ENHANCING MULTI-MODEL INFERENCE WITH NATURAL SELECTION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Ching-Wei Cheng

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Guang Cheng, Chair

Department of Statistics

Dr. Hao Zhang

Department of Statistics

Dr. Michael Levine

Department of Statistics

Dr. Lingsong Zhang

Department of Statistics

Approved by:

Dr. Jun Xie

Graduate Chair of the Department of Statistics

The love of my family made everything possible.

ACKNOWLEDGMENTS

First and the foremost, I would like to express my sincere and deepest gratitude to my advisor, Professor Guang Cheng for guiding me to fulfill my doctorate training. Out of stepping outside of my comfort zone, I wanted to learn from a brilliant researcher who is very different from me. Thanks to his patience, motivation, and immense knowledge, I am able to achieve such a transformative milestone and complete my PhD degree. By taking him as a model, I will keep advancing and pushing myself towards a successful, rigorous and independent researcher.

I would also like to thank my committee members, Professor Hao Zhang, Professor Michael Levine, and Professor Lingsong Zhang, for their continuous support and expert advice. I especially would like to thank Professor Longsong Zhang for having me involved into an applied research project. That was my first experience getting my hands on a real-world big data problem. It does not only make me learn how may we apply machine learning techniques to help getting the world better, but also strengthen my resume. He also provided constructive advice when I was struggling with finding my thesis topic. Professor Michael Levine extended my enthusiasm in time series analysis, and Professor Hao Zhang triggered my interest in spatial statistics. Their support and generosity are crucial to any achievement I was able to make in the Purdue Statistics family.

My sincere appreciation also goes to many others at Purdue. I thank all the members of Professor Guang Cheng's Big Data Theory Group, where I enjoyed academic and leisure discussions with Wei Sun, Zhuqing Yu, Meimei Liu, Botao Hao, Yang Yu, Jincheng Bai, Jexin Duan, Tianyang Hu, Fan Wu, Chi-Hua Wang, Shih-Kang Chao, Qing Yang, and Yao Zheng, among others. My earnest acknowledgment goes to Professor Greg Arling and Professor Zachary Haas at School of Nursing, and Professor Bruce Graig, Ce-Ce Furtner, and other hardworking colleagues in the Statistical Con-

sulting Service, for their support and logistics in consulting service work. To Ji Hwan Oh, Simeng Qu, Yixuan Qiu, among others who joined the department in the same year as I did, I always much appreciate the enjoyable times we hanged out together. In addition, heartfelt appreciation goes to Patti Foster, Mary Sigman and Jesse Wallenfang, among other department secretaries who helped me through all the tedious work in the graduation process and all the tedious administrative business.

Last but not the least, I am grateful to so unprecedentedly many friends for making my life at Purdue so colorful, abundant, and full of joyful memory. Specially thanks go to those volleyball and wallyball fellows for teaming up together and having endured my overly aggressive attitude in the games. I was so lucky enough to have some of you as teammates to conquer so many tournaments and intramural games, including championships in CoRec and Graduate Leagues of the Wallyball Intramural Games in my last year at Purdue. I sincerely hope that the volleyball/wallyball group will live long and prosper, and brings even more enjoyable moments in the future.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AIC	Akaike information criterion
AL	Optimal high-dimensional moving averaging approach of Ando and Li (2014)
BIC	Bayesian information criterion
FDR	False discovery rate
GA	Genetic algorithm
GIC	Generalized information criterion
Lasso	Least absolute shrinkage and selection operator
MCP	Minimax concave penalty
MCS	Model confidence set
PSR	Positive selection rate
RMSE	Root mean squared error
RP	Regularization path
SA	Simulated annealing
SCAD	Smoothly clipped absolute deviation
SMS	Survival model set
SOIL	Sparsity oriented importance learning

ABSTRACT

Cheng, Ching-Wei PhD, Purdue University, December 2019. Enhancing Multi-Model Inference with Natural Selection. Major Professor: Guang Cheng.

Multi-model inference covers a wide range of modern statistical applications such as variable selection, model confidence set, model averaging and variable importance. The performance of multi-model inference depends on the availability of candidate models, whose quality has been rarely studied in literature. In this dissertation, we study genetic algorithm (GA) in order to obtain high-quality candidate models. Inspired by the process of natural selection, GA performs genetic operations such as selection, crossover and mutation iteratively to update a collection of potential solutions (models) until convergence. The convergence properties are studied based on the Markov chain theory and used to design an adaptive termination criterion that vastly reduces the computational cost. In addition, a new schema theory is established to characterize how the current model set is improved through evolutionary process. Extensive numerical experiments are carried out to verify our theory and demonstrate the empirical power of GA, and new findings are obtained for two real data examples.

1. INTRODUCTION

A collection of candidate models serves as a first and important step of multi-model inference, whose spectrum covers variable selection, model confidence set, model averaging and variable importance [1, 2]. The importance of a candidate model set is highlighted in [3]: "all results of the multi-model analyses are conditional on the (candidate) model set." However, in literature, candidate models are either given (e.g., [4, 5]) or generated without any justifications (e.g., [6, 7]). As far as we know, there is no statistical guarantee on the quality of such candidate models, no matter the parameter dimension is fixed or diverges.

In this paper, we study genetic algorithm (GA, [8–10]) in order to search for high-quality candidate models over the whole model space. GA is a class of iterative algorithms inspired by the process of natural selection, and often used for global optimization or search problems; see Figure 1.1. There are two key elements of GA: a genetic representation of the solution domain, i.e., a binary sequence, and a fitness function to evaluate the candidate solutions such as all kinds of information criteria. A GA begins with an initial population of a given size that is improved through iterative application of genetic operations, such as selection, crossover and mutation, until convergence; see Figure 1.2.

Specifically, we employ three basic genetic operations, i.e., selection, crossover and mutation, for the GA. In each generation (the population in each iteration), we adopt elitism and proportional selection so that the fittest model is kept into the next generation, and that fitter models are more likely to be chosen as the "parent" models to breed the next generation, respectively. Uniform crossover is then performed to generate one "child" model by recombining the genes from each pair of parent models. Finally, a mutation operator is applied to randomly alter chosen child genes. Besides the uniform mutation, we propose a new adaptive mutation strategy using the vari-

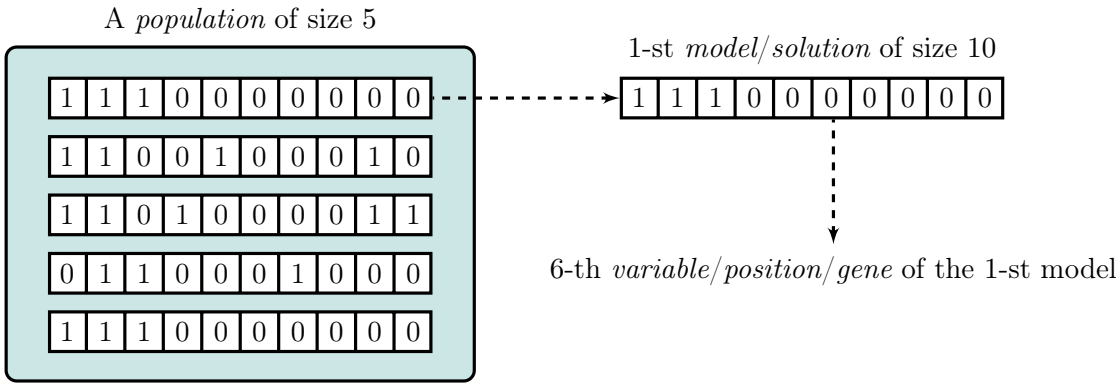

Figure 1.1.: An example of GA terminology. Note that the term *population* in GA is different from what a "population" means in statistics.

able association strength to enhance the variable selection performance. The genetic operations are iteratively performed until the size of the new generation reaches that of the previous one; see Figure 2.2. It is worth noting that the crossover operator generates new models similar to their parents (i.e., local search), while the mutation operator increases the population diversity to prevent GAs from being trapped in local optimum (thus resulting in global search). See Section 2 for more details.

In theory, we investigate the convergence properties of the GA in Theorem 3.1.1 based on the Markov chain theory. A practical consequence is to design an adaptive termination strategy that significantly reduces the computational cost. Furthermore, we prove that a fitter schema (a collection of solutions with specific structures; see Definition 3.2.1) is more likely to survive and be expanded in the next generation, using the schema theory (Theorem 3.2.1 and Corollary 3.2.1). This implies that the average fitness of the subsequent population gets improved, which entitled the "survival of the fittest" phenomenon of the natural selection.

Our results are applied to variable selection and model confidence set (MCS). In the former, the GA generates a manageable number of models (that is much smaller than all models up to some pre-determined size), over which the true model is found; see Proposition 4.1.1. As for the latter, the collected models in the model confidence

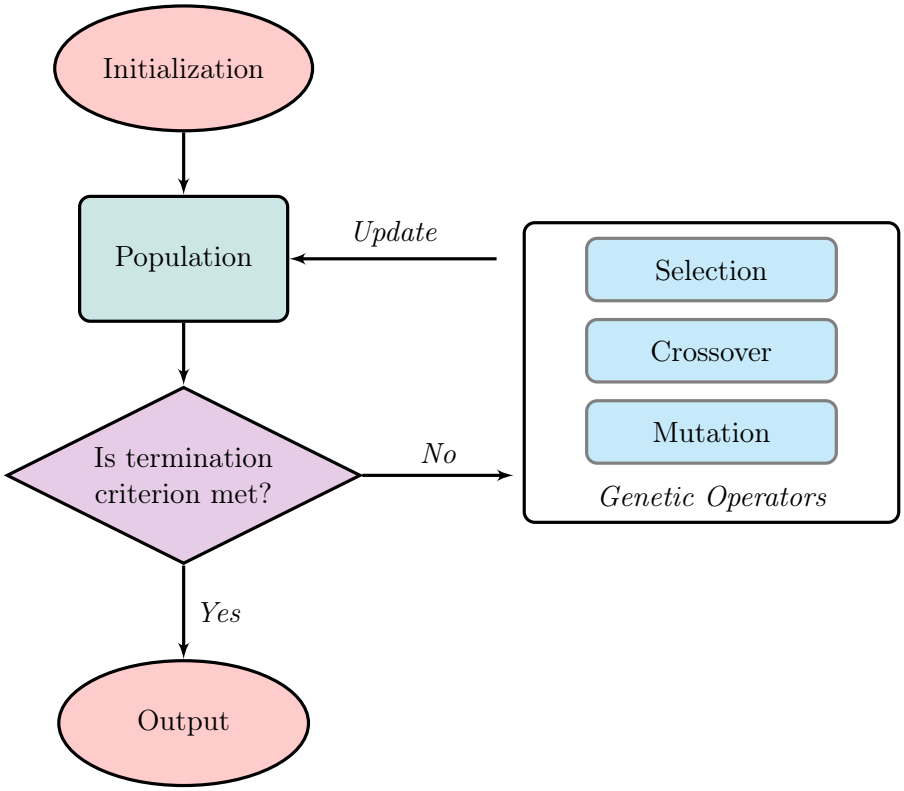

Figure 1.2.: A flowchart of a generic GA. It starts with an initial population and is updated with genetic operations until a termination criterion is met.

sets constructed by the GA are shown to be not statistically worse than the true model with a certain level of confidence; see Proposition 4.2.1.

As far as we are aware, two other methods can also be used to prepare candidate models: (i) collecting distinct models on regularization paths of penalized estimation methods (e.g., Lasso [11], SCAD [12] and MCP [13]), called as "regularization paths (RP)" method; (ii) a simulated annealing (SA) algorithm recently proposed by [14]. The former has no rigorous control on the quality of candidate models since model evaluation is not taken into account, and the latter needs a pre-determined model size and an efficiency threshold to filter out bad models. In comparison, the GA uses information criterion based fitness function to search for good models, and produces models of various sizes. As a result, the candidate models produced by the GA lead to much improved multi-model inference results, as demonstrated in Sections 5 and 6. [6]

and [15] proposed approaches to prepare candidate models that do not work for general multi-model inference applications. Best subset selection and forward stepwise regression can generate solution paths similar to the Lasso [16, 17]. However, the former imposes intractable computational burden and the latter lacks of comprehensive theoretical investigation.

Extensive simulation studies are carried out in Section 5 to demonstrate the power of the GA in comparison with the RP and the SA in terms of computation time, quality of the candidate model set, and performance of multi-model inference applications. In particular, the GA-best model exhibits the best variable selection performance in terms of the high positive selection rate and low false positive rate. For model averaging and variable importance, the GA results in at least comparable performance to the RP and the SA, but exhibits greater robustness than the SA. Additionally, the GA is also shown to possess better applicability than the RP in optimal high-dimensional model averaging.

Two real data examples are next carried out to illustrate the practical utility of the GA. For the riboflavin dataset [18], the GA-best model finds an informative gene which has not stood out in the literature (e.g., [18–22]). For the residential building dataset [23, 24], we identify factors, such as preliminary estimated construction cost, duration of construction, and 1-year delayed land price index and exchange rate, relevant to construction costs. These findings are further confirmed by the variable importance results using the SOIL [7]. Moreover, compared with the aforementioned competing methods, we again find that the GA generates the best candidate model set and results in the best model averaging performance on both datasets.

The rest of this dissertation is organized as follows. In Section 2 we present the GA for global model search, and list several possible ways for improving the implementation. In Section 3 the GA is analyzed using the Markov chain and schema theories. In Section 4 we illustrate how the GA assists multi-model inference tools such as variable selection and model confidence set. Sections 5 and 6 present extensive simulation studies and two real data analysis. In Section 7, we discuss future

works. All proofs and supplementary materials are presented in Sections A and B, respectively, in the appendix.

2. METHODOLOGY

Consider a linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ is the response vector, $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d]$ is the design matrix with $\boldsymbol{X}_j$ representing the j-th column for $j = 1, \ldots, d$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is the noise vector with $\mathsf{E}[\varepsilon_i] = 0$ and $\mathsf{Var}(\varepsilon_i) = \sigma^2$. Suppose $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_d^0)^\top$ is s-sparse (i.e., $\|\boldsymbol{\beta}^0\|_0 = s$) with $s \ll \min(n, d)$. Throughout this paper, s and d are allowed to grow with n.

Genetic representation for variable selection. The genetic representation of a model is defined as a binary sequence of length d, say $u = (u_1, \ldots, u_d)$, and variable j is said to be active (inactive) if $u_j = 1$ ($u_j = 0$). For example, $u = (1, 1, 1, 0, 0, 0, 0, 0, 0, 0)$ denotes the model with $d = 10$ variables but only the first three variables being active. Note that $|u| = \sum_{j=1}^d u_j$ denote the model size. Denote $\boldsymbol{X}_u$ as the submatrix of $\boldsymbol{X}$ subject to u, and $\mathcal{M} = \{0, 1\}^d$ as the model space.

Fitness function. Let $\Psi(t)$ denote the t-th generation of population, and $\overline{\Psi}(t) = \cup_{t'=0}^t \Psi(t')$ the collection of all models that have appeared up to the t-th generation. For any model $u \in \Psi(t)$, the fitness function is then defined as

$$f(u) = \begin{cases} -\,\mathrm{GIC}(u) & \text{if } |u| < n \\ \min\limits_{v \in \overline{\Psi}(t), |v| < n} -\,\mathrm{GIC}(v) & \text{if } |u| \geq n \end{cases}, \tag{2.2}$$

where

$$\mathrm{GIC}(u) = n \log \widehat{\sigma}_u^2 + \kappa_n |u|, \tag{2.3}$$

is the generalized information criterion (GIC, [25, 26]) and

$$\widehat{\sigma}_u^2 = \boldsymbol{Y}^\top \left[\boldsymbol{I}_n - \boldsymbol{X}_u (\boldsymbol{X}_u^\top \boldsymbol{X}_u)^{-1} \boldsymbol{X}_u^\top \right] \boldsymbol{Y}/n$$

is the mean squared error evaluated by the model u. GIC covers many types of information criteria (e.g., AIC [27] with $\kappa_n = 2$, BIC [28] with $\kappa_n = \log n$, modified BIC [29] with $\kappa_n = \log\log|u| \log n$ with $d < n$ and extended BIC [30] with $\kappa_n \asymp \log n + 2\log d$ with $p \geq n$). Since GIC cannot be computed for $|u| \geq n$, we define it as the worst fitness value up to the current generation. The rational is that any model with size larger than n should be unfavorable to all models with size smaller than n given the assumption that $s \ll \min(n, d)$. This definition warrants an unconstrained optimization, which is convenient for subsequent theoretical analysis. This is different from other ways to deal with the "infeasible solutions" in the GA literature, e.g., the "death penalty" in [31] and [32], which lead to constrained optimization.

2.1 A Genetic Algorithm for Candidate Model Search

We propose a genetic algorithm to search for good candidate models in Algorithm 1. Specifically, we use the RP method to generate an initial population, and then adopt proportional selection, uniform crossover and mutation operators to constitute the evolutionary process. Besides uniform mutation, we propose another mutation strategy based on the strength of variable association for improving empirical performances. An adaptive termination strategy is also proposed to enhance the computational efficiency. See Algorithm 1 for the overview of the GA.

Algorithm 1 A Genetic Algorithm for Model Search

Require: Population size K and mutation rate π_m

Generate initial population $\Psi(0) = \{u^1(0), \ldots, u^K(0)\}$
$t \leftarrow 0$
Converge $\leftarrow$ *False*
do
    $t \leftarrow t + 1$
    (Fitness evaluation) Compute fitness values $f\big(u^k(t-1)\big), k = 1, \ldots, K$
    (Elitism selection) Set $u^1(t) = \arg\max_{u \in \Psi(t-1)} f(u)$
    for $k = 2, \ldots, K$ **do**
        (Proportional selection) Randomly select two models from $\Psi(t-1)$ using w_k in (2.4)
        (Uniform crossover) Breed a child model using (2.5)
        (Mutation) Mutate the child genes using (2.6) or (2.7)
    end for
    Set $\Psi(t) = \{u^1(t), \ldots, u^K(t)\}$
    if Convergence criterion (2.8) is met **then**
        $T \leftarrow t$
        Converge $\leftarrow$ *True*
    end if
while *Converge* is *False*
return $\Psi(T) = \{u^1(T), \ldots, u^K(T)\}$

Initialization: The initial population $\Psi(0) = \{u^1(0), \ldots, u^K(0)\}$ only has very minimal requirement as follows: (i) $K \geq 2$ and (ii) $|u^k(0)| < n$ for some $k = 1, \ldots, K$ (i.e., at least one model with commutable GIC). The condition (i) allows the GA to explore through the model space $\mathcal{M}$; see Section 3.1, and (ii) ensures $f\big(u^k(0)\big), k = 1, \ldots, K$, are all available. The choice of K will be discussed in Section 2.2.1. For fast convergence of the GA, we recommend the RP method to generate initial population. Please see Figure 2.1 for how models produced by RP are improved by the GA in terms of GIC.

Given the $(t-1)$-th generation $\Psi(t-1) = \{u^1(t-1), \ldots, u^K(t-1)\}$, the GA produces the next generation $\Psi(t) = \{u^1(t), \ldots, u^K(t)\}$ through proportional selection, uniform crossover and mutation operations. See Figure 2.2 to visualize the evolution process. In what follows, we give details for each step in our main algorithm.

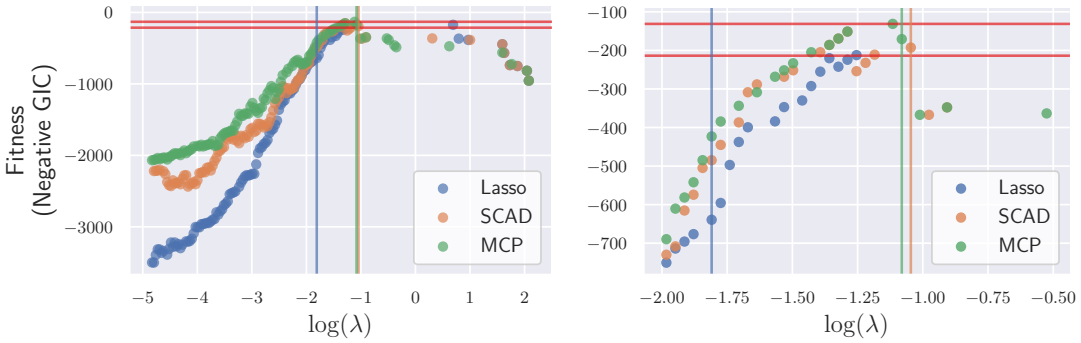

Figure 2.1.: Fitness values of the models obtained from the regularization paths of Lasso, SCAD and MCP. The two red horizontal lines indicate the best and worst fitness values of the GA models, and the vertical lines locate the best λ selected by 10-fold cross-validation. The right panel is a zoomed view of the left panel around the selected λ. Among the 304 RP models, only 10 of them have fitness values not smaller than the GA-worst model. The result is obtained from the first dataset under the simulation Case 1 with $(n, d, s, \rho) = (200, 400, 6, 0.5)$.

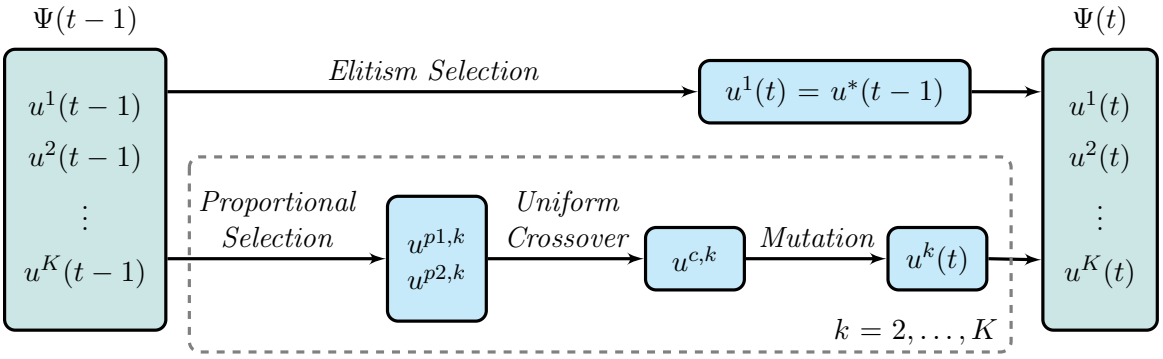

Figure 2.2.: Illustration of the evolution process of the GA. $u^*(t-1)$ denotes the best model in $\Psi(t-1)$. The candidate pool of the proportional selection is the entire $\Psi(t-1)$, which still includes $u^*(t-1)$. For each $k = 2, \ldots, K$, a pair of parent models are selected according to the probability w_k and one child model $u^{c,k}$ is generated through uniform crossover (2.5). Finally, $u^{c,k}$ is processed by the mutation (2.6) or (2.7) to produce $u^k(t)$.

In the **elitism selection** step, we choose $u^*(t-1) := \arg\max_{u \in \Psi(t-1)} f(u)$, i.e., the best model in $\Psi(t-1)$ is kept into $\Psi(t)$, and define it as $u^1(t)$ for simplicity. The **proportional selection** step chooses parent models from $\Psi(t-1)$ (including $u^*(t-1)$) based on the exponentially scaled fitness as follows. Define the fitness

$f_k = f\big(u^k(t-1)\big)$ according to (2.2). For $k = 1, \ldots, K$, first compute the weight w_k for $u^k(t-1)$ as

$$w_k = \frac{\exp\left(f_k/2\right)}{\sum_{l=1}^{K} \exp\left(f_l/2\right)}, \quad k = 1, \ldots, K. \tag{2.4}$$

Then $(K-1)$ pairs of models are randomly selected with replacement from $\Psi(t-1)$, where the probability of selecting $u^k(t-1)$ is w_k. Note that the exponentially scaled information criteria are often used for model weighting in multi-model inference (e.g., [1, 33, 34]).

