Computing, Telecommunication and Control, 2025, Vol. 18, No. 2, Pp. 7–20.
Информатика, телекоммуникации и управление. 2025. Том 18, № 2. С. 7–20.

Intelligent Systems and Technologies, Artificial Intelligence

Интеллектуальные системы и технологии, искусственный интеллект

Research article
DOI: https://doi.org/10.18721/JCSTCS.18201
UDC 519.2

CATEGORICAL SURVIVAL ANALYSIS OF THE REQUIRED JOB EXECUTION TIMES IN THE HYBRID SUPERCOMPUTER CENTER

T.A. Misharina[1], *S.V. Malov*[1,2] ✉

[1] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation;

[2] St. Petersburg Electrotechnical University, St. Petersburg, Russian Federation

✉ tanechkamisharina254@gmail.com

Abstract. According to statistics, the actual execution time of most jobs on a supercomputer cluster differs significantly from the time requested by the user. Investigation of distributions of supercomputer job execution times using statistical or machine learning methods allows optimizing the operation of a supercomputer cluster. We study the results of computational jobs processing in the supercomputer center of Peter the Great St. Petersburg Polytechnic University. We have developed a nonparametric approach for detection and statistical confirmation of weak stochastic orders based on categorical nonparametric framework of contrasts obtained from the Kaplan–Meier estimators obtained from independent groups of right-censored observations. To adjust the confidence level of the detected weak stochastic orders, we apply the Bonferroni correction to all the comparisons under consideration. We perform comparative statistical analysis of the distributions of required execution times to complete successfully the job in different groups of right-censored observations; detect and confirm available weak stochastic orders.

Keywords: survival data, Kaplan–Meier estimator, Wald's type test, stochastic orders, supercomputer cluster, job scheduling

Acknowledgements: The research was partially financially supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state assignment "Development and research of machine learning models for solving fundamental problems of artificial intelligence in the fuel and energy complex" (FSEG-2024-0027). The results of the work were obtained using the computational resources of the shared-use center "Polytechnic Supercomputer Center" of Peter the Great St. Petersburg Polytechnic University (No. 500675, https://ckp-rf.ru/catalog/ckp/500675/).

Citation: Misharina T.A., Malov S.V. Categorical survival analysis of the required job execution times in the hybrid supercomputer center. Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 7–20. DOI: 10.18721/JCSTCS.18201

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.18201
УДК 519.2

КАТЕГОРИАЛЬНЫЙ АНАЛИЗ ВЫЖИВАЕМОСТИ ТРЕБУЕМЫХ ВРЕМЕН ИСПОЛНЕНИЯ ЗАДАЧ В ГИБРИДНОМ СУПЕРКОМПЬЮТЕРНОМ ЦЕНТРЕ

Т.А. Мишарина[1], *С.В. Малов*[1,2] ✉

[1] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация;

[2] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, Российская Федерация;

✉ tanechkamisharina254@gmail.com

Аннотация. Согласно статистике, фактическое время выполнения большинства заданий на суперкомпьютерном кластере существенно отличается от времени, запрошенного пользователем. Исследование распределений времен исполнения задач на суперкомпьютере с использованием статистических методов или методов машинного обучения позволяет оптимизировать работу суперкомпьютерного кластера. Мы изучаем результаты исполнения вычислительных задач в суперкомпьютерном центре Санкт-Петербургского Политехнического университета Петра Великого. Нами разработан непараметрический подход для обнаружения и подтверждения статистической достоверности слабых стохастических порядков. Данный подход основан на категориальном непараметрическом методе сравнений на базе оценок Каплана–Мейера, построенных по независимым группам цензурированных справа наблюдений. Для корректировки уровня достоверности обнаруженных слабых стохастических порядков мы применяем поправку Бонферрони на все рассматриваемые сравнения. Проведен сравнительный статистический анализ распределений времен, необходимых для корректного завершения задач, в различных группах наблюдений, найдены и статистически подтверждены некоторые слабые стохастические порядки.

Ключевые слова: данные типа времени жизни, оценка Каплана–Мейера, критерий типа Вальда, стохастические порядки, суперкомпьютерный кластер, планировщик задач

Финансирование: Исследование выполнено при частичной финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания «Разработка и исследование моделей машинного обучения для решения фундаментальных задач искусственного интеллекта в топливно-энергетическом комплексе» (FSEG-2024-0027). Результаты работы получены с использованием вычислительных ресурсов центра коллективного пользования «Политехнический суперкомпьютерный центр» Санкт-Петербургского политехнического университета Петра Великого (№ 500675, https://ckp-rf.ru/catalog/ckp/500675/).

Для цитирования: Misharina T.A., Malov S.V. Categorical survival analysis of the required job execution times in the hybrid supercomputer center // Computing, Telecommunications and Control. 2025. T. 18, № 2. C. 7–20. DOI: 10.18721/JCSTCS.18201

Introduction

High performance computing is becoming increasingly important in different areas of scientific research and industry. Collective supercomputer centers allow to perform calculations of any complexity to a wide range of users. The operation of a supercomputer center is a complex parallel queuing process of execution of computational jobs over time. An optimal scheduling of entire jobs leads to increasing performance of computations. The most important characteristic of a job is the required supercomputer resource involving the required time for its execution, the number of cores allocated to provide sufficient

Random-Access Memory (RAM) and the job execution quickness. A job scheduled on supercomputer can be divided into computational tasks that can be executed in parallel on different cores. An exit code is obtained at the end of the execution of each job.

