Supplementary Information

GAIT: Gene expression Analysis for Interval Time

Yoojoong Kim[1], Yeong Seon Kang[2], Junhee Seok[1*]

[1]School of Electrical Engineering, Korea University, Seoul, South Korea.

[2]Department of Business Administration, University of Seoul, Seoul, South Korea

Supplementary Methods

1. Detail Methods of GAIT

GAIT is to detect significantly associated genes with interval times of two events in the presence of censoring. Formally, for sample *i*, let T_{i1} and T_{i2} be the occurrence times of event 1 and 2, respectively. Also, let C_{i1} and C_{i2} be the censoring times of event 1 and 2, respectively. In the presence of censoring, instead of true occurrence time T_{i1} and T_{i2}, we observe censored time point X_{i1} and X_{i2}, which are given by $X_{i1} = \min(T_{i1}, C_{i1})$, $X_{i2} = \min(T_{i2}, C_{i2})$, and cersoring indicator Δ_{i1} and Δ_{i2}, which are given by $\Delta_{i1} = I(T_{i1} \leq C_{i1})$, $\Delta_{i2} = I(T_{i2} \leq C_{i2})$ where $I(\cdot)$ is a indicator function. GAIT is to find genes whose expression is significantly associated with interval time $T_2 - T_1$ based on observed data $\{X_1, \Delta_1, X_2, \Delta_2\}$ of *n* samples.

GAIT works through the following steps.

(Step 1) The estimation of joint probabilty distribution of T_1 and T_2, f_{T_1,T_2}.

From observed $\{X_1, \Delta_1, X_2, \Delta_2\}$, the joint distribution of T_1 and T_2 is estimated using the multivariate survival analysis of the optional Polya tree bayseian estimation (Seok *et al*, 2014). While Seok *et al* handles a general *p*-dimensional multivariate problem, GAIT simplifies it as a two-dimensional bivaraite problem for the computation efficiency. Since the detail procedure is fully described in Seok *et al*, here we breifly explain the method focusing on the simplification maded by GAIT.

GAIT uses an optional Polya tree (OPT) to estimate the joint distribution (Wong and Ma, 2010). An OPT is characterized by the likelihood $\Phi(A)$ for region *A* in a sample space Ω. $\Phi(A)$ is recursively calculated by $\Phi(A_{11})$, $\Phi(A_{12})$, $\Phi(A_{21})$, and $\Phi(A_{22})$ where A_{ij} is the *j*-th subregion of *A* when *A* is split at the center point of the T_i axis. Formally,

$$\Phi(A) = \frac{1}{2}\Phi_0(A) + \frac{1}{4}\sum_{i=1}^{2} \frac{B(N(A_{i1}) + 0.5, N(A_{i2}) + 0.5)}{B(0.5, 0.5)} \Phi(A_{i1})\Phi(A_{i2})$$

where $\Phi_0(A)$ is a likelihood when all sample points in *A* are uniformly distributed, $B(\cdot)$ is a beta function, and *N*(*A*) is the number of samples in region *A*. If region *A* has one or no sample, $\Phi(A) = \Phi_0(A)$. By recursively splitting, we can obtain $\Phi(A)$'s for all subregions of Ω obtained by binary splitting.

To calculate the joint distribution of T_1 and T_2, GAIT performs the following steps. For given region *A*,

(1) If $\frac{1}{2}\Phi_0(A) > \frac{1}{4}\sum_{i=1}^{2} \frac{B(N(A_{i1})+0.5, N(A_{i2})+0.5)}{B(0.5,0.5)} \Phi(A_{i1})\Phi(A_{i2})$, *A* is considered to have a uniform distribution. The probability density of *A* is calculated as $\frac{N(A)}{n|A|}$ where *n* is the total number of observed samples and $|A|$ is the area of *A*.

(2) Otherwise, the given region *A* is considered to have an non-uniform distribution and is split further. If

$$\frac{B(N(A_{11})+0.5, N(A_{12})+0.5)}{B(0.5,0.5)} \Phi(A_{11})\Phi(A_{12}) > \frac{B(N(A_{21})+0.5, N(A_{22})+0.5)}{B(0.5,0.5)} \Phi(A_{21})\Phi(A_{22})$$

A is split into A_{11} and A_{12}. Otherwise, it is split into A_{21} and A_{22}.

GAIT recursivly applies these steps for the partitioned subregions. Finally, GAIT partitions the whole sample space into subregions where samples are considered to be uniformly distributed. According to the number of samples in each region, the probability density is determined.

In the presence of censoring, the number of samples in region *A*, $N(A)$, cannot be obtained by counting samples because of missing observations. Since $N(A) = nP_A$ where *n* is the number of total samples and P_A is the

probability mass in A, $N(A)$ can be easily calculated from the joint distribution f_{T_1,T_2}. Let $N(A|f)$ be the estimated number of samples in A given joint distribution f. Also let $\mathrm{OPT}(N(\cdot))$ be the joint distribution from the OPT calculation with number of samples $N(\cdot)$ for all subregions as described in the above paragraph. Here, we have

$$f_{T_1,T_2} = \mathrm{OPT}\left(N(\cdot\,|f_{T_1,T_2})\right)$$

To solve, GAIT uses an iterative approach, which is $f_{T_1,T_2}^{(i+1)} = \mathrm{OPT}\left(N(\cdot\,|f_{T_1,T_2}^{(i)})\right)$. By repeating the iteration until $f_{T_1,T_2}^{(i)}$ converges, GAIT finds the final joint distribution.

The initial distirbution is obtained from the initial estimation of number of samples, $N^{(0)}(A)$. $N^{(0)}(A)$ is estimated assuming that the distribution of T_1 and T_2 are independent in each subregion A. Here, univariate Kaplan-Meier estimators are used to estimate the distribution of T_1 and T_2 within A. The initial distirbution is given by $f_{T_1,T_2}^{(1)} = \mathrm{OPT}\left(N^{(0)}(A)\right)$.

