Computing, Telecommunication and Control, 2026, Vol. 19, No. 2, Pp. 31–43.
Информатика, телекоммуникации и управление. 2026. Том 19, № 2. С. 31–43.

Research article
DOI: https://doi.org/10.18721/JCSTCS.19203
UDC 004.85

SURVIVAL ANALYSIS USING MAHALANOBIS DISTANCE IN KERNELS OF THE BERAN ESTIMATOR

L.V. Utkin ✉, *A.V. Konstantinov*, *S.R. Kirpichenko*, *N.M. Verbova*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ lev.utkin@gmail.com

Abstract. Survival analysis is critical for modeling time-to-event data across medicine, engineering, and economics. The Beran estimator serves as an efficient nonparametric tool for estimating conditional survival functions under censoring. However, the conventional Beran estimator relies on the Euclidean distance for kernel weighting, a metric that fails to account for heterogeneous scales and correlations among covariates. This limitation can result in suboptimal weighting when features exhibit multicollinearity or varying measurement units. To address this, we propose the M-Beran estimator, a generalized formulation that incorporates the Mahalanobis distance into the kernel smoothing scheme. This modification ensures scale invariance and leverages the underlying correlation structure of the feature space, transforming the kernel into an elliptically symmetric form. We evaluate the proposed method through numerical experiments on real-world survival datasets. Results demonstrate that the M-Beran estimator outperforms the conventional Euclidean-distance-based Beran estimator, offering improved predictive accuracy.

Keywords: machine learning, survival analysis, Beran estimator, Mahalanobis distance, Harrell's C-index, survival function

Acknowledgements: The research was 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 for the fuel and energy complex" (FSEG-2024-0027).

Citation: Utkin L.V., Konstantinov A.V., Kirpichenko S.R., Verbova N.M., Survival analysis using Mahalanobis distance in Kernels of the Beran estimator. Computing, Telecommunications and Control, 2026, Vol. 19, No. 2, Pp. 31–43. DOI: 10.18721/JCSTCS.19203

© Utkin L.V., Konstantinov A.V., Kirpichenko S.R., Verbova N.M., 2026. Published by Peter the Great St. Petersburg Polytechnic University

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.19203
УДК 004.85

АНАЛИЗ ВЫЖИВАЕМОСТИ С ИСПОЛЬЗОВАНИЕМ РАССТОЯНИЯ МАХАЛАНОБИСА В ЯДРАХ МОДЕЛИ БЕРАНА

Л.В. Уткин ✉, *А.В. Константинов*, *С.Р. Кирпиченко*, *Н.М. Вербова*

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

✉ lev.utkin@gmail.com

Аннотация. Анализ выживаемости имеет решающее значение для моделирования данных о времени до наступления события в медицине, инженерии и экономике. Оценка Берана служит эффективным непараметрическим инструментом для оценки условных функций выживаемости при цензурированных данных. Однако стандартная оценка Берана основана на евклидовом расстоянии для взвешивания ядер – метрике, которая не учитывает различие в масштабах признаков и корреляции между ними. Эти ограничения могут привести к неоптимальному взвешиванию, когда признаки демонстрируют мультиколлинеарность или различные единицы измерения. Для решения этой проблемы мы предлагаем оценку М-Берана – обобщенную формулировку, которая включает расстояние Махаланобиса в схему сглаживания ядер. Эта модификация обеспечивает инвариантность к масштабу данных и использует лежащую в основе корреляционную структуру пространства признаков, преобразуя ядро в эллиптически симметричную форму. Мы оцениваем предложенный метод с помощью численных экспериментов на реальных наборах данных времен до событий. Результаты показывают, что оценка М-Берана превосходит традиционную оценку Берана на основе евклидова расстояния, обеспечивая более высокую точность предсказаний.

Ключевые слова: машинное обучение, анализ выживаемости, оценка Берана, расстояние Махаланобиса, С-индекс, функция выживаемости

Финансирование: Исследование выполнено при 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).

Для цитирования: Utkin L.V., Konstantinov A.V., Kirpichenko S.R., Survival analysis using Mahalanobis distance in Kernels of the Beran estimator // Computing, Telecommunications and Control. 2026. Т. 19, № 2. С. 31–43. DOI: 10.18721/JCSTCS.19203

Introduction

Survival analysis, the statistical study of time-to-event data, is essential across diverse fields [1–4]. It guides clinical trials and disease modeling in medicine, predicts system failure in engineering, and assesses churn or turnover in economics, social sciences, and marketing. Characterized by censoring, where events may be unobserved, this data complexity motivates the continued development of robust estimation methods. Among the most fundamental tools for flexible survival analysis is the Beran estimator [5]. Given a set of censored survival times and a vector of covariates, the Beran estimator provides a nonparametric estimate of the conditional survival function (SF) by applying kernel smoothing directly to the observed events and censoring processes. Its appeal lies in its ability to model complex, nonlinear relationships without imposing restrictive parametric assumptions such as proportional hazards [6], making it particularly valuable in applications where the underlying mechanisms are poorly

© Уткин Л.В., Константинов А.В., Кирпиченко С.Р., Вербова Н.М., 2026. Издатель: Санкт-Петербургский политехнический университет Петра Великого

understood or where standard models fail. The estimator's formulation relies on a kernel function that assigns weights to neighboring observations based on their proximity in the feature space, typically employing the Euclidean distance.

