Estimating the Probability $\mathbb{P}(X_t < Y_t)$ for $m-$ dependent Processes from Noisy Data with Mixtures of Normal Errors

Thai Phuc Hung[1,*], Nguyen Cao Phong[2]

[1] Faculty of Basics, Soc Trang Community College;
[2] Student Political Affairs Department, Mien Tay Construction University;
*Corresponding author: tphung@stcc.edu.vn

■ Received: 14/11/2024 ■ Revised: 12/12/2024 ■ Accepted: 08/01/2025

ABSTRACT

This study proposes a nonparametric estimation method for computing the probability $\theta = \mathbb{P}(X_t < Y_t)$, where X_t and Y_t are two $m-$dependent stationary processes subject to noise generated from a mixture of two normal distributions. Observations are made at discrete time points $t_j = j\Delta$, with Δ being a positive constant. This method has high practical significance in fields that require handling time-dependent random processes, such as reliability assessment for systems involved in the relationship between applied pressure and capacity (stress-strength model), or in analyzing Receiver Operating Characteristic (ROC) curves. We explore how to address complex noise structures and present results on the convergence rate of the Mean Squared Error (MSE) as well as the asymptotic normality of the estimator. The effectiveness of the method is demonstrated through detailed simulations and an application analysis using real data from Duchenne Muscular Dystrophy (DMD), highlighting that simple noise assumptions based on normal distributions may not be sufficient to accurately capture the complexity observed in real-world scenarios.

Keywords: *Asymptotic normality; Deconvolution; $m-$ dependent; Mean squared error; Mixture distribution; Stationary process.*

1. INTRODUCTION

In statistics, estimating probabilities for random variables influenced by noise is a significant and practical topic, particularly in fields such as economics, medicine, and data science. One commonly encountered model to describe noise is the mixture of two normal distributions, known as the mixture model. This approach allows for the representation of heterogeneity within the data, with different components potentially reflecting different underlying processes within the same dataset.

Consider paired random variables (X_{t_j}, Y_{t_j}) for $j = 1,...,n$ generated from $m-$ dependent stationary processes X_t and Y_t. We are interested in estimating the probability of a specific event:

$$\theta = \mathbb{P}(X_t < Y_t) \tag{1}$$

In this scenario, the variables (X_{t_j}, Y_{t_j}) are observed under the influence of noise as (X'_{t_j}, Y'_{t_j}), where

$$X'_{t_j} = X_{t_j} + \zeta_{t_j},\ Y'_{t_j} = Y_{t_j} + \eta_{t_j}. \tag{2}$$

Here, the random variables ζ_{t_j} and η_{t_j} are modeled as noise, generated from $m-$ dependent stationary processes, each drawn from a mixture of two normal distributions. Specifically, we assume:

$$\begin{aligned} \zeta_{t_j} &\sim \lambda_1 N(\mu_1, \sigma_1) + (1-\lambda_1) N(\nu_1, \delta_1), \\ \eta_{t_j} &\sim \lambda_2 N(\mu_2, \sigma_2) + (1-\lambda_2) N(\nu_2, \delta_2) \end{aligned} \tag{3}$$

with $0 \le \lambda_1, \lambda_2 \le 1$ being the mixing probabilities for each distribution in the mixture. This model allows us to account for the inherent variability and complexity in the noise, providing deeper insights into the nature of the data.

For simplicity, we assume that the observations are made at discrete time points $t_j = j\Delta$ (Δ is a positive constant), and we use j to represent t_j. Then, equation (2) can be rewritten as follows

$$X'_j = X_j + \zeta_j,\ Y'_j = Y_j + \eta_j.$$

The model (1) has key applications in reliability theory and medical diagnostics. In reliability, X and Y represent strength and stress, with θ indicating component reliability (see Kotz, Lumelskii and Pensky [1]). In medicine, θ corresponds to the area under the ROC curve, assessing diagnostic test accuracy (see Zhou [2]). Estimation of θ has garnered attention, particularly with independent and identically distributed (iid) assumptions for X,Y and errors. Extensive studies exist for the error-free case $(\zeta = \eta = 0)$ (see Woodward and Kelley [3], Reiser and Guttman [4], Kundu and Gupta [5], Zhou [2], Motoya and Rubio [6]) and there is also substantial research addressing cases with noise, including work by Reiser [7], Dattner [8], Trong, Nguyen and Phuong [9], Phuong and Thuy [10], Trong and Hung [11].

