Published in final edited form as:

Commun Stat Theory Methods. 2018 ; 47(21): 5163–5195. doi:10.1080/03610926.2017.1388397.

Estimation of Sobol's Sensitivity Indices under Generalized Linear Models

Rong Lu¹, Danxin Wang², Min Wang³, and Grzegorz A. Rempala^{3,4,*}

¹Bioinformatics Core Facility, Department of Clinical Sciences, University of Texas, Southwestern Medical Center, 5323 Harry Hines Blvd. Dallas, TX 75390

²Center for Pharmacogenomics, College of Medicine, The Ohio State University, 333 W. 10th Avenue, Columbus, OH 43210

³Mathematical Bioscience Institute, The Ohio State University, 1735 Neil Ave., Columbus, OH 43210

⁴Biostatistics Division, College of Public Health, The Ohio State University, 1841 Neil Ave., Columbus, OH, 43210

Abstract

We derive explicit formulas for Sobol's sensitivity indices (SSIs) under the generalized linear models (GLMs) with independent or multivariate normal inputs. We argue that the main-effect SSIs provide a powerful tool for variable selection under GLMs with identity links under polynomial regressions. We also show via examples that the SSI-based variable selection results are similar to the ones obtained by the random forest algorithm but without the computational burden of data permutation. Finally, applying our results to the problem of gene network discovery, we identify through the SSI analysis of a public microarray dataset several novel higher-order gene-gene interactions missed out by the more standard inference methods. The relevant functions for SSI analysis derived here under GLMs with identity, log, and logit links are implemented and made available in the R package *SobolSensitivity*.

Keywords

Sobol's indices; global sensitivity analysis; variable ranking; variable selection; generalized linear models; correlated inputs; gene-gene interactions

1 Background

Sensitivity analysis elucidates how different factors/inputs contribute to the uncertainty in the output variables of various input-output models and simulators. In practice, this type of analysis is often performed to gain essential insights into model behavior. Its applications include: 1. Ranking importance of the input variables (Factor Prioritization); 2. Identifying negligible inputs (Factor Fixing or Screening); 3. Determining input variation space that is

*The corresponding author. rempala.3@osu.edu.
Rong.Lu@UTSouthwestern.edu, wang.808@osu.edu, wang.1807@mbi.osu.edu

responsible for certain type of output values of interest (Mapping) [31, 43, 44]. Various forms of indices have been proposed for quantifying sensitivity of different aspects of the model outputs. Some of them are defined using correlation, entropy, regression coefficients, and local derivatives, while others are based on conditional moments (such as the variances of conditional expectation) or the conditional distribution functions. For each specific application, both the forms of the sensitivity indices and the corresponding estimation strategy need to be chosen carefully, keeping in mind not only the analysis goal, the model architecture and computing time, but also the dependance structure among the inputs. Borgonovo and Plischke (2016) [2] provide a methodological review of recent advances in sensitivity analysis for operational research. Pianosi *et al.* [31] give a systematic but less mathematically detailed review of sensitivity analysis of environmental models. Norton [29] discusses sensitivity assessment in models utilizing dynamic simulations. For the general overview of sensitivity analysis approaches using emulation modeling, the interested reader is referred to Ratto *et al.* [33].

Sobol's sensitivity indices, first introduced in 1990 [45], provide a unified way of quantifying output's sensitivity with respect to any subset of the input variables, based on the variance decomposition the response variable. They have become prevalent tools not only for sensitivity analysis [2, 19, 40] but also for other application purposes such as quality assessment of composite indicators [30, 36], variable selection in regression [13, 18, 37], and the multiple criteria analysis [1]. Methods for estimating Sobol's indices have been studied extensively under the assumption that the input factors are independent of each other. This line of work includes Sobol [45, 46, 47], Saltelli [38], Tarantola *et al.* [52], Liburne and Tarantola [25], Saltelli *et al.* [39], Xue *et al.* [55], etc. In order to perform Sobol's sensitivity analysis on systems with correlated inputs, a lot of effort was devoted to approximate the underlying true model in the ways that ensure certain type of orthogonality among their decomposition components. Relevant ideas include decorrelating the inputs with the Gram-Schmidt procedure before approximating the model [27], decomposing the output variance into partial correlated and uncorrelated components [24, 54] using copula theory to model the correlation structure in polynomial chaos expansion [5], using revised Hoeffding decomposition with projection operators to construct hierarchically orthogonal component functions [6], and applying Gram Schmidt procedure recursively to construct hierarchical orthogonal basis [7].

However, all the above ideas for handling correlated inputs only work well when the sample size is much larger than the number of inputs, simply because reconstructing the full input-to-output map is not feasible if the sample size is small, especially when higher orders of input interactions are of main interest. While large datasets are easy to obtain in certain research fields, it's generally not the case in genomic studies. For example, our research group is interested in studying gene-gene interactions that affect enzyme CYP3A4. We want to analyze some existing data to come up with a few promising gene triplet or quadruplet candidates before performing series of expensive and time-consuming lab experiments on them. Since CYP3A4 only expresses in liver, the largest dataset we can find is a microarray data with 427 samples. After extensive literature review, the input list is narrowed down to 46 gene, corresponding to 78 measurements in this dataset. 78 measurements give us 76,076 different triplets and 1,426,425 quadruplets, which makes it impossible to obtain a good

approximation of the full map using only 427 samples. In order to rank these huge number of gene triplets and quadruplets using only 427 samples, we are motivated to study the estimation strategy of Sobol's indices when the available information is not sufficient for approximating the full model containing all correlated inputs. Intuitively, we will have less power to estimate the higher order interactions when sample size is small and therefore we must be willing to make certain assumptions about the underlying data structure. If the ultimate goal is only to estimate Sobol's indices instead of building a predictive model, it may not be necessary to reconstruct the full model containing all input factors. After all, the evaluation of each specific Sobol's index only concerns the conditional expectation with respect to a particular subset of inputs. When the number of input factors is large, we think it's more convenient and computationally more efficient to approximate each partial conditional expectation function directly and independently, while parallelizing the computation. We propose using GLMs as the meta-model to approximate these subspace projections because, as we show below, that evaluation of Sobol's indices is very efficient under GLMs with some not too-restrictive assumptions. To the best of our knowledge the estimation strategy of Sobol's indices has never been considered within the generalized linear model (GLM) framework, despite the fact that GLM is one of the most commonly used statistical models in medical and clinical researches nowadays.

In this paper, we discuss the estimation strategies of Sobol's indices under GLMs with the inputs which are either independent or follow a multivariate normal distribution. For linear GLMs where the systematic component is a linear function of the inputs the closed formulas of Sobol's indices are derived in Section 2 under, respectively, the identity link, the log link, and the logit link functions. We also show that under multivariate polynomial GLMs, if the inverse link function is bounded, continuous and real-valued, each Sobol's index can be approximated independently using empirical variance of a polynomial function containing only the subset of inputs specified by the index. In addition, if the inputs can be further assumed to be either independent or follow a multivariate normal distribution, we show that the proposed estimation strategy works for polynomial GLMs with commonly used link functions, such as the identity link, log link, logit link, complementary log-log link etc. Simulation studies are performed in Section 3 to assess the performance of these Sobol's index estimates in variable selection, not only in terms of their accuracy, but also in terms of type I and II errors and the false discovery rate. We also present an example of Sobol's index application to mapping gene-gene interactions under the multivariate polynomial Gaussian model. The discussion and some conclusions are presented in Sections 4 and 5 with most the technical details of the derivations in Section 3 provided in the Appendix (available on the journal website).

2 Derivations for Sobol's Indices under GLMs

2.1 Generalized Linear Models

GLMs are generalizations of the multivariate linear regression [28]. 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:

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

where $\mathbf{X} = (X_1, \dots, X_n)^T$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^T$. 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. $\mathbf{X}^T \boldsymbol{\beta}$ 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 modeled by specifying different distributions within the exponential family, such as Poisson, Binomial, Gamma, Exponential, Inverse Gaussian, Negative Binomial, etc. Due to their flexibility, GLMs have become one of the most popular modeling techniques in the modern frequentist statistical approach [10, 12, 16].

2.2 Definition of Sobol's Indices

In the literature, there are two ways of defining Sobol's sensitivity indices. One is based on the integrals of HDMR component functions [7, 21, 47]. The other (and equivalent one) is based on variances of conditional expectations of the response given input subsets [38, 44]. Since the variance-based definition fits well into the GLM framework, we use it throughout the paper tacitly assuming that all required second moments exist.

