Statistical Inference and Prediction Across Sparse Model Sets

by

Yagmur Yavuz Ozdemir

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 8, 2026

Keywords: Rashomon sets, High dimensional data, Model selection, p-value aggregation, fMRI, ADHD, Sleep, Epigenetics

Approved by

Roberto Molinari, Chair, Associate Professor of Statistics
Ash Abebe, Professor of Statistics
Jingyi Zheng, Associate Professor of Statistics
Songling Shan, Associate Professor of Mathematics
Francesca Chiaromonte, Professor of Statistics

Abstract

High-dimensional scientific data often support many sparse models with similar predictive performance. In such settings, selecting one final model can hide substantial model uncertainty, especially when predictors are correlated, signals are weak, or several variable combinations provide comparable explanations. This dissertation develops inferential and predictive tools for working with sparse model sets produced by the Sparse Wrapper Algorithm (SWAG). The recovered SWAG library is interpreted as an empirical approximation to a Rashomon Variable Set: a collection of competitive variable subsets rather than a single selected support. The dissertation makes three contributions. First, it develops inference for sparse model sets. A global permutation test assesses whether the recovered library has more concentrated network structure than expected under a null relationship between the response and predictors. Conditional on global evidence of signal, model-specific p-values are aggregated across the library to summarize variable-level evidence. Simulations show approximately nominal type I error, increasing power with signal strength, and more precise variable recovery from George p-value aggregation than from LASSO in the settings considered. Second, the dissertation studies prediction over sparse model sets by stacking the predictions of SWAG library members. Simulations and public-data benchmarks show that stacked sparse model sets can improve held-out prediction while preserving an interpretable representation. Third, the framework is applied to biomedical data, including ADHD classification from resting-state functional connectivity and DNA methylation analysis of childhood sleep timing. Across these examples, sparse model sets provide a practical way to summarize prediction, stability, co-selection, direction of association, and biological follow-up in high-dimensional scientific problems.

Artificial Intelligence (AI) Use Disclosure Statement

In the preparation of this thesis/dissertation, the following Artificial Intelligence (AI) tools were used: ChatGPT and Gemini. These tools were used primarily to assist with editing, brainstorming, code debugging, and summarizing articles. The author acknowledges full responsibility for the intellectual content of this work and has ensured that all AI-assisted sections have been reviewed and revised for accuracy and appropriate academic style. All AI-generated content was reviewed and validated for relevance, appropriateness, and accuracy before incorporation into the final document to maintain scholarly integrity of this research.

Digital Accessibility Disclosure Statement

In the preparation of this dissertation, the following digital accessibility tools were used to ensure this document complies with federal requirements: PAC (PDF Accessibility Checker) and Adobe Acrobat. The author acknowledges full responsibility for the intellectual content of this work and has made a good faith effort to comply with digital accessibility requirements in publishing, wherein the nature of the content does not significantly change in order to do so. Furthermore, all content has been reviewed and revised to meet these requirements prior to final publication.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Roberto Molinari, for making this research possible. He has been not only my academic advisor but also my mentor, always helping me stay on course whenever I felt like giving up. He has supported me unconditionally, believed in me as a researcher, and encouraged me to grow with confidence. I have always looked forward to our meetings, where I learned not only about statistics but also about curiosity, perseverance, and academic integrity. He has been, and will always be, my academic hero.

I would also like to sincerely thank my committee members, Dr. Ash Abebe, Dr. Jingyi Zheng, Dr. Songling Shan, and Dr. Francesca Chiaromonte, and outside reader Dr. Shuai Huang for their guidance, encouragement, and valuable feedback throughout this journey. Their thoughtful comments and support have strengthened this dissertation and helped shape me as a researcher.

Academic life has been challenging in many ways, and as an international student, there is often the additional pressure of wondering whether you are enough. My friends, Busra, Keziban, and Melike, made my years in Auburn unforgettable. Their unwavering emotional support carried me through the most difficult moments of this journey. I cherish our simple, honest, supportive, and nonjudgmental friendship and the incredible girl power we share. Every chapter of this dissertation carries a piece of them.

There are simply not enough pages to express my gratitude to my one and only Ridvan—my husband, my best friend, and my hero. He has done everything possible to make this life worth living. When we met nine years ago, I was at one of the lowest points in my life, and without hesitation, he helped me find my way back. He has loved me unconditionally and supported me in every way imaginable. He listened patiently as I practiced presentations, helped me prepare for teaching, and even supported my students. He was the mathematician I once dreamed of becoming; through his inspiration, I found my own path and became the statistician I was meant to be. Thank you, Ridvan, for believing in me before I believed in myself and for inspiring me

every single day.

Last but certainly not least, I cannot thank my parents enough for their endless love and unwavering support. They tell me how proud they are of me, but I am the one who is truly proud to be their daughter. Although neither of them had the opportunity to continue their education beyond elementary school, they raised a daughter who dared to dream big. Everything I accomplish is a reflection of their sacrifices, and I will spend my life trying to make them proud. My mother, a little girl who was never given the chance to pursue her own education, has always been my greatest inspiration. Mom, I will keep studying for both of us.

Finally, this dissertation is dedicated to all girls and women who were never given the opportunity to study, to dream freely, or to choose the lives they wanted to live. I hope that one day education will never be a privilege but a right that belongs equally to everyone.

Table of Contents

Abstract 2

Artificial Intelligence (AI) Use Disclosure Statement 3

Digital Accessibility Disclosure Statement 4

Acknowledgments 5

List of Tables 9

List of Figures 11

Chapter 1 Dissertation Overview 13

1.1 Main Contributions 14

1.2 Organization of the Dissertation 15

Chapter 2 Literature Review and Positioning 16

2.1 Model Selection, Model Uncertainty, and Replicability 16

2.2 Rashomon Sets and Sparse Model Sets 17

2.3 Inference After Adaptive Search 17

2.4 Prediction, Uncertainty, and Scientific Applications 18

Chapter 3 Statistical Inference for Sparse Model Sets 20

3.1 Introduction 20

3.2 Sparse Wrapper Algorithm (SWAG) 22

3.3 Global Inference for the SWAG Library 24

3.4 Variable-Level Inference via Aggregated P-Values 26

3.5 Validity and Approximation of the Global Test 29

3.6 Simulation Studies 32

3.6.1 Type I Error under the Null 33

3.6.2 Power under Increasing Signal Strength 35

3.6.3 Variable Recovery Using George P-Values 36

3.7 Conclusion 37

Chapter 4 Prediction over Sparse Model Sets 38
4.1 Introduction 38
4.2 Prediction over a SWAG Library 39
4.2.1 Stacking Operator 40
4.2.2 Why Stacking Helps Sparse Model Sets 41
4.2.3 Sparse Consensus Prediction 42
4.2.4 Prediction Disagreement and Uncertainty 43
4.3 Simulation Study 44
4.3.1 Benchmark Evidence on Public Data 47
4.4 Conclusion 48
Chapter 5 Biomedical Applications 50
5.1 Introduction 50
5.2 Neuroimaging Classification of ADHD 51
5.2.1 Data and Preprocessing 51
5.2.2 SWAG Analysis 52
5.2.3 Prediction and Network Interpretation 54
5.2.4 Aggregated P-Values for ADHD Connectivity Features 56
5.3 DNA Methylation and Sleep Timing in Children 58
5.3.1 Study Population 58
5.3.2 SWAG Analysis and Genomic Annotation 59
5.3.3 Methylation Patterns and Biological Follow-Up 60
5.4 Conclusion 62
References 66

List of Tables

3.1 Empirical type I error rates of the entropy-based SWAG test under the null setting. 34
3.2 Empirical power of the entropy-based SWAG test across signal-to-noise ratios (SNR). Power is estimated from 500 Monte Carlo replications at $n = 200$, $p = 50$, and nominal level 0.05. 35
3.3 Comparison of George p-values and LASSO variable selection across SNR levels. All values are simulation averages. 36

4.1 Average test-set R^2 over 500 simulation replications for different signal-to-noise ratios. The SWAG library is stacked using ridge, lasso, XGBoost, and linear meta-models, and is compared with standard lasso and ridge regression as well as individual SWAG models. 45
4.2 Held-out performance of Stacked-SWAG and benchmark methods on public datasets. For regression datasets the metric is R^2; for classification datasets the metric is AUC. Larger values are better. 47

5.1 Site distribution for the ADHD data used in this study. 51
5.2 Most frequent functional-connectivity features in the LOGIT-SWAG library. The second column is the proportion of the 381 models containing the feature. . . . 53
5.3 Region-level frequency summary for the LOGIT-SWAG library. Connectivity features are aggregated by their endpoint brain regions. 55
5.4 Test accuracy for the three classifiers with and without SWAG. Since SWAG returns a set of models, the last column reports the range of test accuracies across the recovered library. 55
5.5 Connectivity features with recorded George-aggregated p-values below 0.05 in the LOGIT-SWAG library. 58
5.6 Characteristics of the sleep-timing study population. 59

5.7 Top 10 CpG target IDs associated with sleep timing from the SWAG analysis. . . . 60

List of Figures

3.1 Empirical null distribution of the entropy of variable selection frequencies obtained from 500 Monte Carlo simulations under the null hypothesis. The light blue histogram represents the distribution of entropy values computed after rerunning the full SWAG procedure for each simulated dataset. The solid red curve shows the Normal approximation derived using the delta method. The close agreement between the histogram and the Normal curve suggests that the delta-method approximation provides a reasonable description of the sampling distribution of the entropy statistic under the null hypothesis. 33

3.2 Empirical CDFs of permutation p-values under the null hypothesis for varying sample sizes. Under correct calibration, p-values follow a Uniform$(0, 1)$ distribution, shown by the diagonal reference line. 34

3.3 Empirical power curve for the entropy-based SWAG test at $n = 200$ and $p = 50$, based on 500 replications. Error bars indicate 95% Monte Carlo standard error. . 35

5.1 Sagittal-view representation of the most frequent LOGIT-SWAG connectivity features. Points represent brain regions and lines represent selected functional-connectivity features between regions. 54

5.2 The corresponding feature names can be found in Table 5.2. The color of the node illustrates the sign of the median of the estimated β coefficients (i.e. the median of all coefficients that a feature takes in each of the SWAG models). The size of each node is proportional to the percentage of models that contain that feature among all SWAG models. The thickness of each link between different nodes is proportional to the percentage of times two features are present together among all SWAG models. The color of the link shows the value of the Spearman correlation coefficient between two different features (blue for positive correlation and red for negative). The grey links represent the connection between features that appear together with less frequency than the ones with red/blue links between them. 57
5.3 Genomic distribution of significant CpG target IDs identified by SWAG. The plot summarizes counts across genomic contexts such as CpG islands, shores, shelves, open sea, and unmapped loci. 61
5.4 Selection frequencies for the top 10 CpG target IDs in the sleep-timing SWAG analysis. Higher percentages indicate greater recurrence across competitive sparse models. 62
5.5 DNA methylation differences between early and late bedtime groups. Panel (a) shows row-wise standardized methylation values for the top 10 genes, with values on the scale -3 to $+3$. Panel (b) shows absolute methylation values for early and late bedtime groups, with p-values from independent t-tests displayed above each plot. 63
5.6 Workflow for cross-referencing SWAG-derived sleep-timing genes with sleep- and obesity-related gene sets. 64

Chapter 1

Dissertation Overview

High-dimensional data analysis often begins with an apparently simple question: which variables matter? In many scientific applications, this question is not answered well by a single selected model. When the number of candidate predictors is large, when predictors are correlated, and when several distinct combinations of variables achieve similar predictive performance, a single sparse model can give a misleading sense of certainty. This instability matters because selected variables are often interpreted as candidate mechanisms, biomarkers, or targets for follow-up study.

This dissertation studies sparse model sets as a response to this problem. The guiding idea is that, in many high-dimensional problems, the object of interest should be a collection of competitive sparse models rather than one final model. The population-level motivation is a Rashomon Variable Set (RVS): a set of variable subsets whose corresponding models achieve performance within a specified tolerance of an optimal or benchmark model. Because an RVS cannot usually be enumerated, especially in large feature spaces, the dissertation focuses on the sparse model set recovered by the Sparse Wrapper Algorithm (SWAG). The recovered SWAG library is treated as a finite, algorithm-dependent approximation to the part of the RVS reached by the search.

This distinction between the scientific target and the recovered library motivates the main questions of the dissertation. If SWAG returns a concentrated group of sparse models, is that concentration evidence of signal, or could it arise from adaptive search under noise? If a variable appears often across the library, is it repeatedly supported by model-specific evidence, or is it mainly a substitute for another correlated variable? If several sparse models predict well, should prediction be based on one of them, an average of them, or a learned combination of their predictions?

1.1 Main Contributions

Chapter 2 first places the dissertation within the literature on model selection, Rashomon sets, replicability, post-selection inference, ensemble prediction, and uncertainty quantification. Chapter 3 then develops an inferential framework for sparse model sets. The SWAG library is represented as a weighted variable network, where nodes are selected variables and edges record co-selection across models. A global permutation test compares the observed library with libraries obtained by rerunning the full SWAG procedure after permuting the response. Simulations show that this test has empirical type I error close to the nominal 0.05 level and increasing power as the signal-to-noise ratio grows. Conditional on global evidence of signal, the chapter aggregates model-specific p-values using the Mudholkar-George logit method. In the simulation study, this aggregation procedure selects fewer variables than LASSO and achieves substantially higher precision, while LASSO is more sensitive but less sparse.

Chapter 4 studies prediction over a sparse model set. Instead of selecting one model from the SWAG library, the predictions of the library members are treated as second-level features and combined through stacked generalization. This construction preserves the interpretability of the base representation because each coordinate of the stacked predictor is an explicit sparse model with a known variable support. The chapter also introduces related summaries, including sparse consensus prediction and a local disagreement diagnostic that records whether library members agree around a new observation. In controlled simulations, the stacked predictor achieves higher held-out R^2 than both the average individual SWAG model and the best individual SWAG model across all signal-to-noise ratios considered. Public-data benchmarks show that Stacked-RVS often improves over regularized generalized linear model baselines and remains competitive with more flexible machine-learning methods, while retaining representation-level interpretability.

Chapter 5 applies the framework to two biomedical problems. The first application uses resting-state functional magnetic resonance imaging data from the ADHD-200 consortium, where each candidate feature is a functional connectivity measure between two brain regions. SWAG reduces 1179 connectivity features to compact logistic and support-vector-machine libraries

involving fewer than 60 distinct features. The logistic-regression library supports interpretation through inclusion frequencies, coefficient signs, co-selection patterns, and aggregated p-values, with recurring evidence involving frontal and temporal connectivity. The second application studies DNA methylation and childhood sleep timing. SWAG screens 865,926 CpG sites and identifies 1006 target IDs mapping to 571 genes. The selected loci are summarized by genomic context and compared with external sleep- and obesity-related gene sets, providing biological contextualization for the recovered sparse model set.

