On Variability due to Local Minima and K-Fold Cross Validation

CAREN MARZBAN, a,b JUEYI LIU, c AND PHILIPPE TISSOT d

a Applied Physics Laboratory, University of Washington, Seattle, Washington
b Department of Statistics, University of Washington, Seattle, Washington
c Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts
d Conrad Blucher Institute, Texas A&M University-Corpus Christi, Corpus Christi, Texas

(Manuscript received 28 November 2021, in final form 29 August 2022)

ABSTRACT: Resampling methods such as cross validation or bootstrap are often employed to estimate the uncertainty in a loss function due to sampling variability, usually for the purpose of model selection. In models that require nonlinear optimization, however, the existence of local minima in the loss function landscape introduces an additional source of variability that is conflated with sampling variability. In other words, some portion of the variability in the loss function across different resamples is due to local minima. Given that statistically sound model selection is based on an examination of variance, it is important to disentangle these two sources of variability. To that end, a methodology is developed for estimating each, specifically in the context of K-fold cross validation, and neural networks (NN) whose training leads to different local minima. Random effects models are used to estimate the two variance components—that due to sampling and that due to local minima. The results are examined as a function of the number of hidden nodes, and the variance of the initial weights, with the latter controlling the “depth” of local minima. The main goal of the methodology is to increase statistical power in model selection and/or model comparison. Using both simulated and realistic data, it is shown that the two sources of variability can be comparable, casting doubt on model selection methods that ignore the variability due to local minima. Furthermore, the methodology is sufficiently flexible so as to allow assessment of the effect of other/any NN parameters on variability.

KEYWORDS: Neural networks; Statistical techniques; Statistics; Machine learning; Regression

1. Introduction

Analysis of variance (ANOVA) is important for a wide range of problems. For example, in deciding which of two models is better in terms of some measure of performance, the variability of that measure is a quantity that plays a prominent role at all levels of analysis, from a simple comparison of boxplots, to performing a two-sample test of statistical significance. In general, larger variability is associated with reduced statistical power. In other words, even if there is a difference between the performance measures for the two models, larger variability renders it more difficult to detect. Indeed, for sufficiently large variability, the comparison of models becomes inconclusive, at best. Therefore, it is important to account for as many sources of variability as possible for the purpose of effectively reducing the variability. Although there exist other arenas in which an accurate assessment of variability is important, here the focus will be on model selection and/or model comparison.

Although the sources of variability are numerous, two that often attract attention are sampling variability and variability due to measurement error (Montgomery 2009; Fuller 1987; Buonaccorsi 2010). However, in problems whose models involve nonlinear optimization, there exist other sources of variability, hereinafter referred to as computational variability. One example is variability due to the existence of local minima. Although its existence is well known, its consequences are often not taken into account, at least not systematically. This source of variability is often a consequence of the sensitivity of the optimization routine with respect to the initial values of the parameters being optimized. In addition, there are indications that underspecification may also contribute to the existence of local minima (D’Amour et al. 2020). In neural networks (NN), for example, the parameters being optimized are the weights, and their estimation involves initializing them randomly. As a result, repeated training can lead to different local minima of the loss function.

Cross validation is often employed to assess the extent to which a model performs beyond the dataset on which it is trained. In one of its many variations, the so-called K-fold cross validation, a sample is partitioned into K nonoverlapping segments, a model is trained on K−1 of the segments, and then it is validated on the unused segment. The sample mean across the K validation performance measures, for example, mean-square error (MSE), is then taken to estimate the expected test error, also known as prediction error. In model selection, the model with the lowest estimated prediction error is selected.

The above model-selection criterion, however, ignores the fact that the estimated prediction error is only an estimate of the true (or expected) prediction error. A number of attempts have been made to build a confidence interval for the latter, but the results are preliminary, at best, because of the lack of independence between the various K−1 training segments (Hastie et al. 2001). There is even evidence that all estimators of variance in K-fold cross validation are biased (Bengio and Grandvalet 2004). A quantity that does play an important

Corresponding author: Caren Marzban, marzban@stat.washington.edu

DOI: 10.1175/AIES-D-21-0004.1 e210004
© 2022 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
role in building a confidence interval is the sample variance of
the $K$ validation performance measures (James et al. 2015). It
is, therefore, important to understand the sources of this
variability.

