Out of the Box

Statistical Methods for the Analysis of Automated Home Cage Experiments

Nadia J. Vendrig

Thesis committee

Promotor:
Prof. Dr C.J.F. ter Braak
Personal chair at the Mathematical and Statistical Methods Group
Wageningen University & Research

Co-promotor:
Dr L. Hemerik
Associate Professor at the Mathematical and Statistical Methods group
Wageningen University & Research

Other members:
Prof. Dr ir P.J. van den Brink, Wageningen University & Research
Dr P. Haccou, Leiden University
Dr L.P.J.J. Noldus, Noldus Information Technology BV, Wageningen
Dr M. Loos, VU Amsterdam and Sylics (Synaptologics B.V.), Amsterdam

This research was conducted under the auspices of the C.T. de Wit Graduate School of Production Ecology & Resource Conservation (PE&RC)

Out of the Box

Statistical Methods for the Analysis of Automated Home Cage Experiments

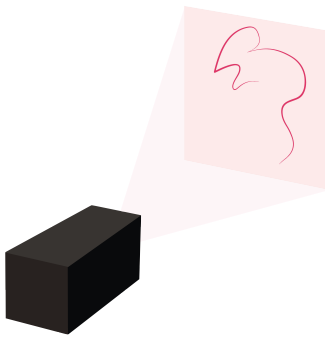
Nadia J. Vendrig

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof.Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 25 May 2018
at 4 p.m. in the Aula.

Nadia J. Vendrig
Out of the Box - Statistical Methods for the Analysis of Automated Home Cage Experiments, 192 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2018)
With references, with summaries in English and Dutch
ISBN 978-94-6343-262-7
DOI https://doi.org/10.18174/441414

Contents



1

General Introduction

...the young and the old of widely different races, both with man and animals, express the same state of mind by the same movements

Charles Darwin

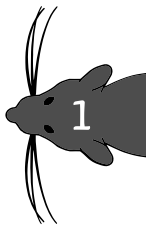
1.1. Behavioural Phenotyping

The epigraph to this chapter is taken from the concluding chapter of "The Expression of the Emotions in Man and Animals" (Darwin 1872, p.351). This book is Darwin's third major work of evolutionary theory and it introduces some of the concepts on which the field of behavioural genetics would be built. Humans and higher animals show emotional states such as fear, pain, and excitement in similar fashions and they exhibit similar behavioural patterns such as playing, fighting, and exploration. These homologies across species enable us to experiment on one model species to learn about behaviour and the organisation of behaviour in general. It also enables us to use one model species to learn about behaviour of another species. In this thesis, we focus on laboratory rats and mice as model species to study human psychiatric and neurological disorders in the broadest sense.

Modelling the complete complexity of human psychiatric disorders in animals in its entirety remains wishful thinking. "*Unanticipated breakthroughs would be required to convincingly model phenomena such as guilt, religiosity, grandiosity, envy, delusions, hallucinations, grief, body image distortion, and multiple personality in the mouse.*"(Tecott and Nestler 2004). In spite of the difficulties, important progress has been made in treating and understanding human psychiatric disorders using laboratory rats and mice. For depression for instance, human symptoms have been replicated in animals and animal behaviour experiments have been developed to identify potential anti-depressants (*e.g.* Porsolt, Le Pichon, and Jalfre 1977; Porsolt et al. 1978; Steru et al. 1985; Willner, Muscat, and Papp 1992). These types of studies have resulted in increased understanding of the disease in humans (*e.g.* Heim and Nemeroff 2001; Shelton 2007).

The key component in animal behavioural research is behavioural phenotyping: the characterisation of the set of observable behavioural characteristics of individuals. A behavioural phenotype, *i.e.* the behaviour shown in a certain situation, is an expression of the interaction between genetic background, the brain, and the environment. "*No behavioral phenotype exists separately from a test situation, because behavior is a reaction to something. It is this reaction that we seek to measure.*" (Wahlsten et al. 2003). Behavioural phenotyping is necessary in all three major categories of psycho-pharmaceutical animal behaviour studies: animal models, behavioural screening, and behavioural bioassays (Willner 1991, defintitions in Box 1.1).

Behaviour is the ultimate and most complex output of the brain (Spruijt and Visser 2006). Behavioural phenotypes are, compared to other phenotypical traits such as hair colour and body weight, not as easy to observe and not as stable. Optimal behavioural phenotyping with maximum validity (definitions in Box 1.2) requires high levels of complexity in research design and description of behaviour.



However, practical feasibility requires abstraction through simplification of research designs and parametrised behaviour.

Box 1.1: Definitions

Behavioural phenotyping is the characterisation of the set of observable behavioural characteristics of an individual resulting from the interaction of its genotype with the environment.

Behavioural tests are used to detect the phenotypic differences and effects of the animal models. For example, in the forced swim test animals are placed in a cylinder of water from which it cannot escape. After some time, rats stop trying to escape and stay immobile in the water (Yankelevitch-Yahav et al. 2015). Rats deprived of maternal care (animal model) show a different behavioural phenotype compared to the control group as they swim less, spend more time immobile, and less time trying to escape (Réus et al. 2011).

Animal models are experimental set-ups or protocols (sometimes also called "a paradigm") that alter the phenotype of the (laboratory) animals to replicate (sets of) symptoms from psychiatric diseases. For example, an animal model for depression in rats is to deprive pups of maternal care. Other examples are administering a drug treatment or a line of animals with a certain genetic defect.

Behavioural screening tests are used to establish the effect of genetic manipulations or drugs on the behavioural phenotype. Technological advances have dramatically increased the possibilities for targeted genetic modifications, first in mice but increasingly in rats. This allows us to study the role of the modified gene in the context of a living organism to increase insight into the functional effect of individual genes.

Behavioural bioassays are used to measure the activity of neural pathways using behavioural parameters as indicators. The behaviour here is used as a means to quantify an effect, for instance in a dose-response experiment.

The most common way of behavioural phenotyping is by means of the so-called classical tests such as the Open Field test (Hall 1934; Hall and Ballachey 1932, Test explanation in Box 1.3) and the Elevated Plus Maze (Test explanation

Box 1.2: Definitions: Validity

Validity of animal models and tests is typically described using three criteria: face, predictive, and construct validity although the exact definition differs between authors (Belzung and Lemoine 2011).

Face validity is the similarity of the symptoms in the animal model to the symptoms in the disorder in humans. Here, these symptoms are mostly the behavioural and cognitive response.

Predictive validity entails that an effective (drug) treatment for the disorder in humans is also effective in the animal model.

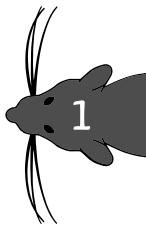
Construct validity is the extent to which the animal model measures what it is intended to measure. This implies that there is a theoretical background that explains the behaviour in the animal model and in the disorder.

in Box 1.4). A typical classical test aims at quantifying a single behavioural construct (*e.g.* anxiety, activity, exploration) using abstract parameters. In the Open Field test for example, the construct of interest is anxiety and it is quantified using abstract parameters such as "Time spent in the middle of the open field". The parameters do not need to have a direct ethological interpretation. Classical tests are cheap and easy to implement and have been and are still widely applied. In the field of anxiety in 2010 and 2012, 80% of studies were based on 14 classical behavioural tests (Haller and Alicki 2012). Although the classical tests have resulted in scientific progress in the past, there are concerns regarding their replicability between laboratories and their validity (*e.g.* Crabbe 1999; Haller and Alicki 2012; Kafkafi et al. 2005; Mandillo et al. 2008; McClearn 2004). In the next section the replicability problem is introduced and an overview of the causes of the limited replicability and validity is provided.

1.2. Limited replicability in behavioural phenotyping

1.2.1. Replicability in science

Science is experiencing a replication crisis. Findings from peer reviewed papers are often not replicable. An experimental result that cannot be reproduced is scientifically useless and the flooding of literature with non-replicable results is detrimental to scientific progress. A survey of over 1,567 scientists executed by Nature revealed that 70% have tried and failed to replicate experiments by another scientist, and 50% have tried and failed to replicate their own (Baker and



Box 1.3: Open Field Test

The Open Field Test is a classical test that is widely applied in mice and rats to measure exploratory behaviour and general activity. The Open Field Test is usually a square, rectangular, or round arena surrounded by a wall that inhibits escape. The mouse or rat is placed in the arena and its behaviour is recorded for typically 10 to 15 minutes. The response variables used vary widely; some of the more common are distance moved, time spent in the centre of the arena, time spent moving, and incidence of rearing (*i.e.* standing on rear limbs). The Open Field Test has multiple uses, it is used to asses overall activity and exploration, anxiety, and to assess the sedative, toxic, or stimulant effect of compounds (text adjusted from Gould, Dao, and Kovacsics 2009).

Figure 1.1: Examples of Open Field apparatuses used for behavioural testing in different laboratories (images obtained from Spruijt et al. 2014).

Box 1.4: Elevated Plus Maze

The Elevated Plus Maze is a classical test that is widely applied in mice and rats to quantify anxiety-related behaviour. The Elevated Plus Maze (Figure 1.2) consists of a maze with four arms of which two are open and two provide shelter. Mice or rats are placed at the junction of the four arms of the maze facing an open arm. The main response variables are the number of entries and the time spent in each arm. In addition other behaviour such as rearing and head-dips can be recorded. An increase of time spent in the open arms of the maze is thought to be indicative of a decrease in anxiety. The Elevated Plus Maze can be used to quantify the anti-anxiety effects of drugs and to identify brain regions and mutations related to anxiety-related behaviour (text adjusted from Gould, Dao, and Kovacsics 2009).

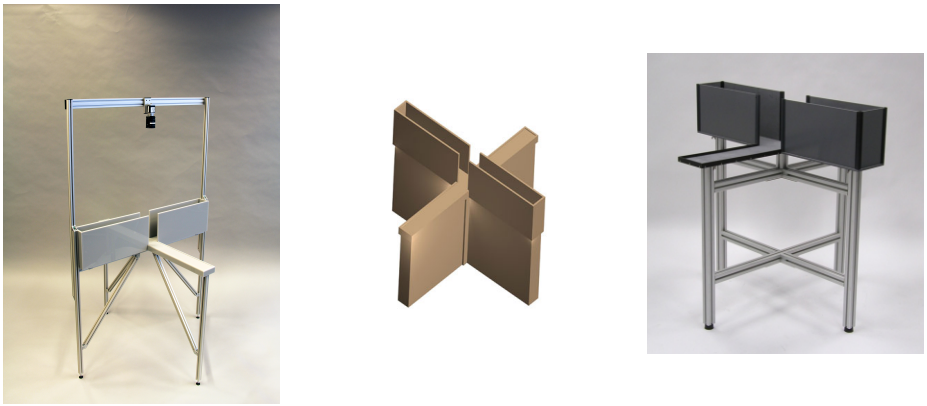
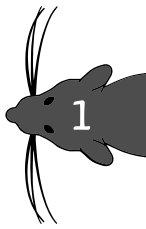


Figure 1.2: Examples of Elevated Plus Maze apparatus used for behavioural testing in different laboratories

Penny 2016). Researchers in the domains Biology and Medicine, relevant for the scope of this thesis, are among the least confident. The lack of replicability is especially alarming in drug-discovery research; pharmaceutical company Bayer has reported to be able to replicate only about 25% of results they read in literature (Prinz, Schlange, and Asadullah 2011).

1.2.2. Replicability in animal behaviour experiments

When conducting animal experiments, the replicability issue is not only pressing from a scientific point-of-view. Ethics and animal welfare regulations prevent scientists from performing unnecessary experiments and thus also from unnec-



essarily duplicating previous experiments. Numerous studies on the replicability of the classical tests for animal behaviour have been conducted that show that results between laboratories truly are poorly replicable. An extensive and highly cited study by Crabbe (1999) in which the researchers "*...went to extraordinary lengths to equate test apparatus, testing protocols, and all possible features of animal husbandry*" found large differences between laboratories for nearly all tested variables and found that the pattern of differences between the tested strains varied substantially among the sites. In a follow-up paper with more extensive analysis the authors emphasize that the main problem was in recovering strain effects of moderate size (Wahlsten et al. 2003). The replicability of behavioural parameters seems to widely vary between parameters as has been shown in many studies (Ennaceur and Chazot 2016). Differences in locomotor activity between inbred strains in the open field test have been found replicable across laboratories and also over decades (Wahlsten et al. 2006) whereas laboratory x strain interactions effects are for instance commonly reported for the Rotarod test (Mandillo et al. 2008).

1.2.3. Replicability in clinical trials

Animal experiments are a crucial and obligatory step towards developing drugs for use in humans. Results of animal experiments thus must not only be replicable between laboratories but also between species. Before being allowed for use in humans, new drugs are extensively tested in the multiple phases of clinical trials. Only a small percentage of drugs that enters the first phase of a clinical trial is ever approved for marketing in humans. For psychiatric drugs the success-rate is especially low (Thomas et al. 2016, Definitions and succes-rates per phase in Table 1.2.3).

Over the last decade, in the domain of psychiatry, the chance for a drug that entered Phase I of clinical trials to be approved for the American market was 6.2%. Only the domain of oncology (5.1%) had a lower percentage of approved drugs; in other domains the percentages reached up to 26.1% (haematology). Phase success for psychiatric drugs is especially low in Phase I and II, it is the lowest of all disease domains. As Phase I is the first safety test on humans and Phase II is the first proof-of-principle, the lack of success in these phases indicates that results obtained in tests on laboratory animals translate poorly to humans.

In conclusion, there exists a replicability problem in science in general. This holds especially for animal behaviour experiments and classical tests. Alongside this issue, results from animal experiments only limited result in the development of safe and effective drugs for psychiatric diseases in humans. These issues are alarming. In the next section of this introduction we discuss the limitations

of classical tests specifically. Thereafter we introduce the automated home cage experiments that have been proposed as a solution.

1.3. Limitations of classical tests

Lack of validity and replicability in the classical behavioural tests has been attributed to numerous causes. Here we provide an overview of the different concerns and explanations given in the literature for the limited reliability and replicability of results from behavioural phenotyping using classical tests. These issues have been subdivided into conceptual issues that threaten the interpretation of results, issues regarding the execution and design of the tests, issues relating to the recording and description of the observed behaviour, and issues regarding the analysis of results.

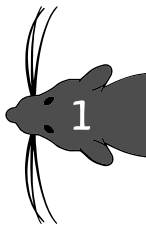
1.3.1. Conceptual

Misinterpretation of results

It has been argued that classical behavioural tests are not suitable in all scientific contexts of behavioural phenotyping (Spruijt et al. 2014). As mentioned earlier, in classical tests the complexity of behaviour is summarized into abstract parameters without a direct ethological interpretation. This makes classical behavioural tests a suitable instrument for behavioural bio-assays in which behaviour is used mainly to quantify the effects of an intervention. For instance, when one is interested in determining the dose-response curve of a certain drug. In these studies, the behavioural parameters do not need to have an ethological interpretation. Often however, we wish to gain insight into the biological function of the changes in behaviour as a result of the treatment, not in a mere quantification of the effect

Table 1.1: Results of analysis of clinical development success rates in the period 2006-2015 executed by the Biotechnology Innovation Organization (Thomas et al. 2016). Percentages are given for the domain of psychiatric diseases and in parenthesis for all disease areas combined. Phase success is the percentage of drugs that moved to the next stage and Approval rate indicates the percentage of drugs in this state that eventually got approved for marketing in the USA

Definition	Phase success	Approval rate
Phase I: Testing on healthy volunteers; safety and dose-ranging	53.9% (63.2%)	6.2% (9.6%)
Phase II: Testing on patients; safety and proof-of-principle	23.7% (30.7%)	11.6% (15.3%)
Phase III: Testing on patients; determine therapeutic effect	55.7% (58.1%)	49.0% (49.6%)
Application for approval to the FDA	87.9% (85.3%)	87.9% (85.3%)
Approved for marketing (USA)		



a single parameter. For instance, when the effect of a mutation on a certain gene is studied.

Looking at behaviour in these contexts using classical tests can lead to misinterpretation of results. For example, in the Elevated Plus Maze, the parameter "time spent in the open arm" has been shown to differentiate between more or less anxious individuals. This however, does not imply that an increase of time spent in the open arm can always be interpreted as the animals being less anxious. It also does not imply that the only reason that animals show different behaviour is because of differences in anxiety (Tecott and Nestler 2004).

The former mechanism can be illustrated using freezing behaviour. Freezing behaviour is a characteristic response to threatening stimuli in rodents. Some mice exhibit freezing behaviour when placed directly in the open arm of the Elevated Plus Maze. Mice showing freezing behaviour spend more time in the open arm but should not be considered to be less anxious. The latter mechanism is caused by classical tests being targeted to and validated to estimate single behavioural constructs. When using the Elevated Plus Maze for behavioural screening of a new mutant line, deviating results could be caused by reduced or increased anxiety compared to the control. There could also be numerous other reasons for instance: cognitive impairment that causes problems when processing contextual cues; general activity levels; locomotor issues; or blindness. The risk of misinterpretation of behavioural test results is highest when their results are interpreted on a test by test basis without relating their results to the results of other tests.

What are we really measuring?

A behavioural phenotype for a certain construct, *e.g.* anxiety, is thought to be the product of state and trait. State anxiety is behavioural response to the testing circumstances whereas trait anxiety is a stable biological trait that is present regardless of the testing circumstances (Andreatini and Bacellar 2000; Lister 1990). State anxiety is highly influenced by environmental variables and the influences of these environmental variables can interact. The relative importance of interaction effects of environmental variables has been found to be much greater compared to their main effects, in behavioural a well as physiological parameters (Valdar et al. 2006). Interactions occur between the different environmental effects but also between genotype and environment. A photo-sensitive albino strain would be affected more by brightly illuminated testing environment than a non-albino strain. And to complicate the situation more, state anxiety can also be affected by the interaction of environmental varibles and trait anxiety. More anxious individuals will have been anxious in their lifetime prior to testing more often than less anxious individuals. In other words, more anxious individuals are

more experienced at dealing with anxiety than less anxious individuals, which has effects on behavioural phenotype (Fonio, Benjamini, and Golani 2012).

Ethological validity

The life of a laboratory rat or mouse is by no means comparable to how rats and mice live in nature. For behavioural phenotyping however, citetPeters2015b argue that some degree of "ethological validity" is required. Examples of lack of ethological validity in behavioural phenotyping are widespread. Mice in nature do not swim whereas Water Maze tests are routinely being performed in mice (Bannerman et al. 2014; Webster et al. 2014) and both rats and mice are nocturnal animals and should thus not be tested in brightly illuminated testing rooms.

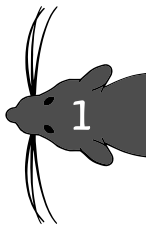
1.3.2. Execution and design of experiments

Environmental Confounding factors

As mentioned before, environmental factors can severely influence animal behaviour (Sousa, Almeida, and Wotjak 2006). These factors can be used to influence the behavioural phenotype intentionally, *e.g.* using deprivation of maternal care as an animal model for depression. Environmental factors also are a source of confounding and decreased replicability between experiments. Confounding environmental variables in classical tests can be divided in four categories:

Ontogeny and development Environmental factors, even in early-life, can influence the (development of) the central nervous system and thus change the behavioural phenotype (Kempermann, Kuhn, and Gage 1997; Laviola 1996; Pacteau, Einon, and Sinden 1989). For example, enrichment of the home cage has been shown to increase the number of neurons in the hippocampus and improve results in the Morris Water Maze test (Kempermann, Kuhn, and Gage 1997). These effects remain present long after the enrichment has been removed: enrichment of the home cage for eight weeks has been shown to influence results in the Open Field test six months later (Amaral et al. 2008).

Housing and laboratory The effects of factors such as cage size, lighting conditions, temperature, diet, background noise, and olfactory cues in the laboratory on outcomes of behavioural tests have been long recognized (Crabbe 1999; McClearn 2004; Walsh and Cummins 1976). For instance, alarm pheromones in the urine of fearful mice cause increased locomotion (Cocke et al. 1993; Whittier and McReynolds 1965). Laboratories vary widely in animal husbandry routines and thus also in their effort towards *e.g.* preventing transmission of smells (López-Salesansky et al. 2016) which increases variability.



Human handling and presence Human interference and animal handling is an important component of classical behavioural tests because animals need to be picked up from their home cage and transported towards the testing environment. As a reaction to handling, laboratory animals show indicators of stress and emotional distress (Brudzynski and Ociepa 1992; Gärtner et al. 1980). Habituation to gentle human handling has significant effects on performance in the Elevated Plus Maze (Hogg 1996). Stress and fear caused by handling can limit the validity of research data (Sherwin, 2004). In addition to the effects of human presence altogether, experimenter identity has been shown to exceed genetic effects as the most prominent explanatory factor (Chesler et al. 2002). For instance, presence of a familiar caretaker increases time spent on that side of the Open Field Test (*i.e.*caretaker effect, McCall, Lester, and Corter 1969).

Testing circumstances The exact circumstances under which a behavioural test is performed are crucial for the results: "*If two test situations are substantially different, such as a small, square box in the dark and a large, round open field under bright lights,activity in the two situations may be thought of as two different phenotypes.*"(Wahlsten et al. 2003).

Standardization

The above-mentioned confounding covariates influence the results of behavioural experiments. Variability between laboratories and experiments with regards to these factors thus increase the variability of results and decreases replicability. Standardization has often been proposed as a solution to reduce this variation and improve replicability of results (Beynen, Gärtner, and Zutphen 2001; Wahlsten 2001). Even after extensive standardization however, variation in results between laboratories persists (Crabbe 1999). Standardization has also been proposed as a factor that reduces validity of animal experiments. Complete standardization might result in reproducible results under very specific circumstances but also in results that do not extrapolate well (Wurbel 2000). A more pragmatic approach towards handling environmental influences is extensive documentation of and reporting on all potential confounding factors when reporting results. The disadvantage is that not all potential confounding factors are known and that not all reported factors are relevant (Wurbel 2002). It has been suggested to increase replicability of results by specifically modelling the genotype x laboratory component via a mixed modelling approach (Kafkafi et al. 2005). This approach has since been expanded to a method to calculate genotype x laboratory adjusted p-values that indicates the probability of replicating the result across laboratories (Kafkafi et al. 2017). Adjusting p-values, and thus reducing statistical power however, does not address more fundamental issues with animal behaviour testing.

Limited observation times

Duration of behavioural tests is often short. Initially, the Open Field Test protocol suggested a mere three minute observation time (Hall and Ballachey 1932). Nowadays, observations typically last ten to fifteen minutes. The short observation times prevent detection of effects of an intervention in the long-term and prevent the full observation of behavioural processes that last longer than the test period. The habituation process of mice for instance, has been shown to last several days (Spruijt et al. 2014). Short observation times cannot take into account circadian rhythm which is an important driving force in animal behaviour. Time of day of testing has been shown to influence results. Relating back to misinterpretation of results, the short duration of classical tests also causes the novelty effect of the test environment to interact with the behavioural construct of interest. The difference in anxiety between strains interacts with duration of the trial (Fonio, Benjamini, and Golani 2012).

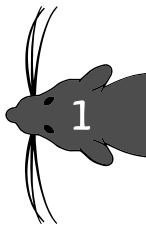
1.3.3. Data recording

Human error and bias

Classical tests usually require the experimenter to record and classify behaviour. As mentioned previously, experimenter identity is an important source of variation in animal experiments. This is not only due to the confounding effects described earlier but also due to limitations of human observers. Human observers can disagree, make mistakes, and be biased (Bohlen et al. 2014). Experimenter disagreement occurs when the same behaviour is judged differently by different observers. For example, in the seemingly easy tasks of classifying behaviours into categories such as drinking, eating, and grooming an observed inter-experimenter agreement around 70% is not uncommon (Jhuang et al. 2010). Consistency between experimenters decreases when experimenters are unfamiliar with the animals (Driel and Talling 2005). Alongside the random error caused by experimenter disagreement, human experimenters can also invoke bias. Experimenter bias occurs when hypotheses, involuntarily or otherwise, influence the results and is an important, common, and often overlooked issue (Strickland and Mercier 2014). Experimenter bias is especially relevant for classical tests because double blind studies are uncommon and impractical and because interpreting behaviour is a subjective task.

Response Variables

The choice of response variables is an important factor in animal behaviour research. Behavioural phenotypes can only be compared if they are recorded correctly and completely. Several factors have been suggested to improve the quality of description of animal movement (Benjamini et al. 2010).



Use of *ad hoc* criteria Classification of behaviour is often done using *ad hoc* cutoff points. A behaviour is for instance classified as "sitting still" when animal stays in place for 0.5 seconds or as "rearing" if the animal executes movements as described in an ethogram. By using those cutoff points however, nuances of behaviour are lost. Behaviours can vary however between animals and treatments, "stopping" can for instance encompass varying behaviours such as stepping in place and scanning the environment (Kafkafi, Pagis, et al. 2003). In addition, the cutoff point between "stopping" and "moving" in itself can be a variable response variable.

Too few response variables Behaviour is a lot more complex than some indicator parameters. The most frequently used outputs of the open field test, distance moved and time spent in the centre, explain less than 10% of the variability in open field behaviour (Lipkind 2004). Reproducibility of results of open field tests across laboratories has been shown to increase through detailed analysis of the results (Kafkafi et al. 2005; Kafkafi, Lipkind, et al. 2003; Kafkafi, Pagis, et al. 2003).

1.4. Automated home cage experiments

1.4.1. Automated home cage experiments

Automated home cage experiments have been proposed as a solution to some of the issues described in Section 1.2 and thus as an alternative to the classical tests (Gerlai 2002; Kas and Van Ree 2004; Spruijt and Visser 2006; Tecott and Nestler 2004; Wurbel 2002). Automated home cage experiments have two main distinguishing characteristics compared to classical tests: 1) the experiment is conducted in a home cage; 2) behaviour is recorded automatically. Due to these characteristics, experiments can be conducted without human interference and can last several hours, days, or even weeks. Automated home cage experiments provide several advantages which we will discuss in further detail in Section 1.4.2.

Multiple (commercial) implementations and variations of automated home cage systems have been introduced. All data incorporated in this thesis was obtained using the PhenoTyper® (Noldus Information Technology, Wageningen, The Netherlands; Box 1.5). The PhenoTyper system tracks mice or rats using a top-view camera with a resolution of up to 25 frames per second. From this tracking data numerous response variables can be calculated *e.g.* distance moved, number of visits to the shelter, and time spent sitting still per time interval. In this thesis, we focus on these location-based response variables for the description of spontaneous behaviour.

The PhenoTyper system does allow for built-in experiments to study behavioural domains beyond spontaneous behaviour. Several protocols have been validated

to serve as tests for avoidance learning (Maroteaux et al. 2012), instrumental learning (Remmelink et al. 2015), anxiety (Aarts et al. 2015), and to measure attention and impulsivity (Remmelink et al. 2017). As an alternative to location-based parameters, progress has been made in classifying activity patterns into behavioural categories such as rearing, eating, and grooming (Dam et al. 2013; Jhuang et al. 2010).

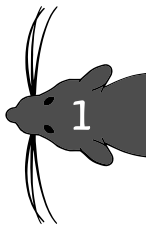
1.4.2. Advantages of automated home cage experiments

The use of automated home cage experiments has several conceptual and practical advantages compared to the classical tests. Conducting experiments in the home cage prevents confounding and stress from animal handling and transport. And it prevents the effect of introduction to a novel environment to interact with the effect of the experimental treatment. Rats for instance have been shown to exhibit dose-response effects of stimulants and sedatives more strongly in their home cage environment than in the open field test (Dunne, O'Halloran, and Kelly 2007).

Automatic recording of behaviour allows for experimenting without human presence. This is advantageous as human presence is an important source of confounding, and because human observers introduce bias and inter-observer differences. Automated recording of behaviour might not be flawless, it is consistent and transparent in its decision making. In addition, automated observation systems can record behaviour for prolonged periods of time and can record numerous response variables simultaneously. EthoVision® (Noldus Information Technology, Wageningen, The Netherlands), the software package associated with the Phenotyper, is a video-based system. Video-based systems provide the benefit of allowing for re-analysis of the video as often as required. This allows for extracting additional response variables from the data after the experiment was conducted, for instance because of new scientific insights or technological advances.

Experimenting in the home cage environment allows animals to perform their full behavioural repertoire uninterrupted which can result in unexpected behaviours. Because automated recording of behaviour allows for registering numerous response variables simultaneously, and for re-analysis if so required, this system is equipped to account for unexpected results. An example of an unexpected result is the remarkably large between-strain difference in the amount of time rats spend on top of their shelter (Loos et al. 2014). This preference is unrelated to motor performance.

Automated home cages are suitable for experiments with much longer durations than classical tests because home cages are designed for long-term housing of animals and automated observation can continue indefinitely. Long term observations are advantageous to eliminate effects of habituation (Fonio, Benjamini,



Box 1.5: PhenoTyper

The PhenoTyper is an observation cage for rats or mice that is completely optimized for video tracking. It has a top unit with a fully integrated infrared sensitive camera, infrared LED lights, an audio stimulus and white and yellow lights that can be controlled automatically. The infrared light makes tracking possible in the dark phase of the animal and makes your setup independent of the light conditions in the lab. The camera images can be used in EthoVision® XT tracking and analysis software and The Observer® XT scoring and analysis software. The lights can be controlled with commands fom EthoVision, for example when the animal enters a certain zone.

Figure 1.3: A mouse in PhenoTyper equipped with a pellet dispenser and water bottle (left) and a shelter (right).

and Golani 2012) and to detect changes in circadian rhythm (Rudenko et al. 2009). Automated home cage systems have been used to show altered habituation phenotype in Rhett syndrom rats over a three hour period whereas no difference in habituation was observed in a ten minute open field (Robinson et al. 2013).

Long-term observations in the home cage also allow for incorporation of baseline behaviour in the analysis. Home cage data has been used in the past as baseline or reference data to help interpret results from the "real" experiment (*i.e.* classical tests) (Ganea et al. 2007; Tang, Orchard, and Sanford 2002; Tang and Sanford 2005). Running wheel activity in the home cage has been used for instance to estimate differences in overall activity between strains of rats to help interpret results of open field tests. In automated home cage experiments the baseline behaviour and the effects of the intervention are measured in the same environment using the same response variables which allows for direct correction.

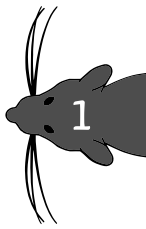
Long-term automated observation of behaviour also allows for the use of individual cut-off values as suggested by Benjamini et al. (2010). Here baseline behaviour is not only used to correct the results, it is used to define parameters. An example is to use an individual's distribution of maximum velocity per movement segment to discriminate between movement segments with a low (lingering) and segments with a high (progressing) velocity. Use of individual cut-off values increases replicability between laboratories in the Open Field test (Lipkind 2004).

1.5. Aim and Scope of the thesis

Automated home cage systems promise to overcome some of the pitfalls of the classical tests. Because the experimental methodology is fundamentally different the collected data are fundamentally different as well. The abstract parameters in the classical tests, with varying success, are aimed to translate into or quantify an ethologically meaningful concept whereas automated home cage experiments allow for the collection of numerous parameters without a direct interpretation. Extensive validation and further work is necessary for the accurate interpretation of the functional relevance of observed effects on behaviour in automated home cage experiments (Tecott and Nestler 2004).

The matter of "how" to best analyse these data however, remains unresolved. This is an issue that has been raised and is continued to be raised ever since automated home cage systems were introduced. For instance by Gerlai (2002): "*The amount of data one gathers using such devices can be staggering. Bioinformatics tools, multivariate statistical methods and pattern analysis can be required to extract information from these complex behavioral experiments properly and concisely.*"; by Spruijt and Visser (2006): "*the issue of obtaining and analysing the appropriate data for behaviour recognition and pattern analysis is still not properly addressed*"; and by Kas et al. (2014): "*However, a caveat of automated home cage testing is that it can yield massive amounts of complex data. This calls for novel analysis methods and endpoints*".

The statistical toolbox of those analysing data from automated home cage experiments typically contains univariate analysis (such as Anova) on single parameters (e.g. Aziriova et al. 2016; Loos et al. 2013; Robinson et al. 2013). Sometimes



combined with more advanced data pre-processing methods such as smoothing, threshold statistics, and PCA.

Spruijt and Visser (2006) have even suggested that the use of less complex and sophisticated techniques to study these data is caused by a psychological barrier. Behaviour can be observed easily and explained intuitively and thus its complexity and the need for complex analysis methods is underestimated.

In this thesis I explore, apply, and expand on some more elaborate methodology to analyse results of automated home cage experiments. The aim is to showcase the potential of these more sophisticated analyses that are more suitable for the complexity of the data at hand.

In **Chapter 2**, I propose and illustrate the use of Redundancy Analysis (RDA) and Principal Response Curves (PRC) for analysing data from automated home cage experiments. Both are multivariate analyses that allow us to describe effects of a treatment on all response variables simultaneously. These techniques have been well established in other applications and are easy to implement and interpret. I show that the same conclusions can be drawn from univariate and multivariate analysis, and that the multivariate analysis has the added advantage of visualisation of the data and the potential to detect effects in more than one dimension.

In **Chapter 3**, I propose an extension to PRC that allows for response variable selection using permutation testing. PRC has been widely applied in (aquatic) ecology and microbiology to visualize the overall effect of a experimental treatment over time over a collection or response variables. In its traditional applications, these response variables are often-times abundances of species or taxa of micro-organisms. For behavioural parameters, it could be desirable to reduce the model such that only those response variables that correlate to the overall response in the data set remain.

In **Chapter 4**, I analyse data from an automated home cage experiment with two rats per cage and two streams of information: activity data per rat and Ultrasonic Vocalisations (USVs) per cage. Rats in the same cage could either interact or were separated by a screen. The aim was to predict the USV-rate per rat given its activity and study the effect of social interaction. I demonstrate that the underlying mechanistic model that links USVs to activity fundamentally differs between rats that could interact and rats that could not. This chapter illustrates the power of combining mechanistic and statistical modelling to pinpoint effects that cannot be observed from the data set as such.

In **Chapter 5**, I describe the results of a simulation study in which Targeted Maximum Likelihood Estimation (TMLE) is used in sub-optimal circumstances.

TMLE has been proposed as a solution for estimation of average causal effects in presence of confounding covariates and is indeed efficient when the theoretical assumptions are not violated and data sets are sufficiently large. In practice however, the assumptions can be easily violated and some cannot even be checked. This chapter shows that violations of the positivity assumption can have detrimental effects on bias, RMSE, and coverage of the estimates. And that the size of "sufficiently large" is well beyond the scale of what would be reasonable to see in animal experiments. It serves as a cautionary tale for application of more advanced data and also illustrates the power of simulation studies.

In **Chapter 6**, the General Discussion, I summarise and discuss the results in this thesis.

References

Aarts, Emmeke, Gregoire Maroteaux, Maarten Loos, Bastijn Koopmans, Jovana Kovačević, August B. Smit, Matthijs Verhage, and Sophie van der Sluis. 2015. "The light spot test: Measuring anxiety in mice in an automated home-cage environment." *Behavioural Brain Research* 294 (November): 123–130.

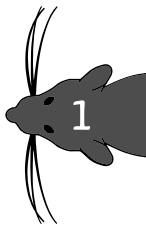
Amaral, Olavo B., Rafael S. Vargas, Gisele Hansel, Iván Izquierdo, and Diogo O. Souza. 2008. "Duration of environmental enrichment influences the magnitude and persistence of its behavioral effects on mice." *Physiology and Behavior* 93 (1-2): 388–394.

Andreatini, Roberto, and Leila F.S. Bacellar. 2000. "Animal models: Trait or state measure? The test-retest reliability of the elevated plus-maze and behavioral despair." *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 24, no. 4 (May): 549–560.

Aziriova, S., K. Repova, K. Krajcirovicova, T. Baka, S. Zorad, V. Mojto, P. Slavkovsky, et al. 2016. "Effect of ivabradine, captopril and melatonin on the behaviour of rats in L-nitro-arginine methyl ester-induced hypertension." *Journal of Physiology and Pharmacology* 67 (6): 895–902.

Baker, Monya, and Dan Penny. 2016. "1,500 scientists lift the lid on reproducibility." *Nature* 533 (7604): 452–454.

Bannerman, David M., Rolf Sprengel, David J. Sanderson, Stephen B. McHugh, J. Nicholas P. Rawlins, Hannah Monyer, and Peter H. Seeburg. 2014. "Hippocampal synaptic plasticity, spatial memory and anxiety." *Nature Reviews Neuroscience* 15, no. 3 (February): 181–192.



Belzung, Catherine, and Maël Lemoine. 2011. "Criteria of validity for animal models of psychiatric disorders: focus on anxiety disorders and depression." *Biology of Mood & Anxiety Disorders* 1 (1): 9.

Benjamini, Yoav, Dina Lipkind, Guy Horev, Ehud Fonio, Neri Kafkafi, and Ilan Golani. 2010. "Ten ways to improve the quality of descriptions of whole-animal movement." *Neuroscience & Biobehavioral Reviews* 34, no. 8 (July): 1351–1365.

Beynen, A.C., K. Gärtner, and L.F.M. van Zutphen. 2001. "Standardization of animal experimentation." In *Principles of Laboratory Animal Science: A Contribution to the Humane Use and Care of Animals and to the Quality of Experimental Results, revised edition,* edited by L F M van Zutphen, V Baumans, and A C Beynen, 103–110. Van Zutphen, Principles of Laboratory Animal Science. Amsterdam: Elsevier.

Bohlen, Martin, Erika R. Hayes, Benjamin Bohlen, Jeremy D. Bailoo, John C. Crabbe, and Douglas Wahlsten. 2014. "Experimenter effects on behavioral test scores of eight inbred mouse strains under the influence of ethanol." *Behavioural Brain Research* 272, no. 10 (October): 46–54.

Brudzynski, Stefan M., and Dorota Ociepa. 1992. "Ultrasonic vocalization of laboratory rats in response to handling and touch." *Physiology & Behavior* 52, no. 4 (October): 655–660.

Chesler, Elissa J., Sonya G. Wilson, William R. Lariviere, Sandra L. Rodriguez-Zas, and Jeffrey S. Mogil. 2002. "Influences of laboratory environment on behavior." *Nature Neuroscience* 5, no. 11 (November): 1101–1102.

Cocke, R., J.A. Moynihan, N. Cohen, L.J. Grota, and R. Ader. 1993. "Exposure to Conspecific Alarm Chemosignals Alters Immune Responses in BALB/c Mice." *Brain, Behavior, and Immunity* 7, no. 1 (March): 36–46.

Crabbe, John C. 1999. "Genetics of Mouse Behavior: Interactions with Laboratory Environment." *Science* 284, no. 5420 (June): 1670–1672.

Dam, Elsbeth A. van, Johanneke E. van der Harst, Cajo J F Ter Braak, Ruud A J Tegelenbosch, Berry M. Spruijt, and Lucas P J J Noldus. 2013. "An automated system for the recognition of various specific rat behaviours." *Journal of Neuroscience Methods* 218 (2): 214–224.

Darwin, Charles. 1872. *The expression of the emotions in man and animals.,* chap. Chapter 14, vol. 3rd ed, 351. London: John Murray.

Driel, Katja S van, and Janet C. Talling. 2005. "Familiarity increases consistency in animal tests." *Behavioural Brain Research* 159, no. 2 (April): 243–245.

Dunne, Fergal, Ambrose O'Halloran, and John P. Kelly. 2007. "Development of a home cage locomotor tracking system capable of detecting the stimulant and sedative properties of drugs in rats." *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 31, no. 7 (October): 1456–1463.

Ennaceur, Abdelkader, and Paul L. Chazot. 2016. "Preclinical animal anxiety research - flaws and prejudices." *Pharmacology Research & Perspectives* 4, no. 2 (April): e00223.

Fonio, Ehud, Yoav Benjamini, and Ilan Golani. 2012. "Short and long term measures of anxiety exhibit opposite results." Edited by Abraham A. Palmer. *PloS one* 7, no. 10 (January): e48414.

Ganea, K., C. Liebl, V. Sterlemann, M.B. Müller, and M.V. Schmidt. 2007. "Pharmacological validation of a novel home cage activity counter in mice." *Journal of Neuroscience Methods* 162, nos. 1-2 (May): 180–186.

Gärtner, K., D. Büttner, K. Döhler, R Friedel, J Lindena, and I Trautschold. 1980. "Stress response of rats to handling and experimental procedures." *Laboratory Animals* 14, no. 3 (July): 267–274.

Gerlai, Robert. 2002. "Phenomics: fiction or the future?" *Trends in Neurosciences* 25, no. 10 (October): 506–509.

Gould, Todd D, David T Dao, and Colleen E Kovacsics. 2009. "The Open Field Test." In *Mood and Anxiety Related Phenotypes in Mice: Characterization Using Behavioral Tests,* edited by Todd D Gould, 1–20. Totowa, NJ: Humana Press.

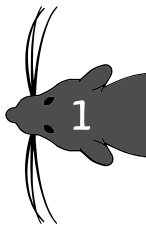
Hall, C. S. 1934. "Drive and emotionality: factors associated with adjustment in the rat." *Journal of Comparative Psychology* 17 (1): 89–108.

Hall, C., and E. L. Ballachey. 1932. *A study of the rat's behavior in a field. A contribution to method in comparative psychology.,* 1–12.

Haller, Jozsef, and Mano Alicki. 2012. "Current animal models of anxiety, anxiety disorders, and anxiolytic drugs." *Current Opinion in Psychiatry* 25, no. 1 (January): 59–64.

Heim, Christine, and Charles B. Nemeroff. 2001. "The role of childhood trauma in the neurobiology of mood and anxiety disorders: preclinical and clinical studies." *Biological Psychiatry* 49, no. 12 (June): 1023–1039.

Hogg, Sandy. 1996. "A review of the validity and variability of the Elevated Plus-Maze as an animal model of anxiety." *Pharmacology Biochemistry and Behavior* 54, no. 1 (May): 21–30.



Jhuang, Hueihan, Estibaliz Garrote, Jim Mutch, Xinlin Yu, Vinita Khilnani, Tomaso Poggio, Andrew D Steele, and Thomas Serre. 2010. "Automated home-cage behavioural phenotyping of mice." *Nature communications* 1 (January): 68.

Kafkafi, Neri, Yoav Benjamini, Anat Sakov, Greg I Elmer, and Ilan Golani. 2005. "Genotype-environment interactions in mouse behavior: a way out of the problem." *Proceedings of the National Academy of Sciences of the United States of America* 102 (12): 4619–4624.

Kafkafi, Neri, Ilan Golani, Iman Jaljuli, Hugh Morgan, Tal Sarig, Hanno Würbel, Shay Yaacoby, and Yoav Benjamini. 2017. "Addressing reproducibility in single-laboratory phenotyping experiments." *Nature Methods* 14 (5): 462–464.

Kafkafi, Neri, Dina Lipkind, Yoav Benjamini, Cheryl L Mayo, Gregory I Elmer, and Ilan Golani. 2003. "SEE locomotor behavior test discriminates C57BL/6J and DBA/2J mouse inbred strains across laboratories and protocol conditions." *Behavioral Neuroscience* 117 (3): 464–477.

Kafkafi, Neri, Michal Pagis, Dina Lipkind, Cheryl L. Mayo, Yoav Bemjamini, Ilan Golani, and Gregory I. Elmer. 2003. "Darting behavior: a quantitative movement pattern designed for discrimination and replicability in mouse locomotor behavior." *Behavioural Brain Research* 142, nos. 1-2 (June): 193–205.

Kas, Martien J., Jeffrey C. Glennon, Jan Buitelaar, Elodie Ey, Barbara Biemans, Jacqueline Crawley, Robert H. Ring, et al. 2014. "Assessing behavioural and cognitive domains of autism spectrum disorders in rodents: Current status and future perspectives." *Psychopharmacology* 231, no. 6 (March): 1125–1146.

Kas, Martien J.H., and Jan M. Van Ree. 2004. "Dissecting complex behaviours in the post-genomic era." *Trends in Neurosciences* 27, no. 7 (July): 366–369.

Kempermann, G., H. G. Kuhn, and F. H. Gage. 1997. "More hippocampal neurons in adult mice living in an enriched environment." *Nature* 386 (6624): 493–495.

Laviola, Giovanni. 1996. "On mouse pups and their lactating dams: Behavioral consequences of early exposure to oxazepam and interacting factors." *Pharmacology Biochemistry and Behavior* 55 (4): 459–474.

Lipkind, Dina. 2004. "New replicable anxiety-related measures of wall vs. center behavior of mice in the open field." *Journal of Applied Physiology* 97, no. 1 (March): 347–359.

Lister, Richard G. 1990. "Ethologically-based animal models of anxiety disorders." *Pharmacology & Therapeutics* 46, no. 3 (January): 321–340.

Loos, M., J. Staal, A.B. Smit, T.J. de Vries, and S. Spijker. 2013. "Enhanced alcohol self-administration and reinstatement in a highly impulsive, inattentive recombinant inbred mouse strain." *Frontiers in Behavioral Neuroscience,* no. OCT.

Loos, Maarten, Bastijn Koopmans, Emmeke Aarts, Gregoire Maroteaux, Sophie van der Sluis, Matthijs Verhage, and August B. Smit. 2014. "Sheltering Behavior and Locomotor Activity in 11 Genetically Diverse Common Inbred Mouse Strains Using Home-Cage Monitoring." Edited by Yann Herault. *PLoS ONE* 9, no. 9 (September): e108563.

López-Salesansky, Noelia, Nur H Mazlan, Lucy E Whitfield, Dominic J Wells, and Charlotte C Burn. 2016. "Olfaction variation in mouse husbandry and its implications for refinement and standardization: UK survey of animal scents." *Laboratory Animals* 50, no. 5 (October): 362–369.

Mandillo, Silvia, Valter Tucci, S. M. Holter, Hamid Meziane, Mumna Al Banchaabouchi, Magdalena Kallnik, Heena V Lad, et al. 2008. "Reliability, robustness, and reproducibility in mouse behavioral phenotyping: a cross-laboratory study." *Physiological Genomics* 34, no. 3 (June): 243–255.

Maroteaux, G, M Loos, S van der Sluis, B Koopmans, E Aarts, K van Gassen, A Geurts, et al. 2012. "High-throughput phenotyping of avoidance learning in mice discriminates different genotypes and identifies a novel gene." *Genes, brain, and behavior* 11, no. 7 (October): 772–84.

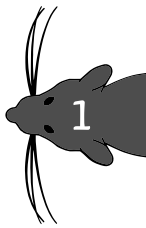
McCall, R B, M L Lester, and C M Corter. 1969. "Caretaker effect in rats." *Developmental Psychology* 1 (6 PART 1): 771.

McClearn, Gerald E. 2004. "Nature and nurture: Interaction and coaction." *American Journal of Medical Genetics* 124B, no. 1 (January): 124–130.

Pacteau, C., D. Einon, and J. Sinden. 1989. "Early rearing environment and dorsal hippocampal ibotenic acid lesions: long-term influences on spatial learning and altenation in the rat." *Behavioural Brain Research* 34 (1-2): 79–96.

Porsolt, R. D., M. Le Pichon, and M. Jalfre. 1977. "Depression: a new animal model sensitive to antidepressant treatments." *Nature* 266, no. 5604 (April): 730–732.

Porsolt, Roger D., Guy Anton, Nadine Blavet, and Maurice Jalfre. 1978. "Behavioural despair in rats: A new model sensitive to antidepressant treatments." *European Journal of Pharmacology* 47, no. 4 (February): 379–391.



Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe it or not: how much can we rely on published data on potential drug targets?" *Nature Reviews Drug Discovery* 10, no. 9 (August): 712–712. eprint: 9907372v1 (cond-mat).

Remmelink, Esther, Uyen Chau, August B. Smit, Matthijs Verhage, and Maarten Loos. 2017. "A one-week 5-choice serial reaction time task to measure impulsivity and attention in adult and adolescent mice." *Scientific Reports* 7 (February): 42519.

Remmelink, Esther, Maarten Loos, Bastijn Koopmans, Emmeke Aarts, Sophie van der Sluis, A.B. August B. Smit, Matthijs Verhage, et al. 2015. "A 1-night operant learning task without food-restriction differentiates among mouse strains in an automated home-cage environment." *Behavioural Brain Research* 283 (April): 53–60.

