EVALUATION AND COMPARISON OF STATISTICAL MODELS FOR INTENSIVE LONGITUDINAL DATA

A THESIS
SUBMITTED TO THE PROGRAM IN SYMBOLIC SYSTEMS
AT STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Julia R. Fischer
June 2025

To the Directors of the Program in Symbolic Systems:
I certify that I have read the thesis of Julia Fischer in its final form for submission and have found it to be satisfactory for the degree of Master of Science.

Signed electronically

6/6/2025

Nilam Ram
Principal Advisor
Departments of Psychology and Communication

To the Directors of the Program in Symbolic Systems:
I certify that I have read the thesis of Julia Fischer in its final form for submission and have found it to be satisfactory for the degree of Master of Science.

Signed electronically

6/6/2025

Tobias Gerstenberg
Second Reader
Department of Psychology

Abstract

Recent advances in machine learning, data availability, and computational resources have given rise to complex nonlinear models of psychological processes. Though theory can help guide model selection, there is no widely agreed-upon framework for identifying models that are appropriately complex and accurate for examining the phenomenon of interest. This thesis develops a framework for selecting statistical models for intensive longitudinal data in a principled manner. Drawing on literature from both social science and machine learning, I argue that neither qualitative model assessment nor automated accuracy testing is sufficient for selecting an appropriate longitudinal model. I instead offer a structured model selection framework that integrates three model properties – complexity, efficacy, and interpretability – to provide clarity on which models are best aligned with a researcher's analysis goals. I outline how to measure the three properties through both existing and newly proposed metrics. I illustrate how complexity and efficacy can be comprehensively measured using well-known metrics, like training time and prediction error. For interpretability, I argue that existing measurement methods are insufficient. Drawing on the perspectives of modeling experts, I present a new definition of interpretability and break it down into measurable facets. To illustrate the practical utility of the model selection framework, I apply it to two example research inquiries. The first example focuses on the time-oriented process of cognitive skill acquisition and uses simulated data to allow for testing of model assumptions. The second example examines an empirical dataset of intensive longitudinal psychophysiological data and explores the intricacies of model selection when there is no clearly defined data-generating function. After walking through these examples, I set forth practical guidelines to aid researchers in applying principled model selection to their own work. These guidelines include documenting the model selection process and letting one's research question guide the interpretation of model evaluation results. I also identify general trends indicative of theoretical relationships among complexity, efficacy, and interpretability.

Acknowledgments

I would like to thank the many people who have supported me, both during my time at Stanford and long before I arrived on campus. I am grateful to my primary advisor, Nilam Ram, and my lab manager, A Garron Torres, for guiding me during my three years as a member of The Change Lab at Stanford. In addition to their wonderful scientific mentorship, they gave me the confidence to believe in myself as a researcher. I would also like to thank the undergraduate, graduate, and postdoctoral members of The Change Lab for their invaluable exchange of ideas and welcoming community. Thank you to Michael Bernstein, Todd Davies, Hyowon Gweon, Michael Frank, Tobias Gerstenberg, Bonnie Krejci, and Dacien Sims for advising me throughout my time in the Stanford Symbolic Systems Program. I thank my fellow Symbolic Systems MS students for providing a supportive scholarly community. Finally, thank you to my loved ones for their unwavering support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the past several decades, quantitative methodology in psychology has evolved from simple statistical tests to an ever-growing set of statistical and mathematical models. Gone are the days of simple hypothesis testing and t-tests. To keep pace with the current state of the field, and arguably, to make useful contributions to science, psychological researchers must increasingly adopt complex modeling techniques. Among these techniques are a core set of theory-driven *data models* (Breiman, 2001b), like linear regression and structural equation models. When we engage in data modeling, we specify a data-generating function, then use a dataset to estimate the best parameter values for this function. As we enter into a new age of statistics, researchers must also consider the advantages of machine learning models – or, in Breiman's ontology, *algorithmic models* – with their associated nonlinearities and high-dimensional feature spaces. Algorithmic modeling takes a black-box approach in which no function is specified in advance. Instead, both the function and the parameter values are learned from data to maximize predictive accuracy.

For some applications, the choice between data modeling and algorithmic modeling is straightforward. If we want to determine whether a predictor variable X_j is linearly related to an outcome Y, a simple data model will suffice. If we want to predict the next word in a sentence, an algorithmic language model is necessary. However, consider the setting of modeling dynamic psychological *processes*, i.e., psychological phenomena that are characterized by within-person change over time. When studying psychological processes, we have competing tensions of eliciting a theoretically relevant functional form and ensuring good predictive accuracy for future observations. For example, if we are studying how people's emotions change in response to stressful life circumstances, we may want both interpretable cognitive parameters that explain between-person differences in stress reactivity and the ability to predict when a person's emotions will return to their baseline. The choice between data models, with their high interpretability, and algorithmic models, with their accurate predictions, becomes unclear.

To study psychological processes, some researchers collect and analyze *intensive longitudinal*

data, which consists of variables repeatedly measured many times at short time intervals (Hamaker & Wichers, 2017). This type of data allows us to examine the moment-to-moment changes in people's cognitions, emotions, and behavior. Intensive longitudinal data thus allow us to tackle research questions that are difficult to answer with data collected infrequently or on only one occasion. Some examples of psychological processes that intensive longitudinal data can help illuminate are regulating one's emotional state (Gross, 2014), task-switching (Monsell, 2003), and acquisition of cognitive skills (VanLehn, 1996).

Once intensive longitudinal data have been collected, the researcher must decide which model or models to use to make inferences about the psychological process of interest. Whether to use a more traditional psychometric model versus a theory-agnostic machine learning model is unclear for many researchers, and the choice depends greatly on the researcher's analysis goals: prediction versus inference, theory discovery versus theory confirmation, dimensionality reduction versus producing a high-dimensional latent representation, analysis of intraindividual (within-person) change versus interindividual (between-person) differences, among others.

My goal in this thesis is to develop a framework to help researchers reason through finding the optimal model for analysis of intensive longitudinal data – with "optimal" being defined relative to the researcher's specific analysis goals. To motivate this model selection framework, I propose a conceptualization of two model properties – efficacy and interpretability – as a function of a third model property, complexity (Figure 1.1). I argue that the optimal model lies at the point along the complexity axis which balances efficacy and interpretability, each weighted by their importance for achieving the researcher's analysis goals.

While this model selection framework is intended to be straightforward to apply, it is not meant to be entirely prescriptive. Using the framework will not necessarily lead to a single obvious choice of model. Recommending a single best model would be antithetical to the goal of helping researchers make appropriate modeling choices for their particular research goals. Instead of being an automated method of model selection, this framework is designed as a way to give researchers access to better information about their models to make more informed, principled modeling decisions. While the framework provides conceptual clarity and tools for the quantitative measurement of model performance, researchers must ultimately draw upon their domain knowledge and research aims to make their final modeling choices.

In addition to proposing a new model selection framework, this thesis has several other goals. First, I hope to develop a comprehensive definition and set of facets to assess interpretability. Methods to evaluate model complexity (e.g., Höge et al. (2018)) and efficacy (e.g., McDonald & Marsh (1990)) have been comprehensively addressed in prior work, but the assessment of interpretability has minimal coverage in the literature. Another goal of this thesis is to provide practical guidance for researchers when selecting models for longitudinal data. To this end, the Discussion section of this thesis sets forth written guidelines for researchers.

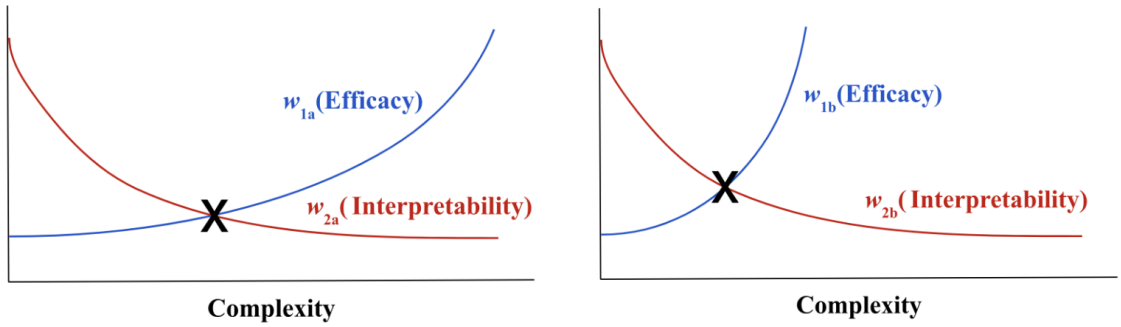

Figure 1.1: Heuristic for conceptualizing model efficacy and interpretability as functions of complexity, weighted (w) by importance to the researcher, with the intersection point (X) indicating the optimal model for a specific application.

1.1 Motivating Example

To illustrate the need for a principled model selection framework for intensive longitudinal data, I provide a toy example using a simple simulated dataset. Let's imagine that we are trying to teach a dog to roll over. We use treats to encourage the dog to reach intermediate positions between standing and rolling on their back. Thus, each "trial" of rolling over is associated with a certain number of treats given to the dog. Trials that require more treats suggest that the dog has a low ability to roll over. As the number of treats per trial decreases, we assume that the dog is learning how to roll over more effectively. We can plot how many treats were required for each trial (Figure 1.2).

We now want to identify a statistical model that represents the underlying process of the dog's learning curve. Perhaps we try a linear regression, a polynomial regression, and a locally estimated scatterplot smoothing (loess) regression (Cleveland & Devlin, 1988). We plot the three models against the data in Figure 1.3.

