$ of observation indices (or observations, for short). Each subset is called a *block* or *learning set*.
- A block design $\mathscr{S}$ is called Balanced Incomplete Block Design when the following two conditions are satisfied:
  (1) Each observation is contained in the same number r (where $1 \leq r < K$) of learning sets.
  (2) Each pair of distinct observations is contained in the same number λ of learning sets.

The second property is the "balancedness". The design is called "incomplete" because $g < n$.

In general, there is no algorithm to decide for a given constellation (n,g,K) whether there exists a BIBD or not. However, in practice one is given the sample size n and would typically like to quickly find a reasonable learning set size g and a design size K together with an explicitly given BIBD for the constellation (n,K,r,g,λ). In this section, we will show that this is in general a feasible task.

For a given design $\mathscr{R}$, we define the *complementary* design as the design on the same sample size n where each block is replaced by the complement in the set $\{1,\ldots,n\}$. Thus, the complement of an (n,g,K)-design is an $(n,n-g,K)$-design. We denote the complement of the design $\mathscr{R}$ by $\mathscr{R}^c$. One can show [Stinson, 2004, 1.32] that the complement of a BIBD is another BIBD. More precisely, the complement of a (n,K,r,g,λ)-design is a design with parameters

$$(n,K,K-r,n-g,K-2r+\lambda). \tag{6.1}$$

Loosely speaking, the process of passing to the complementary design is the reversal of the learning and testing roles of the observations.

A Steiner triple system is a BIBD for a constellation (n,K,r,g,λ) with $g=3$ and $\lambda=1$. Thus, each block has size three, and is called a *Steiner triple*, or triple for short. The following classical fact guarantees the existence of Steiner triple systems.

Theorem 3. *A Steiner triple system exists if and only if the rest of the division of n by six is one or three. The number K of triples is $n(n-1)/6$. The number r of occurences of a given observation is necessarily $(n-1)/2$.*

Thus, $n \in 7,9,13,15,19,21$. By passing to the complementary design and using the fact that the complement of a BIBD is another BIBD, we easily arrive at the conclusion:

Corollary 1. *A BIBD for the constellation (n,K,g) where n and $g=n-3$ are fixed and K may be chosen exists if the rest of the division of n by six is one or three, namely the complement of a Steiner triple system. Since the Steiner triple has $n(n-1)/6$ blocks and the number of blocks does not change when passing to a complement, we have $K=n(n-1)/6$ as well.*

This provides a convenient way for obtaining BIBDs in practice: By omitting observations, we can arrange for the condition on n, and the learning set size $g=n-3$ lends itself for practical implementation.

More generally, there is a general construction due to Stinson [2004] for designs with $g=4$ or $g=5$, and $\lambda=1$. Taking the complement, we arrive at a general construction for test set sizes $n-g=4$ or $n-g=5$. A BIBD for $g=4$ exists when the rest of the division of n by 12 is one or four, and a BIBD with $g=5$ exists when the rest of the division by 20 is one or five. For instance, for the typical sample size range $n \in \{100,\ldots,200\}$, one can

easily check that there are 46 values that satisfy one out of these criteria. Thus, in these cases one can directly carry out a BIBD-design analysis. This may be compared with the classical cross-validation, which is possible only if n is divisible by K. Thus, a BIBD has a higher chance of being possible. By taking the complementary design, SAGE can be used to construct a balanced incomplete block design whenever $n-g \leq 5$, i.e. for test set sizes up to five, whenever it exists. In general, we tested for every $n \in \{100,\ldots,200\}$ and every $n-g \in \{2,3,4,5\}$ and every λ whether a BIBD exists. The interested reader may find details for the case $k=6$ in Wallis [1996].

