Statistical Methods for Functional Genomics Studies Using Observational Data

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Rong Lu, B.S., M.S.

Graduate Program in Biostatistics

The Ohio State University

2016

Dissertation Committee:

Grzegorz A. Rempala, Advisor
Wolfgang Sadee
Shili Lin
© Copyright by

Rong Lu

2016
Abstract

In functional genomics studies, human tissue samples are always difficult to get access to, and the lab experiments are expensive to implement and time-consuming. Data mining in existing databases is an essential step in building scientific hypotheses for designing well-targeted lab experiments. Therefore, it is important to study statistical methods that can better utilize observational data in functional genomics studies.

Measuring allele-specific RNA expression provides valuable insights into cis-acting genetic and epigenetic regulation of gene expression. Widespread adoption of high-throughput sequencing technologies for studying RNA expression permits measurement of allelic RNA expression imbalance at heterozygous single nucleotide polymorphisms (SNPs) across the entire transcriptome, and this approach has become especially popular with the emergence of large databases, such as GTEx. However, the existing methods used to model allelic expression from RNA-seq often assume a strong negative correlation between reference and variant allele reads, which may not be reasonable biologically. In Chapter 2, a folded Skellam mixture model is proposed for AEI analysis using RNA-seq data. Under the null hypothesis of no AEI, a group of SNPs (possibly across multiple genes) is considered comparable if their respective total sums of the allelic reads are of similar magnitude. Within each group of comparable SNPs, we identify SNPs with AEI signal by fitting a mixture of folded Skellam
distributions to the absolute values of read differences. By applying this methodology to RNA-Seq data from human autopsy brain tissues, we identified numerous instances of moderate to strong imbalanced allelic RNA expression at heterozygous SNPs. Findings with SLC1A3 mRNA exhibiting known expression differences are discussed as examples.

In the theory of complex systems, the Sobol sensitivity indices are typically introduced under the high dimension model representation (HDMR, also known as functional ANOVA), assuming all the inputs are independent uniform random variables. The variance-based definitions of Sobol indices are available for analyzing systems with correlated or non-uniform inputs. The existing algorithms for estimating Sobol indices with correlated inputs mostly start with approximating the underlying full model by meta-models with certain type of orthogonality among the decomposition components, which is computationally expensive to implement especially when the number of inputs is large. In Chapter 3, a simple strategy for estimating Sobol indices is proposed under the generalized linear models with independent or multivariate normal inputs. If the ultimate goal is only to estimate Sobol indices for variable selection instead of building a predictive model, it may be more convenient to approximate conditional expectations of the response with respect to different input subsets separately, without reconstructing the complete input-output map. It can be shown that under a large group of GLMs, Sobol sensitivity indices can be either estimated directly using closed analytic formulas or approximated numerically using empirical variance estimates to any level of desired accuracy, without requiring the knowledge of the underlying true model or its HDMR. The usage of this method is
illustrated in the application example of selecting genes that are co-expressed with a target gene of interest, CYP3A4.
This is dedicated to my parents,

♥ Yuwen Lu and Jin Zhang ♥,

for their endless love, support, and trust.
Acknowledgments

First, I’d like to thank my dissertation advisor, Professor Grzegorz A. Rempala, for his guidance and unending patience with me. He has not only taught me statistics, but also help me to become a better thinker and researcher. The tangible and intangible advantages of working with him are too many to enumerate. This dissertation would not have been possible without his help. I also would like to express my sincerest thanks to Professor Wolfgang Sadee and Professor Danxin Wang for helping me understand many pharmacogenomics concepts. Exposure to their team at the center of pharmacogenomics has left me in awe of their passion for their field. Last but not least, I thank Professor Shili Lin for serving on my dissertation committee and taking time evaluating my work. I thank Dr. Min Wang for helping implementing the Sobol index formulas in R package SobolSensitivity. Additionally, I must also acknowledge that the AEI project discussed in Chapter 2 is supported by the National Institute of General Medical Sciences (U01GM092655), the US National Science Foundation (DMS-1318886), and the US National Cancer Institute (R01-CA152158). The CYP3A4 project in Chapter 3 is supported by the National Institute of General Medical Sciences (U01GM092655) and the US National Cancer Institute (R01-CA152158). Both projects received allocation of computing time from the Ohio Supercomputer Center.
Vita

1986 .............................................Born - Dafeng, Yancheng, China

2008 ...........................................B.S. Computational Mathematics

2011 ...........................................M.S. General Mathematics

2014 ...........................................M.S. Statistics

2013-present .................................Graduate Associate, The Ohio State University.

Publications

Research Publications


Fields of Study

Major Field: Biostatistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>v</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>vi</td>
</tr>
<tr>
<td>Vita</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Allele Expression Imbalance</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 RNA Sequencing</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 AEI Signal on Nucleotide Level</td>
<td>3</td>
</tr>
<tr>
<td>1.1.3 Confounding between AEI and Genomic Imprinting</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Gene Activity Analysis</td>
<td>6</td>
</tr>
<tr>
<td>1.2.1 Epistasis</td>
<td>6</td>
</tr>
<tr>
<td>1.2.2 Co-regulated Genes</td>
<td>8</td>
</tr>
<tr>
<td>1.3 Organization of this Thesis</td>
<td>9</td>
</tr>
<tr>
<td>2. AEI Signal Detection Using Mixture Models</td>
<td>11</td>
</tr>
<tr>
<td>2.1 Human Brain RNA-Seq</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Existing Methods for Observational Studies</td>
<td>15</td>
</tr>
<tr>
<td>2.3 Using Folded Skellam Mixture in AEI Analysis</td>
<td>17</td>
</tr>
<tr>
<td>2.3.1 Folded Skellam Mixture Model</td>
<td>17</td>
</tr>
<tr>
<td>2.3.2 Mixture Model Pipeline</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Model Fitting Results</td>
<td>22</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Poisson Mixture Fitting Results</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Folded Skellam Mixture Fitting Results</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Mixture Model Pipeline Performance Analysis</td>
</tr>
<tr>
<td>2.5</td>
<td>Investigation of Identified AEI Signals</td>
</tr>
<tr>
<td>2.5.1</td>
<td>SNP-level AEI Signals on Gene SLC1A3</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Signal Designation Consistency Across Brain Tissues</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Mixture Model Pipeline vs. Whole Gene Filtering Method</td>
</tr>
<tr>
<td>2.5.4</td>
<td>Parallels Between AEI and eQTLs</td>
</tr>
<tr>
<td>3.1</td>
<td>Sensitivity Analysis</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Local and Global Sensitivity Measurements</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Estimation of Sobol Indices with Independent Inputs</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Estimation of Sobol Indices with Correlated Inputs</td>
</tr>
<tr>
<td>3.2</td>
<td>Generalized Linear Models</td>
</tr>
<tr>
<td>3.3</td>
<td>Sobol Indices under GLMs</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Variance-based Definition of Sobol Indices</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Sobol Indices under Linear GLMs</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Sobol Indices under Polynomial GLMs</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Multiple Testing of Sobol Indices</td>
</tr>
<tr>
<td>3.4</td>
<td>Simulation Studies</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Simulations under Gaussian Models</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Simulation under Poisson Models</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Variable Ranking Comparison</td>
</tr>
<tr>
<td>3.5</td>
<td>Application Example: Identifying Co-expressed Genes</td>
</tr>
<tr>
<td>3.6</td>
<td>Other Possible Applications in Gene Activity Analysis</td>
</tr>
<tr>
<td>4.</td>
<td>Contributions and Future Work</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
</tr>
<tr>
<td></td>
<td>Appendices</td>
</tr>
<tr>
<td>A.</td>
<td>Additional Figures and Tables of AEI Analysis</td>
</tr>
<tr>
<td>B.</td>
<td>Proofs of Inverse-logit Function Expectations</td>
</tr>
</tbody>
</table>
C. Proofs of Sobol Index Formulas under Linear GLMs . . . . . . . . . . . . 144

D. Proofs of Sobol Index Estimation under Polynomial GLMs . . . . . . . 153

E. Gaussian Model Simulation with Less Dependent Inputs . . . . . . . . . 157

F. Poisson Model Simulation with Less Dependent Inputs . . . . . . . . . . 160
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Poisson Mixture Model Parameter Estimates and SNPs Classification Results</td>
<td>23</td>
</tr>
<tr>
<td>2.2 Poisson Mixture Comp.1 SNP Counts by Gene Regions</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Folded Skellam Mixture Parameter Estimates And Results of AEI LRTs</td>
<td>27</td>
</tr>
<tr>
<td>2.4 Percentiles of Absolute Reads Ratio</td>
<td>30</td>
</tr>
<tr>
<td>3.1 Canonical Link Functions of Commonly Used GLMs</td>
<td>40</td>
</tr>
<tr>
<td>3.2 Outcome Summary of $M$ Significance Tests</td>
<td>53</td>
</tr>
<tr>
<td>3.3 Quantiles of Relative Difference between SI Estimates and the Corresponding Exact Estimates under Gaussian Model ($\rho = 0.8$)</td>
<td>59</td>
</tr>
<tr>
<td>3.4 Type I Error, Power, and FDR Estimates ($\rho = 0.8$)</td>
<td>65</td>
</tr>
<tr>
<td>3.5 Type I Error, Power, and FDR Estimates ($\rho = 0.3$)</td>
<td>67</td>
</tr>
<tr>
<td>3.6 Quantiles of Relative Difference between SI Estimates and the Corresponding Correct Estimates under Poisson Model with Identity Link ($\rho = 0.8$)</td>
<td>74</td>
</tr>
<tr>
<td>3.7 Quantiles of Relative Difference between SI Estimates and the Corresponding Exact Estimates under Poisson Model with Log Link ($\rho = 0.8$)</td>
<td>85</td>
</tr>
<tr>
<td>3.8 Variable Ranking Comparison by Mean Spearman Rho Under Correct Model</td>
<td>94</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Simulation under Fitted Folded Skellam Mixture Model</td>
<td>25</td>
</tr>
<tr>
<td>3.1</td>
<td>Sobol Index Estimates for Linear Gaussian Model with Identity Link</td>
<td>58</td>
</tr>
<tr>
<td>3.2</td>
<td>Variable Selection Methods Comparison under Multivariate Linear Gaussian Model (inputs correlation $\rho = 0.8$)</td>
<td>62</td>
</tr>
<tr>
<td>3.3</td>
<td>Sobol Index Significance Test under Multivariate Linear Gaussian Model (inputs correlation $\rho = 0.8$)</td>
<td>63</td>
</tr>
<tr>
<td>3.4</td>
<td>ROC Curves for Method Comparison under Multivariate Linear Gaussian Model</td>
<td>66</td>
</tr>
<tr>
<td>3.5</td>
<td>Total-effect Sobol Indices under Multivariate Linear Gaussian Model with Inputs Correlation $\rho = 0.8$</td>
<td>69</td>
</tr>
<tr>
<td>3.6</td>
<td>Sobol Index Estimates for Linear Poisson Model with Identity Link</td>
<td>72</td>
</tr>
<tr>
<td>3.7</td>
<td>Variable Selection Methods Comparison under Linear Poisson Model with Identity Link and Inputs Correlation $\rho = 0.8$</td>
<td>77</td>
</tr>
<tr>
<td>3.8</td>
<td>Sobol Index Significance Test under Linear Poisson Model with Identity Link and Inputs Correlation $\rho = 0.8$</td>
<td>78</td>
</tr>
<tr>
<td>3.9</td>
<td>ROC Curves for Method Comparison under Linear Poisson Model with Identity Link</td>
<td>80</td>
</tr>
<tr>
<td>3.10</td>
<td>Sobol Index Estimates for Linear Poisson Model with Log Link</td>
<td>84</td>
</tr>
<tr>
<td>3.11</td>
<td>Variable Selection Methods Comparison under Linear Poisson Model with Log Link and Inputs Correlation $\rho = 0.8$</td>
<td>88</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>3.12 Sobol Index Significance Test under Linear Poisson Model with Log Link and Inputs Correlation $\rho = 0.8$</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>3.13 ROC Curves for Method Comparison under Linear Poisson Model with Log Link</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>3.14 Variable Ranking Comparison Example Under Contaminated Gaussian Model ($\rho = 0.8$)</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>3.15 Variable Ranking Comparison Example Under Contaminated Gaussian Model ($\rho = 0.3$)</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>3.16 CYP3A4 Sensitivity Network with the Top Gene Quadruplets</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>3.17 Gene Quadruplet with Smallest Residual Deviances</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td>4.1 Likelihood Ratio Test Statistics Calculated Using Moment Estimates</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td>A.1 Scatter Plots of RNA-seq Read Pairs</td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>A.2 Histogram of Observed Absolute Read Differences with Signal Classification</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td>A.3 Q-Q Plots for Checking Folded Skellam Model Fitting</td>
<td>132</td>
<td></td>
</tr>
<tr>
<td>E.1 Variable Selection Methods Comparison (inputs correlation $\rho = 0.3$)</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td>E.2 Sobol Index Significance Test versus Other Methods (inputs correlation $\rho = 0.3$)</td>
<td>159</td>
<td></td>
</tr>
<tr>
<td>F.1 Sobol Index Significance Test under Linear Poisson Model with Log Link and Inputs Correlation $\rho = 0.3$</td>
<td>161</td>
<td></td>
</tr>
<tr>
<td>F.2 Sobol Index Significance Test under Linear Poisson Model with Log Link and Inputs Correlation $\rho = 0.3$</td>
<td>162</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

Given the current technologies, lab experiments for functional genomics studies are still very expensive and time-consuming to implement. Since designed experiments can only test one or two hypothesis at a time and human samples are always difficult to collect, it’s an indispensable step to research existing human databases to construct scientific hypotheses that can have high chances to be confirmed in designed experiments. In order to use the existing databases to the greatest advantage, it is important to continue studying statistical methods that can better utilize the observational data. In this thesis, we will focus on investigating statistical methods for identifying allele expression imbalance signal in observational RNA-seq data and on exploring the usage of Sobol sensitivity indices in gene activity analysis.

1.1 Allele Expression Imbalance

Allele expression imbalance (AEI) or alternatively allele-specific gene expression (ASE) are used to describe the phenomenon when one parental copy of a given autosomal gene is preferentially expressed over the other in the corresponding RNA transcripts. Gene imprinting is an epigenetic process that silences one copy of the gene completely, resulting an extreme case of AEI. But, more often, we observe less dramatic AEI cases where both parental copies get expressed but the expression levels
differ significantly [29, 133]. *Cis*-acting polymorphisms are believed to be the cause of these AEI cases, because such mutations may change promoter/enhancer regions of the gene, alter transcription factor binding sites, or affect RNA stability [71, 98]. Measuring allele-specific RNA expression provides valuable insights into *cis*-acting genetic and epigenetic regulation of gene expression. The goal of AEI analysis is to separate the true signals (imbalanced expression due to biological mechanisms) from the noises (imbalanced expression due to instrumental variations and experimental biases). Since imbalanced expression levels are used as the phenotype for identifying the responsible genetic variants, it is crucial to be able to get stable AEI analysis results without making unrealistic model assumptions.

1.1.1 RNA Sequencing

RNA Sequencing (RNA-Seq) is an application of next-generation sequencing [54] technology, which sequences and quantifies complementary DNAs (cDNAs) generated from RNA. It can provide a snapshot of RNA presence with much higher resolution than microarray-based methods [39, 117]. The basic RNA-Seq strategies include isolating RNA from the cell, preparing a library containing amplified fragments of cDNA, and sequencing. The detailed protocols can vary, depending on whether only polyadenylated RNA is isolated, whether the amplification is done via polymerase chain reaction [26], whether the sequencing is single end or paired ends, etc. The output of RNA-Seq is generally in the form of a FASTQ file [77], which contains the read name, the raw sequence, and information on the quality of each base call. Several alignment software, such as GSNAP [90], MapSplice [89] and STAR [114], can be used to match the reads to a reference genome.
Alignment error occurs when reads cannot be uniquely mapped, or when too many non-reference SNP alleles are observed within one read. Depending on the specific alignment algorithm being used, reads which cannot be uniquely mapped may be assigned to multiple locations with different probabilities or excluded entirely from later analysis, while the reads with too many variants are discarded as experimental errors most of the time. Therefore, the reads of the variant alleles at heterozygous loci generally have lower probability to be mapped correctly than that of the reference alleles. Such bias is referred as the “reference bias” in the literature. IUPAC ambiguity codes and longer reads can be used to help minimize the reference bias. When RNA-seq read counts are used for AEI analysis, appropriate statistical methods are needed to avoid classifying count differences due to experimental bias as the real signals.

1.1.2 AEI Signal on Nucleotide Level

Due to high complexity of observational data often the best we can do in practice is to find potential AEI signals that are strong enough to stand out in massive background noises. Majority of published AEI studies have focused on searching for AEI genes instead of AEI SNPs. There are two reasons why genes are primarily used as the study units in AEI studies: 1. The concept of AEI is originally introduced in terms of a gene. 2. Researchers used to analyze microarray data for AEI analysis, and those microarray-based technologies can not provide nucleotide-level resolution on gene expression. However, it is important to look for AEI signals at the nucleotide level for the following reasons:
Firstly, AEI signals are not always observable at gene level. For moderate AEI genes, the signals are hardly seen consistently at all heterozygous loci across the entire gene and across all tissue samples. This is not only because of the massive experimental noises but also because of the biological differences between different tissue samples. In the process of gene expression, different messenger RNAs (mRNAs) are generated from the same gene due to a mechanism called alternative splicing. A particular exon of the gene may be included in one mRNA isoform but excluded from another, which also prevents us observing consistent AEI signal at gene level.

Secondly, if multiple SNPs in the same gene show significant differences between the read counts of the reference and variant alleles consistently across subjects or organ tissue, this gene is much more likely to have real AEI signals. The percentage of AEI SNPs out of all SNPs observed in the same gene can be viewed as an estimate of the probability that this gene has AEI. As long as the pattern found on SNPs is consistent either across subjects or across organ tissue, the presence of consistent pattern by itself is valuable and worth studying.

1.1.3 Confounding between AEI and Genomic Imprinting

In section 1.1.1, we define the AEI signal as the asymmetric expression of two alleles at the same locus regardless of the cause. But sometimes in literature AEI only refers to asymmetric expression due to a specific allele type. This means we can only observe this type of AEI signals at the heterozygous loci. And if the allele had a different allele type at this locus or both alleles had the same allele type we would not be able to see the asymmetric expression. Recent work focusing on identifying
this type of AEI signal includes Zhang et al. 2009 [74], Fontanillas et al. 2010 [78], Xu et al. 2011 [102] etc.

But allele type is not the only factor that can result in imbalanced expression at the same locus. Some genes only express the allele inherited from the father (mother) and silence the other copy. This phenomenon is called maternal (paternal) imprinting. Classical paternal imprinted genes in human include OBSCN on chromosome 1, HES1 on chromosome 3, PLAGL1 on chromosome 6, COPG2IT1 on chromosome 7, PURG on chromosome 8, IGF2 on chromosome 11, etc. And some of the maternal imprinted genes in human include ZFP36L2 on chromosome 2, MAGI2 on chromosome 8, KCNK9 on chromosome 8, PHPT1 on chromosome 9, VENTX on chromosome 10, KCNQ1 on chromosome 11, etc. To this date, hundreds of genes are known to be genomically imprinted in different species.

In animal or plant studies, imprinting effects can be easily detected by creating large sample of filial 1 hybrids by exchanging the distinct homogeneous parental types. This design is called reciprocal cross design in literature [59, 47, 81, 141]. In human studies, if the phasing information is known and accurate, the imprinting effects can be tested using pedigree data or family trio data under different model assumptions, such as no maternal effect or the quantitative traits must be normally distributed [13, 17, 75, 95, 94, 113]. So without accurate phasing information or family data, we often cannot distinguish the asymmetric expression due to specific allele type and that due to genomic imprinting. In addition, more and more genes are identified to have asymmetric expression affected by both the allele type and the parent-of-origin. Therefore, in Chapter 2 of this thesis, We will not differentiate this two type
of imbalanced expression signals, and focus on methods applicable to population data without pedigree information.

1.2 Gene Activity Analysis

Gene activity analysis includes the full spectrum research on the functionality of a gene, including its interactions with other genes, the genetic control of the gene expression in different cell types, regulatory mechanisms at transcriptional, translational, and post-translational levels that can cause variations in the gene expression across different individuals, etc [48, 69, 92]. In chapter 3, we will illustrate the advantages of using Sobol indices estimated via fitting generalized linear models, in the context of studying two specific types of gene activities: epistasis and gene co-regulation.

1.2.1 Epistasis

Gene-gene interaction happens when a phenotypic trait is affected by two or more genes. More specificity, if the phenotypic trait can be directly linked to the genotypes of two or more genes, we call this type of gene-gene interaction the epistatic effects [30, 53, 55, 73]. There are four types of functional epistasis discussed the most in literature. One is called the duplicate gene actions, in which we observe the phenotypic trait whenever at least one of a group of genes have its dominant allele. The second type is called the complementary gene actions, in which all relevant genes need to have their dominant alleles to produce the phenotypic trait. The third type of epistasis is called the dominant suppression or dominant epistasis. In a simple dominant suppression case, the dominant allele type of one particular gene (the epistatic gene) can mask or alter the phenotypical manifestation of another gene (the hypostatic genes). In other words, different phenotypic traits determined by the hypostatic gene will reveal
only themselves when the epistatic gene is recessive homozygous. The fourth type is the recessive suppression or recessive epistasis, in which the expression of hypostatic genes are masked only when the epistatic gene is recessive homozygous.

In observational studies, there is another related concept, termed statistical epistasis, which can mean different things depending on what method is used for identifying the “interaction effects” [20, 46, 57, 63, 64, 135]. For example, in most regression-based approaches, statistical epistasis means there are statistically significant product terms in model fitting; In linkage disequilibrium (LD) based methods, it means there are statistically significant differences in LD between cohorts classified according to a categorical trait; Similar to the LD-based methods, if the tests are performed on a contingency table of haplotype frequencies, the statistical epistasis often means statistically significant odds ratio for comparing genotype frequencies between cohorts with different type of traits.

Various software are publicly available for genome-wide scan of statistical epistasis, such as PLINK epistasis module (the benchmark approach for new application development), SNP-SNP interactions (based on regression), eCEO (regression based and implemented by bitwise cloud computing), SIXPAC (LD based), EPIBLASTER (use both LD-based screening and logistic regression), IndOR (based on odds ratio test), etc. Statistical epistasis only helps to infer potential functional epistasis with higher probabilities. Observing statistical epistasis is neither sufficient nor necessary when functional epistasis present. To claim a functional epistasis discovery, we still need verification from well-designed biological experiments.
1.2.2 Co-regulated Genes

Co-regulated genes are the group of genes that are required to express coordinately to complete a complex regulation process [65]. In observational studies, one way of inferring gene co-regulation is to search for genes with dependent expression patterns, i.e. co-expressed genes [51, 52]. This is because genes targeted by the same transcription factors are believed to be more likely to show dependency in expression levels. For example, Yu et. al. (2003) integrated a yeast regulation dataset with the expression data of the corresponding 3,474 target genes. And they found 3.3% target gene pairs are co-expressed, which is 4 times greater than the random expectation.

Conventionally, the dependence of gene expression is defined as the absolute value of Pearson correlation on gene pairs [36, 38]. But other measures, such as Euclidean distances and mutual information, also have been applied to quantify the similarity in expression patterns. Once the dependency measure is defined, co-expressed genes can be grouped or organized in hierarchical tree structures by applying different clustering algorithms such as K-means or other regression-based classifiers [51, 52, 65].

However, majority of co-regulated genes have dependent expression that is not detectable by simple linear correlation or other similarity measures defined on gene pairs. This is because most regulation processes involve more than one pair of genes. Or in other words, very few regulation processes are dominated by only two genes. Therefore, in addition to construct gene co-expression networks, researchers also have applied ordinary differential equation models (ODE) to track mRNA decay rates in the processes of cell growth and division, i.e. the so-called regulated flux balance analysis (rFBA) [15, 21, 27, 31]. If time-course expression data is available, dynamic
Bayesian networks are commonly used to infer time-dependent activities among transcription factors [11, 28].

1.3 Organization of this Thesis

The rest of this thesis is organized as follows. Chapter 2 will investigate appropriate statistical methods for identifying AEI SNPs in human brain tissues. Section 2.1 will introduce the RNA-seq data and discuss data characteristics that may affect the model choice and data preprocessing steps. Section 2.2 will discuss the drawbacks of using currently available methods and the motivation of constructing the folded Skellam mixture model. Details of the proposed mixture model pipeline is presented in Section 2.3. The corresponding model fitting results is discussed in Section 2.4, where we will also compare the performance of proposed approach to ratio based method and the basic binomial test. In Section 2.5, we will further investigate the identified AEI SNPs on a known AEI gene SLC1A3, check the signal pattern across different brain tissues, and look for eQTLs within the identified AEI genes.

In Chapter 3, we will mainly focus on exploring the usage of Sobol sensitivity indices in identifying co-expressed genes. As an example, we will investigate the relationship between 46 pre-selected genes and a target gene, CYP3A4, using a published microarray dataset. To facilitate the methodology discussion, Section 3.1 will firstly introduce some basic concepts in sensitivity analysis and review existing methods for estimating Sobol indices. Section 3.2 will give a short introduction to the generalized linear models. In Section 3.3, a new idea for estimating Sobol indices is discussed under the generalized linear models, with the assumption that the inputs are either independent or follow a multivariate normal distribution. This new estimation
strategy is then illustrated and examined in simulation studies in Section 3.4. After applying the proposed method, we identified several gene quadruplets which appear to explain about 68% of variation in CYP3A4 expression across different individual. The detailed results is reported in Section 3.5. At last, we discuss other possible applications of Sobol indices in gene activity analysis in Section 3.6.
Chapter 2: AEI Signal Detection Using Mixture Models

High-throughput DNA sequencing technology, when used for measuring RNA expression (RNA-Seq), provides nucleotide-level resolution of gene expression across the entire transcriptome in a single experiment. This enhanced resolution provides a wealth of detail about gene expression not available through microarray-based technologies. One important goal is to identify regulatory variants that affect transcription and RNA processing. Use of RNA expression arrays and RNA-Seq to determine transcript levels in multiple samples, combined with single nucleotide polymorphism (SNP) chip genotyping, can reveal expression quantitative trait loci (eQTLs) acting either in cis (located at the target gene locus) or in trans [70]. A major caveat of eQTLs is their sensitivity to trans-acting factors, sometimes making it difficult to attribute changes in expression to a causative variant. On the other hand, allelic mRNA ratios reduce the effect of trans-acting factors, revealing the presence of allele-specific regulatory factors acting in cis when allelic ratios in the RNA differ from that in gDNA, termed here “allelic RNA expression imbalance” (AEI) [70].

In the literature, the terms AEI or alternatively allele-specific gene expression (ASE) are used to describe the phenomenon when one parental copy of a given autosomal gene is preferentially expressed over the other in the corresponding RNA transcript. Commonly, regulatory variants cause AEI, but epigenetic processes can
also be allele-selective, such as with imprinting. Recent studies have taken advantage of the single-base resolution afforded by RNA-Seq to measure allelic RNA expression at heterozygous single nucleotide polymorphisms (SNPs) in the brain [119, 120] and liver [123], among other human tissues [108, 83]. Genomic regions subject to epigenetic programming, such as imprinting, which typically results in large (> 10-fold) AEI because of near-complete silencing of one allele, have been identified from RNA-Seq studies of allelic RNA expression in combination with gDNA genotyping [104, 106]. RNA editing can also result in large allelic RNA ratios [119, 120]. Smaller changes in allelic expression can also have biological relevance. However, RNAseq data yield allelic ratios with relatively high noise; therefore, rigorous statistical methods are needed to identify a signature of AEI in transcriptome-wide analyses.

Less extreme AEI ratios resulting from cis-acting regulatory variants influence a variety of phenotypes [133], including therapeutic drug response [101, 134], complex genetic disease risk [119, 120, 125, 110], risk for drug dependence [96, 100], cognitive processes [45], and lethal drug overdose [121]. However, current methods for analyzing allelic RNA expression from RNA-Seq have substantial drawbacks when attempting to reliably identify modest allelic differences (< 2.5-fold). The main ones are experimental and instrumental noise [97] as well as high read-depth requirements [99]. Even under high-stringency conditions and after grouping allelic ratios from multiple SNPs from the same gene together, our ability to predict modest AEI at low coverage is subject to a considerable false discovery rate [119, 120].
2.1 Human Brain RNA-Seq

The RNA-seq data analyzed in this chapter is collected after sequencing human autopsy brain regions provided from an archived biorepository (University of Miami, Miami, FL, USA), as described in Mash et al., 2007 [43]. Ten subjects (age ranging from 16 to 47 years, five African-American, three European-American, one Pacific Islander, one mixed race) were selected from accidental or cardiac sudden deaths with negative urine screens for illicit drugs, with no history of psychiatric disorders or licit or illicit drug use prior to death; five subjects had a history of cigarette smoking. From each subject, ten different brain regions were obtained: frontopolar cortex (Brodmann Area 10; BA10), Wernicke’s area (BA22), anterior cingulate cortex (BA24), dorsolateral prefrontal cortex (BA46), insular cortex, hippocampus, amygdala, posterior putamen, cerebellum, and brainstem raphe nuclei. In total, our dataset included 98 tissue samples (analysis of two tissues failed). These samples are de-identified prior to attainment.