Despite its flexibility, the conventional Beran estimator inherits a fundamental limitation from its reliance on the Euclidean metric. In practical applications, covariate vectors often exhibit heterogeneous scales, complex correlations, and multicollinearity – features that the Euclidean distance is ill-equipped to handle. Consider, for instance, a clinical study of cancer survival where covariates include tumor size (measured in millimeters), patient age (in years), and multiple correlated biomarkers; in engineering reliability, correlated sensor measurements from a monitoring system. In each case, the Euclidean metric treats each covariate dimension independently and equally. Moreover, it ignores the underlying correlation structure of the covariate distribution, which can lead to suboptimal weighting: two patients who appear similar in Euclidean terms may, in fact, be statistically dissimilar if they deviate from the dominant correlation patterns of the biomarker panel, potentially compromising the accuracy of survival predictions that inform critical clinical decisions.

A solution to these challenges lies in replacing the Euclidean distance with the Mahalanobis distance [7].

The Mahalanobis distance offers several advantages that are directly relevant to real-world survival analysis applications. First, it is scale-invariant, automatically accounting for differences in measurement units, whether tumor size in millimeters, gene expression on a log scale, or sensor voltages in engineering systems, and eliminating the need for separate bandwidths per covariate. Second, and more importantly, it incorporates the correlation structure of the data, so that distances are measured in terms of statistical fluctuations rather than raw coordinate differences. Observations that lie along directions of high variance are considered closer, while those that deviate against the correlation structure are appropriately penalized. This property is valuable in applications with redundant or highly correlated predictors, such as the analysis of multiple clinical biomarkers, correlated survey responses in social science event history studies, or multi-sensor reliability data.

In the context of kernel-based nonparametric estimation, substituting the Euclidean distance with the Mahalanobis distance transforms the kernel from a product of independent univariate kernels into a single, elliptically symmetric kernel defined on the standardized and decorrelated feature space. This simple modification carries profound implications for applied survival analysis. It can help to maximize the predictive performance of the Beran estimator, such as by maximizing the conditional likelihood of the observed survival times. While the original Beran estimator and its theoretical properties have been extensively studied [5, 8, 9], the systematic incorporation of the Mahalanobis distance into its kernel structure remains relatively underexplored.

The Mahalanobis distance attracts attention in tasks of survival analysis. For example, a nonparametric weighting approach was proposed in [10], which balances the covariates between treatment groups by Kernel Mean Matching (KMM) with the Mahalanobis distance kernels. A linear-mixed effects model and the Cox model for survival time are considered in [11, 12], where the Mahalanobis distance is used to detect outliers in survival data. In [13], a Moving Multivariate Distance (MMD), based on the Mahalanobis distance, was proposed to quantify the variability of the multivariate biomarker profile as a whole from one visit to the next. In [14], flexible machine learning models were evaluated, capable of incorporating interactions and nonlinear effects from high-dimensional data using the Mahalanobis distance to estimate individualized treatment recommendations in trials with time-to-event outcomes. In [15], a method was proposed to match each untreated patient with the nearest treated patient based on the Mahalanobis distance between patient characteristics.

Techniques based on the Mahalanobis distance and applied in different fields of chemometrics such as in multivariate calibration, pattern recognition and process control are explained and discussed in [16]. The most important application of the Mahalanobis distance is related to out-of-distribution

detection and outlier detection [17–20]. Application of the Mahalanobis distance to a k-means clustering algorithm was studied in [21]. In [22], it is proposed using a Mahalanobis distance metric for computing the attention weights to stretch the underlying feature space in directions of high contextual relevance.

However, to the best of our knowledge, at the moment, no approach is known that incorporates the Mahalanobis distance to the Beran estimator. Therefore, following [22], we propose a generalized Beran estimator that incorporates the Mahalanobis distance into its kernel weighting scheme. We called the proposed model as the *M-Beran estimator*. Through real data, it is shown that the M-Beran estimator outperforms the conventional Beran estimator, using the Euclidean distance.