The optimization of jobs management systems was discussed by a number of authors. The most famous subject of interest is the job running time and its prediction time given by user. In most cases user overestimate significantly job running time that implies non optimality in job scheduling. A machine learning regression-based method to predict mean running time by a vector of observed futures was studied in [5]. It was shown that the prediction of job running time allows the correction of running times obtained from users that increases sufficiently efficiency of job scheduling. Applications of supervised machine learning algorithms to predict the job running time based on information submitted by user at high performance computing centers was discussed in [17]. Machine learning classification methods to predict a class of running time distribution was under consideration in [4, 18]. The underestimation effect of running time by user was studied in [6]. Note that the most important characteristic of job processing is the required execution time to complete successfully the job, which can be equal to the running time or not available, if the job is terminated. Using the observed running time instead of the required execution time as well as just removing jobs, which was not completed successfully, lead to sufficient underestimation of the required execution time if the number of "unsuccessful" jobs is valuable in compare with the total number of jobs. The right-censored survival data model, which is also applicable in the reliability theory, allows to estimate correctly the distribution of required execution time by using the running time and the indicator, which displays, if the job is completed successfully. Machine learning algorithms to predict the distribution of required execution time and its characteristics based on semiparametric and nonparametric regression models of survival analysis were studied in [14, 21].

Note that machine learning methods are more flexible in compare with statistical ones. The statistical conclusions are restricted to the probabilistic model of the experiment, but the statistical conclusions yield another kind of reliability of obtained results.

Categorical methods for survival right-censored data analysis are widely presented in the literature. In [20], likelihood ratio test for right-censored grouped survival data was studied and the chi-square limit distribution of the likelihood ratio test statistic was obtained. In [10], the chi-square test and the Wald's type test for right-censored survival data was obtained and the comparative analysis of the tests under Pitman alternatives was performed. A parametric Pearson's type test for right-censored survival was studied in [8]. In [1], modified versions of goodness of fit chi-square tests for simple and composite parametric null hypotheses in the nonparametric survival data models under presence or absence of the right censoring were obtained. Presence of one extra degree of freedom of the limit distribution in compare with the classical version of the chi-square test is noted and some examples are given. In [12], another approach was used to obtain chi-square test for complex parametric null hypothesis and the comparative analysis in Pitman's efficiency of the test with another version of chi-square test obtained in [1] was performed. An adaptive version of the chi-square test obtained in [10] with a random data-based choice of grouping intervals is given in [2]. The chi-square tests for testing the null hypothesis that the failure time distributions agree with some known parametric model for hazard rates was given in [9], and the chi-square test for agreement of the failure time distributions with some known semiparametric regression model (e.g., the semiparametric accelerated failure time model) in general case under time dependent covariate was obtained in [3]. Wald-type categorical tests for testing homogeneity null hypotheses in the nonparametric right-censored survival data model used in this work was studied in [15], which is universal and can be more efficient than the linear rank tests commonly used in nonparametric survival analysis under some alternatives.

Moreover, machine learning methods also use for survival right-censored data analysis. Random forest-based algorithms were used to analyze right-censored survival data in [7, 11]. In [19], the authors

propose a new Transformer-based survival model, which estimates the patient-specific survival distribution. Another example of the application of machine learning methods to the analysis of randomly censored survival data is presented in [13]. Here, the authors propose a method based on the Beran estimator using neural kernels to estimate the conditional average treatment effect.

In this work, we study the results of users' jobs processing in the supercomputer center of Peter the Great St. Petersburg Polytechnic University. Since the limitations for the running time and the resource of supercomputer used to execute a job should be determined in advance, it is important to evaluate distributions of main characteristics of a job, which cannot be predicted exactly. We are interested in two important characteristics of the required resource: the execution time required to complete successfully the job in seconds, without taking into account the number of cores allocated, and the required computer execution time that is obtained by multiplying the required execution time by the number of cores allocated. We investigate the distributions of the required execution times and required computer times and perform a comparative analysis of the distributions in different groups of users.

Each observation contains the supercomputer resource used to perform the job: the job processing time (observed execution time), the number of cores and the amount of memory allocated and the exit code, which indicates whether the job was completed successfully or it was interrupted due to an error, user request, lack of memory or time allocated for the job execution. In the latter cases, the job is incomplete and the job execution time is assumed to be censored. We relate the execution time (or computer time) required to complete successfully the user's job with the failure time in right-censored survival data model. Then the job execution time and the indicator of successful completion of user's job, which is determined by the exit code, is the right-censored observation.

We apply categorical nonparametric statistical framework based on contrasts obtained from the Kaplan–Meier estimators in d independent groups of right-censored observations to obtain advanced statistical conclusions on distributions of failure times in different groups of observations. Let T be the job execution time or computer time and U be the censoring time. Each observation (X, δ) contains the job processing time $X = \min(T, U)$ and the indicator $\delta = \mathrm{I}_{\{T \le U\}}$, that is equal to 1, if the exit code indicates that the computational task is completed successfully and 0 otherwise. We study the distribution of T and its dependence on the grouping factor, which reflects the user's area of expertise. We create advanced categorical methods for right-censored survival data and apply them to perform comparative analysis of the distributions of job execution times and computer times T in 11 groups of user's domain of scientific expertise [16].

Wald's type categorical tests for survival data

The main object of statistical analysis is the distribution of the required execution time (or computer time) T. The required execution time T is not observed exactly, if the corresponding job is not completed successfully, in this case, we explain the true execution time of the job as an independent censoring time. A single observation consists of the true execution time $X = \min(T, U)$ and the binary job exit code $\delta = \mathrm{I}_{\{T \le U\}}$, which indicates whether the job was completed successfully or censored. We allow the distributions of the required execution times to differ in different groups of jobs and use a categorical covariate $z \in \{1, ..., d\}$ for grouping the data. The observed data contain the true execution times $X_i = \min(T_i, U_i)$ and the binary exit codes $\delta_i = \mathrm{I}_{\{T_i \le U_i\}}$, where T_i is the required execution time of i-th job, U_i is the independent random censoring time, and the covariate $z_i \in \{1, ..., d\}$, which determines the group, to which a corresponding observation belongs, $i = 1, 2, ..., n$. Let $S_z(t) = \mathbb{P}_z(T > t)$ be a completely unknown survival function of the required execution time in group z, $z = 1, 2, ..., d$. The homogeneity null hypothesis is as follows:

$$H_0 : S_1(t) = S_2(t) = \cdots = S_d(t), \quad t \in (-\infty, \infty).$$

We consider a weaker version of the null hypothesis, which requires the equalities of the survival functions S_j at some fixed points:

$$H_0^*: S_1(\vec{t}) = S_2(\vec{t}) = \cdots = S_d(\vec{t}), \quad \vec{t} = (t_1, \ldots, t_k)^T.$$