Réus, Gislaine Z., Roberto B. Stringari, Karine F. Ribeiro, Andreza L. Cipriano, Bruna S. Panizzutti, Laura Stertz, Camila Lersch, Flávio Kapczinski, and João Quevedo. 2011. "Maternal Deprivation Induces Depressive-like Behaviour and Alters Neurotrophin Levels in the Rat Brain." *Neurochemical Research* 36, no. 3 (March): 460–466.

Robinson, Lianne, Andrea Plano, Stuart Cobb, and Gernot Riedel. 2013. "Long-term home cage activity scans reveal lowered exploratory behaviour in symptomatic female Rett mice." *Behavioural Brain Research* 250 (August): 148–156.

Rudenko, Olga, Vadim Tkach, Vladimir Berezin, and Elisabeth Bock. 2009. "Detection of early behavioral markers of Huntington's disease in R6/2 mice employing an automated social home cage." *Behavioural Brain Research* 203, no. 2 (November): 188–199.

Shelton, Richard C. 2007. "The Molecular Neurobiology of Depression." *Psychiatric Clinics of North America* 30, no. 1 (March): 1–11.

Sousa, N., O. F X Almeida, and C. T. Wotjak. 2006. "A hitchhiker's guide to behavioral analysis in laboratory rodents." *Genes, Brain and Behavior* 5, no. SUPPL. 2 (June): 5–24.

Spruijt, Berry M., Suzanne M. Peters, Raymond C. de Heer, Helen H.J. Pothuizen, and Johanneke E. van der Harst. 2014. "Reproducibility and relevance of future behavioral sciences should benefit from a cross fertilization of past recommendations and today's technology: "Back to the future"." *Journal of Neuroscience Methods* 234 (August): 2–12.

Spruijt, Berry M., and Leonie de Visser. 2006. "Advanced behavioural screening: automated home cage ethology." *Drug Discovery Today: Technologies* 3, no. 2 (June): 231–237.

Steru, Lucien, Raymond Chermat, Bernard Thierry, and Pierre Simon. 1985. "The tail suspension test: A new method for screening antidepressants in mice." *Psychopharmacology* 85, no. 3 (March): 367–370.

Strickland, Brent, and Hugo Mercier. 2014. "Bias neglect: A blind spot in the evaluation of scientific results." *The Quarterly Journal of Experimental Psychology* 67, no. 3 (March): 570–580.

Tang, Xiangdong, Stuart M Orchard, and Larry D Sanford. 2002. "Home cage activity and behavioral performance in inbred and hybrid mice." *Behavioural Brain Research* 136, no. 2 (November): 555–569.

Tang, Xiangdong, and Larry D Sanford. 2005. "Home cage activity and activity-based measures of anxiety in 129P3/J, 129X1/SvJ and C57BL/6J mice." *Physiology & Behavior* 84, no. 1 (January): 105–115.

Tecott, Laurence H, and Eric J Nestler. 2004. "Neurobehavioral assessment in the information age." *Nature Neuroscience* 7, no. 5 (May): 462–466.

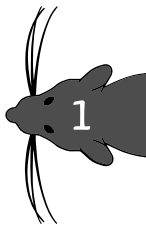
Thomas, David W., Justin Burns, John Audette, Adam Carroll, Corey Dow-Hygelund, and Michael Hay. 2016. *Clinical development success rates 2006-2015.* Technical report. Biotechnology Innovation Organization (BIO), Washington, May.

Valdar, William, Leah C. Solberg, Dominique Gauguier, William O. Cookson, J. Nicholas P Rawlins, Richard Mott, and Jonathan Flint. 2006. "Genetic and Environmental Effects on Complex Traits in Mice." *Genetics* 174, no. 2 (September): 959–984.

Wahlsten, Douglas. 2001. "Standardizing tests of mouse behavior: Reasons, recommendations, and reality." *Physiology & Behavior* 73, no. 5 (August): 695–704.

Wahlsten, Douglas, Alexander Bachmanov, Deborah a. Finn, and John C. Crabbe. 2006. "Stability of inbred mouse strain differences in behavior and brain size between laboratories and across decades." *Proceedings of the National Academy of Sciences* 103, no. 44 (October): 16364–16369.

Wahlsten, Douglas, Pamela Metten, Tamara J. Phillips, Stephen L. Boehm, Sue Burkhart-Kasch, Janet Dorow, Sharon Doerksen, et al. 2003. "Different data from different labs: Lessons from studies of gene-environment interaction." *Journal of Neurobiology* 54 (1): 283–311.



Walsh, Roger N, and Robert A Cummins. 1976. "The open-field test: A critical review." *Psychological Bulletin* 83, no. 3 (May): 482–504.

Webster, Scott J., Adam D. Bachstetter, Peter T. Nelson, Frederick A. Schmitt, and Linda J. Van Eldik. 2014. "Using mice to model Alzheimer's dementia: an overview of the clinical disease and the preclinical behavioral changes in 10 mouse models." *Frontiers in Genetics* 5, no. APR (April): 1–23.

Whittier, James L, and Paul McReynolds. 1965. "Persisting odors as a biasing factor in open-field research with mice." *Canadian Journal of Psychology/Revue canadienne de psychologie* 19, no. 3 (September): 224–230.

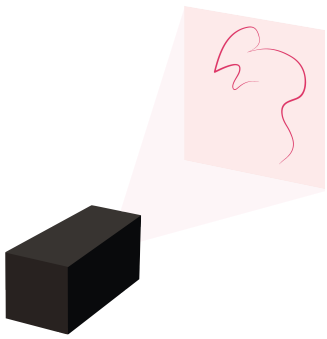
Willner, P. 1991. *Behavioural Models in Psychopharmacology: Theoretical, Industrial and Clinical Perspectives.* Cambridge University Press.

Willner, Paul, Richard Muscat, and Mariusz Papp. 1992. "Chronic mild stress-induced anhedonia: A realistic animal model of depression." *Neuroscience & Biobehavioral Reviews* 16, no. 4 (January): 525–534.

Wurbel, H. 2000. "Behaviour and the standardization fallacy." *Nature genetics* 26 (3): 263.

Wurbel, H. 2002. "Behavioral phenotyping enhanced - beyond (environmental) standardization." *Genes, Brain and Behavior* 1, no. 1 (January): 3–8.

Yankelevitch-Yahav, Roni, Motty Franko, Avrham Huly, and Ravid Doron. 2015. "The Forced Swim Test as a Model of Depressive-like Behavior." *Journal of Visualized Experiments,* no. 97 (March): 1–7.



2

Multivariate analysis of automated home cage experiments

Nadia J. Vendrig

Lia Hemerik

Linde Boekhoudt[1]

Neuro-BSIK Mouse Phenomics Consortium[2]

Cajo J.F. ter Braak

1. Brain Center Rudolf Magnus, Department of Translational Neuroscience, University Medical Center Utrecht, Utrecht, The Netherlands

2. Membership of the consortium is provided in the Acknowledgments.

Data from Automated Home Cage Experiments is predominantly analysed using univariate statistics on one or a few response variables. One of the main advantages of Automated Home Cage Experiments is the potential to gather large numbers of response variables. Using multivariate statistics to analyse these data allows for analysis of all the response variables simultaneously. Here, we introduce the multivariate methods Redundancy Analysis (RDA) and Principal Response Curves (PRC) and demonstrate their potential in two case studies. Both RDA and PRC are frequently used in (aquatic) ecology, toxicology, and microbiology. RDA is a constrained form of Principal Components Analysis (PCA). RDA describes the underlying structure of a data set in terms of the explanatory variables (such as experimental treatment). It quantifies the proportion of variance in the data set that can be described using these explanatory variables. PRC is a special case of RDA used to describe experimental multivariate longitudinal data. It estimates differences among treatments on a collection of response variables over time and the extent to which the response of those individual response variables resembles the overall response. In both case studies, the multivariate analyses were able to draw the same main conclusions as the contrasting univariate analyses. The advantages of using a multivariate analysis rather than a univariate analysis on a single response variable is that the multivariate methods provide a graphical representation of the data set, are easy to interpret, and allow for estimation of the relation between response variables.

2.1. INTRODUCTION

Automated home cage experiments have been proposed as an alternative to classical behavioural tests in animals (*e.g.* the open field test and elevated plus maze) to meet some of the concerns regarding the interpretation and reproducibility of animal behaviour experiments (Gerlai 2002; Kas and Ree 2004; Spruijt and Visser 2006; Tecott and Nestler 2004; Wurbel 2002). Automated home cage systems allow for long-term continuous monitoring of home cage behaviour with minimal human intervention.

Several automated home cage systems for rats and mice have been developed. Animal movement can be tracked using video observation (PhenoTyper®; Visser, Bos, and Spruijt 2005), infra-red sensors (PhenoMaster), or via registration of micro-chips implanted in the animals (IntelliCage; Krackow et al. 2010). Whilst these systems have different technical implementations, their results seem relatively robust between systems (Robinson and Riedel 2014). Here, we focus on analysis of data from the PhenoTyper system, which, due to the higher spatial resolution of video-tracking, has been suggested to be more sensitive to drug treatments (Robinson and Riedel 2014).

Automated home cage experiments have several advantages in comparison to classical behavioural tests. Experimenting within the home cage rather than in a separate test environment reduces the confounding and stress introduced by handling and transportation, and habituation to the testing environment. Drug testing in familiar rather than novel environments has been shown to affect sensitivity to several drugs (Carey, DePalma, and Damianopoulos 2005; Dunne, O'Halloran, and Kelly 2007; Harkin et al. 2000; Joyce and Mrosovsky 1964). Furthermore, in home cage experiments baseline behaviour is collected and can be incorporated into the analysis. This is advantageous because animals can serve as their own control. Another advantage of baseline data is its potential for obtaining animal-specific rather than *ad hoc* cut-off values. These cut-off values are necessary to determine *e.g.* whether a movement bout is long or short. Basing the cut-off values on the statistical properties of the data per animal helps to increase replicability of animal experiments (Benjamini et al. 2010; Lipkind 2004).

In the PhenoTyper system, rats or mice are tracked continuously from atop. The resulting raw data is an overview of the exact location of the animal on the cage surface over time. These coordinates can be used to calculate not only the distance moved per unit of time, but also a large set of response variables (RVs) such as number of stops, total duration of lingering bouts, and mean velocity while progressing. Describing activity using multiple variables rather than merely the distance an animal travelled enhances the discriminability and the variability between treatments (Benjamini et al. 2010; Spruijt et al. 2014). Recent technological advances have opened up the possibility to simultaneously collect behavioural

data and data from other streams such as physiological parameters (*e.g.* Aziriova et al. 2016) and ultrasonic vocalisations (*e.g.* Peters et al. 2017) in the home cage. Combining these parameters with the set of behavioural RVs will enrich the data collected from automated home cages even further.

Data from automated home cage experiments are typically analysed using univariate statistical models such as generalized linear models and mixed models. Analysing all RVs using separate models results in many outputs which 1) is impractical; 2) results in a loss of statistical power due to the necessary corrections for multiple testing; and 3) does not take into account the correlations between the RVs (*e.g.* the distance moved is strongly negatively correlated to the total duration of stops). Researchers thus often opt to analyse (or present data on) only one or a small subset of RVs, most typically on the distance moved. Analysing a subset of the data solves the before-mentioned issues only partially and does not utilize the full potential of having a rich description of activity available.

In this paper, we propose to analyse larger sets of RVs simultaneously using multivariate statistics. Multivariate techniques, as opposed to univariate techniques, allow for integrated analysis of multiple RVs in a single model. The multivariate method Principal Component Analysis (PCA) has been applied to automated home cage data before, but merely for data reduction purposes (*e.g.* Loos et al. 2014; Visser et al. 2006) and not for statistical inference. In this paper we will describe and apply two methods derived from PCA that *do* allow for hypothesis testing: Redundancy Analysis (RDA) and Principal Response Curves analysis (PRC) (Brink and Braak 1999). Each of the methods is applied in a case study and contrasted to a univariate linear mixed model.

2.2. Case Study 1: Locomotor effects of chemogenetic activation of dopamine neurons in rats, with multiple control groups

2.2.1. Experimental Design

The data of the first case study was obtained from a trial on the locomotor effects of chemogenetic activation of midbrain dopamine neurons in rats (Boekhoudt et al. 2016). We present here a simplified version. In this study, Designer Receptors Exclusively Activated by Designer Drugs (DREADDs, Rogan and Roth 2011) were expressed on dopaminergic neurons in the midbrain, making these cells sensitive to designer drug Clozapine-N-Oxide (CNO). TH:Cre transgenic rats were injected with a Cre-dependent DREADD virus, so that Cre-positive rats expressed the hM3D(Gq) designer receptor, whilst their Cre-negative littermates did not. In total, eight TH:Cre positive rats, expressing hM3D(Gq) (further: Gq^+), and seven TH:Cre negative rats, without hM3D(Gq) (further: Gq^-), were included in the study.

Administration of CNO selectively activates hM3D(Gq), and thereby temporarily activates dopaminergic neurons in the Gq^+ rats, but not Gq^- rats. Thus, an effect on locomotor behaviour is only expected in the Gq^+-group after treatment with CNO. We expect no effect of CNO in the Gq^--group, and we expect that the Gq^+-group and Gq^--group differ only in the presence of CNO.

At least four weeks after the injection with the DREADD virus all rats were treated three times with CNO (0.3 mg/kg, intraperitoneally) and three times with a Control vehicle (saline solution). Rats were individually housed in a PhenoTyper 9000 (Noldus Information Technology, Wageningen, The Netherlands). Locomotor activity was recorded at 25 samples per second, and was analysed with EthoVision XT9 and XT11 (Noldus Information Technology, Wageningen, the Netherlands). Movement tracks of the animals' centre point were smoothed by locally weighted scatterplot smoothing. For this trial, we analysed data on fourteen activity RVs (Table 2.1) in two time-intervals: one hour before treatment and between 20 minutes after and 1 hour and 20 minutes after treatment. The 20 minute time-interval directly after treatment was discarded because 1) it showed increased activity, most likely caused by the disturbance from animal handling and injection; and 2) CNO-induced behavioural effects start following approximately 20 minutes after injection.

The data was analysed separately for the Gq^+-group and the Gq^--group using R version 3.2 (R Core Team 2016). Each RV was log-transformed and normalized by subtracting the mean and dividing by the standard deviation.

2.2.2. Univariate Analysis

2.2.2.1. Statistical Analysis

All 14 RVs were individually analysed using a linear mixed model (Bates et al. 2015) with the fixed effects:

Timing: Binary variable, indicating if the observation was before treatment (BT) or after treatment (AT)

Interaction of Timing x Drug; where Drug is a categorical variable with 2 levels: CNO or Control

Note that the predicted values from this model are equivalent to a model without the main-effect for Timing. We defined four treatment groups for which we compute least square means estimates: CNO_{BT}, CNO_{AT}, $Control_{BT}$, and $Control_{AT}$.

The random part of the model contained a random intercept per Animal. The random intercept allowed the baseline value for the RV to vary between the animals. Inspection of the residuals indicated that the strength of the response to CNO differed between rats. Therefore the response to CNO was modelled using an additional random effect of CNO per animal. This means that we assumed that

Table 2.1: Overview of the response variables used in case studies 1 and 2

Abbreviation	Parameter	Case Study 1
DM / Dist. Mov	Distance Moved	1,2
DML	Distance Moved Lingering	1
DMP	Distance Moved Progressing	1
S mov. dist.	Distance Moved Short Movements	2
L mov. dist.	Distance Moved Long Movements	2
# activity bouts	Frequency Movements	2
FL	Frequency Lingering	1
FP	Frequency Progressing	1
S mov. #	Frequency Short Movements	2
L mov. #	Frequency Long Movements	2
# s arrests	Frequency Short Stops	2
# L arrests	Frequency Long Stops	2
FSW	Frequency Switching	1
# s SV	Frequency Short Shelter Visits	2
MVL	Mean Velocity Lingering	1
MVP	Mean Velocity Progressing	1
DSS / Dur. s arrests	Duration Short Stops	1,2
DLS / Dur. l arrests	Duration Long Stops	1,2
Dur.activity	Duration Movements	2
DSM	Duration Short Movements	1
DLM	Duration Long Movements	1
DL	Duration Lingering	1
DP	Duration Progressing	1
Dur. s SV	Duration Short Shelter Visits	2
Dur. Spout z	Duration Spout Zone	2
Dur. feeding z	Duration Feeding Zone	2
Dur. OnShelter z	Duration On Shelter	2

the differences between CNO_{AT} and the other treatment groups per rat form a normal distribution and that these differences for the rats in our experiment are thus randomly drawn samples from this normal distribution. Statistical significance of the Timing x Drug interaction was determined by comparing the full model with interaction to an identical reduced model without the interaction using an approximate F-test based on the Kenward-Roger approach (Halekoh and Højsgaard 2014; Kuznetsova, Bruun Brockhoff, and Haubo Bojesen Christensen 2016). If the Timing x Drug term was significant ($\alpha = 0.05$), the Differences of Least Squares Means between BT and AT (LSdiff) were calculated for CNO and the Control separately (Lenth 2016). If the Timing x Drug term was not significant, it was removed from the full model and statistical significance of the Timing term was determined using the approximate F-test as described above. If the Timing term was significant ($\alpha = 0.05$), the LSdiff was calculated for CNO and Control combined as described above.

2.2.2.2. Results

Distance Moved

Treatment with CNO in the Gq^+-group, resulted in a significantly higher increase in the Distance Moved than the Control treatment (Difference of Least Squares Means: LSdiff). On average over all the subjects, the increase in Distance Moved was 4.7 times higher for CNO than for the Control (Table 2.2). In the Gq^--group, the there was no difference in the effect of the CNO and the Control treatment on the Distance Moved.

All response variables

In the Gq^+-group, seven out of thirteen RVs showed a significant effect similar to that of Distance Moved: the difference in expected values before and after treatment were larger after treatment with CNO than after the Control treatment (Table 2.2). One RV (DLS) showed a reverse effect, the expected values decreased rather than increased after treatment. For these eight RVs, the effect sizes were considerably larger for CNO (all absolute LSdiffs> 1.898) than for the Control (all absolute LSdiffs <0.404). In five out of thirteen RVs we found no significant Timing x Drug interaction, no significant Treatment term, and thus no difference in expected values before and after treatment with CNO or the Control treatment.

In the Gq^--group, no significant Timing x Drug interactions were found, which confirmed that there were no significant differences in locomotor activity following CNO treatment compared to Control (Table 2.2). The Timing term in the reduced model was significant for six RVs indicating a significant difference between BT and AT that did not differ between CNO and the Control. If the RVs are consistent, we would expect that if an RV shows an effect in the Gq^--group we

Table 2.2: Overview of relevant differences between least squares means estimates before and after treatment for all response variables per group. If in the full model the interaction term Timing x Drug was significant the LSdiffs are given ($\pm sd$) for CNO and the Control treatment separately, else the model was reduced. If the Timing term in the reduced model was significant the LSdiffs are given for CNO and the Control treatment combined else there was no significant effect and no LSdiffs are given. Significant LSdiffs are indicated by a star

			CNO Before - After	Control Before - After	Both Treatments Before - After
DL	GQ^-	Timing	.	.	-0.436*±0.190
	GQ^+	No effect	.	.	.
DLM	GQ^-	Timing	.	.	-0.495*±0.157
	GQ^+	Timing * Drug	-1.898*±0.155	-0.404*±0.135	.
DLS	GQ^-	Timing	.	.	0.435*±0.185
	GQ^+	Timing * Drug	2.050*±0.205	0.234 ±0.100	.
DM	GQ^-	Timing	.	.	-0.322*±0.150
	GQ^+	Timing * Drug	-1.912*±0.237	-0.244 ±0.112	.
DML	GQ^-	Timing	.	.	-0.391*±0.171
	GQ^+	No effect	.	.	.
DMP	GQ^-	No effect	.	.	.
	GQ^+	Timing * Drug	-1.904*±0.242	-0.248 ±0.114	.
DP	GQ^-	Timing	.	.	-0.430*±0.151
	GQ^+	Timing * Drug	-1.966*±0.183	-0.377*±0.126	.
DSM	GQ^-	No effect	.	.	.
	GQ^+	No effect	.	.	.
DSS	GQ^-	No effect	.	.	.
	GQ^+	Timing * Drug	-0.483*±0.127	-0.201 ±0.105	.
FL	GQ^-	No effect	.	.	.
	GQ^+	No effect	.	.	.
FP	GQ^-	No effect	.	.	.
	GQ^+	Timing * Drug	-1.794*±0.174	-0.230 ±0.166	.
FSW	GQ^-	No effect	.	.	.
	GQ^+	No effect	.	.	.
MVL	GQ^-	No effect	.	.	.
	GQ^+	Timing * Drug	-1.079*±0.263	-0.194 ±0.135	.
MVP	GQ^-	Timing	.	.	1.052*±0.319
	GQ^+	Timing * Drug	-1.535*±0.402	0.098 ±0.103	.

would also find it in the control treatment of the Gq^+-group. We could however, only replicate this effect in four out of six RVs.

2.2.3. Multivariate Analysis

2.2.3.1. Brief introduction to RDA

RDA is a constrained form of Principal Components Analysis (PCA). Understanding PCA helps to understand RDA. Therefore we first provide an introduction to PCA and then introduce RDA. A more elaborate and intuitive interpretation of PCA is given in S.2.A.

PCA, as the name implies, is used to find the principal components: the underlying structure of a data set. The principal components (PC) are linear combinations of the RV, more specifically: the eigenvectors of the data set. All PC are uncorrelated and the total number of PC is equal to the number of RV. The higher the eigenvalue of a PC, the more variance in the data set it explains. The first principal component (PC1) of a data set is the eigenvector with the highest eigenvalue. The second and subsequent PCs are the linear combinations that, after removing the effect of (the) previous PC(s), describe the maximum amount of variance in the data set.

RDA is similar to PCA. The difference is that PCA considers all variance in the data set and RDA considers only the variance that can be explained by the explanatory variables (such as experimental treatment). In RDA, we do not apply PCA directly to the data set but to the the data described in terms of the explanatory variables. First for each RV a linear model is fitted that regresses the RV on the explanatory variables, then the fitted values from these models are combined in a new data set. PCA is then applied to this data set of fitted values.

In this case study we use partial RDA. In partial RDA (as in partial PCA), the data is conditioned on (*i.e.* corrected for) one of the explanatory variables. The variance that originates from that covariate is removed from the data set before conducting the actual RDA (or PCA). Conditioning on a covariate in essence means setting the average of all observations with the same covariate level to zero.

The main types of output of RDA are the variances explained by its components and (the plot of) the sets of scores of RV, observations, and explanatory variables. RDA also partitions the total variance of a data set in conditioned variance (if present), constrained variance, and unconstrained (or residual) variance. In general, both the variances itself and the proportional variances are given. The statistical significance of an RDA model is usually determined by a permutation test. In a significant model, the constrained variance (the variance explained by the explanatory variables) is large relative to the unconstrained variance (the residual variance). The exact interpretation of the scores of RV, observation, and explanatory variables depends on the scaling used. The interpretation of the vari-

ance partitioning and the different scores is discussed in the results section of the case study (Section 2.2.3.3).

2.2.3.2. Statistical Analysis

Partial RDA (rda-function, Vegan-package, Oksanen et al. 2017) was performed on the data sets for the Gq^+-group and Gq^--group separately using Animal ID as a conditioning variable. The procedure thus corrects for the difference between animals prior to performing the actual RDA. The RDA-procedure answers a different research question than the univariate approach. Recall that in the linear mixed model we tested, for each of the RVs, if the expected value for CNO_{AT} differed from CNO_{BT}. In the RDA-procedure we tested if the constraining variable explained a significant proportion of the variation in the data. We defined the constraining variable as one factor with four levels: CNO_{BT}, CNO_{AT}, $Control_{BT}$, and $Control_{AT}$. If the RDA model with a four level factor was significant, we subsequently tested whether or not the four level factor (full model) could be reduced to a binary factor with two levels: CNO_{AT} and Not CNO_{AT}. Significance of the RDA-models as a whole, the axis, and the model terms was determined using restricted permutation tests. We permuted within animals keeping data collected on the same day together.

2.2.3.3. Results

The main types of output of RDA are the distribution of variance and the sets of scores of RV, observation, and explanatory variables. The distribution of variance in the RDA-model with the four treatment groups as constraining factor differed clearly between the Gq^+-group and the Gq^--group (Table 2.3).

In the Gq^+-group, the conditional variance was 18% of the total variance which means that 18% of the variation in the data set could be attributed to differences between animals. The constrained variance was 40% of total variance which means that 40% of variance in the data could be attributed to differences between

Table 2.3: Distribution of variance (absolute and proportional to the total) within the RDA-analysis for the Gq^+ group and Gq^- group.

	GQ^+		GQ^-	
Type of variance	Proportion explained	Rank	Proportion explained	Rank
Conditional	0.1813	7	0.5360	6
Constrained	0.4041	3	0.0315	3
Unconstrained	0.4146	11	0.4325	12
Total	1.0000		1.0000	

the four treatment groups. The unconstrained variance was 41% of the total variance, which means that 41% of the variation in the data set could not be explained by either the differences between animals or the differences between treatments. In the Gq^--group compared to the Gq^+-group, the percentage of explained conditional variance was much higher (53% of total), the percentage of explained constrained variance was much lower (3.1%), and the unconstrained variance was similar (43%). Note that the total variation in the Gq^- and Gq^+ data sets is equal because both sets have the same number of RVs, all of which are standardized to have standard deviation of 1. The large percentage of explained conditional variance in the Gq^--group thus does not imply that the animals in this group differed more from each-other than in the Gq^+-group.

The results of the permutation tests confirm the hypotheses. The RDA-model for the Gq^--group was not significant as a whole ($p<0.556$), which indicates that there were no differences between the four treatment groups. The RDA-model for the Gq^+-group using the four level factor was significant as a whole ($p<0.001$). The subsequent sequential permutation test confirmed that the four level factor can be simplified to a binary factor with the levels CNO_{AT} or Not CNO_{AT}. This indicates that the CNO treatment had an effect in the Gq^+-group and that, in this group, the rats' behaviour was similar before the CNO treatment, before the Control treatment and after the Control treatment (Figure 2.1. We plot the results of the full model using RDA1 on the x-axis (displaying fitted values) and the first PCA-axis of the residuals on the y-axis (Figure 2.1). We do not show RDA2 here because it explains very little variation (*i.e.* the model can be reduced to one binary factor resulting in a one dimensional RDA with one RDA-axis).

In Figure 2.1 and Figure S.2.B.1 the observations are represented by circles, coloured by Timing x Drug interaction. Envelopes are drawn around the outside observations of each group. Qualitative variables are plotted here as triangles (centroids; one for each level), where the location of a triangle indicates the mean position of all observations with the same variable level. The interpretation of plots of RDA-scores depends on the type of scaling that is used. The exact scaling between implementations of RDA, but the interpretation is comparable. After Type 1 scaling (or object focused or site-scaling), the scores for observations and qualitative variables are such that when plotted the distances between them represent their similarity (Euclidean distances). The interpretation of all points, observations and centroids, is equal. The closer the points are together, the more similar we expect their set of RVs to be. The RDA-plot of the Gq^+-group (Figure 2.1) shows, in line with the permutation tests, that observations from all groups except for CNO_{AT} are mixed and that the observations from group CNO_{AT} are clearly apart from the other groups. The RDA-plot of the Gq^--group (Figure S.2.B.1) shows that observations from all groups are mixed and that the cen-

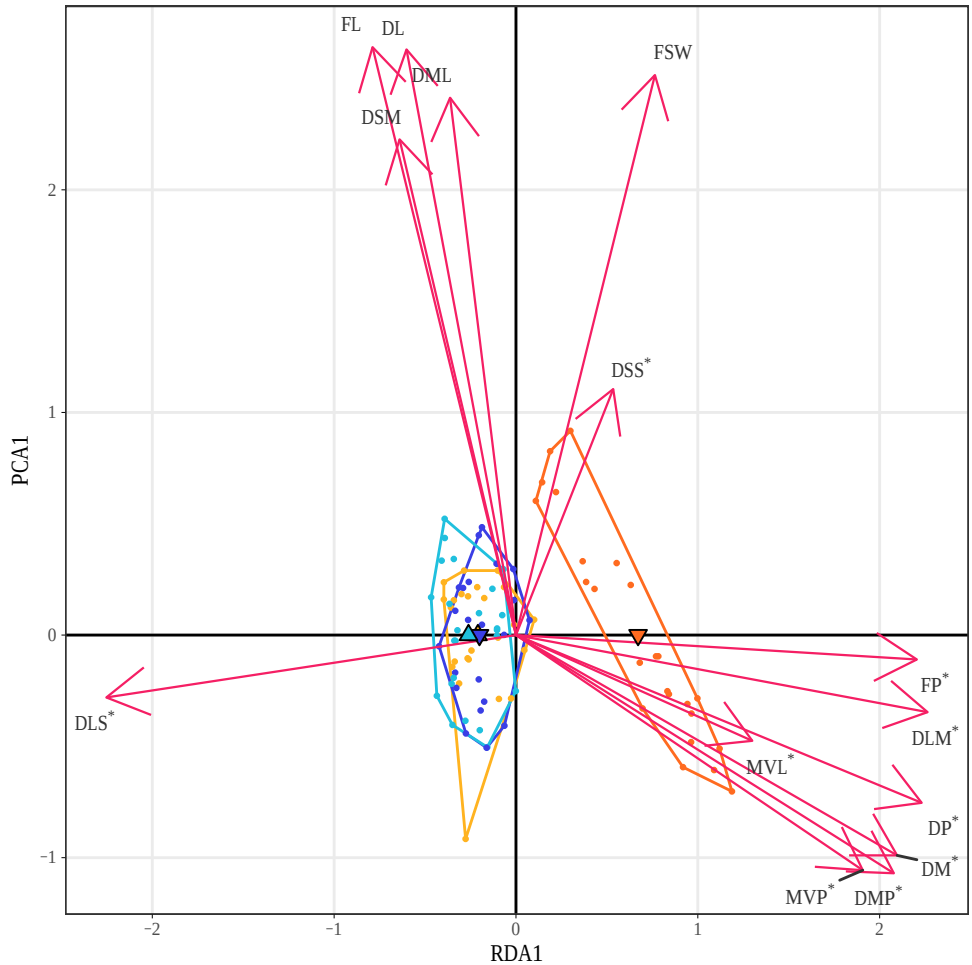


Figure 2.1: Results of the RDA-analysis on the Gq^+-group using scaling = 1. Scores of the first RDA-axis are plotted on the x-axis and scores of the first PCA-axis are plotted on the y-axis. Individual observations are represented by points that are coloured by Drug (CNO: orange, Control: blue) and Timing (Before Treatment: light hues, After Treatment: dark hues). All individual observations in the same treatment group are enveloped. Centroids for the treatment groups are represented by triangles (Before Treatment: upward, After Treatment: downward; coloured as the observations). RVs are represented by named arrows. Names of RVs that had a significant Timing x Drug interaction in the univariate model are indicated by a star

troids of all groups are close together, and thus, as concluded from the permutation tests, that the differences between observations are not explained by their group. As none of the RDA-axis for the Gq^--group were significant we will only discuss the plot of the Gq^+-group here.

In RDA-plots, RVs are usually represented as arrows. The points (centroids and observations) relate to the RVs via right angled projections (RAPs). The RAP of a point on an RV is the position on the RV-arrow from which the distance to the point is the shortest, the connection from the RAP to the point is perpendicular to the RV-arrow. The RAP of an observation on an RV represents its expected value.

In the Gq^+-group, observations from the CNO_{AT}-group are in the positive direction and observations Not CNO_{AT}-group are in the negative direction of the RDA1 axis. This means that observations in the CNO_{AT}-group, compared to the average of all observations, have higher expected values for RV with a positive and lower expected values for RV with a negative loading for RDA1.

The RDA1-axis is targeted towards separation of the CNO_{AT}-group from the Not CNO_{AT}-group. RVs are useful for this separation if the RAPs of observations within these two groups are in distinct groups. This indicates that there is large variation between groups, and not within groups (*e.g.* FP and DLS). RVs are less useful for distinguishing between groups if they vary less between groups and more within groups (*e.g.* FSW and DL). Note that in Figure 2.1 the y-axis represents PC1, the first component of the PCA on the residuals. This axis is targeted towards displaying the maximum amount of variation in the data set that remains after the effect of the first RDA-axis is removed *i.e.* variation that cannot be explained by difference between groups.

The other type of RDA-scaling is Type 2 scaling (or RV focussed or species species-scaling), where the main interest is in the RV-arrows and less on the observations. The main advantage of Type 2 scaling is that angles between all arrows (both RV and quantitative variables) represent their correlation. The interpretation of the relationships between points and RV arrows is equal to that of Type 1 scaling; the RAP of a point on an RV arrow represents the value of that RV for that observation or centroid. However, in contrast to Type 1 scaling, we cannot interpret the distances between points (observations and centroids) as Euclidean distances. In this case study the main interest is in the differences between the Treatment groups and there are no quantitative variables. Therefore Type 1 scaling is more appropriate.

2.2.4. Conclusion of univariate and multivariate analyses

The main conclusion from the univariate and multivariate analysis is the same: 1) in the Gq^+-group, activation of the designer receptors by the designer drug CNO caused hyperactivity, and 2) in the Gq^--group, CNO did not cause behavioural ef-

fects compared to the Control treatment. From the mixed model approach we can conclude that in the Gq^+-group, the increase in DM After Treatment was larger for CNO than for the Control and that in the Gq^--group there was no difference between the treatments. From the RDA approach we can conclude that the treatment explained a significant part of the variation of the data of the Gq^+-group and not of the Gq^--group.

The RVs that were most important in the RDA (the highest absolute scores) were the same RV with significant Timing x Treatment interactions in the mixed model. Distance Moved seems a reasonable RV to quantify the hyperactivity effect as it scored highly on the RDA-axis, although it also scores highly on the PCA-axis which implies it has variability unexplained by the difference between treatment groups. FP or DLM have comparably high RDA-scores and lower PCA-scores which implies smaller variation within treatment groups than for DM. The RVs that seem least relevant in the RDA-plot were the same RV for which the mixed model did not find a significant interaction effect. These RVs have high scores on the PCA-axis and lower absolute scores on the RDA-axis which implies that these RVs are variable independent from the experimental treatment.

An advantage of the linear mixed model compared to the RDA-analysis is that linear mixed models can, in contrast to RDA, provide effect-size estimates with standard errors per RV between CNO_{AT} and CNO_{BT}. RDA-analysis can provide a standardized effect size without a standard error, but only if the constraining variable is binary.

An advantage of the RDA-analysis compared to the linear mixed model is that it provides results of all RVs simultaneously in one plot. This plot, when using the appropriate scaling, also indicates correlation between RVs. RDA can applied to as many RVs as are expected to be affected by a treatment without major consequences for the complexity of the results and without post-hoc adjustments for multiple testing. RDA is therefore useful in situations when the overall effect of an experimental intervention or the overall difference between groups is of importance and the exact size of the effects per RV are not.

2.3. Case Study 2: Direction and strength of effect of a H3 histamine receptor inverse agonist and Diazepam on spontaneous activity

2.3.1. Experimental Design

The data of the second case study was obtained by merging data from two unpublished trials with a different drug treatment in an otherwise identical experimental protocol. The drug treatments were a H3 histamine receptor inverse agonist (GSK 189254, synthesized by Griffin Discoveries, Amsterdam, The Netherlands) in

the first trial (further: H3 trial) and the anxiolytic drug Diazepam in the second trial (further: Diazepam trial). Via a voluntary oral administration procedure (as described in Aarts et al. 2015), each mouse was given each of four doses (a Control, and low, medium, and high doses of the drug) in different orders before the onset of the dark phase (Personal communications: Maarten Loos, Sylics (Synaptologics BV), Amsterdam, Netherlands).

Mice were individually housed in a PhenoTyper 3000 (Noldus Information Technology, Wageningen, The Netherlands). Food and water were provided ad libitum. Locomotor activity was recorded at 15 samples per second, and was analysed with EthoVision® software (EthoVision HTP 2.1.2.0, based on EthoVision XT 4.1, Noldus Information Technology, Wageningen, The Netherlands) and processed to generate behavioural parameters using AHCODA™ analysis software (Synaptologics BV, Amsterdam, The Netherlands). Movement tracks of the animals' centre point were smoothed by locally weighted scatterplot smoothing. For this trial, we analysed data on Distance Moved and fifteen other activity RVs (Table 2.1) in twelve one-hour intervals.

Both trials were executed as a balanced trial but due to technical issues the data sets are not balanced. The H3 trial has records on 23 mice of which 12 with complete records (eight mice with 1 dose missing; two with 2 doses missing; one with 3 doses missing). The Diazepam trial has records on 24 mice of which 12 with complete records (nine mice with 1 dose missing; two with 2 doses missing: 2; one with 3 doses missing).

2.3.2. Univariate Analysis

2.3.2.1. Statistical Analysis

All 16 RVs were analysed using a linear mixed model with the fixed effects:

Time: Categorical variable with 12 levels, indicating the hour;

Trial: Categorical variable with 2 levels, indicating whether observations were from the H3 or Diazepam trial;

Interaction of Time x Treatment: where Treatment is a categorical variable with 7 levels: Saline (combined for H3 and Diazepam trial), three doses of H3, and three doses of Diazepam.

We defined a random intercept model with factor Animal. This random intercept allowed the baseline value for the RV to vary between animals.

Statistical significance of the interaction of Time x Treatment interaction were determined by comparing the full model with interaction to an identical reduced model without the interaction using an approximate F-test based on the Kenward-Roger approach (KRmodcomp-function, pbkrtest-package, R 3.3 Halekoh and Højsgaard 2014). If the Time x Treatment term was significant ($\alpha = 0.05$), the Least

Squares Means were calculated for each combination of Time x Treatment. LSdiffs were calculated within each level of Time, for each Treatment versus the Control with (approximate) Dunnet adjustment for multiple (six) comparisons (lsmeans-function, lsmeans-package, R 3.3 Lenth 2016).

2.3.2.2. Results

Distance Moved

Distance Moved was reduced after treatment with H3 and Diazepam compared to the Control. As the Time x Interaction was significant ($F(72, 1757.60)$, $p < 0.001$), the magnitude of the effect varies between combinations of Time and Treatment. For Diazepam, significant decreases in Distance Moved were found for the medium and high dose at hour 4 and 5 of the dark phase (Figure 2.2). For H3, significant decreases in Distance Moved after treatment were found for all doses between hour 6 and 8 of the dark phase and for the medium and high dose also after hour 10.

All parameters

The Treatment x Time interaction was significant in the univariate linear mixed models of all other RVs. This indicates that for all RVs, the expected value is different between Treatments and that the magnitude of this difference is not the same for the different levels of Time. For each point in time, the LSdiff between each of the Treatments and the Control was calculated. Summed over all RVs, we found more significant LSdiffs for H3 than for Diazepam (55 and 161 respectively; Table 2.4). The overall pattern off all RVs was similar to that of Distance Moved. For Diazepam, the vast majority (47 out of 55) of significant LSdiffs were found at hours 3 to 5 and for H3, all but one significant LSddiffs were found at hours 6 to 8, 11, and 12. For both Diazepam and H3, the number of significant LSdiffs was clearly lower for the lowest dose than for the medium and high dose. Not all RVs were equally sensitive in detecting the effect of treatment with H3 and Diazepam (Table S.2.B.1). The number of significant LSdiffs per RV combined for H3 and Diazepam ranged from 3 (Dur. Spout z) to 18 (L mov. #) (Table S.2.B.1). The difference in sensitivity between RVs appears larger for Diazepam than for H3 (Figure 2.2).

2.3.3. Multivariate Analysis

2.3.3.1. Brief introduction to PRC

PRC is a special case of RDA optimized for describing experimental data that has been collected over time (Brink and Braak 1999). The underlying calculations are equivalent to performing an RDA with the Time x Treatment interaction as constraining factor, and Time as conditioning factor. PRC describes the overall effect of treatments on a collection of RVs simultaneously and the extent to which

Table 2.4: Number of 16 univariate linear mixed models in which the difference in least square means between each of the treatments and the control treatment was significant at that time point. For clarity, zeros are replaced by a period. Shaded cells indicate significant effects found in the PRC-model

		Time												
Drug	Dose	1	2	3	4	5	6	7	8	9	10	11	12	All
Diazepam	0.5	.	.	3	.	.	.	.	.	2	.	.	.	5
	1.0	1	.	2	15	7	.	.	.	.	.	.	.	25
	2.0	1	.	.	11	9	1	.	.	.	.	1	2	25
H3	0.3	.	.	.	.	1	14	.	12	.	.	.	3	30
	1.0	.	.	.	.	.	15	15	16	.	.	11	14	71
	3.0	.	.	.	.	.	14	16	15	.	.	.	15	60

the treatment effect on each of those RVs resembles the overall response. PRC outputs two sets of coefficients that can be plotted in easily interpretable graphs. The first set are the dose-time coefficients (c_{dt}s) which represent the size of the effect of treatment d at time t relative to the control treatment at the same time. For the control treatment, c_{dt}s is thus always zero. The c_{dt}s are typically plotted in a line-plot against time. The second set of coefficients are the RV-weights (b_ks). If b_k of an RV is zero there is no correlation between the RV and the overall response pattern. The further b_k is from zero, the more the response pattern of the RV resembles the overall response pattern (or the negative response pattern if $b_k < 0$).

2.3.3.2. Statistical Analysis

We performed a Principal Response Curves Analysis (PRC) with Animal as extra covariate (adjusted version of prc-procedure, vegan-package, R3.3, available via Electronic Supplement). The implication of using Animal as extra covariate is the same as in RDA, the mean of all observations of an Animal are set to zero (prc-function, vegan-package, Oksanen et al. 2017). Because in PRC we also condition on Time, in effect the mean of all observations of an Animal at the same Time are set to zero. Statistical significance of the PRC-axis and the model as a whole was determined using constrained permutation. We restrict to permutation within Animal (and thus also Trial), and keep observations from the same day together. Balanced data is necessary to do permutation, therefore missing observations were imputed using the average value within Animal for the days that data was available. Only Animals with records on the Control and at least two other doses or the Control and the High dose were included in the imputed data set. These imputed data sets were only used for the significance testing of the whole

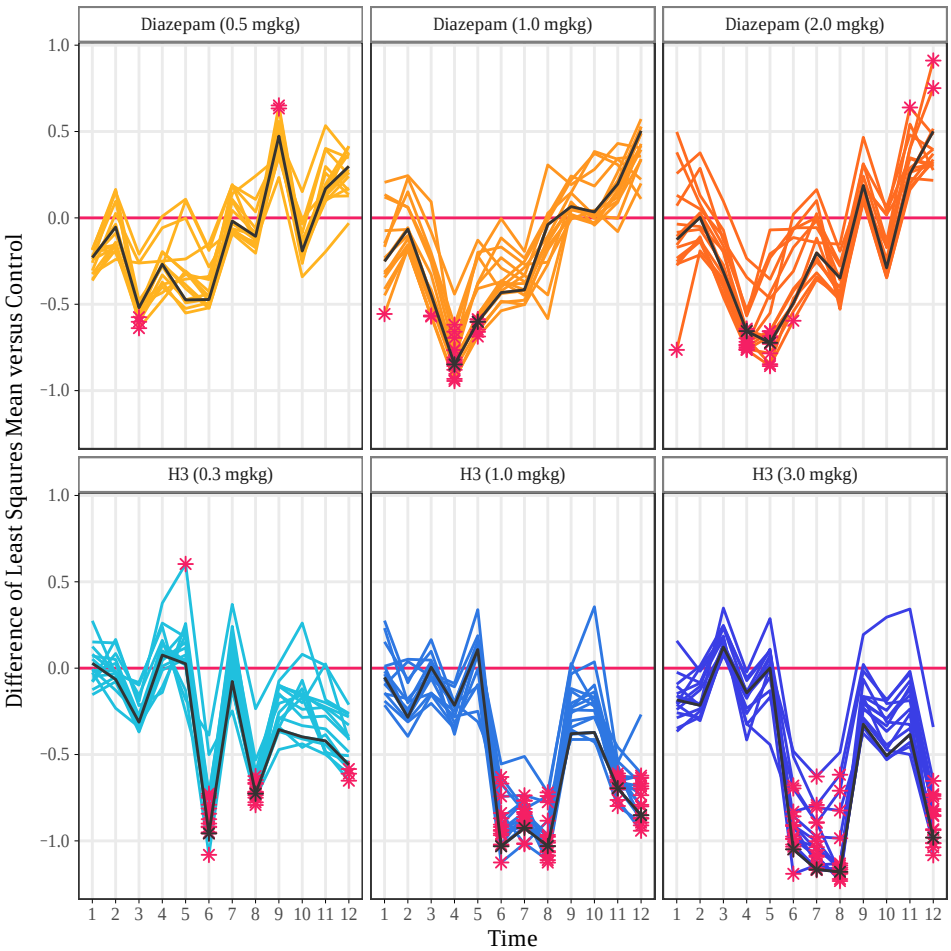


Figure 2.2: Differences of Least Squares Means between the Treatments and the Control are plotted with Distance Moved in grey and the other 15 RVs coloured by Drug (Diazepam: Orange, H3: Blue) and Dose (lowest dose: lightest hue, medium dose: medium hue, highest dose: darkest hue). Stars indicate significant differences ($\alpha = 0.05$; Williams test procedure).

model, the other presented data is from the data set as observed. A RV-selection approach based on permutation testing was used to determine which RVs are statistically relevant for describing the overall response pattern (Vendrig, Hemerik, and Braak 2017).

For each level of Time, we test whether or not the coefficients of the levels of Treatment differ from the control treatment using an adjusted version of the Williams test. The first step is extracting the scores of the first axis of a PCA on the RVs, conditioning on Time and Animal (rda-function, vegan-package Oksanen et al. 2017). Then, for each level of time, we regress the PCA scores on Treatment, and test for differences between each Treatment and the Control with Dunnet adjustment for multiple (six) comparisons (glht-function, multcomp-package Hothorn, Bretz, and Westfall 2008).

2.3.3.3. Results

The first axis of the PRC was significant indicating that there is a difference between the Treatments. The PRC-diagram shows that treatment with Diazepam and H3 resulted in lower c_{dt}-estimates compared to the Control (Figure 2.3). The b_k-scores of all RVs were positive, lower c_{dt}-estimates thus indicate lower expected values for the activity RVs and thus represent a decrease in activity compared to the Control. The Diazepam treatment shows a decrease in activity from hour 2 to hour 9 of the dark phase, after which the activity is similar to or higher than that of the Control. The difference in activity between the H3 treatment and the Control occurs later, it starts after hour 5.

The direction of the effect was the same for all RVs since all b_k-scores were positive. No RVs were excluded based on the RV-selection procedure yet the magnitude of the contribution to the overall treatment effect differed per RV. Distance Moved is one of the RVs with the highest absolute b_k-scores indicating that it strongly contributes to the overall treatment effect and Duration Spout Zone has the lowest absolute b_k-score indicating that it contributes the least of all RVs to the overall treatment effect. The second PRC-axis was not significant, suggesting that there was no secondary response pattern and thus that *e.g.* the nature of the effect of H3 and Diazepam was similar.

2.3.4. Conclusion of univariate and multivariate analyses

The main conclusions from the univariate and multivariate analyses were the same: 1) Diazepam and H3 decrease the activity compared to the Control, 2) the effect of Diazepam occurs earlier after treatment than that of H3, and 3) the effect of Diazepam on activity appears to be weaker than that of H3. The PRC-analysis provides a direct ranking of RVs in terms of correlation to the overal response pattern via the b_k-scores. Those RVs that have high absolute b_k-scores in the PRC

analysis also have a high number or significant LSdiffs in the linear mixed model (Table S.2.B.1). PRC and the univariate analysis of Distance Moved appear to have similar power. In the PRC analysis, a significant effect for the medium and high dose of Diazepam was detected at hour 3 which was not detected in the univariate model for Distance Moved, whereas a significant effect for the lowest dose of H3 at hour 8 was found in the univariate model for Distance Moved but not for the PRC analysis.