Background

The Mahalanobis Distance

The Mahalanobis distance is a measure of distance that accounts for the scale and correlation structure of multivariate data [7]. Unlike the Euclidean distance, which treats all dimensions independently, the Mahalanobis distance provides a metric that is invariant to affine transformations of the covariate space. Let $\boldsymbol{X} = \left(x_1, \ldots, x_p\right)^T$ be a p-dimensional random vector with mean vector $\boldsymbol{\mu} = \mathbb{E}\left[\boldsymbol{X}\right]$ and positive-definite covariance matrix

$$\boldsymbol{M} = \mathrm{Cov}\left(\boldsymbol{X}\right) = \mathbb{E}\left[\left(\boldsymbol{X} - \boldsymbol{\mu}\right)\left(\boldsymbol{X} - \boldsymbol{\mu}\right)^T\right].$$

For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, the squared Mahalanobis distance is defined as

$$d_M^2\left(\boldsymbol{x}, \boldsymbol{y}\right) = \left(\boldsymbol{x} - \boldsymbol{y}\right)^T \boldsymbol{M}^{-1}\left(\boldsymbol{x} - \boldsymbol{y}\right).$$

When comparing an observation $\boldsymbol{x}$ to the population mean $\boldsymbol{\mu}$, the squared Mahalanobis distance simplifies to

$$d_M^2\left(\boldsymbol{x}, \boldsymbol{\mu}\right) = \left(\boldsymbol{x} - \boldsymbol{\mu}\right)^T \boldsymbol{M}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}\right).$$

The Mahalanobis distance possesses several properties that make it particularly attractive for statistical applications. In practice, the population parameters $\boldsymbol{\mu}$ and $\boldsymbol{M}$ are unknown and can be estimated from data.

Concepts of Survival Analysis

Unlike standard regression settings, survival data are characterized by the presence of censoring, which occurs when the event of interest is not observed for some subjects during the study period [23]. Let $T \geq 0$ denote a non-negative random variable representing the survival time, the time until the occurrence of the event of interest. The SF $S(t)$ is defined as the probability that an individual survives beyond time t: $S(t) = \mathbb{P}(T > t)$. Due to censoring, the exact survival time may not be observed for a part of individuals. The most common form is right censoring, where the study ends before the event occurs or the individual is lost to follow-up. Let C denote the censoring time, and define the observed time $Y = \min(T, C)$ and the censoring indicator $\delta = \mathbb{I}\left(T \leq C\right)$, where $\delta = 1$ indicates that the event was observed and $\delta = 0$ indicates censoring.

When no covariates are present, the SF $S(t)$ is nonparametrically estimated by the Kaplan-Meier estimator [24]:

$$\hat{S}_{\mathrm{KM}}\left(t\right) = \prod_{j:t_{(j)} \leq t}\left(1 - \frac{d_j}{r_j}\right),$$

where d_j is the number of events at time $t_{(j)}$; r_j is the number of individuals at risk just before time $t_{(j)}$, i.e., the number of individuals with $Y > t_{(j)}$.

In most applications, interest lies not in the marginal survival distribution but in how survival depends on covariates $\boldsymbol{x} \in \mathbb{R}^p$. The conditional SF is defined as

$$S\left(t\middle|\boldsymbol{x}\right) = \mathbb{P}\left(T > t\middle|\boldsymbol{X} = \boldsymbol{x}\right).$$

Parametric approaches assume a specific functional form for $S(t \mid \boldsymbol{x})$, most notably the Cox proportional hazards model [6]. While the Cox model is widely used due to its interpretability and computational convenience, its proportional hazards assumption may be violated in practice. Nonparametric alternatives, such as the Beran estimator, offer greater flexibility by avoiding such parametric restrictions.

The Beran Estimator

The Beran estimator [5], also known as the conditional Kaplan-Meier estimator, provides a nonparametric estimate of the conditional SF $S(t \mid \boldsymbol{x})$ using kernel smoothing. It extends the Kaplan-Meier estimator to incorporate feature information by weighting observations according to their proximity in the feature space.

If our goal is to estimate SF $S(t \mid \boldsymbol{x}_0)$ for a feature vector $\boldsymbol{x}_0$, then observations with feature values close to $\boldsymbol{x}_0$ should receive greater weights than those far away, as they provide more relevant information about the conditional distribution. The Beran estimator allows us to extend the Kaplan-Meier estimator for accounting this idea.

Let $(\boldsymbol{x}_i, \delta_i, t_i)$, $i = 1, ..., n$, be instances from the dataset, where $t_i = \min(T_i, C_i)$ and $\delta_i = \mathbb{I}\left(T_i \le C_i\right)$. We assume that $t_1 < t_2 < ... < t_n$ is an ordered sequence of times. Define a kernel function K_τ: $\mathbb{R}^p \to \mathbb{R}_+$ with bandwidth parameter $\tau > 0$ of the kernel. The kernel assigns higher weights to observations with $\boldsymbol{x}_i$ close to a target feature vector $\boldsymbol{x}$. The Beran estimator of the conditional SF $\hat{S}_n\left(t\middle|\boldsymbol{x}\right)$ is defined as:

$$\hat{S}_n\left(t\middle|\boldsymbol{x}\right) = \prod_{t_i \le t}\left\{1 - \frac{\alpha\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right)}{1 - \sum_{j=1}^{i-1}\alpha\left(\boldsymbol{x}, \boldsymbol{x}_j, \tau\right)}\right\}^{\delta_i},$$

where the weight $\alpha(\boldsymbol{x}, \boldsymbol{x}_i, \tau)$ can be defined through kernels as

$$\alpha\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right) = \frac{K\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right)}{\sum_{j=1}^{n} K\left(\boldsymbol{x}, \boldsymbol{x}_j, \tau\right)}.$$

In particular, the "softmax" operation with a parameter τ is obtained when the Gaussian kernel is used, i.e., there holds

$$\alpha\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right) = softmax\left(-\frac{\boldsymbol{x} - \boldsymbol{x}_i^2}{\tau}\right).$$

The practical performance of the Beran estimator critically depends on the choice of bandwidth τ. Cross-validation methods are commonly employed to optimize the model parameter τ.

