6 Summary, discussion, and future directions

Statistical analysis is a critical element in the research process: it allows one to draw appropriate research conclusions from sets of collected data. Using the correct statistical approach, i.e., one that fits the nature and structure of the data, is of utter importance in this process. The ever-increasing complexity of data, prompted by advances in experimental techniques available to the field of neuroscience, calls for statistical approaches that go beyond the standard statistical tests. To optimally exploit the information present in experimental data, the statistical methods of choice should not only ensure the reliability and validity of the research conclusions, but also optimally describe and/or accommodate the complexity of the data. In this PhD-project, we aimed to elucidate statistical methods that optimally fit the complex data obtained presently within the field of neuroscience, and to develop a novel statistical model that fully exploits the information contained within intensive longitudinal behavioral mouse data. In addition, we describe a novel method for testing anxiety in a home-cage environment (PhenoTyper).

In **Chapter 2** and **Chapter 3** of this thesis, we demonstrated that it is crucial to accommodate in statistical analyses the clustered nature of data, which arises when multiple observations are collected from each research object. This not only prevents an increased false positive rate but also optimizes statistical power. In case of a study design in which all observations within a research object pertain to the same experimental condition (design A), it has been pointed out before that the false positive rate increases when the clustered nature of the data is not accommodated in the analysis, both within the neuroscience literature and beyond [72, 73, 75, 91, 93, 94]. However, the prevalence of nested data, and the amount of dependency due to nestedness that can be expected in the field of neuroscience had not been assessed previously. For a study design in which the obtained observations within a research object can pertain to different experimental conditions (design B), the discussion in neuroscience literature was limited to the gain in statistical power when accommodating variation in the average baseline outcome [72, 91]. However, in design B not only the average baseline outcome, but also the effect of the experimental manipulation may vary over research objects. Not accommodating variation in the experimental effect may result in an increased false positive rate. By means of a simulation study, we demonstrated the degree of inflation given systematic variation in either only the experimental effect, or in both the experimental effect and the baseline condition. These results are a valuable addition to the few previous (theoretical) studies [75, 151, 152] in which researchers showed with a example case or cases, or by considering the equation of the standard error of the experimental effect, that not accommodating this variation may result in an increased false positive rate.

In **Chapter 4**, we described and pharmacologically validated a new anxiety test that allows for unsupervised, automated, high-throughput testing of mice in a home-cage system. The development of this test was motivated by a pressing need for reliable, high-throughput methods for comprehensive behavioral phenotyping to optimally benefit from the increasing availability of experimentally engineered mouse lines as expressed by e.g. [24, 34, 35], and nicely adds to the automated home-cage task developed by Kas et al. [52] to assess anxiety related behaviors.

In **Chapter 5**, a statistical tool based on Markov modeling - a hierarchical hidden semi Markov model (HSMM) - was developed and implemented in a Bayesian context to describe the temporal organization of behavior that can be observed when mice are studied in home-cage systems over a prolonged period of time. While simulation studies showed that the developed model still requires some adjustment if it is to be applied to data that resemble the observed mouse data, a real data example, comparing the behavioral pattern of young adult and aged C57BL/6J mice already clearly illustrated the advantage of the hierarchical HSMM over standard summary statistical tests. A Markov model including hidden behavioral states has been used once before to analyze longitudinal mouse data [59]. These researchers did not use a hierarchical model. In contrast, they used a two-step procedure in which they first assume that the underlying model that generates the observed behavior is similar over all mice in all groups, but then continue to investigate possible differences between groups based on the parameters obtained in the first step. The hierarchical model that we developed, however, allows for heterogeneity in model parameters both within and between groups. As a consequence, more information on individual differences between mice is retained, and group differences are better discernible and can be tested formally. In addition, the model we developed is not based on the generally untenable assumption that the probability of spending more time in the current behavioral state does not depend on the time already spent in that state. Moreover, although HMMs with a hierarchical structure have received some attention in literature [133–136], a hierarchical HSMM, allowing for random effects in all model parameters while utilizing the favorable properties of the Gibbs sampler, has not been presented before [132, 137].

All in all, the studies reported in these four chapters demonstrate the importance of applying statistical and methodological methods that fully exploit the complex structure of data generated by the novel experimental techniques that conquer the field of neuroscience. In the following paragraphs, I will discuss challenges and opportunities when collecting multiple observations from research objects, validity of automated home-cage systems and recommended developments, and the added value and future directions of the developed hierarchical hidden semi Markov model.

