eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Department of Computer Science and Artificial Intelligence

Selection Procedures in Machine Learning: Scores, Estimations, and Reproducibility

Online supplementary documentation

Ari Urkullu Villanueva

Advisors
Aritz Pérez Martínez

Borja Calvo Molinos

Contents

1 Supplementary material for: Model Selection using crowdsourced data

1.1 Introduction

Here we gather all the plots corresponding to the results deriving from the experimentation carried out in Chapter 2 of the thesis manuscript. Briefly, three kind of plots are collected:

- Difficulty plots: Boxplots in which the AUCs of the classifiers throughout different runs are shown in order to illustrate the degree of the difficulty of ranking them correctly.
- Boxplots of the estimators of the AUC: Boxplots composed of the swap errors of the estimators of the AUC regarding the AUC.
- Scatter plots of the estimators of the AUC: Scatter plots composed of the swap errors of the estimators of the AUC regarding the AUC.

1.2 Difficulty plots

The difficulty plots are shown in Figures 1 to 10.

1.3 Boxplots of the estimators of the AUC

The boxplots of the estimators of the AUC are shown in Figures 11 to 30.

1.4 Scatter plots of the estimators of the AUC

The scatter plots of the estimators of the AUC are shown in Figures 31 to 50.

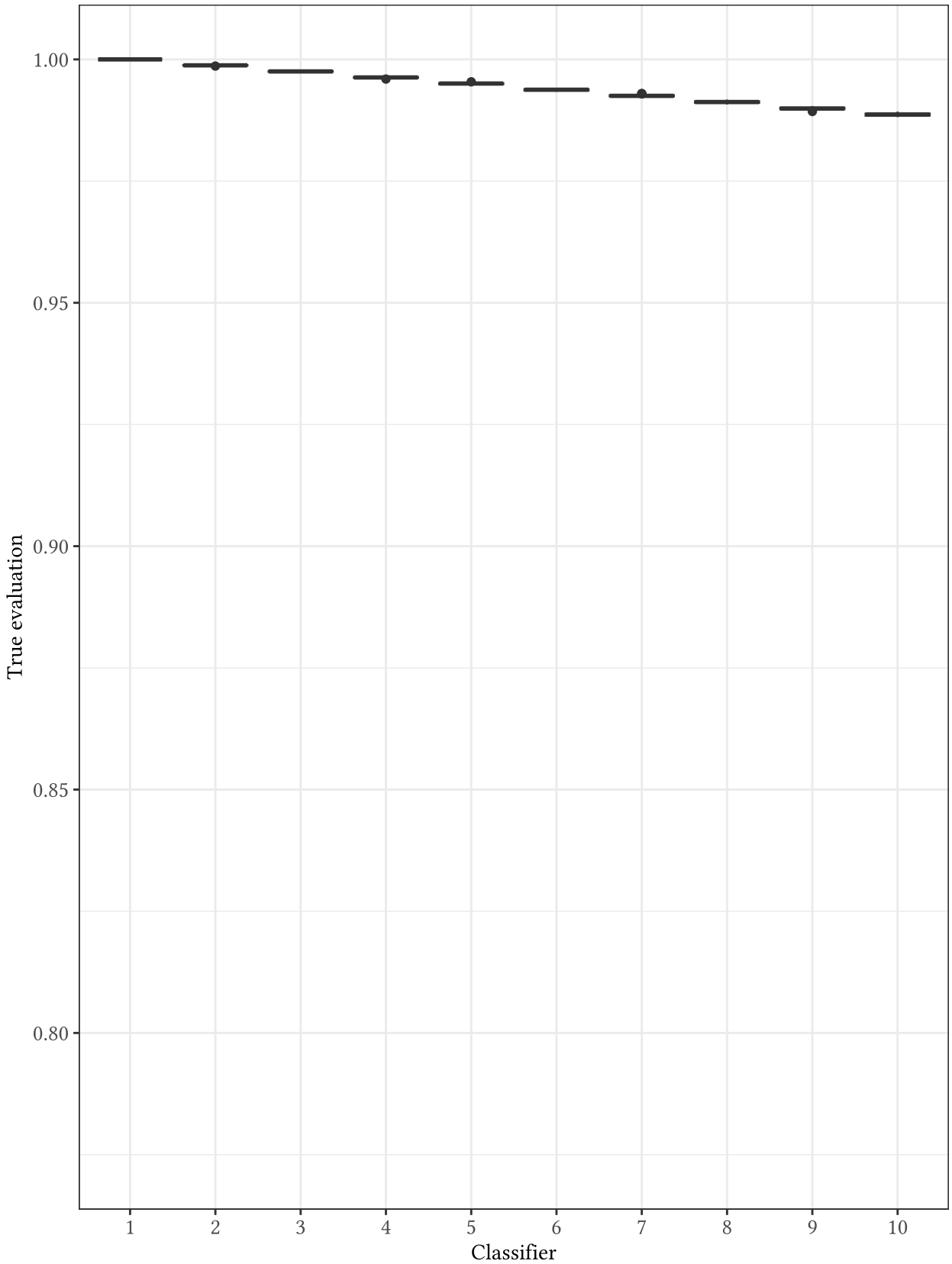


Figure 1: Boxplots of the AUCs of the classifiers $k = 1$ to $k = 10$ throughout the different runs.

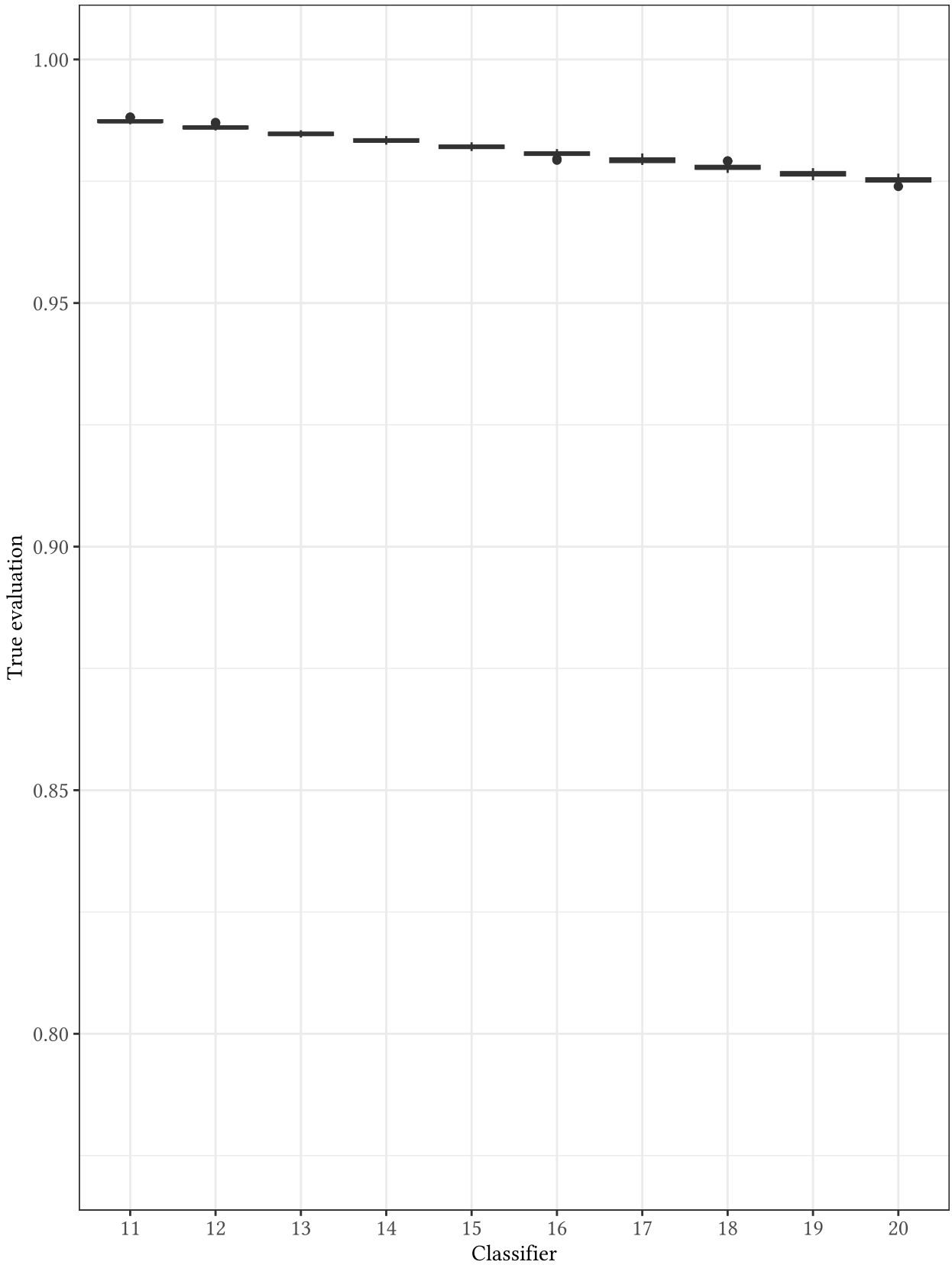


Figure 2: Boxplots of the AUCs of the classifiers $k = 11$ to $k = 20$ throughout the different runs.

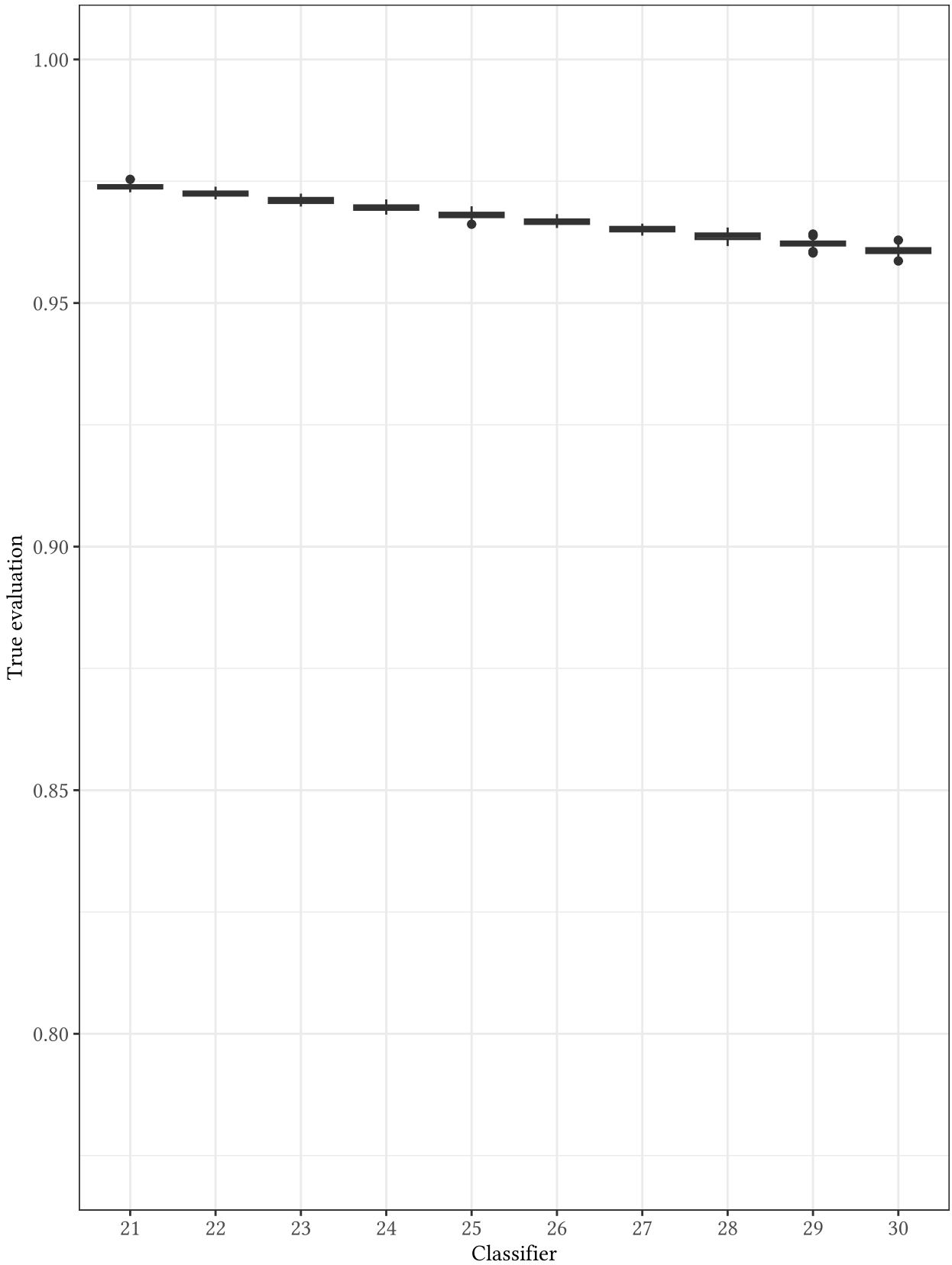


Figure 3: Boxplots of the AUCs of the classifiers $k = 21$ to $k = 30$ throughout the different runs.

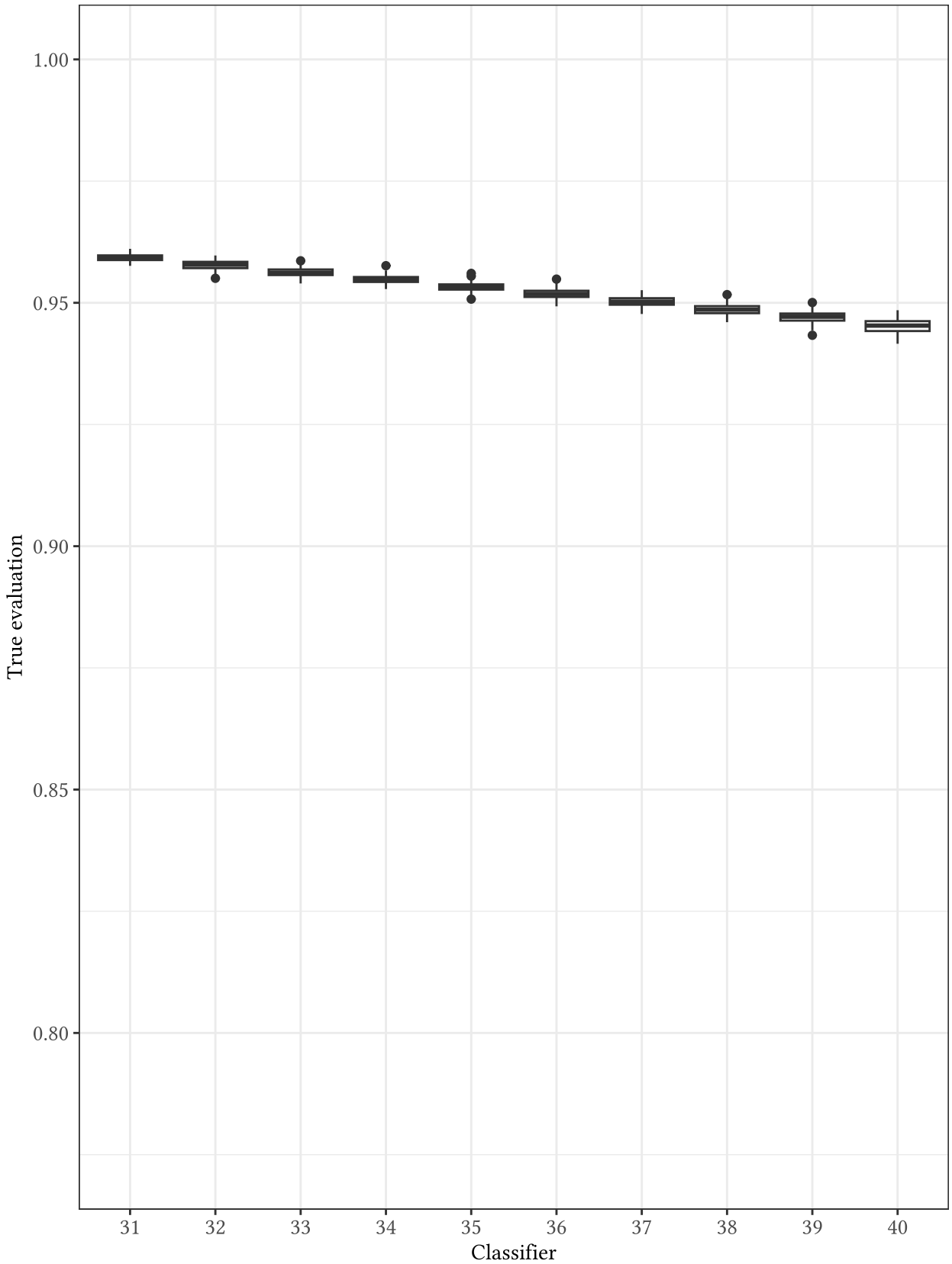


Figure 4: Boxplots of the AUCs of the classifiers $k = 31$ to $k = 40$ throughout the different runs.

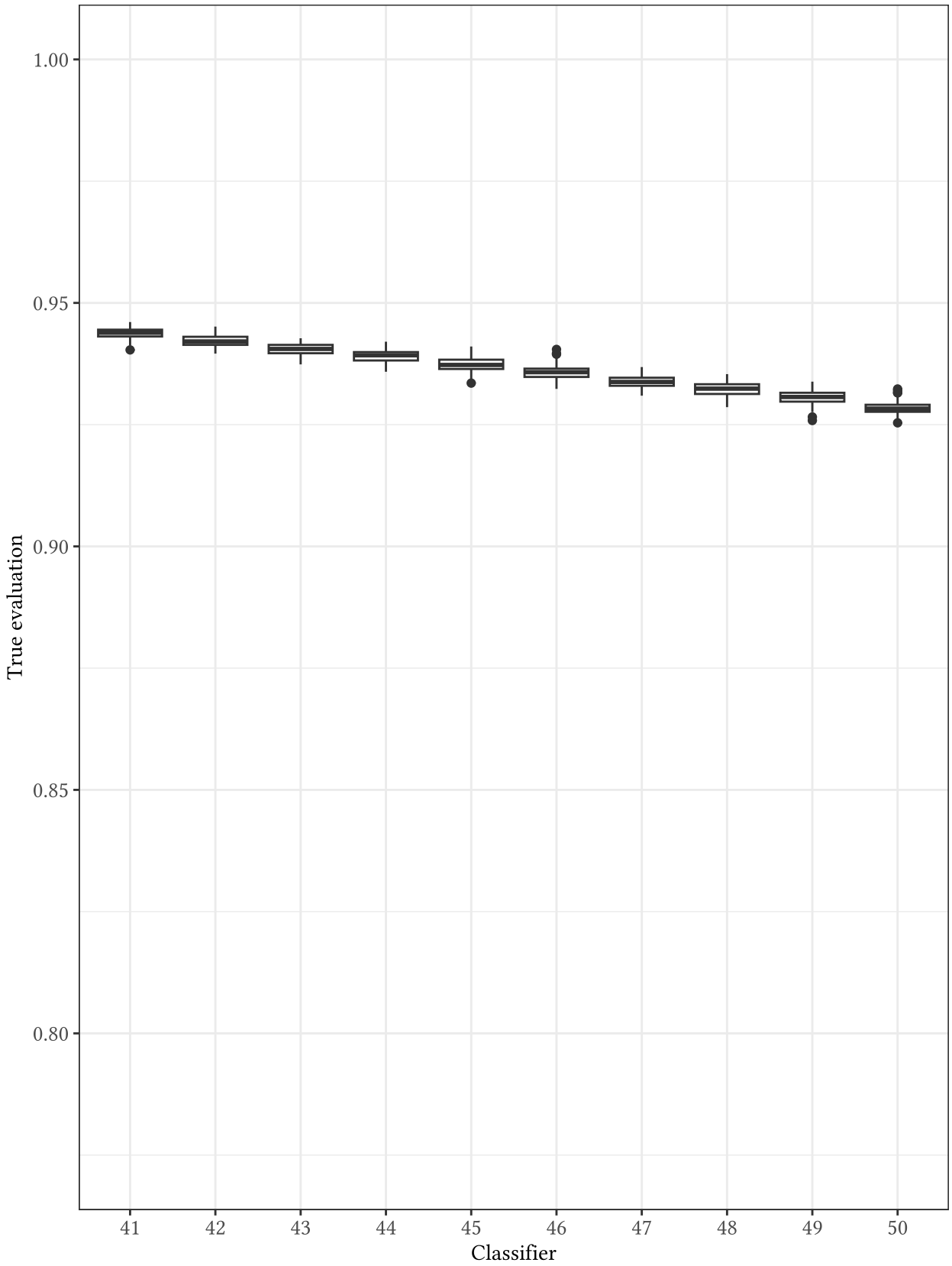


Figure 5: Boxplots of the AUCs of the classifiers $k = 41$ to $k = 50$ throughout the different runs.

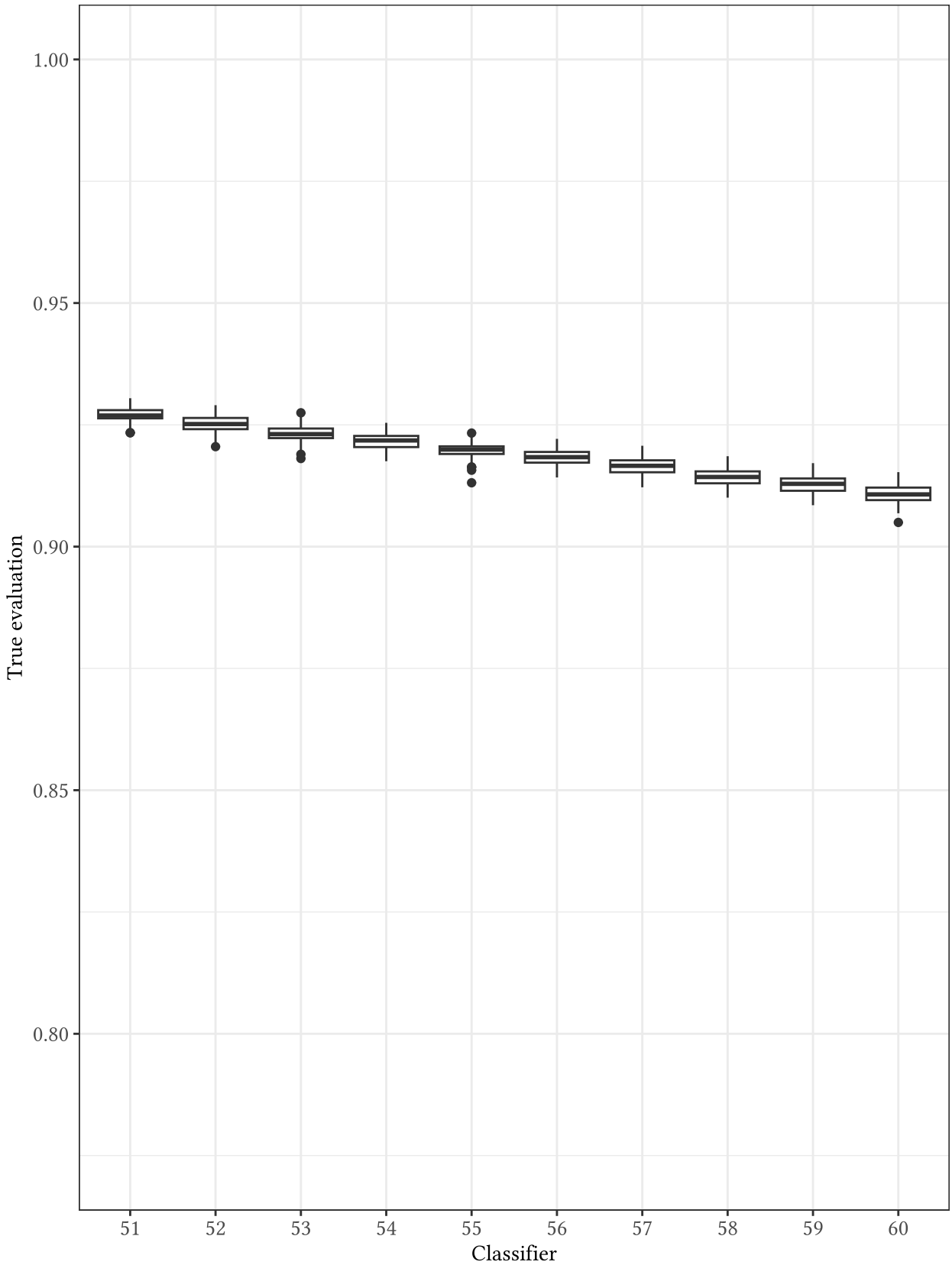


Figure 6: Boxplots of the AUCs of the classifiers $k = 51$ to $k = 60$ throughout the different runs.

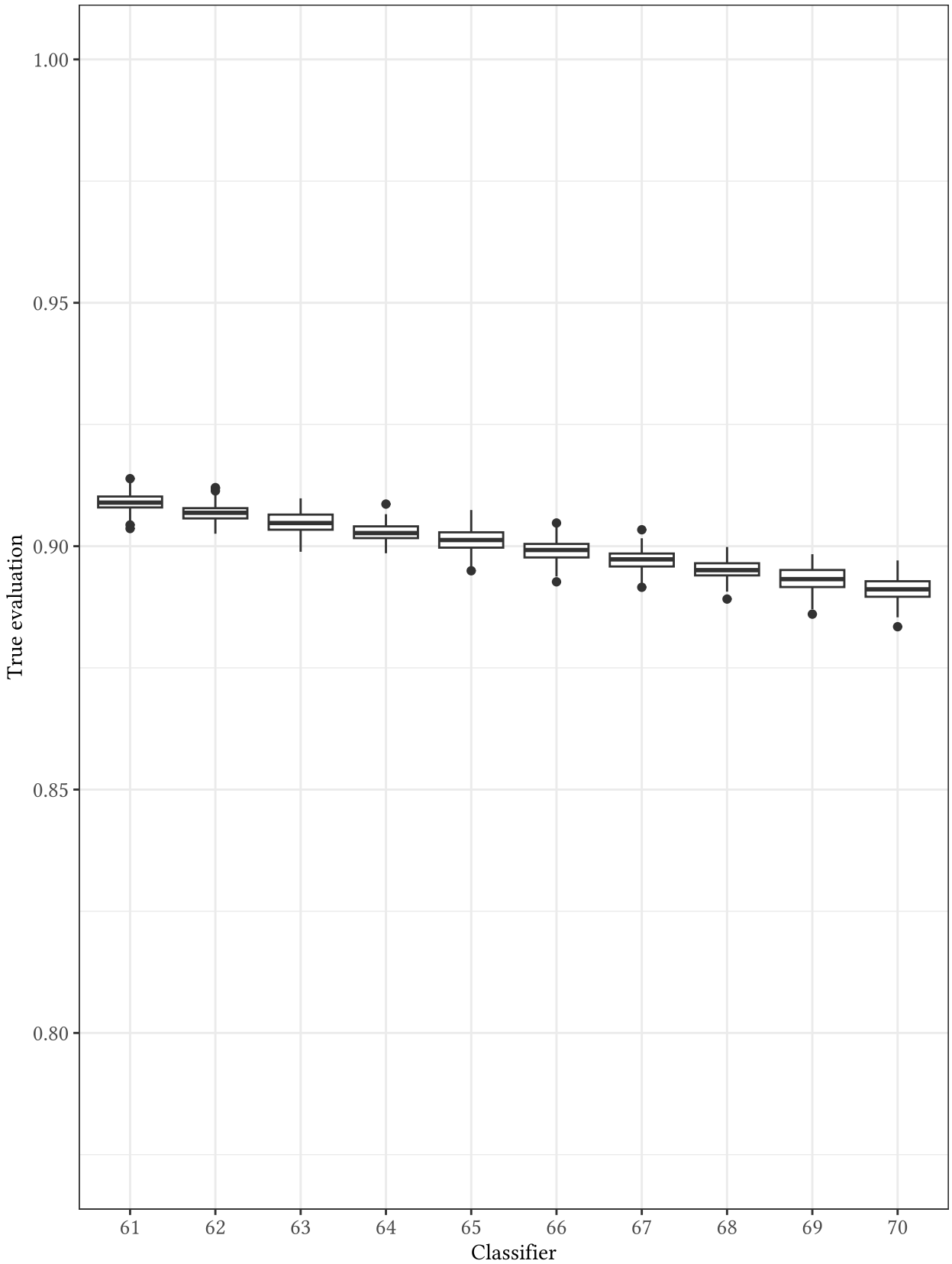


Figure 7: Boxplots of the AUCs of the classifiers $k = 61$ to $k = 70$ throughout the different runs.

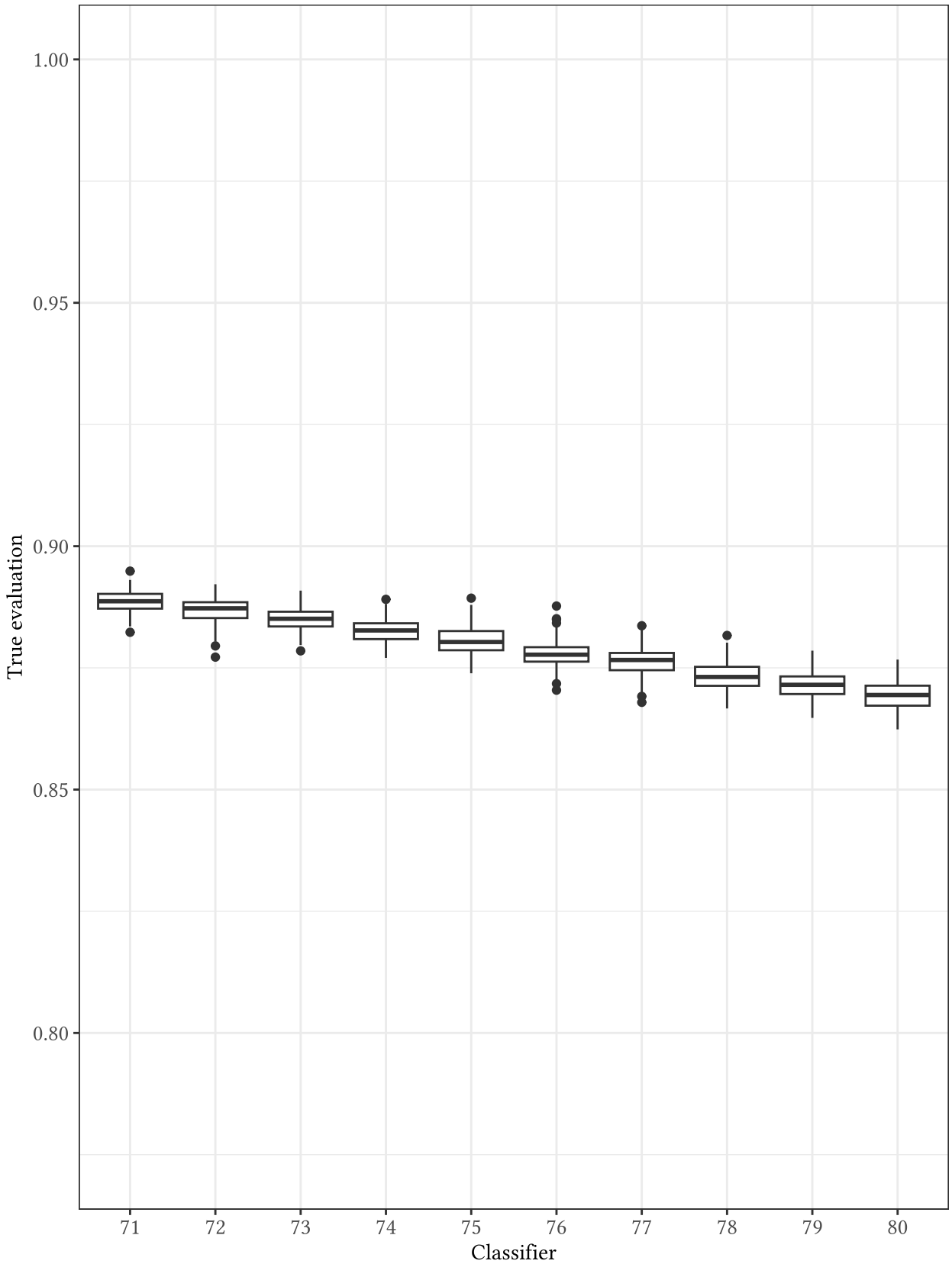


Figure 8: Boxplots of the AUCs of the classifiers $k = 71$ to $k = 80$ throughout the different runs.

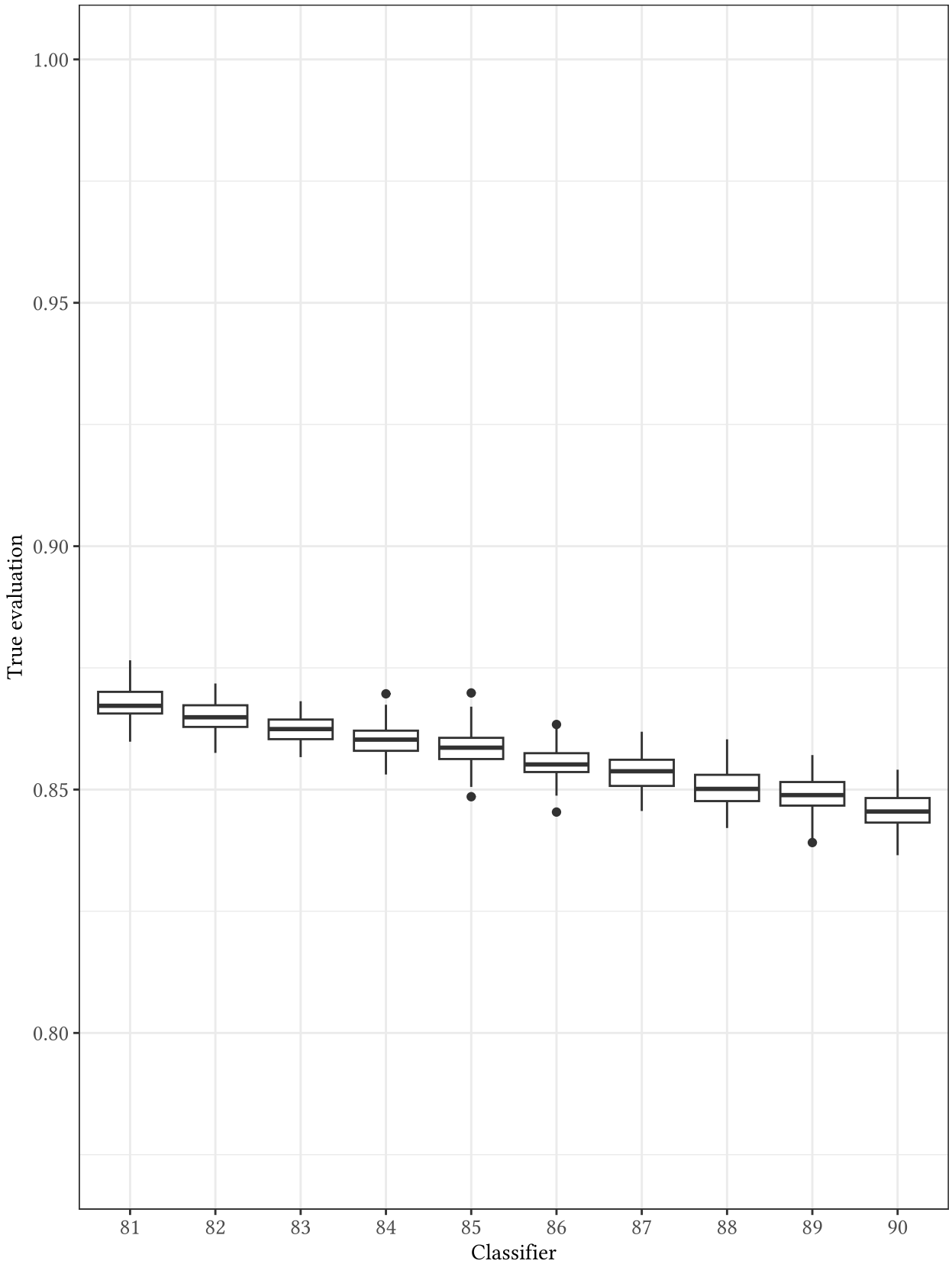


Figure 9: Boxplots of the AUCs of the classifiers $k = 81$ to $k = 90$ throughout the different runs.

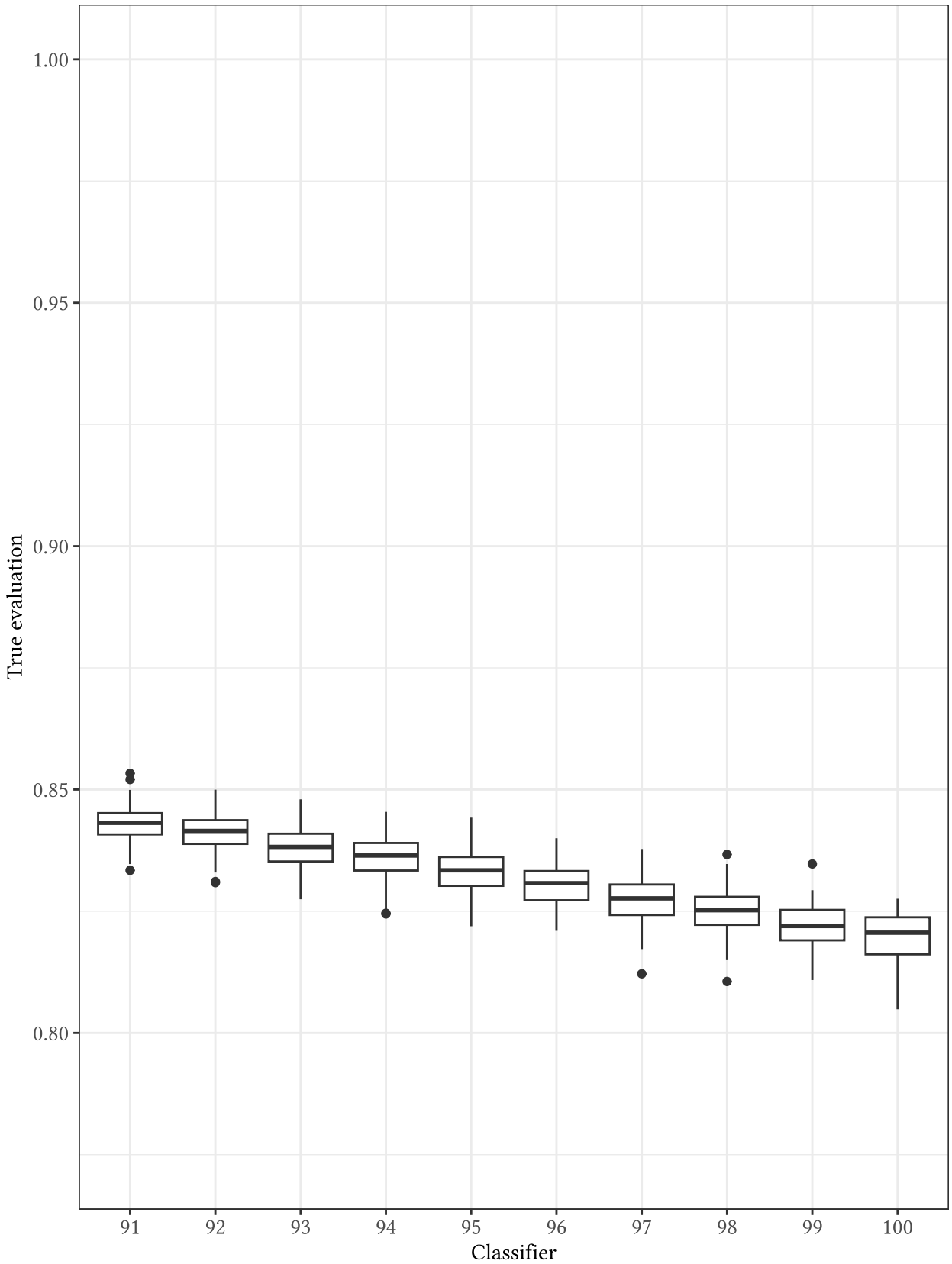


Figure 10: Boxplots of the AUCs of the classifiers $k = 91$ to $k = 100$ throughout the different runs.

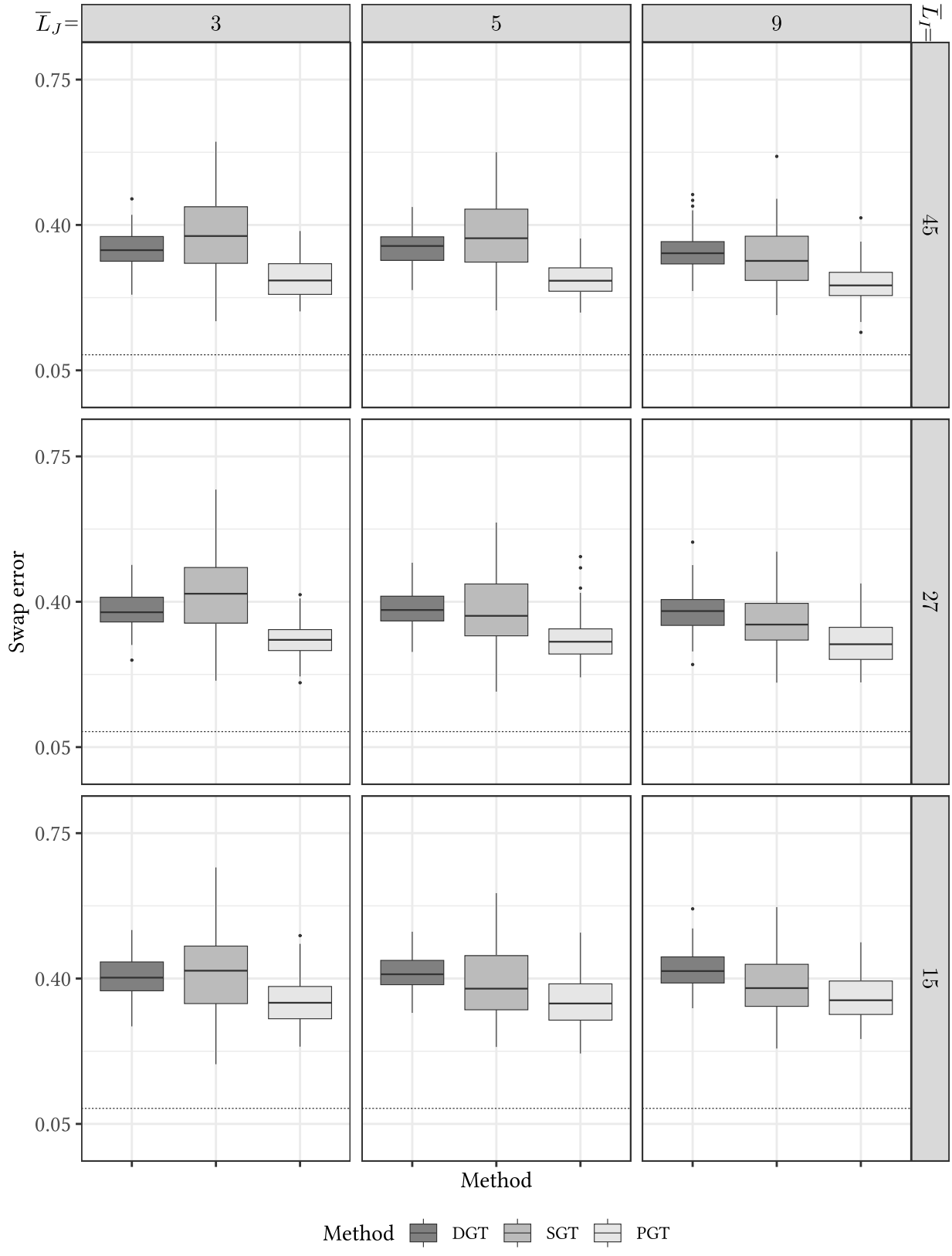


Figure 11: Boxplots for the "extreme" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 15, 27, and 45.

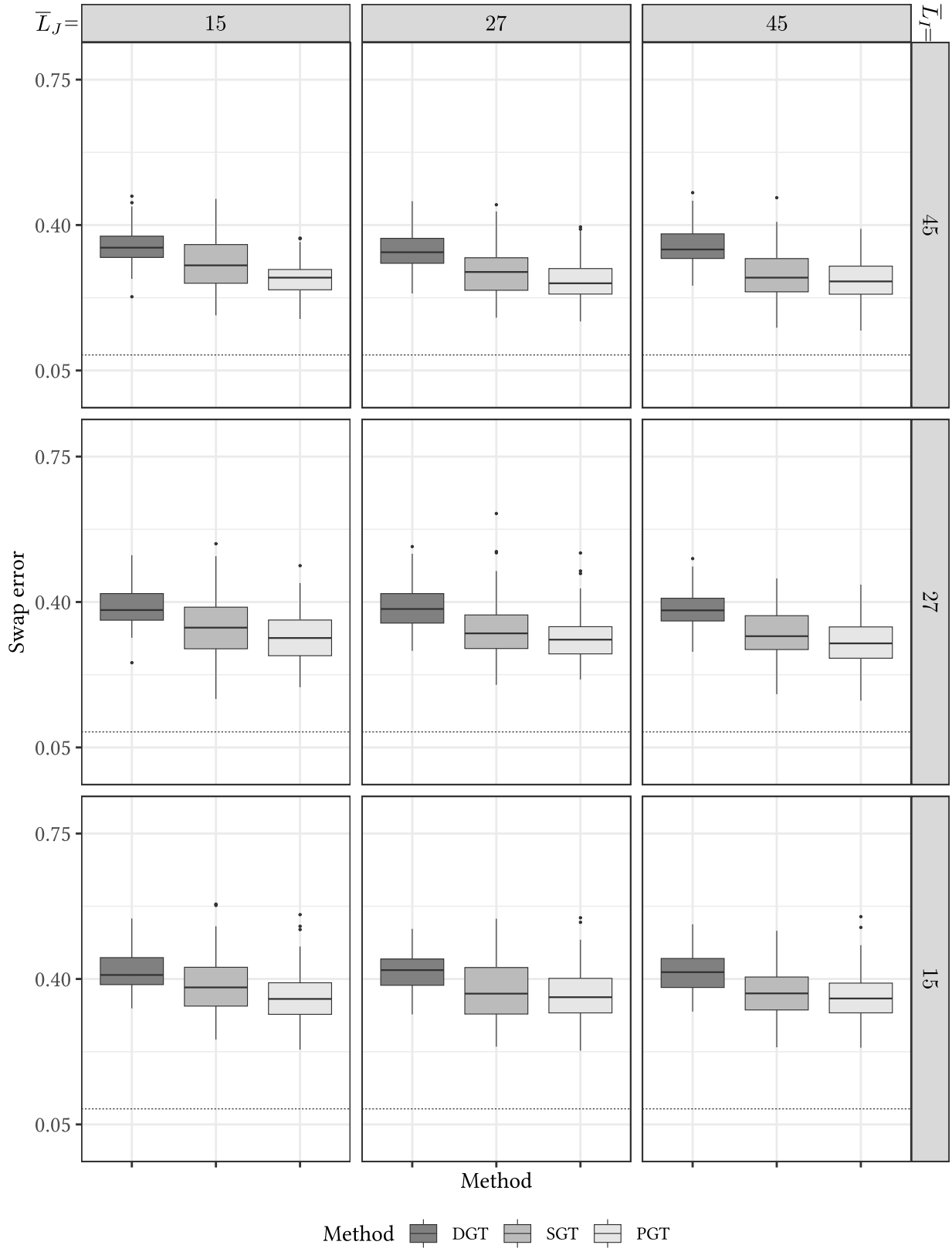


Figure 12: Boxplots for the "extreme" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 15, 27, and 45.