1.2 Organization of the Dissertation

The dissertation is organized around the progression from literature and motivation, to inference, to prediction, to applications. Chapter 2 reviews the methodological literature and clarifies how this dissertation differs from existing approaches. Chapter 3 introduces sparse model sets, defines the SWAG library as the main recovered object, develops the global permutation test, and studies variable-level evidence through George p-value aggregation. Chapter 4 uses the same sparse model set as a prediction space, developing stacked prediction, consensus summaries, and disagreement diagnostics. Chapter 5 demonstrates the framework in neuroimaging and epigenomics, emphasizing how prediction, stability, co-selection, coefficient summaries, and variable-level evidence can be combined in applied scientific interpretation.

Across these chapters, the central message is that model multiplicity should be used rather than suppressed. In high-dimensional scientific problems, the existence of many competitive sparse models contains information about stability, substitutability, interaction among variables, and prediction uncertainty. By treating the recovered sparse model set as the primary object of analysis, this dissertation provides a framework for moving from adaptive sparse search to interpretable inference, stable prediction, and scientifically useful summaries.

Chapter 2

Literature Review and Positioning

This chapter situates the dissertation within several related literatures. The work builds on classical model selection, model uncertainty, Rashomon-set analysis, post-selection inference, ensemble prediction, veridical data science, and distribution-free uncertainty quantification. Its position is at the intersection of these areas: rather than selecting one model, averaging over a black-box collection, or treating multiplicity only as a nuisance, the dissertation treats a recovered collection of sparse variable-subset models as the main statistical object.

2.1 Model Selection, Model Uncertainty, and Replicability

Classical statistical analysis often proceeds by selecting a single model and then interpreting that model as the basis for inference or prediction. Information criteria such as AIC, regularization methods such as the LASSO, and related sparse-selection procedures provide useful tools for obtaining parsimonious models [3, 18]. However, the uncertainty introduced by model selection has long been recognized as a central difficulty [13, 15, 21]. When many models fit nearly equally well, conclusions drawn from one selected model may depend strongly on sampling variation, tuning choices, or the precise search path. Multimodel inference and model averaging respond to this issue by retaining more than one candidate model, but the final target is often still a weighted estimate or prediction rather than a structured analysis of the variables and model relationships that make up the candidate set [12].

The same concern appears in the literature on replicability and researcher degrees of freedom. Multiverse analyses, specification curves, and many-analyst studies show that reasonable analytical choices can lead to different conclusions from the same data [19, 41, 40]. Stability selection addresses one part of this problem by using subsampling to identify variables that are repeatedly selected [27]. The PCS framework similarly emphasizes that reliable data analysis should be assessed through predictability, computability, and stability across reasonable per-

turbations [53, 51]. This dissertation adopts the same broad motivation, but focuses on a more specific object: an algorithmically recovered sparse model set whose variables, co-occurrences, coefficients, p-values, and predictions can all be summarized across competitive models.

2.2 Rashomon Sets and Sparse Model Sets

The dissertation is also directly connected to the Rashomon effect. Breiman used this term to describe settings in which many different functions achieve nearly the same predictive performance [11]. Recent work has formalized Rashomon sets and shown that collections of good models can support interpretability, variable importance, model simplification, fairness, and user choice more effectively than a single fitted model [16, 38, 39]. Several strands of this literature study specific model classes, including sparse decision trees, generalized additive models, and sparse risk scores [49, 54, 26, 35].

This dissertation narrows the Rashomon-set perspective to variable-subset representations. The motivating population object is a Rashomon Variable Set (RVS): a collection of variable combinations whose models are competitive under a specified loss criterion. In practice, the full RVS is not observable in high-dimensional problems. The Sparse Wrapper Algorithm (SWAG) provides a computationally feasible way to recover a finite collection of sparse, high-performing models [29, 32]. The dissertation therefore works with the recovered SWAG library as an empirical approximation to a reachable part of the RVS. This differs from Rashomon-set work that seeks to characterize all good models in a fixed class, and from traditional variable selection that seeks one final support. The recovered set is not treated as complete, but it is treated as informative: it defines a network of variables and co-occurrences that can be studied statistically.

2.3 Inference After Adaptive Search

Inference after model selection is difficult because the selected model is data-dependent. Post-selection and selective-inference methods address this issue by conditioning on, or otherwise accounting for, the selection event [9, 43]. These approaches are important because they

clarify why ordinary p-values from a selected model are not generally valid as if the model had been fixed in advance. The present dissertation shares this concern, but changes the target of inference. Instead of asking only for valid inference within one selected model, it asks whether the recovered sparse model set has structure that is unlikely under a null relationship between the response and predictors, and which variables receive repeated evidence across the set.

Permutation testing provides a natural calibration strategy for this purpose because the entire recovery procedure can be repeated after removing predictive signal while preserving the predictor structure [20, 34, 46]. The global test developed in Chapter 3 uses this idea to calibrate network summaries of the recovered SWAG library. Variable-level inference then aggregates model-specific p-values across models containing the same variable. This connects the dissertation to classical p-value combination methods, including Fisher, Stouffer, Tippett, and Mudholkar-George aggregation [17, 42, 44, 30]. The dissertation's contribution is to place these tools inside a model-set workflow, where the dependence among p-values is induced by overlapping sparse models and by the adaptive search that produced them.

2.4 Prediction, Uncertainty, and Scientific Applications

The prediction component of the dissertation builds on ensemble learning. Stacked generalization combines predictions from multiple base learners using out-of-sample performance, and is closely related to Super Learner and broader ensemble methods [47, 10, 23]. The RVS viewpoint changes the interpretation of the ensemble. The base learners are not arbitrary black-box components; they are explicit sparse models indexed by variable subsets. Thus, the stacked predictor can be studied through its base supports, model weights, inclusion frequencies, and disagreement across sparse explanations. This connects to veridical data-science guidance for final predictive modeling and to recent work on uncertainty quantification under the PCS framework [52, 2, 5].

The uncertainty literature also motivates the dissertation's emphasis on disagreement across competitive representations. Classical discussions of model uncertainty emphasize that

predictive intervals or scientific conclusions conditional on one selected model can understate uncertainty [13, 15, 21]. Conformal prediction and related distribution-free procedures provide finite-sample predictive guarantees under exchangeability for broad classes of prediction algorithms [45, 25, 4, 6]. This dissertation does not fully develop conformal calibration, but it introduces model-set disagreement diagnostics as a step toward uncertainty summaries that reflect both residual error and representation uncertainty.

Finally, the dissertation is motivated by high-dimensional biomedical applications where interpretability is part of the scientific goal. Neuroimaging studies of ADHD require prediction from many functional-connectivity features and benefit from summaries of recurring brain-region relationships [1, 14]. Epigenomic studies of sleep timing involve hundreds of thousands of CpG sites and require screening, annotation, and biological follow-up rather than only black-box prediction [37, 22, 33]. Related genomic work shows that meaningful biological signal may depend on combinations of variables and interaction structures rather than isolated marginal effects [7]. Across these settings, the dissertation positions sparse model sets as a practical middle ground: they preserve interpretable model components, acknowledge multiplicity, and allow inference and prediction to be summarized over a structured collection of competitive explanations.

Chapter 3

Statistical Inference for Sparse Model Sets

This chapter develops an inferential framework for sparse model sets produced by the Sparse Wrapper Algorithm (SWAG). As described in more detail further on, SWAG constructs a library of low-dimensional predictive models rather than selecting a single final model. In the language of Rashomon analysis [38], this library can be viewed as an empirical approximation to a collection of competitive variable-subset models, or Rashomon Variable Set (RVS) which we define further on. The library is useful for interpretation because it shows which variables repeatedly appear in competitive models and which variables tend to appear together. However, the same adaptive search that makes SWAG useful also raises an inferential question: how can we decide whether the recovered model library reflects signal in the data rather than structure induced by noise, finite-sample variation, or the search procedure itself?

The chapter addresses this question in two stages. First, it introduces a global test for the SWAG library based on network summaries of the recovered models. The test compares the observed library to libraries obtained after permuting the response, thereby calibrating the entire search procedure under a null hypothesis of no association between the response and the predictors. Second, conditional on evidence that the library contains signal, the chapter studies variable-level inference by aggregating model-specific p-values across the SWAG library. These steps turn SWAG from a model-discovery tool into an inferential procedure for sparse, interpretable model sets and provide the inferential foundation for the prediction framework developed in the next chapter.

3.1 Introduction

In many predictive modeling problems, several distinct models can achieve nearly the same predictive performance on the same data. Breiman described this phenomenon as the Rashomon effect: a setting in which many different explanations fit the data almost equally well

[11]. More recent work on Rashomon sets has developed this idea into a formal object of study, emphasizing that collections of good models can support interpretability, variable importance, and uncertainty assessment more effectively than a single selected model [16, 49, 54, 26, 35]. This perspective is particularly relevant in scientific settings, where the goal is often not only prediction but also understanding which variables are associated with the response and how stable that association is across plausible analyses.

Sparse learners address interpretability by restricting attention to a small subset of predictors. They are attractive in high-dimensional settings because a model involving a few variables is easier to interpret, communicate, and validate than a model involving many variables. Nevertheless, sparse model selection is unstable when predictors are correlated, the signal is weak, or many models have similar predictive performance. Classical selection procedures may return one model, but that model can obscure the fact that other sparse models provide nearly equivalent evidence [28, 12]. Stability selection and post-selection inference address parts of this problem, but they still heavily rely on the assumption that a single combination of variables can well approximate the underlying data generating process [27, 43]. This dissertation therefore defines and focuses on a new inferential object which assumes that the data generating process is better approximated by a collection of models with different variable combinations. We call this collection a Rashomon Variable Set (RVS) and define it as

$$\mathcal{R}_\tau = \left\{ S \subseteq \mathcal{P} : |S| \leq p_{\text{max}},\ L(S) \leq L^\star + \tau \right\},$$

where $\mathcal{P} = \{1, \ldots, p\}$ is the set of candidate variables and $L(S)$ is the population loss for the subset S. Additionally, $L^\star$ is a reference loss, and τ is a tolerance defining which sparse models are competitive. The reference loss may be the best loss within the sparse class or a benchmark specified independently. This definition emphasizes that the object of interest is not a single support, but a set of variable combinations that provide comparable predictive performance. This dissertation uses the term sparse model set for the recovered object and RVS for the population-

level motivation. The distinction is useful. The scientific target is a set of competitive variable combinations $\mathcal{R}_\tau$, but in finite samples and high dimensions only an algorithmically recovered subset is available. Inference must therefore account for both model multiplicity and the search mechanism used to recover the set.

The Sparse Wrapper Algorithm (SWAG) was developed to construct collections of sparse, high-performing models through a forward-search wrapper procedure [29]. SWAG can be viewed as a structured, computationally feasible way to explore a sparse subset of a Rashomon set: it does not attempt to enumerate all near-optimal models, but it searches for low-dimensional models that perform well under a chosen predictive criterion. SWAG has been applied in biomedical and data-science applications where the number of candidate predictors is large and interpretability is essential [29, 32]. The multi-model output of SWAG motivates two inferential tasks. The first is a global task: determine whether the recovered library has a structure that would be unlikely if the response were unrelated to the predictors. Testing each selected model separately is not satisfactory because it creates a large multiple-testing problem and ignores the dependence induced by the adaptive search. The second task is a variable-level task: after the library has passed a global test, identify which variables receive consistent evidence across the collection of models. This chapter develops both components.

3.2 Sparse Wrapper Algorithm (SWAG)

In practice, $\mathcal{R}_\tau$ cannot usually be enumerated, so the observed SWAG library (which we denote as $\widehat{\mathcal{M}}$) is treated as a finite, data-dependent approximation to the part of this set reached by the algorithm. To formalize the SWAG mechanism, let $\mathbf{y} = (y_1, \dots, y_n)^\top$ denote the response vector and let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$ denote the design matrix. Moreover, we let M_S denote the model fit using the variables indexed by the subset S. Depending on the learning problem, the user can specify $\hat{L}(S)$ as appropriate: it may be a cross-validated prediction error, AIC, misclassification error, or another user-specified criterion. Smaller values of $\hat{L}(S)$ correspond to better performance. The user also specifies a maximum model size

$p_{\max} < p$, a screening proportion $\alpha \in (0, 1)$, and a maximum number of candidate expansions m at each dimension. Given these parameters, SWAG proceeds by building models dimension by dimension as described below.

Screening step. SWAG first fits all one-variable models $M_{\{j\}}$, $j = 1, \ldots, p$, and records their losses $\hat{L}(\{j\})$. Let $q_\alpha^{(1)}$ be the α-quantile of these losses. The retained one-variable model set is

$$\widehat{\mathcal{M}}_1 = \left\{ \{j\} : \hat{L}(\{j\}) \leq q_\alpha^{(1)} \right\}.$$

The corresponding screened variable set is $\hat{\mathcal{S}}_1 = \{j : \{j\} \in \widehat{\mathcal{M}}_1\}$.

General step. For each dimension $d = 2, \ldots, p_{\max}$, SWAG expands retained models from the previous dimension. Candidate models have the form $S \cup \{j\}$, where $S \in \widehat{\mathcal{M}}_{d-1}$ and $j \in \hat{\mathcal{S}}_1 \setminus S$. When the number of possible expansions is larger than the computational budget, SWAG evaluates at most m candidate expansions according to the algorithm's sampling rule. Let $\mathcal{C}_d$ denote the candidate set evaluated at dimension d, and let $q_\alpha^{(d)}$ be the α-quantile of $\{\hat{L}(S) : S \in \mathcal{C}_d\}$. The retained models at dimension d are

$$\widehat{\mathcal{M}}_d = \left\{ S \in \mathcal{C}_d : \hat{L}(S) \leq q_\alpha^{(d)} \right\}.$$

The final SWAG library is the union

$$\widehat{\mathcal{M}} = \bigcup_{d=1}^{p_{\max}} \widehat{\mathcal{M}}_d.$$

Thus, $\widehat{\mathcal{M}}$ is a collection of sparse models that performed well relative to other candidate models considered during the search. It is not guaranteed to equal the full RVS. Some competitive models may be missed because their lower-dimensional prefixes are not retained during forward expansion, while others may be missed because the finite expansion budget does not explore their part of the search graph. This creates two useful distinctions: reachability, which concerns

whether a competitive model can be accessed by the search path, and exploration, which concerns whether a reachable model is actually sampled and retained. The inferential procedures below are therefore calibrated to the recovered library and the algorithm that produced it, rather than to an idealized exhaustive set.