Let $\hat{S}_z$ be the Kaplan–Meier estimator of the survival function S_z in different groups, $z = 1, 2, ..., d$. The asymptotic properties of the Kaplan–Meier estimators imply weak convergence

$$\sqrt{n_z}\left(\hat{S}_z(\vec{t}) - S_z(\vec{t})\right) \Rightarrow N(0, \Sigma_j), \quad z = 1, 2, \ldots, d$$

for any fixed vector of time points , where $N(0, \Sigma_j)$ is the mean zero Gaussian distribution with the matrix of covariance Σ_z, n_z is the number of observations in group z.

The matrix of covariance Σ_z has the following form:

$$\Sigma_z = \left(\sigma_{vu}^{(z)}\right)_{v=1,u=1}^{k,k},$$

where $\sigma_{vu}^{(z)}$ is the element of the limit covariance matrix of the values of the Kaplan–Meier estimator at time points t_v and t_u, $v, u = 1, ..., k$.

The term $\sigma_{vu}^{(z)}$ of the covariance matrix can be estimated by the following Greenwood formula:

$$\hat{\sigma}_{vu}^{(z)} = n_z \hat{S}_z(t_v)\hat{S}_z(t_u)\sum_{l=1}^{\min(v,u)} \frac{D_l}{Y_l^*\left(Y_l^* - D_l\right)}, \quad v, u = 1, \ldots, k,$$

where D_l is the number of jobs completed successfully at T_l, Y_l^* is the number of jobs not completed and not censored before T_l.

Then the estimate of the covariance matrix Σ_z has the following form:

$$\hat{\Sigma}_z = \left(\hat{\sigma}_{vu}^{(z)}\right)_{v=1,u=1}^{k,k}.$$

Taking into account the independence of observations in different groups, we obtain the following joint weak convergence:

$$\sqrt{n}\left(\hat{S}^*(\vec{t}) - \hat{S}^*(\vec{t})\right) \Rightarrow N\left(0, D\Sigma^* D\right),$$

where $\hat{S}^* = \left(\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_d\right)$ is the Kaplan–Meier estimator of the vector of survival functions $S^* = (S_1, S_2, \ldots, S_d)^T$;

$$\Sigma^* = \begin{pmatrix} \Sigma_1 & \mathrm{O} & \ldots & \mathrm{O} \\ \mathrm{O} & \Sigma_2 & \ldots & \mathrm{O} \\ \ldots & \ldots & \ddots & \ldots \\ \mathrm{O} & \mathrm{O} & \ldots & \Sigma_d \end{pmatrix},$$

$D = \mathrm{diag}(n^*)$ is the normalizing diagonal matrix with the elements of the vector $n^* = \left(n_1^*, \ldots, n_d^*\right): n_z^* = \sqrt{(n_z, n_z, \ldots, n_z)/n}$, $z = 1, \ldots, d$, at the diagonal; $D\Sigma^*D$ is the limit covariance matrix; O is the matrix of zeroes of size $k \times k$.

Denote $S^*\left(\vec{t}\right) = \theta^* = \left(\theta_1,\ \theta_2,\ \ldots,\ \theta_d\right)^T$ and $\theta_z = \left(\theta_{z1},\ \theta_{z2},\ \ldots,\ \theta_{zk}\right)^T$, $z = 1, ..., d$. Then the null hypothesis can be written as follows:

$$H_0^* : \theta_{11} = \theta_{21} = \cdots = \theta_{d1},\ \ldots,\ \theta_{1k} = \theta_{2k} = \cdots = \theta_{dk}.$$

Let $\psi = C^T \theta^*$, where C^T is the matrix of contrasts of size $q \times m$, $\theta^* = (\theta_1, \theta_2, ..., \theta_m)^T$, is the vector of size $m \times 1$. The contrasts matrix $\psi = (\psi_1, ..., \psi_q)^T$ contains linear functions of parameter ψ_j such that:

$$\psi_j = c_{1j}\theta_1 + c_{2j}\theta_2 + \cdots + c_{mj}\theta_m,\ \ \sum_{i=1}^{m} c_{ij} = 0,$$

where c_{ij} are the elements of the matrix C, $i = 1, 2, ..., m$, $j = 1, 2, ..., q$.

We use the following pairwise contrasts:

$$\begin{gathered}
\psi_1 = S_1\left(t_1\right) - S_2\left(t_1\right);\\
\cdots\\
\psi_k = S_1\left(t_k\right) - S_2\left(t_k\right);\\
\psi_{k+1} = S_2\left(t_1\right) - S_3\left(t_1\right);\\
\cdots\\
\psi_{2k} = S_2\left(t_k\right) - S_3\left(t_k\right);\\
\vdots\\
\psi_{(d-1)k} = S_{d-1}\left(t_k\right) - S_d\left(t_k\right).
\end{gathered}$$

In terms of the parameters $\theta_{ij} = S_i(t_j)$, $i = 1, 2, ..., d$, $j = 1, 2, ..., k$, the vector function of contrasts ψ can be rewritten in the following matrix form:

$$\psi = C^T\theta^* = \begin{pmatrix} E & -E & \mathrm{O} & \cdots & \mathrm{O} & \mathrm{O} \\ \mathrm{O} & E & -E & & \mathrm{O} & \mathrm{O} \\ & \vdots & & \ddots & & \vdots \\ \mathrm{O} & \mathrm{O} & \mathrm{O} & \cdots & E & -E \end{pmatrix} \cdot \begin{pmatrix} \theta_{11} \\ \vdots \\ \theta_{1k} \\ \theta_{21} \\ \vdots \\ \theta_{2k} \\ \vdots \\ \theta_{d1} \\ \vdots \\ \theta_{dk} \end{pmatrix},$$

where E is the identity matrix of $k \times k$, O is the $k \times k$-matrix of zeroes, the matrix of contrasts C^T is of size $kd \times k(d-1)$, the parameter θ^* is of size $kd \times 1$ and the contrasts vector ψ is of size $k(d-1) \times 1$.