RNA-Seq transcriptomes were generated from all ten human brain regions in ten different individuals. For each individual, genomic DNA (gDNA) was isolated from the cerebellum and used for genome-wide genotyping with the HumanOmni5Exome BeadChip (Illumina, Inc., San Diego, CA), performed at the University of Utah Genomics Core facility. Total RNA was isolated by homogenizing each tissue in TRIzol, mixing thoroughly with chloroform, and precipitating RNA from the aqueous phase using isopropanol. Total RNA was further purified using SpinSmart Total RNA columns (Denville Scientific, Inc, South Plainfield, NJ), and latent genomic DNA (gDNA) was digested on-column with DNase I (QIAGEN Inc., Valencia, CA). Complementary DNA (cDNA) was reverse transcribed from 25 ng total RNA using the
Ovation RNA-Seq System v2 (NuGen), which suppresses ribosomal RNA conversion to cDNA and employs both poly-dT and random hexamer primers, capturing all RNA species (including non-poly-adenylated RNAs and intronic fragments). This cDNA was used to construct libraries for massively parallel sequencing using the NEBNext DNA Library Prep Set for SOLiD (New England Biolabs, NEB, Ipswich, MA), per manufacturer’s instructions.

Sequenced reads from a 5500 SOLiD System (LifeTechnologies, Menlo Park, CA) (40 million reads per tissue) were mapped to a modified human genome containing IUPAC ambiguous nucleotide characters for each annotated SNP in dbSNP 135, downloaded from the UCSC Genome Browser, using LifeScope Genome Analysis Software v2.5.1 (Life Technologies, Menlo Park, CA). This method greatly attenuates reference bias alignment, as previously described [119, 120]. Single nucleotide variants were identified with Samtools v0.1.16 [66], which provides a count of the aligned reads containing the reference or variant allele. Identified SNP locations were annotated based on UCSC annotation databases and dbSNP using annovar annotation software [88]. Those polymorphisms confirmed as heterozygous by high-density gDNA genotyping were subsequently included in analyses. Based on annotation, each SNP was assigned to a location within a gene locus whether exonic, intronic, intergenic, UTR, or upstream/downstream (within 1 kb of the coding region). Exonic, UTR, and intronic counts from coding and non-coding genes were used to calculate allelic RNA expression.

**Ethics statement:** The Office of Responsible Research Practices at The Ohio State University has determined that our study does not meet the federal definition of human subjects research under 45 CFR 46.102(f) [also 32 CFR 219.102(f)]. Therefore,
it is waived from further IRB review. This determination is consistent with The Ohio State University Human Research Protection Program (HRPP) policy on human subjects research, found at http://orrp.osu.edu/irb/osupolicies/documents/ResearchInvolvingHumanSubjects.pdf.

2.2 Existing Methods for Observational Studies

Several methods have been proposed for identifying genes with AEI using RNA-seq data. One class of methods focuses on modeling and correcting for bias involved in generating read counts, such as mapping bias favoring the reference alleles [127, 132, 128]. The other class of methods focuses on modeling over-dispersion in read counts, by means of models such as negative-binomial model, Poisson-Gamma model, beta-binomial model, and two-component mixture of beta-binomial model [99, 112, 136, 131, 137]. Our method falls into the second class of AEI detection methods and aims to resolve the two problems described in detail below that are difficult to overcome with other existing methods in the same category.

The first problem arises when modeling AEI signals in genes with very few SNPs (< 10). To the best of our knowledge, existing models are proposed as single-gene-based methods, with each gene’s reads investigated separately. Based on the rule of thumb (via the cross-validation considerations, see [143]) that estimation of each model parameter requires at least ten observations on average, any single-gene-based model with more than one parameter is only applicable to genes with at least ten heterozygous SNPs, or when data from multiple subjects is available. Taking the human brain dataset analyzed in this paper with RNA-seq (308,912 SNPs called from 98 human brain tissues across ten subjects; SNPs with the same rs number in
different brain tissues are counted multiple times), 78% of genes have 4 SNPs or less in the RNA-seq reads. One can extend the single-gene-based models by aggregating the reads within each gene and applying the models to multiple genes. But in that case, genes with different number of SNPs are treated as directly comparable with each other, ignoring uneven SNP numbers within each gene. Here we use mixture model to group SNPs with similar read coverage across many genes, instead of grouping them by genes. Our approach consists of two modeling stages, one for defining comparable SNP groups and the other for detecting AEI signals within each SNP group.

Another issue with the existing methods for AEI detection is that all the binomial-type models assume a strong negative correlation between reference and variant allele reads. In theory, the RNA expression level of the paternal copy of the gene is independent of the maternal one, but because they are subject to the same cellular environment regulation, the expression levels of the two alleles are likely to be highly positively correlated in the absence of cis-acting regulatory variants. Indeed, we observe high correlations between reference and variant read counts in RNA-seq. For instance, in our human autopsy brain tissue dataset discussed below the overall sample correlation between two allele reads is estimated to be 0.92 (see Figure A.1 in Appendix F). Even after excluding a group of SNPs with the highest read counts, we still see linear correlation around 0.71 between reference and variant reads. The assumption that the reference allele reads follow binomial distribution implies that the theoretical correlation between the reference and variant reads is -1, which is opposite to what is observed in RNA-seq data. The approach taken here is more flexible as it does not assume any specific direction of correlation between reference and variant
reads. Note that since our model makes different assumptions than the binomial-type models, it is not easily directly comparable with them via simulation studies.

2.3 Using Folded Skellam Mixture in AEI Analysis

2.3.1 Folded Skellam Mixture Model

The Skellam random variable [1] (and the corresponding distribution) is defined as the difference of two independent Poisson random variables and has various applications, for example in image reconstruction [41], financial mathematics [130], and genetics [129]. The term “folded Skellam” refers to the absolute value of the Skellam random variable. In the following model description, we denote the SNP allele reads from the paternal copy of a gene as $P$ and that from the maternal copy as $M$. Let $R$ and $V$ be the reference and variant reads respectively. Although the parental origin of reads is not available in our RNA-seq data, introducing the hidden pair $(P, M)$ will help us in justifying the model for analyzing $(R, V)$.

One approach to modeling $(P, M)$ is to use some discrete bivariate distribution with certain correlation structure. For example, we can assume $(P, M)$ follows a mixture of bivariate Poisson distributions. Within each mixture component, the correlation between $P$ and $M$ is modelled by introducing an additive Poisson component, i.e.

$$ P = Y_1 + Z, \quad M = Y_2 + Z $$

where $Y_1, Y_2, Z$ are three independent Poisson random variables. However, the bivariate Poisson mixture model may be not ideal for modeling reads from RNA-seq, as it leads to a restrictive requirement that the marginal distributions have to be univariate Poisson mixtures. In order to be more flexible, in our current approach we
only assume that $Y = P - M = Y_1 - Y_2$ follows a Skellam mixture distribution with unknown fixed number of mixture components $K$. That is, we make no distribution assumption on the shared additive component $Z$. Consequently, the joined density of $(P, M)$ is

$$f_{P,M}(p, m | \pi, \Lambda) = \sum_{i=1}^{K} \left[ \sum_{z=1}^{\min(p, m)} \pi_i \text{Poisson} \left( p - z | \lambda_{i,p} \right) \text{Poisson} \left( m - z | \lambda_{i,m} \right) f_{Z_i}(z) \right]$$

where

$$\pi = (\pi_1, \cdots, \pi_K)$$

$$\Lambda = \left( \begin{pmatrix} \lambda_{1,p} \\ \lambda_{1,m} \end{pmatrix}, \cdots, \begin{pmatrix} \lambda_{K,p} \\ \lambda_{K,m} \end{pmatrix} \right)$$

are the model parameters and $\left\{ f_{Z_i}(z) \right\}_{i=1}^{K}$ is a set of unknown probability mass functions. Since we expect to have $|R - V| = |P - M|$ it follows that $|R - V|$ should have the same folded Skellam mixture distribution as $|P - M|$ in our setting. Since the mean of the Skellam variable equals the difference of two corresponding Poisson means, testing the null hypothesis of no AEI signal within a mixture component is equivalent to testing whether the means of two independent Poisson variables are equal. That is, if the component $i$ is a “no AEI signal” component, then under our model $\lambda_{i,p} = \lambda_{i,m} = \lambda$ and we can estimate $\lambda$ by the method of moments using the fact that $E(R - V)^2 = E(|R - V|)^2 = 2\lambda$.

### 2.3.2 Mixture Model Pipeline

AEI is often measured using the ratio of reads aligned to the reference and the variant allele. The ratios in RNA from autosomal genes observed to deviate significantly from unity are considered as AEI signals. The reliability of many currently
applied AEI measures depends on the stringency of the threshold for assigning AEI, and we have previously used allelic differences of 1.5-fold or greater to assign possible AEI [119, 120]. However such arbitrary threshold may not be very efficient in optimizing the missed and false discovery rates for AEI calls. Since the Skellam mixture model described above takes advantage of read counts information across all genes, including those with small number of SNPs (< 10), it is expected to have better ability to detect AEI.

Under the null hypothesis of no AEI signal, we assume that the fluctuations in sequence read differences (between reference and variant alleles) across multiple SNPs are comparable with each other when the sequencing coverage (i.e., the sum of reference and variant allele reads) is of similar magnitude across these SNPs. We refer to such SNPs as “comparable”. Accordingly, we first categorize the comparable SNPs based on the sequencing coverage counts (rescaled after library size adjustments) using a finite mixture of univariate Poisson distributions, and subsequently search for AEI signals within each group of comparable SNPs by fitting a folded Skellam mixture model to the absolute values of rescaled read differences. This approach provides an alternative way of making AEI signal calls in a manner which is more reflective of the noise structure in the RNA-seq data and thus enables considerations of AEI under improved signal to noise ratio, without overly restrictive a priori fold-change thresholds like 1.5, etc.

Although in most genetic applications one does prefer to represent AEI as a read count ratio rather than a read count difference, under our additive interaction model between P and M there is a clear advantage in considering the latter along with the former. To compensate for the relatively noisy raw read counts differences, we propose
to include library-size adjustments of the originally observed read pairs (the reads of reference and variant alleles at the same locus are considered a pair) while preserving the ratios of the raw counts, and group comparable SNPs before modeling the differences of adjusted read counts. The major advantage of using discrete distributions like Poisson and Skellam in our modeling is that we can fit low counts data well, unlike most smoothing techniques and Gaussian-type approximations. This is important, since, for instance, in our human brain dataset 95% of all 10,702 pairs of read counts at identified SNP sites are low counts (<33 reads) (summary statistics are provided in Table A.1 in Appendix F). Below we describe the Skellam-based pipeline for detecting AEI signals in the brain whole transcriptome sequencing datasets.

Step 1: Library size adjustment

To account for differences in the depth at which each tissue sample was sequenced, we multiply each pair of read counts by the ratio of the median total number of reads across all tissue samples to the total number of reads for the specific sample from which the reads are generated. The scatter plots of read pairs, with and without library size adjustment, are presented in Figure A.1 in Appendix F. Note that adjusting for the library sizes does not alter the ratio between two reads in the original dataset.

Step 2: Classifying the sum of read counts

To facilitate AEI signal detection in read pairs with different magnitudes, we first group SNPs according to the sequencing coverage. By treating each gene from subject-specific brain tissue as a unit, we first average the sum of adjusted reads within each unit, and then fit a finite Poisson mixture model to those reads-sum averages. We use the Expectation-Maximization (EM) algorithm for fitting the Poisson mixture [42], and use Bayesian information criterion (BIC) to set the optimal number of mixture
components (i.e. the number of SNP groups). Based on the fitted model (see Table 2.1 on page 23), each of the subject-and-brain-region-specific gene units can be classified into the Poisson mixture components. Therefore, for instance, genes with very few SNPs are grouped with other genes with similar number of averaged total reads.

Step 3: Classifying the differences of read counts

Before analyzing count differences between variant and reference reads, we further divide the set of count pairs within each Poisson mixture component into another four smaller subsets of read pairs according to their location within a gene: 3’ UTR, 5’ UTR, intron, or exon. This step of the algorithm accounts for the fact that the read count differences or ratios from different genetic regions can differ in magnitude. For example, introns are expected to have lower expression than exons. Furthermore, read ratio differences between these regions can occur due to RNA isoforms generated by alternative splicing or different UTR usage at a given gene locus. Accordingly, further statistical analyses are done separately within each subpopulation. For example, we can first evaluate the subset of all adjusted count pairs that are classified into the first Poisson mixture component and also labeled as reads from the 3’ UTR. We use mixture of folded Skellam distributions to model absolute values of these rescaled read differences and classify data into separate folded Skellam components. For fitting the folded Skellam, we used a likelihood-free Markov chain Monte Carlo (MCMC) method [25], which can be also viewed as an Approximate Bayesian Computation (ABC) type of method [124].

Step 4: Testing for signal significance

We define AEI signals as the count pairs being classified into folded Skellam mixture component with significantly different Poisson means. A likelihood ratio testing
(LRT) procedure is used for assessing significant differences in the two parameters of a folded Skellam distribution. Given the subset of count pairs classified into one folded Skellam mixture component, the folded Skellam parameter (equal Poisson means) under the null hypothesis can be estimated using the method of moments (see the previous section on folded Skellam mixture model), and then the log-likelihood of observing such set of differences under the null hypothesis can be calculated accordingly. To evaluate the log-likelihood without the null hypothesis constraints, we used the corresponding parameter estimates obtained in the process of fitting the overall folded Skellam mixture model. The LRT statistics are compared to a chi-square distribution with one degree of freedom.

2.4 Model Fitting Results

To present the potential of decomposing signals from RNA-seq data using the mixture model pipeline, we consider the dataset described above in which we focus only on pairs of counts with at least 3 reads for the allele with lower expression ($\min(R, V) \geq 3$) and exclude intergenic SNPs.

2.4.1 Poisson Mixture Fitting Results

After normalizing the RNA-seq dataset (see pipeline step 1), we fit the Poisson mixture model and find the optimal number of seven components using the BIC criterion. We note that since the Poisson mixture model is expected to reflect the experiment-specific RNA-seq frequency patterns, the particular number of components does not seem to have any meaningful (biological) interpretation. Overall, as long as the mixture model reasonably well fits the data, our downstream analysis is
Table 2.1: Poisson Mixture Model Parameter Estimates and SNPs Classification Results

<table>
<thead>
<tr>
<th>Mixture Component</th>
<th>Proportion</th>
<th>Poisson Mean</th>
<th>No. of SNPs</th>
<th>No. of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp.1</td>
<td>0.030 (0.029, 0.031)</td>
<td>43.11 (42.54, 43.84)</td>
<td>18367</td>
<td>784</td>
</tr>
<tr>
<td>Comp.2</td>
<td>0.0011 (0.0010, 0.0012)</td>
<td>152.37 (146.08, 166.13)</td>
<td>519</td>
<td>37</td>
</tr>
<tr>
<td>Comp.3</td>
<td>0.186 (0.182, 0.190)</td>
<td>20.34 (20.20, 20.49)</td>
<td>82,963</td>
<td>3,892</td>
</tr>
<tr>
<td>Comp.4</td>
<td>0.003 (0.0025, 0.0033)</td>
<td>108.14 (105.13, 115.60)</td>
<td>2,073</td>
<td>89</td>
</tr>
<tr>
<td>Comp.5</td>
<td>0.0006 (0.0004, 0.0008)</td>
<td>201.01 (196.15, 209.71)</td>
<td>425</td>
<td>27</td>
</tr>
<tr>
<td>Comp.6</td>
<td>0.0073 (0.0069, 0.0077)</td>
<td>74.60 (72.56, 78.08)</td>
<td>5,156</td>
<td>202</td>
</tr>
<tr>
<td>Comp.7</td>
<td>0.771 (0.769, 0.775)</td>
<td>7.82 (7.78, 7.85)</td>
<td>198,889</td>
<td>11,174</td>
</tr>
</tbody>
</table>

NOTE: The Poisson mixture model was fitted to the averaged total reads within tissue-specific genes (62,326 tissue-specific genes in total, i.e. sample size=62,326; overall log-likelihood=-21,6846; BIC=43,3836). Genes with the same rs number but from different brain region were considered as different tissue-specific genes. We found the optimal number of mixture components to be 7, meaning that we could classify all SNPs into 7 “comparable” SNP groups. Most SNPs in the gene of our interest (SLC1A3) were classified into the mixture component Comp.1. The SNPs in Comp.1 were used to fit the folded Skellam mixture model.
Table 2.2: Poisson Mixture Comp.1 SNP Counts by Gene Regions

<table>
<thead>
<tr>
<th></th>
<th>3’ UTR</th>
<th>Exon</th>
<th>Intron</th>
<th>5’ UTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of SNPs</td>
<td>10702</td>
<td>4694</td>
<td>2142</td>
<td>269</td>
</tr>
<tr>
<td>No. of Genes</td>
<td>531</td>
<td>405</td>
<td>236</td>
<td>43</td>
</tr>
</tbody>
</table>

NOTE: In total, 18,367 SNPs were classified into the Poisson mixture component 1 and 10,702 of them were in 3’ UTR of 531 genes. Fitting of the folded Skellam mixture model only used the 10,702 SNPs in 3’ UTR.

expected to be robust with respect to the number of components. For practical reasons, we remove the 0.1 percent of the highest average of scaled counts over different gene by tissue categories. Table 2.1 on page 23 presents the results of this fitting procedure. We note that over 90 % of the genes are contained in mixture components Comp.3 and Comp.7. Accordingly, we expect these two components to contain most of the genome-wide signal.

In order to compare our final AEI predictions against those previously reported in the literature in the same dataset [119, 120], we limit ourselves only to the variants in genes from the first Poisson mixture component (Comp.1) and select the genetic location with the highest number of heterozygous positions aligned, namely the 3’UTR, as noted in Table 2.2 on page 24. In many genes, read counts are greatest in the 3’-UTR because of the use of poly-dT primes in addition to random hexamers, facilitating detection of AEI in the 3’-UTR.
NOTE: Histogram of the simulation from the folded Skellam mixture (sample size=10^5). Different mixture components are indicated by different colors. The two mixture components Mix1 and Mix6 which are closest to zero are considered the two no AEI signal components. The right tail (> 50) with relatively smaller frequencies is enlarged and presented in the inner panel.
2.4.2 Folded Skellam Mixture Fitting Results

We fit the folded Skellam mixture model to the adjusted read pairs classified into the first Poisson mixture component, and only use SNPs on the 3’ UTR. After performing classification of these SNPs, we identify two AEI signal components (Mix2 and Mix4) and two no AEI signal components (Mix1 and Mix6) (see Table 2.3 on page 27) by using the LRT (see pipeline step 4). To help visualize the fitted mixture model, we simulated 105 counts from the fitted folded Skellam mixture where we represented different mixture components with different colors (see Figure 2.1) on page 25. The histograms of the observed absolute read differences indicating classification to the mixture components are available in Figure A.2 in Appendix F. The goodness-of-fit analysis for the mixture model was performed by plotting the percentiles of absolute read differences against those of counts simulated from the fitted model. Since the absolute read differences from 10,702 SNPs have a long and sparse tail on the right-hand side (95th percentile is 29 while the maximum is 221), we expect the fit in the tail to be relatively poor. Note that this should not, however, adversely affect the quality of the AEI calls since the large values are most likely to be classified as AEI SNPs anyway. In the context of screening for AEI signal, the key to fitting the folded Skellam mixture is to get accurate fit on data points that are close to zero (i.e., to identify the smallest AEI signal component). Based on the Q-Q plots (see Figure A.3 in Appendix F) we conclude that the fitting is reasonably good up to the 94th percentile of the data.

We do not use LRTs for mixture component Mix3 and Mix5 because there are too few SNPs (5 SNPs in total) being classified into these two components. However, since both Mix3 and Mix5 are even further away from zero than Mix2, which is already
Table 2.3: Folded Skellam Mixture Parameter Estimates And Results of AEI LRTs

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mix1</th>
<th>Mix2</th>
<th>Mix3</th>
<th>Mix4</th>
<th>Mix5</th>
<th>Mix6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi_i )</td>
<td>0.54 (0.54, 0.55)</td>
<td>0.1 (0.10, 0.11)</td>
<td>0.0065 (0.0064, 0.0066)</td>
<td>0.037 (0.036, 0.038)</td>
<td>0.0003 (0.0003, 0.00035)</td>
<td>0.3 (0.3, 0.31)</td>
</tr>
<tr>
<td>( \lambda_{i,1} )</td>
<td>65.7 (65.4, 66.5)</td>
<td>83.8 (82.6, 84.2)</td>
<td>268 (263.3, 269.4)</td>
<td>92.7 (91.4, 93.1)</td>
<td>214.8 (212.2, 216.3)</td>
<td>4.81 (4.75, 4.84)</td>
</tr>
<tr>
<td>( \lambda_{i,2} )</td>
<td>69.2 (69.2, 70.2)</td>
<td>106 (105, 107)</td>
<td>80.3 (79.9, 81.5)</td>
<td>166 (165.9, 169.1)</td>
<td>78.1 (77.0, 78.5)</td>
<td>5.39 (5.29, 5.40)</td>
</tr>
<tr>
<td>( L_0 )</td>
<td>-17,852</td>
<td>-2,074</td>
<td>NA</td>
<td>-650</td>
<td>NA</td>
<td>-7,860</td>
</tr>
<tr>
<td>( L_1 )</td>
<td>-17,864</td>
<td>-1,967</td>
<td>NA</td>
<td>-522</td>
<td>NA</td>
<td>-8,233</td>
</tr>
<tr>
<td>P-value</td>
<td>1</td>
<td>&lt;0.00001</td>
<td>&lt;0.00001</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of SNPs</td>
<td>5,459</td>
<td>482</td>
<td>3</td>
<td>130</td>
<td>2</td>
<td>4,626</td>
</tr>
<tr>
<td>No. of Genes</td>
<td>471</td>
<td>165</td>
<td>3</td>
<td>72</td>
<td>2</td>
<td>407</td>
</tr>
</tbody>
</table>

NOTE: Only SNPs on 3' UTR and classified into Poisson mixture component 1 were used for fitting the folded Skellam mixture (overall log-likelihood = -34,979; BIC = 70,117; sample-size = 10,702; \( (\lambda_{i,1}, \lambda_{i,2}) \) is estimate of the ordered pair \( (\lambda_{i,P}, \lambda_{i,M}) \). NAs indicate insufficient sample sizes for LRTs.
designated as the AEI signal component by LRT, it is reasonable to call Mix3 and Mix5 the AEI signal components as well. Accordingly, we consider 5 SNPs in Mix3 and Mix5 as AEI signal SNPs. Table A.2 in Appendix F lists the raw read counts of these 5 SNPs, along with the mixture probabilities of these 5 SNPs belonging to each of the six folded Skellam distributions, all with relatively high read coverage and absolute ratio of read counts above 2. The mixture probabilities of these 5 SNPs belonging to Mix1 or Mix6 (the two no AEI signal components) are all zero, indicating the significant AEI signals.

Overall, since the two no AEI mixture components contain about 84% of the data, we conclude that the remaining 16% of tested SNPs (1,712 out of 10,702) appear to carry statistically significant AEI signals under the model assumptions. However, by classifying SNPs into folded Skellam mixture components according to the largest mixture probabilities, we only identified 617 AEI signal SNPs out of the total 10,702 “comparable” SNPs, indicating that only about 6% of tested SNPs can be designated as AEI signal with the classification done according to the maximum value of the six mixture probabilities. The remaining 10% cannot be considered as statistically significant AEI signal sources, although according to our model they did display some evidence of AEI.

2.4.3 Mixture Model Pipeline Performance Analysis

To understand better the characteristics of AEI SNPs that stand out in the screening of our mixture model pipeline, and to investigate the relationship between mixture model pipeline and the commonly employed allele ratio threshold, we first tabulate separately the percentiles of absolute read ratios (i.e. $\max(R,V)/\min(R,V)$) for the
617 AEI SNPs and all remaining 10,085 SNPs (in Mix1 and Mix6, mix of 10 % uncertain AEI signal SNPs and no AEI signal SNPs) (see Table 2.4 on page 30). Approximately 90 % of these 617 AEI SNPs have absolute read ratios above 1.54, while 60 % of the 10,085 mixture SNPs have absolute read ratios below 1.54. Since 10,085 mixture SNPs contain approximately 10 % uncertain AEI signal SNPs (1,712 - 617 = 1,095 uncertain AEI SNPs), high absolute read ratios (> 2.5) are also expected in the 10,085 SNPs mixture.

To investigate further the behavior of our mixture model based AEI detection pipeline, we additionally analyze SNPs designated as having AEI despite a low ratio between the alleles and those designated as not having AEI despite a high ratio between the alleles. Among the 617 AEI signal SNPs, there are 51 SNPs with absolute read ratios less than or equal to 1.5 and 9 with absolute read ratios less than or equal to 1.3. In the 10,085 SNPs mixture, 1,003 SNPs have absolute allelic ratio above 2.5, while 10 have absolute read ratios above 7. Detail information of the 9 AEI signal SNPs with the smallest ratio values and the 10 uncertain mixture SNPs with the largest ratio values are listed in Table A.3 and Table A.4 in Appendix F, respectively. None of the 9 AEI signal SNPs has more than 75 % aggregated probability of being in the signal components (Mix2 through Mix5). If the mixture component classifications were done using 80 % probability being in signal components as the criterion, none of the 9 SNPs would be classified as AEI signal SNP. Obviously, the higher required confidence level, the fewer AEI signal SNPs can be identified.

For the uncertain mixture SNPs in Table A.4, the main reason for SNPs with very high read ratios failing our pipeline screening is that the raw read counts are too low. The minimum values of these SNP read pairs are either exactly three (threshold
Table 2.4: Percentiles of Absolute Reads Ratio

<table>
<thead>
<tr>
<th>SNP category</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>617 AEI Signal SNPs</td>
<td><strong>1.54</strong></td>
<td>1.71</td>
<td>1.88</td>
<td>2.08</td>
<td>2.32</td>
<td><strong>2.64</strong></td>
<td>3.06</td>
<td>3.67</td>
<td>4.85</td>
</tr>
<tr>
<td>10,085 SNPs Mixture</td>
<td>1.05</td>
<td>1.13</td>
<td>1.2</td>
<td>1.29</td>
<td>1.4</td>
<td><strong>1.54</strong></td>
<td>1.71</td>
<td>2</td>
<td><strong>2.5</strong></td>
</tr>
</tbody>
</table>

NOTE: Absolute read ratios were calculated using the formula Max(reference, variant) / Min(reference, variant). The 617 AEI signal SNPs were designated according to the largest mixture probability. The remaining 10,085 SNPs included 10% uncertain AEI signal SNPs and 84% no AEI signal SNPs.

for calling a SNPs) or only one or two reads higher. Additionally, some of these small read differences have even smaller library-size-adjusted differences because the corresponding library sizes are above the median level. On the other hand, there are 143 SNPs (see Table available at Download Link 1) out of the total 617 AEI signal SNPs (see Table available at Download Link 2) that have more than 99% probability of carrying AEI signals under the folded Skellam mixture model. For these 143 99% confident AEI signal SNPs, the mean (median) raw reads of reference and variant alleles are 120 (105) and 75 (31) respectively, while the mean (median) read ratio is around 3.36 (3.21). Therefore, in general, SNPs need both high reads ratio and high reads coverage to pass our mixture model based for robust AEI signals.

### 2.5 Investigation of Identified AEI Signals

#### 2.5.1 SNP-level AEI Signals on Gene SLC1A3

Smith et al. (2013b)[120] previously characterized allelic RNA expression using nine brain regions from a single sample from the same dataset (MB011), finding large and consistent allelic differences for multiple genes, including SLC1A3. AEI in
this gene was confirmed using a targeted PCR-based SNaPshot method to measure allelic RNA ratios [120]. Our mixture model pipeline classifies ten subject-and-tissue-specific SNPs on this gene into AEI signal components. Within subject MB059, SNP rs2269272 in SLC1A3 is identified twice as being (with 99 % confidence) AEI signal SNP in two brain regions, insula and amygdala. Within subject MB052, the same SNP (rs2269272) is again identified as AEI SNP with relatively less confidence, but in the same two brain regions (insula and amygdala). Additionally, SNPs rs1049524, rs104922 and rs10428531 in SLC1A3 are also classified as AEI signal SNPs in one or more brain regions in different subjects including MB011, consistent with previous results [120]. Together, these findings argue for the presence of at least one cis-acting regulatory genetic variant that changes expression of SLC1A3 mRNA.