In practical applications, iid assumptions are unrealistic for time-dependent random sequences like X_t and Y_t, which are often generated by stochastic processes. Jose and Drisya estimated stress-strength reliability over time (see Jose and Drisya [12,13]), and Kamarudin et al. examined time-varying ROC curves (see Kamarudin, Cox and Kolamunnage-Dona [14]). However, these studies assume error-free data and lack rigorous proofs.

To address limitations within the context of $\alpha - mixing$ stationary processes, Trong and Hung [11] studied the estimation of θ in model (1) where X_t, Y_t, ζ_t and η_t are $\alpha - mixing$ stationary processes, with errors in the super-smooth density class. Building on this work, the present paper focuses on the estimation of θ in model (1) where X_t and Y_t are $m-$dependent stationary processes, and the errors are a mixture of two normal distributions. Specifically, we assume conditions

$$f_\varsigma^{ft}(p) = \lambda_1\varphi(\mu_1,\sigma_1) + (1-\lambda_1)\varphi(\nu_1,\delta_1),$$

$$f_\eta^{ft}(p) = \lambda_2\varphi(\mu_2,\sigma_2) + (1-\lambda_2)\varphi(\nu_2,\delta_2),$$

where $\varphi(\mu,\sigma) = \exp\left(i\mu p - \sigma^2 p^2/2\right)$ and $f_\varsigma^{ft}, f_\eta^{ft}$ are the Fourier transforms of ζ,η. This study is important as, in some practical scenarios, noise is generated from a mixture of complex distributions rather than a single distribution, as illustrated by the DMD data.

The structure of the paper is as follows: Section 2 introduces the proposed estimator, Section 3 presents the key results on MSE convergence rate and the asymptotic normality of the estimator, Section 4 is devoted to simulation studies, Section 5 explores empirical applications with a particular focus on the analysis of DMD data, and the paper concludes with a summary of the key findings in Section 6.

2. THE PROPOSED ESTIMATOR

For $\varrho \geq 1$, let L_ϱ be the set of Lebesgue measurable functions f that satisfy the condition $\| f \|_\varrho = \left(\int_{-\infty}^{+\infty} |f(x)|^\varrho dx\right)^{1/\varrho} < \infty$. For $f \in L_1$, define the Fourier transform of f as $f^{ft}(p) := \int_{-\infty}^{+\infty} e^{ipx} f(x) dx$. In this paper, we focus on $m-$ dependent processes. To formally define an $m-$dependent process, we conceptualize it as a dependence structure where the distance between random variables is the key factor. Specifically, a (discrete-time) stationary process $\{X_j\}_{j\in\mathbb{Z}}$ is considered an $m-$dependent stationary process if two subsets of random variables, $\{\ldots, X_{k-1}, X_k\}$ and $\{X_h, X_{h+1}, \ldots\}$, are independent whenever $h-k > m$. To retain as much information as possible from the data, it is essential to appropriately aggregate the relevant terms. A powerful method to handle the dependence structure and improve computational efficiency involves pairing the data from X_t and Y_t.

We subsequently introduce the following estimator for the invariant density function $f_Z(x)$ of the process $Z_t = X_t - Y_t$:

$$f_n(x) = \frac{1}{2\pi n}\sum_{j=1}^{n}\psi(x) \quad \forall x \in \mathbb{R},$$

where

$$\psi(x) = \int_{-\infty}^{\infty} \exp\left[ip\left(X_j' - Y_j' - x\right)\right]\frac{K^{ft}(pb_n)}{f_\varsigma^{ft}(p)\overline{f_\eta^{ft}(p)}}dp,$$

i is the imaginary unit $\left(i^2 = -1\right)$, b_n is the bandwidth satisfying $b_n > 0 \quad \forall n \in \mathbb{N}$, K denotes a known kernel function and K^{ft} represents the Fourier transform of K.