The distribution of local minima has been studied in a num-
ber of articles. For example, in the limit of a large number of
hidden nodes, certain distributional models for the training
error and the weights suggest that the distribution of local
minima tend to cluster around global minima (Choromanska et al. 2015). It has also been shown that the majority of critical
points in NN optimization are in fact not local minima but
saddle points (Dauphin et al. 2014). Both of these works are
theoretical in nature in that they examine theoretical limits,
for example, large number of hidden nodes, in the former,
and very high-dimensional optimization, in the latter. The
scope of the present work is entirely practical, and for that
reason the variance of performance measure due to local mi-
nima is estimated empirically, by retraining an NN starting
from different and random initial weights.

In the most common implementation of $K$-fold cross valida-
tion, the initial weights of each of the $K$ NNs are randomly
set. (The alternative, of initializing all of the $K$ NNs from a
unique set of weights is undesirable because it does not allow
an exploration of weight space.) Consequently, the variabil-
ity across the $K$ folds confounds the variability due to sam-
ping and that due to initialization. In this paper, the latter is
denoted $\sigma^2_{IW}$ [initialization of weights (IW)]. In what may
seem like confusing nomenclature, the variability due to sam-
ping is denoted $\sigma^2_{CV}$ [cross validation (CV)], and the
total variability observed in cross validation is the sum of
these two components.

Given this joint contribution to the total variability, it is
important that the two components are estimated using a mul-
tivariate model. Here, a methodology is developed for decom-
posing the total variability in the loss function (both training
and validation) into two components representing variability
due to sampling and due to local minima. The methodology is
applied to simulated as well as real data, and it is shown that
the two sources of variability can be comparable in magni-
tude. That finding highlights the importance of taking into
account both sources of variability when performing model
selection and/or comparison.

The importance of variance

Consider the central limit theorem (CLT), according to
which, under relatively general conditions, the quantity
$$z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}}$$
has a standard normal distribution, where $\bar{y}$ denotes the mean
of a sample of size $n$ drawn from a distribution with mean and
variance given by $\mu$ and $\sigma^2$, respectively. Let $y$ denote some
performance measure with an approximately normal distri-
bution [e.g., log(MSE)], in which case $\mu$ denotes the true
(population) mean of that performance measure. Then, the
CLT becomes the central foundation for comparing different
models—for example, NNs with different number of hidden
nodes—in terms of their performance. The appearance of $\sigma$
in the standard variate $z$ implies that the confidence interval
for $\mu$ becomes wider, and the $p$ value of a hypothesis test
becomes larger, as $\sigma$ increases. Consequently, an increase in
$\sigma$ is associated with a decrease in statistical power (i.e., the
probability of detecting an effect, if it truly does exist). Said
differently, a large $\sigma$ can lead to an inability to distinguish be-
tween models. Therefore, it is important to take appropriate
measures to reduce $\sigma$.

Note that $\sigma^2$ denotes the population variance and not the
sample variance. As such, it cannot be reduced through sam-
ping techniques. The only way it can be “reduced,” effect-
ively, is through its decomposition into other, smaller
components. For example, if $\sigma^2 = \sigma^2_1 + \sigma^2_2 + \cdots$, then each variance component $\sigma^2_i$, representing the variance of an iden-
tifiable source of variability in $y$, will be smaller than $\sigma^2$. Of
relevance to the present work, $\sigma^2_1$ and $\sigma^2_2$ may measure vari-
ability due to sampling and due to initialization, respectively.
With smaller variance components, then comparison of differ-
ent models in terms of each component, is guaranteed to have
higher statistical power.

It is extremely important to also note that statistical power is
the probability of detecting an effect, if it exists. In other
words, the aforementioned guarantee is valid only in the long
run, across multiple applications of inference, across multiple
samples, that is, only in expectation. Increased power cannot
be observed in an application of inference (confidence inter-
vals and hypothesis tests) to a single sample. As such, given
that in this paper the proposed methodology is applied to a
single sample, increased power cannot be demonstrated. In-
deed, demonstration of increased power is not the goal of this
work. The main goal of this work is to demonstrate that vari-
bility due to local minima and variability due to sampling
contribute comparably to the total variance observed in cross
validation. It then follows that separating the two variance
components leads to a “reduction” of the total variance and,
therefore, to higher statistical power.

