universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Disseratation / Title of the Doctoral Thesis

„Statistical analysis and development of inference procedures post-model-selection“

verfasst von / submitted by

Mag. Danijel Kivaranovic BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2020 / Vienna, 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 794 370 136
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Statistik und Operations Research
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Hannes Leeb

Abstract

Properties of inference procedures from traditional inferential statistics were very well studied in the last decades. On the other hand, the development of prediction procedures that allow for valid inference post-model-selection is an emerging and highly active research area. It is very important to better understand these procedures from analytical perspective. In this thesis we propose new and analyze existing inference procedures post-model-selection. We consider inference procedures for population parameters and new observations of the underlying data generating process. The latter is also known as predictive inference. We provide analytical results on the length of confidence intervals and we propose prediction intervals for complex learning algorithms and non-standard targets. Our results are based on the theory of Selective Inference and Conformal Inference.

Kurzfassung

Eigenschaften von Inferenzprozeduren der traditionellen Inferenzstatistik wurden in den letzten Jahrzehnten sehr detailiert analysiert. Auf der anderen Seite sind Inferenzprozeduren, die valide Inferenz nach Modellselektion ermöglichen, noch ein sehr junges und aktives Forschungsgebiet. Es ist sehr wichtig, diese Prozeduren aus analytischer Sicht zu verstehen. In dieser Doktorarbeit entwickeln wir neue und analysieren bestehende Inferenzprozeduren nach Modellselektion. Dabei betrachten wir Inferenzprozeduren für Populationsparameter und neue Beobachtungen eines zugrundeliegenden Datengenerierungsprozesses. Letzteres ist auch als prädiktive Inferenz bekannt. Wir haben analytische Ergebnisse zu der Länge von Konfidenzintervallen und wir entwickeln prädiktive Intervalle für komplexe Lern-Algorithmen und untypische Zielvariablen. Unsere Resultate basieren auf der Theorie der selektiven Inferenz sowie der konformalen Inferenz.

Contents

1 Introduction

It is common practice that scientists first collect data, then select an appropriate statistical model based on the data and finally draw inference from the selected model. In such a scenario, it is well known that methods from standard inferential statistics can lead to highly misleading conclusions because these standard methods require an a priori known statistical model. A very illustrative example of the risks of methods that neglect the randomness of the selected model was given by Leeb and Pötscher (2005). However, in the last decades and even today, many empirical studies still do not appropriately adjust for model selection step when performing inference. The development and understanding of valid inference procedures post-model selection is very active research area that significantly contributes to resolving the replication crisis that affects most empirical research areas (Ioannidis, 2005).

By a valid inference procedure we mean a procedure which outputs a confidence interval that covers a population parameter or a new observation of the data generating process with probability at least $1 - \alpha$, where $\alpha \in (0, 1)$ is the prespecified significance level. Usually there is a one-to-one relation between a confidence interval and a hypothesis test that rejects a certain null hypothesis with probability at least $1 - \alpha$. In this thesis we solely concentrate on the construction of confidence intervals and their properties and do not discuss hypothesis tests any further. The thesis addresses both, valid confidence intervals after model selection for population parameters and new observations. The latter are also called prediction intervals. We briefly discuss the current state of reasearch for both types of intervals in the following.

There are two promising approaches for inference on population parameters post-model-selection, namely, inference conditional on a selected model and inference simultaneous over all potential models. The former is known as Selective Inference and was introduced by Lee et al. (2016), while the latter is known as POSI and was introduced by Berk et al. (2013). The method of Lee et al. (2016) unfortunately has the undesirable property that the confidence intervals have infinite conditional expected length if the conditioning event is "small" (Kivaranovic and Leeb, 2020a). POSI intervals do not suffer such issues, however these intervals are still very conservative in many situations because they provide universal protection against any model selection procedure. It is important to note that in both approaches, Selective Inference and POSI, non-standard population parameters are considered that depend on the selected model. Covering the true underlying population parameters after model selection is a more difficult task as shown by the impossibility results of Leeb and Pötscher (2006, 2008). In this thesis we only address the confidence intervals from Selective Inference. More precisely, we analyze the interval length of variations of the original method of Lee et al. (2016) which have been proposed by Tian et al. (2018) and Fithian et al. (2015). So far, there are no analytical results on the interval length of these variations, only simulation results that suggest that confidence intervals of these variations are considerably shorter than the original method. In Section 1.2 we give a brief introduction to Selective Inference.

The conformal method (Vovk et al., 2005) has become a very popular and powerful tool to construct valid and accurate prediction intervals. The conformal method is a non-parametric approach that can be applied with any model selection procedure and any learning algorithm to construct valid prediction intervals. One limitation of the conformal method is that it only controls coverage in the marginal sense, meaning, that we average over all training data sets and all new observation. Another limitation is that the computation of the prediction intervals is very expensive or even infeasible if the dataset is large or a complex learning is used such as a neural network. To reduce the computational cost, Lei et al. (2018) proposed the split-conformal approach, which splits the training data into two parts where one part is used for training the model and the other for construction the prediction intervals based on the fitted model. In this thesis we propose a conformal procedure that is tailored to neural networks that outperforms many existing predictive inference procedures in terms of interval length. We further show how to apply the conformal method to more complex situations. More precisely, we modify the conformal method to cover the individual treatment effect, a quantity that is not observable in the training data. The individual treatment effect is a quantity that, for example, measures the efficacy of a new drug in randomized clinical trial (Rubin, 2005). In Section 1.2 we give a brief introduction to Conformal Inference.

1.1 Selective Inference

First we describe the regression setting. Let $n \in \mathbb{N}$ denote the sample size. Let Y denote the $N(\theta, \sigma^2 I_n)$-distributed response vector, where $\theta \in \mathbb{R}^n$, $\sigma^2 \in (0, \infty)$ and I_n is the n-dimensional identity matrix. Let $A \in \mathbb{R}^{n \times d}$ be the non-stochastic regressor matrix, where $d \in \mathbb{N}$ denotes the number of variables. The mean vector θ is not further restricted, meaning that θ is not necessarily a linear function of A. A model selection procedure $\hat{m}$ is a measurable function of Y with values in the power set of $\{1, \dots, d\}$. For a model $m \subseteq \{1, \dots, d\}$, denote by A_m the matrix consisting of the columns of A that corresponding to the model m. The model-dependent target in the Selective Inference literature is defined as

$$\beta_m = \text{argmin}_{b \in \mathbb{R}^{|m|}} \|\theta - A_m b\|_2^2 = (A_m' A_m)^{-1} A_m' \theta;$$

this means, β_m denotes that coefficients of the best linear approximation of μ in the model m. We note that the target is well-defined if A_m has full column rank. The most common model selection procedures, such as the Lasso, only select models such that A_m has full column rank. A natural estimator of β_m is

$$\hat{\beta}_m = \text{argmin}_{b \in \mathbb{R}^{|m|}} \|Y - A_m b\|_2^2 = (A_m' A_m)^{-1} A_m Y;$$

this means, the least squares solution restricted to the sub-model m. For any $\gamma \in \mathbb{R}^{|m|}$, the method of Lee et al. (2016) (also called polyhedral method) allows to construct a confidence interval for $\gamma' \beta_m$ conditional on the event $\{\hat{m} = m\}$ if the event $\{\hat{m} = m\}$ can be written as a finite union of sets of the form $\{BY \leq b\}$, where B and b are a matrix and a vector, respectively. Many model selection procedures, such as the Lasso or Stepwise Regression (Lee et al., 2016; Tibshirani et al., 2016), satisfy this polyhedral condition.

The construction of the confidence interval is based on the conditional distribution of

$$\gamma' \hat{\beta}_m \mid \{\hat{m} = m, (I_n - P_m) Y\},$$

where P_m is the orthogonal projection on the space spanned by $A_m(A_m'A_m)^{-1}\gamma$. It turns out that this conditional distribution is a truncated normal distribution. Kivaranovic and Leeb (2020a) showed that confidence intervals based on a truncated normal distribution have infinite expected length if the truncation set is bounded from below and/or above. They further showed that the truncation sets generated by the Lasso are typically bounded from above and/or below. This means that the polyhedral method typically produces intervals with infinite expected length if the Lasso is used as model selector.

To reduce interval length, Tian et al. (2018) proposed to add noise in the model selection step, meaning, for an n-dimensional noise vector ω with known distribution, model selection should be performed on $Y + \omega$ instead of Y. Alternatively, Fithian et al. (2015) proposed to perform model selection only on a subset of the data instead of the full data set. Simulation results indicate that both variations produce considerably shorter intervals than the original method of Lee et al. (2016) at the price of accuracy in the model selection step.

1.2 Conformal Inference

We again describe the regression setting first. Let n be the sample size as before. We denote by X the random covariate vector with values in $\mathbb{R}^d$, where d is again the number of variables. Let Y denote the real-valued outcome of interest. We can decompose Y into

$$Y = f(X) + e_X,$$

where $f(X) = \mathbb{E}(Y|X)$ is called the regression function and $e_X = Y - \mathbb{E}(Y|X)$ the error. Let $D_n = ((X_i, Y_i))_{1 \leq i \leq n}$ denote a data set of n observations, where the (X_i, Y_i)s are i.i.d. copies of (X, Y). In the following, (X, Y) denotes a new observation that is independent of D_n. Let $f_{D_n}(X)$ be an estimator of the regression function $f(X)$. The most frequently used non-conformity measure $S_{D_n}(x, y)$, $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$, is the absolute error, that is,

$$S_{D_n}(x, y) = |y - f_{D_n}(x)|.$$

A conformal prediction interval (Vovk et al., 2005) for Y is now constructed as follows: Let $D_{n+1}(y)$ denote the augmented dataset which contains D_n and the pseudo observation (X, y). Because of the i.i.d. assumption it is easy to see that the observations

$$S_{D_{n+1}(Y)}(X_1, Y_1), \ldots, S_{D_{n+1}(Y)}(X_n, Y_n), S_{D_{n+1}(Y)}(X, Y)$$

are exchangeable. Denote by $R_{n+1}(Y)$ the rank of $S_{D_{n+1}(Y)}(X, Y)$ among the $n+1$ random variables above. Because of exchangeability it follows that $R_{n+1}(Y)$ has a discrete uniform distribution on $\{1, \ldots, n+1\}$. But this implies that

$$1 - \alpha \leq \mathbb{P}\left(R_{n+1}(Y) \leq \lceil (1-\alpha)(n+1) \rceil\right) \leq 1 - \alpha + 1/(n+1).$$

This inequality allows us to define the set

$$\Gamma_{D_n}(X) = \{y : R_{n+1}(y) \leq \lceil (1-\alpha)(n+1) \rceil\}$$

that satisfies

$$1 - \alpha \leq \mathbb{P}(Y \in \Gamma_{D_n}(X)) \leq 1 - \alpha + 1/(n+1).$$

In general, $\Gamma_{D_n}(X)$ is not necessarily an interval but obviously $[\inf\ \Gamma_{D_n}(X), \sup\ \Gamma_{D_n}(X)]$ is always a prediction interval.

Observe that this procedure is valid for any estimator $f_{D_n}(X)$. Hence we can apply any learning algorithm and any model selection procedure to obtain a valid prediction interval. The length of the prediction interval depends on how accurate $f_{D_n}(X)$ is estimating $f(X)$. Complex machine learning algorithms, such as a neural network, are known to outperform simple algorithms, such as linear regression, if the regression function $f(X)$ is non-linear. However, note that to compute $\Gamma_{D_n}(X)$, we need to fit $f_{D_{n+1}(y)}(X)$ for all $y \in \mathbb{R}$. This is computationally challenging if the computation time of a single function $f_{D_{n+1}(y)}(X)$ is not negligible. Lei et al. (2018) proposed the split-conformal approach which reduces computation time drastically at the price of prediction accuracy. In the split-conformal case, only one model $f_{D_m}(X)$ needs to be fitted for a $m < n$. The non-conformity measure is then computed on the remaining $n - m$ observations.

1.3 Organization of the Thesis

The thesis is written in a cumulative way, meaning that the thesis consists of the following three individual manuscripts: Kivaranovic and Leeb (2020b), Kivaranovic et al. (2020a) and Kivaranovic et al. (2020b). The first manuscript corresponds to the Selective Inference literature, while the remaining two to the Conformal Inference literature. Each manuscript has its own list of references and its own page numbering. In addition we have a page numbering for the whole document which displayed in the bottom right or left corner. In the following, we briefly describe the contributions of each of the tree manuscripts.

In Chapter 2, we present the manuscript of Kivaranovic and Leeb (2020b). Here we derive an upper bound for the length of confidence intervals based on a distribution from a certain class of conditional distributions of normal random variables. We show that this bound is tight if the truncation set of the conditioning event is bounded from below and/or above. This result has significant importance for the Selective Inference literature because the considered class of conditional distributions frequently arises there (Tian et al., 2018; Fithian et al., 2015). The truncated normal distribution is a limit case of this class and our results imply that confidence intervals based on the truncated normal distribution can only be bounded by ∞. Kivaranovic and Leeb (2020a) already showed that confidence intervals based on the truncated normal distribution has infinite expected length if the truncation set is bounded from below and/or above.

In Chapter 3, we present the manuscript of Kivaranovic et al. (2020a). Here we show how to a construct prediction intervals that are tailored to neural networks. We propose prediction intervals with two different coverage guarantees: the usual average coverage as in the conformal theory and a new approximate conditional coverage given the training data. At the core of both coverage guarantees is a novel loss function that allows to accurately estimate prediction intervals that take the heteroscedasticity of the response into account. For the average coverage guarantee, we define a non-conformity measure that is very specific to this new loss function and for the approximate conditional coverage guarantee we need to select a data-dependent parameter of the loss function that controls the length of the intervals. We also demonstrate in simulations that these prediction intervals are close to optimal, meaning that the confidence bounds approximate the conditional quantiles of the response given the covariates in the limit.

In Chapter 4, we present the manuscript of Kivaranovic et al. (2020b). In this manuscript, we show how to apply the conformal theory to construct prediction intervals for the individual treatment effect, a measure that is for example of high interest in randomized clinical trials. The individual treatment effect of a patient is the difference of the outcome under a new treatment and the outcome under the current standard treatment. The difficulty is that the individual treatment effect is not observable because the outcome of each patient is only observed for one of the two treatments and never for both. We show how to modify the conformal method to construct intervals that control conditional coverage of the outcome given either the new or standard treatment and than combine these two intervals to obtain a prediction interval for the individual treatment effect. We provide prediction intervals that are valid under weak assumptions but they are typically conservative because these intervals are valid even in the worst-case scenario. We also provide prediction intervals under stronger assumptions which are considerable shorter.

In the following, we provide the submission status and the relative contribution of the authors of each of the three manuscripts.

- Kivaranovic and Leeb (2020b):
  - submission status: working paper (not submitted);
  - relative contribution: 85% Danijel Kivaranovic, 15% Hannes Leeb.
- Kivaranovic et al. (2020a):
  - submission status: published in the *Proceedings of Machine Learning Research* (accepted on January 7th, 2020);
  - relative contribution: 70% Danijel Kivaranovic, 20% Kory Johnson, 10% Hannes Leeb.
- Kivaranovic et al. (2020b):
  - submission status: under review at the *Journal of the American Statistical Association* (submitted on June 1st, 2020);
  - relative contribution: 70% Danijel Kivaranovic, 10% Robin Ristl, 10% Martin Posch, 10% Hannes Leeb.

Bibliography

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41:802–837.

Fithian, W., Taylor, J., Tibshirani, R., and Tibshirani, R. J. (2015). Selective sequential model selection. *arXiv preprint arXiv:1512.02565*.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2:e124.

Kivaranovic, D., Johnson, K. D., and Leeb, H. (2020a). Adaptive, distribution-free prediction intervals for deep neural networks. *Proceedings of Machine Learning Research*, 108:4346–4356.

Kivaranovic, D. and Leeb, H. (2020a). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association*, forthcoming.

Kivaranovic, D. and Leeb, H. (2020b). A (tight) upper bound for the length of confidence intervals with conditional coverage. *arXiv preprint arXiv:2007.12448*.

Kivaranovic, D., Ristl, R., Posch, M., and Leeb, H. (2020b). Conformal prediction intervals for the individual treatment effect. *arXiv preprint arXiv:2006.01474*.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44:907–927.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21:21–59.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34:2554–2591.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24:338–376.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100:322–331.

Tian, X., Taylor, J., et al. (2018). Selective inference with a randomized response. *The Annals of Statistics*, 46:679–710.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111:600–620.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

2 A (tight) upper bound for the length of confidence intervals with conditional coverage

A (tight) upper bound for the length of confidence intervals with conditional coverage

Danijel Kivaranovic[a] Hannes Leeb[a,b]

[a]Department of Statistics and Operations Research, University of Vienna

[b]Data Science @ Uni Vienna, University of Vienna

Abstract

Kivaranovic and Leeb (2020) showed that confidence intervals based on a truncated normal distribution can have infinite expected length. In this paper, we generalize the results of Kivaranovic and Leeb (2020) by considering a larger class of conditional distributions of normal random variables of which the truncated normal distribution is a limit case of. We derive a finite upper bound for the length of confidence intervals based on a distribution from this class. This means the length of these intervals is always bounded. Furthermore, we show that this upper bound is tight if the conditioning event is bounded. We apply our result to two popular selective inference procedures which are known as data carving (Fithian et al. (2015)) and Lasso selection with a randomized response (Tian et al. (2018b)) and show that these two procedures dominate the sample splitting strategy in terms of interval length.

Keywords: Inference after model selection, confidence interval, interval length

1

1 Introduction

Kivaranovic and Leeb (2020) showed that confidence intervals based on a truncated normal distribution can have infinite expected length. As a consequence, methods, such as the polyhedral method of Lee et al. (2016), typically produce confidence intervals with infinite expected length. In this paper, we generalize the results of Kivaranovic and Leeb (2020) by considering a larger class of conditional distributions of normal random variables of which the truncated normal distribution is a limit case of. We derive a finite upper bound for the length of confidence intervals based on a distribution from this class. This means, the length of these intervals is always bounded. Furthermore, we show that this upper bound is tight if the conditioning event is bounded (meaning that we can make the probability of this event arbitrarily small by choosing appropriate parameters). Such distributions frequently arise in the selective inference literature. For sake of demonstration, we apply our result to two popular procedures which are known as data carving (Fithian et al. (2015)) and Lasso selection with a randomized response (Tian et al. (2018b)). In these two papers, it is reported that inference after selection on a subset (Fithian et al. (2015)) or selection with noisy data (Tian et al. (2018b)) is more powerful (in terms of power and interval length) than the polyhedral method of Lee et al. (2016), which performs selection and inference on the same data. However, no formal proofs of these observations were given so far. In this paper, we do not only derive a (tight) upper bound for the length of confidence interval of the two aforementioned procedures, but we also prove that these two procedures dominate the popular sample splitting strategy in terms of interval length.

There are several ongoing developments of inference procedures after model selection. Roughly speaking, one can divide them into two branches: inference conditional on a selected model and inference simultaneous over all potential models. As pioneer works in these two areas we can mention Lee et al. (2016) and Berk et al. (2013), respectively. Our main result addresses interval procedures in the first of these two branches, namely, inference procedures that evolved from the polyhedral method of

Lee et al. (2016). See, among others, the works of Fithian et al. (2015); Markovic et al. (2018); Panigrahi et al. (2018); Panigrahi and Taylor (2019); Reid et al. (2017, 2018); Taylor and Tibshirani (2018); Tian et al. (2016); Tian and Taylor (2017); Tian et al. (2018a,b); Tibshirani et al. (2016). For related literature to the second of these two branches, see among others, Bachoc et al. (2019, 2020); Kuchibhotla et al. (2018a,b). We also want to note that all these works address model-dependent targets and not the true underlying parameters in the classical sense. Valid inference for an underlying true parameter is a more challenging task, as demonstrated by the impossibility results in Leeb and Pötscher (2006, 2008).

This paper is structured as follows. In Section 2 we present the main technical result, Theorem 1. Here we show that a confidence interval based on a particular conditional distribution of normal random variables has bounded length. We also perform simulations that provide deeper insights. Theorem 1 may not appear relevant as an independent result, however, as we show in Section 3, the conditional distributions we consider frequently arise in selective inference. In particular, we can apply Theorem 1 to the two aforementioned procedures from selective inference to show that they have bounded length and dominate the popular sample splitting strategy in terms of interval length. The paper ends with a discussion in Section 4. The proofs of the results are given in the appendix.

2 Main technical result

Before we present the main technical result, we briefly explain the importance of this section. Our result is a (tight) upper bound for confidence intervals that are constructed based on the conditional distribution of

$$X \mid X + U \in T, \tag{1}$$

where X is a $N(\mu, \sigma^2)$-distributed random variable with $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$, U is a $N(0, \tau^2)$-distributed random variable with $\tau^2 \in (0, \infty)$ and independent of X and

$T \subseteq \mathbb{R}$ is a truncation set of the form

$$T \;=\; \bigcup_{i=1}^{k} (a_i, b_i), \tag{2}$$

where $k \in \mathbb{N}$ and $\infty \leq a_1 < b_1 < \cdots < a_k < b_k \leq \infty$. In the subsequent Section 3, we show that confidence intervals of the two selective inference procedures of Fithian et al. (2015) and Tian et al. (2018b) are constructed based an the exact same conditional distribution. This means, the properties of the confidence intervals considered in this section can be directly carried over to the confidence intervals of these two selective inference procedures. Hence it is helpful to understand first the simplified case of this section before going over to the more complex settings in the next section. We continue now with the presentation of our main technical result

Let $\Phi(x)$ be the c.d.f. of the standard normal distribution and denote by $F^{\tau^2}_{\mu,\sigma^2}(x)$ the conditional c.d.f. of the random variable in (1). For sake of readability, we do not show the dependence of the c.d.f. on T in the notation. For $q \in (0,1)$, let $\mu_q(x)$ satisfy

$$F^{\tau^2}_{\mu_q(x),\sigma^2}(x) \;=\; 1-q. \tag{3}$$

Such an $\mu_q(x)$ exists, is unique and is strictly increasing in q for every $x \in \mathbb{R}$ (this is guaranteed by Lemma A.4). For all $q_1, q_2 \in (0,1)$, such that $q_1 \leq q_2$, we have

$$\mathbb{P}(\mu \in [\mu_{q_1}(X), \mu_{q_2}(X)] \mid X + U \in T) \;=\; q_2 - q_1.$$

This equation is a textbook result for confidence bounds (see chapter 3.5 in Lehmann and Romano (2006)). A common choice is to set $q_1 = \alpha/2$ and $q_2 = 1-\alpha/2$ such that, conditional on $\{X+U \in T\}$, $[\mu_{q_1}(X), \mu_{q_2}(X)]$ is an equal-tailed confidence interval for μ at level $1-\alpha$. Another option is to choose q_1 and q_2 such that, conditional on $\{X+U \in T\}$, $[\mu_{q_1}(X), \mu_{q_2}(X)]$ is an unbiased confidence interval at level $1-\alpha$ (see chapter 5.5 in Lehmann and Romano (2006)).