(Step 2) The Mote Carlo calculation of $\mathrm{E}[T_2 - T_1|X_1, \Delta_1, X_2, \Delta_2]$

From the estiamted joint distribution of T_1 and T_2, GAIT obtains the conditional distribution of T_1 and T_2 given the observed $\{X_{i1}, \Delta_{i1}, X_{i2}, \Delta_{i2}\}$ for sample i. There are four cases:

(1) $\Delta_{i1} = 1$ and $\Delta_{i2} = 1$: Since two events are observed, T_1 and T_2 are determined as X_{i1} and X_{i2}.

(2) $\Delta_{i1} = 1$ and $\Delta_{i2} = 0$: $T_{i1} = X_{i1}$, and $\Pr[T_{i2}|\,T_{i1} = X_{i1}, T_{i2} > X_{i2}]$ is obtained.

(3) $\Delta_{i1} = 0$ and $\Delta_{i2} = 1$: $T_{i2} = X_{i2}$, and $\Pr[T_{i1}|\,T_{i1} > X_{i1}, T_{i2} = X_{i2}]$ is obtained.

(4) $\Delta_{i1} = 0$ and $\Delta_{i2} = 0$: $\Pr[T_{i1}, T_{i2}|\,T_{i1} > X_{i1}, T_{i2} > X_{i2}]$ is obtained.

Then, $\mathrm{E}[T_2 - T_1|X_1, \Delta_1, X_2, \Delta_2]$ is obtained empirically because the software implementation of analytical expectation calculation is not straightforward. From the conditional distribution of T_1 and T_2, the pairs of (T_1, T_2) are randomly generated. By numerically averaging the randomly sampled $T_2 - T_1$, the empirical expecation is calculated.

(Step 3) Statistical inference

Let g_{ij} be the expression index of gene j of sample i. Also let y_i be $\mathrm{E}[T_{i2} - T_{i1}|X_{i1}, \Delta_{i1}, X_{i2}, \Delta_{i2}]$, the expected interval time obtained in Step 2. From the pairs of (y_i, g_{ij}) for $i = 1, 2, \ldots, n$, the statistical association between the gene expression of gene j and the expected interval time is estimated based on a simple linear model, $y_i = \beta_0 + \beta_1 g_{ij} + \epsilon_{ij}$ where ϵ_{ij} is a noise.

2. Simulation Settings

The simulations were performed by the following steps in three different settings.

Step 1. Simulation data generation

500 samples are generated for the simulations. Each sample consists of two event times (T_1, T_2) considered as true event times. The event times in each setting follows additive exponential distribution, log-normal distribution, and clayton model (Clayton, 1978) respectively. Censoring time points (C_1, C_2) are also generated following either independent exponential distirbution or log-normal distribution. The sample distributions are summarized in the below table.

	T	C
Additive exponential	$T_1 \sim Exp(1)$ $T_2 \sim T_1 + Z$ $Z \sim Exp(1)$, $Z \perp T_1$	$C_1, C_2 \sim Exp(0.5)$ $C_1 \perp C_2$
Log-normal	$\log\begin{pmatrix}T_1\\T_2\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right)$	$\log\begin{pmatrix}C_1\\C_2\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$
Clayton	$T_1, T_2 \sim S(t_1, t_2) = \{e^{t_1/\theta} + e^{t_2/\theta} - 1\}^{-\theta}$, where $\theta = 1$	$C_1, C_2 \sim Exp(0.5)$ $C_1 \perp C_2$

- $N(\mu, \Sigma)$ is a bivariate normal distribution with mean μ and covariance Σ, and $S(\cdot)$ denotes a bivariate survival function, i.e. $S(t_1, t_2) = \Pr[T_1 > t_1, T_2 > t_2]$.

Then, the 500 samples are censored by comparing the true event times and censoring time points. The censored time point X_1 and X_2 are given by $X_1 = \min(T_1, C_1)$, $X_2 = \min(T_2, C_2)$, and cersoring indicator Δ_1 and Δ_2 are given by $\Delta_1 = I(T_1 \le C_1)$, $\Delta_2 = I(T_2 \le C_2)$, where $I(\cdot)$ is a indicator function. Finally, the set of $\{X_1, \Delta_1, X_2, \Delta_2\}$ is provided as a data matrix for the simulations.

Step 2. Gene expression generation

A gene expression matrix with 1,000 genes is generated for the 500 samples in Step 1. The 1,000 genes are classified into four groups; group 1 (n=100) is correlated with T_1, group 2 (n=100) is correlated with T_2, group 3 (n=100) is correlated with interval times ($T_2 - T_1$), and group 4 (n=700) is random noise. The details are following.

Group	Number	Distribution
1	100	$g_{ij} = T_{i1} + Z_{ij}$
2	100	$g_{ij} = T_{i2} + Z_{ij}$
3	100	$g_{ij} = T_{i2} - T_{i1} + Z_{ij}$
4	700	$g_{ij} = Z_{ij}$

Here, g_{ij} is the gene expression index of sample *i* and gene *j*. T_{i1} and T_{i2} are the true event time T_1 and T_2 of sample *i*, respectively. Z_{ij} is Gaussian noise of which mean is 0 and varaince is 2^2. After then, the generated gene expression indices are standardized to have mean 0 and varaince 1^2.

Step 3. Statistical inference

The goal of the proposed GAIT is to sort out the group 3 genes associated with interval time among the whole genes. The group 3 genes are considered as positives to be detected, and the rest genes are considered as negatives to be neglected. The performance of GAIT is compared with the following methods.

1. Univarite T_1: This method finds genes associated with only T_1. $\{X_1, \Delta_1\}$ is regressed by gene expression indices with a Cox propotional harzard model. Associated genes are selected by the likelihood of the model.

2. Univarite T_2: This method finds genes associated with only T_2. Similarly with Univariate T_1, this method use Cox models on $\{X_2, \Delta_2\}$.