Building on this estimator, we present an estimator for the parameter θ as follows:

$$\theta_n = 1 - \frac{1}{2\pi n}\sum_{j=1}^{n}\int_0^{\infty}\psi(x)dx \quad \forall x \in \mathbb{R}. \qquad (4)$$

3. MAIN RESULTS

3.1. Some assumptions

(i) *The kernel function* $K \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$, $\int_{-\infty}^{+\infty} K(u)du = 1$, $\int_{-\infty}^{+\infty} uK(u)du = 0$, $\int_{-\infty}^{+\infty} u^2K(u)du < \infty$, $xK(x) \to 0$ as $x \to \infty$ and the Fourier transform K^{ft} is symmetric function, supported on $[-1;1]$.

(ii) Let $0 \le \lambda_1, \lambda_2 \le 1$; $\mu_1, \nu_1, \mu_2, \nu_2 \in \mathbb{R}$; $\sigma_1, \delta_1, \sigma_2, \delta_2 > 0$. For $p \in \mathbb{R}$, the mixtures of normal noise densities f_ς and f_η have Fourier transforms satisfying:

$$f_\varsigma^{ft}(p) = \lambda_1\varphi(\mu_1, \sigma_1) + (1-\lambda_1)\varphi(\nu_1, \delta_1),$$

$$f_\eta^{ft}(p) = \lambda_2\varphi(\mu_2, \sigma_2) + (1-\lambda_2)\varphi(\nu_2, \delta_2),$$

where $\varphi(\mu, \sigma) = \exp\left(i\mu p - \sigma^2 p^2/2\right)$.

(iii) Let $m_X, m_Y, m_\varsigma, m_\eta$ be positive integer constants. The processes X_t, Y_t, ζ_t and η_t are $m_X-, m_Y-, m_\varsigma-$ and $m_\eta-$dependent stationary processes , respectively.

(iv) Let $Z_j' = X_j' - Y_j'$. The invariant density function $f_{Z'}(x)$ of the process $\{Z_j'\}$ is bounded and $\int_0^{\infty}\left[f_{Z'}(x)\right]^{1/2}dx < \infty$. Furthermore, the 2-dimensional probability density function $f_{Z_j', Z_k'}(u,v)$ exists and is bounded for all $1 \le j, k \le n$.

3.2. Theorem 1 (MSE convergence)

Let $\rho_{b_n} \in \left(0, 1\Big/\left(\min\left\{\sigma_1^2, \delta_1^2\right\} + \min\left\{\sigma_2^2, \delta_2^2\right\}\right)\right)$ and let the assumptions $(i)-(iv)$ hold. If $f_Z(x)$ is twice differentiable and $f_Z''(x)$ is continuous and bounded, by choosing the bandwidth $b_n = \rho_{b_n}^{-1/2}(\ln n)^{-1/2}$,we have

$$\mathbb{E}\left(\theta_n - \theta\right)^2 \le O\left((\ln n)^{-2}\right)$$

when n is large enough.

Remark 1. Theorem 1 demonstrates that the convergence rate of the MSE for the estimator θ_n izn $m-$dependent stationary processes is equivalent to the convergence rate of the MSE for the estimator $f_n(x)$ in Theorem 3.2 (with $=2$) of Hung and Phong [15] in the $\alpha-mixing$ setting. This rate is also comparable to the one that Fan [16] proved to be optimal for iid observations. This similarity highlights the efficiency of the estimator θ_n and affirms that it can achieve a rapid convergence rate similar to that in the iid setting.

Proof of Theorem 1. Let

$$m = \max\left\{m_X, m_Y, m_\varsigma, m_\eta\right\}.$$

Then,

$$\alpha_{X_t}(k) = \alpha_{Y_t}(k) = \alpha_{\zeta_t}(k) = \alpha_{\eta_t}(k) = 0$$

when $k > m$. Where

$$\alpha_{\nabla_t}(k) = \sup_{j\in\mathbb{N}}\sup_{A\in\mathcal{F}_1^j, B\in\mathcal{F}_{j+k}^{\infty}}\left|\mathbb{P}[AB] - \mathbb{P}[A]\mathbb{P}[B]\right|$$

and F_l^h denote the $\sigma-algebra$ of events generated by the random variables $\{\nabla_j, 1 \le l \le j \le h\}$. Therefore, for $\varepsilon > 0$, by selecting the appropriate coefficients, we obtain

$$\alpha_{X_t}(k), \alpha_{Y_t}(k), \alpha_{\zeta_t}(k), \alpha_{\eta_t}(k) \le O\left(1/k^{2+\varepsilon}\right)$$