In terms of the contrasts ψ the null hypothesis can be written as follows:

$$H_0^* : \psi_1 = \psi_2 = \cdots = \psi_{k(d-1)} = 0.$$

The asymptotic normality of the estimators $\hat{\theta}^* = \hat{S}^*(\vec{t})$ implies immediately the asymptotic normality of the estimators $\hat{\psi}$ of the corresponding contrasts ψ:

$$\sqrt{n}(\hat{\psi} - \psi) \Rightarrow N(0, C^T D \Sigma^* DC). \quad (1)$$

Let $\Gamma_\psi = C^T D \Sigma^* DC$ and $\hat{\Gamma}_\psi = C^T D \hat{\Sigma}^* DC$, where $\hat{\Sigma}^*$ is a consistent estimate of the asymptotic variance of the estimator $\hat{\psi}$ under the null hypothesis. The Wald's type test statistic for testing H_0^*.

$$n\hat{\psi}^T \hat{\Gamma}_\psi^{-1} \hat{\psi} \Rightarrow \chi_q^2,$$

has asymptotical χ_q^2-distribution under the null hypothesis. Under the fixed alternative $H_A^* : \psi = \psi_0$ the Wald's type test statistic has an asymptotical non-central $\chi_{\mu,q}^2$-distribution with the non-centrality parameter

$$\mu = n\psi_0^T \Gamma_\psi^{-1} \psi_0.$$

Detection and testing significance of stochastic orders

Let X and Y be random variables. The random variable X is stochastically less than the random variable Y $\left(X \le^{st} Y\right)$, if

$$F_X(t) \ge F_Y(t) \quad \text{for all} \quad t \in (-\infty, \infty),$$

where $F_X(t) = P\{X \le t\}$ and $F_Y(t) = P\{Y \le t\}$, $t \in (-\infty, \infty)$, are the distribution functions of X and Y, respectively. In case of X and Y are failure times, it is convenient to rewrite the same property as follows:

$$S_X(t) \le S_Y(t) \quad \text{for all} \quad t \in (-\infty, \infty),$$

where $S_X(t) = P\{X > t\}$ and $S_Y(t) = P\{Y > t\}$, $t \in (-\infty, \infty)$, are the survival functions of X and Y, respectively. The relation $X \le^{st} Y$ determines the partial non-strict order on the set of distributions of random variables. In a similar manner, we say the random variables $X_1, X_2, ..., X_d$ are completely stochastically ordered

$$X_1 \le^{st} X_2 \le^{st} \ldots \le^{st} X_d$$

if $X_i \le^{st} X_{i+1}$ for $i = 1, ..., d-1$. By the transitivity property of the stochastic order the random variables are stochastically ordered $X_1 \le^{st} X_2 \le^{st} \ldots \le^{st} X_d$, if $X_i \le^{st} X_j$ for all $1 \le i < j \le d$. We say that random variables $X_1, X_2, ..., X_d$ are completely stochastically ordered, if there exists a permutation $(\sigma_1, \sigma_2, ..., \sigma_d)$ of indices $(1, 2, ..., d)$, such that $X_{\sigma_1} \le^{st} X_{\sigma_2} \le^{st} \ldots \le^{st} X_{\sigma_d}$. If the stochastic orders $X_{\sigma_i} \le^{st} X_{\sigma_j}$ hold for some pairs of σ_i and σ_j only, then we report the incomplete stochastic order.

We use a special nonparametric approach to state stochastic orders of failure times in different groups of observations with high reliability. Since the survival function of failure time is equal to 1 at point zero and is tending to 0 as the argument is tending to infinity, the stochastic order cannot be checked in the nonparametric model. The stochastic ordering condition can be relaxed. We say that the random variable X is stochastically smaller than the random variable Y in the weak sense $\left(X \le_\Delta^{st} Y\right)$ with respect to the set Δ, if

$$S_X(t) \le S_Y(t) \quad \text{for all} \quad t \in \Delta, \quad (2)$$

where Δ is some bounded set of positive real numbers. Similarly, the complete stochastic order $X_1 \le^{st} X_2 \le^{st} \ldots \le^{st} X_d$ holds, if

$$S_1(t) \le S_2(t) \le \cdots \le S_d(t) \quad \text{for all} \quad t \in (-\infty, \infty),$$

whereas the corresponding weak stochastic order $X_1 \le_{\Delta}^{st} X_2 \le_{\Delta}^{st} \cdots \le_{\Delta}^{st} X_d$ with respect to Δ holds, if

$$S_1(t) \le S_2(t) \le \cdots \le S_d(t) \quad \text{for all} \quad t \in \Delta,$$

The weak stochastic order $X_1 \le_{\Delta}^{st} X_2 \le_{\Delta}^{st} \cdots \le_{\Delta}^{st} X_d$ can be obtained from $d-1$ pairwise stochastic orders $X_i \le_{\Delta}^{st} X_{i+1}$, $i = 1, ..., d-1$, or, in terms of survival functions,

$$S_{X_i}(t) \le S_{X_{i+1}}(t) \quad \text{for all} \quad t \in \Delta \quad \text{and} \quad i = 1, \ldots, d-1.$$