Definition 2.1. Suppose Y is a univariate random variable, and its mean is determined by a set of random variables $\mathbf{X} = (X_1, \dots, X_n)^T$. Then the main-effect Sobol's index of Y with respect to X_i is:

$$S_i = \frac{\text{Var}(E(Y|X_i))}{\text{Var}(Y)}$$

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

$$S_i^T = \frac{E(\text{Var}(Y|X_{-i}))}{\text{Var}(Y)} = 1 - \frac{\text{Var}(E(Y|X_{-i}))}{\text{Var}(Y)}$$

where $\mathbf{X}_{-i}$ denotes a vector of inputs excluding X_i . Similarly, the main-effect Sobol's index of Y with respect to input subset $\mathbf{X}_P = \{X_{i_1}, X_{i_2}, \dots, X_{i_p}\}$ is:

$$S_P = \frac{\text{Var}(E(Y|X_{i_1}, X_{i_2}, \dots, X_{i_p}))}{\text{Var}(Y)}$$

and the corresponding total-effect index is

$$S_P^T = \frac{E(\text{Var}(Y|X_{-P}))}{\text{Var}(Y)} = 1 - \frac{\text{Var}(E(Y|X_{-P}))}{\text{Var}(Y)}$$

where $\mathbf{X}_{-p}$ is the subset of inputs excluding $\mathbf{X}_p$

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 [21, 38, 47]. Since both the total-effect indices and the interaction-effect indices can be viewed as functions of the main-effect indices, the key to Sobol's index estimation is to estimate the main-effect indices, which comes down to evaluating the variances of conditional expectations with respect to the different input subsets. If the conditional expectations can be written out in explicit expressions of the inputs being conditioned on, we can then estimate Sobol's indices using the empirical estimates of these expressions.

2.3 Formulas of Sobol's Indices Under Linear GLMs

In the following three results, closed formulas of the main-effect Sobol's 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. The proofs of these results are provided in Appendix ?? (available on the journal website).

Proposition 2.1. Sobol's indices under Linear GLMs with Identity Link—

Suppose $E[Y|\mathbf{X}] = \beta_0 + \mathbf{X}^T \boldsymbol{\beta}$. If the inputs $\mathbf{X}$ follow a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T$, $\Sigma_{ii} = \sigma_i^2$, $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, then the main-effect Sobol's index with respect to X_i has the following closed form:

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

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

$$\frac{\text{Var}(E(Y|\mathbf{X}_p))}{\text{Var}(Y)} = \frac{\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{pp} \boldsymbol{\eta}}{\text{Var}(Y)}$$

where

$$\boldsymbol{\eta} = \boldsymbol{\beta}_p + \boldsymbol{\Sigma}_{pp}^{-1} \boldsymbol{\Sigma}_{pQ} \boldsymbol{\beta}_Q$$

and $\begin{bmatrix} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{bmatrix}$ is the partition of Σ corresponding to input vector partition $\mathbf{X} = (\mathbf{X}_P^T, \mathbf{X}_Q^T)^T$. If the inputs are independent random variables from any distribution with finite second moment, the main-effect Sobol's indices with respect to a single input simplifies to

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

and the main-effect Sobol's index with respect to input subset $\mathbf{X}_P$ simplifies to

$$\frac{\text{Var}(E(Y|\mathbf{X}_P))}{\text{Var}(Y)} = \frac{\sum_{j \in P} \beta_j^2 \cdot \text{Var}(X_j)}{\text{Var}(Y)}.$$

Proposition 2.2. Sobol's indices under Linear GLMs with Log Link—Suppose $\ln(E[Y|\mathbf{X}]) = \beta_0 + \mathbf{X}^T \beta$. If the inputs $\mathbf{X}$ follow a multivariate normal distribution $N(\mu, \Sigma)$ where $\mu = (\mu_1, \mu_2, \dots, \mu_n)^T$, $\Sigma_{ii} = \sigma_i^2$, $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, the main-effect Sobol's index with respect to X_i has the following closed form:

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

where

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

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

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

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

$$\frac{\text{Var}(E(Y|\mathbf{X}_P))}{\text{Var}(Y)} = \frac{1}{\text{Var}(Y)} \left(e^{\sigma_{**}^2} - 1 \right) e^{2\beta_0 + 2W^{(P)} + 2\mu_{**} + \sigma_{**}^2}$$

where

$$\mu_{**} = \boldsymbol{\mu}_P^T (\boldsymbol{\beta}_P + \boldsymbol{\Sigma}_{PP}^{-1} \boldsymbol{\Sigma}_{PQ} \boldsymbol{\beta}_Q)$$

$$\sigma_{**}^2 = (\boldsymbol{\beta}_P + \boldsymbol{\Sigma}_{PP}^{-1} \boldsymbol{\Sigma}_{PQ} \boldsymbol{\beta}_Q)^T \boldsymbol{\Sigma}_{PP} (\boldsymbol{\beta}_P + \boldsymbol{\Sigma}_{PP}^{-1} \boldsymbol{\Sigma}_{PQ} \boldsymbol{\beta}_Q)$$

$$W^{(P)} = (\boldsymbol{\mu}_Q - \boldsymbol{\Sigma}_{QP} \boldsymbol{\Sigma}_{PP}^{-1} \boldsymbol{\mu}_P)^T \boldsymbol{\beta}_Q + \frac{1}{2} \boldsymbol{\beta}_Q^T (\boldsymbol{\Sigma}_{QQ} - \boldsymbol{\Sigma}_{QP} \boldsymbol{\Sigma}_{PP}^{-1} \boldsymbol{\Sigma}_{PQ}) \boldsymbol{\beta}_Q$$

and $\begin{bmatrix} \boldsymbol{\Sigma}_{PP} & \boldsymbol{\Sigma}_{PQ} \\ \boldsymbol{\Sigma}_{QP} & \boldsymbol{\Sigma}_{QQ} \end{bmatrix}$ is the partition of $\boldsymbol{\Sigma}$ corresponding to the input vector partition $\mathbf{X} = (\mathbf{X}_P^T, \mathbf{X}_Q^T)^T$.

If the inputs are independent random variables from any distribution with finite second moment, the main-effect Sobol's indices with respect to a single input can be estimated using the following expression

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

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

$$\frac{\text{Var}(E(Y|\mathbf{X}_P))}{\text{Var}(Y)} = \frac{1}{\text{Var}(Y)} \left[\exp(\beta_0) \prod_{j \notin P} E[\exp(\beta_j X_j)] \right]^2 \text{Var} \left[\exp \left(\sum_{i \in P} \beta_i X_i \right) \right]$$

Proposition 2.3. Sobol's indices under Linear GLMs with Logit Link—If $\ln$

$\left(\frac{E[Y|\mathbf{X}]}{1 - E[Y|\mathbf{X}]} \right) = \beta_0 + \mathbf{X}^T \boldsymbol{\beta}$ and the inputs follow a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T$, $\Sigma_{ii} = \sigma_i^2$, $\Sigma_{ij} = \rho_{ij} \sigma_i \sigma_j$, the main-effect Sobol's index with respect to X_i is:

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

$$s = \begin{cases} 1 + \frac{\tilde{\mu}}{\tilde{\sigma}^2}, & \text{if } \frac{\tilde{\mu}}{\tilde{\sigma}^2} \in \mathbf{Z}^+ \\ -\frac{\tilde{\mu}}{\tilde{\sigma}^2}, & \text{if } \frac{\tilde{\mu}}{\tilde{\sigma}^2} \in \mathbf{Z}^- \end{cases} \quad (2)$$

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

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

where

$$Z \sim \ln N(0, \tilde{\sigma}^2)$$

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

$$\tilde{\sigma}^2 = Var(\mathbf{X}^T \beta | X_i) = \beta_{-i}^T \Sigma_{P|Q} \beta_{-i}$$

$$\boldsymbol{\beta}_{-i} = (\beta_1, \beta_2, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_n)^T, \Sigma_{P|Q} = \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{PQ}; \begin{bmatrix} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{bmatrix} \text{ is the partition of } \Sigma$$

corresponding to input vector partition $\mathbf{X} = (\mathbf{X}_P^T = \mathbf{X}_i^T, \mathbf{X}_Q^T = \mathbf{X}_{-i}^T)^T$.