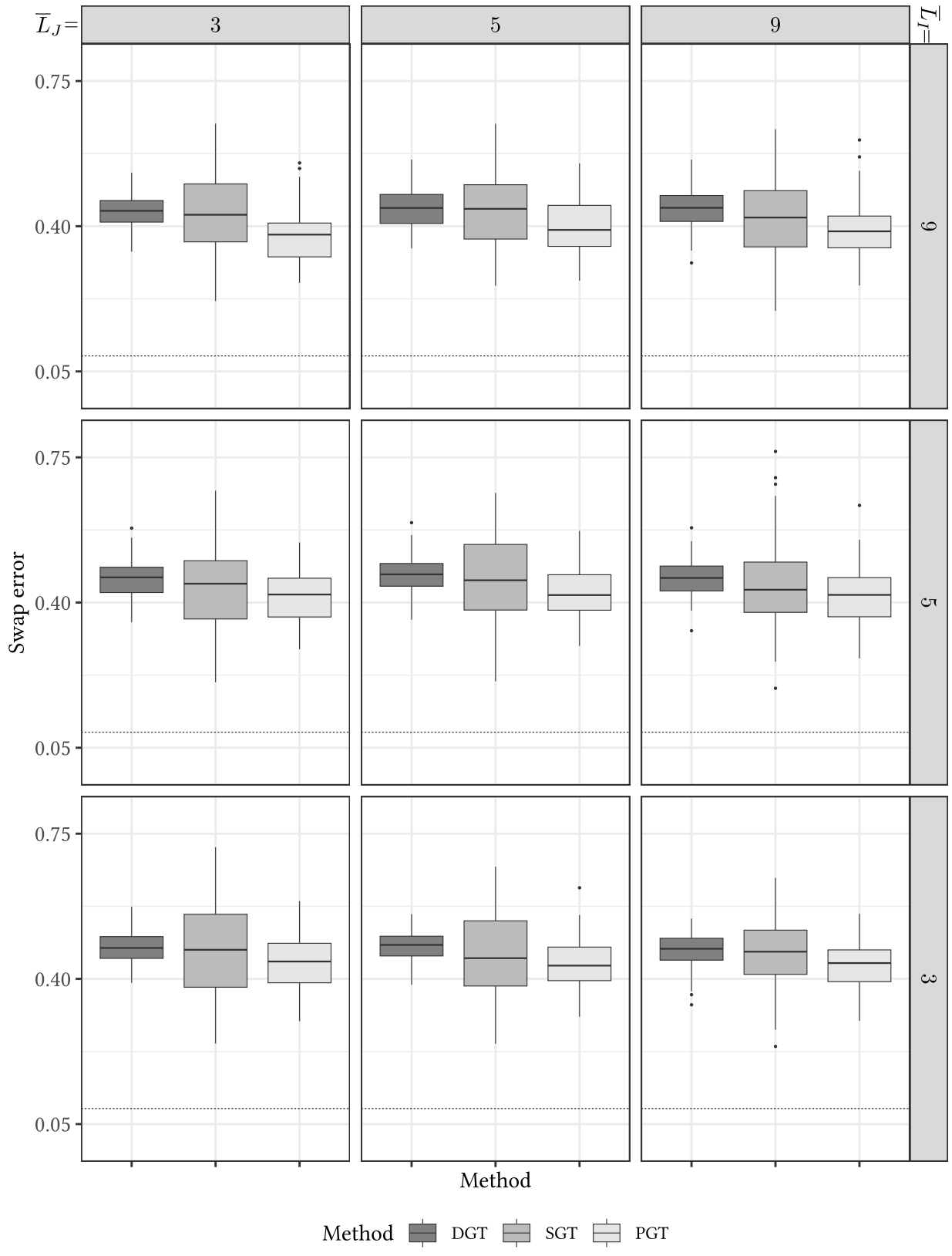


Figure 13: Boxplots for the "extreme" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 3, 5, and 9.

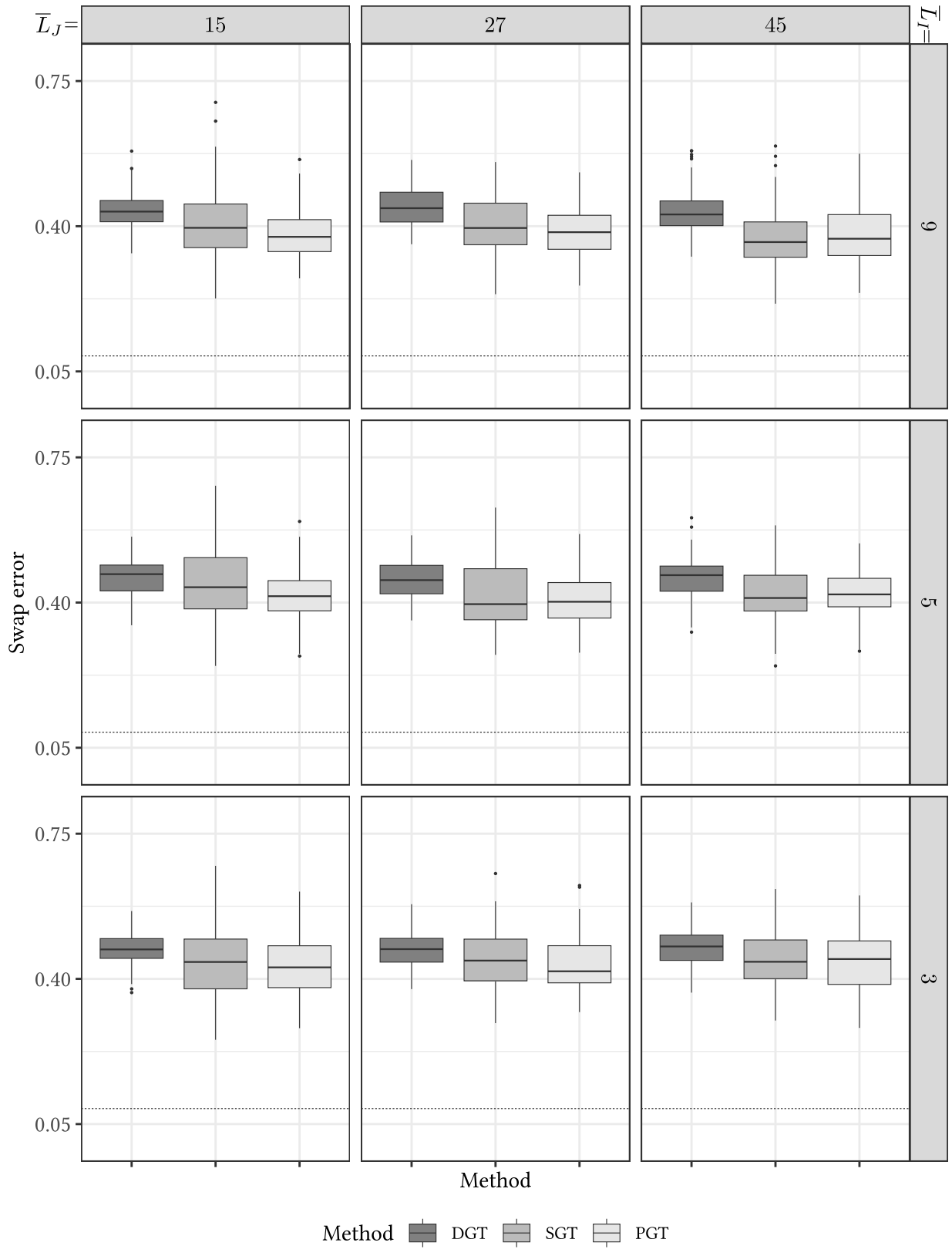


Figure 14: Boxplots for the "extreme" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 3, 5, and 9.

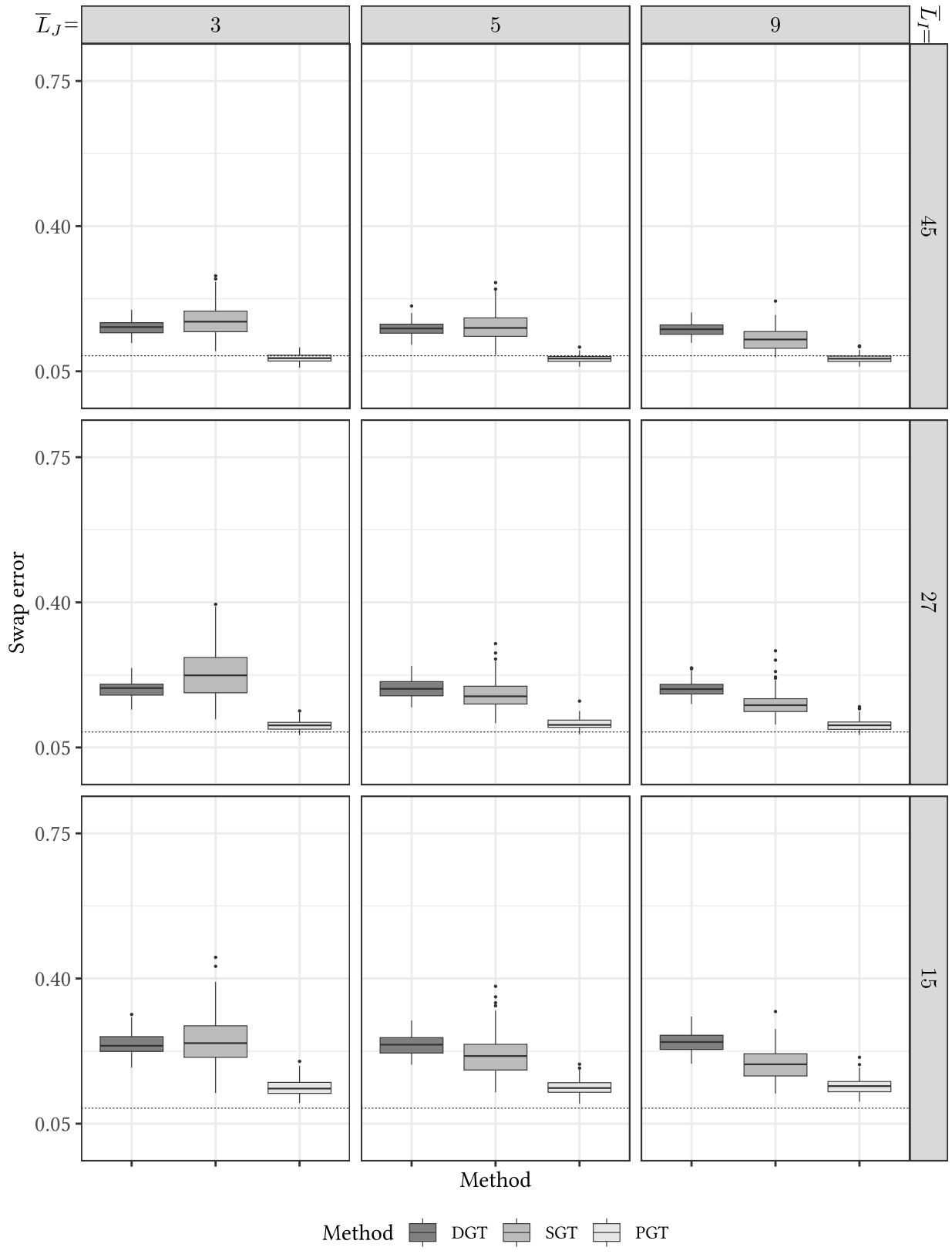


Figure 15: Boxplots for the "bad" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 15, 27, and 45.

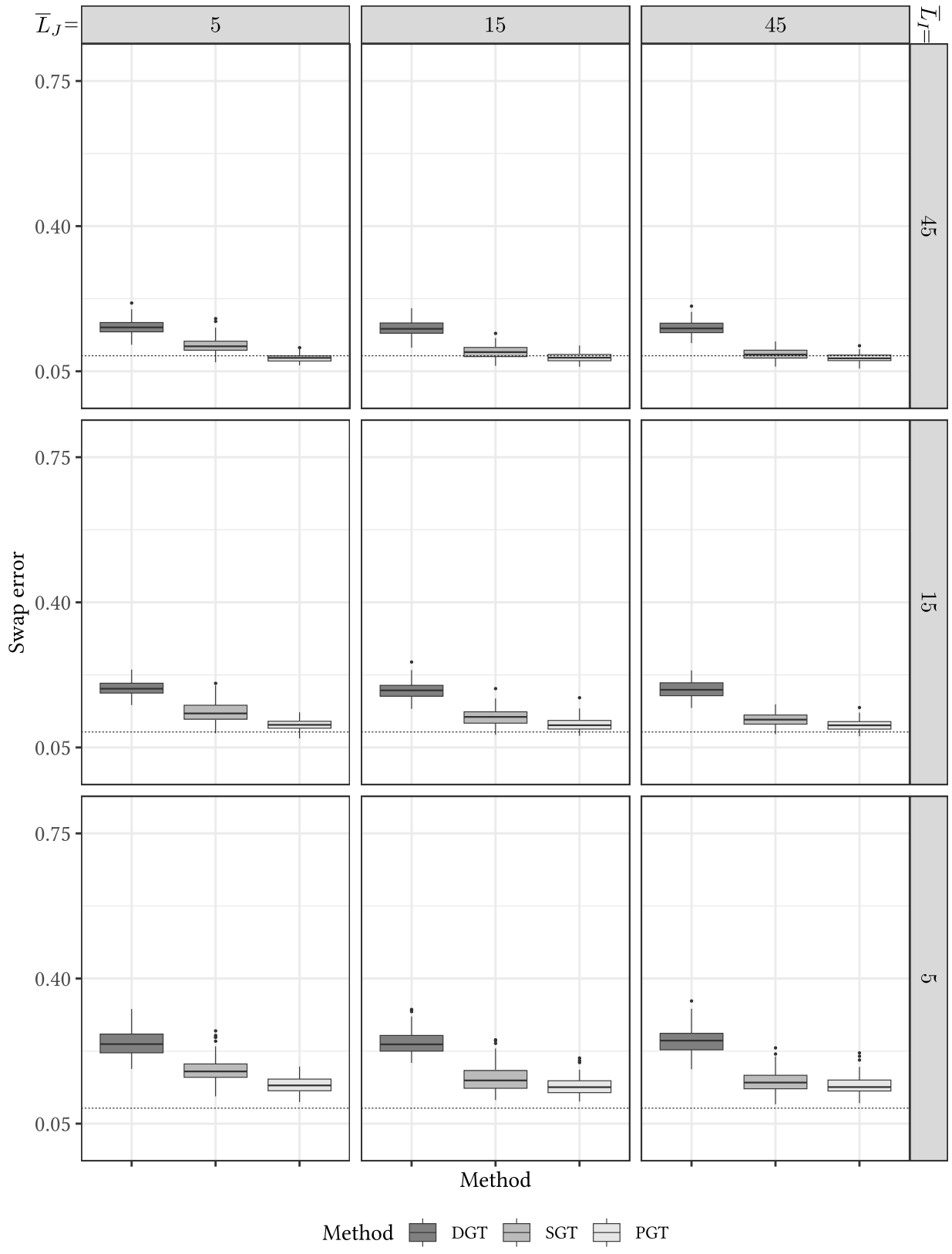


Figure 16: Boxplots for the "bad" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 15, 27, and 45.

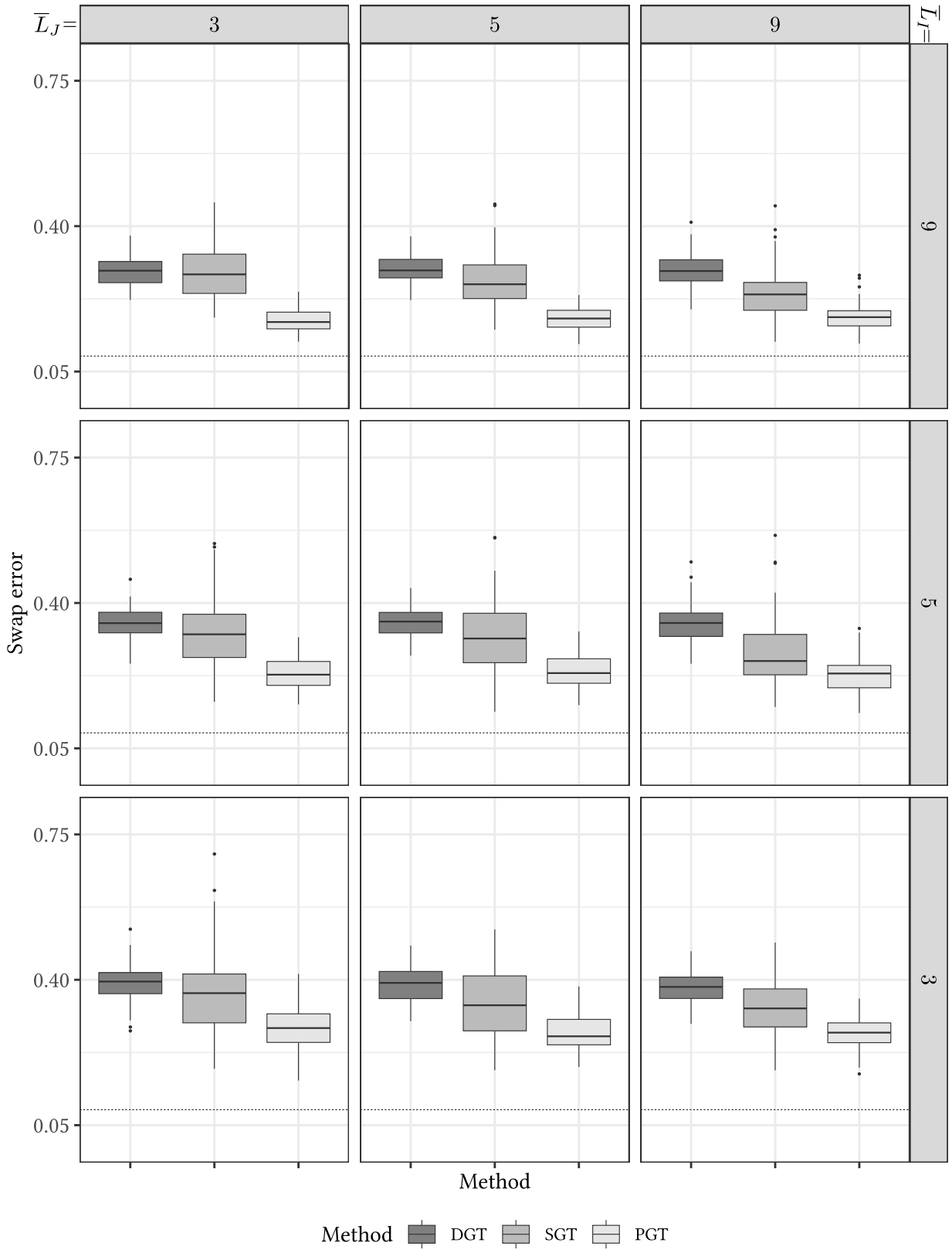


Figure 17: Boxplots for the "bad" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 3, 5, and 9.

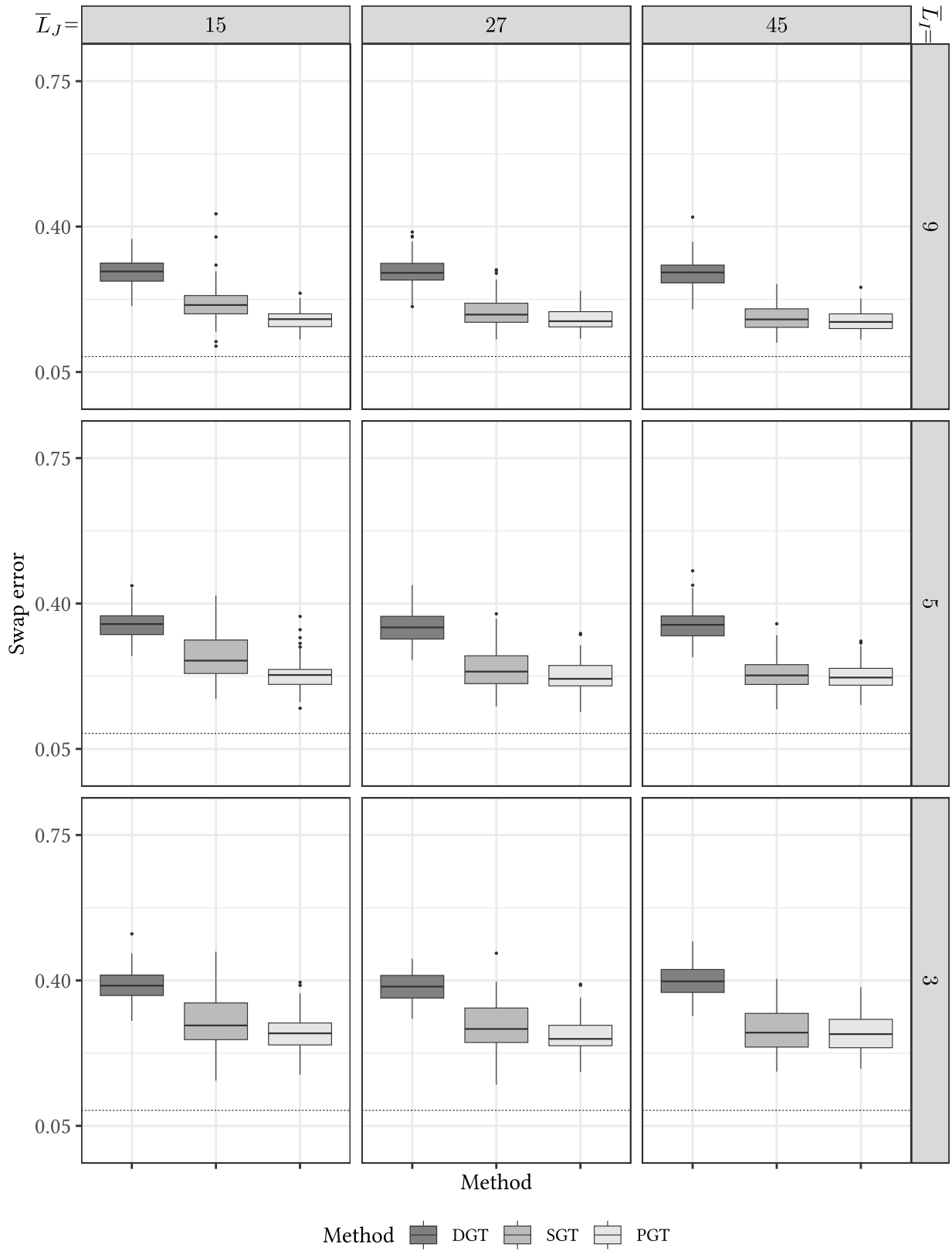


Figure 18: Boxplots for the "bad" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 3, 5, and 9.

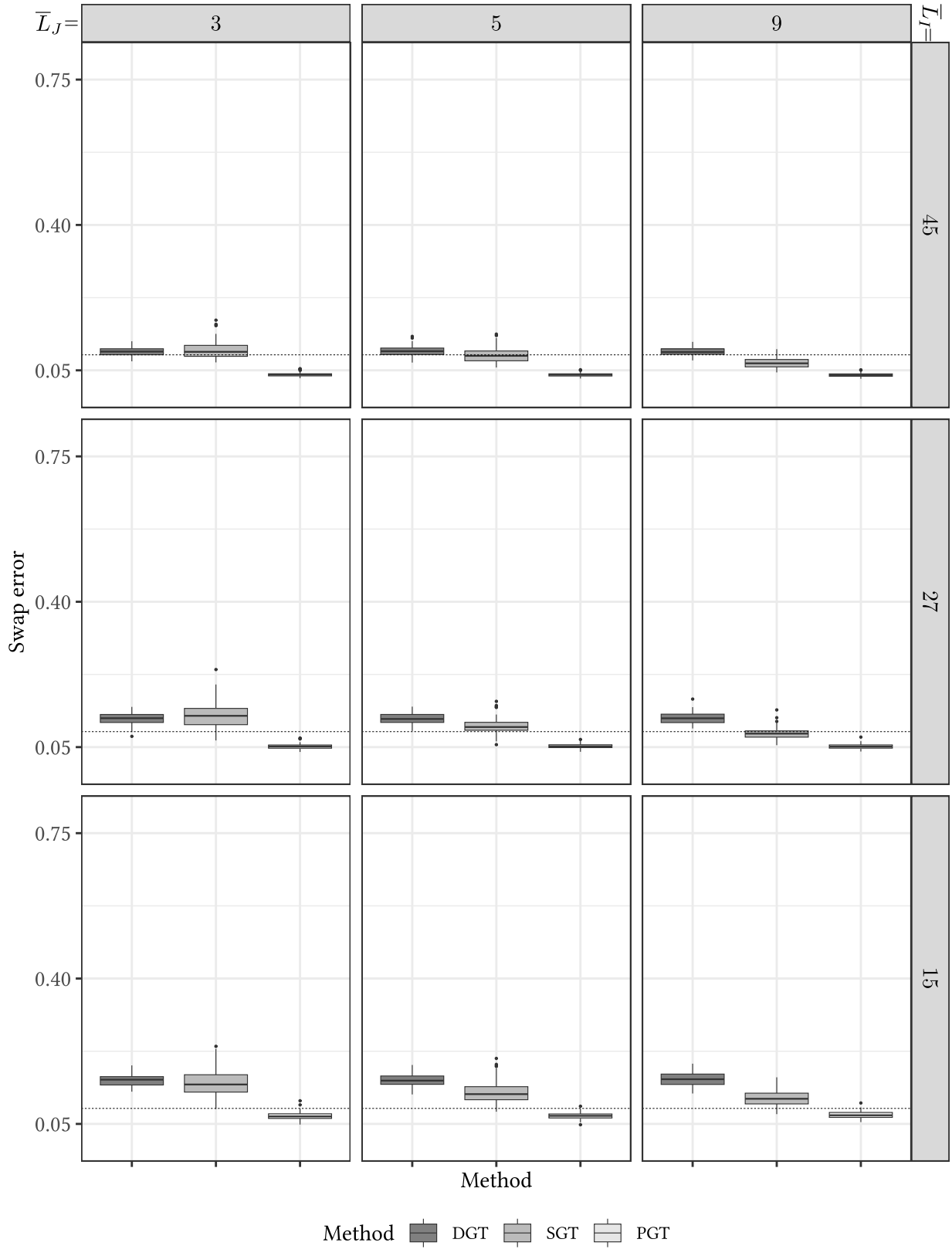


Figure 19: Boxplots for the "average" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 15, 27, and 45.

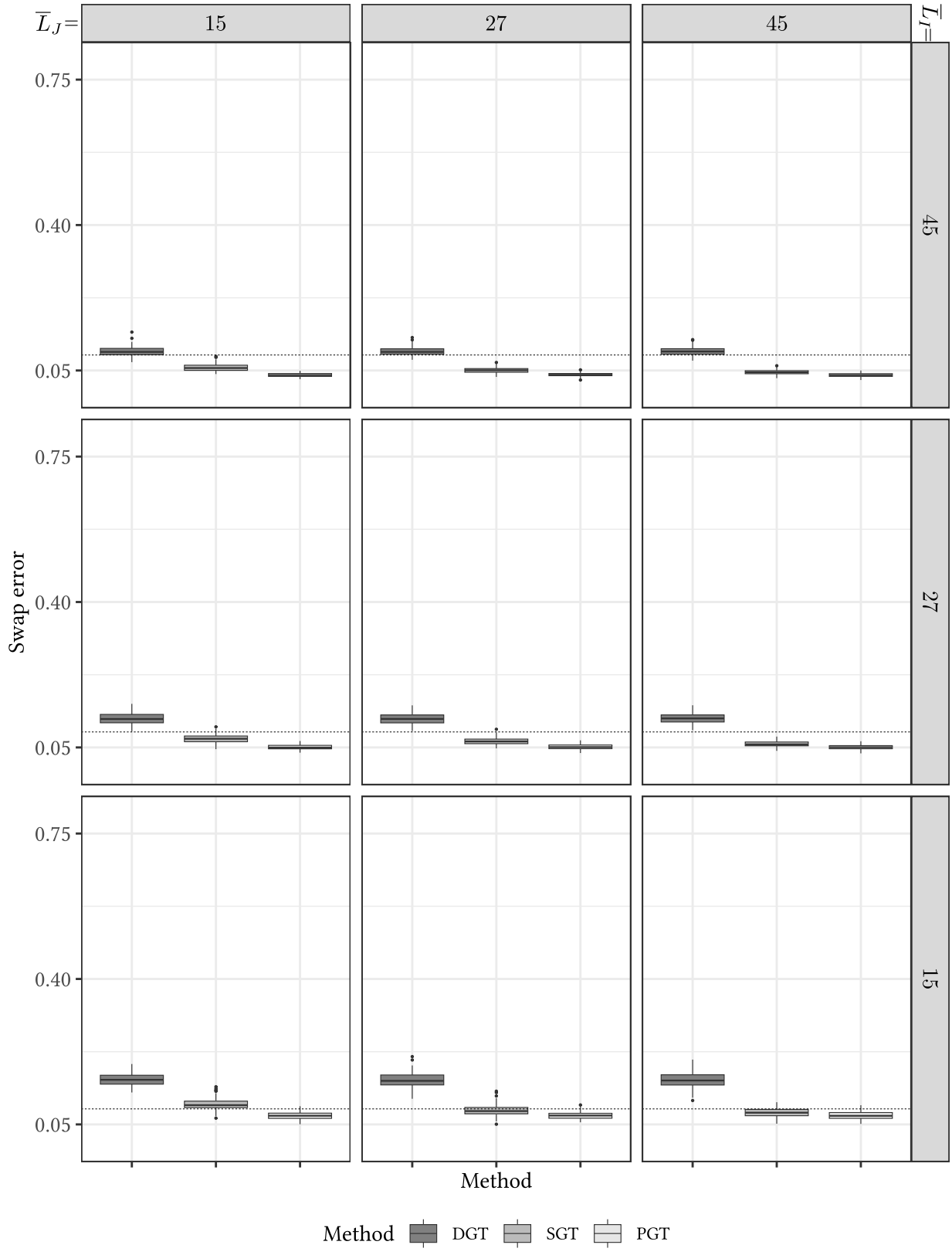


Figure 20: Boxplots for the "average" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 15, 27, and 45.

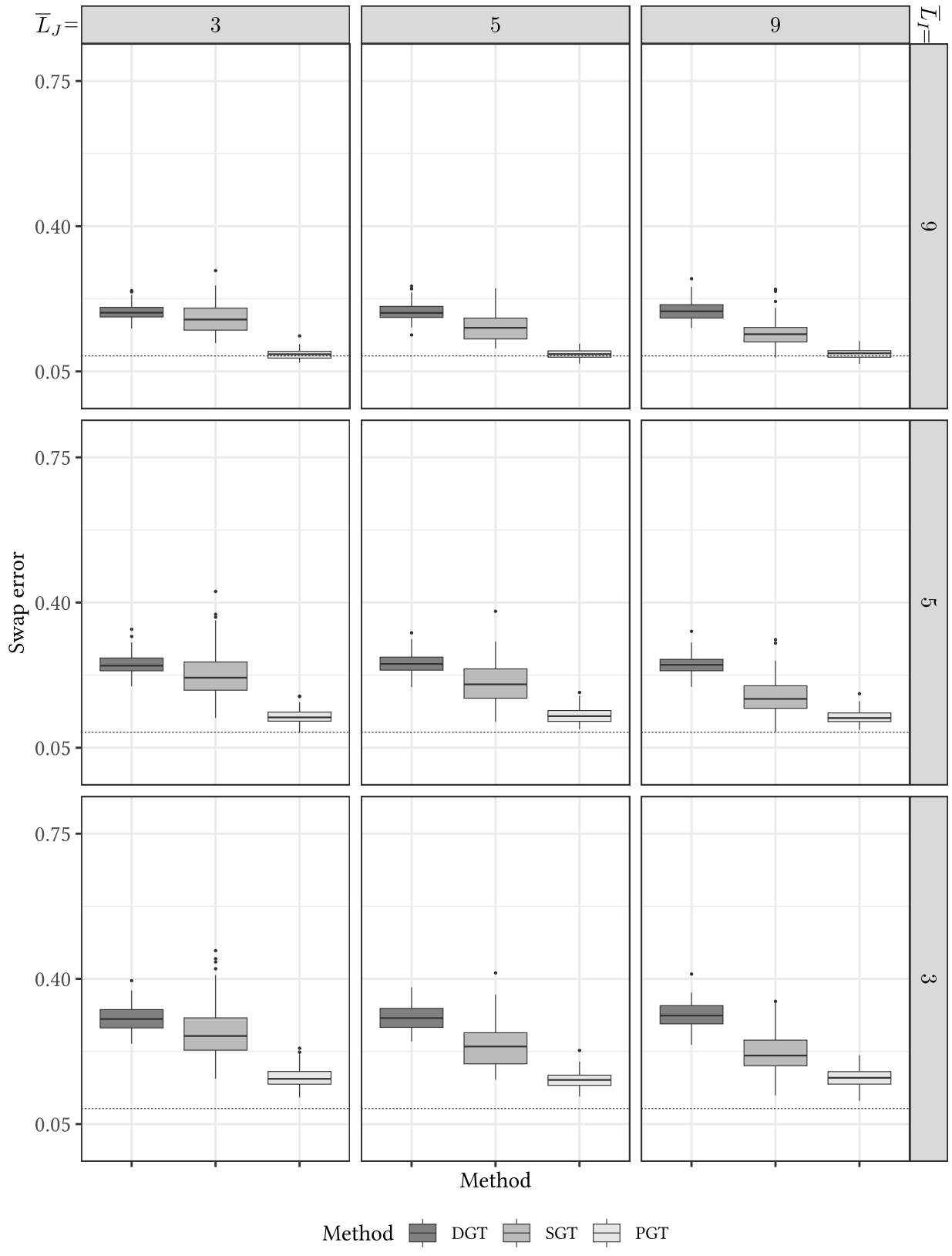


Figure 21: Boxplots for the "average" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 3, 5, and 9.

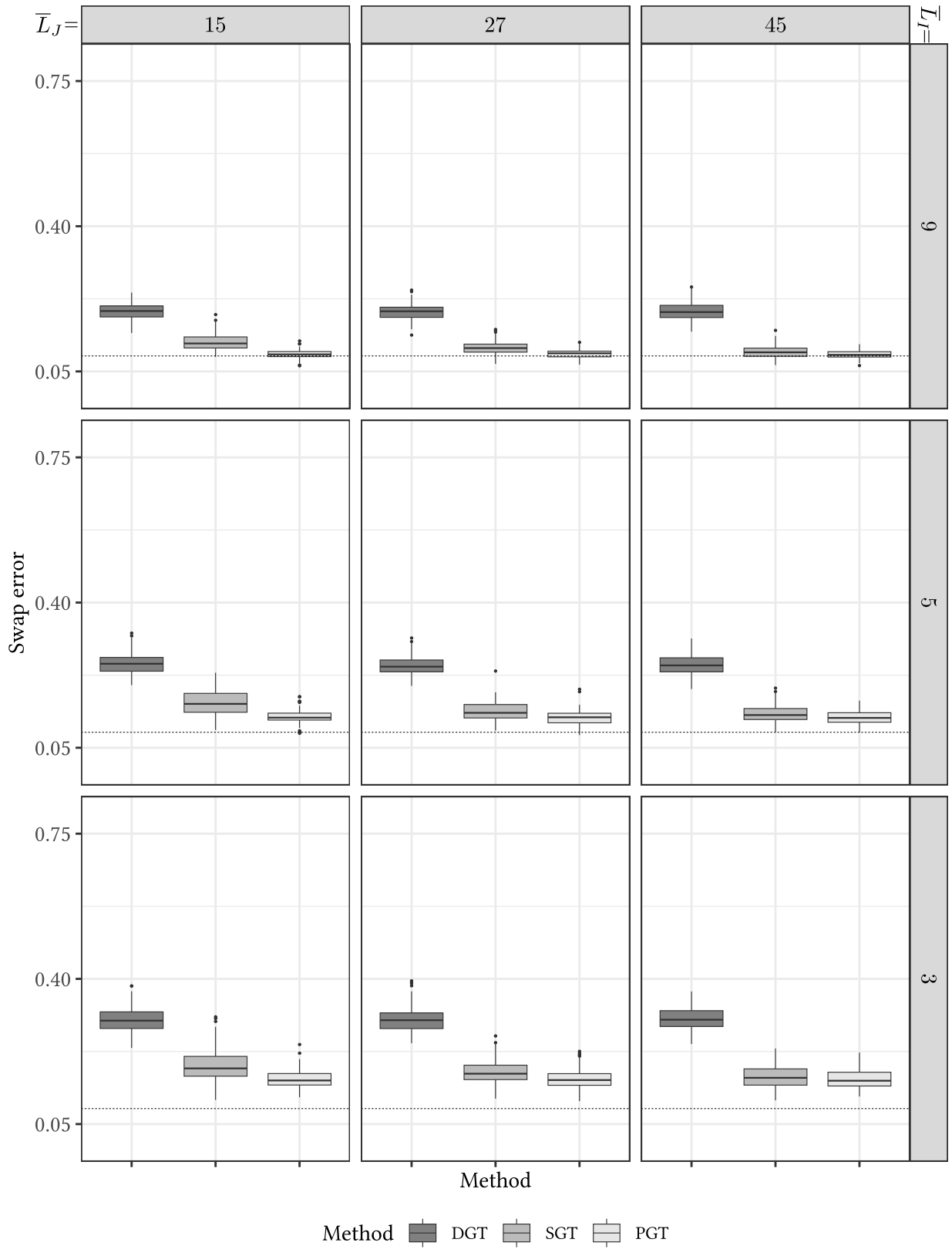


Figure 22: Boxplots for the "average" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 3, 5, and 9.

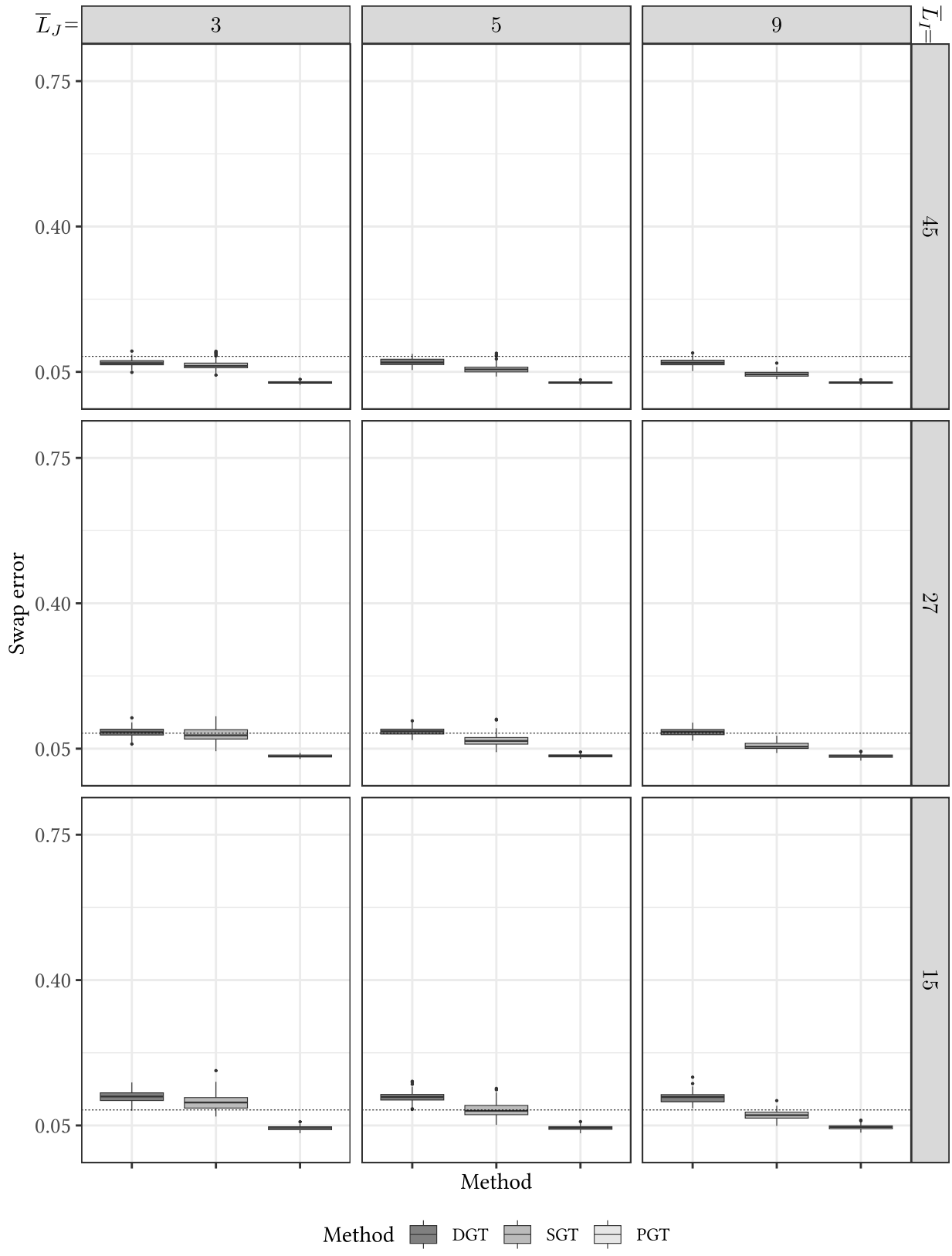


Figure 23: Boxplots for the "good" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 15, 27, and 45.

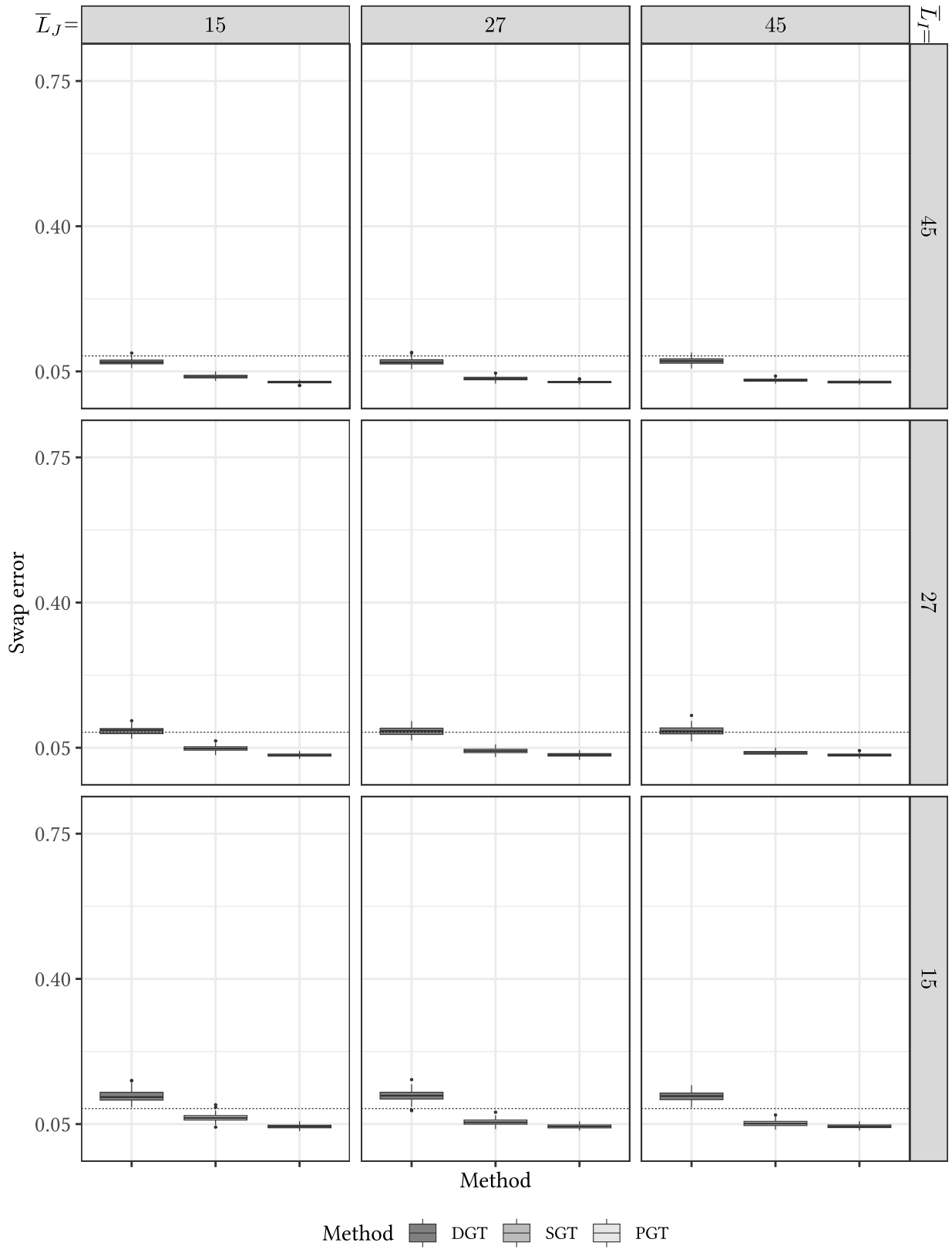


Figure 24: Boxplots for the "good" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 15, 27, and 45.

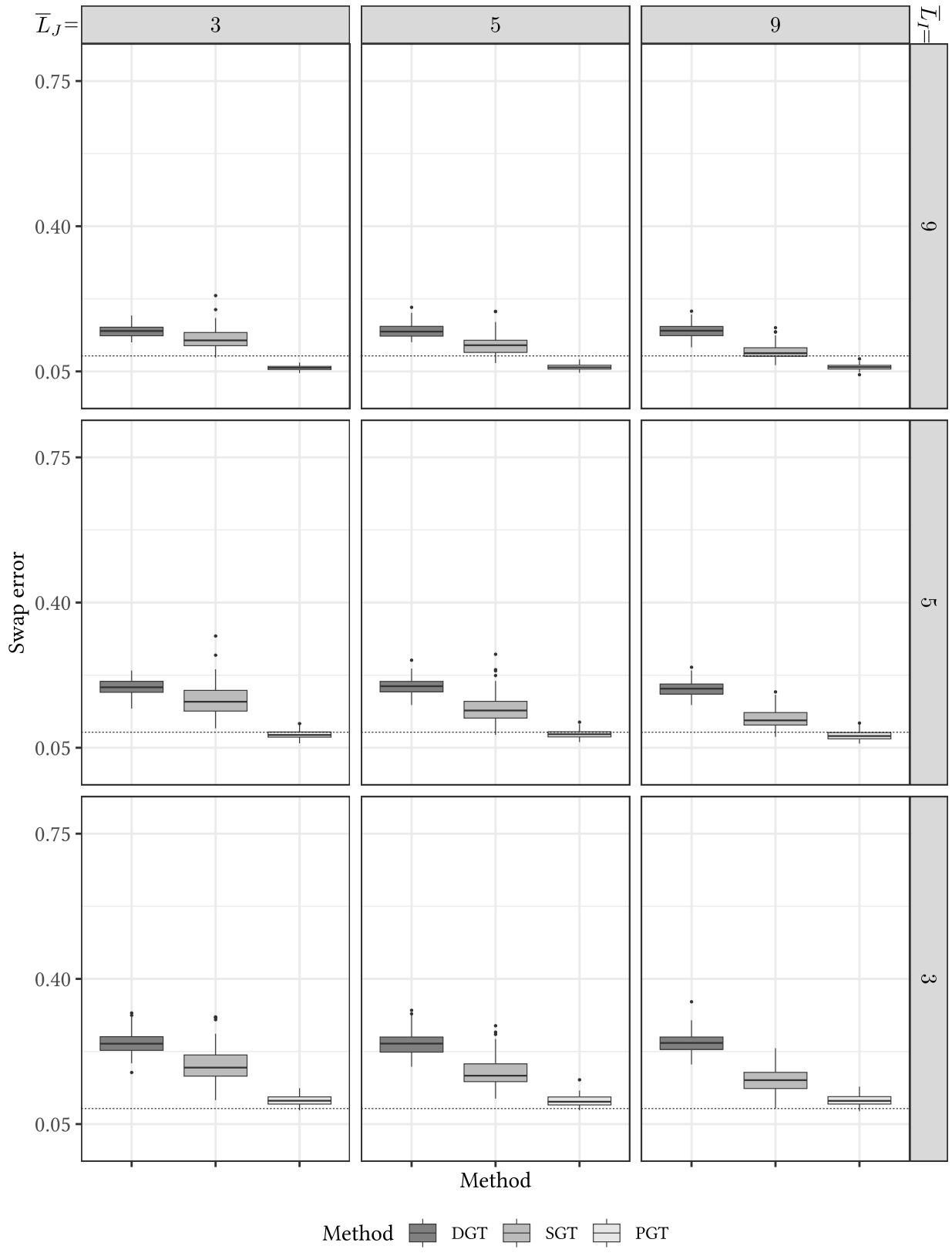


Figure 25: Boxplots for the "good" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 3, 5, and 9.