Now, let

$$A_1 = \min\{\lambda_1, 1-\lambda_1\}, B_1 = 1, \beta = \min\left\{\frac{\sigma_1^2}{2}, \frac{\delta_1^2}{2}\right\},$$

$$A_2 = \min\{\lambda_2, 1-\lambda_2\}, B_2 = 1, \gamma = \min\left\{\frac{\sigma_2^2}{2}, \frac{\delta_2^2}{2}\right\}.$$

We have

$$A_1 \exp\left(-\beta |p|^2\right) \le \left|f_\zeta^{ft}(p)\right| \le B_1 \exp\left(-\beta |p|^2\right),$$

$$A_2 \exp\left(-\gamma |p|^2\right) \le \left|f_\eta^{ft}(p)\right| \le B_2 \exp\left(-\gamma |p|^2\right).$$

Using Theorem 3.4 of Trong and Hung [11], by choosing the bandwidth $b_n = \rho_{b_n}^{-1/2} (\ln n)^{-1/2}$, we conclude

$$\mathbb{E}\left(\theta_n - \theta\right)^2 \le O\left((\ln n)^{-2}\right)$$

when n is large enough.

3.3. Theorem 2 (Asymptotic normality)

Let $\omega_{b_n} = \exp\left[-\left(\min\{\sigma_1^2, \delta_1^2\} + \min\{\sigma_2^2, \delta_2^2\}\right) b_n^{-2}\right]$ and let the assumptions $(i)-(iv)$ hold. If the bandwidth b_n is choosen so that $nb_n\omega_{b_n}^2 \to \infty$ and $s_n := \left\lfloor \left(nb_n^\tau\right)^{1/2} \right\rfloor \to \infty$ as $n \to \infty$ for some $\tau > 1$, we have

$$\limsup\left\{n^{1/2} b_n^{1/2} \omega_{b_n} \left[\theta_n - \mathbb{E}\left(\theta_n\right)\right]\right\} \xrightarrow{D} N\left(0; \sigma^2\right).$$

Where $\sigma^2 = \limsup\left[nb_n\omega_{b_n}^2 Var\left(\theta_n\right)\right] \le C,$

and $C = \left[1 \Big/ \left(2\pi \min\{\lambda_1, \lambda_2, (1-\lambda_1), (1-\lambda_2)\}\right)\right]$

$$\left\{\int_0^\infty \left[f_{Z'}(x)\right]^{1/2} dx\right\}^2 \int_{-\infty}^{\infty} \left|K^{ft}(p)\right|^2 dp.$$

Remark 2. If we choose the bandwidth $b_n = \rho_{b_n}^{-1/2} (\ln n)^{-1/2}$ for a particular choice of $\rho_{b_n} \in \left(0, 1 \Big/ \left(\min\{\sigma_1^2, \delta_1^2\} + \min\{\sigma_2^2, \delta_2^2\}\right)\right)$ as stated in Theorem 1, then the condition $s_n := \left\lfloor \left(nb_n^\tau\right)^{1/2} \right\rfloor \to \infty$ as $n \to \infty$ for some $\tau > 1$ is automatically satisfied.

Proof of Theorem 2. The proof of this Theorem is lengthy and quite complex, and it is not feasible to present it in full within the scope of this paper. However, the result of the Theorem can be obtained by making some modifications to the proof of Theorem 3.5 of Trong and Hung [11].

4. SIMULATION STUDY

The $m-$dependent processes play a vital role in modeling time-based dependencies across diverse real-world scenarios. In finance, they aid in simulating complex interactions within stock markets, interest rates, and macroeconomic trends. In meteorology, they effectively capture weather dynamics, considering the interrelations between temperature, atmospheric pressure, and rainfall patterns over time. In data science, $m-$dependent processes are invaluable for analyzing time series data, enabling the modeling of sequential dependencies that frequently arise in practical datasets.