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

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

$$\tilde{\sigma}^2 = \text{Var}(X^T \beta | \mathbf{X}_P) = \beta_Q^T (\Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{PQ}) \beta_Q$$

Under the logit link GLM, 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's indices corresponding to integer valued $\frac{\tilde{\mu}}{\tilde{\sigma}^2}$ or $\frac{\tilde{\mu}}{\tilde{\sigma}^2}$, we only need to evaluate the sample variance of the conditional expectation expression provided above to get exact estimates of these indices. For Sobol's indices with non-integer valued $\frac{\tilde{\mu}}{\tilde{\sigma}^2}$ or $\frac{\tilde{\mu}}{\tilde{\sigma}^2}$, the efficiency of estimating Sobol's 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 containing 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's indices using the above formula can be fairly quick. The simplest way of generating such table is to calculate the empirical mean from simulated samples of $\frac{Z^s - \lfloor s \rfloor}{1 + Z}$, or to use numerical integration techniques. Meanwhile, since $s - \lfloor s \rfloor$ is bounded between 0 and 1, 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 $\tilde{\sigma}^2$ are actually constants, while $\tilde{\mu}$ and $\tilde{\mu}$ 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 Sobol's indices by evaluating the expression of conditional mean at each observed value of X_j or $\mathbf{X}_P$ and then calculate 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 a closed formula for $E(Y|X_j)$ as a function of X_j when the distributions of $\mathbf{X}_{-j}$ are unknown. Under the logit link and independence assumption,

$$\frac{\text{Var}(E(Y|X_i))}{\text{Var}(Y)} = \frac{E^2 \left[\exp \left(\sum_{j \neq i} \beta_j X_j \right) \right]}{\text{Var}(Y)} \text{Var} \left[E \left[\frac{\exp(\beta_i X_i)}{1 + \exp \left(\sum_{j \neq i} \beta_j X_j \right) \exp(\beta_i X_i)} \middle| X_i \right] \right]$$

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

2.4 Sobol's Indices under Polynomial GLMs

Having settled the case of GLMs with linear systematic components, we will now discuss the estimation of Sobol's indices under GLMs for which the systematic components are multivariate polynomial functions of input variables. Proofs of the following two results are presented in Appendix ?? available on the journal website.

Proposition 2.4. Sobol's indices under Polynomial GLMs with Identity Link—

Suppose the conditional expectation of response Y with respect to all inputs $\mathbf{X} = (X_1, \dots, X_n)$ is a multivariate polynomial function with degree $K \in \mathbb{Z}^+$,

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

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

$$\begin{aligned} \forall \mathbf{X}_p &= (X_{i_1}, X_{i_2}, \dots, X_{i_p}), \quad 1 \leq p \leq n \\ E[Y|\mathbf{X}_p] &= \text{Poly}^{(K')}(\mathbf{X}_p, \beta'), \quad 1 \leq K' \leq K \end{aligned} \quad (3)$$

which means that estimation of the exact Sobol's index with respect to any input subset $\mathbf{X}_p$ only requires fitting the smaller model (3) and then evaluating the empirical variance of $\text{Poly}^{(K')}(\mathbf{X}_p, \beta')$ and Y :

$$\frac{\text{Var}(E(Y|\mathbf{X}_p))}{\text{Var}(Y)} = \frac{\text{Var}[\text{Poly}^{(K')}(\mathbf{X}_p, \beta')]}{\text{Var}(Y)}$$

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's index S_1 , we only need to fit Y as a quadratic function of X_1 . The R^2 (coefficient of determination [9]) of the 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's 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's 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's index with respect to X_2 under the true model; the total-effect Sobol's 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 the main-effect Sobol's index with respect to X_3 under the true model; the total-effect Sobol's 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's 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 GLM 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's 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, Proposition 2.4 provides a unified way of estimating all Sobol's 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, Sobol's 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's indices with any level of desired accuracy.

Proposition 2.5. Approximation of Sobol's indices under Polynomial GLMs with Non-identity Inverse Link—*Suppose $\mathbf{X} = (X_1, \dots, X_n)$ 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 $K \in \mathbb{Z}^+$,*

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

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

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

and is also Lipschitz-continuous, by directly applying the Stone-Weierstrass Theorem [49, 53] we know that $\forall \epsilon > 0$, we may find a $\text{Poly}^{(K')}(\mathbf{X}_p)$, s.t.

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

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

If we assume, for instance, that 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 that satisfy the requirement of the link function in Proposition 2.5, such as the logit, the probit, the arcsin, the arccos, the inverse hyperbolic tangent, the inverse hyperbolic secant, etc. Since the inverse of these functions are all bounded and Lipschitz-continuous on any closed real interval, $E[Y|\mathbf{X}_P]$ is guaranteed to exist. Since the inputs' density is Lipschitz-continuous and defined on compact space, $E[Y|\mathbf{X}_P]$ is guaranteed to be Lipschitz-continuous as well.

2.5 Multiple Testing for Sobol's Indices

If we want to use Sobol's 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.* [11]. Suppose that M Sobol's indices are being tested for significance simultaneously. The null hypothesis of each test is that Sobol's index equals zero. Suppose the outcomes of the M tests are summarized in the way as shown in Table 1. The type I error of these M tests is defined as $E(V)/M_0$. The power is defined as $E(S)/M_1$. The false discovery rate (FDR) is defined as $E(V/R)$.

In simulation studies, given a set of decision rules for the M tests, we approximate the quantities above by the empirical proportions. 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.* [11]. This algorithm can be easily rewritten in terms of the Sobol's indices as shown in the Algorithm 1 below.

Algorithm 1 The Plug-in Estimate of the False Discovery Rate

1. Calculate M Sobol's indices, $S_j, j = 1, \dots, M$ 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's indices, $S_j^k, j = 1, \dots, M$ for each permutation sample $k = 1, \dots, K$.
4. Given M thresholds $C_{jj} = 1, \dots, M$ for the M Sobol's indices, calculate:

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

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

3 Applications to Variables Selection and Gene Inference

3.1 Simulation Studies

In the following simulation studies, we will compare the performance of several different methods, including Sobol's sensitivity indices, in variable selection and variable ranking. Since all Sobol's indices have the same denominator $Var(Y)$, we will only estimate the numerators of Sobol's indices and use them to rank the importance of the inputs.

Simulation Setup.—We first simulate a sample of 40 input variables from a multivariate normal 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 a normal distribution with different mean calculated according to $E(Y|\mathbf{X}) = \beta_0 + \mathbf{X}^T \boldsymbol{\beta}$, and fixed standard deviation that is one random draw from a uniform distribution on $[3.5, 5]$. In order to test the performance of the estimation methods, 20 fake input variables are also simulated from the same multivariate normal distribution that is used to generate the 40 true inputs. The inputs are considered “fake” when 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 drawn independently from the same multivariate normal distribution.

Variables Selection Comparison.—The fact that under our simulation setup, the first order main effect Sobol's indices can be accurately estimated by only fitting univariate models, implies that the univariate analyses are sufficient for variable selections (or singleton feature selections) in such scenarios.

For example, 1000 data sets are generated according to the simulation setup described earlier, with pair-wise correlation among the inputs fixed at 0.8. Each data set contain 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 are available for performing variable selection procedures. For each data set, we perform variable selection using all of the following techniques: 1) the univariate linear regression; 2) the Kendall Tau analysis; 3) the analysis of variance (ANOVA) on multivariate model containing 20 true inputs and 20 fake inputs; 4) the multivariate linear regression containing 20 true inputs and 20 fake inputs; 5) Sobol's index with regression coefficients estimated under the incorrect multivariate linear model (containing 20 true inputs and 20 fake inputs) using iteratively reweighted least square (IRLS); 6) Sobol's index with regression coefficients estimated under the incorrect multivariate model using coordinate descent with Lasso penalty (CD-Lasso); 7) Sobol's indices with regression coefficients estimated using coordinate descent with Ridge penalty (CD-ridge); 8) Sobol's indices with regression coefficients estimated by coordinate descent with Elastic-Net penalty (CD-ElasticNet).

Table 2 lists the type I error, power, and false discovery rate (FDR) of these eight methods, estimated empirically using the 1000 simulation samples and a 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's 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 Sobol's indices estimated using coefficients obtained by IRLS. We can also vary the threshold on the p-values to generate the ROC curves for method comparison (shown in the left panel of Figure 1). From this figure, we can see that all univariate analyses perform almost equally well, and also outperform the multivariate analyses dramatically.

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. 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). The ROC curves for this scenario are plotted in the right panel of Figure 1. Based on these results, we conclude that when the inputs are weakly correlated, Sobol's 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.