It is easy to see that, as τ^2 goes to 0, $F^{\tau^2}_{\mu,\sigma^2}(x)$ converges weakly to the c.d.f. of the truncated normal distribution with mean μ, variance σ^2 and truncation set T as considered in Kivaranovic and Leeb (2020). In the case where the intervals are based on

$\lim_{\tau^2 \to 0} F^{\tau^2}_{\mu,\sigma^2}(x)$, Kivaranovic and Leeb (2020) showed that the length of $[\mu_{q_1}(x), \mu_{q_2}(x)]$ is unbounded if the truncation set T is bounded from above and/or below and the length is bounded otherwise. On the other hand, as τ^2 goes to ∞, it is not hard to show that $F^{\tau^2}_{\mu,\sigma^2}(x)$ converges weakly to the c.d.f. of the normal distribution with mean μ and variance σ^2. Hence, in the case where the intervals are based on $\lim_{\tau^2 \to \infty} F^{\tau^2}_{\mu,\sigma^2}(x)$, the length of $[\mu_{q_1}(x), \mu_{q_2}(x)]$ is equal to $\sigma\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right)$. These observations suggest that, in the case where $\tau^2 \in (0, \infty)$, the length of $[\mu_{q_1}(x), \mu_{q_2}(x)]$ is bounded somewhere between $\sigma\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right)$ and ∞, which is unfortunately not informative.

In the following result we show that, if $\tau^2 \in (0, \infty)$, the length of $[\mu_{q_1}(x), \mu_{q_2}(x)]$ is always bounded. We provide an upper bound that does not depend on the truncation set T and we show that this upper bound is tight if the truncation set T is bounded from above or below (or both).

Theorem 1. *Let $\tau^2 \in (0, \infty)$ and let T be defined in (2). Let $F^{\tau^2}_{\mu,\sigma^2}(x)$ be the c.d.f. of (1) and let $\mu_q(x)$ satisfy (3). Set $\rho^2 = \tau^2/(\sigma^2 + \tau^2)$. Then, for all $x \in \mathbb{R}$ and all $q_1, q_2 \in (0, 1)$ such that $q_1 \leq q_2$,*

$$\mu_{q_2}(x) - \mu_{q_1}(x) \;<\; \frac{\sigma}{\rho}\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right).$$

If $\sup T = b_k < \infty$, the length converges to the upper bound as $x \to \infty$. If $\inf T = a_1 > -\infty$, the length converges to the upper bound as $x \to -\infty$.

Note that the upper bound in the theorem diverges as τ^2 goes to 0. On the other hand, as τ^2 goes to ∞, the upper bound converges to $\sigma\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right)$. This means, this result perfectly fills the gap between the $\tau^2 = 0$ and the $\tau^2 = \infty$ case.

In Figure 2, we plotted the length of $[\mu_{q_1}(x), \mu_{q_2}(x)]$ as a function of x for several truncation sets T. In the left panel the truncation sets are of the form $(-a, a)$ and in the right panel the truncation sets are of the form $(-\infty, -a) \cup (a, \infty)$. The different values for a are shown below the plot. We set $q_1 = 1 - q_2 = 0.025$ and $\sigma^2 = \tau^2 = 1$. The dashed line on top denotes our upper bound $\sigma/\rho\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right)$. The other dashed line denotes $\sigma\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right)$, which corresponds to the length of

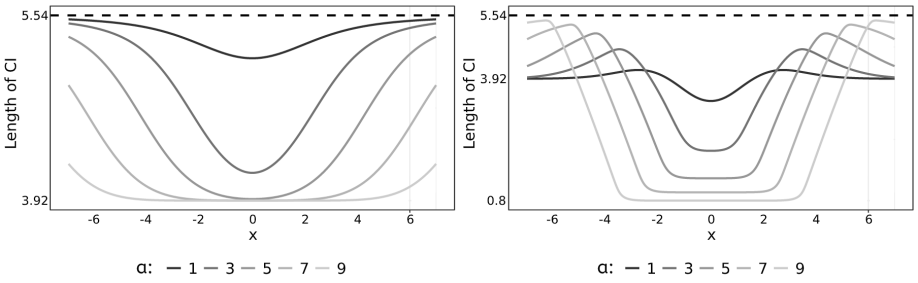

Figure 1: The length $[\mu_{q_1}(x), \mu_{q_2}(x)]$ is plotted as a function of x. In the left panel the truncation sets are of the form $(-a, a)$ and in the right panel truncation sets are of the form $(-\infty, -a) \cup (a, \infty)$. The different values for a are shown below the plot. The remaining parameters where: $q_1 = 1 - q_2 = 0.025$ and $\sigma^2 = \tau^2 = 1$.

the confidence interval with unconditional coverage. On the l.h.s., we see that $\mu_{q_2}(x) - \mu_{q_1}(x)$ approximates the upper bound as $|x|$ diverges. The smaller the truncation set (this means the smaller a), the faster the convergence is. Also in this case, where the truncation set is an interval, the left panel indicates that the length bounded from below by the $\sigma\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right)$. In the right panel we see that our upper bound is not tight when the truncation set is unbounded. In fact the length converges to $\sigma\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right)$ as x diverges. However, as the gap of the truncation set becomes larger (this means as a grows), we see that the length approximates the upper bound for values around a and $-a$. Surprisingly, $\sigma\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right)$ is not an lower bound in this case, as we can see that the length is considerably below it for values in the gap $(-a, a)$. In fact, it seems that the length converges to 0 for values of x around 0 as a diverges.

In Figure 2, we plotted the conditional expected length of $[\mu_{q_1}(X), \mu_{q_2}(X)]$ as a function of μ for several truncation sets T. For this simulations, we drew 2000 random variables as in (1), computed the confidence interval for each of these observations and estimated the conditional expected length by the sample mean of the lengths. In the left panel we again have truncation sets of the form $(-a, a)$ and in the right panel we have truncation sets of the form $(-\infty, -a) \cup (a, \infty)$. All parameters were identical to

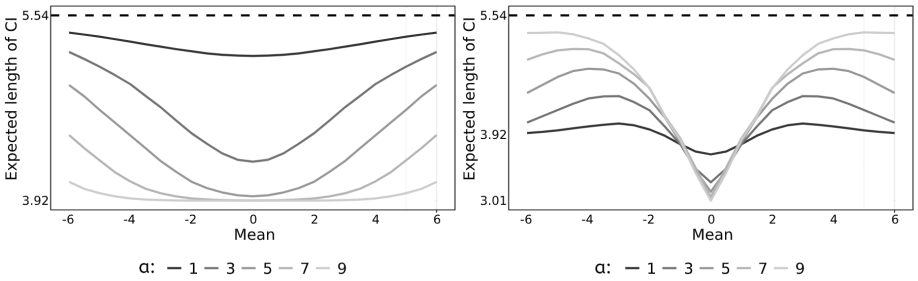

Figure 2: The conditional expected length of $[\mu_{q_1}(X), \mu_{q_2}(X)]$ is plotted as a function of μ. In the left panel the truncation sets are of the form $(-a, a)$ and in the right panel the truncation sets are of the form $(-\infty, -a) \cup (a, \infty)$. The different values for a are shown below the plot. The remaining parameters where: $q_1 = 1 - q_2 = 0.025$ and $\sigma^2 = \tau^2 = 1$.

the setting of Figure 2. In the left panel we see that the conditional expected length is minimized at μ equal to 0 and converges to the upper bound as μ diverges. We also see that the smaller the truncation set, the larger the conditional expected length. This means, the conditional expected length is close to the upper bound if the probability of the conditioning event is small. In the right panel we again see that the conditional expected length is minimized at μ equal to 0. However, it does not converge the upper bound but to $\sigma\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right)$ as μ diverges. Surprisingly, the conditional expected length at μ equal 0 decreases as a increases and becomes significantly smaller than $\sigma\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right)$. This means, in contrast to the left panel, we observe that the conditional expected length decreases as the probability of the conditioning event decreases.

3 Application to selective inference

In selective inference, distributions of the form (1) frequently arise. We show this based on two interesting examples: the first one is data carving (Fithian et al. (2015)), the second one is Lasso selection with a randomized response (Tian et al. (2018b)). We

briefly describe the two applications in the corresponding subsection. For more details, we refer to the aforementioned papers.

3.1 Data carving

Data carving means that only a subset of the data used for selection or exploratory analysis and the full data for inference. In the following, we describe a simple sample mean setting. Let $n \in \mathbb{N}$ and let $X_1, \dots, X_n$ be i.i.d. random variables of the normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$. Let T and $F^{\tau^2}_{\mu,\sigma^2}(x)$ be defined as in Section 2. Let $\delta \in (0,1)$ be such that δn is a positive integer. For any $m \in \{1, \dots, n\}$, let $\bar{X}_m$ denote the sample mean of the first m observations. The exploratory check that is performed on the first δn observations can typically be described as an event of the form $\{\bar{X}_{\delta n} \in T\}$. Inference for μ is based on the conditional distribution of

$$\bar{X}_n \mid \bar{X}_{\delta n} \in T \tag{4}$$

Proposition 3.1. *The conditional c.d.f. of the random variable in* (4) *is equal to* $F^{\tilde{\tau}^2}_{\mu,\tilde{\sigma}^2}(x)$ *with truncation set* T *and*

$$\tilde{\sigma}^2 \;=\; \frac{\sigma^2}{n} \quad \textit{and} \quad \tilde{\tau}^2 \;=\; \frac{\sigma^2}{n}\frac{1-\delta}{\delta}.$$

This is a rather trivial result that uses elementary properties of the normal distribution, therefore, the proof is omitted here. Let $\tilde{\mu}_q(x)$ satisfy $F^{\tilde{\tau}^2}_{\mu,\tilde{\sigma}^2}(x) = 1 - q$, where $\tilde{\sigma}^2$ and $\tilde{\tau}^2$ are as in the proposition. Theorem 1 implies that

$$\tilde{\mu}_{q_2}(x) - \tilde{\mu}_{q_1}(x) \;<\; \frac{\sigma}{\sqrt{(1-\delta)n}} \left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right).$$

This means, in terms of length, the confidence interval $[\tilde{\mu}_{q_1}(\bar{X}_n), \tilde{\mu}_{q_2}(\bar{X}_n)]$ dominates the sample splitting strategy because the confidence interval based on the observations $X_{\delta n+1}, \dots, X_n$ always has length $\sigma(\Phi^{-1}(q_2) - \Phi^{-1}(q_1))/\sqrt{(1-\delta)n}$ (since these observations are independent of the event $\{\bar{X}_{\delta n} \in T\}$). Furthermore, we see that the length of $[\tilde{\mu}_{q_1}(\bar{X}_n), \tilde{\mu}_{q_2}(\bar{X}_n)]$ shrinks at the same $\sqrt{n}$-rate as in the unconditional case. The price

of conditioning is at most the factor $1/\sqrt{1-\delta}$. But if δ is equal to 1, the upper bound is equal to ∞ and the interval is known to have infinite conditional expected length (Kivaranovic and Leeb (2020)).

3.2 Selection with a randomized response

Let $n \in \mathbb{N}$, let Y be $N(\theta, \sigma^2 I_n)$-distributed random vector, where $\theta \in \mathbb{R}^n$, $\sigma^2 \in (0, \infty)$ and I_n is the n-dimensional identity matrix. Let $A \in \mathbb{R}^{n \times d}$, where $d \in \mathbb{N}$. Finally, let ω be $N(0, \tau^2 I_n)$, where $\tau^2 \in (0, \infty)$, and assume that Y and ω are independent. We call Y the response vector, A the (non-stochstic) regressor matrix and ω the (normal) noise vector.

The Lasso estimator, denoted by $\hat{\beta}(y)$, is a minimizer of the least squares problem with an additional penalty on the absolute size of the regression coefficients (Frank and Friedman (1993); Tibshirani (1996)):

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{2}\|y - A\beta\|_2^2 + \lambda\|\beta\|_1, \quad y \in \mathbb{R}^n, \ \lambda \in (0, \infty).$$

The Lasso has the property that some coefficients of $\hat{\beta}(y)$ are zero with positive probability. We assume that the columns of A are in general position which assures that $\hat{\beta}(y)$ is the unique minimizer of the Lasso objective function (Tibshirani (2013)). Let $\hat{m}(y) \subseteq \{1, \dots, d\}$ denote the non-zero variables selected by the Lasso and let $\hat{s}(y) \in \{-1, 1\}^{|\hat{m}(y)|}$ denote the sign vector of the non-zero coefficients of $\hat{\beta}(y)$. Let $m \subseteq \{1, \dots, d\}$ denote a model and $s \in \{-1, 1\}^{|m|}$ denote a sign vector. For $\gamma \in \mathbb{R}^{|m|}$, set

$$\eta_m = A_m (A_m' A_m)^{-1} \gamma$$

and denote by P_{η_m} the orthogonal projection on the space spanned by η_m. Note that $\eta_m' Y$ is simply a linear combination of the least squares solution restricted to the submodel m. This means, if γ is equal to the ith basis vector than $\eta_m' Y$ is the ith component of the restricted least squares solution. The polyhedral method of Lee et al. (2016) allows to construct a confidence interval for $\eta_m' \theta$ with coverage control conditional on

$\{\hat{m}(Y) = m\}$ or on $\{\hat{m}(Y) = m, \hat{s}(Y) = s\}$. In either case, the construction of the intervals is based on the conditional distribution of

$$\eta_m' Y \mid \eta_m' Y \in T(z),\ (I_n - P_{\eta_m})Y = z,$$

where $T(z)$ is a union of open intervals if we condition only on the model or a single open interval if we condition on the model and on the signs (the endpoints of the set $T(z)$ may be infinite). See Lee et al. (2016) for the derivation of this pivotal quantity. Note that in either case the set $T(z)$ has the same form as the truncation set T defined in (2). In case of conditioning on the model only, the set $T(z)$ is typically bounded from above or below (or both) (Figure 3 in Kivaranovic and Leeb (2020)). On the other hand, in case of conditioning on the model and the signs, $T(z)$ is always bounded from above and/or below (Proposition 2 in Kivaranovic and Leeb (2020)). Because $(I_n - P_{\eta_m})Y$ is by construction independent of $\eta_m' Y$, the conditional distribution of the random variable above is equal to a truncated normal distribution with mean $\eta_m'\theta$, variance $\sigma^2 \eta_m' \eta_m$ and truncation set $T(z)$. Kivaranovic and Leeb (2020) showed that confidence intervals based on a truncated normal distribution have infinite conditional expected length if the truncation set is bounded from above or below (or both). Because, this is typically the case for Lasso selection, the intervals of the polyhedral method are quite long.

To improve the performance of the polyhedral method, Tian et al. (2018b) proposed to add some noise ω in the selection step and construct a confidence interval for $\eta_m'\theta$ with coverage control conditional on $\{\hat{m}(Y+\omega) = m\}$ or on $\{\hat{m}(Y+\omega) = m, \hat{s}(Y+\omega) = s\}$. In this case, the construction of the intervals is based on the conditional distribution of

$$\eta_m' Y \mid \eta_m'(Y+\omega) \in T(z),\ (I_n - P_{\eta_m})(Y+\omega) = z, \tag{5}$$

where $T_m(z)$ is as before. Observe that because of linearity, we can write $\eta_m'(Y+\omega)$ as $\eta_m' Y + \eta_m'\omega$. Also, because ω is independent of Y it follows that $\eta_m' Y$ is independent of $(I_n - P_{\eta_m})(Y+\omega)$.

Proposition 3.2. *The conditional c.d.f. of the random variable in* (5) *is equal to*

$F^{\bar{\tau}^2}_{\bar{\mu},\bar{\sigma}^2}(x)$ *with truncation set* $T(z)$ *and*

$$\bar{\mu} \;=\; \eta_m'\theta, \quad \bar{\sigma}^2 \;=\; \sigma^2\eta_m'\eta_m, \quad \bar{\tau}^2 \;=\; \tau^2\eta_m'\eta_m.$$

This result is again trivial, because the random variable in (5) has the exact same form as the random variable in (1) because the truncation set $T(z)$ is fixed and $(I_n - P_{\eta_m})(Y + \omega)$ is independent of $\eta_m' Y$. Let $\bar{\mu}_q(x)$ satisfy $F^{\bar{\tau}^2}_{\bar{\mu},\bar{\sigma}^2}(x) = 1 - q$, where $\bar{\mu}$, $\bar{\sigma}^2$ and $\bar{\tau}^2$ are as in the proposition. Theorem 1 implies that

$$\bar{\mu}_{q_2}(x) - \bar{\mu}_{q_1}(x) \;<\; \frac{\sigma\sqrt{\eta_m'\eta_m}}{\rho}\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right),$$

where ρ is again equal to $\tau^2/(\sigma^2 + \tau^2)$. In the case where $[\bar{\mu}_{q_1}(\eta_m' Y), \bar{\mu}_{q_2}(\eta_m' Y)]$ controls coverage conditional on $\{\hat{m}(Y + \omega) = m, \hat{s}(Y + \omega) = s\}$, the upper bound is tight, because truncation set $T(z)$ is always bounded from above and/or below (Proposition 2 in Kivaranovic and Leeb (2020)). On the other hand, when we control coverage conditional on $\{\hat{m}(Y + \omega) = m\}$, the upper bound is not necessarily tight because the truncation set $T(z)$ may be unbounded (although the truncation set is typically bounded as shown in Figure 3 in Kivaranovic and Leeb (2020)).

We see that adding noise in the selection step can drastically improve the performance of the confidence interval $[\bar{\mu}_{q_1}(\eta_m' Y), \bar{\mu}_{q_2}(\eta_m' Y)]$ in terms of length. The length of this intervals shrinks at the same rate as in the unconditional case. The price of conditioning is at most the factor $1/\rho$. Compared to the sample splitting strategy, the interval $[\bar{\mu}_{q_1}(\eta_m' Y), \bar{\mu}_{q_2}(\eta_m' Y)]$ always dominates (up to a negligible multiplicative constant that converges to 1 as n goes to ∞) the interval where $1 - \rho^2$ percent of the data is used for selection and the remaining ρ^2 percent for inference.

4 Discussion

We proved that the length of confidence intervals with conditional coverage drastically improves if some noise is added to the selection step (or in form of subsampling). Our result clearly supports the observation that an significant increase in power is

obtained by decreasing the 'power' of the model selection step itself. Selection and inference on the same data is not favorable if one knows that the events produced by the selection procedure are bounded, as this will lead to intervals with infinite expected length. However, we also want to emphasize that this does not mean that we do not recommend selection and inference on the same data. For example, in the setting of Heller et al. (2019), where first a global hypothesis is tested and subsequent tests are only performed if the global hypothesis is rejected, bounded selection regions do not arise and excessively long intervals are not an issue there. Hence, we rather recommend to be cautious about the selection procedure one chooses. In some situations, adding noise in the selection step (or in form of subsampling) may be beneficial, in other situations it may not be necessary.

Even though, we only considered Gaussian random variables here, in the examples of Section 3, the central limit theorem justifies the asymptotic validity of our upper bound as long as the conditioning event has positive probability in the limit.

Appendix A Proof of Theorem 1

Throughout this section let $\tau \in (0,1)$ and let T be defined in (2). We simplify the notation by setting $F_\mu(x) = F^{\tau^2}_{\mu,\sigma^2}(x)$, where $F^{\tau^2}_{\mu,\sigma^2}(x)$ is the conditional c.d.f. of the random variable in (1). The results hold for any fixed σ^2 and τ^2 in $(0,\infty)$. Additionally, we will denote by $f_\mu(x) = f^{\tau^2}_{\mu,\sigma^2}(x)$ the conditional p.d.f. of the random variable in (1). We set $\rho^2 = \tau^2/(\sigma^2+\tau^2)$. For $q \in (0,1)$ let $\mu_q(x)$ be the solution of equation (3). We denote by $\phi(x)$ and $\Phi(x)$ the p.d.f. and c.d.f. of the standard normal distribution. We use the usual convention that $\phi(-\infty) = \phi(\infty) = \Phi(-\infty) = 1 - \Phi(\infty) = 0$.

The rest of this section is structured as follows. First we provide some intuition behind the theorem. Second, we state Proposition A.1 and A.2 which are the two core results which the proof of Theorem 1 relies on. Third, we prove Theorem 1 with the help of these two propositions. In Section A.1 and A.2 we prove Proposition A.1 and A.2, respectively. The first of these two propositions is considerably more difficult to prove. In Section A.3 we have collected several auxiliary results which are required for the proofs of the main results.

Observe that the random vector $(X, X+U)'$ has a two-dimensional normal distribution with mean $(\mu,\mu)'$, variance $(\sigma^2, \sigma^2+\tau^2)'$ and covariance σ^2. It is elementary to show that for any $v \in \mathbb{R}$,

$$X \mid X + U = v \;\sim\; N(\rho^2\mu - (1-\rho^2)v, \sigma^2\rho^2). \tag{6}$$

Let $G_\mu(x,v)$ denote the c.d.f. of this normal distribution, this means,

$$G_\mu(x,v) = \Phi\left(\frac{x - (\rho^2\mu + (1-\rho^2)v)}{\sigma\rho}\right). \tag{7}$$

By definition of $F_\mu(x)$, we can write

$$F_\mu(x) = \mathbb{E}\left[G_\mu(x, V_\mu)\right], \tag{8}$$

where V_μ is a random variable that is truncated normally distributed with mean μ, variance $\sigma^2+\tau^2$ and truncation set T. Observe that if T is equal to the singleton

$\{v\}$ (this case is excluded by definition of T in (2)), the c.d.f.s $G_\mu(x, v)$ and $F_\mu(x)$ coincide. Suppose for a moment that T is the singleton $\{v\}$. Because $F_\mu(x)$ is then the normal c.d.f. $G_\mu(x, v)$, it is easy to show that the length of $[\mu_{q_1}(x), \mu_{q_2}(x)]$ is equal to $(\sigma/\rho)(\Phi^{-1}(q_2)-\Phi^{-1}(q_1))$, which is exactly the upper bound in the inequality of Theorem 1. This means, the theorem implies that confidence intervals only become shorter if we condition on a set T with positive Lebesgue measure instead of a singleton. On the other hand, if T is equal to $\mathbb{R}$, it is clear that the length of $[\mu_{q_1}(x), \mu_{q_2}(x)]$ is equal to $\sigma(\Phi^{-1}(q_2) - \Phi^{-1}(q_1))$. Surprisingly, this is not necessarily a lower bound for a proper subset $T \subset \mathbb{R}$ as you can see on the r.h.s. of Figure 2 or in Figure 1 of Kivaranovic and Leeb (2020) in the case where τ is equal to 0.

Proposition A.1. *For all $x \in \mathbb{R}$ and all $\mu \in \mathbb{R}$, we have*

$$\frac{\partial \Phi^{-1}(F_\mu(x))}{\partial \mu} \;<\; -\frac{\rho}{\sigma}. \tag{9}$$

We briefly consider the limit of the derivative as τ^2 goes to ∞ and 0, respectively. In the case $\tau^2 \to \infty$, it is not hard show that $F_\mu(x)$ converges $\Phi((x-\mu)/\sigma)$ and, therefore, the derivative converges to $-1/\sigma$. On the other hand, as $\tau^2 \to 0$, $F_\mu(x)$ converges to the c.d.f. of the truncated normal distribution with mean μ, variance σ^2 and truncation set T. In this limit case, Proposition 1 of Kivaranovic and Leeb (2020) implicitly implies that the derivative can only be bounded by 0 if T is bounded from above and/or below.

Proposition A.2. *Let $G_\mu(x, v)$ be defined as in (7) and let $q \in (0, 1)$. If $\sup T = b_k < \infty$, then*

$$\lim_{x \to \infty} G_{\mu_q(x)}(x, b_k) = 1 - q.$$

If $\inf T = a_1 > -\infty$, then

$$\lim_{x \to -\infty} G_{\mu_q(x)}(x, a_1) = 1 - q.$$

This proposition means that $F_{\mu_q(x)}(x)$ and $G_{\mu_q(x)}(x, b_k)$ are asymptotically equivalent as $x \to \infty$ if the truncation set T is bounded from above (and analogous in the case where T is bounded from below). We continue now with the proof of Theorem 1.