3. Ignore censoring: This method ignores censoring statues and considers X_1 and X_2 as T_1 and T_2. In other words, it finds significantly associated genes with $X_2 - X_1$ instead of $T_2 - T_1$ based on simple linear regression.

4. Without censoring: This method considers only a subset of samples, whose events are all observed. For these samples, $\Delta_1 = 1$ and $\Delta_2 = 1$. Consequently, $X_1 = T_1$ and $X_2 = T_2$. Significantly associated genes are selected by simple linear regression models on $X_2 - X_1$.

5. Multi-state model (MSM): The censored data of two events can be considered as panel data with four states. State 1 is when both events don't occur, state 2 is when only event 1 occurs, state 3 is when only event 2 occurs, and state 4 is when both event occur. Recently developed Markov multi-state models can be used to find significantly associated covariates to each state transition in the presence of censoring. Here, we used msm R-package for this analysis (Jackson *et al*, 2011).

The expression indices of a gene is provided as a covariate, and its association with state transition is measured. Since the goal is finding genes associated with interval time $T_2 - T_1$, the transitions from state 2 to 4 and from 3 to 4 are of interest. Let p_1 and p_2 be the p-values for the associations with transition from state 2 to 4 and from state 3 to 4, respectively. The overall significance of the association is measured by $\min(p_1, p_2)$.

Step 4. Evaluation of the detection performance

Group 3 genes, of which expression indices are generated to be correlated with interval times, are considered as conditional positives. The rest genes are conditional negatives. Based on the estimated significance in Step 3, a gene is detected to be positive if its p-value is less than a pre-defined significance level. Otherwise, it is considered to be negative. The true-positives (TP) are when a conditionaly positive genes are detected as positive. The false-positives (FP) are when a conditional negative genes are detected to be positive. For various significance levels, the true-positive rates and false positive rates are calcualted, and accordingly ROC curves are drawn and AUCs are calculated.

This simulation study was performed with 6 CPU cores which are Intel® Xeon® E5-2630 v2 @ 2.60GHz and 128GB RAM and the average elapsed times for GAIT were 63.9, 89.3 and 34.7 seconds in three different settings, respectively. All simulation codes are available at http://cdal.korea.ac.kr/GAIT.

3. Simulations with Structured Gene Expression Data

The simulations in the previous section assume that genes are indepdent to each other except the association with phenotypes. However, in a real situation, genes are correlated with each other because of many biological functions such as canonical pathways. To consider it, additional simulations were performed with structured expression patterns.

The simulation settings are similar with what is described in the previous section except the way generating gene expression (Step 2). Briefly, 20 genes are randomly selected to be correlated each other and a common random signal is added to their gene expression indices. This selection is repeated by 10 times. Finally, 10 groups of genes are formed, and 20 genes in the same group are correlated with each other. The details are given in the below.

Step 2. Gene expression generation

A gene expression matrix with 1,000 genes is generated for the 500 samples in Step 1. The 1,000 genes are classified into four groups; group 1 (n=100) is correlated with T_1, group 2 (n=100) is correlated with T_2, group 3 (n=100) is correlated with interval times ($T_2 - T_1$), and group 4 (n=700) is random noise. The details are following.

Group	Number	Distribution
1	100s	$g_{ij} = T_{i1} + \sum_k V_{ik} I(j \in P_k) + Z_{ij}$
2	100	$g_{ij} = T_{i2} + \sum_k V_{ik} I(j \in P_k) + Z_{ij}$
3	100	$g_{ij} = T_{i2} - T_{i1} + \sum_k V_{ik} I(j \in P_k) + Z_{ij}$
4	700	$g_{ij} = \sum_k V_{ik} I(j \in P_k) + Z_{ij}$

Here, g_{ij} is the gene expression index of sample *i* and gene *j*. T_{i1} and T_{i2} are the true event time T_1 and T_2 of sample *i*, respectively. Z_{ij} is Gaussian noise of which mean is 0 and varaince is 2^2.

For sample *i*, V_{ik} is a common random signal of group *k*, which is generated from a normal distribution with mean 0 and variance (σ_v^2) 1^2 or 2^2. P_k is a set of genes, and $I(j \in P_k) = 1$ if gene j is in P_k and otherwise it is 0. Each P_k consists of 20 genes randomly selected from the whole 1,000 genes. Here, 10 groups of genes are assumed.

After then, the generated gene expression indices are standardized to have mean 0 and varaince 1^2.