We also performed similar variable selection methods comparisons under Poisson models with the log link. Figure 2 presents the ROC curves using simulation data from Poisson models with the log link and multivariate normal inputs with different input correlation. From this Figure we can see that when the input variables are not strongly correlated, the performance of main-effect Sobol's indices in variable selection might sometimes be outperformed by a powerful non-parametric independence test like Kendall's Tau.

Variables Ranking Comparison.—In this section, we compare the ranking of input variables based on five different importance measures: 1) p-values of the Kendall's Tau independence test; 2) The main-effect Sobol's indices; 3) The totaleffect Sobol's indices; 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 threshold on the input variable. In the following discussion we will use *Kendall's_Tau*, *Main_SI*, *Total_SI*, *RF_1* and *RF_2* when referring to 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 [3, 4, 51]. In our simulations we use their implementation from the *randomForest* R package.

Table 4 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. To help visualize the results, we also presented the corresponding heat map, shown in Figure 3. The distance defined in the heat-map is $0.5 \times (1 - \text{Spearman Rho})$. From the heat map, we can see that if the correct models are fitted, rankings produced by the maineffect Sobol's indices and Kendall's Tau p-values match up each other very well and so do the rankings by total-effect Sobol's indices and random forest importance measures. We also note that the rankings produced by main-effect and total-effect Sobol's indices are not very similar.

Table 5 lists the Mean Spearman Rho estimates calculated under the contaminated models. These contaminated models include half of the true and half of the fake inputs. From the 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 and the total-effect Sobol's indices. It hence appears that all five importance measures can reasonably well distinguish most of time 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 ?? and ?? in Appendix ?? on the journal website.

Figure 4 shows the heat map for assessing the similarity among the input variables ranked within top 20 by different importance measures. The distance in this heat map is defined via the usual Jaccard index [34]. Due to our setup, we only focus on common input variables among the top 20 picks. From the heat map, we can see that main-effect Sobol's indices perform very similar to Kendall's Tau, and the performance of the total-effect Sobol's indices is not similar to any other method.

Based on the simulations shown in this section, we conclude that the total-effect Sobol's indices appear more capable of correctly ranking the importance of input variables if we define the importance on the basis of the percentage of variance explained.

3.2 Identifying Gene-Gene Interactions on CYP3A4

Polymorphic drug metabolizing enzymes are the major causes of adverse drug reactions. Cytochrome P450s (CYPs) make 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's sensitivity indices to identify genes that are coexpressed with CYP3A4 using a published 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.* [56]. 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 the literature). For each of the 78 measurements, missing values are imputed by using the empirical mean of the observed data.

In the first order analysis, we estimate all Sobol's indices with respect to a single candidate gene by fitting a univariate polynomial model of degree 3. Although this degree choice is arbitrary, we verified that the analysis results did not noticeably change for higher degrees. For each candidate gene, we start with the full polynomial model, meaning the model contains all linear terms, quadratic terms, and cubic terms. Then we use a backward-forward stepwise procedure to select a polynomial form that has the best fit. The main-effect Sobol's indices are estimated by the empirical variances of the best fitting polynomial expressions (may have the highest degrees lower than 3). These Sobol's 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's indices with respect to a pair of genes. For each pair, we again start with the full cubic polynomial model, 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 as before use the backward-forward stepwise procedure to obtain an importance ranking for all possible gene pairs and also proceed similarly for all possible gene triplets. Because the total number of possible gene quadruplets is too large (1,426,425), in this example we only estimated Sobol's indices with respect to each of the 194,580 gene quadruplets from the top 200 gene triplets picked out in the third order analysis.

In comparison to ranking gene combinations according to the main-effect Sobol's indices, we can also rank gene combinations 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. By comparing the gene combination rankings based on these two different important measures, we find that the first order analyses based on these two measures give very similar importance ranking. But when it comes to decomposing the system into quadruplets, the structure deduced using deviance difference is way too concentrated around a few genes, and many important genes picked out in the first order analyses are not linked into this structure, which might not be biologically reasonable.

Figure 5 shows the common structure extracted out by both the main-effect Sobol's indices and the deviance difference. Figure 6 shows the structure only picked out by the main-effect Sobol's indices. In both figures, each node represent a candidate gene. The size of each node is proportional to the importance measure, either the main-effect Sobol's indices with respect to the single gene or the residual difference in univariate polynomial models. The edges connect top 3% gene quadruplets with the highest importance measures. The threshold of 3% is selected for readability of the graphical summaries. The experimental validation of our results is currently being performed and its detailed results will be discussed elsewhere.

4 Discussion

When Sobol's indices are used for identifying gene-gene interactions, there are two types of model averaging techniques which can be implemented to improve their performance in practice.

First, we can average multiple estimates of the same Sobol's index to get a more robust estimate of this index. One way to do so is to fit multiple different polynomials on the same original dataset and then estimate Sobol's index under each of these different models. In order to do so, all these polynomials have to contain the particular subset of input genes specified by this Sobol's index. These models can include more genes other than the ones specified by the index, but not less. The goal of the averaging is to improve the accuracy of Sobol's index estimate when there is only a fixed number of samples available for analysis.

Second, we can average multiple estimates of single-gene importance ranking to alleviate the bias induced from enforcing different polynomial structures on limited data. Taking our CYP3A4 application as an example, four different single-gene importance rankings can be computed using the ranking of main-effect Sobol's indices conditioning on single gene, gene pair, gene triplet, and gene quadruplet, respectively. It is easy to see that the ranking of main-effect Sobol's indices obtained by fitting univariate polynomials (i.e. conditioning on single input gene) gives us one possible single-gene importance ranking. Perhaps less obviously, we can also generate different single-gene importance rankings by counting how many times each gene appears in the top gene pairs, the top gene triplets, or the top gene quadruplets. For example, we can define the top gene pairs as the ones associated with second-order main-effect Sobol's index estimate values above certain threshold.

In addition to being used for ranking gene-gene interactions, Sobol's indices may have other potential applications in gene activity analysis. For example, we may define statistical epistasis via Sobol's indices. The conventional way of identifying 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, see, e.g., [14, 17]). 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 Sobol's index with respect to genotype indicators at all loci (including the product terms of these indicators) and the sum of Sobol's indices with respect to genotype indicators at each single locus. That is, one could claim statistical epistasis if Sobol's index for the interaction effect is significant. The advantage of such a method of assessing statistical epistasis is that estimation of each Sobol's index only requires fitting the corresponding lower-dimension projection in a fairly large class of GLMs. If the true model has the identity link, inference based on Sobol's 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, inference based on Sobol's indices is valid as long as the input variables are real-valued.

Sobol's indices also can be used to quantify the effects of any combination of gene regulators in dynamic Bayesian network frameworks. In the conventional dynamic Bayesian networks, the time dependency is modeled 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, for instance, in Kim *et al.* [20], 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, for instance, the time-varying dynamic Bayesian networks such models are difficult to interpret (see, e.g., [48]). Sobol's indices can offer an alternative, simple way to quantify regulation effects of any combination of the parent genes under the fitted model of dynamic Bayesian network.

5 Summary and Conclusions

In this paper, we showed that for a large group of GLMs, Sobol's indices can be estimated from data either by evaluating closed formulas or by numerically fitting simpler models containing only partial inputs. To facilitate the usage of our methods in practice, we implemented Sobol's index formulas that derived under GLMs with identity, log, logit link and linear systematic component in an R package named *SobolSensitivity*. The installation files are available for downloading at <http://R-Forge.R-project.org>. 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 the covariates 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 covariates are defined on a compact space and have Lipschitz-continuous conditional densities. We also note that our SI estimation strategy does not assume nor does it require the knowledge of the full underlying model.

Our analysis suggests that reliable estimation of Sobol's indices can be achieved by finding good polynomial approximations of lower order projections (that is, the models containing only partial inputs). Our results for polynomial GLMs summarized in Propositions 2.4 and 2.5 can be easily generalized to GLMs for which the inverse-link transformed systematic component can be well approximated by piece-wise polynomials, since the proofs will still hold locally where the model is just a polynomial function. It further follows that, if a GLM can be approximated by another GLM with identity link and piece-wise polynomial systematic component, Sobol's indices under the true model can be estimated through fitting simpler models with identity link and piece-wise polynomial systematic components of partial inputs. Finally, we note that all the derived formulas and the approximation results are also applicable to multi-response models (where the response is a vector) under the identity link if the inputs are either independent or multivariate normal. This follows due to the well-known phenomenon of decoupling of the Gaussian multi-response regression models into univariate ones.