Because SWAG returns a collection of models, it is natural to summarize the output as a weighted variable network. The nodes are variables appearing in at least one selected model. For each variable j, define its selection frequency

$$f_j = \sum_{S \in \widehat{\mathcal{M}}} \mathbf{1}\{j \in S\}.$$

For each pair of variables (j, k), define the co-selection frequency

$$A_{jk} = \sum_{S \in \widehat{\mathcal{M}}} \mathbf{1}\{j \in S,\ k \in S\}.$$

The matrix $A = (A_{jk})$ is the weighted adjacency matrix of the SWAG network. Node weights summarize how often variables appear across the library, while edge weights summarize how often variables appear together. Additional summaries can be layered on the same network, including loss-weighted edge strengths, average coefficients for variables in generalized linear models, and centrality measures for variables that connect otherwise distinct groups of competitive models. This representation is central to the inferential approach below because it converts a set of sparse models into a structured object whose concentration, connectivity, co-occurrence, and centrality can be compared against an appropriate null distribution.

3.3 Global Inference for the SWAG Library

The global inferential question is whether the SWAG library contains more structure than would be expected if the response were unrelated to the predictors. The scientific null hypothesis

is

$$H_0 : Y \perp\!\!\!\perp \mathbf{X},$$

meaning that the predictors do not contain information about the response. Under this null hypothesis, SWAG may still select models that appear predictive because it searches adaptively over many candidates. Therefore, the null distribution must account for the entire SWAG procedure, not only for a fixed fitted model.

It is useful to separate the scientific null from the calibration null. The scientific null states that the response is unrelated to the predictors. The calibration null is the distribution of the recovered SWAG network obtained after removing predictive signal while preserving the dependence structure of the predictors, the sample size, the tuning parameters, and the adaptive search rule. The statistic used for testing can then be any network functional of the recovered library, provided that the same functional is recomputed after each null recovery run. A simple and interpretable global statistic is based on the concentration of variable selection frequencies. Let

$$F = \sum_{j=1}^{p} f_j \qquad \text{and} \qquad \widehat{\pi}_j = \frac{f_j}{F}, \quad j = 1, \dots, p,$$

where variables not appearing in the library have $f_j = 0$. The entropy of the selection-frequency distribution is

$$H(\widehat{\pi}) = -\sum_{j:\widehat{\pi}_j > 0} \widehat{\pi}_j \log(\widehat{\pi}_j).$$

Low entropy indicates that the library concentrates around a small subset of variables; high entropy indicates a more diffuse library. Under the alternative hypothesis, if only a few variables carry signal, the SWAG library should repeatedly select models involving those variables and the observed entropy should tend to be smaller than under the null.

Other network summaries can also be used. For example, eigenvector centrality can identify variables that are connected to other frequently co-selected variables. If A is the weighted

adjacency matrix and $\lambda_{\max}$ is its largest eigenvalue, the eigenvector centrality vector $\mathbf{c}$ satisfies

$$A\mathbf{c} = \lambda_{\max}\mathbf{c}.$$

Entropy provides the main statistic studied in this chapter, while frequency, co-selection, and centrality provide complementary summaries of the selected library. In applications, these summaries answer different scientific questions. Inclusion frequencies identify variables that appear often, co-selection frequencies identify variables that act together, and centrality measures can identify variables that bridge otherwise different predictive explanations.

The global test is calibrated by permuting the response. For $b = 1, \dots, B$, let $\mathbf{y}^{(b)}$ be a random permutation of $\mathbf{y}$. SWAG is then rerun on $(\mathbf{X}, \mathbf{y}^{(b)})$ using the same tuning parameters and the same algorithmic protocol used for the observed data. Let $T_{\text{obs}} = H(\widehat{\pi})$ be the observed entropy and let T_b be the entropy from the bth permuted dataset. Since smaller entropy corresponds to stronger concentration, the one-sided Monte Carlo permutation p-value is

$$\widehat{p}_{\text{global}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T_b \le T_{\text{obs}}\}}{B + 1}.$$

Small values of $\widehat{p}_{\text{global}}$ indicate that the observed SWAG library is more concentrated than expected under the null. This test evaluates the model library as a whole and avoids testing each selected model separately. The validity of this procedure is discussed further on while in the next section we discuss inference on single variables once this global test has confirmed the presence of predictive signal.

3.4 Variable-Level Inference via Aggregated P-Values

A significant global test indicates that the SWAG library contains structure, but it does not identify which variables are responsible for that structure. Variable-level inference is therefore a second-stage analysis. This stage is most directly available when the base learner provides model-specific p-values, as in linear or generalized linear models.

For variable j, define the subset of SWAG models containing that variable as

$$\widehat{\mathcal{M}}(j) = \{S \in \widehat{\mathcal{M}} : j \in S\}.$$

For each $S \in \widehat{\mathcal{M}}(j)$, let $p_{j,S}$ denote the model-specific p-value for testing the coefficient of variable j in model M_S. The goal is to aggregate the collection

$$\{p_{j,S} : S \in \widehat{\mathcal{M}}(j)\}$$

into a single variable-level measure of evidence.

This chapter uses the logit combination method of Mudholkar and George [31, 30]. For variable j, define

$$G_j = \sum_{S \in \widehat{\mathcal{M}}(j)} \log\left(\frac{1 - p_{j,S}}{p_{j,S}}\right).$$

Large positive values of G_j correspond to repeated small p-values across models containing j. Under independence and uniformly distributed null p-values, logit transformations have known distributional properties. In SWAG, however, the p-values are generally dependent because models overlap, share the same data, and are produced by an adaptive search. For this reason, the George statistic should be interpreted either with permutation calibration or, as is the case for this thesis, as a practical aggregation statistic whose empirical behavior will be studied later.

The permutation-calibrated version repeats the complete variable-level calculation under permuted responses at the same time as for the global test. If $G_j^{(b)}$ is the George statistic for variable j in the bth permuted dataset, then an empirical p-value is

$$\hat{p}_j^G = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{G_j^{(b)} \geq G_j\}}{B + 1}.$$

Variables with small $\hat{p}_j^G$ show stronger repeated evidence than expected under the null SWAG search. This approach differs from selecting variables by frequency alone: a variable may appear

often but have weak model-specific evidence, or appear less often but have consistently strong evidence when it is selected.

Other p-value combination methods provide useful comparisons. Fisher's method emphasizes very small p-values [17]; Stouffer's method combines inverse normal scores and can incorporate weights [42]; Tippett's method focuses on the minimum p-value [44]. These methods also correspond to different scientific notions of relevance. A minimum-p approach asks whether a variable is strongly supported in at least one competitive representation, while George aggregation asks whether evidence accumulates across many representations. The logit-based George statistic is attractive in the present setting because it uses evidence from all models containing a variable while remaining sensitive to consistently small p-values.

The same model-set representation also permits conditional versions of variable-level inference. For example, let $\mathcal{C}$ denote a condition on the models in the library, such as containing a selected neighbor of j, belonging to a network community, or including a possible substitute variable. Define

$$\widehat{\mathcal{M}}(j;\mathcal{C}) = \{S \in \widehat{\mathcal{M}} \,:\, j \in S,\ S \text{ satisfies } \mathcal{C}\}.$$

A conditional aggregated statistic has the form

$$P_j(\mathcal{C}) = \Psi\{p_{j,S} \,:\, S \in \widehat{\mathcal{M}}(j;\mathcal{C})\},$$

where Ψ is a p-value aggregation rule. This formulation distinguishes variables whose evidence is stable across many contexts from variables that are significant only when they appear with particular partners or without particular competitors. The simulations below focus on marginal aggregation, but the conditional formulation explains why the network representation is more informative than a list of selected variables.

3.5 Validity and Approximation of the Global Test

The global SWAG test is a permutation test applied to a statistic computed after an adaptive model-building procedure. The key point is that permutation validity does not require the statistic to be simple. It requires the statistic to be recomputed in the same way under each permuted response. More specifically, let G_n denote the group of all permutations of the response indices. For a permutation $g \in G_n$, let $g\mathbf{y}$ denote the permuted response vector. Let

$$T(\mathbf{X}, \mathbf{y})$$

be any statistic produced after applying the full SWAG procedure to $(\mathbf{X}, \mathbf{y})$, such as the entropy of selection frequencies. Under H_0, $(\mathbf{X}, \mathbf{y})$ and $(\mathbf{X}, g\mathbf{y})$ have the same distribution for every $g \in G_n$.

Proposition 3.1 (Finite-sample validity under exchangeability). Assume that H_0 holds and that the response is exchangeable with respect to G_n conditional on $\mathbf{X}$. If the full SWAG procedure and the statistic T are applied to every permuted response $g\mathbf{y}$, then the exact permutation test based on the resulting values has finite-sample level α.

Proof. Under the null, the joint distribution of $(\mathbf{X}, \mathbf{y})$ is invariant under permutations of $\mathbf{y}$. Therefore the collection $\{T(\mathbf{X}, g\mathbf{y}) : g \in G_n\}$ is exchangeable under H_0. The standard randomization-test argument then implies that the rank of the observed statistic among its permutation values is uniformly distributed, up to ties, and the resulting exact permutation test controls type I error at level α [24, 20]. □

In practice, running *all* permutations is often infeasible except for very small sample sizes. The Monte Carlo version samples B random permutations and uses the corrected p-value

$$\widehat{p} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T_b \leq T_{\text{obs}}\}}{B + 1}$$

for lower-tail statistics such as entropy. The $B + 1$ correction prevents zero p-values and gives

conservative finite-sample behavior for random permutation samples [34]. As B increases, the Monte Carlo p-value converges to the exact permutation p-value.

The computational cost of rerunning SWAG many times motivates approximate calibration. One useful approximation is based on the empirical selection-frequency vector. Suppose that, under a fixed null mechanism, the total number of variable appearances is

$$F = \sum_{j=1}^{p} f_j,$$

and let π_0 denote the corresponding null selection-frequency probabilities. If the frequency vector behaves approximately like a multinomial count vector, then

$$\sqrt{F}(\hat{\pi} - \pi_0) \xrightarrow{d} N\{0, \Sigma(\pi_0)\}, \qquad \Sigma(\pi_0) = \mathrm{diag}(\pi_0) - \pi_0 \pi_0^{\top}.$$

For the entropy function $H(\pi) = -\sum_j \pi_j \log(\pi_j)$, the delta method gives

$$\sqrt{F}\{H(\hat{\pi}) - H(\pi_0)\} \xrightarrow{d} N\Big(0, \nabla H(\pi_0)^{\top} \Sigma(\pi_0) \nabla H(\pi_0)\Big).$$

This approximation is not a substitute for permutation calibration in all settings, because SWAG creates dependence through adaptive search. It is nevertheless useful as a diagnostic and as a possible computational shortcut when full permutation calibration is expensive.

To investigate the accuracy of the delta-method approximation, we conducted a Monte Carlo simulation study under the global null hypothesis. The simulation was designed to mimic the setting considered throughout this dissertation.

A predictor matrix with $n = 2000$ observations and $p = 500$ predictors was generated from the multivariate normal distribution

$$\mathbf{X}_i \sim N_p\left(\mathbf{0}, \frac{1}{p}\mathbf{I}_p\right),$$

and held fixed throughout the experiment. The regression coefficients were set to

$$\beta = \mathbf{0},$$

so that no predictor was associated with the response. An initial binary response vector was generated from the intercept-only logistic model

$$\Pr(Y_i = 1 \mid \mathbf{X}_i) = \frac{\exp(1)}{1 + \exp(1)},$$

which corresponds to the null hypothesis.

For each of $H = 500$ Monte Carlo replicates, a bootstrap sample of the null response vector was obtained by sampling the responses with replacement. SWAG was then rerun using $p_{\max} = 10$ and $\alpha = 0.05$, and the resulting sparse model library was used to compute an empirical selection-frequency vector. Let

$$\mathbf{f}^{(h)} = (f_1^{(h)}, \dots, f_p^{(h)})$$

denote the number of times each variable appeared in the selected models during replicate h. These counts were normalized to obtain the empirical selection probabilities

$$\hat{\pi}^{(h)} = \frac{\mathbf{f}^{(h)}}{\sum_{j=1}^{p} f_j^{(h)}},$$

from which the entropy

$$H\left(\hat{\pi}^{(h)}\right) = -\sum_{j=1}^{p} \hat{\pi}_j^{(h)} \log \hat{\pi}_j^{(h)}$$

was computed. Repeating this procedure yielded 500 entropy values, whose empirical distribution forms the histogram in Figure 3.1.

To construct the Normal approximation, the empirical selection probabilities from one

representative simulated SWAG library were used to estimate the null selection probabilities,

$$\widehat{\pi}_0.$$

Under the multinomial approximation, the covariance matrix was estimated by

$$\widehat{\Sigma} = \mathrm{diag}(\widehat{\pi}_0) - \widehat{\pi}_0\widehat{\pi}_0^{\top}.$$

The corresponding delta-method variance was computed as

$$\hat{\sigma}^2 = \frac{\nabla H(\widehat{\pi}_0)^{\top}\widehat{\Sigma}\nabla H(\widehat{\pi}_0)}{F},$$

where F denotes the total number of variable appearances in the corresponding SWAG library. For visualization, the Normal curve was centered at the empirical mean of the simulated entropy values and assigned variance $\hat{\sigma}^2$. As shown in Figure 3.1, the empirical distribution of the entropy is reasonably well approximated by the delta-method Normal distribution, providing preliminary evidence that the asymptotic approximation captures the finite-sample behavior of the entropy statistic under the null.

3.6 Simulation Studies

In this section simulation studies evaluate two aspects of the proposed inferential framework. The first is the calibration and power of the entropy-based global SWAG test. The second is the variable-selection behavior of George p-value aggregation relative to LASSO. For each simulation replicate, the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ has independent standard normal entries. The response is generated from

$$Y = \mathbf{X}\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I_n).$$

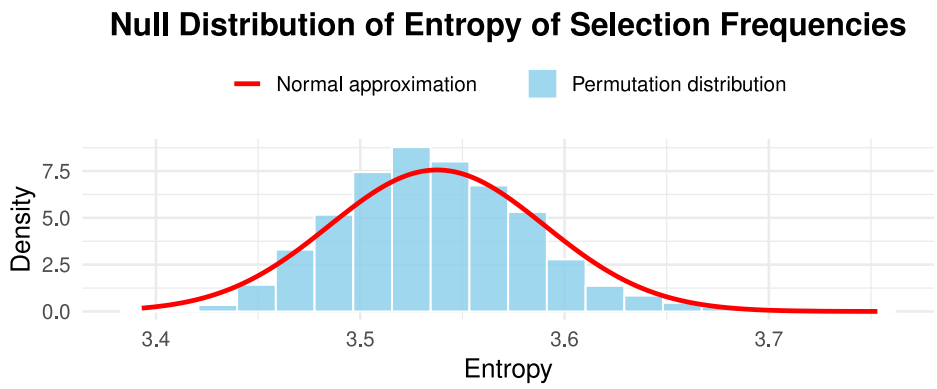

Figure 3.1: Empirical null distribution of the entropy of variable selection frequencies obtained from 500 Monte Carlo simulations under the null hypothesis. The light blue histogram represents the distribution of entropy values computed after rerunning the full SWAG procedure for each simulated dataset. The solid red curve shows the Normal approximation derived using the delta method. The close agreement between the histogram and the Normal curve suggests that the delta-method approximation provides a reasonable description of the sampling distribution of the entropy statistic under the null hypothesis.