Proof of Theorem 1. By definition of $\mu_{q_1}(x)$, we have $\Phi^{-1}(F_{\mu_{q_1}(x)}(x)) = \Phi^{-1}(1-q_1)$. Because Proposition A.1 holds for any x and μ, it follows that for any $c \in (0,\infty)$, we have

$$\Phi^{-1}(F_{\mu_{q_1}(x)+c}(x)) < \Phi^{-1}(1-q_1) - \rho c/\sigma.$$

We plug $c = \sigma(\Phi^{-1}(q_2) - \Phi^{-1}(q_1))/\rho$ into the inequality, apply the strictly increasing function Φ to both sides and use the symmetry $\Phi^{-1}(q) = -\Phi^{-1}(1-q)$ to obtain

$$F_{\mu_{q_1}(x)+\sigma(\Phi^{-1}(q_2)-\Phi^{-1}(q_1))/\rho}(x) < 1 - q_2.$$

Because $F_\mu(x)$ is strictly decreasing in μ by Lemma A.4 and $\mu_{q_2}(x)$ satisfies the equation $F_{\mu_{q_2}(x)}(x) = 1 - q_2$, the previous inequality implies that

$$\mu_{q_2}(x) < \mu_{q_1}(x) + \frac{\sigma}{\rho}\left(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\right).$$

Subtracting $\mu_{q_1}(x)$ on both sides gives the inequality of the theorem. It remains to show that this upper bound is tight if the truncation set T is bounded. We only show the case $\sup T = b_k < \infty$. The case $\inf T = a_1 > -\infty$ follows by similar arguments. In view of the definition of $G_\mu(x,v)$ in (7), Proposition A.2 and the symmetry $\Phi^{-1}(q) = -\Phi^{-1}(1-q)$ imply that

$$\lim_{x\to\infty} \frac{x - (\rho^2\mu_{q_1}(x) + (1-\rho^2)b_k)}{\sigma\rho} = -\Phi^{-1}(q_1)$$

and

$$\lim_{x\to\infty} \frac{x - (\rho^2\mu_{q_2}(x) + (1-\rho^2)b_k)}{\sigma\rho} = -\Phi^{-1}(q_2).$$

Subtracting the second limit from the first limit and multiplying by σ/ρ gives

$$\lim_{x\to\infty} \mu_{q_2}(x) - \mu_{q_1}(x) = \frac{\sigma}{\rho}(\Phi^{-1}(q_2) - \Phi^{-1}(q_1)).$$

Hence the upper bound is tight. □

A.1 Proof of Proposition A.1

The proof of the Proposition is split up into a sequence of lemmas that are directly proven here.

Lemma A.1. *For all $x \in \mathbb{R}$ and all $\mu \in \mathbb{R}$, we have*

$$\frac{\partial \Phi^{-1}(F_\mu(x))}{\partial x} \;<\; \frac{1}{\sigma\rho}. \tag{10}$$

Proof. By the inverse function theorem, we have $\partial\Phi^{-1}(q)/\partial q = 1/\phi(\Phi^{-1}(q))$. This equation and the chain rule imply that

$$\frac{\partial \Phi^{-1}(F_\mu(x))}{\partial x} = \frac{f_\mu(x)}{\phi(\Phi^{-1}(F_\mu(x)))}.$$

In the following, we show that the numerator is bounded from above by $\phi(\Phi^{-1}(F_\mu(x)))/(\sigma\rho)$. It is easy to see that the inequality of the lemma then follows.

In view of equation (8) and (7), Leibniz's rule implies that the p.d.f. $f_\mu(x)$ is equal to

$$f_\mu(x) \;=\; \frac{1}{\sigma\rho}\mathbb{E}\left[\phi\left(\frac{x-(\rho^2\mu-(1-\rho^2)V_\mu)}{\sigma\rho}\right)\right] \;=\frac{1}{\sigma\rho}\mathbb{E}\left[\phi\left(\Phi^{-1}(G_\mu(x,V_\mu)\right)\right].$$

Observe now that it is sufficient to show that the function $\phi(\Phi^{-1}(q))$ is strictly concave because by Jensen's inequality it follows then that $f_\mu(x)$ is bounded from above by

$$\frac{1}{\sigma\rho}\phi\left(\Phi^{-1}\left(\mathbb{E}\left[G_\mu(x,V_\mu)\right]\right)\right) \;=\; \frac{1}{\sigma\rho}\phi(\Phi^{-1}(F_\mu(x))).$$

Elementary calculus shows that

$$\frac{\partial^2 \phi(\Phi^{-1}(q))}{\partial q^2} = \frac{-1}{\phi(\Phi^{-1}(q))}.$$

Because $\phi(\Phi^{-1}(q))$ is positive for all $q \in (0,1)$, it follows that the second derivative is negative for all $q \in (0,1)$. This means, $\phi(\Phi^{-1}(q))$ is strictly concave and the proof is complete. □

Note that the inequality (10) of this lemma resembles inequality (9) of Proposition A.1. While inequality (10) is surprisingly easy to prove, inequality (9) is more difficult. Equation (8) provides intuition why this is the case: The distribution of the random variable V_μ does not depend on x but it depends on μ. Hence to prove inequality (9), we cannot exchange integral and differential and we cannot apply Jensen's inequality

as we did in the proof of Lemma A.1. We did not find a direct proof of inequality (9). However, in the following we show that inequality (10) implies (9).

To show this implication, we need a different representation of $f_\mu(x)$. Elementary calculus and properties of the conditional normal distribution imply that the conditional p.d.f. $f_\mu(x)$ can also be written as

$$f_\mu(x) \;=\; \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\frac{\sum_{i=1}^k \Phi\left(\frac{b_i-x}{\tau}\right) - \Phi\left(\frac{a_i-x}{\tau}\right)}{\sum_{i=1}^k \Phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right) - \Phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)}. \tag{11}$$

Lemma A.2. *Let $G_\mu(x,v)$ be defined as in (7). For all $x \in \mathbb{R}$ and all $\mu \in \mathbb{R}$, we have*

$$\frac{\partial \Phi^{-1}(F_\mu(x))}{\partial \mu} = \frac{-f_\mu(x) + \sum_{i=1}^k h(b_i)(F_\mu(x) - G_\mu(x,b_i)) - h(a_i)(F_\mu(x) - G_\mu(x,a_i))}{\phi(\Phi^{-1}(F_\mu(x)))},$$

where

$$h(v) = \frac{\frac{1}{\sqrt{\sigma^2+\tau^2}}\phi\left(\frac{v-\mu}{\sqrt{\sigma^2+\tau^2}}\right)}{\sum_{i=1}^k \Phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right) - \Phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)}.$$

Proof. The chain rule implies that

$$\frac{\partial \Phi^{-1}(F_\mu(x))}{\partial \mu} = \frac{\partial \Phi^{-1}(F_\mu(x))}{\partial F_\mu(x)}\frac{\partial F_\mu(x)}{\partial \mu}.$$

The inverse function theorem implies that the first derivative on the r.h.s. is equal to $1/\phi(\Phi^{-1}(F_\mu(x)))$. This means, it remains to show that the second derivative is equal to the numerator on the r.h.s. of the equation of this lemma. Leibniz's rule implies that $\partial F_\mu(x)/\partial\mu = \int_{-\infty}^x \partial f_\mu(u)/\partial\mu \; du$. Therefore, we compute $\partial f_\mu(x)/\partial\mu$ first. Lemma A.5 implies that

$$\frac{\partial f_\mu(x)}{\partial \mu} = \frac{x-\mu}{\sigma^2} f_\mu(x) + f_\mu(x)\sum_{i=1}^k h(b_i) - h(a_i).$$

Lemma A.6 implies that the integral of the first summand is equal to

$$-f_\mu(x) - \sum_{i=1}^k h(b_i)G_\mu(x,b_i) - h(a_i)G_\mu(x,a_i).$$

Because the second summand on the r.h.s. depends on x only through $f_\mu(x)$, it is easy to see the integral of this function is equal to

$$F_\mu(x)\sum_{i=1}^k h(b_i) - h(a_i).$$

The sum of the last two expressions is equal to the numerator which completes the proof. □

This lemma implies that proving Proposition A.1 is equivalent to showing that $B_\mu(x) < 0$, where

$$B_\mu(x) = \frac{\rho}{\sigma}\phi(\Phi^{-1}(F_\mu(x))) - f_\mu(x) + \sum_{i=1}^{k} h(b_i)(F_\mu(x) - G_\mu(x, b_i)) - h(a_i)(F_\mu(x) - G_\mu(x, a_i)). \quad (12)$$

Observe that

$$\begin{aligned} 0 &= \lim_{|x|\to\infty} f_\mu(x) = \lim_{|x|\to\infty} \phi(\Phi^{-1}(F_\mu(x)) \\ &= \lim_{x\to-\infty} F_\mu(x) = \lim_{x\to\infty} 1 - F_\mu(x) \\ &= \lim_{x\to-\infty} G_\mu(x, v) = \lim_{x\to\infty} 1 - G_\mu(x, v). \end{aligned}$$

This equation chain implies that $B_\mu(x)$ converges to 0 as $|x| \to \infty$. Holding μ, σ^2 and τ^2 fixed, this is the same as saying that $B_\mu(x)$ converges to 0 as $F_\mu(x)$ goes to 0 or 1. Let the function $F_\mu^{-1}(q)$ be defined by the equation

$$F_\mu(F_\mu^{-1}(q)) = q. \quad (13)$$

Clearly, $F_\mu^{-1}(q)$ is well-defined for all $q \in (0, 1)$ and we have that $F_\mu^{-1}(F_\mu(x)) = x$ for all $x \in \mathbb{R}$. To prove Proposition A.1, it is sufficient to show that $B_\mu(F_\mu^{-1}(q))$ is strictly convex as a function of q for any fixed μ, σ^2 and τ^2.

Lemma A.3. *Let $B_\mu(x)$ be defined in (12) and $F_\mu^{-1}(q)$ in (13). Let μ, σ^2 and τ^2 be fixed. Then, for all $x \in \mathbb{R}$,*

$$\frac{\partial^2 B_\mu(F_\mu^{-1}(q))}{\partial q^2}\Big|_{q=F_\mu(x)} = -\frac{\rho}{\sigma\phi(\Phi^{-1}(F_\mu(x)))} + \frac{1}{\sigma^2 f_\mu(x)}.$$

Proof. We start by computing the first derivative of $B_\mu(F_\mu^{-1}(q))$ with respect to q. The chain rule implies that

$$\frac{\partial B_\mu(F_\mu^{-1}(q))}{\partial q} = \frac{\partial B_\mu(F_\mu^{-1}(q))}{\partial F_\mu^{-1}(q)} \frac{\partial F_\mu^{-1}(q)}{\partial q}.$$

The inverse function theorem implies that the second derivative on the r.h.s. is equal to $1/f_\mu(F_\mu^{-1}(q))$. In view of the definitions of $B_\mu(x)$ and $F_\mu^{-1}(q)$ in (12) and (13), we see that to compute the first derivative on the r.h.s. we need to compute the derivatives of q, $\phi(\Phi^{-1}(q))$, $f_\mu(F_\mu^{-1}(q))$ and $G_\mu(F_\mu^{-1}(q), v)$ with respect to $F_\mu^{-1}(q)$. Because $\partial F_\mu^{-1}(q)/\partial q$ is equal to $1/f_\mu(F_\mu^{-1}(q))$, it follows that the derivative of q with respect to $F_\mu^{-1}(q)$ is equal to $f_\mu(F_\mu^{-1}(q))$. The chain rule implies that the derivative of $\phi(\Phi^{-1}(q))$ with respect to $F_\mu^{-1}(q)$ is equal to $-\Phi^{-1}(q)f_\mu(F_\mu^{-1}(q))$. Lemma A.7 implies that the derivative of $f_\mu(F_\mu^{-1}(q))$ with respect to $F_\mu^{-1}(q)$ is equal to

$$-\frac{F_\mu^{-1}(q)-\mu}{\sigma^2}f_\mu(F_\mu^{-1}(q)) - f_\mu(F_\mu^{-1}(q))\sum_{i=1}^{k} l(F_\mu^{-1}(q), b_i) - l(F_\mu^{-1}(q), a_i),$$

where $l(x, v)$ is defined in Lemma A.7. And finally, the Lemma A.8 implies that the derivative of $G_\mu(F_\mu^{-1}(q), v)$ with respect to $F_\mu^{-1}(q)$ is equal to

$$\frac{f_\mu(F_\mu^{-1}(q))l(F_\mu^{-1}(q), v)}{h(v)},$$

where $h(v)$ is defined in Lemma A.2 and $l(x, v)$ is defined in Lemma A.7. From the previous four derivatives and the definition of $B_\mu(x)$ in (12), we can conclude that

$$\frac{\partial B_\mu(F_\mu^{-1}(q))}{\partial F_\mu^{-1}(q)} = f_\mu(F_\mu^{-1}(q))\left(-\frac{\rho}{\sigma}\Phi^{-1}(q) + \frac{F_\mu^{-1}(q)-\mu}{\sigma^2} + \sum_{i=1}^{k} h(b_i) - h(a_i)\right)$$

and therefore

$$\frac{\partial B_\mu(F_\mu^{-1}(q))}{\partial q} = -\frac{\rho}{\sigma}\Phi^{-1}(q) + \frac{F_\mu^{-1}(q)-\mu}{\sigma^2} + \sum_{i=1}^{k} h(b_i) - h(a_i).$$

Now it is easy to see that

$$\frac{\partial^2 B_\mu(F_\mu^{-1}(q))}{\partial q^2} = -\frac{\rho}{\sigma\phi(\Phi^{-1}(q))} + \frac{1}{\sigma^2 f_\mu(F_\mu^{-1}(q))}.$$

The claim of the lemma follows by evaluating the second derivative at $q = F_\mu(x)$. □

To prove Proposition A.1, it remains to show that

$$-\frac{\rho}{\sigma\phi(\Phi^{-1}(F_\mu(x)))} + \frac{1}{\sigma^2 f_\mu(x)} > 0.$$

But this is the same as showing

$$\frac{f_\mu(x)}{\phi(\Phi^{-1}(F_\mu(x)))} < \frac{1}{\sigma\rho}.$$

The l.h.s. is equal to $\partial\Phi^{-1}(F_\mu(x))/\partial x$ and in Lemma A.1 we showed that this inequality is true. This completes the proof of Proposition A.1.

A.2 Proof of Proposition A.2

Let $\epsilon > 0$. We will only consider the case where $\sup T = b_k < \infty$. The other case follows by the same arguments. Recall the definition of $F_\mu(x)$ in (8). By the law of total probability, we can write $F_\mu(x)$ as

$$\mathbb{P}(V_\mu > b_k - \epsilon)\mathbb{E}\left[G_\mu(x, V_\mu)|V_\mu > b_k - \epsilon\right] + \mathbb{P}(V_\mu \leq b_k - \epsilon)\mathbb{E}\left[G_\mu(x, V_\mu)|V_\mu \leq b_k - \epsilon\right].$$

Note that both conditional expectations are bounded by 1. Because the random variable V_μ (defined under equation (8)) is a truncated normal with mean μ, variance $\sigma^2 + \tau^2$ and truncation set T, it follows that V_μ converges in probability to b_k as μ goes to ∞. Now as x goes to ∞, it follows that $\mu_q(x)$ goes to ∞. This implies that

$$\lim_{x\to\infty} F_{\mu_q(x)}(x) = \lim_{x\to\infty} \mathbb{E}\left[G_{\mu_q(x)}(x, V_{\mu_q(x)})|V_{\mu_q(x)} > b_k - \epsilon\right].$$

Note that $G_\mu(x, v)$ is strictly decreasing in v. This means that $F_\mu(x)$ is bounded from below by $G_\mu(x, b_k)$. But this also means that $\lim_{x\to\infty} F_{\mu_q(x)}(x)$ is bounded from below by $\lim_{x\to\infty} G_{\mu_q(x)}(x, b_k)$. On the other hand, observe that the conditional expectation on the l.h.s. of the previous expression is bounded from above by $G_\mu(x, b_k - \epsilon)$. This means that $\lim_{x\to\infty} F_{\mu_q(x)}(x)$ is bounded from above by $\lim_{x\to\infty} G_{\mu_q(x)}(x, b_k - \epsilon)$. Since $F_{\mu_q(x)}(x)$ is equal to $1 - q$ for all $x \in \mathbb{R}$, it follows that

$$\lim_{x\to\infty} G_{\mu_q(x)}(x, b_k) \leq 1 - q \leq \lim_{x\to\infty} G_{\mu_q(x)}(x, b_k - \epsilon).$$

Because ϵ was arbitrary and $G_\mu(x, v)$ is simply a normal c.d.f. (the normal c.d.f. is uniformly continuous), the claim of the lemma follows.

A.3 Auxiliary results

Lemma A.4. *For every $x \in \mathbb{R}$, $F_\mu(x)$ is continuous and strictly decreasing in μ and satisfies*

$$\lim_{\mu\to\infty} F_\mu(x) = \lim_{\mu\to-\infty} 1 - F_\mu(x) = 0.$$

Proof. Continuity is obvious. For monotonicity, it is sufficient to show that $f_\mu(x)$ has monotone likelihood ratio because Lee et al. (2016) already showed that monotone likelihood ratio implies monotonicity. This means for $\mu_1 < \mu_2$, we need to show that $f_{\mu_2}(x)/f_{\mu_1}(x)$ is strictly increasing in x. In view of the definition of $f_\mu(x)$ in (11), it is easy to see that $f_{\mu_2}(x)/f_{\mu_1}(x)$ can be written as $c\exp((\mu_2-\mu_1)x/\sigma)$, where c is a positive constant that does not depend on x. Because $\mu_1 < \mu_2$ and the exponential function is strictly increasing, it follows that $f_\mu(x)$ has monotone likelihood ratio. Finally, we show that $\lim_{\mu\to\infty} F_\mu(x) = 0$. The other part of this equation follows by similar arguments. Let $M < b_k$. Recall the definition of $F_\mu(x)$ in (8). By the law of total probability, we can write $F_\mu(x)$ as

$$\mathbb{P}(V_\mu > M)\mathbb{E}\left[G_\mu(x, V_\mu)|V_\mu > M\right] + \mathbb{P}(V_\mu \le M)\mathbb{E}\left[G_\mu(x, V_\mu)|V_\mu \le M\right].$$

Note that both conditional expectations are bounded by 1. Because the random variable V_μ (defined under equation (8)) is a truncated normal with mean μ, variance $\sigma^2 + \tau^2$ and truncation set T, it follows that $\lim_{\mu\to\infty} \mathbb{P}(V_\mu > M) = 1 - \lim_{\mu\to\infty} \mathbb{P}(V_\mu \le M) = 1$. This implies that

$$\lim_{\mu\to\infty} F_\mu(x) = \lim_{\mu\to\infty} \mathbb{E}\left[G_\mu(x, V_\mu)|V_\mu > M\right].$$

Because $G_\mu(x, v)$ is strictly decreasing in v, it follows that the conditional expectation on the r.h.s. is bounded from above by $G_\mu(x, M)$. But this means that $\lim_{\mu\to\infty} F_\mu(x)$ is bounded by $\lim_{\mu\to\infty} G_\mu(x, M)$. Because latter limit is equal to 0, the same follows for the former limit. □

This lemma ensures that the function $\mu_q(x)$ is well defined, continuous, strictly increasing in x and q and that $\lim_{x\to\infty} \mu_q(x) = \lim_{x\to-\infty} -\mu_q(x) = \infty$ (see Lemma A.3 in Kivaranovic and Leeb (2020)).

Lemma A.5. *Let the function $h(v)$ be defined as in Lemma A.2. For all $x \in \mathbb{R}$ and all $\mu \in \mathbb{R}$, we have*

$$\frac{\partial f_\mu(x)}{\partial \mu} = \frac{x-\mu}{\sigma^2} f_\mu(x) + f_\mu(x) \sum_{i=1}^{k} h(b_i) - h(a_i).$$

Proof. In view of the definition of $f_\mu(x)$ in (11), the chain rule and product rule imply that the derivative of $f_\mu(x)$ with respect to μ is equal to

$$\begin{aligned}
&\frac{x-\mu}{\sigma^2}\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\frac{\sum_{i=1}^{k}\Phi\left(\frac{b_i-x}{\tau}\right)-\Phi\left(\frac{a_i-x}{\tau}\right)}{\sum_{i=1}^{k}\Phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)-\Phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)} \\
+&\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\frac{\left(\sum_{i=1}^{k}\Phi\left(\frac{b_i-x}{\tau}\right)-\Phi\left(\frac{a_i-x}{\tau}\right)\right)\left(\frac{1}{\sqrt{\sigma^2+\tau^2}}\sum_{i=1}^{k}\phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)-\phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)\right)}{\left(\sum_{i=1}^{k}\Phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)-\Phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)\right)^2}.
\end{aligned}$$

In view of the definition of $f_\mu(x)$, it is easy to see that the first summand is equal to

$$\frac{x-\mu}{\sigma^2} f_\mu(x)$$

and, in view of the definition of $f_\mu(x)$ and $h(v)$, that the second summand is equal to

$$f_\mu(x) \sum_{i=1}^{k} h(b_i) - h(a_i).$$

□

Lemma A.6. *Let $G_\mu(x,v)$ be defined as in (7) and $h(v)$ as in Lemma A.2. For all $x \in \mathbb{R}$ and all $\mu \in \mathbb{R}$, we have*

$$\int_{-\infty}^{x} \frac{u-\mu}{\sigma^2} f_\mu(u)du \;=\; -f_\mu(x) - \sum_{i=1}^{k} h(b_i)G_\mu(x,b_i) - h(a_i)G_\mu(x,a_i).$$

Proof. By definition of $f_\mu(x)$ in (11), the integral can be written as

$$\frac{\sum_{i=1}^{k}\int_{-\infty}^{x}\frac{u-\mu}{\sigma^3}\phi\left(\frac{u-\mu}{\sigma}\right)\Phi\left(\frac{b_i-u}{\tau}\right)du - \int_{-\infty}^{x}\frac{u-\mu}{\sigma^3}\phi\left(\frac{u-\mu}{\sigma}\right)\Phi\left(\frac{a_i-u}{\tau}\right)du}{\sum_{i=1}^{k}\Phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)-\Phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)}.$$

Note that all integrands in the numerator are of the same form. They only differ in the constants $a_1, b_1, \ldots, a_k, b_k$. This means, we can apply Equation 10,011.1 of Owen

(1980) to each integral. This equation implies that the numerator is equal to

$$\sum_{i=1}^{k} \frac{-1}{\sqrt{\sigma^2+\tau^2}}\phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)\Phi\left(\frac{x-(\rho^2\mu+(1-\rho^2)b_i)}{\sigma\rho}\right)-\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\frac{b_i-x}{\tau}\right)$$
$$-\left(\frac{-1}{\sqrt{\sigma^2+\tau^2}}\phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)\Phi\left(\frac{x-(\rho^2\mu+(1-\rho^2)a_i)}{\sigma\rho}\right)-\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\frac{a_i-x}{\tau}\right)\right).$$

(We note that this equation can easily be verified by differentiation of the antiderivative.) In view of the definitions of $f_\mu(x)$, $h(v)$ and $G_\mu(x,v)$, we can see that the claim of the lemma is true. □