2. Data

The term data may refer to several completely different
quantities. On the one hand, it can refer to data in the tradi-
tional sense of the word, generally characterized by the pres-
ence of experimental error in the observations. In this paper,
references to simulated data and real data are both instances
of traditional data. Each of these datasets will be used in per-
foming several experiments involving NNs. The results of the
experiments constitute data as well, but data of a different
kind. The phrase computer data has been proposed to de-
scribe this type of data (Sacks et al. 1989; Santner et al. 2003;
Fang et al. 2006; Welch et al. 1992). It is characterized by the
absence of experimental error, because performing a com-
puter experiment repeatedly leads to identical results. As dis-
cussed below, it is such computer data that are employed to
estimate the aforementioned variance components. Hence-
forth, and to minimize confusion, the term data will be used
to describe traditional data, and the phrase computer data will
refer to the MSE values generated from the experiments on the NN.

a. Simulated and real data

The simulated dataset involves 101 cases on one predictor $x$ and one response $y$. It is generated by varying the input of an NN with six hidden nodes on one hidden layer from $-1$ to $+1$, in 100 steps and recording the output. The weights of the NN are a random sample taken from a uniform distribution between $-10$ and $10$. The true/underlying relationship is shown in Fig. 1. Then, the simulated dataset is generated by adding to the $y$ values a random sample taken from a normal distribution with $\mu = 0$ and $\sigma = 1$; these data are displayed as the circles in Fig. 1. This way of generating data assures that the underlying function is one that is learnable by an NN with six hidden nodes.

The real dataset is the correction of a digital elevation model for a coastal marsh, estimated from a 3D point cloud measured by a terrestrial lidar survey (Nguyen et al. 2019). The aim of the NN is to predict the difference between the land cover. The number of cases is 607.

b. Computer data

The computer data necessary for estimating the variance components are generated as follows:

1) For an NN with a given number of hidden nodes $H$, the initial weights are sampled from a zero-mean normal distribution with standard deviation $\sigma_w$ (whose value is addressed below). Training is initiated on the training set (simulated and real), and it is continued until convergence has occurred. The values of MSE on the training and validation sets are recorded.

2) Step 1 is repeated for all $K = 10$ training sets in $K$-fold cross validation, with the value of MSE recorded for both the training and the validation sets. The initial weights are fixed for each of the $K = 10$ runs.

3) Steps 1 and 2 are repeated for $W = 11$ different initial weights. The resulting computer data can be organized in the form of the $K \times W$ matrix shown in Table 1, where only the training MSE values are shown. A similar matrix exists (not shown) for the validation MSE values.

4) Steps 1–3 are performed for a range of $H$ values (0, 2, 4, 6, 8, and 10 for the simulated dataset, and 0, 4, 8, and 12 for the real dataset). Varying $H$ allows one to examine how the histogram of local minima (i.e., the histogram of the training MSE values) depends on the complexity of NN. The histogram of the validation MSE values can also be monitored for the purpose of model selection.

Note that step 2 is contrary to how traditional cross validation is performed, wherein the weights are randomly initialized for each of the $K$ folds. In the present design, step 3 is where weights are randomized initially. The advantage of this design is that traditional cross validation is included as a subset. Specifically, the diagonal elements of Table 1 are MSE values that would be generated in traditional cross validation. This design is necessary for decomposing the total variability into its components.

The above four steps are sufficient for the development of models that estimate the variability of the MSE values due to sampling and due to local minima. However, given that $\sigma_w$ is a user-specified parameter in NN development, steps 1–4 are repeated for two different values of $\sigma_w$. Ideally, to make a deeper local minimum more likely, one should increase the density of points sampled in weight space. To increase that density, one can either increase the number of sampled points, or decrease $\sigma_w$, or both. For a given $\sigma_w$, increasing the number of sampled points in weight space grows with the dimensionality of that space and, therefore, with $H$, making it difficult to compare the results across different $H$ values. To avoid that complexity, here the number of points in weight space is fixed at $W = 11$, and only $\sigma_w$ is varied. To assure that different $\sigma_w$ values do lead to NNs that occupy distinct regions in weight space, with distinctly different MSE values, the histogram of MSE is generated across 500 random initializations, at two different values of $\sigma_w$ (0.1 and 10). The positions of these two histograms convey which of the two $\sigma_w$
values is associated with generally deeper local minima. It is important to emphasize that varying $\sigma_a$ is not necessary for the estimation of the two variance components. The $W = 500$ initializations with different values of $\sigma_a$ are examined here only for the purpose of better understanding how the proposed methodology depends on that user-specified quantity.