Acknowledgments

This work was partially supported by the US National Science Foundation (DMS-1318886), and the US National Cancer Institute (R01-CA152158). The authors have also received generous allocation of computing time from the Ohio Supercomputer Center at OSU as well additional logistical support from the Mathematical Biosciences Institute at OSU which receives its support under the US National Science Foundation grant DMS 1440386. The authors would like to acknowledge the comments of the associate editor and the referees which helped them improve the original manuscript.

APPENDIX

Estimation of Sobol's Sensitivity Indices under Generalized Linear Models

A Calculations for Linear GLMs

Proposition 2.1. Sobol's indices under Linear GLMs with Identity Link—Since $E(Y|X_1, \dots, X_n) = E(Y|\mathbf{X}^T \boldsymbol{\beta})$ under GLMs, we have:

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

Thus,

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

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

$$\begin{aligned}
E(Y|\mathbf{X}_P) &= E(E(Y|X_1, \dots, X_n)|\mathbf{X}_P) = E(\mathbf{X}^T\beta|\mathbf{X}_P) \\
&= \beta_0 + \mathbf{X}_P^T\beta_P + E(\mathbf{X}_Q^T\beta_Q|\mathbf{X}_P) \\
&= \beta_0 + \mathbf{X}_P^T\beta_P + E(\mathbf{X}_Q|\mathbf{X}_P)^T\beta_Q \\
&= \beta_0 + \mathbf{X}_P^T\beta_P + (\boldsymbol{\mu}_Q + \boldsymbol{\Sigma}_{QP}\boldsymbol{\Sigma}_{PP}^{-1}(\mathbf{X}_P - \boldsymbol{\mu}_P))^T\beta_Q \\
&= \text{constant} + \mathbf{X}_P^T(\beta_P + \boldsymbol{\Sigma}_{PP}^{-1}\boldsymbol{\Sigma}_{PQ}\beta_Q)
\end{aligned}$$

we have

$$\frac{\text{Var}(E(Y|\mathbf{X}_P))}{\text{Var}(Y)} = \frac{(\beta_P + \boldsymbol{\Sigma}_{PP}^{-1}\boldsymbol{\Sigma}_{PQ}\beta_Q)^T \boldsymbol{\Sigma}_{PP}(\beta_P + \boldsymbol{\Sigma}_{PP}^{-1}\boldsymbol{\Sigma}_{PQ}\beta_Q)}{\text{Var}(Y)}$$

Proposition 2.2. Sobol's indices under Linear GLMs with Log Link—Since $E(Y|X_1, \dots, X_n) = E(Y|\mathbf{X}^T\beta)$ under GLMs, we have:

$$\begin{aligned}
E(Y|X_i) &= E(E(Y|X_1, \dots, X_n)|X_i) = E(E(Y|\mathbf{X}^T\beta)|X_i) \\
&= E(\exp(\mathbf{X}^T\beta)|X_i) = \exp(\beta_0 + \beta_i X_i) E\left(\exp\left(\sum_{j \neq i}^n \beta_j X_j\right) \middle| X_i\right) \\
&= \exp(\beta_0 + \beta_i X_i) E(\exp(\mathbf{X}_{-i}^T\beta_{-i})|X_i).
\end{aligned}$$

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

$$E \left(e^{\mathbf{X}_{-i}^T \beta_{-i} | X_i} \right) = e^{E(\mathbf{X}_{-i}^T \beta_{-i} | X_i) + \frac{1}{2} \text{Var}(\mathbf{X}_{-i}^T \beta_{-i} | X_i)}.$$

$$E(\mathbf{X}_{-i}^T \beta_{-i} | X_i) = E(\mathbf{X}_{-i} | X_i)^T \beta_{-i}$$

Since

$$\begin{aligned} &= \sum_{j \neq i}^n \beta_j \left(\mu_j + \rho_{ji} \frac{\sigma_j}{\sigma_i} (X_i - \mu_i) \right) \\ &= \frac{\sum_{j \neq i}^n \beta_j \rho_{ji} \sigma_j}{\sigma_i} X_i + \sum_{j \neq i}^n \beta_j \left(\mu_j - \mu_i \rho_{ji} \frac{\sigma_j}{\sigma_i} \right) \\ &= \frac{\sum_{j \neq i}^n \beta_j \rho_{ji} \sigma_j}{\sigma_i} X_i + K_1^{(i)} \end{aligned}$$

$$\text{Var}(\mathbf{X}_{-i}^T \beta_{-i} | X_i) = \beta_{-i}^T \Sigma_{Q|P} \beta_{-i}$$

where

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

$\begin{bmatrix} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{bmatrix}$ is the partition of Σ corresponding to input vector partition

$\mathbf{X} = (\mathbf{X}_P = \mathbf{X}_i, \mathbf{X}_Q = \mathbf{X}_{-i}^T)^T$, we have:

$$\begin{aligned} E(Y | X_i) &= e^{\beta_0 + \beta_i X_i + \frac{\sum_{j \neq i}^n \beta_j \rho_{ji} \sigma_j}{\sigma_i} X_i + K_2^{(i)}} \\ &= e^{\beta_0 + K_2^{(i)}} e^{\left(\beta_i + \frac{\sum_{j \neq i}^n \beta_j \rho_{ji} \sigma_j}{\sigma_i} \right) X_i} \end{aligned}$$

where $K_2^{(i)} = K_1^{(i)} + \frac{1}{2}\beta_{-i}^T \Sigma_{Q|P} \beta_{-i}$. Note that the values of $K_2^{(i)}$ will differ depending on which input variable is chosen to be X_i .

Since

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

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

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

$$\text{Var}(X_*) = \left(e^{\frac{\sigma_*^2}{2}} - 1\right) e^{2\mu_* + \sigma_*^2}$$

Therefore,

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

Similarly, since

$$\begin{aligned} E(Y|X_P) &= E(E(Y|X_1, \dots, X_n)|X_1, \dots, X_{i_p}) \\ &= E\left(e^{X^T \beta} | X_{i_1}, \dots, X_{i_p}\right) \\ &= e^{\beta_0 + X_P^T \beta_P} E\left(e^{X_Q^T \beta_Q} | X_P\right) \\ &= e^{\beta_0 + X_P^T \beta_P} \times e^{E(X_Q^T \beta_Q | X_P)} + \frac{1}{2} \text{Var}(X_Q^T \beta_Q | X_P) \end{aligned}$$

$$\begin{aligned}
E\left(X_Q^T \beta_Q | X_P\right) &= E\left(X_Q | X_P\right)^T \beta_Q \\
&= \left(\mu_Q + \Sigma_{QP} \Sigma_{PP}^{-1} (X_P - \mu_P)\right)^T \beta_Q \\
&= X_P^T \Sigma_{PP}^{-1} \Sigma_{PQ} \beta_Q + \left(\mu_Q - \Sigma_{QP} \Sigma_{PP}^{-1} \mu_P\right)^T \beta_Q
\end{aligned}$$

$$\begin{aligned}
\text{Var}\left(X_Q^T \beta_Q | X_P\right) &= \beta_Q^T \text{Var}\left(X_Q | X_P\right) \beta_Q \\
&= \beta_Q^T \left(\Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{PQ}\right) \beta_Q
\end{aligned}$$

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

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

Therefore,

$$\begin{aligned}
\text{Var}(E(Y | X_P)) &= e^{2\beta_0 + 2K_2^{(P)}} \times \text{Var}\left(e^{X_P^T \eta}\right) \\
&= e^{2\beta_0 + 2K_2^{(P)}} \times \left(e^{\sigma_{**}^2} - 1\right) e^{2\mu_{**} + \sigma_{**}^2}
\end{aligned}$$

where

$$\mu_{**} = E\left(X_P^T \eta\right) = \mu_P^T \eta$$

$$\sigma_{**} = \text{Var}\left(X_P^T \eta\right) = \eta^T \Sigma_{PP} \eta$$

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

Therefore,

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

Proposition 2.3. Sobol's indices under Linear GLMs with Logit Link—Note

$$E(Y|X_i) = E \left(\frac{e^{\mathbf{X}^T \beta}}{1 + e^{\mathbf{X}^T \beta}} \middle| X_i \right) = E \left(\frac{e^{\tilde{X}}}{1 + e^{\tilde{X}}} \right)$$

where

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

$$\begin{aligned} \tilde{\mu} &= E(\mathbf{X}^T \beta | X_i) \\ &= \beta_i X_i + \sum_{j \neq i}^n \beta_j E(X_j | X_i) \\ &= \beta_i X_i + \sum_{j \neq i}^n \beta_j \left(\mu_j + \rho_{ij} \frac{\sigma_j}{\sigma_i} (X_i - \mu_i) \right) \\ &= \left(\beta_i + \sum_{j \neq i}^n \beta_j \rho_{ij} \frac{\sigma_j}{\sigma_i} \right) X_i + \sum_{j \neq i}^n \beta_j \left(\mu_j - \mu_i \rho_{ij} \frac{\sigma_j}{\sigma_i} \right) \end{aligned}$$

$$\tilde{\sigma}^2 = \text{Var}(\mathbf{X}^T \beta | X_i) = \text{Var} \left(\sum_{j \neq i}^n \beta_j X_j \middle| X_i \right) = \beta_{-i}^T \Sigma_{Q|P} \beta_{-i}$$

$$\beta_{-i} = (\beta_1, \beta_2, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_n)^T$$