Lemma A.7. *For all $x \in \mathbb{R}$ and all $\mu \in \mathbb{R}$, we have*

$$\frac{\partial f_\mu(x)}{\partial x} = -\frac{x-\mu}{\sigma^2}f_\mu(x) - f_\mu(x)\sum_{i=1}^{k} l(x,b_i) - l(x,a_i),$$

where

$$l(x,v) = \frac{\frac{1}{\tau}\phi\left(\frac{v-x}{\tau}\right)}{\sum_{i=1}^{k}\Phi\left(\frac{b_i-x}{\tau}\right)-\Phi\left(\frac{a_i-x}{\tau}\right)}.$$

Proof. In view of the definition of $f_\mu(x)$ in (11), the chain rule and product rule imply that the derivative of $f_\mu(x)$ with respect to x is equal to

$$-\frac{x-\mu}{\sigma^2}\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\frac{\sum_{i=1}^{k}\Phi\left(\frac{b_i-x}{\tau}\right)-\Phi\left(\frac{a_i-x}{\tau}\right)}{\sum_{i=1}^{k}\Phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)-\Phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)}$$
$$-\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\frac{\frac{1}{\tau}\sum_{i=1}^{k}\phi\left(\frac{b_i-x}{\tau}\right)-\phi\left(\frac{a_i-x}{\tau}\right)}{\sum_{i=1}^{k}\Phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)-\Phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)}.$$

It is easy to see that the first summand is equal to

$$-\frac{x-\mu}{\sigma^2}f_\mu(x)$$

and, in view of the definitions of $f_\mu(x)$ and $l(x,v)$, that the second summand is equal to

$$-f_\mu(x)\sum_{i=1}^{k} l(x,b_i) - l(x,a_i).$$

□

Lemma A.8. *Let $h(v)$ be defined as in Lemma A.2 and $l(x,v)$ as in Lemma A.7. For all $x \in \mathbb{R}$ and all $\mu \in \mathbb{R}$, we have*

$$\frac{\partial G_\mu(x,v)}{\partial x} = \frac{f_\mu(x) l(x,v)}{h(v)}.$$

Proof. By definition of $G_\mu(x,v)$ in (7), we have

$$\frac{\partial G_\mu(x,v)}{\partial x} = \frac{1}{\sigma\rho}\phi\left(\frac{x-(\rho^2\mu+(1-\rho^2)v)}{\sigma\rho}\right).$$

We claim that

$$\frac{1}{\sigma\rho}\phi\left(\frac{x-(\rho^2\mu+(1-\rho^2)v)}{\sigma\rho}\right) = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\tau}\phi\left(\frac{v-x}{\tau}\right)}{\frac{1}{\sqrt{\sigma^2+\tau^2}}\phi\left(\frac{v-\mu}{\sqrt{\sigma^2+\tau^2}}\right)}.$$

To see this note that we can write the l.h.s. as $c_1\exp(-d_1/2)$ and r.h.s. as $c_2\exp(-d_2/2)$, where

$$c_1 = \frac{1}{\sigma\rho\sqrt{2\pi}} = \frac{\sqrt{\sigma^2+\tau^2}}{\sigma\tau\sqrt{2\pi}} = c_2$$

and

$$\begin{aligned} d_1 &= \left(\frac{x-(\rho^2\mu+(1-\rho^2)v)}{\sigma\rho}\right)^2 \\ &= \frac{x^2-2\rho^2\mu x-2(1-\rho^2)vx+\rho^4\mu^2+(1-\rho^2)^2v^2+2\rho^2(1-\rho^2)\mu v}{\sigma^2\rho^2} \\ &= \frac{\rho^2(x-\mu)^2+(1-\rho^2)(v-x)^2-\rho^2(1-\rho^2)(v-\mu)^2}{\sigma^2\rho^2} \\ &= \left(\frac{x-\mu}{\sigma}\right)^2+\left(\frac{v-x}{\tau}\right)^2-\left(\frac{v-\mu}{\sqrt{\sigma^2+\tau^2}}\right)^2 = d_2. \end{aligned}$$

Hence the claimed equation is true. Observe that the r.h.s. of this equation can be written as

$$\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\frac{\sum_{i=1}^k \Phi\left(\frac{b_i-x}{\tau}\right)-\Phi\left(\frac{a_i-x}{\tau}\right)}{\sum_{i=1}^k \Phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)-\Phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)}\frac{\frac{1}{\tau}\phi\left(\frac{v-x}{\tau}\right)}{\sum_{i=1}^k \Phi\left(\frac{b_i-x}{\tau}\right)-\Phi\left(\frac{a_i-x}{\tau}\right)}\frac{\sum_{i=1}^k \Phi\left(\frac{b_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)-\Phi\left(\frac{a_i-\mu}{\sqrt{\sigma^2+\tau^2}}\right)}{\frac{1}{\sqrt{\sigma^2+\tau^2}}\phi\left(\frac{v-\mu}{\sqrt{\sigma^2+\tau^2}}\right)}.$$

In view of the definitions of $f_\mu(x)$, $h(v)$ and $l(x,v)$, it is easy to see that the previous expression is equal to $f_\mu(x)l(v)/h(x,v)$. Hence the derivative of $G_\mu(x,v)$ with respect to x is of the claimed form. □

References

Bachoc, F., Leeb, H., and Pötscher, B. M. (2019). Valid confidence intervals for post-model-selection predictors. *Annals of Statistics*, 47:1475–1504.

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2020). Uniformly valid confidence intervals post-model-selection. *Annals of Statistics*, 48:440–463.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41:802–837.

Fithian, W., Taylor, J., Tibshirani, R., and Tibshirani, R. J. (2015). Selective sequential model selection. *arXiv preprint arXiv:1512.02565*.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135.

Heller, R., Meir, A., and Chatterjee, N. (2019). Post-selection estimation and testing following aggregate association tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:547–573.

Kivaranovic, D. and Leeb, H. (2020). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association*, forthcoming.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2018a). A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. *arXiv preprint arXiv:1802.05801*.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2018b). Valid post-selection inference in assumption-lean linear regression. *arXiv preprint arXiv:1806.04119*.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44:907–927.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34:2554–2591.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24:338–376.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Markovic, J., Xia, L., and Taylor, J. (2018). Unifying approach to selective inference with applications to cross-validation. *arXiv preprint arXiv:1703.06559*.

Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9:389–419.

Panigrahi, S. and Taylor, J. (2019). Approximate selective inference via maximum-likelihood. *arXiv preprint arXiv:1902.07884*.

Panigrahi, S., Zhu, J., and Sabatti, C. (2018). Selection-adjusted inference: an application to confidence intervals for *cis*-eQTL effect sizes. *arXiv preprint arXiv:1801.08686*.

Reid, S., Taylor, J., and Tibshirani, R. (2017). Post-selection point and interval estimation of signal sizes in gaussian samples. *Canadian Journal of Statistics*, 45:128–148.

Reid, S., Taylor, J., and Tibshirani, R. (2018). A general framework for estimation and inference from clusters of features. *Journal of the American Statistical Association*, 113:280–293.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for l_1-penalized likelihood models. *Canadian Journal of Statistics*, 46:41–61.

Tian, X., Loftus, J. R., and Taylor, J. E. (2018a). Selective inference with unknown variance via the square-root lasso. *Biometrika*, 105:755–768.

Tian, X., Panograhi, S. a. M. J., Bi, N., and Taylor, J. (2016). Selective sampling after solving a convex problem. *arXiv preprint arXiv:1609.05609*.

Tian, X. and Taylor, J. (2017). Asymptotics of selective inference. *Scandinavian Journal of Statistics*, 44:480–499.

Tian, X., Taylor, J., et al. (2018b). Selective inference with a randomized response. *The Annals of Statistics*, 46:679–710.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111:600–620.

3 Adaptive, Distribution-Free Prediction Intervals for Deep Neural Networks

Adaptive, Distribution-Free Prediction Intervals for Deep Neural Networks

Danijel Kivaranovic[a] Kory D. Johnson[b] Hannes Leeb[a,c]

[a]Department of Statistics and Operations Research, University of Vienna

[b]Institute for Statistics and Mathematics,
Vienna University of Economics and Business

[c]Data Science @ Uni Vienna, University of Vienna

Abstract

The machine learning literature contains several constructions for prediction intervals that are intuitively reasonable but ultimately ad-hoc in that they do not come with provable performance guarantees. We present methods from the statistics literature that can be used efficiently with neural networks under minimal assumptions with guaranteed performance. We propose a neural network that outputs three values instead of a single point estimate and optimizes a loss function motivated by the standard quantile regression loss. We provide two prediction interval methods with finite sample coverage guarantees solely under the assumption that the observations are independent and identically distributed. The first method leverages the conformal inference framework and provides average coverage. The second method provides a new, stronger guarantee by conditioning on the observed data. Lastly, our loss function does not compromise the predictive accuracy of the network like other prediction interval methods. We demonstrate the ease of use of our procedures as well as its improvements over other methods on both simulated and real data. As most deep networks

can easily be modified by our method to output predictions with valid prediction intervals, its use should become standard practice, much like reporting standard errors along with mean estimates.

2

1 Introduction

Deep neural networks have gained tremendous popularity in the last decade due to their superior predictive performance over other machine learning algorithms. They have become the state-of-the-art algorithm in many challenging tasks such as computer vision (Krizhevsky et al., 2012; Karpathy et al., 2014), speech recognition (Hinton et al., 2012), natural language processing (Collobert and Weston, 2008), and bioinformatics (Alipanahi et al., 2015). Despite these successes, there is a paucity of research on the uncertainty of neural network predictions on new samples.

The development of accurate prediction intervals (PIs) for neural networks is a challenging task that is only beginning to gain research interest. Several authors have provided motivation for modified loss functions intended to encourage desirable properties (Nix and Weigend, 1994; Heskes, 1997; Khosravi et al., 2011; Lakshminarayanan et al., 2017; Pearce et al., 2018; Keren et al., 2018; Tagasovska and Lopez-Paz, 2018; Kuleshov et al., 2018). That being said, some provide a PI without a point estimate (Pearce et al., 2018; Keren et al., 2018) or use loss functions which cannot be optimized with stochastic gradient descent (Khosravi et al., 2011). Others have distributional assumptions (Nix and Weigend, 1994; Lakshminarayanan et al., 2017) or provide intervals in which the lower bound is not guaranteed to be smaller than the upper bound (Tagasovska and Lopez-Paz, 2018). These methods also do not provide rigorous guarantees that are desired for a PI. Uncertainty estimation of neural networks from Bayesian perspective has been analyzed by (MacKay, 1992; Gal and Ghahramani, 2016).

The statistics literature is full of confidence interval constructions with provable performance guarantees; however, these methods often cannot be used efficiently with modern learning machines. We use variations of well-known statistical techniques to propose two neural-network-based PIs with valid coverage claims that outperform existing methods. The core of both procedures is a neural network that outputs three values and that optimizes a modified quantile regression loss function. The first method is based on the conformal method introduced by Vovk et al. (2005). Papadopoulos and

Haralambous (2011) applied standard conformal inference to neural networks to construct prediction intervals with finite sample coverage guarantee. These intervals are not specific to neural networks and do not leverage their ability to adapt to a wider class of distributions. We propose a new conformity score that is tailored to neural networks that fully exploits their flexibility. A related conformity score was simultaneously proposed by Romano et al. (2019).

In practice, it is often desirable to have a stronger coverage guarantee than average coverage. We propose a second neural-network-based PI with an approximate conditional coverage claim, where we condition on the observed data. Both proposed procedures use a sample splitting strategy, where one part of the data is used to fit a network that outputs predictions and intervals while the other part is used to adjust these intervals to be valid PIs.

In Section 1.1, we formally set up the prediction problem and explain in detail what we mean by valid PIs. This includes standard concepts such as average coverage but also the new approximate conditional coverage claim which we term Probably-Approximately-Valid (PAV). Section 2 introduces neural-network-based PIs and the algorithms to produce valid prediction intervals. Extensive simulations and real-data examples in Section 3 highlight the improvements our methods make over competing algorithms: Other methods either fail to provide adequate coverage, compromise predictive accuracy, or fail to account for heteroskedasticity.

1.1 Prediction Intervals

We consider a non-parametric regression setting, where X denotes the $\mathbb{R}^d$-valued covariate vector and Y the $\mathbb{R}$-valued response. The data set $D = (X_i, Y_i)_{1 \leq i \leq n}$ contains n i.i.d. copies of the random variable (X, Y). Throughout the paper, we will write (X, Y) as new random variables that are independent of our data D.

In general, a prediction interval $\Gamma_\alpha(X) = \Gamma_{D,\alpha}(X)$ is an interval-valued function of X, the data D, and the confidence level α, such that, loosely speaking, Y falls within

the interval with probability at least $1 - \alpha$. This loose definition can be made precise in various ways.

Definition 1. *The prediction interval* $\Gamma_\alpha(X)$ *controls average coverage if*

$$\mathbb{P}(Y \in \Gamma_\alpha(X)) \geq 1 - \alpha. \tag{1}$$

Note that the probability in equation (1) is over all of the random variables included: the new observation (X, Y) *and the data* D. In Section 2.1, we propose a neural-network-based inference procedure that controls average coverage.

One may object to average coverage because one would like a coverage guarantee given a particular data set instead of averaging over all potential data sets. Our second prediction interval method takes steps toward alleviating this concern. We define a notion of prediction interval validity called Probably Approximately Valid (PAV), which is inspired by the theory of probably approximately correct (PAC) learning (Valiant, 1984). A task is PAC-learnable if, loosely speaking, regardless of the data generating distribution, one can approximate the task arbitrarily well with high probability given sufficient data. Similarly, a PAV interval provides a conditional coverage guarantee with high probability regardless of the data generating distribution. To this end, let $p(\Gamma_\alpha|D)$ denote the conditional probability that $\Gamma_\alpha(X)$ contains Y conditional on D, i.e.,

$$p(\Gamma_\alpha|D) = \mathbb{P}(Y \in \Gamma_\alpha(X)|D).$$

Definition 2. *The prediction interval* $\Gamma_\alpha(X)$ *is* Probably Approximately Valid (PAV) *if for all* $\epsilon > 0$*, all* $\delta > 0$ *and all* $n \geq n_0(\epsilon, \delta)$*,*

$$\mathbb{P}\left(p(\Gamma_\alpha|D) \geq 1 - \alpha - \epsilon\right) \geq 1 - \delta. \tag{2}$$

This means, $\Gamma_\alpha(X)$ is PAV if the conditional probability of $\Gamma_\alpha(X)$ covering Y given D is at least $1 - \alpha - \epsilon$ for all but $\delta 100\%$ of data sets D. In Section 2.2, we provide a greedy algorithm that selects a neural-network-based PI that is PAV such that $n_0(\epsilon, \delta)$ is of order $-\log(\delta)/\epsilon^2$. Corollary 1 shows that PAV prediction intervals can also control average coverage with proper choice of α, ϵ, and δ.

Both proposed PIs are the result of a prediction-interval-specific neural network that outputs a three-dimensional vector and optimizes a loss function used in quantile regression. The next subsection describes and motivates this loss function.

1.2 The PI-specific loss function

Let $N : \mathbb{R}^d \to \mathbb{R}^3$ be a network such that $N(x) = (l(x), m(x), u(x))$, where $l, m, u : \mathbb{R}^d \to \mathbb{R}$ with the restriction that $l(x) \leq m(x) \leq u(x)$ for all $x \in \mathbb{R}^d$. We use $m(x)$ to estimate the median of Y given X, while $l(x)$ and $u(x)$ estimate the lower and upper bounds of our PI, respectively. The monotonicity, $l(x) \leq m(x) \leq u(x)$, is easily enforced by modifying any network that outputs a triple (z_1, z_2, z_3) to output (z_1', z_2', z_3') given by $z_1' = z_1$, $z_2' = z_1' + \text{ReLU}(z_2 - z_1')$, and $z_3' = z_2' + \text{ReLU}(z_3 - z_2')$. Here, $\text{ReLU}(\cdot)$ is the rectified linear unit, $\max(0, \cdot)$.

For $\tau \in [0, 1]$ and $u \in \mathbb{R}$, let $h_\tau(u) = (\tau - \mathbb{1}_{u \leq 0})u$ be the standard loss function to estimate the τth quantile. We define the level-τ loss function evaluated on N at (x, y) by

$$\begin{aligned} \mathcal{L}_\tau(N(x), y) = h_{\tau/2}(y - l(x)) + h_{1/2}(y - m(x)) \\ + h_{1-\tau/2}(y - u(x)). \end{aligned} \tag{3}$$

Letting $|D'|$ denote the cardinality of a set D', the empirical risk of network N on $D' \subseteq D$ is

$$\mathcal{R}_{D',\tau}(N) = \frac{1}{|D'|} \sum_{(X_i, Y_i) \in D'} \mathcal{L}_\tau(N(X_i), Y_i). \tag{4}$$

Definition 3. *Denote the neural network $N_{D',\tau}(x) = (l_{D',\tau}(x), m_{D',\tau}(x), u_{D',\tau}(x))$ to be one that is fit on $D' \subseteq D$ by minimizing the empirical risk $\mathcal{R}_{D',\tau}(N)$. In the trivial case where $\tau = 0$, we set $l_0(x) = -\infty$ and $u_0(x) = \infty$ for all $x \in \mathbb{R}^d$.*

By the strong law of large numbers, the empirical risk $\mathcal{R}_{D,\tau}(N)$ converges almost surely to $\mathbb{E}[\mathcal{L}_\tau(N(X), Y)]$ as $|D| \to \infty$. Let $q_\tau(x) = \inf\{y : \mathbb{P}(Y \leq y \mid X = x) \geq \tau\}$ be the conditional τ-quantile of Y given $X = x$, e.g., $q_{1/2}(x)$ is the conditional

median of Y given $X = x$. Following standard texts (Koenker, 2005), the triple $(q_{\tau/2}(x), q_{1/2}(x), q_{1-\tau/2}(x))$ is the minimizer of $\mathbb{E}[\mathcal{L}_\tau(N(X), Y)]$. For a given confidence level $\alpha \in (0, 1)$, $(q_{\alpha/2}(x), q_{1/2}(x), q_{1-\alpha/2}(x))$ has the desirable properties that

$$q_{1/2}(\cdot) \;=\; \underset{f:\mathbb{R}^d \to \mathbb{R}}{\arg\min}\, \mathbb{E}|Y - f(X)| \tag{5}$$

and

$$1 - \alpha \;=\; \mathbb{P}(q_{\alpha/2}(X) \leq Y \leq q_{1-\alpha/2}(X)) \mid X) \tag{6}$$

under minimal assumptions on the joint distribution of (X, Y). If the problem is constrained to linear regression or M-estimation, then the estimators resulting from empirical risk minimization are consistent (Wooldridge, 2001); however, under our minimal assumptions, it is not known whether $N_{D,\tau}(x)$ consistently estimates $(q_{\alpha/2}(x), q_{1/2}(x), q_{1-\alpha}(x))$. As the network $N_{D,\alpha}(x)$ does not generally provide the desired properties in finite samples, Section 2 provides two modifications with finite sample coverage guarantees based on sample splitting.

1.3 Comments on the network architecture

In the definition of the network $N_{D,\tau}(x)$, we only specified the output layer of the neural network and the loss function it is minimizing. The remaining network architecture can be chosen by the users in order to achieve their predictive goals. The term architecture comprises the network design (e.g. number and depth of hidden layers or dropout layers) and training parameters (e.g. number of epochs, batch size, or regularization parameters). See Goodfellow et al. (2016) and citations therein for an overview on network architecture, regularization methods, and optimization algorithms. In fact, one can use a pre-trained neural network in combination with our proposed output layer and loss function to fit a neural network with accurate predictive performance and valid prediction intervals. See Section 3 for an application to a real image dataset.

We advocate the use of a single network that outputs both the prediction and the prediction interval to ensure that the prediction is always contained in the interval.

Crossing quantiles is a well-known problem in the literature (see He (1997)) that can be easily circumvented in this way without any loss in predictive accuracy (see Section 3).

In fact, even our proposed loss function can be modified to the users needs. For example, the midpoint estimate is not restricted to be the median; one can also use the mean squared error instead of the absolute error in the loss function in order to estimate the conditional mean. If the underlying distribution is highly skewed, however, the conditional mean is not guaranteed to be lie between $q_{\alpha/2}(x)$ and $q_{1-\alpha/2}(x)$. We note that the mean squared error of the point estimate can be written as $h_{1/2}(y - m(x))^2$, so that squaring the other terms of equation (3) puts them on the same scale.

The supplemental materials demonstrate the validity of our methods over various architectures and loss functions. As this is guaranteed by our theorems, these examples are not included in the paper.

2 Construction of valid PIs with neural networks

Let D_1 and D_2 be an arbitrary partition of D into two disjoint sets. D_1 is used to select and train the network and D_2 is used to adjust the resulting intervals to provide coverage guarantees. We emphasize that all results are still valid for any data-dependent architecture, as long as the dependency is only on D_1. As all networks in this paper are fit using D_1, we simplify our notation, setting $N_\tau(x) = N_{D_1,\tau}(x)$ and analogously for $l_\tau(x)$, $m_\tau(x)$ and $u_\tau(x)$.

2.1 Average Coverage

To achieve average coverage, we use methods derived from conformal inference (Vovk et al., 2005; Lei et al., 2018). In general, conformal prediction intervals require a fixed prediction procedure that is refit on an augmented data set. As it would clearly be infeasible to refit a large network many times, this process can be simplified by using

Algorithm 1 Split Conformal Prediction Intervals

Input: Holdout data D_2, network $N_\tau(x)$.

Output: Scaled network $N_\tau^{\hat{c}}(x)$ with parameter $\hat{c}$.

Set: $c_i = \max\left(\frac{m_\tau(X_i)-Y_i}{m_\tau(X_i)-l_\tau(X_i)}, \frac{Y_i-m_\tau(X_i)}{u_\tau(X_i)-m_\tau(X_i)}\right)$ for all $(X_i, Y_i) \in D_2$.

Set: $\hat{c} = c_{(k)}$, $k = \lceil (1-\alpha)(|D_2|+1) \rceil$ and $c_{(k)}$ the kth order statistic.

Set: $l_\tau^{\hat{c}}(x) = m_\tau(x) - \hat{c}(m_\tau(x) - l_\tau(x))$ and $u_\tau^{\hat{c}}(x) = m_\tau(x) + \hat{c}(u_\tau(x) - m_\tau(x))$.

Return: $N_\tau^{\hat{c}}(x) = (l_\tau^{\hat{c}}(x), m_\tau(x), u_\tau^{\hat{c}}(x))$.

classical sample slitting. Lei et al. (2018) refer to such methods as split-conformal.

For a given $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$, the interval $[l_\tau(x), u_\tau(x)]$ may not contain y; however, with c given by

$$c = \max\left(\frac{m_\tau(x) - y}{m_\tau(x) - l_\tau(x)}, \frac{y - m_\tau(x)}{u_\tau(x) - m_\tau(x)}\right),$$

y is an endpoint of the interval $\Gamma_\tau^c(x) = [m_\tau(x) - c(m_\tau(x) - l_\tau(x)), m_\tau(x) + c(u_\tau(x) - m_\tau(x))]$, and hence is contained in $\Gamma_\tau^c(x)$. These facts can be used in a conformal inference procedure to calibrate the PI. In essence, a constant $\hat{c} \in (0, \infty)$ is chosen such that at least (1-α)100% of the observations *in the hold-out data*, D_2, are contained in the interval $\Gamma_\tau^{\hat{c}}(X)$. Details are given in Algorithm 1.