While the Beran estimator with product kernel is widely used, its reliance on the Euclidean metric imposes several limitations. First, it assumes that the influence of each feature is separable and that features are independent in the distance calculation. As a results, the Euclidean metric ignores correlations among features, which can lead to suboptimal weighting in the presence of multicollinearity. Second, it is sensitive to differences in scale: features with larger magnitudes dominate the kernel weights unless carefully standardized. These limitations motivate the extension to the M-Beran estimator.

The M-Beran Estimator

The methodological core of our work is the incorporation of the Mahalanobis distance into the kernel weighting scheme of the Beran estimator. It is proposed to replace the Euclidean distance in the kernel functions of the Beran estimator with the Mahalanobis distance.

Let $\boldsymbol{M}$ be a positive-definite covariance matrix and $\tau > 0$ the bandwidth parameter. The M-Beran estimator of the conditional SF $S(t \mid \boldsymbol{x})$ is:

$$\hat{S}_n\left(t \mid \boldsymbol{x}\right) = \prod_{t_i \le t} \left\{ 1 - \frac{\alpha^M\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right)}{1 - \sum_{j=1}^{i-1} \alpha^M\left(\boldsymbol{x}, \boldsymbol{x}_j, \tau\right)} \right\}^{\delta_i},$$

where $\alpha^M(\boldsymbol{x}, \boldsymbol{x}_j, \tau)$ is a weight determined by the Mahalanobis distance from the target feature vector $\boldsymbol{x}$ to each observation's feature vector and determined for the Gaussian kernel as

$$\alpha\left(\boldsymbol{x}, \boldsymbol{x}_i, \tau\right) = softmax\left(-\frac{d_M^2\left(\boldsymbol{x}, \boldsymbol{x}_i\right)}{\tau} \right).$$

Given a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we first estimate the mean vector of features $\hat{\boldsymbol{\mu}}$:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i.$$

Second, the sample covariance matrix $\boldsymbol{M}$ is estimated using the mean vector as:

$$\hat{\boldsymbol{M}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\right)\left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\right)^T.$$

In high-dimensional settings where $p > n$, $\hat{\boldsymbol{M}}$ becomes singular, requiring regularization techniques. The next step is to compute the matrix $\boldsymbol{M}^{-1}$ by inverting the matrix $\hat{\boldsymbol{M}}$. The matrix $\hat{\boldsymbol{M}}$ in high-dimensional settings where $p > n$, may be not invertible. In this case, regularization can be used as follows:

$$\boldsymbol{M}_{reg} = \boldsymbol{M} + \varepsilon \boldsymbol{I},$$

were where ε is a small positive number, and $\boldsymbol{I}$ is the identity matrix.

Then we compute the Mahalanobis distance as:

$$d_M^2\left(\boldsymbol{x}, \boldsymbol{x}_i\right) = \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\right)^T \boldsymbol{M}^{-1} \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}\right).$$

The M-Beran estimator operates as follows: at each distinct event time t_j, we compute a kernel-weighted proportion of events, where the weights are determined by the Mahalanobis distance. Observations with $\boldsymbol{x}_i$ close to $\boldsymbol{x}$ in the Mahalanobis sense receive higher weight. The product of one minus these weighted proportions over all failure times up to t yields the estimated conditional survival probability.

The Mahalanobis distance generalizes the Euclidean distance, namely, if $\boldsymbol{M}$ is a diagonal matrix $\boldsymbol{M} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$, i.e., its diagonal consists of variances of individual features $\left(\boldsymbol{M} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)\right)$ and other elements are zero, then the Mahalanobis distance is equal to the standardized Euclidean distance.

Numerical results

To compare the M-Beran estimator with the standard Beran estimator, which uses the Euclidean distance in kernels, the following real benchmark datasets were used:

- Veterans' Administration Lung Cancer Study (*Veteran*) Dataset: 137 observations, 6 features;
- *Breast Cancer* Dataset: 198 observations, 80 features;
- *Lung Cancer* Dataset: 228 observations, 9 features;
- German Breast Cancer Study Group 2 (*GBSG2*) Dataset: 686 observations, 8 features;
- Monoclonal Gammopathy (*MGUS*) Dataset: 1384 observations, 5 features;
- Diabetic *Retinopathy* Dataset: 394 observations, 9 features;
- Ovarian Cancer Survival Data (*Ovarien*): 26 observations, 4 features.
- Chronic Granulomatous Disease (*CGD*) Dataset: 203 observations, 16 features.