Each pair of parent models produces a child model by performing **uniform crossover** with equal mixing rate (i.e., each child position has equal chance to be passed from the two parents). That is, let $u^{p1,k} = (u_1^{p1,k}, \ldots, u_d^{p1,k})$ and $u^{p2,k} = (u_1^{p2,k}, \ldots, u_d^{p2,k})$ be the chosen parent models, and then the genes in the child model $u^{c,k} = (u_1^{c,k}, \ldots, u_d^{c,k})$ is determined by

$$u_j^{c,k} = \begin{cases} u_j^{p1,k} & \text{with probability } 1/2 \\ u_j^{p2,k} & \text{otherwise} \end{cases}, \quad j = 1, \ldots, d. \tag{2.5}$$

In the last step, we apply **mutation** to the child model $u^{c,k}$. Given a mutation probability π_m (usually low, such as $\pi_m = 0.01$ or $1/d$), we consider the following two mutation schemes. Denote by $u^k(t) = \big(u_d^1(t), \ldots, u_d^k(t)\big)$ the resulting model after mutation being applied to $u^{c,k}$.

- **Uniform mutation:** Genes in $u^{c,k}$ are randomly flipped with probability π_m, i.e.,

$$u_j^k(t) = \begin{cases} 1 - u_j^{c,k} & \text{with probability } \pi_m \\ u_j^{c,k} & \text{otherwise} \end{cases}, \quad j = 1, \ldots, d. \tag{2.6}$$

- **Adaptive mutation:** We propose a data-dependent mutation operator based on the variable association measures γ_j. For example, γ_j can be either the

marginal correlation learning $\left|\widehat{\mathsf{Cor}}(\boldsymbol{X}_j, \boldsymbol{Y})\right|$ [35] or the high-dimensional ordinary least-squares projection $\left|\boldsymbol{X}_j(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{Y}\right|$ [36] (available only for $d \geq n$). Let $V_+^k = \{j : u_j^{c,k} = 1\}$ and $V_-^k = \{j : u_j^{c,k} = 0\}$. Define the mutation probability for the $u_j^{c,k}$ as

$$\bar{\pi}_{m,j}^k = \begin{cases} \dfrac{\gamma_j^{-1}}{\sum_{l \in V_+^k} \gamma_l^{-1}} |V_+^k| \pi_m & \text{if } j \in V_+^k \\ \dfrac{\gamma_j}{\sum_{l \in V_-^k} \gamma_l} |V_-^k| \pi_m & \text{if } j \in V_-^k \end{cases}.$$

Then the proposed mutation operation is performed by

$$u_j^k(t) = \begin{cases} 1 - u_j^{c,k} & \text{with probability } \bar{\pi}_{m,j}^k \\ u_j^{c,k} & \text{otherwise} \end{cases}, \quad j = 1, \ldots, d. \tag{2.7}$$

By defining $\bar{\pi}_{m,j}^k$ this way, unimportant active variables are more likely to be deactivated, and important inactive variables are more likely to be activated. Also, it can be easily seen that this mutation operation results in the same expected number of deactivated and activated genes as those of uniform mutation operation. As far as we are aware, this is the first data dependent mutation method in the GA literature.

In numerical experiments, we note that the adaptive mutation performs slightly better than the uniform mutation. For space constraint, we just focus on the adaptive mutation with $\pi_m = 1/d$.

As for **termination**, we propose an adaptive criterion by testing whether the average fitness becomes stabilized; see Section 2.2.2 for more details. This is very different from the user specified criteria used in GA literature such as the largest number of generations (e.g., [37]) or the minimal change of the best solution (e.g., [38]).

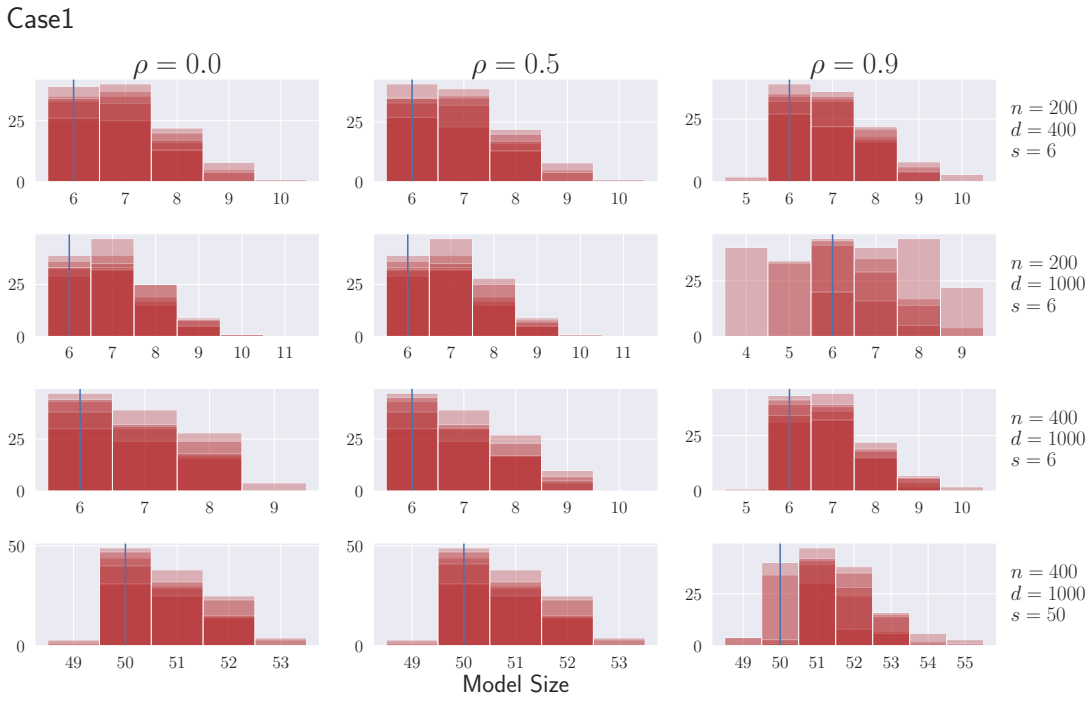

Figure 2.3.: The overlapped distributions of the sizes of the final candidate models collected by the GA, with the blue vertical line indicating the true model size. The results are obtained under the simulation Case 1 (see Section 5 for more details) and other cases exhibit similar patterns.

Remark 2.1.1 *We note that the models collected by the GA are in nature sparse since their sizes are around the true model size s; see Figure 2.3 for example. This empirically appealing feature allows us to construct GA-based sparse model confidence sets in the later sections.*

2.2 Computational Considerations

Computational concern has been the major critiques that prevent GAs from being popular over other optimization methods such as gradient descent in machine learning and statistical communities. In our experience, the most computational cost is taken by the calculation of the fitness evaluation, which could be alleviated by reducing the population size (Section 2.2.1) and the number of generations (Section 2.2.2).

2.2.1 Population Sizing

The population size K plays an important role in GAs. It is obvious that larger population makes GAs computationally more expensive. On the other hand, empirical results indicate that small population would jeopardize performance (e.g., [39–41]). We found that the minimum population size suggested in [42] makes a good balance. The idea is to have a population such that every possible solution in the search space should be reachable from an randomly generated initial population by crossover only. In binary gene coding cases, it means that the solutions in the initial population cannot be all 0 or 1 for any position. For any K, the probability of such an event can be found by

$$P^* = (1 - 1/2^{K-1})^d = \exp\left[d \log(1 - 1/2^{K-1})\right] \approx \exp(-d/2^{K-1}).$$

Accordingly, for every given P^*, we can calculate the minimum population size

$$K^* \approx \left\lceil 1 + \log(-d/\log P^*)/\log 2 \right\rceil,$$

where $\lceil a \rceil$ is the smallest integer larger than $a \in \mathbb{R}$. For example, a population of size $K = 25$ is enough to ensure that the required probability exceeds 99.99% when $d = 1{,}000$.

In our implementation, we conservatively use

$$K = 4\left\lceil 1 + \log(-d/\log P^*)/\log 2 \right\rceil$$

with $P^* = 0.9999$, to specify the population size according to model dimension.

2.2.2 Adaptive Termination

To adaptively terminate, we perform an independent two-sample t-test on whether the average fitness of $\Psi(t)$ and $\Psi(t-10)$ are the same at a significance level 0.05:

$$H_0^t: \ \bar{f}\big(\Psi(t)\big) = \bar{f}\big(\Psi(t-10)\big) \quad \text{v.s.} \quad H_1^t: \ \bar{f}\big(\Psi(t)\big) \neq \bar{f}\big(\Psi(t-10)\big), \tag{2.8}$$

where $\bar{f}\big(\Psi(t)\big)$ is the average fitness of the t-th generation. The T-th generation is set to be the final generation if T is the smallest $t \geq 10$ such that H_0^t is rejected. The generation gap 10 is meant to weaken the correlation between the two generations being tested. Note that the GA can be regarded as a Markov chain (see Section 3.1 for details) and therefore there exists dependence among generations. Hence, it is not appropriate to perform two-sample t-test of the average fitness from two consecutive generations.

Remark 2.2.1 *This termination criterion is constructed based on the limiting distribution derived for the associated Markov chain (see the discussion below Theorem 3.1.1 for more details) and results in huge computational efficiency. In the literature, the GA iteration is often terminated at a fixed, predetermined number of generations, say $T_{\max}$, which is usually large such as $50, 100$ or even larger (e.g., [43, 44]). Our termination criterion, on the other hand, entitles a scientific check for the convergence. With the RP used for generating the initial population, the GA enters the stationary distribution (as the average fitness is tested to be stabilized) in just a few generations (say, around 20 generations). In addition, we note that the computational cost incurred by the independent two-sample t-test (2.8) is negligible, as the fitness values are computed as the models are generated.*

3. THEORETICAL PROPERTIES

In this section, we study the theoretical properties of the GA, which belongs to the so-called canonical GA (CGA) family[1] [8]. The proposed GA is a CGA that specifically employs elitism and proportional selection, uniform crossover and uniform or adaptive mutation as described in Section 2. We first investigate the convergence properties for a general CGA family based on Markov chain theory, i.e., Theorem 3.1.1. Furthermore, Theorem 3.1.2 presents a brand new theoretical framework to construct MCSs for the globally best model. We next establish a new schema theory (Theorem 3.2.1 and Corollary 3.2.1) to elicit the evolutionary mechanism for the GA. It is worthy noting that the theoretical results established in this section apply to the general CGA framework and hence not restricted to the specific variable selection problem.

3.1 Convergence Analysis

In this section, we show that the Markov chain associated with a CGA class has a unique stationary distribution from which the asymptotic inclusion of the globally best model, i.e., global convergence, can be deduced. Such a result justifies the adaptive termination rule in Section 2.2.2, and can be used to reduce the search space for variable selection problems; see Proposition 4.1.1. Note that the theoretical results obtained in this section hold for any finite sample size.

Recall that $\Psi(t) = \left\{u^1(t), \ldots, u^K(t)\right\}$ represents the t-th generation of the population, and denote by $\left\{\Psi(t)\right\}_{t\geq 0}$ the associated Markov chain with values on the finite state (population) space $\mathcal{M}^K$. The corresponding transition matrix is defined

[1]CGAs are also called as simple or standard GAs in the literature. CGA uses binary sequence for solution representation, and updates a fixed-sized population via selection, crossover and mutation operators.

as $\boldsymbol{P} = \left[P_{\boldsymbol{uv}}\right]_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{M}^K}$, where $P_{\boldsymbol{uv}} = P\big(\Psi(t+1) = \boldsymbol{v}\big|\Psi(t) = \boldsymbol{u}\big)$. We need the following definitions for our subsequent analysis.

Definition 3.1.1 *A square matrix* $\boldsymbol{A} = [a_{ij}] \in \mathbb{R}^{K\times K}$ *is said to be* non-negative (positive) *if* $a_{ij} \geq 0$ *(*$a_{ij} > 0$*) for all* $i, j \in \{1, \ldots, d\}$*. A non-negative square matrix* $\boldsymbol{A}$ *is said to be*

(a) primitive *if there exists a positive integer* k *such that* $\boldsymbol{A}^k$ *is positive;*

(b) reducible *if there exists two square matrices* $\boldsymbol{A}_{11}$ *and* $\boldsymbol{A}_{22}$ *and a matrix* $\boldsymbol{A}_{21}$ *with suitable dimensions such that* $\boldsymbol{A}$ *can be expressed as the form*

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{O} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix},$$

where $\boldsymbol{O}$ *denotes a zero matrix with suitable dimensions, by applying the same permutations to rows and columns;*

(c) irreducible *if it is not reducible;*

(d) stochastic *if* $\sum_{j=1}^{K} a_{ij} = 1$ *for all* $i = 1, \ldots, K$*.*

Let

$$u^* := \arg\max_{u\in\mathcal{M}} f(u)$$

denote the best model in $\mathcal{M}$ and suppose it is unique, i.e., $f(u^*) > f(u)$ for all $u \in \mathcal{M} - \{u^*\}$. Moreover, denote the collection of states that contains u^* by

$$\mathcal{M}_{\max} = \Big\{\boldsymbol{u} = \big\{u^1, \ldots, u^K\big\} \in \mathcal{M}^K : u^* \in \boldsymbol{u}\Big\}. \tag{3.1}$$

The following theorem describes two important convergence properties.

Theorem 3.1.1 *Let* $\boldsymbol{P}$ *denote the transition probability matrix of the Markov chain associated with a CGA with elitism selection, population size* $K \geq 2$ *and mutation rate* $\pi_m \in (0, 1)$*.*

(a) There exists a unique stationary distribution $\boldsymbol{\pi} = \big(\pi(\boldsymbol{u}) : \boldsymbol{u} \in \mathcal{M}^K\big)^\top$ *that satisfies* $\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \boldsymbol{P}$ *and* $\pi(\boldsymbol{u}) = \lim_{t\to\infty} P\big(\Psi(t) = \boldsymbol{u}\big)$ *with* $\pi(\boldsymbol{u}) > 0$ *for* $\boldsymbol{u} \in \mathcal{M}_{\max}$ *and* $\pi(\boldsymbol{u}) = 0$ *for* $\boldsymbol{u} \notin \mathcal{M}_{\max}$.

(b) (Theorem 6 of [45]) We have

$$\lim_{t\to\infty} P\big(u^* \in \Psi(t)\big) = 1. \tag{3.2}$$

As far as we are aware, the existence of the stationary distribution stated in Theorem 3.1.1 (a) for CGAs *with elitism selection* is new, even though similar results for non-elitist CGAs has been presented for over decades (e.g., [45, 46]). This is in contrast with the GA literature that typically concerns global convergence (e.g., [45–47]) rather than the stationary distribution. As for Theorem 3.1.1 (b), the elitism selection is a necessary condition [45, 47] and it is different from the path-consistency property of non-convex penalization approaches (e.g., [48, 49]) in that the former captures the *best* model for any sample size n as $t \to \infty$ and the latter targets at the *true* model as $n \to \infty$. Later, Theorem 3.1.1 (b) is extended to a selection consistency result as $n \to \infty$; see Proposition 4.1.1.

Part (a) of Theorem 3.1.1 has the following implication. Recall that $\bar{f}(\boldsymbol{u})$ is the average fitness of any population $\boldsymbol{u}$, and thus we have

$$\lim_{t\to\infty} \mathsf{E}\big[\bar{f}\big(\Psi(t)\big)\big] = \sum_{\boldsymbol{u}\in\mathcal{M}^K} \pi(\boldsymbol{u})\bar{f}(\boldsymbol{u}),$$

which is a constant given data $(\boldsymbol{X}, \boldsymbol{Y})$. This indicates that the average fitness oscillates around a constant in the long run, as $\Psi(t)$ becomes stabilized (i.e., the associated Markov chain converges). This justifies the termination check in (2.8).

Remark 3.1.1 *It is worth noting that Theorem 3.1.1 does not only apply to the GA but also any CGA with elitism selection. The key reason is that the child solutions generated through selection, crossover and mutation operators always remain in the search space for unconstrained optimization or search problems. Accordingly, instead*

of the ones mentioned in Section 2, Theorem 3.1.1 still holds for any other selection, crossover and mutation operations (e.g., rank-based or tournament selection [50] and the newly proposed crossover and mutation operations developed in [51] and [52], respectively).

In contrast to the asymptotic result in Theorem 3.1.1 (b) as $t \to \infty$, it is also of practical relevance to construct a $100(1-\alpha)\%$ MCS that covers the best model u^* after a finite number of generations. A particularly appealing feature is that every model in this set is sparse. This is conceptually different from the MCS constructed based on the debiased principle [53–55], which mostly contains dense models.

Theorem 3.1.2 *Let $\Psi(t)$ denote the t-th population of a CGA with elitism selection, $K \geq 2$ and $\pi_m \in (0,1)$. Then for any $\alpha \in (0,1)$ there exists a positive integer T_α such that*

$$P\big(u^* \in \Psi(t)\big) \geq 1-\alpha \tag{3.3}$$

for any $t \geq T_\alpha$.

The proof of Theorem 3.1.2 implies the global convergence property described in Theorem 3.1.1 (b) by letting $\alpha = 0$ and thus $T_0 = \infty$. From the proof of Theorem 3.1.2, we note that obtaining the value of T_α requires the knowledge of the constant ξ as defined in (A.4), which is often unknown. By definition, ξ can be obtained by estimating the submatrix $\boldsymbol{R}$ in the transition matrix $\boldsymbol{P}$. That is, ξ is the smallest of the row sums of $\boldsymbol{R}$. Since $\boldsymbol{R}$ has $\big|\mathcal{M}^K\big| - \big|\mathcal{M}_{\max}\big| = \binom{K+2^d-1}{K} - \binom{K+2^d-2}{K-1}$ rows and $\big|\mathcal{M}_{\max}\big| = \binom{K+2^d-2}{K}$ columns, the size of $\boldsymbol{R}$ is massive. For instance, when $(K,d) = (10,5)$, there are about 3×10^{17} elements in $\boldsymbol{R}$. Albeit [56] provides a useful formula to compute the elements in the $\boldsymbol{P}$, the computational cost is too large to be carried out in practice. Hence, we leave an accurate estimation or approximation of ξ to future study.

3.2 Evolvability Analysis

In this section, we establish a schema theorem to study the evolution process of the GA. Specifically, it is proven that the average fitness gets improved over generations. To the best of our knowledge, we are the first to develop a schema theorem for GAs with proportional selection, uniform crossover and uniform mutation at the same time in the GA literature. The most closely related schema theorems are provided by [57, 58] for GAs with proportional selection and one-point crossover, and by [59] for GAs with uniform crossover alone.

In the following we give the definition of a schema with general GA terminology (i.e., using "solutions" instead of "models"), followed by an example as illustration.

Definition 3.2.1 *A* schema $H = (H_1, \ldots, H_d) \in \{0, 1, *\}^d$ *is a ternary sequence of length* d*, where the "*$*$*" is a wildcard symbol, meaning that we do not care whether it is* 0 *or* 1*. The indices where the schema has a* 0 *or* 1 *are called the* fixed positions*. We say a solution* $u = (u_1, \ldots, u_d)$ *matches* H *if all fixed positions of* H *are the same as the corresponding positions in* u*. The* order of a schema H*, denoted by* $\mathrm{ord}(H)$*, is defined by the number of fixed positions in* H*. Moreover, by adopting the notations used in the order theory (e.g., [60]), for any schema* H *we define the* expansion operator $\uparrow(H)$ *to map* H *to the set of all possible solutions that match* H*, i.e.,*

$$\uparrow(H) = \big\{u \in \mathcal{M} : u_j = H_j \text{ or } H_j = * \text{ for each } j = 1, \ldots, p\big\}.$$

Example 3.2.1 *Suppose a schema* $H = (1, 0, *, 0, *)$*. In this case,* $\mathrm{ord}(H) = 3$*, and* $\uparrow(H) = \big\{(1,0,0,0,0), (1,0,0,0,1), (1,0,1,0,0), (1,0,1,0,1)\big\}$.

Let $m(H, t)$ denote the number of solutions that match a schema H in the t-th generation, and $\alpha(H, t)$ the probability that the schema H survives or is created after

the t-th generation. [61] noted that $m(H, t+1)$ follows a binomial distribution with the number of trials K and success probability $\alpha(H, t)$, i.e., (K is the population size)

$$m(H, t+1) \sim Binomial\big(K, \alpha(H, t)\big). \tag{3.4}$$

Hence, we have $\mathsf{E}\big[m(H, t+1)\big] = K\alpha(H, t)$. Accordingly, higher $\alpha(H, t)$ leads to higher $\mathsf{E}\big[m(H, t+1)\big]$ and thus tends to result in more solutions of H in the next generation. Since the population size is fixed, more solutions of a fitter schema imply higher average fitness in the subsequent generation. Hence, we will show that $\alpha(H^1, t)$ is larger than $\alpha(H^2, t)$ if the average fitness of H^1 is larger than that of H^2.

To prove the above result, we need to define the following different notions of Hamming distance. The first concerns two models $u = (u_1, \ldots, u_p)$ and $v = (v_1, \ldots, v_p)$, i.e., $\delta(u, v) = \sum_{j=1}^{p} \mathbb{1}(u_j \neq v_j)$, while the second type of Hamming distance is between a model and a schema H on the fixed positions: $\delta(u, H) = \sum_{j:H_j \neq *} \mathbb{1}(u_j \neq H_j)$. The last one is Hamming distance between models u and v with respect to the fixed positions of any schema H: $\delta_H(u, v) = \sum_{j:H_j \neq *} \mathbb{1}(u_j \neq v_j)$.

We are now ready to characterize $\alpha(H, t)$ explicitly for the GA with uniform mutation. Recall from (2.4) that w_k denotes the probability that model u^k is selected as a parent model.

Theorem 3.2.1 *Given the t-th generation $\Psi(t) = \{u^1, \ldots, u^K\}$ and a schema H, define the probability that a solution matching H is selected by the proportional selection operator as*

$$\alpha_{sel}(H, t) = \sum_{k:u^k \in \uparrow(H)} w_k.$$

For the GA with uniform mutation, we have

$$\alpha(H,t) = \alpha_{sel}^2(H,t)(1-\pi_m)^{\mathrm{ord}(H)} + \alpha_{sel}(H,t)\sum_{l:u^l\notin\uparrow(H)} w_l \frac{(1-\pi_m)^{\mathrm{ord}(H)}}{\big[2(1-\pi_m)\big]^{\delta(u^l,H)}} + \sum_{k,l:u^k,u^l\notin\uparrow(H)} w_k w_l \frac{(2\pi_m)^{h_{kl}}(1-\pi_m)^{\mathrm{ord}(H)}}{\big[2(1-\pi_m)\big]^{\delta_H(u^k,u^l)}}, \tag{3.5}$$

where $h_{kl} = \big|\{j : H_j \neq *, u_j^k = u_j^l \neq H_j\}\big|$.

The general result in Theorem 3.2.1 provides an exact form of $\alpha(H,t)$, which is quite difficult to interpret and analyze. Accordingly, we derive a simple-to-analyze lower bound for $\alpha(H,t)$.