We first try selecting a model using subjective impressions. Perhaps we are cautious of overfitting to the data and thus choose the linear regression. If we rely on subjective impressions to do model selection, we may overlook crucial quantitative advantages of other models, like goodness of fit to the data. Next, we try a different strategy and select the model with the lowest error. In this case, we choose the loess regression. If we only make use of quantitative metrics, we may end up selecting a model that is difficult to interpret and gives us little information about the data-generating process we are trying to understand. When we use either subjective impressions or quantitative metrics alone, we are prone to weighting one model property over the others, instead of taking a balanced approach that seeks to optimize all three. By defining model properties, measuring them with rigor, and weighting them based on our unique research needs, we can make more informed modeling decisions. My framework aims to make this process of principled model selection more approachable

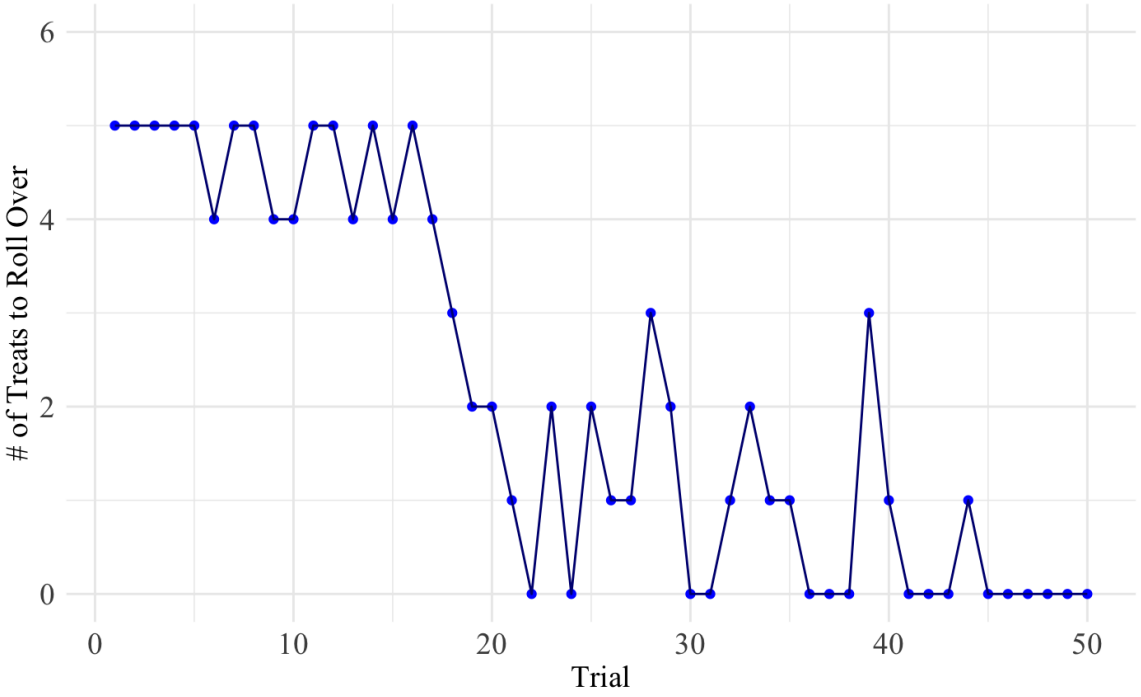

Figure 1.2: Toy example simulated data: number of treats required in each trial of teaching a dog to roll over.

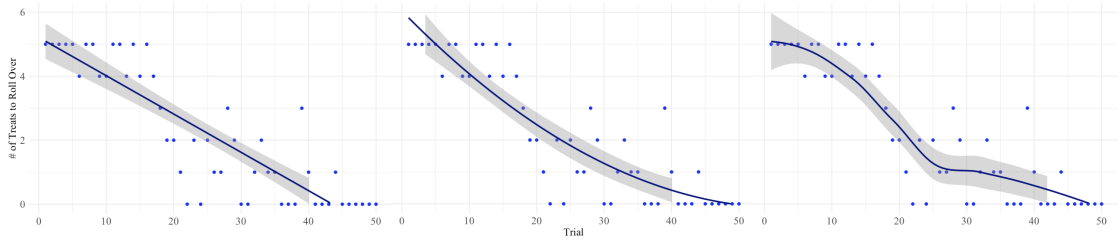

Figure 1.3: Toy example regression models: linear, polynomial, and loess.

to all researchers.

1.2 Complexity, Efficacy, and Interpretability

I next provide definitions of complexity, efficacy, and interpretability in the context of this framework. I aim to give holistic definitions that underscore the unique contributions of each model property. Specific metrics for measuring these model properties are given in the Method section and in Figure

1.4.

Complexity refers to how strongly the assumptions a model makes constrain the model's shape, parameterization, and expressivity. Complexity encompasses traditional notions of computational, time, and space complexity.

Efficacy refers to a model's ability to represent and predict data in detail and with high accuracy. Efficacy encompasses traditional notions of accuracy and goodness of fit.

Interpretability refers to how easily one can understand a model's transformation of inputs to outputs and obtain meaningful substantive insights from the model. Interpretability encompasses notions of explainability, transparency, and trustworthiness.

It is worth noting that these three properties may not be completely orthogonal. For example, within a certain class of models, such as deep learning models, efficacy may be directly correlated with complexity. Nevertheless, each property provides new information about a model not derivable from the other two properties. Complexity has the distinct quality of being an innate property of a model's specification. Efficacy and interpretability are causally downstream of complexity, the data being used, and the research question (at least in the case of interpretability). In other words, a model may become more efficacious or less interpretable *by virtue of* its increased complexity.

The final goal of this thesis is to uncover fundamental relationships among these three model properties. Fixing complexity as the independent variable, we can consider how the dependent variables of efficacy and interpretability vary with complexity. As more complex models have more parameters to represent richer information structures, I hypothesize that efficacy will **increase** with complexity, assuming appropriate regularization to guard against overfitting. In contrast, I anticipate that interpretability will **decrease** with complexity. As models become more complex, it becomes more difficult to ascertain the meaning and contribution of specific parameters, making the overall model potentially less interpretable. However, with high enough complexity, it is possible that a new form of interpretability emerges. While the inner workings of large language models (LLMs) are not always easy to interpret, they exhibit emergent communicative and reasoning behavior that allows us to interact with them as independent agents. We can ask these agents to explain how they reach conclusions and make decisions, affording these models an ability that I define as *emergent interpretability*.

Model Property	Definition	Evaluation Metrics
Complexity	(Lack of) constraining assumptions on model shape, parameterization, etc.; time/space complexity	Number of learned parameters Linear vs. nonlinear Pre-fixed vs. learned functional form Training time Prediction time (on test set)
Efficacy	Model's ability to represent and predict data in detail and with high accuracy	Training set mean squared error (MSE) Test set mean squared error (MSE) Akaike information criterion (AIC) Bayesian information criterion (BIC) Widely applicable information criterion (WAIC) Leave-one-out cross-validation information criterion (LOOIC)
Interpretability	Ease of understanding model's transformation of inputs to outputs and obtaining meaningful substantive insights from model	Qualitative analysis or quantitative scoring of models on interpretability facets by domain and/or modeling experts

Figure 1.4: Definitions and metrics to assess the three model properties of complexity, efficacy, and interpretability.

Chapter 2

Background

Prior work mapping the space of longitudinal models focuses on traditional quantitative psychology approaches, like structural equation modeling and growth curve models (Collins, 2006; Asparouhov & Muthén, 2020). There is a separate line of work focusing on machine learning methods for longitudinal data (e.g., Hu & Szymczak (2023)), but this work does not attempt to compare machine learning with traditional statistical approaches. Further, model selection research typically focuses on a particular model property, often efficacy, as in Zucchini (2000), and gives little advice about how to consider multiple model properties simultaneously.

With existing model selection criteria, it is unclear whether researchers apply these criteria in a principled way when doing model selection in practice. Some research articles motivate their choice of model based on what has worked in the past or what prior theory suggests as a plausible model. Yet other articles do not address how they selected their model at all. The lack of clarity surrounding model selection motivates the need for a robust model selection framework using a justifiable set of model properties.

2.1 Determining Model Properties

The three model properties – complexity, efficacy, and interpretability – were selected because they (a) are meaningfully related to what makes a good model, (b) each capture distinct aspects of what makes a model "good", and (c) are quantifiable and measurable.

Comparison with existing frameworks for model selection reveals the fundamental and comprehensive nature of this proposed set of properties. Past work in statistical modeling literature has focused on the tradeoffs of bias versus variance and parsimony versus goodness of fit. In the bias-variance tradeoff (Geman et al., 1992), *bias* refers to how incorrect the functional form of a model is. A model with high bias will be too simple to accurately capture the phenomenon of interest, leading to underfitting and poor accuracy on both training and test data. *Variance* refers to how closely

a model is fit to its specific training dataset. A model with high variance will be too sensitive to sample-specific fluctuations, which can lead to overfitting and lack of generalization to the population of interest. In model selection, we are often choosing between a higher bias, lower variance model and a lower bias, higher variance model. This tradeoff decision is similar to the decision between a low complexity model and a high complexity model. Bias and variance are also related to efficacy, with each of high bias and high variance contributing to low efficacy in a different way. While bias and variance are helpful for analyzing sources of model error, they are often too closely (inversely) related to be considered as two distinct dimensions in a model selection framework. Further, in the space of highly complex neural network models, the bias-variance tradeoff no longer seems to hold; when we have highly parameterized models and lots of training data, both bias and variance seem to decrease with complexity (Belkin et al., 2019; B. Neal et al., 2019). Thus, perhaps bias and variance are better thought of as contributors to efficacy that are related to the more fundamental dimension of complexity.

Similar issues arise with the tradeoff of parsimony versus goodness of fit, which refers to the choice between a simpler, more interpretable model and a more complex, better fitting model (Marsh & Hau, 1996). The relationship between parsimony and goodness of fit is less straightforward than once believed, especially with the emergence of deep learning models. Further, pitting parsimonious models versus well fitting models as a dichotomy can be misleading – the parsimonious model may be the better fitting model, especially when we are trying to identify a generalizable functional form for a time-varying process. To some extent, parsimony measures complexity and interpretability, while goodness of fit measures efficacy. The properties of complexity, efficacy, and interpretability thus cleanly capture the underlying dimensions of the bias-variance and parsimony-goodness of fit tradeoffs without entangling us in false dichotomies.