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

$\begin{bmatrix} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{bmatrix}$ is a partition of Σ corresponding to the input vector partition

$$\mathbf{X} = (\mathbf{X}_P = \mathbf{X}_i, \mathbf{X}_Q^T = \mathbf{X}_{-i}^T)^T.$$

Then by applying Proposition A.1 below, we have the final expression of the maineffect index with respect to X_i :

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

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

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

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

where

$$Z \sim \ln N(0, \tilde{\sigma}^2)$$

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

$$\tilde{\sigma}^2 = \text{Var}(\mathbf{X}^T \beta | X_i) = \beta_{-i}^T \Sigma_{P|Q} \beta_{-i}$$

$\beta_{-i} = (\beta_1, \beta_2, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_n)^T$, $\Sigma_{P|Q} = \Sigma_{QQ} - \Sigma_{QP} \Sigma_{PP}^{-1} \Sigma_{PQ}$; $\begin{bmatrix} \Sigma_{PP} & \Sigma_{PQ} \\ \Sigma_{QP} & \Sigma_{QQ} \end{bmatrix}$ is the partition of Σ

corresponding to input vector partition $\mathbf{X} = (\mathbf{X}_P = \mathbf{X}_i, \mathbf{X}_Q = \mathbf{X}_{-i})^T$. The proof for main-effect index with respect to $\mathbf{X}_P$ is similar and hence omitted.

The following result may be obtained by direct computation.

Proposition A.1. 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:

$$\begin{aligned}
E\left(\frac{e^X}{1+e^X}\right) &= E\left(\frac{U}{1+U}\right) = E\left(\frac{1}{1+V}\right) \\
&= e^{-\frac{\mu^2}{2\sigma^2}} E\left(\frac{1+\frac{\mu}{\sigma^2}}{\frac{Z}{1+Z}}\right) = e^{-\frac{\mu}{2\sigma^2}} E\left(\frac{-\frac{\mu}{\sigma^2}}{\frac{Z}{1+Z}}\right), \quad \forall \frac{\mu}{\sigma^2} \in \mathbf{R} \\
&= e^{-\frac{\mu^2}{2\sigma^2}} \left[(-1)^{s-1} \frac{1}{2} + \sum_{i=1}^{s-1} (-1)^{i-1} e^{\frac{1}{2}(s-i)^2 \sigma^2} \right],
\end{aligned}$$

$$\begin{aligned}
s &= \begin{cases} 1 + \frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbf{Z}^+ \\ -\frac{\mu}{\sigma^2}, & \text{if } \frac{\mu}{\sigma^2} \in \mathbf{Z}^- \end{cases} \\
&= e^{-\frac{\mu^2}{2\sigma^2}} \left[(-1)^{\lfloor s \rfloor} E\left(\frac{Z^s - \lfloor s \rfloor}{1+Z}\right) + \sum_{i=1}^{\lfloor s \rfloor} (-1)^{i-1} e^{\frac{1}{2}(s-i)^2 \sigma^2} \right],
\end{aligned}$$

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

B Calculations for Polynomial GLMs

Proposition 2.4. Sobol's indices under Polynomial GLMs with Identity Link

Independent Inputs: For any $\mathbf{X}_P$ of interest, we can rearrange $\mathbf{X} = (\mathbf{X}_P, \mathbf{X}_{-P})$ to make $\mathbf{X}_P = (X_1, X_2, \dots, X_p)$, where $\mathbf{X}_{-P}$ is the complement set of $\mathbf{X}_P$.

Since $X_1, \dots, X_p$ are independent,

$$E\left[X_1^{k_1} X_2^{k_2} \dots X_n^{k_n} \middle| \mathbf{X}_P\right] = \prod_{i=1}^p X_i^{k_i} \prod_{i=p+1}^n E\left[X_i^{k_i}\right], \quad \forall 0 \leq k_1 + k_2 + \dots + k_n \leq K$$

Therefore,

$$E\left[\text{Poly}^{(K)}(\mathbf{X}) \middle| \mathbf{X}_P\right] = \text{Poly}^{(K')}(\mathbf{X}_P), \quad 1 \leq K' \leq K$$

Multivariate Normal Inputs: Approximation of Sobol's indices under Polynomial GLMs with Non-identity Inverse Link.

Let $\Sigma = \mathbf{L}\mathbf{D}\mathbf{L}^T$ where $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$ and $\mathbf{L}$ is a lower unit triangular.

$$\mathbf{W} = \mathbf{L}^{-1}\mathbf{X} \sim MN(\mathbf{L}^{-1}\boldsymbol{\mu}, \mathbf{D}).$$

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 $\mathbf{X}_P = \{X_1\}$,

$$\begin{aligned} E(Y|\mathbf{X}_P) &= E[E(Y|X_1, \dots, X_n)|X_1] \\ &= E\left[\text{Poly}^{(K)}(\mathbf{X}, \beta) \middle| X_1\right] \\ &= E\left[\text{Poly}^{(K)}(\mathbf{L}\mathbf{W}, \beta) \middle| W_1\right] \\ &= E\left[\text{Poly}^{(K)}(\mathbf{W}, \beta^*) \middle| W_1\right] \end{aligned}$$

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

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

$$\forall 0 \leq k_1 + k_2 + \dots + k_n \leq K$$

Therefore,

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

Since any X_j can be chosen as the X_1 , we already proved the result for any singleton $\mathbf{X}_P = \{X_j\}$, $\forall 1 \leq j \leq n$.

For any $\mathbf{X}_P$ containing more than one input variable, we can choose these inputs as the first p variables in $\mathbf{X}$. Thus,

$$\begin{aligned} E(Y|\mathbf{X}_P) &= E[E(Y|X_1, \dots, X_n)|X_1, \dots, X_p] \\ &= E\left[\text{Poly}^{(K)}(\mathbf{X}, \beta) \middle| \mathbf{X}_P\right] = E\left[\text{Poly}^{(K)}(\mathbf{L}\mathbf{W}, \beta) \middle| (\mathbf{L}\mathbf{W})_P\right] = E\left[\text{Poly}^{(K)}(\mathbf{L}\mathbf{W}, \beta) \middle| \mathbf{W}_P\right] \end{aligned}$$

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

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

is because

$$\begin{aligned}
 E \left[\text{Poly}^{(K)}(\mathbf{LW}) \middle| (\mathbf{LW})_P \right] &= \int \text{Poly}^{(K)}(\mathbf{LW}) f_{\mathbf{LW} | (\mathbf{LW})_P} d(\mathbf{LW})_{-P} \\
 &= \int \text{Poly}^{(K)}(\mathbf{LW}) f_{\mathbf{LW} | (\mathbf{LW})_P} |\mathbf{L}_{-P}^{-1}|^{-1} d\mathbf{W}_{-P}
 \end{aligned}$$

Since $|\mathbf{L}_{-P}^{-1}| = 1$ and

$$\begin{aligned}
 f_{\mathbf{LW} | (\mathbf{LW})_P} &= \frac{f_{\mathbf{LW}}}{\int f_{\mathbf{LW}} d(\mathbf{LW})_{-P}} \\
 &= \frac{f_{\mathbf{LW}}}{\int f_{\mathbf{LW}} |\mathbf{L}_{-P}^{-1}|^{-1} d\mathbf{W}_{-P}} \\
 &= \frac{f_{\mathbf{LW}}}{\int f_{\mathbf{LW}} d\mathbf{W}_{-P}} \\
 &= f_{\mathbf{LW} | \mathbf{W}_P}
 \end{aligned}$$

we have:

$$\begin{aligned}
 E \left[\text{Poly}^{(K)}(\mathbf{LW}) \middle| (\mathbf{LW})_P \right] &= \int \text{Poly}^{(K)}(\mathbf{LW}) f_{\mathbf{LW} | (\mathbf{LW})_P} d\mathbf{W}_{-P} \\
 &= \int \text{Poly}^{(K)}(\mathbf{LW}) f_{\mathbf{LW} | \mathbf{W}_P} d\mathbf{W}_{-P} \\
 &= E \left[\text{Poly}^{(K)}(\mathbf{LW}, \beta) \middle| \mathbf{W}_P \right]
 \end{aligned}$$

where $\mathbf{LW}$ is the joint probability density function of $\mathbf{LW}$. Since $\{W_i, i = 1, \dots, n\}$ are independent,

$$\begin{aligned}
 E(Y | \mathbf{X}_P) &= E \left[\text{Poly}^{(K)}(\mathbf{LW}, \beta) \middle| \mathbf{W}_P \right] \\
 &= \text{Poly}^{(K')}(\mathbf{W}_P, \beta^*) \\
 &= \text{Poly}^{(K')}(\mathbf{X}_P, \beta'), \quad 1 \leq K' \leq K
 \end{aligned}$$

Once we know the coefficients in $\text{Poly}^{(K')}(\mathbf{X}_P)$, the main-effect Sobol's 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 .