Under the null setting, all entries of β are zero. Under the alternative setting, the first five coefficients are nonzero, with $\beta_1 = \cdots = \beta_5 = 2$, and the remaining coefficients are zero. This design makes the true signal known, which allows direct evaluation of both global detection and variable recovery.

3.6.1 Type I Error under the Null

The type I error simulation evaluates whether the entropy-based test maintains the nominal significance level when no true signal is present. The empirical rejection rates in Table 3.1 are close to the nominal 0.05 level for all three sample sizes. The small deviations are consistent with Monte Carlo uncertainty and suggest that the test is slightly conservative in these settings. The validity of this approach is supported by Figure 3.2 where we observe that the p-values of this test follow a standard uniform distribution, as would be expected under the null hypothesis.

Table 3.1: Empirical type I error rates of the entropy-based SWAG test under the null setting.

Sample size (n)	Nominal level	Empirical type I error
100	0.05	0.040
200	0.05	0.042
500	0.05	0.036

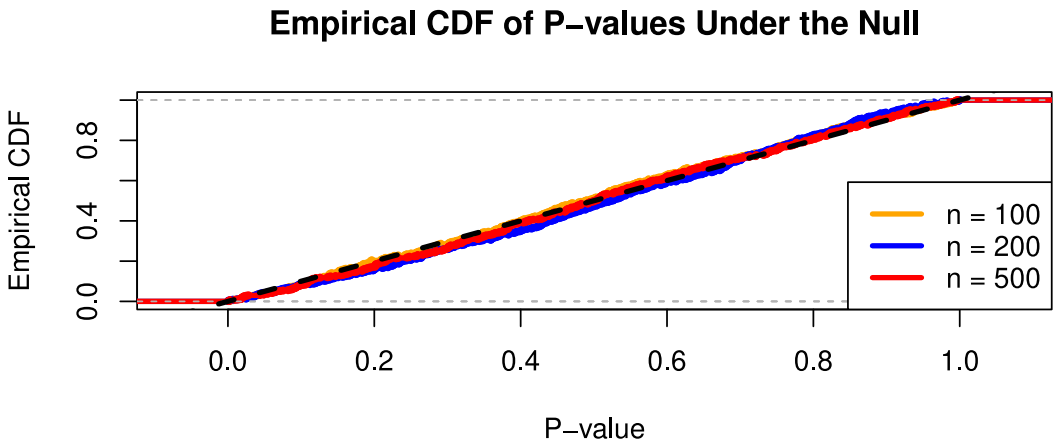

Figure 3.2: Empirical CDFs of permutation p-values under the null hypothesis for varying sample sizes. Under correct calibration, p-values follow a Uniform$(0, 1)$ distribution, shown by the diagonal reference line.

3.6.2 Power under Increasing Signal Strength

To study power, we fix $n = 200$ and $p = 50$ and vary the signal-to-noise ratio

$$\text{SNR} = \frac{\text{Var}(\mathbf{X}\beta)}{\sigma^2}.$$

The SNR values are $0.1, 1, 10, 100$, and 1000. The resulting empirical powers are shown in Table 3.2. Overall, power increases as the signal becomes stronger, reaching 0.906 at SNR $= 1000$. The nonmonotonic value at SNR $= 10$ reflects finite simulation variability and the adaptive nature of the SWAG search, but the broader trend confirms that the entropy statistic is sensitive to signal concentration in the recovered library.

Table 3.2: Empirical power of the entropy-based SWAG test across signal-to-noise ratios (SNR). Power is estimated from 500 Monte Carlo replications at $n = 200$, $p = 50$, and nominal level 0.05.

SNR	Empirical power	MCSE
0.1	0.342	0.0217
1	0.514	0.0223
10	0.464	0.0223
100	0.692	0.0203
1000	0.906	0.0127

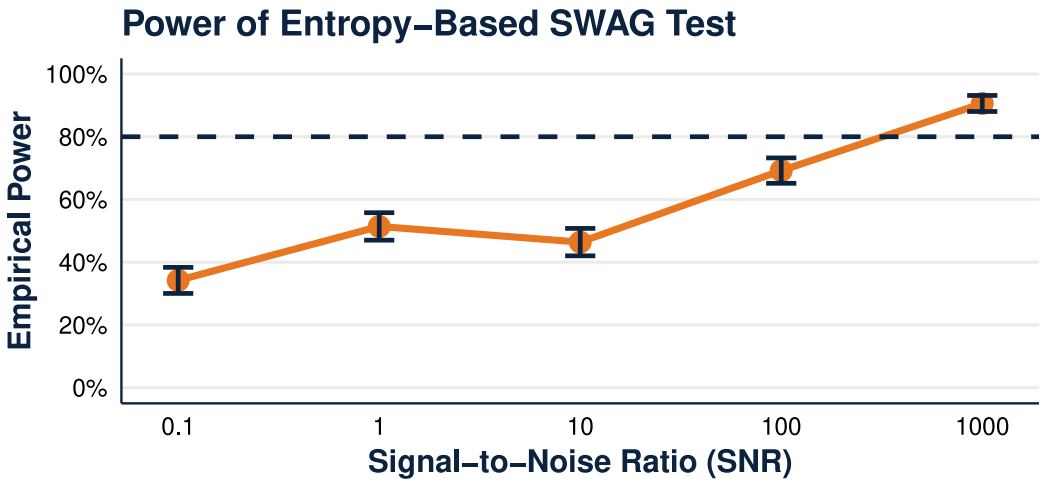

Figure 3.3: Empirical power curve for the entropy-based SWAG test at $n = 200$ and $p = 50$, based on 500 replications. Error bars indicate 95% Monte Carlo standard error.

3.6.3 Variable Recovery Using George P-Values

Table 3.3 compares the George p-value aggregation procedure with LASSO variable selection. The George procedure selects fewer variables on average and has substantially higher precision across all SNR levels. LASSO has higher true positive rates, especially for SNR ≥ 1, but it selects many more noise variables, leading to lower precision and lower F1 scores in most settings. These results illustrate the intended role of George aggregation: it is not designed to select every possible signal, but to identify variables with repeated evidence across the SWAG library while maintaining a sparse and interpretable selection.

Table 3.3: Comparison of George p-values and LASSO variable selection across SNR levels. All values are simulation averages.

Method	Metric	SNR				
		0.1	**1**	**10**	**100**	**1000**
George	Mean selected	2.345	4.432	4.595	5.147	5.611
George	TPR	0.388	0.800	0.800	0.800	0.800
George	Precision	0.828	0.903	0.871	0.777	0.713
George	F1 score	0.529	0.848	0.834	0.788	0.754
LASSO	Mean selected	8.69	16.16	16.17	16.30	10.89
LASSO	TPR	0.555	1.000	1.000	1.000	1.000
LASSO	Precision	0.319	0.309	0.309	0.307	0.459
LASSO	F1 score	0.406	0.473	0.472	0.470	0.629

At the weakest signal level, George aggregation selects an average of 2.345 variables, with precision 0.828 and F1 score 0.529. LASSO recovers more true variables but selects many more variables overall, resulting in lower precision. For SNR ≥ 1, LASSO reaches TPR 1.000, but its mean selected model size remains much larger than the true support size. George aggregation maintains TPR 0.800 while producing a more compact set of selected variables. This trade-off is desirable in settings where the scientific goal is to identify a smaller set of variables with stable evidence rather than to maximize sensitivity at the cost of many false positives.

3.7 Conclusion

This chapter developed an inferential framework for sparse model sets produced by SWAG. The framework begins with a global permutation test that asks whether the recovered SWAG library has more concentrated network structure than expected under a null relationship between the response and the predictors. By rerunning the entire SWAG procedure after permuting the response, the test calibrates the adaptive search itself rather than treating the selected models as fixed.

The chapter then introduced a variable-level procedure based on George p-value aggregation. This second-stage analysis combines model-specific evidence across all SWAG models containing a given variable. Because SWAG models overlap and are generated adaptively, the aggregated statistics require empirical study and, when possible, permutation calibration. In simulations, the George procedure produced substantially sparser selections and higher precision than LASSO, while maintaining stable recovery of the strongest signals.

Together, the global SWAG test and the George aggregation procedure provide a path from sparse model discovery to statistical inference. The global test assesses whether the model library contains meaningful structure, while the variable-level analysis identifies variables that receive repeated support across the library. The same framework also indicates how future conditional tests could assess co-occurrence, substitution, and community-level evidence. These tools support the broader goal of using sparse model sets not only for prediction, but also for stable and interpretable scientific discovery.

Chapter 4

Prediction over Sparse Model Sets

This chapter studies how a collection of sparse models can be used for prediction. The previous chapter treated the SWAG library as an inferential object: the goal was to decide whether the selected variables and the corresponding network structure contain more information than would be expected under a null relationship between the response and the predictors. Here the goal is different. We assume that a sparse model set has been statistically validated and ask how its members can be combined to produce accurate and stable predictions.

4.1 Introduction

SWAG was introduced as a method for constructing a library of low-dimensional models with strong predictive performance [29]. This library is useful for interpretation because it displays which variables recur across competitive sparse models and which variables tend to appear together. In the framework of the previous chapter, the library is also an empirical representation of a recovered RVS: each coordinate of the representation is an interpretable predictive model indexed by an explicit variable subset. However, a sparse model set should not only be interpretable. In applied domains such as genetics, biology, medicine, and neuroimaging, the selected models must also deliver reliable predictions on new observations. If the recovered library contains several competitive models, selecting only one of them discards information about the remaining models and reintroduces the instability that motivated the multi-model perspective in the first place.

The central idea of this chapter is to use the sparse model set as a prediction space. Rather than treating the models in the SWAG library as competitors, we treat their predictions as new features and combine them through stacked generalization. Stacking was introduced by Wolpert [47] and developed for regression by Breiman [10]; the same principle underlies broader ensemble and Super Learner constructions [23]. Its purpose is to learn how to combine

a set of base learners using out-of-sample predictions, so that the final predictor can exploit complementary information across the library. This is closely related to the broader shift from selecting a single best model to working with multiple good models, as in multimodel inference and Rashomon-set analysis [12, 11, 16].

This perspective is particularly natural for SWAG. Each model in the library uses a small subset of variables and therefore gives a simple predictive explanation. Different models may obtain similar accuracy because they capture different parts of the signal, use correlated substitutes, or perform better in different regions of the predictor space. Stacking does not require us to decide in advance which sparse model is the final model. Instead, it estimates how much each model should contribute to prediction. In this way, the final predictor is no longer a single sparse model, but a weighted prediction over a sparse model set.

The approach also aligns with the stability principle in the Predictability - Computability - Stability (PCS) framework [51]. A conclusion based on one selected model may be sensitive to small changes in the data or to the details of the selection procedure. A stacked predictor uses cross-validated predictions from many sparse models and therefore provides a way to aggregate across model uncertainty. This does not remove the need for validation, but it makes the prediction step consistent with the broader goal of obtaining conclusions that are stable across reasonable analytical choices. The resulting object is not a black-box representation in the usual sense: its inputs are predictions from explicit sparse models, so the fitted ensemble can still be studied at the level of the base representations and the variables that define them.

4.2 Prediction over a SWAG Library

Let

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$$

denote the observed data, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ in the regression setting considered in this chapter. As in the previous chapter, let $\mathcal{P} = \{1, \dots, p\}$ denote the index set of available predictors.

A SWAG run returns a library

$$\widehat{\mathcal{M}} = \{S_1, \dots, S_K\},$$

where each $S_k \subseteq \mathcal{P}$ satisfies $|S_k| \leq p_{\max}$. The fitted prediction rule from model M_{S_k} is denoted by $\widehat{f}_k$. If the models in the SWAG library are based on linear regression, for example, then we have

$$\widehat{f}_k(\mathbf{x}) = \hat{\beta}_{0k} + \mathbf{x}_{S_k}^{\top} \hat{\beta}_{S_k},$$

where $\mathbf{x}_{S_k}$ is the subvector of predictors indexed by S_k.

The SWAG library can be viewed as an empirical sparse model set: it is not necessarily the full Rashomon set or the full RVS, but it is a structured subset of competitive sparse models found by a forward wrapper search. This distinction is important. The prediction method developed here does not require the library to contain every near-optimal model. It only requires a collection of accurate and sufficiently diverse sparse models whose prediction errors are not perfectly redundant. The prediction matrix built from the functions $\widehat{f}_1, \dots, \widehat{f}_K$ is therefore a learned representation of the observations in the space of competitive sparse explanations.

4.2.1 Stacking Operator

Stacking constructs a second-level regression problem using out-of-sample predictions from the base models. Let $\widehat{f}_k^{(-i)}(\mathbf{x}_i)$ denote the prediction for observation i from model M_{S_k} when the fitting procedure for M_{S_k} is trained without using observation i, or more generally without using the validation fold containing observation i. Define the level-one prediction matrix

$$\mathbf{Z} = \begin{bmatrix} \widehat{f}_1^{(-1)}(\mathbf{x}_1) & \cdots & \widehat{f}_K^{(-1)}(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \widehat{f}_1^{(-n)}(\mathbf{x}_n) & \cdots & \widehat{f}_K^{(-n)}(\mathbf{x}_n) \end{bmatrix}.$$

The stacked estimator is then

$$\widehat{f}_{\text{stack}}(\mathbf{x}) = \sum_{k=1}^{K} \widehat{\alpha}_k \widehat{f}_k(\mathbf{x}),$$

where the weight vector $\widehat{\alpha} = (\widehat{\alpha}_1, \dots, \widehat{\alpha}_K)^\top$ is estimated by minimizing cross-validated squared error,

$$\widehat{\alpha} = \arg\min_{\alpha \in \mathcal{A}} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} \alpha_k \widehat{f}_k^{(-i)}(\mathbf{x}_i) \right)^2 .$$

The set $\mathcal{A}$ specifies the allowed stacking weights. A common choice is the convex set

$$\mathcal{A} = \left\{ \alpha : \alpha_k \geq 0, \sum_{k=1}^{K} \alpha_k = 1 \right\},$$

which keeps the stacked prediction inside the span of the base predictions and gives the weights a direct interpretation. One may also use unconstrained least squares or regularized least squares when the number of base models is large or when their predictions are highly correlated. In the simulations below, the meta-learner is linear and is trained using the same training/test split used to evaluate the SWAG library.