Let $\Delta = \{t_1, t_2, ..., t_k\}$ be a finite set. Then (2) can be rewritten as follows:

$$S_X(t_s) \le S_Y(t_s), \quad s = 1, \ldots, k.$$

It seems natural to choose the checkpoints t_s, which cover each of the supports of X and Y, but the statistical framework is inefficient, if the distribution of X is highly biased with respect to Y in this case. In order to increase the efficiency of the statistical analysis, we choose the checkpoints empirically from the combined distribution for each pair of distributions X_i and X_j, $i \ne j$. Let $(\sigma_1, \sigma_2, ..., \sigma_s)$, $s \le d$, be an arrangement of the indices (1, 2, ..., d); $\left\{\Delta^{(i,j)}\right\}_{i,j=1}^{d} : \Delta^{(i,j)} = \Delta^{(i,j)} \subset (0, \infty)$ is determined for each pair of distributions of X_i and X_j, $i, j \in \{1, ..., d\}$. We say the conditional weak stochastic order $X_{\sigma_1} \le_{\Delta(\sigma_1,\sigma_2)}^{st} \ldots \le_{\Delta(\sigma_{s-1},\sigma_s)}^{st} x_{\sigma_s}$ with respect to $\left\{\Delta^{(i,j)}\right\}_{i,j=1}^{d}$ holds, if

$$S_{\sigma_i}(\vec{t}) \le S_{\sigma_j}(\vec{t}) \quad \text{for all} \quad \vec{t} \in \Delta^{(\sigma_i,\sigma_j)}, \quad 1 \le i < j \le s. \tag{3}$$

In other words, the conditional weak stochastic order $X_{\sigma_1} \le_{\Delta(\sigma_1,\sigma_2)}^{st} \ldots \le_{\Delta(\sigma_{s-1},\sigma_s)}^{st} x_{\sigma_s}$ with respect to $\left\{\Delta^{(i,j)}\right\}_{i,j=1}^{d}$ holds, if $S_{\sigma_i} \le_{\Delta(\sigma_i,\sigma_j)}^{st} S_{\sigma_j}$ for all $1 \le i < j \le s$. Note that the pairwise relation of the conditional stochastic order is not transitive in general case, since the stochastic order in each pair is determined in a different weak sense.

Let the survival data contain d groups of independent right-censored observations with failure times T_i having completely unknown survival functions S_i within i-th group, $i = 1, ..., d$. Set $\vec{t}^{(i,j)} = \left(t_1^{(i,j)}, t_2^{(i,j)}, \ldots, t_k^{i,j}\right)$ are checkpoints (that can be data based in general case) for weak pairwise stochastic ordering of distributions T_i and T_j for each pair of groups i and j; $\Delta^{(i,j)} = \left(t_1^{(i,j)}, t_2^{(i,j)}, \ldots, t_k^{i,j}\right)$.

We confirm the pairwise weak stochastic order $T_i \le_{\Delta(i,j)}^{st} T_j$ at the confidence level $1 - \alpha$, if each of the particular left sided confidence intervals of level $1 - \alpha/d$ for all of k contrasts $\psi_1^{(i,j)}, \ldots, \psi_k^{(i,j)}$, where $\psi_s^{(i,j)} = S_i\left(t_s^{(i,j)}\right) - S_j\left(t_s^{(i,j)}\right)$, are located entirely to the left of zero. In other words, we obtain joint confidence intervals for the contrasts by using the Bonferroni method. The particular right sided asymptotic confidence intervals are obtained from the asymptotic normality (1). The conditional stochastic order including more than two groups can be confirmed at some confidence level in a similar manner by using the right sided confidence intervals for the contrasts related to all the pairwise orders that determine the conditional stochastic order with the Bonferroni correction on the total number of contrasts.

We also report the p-value that allows to estimate the true confidence of the statistical conclusion. Note that a pairwise weak stochastic order can be confirmed with some confidence, only if (2) holds for the Kaplan–Meier estimators for each $t \in \Delta$. If the pairwise weak stochastic order for the Kaplan–Meier estimators fail, we report the p-value is equal to 1 and the corresponding confidence is estimated equal to 0. Otherwise, the p-value is determined as the infimum of α, such that the pairwise weak stochastic order holds at the confidence level $1 - \alpha$. Since a conditional weak stochastic order of 3 and more distributions is determined by the corresponding weak pairwise stochastic orders, the p-value of the conditional weak stochastic order is given as the maximal of p-values of the pairwise stochastic orders. The pairwise stochastic orders are obtained by using the confidence intervals for the contrasts with the correction to the number of weak pairwise stochastic orders that determines the conditional weak stochastic order. Finally, the estimator for the true confidence of a weak stochastic order is equal to $1 - p$-value.

Another application of the contrasts method is the detection of available stochastic orders and the confirmation. Note that the whole range of (non-conditional) weak stochastic orders can be determined by using $d(d-1)/2$ pairwise weak stochastic orders and both the alternative pairwise stochastic orders $X_i \leq_{\Delta}^{st} X_j$ and $X_j \leq_{\Delta}^{st} X_i$ related to the same pair (i, j) can be obtained by using the same k contrasts $\psi_s^{(i,j)}$, $s = 1, ..., k$. If all the two-sided joint confidence intervals for the contrasts lie entirely to the left of zero, we confirm that $T_i \leq_{\Delta(i,j)}^{st} T_j$, whereas if all they lie entirely to the right of zero, we confirm that $T_j \leq_{\Delta(i,j)}^{st} T_i$ with the same confidence as the joint confidence level of the intervals. Hence, only $\frac{kd(d-1)}{2}$ contrasts are required to detect all the pairwise weak stochastic orders for any distributions of $T_1, ..., T_d$.