4.1. Simulation with Gamma Distribution

We now consider $\{X_j\}_{j\in\mathbb{Z}}$ and $\{Y_j\}_{j\in\mathbb{Z}}$ as $30-$ dependent and $20-$ dependent stationary processes, where their invariant distribution follow Gamma distributions with shape parameters α_{X_t} and α_{Y_t} detailed in Tables 1 and 2, and rate parameters $\beta_{X_t} = \beta_{Y_t} = 1$. For simplicity, the noises ζ_j and η_j are iid random variables generated from a mixture of two normal distributions as described in (3), with $\lambda_2 = 0.5, \sigma_2^2 = 0.5$, $\mu_1 = \nu_1 = \mu_2 = \nu_2 = 0, \delta_1^2 = \delta_2^2 = 1$. The values of λ_1 and σ_1^2 are given in Tables 1 and 2.

To simulate data for $\{X_j\}_{j\in\mathbb{Z}}$, we begin by generating 30 independent data points, each sampled from a Gamma distribution with shape parameter α_{X_t} and rate parameter β_{X_t}.The 31st data point is then chosen either from the first 30 data points with a probability of 0.5, or as a new independent sample from the Gamma distribution with the same parameters, also with a probability of 0.5. For data points indexed by $(30+l)$, with $l \ge 2$, each point is selected from one of the previous 30 data points with a probability of 0.5. If the $(30+l-1)$th data point matches the $(l-1)$ th point, it is discarded. Alternatively, it can also be independently drawn from a Gamma distribution with the same parameters with a

probability of 0.5. The same approach is used to simulate data for $\{Y_k\}_{k\in\mathbb{Z}}$. We also generate iid data for noises ζ and η then construct the dataset $\{X'_j = X_j + \zeta_j, Y'_j = Y_j + \eta_j\}_{j=\overline{1,n}}$. Finally, we employ the proposed estimator given in (4) to estimate θ. The authors choose the bandwidth $b_n = \left[1/(1.1\upsilon)\right]^{-1/2}(\ln n)^{-1/2}$, with $\upsilon = \left(\min\{\sigma_1^2, \delta_1^2\} + \min\{\sigma_2^2, \delta_2^2\}\right)$, according to Theorem 1 and the kernel function

$$K(x) = \frac{48\cos x}{\pi x^4}\left(1 - 15/x^2\right) - \frac{144\sin x}{\pi x^5}\left(2 - 5/x^2\right),$$

with $K^{ft}(p) = \left(1-p^2\right)^3 I_{[-1,1]}(p)$, where

$$I_{[-1,1]}(p) = \begin{cases} 1 & \text{when } p \in [-1,1] \\ 0 & \text{when } p \notin [-1,1] \end{cases}.$$

These simulations include three different sample sizes: $n = 100$, $n = 200$, and $n = 500$, enabling an analysis of the estimation's sensitivity to sample size. Each simulation consists of 100 replications of observations for the processes $\{X_j\}_{j\in\mathbb{Z}}$ and $\{Y_k\}_{k\in\mathbb{Z}}$. The empirical Mean and MSE (multiplied by 100 for easier interpretation) are presented in Tables 1 and 2.

Table 1. Empirical Mean and MSE (x100) for Estimator (4) at $\theta = \mathbb{P}(X_j < Y_j) = 0.5$ with Sample Sizes $n \in \{100, 200, 500\}$ for Gamma Invariant Distributions

$\alpha_{X_t} = 5, \alpha_{Y_t} = 5$			$\theta = 0.5$		
λ_1	σ_1^2	n =	100	200	500
0.5	0.25	Mean	0.496	0.503	0.501
		MSE	0.474	0.281	0.105
0.5	0.5	Mean	0.495	0.497	0.498
		MSE	0.358	0.243	0.087
0.75	0.25	Mean	0.495	0.498	0.502
		MSE	0.412	0.298	0.096
0.75	0.5	Mean	0.493	0.505	0.499
		MSE	0.365	0.224	0.102

Table 2. Empirical Mean and MSE (x100) for Estimator (4) at $\theta = \mathbb{P}(X_j < Y_j) = 0.637$ with Sample Sizes $n \in \{100, 200, 500\}$ for Gamma Invariant Distributions

$\alpha_{X_t} = 4, \alpha_{Y_t} = 5$			$\theta = 0.637$		
λ_1	σ_1^2	n =	100	200	500
0.5	0.25	Mean	0.628	0.631	0.633
		MSE	0.489	0.218	0.113
0.5	0.5	Mean	0.624	0.632	0.63
		MSE	0.354	0.228	0.099
0.75	0.25	Mean	0.626	0.634	0.629
		MSE	0.472	0.254	0.093
0.75	0.5	Mean	0.627	0.621	0.636
		MSE	0.405	0.237	0.097