The advantage of the PRC approach compared to univariate analysis is that PRC incorporates the complete set of RVs without adding complexity to the analysis and the interpretation of its results. Interpretation of the complete set of univariate models is not straightforward. Here, both PRC and the set of linear mixed models suggest that 1) the response pattern to H3 and Diazepam across all RVs is similar and merely shifted over time and 2) that all RVs show a similar response pattern over time with a varying sensitivity to Treatment effect. The PRC-procedure provides a statistical foundation for this conclusion because the second PRC-axis was not significant whereas the univariate procedure does not allow for statistical inference on the relationship between RVs.

2.4. Discussion

For both case studies, we were able to replicate the main conclusions of the univariate analysis of Distance Moved using the appropriate multivariate method. These methods are hardly more complicated to apply and interpret compared to univariate methods and are available in many software packages (*e.g.* Canoco (Smilauer and Lepš 2014), R (Oksanen et al. 2017; R Core Team 2016), and Microsoft Excel (Addinsoft 2007)). Both RDA (*e.g.* Mayor et al. 2017; Ruff et al. 2015; Storkey et al. 2015) and PRC (*e.g.* Ferrenberg, Reed, and Belnap 2015; Fuentes et al. 2014; Guo et al. 2014) have been widely applied in fields such as ecology, toxicology and microbiology. In these fields, data that, like data from automated home cage experiments, have many correlated RVs and relatively few experimental units are common.

Using multivariate methods as opposed to univariate methods, data of multiple RVs can be assessed simultaneously which eliminates the need for *a priori* RV-selection. RDA and PRC provide a graphical representation of the data on all the RVs in one (or two in case of PRC) easily interpretable plots and provides statistical inference on the complete set of RVs. Analysing the complete set of RV using separate univariate analyses results in a set of results that can only be interpreted separately. The results can be summarized per RV, but not for the set of RV together.

Results of RDA and PRC are not overly more complicated to interpret than results of a linear mixed model. The exact interpretation of results of an RDA-

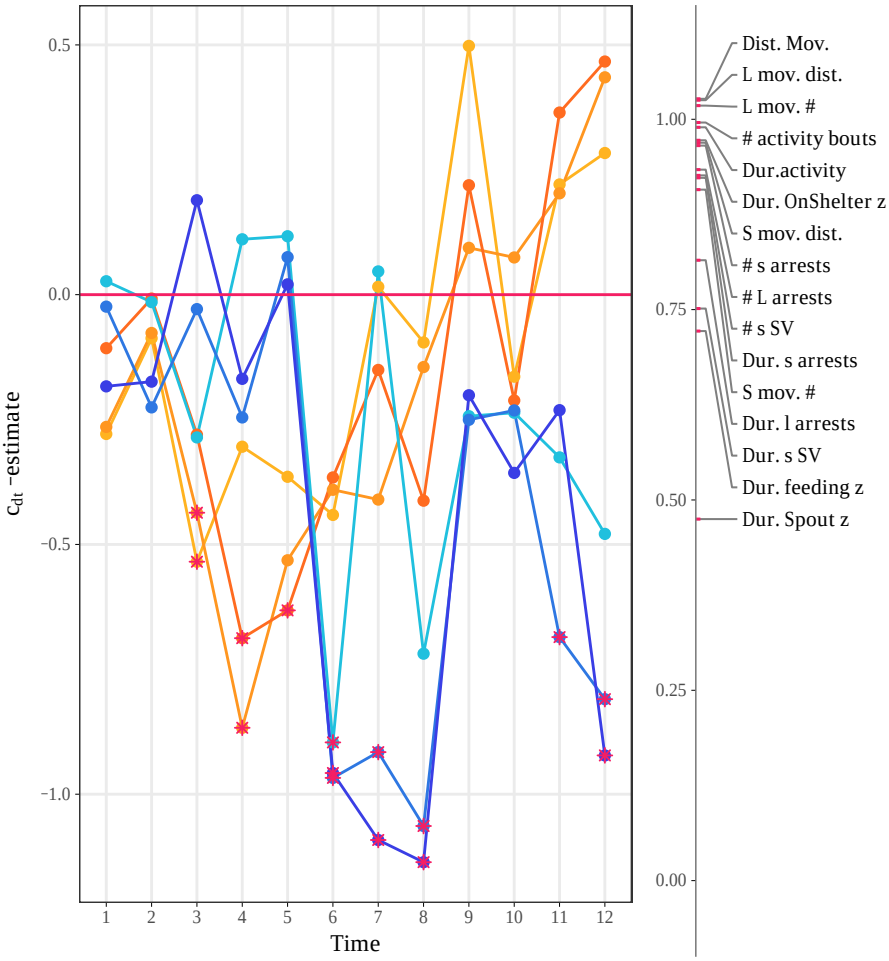


Figure 2.3: Results of PRC analysis using Animal as an extra covariate. The Principal Response Curves are given on the left and b_k-scores are given on the right side of the plot. Principal Response Curves are coloured by Drug (Diazepam: Orange, H3: Blue, Control: Pink) and Dose (lowest dose: lightest hue, medium dose: medium hue, highest dose: darkest hue). Pink stars indicate significant difference ($\alpha = 0.05$; Williams test procedure) between the treatment and the control at that time point.

analysis requires some instruction but the crude interpretation of results of an RDA-analysis and a plot of RDA-scores are fairly intuitive: 1) an explanatory variable is important if it explains a lot of variance compared to the residual variance; and 2) points that are close together in the RDA-plot are more similar than points that are far apart. The crude interpretation of PRC-plots is even more intuitive than the interpretation of RDA-plots. The line-plot indicates the overall response of all RV to the treatment and that the relative importance of the RV is indicated by the b_k-scores. The interpretation is aided because the principal response curves visually resemble a straightforward plot of the LSdiffs of one RV over time relative to the control group. Comparing the principal response curves in Figure 2.3 to the somewhat equivalent plotting of the LSdiffs of all individual RVs in Figure 2.2; Figure 2.2 is larger and more cluttered without being more informative than Figure 2.3.

In the first case study, the univariate analysis showed that not all RVs were affected by the treatment. In the second case study, the univariate analysis showed that all RVs were affected by the treatment. The univariate analyses however do not provide information on the relations between the RVs.

The results of the univariate analyses were confirmed and clearly visualized in the multivariate analyses. The RDA-plot of the first case study visualizes which RVs were important in describing the difference between treatment groups and which RVs varied within animals regardless of treatment. The RDA-plot also depicts the individual observations coloured by treatment and thus the direction of variance between and within treatment groups. The PCA-plot of the second case study depicts the principal response curves which visualize the difference in behaviour between the control and the other Treatments over time and relative resemblance of individual RVs to this response pattern. Multivariate methods do not assign a p-value to individual RVs. If so required however, a permutation testing approach has been developed to identify RVs that do not contribute significantly to the overall response pattern (Vendrig, Hemerik, and Braak 2017).

The set of RVs in these case studies were all highly correlated activity parameters. We were not able to identify a secondary response pattern. An example of a hypothetical secondary response patterns in the first case study is a behavioural response to the control treatment, *e.g.* an effect in some of the RVs due to stress after handling. In our case study, such an effect was not detected as indicated by the fact that we could simplify the model with four treatment groups to a model with two treatment groups. An example of a secondary response pattern in the second case study is that the effect of the treatment differs in nature over time, such as when animals first stay active but move slower (increased duration lingering, no change in duration stopping) and subsequently fall asleep (decreased duration lingering, increased duration stopping). If such effects are present, mul-

tivariate methods show them instantly whereas univariate methods will not. Such secondary response patterns have regularly been detected using PRC in ecology, for instance to show that the difference in abundances of macro-invertebrates (the RVs) over time between two different sites exist of two distinct patterns that occurred simultaneously with different dominant RVs (Brink et al. 2009).

Multivariate methods have been applied before to data from automated home-cage experiments but merely for data description (*e.g.* Dam et al. 2013) and reduction purposes (Loos et al. 2014; Visser et al. 2006). This paper proposes multivariate methods that can also be used for statistical inference. We have shown that these methods replicate the results of univariate analyses and in addition have the potential to discover secondary response patterns. The advances in the field will continue to provide more and more complicated data which only increase the need for more advanced data analysis methods (Spruijt and Visser 2006). Results of such methods however should allow for ethological interpretations (Spruijt et al. 2014). We believe that the multivariate approaches presented here provide a valuable addition to the statistical toolbox of neuroscientists as they can translate data sets with many RVs into integrated and easily interpretable outputs.

Acknowledgements

The Neuro-BSIK Mouse Phenomics consortium: A.B. Brussaard[a], J.G. Borst[b], Y. Elgersma[b], N. Galjart[c], G.T. van der Horst[c], C.N. Levelt[d], C.M. Pennartz[e], A.B. Smit[f], B.M. Spruijt[g], M. Verhage[h] and C.I. de Zeeuw[b], and the companies Noldus Information Technology (www.noldus.com) and Sylics (Synaptologics BV; www.sylics.com).
[a] Department of Integrative Neurophysiology, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University Amsterdam, Amsterdam, The Netherlands. [b] Department of Neuroscience, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [c] Department of Cell Biology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [d] Netherlands Institute for Neuroscience, Amsterdam, The Netherlands. [e] Swammerdam Institute for Life Sciences–Center for Neuroscience, University of Amsterdam, Amsterdam, the Netherlands. [f] Department of Molecular and Cellular Neurobiology, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University Amsterdam, Amsterdam, The Netherlands. [g] Department of Biology, University of Utrecht, Utrecht, The Netherlands. [h] Department of Functional Genomics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University Amsterdam, Amsterdam, The Netherlands.

References

Aarts, Emmeke, Gregoire Maroteaux, Maarten Loos, Bastijn Koopmans, Jovana Kovačević, August B. Smit, Matthijs Verhage, and Sophie van der Sluis. 2015. "The light spot test: Measuring anxiety in mice in an automated home-cage environment." 294 (November): 123–130.

Addinsoft. 2007. *XLSTAT 2007.* Paris, France.

Aziriova, S., K. Repova, K. Krajcirovicova, T. Baka, S. Zorad, V. Mojto, P. Slavkovsky, et al. 2016. "Effect of ivabradine, captopril and melatonin on the behaviour of rats in L-nitro-arginine methyl ester-induced hypertension." *Journal of Physiology and Pharmacology* 67 (6): 895–902.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using {lme4}." *Journal of Statistical Software* 67 (1): 1–48.

Benjamini, Yoav, Dina Lipkind, Guy Horev, Ehud Fonio, Neri Kafkafi, and Ilan Golani. 2010. "Ten ways to improve the quality of descriptions of whole-animal movement." 34, no. 8 (July): 1351–1365.

Boekhoudt, Linde, Azar Omrani, Mieneke C M Luijendijk, Inge G. Wolterink-Donselaar, Ellen C. Wijbrans, Geoffrey van der Plasse, and Roger A H Adan. 2016. "Chemogenetic activation of dopamine neurons in the ventral tegmental area, but not substantia nigra, induces hyperactivity in rats." *European Neuropsychopharmacology* 26 (11): 1784–1793.

Brink, Paul J. van den, Piet J. den Besten, Abraham bij de Vaate, and Cajo J F ter Braak. 2009. "Principal response curves technique for the analysis of multivariate biomonitoring time series." *Environmental Monitoring and Assessment* 152, nos. 1-4 (May): 271–281.

Brink, Paul J. van den, and Cajo J. F. ter Braak. 1999. "Principal response curves: Analysis of time-dependent multivariate responses of biological community to stress." *Environmental Toxicology and Chemistry* 18, no. 2 (February): 138–148.

Carey, Robert J., Gail DePalma, and Ernest Damianopoulos. 2005. "Acute and chronic cocaine behavioral effects in novel versus familiar environments: open-field familiarity differentiates cocaine locomotor stimulant effects from cocaine emotional behavioral effects." *Behavioural Brain Research* 158, no. 2 (March): 321–330.

Dam, Elsbeth A. van, Johanneke E. van der Harst, Cajo J F ter Braak, Ruud A J Tegelenbosch, Berry M. Spruijt, and Lucas P J J Noldus. 2013. "An automated system for the recognition of various specific rat behaviours." *Journal of Neuroscience Methods* 218 (2): 214–224.

Dunne, Fergal, Ambrose O'Halloran, and John P. Kelly. 2007. "Development of a home cage locomotor tracking system capable of detecting the stimulant and sedative properties of drugs in rats." 31, no. 7 (October): 1456–1463.

Ferrenberg, Scott, Sasha C Reed, and Jayne Belnap. 2015. "Climate change and physical disturbance cause similar community shifts in biological soil crusts." *Proceedings of the National Academy of Sciences* 112, no. 39 (September): 12116–12121.

Fuentes, Susana, Els van Nood, Sebastian Tims, Ineke Heikamp-de Jong, Cajo Jf F ter Braak, Josbert J Keller, Erwin G Zoetendal, and Willem M de Vos. 2014. "Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent Clostridium difficile infection." *The ISME journal* 8, no. 8 (August): 1621–1633.

Gerlai, Robert. 2002. "Phenomics: fiction or the future?" *Trends in Neurosciences* 25, no. 10 (October): 506–509.

Guo, Yanyan, Yanjie Feng, Yang Ge, Guillaume Tetreau, Xiaowen Chen, Xuehui Dong, and Wangpeng Shi. 2014. "The cultivation of Bt corn producing Cry1Ac toxins does not adversely affect non-target arthropods." *PLoS ONE* 9 (12): 1–17.

Halekoh, Ulrich, and Søren Højsgaard. 2014. "A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The {R} Package {pbkrtest}." *Journal of Statistical Software* 59 (9): 1–30.

Harkin, Andrew, John P. Kelly, John Frawley, James M. O'Donnell, and Brian E. Leonard. 2000. "Test Conditions Influence the Response to a Drug Challenge in Rodents." *Pharmacology Biochemistry and Behavior* 65, no. 3 (March): 389–398.

Hothorn, Torsten, Frank Bretz, and Peter Westfall. 2008. "Simultaneous Inference in General Parametric Models." *Biometrical Journal* 50 (3): 346–363.

Joyce, Daphné, and N. Mrosovsky. 1964. "Eating, drinking and activity in rats following 5-hydroxytryptophan (5-HTP) administration." *Psychopharmacologia* 5, no. 6 (November): 417–423.

Kas, Martien J.H., and Jan M. van Ree. 2004. "Dissecting complex behaviours in the post-genomic era." 27, no. 7 (July): 366–369.

Krackow, S, E Vannoni, a Codita, a H Mohammed, F Cirulli, I Branchi, E Alleva, et al. 2010. "Consistent behavioral phenotype differences between inbred mouse strains in the IntelliCage." *Genes, brain, and behavior* 9, no. 7 (October): 722–31.

Kuznetsova, Alexandra, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. 2016. *lmerTest: Tests in Linear Mixed Effects Models.*

Lenth, Russell V. 2016. "Least-Squares Means: The {R} Package {lsmeans}." *Journal of Statistical Software* 69 (1): 1–33.

Lipkind, Dina. 2004. "New replicable anxiety-related measures of wall vs. center behavior of mice in the open field." 97, no. 1 (March): 347–359.

Loos, Maarten, Bastijn Koopmans, Emmeke Aarts, Gregoire Maroteaux, Sophie van Der Sluis, Matthijs Verhage, August B. Smit, et al. 2014. "Sheltering behavior and locomotor activity in 11 genetically diverse common inbred mouse strains using home-cage monitoring." *PLoS ONE* 9 (9): 1–9.

Mayor, Jordan R, Nathan J Sanders, Aimée T Classen, Richard D Bardgett, Jean-Christophe Clément, Alex Fajardo, Sandra Lavorel, et al. 2017. "Elevation alters ecosystem properties across temperate treelines globally." *Nature Publishing Group* 542 (7639): 91–95.

Oksanen, Jari, F Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R Minchin, et al. 2017. *vegan: Community Ecology Package.*

Peters, Suzanne M., Joe A. Tuffnell, Ilona J. Pinter, Johanneke E. van der Harst, and Berry M. Spruijt. 2017. "Short- and long-term behavioral analysis of social interaction, ultrasonic vocalizations and social motivation in a chronic phencyclidine model." *Behavioural Brain Research* 325:34–43.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Robinson, Lianne, and Gernot Riedel. 2014. "Comparison of automated home-cage monitoring systems: Emphasis on feeding behaviour, activity and spatial learning following pharmacological interventions." *Journal of Neuroscience Methods* 234:13–25.

Rogan, Sarah C., and Bryan L. Roth. 2011. "Remote control of neuronal signaling." *Pharmacological Reviews* 63 (2): 291–315.

Ruff, S Emil, Jennifer F Biddle, Andreas P Teske, Katrin Knittel, Antje Boetius, and Alban Ramette. 2015. "Global dispersion and local diversification of the methane seep microbiome." *Proceedings of the National Academy of Sciences* 112, no. 13 (March): 4015–4020.

Smilauer, Petr, and Jan Lepš. 2014. *Multivariate Analysis of Ecological Data using CANOCO 5,* 1–362.

Spruijt, Berry M., Suzanne M. Peters, Raymond C. de Heer, Helen H.J. Pothuizen, and Johanneke E. van der Harst. 2014. "Reproducibility and relevance of future behavioral sciences should benefit from a cross fertilization of past recommendations and today's technology: "Back to the future"." 234 (August): 2–12.

Spruijt, Berry M., and Leonie de Visser. 2006. "Advanced behavioural screening: automated home cage ethology." *Drug Discovery Today: Technologies* 3 (2): 231–237.

Storkey, J, A J Macdonald, P R Poulton, T Scott, I H Köhler, H Schnyder, K W T Goulding, and M J Crawley. 2015. "Grassland biodiversity bounces back from long-term nitrogen addition." *Nature* 528, no. 7582 (December): 401–404.

Tecott, Laurence H, and Eric J Nestler. 2004. "Neurobehavioral assessment in the information age." 7, no. 5 (May): 462–466.

Visser, L de, R van den Bos, W W Kuurman, M J H Kas, and B M Spruijt. 2006. "Novel approach to the behavioural characterization of inbred mice: automated home cage observations." *Genes, brain, and behavior* 5, no. 6 (August): 458–66.

Visser, Leonie de, Ruud van den Bos, and Berry M Spruijt. 2005. "Automated home cage observations as a tool to measure the effects of wheel running on cage floor locomotion." *Behavioural brain research* 160, no. 2 (May): 382–8.

Wurbel, H. 2002. "Behavioral phenotyping enhanced - beyond (environmental) standardization." 1, no. 1 (January): 3–8.

S.2.A. Intuitive interpretation of PCA

PCA is a method for summarizing observations in a data set in terms of the RV. The aim is to describe the data set using less RV than the original data sets. Rather than selecting the most appropriate RV and removing the others, new RVs are constructed using linear combinations of the original RVs. These linear combinations are called Principal Components (PCs). A data set has as many PCs as there are RVs. The best linear combination to describe the data is the first principal component (PC1). PC1 is defined such that it summarizes the most variation between observations. This aim coincides with it being best able to approximate the original data set in one dimension.

In this supplement we will illustrate PCA graphically. For simplicity, we use a subset of the data of the case study (Section 2.2.1). This subset has only 2 RV (DML and MVP) and contains only BT observations from the Gq^--group (Figure S.2.A.1). The observations are numbered from lowest value for DML to highest and enveloped.

Because the data set has only two RVs, PCA will result in 2 PCs. Most variation in the cloud of data points is seen roughly along the line of the lower left corner (observations 2 and 4) to the upper right corner (below observation 39). PCA will rotate (or project) the data such that the largest amount of variation in the data set is along PC1 (Figure S.2.A.2). PCs are by definition uncorrelated and thus PC2 will be perpendicular to PC1.

As mentioned before, PCs are linear combinations of RVs. For every observation in this example we can calculate the PC-scores (*i.e.* the position in the PCA plot) using the: $\text{PC1} = 0.71 * \text{DML} + 0.71 * \text{MVP}$ and $\text{PC2} = -0.71 * \text{DML} + 0.71 * \text{MVP}$. The coefficients of a RV to calculate the PC-score is the loading of that RV. If a RV has a positive loading it is positively associated with that PC, if a RV has a negative loading it is negatively associated with that RV.

The PCA plot (Figure S.2.A.2) is a rotated version of the original data (Figure S.2.A.1). The interpretation of the axes thus changes, but the shape of the cloud of points does not. The arrow heads for the RVs in Figure S.2.A.2 are the rotated version of the points (1,0) and (0,1) in Figure S.2.A.1.

PCA can be extended from 2 dimensional to more dimensional situations. A graphical representation of what happens becomes more difficult because data is now in 3d (Figure S.2.A.3). The grid or coordinate system that holds the cloud of points is no longer a square on a piece of paper, it is a cube. The cloud of point in the cube however, can still be rotated such that the maximum amount of variation is on PC1 (in one dimension), or on PC1 and PC2 (in two dimensions). Again, the shape of the cloud does not change in three dimensions, but we loose some detail when viewing it in two dimensions (Figure S.2.A.3).

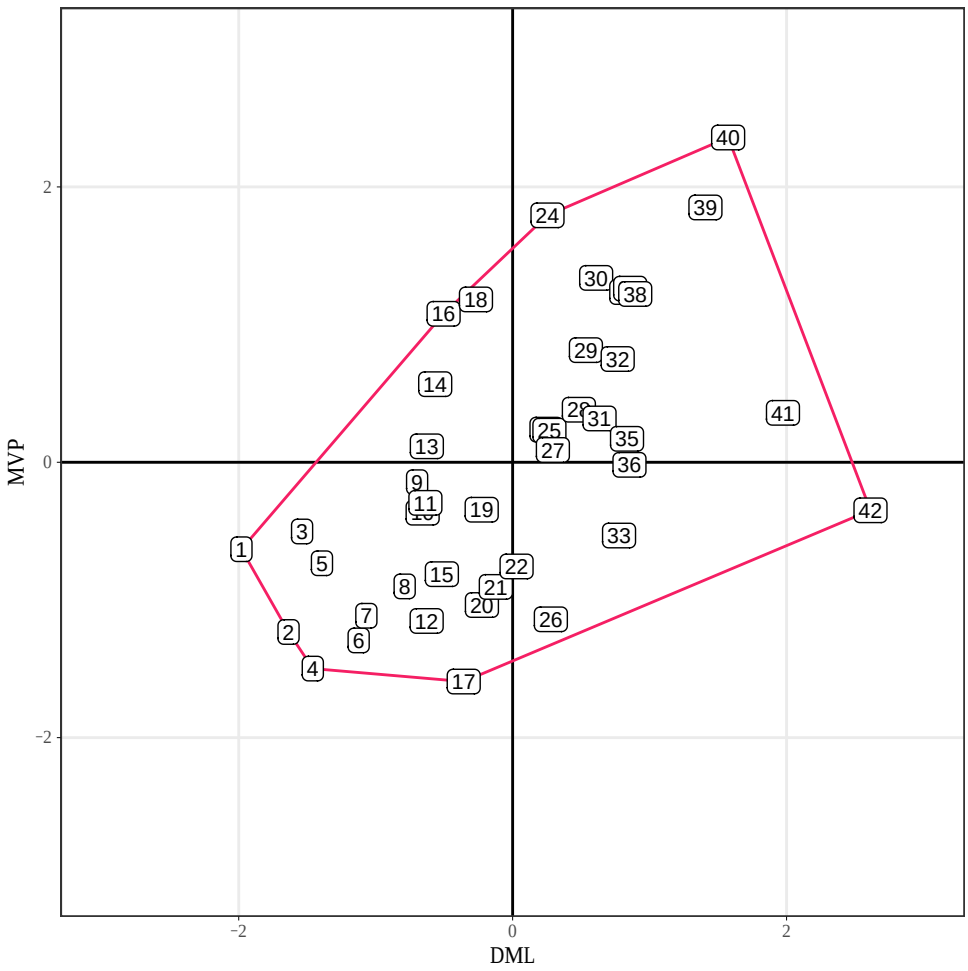


Figure S.2.A.1: Scatter plot of response variables DML and MVP of the Gq^{-}-group before treatment. Observations are labelled from lowest to highest value for DML. An envelop is drawn around the most outward observations

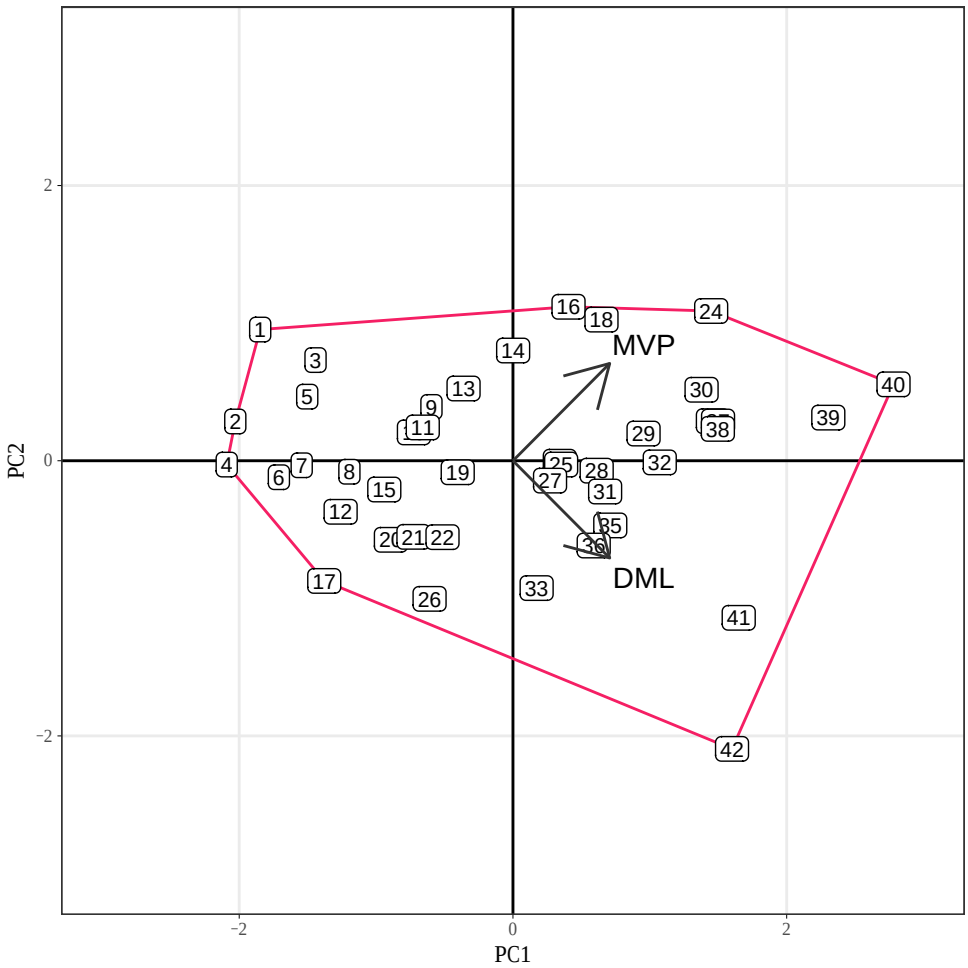


Figure S.2.A.2: PCA plot of the first and second principal component of the data in Figure S.2.A.1. The axis have been rotated such that PC1 explains the highest amount of variance possible. The arrows for MVP and DML represent the coefficients in the linear combination. MVP has a positive loading for both the PC1 and the PC2, DML has a positive loading for PC1 and a negative loading for PC2.

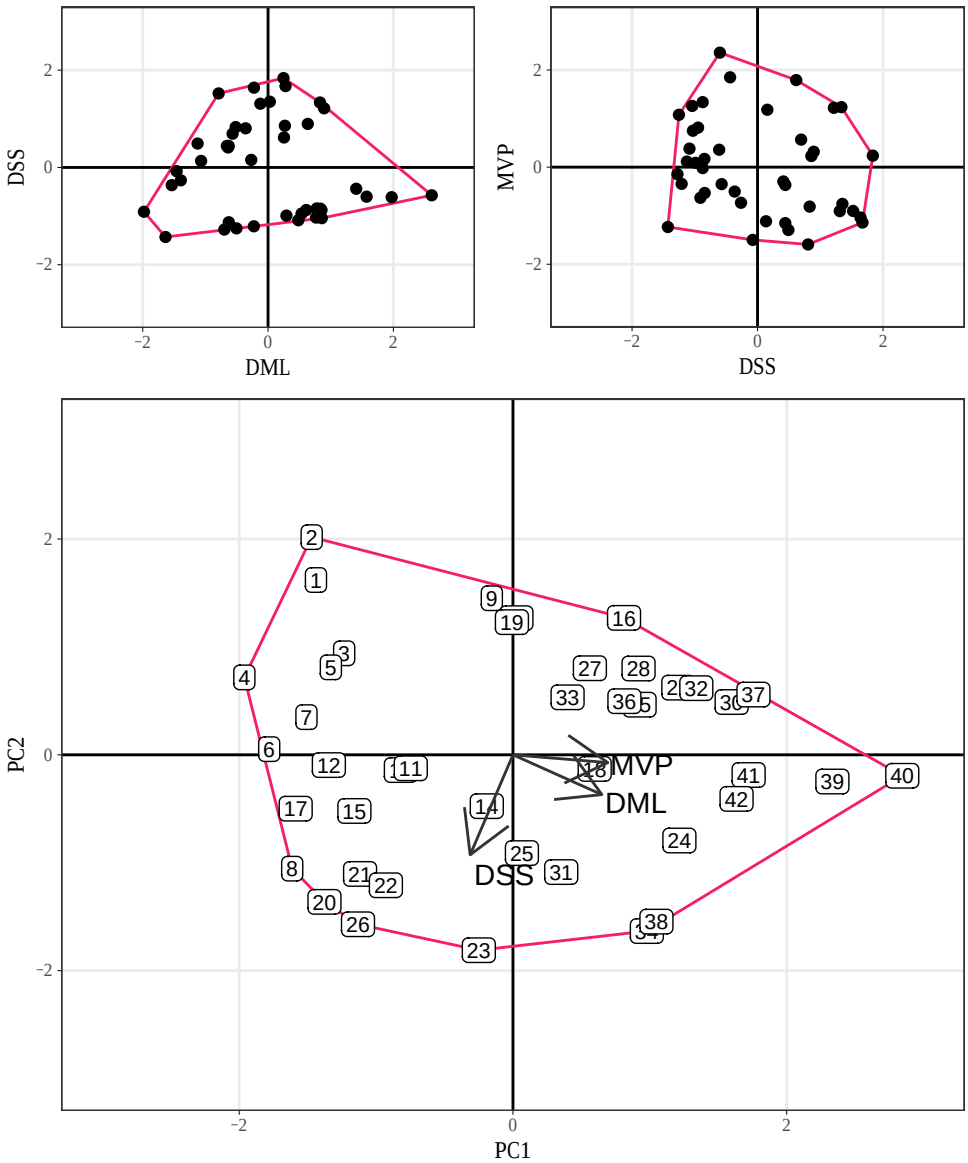


Figure S.2.A.3: PCA on the data in Figure S.2.A.1 with a third RV: DSS. Top row: Scatter plots of response variables DML, MVP, and DSS of the Gq^{-}-group before treatment. Bottom row: PCA plot of the first and second principal component. The axis have been rotated such that PC1 explains the highest amount of variance possible. The arrows for DML, MVP, and DSS represent the coefficients in the linear combination. MVP has a positive loading for PC1 and a small negative loading for PC2, DML has a positive loading for PC1 and a negative loading for PC2, DSS has a negative loading for both PC1 and PC2.

S.2.B. Supplementary Figures and Tables

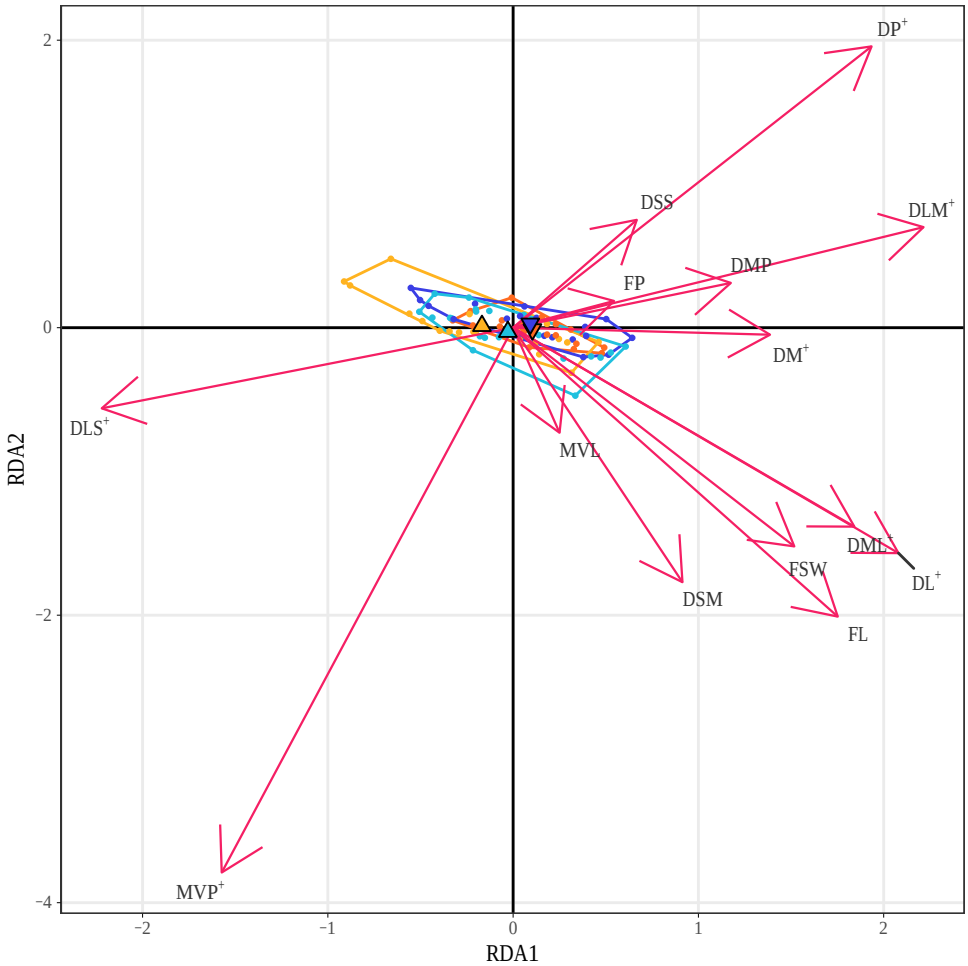
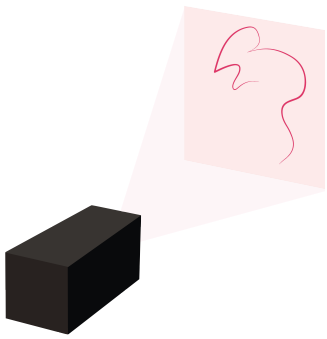


Figure S.2.B.1: Results of the RDA-analysis on the Gq^--group using scaling = 1. Scores of the first RDA-axis are plotted on the x-axis and scores of the second RDA-axis are plotted on the y-axis. Individual observations are represented by points that are coloured by Drug (CNO: orange, Control: blue) and Timing (Before Treatment: light hues, After Treatment: dark hues). All individual observations in the same treatment group are enveloped. Centroids for the treatment groups are represented by triangles (Before Treatment: upward, After Treatment: downward; coloured as the observations). RVs are represented by named arrows. Names of RVs that had a significant Timing-effect interaction in the univariate model are indicated by a plus

Table S.2.B.1: Number of Significant Differences of Least Squares Means (LSdiffs) between Treatment and Control per Time-point, per RV, summed for the three doses of Diazepam and H3. RVs are ordered based on the total number of significant LSdiffs for all Treatments

		Time										
Drug		1	3	4	5	6	7	8	9	11	12	All
Diazepam	L mov. #	.	2	2	2	1	.	.	.	.	.	7
	Dur. OnShelter z	2	1	2	.	.	.	.	.	1	.	6
	# activity bouts	.	1	2	2	.	.	.	.	.	.	5
	L mov. dist.	.	.	2	2	.	.	.	.	.	.	4
	# s arrests	.	.	2	2	.	.	.	.	.	.	4
	Dist. Mov.	.	.	2	2	.	.	.	.	.	.	4
	Dur.activity	.	.	2	2	.	.	.	.	.	.	4
	S mov. dist.	.	.	2	1	.	.	.	1	.	.	4
	Dur. s arrests	.	.	2	2	.	.	.	.	.	.	4
	S mov. #	.	.	2	1	.	.	.	1	.	.	4
	# L arrests	.	.	2	.	.	.	.	.	.	.	2
	# s SV	.	.	1	.	.	.	.	.	.	1	2
	Dur. feeding z	.	1	1	.	.	.	.	.	.	.	2
	Dur. l arrests	.	.	1	.	.	.	.	.	.	.	1
	Dur. s SV	.	.	.	.	.	.	.	.	.	1	1
	Dur. Spout z	.	.	1	.	.	.	.	.	.	.	1
H3	L mov. #	.	.	.	.	3	2	3	.	1	2	11
	Dur. OnShelter z	.	.	.	1	1	2	3	.	1	3	11
	# activity bouts	.	.	.	.	3	2	3	.	1	2	11
	L mov. dist.	.	.	.	.	3	2	3	.	1	3	12
	# s arrests	.	.	.	.	3	2	3	.	1	2	11
	Dist. Mov.	.	.	.	.	3	2	3	.	1	2	11
	Dur.activity	.	.	.	.	3	2	3	.	1	2	11
	S mov. dist.	.	.	.	.	3	2	3	.	1	2	11
	Dur. s arrests	.	.	.	.	3	2	2	.	1	2	10
	S mov. #	.	.	.	.	3	2	2	.	1	2	10
	# L arrests	.	.	.	.	3	2	3	.	.	3	11
	# s SV	.	.	.	.	3	2	3	.	1	2	11
	Dur. feeding z	.	.	.	.	3	2	3	.	.	1	9
	Dur. l arrests	.	.	.	.	3	2	3	.	.	2	10
	Dur. s SV	.	.	.	.	3	2	2	.	.	2	9
	Dur. Spout z	.	.	.	.	.	1	1	.	.	.	2



3

Response Variable selection in Principal Response Curves using permutation testing

Nadia J. Vendrig

Lia Hemerik

Cajo J.F. ter Braak

Published as:

Nadia J. Vendrig, Lia Hemerik, and Cajo J. F. ter Braak. 2016. "Response variable selection in principal response curves using permutation testing"

Principal Response Curves analysis (PRC) is widely applied to experimental multivariate longitudinal data for the study of time-dependent treatment effects on the multiple outcomes or response variables (RVs). Often, not all of the RVs included in such a study are affected by the treatment and RV-selection can be used to identify those RVs and so give a better estimate of the principal response. We propose four backward selection approaches, based on permutation testing, that differ in whether coefficient size is used or not in ranking the RVs. These methods are expected to give a more robust result than the use of a straightforward cut-off value for coefficient size. Performance of all methods is demonstrated in a simulation study using realistic data. The permutation testing approach that uses information on coefficient size of RVs speeds up the algorithm without affecting its performance. This most successful permutation testing approach removes roughly 95% of the RVs that are unaffected by the treatment irrespective of the characteristics of the data set and, in the simulations, correctly identifies up to 97% of RVs affected by the treatment.

3.1. Introduction

In ecological research, the effect of a treatment is often assessed for several response variables (RVs) at several points in time. This results in multivariate longitudinal data, also called multivariate time series data. For instance, if we wish to assess how invertebrate communities in ditches change as a result of a single application of a certain pesticide, we would select a number of ditches (experimental sites), assign every ditch to a treatment of a dose of pesticide or a control treatment, and measure the abundances of the invertebrate species living in the ditches at several times before and after treatment. Abundance of invertebrates is not only influenced by our treatment but also by the moment of sampling due to external factors such as the time of year. Principal Response Curves analysis (PRC) (Brink and Braak 1998; Brink and ter Braak 1999) removes these unwanted time effects; succinctly describes the time-dependent overall-response of the community to the treatment(s) relative to the control treatment; and indicates for each of the species whether their response is positively or negatively correlated to the overall-response and to which extent.

PRC is a special case of Redundancy Analysis (RDA) used to describe experimental multivariate longitudinal data. It estimates differences among treatments on a collection of RVs over time and the extent to which the response of those individual RVs resembles the overall response. PRC has been widely applied in aquatic ecology and ecotoxicology (*e.g.* Cuppen et al. 2000; Duarte et al. 2008; Hartgers et al. 1998; Roessink et al. 2006; Verdonschot et al. 2015), terrestrial ecology and ecotoxicology (*e.g.* Britton and Fisher 2007; Heegaard and Vandvik 2004; Moser et al. 2007; Pakeman 2004), microbiology (*e.g.* Andersen et al. 2010; Fuentes et al. 2014) and soil science (*e.g.* Cardoso et al. 2008; Kohler et al. 2006).

The main result of PRC are two sets of coefficients visualized in two easily interpretable graphs. The first set consists of the dose-time coefficients (c_{dt}s) estimated for each combination of the treatment-levels ($d = 1, \ldots, D$) and the time-points ($t = 1, \ldots, T$). The c_{dt}s represent the effect-size of treatment d at time t relative to the reference treatment at the same time. Thus, by definition, $c_{dt} = 0$ for the reference treatment. The reference treatment is often the control treatment, but the choice of reference treatment does not affect the estimates of differences between treatments; it merely defines the baseline, *i.e.* relative to which treatment the results are presented. The c_{dt}s are depicted in the Principal Response Curves, a line-plot of c_{dt}s against time grouped by treatment (Figure 3.1). The second set of coefficients are the weights for the RVs (b_ks) estimated for each of the RVs ($k = 1, \ldots, K$). They represent the resemblance of RV k to the overall response pattern specified by the Principal Response Curves (*i.e.* the c_{dt}s) and are typically depicted on a vertical bar alongside the line-plot. The further b_k is from zero, the more the response pattern of RV k resembles the overall re-

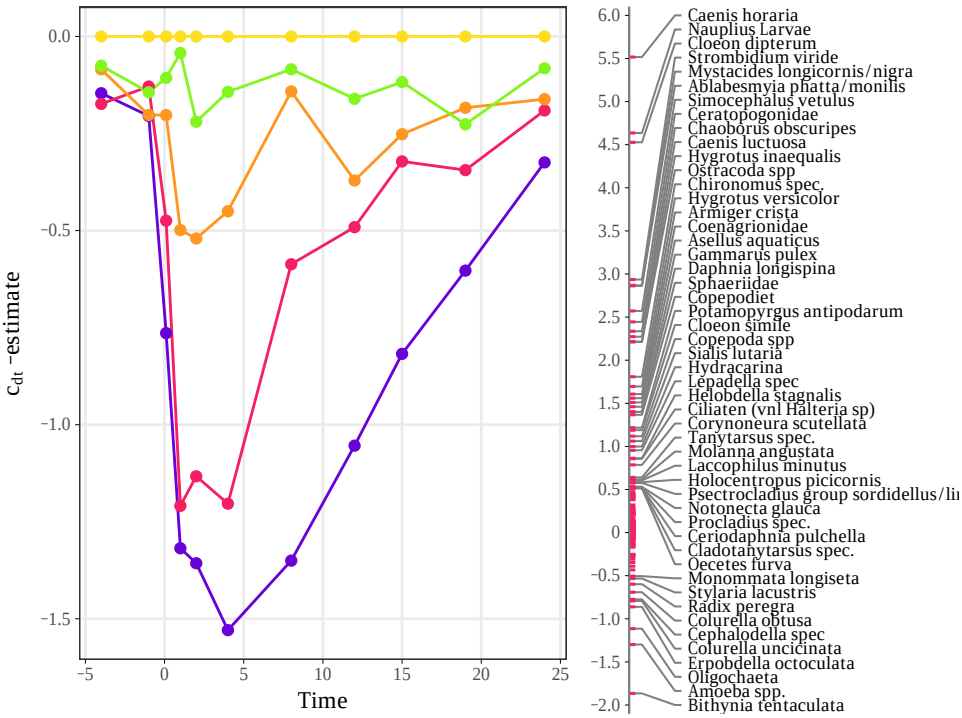


Figure 3.1: Principal Response Curves (left) for the Pyrifos data (Brink and ter Braak 1999) for the different doses of Chlorpyrifos (0: ●, 0.1: ●, 0.9: ●, 6: ●, and 44μg/L: ●) with b_k-estimates (right). Only RVs with an absolute b_k-estimate above 0.5 are labelled

sponse pattern (if $b_k > 0$) or the negative overall response pattern (if $b_k < 0$). A b_k of zero indicates that the expected value of RV k at time t does not differ between treatments or is uncorrelated with the overall response pattern. The c_{dt}s and b_ks can be used to rank dose-time combinations or RVs respectively. For instance, if $|c_{23}| > |c_{24}|$, the estimated treatment effect for treatment 2 is larger at time-point 3 than at time-point 4. The coefficients however, do neither have a unit nor a direct interpretation. The coefficients are estimated under the assumption that $\pi_{tdk} = c_{dt}b_k$, where π_{tdk} is the difference in expected value of RV k, at time t between treatment d and the reference treatment. The expected value y_{tdk} of RV k, at time t under treatment d is thus estimated as $y_{tdk} = a_{tk} + \pi_{tdk}$, where a_{tk} is the expected value in the reference group.

Standard PRC assumes that only one factor (*e.g.* treatment) is relevant, while other (environmental) factors are either as similar as possible or, if not, randomized by design of the experiment. PRC has also been applied to monitoring sites where this assumption is more problematic. It should be noted that it is possible to adjust for unwanted variation between sites if this variation is due to one

or more measured environmental variables. The environmental variables can be included as covariates in addition to the factor time which is the default covariate in PRC. This possibility is not yet available in Vegan (Oksanen et al. 2015), a much used R-package that includes a PRC-function, but it is available in Canoco 5 (Smilauer and Lepš 2014), a computer program for multivariate statistical analysis using ordination. An example is given in Fuentes et al. (2014).

When PRC is applied in aquatic ecology, the research interest typically is the response of a community as a whole to a treatment and the set of RVs thus typically consists of abundance data on all available species or taxa (*e.g.* all taxa of invertebrates) at an experimental site. RVs are included irrespective of their expected susceptibility to the treatment beforehand, and a large proportion of the included RVs could thus be unaffected by the treatment. PRC handles RVs that do not follow the response pattern (Noise-RVs) by assigning these RVs b_k-estimates close to zero which is advantageous in contrast to the use of *e.g.* Bray-Curtis Similarity (Bray and Curtis 1957) which is calculated with equal weights for all RVs (Brink and Braak 1998). But although inclusion of Noise-RVs in PRC does not add bias to c_{dt}-estimates, their inclusion introduces extra noise into the data set which adds extra imprecision to the estimates and reduces power. It would be advantageous to be able to point-out which RVs are Noise-RVs. Reducing the data set accordingly would not only improve c_{dt}-estimation, it would also improve comparability of results of PRC between studies. As of yet this is difficult because the coefficients have no unit so only the shape of the principal response curves and the order of the species weights can be compared between studies. Reduction of the number of RVs in the analysis would also improve the readability of RV-weights graphs. At present, authors improve readability of the RV-weights graph by showing only RVs that exceed a certain threshold (mostly 0.5). Although effective in reducing the number of RVs, this practice is at best sub-optimal because b_k values (1) depend on the extent to which other RVs in the same data set are affected by the treatment, (2) are affected by the type of scaling used, and (3) are affected by the choice of standardization (details and illustrated examples on effect of these factors on b_k-estimates are given in Section S.3.A.1).

In this paper, we propose permutation testing approaches as an improved method for RV-selection in PRC. We further show that these approaches are robust to high residual correlation between RVs, and to adding additional RVs with strong effect (very high b_k) or adding many RVs with no effect ($b_k = 0$) to the data set. We specifically show that information obtained from ranking RVs based on b_k scores of the full model, can help accelerate the algorithm for variable selection without performance loss.