Corollary 3.2.1 *Suppose conditions in Theorem 3.2.1 hold. For* $\pi_m \leq 0.5$*, we have*

$$\alpha(H,t) \geq (1-\pi_m)^{\mathrm{ord}(H)}\alpha_{sel}(H,t)^2 + 2^{-\mathrm{ord}(H)}\alpha_{sel}(H,t)\big[1-\alpha_{sel}(H,t)\big] + \big[1-\alpha_{sel}(H,t)\big]^2\pi_m^{\mathrm{ord}(H)}. \tag{3.6}$$

It can be seen from (3.6) that the lower bound of $\alpha(H,t)$ gets larger when the schema selection probability $\alpha_{sel}(H,t)$ increases or the schema H has lower order (i.e., $\mathrm{ord}(H)$ is small). By definition, fitter schema H leads to larger $\alpha_{sel}(H,t)$ and therefore higher $\alpha(H,t)$ and $\mathsf{E}\big[m(H,t+1)\big]$. Since an expansion of the fitter schema H is expected in a fixed-size population, fitter models matching H are more likely to be generated in place of weaker models; see Section 5.2 for a numerical verification. Accordingly, the subsequent generation is anticipated to have higher average fitness. This entitles the "survival of the fittest" phenomenon of the natural selection and acknowledges the evolvability of the GA.

4. GA-ASSISTED MULTI-MODEL INFERENCE

In this section, we describe how the GA helps multi-model inferences. Note that existing information-criteria based variable selection (e.g., [30, 62, 63]) and MCS procedures (e.g., [4,64,65]) typically concern the true model rather than the globally best model, which is the target of the GA. To bridge this gap, we first present a lemma suggesting that the true model indeed possess the lowest GIC value and therefore become the globally best model in large samples.

The following regularity condition is needed.

Assumption 4.0.1 *(A1) There exists a positive constant C_1 such that $\lambda_{\min}\big(\boldsymbol{X}_{u^0}^\top \boldsymbol{X}_{u^0}/n\big) > C_1$ for all n, where u^0 denotes the true model;*

(A2) There is a positive constant C_2 such that

$$\inf_{u \neq u^0, |u| < \tilde{s}} \boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{H}_u)\boldsymbol{\mu} \geq C_2 n,$$

where $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}^0$ and $\boldsymbol{H}_u = \boldsymbol{X}_u(\boldsymbol{X}_u^\top \boldsymbol{X}_u)^{-1}\boldsymbol{X}_u^\top$ denotes the hat matrix of the model u, for some positive integer $\tilde{s}$ with $s \leq \tilde{s} < n$.

Condition (A1) ensures the design matrix of the true model is well-posed and Condition (A2) is the asymptotic identifiability condition used in [30], indicating that the model is identifiable if no model with comparable size can predict as well as the true model.

Recall that κ_n is defined in the GIC formulation (2.3).

Lemma 4.0.1 *Suppose Assumption 4.1 holds,* $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, $\log d = O(n^\tau)$, $\kappa_n = O(n^\tau)$ *and* $\kappa_n \to \infty$ *for some positive constant* $\tau < 1$. *Then for any positive integer* $\tilde{s}$ *satisfying* $\tilde{s} \geq s$ *and* $\tilde{s} \log d = o(n)$, *we have, as* $n \to \infty$,

$$\min_{u \in \mathcal{M}_{\tilde{s}} - \{u^0\}} \mathrm{GIC}(u) - \mathrm{GIC}(u^0) > 0, \tag{4.1}$$

where $\mathcal{M}_{\tilde{s}} = \big\{u \in \mathcal{M} : |u| \leq \tilde{s}\big\}$.

4.1 Variable Selection

The GA offers a practical way to perform variable selection by only searching the models generated by the GA instead of the whole model space. The existing information-criterion based selection methods search a constrained model space $\mathcal{M}_{\tilde{s}}$ for some $s \leq \tilde{s} \ll n$. However, by using the GA, we only need to evaluate at most $K \times T$ models (recall that K and T are the population size and the number of generations to convergence, respectively). For example, under the simulation Case 1 with $(n, d, s, \rho) = (200, 400, 6, 0.5)$ (see Section 5.1), it is nearly impossible to go through $\binom{400}{6} \approx 5.5 \times 10^{12}$ models with the true size 6, not to mention to compare all the models with sizes at most $\tilde{s}$ for some $6 \leq \tilde{s} \ll n$. On the other hand, the GA searches for the true model in all 500 simulation runs, each with less than 1,750 models evaluated ($K = 92$ and $T \leq 19$ generations to convergence).

By combining Theorem 3.1.1 (b) and Lemma 4.0.1, Proposition 4.1.1 shows that the true model becomes the best model in large samples and is eventually captured by the GA. Let $\Psi_{\tilde{s}}(t)$ denote the t-th generation of a GA population on the constrained model space $\mathcal{M}_{\tilde{s}}$. The fitness function (2.2) makes models of sizes at least n nearly impossible to be generated. Accordingly, it is equivalent to setting $\tilde{s} = n - 1$.

Proposition 4.1.1 *Suppose conditions in Lemma 4.0.1 hold and $\Psi_{\tilde{s}}(t)$ satisfies the conditions in Theorem 3.1.1. Define*

$$\widehat{u}(t) = \arg\min_{u \in \Psi_{\tilde{s}}(t)} \mathrm{GIC}(u).$$

Then we have

$$\lim_{t\to\infty} \lim_{n\to\infty} P\left(\widehat{u}(t) = u^0\right) = 1. \tag{4.2}$$

4.2 Model Confidence Set

In this section, we construct practically feasible model confidence sets with the aid of the GA, in comparison with the one based on Theorem 3.1.2. The main idea is to employ the two-step testing procedure of [66], given that the candidate models are produced by GA.

Given a candidate model set $\Psi = \{u^1, \ldots, u^K\}$, let $u^{\#} = \arg\min_{u\in\Psi} \mathrm{GIC}(u)$ denote the best candidate model in Ψ. Collect

$$\mathcal{A}_\alpha = \{u \in \Psi : H_{0,u} \text{ is not rejected at a significance level } \alpha\} \tag{4.3}$$

by performing the hypothesis testing

$$H_{0,u}: \text{ Model } u \text{ is not worse than } u^{\#} \quad \text{vs.} \quad H_{1,u}: \text{ Model } u \text{ is worse than } u^{\#}. \tag{4.4}$$

for every $u \in \Psi - \{u^{\#}\}$ at significance level α. We name the model confidence set $\mathcal{A}_\alpha$ as survival model set (SMS) since the models therein survive the elimination testing (4.4). Recall from Section 3.2 that the GA models, even after the globally best model is found, keep being improved until convergence. Accordingly, a manageable number of good (and sparse) models are included in the SMS when the GA is used to provide

candidate models. Later, we use the relative size $|\mathcal{A}_\alpha|/|\Psi|$ to measure the quality of the candidate model set in Section 5.3.3.

To perform the hypothesis testing (4.4) where u and $u^\#$ may not be nested, we employ the two-step procedure of [66] by decomposing (4.4) as first model distinguishability test

$$H_{0,u}^{dis}: \ u \text{ and } u^\# \text{ are indistinguishable} \quad \text{vs.} \quad H_{1,u}^{dis}: \ u \text{ and } u^\# \text{ are distinguishable} \tag{4.5}$$

and if $H_{0,u}^{dis}$ is rejected, then a superiority test

$$H_{0,u}^{sup}: \ \mathsf{E}\big[\,\mathrm{GIC}(u)\big] \le \mathsf{E}\big[\,\mathrm{GIC}(u^\#)\big] \quad \text{vs.} \quad H_{1,u}^{sup}: \ \mathsf{E}\big[\,\mathrm{GIC}(u)\big] > \mathsf{E}\big[\,\mathrm{GIC}(u^\#)\big]. \tag{4.6}$$

The rejection of $H_{0,u}$ at significance level α is equivalent to that $H_{0,u}^{dis}$ and $H_{0,u}^{sup}$ are both rejected at significance level α. We note that the original superiority test in [66] is based on likelihood ratio, and therefore certain adjustment is needed for our case; see Section B.1.1 for detailed description. The R package **nonnest2** [67] is used to test (4.5) and extract necessary quantities for the GIC-based superiority test (4.6).

The following proposition justifies the asymptotic validity of the constructed SMS.

Proposition 4.2.1 *Suppose conditions in Proposition 4.1.1 hold and $\Psi_{\tilde{s}}(t)$ satisfies the conditions in Theorem 3.1.1. Let $\mathcal{A}_\alpha(t)$ denote a $100(1-\alpha)\%$ SMS with $\Psi_{\tilde{s}}(t)$ serving as the candidate model set. Then we have*

$$\lim_{t\to\infty}\lim_{n\to\infty} P\big(u \in \mathcal{A}_\alpha(t)\big) \ge 1-\alpha$$

for all $u \in \Psi_{\tilde{s}}(t) - \{u^0\}$ such that $H_{0,u}$ is not rejected at significance level α.

5. SIMULATION STUDIES

In this section, we conduct extensive simulation studies to provide numerical support for the new schema theory supplied in Section 3.2 and show that the GA outperforms the RP method and the SA algorithm of [14]. The simulated data were generated based on the linear model (2.1) with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I})$. Each row of the design matrix $\boldsymbol{X}$ was generated independently from $\mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a Toeplitz matrix with the (k, l)-th entry $\Sigma_{kl} = \rho^{|k-l|}$ for $\rho = 0, 0.5$ and 0.9. Results were obtained based on 500 simulation replicates.

5.1 Simulation Settings

We consider six simulation cases below with both high-dimensional (i.e., $d \geq n$; Cases 1–4) and low-dimensional (Cases 5 and 6) settings. Cases 1 and 2 with $\rho = 0$ refers to the first two simulation cases used in [7]. Case 3 is inspired from the simulations settings used in [14], but our settings ensure X_{s+1} and X_{s+2} are always marginally distributed as $\mathcal{N}(0, 1)$ for any $\rho \in [0, 1)$. Case 4 is similar to Case 3 but with weak signals. Cases 5 and 6 refers to the simulation example 2 of [29], with weak signals in Case 6.

Let $\mathbf{1}_p$ and $\mathbf{0}_p$ denote the p-dimensional vectors of 1's and 0's, respectively.

Case 1: $\boldsymbol{\beta}^0 = (4\mathbf{1}_{s-2}^\top, -6\sqrt{2}, \frac{4}{3}, \mathbf{0}_{d-s}^\top)^\top$.

Case 2: $\boldsymbol{\beta}^0$ as in Case 1. Re-define $X_{s+1} = 0.5X_1 + 2X_{s-2} + \eta_1$, where $\eta_1 \sim \mathcal{N}(0, 0.01)$.

Case 3: $\boldsymbol{\beta}^0 = (3\mathbf{1}_s^\top, \mathbf{0}_{d-s}^\top)^\top$. Re-define $X_{s+1} = \frac{2}{3\sqrt{(1+\rho)}}(X_1 + X_2) + \eta_2$ and $X_{s+2} = \frac{2}{3\sqrt{(1+\rho)}}(X_3 + X_4) + \eta_3$, where $\eta_1, \eta_2 \overset{iid}{\sim} \mathcal{N}(0, 1/9)$.

Case 4: $\boldsymbol{\beta}^0 = (3\log(n)/\sqrt{n}\mathbf{1}_s^\top, \mathbf{0}_{d-s}^\top)^\top$. Re-define X_{s+1} and X_{s+2} as in Case 3.

Two cases are set up for moderate dimensional (i.e., $d < n$) scenarios:

Case 5: $\beta_1^0 \geq \cdots \geq \beta_s^0$ are iid *Uniform*$(0.5, 1.5)$, sorted decreasingly, and $\beta_j = 0$ for $j > s$.

Case 6: $\boldsymbol{\beta}^0 = (3\log(n)/\sqrt{n}\mathbf{1}_s^\top, \mathbf{0}_{d-s}^\top)^\top$.

For the GA implementation, we use

$$\mathrm{GIC}(u) = n \log \widehat{\sigma}_u^2 + 3.5|u| \log d, \tag{5.1}$$

to evaluate models. This choice of $\kappa_n = 3.5 \log d$ makes the GIC coincide with the pseudo-likelihood information criterion [68] and the high-dimensional BIC [63]. The penalization constant 3.5 is specifically used due to the superior performance shown in the simulation studies in [63]. It should be mentioned that (5.1) works well regardless the relationship between n and d (e.g., [63, 68]). Our Python implementation of the GA, the RP and the SA is publicly available in the Github repository `https://github.com/aks43725/cand`.

5.2 Schema Evolution

The discussion followed by Corollary 3.2.1 concludes that fitter schema H leads to larger $\alpha_{sel}(H, t)$ (the probability of selecting a model matching H in the t-th generation) and hence larger $\mathsf{E}\big[m(H, t+1)\big]$ (the expected number of models matching H in the $(t+1)$-th generation). In the following we provide empirical evidence by observing that $m(H, t+1)$ aligns with $\alpha_{sel}(H, t)$ for three schemata:

$$H^1 = (\mathbf{1}_s, \mathbf{0}_{2s}, *, \ldots, *), \quad H^2 = (\mathbf{1}_{s+2}, *, \ldots, *) \quad \text{and} \quad H^3 = (\mathbf{1}_{s-1}, 0, *, \ldots, *),$$

which represent good, fair and bad performing schemata, respectively. In particular, H^1 is expected to perform the best by covering good models such as the true model. The $2s$ 0's are placed to deteriorate its overall performance through ruling out some

models that are too good to observe the evolution of $m(\cdot, t)$ and $\alpha_{sel}(\cdot, t)$. H^2 is expected to be slightly worse than H^1 because models matching it are all overfitting by having at least two false discoveries. We anticipate H^3 to have the worst performance due to missing one true signal. Note that $\uparrow(H^1) \cap \uparrow(H^2) \cap \uparrow(H^3)$ does not cover the whole model space $\mathcal{M}$. For implementation, we used uniform mutation as needed in the theoretical conditions. Moreover, since the GA with initial population provided by the RP is too good to observe the evolution process, we used an approach proposed in Section B.1.3 to randomly generate an initial population.

Figure 5.1 is obtained under Case 3 with $(n, d, s, \rho) = (200, 400, 6, 0.5)$ and serves as a representative example since other cases (included in supplementary, Section B.2.1) exhibits similar patterns. The upper panel confirms our performance assertion on the overall schema performance, i.e., H^1 is slightly better than H^2 and H^3 is the worst. From the lower panel, it is evident that the pattern of $m(H^1, t+1)$ aligns with that of $\alpha_{sel}(H^1, t)$ in all cases. In addition, the strong schema H^1 evolves to take over the whole population eventually even it is a minority at the beginning, and vise versa for the weaker schema H^2. On the other hand, the evolution process of H^3 illustrates a typical example that a particularly weak schema extincts soon and never rises again. In summary, a good schema expands and a weak one diminishes over generations, resulting in an improved average fitness until convergence.

5.3 Comparison with Existing Methods

In this section, we compare the GA with the RP and the SA in terms of computation time, quality of candidate model sets, and performance of multi-model inference applications such as variable importance, model confidence set and model averaging. For the RP, we collect the unique models on the regularization paths of Lasso, SCAD and MCP using the Python package **pycasso**. Recall that the GA takes the RP for the initial population. The SA is implemented to search for models of sizes appeared in the last GA generation, and the best K models are kept as the final candidate

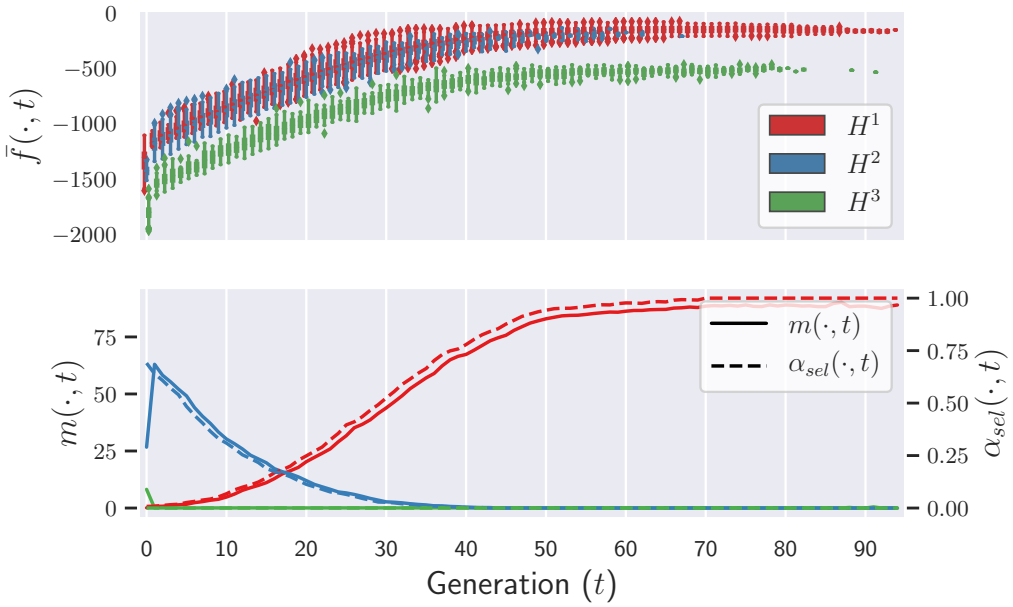

Figure 5.1.: Schema performance (upper panel) and evolution (lower panel) under Case 3 with $(n, d, s, \rho) = (200, 400, 6, 0.5)$.

model set. Other tuning parameters are settled according to the simulation settings in [14].

In the following, we show that the GA evidently improve the models generated by the RP in reasonable computation time, and that the SA takes a long time to implement but produces at most comparable results to those of the GA. In particular, the GA exhibits the best performance in all cases in terms of variable selection and quality of candidate model set. In terms of model averaging and variable importance, the GA performs at least comparably to the RP and the SA in high-dimensional cases, while just comparably under low-dimensional settings.

5.3.1 Computation Time

The averaged computation time for the three methods are depicted in Figure 5.2. It is obvious that the GA is a bit slower than the RP but way much (like more than 10 times) faster than the SA.

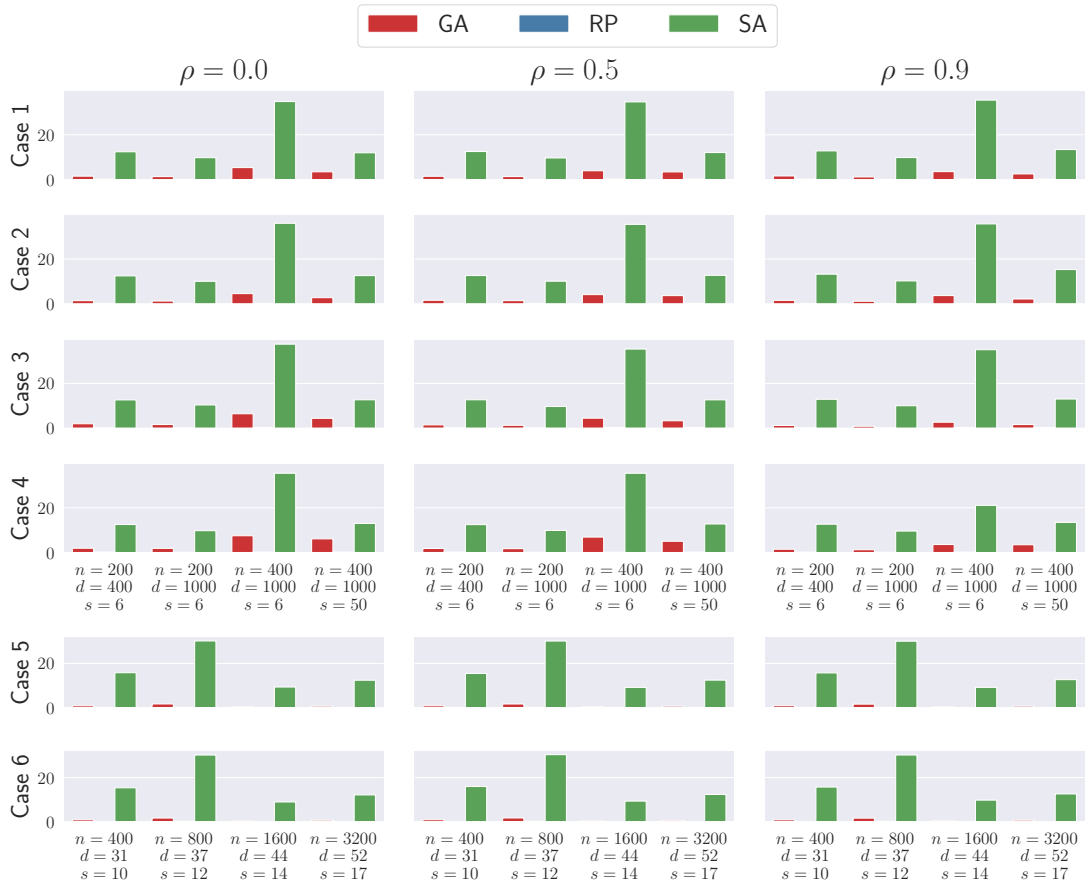

Figure 5.2.: Computation time. (The RP method is too fast to be visualized.)

5.3.2 Variable Selection

To evaluate the performance of variable selection, the boxplots of the positive selection rate (PSR, the proportion of true signals that are active in the best model) and the false discovery rate (FDR, the proportion of false signals that are active in the best model) are drawn in Figure 5.3 and Figure 5.4, respectively. We see that the GA-best model gives fairly high PSR and low FDR in all cases, demonstrating excellent variable selection performance. Under high-dimensional settings (Cases 1–4), the RP produces high PSR but also high FDR, while the SA results in the opposite (PSR and FDR are both low). For moderate dimensional cases (Cases 5 and 6), both of the

Figure 5.3.: Positive selection rate (PSR) of the best model.

RP and the SA give low PSR and FDR. In summary, the GA-best model possesses much better variable selection performance than those from the RP and the SA.

