Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 2421

Robust inference for systems under distribution shifts

LUDVIG HULT

ACTA UNIVERSITATIS
UPSALIENSIS
2024

ISSN 1651-6214
ISBN 978-91-513-2178-3
urn:nbn:se:uu:diva-533683

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in 10134, Polhemsalen, Lägerhyddevägen 1, Uppsala, Friday, 20 September 2024 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Antoine Chambaz (Université Paris Cité).

Abstract

Hult, L. 2024. Robust inference for systems under distribution shifts. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2421. 45 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2178-3.

We use statistics and machine learning to make advanced inferences from data. Challenges may arise, invalidating inferences, if the context changes. Situations where the data generating process changes from one context to another is known as distribution shift, and may arise for several reasons. This thesis presents five articles on the topic of making robust inferences in the presence of distribution shifts.

Paper 1 to 3 develop mathematical methods for robust inference. Paper 1 adresses the problem that when there is uncertainty about the structue of the underlying data generating process, confidence intervals are not generally valid for estimating the impact of interventions. We propose a method for constructing valid confidence intervals for the average treatment effect using linear structural causal models. Paper 2 addresses the problem of model evaluation under distribution shift, using nonparametric statistics. We show that with a small validation sample, one can make finite-samplevalid inference about a machine learning model performance on a new data set despite distribution shift. Paper 3 addresses the problem that inventory control policies may become invalid without assumptions on the demand. Using a deterministic feedback mechanism, we construct an order policy that guarantees any prescribed service level, with weak assumptions on the demand, allowing distribution shift.

Paper 4 and 5 focus on applications to neurocritical care data. Paper 4 uses machine learning to predict intracranial pressure insults in neurocritical care. Since distribution shift may occur between patients and/or years, the validation methods takes this into account. Paper 5 explores the use of causal inference on neurointensive care data. While this may eventually lead to inferences valid under intervention distribution shift, several obstacles to effective application are identified and discussed.

Ludvig Hult, Department of Information Technology, Division of Systems and Control, Box 337, Uppsala University, SE-75105 Uppsala, Sweden. Department of Information Technology, Artificial Intelligence, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-2178-3
URN urn:nbn:se:uu:diva-533683 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-533683)

Dedicated to Sara, with whom I share my life

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Inference of causal effects when control variables are unknown** *Ludvig Hult, Dave Zachariah* Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, PMLR 161:1300-1309, 2021.

II **Diagnostic Tool for Out-of-Sample Model Evaluation** *Ludvig Hult, Dave Zachariah, Petre Stoica* Transactions in Machine Learning 2023

III **Certified Inventory Control of Critical Resources** *Ludvig Hult, Dave Zachariah, Petre Stoica* arXiv:2405.1505. Under review.

IV **Machine learning based prediction of imminent ICP insults during neurocritical care of traumatic brain injury** *Peter Galos, Ludvig Hult, Dave Zachariah, Anders Lewén, Anders Hånell, Timothy Howells, Thomas Schön, Per Enblad* Under review.

V **Challenges in CATE estimation for neurointensive care** *Ludvig Hult, Peter Galos, Per Enblad, Dave Zachariah* Manuscript.

Contents

Notation

$\mathbf{1}\{A\}$	Indicator; 1 if A is true, 0 otherwise
$\mathbb{E}[X]$	Expectation of X
$\mathbb{E}[X\|Y]$	Conditional expectation of X given Y
$\mathbb{P}[A]$	Probability of A
$\mathbb{P}[A\|B]$	Conditional probability of A given B
i.i.d.	Independent and identically distributed
$X \perp Y$	Independence of X and Y
$X \perp Y \,\|Z$	Conditional independence of X and Y, given Z
$(X)^+$	$\max\{X, 0\}$
$(X)^-$	$\max\{-X, 0\}$
$\xrightarrow{d}$	Convergence in distribution
$\xrightarrow{p}$	Convergence in probability

1. Introduction

1.1 Motivation

Not a day goes by without you making decisions, choosing between different options, all with unknown outcomes. Some decisions are less severe, like what to eat for dinner. Others have larger consequences, like what school to put your child in. We employ our experiences and judgement, and it often leads us right. For small-consequence decisions we don't think more of it, but when the stakes are high we employ as much rigor as we can to form a well-founded decision.

Systematizing the process of understanding the consequences of decisions and actions is a central tenet of science, with experimentation and controlling all influencing variables as the main tool. Randomness in outcomes may still be present, and the development of statistics in the 20th century allowed researchers to distinguish between random variation and actual effects. If the conditions change from the experimental context to an applied one, the conclusions are still valid. The laws of electromagnetism govern on the moon even if they are derived from data on earth. Experiments help us find patterns in data that generalize from experiment conditions to new contexts.

Sometimes experimentation is not feasible, e.g., by being unethical or impractical. If we worry that smoking causes cancer, should we really randomize people to heavy smoking? If catching COVID-19 increases the risk of cancer over a lifetime, should we wait 100 years to see who gets cancer before attempting to draw conclusions? Observational data or imperfect experimental data may be available. Causal inference enables conclusions from such data that do generalize correctly. The problem can also be broadened. Generally, if the conditions are different between the context in which we make inference and the context in which we want to apply them, we call it distribution shift.

This thesis explores theoretical aspects and practical implementations of methods of causal inference and distribution shift in machine learning.

1.2 Outline

This thesis consists of several research papers and a comprehensive summary that binds them together. The purpose of the summary is to provide

the terms and definitions needed to approach the papers. A brief summary of each paper is presented in Section 1.3. Chapter 2 introduces how we model decision making from data using probability, causal models, and deterministic feedback control. Chapter 3 introduces terminology and concepts from machine learning. Chapter 4 describes the problem of distribution shifts and how it can be formalized. Finally, Chapter 5 describes how we utilize various techniques to make decisions from data in the presence of distribution shift.

Throughout the summary, connections to the papers are made so that key concepts are related to the research papers as they are introduced.

1.3 Contributions

This dissertation contains five papers with contributions to the research field. Paper I, II and III concerns methodological advances. Paper IV and Paper V are applied work on inference in neurocritical care data.

Paper I: Inference of causal effects when control variables are unknown

Summary: This paper presents methods for estimating a statistical parameter that generalizes the Average Treatment Effect (ATE). Estimating the ATE under standard conditions with linear models requires knowledge of control variables. If this is not known a priori, few methods can provide valid confidence intervals for the parameter of interest. This paper utilizes a smooth characterization of graph constraints to estimate a set of control variables and provide a valid confidence interval.

Contribution: Dave proposed the idea of leveraging the smooth constraints for acyclic graphs to obtain confidence intervals of the ATE. We developed this together, with me formulating the proofs, the numerical experiments, the optimization algorithm, and the benchmarks. I wrote the draft, and Dave supervised the writing and the editing.

Paper II: Diagnostic Tool for Out-of-Sample Model Evaluation

Summary: In this paper, we posit the statistical measure, the 'level-alpha-loss' or 'LAL'. This measure can be computed from a trained machine learning model and a dataset exchangeable with out-of-sample test data and provide a probabilistic upper bound on losses for out-of-sample test data. We show that the measure is similar to other established metrics for characterization of either statistical models (via the complementary cumulative distribution function) or distributions (via the value-at-risk).

Contribution: Dave proposed the idea to use conformal prediction as a post-hoc analysis tool for machine learning models. After preliminary investigations from my side, we formulated the LAL measure together,

and I investigated its properties. I proposed the numerical experiments and did the draft writing. Dave supported with editing and writing. Petre provided feedback on the writing.

Paper III: Certified Inventory Control of Critical Resources
Summary: This paper study inventory control. Decisions to purchase stock are based on the unknown future demand, and policies are typically developed with models that require strong assumptions to be valid. We show that by use of feedback control, any such model can be provided with an adaptive safety margin to provide service level guarantees valid under weaker assumptions. The guarantee is valid even under arbitrary distribution shift. This kind of conservative policy may incur excess costs, so the method also allows estimation of future cost with a similar guarantee.

Contribution: Dave proposed to investigate theory with weak assumptions to tackle a critical stock inventory problem. I did the investigations, exploring conformal inference, semiparametric modeling and dynamic programming under distributional uncertainty before setting for this non-random approach. The proofs, derivations, and numerical experiments are mine. Dave provided feedback on presentation, structuring of proofs, writing, and editing. Petre provided feedback on the writing.

Paper IV: Machine learning based prediction of imminent ICP insults during neurocritical care of traumatic brain injury
Summary: This paper is an improvement on existing machine learning methodology in neurocritical care for patients with traumatic brain injury. The lead time from symptoms of suspected intracranial hypertension to requiring acute care can be very short, motivating the interest in machine learning prediction based alarm systems. We improved the state of the art by showing how methods with simple implementation (e.g., XGBoost) provides performance on par with customized Gaussian processes requiring larger development implementation efforts. To improve the robustness of conclusions with respect to distribution shifts, the model performance was quantified by grouped cross validation and prevalence invariant metrics.