2.2 Model Selection for Psychological Processes

In the context of time-varying psychological processes, there are several special considerations we need to keep in mind when doing model selection. First, we must consider the granularity at which we want to study a psychological process. In some cases, we are interested in the high-level dynamics of a process and finding a functional form to characterize these dynamics. In other cases, we are more interested in within-person variation and moment-to-moment changes, as opposed to getting a high-level view of the process. In psychology, we are often working at a level of analysis that is somewhat abstracted away from the physical or biological underpinnings of an experience. This is not a necessarily a bad thing, and it often helps us understand phenomena more holistically. A good model is not necessarily one that captures every detailed fluctuation throughout a process. Thus, automated model selection frameworks that optimize for goodness of fit will likely choose a model that is too complex for our research question.

By definition, intensive longitudinal data contain many observations and often many participants and variables. Computational complexity is thus a crucial consideration when studying time-varying psychological processes. In particular, longitudinal Bayesian models can be quite computationally expensive. However, they allow for the influence of cumulative theory and offer useful information about parameters through posterior distributions. Choosing whether or not to use a Bayesian model in an intensive longitudinal setting requires a thorough evaluation of candidate models' practical complexity, estimation efficacy, and interpretability specific to the guiding research question.

Chapter 3

Method

3.1 Measurement of Model Properties

A key aim of this paper is to construct an evaluation framework to aid in researchers' selection of models for intensive longitudinal data by evaluating models on the properties of complexity, efficacy, and interpretability. In this section, I seek to operationalize each of these properties through several measurable quantities.

3.1.1 Measuring Complexity

To measure model complexity, or how constrained versus expressive our model is, we can deploy metrics that measure the complicatedness of our model specification and fitted model. It is important to observe that many of these metrics are a function of both the type of model *and* the size of the dataset being used. Thus, between-model comparison should only be done with models fit to the same data.

Number of learned parameters

The number of learned parameters characterizes the functional complexity of a model. It is measured by counting up all model parameters that need to be stored in order to make predictions on new data. For example, the coefficients β and the standard deviation of the error ϵ in a linear regression model count as parameters, while the maximum tree depth or minimum bucket size hyperparameters in a regression tree model do not.

Linearity

The linearity of a model is determined by how the predictor variables are related to the outcome variable. For models that are linear, all predictors are linearly related to the outcome. For models

that are nonlinear, at least one predictor is related to the outcome via a nonlinear function.

Functional form

Functional form describes how the parameters of a model figure into the model equation (Myung, 2000). We are specifically interested in whether the model's functional form is fixed in advance of model fitting or learned during the model training process. For example, a linear regression model has a pre-fixed functional form, as its model equation is specified in advance, while a regression tree model learns its exact functional form during training by determining the optimal tree splits.

Training time

Training time refers to how long it takes to learn the values of the model parameters from the data. For models with a separate regularization step, training time does not include the time taken to perform regularization.

Prediction time

Prediction time refers to how long it takes to make predictions on a held out test dataset, given a stored model object.

3.1.2 Measuring Efficacy

To measure model efficacy, or how accurately and comprehensively the model represents the phenomenon of interest, we can deploy metrics that measure how well the model performs on both training and test data. Like with the complexity metrics, the efficacy metrics are a function of both the type of model and the particular dataset. Thus, between-model comparison should only be done with models fit to the same data. Importantly, these efficacy metrics seek to both assess goodness of fit to the training data and penalize overfitting by considering fit to the test set. For applications in which preventing overfitting is crucial, it may be necessary to prioritize test set efficacy over all metrics of training set efficacy.

Training set mean squared error (MSE)

Training set MSE (Equation 3.1) measures the discrepancy between each observation's outcome variable value y_i in the training dataset and the model's predicted outcome $\hat{y}_i$ for that observation.

$$\text{MSE}_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.1}$$

Test set mean squared error (MSE)

Test set MSE (Equation 3.2) measures the discrepancy between each observation's outcome variable value y_j in the unseen test dataset and the model's predicted outcome $\hat{y}_j$ for that observation.

$$\text{MSE}_{\text{test}} = \frac{1}{m} \sum_{j=1}^{m} (y_j - \hat{y}_j)^2 \tag{3.2}$$

Akaike information criterion (AIC) and Bayesian information criterion (BIC)

AIC (Equation 3.3) is a measure of goodness of fit. It gives credit for a high maximized likelihood function $\hat{L}$ and penalizes for a greater number of parameters k, thereby guarding against overfitting (Akaike, 1998). A lower AIC value indicates a better fitting model.

$$\text{AIC} = 2k - 2\ln(\hat{L}) \tag{3.3}$$

BIC (Equation 3.4) is another measure of goodness of fit. It is similar to BIC except that it places a stronger penalty on overfitting by incorporating the number of observations n (Schwarz, 1978). A lower BIC value indicates a better fitting model.

$$\text{BIC} = k\ln(n) - 2\ln(\hat{L}) \tag{3.4}$$

AIC and BIC are only computed for frequentist regression models, as they are not directly applicable to Bayesian regression and nonparametric models.

Widely applicable information criterion (WAIC) and approximate leave-one-out cross-validation information criterion (LOOIC)

WAIC (Equations 3.5 and 3.6) is a measure of how well a model is predicted to perform on unseen data. It considers the likelihood p of each observation y_i given the model parameters θ. It also penalizes against overfitting through a penalty term, which takes into consideration θ and the number of observations n (Vehtari et al., 2017). The p_{WAIC} term is referred to as the effective number of parameters in the model.

$$\text{WAIC} = -2\left(\sum_{i=1}^{n} \ln E_\theta \left[p(y_i \mid \theta)\right]\right) + 2p_{\text{WAIC}} \tag{3.5}$$

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_\theta \left(\ln p(y_i \mid \theta)\right) \tag{3.6}$$

LOOIC (Equation 3.7) is another measure of how well a model is predicted to perform on unseen data. It is based on the idea of leave-one-out cross-validation, in which we average over the prediction

error we get when predicting one held-out observation based on all other observations. Like WAIC, LOOIC considers the likelihood p of each observation y_i given the model parameters θ and penalizes against overfitting (Vehtari et al., 2017).

$$\text{LOOIC} = -2\sum_{i=1}^{n} \ln\left(\mathbb{E}_\theta\left[p(y_i \mid \theta)^{-1}\right]^{-1}\right) \tag{3.7}$$

WAIC and LOOIC are only computed for Bayesian regression models, as they are not directly applicable to frequentist regression and nonparametric models.

3.1.3 Measuring Interpretability

Challenges in Measuring Interpretability

Interpretability, or how easily one can obtain meaningful substantive insights from a model, differs from the other two model properties in that it cannot be assessed in a purely quantitative manner. Since interpretability involves both a model and a human doing the interpreting, any interpretability rating system will, by definition, involve some element of human judgment. Further, whether or not a model is interpretable depends highly on the context of one's data and research question. To this end, it would be misguided to prescribe a definitive system for evaluating interpretability across all contexts. Instead, I provide a set of expert-guided interpretability facets that researchers can use to assess their models.

There is limited work on how to consistently assess the interpretability of very simple models on the same scale as very complex models. Different academic disciplines fail to converge on even a basic definition for model interpretability (Lipton, 2016). This lack of agreement appears to stem from the different types of models used in different disciplines, as well as what purpose those models serve for their users. For behavioral and social scientists, models are most often a tool to inform domain knowledge, theory, and practice through a rigorous, data-informed process. For machine learning researchers, models are much more complex, with modern LLMs boasting hundreds of billions of parameters (Meta AI, 2025). The goals of computer scientists when working with these models are varied, but often revolve around achieving humanlike reasoning abilities and other intelligent behavior. In a machine learning context, interpretability refers not to the mapping of model results to substantive takeaways, but rather to the transparency of the model's decisions (Singh et al., 2024). The interpretability of LLMs can be assessed by directly asking the model to describe its reasoning – a feature that is not available in smaller statistical models.

Expert-Guided Interpretability Framework

To move towards an interpretability evaluation framework that is applicable across model complexity and domain, I surveyed twelve statistical modeling experts, all faculty members and postdoctoral

scholars at major research universities. Participants were contacted via email and asked to complete an anonymous online survey. In the survey, they answered questions about how they define model interpretability (free response) and how strongly various factors contribute to their assessment of a model's interpretability (11-point Likert scale). Data collection procedures were approved by the Stanford University Institutional Review Board (eProtocol #80480).

When asked how they would define model interpretability, several experts independently converged on similar points:

1. Understanding the transformation of inputs to outputs and how each input contributes to the outputs (7/12 respondents)
2. Ease of mapping parameters to real-world quantities and theoretical constructs (6/12 respondents)
3. Ability to break down the intermediate steps in the transformation of inputs to outputs (3/12 respondents)
4. Understanding how the model makes decisions and predictions (3/12 respondents)

I integrated these key points into my definition of interpretability given in this thesis: ease of understanding a model's transformation of inputs to outputs and obtaining meaningful substantive insights from the model. The process of interpreting a model involves understanding both the model's mechanisms of action (points 1, 3, and 4) and what the model's parameters represent (point 2).

Participants were also asked to rate nine proposed facets of interpretability (presented in randomized order) on how relevant they felt each facet was for determining how interpretable a model is. Each rating was made using a Likert scale of 0 (not at all relevant) to 10 (extremely relevant). I curated the set of facets by reviewing literature and consulting fellow quantitative social science researchers in my lab. From most to least important, according to expert rankings, the nine facets (and their mean importance ratings out of 10) are:

1. Simplicity of the parametric (or learned nonparametric) form of the model (5.50)
2. Ability to probe causal relationships within the fitted model (4.67)
3. Ease with which one can relate the model results to preexisting domain knowledge (4.50)
4. Approachability of graphical representations to visualize the model (4.00)
5. Degree to which the model informs your understanding of the overall dynamics of a process, regardless of how meaningful individual model parameters are (3.92)
6. Degree to which the model can help inform real-world matters (3.67)

7. Ability to interact with the model as an independent agent (e.g., in the way that one might interact with a large language model through natural language prompting) (3.50)
8. Approachability of quantitative results metrics by which to evaluate the model (3.00)
9. Ease with which one can describe the substantive meaning of specific model parameters (2.75)

The most highly rated facet (simplicity of the model's form) indicates that, at least from these experts' perspectives, less complex models are more interpretable. Additionally, other highly rated facets suggest that understanding causal structures within a model and being able to connect model results to domain knowledge contribute to interpretability.