6.1.1 Challenges and opportunities when collecting multiple observations from each research object

The possibility to study increasingly smaller biological entities facilitated by advances in experimental techniques have shifted the n from the animal or tissue level to the cellular or even subcellular level. This makes it possible to obtain multiple measurements from each animal (for example, measurements on multiple neurons harvested from the same animal), allowing researchers to reduce the number of sacrificed animals while still obtaining a reasonable number of observations. Measuring at this lower level is thus advantageous for scientific, economic, and ethical reasons. Experimental designs, in which multiple observations or measurements are clustered within research objects, are, as we showed in **Chapter 2**, common in neuroscience research. Following our study presented in **Chapter 2**, we observe that in neuroscience and related scientific areas the interest in multilevel analysis has grown, as did the awareness of the significance of accommodating dependency resulting from nested designs. Boisgontier and Cheval [153] underline the need to transition towards multilevel models instead of classical analysis methods (i.e., ANOVA) to increase the reliability of the field of neuroscience, and point out that although the ratio between using these two methods have started to increase, the field of neuroscience is lagging behind compared to other scientific areas. In the field of primatology, Pollet et al. [154] discuss the conceptual advantages of using multilevel models compared to not accommodating the nested structure of the data or aggregating multiple observations taken from the same individual. Magezi [155] stresses the need to use multilevel models in case of within-participant data and presents a free, simple, graphical user interface to do so, and Moen et al. [156] discuss the need to account for multiple observations nested within a study participant in detail.

The use of multilevel analysis also provides opportunities in terms of novel research questions that can be probed, and has implications for the number of observations that need to be collected. In addition, collecting multiple observations per research object raises several theoretical and statistical issues, some of which will be discussed below.

People or plants? When it comes to nested designs, an important theoretical question is what the actual biological unit of interest is: are we interested in what happens at the level of the cluster (e.g., mouse), or are we actually interested in what happens at the lower level, i.e., the level of the observations within the clusters (e.g., neuron or cell). The tradition of multilevel analysis originates from the social sciences, in which nearly always individual people constitute the lowest level in a hierarchical model. Despite the fact that the people are clustered – e.g., children clustered in schools, or patients clustered in treatment facilities – the people are generally the unit of interest within these studies, not the higher order clusters. That is, social science researchers are interested in drawing conclusions about people while statistically accommodating the fact that the people from the same cluster can show additional similarity, which violates the assumption of independence. This contrasts with the statistical tradition in biological sciences, where much of ANOVA was developed and first applied within agricultural research [157, 158]. Here, early research focused on measuring e.g. the yield of a plot of land as a function of various experimental interventions that affected individual plants on the plot. Within this research design, the total yield of the plot of land is the unit of interest, and not the yield of the individual plants in the plot. In this context, measurements from plants are often referred to as subsamples, or pseudoreplicates, and are assumed not to contribute any new information. Hence, in these analyses the number of observations equals the number of clusters (i.e., plots of land), and standard statistical analysis is performed on the aggregated individual measurements over clusters [92].

Within neuroscience studies, the question whether the individual observations (e.g., neurons, cells) are the unit of interest, possibly contributing individual information, or whether these rather represent pseudo-replicated measurements of the same research object (e.g., animal; viewing them as plants), can be topic of discussion. However, the rise of, for example, single-cell biology strongly advocates the stance that neurons and cells can contribute individual information and should at least in some studies be treated as the unit of interest [159]. Until recently, it was largely unknown how heterogeneous phenotypically/morphologically similar cell populations are. However, advances in experimental techniques have allowed researchers to probe variability at the molecular level. For example, transcriptome in vivo analysis (TIVA) allows determination of gene-expression patterns at the single cell level and even of subcellular compartments [160], multiplexed error-robust fluorescence in situ hybridization (MERFISH) not only allows the quantification of RNA transcripts for single cells, but also its location [161], and recent advances in mass spectrometry methods allows one to quantify metabolites at the single-cell level (see e.g. [162]). These techniques have demonstrated that subtle differences at the molecular level can yield significant differences in cellular behavior (see e.g., [163, 164]), and that cell populations are much more heterogeneous than previously thought. In addition, in many neuroscience studies not all observations within a cluster receive by definition the same treatment, as is generally the case in the context of subsampling or pseudoreplicates. For example, when vesicle release is investigated in response to trains of electricity, the manipulation (i.e., trains of electricity) is conducted on each individual cell separately, and not on a group of cells. As such, we adhere to the stance that the unit of interest in neuroscience studies is typically at the level of the individual observations (i.e., we are truly interested in characteristics of the neuron or cell itself) in this PhD-project.