3.2. Materials and Methods

3.2.1. Principal Response Curves Analysis

PRC models the expected value of RV k at time t in treatment level d as the sum of three effects: (1) the expected value of the RV in the reference group a_{tk}, (2) the time-specific effect of treatment level (π_{tdk}), and (3) an error-term (ε_{ik}). The (multivariate) regression model for y_{ik}, *i.e.* the observed value of RV k in observation i (where $i = 1, \ldots I$, with $I = T{\cdot}$number of experimental sites), is:

$$y_{ik} = \sum_{t=1}^{T} a_{tk} w_{it} + \sum_{t=1}^{T} \sum_{d=1}^{D} \pi_{tdk} z_{idt} + \varepsilon_{ik} \tag{3.2}$$

where w_{it} and z_{idt} are indicator variables (0/1 or dummy variables) that indicate, respectively, whether (1) or not (0) observations are in the reference treatment and whether or not observations received dose d at time t. The general assumption of PRC is that $\pi_{tdk} = b_k c_{dt}$ which implies that b_k and c_{dt} can be estimated by partial RDA (*i.e.* reduced rank regression with concomitant variables) (Davies and Tso 1982) using equation 3.2. Note that, in contrast to what is written in Smilde et al. (2012) and in the appendix of Timmerman and ter Braak (2008), a_{tk} is a free, unknown parameter of the model that is estimated by the partial RDA. Note that the estimation procedure also works with unbalanced data, as PRC fits in the regression framework which is more general than the ANOVA framework used by Smilde et al. (2012).

The estimates for c_{dt} and b_k are determined on an arbitrary scale because $c_{dt} b_k = \beta b_k * \frac{c_{dt}}{\beta}$, where β is an arbitrary scalar (*i.e.* any real number). As a result, the coefficients lack a unit and a direct interpretation and the scalar can be chosen such that it gives the coefficients the desired properties. In Canoco (Smilauer and Lepš 2014), the first software package to include PRC, the default is to scale coefficients such that the mean square of b_k-estimates is 1 and we used this scaling in Figure 3.1. The result is that, *ceteris paribus*, larger true treatment effects result in larger absolute estimates of c_{dt}. The b_k-estimates are expected to fall roughly between -3 and 3, independent of treatment effect. Therefore, when applying this scaling one could opt to select RVs based on a cut-off value of absolute b_k (usually 0.5).

This approach, which we will refer to as Naive RV-selection (Naive RVS), has some pitfalls. We wish to distinguish RVs affected by the treatment (Effect-RVs) from RVs that are uncorrelated to the overall response pattern. Such RVs are either unaffected by the treatment (Noise-RVs) or contribute to minor response patterns. In a situation with only Noise-RVs however, due to scaling, some Noise-RVs will get a b_k-estimate above the cut-off value. *Vice versa*, scaling causes the b_k-estimate of an Effect-RV to be lower when a very strongly affected Effect-RV

is in the data set than when that strongly affected RV is not in the data set. As a result, including a very strongly affected RV to the data set could result in b_k-estimates of other RVs to drop below the cut-off value. Another pitfall is that Naive RVS has little value when coefficients are scaled differently. Coefficients could for instance be scaled such that mean square of $\tilde{c}_{dt}$s is 1, where $\tilde{c}_{dt}$s are a centered version of the c_{dt}s. In Vegan (Oksanen et al. 2015) the default option scales the coefficients differently with both the b_ks and c_{dt}s showing effect sizes. For any of these scaling-methods, choosing a cut-off value in advance does not make sense.

3.2.2. Response Variable Selection protocols

Ideally, an RVS-protocol would make perfect predictions and thus remove all the Noise-RVs from the model and keep all the Effect-RVs in the model. Such a result is not feasible in practice. Therefore we aim at achieving an optimal, yet realistic method for RVS, in which every Noise-RV has a 1-α probability to be removed from the model (*e.g.* $\alpha = 0.05$) whilst keeping as many Effect-RVs in the model as possible. With this aim there is no need to correct for multiplicity in statistical testing of RVs (such as Bonferroni) in the RVS-protocols that we propose.

For any RV k, the hypothesis that its expected value is independent from the treatment (*i.e.* whether or not $b_k = 0$) can be tested by calculating a permutation p-value and comparing it to α. A permutation p-value for RV k is obtained by performing 500 permutations in which time-series of observations from RV k on the same experimental unit (*e.g.* ditch, plot, or site) are permuted between treatments (including the control treatment). We estimate b_k in PRC on non-permuted data and on all 500 permuted data sets. The permutation p-value is the proportion of the 501 estimated b_ks (including the b_k from non-permuted data) greater than or equal to the estimated b_k from PRC with non-permuted data, if the estimated b_k from the PRC with non-permuted data is positive. If the estimated b_k from PRC with non-permuted data is negative, the proportion equal or lower is used. The number of 500 is large enough to provide sufficient power at $\alpha = 0.05$ and is still acceptable in terms of computing time.

As an alternative to Naive-RVS, we propose four RVS-protocols based on permutation testing (in short: permutation RVS-protocols) that all incorporate permutation p-value calculation as described above. All four permutation RVS-protocols are backward procedures, indicating that they start with the whole set of RVs and predict which of those are Noise-RVs that can be removed from the model and which are Effect-RVs that should be kept.

Two Step RVS

The most thorough permutation RVS-protocol is the Two Step RVS. In this protocol, we calculate a permutation p-value for all RVs in the data set. If any of

the permutation p-values is higher or equal to α, the RV with the highest permutation p-value is removed from the model. Thereafter, we repeat the procedure with the remaining RVs and keep repeating until only RVs with a permutation p-value lower than α remain. The advantage of this elaborate approach is that it accounts for RVs being correlated. The pitfall is that it is computationally intensive because many permutation p-values need to be calculated (*e.g.* for $K = 200$; as many as $0.5(K^2 + K) = 20,100$).

Screening RVS

We could do with a less computationally intensive protocol if it would be reasonable to assume that the permutation p-value of an RV is independent of the other RVs in the data set. This simpler protocol, called the Screening RVS protocol, calculates a permutation p-value once for each RV in the data set using the full model. All RVs with permutation p-values higher or equal to α are removed from the model at once.

Stepwise RVS

Importantly, estimated b_ks of Noise-RVs are expected to be closer to zero than estimated b_ks of Effect-RVs. Thus, to incorporate this information, a third RVS approach uses an even less computationally intensive procedure. This protocol, called the Stepwise RVS protocol, performs PRC on the data set, selects the RV with the estimated b_k closest to zero, and calculates a permutation p-value for that RV. If that permutation p-value is higher or equal to α, it removes the RV from the model. If it is not, it keeps the RV in the model and calculates the permutation p-value of the RV with the estimated b_k second closest to zero. Once an RV is kept in the model, its permutation p-value is not calculated again. Stepwise RVS is computationally less intensive than Screening RVS because the PRC-procedure, which is performed 501 times per permutation p-value, gets faster with a smaller number of RVs in the model. In Stepwise RVS, permutation p-values are calculated using PRC on the reduced model with increasingly less RVs as the procedure progresses, whereas, in Screening RVS, all permutation p-values are calculated using PRC on the full set of RVs.

Stepwise Stop RVS

When we are willing to assume that all RVs with an absolute estimated b_k under a certain threshold are Noise-RVs, we can make an even faster version of the Stepwise RVS protocol: the Stepwise Stop RVS protocol. This protocol is the same as the Stepwise RVS protocol, except that it stops entirely when the first permutation p-value lower than α is encountered.

3.2.3. Simulation Study

We evaluated the performance of the four permutation testing protocols and Naive RVS in a simulation study. The data used in this simulation study was modelled after the so-called Pyrifos data set. The Pyrifos data set, used as example throughout this paper, consists of log-transformed abundance data obtained from a toxicological experiment in outdoor experimental ditches, explained in detail by Wijngaarden et al. (1996) and Brink et al. (1996). In the experiment, experimental ditches were randomly allocated to the reference treatment or a dose of insecticide chlorpyrifos. The RVs are abundances of species of invertebrates. In this simulation study, we generated data from scenarios inspired by the Pyrifos-experiment. In the Pyrifos-like data scenario, an experiment was conducted in which the effects of three levels of treatment (reference, low and high dose) were measured on four independent locations per treatment at five different time-points. The Pyrifos-like data contains abundance data of 100 RVs, 50 of which are noise RVs which are unaffected by the treatment ($b_k = 0$) and 50 are effect-RVs which have a low, medium, high or reversed low treatment effect ($b_k = 1$, 2, 3, or -1). Covariance between time-points is auto-regressive and covariance between RVs resembles covariance in the Pyrifos data set. Error terms were simulated using a multivariate normal distribution. We back-transformed the sum of the structural effect and the error term to the abundance-scale, used it as expected value for a random draw from a Poisson-distribution, and log-transformed the result (more details in Section S.3.B).

To provide additional experimental outcomes that approximated the range of treatment effects in the literature, we also generated data based on 17 data scenarios similar to the Pyrifos-like data scenario with one or two parameters manipulated. We manipulated the composition of the set of Effect-RVs, the number of Noise-RVs, the number of ditches, the amount of covariance between RVs, and the treatment-effect size. For an overview see Table 3.1.

For each of the 18 data scenarios, 100 data sets were generated which were centered before analysis (Centering). We also analyzed each data set after standardizing data per RV (Standardization) resulting in another 18 simulation scenarios. Standardization in addition to Centering is useful when it is of interest whether RVs are affected by a treatment (positively, negatively, or not at all) and not so much what the size of the difference in effect between RVs is. For Naive-RVS, coefficients were scaled such that mean squares of b_k are 1 as this is the only scaling that is sensible for this protocol. Scaling of coefficients does not affect the RV selection in the permutation RVS-protocols.

Performance of the RVS-protocols was evaluated using sensitivity and specificity. Sensitivity is the number of Effect-RVs kept in the model divided by the total number of Effect-RVs in the data set. Specificity is the number of Noise-RVs

Table 3.1: Overview of data scenarios in the simulation study with three treatments, incl. control, at five time-points

Data Scenario	Description
Pyrifos-like	as described in section 3.2.3 (4 replications, 50 Effect-RVs, and 50 Noise-RVs)
More Ditches	as Pyrifos-like, with 4 additional ditches per treatment (8 total)
Most Ditches	as Pyrifos-like, with 8 additional ditches per treatment (12 total)
Weak Effect-RVs	as Pyrifos-like, with Effect-RVs consisting of 38 RVs with $b_k = 1$ and 12 RVs with $b_k = -1$
Strong Effect-RVs	as Pyrifos-like, with 12 additional strong Effect-RVs (b_k=10)
One Noise-RV	as Pyrifos-like, with only 1 Noise-RV
Many Noise-RVs	as Pyrifos-like, with 150 additional Noise-RVs (200 total)
No Covariance	as Pyrifos-like, except there is no covariance between RVs
More Covariance	as Pyrifos-like, with 40% higher correlation between RVs
<name of data scenario>$^+$	all nine data scenarios described above, with a larger treatment effect ($c_{dt}{}^+ = 4c_{dt}$)

removed from the model divided by the total number of Noise-RVs in the data set. Permutation method are expected to have a specificity of 0.95 with $\alpha = 0.05$, indicating that 5% of saved RVs could in fact be Noise-RVs. In the ideal situation, sensitivity would be 1, indicating that all effect-RVs are identified. In practice, we would expect sensitivity to increase with increasing power, *e.g.* with larger effect-size or more observations.

There is a trade-off between specificity and sensitivity which becomes apparent when comparing both Stepwise RVS procedures. All RVs removed in the Stepwise Stop RVS procedure are also removed in the Stepwise RVS procedure. In the Stepwise RVS procedure some additional RVs could be removed. Stepwise Stop RVS thus always keeps the same or more Effect-RVs in the model than Stepwise RVS and thus has an equal or higher sensitivity. Stepwise Stop RVS always removes the same number or less Noise-RVs from the model than Stepwise RVS and thus has an equal or lower specificity.

The overall quality of RVS-protocols was evaluated with the Matthews correlation coefficient (M_c) (Matthews 1975) which is a correlation coefficient between a prediction and the reality:

$$M_c = \frac{TP * TN - FP * FN}{(TP + FN)(TN + FP)(TP + FP)(TN + FN)} \tag{3.2}$$

where TP (true positives) is the number of kept Effect-RVs, TN (true negatives) is the number of removed Noise-RVs, FP (false positives) is the number of kept Noise-RVs, and FN (false negatives) is the number of removed Effect-RVs. The M_c ranges between -1 and 1 where 1 indicates perfect prediction (*i.e.* all Noise-RVs removed, all Effect-RVs kept), 0 indicates prediction no better than random, and -1 indicates total disagreement between prediction and reality (*i.e.* all Noise-RVs kept, all Effect-RVs removed).

The effect of RVS on model fit was evaluated in terms of difference in residual mean squared error ($\text{RMSE}_{\text{diff}}$). RMSE of the reduced model ($\text{RMSE}_{\text{reduced}}$) was compared to RMSE of the reduced set of RVs calculated using fitted values from the full model ($\text{RMSE}_{\text{full}}$).

After evaluating performance of the RVS-protocols we applied the best protocol to the Pyrifos data as a case study. In order to better compare the shapes of PRC on the full and the reduced data set, we scaled such that the population variance of all available case scores $\{x_i = c_{dt} z_{idt}\}$ was 1. For balanced data, this corresponds to setting the mean square of $\tilde{c}_{dt}$s to 1. The scaling such that the mean square of b_k is 1 always results in higher b_k-estimates and lower c_{dt}-estimates when comparing results before to after removing Noise-RV, because Noise-RV typically have low b_k-estimates. All data simulations and analyses were performed in R 3.1.0. The scripts to replicate the case-study are available as Online Resource 3 to Vendrig, Hemerik, and Braak (2016).

3.3. Results

3.3.1. General Results

In our simulation study, we assessed sensitivity, specificity, and M_c of the Two Step, Screening, Stepwise, and Stepwise Stop permutation RVS-protocols and Naive RVS. The aim was to find an RV-selection method that is 0.95 specific whilst being as sensitive as possible. Computing time of the Two Step RVS-protocol was extremely long. Analysis of one data set generated using the Pyrifos-like data scenario took on average 2 hours and 24 minutes, whereas Screening RVS took 3 minutes 50 seconds, Stepwise RVS took 2 minutes and 48 seconds, and Naive RVS took less than a second. Therefore Two Step RVS was run on 12 rather than 100 data sets per scenario. The results thereof gave no reason to assume that Two Step RVS outperformed Screening or Stepwise RVS. On the contrary, based on confidence intervals around the mean, we found that mean specificity in the Two Step RVS was different from 0.95 in 7 out of 36 simulation scenarios whereas for Screening and Stepwise RVS, also based on 12 iterations, mean specificity was different from 0.95 in respectively 3 and 0 out of 36 data scenarios. As a result, we decided to base results of the Two Step RVS on 12 iterations and not report the results in text.

Based on 100 data sets per scenario, we concluded that Screening and Stepwise RVS hardly differed in specificity and sensitivity. Per scenario, the difference between methods in mean specificity ranged from -0.020 to 0.030 and the difference in mean sensitivity ranged from -0.011 to 0.006. The Stepwise Stop RVS protocol did not meet the requirement of being 0.95 specific. The 95% confidence interval of mean specificity excluded 0.95 in all of the 36 simulation scenarios. Therefore, we will only report on results from Stepwise RVS in text which we will compare to results from Naive RVS. Full results for all methods and all simulation scenarios can be found in Section S.3.C.

The overall quality of prediction M_c of both Stepwise RVS and Naive RVS (from 0.25 to 0.92) was moderately to highly positive except in the Weak Effect-RVs data scenarios (due to very low power) and One Noise-RV data scenarios (due to specificity of either 0 or 1) for both Stepwise and Naive RVS, and in Many Noise-RVs data scenarios using Naive RVS. $RMSE_{diff}$, the difference between $RMSE_{full}$ and $RMSE_{reduced}$, was not large and did not differ much between the RVS-protocols, indicating that removing RVs from the model with RV-selection did not influence model predictions for RVs kept in the model much. In the data scenarios with Pyrifos-like treatment effect, $RMSE_{diff}$ ranged from -0.142 to 0.066 and in the data scenarios with increased treatment effects (such as Pyrifos-like^{+}) $RMSE_{diff}$ ranged from -0.341 to 0.068.

Comparing mean M_c within the same simulation scenario, M_c of Stepwise RVS was higher than Naive RVS in all but 5 out of 36 simulation scenarios (dif-

ference from -0.05 to 0.25, mean=0.05). The main difference in performance of both methods lies in the trade-off between specificity and sensitivity. Stepwise RVS was more successful than Naive RVS in identifying the vast majority of Noise-RVs, as judged from the mean specificity results per simulation scenario. Mean specificity of Stepwise RVS was consistently high (from 0.87 to 0.95) and its 95% confidence interval included 0.95 in 23 out of 36 simulation scenarios whereas mean specificity of Naive RVS was highly varying (from 0.37 to 1) and its 95% confidence interval never included 0.95. For both Stepwise RVS and Naive RVS, mean specificity approached 0.95 more closely with increasing power. In Stepwise RVS, the 95% confidence interval included 0.95 more often in data scenarios with larger treatment effect (16 out of 18) than in data scenarios with Pyrifos-like treatment effect (7 out of 18). For Naive RVS, mean specificity of scenarios with was higher than of scenarios without larger treatment effects (*e.g.* compare Pyrifos^{+} to Pyrifos-like), the difference ranged from 0.08 to 0.43 (mean 0.31). Mean specificity also increased with increasing sample size (difference between Pyrifos-like, More Ditches, and Most Ditches data scenarios; Figure S.3.C.1). Mean sensitivity is highly variable for both Stepwise (from 0.17 to 0.97) and Naive RVS (from 0.35 to 0.95). For Stepwise RVS, mean sensitivity increases when the analysis has more power (due to larger treatment effects or increased sample size). Such a straightforward relationship could not be found for Naive RVS. Mean sensitivity between simulation scenarios with and without larger treatment effects did not increase in all cases and was not clearly affected by increasing the sample size.

Standardization rather than only Centering did not affect results of Stepwise RVS regarding specificity (difference -0.06 to 0.0006) and sensitivity (from -0.006 to 0.017) to great extent. For Naive RVS, Standardization in addition to Centering resulted in lower mean specificity (from -0.02 to -0.25; mean -0.10) and higher mean sensitivity (from 0.01 to 0.32; mean 0.11).

Results of Stepwise RVS are more robust to changes in the composition of the set of RVs than results of Naive RVS. Mean specificity and sensitivity changed less than 0.05 point after adding additional strong Effect-RVs to the Pyrifos-like data set (Strong Effect-RVs; Figure 3.2) and after removing or adding Noise-RVs (One Noise-RV and Many Noise-RVs; Figure S.3.C.2). Note that we calculated specificity and sensitivity of the Strong Effect-RVs data scenario without including results on the additional strong Effect-RVs as to better compare results to the Pyrifos-like data scenario. Using Naive RVS, specificity increased and sensitivity decreased comparing Pyrifos-like to Strong Effect-RVs simulations scenarios. Comparing the One Noise-RV to the Many Noise-RVs data scenario, specificity decreased and sensitivity slightly increased. These changes are smaller when using Standardization in addition to Centering.

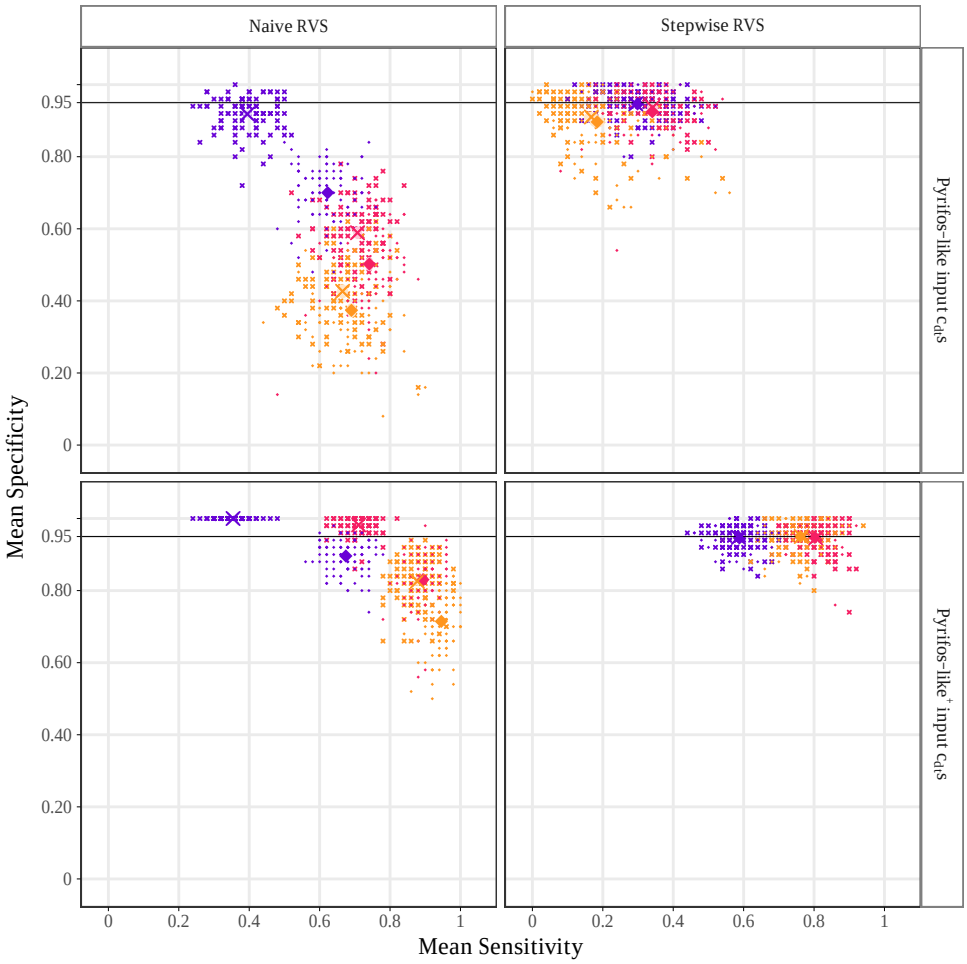


Figure 3.2: Specificity and Sensitivity of Naive and Stepwise RVS when applied to Standardized (points) or Centered (crosses) data generated using the Pyrifos-like/Pyrifos-like$^+$ (• top row/bottom row), Strong Effects RVs/Strong Effects RVs$^+$ (•), and Weak Effects RVs/Weak Effects RVs$^+$ (•) data scenarios. Mean Specificity and Sensitivity over 100 simulations are represented by large symbols, and Specificity and Sensitivity per simulation are represented by small symbols. Ellipses indicate the 95% confidence region of the mean of the estimates. As the confidence regions are small the ellipses are difficult to see

We found that both Stepwise and Naive RVS do not differ in performance between the No Covariance, Pyrifos-like, and More Covariance data scenarios (Figure S.3.C.3). This indicates that covariance in the residuals is not reflected in the b_k-estimates which confirms that PRC deals with this issue well.

3.3.2. Case-study

Stepwise RVS on the Pyrifos data reduced the set of RVs from 178 to 38 species (Figure 3.4). The shape of the Principal Response Curves was mildly affected (Figure 3.3). In general, the shape after RVS seems slightly smoother and the unexpected W-shape around Time=2 of the 6μg/L dose before RVS has disappeared.

When scaling such that mean square of b_k is 1, species with an absolute b_k-estimate over 0.5 in the full-model were more likely to be in the reduced model (26 out of 50; 52%) than species with an absolute b_k-estimate under 0.5 (12 out of 128; 9.4%).

3.4. Discussion

The main reason to apply response variable selection (RVS) in PRC is to be able to distinguish between those species that do follow the principal response and those that do not. Standard PRC usually gives small coefficients to species of the latter group. By setting these coefficients actually to zero, that is, by removing these species, the noise in the data caused by these species is removed from the estimation of the principal response curves. The result is a better estimate of the true response when there were many noise variables and as visibly suggested in the case study where the response curves were smoother after RVS.

One may argue that PRC after selection of response variables is a PRC of a subset of the species only and no longer the PRC of the whole community. We argue that it is still the PRC of the whole community, but one in which non-responding species received a zero coefficient.This differential weighing of species was already an advantage of PRC over similarity analysis (Brink and Braak 1998), but is an even bigger advantage in PRC with Stepwise RVS.

We found no differences in performance between the Two Step, Screening, and Stepwise RVS protocols. In Two Step RVS, RVs were removed from the model one at a time, based on permutation p-values that were recalculated every time an RV was removed from the model. In Screening RVS, permutation p-values were calculated once for every RV using the full model. As Two Step RVS did not yield better results than Screening RVS, we concluded that calculating permutation p-values based on models with increasingly less Noise-RVs did not enhance performance. This conclusion was supported by the finding that adding additional Noise-RVs to or removing Noise-RVs from the data did not affect specificity and sensitivity of permutation RVS-protocols.

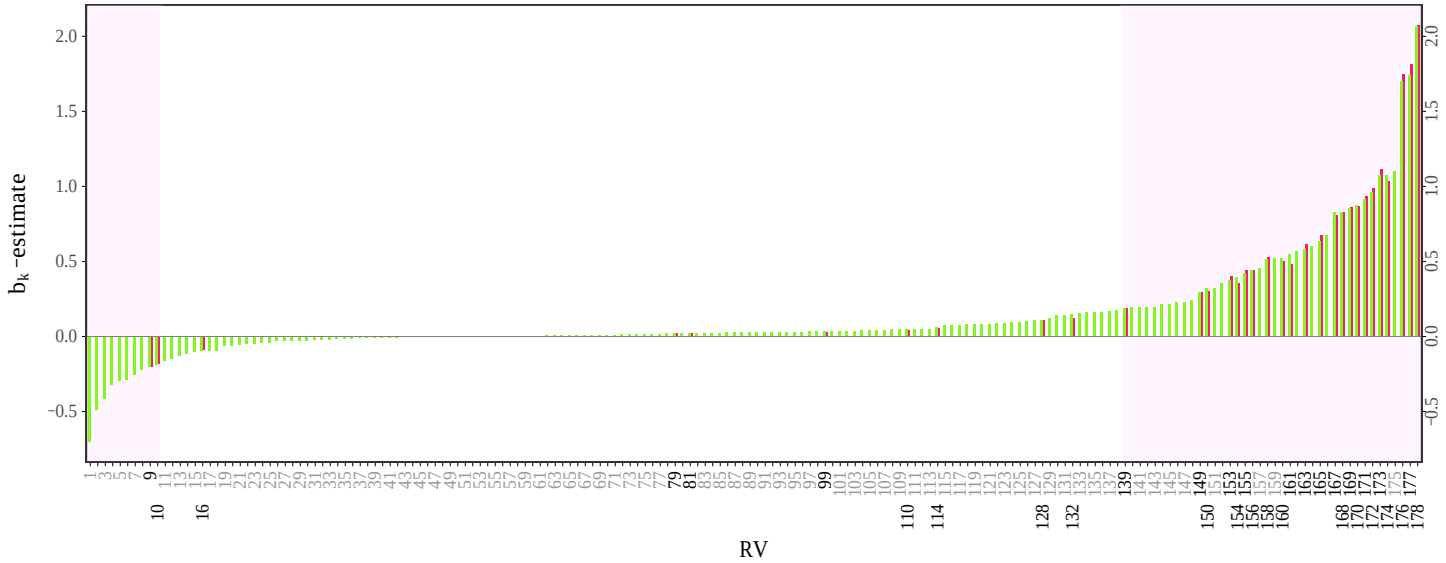


Figure 3.4: b_k-Estimates for the Pyrifos data set (Brink and ter Braak 1999) before (●) and after RV-selection using Stepwise RVS (●) (scaled such that mean square of $\tilde{c}_{dt}$s is 1). Shaded areas represent which RV would be kept when using Naive RVS (with the appropriate scaling). Index numbers of the species are printed in black if kept and printed in grey if removed from the model The kept RV are 9: Stylaria lacustris; 10: Monommata longiseta; 16: Sigara spec.; 79: Alonella nana; 81: Athripsodes aterrimus; 99: Grabtoleberis testudinaria; 110: Pleuroxus aduncus; 114: Alona costata; 128: Alona rectangula; 132: Alona affinis; 139: Oecetes furva; 149: Corynoneura scutellata ; 150: Ciliaten (vnl Halteria sp); 153: Hydracarina; 154: Sialis lutaria; 155: Copepoda spp; 156: Cloeon simile; 158: Copepodiet; 160: Daphnia longispina; 161: Gammarus pulex; 163: Coenagrionidae; 165: Hygrotus versicolor; 167: Ostracoda spp; 168: Hygrotus inaequalis; 169: Caenis luctuosa; 170: Chaoborus obscuripes; 171: Ceratopogonidae; 172: Simocephalus vetulus; 173: Ablabesmyia phatta/monilis; 174: Mystacides longicornis/nigra; 176: Cloeon dipterum; 177: Nauplius Larvae; 178: Caenis horaria. A fully annotated version of this Figure is available via: `http://rdcu.be/JXGP`

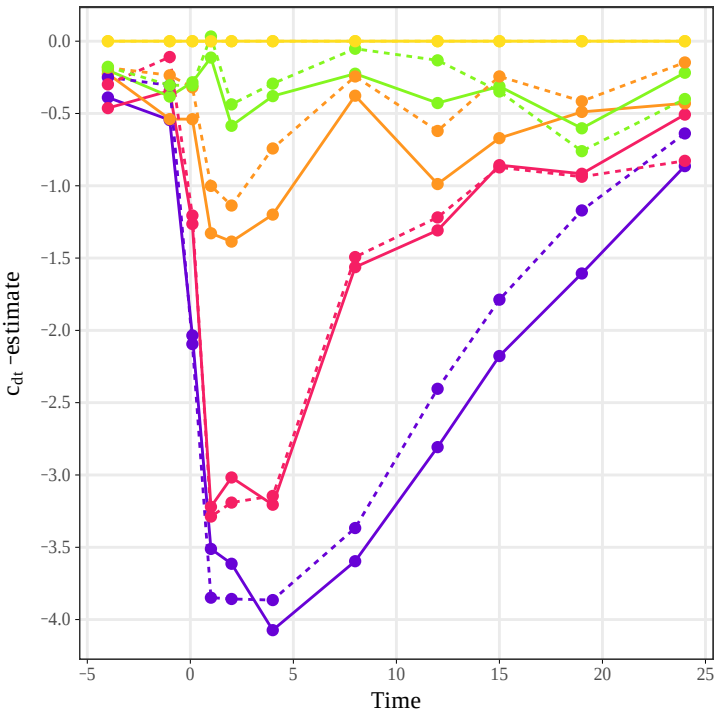


Figure 3.3: Principal Response Curves for the Pyrifos data (Brink and ter Braak 1999) before (solid line) and after RV-selection using Stepwise RVS (dashed line) for the different doses of Chlorpyrifos (0: ●, 0.1: ●, 0.9: ●, 6: ●, and 44μg/L: ●) scaled such that mean square of $\tilde{c}_{dt}$s is 1. Note that the shape of the PRC before RV-selection is identical to the shape in Figure 3.1

We concluded that permutation p-values of RVs were independent of other RVs in the data set because the performance of Screening RVS did not differ from the other protocols. Furthermore, we found that adding additional residual covariance did not affect the quality of b_k-estimates. So we confirmed that PRC is robust against between-species covariance in the residual, even though residual covariance between species is ignored in estimating the PRC coefficients as PRC uses simple least-squares. This is in contrast to what we would expect when selecting predictors rather than RV, such as in multiple regression. In that situation one would expect coefficients, and thus their p-values, and model predictions to be altered as a result of selection.

Performance of Stepwise RVS did not differ from performance of Screening RVS except for being computationally less intensive. It is less intensive, as calculating permutation p-values is faster in data sets with a smaller number of RVs, and Stepwise RVS calculates permutation p-values using an ever smaller set of RVs. The order of deleted RVs was determined based on estimated b_k, which is

a reasonable indicator of effect-size. The Stepwise Stop RVS-protocol was computationally even less intensive than Stepwise RVS. This method however does not meet the goal of 0.95 specificity. Therefore Stepwise RVS was selected as the preferred permutation RVS protocol.

Stepwise RVS combined a stable high specificity with a sensitivity that increased with power. Its performance was unaffected by the number of Noise-RVs in the data set, additional covariance in the residuals, adding additional strong Effect-RVs, and the choice of centering or standardization of the data. In contrast, Naive RVS was highly variable in specificity and sensitivity and was affected by number of Noise-RVs in the data set and adding additional strong Effect-RVs. Because true b_k of RVs in data from practice are unknown, so is the performance of Naive RVS in terms of specificity and sensitivity. We therefore advise Stepwise RVS as the preferred method for RVS in PRC over Naive RVS. We see Stepwise RVS in PRC as an easy applicable and interpretable tool to enhance the insight in the response to treatment of a community over time.

References

Andersen, Roxane, Laurent Grasset, Markus N. Thormann, Line Rochefort, and André-Jean Francez. 2010. "Changes in microbial community structure and function following Sphagnum peatland restoration." *Soil Biol. Biochem.* 42, no. 2 (February): 291–301.

Bray, J. Roger, and J. T. Curtis. 1957. "An Ordination of the Upland Forest Communities of Southern Wisconsin." *Ecol. Monogr.* 27, no. 4 (February): 325–349.

Brink, Paul J. van den, and Cajo J F ter Braak. 1998. "Multivariate analysis of stress in experimental ecosystems by principal response curves and similarity analysis." *Aquat. Ecol.* 32 (2): 163–178.

Brink, Paul J. van den, and Cajo J. F. ter Braak. 1999. "Principal response curves: Analysis of time-dependent multivariate responses of biological community to stress." 18, no. 2 (February): 138–148.

Brink, Paul J. van den, René P.A. van Wijngaarden, Wil G.H. Lucassen, Theo C.M. Brock, and Peter Leeuwangh. 1996. "Effects of the insecticide Dursban® 4E (active ingredient chlorpyrifos) in outdoor experimental ditches: II. Invertebrate community responses and recovery." *Environ. Toxicol. Chem.* 15 (7): 1143–1153.

Britton, A. J., and J. M. Fisher. 2007. "Interactive effects of nitrogen deposition, fire and grazing on diversity and composition of low-alpine prostrate Calluna vulgaris heathland." *J. Appl. Ecol.* 44, no. 1 (November): 125–135.

Cardoso, P.G., D. Raffaelli, A.I. Lillebø, T. Verdelhos, and M.A. Pardal. 2008. "The impact of extreme flooding events and anthropogenic stressors on the macrobenthic communities' dynamics." *Estuar. Coast. Shelf Sci.* 76, no. 3 (February): 553–565.

Cuppen, Jan G.M., Paul J. van den Brink, Edith Camps, Kristiaan F Uil, and Theo C.M. Brock. 2000. "Impact of the fungicide carbendazim in freshwater microcosms. I. Water quality, breakdown of particulate organic matter and responses of macroinvertebrates." *Aquat. Toxicol.* 48, nos. 2-3 (March): 233–250.

Davies, P. T., and M. K-S. Tso. 1982. "Procedures for Reduced-Rank Regression." *Appl. Stat.* 31 (3): 244.

Duarte, Sofia, Cláudia Pascoal, Artur Alves, António Correia, and Fernanda Cássio. 2008. "Copper and zinc mixtures induce shifts in microbial communities and reduce leaf litter decomposition in streams." *Freshw. Biol.* 53, no. 1 (September): 91–101.

Fuentes, Susana, Els van Nood, Sebastian Tims, Ineke Heikamp-de Jong, Cajo J F ter Braak, Josbert J Keller, Erwin G Zoetendal, and Willem M de Vos. 2014. "Reset of a critically disturbed microbial ecosystem: faecal transplant in recurrent Clostridium difficile infection." 8, no. 8 (August): 1621–33.

Hartgers, Elizabeth M., G. H. Aalderink, Paul J. van den Brink, Ronald Gylstra, J. Wilfred F Wiegman, and Theo C M Brock. 1998. "Ecotoxicological threshold levels of a mixture of herbicides (atrazine, diuron and metolachlor) in freshwater microcosms." *Aquat. Ecol.* 32 (2): 135–152.

Heegaard, Einar, and Vigdis Vandvik. 2004. "Climate change affects the outcome of competitive interactions?an application of principal response curves." *Oecologia* 139, no. 3 (May): 459–466.

Kohler, Florian, François Gillet, Jean-michel Gobat, Alexandre Buttler, Florian Kohler, Francois Gillet, and Jean-michel Gobat. 2006. "Effect of Cattle Activities on Gap Colonization in Mountain Pastures." *Folia Geobot.* 41 (3): 289–304.

Matthews, B.W. W. 1975. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." *Biochim. Biophys. Acta - Protein Struct.* 405, no. 2 (October): 442–451.

Moser, Thomas, Jörg Römbke, Hans-Joachim Schallnass, and Cornelis a M van Gestel. 2007. "The use of the multivariate Principal Response Curve (PRC) for community level analysis: a case study on the effects of carbendazim on enchytraeids in Terrestrial Model Ecosystems (TME)." *Ecotoxicology* 16, no. 8 (November): 573–83.

Oksanen, Jari, F Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R Minchin, R B O'Hara, Gavin L Simpson, Peter Solymos, M Henry H Stevens, and Helene Wagner. 2015. *vegan: Community Ecology Package.*

Pakeman, Robin J. 2004. "Consistency of plant species and trait responses to grazing along a productivity gradient: A multi-site analysis." *J. Ecol.* 92, no. 5 (October): 893–905.

Roessink, I, S J H Crum, F Bransen, E van Leeuwen, F van Kerkum, a a Koelmans, and T C M Brock. 2006. "Impact of triphenyltin acetate in microcosms simulating floodplain lakes. I. Influence of sediment quality." *Ecotoxicology* 15, no. 3 (April): 267–93.

Smilauer, Petr, and Jan Lepš. 2014. *Multivariate Analysis of Ecological Data using CANOCO 5.* 2nd Editio. 373. Cambridge University Press.

Smilde, A. K., M. E. Timmerman, M. M. W. B. Hendriks, J. J. Jansen, and H. C. J. Hoefsloot. 2012. "Generic framework for high-dimensional fixed-effects ANOVA." *Brief. Bioinform.* 13 (5): 524–535.

Timmerman, Marieke E., and Cajo J.F. ter Braak. 2008. "Bootstrap confidence intervals for principal response curves." *Comput. Stat. Data Anal.* 52, no. 4 (January): 1837–1849.

Verdonschot, Ralf C. M, Agata M. van Oosten-Siedlecka, Cajo J F ter Braak, and Piet F M Verdonschot. 2015. "Macroinvertebrate survival during cessation of flow and streambed drying in a lowland stream." *Freshw. Biol.* 60, no. 2 (February): 282–296.

Wijngaarden, René P. A. van, Paul J. van den Brink, Steven J. H. Crum, Theo C. M. Brock, Peter Leeuwangh, and Jan H. Oude Voshaar. 1996. "Effects of the insecticide dursban® 4E (active ingredient chlorpyrifos) in outdoor experimental ditches: I. Comparison of short-term toxicity between the laboratory and the field." *Environ. Toxicol. Chem.* 15, no. 7 (July): 1133–1142.

S.3.A. Effect of coefficient scaling, standardization, and composition of data set on b_k-estimates

S.3.A.1. Introduction

In this section, we explain and illustrate how b_k-estimates were influenced by the coefficient scaling used, type of standardization applied to the data set, and number of other RVs present in the data set.

For illustration, data sets were generated similar to the procedure described in in Electronic Supplement 2 using eight data scenarios (Table S.3.A.1). For every data scenario described, we generated one data set using the input c_{dt} such as in Electronic Supplement 2 and four others with smaller and larger treatment effects by multiplying the input c_{dt} by a factor ranging from 0.01 to 16.

S.3.A.2. Coefficient Scaling

When scaling coefficients such that mean squares of b_ks are 1, the estimates of b_k will range roughly from -3 to 3 regardless of the size of treatment effect (Figure S.3.A.1, bottom row). In this situation, using Naive RVS with a cut-off value of 0.5 is defensible. When scaling coefficients such that mean square of centered, and with unbalanced data weighted, c_{dt}s is 1, the estimates of b_k increased with rising input c_{dt}s (Figure S.3.A.1, top row). Whichever scaling method was chosen, larger input c_{dt}s resulted in a better separation between RVs with lower and higher input b_k.

S.3.A.3. Standardization of data set

Data can be centered or also standardized to have a standard deviation of 1 per RV before applying PRC. Note that standardizing all RVs together would have no effect on results of PRC because coefficients are scaled. When data was centered, RVs with higher input b_k had higher b_k-estimates (Figure S.3.A.1, left column). When data was standardized, Noise-RVs had b_k-estimates around zero whereas Effect-RVs had b_k-estimates further away from zero. The b_k-estimates of Effect-RVs with higher and lower input b_ks did systematically differ (Figure S.3.A.1, right column).

Table S.3.A.1: Overview of data scenarios in the simulation study

Data scenario	Response Variables
Pyrifos-like	50 Noise-RVs and 50 Effect-RVs ($13 \cdot b_k = -1$; $13 \cdot b_k = 1$; $12 \cdot b_k = 2$; $12 \cdot b_k = 3$)
No Effect-RVs	50 Noise-RVs
Few Effect-RVs	50 Noise-RVs and 4 Effect-RVs ($1 \cdot b_k = -1$; $1 \cdot b_k = 1$; $1 \cdot b_k = 2$; $1 \cdot b_k = 3$)
No Noise-RVs	as Pyrifos-like, without Noise-RVs
Weak Effect-RVs	50 Noise-RVs and 50 Effect-RVs ($13 \cdot b_k = -1$; $37 \cdot b_k = 1$)
One Strong Effect-RV	as No Noise-RVs, with 1 additional strong Effect-RV ($b_k = 10$)
Strong Effect-RVs	as No Noise-RVs, with 15 additional strong Effect-RVs ($b_k = 10$)

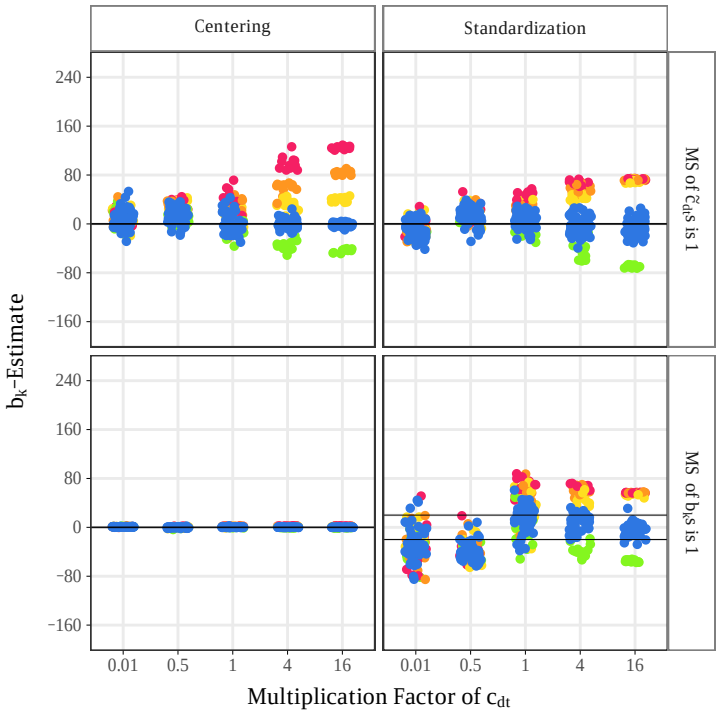


Figure S.3.A.1: Plots of estimated b_ks colored by input b_k (-1: ●, 0: ●, 1: ●, 2: ●, and 3: ●) obtained from PRC on Pyrifos-like data sets with varying input c_{dt}s. Either Centering or Standardization was applied (columns) and mean square of either $\tilde{c}_{dt}$s or b_ks was set to 1 (rows). In the latter situation, a cut-off value of 0.5 is indicated by a black line to simulate Naive RVS

S.3.A.4. Other RVs in the data set

The selection of RVs included in the data set may affect the b_k-estimates. In this section, the difference in b_k-estimates between data sets with different sets of input b_k is illustrated. When scaling such that mean squares of b_k are 1, b_k-estimates ranged between -3 and 3, even when all RVs were Noise-RVs (Figure S.3.A.2, No Effect-RVs). Adding a small number of Effect-RVs to a data set with sufficiently large treatment effect (Figure S.3.A.2, Few Effect-RVs) resulted in much smaller estimated b_k for the Noise-RVs than in a data set with only Noise-RVs. Comparing a data set without Noise-RVs (Figure S.3.A.2, No Noise-RVs) to the same data set with 50 Noise-RVs (Figure S.3.A.1, Pyrifos-like), the data set without Noise-RVs had slightly higher b_k-estimates. Adding only 1 Effect-RV with a much higher input b_k (Figure S.3.A.2, One Strong Effect-RV) to the data set resulted in lower b_k-estimates for the other RVs. This effect was larger when more Effect-RVs with a higher input b_k were added (Figure S.3.A.2, Strong Effect-RVs). The performance of Naive RVS varied greatly with the changing input b_k. When no Effect-RVs were in the data set many Noise-RVs had an estimated b_k above the threshold and when strong Effect-RVs were present many Effect-RVs had estimated b_k below the threshold. Because this issue is inherent to the scaling, which forces the mean square of estimated b_ks to be 1 even when all input b_k would in fact be 0, it cannot be solved by estimating parameters more accurately, and thus not by increasing power (*e.g.* increasing sample size).

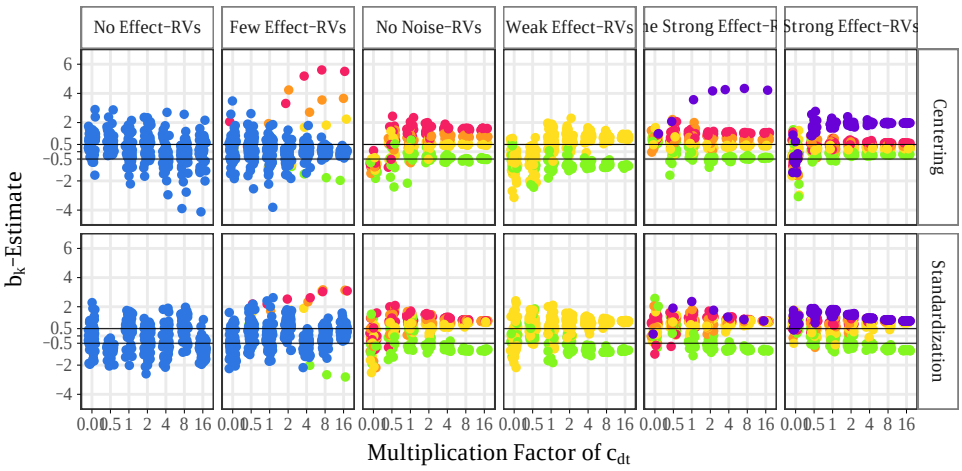


Figure S.3.A.2: Plots of estimated b_ks colored by input b_k (-1: ●, 0: ●, 1: ●, 2: ●, 3: ●, and 10: ●) obtained from PRC on several different data sets (columns; see Table S.3.A.1) with varying input c_{dt}s. Either Centering or Standardization was applied (rows) and mean squares of b_k were set to 1. A cut-off value of 0.5 is indicated by a black line to show effect of Naive RVS

S.3.B. Data Generation

The simulation experiment used in this paper is a simplified version of the Pyrifos experiment, explained in detail by Wijngaarden et al. (1996) and Brink et al. (1996). We simulated three treatment levels, each applied to the same number of ditches. Abundances were recorded on five occasions and only after the treatment was applied.

Sensible estimates for effect-size and covariance structure were based on results of PRC performed on a subset of the Pyrifos data set containing only the 21 RVs with less than 40% zeros. The structural part of the simulated data consisted of the sum of a_{tk} and the product of c_{dt} and b_k. As input a_{tk}, we used a random sample from the predicted values in the control group. As input c_{dt} we used the c_{dt} estimates of the first five sample times after treatment of the second and highest dose. As input b_k for the effect-RVs, we used -1, 1, 2, and 3 which reflected the range of estimated b_ks.