SAGE can construct a BIBD in all these cases where $n-g \leq 5$, and often when $n-g>5$, see Figure 2. This graph plots, for each sample size n on the horizontal axis, a test set size $n-g$ on the vertical axis with the property that a Balanced Incomplete Block Design can be found with the command
`sage.combinat.designs.bibd.balanced_incomplete_block_design(n, n-g).blocks()`
in the numerical software SAGE, v.6.9. Since one has to pass to the complementary design, the output of the command specifies the indices of the test observations rather than of those of the learning observations. For instance, there readily exists a BIBD with $n=101$ and $n-g=5$, and so on. Before taking the complement, these are BIBDs whose parameter λ is equal to one. After taking the complement, λ is given by Formula (6.1). For this reason, the plot is non-exhaustive: there will be more BIBDs for other values of λ. The ones with $\lambda=1$ are particularly easy to find, and sufficiently abundant, as the shows.

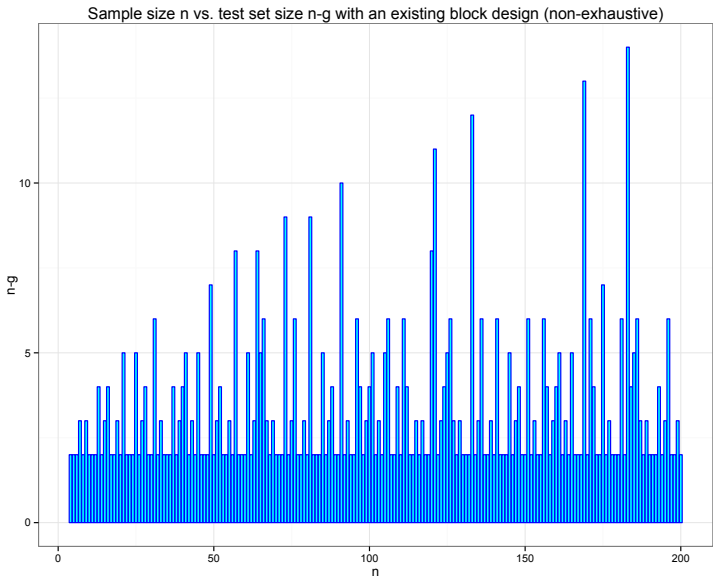


Figure 2. Plot of sample size n vs. test set size $n-g$ for which a BIBD exists. Those designs are depicted which arise as complementaries of designs with $\lambda=1$ and can easily be found using the command `sage.combinat.designs.bibd.balanced_incomplete_block_design(n, n-g).blocks()`.

SUPPLEMENT: AN EXEMPLARY COMPUTATION OF THE QUANTITIES $\tau_d^{(i)}, \xi_c$, AND α_c

Here, we consider the case where the distribution P is specified by: 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 $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ with β_0 and σ^2 unknown. The goal is to compute the quantities $\tau_d^{(i)}, \xi_c$, and α_c analytically. Since $\beta_0 = \mathbb{E}(Y - \beta_1 X)$, one has $\widehat{\beta_0} = g^{-1}\sum_{i=1}^{g}(Y_i - \beta_1 X_i)$ and Γ takes the form

$$\begin{aligned}\Gamma(1,\ldots,g;g+1) &= G((X_1,Y_1),\ldots,(X_g,Y_g);(X_{g+1},Y_{g+1})) \\ &= \big(g^{-1}\sum_{i=1}^{g}(Y_i - \beta_1 X_i) + \beta_1 X_{g+1} - Y_{g+1}\big)^2,\end{aligned}$$

using the mean squared error as the loss function. By a slight abuse of notation, let us write $Z_i := Y_i - \beta_1 X_i = \beta_0 + \varepsilon_i$. (Correctly, one would have to use yet another notation, say W_i instead of Z_i; however, one would then obtain

$$G(Z_1,\ldots,Z_g;Z_{g+1}) = G(W_1,\ldots,W_g;W_{g+1})$$

as equality of random variables on the entire probability space which is why we use the notation Z_i in the first place.) Then, Z_i is *i.i.d.* from $Z \sim \mathcal{N}(\beta_0,\sigma^2)$ and Γ can be written in terms of these variables as

$$\Gamma(1,\ldots,g;g+1) = G(Z_1,\ldots,Z_g;Z_{g+1}) = (g^{-1}\sum_{i=1}^{g} Z_i - Z_{g+1})^2 = (g^{-1}\sum_{i=1}^{g}\varepsilon_i - \varepsilon_{g+1})^2.$$

Therefore, Γ is $\sigma^2(1/g+1)$ times a chi-square variable with one degree of freedom. Moreover, $\Theta = \mathbb{E}\Gamma = \mathbb{V}(g^{-1}\sum_{i=1}^{g}\varepsilon_i - \varepsilon_{g+1}) = \sigma^2(1+g^{-1})$. This formula is similar to Zhang and Qian [2013, (9), (10)].