Theorem 1. *Let $N_\tau(x) = N_{D_1,\tau}(x)$ be as in Definition 3 and $N_\tau^{\hat{c}}(x)$ be the result of Algorithm 1. Set $\Gamma_\alpha^{\hat{c}}(x) = [l_\tau^{\hat{c}}(x), u_\tau^{\hat{c}}(x)]$. Then $\mathbb{P}\left(Y \in \Gamma_\alpha^{\hat{c}}(X)\right) \geq 1 - \alpha$.*

Note that Theorem 1 also holds for any data-dependent $\hat{\tau}$ as long as the dependence of $\hat{\tau}$ on the data is only through D_1. As τ controls the width of the estimated interval and that is precisely what $\hat{c}$ is selected to calibrate, we suggest setting $\tau = \alpha$ in practice. In the simulations of Section 3, we typically observe $\hat{c} \geq 1$ for $N_\alpha(x)$, but $\hat{c} \to 1$ as $|D_1|$ increases, suggesting that $N_\alpha(x)$ has asymptotic $1 - \alpha$ average coverage.

Proof of Theorem 1. Because the network $N_\tau(x)$ is fit on D_1 which is independent of

D_2, the statistics c_i are i.i.d. conditional on D_1. For the new observation (X, Y), set

$$c' = \max\left(\frac{m_\tau(X) - Y}{m_\tau(X) - l_\tau(X)}, \frac{Y - m_\tau(X)}{u_\tau(X) - m_\tau(X)}\right).$$

Conditional on D_1, c' is independent of the c_is, has the same distribution as the c_is, and the rank of c' among the c_is is uniform over the set of integers $\{1, 2, \ldots, |D_2| + 1\}$. Therefore, $\mathbb{P}(c' > c_{(k)} \mid D_1) \leq \alpha$, where $k = \lceil (1-\alpha)(|D_2|+1) \rceil$ and $c_{(k)}$ is the kth order statistic of the $c_1, \ldots, c_n$. But this implies $\mathbb{P}(Y \in \hat{\Gamma}^c_\alpha(X) \mid D_1) \geq 1 - \alpha$. □

If ties among c_is only happen on a set of measure 0, or if we use a random tie-breaking rule, the probability of coverage can be upper bounded by $1-\alpha+1/(|D_1|+1)$, meaning that the intervals are not unnecessarily wide (cf. Theorem 2.2 of Lei et al. (2018)). Algorithm 1 is a modified version of the Split-Conformal algorithm of Lei et al. (2018). Our statistics c_i are tailored to our neural network which outputs an ordered triple, whereas previous algorithms use the residuals $R_i = |Y_i - m(X_i)|$ or standardized variants.

2.2 PAV

In order to provide PAV intervals, we select $\hat{\tau}$ using D_2 so that $l_{\hat{\tau}}(x)$ and $u_{\hat{\tau}}(x)$ adaptively estimate the quantiles $q_{\alpha/2}(x)$ and $q_{1-\alpha/2}$. This adaptation is required as we are not guaranteed that $l_\alpha(x)$ and $u_\alpha(x)$ accurately estimate these quantities. Let $\hat{p}(N_\tau, D_2)$ be the empirical coverage probability of the neural network N_τ on the data set D_2, i.e.,

$$\hat{p}(N_\tau, D_2) = \frac{1}{|D_2|} \sum_{(X_i, Y_i) \in D_2} \mathbb{1}_{Y_i \in [l_\tau(X_i), u_\tau(X_i)]}.$$

We choose a parameter $\hat{\tau}$ over a grid G_K, where

$$G_K = \{\tau_{(1)}, \ldots, \tau_{(K)})\}, \quad 1 > \tau_{(1)} > \ldots \tau_{(K)} > 0,$$

such that the network $N_{\hat{\tau}}(x)$ has $1 - \alpha$ coverage on D_2. Details are given in Algorithm 2.

Algorithm 2 PAV Prediction Intervals

Input: Holdout data D_2, a grid G_K, Networks $N_\tau(x)$ with $\tau \in G_K$.

Output: Network $N_{\hat{\tau}}(x)$.

Set: $\hat{\tau} = 0$.

For $i = 1$ to K:

If $\hat{p}(N_{\tau(i)}, D_2) \geq 1 - \alpha$**:**

Set: $\hat{\tau} = \tau_{(i)}$**.**

End.

Return: $N_{\hat{\tau}}(x)$**.**

Theorem 2. *Let $N_\tau(x) = N_{D_1,\tau}(x)$ be as in Definition 3 and $N_{\hat{\tau}}(x)$ be the result of Algorithm 2. Set $\Gamma^{\hat{\tau}}_\alpha(x) = [l_{\hat{\tau}}(x), u_{\hat{\tau}}(x)]$. Set $p(\Gamma^{\hat{\tau}}_\alpha|D) = \mathbb{P}(Y \in \Gamma^{\hat{\tau}}_\alpha(X) \mid D)$ and $n_2 = |D_2|$. Then,*

$$\mathbb{P}\left(p(\Gamma^{\hat{\tau}}_\alpha|D) \leq 1 - \alpha - \epsilon\right) \leq K \exp(-2\epsilon^2 n_2).$$

The theorem continues to hold for any data-dependent grid, $\hat{G}_K$, as long as the dependence of $\hat{G}_K$ on the data is only through D_1. Given the tendency of flexible models such as neural networks to over-fit the data, $[l_\alpha(x), u_\alpha(x)]$ typically under-covers in finite samples. As such, we suggest setting $G_K \subseteq [0, \alpha]$, and in the simulations of Section 3 we typically observe $\hat{\tau} \leq \alpha$ but that $\hat{\tau} \to \alpha$ from below as $|D_1|$ increases. Note that by solving the equation $\delta = K \exp(-2\epsilon^2 n_2)$, we get that for all $n_2 \geq n(\epsilon, \delta, K) = -\log(\delta/K)/(2\epsilon^2)$,

$$\mathbb{P}\left(p(\Gamma^{\hat{\tau}}_\alpha|D) \geq 1 - \alpha - \epsilon\right) \geq 1 - \delta.$$

Proof of Theorem 2. By definition of $N_{\hat{\tau}}(x)$, we have

$$\begin{aligned}
&\mathbb{P}\left(p(\Gamma^{\hat{\tau}}_\alpha|D) \leq 1 - \alpha - \epsilon\right) \\
&\quad \leq \mathbb{P}\left(p(\Gamma^{\hat{\tau}}_\alpha|D) \leq \hat{p}(N_{\hat{\tau}}, D_2) - \epsilon\right) \\
&\quad = \mathbb{E}\left[\mathbb{P}\left(p(\Gamma^{\hat{\tau}}_\alpha|D) \leq \hat{p}(N_{\hat{\tau}}, D_2) - \epsilon \mid D_1\right)\right].
\end{aligned}$$

We use the union bound to bound the conditional probability within the expectation from above by

$$\sum_{\tau \in G_K \cup \{0\}} \mathbb{P}\left(p([l_\tau(X), u_\tau(X)] \mid D) \leq \hat{p}(N_\tau, D_2) - \epsilon \mid D_1\right).$$

For $\tau = 0$, we have $[l_\tau(X), u_\tau(X)] = [-\infty, \infty]$ and the corresponding summand in the previous expression is equal to 0. For $\tau \neq 0$, we have $p([l_\tau(X), u_\tau(X)] \mid D) = p([l_\tau(X), u_\tau(X)] \mid D_1)$, because the event in the first conditional probability is independent of D_2. Observe that, conditional on D_1, $\hat{p}(N_\tau, D_2)$ is the mean of n_2 Bernoulli-trails with mean $p([l_\tau(X), u_\tau(X)] \mid D_1)$. By Hoeffding's inequality, the corresponding summand in the previous expression is bounded by $\exp(-2\epsilon^2 n_2)$. □

Even though the PAV interval $\Gamma^{\hat{\tau}}_\alpha(X)$ does not generally provide average coverage, this can be achieved by defining a more conservative PAV interval.

Corollary 1. *Fix $\epsilon > 0$ such that $\alpha - \epsilon > 0$. Let $\Gamma^{\hat{\tau}}_{\alpha-\epsilon}(x)$ be defined as in Theorem 2 with $\alpha - \epsilon$ replacing α. For all*

$$n_2 \geq \frac{-2\log(\epsilon/(2K(1-\alpha+\epsilon/2)))}{\epsilon^2}$$

we have

$$\mathbb{P}\left(Y \in \Gamma^{\hat{\tau}}_{\alpha-\epsilon}(X)\right) \geq 1 - \alpha.$$

To prove the result observe that $\mathbb{P}(Y \in \Gamma^{\hat{\tau}}_{\alpha-\epsilon}(X)) \geq (1 - \alpha + \epsilon/2)(1 - K\exp(-\epsilon^2 n_2/2))$.

2.3 Discussion of contributions

The split-conformal and adaptive selection approaches of the previous subsections admittedly constitute a revival of sample splitting. Practitioners have long used sample-splitting as a valid, tractable method for tuning or testing and can easily implement it. The methodology followed in this paper differs subtly but significantly from the standard use of sample splitting or conformal inference. Typically one creates either

residuals or standardized residuals in order to estimate an additive adjustment factor for creating a PI of the form estimate $\pm$ error. It is clear that such intervals may not be appropriate for skewed or multimodal distributions. Instead of designing more complex standardizations for residuals, this paper leverages the power of NN in order to estimate the quantiles directly, using subsequent adjustments for finite sample guarantees.

We conjecture that the desirable properties in equations (5) and (6) are satisfied in a well-defined asymptotic setting: First, neural networks are universal approximators (cf. Cybenko (1989); Hornik (1991)). Second, given a sufficiently large data set, and under the assumption that we can minimize the empirical risk, a neural network is able to learn the conditional mean of Y given X (cf. Bauer and Kohler (2017)). This conjecture is supported by simulation evidence presented in the supplement where we observe that $N_{D,\tau}(x) \to (q_{\alpha/2}(x), q_{1/2}(x), q_{1-\alpha/2}(x))$ as $|D| \to \infty$.

We propose PIs with average coverage control and an approximate conditional coverage control (PAV). A stronger coverage claim would be to control the conditional coverage probability, conditional on the input X. However, this is only achievable under much more restrictive assumptions. Also, we note that our only (and crucial) assumption is i.i.d.-ness of the data. This means, our PIs are in general not valid for out-of-distribution samples.

3 Simulations

Throughout this section, we refer to $\Gamma^{\hat{c}}_{\alpha}(X)$ from Section 2.1 as *conf-nn* and $\Gamma^{\hat{\tau}}_{\alpha}(X)$ from Section 2.2 as *pav*. We compare our methods to 5 different procedures that use different types of loss functions that are designed to fit neural networks that output PIs. The first (*conf-fw*) is the classical, "fixed-width" conformal method that uses the absolute residuals as conformity score to create valid PIs (Papadopoulos and Haralambous, 2011). The second method (*high-q*) is the so-called "high-quality" driven method of Pearce et al. (2018). This method uses a loss function that penalizes in-sample mis-coverage and interval length. The third method (*neg-ll*) uses the Gaussian negative-log-likelihood

to estimate the mean and variance of the target (Nix and Weigend, 1994). The fourth method is the calibrated regression model of Kuleshov et al. (2018), where the Gaussian negative-log-likelihood is used to estimate the distribution and isotonic regression for calibration. Because of its probabilistic approach, we refer to this method as *bayes*. Finally, we also consider the interval $\Gamma_\alpha(X)$ (*qreg-un*), i.e., the unadjusted interval function of the network $N_\alpha(X)$.

In all data examples, we split the data D into a training set D_1, a validation set D_2, and a test set D_3. D_1 is used to train the network, D_2 is used to calibrate the intervals for *pav*, *conf-nn*, *conf-fw* and *bayes*. Because *neg-ll* and *high-q* were very sensitive to the choice of hyperparameters in our experiments, we used D_1 to train these networks and D_2 for hyperparameter selection. All data experiments set $\alpha = .1$ and calculate results using D_3, which was unseen by the models. The $\hat{\tau}$ in $\Gamma^{\hat{\tau}}_\alpha(X)$ is selected on the grid $G_{10} = \{.1, .09, \ldots, .01\}$ and we set $\tau = \alpha$ when constructing $\Gamma^{\hat{c}}_\alpha(X)$. In all experiments, we used the Adam optimizer with learning rate 0.01 and no learning rate decay. The number of epochs was chosen by cross-validation on D_1. In each data example, the same network architecture is used for all methods and is trained in Python using the PyTorch library (Paszke et al., 2017). These architectures are described along with the data examples. All computations were performed on a Google-Cloud-Platform instance with a NVIDIA Tesla K80 GPU.

3.1 Artificial data

We simulate covariates $X \in [0, 1]^{100}$, where each entry is drawn i.i.d. from a standard uniform distribution. The response is given by

$$Y = f(\beta' X) + \epsilon, \quad \epsilon \sim N(0, 1 + (\beta' X)^2),$$

where $f(x) = 2\sin(\pi x) + \pi x$ and only the first 5 components of $\beta \in \{0, 1\}^{100}$ are non-zero and equal to 1. This is a challenging setting because the model is sparse, non-linear in X, and heteroskedastic in Y given X. We compare all the methods to an oracle that knows this data-generating process and thus the true conditional median, $q_{1/2}(X) =$

$f(\beta' X)$, as well as the uniformly most accurate, unbiased PI, $f(\beta' X) \pm z_{\alpha/2}\sqrt{1+(\beta' X)^2}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. We generate $|D_1 \cup D_2| = 100,000$ and used 3/4 of the data for training and 1/4 for validation. In the supplement, we provide results for data set sizes ranging between 5,000 and 100,000 with comparable results. We repeated the experiment 10 times. For each method, we train a neural network with one hidden layer and 200 hidden nodes using 80-100 passes through the data. An independent data set of size 20,000 is used for testing.

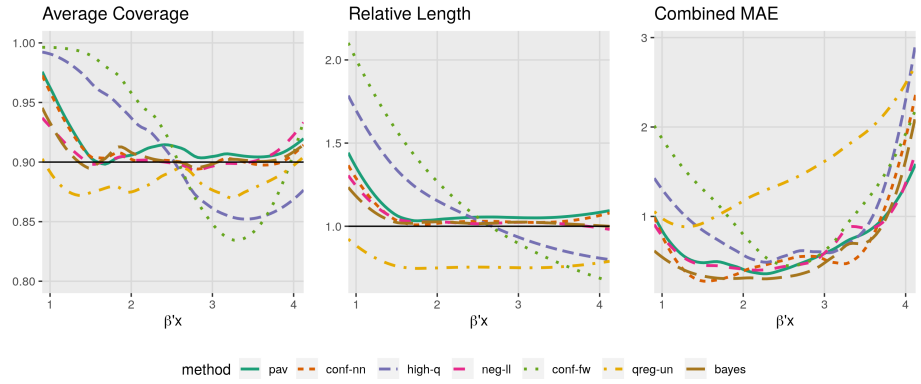

Figure 1: Results on Simulated Data. Each measure is plotted as a function of the true linear component, $\beta' X$. The combined MAE is the mean absolute error for estimating the oracle: $(q_{1/2}(X),\ q_{\alpha/2}(X),\ q_{1-\alpha/2}(X))$.

Simulation results are summarized in Figure 3.1, which shows the empirical coverage, the interval length relative to the oracle, and the mean absolute error (MAE) from estimating $(q_{1/2}(X),\ q_{\alpha/2}(X),\ q_{1-\alpha/2}(X))$. Each measure is plotted as a function of $\beta' X$ in order to observe how various methods adapt to heteroskedasticity. Both *conf-nn* and *pav* provide coverage close to the nominal level throughout the majority of the domain of $\beta' X$. Only for small $\beta' X$, the two methods are too wide. This can be explained by the fact that in this area only a few observations are available (see the data generation process) to estimate the conditional quantiles. This demonstrates that both procedures are accurately estimating the true underlying distribution on a wide

range of the input space. *bayes* and *neg-ll* perform equally well, which is not surprising because both procedures minimize the Gaussian negative-log-likelihood. In terms of length and estimation error, *conf-nn*, *pav*, *bayes* and *neg-ll* all provide quality estimates for a wide range of $\beta' X$.

conf-fw and *high-q* on the other hand over-cover in the left tail and undercover in the right tail as they do not capture the heteroscedasticity of the true distribution. This is observed in their lengths relative to the oracle. We see that *qreg-un* undercovers through the entire input space, which confirms the need for calibration.

3.2 Real Data

In each of the real data examples, the ratio of observations in D_1, D_2 and D_3 was equal to 3:1:1. Each experiment was repeated 20 times, i.e., the data set is randomly partitioned into three parts, then the model is trained on D_1, tuned on D_2, and evaluated on D_3. This process constitutes one repetition. We discuss the results from four real datasets here, about which further information is given in the supplement. When applicable, covariates were standardized to have mean 0 and variance 1.

We consider two datasets from the data science platform Kaggle and two from the UCI machine learning repository (Dua and Graff, 2017). The first is a bike share dataset from UCI that contains the hourly count of bike rentals in the Capital bikeshare system along with weather information between 2011 and 2012 (Fanaee-T and Gama, 2013). The task is to predict the number of bike rentals based on weather information. The other UCI dataset provides hourly traffic volume, westbound on I-94, in Minneapolis-St Paul, Minnesota between 2012 and 2018. The dataset also includes weather and holiday information. For both datasets, we trained a network with one hidden layer and 100 hidden nodes.

From Kaggle, we consider an image data set and a standard regression task. The image dataset contains 12,611 observations, each consisting of an x-ray of a patient's hand. The task is to predict the patient's age using the x-ray. We use an intermediate

layer of the pre-trained Inception V3 network as the feature extractor (Szegedy et al., 2016). We trained a neural network on the extracted features with one hidden layer and 300 hidden nodes. We used data augmentation (random rotation and horizontal flips of the images) on D_1 to reduce over-fitting. The regression dataset consists of 21,613 sale prices for homes in King County, Washington, between May 2014 and May 2015. There are 19 covariates describing the features of the house which are used to predict log sale price. For each method we trained a neural network with one hidden layer, with 100 hidden nodes, and between 80 and 200 passes through the data.

Figure 2 summarizes the results on these data. It shows the empirical coverage, the average length, and the mean absolute error (MAE) of each method for predicting the target. Starting with average coverage, we see that *conf-nn* and *pav* provide average coverage at nearly exact nominal level for all repetitions and all datasets as guaranteed by our theorems. We can see that *pav* is more conservative, due to its stronger guarantee, though potentially also due to the coarse selection grid for τ. The unadjusted method, *qreg-un*, does not provide coverage, which confirms our presumption that the neural network $N_\alpha(X)$ fails to accurately estimate the conditional quantiles in finite samples. This also strongly affirms the necessity to adjust neural network based intervals. *bayes* also provides precise coverage and the remaining methods primarily over-cover, partitulcarly *neg-ll*. We see that *high-q* does not provide consistent coverage on the traffic data and found it to be highly-sensitive to hyperparameters.

Among the valid methods, *conf-nn* and *pav* are always among the shortest intervals. *bayes* performs significantly worse on the Bike Share and traffic datasets, with many repetitions having double or triple the length and error of *conf-nn* and *pav*. While *conf-fw* provides surprisingly short intervals on average, we note that in the presence of heteroscedasticity it can occur that inferior fixed-width intervals are narrower, on average, than the optimal intervals. *high-q* is slightly wider and *neg-ll* is considerably wider on average, also showing they they can be too conservative.

In terms of the estimation error *conf-nn*, *pav* and *conf-fw* have similar performance on all datasets. We emphasize this as *conf-fw* fits a neural network that minimizes

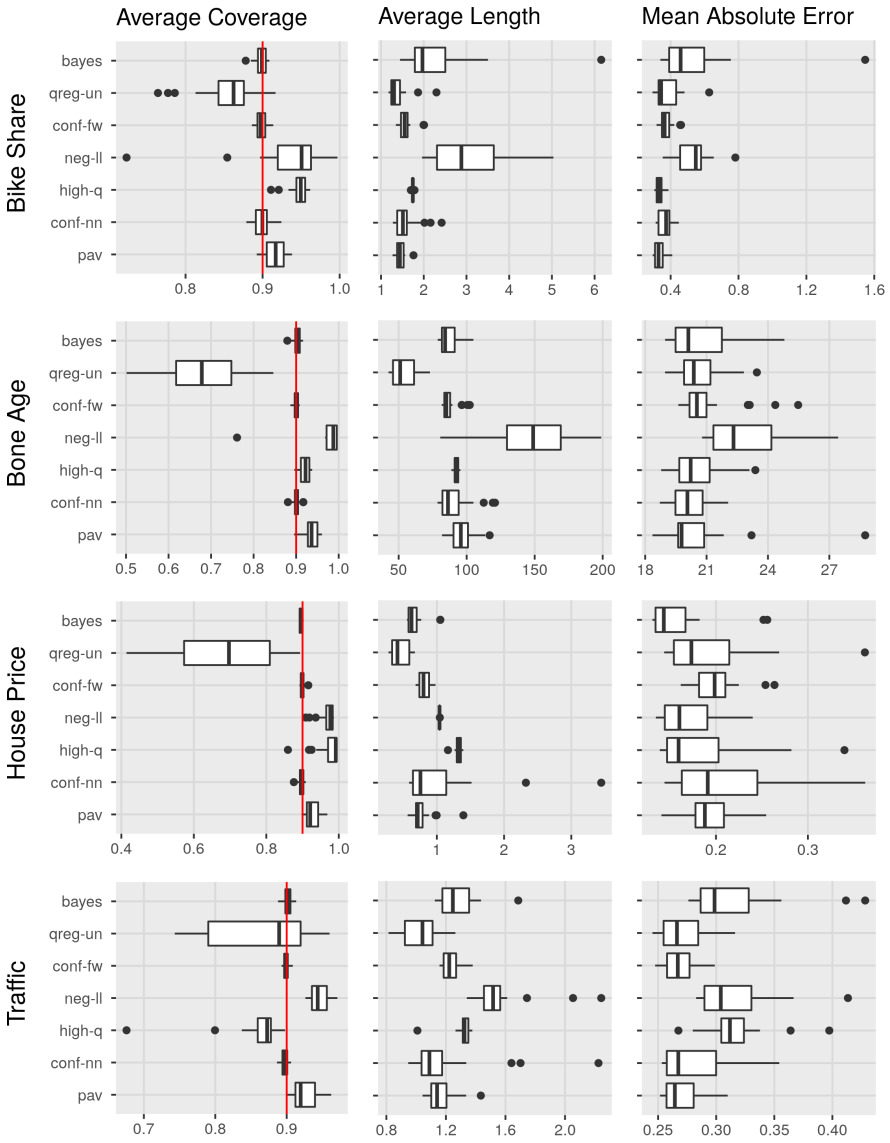

Figure 2: Summary of results on real data. Each row corresponds to a different data set and are organized alphabetically. The red line in the left-most plots denote the nominal confidence level.

mean absolute error directly, while *conf-nn* and *pav* both use the loss function given in (3). This demonstrates that our proposed loss function does not comprise the predictive accuracy of the fitted networks.

4 Conclusion

In this paper we demonstrated how to use standard statistical techniques such as sample splitting and quantile regression to fit a neural network that outputs accurate predictions and valid prediction intervals. We proposed two procedures with provable coverage guarantees that improve current state-of-the-art methods in terms of interval length and predictive accuracy. The ease and transparency of the two procedures advocate their application to many deep neural networks in order to express the uncertainty of their predictions. As data sets grow in size, the cost of using a validation set decreases, making the observation of Barnard (1974) even more accurate today:

> The simple idea of splitting a sample into two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics, if we measure the degree of neglect by the ratio of the number of cases where a method could give help to the number of cases where it is actually used.

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33:831.

Barnard, G. A. (1974). Discussion of "cross-validatory choice and assessment of statistical predictions," by M. Stone. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36:133–135.