The use of out-of-sample predictions is essential. If the second-level regression were trained on in-sample predictions, the meta-learner could exploit overfitting in the base models and produce optimistic estimates of prediction error. Cross-validation separates the construction of the base predictions from the estimation of the stacking weights, following the logic of stacked generalization [47, 10].

4.2.2 Why Stacking Helps Sparse Model Sets

Stacking is useful when the library contains models that are both accurate and diverse. If all models in $\widehat{\mathcal{M}}$ produce nearly identical predictions, then the level-one matrix $\mathbf{Z}$ has little additional information beyond any one column, and stacking can offer little improvement over selecting a single model. If, however, different sparse models capture different components of the signal, the meta-learner can combine them to reduce prediction error.

This can be stated in terms of an oracle stacked predictor. Let

$$R(\alpha) = E\left[\left(Y - \sum_{k=1}^{K} \alpha_k \widehat{f}_k(\mathbf{X})\right)^2\right]$$

denote the prediction risk for fixed fitted base learners. If the feasible set $\mathcal{A}$ contains the unit vectors, then the best population-level stacked combination satisfies

$$\min_{\alpha \in \mathcal{A}} R(\alpha) \leq \min_{1 \leq k \leq K} R(\mathbf{e}_k),$$

where $\mathbf{e}_k$ places all weight on model k. This inequality is an oracle comparison: it says that the best possible combination over the chosen weight class is at least as good as the best individual base model within that same population criterion. The estimated stacked predictor may not achieve this inequality exactly in finite samples, which is why cross-validation, constraints, and external test evaluation are important.

From the sparse-model-set point of view, stacking has two roles. First, it provides a predictive use of the whole library rather than forcing a final model-selection step. Second, the estimated weights can help summarize which sparse representations contribute most to prediction after accounting for the other models. These weights should not be interpreted as posterior probabilities of the models. They are prediction weights learned from out-of-sample performance. Nevertheless, when the library contains several models with overlapping but non-identical supports, the stacking weights provide a useful description of how prediction is distributed across the recovered sparse model set.

4.2.3 Sparse Consensus Prediction

We refer to $\widehat{f}_{\text{stack}}$ as a sparse consensus predictor because it combines predictions from many sparse models while keeping each base representation interpretable. The final predictor itself may involve many variables indirectly, since different base models can use different

supports. However, the prediction remains decomposable into sparse components:

$$\widehat{f}_{\text{stack}}(\mathbf{x}) = \sum_{k=1}^{K} \widehat{\alpha}_k \left(\hat{\beta}_{0k} + \mathbf{x}_{S_k}^{\top} \hat{\beta}_{S_k} \right).$$

Thus, the prediction can be studied at two levels. At the model level, one can examine which sparse models receive large weights. At the variable level, one can combine the stacking weights with the inclusion frequencies or coefficient summaries from the previous chapter to identify variables that are frequently selected, strongly weighted, or both. This connection is important because it links prediction to the interpretability of the SWAG network and to the variable-importance summaries developed earlier.

As the signal-to-noise ratio increases, one expects accurate models that contain the active variables, or good substitutes for them, to receive larger weights. This should be understood as a qualitative prediction mechanism rather than a formal support-recovery theorem. Establishing consistency of the stacking weights or support recovery would require additional assumptions about the design matrix, the SWAG search procedure, the base learners, and the growth of K with n. The purpose of this chapter is more modest: to show that the recovered sparse model set can be used as a practical prediction space and that stacking can improve prediction relative to using individual sparse models alone.

4.2.4 Prediction Disagreement and Uncertainty

The sparse-model-set view also clarifies a source of predictive uncertainty that is usually hidden when one final model is selected. A conventional prediction interval is often constructed conditionally on a fitted model or a fitted architecture. When many sparse models are competitive, however, there is additional uncertainty associated with representation choice: two models may have similar validation performance but produce different predictions for a particular observation [13, 15, 21].

One descriptive measure of local representation uncertainty is the weighted disagreement

$$U(\mathbf{x}) = \sum_{k=1}^{K} \widehat{\alpha}_k \left\{ \widehat{f}_k(\mathbf{x}) - \widehat{f}_{\text{stack}}(\mathbf{x}) \right\}^2 .$$

Large values of $U(\mathbf{x})$ indicate that the competitive sparse models disagree around $\mathbf{x}$, even after the models are combined by stacking. This quantity should not be interpreted as a calibrated prediction interval. Rather, it is a diagnostic for local instability of the predictive representation. It can be reported together with the stacked prediction to show where the prediction is driven by broad agreement across the library and where it depends on combining conflicting sparse explanations.

This diagnostic also suggests future calibration strategies. For example, held-out or cross-fitted conformal procedures could build conformity scores from the stacked residuals together with disagreement terms such as $U(\mathbf{x})$, so that prediction intervals reflect both residual error and representation uncertainty [45, 25, 4, 6]. Such calibration is outside the simulation study below, but it is a natural extension of treating the recovered sparse model set as a representation space.

4.3 Simulation Study

We conduct a simulation study using a linear regression model with a known sparse signal. The objective is to evaluate whether stacking over the SWAG library can provide competitive test-set prediction relative to standard linear benchmarks and to individual SWAG models. This setting is intentionally favorable to classical sparse linear methods, since the data-generating model is exactly linear, the signal is sparse, and the candidate benchmark methods include lasso and ridge regression.

For each simulation replicate, we generate a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with independent standard normal entries,

$$X_{ij} \sim N(0, 1).$$

The response is generated according to

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

We set $n = 200$ and $p = 50$. The signal is sparse: the first five coefficients are set equal to 2, so that $\beta_1 = \cdots = \beta_5 = 2$, and the remaining $p - 5$ coefficients are set equal to zero. The target signal-to-noise ratio is defined as

$$\text{SNR} = \frac{\text{Var}(\mathbf{X}\beta)}{\sigma^2}.$$

We consider

$$\text{SNR} \in \{0.1, 1, 10, 100, 1000\}.$$

For each value of SNR, the noise variance σ^2 is chosen so that the desired signal-to-noise ratio is achieved.

For each replicate, the data are split into training, validation, and test sets, containing 60%, 20%, and 20% of the observations, respectively. SWAG is run on the training set to construct the base model library. The stacked predictors are trained using the validation-set predictions from the SWAG models and are then evaluated on the held-out test set. We use $B = 500$ Monte Carlo replications. Predictive performance is measured by the test-set coefficient of determination, R^2.

Table 4.1: Average test-set R^2 over 500 simulation replications for different signal-to-noise ratios. The SWAG library is stacked using ridge, lasso, XGBoost, and linear meta-models, and is compared with standard lasso and ridge regression as well as individual SWAG models.

SNR	Ridge stack	Lasso stack	XGB stack	Linear stack	Lasso	Ridge	Best SWAG	Mean SWAG
0.1	-0.02	-0.03	-0.03	-0.09	-0.02	-0.02	0.08	-0.04
1	0.36	0.34	0.31	0.31	0.42	0.33	0.41	0.23
10	0.71	0.71	0.67	0.69	0.89	0.87	0.73	0.48
100	0.78	0.78	0.75	0.77	0.99	0.98	0.80	0.53
1000	0.80	0.79	0.76	0.78	0.99	1.00	0.81	0.53

Table 4.1 summarizes the average test-set performance of the different procedures. As

expected, the standard lasso and ridge regressions perform extremely well in this simulation design, especially at moderate and high signal-to-noise ratios. This is not surprising, since the data are generated from a sparse linear model with independent Gaussian predictors, precisely the type of setting in which these methods are expected to be highly competitive. At SNR $= 10$, for example, lasso and ridge achieve average test-set R^2 values of 0.89 and 0.87, respectively, while at SNR $= 100$ and 1000 they approach nearly perfect prediction.

The SWAG stacking procedures do not uniformly outperform these standard linear benchmarks in this setting. Nevertheless, they remain competitive, particularly given the difficulty of the stacking problem induced by the full SWAG library. With $n = 200$ and a 60-20-20 split, the meta-model is trained on only about 40 validation observations, while the SWAG library contains roughly 2000 candidate models. Thus, the stacking stage is effectively asked to learn from a highly overparameterized validation design, with approximately 40 observations and 2000 model-prediction features. Despite this, the ridge and lasso stacking procedures achieve average R^2 values of 0.71 at SNR $= 10$, 0.78 at SNR $= 100$, and about 0.79–0.80 at SNR $= 1000$. These values are below the oracle-favorable lasso and ridge benchmarks, but they are close to or better than the best individual SWAG model across several signal regimes.

The comparison with the individual SWAG models is especially informative. The mean SWAG model performs substantially worse than the stacked versions, reaching only 0.48 at SNR $= 10$ and about 0.53 at the two largest signal-to-noise ratios. The best individual SWAG model performs better, with average R^2 values of 0.73, 0.80, and 0.81 for SNR $= 10, 100, 1000$, respectively. However, the ridge and lasso stacked SWAG predictors are broadly comparable to the best single SWAG model, and in some cases improve upon it. This suggests that the SWAG library contains useful predictive information spread across multiple sparse models, even when no single model fully matches the performance of the standard lasso or ridge estimator.

These results should therefore be interpreted as a conservative assessment of SWAG stacking. The simulation favors standard linear estimators, while the stacking procedure is deliberately applied to a large, unfiltered SWAG library. In practice, stronger performance may

be obtained by post-processing the SWAG library before stacking, for example by selecting a smaller set of high-performing and less-correlated models. Such de-correlation or diversity-based filtering would reduce the dimension of the meta-learning problem and may allow the stacking model to combine complementary sparse representations more effectively. Importantly, even in this unfavorable setting, SWAG stacking preserves stable predictive performance while retaining the interpretability advantages of the SWAG framework: prediction is still built from an explicit library of sparse models, allowing the analyst to study variable inclusion, model diversity, and alternative sparse explanations rather than relying only on a single selected model.

4.3.1 Benchmark Evidence on Public Data

As a complementary empirical check, stacked prediction over recovered sparse model sets was also evaluated on several public regression and classification datasets. The goal of these experiments was not to exhaustively tune every benchmark, but to assess whether SWAG library representations could improve over regularized generalized linear model baselines while remaining competitive with standard machine-learning predictors. For each dataset, the recovered sparse models were used as base learners, their predictions were passed to a stacking procedure, and performance was evaluated on held-out test data.

Table 4.2: Held-out performance of Stacked-SWAG and benchmark methods on public datasets. For regression datasets the metric is R^2; for classification datasets the metric is AUC. Larger values are better.

Dataset	Metric	Stacked-SWAG	Stacked-Lasso	Lasso	Ridge	RF	XGBoost	NN
Diabetes	R^2	0.538	0.351	0.474	0.482	0.497	0.425	0.532
California Housing	R^2	0.771	0.568	0.425	0.125	0.809	0.811	0.802
Friedman #1	R^2	0.925	0.002	0.893	0.891	0.824	0.916	0.883
Bike Sharing	R^2	0.822	0.366	0.665	0.665	0.939	0.930	0.944
Breast Cancer	AUC	0.989	0.973	1.000	0.990	0.995	0.999	0.988
Adult Income	AUC	0.914	0.891	0.910	0.910	0.916	0.923	0.887
Bank Marketing	AUC	0.915	0.853	0.721	0.909	0.927	0.929	0.877
Credit Default	AUC	0.776	0.748	0.731	0.756	0.780	0.784	0.723

Table 4.2 shows the same qualitative pattern as the controlled simulation. Stacked-SWAG improves over Lasso, Ridge, and Stacked-Lasso on most datasets, while often remaining close to the strongest black-box benchmarks. It is the best-performing method on Diabetes and Friedman #1, and it is competitive on Adult Income, Bank Marketing, and Credit Default. On Bike Sharing, random forests, XGBoost, and neural networks are substantially stronger, indicating that the sparse linear representations used here do not capture all nonlinear structure in that dataset.

These benchmark results should be interpreted as evidence for the usefulness of the representation rather than as a claim of universal dominance. The important point is that stacking over compact sparse models can recover a large part of the predictive performance of more flexible procedures while preserving a representation-level interpretation. Each coordinate of the stacked predictor corresponds to a known sparse model, so the ensemble can be inspected through model weights, variable supports, inclusion frequencies, co-occurrence patterns, and disagreement diagnostics.

4.4 Conclusion

This chapter showed how the collection of sparse models produced by SWAG can be used for prediction. Instead of selecting one final model from the library, we treated the predictions of the SWAG models as a second-level representation and combined them using stacked generalization. This reframes SWAG as more than a variable-selection procedure: it becomes a mechanism for constructing a sparse prediction space whose coordinates remain interpretable.

The controlled simulation showed that the stacked predictor consistently achieved higher test-set accuracy than both the average individual SWAG model and the best individual SWAG model across all signal-to-noise ratios considered. The improvement was especially clear in the moderate-signal settings, where no single sparse model fully captured the signal but several sparse models contributed useful information. In high-signal settings, the best individual model already performed well, but stacking still achieved the highest average R^2. The public-data benchmarks gave a complementary picture: Stacked-RVS often improved over regularized gener-

alized linear model baselines and, in several cases, approached the performance of more flexible machine-learning methods.

These results complement the inferential perspective developed in the previous chapter. The same sparse model set that supports variable-frequency summaries, network representations, permutation-based inference, and conditional evidence summaries can also be used to build accurate predictions. Thus, interpretability and prediction do not need to be treated as separate goals. A sparse model set can provide interpretable components, while stacking combines those components into a stable predictive rule.

Future work may study regularized or nonlinear meta-learners, extensions to generalized linear models and classification, and uncertainty quantification for predictions obtained over sparse model sets. In particular, local disagreement measures and conformal calibration could be used to determine when predictions are stable across competitive sparse representations and when they depend on unresolved model uncertainty. Such extensions would connect directly to the broader goal of using Rashomon-style model multiplicity not only to interpret data, but also to improve prediction while accounting for model uncertainty.