### 3. Method

As explained previously, the computer data are employed for the estimation of the variance components. Therefore, the experimental design adopted here includes one response variable MSE and two discrete factors denoted CV and IW, which, respectively, take values $i = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, W$. Here, $K = 10$ is the number of partitions in $K$-fold cross validation and $W = 11$ is the number of different NN initializations. Consequently, the computer data generated in this design can be organized in the form of a $K \times W$ matrix with each element containing the corresponding MSE value (e.g., Table 1).

The model for the computer data is

$$\text{MSE}_{ij} = \mu + CV_i + IW_j + \epsilon_{ij}, \quad (1)$$

where the response $\text{MSE}_{ij}$ denotes the value of MSE when the factors CV and IW take values $i$ and $j$, respectively. The term $\mu$ denotes the true mean of the response across all values of the factors, and $\epsilon$ accounts for any factor that has not been included in the model.\(^3\) We point out that this model is not a traditional regression model mapping the factors CV and IW to MSE, because that would require the factors to be quantitative variables. Here, although the factors take integer values, those values are simply labels; the factors are in fact categorical variables. The terms $CV_i$ and $IW_j$ that appear on the right-hand side of this model denote the conditional mean of the response MSE at different levels of the factors CV and IW. Models of this type are often called linear models (or ANOVA-type models) (Montgomery 2009).

In one realization of linear models the factors on the right-hand side of the model are fixed (nonrandom) quantities, with the exception of the error term, which is assumed to be a zero-mean, normally distributed random variable with variance $\sigma^2_\epsilon$. Such models are called fixed-effects models, and statistical tests exist for testing whether any of the factors, or a given factor, has an effect on the response. In fixed-effects models, the results of the tests (e.g., reject or not reject the null hypothesis) apply to only the specific values/levels of the factors appearing in the computer data. For example, if the CV factor is found to have a small $p$ value, all one can conclude is that there is evidence that the true mean of the response varies across the specific values/levels taken by that factor in the computer data. To generalize that conclusion to all possible values/levels, one must treat the factors as random variables.

When all of the factors on the right-hand side (except $\mu$) are treated as random variables, then the model is called a random-effects model. The simplest probability model for CV and IW is that they are zero-mean, normally distributed variables, with corresponding variances satisfying

$$\sigma^2_{\text{MSE}} = \sigma^2_{\text{CV}} + \sigma^2_{\text{IW}} + \sigma^2_{\epsilon}. \quad (2)$$

The various terms on the right side of this expression are called variance components. It is estimates of these variance components that quantify the contribution of the two sources of variability to the total variability of MSE. Two common estimators for the variance components are called the ANOVA estimator, and the restricted-maximum-likelihood (REML) estimator (Montgomery 2009); the former is the simpler of the two, but it has the defect of sometimes leading to negative values for the estimates of the variance components—an undesirable occurrence. For this reason, the latter estimator is used in the present study.\(^4\)

Given that the variance components are population parameters, one can build confidence intervals for them. Although analytic formulas exist for the confidence interval when the ANOVA estimator is used, analytic results for the REML-based confidence intervals are not known; however, approximate formulas do exist (Montgomery 2009), and they are used here. Also, given that the main focus of the present work

---

\(^3\) The $\epsilon_{ij}$ term can also be interpreted as an interaction term; see section 5.

\(^4\) It turns out that the final results are similar for the two estimators.
is the relative magnitude of the two variance components, the final results are reported in terms of the ratio $\sigma^2_{w}/\sigma^2_{CV}$. The confidence interval for that ratio is straightforward to compute analytically, because the ratio of the respective sample variances follows the $F$ distribution (Montgomery 2009).