5.3.3 Quality of Candidate Models

We evaluate the quality of candidate model sets using two criteria: (i) the average fitness and (ii) the relative size of 95% SMSs (see Section 4.2 for the SMS construction) to the original candidate model set. Figure 5.5 exhibits the boxplots of average fitness and suggests that the GAs produce the fittest models in all cases. The SA takes the second place in high-dimensional cases (Cases 1–4), yet is outperformed by the RP in

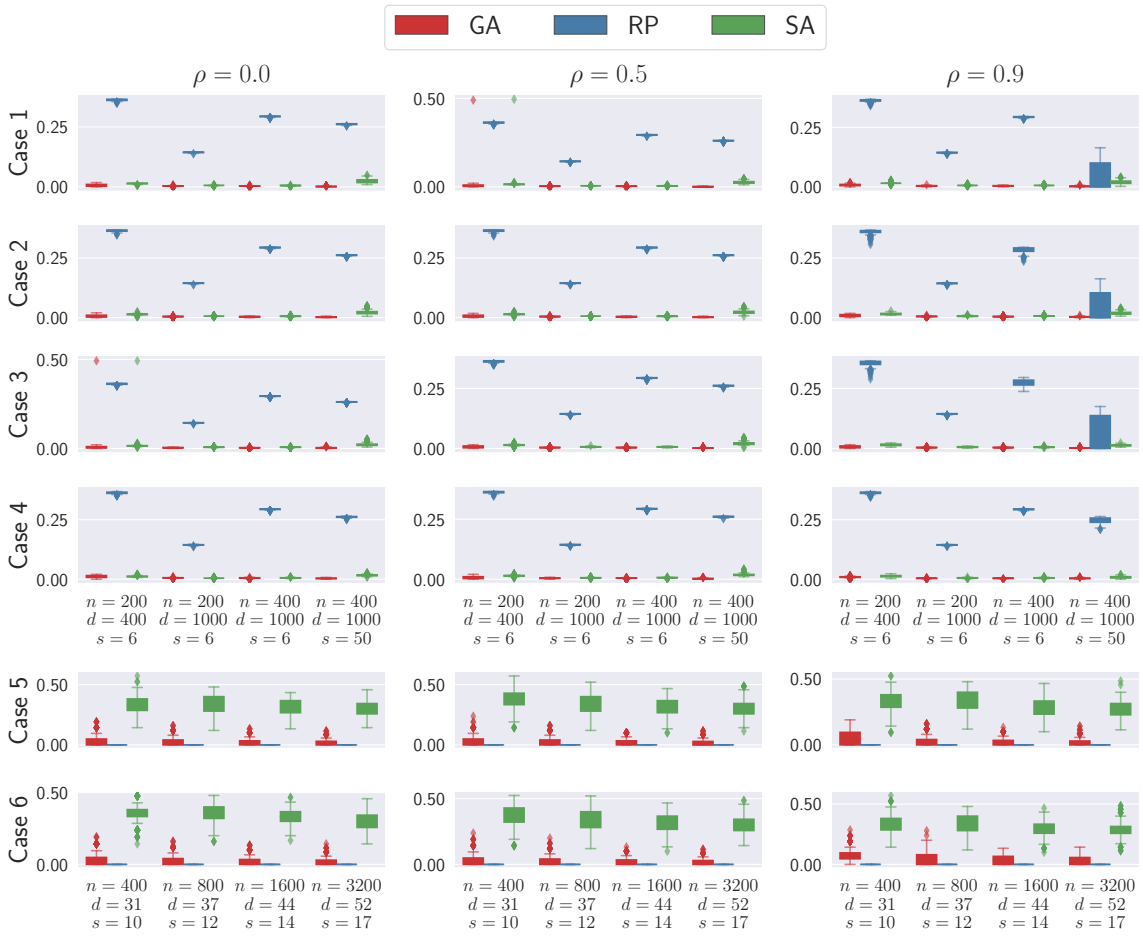

Figure 5.4.: False discovery rate (FDR) of the best model.

moderate dimensional cases (Cases 5 and 6) with $\rho = 0$ and 0.5, where the covariates are not strongly correlated. To conclude, the candidate model set produced by the GA possesses the best quality among the three approaches.

Figure 5.6 displays the boxplots of the relative size of 95% SMSs $\mathcal{A}_{0.05}$ against the original candidate model set Ψ, i.e., $|\mathcal{A}_{0.05}|/|\Psi|$, where larger values indicate better quality of Ψ. We see that the relative sizes for the GA are typically higher than those from the RP and SA in all cases, and are close to 1 in high-dimensional settings (e.g., Cases 1–4). This supports the conclusion we made about the quality of candidate models in the previous paragraph.

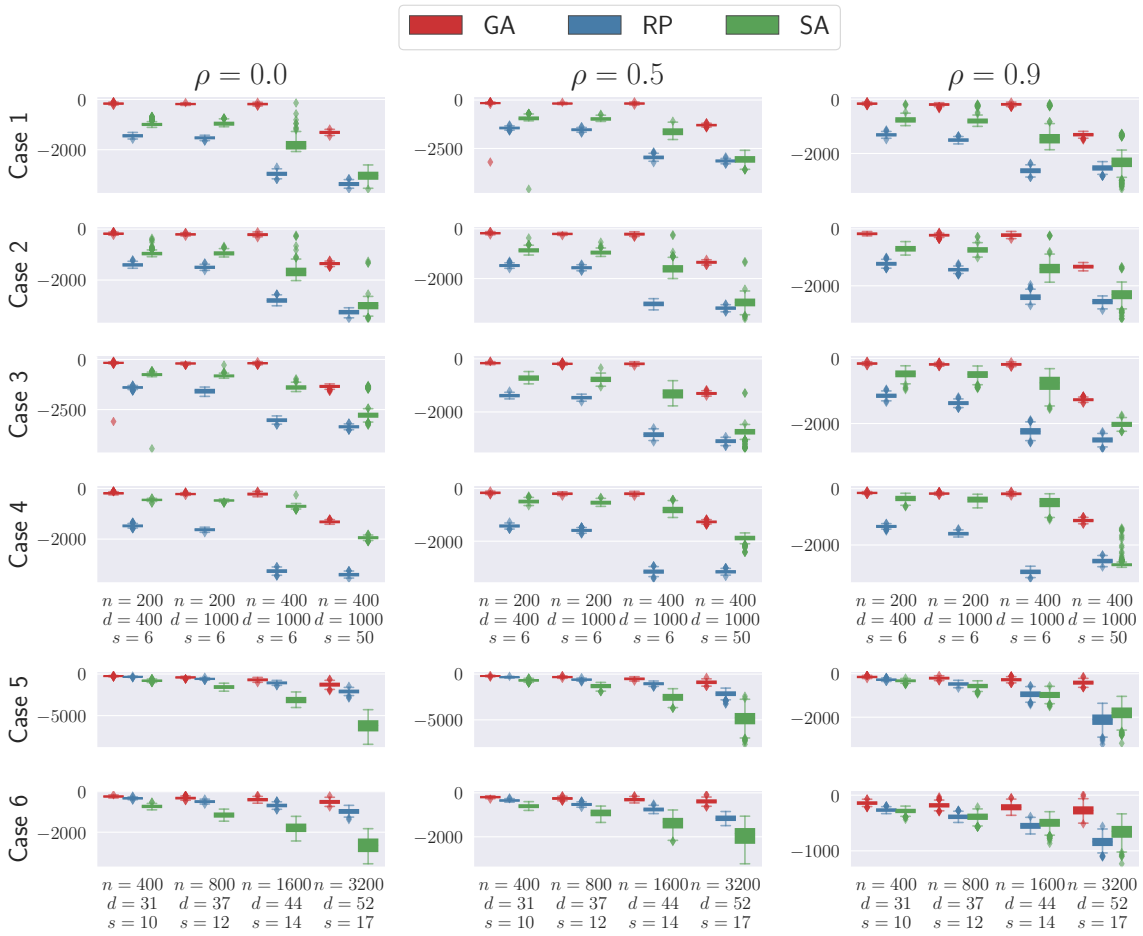

Figure 5.5.: Boxplots of the average fitness of the candidate model sets.

5.3.4 Model Averaging

Model averaging, especially in high-dimensional predictive analysis, is a prominent application of multi-model inference. The GA does not perform significantly better than the RP and the SA in model averaging, but exhibits better applicability than the RP, and greater robustness than the SA.

Given a candidate model set $\Psi = \{u^1, \ldots, u^K\}$, the model averaging predictor is defined by

$$\widehat{\boldsymbol{Y}} = \sum_{k=1}^{K} w_k \widehat{\boldsymbol{Y}}_{u^k}, \tag{5.2}$$

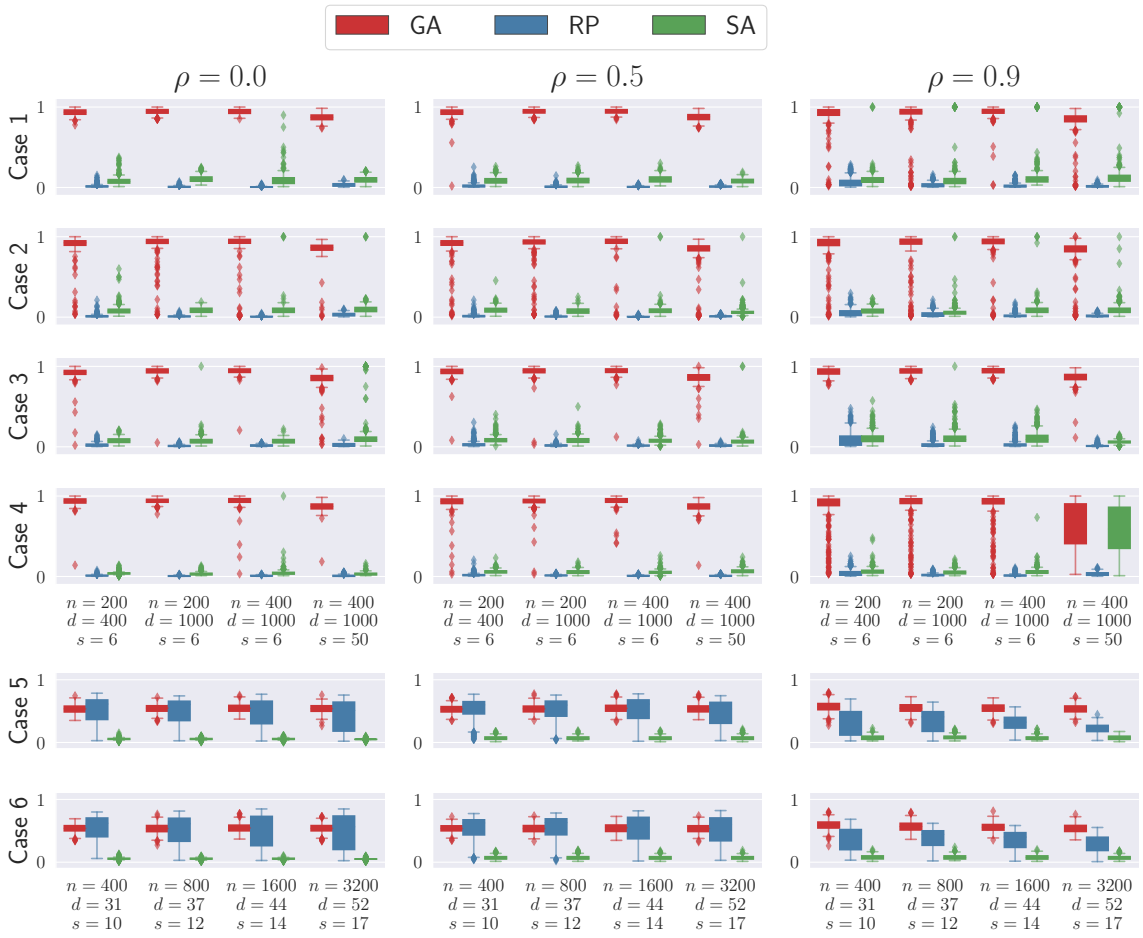

Figure 5.6.: Relative size of 95% SMS over the original candidate model set.

where $\widehat{\boldsymbol{Y}}_{u^k} = \boldsymbol{X}_u(\boldsymbol{X}_u^\top \boldsymbol{X}_u)^{-1}\boldsymbol{X}_u^\top \boldsymbol{Y}$ are the least-squares predictors and w_k with $0 \leq w_k \leq 1$ denote the model weights of u^k for $k = 1, \ldots, K$. We use the root mean squared error (RMSE) defined by

$$\sqrt{n^{-1}(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})}$$

to assess the performance of model averaging.

Two model weighting schemes are considered to obtain the model weights w_k: (i) GIC-based weights as in (2.4) with f_k replaced by $-\,\mathrm{GIC}(u^k)$, and (ii) the weighting approach proposed by [6], which we called it the "AL weighting" hereafter (see Sec-

tion B.1.2 for detailed construction). We note that (i) is the the most commonly used model weighting scheme in multi-model inference (e.g., Akaike weights [1, 33, 69, 70] and Bayesian model averaging [34]), and (ii) is developed for optimal predictive performance in high-dimensional model averaging.

Figure 5.7 displays the boxplots of the RMSE using the GIC-based model weighting, showing that the GA exhibits good and robust (in contrast to the wildly high RMSE by SA in Case 4 with $(n, d, s, \rho) = (400, 1000, 50, 0.9)$; see Remark 5.3.1 for more details) results over all cases. On the other hand, the RP is obviously worse than the GA in Case 2, and the SA's performance is just comparable to that of the GA. The three methods perform similarly in the rest cases (i.e., Cases 1, 3, 5 and 6).

The RMSEs obtained by the AL weighting are shown in Figure 5.8. Different from the results using the GIC-based model weights, the GA behaves slightly better than SA in some cases (e.g., Case 1 with $\rho = 0.0$ and 0.5 and Case 3 with $\rho = 0.0$) and comparably in the rest. Yet similarly, the GA performs robustly and the SA has wildly high RMSE in Case 4 with $(n, d, s, \rho) = (400, 1000, 50, 0.9)$. On the other hand, the results for the RP are omitted due to the computational infeasibility (inverting a singular matrix) in generating the AL weights. Accordingly, the GA is shown to possess better applicability in optimal high-dimensional model averaging.

5.3.5 Variable Importance

To evaluate the performance of high-dimensional variable importance, we employ the sparsity oriented importance learning (SOIL, [7]) defined by

$$\mathrm{SOIL}_j \equiv \mathrm{SOIL}(j; \boldsymbol{w}, \Psi) = \sum_{k=1}^{K} w_k \mathbb{1}(u_j^k = 1)$$

with the GIC-based model weights w_k given in (2.4). It can well separate the variables in the true model from the rest in the sense that SOIL_j rarely gives 0 (1) if the variable j is (not) in the true model. Moreover, it rarely gives variables not in the true model

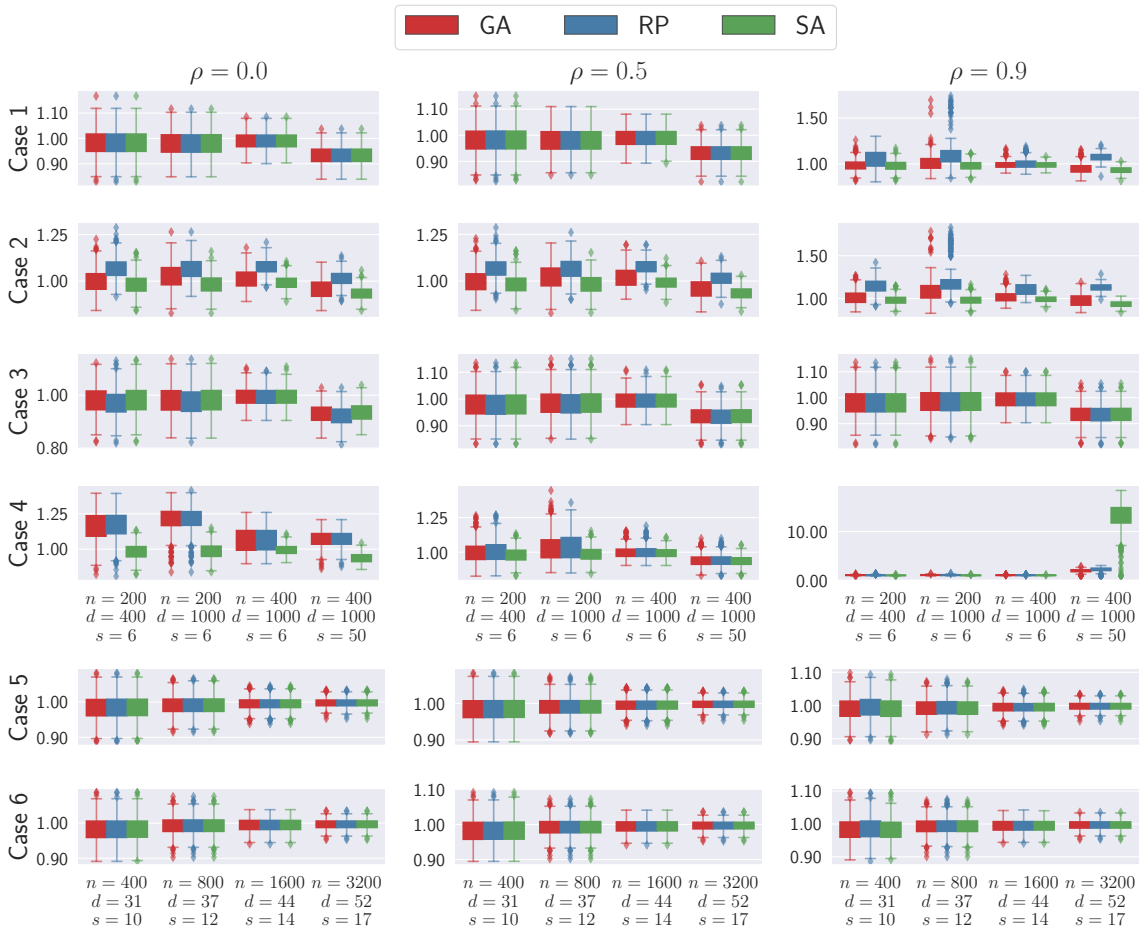

Figure 5.7.: Boxplots of the RMSE obtained by model averaging using the GIC-based weighting.

significantly higher values than those in the true model even if the signal is weak. In the original work [7], the candidate models were generated using the RP method. Our results indicate that the GA performs at least comparably to the SA and the RP in separating the true signals from the rest.

Figure 5.9 and Figure 5.10 depict the averaged SOIL values for the first $2s$ variables for Cases 2 and 4, respectively, where the active ones are before the vertical gray line and the rest are not shown due to $\text{SOIL}_j \approx 0$ for $j > 2s$ no matter which method was used for candidate model preparation. Results for Cases 1, 3, 5 and 6 are presented in supplementary (Section B.2.2) due to high similarity among the three methods.

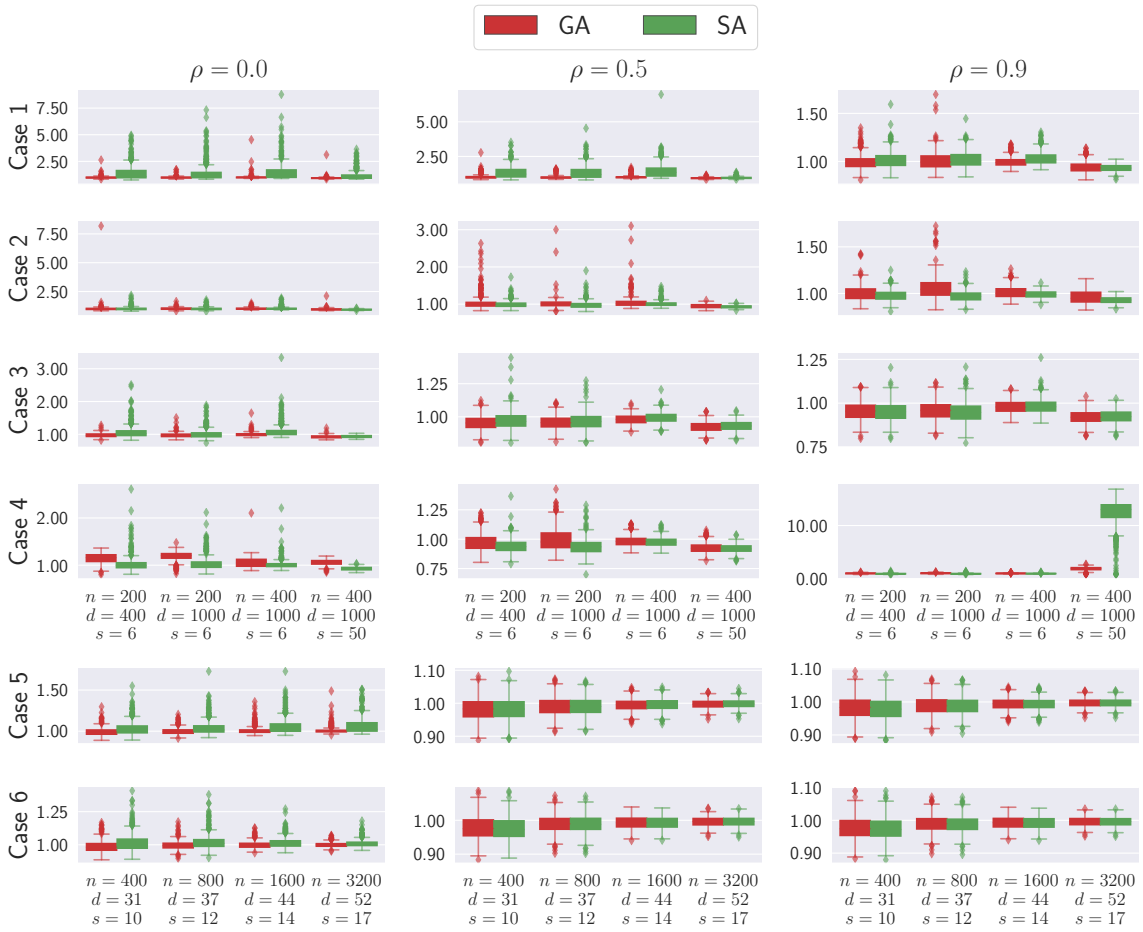

Figure 5.8.: Boxplots of the root mean squared error obtained by high-dimensional model averaging approach of [6]. The RP method fails to perform in all cases and thus is not shown.

the GA exhibits the best performance that separate the true signals from the rest. Specifically, the resulting SOIL values are by no means close to 0 and 1 for truly active and inactive variables, respectively. On the other hand, in Case 2 the RP results in $\text{SOIL}_{s-2} \equiv 1$ and $\text{SOIL}_{s+1} = 0$, where X_{s-2} is a true signal and X_{s+1} is not. Moreover, in Case 4 with $(n, d, s, \rho) = (200, 1000, 50, 0.9)$, since the SA results in $\text{SOIL}_j \leq 0.03$ for $j = 38, \ldots, 50$, these 13 true signals may easily be regarded as not important.

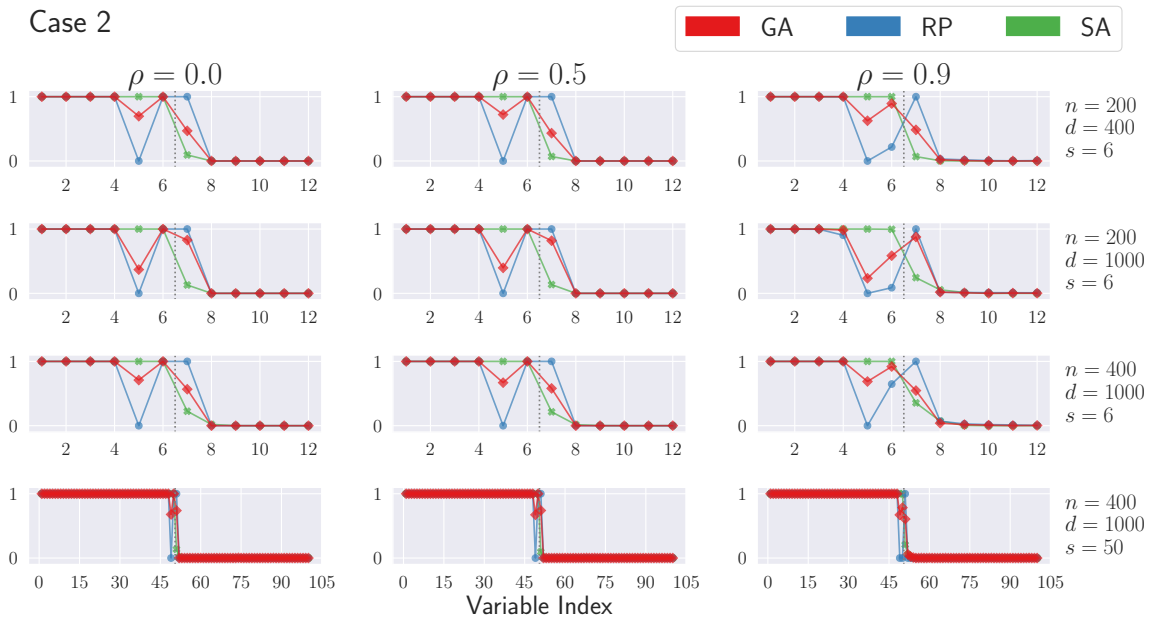

Figure 5.9.: (Case 2) Averaged SOIL measures.