In order to optimize the model temperature parameters τ of the estimators, we perform a 5-fold cross-validation on the training set. The cross-validation is performed with 100 repetitions. This procedure is realized by considering all possible values of τ in a predefined grid. The models are implemented by using the R software with the main R package "survival".

Table 1 presents a comparative analysis of the Mahalanobis-based Beran estimator (M-Beran) versus the standard Beran estimator across eight real survival datasets, using the C-index as the evaluation metric. The M-Beran estimator outperforms the standard Beran estimator on six out of eight datasets, demonstrating the general effectiveness of incorporating Mahalanobis distance into kernel-based survival estimation. The M-Beran achieves its highest C-index on the datasets which likely contain meaningful correlation structures among features that the Mahalanobis distance effectively captures. We can also note that the dataset *Breast Cancer* with the worse C-index of the M-Beran estimator has the large number of features (80 features) in comparison with other datasets, which may affect the estimation of the covariance matrix. The results indicate that the Mahalanobis distance is a valuable enhancement to the Beran estimator. The negative result on the *Breast Cancer* dataset serves as an important caveat, highlighting that the method is not universally superior and that the choice of distance metric should be guided by the underlying data structure.

Table 1

Results of comparison using the C-index metric for the M-Beran estimator with the standard Beran estimator on real datasets

Dataset	Standard Beran	M-Beran	Difference
Veteran	0.706	0.713	0.007
Breast Cancer	0.517	0.463	–0.054
GBSG2	0.513	0.523	0.01
Retinopathy	0.582	0.588	0.006
MGUS	0.724	0.742	0.018
Lung Cancer	0.646	0.668	0.022
Ovarien	0.774	0.798	0.024
CGD	0.583	0.724	0.141

In order to illustrate the difference in SFs predicted by the standard Beran estimator and the M-Beran estimator for the *Veteran* dataset, we show the corresponding SFs in Fig. 1, where SFs of the standard Beran estimator are depicted by solid lines, SFs of the M-Beran estimator are depicted by dashed lines. The left graph corresponds to the patients with: Karnofsky score – 90, Cell type – Squamous, Age – 55, Treatment – Standard, Prior therapy – No. The right graph corresponds to the patients with:

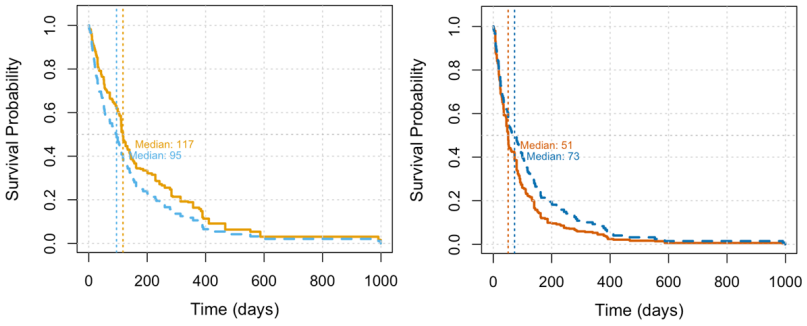


Fig. 1. SFs predicted by the standard Beran estimator and the M-Beran estimator for two patients from the Veteran dataset

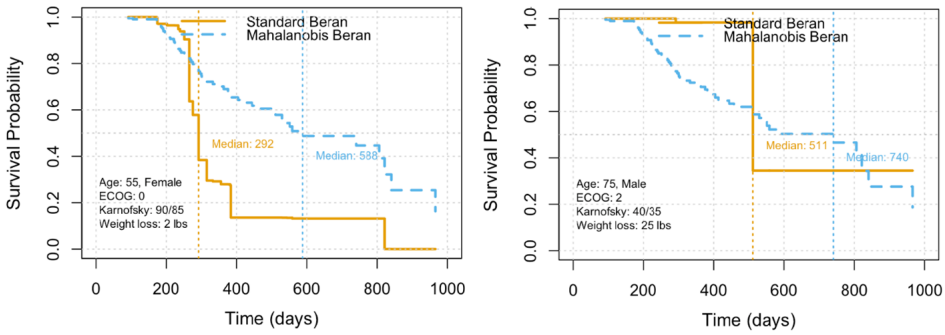


Fig. 2. SFs predicted by the standard Beran estimator and the M-Beran estimator for two patients from the Lung Cancer dataset

Karnofsky score – 40, Cell type – Adenocarcinoma, Age – 70, Treatment – Test, Prior therapy – Yes. It can be seen from Fig. 1 that the M-Beran estimator corrects SFs in accordance with the patient features.