4. References

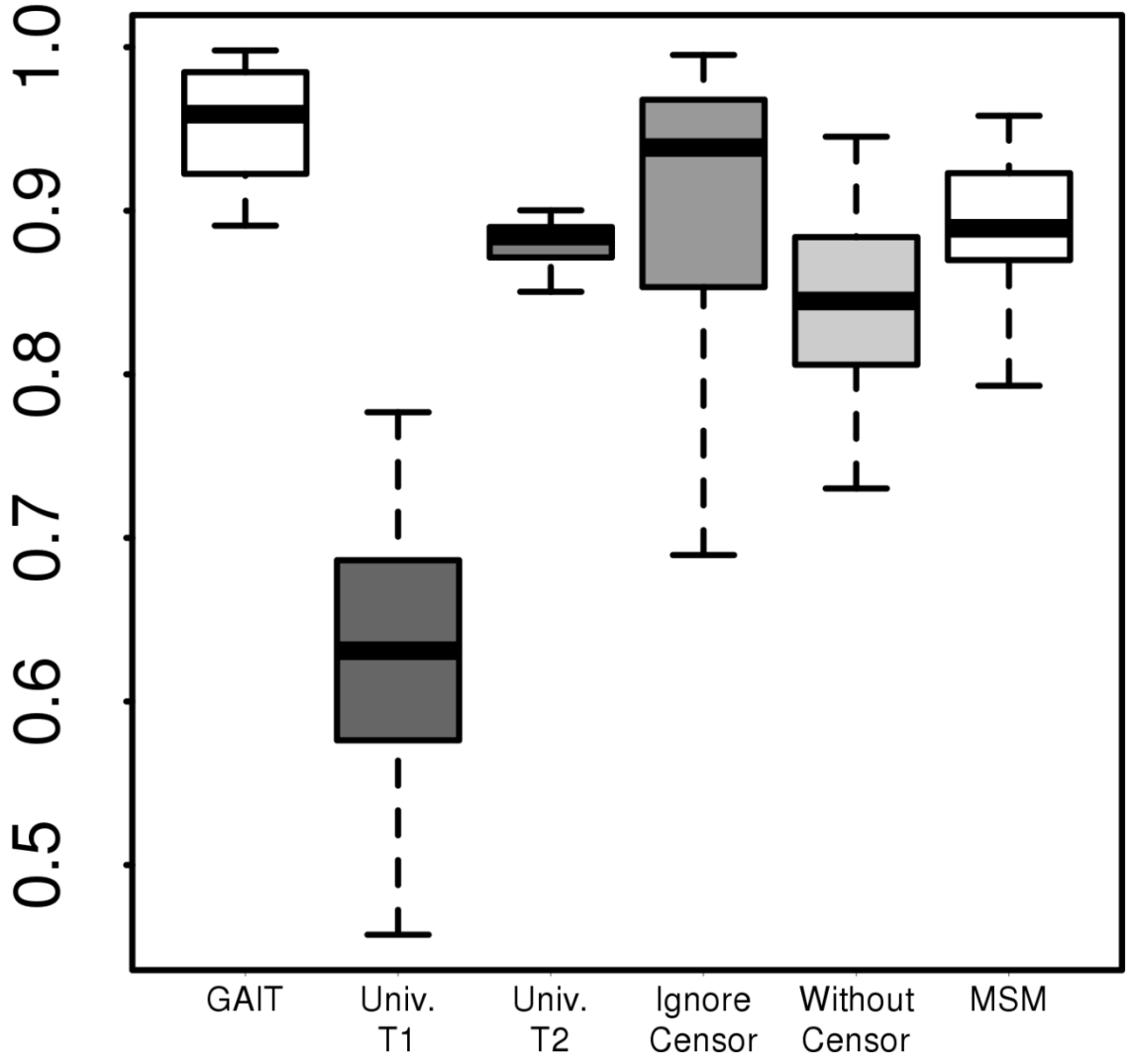
Clayton, D.G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65, 141-151.

Jackson, C.H. (2011) Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, 38, 1–29.

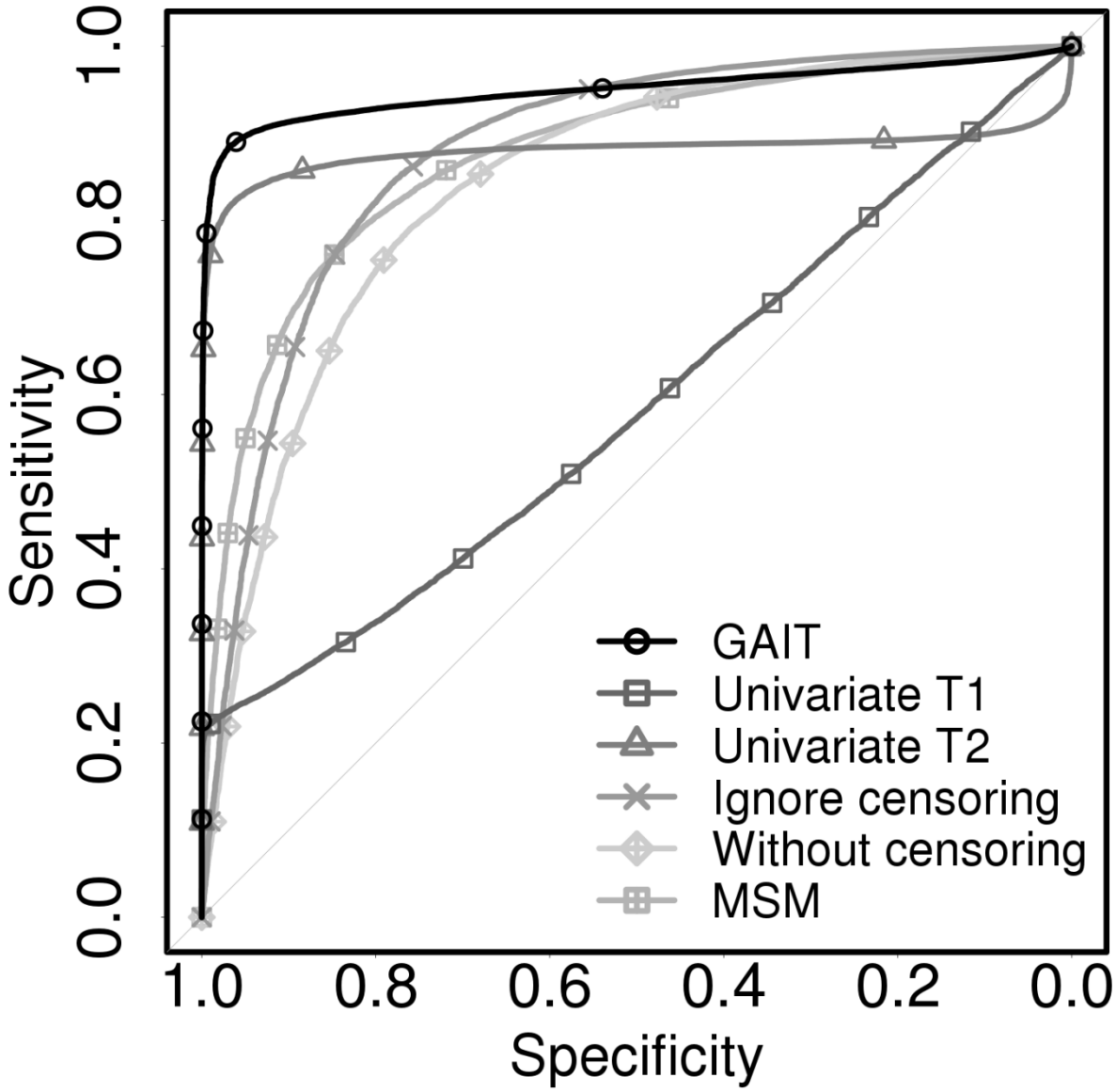
Seok, J. *et al*. (2014) Density estimation on multivariate censored data with optional Pólya tree. *Biostatistics*, 15, 182-195.