Chapter 5

Biomedical Applications

5.1 Introduction

The previous chapters developed sparse model sets as inferential and predictive objects. This chapter studies how that framework behaves in two biomedical applications where the number of candidate predictors is large, the scientific interpretation matters, and a single selected model would give an incomplete picture of the evidence. The goal is not only to report predictive accuracy, but also to examine whether SWAG produces compact libraries of models whose variables, co-occurrences, coefficients, and aggregated evidence can be interpreted scientifically. The results of these applications have been published in peer-reviewed journals [32, 36].

The first application concerns classification of Attention Deficit Hyperactivity Disorder (ADHD) using resting-state functional magnetic resonance imaging (Rs-fMRI) data from the ADHD-200 consortium [1]. In this setting, each feature is a functional connectivity measure between two brain regions. A sparse model set is useful because it identifies recurring brain connections and the groups of connections that appear together across competitive classifiers.

The second application concerns DNA methylation signatures associated with sleep timing in children [37]. Here the original feature space contains $865,926$ CpG sites measured using the Illumina Infinium MethylationEPIC BeadChip. SWAG is used as a screening and model-set recovery procedure to identify a much smaller set of loci whose methylation patterns distinguish early and late bedtime groups.

Together, the two applications illustrate complementary roles of sparse model sets. In neuroimaging, the recovered library supports prediction and network interpretation of functional connectivity. In epigenomics, the recovered library reduces a very high-dimensional molecular search space to a set of loci and genes that can be followed by biological annotation.

5.2 Neuroimaging Classification of ADHD

5.2.1 Data and Preprocessing

The ADHD data were obtained from the ADHD-200 consortium [1]. The original diagnostic labels distinguish three ADHD subtypes: ADHD-I, characterized primarily by inattention; ADHD-H, characterized primarily by hyperactivity or impulsivity; and ADHD-C, which combines both symptom profiles. Because the goal of this analysis is to study sparse connectivity patterns associated with ADHD status, the three subtypes are combined into a single ADHD class and compared with healthy controls. This binary formulation does not address subtype-specific mechanisms, but it allows the analysis to focus on whether functional connectivity features contain stable predictive information for ADHD detection.

The dataset contains 930 subjects, including 573 healthy controls and 357 subjects with ADHD. The subjects were scanned at seven acquisition sites, summarized in Table 5.1. Additional acquisition details are available from the ADHD-200 consortium website at `http://fcon_1000.projects.nitrc.org/indi/adhd200/`.

Table 5.1: Site distribution for the ADHD data used in this study.

Imaging site	Control	ADHD	Total
Peking University	143	102	245
Kennedy Krieger Institute	69	25	94
NeuroIMAGE Sample	37	36	73
New York University Child Study Center	110	147	257
Oregon Health & Science University	70	43	113
University of Pittsburgh	94	4	98
Washington University	50	0	50
Total	573	357	930

A standard preprocessing pipeline was applied to the Rs-fMRI data using the Data Processing Assistant for Resting-State fMRI Toolbox (DPARSF) [50]. The preprocessing included removal of the first five volumes, slice-timing correction, motion realignment, co-registration

of anatomical and functional images, nuisance regression, normalization to the MNI template, blind deconvolution to estimate latent neural signals [48], and bandpass filtering between 0.01 and 0.1 Hz. Mean time series were extracted from 190 functionally homogeneous regions defined by the CC200 atlas [14]. Functional connectivity was then computed using Pearson correlations between pairs of regional time series. A liberal preliminary t-test screen ($p < 0.05$, uncorrected) retained 1179 connectivity features for the SWAG analysis.

5.2.2 SWAG Analysis

SWAG was applied to the ADHD classification task using three base learners: logistic regression (LOGIT), linear-kernel support vector machines (SVM-L), and radial-kernel support vector machines (SVM-R). This comparison allows the recovered sparse model sets to be examined across both interpretable and less directly interpretable learning mechanisms. Following the rules of thumb in [29], the maximum model size was set to $p_{\max} = 20$, and the screening proportion was set to $\alpha = 0.05$. The initial screen therefore retained approximately $1179 \times 0.05 \approx 59$ one-feature models. The candidate expansion budget was set to $m = \binom{59}{2} = 1711$, so that the two-dimensional search space induced by the screened features could be explored.

The LOGIT-based SWAG library is the main object of interpretation because it provides both sparse supports and model-specific coefficient estimates. After post-processing, LOGIT-SWAG produced 381 models, with dimensions ranging from 9 to 20, built from 53 distinct connectivity features. The SVM-L and SVM-R libraries produced 114 and 146 models, respectively, and led to the same broad conclusion that frontal and temporal regions are central to classification. Since SVM coefficients are not directly comparable to logistic-regression coefficients in this setting, the detailed interpretation below focuses on the LOGIT library.

Each feature is a Pearson correlation between two brain regions and is denoted by $A \leftrightarrow B$. Table 5.2 reports the most frequent LOGIT-SWAG features, the proportion of models containing each feature, the proportion of positive coefficient estimates among models containing the feature, and the median coefficient. These quantities correspond to the model-set summaries

developed in Chapter 3: inclusion frequency measures recurrence across competitive models, while coefficient sign and magnitude summarize conditional direction within the recovered library.

Table 5.2: Most frequent functional-connectivity features in the LOGIT-SWAG library. The second column is the proportion of the 381 models containing the feature.

Feature	Prop. Models	Prop. Positive β	Median β
Insula-R ↔ Paracentral-Lobule-R (A)	1.000	0.000	-0.533
Frontal-Sup-R ↔ Frontal-Mid-R (B)	1.000	1.000	0.675
Temporal-Mid-L ↔ Lingual-L (C)	1.000	1.000	0.508
Insula-L ↔ Frontal-Mid-L (D)	1.000	0.000	-0.977
Frontal-Sup-R ↔ Frontal-Inf-Orb-R (E)	0.989	0.003	-0.313
Supp-Motor-Area-L ↔ SupraMarginal-L (F)	0.989	1.000	0.357
Frontal-Sup-R ↔ ParahippocampaGyrus (G)	0.953	0.000	-2.081
Vermis-1-2 ↔ Temporal-Mid-L-4 (H)	0.953	1.000	0.781
Vermis-4-5 ↔ Frontal-Sup-Orb-R (I)	0.853	0.000	-1.099
Frontal-Sup-Medial-L ↔ Cingulum-Mid-L (J)	0.745	0.000	-1.007
Temporal-Mid-L ↔ Pallidum-R (K)	0.738	0.979	0.222
Occipital-Mid-L ↔ Cerebelum-Crus2-L (L)	0.700	1.000	2.261
Cingulum-Ant-L ↔ Paracentral-Lobule-R(M)	0.677	0.000	-1.104
Temporal-Pole-Sup-R ↔ Frontal-Mid-L (N)	0.617	0.000	-1.072
Frontal-Mid-L ↔ Caudate-R (O)	0.543	1.000	1.162
Frontal-Mid-L ↔ Frontal-Sup-L (P)	0.528	1.000	0.939

For example, the feature Frontal-Sup-R ↔ Frontal-Inf-Orb-R appears in 98.9% of the LOGIT-SWAG models. Among the models containing it, the coefficient is positive only 0.3% of the time, with median coefficient -0.313. Thus, conditional on the other features in the sparse model set, larger connectivity between these two frontal regions is usually associated with a lower predicted probability of ADHD. This interpretation is more informative than a single coefficient from one selected model because it records how the association behaves across many competitive sparse representations.

Figure 5.1 displays the top 35% most frequent LOGIT-SWAG features, corresponding to 19 of the 53 selected connectivity features. In this display, nodes are brain regions and edges are selected connectivity features. Connections involving the frontal region are particularly frequent,

consistent with the region-level summary in Table 5.3.

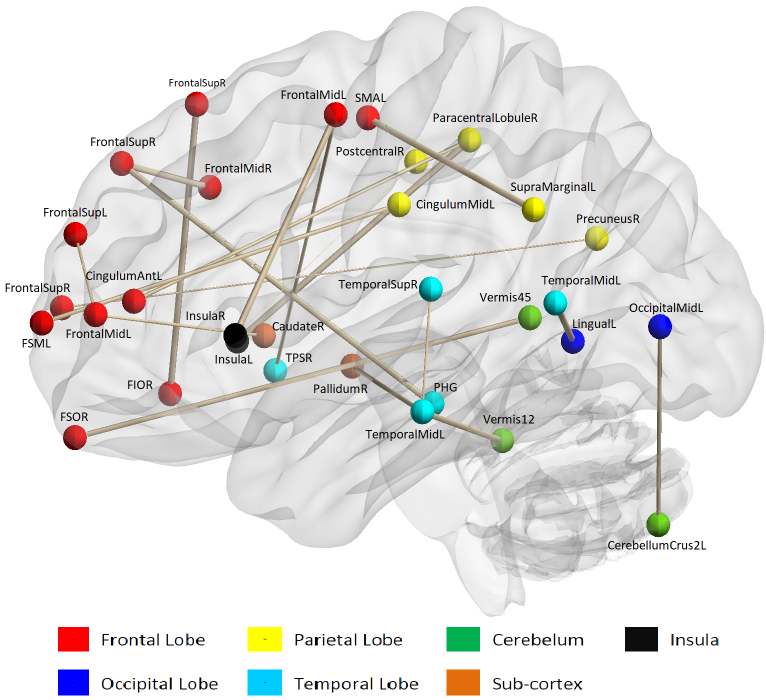

Figure 5.1: Sagittal-view representation of the most frequent LOGIT-SWAG connectivity features. Points represent brain regions and lines represent selected functional-connectivity features between regions.

To summarize selected connectivity features at the level of brain regions, each feature contributes one count to each endpoint region. For example, Frontal-Sup-R ↔ Frontal-Inf-Orb-R contributes two counts to the frontal region, while Insula-L ↔ Frontal-Mid-L contributes one count to the insula region and one to the frontal region. Table 5.3 shows that frontal and temporal regions dominate the selected LOGIT-SWAG library.

5.2.3 Prediction and Network Interpretation

Table 5.4 compares the test accuracy of the base learners fitted without SWAG and the accuracy ranges obtained from the corresponding SWAG libraries. For LOGIT, the non-SWAG benchmark is a Lasso logistic regression fit using [18], which gives a more appropriate sparse baseline than an unregularized high-dimensional logistic model.

Table 5.3: Region-level frequency summary for the LOGIT-SWAG library. Connectivity features are aggregated by their endpoint brain regions.

Brain region	Frequency
Frontal	2903
Temporal	1179
Insula	762
Vermis	688
Paracentral	639
Cingulum	542
Lingual	381

Table 5.4: Test accuracy for the three classifiers with and without SWAG. Since SWAG returns a set of models, the last column reports the range of test accuracies across the recovered library.

Classifier	Without SWAG	With SWAG
Lasso logistic / LOGIT-SWAG	0.544	$(0.532, 0.614)$
Linear SVM / SVM-L-SWAG	0.567	$(0.520, 0.608)$
Radial SVM / SVM-R-SWAG	0.626	$(0.561, 0.632)$

The ADHD-200 benchmark accuracies reported for other classification approaches range from 0.374 to 0.625 [1]. The SWAG libraries therefore achieve accuracy in the same range while using far fewer predictors. Lasso logistic regression selected 123 features, and the SVM classifiers use the full feature set unless combined with an external selection procedure. In contrast, each SWAG library uses fewer than 60 distinct features out of 1179 candidates. Thus, the main gain is not only predictive accuracy, but the recovery of a compact set of competitive models whose structure can be inspected.

Figure 5.2 represents the LOGIT-SWAG library as a feature network. Each node is a selected connectivity feature, and each edge records how often two connectivity features appear together in the same sparse model. Node size is proportional to inclusion frequency. Node color represents the sign of the median logistic-regression coefficient, with positive and negative signs interpreted conditionally on the other features in the models containing that node. Edge thickness represents co-occurrence frequency, while edge color represents the Spearman correlation between the two connectivity features. This network view adds information that is not available from marginal feature ranking alone: it shows which connectivity features form recurring groups of predictive explanations.

5.2.4 Aggregated P-Values for ADHD Connectivity Features

The LOGIT-SWAG library also allows the variable-level procedure from Chapter 3 to be applied directly. For each connectivity feature, model-specific p-values are aggregated across the SWAG models that contain that feature using the Mudholkar-George logit aggregation rule [30]. This step asks a different question from inclusion frequency. A feature can appear often because it is a useful substitute or appears in many competitive contexts, while aggregated p-values measure whether its coefficient receives repeated model-specific evidence across those contexts.

The LOGIT-SWAG library contains 53 distinct connectivity features. At significance level 0.05, the available George-aggregated results identify the features listed in Table 5.5. Several of

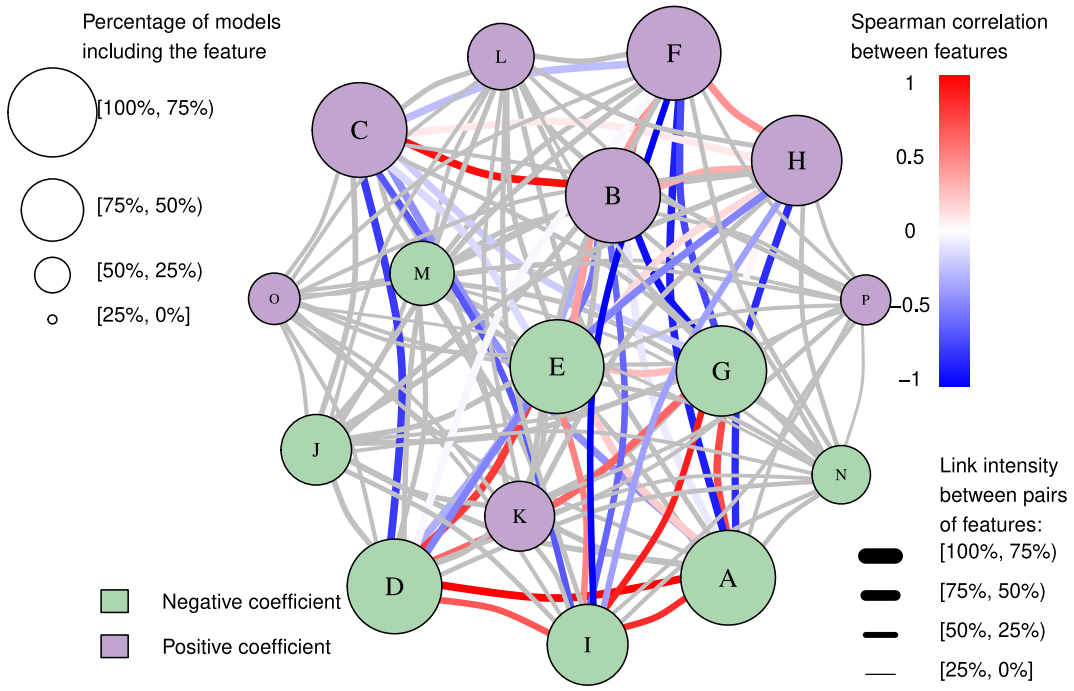

Figure 5.2: The corresponding feature names can be found in Table 5.2. The color of the node illustrates the sign of the median of the estimated β coefficients (i.e. the median of all coefficients that a feature takes in each of the SWAG models). The size of each node is proportional to the percentage of models that contain that feature among all SWAG models. The thickness of each link between different nodes is proportional to the percentage of times two features are present together among all SWAG models. The color of the link shows the value of the Spearman correlation coefficient between two different features (blue for positive correlation and red for negative). The grey links represent the connection between features that appear together with less frequency than the ones with red/blue links between them.

these features involve frontal, temporal, cerebellar, cingulate, or insular regions, consistent with the network and region-frequency summaries above.