We emphasize that the individual observations anticipated in this thesis are of a different order than (technical) replicates often taken to monitor the performance of the experiment, for example repeating western blots or measuring mRNA levels multiple times within the same animal, as described by for example Vaux et al. [165, 166]. In the case of technical replicates, treating measurements from the same animal as pseudoreplicates is indeed the fitting procedure. We acknowledge that in biological neuroscience the boundaries between true technical replicates and multiple measurements that can contribute individual information and are themselves the unit of interest, can be fuzzy. We note, however, that in the specific case that individual observations originating from the same cluster are indeed exact replicates, multilevel analysis will pick this up and correct for this by setting the effective sample size equal to the number of clusters, as such rendering the clusters (rather than e.g. the neurons within clusters) the focus of the analysis. However, if not, individual observations originating from the same cluster, either or not intended as pseudo- or technical replicates, are allowed to contribute unique information.

More than 2 levels For the sake of simplicity, we only discussed hierarchical models with 2 levels in **Chapter 2** and **Chapter 3**, for example measuring multiple neurons within each animal. However, the possibility to study increasingly smaller biological entities results in data that sometimes comprises more than 2 levels of nesting. For instance, when collecting multiple measurements per cell, often more than one cell is obtained per mouse. In that case, characteristics of the cell may depend on the pup the cells were harvested from, rendering pups a third level. As advances in experimental techniques proceed, it only becomes more likely that data comprises more than 2 levels. As such, the question of how to deal with a more extensive multileveled structure becomes essential. Fortunately, multilevel techniques allow for the classification of many different levels. Of course, for stable parameter estimates, a sufficient number of observations are required at every level. If the number of third level clusters is small, another possibility is to use the third level variable (e.g., pup) as a covariate. This is an adequate way to accommodate the dependency if the number of clusters at the third level is too small (e.g., not more than 4) to model properly.

Should all levels be incorporated in the analysis? To model the effects of all the levels properly, identifiers of all possible levels of nesting need to be present in the data. In addition, when ignoring a level of nesting does not influence the reliability of the estimated effect, we might overcomplicate our model by including all levels. Some work has been done on the consequences of ignoring a level of nesting [167], demonstrating that ignoring a level of nesting can decrease the power to detect the experimental effect of interest. That is, when a level of nesting is ignored, variation between clusters at this level cannot be accommodated in the statistical model. The ignored variation ends up as noise in the lower modeled level of nesting, yielding an increased standard error of estimated effects at this level. Using the current example, we may ignore the cell level (i.e., the intermediate level) or the pup level (i.e., the top level). Let us assume that the experimental effect does not vary over clusters when the experimental manipulation is performed at one of the lower levels. In that case, ignoring the intermediate level results in a decreased power to detect the effect of a predictor at the lowest level (i.e., the experimental manipulation is varied at the level of the individual observations within the cell), but does not affect a predictor at the top level (i.e., the experimental manipulation is varied at the mouse level). Ignoring the top level results in a decreased power to detect the effect of a predictor at the intermediate level, but does not affect the estimation of effects of predictors at the lowest level.

However, more research is needed in case that the experimental effect does vary over clusters. For instance, in case of design B data, it is as yet not known how ignoring a level of nesting influences the estimation of the experimental effects at lower levels. As the complexity of data continues to increase, questions like these becomes more urgent, especially as it will often proof impossible (financially, ethically, time-wise) to collect data that contain sufficient clusters at all levels to allow comprehensive multilevel modeling.