2.5.2 Signal Designation Consistency Across Brain Tissues

Generally speaking, within the same subject, when one SNP locus in one brain region is showing AEI we expect to see the same SNP locus showing AEI signals consistently across most of the other brain regions, unless the regulatory effects are tissue or brain region selective. Using the maximum mixture probability as the criterion, we can compare the number of times that a specific SNP locus is identified as AEI signal across multiple brain regions with the total number of times it is expressed within the same subject. By including only SNPs with read coverage observed in at least two brain regions from the same subjects, we find that there are 114 subject-specific SNPs showing AEI signals in at least half of the brain regions where we have observed expressions. Among these 114 SNPs, over 50 % SNPs show consistent AEI signals in more than one region, while some show consistent AEI signals in all regions that
the gene expresses. For example, SLC24A2 SNP rs7872265 expresses in five brain regions (brain region BA10, BA22, BA24, raphaenucleus, and BA46) and shows AEI in all five regions in MB011. Any inconsistent results in different brain regions may be caused by relative low count coverage in one or more regions and/or lower AEI ratios. We also cannot rule out the possibility of different splice variants or 3’UTR usage in different brains regions, which can confound AEI analysis.

2.5.3 Mixture Model Pipeline vs. Whole Gene Filtering Method

An alternative analysis for the AEI detection known as the whole gene filtering method (described fully in Smith et al., 2013b [120]) was carried out on the same brain tissue samples analyzed above, with some additional replicate sequencing runs. The main differences between the two methods are summarized as follows: 1. The mixture model pipeline scans for AEI signals at the SNP level, while the whole gene filtering method scans for AEI signals at the gene level; 2. For the whole gene filtering method, the read ratios of SNPs in all genetic regions (3’ UTR, exon, intron, and 5’ UTR, etc.) on the same gene are averaged to get a gene-level expression imbalance measurement, while fluctuations in SNPs from different genetic regions are considered non-comparable in the mixture model and modeled separately. 3. SNPs are not called in the whole gene filtering method if the corresponding genes have only one SNP expressed, while these SNPs are still used and classified in the mixture model pipeline as long as both the reference and variant allele read counts are above 3 (the predetermined threshold). Overall in our comparisons the mixture model appears to be more sensitive to identifying AEI signal than the whole gene filtering method, yielding more AEI signal SNPs. For example, the 592 SNPs identified by the mixture
model pipeline with AEI were not identified by the alternative method, likely because their limited coverage or SNP calls across the gene. These 592 instances include 287 unique SNPs present in 175 genes. On the other hand, 90 SNPs identified by the whole gene filtering method failed to be detected in the mixture model pipeline. Interestingly, 84% of these were assigned into the first folded Skellam mixture component (Mix1) indicating that there was a notable difference between allele counts, but not enough evidence for the final AEI designation, possibly caused by low coverage or low AEI signal as discussed above. Since the mixture model method used only SNPs in 3’UTR, while the genome filter method used all SNPs along the expressed gene locus (from 5’ to 3’UTR), the discrepancy could also be caused by different 3’UTR usage or overlapping neighboring genes.

2.5.4 Parallels Between AEI and eQTLs

The goal of AEI analysis is to identify functional regulatory variants, which are speculated to underline many association signals in genome-wide association studies or eQTL analyses. We have used the Genotype-Tissue Expression Project (GTEx) data to test for the potential of the AEI signal SNPs to reveal the presence of eQTLs. The eQTLs were extracted from transcript counts over all tissues and individuals available in the first release of the GTEx data (56 tissues; 216 individuals). We have normalized the transcript read counts using the function estimateSizeFactors’ in the Bioconductor package DESeq’ (http://bioconductor.org/packages/release/bioc/html/DESeq.html), and to make our analysis more robust to low counts, we have summed all transcript reads in a given gene, obtaining a single expression value for each gene across all tissues. Next, we have
stratified individuals by genotype (homozygous major, heterozygous, and homozygous minor) for each SNP with available genotype data (genotyping was performed on Illumina 5 M and Illumina exome chips) - here we did not use imputation to avoid losing statistical power. Finally, we used standard linear regression to test whether the expression level is dependent on the genotype. Of AEI SNPs (in components Mix2 and Mix4) that were directly genotyped 17.6 % (18) reached the standard statistical level of significance (0.05) in the linear regression model (see Table available at https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-015-1749-0/MediaObjects/12864_2015_1749_MOESM10_ESM.doc). Of SNPs without evidence for AEI (in component Mix6), a much lower percentage, 9 % (37), were statistically significant eQTLs. Using the sm package in R (http://www.r-project.org), we compared the distributions of p-values for association with gene expression between AEI and no AEI SNPs. Overall we observed a non-significant trend of lower p-values among AEI SNPs.
Chapter 3: Quantification of Gene Activity Dependency via Sobol Indices

3.1 Sensitivity Analysis

In order to understand the behaviour of a complex system, sensitivity analysis is often performed to investigate what factors (the inputs) contribute to the uncertainty of a variable of interest (the output) and by how much.

3.1.1 Local and Global Sensitivity Measurements

The traditional sensitivity analysis, also called local sensitivity analysis, uses partial derivatives to quantify the uncertainty contributions from each of the input variables when the true mapping function from the inputs to the output is known and smooth enough [32, 56]. The results of local sensitivity analysis depends on the actual input values being specified (the location of interest), and this approach can only examine sensitivity with respect to input variables one at a time (default assuming the inputs vary independently). The local sensitivity analysis has been widely used for exploring the parameter identifiability of biological models, especially in ecologic [5, 34, 50, 62]. Most local sensitivity methods perform eigenvalue decomposition on sensitivity matrix (or the sensitivity matrix transpose times the sensitivity matrix if
the sensitivity matrix is not a square matrix) and then report the eigenvalues after normalizing with respect to the maximum eigenvalue. They are useful for distinguishing identifiable and non-identifiable parameters. To identify redundant or linearly dependent parameters, the approaches based on the singular value decomposition [118] are more efficient.

The global sensitivity analysis [4] uses Sobol’ indices [7, 8, 35, 56] instead of the directional partial derivatives to measure the inputs contribution to the overall output uncertainty over the entire domain of input parameters. The Sobol’ indices, firstly introduced in 1990 [4], provide a unified way of quantifying output’s sensitivity with respect to any subset of input variables. They have become prevalent tools not only for sensitivity analysis [87, 139, 142] but also for other application purposes such as quality assessment of composite indicators [37, 116], variable selection in regression [9, 10, 80], and basis of multiple criteria analysis [93].

3.1.2 Estimation of Sobol Indices with Independent Inputs

Estimation methods of Sobol indices have been studied extensively under the assumption that the inputs variables are independent of each other. If the mapping function from the inputs to the output is known and the independence assumption holds, the Sobol indices can be estimated by generating a large number of input random samples (or quasi-random samples) and then varying inputs (or input subsets) one at a time to obtain the corresponding output values. Different Monte-Carlo estimation formulas have been proposed based on this “vary one at a time” scheme or its variations. This line of work includes Sobol (1990) [4], Sobol (1994) [6], Sobol (2001) [18], Saltelli (2002) [24], Tarantola et al. (2007) [44], Liburne and Tarantola
(2009) [67], Saltelli et al. (2010) [86], Xue et al. (2010) [91], etc. The convergence of Monte-Carlo based estimation methods is discussed in Yang (2011) [103].

There is also an interesting publication on how to estimate Sobol indices using Pearson correlations in a similar Monte-Carlo manner when the true mapping function is known and inputs are essentially independent (may allow small spurious correlation among inputs) [107]. Since the Sobol indices were originally introduced under the high dimensional model representation (HDMR) setup, another group of researchers focused on estimating Sobol indices by constructing the HDMR using different metamodeling techniques, such as random sampling with orthonormal polynomial basis or cubic B spine basis [23, 40], interpolation through model values on cut-HDMR expansion [12], polynomial chaos expansion by Smolyak’s cubature projections [61], Gaussian process metamodel [68], etc.

3.1.3 Estimation of Sobol Indices with Correlated Inputs

In order to perform global sensitivity analysis on systems with correlated inputs, a lot of effort was devoted to approximate the underlying true model in ways that the reconstructions can have certain type of orthogonality among their decomposition components. Relevant ideas include decorrelating the inputs with the Gram-Schmidt procedure before approximate the model [109], decomposing the output variance into partial correlated and uncorrelated components [60, 85], using copula theory to model the correlation structure in polynomial chaos expansion [76], using revised Hoeffding decomposition with projection operators to construct hierarchically orthogonal component functions [105], and applying Gram-Schmidt procedure recursively to construct hierarchical orthogonal basis [138].
If the ultimate goal is only to estimate Sobol indices instead of building a predictive model, it may be more convenient to use multivariate polynomial GLMs as the meta-model to obtain approximated expressions of the conditional expectations of the output given different input subsets separately, instead of approximating the complete mapping from all inputs to the output first and then trying to figure out the partial conditional expectations based on the approximated full map. In this chapter, we will discuss the estimation strategies of Sobol indices under the generalized linear models (GLMs) with the assumption that the inputs are either independent or follow a multivariate normal distribution. For linear GLMs (that is, when the systematic component is a linear function in terms of the inputs), analytic formulas of Sobol indices are derived respectively under the identity link, the log link and the logit link functions. For multivariate polynomial GLMs (that is, when the systematic component is a multivariate polynomial function of the inputs), if the inverse link function is bounded, continuous, and real-valued, a simple estimation strategy based on empirical variance estimates of subspace projections is proposed for estimating Sobol indices with any level of desired accuracy. In addition, if the inputs can be further assumed to be either independent or follow multivariate normal distribution, we can show the proposed estimation strategy also works for polynomial GLMs with identity link. Simulation studies are performed to access the performance of these Sobol index estimates, both in terms of the accuracy and the power, type I error and false discovery rate when used for variable selection. We will finish up the discussion of this chapter with an application example of mapping gene-gene interactions. The importance of gene-gene interactions is ascertained by ranking the candidate
gene subsets according to the Sobol index estimates under multivariate polynomial Gaussian models.

3.2 Generalized Linear Models

The generalized linear models (GLMs) are generalizations of the multivariate linear regression[3]. Instead of modelling the conditional expectation of the response directly, GLMs model transformations of the response conditional expectation. For example, a simple GLM can be specified in the following way:

\[
g(\mu) = g(\mathbb{E}[Y|X]) = X^T \beta
\]

where \( X = (X_1, \ldots, X_n)^T, \beta = (\beta_1, \ldots, \beta_n)^T \). The function \( g(\cdot) \) used for performing the transformation on response expectation is the “link function”. And the regression model on the inputs, i.e. \( X^T \beta \) in this case, is called the “systematic component”.

Another generalization from multivariate linear regression to GLMs is that the error distribution (or the conditional distribution of the response given all the inputs) is no longer restricted to be Normal. Different relationship between the response mean and variance (conditioning on all the inputs) can be modelled by specifying different distributions within the exponential family, such as Poisson, Binomial, Gamma, Exponential, Inverse Gaussian, Negative Binomial, etc.

The univariate exponential dispersion family is a class of distributions that can be written in the following unified form:

\[
f(y; \theta, \phi) = \exp\left\{\frac{[y\theta - b(\theta)]}{a(\phi)} + c(y; \phi)\right\}
\]

where \( \theta \) is called the natural parameter or canonical parameter, \( \phi \) is the scale or dispersion parameter. And \( E(Y) = b'(\theta), Var(Y) = a(\phi)b''(\theta) \) for all univariate
exponential family distributions. The link function $g(\cdot)$ become the “canonical link” if it satisfies $g(\mu) = \theta$. Table 3.1 on page 40 lists the canonical link functions of the commonly used GLMs. In applications, researchers often prefer to use the canonical link functions because they can help simplify the derivation of the maximum likelihood estimates of $\theta$. But other non-canonical links can also be very useful for improving the overall model fitting when the canonical links can not fit the real data well. In real-world applications, GLM is one of the most popular modelling techniques nowadays[22, 79, 115].

**Table 3.1: Canonical Link Functions of Commonly Used GLMs**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Support</th>
<th>Link Name</th>
<th>Canonical Link</th>
<th>Mean Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$(-\infty, +\infty)$</td>
<td>Identity</td>
<td>$\mu = X^T \beta$</td>
<td>$\mu = X^T \beta$</td>
</tr>
<tr>
<td>Bernoulli</td>
<td>{0, 1}</td>
<td>Logit</td>
<td>$\ln \left( \frac{\mu}{1-\mu} \right) = X^T \beta$</td>
<td>$\mu = \frac{\exp(X^T \beta)}{1+\exp(X^T \beta)}$</td>
</tr>
<tr>
<td>Poisson</td>
<td>{0, 1, 2, \ldots}</td>
<td>Log</td>
<td>$\ln(\mu) = X^T \beta$</td>
<td>$\mu = \exp(X^T \beta)$</td>
</tr>
<tr>
<td>Negative Binomial</td>
<td>{0, 1, 2, \ldots}</td>
<td>Log-difference</td>
<td>$\ln \left( \frac{\mu}{\mu+r} \right) = X^T \beta$</td>
<td>$\mu = \frac{r \exp(X^T \beta)}{1+\exp(X^T \beta)}$</td>
</tr>
<tr>
<td>Exponential Gamma</td>
<td>$(-\infty, +\infty)$</td>
<td>Inverse</td>
<td>$\frac{1}{\mu} = X^T \beta$</td>
<td>$\mu = (X^T \beta)^{-1}$</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>$(0, +\infty)$</td>
<td>Inverse-square</td>
<td>$\frac{1}{\mu^2} = X^T \beta$</td>
<td>$\mu = \left(\sqrt{X^T \beta}\right)^{-1}$</td>
</tr>
</tbody>
</table>

**NOTE:** $\mu$ is the expected value of response $Y$; $r$ is the parameter denoting the number of failures pre-specified in Negative Binomial distribution.
3.3 Sobol Indices under GLMs

3.3.1 Variance-based Definition of Sobol Indices

In the literature, there are two equivalent ways of defining the Sobol sensitivity indices. One is based on the integrals of HDMR component functions \([18, 35, 138]\). The other is based on variances of conditional expectations of the response given input subsets \([24, 56]\). To facilitate our discussion under the GLMs with multivariate normal inputs, we will work with the variance-based definitions stated as follows:

**Definition 3.3.1.** Suppose \(Y\) is a univariate random variable, and its mean is determined by a set of random variables \(X = (X_1, \ldots, X_n)^T\). Then the main-effect Sobol index of \(Y\) with respect to \(X_i\) is:

\[
S_i = \frac{Var(E(Y|X_i))}{Var(Y)} \tag{3.1}
\]

The total-effect Sobol index of \(Y\) with respect to \(X_i\) is:

\[
S^T_i = \frac{E(Var(Y|X_{-i}))}{Var(Y)} = 1 - \frac{Var(E(Y|X_{-i}))}{Var(Y)} \tag{3.2}
\]

where \(X_{-i}\) denotes the vector of inputs excluding \(X_i\). Similarly, the main-effect Sobol index of \(Y\) with respect to input subset \(X_P = \{X_{i_1}, X_{i_2}, \ldots, X_{i_p}\}\) is:

\[
S_P = \frac{Var(E(Y|X_{i_1}, X_{i_2}, \ldots, X_{i_p}))}{Var(Y)} \tag{3.3}
\]

and the corresponding total-effect index is

\[
S^T_P = \frac{E(Var(Y|X_{-p}))}{Var(Y)} = 1 - \frac{Var(E(Y|X_{-p}))}{Var(Y)} \tag{3.4}
\]

where \(X_{-p}\) is the subset of inputs excluding \(X_P\).
In addition, various interaction-effect indices can be defined by subtracting lower order main-effect indices from the higher order main-effect indices. For example, the two-way interaction index between $X_1$ and $X_2$ is:

$$S_{1,2} = \frac{\text{Var}(E(Y|X_1, X_2))}{\text{Var}(Y)} - \frac{\text{Var}(E(Y|X_1))}{\text{Var}(Y)} - \frac{\text{Var}(E(Y|X_2))}{\text{Var}(Y)}$$

Higher order interaction-effect indices are defined in similar manners. Since both the total-effect indices and the interaction-effect indices can be viewed as functions of main-effect indices, the key to Sobol index estimation is to estimate main-effect indices which comes down to evaluating the variances of conditional expectations with respect to different input subsets. If the conditional expectations can be written out in explicit expressions of the inputs being conditioned on, we can then estimate the Sobol indices using empirical estimates of these expressions.

The following is a simple example for helping understand the definitions of two types of Sobol indices:

**Example 3.3.1.** Suppose we have a multivariate linear regression model $E[Y|X] = \beta_0 + X^T \beta$ and the inputs are independent. Then the main-effect Sobol index of $Y$ with respect to $X_i$ is:

$$S_i = \frac{\text{Var}(E(Y|X_i))}{\text{Var}(Y)} = \frac{\beta_i^2}{\text{Var}(Y)} \cdot \text{Var}(X_i)$$

The total-effect Sobol index of $Y$ with respect to $X_i$ is:

$$S_i^T = \frac{E(\text{Var}(Y|X_{-i}))}{\text{Var}(Y)} = 1 - \frac{\text{Var}(E(Y|X_{-i}))}{\text{Var}(Y)} = 1 - \frac{\sum_{j \neq i} \beta_j^2 \cdot \text{Var}(X_j)}{\text{Var}(Y)}$$

where $X_{-i}$ denotes the vector of inputs excluding $X_i$. Similarly, the main-effect Sobol index of $Y$ with respect to input subset $X_P = \{X_{i_1}, X_{i_2}, \ldots, X_{i_p}\}$ is:

$$S_P = \frac{\text{Var}(E(Y|X_{i_1}, X_{i_2}, \ldots, X_{i_p}))}{\text{Var}(Y)} = \frac{\sum_{j \in P} \beta_j^2 \cdot \text{Var}(X_j)}{\text{Var}(Y)}$$
and the corresponding total-effect index is

\[ S_P^T = \frac{E(Var(Y|X_P))}{Var(Y)} = 1 - \frac{Var(E(Y|X_{-P}))}{Var(Y)} = 1 - \frac{\sum_{j \notin P} \beta_j^2 \cdot Var(X_j)}{Var(Y)} \]

where \( X_{-P} \) is the subset of inputs excluding \( X_P \).

### 3.3.2 Sobol Indices under Linear GLMs

In the following three results, analytic expressions of main-effect Sobol indices are presented under linear GLMs with identity link, log link, and logit link respectively, while assuming the inputs are independent random variables with unknown distributions or follow multivariate Normal distributions. With Dr. Min Wang’s help, the formulas for estimating Sobol indices under GLMs with identity, log, logit link and linear systematic component are implemented in the R package SobolSensitivity. The installation files are available for downloading at http://R-Forge.R-project.org. The proofs of these results are provided in Appendix C.

#### Result 3.3.1. Sobol Indices under Linear GLMs with Identity Link

Suppose \( E[Y|X] = \beta_0 + X^T \beta \). If the inputs \( X \) follow a multivariate normal distribution \( N(\mu, \Sigma) \) where \( \mu = (\mu_1, \mu_2, \ldots, \mu_n)^T \), \( \Sigma_{ii} = \sigma_i^2, \Sigma_{ij} = \rho_{ij} \sigma_i \sigma_j \), then the main-effect Sobol index with respect to \( X_i \) has the following closed form:

\[ \frac{Var(E(Y|X_i))}{Var(Y)} = \left( \beta_i + \frac{1}{\sigma_i} \sum_{j \neq i} \beta_j \rho_{ij} \sigma_j \right)^2 \frac{Var(X_i)}{Var(Y)} \quad (3.5) \]

Let \( X_P = (X_{i_1}, \ldots, X_{i_p})^T \) be a subset of inputs, and \( X_Q \) be the input vector containing the remaining X’s. Then the main-effect Sobol index with respect to input subset \( X_P \) has the following closed form:

\[ \frac{Var(E(Y|X_P))}{Var(Y)} = \frac{\eta^T \Sigma_{PP} \eta}{Var(Y)} \quad (3.6) \]
where
\[ \eta = \beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q \]
and \[
\begin{bmatrix}
\Sigma_{PP} & \Sigma_{PQ} \\
\Sigma_{QP} & \Sigma_{QQ}
\end{bmatrix}
\]
is the partition of \( \Sigma \) corresponding to input vector partition \( \mathbf{X} = (\mathbf{X}_P^T, \mathbf{X}_Q^T)^T \).

Result 3.3.2. **Sobol Indices under Linear GLMs with Log Link**

Suppose \( \ln(E[Y|X]) = \beta_0 + \mathbf{X}^T \mathbf{\beta} \). If the inputs \( \mathbf{X} \) follow a multivariate normal distribution \( N(\mu, \Sigma) \) where \( \mu = (\mu_1, \mu_2, \cdots, \mu_n)^T \), \( \Sigma_{ii} = \sigma_i^2, \Sigma_{ij} = \rho_{ij} \sigma_i \sigma_j \), the main-effect Sobol index with respect to \( X_i \) has the following closed form:

\[
\frac{\text{Var}(E(Y|X_i))}{\text{Var}(Y)} = \frac{1}{\text{Var}(Y)} \left( e^{\sigma_i^2} - 1 \right) e^{2\beta_0 + 2W(i) + 2\mu_0 + \sigma^2} \quad (3.7)
\]

where
\[
\mu_* = \left( \beta_i + \frac{\sum_{j \neq i} \beta_j \rho_{ij} \sigma_j}{\sigma_i} \right) \mu_i, \quad \sigma_*^2 = \left( \beta_i + \frac{\sum_{j \neq i} \beta_j \rho_{ij} \sigma_j}{\sigma_i} \right)^2 \sigma_i^2
\]
\[
W(i) = \sum_{j \neq i} \beta_j \left( \mu_j - \mu_i \rho_{ji} \sigma_j \sigma_i \right) + \frac{1}{2} \beta_j^T \left( \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{PQ} \right) \beta_j - \frac{1}{2} \beta_i^T \left( \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{PQ} \right) \beta_i
\]

\( \beta_{-i} = (\beta_1, \beta_2, \cdots, \beta_{i-1}, \beta_{i+1}, \cdots, \beta_n)^T \) and \[
\begin{bmatrix}
\Sigma_{PP} & \Sigma_{PQ} \\
\Sigma_{QP} & \Sigma_{QQ}
\end{bmatrix}
is the partition of \( \Sigma \) corresponding to the input vector partition \( \mathbf{X} = (\mathbf{X}_P^T, \mathbf{X}_Q^T) \).

Let \( \mathbf{X}_P = (X_{i_1}, \cdots, X_{i_p})^T \) be a subset of inputs, and \( \mathbf{X}_Q \) be the input vector containing the remaining \( \mathbf{X} \)'s. Then the main-effect Sobol index with respect to input subset \( \mathbf{X}_P \) has the following closed form:

\[
\frac{\text{Var}(E(Y|X_P))}{\text{Var}(Y)} = \frac{1}{\text{Var}(Y)} \left( e^{\sigma_*^2} - 1 \right) e^{2\beta_0 + 2W(P) + 2\mu_* + \sigma_*^2} \quad (3.8)
\]

where
\[
\mu_* = \mu_P^T \left( \beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q \right),
\]
\[
\sigma_*^2 = (\beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q)^T \Sigma_{PP} (\beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q)
\]
\[ W^{(P)} = (\mu_Q - \Sigma_{QP} \Sigma_{PP}^{-1} \mu_P)^T \beta_Q + \frac{1}{2} \beta_Q^T (\Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{QP}) \beta_Q \]

and \[ \begin{bmatrix} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{bmatrix} \]

is the partition of \( \Sigma \) corresponding to the input vector partition \( X = (X_P^T, X_Q^T)^T \).

If the inputs are independent random variables with unknown distributions, the main-effect Sobol indices with respect to single input can be estimated using the following expression:

\[
\frac{Var(E(Y|X_i))}{Var(Y)} = \frac{1}{Var(Y)} \left[ \exp(\beta_0) \prod_{j \neq i} E[\exp(\beta_j X_j)] \right]^2 Var[\exp(\beta_i X_i)] \tag{3.9}
\]

since it’s easy to obtain empirical estimates of \( E(\exp(\beta_j X_j)) \) and \( Var[\exp(\beta_i X_i)] \) given an input sample. Similarly, the main-effect Sobol indices with respect to multiple inputs can be estimated using:

\[
\frac{Var(E(Y|X_P))}{Var(Y)} = \frac{1}{Var(Y)} \left[ \exp(\beta_0) \prod_{j \notin P} E[\exp(\beta_j X_j)] \right]^2 Var\left[\exp\left(\sum_{i \in P} \beta_i X_i\right)\right] \tag{3.10}
\]

Result 3.3.3. Sobol Indices under Linear GLMs with Logit Link

If \( \ln\left(\frac{E[Y|X]}{1-E[Y|X]}\right) = \beta_0 + X^T \beta \) and the inputs follow a multivariate normal distribution \( N(\mu, \Sigma) \) where \( \mu = (\mu_1, \mu_2, \ldots, \mu_n)^T, \Sigma_{ii} = \sigma_i^2, \Sigma_{ij} = \rho_{ij} \sigma_i \sigma_j \), the main-effect Sobol
index with respect to $X_i$ is:

\[
\frac{\text{Var}(E(Y|X_i))}{\text{Var}(Y)} = \frac{1}{\text{Var}(Y)} \text{Var} \left\{ e^{-\frac{\hat{\mu}^2}{2\hat{\sigma}^2}} \left[ (-1)^{s-1} \frac{1}{2} + \sum_{k=1}^{s-1} (-1)^{k-1} e^{\frac{1}{2}(s-k)^2\hat{\sigma}^2} \right] \right\},
\]

(3.11)

\[
s = \begin{cases} 
1 + \frac{\hat{\mu}}{\hat{\sigma}^2}, & \text{if } \frac{\hat{\mu}}{\hat{\sigma}^2} \in \mathbb{Z}^+ \\
-\frac{\hat{\mu}}{\hat{\sigma}^2}, & \text{if } \frac{\hat{\mu}}{\hat{\sigma}^2} \in \mathbb{Z}^-
\end{cases}
\]

where

\[
Z \sim \ln N(0, \hat{\sigma}^2)
\]

\[
\tilde{\mu} = E(X^T\beta|X_i) = \left( \beta_i + \sum_{j \neq i} \beta_j \rho_{ij} \frac{\sigma_j}{\sigma_i} \right) X_i + \sum_{j \neq i} \beta_j \left( \mu_j - \mu_i \rho_{ij} \frac{\sigma_j}{\sigma_i} \right)
\]

\[
\tilde{\sigma}^2 = \text{Var} \left( X^T\beta|X_i \right) = \beta_i^T \Sigma_{P|Q} \beta_i
\]

\[
\beta_{-i} = (\beta_1, \beta_2, \cdots, \beta_{i-1}, \beta_{i+1}, \cdots, \beta_n)^T, \quad \Sigma_{P|Q} = \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{QP}; \quad \left[ \begin{array}{cc} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{array} \right]
\]

is the partition of $\Sigma$ corresponding to input vector partition $X = (X_p = X_i, X_q = X_{-i})^T$.

Let $X_p = (X_{i_1}, \cdots, X_{i_p})^T$ be a subset of inputs, and $X_q$ be the input vector containing the remaining $X$'s. Then the main-effect index with respect to input subset $X_p$ has the same form as expression (3.11) after replacing $\tilde{\mu}$ and $\tilde{\sigma}^2$ with the following $\tilde{\mu}$ and $\tilde{\sigma}^2$:

\[
\tilde{\mu} = E \left( X^T\beta | X_p \right) = \beta_0 + X_p (\beta_p + \Sigma_{PP}^{-1} \Sigma_{QP} \beta_Q) + (\mu_Q - \Sigma_{QP} \Sigma_{PP}^{-1} \mu_P) \beta_Q
\]

\[
\tilde{\sigma}^2 = \text{Var} \left( X^T\beta | X_p \right) = \beta_p^T (\Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{QP}) \beta_Q
\]
Under the logit link, the idea is to provide a recursive formula to reduce the computational burden of evaluating the conditional expectations with respect to different input combinations. For Sobol indices corresponding to integer valued $\tilde{\sigma}^2$ or $\tilde{\tilde{\sigma}}^2$, we only need to evaluate the sample variance of the conditional expectation expression provided above to get exact estimates of these indices. For Sobol indices with non-integer valued $\tilde{\mu}$ or $\tilde{\tilde{\mu}}$, the efficiency of estimating Sobol indices using above formula largely depends on the efficiency of evaluating $E\left[\frac{Z^{s-\lfloor s\rfloor}}{1+Z}\right]$. If a table contain values of $E\left[\frac{Z^{s-\lfloor s\rfloor}}{1+Z}\right]$ evaluated at different $s - \lfloor s \rfloor$ is generated in advance, estimating Sobol indices using the above formula can be fairly quick. The simplest way of generating such table is to calculate empirical mean from simulation samples of $\frac{Z^{s-\lfloor s\rfloor}}{1+Z}$, or use numerical integration techniques. Meanwhile, since $s - \lfloor s \rfloor$ is bounded between 0 and 1, so the $E\left[\frac{Z^{s-\lfloor s\rfloor}}{1+Z}\right]$ table needed to achieve a reasonable estimation accuracy will not be huge. Note that in this result, $\tilde{\sigma}^2$ and $\tilde{\tilde{\sigma}}^2$ are actually constants. But, $\tilde{\mu}$ and $\tilde{\tilde{\mu}}$ are random, because they are functions of inputs. Given a sample of the inputs and the estimated regression coefficients, we can obtain empirical estimates of the Sobol indices by evaluating the expression of conditional mean at each observed value of $X_i$ or $X_P$ and then calculating the sample variance of this expression.