The reliability of the confidence intervals requires assuming that the response MSE has a normal distribution with a constant variance within every possible combination of $i$ and $j$ in Eq. (1). To that end, a Box–Cox transformation is performed on the MSE values to maximize normality. Then, Q–Q plots are employed to check the validity of the assumptions.

4. Results

As mentioned in section 2, although the estimation of the variance components requires the computer data generated by the above experimental design, the training of the NNs requires data as well. Recall that two datasets will be examined: 1) a simulated, bivariate dataset constructed such that the true/underlying relationship between the predictor and response is learnable by an NN with $H = 6$ and 2) a real, multivariate dataset consisting of 12 predictors and 1 response variable. The statistics of these two datasets will not be presented, because the focus of the present study is the modeling of the computer data on MSE and the factors CV and IW. Suffice it to say that all of the usual preprocessing steps for the development of NNs are performed (Masters 1993); specifically, the inputs to the NNs are log transformed to assure bell-shaped histograms and then are standardized to have 0 mean and a standard deviation of 1. Also, collinearity is examined to assure no two predictors are highly correlated. As for the computer data, Q–Q plots are used to confirm that the MSE values satisfy the assumptions necessary for the construction of confidence intervals.

As explained in section 3, two values of $\sigma_w$ are selected to examine the sensitivity of the results on the depth of the local minima. Figure 2 shows the histogram of the training MSE values across 500 initializations of NNs with $H = 12$ trained on the real data. This pattern is observed for $H = 4$ and $H = 8$, as well (not shown). It is evident that the smaller value of $\sigma_w$ is associated with distinctly lower values of MSE. The extent to which $\sigma_w$ controls the overall depth of the local minima is addressed in section 5.

Before estimating the variance components in Eq. (2), it is instructive to examine the computer data on MSE, CV, and IW, for different values of $H$ and of the standard deviation $\sigma_w$ of the initial weights. To be specific, consider the computer data on training MSE values shown in Table 1, corresponding to the simulated dataset. The marginal medians of this matrix convey useful information: the histogram of the column medians conveys information about the distribution of local minima; similarly, the histogram of the row medians displays information about the sampling variability of MSE. Both of these histograms pertain to the training MSE values; there are two analogous histograms for the validation MSE values. Then, summarizing each histogram with a boxplot leads to four boxplots denoted trn.CV, trn.IW, vld.CV, and vld.IW, for every experiment performed here (i.e., every combination of $H$ and $\sigma_w$).

Figure 3 shows all of the boxplots for different values for the number of hidden nodes $H$, and the standard deviation of the initial weights $\sigma_w$. To enhance visual acuity, the logarithm of MSE is shown. Evidently, for a given value of $\sigma_w$, the training MSE values (trn.CV and trn.IW) generally decrease as $H$ increases. This finding is consistent with the expectation that a more complex NN leads to lower values of training MSE. For small $\sigma_w$ (i.e., 0.1), associated with deeper local minima (see Fig. 2), the validation boxplots are consistent with the commonly expected “U” pattern, suggesting that larger values of $H$ may lead to overfitting. The true/known value of $H = 6$ is also consistent with the observed pattern. For $\sigma_w = 10$, the validation boxplots do not follow the expected U pattern; although this is somewhat surprising, there has been recent evidence that such expectations are unwarranted when gradient methods are invoked for NN training (Sankararaman et al. 2019).

Of particular interest for the focus of this study are the spreads of the boxplots shown in Fig. 3. Intuitively, it is these that approximate the variance components $\sigma^2_{CV}$ and $\sigma^2_{IW}$ that one seeks to estimate. Although the estimation of these variance components involves all of the MSE values in the computer data (e.g., Table 1) and not the marginal medians only, Fig. 3 already allows one to qualitatively compare the variability of MSE due to sampling with that due to local minima. For example, consider the validation boxplots in Fig. 3; first, note that for larger values of $H$, say 8 and 10, the large spread of the boxplots precludes any meaningful comparisons. However, for smaller $H$ values, including the true value $H = 6$, the blue boxplots for $\sigma_w = 10$ are generally wider than those for $\sigma_w = 0.1$. Indeed, when the initial weights are sampled from a wider distribution, sampling variability is generally comparable to that due to local minima.
Figure 4 shows the analogous results for the real data. First, focusing on the training boxplots (top row), it can be seen that a larger variance for the initial weights is generally associated with larger MSE values. This observation is consistent with the finding in Fig. 1 based on 500 initializations (and no cross validation).