Implications and opportunities Post hoc choosing a statistical analysis method suited to the data collected within a specific research design is important in optimally exploiting research data and in the acquisition of trustworthy research conclusions. However, a priori considering the possibilities and requirements of the envisioned statistical method when designing and conducting a study, is essential for the success of a scientific study. For instance, neuroscience studies that utilize a nested design currently generally encompass several to many observations within each cluster, collected over only a few clusters. However, not only does the statistical power to detect the experimental effect of interest largely depend on the number of clusters rather than the number of observations within each cluster, a minimum of 10 clusters is recommended to obtain unbiased estimates of the overall experimental effect and its standard error [76, 97]. As multilevel analysis requires more clusters than are presently conventionally sampled, it is crucial to recognize a nested design before data collection starts such that the inclusion of a sufficient number of clusters can be warranted. The number of clusters should be such that not only unbiased estimates can be obtained, but also such that statistical power is sufficient to detect the experimental effect of interest. In practice, this may imply that more animals need to be sacrificed than is currently customary for a study to obtain reliable parameter estimates and reach adequate statistical power. Of course, the appropriate balance between using as few animals as possible and obtaining reliable research conclusions, should be topic of debate. However, we believe it is crucial to appreciate the waste associated with studies that yield unreliable research conclusions due to a (too) small sample size, and/or with studies that are characterized by low statistical power, i.e., a too low probability to detect experimental effects that are actually present [14–16]. Multilevel analysis can provide more than a means to obtain reliable research results in case of nested designs, and it can be rewarding in future studies to adjust the study design and data collection such that it is possible to explore these venues. Multilevel analysis, for instance, allows one to quantify and test the statistical significance of the cluster-related variation in the experiential effect, which provides a valuable test of the generalizability of the experimental effect [95] (i.e., whether the impact of the experimental manipulation is similar across (biologically intrinsically) different settings). One can also quantify and assess the statistical significance of cluster-related variation in the average (baseline) outcome. Both types of cluster-related variation are indicative of structural variation over clusters, which can be scientifically and biologically interesting. As such, it would be useful to gain knowledge on the amount of cluster-related variation typically observed in certain experimental designs. In addition to being scientifically and biologically interesting, this will also aid a priori power analysis. Testing the statistical significance of cluster-related variation, however, requires a minimum of 30 clusters [76, 97].

As recently also pointed out by [156], even more interesting is that multilevel analysis facilitates studies into the factors underlying the cluster related variation by including them in the model as a covariate. An important implication of this possibility within the context of neuroscience studies is that it enables the research of gene/gene and gene/environment interactions. For example, mice can be randomized over various environmental settings, while knocking down a specific gene in some, but not all, of the neurons in each mouse (e.g., using shRNA). Infection of the neurons can be quantified by using a fluorescent marker. Subsequently, one can assess whether the effect of the gene (i.e., the difference between infected and not infected neurons) is influenced by the varying environmental settings. Traditional studies typically investigate the effect of only one genetic or one environmental factor, but it becomes more and more clear that the interaction between genetic factors, and the interaction between genes and environment, are critical in normal and abnormal functioning of the central nervous system. Multilevel analysis provides a means to study these questions directly.

6.1.2 Validity of automated home-cage systems and recommended developments

Automated home-cage systems allow the study of various aspects of spontaneous behavior, and yield unbiased long-term continuous observations of both novelty-induced and habituated behavior in mice with minimal human intervention [24]. However, these systems have only been developed recently, and as such their validity still needs to be established. In the following paragraphs, I will evaluate how well automated home-cage testing, and the PhenoTyper in particular, translates observed behavior into measurable factors, and how this can possibly be improved. I will distinguish between *ecological validity* – is the mouse behavior observed in the home-cage truly representative for mouse behavior in natural conditions, *construct validity* – are the measurements an accurate and a sufficient representation of mouse behavior, and *criterion validity* – how well does information obtained in the home-cage correspond to information obtained using other, already validated, instruments.

Ecological validity Automated home-cage testing allows for an animal-centered behavioral phenotyping method: the mouse is allowed to actively manage the timing and amount of participation to the test [33]. This contrasts with the experimenter-centered approach of classical behavioral tests, in which mice are introduced to the test setting at the convenience of the experimenter. In this perspective, automated home-cage testing in general, and the light spot test described in **Chapter 4** in particular, can be considered more ecologically valid compared to classical behavioral tests: in natural settings mice also manage the timing and amount of exposure to various stimuli themselves.