Recall that the covariance between two chi-square random variables can be computed as follows. Let (P,Q) be a bivariate normal distribution with covariance matrix $\left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ and mean $(0,0)^T$. Then, $Cov(P^2,Q^2) = 2b^2$. Hence, all $\tau_d^{(i)}$ are non-negative in this case.

Some care has to be taken: the degree of Θ is two rather than $g+1$; thus, Assumption 1 is not valid in this case. However, in this chapter we will only make use of the non-degeneracy of the associated U-statistic which is a slightly weaker statement than the assumption; non-degeneracy still remains valid. On a related note, let s^2 denote the usual unbiased variance estimator for σ^2, which is a U-statistic of degree two. Then one can check that the symmetrized form Γ_0 of Γ coincides with $s^2(1+g^{-1})$, which also follows from the uniqueness of the U-statistic for a regular parameter.

Another possibility to resolve the issue would be to add a negligibly small term of degree $g+1$ to Γ; in other words, the collection of choices of Γ such that the assumption is violated is a null set in some sense.

Let us abbreviate $A = \sum_{i=1}^{d}\varepsilon_i, C = \sum_{i=d+1}^{g}\varepsilon_i, D = \sum_{i=g+2}^{2g-d+1}\varepsilon_i$. Then, $A \sim (d\sigma^2)^{1/2}\mathcal{N}(0,1)$, $C \sim ((g-d)\sigma^2)^{1/2}\mathcal{N}(0,1), D \sim ((g-d)\sigma^2)^{1/2}\mathcal{N}(0,1)$. Furthermore, $\mathbb{E}A^4 = 3(d\sigma^2)^2$ due to the normal kurtosis, $\mathbb{E}A^3 = \mathbb{E}A = \mathbb{E}C = \mathbb{E}D = 0, \mathbb{E}A^2 = d\sigma^2, \mathbb{E}C^2 = \mathbb{E}D^2 = (g-d)\sigma^2$.

Note that for type one, the overlap is only between the two learning sets, thus $d = c$, and we only use the letter d. Making use of the mutual independences between A, C, D, ε_{g+1},

ε_{2g+2-d}, we obtain:

$$\begin{aligned}\tau_c^{(1)} &= Cov((g^{-1}(A+C)-\varepsilon_{g+1})^2,(g^{-1}(A+D)-\varepsilon_{2g+2-c})^2) = \\ &= 2\big(Cov(g^{-1}(A+C)-\varepsilon_{g+1},g^{-1}(A+D)-\varepsilon_{2g+2-c})\big)^2 = \\ &= c^2[2g^{-4}\sigma^4]\end{aligned}$$

This is remarkable because there seem to be few places in the literature where the quantities σ_d of a U-statistic are explicitly calculated. In particular, no variance formulae for the leave-p-out error of linear regression are known, except in the "leave-one-out"-case.
For type two, we have $d = c+1$ and it is convenient to choose the following abbreviations: $A = \sum_{i=1}^{c}\varepsilon_i$, $C = \sum_{i=c+1}^{g}\varepsilon_i$, and $D = \sum_{i=g+2}^{2g-c}\varepsilon_i$. Note that the symmetry between C and D is lost and we have $\mathbb{E}C^2 = (g-c)\sigma^2$ and $\mathbb{E}D^2 = (g-c-1)\sigma^2$. We prefer to perform the index shift $c+1$ on the left hand-side of the equation in order to stress the analogy of the computation with type one above. We then have

$$\begin{aligned}\tau_{c+1}^{(2)} &= 2Cov\big[(g^{-1}(A+C)-\varepsilon_{g+1}),(g^{-1}(A+\varepsilon_{g+1}+D)-\varepsilon_{2g-c+1})\big]^2 \\ &= c^2[2g^{-4}\sigma^4]+c[-4g^{-3}\sigma^4]+2g^{-2}\sigma^4.\end{aligned} \tag{6.2}$$