Wong, W. H. and Ma, L. (2010) Optional Polya tree and Bayesian inference. *The Annals of Statistics*, 38(3): 1433–1459.

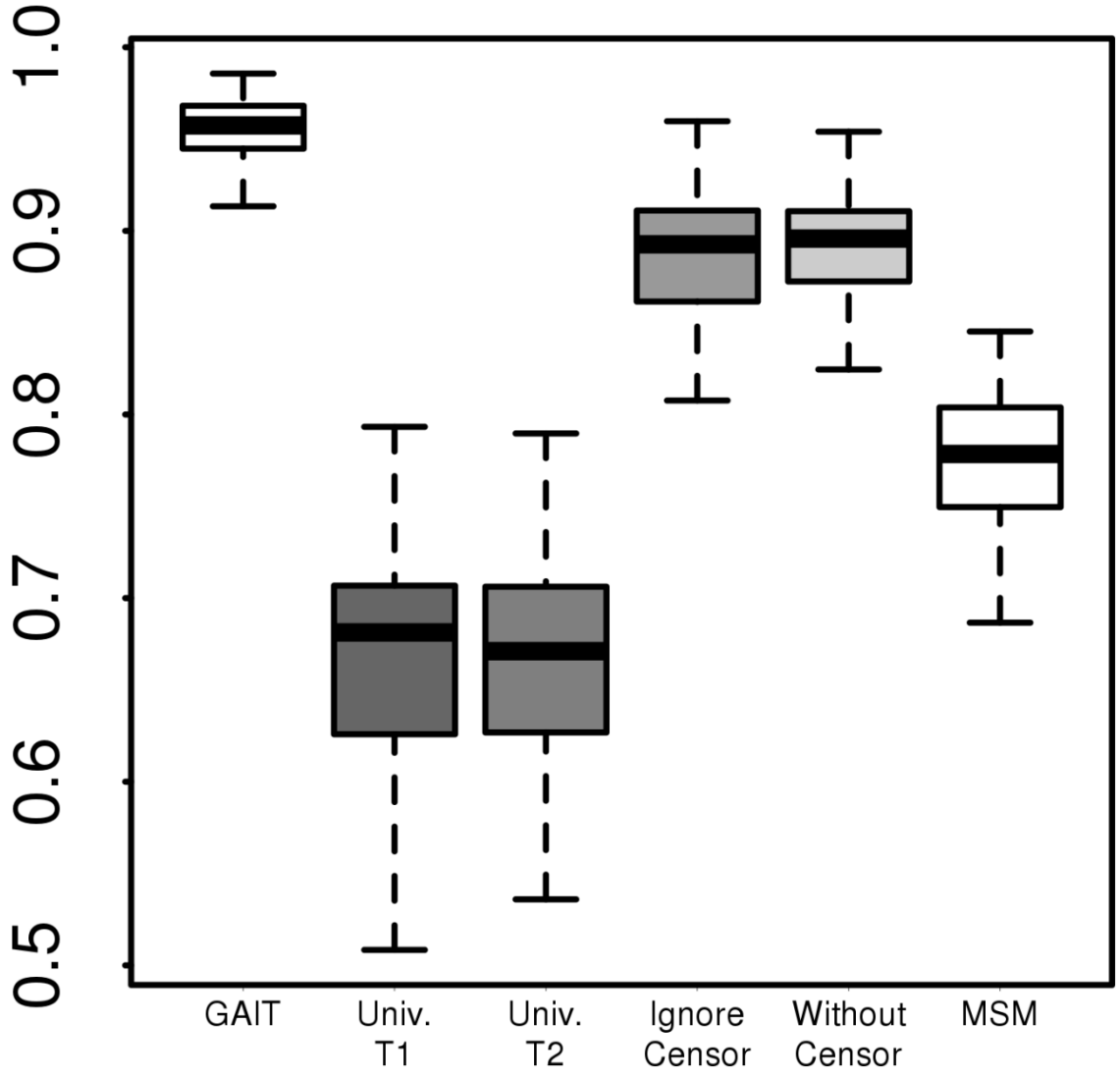
Supplementary Figures



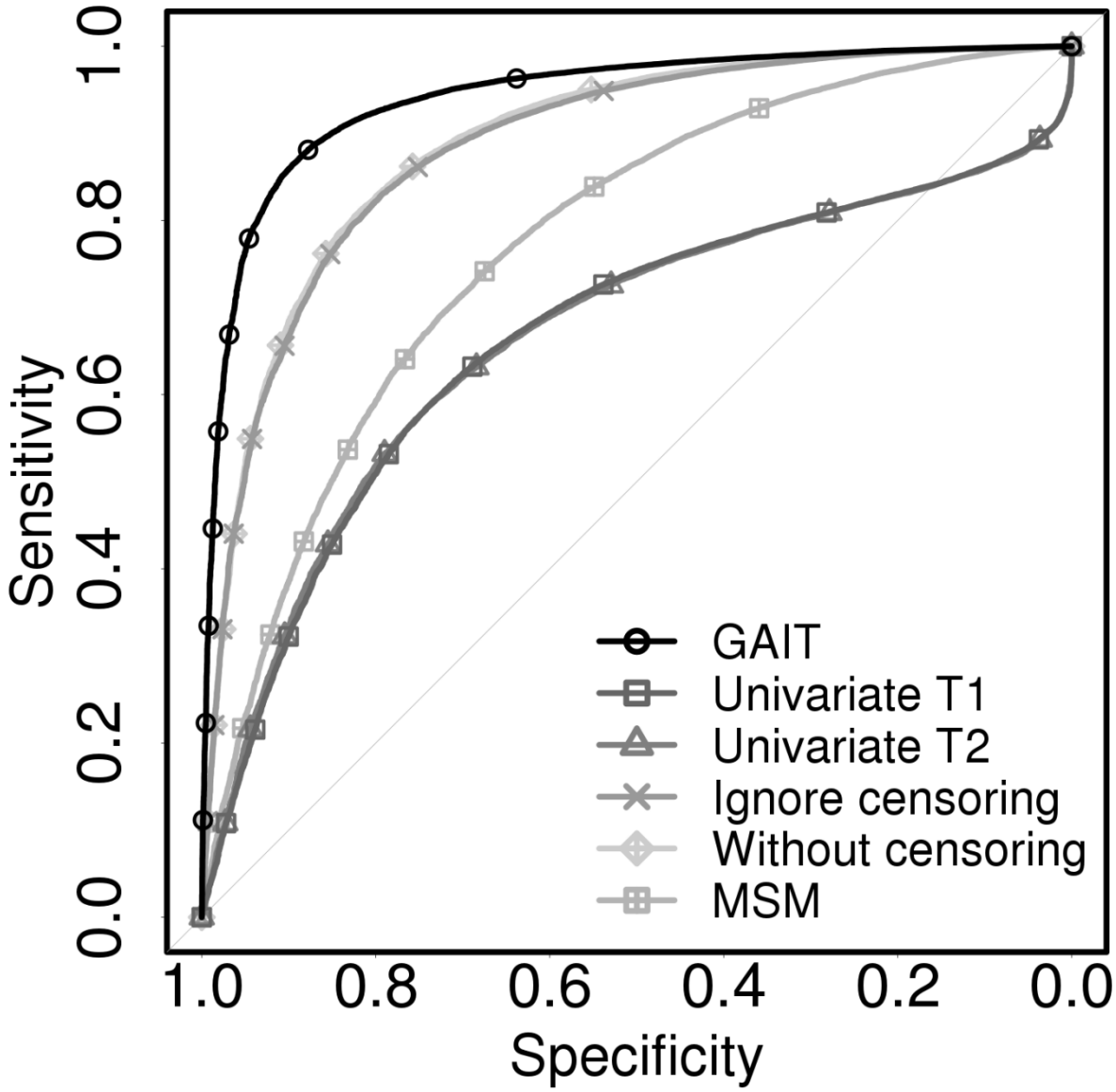