C Additional Simulation Results

In Figure 7 we present the results of all eight methods after analyzing the same simulation sample with inputs correlation fixed at 0.8. 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's 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 7 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. 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 Sobol's index estimates obtained by using formula (1) with coefficients estimated using different fitting algorithms. Note that all the coefficients used in calculating these Sobol's indices are estimated under the incorrect multivariate model. But these approximated Sobol's indices still present a clear separation between the true and fake inputs, and all fake inputs have Sobol's indices close to zero.

We can also estimate p-values for the significance tests of Sobol's indices, where the null hypothesis is that Sobol's index equals zero. By permuting the response values to match up different inputs observations, we can repeatedly estimate Sobol's indices using different permutation samples to approximate the distribution of each Sobol's index under the null hypothesis given the specified model. The p-value is approximately the percentage of Sobol's index estimates that are larger than the one estimated under the original sample. Figure 8 gives the p-value estimates corresponding to Sobol's indices shown in Figure 7. The p-values of Sobol's 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.

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. The corresponding plots (analogous to Figure 7 and 8) are given in Figure 9 and 10.

List of Abbreviations

CD	ElasticNet: Coordinate Descent with Elastic Net Penalty
CD-Lasso	Coordinate Descent with Lasso Penalty
CD-Ridge	Coordinate Descent with Ridge Regression Penalty
CYPs	Cytochrome P450s
F-ANOVA	Functional ANOVA Decomposition
FDR	False Discovery Rate HDMR: High Dimensional Model Representation
IRLS	Iteratively Reweighted Least Square
Main SI	Main-effect Sobol's indices
Total SI	Total-effect Sobol's indices
RF_1	Random Forest Importance Measure 1, averaged increase in prediction error (measured by mean square error) after permuting the input variable

RF_2	Random Forest Importance Measure 1, total increase in node impurities (measured by Gini index for classification and measure by residual sum of squares for regression) after splitting the data by setting threshold on the input variable
ROC	Receiver Operating Characteristic
SI	Sobob Sensitivity Indices

References

- [1]. Annoni P, Brüggemann R and Saltelli A, (2011). Partial order investigation of multiple indicator systems using variance-based sensitivity analysis. *Environmental Modelling & Software*, 26(7), 950–958.
- [2]. Borgonovo E and Plischke E, (2016). Sensitivity analysis: A review of recent advances. *European Journal of Operational Research*, 248(3), 869–887.
- [3]. Breiman L, (2001). Random forests. *Machine learning*, 45(1), 5–32.
- [4]. Breiman L, (2002). Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA.
- [5]. Caniou Y and Sudret B, (2010). Distribution-based global sensitivity analysis using polynomial chaos expansions. *Procedia-Social and Behavioral Sciences*, 2(6), 7625–7626.
- [6]. Chastaing G, Gamboa F and Prieur C, (2012). Generalized hoeffdingsobol decomposition for dependent variables-application to sensitivity analysis. *Electronic Journal of Statistics*, 6, 2420–2448.
- [7]. Chastaing G, Gamboa F and Prieur C, (2015). Generalized Sobol sensitivity indices for dependent variables: numerical methods. *Journal of Statistical Computation and Simulation*, 85(7), 1306–1333.
- [8]. Crestaux T, Le Maitre O and Martinez JM, (2009). Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering & System Safety*, 94(7), 1161–1172.
- [9]. Draper NR and Smith H, (2014). *Applied regression analysis*. John Wiley & Sons.
- [10]. Fahrmeir L and Tutz G, (2013). *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media.
- [11]. Friedman J, Hastie T and Tibshirani R, (2001). *The elements of statistical learning* (Ed. 1). Springer, Berlin: Springer series in statistics.
- [12]. Friedman J, Hastie T and Tibshirani R, (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1–10. [PubMed: 20808728]
- [13]. Genuer R, Poggi JM and Tuleau-Malot C, (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225–2236.
- [14]. Gjuvslund AB, Hayes BJ, Omholt SW, and Carlborg O (2007). Statistical Epistasis Is a Generic Feature of Gene Regulatory Networks. *Genetics*, 175(1), 411–420. [PubMed: 17028346]
- [15]. Glen G and Isaacs K, (2012). Estimating Sobol sensitivity indices using correlations. *Environmental Modelling & Software*, 37, 157–166.
- [16]. Guisan A, Edwards TC and Hastie T, (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157(2), 89–100.
- [17]. Hartmann K, Seweryn M, Handleman SK, Rempala GA, and Sadee W (2016). Non-linear interactions between candidate genes of myocardial infarction revealed in mRNA expression profiles. *BMC Genomics*, 17, 738. [PubMed: 27640124]
- [18]. Hoover KD and Perez SJ, (1999). Data mining reconsidered: encompassing and the general to specific approach to specification search. *The econometrics journal*, 2(2), 167–191.
- [19]. Iooss B and Lematre P, (2015). A review on global sensitivity analysis methods In *Uncertainty Management in Simulation-Optimization of Complex Systems* (pp. 101–122). Springer US.

- [20]. Kim SY, Imoto S and Miyano S, (2003). Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Briefings in bioinformatics*, 4(3), 228–235. [PubMed: 14582517]
- [21]. Kucherenko SS, (2005). Global sensitivity indices for nonlinear mathematical models, *Review. Wilmott Mag*, 1, 56–61.
- [22]. Li G, Wang SW and Rabitz H, (2002). Practical approaches to construct RS-HDMR component functions. *The Journal of Physical Chemistry A*, 106(37), 8721–8733.
- [23]. Li G, Hu J, Wang SW, Georgopoulos PG, Schoendorf J and Rabitz H, (2006). Random sampling-high dimensional model representation (RSHDMR) and orthogonality of its different order component functions. *The Journal of Physical Chemistry A*, 110(7), 2474–2485. [PubMed: 16480307]
- [24]. Li G, Rabitz H, Yelvington PE, Oluwole OO, Bacon F, Kolb CE and Schoendorf J, (2010). Global sensitivity analysis for systems with independent and / or correlated inputs. *The Journal of Physical Chemistry A*, 114(19), 6022–6032. [PubMed: 20420436]
- [25]. Lilburne L and Tarantola S, (2009). Sensitivity analysis of spatial models. *International Journal of Geographical Information Science*, 23(2), 151–168.
- [26]. Marrel A, Iooss B, Laurent B and Roustant O, (2009). Calculations of Sobol indices for the Gaussian process metamodel. *Reliability Engineering & System Safety*, 94(3), 742–751.
- [27]. Mara TA and Tarantola S, (2012). Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering & System Safety*, 107, 115–121.
- [28]. McCullagh P and Nelder JA, (1989). *Generalized linear models* (Vol. 37). CRC press.
- [29]. Norton J, (2015). An introduction to sensitivity assessment of simulation models. *Environ. Model. Softw*, 69, 166–174.
- [30]. Paruolo P, Saisana M and Saltelli A, (2013). Ratings and rankings: voodoo or science. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(3), 609–634.
- [31]. Pianosi Francesca, et al. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, 79, 214–232.
- [32]. Rabitz H, Ali OF, Shorter J and Shim K, (1999). Efficient input–output model representations. *Computer Physics Communications*, 117(1), 11–20.
- [33]. Ratto M, Castelletti A, and Pagano A, (2012). Emulation techniques for the reduction and sensitivity analysis of complex environmental models. *Model. Softw*, 34, 1–4.
- [34]. Real R and Vargas J (1996). The Probabilistic Basis of Jaccard's Index of Similarity. *Syst. Biol* 45 (3):380–385.
- [35]. Rosolem R, Gupta HV, Shuttleworth WJ, Zeng X and Gonalves LGG, (2012). A fully multiple criteria implementation of the Sobol method for parameter sensitivity analysis. *Journal of Geophysical Research: Atmospheres*, 117(D7).
- [36]. Saisana M, Saltelli A and Tarantola S, (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2), 307–323.
- [37]. Sala-i-Martin XX, (1997). I just ran two million regressions. *The American Economic Review*, 178–183.
- [38]. Saltelli A, (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2), 280–297.
- [39]. Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M and Tarantola S, (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2), 259–270.
- [40]. Saltelli A and Annoni P, (2010). How to avoid a perfunctory sensitivity analysis. *Environmental Modelling & Software*, 25(12), 1508–1517.
- [41]. Saltelli A, Sobol IM, (1995). About the use of rank transformation in sensitivity analysis of model output, *Reliability Engineering & System Safety*, 50(3), 225–239.
- [42]. Archer GEB, Saltelli A, Sobol IM, (1997). Sensitivity measures, ANOVA-like techniques and the use of bootstrap, *Journal of Statistical Computation and Simulation*, 58(2), 99–12.