The random part of the data was estimated using a multivariate normal distribution which requires an input-covariance structure. To obtain Pyrifos-like covariance structure estimates, we first calculated residuals by subtracting fitted values from observed values. For *e.g.* species covariance structure, we estimated covariance between residuals obtained at the same time in the same ditch for the different species; and subtracted the mean covariance for that combination of time and ditch. Covariance structures for time and ditches were calculated in a similar fashion. As we assume ditches are independent, the variance-covariance matrix for ditches contained only variance estimated on the diagonal. Based on the observations, the variance-covariance matrix for time was assumed to have a first-order heterogeneous structure with ρ=0.4. We obtained only 21 estimates for species covariance-structures whilst we needed many more for the simulation study. To ensure that the variance-covariance matrix of the simulated data set, like that of the Pyrifos data set, contained clusters of species, we assigned a parent-species from the Pyrifos data set to every species in the simulation data set. The correlation structure of the species is the same as that of the parent-species. For species with the same parent species, we set the correlation to 125% of the maximum correlation observed between species. In the covariance-matrix constructed from the correlation matrix, diagonals were increased by 40% to reduce probability of the matrix not being positive definite. The make.positive.definite function in the corpcor package (Schäfer et al. 2015) was used to ensure all matrices were positive definite. Variance-Covariance matrices for species and time were combined into one data set Variance-Covariance matrix and used to generate error estimates per ditch (mvrnorm function in the MASS package (Venables and Ripley 2002)).

The range of the data simulated using the method described above was wider than that in the Pyrifos data set. Therefore, the estimates were multiplied by a factor to have the same maximum value as in the Pyrifos data set. Thereafter, the estimates were transformed back to abundance scale ($0.1 * (\exp(x))$) and used as expected values for a Poisson distribution.

References

Brink, Paul J. van den, René P.A. van Wijngaarden, Wil G.H. Lucassen, Theo C.M. Brock, and Peter Leeuwangh. 1996. "Effects of the insecticide Dursban® 4E (active ingredient chlorpyrifos) in outdoor experimental ditches: II. Invertebrate community responses and recovery." *Environ. Toxicol. Chem.* 15 (7): 1143–1153.

Schäfer, Juliane, Rainer Opgen-Rhein, Verena Zuber, Miika Ahdesmäki, A Pedro Duarte Silva, and Korbinian Strimmer. 2015. *corpcor: Efficient Estimation of Covariance and (Partial) Correlation.*

Venables, W N, and B D Ripley. 2002. *Modern Applied Statistics with S.* Fourth. New York: Springer.

Wijngaarden, René P. A. van, Paul J. van den Brink, Steven J. H. Crum, Theo C. M. Brock, Peter Leeuwangh, and Jan H. Oude Voshaar. 1996. "Effects of the insecticide dursban® 4E (active ingredient chlorpyrifos) in outdoor experimental ditches: I. Comparison of short-term toxicity between the laboratory and the field." *Environ. Toxicol. Chem.* 15, no. 7 (July): 1133–1142.

S.3.C. Supplementary Results

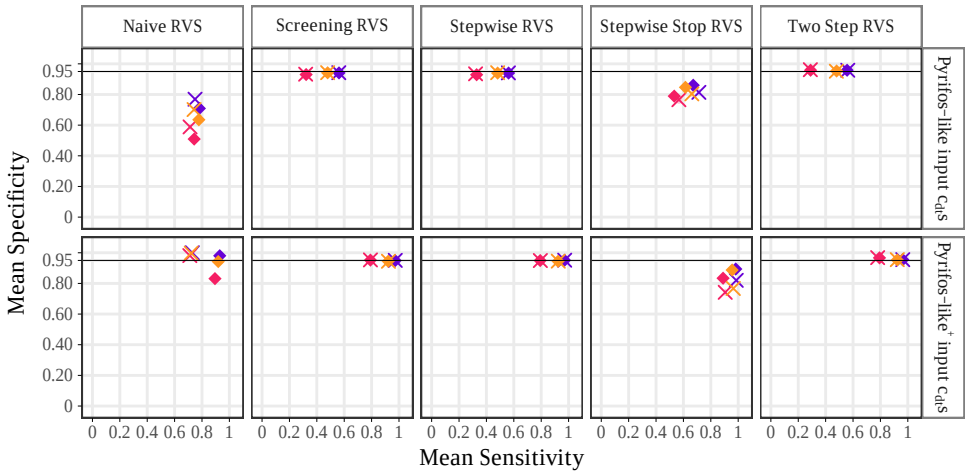


Figure S.3.C.1: Mean Specificity and mean Sensitivity of Naive, Screening, Stepwise, Stepwise Stop, and Two Step RVS when applied to Standardized (points) or Centered (crosses) data generated using the Pyrifos-like/Pyrifos-like$^+$ (• top row/bottom row), More Ditches/More Ditches$^+$ (•), and Most Ditches/Most Ditches$^+$ (•) data scenarios. Ellipses indicate the 95% confidence region of the mean of the estimates. As the confidence regions are small the ellipses are difficult to see

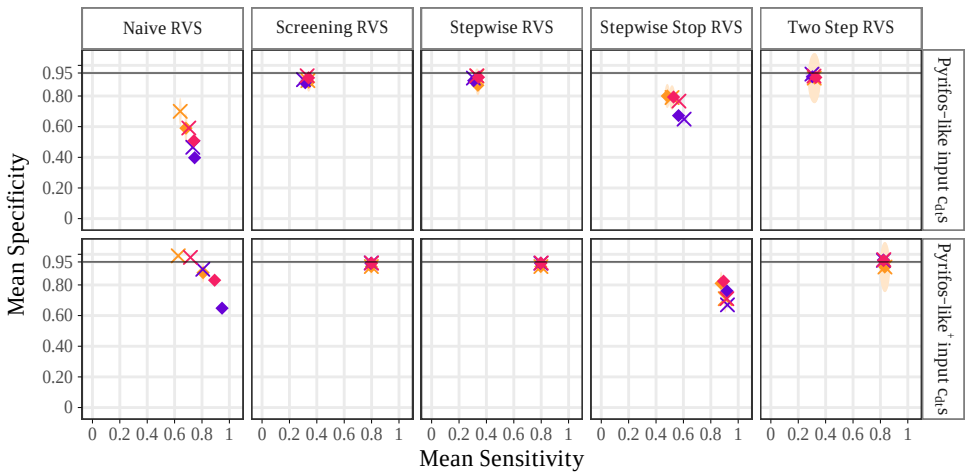


Figure S.3.C.2: Mean Specificity and mean Sensitivity of Naive, Screening, Stepwise, Stepwise Stop, and Two Step RVS when applied to Standardized (points) or Centered (crosses) data generated using the Pyrifos-like/Pyrifos-like$^+$ (• top row/bottom row), One Noise-RV/One Noise-RV$^+$ (•), and Many Noise-RVs/Many Noise-RVs$^+$ (•) data scenarios. Ellipses indicate the 95% confidence region of the mean of the estimates. As the confidence regions are small the ellipses are difficult to see

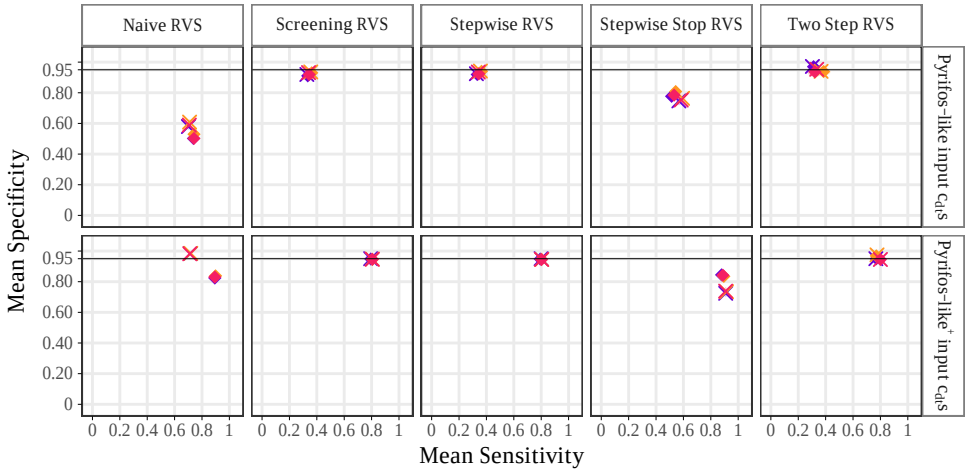


Figure S.3.C.3: Mean Specificity and mean Sensitivity of Naive, Screening, Stepwise, Stepwise Stop, and Two Step RVS when applied to Standardized (points) or Centered (crosses) data generated using the Pyrifos-like/Pyrifos-like$^+$ (● top row/bottom row), No Covariance/No Covariance$^+$ (●), and More Covariance/More Covariance$^+$ (●) data scenarios. Ellipses indicate the 95% confidence region of the mean of the estimates. As the confidence regions are small the ellipses are difficult to see

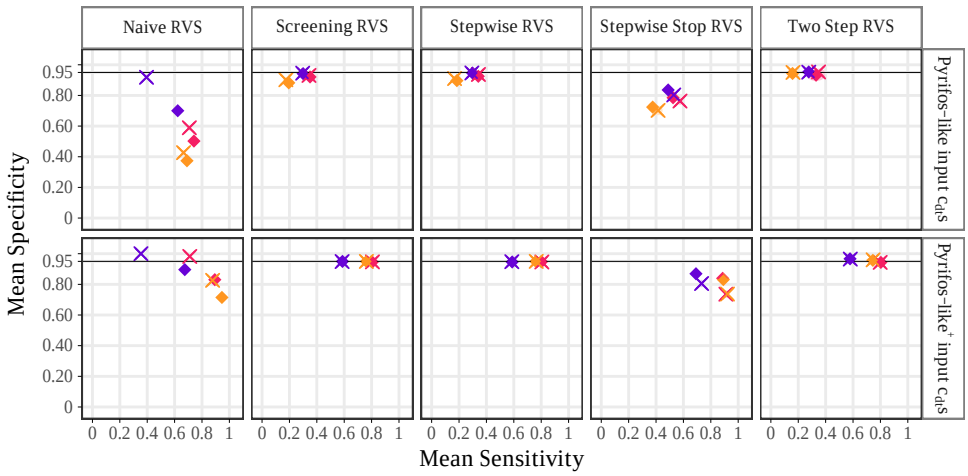


Figure S.3.C.4: Mean Specificity and mean Sensitivity of Naive, Screening, Stepwise, Stepwise Stop, and Two Step RVS when applied to Standardized (points) or Centered (crosses) data generated using the Pyrifos-like/Pyrifos-like$^+$ (● top row/bottom row), Strong Effects RVs/Strong Effects RVs$^+$ (●), and Weak Effects RVs/Weak Effects RVs$^+$ (●) data scenarios. Ellipses indicate the 95% confidence region of the mean of the estimates. As the confidence regions are small the ellipses are difficult to see

Table 3.1: Mean ±1.96·SEM specificity, sensitivity, M_c, and $RMSE_{diff}$ of the RVS-protocols on the different data scenarios using Centering

Data Scenario	RVS-protocol	Specificity	Sensitivity	M_c	$RMSE_{diff}$
Pyrifos-like	Naive RVS	0.59±0.02	0.71±0.01	0.30±0.01	0.03±0.03
Pyrifos-like	Screening RVS	0.93±0.01	0.34±0.01	0.34±0.01	0.03±0.02
Pyrifos-like	Stepwise RVS	0.94±0.01	0.34±0.01	0.35±0.01	0.03±0.02
Pyrifos-like	Stepwise Stop RVS	0.76±0.02	0.57±0.01	0.35±0.01	-0.02±0.03
Pyrifos-like	Two Step RVS	0.95±0.02	0.35±0.02	0.38±0.02	0.07±0.03
More Ditches	Naive RVS	0.70±0.02	0.74±0.01	0.45±0.01	0.00±0.02
More Ditches	Screening RVS	0.94±0.01	0.48±0.01	0.48±0.01	0.06±0.02
More Ditches	Stepwise RVS	0.94±0.01	0.48±0.01	0.48±0.01	0.06±0.02
More Ditches	Stepwise Stop RVS	0.80±0.02	0.66±0.01	0.48±0.01	-0.04±0.03
More Ditches	Two Step RVS	0.95±0.02	0.48±0.02	0.49±0.02	0.09±0.05
Most Ditches	Naive RVS	0.77±0.01	0.75±0.01	0.52±0.01	-0.02±0.02
Most Ditches	Screening RVS	0.94±0.01	0.56±0.01	0.55±0.01	0.05±0.02
Most Ditches	Stepwise RVS	0.94±0.01	0.56±0.01	0.55±0.01	0.05±0.02
Most Ditches	Stepwise Stop RVS	0.81±0.02	0.71±0.01	0.54±0.01	-0.07±0.02
Most Ditches	Two Step RVS	0.96±0.02	0.56±0.02	0.57±0.02	0.08±0.08
Weak Effect-RVs	Naive RVS	0.43±0.02	0.67±0.01	0.09±0.01	-0.05±0.03
Weak Effect-RVs	Screening RVS	0.90±0.02	0.17±0.01	0.11±0.01	0.02±0.02
Weak Effect-RVs	Stepwise RVS	0.91±0.01	0.17±0.01	0.11±0.01	0.00±0.02
Weak Effect-RVs	Stepwise Stop RVS	0.70±0.03	0.41±0.02	0.13±0.01	-0.02±0.03
Weak Effect-RVs	Two Step RVS	0.95±0.02	0.16±0.02	0.18±0.04	-0.07±0.06
Strong Effect-RVs	Naive RVS	0.92±0.01	0.39±0.01	0.37±0.01	-0.04±0.02
Strong Effect-RVs	Screening RVS	0.95±0.01	0.30±0.01	0.32±0.01	0.04±0.01
Strong Effect-RVs	Stepwise RVS	0.95±0.01	0.29±0.01	0.32±0.01	0.04±0.01
Strong Effect-RVs	Stepwise Stop RVS	0.80±0.02	0.53±0.01	0.36±0.01	-0.01±0.02

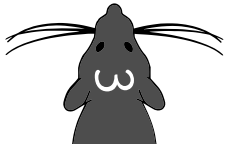
Continued on next page

Data Scenario	RVS-protocol	Specificity	Sensitivity	M_c	$\text{RMSE}_{\text{diff}}$
Strong Effect-RVs	Two Step RVS	0.95±0.02	0.28±0.03	0.31±0.02	0.04±0.03
One Noise-RV	Naive RVS	0.70±0.09	0.64±0.01	0.10±0.01	-0.01±0.03
One Noise-RV	Screening RVS	0.90±0.06	0.33±0.01	0.07±0.01	0.01±0.02
One Noise-RV	Stepwise RVS	0.93±0.05	0.33±0.01	0.08±0.01	0.01±0.02
One Noise-RV	Stepwise Stop RVS	0.79±0.08	0.52±0.01	0.09±0.01	0.00±0.03
One Noise-RV	Two Step RVS	0.92±0.16	0.32±0.02	0.07±0.03	0.01±0.05
Many Noise-RVs	Naive RVS	0.47±0.02	0.73±0.01	0.16±0.01	0.01±0.03
Many Noise-RVs	Screening RVS	0.91±0.01	0.30±0.01	0.26±0.02	0.03±0.02
Many Noise-RVs	Stepwise RVS	0.92±0.01	0.30±0.01	0.28±0.02	0.05±0.02
Many Noise-RVs	Stepwise Stop RVS	0.65±0.03	0.60±0.01	0.22±0.01	-0.01±0.02
Many Noise-RVs	Two Step RVS	0.94±0.02	0.30±0.02	0.32±0.03	0.06±0.04
No Covariance	Naive RVS	0.61±0.02	0.71±0.01	0.32±0.01	-0.07±0.03
No Covariance	Screening RVS	0.94±0.01	0.36±0.01	0.36±0.01	0.03±0.02
No Covariance	Stepwise RVS	0.94±0.01	0.36±0.01	0.37±0.01	0.03±0.02
No Covariance	Stepwise Stop RVS	0.76±0.03	0.59±0.01	0.38±0.01	-0.08±0.03
No Covariance	Two Step RVS	0.94±0.03	0.37±0.02	0.38±0.02	0.06±0.04
More Covariance	Naive RVS	0.58±0.02	0.70±0.01	0.28±0.01	-0.05±0.03
More Covariance	Screening RVS	0.92±0.01	0.33±0.01	0.31±0.01	0.03±0.02
More Covariance	Stepwise RVS	0.92±0.01	0.33±0.01	0.32±0.01	0.03±0.02
More Covariance	Stepwise Stop RVS	0.75±0.03	0.57±0.01	0.33±0.01	-0.07±0.03
More Covariance	Two Step RVS	0.97±0.01	0.30±0.02	0.37±0.02	-0.01±0.04

Table 3.2: Mean ±1.96·SEM specificity, sensitivity, M_c, and $RMSE_{diff}$ of the RVS-protocols on the different data scenarios using Centering

Data Scenario	RVS-protocol	Specificity	Sensitivity	M_c	$RMSE_{diff}$
Pyrifos-like$^+$	Naive RVS	0.98±0.00	0.71±0.00	0.72±0.00	-0.04±0.01
Pyrifos-like$^+$	Screening RVS	0.95±0.01	0.81±0.01	0.76±0.01	0.04±0.01
Pyrifos-like$^+$	Stepwise RVS	0.95±0.01	0.80±0.01	0.76±0.01	0.06±0.01
Pyrifos-like$^+$	Stepwise Stop RVS	0.74±0.03	0.91±0.01	0.66±0.01	0.01±0.01
Pyrifos-like$^+$	Two Step RVS	0.94±0.02	0.80±0.02	0.75±0.01	0.10±0.02
More Ditches$^+$	Naive RVS	1.00±0.00	0.72±0.00	0.75±0.00	-0.02±0.01
More Ditches$^+$	Screening RVS	0.94±0.01	0.92±0.00	0.87±0.00	0.01±0.01
More Ditches$^+$	Stepwise RVS	0.95±0.01	0.92±0.00	0.87±0.00	0.03±0.01
More Ditches$^+$	Stepwise Stop RVS	0.77±0.02	0.96±0.00	0.75±0.01	-0.04±0.01
More Ditches$^+$	Two Step RVS	0.96±0.02	0.92±0.01	0.88±0.01	0.01±0.03
Most Ditches$^+$	Naive RVS	1.00±0.00	0.73±0.00	0.76±0.00	-0.01±0.01
Most Ditches$^+$	Screening RVS	0.95±0.01	0.97±0.00	0.92±0.00	0.04±0.01
Most Ditches$^+$	Stepwise RVS	0.95±0.01	0.97±0.00	0.92±0.00	0.03±0.01
Most Ditches$^+$	Stepwise Stop RVS	0.82±0.02	0.99±0.00	0.82±0.01	-0.03±0.01
Most Ditches$^+$	Two Step RVS	0.96±0.01	0.96±0.01	0.92±0.01	0.03±0.03
Weak Effect-RVs$^+$	Naive RVS	0.83±0.01	0.88±0.00	0.71±0.01	-0.09±0.02
Weak Effect-RVs$^+$	Screening RVS	0.95±0.01	0.76±0.01	0.73±0.01	0.04±0.02
Weak Effect-RVs$^+$	Stepwise RVS	0.95±0.01	0.76±0.01	0.73±0.01	0.05±0.02
Weak Effect-RVs$^+$	Stepwise Stop RVS	0.74±0.03	0.92±0.00	0.68±0.01	-0.07±0.02
Weak Effect-RVs$^+$	Two Step RVS	0.96±0.01	0.75±0.02	0.72±0.02	0.03±0.05
Strong Effect-RVs$^+$	Naive RVS	1.00±0.00	0.35±0.00	0.46±0.00	0.00±0.00
Strong Effect-RVs$^+$	Screening RVS	0.95±0.01	0.58±0.01	0.57±0.01	0.06±0.00
Strong Effect-RVs$^+$	Stepwise RVS	0.95±0.01	0.59±0.00	0.57±0.01	0.06±0.00
Strong Effect-RVs$^+$	Stepwise Stop RVS	0.81±0.02	0.73±0.01	0.55±0.01	0.03±0.00

Continued on next page

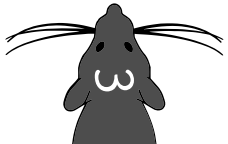


Data Scenario	RVS-protocol	Specificity	Sensitivity	M_c	$\text{RMSE}_{\text{diff}}$
Strong Effect-RVs$^+$	Two Step RVS	0.96±0.02	0.58±0.01	0.59±0.01	0.06±0.01
One Noise-RV$^+$	Naive RVS	0.99±0.02	0.63±0.00	0.18±0.00	-0.02±0.01
One Noise-RV$^+$	Screening RVS	0.92±0.05	0.80±0.01	0.25±0.01	0.02±0.01
One Noise-RV$^+$	Stepwise RVS	0.92±0.05	0.80±0.01	0.25±0.01	0.02±0.01
One Noise-RV$^+$	Stepwise Stop RVS	0.71±0.09	0.91±0.01	0.31±0.02	0.02±0.01
One Noise-RV$^+$	Two Step RVS	0.92±0.16	0.83±0.02	0.28±0.03	-0.01±0.04
Many Noise-RVs$^+$	Naive RVS	0.90±0.01	0.81±0.01	0.67±0.01	-0.02±0.02
Many Noise-RVs$^+$	Screening RVS	0.94±0.01	0.80±0.01	0.74±0.01	0.06±0.01
Many Noise-RVs$^+$	Stepwise RVS	0.94±0.01	0.80±0.01	0.73±0.01	0.07±0.01
Many Noise-RVs$^+$	Stepwise Stop RVS	0.67±0.02	0.92±0.00	0.49±0.01	0.02±0.02
Many Noise-RVs$^+$	Two Step RVS	0.96±0.01	0.82±0.02	0.79±0.02	0.04±0.03
No Covariance$^+$	Naive RVS	0.99±0.00	0.72±0.00	0.73±0.00	-0.05±0.01
No Covariance$^+$	Screening RVS	0.95±0.01	0.80±0.01	0.76±0.00	0.05±0.01
No Covariance$^+$	Stepwise RVS	0.95±0.01	0.80±0.00	0.76±0.00	0.05±0.01
No Covariance$^+$	Stepwise Stop RVS	0.73±0.02	0.91±0.01	0.67±0.01	-0.03±0.02
No Covariance$^+$	Two Step RVS	0.98±0.01	0.78±0.02	0.77±0.01	0.07±0.01
More Covariance$^+$	Naive RVS	0.98±0.00	0.72±0.00	0.73±0.00	-0.02±0.01
More Covariance$^+$	Screening RVS	0.95±0.01	0.79±0.00	0.75±0.01	0.06±0.01
More Covariance$^+$	Stepwise RVS	0.95±0.01	0.80±0.00	0.76±0.01	0.05±0.01
More Covariance$^+$	Stepwise Stop RVS	0.72±0.02	0.91±0.01	0.65±0.01	-0.02±0.01
More Covariance$^+$	Two Step RVS	0.95±0.01	0.77±0.01	0.73±0.02	0.05±0.02

Table 3.3: Mean ±1.96·SEM specificity, sensitivity, M_c, and $RMSE_{diff}$ of the RVS-protocols on the different data scenarios using Standardization

Data Scenario	RVS-protocol	Specificity	Sensitivity	M_c	$RMSE_{diff}$
Pyrifos-like	Naive RVS	0.50±0.02	0.74±0.01	0.25±0.01	-0.06±0.02
Pyrifos-like	Screening RVS	0.92±0.01	0.35±0.01	0.33±0.01	-0.06±0.01
Pyrifos-like	Stepwise RVS	0.92±0.01	0.34±0.01	0.33±0.01	-0.05±0.01
Pyrifos-like	Stepwise Stop RVS	0.78±0.03	0.52±0.01	0.33±0.01	-0.07±0.01
Pyrifos-like	Two Step RVS	0.93±0.04	0.33±0.03	0.32±0.05	-0.04±0.03
More Ditches	Naive RVS	0.64±0.02	0.78±0.01	0.42±0.01	-0.05±0.01
More Ditches	Screening RVS	0.94±0.01	0.48±0.01	0.48±0.01	-0.08±0.01
More Ditches	Stepwise RVS	0.94±0.01	0.48±0.01	0.48±0.01	-0.06±0.01
More Ditches	Stepwise Stop RVS	0.85±0.02	0.61±0.01	0.48±0.01	-0.10±0.01
More Ditches	Two Step RVS	0.95±0.02	0.48±0.01	0.49±0.02	-0.04±0.04
Most Ditches	Naive RVS	0.71±0.02	0.78±0.01	0.49±0.01	-0.09±0.01
Most Ditches	Screening RVS	0.94±0.01	0.56±0.01	0.55±0.01	-0.06±0.01
Most Ditches	Stepwise RVS	0.94±0.01	0.56±0.01	0.55±0.01	-0.08±0.01
Most Ditches	Stepwise Stop RVS	0.86±0.02	0.67±0.01	0.55±0.01	-0.10±0.01
Most Ditches	Two Step RVS	0.96±0.02	0.56±0.02	0.56±0.02	-0.11±0.04
Weak Effect-RVs	Naive RVS	0.37±0.02	0.69±0.01	0.07±0.01	-0.03±0.01
Weak Effect-RVs	Screening RVS	0.88±0.02	0.20±0.01	0.11±0.01	0.00±0.01
Weak Effect-RVs	Stepwise RVS	0.90±0.02	0.18±0.01	0.12±0.01	-0.01±0.01
Weak Effect-RVs	Stepwise Stop RVS	0.72±0.03	0.38±0.02	0.11±0.01	-0.01±0.01
Weak Effect-RVs	Two Step RVS	0.94±0.03	0.16±0.02	0.16±0.04	0.00±0.04
Strong Effect-RVs	Naive RVS	0.70±0.01	0.62±0.01	0.33±0.01	-0.05±0.01
Strong Effect-RVs	Screening RVS	0.94±0.01	0.30±0.01	0.32±0.01	-0.06±0.01
Strong Effect-RVs	Stepwise RVS	0.95±0.01	0.30±0.01	0.32±0.01	-0.07±0.01
Strong Effect-RVs	Stepwise Stop RVS	0.84±0.02	0.49±0.01	0.36±0.01	-0.07±0.01

Continued on next page

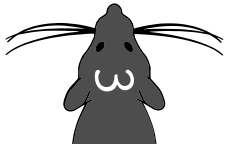


Data Scenario	RVS-protocol	Specificity	Sensitivity	M_c	$\text{RMSE}_{\text{diff}}$
Strong Effect-RVs	Two Step RVS	0.95±0.01	0.28±0.03	0.31±0.03	-0.13±0.03
One Noise-RV	Naive RVS	0.59±0.10	0.68±0.01	0.08±0.02	-0.06±0.02
One Noise-RV	Screening RVS	0.89±0.06	0.34±0.01	0.07±0.01	-0.04±0.01
One Noise-RV	Stepwise RVS	0.87±0.07	0.34±0.01	0.06±0.01	-0.03±0.01
One Noise-RV	Stepwise Stop RVS	0.80±0.08	0.48±0.01	0.08±0.01	-0.05±0.02
One Noise-RV	Two Step RVS	0.92±0.16	0.32±0.02	0.07±0.03	-0.01±0.04
Many Noise-RVs	Naive RVS	0.40±0.01	0.75±0.01	0.12±0.01	-0.01±0.02
Many Noise-RVs	Screening RVS	0.89±0.02	0.32±0.01	0.24±0.02	-0.06±0.01
Many Noise-RVs	Stepwise RVS	0.90±0.02	0.31±0.01	0.25±0.02	-0.03±0.01
Many Noise-RVs	Stepwise Stop RVS	0.67±0.03	0.56±0.01	0.21±0.02	-0.06±0.01
Many Noise-RVs	Two Step RVS	0.93±0.04	0.30±0.02	0.30±0.04	-0.02±0.03
No Covariance	Naive RVS	0.53±0.02	0.74±0.01	0.28±0.01	-0.07±0.01
No Covariance	Screening RVS	0.93±0.01	0.36±0.01	0.36±0.01	-0.09±0.01
No Covariance	Stepwise RVS	0.94±0.01	0.36±0.01	0.37±0.01	-0.08±0.01
No Covariance	Stepwise Stop RVS	0.81±0.02	0.54±0.01	0.38±0.01	-0.10±0.01
No Covariance	Two Step RVS	0.94±0.02	0.38±0.03	0.39±0.02	-0.07±0.04
More Covariance	Naive RVS	0.50±0.02	0.74±0.01	0.24±0.01	-0.05±0.01
More Covariance	Screening RVS	0.91±0.01	0.33±0.01	0.30±0.01	-0.07±0.01
More Covariance	Stepwise RVS	0.92±0.01	0.33±0.01	0.31±0.01	-0.05±0.01
More Covariance	Stepwise Stop RVS	0.78±0.03	0.51±0.01	0.31±0.01	-0.08±0.01
More Covariance	Two Step RVS	0.97±0.02	0.31±0.02	0.37±0.02	-0.08±0.05

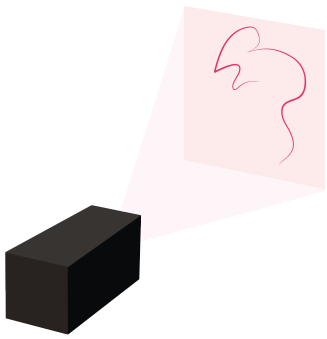
Table 3.4: Mean ±1.96·SEM specificity, sensitivity, M_c, and $RMSE_{diff}$ of the RVS-protocols on the different data scenarios using Standardization

Data Scenario	RVS-protocol	Specificity	Sensitivity	M_c	$RMSE_{diff}$
Pyrifos-like$^+$	Naive RVS	0.83±0.01	0.89±0.00	0.73±0.01	-0.29±0.01
Pyrifos-like$^+$	Screening RVS	0.95±0.01	0.80±0.00	0.76±0.01	-0.28±0.01
Pyrifos-like$^+$	Stepwise RVS	0.95±0.01	0.80±0.01	0.76±0.00	-0.27±0.01
Pyrifos-like$^+$	Stepwise Stop RVS	0.84±0.02	0.89±0.01	0.73±0.01	-0.29±0.01
Pyrifos-like$^+$	Two Step RVS	0.94±0.02	0.80±0.02	0.76±0.01	-0.27±0.05
More Ditches$^+$	Naive RVS	0.94±0.01	0.92±0.00	0.86±0.01	-0.34±0.01
More Ditches$^+$	Screening RVS	0.94±0.01	0.92±0.00	0.87±0.00	-0.31±0.01
More Ditches$^+$	Stepwise RVS	0.94±0.01	0.92±0.00	0.87±0.01	-0.31±0.01
More Ditches$^+$	Stepwise Stop RVS	0.89±0.01	0.96±0.00	0.85±0.01	-0.32±0.01
More Ditches$^+$	Two Step RVS	0.95±0.02	0.92±0.01	0.88±0.01	-0.24±0.04
Most Ditches$^+$	Naive RVS	0.98±0.00	0.93±0.00	0.91±0.00	-0.33±0.01
Most Ditches$^+$	Screening RVS	0.95±0.01	0.97±0.00	0.92±0.00	-0.32±0.01
Most Ditches$^+$	Stepwise RVS	0.95±0.01	0.97±0.00	0.92±0.00	-0.32±0.01
Most Ditches$^+$	Stepwise Stop RVS	0.89±0.01	0.98±0.00	0.88±0.01	-0.33±0.01
Most Ditches$^+$	Two Step RVS	0.96±0.01	0.96±0.01	0.92±0.01	-0.33±0.03
Weak Effect-RVs$^+$	Naive RVS	0.71±0.02	0.95±0.00	0.68±0.01	-0.16±0.01
Weak Effect-RVs$^+$	Screening RVS	0.95±0.01	0.76±0.01	0.73±0.01	-0.12±0.01
Weak Effect-RVs$^+$	Stepwise RVS	0.95±0.01	0.76±0.01	0.73±0.01	-0.11±0.01
Weak Effect-RVs$^+$	Stepwise Stop RVS	0.83±0.02	0.89±0.01	0.73±0.01	-0.16±0.01
Weak Effect-RVs$^+$	Two Step RVS	0.96±0.02	0.74±0.03	0.72±0.03	-0.14±0.04
Strong Effect-RVs$^+$	Naive RVS	0.90±0.01	0.67±0.00	0.59±0.01	-0.22±0.01
Strong Effect-RVs$^+$	Screening RVS	0.95±0.01	0.59±0.01	0.58±0.01	-0.24±0.01
Strong Effect-RVs$^+$	Stepwise RVS	0.95±0.01	0.59±0.00	0.57±0.01	-0.24±0.01
Strong Effect-RVs$^+$	Stepwise Stop RVS	0.87±0.02	0.69±0.01	0.58±0.01	-0.22±0.01

Continued on next page



Data Scenario	RVS-protocol	Specificity	Sensitivity	M_c	$\text{RMSE}_{\text{diff}}$
Strong Effect-RVs$^+$	Two Step RVS	0.97±0.01	0.58±0.01	0.59±0.01	-0.24±0.02
One Noise-RV$^+$	Naive RVS	0.88±0.06	0.81±0.01	0.25±0.01	-0.22±0.02
One Noise-RV$^+$	Screening RVS	0.92±0.05	0.80±0.01	0.25±0.01	-0.20±0.02
One Noise-RV$^+$	Stepwise RVS	0.92±0.05	0.80±0.01	0.25±0.01	-0.21±0.02
One Noise-RV$^+$	Stepwise Stop RVS	0.81±0.08	0.88±0.01	0.29±0.02	-0.25±0.02
One Noise-RV$^+$	Two Step RVS	0.92±0.16	0.83±0.02	0.27±0.03	-0.14±0.04
Many Noise-RVs$^+$	Naive RVS	0.65±0.01	0.95±0.00	0.48±0.01	-0.16±0.01
Many Noise-RVs$^+$	Screening RVS	0.94±0.01	0.79±0.01	0.73±0.01	-0.24±0.01
Many Noise-RVs$^+$	Stepwise RVS	0.94±0.01	0.80±0.01	0.73±0.01	-0.25±0.01
Many Noise-RVs$^+$	Stepwise Stop RVS	0.76±0.02	0.92±0.00	0.57±0.01	-0.20±0.01
Many Noise-RVs$^+$	Two Step RVS	0.96±0.01	0.82±0.02	0.79±0.01	-0.23±0.03
No Covariance$^+$	Naive RVS	0.84±0.01	0.90±0.00	0.74±0.01	-0.30±0.01
No Covariance$^+$	Screening RVS	0.95±0.01	0.81±0.01	0.76±0.01	-0.30±0.01
No Covariance$^+$	Stepwise RVS	0.95±0.01	0.80±0.00	0.76±0.00	-0.29±0.01
No Covariance$^+$	Stepwise Stop RVS	0.83±0.02	0.90±0.00	0.74±0.01	-0.29±0.01
No Covariance$^+$	Two Step RVS	0.97±0.01	0.77±0.02	0.75±0.01	-0.23±0.06
More Covariance$^+$	Naive RVS	0.82±0.01	0.89±0.00	0.72±0.01	-0.29±0.01
More Covariance$^+$	Screening RVS	0.95±0.01	0.79±0.00	0.75±0.01	-0.29±0.01
More Covariance$^+$	Stepwise RVS	0.95±0.01	0.79±0.00	0.75±0.01	-0.30±0.01
More Covariance$^+$	Stepwise Stop RVS	0.84±0.02	0.88±0.01	0.73±0.01	-0.27±0.01
More Covariance$^+$	Two Step RVS	0.95±0.01	0.77±0.01	0.73±0.02	-0.24±0.04



4

Relating ultrasonic vocalisations from a pair of rats to individual behaviour: a Composite Link Model approach

Nadia J. Vendrig

Lia Hemerik

Ilona Pinter[1]

Cajo J.F. ter Braak

Submitted to Statistica Neerlandica (in revision)

1. Delta Phenomics BV, Schaijk, The Netherlands

Ultrasonic Vocalisations (USVs) are crucial in social behaviour of rats. We aim to relate USV-rates of pairs of rats to individual activity in an automated home-cage (PhenoTyper) where USVs are recorded per pair and not per individual. We propose a Composite Link Model (CLM) approach to parametrise a mechanistic "sum-of-rates" model in which the pair's USV-rate is the sum of the USV-rates of individuals depending on their own behaviour. In generalized linear models (GLM) the individual's USV-rates are multiplied. We verified through simulation that CLM gave lower Poisson Deviance than GLM. We analysed data from an experiment in which half of the cages did allow the pairs to interact (Pair Housing) and the other half did not (Individual Housing). The "sum-of-rates" model fit best for Individual Housing and GLM for Pair Housing. An additional simulation study strongly suggests that interaction between rats changes the underlying mechanism for vocalisation behaviour.

4.1. Introduction

Automated home-cage systems have made it possible to study spontaneous behaviour of laboratory rodents. Studying laboratory rodents (here, rats) in pairs is an essential step towards understanding their social behaviour and to use rats as a model species for studying psychiatric disorders with social impairment such as autism and depression. Recently, important steps have been made to improve individual tracking of multiple rats in a home-cage. These advances allow researchers to study vital social behaviour such as playing and fighting. The PhenoTyper® system (Noldus Information Technology, Wageningen, The Netherlands) tracks the whereabouts of both rats in the cage using a top-view camera and obtains their exact location per video-frame. The distance that the rat covers in the interval between two frames (further: time interval, typically 1/25 second) is used to calculate several parameters such as its velocity. With such a system we can study multi-dimensional aspects of behaviour (e.g. repetitive and stereotypic behaviours and social interaction; relevant to autism) for longer periods of time without human interference (Kas et al. 2014).

In addition to movement of rats, Ultrasonic Vocalisations (USVs) are assumed to be an important part of rat behaviour. USVs are indicators of the emotional state of the vocalizing rat (Brudzynski 2009; Burgdorf et al. 2008), important for establishing and maintaining social contact such as in sexual and playing behaviour (Himmler et al. 2014; Wöhr and Schwarting 2013), and USVs can invoke response of other rats e.g. transmission of fear (Kim et al. 2010). Emission of USVs in social settings is one of the most widely used means to study social communication in rodents (Servadio, Vanderschuren, and Trezza 2015) and can provide more insight into the functional meaning of social behaviour (Peters, Pothuizen, and Spruijt 2015).

Using the Sonotrack® recording system (Metris, Hoofddorp, The Netherlands) we can continuously record vocalizations of rats in the PhenoTyper. Interpretation of these data streams for studying autism, for example, requires development of analytical methods to establish the relation between USVs and behaviour (Kas et al. 2014). As of yet, it has not been possible to assign recorded calls to individual rats when multiple rats are housed in the PhenoTyper.

Allocating USVs to individual rats is difficult because "voices" of rats do not noticeably differ between individuals and that USVs recorded in (automated home) cages cannot be traced back to their location of origin because echoes of USVs can be as loud as or louder than the original USVs (R. Bulthuis, Metris, personal communication).

Being able to record activity per individual rat and not being able to record USVs per individual rat poses a challenge when integrating both data streams. Because of this challenge, Ågmo and Snoeren (2015), for example, could not anal-

yse the full data of their experiment on the effect of USVs on mating behaviour. We address the challenge by proposing a simple mechanistic model: sound is recorded when either one of the rats produces sound and the recorded vocalization rate is thus the sum of the two individual rates. We then assume as a statistical model a generalized linear model for the individual vocalization rate in relation to the individual activity. In this way, the model becomes a Composite Link model (CLM), which we then extend to be applicable to fine-scaled video-frame data.

In the first part of the paper, we show that the mechanistic model performs better in simulations than a traditional Poisson generalized linear model (GLM). In the second part of the paper, the model is applied in a case study. In the case study, data from pairs of rats that could or could not interact is analysed. We show that this potential for interaction between rats fundamentally changes the relation between activity and vocalisation behaviour, and thus that the mechanistic model proposed in the first part of the paper does not hold for rats with the potential to interact, which strengthens the claim that vocalisation behaviour plays a role in social interaction.

4.2. Statistical modelling

4.2.1. Generalized Linear Model approach

We wish to integrate USVs and the activity of two rats. For simplicity, let us suppose the activity per rat (A_1 and A_2 for rat 1 and 2 respectively) is recorded for every time interval in three categories: S, L, and P, denoting Stopping (sitting still), Lingering (moving slowly) and Progressing (moving quickly), respectively. For every time interval, we evaluate whether or not a USV was detected. There exist nine combinations of activity states ($A_1A_2 \in States$, with $States = \{SS, SL, SP, LS, LL, LP, PS, PL, PP\}$. For all these combinations we can sum the observed number of USVs ($y_{A_1A_2}$ with $A_1A_2 \in States$) and the total number of frames ($F_{A_1A_2}$ with $A_1A_2 \in States$).

The cage-USV rate ($\mu_{A_1A_2}$ with $A_1A_2 \in States$) (i.e. the expected number of USVs per time interval) can now easily be estimated using a Poisson GLM with responses $y_{A_1A_2}$, a log-link function, nine parameters (Table 4.1) and an offset equal to the logarithm of $F_{A_1A_2}$.

If both rats vocalize identically, $\mu_{SL} = \mu_{LS}$, $\mu_{SP} = \mu_{PS}$, and $\mu_{LP} = \mu_{PL}$, so that the number of parameters reduces from nine to six. In that case, the data can be represented as in Table 4.2 and the corresponding model is:

Combined Activity: $$\mu_{A_1A_2} = h(\beta_{A_1A_2}) = \text{EXP}(\beta_{A_1A_2}), \tag{4.2}$$

Table 4.1: Vocalisation rate per combination of activity categories

		Rat 1		
	Activity State	Stopping	Lingering	Progressing
Rat 2	Stopping	μ_{SS}	μ_{LS}	μ_{PS}
	Lingering	μ_{SL}	μ_{LL}	μ_{PL}
	Progressing	μ_{SP}	μ_{LP}	μ_{PP}

where h is the log-link function ($h(x) = \text{EXP}(x)$) and $\beta_{A_1A_2}$ represents the parameter to be estimated. This Poisson GLM approach (Combined Activity GLM) is easy to implement and flexible, yet we have to estimate a separate parameter for every combination of activity states and we completely disregard the fact that there exist two individual rats that both emit USVs rather than a single source.

Table 4.2: Input data for Contingency table GLM Poisson assuming equal vocalization parameters

Activity Combination	USVs	Frequency
Stopping - Stopping	y_{SS}	F_{SS}
Stopping - Lingering	$y_{SL} + y_{LS}$	$F_{SL} + F_{LS}$
Stopping - Progressing	$y_{SP} + y_{PS}$	$F_{SP} + F_{PS}$
Lingering - Lingering	y_{LL}	F_{LL}
Lingering - Progressing	$y_{LP} + y_{PL}$	$F_{LP} + F_{PL}$
Progressing - Progressing	y_{PP}	F_{PP}

An alternative Poisson GLM model (Count GLM) is to use the activity states per rat as explanatory variables, rather than the activity combinations. If vocalization parameters of both rats are identical, we require only three parameters and can represent the data as in Table 4.3 (β_S, β_L, and β_P).

The corresponding model is:

$$\text{Count GLM:} \qquad \mu_{A_1A_2} = h(\beta_{A_1} + \beta_{A_2}) = \text{EXP}(\beta_{A_1} + \beta_{A_2}) = \text{EXP}(\beta_{A_1}) \cdot \text{EXP}(\beta_{A_2}). \tag{4.2}$$

The Count GLM model reduces the number of parameters from six to three. It still disregards the fact that there are two individual rats that both vocalize. Note that, due to the log-link function, the Count GLM model estimates $\mu_{A_1A_2}$ as a product of (what we would like to interpret as) rates, where the rates are the exponentiated parameters. However neither parameter has an interpretation as a rate, because rates should be summed rather than multiplied. In other words, this model gives no estimate for e.g. the USV-rate of a lingering rat.

Table 4.3: Input data for GLM Poisson assuming equal vocalization parameters

Activity			USV	
Stopping	Lingering	Progressing	USVs	Frequency
2	0	0	y_{SS}	F_{SS}
0	2	0	y_{LL}	F_{LL}
0	0	2	y_{PP}	F_{PP}
1	1	0	$y_{SL} + y_{LS}$	$F_{SL} + F_{LS}$
1	0	1	$y_{SP} + y_{PS}$	$F_{SP} + F_{PS}$
0	1	1	$y_{LP} + y_{PL}$	$F_{LP} + F_{PL}$

4.2.2. Composite Link Model approach

Mechanistically, the USV-rate of a pair of rats is the *sum* of the USV-rates of both rats. So a more mechanistic model would estimate the USV-rates per rat based on their activity, and sum these estimates:

Sum-of-rates CLM: $$\mu_{A_1 A_2} = \text{EXP}(\beta_{A_1}) + \text{EXP}(\beta_{A_2}). \tag{4.2}$$

This "sum-of-rates" model can be recognized as a composite link model (CLM) (Thompson and Baker 1981). In CLM, one observation can be linked to multiple linear predictors which allows us to link the observed USVs of a cage to a separate linear predictor per rat. The exponents of the three estimated parameters $\text{EXP}(\beta_S)$, $\text{EXP}(\beta_L)$, and $\text{EXP}(\beta_P)$, can be interpreted directly as the USV-rate of a rat in the Stopping, Lingering, or Progressing activity state, respectively. An alternative mechanistic interpretation for the parameters of the same statistical model is described in Section S.4.A.

The concept of the Composite Link Model (Thompson and Baker 1981) is easiest to grasp in matrix notation. The CLM model is written as:

$$\mu = Ch(X\beta)$$

where C is a so-called link-matrix, h is a link-function (log-link in our situation), matrix X contains the observations, and vector β contains the parameters. If the link-matrix is the identity matrix, the model is equivalent to a GLM. The "sum-of-rates" model is a CLM with a C-matrix that consists of two adjacent identity matrices and an X-matrix with double the number of observations, one for every rat rather than one for every cage (more extensive explanation in Section S.4.B).

In the special case of a model where the linear predictors are the same for both rats, the "sum-of-rates" model is equivalent to a Poisson GLM with a constant difference of log(2). For example a CLM model for predicting USV-rate as a

function of cage-temperature T:

$$\mu_T = \text{EXP}(\beta_i + \beta_T T) + \text{EXP}(\beta_i + \beta_T T) = 2\,\text{EXP}(\beta_i + \beta_T T) = \text{EXP}(\text{LOG}(2) + \beta_i + \beta_T T) = \text{EXP}(\beta_i^* + \beta_T T), \quad (4.2)$$

where β_i is the intercept in the "sum-of-rates" model and β_i^* is the updated intercept in the Poisson GLM approach.

4.2.3. EXTENSION TO MULTIPLE CAGES

Pairs of rats may vocalize with different overall frequencies. When extending our approach for multiple cages, we would therefore like to allow for a cage-specific intercept for each of the K cages ($k = 1, \ldots, K$). As a consequence, one of the activity states (Stopping) is chosen as reference category and β_S is set to zero. In the GLM approach, the exponent of the cage-specific intercept β_k can be interpreted as the USV-rate of the cage when both rats are in activity state Stopping. For the CLM approach (further: Count CLM), β_k is included in the linear predictor of both rats and $\text{EXP}(\beta_k)$ can thus be interpreted as the USV-rate of one rat in activity state Stopping. This extension is used in the case study below.

4.2.4. FROM COUNTS TO BINARY DATA

The approach we have taken so far is based on counts observed in intervals: it estimates a USV-rate assuming that $y_{A_1A_2}$ follows a Poisson distribution with a USV-rate $\mu_{A_1A_2}$ per time interval and an interval length $F_{A_1A_2}$. However, we record USV's as a binary variable per video-frame: $y_{A_1A_2}$ for a frame is either 1 or 0, depending on whether USV is or is not detected within the frame. The problem with the count approach is that multiple USVs could occur within one frame and the count over frames does not account for this. This problem can be overcome by modelling the binary response per video-frame using a binomial distribution.

For ease of notation we drop the indices A_1A_2 for the moment. The probability that $y = 1$ in a frame is equal to the probability of detecting at least one USV, that is, the probability that the number of vocalizations (N) is not equal to zero. As N is Poisson distributed,

$$p(y = 1) = p(N > 0) = 1 - P(N = 0) = 1 - \text{EXP}(-\mu t) \quad (4.2)$$

where μ is the vocalization rate and t the duration of the video-frame. In a GLM model for binary data (binary GLM model), this equation results in the complementary log-log-link function ($H(x) = \text{LOG}(-\text{LOG}(1 - x))$), rather than the customary logistic function. We note for our model that the vocalization rate in the equation is the total rate, which is the sum of the rates of the two individual rats (Equation 4.2). The resulting model can be expressed as a CLM-model with two

link functions, a log-link that relates predictors to the rates and a inverse-link function (Equation 4.2) that links the total rate to the binary observation:

$$p(y = 1) = E(y) = h_1(C(h(X\beta))) \tag{4.2}$$

where $h_1(x) = 1 - \text{EXP}(-x)$ and $h(x) = \text{EXP}(x)$. This "sum-of-rates" model for binary data, briefly referred to as the binary CLM model, is a special case of the bilinear composite link model of Thompson and Baker (1981).