Supplementary Figure 1. The distribution of AUCs for 100 times of simulations in Log-normal distribution setting.



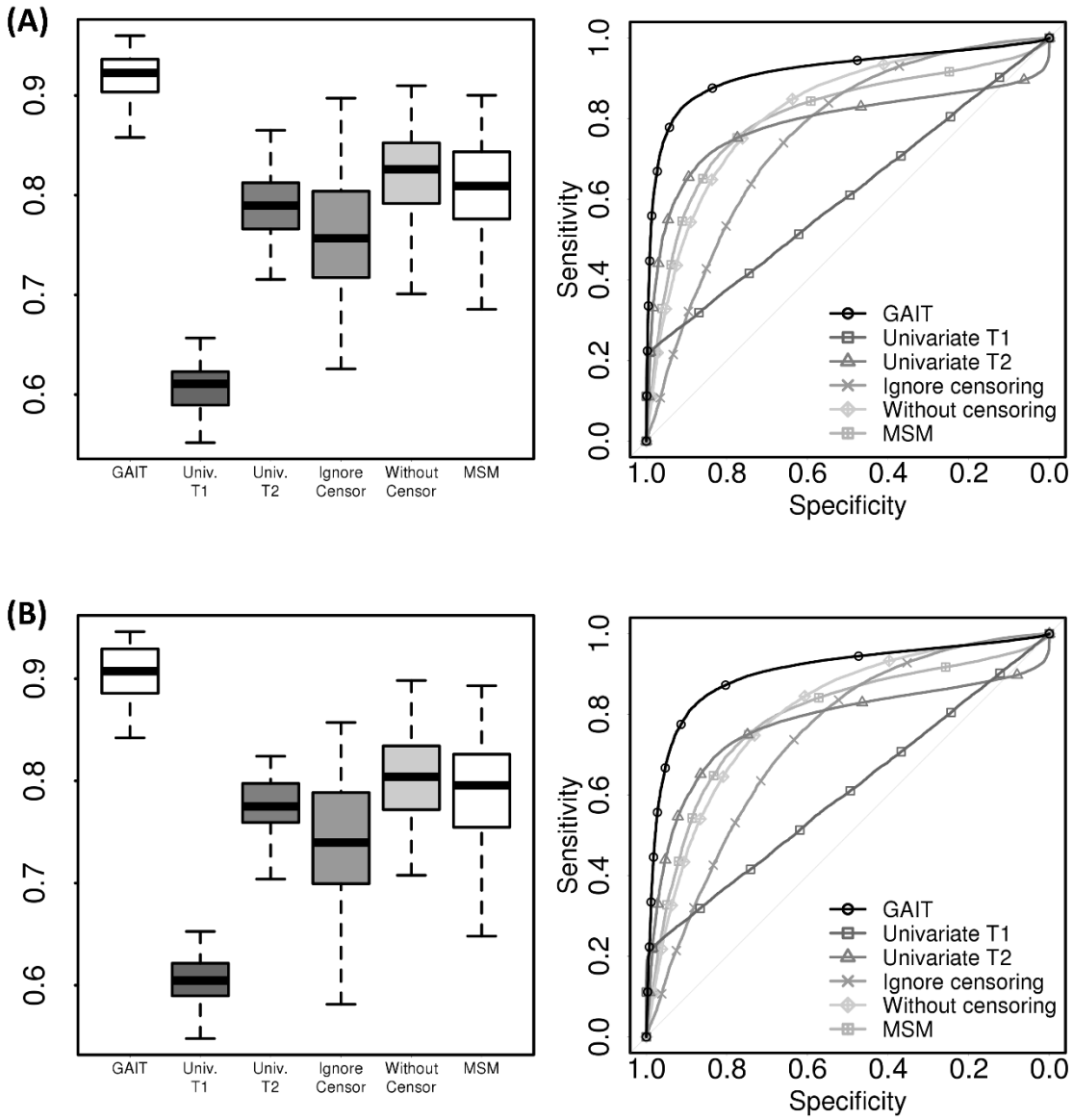
Supplementary Figure 2. The ROC curves for 100 times of simulations in Log-normal distribution setting (Supplementary Figure 1).