The reason why there are no formulas derived for independent inputs in this case is that under the logit link it’s impossible to write out an analytic expression of $E(Y|X_i)$ as a function of $X_i$ when the distributions of $X_{-i}$ are unknown. Under the logit link and independence assumption,

$$\frac{Var(E(Y|X_i))}{Var(Y)} = E^2\left[\frac{\exp\left(\sum_{j\neq i} \beta_j X_j\right)}{Var(Y)}\right] Var\left[ E\left[ \frac{\exp(\beta_i X_i)}{1 + \exp\left(\sum_{j\neq i} \beta_j X_j\right) \exp(\beta_i X_i)} \right| X_i \right] \right]$$
where $E \left[ \frac{\exp(\beta_i X_i)}{1 + \exp(\sum_{j \neq i} \beta_j X_j) \exp(\beta_i X_i)} \big| X_i \right]$ cannot be written as a function of $X_i$ if distribution of $\sum_{j \neq i} \beta_j X_j$ is unknown. But this does not mean the Sobol indices cannot be estimated in such scenarios. In the next section, a simple strategy is proposed for estimating Sobol indices under any polynomial GLMs (including linear GLMs with Logit link). Given sufficient amount of data, estimates obtained by the this proposed strategy can achieve any level of desired accuracy, if the inputs follow a multivariate normal distribution or independently follow some other distributions.

### 3.3.3 Sobol Indices under Polynomial GLMs

Now, instead of GLMs with linear systematic components, we will discuss the estimation of Sobol indices under GLMs of which the systematic components are multivariate polynomial functions of the input variables. Proofs of the following two results are presented in Appendix D.

**Result 3.3.4. Sobol Indices under Polynomial GLMs with Identity Link.** Suppose the conditional expectation of response $Y$ with respect to all inputs $X = (X_1, \ldots, X_n)$ is a multivariate polynomial function with degree $K \in \mathbb{Z}^+$,

$$E [Y \mid X] = \text{Poly}^{(K)} (X, \beta) = \sum_{|k| \leq K} \beta_k X^k$$

where $k = (k_1, k_2, \ldots, k_n) \in \mathbb{Z}^n$. If $X$’s are independent random variables with unknown distributions or follow multivariate normal distributions, we can show that

$$\forall \ X_P = (X_{i_1}, X_{i_2}, \ldots, X_{i_p}), 1 \leq p \leq n$$

$$E [Y \mid X_P] = \text{Poly}^{(K')} (X_P, \beta'), \ 1 \leq K' \leq K$$

(3.12)

which means estimation of the exact Sobol indices with respect to any input subset $X_P$ only requires fitting the smaller model (3.12) and then evaluating the empirical
variance of $\text{Poly}^{(K')} (X_P, \beta')$ and $Y$:

$$\frac{\text{Var}(E(Y|X_P))}{\text{Var}(Y)} = \frac{\text{Var}\left[\text{Poly}^{(K')} (X_P, \beta')\right]}{\text{Var}(Y)}$$

(3.13)

For example, if the true model is a multivariate Gaussian regression with 3 independent inputs and $E[Y|X_1, X_2, X_3] = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_{1,3} X_1^2 X_3 + \beta_3 X_3^3$, then to estimate Sobol index $S_1$, we only need to fit $Y$ as a quadratic function of $X_1$. The $R^2$ (coefficient of determination [126]) of regression model $\text{Poly}^{(2)} (X_1) = \beta'_0 + \beta'_1 X_1 + \beta'_2 X_1^2$ is exactly the empirical estimate of the main-effect Sobol index with respect to $X_1$ under the true model; $1 - R^2$ of the $\text{Poly}^{(2)} (X_1)$ is the empirical estimate of total-effect Sobol index with respect to $(X_2, X_3)$ under the true model. The $R^2$ of a $\text{Poly}^{(2)} (X_2) = \beta''_0 + \beta'_1 X_2^2$ is the empirical estimate of main-effect Sobol index with respect to $X_2$ under the true model; the total-effect Sobol index with respect to $(X_1, X_3)$ under the true model can be estimated by $1 - R^2$ of the $\text{Poly}^{(2)} (X_2)$. The $R^2$ of a $\text{Poly}^{(3)} (X_3) = \beta'''_0 + \beta'_1 X_3^3$ is the empirical estimate of main-effect Sobol index with respect to $X_3$ under the true model; the total-effect Sobol index with respect to $(X_1, X_2)$ under the true model can be estimated by $1 - R^2$ of the model $\text{Poly}^{(3)} (X_3)$. The empirical estimate of the interaction-effect Sobol index with respect to $(X_1, X_3)$ can be obtained by subtracting the $R^2$ of $\text{Poly}^{(2)} (X_1)$ and the $R^2$ of $\text{Poly}^{(3)} (X_3)$ from the $R^2$ of $\text{Poly}^{(3)} (X_1, X_3)$.

If instead of a Gaussian model, we have a Poisson model $E[Y|X_1, X_2, X_3] = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_{1,3} X_1^2 X_3 + \beta_3 X_3^3$ (identity link) with multivariate normal inputs, then the interaction-effect Sobol index with respect to $(X_1, X_3)$ is exactly the largest main-effect index under the Poisson model $E[Y|X_1, X_3] = \text{Poly}^{(3)} (X_1, X_3)$.
(identity link) subtracting the largest main-effect index under the Poisson model
\[ E[Y|X_1] = \text{Poly}^{(2)}(X_1) \] and the largest main-effect index under Poisson model
\[ E[Y|X_3] = \text{Poly}^{(3)}(X_3). \] Since we can always estimate the largest main-effect index of any GLM with identity link empirically by calculating the sample variance of the systematic component, Result 3.3.4 provides a unified way of estimating all the Sobol indices under any polynomial GLMs with identity link and independent or multivariate normal inputs.

This estimation strategy also potentially gives us more power in real-world applications, because this method does not always require reconstructing the complete input-output map using the limited data. If the goal is just to rank the inputs one by one, we argue that only the univariate models need to be fitted or approximated, which in theory should be a much simpler task comparing to reconstructing the full map. We note that for most real-world problems, we can hardly claim any model we fit is exact, or the inputs we observe contain all the variables affecting the outputs anyway.

Next, we show that for some link functions other than the identity link, the Sobol indices can still be estimated using a similar strategy to that proposed above. Give sufficient amount of data, the method presented below allows us to provide approximations of Sobol indices with any level of desired accuracy.

Result 3.3.5. **Approximation of Sobol Indices under Polynomial GLMs with Non-identity Inverse Link.** Suppose \( X = (X_1, \cdots, X_n) \) are defined on a compact space \( D^n \subset \mathbb{R}^n \), and the conditional expectation of the response after being transformed by the link function is a multivariate polynomial function of the input
variables with degree $K \in \mathbb{Z}^+$,

$$g(E[Y|X]) = \text{Poly}^{(K)}(X, \beta) = \sum_{|k| \leq K} \beta_k X^k$$

where $k = (k_1, k_2, \cdots, k_n) \in \mathbb{Z}^n$. Let $X_P = (X_{i_1}, X_{i_2}, \cdots, X_{i_p}), 1 \leq p \leq n$ be a subset of $X$. If the inverse link function $g^{-1}(\cdot)$ is Lipschitz-continuous and

$$E[Y|X_P] = E\left[g^{-1}\left(\text{Poly}^{(K)}(X, \beta)\right)\bigg| X_P\right] < \infty, \quad \forall X_P \in D^p$$

and is also Lipschitz-continuous, by directly applying the Stone-Weierstrass Theorem [2, 122] we know that $\forall \epsilon > 0$, we may find a $\text{Poly}^{(K')} (X_P)$, s.t.

$$\left| E[Y|X_P] - \text{Poly}^{(K')} (X_P) \right| \leq \epsilon, \quad \forall X_P \in D^p \subset R^p$$

Based on the above result, we know that if $E[Y|X_P]$ is Lipschitz-continuous and the Sobol indices exist, for any fixed accuracy level $\epsilon > 0$ there exists a polynomial function of $X_P$ can approximate $E(Y|X_P)$ with the required accuracy at all observed values of $X_P$. Therefore, as long as the sample is big enough, we can obtain a numerical approximation of $\frac{\text{Var}(E(Y|X_P))}{\text{Var}(Y)}$ for any level of desired accuracy only by fitting $E[Y|X_P] = \text{Poly}^{(K')} (X_P)$. It also worth noting that the above result is applicable to systems with inputs from any multivariate distribution defined on a compact space, as long as $E[Y|X_P]$ exist and is Lipschitz-continuous. The proof of the above result also does not require the inverse link to be differentiable.

If we consider the inputs are independent uniform random variables defined on closed intervals, or independent two-way truncated normal random variables, there are many commonly used functions satisfy the requirement of the link function in Result 3.3.5, such as the logit, the probit, the arcsin, the arccos, the inverse hyperbolic
tangent, the inverse hyperbolic secant, etc. Since inverse of these functions are all bounded and Lipschitz-continuous on any closed real interval, \( E[Y|X_P] \) is guaranteed to exist. Since the inputs’ density is Lipschitz-continuous and defined on compact space, \( E[Y|X_P] \) is guaranteed to be Lipschitz-continuous as well.

We can also construct other more complicated link functions so that they can satisfy the condition and also have appropriate range and domain for the distribution family being specified. For example, we can build a GLM with Binomial family and the inverse link being \( \frac{e^{\sin(x)} - e^{-1}}{e - e^{-1}} \) or \( \frac{e^{\cos(x)} - e^{-1}}{e - e^{-1}} \). Although for some of these complicated links, it might not be feasible to fit the corresponding GLMs directly using the software that is currently available, the point is when we estimate the Sobol indices using the proposed method, no specific form of the true complete map is assumed. The true complete model can be GLMs with any link function that guarantees the existence and Lipschitz-continuity of \( E[Y|X_P] \) (the lower-dimensional projections).

### 3.3.4 Multiple Testing of Sobol Indices

If we want to use Sobol indices for feature selection, it’s necessary to discuss the estimation of type I error, power, and false discovery rate in the framework of multiple testing. For example, we can use the same definitions of type I error, power, and false discovery rate that are stated in Section 18.7.1 of Friedman et. al. (2001) [16]. Suppose \( M \) Sobol indices are being tested for significance simultaneously. The null hypothesis of each test is that the Sobol index equals zero. Suppose the outcomes of the \( M \) tests are summarized in the way as shown in Table 3.2. The type I error of
Table 3.2: Outcome Summary of $M$ Significance Tests

<table>
<thead>
<tr>
<th></th>
<th>Called Not Significant</th>
<th>Called Significant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ True</td>
<td>U</td>
<td>V</td>
<td>$M_0$</td>
</tr>
<tr>
<td>$H_1$ True</td>
<td>T</td>
<td>S</td>
<td>$M_1$</td>
</tr>
<tr>
<td>Total</td>
<td>M-R</td>
<td>R</td>
<td>M</td>
</tr>
</tbody>
</table>

The type I error corresponding to this set of decision rules can be estimated by sample mean of $V$ divided by $M_0$. The corresponding power estimate is the sample mean of $S$ divided by $M_1$. The corresponding FDR can be estimated by the sample mean of $V/R$.

In simulation studies, given a set of decision rules for the $M$ tests, we can simulate large number of response samples independently from the inputs and perform the $M$ significance tests using this set of decision rules repeatedly on each sample. Then the type I error corresponding to this set of decision rules can be estimated by sample mean of $V$ divided by $M_0$. The corresponding power estimate is the sample mean of $S$ divided by $M_1$. The corresponding FDR can be estimated by the sample mean of $V/R$.

However, in observational studies, we often don’t know the underlying true model, and thus do not have the ability of simulating large number of response samples from the true model to estimate FDR in the same way as described above. So the idea for handling observational data is to use permutation samples instead of the simulation samples. The FDR estimate obtained from the permutation samples is called the “Plug-in” estimate in Algorithm 18.3 in Friedman et. al. (2001) [16]. This algorithm can be rewritten in terms of Sobol indices as follows:
Algorithm 1. The Plug-in Estimate of the False Discovery Rate

1. Calculate $M$ Sobol indices, $S_j$, $j = 1, \ldots, M$ based on the observed data.

2. Create $K$ permutations of the observed data, by fixing the inputs and permuting the response $K$ times.

3. Calculate $M$ Sobol indices, $S^k_j$, $j = 1, \ldots, M$ for each permutation sample $k = 1, \ldots, K$.

4. Given $M$ thresholds $C_j$, $j = 1, \ldots, M$ for the $M$ Sobol indices, calculate:

   $$R_{obs} = \sum_{j=1}^{M} I(S_j > C_j), \quad \hat{E}(V) = \frac{1}{K} \sum_{j=1}^{M} \sum_{k=1}^{K} I(S^k_j > C_j)$$

5. The plug-in estimate of FDR is $\hat{FDR} = \hat{E}(V)/R_{obs}$

end

3.4 Simulation Studies

In the following simulation studies, we will assess the accuracy of Sobol index estimates obtained by the empirical variances of separate lower dimensional projections, and compare the performance of Sobol indices with other variable selection methods. Since all Sobol indices have the same denominator $Var(Y)$, we will only estimate the numerators of Sobol indices and use them to rank the importance of the inputs.
3.4.1 Simulations under Gaussian Models

3.4.1.1 Simulation Setup

We first simulate a sample of 40 input variables from a multivariate normal (MVN) distribution with the mean values generated from the uniform distribution with domain \([-50,50]\), the marginal standard deviations generated from the uniform on \([1,10]\), and all pairwise correlations set as 0.8. The sample size is 1000. Then we generate the true regression coefficients of these inputs from the uniform distribution defined on \([-1, -0.1] \cup [0.1, 1]\). To make sure all the inputs are important to the response, values near zero are avoided when generating the true regression coefficients. Using the inputs and the coefficients generated above, the responses are simulated from normal distribution with different mean calculated according to \(E(Y|X) = \beta_0 + X^T\beta\), and fixed standard deviation that is one randomly draw from uniform on \([3.5, 5]\).

In order to test the performance of the estimation methods, another 20 fake input variables are also simulated from the same multivariate normal distribution that is used to generated the 40 true inputs. The fake inputs are considered “fake” because they are not used in generating the response values, and thus have no relation to the response variable. They are only similar to the true inputs in the sense that both the true inputs and the fake inputs were draw independently from the same multivariate normal distribution.
3.4.1.2 Accuracy Assessment of Sobol Index Estimates

For linear Gaussian regression we have derived exact analytic expressions of Sobol indices in section 3.3.2, which can be used to check the accuracy of the Sobol index estimates obtained by evaluating the empirical variances of lower dimension projections of the response on partial inputs.

For example, given one simulation sample, we can first estimated the exact first-order main effect Sobol indices by evaluating formula (3.5) with coefficients estimates obtained from fitting the correct full model (multivariate linear regression containing all 40 true inputs). To obtain the corresponding empirical variance estimates, we need to fit separate univariate linear regression in terms of each input, and then calculate the empirical variances of each fitted univariate linear function. The exact theoretical Sobol indices (obtained by using formula (3.5)) are plotted in Figure 3.1 panel (a) (without scaling by the response variance). In this case, since the correct form of the model is known, there are no Sobol indices estimated for the fake inputs (inputs with id from 41 to 60). When we estimate Sobol indices using empirical variances of lower dimension projections, we don’t have to know which inputs are involved in the true model. We can obtain Sobol index estimates for all 60 inputs (including the 40 true inputs actually used to generate response, plus 20 fake ones that have no relation to the response), by fitting 60 univariate linear regression models and then calculate the sample variances of each of these 60 univariate linear functions. These Sobol index estimates are plotted in Figure 3.1 panel (b) (without scaling by the response variance as well). By comparing panel (a) and (b), we can see that the empirical variance estimates of Sobol indices are very accurate. If a input is not included in the
underlying true mode (the fake input), its empirical variance estimate of Sobol index should be very close to zero.

According to Result 3.3.4, we know that if the true model is a multivariate linear regression, any lower dimensional projection (the conditional expectation with respect to any input subset) is also a multivariate linear function of the partial inputs. Therefore, under the multivariate linear models with MVN inputs setup, formula (3.5) will always return correct Sobol index estimates regardless whether the model is correctly specified or not, or in other words, regardless whether the fitted model contain all true inputs or not.

To confirm this in our simulation, we can pretend that the true model is mistaken to be a multivariate regression of the first 20 true inputs and the 20 fake inputs. The other 20 true inputs (with input id from 21 to 40) are not observed in data collection. So we obtain another set of Sobol index estimates by fitting a wrong multivariate regression and then evaluate formula (3.5) with 40 incorrect coefficient estimates obtained from fitting this wrong model. Sobol index estimates obtained this way are plotted in Figure 3.1 panel (c). By comparing panel (a) and (c), we can see that the Sobol indices estimated under the incorrect multivariate model also turn out to be fairly accurate. And the indices estimated for the fake inputs are all very close to zero. This is because although the model is mis-specified, the point estimates of the regression coefficients are still reasonable. The estimated coefficients for the fake inputs are all essentially zero. But the p-values inferred under the incorrect multivariate model will no longer be valid for variable selection because the independence assumption on the inputs are severely violated (the pairwise correlation among inputs were fixed at 0.8).
Figure 3.1: Sobol Index Estimates for Linear Gaussian Model with Identity Link
Table 3.3: Quantiles of Relative Difference between SI Estimates and the Corresponding Exact Estimates under Gaussian Model ($\rho = 0.8$)

<table>
<thead>
<tr>
<th>RD-Quantiles</th>
<th>10%</th>
<th>30%</th>
<th>50%</th>
<th>70%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI-UM</td>
<td>$5.5 \times 10^{-16}$</td>
<td>$1.6 \times 10^{-15}$</td>
<td>$2.9 \times 10^{-15}$</td>
<td>$4.8 \times 10^{-15}$</td>
<td>$1.0 \times 10^{-14}$</td>
</tr>
<tr>
<td>SI-CMM</td>
<td>$2.1 \times 10^{-16}$</td>
<td>$7.2 \times 10^{-16}$</td>
<td>$1.3 \times 10^{-15}$</td>
<td>$2.3 \times 10^{-15}$</td>
<td>$6.1 \times 10^{-15}$</td>
</tr>
</tbody>
</table>

NOTE: "SI-UM" stands for Sobol index estimates obtained by fitting univariate models. "SI-CMM" stands for Sobol index estimates obtained by fitting contaminated multivariate model. The accuracy of "SI-UM" is quantified by the following relative difference formula: $\frac{\text{abs}(\text{SI-UM} - \text{SI-EX})}{\text{SI-EX}}$, where "SI-EX" stands for the exact Sobol index estimates obtained by fitting the correct multivariate model. The quantile estimates are obtained based on 1000 simulations (each with sample size 1000) under the Gaussian model with input correlation 0.8. "RD-Quantiles" stands for quantile estimates of the relative differences.

To measure the accuracy of Sobol index estimates based on all 1000 simulations (each with sample size 1000), we can calculate the relative difference between the exact estimates and the estimates obtained by different methods. The percentiles of these relative differences are presented in Table 3.3. From this table, we can see that Sobol indices obtained by fitting univariate model are very accurate. The Sobol indices obtained by fitting multivariate model are slightly more accurate, although the model is misspecified.

We can also re-run above simulations with less correlated inputs. For example, we can use exactly the same simulation setup as above, except that the pair-wise correlations among the inputs are fixed at 0.3 this time. Sobol indices can be estimated again by: 1) fitting the true multivariate model and then evaluating formula (3.5); 2) fitting separate univariate models and then computing empirical variances of lower dimension projections; and 3) by fitting an incorrect multivariate model and then evaluating formula (3.5). The corresponding Sobol index estimates are plotted...
in Figure 3.1 panel (d) to (f). Since the inputs are less correlated in this scenario, the Sobol index estimates of the true inputs have larger variation compared to that with highly correlated inputs in panel (a) to (d). But the index estimates of the fake inputs are all very close to zero and clearly separated from the true inputs. We can also calculate the quantiles of relative difference between Sobol index estimates and the corresponding exact estimates using all 1000 simulations. The summary table is shown in Table E.1 in Appendix E, which again confirms high accuracy of both the Sobol index estimates based on univariate model and that based on contaminated multivariate model.

To summarize, simulations shown in this section indicate that the Sobol index estimates obtained by empirical variances of lower dimension projections are as accurate as using the exact analytic expression of Sobol indices with point estimates of regression coefficients, when the sample size is large enough. In addition, if the true model is multivariate linear regression, Sobol index estimates obtained using formula (3.5) will always give the same value for the same input variable, regardless which or how many input variables are included in model fitting.

### 3.4.1.3 Variable Selection Method Comparison

The fact that the first order main effect Sobol indices can be accurately estimated by only fitting univariate models also implies that the univariate analyses are generally sufficient for variable selections (or singleton feature selections) in real-world applications. In this section, we will use simulation examples to show that the univariate analyses are generally better than the multivariate analyses for the purpose of variable selections when the underlying true models are unknown.
For example, 1000 samples are generated from the same simulation setup used before, with pair-wise correlation among the inputs being fixed at 0.8. Each sample still have 1000 observations. Although in total 40 true inputs are generated and used for simulating the response, we pretend only the first 20 true inputs and the 20 fake inputs (simulated independently with the response, and have no relation to the response) are available for performing variable selection procedures. For each sample, we perform variable selection using all of the following techniques: 1) the univariate linear regression; 2) the Kendall’s Tau Tests; 3) the analysis of variance (ANOVA) on multivariate model contain 20 true inputs and 20 fake inputs; 4) the multivariate linear regression contain 20 true inputs and 20 fake inputs; 5) Sobol index with regression coefficients estimated under the incorrect multivariate linear model (contain 20 true inputs and 20 fake inputs) using iteratively reweighed least square (IRLS); 6) Sobol index with regression coefficients estimated under the incorrect multivariate model using coordinate descent with Lasso penalty (CD-Lasso); 7) Sobol indices with regression coefficients estimated using coordinate descent with Ridge penalty (CD-Ridge); 8) Sobol indices with regression coefficients estimated by coordinate descent with Elastic Net penalty (CD-ElasticNet).

Figure 3.2 plotted the results of all eight methods after analyzing the same simulation sample. In the plots on the first row, the green horizontal lines are the 0.05 significance threshold for p-values. The green lines in the second row plots indicate the maximum Sobol index value among the fake inputs. The red vertical lines are used to separate the true and fake inputs. From the first row plots in Figure 3.2 we can see that when inputs are highly correlated the univariate analyses (the univariate linear regression and the Kendall’s Tau test) picked up all the true inputs.
Figure 3.2: Variable Selection Methods Comparison under Multivariate Linear Gaussian Model (inputs correlation $\rho = 0.8$)
Figure 3.3: Sobol Index Significance Test under Multivariate Linear Gaussian Model (inputs correlation $\rho = 0.8$)
But the multivariate analyses (ANOVA and multivariate linear regression) failed to pick out all the true inputs and also picked up one fake input. This is due to the mis-specification of the model and the violation of the independence assumption on the inputs.

The four plots on the second row are the Sobol index estimates obtained by using formula (3.5) with coefficients estimated using different fitting algorithms. Note that all the coefficients used in calculating these Sobol indices are estimated under the incorrect multivariate model. But these approximated Sobol indices still present a clear separation between the true and fake inputs, and all fake inputs have Sobol indices close to zero.

We can also estimate p-values for the significance tests of Sobol indices, where the null hypothesis is that the Sobol index equals zero. By permuting the response values to match up different inputs observations, we can repeatedly estimate Sobol indices using different permutation samples to approximate the distribution of each Sobol index under the null hypothesis given the specified model. The p-value is approximately the percentage of Sobol index estimates that are larger than the one estimated under the original sample. Figure 3.3 gives the p-value estimates corresponding to the Sobol indices shown in Figure 3.2. The p-values of Sobol indices under the incorrect multivariate linear regression turn out to be roughly the same as the ones under univariate linear regression, regardless which fitting algorithm (IRLS, CD-Lasso, CD-Ridge, or CD-ElasticNet) is used to obtain the regression coefficients.

These comparison conclusions are also confirmed by analyzing all 1000 simulation samples. Table 3.4 lists the type I error, power, and false discovery rate (FDR) of these eight methods, estimated empirically using the 1000 simulation samples and a
Table 3.4: Type I Error, Power, and FDR Estimates ($\rho = 0.8$)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Univariate Regression</th>
<th>Kendall’s Tau</th>
<th>ANOVA</th>
<th>Multivariate Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I Error</td>
<td>0.0290</td>
<td>0.0303</td>
<td>0.0224</td>
<td>0.0218</td>
</tr>
<tr>
<td>Power</td>
<td>0.9721</td>
<td>0.9702</td>
<td>0.7815</td>
<td>0.7219</td>
</tr>
<tr>
<td>FDR</td>
<td>0.0289</td>
<td>0.0302</td>
<td>0.0279</td>
<td>0.0293</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I Error</td>
<td>0.0287</td>
<td>0.0289</td>
<td>0.0282</td>
<td>0.0291</td>
</tr>
<tr>
<td>Power</td>
<td>0.9720</td>
<td>0.9719</td>
<td>0.9735</td>
<td>0.9719</td>
</tr>
<tr>
<td>FDR</td>
<td>0.0287</td>
<td>0.0289</td>
<td>0.0281</td>
<td>0.0291</td>
</tr>
</tbody>
</table>

NOTE: Both the univariate and multivariate models are fitted by R function `glm`; The Kendall’s Tau test are performed using R function `cor.test`; The ANOVA analysis is executed using R function `anova`; Model fitting using coordinate decent algorithm with different penalties are executed using R function `glmnet`.

fixed threshold on p-values. These type I error, power, and FDR estimates are calculated after adjusting the p-values for multiple testing using the Benjamini-Hochberg’s procedure. The significance level of 0.05 is used as the selection threshold after performing the Benjamini-Hochberg’s procedure. From this table we can see that when the inputs are highly correlated, using Sobol indices with regression coefficients obtained by CD-Ridge appears to be the best among all. But it’s only slightly better than using the univariate regression or using the Sobol indices estimated using coefficients obtained by IRLS. We can also vary threshold on p-values to generate the ROC curves for method comparison (shown in the left panel of Figure 3.4). From this figure we can clearly see that all univariate analyses perform almost equally well, and also outperform the multivariate analyses dramatically.
Figure 3.4: ROC Curves for Method Comparison under Multivariate Linear Gaussian Model
Table 3.5: Type I Error, Power, and FDR Estimates ($\rho = 0.3$)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Univariate Regression</th>
<th>Kendall’s Tau</th>
<th>ANOVA</th>
<th>Multivariate Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I Error</td>
<td>0.0226</td>
<td>0.0222</td>
<td>0.0227</td>
<td>0.0214</td>
</tr>
<tr>
<td>Power</td>
<td>0.7917</td>
<td>0.7791</td>
<td>0.7784</td>
<td>0.7501</td>
</tr>
<tr>
<td>FDR</td>
<td>0.0277</td>
<td>0.0276</td>
<td>0.0283</td>
<td>0.0277</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I Error</td>
<td>0.0220</td>
<td>0.0215</td>
<td>0.0207</td>
<td>0.0215</td>
</tr>
<tr>
<td>Power</td>
<td>0.7910</td>
<td>0.7918</td>
<td>0.7901</td>
<td>0.7918</td>
</tr>
<tr>
<td>FDR</td>
<td>0.0271</td>
<td>0.0264</td>
<td>0.0255</td>
<td>0.0264</td>
</tr>
</tbody>
</table>

NOTE: Both the univariate and multivariate models are fitted by R function `glm`; The Kendall’s Tau test are performed using R function `cor.test`; The ANOVA analysis is executed using R function `anova`; Model fitting using coordinate decent algorithm with different penalties are executed using R function `glmnet`.

To investigate scenarios where the inputs are weakly correlated, we repeat the above simulation with inputs correlation fixed at 0.3 instead of 0.8. Comparison figures similar to Figure 3.2 and 3.3 are plotted based on one simulation sample when the inputs are less correlated (see Figure E.1 and E.2 in Appendix E). The type I error, power, and FDR of the same eight methods are again estimated using 1000 simulation samples and 0.05 threshold on p-values (see Table 3.5). The ROC curves for this scenario are plotted in the right panel of Figure 3.4. Based on these results, we conclude that when the inputs are weakly correlated, the Sobol indices with regression coefficients estimated by CD-Lasso or CD-ElasticNet have slightly better performance. All univariate analyses still behave almost equally well, and only have slightly better performance than the multivariate methods.
To summarize, simulations shown in this section suggest that multivariate analyses should not be used for variable selection (or singleton feature selection) in real-world applications, since the correct form of the underlying true models are impossible to know in advance. But they are useful for variable combination selections because estimation of higher order Sobol indices with respect to more than one input rely on fitting multivariate models. In addition, Sobol indices estimated by fitting incorrect multivariate models perform almost equally well for the purpose of variable selection, regardless what fitting algorithm is used for obtaining the coefficient estimates and regardless how strong the correlation is among the inputs.