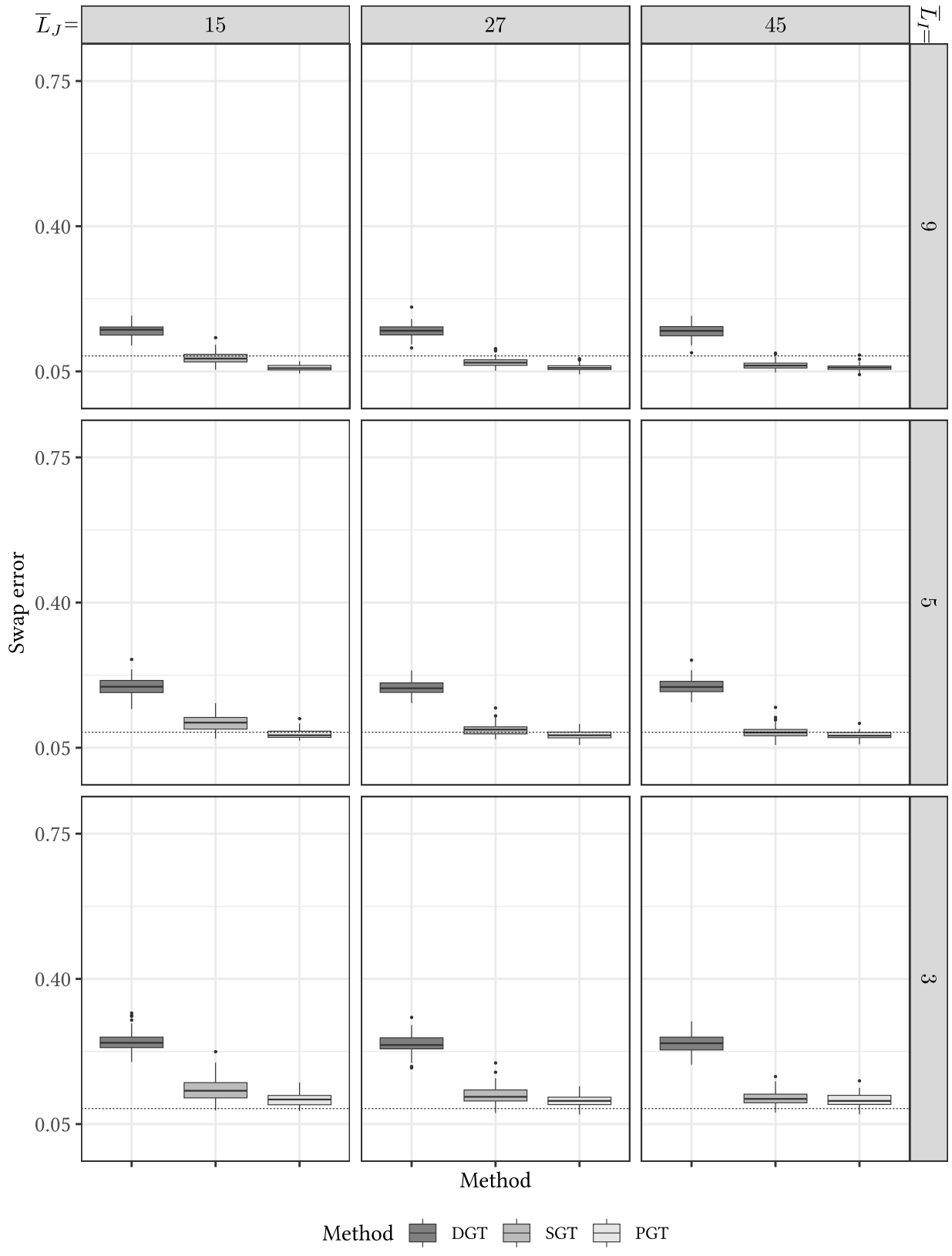


Figure 26: Boxplots for the "good" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 3, 5, and 9.

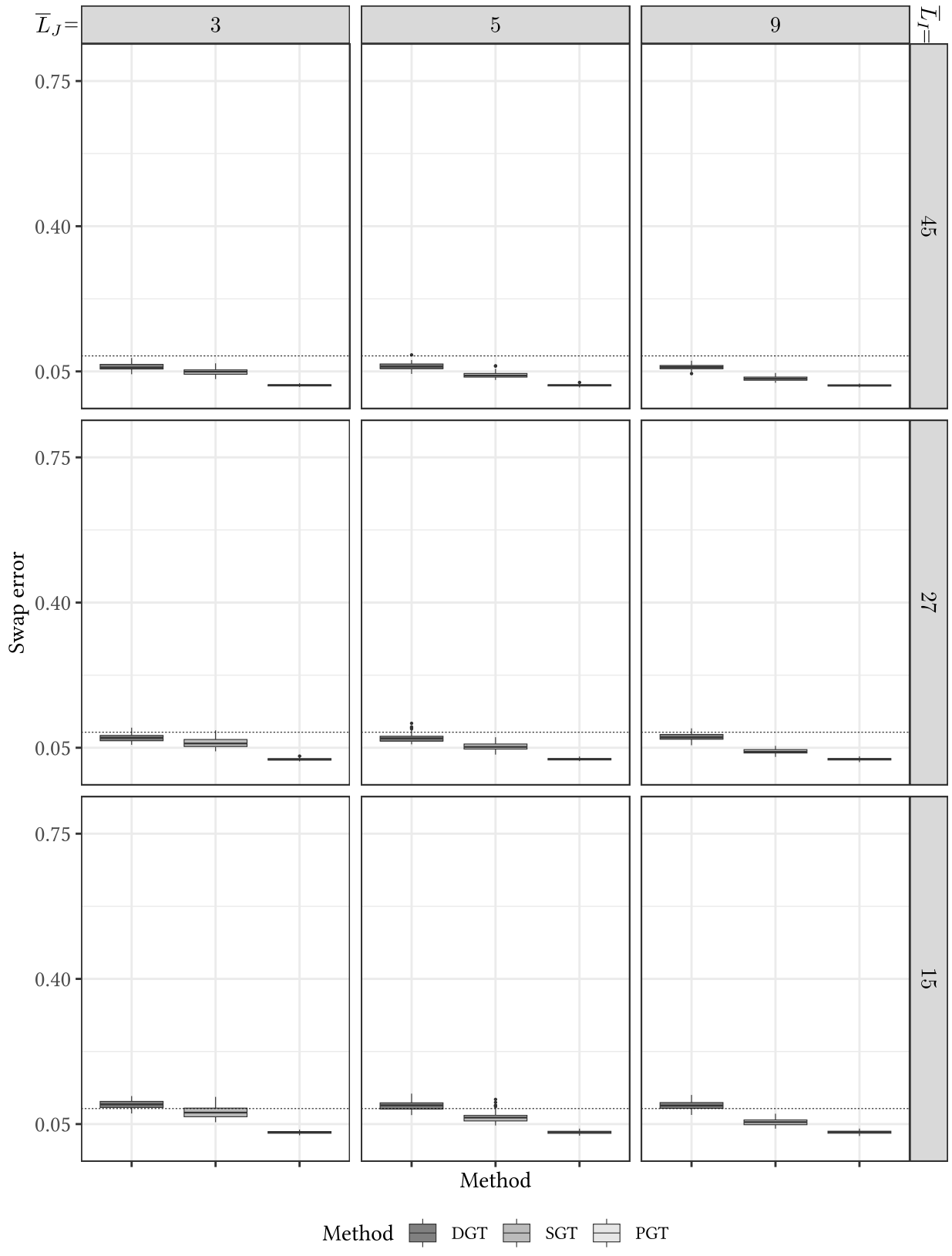


Figure 27: Boxplots for the "outstanding" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 15, 27, and 45.

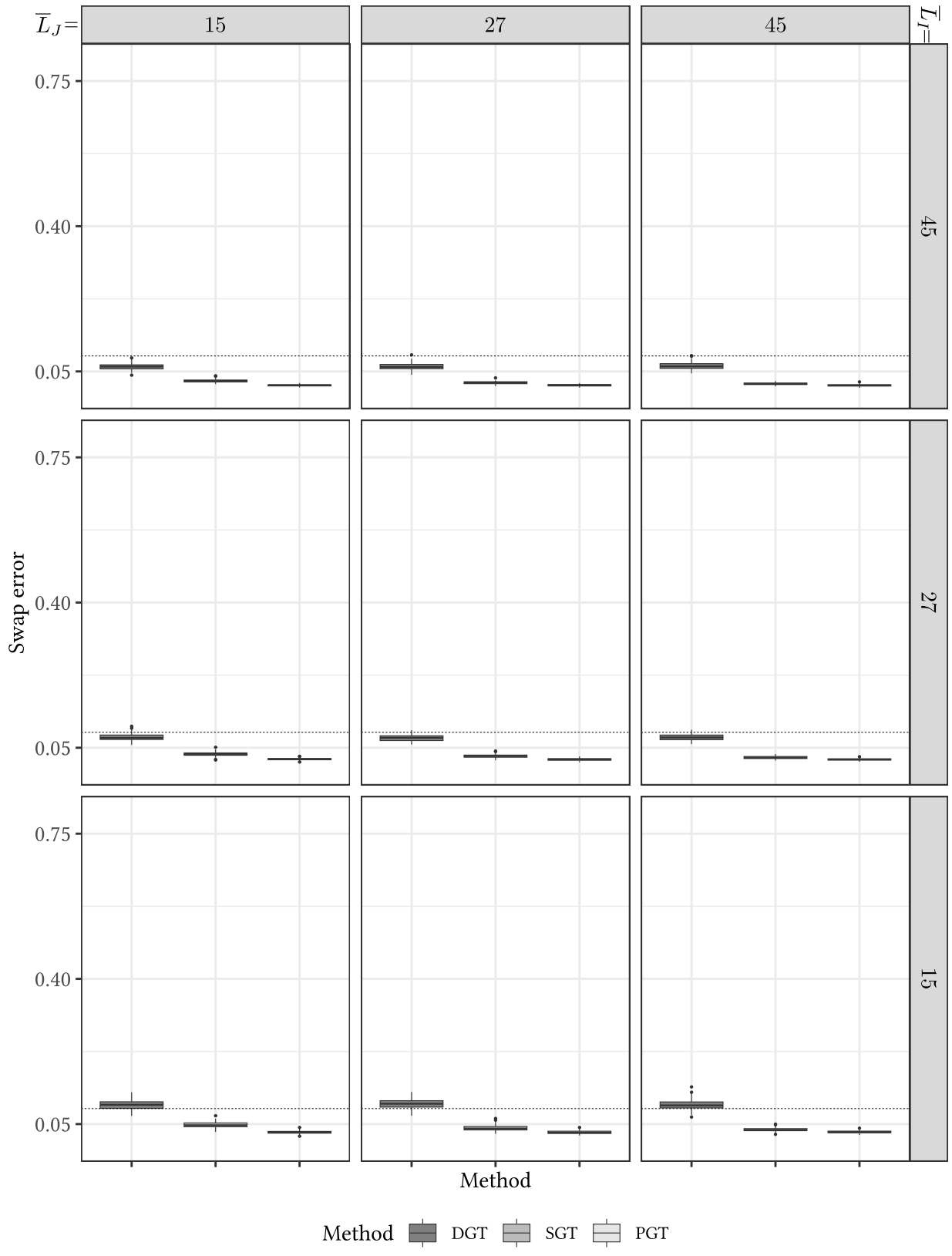


Figure 28: Boxplots for the "outstanding" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 15, 27, and 45.

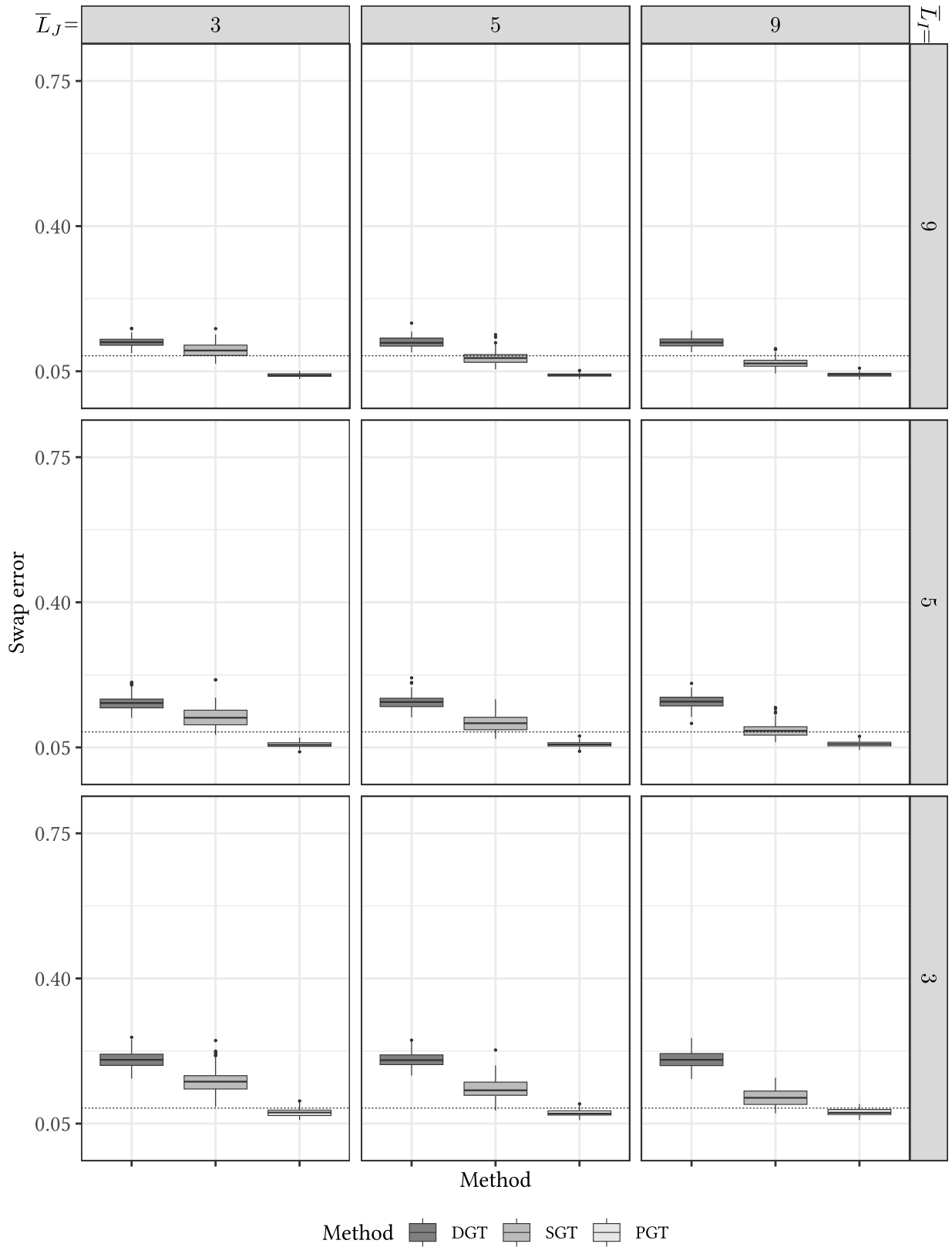


Figure 29: Boxplots for the "outstanding" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 3, 5, and 9.

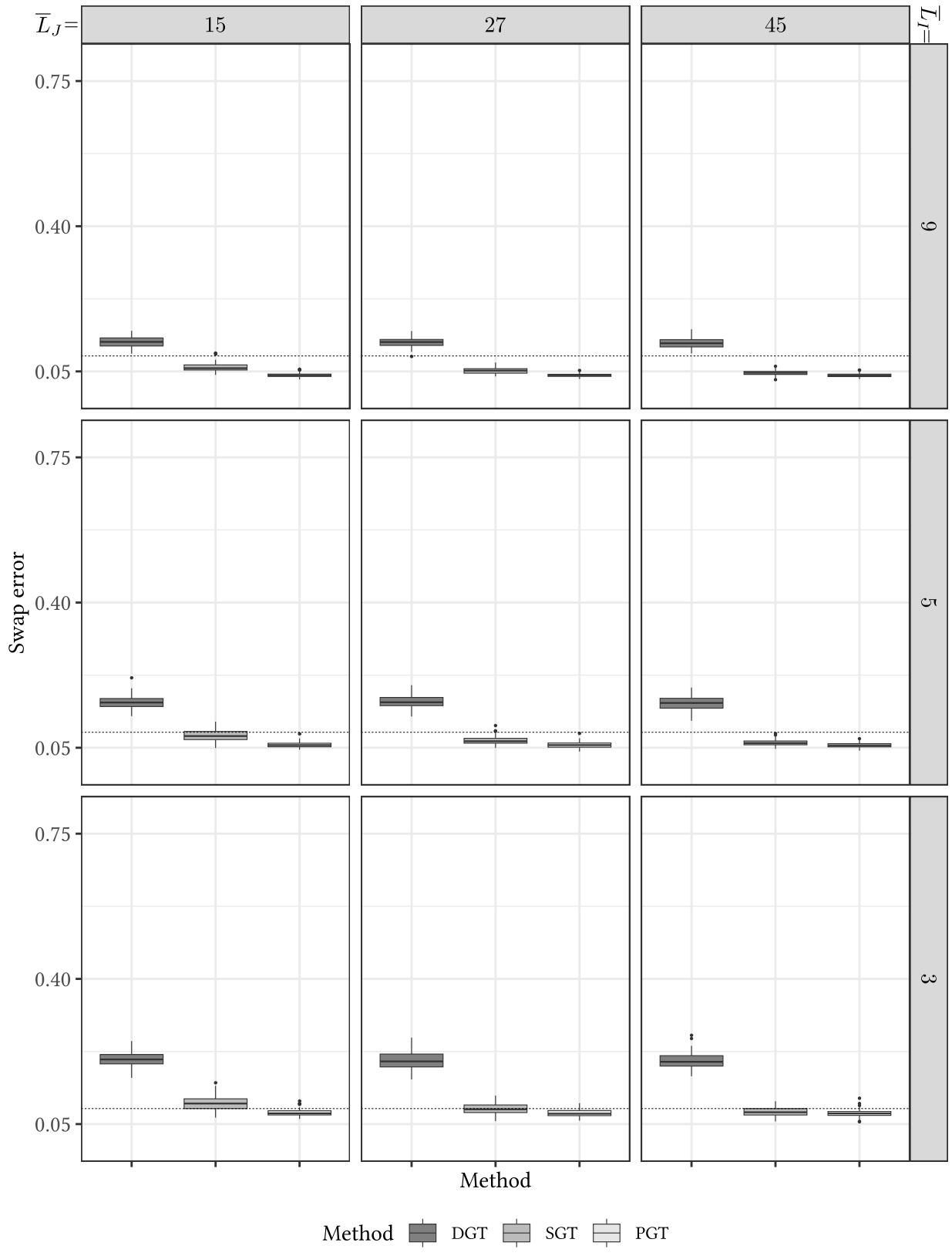


Figure 30: Boxplots for the "outstanding" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 3, 5, and 9.

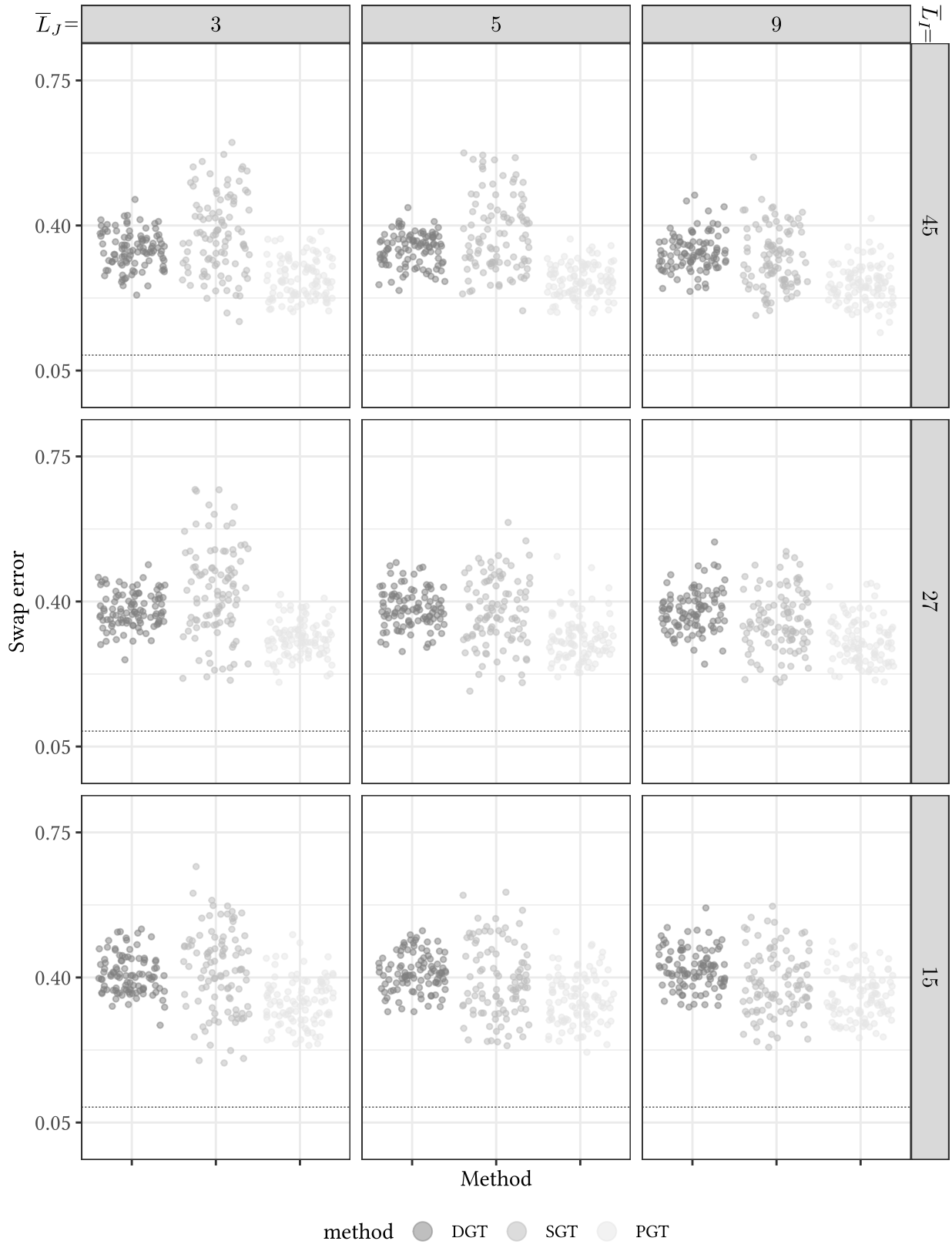


Figure 31: Scatter plots for the "extreme" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 15, 27, and 45.

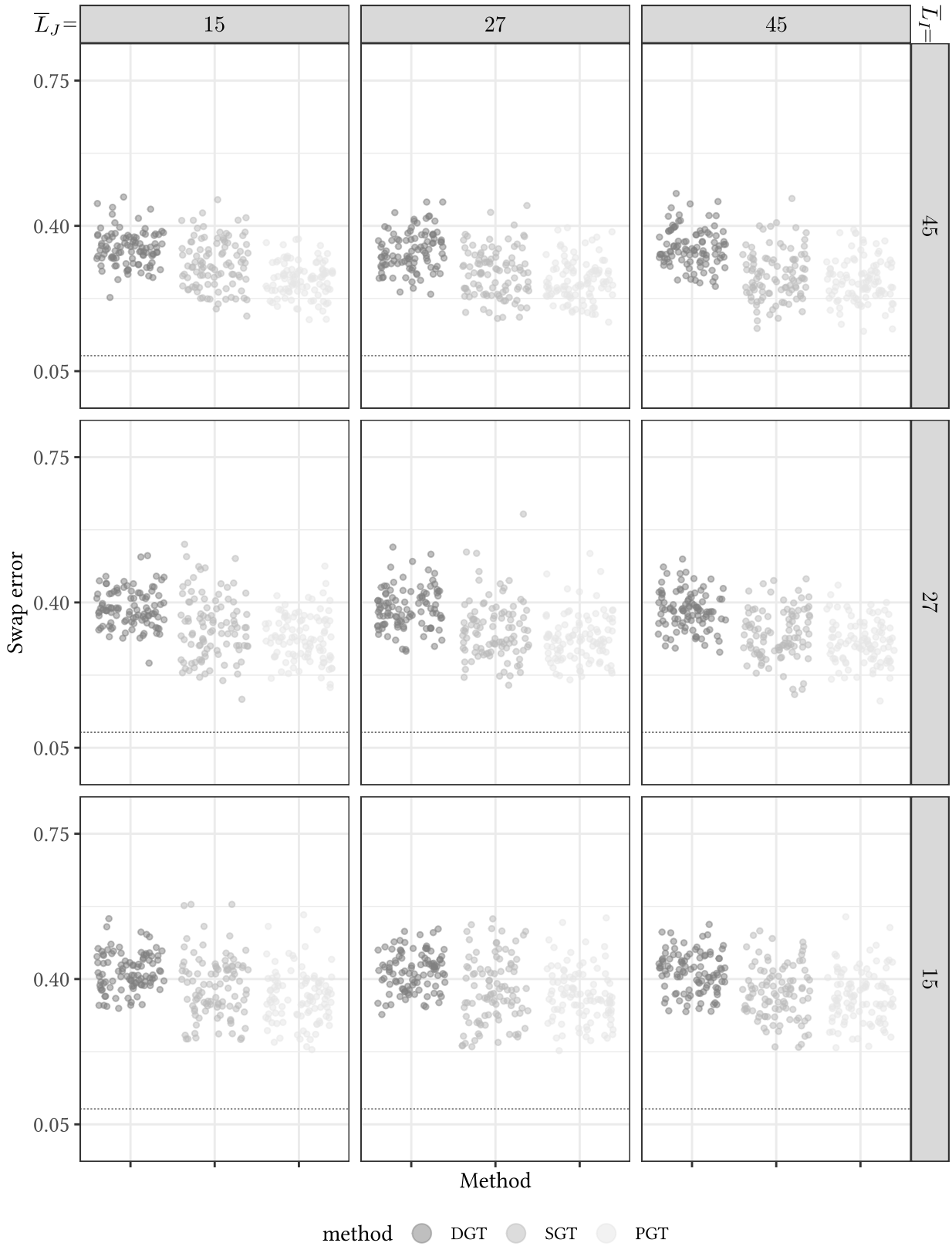


Figure 32: Scatter plots for the "extreme" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 15, 27, and 45.

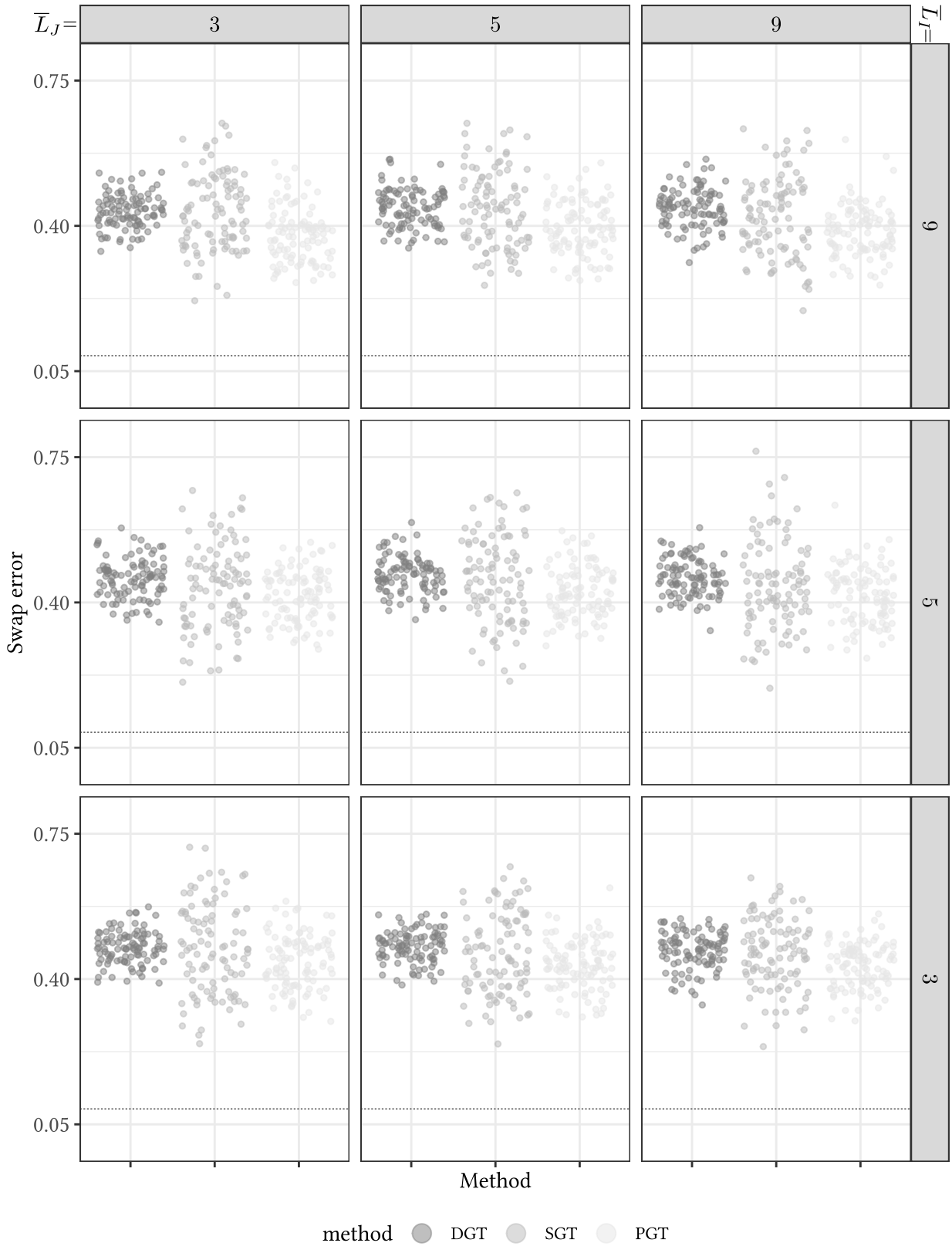


Figure 33: Scatter plots for the "extreme" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 3, 5, and 9.

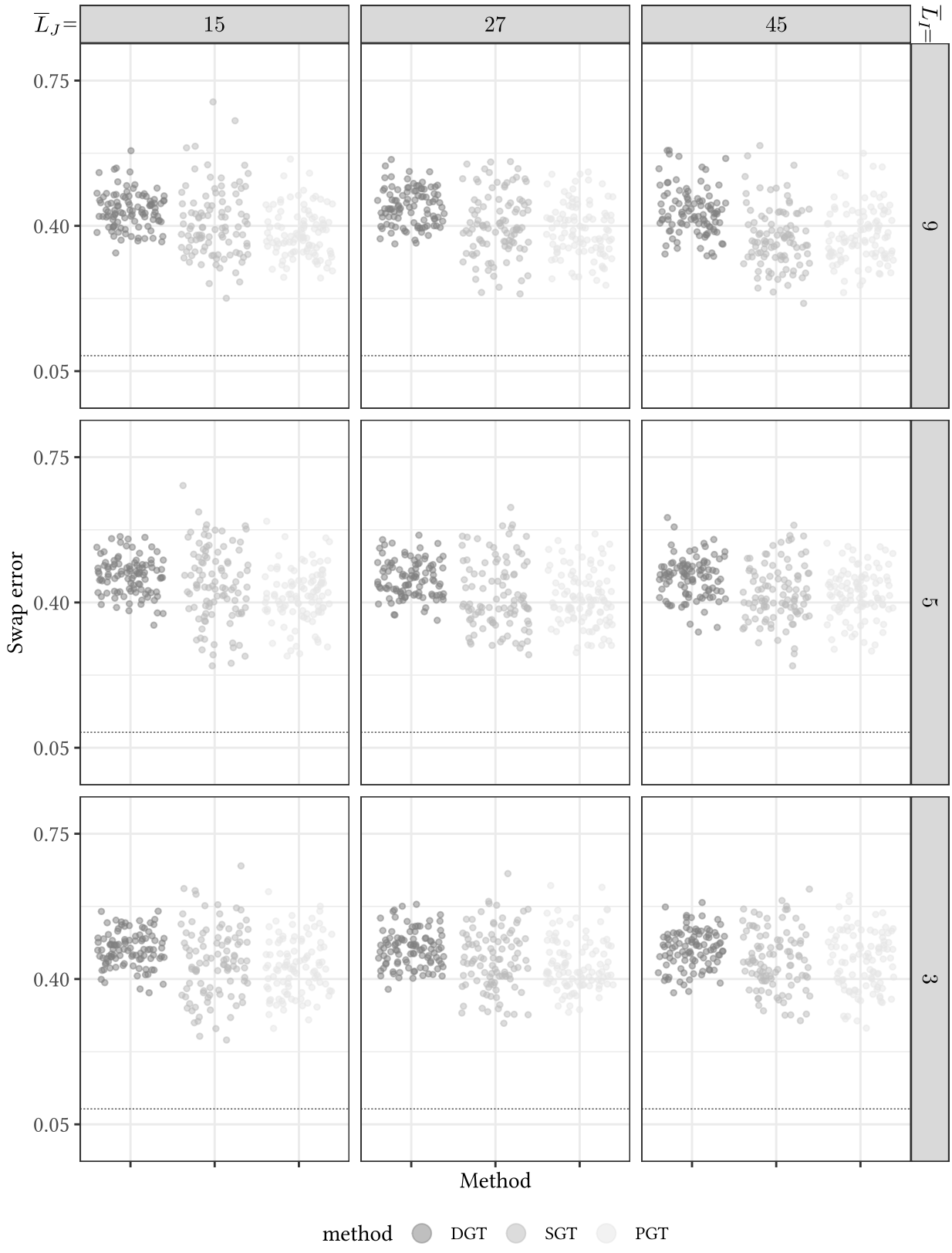


Figure 34: Scatter plots for the "extreme" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 3, 5, and 9.

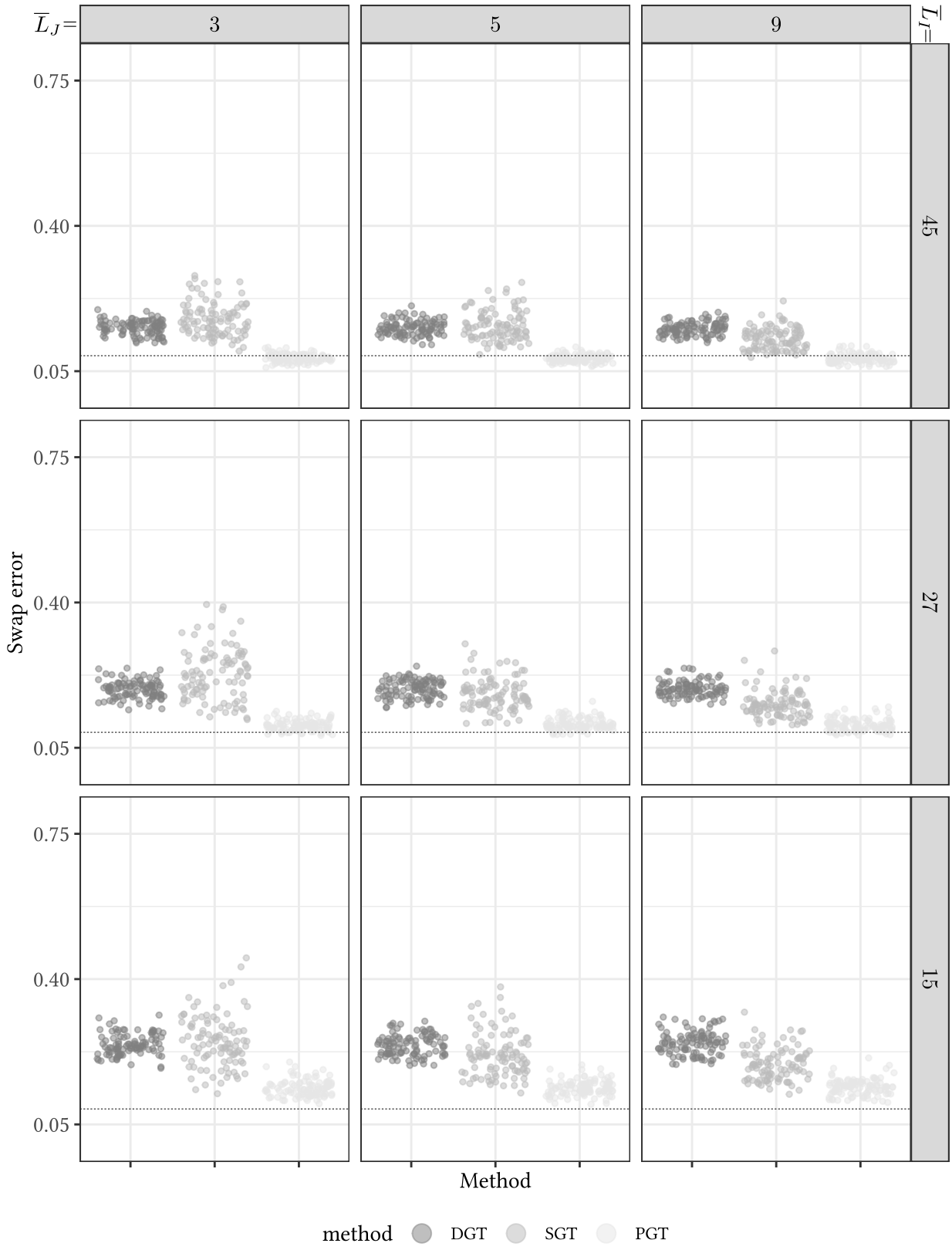


Figure 35: Scatter plots for the "bad" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 15, 27, and 45.

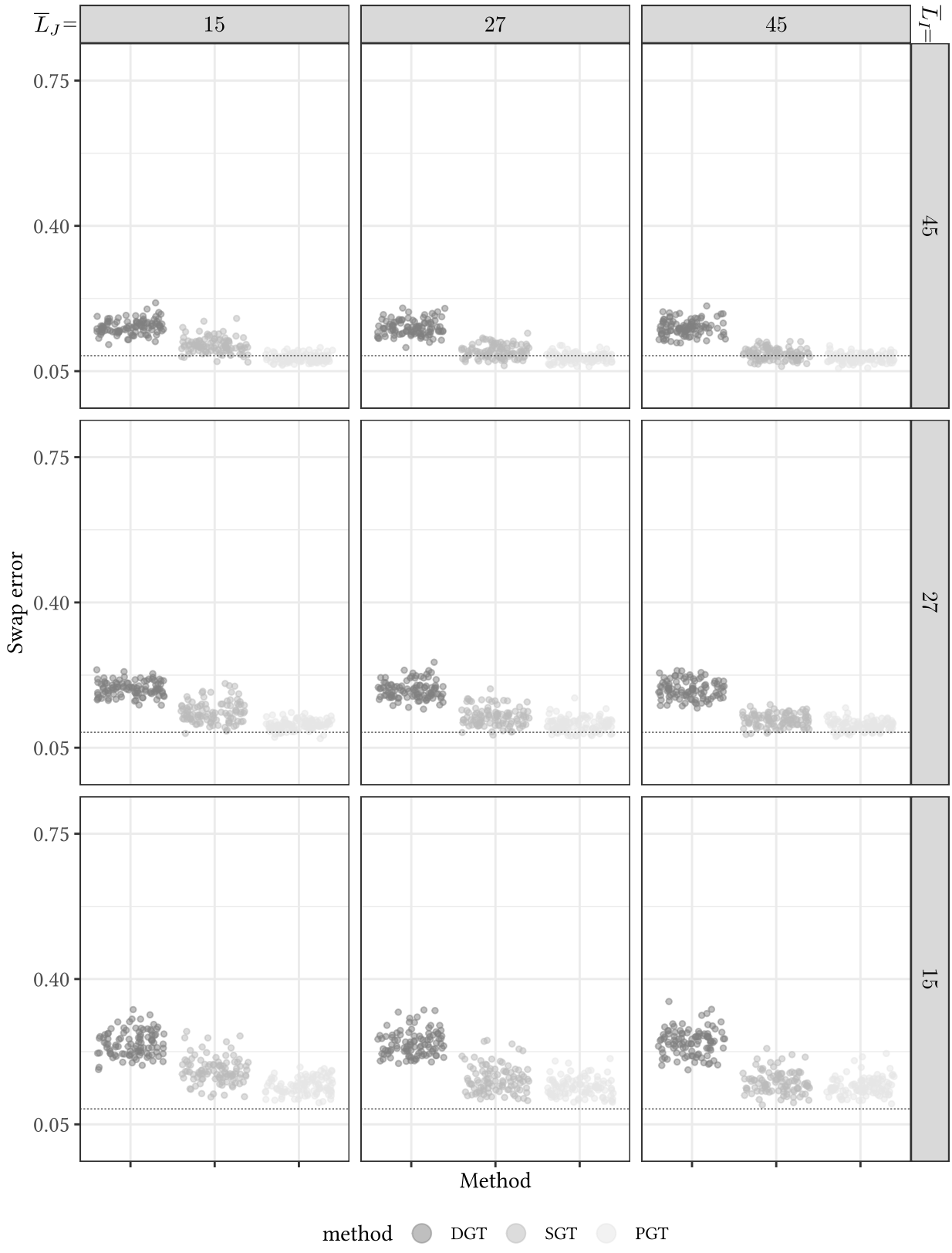


Figure 36: Scatter plots for the "bad" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 15, 27, and 45.

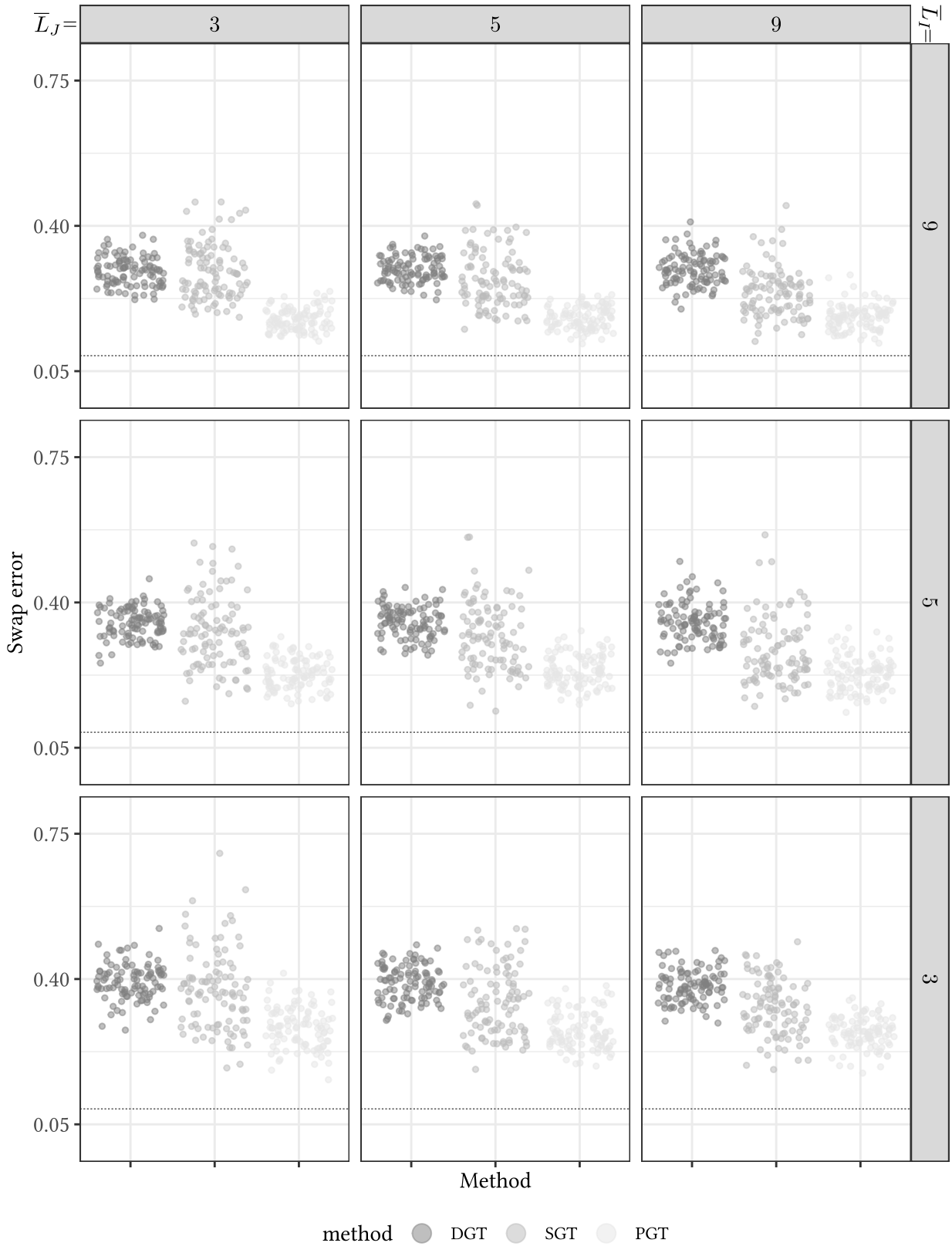


Figure 37: Scatter plots for the "bad" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 3, 5, and 9.