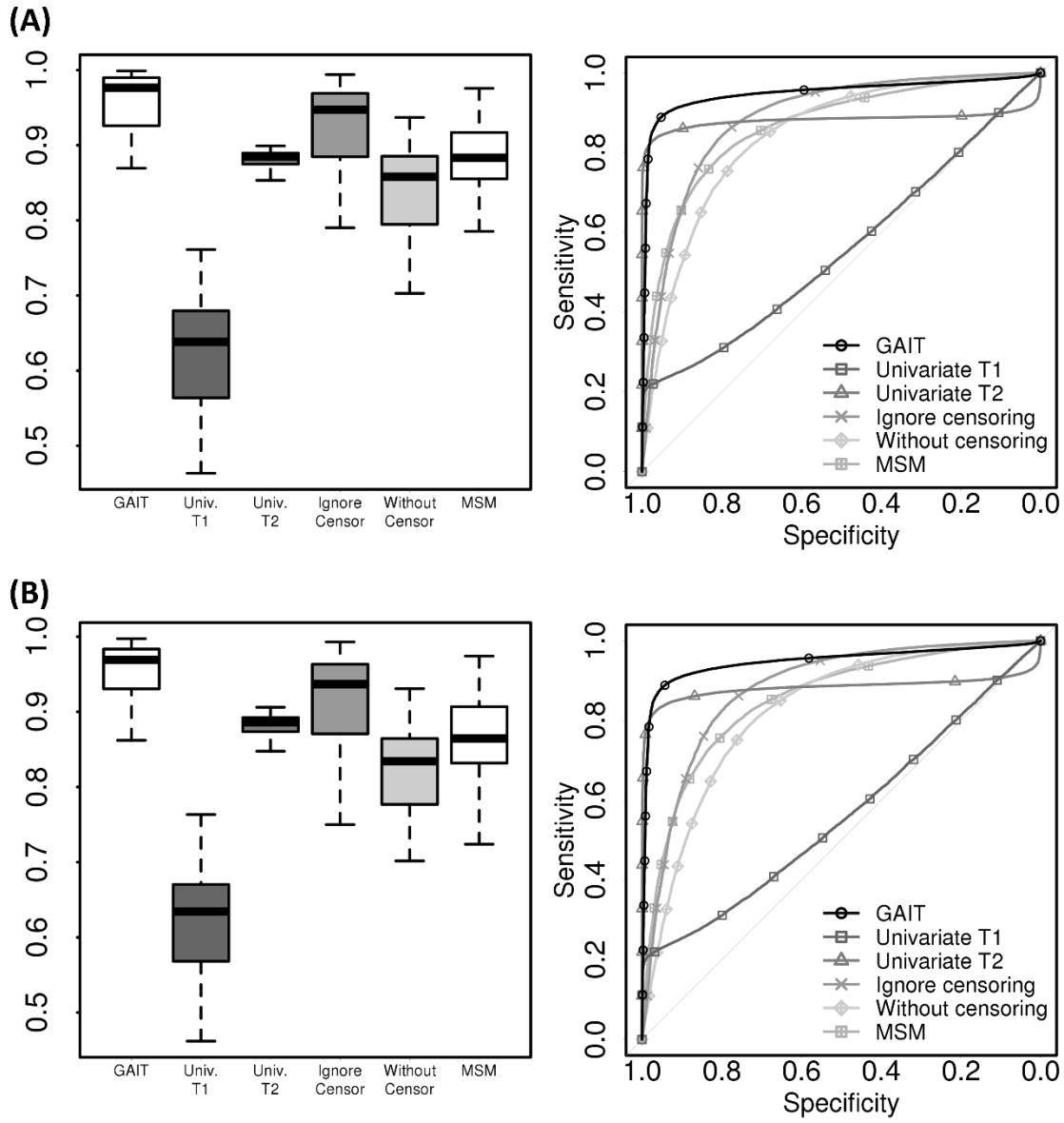
Supplementary Figure 3. The distribution of AUCs for 100 times of simulations in Clayton model setting.