Table 5.5: Connectivity features with recorded George-aggregated p-values below 0.05 in the LOGIT-SWAG library.

Connectivity feature	Aggregated p-value
Frontal-Sup-R ↔ ParahippocampaGyrus	0.002
Supp-Motor-Area-L ↔ SupraMarginal-L	0.039
Vermis-1-2 ↔ Temporal-Mid-L	0.024
Occipital-Mid-L ↔ Cerebelum-Crus2-L	0.002
Frontal-Mid-L ↔ Cingulum-Ant-L	0.023
Insula-R ↔ Pallidum-R	0.026
Rolandic-Oper-L ↔ Temporal-Inf-L	0.026
Temporal-Inf-L ↔ Lingual-L	0.031

These results should be interpreted as library-level evidence, not as post-selection p-values from a single final model. The variables in Table 5.5 are features whose model-specific evidence persists across multiple competitive logistic models. In this sense, the ADHD analysis illustrates the distinction developed in Chapter 3: frequency, co-occurrence, coefficient stability, and aggregated evidence are complementary summaries of the same recovered sparse model set.

5.3 DNA Methylation and Sleep Timing in Children

5.3.1 Study Population

Participants were grouped by bedtime habits into an early-bedtime group, defined as bedtime before 8:30 PM, and a late-bedtime group, defined as bedtime after 8:31 PM. Table 5.6 summarizes demographic and anthropometric characteristics. The mean age was 8.55 years. BMI z-score, waist-circumference z-score, and waist-to-height-ratio z-score were higher on average in the late-bedtime group, but these differences were not statistically significant. The study design and broader analytical workflow are described in [37].

Table 5.6: Characteristics of the sleep-timing study population.

Parameter	Total	Early bedtime	Late bedtime	p-value
		Before 8:30 PM	**After 8:31 PM**	
Total participants	31	15	16	–
Sex (male/female)	17/14	9/6	8/8	–
Age (years)	8.55 ± 0.24	8.22 ± 0.39	8.86 ± 0.29	0.201
Weight (kg)	36.05 ± 2.27	33.58 ± 3.58	38.36 ± 2.84	0.305
Height (cm)	134.35 ± 2.23	132.24 ± 3.36	136.33 ± 2.97	0.369
BMI (kg/m^2)	19.42 ± 0.64	18.51 ± 0.98	20.28 ± 0.81	0.174
BMI z-score	1.27 ± 0.22	0.91 ± 0.36	1.62 ± 0.23	0.111
WC z-score	0.91 ± 0.11	0.84 ± 0.18	0.97 ± 0.14	0.556
WHtR z-score	0.69 ± 0.12	0.63 ± 0.18	0.75 ± 0.17	0.637

Values are presented as mean ± standard deviation. BMI = body mass index; WC = waist circumference; WHtR = waist-to-height ratio.

5.3.2 SWAG Analysis and Genomic Annotation

To identify methylation markers associated with sleep timing, SWAG was applied to $865,926$ CpG sites profiled using the Illumina Infinium MethylationEPIC BeadChip. The analysis selected 1006 target IDs, corresponding to 840 unique loci after duplicate target IDs were removed. Among the unique loci, 352 were hypermethylated and 488 were hypomethylated in children with early bedtime compared with late bedtime ($p < 0.001$). These loci mapped to 571 unique genes.

The significant CpG sites were annotated using the array manifest to genomic features, including promoter regions, gene bodies, untranslated regions, CpG islands, shores, shelves, and other regulatory elements [22]. Of the 610 annotated significant target ID sites, 403 (66%) were located within CpG islands, which are genomic regions often enriched near transcription start sites and associated with transcriptional regulation. Among island-associated CpGs, 149 (37%) were within 0–200 bp of the transcription start site (TSS200), 53 (13.2%) were within 200–1500 bp upstream of the transcription start site (TSS1500), 79 (19.6%) were in the 5′UTR, 75 (18.6%) were in gene bodies, 3 (0.7%) were in the 3′UTR, 31 (7.7%) were intragenic, and 13 (3.2%) were categorized as unmapped. CpG island context analysis further indicated that 40% of annotated

sites were located within CpG islands, with additional sites mapping to north shores (8.8%), south shores (6.9%), north shelves (2.3%), and south shelves (2.6%). The remaining 39.4% could not be assigned to a defined CpG island context.

Figure 5.3 summarizes the genomic distribution of the significant CpG target IDs. Several highly ranked target IDs, including cg26811976 (*PRAMEF4*), cg07891983 (*CREM*), cg04402799 (*CDH4*), and cg00136968 (*SDK1*), were categorized as unmapped by the current reference annotation. Their recurrence in the SWAG results despite incomplete annotation suggests that unmapped loci should not be automatically discarded during high-dimensional screening.

Table 5.7: Top 10 CpG target IDs associated with sleep timing from the SWAG analysis.

No.	Target ID	Gene	CHR	Location
1	cg09760986	*ABCG2*	4	S_Shore
2	cg22792063	*ABHD4*	14	Island
3	cg00807892	*MOBKL1A*	4	Island
4	cg25282780	*AK3*	9	Island
5	cg09321097	*SDE2*	1	Island
6	cg26811976	*PRAMEF4*	1	∼
7	cg07891983	*CREM*	10	∼
8	cg04402799	*CDH4*	20	∼
9	cg03232960	*BRAT1*	7	Island
10	cg00136968	*SDK1*	7	∼

CHR = chromosome; S_Shore = south shore; Island = CpG island; ∼ denotes loci with unmapped or unspecified genomic positions in the Infinium MethylationEPIC manifest.

Table 5.7 lists the top 10 target IDs from the SWAG analysis. These targets were selected because they combine statistical evidence, recurrence across the recovered sparse model set, and consistent methylation differences between early and late bedtime groups.

5.3.3 Methylation Patterns and Biological Follow-Up

Figure 5.4 displays the selection frequencies of the top target IDs. Figure 5.5 then compares methylation patterns between early and late bedtime groups. The heatmap displays row-wise standardized methylation values, while the boxplots display absolute β values. This

distinction is important: standardized values are useful for visualizing participant-level clustering, whereas absolute β values show the direction and magnitude of group differences.

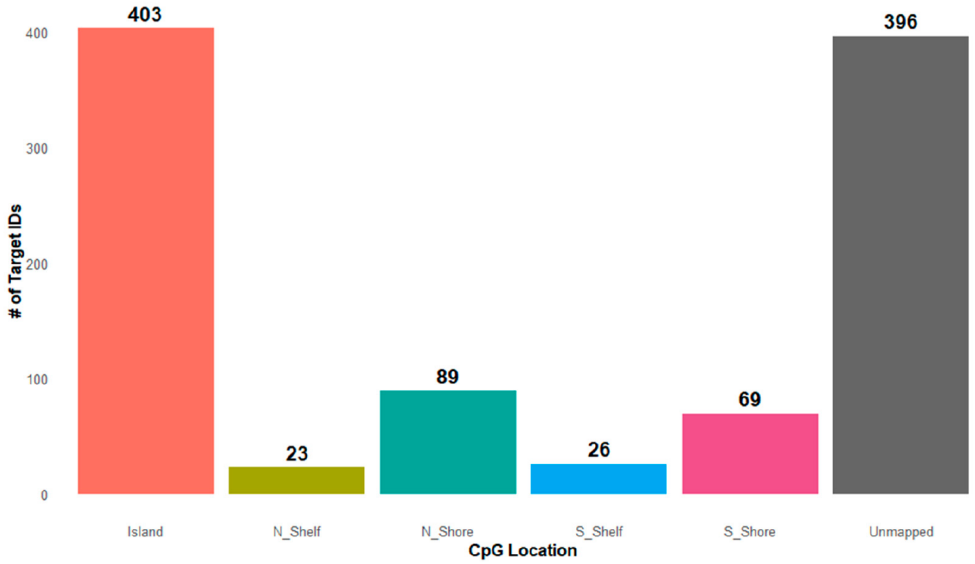

Figure 5.3: Genomic distribution of significant CpG target IDs identified by SWAG. The plot summarizes counts across genomic contexts such as CpG islands, shores, shelves, open sea, and unmapped loci.

Among the top targets, *ABCG2* (cg09760986), *ABHD4* (cg22792063), and *MOBKL1A* (cg00807892) exhibited the largest absolute methylation differences, with cg09760986 showing $\Delta\beta = 0.115$. The top targets also included *AK3*, *SDE2*, *PRAMEF4*, *CREM*, *CDH4*, *BRAT1*, and *SDK1*. As a descriptive follow-up, independent t-tests gave small p-values across the top sites ($p < 0.001$). The absolute β value distributions show that *ABCG2* and *PRAMEF4* were hypermethylated in children with late bedtime, whereas *ABHD4*, *MOBKL1A*, *AK3*, *SDE2*, *CREM*, *CDH4*, *BRAT1*, and *SDK1* were hypomethylated in children with late bedtime compared with early bedtime.

Building on prior research identifying obesity-associated DNA methylation patterns in children, the SWAG-derived gene list was compared with external gene-disease association sets related to sleep and obesity [33]. The comparison used DisGeNET (`https://www.disgenet.com/`, accessed on 27 October 2025), a curated platform for gene-disease associations, with one gene set associated with sleep regulation ($n = 1639$) and another associated with obesity ($n = 2822$). The SWAG-derived genes overlapped with 72 sleep-related genes and 81 obesity-

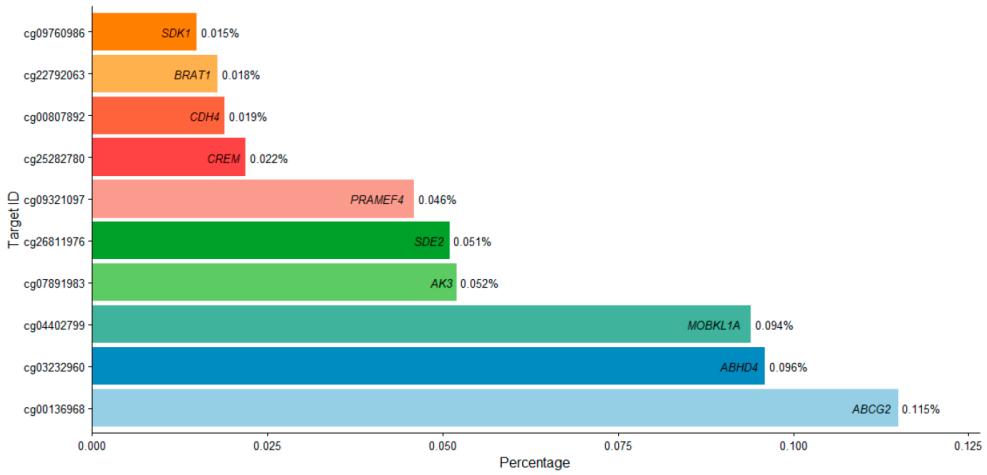

Figure 5.4: Selection frequencies for the top 10 CpG target IDs in the sleep-timing SWAG analysis. Higher percentages indicate greater recurrence across competitive sparse models.

related genes. This comparison does not validate the selected loci by itself, but it places the sparse-model-set results in a broader biological context. The analytical workflow, from SWAG filtering to gene-set intersection, is summarized in Figure 5.6.

Focusing on the intersection of the sleep and obesity gene sets, 10 genes were both differentially methylated between early and late bedtime groups and represented in both external categories: *CDH4*, *NR3C2*, *ACTG1*, *COG5*, *CAT*, *HDAC4*, *FTO*, *DOK7*, *OCLN*, and *ATXN1*. Notably, *CDH4* appeared both among the highly ranked SWAG target IDs and in this intersection set. These genes remained significant after applying the Benjamini-Yekutieli procedure for multiple testing correction [8] (FDR-adjusted $p < 0.05$).

5.4 Conclusion

This chapter applied SWAG to two high-dimensional biomedical problems: neuroimaging classification for ADHD and DNA methylation profiling for childhood sleep timing. The two examples serve different purposes. The ADHD analysis emphasizes prediction, model-set interpretation, and feature co-occurrence among functional-connectivity variables. The sleep analysis emphasizes high-dimensional screening, annotation, and biological follow-up for CpG sites. Together, they show how the sparse model set framework developed in the previous

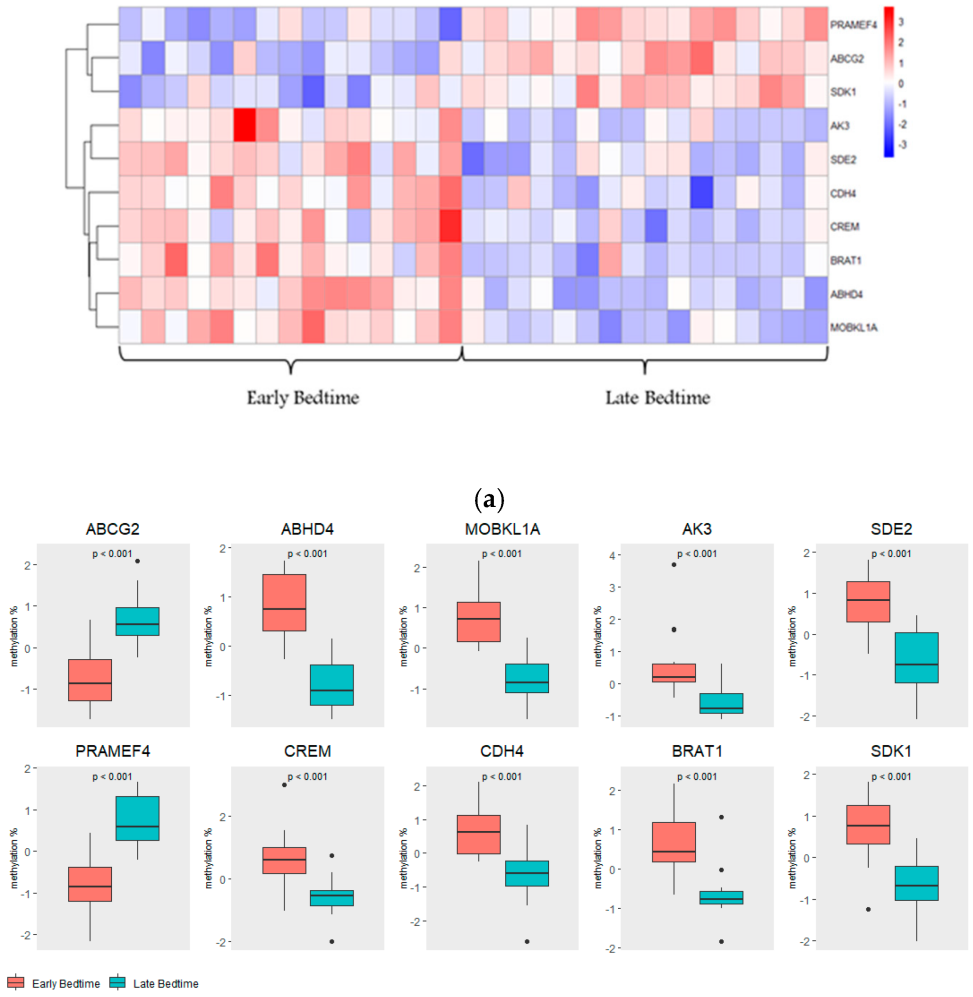

Figure 5.5: DNA methylation differences between early and late bedtime groups. Panel (a) shows row-wise standardized methylation values for the top 10 genes, with values on the scale -3 to $+3$. Panel (b) shows absolute methylation values for early and late bedtime groups, with p-values from independent t-tests displayed above each plot.