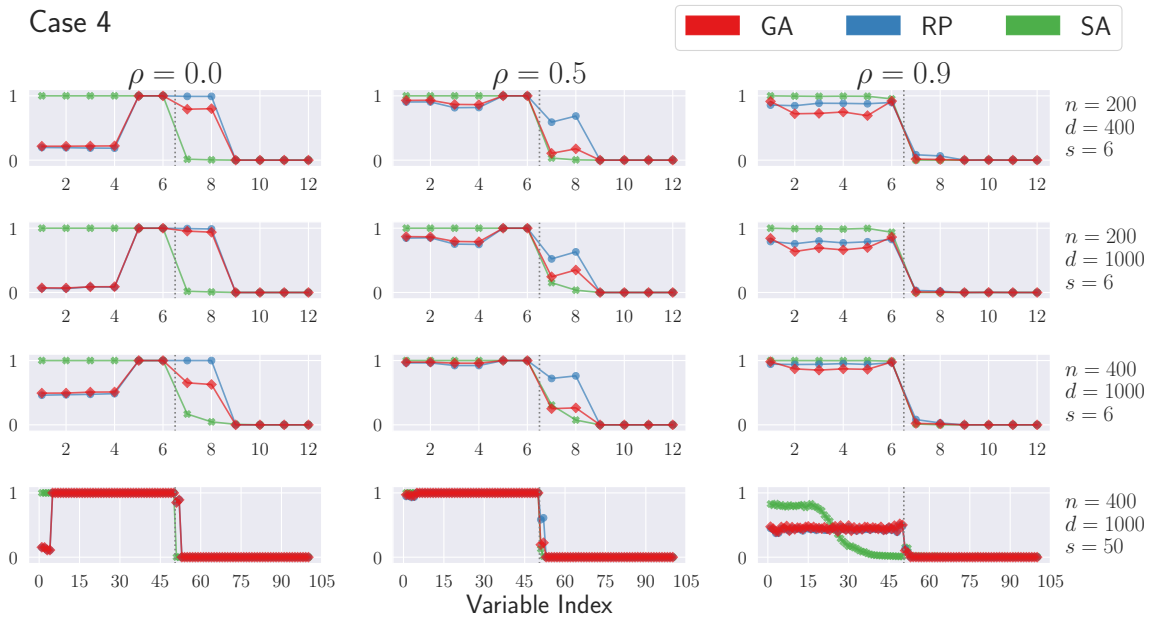

Figure 5.10.: (Case 4) Averaged SOIL measures.

Remark 5.3.1 *From Case 4 with* $(n, d, s, \rho) = (200, 1000, 50, 0.9)$*, we note that the SA's performance in model averaging and variable importance critically depends on*

the model size specification. Recall that the SA only searches for the models with sizes resulting from the GA candidate models. For this simulation case, the GA model sizes are around a half of the number of strong signals, i.e., $s/2$. Such model size misspecification causes the SA to perform poorly in model averaging and variable importance. On the other hand, the GA still behaves well even when all of its resulting candidate models miss certain number of true signals.

6. REAL DATA EXAMPLES

In this section, we present two real data examples to exhibit the usefulness of the proposed GA. Additionally, hypothesis testing (4.4) was conducted to compare models in terms of the GIC.

6.1 The Riboflavin Dataset

We first introduce the riboflavin (vitamin B) production dataset that was widely studied in high-dimensional variable selection literature (e.g., [18–22]). The response variable is the logarithm of the riboflavin production rate in Bacillus subtilis for $n = 71$ samples and the covariates are the logarithm of the expression level of $p = 4{,}088$ genes. Please see more details in Supplementary Section A.1 of [18].

The proposed GA delivers new insights by yielding better variable selection results than the existing works. From Table 6.1, the GA-best model contains only one active gene XHLA-at which was not identified by previous approaches. However, it turns out to be the fittest model (with all p-values < 0.0001) among those listed. Moreover, the importance of the gene XHLA-at is confirmed by having $\text{SOIL}_{\text{XHLA-at}} = 1$ and all other SOIL values less than 0.01. Accordingly, we suggest a further investigation on the gene XHLA-at is needed from scientists.

Table 6.2 summarizes the results of 95% SMSs (see Section 4.2), and shows the GA outperforms the RP and the SA in terms of the quality of candidate model set and model averaging. For the former, besides the much fittest (i.e., lowest GIC) model, the GA also gives the highest relative size of 95% SMSs of $56/67 = 83.58\%$ (compared to $1/54 = 1.85\%$ for the RP and $11/16 = 68.75\%$ for the SA). For model averaging, the GA results in the smallest RMSE using the GIC-based weighting. Moreover, as the only method leading to successful AL weighting (see Section 5.3.4) computation,

Table 6.1.: Variable selection results and GIC values of the selected models for the riboflavin dataset.

Method	Active Covariates	GIC
Proposed GA	XHLA-at	−20.520
Multisplit procedure [71]†	YXLD-at	−14.357
Stability selection [72]†	YXLD-at, YOAB-at, LYSC-at	−1.431
Debiased Lasso [19]	YXLD-at, YXLE-at	15.643
B-TREX [20]	YXLD-at, YOAB-at, YXLE-at	10.624
AV_∞ [21]	YXLD-at, YOAB-at, YEBC-at, ARGF-at, XHLB-at	−5.681
RP, SA, and Ridge-type projection [73]†	None	−11.775

†Obtained by [18] using the R package **hdi**.

Table 6.2.: Results of the relative size of 95% SMSs and model averaging for the riboflavin dataset.

			RMSE of Model Averaging	
Method	#(Candidate Models)	#(Models in 95% SMS)	GIC-based	AL
GA	67	56	0.6941	0.6162
RP	54	1	0.9139	N/A
SA	16	11	0.9139	N/A

the GA is shown to possess better applicability in optimal high-dimensional model averaging.

6.2 Residential Building Dataset

The second dataset was used to study $n = 372$ residential condominiums from as many 3- to 9-story buildings constructed between 1993 and 2008 in Tehran, Iran [23, 24]. Construction cost, sale price, 8 project physical and financial (PF) variables and 19 economic variables and indices (EVI) with up to 5 time lags before the construction were collected on the quarterly basis. Similar to the analysis in [24], we study how construction cost is influenced by the PF and delayed EVI factors, but exclude the only categorical PF variable, project locality. Accordingly, we have $d = 7 + 19 \times 5 = 102$ covariates. We define the variable coding in Table B.1 for the ease of presentation.

Table 6.3.: Summary of the best models for the residential building dataset.

Method	Active Variables	GIC
GA	PF-5, PF-7, EVI-05-Lag1, EVI-07-Lag4, EVI-12-Lag5, EVI-13-Lag4	2571.49
RP	(None)	3788.05
SA	PF-2, PF-3, PF-4, PF-5, PF-6, PF-7	2699.63

Table 6.4.: SOIL values of the important variables for the residential building dataset. SOIL values less than 0.05 are not listed.

	SOIL		
Variable Code	GA	RP	SA
PF-2			1.000
PF-3			1.000
PF-4			1.000
PF-5	1.000		1.000
PF-6			1.000
PF-7	1.000		1.000
EVI-05-Lag1	1.000		
EVI-07-Lag4	1.000		
EVI-12-Lag5	1.000		
EVI-13-Lag4	1.000		
EVI-19-Lag1		1.000	

Table 6.3 and Table 6.4 respectively summarize the variable selection and variable importance results of the GA, the RP and the SA. From the former, we see that the GA-best model gives the best performance (i.e., lowest GIC), and its variable structure agrees with the findings by [24], which suggest that PF and EVI factors (especially 4-quarter delayed ones) be informative. Moreover, the second column in Table 6.4 confirms the relevance of PF-5, PF-7, 1-quarter delayed EVI-05, 4-quarter delayed EVI-07 and EVI-13, and 5-quarter delayed EVI-12. We also note that the RP- and SA-best models do not consist of sensible variable structures and are significantly worse than the GA-best model (p-values < 0.0001).

Figure 6.1 and Table 6.5 respectively display the boxplots of the fitness values of the candidate models and the multi-model analysis results to evaluate the quality of candidate model sets and model averaging. The former (Figure 6.1) suggests that the

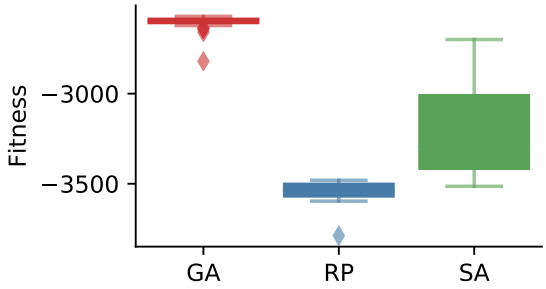

Figure 6.1.: Boxplots of the fitness values of the candidate models for the residential building dataset.

Table 6.5.: Results of relative size of 95% SMSs and model averaging for the residential building dataset.

			RMSE of Model Averaging	
Method	#(Candidate Models)	#(Models in 95% SMS)	GIC-based	AL
GA	48	41	27.5553	28.4411
RP	11	3	104.9914	N/A
SA	84	13	32.7367	32.2841

GA models generally possess higher fitness (i.e., lower GIC) values. Again, the GA is shown to produce the best candidate model set by having the fittest best model (all p-values < 0.0001) and the highest relative size of 95% SMS of $41/48 = 85.41\%$ (compared to approximately 14% for the RP and the SA). In addition to generating the best candidate model set, the GA also results in the lowest RMSE of model averaging using both the GIC-based and AL weighting methods. These results suggest that good candidate models be helpful in enhancing the performance of multi-model inference.

To further investigate the predictive performance via model averaging with the AL weighting, we randomly split the dataset using five ratios of validation to training (RVTs) of $10\%, 20\%, 30\%, 40\%$ and 50%. For each RVT, 100 randomly selected validation and training datasets were generated by splitting the original dataset, and the boxplots of RMSE are drawn in Figure 6.2. In summary, the GA generally results in lower RMSE, suggesting its superior predictive performance.

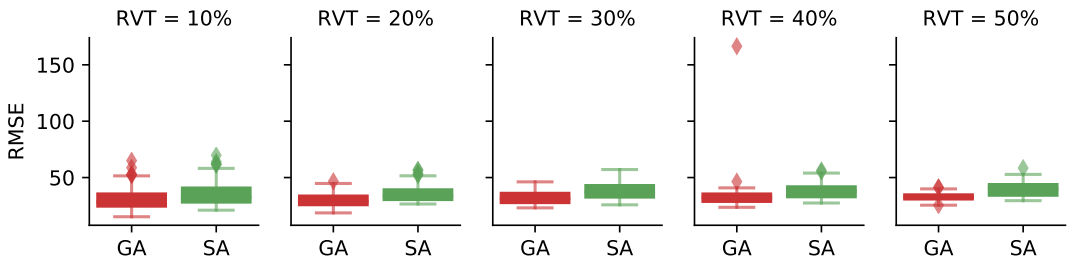

Figure 6.2.: Boxplots of RMSE of model averaging using the AL weighting for the residential building dataset. The RP method failed in weight calculations in all cases and therefore is not shown.

7. DISCUSSION

In the end, we propose three future directions. Firstly, we are interested in developing more implementable algorithms for Theorem 3.1.2 to construct the proposed MCS procedure. Secondly, we believe that incorporating GAs into modern computational tools such as neural networks may produce more powerful statistical inference procedures. For instance, the *deep neuroevolution* developed by the Uber AI Labs uses GAs to train deep reinforcement learning (DRL) models and demonstrates amazing performance on hard DRL benchmarks such as Atari and Humanoid Locomotion (e.g., [74–76]); see `https://eng.uber.com/deep-neuroevolution/` for a comprehensive introduction. Lastly, we want to investigate more advanced GA variants (e.g., adaptive GAs (e.g., [77–80]), the immune GAs (e.g., [32, 81, 82]) or the hybrid GAs (e.g., [83–86])) from statistical and machine learning perspectives.

REFERENCES

[1] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag New York, 2004.

[2] David R. Anderson. *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer-Verlag New York, 2008.

[3] J. Lavou and P. O. Droz. Multimodel inference and multimodel averaging in empirical modeling of occupational exposure levels. *The Annals of Occupational Hygiene*, 53(2):173–180, 2009.

[4] Peter R. Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011.

[5] Bruce E. Hansen. Model averaging, asymptotic risk, and regressor groups. *Quantitative Economics*, 5(3):495–530, 2014.

[6] Tomohiro Ando and Ker-Chau Li. A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109(505):254–265, 2014.

[7] Chenglong Ye, Yi Yang, and Yuhong Yang. Sparsity oriented importance learning for high-dimensional linear regression. *Journal of the American Statistical Association*, 113(524):1797–1812, 2018.

[8] John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. University of Michigan Press, Ann Arbor, MI, 1975.

[9] Melanie Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA, USA, 1996.

[10] Darrell Whitley. A genetic algorithm tutorial. *Statistics and Computing*, 4(2):65–85, 1994.

[11] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288, 1996.

[12] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[13] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[14] Daniel Nevo and Ya'acov Ritov. Identifying a minimal class of models for high-dimensional data. *Journal of Machine Learning Research*, 18(24):1–29, 2017.

[15] Wei Lan, Yingying Ma, Junlong Zhao, Hansheng Wang, and Chih-Ling Tsai. Sequential model averaging for high dimensional linear regression models. *Statistica Sinica*, 28:449–469, 2018.

[16] Ryan J. Tibshirani. A general framework for fast stagewise algorithms. *Journal of Machine Learning Research*, 16:2543–2588, 2015.

[17] Trevor Hastie, Robert Tibshirani, and Ryan J. Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the Lasso. arXiv preprint, 2017.

[18] Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278, 2014.

[19] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

[20] Johannes Lederer and Christian L. Muller. Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2729–2735. AAAI Press, 2015.

[21] Michael Chichignoud, Johannes Lederer, and Martin J. Wainwright. A practical scheme and fast algorithm to tune the Lasso with optimality guarantees. *Journal of Machine Learning Research*, 17(231):1–20, 2016.

[22] Haileab Hilafu and Xiangrong Yin. Sufficient dimension reduction and variable selection for large-p-small-n data with highly correlated predictors. *Journal of Computational and Graphical Statistics*, 26(1):26–34, 2017.

[23] Mohammad Hossein Rafiei and Hojjat Adeli. A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management*, 142(2):04015066, 2016.

[24] Mohammad Hossein Rafiei and Hojjat Adeli. Novel machine-learning model for estimating construction costs considering economic variables and indexes. *Journal of Construction Engineering and Management*, 144(12):04018106, 2018.

[25] Ryuei Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12(2):758–765, 1984.

[26] Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 25:221–264, 1997.

[27] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In Frigyes Csáki Nikolaevich Petrov, editor, *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2–8, 1971*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[28] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[29] Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009.

[30] Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

[31] Adam Chehouri, Rafic Younes, Jean Perron, and Adrian Ilinca. A constraint-handling technique for genetic algorithms using a violation factor. *Journal of Computer Science*, 12(7):350–362, 2016.

[32] Yalong Zhang, Hisakazu Ogura, Xuan Ma, Jousuke Kuroiwa, and Tomohiro Odaka. A genetic algorithm using infeasible solutions for constrained optimization problems. *The Open Cybernetics & Systemics Journal*, 8:904–912, 2014.

[33] Eric-Jan Wagenmakers and Simon Farrell. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1):192–196, 2004.

[34] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T.Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

[35] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[36] Xiangyu Wang and Chenlei Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):589–611, 2016.

[37] David Murrugarra, Jacob Miller, and Alex N. Mueller. Estimating propensity parameters using Google PageRank and genetic algorithms. *Frontiers in Neuroscience*, 10:513, 2016.

[38] Alexander Aue, Rex C. Y. Cheung, Thomas C. M. Lee, and Ming Zhong. Segmented model selection in quantile regression using the minimum description length principle. *Journal of the American Statistical Association*, 109(507):1241–1256, 2014.

[39] V. K. Koumousis and C. P. Katsaras. A saw-tooth genetic algorithm combining the effects of variable population size and reinitialization to enhance performance. *IEEE Transactions on Evolutionary Computation*, 10(1):19–28, 2006.

[40] Alan Piszcz and Terence Soule. Genetic programming: Optimal population sizes for varying complexity problems. In *Proceedings of the Genetic and Evolutionary Computation Conference, 2006*, pages 953–954, 2006.

[41] Fernando G. Lobo and Cláudio F. Lima. A review of adaptive population sizing schemes in genetic algorithms. In *Proceedings of the 7th Annual Workshop on Genetic and Evolutionary Computation*, GECCO '05, pages 228–234, New York, NY, USA, 2005. ACM.

[42] Colin R. Reeves, editor. *Modern Heuristic Techniques for Combinatorial Problems*. John Wiley & Sons, Inc., New York, NY, USA, 1993.

[43] Henrik Höglund. Tax payment default prediction using genetic algorithm-based variable selection. *Expert Systems with Applications*, 88:368–375, 2017.

[44] Samad Jafar-Zanjani, Sandeep Inampudi, and Hossein Mosallaei. Adaptive genetic algorithm for optical metasurfaces design. *Scientific Reports*, 8(1):11040, 2018.

[45] Günter Rudolph. Convergence analysis of canonical genetic algorithms. *IEEE Transactions on Neural Networks*, 5(1):96–101, 1994.

[46] Chang C. Y. Dorea, Judinor A. Guerra Jr., Rafael Morgado, and Andre G. C. Pereira. Multistage Markov chain modeling of the genetic algorithm and convergence results. *Numerical Functional Analysis and Optimization*, 31(2):164–171, 2010.

[47] Alexandru Agapie. Genetic algorithms: Minimal conditions for convergence. In Jin-Kao Hao, Evelyne Lutton, Edmund Ronald, Marc Schoenauer, and Dominique Snyers, editors, *Artificial Evolution: Third European Conference AE '97 Nîmes, France, October 22–24, 1997 Selected Papers*, pages 181–193. Springer Berlin Heidelberg, 1998.

[48] Yongdai Kim and Sunghoon Kwon. Global optimality of nonconvex penalized estimators. *Biometrika*, 99(2):315–325, 2012.

[49] Lan Wang, Yongdai Kim, and Runze Li. Calibrating nonconvex penalized regression in ultra-high dimension. *The Annals of Statistics*, 41(5):2505–2536, 2013.

[50] A. Shukla, H. M. Pandey, and D. Mehrotra. Comparative review of selection techniques in genetic algorithm. In *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, pages 515–519, 2015.

[51] Ahmad B. A. Hassanat and Esra'a Alkafaween. On enhancing genetic algorithms using new crossovers. arXiv preprint, 2018.

[52] Ahmad B. A. Hassanat, Esra'a Alkafaween, Nedal A. Al-Nawaiseh, Mohammad A. Abbadi, Mouhammd Alkasassbeh, and Mahmoud B. Alhasanat. Enhancing genetic algorithms using multi mutations. arXiv preprint, 2018.

[53] Sara van de Geer, Peter Bühlmann, YaâĂŹacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[54] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[55] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

[56] Michael D. Vose. Modeling simple genetic algorithms. In L. Darrell Whitley, editor, *Foundations of Genetic Algorithms*, volume 2 of *Foundations of Genetic Algorithms*, pages 63–73. Elsevier, 1993.

[57] Riccardo Poli. Recursive conditional schema theorem, convergence and population sizing in genetic algorithms. In Worthy N. Martin and William M. Spears, editors, *Foundations of Genetic Algorithms 6*, pages 143–163. Morgan Kaufmann, San Francisco, 2001.

[58] Riccardo Poli. Exact schema theory for genetic programming and variable-length genetic algorithms with one-point crossover. *Genetic Programming and Evolvable Machines*, 2(2):123–163, 2001.

[59] Liang Ming, Yu-Ping Wang, and Yu ming Cheung. A new schema theorem for uniform crossover based on ternary representation. In *Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference*, pages 235–239, 2004.

[60] Jack McKay Fletcher and Thomas Wennekers. A natural approach to studying schema processing. arXiv preprint, 2017.

[61] Riccardo Poli, W. B. Langdon, and U.-M. O'Reilly. Analysis of schema variance and short term extinction likelihoods. In *Genetic Programming: Proceedings of the Third Annual Conference*, pages 284–292. Morgan Kaufmann, 22–25 July 1998.

[62] Yongdai Kim, Sunghoon Kwon, and Hosik Choi. Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13:1037–1057, 2012.

[63] Tao Wang and Lixing Zhu. Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, 102(7):1141–1151, 2011.

[64] Davide Ferrari and Yuhong Yang. Confidence sets for model selection by F-testing. *Statistica Sinica*, 25:1637–1658, 2015.

[65] Chao Zheng, Davide Ferrari, and Yuhong Yang. Model selection confidence sets by likelihood ratio testing. *Statistica Sinica*, 2018+. To appear.

[66] Quang H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989.

[67] Edgar Merkle and Dongjun You. **nonnest2***: Tests of non-nested models*, 2018. R package version 0.5-1.

[68] Xin Gao and Raymond J. Carroll. Data integration with high dimensionality. *Biometrika*, 104(2):251–272, 2017.

[69] Hirotugu Akaike. On the likelihood of a time series model. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 27(3/4):217–235, 1978.

[70] Hamparsum Bozdogan. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[71] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. p-Values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[72] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[73] Peter Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 09 2013.

[74] Felipe Petroski Such, Vashisht Madhavan, Edoardo Conti, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. arXiv preprint, 2018.

[75] Xingwen Zhang, Jeff Clune, and Kenneth O. Stanley. On the relationship between the OpenAI evolution strategy and stochastic gradient descent. arXiv preprint, 2017.

[76] Edoardo Conti, Vashisht Madhavan, Felipe Petroski Such, Joel Lehman, Kenneth Stanley, and Jeff Clune. Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5027–5038. Curran Associates, Inc., 2018.

[77] Hongcheng Tang. An improved adaptive genetic algorithm. In Honghua Tan, editor, *Knowledge Discovery and Data Mining*, pages 717–723. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[78] Xudong Song and Yunlong Xiao. An improved adaptive genetic algorithm. In Peter Li, editor, *Proceedings of the 2013 Conference on Education Technology and Management Science (ICETMS 2013)*, pages 816–819. Atlantis Press, 2013.

[79] B. R. Rajakumar and Aloysius George. APOGA: An adaptive population pool size based genetic algorithm. *AASRI Procedia*, 4:288 – 296, 2013. 2013 AASRI Conference on Intelligent Systems and Control.

[80] G. J. LaPorte, J. Branke, and C. H. Chen. Adaptive parent population sizing in evolution strategies. *Evolutionary Computation*, 23(3):397–420, 2015.

[81] Licheng Jiao and Lei Wang. A novel genetic algorithm based on immunity. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(5):552–561, 2000.

[82] Yang Yu and Zhi-Hua Zhou. On the usefulness of infeasible solutions in evolutionary search: A theoretical study. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pages 835–840, 2008.

[83] Felix T.S. Chan, S.H. Chung, and Subhash Wadhwa. A hybrid genetic algorithm for production and distribution. *Omega*, 33(4):345–355, 2005.

[84] Yi-Tung Kao and Erwie Zahara. A hybrid genetic algorithm and particle swarm optimization for multimodal functions. *Applied Soft Computing*, 8(2):849–857, 2008.

[85] Po-Han Chen and Seyed Mohsen Shahandashti. Hybrid of genetic algorithm and simulated annealing for multiple project scheduling with multiple resource constraints. *Automation in Construction*, 18(4):434 – 443, 2009.

[86] K. Zhu, H. Song, L. Liu, J. Gao, and G. Cheng. Hybrid genetic algorithm for cloud computing applications. In *2011 IEEE Asia-Pacific Services Computing Conference*, pages 182–187, 2011.