For researchers looking to evaluate model interpretability in their own work, this set of facets can be easily converted into a quantitative model scoring system. A group of domain experts can rate specific models on each of the facets using Likert scales or rank ordering. Ratings for each model can be made on the basis of that model's equations, code, graphical visualizations, and other materials derived from the modeling process. However, since interpretability is such a context-specific property, a quantitative scoring system should be driven by the research question one seeks to answer, as well as who the target audience of the inquiry is. For example, if a study is run in service of informing a policy decision, it makes sense to weight facet 6 (degree to which model helps inform real-world matters) higher than facet 1 (simplicity of the model's form). When working with LLMs, facet 7 (ability to interact with model) becomes much more relevant, while other facets may become less relevant.

3.2 Models

I now describe the nine models to which I will later apply the model evaluation framework.

3.2.1 Frequentist Linear Model

The frequentist linear model is a standard, non-Bayesian linear regression without mixed effects. It is characterized by an intercept term and linear predictors, as given in Equation 3.8. Subscripts i and t indicate individual- and time-varying coefficients and variables, respectively. The frequentist linear models in this thesis were estimated using ordinary least squares regression with the `lm()` function from the `stats` package in base R (R Core Team, 2025).

$$Y_{it} = \beta_0 + \beta_1 time_t + \beta_2 X_{2it} + \cdots + \beta_n X_{nit} + \epsilon_{it} \tag{3.8}$$

3.2.2 Frequentist Linear Multilevel Model (MLM)

The frequentist linear multilevel model is a non-Bayesian linear regression with mixed effects. This means that both standard linear regression coefficients and group-level random effect terms are estimated. Since we are interested in within-person change over time, we group by participant ID. This model is characterized by an intercept term, linear predictors, and group-level random intercepts and slopes as given in Equations 3.9 and 3.10. The frequentist linear multilevel models in this thesis were estimated using restricted maximum likelihood (REML) estimation with the `lmer()` function from the `lme4` package in R (Bates et al., 2015).

$$Y_{it} = \beta_{0i} + \beta_{1i} time_t + \beta_{2i} X_{2it} + \cdots + \beta_{ni} X_{nit} + \epsilon_{it} \tag{3.9}$$

$$\beta_{ni} = \gamma_{n0} + u_{ni} \tag{3.10}$$

3.2.3 Frequentist Nonlinear Multilevel Model (MLM)

The frequentist nonlinear multilevel model is a non-Bayesian nonlinear regression with mixed effects. The exact functional form of the nonlinearity varies based on the dataset and research question in our examples, but is always a nonlinear function of time. Similar to the linear multilevel model, both standard linear regression coefficients and group-level random effect terms are estimated, again grouping by participant ID. This model is characterized by an intercept term, a nonlinear time predictor, other linear predictors, and group-level random intercepts and slopes as given in Equations 3.11 and 3.12. The frequentist nonlinear multilevel models in this thesis were estimated using restricted maximum likelihood (REML) estimation with either the `nlme()` function from the `nlme` package in R (Pinheiro et al., 2025) or the `lmer()` function from the `lme4` package in R (Bates et al., 2015).

$$Y_{it} = \beta_{0i} + \beta_{1i} \text{nonlinearity}(time_t) + \beta_{2i} X_{2it} + \cdots + \beta_{ni} X_{nit} + \epsilon_{it} \tag{3.11}$$

where `nonlinearity()` is some nonlinear function, such as a logarithmic or exponential function

$$\beta_{ni} = \gamma_{n0} + u_{ni} \tag{3.12}$$

3.2.4 Automatic Differentiation Variational Inference (ADVI) Nonlinear Multilevel Model (MLM)

The ADVI nonlinear multilevel model is exactly the same as the frequentist nonlinear multilevel model (Equations 3.11 and 3.12) with the exception of the estimation method. The model is estimated using an approximate Bayesian variational inference algorithm. Variational inference approximates posterior parameter distributions by using gradient ascent to maximize the evidence lower bound (ELBO) on the log-likelihood of the observed data Equation 3.13. Variational inference turns the computation of an intractable integral into an optimization problem, making the integral less computationally expensive to compute, but often sacrificing some accuracy. The ADVI nonlinear multilevel models in this thesis were estimated with the `brm()` function, specifying the argument `algorithm="meanfield"`, from the `brms` package in R (Bürkner, 2017). `brms` uses a specific version of variational inference called automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2017). ADVI automatically determines which specific variational inference algorithm to use. A weakly informative prior of $\mathcal{N}(\mu = 0, \sigma = 1)$ was used for all parameters.

$$\text{ELBO}(x; Q, \theta) = \sum_{z} Q(z) \log \frac{p(x, z; \theta)}{Q(z)} \tag{3.13}$$

3.2.5 Hamiltonian Monte Carlo (HMC) Nonlinear Multilevel Model (MLM)

The HMC nonlinear multilevel model is exactly the same as the frequentist and ADVI nonlinear multilevel models (Equations 3.11 and 3.12) with the exception of the estimation method. The model is estimated using a Bayesian sampling algorithm. The HMC algorithm is a variant of the Markov chain Monte Carlo (MCMC) algorithm, in which we approximate a distribution by using a Markov process to draw samples. As more samples are drawn, the distribution of the samples becomes closer and closer to the distribution we are approximating (Hastings, 1970; Metropolis et al., 1953). The Hamiltonian Monte Carlo algorithm draws upon the idea of conservation of energy in physical systems to propose new samples (R. M. Neal, 2011). The HMC nonlinear multilevel models in this thesis were estimated with the `brm()` function, specifying the argument `algorithm="sampling"`, from the `brms` package in R (Bürkner, 2017). `brms` uses a specific version of the HMC algorithm called the No-U-Turn Sampler (NUTS) (Hoffman & Gelman, 2011). NUTS automatically sets the number of steps taken in the random walk, thus increasing the efficiency of HMC. A weakly informative prior of $\mathcal{N}(\mu = 0, \sigma = 1)$ was used for all parameters.

3.2.6 Multilevel Regression Tree

The multilevel regression tree is a multilevel extension of a standard regression tree, in which data are recursively partitioned based on predictor values into increasingly smaller nodes. A final multilevel linear regression (Equations 3.9 and 3.10) prediction is made for all observations within a

particular terminal node. The multilevel implementation of the regression tree uses fixed effects predictors to determine optimal splits, then includes person-level (or group-level) random effects in the final regression predictions. The multilevel regression trees in this thesis were estimated with the `lmertree()` function from the `glmertree` package in R (Fokkema et al., 2018; Fokkema & Zeileis, 2024). This function uses restricted maximum likelihood (REML) estimation and the Nelder-Mead algorithm for optimization. It determines the optimal tree by iteratively estimating a linear regression tree without mixed effects, then fitting the mixed effects regression from that tree's terminal nodes (Fokkema et al., 2018). To guard against overfitting, the multilevel regression trees in this thesis were regularized by tuning the maximum tree depth hyperparameter using a held-out validation dataset.

3.2.7 Boosted Multilevel Regression Tree

The boosted multilevel regression tree is an extension of the multilevel regression tree that uses boosting. Boosting is an ensemble tree method in which weak learners, often simple regression trees, are iteratively fit to achieve a strong learner, i.e., the final boosted tree (Schapire, 1999). Each weak learner prioritizes correcting errors made by the previous weak learner by weighting data points with higher previous prediction error more highly. This allows the final strong learner to be robust to overfitting and adequately deal with nonlinearities in the data. The boosted multilevel regression trees in the thesis were estimated with the `gpboost()` function from the `gpboost` package in R (Sigrist, 2022). To further guard against overfitting, the boosted multilevel regression trees in this thesis were regularized by tuning the maximum tree depth hyperparameter using a held-out validation dataset. 10,000 weak learners and a learning rate of 0.001 were used to achieve sufficient fit to the training data.

3.2.8 Longitudinal Random Forest

The longitudinal random forest model is an extension of the random forest technique, in which an ensemble of decision trees are aggregated. Each tree is trained using random subsets of the training observations and predictor variables, and final regression predictions are made by averaging over the individual trees' predictions (Breiman, 2001a). The longitudinal implementation of random forest relaxes constraints to allow the model's covariance structure to change over time (Capitaine et al., 2020). The longitudinal random forest models in this thesis were estimated with the `MERF()` function from the `LongituRF` package in R (Capitaine, 2025). This function models an individual's within-person process as the sum of a fixed-effects random forest model, a random effects term, and a person-specific stochastic process, which models serial correlations of that individual's outcome observations using Brownian motion. An expectation-maximization (EM) algorithm is used to iteratively fit the random effects and the random forest model (Hajjem et al., 2012). To guard against overfitting,

the longitudinal random forest models in this thesis were regularized by tuning the maximum tree depth hyperparameter using a held-out validation dataset.

3.2.9 Multilayer Perceptron

The multilayer perceptron is a fully-connected feedforward neural network, meaning that all neurons in one layer are connected to all neurons in the next layer, and all connections flow in the direction of input to output. In this thesis, all multilayer perceptrons include an input layer, 2 hidden layers, and an output layer. The size of the hidden layers were determined using a validation set, with the second hidden layer being half the size of the first hidden layer in all cases. The `torch` R package (Falbel & Luraschi, 2025), a machine learning framework based on `PyTorch` (Paszke et al., 2019), was used to specify and train neural network models. The multilayer perceptron was trained for 5000 epochs of the gradient-based Adam optimizer (Kingma & Ba, 2015) using `torch`'s built-in automatic differentiation module, `autograd`, and a mean squared error loss function.

Chapter 4

Example 1: Simulated Cognitive Skill Acquisition Data

I next apply the model evaluation framework to a set of simulated data.

4.1 Example 1 Data

This dataset uses a logarithmic function with added noise to simulate individuals' processes of cognitive skill acquisition, which is the development of a cognitive skill through repetitive practice over time (Anderson, 1982). Task completion time has been theorized to take an exponential form, which is conceptually consistent with a logarithmic model of cognitive ability over time (Ritter & Schooler, 2001). This logarithmic example data is useful because (a) it is nonlinear, and thus facilitates a comparison of linear versus nonlinear models; (b) it is relatively simple in a mathematical sense, making it accessible to a wide range of audiences; and (c) it is prevalent in psychological theory, so the simulated data may mirror some real-world datasets of interest to researchers.