3.4.1.4 Variable Selection by Total-effect Sobol Indices

Since when all the inputs are positively correlated with each other, the total-effect Sobol index with respect to a single input variable is strictly greater than the corresponding main-effect Sobol index. So we think it’s interesting to investigate the performance of total-effect Sobol indices in variable selection tasks. Figure 3.5 plotted the total-effect Sobol indices estimated under a Gaussian model with input correlation equal 0.8. From this figure, we can see that the total-effect Sobol indices for fake inputs are no longer approximately zero. The total-effect Sobol indices do not seem to perform better in variable selection tasks, due to the contamination in the multivariate model that was used for estimating the regression coefficients.

3.4.2 Simulation under Poisson Models

In this section, we will use simulation examples based on two types of Poisson models to test the performance of the derived expression of Sobol indices under GLMs with log link (Result 3.3.2 in Section 3.3.2). Both models used for simulation have
Figure 3.5: Total-effect Sobol Indices under Multivariate Linear Gaussian Model with Inputs Correlation $\rho = 0.8$. 
multivariate normal inputs. But the first model uses the identity link to simulate the response observations, while the second one uses the log link.

### 3.4.2.1 Simulation under Poisson Model with Identity Link

**Simulation Setup** We first simulate a sample of 40 input variables from a multivariate normal (MVN) distribution with the mean values generated from the uniform distribution with domain [-50,50], the marginal standard deviations generated from the uniform on [0,10], and all pairwise correlations set as 0.8. The sample size is 1000. Then we generate the true regression coefficients of these inputs from the uniform distribution defined on [-1, 1]. To make sure the response variable, i.e. the Poisson random variable, have a positive mean, the intercept coefficient $\beta_0$ is generated after all 1000 observations of $\sum_{i=1}^{40} \beta_i X_i$ are generated. The value of $\beta_0$ is set as a positive number (drawn from uniform on [0, 1]) plus the largest absolute value of $\sum_{i=1}^{40} \beta_i X_i$ across all 1000 observations. Using the inputs and the coefficients generated above, the responses are simulated from Poisson distribution with different mean calculated according to $\lambda = E(Y|X) = \beta_0 + X^T \beta$.

In order to test the performance of the estimation methods, another 20 fake input variables are also simulated from the same multivariate normal distribution that is used to generated the 40 true inputs. The fake inputs are considered “fake” because they are not used in generating the response values, and thus have no relation to the response variable. They are only similar to the true inputs in the sense that both the true inputs and the fake inputs were draw independently from the same multivariate normal distribution.
Accuracy Assessment of Sobol Index Estimates

Since the underlying true model uses identity link, based on the simulation study under the Gaussian model and Result 3.3.1 and 3.3.4, we know that either using the formula in Result 3.3.1 or fitting univariate regression can help us to obtain accurate Sobol index estimates for Poisson model with identity link. In this section, we will use the Sobol index estimates obtained by fitting univariate Poisson regression with identity link to check the accuracy of Sobol index estimates obtained by using other methods.

For linear Poisson regression with log link, we have derived exact analytic formulas of Sobol indices in Result 3.3.2. In this section, we will show that these formulas (incorporated with coefficients estimated by fitting Poisson model with log link) can be used to obtain the correct Sobol index estimates for Poisson Model with identity link. For example, given one simulation sample, we can first obtain the correct estimates of the first-order main effect Sobol indices by fitting univariate regressions and then calculate the sample variances of each of these univariate functions. These estimates based on univariate models are plotted in Figure 3.6 panel (a) (without scaling by the response variance). We can then obtain another set of Sobol index estimates by evaluating formula (3.7) with coefficient estimates obtained from fitting the Poisson model containing all 40 true inputs with the log link. These estimates based on fitting Poisson model with log link are plotted in Figure 3.6 panel (b) (without scaling by the response variance as well). By comparing panel (a) and (b), we can see that the Sobol index estimates obtained by applying formula (3.7) are as accurate as the ones obtained by fitting separate univariate regressions.

According to Result 3.3.4, we know that if the true model has a linear systematic component, the identity link and MVN inputs, any lower dimensional projection
Figure 3.6: Sobol Index Estimates for Linear Poisson Model with Identity Link
(the conditional expectation with respect to any input subset) is also a multivariate linear function of the partial inputs. Therefore, under the linear Poisson model with identity link and MVN inputs, Sobol indices can also be accurately estimated by fitting models containing only partial inputs. The following simulation indicates that under linear Poisson model with identity link and MVN inputs, Sobol indices can even be estimated accurately using the formula derived under log link, incorporating with coefficients estimated by fitting linear Poisson model with log link on mixture of partial inputs and noises.

In this simulation, we pretend that the true model is mistaken to be a linear Poisson regression with log link that contains the first 20 true inputs and the 20 fake inputs. The other 20 true inputs (with input id from 21 to 40) are not observed in data collection. Then we obtain a set of Sobol index estimates by fitting the linear Poisson model with log link on this mixture of partial (20 out of 40) true inputs and 20 fake inputs, and then evaluate formula (3.7) with 40 incorrect coefficient estimates obtained from fitting this contaminated model. Sobol index estimates obtained this way are plotted in Figure 3.6 panel (c). By comparing panel (a) and (c), we can see that the Sobol indices estimated under the contaminated model also turn out to be very accurate. And the indices estimated for the fake inputs are all still very close to zero, clearly separated from the estimates for the true inputs.

To measure the accuracy of Sobol index estimates based on all 1000 simulations (each with sample size 1000), we can calculate the relative difference between the exact estimates and the estimates obtained by different methods. The percentiles of these relative differences are presented in Table 3.6. From this table, we can see
Table 3.6: Quantiles of Relative Difference between SI Estimates and the Corresponding Correct Estimates under Poisson Model with Identity Link ($\rho = 0.8$)

<table>
<thead>
<tr>
<th>RD-Quantiles</th>
<th>10%</th>
<th>30%</th>
<th>50%</th>
<th>70%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI-MML</td>
<td>3.2×10^{-3}</td>
<td>1.0×10^{-2}</td>
<td>2.2×10^{-2}</td>
<td>4.3×10^{-2}</td>
<td>1.1×10^{-1}</td>
</tr>
<tr>
<td>SI-CMML</td>
<td>2.5×10^{-3}</td>
<td>9.5×10^{-3}</td>
<td>2.1×10^{-2}</td>
<td>3.7×10^{-2}</td>
<td>9.4×10^{-2}</td>
</tr>
</tbody>
</table>

NOTE: "SI-MML" stands for Sobol index estimates obtained by fitting the multivariate models with all true inputs and the log link. "SI-CMML" stands for Sobol index estimates obtained by fitting contaminated multivariate model with log link. The accuracy of "SI-MML" is quantified by the following relative difference formula: $\text{abs}(\text{"SI-MML" - "SI-UM"}) / \text{"SI-UM"}$, where "SI-UM" stands for the correct Sobol index estimates obtained by fitting the univariate model. The quantile estimates are obtained based on 1000 simulations (each with sample size 1000) from the Poisson model with identity link and input correlation 0.8. "RD-Quantiles" stands for quantile estimates of the relative differences.

that Sobol indices obtained by fitting multivariate models with log link are still fairly accurate.

We can also re-run above simulations with less correlated inputs. For example, we can use exactly the same simulation setup as above, except that the pair-wise correlations among the inputs are fixed at 0.3 this time. Sobol indices can be estimated again by: 1) fitting separate univariate regression and then computing empirical variances of lower dimension projections; 2) fitting the linear Poisson model containing all 40 true inputs with log link, and then evaluating formula (3.7); and 3) by fitting linear Poisson model containing partial true inputs and some fake inputs with log link, and then evaluating formula (3.7). The corresponding Sobol index estimates are plotted in Figure 3.1 panel (d) to (f). Similar to the simulations under Gaussian models, since the inputs are less correlated in this scenario, the Sobol index estimates for the true inputs have larger variation compared to that with highly correlated inputs.
in panel (a) to (d). But the index estimates of the fake inputs all consistently stay close to zero. We can also calculate the quantiles of relative difference between Sobol index estimates and the corresponding correct estimates using all 1000 simulations. The summary table is shown in Table F.1 in Appendix F, which again confirms the accuracy of the Sobol index estimates obtained by fitting multivariate linear Poisson model with log link.

To summarize, simulations shown in this section indicate that if the underlying true model is a linear Poisson model with identity link and multivariate normal inputs, Sobol indices can be accurately estimated by applying the formulas (3.7) derived under the log link, regardless whether the model is contaminated by noise variables or not, as long as the coefficients used for evaluating formula (3.7) are obtained by fitting linear Poisson model with the log link.

**Variable Selection Method Comparison**  In this section, we will compare variable selection methods under the linear Poisson model with the identity link and multivariate normal inputs. 1000 samples are generated from the same Poisson model used in the previous section, with pair-wise correlation among the inputs being fixed at 0.8. Each sample still have 1000 observations. Although in total 40 true inputs are generated and used for simulating the response, we pretend only the first 20 true inputs and the 20 fake inputs (simulated independently with the response, and have no relation to the response) are available for performing variable selection procedures. For each sample, we perform variable selection using all of the following techniques: 1) the univariate linear Poisson regression with log link; 2) the Kendall’s Tau Tests; 3) the analysis of variance (ANOVA) on multivariate model contain 20 true inputs and
20 fake inputs; 4) the multivariate linear Poisson regression contain 20 true inputs and
20 fake inputs with log link; 5) first-order main effect Sobol indices with regression
coefficients estimated under the incorrect linear Poisson model with log link (contain
20 true inputs and 20 fake inputs) using iteratively reweighed least square (IRLS);
6) Sobol indices with regression coefficients estimated under the same incorrect Pois-
son model using coordinate descent with Lasso penalty (CD-Lasso); 7) Sobol indices
with regression coefficients estimated using coordinate descent with Ridge penalty
(CD-ridge) under the same incorrect Poisson model; 8) Sobol indices with regression
coefficients estimated by coordinate descent with Elastic Net penalty (CD-ElasticNet)
under the same incorrect Poisson model.

Figure 3.7 plotted the results of all eight methods after analyzing the same sim-
ulation sample. In the plots on the first row, the green horizontal lines are the 0.05
significance threshold for p-values. The green lines in the second row plots indicate
the maximum Sobol index value among the fake inputs. The red vertical lines are
used to separate the true and fake inputs. From the first row plots in Figure 3.7 we
can see that when the underlying true model is a multivariate linear Poisson model
with identity link, the univariate linear Poisson model with log link is doing almost
as good as the Kendall’s Tau test. Both these two univariate approaches picked out
all true inputs correctly. But the multivariate analyses (ANOVA and multivariate
linear Poisson regression with log link) failed to pick out all the true inputs. And the
contaminated multivariate Poisson model also made two false discoveries in this case.

The four plots on the second row present the Sobol index estimates obtained by
using formula (3.7) with coefficients estimated using different algorithms for fitting the
contaminated multivariate Poisson model with log link. Note that all the coefficients
Figure 3.7: Variable Selection Methods Comparison under Linear Poisson Model with Identity Link and Inputs Correlation $\rho = 0.8$. 

- **p.value_multivariate**: shows the p-values for the multivariate selection method. 
- **p.value_ANOVA**: displays the p-values for the ANOVA-based selection method. 
- **p.value_Kendall's Tau**: illustrates the p-values for Kendall's Tau method. 
- **p.value_univariate**: presents the p-values for the univariate selection method. 
- **Sobol_index_CD.Frechet**: plots the Sobol indices for the CD.Frechet method. 
- **Sobol_index_CD.Ridge**: exhibits the Sobol indices for the CD.Ridge method. 
- **Sobol_index_CD.Lasso**: shows the Sobol indices for the CD.Lasso method. 
- **Sobol_index_IRLS**: displays the Sobol indices for the IRLS method.

Each graph compares the performance of these selection methods under the specified conditions.
Figure 3.8: Sobol Index Significance Test under Linear Poisson Model with Identity Link and Inputs

Correlation $\rho = 0.8$.
used in calculating these Sobol indices are estimated under the incorrect link function. Regardless which fitting algorithm is used, these first-order main-effect Sobol indices present a clear separation between the true and fake inputs, and all fake inputs have approximated zero-valued Sobol indices.

Figure 3.8 gives the p-value estimates corresponding to the Sobol indices shown in Figure 3.7. The p-values of Sobol indices (estimated using the coefficients estimates of the contaminated multivariate linear Poisson model with log link) present a slightly better separation of true inputs and the fake inputs than the Kendall’s Tau test, regardless which fitting algorithm (IRLS, CD-Lasso, CD-Ridge, or CD-ElasticNet) is used to obtain the regression coefficients.

These comparison conclusions are also confirmed by analyzing all 1000 simulation samples. The left panel in Figure 3.9 shows the ROC curves for the first five variable selection methods discussed above: 1) univariate linear Poisson with log link; 2) the Kendall’s Tau test; 3) ANOVA; 4) contaminated multivariate linear Poisson model with log link; 5) Sobol indices estimated using coefficients obtained from fitting the contaminated Poisson model with log link. From this figure we can clearly see that all univariate analyses perform almost equally well, and also outperform the multivariate analyses dramatically.

To investigate scenarios when the inputs are weakly correlated, we repeat the above simulation with inputs correlation fixed at 0.3. The corresponding ROC curves are presented in the right panel of Figure 3.9. From this plot, we can see that similar to the cases where the input correlation equal 0.8, all univariate analyses perform almost equally well. The best method is using the first-order main-effect Sobol indices. ANOVA is better than the multivariate Poisson model with log link.
Figure 3.9: ROC Curves for Method Comparison under Linear Poisson Model with Identity Link
But the performance of these two multivariate analyses are much worse than the univariate analyses.

To summarize, simulations shown in this section again suggest that univariate analyses are preferred for variable selection or singleton feature selection, if the observed inputs are likely to contain a lot of noise variables. The usage of Sobol index formulas derived under log link is not limited to cases where the true models in fact use log link. It’s interesting to see that under multivariate linear Poisson model with identity link, Sobol indices can still be accurately estimated using the formula derived under log link. Although the Sobol indices were estimated by fitting contaminated models with a incorrect link (the log link), they still appear to have the best performance among all five variable selection methods being compared, regardless what fitting algorithm is used to obtain the coefficient estimates, and regardless how strong the correlation is among the inputs.

### 3.4.2.2 Simulation under Poisson Model with Log Link

**Simulation Setup** We first simulate a sample of 40 input variables from a multivariate normal (MVN) distribution with the mean values generated from the uniform distribution with domain [-1,1], the marginal standard deviations generated from the uniform on [0.1,0.3], and all pairwise correlations set as 0.8. The sample size is 1000. Then we generate the true regression coefficients of these inputs from the uniform distribution defined on [-1, 1]. Using the inputs and the coefficients generated above, the responses are simulated from Poisson distribution with different mean calculated according to $\lambda = \exp(E[Y|X]) = \exp(\beta_0 + X^T \beta)$.

In order to test the performance of the estimation methods, another 20 fake input variables are also simulated from the same multivariate normal distribution that is
used to generated the 40 true inputs. The fake inputs are considered “fake” because they are not used in generating the response values, and thus have no relation to the response variable. They are only similar to the true inputs in the sense that both the true inputs and the fake inputs were draw independently from the same multivariate normal distribution.

**Accuracy Assessment of Sobol Index Estimates** Since the underlying true model uses log link, we know that using the formula in Result 3.3.2 can help us to obtain accurate Sobol index estimates for Poisson model with log link. In addition, according to Result 3.3.5, we know that fitting univariate polynomial functions on each input variable can also provide approximations for Sobol index with the accuracy depending on the sample size and the degree of polynomial function. In this section, we will use the Sobol index estimates obtained by applying formula (3.7) to check the accuracy of Sobol index estimates obtained by using other methods, including fitting univariate polynomial functions with degree 3.

In the following simulation example, we will show that when the underlying true model is a multivariate linear Poisson model with log link, estimating Sobol indices by fitting univariate polynomial function with degree 3 is decent for variable selection, but not for importance ranking. For example, given one simulation sample, we can first obtain the exact estimates of the first-order main-effect Sobol indices by fitting the correct multivariate Poisson model and then use its coefficients estimates to evaluate formula (3.7). These exact estimates based on the correct multivariate Poisson model are plotted in Figure 3.10 panel (a) (without scaling by the response variance).
We can then obtain another set of Sobol index estimates by fitting separate univariate polynomial functions with degree 3 on each one of the input variables and then evaluate the empirical variances of fitted polynomial functions. These approximated Sobol index estimates based on univariate model fitting are plotted in Figure 3.10 panel (b) (without scaling by the response variance as well). By comparing panel (a) and (b), we can see that although the approximated Sobol index estimates are not very accurate, they in fact clearly separated the true inputs and fake ones, and the fake ones still have Sobol index estimates that are approximately zero.

Since the underlying true model does not use the identity link, Result 3.3.4 is no longer held for this Poisson model simulation. So if the model is contaminated and contain only partial true inputs, we should expect inaccurate Sobol index estimates obtained by applying formula (3.7). The following simulation confirms that estimation of Sobol indices under models with non-identity link is very sensitive to model specification. But these inaccurate Sobol indices still seem to be quite sufficient for the purpose of variable selection.

In this simulation, we pretend that the true model is mistaken to be a linear Poisson regression with log link that contains the first 20 true inputs and the 20 fake inputs. The other 20 true inputs (with input id from 21 to 40) are not observed in data collection. Then we obtain a set of Sobol index estimates by fitting the multivariate linear Poisson model with log link on this mixture of true and fake inputs, and then evaluate formula (3.7) with 40 incorrect coefficient estimates obtained from fitting this contaminated model. Sobol index estimates obtained this way are plotted in Figure 3.10 panel (c). By comparing panel (a) and (c), we can see that the Sobol indices estimated under the contaminated model are no longer accurate. But the indices
Figure 3.10: Sobol Index Estimates for Linear Poisson Model with Log Link
Table 3.7: Quantiles of Relative Difference between SI Estimates and the Corresponding Exact Estimates under Poisson Model with Log Link ($\rho = 0.8$)

<table>
<thead>
<tr>
<th>RD-Quantiles</th>
<th>10%</th>
<th>30%</th>
<th>50%</th>
<th>70%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI-UM</td>
<td>$1.9 \times 10^{-2}$</td>
<td>$6.1 \times 10^{-2}$</td>
<td>$1.2 \times 10^{-1}$</td>
<td>$2.5 \times 10^{-1}$</td>
<td>$9.5 \times 10^{-1}$</td>
</tr>
<tr>
<td>SI-CMM</td>
<td>$7.4 \times 10^{-3}$</td>
<td>$2.4 \times 10^{-2}$</td>
<td>$4.5 \times 10^{-2}$</td>
<td>$8.6 \times 10^{-2}$</td>
<td>$3.1 \times 10^{-1}$</td>
</tr>
</tbody>
</table>

NOTE: "SI-UM" stands for Sobol index estimates obtained by fitting univariate models. "SI-CMM" stands for Sobol index estimates obtained by fitting contaminated multivariate model. The accuracy of "SI-UM" is quantified by the following relative difference formula: $\frac{\text{abs(\"SI-UM\" - \"SI-EX\")}}{\text{\"SI-EX\"}}$, where "SI-EX" stands for the exact Sobol index estimates obtained by fitting the correct multivariate model. The quantile estimates are obtained based on 1000 simulations (each with sample size 1000) under the Poisson model with log link and input correlation 0.8. "RD-Quantiles" stands for quantile estimates of the relative differences.

estimated for the fake inputs are all still very close to zero, and clearly separated from the estimates for the true inputs.

To measure the accuracy of Sobol index estimates based on all 1000 simulations (each with sample size 1000), we can calculate the relative difference between the exact estimates and the estimates obtained by different methods. The percentiles of these relative differences are presented in Table 3.7. From this table, we can see that Sobol indices obtained by fitting univariate model are no longer very accurate. The Sobol indices obtained by fitting contaminated multivariate model are still slightly more accurate, compared to estimates obtained by fitting univariate models.

We can also perform a similar simulation with the pair-wise inputs correlation fixed at 0.3, and again estimate Sobol indices using the same set of methods: 1) fitting the correct multivariate linear Poisson model containing all 40 true inputs with log link, and then evaluate formula (3.7); 2) fitting separate univariate polynomial functions
with degree 3 and then compute empirical variances of these fitted lower-dimension projections; and 3) by fitting contaminated multivariate linear Poisson model with log link on partial true inputs and some fake inputs, and then evaluate formula (3.7). The corresponding Sobol index estimates are plotted in Figure 3.10 panel (d) to (f). Similar to the simulations under Gaussian models, since the inputs are less correlated in this scenario, the Sobol index estimates for the true inputs in panel (d) have larger variation compared to that with highly correlated inputs in panel (a). Both fitting univariate functions and fitting contaminated multivariate model produce inaccurate Sobol index estimates. But these inaccurate estimates appear to be sufficient for variable selection. We can also calculate the quantiles of relative difference between Sobol index estimates and the corresponding exact estimates using all 1000 simulations. The summary table is shown in Table F.2 in Appendix F, which again indicates that the estimation of Sobol indices under log link is very sensitive to model specification.

To summarize, simulations shown in this section indicate that if the underlying true model is a multivariate linear Poisson model with the log link and multivariate normal inputs, Sobol indices can be accurately estimated by applying the formulas (3.7), only if the correct model is fitted for providing the coefficients estimates. But if the model is contaminated by noise variables, the inaccurate estimates obtained by formulas (3.7) still seem to be quite sufficient for the task of variable selection. So are the inaccurate estimates obtained by fitting univariate polynomial functions with low degree.
Variable Selection Method Comparison  In this section, we will compare variable selection methods under the linear Poisson model with the log link and multivariate normal inputs. 1000 samples are generated from the same Poisson model used in the previous section, with pair-wise correlation among the inputs being fixed at 0.8. Each sample still have 1000 observations. Although in total 40 true inputs are generated and used for simulating the response, we pretend only the first 20 true inputs and the 20 fake inputs (simulated independently with the response, and have no relation to the response) are available for performing variable selection procedures. For each sample, we perform variable selection using all of the following techniques: 1) the univariate linear Poisson regression with log link; 2) the Kendall’s Tau Tests; 3) the analysis of variance (ANOVA) on multivariate model contain 20 true inputs and 20 fake inputs; 4) the multivariate linear Poisson regression contain 20 true inputs and 20 fake inputs with log link; 5) first-order main effect Sobol indices with regression coefficients estimated under the incorrect linear Poisson model with log link (contain 20 true inputs and 20 fake inputs) using iteratively reweighed least square (IRLS); 6) Sobol indices with regression coefficients estimated under the same incorrect Poisson model using coordinate descent with Lasso penalty (CD-Lasso); 7) Sobol indices with regression coefficients estimated using coordinate descent with Ridge penalty (CD-ridge) under the same incorrect Poisson model; 8) Sobol indices with regression coefficients estimated by coordinate descent with Elastic Net penalty (CD-ElasticNet) under the same incorrect Poisson model.

Figure 3.11 plotted the results of all eight methods after analyzing the same simulation sample. In the plots on the first row, the green horizontal lines are the 0.05 significance threshold for p-values. The green lines in the second row plots indicate
Figure 3.11: Variable Selection Methods Comparison under Linear Poisson Model with Log Link and Inputs Correlation $\rho = 0.8$.
the maximum Sobol index value among the fake inputs. The red vertical lines are used to separate the true and fake inputs. From the first row plots in Figure 3.11 we can see that when the underlying true model is a multivariate linear Poisson model with log link, the univariate analyses are still preferred over the multivariate analyses. Both the parametric and nonparametric univariate approaches picked out all true inputs correctly. But the multivariate analyses (ANOVA and multivariate linear Poisson regression with log link) not only failed to pick out all the true inputs, but also made false signal discoveries in this simulation example.

The four plots on the second row present the Sobol index estimates obtained by using formula (3.7) with coefficients estimated using different algorithms for fitting the contaminated multivariate Poisson model with log link. Note that regardless which fitting algorithm is used, these first-order main-effect Sobol indices present clear separation between the true and fake inputs, and the approximated Sobol indices for all fake inputs are estimated to be essentially zero.

Figure 3.12 gives the p-value estimates corresponding to the Sobol indices shown in Figure 3.11. The p-values of Sobol indices (estimated using the coefficients estimates of the contaminated multivariate linear Poisson model with log link) made clean classification of true inputs and fake inputs, regardless which fitting algorithm (IRLS, CD-Lasso, CD-Ridge, or CD-ElasticNet) is used to obtain the regression coefficients.

These comparison conclusions are also confirmed by analyzing all 1000 simulation samples. The left panel in Figure 3.13 shows the ROC curves for the first five variable selection methods discussed above: 1) univariate linear Poisson with log link; 2) the Kendall’s Tau test; 3) ANOVA; 4) contaminated multivariate linear Poisson model with log link; 5) Sobol indices estimated using coefficients obtained from fitting the
Figure 3.12: Sobol Index Significance Test under Linear Poisson Model with Log Link and Inputs Correlation $\rho = 0.8$. 
contaminated Poisson model with log link. From this figure we can clearly see that Sobol indices and the Kendall’s Tau perform almost equally well. The univariate linear Poisson model with log link is not as good as Sobol indices and Kendall’s Tau test, but still outperform the multivariate analyses. It’s worth noting that the Sobol indices with the best variable-selection performance are estimated by fitting the contaminated multivariate Poisson model that has the worst variable-selection performance.

To investigate scenarios when the inputs have small correlations, we perform another similar simulation with inputs correlation fixed at 0.3. Comparison figures similar to Figure 3.11 and 3.12 are plotted based on one simulation sample where the inputs are simulated use correlation 0.3 (see Figure F.1 and F.2 in Appendix F). Note that the univariate linear Poisson model with log link did not show any strength in detecting the true input, compared to the contaminated multivariate Poisson model. The corresponding ROC curves are presented in the right panel of Figure 3.13. From this plot, we can see that the best variable-selection method is the Kendall’s Tau. Sobol indices are doing better than the ANOVA. And ANOVA is better than the multivariate Poisson model with log link and the univariate linear Poisson model with log link.

To summarize, simulations shown in this section suggest that estimation of Sobol indices under the multivariate linear Poisson model with log link require correct model specification in model fitting. When inputs are highly correlated, both univariate models and contaminated multivariate model produce inaccurate Sobol index estimates. But these estimates are sufficient for variable selection task. However, when
Figure 3.13: ROC Curves for Method Comparison under Linear Poisson Model with Log Link
the inputs have weak correlations, in terms of variable selection, Kendall’s Tau out-
perform Sobol indices estimated by fitting contaminated models. And the univariate
model is no longer doing better than the contaminated multivariate model.

3.4.3 Variable Ranking Comparison

In this section, we will compare the ranking of input variables based on five differ-
ent importance measures: 1) p-values of the Kendall’s Tau independence test; 2) main-
effect Sobol indices; 3) total-effect Sobol indice; 4) averaged increase in prediction
error (measured by Mean Squared Error) after permuting the input variable; 5) total
decrease in node impurities (measured by Gini index for classification and measured
by residual sum of squares for regression) after splitting the data by setting thresh-
old on the input variable. In the following discussion we will use “Kendall’s Tau”,
“Main_SI”, “Total_SI”, “RF_1” and “RF_2” to refer the variable ranking based on
these five importance measures. The reason that importance measure 4) and 5) are
referred by “RF_1” and “RF_2” here is because these two measures are commonly used
in variable selection based on random forest ensemble learning algorithms [14, 19, 33]
and implemented in the following simulations using R package randomForest.