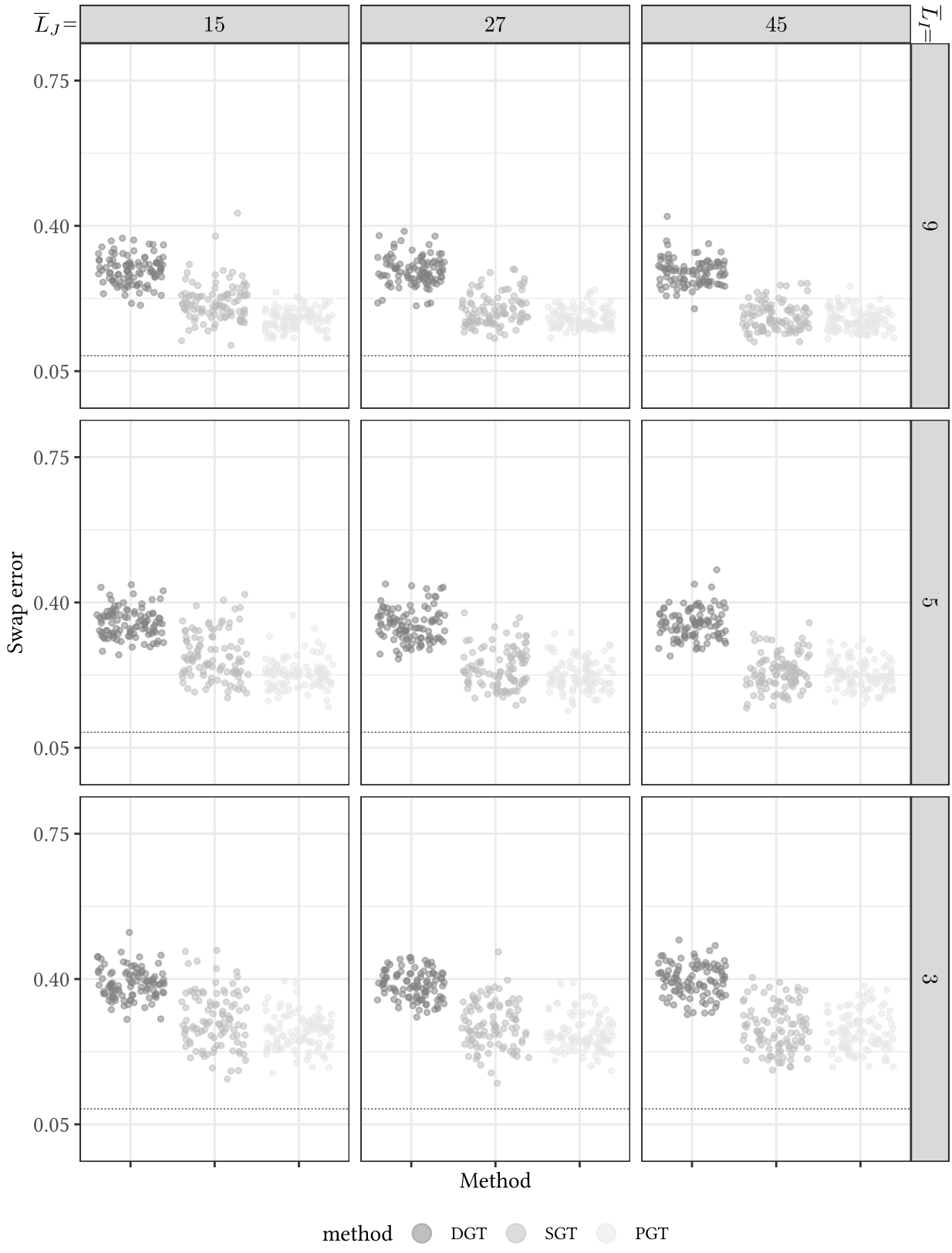


Figure 38: Scatter plots for the "bad" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 3, 5, and 9.

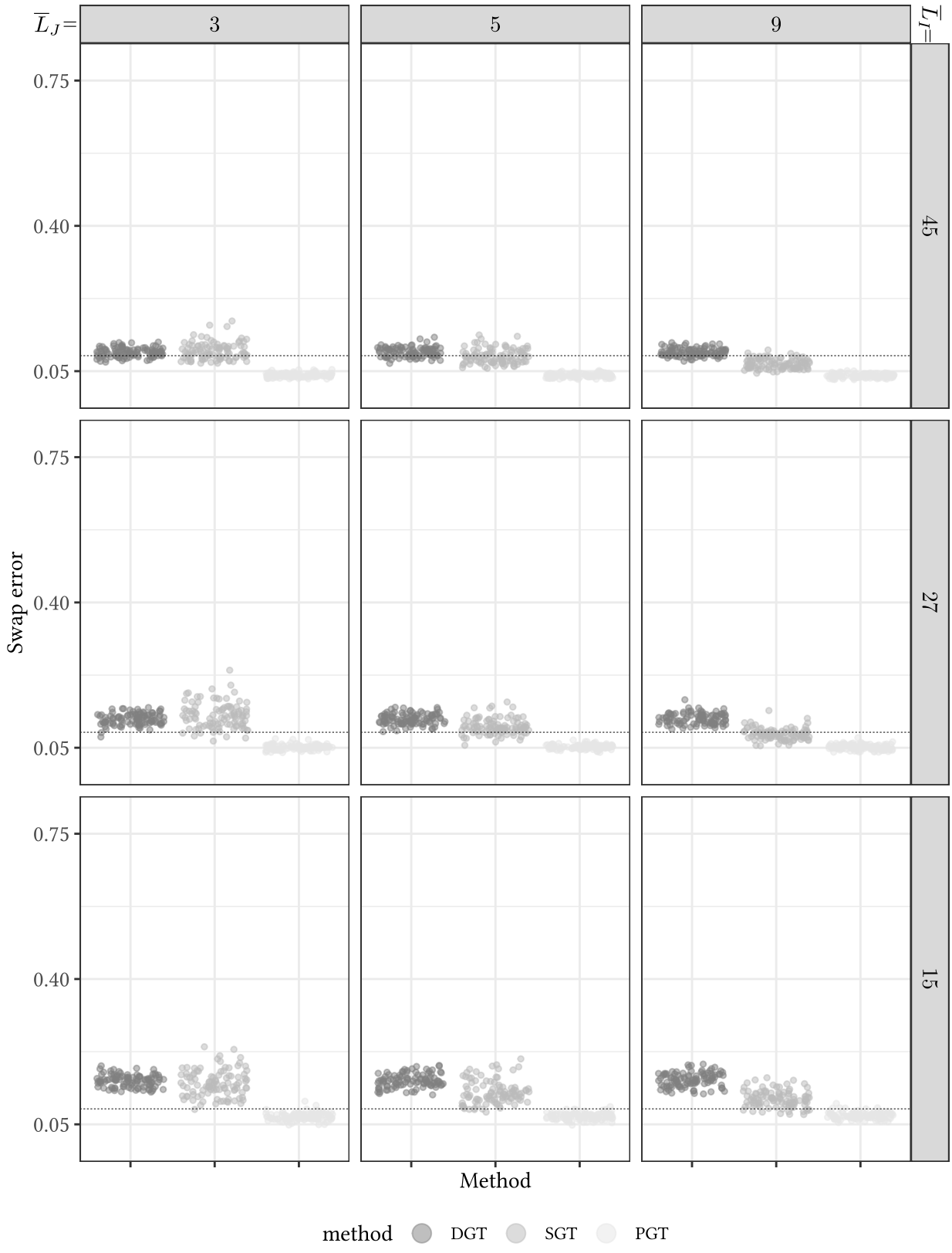


Figure 39: Scatter plots for the "average" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 15, 27, and 45.

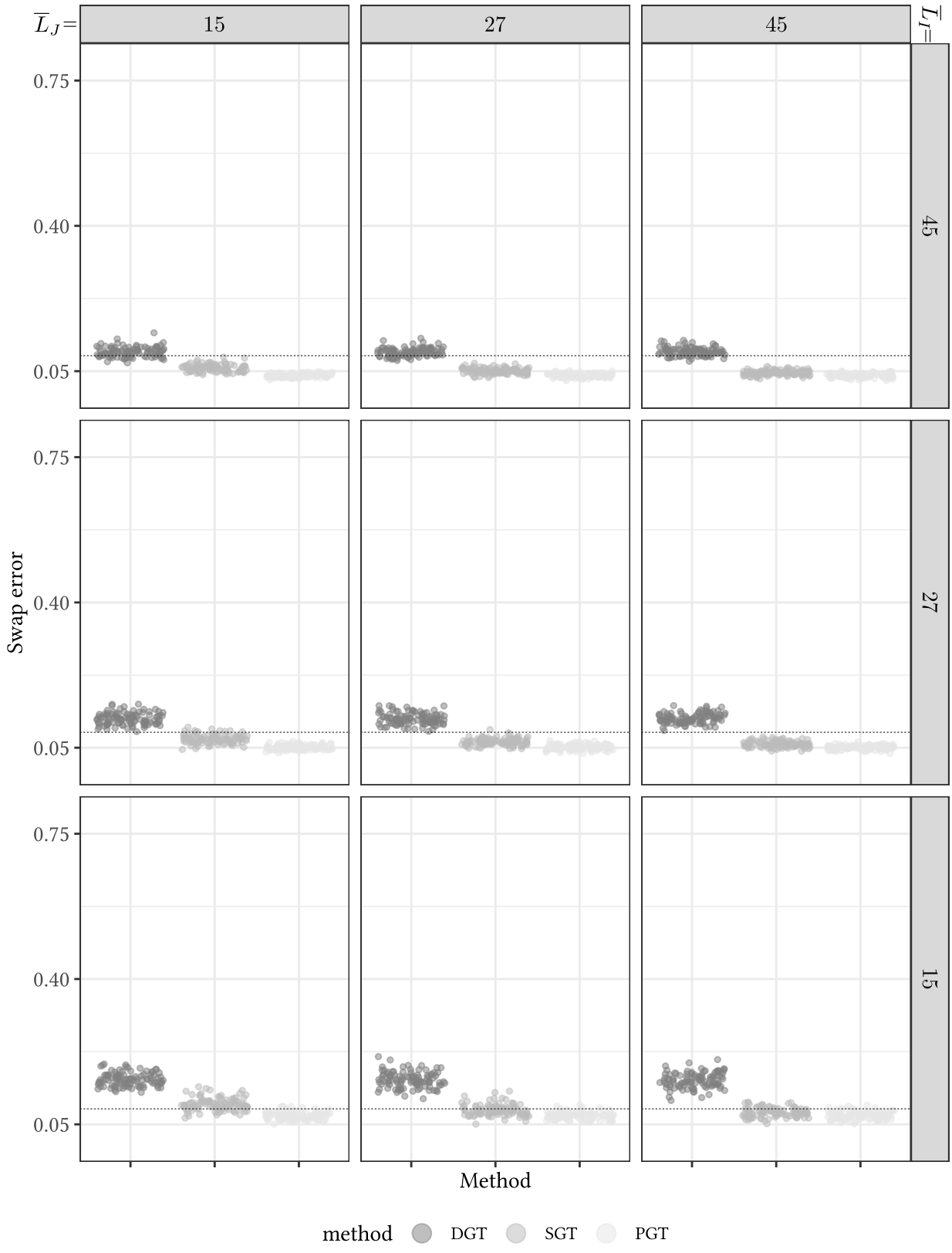


Figure 40: Scatter plots for the "average" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 15, 27, and 45.

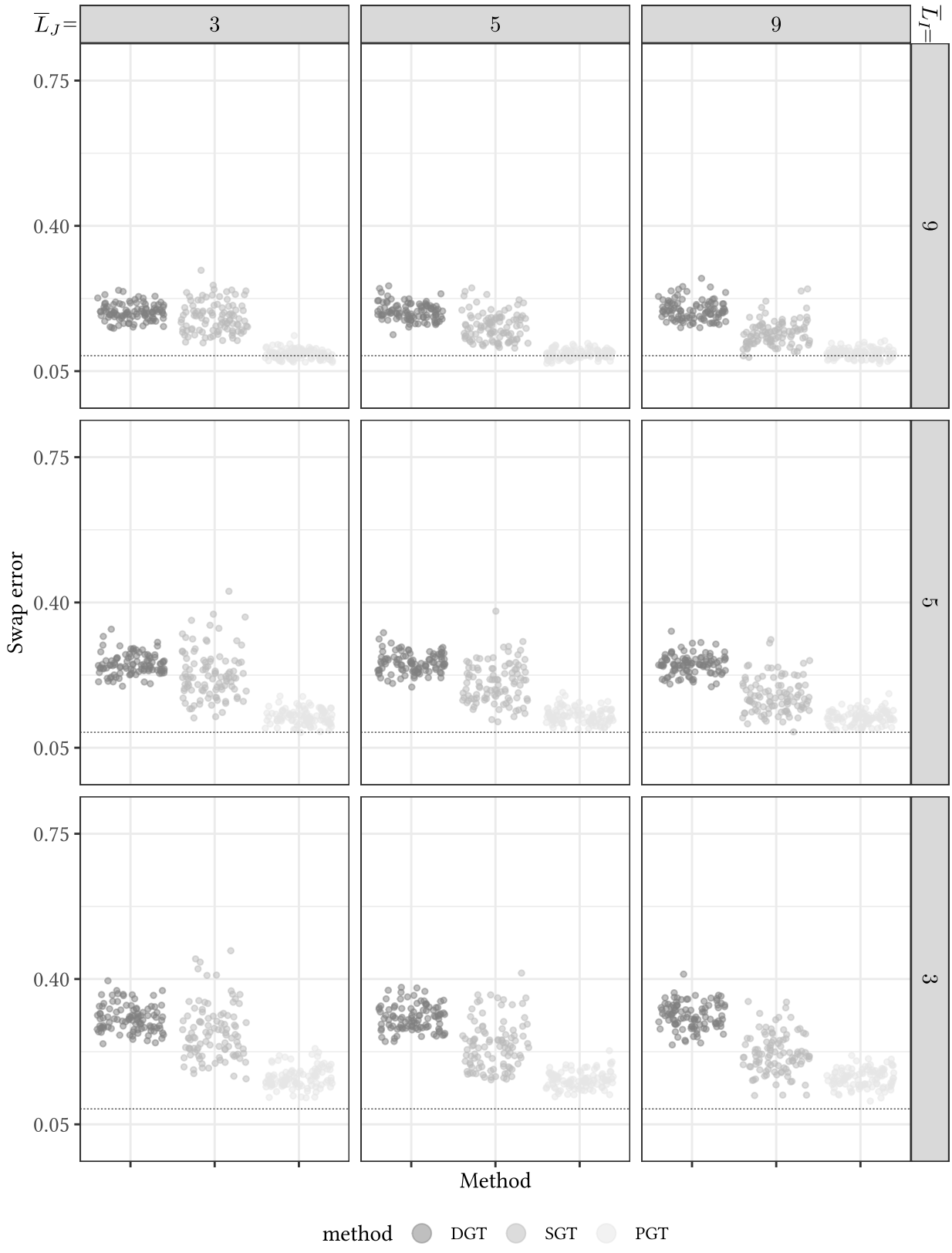


Figure 41: Scatter plots for the "average" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 3, 5, and 9.

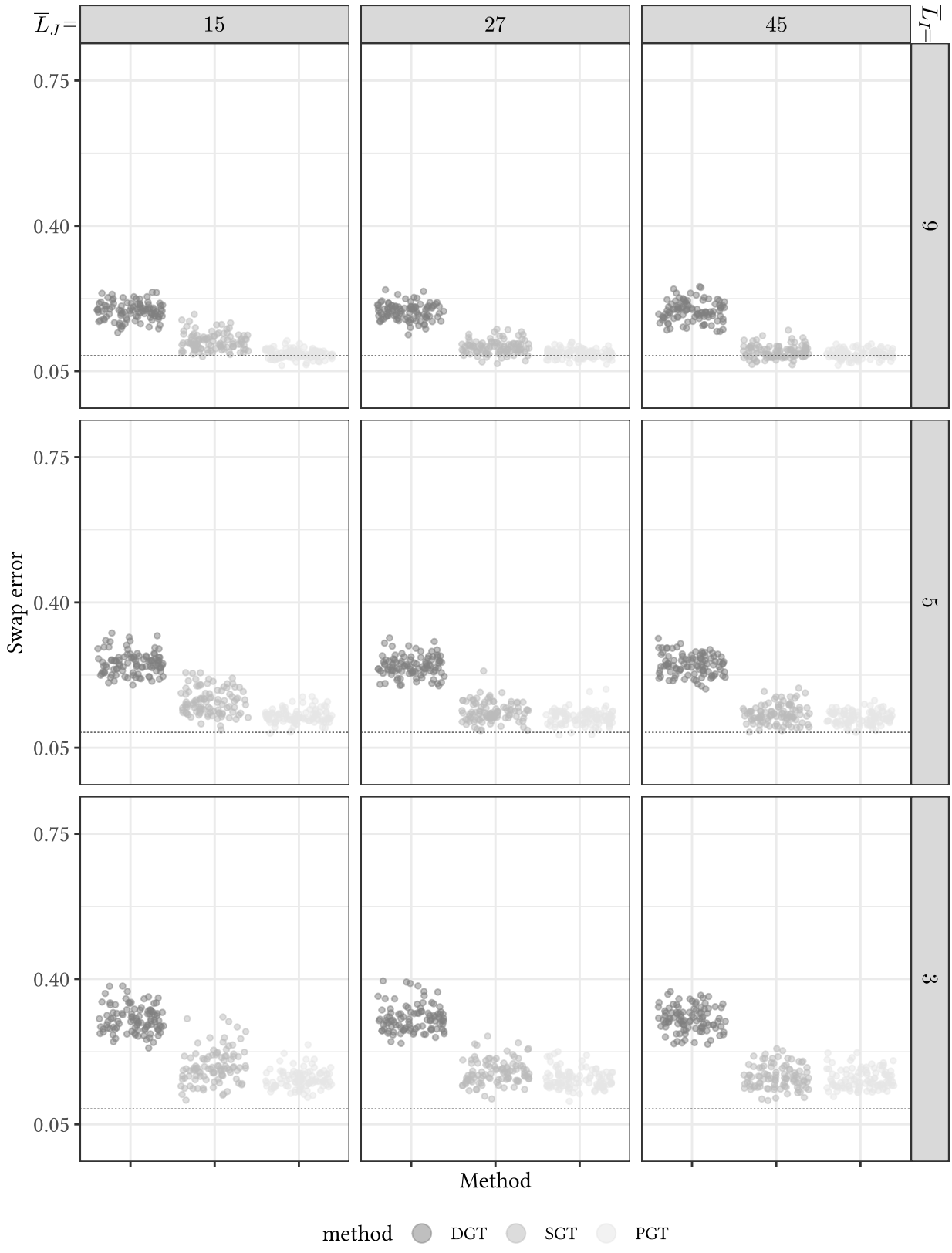


Figure 42: Scatter plots for the "average" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 3, 5, and 9.

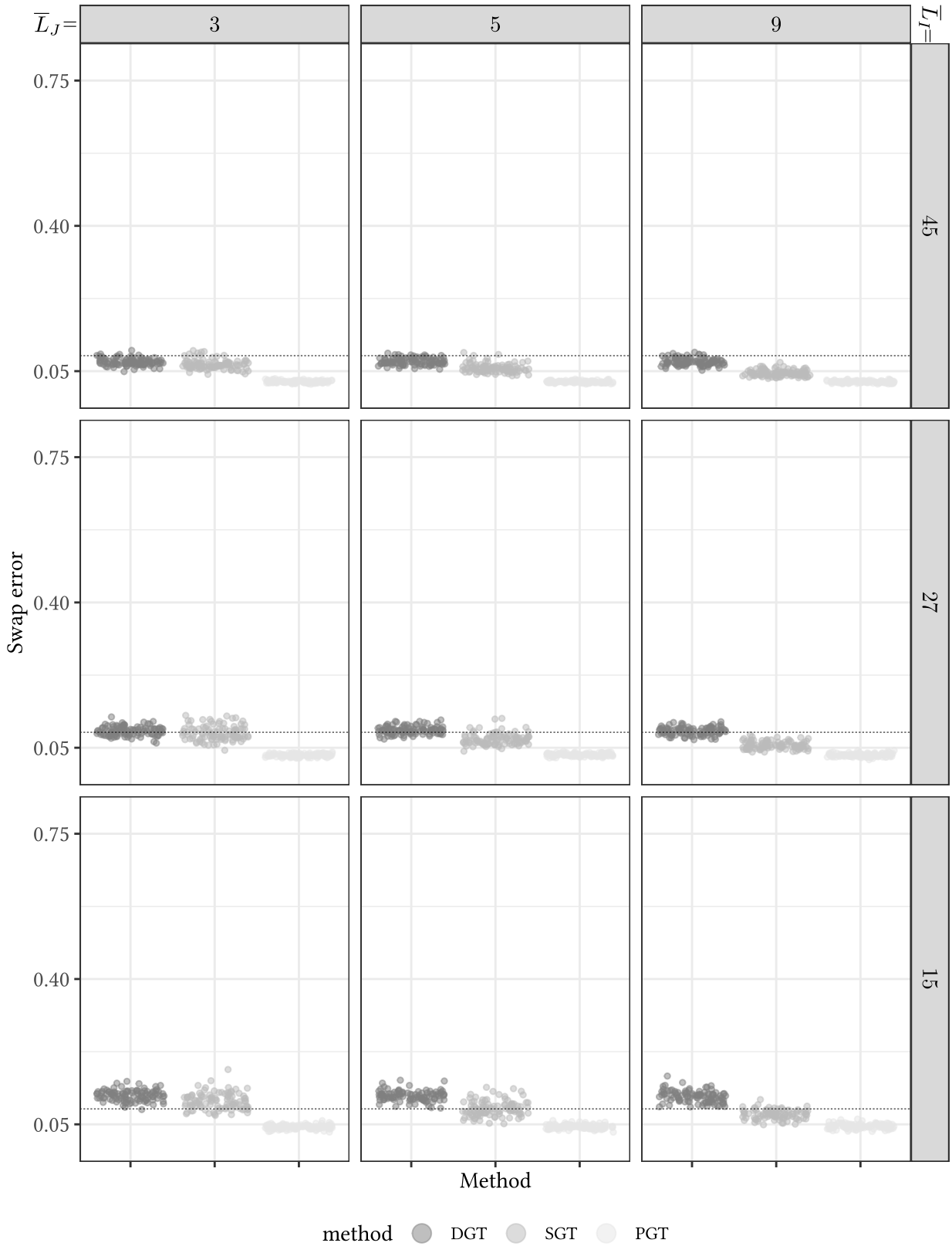


Figure 43: Scatter plots for the "good" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 15, 27, and 45.

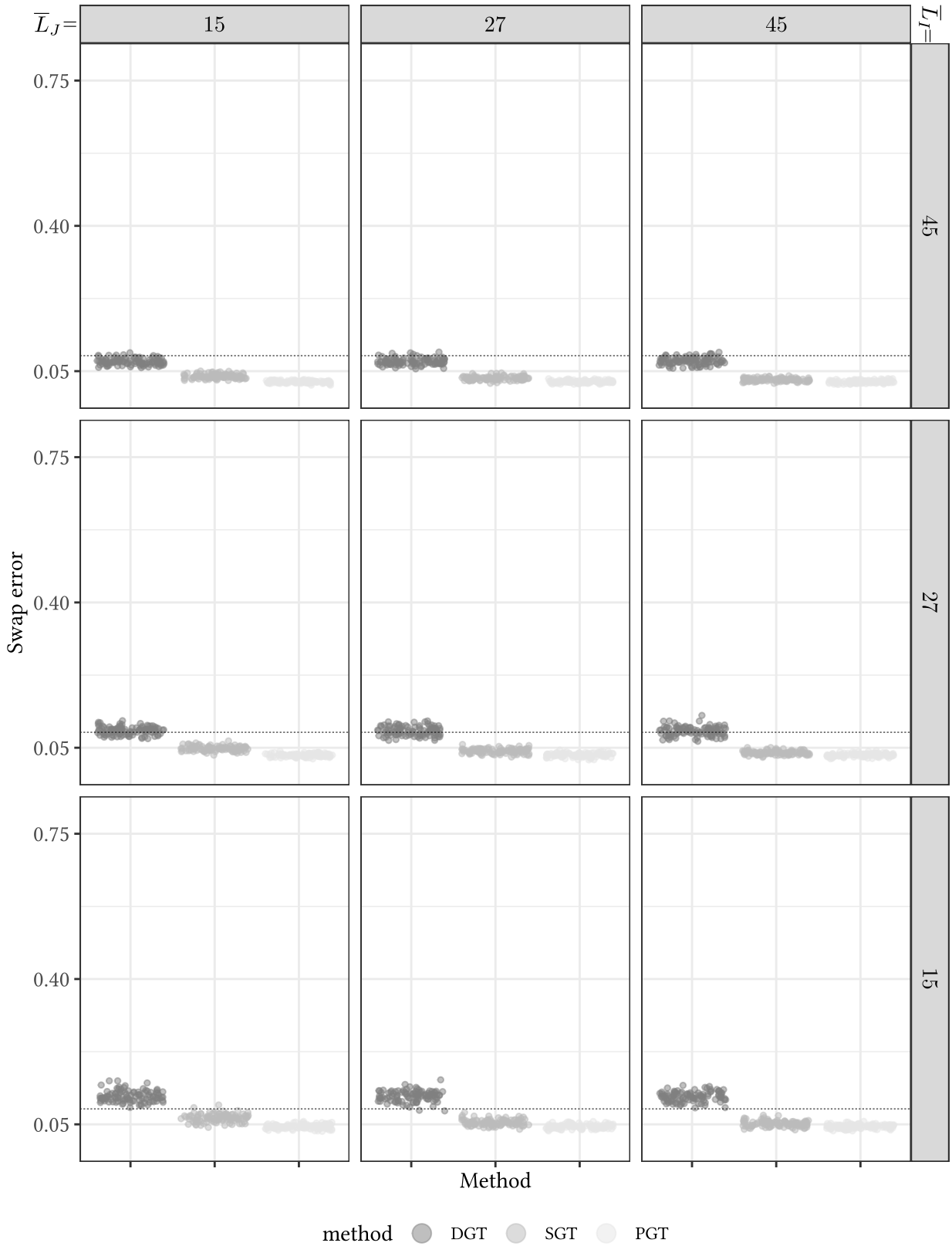


Figure 44: Scatter plots for the "good" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 15, 27, and 45.

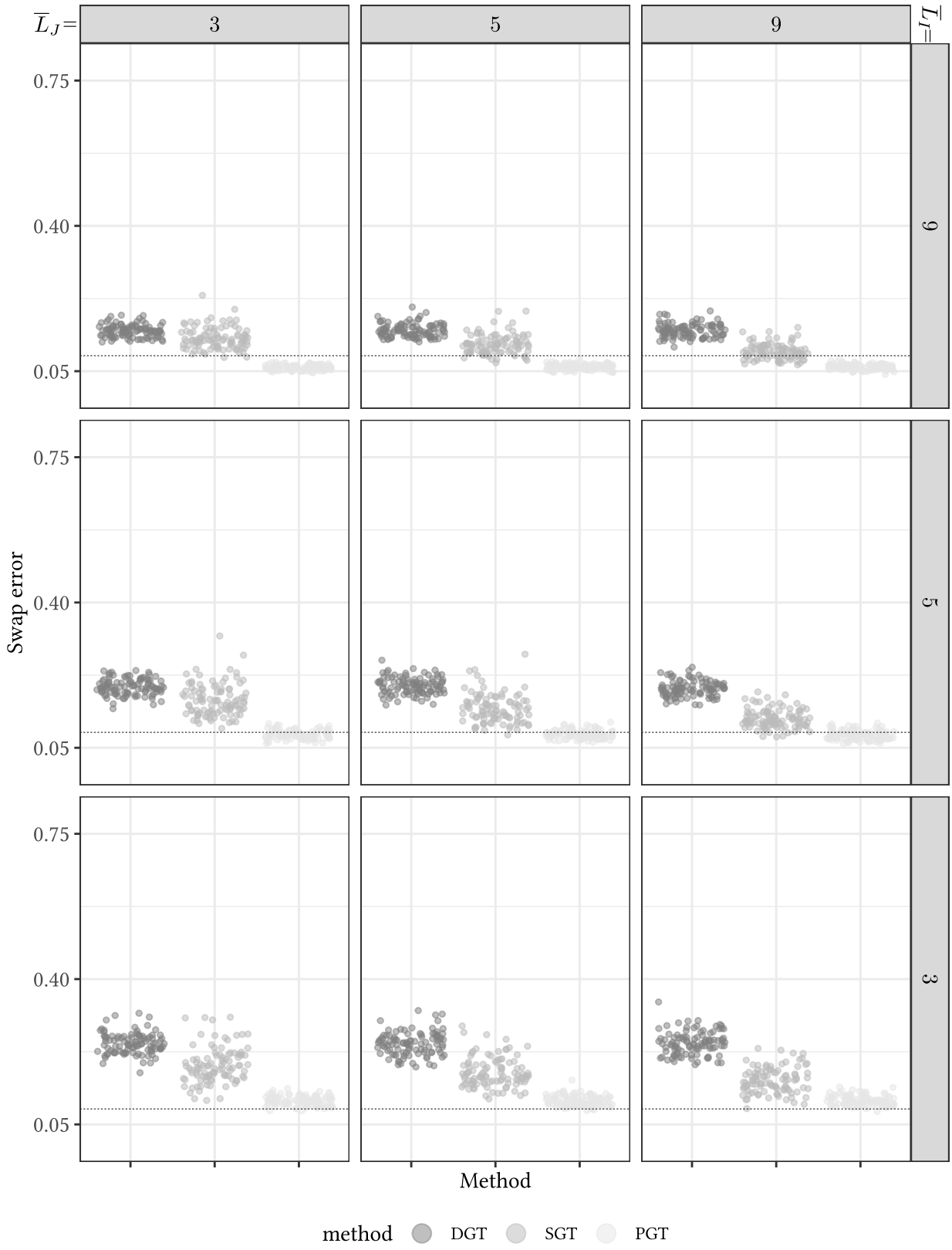


Figure 45: Scatter plots for the "good" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 3, 5, and 9.

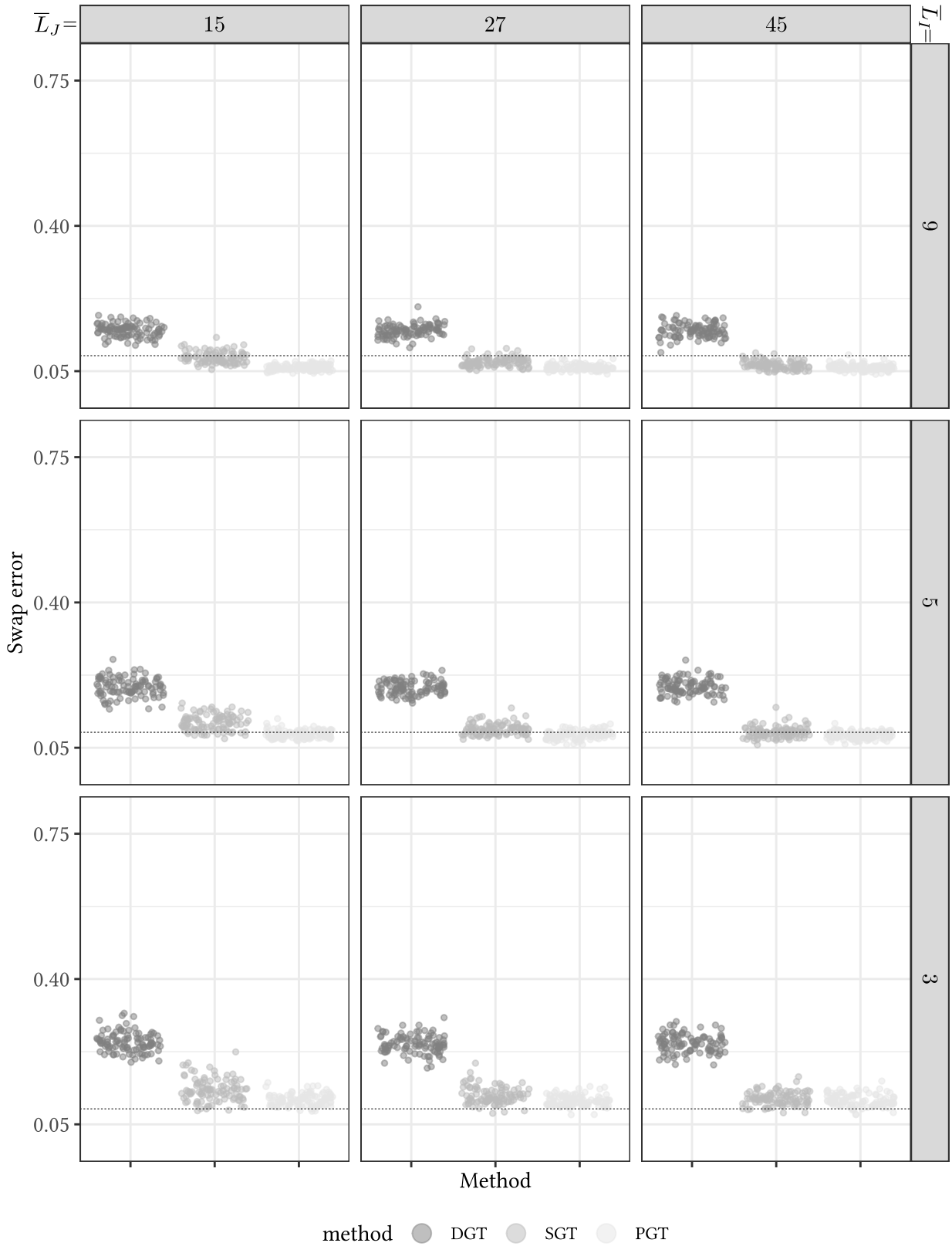


Figure 46: Scatter plots for the "good" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 3, 5, and 9.

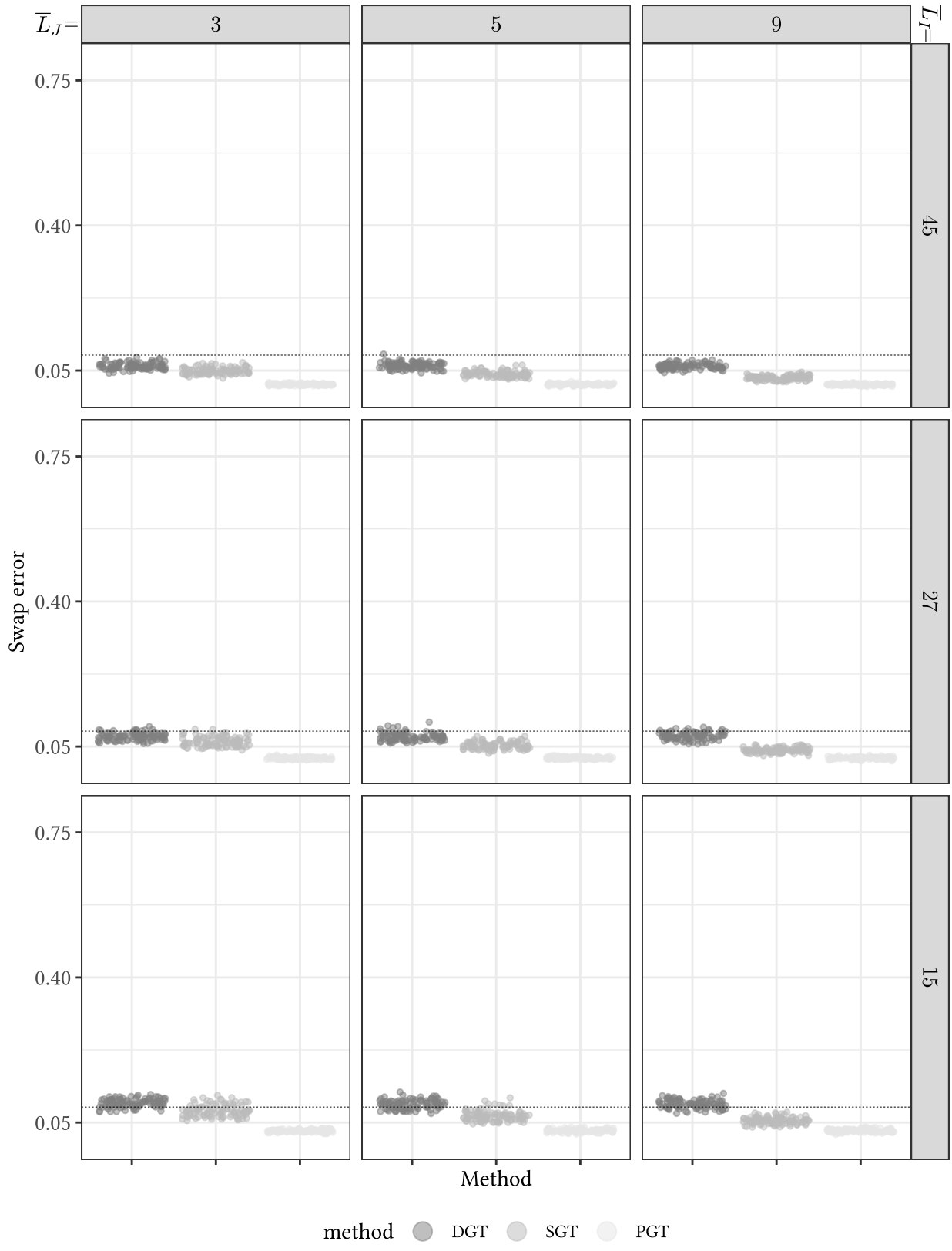


Figure 47: Scatter plots for the "outstanding" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 15, 27, and 45.

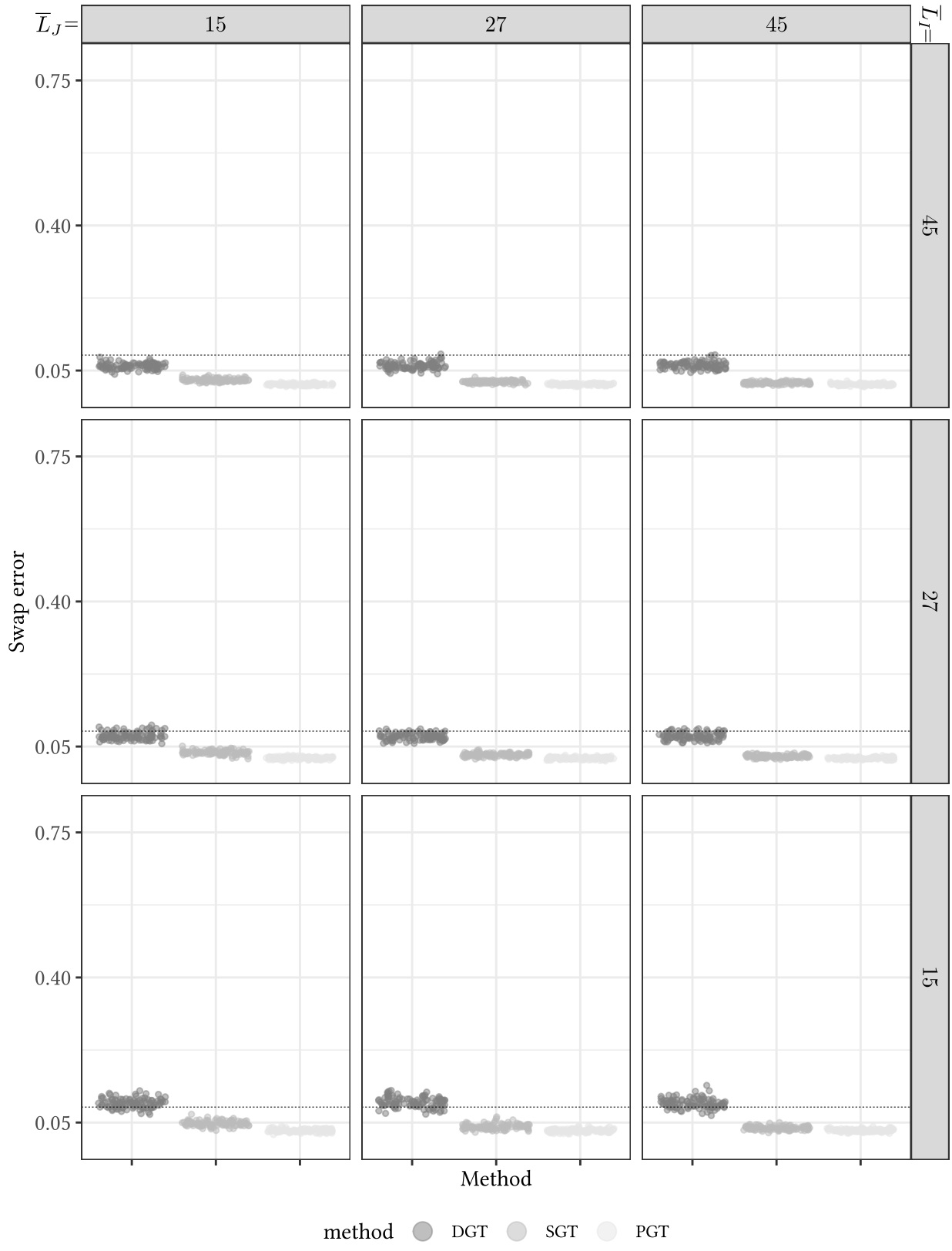


Figure 48: Scatter plots for the "outstanding" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 15, 27, and 45.

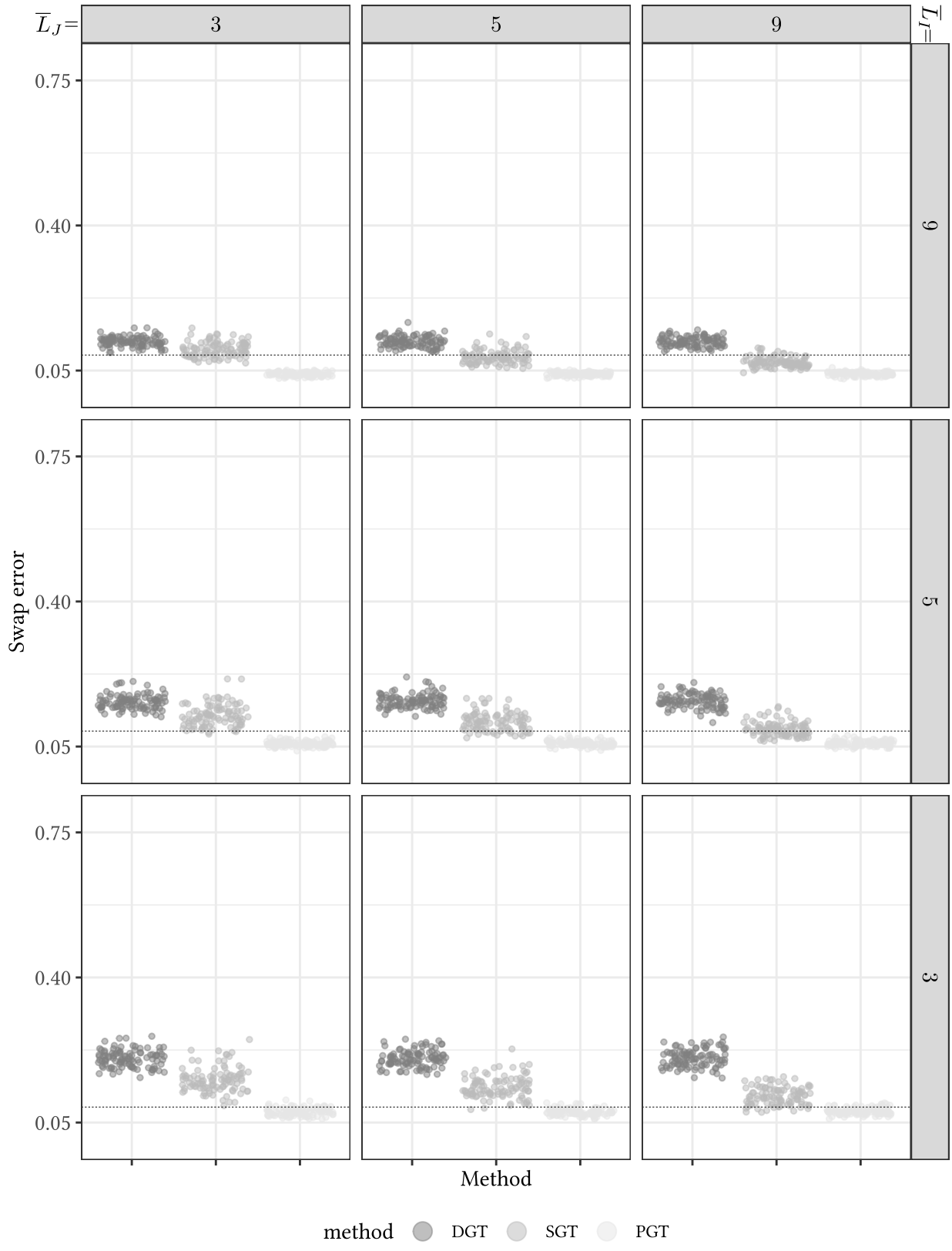


Figure 49: Scatter plots for the "outstanding" case, when $\overline{L}_J$ takes the values 3, 5, and 9, and when $\overline{L}_I$ takes the values 3, 5, and 9.

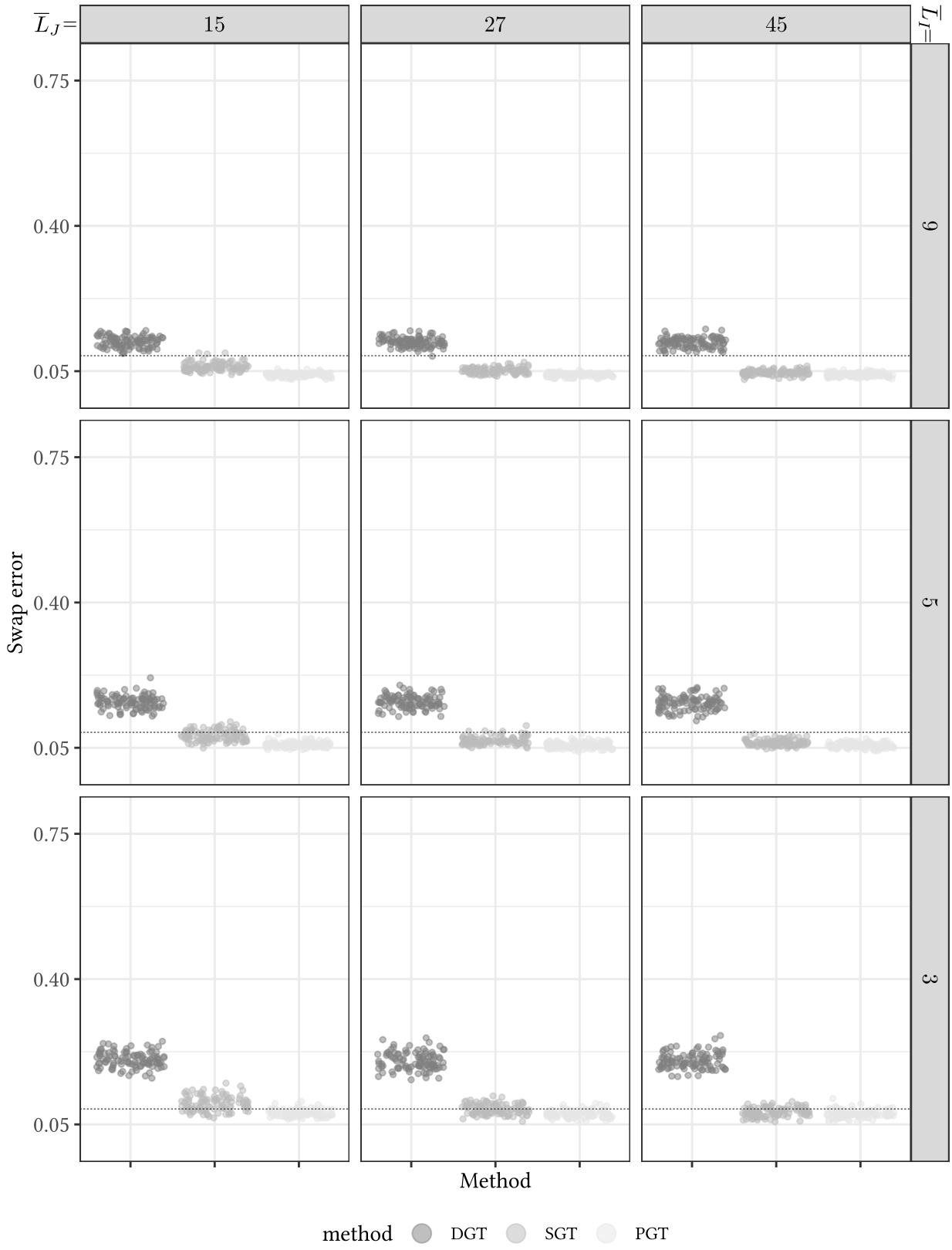


Figure 50: Scatter plots for the "outstanding" case, when $\overline{L}_J$ takes the values 15, 27, and 45, and when $\overline{L}_I$ takes the values 3, 5, and 9.

2 Supplementary material for: Statistical Model for Reproducibility in Ranking-Based Feature Selection

2.1 Introduction

Here, we gather additional documentation which is not presented in Chapter 3 of the thesis manuscript. Briefly, this additional documentation consists of:

- An extensive and detailed description of the algorithm that enables to find the best sequence.
- A specification of the values of the parameters of all the distributions used during the experimentation with synthetic data.
- All the plots and tables derived from the experimentation with synthetic data.
- A description of each real database used during the experimentation with real data.
- The preprocessings applied to the different real databases.
- The stratification process applied after the preprocessing has been carried out when the sampling of the data is tackled in order to derive $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$.
- All the plots and tables derived from the experimentation with real data.