Using Equation 4.1, I simulated a dataset with 200 participants and 20 observations per participant. This equation generates a person- and time-varying *skill* value as a function of *time* and person-specific, time-invariant predictors *age*, *iq* (intelligence quotient), and *distracted* (score on a hypothetical distractedness scale). I split each person's time series data into training, development, and test observations with a 10:5:5 ratio, respectively. Observations were assigned to splits at random. I combined all participants' training observations to create the training dataset, and so on for the development and test datasets. The three datasets are visualized in Figure 4.1.

$$
\begin{aligned}
skill_{it} = &(g_{00} + g_{01}\, age_i + g_{02}\, iq_i + g_{03}\, distracted_i + u_{0i}) \\
&+ (g_{10} + g_{11}\, age_i + g_{12}\, iq_i + g_{13}\, distracted_i + u_{1i}) \log(time_t) \\
&+ \epsilon_{it}
\end{aligned} \tag{4.1}
$$

The data-generating function uses values $\{g_{00} = 5, g_{01} = 0.25, g_{02} = 0.5, g_{03} = -0.75, g_{10} = 3, g_{11} = 0.1, g_{12} = 0.4, g_{13} = -0.65\}$. In general, *time*, *age*, and *iq* enable cognitive skill acquisition, while *distracted* inhibits cognitive skill acquisition. *age*, *iq*, and *distracted* are distributed as $\mathcal{N}(\mu = 0, \sigma = 1)$ to simulate the common practice of standardizing predictor variables. The person-specific, time-invariant random effects intercepts are distributed as $u_{0i} \sim \mathcal{N}(\mu = 0, \sigma = 0.7)$ and $u_{1i} \sim \mathcal{N}(\mu = 0, \sigma = 0.5)$. The person-specific, time-varying error ϵ_{it} is distributed as $\mathcal{N}(\mu = 0, \sigma = 2.2)$.

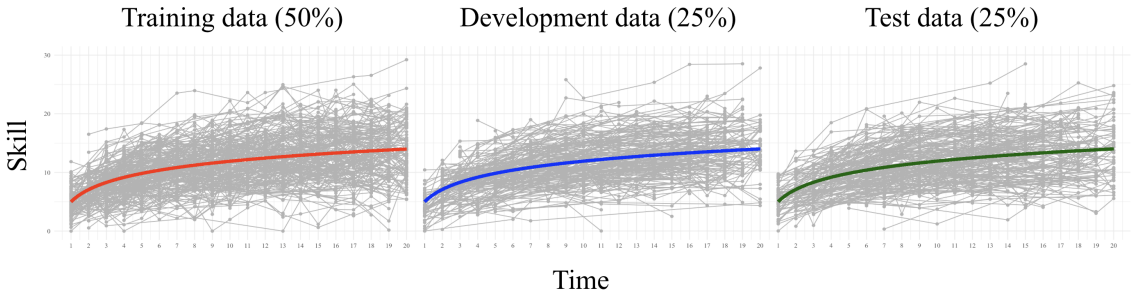

Figure 4.1: Training, development, and test splits for the simulated logarithmic cognitive skill acquisition data.

4.2 Example 1 Research Question

A dataset alone is insufficient to guide model selection. To appropriately synthesize the findings of this model selection framework, we must have a specific research question or goal in mind. The type of question we seek to answer with our model can significantly influence our choice. For example, an exploratory research question about a little studied phenomenon may lead us to weight high complexity and high efficacy over high interpretability. In contrast, if we are testing a very specific hypothesis about the dynamics of a process, an interpretable parametric model may be favored, even if it is less efficacious than nonparametric models.

The research question we are interested in for this example inquiry is, "How do person-specific attributes affect the increase and plateau observed in the process of cognitive skill acquisition?" As we move on to evaluating and comparing models, it is helpful to keep in mind that we seek to

understand (a) the relative contributions of specific predictors and (b) the functional form of the process.

4.3 Example 1 Results

4.3.1 Complexity Results

Complexity metrics assessed for all models are presented in Table 4.1. The number of learned parameters per model varies widely among models, with nonparametric models often having several orders of magnitude more parameters that parametric models. There is also noticeable variability in training time, with times ranging from 0.002 seconds for the frequentist linear model to 648.0 seconds for the multilayer perceptron. While the test set prediction times for the Hamiltonian Monte Carlo logarithmic multilevel model (2.366 seconds) and the boosted multilevel regression tree (1.585 seconds) were greater than that of the other models, none of the models took prohibitively long to make predictions.

Model	# Param.	Linearity	Func. form	Train time	Test time
Frequentist parametric					
Frequentist linear model	6	Linear	Pre-fixed	0.002 s	0.008 s
Frequentist linear MLM	12	Linear	Pre-fixed	0.089 s	0.006 s
Frequentist logarithmic MLM	12	Nonlinear	Pre-fixed	0.193 s	0.002 s
Bayesian parametric					
ADVI logarithmic MLM	12	Nonlinear	Pre-fixed	30.20 s	0.983 s
HMC logarithmic MLM	12	Nonlinear	Pre-fixed	110.6 s	2.366 s
Nonparametric					
Multilevel regression tree	20	Nonlinear	Learned	0.476 s	0.005 s
Boosted multilevel regression tree	1,270,003	Nonlinear	Learned	1.618 s	1.585 s
Longitudinal random forest	159,603	Nonlinear	Learned	16.69 s	0.033 s
Multilayer perceptron	4591	Nonlinear	Learned	648.0 s	0.006 s

Table 4.1: Complexity metrics assessed for all models in Example 1.

4.3.2 Efficacy Results

Efficacy metrics assessed for all models are presented in Table 4.2. Training set MSE, AIC/BIC, and WAIC/LOOIC generally decrease with model complexity. Test set MSE generally decreases, then increases with model complexity, likely due to nonparametric models being overly complex compared to the data-generating function.

Since this example uses a simulated dataset, we can examine the bias between the fitted logarithmic multilevel models and the actual data-generating function. Table 4.3 details the bias for each parameter in the three correctly specified models: frequentist, Bayesian ADVI, and Bayesian HMC.

Model	Train MSE	Test MSE	AIC	BIC
Frequentist parametric				
Frequentist linear model	7.74	8.12	9781.61	9815.22
Frequentist linear MLM	4.89	6.34	9412.73	9479.94
Frequentist logarithmic MLM	4.04	5.16	9044.14	9111.35
			WAIC	**LOOIC**
Bayesian parametric				
ADVI logarithmic MLM	4.19	5.24	8973.55	8981.31
HMC logarithmic MLM	4.03	5.17	8862.20	8866.17
Nonparametric				
Multilevel regression tree	4.05	5.35	–	–
Boosted multilevel regression tree	4.08	6.04	–	–
Longitudinal random forest	3.67	6.87	–	–
Multilayer perceptron	3.93	7.79	–	–

Table 4.2: Efficacy metrics assessed for all models in Example 1.

For all three models, bias is somewhat spread out among the various parameters. The frequentist and HMC models have a similar magnitude (L2 norm across parameters) of bias (0.37 and 0.39, respectively). The ADVI model has noticeably higher magnitude of bias (0.74), suggesting that the approximate posterior inference algorithm, while saving time, costs us some efficacy.

Model	$\boldsymbol{\beta_{00}}$	$\boldsymbol{\beta_{01}}$	$\boldsymbol{\beta_{02}}$	$\boldsymbol{\beta_{03}}$	$\boldsymbol{\beta_{10}}$	$\boldsymbol{\beta_{11}}$	$\boldsymbol{\beta_{12}}$	$\boldsymbol{\beta_{13}}$	$\boldsymbol{u_{00}}$	$\boldsymbol{u_{10}}$	$\boldsymbol{\sigma}$	$\|$Bias$\|$
Freq.	-0.05	-0.02	0.20	0.27	-0.02	0.08	0.08	0.03	0.06	0.01	-0.09	0.37
ADVI	-0.13	0.02	0.19	0.25	0.01	0.06	0.10	0.03	-0.63	0.08	-0.09	0.74
HMC	-0.13	-0.02	0.19	0.26	0.00	0.08	0.09	0.04	0.09	0.04	-0.09	0.39

Table 4.3: Bias in parameter estimates of correctly specified models in Example 1. The rightmost column provides the magnitude, or L2 norm, of bias across parameters.

4.3.3 Interpretability Results

Thinking back to the criteria for assessing model interpretability, we are interested in how each of these models lends itself to (a) an understanding of the transformation of inputs to outputs and (b) a practical understanding of the model results. The frequentist and Bayesian parametric models are easily translated to single equations, with each parameter directly mapping onto some theoretical construct. In contrast, the nonparametric models are much more variable in their internal structure and cannot be summarized in a single equation. Tools like LIME (Ribeiro et al., 2016) and feature importance (Fisher et al., 2019) can help us assess the differential contributions of our predictors to the outcome of cognitive skill in nonparametric models. However, we are specifically

interested in how person-specific predictors contribute to the dynamics of a person's cognitive skill acquisition – that is, how quickly their skill increases, and with what degree of variability. The logarithmic parametric models are highly interpretable with respect to this research question. They have specific parameters that map predictor variables to characteristics of the outcome variable's functional form. We can clearly trace the path of predictors to outcome by examining the model equation. Thus, in the context of our research question, the three logarithmic parametric models appear to be the most interpretable.

4.4 Example 1 Discussion

Now that we have assessed our set of models on the three model properties, we need to synthesize these evaluations to identify our strongest candidate models. Since complexity and efficacy are scored on quantitative scales, we can plot each model on the axes of efficacy and complexity, as in Figure 4.2. To compute overall complexity and efficacy scores, each metric was individually centered (mean = 0) and scaled (standard deviation = 1). Metrics that are inversely correlated with the overall property (e.g., low error is correlated with high efficacy) were multiplied by -1. All complexity metric columns were summed to get each model's overall complexity score. Train and test MSE columns were summed to get each model's overall efficacy score. AIC, BIC, WAIC, and LOOIC were excluded from efficacy score calculation because they were only computed for a subset of the models. Examining Figure 4.2, we observe that efficacy initially increases with complexity, then plateaus. For this dataset, there are diminishing returns on increased model complexity beyond the correctly specified logarithmic models. We identify the set of logarithmic models, as well as the multilevel regression tree, as maximizing efficacy while maintaining reasonable complexity. Our interpretability analysis indicated that the logarithmic models were most interpretable for our research question. Thus, all three model properties converge to suggest that the logarithmic models are our best options. This conclusion makes sense, as our data were simulated from a logarithmic function. In the next example, we will work with a set of empirical data to see how this model evaluation framework fares in a less straightforward setting.