[87] Yuhong Yang. Model selection for nonparametric regression. *Statistica Sinica*, 9:475–499, 1999.

A. PROOFS

A.1 Proofs for Section 3

A.1.1 Proof of Theorem 3.1.1

To prove (a), first note that given $u^* \in \Psi(t)$ the subsequent generations cannot travel to any state that does not contain u^* due to elitism selection. This means $\boldsymbol{P}$ is reducible, and $\mathcal{M}_{\max}$ is closed (in the sense that $P\big(u^* \notin \Psi(t') \big| u^* \in \Psi(t)\big) = 0$ for all $t' > t$).

Without loss of generality, there exists square matrices $\boldsymbol{A}$ and $\boldsymbol{T}$, and a matrix $\boldsymbol{R}$ with suitable dimensions such that

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{O} \\ \boldsymbol{R} & \boldsymbol{T} \end{bmatrix},$$

where $\boldsymbol{A}$ is a $|\mathcal{M}_{\max}| \times |\mathcal{M}_{\max}|$ transition probability submatrix corresponding to the states in $\mathcal{M}_{\max}$. According to Lemma A.3.1 (Theorem 2 of [45]), it suffices to show that $\boldsymbol{A} = [a_{\boldsymbol{uv}}]_{\boldsymbol{u},\boldsymbol{v} \in \mathcal{M}_{\max}}$ is stochastic and primitive, and $\boldsymbol{R}$ and $\boldsymbol{T}$ are not zero matrices.

To show $\boldsymbol{A}$ is stochastic and primitive, first note that $\boldsymbol{A}$ corresponds to the transition probability matrix for the states $\boldsymbol{u} \in \mathcal{M}_{\max}$. Since any $P\big(\Psi(t+1) \notin \mathcal{M}_{\max} \big| \Psi(t) \in \mathcal{M}_{\max}\big) = 0$ for any $t \geq 0$, we must have $\sum_{\boldsymbol{v} \in \mathcal{M}_{\max}} a_{\boldsymbol{uv}} = 1$. This indicates that $\boldsymbol{A}$ is stochastic.

For any fixed-size population $\boldsymbol{u}$, the child models generated by selection and crossover operations still belong to $\mathcal{M}$, and they can be transformed to any other models through the mutation operator with $\pi_m \in (0, 1)$. In other words, any model $u \in \boldsymbol{u}$ with $u \neq u^*$ can be mapped to any $v \in \mathcal{M}$. This implies any state in $\mathcal{M}_{\max}$

can travel to any other state in $\mathcal{M}_{\max}$ with positive probability. Accordingly, $\boldsymbol{A}$ is positive and thus primitive.

Similar argument yields that $P_{\boldsymbol{uu}} = P\big(\Psi(t+1) = \boldsymbol{u} \big| \Psi(t) = \boldsymbol{u}\big) > 0$ for all $\boldsymbol{u} \in \mathcal{M}^K$, and therefore $\boldsymbol{T}$, the transition probability matrix corresponding to the states not in $\mathcal{M}_{\max}$, is not zero. Moreover, since the generational best model can only be improved, any model u can be transformed to u^* with positive probability due to the mutation operator with $p_m \in (0,1)$. Hence for any $t \geq 0$ we have

$$P\big(\Psi(t+1) = \boldsymbol{v} \big| \Psi(t) = \boldsymbol{u}\big) > 0 \quad \text{for all } \boldsymbol{u} \notin \mathcal{M}_{\max} \text{ and } \boldsymbol{v} \in \mathcal{M}_{\max}. \tag{A.1}$$

Note that the entries of $\boldsymbol{R}$ collects all such transition probabilities. Consequently, it is a positive, and thus nonzero matrix.

The result of (b) is a straightforward consequence of (a). That is, since $\boldsymbol{\pi}$ is a distribution over $\mathcal{M}^K$ and $\pi(\boldsymbol{u}) = 0$ for all $\boldsymbol{u} \notin \mathcal{M}_{\max}$, we have $\sum_{\boldsymbol{u} \in \mathcal{M}_{\max}} \pi(\boldsymbol{u}) = 1$. By the definition of $\mathcal{M}_{\max}$, it further implies the asymptotic inclusion of the best model as $t \to \infty$.

A.1.2 Proof of Theorem 3.1.2

It suffices to show that

$$P\big(\Psi(T_\alpha) \in \mathcal{M}_{\max}\big) \geq 1 - \alpha. \tag{A.2}$$

Since the GA with elitism selection satisfies

$$\Big\{\Psi(t) \in \mathcal{M}_{\max}\Big\} \subset \Big\{\Psi(t+1) \in \mathcal{M}_{\max}\Big\} \quad \text{for all } t \geq 0,$$

it suffices to show that there exists a positive integer T_α such that

$$P\left(\bigcup_{t=1}^{T_\alpha} \Big\{\Psi(t) \in \mathcal{M}_{\max}\Big\} \,\middle|\, \Psi(0) = \boldsymbol{u}\right) \geq 1 - \alpha \quad \text{for any } \boldsymbol{u} \in \mathcal{M}^K. \tag{A.3}$$

Let

$$P_{\boldsymbol{u}\mathcal{M}_{\max}} = \sum_{\boldsymbol{v}\in\mathcal{M}_{\max}} P\big(\Psi(t+1)=\boldsymbol{v}\big|\Psi(t)=\boldsymbol{u}\big)$$

denotes the total probability that a population $\boldsymbol{u}$ is transmitted into any population with the best solution in one iteration. According to (A.1), define

$$\xi := \inf_{\boldsymbol{u}\in\mathcal{M}^K} P_{\boldsymbol{u}\mathcal{M}_{\max}} = \inf_{\boldsymbol{u}\in\mathcal{M}^K} \sum_{\boldsymbol{v}\in\mathcal{M}_{\max}} P\big(\Psi(t+1)=\boldsymbol{v}\big|\Psi(t)=\boldsymbol{u}\big) > 0. \tag{A.4}$$

Note that, for all $\boldsymbol{u}\in\mathcal{M}^K$ and positive integer t,

$$1-\xi \geq P\Big(\Psi(t)\notin\mathcal{M}_{\max}\,\Big|\,\Psi(0)=\boldsymbol{u}\Big) = \mathsf{E}\Big[\mathbb{1}\big(\Psi(t)\notin\mathcal{M}_{\max}\,\big|\,\Psi(0)=\boldsymbol{u}\big)\Big].$$

It then holds, for any $\boldsymbol{u}\in\mathcal{M}^K$ and positive integer T,

$$\begin{aligned}
&P\left(\bigcap_{t=1}^{T}\Big\{\Psi(t)\notin\mathcal{M}_{\max}\Big\}\,\middle|\,\Psi(0)=\boldsymbol{u}\right)\\
&\quad=\mathsf{E}\left[\mathbb{1}\left(\bigcap_{t=1}^{T}\Big\{\Psi(t)\notin\mathcal{M}_{\max}\Big\}\right)\middle|\,\Psi(0)=\boldsymbol{u}\right]\\
&\quad=\mathsf{E}\left[\prod_{t=1}^{T}\mathbb{1}\big(\Psi(t)\notin\mathcal{M}_{\max}\big)\middle|\,\Psi(0)=\boldsymbol{u}\right]\\
&\quad=\mathsf{E}\left[\mathsf{E}\Big[\mathbb{1}\big(\Psi(T)\notin\mathcal{M}_{\max}\big)\,\Big|\,\Psi(T-1)\Big]\prod_{t=1}^{T-1}\mathbb{1}\big(\Psi(t)\notin\mathcal{M}_{\max}\big)\middle|\,\Psi(0)=\boldsymbol{u}\right]\\
&\quad\leq(1-\xi)\mathsf{E}\left[\prod_{t=1}^{T-1}\mathbb{1}\big(\Psi(t)\notin\mathcal{M}_{\max}\big)\middle|\,\Psi(0)=\boldsymbol{u}\right],
\end{aligned}$$

where the third equality is due to the Markov property. By keeping doing this we obtain

$$P\left(\bigcap_{t=1}^{T}\Big\{\Psi(t)\notin\mathcal{M}_{\max}\Big\}\,\middle|\,\Psi(0)=\boldsymbol{u}\right)\leq(1-\xi)^T.$$

Since $\xi \in (0,1)$, there exists a positive integer T_α such that $(1-\xi)^{T_\alpha} \leq \alpha < (1-\xi)^{T_\alpha - 1}$. Accordingly, the desired confidence statement (A.3) follows.

A.1.3 Proof of Theorem 3.2.1

We first characterize the individual probabilities caused by the selection, crossover and mutation operations. Firstly, it is obvious that the probability that models u^k and u^l are selected is $w_k w_l$. Secondly, the probability that the uniform mutation operation transforms a given model v into a solution that matches H is $\pi_m^{\delta(v,H)}(1-\pi_m)^{\mathrm{ord}(H)-\delta(v,H)}$.

Finally, we discuss the effect of the uniform crossover operation, given two parent models u^k and u^l are selected. Due to the mechanism of the uniform crossover, all possible child models has equal probabilities to be generated. This allows us to focus on the fixed positions of H. Note that it is possible that u^k and u^l can never generate a child model that is a solution that matches H. Therefore, we define

$$h_{kl} = \left|\{j : H_j \neq *, u_j^k = u_j^l \neq H_j\}\right|$$

as the minimum $\delta(v,H)$ among all the child models v produced by the uniform crossover with parent models u^k and u^l. Now, suppose v is a model generated through uniform crossover with u^k and u^l, we have

$$P\big(\delta_H(v,H) = h + h_{kl} \,\big|\, \text{parent models } u^k, u^l\big) = \frac{\binom{\delta_H(u^k,u^l)-h_{kl}}{h}}{2^{\delta_H(u^k,u^l)}} \quad \text{for } h = 0,1,\ldots,\delta_H(u^k,u^l) - h_{kl}.$$

Accordingly, a general form of $\alpha(H,t)$ can be written by

$$\begin{aligned}
\alpha(H,t) &= \sum_{k,l:u^k,u^l\in\Psi(t)} w_k w_l \left[\sum_{h=0}^{\delta_H(u^k,u^l)-h_{kl}} \frac{\binom{\delta_H(u^k,u^l)-h_{kl}}{h}}{2^{\delta_H(u^k,u^l)-h_{kl}}} \pi_m^{h+h_{kl}} (1-\pi_m)^{\mathrm{ord}(H)-h-h_{kl}}\right] \\
&= \sum_{k,l:u^k,u^l\in\Psi(t)} w_k w_l \frac{\pi_m^{h_{kl}}(1-\pi_m)^{\mathrm{ord}(H)-\delta_H(u^k,u^l)}}{2^{\delta_H(u^k,u^l)-h_{kl}}} \\
&\qquad\qquad \times \left[\sum_{h=0}^{\delta_H(u^k,u^l)-h_{kl}} \binom{\delta_H(u^k,u^l)-h_{kl}}{h} \pi_m^{h+h_{kl}} (1-\pi_m)^{\delta_H(u^k,u^l)-h-h_{kl}}\right] \\
&= \sum_{k,l:u^k,u^l\in\Psi(t)} w_k w_l \frac{\pi_m^{h_{kl}}(1-\pi_m)^{\mathrm{ord}(H)-\delta_H(u^k,u^l)}}{2^{\delta_H(u^k,u^l)-h_{kl}}} \qquad \text{(A.5)}
\end{aligned}$$

Note that on the right hand side of (A.5), the summation can be tore apart to three cases based on whether the parents are solutions that match H. That is,

$$\alpha(H,t) = P\big(\text{Case 1}\big) + P\big(\text{Case 2}\big) + P\big(\text{Case 3}\big), \qquad \text{(A.6)}$$

where Cases 1, 2 and 3 refer to the events that the final child model after crossover and mutation is a solution that matches H given that

1. both parents match H (i.e., k, such that $u^k, u^l \in \uparrow(H)$),
2. only one of the parents matches H (i.e., k such that $u^k \in \uparrow(H)$ and $l : u^l \notin \uparrow(H)$), and
3. neither of the parents matches H (i.e., k, l such that $u^k, u^l \notin \uparrow(H)$),

respectively.

For Case 1, since both parents belong to $\uparrow(H)$, it follows that $\delta_H(u^k, u^l) = 0$ and $h_{kl} = 0$, and hence

$$\begin{aligned} P\big(\text{Case 1}\big) &= \sum_{k,l:u^k,u^l\in\uparrow(H)} w_k w_l (1-\pi_m)^{\mathrm{ord}(H)} \\ &= \left(\sum_{k:u^k\in\uparrow(H)} w_k\right)^2 (1-\pi_m)^{\mathrm{ord}(H)} \\ &= \alpha_{sel}(H,t)^2 (1-\pi_m)^{\mathrm{ord}(H)}. \end{aligned} \tag{A.7}$$

For Case 2, since one of the parents matches H, it holds $h_{kl} = 0$ and $\delta_H(u^k, u^l) = \delta(u^l, H)$. It then holds that

$$\begin{aligned} P\big(\text{Case 2}\big) &= \sum_{\substack{k:u^k\in\uparrow(H)\\ l:u^l\notin\uparrow(H)}} w_k w_l \frac{(1-\pi_m)^{\mathrm{ord}(H)}}{\big[2(1-\pi_m)\big]^{\delta(u^l,H)}} \\ &= \alpha_{sel}(H,t) \sum_{l:u^l\notin\uparrow(H)} w_l \frac{(1-\pi_m)^{\mathrm{ord}(H)}}{\big[2(1-\pi_m)\big]^{\delta(u^l,H)}}. \end{aligned} \tag{A.8}$$

For Case 3, there seems no simplification available, and therefore we have

$$P\big(\text{Case 3}\big) = \sum_{k,l:u^k,u^l\notin\uparrow(H)} w_k w_l \frac{(2\pi_m)^{h_{kl}}(1-\pi_m)^{\mathrm{ord}(H)}}{\big[2(1-\pi_m)\big]^{\delta_H(u^k,u^l)}}. \tag{A.9}$$

The proof is then complete by plugging (A.7), (A.8) and (A.9) into (A.6).

A.1.4 Proof of Corollary 3.2.1

First note that $2(1-\pi_m) > 1$ since $\pi_m \leq 0.5$. Since $\delta(u^l, H) \leq \text{ord}(H)$ for all $u^l \notin \uparrow(H)$, it follows that

$$\begin{aligned} P(\text{Case 2}) &= \alpha_{sel}(H,t) \sum_{l:u^l \notin \uparrow(H)} w_l \frac{(1-\pi_m)^{\text{ord}(H)}}{\big[2(1-\pi_m)\big]^{\delta(u^l,H)}} \\ &\geq 2^{-\text{ord}(H)} \alpha_{sel}(H,t) \sum_{l:u^l \notin \uparrow(H)} w_l \\ &= 2^{-\text{ord}(H)} \alpha_{sel}(H,t)\big[1-\alpha_{sel}(H,t)\big]. \end{aligned}$$

Similarly, since $2\pi_m < 1$, $h_{kl} \leq \text{ord}(H)$ and $\delta_H(u^k, u^l) \leq \text{ord}(H)$ for all $u^k, u^l \notin \uparrow(H)$, we have

$$\begin{aligned} P(\text{Case 3}) &= \sum_{k,l:u^k,u^l \notin \uparrow(H)} w_k w_l \frac{(2\pi_m)^{h_{kl}}(1-\pi_m)^{\text{ord}(H)}}{\big[2(1-\pi_m)\big]^{\delta_H(u^k,u^l)}} \\ &\geq \pi_m^{\text{ord}(H)}\big[1-\alpha_{sel}(H,t)\big]^2. \end{aligned}$$

Accordingly, we have the desired result (3.6).

A.2 Proof for Section 4

A.2.1 Proof of Lemma 4.0.1

Without loss of generality, let u^0 denote the binary sequence with first s genes active and the rest inactive and $\sigma^2 = 1$. Recall that $\boldsymbol{X}_u$ denotes the submatrix of $\boldsymbol{X}$ subject to the active variable indices in u. Let $\boldsymbol{H}_u = \boldsymbol{X}_u(\boldsymbol{X}_u^\top \boldsymbol{X}_u)^{-1}\boldsymbol{X}_u^\top$ the projection matrix of the submatrix $\boldsymbol{X}_u$.

We first consider the case $u \not\supseteq u^0$, i.e., model u misses at least one relevant variable. We can write

$$\begin{aligned}\mathrm{GIC}(u) - \mathrm{GIC}(u^0) &= n \log\left(1 + \frac{\mathrm{RSS}(u) - \mathrm{RSS}(u^0)}{\mathrm{RSS}(u^0)}\right) + \kappa_n\big(|u| - s\big) \\ &\geq n \log\left(1 + \frac{\mathrm{RSS}(u) - \mathrm{RSS}(u^0)}{\mathrm{RSS}(u^0)}\right) - \kappa_n s.\end{aligned}$$

Note that

$$\mathrm{RSS}(u^0) = \boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H}_{u^0})\boldsymbol{Y} = \boldsymbol{\varepsilon}^\top(\boldsymbol{I} - \boldsymbol{H}_{u^0})\boldsymbol{\varepsilon} = \sum_{i=1}^{d-s} Z_i^2 = n(1 + o(1)), \tag{A.10}$$

where the Z_i are independent $\mathcal{N}(0,1)$ variables, and

$$\begin{aligned}\mathrm{RSS}(u) - \mathrm{RSS}(u^0) &= \boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H}_u)\boldsymbol{Y} - \boldsymbol{\varepsilon}^\top(\boldsymbol{I} - \boldsymbol{H}_{u^0})\boldsymbol{\varepsilon} \\ &= \boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H}_{u^0})\boldsymbol{\mu} + 2\boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H}_u)\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^\top\boldsymbol{H}_u\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^\top\boldsymbol{H}_{u^0}\boldsymbol{\varepsilon}, \quad \text{(A.11)}\end{aligned}$$

where $\boldsymbol{\mu} = \boldsymbol{X}_{u^0}\boldsymbol{\beta}^0_{u^0}$. By Condition (A2), uniformly over u with $|u| \leq \tilde{s}$, it holds

$$\min_{u \in \mathcal{M}_{\tilde{s}} - \{u^0\}} \boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H}_u)\boldsymbol{\mu} \geq C_2 n. \tag{A.12}$$

Write

$$\boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H}_u)\boldsymbol{\varepsilon} = \sqrt{\boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H}_u)\boldsymbol{\mu}}Z_u, \quad \text{where } Z_u = \frac{\boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H}_u)\boldsymbol{\varepsilon}}{\sqrt{\boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H}_u)\boldsymbol{\mu}}} \sim \mathcal{N}(0,1).$$

Note that for any model u with $|u| \leq \tilde{s}$, there exists a positive constant C_3 such that

$$P\big(|Z_u| > t\big) = C_3 \exp\left(-\frac{t^2}{2}\right).$$

By the union bound, it follows that

$$P\left(\max_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\not\supseteq u^0}|Z_u|>t\right)\le\sum_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\not\supseteq u^0}P\Big(|Z_u|>t\Big)\le 2^{\tilde{s}}C_3\exp\left(-\frac{t^2}{2}\right).$$

Let $t=\sqrt{2s\log d}$, we arrive at

$$P\left(\max_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\not\supseteq u^0}|Z_u|>t\right)\le C_3\left(\frac{2}{d}\right)^{\tilde{s}}\to 0$$

as $n\to\infty$. Accordingly,

$$\max_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\not\supseteq u^0}|Z_u|=O_P\big(\sqrt{\tilde{s}\log d}\big)=o_P\big(\sqrt{n}\big),$$

and therefore we have

$$\begin{aligned}\max_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\not\supseteq u^0}\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{H}_u)\boldsymbol{\varepsilon}&\le\sqrt{\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{H}_u)\boldsymbol{\mu}}\max_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\not\supseteq u^0}Z_u\\&=\sqrt{\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{H}_u)\boldsymbol{\mu}}\;o_P(\sqrt{n})\\&=o_P\left(\sqrt{\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{H}_u)\boldsymbol{\mu}}\right).\end{aligned}\tag{A.13}$$

Now we deal with the last two terms in (A.11). Note that we can write

$$\boldsymbol{\varepsilon}^\top\boldsymbol{H}_u\boldsymbol{\varepsilon}=\sum_{i=1}^{|u|}Z_i^2\sim\chi^2_{|u|},$$

where Z_i are some independent $\mathcal{N}(0,1)$ variables. By the union bound, it then holds

$$P\left(\max_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\not\supseteq u^0}\boldsymbol{\varepsilon}^\top\boldsymbol{H}_u\boldsymbol{\varepsilon}>t\right)\le\sum_{j=1}^{\tilde{s}}\binom{d}{j}P(\chi^2_j>t)\le d^{\tilde{s}}P(\chi^2_{\tilde{s}}>t).$$

It is east to see that (see, for example, [87])

$$P(\chi^2_{\tilde{s}}>t)\le\exp\left(-\frac{t-\tilde{s}}{2}\right)\left(\frac{t}{\tilde{s}}\right)^{\tilde{s}/2}.$$

Let $t = 3s \log d$, we arrive at

$$P\left(\max_{u \in \mathcal{M}_{\tilde{s}} - \{u^0\}, u \not\supseteq u^0} \boldsymbol{\varepsilon}^\top \boldsymbol{H}_u \boldsymbol{\varepsilon} > t\right) \leq \left(\frac{e \log d}{d}\right)^{\tilde{s}/2} \to 0$$

as $n \to \infty$. Consequently, we have

$$\max_{u \in \mathcal{M}_{\tilde{s}} - \{u^0\}, u \not\supseteq u^0} \boldsymbol{\varepsilon}^\top \boldsymbol{H}_u \boldsymbol{\varepsilon} = O_P\big(\tilde{s} \log d\big) = o_P(n). \tag{A.14}$$

Similarly,

$$\boldsymbol{\varepsilon}^\top \boldsymbol{H}_{u^0} \boldsymbol{\varepsilon} = o_P(n). \tag{A.15}$$

By (A.12), (A.13), (A.14) and (A.15), it is easy to see that $\mathrm{RSS}(u) - \mathrm{RSS}(u^0)$ is dominated by $\boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{H}_{u^0}) \boldsymbol{\mu}$. Coupled with (A.10), there is a positive constant C_4 such that

$$\log\left(1 + \frac{\mathrm{RSS}(u) - \mathrm{RSS}(u^0)}{\mathrm{RSS}(u^0)}\right) \geq \log(1 + C_4)$$

in probability. Since $\kappa_n = o(n)$, we conclude that

$$\min_{u \in \mathcal{M}_{\tilde{s}} - \{u^0\}, u \not\supseteq u^0} \mathrm{GIC}(u) - \mathrm{GIC}(u^0) \geq n \log(1 + C_4) - \kappa_n s > 0 \tag{A.16}$$

as $n \to \infty$.