4.3. Simulation Study

4.3.1. Material and Methods

We simulate an experiment with one cage of two rats according to the binary "sum-of-rates" model. For each rat, activity data was generated as a Markov-process with the three activity states as states using an input transition matrix based on the case study. USVs were generated per rat, using a Poisson process with a length of 1 frame, and thereafter combined yielding a '0' if no USVs were detected in the frame and a '1' otherwise (the data are thus truncated at 1). The input USV-rates for the Poisson process per activity state are 0.01, 0.14, and 0.04 for Stopping, Lingering and Progressing respectively. These USV-rates were loosely based on the real data, namely, the Pair Housing group of the case study (subsection Case study). Every data set was analysed using a Combined Activity GLM; a count and a binary GLM approach; and a count data and a binary CLM approach. The simulation study was repeated using ten times higher input USV-rates.

Performance of the five models was quantified in terms of deviation of the model predictions from the expected values based on the true underlying data distribution. More specifically, we calculated the Poisson deviance of the model predictions as:

$$PD = 2 \sum_{A_1 \in \{S,L,P\}} \sum_{A_2 \in \{S,L,P\}} (p_{A_1A_2} \cdot \text{LOG}(\frac{p_{A_1A_2}}{q_{A_1A_2}}) - p_{A_1A_2} + q_{A_1A_2}), \tag{4.3}$$

where $p_{A_1A_2}$ is the expected number of USVs based on the true underlying data distribution and $q_{A_1A_2}$ is the model prediction of number of USVs. We also calculated the PD of the observed data.

4.3.2. Results

The Poisson deviances of the CLM models were lower than those of the corresponding GLM models, whereas the Poisson deviance of the Combined Activity GLM was in between (Figure 4.1, top row). With ten times higher input USV-rates, the binary models compared to the count models becomes apparent (Figure 4.1, bottom row). This is logical, as with higher USV-rates, multiple USVs per frame

become more frequent than with lower USV-rates and the binary approach takes account of the truncation at 1 per frame. The simulation results also confirm that the activity state coefficients of the CLM models can be directly interpreted as USV-rate per rat. These estimates are close to the input USV-rates. Over all replications, all activity states, and both scenarios, the maximum observed squared error of a USV-rate estimate of the binomial CLM model was less than 0.001.

4.4. Case study

4.4.1. Experimental design and analysis

We observed 16 pairs of rats in automated home-cages (PhenoTyper® 9000, Noldus Information Technology, Wageningen, The Netherlands) equipped with a Sonotrack® recording system (Metris, Hoofddorp, The Netherlands). Behaviour was recorded using EthoVision® (Noldus, Information Technology, Wagenignen, The Netherlands). In the first eight home-cages, the two rats are housed together and can interact (Pair Housing). In the second eight, they are housed in a similar home-cage but with a separator in the middle which does not allow interaction (Individual Housing). For each home-cage, for every frame, we record whether or not a USV was observed, and also the activity of each of the two rats (Stopping, Lingering, Progressing). Data is analysed using four candidate binomial CLM models:

Intercept-only The USV-rate per frame differs between the cages and is independent of activity of the rats in that cage.

$$\mu_k = \text{EXP}(\beta_k) + \text{EXP}(\beta_k), \tag{4.4}$$

where μ_i is the USV-rate per frame for each of the K cages ($k = 1, \cdots, K$) and β_k is the intercept term per cage.

Current Activity The USV-rate per frame differs between the cages and is dependent on the activity of rat 1 and activity of rat 2 in that frame:

$$\mu_{kA_1A_2} = \text{EXP}(\beta_k + \beta_L X_{L_1} + \beta_P X_{P_1}) + \text{EXP}(\beta_k + \beta_L X_{L_2} + \beta_P X_{P_2}), \tag{4.4}$$

where $\mu_{kA_1A_2}$ is the USV-rate per frame for each of the K cages ($k = 1, \cdots, K$) given the activity of rat 1 ($A_1 = S, L$, or P) and rat 2 ($A_2 = S, L$, or P), β_k is the intercept term per cage, β_L and β_P are the regression coefficients for Lingering and Progressing respectively, and X_{L_1}, X_{L_2}, X_{P_1}, and X_{P_2} are indicator variables that indicate whether or not rat 1 (X_{A_1}) and rat 2 (X_{A_2}) are lingering (X_{L_n}) or progressing (X_{P_n}). Note that, as is customary for regression models with nominal variables (factors), the intercept per cage is

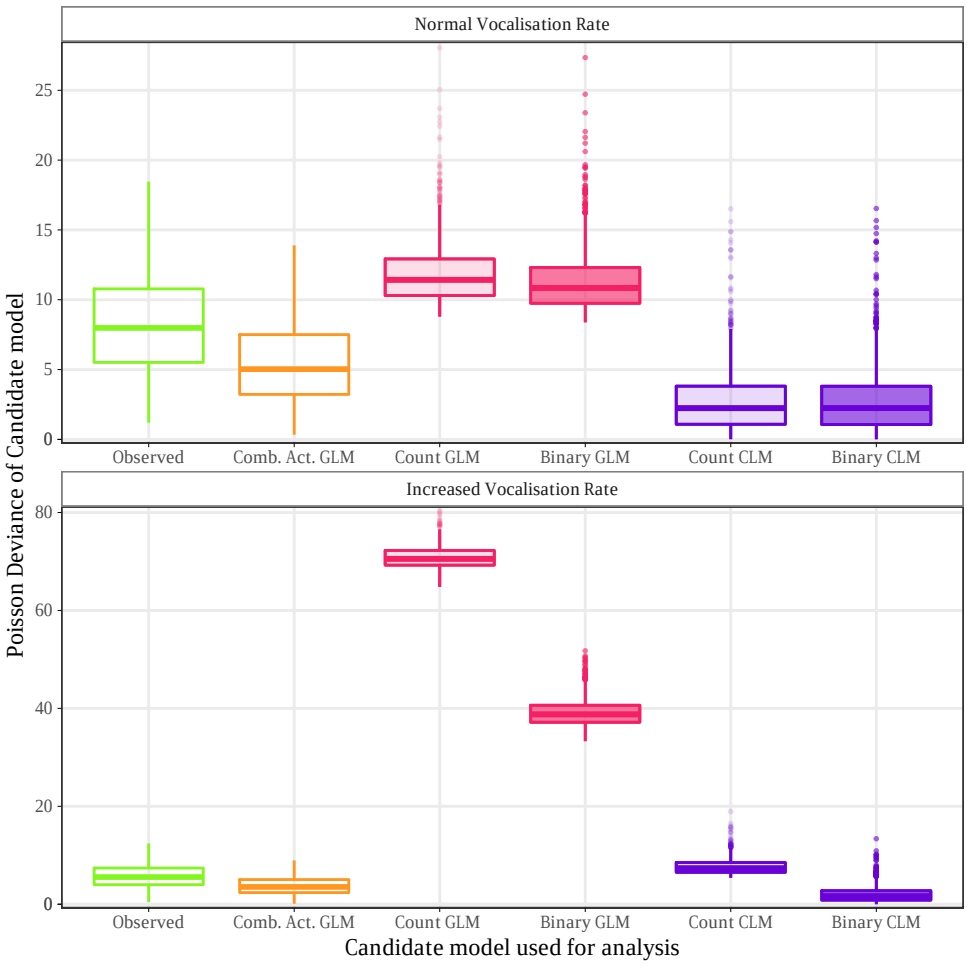


Figure 4.1: Poisson Deviance of observed and modelled number of USVs on simulated data with realistic vocalisation rates (top row) and data with increased vocalisation rates (bottom row). The Combined Activity (Comb. Act.) GLM model has six and the count and binary models have three parameters.

here the expected USV-rate per frame when the rat is Stopping (i.e. the baseline USV-rate) and the regression coefficients for Lingering measures the difference in log-vocalization rate between Lingering and Stopping and similarly for Progressing.

Past Activity The USV-rate per frame differs between the cages and is dependent on the activity of rat 1 and activity of rat 2 in the second (25 frames) before that frame. The model equation is the same as Equation 4.4 with a different definition of all parameters except β_k, β_L, and β_P. In the Past Activity model, μ_{if} is the USV-rate per time interval for each of the K cages ($i = 1, \cdots, K$) and ($f = 25, \cdots, F$) time intervals given the activity of rat 1 and rat 2 in the second before that frame, and X_{L_1}, X_{L_2}, X_{P_1}, and X_{P_2} are the proportion of frames in the previous second rat 1 (X_{A_1}) and rat 2 (X_{A_2}) were lingering (X_{L_n}) and progressing (X_{P_n}). The first 24 time intervals in the data set cannot be used because we do not have recorded activity in the second before.

Averaged Past Activity The USV-rate per frame differs between the cages and is dependent on the combined activity of both rats in the second (25 frames) before that frame. The model equation is the same as Equation 4.4 different interpretation of all parameters except β_k, β_L, and β_P. In the Averaged Past Activity model, μ_{if} is the USV-rate per frame for each of the K cages ($i = 1, \cdots, K$) and ($f = 25, \cdots, F$) given the averaged activity of both rats in the second (25 frames) before that frame; X_{L_1} equals X_{L_2}, and X_{P_1} equals X_{P_2} which are the proportion of frames in the previous second rat 1 and rat 2 were lingering and progressing. The model equation can be rewritten as a GLM model where $\beta_k^* = \beta_k + \text{LOG}(2)$ (as in Equation 4.2).

The best candidate model for the data set is chosen based on AIC (the lower the better). The Individual Housing and Pair Housing data set have a different best candidate model (Table 4.4). In the Individual Housing data, the best candidate model was the Current Activity model. This model is similar to the one used in the simulation study. For the Pair Housing data, the best candidate model was the Averaged Past Activity Model. This model has equal linear predictors for both rats and thus has an equivalent GLM model (see equation 4.2).

4.5. Discussion

In this paper we presented two approaches to relate the USV-rate per cage to the activities of two rats which we evaluated in a small simulation study. In the CLM

Table 4.4: AIC, AIC difference from the Intercept Only model, and rank based on lowest AIC of the four candidate models on the Individually Housed and the Pair Housed dataset. Data of the model with the lowest AIC per dataset is printed in bold.

	Individual Housing			Pair Housing		
	AIC	ΔAIC	rank	AIC	ΔAIC	rank
Intercept Only	15852		4	30819		4
Current Activity	**15750**	**− 102**	**1**	29104	− 1715	3
Past Activity	15759	− 94	3	28830	− 1989	2
Averaged Past Activity	15756	− 96	2	**28700**	**− 2119**	**1**

approach, like in the mechanistic model, the activity of a rat predicts its USV-rate and the combination of USV-rates of both rats gives the USV-rate per cage ("sum-of-rates"). In contrast, in the GLM approach, the combined activity of both rats predicts the USV-rate per cage. On data simulated using the mechanistic model, the Count and Binary CLM models were better able to predict USV-rate per cage than the Count and Binary GLM.

Application of the CLM approach in a case study on data of rats that could not interact, Individual Housing data, resulted in the same conclusion. The best model for this data in terms of AIC is the "sum-of-rates" model (Current Activity). When analysing data from rats that could interact (Pair Housing data) however, we found that the "sum-of-rates" model did not provide the best fit in terms of AIC. Instead, a CLM model with the averaged activity of both rats (Averaged Past Activity; which has an equivalent GLM model) as predictor fitted better.

As the AIC is not a formal test but rather a guideline for model selection, we verified that we have enough power in our analysis to select the "true" model via a second simulation study. In this simulation study, we alternately assume one of the four candidate models is true and generate data accordingly. All generated data was analysed using all four candidate models. In almost all instances the best candidate model was the candidate model used to generate the data (more details and results in Section S.4.C). From this simulation study we can thus conclude that it is highly unlikely that data with an underlying model structure from one of the candidate models, results in another best candidate model. Note also the large size of the differences between AICs of the candidate models.

We posed a mechanistic model for the vocalization rate observed in a cage in relation to the behaviour of the individual rats. This model led to a composite link model (equation 4.2). We thus went from a mechanistic model to a statistical model. However, the mechanistic model is not the only one leading to this specific statistical model. Another mechanistic model would be that Rat 1 vocalizes only

in response to the behaviour of Rat 2 and *vice versa*. This leads also to the statistical model of equation 4.2. For the full set of models see Section S.4.A.

We can nevertheless still infer that rats that *can* and rats that *cannot* interact show a different relation between USVs and activity. More specifically, the USV-rates estimated from individual behaviour of rats gives the best predictor when rats are housed individually, but when rats are housed in pairs, such an individualistic model no longer gives the best predictor and a predictor based on the combination of the rats behaviour performs better. The USV-rate of a pair of interacting rats is thus shown to be different from the sum of its parts. We conclude from this experiment that social interaction between rats changes the relation between activity and USV-rate of the rats.

With this application we have once more demonstrated the utility of the composite link model approach. A small R library implementing this approach is available for download[2].

References

Ågmo, Anders, and Eelke M S Snoeren. 2015. "Silent or vocalizing rats copulate in a similar manner." *PLoS ONE* 10 (12): 1–13.

Brudzynski, S M. 2009. "Communication of Adult Rats by Ultrasonic Vocalization: Biological, Sociobiological, and Neuroscience Approaches." *ILAR Journal* 50, no. 1 (January): 43–50.

Burgdorf, Jeffrey, Roger A Kroes, Joseph R Moskal, James G Pfaus, Stefan M Brudzynski, and Jaak Panksepp. 2008. "Ultrasonic vocalizations of rats (Rattus norvegicus) during mating, play, and aggression: behavioral concomitants, relationship to reward, and self-administration of playback." *Journal of Comparative Psychology* 122 (4): 357–367.

Himmler, BT, TM Kisko, D Euston, B Kolb, and SM Pellis. 2014. "Are 50-kHz calls used as play signals in the playful interactions of rats? I. Evidence from the timing and context of their use." *Behavioural processes* 106:60–66.

Kas, Martien J., Jeffrey C. Glennon, Jan Buitelaar, Elodie Ey, Barbara Biemans, Jacqueline Crawley, Robert H. Ring, et al. 2014. "Assessing behavioural and cognitive domains of autism spectrum disorders in rodents: Current status and future perspectives." *Psychopharmacology* 231, no. 6 (March): 1125–1146.

Kim, Eun Joo, Earnest S. Kim, Ellen Covey, and Jeansok J. Kim. 2010. "Social transmission of fear in rats: The role of 22-kHz ultrasonic distress vocalization." *PLoS ONE* 5 (12).

2. https://figshare.com/s/43a0014c3c47054ee93d

Peters, Suzanne M., Helen H.J. Pothuizen, and Berry M. Spruijt. 2015. "Ethological concepts enhance the translational value of animal models." *European Journal of Pharmacology* 759 (July): 42–50.

Servadio, Michela, Louk J M J Vanderschuren, and Viviana Trezza. 2015. "Modeling autism-relevant behavioral phenotypes in rats and mice: Do 'autistic' rodents exist?" *Behavioural pharmacology* 26 (6): 522–40.

Thompson, R, and R J Baker. 1981. "Composite Link Function in Generalized Linear Models." *Applied Statistics* 30 (2): 125–131.

Wöhr, Markus, and Rainer K W Schwarting. 2013. "Affective communication in rodents: ultrasonic vocalizations as a tool for research on emotion and motivation." *Cell and Tissue Research* 354, no. 1 (October): 81–97.

S.4.A. Alternative mechanistic interpretation of the statistical model

We posed a mechanistic model for the USV-rate observed in a cage in relation to the behaviour of the individual rats. This model led to a composite link model (equation 4.2 in the main text). We thus went from a mechanistic model to a statistical model. A valid question is whether the mechanistic model is the only one leading to this statistical model. In this section we show that it is not.

In this paper, we have assumed a mechanistic model in which the two rats in the same cage have a USV-rate depending on their activity state, which sums op to the USV-rate of the cage. In this model, we can interpret $\text{EXP}(\beta_{A_1})$ as the USV-rate of the first rat and $\text{EXP}(\beta_{A_2})$ as the USV-rate of the second rat. This is equivalent to saying that $\text{EXP}(\beta_{A_1})$ is the USV-rate of the cage conditional on the activity of the first rat and $\text{EXP}(\beta_{A_2})$ is the USV-rate of the cage conditional on the activity of the second rat.

But this is not the only possible interpretation for the parameters from the statistical model. Now let us assume a mechanistic model in which USVs of rats are not only related to their own activity, but also to the activity of the other rat in the cage. In the most extreme case, rat 1 would vocalize solely conditional on the activity of rat 2 and vice versa. Because we have no way of detecting which rat uttered which USV, this assumption would lead to exactly the same statistical model and exactly the same estimated model parameters. The full collection of models can be written as:

$$\mu_{A_1A_2} = \text{EXP}(p\beta_{A_1} + (1-p)\beta_{A_1}) + \text{EXP}(q\beta_{A_2} + (1-q)\beta_{A_2}) = \\ \text{EXP}(\beta_{A_1}) + \text{EXP}(\beta_{A_2})$$

where p and q range from 1 when USVs are solely related to a rat's own behaviour and 0 when USVs of a rate are solely related to the other rat's behaviour (Table S.4.A.1). The parameters p and q are inestimable when the observed USVs cannot be allocated to one of the rats.

Table S.4.A.1: Vocalisation rate per combination of activity categories

		USVs		
		Rat 1	Rat 2	Vocalisation Rate
Activity	Rat 1 A_1	$p\beta_{A_1}$	$(1-p)\beta_{A_1}$	$\text{EXP}(\beta_{A_1})$
	Rat 2 A_2	$(1-q)\beta_{A_2}$	$q\beta_{A_2}$	$\text{EXP}(\beta_{A_2})$
	sum			$\mu_{A_1A_2}$

Now the USV-rates of the cage conditional on the activity of the first or second rat respectively can still be estimated by $\exp(\beta_{A_1})$ and $\exp(\beta_{A_2})$ respectively. But the USV-rates of rat 1 and rat 2 are given by $\exp(p\beta_{A_1} + (1-q)\beta_{A_1})$ and $\exp((1-p)\beta_{A_2} + q\beta_{A_2})$ respectively which we cannot estimate.

S.4.B. CLM MODEL

The concept of the Composite Link Model (Thompson and Baker 1981) is easiest to grasp in matrix notation. In this supplement, we first write the Count GLM model in CLM notation and thereafter the "sum-of-rates" model. For brevity, we will show matrices as if there exist only two activity states (S and L). In that case, the model matrices for the Count GLM model are:

$$\mu = \begin{bmatrix} \mu_{SS} \\ \mu_{SL} \\ \mu_{LS} \\ \mu_{LL} \end{bmatrix}; C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}; X = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 2 \end{bmatrix}; \text{ and } \beta = \begin{bmatrix} \beta_S \\ \beta_L \end{bmatrix}.$$

Here C is the identity-matrix and thus $Ch(X\beta) = h(X\beta)$; so that

$$\mu = h(X\beta) \begin{bmatrix} \mu_{SS} \\ \mu_{SL} \\ \mu_{LS} \\ \mu_{LL} \end{bmatrix} = \text{EXP}(\begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \beta_S \\ \beta_L \end{bmatrix}) = \text{EXP}(\begin{bmatrix} 2 \cdot \beta_S \\ \beta_S + \beta_L \\ \beta_L + \beta_S \\ 2 \cdot \beta_L \end{bmatrix}) =$$

$$\begin{bmatrix} \text{EXP}(\beta_S + \beta_S) \\ \text{EXP}(\beta_S + \beta_L) \\ \text{EXP}(\beta_L + \beta_S) \\ \text{EXP}(\beta_L + \beta_L) \end{bmatrix} = \begin{bmatrix} \text{EXP}(\beta_S) \cdot \text{EXP}(\beta_S) \\ \text{EXP}(\beta_S) \cdot \text{EXP}(\beta_L) \\ \text{EXP}(\beta_L) \cdot \text{EXP}(\beta_S) \\ \text{EXP}(\beta_L) \cdot \text{EXP}(\beta_L) \end{bmatrix},$$

which is equivalent to Equation 4.2.1.

The link-matrix allows us to link multiple observations to one linear predictor. In order to obtain the "sum-of-rates" model of Equation 4.2, we set matrices X and C to:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } C = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix},$$

so that

$$h(X\beta) = \text{EXP}(\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_S \\ \beta_L \end{bmatrix}) = \begin{bmatrix} \text{EXP}(\beta_S) \\ \text{EXP}(\beta_S) \\ \text{EXP}(\beta_L) \\ \text{EXP}(\beta_L) \\ \text{EXP}(\beta_S) \\ \text{EXP}(\beta_L) \\ \text{EXP}(\beta_S) \\ \text{EXP}(\beta_L) \end{bmatrix}$$

and

$$\begin{bmatrix} \mu_{SS} \\ \mu_{SL} \\ \mu_{LS} \\ \mu_{LL} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \text{EXP}(\beta_S) \\ \text{EXP}(\beta_S) \\ \text{EXP}(\beta_L) \\ \text{EXP}(\beta_L) \\ \text{EXP}(\beta_S) \\ \text{EXP}(\beta_L) \\ \text{EXP}(\beta_S) \\ \text{EXP}(\beta_L) \end{bmatrix} = \begin{bmatrix} \text{EXP}(\beta_S) + \text{EXP}(\beta_S) \\ \text{EXP}(\beta_S) + \text{EXP}(\beta_L) \\ \text{EXP}(\beta_L) + \text{EXP}(\beta_S) \\ \text{EXP}(\beta_L) + \text{EXP}(\beta_L) \end{bmatrix},$$

which is equivalent to Equation 4.2.

References

Thompson, R, and R J Baker. 1981. "Composite Link Function in Generalized Linear Models." *Applied Statistics* 30 (2): 125–131.

S.4.C. Simulation Study for AICs

In the case study we observed that out of four candidate models the Current Activity model had the lowest AIC on the Individual Housing data and the Average Past Activity model had the lowest AIC on the Pair Housing data. In this simulation study we show that it is highly unlikely that, assuming that the true model is amongst the candidate models, the same model holds for the Individual Housing and Pair Housing data and we show that AIC is a suitable method for model selection in this type of data.

In our simulations study, for every combination of the four candidate models and the two experimental data sets (Individual Housing and Pair Housing):

(i) Fit the candidate model to the experimental data set to obtain an Input Model.

(ii) Predict the expected USV-rate for every frame in the data set using the model predictions from the Input Model.

(iii) Generate 1000 simulation data sets using the expected USV-rates from the previous step.

(iv) Fit all four candidate models to the 1000 generated simulation data sets.

(v) For every of the simulation data sets, compare the AICs of the four candidate models used for analysis and determine which candidate model has the lowest AIC.

Note that a candidate model here indicates the structure of the statistical model, not the size of coefficients. When comparing AICs of the candidate models, we always use the four AICs calculated in step *iv* of the procedure. The correct candidate model is thus the candidate model that has the same structure as the Input Model, not the exact same model.

The results of the simulation study show that AIC is a reliable tool for selecting the correct candidate model. In the vast majority of 8,000 simulation data sets, the correct candidate model had the lowest AIC. More often in data based on the Pair Housing (95.6%) than in the Individual Housing data (80.8%), and depending on which was the correct candidate model (Table S.4.C.1). In simulation data based on the Pair Housing case study, the correct candidate model always had the lowest AIC except when the correct candidate model was the Intercept Only model. In simulation data based on the Individual Housing case study, the Current Activity model always had the lowest AIC when it was correct.

Differences in AIC between the correct candidate model and the other three models (Figure S.4.C.1) were larger in the Pair Housing than the Individual Housing set and differed between the different candidate models used to estimate the

Input model. The order of magnitude of the AIC-differences when the correct candidate model had the lowest AIC was the same as the AIC-differences observed in the case study. The maximum AIC-difference observed when the wrong candidate model was identified was 13.3 which is well below the observed AIC-differences for the case study which indicates that.

Secondly, we use the simulation to show it is unlikely that the two data sets from the case-study have the same underlying model, given that a different candidate model had the lowest AIC when analysing this data (Current Activity for Individual Housing versus Average Past Activity for Pair Housing). For this aim, we determine how likely is it that the candidate model with the lowest AIC is in fact the correct candidate model. If it is very likely that the candidate model with the lowest AIC is the correct candidate model, and the candidate models with the lowest AIC are not the same for both data sets, it is unlikely that the underlying "true" model is the same.

In the simulation study, in the Individual Housing data sets, when the AIC of the Current Activity model is lowest that candidate model was the true candidate model in 89% of simulation data sets (Table S.4.C.2). In none of the simulation data sets in which the Current Activity model had the lowest AIC was the Average Past Activity the true candidate model. In the Pair Housing simulation data sets, when the AIC of the Average Past Second model is lowest that candidate model was the true candidate model in 96% of simulation data sets. In none of the simulation data sets in which the Average Past Activity model had the lowest AIC was the Current Activity the true candidate model. It is thus both unlikely that in the case study the true candidate model of the Individual Housing data set was the Average Past Activity model and that in the Pair Housing data set the true candidate model was the Current Activity model.

Table S.4.C.1: For each of the eight simulation scenarios (two experimental datasets x four candidate models), for each of the four candidate models used for analysis, the proportion of 1000 simulations in which each candidate model has the lowest AIC

Simulation data set		Candidate model used for analysis			
		IC Only	Cur. Act.	Past Act.	Ave. Past Act.
Ind. Housing	IC Only	0.772	0.100	0.058	0.070
	Cur. Act.	0.000	1.000	0.000	0.000
	Past Act.	0.000	0.000	0.716	0.284
	Ave. Past Act.	0.000	0.000	0.274	0.726
Pair Housing	IC Only	0.761	0.122	0.061	0.056
	Cur. Act.	0.000	1.000	0.000	0.000
	Past Act.	0.000	0.000	1.000	0.000
	Ave. Past Act.	0.000	0.000	0.000	1.000

Table S.4.C.2: For data simulated based on Individual Housing and Pair Housing, given that a candidate model has the lowest AIC, what is the observed probability that each of the four candidate models were used to generate the data set

Simulation data set		Candidate model with lowest AIC			
		IC Only	Cur. Act.	Past Act.	Ave. Past Act.
Ind. Housing	IC Only	1.000	0.091	0.055	0.065
	Cur. Act.	0.000	0.909	0.000	0.000
	Past Act.	0.000	0.000	0.683	0.263
	Ave. Past Act.	0.000	0.000	0.261	0.672
Pair Housing	IC Only	1.000	0.109	0.057	0.053
	Cur. Act.	0.000	0.891	0.000	0.000
	Past Act.	0.000	0.000	0.943	0.000
	Ave. Past Act.	0.000	0.000	0.000	0.947

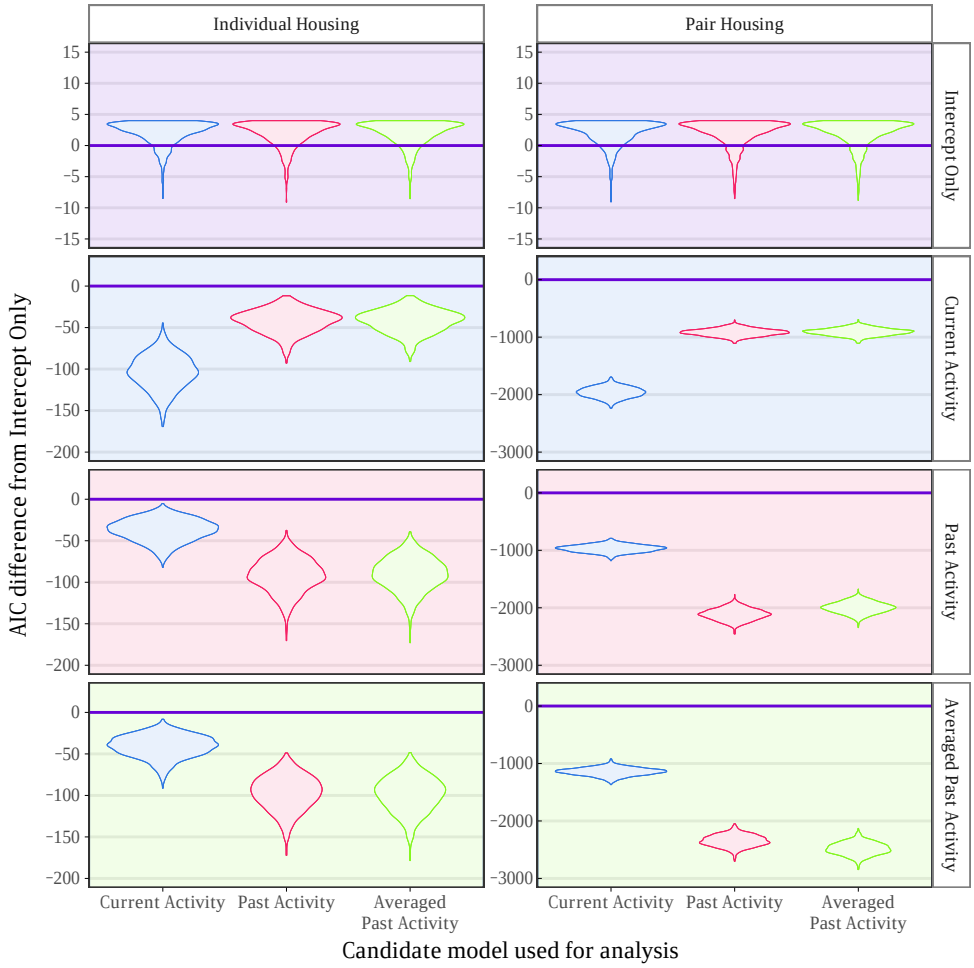
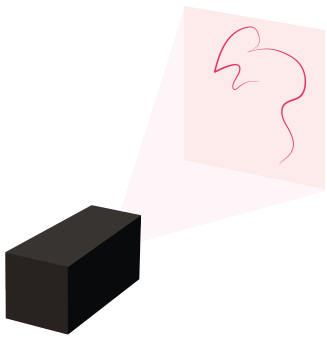


Figure S.4.C.1: Results of the simulation study for AICs. Each panel represents results on 1000 simulation data sets based on the Individual Housing (left column) or Pair Housing (right column) experiment, using each of the four candidate models (rows) to generate input data. Violin plots indicate the difference in AIC between the Intercept Only model and the candidate model used for analysis



5

The promises of targeted learning examined

Nadia J. Vendrig

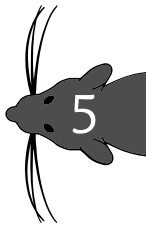
Lia Hemerik

Cajo J.F. ter Braak

Doubly robust methods have been proposed to better estimate causal effects from observational studies. Targeted Maximum Likelihood Estimation (TMLE) is a new and promising doubly robust method, that can use machine learning methods to increase performance. In theory TMLE should provide unbiased causal effect estimators even when either the treatment outcome or treatment assignment model is misspecified. When applying TMLE in practice however, the required theoretical assumptions such as the positivity assumption and no unobserved confounders can be violated. In this simulation study we illustrate the effects of unobserved (non-)confounding covariates and noise covariates on bias, RMSE, and coverage of TMLE on near-balanced data sets (with low risk of positivity violations) and unbalanced data sets (with higher risks of positivity violations). TMLE is able to estimate average causal effect with low bias and MSE, compared to the golden standard linear regression, given that the sample size is large, the data set is near-balanced, and the assignment model is specified correctly. In unbalanced data sets TMLE did not live up to expectations in this small simulation study. Also in data sets in which the positivity assumption was not violated.

5.1. Introduction

Doubly robust methods have been proposed to better estimate causal effects from observational studies (Laan and Robins 2003; Robins, Rotnitzky, and Zhao 1994) than traditional methods such as regression and ANCOVA methods. Both doubly robust and traditional models estimate the relation between the outcome Y, the treatment variable A, confounding covariates W, and non-confounding covariates (V) by means of the treatment outcome model $Q = E(Y|A, W, V)$ (further: outcome model), where $E(.)$ denotes expectation. Doubly robust methods differ from traditional models by also incorporating a propensity score weighting approach (Rosenbaum and Rubin 1983). In propensity score weighting, weights are assigned to subjects based on a treatment assignment model $g = P(A|W)$ (further: assignment model) which determines the probability a subject has to be in a treatment group given the confounding covariates. The advantage of using the double amount of models (treatment and assignment) is that doubly robust models are robust to misspecification of one of the models, given that the other is specified correctly.



Using doubly robust methods for causal inference in observational data requires the assumption of exchangeability, positivity, and consistency. Briefly, the exchangeability assumption entails that if the distribution of confounding covariates is the same in the treated and the non-treated group, the groups differ only in terms of the treatment. The exchangeability assumption thus implies there are no unobserved confounders and that there is no selection bias. The positivity assumption entails that all combinations of covariates in the study, have a positive probability to appear in the treated and the non-treated group. The consistency assumption (Cole and Frangakis 2009; VanderWeele 2009) or consistency rule (Pearl 2010) entails that the observed effect of the treatment on a subject is the only possible effect the treatment could have on that subject. Theoretically, these assumptions seem reasonable and straightforward. In practice however, it can be difficult to test whether the assumptions hold (*e.g.* Cole and Frangakis 2009; Petersen et al. 2012). For instance, the assumption of no unobserved confounders is inertly non-testable and the positivity assumption can hold theoretically while in the data set, by chance, certain combinations of covariates do not appear in both the treated and the non-treated group. Targeted Maximum Likelihood Estimation (TMLE) (Laan and Rose 2011; Laan and Rubin 2006) is a recent and promising doubly robust method, that can use machine learning methods to increase performance. TMLE will theoretically provide unbiased precise estimates of the parameter of interest when all assumptions are met. The method and its derivations are increasingly used for practical applications in a variety of domains (*e.g.* Arnold et al. 2017; Kotwani et al. 2014; Xu and Archambault 2015). The aim of this paper is to show what the effects are, in terms of bias, RMSE, and

coverage, of applying TMLE in less than optimal conditions. More specifically, we focus here on the effect of model misspecification and violation of the positivity assumption in the data set due to chance because these are issues that are beyond the researchers control and cannot be prevented when designing the study. Our approach was to generate data in a simulation study and analyse these data sets using several scenarios that differ in "prior knowledge" with regard to the studied system.

We start this paper with an illustrative example to introduce causal effect estimation (specifically TMLE). We explore the advantageous double robustness property of TMLE and describe the positivity assumption in more detail. Thereafter, we perform a simulation study to explore the robustness of TMLE in terms of bias, RMSE and (bootstrapping) coverage to several realistic flaws in analysis or data. We show the effect of incorrect model specifications by omitting confounding or non-confounding covariates and by adding uninformative covariates to the model. We explore the effect of (near) violations of the positivity assumption by generating data sets with more extreme data generating distributions and by generating data sets with very few observations. And we explore whether or not it is advantageous to use TMLE rather than a non-doubly robust model when no prior information on the treatment assignment mechanism is available.

5.2. Theory and Methods

5.2.1. Some background on causal effect estimation and doubly robustness

In this section we provide some background on causal effect estimation by way of a illustrative example.

Consider as example the estimation of the possible cardioprotective effect of moderate consumption of red wine, which has been debated for years (*e.g.* Streppel et al. 2009). Let us suppose one wishes to estimate the average causal effect of drinking wine versus not drinking wine on heart health. Heart health Y is measured on some continuous scale and predicted by a set of covariates W and by wine consumption, which is here the treatment or exposure variable A. Subjects included in the observational study are randomly selected, classified as wine drinkers (wine group; A=1) or not (no wine group; A=0), and have their heart health and covariates recorded. The difference in average heart health of the wine and the no wine group however, is not a valid estimator for the average causal effect, because subjects are not randomly assigned to the wine or the no wine group. Confounding factors, such as dietary preference and economic status, influence both the heart health and the probability for a subject to drink wine. Subjects in the wine and the no wine group have, on average, different baseline characteristics and different mean heart health regardless of wine consumption.

We now describe the role of the outcome model and the assignment model in this example.

The outcome model relates the heart health Y to the wine consumption A, a set of confounding covariates W, and a set of non-confounding covariates V. Subsequently, the model is used to predict what the heart health of that subject would have been, if that subject would have been in the other group (wine or no wine group). For every subject we now have two heart health values available: an observed heart health for the group the subject is in and an estimated, or counterfactual, heart health for the group the subject is not in. For every subject, the difference between the heart health in the wine group and in the no wine group is calculated. One of the outcomes is observed and the other is the counterfactual outcome. The average of these differences in heart health is an unbiased estimator for average causal effect (Rubin 1974).

The second model, the assignment model, relates the treatment (wine or no wine group) to the covariates. The assignment model is used to correct for differences in group composition in the wine and the no wine group using a method named inverse propensity weighting (IPW). The assignment model is used to estimate the propensity of subjects for being in the wine group or the no wine group based on the subject's baseline characteristics. Subjects with a high propensity to be in the wine group, have a set of baseline characteristics that are assumed to be overrepresented in the wine group and under-represented in the no wine group and *vice versa* subjects with a low propensity to be in the wine group, have a set of baseline characteristics that are assumed to be under-represented in the wine group and overrepresented in the no wine group. Inverse propensity score weighting corrects for over- and under-representation of sets of baseline characteristics by weighting subjects; subjects in the wine group with a low propensity to be in the wine group are weighted more heavily than subjects in the wine group with a low propensity to be in the wine group and *vice versa*. The weighting (nearly) balances out the differences in group composition and the average causal effect can be estimated as the difference between the weighted mean heart health of the wine and no wine group (Rosenbaum and Rubin 1983).

Both the method based on outcome model and the method based on assignment model result in unbiased estimates for average causal effect given that they are specified correctly. In doubly robust methods we incorporate both approaches. The benefit is that specifying either the outcome model *or* the assignment model correctly results in unbiased estimates of average causal effect. And thus that we allow for misspecification of one of the outcome model or the assignment model. This is a beneficial property in contrast to the traditional approach as it can never be checked whether an outcome or assignment model was specified correctly (i.e. all covariates measured and included in the correct form in the model). Do note

that doubly robust methods are not robust to unobserved confounding covariates. Confounding covariates are necessary for the correct specification of the assignment and the treatment model and thus cannot be unobserved.

5.2.2. Positivity assumption

A key assumption of TMLE and other doubly robust methods is the positivity assumption. The positivity assumption concerns the assignment model and entails that the probability of a subject belonging to any treatment group (here wine or no wine) should be positive (i.e. non-zero) for all combinations of covariates in the population studied. Violations of the positivity assumption can occur structurally when certain combinations are logically impossible. In the wine/no wine example for instance, if one of the covariates is whether a subject abstains from all alcohol, the probability for an abstainer to be in the wine group is zero. This is a violation of the positivity assumption in the strict sense. Even when the positivity assumption is not violated in the strict sense, it can still be violated in the data set. Violation of the positivity assumption in the data set occurs when some combinations of covariates are missing from either the treated or the non-treated group due to chance. In the remainder of this paper we will interpret adopt this broader interpretation of the positivity assumption. We did not incorporate violations of the positivity in the strict sense in our simulation studies.

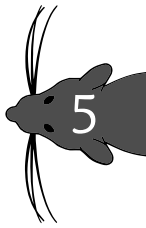
Several methods have been proposed for dealing with positivity assumption violations (see Petersen et al. (2012) for an overview of methods of diagnosing and handling the issue). Decreasing the number of covariates is an effective solution to reduce risk of violation and in some cases this comes at low costs of extra bias. Reducing the number of covariates however, also increases the risk of unobserved confounders. Another option is to check the positivity assumption for all combinations of covariates separately and remove all observations that have a combination of covariates for which the positivity assumption is violated. The consequence is that the conclusions drawn from the study generalize to only a subset of the population. In the wine example, we would remove all abstainers from the no wine group as there can be no abstainers in the wine group. The consequence would be that the estimated average causal effect is only valid for non-abstainers rather than for the full general population. Rather than limiting the population we could also limit the range of the target parameter (see next paragraph) by adjusting the projection function used to estimate the causal effect. In our example, we would limit the target parameter to estimate the effect of drinking wine in addition to drinking other alcoholic beverages rather then estimating the effect of drinking wine.

5.2.3. Targeted Maximum Likelihood Estimation

In normal maximum likelihood estimation all parameters in the outcome are simultaneously estimated, irrespective whether a parameter is a nuisance parameter or a parameter of interest. In contrast, TMLE is targeted towards the parameter of interest, the so-called target parameter. It aims at the best possible bias/variance trade-off for the target parameter, even if this would give larger bias and variance on the estimates for the covariates. The average causal effect (ACE) is defined based on the outcomes model. It is the expected difference between the outcome value for a subject and its counterfactual outcome. The estimate of the average causal effect ($\hat{ACE}$) is the simple analogue thereof, the difference between the fitted outcome value for a subject and its counterfactual outcome ($\hat{CE}_i$), averaged over all subjects ($i = 1, \dots, n$) in the study:

$$\hat{ACE} = \frac{1}{n}\sum_{i=1}^{n} \hat{CE}_i = \frac{1}{n}\sum_{i=1}^{n} \hat{Q}(A_i = 1, W_i, V_i) - \hat{Q}(A_i = 0, W_i, V_i) \tag{5.2}$$

where $\hat{Q}(A, W, V)$ is an estimate of the outcome model $Q(A, W, V) = E(Y|A, W, V)$. The initial estimate of $\hat{Q}(A, W, V)$ is obtained by traditional maximum likelihood which in the simplest case reduces to multiple linear regression.



The initial estimate of the assignment model ($g(A, W) = P(A|W)$) can be obtained using super learning and is used to define the clever covariate H_n. The clever covariate has values:

$$H_i \equiv \frac{1}{\hat{g}(1|W_i)} \text{ if } A_i = 1 \text{ and} \tag{5.2}$$

$$H_i \equiv -1 * \frac{1}{\hat{g}(0|W_i)} \text{ if } A_i = 0 \tag{5.2}$$

Now we regress the outcome parameter on the clever covariate using $\hat{CE}_i$ as an offset. The obtained coefficient for the clever covariate (ϵ_n) is used to update the estimate of the outcome model ($\hat{Q}_{up}(A, W, V)$). Then $\hat{Q}_{up}(A, W, V)$ is used to obtain an updated estimate for $\hat{ACE}$:

$$\hat{ACE}_{up} = \frac{1}{n}\sum_{i=1}^{n} \hat{Q}_{up}(A_i = 1, W_i, V_i) - \hat{Q}_{up}(A_i = 0, W_i, V_i) \tag{5.2}$$

where

$$\hat{Q}_{up}(A, W, V) = \hat{Q}(A, W, V) + \epsilon_n H_n$$

As proposed by Gruber and Van der Laan (2011), the regression is in practice a logistic regression after transforming both the outcome Y and the initial estimate $Q^0(A, W)$ by a logistic transformation. This is done to avoid outliers. The fitted values and associated counterfactuals of this model are then backtransformed.

5.3. Simulation Study

5.3.1. Data generation

In this simulation study, which resembles the one in Gruber and Van der Laan (2011), we generate data that mimic an epidemiological study which attempts to estimate the average causal effect of wine consumption on heart health. In the data generating model, the continuous variable heart health (Y) is dependent on wine consumption (A), three confounding covariates (W_{1-3}), and, three non-confounding covariates (V_{1-3}). The true outcome model is:

$$Y = A + 2W_1 + 3W_2 + 4W_3 + 2V_1 + 3V_2 + 4V_3 + e,\ e \sim \mathcal{N}(0, 1), \tag{5.3}$$

where the confounding covariates W_{1-3} and non-confounding covariates V_{1-3} are independent binary random variables:

$$W_{1-3} \sim Bernoulli(0.5);\ V_{1-3} \sim Bernoulli(0.5)$$

and W_{1-3} affect A through the assignment mechanism:

$$g_0(A = 1|W) = \frac{exp(\beta W_1 + \delta W_2 + \gamma W_3)}{1 + exp(\beta W_1 + \delta W_2 + \gamma W_3)}. \tag{5.3}$$

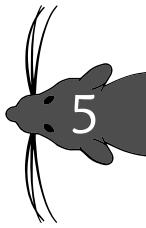
Following Gruber and Van der Laan (2011), we define two treatment assignment mechanisms: $g_{0,1}$ ($\beta_1 = 0.5, \delta_1 = 1.5, \gamma_1 = -1$) and $g_{0,2}$ ($\beta_1 = 1.5, \delta_1 = 4.5, \gamma_1 = -3$). Under $g_{0,1}$, the probability to be in the wine group ranges from 0.37 to 0.88 for the different combinations of W_{1-3}. We refer to data generated using this mechanism as "near-balanced" because it is likely that for all combinations of W_{1-3} there are observations in the wine and the no wine group. Under $g_{0,2}$, the probability to be in the wine group ranges from 0.05 to 0.997 for the different combinations of W_{1-3}. We refer to data generated using this mechanism as "unbalanced" because for some combinations of W_{1-3} it is likely there are only observations in either the wine or the no wine group, not both. In other words, in unbalanced data sets the potential for (near) violation of the positivity assumption is high.

A thousand data sets consisting of 100, 300, 1000, or 3000 observations were generated using the near-balanced assignment mechanism and another thousand using the unbalanced assignment mechanism. Data sets were restricted to have at least five observations with A=0 and five observations with A=1.

5.3.2. Statistical analysis

We specify four outcome models which represent differences in hypothetical prior knowledge. The four outcome models were (abbreviation between parenthesis): 1) the ideal situation in which all confounding and non-confounding covariates are known and no uninformative variables are supplied (Correct); 2) the situation in which the non-confounding covariates are unknown (MisNonConf); 3) the situation in which the confounding covariates are unknown (MisConf); and 4) the situation in which all covariates are known but also ten additional, uninformative variables are supplied (Noise), with model specifications:

1) Correct: $Y \sim A + W_1 + W_2 + W_3 + V_1 + V_2 + V_3$
2) MisNonConf: $Y \sim A + W_1 + W_2 + W_3$
3) MisConf: $Y \sim A + V_1 + V_2 + V_3$
4) Noise: $Y \sim A + W_1 + W_2 + W_3 + V_1 + V_2 + V_3 + N_1 + \ldots + N_{10}$, where N_{1-10} are independent binary variables: $N_{1-10} \sim Bernoulli(0.5)$,



Each of the four outcome model was combined with three statistical models. The statistical models represent differences in hypothetical prior knowledge on the assignment mechanism.

1) TMLEmodel: a TMLE that incorporates the correctly specified assignment model.
2) SuperTMLE: a TMLE with a SuperLearner-procedure estimating the assignment model using the variables in the outcome model, which represents having the suspicion that confounding covariates exist but having no prior knowledge on the assignment model.
3) LR: a multiple linear regression procedure which does not take any assignment mechanism into account.

Each of the data sets was analysed using each of the twelve combinations of the four outcome models and TMLEmodel, SuperTMLE, and LR (named *e.g.* Correct LR and MisConf SuperTMLE). Parameter estimates were assessed based on mean bias, bias of the median, root mean squared error (RMSE), coverage of the 95% confidence region (shortly denoted as coverage), and coverage of the 95% bootstrapping confidence interval (short denoted as bootstrapping coverage; calculated for TMLEmodel and SuperTMLE only).

Based on the theory, the LR analyses were predicted to result in unbiased estimates of the average causal effect and good coverage, except MisConf LR. MisConf LR was expected to cause biased estimates with unknown consequence for

the coverage. Correct LR was regarded as the golden standard when analysing this data, as the outcome model used to generate the data sets was linear and the Correct outcome model contained all relevant covariates and no others. TMLEmodel was predicted to be almost unbiased in all cases because TMLE is doubly robust and the correct assignment model was always specified. SuperTMLE was predicted to be almost unbiased with the correct outcome model.