Similar SFs of two patients from the *Lung Cancer* dataset are shown in Fig. 2. The left graph details a 55-year-old female with favorable health metrics, where the M-Beran estimator predicts a median survival of 588 days, nearly double the standard Beran model's 292 days. In this first scenario, the blue dashed line stays consistently higher than the orange line, indicating a superior predicted survival probability for the M-Beran approach. The right plot shows a 75-year-old male with poorer health indicators, where the M-Beran estimator again forecasts a longer median survival of 740 days compared to the standard model's 511 days. While the overall SF drops more rapidly for the second patient, the M-Beran estimator consistently estimates better outcomes than the standard Beran estimator in both distinct cases.

Fig. 3 illustrates the relationship between instances in the *Veteran* (the left graph) and *Lung Cancer* (the right graph) datasets by using the well-known procedure of the dimensionality reduction called t-SNE [25].

One can see from Fig. 3 that instances in the both datasets are highly correlated. This implies that the M-Beran estimator provides better results than the standard Beran estimator. The same can be seen from Table 1.

Fig. 4 shows the same relationship between instances in the *Breast Cancer* dataset. As a result, the standard Beran estimator outperforms the M-Beran estimator as it is also follows from Table 1.

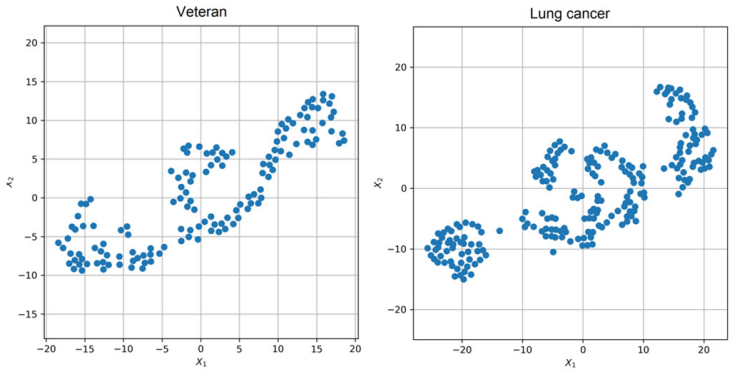


Fig. 3. T-SNE representation of the Veteran and Lung cancer datasets

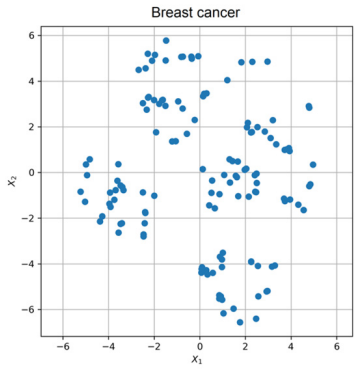


Fig. 4. T-SNE representation of the Breast Cancer dataset

It is interesting to note that the M-Beran estimator performance depends also on the kernel type. For example, let us consider the triangular kernel defined as follows:

$$K(u) = \max(0, 1 - |u|).$$

Hence, the weight α in the M-Beran estimator is

$$\alpha(\boldsymbol{x}, \boldsymbol{x}_i, \tau) = \frac{\max\left(0, 1 - d_M^2(\boldsymbol{x}, \boldsymbol{x}_i)/\tau\right)}{\sum_{j=1}^{n} \max\left(0, 1 - d_M^2(\boldsymbol{x}, \boldsymbol{x}_j)/\tau\right)}.$$

The use of this kernel significantly improves the M-Beran estimator performance for the *Veteran* dataset. The C-indices in this case for the standard Beran and M-Beran estimators are 0.582 and 0.724, respectively. At the same time, results for the *Lung Cancer* dataset (0.525 and 0.497) demonstrate deterioration in accuracy characteristics. This implies that the performance of the M-Beran estimator depends on many parameters, including the kernel type.

Discussion

Advantages

In the context of survival analysis, particularly when using nonparametric methods like the Beran estimator, the choice of distance metric is critical because it dictates how "similarity" between patients

or systems is defined. The Mahalanobis distance is important in survival analysis for the following key reasons:

1. Survival datasets often contain features measured in vastly different units (e.g., age in years, tumor size in millimeters, gene expression on a log scale, or sensor voltage). The Euclidean distance is sensitive to the scale of measurement. A variable with a large range will dominate the distance calculation over a variable with a small range, even if the latter is more predictive of survival. By incorporating the covariance matrix, it automatically standardizes the variables, ensuring that each covariate contributes appropriately to the distance metric regardless of its original unit or magnitude. This eliminates the need for arbitrary manual standardization or tuning separate bandwidths for each covariate.

2. In many survival applications (e.g., genomics, multi-sensor engineering monitoring), covariates are highly correlated. The Euclidean distance assumes orthogonality between dimensions. When two biomarkers are closely related, Euclidean distance counts the same information twice, making the similarity calculation inaccurate. The Mahalanobis distance accounts for the correlation structure of the data. It measures distance in terms of standard deviations from the mean, penalizing deviations that go against the natural correlation pattern of the data. This ensures that redundant information does not artificially inflate the similarity between observations.