Contribution: Peter held the project together, ensured permits for ethics and compute access, and did much of the writing and editing. I did the data quality management, performed all coding, selected models and training methods, the evaluation metrics and contributed in writing. The idea of using grouped cross validation and focusing on prevalence invariant performance metrics were mine. Per supervised the writing and editing. Anders L, Anders H and Timothy provided access to data. Thomas and Dave provided feedback on the analysis methodology.

Paper V: Challenges in CATE estimation for neurocritical care

Summary: This paper investigates estimation of conditional average treatment effect (CATE) to neurocritical care data, with the purpose of assessing readiness of methods and datasets. From the clinical point of view, we find challenges in lacking overlap, outcome definitions, and small sample sizes, highlighting the need for more systematic data collection and established criteria to define the study cohort and outcomes. From the data analytical point of view, more effort is needed in developing time series models that manage missing data, practical transfer learning for CATE estimation, and algorithms with valid confidence intervals for very high dimensional data.

Contribution: The idea to use treatment information for causal analysis of neurocritical care data came up in discussions between me, Dave and Peter. I formulated the research question in discussions with Dave, did the investigations, model selection, data cleaning and writing. Peter provided data access. Per and Peter both provided feedback on the writing from the clinical perspective. Dave provided feedback on editing and writing, as well as supervision in the investigations.

2. Decision Making from Data

When making decisions based on observations in the world, we must deal with randomness. Measurements are noisy and outcomes are unreliable. We primarily use statistics and probability theory to mathematically formalize random events and how to make inference from observations. Some concepts in causal inference are difficult to formalize using statistics and probability alone, leading to the concept of a structural causal model. Finally, there are cases where we may rid ourselves of modeling the randomness, and still take appropriate decisions from data of an unknown nature. These three paradigms are the focus of this chapter.

We observe data generated by some process. It can be the outcome of the real world, such as rolling physical dice or measure the temperature outside. It can be the outcome of a parametric simulation, such as running Monte Carlo simulations. It may or may not have a simple mathematical description. In all cases, we call this the data generating process (DGP).

2.1 Statistics and Probability

This section describes the key terms and concepts in statistical inference, and how assumptions about the DGP affect how we can draw conclusions from data.

Any DGP can be described by a probability distribution P. In many cases P is not known. We define a statistical estimand to be a function $\tau(\cdot)$ of that distribution that represents some property of interest. We have access to data $\mathcal{D}$ generated by P, and we want to estimate $\tau(P)$ from $\mathcal{D}$. An estimator $\hat{\tau}(\cdot)$ is a function that maps data to an estimate $\hat{\tau}(\mathcal{D})$. These concepts are illustrated in Figure 2.1. The estimate may be either a point estimate or an interval (or a set more generally). Three statistical intervals (or sets) relates to this thesis.

Definition 1 (Statistical intervals [38])**.** A valid *confidence interval* $C(\mathcal{D})$ contains the estimand τ with a probability $1 - \alpha$;

$$\mathbb{P}[\tau \in C(\mathcal{D})] \geq 1 - \alpha. \tag{2.1}$$

When data comes in pairs $Z = (X, Y)$, a valid *prediction interval* $C(\mathcal{D}; X)$ at level α if it fulfills

$$\mathbb{P}[Y \in C(\mathcal{D}; X)] \geq 1 - \alpha. \tag{2.2}$$

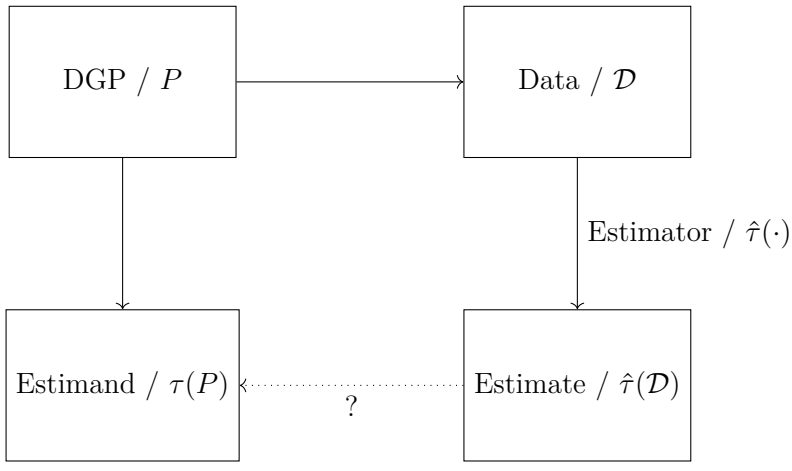


Figure 2.1. Central concepts in statistical inference. Statistical theory guides us on how to construct estimators that describe the estimand in a reliable way.

A valid *tolerance interval* $C(\mathcal{D})$ is such that at a fraction $1-\beta$ of the distribution of future observations Z should be contained in the interval with probability at least $1-\alpha$.

$$\mathbb{P}\left\{\mathbb{P}[Z \in C(\mathcal{D})|\mathcal{D}] \geq 1-\beta\right\} \geq 1-\alpha. \tag{2.3}$$

Example 1. A prototypical case is when P describes n i.i.d. samples from a Gaussian distribution with unknown mean μ and known variance σ^2; $\mathcal{D} = (X_i)_{i=1}^n$, $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$. The estimand is the unknown mean, $\tau(P) = \mu$. One point estimate is the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$. It is unbiased ($\mathbb{E}[\hat{\mu}] = \mu$) and consistent ($\hat{\mu} \xrightarrow{p} \mu$ as $n \to \infty$). A valid confidence interval can be constructed as $\hat{\mu} \pm z_{1-\alpha/2}\sigma^2$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

We can use these concepts for decision making. The DGP is a mechanism that we are interested in and that describes what happens when we make decisions. The estimand describes a quantity of interest related to the decision, and informs us whether we should take this decision. Since the DGP is unknown, we estimate the estimand $\tau(P)$ from data. The whole process is called making inference.

Example 2. You are invited to gamble some money. It costs 100 SEK to play. You roll a six-sided die, and if you get a six, you get 700 SEK back. Otherwise, you get nothing. The person offering you the game seems dubious, and you suspect the die is loaded. The data generating process means rolling the die, taking 6 to mean a win ($X = 1$), and otherwise it is a loss ($X = 0$). You decide that if the chance of winning is at least 20%,

you will play the game. The win rate is your estimand, $\tau(P) = \mathbb{E}[X]$. Some friends of yours with money to spare play the game 20 times and win 4 or them. This is your data set $\mathcal{D}$. You compute a 95% confidence interval to be $[0\%, 40\%]$. Based on this, you are not convinced that the win rate is high enough, and you refrain from playing.

The estimand may be a parameter of a parametric distribution, as was the case in Example 1 and 2. This assumes that the DGP is a member of that parametric family, which may be a strong assumption, and any inference based on that assumption will be invalid if the assumption does not hold. We need methods that apply more generally. This motivates estimands that are not based on narrow distributional assumptions [5, 62, 65]. We use this idea in Paper I when we define the estimand γ as the best linear approximation of change in mean outcome for a variable Y as we intervene on X. The quantity is well defined even if the DGP is not linear.

A prominent example of assumtion lean inference is M-estimation [61, 66]. The estimand be defined as a minimizer of a loss function ℓ that depends on a data point Z and a parameter of interest θ,

$$\tau(P) = \arg\min_{\theta} \mathbb{E}[\ell(\theta, Z)]\,. \tag{2.4}$$

If observing n i.i.d. data points drawn from P, under weak assumptions, the estimate

$$\hat{\tau}(\mathcal{D}) = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, x_i) \tag{2.5}$$

will be asymptotically normally distributed. M-estimation is the method for inference behind Paper I.

Example 3. M-estimation can change how we can think about Example 1. The estimand is the population mean, and we can define it as the minimizer of the squared loss function,

$$\tau(P) = \mathbb{E}[X] = \arg\min_{\mu} \mathbb{E}\Big[(\mu - X)^2\Big]\,. \tag{2.6}$$

The point estimate is still the sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The confidence interval is constructed as $\hat{\mu} \pm z_{1-\alpha/2}\hat{\sigma}^2$ where $\hat{\sigma}^2$ is the sample variance. We have won something since we can relax the assumption of normality for P, but we lost something since the results are now only asymptotically valid.

2.2 Causal Models

Causal models formalize not only the random process that generate data, but also encode how it would change if interventions were made. Two major frameworks are in use. The Potential Outcomes by Donald Rubin and the Structural Causal Models of Judea Pearl. Since both frameworks are used in the papers of this thesis, they are presented here.

Potential Outcomes

The Potential Outcomes framework is a framework of causality within mathematical probability.[26, 29] From it, sufficient conditions for drawing causal conclusions from data can be put forward. Here we present a special case.