LR and SuperTMLE are equivalent in terms of required prior knowledge on the model parameters. In both cases we only specified the outcome model, not the assignment model. Therefore we can directly compare the results of LR and SuperTMLE with the same model specifications. For TMLEmodel, we always specified a correct assignment model and thus we incorporated more prior knowledge on the model parameters in TMLEmodel than in LR and SuperTMLE. This difference in prior knowledge is especially relevant when evaluating performance of MisConf TMLEmodel to MisConf LR (and SuperTMLE). The hypothetical statistician that specified the MisConf TMLEmodel had accurate prior knowledge on the assignment model and was unaware of the role of the covariates in the outcome model. The confounding covariates were thus *not* truly unobserved but merely omitted from the outcome model. When the same hypothetical statistician incorporates the same prior knowledge in a LR model, she would always opt to include the covariates that she knows to be important for treatment assignment in the LR model to correct for them. Therefore, we deemed comparing results of MisConf TMLEmodel to MisConf LR irrelevant and compared results of MisConf TMLEmodel to Correct LR and correct SuperTMLE instead.

Analyses were carried out in the R statistical programming environment (version 3.3.1), using the lm-procedure from the R base package for LR and the tmle-procedure from the R package *tmle* (version 1.2.0-5) (Gruber and Van der Laan 2012) for TMLEmodel and SuperTMLE. Default settings were used in the lm-procedure. In the TMLEmodel and SuperTMLE procedure, in accordance to Gruber and Van der Laan (2011), the initial estimate of the conditional mean of $Y|A, W$ was bounded away from 0 and 1 by truncating at $(\alpha, 1-\alpha)$ with $\alpha = 0.005$ and predicted values for $g_n(A|W)$ were bounded away from 0 and 1 by truncating at $(p, 1-p)$ with $p = 0.01$. For TMLEmodel, adjusted bootstrap percentile intervals (bca) intervals based on 1000 replications were obtained using the R package *boot* (version 1.3-19) (Canty and Ripley 2013).

5.4. Results

5.4.1. Frequency of violation of the positivity assumption

The positivity assumption in the strict sense was satisfied throughout this simulation study because for all combinations of confounders there was a positive probability to be in both treatment groups. The positivity assumption in the

Table 5.1: Proportion of simulation data sets without violation of the positivity assumption (i.e. in which both outcomes of A are observed for each of the possible combination of the confounding covariates), split out by near-balanced and unbalanced treatment assignment mechanism and sample size

Observations	Near-balanced data	Unbalanced data
100	0.622	0.001
300	0.990	0.023
1000	1.000	0.197
3000	1.000	0.599

broader sense was however frequently violated in the generated data sets, as we frequently did not observe both treatments for all combinations of confounders (Table 5.1). The positivity assumption was more often violated in Unbalanced than in Near-balanced data sets, and more often in small than in larger data sets. In the unbalanced scenario with 100 or 300 less observations, more than 97% of the data sets violated the positivity assumption. In the Near-balanced scenario with 300, 1000, or 3000 observations, 1% or less of the data sets violated the positivity assumption.

5.4.2. Large near-balanced data sets

This subsection reports the results for near-balanced data with 1000 or 3000 observations. For these data sets the positivity assumption was never violated (Table 5.1).

As expected, TMLEmodel performed close to the golden standard and provided an unbiased accurate estimate of the ACE of A with all outcome models on these data sets (Figure S.5.A.1). Absolute bias was always under 0.01 and RMSE was always under 0.21 (Table 5.2). RMSE of the MisNonConf TMLEmodel was over double that of the other specifications of TMLEmodel. Coverage of all TMLEmodel outcome models was good, except for the MisConf TMLEmodel which had a coverage of 100% for both data sets. The coverage was correct (i.e. 95%) after bootstrapping.

Correct, MisNonConf, and Noise SuperTMLE performed similarly to the TMLEmodel with the same outcome model and thus produced accurate, unbiased average with good coverage. The estimates of MisConf SuperTMLE were strongly biased (mean and median bias over 0.35), had large error (RMSE 0.37), and low coverage (coverage < 51%). The poor results were expected as in this scenario the confounding covariates were unobserved, and thus both the assignment model and the treatment model were misspecified.

For all outcome models, results of SuperTMLE were similar to that of LR in terms of bias, RMSE, and coverage. RMSE of SuperTMLE was consistently (slightly) higher, and bias of SuperTMLE was consistently (slightly) lower than that of LR.

5.4.3. Small near-balanced data sets

This subsection reports the results for near-balanced data with 100 and 300 observations. The positivity assumption was violated for 38% of data sets sets with 100 observations and 1% of data sets with 300 observations (Table 5.1).

Bias of TMLEmodel was only slightly higher in small than in large near-balanced data sets (Figure S.5.A.1). Absolute mean and median bias for all outcome models and sample sizes was under 0.02. The effect on RMSE was more prominent, the RMSEs of TMLEmodel with the Correct outcome model for data sets with 100 and 300 observations were respectively 6.19, and 3.3 times larger than the RMSE for data sets with 3000 observations. Coverage of the MisConf TMLEmodel was too high (99.8%) and coverage of TMLEmodel with the other outcome models was too low (covarage < 91%). The coverages were correct after bootstrapping.

RMSE of SuperTMLE was slightly yet consistently larger than for LR (8 out of 8 data sets) and coverages and bootstrapping coverages of SuperTMLE were slightly yet consistently further from 95% than for LR (both 7 out of 8 data sets). The difference between small and large near balanced data sets was larger in SuperTMLE than in LR. The range of estimates and errors was larger for Noise SuperTMLE than for Noise LR, especially for data sets with 100 observations (Figure S.5.A.1).

5.4.4. Unbalanced treatment assignment mechanism

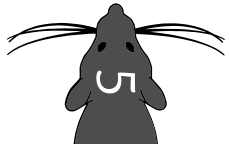
This subsection reports the results for small and large unbalanced data. In these data sets, violations of the positivity assumption occurred frequently (Table 5.1). TMLEmodel performed poorer on unbalanced than near-balanced data with the same number of observations in terms of RMSE, bias, and coverage (Table 5.2 and Figure S.5.A.2). RMSE was most severely affected, for example, RMSE of the MisConf TMLEmodel on unbalanced data was 10.3 times larger than on near-balanced data with 3000 observations (0.467 and 0.045 for unbalanced and near-balanced respectively). Coverage was too low (between 36% and 84%) for all outcome models and all sample sizes. Coverage increased with the number of observations. All coverages were higher after bootstrapping yet none reached up to 95%.

The MisConf TMLEmodel yielded biased estimates on unbalanced data. Bias decreased with increasing sample sizes, yet was still apparent on the largest data set with 3000 observations (mean bias: -0.20 and bias of the median: -0.11). The positivity assumption was violated in 40.1% of the unbalanced data sets with 3000 observations. The bias of the MisConf TMLEmodel reduced but did not disappear after excluding those data sets with a violated positivity assumption (mean bias: -

Table 5.2: RMSE, mean bias, bias of the median, coverage and bootstrapping coverage of effect-size estimates for near-balanced and unbalanced data with 100, 300, 1000, or 3000 observations

Data set and model		RMSE	Mean bias	Median bias	Coverage	Bootstrap coverage
Near-balanced data, 100 observations						
TMLE	Correct	0.2640	−0.009628	−0.00739	0.888	0.927
	MisNonConf	0.7094	0.006867	0.01637	0.912	0.946
	MisConf	0.3335	−0.004214	0.00876	0.998	0.973
	Noise	0.2709	−0.006234	−0.00389	0.864	0.946
SuperTMLE	Correct	0.2756	−0.007317	−0.00894	0.867	0.988
	MisNonConf	0.7172	0.008941	0.01495	0.909	0.960
	MisConf	0.6961	0.354318	0.36705	0.888	0.924
	Noise	0.3456	0.001022	−0.00737	0.738	0.988
LR	Correct	0.2434	−0.005832	−0.00432	0.935	
	MisNonConf	0.6722	0.015976	0.00613	0.945	
	MisConf	0.6957	0.351357	0.37077	0.907	
	Noise	0.2540	−0.003385	−0.00423	0.940	
Near-balanced data, 300 observations						
TMLE	Correct	0.1366	0.003276	−0.00118	0.943	0.951
	MisNonConf	0.3611	0.013273	0.01080	0.955	0.960
	MisConf	0.1485	0.008879	0.00921	1.000	0.964
	Noise	0.1378	0.002132	−0.00062	0.937	0.955
SuperTMLE	Correct	0.1369	0.003394	0.00087	0.939	0.976
	MisNonConf	0.3612	0.013459	0.01148	0.954	0.968
	MisConf	0.4924	0.362164	0.35806	0.809	0.816
	Noise	0.1401	0.002353	−0.00066	0.920	0.976

Continued on next page

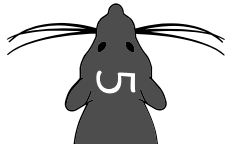


Data set and model		RMSE	Mean bias	Median bias	Coverage	Bootstrap coverage
LR	Correct	0.1315	0.004426	0.00178	0.950	
	MisNonConf	0.3478	0.012537	0.00033	0.959	
	MisConf	0.4919	0.361893	0.35937	0.821	
	Noise	0.1332	0.003502	0.00079	0.957	
Near-balanced data, 1000 observations						
TMLE	Correct	0.0726	0.000595	−0.00358	0.952	0.949
	MisNonConf	0.2067	0.000866	−0.00621	0.954	0.954
	MisConf	0.0786	0.001725	−0.00060	1.000	0.952
	Noise	0.0731	0.000737	−0.00471	0.950	0.950
SuperTMLE	Correct	0.0730	0.000607	−0.00284	0.950	0.960
	MisNonConf	0.2067	0.000483	−0.00521	0.954	0.936
	MisConf	0.4007	0.355728	0.35302	0.506	0.468
	Noise	0.0732	0.000732	−0.00356	0.949	0.964
LR	Correct	0.0705	0.000798	−0.00178	0.951	
	MisNonConf	0.1966	0.001760	−0.00720	0.957	
	MisConf	0.4005	0.355735	0.35267	0.514	
	Noise	0.0710	0.000915	−0.00196	0.951	
Near-balanced data, 3000 observations						
TMLE	Correct	0.0423	0.000624	0.00114	0.961	0.959
	MisNonConf	0.1241	−0.002080	−0.00576	0.946	0.947
	MisConf	0.0445	0.001773	0.00226	1.000	0.955
	Noise	0.0424	0.000509	0.00144	0.957	0.957
SuperTMLE	Correct	0.0423	0.000694	0.00147	0.961	0.980
	MisNonConf	0.1240	−0.002079	−0.00437	0.946	0.948
	MisConf	0.3776	0.360556	0.35959	0.089	0.108
	Noise	0.0424	0.000487	0.00131	0.958	0.972

Continued on next page

Data set and model		RMSE	Mean bias	Median bias	Coverage	Bootstrap coverage
LR	Correct	0.0404	0.000761	0.00234	0.949	
	MisNonConf	0.1187	−0.001405	−0.00297	0.944	
	MisConf	0.3775	0.360466	0.35959	0.091	
	Noise	0.0404	0.000646	0.00190	0.953	
Unbalanced data, 100 observations						
TMLE	Correct	1.1414	−0.01737	−0.0043	0.369	0.860
	MisNonConf	2.2302	−0.03605	−0.0198	0.412	0.843
	MisConf	2.2488	−0.57086	−0.8447	0.390	0.802
	Noise	1.1194	−0.01764	−0.0277	0.368	0.854
SuperTMLE	Correct	1.3274	0.02914	0.0399	0.302	0.924
	MisNonConf	2.3134	−0.02294	−0.0169	0.398	0.852
	MisConf	0.9730	0.75248	0.7263	0.733	0.724
	Noise	1.5233	0.06695	0.0394	0.229	0.677
LR	Correct	0.3097	−0.01824	−0.0181	0.955	
	MisNonConf	0.8586	−0.03130	−0.0500	0.954	
	MisConf	0.9648	0.75236	0.7311	0.784	
	Noise	0.3291	−0.01748	−0.0187	0.948	
Unbalanced data, 300 observations						
TMLE	Correct	0.5185	0.01884	0.0082	0.500	0.903
	MisNonConf	1.3469	−0.10625	−0.0739	0.504	0.892
	MisConf	1.5560	−0.73600	−0.5941	0.427	0.721
	Noise	0.5091	0.01997	0.0103	0.488	0.905
SuperTMLE	Correct	0.5591	0.01245	0.0096	0.452	0.924
	MisNonConf	1.3893	−0.11193	−0.0711	0.477	0.888
	MisConf	0.8023	0.73276	0.7328	0.382	0.360
	Noise	0.6461	0.02919	0.0232	0.367	0.848

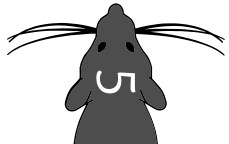
Continued on next page



Data set and model		RMSE	Mean bias	Median bias	Coverage	Bootstrap coverage
LR	Correct	0.1720	−0.00460	−0.0061	0.959	
	MisNonConf	0.4859	−0.03147	−0.0706	0.954	
	MisConf	0.8008	0.73201	0.7308	0.443	
	Noise	0.1756	−0.00401	−0.0075	0.958	
Unbalanced data, 1000 observations						
TMLE	Correct	0.2373	−0.00384	−0.0054	0.683	0.936
	MisNonConf	0.6761	−0.01847	−0.0474	0.671	0.935
	MisConf	0.9777	−0.44701	−0.1832	0.748	0.816
	Noise	0.2369	−0.00347	−0.0069	0.680	0.936
SuperTMLE	Correct	0.2393	−0.00422	−0.0015	0.672	0.932
	MisNonConf	0.6864	−0.01856	−0.0431	0.648	0.960
	MisConf	0.7496	0.72824	0.7288	0.016	0.016
	Noise	0.2496	−0.00566	−0.0101	0.629	0.928
LR	Correct	0.0944	−0.00328	−0.0055	0.949	
	MisNonConf	0.2661	−0.00310	−0.0120	0.957	
	MisConf	0.7493	0.72806	0.7311	0.018	
	Noise	0.0949	−0.00306	−0.0051	0.949	
Unbalanced data, 3000 observations						
TMLE	Correct	0.1268	0.00152	0.0014	0.763	0.925
	MisNonConf	0.3603	−0.01228	−0.0081	0.794	0.932
	MisConf	0.4666	−0.19601	−0.1112	0.840	0.924
	Noise	0.1269	0.00173	0.0017	0.758	0.930
SuperTMLE	Correct	0.1278	0.00204	0.0029	0.762	0.936
	MisNonConf	0.3610	−0.01135	−0.0054	0.787	0.928
	MisConf	0.7355	0.72817	0.7301	0.000	0.000
	Noise	0.1281	0.00117	0.0020	0.753	0.932

Continued on next page

Data set and model		RMSE	Mean bias	Median bias	Coverage	Bootstrap coverage
LR	Correct	0.0560	0.00123	0.0023	0.946	
	MisNonConf	0.1550	−0.00167	0.0040	0.954	
	MisConf	0.7354	0.72809	0.7301	0.000	
	Noise	0.0561	0.00116	0.0028	0.944	



0.12 and median bias: -0.08; Table 5.3). We observed slightly less frequent and less extreme outliers in the data sets without a violation of the positivity assumption than in the excluded data sets (Figure 5.1).

Performance of SuperTMLE with the different outcome models resembled that of TMLEmodel with the exception of MisConf SuperTMLE. As expected, MisConf SuperTMLE yielded severely biased estimates for all sample sizes (mean bias and bias of the median > 0.7). Coverage was too low and ranged from 73% with 100 observations to 0 with 3000 observations. Bootstrapping coverages were not better.

MisConf LR performed similarly to MisConf SuperTMLE in terms of bias, RMSE, and coverage. For all other outcome models, RMSE of SuperTMLE and TMLEmodel were consistently higher than that of LR (between 2.01 and 4.7 times) for the same outcome model on the same data set. Coverage of LR was near 95% throughout whereas coverages of SuperTMLE and TMLEmodel were too low.

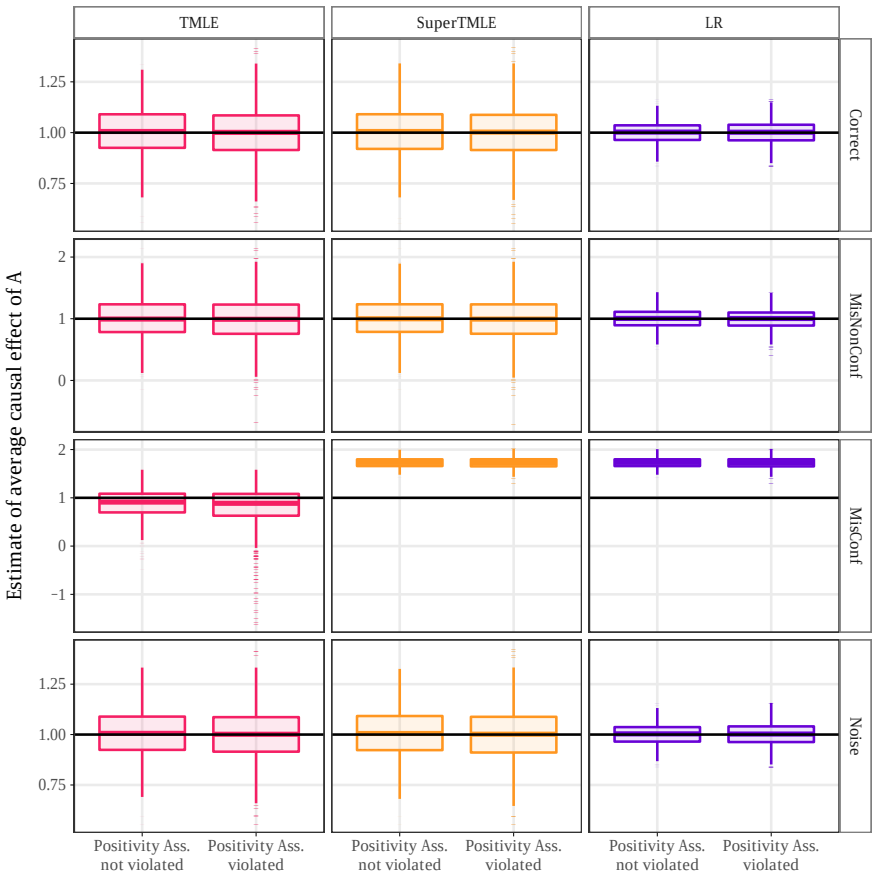


Figure 5.1: Boxplots of point estimates of the average causal effect of A estimated by TMLEmodel (pink), SuperTMLE (orange), and LR (purple) with Correct, MisNonConf, MisConf, and Noise treatment assignment models on unbalanced data with 3000 observations. Split out for data sets without (left) and with (right) violation of the positivity assumption

5.5. Discussion

Performance of doubly robust models has been evaluated in several studies (*e.g.* Carpenter, Kenward, and Vansteelandt 2006; Kang and Schafer 2007; Waernbaum 2012)) and is mostly based on MSE and bias. In our study, in addition to MSE and bias, the coverage of the 95% (bootstrapping) confidence interval was evaluated. Our simulation study was focussed on practical applicability of the results and aims at the decision making process.

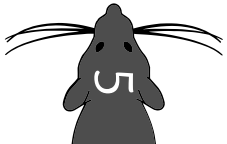
Under the optimal circumstances of near-balanced data with sufficient observations and no violation of the positivity assumption, and with the correct Assignment model specified, TMLEmodel proved to be doubly-robust by providing unbiased and precise estimates regardless of whether the Correct, Noise, MisConf, or MisNonConf outcome model was specified. Under these circumstances, the quality of prediction was very near to that of the golden standard Correct LR.

The doubly robust property of TMLE disappeared when risk of observed violation of the positivity assumption was higher: estimates of MisConf TMLEmodel were clearly biased, RMSE of all outcome models was higher than that of LR, and (bootstrapped) coverages were too low. These issues were not resolved by excluding data sets in which the positivity assumption was violated. We confirm results in previous studies that warn for detrimental effects of positivity assumptions violation (Petersen et al. 2012). The low coverages (80% - 90%) for doubly robust methods have been previously reported by Schafer and Kang (2008). In contrast, a study by Funk et al. (2011) did not reveal any coverage issues. In both studies substantial sample sizes were used and smaller sample sizes were not taken into account. Sample sizes of 100 or 300 subjects however, are not unusually small when estimating average causal effects using doubly robust methods (*e.g.* Rosenblum et al. 2009).

SuperTMLE provided unbiased and precise estimates on near-balanced data with sufficient observations and no violation of the positivity assumption, given that the confounding covariates were included in the outcome model. In other words, adding extra complexity by using a TMLE rather than the most simple accurate model (LR in our study) does not harm the estimates. TMLE comes with the advantage of having no assumptions on the relation between the covariates. In less advantageous circumstances (*e.g.* low sample sizes, noise covariates, unbalanced data), RMSE on the estimates of SuperTMLE was considerably higher than that of the golden standard LR model. When confounding covariates are unobserved, SuperTMLE produces estimates that are just as biased as those of LR. Using a SuperLearner procedure without prior knowledge on the assignment model in our situation thus did not help reduce bias but it does come at the cost of extra RMSE.

Table 5.3: As Table 5.2 for unbalanced data with 3000 observations with and without observed violation of the positivity assumption

Data set and model		RMSE	Mean bias	Median bias	Coverage	Bootstrap coverage
Positivity assumption not violated						
TMLE	Correct	0.1223	0.006318	0.00565	0.78	0.92
	MisNonConf	0.3391	0.003994	−0.00499	0.82	0.93
	MisConf	0.3368	−0.128897	−0.08793	0.88	0.94
	Noise	0.1226	0.006447	0.00604	0.78	0.93
SuperTMLE	Correct	0.1236	0.006841	0.00560	0.78	0.95
	MisNonConf	0.3390	0.006334	−0.00072	0.81	0.94
	MisConf	0.7372	0.730312	0.73428	0.00	0.00
	Noise	0.1234	0.005795	0.00544	0.76	0.94
LR	Correct	0.0547	0.000922	0.00282	0.95	
	MisNonConf	0.1583	0.001857	0.00557	0.95	
	MisConf	0.7371	0.730217	0.73457	0.00	
	Noise	0.0548	0.000820	0.00347	0.95	
Positivity assumption violated						
TMLE	Correct	0.1271	0.001216	0.00134	0.76	0.93
	MisNonConf	0.3616	−0.013319	−0.00865	0.79	0.93
	MisConf	0.4737	−0.200284	−0.11343	0.84	0.92
	Noise	0.1272	0.001427	0.00165	0.76	0.93
SuperTMLE	Correct	0.1281	0.001735	0.00278	0.76	0.94
	MisNonConf	0.3623	−0.012478	−0.00582	0.79	0.93
	MisConf	0.7354	0.728038	0.72998	0.00	0.00
	Noise	0.1284	0.000876	0.00194	0.75	0.93
LR	Correct	0.0560	0.001246	0.00220	0.95	
	MisNonConf	0.1548	−0.001896	0.00394	0.95	
	MisConf	0.7353	0.727956	0.73011	0.00	
	Noise	0.0562	0.001182	0.00270	0.94	



As shown in multiple studies, and confirmed in this paper, violation of the positivity assumption has serious consequences for performance of TMLE, and other doubly robust estimators (*e.g.* Carpenter, Kenward, and Vansteelandt 2006; Kang and Schafer 2007; Porter et al. 2011; Waernbaum 2012). Diagnosing and handling violations of the positivity assumption is not straightforward and opposing views exist on best practices (*e.g.* Cheng et al. 2010; Oakes, Messer, and Mason 2010; Petersen et al. 2012; Westreich and Cole 2010). The positivity issue is even more pressing in data from practice then in simulation data sets. The present simulation study had only three categorical confounders with two levels each and equal probability for each confounder level. Yet even when the assignment mechanism was near-balanced, with 0.88 as most extreme probability for A, having as many as 300 observations was not sufficient to prevent violations of the positivity assumption. And over 40% of the unbalanced data sets with 3000 observations showed positivity violations. Real-life data can be expected to have more complicated assignment mechanisms and *e.g.* continuous and multi-level categorical data which causes the chances that the positivity assumption holds even less likely. With the increasing use of TMLE in practice, resulting in publications in more applied journals, the importance of stressing the limitations of TMLE and other doubly robust approaches alongside their benefits remains an important task.

5.6. Conclusion

TMLE is able to estimate average causal effect with low bias and MSE compared to linear regression, given that the sample size is large, the data set is near-balanced, and the assignment model is specified correctly. Coverage is however not always sufficient; 95% bootstrapping confidence intervals provide sufficient coverage under these circumstances. In unbalanced data sets, TMLE did not live up to expectations in this small simulation study.

References

Arnold, Benjamin F., Mark J van der Laan, Alan E Hubbard, Cathy Steel, Joseph Kubofcik, Katy L Hamlin, Delynn M Moss, Thomas B Nutman, Jeffrey W Priest, and Patrick J Lammie. 2017. "Measuring changes in transmission of neglected tropical diseases, malaria, and enteric pathogens from quantitative antibody levels." Edited by Mathieu Picardeau. *PLOS Neglected Tropical Diseases* 11, no. 5 (May): e0005616.

Canty, Angelo, and B D Ripley. 2013. *boot: Bootstrap R (S-Plus) Functions.*

Carpenter, James R., Michael G. Kenward, and Stijn Vansteelandt. 2006. "A comparison of multiple imputation and doubly robust estimation for analyses

with missing data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169, no. 3 (July): 571–584.

Cheng, Yvonne W., Alan Hubbard, Aaron B. Caughey, and Ira B. Tager. 2010. "Cheng et al. Respond to "Positivity in practice"." *American Journal of Epidemiology* 171 (6): 678–679.

Cole, Stephen R., and Constantine E. Frangakis. 2009. "The Consistency Statement in Causal Inference." *Epidemiology* 20 (1): 3–5.

Funk, Michele Jonsson, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. "Doubly robust estimation of causal effects." *American journal of epidemiology* 173, no. 7 (April): 761–767.

Gruber, Susan, and Mark J. Van der Laan. 2011. In Laan and Rose 2011, chap. 7.3 Simulations.

Gruber, Susan, and Mark J Van der Laan. 2012. "tmle: An R Package for Targeted Maximum Likelihood Estimation." *Journal of Statistical Software* 51 (13): 1–35.

Kang, Joseph D.Y., and Joseph L. Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical science* 22, no. 4 (January): 523–539.

Kotwani, Prashant, Laura Balzer, Dalsone Kwarisiima, Tamara D. Clark, Jane Kabami, Dathan Byonanebye, Bob Bainomujuni, et al. 2014. "Evaluating linkage to care for hypertension after community-based screening in rural Uganda." *Tropical Medicine and International Health* 19 (4): 459–468.

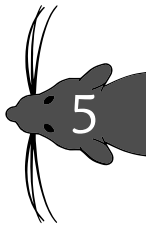
Laan, Mark J. van der, and James M. Robins. 2003. *Unified Methods for Censored Longitudinal Data and Causality.* Springer Series in Statistics. New York: Springer.

Laan, Mark J. van der, and Sherrie Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer Series in Statistics. New York: Springer.

Laan, Mark J. van der, and Daniel Rubin. 2006. "Targeted Maximum Likelihood Learning." *The International Journal of Biostatistics* 2, no. 1 (January): art. no. 11.

Oakes, J. Michael, Lynne C. Messer, and Susan Mason. 2010. "Messer et al. Respond to "Positivity in Practice"." *American Journal of Epidemiology* 171, no. 6 (March): 680–681.

Pearl, Judea. 2010. "On the Consistency Rule in Causal Inference." *Epidemiology* 21 (6): 872–875.



Petersen, Maya L., Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. van der Laan. 2012. "Diagnosing and responding to violations in the positivity assumption." *Statistical methods in medical research* 21 (1): 31–54.

Porter, Kristin E., Susan Gruber, Mark J. van der Laan, and Jasjeet S. Sekhon. 2011. "The Relative Performance of Targeted Maximum Likelihood Estimators." *The International Journal of Biostatistics* 7, no. 1 (January): 1–34.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–866.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70 (1): 41–55.

Rosenblum, Michael, Nicholas P Jewell, Mark van der Laan, Steve Shiboski, Ariane van der Straten, and Nancy Padian. 2009. "Analyzing Direct Effects in Randomized Trials with Secondary Interventions: An Application to HIV Prevention Trials." *Journal of the Royal Statistical Society. Series A, (Statistics in Society)* 172, no. 2 (April): 443–465.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66 (5): 688–701.

Schafer, Joseph L, and Joseph Kang. 2008. "Average causal effects from nonrandomized studies: a practical guide and simulated example." *Psychological methods* 13, no. 4 (December): 279–313.

Streppel, M T, M C Ocke, H C Boshuizen, F J Kok, and D Kromhout. 2009. "Long-term wine consumption is related to cardiovascular mortality and life expectancy independently of moderate alcohol intake: the Zutphen Study." *Journal of Epidemiology & Community Health* 63, no. 7 (July): 534–540.

VanderWeele, Tyler J. 2009. "Concerning the Consistency Assumption in Causal Inference." *Epidemiology* 20, no. 6 (November): 880–883.

Waernbaum, Ingeborg. 2012. "Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation." *Statistics in medicine* 31, no. 15 (July): 1572–1581.

Westreich, Daniel, and Stephen R. Cole. 2010. "Invited commentary: Positivity in practice." *American Journal of Epidemiology* 171 (6): 674–677.

Xu, Ziyun, and Éric Archambault. 2015. "Chinese interpreting studies: structural determinants of MA students' career choices." *Scientometrics* 105, no. 2 (November): 1041–1058.

S.5.A. Supplementary Figures

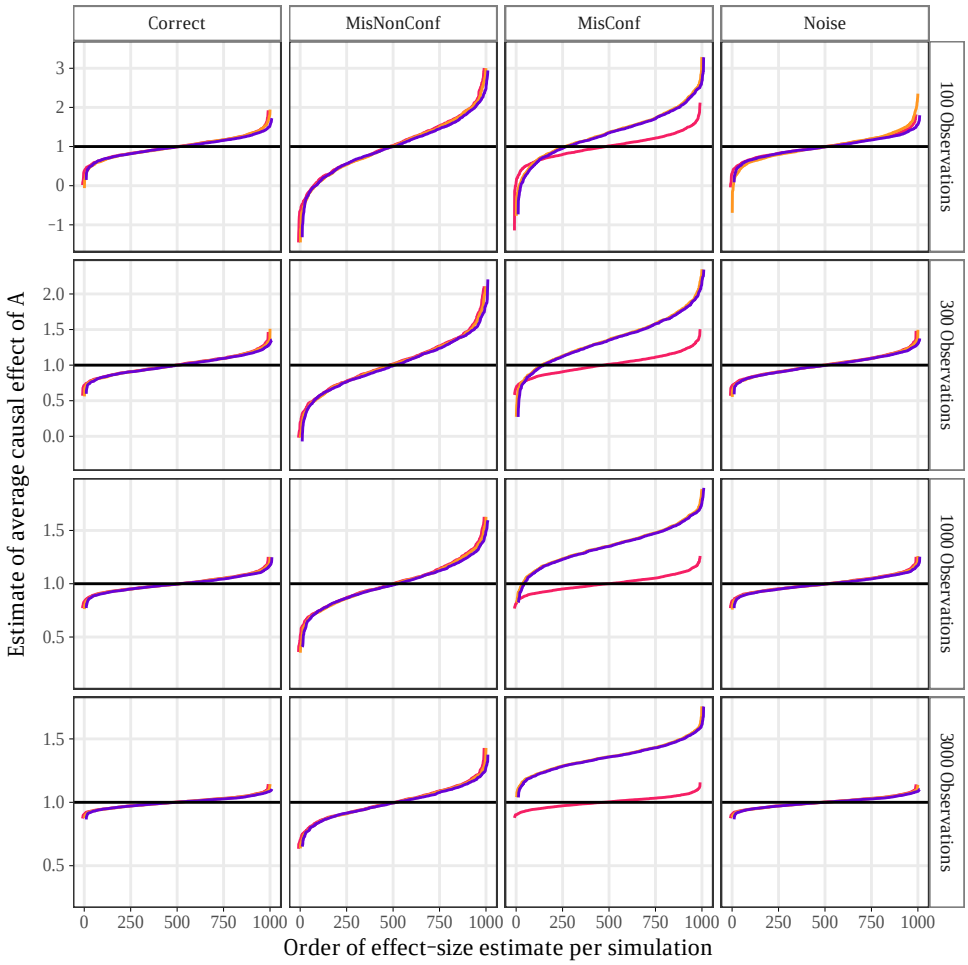


Figure S.5.A.1: Point estimates of the average causal effect of A, ordered from small to large, estimated by TMLEmodel (pink), SuperTMLE (orange), and LR (purple) with Correct, MisNonConf, MisConf, and Noise treatment assignment models on near-balanced data with 100, 300, 1000, or 3000 observations. For visibility, curves of TMLEmodel are shifted to the left and curves of SuperTMLE are shifted to the right

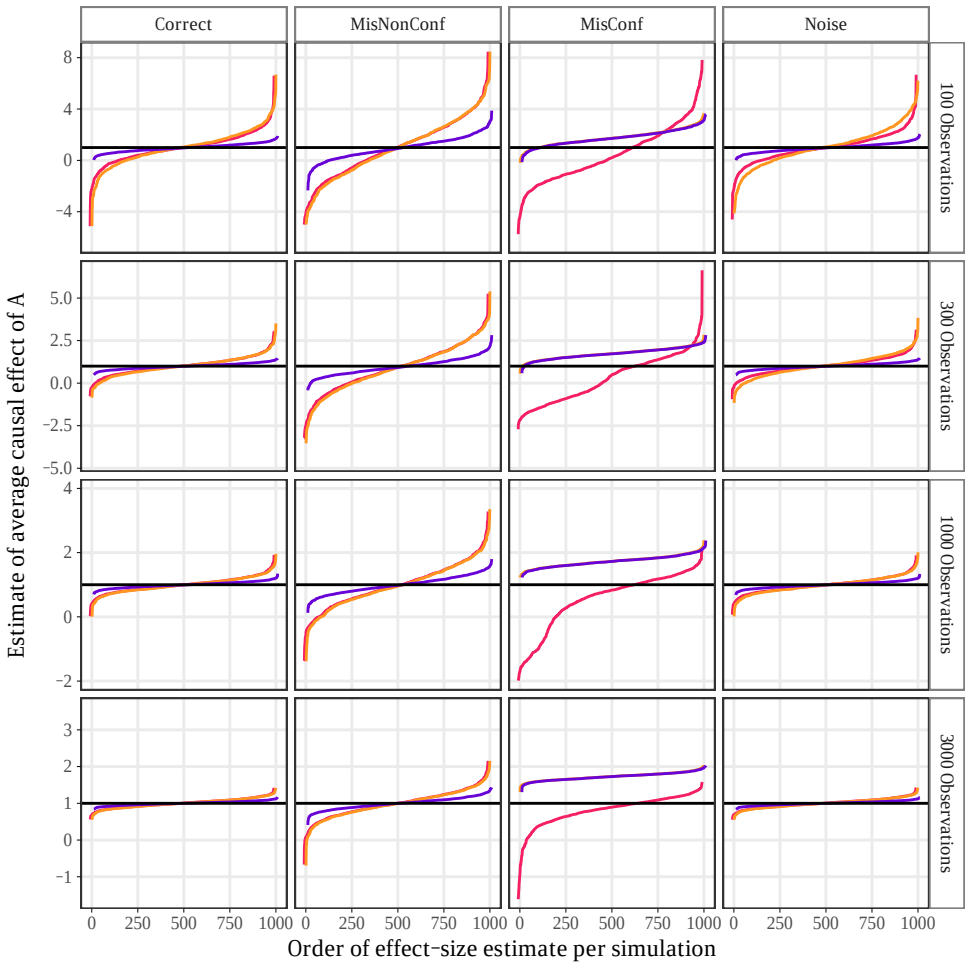
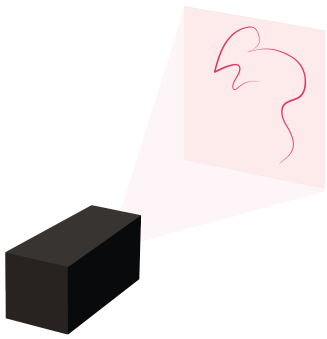


Figure S.5.A.2: Point estimates of the average causal effect of A, ordered from small to large, estimated by TMLEmodel (pink), SuperTMLE (orange), and LR (purple) with Correct, MisNonConf, MisConf, and Noise treatment assignment models on unbalanced data with 100, 300, 1000, or 3000 observations. For visibility, curves of TMLEmodel are shifted to the left and curves of SuperTMLE are shifted to the right



6

General Discussion

As soon as I have got flying to perfection,
I have got a scheme about a steamengine

Ada Lovelace

6.1. Introduction

Automated home cage experiments have been proposed as a solution to some of the pitfalls of the classical tests. The data from automated home cage experiments is fundamentally different from the data from most of the classical tests. Data from automated home cage experiments can have numerous response variables that do not necessarily quantify a meaningful ethological concept. The aim of this thesis was to explore, apply, and expand on methodology for better analysis of data from automated home cage experiments. This general discussion first provides an overview of how the results from earlier chapters contribute to this aim. Thereafter it highlights some topics of interest relating to statistical analysis of automated home cage experiments and the future work in this area.

6.2. Results of the thesis

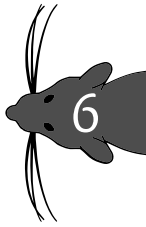
Automated home cage systems can produce large sets of behavioural response variables which might not be directly interpretable by the human observer. Analysing all these response variables using univariate methods such as generalized linear models or mixed models results either in analysis of only one or a few response variables (while disregarding the others) or in the analysis of a large collection of models for highly correlated variables. Neither of these options is desirable. Data with similar characteristics as data from automated home cage experiments is found in fields such as microbiology, (aquatic) ecology, and toxicology. In these fields, multivariate methods such as Redundancy Analysis (RDA) and Principal Response Curves (PRC) are commonly used for analysis. The validity of these methods has been widely accepted and they are available from multiple sources (including open source and free of cost options).

In Chapter 2, the use of these two established methods was introduced for the analysis of automated home cage experiments using two case studies. In these case studies, the multivariate methods replicated the results of the more traditional univariate analyses. As an additional advantage, RDA and PRC visualize the treatment effect of the complete set of response variables in easily interpretable plots. These plots also visualize the relation between response variables. Correlations are visualized in RDA plots and relative importance is visualized in PRC. These methods also provide the potential to find secondary response patterns in the data although this could not be illustrated in the case studies presented here as no such pattern was present.

In Chapter 3, a newly developed protocol for response variable selection was presented for PRC. This extension was demonstrated on a case study from aquatic ecology because of the frequent use of PRC in this field. The idea to develop a response variable selection methodology however arose from the application in automated home cage experiments. The potential to decrease the set of response

variables to only those that are necessary to describe the behavioural response was thought to aid the interpretation of that response. Unfortunately, the set of response variables analysed using PRC in Chapter 2 could not be reduced using the developed selection protocol. This implies that all of the response variables are linked to the main response pattern to the treatment. Because the response variables in this set were pre-selected to describe activity this result is not unexpected. In data sets with response variables relating to different behavioural and physiological domains we *would* expect to see an effect of response variable selection that could for instance result in different sets of selected response variables depending on the treatment.

Technological advances have made it possible to integrate response variables from automated home cages with response variables from other sources. This integration causes interesting new challenges. In Chapter 4, we describe and analyse results from an experiment on social interaction between pair of rats. The hypothesis posed in this experiment could not be tested using a standard statistical analysis. Chapter 4 demonstrated that a combination of statistical techniques can provide new insight into the mechanisms underlying animal behaviour. The response variables in the case study were an activity parameter per individual and the number of ultrasonic vocalisations per pair of rats. In the first part of Chapter 4 it was shown that a standard generalized linear model would not allow for accurate estimation of the vocalisation rate per animal. It was also shown that a composite link model, a simple and elegant extension to the link function in a generalized linear model, was suitable for this task. In the second part of Chapter 4, the analysis of the data from the case study was combined with a second simulation study which supported the hypothesis that animals that can interact, behave truly different from animals that cannot. The underlying model that links activity and USVs was structurally different between individually housed animals and pair-wise housed animals.



Chapter 4 is an example of the power of simulation studies for statistical inference. In Chapter 5 simulation studies were used in a more traditional setting: to evaluate the performance of a statistical technique under a set of circumstances. The performance of Targeted Maximum Likelihood Estimation (TMLE) was evaluated, a promising method that promises causal effect estimates even from observational data. This method was expected to be robust to confounding covariates. Robustness to confounding is a very interesting characteristic in the context of automated home cage experiments because environmental confounding is an important issue in behavioural data. Unfortunately, TMLE proved to be very sensitive to unobserved confounding in this simulation. The effect was most apparent for smaller sample sizes. The conclusion thus was that the sample sizes neces-

sary to obtain unbiased estimates from TMLE were well beyond the reasonable for animal experiments.

Throughout this thesis, methods have been proposed and demonstrated that allow for direct use in the field. This thesis thus fulfils its aim of expanding the statistical toolbox for the use in automated home cage experiments. In the subsequent sections of this general discussion some important topics regarding statistical analysis of automated home cage experiments are discussed.

6.3. Sample size reduction

6.3.1. Replacement, reduction, refinement

Animal experiments for scientific purposes are necessary, yet raise societal and ethical concerns. The use of animals for scientific experiments within the European Union is regulated based on the concepts of Replacement, Reduction, and Refinement (European Union 2010, Directive 2010/63/EU). Animal experiments can only be performed provided that they comply with all three conditions.

Replacement: There are no scientifically satisfactory methods available that can replace the animal experiment

Reduction: The number of animals is reduced to a minimum without compromising the objectives of the project

Refinement: Suffering, pain, distress, and lasting harm are avoided or reduced to a minimum in the animal experiment as well as during breeding, housing, and animal care.

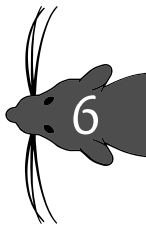
Compared to classical tests, automated home cage experiments contribute to the aim of Refinement because they require less of the stressful animal handling and transport. It has also been argued that the use of automated home cage experiments will help Reducing the number of animals because they allow for more precise observations per animal. Automated home cage experiments can increase the precision of estimates per animals because they accommodate for experiments of longer duration and the incorporation of baseline behaviour. For instance in the first case study in Chapter 2, both the Control and the CNO treatment were given three times to each animal and behaviour was monitored for three hours, starting one hour before the treatment. As a result, we were able to correct the analysis for baseline differences in activity between animals and able to detect that the effect-size of the CNO-treatment varied widely between animals. Such a set-up would be extremely labour intensive using human observers and an Open Field Test.

6.3.2. Improved precision of estimates

The precision of the estimate per animal is one of the three major factors that determine the power of an experiment. The power, or the probability to correctly reject the null-hypothesis, is also determined by the statistical significance criterion and the magnitude of the effect of interest. When estimating average causal effects in automated home cage experiments, the precision of the estimate is influenced by 1) the variability of observed response variable within animals, 2) the variability of the observed response variable between animals, 3) the sample size, and 4) the experimental design.

A toy example that quantifies these effects is given in Box 6.1. In brief, *ceteris paribus*, increasing the number of animals always increases the precision of an average causal effect estimate and thus the statistical power of an experiment. Increasing the precision of observations on an individual also increases the statistical power. The effect of improved precision per animal on the power of the experiment is largest when the within-animal variability is large compared to the population variability. One could argue that an increase of precision on observations on individuals allows for a reduction of the number of animals without a loss of statistical power. Sample sizes in animal experiments however are typically already very small (6-8 animals per group) and reducing these numbers even further would create vulnerability to unexpected events such as technical malfunction and animal drop-out due to disease or death.

As becomes apparent from the toy example, the variance of an average causal effect estimation depends both on the variance of the estimates on the individual level and the population level. To obtain the same statistical power, an increase in precision of estimates on the individual level thus allows for a larger variation in the population. Increased heterogeneity in populations has been proposed as a means to increase extrapolability of results of animal experiments (Wurbel 2000, 2002). For instance by allowing more diverse genetic backgrounds, enriched housing, and variation in age and sex.



6.4. Description of behaviour

6.4.1. Behavioural categories

In section 6.3 we reflected on the precision of measuring a single response variable. Behaviour however, is not directly translated into one response variable. Automated home cage experiments allow for a detailed description of behaviour using multiple response variables. As mentioned in the Introduction of the thesis, a more precise description of animal behaviour has been shown to increase replicability (Kafkafi et al. 2005; Kafkafi, Lipkind, et al. 2003; Kafkafi, Pagis, et al. 2003).

In this thesis we have analysed data sets in which behaviour has been defined in term of the location of the animal. We define behaviour in terms of movement

Box 6.1: Sample size, number of observations, and causal effect estimates

Toy example Let us suppose we wish to estimate the difference between two treatments on response variable Y. An experiment is performed on two treatment groups of six animals each. We assume that the baseline value for Y varies per animal and that there is day-to-day variation in the behaviour of animals. The best possible estimator for the treatment effect on Y is the Average Treatment Effect (ATE): the difference between the average of Y in the Treatment group and the Control group.

$$\hat{Y}_{ij} = \mu + x_i + a_j \text{ and } Y_{ij} = \mu + x_i + a_j + \varepsilon_{ij},$$

where $\hat{Y}_{ij}$ is the expected and Y_{ij} is the observed value for Y for treatment i on individual j, with $i = 0$ for the Control group and $i = 1$ for the Treatment group and with j is 1, ..., 12;
μ is the population average of the baseline level for Y (set to $\mu = 0$);
x_i is the treatment effect (set to $x_0 = 0$ and $x_1 = 1$);
a_j is the deviation from μ for individual j (set to $a_j \sim \mathcal{N}(\text{mean}, \sigma^2) = \mathcal{N}(0, 0.5)$;
and ε_{ij} is the deviation between the observed and expected value of Y (set to $\varepsilon_{ij} \sim \mathcal{N}(0, 0.5)$).

Variance of ATE In our example we assume that the observations of Y are drawn from two normal distributions: $Y_{0j} \sim \mathcal{N}(0, 1)$ and $Y_{1j} \sim \mathcal{N}(1, 1)$. The variance underlying these distributions is the sum of the variance of a_j and ε_{ij}: the random difference between individuals and the random difference within individuals.

The observed average within a treatment group ($\bar{Y}_i$) has a variance of $1/n = 1/6$, the variance of one observation divided by the number of observations. The ATE ($\bar{Y}_1 - \bar{Y}_0$) thus has a variance of 2/6, the sum of the variances of both estimates.

The practical implication is that if we perform the experiment as described here, we are 90% confident we will observe an ATE between 0.05 and 1.95.

Continued on next page

Box 6.1: Sample size, number of observations, and causal effect estimates

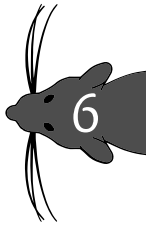
Increase the sample size Increasing the number of animals is not a desirable solution because of ethical and practical reasons. The estimate of the ATE however, will improve by increasing the number of animals. If we increase the number of animals per group from six to twelve, the variance of $\bar{Y}_0$ and $\bar{Y}_1$ decreases from 1/6 to 1/12. It follows that the variance of the ATE decreases from 2/6 to 2/12. If we perform the experiment as described with the double amount of animals per treatment group, we are 90% confident we will observe an ATE between 0.33 and 1.67.

Paired samples Rather than using two treatment groups it is, for some experimental treatments, also possible to give the Control and the Treatment to the same animals. The treatment effect per animal is now estimated as $Y_{dj} = Y_{1j} - Y_{0j} = 1 + \varepsilon_{1j} - \varepsilon_{0j}$ which has a variance of 1. It follows that the variance of the ATE decreases from 1/6 to 1/12. In this simplified example, the effect of administering both treatments to each of the animals thus has the same effect on the variance of the ATE as doubling the sample size per group.

Decrease within-animal variation Automated home cages allow for more precise observations per animal. For instance via prolonged observation times and inclusion of baseline behaviour. In our toy example, we simulate the effect of more precise estimates by increasing the number of observations per animal from one to four and keep using six animals per group.

The value of $\hat{Y}_{ij}$ will be the same for each of the k (with k is $1, \ldots, 4$) repetitions per animal. The estimate for Y_{ij} however, is now the average of 4 repetitions per animal. For simplicity we make the unrealistic assumption that ε_{ijk} within an animal is independent. In that case the the variance of ε_{ij} is decreased to $0.5/4 = 0.125$.