The conditional weak stochastic orders of 3 and more distributions can be obtained from pairwise weak stochastic orders. We consider the combination of all pairs for which there are the arrangements $(\sigma_1, \sigma_2, ..., \sigma_s)$, $s \leq d$ of indices $(1, 2, ..., d)$, such that (3) holds. For example, based on pairwise weak stochastic orders $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_2)}^{st} T_{\sigma_2}$, $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_3)}^{st} T_{\sigma_3}$, $T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_3)}^{st} T_{\sigma_3}$, we construct conditional weak stochastic order $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_2)}^{st} T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_3)}^{st} T_{\sigma_3}$. In addition, we confirm the following pairwise weak stochastic orders: $T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_4)}^{st} T_{\sigma_4}$, $T_{\sigma_3} \leq_{\Delta(\sigma_3,\sigma_4)}^{st} T_{\sigma_4}$. Then we obtain the conditional weak stochastic orders $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_2)}^{st} T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_3)}^{st} T_{\sigma_3}$ and $T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_3)}^{st} T_{\sigma_3} \leq_{\Delta(\sigma_3,\sigma_4)}^{st} T_{\sigma_4}$, but not the conditional weak stochastic order $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_2)}^{st} T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_3)}^{st} T_{\sigma_3} \leq_{\Delta(\sigma_3,\sigma_4)}^{st} T_{\sigma_4}$, since we do not confirm the pairwise weak stochastic order $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_4)}^{st} T_{\sigma_4}$.

We consider the whole range of the two-sided joint confidence intervals for all $\frac{kd(d-1)}{2}$ the contrasts and confirm all available conclusions on the pairwise weak stochastic orders at the confidence level $(1 - \alpha)$. Since we detect conditional stochastic orders, we are ready to reject all inconsistent pairwise stochastic orders.

Statistical data and planning of statistical analysis

The statistical data contains the results of users' jobs processing at the supercomputer center of Peter the Great St. Petersburg Polytechnic University. For each run initiated by the corresponding user's job, we have the processing time of computational task, the number of cores allocated and the exit code,

which allows to determine, whether the user's job was completed successfully. Runs of duration less than 5 seconds were removed. Finally, we use information on 1338565 runs from 01.09.2021 to 31.08.2023. Runs that were not completed or not completed successfully are assumed to be censored.

We analyze distributions of the execution time required to complete successfully user's job, in seconds, and the computer time (spent processor time), in processor seconds (sec.* CPU). All user jobs and corresponding runs were classified to 11 groups by user's area of expertise:

- astrophysics;
- bioinformatics;
- biophysics;
- energetics;
- geophysics;
- IT;
- mechanical engineering;
- mechanics;
- physics;
- radiophysics;
- a special group called geovation [10].

The Kaplan–Meier estimators of the survival functions of the required times and computer times to complete successfully user's job are visualized in Figs. 1 and 2.

First, we test the homogeneity null hypothesis that the distributions of the required times (or computer times) to complete successfully user's job in different groups are all the same, as well as the pairwise homogeneity null hypotheses each pair of groups separately by using Wald's type tests. If the null hypothesis of homogeneity is rejected, we perform advanced statistical analysis using contrasts method for each pair of the groups, for which significant differences in distributions of the required times (or computer times) to complete successfully user's job were found, adjusted to the total number of pairs. The conditional weak stochastic orders of 3 and more distributions are obtained from the confirmed pairwise stochastic orders according to (3).

The checkpoints for pairwise homogeneity testing and further advanced analysis of contrasts are obtained in the following way:

1. We obtain $t_{\max}$ is the largest observed failure time of for each of samples.
2. The group with the smaller value of $t_{\max}$ is assumed to be a baseline group.
3. The checkpoints are defined as 7 octiles (12.5%; 25%; 37.5%; 50%; 62.5%; 75%; 87.5%) of the Kaplan–Meier estimator related to the baseline group and the midpoint between the last octile and $t_{\max}$, totally $k = 8$ of the checkpoints.

The checkpoints are consistent estimates of the corresponding numerical characteristics that depend on the joint distribution of failure and censoring times. Then the asymptotic normality of the Kaplan–Meier estimators at the checkpoints is preserved under the null hypothesis and under a fixed alternative.

We use the significance level $\alpha = 0.05$ (5%) and the joint confidence level $1 - \alpha = 0.95$ for all statistical conclusions.

Results of statistical analysis

Testing the homogeneity null hypotheses displays significant differences in distributions of the required times and computer times to complete successfully user's job both with the p-value not exceeding 10^{-300}, the minimal available value in R. Testing the pairwise homogeneity null hypothesis displays highly significant differences in distributions of the required times and computer times to complete successfully user's job for each pair of times and computer times as well with the maximal p-value $8.3 \cdot 10^{-67}$ for times in astrophysics and mechanics groups.