Consider a DGP that under observation generates i.i.d. observations of random triplets (X, Y, T) with background variables X, outcomes Y and binary treatments T. There are also the interventional random variables $Y(1)$ and $Y(0)$, describing the outcomes that would be observed if we somehow intervened on T to make it 0 or 1. They are called the *potential outcomes*. Under this framework, we can define estimands of causal interest. The linear contrast is a standard individual treatment effect (ITE) measure,

$$\text{ITE} = Y(1) - Y(0), \tag{2.7}$$

which is a random variable itself. Other ITE definitions, such as $Y(1)/Y(0)$ are also thinkable, and the choice is contextual and application dependant. We will only consider the linear contrast. The conditional average treatment effect

$$\text{CATE}\,(x) = \mathbb{E}[Y(1) - Y(0)|X = x] \tag{2.8}$$

and the average treatment effect

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] \tag{2.9}$$

are the causal estimands of this thesis.

Example 4. Consider whether you should get a vaccine for the seasonal strain of influenza. The DGP is a combination of patient characteristics X, whether the vaccine were taken or not T, and whether the patient got influenza or not Y. The CATE describes the average change in infection rates for patients with certain characteristics $X = x$ if they take the vaccine or not. The ATE describes the average effect on the population as a whole.

With these estimands in place, we may ask how to estimate them from data. Neither $Y(1)$ nor $Y(0)$ are observed when no intervention is made,

and even if we make interventions, only one of the two can be observed at a time. This is the fundamental problem of causal inference, and is sometimes phrased as causal inference being a "missing data problem". Appropriate assumptions are needed.

Definition 2. The DGP has interventional *consistency* if $Y = Y(T)$, the observed outcome is the same as the potential outcome for the observed treatment.

The DGP has a *conditionally ignorable treatment assignment* if when X is conditioned on, it does not matter for the interventional outcomes how the treatment was assigned. I.e. $Y(0), Y(1) \perp T \,|\, X$. The property is also called conditional ignorability. The variables X are called a valid set of control variables, or an adjustment set.

The DGP has interventional *positivity* if T is not deterministic, i.e. $0 < \mathbb{P}[T = 1|X = x] < 1$ for all x.

The three properies of Definition 2 are sufficient for identifiability of the causal estimands. The proof is direct and instructive.

Theorem 1. *Under consistency, positivity and conditionally ignorable treatment assignment, the CATE and the ATE are identifiable from observational data.*

Proof. First we use the conditional ignorability and consistency to rewrite the CATE as

$$\begin{aligned} \text{CATE}\,(x) = \mathbb{E}[Y(1)|X = x, T = 1] - \mathbb{E}[Y(0)|X = x, T = 0] = \\ \mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x, T = 0]\,. \end{aligned} \tag{2.10}$$

These quantities can be estimated from data using the law of large numbers assuming that both $T = 1$ and $T = 0$ are observed for all $X = x$, i.e. under positivity.[1] The ATE is obtained by marginalizing the estimated CATE over X.

$$\text{ATE} = \mathbb{E}[\text{CATE}\,(X)] = \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0] \tag{2.11}$$

□

The assumptions of positivity, consistency, and conditional ignorability are strong, and do not always hold. In Paper V we assume that they hold approximately, and investigate whether we obtain useful estimates of the CATE.

[1] The conditional probability $\mathbb{E}[Y|X = x, T = t]$ is not even defined when positivity is violated, so the characterization of the CATE as $\text{CATE}\,(x) = \mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x, T = 0]$ would be meaningless.

Structural Causal Models

Structural Causal Models (SCM) is another framework for drawing causal conclusions [44, 46, 56]. SCMs have also been used for reasoning, and has a connection to the philosophical study of causality [22, 23, 11].

A structural causal model is defined by the triplet $\mathcal{S} = \langle \mathcal{V}, \mathcal{E}, F \rangle$, consisting of a set of independent random variables $\mathcal{E} = \{e_1, ..., e_n\}$, symbols $\mathcal{V} = \{v_1, ..., v_n\}$, and for each such symbol, functions f_j with "causal parents" $\text{pa}_j \subset \mathcal{V} \setminus \{v_j\}$ encoded as $F = \left\{ (f_j, \text{pa}_j) \middle| j = 1, ..., n \right\}$. The model is finally obtained as any solution to the simultaneous system of equations

$$v_j = f_j(\text{pa}_j, e_j), \quad \forall j \in \{1, ..., n\} \tag{2.12}$$

such that $\mathcal{V}$ are random variables. Such a solution means that $\mathcal{S}$ also induces a probability distribution P over $\mathcal{V}$. The model $\mathcal{S}$ also defines a graph $\mathcal{G}$ with nodes $\mathcal{V}$ and an edge from v_i to v_j if $v_i \in \text{pa}_j$. A sufficient (but not necessary) criterion for P to be defined is that $\mathcal{G}$ is a directed and acyclic graph (DAG) [44].

One purpose with SCMs is to facilitate causal inference. To model simple interventions in a SCM, we use so called atomic do-interventions [44]. An atomic do-intervention on a vertex v_i means replacing the function f_i with a constant c_i. This also changes the parental set, so that $\text{pa}_i = \emptyset$. This new model induces a new distribution, $P^{do(v_i = c_i)}$, which is the interventional distribution.

The arguably strongest assumption of causal inference is the assumption of conditional ignorability. Under the SCM framework, this assumption can be verified from the causal graph $\mathcal{G}$ algorithmically. By assuming the causal graph $\mathcal{G}$, applied researchers can derive assumptions about the distribution P that are needed for causal identification. This is done in Paper V. If the graph is not known, *causal discovery* is the process of estimating it from data. Traditional methods are based on hypothesis testing and combinatorial search [56] which does not lend itself to uncertainty quantification. In Paper I we obtain valid confidence intervals for the ATE, using the SCM framework and a smooth characterization of the graphical constraints.

2.3 Non-random Alternatives

Automatic Control is explicitly concerned with automated decision making based on data. One formulation in discrete time is as follows [7]. A dynamical system is at time t in state X_t. You can take a control decision U_t. There are uncontrollable inputs to the system at the same time, W_t, called process noise. The system evolves according to a function f. The

initial state is X_0.

$$X_{t+1} = f(X_t, U_t, W_t) \quad t = 1, ..., T \tag{2.13}$$

We want to find a control law $U_t = \pi(X_t)$ that minimizes some cost functional J over a time horizon T. The cost function at time t is $g_t(X_t, U_t, W_t)$. To account for the fact that W_t might be modelled as random, the cost functional for finite time control is

$$J = \sum_{t=0}^{T-1} \mathbb{E}[g_t(X_t, U_t, W_t)] \tag{2.14}$$

and the optimal control policy is

$$\hat{\pi} = \arg\min_{\pi} J. \tag{2.15}$$

By adjusting the cost functional, we can impose constraints on the policy or on the state trajectory. E.g., $g_t(X_t, U_t, W_t) = \frac{1}{X_{t+1} - C}$ incurs arbitrarily high costs as X_{t+1} approaches C, effectively excluding all policies π for which there is a nonzero probability that $X_{t+1} = C$. To obtain J, despite the unknown process noise ($W_t \sim P_t$), one needs to make modeling assumptions, estimate the system from data, or do both.

Example 5 (The Newsvendor problem)**.** Let X_t be the number of newspaper in stock on a morning. Let W_t be the number of newspapers in demand. All items not sold are discarded at the end of day. The transition function is

$$X_{t+1} = 0. \tag{2.16}$$

The cost each day depends on the number of papers bought, sold, and the unfulfilled demand

$$g_t(X_t, U_t, W_t) = cU_t + h\left(W_t - U_t\right)^+ + p\left(W_t - U_t\right)^- . \tag{2.17}$$

Recent developments concerns optimal policy selection under the uncertainty from estimating the demand [34] Inventory control problems like the one in Paper III is a natural extension to this problem where $X_{t+1} = (X_t - W_t + U_t)^+$.

One may instead consider W_t unknown but deterministic. This would provide a non-random modeling approach. Any guarantee about the cost or the state trajectory would be valid for any noise process W_t, even those that are random with distribution shifts or feedback. This framework has been used in combination with statistical estimation [67, 2]. We use it in Paper III as a means to provide guarantees for inventory control valid under distribution shift.

3. Machine Learning Fundamentals

Exact definitions of machine learning varies between sources, but they often revolve around computer programs that solves a task, and improves on that tasks by processing data [35, 40]. Common tasks are to predict a future outcome, estimate a value or make a decision. One distinction between machine learning and statistics is that machine learning is generally not interested in finding a "true" statement about the world (such as inferring an estimand), but rather to find a model that performs well on the task given. In practice, these two goals often overlap.

In this chapter, we introduce terminology in machine learning and discuss how we measure performance of machine learning models.

3.1 Supervised Machine Learning

Supervised machine learning is arguably the largest branch of machine learning. The task is to learn a parametric function f_θ. We measure the performance of the function on a data point (X, y) using a loss $\ell(f_\theta(X), y)$, on new data where smaller is better. We call the inputs X *covariates* or *features*, and y are called *labels*.

We learn the specific values of the parameters $\hat{\theta}$ from a dataset $\mathcal{D}_{\text{train}} = \{(X_i, y_i)\}_{i=1}^n$, using a learning algorithm $\mathcal{A}$

$$\hat{\theta} = \mathcal{A}\left(\mathcal{D}_{\text{train}}\right). \tag{3.1}$$

The Empirical Risk Minimizer algorithm

$$\mathcal{A}\left(\mathcal{D}_{\text{train}}\right) = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(X_i), y_i) \tag{3.2}$$

provides a common example of a learning algorithm, and also shows a clear connection to the M-estimation framework of (2.5).