2.2 Detailed description of the algorithm to find the best sequence

The pseudo-code that enables the calculation of $\boldsymbol{a}^*$ for a given $\widehat{\boldsymbol{\rho}}$ is presented in Algorithm 2.1.

Next, we provide a brief description of Algorithm 2.1 in order to summarize and provide a general idea of how it works:

- The variables: n stores the amount of balls, A stores in its columns the best solutions for the different subproblems, ϵ stores the errors of the best solutions for the different subproblems, E_i stores the cumulative errors described in Eq. 3.7 of the thesis manuscript (which we present here again in Eq. 1) for the ith subproblem, P_i store the paths that enable the retrieval of the best solution $\boldsymbol{a}$ for the ith subproblem, and $\boldsymbol{a}^*$ stores the best solution for the whole problem:

$$E_{a_k}^{(k)}(\widehat{\boldsymbol{\rho}}) = \epsilon_k(\widehat{\rho}_k, a_k, a_n) + \min(E_{a_k}^{(k-1)}(\widehat{\boldsymbol{\rho}}), E_{a_k-1}^{(k-1)}(\widehat{\boldsymbol{\rho}})). \tag{1}$$

- The loops: Algorithm 2.1 solves the problem using three nested loops: The outer one iterates through different subproblems, the middle one through different positions of the sequences of the amounts of relevant balls, and the inner one through different amounts of relevant balls. The outer loop only needs to cover half of the subproblems due to the aforementioned symmetry regarding the relative amount of relevant and irrelevant balls.
- The steps:
  - n, A and ϵ are initialized.
  - Inside the outer loop, in each iteration:
    - E_i and P_i are initialized.
    - The middle and inner loops are executed to solve the ith subproblem, filling E_i and P_i accordingly (considering Eq. 1).
    - The best error achieved in the ith subproblem is stored in ϵ.

Computing $\boldsymbol{a}^*$

input: Estimated expected reproducibility curve $\widehat{\boldsymbol{\rho}}$.
output: Sequence of the amounts of relevant balls $\boldsymbol{a}^*$ that minimizes the cumulative error function.
$n = \text{Length}(\widehat{\boldsymbol{\rho}})$
$A = \text{Zeros}(n, \lfloor n/2 \rfloor + 1)$
$\epsilon = \text{Zeros}(\lfloor n/2 \rfloor + 1)$
for $i = 0$ **to** $\lfloor n/2 \rfloor$
    $E_i = \text{Infinites}(n+1, i+1)$
    $E_i[0,0] = 0$
    $P_i = \text{Zeros}(n+1, i+1)$
    for $j = 1$ **to** n
        for $k = 0$ **to** $\min(j, i)$
            if $k = 0$ **then**
                $E_i[j,k] = E_i[j-1,k] + \epsilon_k(\widehat{\rho}_k, k, i)$
                $P_i[j,k] = k$
            fi
            if $k = j$ **then**
                $E_i[j,k] = E_i[j-1,k-1] + \epsilon_k(\widehat{\rho}_k, k, i)$
                $P_i[j,k] = k-1$
            fi
            if $k \neq 0$ and $k \neq j$ **then**
                if $E_i[j-1,k] < E_i[j-1,k-1]$ **then**
                $E_i[j,k] = E_i[j-1,k] + \epsilon_k(\widehat{\rho}_k, k, i)$
                $P_i[j,k] = k$
                else
                $E_i[j,k] = E_i[j-1,k-1] + \epsilon_k(\widehat{\rho}_k, k, i)$
                $P_i[j,k] = k-1$
                fi
            fi
        rof
    rof
    $\epsilon[i] = E_i[i,n]$
    $A[.,i] = \text{Get\_subproblem\_best\_solution}(P_i)$
rof
$\boldsymbol{a}^* = \text{Get\_problem\_best\_solution}(A, \epsilon)$
return $\boldsymbol{a}^*$

Algorithm 2.1: The pseudo-code for computing the sequence of the amounts of relevant balls $\boldsymbol{a}^*$ of minimum error.

    * From the filled matrix of paths P_i the sequence $\boldsymbol{a}$ that is the best solution for subproblem i is derived and stored in the ith column of A.
- Given ϵ, the best solution for the whole problem, $\boldsymbol{a}^*$, can be found within the solutions stored in A.

Now, we provide a detailed explanation of how multiple lines of Algorithm 2.1 work in order to ease a

full understanding of how Algorithm 2.1 works:

- Line 3: Sets n to be the amount of balls, which is equal to the amount of features of the estimated expected reproducibility curve $\widehat{\boldsymbol{\rho}}$.
- Line 4: Creates the matrix A of n rows and $\lfloor n/2 \rfloor + 1$ columns filled with zeros. This matrix is dedicated to storing in each column the best solution (a sequence $\boldsymbol{a}$) of a different subproblem of the $\lfloor n/2 \rfloor + 1$ subproblems.
- Line 5: Creates the vector ϵ of length $\lfloor n/2 \rfloor + 1$ filled with zeros. This vector is dedicated to storing the $\lfloor n/2 \rfloor + 1$ errors associated to the $\lfloor n/2 \rfloor + 1$ best solutions of the $\lfloor n/2 \rfloor + 1$ subproblems.
- Line 6: Starts the outer loop in which each iteration is dedicated to a different subproblem of the $\lfloor n/2 \rfloor + 1$ subproblems.
- Line 7: Creates the matrix E_i of $n + 1$ rows and $i + 1$ columns filled with infinites. This matrix is dedicated to storing the cumulative errors described in Eq. 1.
- Line 8: Initializes the trivial case of E_i, in which the cumulative error is always 0 by definition.
- Line 9: Creates the matrix P_i of $n + 1$ rows and $i + 1$ columns filled with zeros. This matrix is dedicated to storing the paths that enable the retrieval of the best solution $\boldsymbol{a}$ for the ith subproblem. Specifically, for a given cell in row j and in column k, it stores information regarding the best solution $\boldsymbol{a}$ for the ith subproblem belonging to the subset of solutions that fulfill $a_k = k$. Briefly, the value of that cell, $P_i[j, k]$, specifies both the value of a_{k-1} for that solution and the column of the previous row of P_i in which the specification of the value of a_{k-2} can be located.
- Line 10: Starts the middle loop in which each iteration is dedicated to a different row of the ith subproblem, given that each row is associated to a different position of the sequences of the amounts of relevant balls.
- Line 11: Starts the inner loop in which each iteration is dedicated to a different column of the ith subproblem, given that each column is associated to a different amount of relevant balls.
- Lines 12 to 15: Given that $E_i[j-1, k-1]$ is not an option because $k = 0$ (line 12), the only possibility is to set the cumulative error $E_i[j, k]$ to be $\epsilon_k(\widehat{\rho}_k, k, i) + E_i[j-1, k]$, following the essence of Eq. 1. $P_i[j, k]$ is updated in consonance to be k.
- Lines 16 to 19: Given that $E_i[j-1, k]$ is not an option because $k = j$ (line 16), the only possibility is to set the cumulative error $E_i[j, k]$ to be $\epsilon_k(\widehat{\rho}_k, k, i) + E_i[j-1, k-1]$, following the essence of Eq. 1. $P_i[j, k]$ is updated in consonance to be $k - 1$.
- Lines 20 to 28: Given that $k \neq 0$ and that $k \neq j$ (line 20) there are two possibilities (either $E_i[j-1, k] < E_i[j-1, k-1]$ is satisfied or not). Each possibility corresponds to a different option in Eq. 1. Consequently, each possibility implies a different update of $E_i[j, k]$ and $P_i[j, k]$, according to Eq. 1.
- Line 31: The best error achieved in the ith subproblem is stored in the ith position of ϵ.
- Line 32: Given the matrix of paths P_i, the sequence $\boldsymbol{a}$ that is the best solution for subproblem i is derived. First, for this sequence $\boldsymbol{a}$, it is already known that $a_n = i$. Secondly, the rest of the solution $\boldsymbol{a}$ is derived starting from $P_i[i, n]$. Let us recall that the value $P_i[i, n]$ specifies both the value a_{n-1}

of the solution $\boldsymbol{a}$ and the column of the previous row of P_i in which the value of a_{n-2} can be located. In conclusion, departing from $P_i[i, n]$ and proceeding recursively, the sequence $\boldsymbol{a}$ is derived. Finally, the retrieved sequence $\boldsymbol{a}$ is stored in the ith column of A.

- Line 34: First, the position in ϵ that stores the minimum value is located. Secondly, the sequence $\boldsymbol{a}$ stored in the corresponding column of A is stored in $\boldsymbol{a}^*$.

2.3 Parameters of the synthetic experimentation

First of all, for the sake of clarity, let us show again in Figure 51 the distributions from which the data are sampled in the experimentation with synthetic data (this figure is also shown in Chapter 3 of the thesis manuscript).

In Table 1 we show the specific values of the parameters of each of the distributions shown in Figure 51. In Table 1 each of those distributions is identified by its associated scenario (differences in location or differences in both location and spread) and difficulty as shown in Figure 51.

2.4 Plots and tables of the experimentation with synthetic data

In Figures 52 to 93 the plots of the experimentation with synthetic data can be seen. Specifically, the plots are shown in order, first showing those corresponding to the scenario of differences in location and then showing those corresponding to the scenario of differences in both location and spread. Additionally, the plots belonging to the same scenario are shown in order, first showing those corresponding to difficulty 1 and lastly showing those corresponding to difficulty 21.

In Tables 2 and 3 the weights and AUC values of the experimentation with synthetic data can be seen.

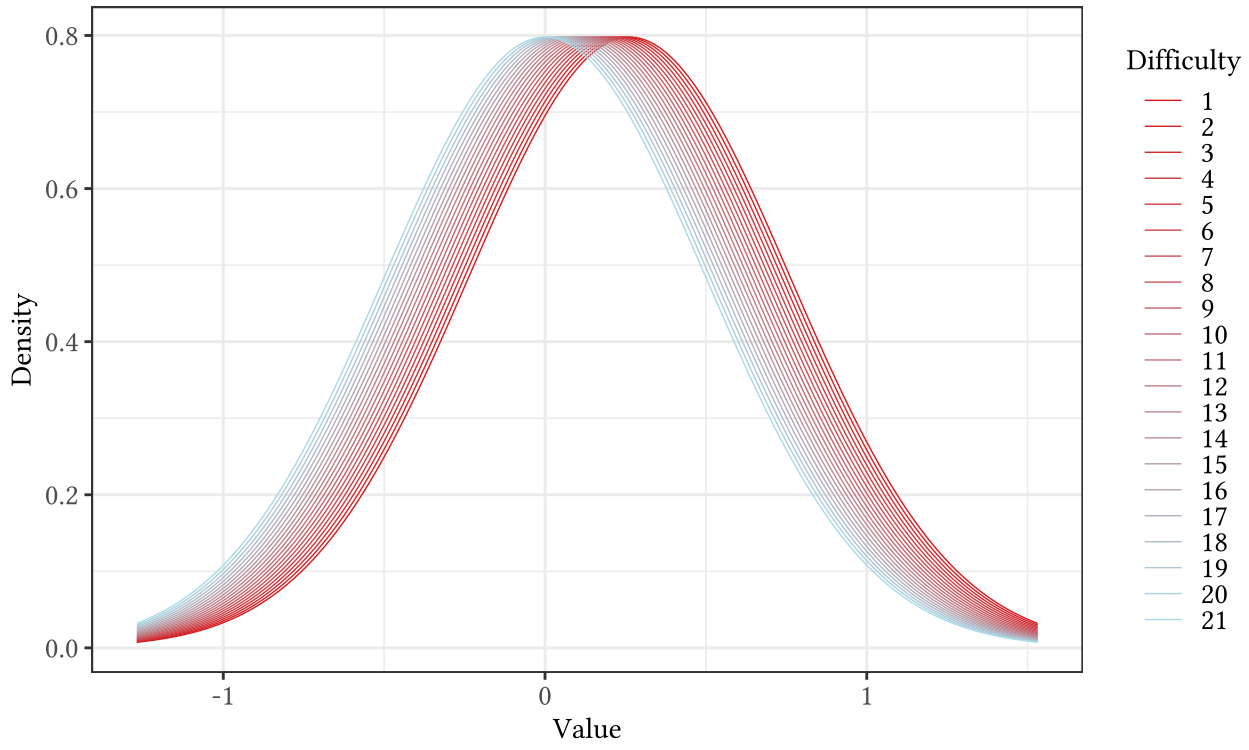
2.5 Descriptions of the real databases

- Breast database [1, 2]: This dataset includes 30 features and 569 instances. Specifically, the features consist of visual characteristics of the cell nuclei present in digitized images from patients with breast cancer. The instances of this dataset are breast masses of women with breast tumors, 357 women with benign tumors and 212 women with malignant tumors.

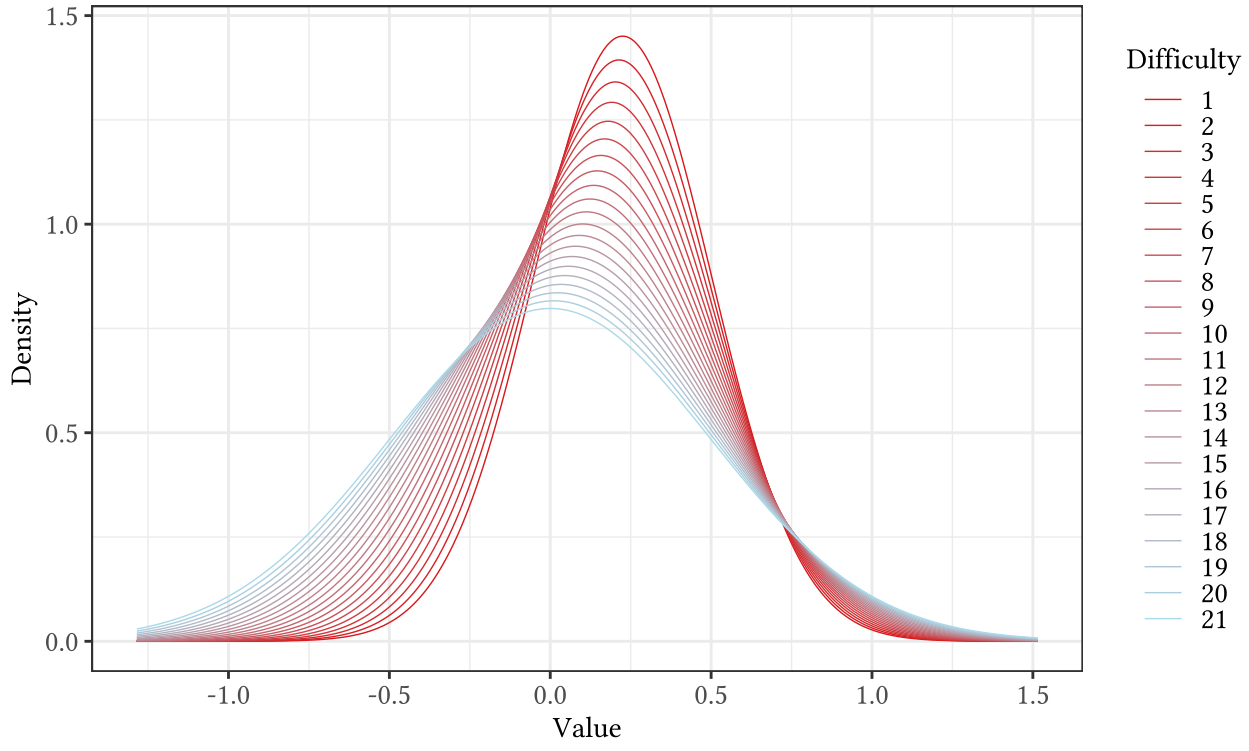
- Mice database [3]: This dataset contains 77 features and 1080 instances. The features consist of protein expression levels of 77 proteins, while the instances correspond to different measurements in 38 control mice and 34 mice with Down's syndrome (multiple measurements of each protein were carried out for each mouse).

- SECOM database [4]: This dataset has 591 features and 1567 instances. Briefly, this dataset is a semiconductor manufacturing process dataset in which the features correspond to process signals. Regarding the instances, each of them corresponds to a different production entity, the instances being divided into 1463 correct productions and 104 faulty productions.

- Arcene database [5]: This dataset includes 10000 features and 900 instances. Specifically, this dataset consists of mass-spectrometric data of biological samples. In particular, 7000 features are measurements derived from the biological samples, while the remaining 3000 features are non-real features added as a distracting factor. Besides, the 900 instances consists of individuals that can be divided into two groups, 398 patients with cancer and 502 healthy individuals.



(a)



(b)

Figure 51: Distributions used in the scenario of differences in location (51a) and in the scenario of differences in both location and spread (51b)

Table 1: Values of the parameters of each of the distributions used during the experimentation with synthetic data.

Scenario	Difficulty	Parameters	
		μ	σ^2
Location	1	0.262500	0.500000^2
Location	2	0.249375	0.500000^2
Location	3	0.236250	0.500000^2
Location	4	0.223125	0.500000^2
Location	5	0.210000	0.500000^2
Location	6	0.196875	0.500000^2
Location	7	0.183750	0.500000^2
Location	8	0.170625	0.500000^2
Location	9	0.157500	0.500000^2
Location	10	0.144375	0.500000^2
Location	11	0.131250	0.500000^2
Location	12	0.118125	0.500000^2
Location	13	0.105000	0.500000^2
Location	14	0.091875	0.500000^2
Location	15	0.078750	0.500000^2
Location	16	0.065625	0.500000^2
Location	17	0.052500	0.500000^2
Location	18	0.039375	0.500000^2
Location	19	0.026250	0.500000^2
Location	20	0.013125	0.500000^2
Location	21	0.000000	0.500000^2
Location & spread	1	0.225000	0.275000^2
Location & spread	2	0.213750	0.286250^2
Location & spread	3	0.202500	0.297500^2
Location & spread	4	0.191250	0.308750^2
Location & spread	5	0.180000	0.320000^2
Location & spread	6	0.168750	0.331250^2
Location & spread	7	0.157500	0.342500^2
Location & spread	8	0.146250	0.353750^2
Location & spread	9	0.135000	0.365000^2
Location & spread	10	0.123750	0.376250^2
Location & spread	11	0.112500	0.387500^2
Location & spread	12	0.101250	0.398750^2
Location & spread	13	0.090000	0.410000^2
Location & spread	14	0.078750	0.421250^2
Location & spread	15	0.067500	0.432500^2
Location & spread	16	0.056250	0.443750^2
Location & spread	17	0.045000	0.455000^2
Location & spread	18	0.033750	0.466250^2
Location & spread	19	0.022500	0.477500^2
Location & spread	20	0.011250	0.488750^2
Location & spread	21	0.000000	0.500000^2

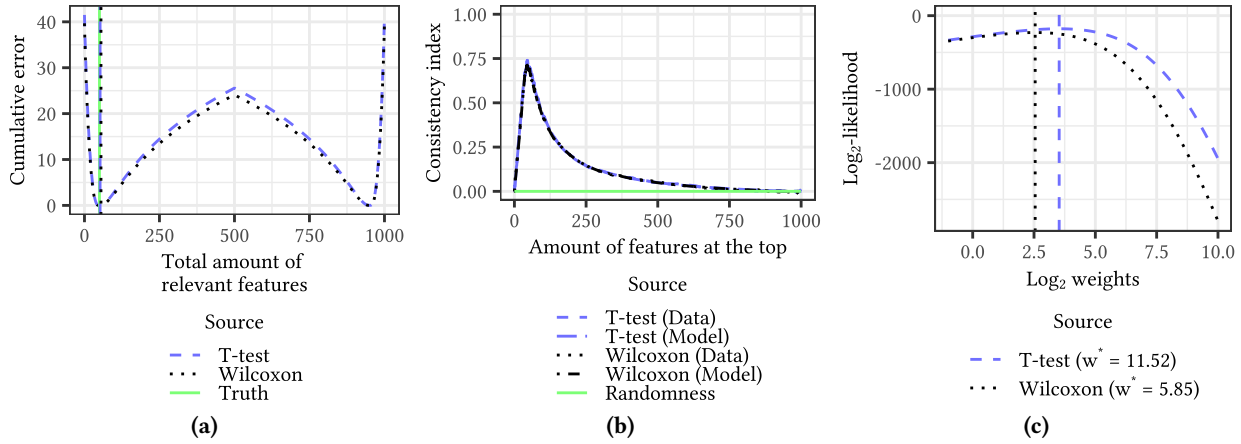


Figure 52: Error plot (52a), reproducibility plot (52b) and weight plot (52c) for the difficulty configuration 1, in the differences in location scenario

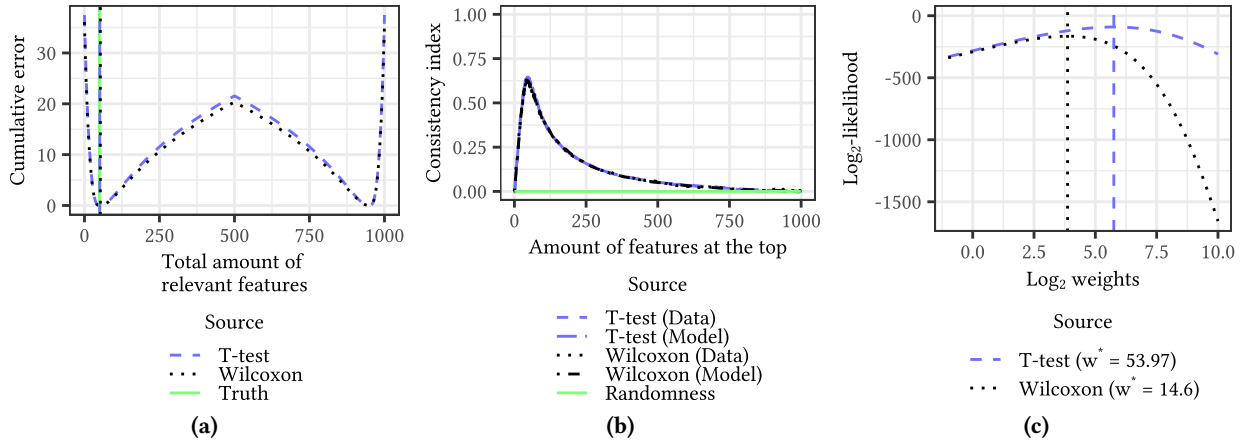


Figure 53: Error plot (53a), reproducibility plot (53b) and weight plot (53c) for the difficulty configuration 2, in the differences in location scenario

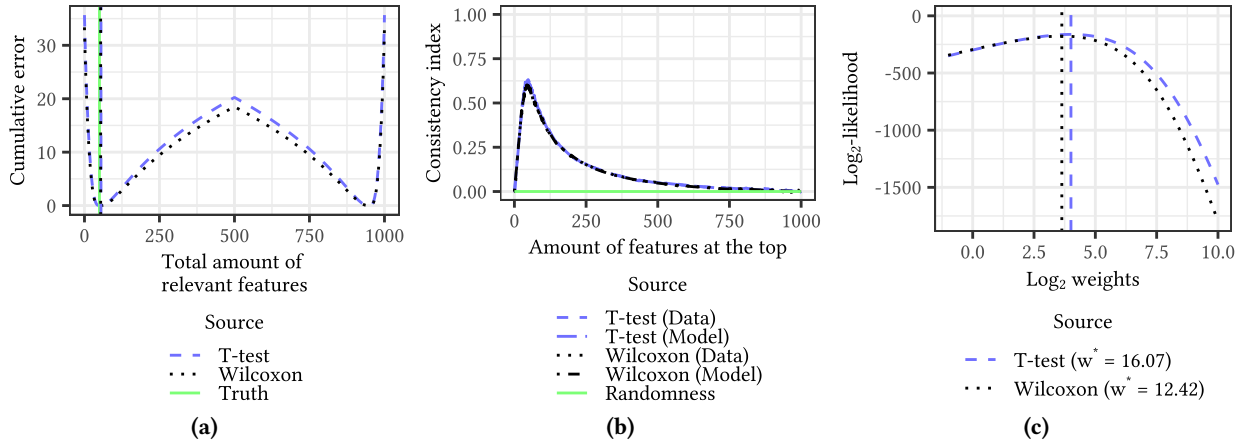


Figure 54: Error plot (54a), reproducibility plot (54b) and weight plot (54c) for the difficulty configuration 3, in the differences in location scenario

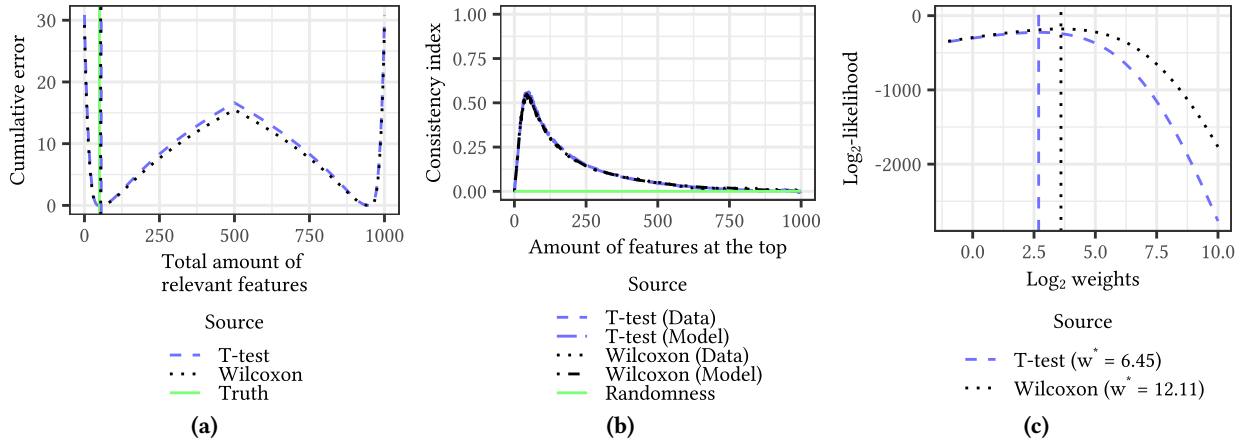


Figure 55: Error plot (55a), reproducibility plot (55b) and weight plot (55c) for the difficulty configuration 4, in the differences in location scenario

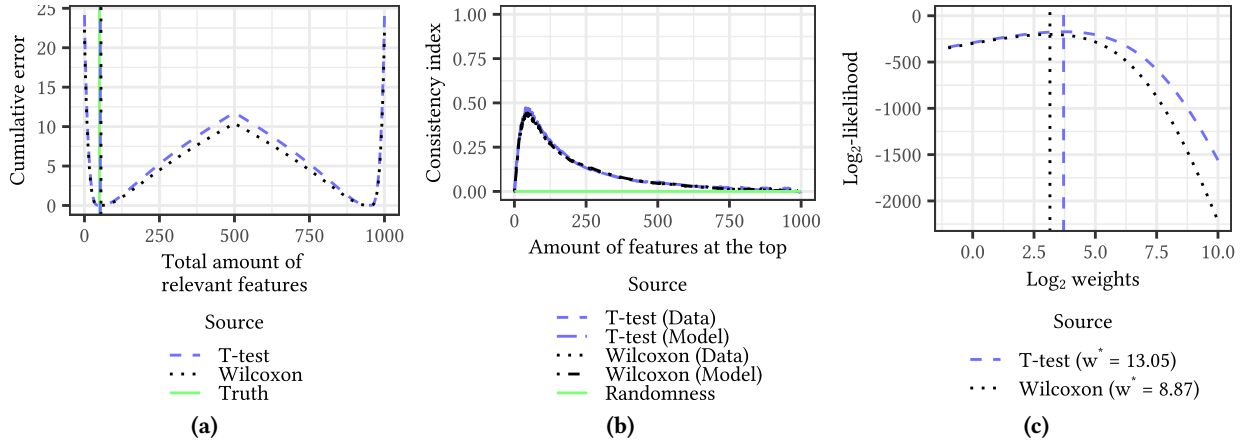


Figure 56: Error plot (56a), reproducibility plot (56b) and weight plot (56c) for the difficulty configuration 5, in the differences in location scenario

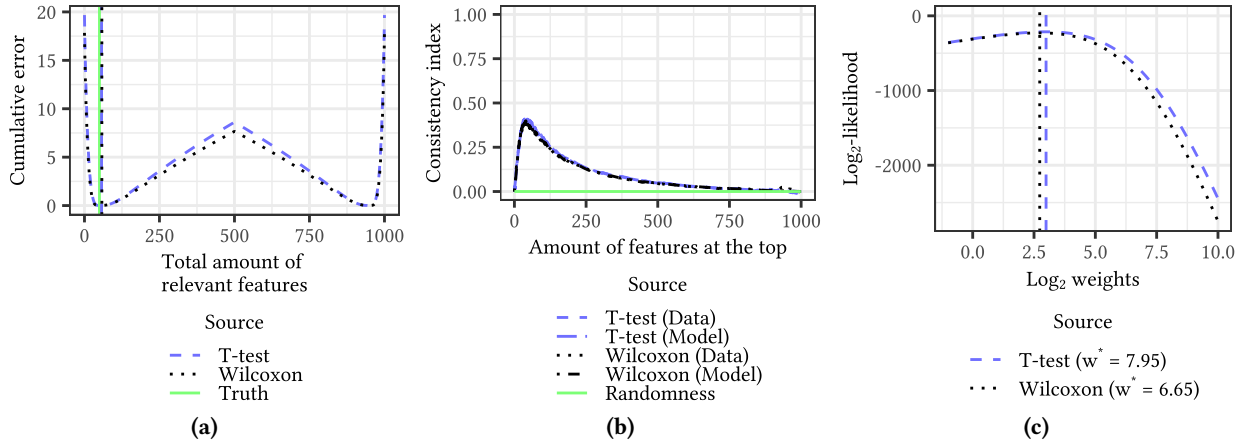


Figure 57: Error plot (57a), reproducibility plot (57b) and weight plot (57c) for the difficulty configuration 6, in the differences in location scenario

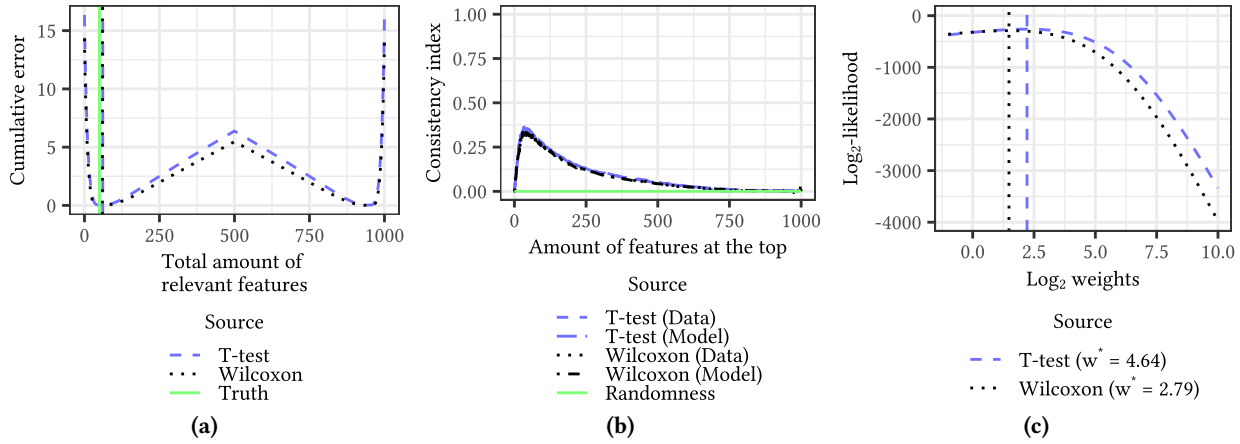


Figure 58: Error plot (58a), reproducibility plot (58b) and weight plot (58c) for the difficulty configuration 7, in the differences in location scenario

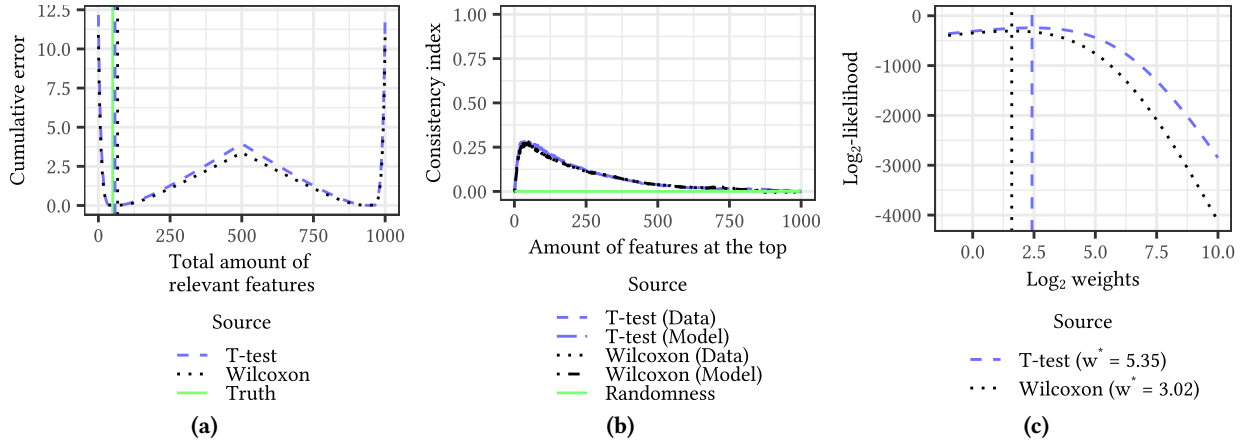


Figure 59: Error plot (59a), reproducibility plot (59b) and weight plot (59c) for the difficulty configuration 8, in the differences in location scenario

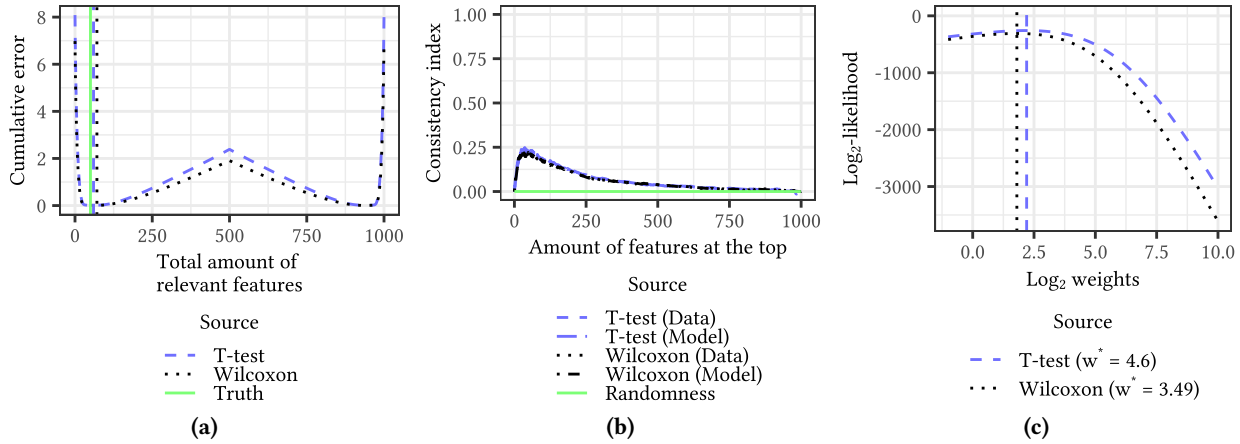


Figure 60: Error plot (60a), reproducibility plot (60b) and weight plot (60c) for the difficulty configuration 9, in the differences in location scenario

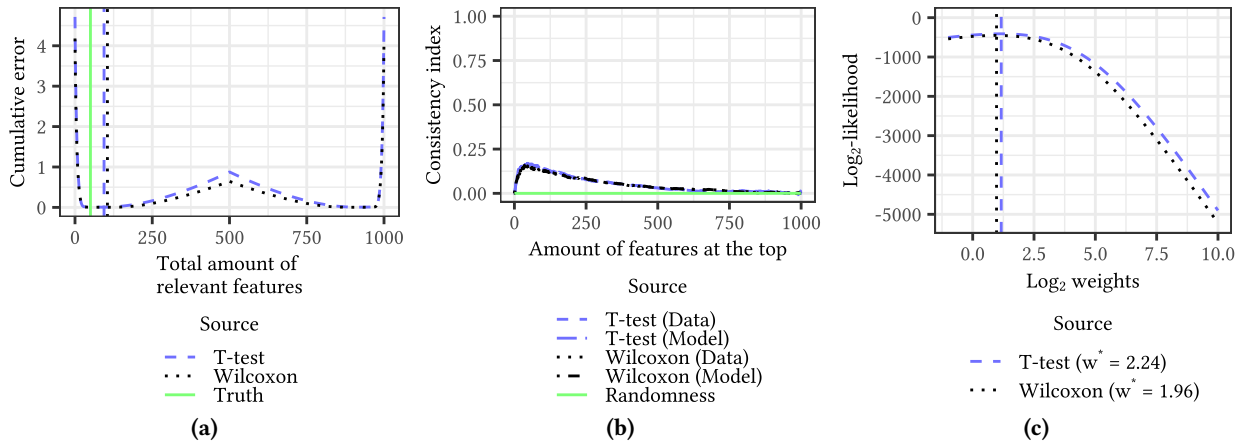


Figure 61: Error plot (61a), reproducibility plot (61b) and weight plot (61c) for the difficulty configuration 10, in the differences in location scenario

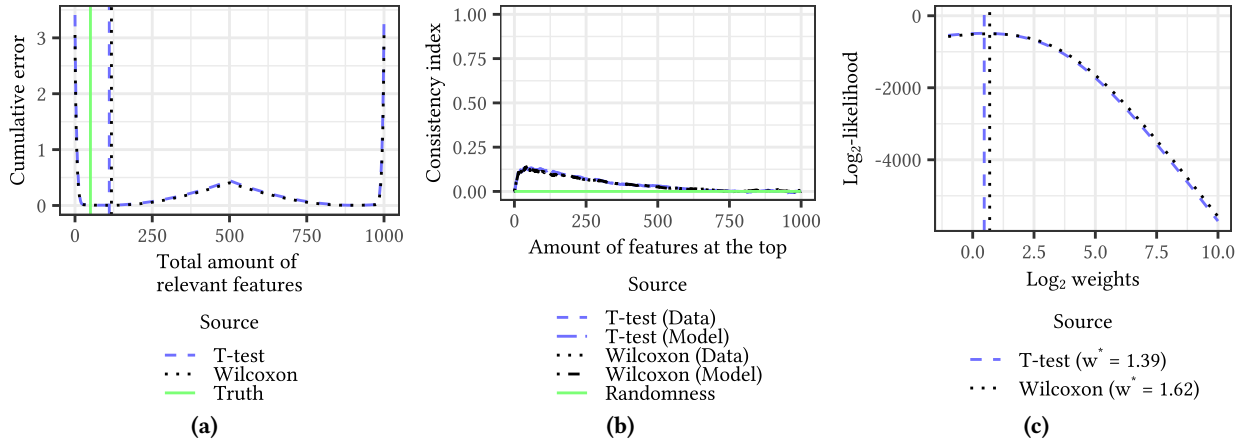


Figure 62: Error plot (62a), reproducibility plot (62b) and weight plot (62c) for the difficulty configuration 11, in the differences in location scenario

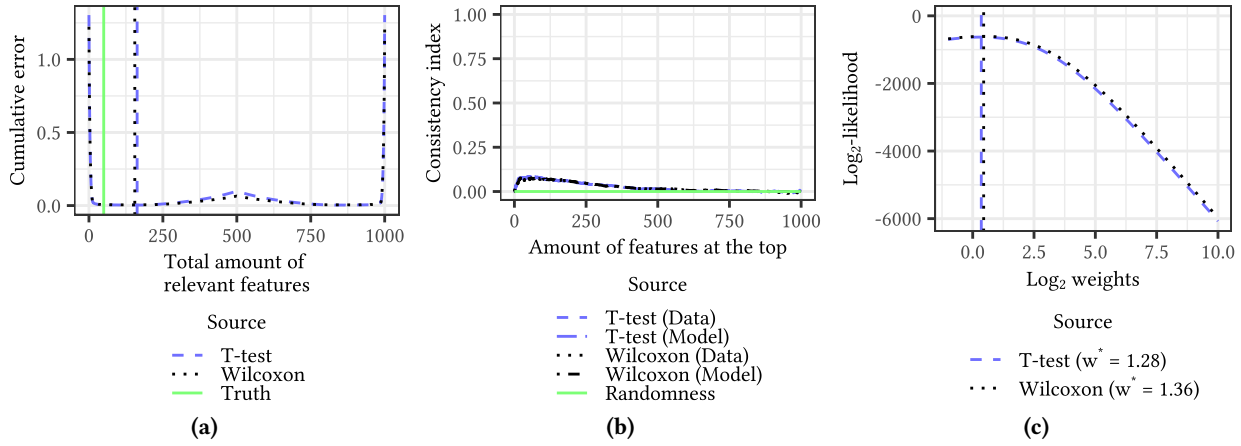


Figure 63: Error plot (63a), reproducibility plot (63b) and weight plot (63c) for the difficulty configuration 12, in the differences in location scenario

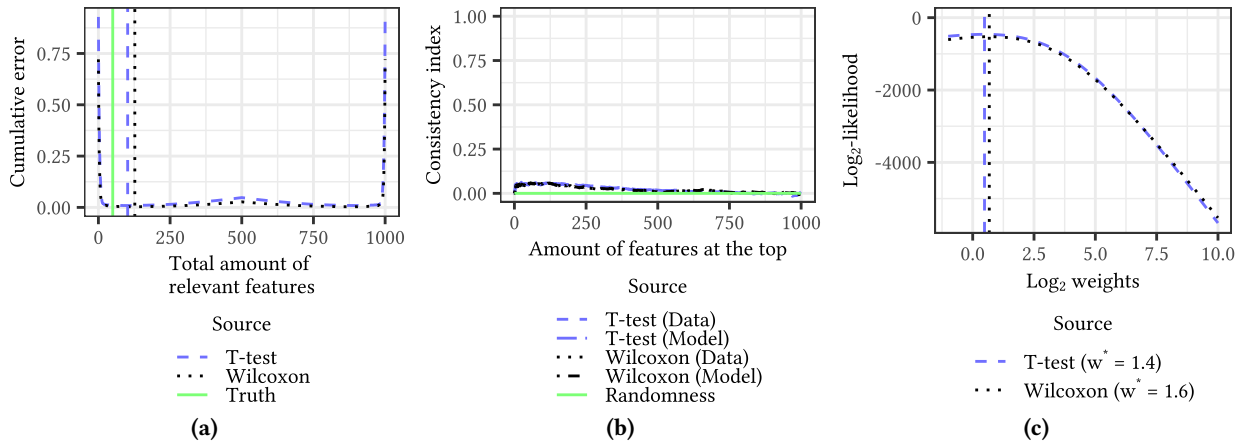


Figure 64: Error plot (64a), reproducibility plot (64b) and weight plot (64c) for the difficulty configuration 13, in the differences in location scenario

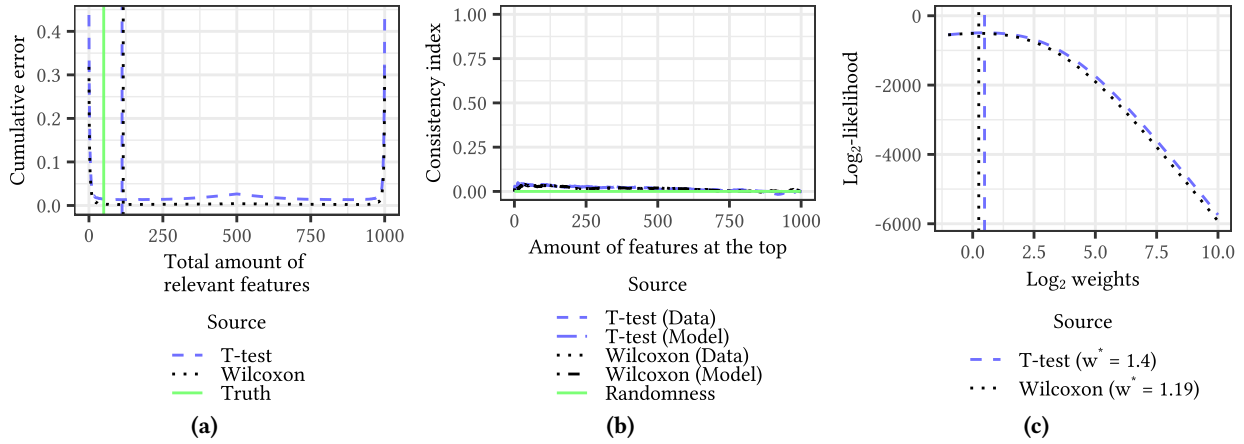


Figure 65: Error plot (65a), reproducibility plot (65b) and weight plot (65c) for the difficulty configuration 14, in the differences in location scenario

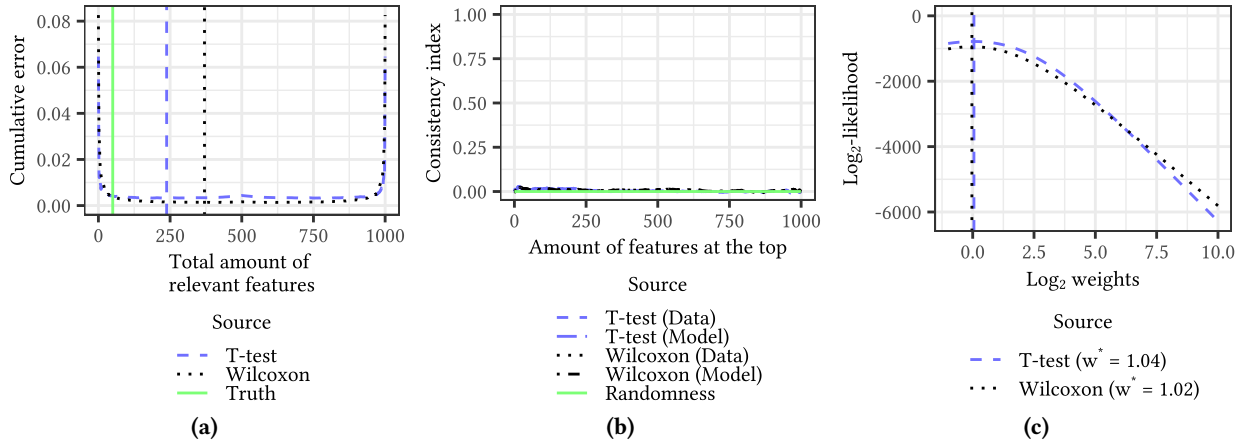


Figure 66: Error plot (66a), reproducibility plot (66b) and weight plot (66c) for the difficulty configuration 15, in the differences in location scenario

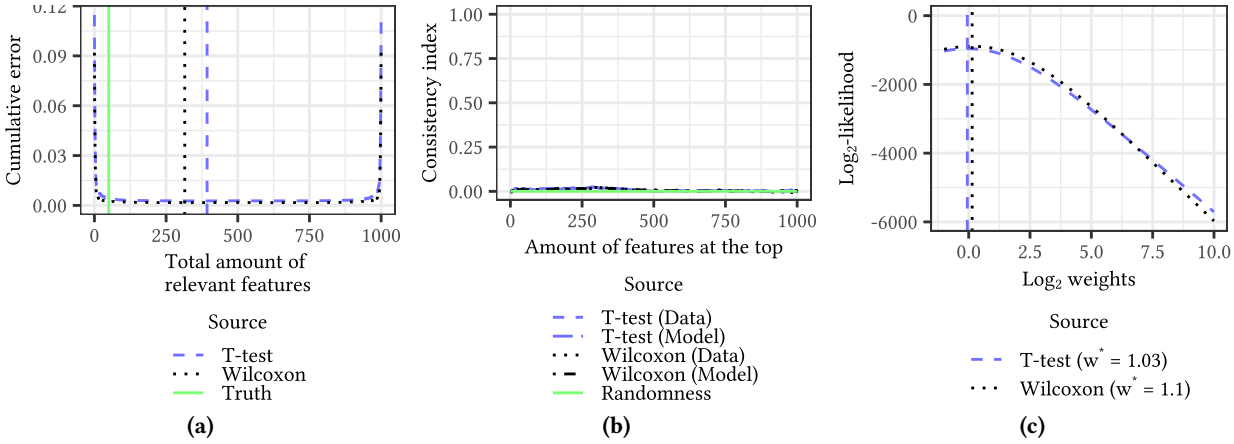


Figure 67: Error plot (67a), reproducibility plot (67b) and weight plot (67c) for the difficulty configuration 16, in the differences in location scenario