4.2. Simulation with Normal Distribution

We now examine $\{X_j\}_{j\in\mathbb{Z}}$ and $\{Y_j\}_{j\in\mathbb{Z}}$ as stationary processes that are $15-$dependent and $10-$dependent, respectively. Their invariant distributions are normal, with means μ_{X_t} and μ_{Y_t} outlined in Tables 3 and 4, and variances $\sigma^2_{X_t} = \sigma^2_{Y_t} = 2$. The noise terms ζ_j and η_j are iid random variables generated from a mixture of two normal distributions as described in (3), with parameters $\lambda_1 = 0.5, \sigma_1^2 = 0.5, \mu_1 = 0.1, \quad \nu_1 = -0.1$, and $\mu_2 = \nu_2 = 0, \delta_1^2 = \delta_2^2 = 1$. The values for λ_2 and σ_2^2 can be found in Tables 3 and 4.

First, we generate iid normal random variables $\{U_j\}_{j\in\mathbb{Z}}$, $\{V_j\}_{j\in\mathbb{Z}}$. Subsequently, $15-$ and $10-$dependent stationary processes $\{X_j\}_{j\in\mathbb{Z}}$ and $\{Y_j\}_{j\in\mathbb{Z}}$ are constructed with $X_j = (1/15)(U_j + \ldots + U_{j+14}), j = 1,\ldots,n$ and $X_j = (10)(V_j + \ldots + V_{j+9}), j = 1,\ldots,n.$ Using the same bandwidth b_n and kernel function detailed in Subsection 4.1, simulations are performed for three sample sizes: $n = 100$, $n = 200$, and $n = 500$ to assess the effect of sample size on estimation sensitivity. Each simulation is repeated 100 times, and the results are summarized in Tables 3 and 4.

Table 3. Empirical Mean and MSE (x100) for Estimator (4) at $\theta = \mathbb{P}\left(X_j < Y_j\right) = 0.637$ with Sample Sizes $n \in \{100, 200, 500\}$ for Normal Invariant Distributions

$\mu_{X_t} = 5, \mu_{Y_t} = 2$			$\theta = 0.067$		
λ_1	σ_1^2	n =	100	200	500
0.5	0.25	Mean	0.069	0.066	0.068
		MSE	0.549	0.251	0.094
0.5	0.5	Mean	0.068	0.065	0.066
		MSE	0.556	0.313	0.103
1	0.25	Mean	0.07	0.067	0.064
		MSE	0.538	0.229	0.101
1	0.5	Mean	0.072	0.064	0.07
		MSE	0.466	0.268	0.098

Table 4. Empirical Mean and MSE(x100) for Estimator (4) at $\theta = \mathbb{P}\left(X_j < Y_j\right) = 0.5$ with Sample Sizes $n \in \{100, 200, 500\}$ for Normal Invariant Distributions

$\mu_{X_t} = 5, \mu_{Y_t} = 5$			$\theta = 0.5$		
λ_1	σ_1^2	n =	100	200	500
0.5	0.25	Mean	0.508	0.503	0.498
		MSE	1.276	0.721	0.329
0.5	0.5	Mean	0.505	0.502	0.499
		MSE	1.489	0.711	0.381
1	0.25	Mean	0.494	0.509	0.495
		MSE	1.412	0.696	0.344
1	0.5	Mean	0.492	0.503	0.501
		MSE	1.342	0.745	0.317

5. DMD DATA

The dataset is from a study on Duchenne muscular dystrophy (DMD), a severe genetic disorder affecting children and typically fatal by the early 20s (see Reiser [7]). Screening for female carriers is crucial due to the lack of effective treatment. The dataset includes blood samples from 38 DMD carriers and 87 healthy individuals, focusing on serum creatine kinase. Subjects provided between 1 and 7 samples each, resulting in an unbalanced dataset. Upon examining the data, significant skewness in the marker values was observed, prompting the application of a log transformation to improve normality.