These results can be employed for two purposes—1) model selection and 2) the comparison of variability due to sampling and local minima. For the former, it is evident that the “best” model is that with $H = 50$ (i.e., a linear regression model), because the lowest validation errors are generally associated with $H = 0$; and that conclusion is valid for both choices of $s_w$, and regardless of the source of variability.

Another indication of overfitting—one that involves $s_w$—is also evident in these figures. Consider the $H = 12$ boxplots: the training MSE values for $s_w = 0.1$ are significantly lower than those for $s_w = 10$. Again, this is consistent with Fig. 1. However, the corresponding validation MSE values follow the reverse order, that is, with $s_w = 0.1$ boxplots being significantly higher than those for $s_w = 10$. This implies that a smaller $s_w$ leads to overfitting. This, too, may seem surprising, but it simply highlights the importance of $s_w$ in NN training.

As for the comparison of the two sources of variability, Fig. 4 implies that for small $s_w$, variability due to initialization are comparable in magnitude, and regardless of $s_w$. This finding is the reason for why it is important to take into account both sources of variability.

All of the above conclusions are based on qualitative comparisons of boxplots that summarize the computer data (e.g., Table 1). Moreover, only the row and column marginals of the computer data are used in the boxplots. The random effects models described in section 3 (method) can be employed to estimate the variance components, rigorously, and from all of the elements in the table (not the marginals only). Also, given that the main focus of this work is to compare the two sources of variability, the results are shown in terms of the ratio $s_{IW}^2/s_{CV}^2$.

Figure 5 shows the variance ratio $s_{IW}^2/s_{CV}^2$ inferred from the computer data generated from the simulated data. The approximate 95% confidence intervals are also shown. We also point out that in such figures the variance ratio at $H = 0$ is identically zero, because an NN with $H = 0$ (i.e., linear regression) has a strictly convex error function, and so there exists a unique global minimum.

A glance at the range of the $y$ axis in the two panels in Fig. 5 reveals that a larger spread in the initial weights of the NN
leads to larger variance ratios. More specifically, for \( \sigma_w = 0.1 \), variability due to local minima is less than that due to sampling. For \( \sigma_w = 10 \), however, the pattern is generally reversed, for all values of \( H \), with variability due to local minima dominating sampling variability by as much as an order of magnitude. Indeed, in such extreme limits, one would be justified to perform model selection based on variability due local minima, only. The apparent trend in the bottom panel is addressed in section 5; to that end, we note that the confidence intervals in the bottom panel are generally wider than those in the top panel.

The analogous results for the experiments based on real data are shown in Fig. 6. As in the case of the simulated data, a larger \( \sigma_w \) is generally associated with a larger variance ratio. The one possible exception is when \( H = 4 \), in which case the variance ratio is consistent with the value 1, for both values of \( \sigma_w \). Further analysis of these results is presented in the next section.

5. Conclusions and discussion

The local minima in the loss function landscape of NNs are examined, and it is shown that variability due to local minima can be comparable to, or even larger than, that due to sampling in \( K \)-fold cross validation. This implies that both sources of variability must be taken into account when performing model selection and/or comparison. A methodology is developed for estimating the contribution of the two sources of variability. Moreover, the methodology is sufficiently general to allow for the estimation of other sources of computational variability.

One may question the practical utility of estimating variance components. After all, at no point in this work is a
That variance ought to be decomposed into components

1) The variance of the performance measure across cross

2) That variance ought to be decomposed into components

3) The row and column marginals of the computer data table

4) Random effects models of the computer data allow for

In the real data examined here, two values of \( \sigma_w \) are identified as having significantly different values of MSE, and the behavior of the variance components, as a function of \( H \), is

examined at those values. In other words, the user specified \( \sigma_n \) is used to “control” the depth of the local minima. Specifically, it was found that the two values \( \sigma = 0.1 \) and \( 10 \) correspond to deeper and shallower local minima, respectively. The study of the precise relationship between \( \sigma_n \) and the depth of the local minima is beyond the scope of this work, but it is important to note that the above guidance applies regardless of that specific relationship. Said differently, to estimate the two variance components, it is not necessary to vary \( \sigma_n \) at all. That said, it makes for good practice to do so.