Bauer, B. and Kohler, M. (2017). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, Submitted for publication.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2:303–314.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Fanaee-T, H. and Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, pages 1–15.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, 51:186–192.

Heskes, T. (1997). Practical confidence and prediction intervals. In *Advances in neural information processing systems*, pages 176–182.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4:251–257.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Keren, G., Cummins, N., and Schuller, B. (2018). Calibrated prediction intervals for neural network regressors. *IEEE Access*, 6:54033–54041.

Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F. (2011). Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE transactions on neural networks*, 22:337–346.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111.

MacKay, D. J. (1992). A practical bayesian framework for backpropagation networks. *Neural computation*, 4:448–472.

Nix, D. A. and Weigend, A. S. (1994). Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60. IEEE.

Papadopoulos, H. and Haralambous, H. (2011). Reliable prediction intervals with regression neural networks. *Neural Networks*, 24:842–851.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

Pearce, T., Brintrup, A., Zaki, M., and Neely, A. (2018). High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *ICML*.

Romano, Y., Patterson, E., and Candès, E. J. (2019). Conformalized quantile regression. *arXiv preprint arXiv:1905.03222*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Tagasovska, N. and Lopez-Paz, D. (2018). Frequentist uncertainty estimates for deep learning. *CoRR*, abs/1811.00908.

Valiant, L. G. (1984). A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg.

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data*, volume 1 of *MIT Press Books*. The MIT Press.

Supplemental Material: Adaptive, Distribution-Free Prediction Intervals for Deep Neural Networks

Danijel Kivaranovic[a] Kory D. Johnson[b] Hannes Leeb[a,c]

[a]Department of Statistics and Operations Research, University of Vienna

[b]Institute for Statistics and Mathematics,
Vienna University of Economics and Business

[c]Data Science @ Uni Vienna, University of Vienna

In this supplement, we present further results on simulated data. Figure 1 shows how the algorithms behave as a function of the number of observations. We see that our two methods, *pav* and *conf-nn*, control average coverage for all n (left panel). Also both methods approximate the oracle as n grows (middle and right panel). We observed that the intervals of *pav* are more conservative and slightly longer than the intervals of *conf-nn*. The prediction intervals of *neg-ll* are very wide for smaller sample sizes but perform similarly in larger samples. The *bayes* method performs similarly well as our two methods. The *conf-fw* and *high-q* methods also control average coverage, however the intervals are slightly longer (middle panel) compared to the intervals of our two methods. The reason is that both methods do not adapt to the heteroskedasticity of the data. Only *qreg-un* does not control average coverage (left panel). This demonstrates that adjustment of the intervals is necessary to control average coverage. While our methods, *pav* and *conf-nn*, do not assume normality, we achieve results that are close to the optimal prediction intervals of the oracle.

In Figure 1 of the main paper, we show more detailed results for sample size equal

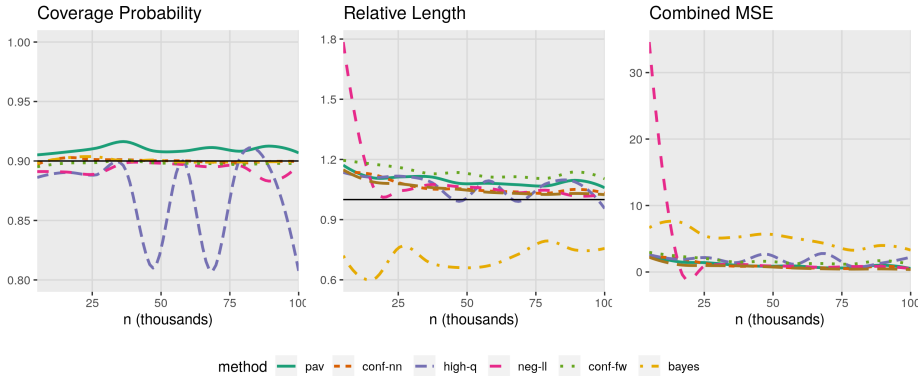

Figure 1: Asymptotic performance of procedures. Both relative length and combined MSE are computed relative to the oracle predictions and prediction intervals.

to 100,000. In Figure 2 we show the same simulation results for sample size equal to 5,000 (top) and 47,222 (bottom). In the left panel, we see that even in small samples the methods *pav* and *conf-nn* provide coverage close to the the nominal level for a wide range of $\beta'X$. In the middle panel, we see that the relative length approaches 1 (i.e. it approaches the length of the oracle) as the sample sizes grows from 5,000 to 47,222. In the right panel, we see that the combined prediction error approaches 0 as the sample size increases, demonstrating that it is accurately estimating the oracle. On the other hand, the methods *conf-fw* and *high-q* undercover when $\beta'X$ is large and overcover when $\beta'X$ is small. The reason for this is that these two methods do not adapt to the heteroskedasticity of the data as seen in the middle and right panel. We see that the method *neg-ll* performs considerably worse when the sample size is small, but is comparable to *pav* and *conf-nn* when the sample size is larger. This suggests that this method needs more data to accurately estimate the distribution of the data. This is surprising, since *neg-ll* assumes that the data is Gaussian, which is the case in our artificial data example. We again see that *qreg-un* is consistently below the nominal level due to overfitting. The *bayes* method performs similary to *pav* and *conf-nn*, but we again note that *bayes* implicitly assumes that the data is Gaussian.

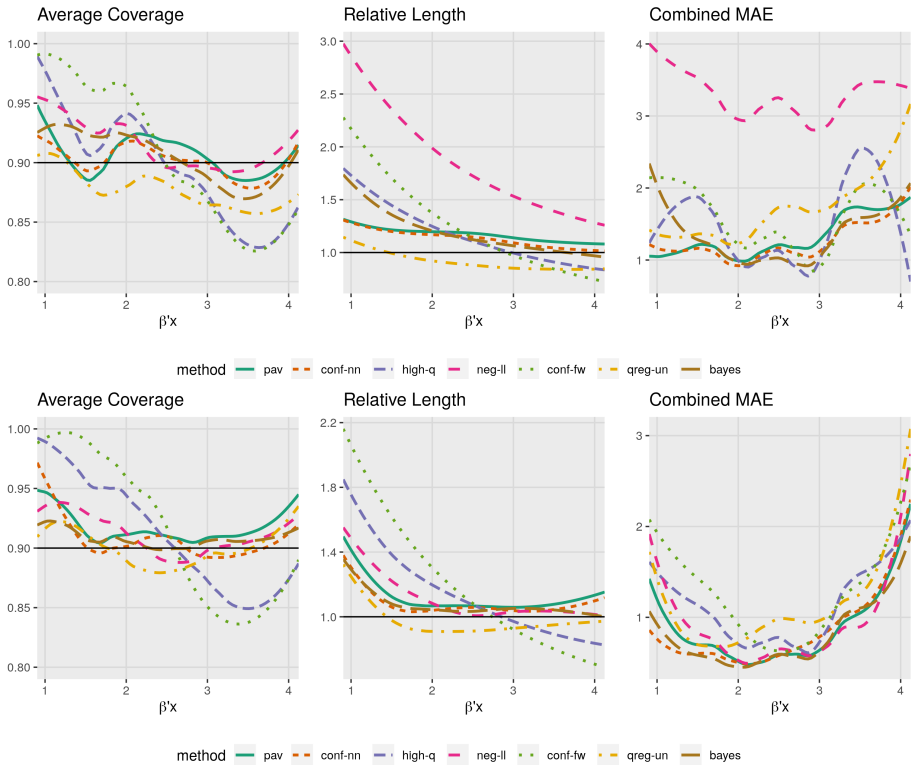

Figure 2: Simulation summaries for $n = 5,000$ (top) and $n = 47,222$ (bottom).

Figure 3 explores the effect of architectures and loss function on the performance of our procedure. We use the same simulated data with sample size fixed to 50,000. For simplicity we only consider the conformal intervals given elsewhere as *conf-nn*. We considered two architectures: a network with one hidden layer (dep1) and a network with two hidden layers (dep2). We also considered two different loss functions: the quantile loss function of the main paper (loss_l1) and the modified version of it where we square each term in the loss function (loss_l2). The squared version is supposed to estimate the conditional mean instead of the conditional median. However, note that in our Gaussian setting the conditional mean is equal to the conditional median.

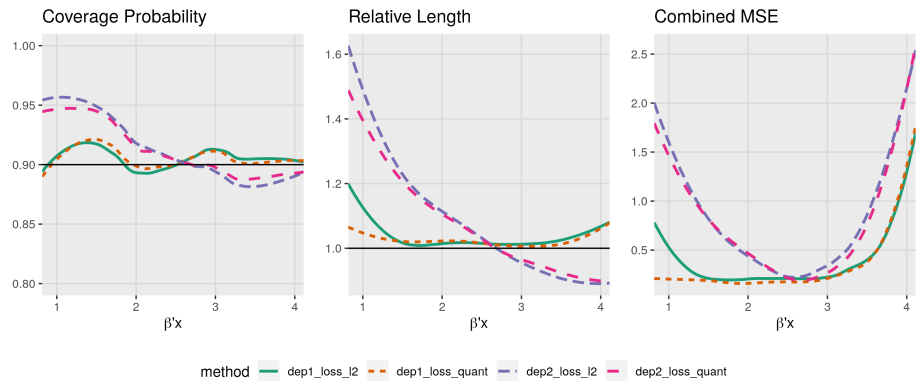

Figure 3: Results for different architectures and loss functions. The networks with one and two hidden layers are denoted by dep1 and dep2, respectively. The loss functions with the $l1$-type and $l2$-type loss are denoted by loss_l1 and loss_l2, respectively. Each measure is plotted as a function of the true linear component, $\beta'X$. Both length and combined mean absolute error (MAE) are computed relative to the oracle predictions and prediction intervals.

Therefore it is not surprising that there is little difference in performance as a result of switching to the $l2$-loss. On the other hand, we observed that the networks with two hidden layers perform worse than the networks with one hidden layer. In the left panel, we see that these prediction intervals adapt much worse to heteroskedasticity of the data (they overcover considerably when $\beta'X$ is small and undercover when $\beta'X$ is large). Therefore, the intervals are either much longer or much shorter (middle panel). This can be explained by the fairly simple structure of the data which does not require more complex neural networks.

4 Conformal prediction intervals for the individual treatment effect

Conformal prediction intervals for the individual treatment effect

Danijel Kivaranovic[a] Robin Ristl[b] Martin Posch[b] Hannes Leeb[a,c]

[a]Department of Statistics and Operations Research, University of Vienna

[b]Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna

[c]Data Science @ Uni Vienna, University of Vienna

Abstract

We propose several prediction intervals procedures for the individual treatment effect with either finite-sample or asymptotic coverage guarantee in a non-parametric regression setting, where non-linear regression functions, heteroskedasticity and non-Gaussianity are allowed. To construct the prediction intervals we use the conformal method of Vovk et al. (2005). In extensive simulations, we compare the coverage probability and interval length of our prediction interval procedures. We demonstrate that complex learning algorithms, such as neural networks, can lead to narrower prediction intervals than simple algorithms, such as linear regression, if the sample size is large enough.

Keywords: Conformal inference, individual treatment effect, machine learning.

1 Introduction

In usual randomized controlled clinical trials, the average treatment effect, averaged over the population under study, is estimated and used as predictor for the expected treatment effect in a new patient. However, treatment effects may depend on individual patient characteristics and the personalized medicine paradigm aims to tailor treatment decisions to these characteristics (Pellegrini et al., 2019; Graf et al., 2020). When comparing an experimental treatment to a control treatment, the relevant quantity to establish a personalized treatment decision is the so called *individual treatment effect*, defined as the difference between the patient's potential outcomes under treatment and under control (Rubin, 1974, 2005). In this paper, we propose several prediction intervals for the individual treatment effect with either finite-sample or asymptotic coverage guarantee. Our intervals are valid in a non-parametric regression setting, where non-linear regression functions, heteroskedasticity and non-Gaussianity are allowed. The construction of our prediction intervals relies on the conformal method of Vovk et al. (2005); Shafer and Vovk (2008).

To predict individual treatment effects based on a set of co-variates, regression models for, both, the response under treatment and under control can be fitted. This can be achieved, for example, if interaction terms for the treatment group indicator and the considered predictive variables are included or if separate regression models for each treatment group are fitted. Alternatively, more complex non-parametric or machine-learning models can be applied to model non-linear associations (Lamont et al., 2018). The expected individual treatment effect can then be estimated as difference of predictions under treatment and under control, given a patient's co-variate values.

In order to reliably apply this concept for decision making, the uncertainty attached to an estimated individual treatment effect needs to quantified. Ballarini et al. (2018) studied the application of confidence intervals for expected individual treatment effects to select subgroups with improved treatment benefit. While this approach covers uncertainty due to the prediction model being estimated from a finite training sample,

a complete assessment of uncertainty in an individual prediction needs to take into account the residual distribution of the effect, which comprises unexplained between-patient variability and within-patient variability. (The latter reflects the possibility that the same patient may respond differently to a treatment at different occasions; but only in rare settings which allow for repeated observations of the effect in the same patient these sources of variability may be identified separately (Senn, 2016).)

A suitable way to communicate the extent of remaining uncertainty is in terms of prediction intervals. The calculation of individual prediction intervals from models for the individual treatment effect is complicated by the fact that usually each patient in the training data set is observed under one of the two treatment conditions only. Hence, residuals of the individual treatment effect cannot be observed. To overcome this problem, in this manuscript we propose a method to calculate prediction intervals for the individual treatment effect by an extension of the conformal inference framework towards the difference in hypothetical outcomes under two treatment conditions. Conformal inference allows to estimate intervals in the residual distribution of an arbitrary prediction model without distributional assumptions and take into account uncertainty in the model estimation procedure. The resulting intervals are exact in the sense that the average coverage probability across the studied population is controlled.

In Section 2 we introduce the notation and present our main results: Theorem 1 and 2 provide prediction intervals with finite-sample coverage, and Theorem 3 provides prediction intervals with asymptotic coverage. Theorem 1 and 2 have weaker assumptions but are therefore more conservative. Theorem 3 has stronger assumptions, namely consistency of the prediction procedure and Gaussian error distribution, but provides considerably shorter intervals. All three theorems assume the existence of a prediction interval procedure for the outcome conditional on the assigned treatment. In Section 3, we show how to construct such a prediction interval procedure using the conformal method. In Section 4 we compare the intervals in several simulation settings. The paper ends with a discussion in Section 5. The proofs of all results are given in the appendix.

2 Prediction intervals for the individual treatment effect

2.1 Notation and definitions

We denote by X the random covariate vector with values in $\mathbb{R}^d$, $d \in \mathbb{N}$. Let T be the $\{-1, 1\}$-valued random variable that indicates whether a patient was assigned to the treatment group ($T = 1$) or control group ($T = -1$). We assume that the random variables X and T are independent. This means, patients are assigned to treatment or control group independent of their condition, which is described by X. (We note that this setting excludes stratified sampling, because it would introduce a dependence structure between X and T.) Let Y denote the real-valued outcome of interest. We can decompose Y into

$$Y = f(X, T) + e_{X,T},$$

where $f(X, T) = \mathbb{E}(Y|X, T)$ is called the regression function and $e_{X,T} = Y - \mathbb{E}(Y|X, T)$ the error. The X and T in the subscript of $e_{X,T}$, denote the dependence of $e_{X,T}$ on X and T. Each patient is only assigned to one of the two treatment groups, which means that the random variable

$$Y' = f(X, -T) + e_{X,-T}$$

is not observable. Note that Y' denotes the outcome of the patient assigned to the opposite treatment group. The individual treatment effect $\tau(X)$ is defined as $T(Y - Y')$ or, equivalently,

$$\tau(X) = f(X, 1) - f(X, -1) + e_{X,1} - e_{X,-1}.$$

Clearly, because Y' is not observable, $\tau(X)$ is also not observable. We note that $\tau(X)$ is not to be confused with $f(X, 1) - f(X, -1)$ which is also sometimes referred to as (predicted) individual treatment effect in the literature. Conditional on X, $f(X, 1) - f(X, -1)$ is deterministic, while $\tau(X)$ is still random because of the error $e_{X,1} - e_{X,-1}$. This also means that, in a well-defined setting, the length of confidence

intervals for $f(X,1) - f(X,-1)$ converge to 0 as sample size grows while the length of prediction intervals for $\tau(X)$ cannot converge to 0 unless the variance of $e_{X,1} - e_{X,-1}$ conditional on X is zero. In Theorem 1 we make no assumption about the distribution of $e_{X,1}$ and $e_{X,-1}$, while Theorem 2 requires conditional independence and Theorem 3 conditional positive correlation given X, respectively. We note that conditional independence between $e_{X,1}$ and $e_{X,-1}$ given X is often considered as worst-case scenario, because unobserved explanatory variables typically cause a positive correlation structure between the errors.

Let $D_n = ((X_i, Y_i, T_i))_{1 \leq i \leq n}$ denote a dataset of $n \in \mathbb{N}$ observations, where the (X_i, T_i, Y_i)s are i.i.d. copies of (X, Y, T). We only consider the non-degenerate case where $0 < \mathbb{P}(T = 1) < 1$. We denote by $\mathbb{D}^n = (\mathbb{R}^d \times \mathbb{R} \times \{-1, 1\})^n$ the range of D_n. Throughout this paper, (X, Y, T) denotes a new observation that is independent of D_n. Let $\alpha \in (0, 1)$ be the confidence level. We cannot define a prediction interval procedure for the individual treatment effect $\tau(X)$ based on the unobservable vector $(\tau(X_1), \ldots, \tau(X_n))'$. Therefore, we propose to construct a prediction interval $[l_{D_n}(X, -1), u_{D_n}(X, -1)]$ for Y conditional on the event $\{T = -1\}$ and another prediction interval $[l_{D_n}(X, 1), u_{D_n}(X, 1)]$ for Y conditional on the event $\{T = 1\}$. Then, we combine these two prediction intervals to obtain a prediction interval for $\tau(X)$. In Section 3, we show how to modify the conformal method (Vovk et al. (2005); Shafer and Vovk (2008)) to construct prediction intervals for Y with coverage probability $1-\alpha$ conditional on $\{T = -1\}$ and $\{T = 1\}$, respectively. The construction of these intervals merely requires i.i.d. data and no further assumption on the distribution of the data or the learning algorithm.

The accuracy of the proposed procedure (in terms of coverage probability and interval length), depends on how accurately we can estimate the regression function $f(x, t)$ and how accurately we can estimate the conditional distribution of $e_{X,T}$ conditional on X and T.

Definition 1. *A prediction procedure $\mathcal{A}_n$ is a mapping from $\mathbb{D}^n$ into the set of all*

measurable functions $\{f : \mathbb{R}^d \times \{-1, 1\} \to \mathbb{R}\}$. *We set*

$$f_{D_n} = \mathcal{A}_n(D_n).$$

We say that $\mathcal{A}_n$ *is a consistent prediction procedure if for any compact set* $K \subseteq \mathbb{R}^d$ *and any* $\epsilon > 0$

$$\mathbb{P}\left(\sup_{(x,t)\in K\times\{-1,1\}} |f_{D_n}(x,t) - f(x,t)| < \epsilon\right) \to 1 \quad \text{as } n \to \infty.$$

Typically, the efficacy of the treatment t interacts with the covariate vector x. In the following, we give examples of prediction procedures $\mathcal{A}_n$ that model these interactions. Furthermore, we discuss when these procedures are consistent in the sense of Definition 1:

(i) The least squares method is the most common and most frequently used prediction procedure. The function f_{D_n} is of the form

$$f_{D_n}(x,t) = \hat{\beta}_0 + \hat{\beta}_1 t + \sum_{i=1}^{d} \hat{\gamma}_i x_i + t \sum_{i=1}^{d} \hat{\delta}_i x_i,$$

where the $\hat{\beta}$s, $\hat{\gamma}$s and $\hat{\delta}$s are chosen by minimizing the mean squared error on the dataset D_n. Interactions between x and t are explicitly modeled by the last sum. The linear least squares method can only be consistent if the regression function $f(x,t)$ is a linear function of x and t and the components of x only interact with t.

(ii) Kernel regression is a non-parametric prediction procedure. The function f_{D_n} is defined as

$$f_{D_n}(x,t) = \sum_{i=1}^{n} Y_i \hat{w}_i(x,t),$$

where the $\hat{w}_i$s are D_n-depended weight functions such that $\hat{w}_i(x,t)$ is large if (x,t) is close to (X_i, T_i) and small if (x,t) is far from (X_i, T_i). Interaction between x and t are implicitly modeled by the data-dependent weight functions. Sufficient conditions for consistency of kernel regression methods are given in Stone (1977).

(iii) The neural network is a prediction procedure that has recently gained a lot of attention because of its predictive performance. A standard neural network f_{D_n} with one hidden layer, with $k \in \mathbb{N}$ nodes and non-linear activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is defined as

$$f_{D_n}(x,t) = \hat{W}_2 \; \sigma(\hat{W}_1(x',t)' + \hat{b}_1) + \hat{b}_2,$$

where $\hat{W}_1$ is a D_n-dependent $k \times (d+1)$ weight matrix; $\hat{b}_1$ is a D_n-dependent k-dimensional bias vector; σ is applied component-wise; $\hat{W}_2$ is a D_n-dependent $1 \times k$ weight matrix and $\hat{b}_2$ is a D_n-dependent 1-dimensional bias vector. A neural network is called deep, if the number of hidden layers and hidden nodes is large. The larger the network, the more interactions between x and t can be modeled. Neural networks are widely known as universal approximators (Cybenko, 1989; Hornik et al., 1989) and, more recently, consistency properties from statistical perspective have been analyzed by Bauer and Kohler (2019).

Note that if the sample size n is small, some prediction procedures are not well defined (e.g. if the number of parameters in a linear regression model is larger than n). In such cases, f_{D_n} is simply to be interpreted as the 0 function.

Definition 2. *An interval procedure $\mathcal{P}_n$ is a mapping from $\mathbb{D}^n$ into the set of all pairs of ordered measurable functions $\{(l,u) : l, u : \mathbb{R}^d \times \{-1,1\} \to \mathbb{R},\ l \leq u\}$. We set*

$$(l_{D_n},\ u_{D_n}) = \mathcal{P}_n(D_n).$$

We say that $\mathcal{P}_n$ is a T-conditional prediction-interval procedure at level $1-\alpha$, if for all $n \in \mathbb{N}$

$$\mathbb{P}\left(Y \in [l_{D_n}(X,T), u_{D_n}(X,T)] \mid T = t\right) \geq 1 - \alpha, \quad t \in \{-1,1\}.$$

The standard conformal method allows to construct unconditional prediction intervals requiring only that the data is i.i.d.. In Section 3, we show how to modify the conformal method to obtain a T-conditional prediction-interval procedure in the sense of Definition 2.

2.2 Theoretical results

We propose prediction intervals for $\tau(X)$ that either control finite sample coverage or asymptotic coverage.

Definition 3. *Let $\Gamma_{D_n} : \mathbb{R}^d \to \{[a,b] : a \leq b\}$ be an interval-valued function that depends on D_n. $\Gamma_{D_n}(X)$ is called a prediction interval for the individual treatment effect $\tau(X)$ at level $1-\alpha$, if for all $n \in \mathbb{N}$*

$$\mathbb{P}(\tau(X) \in \Gamma_{D_n}(X)) \geq 1 - \alpha. \tag{1}$$

We say that $\Gamma_{D_n}(X)$ is an asymptotically valid prediction interval for $\tau(X)$ at level $1-\alpha$, if

$$\lim_{n\to\infty} \mathbb{P}(\tau(X) \in \Gamma_{D_n}(X)) \geq 1 - \alpha. \tag{2}$$

In all of the following results, we assume that we have a T-conditional prediction-interval procedure $\mathcal{P}_n$ in the sense of Definition 2. Construction of such a procedure is described in Section 3. We note that the following results hold for any T-conditional prediction-interval procedure $\mathcal{P}_n$ (not necessarily based on the method proposed in Section 3).