Example 6. Ordinary Least Squares is Empirical Risk Minimization. Take $y \in \mathbb{R}$, $X \in \mathbb{R}^d$, $\Theta = \mathbb{R}^d$. The loss is $\ell(f_\theta(X), y) = |z - \theta^\top X|^2$. $f_\theta(X) = \theta^\top X$ If $\mathbf{y}^\top = (y_1, ..., y_n)$, $\mathbf{X}^\top = (X_1, ..., X_n)$, the minimizer is given by the standard formula $\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$.

Other learning algorithms used in this thesis are regularized empirical risk minimizers (e.g. LASSO regression), the Random Forest algorithm, Extreme Gradient Boosting, Gaussian Processes and Neural Networks optimized by Adam and early stopping.

3.2 Out-of-sample Performance Assessments

When learning the parameters θ, we only have access to a sample of data. A key questions is to understand the performance of the model on new data, or *out-of-sample*.

When the out-of-sample data is observed from a DGP described by a distribution P_{test}, the performance of the function is commonly quantified by the expected loss, called the *risk*

$$R(\theta) = \mathbb{E}_{P_{\text{test}}}[\ell(f_\theta(X), y)]\,. \tag{3.3}$$

If the training data $\mathcal{D}$ are i.i.d. samples from P_{test}, and regularity constraints hold, the M-estimation framework of section 2.1 applies, and the learned function will converge to the minimum-risk function. If there is no relation between $\mathcal{D}_{\text{train}}$ and P_{test}, we cannot say anything certain which implies that we must either observe data from P_{test} or some prior knowledge in order to make any statements about the out-of-sample performance.

Hold-out data

The recommended way to estimate out-of-sample performance is to emulate it. Take a dataset $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}$. If the data in $\mathcal{D}_{\text{test}}$ is sampled from P_{test} independently of the $\mathcal{D}_{\text{train}}$, the empirical risk on $\mathcal{D}_{\text{test}}$ is an unbiased estimate of the risk. This is the hold-out method.

$$\hat{R}^{HO}(\theta) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(X_i,y_i)\in\mathcal{D}_{\text{test}}} \ell(f_{\hat{\theta}}(X_i), y_i) \tag{3.4}$$

One early paper on the topic traces the lineage back to the 1930s, and displays the methods efficacy on various statistical tasks such as estimation and prediction [57].

Cross validation

The term Cross Validation today often refers to K-fold Cross Validation algorithm. There is several extensions for it [4, 16]. It can itself be seen as an extension of the hold-out data method above. The basic premise is described below.

The K-fold cross validation algorithm is to take a dataset of K disjoint sub-datasets, $\mathcal{D} = \cup_{k=1}^{K}\mathcal{D}_k$. For each $j = 1, ..., K$, learn a model on $K-1$ subsets $\mathcal{D}^{(-j)} = \cup_{k\neq j}\mathcal{D}_k$ and compute the hold-out risk the j'th subset $\mathcal{D}_j$.

$$\hat{R}_j = \frac{1}{|\mathcal{D}_j|} \sum_{(X_i,y_i)\in\mathcal{D}_j} \ell(f_{\mathcal{A}(\mathcal{D}^{(-j)})}(X_i), y_i). \tag{3.5}$$

We define the K-fold cross validation to be the mean of the hold out risks

$$\hat{R}^{CV} = \frac{1}{K}\sum_{j=1}^{K} \hat{R}_j \tag{3.6}$$

If $|\mathcal{D}_j| = N$ for all j, and $(X, y) \overset{i.i.d.}{\sim} P_{\text{test}}$ for all $(X, y) \in \mathcal{D}$, the estimate $\hat{R}^{CV}$ is unbiased for the risk for a model trained on $(K-1)N$ data points.

Sometimes the held-out data and the out-of-sample data are not independent e.g., when using machine learning to estimate outcome for rare diseases and the population of test subjects is small. Risk estimation via K-fold cross validation for would not work in such cases.

Tail bound on the risk

Sometimes, it is more important to bound the tail probability of the loss, rather than the expected loss. This is captured by the value-at-risk [30]

$$\text{VaR}_\alpha(\mathcal{D}) = \inf_t \{t \,|\, \mathbb{P}[\ell(f_\theta(X), y) \leq t | \mathcal{D}] \geq 1 - \alpha\}. \tag{3.7}$$

If the distributions over the loss is continuous, this is exactly the $1 - \alpha$ quantile of the loss distribution.

In Paper IV we posit a similar measure, the level-alpha-loss, or LAL. In the special case of observing a single new data point sampled from P_{test}, the LAL is defined to obey

$$\mathbb{P}\Big[\ell(f_\theta(X), y) \leq \text{LAL}_\alpha^\beta(\mathcal{D}_{\text{train}})\Big] \geq 1 - \alpha \tag{3.8}$$

where the VaR instead obeys

$$\mathbb{P}[\ell(f_\theta(X), y) \leq \text{VaR}_\alpha(\mathcal{D}_{\text{train}}) | \mathcal{D}_{\text{train}}] \geq 1 - \alpha. \tag{3.9}$$

Three points clarify the distinction. First, the guarantee of the VaR is conditional on the training data $\mathcal{D}_{\text{train}}$, whereas the guarantee for LAL is unconditional. Because of this, we have a general formula for computing it, which is not available for VaR. Secondly, the LAL is generalized to control the tail losses for multiple observations, not just a single one. Similarly to how a tolerance interval generalize the prediction interval to a set of future observations, the LAL generalize the VaR to a set of future observations. Thirdly, we suggest that the LAL should be considered for a range of α values, and not just a single one, leading to the idea of the LAL curve.

4. Describing Distribution Shifts

Decisions are taken for the future based on past experience. This sometimes leads us wrong, since the future is not always like the past. In the statistical terms we have put forward, we say that the DGP which generated $\mathcal{D}_{\text{train}}$ may not be the same as the DGP for P_{test}. This is the problem of *distribution shift* [14] or *concept drift* [37]. We must model this phenomenon to be able to analyze it, and there are multiple ways to model distribution shifts. This section will introduce ways to describe distribution shifts of relevance to the papers in this thesis.

Arbitrary distribution shifts can be measured with statistical distance measures. A chief example is the KL divergence,

$$D_{KL}(P\|Q) = \int \log\left(\frac{dP}{dQ}\right) dP, \tag{4.1}$$

which shows up in many places in machine learning. It has the properties that $D_{KL}(P\|Q) \geq 0$ with equality if and only if $P = Q$. We interpret it such that a small KL-divergence means that the distribution shift between the two is small. Other statistical distance measures found in the literature to describe distribution shifts are f-divergences [27], Mahalanobis distance [43], Maximum Mean Discrepancy [20] and χ^2-distance [19], and many more [37].

It may also be interesting to know what part of the data space in which the shift happens. We say that A is a *drift region* if the distribution shift happens in A, and not outside of it,

$$P(x) \neq Q(x) \quad \forall x \in AP(x) = Q(x) \quad \forall x \notin A \tag{4.2}$$

An example of this would be a failing sensor that generate fatter tails in the noise distribution. Another case is when the data is multidimensional, and only some dimensions of the data shifts. The set A where the distribution shift happens is called a *drift region* and understanding it can benefit in adaptation to distribution shift. [37]

Factor Shifts

In supervised machine learning, we analyze pairs of covariates and labels (X, y). The joint distribution can always be factorized.

$$P(X, y) = P(X|y)P(y) = P(y|X)P(X) \tag{4.3}$$

and the drift region may be localized in either factor or happen in both factors at the same time. This gives rise to terminology.

Definition 3. The three factor shifts in supervised learning are

- *Distribution shift*: $Q \neq P$ generically
- *Label shift*: $Q(X|y) = P(X|y)$, but $Q(y) \neq P(y)$.
- *Covariate shift*: $Q(y|X) = P(y|X)$, but $Q(X) \neq P(X)$.

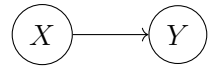
The different shifts have been explored both theoretically and in application. Covariate shift problems have been explored in regression and classification, since the values of X will be available when making predictions on out-of-sample data. [55, 20, 59]. Label shift problems occur in classification tasks where the covariates are generated from the labels. Medical diagnosis is one prominent application [36, 58]. In the medical applications where y indicates disease and $P(y)$ describes the prevalence, it may be called a prevalence shift [54, 12].

Intervention Shifts

Making interventions changes what data we observe. One perspective on this is that there is a system generating the data, and it does so in different ways depending on our the interventions taken. We formalize this differently in potential outcomes or SCMs.