As mentioned in footnote 3, the \( \epsilon_i \) term in the random effects model in Eq. (1) can also be interpreted as an interaction term. In that case, \( \epsilon_i \) measures the extent to which the effect of the one factor (say, CV) on the response MSE varies with the other factor (IW). In ANOVA-type models like Eq. (1), there are two approaches that are generally followed: 1) The degrees of freedom associated with the interaction term are used to perform inference on the main effects of each factor, or 2) treat the \( \epsilon \) term as an interaction, but then give up the ability to perform inference (Montgomery 2009). In the present work, focus has been on the main effects from CV and IW, and the ability to perform inference on them. Another reason why the first approach is followed here is that according to the hierarchy principle, interaction terms generally have a much smaller magnitude than main effects. This phenomenon is generally borne out by several “principles”: the principle of hierarchical ordering, the principle of effect sparsity, and the principle of effect hierarchy; see Montgomery (2009, pp. 192, 230, 272, 314, and 329) and Li et al. (2006, 33–34).

Several extensions of the present work are possible. For instance, among the sources of computational variability another quantity that is often set by the user is the number of training epochs. It is commonly expected that increasing the number of training epochs may lead to a deeper local minimum. However, a natural question that arises then is, what proportion of the total variance is due to the number of training epochs, which in turn suggests an extension of the model in Eq. (2) wherein there is also an “epoch” factor on the right side. This extension of the model, as well as other generalizations will be examined elsewhere.

Some of the panels in Figs. 5 and 6 display trends in the way \( \sigma_{\text{IV}}/\sigma_{\text{CV}} \) varies with \( H \). For example, the bottom panel in Fig. 5 suggests that for the simulated data, when \( \sigma_n \) is large, the variance ratio decreases with \( H \). By contrast, that trend appears to be reversed in Fig. 6 where real data are examined. Such apparent patterns must be tempered by the confidence intervals. Indeed, given the relatively wide confidence intervals in these figures, it is possible that the apparent trends are not statistically significant. One of the causes of the wide confidence intervals is the choice of the REML estimator for the variance components, and the approximations employed here for the resulting confidence intervals. It may be worthwhile to examine other estimators and alternative means of computing confidence intervals, for example those proposed by Vanwinckelen and Blockeel (2012).

The formalism developed here is guided by a comparison of variability due to local minima with that due to sampling in

---

**Fig. 6.** As in Fig. 5, but estimated from the computer data on MSE generated from training NNs on the real data.
10-fold cross validation. \( K \) is set to 10 because there is evidence that 10-fold cross validation constitutes an acceptable trade-off between bias and variance (Hastie et al. 2001). However, most recently, evidence has begun to gather that the usual arguments suggesting that variance increases with \( K \) may be flawed (Bengio and Grandvalet 2004; Burman 1989; Zhang and Yang 2015). As such, it will be interesting to extend the analysis performed here to other values of \( K \) to see whether the relationship between the sources of variability depends on \( K \). Indeed, it will be interesting to compare variability due to local minima with that due to other resampling methods, for example, bootstrap.

Here, the local minima are all minima in the loss function of NNs, trained with a conjugate gradient method. Given that the loss function landscape is a function of both the model and/or optimization algorithm, another useful extension of this work will be to examine other models and/or methods, for example, bootstrap.

It is not immediately clear how much of the substantive conclusions in this work extend to deep learning. For example, is it likely that in some applications of deep learning, the variability due to local minima may be as large as or larger than that due to sampling? We conjecture that the answer is yes simply because deeper NNs are more likely to have more local minima. That said, there is nothing in the proposed methodology itself that prevents it from being applied to deep learning.

Acknowledgments. We are grateful to Xiaopeng Cai, Chuyen Nguyen, and Mike Starek for providing the real data and related discussions and to Zaid Harchaoui for bringing to our attention the literature on the benefits of overparameterizing neural networks.

Data availability statement. The (computer) data generated and used for this study are available upon request from the corresponding author.

REFERENCES