Table 3.8 lists the mean Spearman Rho estimates for comparing rankings of all
ten method pairs, assuming all the true inputs are observed and the correct models
are fitted. The mean Spearman Rho are obtained by averaging 1000 Spearman Rho
estimates calculated using the 1000 simulation samples. Each simulation sample has
size 1000. From this table we can see that if correct models are fitted, rankings
produced by main-effect Sobol indices and Kendall’s Tau p-values match up well
with each other. So are the rankings by total-effect Sobol indices and random forest
<table>
<thead>
<tr>
<th></th>
<th>Gaussian Model ((\text{rho}=0.3))</th>
<th>Poisson Model ((\text{rho}=0.8))</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Main_SI vs. Kendall’s Tau</td>
<td>0.565</td>
<td>0.928</td>
<td>0.942</td>
</tr>
<tr>
<td>RF_1 vs. RF_2</td>
<td>0.804</td>
<td>0.727</td>
<td>0.878</td>
</tr>
<tr>
<td>RF_2 vs. Total_SI</td>
<td>0.619</td>
<td>0.788</td>
<td>0.544</td>
</tr>
<tr>
<td>Main_SI vs. RF_2</td>
<td>0.590</td>
<td>0.782</td>
<td>0.571</td>
</tr>
<tr>
<td>RF_1 vs. Kendall’s Tau</td>
<td>0.647</td>
<td>0.623</td>
<td>0.628</td>
</tr>
<tr>
<td>RF_2 vs. Kendall’s Tau</td>
<td>0.554</td>
<td>0.520</td>
<td>0.545</td>
</tr>
<tr>
<td>Main_SI vs. Total_SI</td>
<td>0.623</td>
<td>0.672</td>
<td>0.551</td>
</tr>
<tr>
<td>RF_1 vs. Kendall’s Tau</td>
<td>0.451</td>
<td>0.503</td>
<td>0.484</td>
</tr>
<tr>
<td>RF_2 vs. Kendall’s Tau</td>
<td>0.453</td>
<td>0.391</td>
<td>0.329</td>
</tr>
<tr>
<td>Main_SI vs. Total_SI</td>
<td>0.334</td>
<td>0.165</td>
<td>0.312</td>
</tr>
<tr>
<td>Kendall’s Tau vs. Total_SI</td>
<td>0.176</td>
<td>0.315</td>
<td>0.117</td>
</tr>
<tr>
<td>94</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.8: Variable Ranking Comparison by Mean Spearman Rho Under Correct Model
Table 3.9: Variable Ranking Comparison by Mean Spearman Rho Under Contaminated Model

<table>
<thead>
<tr>
<th></th>
<th>Gaussian Model (rho=0.8)</th>
<th>Gaussian Model (rho=0.3)</th>
<th>Poisson Model (rho=0.8)</th>
<th>Poisson Model (rho=0.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main_SI vs. Kendall’_Tau</strong></td>
<td>0.939</td>
<td>0.957</td>
<td>0.936</td>
<td>0.955</td>
</tr>
<tr>
<td><strong>RF_1 vs. Total_SI</strong></td>
<td>0.715</td>
<td>0.722</td>
<td>0.581</td>
<td>0.682</td>
</tr>
<tr>
<td><strong>RF_2 vs. Total_SI</strong></td>
<td>0.725</td>
<td>0.757</td>
<td>0.637</td>
<td>0.722</td>
</tr>
<tr>
<td><strong>RF_1 vs. RF_2</strong></td>
<td>0.855</td>
<td>0.780</td>
<td>0.799</td>
<td>0.768</td>
</tr>
<tr>
<td><strong>Kendall’s _Tau vs. Total_SI</strong></td>
<td>0.656</td>
<td>0.662</td>
<td>0.548</td>
<td>0.639</td>
</tr>
<tr>
<td><strong>Main_SI vs. Total_SI</strong></td>
<td>0.666</td>
<td>0.683</td>
<td>0.565</td>
<td>0.661</td>
</tr>
<tr>
<td><strong>Main_SI vs. RF_1</strong></td>
<td>0.777</td>
<td>0.688</td>
<td>0.699</td>
<td>0.651</td>
</tr>
<tr>
<td><strong>Main_SI vs. RF_2</strong></td>
<td>0.839</td>
<td>0.753</td>
<td>0.830</td>
<td>0.729</td>
</tr>
<tr>
<td><strong>RF_1 vs. Kendall’s Tau</strong></td>
<td>0.761</td>
<td>0.678</td>
<td>0.675</td>
<td>0.643</td>
</tr>
<tr>
<td><strong>RF_2 vs. Kendall’s Tau</strong></td>
<td>0.816</td>
<td>0.742</td>
<td>0.808</td>
<td>0.722</td>
</tr>
</tbody>
</table>
Figure 3.14: Variable Ranking Comparison Example Under Contaminated Gaussian Model ($\rho = 0.8$)
Figure 3.15: Variable Ranking Comparison Example Under Contaminated Gaussian Model ($\rho = 0.3$)
Table 3.10: Variable Ranking Accuracy Assessment by Mean Spearman Rho

<table>
<thead>
<tr>
<th></th>
<th>Gaussian Model (rho=0.8)</th>
<th>Gaussian Model (rho=0.3)</th>
<th>Poisson Model (rho=0.8)</th>
<th>Poisson Model (rho=0.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendall’s _ Tau</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Main_SI</td>
<td>1</td>
<td>1</td>
<td>0.996</td>
<td>0.995</td>
</tr>
<tr>
<td>Total_SI</td>
<td>0.859</td>
<td>0.870</td>
<td>0.873</td>
<td>0.876</td>
</tr>
<tr>
<td>RF_1</td>
<td>0.838</td>
<td>0.873</td>
<td>0.822</td>
<td>0.850</td>
</tr>
<tr>
<td>RF_2</td>
<td>0.889</td>
<td>0.912</td>
<td>0.905</td>
<td>0.914</td>
</tr>
</tbody>
</table>
importance measures. But the rankings produced by main-effect and total-effect Sobol indices are not very similar to each other.

Table 3.9 lists the Mean Spearman Rho estimates calculated under the contaminated models. These contaminated models include half of the true inputs and equal number of fake inputs. From this table we can see that if contaminated models are fitted, variable rankings based on different importance measures are fairly similar to each other, including the main-effect Sobol indices versus the total-effect Sobol indices. This is because half of the inputs in each model are fake inputs. All five importance measure can distinguish most of the true inputs from the fake ones. Example scatter plots for comparing rankings under Gaussian Models with different input correlation values are shown in Figure 3.14 and 3.15.

But the five importance measures, except Kendall’s Tau and main-effect Sobol indices (under identity link), are no longer exactly estimated under the contaminated model, which affects the importance ranking to some extent. Table 3.10 shows the mean Spearman Rho estimates for assessing the accuracy of rankings of the true inputs based on the contaminated model fitting. If the rankings obtained by fitting contaminated models are correct, they should be identical to the rankings obtained by fitting the correct models.

3.5 Application Example: Identifying Co-expressed Genes

Polymorphic drug metabolising enzymes are the major causes of adverse drug reactions. Cytochrome P450s (CYPs) is one the most important phase I enzyme family, which metabolizes about 70% of drugs. One valuable enzyme in this family is called CYP3A4, which metabolize 45-60% of currently used drugs. In this section,
we will apply Sobol sensitivity indices to identify genes that are co-expressed with CYP3A4 using a publish dataset.

The CYP3A4 locus is on chromosome 7 with expansion over 281kb. Within this region, there are multiple genes, including CYP3A4 (expressed only in adult livers), CYP3A5, CYP3A7 (only expressed in fetal stage), two pseudogenes and CYP3A43 (expressed in liver but with unknown function). Moreover, CYP3A4 is known to have very large inter-individual variability not only on protein level (40-100 fold) but also in constitutive activities (7-20 fold) and induced activities (about 11 fold). In order to explain these large variability, the genetics of CYP3A4 has been studied for many years. But inside the CYP3A4 gene locus, not many cis-activity polymorphisms have been found. And no particular trans-acting polymorphisms or epigenetic factors have been identified responsible for a large portion of the variability.

The dataset used for this analysis is the same microarray data used in Yang et al. 2010 [92]. This microarray was done using Cy3 and Cy5 fluorescent to label individual samples and pooled control sample. The relative intensity of the two dyes are reported for 427 individuals. The measurements represent the relative abundance of gene expression compared to the pooled control group. Some genes may have multiple measurements reported because there are multiple probes being used. In total, we will investigate 78 measurements collected on 46 candidate genes (pre-selected based on literature review). For each of the 78 measurements, missing values are imputed by the empirical mean of the observed data.

There are many different ways of using Sobol indices to define and weight the edges in co-expression networks. The simplest idea is to stick with the conventional pairwise-dependent structure, in which case we consider each gene as a node in the
network and weight the edge between CYP3A4 and each candidate gene by the Sobol index of CYP3A4 (the model output) with respect to that candidate gene (the input variable). If the underlying true relationship between CYP3A4 and the candidate gene is indeed a univariate linear regression, such pairwise-dependent structure based on Sobol indices is essentially the same as the conventional network constructed according to the squared Pearson correlation. But if the true one-dimensional projection is some other polynomial or piece-wise polynomial form with degree higher than 1, the network constructed on Sobol indices will be a better model since it does not force a misspecified linear relationship and can be robustly estimated as long as other observed or unobserved confounding factors are either independent of the candidate gene expression or correlated with it as two coordinates in a multivariate normal distribution. Since we are only looking at the one-dimensional projections of CYP3A4 expression with respect to each single candidate gene in above analysis, we will refer this analysis procedure as the first-order co-expression analysis in our later discussions.

One obvious drawback of the first-order analysis is that we will not be able to compare how different gene-gene interactions or candidate gene combinations are affecting the expression of CYP3A4. To quantify different gene combination effects (including the effects of the gene-gene interactions within the combination), we can define another type of edge between CYP3A4 and a subset of candidate genes as the Sobol index with respect to that candidate gene subset. By doing so, we can study dependent expression patterns that involve more than two genes. Moreover, regardless how many genes are actually involved in the biological mechanism that is affecting the expression of CYP3A4 and how many of them are actually being
measured, such network constructed on Sobol indices should always provide valid inference as long as the unobserved factors are independent with the observed genes or correlated with the observed ones in the way of multivariate normal distribution.

To summarize, the detailed analysis procedures performed on the CYP3A4 microarray dataset is described as follows. In the first order analysis, we estimate all Sobol indices with respect to a single candidate gene by fitting univariate polynomial model with degree 3. The reason of choosing degree 3 is because the analysis results do not change much after fitting polynomials with higher degrees. For each candidate gene, we start with the full polynomial model with degree 3, meaning the model contains all linear terms, quadratic terms, and cubic terms. Then we use backward-forward stepwise procedure to select a polynomial form that has the best fit. And the main-effect Sobol indices are estimated by the empirical variances of the best fitting polynomial expressions (may have the highest degrees lower than 3). These Sobol indices can be interpreted as the proportion in CYP3A4 variation that can be explained by the candidate genes individually. In the second order analysis, we estimate all the main-effect Sobol indices with respect to a gene pair. For each gene pair, we also start with the full polynomial model with degree 3, meaning the model contains all pairwise product terms in addition to the terms used in the first order analysis. To identify the best fitting form, we also use the backward-forward stepwise procedure. So we obtain an importance ranking of all possible gene pairs. Similarly, in the third order analysis, we generate the ranking of all possible gene triplets according to Sobol indices estimated from fitting polynomial models with 3 inputs and the highest degree less than or equal to 3. Because the total number of possible gene quadruplets is too large (1,426,425), in this example we only estimated
Figure 3.16: CYP3A4 Sensitivity Network with the Top Gene Quadruplets
Figure 3.17: Gene Quadruplet with Smallest Residual Deviances
the Sobol indices with respect to each of the 194,580 gene quadruplets that formed by the top 200 gene triplets picked out in the third order analysis.

To help visualize the analysis results, a sensitivity network is plotted in Figure 3.16. Each node represent a candidate gene. The size of each node is proportional to the main-effect Sobol indices with respect to the single gene. The edges connect top 3% gene quadruplets with the highest Sobol indices. The corresponding Sobol index values range from 68% to 73%. The reason of only plotting the top 3% quadruplet is not because only the top 3% quadruplets are statistically significant. In fact, all 194,580 quadruplets are statistically significant. We will not be able to see the top picks if plot all of them. In Figure 3.16, the strongest dependency structures recovered by the fourth order analysis involved more than 20 genes and 135 edges. Some of the co-expressed genes are detectable in the first order analysis such as ESR1, THR, PPAR, etc. But some of them can only be seen in the higher order analysis, such as VDR.

In comparison, we can also rank these quadruplets according to some other goodness-of-fit measure under the GLM framework, such as the residual deviance. Since small residual deviance means good fit, we can define the strength of dependency as the difference between the null deviance and the residual deviance. So big deviance difference corresponds to strong dependency. Figure 3.17 shows the top 3% quadruplets with the highest deviance differences. Each node still represent a candidate gene. The size of each node is proportional to the residual differences of the univariate models. Simply by comparing the node sizes in Figure 3.16 and Figure 3.17, we can see that the first order analyses based on these two dependency measure give very similar importance ranking. But when it comes to decomposing the system into quadruplets,
the structure in Figure 3.17 is way too concentrated around a few genes, and many
important genes picked out in the first order analyses are not linked into this struc-
ture, which might not be biologically reasonable. One can argue that Figure 3.17 is
less believable because the dependency measure based on deviance emphasizes the
distribution assumption too much. The residual deviances of quadruplets in Figure
3.17 all below 58, while 75% the quadruplets in Figure 3.16 have residual deviances
less than 64.

3.6 Other Possible Applications in Gene Activity Analysis

Sobol indices can be used to define statistical epistasis. One conventional way of
identifying statistical epistasis is to compare the fitting of the saturated regression
model (containing interaction effect indicators in addition to additive main effects)
with the reduced model (containing only the additive effect indicators). Statistical
epistasis is claimed if the saturated model fits the data significantly better than the
reduced model. However, validity of such inference depends on whether the models
are correctly specified, because which and how many confounding effects are adjusted
in model fitting can potentially alter the final conclusion, especially when the tested
loci are close to each other and the genotypes are correlated.

One way to make the inference less dependent on model specification is to define
statistical epistasis as the significant difference between the Sobol index with respect
to genotype indicators at all loci (including the product terms of these indicators)
and the sum of Sobol index with respect to genotype indicators at each single locus.
That is, we claim statistical epistasis if the interaction effect Sobol index is significant.
The advantage of assessing statistical epistasis using Sobol indices is that estimation
of each Sobol index only require fitting the corresponding lower-dimension projection under a large group of GLMs. If the true model has the identity link, the inference based on Sobol indices under lower-dimension projection is valid as long as other confounding factors are either independent or follow a multivariate normal distribution. If the true model has a bounded real-valued continuous inverse link, the inference based on Sobol indices is valid as long as the input variables are real-valued.

Sobol indices can be also used to quantify the effects of any combination of regulators in dynamic Bayesian networks. In the conventional dynamic Bayesian networks, the time dependency is modelled by the normal densities, assuming a child gene expression at time $i$ follows a normal distribution with the mean being a regression model in terms of its parent gene expression at time $i - 1$ (the first-order Markov dependence). The criterion for learning such networks is to estimate the regression coefficients by maximizing the posterior probability of the entire network condition on the observed data. If the regression coefficients are assumed to be independent of time, as assumed in Kim et. al. (2003) [28], such networks actually imply different regression model for different gene, but the same regression expression over time for the same gene. If the regression coefficients are assumed to be time-dependent, as assumed in the time-varying dynamic Bayesian networks [72], the fitted networks imply not only different regression for different gene, but also different regression for the same gene at different time point.

Despite the fact that there are normal densities involved in the network fitting, these induced regression expressions are technically speaking no longer the Gaussian regressions by the conventional definition, because they are not fitted to maximize a single Gaussian density. Nevertheless, we will obtain regression expression for each
gene after learning the network. So the Sobol indices can be estimated to quantify regulation effects of any combination of its parent genes under the fitted dynamic Bayesian network.
Chapter 4: Contributions and Future Work

Chapter 2 provides a novel framework to determine cases of AEI, and hence cis-acting regulatory factors, from RNA-seq data. The method is particularly useful when scanning for AEI signals in RNA-seq datasets having a large number of genes with small number of heterozygous SNPs (<10) from multiple tissues. Our method ensures that all read counts get analyzed simultaneously and all contribute to the AEI classification for each SNP. It also utilizes both the sum and the difference of the adjusted read counts while preserving the raw count ratios throughout the entire analysis. For instance, the mixture model we propose treats a pair of reads (1, 2) differently from (100, 200), while they are viewed exactly the same by ratio statistics. As a consequence, our method can also detect AEI signal that is below the commonly used ratio threshold as long as the signal is consistent and robust, in the sense that there is a sufficient number of large read differences. The robust threshold values typically applied for AEI calling using gene-based criteria seem to result in poor overlap between AEI calls based on the folded Skellam mixture and the ratio threshold approach. However, as long as its model assumptions are valid, our mixture method can make corrections in AEI calls once more data or information becomes available, which is not the case for the predetermined thresholds where the accuracy of AEI classification criterion cannot be improved regardless how much additional data
is collected. Finally, unlike the binomial-type Bayesian models, ours does not assume (or require) a strong negative correlation between reference and variant allele reads. Some drawbacks of using mixture models need to be pointed out as well. Because of the identifiability issues [140], fitting of a mixture model is often computationally challenging and expensive, and the confidence intervals obtained by MCMC or ABC type methods may be sometimes too wide for meaningful interpretation with small amount of reads. Since our mixture model provides an unsupervised AEI detection method, it is sensitive to the underlying parametric assumptions.

By applying the folded Skellam mixture model to RNA-Seq data from human autopsy brain tissues, we find that within a group of 531 “comparable” genes, 16 % SNPs in the 3UTR show AEI, which compares favorably with other similar studies. For instances, Dimas et al. analyzed allelic expression in different HapMap populations, including 60 Caucasians, 45 Chinese, 45 Japanese, and 60 Yoruba, and found approximately 18 % human genes show AEI [49]. Serre et al. performed AEI analysis on more than 80 individuals of European descent for 2,968 SNPs located in 1,380 genes, and found about 20 % human genes show AEI [58]. Most recently, Zhang et al. proposed a two component beta-binomial mixture for AEI analysis, and they concluded that approximately 17 % genes within a single individual show AEI [24]. Our present findings seem to be consistent with these results.

In Chapter 3, we showed that for a large group of GLMs, the Sobol indices can be estimated either by evaluating closed formulas or by fitting simpler models containing only partial inputs. For GLMs with polynomial systematic components, the proposed estimation strategy is as simple as fitting GLMs with observed inputs using identity link and then estimate the variance of the systematic component empirically. If the
true model has the identity link, the proposed estimation method is valid as long as other confounding factors are either independent or follow a multivariate normal distribution. If the true model has a bounded Lipschitz-continuous inverse link, the proposed estimation method is valid as long as the input variables are defined on a compact space and have Lipschitz-continuous conditional densities. In addition, this estimation strategy does not assume any specific form of the underlying complete model. In real-world applications, if the above assumptions hold, the estimation of Sobol indices comes down to finding good polynomial approximations of lower order projects (the models containing only partial inputs). The theoretical results on polynomial GLMs in Result 3.3.4 and 3.3.5 can be easily generalize to GLMs of which the inverse-link transformed systematic component can be well approximated by piece-wise polynomials, since the proofs will still hold on each piece where locally the model is just a polynomial function. That is also saying that, if a GLM can be approximated by another GLM with identity link and piece-wise polynomial systematic component, the Sobol indices under the true model can still be estimated through fitting simpler models with identity link and piece-wise polynomial systematic components on partial inputs. Moreover, all the derived formulas and the approximation results are also applicable to multi-response models (where the response is a vector instead of a scale) if the inputs are still either independent or multivariate normal. This is because these formulas are derived conditioning on knowing the regression coefficients. As long as there is a way to fit the multi-response models, the estimation of Sobol indices can be done in the exactly the same fashion.

For future studies, we can research the effect of using moment estimate in likelihood ratio test (LRT) for AEI detection. Since we evaluated the likelihood under no
AEI assumption using the moment estimate $\hat{\lambda}_{null} = \frac{1}{2n} \sum_{i=1}^{n} z_i^2$, strictly speaking, the likelihood ratio is not guaranteed to be asymptotically Chi-square with one degree of freedom. This is a difficult problem to study, majorly due to the challenges in obtaining the maximum likelihood estimates of Folded Skellam mixture model parameters, under the assumption that all mixture components are AEI component. In our application to human brain RNA-seq data, the mixture model fitting took more than two weeks. The following is a simplified simulation study for accessing the effect of using moment estimates in LRTs. 1000 Skellam random samples with size 1000 are generated with $\lambda_1 = \lambda_2 = 1$. For each sample, the likelihood ratio test statistic is computed using parameter estimates obtained by moment estimation. That is, the Poisson mean under the no AEI assumption is calculated according to $\hat{\lambda}_{null} = \frac{1}{2n} \sum_{i=1}^{n} x_i^2$; and two different Poisson mean values under the AEI assumption are estimated by $\hat{\lambda}_1 = \frac{1}{2}(\text{Mean} + \text{Variance})$ and $\hat{\lambda}_2 = \frac{1}{2}(\text{Variance} - \text{Mean})$. The left panel in Figure 4.1 shows the histogram of the LR test statistics calculated using moment estimates based on the simulation described above. And the right panel in Figure 4.1 compares the empirical cumulative density function of these LR test statistics with the theoretical cumulative density function of chi-square distribution with 1 degree of freedom. From this figure we can see that using the moment estimates of Skellam parameters did not affect much on the behavior of the LR test statistic.

For future studies, we can also continue study whether Sobol indices can be robustly estimated by fitting lower-dimensional polynomial projections when the input variables follow a multivariate skew T distribution. Given the similarity between multivariate normal distribution and the multivariate skew T distribution, we would suspect that the formulas derived for GLMs with multivariate normal inputs could
Figure 4.1: Likelihood Ratio Test Statistics Calculated Using Moment Estimates
provide good approximation of Sobol indices when the inputs are from a multivariate skew T distribution. However, the difficulty of running a sensitivity analysis for such purpose is to find a way of calculating the theoretical/exact Sobol indices under the assumption that inputs are from multivariate skew T distribution. Currently, we do not have any result that can help us obtain those exact estimates.

Other possible directions for future studies also include AEI meta-analysis methods for extracting information from multiple RNA-seq combined datasets, and investigation of Sobol indices’ performance in mixed graphical models (since in this type of models each node conditional distribution is exactly modeled by a GLM).
Bibliography


[38] Zhang, B. and Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology, **4**(1).


Appendices
Appendix A: Additional Figures and Tables of AEI Analysis

Table A.1: Summary Statistics of Reference and Variant Allele Reads Before and After Library Size Adjustment

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1st Qu.</th>
<th>Median</th>
<th>3rd Qu.</th>
<th>Max</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw_ref</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>11</td>
<td>4,667</td>
<td>11.772</td>
<td>1,174.653</td>
</tr>
<tr>
<td>adjusted_ref</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>2,805</td>
<td>8.806</td>
<td>595.878</td>
</tr>
<tr>
<td>raw_var</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>11</td>
<td>3,128</td>
<td>11.025</td>
<td>924.083</td>
</tr>
<tr>
<td>adjusted_var</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>2,413</td>
<td>8.271</td>
<td>507.409</td>
</tr>
</tbody>
</table>

NOTE: The total number of SNPs is 308,912.
Figure A.1: Scatter Plots of RNA-seq Read Pairs
Figure A.2: Histogram of Observed Absolute Read Differences with Signal Classification
Figure A.3: Q-Q Plots for Checking Folded Skellam Model Fitting

1. Histogram of abs.adjusted.reads.diff
2. Histogram of folded.SkellamMix.simulations
3. Q-Q plot (unit: 2%)
4. Q-Q plot (unit: 1%)
Table A.2: SNPs Classified in Folded Skellam Mixture Component Mix3 and Mix5

<table>
<thead>
<tr>
<th>SNP</th>
<th>ref</th>
<th>var</th>
<th>Abs.Ratio</th>
<th>Abs.Adj.Dif</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs73414847</td>
<td>306</td>
<td>79</td>
<td>3.873</td>
<td>186</td>
<td>0</td>
<td>0</td>
<td>0.9988</td>
<td>0</td>
<td>0.0012</td>
<td>0</td>
</tr>
<tr>
<td>rs998754</td>
<td>21</td>
<td>102</td>
<td>4.857</td>
<td>129</td>
<td>0</td>
<td>0</td>
<td>0.0825</td>
<td>0.2236</td>
<td>0.6938</td>
<td>0</td>
</tr>
<tr>
<td>rs77764633</td>
<td>277</td>
<td>55</td>
<td>5.036</td>
<td>133</td>
<td>0</td>
<td>0</td>
<td>0.1652</td>
<td>0.1043</td>
<td>0.7305</td>
<td>0</td>
</tr>
<tr>
<td>rs74074295</td>
<td>411</td>
<td>128</td>
<td>3.211</td>
<td>170</td>
<td>0</td>
<td>0</td>
<td>0.9861</td>
<td>0</td>
<td>0.0139</td>
<td>0</td>
</tr>
<tr>
<td>rs1045450</td>
<td>744</td>
<td>339</td>
<td>2.195</td>
<td>221</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

NOTE: “ref” and “var” are the original read counts of reference and variant alleles without the adjustment for library sizes. Abs.Ratio= Max(ref, var) / Min(ref, var). “Abs.Adj.Dif” is the absolute value of read difference between reference and variant alleles after library size adjustments. \{P_i\}, i=1,2,...,6, are the mixture probabilities corresponding to each of the six folded Skellam mixture components. Only SNPs in 3’ UTR were used for fitting folded Skellam mixture.
Table A.3: AEI Signal SNPs with Absolute Reads Ratio $\leq 1.3$

<table>
<thead>
<tr>
<th>SNPs</th>
<th>ref</th>
<th>var</th>
<th>Abs.Ratio</th>
<th>Adj.Abs.Dif</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs41147</td>
<td>66</td>
<td>58</td>
<td>1.14</td>
<td>28</td>
<td>0.471</td>
<td>0.526</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>rs41147</td>
<td>129</td>
<td>108</td>
<td>1.19</td>
<td>29</td>
<td>0.434</td>
<td>0.561</td>
<td>0</td>
<td>0.004</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>rs41147</td>
<td>189</td>
<td>153</td>
<td>1.24</td>
<td>33</td>
<td>0.284</td>
<td>0.704</td>
<td>0</td>
<td>0.012</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>rs12574994</td>
<td>86</td>
<td>70</td>
<td>1.23</td>
<td>30</td>
<td>0.387</td>
<td>0.607</td>
<td>0</td>
<td>0.005</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>rs3733398</td>
<td>500</td>
<td>429</td>
<td>1.17</td>
<td>33</td>
<td>0.284</td>
<td>0.704</td>
<td>0</td>
<td>0.012</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>rs2021320</td>
<td>88</td>
<td>69</td>
<td>1.28</td>
<td>32</td>
<td>0.316</td>
<td>0.674</td>
<td>0</td>
<td>0.009</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>rs2021320</td>
<td>755</td>
<td>636</td>
<td>1.19</td>
<td>34</td>
<td>0.246</td>
<td>0.738</td>
<td>0</td>
<td>0.016</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>rs2269272</td>
<td>210</td>
<td>163</td>
<td>1.29</td>
<td>34</td>
<td>0.246</td>
<td>0.738</td>
<td>0</td>
<td>0.016</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>rs3749538</td>
<td>179</td>
<td>227</td>
<td>1.27</td>
<td>28</td>
<td>0.471</td>
<td>0.526</td>
<td>0</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

NOTE: “ref” and “var” are the original read counts of reference and variant alleles without the adjustment for library sizes. Abs.Ratio = Max(ref, var) / Min(ref, var). “Abs.Adj.Dif” is the absolute value of read difference between reference and variant alleles after library size adjustments. \( \{P_i\} \), \( i=1,2,...,6 \), are the mixture probabilities corresponding to each of the six folded Skellam mixture components. Only SNPs in 3' UTR were used for fitting folded Skellam mixture.
Table A.4: Uncertain Signal SNPs with Absolute Reads Ratio $\geq 7$

<table>
<thead>
<tr>
<th>SNP</th>
<th>ref</th>
<th>Var</th>
<th>Abs.ratio</th>
<th>Adj.Abs.Dif</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs115141047</td>
<td>30</td>
<td>4</td>
<td>8</td>
<td>24</td>
<td>0.622</td>
<td>0.377</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>rs35674</td>
<td>3</td>
<td>24</td>
<td>8</td>
<td>12</td>
<td>0.898</td>
<td>0.097</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.005</td>
<td>1</td>
</tr>
<tr>
<td>rs1055253</td>
<td>36</td>
<td>5</td>
<td>7</td>
<td>24</td>
<td>0.622</td>
<td>0.377</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>rs7566</td>
<td>5</td>
<td>39</td>
<td>8</td>
<td>24</td>
<td>0.622</td>
<td>0.377</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>rs9347</td>
<td>3</td>
<td>27</td>
<td>9</td>
<td>23</td>
<td>0.661</td>
<td>0.339</td>
<td>0</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>rs1126878</td>
<td>22</td>
<td>3</td>
<td>7</td>
<td>18</td>
<td>0.803</td>
<td>0.196</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>rs7020</td>
<td>3</td>
<td>26</td>
<td>9</td>
<td>27</td>
<td>0.512</td>
<td>0.486</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>rs10898931</td>
<td>3</td>
<td>29</td>
<td>10</td>
<td>21</td>
<td>0.727</td>
<td>0.272</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>rs2016057</td>
<td>3</td>
<td>23</td>
<td>8</td>
<td>14</td>
<td>0.875</td>
<td>0.125</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>1</td>
</tr>
<tr>
<td>rs2554315</td>
<td>3</td>
<td>24</td>
<td>8</td>
<td>14</td>
<td>0.875</td>
<td>0.125</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>1</td>
</tr>
</tbody>
</table>
Appendix B: Proofs of Inverse-logit Function Expectations

Result B.0.1. Expectations of Functions of Univariate Normal with Zero Mean

Suppose \( X \sim N(0, \sigma^2) \), \( Z = e^X \sim \ln N(0, \sigma^2) \), we have:

1. \[ E \left( \frac{e^X}{1+e^X} \right) = E \left( \frac{Z}{1+Z} \right) = E \left( \frac{1}{1+Z} \right) = \frac{1}{2} \]

2. \[ E \left( \frac{e^{kX}}{1+e^X} \right) = E \left( \frac{Z^k}{1+Z} \right) = E \left( \frac{Z^{1-k}}{1+Z} \right) \]

   \[ = (-1)^{|k|} E \left( \frac{Z^{s-[s]}}{1+Z} \right) + \sum_{i=1}^{|s|} (-1)^{i-1} e^{\frac{1}{2}Z(s-i)^2\sigma^2}, \quad s = \begin{cases} k, & \text{if } k > 1 \\ 1-k, & \text{if } k \in \mathbb{R}^- \end{cases} \]

   \[ = (-1)^{s-\frac{1}{2}} + \sum_{i=1}^{s-1} (-1)^{i-1} e^{\frac{1}{2}Z(s-i)^2\sigma^2}, \quad s = \begin{cases} k, & \text{if } k \in \mathbb{Z}^+ \setminus \{1\} \\ 1-k, & \text{if } k \in \mathbb{Z}^- \end{cases} \]

3. \[ E \left( \frac{Z^2}{(1+Z)^2} \right) = E \left( \frac{1}{(1+Z)^2} \right) = \frac{1}{2} - E \left( \frac{Z}{(1+Z)^2} \right) \]

4. \[ E \left( \frac{e^{kX}}{(1+e^X)^2} \right) = E \left( \frac{Z^k}{(1+Z)^2} \right) = E \left( \frac{Z^{2-k}}{(1+Z)^2} \right) \]

   \[ = (-1)^{|k|} E \left( \frac{Z^{s-[s]}}{1+Z} \right) + \sum_{i=1}^{s-2} (-1)^{i-1} e^{\frac{1}{2}Z(s-i)^2\sigma^2} + \sum_{i=1}^{s-1} (-1)^{i-1} (-1)^{s-1} \frac{Z^e}{(1+Z)^2}, \quad s = \begin{cases} k, & \text{if } k > 2 \\ 2-k, & \text{if } k \in \mathbb{R}^- \end{cases} \]

   \[ = (-1)^{s-2-s-\frac{1}{2}} + \sum_{i=1}^{s-2} (-1)^{i-1} e^{\frac{1}{2}Z(s-i)^2\sigma^2} + (-1)^{s-1} E \left( \frac{Z}{(1+Z)^2} \right), \quad s = \begin{cases} k, & \text{if } k \in \mathbb{Z}^+ \setminus \{1, 2\} \\ 2-k, & \text{if } k \in \mathbb{Z}^- \end{cases} \]
Proof.