The first result only requires i.i.d. observations and a T-conditional prediction-interval procedure at level $1-\alpha/2$.

Theorem 1. *Let $\mathcal{P}_n$ be a T-conditional prediction-interval procedure at level $1-\alpha/2$ in the sense of Definition 2. Set $(l_{D_n}, u_{D_n}) = \mathcal{P}_n(D_n)$ and*

$$\Gamma_{D_n}(X) = \left[l_{D_n}(X,1) - u_{D_n}(X,-1), u_{D_n}(X,1) - l_{D_n}(X,-1)\right]. \tag{3}$$

Then $\Gamma_{D_n}(X)$ is a prediction interval for $\tau(X)$ at level $1-\alpha$.

In this result, $\Gamma_{D_n}(X)$ controls coverage at level $1-\alpha$, but $\mathcal{P}_n$ is a T-conditional prediction-interval procedure at level $1-\alpha/2$. We can improve this result (in terms of interval length) by assuming conditional independence of the errors $e_{X,-1}$ and $e_{X,1}$ given X. Under this additional assumption, a T-conditional prediction-interval procedure at

level $\sqrt{1-\alpha}$ (which is strictly smaller than $1-\alpha/2$ for all $\alpha \in (0,1)$) is sufficient such that inequality (1) continues to hold.

Theorem 2. *Let $\mathcal{P}_n$ be a T-conditional prediction-interval procedure at level $\sqrt{1-\alpha}$ in the sense of Definition 2. Assume that $e_{X,-1}$ and $e_{X,1}$ are conditionally independent given X. Set $(l_{D_n}, u_{D_n}) = \mathcal{P}_n(D_n)$ and let $\Gamma_{D_n}(X)$ be defined as in (3). Then $\Gamma_{D_n}(X)$ is a prediction interval for $\tau(X)$ at level $1-\alpha$.*

Both, Theorem 1 and 2, have strong finite-sample coverage claims under minimal assumptions. Thus, it is not surprising that these intervals are conservative in certain situations. This is especially the case if the true regression function $f(x,t)$ is simple and easy to estimate (see the simulations of Section 4). In Theorem 1 and 2 we made no assumption on the accuracy of the estimated regression function $f_{D_n}(x,t)$. This means that these intervals do not only take the variance of the error $e_{X,1} - e_{X,-1}$ but also the estimation error of $f_{D_n}(x,t)$ into account.

The intervals of the next result are much shorter, but the assumptions are considerable stronger and the coverage claim is weaker.

Theorem 3. *Let $\mathcal{A}_n$ a be consistent prediction procedure as given in Definition 1. Let $\mathcal{P}_n$ be a T-conditional prediction-interval procedure at level $1-\alpha$ as given in Definition 2. Set (l_{D_n}, u_{D_n}) and suppose that*

$$\mathbb{P}\left(l_{D_n}(X,T) \leq f_{D_n}(X,T) \leq u_{D_n}(X,T)\right) \;=\; 1. \tag{4}$$

Assume that conditional on X, $(e_{X,1}, e_{X,-1})'$ is multivariate normal-distributed with $\mathrm{Var}(e_{X,1}) = \mathrm{Var}(e_{X,-1})$ and non-negative covariance. For $t \in \{-1,1\}$, set

$$\tilde{l}_{D_n}(X,t) = f_{D_n}(X,t) - \frac{f_{D_n}(X,t) - l_{D_n}(X,t)}{\sqrt{2}};$$
$$\tilde{u}_{D_n}(X,t) = f_{D_n}(X,t) + \frac{u_{D_n}(X,t) - f_{D_n}(X,t)}{\sqrt{2}}$$

and

$$\Gamma_{D_n}(X) = \left[\tilde{l}_{D_n}(X,1) - \tilde{u}_{D_n}(X,-1), \tilde{u}_{D_n}(X,1) - \tilde{l}_{D_n}(X,-1)\right].$$

Then $\Gamma_{D_n}(X)$ is an asymptotically valid prediction interval for $\tau(X)$ at level $1-\alpha$.

An inspection of the proof and the remark after Lemma 1 shows that if the errors $e_{X,-1}$ and $e_{X,1}$ are negatively correlated then Theorem 3 continuous to hold if we omit the factor $\sqrt{2}$ in the definition of $\tilde{l}_{D_n}(X,t)$ and $\tilde{u}_{D_n}(X,t)$. (This means, $\tilde{l}_{D_n}(X,t)$ and $\tilde{u}_{D_n}(X,t)$ reduces to $l_{D_n}(X,t)$ and $u_{D_n}(X,t)$, respectively.)

Because $\mathcal{A}_n$ is a consistent prediction procedure, the proof of the theorem basically reduces to showing an elementary probabilistic inequality for Gaussian random variables which is given in Lemma 1 of the appendix.

3 T-conditional prediction-interval procedures

Conformal prediction intervals control coverage marginally (Vovk et al., 2005). But it is straightforward to construct conformal prediction intervals with coverage conditional on any event with positive probability. Loosely speaking, one only needs to apply the conformal method "conditional" on the event of interest. For completeness, this section describes in detail how to apply the conformal method to construct T-conditional prediction-interval procedures $\mathcal{P}_n$ in the sense of Definition 2. Because computation of conformal prediction intervals is expensive and infeasible in certain situations, we also show how to adapt the "split-conformal" procedure (Lei et al. (2018)). Computation of the "split-conformal" prediction intervals is much cheaper but the resulting intervals are typically longer.

3.1 Nonconformity measures and scores

In the following, we define the terms conformal procedure, nonconformity measure and nonconformity score.

Definition 4. *A conformal procedure $\mathcal{M}_n$ is a mapping from $\mathbb{D}^n$ into the set of measurable functions $\{m : \mathbb{D} \to [0,\infty)\}$. We set*

$$m_{D_n} = \mathcal{M}_n(D_n),$$

where we call m_{D_n} the nonconformity measure and $m_{D_n}(x, y, t)$ the nonconformity score of the observation $(x, y, t) \in \mathbb{D}$.

$\mathcal{M}_n$ is called permutation invariant if, for any $n \in \mathbb{N}$, for any permutation $\pi_n : \{1, \ldots, n\} \to \{1, \ldots, n\}$ and for any $(x, y, t) \in \mathbb{D}$,

$$\mathbb{P}\left(m_{D_n}(x, y, t) = m_{D_{n,\pi_n}}(x, t)\right) \;=\; 1.$$

where $D_{n,\pi_n} = ((X_{\pi_n(1)}, Y_{\pi_n(1)}, T_{\pi_n(1)}), \ldots, (X_{\pi_n(n)}, Y_{\pi_n(n)}, T_{\pi_n(n)}))$.

In the following, we give examples of different nonconformity measures.

Example 1 (Absolute residuals)**.** *The simplest nonconformity measure is given by*

$$m_{D_n}(x, y, t) = |y - f_{D_n}(x, t)|.$$

However, this measure is not optimal if the conditional variance of Y given X and T is not constant, i.e., if the error $e_{X,T}$ is heteroscedastic.

Example 2 (Absolute standardized residuals)**.** *To take heteroscedasticity into account, one can define the nonconformity measure as*

$$m_{D_n}(x, y, t) = |y - f_{D_n}(x, t)|/\sigma_{D_n}(x, t),$$

where $\sigma^2_{D_n}(x, t)$ is an estimator for the conditional variance of Y conditional on X and T. Typically, a kernel-based algorithm is used to obtain $\sigma^2_{D_n}(x, t)$.

Example 3 (ML-based measures)**.** *Modern machine learning algorithms, such as neural networks, are able to estimate the distribution of the error $e_{X,t}$ and one can define nonconformity measures tailored to these algorithms as shown in Kivaranovic et al. (2019); Romano et al. (2019).*

Note that the conformal procedures in these examples are permutation invariant, if the corresponding estimation procedures are permutation invariant.

3.2 T-conditional conformal prediction intervals

Let $D_{n+1}(y)$ denote the augmented dataset which contains D_n and the pseudo observation (X, y, T), where $y \in \mathbb{R}$. Recall that

$$m_{D_{n+1}(y)} = \mathcal{M}_{n+1}(D_{n+1}(y)),$$

is the nonconformity measure of the conformal procedure $\mathcal{M}_{n+1}$ evaluated at $D_{n+1}(y)$. Denote by $D_{n+1,T}(y)$ the subset of observations in $D_{n+1}(y)$ with treatment equal to T. Note that $D_{n+1,T}(y)$ is non-empty because (X, y, T) is always contained. For ease of notation, we set

$$m_{D_{n+1}(y)}(D_{n+1,T}(y)) = \left\{ m_{D_{n+1}(y)}((\tilde{X}, \tilde{Y}, \tilde{T})) : (\tilde{X}, \tilde{Y}, \tilde{T}) \in D_{n+1,T}(y) \right\}, \tag{5}$$

this means, $m_{D_{n+1}(y)}(D_{n+1,T}(y))$ is the set of nonconformity scores of the nonconformity measure $m_{D_{n+1}(y)}$ evaluated on the random variables in $D_{n+1,T}(y)$. Denote by $R_{D_{n+1}(y)}(y)$ the rank of $m_{D_{n+1}(y)}(X, y, T)$ among the random variables in $m_{D_{n+1}(y)}(D_{n+1,T}(y))$. In case of ties, we use a data-independent tie-breaking rule. Let $N_T = \sum_{i=1}^{n} \mathbb{1}_{\{T_i = T\}}$ and set

$$\Gamma^F_{D_n}(X, T) = \left\{ y \in \mathbb{R} : \; R_{D_{n+1}(y)}(y) \leq \lceil (1-\alpha)(N_T + 1) \rceil \right\}. \tag{6}$$

(Note that if $\alpha(N_T + 1) < 1$, then $\Gamma^F_{D_n}(X, T) = (-\infty, \infty)$.) The superscript F emphasizes that the full dataset is used for computation of the nonconformity measure. This is in contrast to the "split-conformal" version of the next section, where only a subset is used for computation of the nonconformity measure.

Theorem 4. *Let $\mathcal{M}_n$ be a permutation invariant conformal procedure as given in Definition 4 . Let $\Gamma^F_{D_n}(X, T)$ be defined as in (6). Let $t \in \{-1, 1\}$ and set $p_t = \mathbb{P}(T = t)$. Then*

$$1 - \alpha \;\leq\; \mathbb{P}\left(Y \in \Gamma^F_{D_n}(X, T) \mid T = t\right) \;\leq\; 1 - \alpha + \frac{1 - (1 - p_t)^{n+1}}{(n+1)p_t}.$$

The upper bound in this theorem converges to $1 - \alpha$ as $n \to \infty$, therefore, the procedure is asymptotically exact. We note that $\Gamma^F_{D_n}(X, T)$ is not an interval in general.

But, obviously, the interval

$$\left[\inf\ \Gamma^F_{D_n}(X,T), \sup\ \Gamma^F_{D_n}(X,T)\right]$$

is a T-conditional prediction interval for Y, because it is a superset of $\Gamma^F_{D_n}(X,T)$. Computation of $\Gamma^F_{D_n}(X,T)$ is very expensive if a brute-force approach is used because then one needs to compute the nonconformity measure $m_{D_{n+1}(y)}$ for all augmented datasets $D_{n+1}(y)$, where $y \in \mathbb{R}$. Of course, practically, we do this only on a sufficiently dense grid for y, but if the dataset is large and one is using a modern learning algorithm, such as a neural network, this is computationally infeasible.

3.3 T-conditional "split-conformal" prediction intervals

For $1 \leq m < n$, D_m is the dataset of the first m observation and denote by $D_{m:n}$ the dataset of the remaining $n-m$ observations. Let $D_{m:(n+1)}(y)$ be the augmented dataset which contains $D_{m:n}$ and the pseudo observation (X, y, T). Denote by $D_{m:(n+1),T}(y)$ the subset of observations in $D_{m:(n+1)}(y)$ with treatment equal to T. Again, $D_{m:(n+1),T}(y)$ is non-empty because (X, y, T) is always contained. Let $m_{D_m}(D_{m:(n+1),T}(y))$ be defined as in (5) with D_m replacing $D_{n+1}(y)$ and $D_{m:(n+1),T}(y)$ replacing $D_{n+1,T}(y)$. This means, $m_{D_m}(D_{m:(n+1),T}(y))$ is the set nonconformity scores of the nonconformity measure m_{D_m} evaluated on the random variables in $D_{m:(n+1),T}(y)$. Denote by $R_{D_m}(y)$ the rank of $m_{D_m}(X, y, T)$ among the random variables in $D_{m:(n+1),T}(y)$. In case of ties, we use a data-independent tie-breaking rule. Let $\tilde{N}_T = \sum_{i=m+1}^{n} \mathbb{1}_{\{T_i=T\}}$ and set

$$\Gamma^S_{D_n}(X,T) = \left\{ y \in \mathbb{R} :\ R_{D_m}(y) \leq \left\lceil (1-\alpha)\left(\tilde{N}_T + 1\right)\right\rceil \right\}. \tag{7}$$

(Note that if $\alpha(\tilde{N}_T+1) < 1$ then $\Gamma^S_{D_n}(X,T) = (-\infty, \infty)$.) The superscript S emphasizes that only a subset of the data is used for computation of the nonconformity measure.

Theorem 5. *Let $\mathcal{M}_n$ be a conformal procedure as given in Definition 4 . Let $\Gamma^S_{D_n}(X,T)$ be defined as in (7). Let $t \in \{-1, 1\}$ and set $p_t = \mathbb{P}(T = t)$,*

$$1-\alpha \ \leq\ \mathbb{P}\left(Y \in \Gamma^S_{D_n}(X,T) \mid T = t\right) \ \leq\ 1-\alpha + \frac{1-(1-p_t)^{n-m+1}}{(n-m+1)p_t}.$$

Again, the upper bound converges to $1-\alpha$ as $n-m \to \infty$, therefore, the procedure is asymptotically exact. Note that in the "split-conformal" case, we do not need that $\mathcal{M}_n$ is permutation invariant. Computation of $\Gamma^S_{D_n}(X,T)$ is much cheaper, because we need to calculate the nonconformity measure only once on D_m. The set $\Gamma^S_{D_n}(X,T)$ is an interval if the nonconformity measure m_{D_n} is convex in y, which is typically the case for the most common nonconformity measures.

4 Simulation

We fixed the covariate dimension $d = 10$ and the significance level $\alpha = 0.1$. Further, we set $\mathbb{P}(T = 1) = 1/2$. We considered 4 different sample sizes: $n = 300, 700, 1200, 2000$. The covariate vector X is $N(0, \Sigma)$-distributed, where Σ is a 10×10 matrix where each element on the diagonal is equal 1 and each element off the diagonal is equal to ρ. We considered two values for ρ: 0.2 (low correlation) and 0.8 (high correlation). We considered two different regression functions:

$$f_1(x,t) = x_1 + x_2 + x_3 + t \quad \text{and} \quad f_2(x,t) = \text{sign}(f_1(x,t)) f_1(x,t)^2.$$

We also considered two different distributions for the error distribution: $e_{X,t}$ conditional on X is either a standard normal distributed or Laplace distributed with variance 1. Considering all possible combinations, we have 8 different data generating processes in total: LOW COR vs. HIGH COR, LINEAR vs. NON LINEAR and NORMAL vs. LAPLACE.

We also considered two different T-conditional prediction interval procedures: The first is the linear least squares method with the absolute error as nonconformity measure (as described in Example 1). Because linear models are very fast to fit, we use the full conformal version as described in Section 3.2 to construct the prediction intervals. The second is a neural network with one hidden layer and ten hidden nodes and the same nonconformity measure as in the previous case. Because fitting a neural network is computationally more expensive, we use the split conformal approach as described in

Section 3.3. We used 2/3 of the data for training of the model and 1/3 for computation of the nonconformity scores. For both methods, we constructed the prediction interval as described in Theorem 2 and 3. This means, in total we have 4 different prediction intervals which we label LM1, LM2, NN1 and NN2. Here LM denotes the linear least squares method, NN the neural network and 1 and 2 denote the construction as described in Theorem 2 and 3, respectively.

For each of the 4 sample sizes, 8 data generating processes and 4 prediction intervals, we repeated the following steps 1000 times: First, we draw the dataset D_n and the new covariate vector X and we compute the prediction interval for $\tau(X)$. Then we draw $\tau(X)$. And, finally, we compute the length of the prediction interval and check if it covers $\tau(X)$. The length of the prediction intervals is computed relative to the length of the oracle which knows the underlying data generating process. To compute the conformal prediction intervals, we took a sufficiently large grid with step size .1 and defined the interval as the smallest and largest value in the grid that satisfied the condition (6). The results are summarized in Figure 1 and 2.

In Figure 1, all data generating processes with Gaussian error distribution are shown. We see that the methods LM1 and NN1 always control coverage. This is in line with Theorem 2 which guarantees finite sample coverage in all these situations. But, because of the strong coverage claim, these intervals are also very conservative, especially when the true regression function is a simple linear model (first and second row). Furthermore, we see that the method NN1 significantly outperforms LM1 (in terms of length) when the true model is nonlinear (third and fourth row) and the sample size grows. This means, the neural network is able to estimate the non linear regression function more accurately as the sample size grows and, therefore, gives more accurate intervals. The linear regression is clearly not able to do this.

On the other hand, the methods LM2 and NN2 are much closer to the nominal level of 90% and their lengths are much closer to the oracle. But we note that the assumptions of Theorem 3 are much stronger. We see that LM2 is nearly optimal (in terms of coverage and length) if the true regression function is linear. The method

NN2 performs only marginally worse in these cases. This is not surprising because both, linear regression and neural networks, are able to consistently estimate linear functions. NN2 performs a bit worse because the complexity of the neural network is not required for the simple structure of the linear regression function. In the cases where the true regression function is non linear (third and fourth row), the method LM2 has coverage consistently below the nominal level. Furthermore, the length of the intervals is up to 4 times larger than of the oracle and does not improve with growing sample size. Clearly, LM2 does not satisfy the assumptions of Theorem 3 and there for it is not surprising that it does not control coverage. However, NN2 performs much better in these situations. The coverage of NN2 is close to the nominal level and the length converges to the length of the oracle as the sample size grows. This shows that the gain of complex models is very high if sufficient data is available and the true regression function not linear. NN2 controls not only the coverage more accurately, the intervals are also shorter by a factor 3 to 4.

To demonstrate the robustness of our results, Figure 2 shows the same data generating processes with Laplace error instead of Gaussian error. All four methods behave very similarly to the Gaussian case. This suggests that Theorem 3 may be valid for a wider class of distributions.

5 Discussion

In this manuscript we proposed several methods to construct prediction intervals for the estimation of individual treatment effects based on complex prediction models such as regression models with variable selection or neural network algorithms. The prediction intervals give bounds for the differences between the outcomes under treatment and control for individual patients. This is in contrast to confidence intervals for the predicted individual treatment effect, which estimate the expected individual treatment effect for patients with specific baseline characteristics.

The intervals of Theorem 1 and 2 control coverage of the individual treatment effect

under minimal assumptions. Theorem 3 provides considerably shorter intervals that are asymptotically valid under stronger assumptions. A limitation in the interpretation of conformal prediction interval is that the coverage probabilities are guaranteed on a population level only, but not conditional on a specific vector of covariates. Especially, they are only meaningful, if predictions are made for a population with the same covariate distribution as the population from which the intervals were computed. This is in contrast to the classical prediction intervals, e.g., for linear models, that also provide conditional coverage, given the value of the covariates.

The width of the prediction intervals can be a useful measure to compare different prediction models. While more flexible models may be able to explain more variability by adjusting for more covariates or more complex functions of the covariates, they also introduce variability because of increasing uncertainties in the model estimates for more extensive models. As demonstrated in the simulation study, if the sample size is low simpler models, even though they are miss-specified, can lead to narrower prediction intervals than more complex models which overfit the data and lead to highly variable model estimates.

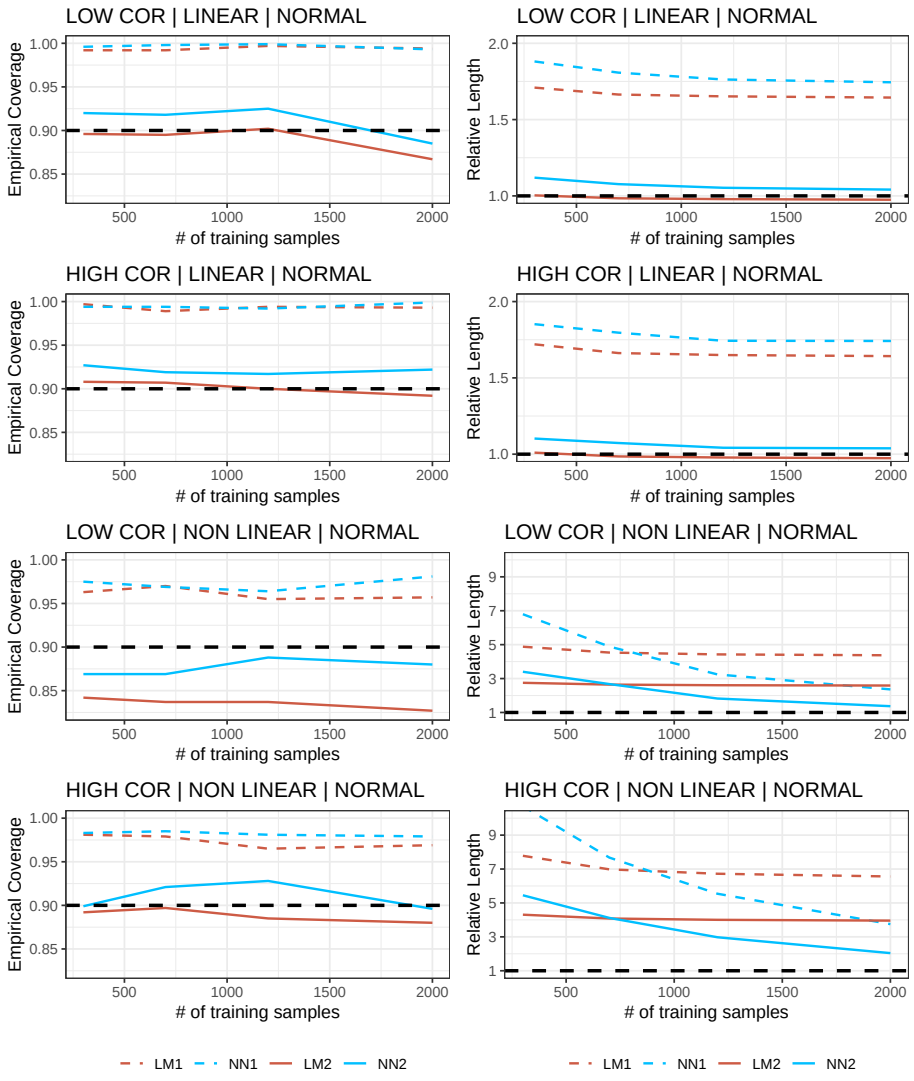

Figure 1: Empirical coverage is shown on the l.h.s. (standard error is approximately 1%). The dotted line at .9 denotes the nominal confidence level. On the r.h.s., the relative length compared to the oracle is shown. The dotted line at 1 denotes the length of the oracle. The data generating process is given above each graph. Each setting is simulated with Gaussian errors.