It follows that the variance of $Y_{ij} = 0.5 + 0.125 = 0.625$ and thus that the variance of $\bar{Y}_0$ and $\bar{Y}_1$ decreases from 1/6 to 0.625/6. The variance of the ATE decreases from 2/6 to 1.25/6. If we perform the experiment as described with a more precise estimate per animal, we are 90% confident we will observe an ATE between 0.25 and 1.75.



and immobility using parameters such as Duration Lingering and Distance Moved per time interval. An alternative approach, which has become increasingly feasible during the time this PhD-project was executed, is to categorise behaviour not only in intervals of stopping, lingering and progressing but also in for instance eating, sniffing, and grooming.

Adding response variables such as Frequency Grooming and Duration Eating to the data sets used in this thesis would not fundamentally change their analysis.

6.4.2. Time Scales

The time intervals used in the data sets in this thesis were all chosen in advance and did not vary within a trial. Choosing an appropriate time scale is important. If a time interval is too short not all behavioural categories are observed in each interval. For instance, the response variable Mean Velocity Lingering cannot be calculated for a time interval in which no Lingering was observed. If a time interval is too long, subtle effects of short duration could go unnoticed. Random variability between time intervals will be larger for shorter time intervals and lower for long time intervals.

Defining the length of time intervals in advance is not necessary from a technological point of view. If the raw data of the automated home cage trial is stored, response variables can be calculated time and time again using different settings. It is thus technologically possible to select the optimal duration of a time interval based on the results of the trial. In addition, there is no need for the duration of the time intervals within a trial to be equal. Smaller time intervals could be used for the times of day that the animals are most active and that behavioural effects of the treatment are most apparent and longer time intervals could be used for times of day when the animals are inactive.

Future research is necessary to explore the best practices for implementing *post hoc* dynamic time scales. Defining crucial parameters of an experiment, such as length of a time interval, based on the results of that same experiment can have detrimental effects for the reliability of the conclusions from that experiment.

6.4.3. Location Based versus time-to-event variables

The optimal choice of time scale remains an open challenge. An alternative approach is to evade using time scales altogether. The location based response variables used in this thesis, are based on the (changes in) location of the animal measured on a near-continuous scale. The emphasis is on how the animal distributes its behaviour over time and less on the order and duration of behavioural events. For instance, the total Distance Moved within a time interval is recorded.

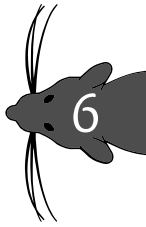
An alternative approach is to describe behaviour as a sequence of events. For instance: Eat (15 seconds)- Progress (5 seconds) - Groom (12 seconds) - Linger (25

seconds). In such a sequence of events the emphasis is not on the total time budget of the animal but on the order and duration of individual events. The order of behaviour and the duration of individual segments can carry valuable information. The analysis of time-to-event data requires a fundamentally different statistical approach compared to continuous time-interval data. Techniques such as (multivariate) survival analysis have potential in this alternative approach.

In this thesis, a sequence of events approach was not an obvious choice because the behavioural categories were limited to Stopping, Lingering, and Progressing with very little variation in order. Stopping segments were by definition always followed by a Lingering or Progressing segment and *vice versa* because the raw data were first classified into Stopping and movement segments, and thereafter the movement segments were divided into Lingering and Progressing.

6.4.4. Analysing Multiple Response Variables

The total number of response variables that can be calculated from automated home cage experiments is near infinite. In Chapter 2, the advantages of multivariate over univariate statistics for automated home cage data with multiple response variables has been discussed at length. Analysing multiple correlated response variables in separate analyses is problematic because of increased difficulty of interpretation and because corrections for multiple testing are necessary which costs statistical power. In addition, it also provides a fundamental problem. In Automated home cage experiments, response variables are typically highly correlated. This indicates that a statistically significant difference that occurred by chance or artefact in one response variable will re-appear in multiple response variables. A normal correction for multiple comparisons (*e.g.* a Bonferroni correction) does not account for this issue. It is trivial that multiple observations from one animal should not be analysed as if it were observations from multiple animals, it should also be trivial that multiple response variables from one experiment should not be analysed if it were response variables from multiple experiments.



Interpretation of large sets of response variables is more complicated compared to smaller sets of single response variables which makes reducing the number of response variables desirable. To prevent data dredging (*i.e.* selecting response variables based on observed effects rather than hypotheses) a response variable selection approach should be decided on prior to the experiment. In Chapter 3, an example of a *post hoc* response variable selection protocol based on permutation testing was proposed for Principal Response Curves. The advantage of *post hoc* response variable selection versus *a priori* selection is that the full set of response variables can be used to describe the overall effect of the treatment

and that response variables. This approach is thus also suitable when the effects of a treatment could be unexpected.

6.5. Machine Learning and bioinformatics

6.5.1. Big Data

Raw data from Automated home cage experiments requires undoubtedly more computer memory to store than data from classical tests. The frame-by-frame information on the exact position of the animal within the cage is large in size but not necessarily rich in information. Even after restructuring the frame-by-frame data into behavioural response variables the data sets are still considerably larger than the typical results of classical tests. Sets of over a hundred response variables have been reported (*e.g.* Loos et al. 2015). Data from Automated home cage experiments could thus be regarded as "Big Data". This makes analysing these data using machine or statistical learning techniques seem like an obvious choice. Lead authors in behavioural phenotyping have suggested the use of such techniques, referred to as "bioinformatics tools" and "pattern analysis"(Gerlai 2002) or "artificial intelligence" (Spruijt and Visser 2006).

In machine learning, computers are not specifically programmed for a task but learn from the data. These algorithms can be trained to make predictions, cluster, or classify data. Machine learning models are referred to as black box models because the decision making process is generally incomprehensible to humans. In supervised learning, a machine learning algorithm is introduced to a set of data with correct labels or outcomes and used the information from this "training set" to learn how to make predictions for or classify on data sets without labels. Such techniques have been applied successfully to data from automated home cage experiments. For instance for the classification of animal behaviour into categories (Dam et al. 2013; Hong et al. 2015; Jhuang et al. 2010). With respect to behavioural phenotyping, is has been shown that machine learning algorithms such as support vector machines can distinguish between strains of rats and assign rats unknown to the algorithm to the correct strain based on response variables from the PhenoTyper®(Fekas et al. 2010).

The open question with respect to black box models is how valuable these techniques are from a behavioural phenotyping perspective. After all, the aim of statistical analysis in animal behavioural experiments is provide numerical foundation for ethological theories (Spruijt and Visser 2006). Building an algorithm that can distinguish between strains of rats based on a set of response variables proves that strains of rats differ in terms of these response variables. But of what use is that knowledge if we cannot comprehend *how* this distinction is made?

To fulfil the aim of providing a numerical foundation for ethological theories hypothesis driven statistical techniques are necessary. In Chapter 5 we have

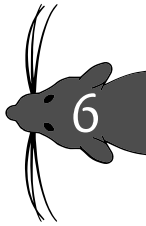
looked into Targeted Maximum Likelihood Estimation, a method for causal effect estimation that incorporates machine learning. Unfortunately we had to conclude that this method has some serious pitfalls. The method is sensitive to unobserved confounding variables which are hard to control in animal experiments. In addition the sample sizes necessary to implement such methods successfully are well beyond what would be reasonable for automated home cage experiments.

6.5.2. Potential for Machine learning

Whilst machine learning techniques will not help us gain insight into animal behaviour directly, these techniques can still be useful in other stages of animal experiments.

Data generation and pre-processing

Machine learning alghoritms, as mentioned earlier, have been successfully trained to classify behaviour into pre-defined categories and have obtained error-rates close to that of human observers (Dam et al. 2013; Hong et al. 2015; Jhuang et al. 2010). As described in section 6.4.1, classifying behaviour into categories could result in interesting sets of new response variables. In addition to learning to do tasks that human observers can do, machine learning algorithms can also learn to do what humans cannot. The case study described in Chapter 4 was complicated to analyse because vocalisations could not be assigned to individual rats based on the available data. If techniques keep developing at the current rate, in the future we might be able to train a computer model to detect which rat is the vocalising individual based on the video-images or audio profiles.



Monitoring and detection of deviant behaviour

In unsupervised machine learning, in contrast to supervised machine learning, the algorithms do not require the user to provide correct answers for learning. Unsupervised learning algorithms can for instance cluster similar observations together and thus create new classes in which behaviour can be classified. One-class algorithms can define which observations are "normal" (*i.e.* in the class) and which are "abnormal" (*i.e.* outside of the class) (Lian 2012). Such a clustering system has the potential for grouping individuals which exhibit similar behaviour and, perhaps more applicable, pointing out individuals which exhibit "abnormal" behaviour. Unsupervised learning algorithms have been successfully applied to detect for instance outlier plant varieties (Dijk et al. 2014). Outlier behaviour in laboratory animals could indicate discomfort, illness, or stress. These types of models thus have the potential to, combined with the automated continuous monitoring via the automated home cage system, be developed into an animal welfare monitor.

Online machine learning algorithms are able to incorporate new data into their predictions as the data becomes available. This allows the algorithms to change their decision making process over time as circumstances change. Building on the concept of an animal welfare monitoring system as proposed in the previous paragraph, an updating would be especially suitable for use in the study of progressive disease (*e.g.* Robinson et al. 2013). In such long-term studies it is important to determine humane end-points (*i.e.* the point at which the suffering of the animal no longer outweighs the benefits of the trial) whilst accounting for the fact that what constitutes as "normal" changes over time.

6.6. Simulation studies

Automated home cage experiments have been designed as a better method for collecting data. In this thesis we have focussed on better methods for analysing these data. In this section we focus on the "creation" of new data without experimenting.

6.6.1. Permutation testing and Bootstrapping

Bootstrapping and permutation testing are methods in which data from an existing data set is re-used. A statistical analysis is repeated using a different random sample from the data set with (bootstrapping) or without (permutation testing) replacement. The parameters of interest that are estimated using the statistical tests will show slightly different results for each sample. The added value of bootstrapping and permutation testing is that the uncertainty distributions of these parameters can be obtained.

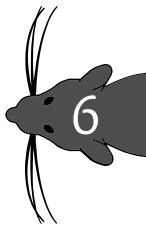
Bootstrapping is mainly used to estimate properties such as the variance of the parameter (*e.g.* treatment effect). In this thesis, bootstrapping has been applied in Chapter 5 to improve confidence intervals of TMLE. Another example of its use is to calculate confidence intervals for PRC, as confidence intervals for PRC cannot be obtained analytically and thus rely on bootstrapping (Timmerman and Braak 2008).

Permutation testing is mainly used for hypothesis testing. The parameter estimate when using the original data set is compared to the distribution of parameter estimates using the permuted data sets (including the original). It is routinely used in RDA and PRC procedures (such as in Chapter 2) for instance to determine significance of axes. In Chapter 3 we propose a permutation testing protocol for Response Variance Selection in PRC. Restricting permutation, for instance such that observations are only permuted within animals, corrects for differences between animals without explicit incorporation of these effects in the statistical model.

6.6.2. Simulation studies

In contrast to the methods described in the previous section, simulation studies do not rely on behavioural experiments for data at all. Data is created using computers solely. This allows for *in silico* experiments. Simulation studies have several useful purposes. In Chapter 3, Chapter 4, and Chapter 5 we have simulated data with known properties to evaluate the performance of statistical methods. The advantage of using simulation studies is that all the properties of the data are known and predetermined. This allows us to test the statistical methods under the desired circumstances and to evaluate the precision of its estimates and validity of its conclusions under those circumstances. In Chapter 3 simulation studies were used to develop a novel method. We selected the optimal protocol amongst several options. In Chapter 4 simulation studies were used to show that, given a set of assumptions on the data, a composite link function approach would perform better than some alternatives. In Chapter 5 we used simulation studies to evaluate the performance of an existing method under circumstances that are relevant in the context of automated home cage experiments. We evaluated for instance what the effect of small sample sizes and unobserved confounding variables would be on the precision and accuracy of effect size estimations.

Simulation studies can also be used for hypothesis testing. In Chapter 4, a second simulation study was used for this purpose. The simulation studies were used to demonstrate that the results of an animal experiment were unlikely given a certain set of assumptions. The logical implication is that the set of assumptions is thus unlikely to be true. In situations for which no formal statistical tests exist (such as comparing different statistical models) simulations studies provide a valuable alternative approach for statistical inference.



6.7. The Future of biostatistics in behavioural phenotyping

A number of trends in complexity of behaviour experiments can be seen. Behavioural data has greatly increased in complexity compared to the times of the first Open Field Trial by (Hall 1934; Hall and Ballachey 1932). Experimental designs will increase in complexity and animal welfare regulations restrain the number of animals researchers can use. Data sets grow larger as automated home cage experiments can yield many response variables (*e.g.* Loos et al. 2015). Data from other sources such as heart rate, body temperature, and ultrasonic vocalisations is also increasingly being incorporated (*e.g.* Aziriova et al. 2016; Peters et al. 2017). With many more sensors in home-cages, each giving a data-stream, another great challenge in deriving meaningful descriptors from these data-streams. Biostatistics and machine learning should contribute here.

This seeming complexity should not confuse the statistician into analysing the data using overly complex tools. The small sample sizes in animal experiments

do not allow for overly complicated methods. Chapter 5 of this thesis serves as a cautionary tale. I believe that the multivariate statistics provided in Chapter 2 and the Composite Link Model in Chapter 4 are examples of a compromise between more advanced analysis whilst keeping the results interpretable. In the field of animal behaviour, the potential for direct interpretation of results is of the highest importance as the aim of behavioural phenotyping is to describe and qualify behaviour.

The potential for implementation of machine learning techniques has been highlighted in section 6.5. The incorporation of machine learning techniques for behavioural classification and identification of individuals is promising. It will also result in new types of challenges. Another important open issue with regard to data analysis of automated home cage experiments is time-to-event data. It would be interesting to explore the potential of multivariate survival analysis for this type of data.

In light of the developments and incorporation of new methodology and analysis, the important role of biostatisticians is to stay critical. As demonstrated in Chapter 5, not all promising innovations are suitable for use in every circumstance. Use of novel techniques can provide great benefits but also brings about more uncertainty. Use of simulation studies is an important tool to check the suitability of a method under a situation and the check assumptions on the data.

Another important task is to refrain from *ad hoc* analysis and cherry picking results and develop strategies for data analysis prior to data exploration. An example of a strategies for *post-hoc* response variable selection is provided in Chapter 3. The need for pre-determined data analysis strategies increases when data-based parameters are used for the analysis, such as the dynamic time scales referred to in section 6.4.2. Such parameters have great potential for improving the efficiency of data analysis but can also harm the reliability of the conclusions if not handled with care.

In short, with the innovations in the field of behavioural phenotyping the role of the biostatistics is growing. More complicated data requires more preprocessing, the procedure of which must be evaluated in view of the aim of the experiment and the final statistical analysis, which may not necessarily be more complicated. All of this requires in depth knowledge on the methodology and its underlying assumptions. The role of biostatistics is to help behavioural and neuroscientists to translate the results from these experiments into comprehensible conclusions that drive the field forwards.

References

Aziriova, S., K. Repova, K. Krajcirovicova, T. Baka, S. Zorad, V. Mojto, P. Slavkovsky, et al. 2016. “Effect of ivabradine, captopril and melatonin on the behaviour

of rats in L-nitro-arginine methyl ester-induced hypertension." 67 (6): 895–902.

Dam, Elsbeth A.van, Johanneke E. van der Harst, Cajo J F ter Braak, Ruud A J Tegelenbosch, Berry M. Spruijt, and Lucas P J J Noldus. 2013. "An automated system for the recognition of various specific rat behaviours." *Journal of Neuroscience Methods* 218 (2): 214–224.

Dijk, Jeroen P. van, Carla Souza de Mello, Marleen M. Voorhuijzen, Ronald C.B. Hutten, Ana Carolina Maisonnave Arisi, Jeroen J. Jansen, Lutgarde M.C. Buydens, Hilko van der Voet, and Esther J. Kok. 2014. "Safety assessment of plant varieties using transcriptomics profiling and a one-class classifier." *Regulatory Toxicology and Pharmacology* 70, no. 1 (October): 297–303.

European Union. 2010. *Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes.* Technical report.

Fekas, Dimitris, Nieuwe Kanaal, Raymond C de Heer, and Berry M Spruijt. 2010. "Machine Learning Techniques for Parameter Selection and Automated Behavioral Classification." In *Measuring Behavior 2010,* edited by Andrew Spink, Fabrizio Grieco, Olga Krips, Leanne Loijens, Lucas Noldus, and Patrick Zimmerman, 2010:171–173. Eindhoven, The Netherlands.

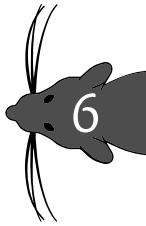
Gerlai, Robert. 2002. "Phenomics: fiction or the future?" 25, no. 10 (October): 506–509.

Hall, C. S. 1934. "Drive and emotionality: factors associated with adjustment in the rat." 17 (1): 89–108.

Hall, C., and E. L. Ballachey. 1932. *A study of the rat's behavior in a field. A contribution to method in comparative psychology.,* 1–12.

Hong, Weizhe, Ann Kennedy, Xavier P. Burgos-Artizzu, Moriel Zelikowsky, Santiago G. Navonne, Pietro Perona, and David J. Anderson. 2015. "Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning." *Proceedings of the National Academy of Sciences* 112 (38): E5351–E5360.

Jhuang, Hueihan, Estibaliz Garrote, Jim Mutch, Xinlin Yu, Vinita Khilnani, Tomaso Poggio, Andrew D Steele, and Thomas Serre. 2010. "Automated home-cage behavioural phenotyping of mice." 1 (January): 68.



Kafkafi, Neri, Yoav Benjamini, Anat Sakov, Greg I Elmer, and Ilan Golani. 2005. "Genotype-environment interactions in mouse behavior: a way out of the problem." 102 (12): 4619–4624.

Kafkafi, Neri, Dina Lipkind, Yoav Benjamini, Cheryl L Mayo, Gregory I Elmer, and Ilan Golani. 2003. "SEE locomotor behavior test discriminates C57BL/6J and DBA/2J mouse inbred strains across laboratories and protocol conditions." 117 (3): 464–477.

Kafkafi, Neri, Michal Pagis, Dina Lipkind, Cheryl L. Mayo, Yoav Bemjamini, Ilan Golani, and Gregory I. Elmer. 2003. "Darting behavior: a quantitative movement pattern designed for discrimination and replicability in mouse locomotor behavior." 142, nos. 1-2 (June): 193–205.

Lian, Heng. 2012. "On feature selection with principal component analysis for one-class SVM." *Pattern Recognition Letters* 33, no. 9 (July): 1027–1031.

Loos, Maarten, Bastijn Koopmans, Emmeke Aarts, Gregoire Maroteaux, Sophie van der Sluis, Matthijs Verhage, and August B. Smit. 2015. "Within-strain variation in behavior differs consistently between common inbred strains of mice." *Mammalian Genome* 26, nos. 7-8 (August): 348–354.

Peters, Suzanne M., Joe A. Tuffnell, Ilona J. Pinter, Johanneke E. van der Harst, and Berry M. Spruijt. 2017. "Short- and long-term behavioral analysis of social interaction, ultrasonic vocalizations and social motivation in a chronic phencyclidine model." 325:34–43.

Robinson, Lianne, Andrea Plano, Stuart Cobb, and Gernot Riedel. 2013. "Long-term home cage activity scans reveal lowered exploratory behaviour in symptomatic female Rett mice." 250 (August): 148–156.

Spruijt, Berry M., and Leonie de Visser. 2006. "Advanced behavioural screening: automated home cage ethology." 3, no. 2 (June): 231–237.

Timmerman, Marieke E., and Cajo J.F. ter Braak. 2008. "Bootstrap confidence intervals for principal response curves." *Computational Statistics & Data Analysis* 52, no. 4 (January): 1837–1849.

Wurbel, H. 2000. "Behaviour and the standardization fallacy." *Nature genetics* 26 (3): 263.

Wurbel, H. 2002. "Behavioral phenotyping enhanced - beyond (environmental) standardization." *Genes, Brain and Behavior* 1, no. 1 (January): 3–8.

Summary

Automated home cage experiments have been proposed as an alternative to the classical tests used for behavioural phenotyping. As the name implies, automated home cage experiments are conducted in home cage environments and the behaviour is recorded automatically. The experiments can thus be conducted without human interference and can last for several days.

All data incorporated in this thesis is collected using a PhenoTyper® system (Noldus Information Technology, Wageningen, The Netherlands). The PhenoTyper is a home cage environment with an integrated top-view camera. The exact location of the rat or mouse is determined for every frame in the video. Behavioural response variables such as Distance Moved or Duration Progressing are extracted from the location.

Data from automated home cage experiments typically consists of multiple response variables that can be highly correlated. In addition to the location-based activity response variables, automated home cage environments have the potential to incorporate data from other sources such as biometric parameters.

The aim of this thesis is to expand the methodology available to analyse these data.

In Chapter 2, the use of multivariate statistics for data from automated home cage experiments is demonstrated in two case studies. Data from automated home cage experiments is pre-dominantly analysed using univariate statistics in which the significance and magnitude of the effect of a treatment on a single response variable is tested. By analysing single response variables the benefit that automated home cage experiments allow for the collection of numerous response variables simultaneously is not fully utilized. The use of multivariate statistics allows for simultaneous analysis of multiple response variables. The multivariate methods described in Chapter 2 are Redundancy Analysis (RDA) and Principal Response Curves (PRC). Both of these methods are frequently used in (aquatic) ecology, toxicology, and microbiology. RDA is a constrained form of Principal Components Analysis (PCA). RDA describes the underlying structure of a data set in terms of the explanatory variables (such as experimental treatment). It quantifies the proportion of variance in the data set that can be described using these explanatory variables. PRC is a special case of RDA used to describe experimental multivariate longitudinal data. It estimates differences among treatments on a collection of response variables over time and the extent to which the response of those individual response variables resembles the overall response. In both case studies, the multivariate analyses were able to draw the same main conclusions as the contrasting univariate analyses. The advantages of using a multivariate analysis rather than a univariate analysis on a single response variable is that the multivariate methods provide a graphical representation of the data set, are easy to interpret, and allow for estimation of the relation between response variables.

In Chapter 3, a novel extension to PRC is presented that allows for response variable selection using permutation testing. Often, not all of the response variables included in PRC are affected by the treatment which can make response variable selection desirable. One approach is to use a straightforward cut-off value for coefficient size. Because coefficient size of response variables are affected by more factors than effect-size alone, results of this approach can be variable between data sets. A backward selection approach was expected to give a more robust result. Four backward selection approaches based on permutation testing were presented. The approaches differ in whether coefficient size is used or not in ranking the response variables to test. The performance of these approaches was demonstrated in a simulation study using a well known data set in the field of aquatic ecology. The permutation testing approach that uses information on coefficient size of RVs sped up the algorithm without affecting its performance. This most successful permutation testing approach removed roughly 95% of the response variables that are unaffected by the treatment irrespective of the characteristics of the data set (which is a desirable property of a statistical test) and, in the simulations, correctly identified up to 97% of response variables affected by the treatment.

In Chapter 4, a case study is used to illustrate the power of combining mechanistic and statistical modelling, and the benefits of simulation studies. In this case study, an integrated analysis of two streams of information: activity response variables per rat and Ultrasonic Vocalisations (USVs) per cage (containing a pair of rats). USVs are crucial in the social behaviour of rats. The aim of the first part of the chapter was to develop methodology to predict the USV-rate of the pair of rats as a function of the activity of the individuals. A mechanistic model is that the USV-rate of the pair of rats is the sum of the USV-rates of the two individuals depending on their own behaviour ("sum-of-rates" model). It turns out that this "sum-of-rates" model can be fitted to data using a Composite Link Model (CLM) approach. In generalized linear models (GLM) the individual's USV-rates are multiplied rather than summed. A simulation study verified that CLM gave a better fit (lower Poisson Deviance) than GLM. In the second part of the chapter, data from an experiment in which half of the cages did allow the rats of the pair to interact (Pair Housing) and the other half did not (Individual Housing). A number of models was fitted to investigate whether there is evidence that interaction between rats affects their behaviour. The "sum-of-rates" model fit best for Individual Housing and GLM for Pair Housing. This difference in fit supports the hypothesis that interaction between rats affects their behaviour. An additional simulation study strongly suggested that this difference was not due to chance and that the underlying mechanism that links activity and USVs structurally differed between Pair Housing and Individual Housing.

In Chapter 5, a simulation study is described that evaluates the performance of a new and promising statistical learning method under circumstances relevant for automated home cage experiments. Targeted Maximum Likelihood Estimation (TMLE) is a new and promising statistical method for causal effect estimation, even in observational studies, that can use machine learning methods to increase performance. The intended role of TMLE in the analysis of home cage experiments was to account for inter-individual variation in behaviour when testing specific treatment effects. TMLE is a doubly robust method, which means that it is robust to misspecification of either the treatment outcome model or the treatment assignment model. A treatment outcome model predicts the effect of a treatment on the response variable given the covariates. A treatment assignment model predicts the probability that an individual is in a treatment group given the covariates. In theory, when all assumptions are correct, TMLE should thus provide unbiased causal effect estimators even when either the treatment outcome or treatment assignment model is misspecified. When TMLE is applied in practice however, it is possible that these required theoretical assumptions such as the positivity assumption and no unobserved confounders are violated. The simulation study in Chapter 5 illustrates the effects of unobserved (non-)confounding covariates and noise covariates on bias, mean square error, and coverage of TMLE on near-balanced data sets (with low risk of positivity violations) and unbalanced data sets (with higher risks of positivity violations). The conclusion was that TMLE is able to estimate average causal effects with low bias and mean square error, compared to the golden standard linear regression, given that the sample size is large, the data set is near-balanced, and the assignment model is specified correctly. In unbalanced data sets TMLE did not live up to expectations, also in data sets in which the positivity assumption was not violated. The conclusion from the simulation study is that TMLE is as yet not suited for the intended use in home cage experiments.

In Chapter 6, the General Discussion, the main findings of the thesis are summarised and discussed in relation to the aim of the thesis. In addition, several hot topics in biostatistics for automated home cage experiments are discussed.

Samenvatting

Geautomatiseerde thuiskooi-experimenten worden gezien als een alternatief voor de "klassieke testen" die nu gebruikt worden voor gedragsfenotypering (het typeren van geobserveerd gedrag). Zoals de naam al zegt, worden geautomatiseerde thuiskooi-experimenten uitgevoerd in een thuiskooi en wordt het gedrag automatisch vastgelegd. Daarom kan het experiment meerdere dagen duren en is inmenging van mensen niet nodig.

Alle gegevens die gebruikt zijn in dit proefschrift werden verzameld met behulp van het PhenoTyper systeem (Noldus IT, Wageningen, The Netherlands). De PhenoTyper is een thuiskooi omgeving waarbij een camera die geïntegreerd is in het dak van bovenaf de bewoner filmt. Voor elk beeld in de video wordt de exacte locatie van de rat of muis bepaald. Uit deze locatiegegevens worden gedragsvariabelen zoals "Afgelegde Afstand" en "Tijd besteed aan Rennen" berekend per tijdsinterval.

Een typische dataset uit een geautomatiseerd thuiskooi-experiment bevat veel gecorreleerde gedragsvariabelen. Een geautomatiseerd thuiskooi-experiment biedt ook de mogelijkheid om gegevens uit andere bronnen te integreren zoals biomedische parameters. Het doel van dit proefschrift is het uitbreiden van de methodologie voor het analyseren van gegevens afkomstig van geautomatiseerde thuiskooi-experimenten.

In Hoofdstuk 2 wordt het gebruik van multivariate statistiek om gegevens uit geautomatiseerde thuiskooi-experimenten te analyseren gedemonstreerd aan de hand van twee praktijkvoorbeelden. Gegevens uit geautomatiseerde thuiskooi-experimenten worden tot nog toe voornamelijk geanalyseerd met univariate methoden. In univariate methoden wordt de aanwezigheid van een behandelingseffect en de grootte ervan op slechts één gedragsparameter getest. Eén van de voordelen van geautomatiseerde thuiskooisystemen is dat er meerdere gedragsparameters tegelijkertijd kunnen worden verzameld. Door er slechts één te analyseren maken we dus onvolledig gebruik van het potentieel van de methode.

Het gebruik van multivariate statistiek staat ons toe om meerdere gedragsparameters tegelijkertijd te analyseren. De multivariate methoden die we beschrijven in Hoofdstuk 2 zijn Redundancy Analysis (RDA; NL: Redundantie-Analyse) en Principal Response Curves (PRC; NL: Hoofdreactie curven). Deze beide methoden zijn veelgebruikt in de (aquatische) ecologie, toxicologie en microbiologie. RDA is een begrensde vorm van Principal Components Analysis (PCA; NL: Hoofdcomponentenanalyse). RDA beschrijft de onderliggende structuur van een dataset in termen van de verklarende variabelen (zoals experimentele behandeling). RDA kwantificeert de fractie van de variantie in de dataset die wordt verklaard door de verklarende variabelen.

PRC is een speciaal geval van RDA dat gebruikt wordt om multivariate gegevens verkregen uit experimenten te beschrijven. De methode schat de verschillen

tussen behandelingen op een verzameling gedragsvariabelen over de tijd en schat in hoeverre de reactie van de individuele gedragsvariabelen overeenkomt met het hoofdpatroon. In beide praktijkvoorbeelden kon met de multivariate analyse dezelfde conclusies getrokken worden als met de univariate analyse. Het voordeel van het gebruik van multivariate analyse boven het gebruik van univariate analyse is dat multivariate methoden de data grafisch weergeven in eenvoudig te interpreteren figuren diede onderlinge relatie tussen gedragsvariabelen weergeven.

In Hoofdstuk 3 wordt een nieuwe uitbreiding van PRC gepresenteerd die ons in staat stelt gedragsvariabelen te selecteren door middel van permutatietesten. Selectie van gedragsvariabelen is wenselijk omdat het vaak voorkomt dat niet alle gedragsvariabelen worden beïnvloed door de behandeling. Een mogelijke methode voor het selecteren van gedragsvariabelen is om een grenswaarde vast te stellen en alleen gedragsvariabelen met een coëfficiënt boven de grenswaarde te selecteren. De resultaten van deze aanpak kunnen verschillen tussen datasets omdat de coëfficiëntgrootte van gedragsvariabelen beïnvloed wordt door meerdere factoren. Wij verwachtten dat een selectieprotocol waarbij vanuit de volledige set gedragsvariabelen terug wordt gesnoeid naar een gereduceerde set (achterwaartse selectie) een robuuster resultaat geeft. In het hoofdstuk worden vier achterwaartse selectieprotocollen gepresenteerd die allemaal gebruik maken van permutatietesten. Het verschil tussen deze protocollen zit in de volgorde waarin de gedragsvariabelen worden getest. De prestatie van deze verschillende aanpakken werd gedemonstreerd in een simulatiestudie die is gebaseerd op een beroemde dataset binnen de aquatische ecologie. Het protocol dat informatie over de coëfficiëntgrootte van de gedragsvariabelen gebruikte om de testvolgorde te bepalen was sneller dan de andere protocollen zonder prestatieverlies. Dit meest succesvolle protocol verwijderde in de simulaties ongeveer 95% van de gedragsvariabelen die niet beïnvloed werden door de experimentele behandeling, ongeacht de eigenschappen van de dataset (wat een goede eigenschap is voor een statistische test). Het protocol identificeerde tot wel 97% van de gedragsvariabelen die wel beïnvloed werden door de behandeling.

In Hoofdstuk 4 wordt een praktijkvoorbeeld gebruikt om de toegevoegde waarde van het combineren van mechanistisch en statistisch modelleren en van simulatiestudies te demonstreren. Dit praktijkvoorbeeld is de geïntegreerde analyse van twee informatiestromen: activiteitsgedragsvariabele per rat en Ultrasone Vocalisaties (USVs) per kooi (met daarin twee ratten). USVs zijn onmisbaar voor het sociale gedrag van ratten. Het doel van het eerste deel van het hoofdstuk was om een methode te ontwikkelen om het aantal USVs per tijdseenheid (USV-rate) voor het paar ratten te voorspellen als functie van hun individuele activiteit.

In Hoofdstuk 5 wordt een simulatiestudie beschreven die de prestatie van een nieuwe en veelbelovende "statistical learning" methode evalueert onder omstandigheden die relevant zijn voor geautomatiseerde thuiskooi-experimenten. Targeted Maximum Likelihood Estimation (TMLE; NL: Gerichte schatting van de meest waarschijnlijke schatter) is een nieuwe en veelbelovende methode voor de schatting van oorzakelijke effecten, ook voor observationele studies, die "machine learning" technieken kan integreren om zijn prestaties te verbeteren. De beoogde toegevoegde waarde van het gebruik van TMLE in de analyse van geautomatiseerde thuiskooi-experimenten was om te corrigeren voor gedragsvariatie binnen individuen tijdens het testen voor specifieke behandelingseffecten. TMLE is een dubbel robuuste methode, dat betekent dat het bestand is tegen missspecificatie van ofwel het model dat behandelingsuitkomsten voorspelt ofwel het model dat behandelingstoewijzing voorspelt. Een behandelingsuitkomstmodel voorspelt het effect van een behandeling op een gedragsvariabele rekening houdend met de andere verklarende variabelen. Een behandelingstoewijzingsmodel voorspelt de kans dat een individu een behandeling krijgt rekening houdend met de andere verklarende variabelen.

Theoretisch gezien, als aan alle aannames wordt voldaan, zou TMLE een schatting van het gemiddelde behandelingseffect kunnen geven zonder structurele afwijking (bias), zelfs als het behandelingsuitkomstmodel of het behandelingstoewijzingsmodel niet correct is. Bij gebruik van TMLE in de praktijk is het mogelijk dat de vereiste theoretische aannames niet correct zijn. De simulatiestudie in Hoofdstuk 5 laat zien wat de effecten zijn van onopgemerkt gebleven (verstorende) verklarende variabelen en van ruisvariabelen op de bias, gemiddelde gekwadrateerde fout, en coverage (dekkingsgraad van het betrouwbaarheidsinterval) van TMLE in datasets met laag (bijna-evenwichtige datasets) of hoog (onevenwichtige datasets) risico om een aanname te schenden.

De conclusie was dat TMLE het gemiddelde oorzakelijke effect kan schatten met lage bias en gemiddelde gekwadrateerde fout in vergelijking tot de gouden standaard (lineaire regressie) op voorwaarde dat er een grote steekproef is genomen, de dataset bijna-evenwichtig is en dat het behandelingstoewijzingsmodel correct gespecificeerd is. In onevenwichtige datasets voldeed TMLE niet aan de verwachtingen, ook niet in datasets waar de positiviteitsaanname wel correct was. De conclusie van deze simulatiestudie is dan ook dat TMLE tot nog toe niet geschikt is om te gebruiken in geautomatiseerde thuiskooi-experimenten.

In hoofdstuk 6, de algemene discussie, worden de hoofdbevindingen van dit proefschrift samengevat en bediscussieerd in het licht van het doel van dit proefschrift. Verder worden verschillende actuele thema's die relevant zijn voor biostatistiek voor geautomatiseerde thuiskooi-experimenten bediscussieerd.

PE&RC Training and Education Statement With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (4.5 ECTS)

- Statistical analysis of automated home cage experiments , BMS-ANed PhD Day 2017

Writing of project proposal (4.5 ECTS)

- Statistical modelling of behaviour and behavioural responses to psychopharmaca of rodent strains and rodent models of psychiatric and neurological disease.

Post-graduate courses (6.6 ECTS)

- Pattern Recognition; NBIC (2013)
- PerClass: Machine learning for R&D specialists; PerClass (2012)
- Bayesian Statistics; PE&RC (2012)
- Multivariate Analysis; PE&RC (2012)
- Zero Inflated Models & GLMM with R; PE&RC (2014)

Laboratory training and working visits (0.3 ECTS)

- Behavioural effects of drugs in Rats (2013)

Deficiency, Refresh, Brush-up courses (3 ECTS)

- Introduction to Mathematical Thinking; Coursera.org (2012)
- Learning to Program: The Fundamentals; Coursera.org (2012)

Competence strengthening / skills courses (3.3 ECTS)

- Project and Time Management; PE&RC (2013)
- Scientific Writing; PE&RC (2015)

PE&RC Annual meetings, seminars and the PE&RC weekend (2.7 ECTS)
- PE&RC First years weekend (2012)
- PE&RC Mid-term weekend (2014)
- PE&RC Last years weekend (2015)
- PE&RC PE&RC day (2013)
- PE&RC PE&RC day (2014)

Discussion groups / local seminars / other scientific meetings (7.5 ECTS)
- Modelling and Statistics Network (MSN); WUR (2012-2016)
- PhenoTyper Meetings; Sylics, VU Amsterdam (2012-2015)

International symposia, workshops and conferences (3.2 ECTS)
- Endo-Neuro-Psycho Meeting; oral presentation; Lunteren, The Netherlands (2014)
- IBS Channel Meeting; oral presentation; Nijmegen, The Netherlands (2015)

Lecturing / supervision of practical's / tutorials (6.3 ECTS)
- Statistics 1 (2012-2017)
- Statistics 2 (2012-2017)
- Canoco Introduction (2015)

Acknowledgements

Getting to the heart of things is never easy

Phillipe Starck

"It takes a village to write a PhD-thesis." My name is the only one on the cover of this booklet but I did not do it, nor could I have done it, alone. Many wonderful people have supported me on my journey. I am deeply grateful for all the help I have received along the way. For people sharing their knowledge, for insightful discussions, for practical help, and for emotional support. So many times I have complained about "dat ellendige proefschrift" and all the things that went wrong. And, luckily, so many times I was able to celebrate the things that went right. To paraphrase a Dutch expression: "Shared pain is half the pain, and shared joy is twice the joy".

In addition to this general "Thank you" to all who have supported me, I would like to thank some people in person. Allereerst mijn promotor Cajo. Dankjewel dat je mij de kans hebt gegeven aan dit project te beginnen, bijna op goed geluk, zonder achtergrond in de statistiek. We zijn heel verschillende persoonlijkheden met heel verschillende manieren van werken en we hebben behoorlijk aan elkaar moeten wennen. Gelukkig hebben we altijd kunnen lachen met elkaar, ook als er weer eens dingen helemaal verkeerd gingen. Jij hebt me geleerd vanuit het simpelste geval te beginnen en van daar verder te bouwen. En om eerst een fundering te leggen voor je gaten in het gebouw begint te schieten.

Als tweede bedank ik mijn co-promotor Lia. Lieve Lia, je hebt dit project vlot getrokken. Zonder jouw daadkracht en positieve benadering had dit proefschrift hier nu niet gelegen. Als Cajo en ik plannen bleven bedenken en nieuwe bezwaren bedachten riep jij ons tot de orde. "Wat moet er nog gedaan? Wat is daarvoor nodig? Hoe lang gaat het duren? Wanneer is het af? Gaan!"

Daarnaast nog dank aan de andere co-auteurs van hoofdstukken in dit proefschrift. Linde, onze samenwerking leidde tot hoofstuk 2. Dankjewel dat je me uitnodigde om te presenteren tijdens de Endo-Neuro-Psycho meeting in Lunteren. Ilona, ook wij hebben fijn samengewerkt. Je kwam met een hele interessante dataset aan. Van de bijbehorende onderzoeksvraag dachten wij achtereenvolgens dat hij heel eenvouding, volslagen onmogelijk, of misschien ten dele te beantwoorden was. Het resultaat staat in hoofdstuk 4.

Mijn project is gefinancierd vanuit het NeuroBasic PharmaPhenomics consortium. Binnen Wageningen University werkten wij op een eilandje aan dit project, maar de projectmeetings van het consortium gaven mij een beeld van het volledige spectrum aan onderzoek dat rond dit thema gedaan werd. Dat hielp mij om in te zien hoe mijn radertje past binnen het grote geheel. Ook de PhenoTyper meetings bij Sylics waren interessant om een beter beeld te krijgen van de praktische aspecten van het uitvoeren van dierexperimenten. Raymond, jij hebt er veel tijd aan besteed om mij op gang te helpen in de beginfase van mijn project. Dankjewel daarvoor. Wil, jij was mijn ingangspunt bij Noldus en was altijd bereid om te helpen.

My dear Biometris colleagues, I was a bit afraid to start working at the Mathematical and Statistical Methods group but I felt welcome from the start. The coffee breaks and the abundance of cake definitely helped me feel at home.

Bas E., Paul G., Gerrit G., ik ben regelmatig jullie kantoor binnengelopen als ik weer eens een (mixed modelling) vraag had. En ik kwam er altijd met een antwoord uit. Elly, bedankt dat ik bij jou mijn oma-boodschappen mocht testen. Saskia, Eric, Ineke, en alle student assistenten, dankzij jullie werd onderwijs geven van een noodzakelijk kwaad een plezier. Ook al was ik het soms helemaal zat. Maikel, Joost, en Gerrit P, dankzij jullie had ik genoeg opslagruimte voor al die enorme datasets en een goede pc om ze te analyseren. Maikel, jij speciaal bedankt voor het oplossen van mijn computerproblemen in de afrondingsfase. Dinie, stralend middelpunt van Biometris, je staat altijd voor iedereen klaar. Dankjewel voor je gezelligheid, voor het organiseren van al die uitjes en nieuwjaarsdiners, voor het eindeloos invullen van formulieren, en voor het helpen bemachtigen van een promotieceremonie op vrijdagmiddag om 16.00u. Jaap, Fred, Gerie, Ron, mede dankzij jullie behoren de medewerkers van Biometris, volgens onafhankelijk onderzoek, tot de gelukkigsten binnen de universiteit.

I consider myself very lucky that so many of my colleagues became my friends. We had a lot of fun outside of the office too. Our Friday afternoon drinks (that sometimes ended on Saturday morning), all of our dinners, (PhD-)parties, the We-days, trips to Ameland and Limburg, thank you all! Jeroen, Cassandra, Simon, Marian, toen ik jullie ontmoette wist ik dat ik het naar mijn zin zou gaan krijgen bij Biometris. Het beste medicijn tegen een PhD-dip is een flinke dosis domme grappen en eindeloze lunches of theepauzes in de zon. Onverstoorbare Jeroen, op jou kun je altijd rekenen. Cassandra, ik kan nooit meer erwtensoep met rookworst eten zonder even te lachen. Simon, ik dank je evenveel voor je goede ideeën en inzichten als voor de onzin en bizarre plannen die je kunt verzinnen. Ik ben blij dat je mijn paranimf wilde zijn. Daniela, my other paranymf, thank you for being my sports buddy, my office mate and my friend. Rianne, mede-AIO-van-Cajo, we hebben het altijd gezellig gehad en zijn samen vele Young Statistician evenementen langsgegaan. Tryntje en Amber, onze tripjes naar Stuttgart met Cassandra en Jeroen waren fantastisch. En wat hebben we daarbuiten ook nog een hoop afgelachen. Thanks to my office mates on the second floor and in the corner office. Thomas, Nurudeen, you are two of the most positive and happy people I know. Thanks for making me laugh. Maggie, your strength during your PhD-project inspired me. Thank you for having me as your paranymf. George A., I will never forget your talent for Pictionary. Vincent, you bring colour to Biometris with your creative outbursts and philosophical insights. Julio, we have joined some truly great parties together. Martin, you are the youngest grumpy old man I know. And thanks to my newest office mates Manya, Peter D., Ruud,

Jurian, Viktor. Namaste to my yoga and dinner buddies Frederik and Emily. It is not solely the yoga that keeps me sane and relaxed, the food, wine, and good conversation are of great help as well. Thanks to George K. and Katarina for the dinners, both the wonderful elaborate moussaka dinners and the improvised dinners after Friday afternoon drinks. Gavin, with your company and a a bottle of wine I can feel "als god in Frankrijk" anywhere.

Many international guests have stayed at Biometris. Some of them have invited me to their homes. Carina, ray of sunshine, thank you for having me as a guest at your wedding. It was an honour. Matthias, thank you for inviting us to the Canstatter Wasen and for teaching me some essential German. You always know the right questions to ask. Han Sen, we shared some epic weekends in Stuttgart and in Rotterdam. Federico, it was lovely visiting Veronica and you in Sicily.

I also would like to thank Guus, Dominique, Yutaka, Stephan, Kevin, Henri, Bas, Sanne, Santosh, Apri, Rumbi, Pieter, Xu Dan, Wenhao, Bader, Rianne, v B., Ricky, Sophie, Amy, Bart-Jan, Johannes, Marco, Antoine D., Antoine L., and all the others that have come and gone.

Thanks to the VVS-OR and the Young Statisticians I was able to meet many interesting people working in statistics throughout the country. Nynke, Sanne, Iris, it was wonderful being in the board of the Young Statisticians with you. Maarten, Kees, Pieter, Birgit, Mia, and all the other members made the activities worthwhile.

Fortunately, I did not have to work all the time. Laura, our trip to the Philippines was amazing! You are amazing. Thank you for being my friend. Mijn lieve nerdy Wageningse vrienden; het was altijd gezellig tijdens het geo-cachen, AIVD-kerstpuzzelen, spelletjes spelen, pub quizzen en borrelen. Tom en Jan, bij jullie heb ik altijd een bank om op in slaap te vallen. Tom, onze vakantie in Estland en Letland was geweldig. Mirthe, jouw praktische inslag maakt van elk gedoe snel een gedoetje. Susan, fijn dat je met Yoannis weer terug naar Wageningen bent verhuisd. Ik voorzie nog vele spelletjesmiddagen.

Peter, we zijn samen aan dit avontuur begonnen en inmiddels onze eigen weg gegaan. Bedankt voor jouw steun bij het afronden van mijn MSc en de eerste fasen van mijn PhD.

Gerard, er is veel veranderd sinds ik als totale noob bij de Internetcie begon en jou ontmoette. Maar onze vriendschap is gebleven.

Wallies, oude en nieuwe, wat hebben we het fantastisch gehad met elkaar. Rinske, Lisanne, Klaas, Maarten, Mechiel, Linda K., Linda S., Maartje, Meike, Joep, Daniel, Felicia, en al die anderen. Wat fijn dat het elke keer dat we elkaar weer zien meteen weer als thuiskomen voelt.

Rozemarijn, eerste huisgenootje op de Churchillweg. Liters witte verf hebben we moeten aanslepen om er iets van te maken. Daarna hebben we vele koppen

thee gedronken en series gebinged samen. Jaume, roomie number 2, I hope you enjoy your new palace. You are always welcome to come over if you miss the cold.

Om al die uren achter mijn computer te compenseren heb ik veel gesport in het sportcentrum en het zwembad van de De Bongerd. Ik had niet altijd zin om te gaan maar dankzij Ingi, Ilona, Tijmen, Matthijs, en de andere instructeurs kwam ik er altijd met een goed humeur vandaan. Met sportmaatjes Jasperina en Jan-Eise was het altijd gezellig. De dinsdagavonden bij Aquifer waren sportief maar bij de toetjesavonden, karacocktail parties, en weekenden werd dat ruimschoots gecompenseerd. Jarst, Johan, Ron, Romy, Romee, Tom, Casper, en alle anderen. Dankjulliewel.

Marijn, Guyonne en Albertine, wat kennen we elkaar al lang. Als we elkaar zien is het altijd veel te lang geleden en voelt het meteen weer vertrouwd.

Families Vendrig en van der Mark, wat ben ik er dankbaar voor dat ik ben opgegroeid met zoveel tantes, ooms, nichten, neven, achternichtjes en achterneefjes.

Wes en Frank, mijn kleine broertjes die al lang niet meer klein zijn. Hoe ouder we worden, hoe dichter we naar elkaar groeien. Frank, we hebben wat legendarische avonden meegemaakt samen. Dankjewel voor je hulp bij het ontwerpen van de kaft en uitnodigingen. Wes en Edwina, dankjulliewel voor alle gezelligheid. Ik kijk er enorm naar uit om het nieuwste Vendrigje te ontmoeten.

Papa en mama, jullie hebben me van kleins af aan al gesteund en gestimuleerd. Zonder jullie lag dit proefschrift hier niet. Ik weet dat ik altijd op jullie terug kan vallen. Dankjewel.

The research described in this thesis was financially supported by the Dutch Fund for Economic Structure Reinforcement (FES), under Grant Agreement Number 0908 (the "NeuroBasic PharmaPhenomics project").

Cover design by Frank Vendrig and Nadia Vendrig

Printed by Digiforce || ProefschriftMaken