1. Since $Z$ and $\frac{1}{Z}$ both follow $\ln N(0, \sigma^2)$, we have

$$E\left(\frac{Z}{1+Z}\right) = E\left(\frac{\frac{1}{Z}}{1+\frac{1}{Z}}\right) = E\left(\frac{1}{1+Z}\right)$$

Additionally, since

$$E\left(\frac{Z}{1+Z}\right) + E\left(\frac{1}{1+Z}\right) = 1$$

we have

$$E\left(\frac{Z}{1+Z}\right) = E\left(\frac{1}{1+Z}\right) = \frac{1}{2}$$

2. Since $Z$ and $\frac{1}{Z}$ both follow $\ln N(0, \sigma^2)$, we have

$$E\left(\frac{Z^k}{1+Z}\right) = E\left(\frac{\frac{1}{Z^k}}{1+\frac{1}{Z}}\right) = E\left(\frac{Z^{1-k}}{1+Z}\right)$$

Since if $k > 1$

$$E\left(\frac{Z^k}{1+Z}\right) = E\left(\frac{Z^{k-1} (1 + Z)}{1+Z} - \frac{Z^{k-1}}{1+Z}\right)$$

$$= E\left(Z^{k-1}\right) - E\left(\frac{Z^{k-1}}{1+Z}\right)$$

$$= E\left(Z^{k-1}\right) - \left[E\left(Z^{k-2}\right) - E\left(\frac{Z^{k-2}}{1+Z}\right)\right]$$

$$= \ldots$$

$$= E\left(Z^{k-1}\right) - \left[E\left(Z^{k-2}\right) - \left[E\left(Z^{k-3}\right) \ldots \right\} - \left[E\left(Z^{k-\lfloor k\rfloor}\right) - E\left(\frac{Z^{k-\lfloor k\rfloor}}{1+Z}\right)\right] \ldots \right\} \right\}, \forall k > 1, k \in \mathbb{R}^+$$

$$= E\left(Z^{k-1}\right) - \left[E\left(Z^{k-2}\right) - \left[E\left(Z^{k-3}\right) \ldots \right\} - \left[E\left(Z\right) - E\left(\frac{Z}{1+Z}\right)\right] \ldots \right\} \right\}, k \in \mathbb{Z}^+ - \{1\}$$

and

$$Z^k \sim \ln N(0, k^2 \sigma^2), \quad E\left(Z^k\right) = e^{\frac{1}{2}k^2\sigma^2}, \quad E\left(\frac{Z}{1+Z}\right) = \frac{1}{2}$$
we have:

\[ E \left( \frac{Z^k}{1 + Z} \right) = e^{\frac{1}{2}(k-1)^2\sigma^2} - \left[ e^{\frac{1}{2}(k-2)^2\sigma^2} - \left[ e^{\frac{1}{2}(k-3)^2\sigma^2} \ldots \right. \right. \]

\[ \left. \left. \left. \left. \left. \ldots - \left( E \left( \frac{Z^{k-\lfloor k \rfloor}}{1 + Z} \right) \right) \ldots \right]\right]\right], \]

if \( k > 1, \ k \in \mathbb{R}^+ \)

\[ E \left( \frac{Z^k}{1 + Z} \right) = e^{\frac{1}{2}(k-1)^2\sigma^2} - \left[ e^{\frac{1}{2}(k-2)^2\sigma^2} - \left[ e^{\frac{1}{2}(k-3)^2\sigma^2} \ldots - \left( e^{\frac{1}{2}\sigma^2 - \frac{1}{2}} \right) \ldots \right]\right]\right], \]

if \( k \in \mathbb{Z}^+ - \{1\} \)

That is,

\[ E \left( \frac{e^{kX}}{1 + e^X} \right) = E \left( \frac{Z^k}{1 + Z} \right) = E \left( \frac{Z^{1-k}}{1 + Z} \right) \]

\[ = (-1)^{|s|} E \left( \frac{Z^{s-|s|}}{1 + Z} \right) + \sum_{i=1}^{|s|} (-1)^{i-1} e^{\frac{1}{2}(s-i)^2\sigma^2}, \quad s = \begin{cases} k, & \text{if } k > 1 \\ 1 - k, & \text{if } k \in \mathbb{R}^- \end{cases} \]

\[ = (-1)^{s-1} \frac{1}{2} + \sum_{i=1}^{s-1} (-1)^{i-1} e^{\frac{1}{2}(s-i)^2\sigma^2}, \quad s = \begin{cases} k, & \text{if } k \in \mathbb{Z}^+ - \{1\} \\ 1 - k, & \text{if } k \in \mathbb{Z}^- \end{cases} \]

3. Since

\[ E \left( \frac{Z^2}{(1 + Z)^2} \right) = E \left( \frac{Z(Z + 1)}{(1 + Z)^2} \right) - E \left( \frac{Z}{(1 + Z)^2} \right) = \frac{1}{2} - E \left( \frac{Z}{(1 + Z)^2} \right) \]

and

\[ E \left( \frac{Z^2}{(1 + Z)^2} \right) = E \left( \frac{Z^2}{(1 + Z)^2} \right) = E \left( \frac{1}{(1 + Z)^2} \right) \]

we have

\[ E \left( \frac{Z^2}{(1 + Z)^2} \right) = E \left( \frac{1}{(1 + Z)^2} \right) = \frac{1}{2} - E \left( \frac{Z}{(1 + Z)^2} \right) \]

4. Since \( Z \) and \( \frac{1}{Z} \) both follow \( \text{ln } N(0, \sigma^2) \), we have

\[ E \left( \frac{Z^k}{(1 + Z)^2} \right) = E \left( \frac{\frac{1}{Z^k}}{(1 + \frac{1}{Z})^2} \right) = E \left( \frac{Z^{2-k}}{(1 + Z)^2} \right) \]
Since if $k > 1$

\[ E \left( \frac{Z^k}{(1+Z)^2} \right) = E \left( \frac{Z^{k-1} (1+Z)}{(1+Z)^2} - \frac{Z^{k-1}}{(1+Z)^2} \right) = E \left( \frac{Z^{k-1}}{1+Z} \right) - E \left( \frac{Z^{k-1}}{(1+Z)^2} \right) = E \left( \frac{Z^{k-1}}{1+Z} \right) - \left( E \left( \frac{Z^{k-2}}{1+Z} \right) - E \left( \frac{Z^{k-2}}{(1+Z)^2} \right) \right) = \cdots \]

\[ = E \left( \frac{Z^{k-1}}{1+Z} \right) - \left[ E \left( \frac{Z^{k-2}}{1+Z} \right) - \left[ E \left( \frac{Z^{k-3}}{1+Z} \right) \cdots - \left( E \left( \frac{Z^{k-[k]}}{1+Z} \right) - E \left( \frac{k-[k]}{(1+Z)^2} \right) \right) \right] \right], \quad k > 1, \ k \in \mathbb{R}^+ \]

\[ = E \left( \frac{Z^{k-1}}{1+Z} \right) - \left[ E \left( \frac{Z^{k-2}}{1+Z} \right) - \left[ E \left( \frac{Z^{k-3}}{1+Z} \right) \cdots - \left( E \left( \frac{Z}{1+Z} \right) - E \left( \frac{Z}{(1+Z)^2} \right) \right) \right] \right], \quad k \in \mathbb{Z}^+ - \{1\} \]

and

\[ E \left( \frac{e^{kX}}{1+e^X} \right) = E \left( \frac{Z^k}{1+Z} \right) = E \left( \frac{Z^{1-k}}{1+Z} \right) \]

\[ = (-1)^{|s|} E \left( \frac{Z^{s-[s]}}{1+Z} \right) + \sum_{i=1}^{[s]} (-1)^{i-1} e^{\frac{1}{2}(s-i)^2 \sigma^2}, \quad s = \begin{cases} k, & \text{if } k > 1 \\ 1-k, & \text{if } k \in \mathbb{R}^- \end{cases} \]

\[ = (-1)^{s-1} \frac{1}{2} + \sum_{i=1}^{s-1} (-1)^{i-1} e^{\frac{1}{2}(s-i)^2 \sigma^2}, \quad s = \begin{cases} k, & \text{if } k \in \mathbb{Z}^+ - \{1\} \\ 1-k, & \text{if } k \in \mathbb{Z}^- \end{cases} \]

we have:

\[ E \left( \frac{e^{kX}}{(1+e^X)^2} \right) = E \left( \frac{Z^k}{(1+Z)^2} \right) = E \left( \frac{Z^{2-k}}{(1+Z)^2} \right) \]

\[ = (-1)^{|s|-1} ([s] - 1) E \left( \frac{Z^{s-[s]}}{1+Z} \right) + \sum_{i=1}^{[s]-1} (-1)^{i-1} e^{\frac{1}{2}(s-i)^2 \sigma^2} + (-1)^{|s|-1} E \left( \frac{Z^{s-[s]+1}}{(1+Z)^2} \right), \quad s = \begin{cases} k, & \text{if } k > 2 \\ 2-k, & \text{if } k \in \mathbb{R}^- \end{cases} \]

\[ = (-1)^{s-2} \frac{s-1}{2} + \sum_{i=1}^{s-2} (-1)^{i-1} e^{\frac{1}{2}(s-i)^2 \sigma^2} + (-1)^{s-1} E \left( \frac{Z}{(1+Z)^2} \right), \quad s = \begin{cases} k, & \text{if } k \in \mathbb{Z}^+ - \{1,2\} \\ 2-k, & \text{if } k \in \mathbb{Z}^- \end{cases} \]
Result B.0.2. *Expectations of Functions of Univariate Normal with non-Zero Mean*

Suppose $X \sim N(\mu, \sigma^2)$, $\mu \neq 0$, $U = e^X \sim \ln N(\mu, \sigma^2)$, $V = \frac{1}{U} = e^{-X} \sim \ln N(-\mu, \sigma^2)$ and $Z \sim \ln N(0, \sigma^2)$, we have:

1. $E\left(\frac{e^X}{1 + e^X}\right) = E\left(\frac{U}{1 + U}\right) = E\left(\frac{1}{1 + V}\right)$

$$= e^{-\frac{\mu^2}{2\sigma^2}} E\left(\frac{Z^{1+\frac{\mu}{\sigma^2}}}{1 + Z}\right) = e^{-\frac{\mu}{\sigma^2}} E\left(\frac{Z^{-\frac{\mu}{\sigma^2}}}{1 + Z}\right), \quad \forall \frac{\mu}{\sigma^2} \in \mathbb{R}$$

$$= e^{-\frac{\mu^2}{2\sigma^2}} \left[ (s - 1) \left\lfloor \frac{s - 1}{2} \right\rfloor + \sum_{i=1}^{s-1} (s - i - 1) e^{\frac{1}{2} (s - i)^2 \sigma^2} \right],$$

$$s = \begin{cases} 1 + \frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbb{Z}^+ \\ -\frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbb{Z}^- \end{cases}$$

$$= e^{-\frac{\mu^2}{2\sigma^2}} \left[ (s - 1) \left\lfloor \frac{s - 1}{2} \right\rfloor E\left(\frac{Z^{s-\left\lfloor \frac{s}{2} \right\rfloor}}{1 + Z}\right) + \sum_{i=1}^{\left\lfloor \frac{s}{2} \right\rfloor} (s - i - 1) e^{\frac{1}{2} (s - i)^2 \sigma^2} \right],$$

$$s = \begin{cases} 1 + \frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbb{R}^+ \\ -\frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} < -1 \end{cases}$$

140
2. $E\left(\frac{e^{2X}}{(1 + e^X)^2}\right) = E\left(\frac{U^2}{(1 + U)^2}\right) = E\left(\frac{1}{(1 + V)^2}\right)$

$= e^{-\frac{\mu^2}{2\sigma^2}} E\left(\frac{Z^{2+\frac{\mu^2}{2\sigma^2}}}{(1 + Z)^2}\right) = e^{-\frac{\mu^2}{2\sigma^2}} E\left(\frac{Z^{-\frac{\mu}{\sigma^2}}}{(1 + Z)^2}\right)$

$= e^{-\frac{\mu^2}{2\sigma^2}} \left[ (-1)^{s-2} s - 1 \frac{1}{2} + \sum_{i=1}^{s-2} (-1)^{i-1} i e^{\frac{1}{2}(s-1-i)^2\sigma^2} \right. $ 

$\left. + (-1)^{s-1} E\left(\frac{Z}{(1 + Z)^2}\right) \right],

$s = \begin{cases} 2 + \frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in Z^+ \\ -\frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in Z^- \setminus \{-1, -2\} \end{cases}$

$= e^{-\frac{\mu^2}{2\sigma^2}} \left[ (-1)^{|s| - 2} (|s| - 1) E\left(\frac{Z^{s-|s|}}{1 + Z}\right) + \sum_{i=1}^{|s|-2} (-1)^{i-1} i e^{\frac{1}{2}(s-1-i)^2\sigma^2} + (-1)^{|s| - 1} E\left(\frac{Z^{s-|s|}}{1 + Z}\right) \right],

$s = \begin{cases} 2 + \frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in R^+ \\ -\frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} < -2 \end{cases}$

Proof.

1. Since $Z$ and $\frac{1}{Z}$ both follow $\ln N(0, \sigma^2)$ and

$E\left(\frac{U}{1 + U}\right) = E\left(\frac{1}{1 + \frac{1}{U}}\right) = E\left(\frac{1}{1 + V}\right)$

we have

$E\left(\frac{U}{1 + U}\right) = E\left(\frac{1}{1 + V}\right) = \int_0^{+\infty} \frac{U}{1 + U} \frac{1}{U\sigma\sqrt{2\pi}} e^{-\frac{(\ln U - \mu)^2}{2\sigma^2}} dU$

$= e^{-\frac{\mu^2}{2\sigma^2}} \left\{ \int_0^{+\infty} \left[ \frac{U}{1 + U} e^{\frac{2\mu \ln U}{2\sigma^2}} \right] \frac{1}{U\sigma\sqrt{2\pi}} e^{-\frac{(-\ln U)^2}{2\sigma^2}} dU \right\}$

$= e^{-\frac{\mu^2}{2\sigma^2}} \left\{ \int_0^{+\infty} \left[ \frac{Z}{1 + Z} e^{\frac{2\mu \ln Z}{2\sigma^2}} \right] \frac{1}{Z\sigma\sqrt{2\pi}} e^{-\frac{(-\ln Z)^2}{2\sigma^2}} dZ \right\}$

$= e^{-\frac{\mu^2}{2\sigma^2}} E\left(\frac{Z}{1 + Z} \frac{\mu}{Z\sigma^2}\right) = e^{-\frac{\mu^2}{2\sigma^2}} E\left(\frac{Z^{1+\frac{\mu}{\sigma^2}}}{1 + Z}\right)$

$= e^{-\frac{\mu^2}{2\sigma^2}} E\left(\frac{1}{1 + \frac{1}{Z}}\right) = e^{-\frac{\mu^2}{2\sigma^2}} E\left(\frac{Z^{-\frac{\mu}{\sigma^2}}}{1 + Z}\right)$
By applying the 2nd bullet of Result B.0.1, we have:
\[
E \left( \frac{e^X}{1 + e^X} \right) = E \left( \frac{U}{1 + U} \right) = E \left( \frac{1}{1 + V} \right)
\]
\[
= e^{-\frac{\mu^2}{2\sigma^2}} E \left( \frac{Z^{1+\frac{\mu}{\sigma^2}}}{1 + Z} \right) = e^{-\frac{\mu^2}{2\sigma^2}} E \left( \frac{Z^{-\frac{\mu}{\sigma^2}}}{1 + Z} \right), \quad \forall \frac{\mu}{\sigma^2} \in \mathbb{R}
\]
\[
= e^{-\frac{\mu^2}{2\sigma^2}} \left[ (-1)^{s-1} \frac{1}{2} + \sum_{i=1}^{s-1} (-1)^{i-1} e^{\frac{1}{2}(s-i)\sigma^2} \right],
\]
\[
s = \begin{cases} 
1 + \frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbb{Z}^+ \\
-\frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbb{Z}^-
\end{cases}
\]
\[
= e^{-\frac{\mu^2}{2\sigma^2}} \left[ (-1)^{|s|} E \left( \frac{Z^{1-|s|}}{1 + Z} \right) + \sum_{i=1}^{|s|} (-1)^{i-1} e^{\frac{1}{2}(s-i)\sigma^2} \right],
\]
\[
s = \begin{cases} 
1 + \frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbb{R}^+ \\
-\frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} < -1
\end{cases}
\]

2. Since \(Z\) and \(\frac{1}{Z}\) both follow \(\ln \mathcal{N}(0, \sigma^2)\) and

\[
E \left( \frac{U}{1 + U} \right) = E \left( \frac{1}{1 + \frac{1}{U}} \right) = E \left( \frac{1}{(1 + V)^2} \right)
\]
we have

\[
E \left( \frac{U^2}{(1 + U)^2} \right) = E \left( \frac{1}{(1 + V)^2} \right) = \int_{0}^{+\infty} \frac{U^2}{(1 + U)^2} \frac{1}{U \sqrt{2\pi}} e^{-\frac{(\ln U - \mu)^2}{2\sigma^2}} dU
\]
\[
= e^{-\frac{\mu^2}{2\sigma^2}} \left\{ \int_{0}^{+\infty} \left[ \frac{U^2}{(1 + U)^2} \right] e^{\frac{2\mu \ln U}{2\sigma^2}} \frac{1}{U \sqrt{2\pi}} e^{-\frac{(\ln U)^2}{2\sigma^2}} dU \right\}
\]
\[
= e^{-\frac{\mu^2}{2\sigma^2}} \left\{ \int_{0}^{+\infty} \left[ \left( \frac{Z^2}{(1 + Z)^2} \right) \right] e^{\frac{2\mu \ln Z}{2\sigma^2}} \frac{1}{Z \sqrt{2\pi}} e^{-\frac{(\ln Z)^2}{2\sigma^2}} dZ \right\}
\]
\[
= e^{-\frac{\mu^2}{2\sigma^2}} E \left( \frac{Z^2}{(1 + Z)^2} \right) = e^{-\frac{\mu^2}{2\sigma^2}} E \left( \frac{Z^{2+\frac{\mu}{\sigma^2}}}{(1 + Z)^2} \right)
\]
\[
= e^{-\frac{\mu^2}{2\sigma^2}} \left( \frac{1}{Z^{2+\frac{\mu}{\sigma^2}}} \right) = e^{-\frac{\mu^2}{2\sigma^2}} E \left( \frac{Z^{-\frac{\mu}{\sigma^2}}}{(1 + Z)^2} \right)
\]

By applying the 4th bullet of Result B.0.1, we have:
\[
E \left( \frac{e^{2X}}{(1 + e^{X})^2} \right) = E \left( \frac{U^2}{(1 + U)^2} \right) = E \left( \frac{1}{(1 + V)^2} \right) \\
= e^{-\mu^2/2\sigma^2} E \left( \frac{Z^{2+\mu/\sigma^2}}{(1 + Z)^2} \right) = e^{-\mu^2/2\sigma^2} E \left( \frac{Z^{-\mu/\sigma^2}}{(1 + Z)^2} \right) \\
= e^{-\mu^2/2\sigma^2} \left[ (-1)^{s-2} \frac{s - 1}{2} + \sum_{i=1}^{s-2} (-1)^{i-1} i e^{\frac{1}{2}(s-1-i)\sigma^2} \right. \\
\left. + (-1)^{s-1} E \left( \frac{Z}{(1 + Z)^2} \right) \right], \\
\\
s = \begin{cases} 
2 + \frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbb{Z}^+ \\
-\frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbb{Z}^- - \{-1, -2\} 
\end{cases} \\
= e^{-\mu^2/2\sigma^2} \left[ (-1)^{|s|} - 2(|s| - 1) E \left( \frac{Z^{s-|s|}}{1 + Z} \right) + \sum_{i=1}^{|s|} (-1)^{i-1} i e^{\frac{1}{2}(s-1-i)\sigma^2} \right. \\
\left. + (-1)^{|s|} - 1 E \left( \frac{Z^{s-|s|}}{(1 + Z)^2} \right) \right], \\
\\
s = \begin{cases} 
2 + \frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbb{R}^+ \\
-\frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} < -2 
\end{cases}
\]

Appendix C: Proofs of Sobol Index Formulas under Linear GLMs

Result C.0.3. Sobol Indices under Linear GLMs with Identity Link. If \( E[Y|X] = X^T\beta \) and the inputs follow a multivariate normal distribution \( N(\mu, \Sigma) \) where \( \mu = (\mu_1, \mu_2, \cdots, \mu_n)^T \), \( \Sigma_{ii} = \sigma_i^2 \), \( \Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j \), the main-effect Sobol index with respect to single input has the following closed form:

\[
\frac{\text{Var}(E(Y|X_i))}{\text{Var}(Y)} = \left( \frac{\beta_i + \frac{1}{\sigma_i} \sum_{j \neq i} \beta_j \rho_{ij} \sigma_j}{\text{Var}(X_i)} \right)^2 \frac{\text{Var}(X_i)}{\text{Var}(Y)} \tag{C.1}
\]

Let \( X_P = (X_{i_1}, \cdots, X_{i_p})^T \), and \( X_Q \) be the input vector containing the remaining X’s. Then the main-effect Sobol index with respect to input subset \( X_P \) has the following closed form:

\[
\frac{\text{Var}(E(Y|X_P))}{\text{Var}(Y)} = \frac{\eta^T \Sigma_{PP} \eta}{\text{Var}(Y)} \tag{C.2}
\]

where

\[
\eta = \beta_P + \Sigma_{PP}^{-1}\Sigma_{PQ} \beta_Q
\]

and \( \begin{bmatrix} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{bmatrix} \) is the partition of \( \Sigma \) corresponding to input vector partition \( X = (X_P^T, X_Q^T)^T \).
Proof. Since \( E(Y|X_1, \cdots, X_n) = E(Y|X^T\beta) \) under GLMs, we have:

\[
E(Y|X_i) = E(E(Y|X_1, \cdots, X_n)|X_i) = E(E(Y|X^T\beta)|X_i) \\
= E(X^T\beta|X_i) = \beta_0 + \beta_i X_i + \sum_{j \neq i}^{n} \beta_j E(X_j|X_i) \\
= \beta_0 + \beta_i X_i + \sum_{j \neq i}^{n} \beta_j \left[ \mu_j + \rho_{ij} \frac{\sigma_j}{\sigma_i} (X_i - \mu_i) \right]
\]

Thus,

\[
E(Y|X_i) = \left( \beta_i + \frac{1}{\sigma_i} \sum_{j \neq i}^{n} \beta_j \rho_{ji} \sigma_j \right) X_i + \left( \beta_0 - \sum_{j \neq i}^{n} \beta_j \mu_i \rho_{ij} \frac{\sigma_j}{\sigma_i} \right)
\]

Therefore,

\[
\frac{\text{Var}(E(Y|X_i))}{\text{Var}(Y)} = \left( \beta_i + \frac{1}{\sigma_i} \sum_{j \neq i}^{n} \beta_j \rho_{ji} \sigma_j \right)^2 \frac{\text{Var}(X_i)}{\text{Var}(Y)}
\]

Similarly, since

\[
E(Y|X_P) = E(E(Y|X_1, \cdots, X_n)|X_P) = E \left( X^T \beta \mid X_P \right) \\
= \beta_0 + X_P^T \beta_P + E \left( X_Q^T \beta_Q \mid X_P \right) \\
= \beta_0 + X_P^T \beta_P + E(X_Q^T X_P)^T \beta_Q \\
= \beta_0 + X_P^T \beta_P + (\mu_Q + \Sigma_{QP} \Sigma_{PP}^{-1} (X_P - \mu_P))^T \beta_Q \\
= \text{constant} + X_P^T (\beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q)
\]

we have

\[
\frac{\text{Var}(E(Y|X_P))}{\text{Var}(Y)} = \frac{(\beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q)^T \Sigma_{PP} (\beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q)}{\text{Var}(Y)}
\]

\[\square\]
Result C.0.4. **Sobol Indices under Linear GLMs with Log Link.** If $\ln(E[Y|X]) = X^T\beta$ and the inputs follow a multivariate normal distribution $N(\mu, \Sigma)$ where $\mu = (\mu_1, \mu_2, \cdots, \mu_n)^T$, $\Sigma_{ii} = \sigma_i^2$, $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, the main-effect Sobol index with respect to single input has the following closed form:

$$
\frac{\text{Var}(E(Y|X_i))}{\text{Var}(Y)} = \frac{1}{\text{Var}(Y)} \left( e^{\sigma_i^2} - 1 \right) e^{2\beta_0 + 2K_2^{(i)}} + 2\mu_\ast + \sigma_i^2
$$

where

$$
\mu_\ast = \left( \beta_i + \frac{\sum_{j \neq i}^n \beta_j \rho_{ij} \sigma_j}{\sigma_i} \right) \mu_i,
\sigma_i^2 = \left( \beta_i + \frac{\sum_{j \neq i}^n \beta_j \rho_{ij} \sigma_j}{\sigma_i} \right)^2 \sigma_i^2
$$

$$
K_2^{(i)} = \sum_{j \neq i}^n \beta_j \left( \mu_j - \mu_i \rho_{ji} \sigma_j \sigma_i \right) + \frac{1}{2} \beta_i^T \left( \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP} \Sigma_{QP} \right) \beta_i
$$

$\beta_{-i} = (\beta_1, \beta_2, \cdots, \beta_{i-1}, \beta_{i+1}, \cdots, \beta_n)^T$ and $\begin{bmatrix} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{bmatrix}$ is the partition of $\Sigma$ corresponding to the input vector partition $X = (X_P, X_Q = X_{-i}^T)^T$.

Let $X_P = (X_{i_1}, \cdots, X_{i_p})^T$, and $X_Q$ be the input vector containing the remaining $X$'s. Then the main-effect Sobol index with respect to input subset $X_P$ has the following closed form:

$$
\frac{\text{Var}(E(Y|X_P))}{\text{Var}(Y)} = \frac{1}{\text{Var}(Y)} \left( e^{\sigma_\ast^2} - 1 \right) e^{2\beta_0 + 2K_2^{(P)}} + 2\mu_\ast + \sigma_\ast^2
$$

where

$$
\mu_{\ast\ast} = \mu_P^T \left( \beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q \right)
$$

$$
\sigma_{\ast\ast}^2 = \left( \beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q \right)^T \Sigma_{PP} \left( \beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q \right)
$$

$$
K_2^{(P)} = \left( \mu_Q - \Sigma_{QP} \Sigma_{PP} \Sigma_{PQ} \beta_Q \right)^T \beta_Q + \frac{1}{2} \beta_Q^T \left( \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP} \Sigma_{QP} \Sigma_{QQ} \right) \beta_Q
$$

and $\begin{bmatrix} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{bmatrix}$ is the partition of $\Sigma$ corresponding to the input vector partition $X = (X_P^T, X_Q^T)^T$.  