- [43]. Saltelli A, Tarantola S, Campolongo F and Ratto M, (2004). Sensitivity analysis in practice: a guide to assessing scientific models. John Wiley & Sons.
- [44]. Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M and Tarantola S, (2008). Global sensitivity analysis: the primer. John Wiley & Sons.
- [45]. Sobol IM, (1990). On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe Modelirovanie*, 2(1), 112–118.
- [46]. Sobol IM, (1994). A primer for the Monte Carlo method. CRC press.
- [47]. Sobol IM, (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation*, 55(1), 271–280.
- [48]. Song L, Kolar M and Xing EP, (2009). Time-varying dynamic Bayesian networks. *Advances in Neural Information Processing Systems*, 1732–1740.
- [49]. Stone MH, (1948). The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 21(5), 237–254.
- [50]. Sudret B (2008). Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 937: 964–979.
- [51]. Svetnik V, Liaw A, Tong C and Wang T, (2004). Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules *In* Multiple Classifier Systems (pp. 334–343). Springer Berlin, Heidelberg.
- [52]. Tarantola S, Gatelli D, Kucherenko SS and Mauntz W, (2007). Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliability Engineering & System Safety*, 92(7), 957–960.
- [53]. Trefethen LN, (2013). Approximation theory and approximation practice. Siam. [Section 14.08 – 14.081]
- [54]. Xu C and Gertner GZ, (2008). Uncertainty and sensitivity analysis for models with correlated parameters. *Reliability Engineering & System Safety*, 93(10), 1563–1573.
- [55]. Xue J, Zartarian VG and Nako S, (2010). The Stochastic Human Exposure and Dose Simulation (SHEDS)-Dietary Model Technical Manual. Prepared for the July, 20–22.
- [56]. Yang X, Zhang B, Molony C, Chudin E, Hao K, Zhu J, Gaedigk A, Suver C, Zhong H, Leeder JS and Guengerich FP, (2010). Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome research*, 20(8), 1020–1036. [PubMed: 20538623]

Figure 1:
ROC Curves for Method Comparison Under Gaussian Model with Identity Link

Figure 2:
ROC Curves for Method Comparison Under Poisson Model with Log Link

Figure 3:
Heat Map for Variable Ranking Comparison Under True Model

Figure 4:
Heat Map for Similarity Assessment among Top Picks

Figure 5:
Gene Quadruplets Picked by Both Sobol's indices and Deviance Difference

Figure 6:
Gene Quadruplets Picked Only by Main-effect Sobol's indices

Figure 7:
Variable Selection Methods Comparison (inputs correlation $p = 0.8$)

Figure 8:
Sobol's index Significance Test (inputs correlation $p = 0.8$)

Figure 9:
Variable Selection Methods Comparison (inputs correlation $p = 0.3$)

Figure 10:
Sobol's index Significance Test versus Other Methods (inputs correlation $\rho = 0.3$)

Figure 11:
Ranking Comparison Example Under Contaminated Gaussian Model($p = 0.8$)

Figure 12:
Ranking Comparison Example Under Contaminated Gaussian Model ($p = 0.3$)

Algorithm 1 The Plug-in Estimate of the False Discovery Rate

-
1. Calculate M Sobol's indices, $S_j, j = 1, \dots, M$ 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's indices, $S_j^k, j = 1, \dots, M$ for each permutation sample $k = 1, \dots, K$.
 4. Given M thresholds $C_{jj} = 1, \dots, M$ for the M Sobol's indices, calculate:

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

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

Table 1:Outcome Summary of M Significance Tests

	Called Not Significant	Called Significant	Total
H_0 True	U	V	M_0
H_1 True	T	S	M_1
Total	M-R	R	M

Table 2:Type I Error, Power, and FDR Estimates ($\rho = 0.8$)

Methods	Univariate Regression	Kendall's Tau	ANOVA	Multivariate Regression
Type I Error	0.0290	0.0303	0.0224	0.0218
Power	0.9721	0.9702	0.7815	0.7219
FDR	0.0289	0.0302	0.0279	0.0293
Methods	SI-IRLS	SI-CD.Lasso	SI-CD.Ridge	SI-CD.ElasticNet
Type I Error	0.0287	0.0289	0.0282	0.0291
Power	0.9720	0.9719	0.9735	0.9719
FDR	0.0287	0.0289	0.0281	0.0291

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*.

Table 3:Type I Error, Power, and FDR Estimates ($\rho = 0.3$)

Methods	Univariate Regression	Kendall's Tau	ANOVA	Multivariate Regression
Type I Error	0.0226	0.0222	0.0227	0.0214
Power	0.7917	0.7791	0.7784	0.7501
FDR	0.0277	0.0276	0.0283	0.0277
Methods	SI-IRLS	SI-CD.Lasso	SI-CD.Ridge	SI-CD.ElasticNet
Type I Error	0.0220	0.0215	0.0207	0.0215
Power	0.7910	0.7918	0.7901	0.7918
FDR	0.0271	0.0264	0.0255	0.0264

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*.

Table 4:

Variable Ranking Comparison by Mean Spearman Rho Under Correct Model

	Gaussian Model ($\rho=0.8$)	Gaussian Model ($\rho=0.3$)	Poisson Model ($\rho=0.8$)	Poisson Model ($\rho=0.3$)
Main_SI vs. Kendall'_Tau	0.965	0.937	0.958	0.942
RF_1 vs. RF_2	0.804	0.916	0.727	0.878
RF_1 vs. TotalSI	0.619	0.788	0.544	0.751
RF_2 vs. TotalLS	0.590	0.782	0.567	0.768
Main_SI vs. RF_2	0.647	0.623	0.625	0.628
RF_2 vs. Kendall's Tau	0.554	0.520	0.545	0.558
Main_SI vs. RF_1	0.451	0.553	0.425	0.551
RF_1 vs. Kendall's Tau	0.379	0.453	0.391	0.484
Main_SI vs. TotalSI	0.176	0.334	0.165	0.329
Kendall's_Tau vs. TotalSI	0.145	0.235	0.179	0.312

Table 5:

Variable Ranking Comparison by Mean Spearman Rho Under Contaminated Model

	Gaussian Model ($\rho=0.8$)	Gaussian Model ($\rho=0.3$)	Poisson Model ($\rho=0.8$)	Poisson Model ($\rho=0.3$)
Main_SI vs. Kendall' _Tau	0.939	0.957	0.936	0.955
RF_1 vs. TotalSI	0.715	0.722	0.581	0.682
RF_2 vs. TotalSI	0.725	0.757	0.637	0.722
RF_1 vs. RF_2	0.855	0.780	0.799	0.768
Kendall's_Tau vs. TotalSI	0.656	0.662	0.548	0.639
Main_SI vs. TotalSI	0.666	0.683	0.565	0.661
Main_SI vs. RF_1	0.777	0.688	0.699	0.651
Main_SI vs. RF_2	0.839	0.753	0.830	0.729
RF_1 vs. Kendall's Tau	0.761	0.678	0.675	0.643
RF_2 vs. Kendall's Tau	0.816	0.742	0.808	0.722