3. Nonparametric survival estimators like the Beran estimator rely on kernel smoothing, where observations closer to the target vector $\boldsymbol{x}$ receive higher weights. Using Euclidean distance creates "spherical neighborhoods". In a correlated feature space, a spherical neighborhood may include observations that are statistically dissimilar while excluding those that are relevant but lie along a correlated axis. The Mahalanobis distance creates "elliptically symmetric neighborhoods" aligned with the data density. This ensures that the kernel weights are assigned to observations that are truly statistically similar, leading to a more accurate estimation of the conditional SF $S(t \mid \boldsymbol{X} = \boldsymbol{x})$.

4. Survival data is characterized by censoring, meaning information about the event time is missing for some subjects. This makes the effective sample size smaller and the estimation problem harder. Because censored data provides less information, it is crucial to maximize the utility of every uncensored neighbor. A superior distance metric ensures that the few available relevant neighbors are identified correctly. By reducing bias in neighbor selection, the Mahalanobis distance helps produce more stable survival curves even when data is sparse due to heavy censoring.

Challenges and disadvantages

While the Mahalanobis distance offers significant advantages over the Euclidean distance in handling scale and correlation, it also introduces specific challenges and disadvantages in the context of survival analysis:

1. The Mahalanobis distance requires the inversion of the covariance matrix $\boldsymbol{M}^{-1}$. In modern survival analysis, the number of features p often may exceed the number of observations n. When $p > n$, the covariance matrix is singular and cannot be inverted. Even when $p < n$, if features are nearly collinear, the matrix becomes ill-conditioned, leading to numerical instability and unreliable distance calculations. This requires regularization techniques which introduce additional tuning parameters.

2. The Mahalanobis distance relies on the sample mean and sample covariance matrix, both of which are highly sensitive to outliers. In survival data, extreme values can disproportionately skew the covariance estimate. This "masking effect" can cause the distance metric to shrink around the outliers, making actual outliers appear normal and distorting the kernel weights for the majority of observations. Robust estimators are required but are computationally more expensive.

3. The Mahalanobis distance typically uses a global covariance matrix estimated from the entire dataset. Survival relationships are often local; the correlation structure between features might differ across subgroups (e.g., the relationship between age and biomarkers might differ for early-stage versus late-stage disease). A global metric assumes the correlation structure is constant everywhere, which may

not align with the local nature of kernel-based estimators like the Beran estimator. A local covariance would be more appropriate but is difficult to estimate reliably due to data sparsity.

4. Computing the inverse covariance matrix has a computational complexity not less than $O(p^2)$, where p is the number of features. For datasets with a large number of features, this calculation becomes burdensome, especially if the distance needs to be recomputed repeatedly (e.g., during cross-validation). This is less of an issue with the Euclidean distance, which requires no matrix inversion.

5. Euclidean distance corresponds to intuitive geometric space (e.g., "5 years difference in age"). Mahalanobis distance measures distance in terms of standard deviations within a multivariate distribution. It is harder for clinicians or engineers to interpret what a specific Mahalanobis distance value means in practical terms. Explaining why two patients are "distant" requires explaining the underlying correlation structure, which can be a barrier to adoption in applied settings.

In sum, the Mahalanobis distance introduces computational burdens, instability in high dimensions, and sensitivity to outliers. In survival analysis, these limitations often require additional methodological steps, such as regularization, robust covariance estimation, or dimensionality reduction, to ensure the metric remains reliable.

REFERENCES

1. **Cox D.R.** *Analysis of survival data*. London: Chapman and Hall/CRC, 2018.

2. **Hosmer D.W., Lemeshow S., May S.** *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New Jersey: John Wiley & Sons, 2008. DOI: 10.1002/9780470258019

3. **Tattar P., Vaman H.J.** *Survival Analysis*. NY: Chapman and Hall/CRC, 2022. DOI: 10.1201/97810-03306979

4. **Wang P., Li Y., Reddy C.K.** Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys* (*CSUR*), 2019, Vol. 51, No. 6, Art. no. 110. DOI: 10.1145/3214306

5. **Beran R.** *Nonparametric regression with randomly censored survival data*, 1981.

6. **Cox D.R.** Regression models and life-tables. *Journal of the Royal Statistical Society: Series B* (*Methodological*), 1972, Vol. 34, No. 2. Pp. 187–220. DOI: 10.1111/j.2517-6161.1972.tb00899.x

7. **Berrendero J.R., Bueno-Larraz B., Cuevas A.** On Mahalanobis Distance in Functional Settings. *Journal of Machine Learning Research*, 2020, Vol. 21, Pp. 1–33.

8. **Dabrowska D.M.** Non-Parametric Regression with Censored Survival Time Data. *Scandinavian Journal of Statistics*, 1987, Vol. 14, No. 3, Pp. 181–197.

9. **Dabrowska D.M.** Uniform Consistency of the Kernel Conditional Kaplan–Meier Estimate. *The Annals of Statistics*, 1989, Vol. 17, No. 3, Pp. 1157–1167.