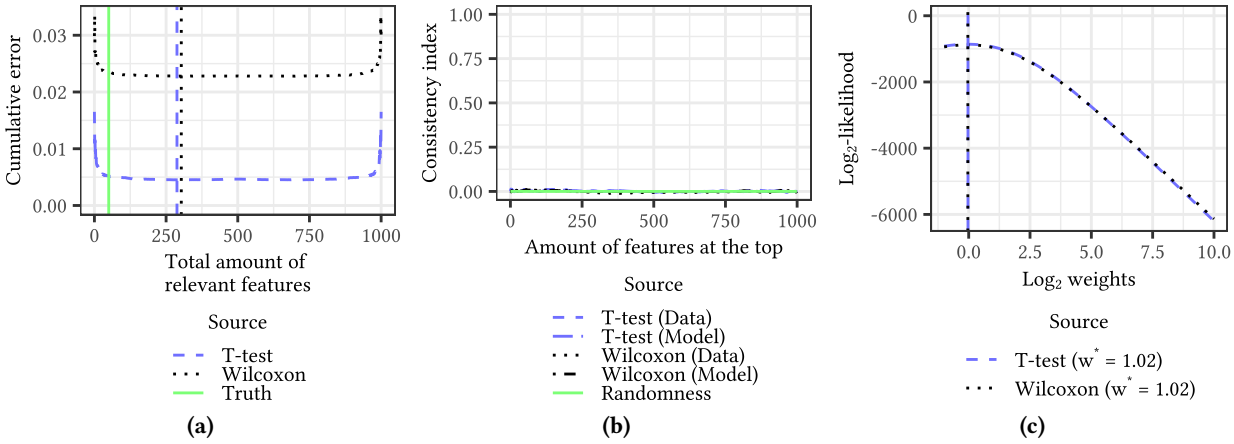


Figure 68: Error plot (68a), reproducibility plot (68b) and weight plot (68c) for the difficulty configuration 17, in the differences in location scenario

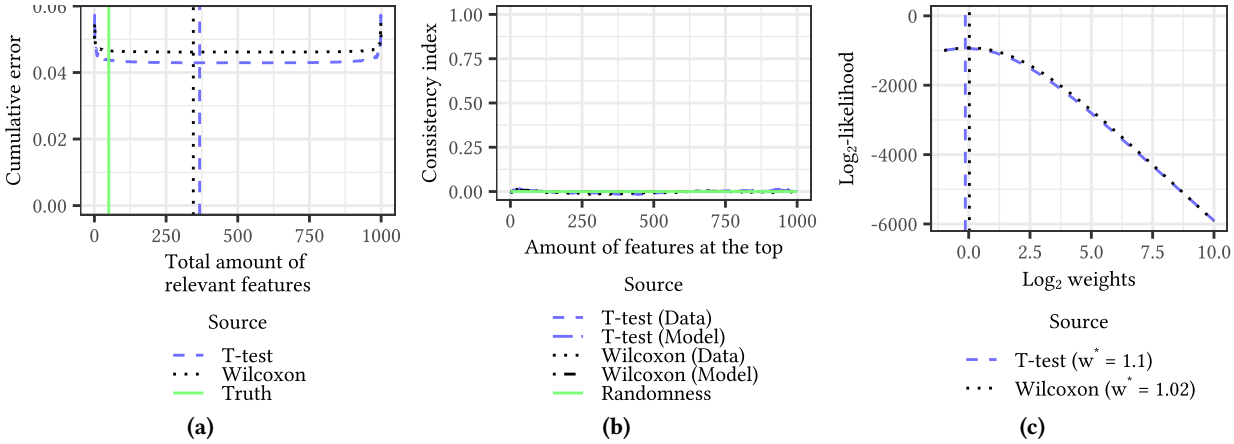


Figure 69: Error plot (69a), reproducibility plot (69b) and weight plot (69c) for the difficulty configuration 18, in the differences in location scenario

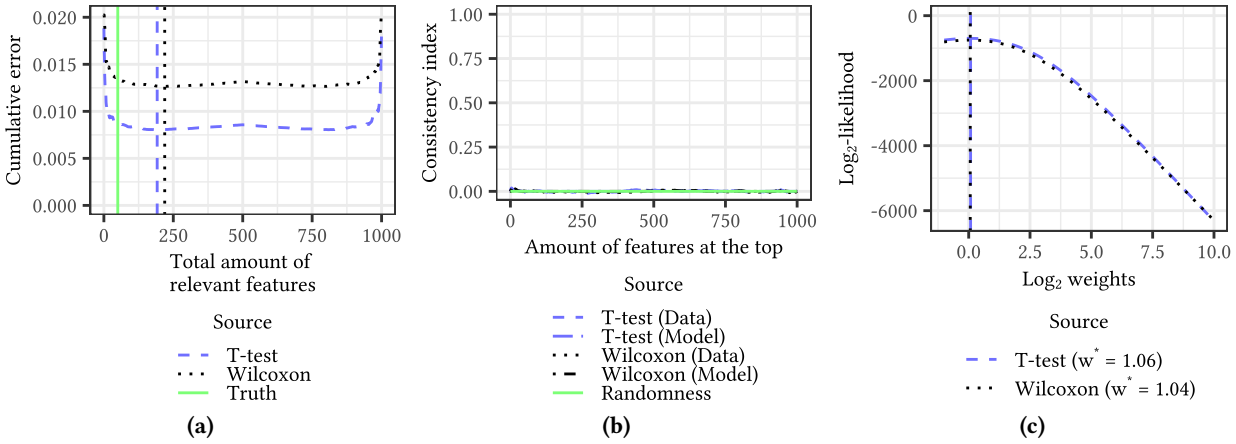


Figure 70: Error plot (70a), reproducibility plot (70b) and weight plot (70c) for the difficulty configuration 19, in the differences in location scenario

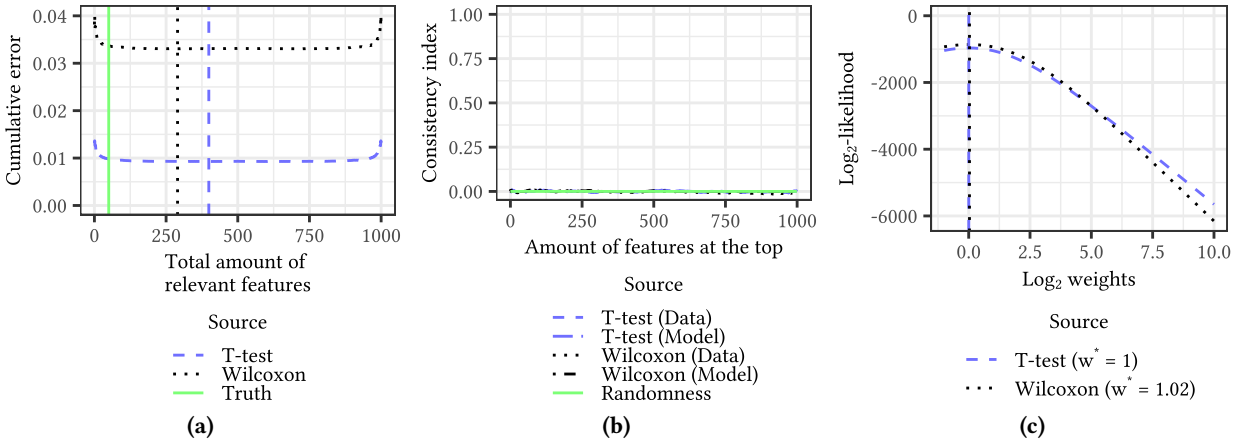


Figure 71: Error plot (71a), reproducibility plot (71b) and weight plot (71c) for the difficulty configuration 20, in the differences in location scenario

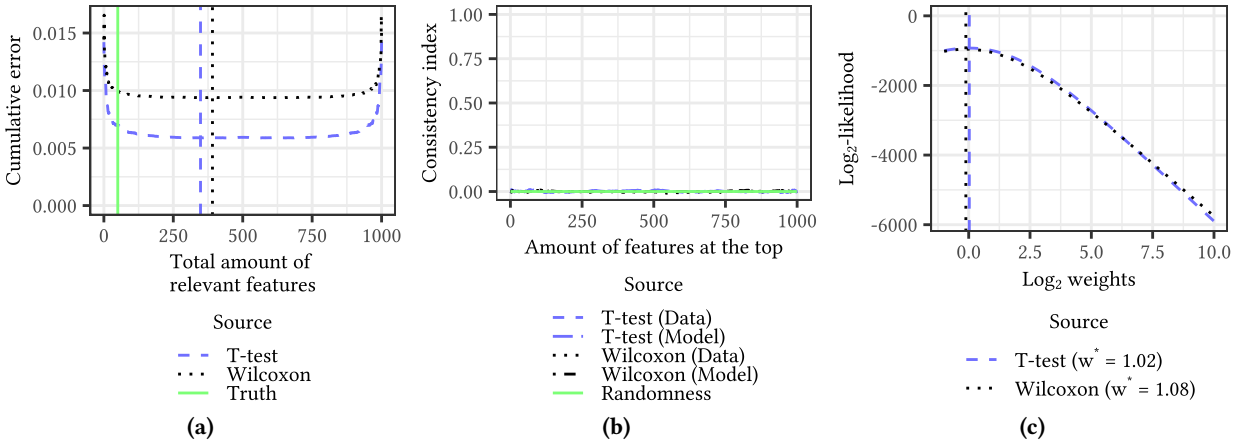


Figure 72: Error plot (72a), reproducibility plot (72b) and weight plot (72c) for the difficulty configuration 21, in the differences in location scenario

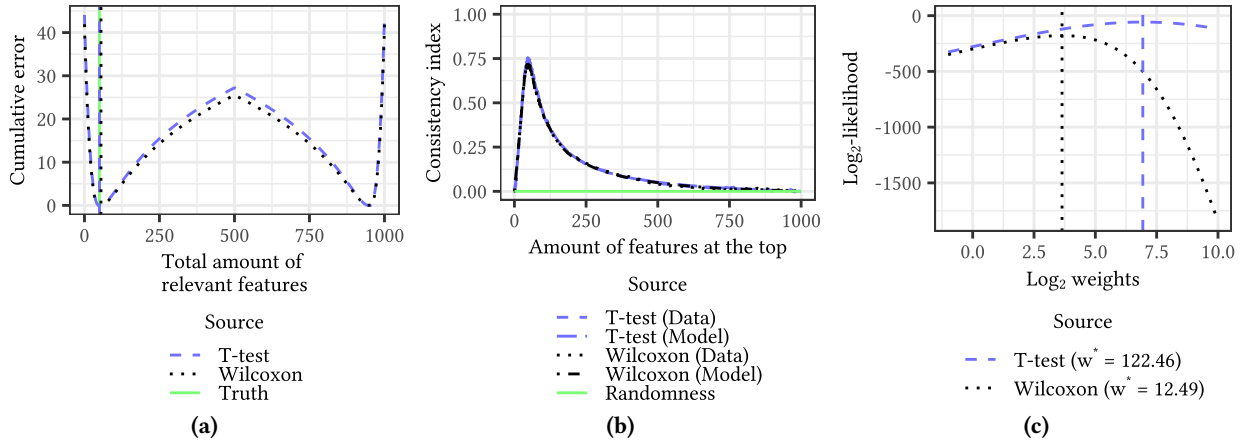


Figure 73: Error plot (73a), reproducibility plot (73b) and weight plot (73c) for the difficulty configuration 1, in the differences in both location and spread scenario

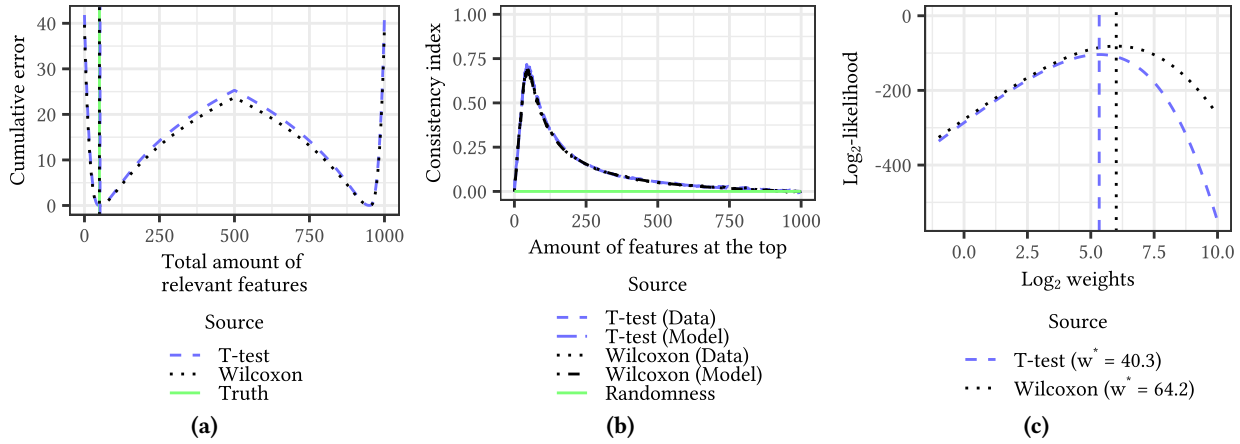


Figure 74: Error plot (74a), reproducibility plot (74b) and weight plot (74c) for the difficulty configuration 2, in the differences in both location and spread scenario

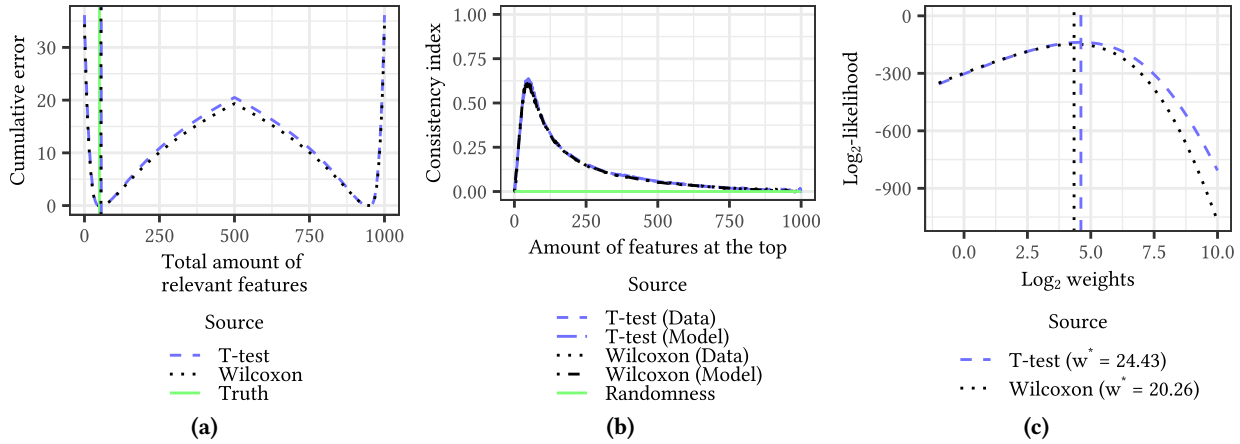


Figure 75: Error plot (75a), reproducibility plot (75b) and weight plot (75c) for the difficulty configuration 3, in the differences in both location and spread scenario

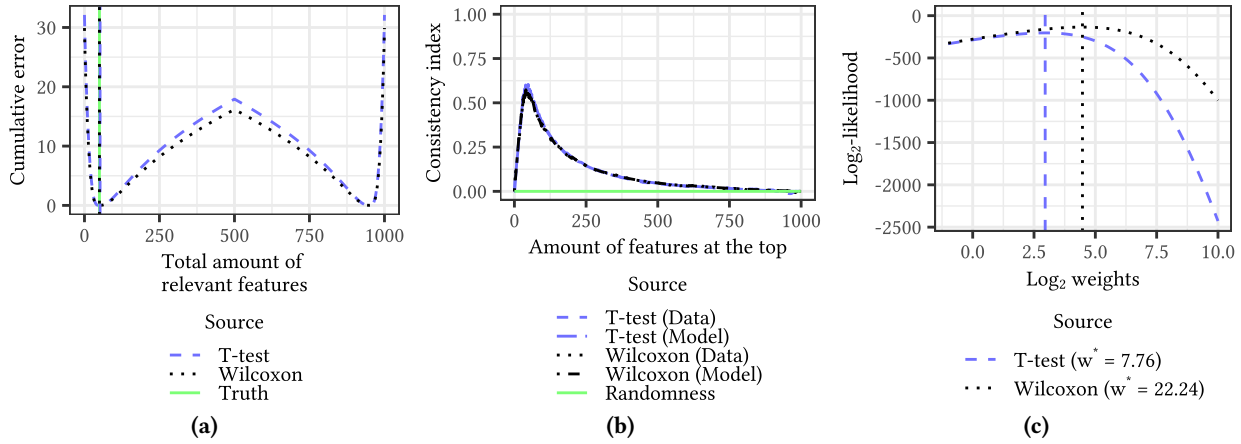


Figure 76: Error plot (76a), reproducibility plot (76b) and weight plot (76c) for the difficulty configuration 4, in the differences in both location and spread scenario

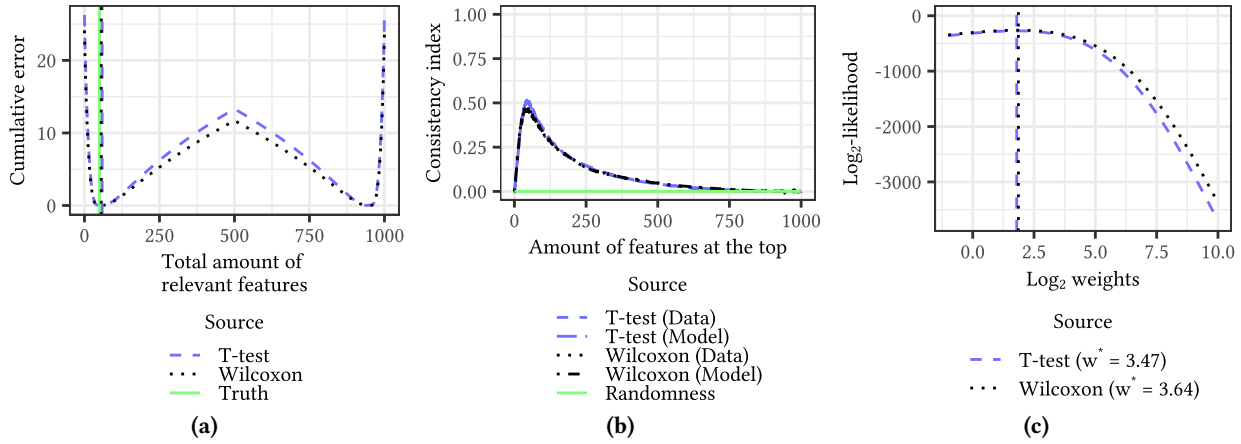


Figure 77: Error plot (77a), reproducibility plot (77b) and weight plot (77c) for the difficulty configuration 5, in the differences in both location and spread scenario

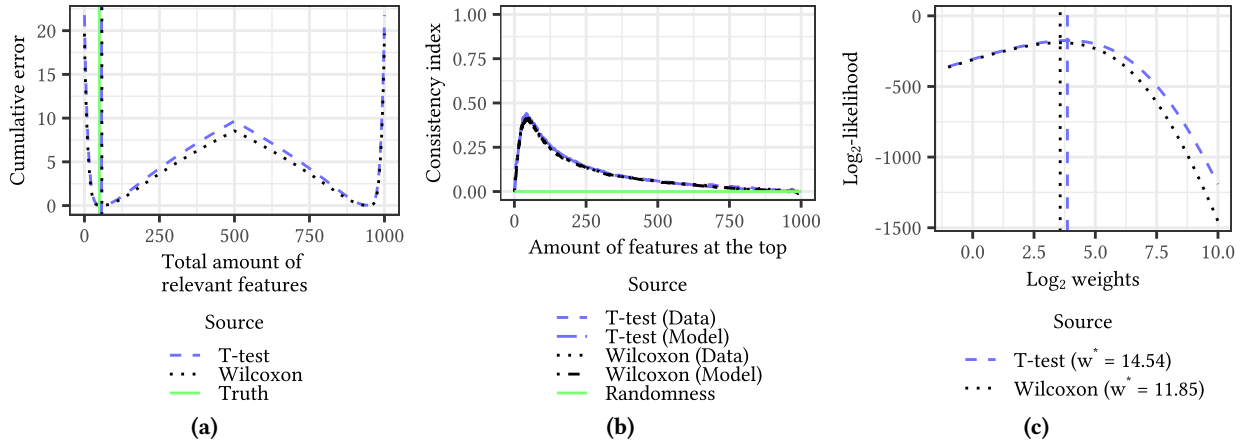


Figure 78: Error plot (78a), reproducibility plot (78b) and weight plot (78c) for the difficulty configuration 6, in the differences in both location and spread scenario

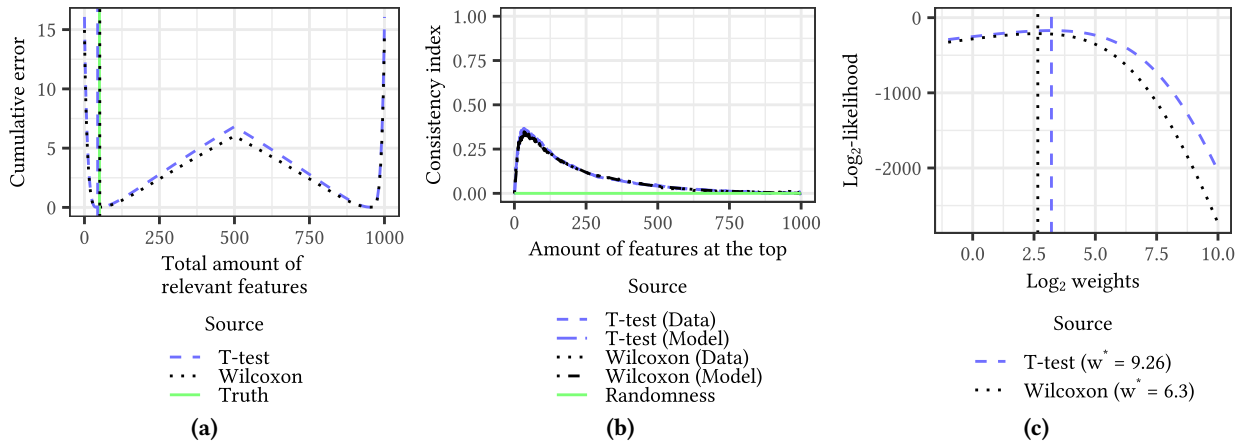


Figure 79: Error plot (79a), reproducibility plot (79b) and weight plot (79c) for the difficulty configuration 7, in the differences in both location and spread scenario

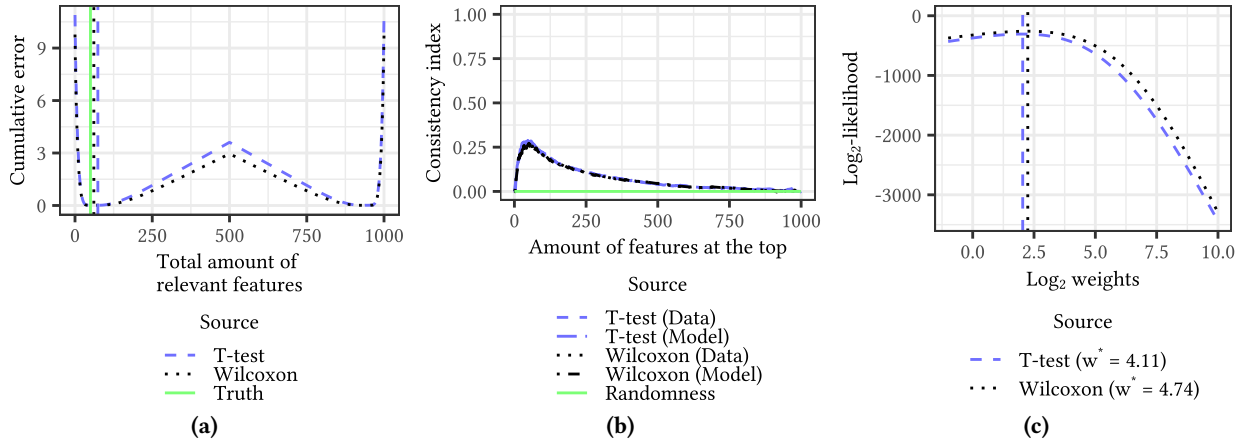


Figure 80: Error plot (80a), reproducibility plot (80b) and weight plot (80c) for the difficulty configuration 8, in the differences in both location and spread scenario

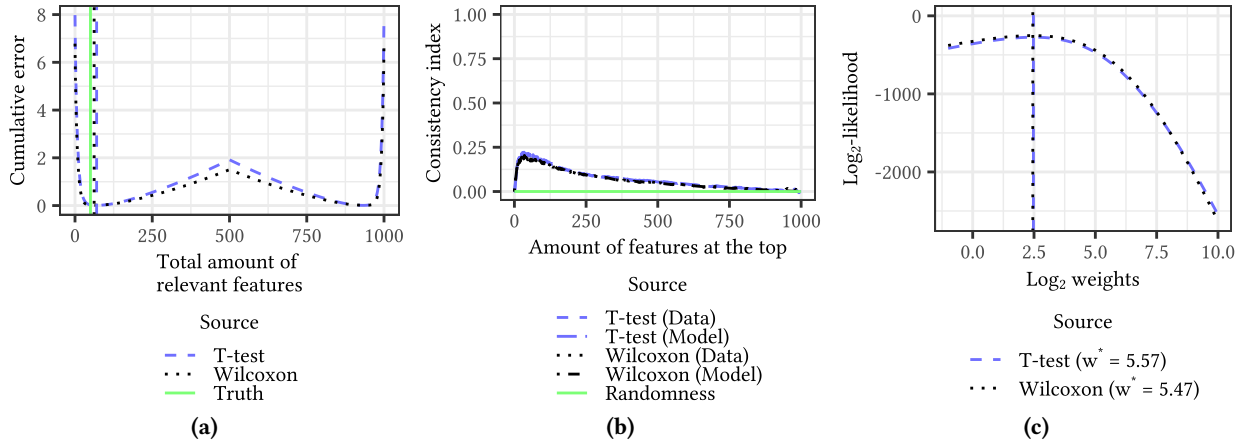


Figure 81: Error plot (81a), reproducibility plot (81b) and weight plot (81c) for the difficulty configuration 9, in the differences in both location and spread scenario

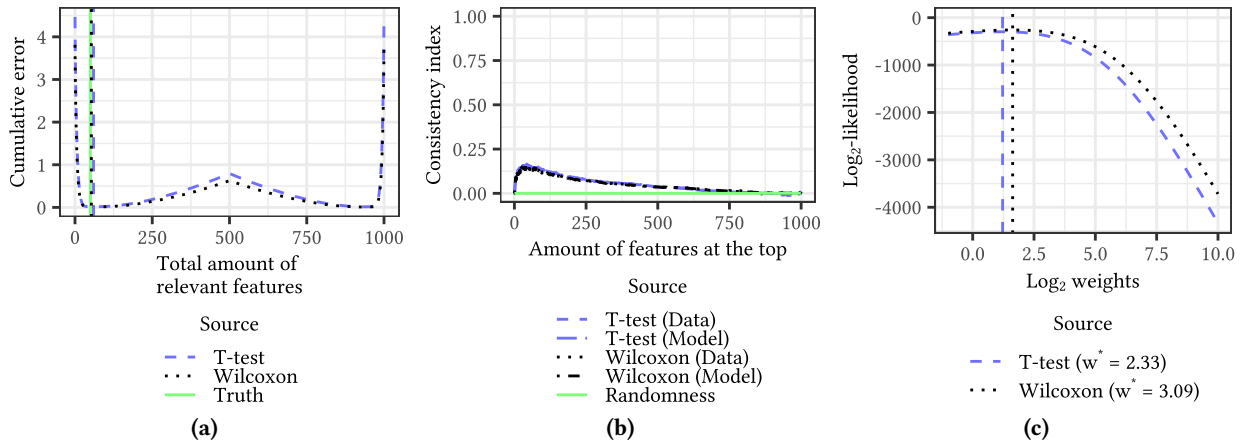


Figure 82: Error plot (82a), reproducibility plot (82b) and weight plot (82c) for the difficulty configuration 10, in the differences in both location and spread scenario

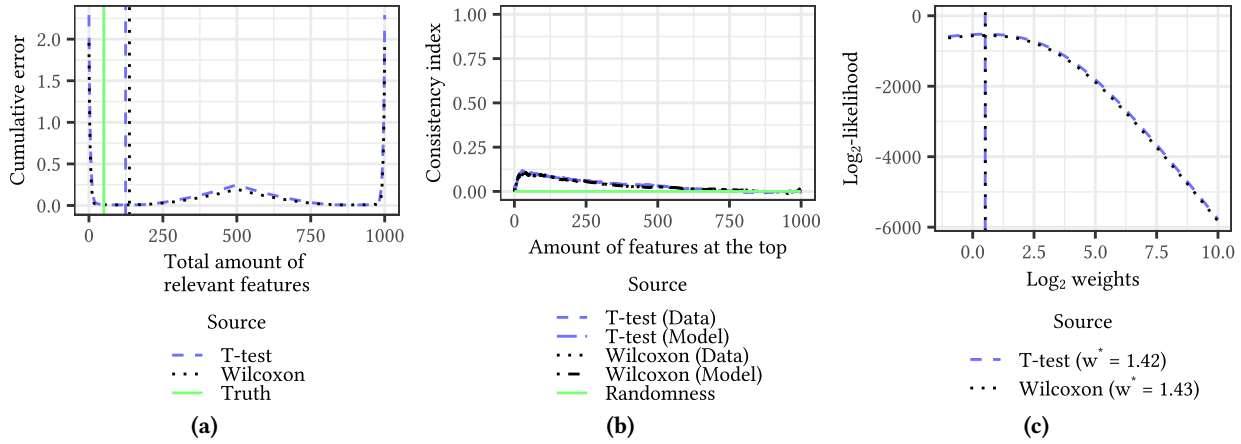


Figure 83: Error plot (83a), reproducibility plot (83b) and weight plot (83c) for the difficulty configuration 11, in the differences in both location and spread scenario

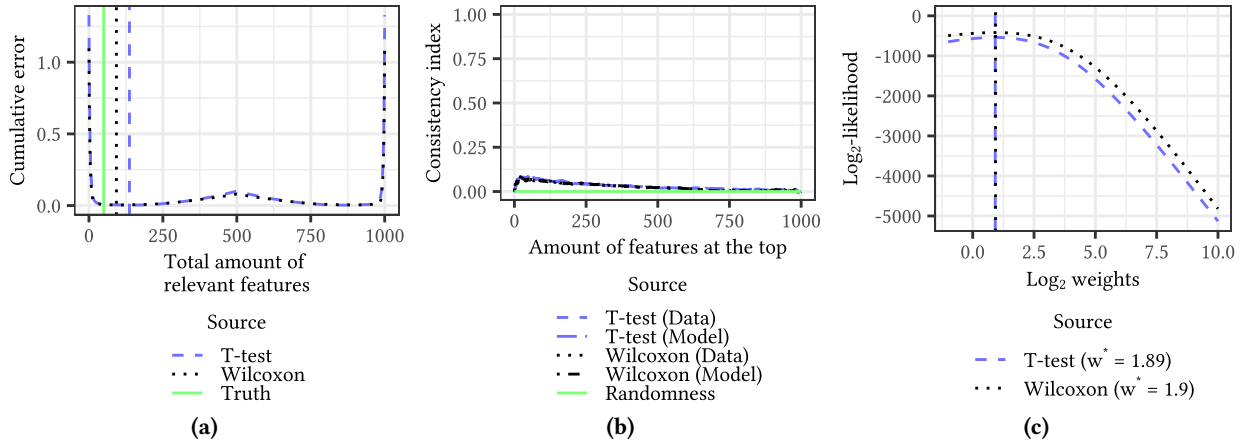


Figure 84: Error plot (84a), reproducibility plot (84b) and weight plot (84c) for the difficulty configuration 12, in the differences in both location and spread scenario

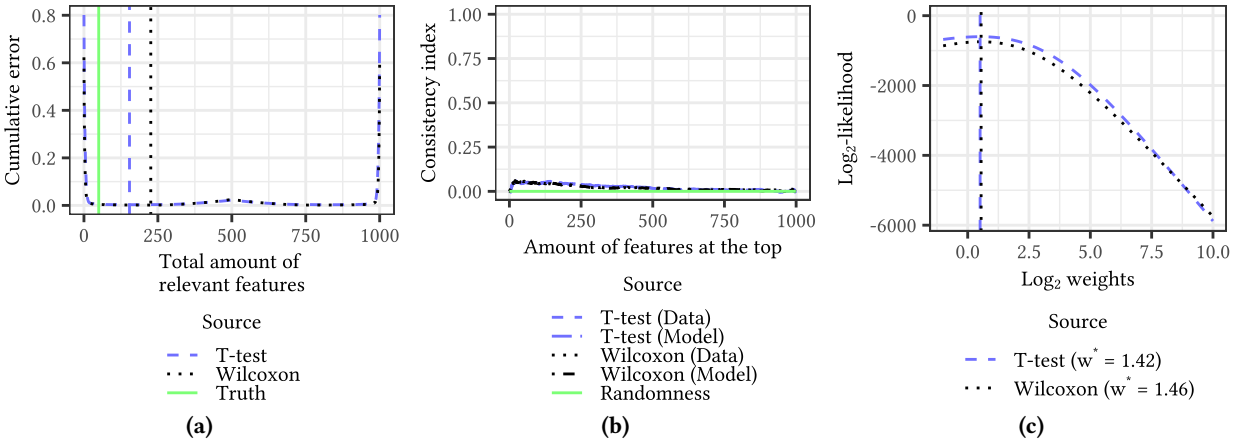


Figure 85: Error plot (85a), reproducibility plot (85b) and weight plot (85c) for the difficulty configuration 13, in the differences in both location and spread scenario

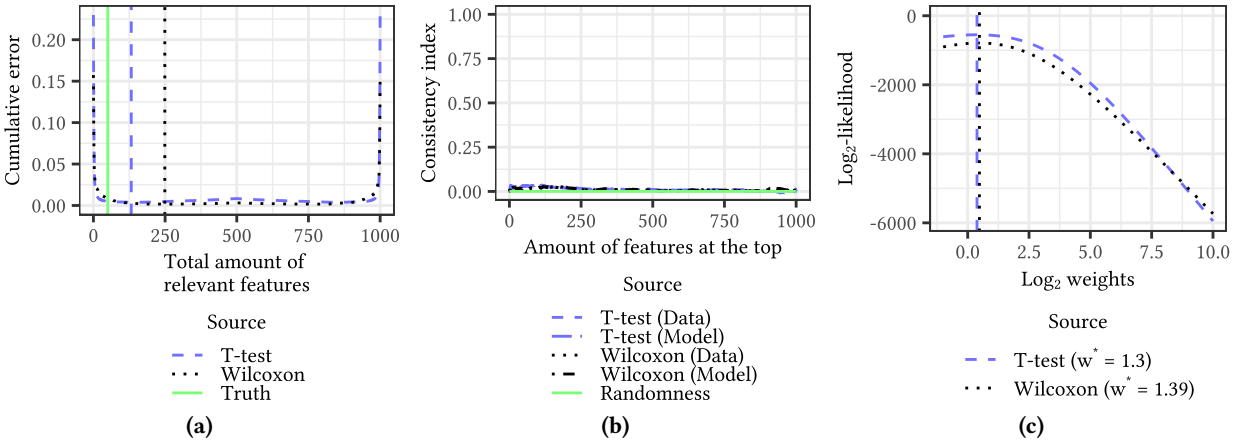


Figure 86: Error plot (86a), reproducibility plot (86b) and weight plot (86c) for the difficulty configuration 14, in the differences in both location and spread scenario

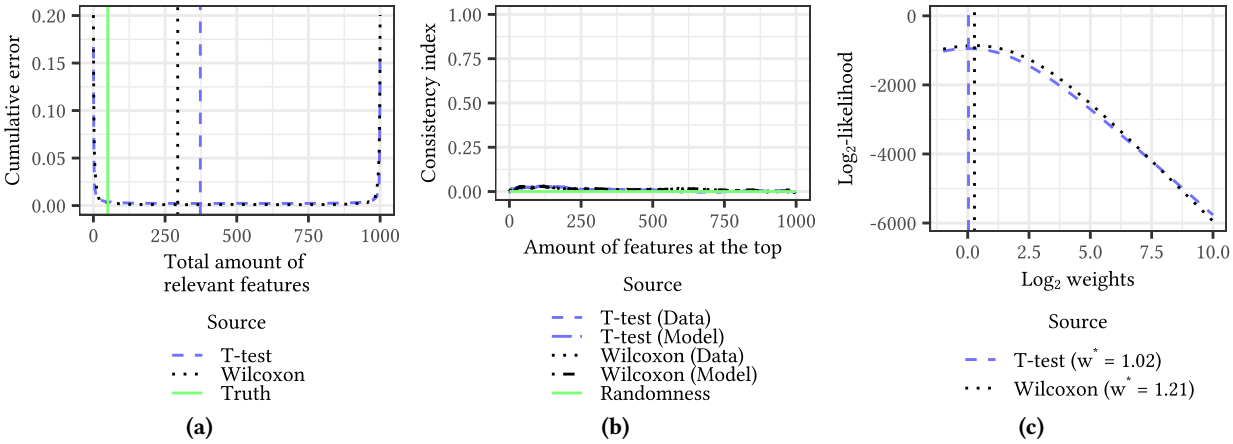


Figure 87: Error plot (87a), reproducibility plot (87b) and weight plot (87c) for the difficulty configuration 15, in the differences in both location and spread scenario

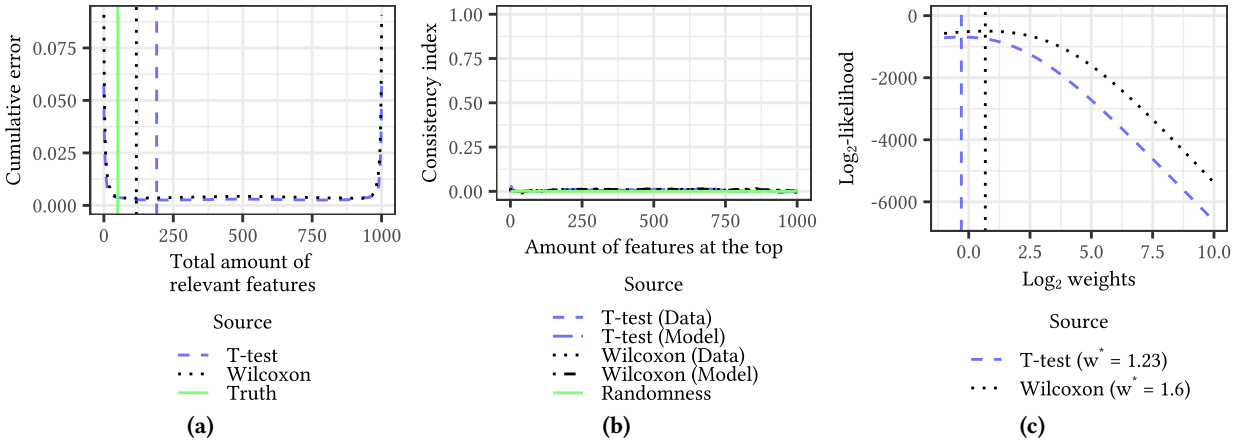


Figure 88: Error plot (88a), reproducibility plot (88b) and weight plot (88c) for the difficulty configuration 16, in the differences in both location and spread scenario

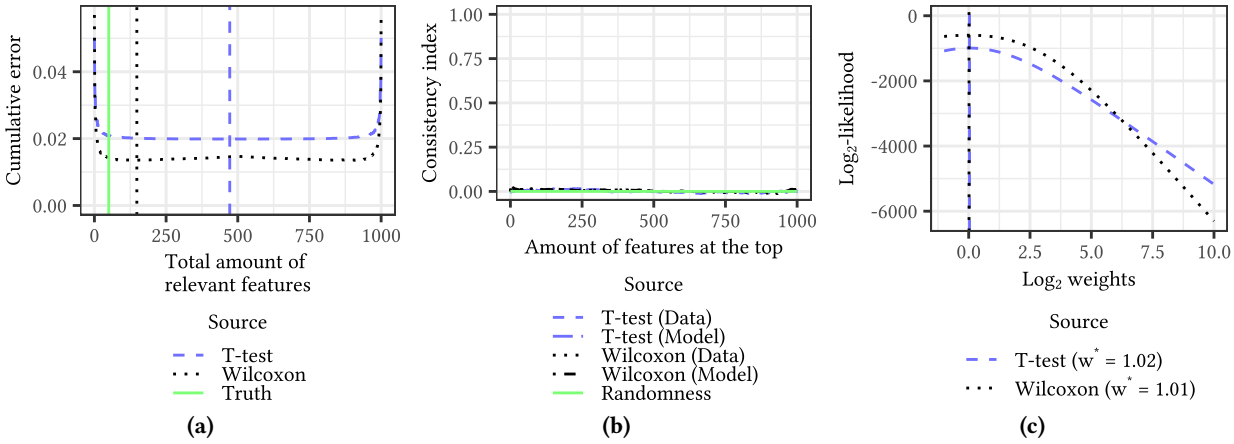


Figure 89: Error plot (89a), reproducibility plot (89b) and weight plot (89c) for the difficulty configuration 17, in the differences in both location and spread scenario

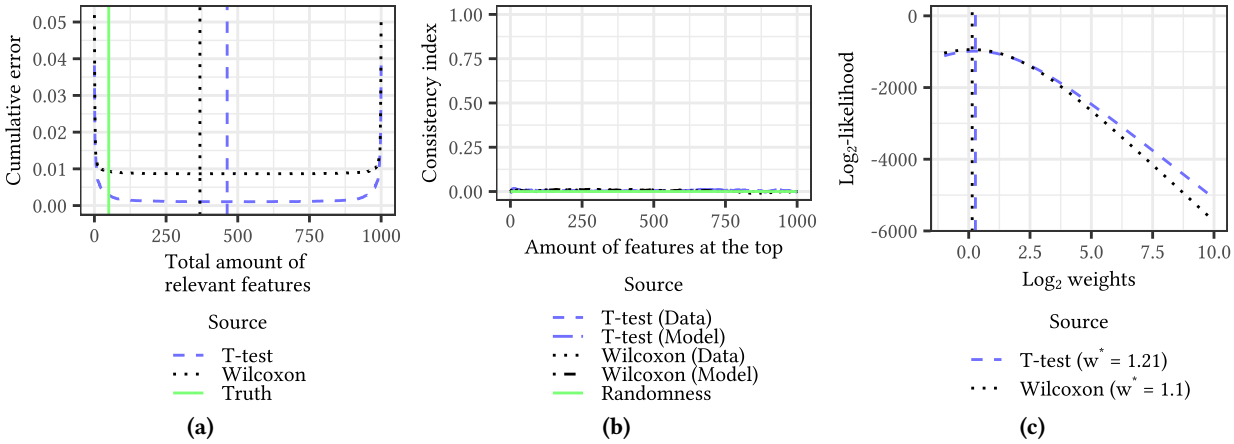


Figure 90: Error plot (90a), reproducibility plot (90b) and weight plot (90c) for the difficulty configuration 18, in the differences in both location and spread scenario

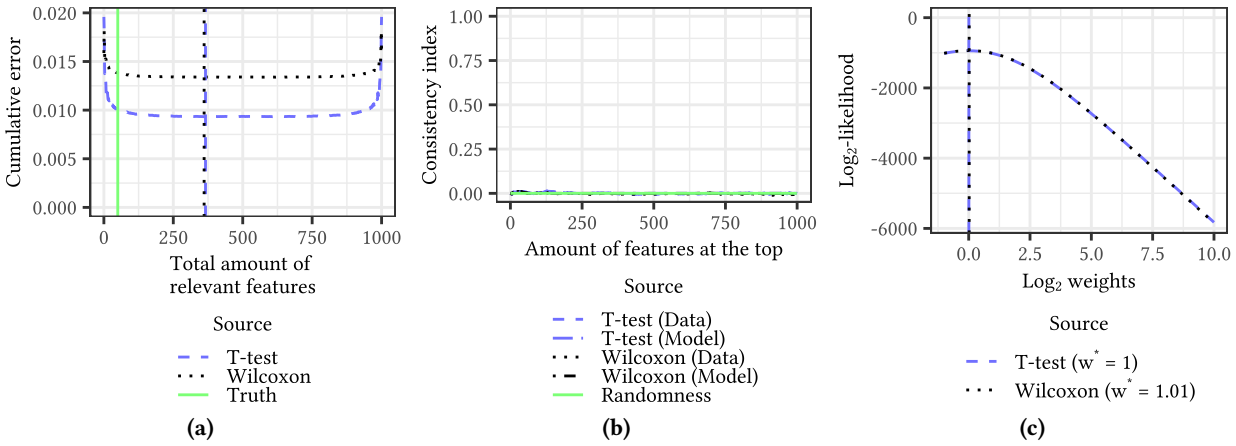


Figure 91: Error plot (91a), reproducibility plot (91b) and weight plot (91c) for the difficulty configuration 19, in the differences in both location and spread scenario

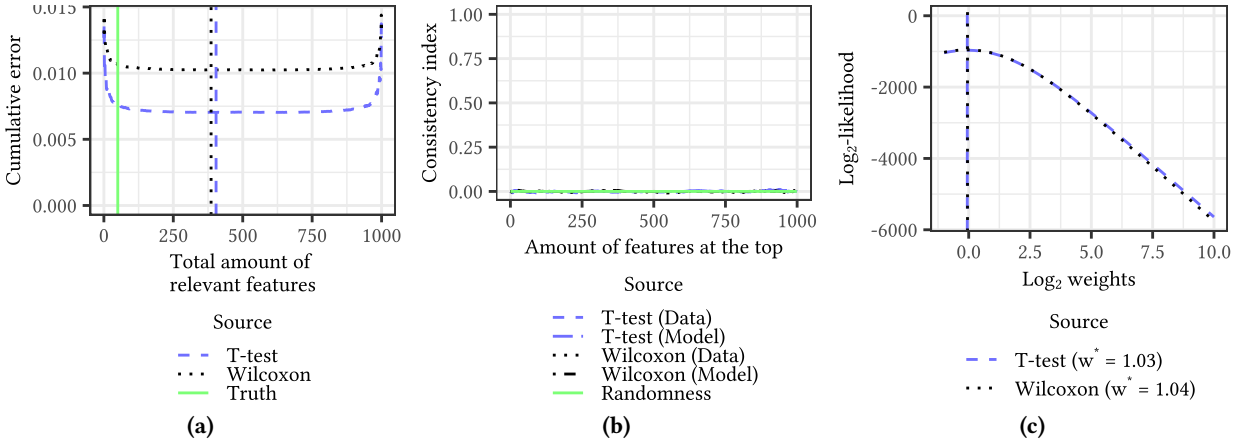


Figure 92: Error plot (92a), reproducibility plot (92b) and weight plot (92c) for the difficulty configuration 20, in the differences in both location and spread scenario

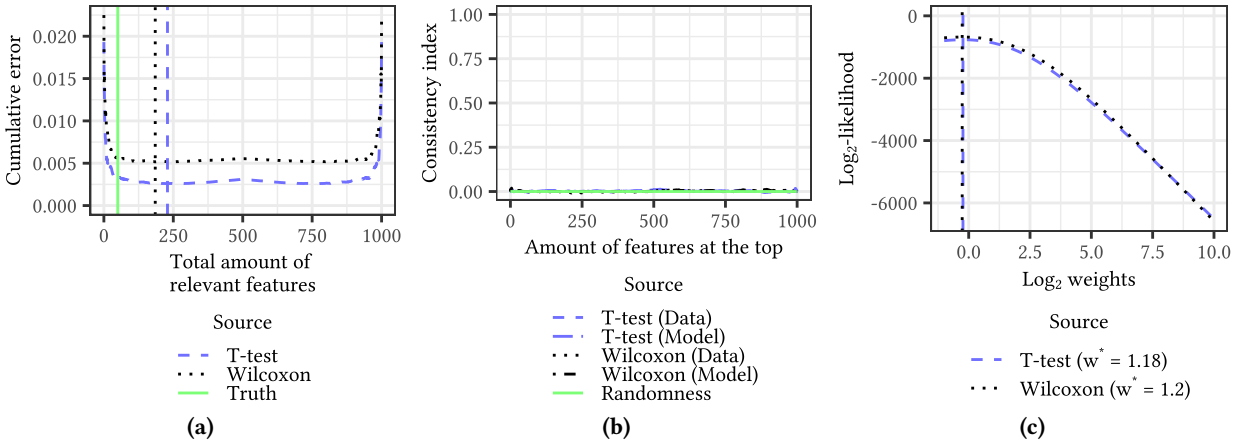


Figure 93: Error plot (93a), reproducibility plot (93b) and weight plot (93c) for the difficulty configuration 21, in the differences in both location and spread scenario

Table 2: w^* and AUC values when the relevant features show differences only in location

Difficulty	Method	w^*	Average data AUC	Model AUC
1	t-test	11.518	0.99148	0.97250
1	Wilcoxon test	5.849	0.98946	0.95259
2	t-test	53.975	0.98500	0.98882
2	Wilcoxon test	14.604	0.98267	0.97205
3	t-test	16.069	0.98233	0.96627
3	Wilcoxon test	12.418	0.97970	0.95985
4	t-test	6.450	0.97515	0.93749
4	Wilcoxon test	12.113	0.97090	0.95435
5	t-test	13.050	0.96332	0.95136
5	Wilcoxon test	8.872	0.95915	0.93693
6	t-test	7.953	0.94897	0.92738
6	Wilcoxon test	6.647	0.94338	0.91295
7	t-test	4.642	0.93835	0.89594
7	Wilcoxon test	2.787	0.93154	0.87626
8	t-test	5.345	0.91565	0.88733
8	Wilcoxon test	3.016	0.90990	0.85423
9	t-test	4.595	0.89733	0.85790
9	Wilcoxon test	3.485	0.88775	0.82226
10	t-test	2.240	0.85830	0.76064
10	Wilcoxon test	1.962	0.85010	0.74880
11	t-test	1.386	0.83564	0.69693
11	Wilcoxon test	1.616	0.82877	0.70229
12	t-test	1.275	0.79122	0.63219
12	Wilcoxon test	1.363	0.78047	0.64459
13	t-test	1.397	0.74516	0.66799
13	Wilcoxon test	1.599	0.73620	0.65572
14	t-test	1.398	0.71663	0.65058
14	Wilcoxon test	1.187	0.70986	0.62836
15	t-test	1.036	0.65179	0.54314
15	Wilcoxon test	1.021	0.64477	0.54778
16	t-test	1.035	0.61097	0.52834
16	Wilcoxon test	1.104	0.60878	0.56503
17	t-test	1.023	0.58113	0.51393
17	Wilcoxon test	1.019	0.57768	0.51062
18	t-test	1.099	0.54310	0.50799
18	Wilcoxon test	1.020	0.54333	0.50883
19	t-test	1.055	0.51729	0.52850
19	Wilcoxon test	1.045	0.51828	0.51656
20	t-test	1.002	0.50171	0.50972
20	Wilcoxon test	1.023	0.50197	0.51516
21	t-test	1.017	0.50194	0.52346
21	Wilcoxon test	1.076	0.50303	0.50471

Table 3: w^* and AUC values when the relevant features show differences both in location and spread

Difficulty	Method	w^*	Average data AUC	Model AUC
1	t-test	122.460	0.99407	0.99536
1	Wilcoxon test	12.492	0.99164	0.96449
2	t-test	40.297	0.99137	0.98715
2	Wilcoxon test	64.199	0.98905	0.99124
3	t-test	24.434	0.98295	0.97447
3	Wilcoxon test	20.265	0.97917	0.97005
4	t-test	7.758	0.97871	0.95781
4	Wilcoxon test	22.241	0.97396	0.97343
5	t-test	3.470	0.96880	0.90336
5	Wilcoxon test	3.639	0.96373	0.92110
6	t-test	14.536	0.95099	0.95008
6	Wilcoxon test	11.848	0.94507	0.93907
7	t-test	9.258	0.93612	0.94695
7	Wilcoxon test	6.296	0.92903	0.91514
8	t-test	4.110	0.91651	0.84912
8	Wilcoxon test	4.742	0.90731	0.87484
9	t-test	5.565	0.88543	0.86763
9	Wilcoxon test	5.469	0.87457	0.86828
10	t-test	2.328	0.85350	0.81743
10	Wilcoxon test	3.093	0.84282	0.84080
11	t-test	1.425	0.81481	0.68420
11	Wilcoxon test	1.429	0.80419	0.67806
12	t-test	1.891	0.76569	0.68487
12	Wilcoxon test	1.905	0.75853	0.69762
13	t-test	1.420	0.73233	0.65252
13	Wilcoxon test	1.461	0.72517	0.62383
14	t-test	1.301	0.68675	0.61855
14	Wilcoxon test	1.393	0.67817	0.58700
15	t-test	1.024	0.64637	0.54806
15	Wilcoxon test	1.211	0.64123	0.59103
16	t-test	1.228	0.60370	0.57824
16	Wilcoxon test	1.600	0.59955	0.61535
17	t-test	1.024	0.57310	0.52380
17	Wilcoxon test	1.010	0.56957	0.55216
18	t-test	1.209	0.53356	0.55400
18	Wilcoxon test	1.104	0.53449	0.55425
19	t-test	1.005	0.51847	0.50532
19	Wilcoxon test	1.014	0.51820	0.50659
20	t-test	1.033	0.50659	0.50427
20	Wilcoxon test	1.041	0.50810	0.50814
21	t-test	1.178	0.49810	0.54628
21	Wilcoxon test	1.204	0.49794	0.53650

- Ovarian database [6]: This dataset contains 27578 features and 540 instances. The features consists of β-values that denote the methylation level of different CpG-sites on the peripheral whole blood of 540 postmenopausal women. The instances of this dataset correspond to 274 healthy women, 131 women with ovarian cancer yet to be treated and 135 women with ovarian cancer already treated, i.e., 274 controls and 266 cases. The groups are age-matched and the range of ages covered by the database is from 49 to 91 years.