For type three, we have $d = c+2$ and it is convenient to choose the following abbreviations: A and D as above, but $C = \sum_{c+2}^{g}\varepsilon_i$. We then have

$$\begin{aligned}\tau_{c+2}^{(3)} &= 2Cov\big[(g^{-1}(A+\varepsilon_{c+1}+C)-\varepsilon_{g+1}),(g^{-1}(A+\varepsilon_{g+1}+D)-\varepsilon_{c+1})\big]^2 \\ &= c^2[2g^{-4}\sigma^4]+c[-8g^{-3}\sigma^4]+8g^{-2}\sigma^4.\end{aligned} \tag{6.3}$$

For type four, we abbreviate $A = \sum_{i=1}^{c}\varepsilon_i, C = \sum_{i=c+1}^{g}\varepsilon_i$, and $D = \sum_{i=g+2}^{2g-c+1}\varepsilon_i$. Using that $\mathbb{E}\varepsilon_{g+1}^3 = 0$ because the third central moment of a normal random variate vanishes, we obtain:

$$\begin{aligned}\tau_{c+1}^{(4)} &= 2Cov\big[(g^{-1}(A+C)-\varepsilon_{g+1}),(g^{-1}(A+D)-\varepsilon_{g+1})\big]^2 \\ &= c^2[2g^{-4}\sigma^4]+c[4g^{-2}\sigma^4]+2\sigma^4.\end{aligned}$$

By (3.5), the expressions for the quantities τ as functions of c yield for ξ_c:

$$\xi_c = 2\sigma^4\left[c-2g+n+\frac{c^2}{g^2}-\frac{2c}{g}+\frac{2c^2n}{g^3}-\frac{4cn}{g^2}+\frac{2n}{g}+\frac{c^2n^2}{g^4}\right].$$

By (3.12), we have

$$\begin{aligned}\alpha_0 &= 2\sigma^4\big[-2g+n+2ng^{-1}\big] \\ \alpha_1 &= 2\sigma^4\big[-\frac{2}{g}-\frac{4n}{g^2}+\frac{1}{g^2}+\frac{2n}{g^3}+\frac{n^2}{g^4}+1\big] \\ \alpha_2 &= 4\sigma^4\big[g^{-2}+2ng^{-3}+n^2g^{-4}\big] \\ \alpha_\gamma &= 0, \quad \gamma \geq 3.\end{aligned}$$

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.

REFERENCES

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

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.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research (JMLR)*, 5:1089–1105, 2003/04. ISSN 1532-4435.

Mathias Fuchs, Roman Hornung, Riccardo De Bin, and Anne-Laure Boulesteix. A u-statistic estimator for the variance of resampling-based error estimators. Technical report, Ludwig Maximilian University of Munich, 2013. URL http://epub.ub.uni-muenchen.de/17654/.

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.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325, 1948. ISSN 0003-4851.

Y. I-Cheng. Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, 29(6):474–480, 2007.

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.

Yoshihiko Maesono. Asymptotic comparisons of several variance estimators and their effects for Studentizations. *Annals of the Institute of Statistical Mathematics*, 50(3): 451–470, 1998. ISSN 0020-3157. doi: 10.1023/A:1003521327411. URL http://dx.doi.org/10.1023/A:1003521327411.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Mach. Learn.*, 52(3):239–281, September 2003. ISSN 0885-6125. doi: 10.1023/A:1024068626366. URL http://dx.doi.org/10.1023/A:1024068626366.

Douglas R. Stinson. *Combinatorial designs*. Springer-Verlag, New York, 2004. ISBN 0-387-95487-2. Constructions and analysis, With a foreword by Charles J. Colbourn.

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

W. D. Wallis, editor. *Computational and constructive design theory*, volume 368 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996. ISBN 0-7923-4015-9. doi: 10.1007/978-1-4757-2497-4. URL http://dx.doi.org/10.1007/978-1-4757-2497-4.

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

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.