In the potential outcomes framework, the data generating process is constant, but the way we observe the data is changed. The underlying joint distribution for data, $(X, Y, T, Y(1), Y(0)) \sim P$ is not affected by interventions. The observation mechanism is often not explicitly modeled in the analysis, so the distribution shift is implicit. This sometimes complicates the analysis since assumptions for identifiability (e.g., Definition 2) is on P, an object that describes both $Y(1)$ and $Y(0)$ at the same time, despite only one of the two can happen in the real world. In this thesis, the 'DGP' sometimes refer to the combination of data generation and observation mechanism, and it should be clear from context when this is the case.

In contrast, the SCM framework models the whole system and its behavior under intervention explicitly. The DGP is described by an SCM, and the distribution of data observed from the system depends on the intervention. We can illustrate this in the case of factor shifts on two variables.



The factorization of the distribution that is implied by this graph, and the interventional distributions are

$$P(x, y) = P(x)P(y|x) \tag{4.4}$$

$$P^{do(X=x)}(x, y) = \mathbf{1}\{X = x\}\, P(y|x) \tag{4.5}$$

$$P^{do(Y=y)}(x, y) = P(x)\mathbf{1}\{Y = y\} \tag{4.6}$$

indicating a difference that under one intervention, the other variable is conditioned (causal shift), or it is not (anticausal shift). This has implications for how to address machine learning tasks on data from this system [53]. In this setting, it becomes clear that interventions induce a distribution shift from the observational distribution P to the interventional distribution $P^{do(X=x)}$.

Whether to chose potential outcomes or the SCMs for analysis finally depends on the context. In Paper V the former approach is used, whereas Paper I uses the latter.

5. Robust Decision Making

This chapter discusses ways to make decisions related to systems that exhibit distribution shift. There are several approaches that can be taken, and only a few are explored in this thesis. Section 5.1 and 5.2 presents two approaches of inference that are lean on assumptions. Section 5.3 presents machine learning of causal effect estimation. Causal parameters are invariant to the distributions shifts that decisions about interventions induce, forming a valid foundation for decision making. Section 5.4 discusses challenges in machine learning that arise in the application of machine learning to data from neurocritical care, where distribution shifts are present.

5.1 Loss Bounds

When decisions have consequences that we want to control, we can make inferences about them directly. The diagnostic tool LAL of Paper II can be used to quantify how well a model performs out-of-sample, even if there is distribution shift between the training data set and the data generating process for future data.

It was inspired by ongoing research in conformal inference, where the goal is to make valid inferences without assuming i.i.d. data. To make this connection clear, below follows a brief presentation of split conformal inference. Several extensions are available [63].

Definition 4. A sequence of random variables $X_1, X_2, \ldots$ is *exchangeable* if the distribution function F_n is invariant under permutations of the indices

$$F_n(x_1, x_2, \ldots, x_n) = F_n(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)}) \tag{5.1}$$

for all n and permutations π over $\{1, 2, \ldots, n\}$.

This forms a relaxation of the i.i.d. assumption, and shows up in e.g., sampling without replacement. We consider a dataset $(x_1, \ldots, x_n, x')$ that is exchangeable, and we want to make a prediction about x' based on the other data points. Select a real valued *conformity score* $\mathcal{B}$ mapping $\mathcal{B}(x_i) = B_i$ and $\mathcal{B}(x') = B'$. Small values of the conformity score $\mathcal{B}(x)$ means the input was x as expected. Since the conformity scores B_i are

real valued and exchangeable, a combinatorial argument can be used to construct a prediction set $C(\{B_i\}_{i=1}^n)$ with valid coverage

$$\mathbb{P}[B' \in C(\{B_i\}_{i=1}^n)] \geq 1 - \alpha \tag{5.2}$$

By mapping back to the original data $\{x_i\}_{i=1}^n$, we obtain a valid prediction set $C'(\{x_i\}_{i=1}^n) = \{x : \mathcal{B}(x) \in C(\{B_i\}_{i=1}^n)\}$

$$\mathbb{P}[x' \in C'(\{x_i\}_{i=1}^n)] \geq 1 - \alpha. \tag{5.3}$$

We use a similar logic in the construction of the LAL in Paper II. It can be seen in the comparison of (3.8) and (5.2). The analysis of a real valued loss function $\ell(\cdot)$ plays the role of the conformity score $\mathcal{B}$. Through a prediction set for the loss on future data, we can place trust in a machine learning model despite distribution shift.

5.2 Feedback Control

One of the drawbacks with the assumption of exchangeability is the challenge to deal with time series data. Despite effort in extending conformal inference to time series, no proposed framework shows the same lean assumptions as conformal inference [13, 41]. One alternative explored in the literature is to shift focus from probabilistic guarantees to guarantees that are deterministic in time instead [18, 2]. The method constructs a sequence of prediction sets $C(\{x_u\}_{u=1}^{s-1})$ with guarantees takes a form

$$\frac{1}{t}\sum_{s=1}^{t} \mathbf{1}\Big\{x_s \in C(\{x_u\}_{u=1}^{s-1})\Big\} \geq 1 - \alpha + o(t), \tag{5.4}$$

possibly involving an upper bound, or incorporating side information about x_s for each prediction set. The asymptotic terms $o(t)$ have explicit formulations, so the guarantee can be made valid for all t.

This guarantee is not probabilistic, and must be valid for all random sequences that fulfil the assumptions of the guarantee 5.4. We utilize closely connected ideas in Paper III to obtain a similar deterministic guarantee about service levels and costs in inventory control.

5.3 Machine Learning in ATE and CATE Estimation

Several advances have been made in combining machine learning with causal inference. Paper V in this thesis explores two recent CATE estimators to assess their applicability to neurocritical care data. We aim to find an estimate $\hat{\tau}$ of the CATE τ.

There might also be nuisance parameters ν that must be estimated in the process, but that you are not specifically interested in. It is sometimes possible to use quite general machine learning methods for estimating ν and still get a valid estimate of τ, only requiring the stronger assumptions only for τ.

The naïve IPW estimator for the ATE is illustrative. One can show the under the identification assumptions in Definition 2, the ATE for a binary treatment is identified by

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[Y \frac{(\pi(X) - T)}{\pi(X)(1 - \pi(X))}\right]. \tag{5.5}$$

Estimate the propensity $\hat{\pi}(X)$ with any machine learning method, then compute the ATE by the plug in estimator of (5.5)

$$\widehat{\text{ATE}} = \frac{1}{|\mathcal{D}|} \sum_{(x,t,y)\in\mathcal{D}} y \frac{(\hat{\pi}(x) - t)}{\hat{\pi}(x)(1 - \hat{\pi}(x))}. \tag{5.6}$$

As long as the propensity score π is estimated consistently by $\hat{\pi}$, the ATE estimator will be consistent as well. The propensity score function π takes the role of a nuisance parameter. The problem with the IPW estimator is the factor $\frac{\pi(X)-T}{\pi(X)(1-\pi(X))}$ wich may be large, inflating the variance of the estimator significantly. This motivates the use of more complex estimators for the ATE and the CATE. We use two specific methods in Paper V.

DR-learner

Meta-learners are a family of estimation techniques where machine learning algorithms are combined according to a higher level algorithm to provide the estimator [32]. One example is the T-learner, where models $\hat{\mu}_t$ are trained to estimate the mean outcomes $\mu_t(x) = \mathbb{E}[Y|X = x, T = t]$. The T-learner estimates the CATE by the difference of the two models

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x) \tag{5.7}$$

and as long as the two models are consistent ($\hat{\mu}_t \xrightarrow{p} \mu_t$), the estimator $\hat{\tau}$ will be consistent as well.

The DR-learner [31] extends this by combining the approach of the IPW and the T-learner to get an estimate of the CATE. As long as either the propensity score or the mean outcome models are accurate, the final estimator will be so. Under certain assumptions on the involved estimators and the data generating process, the DR-learner can provide uncertainty estimates despite no such results being available for the nuisance parameters.

Generalized random forests

The generalized random forests algorithm solves a local estimating equations [6] of the form

$$\mathbb{E}[\psi_{\hat{\tau},\nu}(Z)|X = x] = 0 \tag{5.8}$$

through a specific construction of a random forest. Asymptotically, as the number of data points and the number of trees grow, the estimate $\hat{\tau}$ will obey an asymptotic normality which also provides asymptotically valid confidence intervals. The CATE estimation problem can be posed as solving the minimization

$$\hat{\tau}(x) = \arg\min_{\tau} \min_{\nu} \mathbb{E}\Big[(Y - \tau(X)T - \nu(x))^2 \Big| X = x\Big]. \tag{5.9}$$

From first order optimality, it would need

$$\mathbb{E}\left[\begin{bmatrix}(Y - \hat{\tau}(X)T - \nu(x)) \\ (Y - \hat{\tau}(X)T - \nu(x))T\end{bmatrix}\middle| X = x\right] = \begin{bmatrix}0\\0\end{bmatrix}, \tag{5.10}$$

which is an estimating equation on the form of (5.8), making it possible to solve it with the generalized random forest algorithm.

5.4 Evaluation of Machine Learning Models

In the applied work or Paper IV and Paper V, we are faced with the challenge of evaluating machine learning models in the presence of distribution shift. We observe shifts between patients and between years, and we want to make sure that the models we train are robust to these shifts. This section discusses two aspects of this problem.