In addition, the aversive stimulus used in **Chapter 4** is a mild light spot (2000 Lx, comparable to the light of a typical overcast day). As a result, the manipulation can be considered ecologically valid: such mild aversive stimuli that do no explicitly involve pain or discomfort are encountered in natural settings and the observed response can therefore likely be generalized more easily to natural behavior. The use of such milder, ecologically more valid, experimental manipulations is facilitated by the powerful design of automated home-cage testing: it includes multiple habituation days that can feature as a baseline condition in the evaluation of subsequent experimental manipulations. This within-subjects design, in which the response of a mouse can be evaluated in the light of its own baseline behavior, is statistically more powerful compared to a between-subjects design in which evaluation of experimental manipulations is based on comparisons between different groups.

A limitation of automated home-cage testing in terms of ecological validity is that mice are housed individually, whereas mice are social animals with a structured hierarchy. The problem lies in the tracking of the animals: it is as yet problematic to track multiple animals at once at a detailed x-y coordinates resolution. Until now, all setups providing this detailed resolution use single housing [36–42, 46]. The systems that do allow group housing with individual recognition (e.g. the IntelliCage [43–45]) used RFID (radio-frequency identification) chips implanted subcutaneously in the animals, and only provide information on the animal being in a particular compartment. This information is, for example, not sufficient to classify the recorded behavior into mutually exclusive behavioral acts as done in **Chapter 5**.

Construct validity In **Chapter 5**, we modeled spontaneous behavior, which was operationalized in terms of a set of mutually exclusive behavioral acts: move, linger, sit, eat, on the shelter, short time in shelter, medium time in shelter, and long time in shelter. The question in terms of construct validity is whether this operationalization of spontaneous behavior is accurate and sufficient. Spontaneous behavior of mice is composed of many more components than can currently be recorded by the PhenoTyper, such as grooming, rearing, digging, climbing, and hanging. By recording a more detailed account of murine behavior, construct validity of spontaneous behavior can be improved, facilitating research on mouse models of e.g., OCD and ADHD.

Newly developed automated home-cage systems like the Spectrometer [46] or the system of Jhuang et al. [48] can extract many more behavioral acts (e.g., grooming, rearing, climbing). The Spectrometer includes accelerometers embedded in the floor supports, which capture the mouse's vibrations and infrared beams detecting when an animal rears, enabling a highly detailed representation of the spontaneous behavior of a mouse in an open field. The system of Jhuang et al. includes a front camera, facilitating the use of a trainable computer vision system that automatically annotates complex mouse behaviors. However, also other, less technologically advanced alterations to current home-cage systems would increase the amount of information obtained about spontaneous behavior. For instance, enlarging the cage would allow mice to display running behavior (i.e., with the current dimensions – 30 x 30 cm – running for more than a split second is not possible and as such difficult to record), and thus facilitate the dissection of slow and fast movement as shown in [168,169]. In addition, in a larger cage different distances from the wall categories can be meaningfully dissected, which is indicative of e.g. anxiety [115, 169]. Furthermore, a camera in the shelter would reveal what the mouse does in all the hours that it resides there. And including an infrared beam in front of the bars of the feeding station and the nozzle of the drinking spout would allow a more reliable measurement of eating and drinking behavior.

Criterion validity Theoretically, criterion validity can be further dissected into predictive validity – the degree to which a test predicts what it should predict according to theory, and concurrent validity – the degree to which the test outcomes correlate with other measures of the same construct. The behavioral response to the light spot test discussed in **Chapter 4** was blunted by treatment with an anxiolytic drug (Diazepam), demonstrating predictive validity of the assay, by indicating that the observed behavioral response has a significant anxiety component. To assess the concurrent validity of the light spot test, one could subject mice to both the light spot test and a classical anxiety test, such as the dark-light box or the elevated plus maze [21, 22], and correlate these measures, an experiment we did not perform. Despite the fact that we showed that the light spot is equally effective as classical anxiety tests from a pharmacological perspective, it remains to be tested whether previous results assessing mutant/wild-type differences in classical anxiety tests can be replicated in anxiety tests within an automated home-cage system. In the light spot test, the anxiogenic stimulus is provided within a habituated home cage environment, disentangling unspecific arousal states from the anxiogenic stimulus. In classical anxiety tests, however, both human-animal interactions as well as the general novelty of the apparatus itself will impact the arousal state of the mouse, possibly amplifying the behavioral differences between mutants and wild types. For example, Fonio et al. [31] demonstrated that the difference in anxiety between an inbred strain typically characterized as anxious (BALB/c) and a wild derived strain (CAST) was reversed after habituation. As such, it would be informative in the context of behavioral genetics studies to compare mutant/wild-type differences in both classical anxiety tests and anxiety tests within an automated home-cage system. Both type of tests measure anxiety within a specific context instead of generic anxiety, and as such providing differential information. A low correlation between the outcome measures of classical anxiety tests and newly developed tests like the light spot test would be indicative of this, i.e., displaying low concurrent validity. Important to note, however, is that low concurrent validity does not imply low construct validity of any of the tests involved. It can simply mean that the tests tap the construct of interest from a different, but equally valid, angle.