Now we consider the case $u \supseteq u^0$ but $u \neq u^0$. Since $(\boldsymbol{I} - \boldsymbol{H}_u) \boldsymbol{X}_{u^0} = \boldsymbol{O}$, we have $\boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{H}_u) \boldsymbol{Y} = \boldsymbol{\varepsilon}^\top (\boldsymbol{I} - \boldsymbol{H}_u) \boldsymbol{\varepsilon}$ and

$$\mathrm{RSS}(u^0) - \mathrm{RSS}(u) = \boldsymbol{\varepsilon}^\top (\boldsymbol{H}_u - \boldsymbol{H}_{u^0}) \boldsymbol{\varepsilon} = \sum_{i=1}^{|u|-s} Z_{u,i}^2 \sim \chi^2_{|u|-s},$$

where $Z_{u,i}$ are some independent $\mathcal{N}(0,1)$ variables depending on u. By the union bound we have

$$\begin{aligned}
P\left(\min_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\supseteq u^0}\mathrm{GIC}(u)-\mathrm{GIC}(u^0)\le 0\right) &\le \sum_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\supseteq u^0} P\big(\mathrm{RSS}(u^0)-\mathrm{RSS}(u)\ge \kappa_n(|u|-s)\big)\\
&= \sum_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\supseteq u^0} P\left(\boldsymbol{\varepsilon}^\top(\boldsymbol{H}_u-\boldsymbol{H}_{u^0})\boldsymbol{\varepsilon}\ge \kappa_n(|u|-s)\right)\\
&\le \sum_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\supseteq u^0} \left[\kappa_n\exp(1-\kappa_n)\right]^{\frac{|u|-s}{2}}\\
&= \sum_{j=s+1}^{\tilde{s}}\binom{d-s}{j-s}\left[\kappa_n\exp(1-\kappa_n)\right]^{\frac{j-s}{2}}\\
&\le \sum_{m=0}^{d-s}\binom{d-s}{m}\left[\kappa_n\exp(1-\kappa_n)\right]^{\frac{m}{2}}-1\\
&= \left(1+\sqrt{\frac{e\kappa_n}{\exp\kappa_n}}\right)^{d-s}-1\to 0 \quad \text{as } n\to\infty,
\end{aligned}$$

where the second inequality follows from the sharp deviation bound on the χ^2 distribution (see Lemma 3 of [35]). Hence we have

$$\min_{u\in\mathcal{M}_{\tilde{s}}-\{u^0\},u\supseteq u^0}\mathrm{GIC}(u)-\mathrm{GIC}(u^0)>0 \tag{A.17}$$

with probability tending to 1. Accordingly, the desired result (4.1) follows from (A.16) and (A.17).

A.2.2 Proof of Proposition 4.1.1

From Lemma 4.0.1 we know that the true model u^0 is the best model in the model space $\mathcal{M}_{\tilde{s}}$ with probability tending to 1. Along with Theorem 3.1.1 (b) we have

$$\lim_{t\to\infty}\lim_{n\to\infty} P\big(u^0=u^*\in\Psi_{\tilde{s}}(t)\big)=1.$$

By the definition of $\widehat{u}(t)$, we arrive at

$$\lim_{t\to\infty}\lim_{n\to\infty} P\left(\widehat{u}(t)=u^0\right)=1.$$

This completes the proof.

A.2.3 Proof of Proposition 4.2.1

By the construction of the $\mathcal{A}_\alpha(t)$, we have

$$\lim_{n\to\infty} P\big(u\in\mathcal{A}_\alpha(t)\big)\geq 1-\alpha$$

for all $u\in\Psi_{\tilde{s}}(t)-\{\widehat{u}(t)\}$ with $H_{0,u}$ not rejected and any $t\geq 0$. Along with Proposition 4.1.1, which ensures that $\lim_{t\to\infty}\lim_{n\to\infty} P\left(\widehat{u}(t)=u^0\right)=1$, the desired result then holds.

A.3 Auxiliary Lemmas

In this section we provide technical lemmas, with a bit abuse of notations.

Lemma A.3.1 (Theorem 2 of [45]) *Let $\boldsymbol{P}$ be a $n\times n$ reducible stochastic matrix that can be decomposed into*

$$\boldsymbol{P}=\begin{bmatrix}\boldsymbol{C} & \boldsymbol{O}\\ \boldsymbol{R} & \boldsymbol{T}\end{bmatrix},$$

where $\boldsymbol{C}$ is an $m\times m$ primitive stochastic matrix with $m\leq n$ and $\boldsymbol{R}$ and $\boldsymbol{T}$ are two non-zero matrices with suitable dimensions. Then there exists an $(n-m)\times n$ positive matrix $\boldsymbol{R}_\infty$ such that

$$\boldsymbol{P}^\infty=\lim_{k\to\infty}\boldsymbol{P}^k=\lim_{k\to\infty}\begin{bmatrix}\boldsymbol{C}^k & \boldsymbol{O}\\ \sum_{i=0}^{k-1}\boldsymbol{T}^i\boldsymbol{R}\boldsymbol{C}^{k-i} & \boldsymbol{T}^k\end{bmatrix}=\begin{bmatrix}\boldsymbol{C}^\infty & \boldsymbol{O}\\ \boldsymbol{R}_\infty & \boldsymbol{O}\end{bmatrix}$$

is a stable stochastic matrix with $\boldsymbol{P}^{\infty} = \mathbf{1}\boldsymbol{\pi}^{\top}$*, where* $\mathbf{1} = (1, \ldots, 1)^{\top}$ *is the vector of* 1*'s with suitable length,* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^{\top} = \boldsymbol{\pi}_0^{\top}\boldsymbol{P}^{\infty}$ *is unique regardless of the initial distribution* $\boldsymbol{\pi}_0$*, and* $\boldsymbol{\pi}$ *satisfies* $\pi_i > 0$ *for* $i = 1, \ldots, m$ *and* $\pi_i = 0$ *for* $i = m{+}1, \ldots, n$*.*

B. SUPPLEMENTARY MATERIALS

B.1 Details of the Auxiliary Methods

B.1.1 GIC-Based Superiority Test

A natural test statistic for the GIC-based superiority test (4.6) can be derived based on the difference of the GIC values of models u and $u^{\#}$. Note that the first term in the GIC (2.3) comes from simplifying the log likelihood with Gaussian noise. That is, the general form for GIC can be written as

$$\mathrm{GIC}(u) = -2\log L(\widehat{\boldsymbol{\beta}}_u; \boldsymbol{X}, \boldsymbol{Y}) + \kappa_n |u|,$$

where $L(\boldsymbol{\beta}_u; \boldsymbol{X}, \boldsymbol{Y})$ is the likelihood function of model u evaluated at $\boldsymbol{\beta}_u$ given data $(\boldsymbol{X}, \boldsymbol{Y})$, and $\widehat{\boldsymbol{\beta}}_u = \left(\boldsymbol{X}_u^\top \boldsymbol{X}_u\right)^{-1} \boldsymbol{X}_u \boldsymbol{Y}$ for any model $u \in \mathcal{M}$ with $|u| < n$. As a result, we write

$$\mathrm{GIC}(u) - \mathrm{GIC}(u^{\#}) = \big(|u| - |u^{\#}|\big)\kappa_n - 2\log \frac{L(\widehat{\boldsymbol{\beta}}_u; \boldsymbol{X}, \boldsymbol{Y})}{L(\widehat{\boldsymbol{\beta}}_{u^{\#}}; \boldsymbol{X}, \boldsymbol{Y})}.$$

Note that the first term on the R.H.S. is merely a constant and the sampling variation comes only from the second term. When u and $u^{\#}$ are distinguishable (i.e., $H_{0,u}^{dis}$ in (4.5) is rejected), [66] showed that the normalized log likelihood ratio

$$n^{-1/2} \log \frac{L(\widehat{\boldsymbol{\beta}}_u; \boldsymbol{X}, \boldsymbol{Y})}{L(\widehat{\boldsymbol{\beta}}_{u^{\#}}; \boldsymbol{X}, \boldsymbol{Y})} \Longrightarrow \mathcal{N}(0, \omega_u^2), \tag{B.1}$$

where

$$\omega_u^2 = \mathsf{Var}\left(\log \frac{L(\boldsymbol{\beta}_u^0; \mathbf{X}, \mathsf{Y})}{L(\boldsymbol{\beta}_{u^{\#}}^0; \mathbf{X}, \mathsf{Y})}\right),$$

denotes the population variance of the log likelihood ratio of u and $u^{\#}$, $\boldsymbol{\beta}_u^0$ is the true regression coefficient under model u, and $\mathbf{X}$ and Y are the population counterparts of the design vector and the response scalar, respectively. Accordingly, under $H_{0,u}^{sup}$, the result (B.1) can be used to show that

$$n^{-1/2}\big[\,\mathrm{GIC}(u)-\mathrm{GIC}(u^{\#})\big]=n^{-1/2}\left[\big(|u|-|u^{\#}|\big)\kappa_n-2\log\frac{L(\widehat{\boldsymbol{\beta}}_u;\boldsymbol{X},\boldsymbol{Y})}{L(\widehat{\boldsymbol{\beta}}_{u^{\#}};\boldsymbol{X},\boldsymbol{Y})}\right]$$
$$\Longrightarrow\mathcal{N}(0,4\omega_u^2). \tag{B.2}$$

In practice, we plug-in a consistent estimate of ω_u^2, denoted by $\widehat{\omega}_u^2$ (see [66] for the formula), into (B.2) to perform the test. Accordingly, we reject $H_{0,u}^{\mathrm{sup}}$ if

$$\mathrm{GIC}(u)-\mathrm{GIC}(u^{\#})>2z_{1-\alpha}\widehat{\omega}_u\sqrt{n},$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of standard normal distribution, and the value of $\widehat{\omega}_u^2$ can be extracted from the R package **nonnest2** [67] when implementing the distinguishability test (4.5).

B.1.2 Model Averaging Approach of [6]

Given a candidate model set $\Psi=\{u^1,\ldots,u^K\}$, let $\boldsymbol{D}_k$ be a $n\times n$ diagonal matrix with the l-th element being $(1-h_{kl})^{-1}$, where h_{kl} is the l-th diagonal element of the hat matrix $\boldsymbol{H}_{u^k}=\boldsymbol{X}_{u^k}\big(\boldsymbol{X}_{u^k}^{\top}\boldsymbol{X}_{u^k}\big)^{-1}\boldsymbol{X}_{u^k}^{\top}$, and $\widetilde{\boldsymbol{H}}_k=\boldsymbol{D}_k(\boldsymbol{H}_{u^k}-\boldsymbol{I})+\boldsymbol{I}$. Following [6], the K-dimensional weight vector $\boldsymbol{w}=(w_1,\ldots,w_K)^{\top}$ can be computed by

$$\widehat{\boldsymbol{w}}=\underset{\boldsymbol{w}\in[0,1]^K}{\arg\min}\left(\boldsymbol{Y}^{\top}\boldsymbol{Y}-2\boldsymbol{w}^{\top}\boldsymbol{a}+\boldsymbol{w}^{\top}\boldsymbol{B}\boldsymbol{w}\right), \tag{B.3}$$

where $\boldsymbol{a}=(a_1,\ldots,a_K)^{\top}$ with $a_k=\boldsymbol{Y}^{\top}\widetilde{\boldsymbol{H}}_k\boldsymbol{Y}$, and $\boldsymbol{B}$ is a $K\times K$ matrix with the (k,j)-th element $B_{kl}=\boldsymbol{Y}^{\top}\widetilde{\boldsymbol{H}}_k^{\top}\widetilde{\boldsymbol{H}}_l\boldsymbol{Y}$. Note that the common constraint $\sum_{k=1}^K w_k=1$ for model weights does not necessarily to be imposed. In fact, [6] show that their

weighting approach leads to the smallest possible estimation error of the model averaging predictor (5.2) without the constraint.

B.1.3 A Variable Association Measure Assisted Approach for Generating the Initial Population

Given variable association measures $\gamma_j, j = 1, \ldots, d$ (e.g., the marginal correlation learning $\left|\widehat{\mathsf{Cor}}(\boldsymbol{X}_j, \boldsymbol{Y})\right|$ [35] or the HOLP $\left|\boldsymbol{X}_j(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{Y}\right|$ [36] (available only for $d \geq n$), we introduce an approach to randomly generate the initial population $\{u^0, \ldots, u^0_K\}$ for the GA as follows.

Step 1: Assign the model sizes $|u^0_k|, k = 1, \ldots, K$, by generating K independent $\mathit{HyperGeom}\big(6\min(n,d), 2\min(n,d), \min(n,d)\big)$ random variables, where $\mathit{HyperGeom}(N, M, n)$ denotes the hypergeometric distribution with the probability mass function

$$P\big(\mathit{HyperGeom}(N,M,n) = m\big) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}},$$
$$m = \min(0, n+M-N), \ldots, \min(n, M).$$

Step 2: For $k = 1, \ldots, K$, the active positions of u^0_k are determined by randomly selecting $|u^0_k|$ numbers from $[d]$ without replacement according to the probability distribution $\big\{\gamma_j / \sum_{l=1}^d \gamma_l\big\}_{j=1,\ldots,d}$.

This approach ensures the model sizes are around $\min(n,d)/3$ and never exceed $\min(n,d)$. Moreover, by making use of the variable association measures γ_j, the resulting models are likely to contains the true signals so that their performance are by no means poor.

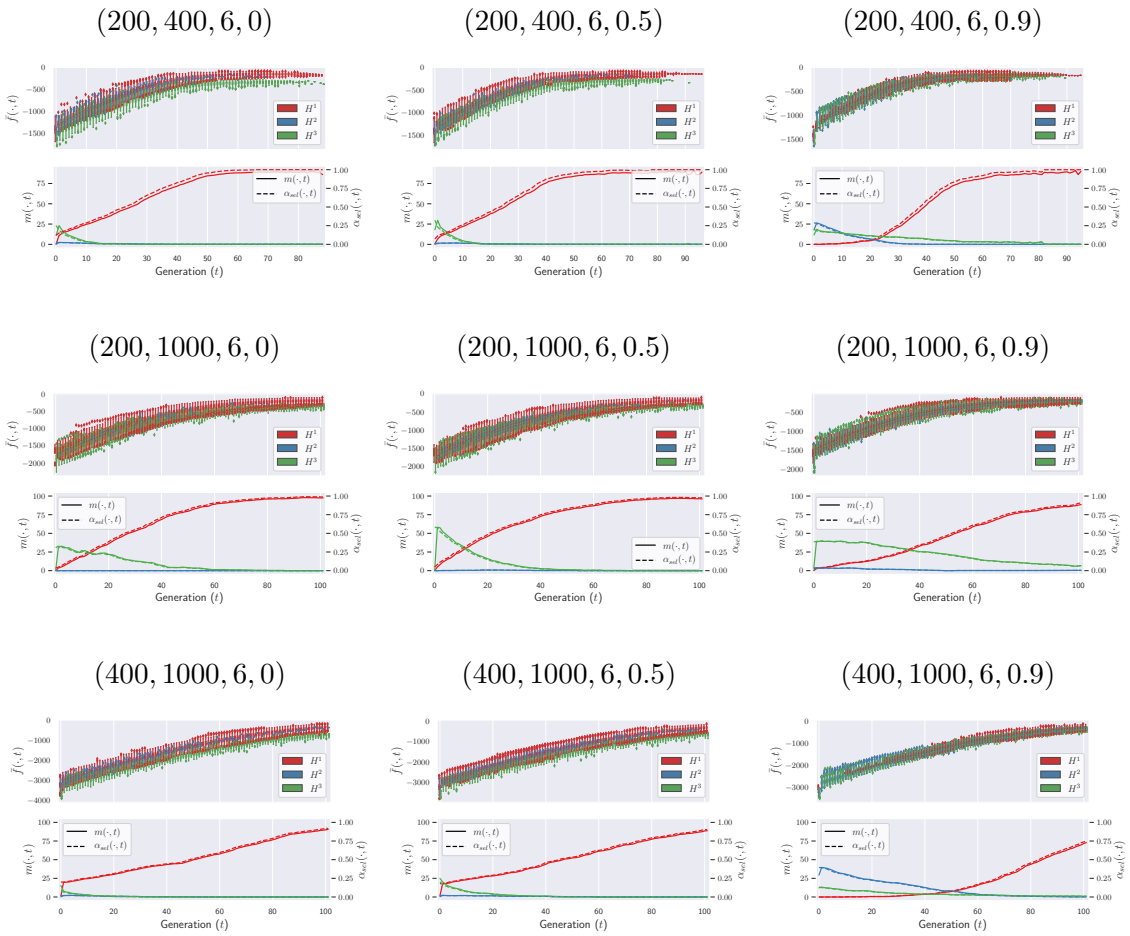

Figure B.1.: (Case 1) Schema performance (upper panel) and evolution (lower panel) obtained from 500 simulation runs. Subfigure titles indicate simulation parameter settings (n, d, s, ρ).

B.2 Supplementary Simulation Results

B.2.1 Schema Evolution

Figure B.1–B.6 present the additional results of schema evolution. The conclusions we draw in Section 5.2 still applies for these results, even though the patterns for high-dimensional (Cases 1–4) and low-dimensional (Cases 5 and 6) results are clearly different.

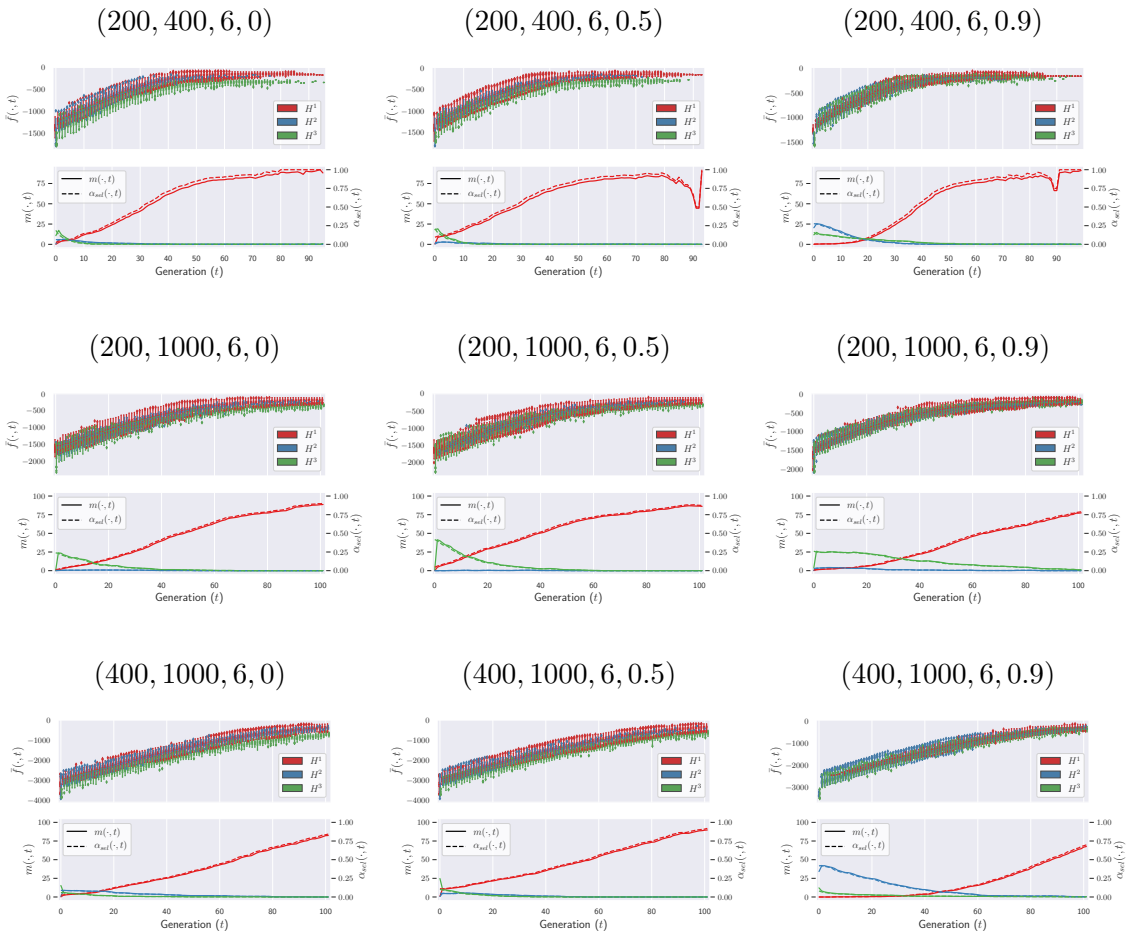

Figure B.2.: (Case 2) Schema performance (upper panel) and evolution (lower panel) obtained from 500 simulation runs. Subfigure titles indicate simulation parameter settings (n, d, s, ρ).

B.2.2 Variable Importance

Figure B.7–B.10 present additional simulation results of variable importance under Cases 1, 3, 5 and 6. In Cases 1 and 3 (Figure B.7 and Figure B.8), we see the results of GA and the SA are comparable, and slightly better than the RP in separating the true signals from the rest in some cases (e.g., X_s under Case 1 with $\rho = 0.9$ and X_{s+1} and X_{s+2} under Case 3 with $\rho = 0$). However, under Cases 5 and 6 (Figure B.9 and Figure B.10) the three methods are just comparable.

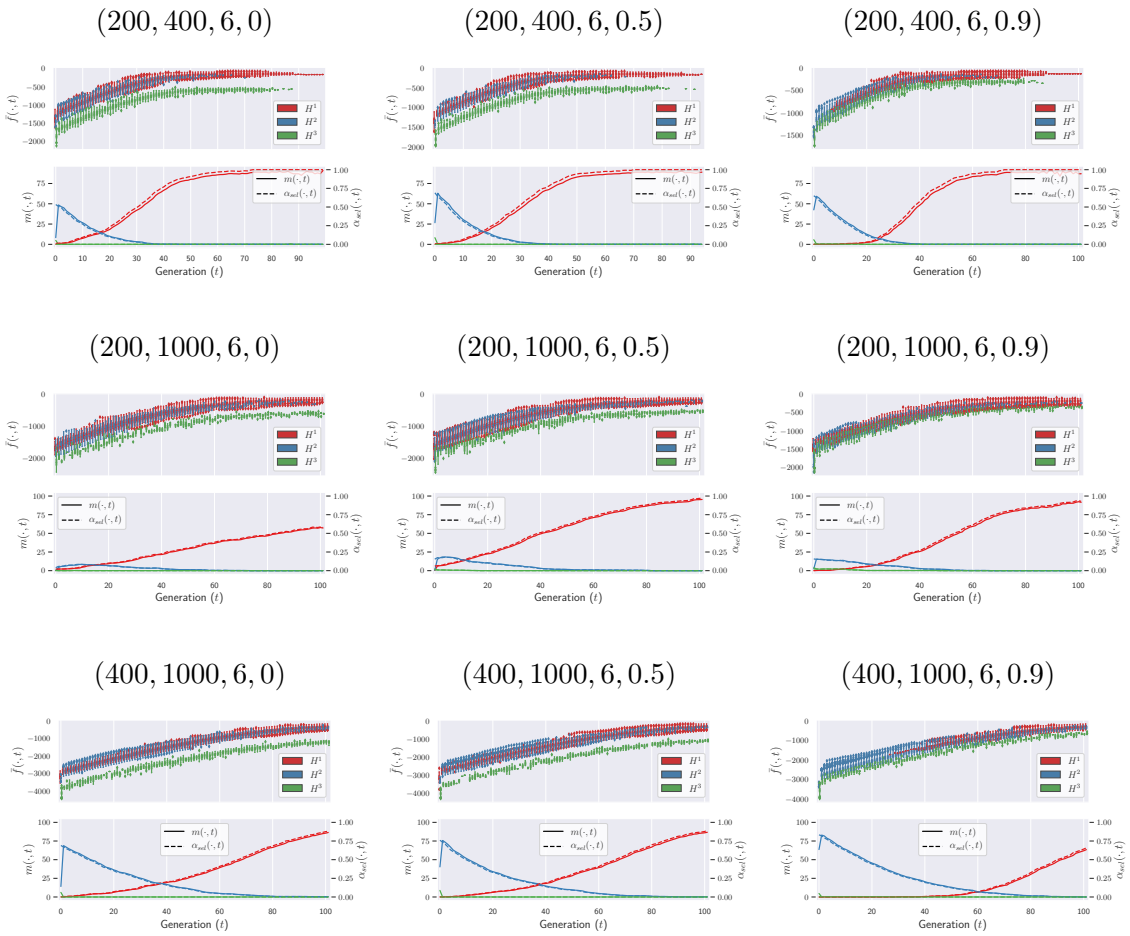

Figure B.3.: (Case 3) Schema performance (upper panel) and evolution (lower panel) obtained from 500 simulation runs. Subfigure titles indicate simulation parameter settings (n, d, s, ρ).