Label Shift Invariant Performance Metrics

Consider the misclassification loss $\ell(f_\theta(X), Y) = \mathbf{1}\{f_\theta(X) \neq Y\}$ for use in binary classification tasks, where $Y \in \{0, 1\}$. The risk of a trained model can be written as

$$R(\theta) = \sum_{y=0,1} \mathbb{P}[Y = y]\, \mathbb{E}[\mathbf{1}\{f_\theta(X) \neq Y\}|Y = y]. \tag{5.11}$$

In situations with label shift, the conditional expectation will be the same both in training and out-of-sample since $P(X|Y)$ is invariant. One can argue that measures of the performance of the trained model should be equally invariant, since any factor that involves the trained model is invariant. However, the factor $\mathbb{P}[Y = y]$ is not invariant, the risk estimated from data may be different from the out-of sample risk, so the inference made about the risk will not be robust against the distribution shift.

To address this, one can use alternative metrics that only depend on the factors that are invariant. The Sensitivity, Specificity and Balanced Accuracy are such metrics for trained models with binary outputs

$$\text{Sens}(\theta) = \mathbb{P}[f_\theta(X) = 1 | Y = 1]\,, \tag{5.12}$$

$$\text{Spec}(\theta) = \mathbb{P}[f_\theta(X) = 0 | Y = 0]\,, \tag{5.13}$$

$$\text{BalAcc}(\theta) = \frac{1}{2}(\text{Sens}(\theta) + \text{Spec}(\theta)). \tag{5.14}$$

The area under the receiver operating characteristic AUROC curve can be defined for binary prediction models that produce a score rather than a 0/1 value. A binary predictor can be constructed from it by thresholding, i.e., $\mathbf{1}\{f_\theta(X) > t\}$ for some threshold t. The AUROC is equivalent to probability that $f_\theta(X_1) > f_\theta(X_0)$ if $X_1 \sim P(X|Y = 1)$ and $X_0 \sim P(X|Y = 0)$ making it invariant under label shift [15]. We use these metrics in Paper IV, since we observe that $Q(y) \neq P(y)$ where P and Q are the distributions for different years of data, but believe that $Q(X|y) = P(X|y)$ approximately.

Group Invariant Learning

The distribution shift between years or patients in medical data can be seen as a grouping of data. Attempts have been made to exploit group structure for machine learning. An incomplete list of methods includes: Decomposing features into conditionally invariant features vs features that vary between groups [25], Hypothesis testing to find invariant features, so that a learned model uses the same sparsity pattern across different contexts [45], Regularizing linear models in a way that generalize instrumental variable regression and linear SCMs [51], regularizing binary classifiers to have smaller variation in risk between groups [27, 19], and adversarially weight the risk so that the obtained ML model is minimax optimal in a class of mixture distribution based on observed groups [42].

We use the group information to do model selection via grouped cross validation. When K-fold cross validation is applied to grouped data, random splitting will not fulfil the assumption that $\mathcal{D}_j$ is independent of $\mathcal{D}^{(-j)}$, so grouped cross validation is preferable [60, 49] to reattain that property and restoring unbiasedness. We use grouped cross validation in Paper IV and Paper V for this reason. In a hold-out performance evaluation, it would be done in the following way. Partition the dataset $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{val} \cup \mathcal{D}_{\text{test}}$ so that each group is present in at exactly one of the subsets. Train various models on $\mathcal{D}_{train}$ and measure the hold-out risk using $\mathcal{D}_{val}$. Select the most promising model. Retrain the model on $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ and evaluate on $\mathcal{D}_{\text{test}}$. The extension to K-fold cross validation is straightforward. This is standard practice in machine learning when partitioning the data randomly, but grouped partitioning is less common.

A final word of warning is needed. Both the training and testing must be treated equally. One study analyze grouped image data and compare different cross validation schemes [24]. They split data into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, either by random sample selection, or in a grouped fashion. They then train a model on $\mathcal{D}_{\text{train}}$ and evaluate it on $\mathcal{D}_{\text{test}}$. The authors state in the article highlights "Random validation showed the best performance". While true, it is not a relevant conclusion. Random validation allows data leakage, and the performance estimate will be biased. While the number indeed is better, it says little about the out-of-sample performance.

Summary in Swedish - Sammanfattning på Svenska

Vi tar ständigt beslut baserat på tidigare erfarenheter. Vissa beslut har små konsekvenser och vi tar dem lättvindigt, men när mycket står på spel göra vi allt för att försäkra oss om att besluten får önskade konsekvenser. Vetenskapliga metoder, och experiment i synnerhet, har lett oss till robusta slutsatser som hjälper oss ta rätt beslut. Tack vare förmågan att kontrollera relevanta ingående variabler har man upptäckt lagar som generaliserar väl.

Dessvärre finns situationer där denna analys försvåras. Dels finns experiment där inte alla kontrollvariabler kan styras fritt. Dels kan slumpeffekter verka olika i experiment och senare i beslutssituationer. Vi behöver statistiska verktyg som hjälper oss att överföra slutsatser dragna från ej perfekta experiment till de verkliga beslutssituationer som vi vill förstå. Om slutsatserna handlar om hur beslut orsakar utfall (till skillnad från an analys om statistisk samvariation) kallar vi denna analys för kausal inferens.

Matematiskt modelleras detta med sannolikhetsteori. Data från experiment skapas utifrån en viss underliggande process. Den verkliga sitationen motsvarar likaså en underliggande process. Dessa beskrivs med sannolikhetsfördelningar. Vid perfekta experiment sammanfaller dessa fördelningar vilket underlättar analysen. När de skiljer sig har vi en *fördelningsförskjutning*. Fördelningsförskjutningar uppstår i flera andra situationer, så som när maskininläning används för att analysera observationsdata. Denna avhandling diskuterar metoder för att dra slutsatser från data som är robusta mot fördelningsförskjutningar. Fem delarbeten presenteras.

Delarbete I *Inference of causal effects when control variables are unknown* (Inferens om kausala effekter trots okända kontrollvariabler) undersöker möjligheten att skapa giltiga konfidensintervall för en kausal effektparameter. Tidigare arbeten kräver att man vet vilka variabler som är kontrollvariabler. Detta delarbete kombinerar skattningen av kausal effekt med skattningen av kontrollvariabler och erhåller asymptotiskt normal fördelning av skattaren. Bland nyckelantagandena är att alla kontrollvariabler finns observerade, och den kausala effektparametern är definierad utifrån en linjär approximation.

Delarbete II, *Diagnostic Tool for Out-of-Sample Model Evaluation* (Diagnostiskt verktyg för utvärdering av modeller på ny data) introducerar

ett mått som kan användas för att analysera hur maskinginlärningsmodeller presterar på data som du inte har tillgång till. Vi använder metoder från ickeparametrisk statistisk teori och visar att om du har tillgång till ett litet dataset som är utbytbart ("exchangeable") med framtida data kan du ändå få garantier. Genom att kvantifiera modellens prestanda kan det bli möjligt att använda maskinginlärningsmodeller trots fördelningsförskjutning.

Delarbete III, *Certified Inventory Control of Critical Resources* (Lagerhantering för kritiska resurser, med garantier) använder sig av återkoppling för att adaptivt styra en lagerprocess. Det är möjligt att ta en enkel regulator och lägga på en adaptiv säkerhetsmarginal för att få garantier kring hur ofta lagret går ner på kritiska nivåer. Då regulatorn måste vara konservativ för att uppnå garantierna, och konservativa strategier ofta har större kostnader, tar vi även fram en metod för att skatta kostnaden för strategin. Den skattaren har garantier av samma typ som den justerade regulatorn.

Delarbete IV, *Machine learning based prediction of imminent ICP insults during neurocritical care of traumatic brain injury* (Maskininlärning för att förutsäga kritiskt förhöjt intrakraniellt tryck hos patienter med traumatisk hjärnskada) syftar till att undersöka möjligheten att utveckla kliniska varningssystem. Neurointensivvårdsavdelningen vid Akademiska Sjukhuset i Uppsala har sammanställt en forskningsdatabas med patienter som lider av traumatisk hjärnskada. Artikeln utgår från tidigare arbeten på liknande data och visar hur detta större dataset möjliggör bättre prestanda än tidigare. Vi visar vidare att modeller som har väl etablerade ramverk (framförallt XGBoost) presterar minst lika bra än modeller med högre kostnad för utveckling (handanpassad gausisk process). Slutligen diskuterar vi några av hindren som kvarstår innan en klinisk användning är möjlig. Genom arbetet såg vi variation i prevalens mellan olika inskrivningsår, vilket indikerar att det finns en fördelningsförskjutning i underlaget. Vi hanterar detta genom att utvärdera våra modeller med utvärderingsmått som är invarianta mot fördelningsförskjutningar i den underliggande prevalensen.