Predictive validity of modeling the pattern of spontaneous behavior as described in **Chapter 5** has been assessed to some extent by comparing the behavior of young adult and aged mice. We showed that aged mice display a less active pattern of behavior compared to young adult mice, in line with what according to theory is expected. Future studies are however needed to fully assess the construct, predictive, and concurrent validity of our approach to modeling the pattern of spontaneous behavior. Further assessment of predictive validity can for example be performed using pharmacological compounds or using well characterized mouse models of neurological, psychiatric, or neurodegenerative disorders. Demonstrating concurrent validity will be challenging, however: our developed statistical method describes behavior from a dynamic angle, while classical statistical methods describe behavior from a static, segregated point of view. It is therefore questionable how informative a correlation between the results of our method and classical analysis methods would be in terms of concurrent validity.

6.1.3 Added value and future directions of the developed hierarchical hidden semi Markov model

The developed hierarchical hidden semi Markov model provides a comprehensive description of mouse behavior over time, resulting in new and more detailed information on behavior, both revealing differences not observed with conventional analysis, and providing information on why differences occur (**Chapter 5**). Using the proposed model, differences in behavior can be established and understood. As such, modeling the dynamics of behavior in mouse models may shed new light on the pathophysiology and treatment of neurological, psychiatric, and neurodegenerative disorders that often characterize changes in day-to-day behavior [53]. The development of models that describe mouse behavior over time has only recently become relevant, as innovations in automated detection of rodent behavior via a tracking system (for example video tracking [3] or transponder technology [43]) have eased the collection of prolonged observations in e.g. automated home-cage systems.

Applicability of the hierarchical HSMM Prompted by technological advances, the collection of (intense) longitudinal data has also become more frequent in other scientific areas. For example, user-friendly wearable measuring devices that can automatically keep track of person related measurements (e.g., the number of steps walked, heart rate, the quality and amount of sleep, or a persons body temperature, see for example the fitbit, https://www.fitbit.com/, and temptraq, https://www.temptraq.com/), or environment related measurements (e.g., temperature of the environment, exposure to light and UV, or air quality, see for example sunsprite, https://www.sunsprite.com/, and TZOA, http://www.tzoa.com/) are now commercially available. In addition, the use of hand held computers, mobile phones, and web interfaces have greatly increased the possibilities to keep a diary of psychological or other (health related) measures, in which the measurements are densely spaced in time and collected over a prolonged period of time.

Typically, time series analysis is used to describe the within-subjects processes captured in such data (see e.g., [170–174]],). Time series analysis is concerned with describing the relationship of a variable with itself over time, and/or with the relationship of a variable with other variables over time (see e.g., [175,176]). Examples are autoregressive (AR) models, which describe the degree to which the current observation can be predicted from the previous observation, and multivariate autoregressive models (VAR), which investigate the causal relationship between measured variables. Another popular time series model is the regime switching model, in which the (V)AR model is extended to include multiple structures (equations) that describe the time series in different "regimes". The regimes are unobserved (i.e., they are inferred from the data), and the switching mechanism between these regimes is a Markov model. Hence, the regime switching model is a hidden Markov model, in which the observations generated in each state (here called regime) are characterized by an (V)AR model. Conventional time series analysis models are suited to describe the data of a single case. However, with the improved possibilities to collect many repeated measures from many individuals, hierarchical extensions of these autoregressive models are highly contemporary and a lively topic in current research (see e.g., [177–183]. These models allow the study of group effects, while allowing the within-subjects dynamics to vary. Consider for example the application of a multilevel autoregressive model to study relationship-specific positive and negative affect in males and females. Here, one can quantify the overall mean levels of relationship-specific positive and negative affect for males and females, and the overall carryover effect of positive and negative affect in males and females, while allowing for heterogeneity between individuals. In addition, one can investigate whether a predictor measured at the individual level (i.e., only measured once for each subject, for example relationship satisfaction) can explain individual differences observed in the mean and carryover effect of positive and negative affect in males and females [183].