Let X and Y represent independent random variables for the marker distribution in diseased and healthy populations, with random samples $X_1,...,X_n$ and $Y_1,...,Y_m$. Reiser notes that these variables are often unobservable and measured with additive normal measurement error. Let X_{j,l_j} and Y_{k,s_k} denote the l_j th replicate for the j th diseased subject and the s_k th replicate for the k th healthy subject, respectively. Thus

$$X_{j,l_j} = X_j + \zeta_{j,l_j}, \quad j = 1,...,n; l_j = 1,...,p_j,$$

$$Y_{k,s_k} = Y_k + \eta_{k,s_k}, \quad k = 1,...,m; s_k = 1,...,q_k,$$

with $\zeta_{j,l_j} \sim N(0,\sigma_1^2), \eta_{k,s_k} \sim N(0,\sigma_2^2)$. In this context, the sequences of random variables $X_{1,1},...,X_{1,p_1},...,X_{n,1},...,X_{n,p_n}$ and $Y_{1,1},...,Y_{1,q_1}$, $...,Y_{m,1},...,Y_{m,q_m}$ represent $p-$ and $q-$ dependent processes, respectively, where $p = \max\{p_1,...,p_n\}$ and $q = \max\{q_1,...,q_m\}$. Under the premise that $X \sim N\left(\mu_X, \sigma_X^2\right)$ and $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$, Reiser employed maximum likelihood estimation (MLE). The estimates, computed using the SAS MIXED procedure, are shown in Table 5.

Table 5. MLE for the DMD data

Parameters	μ_X	μ_Y	σ_X^2	σ_Y^2
Estimates	4.73	3.58	0.66	0.07

However, in practice, the noise may not simply follow a normal distribution, but rather a more complex one, such as a mixture of normal distributions. Using Gaussian Mixture Models (GMM) and the mixtools package in R, we analyzed the data and found that the noise in the measurements from the diseased group (ζ) follows a mixture of two normal distributions with the estimated parameters, as shown in Table 6.

Table 6. Parameters of Noise Variable (ζ) for Diseased Group (GMM)

$\lambda =$	0.5136433	$1-\lambda =$	0.4863567
$\mu_{1=}$	0.009519739	$v_1=$	-0.01005388
$\sigma_1^{2=}$	0.005514073	δ_1^2	0.1829592

We now consider the noise η for the healthy group, which is assumed to follow a normal distribution. The parameters for η are computed using the standard formula $\sigma_\eta^2 = \left[\sum_k (q_k - 1)\right]^{-1} \sum_k \sum_{s_k} \left(Y_{k,s_k} - \overline{Y}_k\right)^2$, resulting in $\mu_\eta = 0$ and $\sigma_\eta^2 = 0.016063469$.

We apply estimator (4) to estimate $\theta = \mathbb{P}\left(X < Y\right)$ using paired data. To balance the diseased group $\left\{X_{j,l_j}\right\}$ and the healthy group $\left\{Y_{k,s_k}\right\}$, we replicate data for individuals in the patient group with fewer than 5 measurements, ensuring the process remains $5-$dependent. Note that estimator (4) does not assume normality or independence among the data, making it more practical for real-world applications. The estimated results for θ, using maximum likelihood estimation and estimator (4) with both mixture normal noise and simple normal noise (as in Trong and Hung [11]), are presented in Table 7. For the estimator (4) under mixture normal noise, we continue to use the same bandwidth and kernel function as detailed in Subsection 4.1.

Table 7. Estimator (θ_n) for the Probability $\theta = \mathbb{P}\left(X < Y\right)$ Based on the DMD Data

Method	MLE (Reiser)	Estimator (4)	
		Simple Noise	Mixture Noise
(θ_n)	0.089156	0.088086	0.089666

6. CONCLUSION

In this study, we proposed a nonparametric estimator for estimating the probability $\theta = \mathbb{P}\left(X < Y\right)$, where X_t and Y_t are two $m-$dependent stationary processes subject to complex noise structures. The noise in our model follows a mixture of two normal distributions, which distinguishes this work from previous studies using simpler noise models.

Key advantages of this study include:

- Stable and efficient estimation: The results from our simulations demonstrate that the proposed estimator performs efficiently and stably, with convergence improving as the sample size increases. Specifically, as we tested sample sizes of 100, 200, and 500, the MSE values decreased with increasing sample size, confirming the estimator's accuracy and effectiveness in various scenarios.