Delarbete V, *Challenges in CATE estimation for neurointensive care* bygger vidare på delarbete IV. En mindre kohort av patienter vid Neurointensivvårdsavdelningen, för vilka behandlingsinformation är nedtecknat, används för att undersöka möjligheten att använda kausal inferens. Specifikt undersöker vi skattare för den betingade genomsnittliga behandlingseffekten. Ingen av skattarna uppvisar acceptabel prestanda, vilket pekar på flera utmaningar som kvarstår innan kausal inferens kan användas i praktiken. Dels krävs ytterligare teoretiska landvinningar för att ta fram giltiga konfidensintervall när kontrollvariablerna är mycket högdimensionella. Empiriska undersökningar av hur man kan kombinera data utan behandlingsinformation med data med behandlingsinformation be-

hövs för att effektivisera skattningarna. Vid kliniken behöver man samla in större dataset med högre kvalitet. Slutligen kan väl etablerade klinska regler för databehandling (till exempel exklusionskriterier) underlätta implementation av kausal inferens.

Acknowledgements

My first gratitude is to my main supervisor, Dave Z. He has taught me that good ideas take time to develop, and that simplicity is critical to communicating ideas. He has always been supportive. My co-supervisor Thomas S is a pillar to the positive spirit at SysCon. He has always cheered on, both as a co-supervisor and as a department head. I thank Petre S for all detailed feedback and advice on the art of scientific writing. I want to thank Per M for being a mentor the first few summer months when the corridors at Pollacksbacken were empty.

I thank Peter G, for stimulating discussions in topics from computers to healthcare and from research to business, and for working on the digitalization of the healthcare system. I also thank Per E and Anders H, my other closest collaborators at Akademiska.

Jonas Peters accepted me as a visitor in his group at ETH in Zürich, gave valuable feedback on my work with Paper V, gave me other perspectives on this research field, and I thank him for this.

My colleagues at SysCon all deserve my thanks. Daniel G, for both being a colleague and a friend. Anna W, Calle A, Calle J, Fredrik K G and Tung have all provided me with stimulating discussions both about work and leisure. Linnea G, for being the desk mate that gave me more perspectives on interdisciplinary research, for ideas on applying for grants, and for being brave. Fabio B, for coming up with fun ideas like tasting julmust and for also discussing time series models. Antônio R, for advice on maths and research over the years. I thank all other PhD students and PostDocs in SysCon, for making the tone a bit brighter, for stimulating talks over lunch, for kareoke, for parties and for the reading groups.

Emil, Tomas and Göran gave me the encouragement to apply for a PhD student position and accept it. I needed it.

My family – Patrik, Hillevi and Sofia – has always encouraged me to make the intellectually stimulating choices. And my great aunt Monica showed me that it is never too late to pursue research.

And I must finally thank Sara. Thank you for everything.

Funding

I am thankful for the funding by the Swedish Research Council that has provided financing for my research. I am also thankful for grants provided by ELISE and by G Research enabling travelling during my studies. Part of Paper V was done while I was visiting ETH Zürich. The Swedish

National Infrastructure for Computing and the National Academic Infrascructure for Supercomputing in Sweden has provided computing infrastructure for part of my work.

Bibliography

[1] Ahmed Alaa, Zaid Ahmad, and Mark van der Laan. "Conformal Meta-learners for Predictive Inference of Individual Treatment Effects." In: arXiv:2308.14895 (Aug. 2023). DOI: 10.48550/arXiv.2308.14895.

[2] Anastasios N. Angelopoulos, Emmanuel J. Candes, and Ryan J. Tibshirani. "Conformal PID Control for Time Series Prediction." In: arXiv:2307.16895 (July 2023). DOI: 10.48550/arXiv.2307.16895.

[3] Martin Arjovsky et al. *Invariant Risk Minimization*. July 2019. arXiv: 1907.02893 [cs, stat]. URL: http://arxiv.org/abs/1907.02893.

[4] Sylvain Arlot and Alain Celisse. "A survey of cross-validation procedures for model selection." In: *Statistics Surveys* 4 (Jan. 2010), pp. 40–79. ISSN: 1935-7516. DOI: 10.1214/09-SS054.

[5] Peter M. Aronow and Benjamin T. Miller. *Foundations of Agnostic Statistics*. 1st ed. Cambridge University Press, Jan. 2019. ISBN: 978-1-316-83176-2. DOI: 10.1017/9781316831762.

[6] Susan Athey, Julie Tibshirani, and Stefan Wager. "Generalized random forests." In: *The Annals of Statistics* 47.2 (Apr. 2019), pp. 1148–1178. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/18-AOS1709.

[7] Dimitri P. Bertsekas. *Dynamic programming and optimal control*. Belmont, Mass: Athena Scientific, 1995. ISBN: 978-1-886529-11-3.

[8] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. New York: Springer, 2006. ISBN: 978-0-387-31073-2.

[9] Peter Bühlmann. *Invariance, Causality and Robustness*. Dec. 2018. arXiv: 1812.08233 [stat]. URL: http://arxiv.org/abs/1812.08233.

[10] Peter Bühlmann. "Invariance, Causality and Robustness." In: *Statistical Science* 35.3 (Aug. 2020), pp. 404–426. ISSN: 0883-4237. DOI: 10.1214/19-STS721.

[11] Nancy Cartwright. "Modularity: It Can-and Generally Does-Fail*." In: *Stochastic Causality*. CSLI Publications, 2001, pp. 65–84. ISBN: 1-57586-321-9.

[12] Daniel C. Castro, Ian Walker, and Ben Glocker. "Causality matters in medical imaging." In: *Nature Communications* 11.1 (July 2020), p. 3673. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17478-w.

[13] Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. "Exact and Robust Conformal Inference Methods for Predictive Machine Learning with Dependent Data." In: *Proceedings of the 31st Conference On Learning Theory*. PMLR, July 2018, pp. 732–749. URL: https://proceedings.mlr.press/v75/chernozhukov18a.html.

[14] *Dataset shift in machine learning*. Neural information processing series. Cambridge, Mass: MIT Press, 2009. ISBN: 978-0-262-17005-5. URL: https://ebookcentral.proquest.com/lib/uu/detail.action?docID=3338975#.

[15] Tom Fawcett. "An introduction to ROC analysis." In: *Pattern Recognition Letters* 27.8 (June 2006), pp. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010.

[16] Tadayoshi Fushiki. "Estimation of prediction error by using K-fold cross-validation." In: *Statistics and Computing* 21.2 (Apr. 2011), pp. 137–146. ISSN: 0960-3174, 1573-1375. DOI: 10.1007/s11222-009-9153-8.

[17] Saurabh Garg et al. "A Unified View of Label Shift Estimation." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 3290–3300. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/219e052492f4008818b8adb6366c7ed6-Paper.pdf.

[18] Isaac Gibbs and Emmanuel Candes. "Adaptive Conformal Inference Under Distribution Shift." In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 1660–1672. URL: https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html.

[19] Jun-ya Gotoh, Michael Jong Kim, and Andrew E.B. Lim. "Robust empirical optimization is almost the same as mean–variance optimization." In: *Operations Research Letters* 46.4 (July 2018), pp. 448–452. ISSN: 01676377. DOI: 10.1016/j.orl.2018.05.005.

[20] Arthur Gretton et al. "Covariate Shift by Kernel Mean Matching." In: *Dataset Shift in Machine Learning*. Ed. by Joaquin Quiñonero-Candela et al. The MIT Press, Dec. 2008, pp. 131–160. ISBN: 978-0-262-17005-5. DOI: 10.7551/mitpress/9780262170055.003.0008.

[21] Allan Gut. *Probability: a graduate course*. Second edition. Springer texts in statistics. New York: Springer, 2013. ISBN: 978-1-4614-4707-8.

[22] J. Y. Halpern and Judea Pearl. "Causes and Explanations: A Structural-Model Approach. Part I: Causes." In: *The British Journal for the Philosophy of Science* 56.4 (Sept. 2005), pp. 843–887. ISSN: 0007-0882, 1464-3537. DOI: 10.1093/bjps/axi147.

[23] J. Y. Halpern and Judea Pearl. "Causes and Explanations: A Structural-Model Approach. Part II: Explanations." In: *The British Journal for the Philosophy of Science* 56.4 (Sept. 2005), pp. 889–911. ISSN: 0007-0882, 1464-3537. DOI: 10.1093/bjps/axi148.

[24] Junjie Han et al. "Evaluation of computer vision for detecting agonistic behavior of pigs in a single-space feeding stall through blocked cross-validation strategies." In: *Computers and Electronics in Agriculture* 204 (Jan. 2023), p. 107520. ISSN: 0168-1699. DOI: 10.1016/j.compag.2022.107520.

[25] Christina Heinze-Deml and Nicolai Meinshausen. "Conditional Variance Penalties and Domain Shift Robustness." In: *arXiv:1710.11469 [cs, stat]* (Oct. 2017). URL: http://arxiv.org/abs/1710.11469.

[26] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020. URL: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

[27] Weihua Hu et al. "Does Distributionally Robust Supervised Learning Give Robust Classifiers?" In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 2029–2037. URL: https://proceedings.mlr.press/v80/hu18a.html.