2.6 Preprocessing of the real databases

We divide this section into two parts. The first one is dedicated to the preprocessing applied to the databases extracted from the UCI repository. The second one is dedicated to the preprocessing applied to the ovarian cancer database.

2.6.1 Preprocessing of the UCI repository databases

The preprocessing applied to the selected databases from the UCI repository consists of the following step:

1. The features that have missing values for more than 50% of the individuals are removed[1].

2.6.2 Preprocessing of the ovarian cancer database

The preprocessing we applied to the ovarian cancer database is based on what was done by Wang et al [7]. The preprocessing consists of applying the following steps sequentially to the matrix of β-values available in the GEO database:

1. Among all the ovarian cancer cases, only those who gave their blood at the time of their diagnosis prior to treatment have been used.
2. Samples whose bisulfite conversion efficiencies are too low (< 4000) have been removed.
3. Data from batches 10-12 have been removed.
4. In order to remove outlier samples, for each sample, we have computed the average of the distances of all the values (using raw total intensities) of its CpG sites regarding their median values across samples. Then, all the generated averages are used to calculate the range $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$. Finally, all those samples whose averages are not within that range are removed.
5. All those individuals that do not cover at least 95% of the CpG sites with a detection p-value smaller than 0.05 are removed.
6. All the CpG sites whose detection p-values are not below 0.05 in all samples are removed.
7. All the CpG sites that do not have numeric values (e.g., NA values) for at least 50 individuals per group are removed.
8. The CpG sites that have missing values for more than 99.55% of the individuals are removed[2].

[1] This preprocessing step was included in order to enable the computation of the RBFS method based on the coefficients of a linear SVM.

[2] This preprocessing step was included in order to enable the computation of the RBFS method based on the coefficients of a linear SVM.

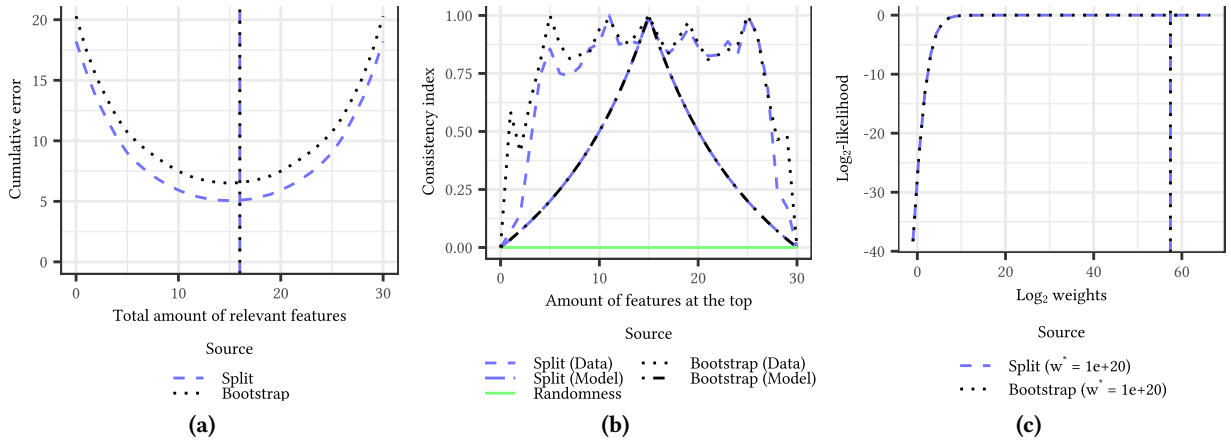


Figure 94: Error plot (94a), reproducibility plot (94b) and weight plot (94c) when the RBFS algorithm based on the mutual information is applied to the breast cancer database

2.7 Stratification of the ovarian cancer database

In the following lines, the stratification procedure is explained for the different sampling procedures:

- Splitting in halves: In this sampling procedure, for each group, the individuals are sorted according to their age. Then, for each group, each of its odd individuals is assigned randomly either to belong to $\boldsymbol{D}^{(1)}$ or to belong to $\boldsymbol{D}^{(2)}$. Finally, for each group, each of its even individuals is assigned to $\boldsymbol{D}^{(1)}$ if its previous odd individual was assigned to $\boldsymbol{D}^{(2)}$ or is assigned to $\boldsymbol{D}^{(2)}$ if its previous odd individual was assigned to $\boldsymbol{D}^{(1)}$.
- Bootstrapping: In this sampling procedure, the individuals for both $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$ are directly randomly sampled with replacement from $\boldsymbol{D}$.

2.8 Plots and tables of the experimentation with real data

In Figures 94 to 113 the plots and tables of the experimentation with real data can be seen.

In Tables 4 and 5 the weights and AUC values of the experimentation with real data can be seen.

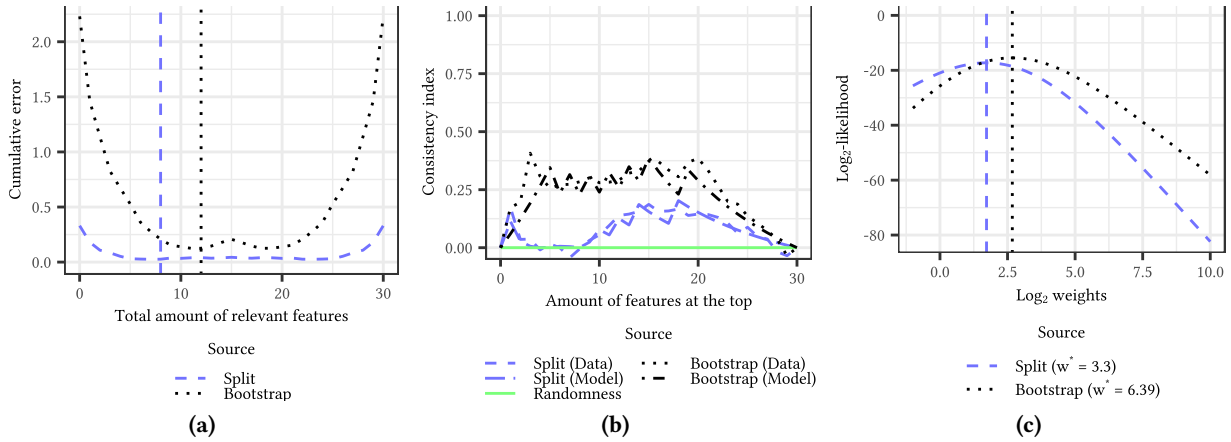


Figure 95: Error plot (95a), reproducibility plot (95b) and weight plot (95c) when the RBFS algorithm based on the linear SVM is applied to the breast cancer database

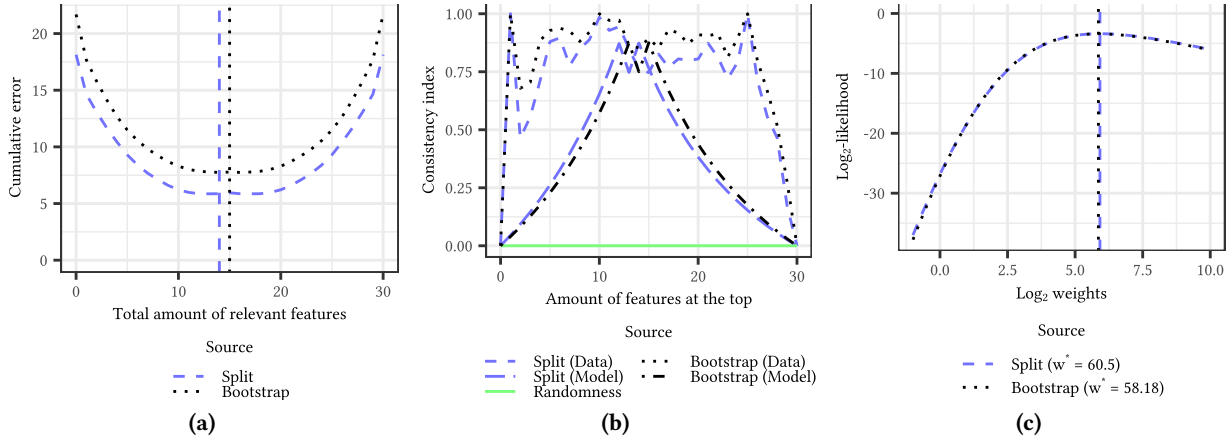


Figure 96: Error plot (96a), reproducibility plot (96b) and weight plot (96c) when the RBFS algorithm based on the t-test is applied to the breast cancer database

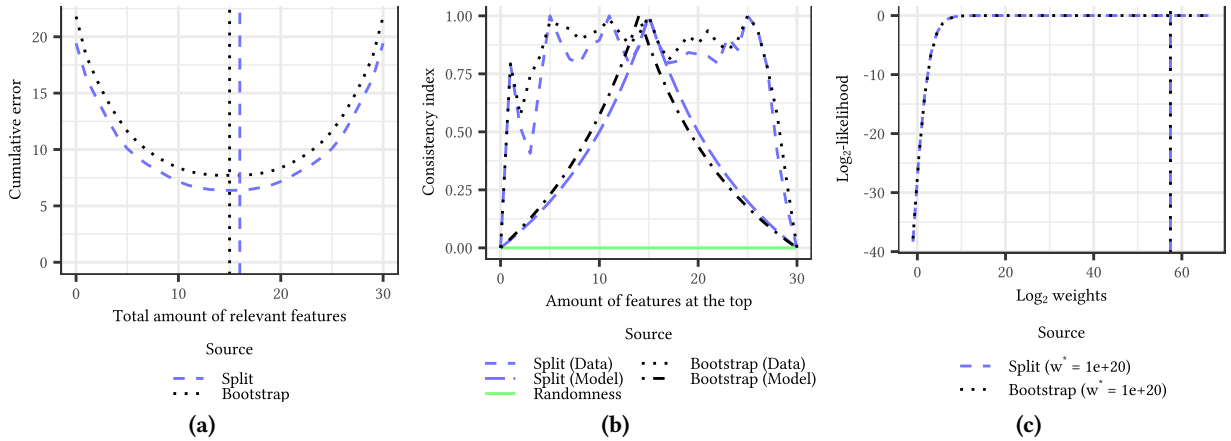


Figure 97: Error plot (97a), reproducibility plot (97b) and weight plot (97c) when the RBFS algorithm based on the Wilcoxon test is applied to the breast cancer database

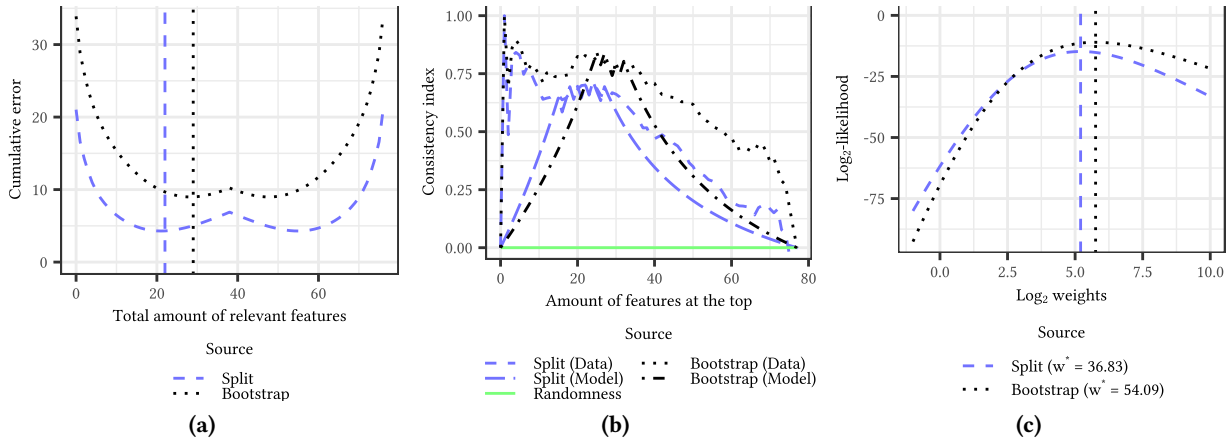


Figure 98: Error plot (98a), reproducibility plot (98b) and weight plot (98c) when the RBFS algorithm based on the mutual information is applied to the mice database

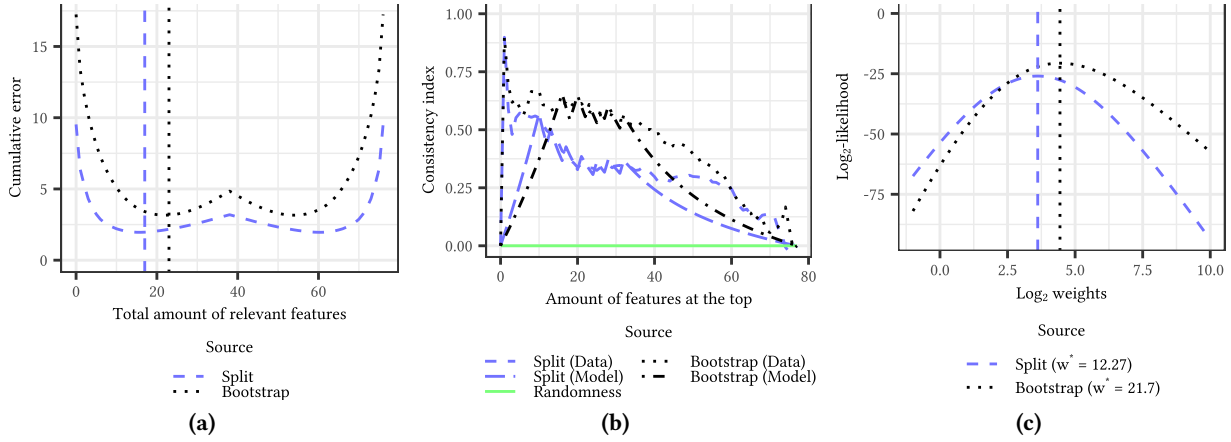


Figure 99: Error plot (99a), reproducibility plot (99b) and weight plot (99c) when the RBFS algorithm based on the linear SVM is applied to the mice database

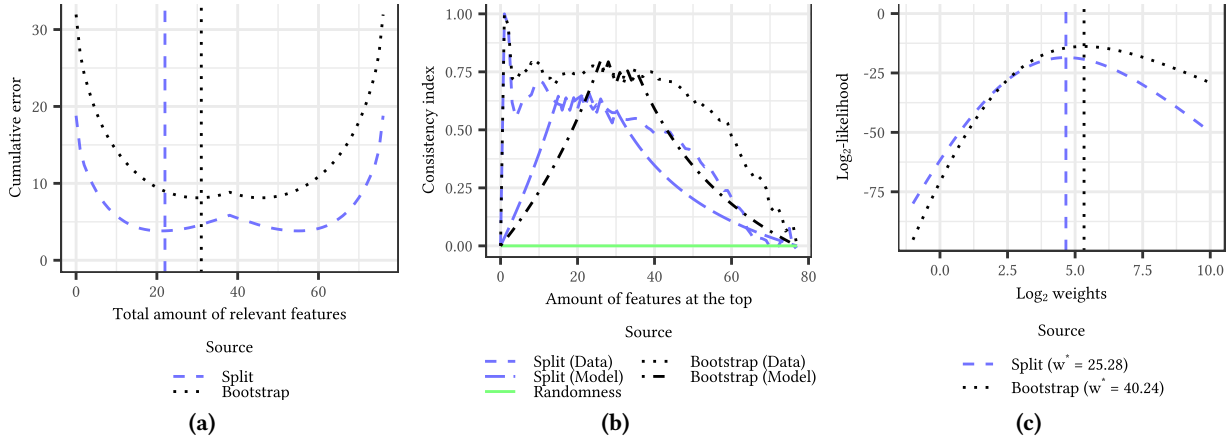


Figure 100: Error plot (100a), reproducibility plot (100b) and weight plot (100c) when the RBFS algorithm based on the t-test is applied to the mice database

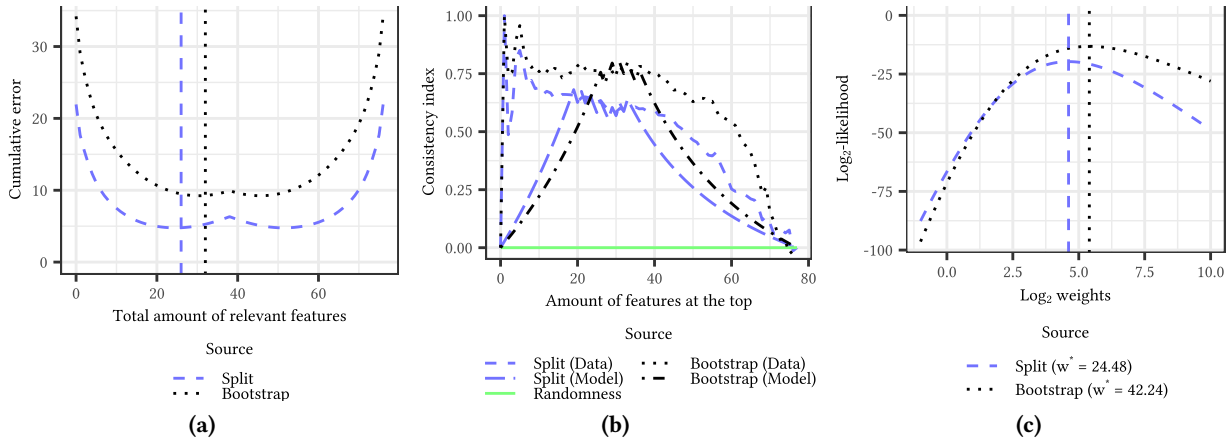


Figure 101: Error plot (101a), reproducibility plot (101b) and weight plot (101c) when the RBFS algorithm based on the Wilcoxon test is applied to the mice database

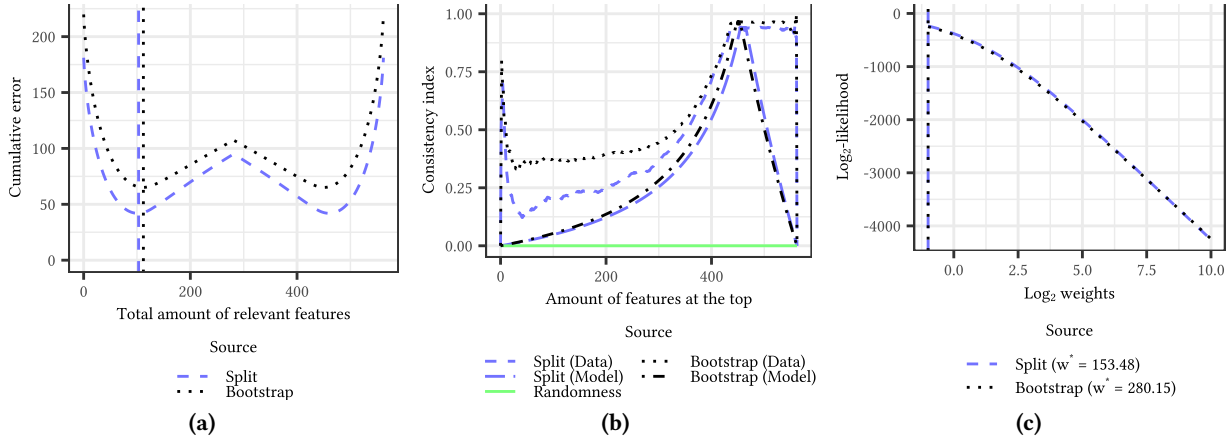


Figure 102: Error plot (102a), reproducibility plot (102b) and weight plot (102c) when the RBFS algorithm based on the mutual information is applied to the SECOM database

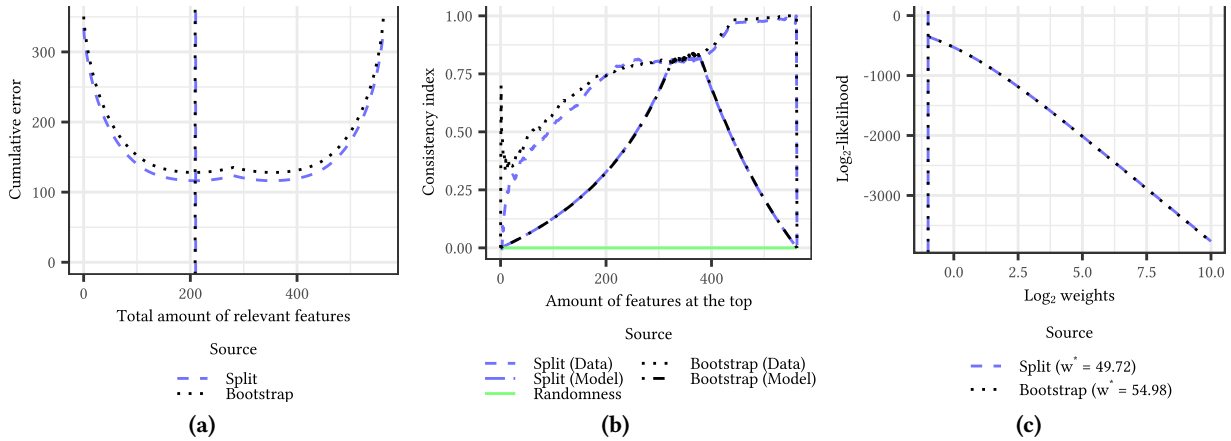


Figure 103: Error plot (103a), reproducibility plot (103b) and weight plot (103c) when the RBFS algorithm based on the linear SVM is applied to the SECOM database

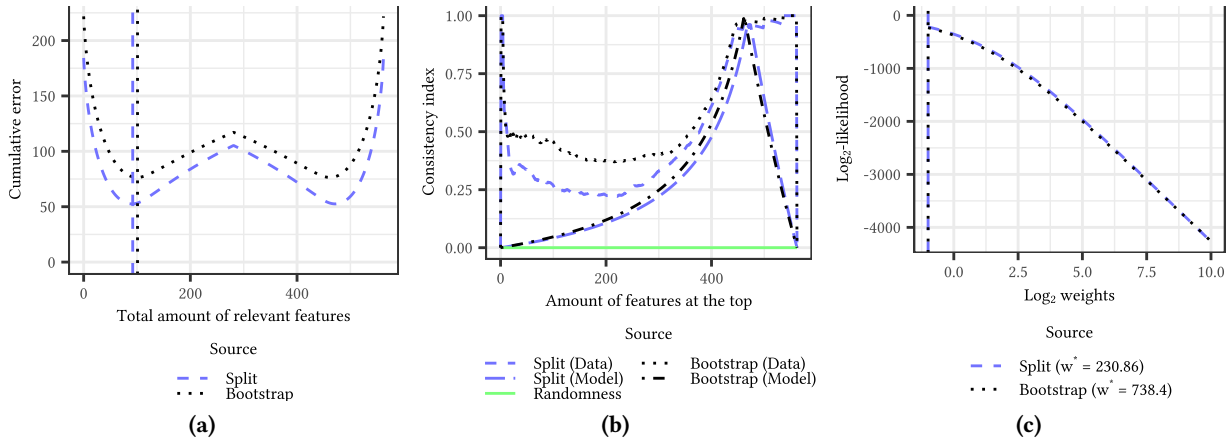


Figure 104: Error plot (104a), reproducibility plot (104b) and weight plot (104c) when the RBFS algorithm based on the t-test is applied to the SECOM database

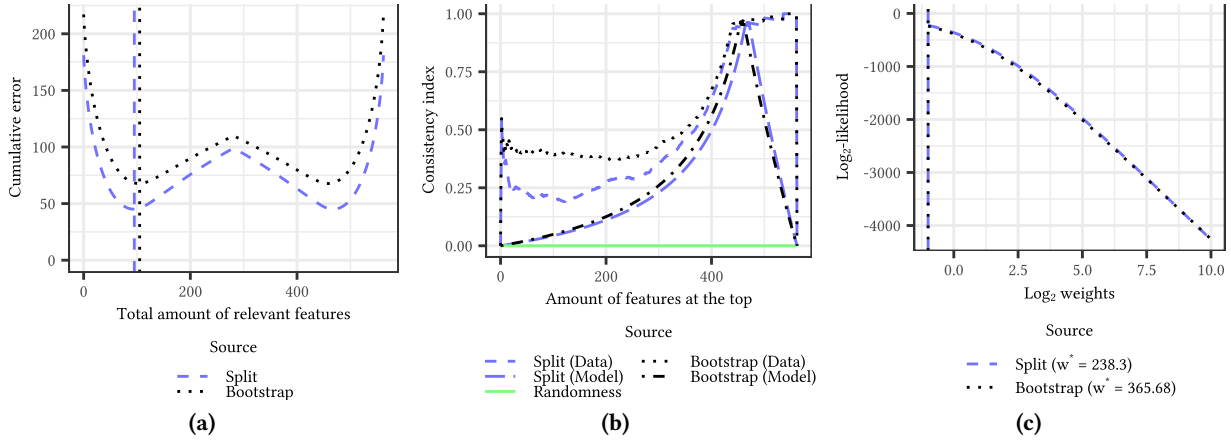


Figure 105: Error plot (105a), reproducibility plot (105b) and weight plot (105c) when the RBFS algorithm based on the Wilcoxon test is applied to the SECOM database

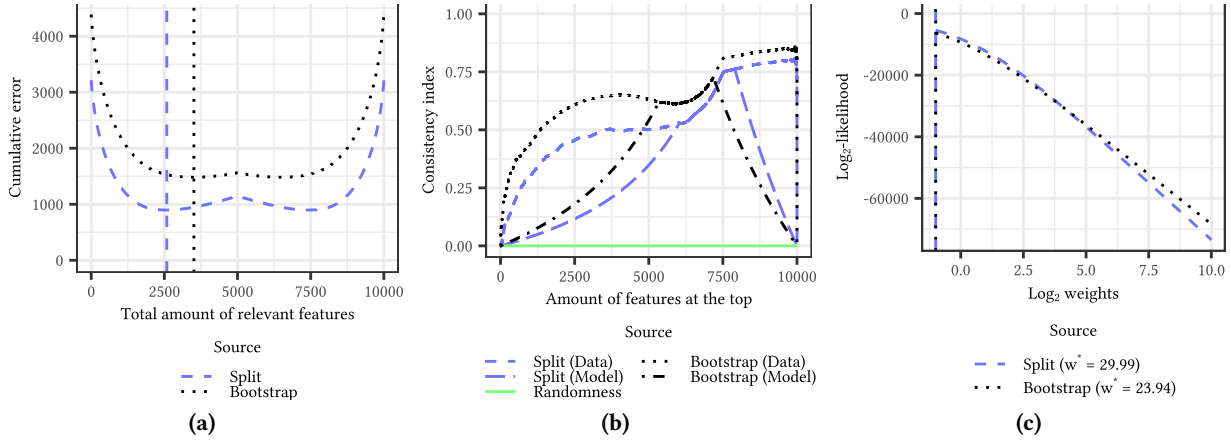


Figure 106: Error plot (106a), reproducibility plot (106b) and weight plot (106c) when the RBFS algorithm based on the mutual information is applied to the arcene database

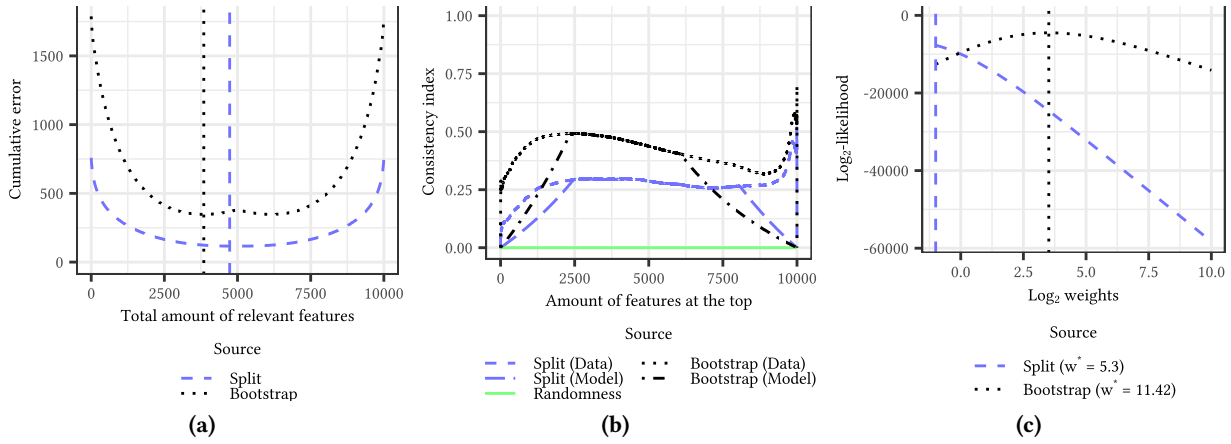


Figure 107: Error plot (107a), reproducibility plot (107b) and weight plot (107c) when the RBFS algorithm based on the linear SVM is applied to the arcene database

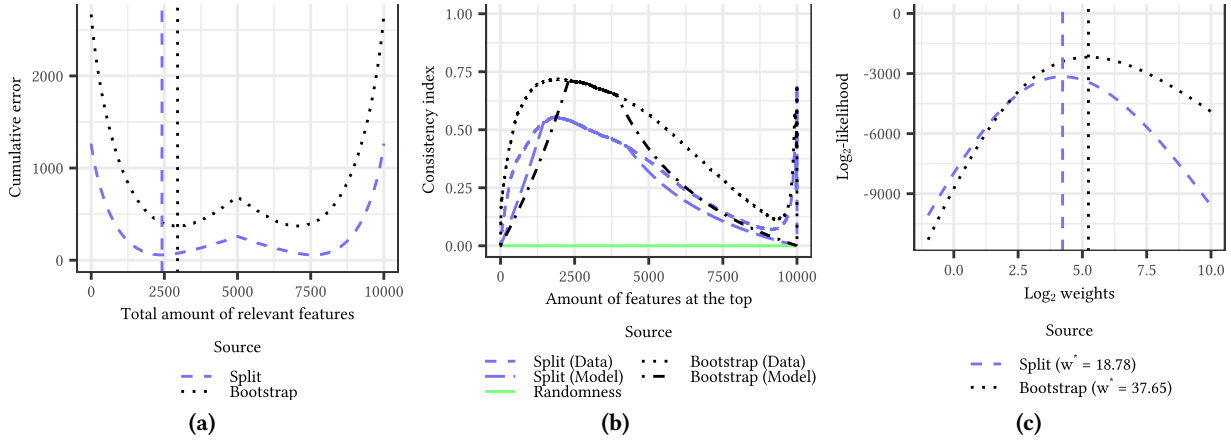


Figure 108: Error plot (108a), reproducibility plot (108b) and weight plot (108c) when the RBFS algorithm based on the t-test is applied to the arcene database

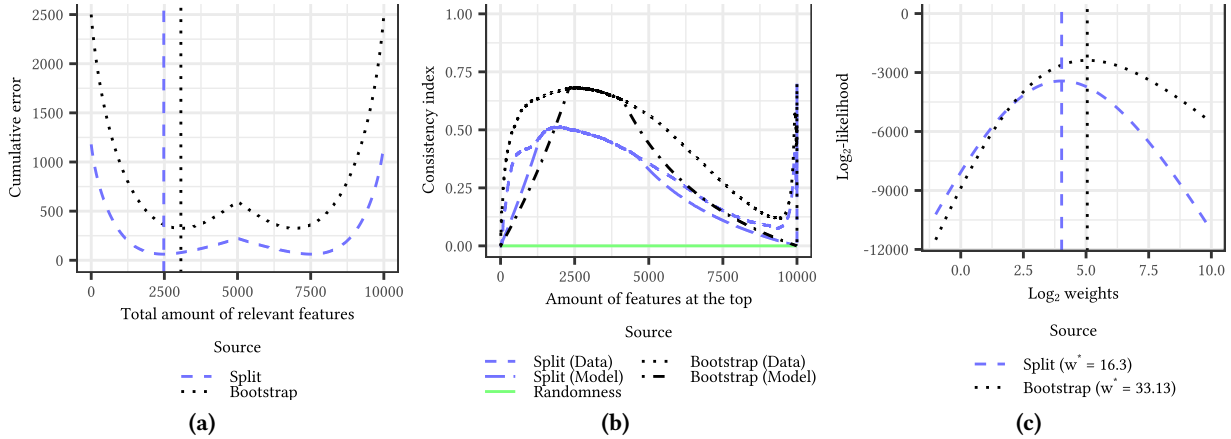


Figure 109: Error plot (109a), reproducibility plot (109b) and weight plot (109c) when the RBFS algorithm based on the Wilcoxon test is applied to the arcene database

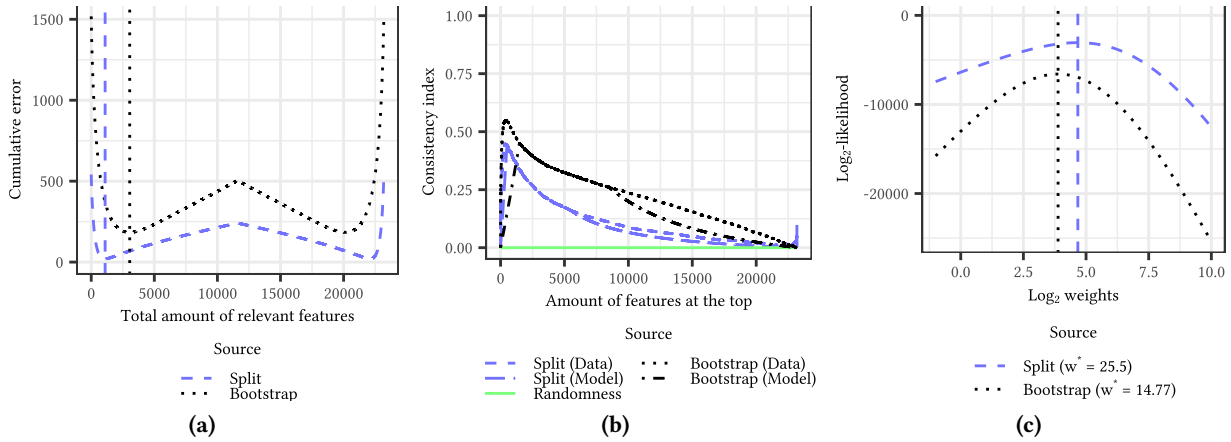


Figure 110: Error plot (110a), reproducibility plot (110b) and weight plot (110c) when the RBFS algorithm based on the mutual information is applied to the ovarian cancer database

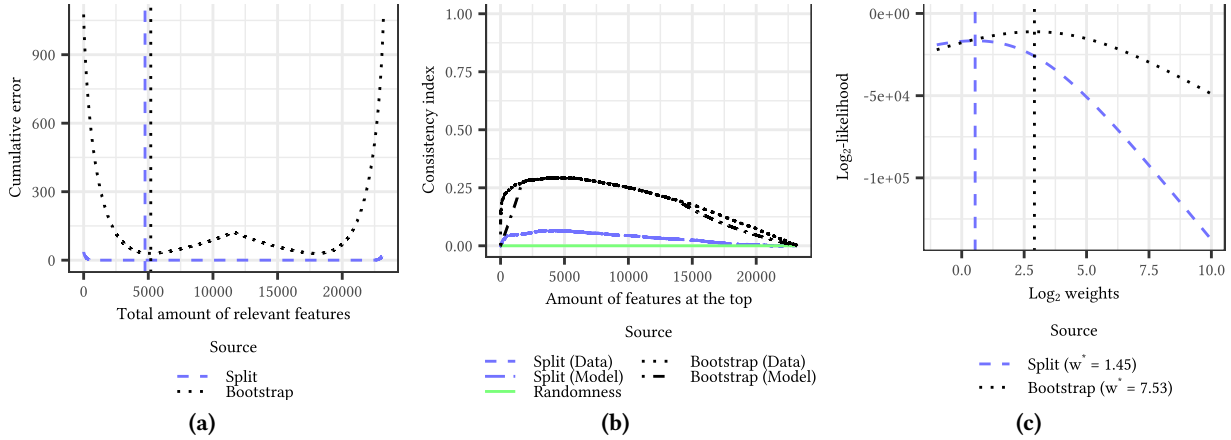


Figure 111: Error plot (111a), reproducibility plot (111b) and weight plot (111c) when the RBFS algorithm based on the linear SVM is applied to the ovarian cancer database

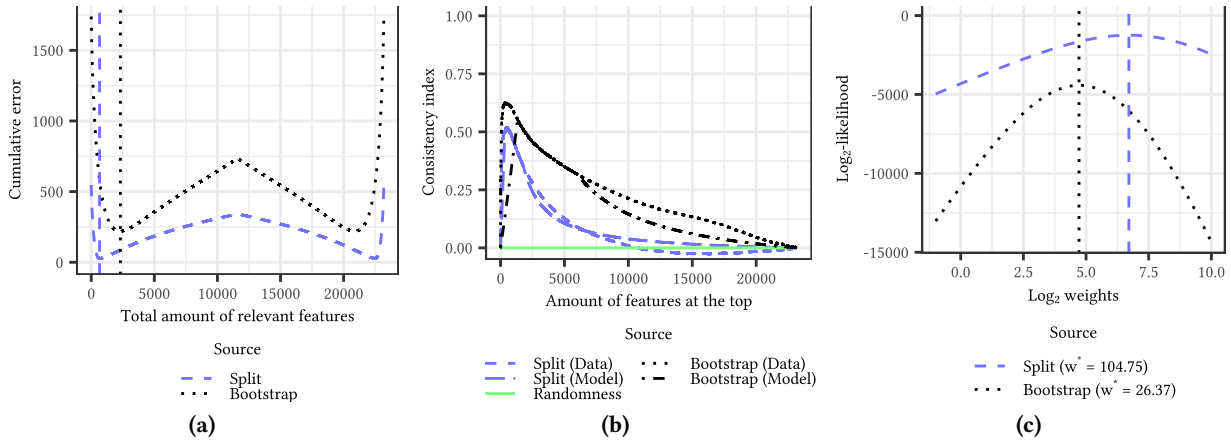


Figure 112: Error plot (112a), reproducibility plot (112b) and weight plot (112c) when the RBFS algorithm based on the t-test is applied to the ovarian cancer database

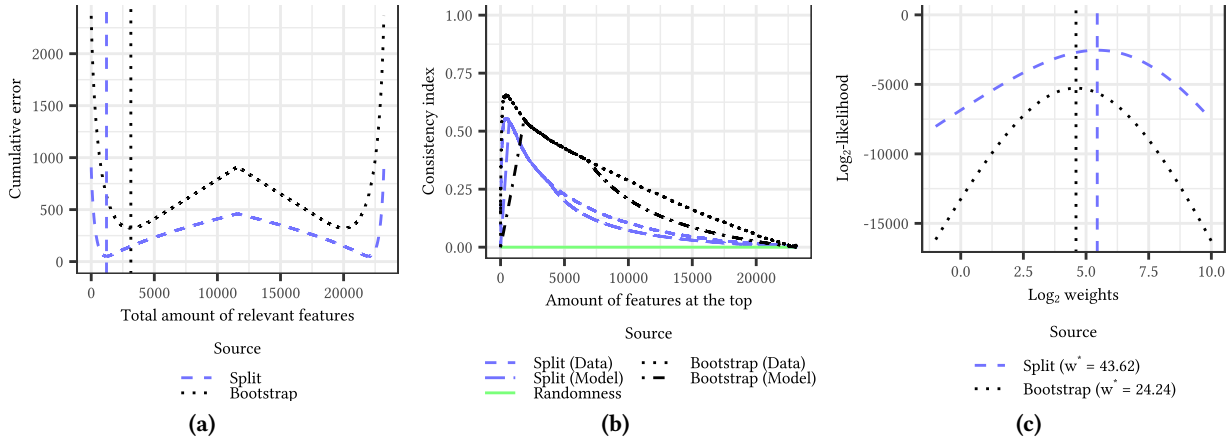


Figure 113: Error plot (113a), reproducibility plot (113b) and weight plot (113c) when the RBFS algorithm based on the Wilcoxon test is applied to the ovarian cancer database

Table 4: w^* values for the RBFS methods when applied to the real datasets ($*$ = the best sequences found by the model have the two types of features completely separated, and, therefore, the higher the w^* the higher the likelihood of it, $\mathcal{L}(\boldsymbol{a}|w)$, without any limit)

Database	Estimation	Mutual information	SVM	t-test	Wilcoxon test
Breast cancer	Random split	∞^*	3.301	60.500	∞^*
Breast cancer	Bootstrap	∞^*	6.392	58.182	∞^*
Mice	Random split	36.834	12.266	25.282	24.479
Mice	Bootstrap	54.094	21.697	40.239	42.238
SECOM	Random split	153.483	49.718	230.861	238.299
SECOM	Bootstrap	280.152	54.985	738.401	365.684
Arcene	Random split	29.988	5.299	18.776	16.297
Arcene	Bootstrap	23.941	11.420	37.650	33.126
Ovarian cancer	Random split	25.504	1.451	104.754	43.624
Ovarian cancer	Bootstrap	14.767	7.525	26.369	24.236

Table 5: Model AUC values for the RBFS methods when applied to the real datasets

Database	Estimation	Mutual information	SVM	t-test	Wilcoxon test
Breast cancer	Random split	1.00000	0.73292	0.99548	1.00000
Breast cancer	Bootstrap	1.00000	0.88995	0.99554	1.00000
Mice	Random split	0.98554	0.94365	0.97874	0.97846
Mice	Bootstrap	0.99417	0.97438	0.99078	0.99158
SECOM	Random split	0.99968	0.99520	0.99986	0.99986
SECOM	Bootstrap	0.99988	0.99575	0.99998	0.99990
Arcene	Random split	0.98496	0.87718	0.96597	0.95932
Arcene	Bootstrap	0.98079	0.94586	0.98705	0.98459
Ovarian cancer	Random split	0.96858	0.66808	0.99278	0.98301
Ovarian cancer	Bootstrap	0.94793	0.89923	0.97239	0.97071

3 Supplementary material for: alternatives to p-value in ranking-based feature selection

Table 6: Parameters of the distributions when Beta distributions are used in the first stage.