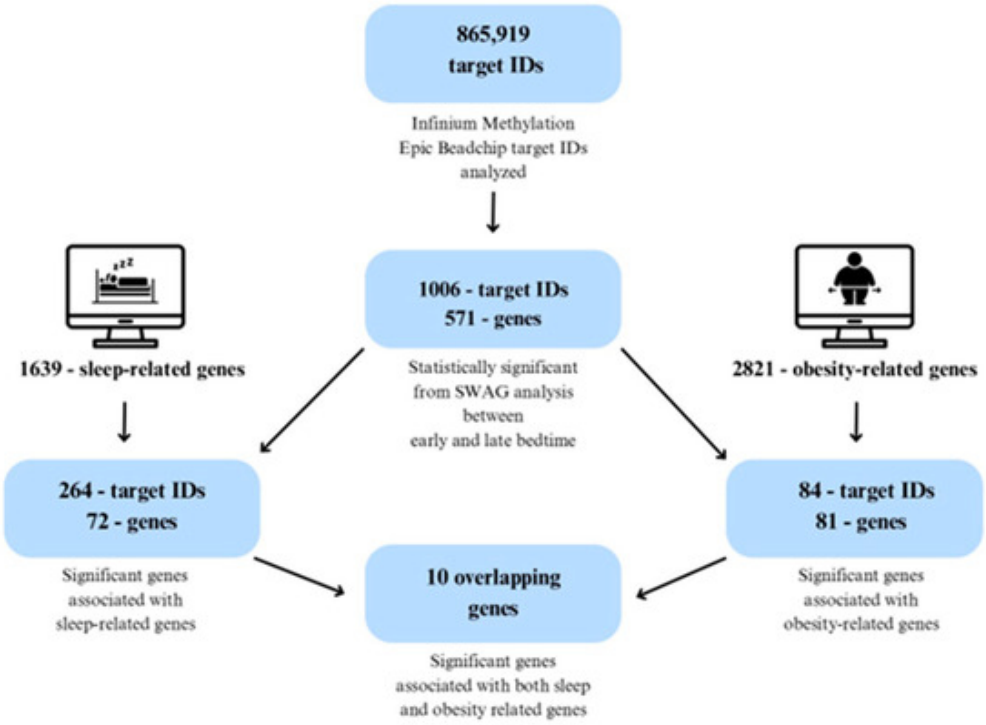

Figure 5.6: Workflow for cross-referencing SWAG-derived sleep-timing genes with sleep- and obesity-related gene sets.

chapters can support both predictive and inferential summaries.

In the ADHD study, SWAG reduced 1179 functional-connectivity features to sparse logistic and support-vector-machine libraries containing fewer than 60 distinct features. The resulting prediction accuracies were comparable to benchmark high-dimensional methods, while the recovered LOGIT-SWAG library retained coefficient signs, inclusion frequencies, feature co-occurrence, and aggregated p-values. These summaries provide a richer description than a single selected model because they identify recurring connectivity patterns rather than treating one sparse model as definitive.

In the sleep-timing study, SWAG screened 865,926 CpG sites and identified 1006 target IDs mapping to 571 genes. The top targets, including *ABCG2*, *ABHD4*, *MOBKL1A*, and *CDH4*, showed clear methylation differences between early and late bedtime groups. Cross-referencing the SWAG-derived gene list with DisGeNET gene sets connected the selected loci to sleep- and obesity-related biological contexts, while also illustrating the need to interpret database overlap as supporting evidence rather than as independent confirmation.

Across both applications, the main contribution is not simply dimensionality reduction. SWAG produces a structured collection of competitive sparse models, and that collection can be summarized through prediction, stability, co-selection, direction of association, and aggregated evidence. This perspective is consistent with the central theme of the dissertation: in high-dimensional biomedical problems, scientific interpretation can be more stable when it is based on the behavior of a model set rather than on a single selected model.

References

[1] C. ADHD-200. “The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience.” In: *Frontiers in systems neuroscience* 6 (2012), p. 62.

[2] A. Agarwal, F. Xiao, R. Barter, O. Ronen, B. Fan, and B. Yu. *PCS-UQ: Uncertainty Quantification via the Predictability-Computability-Stability Framework*. 2025. arXiv: 2505.08784 [stat.ML].

[3] H. Akaike. “A New Look at the Statistical Model Identification.” In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723.

[4] A. N. Angelopoulos and S. Bates. “A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification.” In: *arXiv preprint arXiv:2107.07511* (2021).

[5] I. Azizi, J. Bodik, J. Heiss, and B. Yu. *CLEAR: Calibrated Learning for Epistemic and Aleatoric Risk*. 2025. arXiv: 2507.08150 [stat.ML].

[6] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. “Predictive Inference with the Jackknife+.” In: *The Annals of Statistics* 49.1 (2021), pp. 486–507.

[7] M. Behr, K. Kumbier, A. Cordova-Palomera, M. Aguirre, O. Ronen, C. Ye, E. Ashley, A. J. Butte, R. Arnaout, B. Brown, J. Priest, and B. Yu. “Learning Epistatic Polygenic Phenotypes with Boolean Interactions.” In: *PLOS ONE* 19.4 (2024), e0298906.

[8] Y. Benjamini and D. Yekutieli. “The control of the false discovery rate in multiple testing under dependency.” In: *Annals of statistics* (2001), pp. 1165–1188.

[9] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. “Valid Post-Selection Inference.” In: *The Annals of Statistics* 41.2 (2013), pp. 802–837.

[10] L. Breiman. “Stacked regressions.” In: *Machine learning* 24.1 (1996), pp. 49–64.

[11] L. Breiman. “Statistical modeling: The two cultures (with comments and a rejoinder by the author).” In: *Statistical science* 16.3 (2001), pp. 199–231.

[12] K. P. Burnham and D. R. Anderson. “Multimodel inference: understanding AIC and BIC in model selection.” In: *Sociological methods & research* 33.2 (2004), pp. 261–304.

[13] C. Chatfield. “Model Uncertainty, Data Mining and Statistical Inference.” In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 158.3 (1995), pp. 419–466.

[14] R. C. Craddock, G. A. James, P. E. Holtzheimer III, X. P. Hu, and H. S. Mayberg. “A whole brain fMRI atlas generated via spatially constrained spectral clustering.” In: *Human brain mapping* 33.8 (2012), pp. 1914–1928.

[15] D. Draper. “Assessment and Propagation of Model Uncertainty.” In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 45–97.

[16] A. Fisher, C. Rudin, and F. Dominici. “All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously.” In: *Journal of Machine Learning Research* 20.177 (2019), pp. 1–81.

[17] R. A. Fisher. “Statistical methods for research workers.” In: (1934).

[18] J. Friedman, T. Hastie, and R. Tibshirani. “Regularization paths for generalized linear models via coordinate descent.” In: *Journal of statistical software* 33.1 (2010), p. 1.

[19] A. Gelman and E. Loken. “The Statistical Crisis in Science.” In: *American Scientist* 102.6 (2014), pp. 460–465.

[20] P. Good. *Permutation, parametric and bootstrap tests of hypotheses*. Springer, 2005.

[21] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. “Bayesian Model Averaging: A Tutorial.” In: *Statistical Science* 14.4 (1999), pp. 382–417.

[22] H. S. Jang, W. J. Shin, J. E. Lee, and J. T. Do. “CpG and non-CpG methylation in epigenetic gene regulation and brain function.” In: *Genes* 8.6 (2017), p. 148.

[23] M. J. van der Laan, E. C. Polley, and A. E. Hubbard. “Super Learner.” In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (2007), Article 25.

[24] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer, 2005.

[25] J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. “Distribution-Free Predictive Inference for Regression.” In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1094–1111.

[26] J. Liu, C. Zhong, B. Li, M. Seltzer, and C. Rudin. “FasterRisk: fast and accurate interpretable risk scores.” In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17760–17773.

[27] N. Meinshausen and P. Bühlmann. “Stability selection.” In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.4 (2010), pp. 417–473.

[28] N. Meinshausen and B. Yu. “Lasso-type recovery of sparse representations for high-dimensional data.” In: (2009).

[29] R. Molinari, G. Bakalli, S. Guerrier, C. Miglioli, S. Orso, M. Karemera, and O. Scaillet. “Swag: A Wrapper Method for Sparse Learning.” In: *arXiv preprint arXiv:2006.12837* (2020).

[30] G. S. Mudholkar and E. George. “The logit statistic for combining probabilities-an overview.” In: (1977).

[31] G. S. Mudholkar and E. George. “The logit statistic for combining probabilities-an overview.” In: (1977).

[32] Y. Y. Ozdemir, N. C. P. Nukala, R. Molinari, and G. Deshpande. “A Multi-Model Framework to Explore ADHD Diagnosis from Neuroimaging Data.” In: *Journal of Data Science* 22.2 (2024).

[33] P. Patel, V. Selvaraju, J. R. Babu, X. Wang, and T. Geetha. “Novel differentially methylated regions identified by genome-wide DNA methylation analyses contribute to racial disparities in childhood obesity.” In: *Genes* 14.5 (2023), p. 1098.

[34] B. Phipson and G. K. Smyth. “Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn.” In: *Statistical applications in genetics and molecular biology* 9.1 (2010).

[35] C. Qinyu Zhu, M. Tian, L. Semenova, J. Liu, J. Xu, J. Scarpa, and C. Rudin. “Fast and Interpretable Mortality Risk Scores for Critical Care Patients.” In: *arXiv e-prints* (2023), arXiv–2311.

[36] E. Richter, P. Patel, J. R. Babu, X. Wang, and T. Geetha. “The importance of sleep in overcoming childhood obesity and reshaping epigenetics.” In: *Biomedicines* 12.6 (2024), p. 1334.

[37] E. Richter, P. Patel, Y. Y. Ozdemir, U. V. Nnyaba, R. Molinari, J. R. Babu, and T. Geetha. “Epigenetic Impact of Sleep Timing in Children: Novel DNA Methylation Signatures via SWAG Analysis.” In: *International Journal of Molecular Sciences* 26.21 (2025), p. 10615.

[38] C. Rudin, C. Zhong, L. Semenova, M. Seltzer, R. Parr, J. Liu, S. Katta, J. Donnelly, H. Chen, and Z. Boner. “Amazing Things Come From Having Many Good Models.” In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2024.

[39] L. Semenova, H. Chen, R. Parr, and C. Rudin. “A path to simpler models starts with noise.” In: *Advances in neural information processing systems* 36 (2024).

[40] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. C. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, D. Pope, N. G. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. “Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results.” In: *Advances in Methods and Practices in Psychological Science* 1.3 (2018), pp. 337–356.

[41] S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. “Increasing Transparency Through a Multiverse Analysis.” In: *Perspectives on Psychological Science* 11.5 (2016), pp. 702–712.

[42] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr. “The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.” In: (1949).

[43] J. Taylor and R. J. Tibshirani. “Statistical learning and selective inference.” In: *Proceedings of the National Academy of Sciences* 112.25 (2015), pp. 7629–7634.

[44] L. H. C. Tippett. “The methods of statistics.” In: (1931).

[45] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. New York: Springer, 2005.

[46] A. M. Winkler, G. R. Ridgway, M. A. Webster, S. M. Smith, and T. E. Nichols. “Permutation inference for the general linear model.” In: *Neuroimage* 92 (2014), pp. 381–397.

[47] D. H. Wolpert. “Stacked generalization.” In: *Neural networks* 5.2 (1992), pp. 241–259.

[48] G.-R. Wu, W. Liao, S. Stramaglia, J.-R. Ding, H. Chen, and D. Marinazzo. “A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data.” In: *Medical image analysis* 17.3 (2013), pp. 365–374.

[49] R. Xin, C. Zhong, Z. Chen, T. Takagi, M. Seltzer, and C. Rudin. “Exploring the whole rashomon set of sparse decision trees.” In: *Advances in neural information processing systems* 35 (2022), pp. 14071–14084.

[50] C. Yan and Y. Zang. “DPARSF: a MATLAB toolbox for” pipeline” data analysis of resting-state fMRI.” In: *Frontiers in systems neuroscience* 4 (2010), p. 1377.

[51] B. Yu. “Veridical data science.” In: *Proceedings of the 13th international conference on web search and data mining*. 2020, pp. 4–5.

[52] B. Yu and R. L. Barter. “Producing the Final Prediction Results.” In: *Veridical Data Science: The Practice of Responsible Data Analysis and Decision Making*. The MIT Press, 2024. Chap. 13. URL: https://vdsbook.com/13-final_ml.

[53] B. Yu and K. Kumbier. “Veridical data science.” In: *Proceedings of the National Academy of Sciences* 117.8 (2020), pp. 3920–3929. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1901326117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1901326117.

[54] C. Zhong, Z. Chen, J. Liu, M. Seltzer, and C. Rudin. “Exploring and interacting with the set of good sparse generalized additive models.” In: *Advances in neural information processing systems* 36 (2023), pp. 56673–56699.