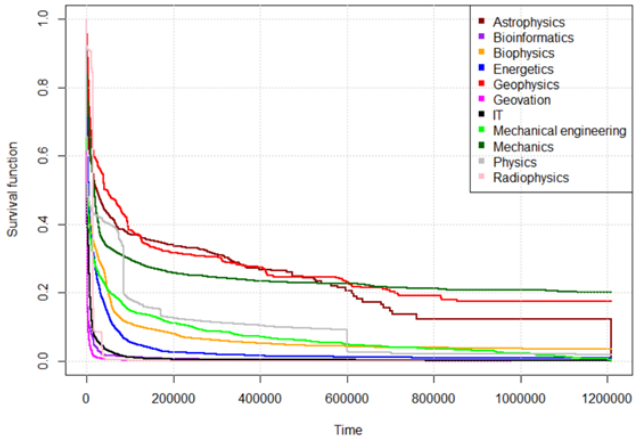

Fig. 1. Kaplan–Meier estimators of the required job execution times

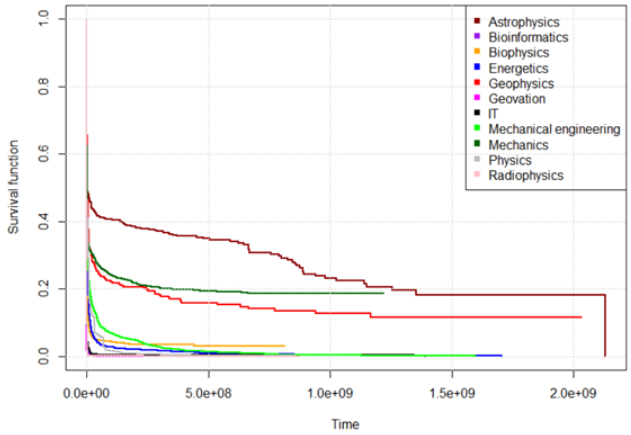

Fig. 2. Kaplan–Meier estimators of the required job execution computer times

The results of detection and confirmation of the obtained pairwise weak stochastic orders in the distributions of the required times and computer times to complete successfully user's job are visualized by using directed graphs in Figs. 3 and 4, respectively: each vertex represents a group by the user's area of expertise; an edge exists, if the stochastic order is confirmed at the joint confidence level of 0.95 adjusted to the total number of pairs; each edge is directed from a larger distribution to a smaller one.

We detect and confirm 21 pairwise weak stochastic orders for the required times to complete successfully user's job and 23 pairwise weak stochastic orders for the required computer times to complete successfully user's job.

The graph shows that we also detect and confirm weak stochastic orders for the three groups, for example, the required times in the geophysics group is stochastically larger than the required times in the biophysics group, which is stochastically larger than that in the geovation group.

We detect and confirm 4 triple weak stochastic orders for the required times to complete successfully user's job and 10 triple weak stochastic orders for the required computer times to complete successfully user's job.

Discussion

In this study, we develop the statistical framework for detection and confirmation, at some confidence level, of weak stochastic orders in distributions of failure times from right-censored survival data.

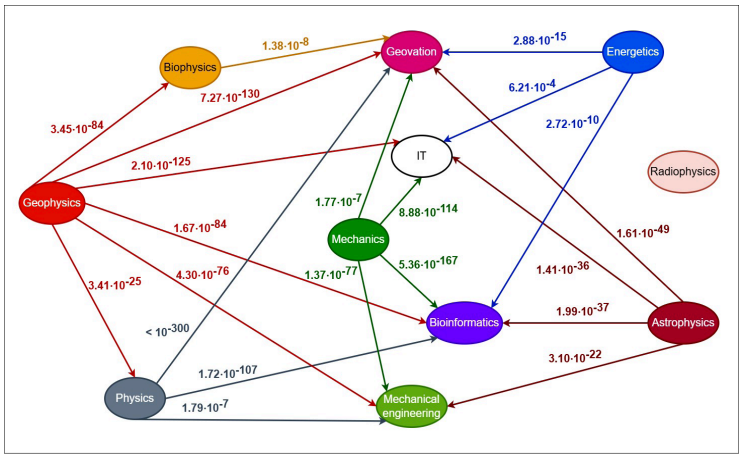

Fig. 3. Significant conditional pairwise weak stochastic orders
in distributions of the required times to complete successfully user's job

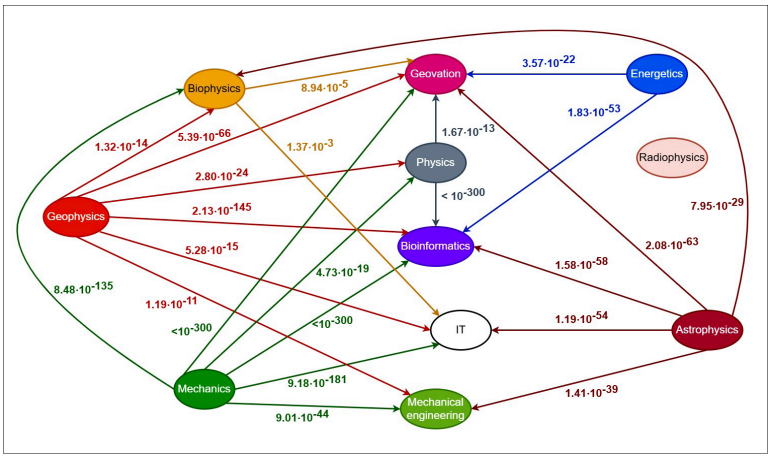

Fig. 4. Significant conditional pairwise weak stochastic orders
in distributions of the required computer times to complete successfully user's job

The set of tools for nonparametric categorical analysis of right-censored survival data based on the Kaplan–Meier and the Nelson–Aalen estimators, as well as the contrasts methods for detection and confirmation of weak stochastic orders, were implemented in the R software development environment. The Bonferroni correction is applied to adjust the confidence level for all contrasts under consideration.

We analyze the results of users' jobs processing obtained at the supercomputer center of Peter the Great St. Petersburg Polytechnic University. We group users' jobs into 11 groups by the user's area of expertise. The main objects of interest are the required times and computer times to complete successfully the user's job in different groups of users. We associate these characteristics with the failure time in right-censored data model. Testing the homogeneity null hypotheses of failure time distributions in different groups of users, as well as each of pairwise homogeneity null hypotheses, reveals non-random differences in the corresponding estimators in different groups of users with extremely high significance, close to absolute, for both required times and computer times to complete successfully user's job.

We detect and confirm at the 95% confidence level 21 pairwise weak stochastic orders for the required times to complete successfully user's job and 23 pairwise weak stochastic orders for the required computer times to complete successfully user's job.

Note that the obtained stochastic orders are a much more informative result than simply establishing the significant differences in the distributions. In particular, $T_1 \leq^{st} T_2$ implies that the mean value and all quantiles of T_1 are smaller than the corresponding characteristics of T_2. In some cases, weak stochastic order does not guarantee the existence of a corresponding stochastic order, but it is useful, because it allows us to draw conclusions for the corresponding quantiles. These conclusions are applicable to optimize user's jobs processing.