Our developed model nicely adds to these novel methods available to describe longitudinally collected data. In the regime switching model, it is implicitly assumed that the probability of spending more time in the current regime does not depend on the time already spent in that regime (i.e., the regime switches are characterized by a Markov process). It is likely that this assumption, just like in mouse behavior, is untenable in many of the longitudinally collected data. It is relatively straightforward to specify an autoregressive state-dependent process in our developed model instead of a model that describes the probability of observing categories of events. In this case, one obtains a hierarchical regime switching model, which allows for the quantification of group effects while allowing for between-person differences, and which models the duration of the regimes explicitly using a duration distributions that can deviate from the exponential distribution.

Even when retaining a categorical state-dependent distribution, the developed model described in **Chapter 5** can provide an interesting description of many types of longitudinally collected data. For example, consider longitudinal data on sleep stages collected in subjects that do or do not suffer from insomnia. In this example, the hidden states can be used to filter out "measurement error" in the sleep stages, advantageous when for example using automatic classification of sleep stages based on EEG measurements. Using the hierarchical HSMM, one can investigate whether in general individuals that do and do not suffer from insomnia differ in the duration of the sleep states and the probability to switch between sleeping states. In addition, the model allows for heterogeneity between individuals in the duration of the different sleep stages, the switching between the sleep stages, and the amount of measurement error. In summary, the developed model is not only relevant to behavioral mouse data, but applicable to a much wider variety of longitudinally collected data.

Future extensions of the hierarchical HSMM The developed hierarchical HSMM model still requires some adjustment when applying it to the observed mouse data. That is, the developed model does not accommodate the durations that characterize the observed behavioral acts, which results in biased parameter estimates. A possible solution is to include the duration of the current act as a covariate in the model, but further research is required to show the viability of this solution. When this extended model is obtained, further studies are required to establish the minimal number of mice, and/or minimal duration of the time series to obtain robust parameter estimates and reasonable levels of statistical power to detect differences between (experimental) groups.

In addition, the developed model assumes that parameters that describe behavior are stable over time (i.e., time-homogeneous). In the current mouse data, this limited the selection of behavioral data to a habituated episode of several hours. To enable the analysis of longer sequences of mouse behavior (e.g., one complete day where the level of activity changes throughout the 24 hours due to changes in circadian rhythm), additional adjustments to the model are required. One option would be to include time-varying covariates: one can design a model in which model parameters (e.g., different transition probabilities between states) vary across different observational periods. Additionally, for some types of data (e.g., longitudinal data collected on humans using wearables) it might be difficult to determine whether the chosen selection is indeed time-homogenous, making extending the model to accommodate non time-homogenous data expedient. Also, the inclusion of time-varying covariates allows for within-subject comparisons of behavior. For example, comparing mouse behavior before and after an experimental manipulation, or comparing behavior of patients before and after medication or therapy.

Moreover, additional studies comparing mutant/wild-type differences are required to demonstrate the usefulness of the developed model in the context of behavioral genetics studies.

6.1.4 Final remarks

The importance of using optimal statistical methods received suitable attention over the past few years (see e.g., [69–71, 184, 185]). Here, the focus is mainly on using the correct analysis method to obtain valid research conclusions. For example, Nature recently issued a special collection devoted to reproducibility [186], including several (new) publications on improving used statistical methods and experimental designs [187, 188], and better understanding of analysis results [189, 190]. As we have shown, going beyond the application of univariate models and optimally utilizing the rich data generated by novel experimental techniques offers additional advantages: besides obtaining correct research conclusions new biologically relevant information can be revealed. The complexity of experiments will continue to grow, as will the required statistical techniques. The work discussed in this thesis demonstrated once again that statistical methodology is truly key in optimizing the extent to which the field of neuroscience can profit from the marvelous technological advances that found scientific development.