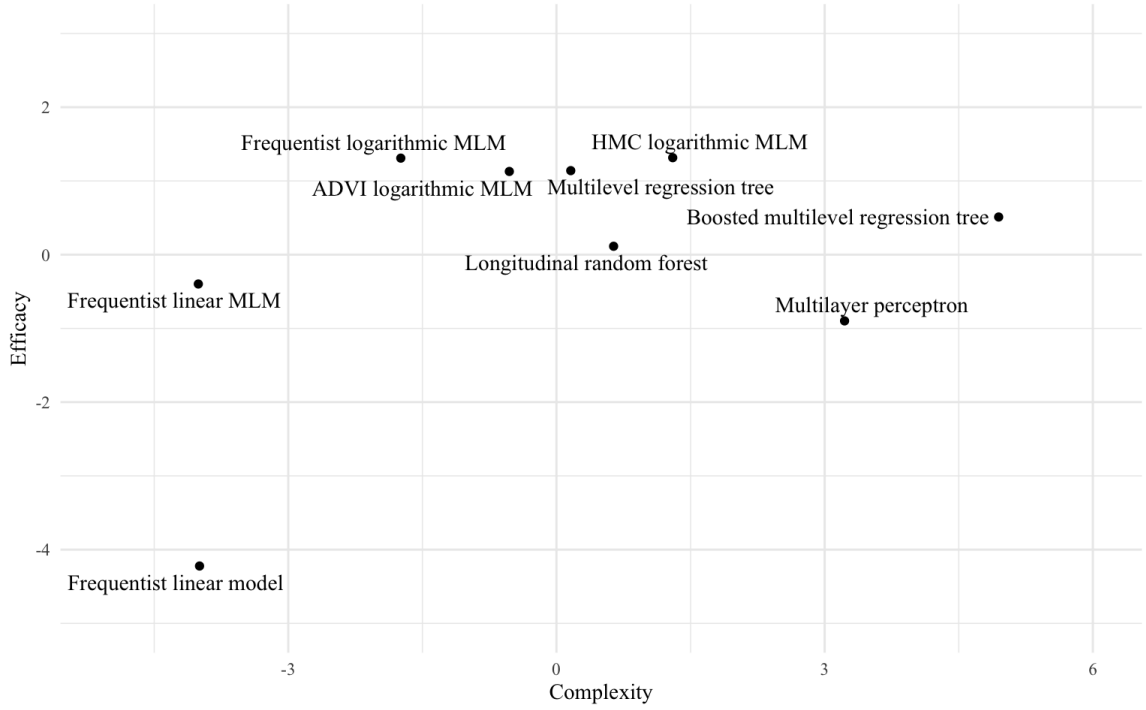

Figure 4.2: Models plotted on the axes of efficacy and complexity. Efficacy and complexity scores were calculated as the sums of relevant scaled and centered columns, with reverse coding of columns where necessary.

Chapter 5

Example 2: Empirical Psychophysiology Data

I next apply the model evaluation framework to a set of empirical data.

5.1 Example 2 Data

This dataset was collected as part of an empirical study of the development of autonomic nervous system function during childhood (Gatzke-Kopp & Ram, 2018). On three separate occasions (kindergarten, first grade, and second grade), researchers collected intensive longitudinal measures of 339 children's sympathetic (cardiac pre-ejection period, PEP; nonspecific skin conductance response, NS-SCR) and parasympathetic (respiratory sinus arrhythmia, RSA) nervous system function while watching emotional film clips. For simplicity, I set aside the developmental aspect of the data and focus only on the time series data from when the children were in first grade. I examine the three physiological variables measured at each of 31 30-second epochs during film viewing. In particular, I examine NS-SCR as the outcome variable and PEP and RSA as time-varying predictor variables. I also consider trait internalizing and externalizing symptoms as person-specific predictors. Like with the Example 1 dataset, I randomly split each person's time series data into training, development, and test observations with a 50:25:25 ratio. I combined all participants' training observations to create the training dataset, and so on for the development and test datasets. The three datasets are visualized in Figure 5.1.

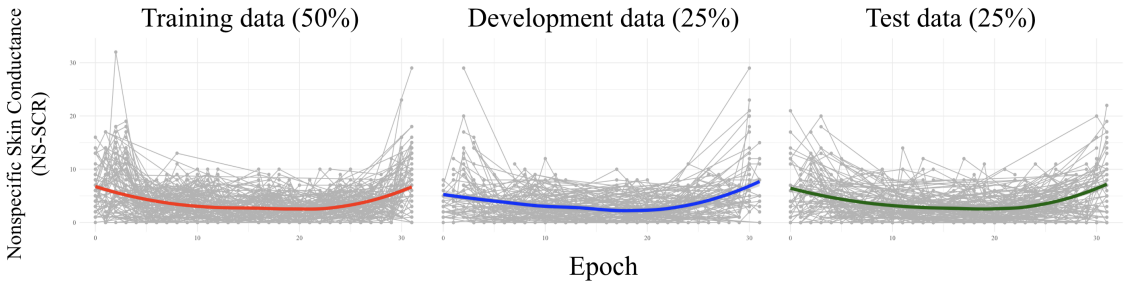

Figure 5.1: Training, development, and test splits for the empirical psychophysiology data.

5.2 Example 2 Research Question

The research question we are interested in for this example inquiry is, "What are the temporal dynamics of children's nonspecific skin conductance response (NS-SCR) during a period of fluctuating emotions, and how are these dynamics related to the changes in other psychophysiological measures?" As we move on to evaluating and comparing models, it is helpful to keep in mind that we seek to understand (a) the relationships among several time-varying constructs and (b) the functional form of our main variable of interest, NS-SCR.

Unlike with the cognitive skill acquisition example, we do not have a strong theory for a functional form guided by previous literature. Instead, our goal is to *find* that functional form by fitting a complex model that will help us uncover the structure of the dynamic process. Once we have identified a suitable functional form, we can go back and fit a simpler parametric model to do more targeted hypothesis testing.

5.3 Example 2 Results

5.3.1 Complexity Results

Complexity metrics assessed for all models are presented in Table 5.1. The number of learned parameters per model varies widely among models, with nonparametric models often having several orders of magnitude more parameters that parametric models. There is also noticeable variability in training time, with times ranging from 0.003 seconds for the frequentist linear model to 228.3 seconds for the multilayer perceptron. These two models also had the minimum and maximum training times, respectively, for the simulated data, suggesting a more general trend in the complexity of these models. While the test set prediction times for the Hamiltonian Monte Carlo logarithmic multilevel model (1.365 seconds) and the boosted multilevel regression tree (1.972 seconds) were greater than that of the other models, none of the models took prohibitively long to make predictions. These two

models also had the highest test set prediction times for the simulated data, suggesting that they are generally more complex models, regardless of the specific data they are fit to.

Model	# Param.	Linearity	Func. form	Train time	Test time
Frequentist parametric					
Frequentist linear model	7	Linear	Pre-fixed	0.003 s	0.001 s
Frequentist linear MLM	17	Linear	Pre-fixed	0.222 s	0.006 s
Frequentist polynomial MLM	20	Nonlinear	Pre-fixed	0.203 s	0.005 s
Bayesian parametric					
ADVI polynomial MLM	20	Nonlinear	Pre-fixed	29.96 s	0.610 s
HMC polynomial MLM	20	Nonlinear	Pre-fixed	40.41 s	1.365 s
Nonparametric					
Multilevel regression tree	40	Nonlinear	Learned	3.159 s	0.006 s
Boosted multilevel regression tree	2,047,0003	Nonlinear	Learned	2.011 s	1.972 s
Longitudinal random forest	198,203	Nonlinear	Learned	2.709 s	0.028 s
Multilayer perceptron	5901	Nonlinear	Learned	228.3 s	0.006 s

Table 5.1: Complexity metrics assessed for all models in Example 2.

5.3.2 Efficacy Results

Efficacy metrics assessed for all models are presented in Table 5.2. Training set MSE, AIC/BIC, and WAIC/LOOIC generally decrease with model complexity. Test set MSE generally decreases, then increases with model complexity. Given that the data do not have a clearly observable parametric form, it is unsurprising that test set MSE is relatively high.

Model	Train MSE	Test MSE	AIC	BIC
Frequentist parametric				
Frequentist linear model	11.2	14.3	8441.80	8479.46
Frequentist linear MLM	7.00	11.2	8215.68	8307.16
Frequentist polynomial MLM	5.94	9.19	7903.90	8011.52
			WAIC	**LOOIC**
Bayesian parametric				
ADVI polynomial MLM	7.46	10.8	8099.14	8110.66
HMC polynomial MLM	5.97	9.18	7765.53	7769.59
Nonparametric				
Multilevel regression tree	5.03	8.68	–	–
Boosted multilevel regression tree	4.31	8.24	–	–
Longitudinal random forest	1.89	11.4	–	–
Multilayer perceptron	1.30	18.4	–	–

Table 5.2: Efficacy metrics assessed for all models in Example 2.

5.3.3 Interpretability Results

Our research question for this example is concerned with discovering how NS-SCR and the other psychophysiology measures co-vary over time. As such, an interpretable model in this context is one that provides rich information about the relationships among variables, including time. While the linear and polynomial parametric models may be more interpretable in a conventional sense, they do not necessarily capture the intricacies of physiological phenomena that are crucial to our research question. The multilevel tree model might be more interpretable for our purposes, as it can be plotted as a decision tree that walks us through the transformation of inputs to outputs. In addition, the smooth functions produced by models like the boosted multilevel regression tree are more complex than the curves of the parametric models, perhaps giving a better approximation to the complex process we are modeling. The intense restrictions on functional form imposed by the parametric models limit expressivity, which in turn hinders our discovery of the process's true functional form.