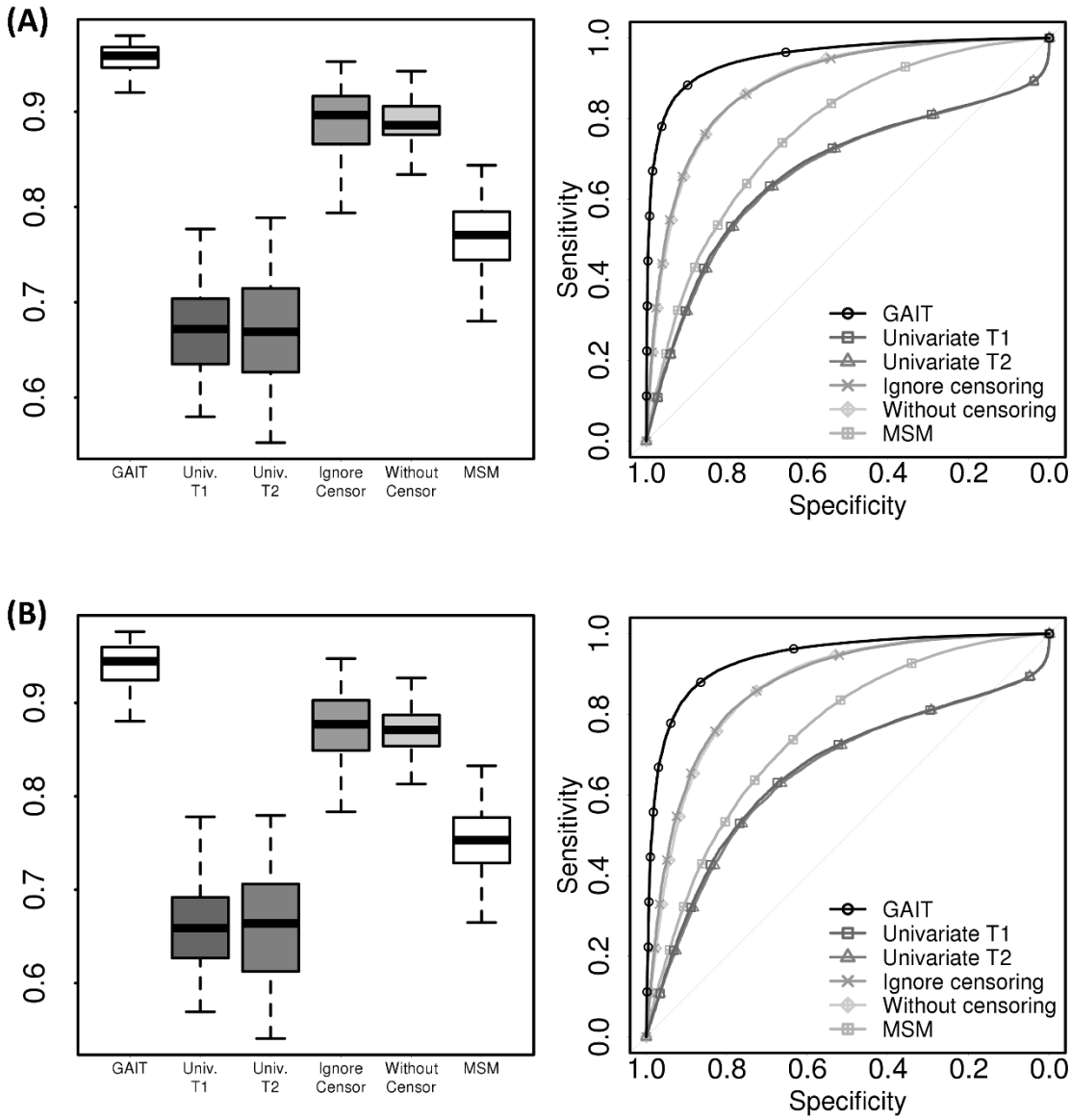
Supplementary Figure 4. The ROC curves for 100 times of simulations in Clayton model setting (Supplementary Figure 3).



Supplementary Figure 5. The distribution of AUCs and ROC curves for 100 times of simulations in Additive Exponential distribution setting and correlated gene structure with **(A)** σ_V =1 and **(B)** σ_V=2.



Supplementary Figure 6. The distribution of AUCs and ROC curves for 100 times of simulations in Log-Normal distribution setting and correlated gene structure with **(A)** σ_V =1 and **(B)** σ_V=2.



Supplementary Figure 7. The distribution of AUCs and ROC curves for 100 times of simulations in Clayton model setting and correlated gene structure with **(A)** σ_V =1 and **(B)** σ_V=2.

Supplementary Tables

Supplementary Table 1. The average AUCs of simulations in uncorrelated gene structures and correlated gene structures.

		GAIT	T_1	T_2	Ignore censoring	Without censoring	MSM
Additive Exponential	uncorrelated structure	0.92	0.61	0.80	0.76	0.83	0.81
	correlated structure ($\sigma_V = 1$)	0.92	0.61	0.79	0.76	0.82	0.81
	correlated structure ($\sigma_V = 2$)	0.90	0.61	0.78	0.74	0.80	0.79
Log-Normal	uncorrelated structure	0.95	0.62	0.88	0.89	0.84	0.89
	correlated structure ($\sigma_V = 1$)	0.95	0.62	0.88	0.90	0.84	0.88
	correlated structure ($\sigma_V = 2$)	0.95	0.62	0.88	0.89	0.83	0.86
Clayton	uncorrelated structure	0.94	0.67	0.67	0.89	0.89	0.78
	correlated structure ($\sigma_V = 1$)	0.95	0.67	0.67	0.89	0.89	0.77
	correlated structure ($\sigma_V = 2$)	0.94	0.66	0.65	0.87	0.87	0.75

- σ_v^2 is the variance of common signals of a gene group

Supplementary Table 2. Using 305 genes only found from applying GAIT for the real dataset of multiple myeloma patients, the five gene sets correlated with the genes were found from the Fisher's exact test. The gene sets are in the cancer modules of MSigDB.

Gene Sets	Odds Ratio	p-value
Neighborhood of Cyclin A2 (CCNA2) gene	9.69	1.25×10^{-6}
Neighborhood of Cell Division Cycle 20 (CDC20) gene	10.57	2.73×10^{-6}
Neighborhood of Cell Division Cycle 2 (CDC2) gene	9.39	5.98×10^{-6}
Neighborhood of Hyaluronan-Mediated Motility Receptor (HAMMR) gene	11.06	8.74×10^{-6}
Neighborhood of cyclin B2 (CCNB2) gene	8.84	3.21×10^{-6}