B.3 Variable Coding for the Residential Building Dataset

The variable coding with descriptions and units for the residential building dataset used in Section 6.2 is listed in Table B.1. Detailed explanations are omitted and can be found in Table 1 of [24].

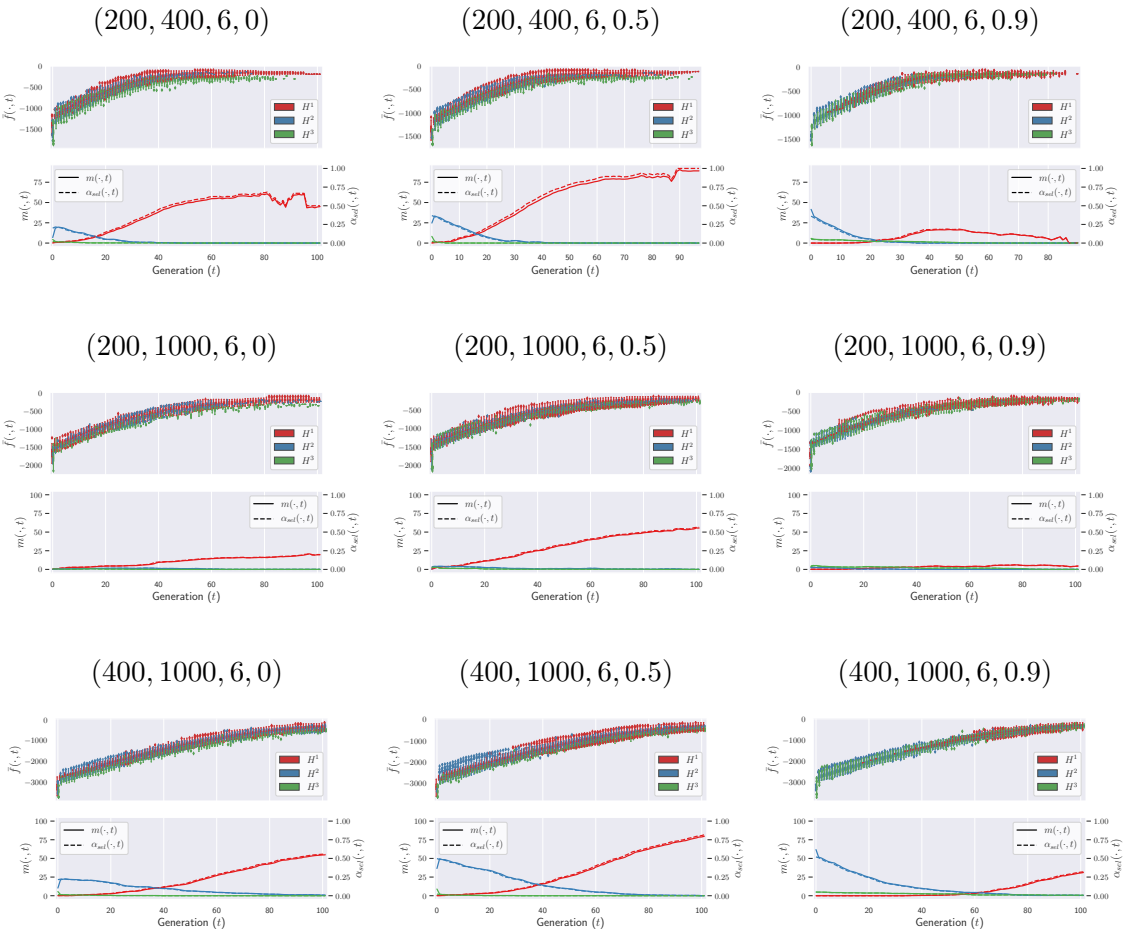

Figure B.4.: (Case 4) Schema performance (upper panel) and evolution (lower panel) obtained from 500 simulation runs. Subfigure titles indicate simulation parameter settings (n, d, s, ρ).

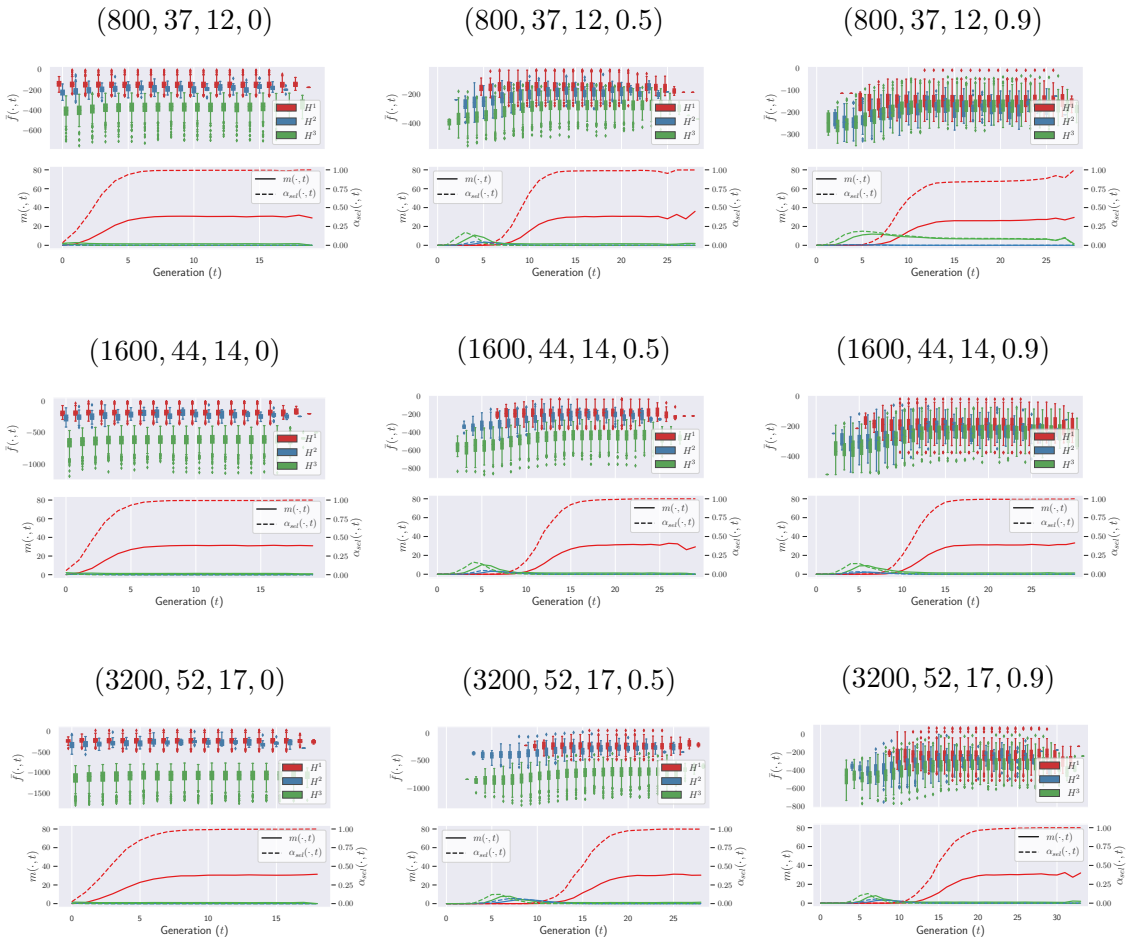

Figure B.5.: (Case 5) Schema performance (upper panel) and evolution (lower panel) obtained from 500 simulation runs. Subfigure titles indicate simulation parameter settings (n, d, s, ρ).

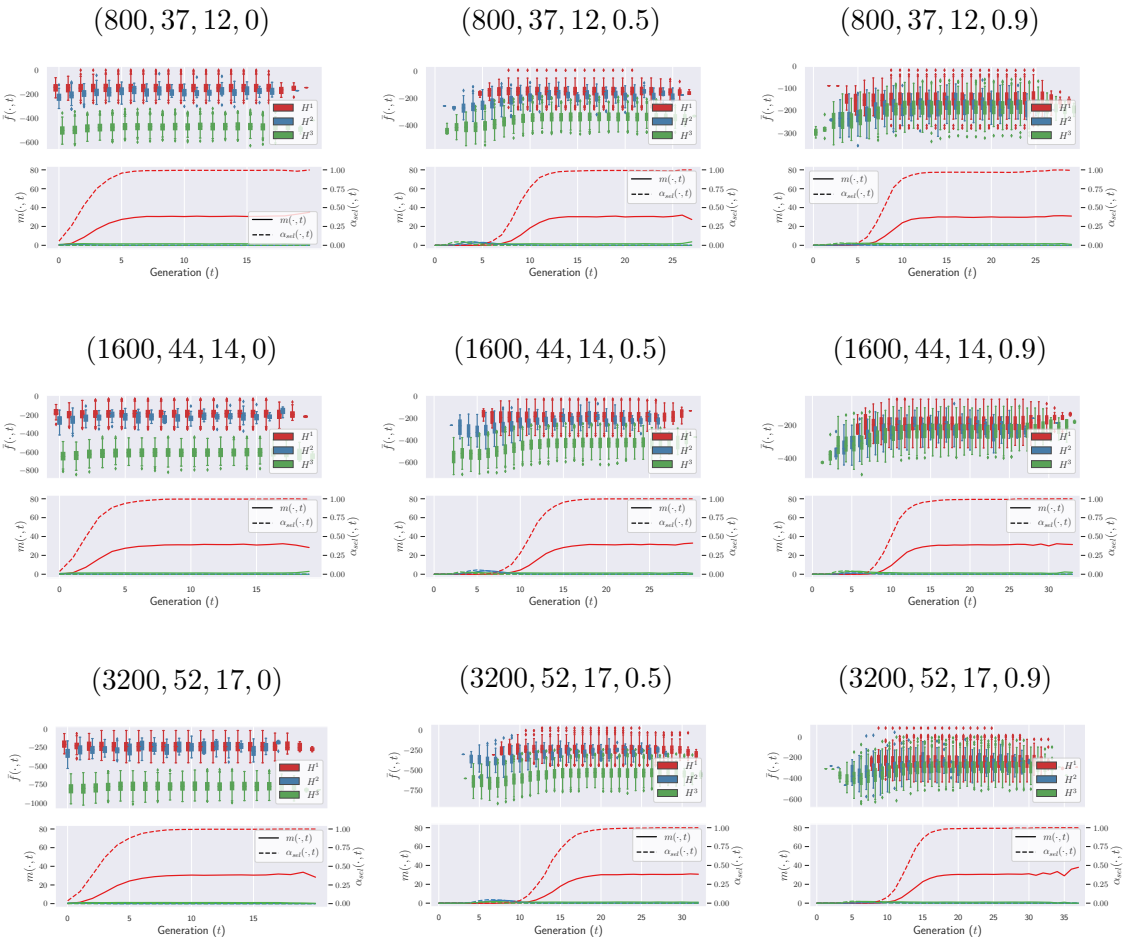

Figure B.6.: (Case 6) Schema performance (upper panel) and evolution (lower panel) obtained from 500 simulation runs. Subfigure titles indicate simulation parameter settings (n, d, s, ρ).

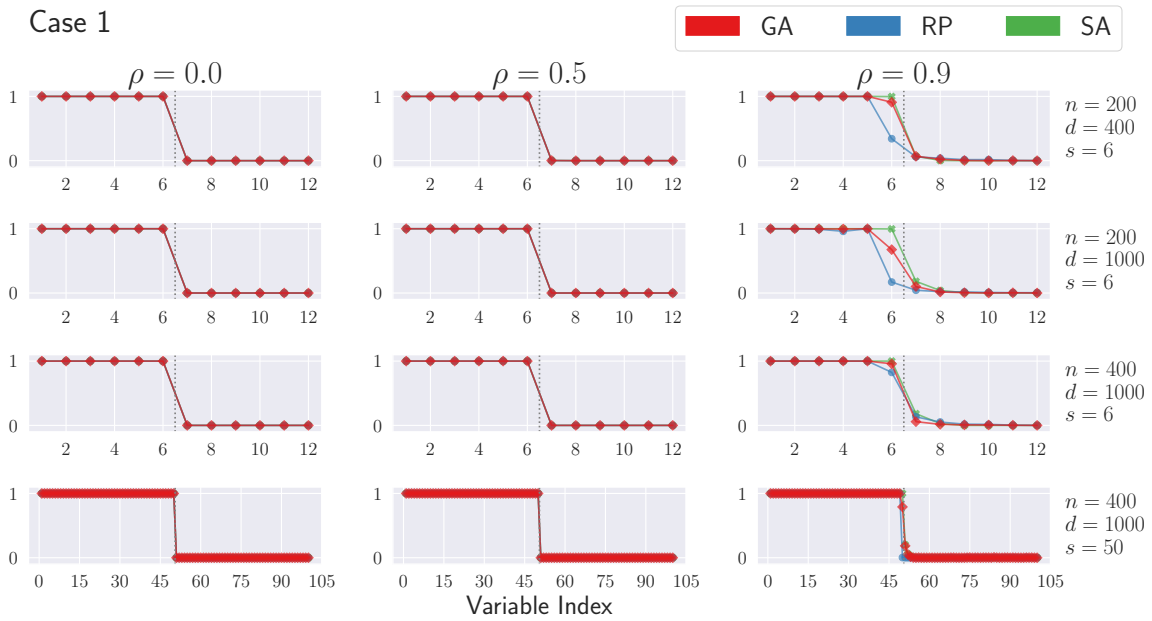

Figure B.7.: (Case 1) Averaged SOIL measures.

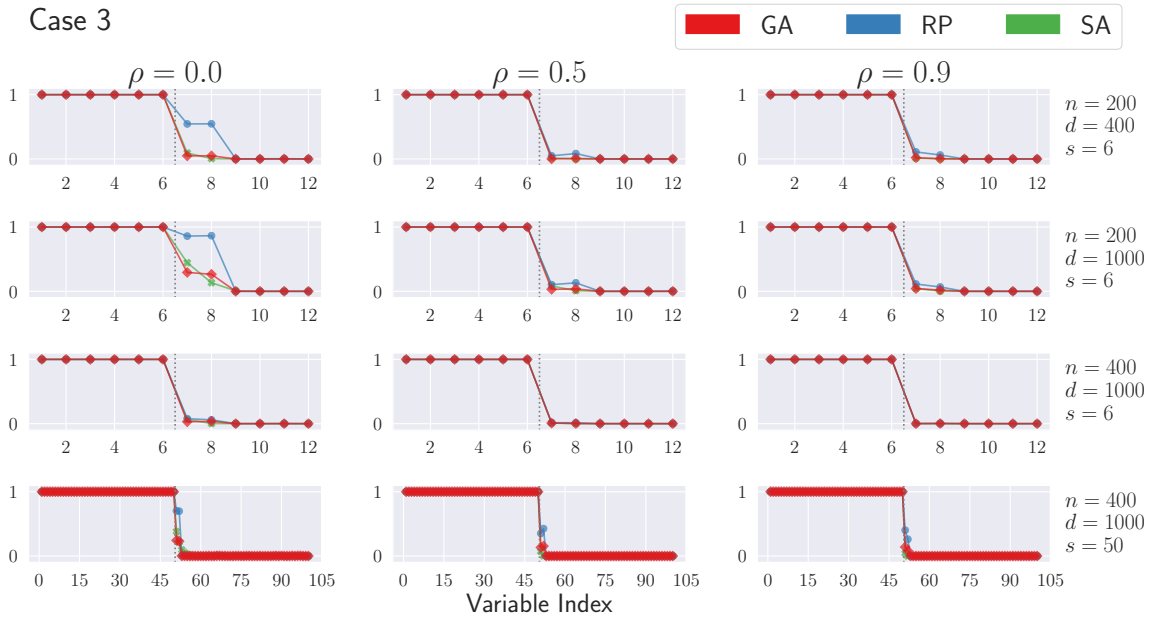

Figure B.8.: (Case 3) Averaged SOIL measures.

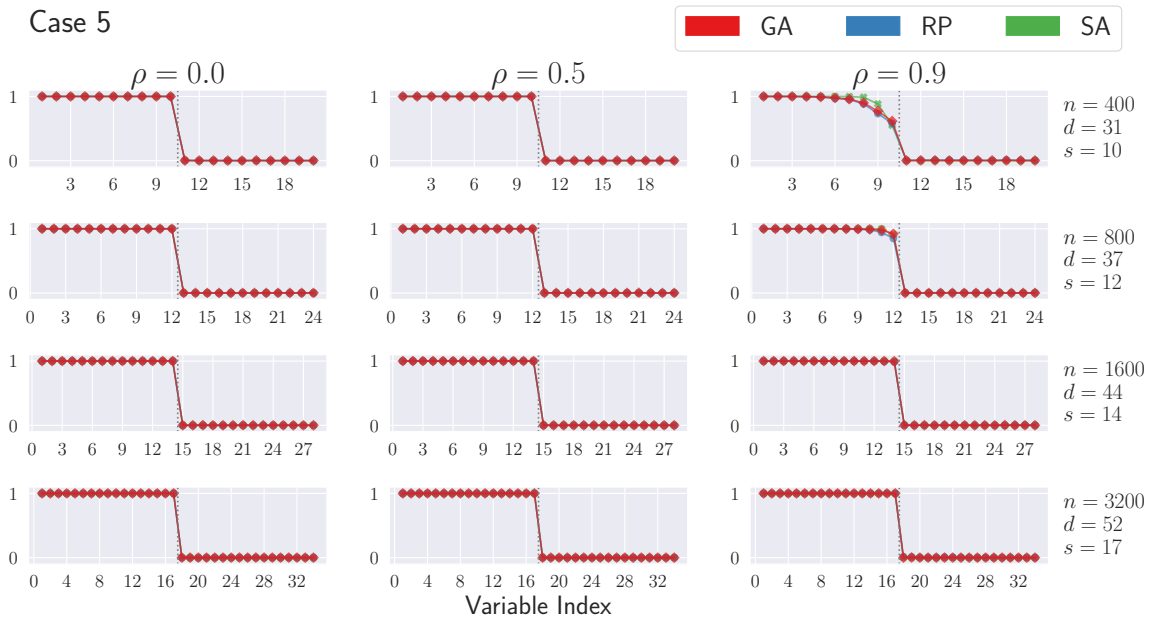

Figure B.9.: (Case 5) Averaged SOIL measures.

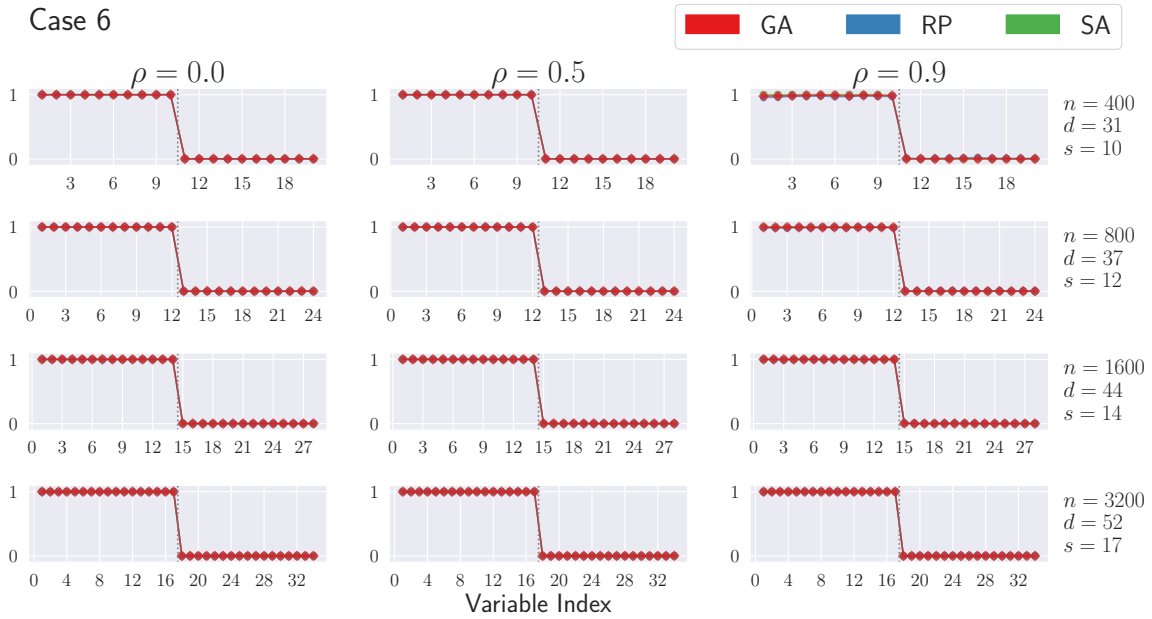

Figure B.10.: (Case 6) Averaged SOIL measures.

Table B.1.: Variable coding for the residential building dataset.

Code	Descriptions	Unit
Project Physical and Financial (PF) Variables		
PF-1	Project locality defined in terms of zip codes	N/A
PF-2	Total floor area of the building	m^2
PF-3	Lot area	m^2
PF-4	Total preliminary estimated construction cost based on the prices at the beginning of the project	10^7 IRR^m
PF-5	Preliminary estimated construction cost based on the prices at the beginning of the project	10^5 IRR^m
PF-6	Equivalent preliminary estimated construction cost based on the prices at the beginning of the project in a selected base year	10^5 IRR^m
PF-7	Duration of construction	Time resolution
PF-8	Price of the unit at the beginning of the project per m^2	10^5 IRR^m
Economic Variables and Indexes (EVI)		
EVI-01	The number of building permits issued	N/A
EVI-02	Building services index (BSI) for a preselected base year	N/A
EVI-03	Wholesale price index (WPI) of building materials for the base year	N/A
EVI-04	Total floor areas of building permits issued by the city/municipality	m^2
EVI-05	Cumulative liquidity	10^7 IRR^m
EVI-06	Private sector investment in new buildings	10^7 IRR^m
EVI-07	Land price index for the base year	10^7 IRR^m
EVI-08	The number of loans extended by banks in a time resolution	N/A
EVI-09	The amount of loans extended by banks in a time resolution	10^7 IRR^m
EVI-10	The interest rate for loan in a time resolution	%
EVI-11	The average construction cost of buildings by private sector at the time of completion of construction	10^5 IRR$^m/m^2$
EVI-12	The average of construction cost of buildings by private sector at the beginning of the construction	10^5 IRR$^m/m^2$
EVI-13	Official exchange rate with respect to dollars	IRR^m
EVI-14	Nonofficial (street market) exchange rate with respect to dollars	IRR^m
EVI-15	Consumer price index (CPI) in the base year	N/A
EVI-16	CPI of housing, water, fuel and power in the base year	N/A
EVI-17	Stock market index	N/A
EVI-18	Population of the city	N/A
EVI-19	Gold price per ounce	IRR^m

VITA

Ching-Wei Cheng was born in Taipei City, Taiwan, on February 25, 1985. He received his bachelor degree in Applied Mathematics from National Dong Hua University in 2007, and master degree in Statistics from National Central University in 2009. After one year of obligational military service, he joined Taiwan Suicide Prevention Center and then the Institute of Statistical Science at Academia Sinica as research assistants to hone his practical and academic techniques for higher statistical training. Afterward, he joined the Department of Statistics at Purdue University, where he expects to receive his PhD degree in 2019.