Concept	Location	Location and spread
Reference distribution, α parameter	100.00000	100.00000
Reference distribution, β parameter	25.00000	25.00000
Alternative distribution, α parameter	92.99085	383.00440
Alternative distribution, β parameter	19.73293	85.32067

Table 7: Parameters of the distributions when Normal distributions are used in the first stage.

Concept	Location	Location and spread
Reference distribution, μ parameter	-1.00	-1.00
Reference distribution, σ^2 parameter	0.50^2	0.50^2
Alternative distribution, μ parameter	-0.65	-0.75
Alternative distribution, σ^2 parameter	0.50^2	0.25^2

3.1 Introduction

Here, we gather additional documentation which is not presented in Chapter 4 of the thesis manuscript. Briefly, this additional documentation gathers:

- Details of the experimentation carried out.
- Descriptions of other alternative methods.

3.2 Details of the experimentation

In this section we explain the remaining details of the experimentation conducted. In order to do so, we have divided this section into four parts, each one dedicated to a different stage of the experimental framework.

However, before going into the details of each stage, we first expose a common property of the first, second and third stages. That property consists of that, for each configuration, the values of the parameters for the reference and alternative distributions in the first stage and the reference and fourth alternative distributions in the second and third stages are based on the experimentation conducted by Chen et al [8].

3.2.1 First stage

The specific values of the parameters of the reference and alternative distributions for each possible configuration depend on its associated type of distributions and its associated nature of differences. These specific values of the parameters are gathered in Table 6 and in Table 7. In addition, in Figure 114 a visual representation of all these distributions is offered.

3.2.2 Second stage

As in the previous stage, once again, the specific values of the parameters of the reference and alternative distributions for each possible configuration depend on its associated type of distributions and its associated nature of differences. These specific values of the parameters are gathered in Table 8 and in Table 9. It is convenient to mention that in each configuration the alternative distributions 1, 2 and 3 are equally

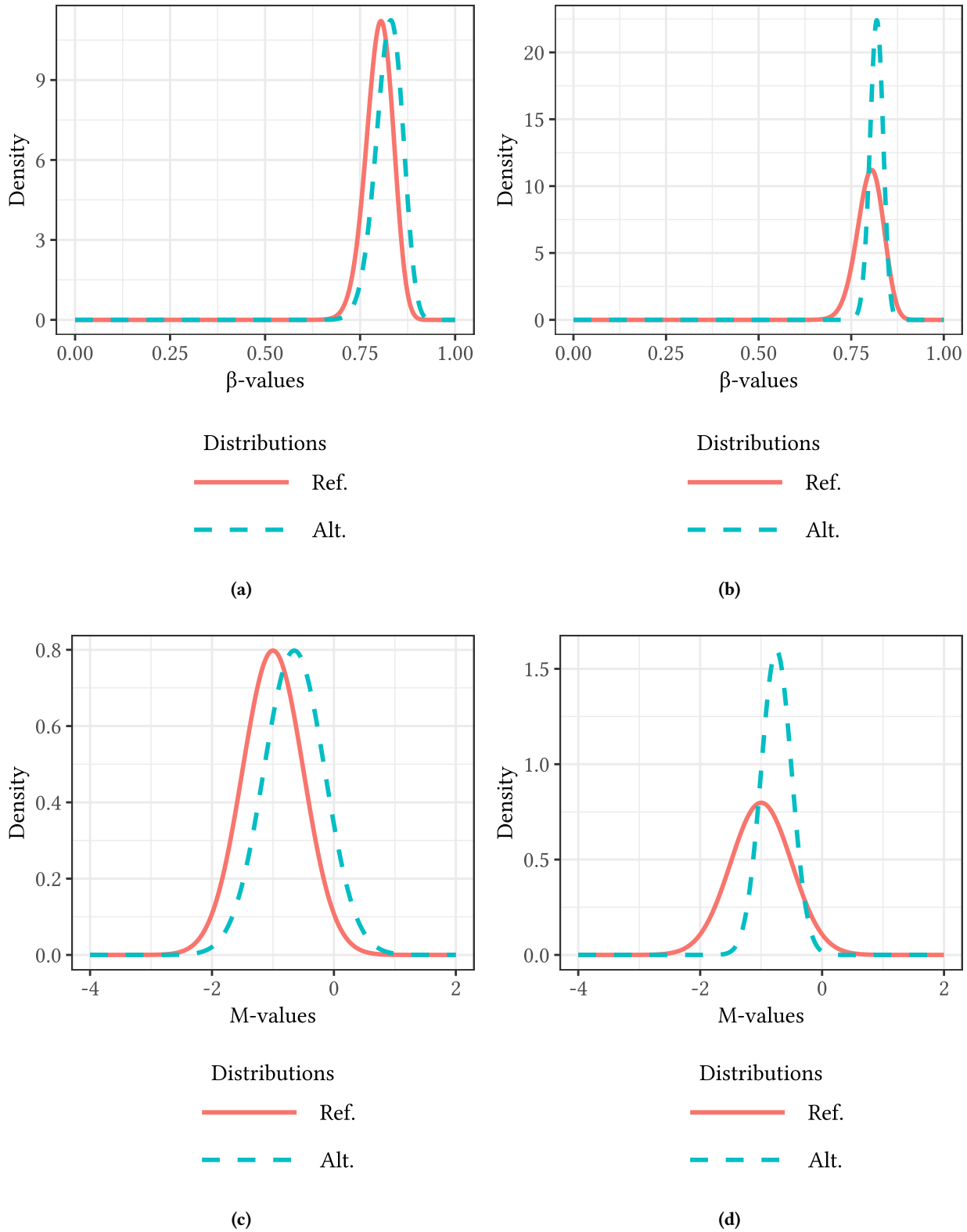


Figure 114: Probability distributions used in the first stage for the combinations 114a Beta distributions and differences in location, 114b Beta distributions and differences in location and spread, 114c Normal distributions and differences in location and 114d Normal distributions and differences in location and spread.

Table 8: Parameters of the distributions when Beta distributions are used in the second stage.

Concept	Location	Location and spread
Reference distribution, α parameter	100.00000	100.00000
Reference distribution, β parameter	25.00000	25.00000
Alternative distribution 1, α parameter	98.37920	129.35790
Alternative distribution 1, β parameter	23.64362	31.44414
Alternative distribution 2, α parameter	96.67151	174.24340
Alternative distribution 2, β parameter	22.31288	41.16213
Alternative distribution 3, α parameter	94.87577	248.10500
Alternative distribution 3, β parameter	21.00894	56.93099
Alternative distribution 4, α parameter	92.99085	383.00440
Alternative distribution 4, β parameter	19.73293	85.32067

Table 9: Parameters of the distributions when Normal distributions are used in the second stage.

Concept	Location	Location and spread
Reference distribution, μ parameter	-1.0000	-1.0000
Reference distribution, σ^2 parameter	0.5000^2	0.5000^2
Alternative distribution 1, μ parameter	-0.9125	-0.9375
Alternative distribution 1, σ^2 parameter	0.5000^2	0.4375^2
Alternative distribution 2, μ parameter	-0.8250	-0.8750
Alternative distribution 2, σ^2 parameter	0.5000^2	0.3750^2
Alternative distribution 3, μ parameter	-0.7375	-0.8125
Alternative distribution 3, σ^2 parameter	0.5000^2	0.3125^2
Alternative distribution 4, μ parameter	-0.6500	-0.7500
Alternative distribution 4, σ^2 parameter	0.5000^2	0.2500^2

spaced between the reference distribution and the alternative distribution 4 in terms of mean and standard deviation. Finally, in Figure 115 all these distributions are represented graphically.

3.2.3 Third stage

In this stage, since only one type of distribution is used, the specific values of the parameters of the reference and alternative distributions for each possible configuration depend only on its associated nature of differences. Table 10 shows all these specific values. Besides, in each configuration, each component of the alternative distributions 1, 2 and 3 is equally spaced between the corresponding components of the reference distribution and the alternative distribution 4 in terms of mean and standard deviation. Finally, all these distributions are displayed in Figures 116, 117 and 118.

3.2.4 Fourth stage

We have divided this section into two parts, one dedicated to the preprocessings of the ovarian cancer database and the other dedicated to the preprocessing of the nephropathy database.

Ovarian cancer database preprocessings The two preprocessings done in our contribution, that have been applied to the ovarian cancer database, are based on what was done by Wang et al [9]. The first

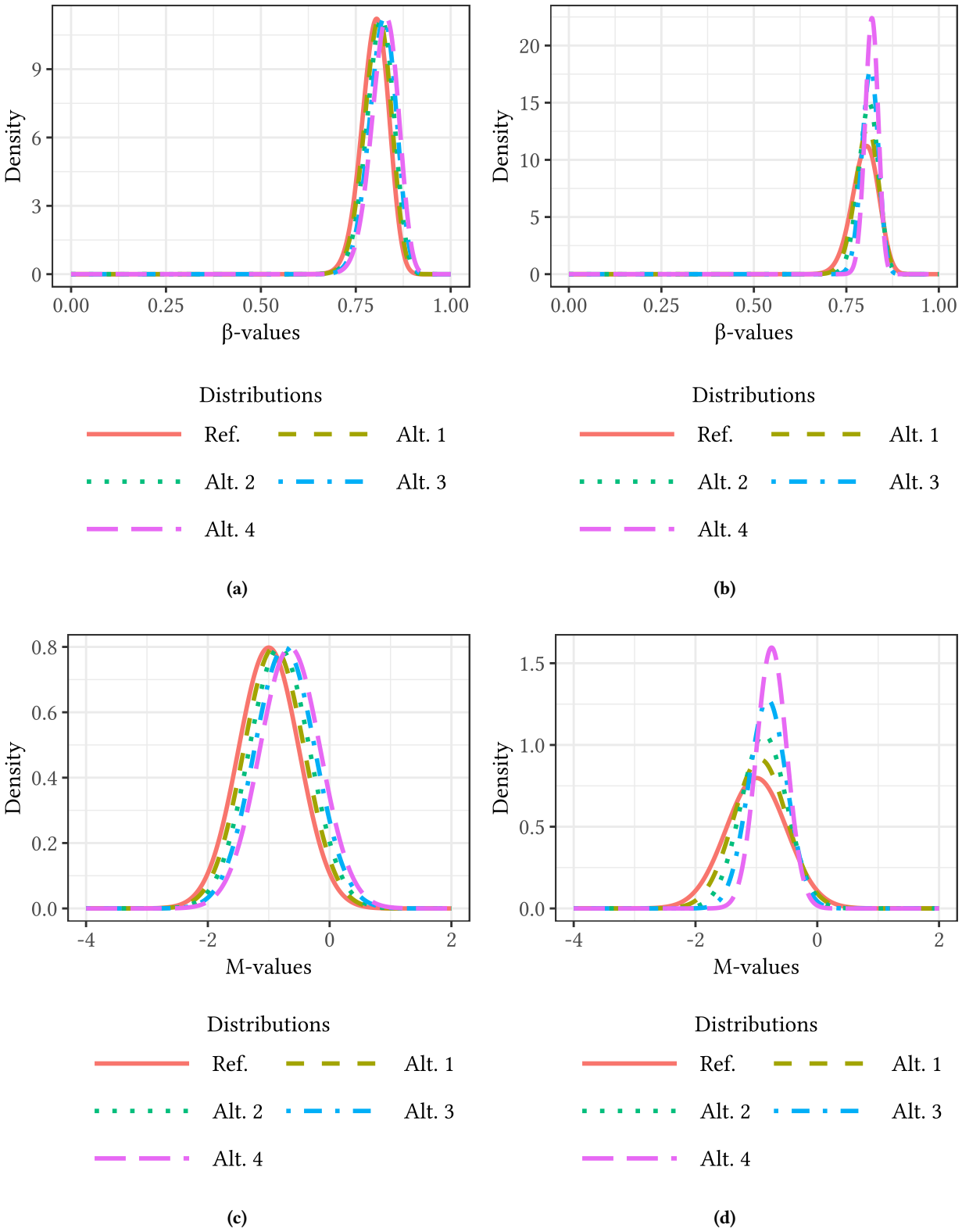


Figure 115: Probability distributions used in the second stage for the combinations 115a Beta distributions and differences in location, 115b Beta distributions and differences in location and spread, 115c Normal distributions and differences in location and 115d Normal distributions and differences in location and spread.

Table 10: Parameters of the distributions used in the third stage.

Concept	Location	Location and spread
Reference distribution, component 1, μ parameter	-3.0000	-3.0000
Reference distribution, component 1, σ^2 parameter	0.5000^2	0.5000^2
Reference distribution, component 2, μ parameter	2.0000	2.0000
Reference distribution, component 2, σ^2 parameter	0.5000^2	0.5000^2
Alternative distribution 1, component 1, μ parameter	-2.9125	-2.9375
Alternative distribution 1, component 1, σ^2 parameter	0.5000^2	0.4375^2
Alternative distribution 1, component 2, μ parameter	2.0875	2.0625
Alternative distribution 1, component 2, σ^2 parameter	0.5000^2	0.4375^2
Alternative distribution 2, component 1, μ parameter	-2.8250	-2.8750
Alternative distribution 2, component 1, σ^2 parameter	0.5000^2	0.3750^2
Alternative distribution 2, component 2, μ parameter	2.1750	2.1250
Alternative distribution 2, component 2, σ^2 parameter	0.5000^2	0.3750^2
Alternative distribution 3, component 1, μ parameter	-2.7375	-2.8125
Alternative distribution 3, component 1, σ^2 parameter	0.5000^2	0.3125^2
Alternative distribution 3, component 2, μ parameter	2.2625	2.1875
Alternative distribution 3, component 2, σ^2 parameter	0.5000^2	0.3125^2
Alternative distribution 4, component 1, μ parameter	-2.6500	-2.7500
Alternative distribution 4, component 1, σ^2 parameter	0.5000^2	0.2500^2
Alternative distribution 4, component 2, μ parameter	2.3500	2.2500
Alternative distribution 4, component 2, σ^2 parameter	0.5000^2	0.2500^2

preprocessing that was applied, the one that does not remove every single outlier systematically, consists of applying the following steps sequentially to the matrix of β-values available in the GEO database:

1. Among all the ovarian cancer cases, only those who gave their blood at the time of their diagnosis prior to treatment have been used.
2. Samples whose bisulfite conversion efficiencies are too low (< 4000) have been removed.
3. Data from batches 10-12 have been removed.
4. In order to remove outlier samples, for each sample, we have computed the average of the distances of all the values (using raw total intensities) of its CpG sites regarding their median values across samples. Then, all the generated averages are used to calculate the range $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$. Finally all those samples whose averages are not within that range are removed.
5. All those individuals that do not cover at least 95% of the CpG sites with a detection p-value smaller than 0.05 are removed.
6. All the CpG sites whose detection p-values are not below 0.05 in all samples are removed.
7. All the CpG sites that do not have numeric values (e.g., NA values) for at least 50 individuals per group are removed.

The other preprocessing is pretty similar to that which has already been presented. Specifically, the next steps are applied sequentially to the matrix of β-values available in the GEO database:

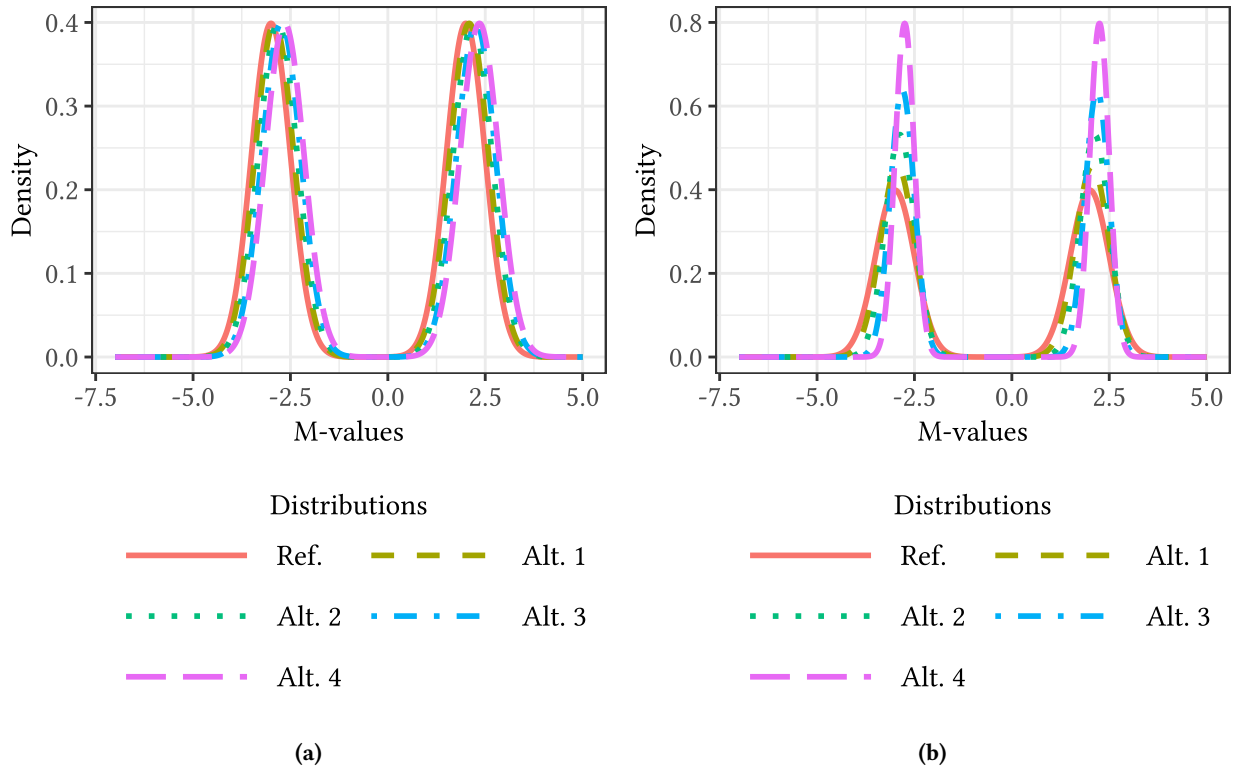


Figure 116: Probability distributions used in the third stage for the combinations 116a weights $(50, 50)$ and differences in location and 116b weights $(50, 50)$ and differences in location and spread.

1. Among all the ovarian cancer cases, only those who gave their blood at the time of their diagnosis prior to treatment have been used.
2. Samples whose bisulfite conversion efficiencies are too low (< 4000) have been removed.
3. Data from batches 10-12 have been removed.
4. In order to remove outlier samples, for each sample, we have computed the average of the distances of all the values (using raw total intensities) of its CpG sites regarding their median values across samples. Then, all the generated averages are used to calculate the range $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$. Finally all those samples whose averages are not within that range are removed.
5. For each CpG site, we have measured which values of total intensity and which β-values lie outside their corresponding ranges defined by $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$. All the corresponding β-values are erased, setting their value to NA.
6. All those individuals that do not cover at least 95% of the CpG sites with a detection p-value smaller than 0.05 are removed.
7. All the CpG sites that do not have numeric values (e.g., NA values) for at least 50 individuals per group are removed.

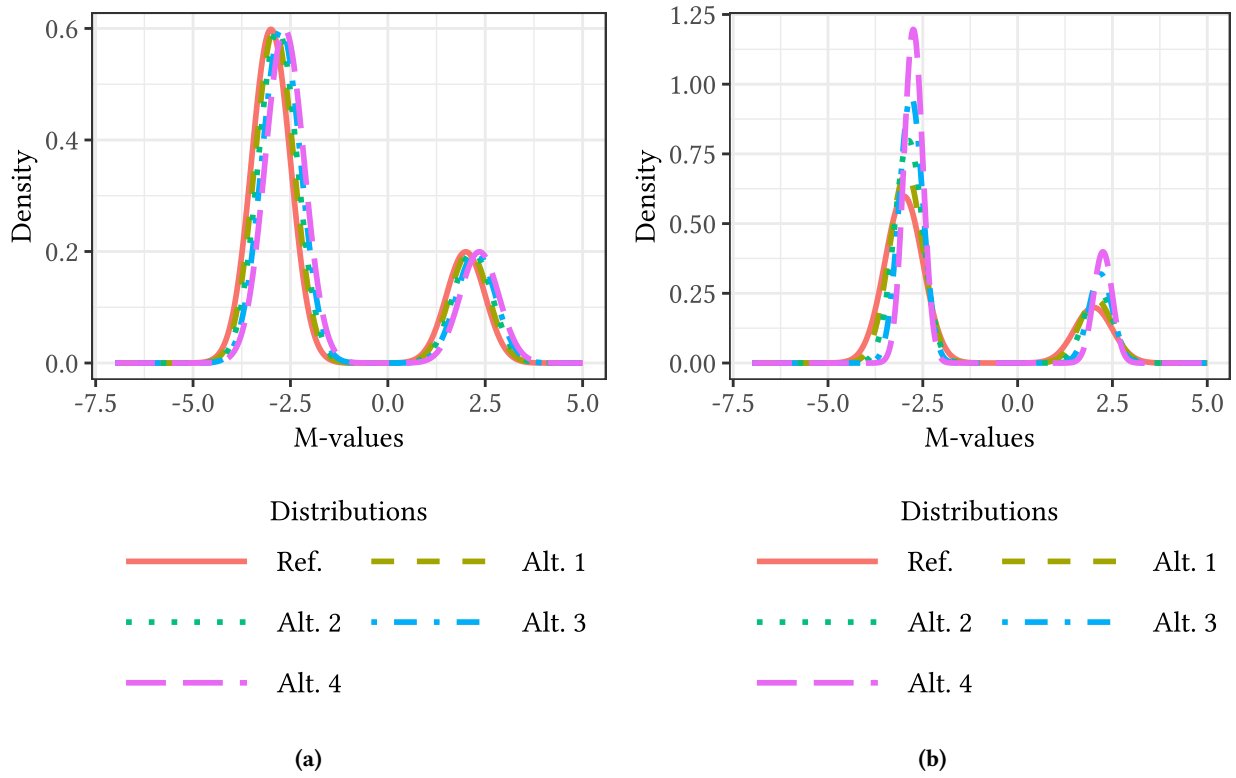


Figure 117: Probability distributions used in the third stage for the combinations 117a weights $(75, 25)$ and differences in location and 117b weights $(75, 25)$ and differences in location and spread.

Nephropathy database preprocessing The preprocessing done in our contribution to the nephropathy cancer database is based on what was done by Teschendorff et al [6]. The preprocessing we applied, consists of applying the following steps sequentially to the matrix of β-values available in the GEO database:

- Samples whose bisulfite conversion efficiencies are too low (lower values outside the range $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$) have been removed.
- In order to remove outlier samples, for each sample, we have computed the average of the distances of all the values (using raw total intensities) of its CpG sites regarding their median values across samples. Then, all the generated averages are used to calculate the range $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$. Finally all those samples whose averages are not within that range are removed.
- All those individuals that do not cover at least 95% of the CpG sites with a detection p-value smaller than 0.05 are removed.
- All the CpG sites whose detection p-values are not below 0.05 in all samples are removed.

3.3 Other alternative methods

We have divided this section into two parts, dedicating each one to a different alternative method. Before presenting the two methods, it is convenient to recall that $\boldsymbol{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(x)}\}$ and $\boldsymbol{Y} = \{\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(y)}\}$

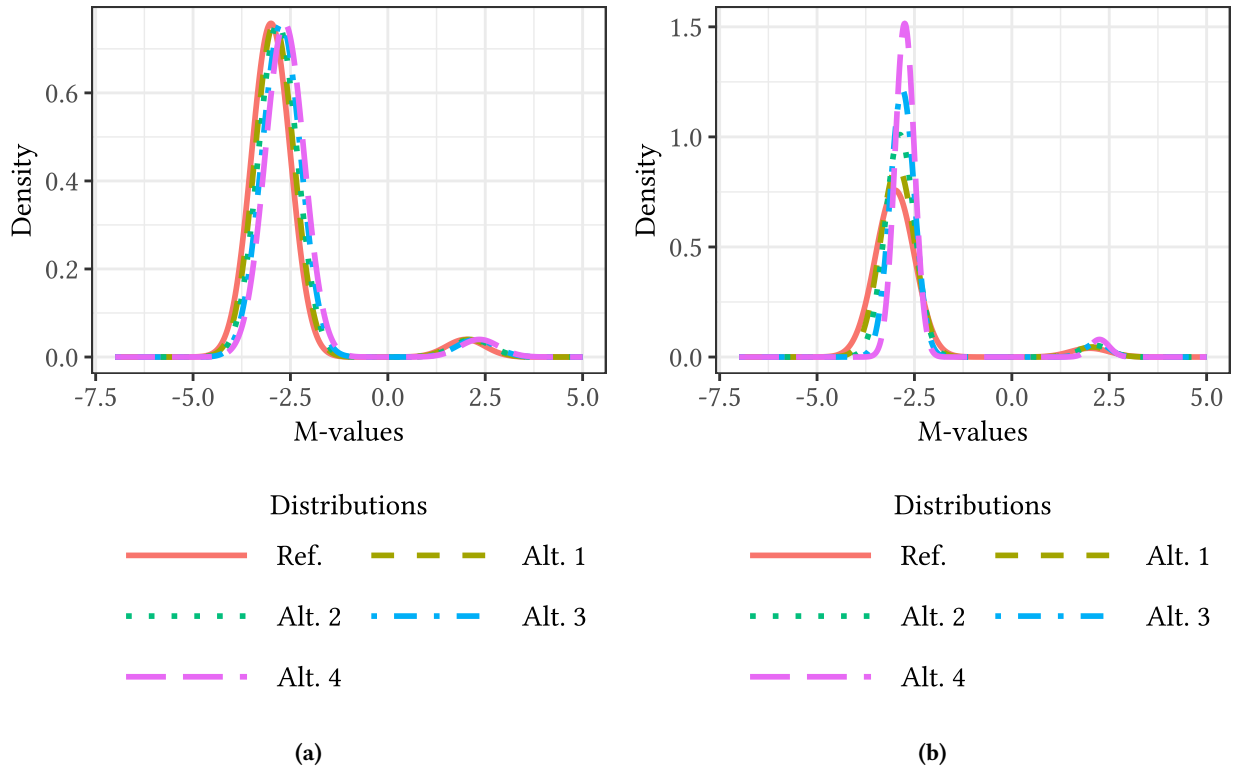


Figure 118: Probability distributions used in the third stage for the combinations 118a weights $(95, 5)$ and differences in location and 118b weights $(95, 5)$ and differences in location and spread.

are two samples of x and y individuals, respectively. We denote as $\boldsymbol{x}_i = \{x_i^{(1)}, \ldots, x_i^{(x)}\}$ and $\boldsymbol{y}_i = \{y_i^{(1)}, \ldots, y_i^{(y)}\}$ the vectors of values for site i in samples $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

3.3.1 Movement of distributions method

The idea behind this method is to attempt to measure somehow how much it costs to transform the estimation of the distribution of one group into the estimation of the distribution of the other group. So as to do that, we considered that at each of the x values of the first group there is a portion of $1/x$ of the whole distribution density (analogously, at each of the y values of the second group we considered there to be a portion of $1/y$). Namely, the method is composed of the following steps:

1. Calculate the least common multiple (LCM) between x and y.
2. Repeat LCM/x times each element of $\boldsymbol{x}_i$ and sort the resulting vector. Analogously, repeat LCM/y times each element of $\boldsymbol{y}_i$ and sort the resulting vector. We denote these two new vectors as $\boldsymbol{u}_i = \{u_i^{(1)}, \ldots, u_i^{(LCM)}\}$ and $\boldsymbol{v}_i = \{v_i^{(1)}, \ldots, v_i^{(LCM)}\}$, respectively.
3. Finally, compute the next value as the outcome of the method, $s(\boldsymbol{x}_i)$ and $s(\boldsymbol{y}_i)$ being the standard deviations of $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$:

$$\frac{\sum_{j=1}^{LCM} |u_i^{(j)} - v_i^{(j)}|}{LCM \cdot (s(\boldsymbol{x}_i) + s(\boldsymbol{y}_i))}.$$

Since this method is sensitive to differences in both location and spread, in Figures 119 and 120 the results during the experimentation are displayed together with the results of the Tl test and the Kolmogorov-Smirnov test.

3.3.2 Differences of distributions method

The essence of this method consists of computing the amount of difference between the areas of the empirically estimated distributions at each of the different points (except for the first one) of $\boldsymbol{x}_i \cup \boldsymbol{y}_i$ taking into account the previous point. $\boldsymbol{x}_i \cup \boldsymbol{y}_i$ being represented through $\{z_i^{(1)}, \ldots, z_i^{(z)}\}$ and it being ordered, $z_i^{(1)} <, \ldots, < z_i^{(z)}$, the outcome of this method is summarized in the following expression:

$$\sum_{j=2}^{x+y} \left| \frac{|\boldsymbol{x}_i < z_i^{(j)}|}{x} - \frac{|\boldsymbol{y}_i < z_i^{(j)}|}{y} \right| \cdot |z_i^{(j)} - z_i^{(j-1)}|.$$

Since this method is sensitive to differences in both location and spread, in Figures 121 and 122 the results during the experimentation are displayed together with the results of the Tl test and the Kolmogorov-Smirnov test.

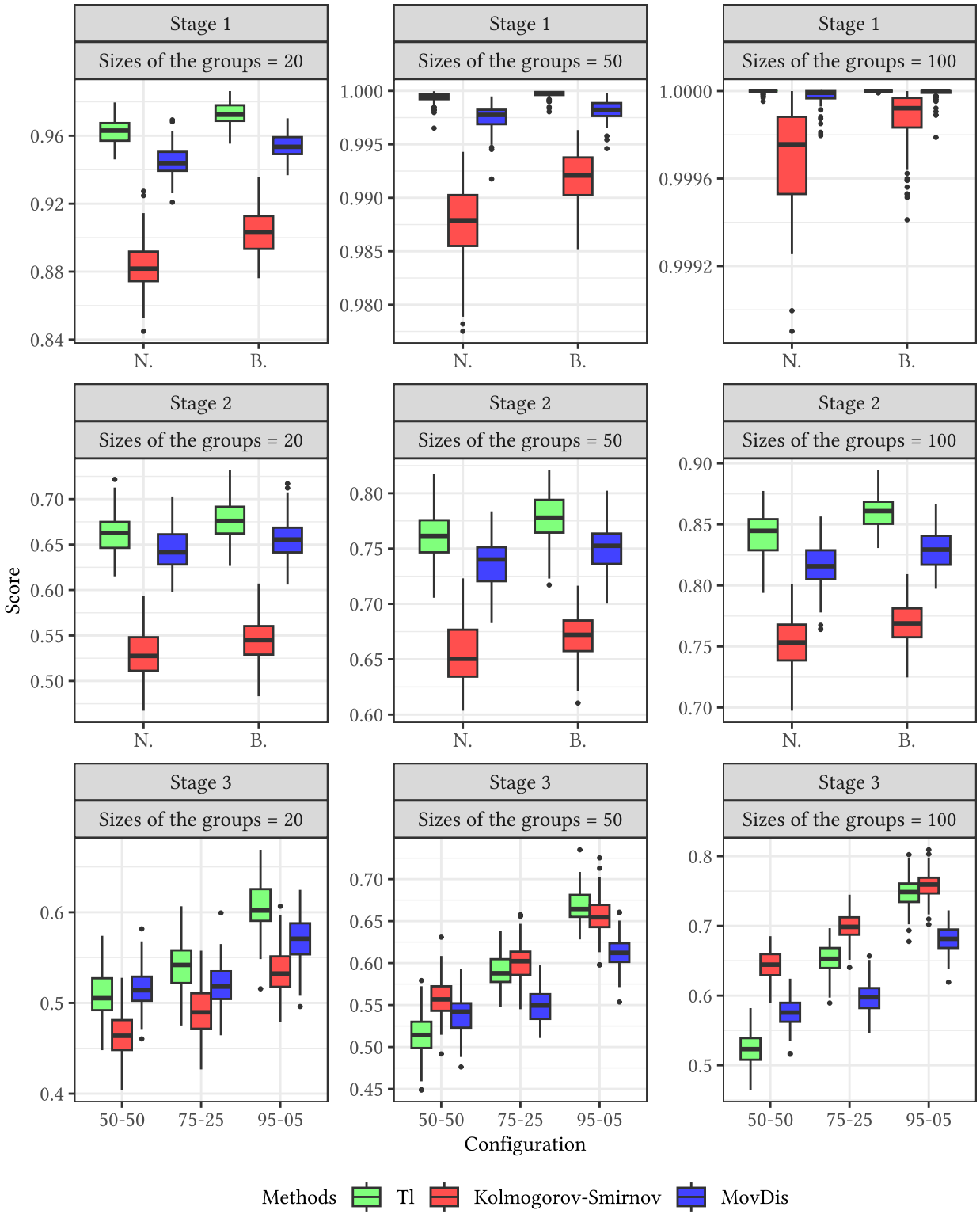


Figure 119: Results of the Tl test, the Kolmogorov-Smirnov test and the MovDist method in the synthetic stages. The labels of the abscissa axes of the boxplots specify information about the distributions used: "N." - Normal distributions, "B." - Beta distributions, "50 — 50", "75 — 25" or "95 — 05" - mixtures of Normal distributions in which the weights of the Normal distributions are equal to the values specified by the corresponding label.

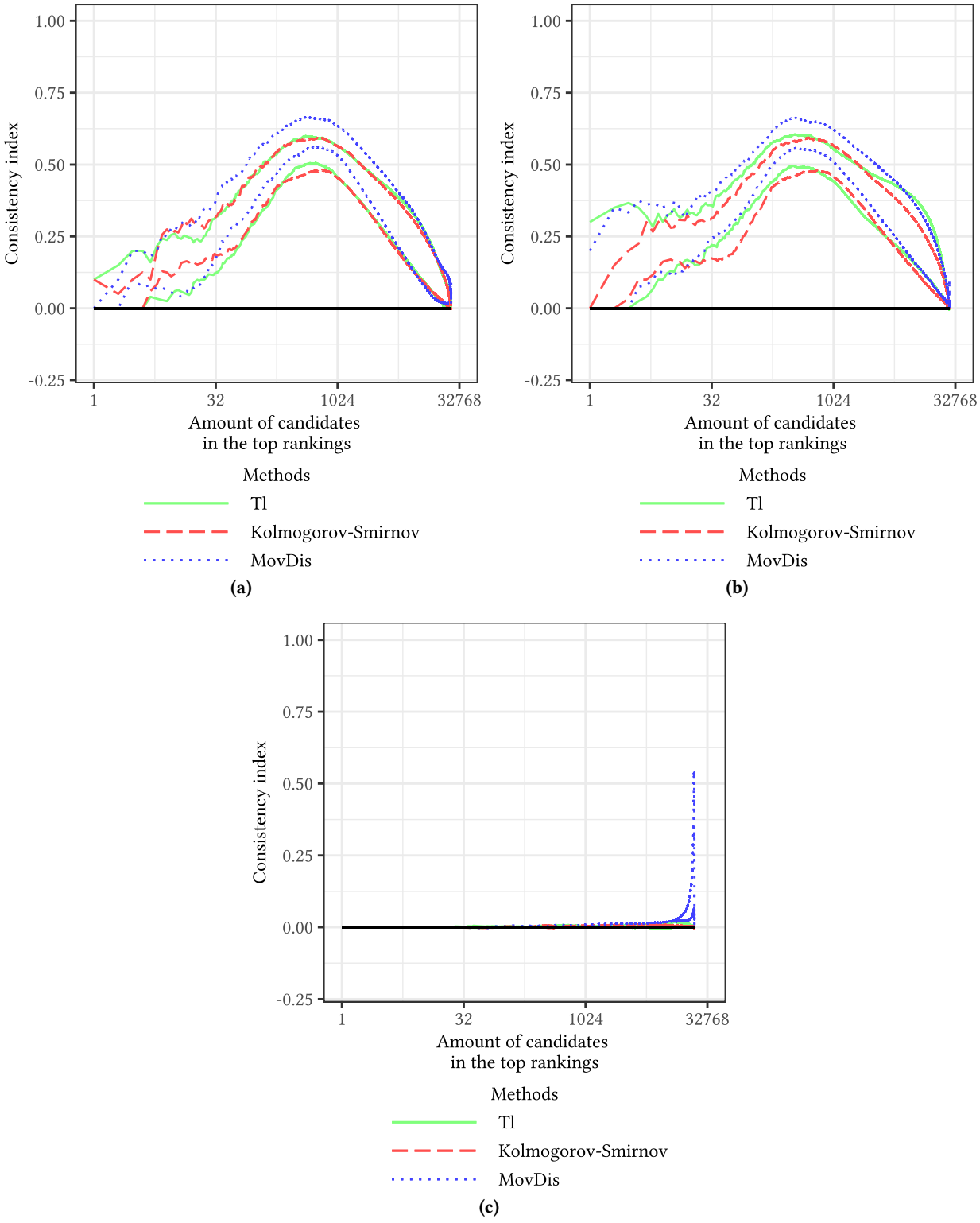


Figure 120: Results in the real stage, when the Tl test, the Kolmogorov-Smirnov test and the MovDist method are applied in 120a the ovarian cancer database (in which the first preprocessing has been executed), 120b the ovarian cancer database (in which the second preprocessing has been executed) and 120c the nephropathy database.

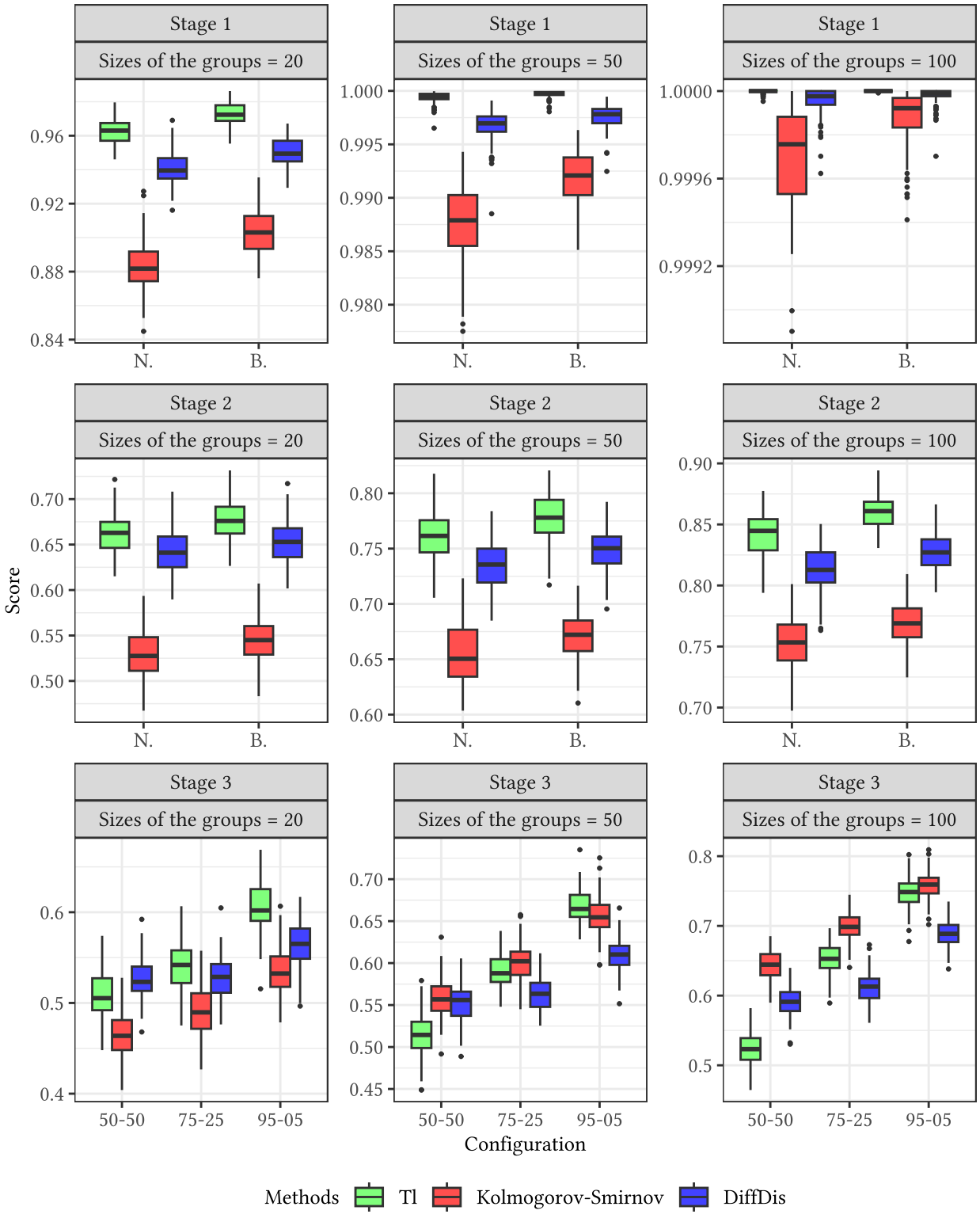


Figure 121: Results of the Tl test, the Kolmogorov-Smirnov test and the DifDist method in the synthetic stages. The labels of the abscissa axes of the boxplots specify information about the distributions used: "N." - Normal distributions, "B." - Beta distributions, "50 − 50", "75 − 25" or "95 − 05" - mixtures of Normal distributions in which the weights of the Normal distributions are equal to the values specified by the corresponding label.

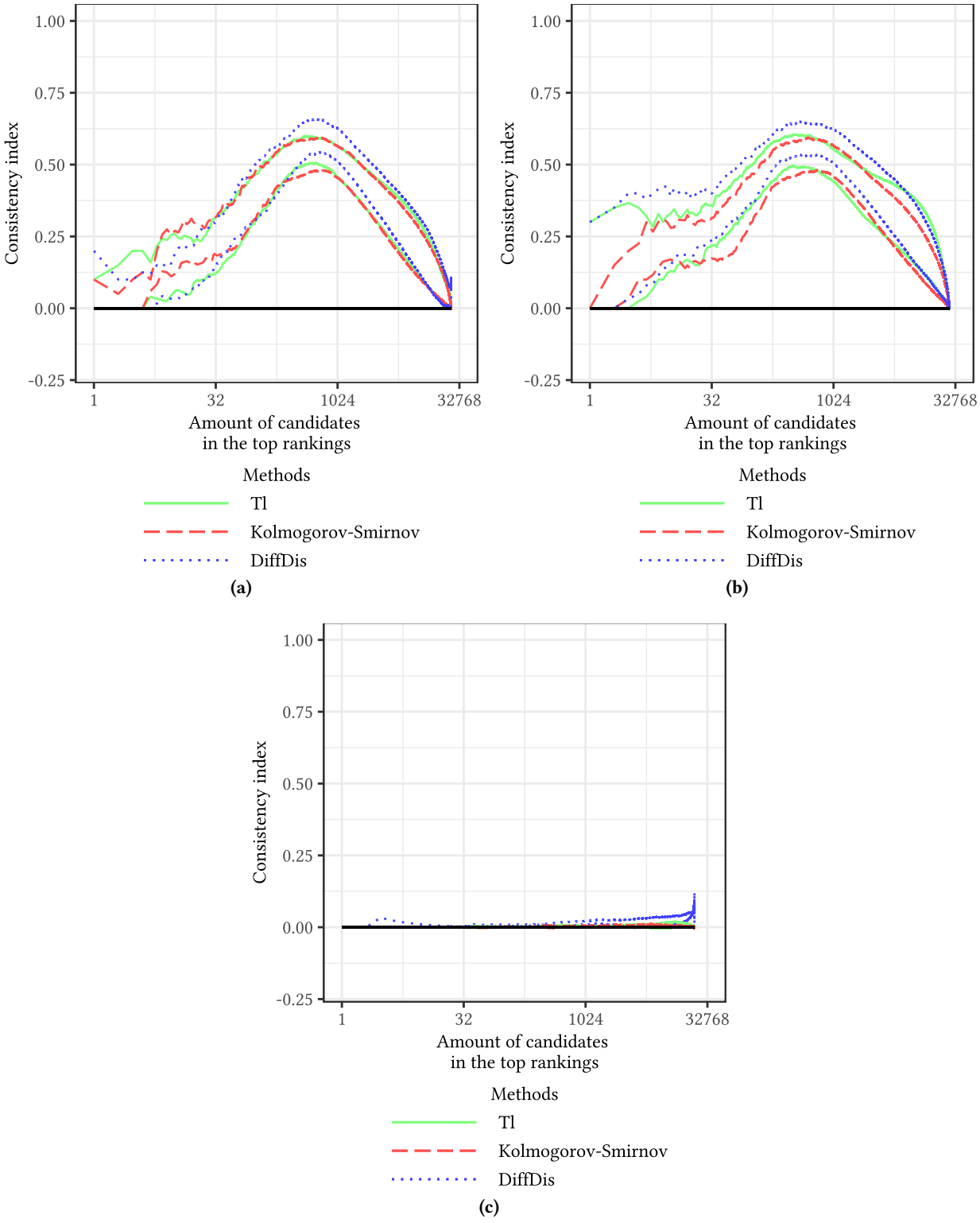


Figure 122: Results in the real stage, when the Tl test, the Kolmogorov-Smirnov test and the DifDist method are applied in 122a the ovarian cancer database (in which the first preprocessing has been executed), 122b the ovarian cancer database (in which the second preprocessing has been executed) and 122c the nephropathy database.

References

[1] Olvi L Mangasarian, W Nick Street, and William H Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.

[2] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. International Society for Optics and Photonics, 1993.

[3] Clara Higuera, Katheleen J Gardiner, and Krzysztof J Cios. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PloS one*, 10(6):e0129126, 2015.

[4] Michael McCann, Yuhua Li, Liam Maguire, and Adrian Johnston. Causality challenge: benchmarking relevant signal components for effective monitoring and process control. In *Causality: Objectives and Assessment*, pages 277–288, 2010.

[5] Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552, 2005.

[6] Andrew E Teschendorff, Usha Menon, Aleksandra Gentry-Maharaj, Susan J Ramus, Daniel J Weisenberger, Hui Shen, Mihaela Campan, Houtan Noushmehr, Christopher G Bell, A Peter Maxwell, et al. Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome research*, 20(4):440–446, 2010.

[7] Shuang Wang. Method to detect differentially methylated loci with case-control designs using illumina arrays. *Genetic epidemiology*, 35(7):686–694, 2011.

[8] Chen, Yong and Ning, Yang and Hong, Chuan and Wang, Shuang. Semiparametric Tests for Identifying Differentially Methylated Loci With Case–Control Designs Using Illumina Arrays. *Genetic Epidemiology*, 38(1):42–50, 2014.

[9] Emily S Wan, Weiliang Qiu, Andrea Baccarelli, Vincent J Carey, Helene Bacherman, Stephen I Rennard, Alvar Agustí, Wayne H Anderson, David A Lomas, and Dawn L DeMeo. Systemic steroid exposure is associated with differential methylation in chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 186(12):1248–1255, 2012.