146
Proof. Since \( E(Y|X_1, \cdots, X_n) = E(Y|X^T\beta) \) under GLMs, we have:

\[
E(Y|X_i) = E(E(Y|X_1, \cdots, X_n)|X_i) = E(E(Y|X^T\beta)|X_i)
\]

\[
= E(\exp(X^T\beta)|X_i) = \exp(\beta_0 + \beta_i X_i)E\left( \exp \left( \sum_{j \neq i} \beta_j X_j \right) | X_i \right)
\]

\[
= \exp(\beta_0 + \beta_i X_i)E \left( \exp \left( X^T_i \beta_{-i} \right) | X_i \right)
\]

Since the conditional distribution of \( X^T_i \beta_{-i} | X_i \) is a univariate normal, the conditional distribution of \( e^{X^T_i \beta_{-i}} | X_i \) is a Log-normal distribution whose expectation can be written as a function of the mean and variance of \( X^T_i \beta_{-i} | X_i \), i.e.

\[
E \left( e^{X^T_i \beta_{-i}} | X_i \right) = e^{E(X^T_i \beta_{-i}|X_i) + \frac{1}{2}Var(X^T_i \beta_{-i}|X_i)}
\]

Since

\[
E \left( X^T_i \beta_{-i} | X_i \right) = E \left( X_{-i} | X_i \right)^T \beta_{-i}
\]

\[
= \sum_{j \neq i}^{n} \beta_j \left( \mu_j + \rho_{ji} \frac{\sigma_j}{\sigma_i} (X_i - \mu_i) \right)
\]

\[
= \sum_{j \neq i}^{n} \beta_j \rho_{ji} \frac{\sigma_j}{\sigma_i} X_i + \sum_{j \neq i}^{n} \beta_j \left( \mu_j - \mu_i \rho_{ji} \frac{\sigma_j}{\sigma_i} \right)
\]

\[
= \sum_{j \neq i}^{n} \beta_j \rho_{ji} \frac{\sigma_j}{\sigma_i} X_i + K_1^{(i)}
\]

\[
Var \left( X^T_i \beta_{-i} | X_i \right) = \beta_i^T \Sigma_{Q|P} \beta_{-i}
\]

where

\[
\Sigma_{Q|P} = \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{QP}
\]

\[
\begin{bmatrix}
\Sigma_{PP} & \Sigma_{PQ} \\
\Sigma_{QP} & \Sigma_{QQ}
\end{bmatrix}
\]

is the partition of \( \Sigma \) corresponding to input vector partition \( X = (X_P = X_i, X_Q = X_{-i})^T \), we have:

\[
E(Y|X_i) = e^{\beta_0 + \beta_i X_i} e^{\sum_{j \neq i}^{n} \beta_j \rho_{ji} \frac{\sigma_j}{\sigma_i} X_i + K_2^{(i)}}
\]

\[
= e^{\beta_0 + K_2^{(i)}} e^{\left( \beta_i + \sum_{j \neq i}^{n} \beta_j \rho_{ji} \frac{\sigma_j}{\sigma_i} \right) X_i}
\]

147
where \( K^{(i)}_2 = K^{(i)}_1 + \frac{1}{2} \beta^T_i \Sigma_{Q|P} \beta_i \). Note that the values of \( K^{(i)}_2 \) will differ depending on which input variable is chosen to be \( X_i \).

Since

\[
X_* = e^{\left( \beta_i + \sum_{j \neq i} \frac{\beta_j \rho_{ij} \sigma_j}{\sigma_i} \right) X_i} \sim \ln N(\mu_*, \sigma^2_*)
\]

\[
\mu_* = E \left[ \left( \beta_i + \sum_{j \neq i} \frac{\beta_j \rho_{ij} \sigma_j}{\sigma_i} \right) X_i \right] = \left( \beta_i + \sum_{j \neq i} \frac{\beta_j \rho_{ij} \sigma_j}{\sigma_i} \right) \mu_i
\]

\[
\sigma^2_* = \text{Var} \left[ \left( \beta_i + \sum_{j \neq i} \frac{\beta_j \rho_{ij} \sigma_j}{\sigma_i} \right) X_i \right] = \left( \beta_i + \sum_{j \neq i} \frac{\beta_j \rho_{ij} \sigma_j}{\sigma_i} \right)^2 \sigma_i^2
\]

\[
\text{Var}(X_*) = (e^{\sigma^2_*} - 1) e^{2\mu_* + \sigma^2_*}
\]

Therefore,

\[
\frac{\text{Var}(E(Y|X_i))}{\text{Var}(Y)} = \frac{1}{\text{Var}(Y)} \left( e^{\beta_0 + K^{(i)}_2} \right)^2 \left( e^{\sigma^2_*} - 1 \right) e^{2\mu_* + \sigma^2_*}
\]

Similarly, since

\[
E(Y|X_P) = E \left( E(Y|X_1, \ldots, X_n)|X_{i_1}, \ldots, X_{i_p} \right)
\]

\[
= E \left( e^{X^T \beta} | X_{i_1}, \ldots, X_{i_p} \right)
\]

\[
= e^{\beta_0 + X^T_P \beta_P} E \left( e^{X^T_Q \beta_Q} | X_P \right)
\]

\[
= e^{\beta_0 + X^T_P \beta_P} \times e^{E(X^T_Q \beta_Q | X_P) + \frac{1}{2} \text{Var}(X^T_Q \beta_Q | X_P)}
\]

\[
E \left( X^T_Q \beta_Q | X_P \right) = E \left( X_Q | X_P \right)^T \beta_Q
\]

\[
= \left( \mu_Q + \Sigma_{QP} \Sigma_{PP}^{-1} (X_P - \mu_P) \right)^T \beta_Q
\]

\[
= X^T_P \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q + (\mu_Q - \Sigma_{QP} \Sigma_{PP}^{-1} \mu_P)^T \beta_Q
\]

\[
\text{Var} \left( X^T_Q \beta_Q | X_P \right) = \beta^T_Q \text{Var} \left( X_Q | X_P \right) \beta_Q
\]

\[
= \beta^T_Q \left( \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{PQ} \right) \beta_Q
\]
we have
\[ E(Y|X_P) = e^{\beta_0 + K_2^{(P)}} \times e^{X_P^T(\beta_P + \Sigma_{PP}^{-1}\Sigma_{PQ}\beta_Q)} \]

where
\[ K_2^{(P)} = (\mu_Q - \Sigma_{QP}\Sigma_{PP}^{-1}\mu_P)^T \beta_Q + \frac{1}{2} \beta_Q^T (\Sigma_{QQ} - \Sigma_{QP}\Sigma_{PP}^{-1}\Sigma_{PQ}) \beta_Q \]

Therefore,
\[
Var(E(Y|X_P)) = e^{2\beta_0 + 2K_2^{(P)}} \times Var(e^{X_P^T\eta}) \\
= e^{2\beta_0 + 2K_2^{(P)}} \times (e^{\sigma^*_*} - 1) e^{2\mu_** + \sigma^*_*}
\]

where
\[
\mu_** = E(X_P^T\eta) = \mu_P^T \eta \\
\sigma_** = Var(X_P^T\eta) = \eta^T \Sigma_{PP} \eta \\
\eta = \beta_P + \Sigma_{PP}^{-1}\Sigma_{PQ}\beta_Q
\]

Therefore,
\[
\frac{Var(E(Y|X_P))}{Var(Y)} = \frac{1}{Var(Y)} = \frac{1}{Var(Y)} \left( e^{\beta_0 + K_2^{(P)}} \right)^2 (e^{\sigma^*_*} - 1) e^{2\mu_** + \sigma^*_*}
\]

\[ \square \]

Result C.0.5. Sobol Indices under Linear GLMs with Logit Link. If \( \ln \left( \frac{E[Y|X]}{1-E[Y|X]} \right) = X^T\beta \) and the joint distribution of all input variables can be reasonably modelled by a multivariate normal distribution \( N(\mu, \Sigma) \) where \( \mu = (\mu_1, \mu_2, \cdots, \mu_n)^T \), \( \Sigma_{ii} = \sigma_i^2, \Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j \), the main-effect Sobol index with respect to single input has
the following form:

\[
\frac{Var(E(Y|X_i))}{Var(Y)} = \frac{1}{Var(Y)} \! \left\{ e^{-\frac{\bar{\mu}^2}{2\bar{\sigma}^2}} \left[ (-1)^{s-1} \frac{1}{2} + \sum_{k=1}^{s-1} (-1)^{k-1} \epsilon_{s-k} \bar{\sigma}^2 \right] \right\},
\]

(C.5)

where

\[
Z \sim \ln N(0, \bar{\sigma}^2)
\]

\[
\bar{\mu} = E(X^T\beta | X_i) = \left( \beta_i + \sum_{j \neq i} \beta_j \rho_{ij} \frac{\sigma_j}{\sigma_i} \right) X_i + \sum_{j \neq i} \beta_j \left( \mu_j - \mu_i \rho_{ij} \frac{\sigma_j}{\sigma_i} \right)
\]

\[
\bar{\sigma}^2 = Var(X^T\beta | X_i) = \beta^T \left( \Sigma_{PP} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{QP} \right) \beta
\]

\[
\beta_{-i} = (\beta_1, \beta_2, \cdots, \beta_{i-1}, \beta_{i+1}, \cdots, \beta_n)^T, \quad \Sigma_{P|Q} = \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{QP}; \quad \left[ \begin{array}{cc} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{array} \right]
\]

is the partition of \( \Sigma \) corresponding to input vector partition \( X = (X_P = X_i, X_Q = X_{-i})^T \).

Let \( X_P = (X_{i_1}, \cdots, X_{i_p})^T \), and \( X_Q \) be the input vector containing the remaining X’s. Then the main-effect Sobol index with respect to input subset \( X_P \) has the same form as expression (C.5) when \( \bar{\mu} \) and \( \bar{\sigma}^2 \) are replaced by the following \( \tilde{\mu} \) and \( \tilde{\sigma}^2 \):

\[
\tilde{\mu} = E(X^T\beta | X_P) = \beta_0 + X_P (\beta_P + \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q) + (\mu_Q - \Sigma_{QP} \Sigma_{PP}^{-1} \mu_P)^T \beta_Q
\]

\[
\tilde{\sigma}^2 = Var(X^T\beta | X_P) = \beta_Q^T \left( \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{QP} \right) \beta_Q
\]
Proof.

\[ E(Y|X_i) = E\left(\frac{e^{X^T\beta}}{1+e^{X^T\beta}}\big| X_i\right) = E\left(\frac{e^{\tilde{X}}}{1+e^{\tilde{X}}}\right) \]

where

\[ \tilde{X} \sim N(\tilde{\mu}, \tilde{\sigma}^2) \]

\[ \tilde{\mu} = E\left(X^T\beta|X_i\right) \]
\[ = \beta_iX_i + \sum_{j\neq i}^{n} \beta_jE(X_j|X_i) \]
\[ = \beta_iX_i + \sum_{j\neq i}^{n} \beta_j\left(\mu_j + \rho_{ij}\frac{\sigma_j}{\sigma_i}(X_i - \mu_i)\right) \]
\[ = \left(\beta_i + \sum_{j\neq i}^{n} \beta_j \rho_{ij}\frac{\sigma_j}{\sigma_i}\right)X_i + \sum_{j\neq i}^{n} \beta_j\left(\mu_j - \mu_i \rho_{ij}\frac{\sigma_j}{\sigma_i}\right) \]

\[ \tilde{\sigma}^2 = Var\left(X^T\beta|X_i\right) = Var\left(\sum_{j\neq i}^{n} \beta_jX_j|X_i\right) = \beta_{-i}^T\Sigma_{Q|P}\beta_{-i} \]

\[ \beta_{-i} = (\beta_1, \beta_2, \cdots, \beta_{i-1}, \beta_{i+1}, \cdots, \beta_n)^T \]

\[ \Sigma_{Q|P} = \Sigma_{PP} - \Sigma_{PQ} \Sigma_{QQ}^{-1}\Sigma_{QP} \]

\[ \left[\begin{array}{cc}
\Sigma_{PP} & \Sigma_{PQ} \\
\Sigma_{QP} & \Sigma_{QQ}\end{array}\right] \]

is a partition of \( \Sigma \) corresponding to the input vector partition \( X = (X_P = X_i, X_Q^T = X_{T-i})^T \).

Then by applying bullet 1 of the Result B.0.2, we have the final expression of the main-effect index with respect to \( X_i \):
\[
\frac{\text{Var}(E(Y|X_i))}{\text{Var}(Y)} = \frac{1}{\text{Var}(Y)} \text{Var} \left\{ e^{-\frac{\hat{\beta}^2}{2\hat{\sigma}^2}} \left[ (-1)^{s-1} \frac{1}{2} + \sum_{k=1}^{s-1} (-1)^{k-1} e^{\frac{1}{2}(s-k)\hat{\sigma}^2} \right] \right\},
\]

\[
s = \begin{cases} 
1 + \frac{\hat{\mu}}{\hat{\sigma}^2}, & \text{if } \frac{\hat{\mu}}{\hat{\sigma}^2} \in \mathbb{Z}^+ \\
-\frac{\hat{\mu}}{\hat{\sigma}^2}, & \text{if } \frac{\hat{\mu}}{\hat{\sigma}^2} \in \mathbb{Z}^-
\end{cases}
\]

\[
\frac{1}{\text{Var}(Y)} \text{Var} \left\{ e^{-\frac{\hat{\beta}^2}{2\hat{\sigma}^2}} \left[ (-1)^{s-1} \frac{1}{2} E \left( \frac{Z^{s-|s|}}{1+Z} \right) + \sum_{k=1}^{|s|} (-1)^{k-1} e^{\frac{1}{2}(s-k)\hat{\sigma}^2} \right] \right\},
\]

\[
s = \begin{cases} 
1 + \frac{\hat{\mu}}{\hat{\sigma}^2}, & \text{if } \frac{\hat{\mu}}{\hat{\sigma}^2} \in \mathbb{R}^+ \\
-\frac{\hat{\mu}}{\hat{\sigma}^2}, & \text{if } \frac{\hat{\mu}}{\hat{\sigma}^2} < -1
\end{cases}
\]

where

\[Z \sim \ln N(0, \hat{\sigma}^2)\]

\[\hat{\mu} = E(X^T \beta | X_i) = \left( \beta_i + \sum_{j \neq i}^n \beta_j \rho_{ij} \frac{\sigma_j}{\sigma_i} \right) X_i + \sum_{j \neq i}^n \beta_j \left( \mu_j - \mu_i \rho_{ij} \frac{\sigma_j}{\sigma_i} \right)\]

\[\hat{\sigma}^2 = \text{Var} (X^T \beta | X_i) = \beta^T (\Sigma_{P|Q} \beta_{-i})\]

\[\beta_{-i} = (\beta_1, \beta_2, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n)^T, \Sigma_{P|Q} = \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{PQ}; \begin{bmatrix} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{bmatrix}\]

is the partition of \( \Sigma \) corresponding to input vector partition \( X = (X_P = X_i, X_Q^T = X^T_{-i})^T \).

The proof for main-effect index with respect to \( X_P \) goes similarly. \(\Box\)
Appendix D: Proofs of Sobol Index Estimation under Polynomial GLMs

Result D.0.6. Sobol Indices under Polynomial GLMs with Identity Link and Independent Inputs. Suppose $X = (X_1, \cdots, X_n)$ are independent random variables and the conditional expectation of response $Y$ with respect to all inputs is a multivariate polynomial function of $X$ with degree $K \in \mathbb{Z}^+$,

$$E[Y|X] = \text{Poly}^{(K)}(X, \beta) = \sum_{|k|_1 \leq K} \beta_k X^k$$

where $k = (k_1, k_2, \cdots, k_n) \in \mathbb{Z}^n$. Then we can show that $\forall \ X_P = (X_{i_1}, X_{i_2}, \cdots, X_{i_p})$, $1 \leq p \leq n$

$$E[Y|X_P] = \text{Poly}^{(K')} (X_P, \beta'), \ 1 \leq K' \leq K$$

which means the estimation of exact Sobol index with respect to any input subset $X_P$ only requires fitting $Y$ as a polynomial function of $X_P$.

Proof. For any $X_P$ of interest, we can rearrange $X = (X_P, X_{-P})$ to make $X_P = (X_1, X_2, \cdots, X_p)$, where $X_{-P}$ is the complement set of $X_P$.

Since $X_1, \cdots, X_n$ are independent,

$$E \left[ X_1^{k_1} X_2^{k_2} \cdots X_n^{k_n} | X_P \right] = \prod_{i=1}^p X_i^{k_i} \prod_{i=p+1}^n E \left[ x_i^{k_i} \right], \ \forall \ 0 \leq k_1 + k_2 + \cdots + k_n \leq K$$
Therefore,
\[ E \left[ \text{Poly}^{(K)}(\mathbf{X}) \bigg| \mathbf{X}_P \right] = \text{Poly}^{(K')} (\mathbf{X}_P), \quad 1 \leq K' \leq K \]

\[ \square \]

Result D.0.7. **Sobol Indices under Polynomial GLMs with Identity Link and Multivariate Normal Inputs.** Suppose \( \mathbf{X} = (X_1, \cdots, X_n) \) follows a Multivariate Normal distribution \( \text{MN}(\mathbf{\mu}, \Sigma) \) and the conditional expectation of response \( Y \) with respect to all inputs is a multivariate polynomial function of \( \mathbf{X} \) with degree \( K \in \mathbb{Z}^+ \),
\[ E[\mathbf{Y} \mid \mathbf{X}] = \text{Poly}^{(K)}(\mathbf{X}, \mathbf{\beta}) = \sum_{|k| \leq K} \beta_k \mathbf{X}^k \]
where \( k = (k_1, k_2, \cdots, k_n) \in \mathbb{Z}^n \). Then we can show that \( \forall \ X_P = (X_{i_1}, X_{i_2}, \cdots, X_{i_p}), \)
\[ 1 \leq p \leq n \]
\[ E[\mathbf{Y} \mid \mathbf{X}_P] = \text{Poly}^{(K')} (\mathbf{X}_P, \mathbf{\beta}'), \quad 1 \leq K' \leq K \]
which means the estimation of exact Sobol index with respect to any input subset \( \mathbf{X}_P \) only requires fitting \( \mathbf{Y} \) as a polynomial function of \( \mathbf{X}_P \).

\( \textbf{Proof.} \) Let \( \Sigma = \mathbf{LDL}^T \) where \( \mathbf{D} = \text{diag}(d_1, \cdots, d_n) \) and \( \mathbf{L} \) is a lower unit triangular. Then we have:
\[ \mathbf{W} = \mathbf{L}^{-1} \mathbf{X} \sim \text{MN}(\mathbf{L}^{-1}\mathbf{\mu}, \mathbf{D}) \]
Since every unit lower triangular matrix is nonsingular and its inverse is also a unit lower triangular matrix, we know \( W_1 = X_1 \). Thus, for singleton \( X_P = \{X_1\} \),

\[
E(Y|X_P) = E\left[E(Y|X_1, \ldots, X_n)|X_1\right]
= E\left[\text{Poly}^{(K)}(X, \beta)|X_1\right]
= E\left[\text{Poly}^{(K)}(\mathbf{LW}, \beta)|W_1\right]
= E\left[\text{Poly}^{(K)}(\mathbf{W}, \beta^*)|W_1\right]
\]

Since \( \{W_i, \ i = 1, \ldots, n\} \) are independent,

\[
E\left(W_1^{k_1}W_2^{k_2}W_3^{k_3} \cdots W_n^{k_n}|W_1\right) = W_1^{k_1}E\left[W_2^{k_2}\right]E\left[W_3^{k_3}\right] \cdots E\left[W_n^{k_n}\right]
\]

\[\forall \ 0 \leq k_1 + k_2 + \cdots + k_n \leq K\]

Therefore,

\[
E(Y|X_1) = E\left[\text{Poly}^{(K)}(\mathbf{W}, \beta)|W_1\right] = \text{Poly}^{(K')} (W_1, \beta') = \text{Poly}^{(k')}(X_1, \beta'), \ 1 \leq k' \leq k
\]

Since any \( X_i \) can be chosen as the \( X_1 \), we already proved the result for any singleton \( X_P = \{X_1\}, \ \forall 1 \leq i \leq n \).

For any \( X_P \) containing more than one input variable, we can choose these inputs as the first \( p \) variables in \( X \). Thus,

\[
E(Y|X_P) = E\left[E(Y|X_1, \ldots, X_n)|X_1, \ldots, X_p\right]
= E\left[\text{Poly}^{(K)}(\mathbf{X}, \beta)|X_P\right]
= E\left[\text{Poly}^{(K)}(\mathbf{LW}, \beta)|(\mathbf{LW})_P\right]
= E\left[\text{Poly}^{(K)}(\mathbf{W}, \beta)|\mathbf{W}_P\right]
\]

where \( \mathbf{W}_P = (W_1, \cdots, W_p) \). The reason why

\[
E\left[\text{Poly}^{(K)}(\mathbf{LW}, \beta)|(\mathbf{LW})_P\right] = E\left[\text{Poly}^{(K)}(\mathbf{LW}, \beta)|\mathbf{W}_P\right]
\]
is because

$$E \left[ \text{Poly}^{(K)}(\mathbf{LW}) \mid (\mathbf{LW})_P \right] = \int \text{Poly}^{(K)}(\mathbf{LW}) \ f_{\mathbf{LW} | (\mathbf{LW})_P} \ d (\mathbf{LW})_P$$

$$= \int \text{Poly}^{(K)}(\mathbf{LW}) \ f_{\mathbf{LW} | (\mathbf{LW})_P} |L^{-1}_P|^{-1} \ d \mathbf{W}_P$$

Since $|L^{-1}_P| = 1$ and

$$f_{\mathbf{LW} | (\mathbf{LW})_P} = \frac{f_{\mathbf{LW}}}{\int f_{\mathbf{LW}} \ d (\mathbf{LW})_P}$$

$$= \frac{f_{\mathbf{LW}}}{\int f_{\mathbf{LW}} |L^{-1}_P|^{-1} \ d \mathbf{W}_P}$$

$$= \frac{f_{\mathbf{LW}}}{\int f_{\mathbf{LW}} \ d \mathbf{W}_P}$$

$$= f_{\mathbf{LW} | \mathbf{W}_P}$$

we have:

$$E \left[ \text{Poly}^{(K)}(\mathbf{LW}) \mid (\mathbf{LW})_P \right] = \int \text{Poly}^{(K)}(\mathbf{LW}) \ f_{\mathbf{LW} | (\mathbf{LW})_P} \ d \mathbf{W}_P$$

$$= \int \text{Poly}^{(K)}(\mathbf{LW}) \ f_{\mathbf{LW} | \mathbf{W}_P} \ d \mathbf{W}_P$$

$$= E \left[ \text{Poly}^{(K)}(\mathbf{LW}, \beta) \mid \mathbf{W}_P \right]$$

where $f_{\mathbf{LW}}$ is the joint probability density function of $\mathbf{LW}$.

Since $\{W_i, \ i = 1, \cdots, n\}$ are independent,

$$E(Y | \mathbf{X}_P) = E \left[ \text{Poly}^{(K)}(\mathbf{LW}, \beta) \bigg\mid \mathbf{W}_P \right]$$

$$= \text{Poly}^{(K')} (\mathbf{W}_P, \beta^*)$$

$$= \text{Poly}^{(K')} (\mathbf{X}_P, \beta'), \ 1 \leq K' \leq K$$

Once we know the coefficients in $\text{Poly}^{(K')} (\mathbf{X}_P)$, the main-effect Sobol index with respect to input subset $\mathbf{X}_P$ can be estimated by the sample variance of $\text{Poly}^{(K')} (\mathbf{X}_P)$ divided by the sample variance of $Y$. 

\[\square\]
Appendix E: Gaussian Model Simulation with Less Dependent Inputs

Table E.1: Quantiles of Relative Difference between SI Estimates and the Corresponding Exact Estimates under Gaussian Model (\(\rho = 0.3\))

<table>
<thead>
<tr>
<th>RD-Quantiles</th>
<th>10%</th>
<th>30%</th>
<th>50%</th>
<th>70%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI-UM</td>
<td>5.7\times10^{-16}</td>
<td>1.7\times10^{-15}</td>
<td>3.2\times10^{-15}</td>
<td>5.5\times10^{-15}</td>
<td>1.5\times10^{-14}</td>
</tr>
<tr>
<td>SI-CMM</td>
<td>2.0\times10^{-16}</td>
<td>7.2\times10^{-16}</td>
<td>1.4\times10^{-15}</td>
<td>2.8\times10^{-15}</td>
<td>9.2\times10^{-15}</td>
</tr>
</tbody>
</table>

NOTE: "SI-UM" stands for Sobol index estimates obtained by fitting univariate models. "SI-CMM" stands for Sobol index estimates obtained by fitting contaminated multivariate model. The accuracy of "SI-UM" is quantified by the following relative difference formula: \(\text{abs}("\text{SI-UM}" - "\text{SI-EX}")/ "\text{SI-EX}"\), where "SI-EX" stands for the exact Sobol index estimates obtained by fitting the correct multivariate model. The quantile estimates are obtained based on 1000 simulations (each with sample size 1000) under the Gaussian model with input correlation 0.3. "RD-Quantiles" stands for quantile estimates of the relative differences.
Figure E.1: Variable Selection Methods Comparison (inputs correlation $\rho = 0.3$)
Figure E.2: Sobol Index Significance Test versus Other Methods (inputs correlation $\rho = 0.3$)
Appendix F: Poisson Model Simulation with Less Dependent Inputs

Table F.1: Quantiles of Relative Difference between SI Estimates and the Corresponding Correct Estimates under Poisson Model with Identity Link ($\rho = 0.3$)

<table>
<thead>
<tr>
<th>RD-Quantiles</th>
<th>10%</th>
<th>30%</th>
<th>50%</th>
<th>70%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI-MML</td>
<td>6.5×10^{-3}</td>
<td>2.3×10^{-2}</td>
<td>4.9×10^{-2}</td>
<td>1.0×10^{-1}</td>
<td>4.8×10^{-1}</td>
</tr>
<tr>
<td>SI-CMML</td>
<td>4.7×10^{-3}</td>
<td>1.8×10^{-3}</td>
<td>3.8×10^{-2}</td>
<td>8.6×10^{-2}</td>
<td>4.4×10^{-1}</td>
</tr>
</tbody>
</table>

NOTE: "SI-MML" stands for Sobol index estimates obtained by fitting the multivariate models with all true inputs and the log link. "SI-CMML" stands for Sobol index estimates obtained by fitting contaminated multivariate model with log link. The accuracy of "SI-MML" is quantified by the following relative difference formula: \text{abs("SI-MML" - "SI-UM")/ "SI-UM"}, where "SI-UM" stands for the correct Sobol index estimates obtained by fitting the univariate model. The quantile estimates are obtained based on 1000 simulations (each with sample size 1000) from the Poisson model with identity link and input correlation 0.3. "RD-Quantiles" stands for quantile estimates of the relative differences.
Figure F.1: Sobol Index Significance Test under Linear Poisson Model with Log Link and Inputs

Correlation $\rho = 0.3$
Figure F.2: Sobol Index Significance Test under Linear Poisson Model with Log Link and Inputs
Correlation $\rho = 0.3$
Table F.2: Quantiles of Relative Difference between SI Estimates and the Corresponding Exact Estimates under Poisson Model with Log Link ($\rho = 0.3$)

<table>
<thead>
<tr>
<th>RD-Quantiles</th>
<th>10%</th>
<th>30%</th>
<th>50%</th>
<th>70%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI-UM</td>
<td>0.23</td>
<td>0.61</td>
<td>0.85</td>
<td>0.99</td>
<td>13.74</td>
</tr>
<tr>
<td>SI-CMM</td>
<td>0.16</td>
<td>0.44</td>
<td>0.67</td>
<td>0.90</td>
<td>3.81</td>
</tr>
</tbody>
</table>

NOTE: "SI-UM" stands for Sobol index estimates obtained by fitting univariate models. "SI-CMM" stands for Sobol index estimates obtained by fitting contaminated multivariate model. The accuracy of "SI-UM" is quantified by the following relative difference formula: $\text{abs}("SI-UM" - "SI-EX")/ "SI-EX"$, where "SI-EX" stands for the exact Sobol index estimates obtained by fitting the correct multivariate model. The quantile estimates are obtained based on 1000 simulations (each with sample size 1000) under the Poisson model with log link and input correlation 0.3. "RD-Quantiles" stands for quantile estimates of the relative differences.