REFERENCES

1. **Akritas M.G.** Pearson-Type Goodness-of-Fit Tests: The Univariate Case. *Journal of the American Statistical Association*, 1988, Vol. 83, No. 401, Pp. 222–230. DOI: 10.2307/2288944

2. **Bagdonavicius V.B., Nikulin M.S.** Chi-squared goodness-of-fit test for right-censored data. *International Journal of Applied Mathematics & Statistics*, 2011, Vol. 24, No. SI-11A, Pp. 30–50.

3. **Bagdonavičius V., Levuliene R, Nikulin M.S., Tran Q.X.** On chi-square type tests and their applications in survival analysis and reliability. *Journal of Mathematical Sciences*, 2014, Vol. 199, Pp. 88–99. DOI: 10.1007/s10958-014-1835-x

4. **Baranov A.V., Nikolaev D.S.** Machine learning to predict the supercomputer jobs execution time. *Software & Systems*, 2020, Vol. 33, No. 2, Pp. 218–228. DOI: 10.15827/0236-235X.130.218-228.

5. **Gaussier E., Glesser D., Reis V., Trystram D.** Improving backfilling by using machine learning to predict running times. SC'15: *International Conference for High Performance Computing, Networking, Storage and Analysis*, 2015, Pp. 1–10. DOI: 10.1145/2807591.2807646

6. **Guo J., Nomura A., Barton R., Zhang H., Matsuoka S.** Machine learning predictions for underestimation of job runtime on HPC system. In: *Supercomputing Frontiers* (eds. R. Yokota, W. Wu), 2018, Vol .10776. DOI: 10.1007/978-3-319-69953-0_11

7. **Ishwaran H., Kogalur U.B., Blackstone E.H., Lauer M.S.** Random survival forests. *The Annals of Applied Statistics*, 2008, Vol. 2, No. 3, Pp. 841–860. DOI: 10.1214/08-AOAS169

8. **Habib M.G., Thomas D.R.** Chi-square goodness-if-fit tests for randomly censored data. *The Annals of Statistics*, 1986, Vol. 14, No. 2, Pp. 759–765. DOI: 10.1214/aos/1176349953

9. **Hjort N.L.** Goodness of fit tests in models for life history data based on cumulative hazard rates. *The Annals of Statistics*, 1990, Vol. 18, No. 3, Pp. 1221–1258. DOI: 10.1214/aos/1176347748

10. **Hollander M., Peña E.A.** A chi-squared goodness-of-fit test for randomly censored data. *Journal of the American Statistical Association*, 1992, Vol. 87, No. 418, Pp. 458–463. DOI: 10.2307/2290277

11. **Hothorn T, Bühlmann P, Dudoit S, Molinaro A., Van der Laan M.J.** Survival ensembles. *Biostatistics*, 2006, Vol. 7, No. 3, Pp. 355–373. DOI: 10.1093/biostatistics/kxj011

12. **Kim J.H.** Chi-square goodness-of-fit tests for randomly censored data. *The Annals of Statistics*, 1993, Vol. 21, No. 3, Pp. 1621–1639. DOI: 10.1214/aos/1176349275

13. **Kirpichenko S., Utkin L., Konstantinov A., Muliukha V.** BENK: The Beran estimator with neural kernels for estimating the heterogeneous treatment effect. *Algorithms*, 2024, Vol. 17, No. 1, Art. no. 40. DOI: 10.3390/a17010040

14. **Konstantinov A.V.** Predictive models and dynamics of estimates of applied tasks characteristics using machine learning methods. *Computing, Telecommunications and Control*, 2024, Vol. 17, No. 3, Pp. 54–60. DOI: 10.18721/JCSTCS.17305

15. **Malov S., O'Brien S.** On survival categorical methods with applications in epidemiology and AIDS research. In: *Applied Methods of Statistical Analysis. Applications in Survival Analysis, Reliability and Quality Control* (eds. B. Lemeshko, M. Nikulin, N. Balakrishnan), 2013, Pp. 173–180.

16. **Malov S.V., Lukashin A.A.** Count time series analysis of jobs scheduling in the hybrid supercomputer center. *Computing, Telecommunications and Control*, 2024, Vol. 17, No. 3, Pp. 42–53. DOI: 10.18721/JCSTCS.17304

17. **McKenna R., Herbein S., Moody A., Gamblin T., Taufer M.** Machine learning predictions of runtime and IO traffic on high-end clusters. *2016 IEEE International Conference on Cluster Computing* (*CLUSTER*), 2016, Pp. 255–258. DOI: 10.1109/CLUSTER.2016.58

18. **Savin G.I., Shabanov B.M., Nikolaev D.S., Baranov A.V., Telegin P.N.** Jobs runtime forecast for JSCC RAS supercomputers using machine learning methods. *Lobachevskii Journal of Mathematics*, 2020, Vol. 41, Pp. 2593–2602. DOI: 10.1134/S1995080220120343

19. **Hu S., Fridgeirsson E., van Wingen G., Welling M.** Transformer-based deep survival analysis. *Proceedings of AAAI Spring Symposium on Survival Prediction – Algorithms, Challenges and Applications*, 2021, Vol. PLMR 146, Pp. 132–148.

20. **Turnbull B.W., Weiss L.** A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, 1978, Vol. 34, No. 3, Pp. 367–375.

21. **Utkin L.V., Konstantinov A.V., Eremenko D.Yu. et al.** Interpretation methods for machine learning models in the framework of survival analysis with censored data: a brief overview. *Computing, Telecommunications and Control*, 2024, Vol. 17, No. 3, Pp. 22–31. DOI: 10.18721/JCSTCS.17302

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Misharina Tatiana A.
Мишарина Татьяна Андреевна
E-mail: tanechkamisharina254@gmail.com

Malov Sergey V.
Малов Сергей Васильевич
E-mail: sergey.v.malov@gmail.com
ORCID: https://orcid.org/0000-0003-0093-6506

Submitted: 09.11.2024; Approved: 21.04.2025; Accepted: 13.05.2025.
Поступила: 09.11.2024; Одобрена: 21.04.2025; Принята: 13.05.2025.