5.4 Example 2 Discussion

We now synthesize the complexity, efficacy, and interpretability results for Example 2. Like with Example 1, we plot each model on the axes of efficacy and complexity (Figure 5.2). Overall complexity and efficacy scores were calculated using the same method described in the Example 1 Discussion section. Examining Figure 5.2, we observe that efficacy initially increases with complexity, then plateaus. This same pattern was observed with Example 1, suggesting a general trend. Unlike with Example 1, the data used in Example 2 do not have a clear parametric form. Thus, it is unsurprising that efficacy continues to increase as we move into the more complex space of parametric models. In other words, when our data are more complex, we can increase model complexity further before we reach a plateau in efficacy. This observation may explain why LLMs, which are highly complex, are still extremely efficacious – their training data, consisting of massive numbers of websites and documents, is so complex that it requires a sufficiently complex model to achieve high efficacy. In the context of Example 2, the most interpretable models are those that capture the idiosyncrasies of psychophysiological processes. The tree-based nonparametric models (multilevel regression tree, boosted multilevel regression tree, and longitudinal random forest) do a good job of describing fluctuations in both the outcome variable and the predictors. These models are also highly efficacious, thus the framework indicates that they are strong candidate models.

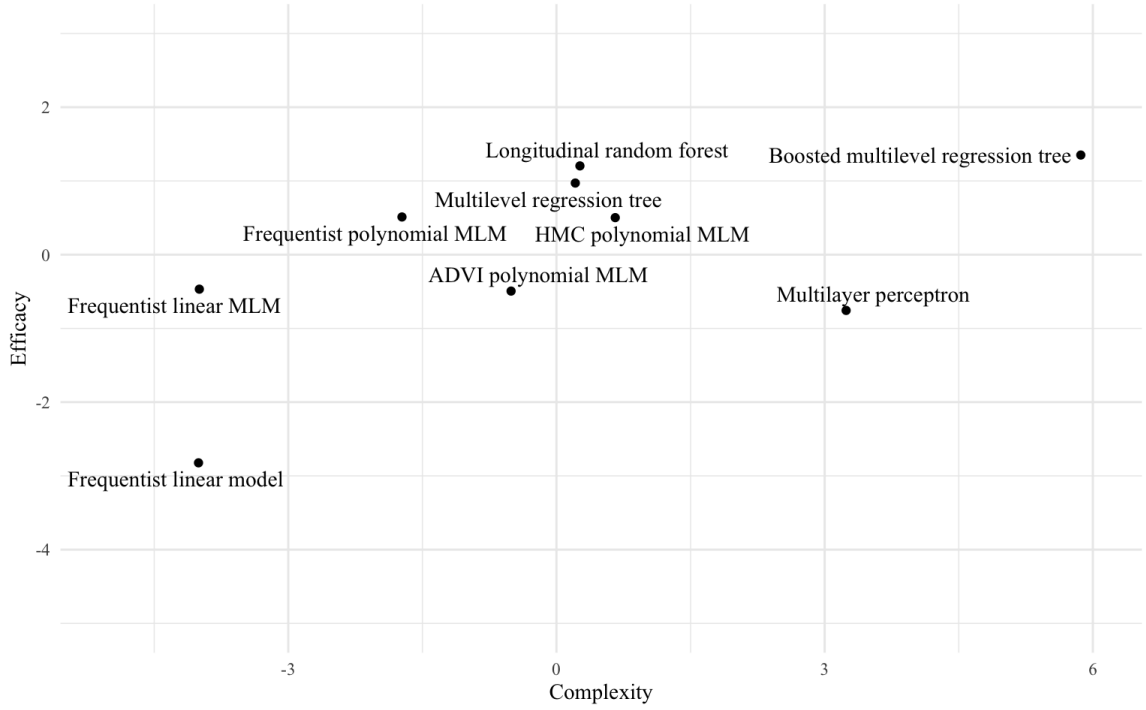

Figure 5.2: Models plotted on the axes of efficacy and complexity. Efficacy and complexity scores were calculated as the sums of relevant scaled and centered columns, with reverse coding of columns where necessary.

Chapter 6

Discussion

6.1 Relationships Among Model Properties

After applying the model evaluation framework to both simulated and empirical data, we can start to think about general relationships among the model properties. We see some support for the hypothesis that efficacy increases with complexity. For both example datasets, however, increasing complexity beyond a certain point did not improve efficacy. At times, overly complex (with respect to a particular dataset) models, like the multilayer perceptron, had lower efficacy than simpler models. Notably, the example inquiries did not examine the performance of LLMs, the most complex models that currently exist. Future work might explore whether efficacy increases again once we reach the level of complexity provided by LLMs.

We also observed that some facets of interpretability decrease with complexity. The ability to map parameters to real-world quantities decreases as we move from simple parametric to complex nonparametric models. However, our ability to understand a model's transformation of inputs to outputs may actually increase with complexity. Simple parametric models may not be expressive enough to capture the dynamics of a complex process, and thereby they obfuscate a more realistic interpretation of how inputs are transformed into outputs. While I did not work directly with LLMs in this thesis, existing literature indicates that LLMs exhibit emergent interpretability through their ability to interact with users as explanatory agents. Future work might investigate whether the interpretability provided by LLMs' self-descriptions is as psychologically meaningful as the interpretability we get from the standard model interpretation process.

6.2 Contributions to Interpretability Literature

Interpretability is a construct often mentioned in computational and social science literature. However, widely accepted definitions of interpretability, as well as methods for measuring it, are rare.

One key contribution of my thesis is an expert-informed definition of interpretability and a set of facets that detail various building blocks of interpretability. I identified two essential features of interpretability: the transparency of the transformation of inputs to outputs, and the mapping of inputs and outputs to external constructs of interest. I then used these features to formulate a new definition of interpretability: "How easily one can understand a model's transformation of inputs to outputs and obtain meaningful substantive insights from the model." My work also indicates that experts prioritize simplicity, causality, and substantive meaning when deciding which models are most interpretable. While assessments of interpretability are still highly dependent on the research question and modeling context, my proposed definition and facets provide first steps toward an objective rating system for interpretability.

6.3 Limitations

I now discuss a few limitations of my framework and my empirical methodology for studying interpretability. The proposed model selection framework does not provide users with a method for identifying the initial set of models to consider. I attempted to provide a relatively comprehensive set of models in my example inquiries, but there are many potential models that I did not include. The choice of which models to consider may significantly impact a researcher's interpretation of complexity, efficacy, and interpretability scores. If one constrains themselves to a small set of very similar models, apparent differences in model property scores may not be meaningful. Examining too many models may become prohibitively time intensive. Due to the extensive diversity in research questions and datasets, it is not possible to prescribe a set of models that will work for every researcher. Nevertheless, my examples provide a starting point for selecting a set of models to evaluate. One might question whether a human-guided model selection framework is necessary when we can ask LLMs to help us select models. While I did not examine the utility of LLMs for doing model selection, it may be a worthwhile line of future work. LLMs may be helpful for reducing the workload of doing model selection using this framework. For example, a researcher might use an LLM to write code for computing complexity and efficacy metrics. However, it is unclear whether LLMs have the domain knowledge and reasoning abilities to assess interpretability. A limitation of the expert-guided interpretability study is that contacted participants were mainly from social science disciplines, such as psychology. Future work might look to include more computational science experts, who bring unique perspectives on what interpretability means.

6.4 Practical Recommendations for Researchers

I now provide some practical recommendations for researchers who want to incorporate more principled model selection into their work.

1. Whatever your model selection process is, document it. Consider adding a paragraph to your method section that outlines what other models you considered and how you eventually settled on your final model.

2. Before you begin computing evaluation metrics, consider how you want to weight complexity, efficacy, and interpretability. Depending on your goals, some properties may be more important than others.

3. Let your research question guide your interpretation of evaluation results, especially for the property of interpretability.

4. Use regularization to prevent complex models from overfitting. Models that have overfit to training data will receive lower efficacy scores than if they were properly regularized.

5. Select efficacy metrics that are most applicable to your models. Be cautious of information criteria like AIC, BIC, WAIC, and LOOIC, which do not apply to all classes of models.

6. Consider assembling a team of raters to assess interpretability. A single individual's ratings of interpretability may be biased by their familiarity with specific models over others. Interpretability raters should ideally be domain experts who understand the substantive meaning of the data.

7. When appropriate, select several models to showcase in your published work. A small set of models with complementary strengths can facilitate a more comprehensive analysis of the data.

Chapter 7

Conclusion

Breiman (2001b) presents us with two classes of models to choose between – simple parametric data models and complex nonparametric algorithmic models. While Breiman's classification provides a helpful distinction, a closer examination reveals that there are many other ways to characterize and compare models. The model properties of complexity, efficacy, and interpretability delineate three crucial dimensions to consider when selecting a model. To move toward more principled model selection, we should start measuring these properties and using the results to inform our decisions. In this thesis, I drew upon existing model evaluation techniques to inform the quantitative assessment of complexity and efficacy. I gathered expert perspectives to inform a new definition and potential measurement system for interpretability, which is less straightforward than the other two properties. Through two example applications of the framework, I provided researchers with guidance on how to apply the framework in their own research. I also discovered potential generalizable relationships among the three model properties. As models continue to get more complex, it will be exciting to see how these relationships evolve.

Appendix A

R Code

R code is provided for the fitting of each model in the first example inquiry. Models for the second example inquiry are largely the same, with the exception of different predictor and outcome variables. Additionally, the nonlinearity in the nonlinear parametric models is quadratic polynomial instead of logarithmic.