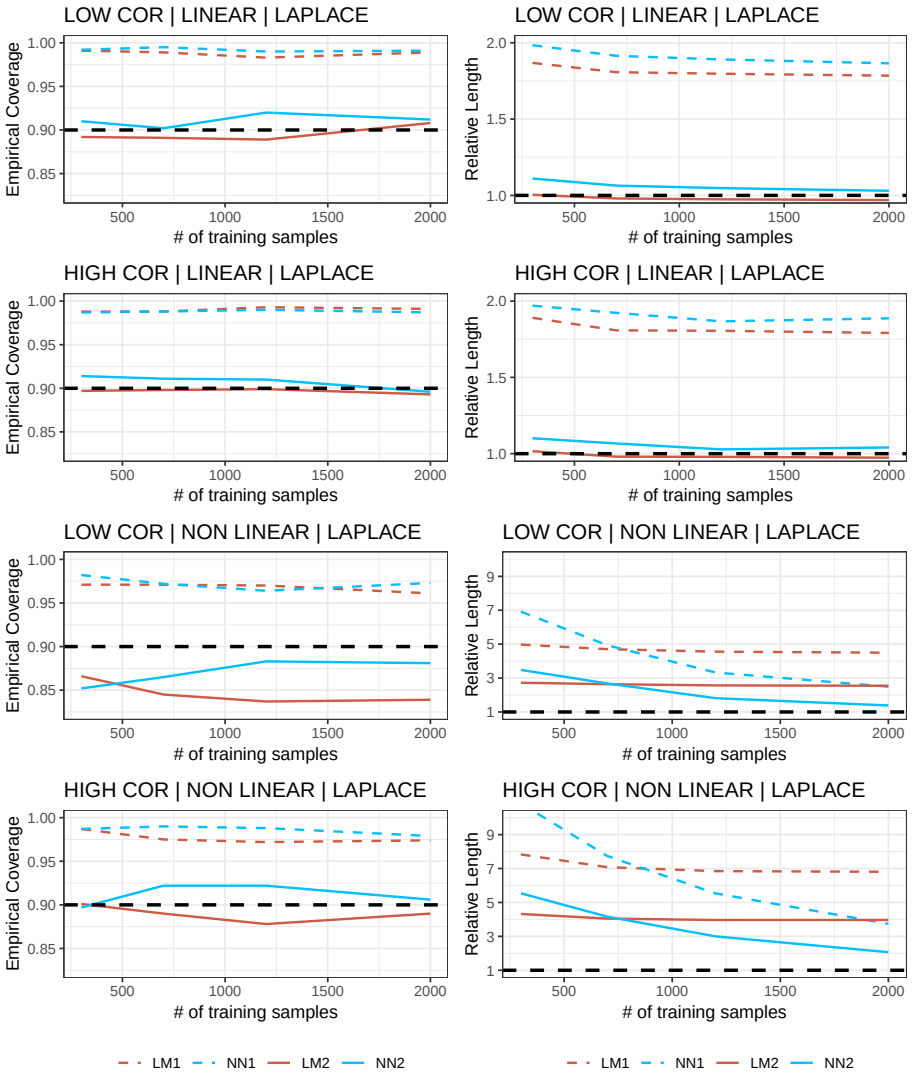

Figure 2: The results of the same simulation settings as in Figure 1 are shown but with Laplace error distribution instead of Gaussian error distribution.

Appendix A Proofs of the results

A.1 Proof of Theorem 1 and 2

Proof of Theorem 1. By definition of $\Gamma_{D_n}(X)$, we have

$$\begin{aligned}
&\{f(X,-1)+\epsilon_{X,-1} \ \in [l_{D_n}(X,-1), u_{D_n}(X,-1)]\} \\
&\cap \ \{f(X,1)+\epsilon_{X,1} \ \in [l_{D_n}(X,1), u_{D_n}(X,1)]\} \\
&\subseteq \{\tau(X) \in \Gamma_{D_n}(X)\}\,.
\end{aligned}$$

This means, it is sufficient to show that the probability of the smaller event is larger than $1-\alpha$. For $t \in \{-1,1\}$, we have

$$\begin{aligned}
&\mathbb{P}\left(Y \ \in [l_{D_n}(X,T), u_{D_n}(X,T)] \mid T=t\right) \\
&= \mathbb{P}\left(f(X,t)+\epsilon_{X,t} \ \in [l_{D_n}(X,t), u_{D_n}(X,t)] \mid T=t\right) \\
&= \mathbb{P}\left(f(X,t)+\epsilon_{X,t} \ \in [l_{D_n}(X,t), u_{D_n}(X,t)]\right),
\end{aligned}$$

where the last equation follows because T is independent of D_n and X. If $\mathcal{P}_n$ is a T-conditional prediction interval procedure at level $1-\alpha/2$, the previous equation implies that

$$\mathbb{P}\left(f(X,t)+\epsilon_{X,t} \ \in [l_{D_n}(X,t), u_{D_n}(X,t)]\right) > 1-\alpha/2$$

for $t \in \{-1,1\}$. Hence the probability of the intersection is bounded from below by $1-\alpha$.

□

Proof of Theorem 2. Set

$$A_t = \{f(X,t)+\epsilon_{X,t} \ \in [l_{D_n}(X,t), u_{D_n}(X,t)]\}.$$

Because of the observation at the beginning of the previous proof, it is sufficient to show that

$$\mathbb{P}(A_{-1} \cap A_1) = \mathbb{E}\left[\mathbb{P}\left(A_{-1} \cap A_1 \mid \ X, D_n\right)\right].$$

is bounded from below by $1-\alpha$. Observe that, conditional on X and D_n, the only random quantities in A_{-1} and A_1 are $e_{X,-1}$ and $e_{X,1}$, respectively. Because $e_{X,-1}$ and $e_{X,1}$ are conditionally independent given X, it follows that

$$\mathbb{P}\left(A_{-1} \cap A_1 \mid \ X, D_n\right) = \mathbb{P}\left(A_{-1} \mid \ X, D_n\right) \mathbb{P}\left(A_1 \mid \ X, D_n\right)$$

and therefore it is sufficient to show that

$$\mathbb{E}\left[\mathbb{P}\left(A_{-1} \mid \ X, D_n\right) \mathbb{P}\left(A_1 \mid \ X, D_n\right)\right].$$

is bounded from below by $1-\alpha$. Because the product function is convex on $[0,1]^2$, the Jensen inequality implies that the previous expression is bounded from below by

$$\mathbb{E}\left[\mathbb{P}\left(A_{-1} \mid \ X, D_n\right)\right] \mathbb{E}\left[\mathbb{P}\left(A_1 \mid \ X, D_n\right)\right] = \mathbb{P}(A_{-1})\mathbb{P}(A_1) \geq 1-\alpha.$$

□

A.2 Proof of Theorem 3

Before we prove the theorem, we need the following two lemmas.

Lemma 1. *Let $(W_1, W_2)'$ be a 2-dimensional centered Gaussian random vector with $\mathrm{Var}(W_1) = \mathrm{Var}(W_2)$ and non-negative covariance. For any $l_1, l_2 \leq 0$ and $u_1, u_2 \geq 0$, we have*

$$\begin{aligned}\mathbb{P}\left(\frac{l_1 - u_2}{\sqrt{2}} \leq W_1 - W_2 \ \leq \ \frac{u_1 - l_2}{\sqrt{2}}\right) \ \geq \ & \frac{1}{2}\,\mathbb{P}\left(l_1 \leq W_1 \leq u_1\right) \\ & + \ \frac{1}{2}\,\mathbb{P}\left(l_2 \leq W_2 \leq u_2\right).\end{aligned}$$

Proof. We set $\sigma^2 = \mathrm{Var}(W_1) = \mathrm{Var}(W_2)$. Because the covariance of W_1 and W_2 is non-negative, it follows that the variance of $W_1 - W_2$ is bounded from above by $2\sigma^2$. This implies that the probability on the l.h.s. of the inequality of the lemma is bounded from below by

$$\Phi\left(\frac{u_1 - l_2}{2\sigma}\right) - \Phi\left(\frac{l_1 - u_2}{2\sigma}\right) = \Phi\left(\frac{u_1 - l_2}{2\sigma}\right) + \Phi\left(\frac{u_2 - l_1}{2\sigma}\right) - 1,$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. Because $\Phi(x)$ is concave on the non-negative real numbers and u_1 and $-l_1$ are non-negative, it follows that

$$\Phi\left(\frac{u_1 - l_2}{2\sigma}\right) \geq \frac{1}{2}\Phi\left(\frac{u_1}{\sigma}\right) + \frac{1}{2}\Phi\left(\frac{-l_2}{\sigma}\right).$$

and likewise for the second summand on the r.h.s. of the previous equation. But this implies the inequality of the lemma. □

If the random variables W_1 and W_2 are negatively correlated, an inspection of the proof shows that the lemma continuous to hold if we replace $(l_1 - u_2)/\sqrt{2}$ and $(u_1 - l_2)/\sqrt{2}$ by $l_1 - u_2$ and $u_1 - l_2$ on the l.h.s. of the inequality, respectively.

Lemma 2. *Let W be a centered Gaussian distribution. Let $l, u \in \mathbb{R}$ with $l \leq u$. Then for all $\epsilon > 0$, we have*

$$\mathbb{P}(l + \epsilon \leq W \leq u - \epsilon) \geq \mathbb{P}(l \leq W \leq u) - 2\mathbb{P}(-\epsilon/2 \leq W \leq \epsilon/2).$$

Proof. We can write the probability on the l.h.s. as

$$\mathbb{P}(l \leq W \leq u) - \mathbb{P}(u - \epsilon \leq W \leq u) - \mathbb{P}(l \leq W \leq l + \epsilon).$$

Because the p.d.f. of a centered Gaussian distribution is symmetric around 0 and is strictly decreasing on the positive reals, it follows that the second and third probability are bounded from above by $\mathbb{P}(-\epsilon/2 \leq W \leq \epsilon/2)$. □

We continue now with the proof of Theorem 3.

Proof. Fix an $\epsilon > 0$. Let $K \subset \mathbb{R}^d$ be compact and set

$$B_1 = \{X \in K\}$$

and

$$B_{2,n} = \left\{ \sup_{(x,t) \in K \times \{-1,1\}} |f_{D_n}(x,t) - f(x,t)| < \epsilon \right\}.$$

Since X is stochastically bounded (it is $\mathbb{R}^d$-valued), we can choose K large enough such that $\mathbb{P}(B_1)$ is arbitrarily close to 1. Because $\mathcal{A}_n$ is a consistent prediction procedure, we can choose n large enough such that $\mathbb{P}(B_{2,n})$ is arbitrarily close to 1. This means, we can choose K and n large enough such that $\mathbb{P}(B_1 \cap B_{2,n})$ is arbitrarily close to 1. Note that proving $\mathbb{P}(\tau(X) \in \Gamma_{D_n}(X)) \geq 1 - \alpha$ as $n \to \infty$ is equivalent to proving that

$$\mathbb{P}\left(\tau(X) \in \Gamma_{D_n}(X) \mid B_1 \cap B_{2,n}\right)$$

is bounded from below by $1 - \alpha$ as $\mathbb{P}(B_1 \cap B_{2,n})$ converges to 1. Observe that we can write the conditional probability as

$$\frac{\mathbb{E}\left[\mathbb{E}[\mathbb{1}_{\{\tau(X) \in \Gamma_{D_n}(X)\}} \mathbb{1}_{B_1 \cap B_{2,n}} \mid X, D_n]\right]}{\mathbb{P}(B_1 \cap B_{2,n})}.$$

Because the event $B_1 \cap B_{2,n}$ only depends on X and D_n (and is therefore measurable with respect to the σ-Algebra generated by X and D_n), the conditional expectation in the numerator is equal to $\mathbb{P}(\tau(X) \in \Gamma_{D_n}(X) | X, D_n) \mathbb{1}_{B_1 \cap B_{2,n}}$. But this implies that the conditional probability is equal to

$$\mathbb{E}\left[\mathbb{P}(\tau(X) \in \Gamma_{D_n}(X) \mid X, D_n) | B_1 \cap B_{2,n}\right]. \tag{8}$$

Observe that

$$\begin{aligned}
\{\tau(X) \in \Gamma_{D_n}(X)\} \;=\; \Big\{ & (\tilde{l}_{D_n}(X,1) - f(X,1)) - (\tilde{u}_{D_n}(X,-1) - f(X,-1)) \\
& \leq\; e_{X,1} - e_{X,-1} \\
& \leq\; (\tilde{u}_{D_n}(X,1) - f(X,1)) - (\tilde{l}_{D_n}(X,-1) - f(X,-1)) \Big\}.
\end{aligned}$$

On the event $B_1 \cap B_{2,n}$, we have $|f(X,t) - f_{D_n}(X,t)| < \epsilon$ for all $t \in \{-1, 1\}$. This means that (8) becomes only smaller if we replace $\{\tau(X) \in \Gamma_{D_n}(X)\}$ by

$$\begin{aligned}
\Big\{ & (\tilde{l}_{D_n}(X,1) - f_{D_n}(X,1)) - (\tilde{u}_{D_n}(X,-1) - f_{D_n}(X,-1)) + 2\epsilon \\
& \leq\; e_{X,1} - e_{X,-1} \\
& \leq\; (\tilde{u}_{D_n}(X,1) - f_{D_n}(X,1)) - (\tilde{l}_{D_n}(X,-1) - f_{D_n}(X,-1)) - 2\epsilon \Big\}.
\end{aligned}$$

By definition of $\tilde{l}_{D_n}(X,t)$ and $\tilde{u}_{D_n}(X,t)$, $t \in \{-1,1\}$, the previous set is equal to

$$\begin{aligned}\Big\{ & -\frac{f_{D_n}(X,1) - l_{D_n}(X,1))}{\sqrt{2}} - \frac{u_{D_n}(X,-1) - f_{D_n}(X,-1)}{\sqrt{2}} + 2\epsilon \\ & \leq\ e_{X,1} - e_{X,-1} \\ & \leq\ \frac{u_{D_n}(X,1) - f_{D_n}(X,1)}{\sqrt{2}} + \frac{f_{D_n}(X,-1) - l_{D_n}(X,-1)}{\sqrt{2}} - 2\epsilon \Big\}.\end{aligned}$$

Observe that conditional on X and D_n, the only random quantities remaining are $e_{X,-1}$ and $e_{X,1}$. For $t \in \{-1,1\}$, we have by assumption that $e_{X,t}$ given X is a centered Gaussian. Because D_n is independent of $e_{X,t}$, the same follows for $e_{X,t}$ given X and D_n. Therefore, Lemma 2 implies that (8) becomes only smaller if we replace the conditional probability in (8) by

$$\begin{aligned}\mathbb{P}\Big(& -\frac{f_{D_n}(X,1) - l_{D_n}(X,1))}{\sqrt{2}} - \frac{u_{D_n}(X,-1) - f_{D_n}(X,-1)}{\sqrt{2}} \\ & \leq\ e_{X,1} - e_{X,-1} \\ & \leq\ \frac{u_{D_n}(X,1) - f_{D_n}(X,1)}{\sqrt{2}} + \frac{f_{D_n}(X,-1) - l_{D_n}(X,-1)}{\sqrt{2}} \mid X, D_n \Big) \\ & + 2\mathbb{P}\left(-\epsilon \leq e_{X,1} - e_{X,-1} \leq \epsilon \mid X, D_n\right).\end{aligned}$$

In view of (4), we have $f_{D_n}(X,t) - l_{D_n}(X,t) \geq 0$ a.s. and $u_{D_n}(X,t) - f_{D_n}(X,t) \geq 0$ a.s. for all $t \in \{-1,1\}$. By assumption, $\epsilon_{X,-1}$ and $\epsilon_{X,1}$, conditional on X, are centered multivariate Gaussian with the same variance and non-negative correlation. Because D_n is independent of $e_{X,t}$, the same follows for $\epsilon_{X,-1}$ and $\epsilon_{X,1}$ given X and D_n. Lemma 1 implies that the first summand in the previous expression is bounded from below by

$$\frac{1}{2} \sum_{t \in \{-1,1\}} \mathbb{P}\left(l_{D_n}(X,t) - f_{D_n}(X,t) \leq e_{X,t} \leq u_{D_n}(X,t) - f_{D_n}(X,t) \mid X, D_n\right).$$

Because $|f(X,t) - f_{D_n}(X,t)| < \epsilon$ on the event $B_1 \cap B_{2,n}$, conditional on $B_1 \cap B_{2,n}$, the previous sum is bounded from below by

$$\begin{aligned}& \frac{1}{2} \sum_{t \in \{-1,1\}} \mathbb{P}\left(l_{D_n}(X,t) - f(X,t) \leq e_{X,t} \leq u_{D_n}(X,t) - f(X,t) \mid X, D_n\right) \\ & - \sum_{t \in \{-1,1\}} \mathbb{P}\left(-\epsilon/2 \leq e_{X,t} \leq \epsilon/2 \mid X, D_n\right).\end{aligned}$$

(We replace the first $f_{D_n}(X,t)$ by $f(X,t)-\epsilon$ and the second by $f(X,t)+\epsilon$ and then apply Lemma 2.) Using the same argument as in the paragraph above (8) (only in the reverse direction), we can conclude that (8) is bounded from below by

$$\begin{aligned}
&\frac{1}{2}\sum_{t\in\{-1,1\}}\mathbb{P}\left(l_{D_n}(X,t)-f(X,t)\leq e_{X,t}\leq u_{D_n}(X,t)-f(X,t)\mid B_1\cap B_{2,n}\right)\\
&-\sum_{t\in\{-1,1\}}\mathbb{P}\left(-\epsilon/2\leq e_{X,t}\leq\epsilon/2\ \mid B_1\cap B_{2,n}\right)\\
&-2\mathbb{P}\left(-\epsilon\leq e_{X,1}-e_{X,-1}\leq\epsilon\ \mid B_1\cap B_{2,n}\right).
\end{aligned}$$

Since we want to show that this is bounded from below by $1-\alpha$ as $\mathbb{P}(B_1\cap B_{2,n})$ converges to 1, it is equivalent to showing that

$$\begin{aligned}
&\frac{1}{2}\sum_{t\in\{-1,1\}}\mathbb{P}\left(l_{D_n}(X,t)-f(X,t)\leq e_{X,t}\leq u_{D_n}(X,t)-f(X,t)\right)\\
&-\sum_{t\in\{-1,1\}}\mathbb{P}\left(-\epsilon/2\leq e_{X,t}\leq\epsilon/2\right)\\
&-2\mathbb{P}\left(-\epsilon\leq e_{X,1}-e_{X,-1}\leq\epsilon\right).
\end{aligned}$$

is bounded from below by $1-\alpha$ as $n\to\infty$. Observe that the first summand is equal to

$$\frac{1}{2}\sum_{t\in\{-1,1\}}\mathbb{P}\left(Y\in[l_{D_n}(X,t),u_{D_n}(X,t)]\mid T=t\right)$$

which is bounded from below by $1-\alpha$ for every n by assumption. Since $(e_{X,-1},e_{X,1})'$ conditional on X is multivariate Gaussian, it follows that $(e_{X,-1},e_{X,1})'$ has no point mass unconditionally. Because ϵ was arbitrary, we can make the second and third summand arbitrarily small.

□

A.3 Proofs of Theorem 4 and 5

Proof of Theorem 4. Set

$$N_t=\sum_{i=1}^{n}\mathbb{1}_{\{T_i=t\}}.$$

Observe that N_t is Binom(n, p_t)-distributed. Conditional on $\{N_t, T = t\}$, the set $D_{n+1,T}(Y)$ has cardinality $N_t + 1$ and the conditional joint distribution of the random variables in $m_{D_{n+1}(Y)}(D_{n+1,T}(Y))$ is exchangeable. This implies that the conditional distribution of $R_{D_{n+1}(Y)}(Y)$ is a discrete uniform distribution on $\{1, \ldots, N_t + 1\}$. Hence, it follows that

$$\begin{aligned} 1 - \alpha &\leq \frac{\lceil (1-\alpha)(N_t+1) \rceil}{N_t + 1} \\ &= \mathbb{P}\left(R_{D_{n+1}(Y)}(Y) \leq \lceil (1-\alpha)(N_t+1) \rceil \mid N_t, T = t \right) \\ &\leq 1 - \alpha + \frac{1}{N_t + 1}. \end{aligned}$$

This means that the conditional probability

$$\mathbb{P}\left(R_{D_{n+1}(Y)}(Y) \leq \lceil (1-\alpha) N_t \rceil \mid T = t \right)$$

is also bounded from below by $1-\alpha$. For the upper bound, we need to compute the first moment of $(N_t + 1)^{-1}$, where N_t is Binom(n, p_t)-distributed. Chao and Strawderman (1972) showed that this is equal to $(1 - (1 - p_t)^{n+1})/((n + 1)p_t)$. □

Proof of Theorem 5. Set

$$\tilde{N}_t = \sum_{i=m+1}^{n} \mathbb{1}_{\{T_i = t\}}.$$

Observe that $\tilde{N}_t$ is Binom$(n - m, p_t)$-distributed. Conditional on $\{\tilde{N}_t, T = t\}$, the set $D_{m:(n+1),T}(Y)$ has cardinality $\tilde{N}_t + 1$ and the random variables in $m_{D_m}(D_{m:(n+1),T}(Y))$ are exchangeable because the dataset D_m is independent of the random variables in $D_{m:(n+1),T}(Y)$). This implies that the conditional distribution of $R_{D_m}(Y)$ is a discrete uniform on $\{1, \ldots, \tilde{N}_t + 1\}$ and therefore

$$\begin{aligned} 1 - \alpha &\leq \frac{\lceil (1-\alpha)(\tilde{N}_t+1) \rceil}{\tilde{N}_t + 1} \\ &= \mathbb{P}\left(R_{D_m}(Y) \leq \lceil (1-\alpha)(\tilde{N}_t+1) \rceil \mid \tilde{N}_t, T = t \right) \\ &\leq 1 - \alpha + \frac{1}{(\tilde{N}_t + 1)}. \end{aligned}$$

The claim now follows by the same argument as in the proof of Theorem 4. □

References

Ballarini, N. M., Rosenkranz, G. K., Jaki, T., König, F., and Posch, M. (2018). Subgroup identification in clinical trials via the predicted individual treatment effect. *PloS one*, 13(10).

Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285.

Chao, M. T. and Strawderman, W. E. (1972). Negative moments of positive random variables. *Journal of the American Statistical Association*, 67(338):429–431.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Graf, A. C., Magirr, D., Dmitrienko, A., and Posch, M. (2020). Optimized multiple testing procedures for nested sub-populations based on a continuous biomarker. *Statistical Methods in Medical Research*, page 0962280220913071.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366.

Kivaranovic, D., Johnson, K. D., and Leeb, H. (2019). Adaptive, distribution-free prediction intervals for deep neural networks. *arXiv preprint arXiv:1905.10634*.

Lamont, A., Lyons, M. D., Jaki, T., Stuart, E., Feaster, D. J., Tharmaratnam, K., Oberski, D., Ishwaran, H., Wilson, D. K., and Van Horn, M. L. (2018). Identification of predicted individual treatment effects in randomized clinical trials. *Statistical methods in medical research*, 27(1):142–157.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Pellegrini, F., Copetti, M., Bovis, F., Cheng, D., Hyde, R., de Moor, C., Kieseier, B. C., and Sormani, M. P. (2019). A proof-of-concept application of a novel scoring approach for personalized medicine in multiple sclerosis. *Multiple Sclerosis Journal*, page 1352458519849513.

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, pages 3538–3548.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Senn, S. (2016). Mastering variation: variance components and personalised medicine. *Statistics in medicine*, 35(7):966–977.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3):371–421.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.