10. **Ma Q., Zeng L., Tu S., Xu L.** Kernel Mean Matching with Mahalanobis Distance for Causal Inference of Time-to-event Outcome. *2022 IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*), 2022, Pp. 509–514. DOI: 10.1109/BIBM55620.2022.9995021

11. **Baghfalaki T., Ganjali M., Berridge D.** Robust joint modeling of longitudinal measurements and time to event data using normal/independent distributions: a Bayesian approach. *Biometrical Journal*, 2013, Vol. 55, No. 6, Pp. 844–865. DOI: 10.1002/bimj.201200272

12. **Baghfalaki T., Ganjali M., Hashemi R.** Bayesian joint modeling of longitudinal measurements and time-to-event data using robust distributions. *Journal of Biopharmaceutical Statistics*, 2014, Vol. 24, No. 4, Pp. 834–855. DOI: 10.1080/10543406.2014.903657

13. **Liu M., Legault V., Fülöp T., Côté A.-M., Gravel D., Blanchet F.G., Leung D.L., Lee S.J., Nakazato Y., Cohen A.A.** Prediction of mortality in hemodialysis patients using moving multivariate distance. Frontiers in Physiology, 2021, Vol. 12, Art. no. 612494. DOI: 10.3389/fphys.2021.612494

14. **Roblin E., Cournède P.-H., Michiels S.** Evaluation of machine-learning models to measure individualized treatment effects from randomized clinical trial data with time-to-event outcomes. *arXiv:2506.12277*, 2025. DOI: 10.48550/arXiv.2506.12277

15. **Maas C.C.H.M., Kent D.M., Hughes M.C., Dekker R., Lingsma H.F., van Klaveren D.** Performance metrics for models designed to predict treatment effect. *BMC Medical Research Methodology*, 2023, Vol. 23, Art. no. 165. DOI: 10.1186/s12874-023-01974-w

16. **De Maesschalck R., Jouan-Rimbaud D., Massart D.L.** The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 2000, Vol. 50, No. 1, Pp. 1–18. DOI: 10.1016/S0169-7439(99)00047-7

17. **Anthony H., Kamnitsas K.** On the use of Mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging* (*UNSURE 2023*), 2023, Vol. 14291, Pp. 136–146. DOI: 10.1007/978-3-031-44336-7_14

18. **Cabana E., Lillo R.E., Laniado H**. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical Papers*, 2021, Vol. 62, Pp. 1583–1609. DOI: 10.1007/s00362-019-01148-1

19. **Colombo P., Dadalto E., Staerman G., Noiry N., Piantanida P.** Beyond Mahalanobis distance for textual OOD detection. *Advances in Neural Information Processing Systems*, 2022, Vol. 35, P. 17744–17759.

20. **Podolskiy A., Lipin D., Bout A., Artemova E., Piontkovskaya I.** Revisiting Mahalanobis distance for transformer-based out-of-domain detection. *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021, Vol. 35, No. 35, Pp. 13675–13682.

21. **Brown P.O., Chiang M.C., Guo S., Jin Y., Leung C.K., Murray E.L., Pazdor A.G.M., Cuzzocrea A.** Mahalanobis distance based k-means clustering. *Big Data Analytics and Knowledge Discovery*, 2022, Vol. 13-428, Pp. 256–262. DOI: 10.1007/978-3-031-12670-3_23

22. **Nielsen S., Abdullaev L., Teo R., Nguyen T.M.** Elliptical attention. *38th Conference in Neural Information Processing Systems* (*NeurIPS*), 2024, Pp. 1–42.

23. **Wiegrebe S., Kopper P., Sonabend R., Bischl B., Bender A.** Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 2024, Vol. 57, Art. no. 65. DOI: 10.1007/s10462-023-10681-3

24. **Kaplan E.L., Meier P.** Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 1958, Vol. 53, No. 282, P. 457–481. DOI: 10.2307/2281868

25. **Van der Maaten L., Hinton G.** Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, Vol. 9, No. 86, P. 2579–2605.

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Lev V. Utkin
Уткин Лев Владимирович
E-mail: lev.utkin@gmail.com
ORCID: https://orcid.org/0000-0002-5637-1420

Andrei V. Konstantinov
Константинов Андрей Владимирович
E-mail: andrue.konst@gmail.com
ORCID: https://orcid.org/0000-0002-1542-6480

Stanislav R. Kirpichenko
Кирпиченко Станислав Романович
E-mail: kirpichenko.sr@gmail.com
ORCID: https://orcid.org/0000-0003-2275-1473

Natalia M. Verbova
Вербова Наталья Михайловна
E-mail: nag00@mail.ru
ORCID: https://orcid.org/0000-0001-8749-9470

Submitted: 07.04.2025; Approved: 17.06.2026; Accepted: 17.06.2026.
Поступила: 07.04.2025; Одобрена: 17.06.2026; Принята: 17.06.2026.