A.1 Frequentist Linear Model

```
freq.lm <- lm(skill ~ time + age + iq + distracted,
    data = log_data_train)
```

A.2 Frequentist Linear Multilevel Model

```
freq.lin.mlm <- lmer(skill ~ (age + iq + distracted)
    * time + (1 + time | id), data = log_data_train)
```

A.3 Frequentist Logarithmic Multilevel Model

```
freq.log.mlm <- nlme(skill ~ (b_00 + b_01 * age + b_02 * iq
    + b_03 * distracted + u_0)
    + (b_10 + b_11 * age + b_12 * iq
    + b_13 * distracted + u_1) * log(time),
    data = log_data_train,
    fixed = b_00 + b_01 + b_02 + b_03
```

```
    + b_10 + b_11 + b_12 + b_13 ~ 1,
    random = u_0 + u_1 ~ 1,
    groups = ~id,
    start = c(0, 0, 0, 0, 0, 0, 0, 0))
```

A.4 ADVI Logarithmic Multilevel Model

```
advi.log.mlm <- brm(brms::bf(skill ~ (b00 + b01 * age + b02 * iq
    + b03 * distracted)
    + (b10 + b11 * age + b12 * iq
    + b13 * distracted) * log(time),
    b00 ~ 1 + (1|id), b01 ~ 1, b02 ~ 1, b03 ~ 1,
    b10 ~ 1 + (1|id), b11 ~ 1, b12 ~ 1, b13 ~ 1, nl = TRUE),
    data = log_data_train,
        prior = c(
        prior(normal(0, 1), nlpar = "b00"),
        prior(normal(0, 1), nlpar = "b01"),
        prior(normal(0, 1), nlpar = "b02"),
        prior(normal(0, 1), nlpar = "b03"),
        prior(normal(0, 1), nlpar = "b10"),
        prior(normal(0, 1), nlpar = "b11"),
        prior(normal(0, 1), nlpar = "b12"),
        prior(normal(0, 1), nlpar = "b13")),
    iter = 2000, algorithm = "meanfield",
    control = list(tol_rel_obj = 0.001))
```

A.5 HMC Logarithmic Multilevel Model

```
hmc.log.mlm <- brm(brms::bf(skill ~ (b00 + b01 * age + b02 * iq
    + b03 * distracted)
    + (b10 + b11 * age + b12 * iq
    + b13 * distracted) * log(time),
    b00 ~ 1 + (1|id), b01 ~ 1, b02 ~ 1, b03 ~ 1,
    b10 ~ 1 + (1|id), b11 ~ 1, b12 ~ 1, b13 ~ 1, nl = TRUE),
    data = log_data_train,
    prior = c(
```

```
        prior(normal(0, 1), nlpar = "b00"),
        prior(normal(0, 1), nlpar = "b01"),
        prior(normal(0, 1), nlpar = "b02"),
        prior(normal(0, 1), nlpar = "b03"),
        prior(normal(0, 1), nlpar = "b10"),
        prior(normal(0, 1), nlpar = "b11"),
        prior(normal(0, 1), nlpar = "b12"),
        prior(normal(0, 1), nlpar = "b13")),
    iter = 2000, warmup = 500, chains = 4, cores = 4)
```

A.6 Multilevel Regression Tree

```
lmer.tree <- lmertree(skill ~ age + iq + distracted + time
    | (1 + time | id) | (age + iq + distracted) * time,
    data = log_data_train,
    lmer.control = lmerControl(optimizer  = "Nelder_Mead"),
    maxdepth = maxdepth_optimal)
```

A.7 Boosted Multilevel Regression Tree

```
gpboost.tree.model <- GPModel(
    group_data = as.numeric(log_data_train$id))
gpboost.tree <- gpboost(
    data = as.matrix(log_data_train[, c("time", "age", "iq",
    "distracted")]),
    label = log_data_train$skill,
    gp_model = gpboost.tree.model,
    objective = "regression_l2",
    learning_rate = 0.001,
    nrounds = 10000,
    max_depth = maxdepth_optimal,
    verbose = 0)
```

A.8 Longitudinal Random Forest

```
long.rf <- MERF(
    X = as.matrix(log_data_train[,c("age", "iq",
    "distracted")]),
    Y = as.matrix(log_data_train$skill),
    Z = cbind(intercept = 1, time = log_data_train$time),
    id = log_data_train$id,
    time = log_data_train$time,
    iter = 100,
    ntree = ntree_optimal,
    sto = "BM",
    delta = 0.001)
```

A.9 Multilayer Perceptron

```
nn.model <- nn_module(
    initialize = function(d_in, d_hidden_1, d_hidden_2, d_out) {
        self$layer_1 <- nn_linear(d_in, d_hidden_1)
        self$layer_2 <- nn_linear(d_hidden_1, d_hidden_2)
        self$layer_3 <- nn_linear(d_hidden_2, d_out)
    },
    forward = function(x) {
        x <- self$layer_1(x)
        x <- nnf_relu(x)
        x <- self$layer_2(x)
        x <- nnf_relu(x)
        x <- self$layer_3(x)
        x
    }
)

nn.fit <- nn.model %>%
    setup(loss = nn_mse_loss(),
    optimizer = optim_adam) %>%
    set_hparams(
        d_in = d_in,
        d_hidden_1 = d_hidden_1,
        d_hidden_2 = d_hidden_2,
```

```
    d_out = d_out
) %>%
fit(data_loader, epochs = 5000)
```

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. *Springer Series in Statistics*, 199–213. doi: 10.1007/978-1-4612-1694-0_15

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369–406. doi: 10.1037/0033-295X.89.4.369

Asparouhov, T., & Muthén, B. (2020). Comparison of models for the analysis of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 275-297. doi: 10.1080/10705511.2019.1626733

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. Retrieved from `https://doi.org/10.18637/jss.v067.i01`

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.*, *116*(32), 15849–15854. doi: 10.1073/pnas.1903070116

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. doi: 10.1214/ss/1009213726

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*. Retrieved from `https://doi.org/10.18637/jss.v080.i01`

Capitaine, L. (2025). LongituRF: Random forests for longitudinal data [Computer software manual]. Retrieved from `https://github.com/sistm/longiturf` (R package version 0.9)

Capitaine, L., Genuer, R., & Thiébaut, R. (2020). Random forests for high-dimensional longitudinal data. *Statistical Methods in Medical Research*, *30*(1), 166–184. doi: 10.1177/0962280220946080

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*(403), 596–610. doi: 10.1080/01621459.1988.10478639

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, *57*, 505–528. doi: 10.1146/annurev.psych.57.102904.190146

Falbel, D., & Luraschi, J. (2025). torch: Tensors and neural networks with 'gpu' acceleration [Computer software manual]. Retrieved from `https://torch.mlverse.org/docs` (R package version 0.14.2)

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81. doi: 10.48550/arXiv.1801.01489

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, *50*, 2016–2034. doi: 10.3758/s13428-017-0971-x

Fokkema, M., & Zeileis, A. (2024). Subgroup detection in linear growth curve models with generalized linear mixed model (GLMM) trees. *Behavior Research Methods*, *56*, 6759–6780. doi: 10.3758/s13428-024-02389-1

Gatzke-Kopp, L., & Ram, N. (2018). Developmental dynamics of autonomic function in childhood. *Psychophysiology*, e13218. doi: 10.1111/psyp.13218

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58. doi: 10.1162/neco.1992.4.1.1

Gross, J. J. (2014). Emotion regulation: Conceptual and empirical foundations. *Handbook of Emotion Regulation*. Retrieved from `https://psycnet.apa.org/record/2013-44085-001`

Hajjem, A., Bellavance, F., & Larocque, D. (2012). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328. doi: 10.1080/00949655.2012.741599

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15. doi: 10.1177/0963721416666518

Hastings, W. K. (1970). Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109. doi: 10.2307/2334940

Hoffman, M. D., & Gelman, A. (2011). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, *15*(1), 1593–1623. doi: 10.48550/arXiv.1111.4246

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, *24*(2), 1–11. doi: 10.1093/bib/bbad002

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, *54*, 1688–1715. doi: 10.1002/2017WR021902

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations*. doi: 10.48550/arXiv.1412.6980

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, *18*, 1–45. doi: 10.48550/arXiv.1603.00788

Lipton, Z. C. (2016). The mythos of model interpretability. *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*. doi: 10.48550/arXiv.1606.03490

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Education*, *64*(4), 364–390. doi: 10.1080/00220973.1996.10806604

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, *107*(2), 247–255. doi: 10.1037/0033-2909.107.2.247

Meta AI. (2025). *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.* Retrieved from `https://ai.meta.com/blog/llama-4-multi` (Accessed: 2025-05-28)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. doi: 10.1063/1.1699114

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. doi: 10.1016/s1364-6613(03)00028-7

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190–204. doi: 10.1006/jmps.1999.1283

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., & Mitliagkas, I. (2019). A modern take on the bias-variance tradeoff in neural networks. *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*. doi: 1810.08591

Neal, R. M. (2011). Handbook of Markov chain Monte Carlo. In (chap. MCMC using Hamiltonian dynamics). Chapman and Hall/CRC. doi: 10.1201/b10905

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (pp. 8024–8035). doi: 10.48550/arXiv.1912.01703

Pinheiro, J., Bates, D., & R Core Team. (2025). nlme: Linear and nonlinear mixed effects models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=nlme` (R package version 3.1-168)

R Core Team. (2025). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. doi: 10.1145/2939672.2939778

Ritter, F. E., & Schooler, L. J. (2001). The learning curve. In *International encyclopedia of the social and behavioral sciences* (pp. 8602–8605). Pergamon Press. Retrieved from `https://www.frankritter.com/papers/ritterS01.pdf`

Schapire, R. E. (1999). A brief introduction to boosting. In *Proceedings of the 16th international joint conference on artificial intelligence - volume 2* (pp. 1401–1406). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. doi: 10.5555/1624312.1624417

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. doi: 10.1214/aos/1176344136

Sigrist, F. (2022). Gaussian process boosting. *Journal of Machine Learning Research*, *23*(232), 1–46. Retrieved from `http://jmlr.org/papers/v23/20-322.html`

Singh, C., Inala, J. P., Galley, M., Caruana, R., & Gao, J. (2024). Rethinking interpretability in the era of large language models. *ArXiv Preprint*. doi: 10.48550/arXiv.2402.01761

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology*, *47*, 513–539. doi: 10.1146/annurev.psych.47.1.513

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. doi: 10.1007/s11222-016-9696-4

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, *44*, 41–61. doi: 10.1006/jmps.1999.1276