[28] Duligur Ibeling and Thomas Icard. "Comparing Causal Frameworks: Potential Outcomes, Structural Models, Graphs, and Abstractions." In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 80130–80141. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/fd83f4e0dcaf1c64ea15bbb1695bb40f-Paper-Conference.pdf.

[29] Guido Imbens and Donald B. Rubin. *Causal inference for statistics, social, and biomedical sciences: an introduction*. New York: Cambridge University Press, 2015. ISBN: 978-0-521-88588-1.

[30] Philippe Jorion. *Value at risk: the new benchmark for managing financial risk*. 3rd ed. New York: McGraw-Hill, 2007. ISBN: 978-0-07-173692-3.

[31] Edward H. Kennedy. "Towards optimal doubly robust estimation of heterogeneous causal effects." In: arXiv:2004.14497 (Aug. 2023). DOI: 10.48550/arXiv.2004.14497.

[32] Sören R. Künzel et al. “Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning.” In: *Proceedings of the National Academy of Sciences* 116.10 (Mar. 2019), pp. 4156–4165. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1804597116.

[33] Lihua Lei and Emmanuel J. Candès. “Conformal inference of counterfactuals and individual treatment effects.” In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 83.5 (Nov. 2021), pp. 911–938. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/rssb.12445.

[34] Retsef Levi, Georgia Perakis, and Joline Uichanco. “The Data-Driven Newsvendor Problem: New Bounds and Insights.” In: *Operations Research* 63.6 (Dec. 2015), pp. 1294–1306. ISSN: 0030-364X, 1526-5463. DOI: 10.1287/opre.2015.1422.

[35] Andreas Lindholm et al. *Machine Learning: A First Course for Engineers and Scientists*. 1st ed. Cambridge University Press, Mar. 2022. ISBN: 978-1-108-91937-1. DOI: 10.1017/9781108919371.

[36] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. “Detecting and Correcting for Label Shift with Black Box Predictors.” In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 3122–3130. URL: https://proceedings.mlr.press/v80/lipton18a.html.

[37] Jie Lu et al. “Learning under Concept Drift: A Review.” In: *IEEE Transactions on Knowledge and Data Engineering* (2018). ISSN: 1041-4347, 1558-2191, 2326-3865. DOI: 10.1109/TKDE.2018.2876857.

[38] William Q. Jr Meeker, Gerald J. Hahn, and Luis A. Escobar. *Statistical intervals: a guide for practitioners and researchers*. 2nd ed. Hoboken: John Wiley et sons, 2017. ISBN: 978-1-118-59484-1. URL: https://ebookcentral.proquest.com/lib/uu/detail.action?docID=4822515.

[39] Rupert G. Miller. “The Jackknife–A Review.” In: *Biometrika* 61.1 (Apr. 1974), p. 1. ISSN: 00063444. DOI: 10.2307/2334280.

[40] Kevin P. Murphy. *Probabilistic machine learning: an introduction*. Adaptive computation and machine learning series. Cambridge, Massachusetts: The MIT Press, 2022. ISBN: 978-0-262-04682-4.

[41] Buddhika Nettasinghe et al. “Extending Conformal Prediction to Hidden Markov Models with Exact Validity via de Finetti’s Theorem for Markov Chains.” In: arXiv:2210.02271 (May 2023). DOI: 10.48550/arXiv.2210.02271.

[42] Muhammad Osama, Dave Zachariah, and Petre Stoica. "Robust Learning in Heterogeneous Contexts." In: arXiv:2105.08532 (Feb. 2022). DOI: 10.48550/arXiv.2105.08532.

[43] Chunjong Park et al. "Reliable and Trustworthy Machine Learning for Health Using Dataset Shift Detection." In: *arXiv:2110.14019 [cs]* (Oct. 2021). URL: http://arxiv.org/abs/2110.14019.

[44] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, Sept. 2009. ISBN: 978-0-521-89560-6.

[45] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. "Causal inference using invariant prediction: identification and confidence intervals." In: *arXiv:1501.01332 [stat]* (Jan. 2015). URL: http://arxiv.org/abs/1501.01332.

[46] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. Adaptive computation and machine learning series. Cambridge, Massachuestts: The MIT Press, 2017. ISBN: 978-0-262-03731-0.

[47] Richard R. Picard and R. Dennis Cook. "Cross-Validation of Regression Models." In: *Journal of the American Statistical Association* 79.387 (1984), pp. 575–583. ISSN: 0162-1459. DOI: 10.2307/2288403.

[48] Thomas Richardson. *Single World Intervention Graphs (SWIGs): Unifying the Counterfactual and Graphical Approaches to Causality*. Aug. 2013. URL: http://people.tuebingen.mpg.de/p/causality-perspect/slides/Thomas_swig.pdf.

[49] David R. Roberts et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." In: *Ecography* 40.8 (2017), pp. 913–929. ISSN: 1600-0587. DOI: 10.1111/ecog.02881.

[50] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. "The Risks of Invariant Risk Minimization." In: *arXiv:2010.05761 [cs, stat]* (Oct. 2020). URL: http://arxiv.org/abs/2010.05761.

[51] Dominik Rothenhäusler et al. "Anchor Regression: Heterogeneous Data Meet Causality." In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.2 (Apr. 2021), pp. 215–246. ISSN: 1369-7412. DOI: 10.1111/rssb.12398.

[52] Donald B. Rubin. "Bayesian Inference for Causal Effects: The Role of Randomization." In: *The Annals of Statistics* 6.1 (1978), pp. 34–58. ISSN: 0090-5364. URL: https://www.jstor.org/stable/2958688.

[53] Bernhard Schölkopf et al. “On causal and anticausal learning.” In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. Ed. by John Langford and Joelle Pineau. ICML ’12. Edinburgh, Scotland, GB: Omnipress, July 2012, pp. 1255–1262. ISBN: 978-1-4503-1285-1. URL: https://icml.cc/Conferences/2012/papers/215.pdf.

[54] Jessica Schrouff et al. “Diagnosing failures of fairness transfer across distribution shift in real-world medical settings.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 19304–19318. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/7a969c30dc7e74d4e891c8ffb217cf79-Paper-Conference.pdf.

[55] Hidetoshi Shimodaira. “Improving predictive inference under covariate shift by weighting the log-likelihood function.” In: *Journal of Statistical Planning and Inference* 90.2 (Oct. 2000), pp. 227–244. ISSN: 03783758. DOI: 10.1016/S0378-3758(00)00115-4.

[56] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Vol. 81. Lecture Notes in Statistics. New York, NY: Springer New York, 1993. ISBN: 978-1-4612-7650-0. DOI: 10.1007/978-1-4612-2748-9.

[57] M. Stone. “Cross-Validatory Choice and Assessment of Statistical Predictions.” In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (Jan. 1974), pp. 111–133. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1974.tb00994.x.

[58] Amos Storkey. “When Training and Test Sets Are Different: Characterizing Learning Transfer.” In: *Dataset Shift in Machine Learning*. Ed. by Joaquin Quiñonero-Candela et al. Neural information processing series. The MIT Press, 2008. ISBN: 9780262255103.

[59] Ryan J Tibshirani et al. “Conformal Prediction Under Covariate Shift.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

[60] Ilias Tougui, Abdelilah Jilbab, and Jamal El Mhamdi. “Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications.” In: *Healthcare Informatics Research* 27.3 (July 2021), pp. 189–199. ISSN: 2093-3681. DOI: 10.4258/hir.2021.27.3.189.

[61] A. W. van der Vaart. *Asymptotic Statistics*. 1st ed. Cambridge University Press, Oct. 1998. ISBN: 978-0-521-49603-2. DOI: 10.1017/CBO9780511802256.

[62] Stijn Vansteelandt and Oliver Dukes. “Assumption-lean Inference for Generalised Linear Model Parameters.” In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.3 (July 2022), pp. 657–685. ISSN: 1369-7412. DOI: 10.1111/rssb.12504.

[63] Vladimir Vovk, A. Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. New York: Springer, 2005. URL: https://doi.org/10.1007/b106715.

[64] Stefan Wager and Susan Athey. “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests.” In: *Journal of the American Statistical Association* 113.523 (July 2018), pp. 1228–1242. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2017.1319839.

[65] Larry Wasserman. *All of nonparametric statistics*. Springer texts in statistics. New York: Springer, 2006. ISBN: 978-0-387-25145-5.

[66] Jeffrey M. Wooldridge. *Econometric analysis of cross section and panel data*. 2nd ed. Cambridge, Mass: MIT Press, 2010. ISBN: 978-0-262-23258-6.

[67] Zitong Yang, Emmanuel Candès, and Lihua Lei. “Bellman Conformal Inference: Calibrating Prediction Intervals For Time Series.” In: arXiv:2402.05203 (Feb. 2024). DOI: 10.48550/arXiv.2402.05203.

[68] Xun Zheng et al. “DAGs with NO TEARS: Continuous Optimization for Structure Learning.” In: *arXiv:1803.01422 [cs, stat]* (Mar. 2018). URL: http://arxiv.org/abs/1803.01422.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2421

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-533683

ACTA UNIVERSITATIS
UPSALIENSIS
2024