- Handling complex noise structures: Unlike traditional methods assuming simple normal noise, our estimator offers the flexibility to handle more intricate noise structures, such as the mixture of normal distributions. This is particularly useful for real-world applications where noise may not follow a simple normal distribution, making the method highly adaptable.

- Application to real-world data: We applied the proposed estimator to real-world DMD data, where we compared our method with Reiser's MLE technique. The results indicated that our estimator performs similarly to MLE in the case of simple normal noise and shows promising results when dealing with the more complex noise structure, such as the mixture of two normal distributions.

Overall, the proposed estimator provides a robust and reliable method for estimating probabilities in the presence of dependent data and complex noise. The flexibility and adaptability of the method make it suitable for a wide range of real-world applications, where noise structures may be more complex than the simple normal distribution. Future research will explore further applications, including handling even more intricate noise models and extending the method to multidimensional processes.

REFERENCES

[1] S. Kotz, Y. Lumelskii and M. Pensky, The stress-strength model and its generalizations theory and applications. Singapore: World Scientific; 2003.

[2] W. Zhou, "Statistical inference for $\mathbb{P}(X < Y)$". Stat Med, 27:257-279, 2008. Doi: 10.1002/sim.v27:2

[3] W.A. Woodward and G.D. Kelley, "Minimum variance unbiased estimation of $\mathbb{P}(X < Y)$ in the normal case". Technometrics, 19(1):95-98, 1977. Doi: 10.1080/00401706. 1977.10489505

[4] B. Reiser and I. Guttman, "Statistical inference for P($X < Y$): the normal case". Technometrics, 28:253-257, 1986. Doi: 10.2307/1269081

[5] D. Kundu and R.D. Gupta, "Estimation of $\mathbb{P}(X < Y)$ for Weibull distribution". IEEE Trans Reliab, 55:270-280, 2006 Doi: 10.1109/ TR. 2006.874918

[6] J.A. Motoya and F.J. Rubio, "Nonparametric inference for $\mathbb{P}(X < Y)$ with paired variables". J Data Sci, 12:359-375, 2014. Doi: 10.6339/JDS. 201404_12(2).0009

[7] B. Reiser, "Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of roc curves". Stat Med, 19(16):2115-29, 2000. Doi: 10.1002/(ISSN)1097-0258

[8] I. Dattner, "Deconvolution of $\mathbb{P}(X < Y)$ with supersmooth error distributions". Stat Probab Lett, 83:1880-1887, 2013. Doi: 10.1016/j.spl. 2013.04.024

[9] D.D. Trong, T.T.Q. Nguyen and C.X. Phuong, "Deconvolution of $\mathbb{P}(X < Y)$ with compactly supported error densities". Stat Probab Lett, 123:171-176, 2016. Doi: 10.1016/j. spl.2016.12.014

[10] C.X. Phuong and L.T.H. Thuy, "Deconvolution of $\mathbb{P}(X < Y)$ with unknown error distributions". Commun Stat Theory Methods, 51(17):5889-5912, 2022. Doi: 10.1080/ 03610926.2020.1849722

[11] D.D. Trong and T.P. Hung, "Deconvolution of $\mathbb{P}(X < Y)$ for stationary processes with supersmooth error distributions". Statistics, 58(6):1463-1487, 2024. Doi: 10.1080/02331888. 2024.2407913

[12] J.K. Jose and M. Drisya, "Time-dependent stress-strength reliability models based on phase type distribution". Comput Stat, 35:1345-1371, 2020. Doi: 10.1007/s00180-020-00991-3

[13] J.K. Jose and M. Drisya, "Stress-strength reliability estimation of time-dependent models with fixed stress and phase type strength distribution". Rev Colombiana De Estad Theor Stat, 44(1):201-223, 2021. Doi: 10.15446/rce. v44n1.86519

[14] A.N. Kamarudin, T. Cox and R. Kolamunnage-Dona, "Time-dependent ROC curve analysis in medical research: current methods and applications". BMC Med Res Methodol, 17(1):1-19, 2017. Doi: 10.1186/ s12874-017-0332-6

[15] T.P. Hung and N.C. Phong, "Deconvolution Estimators for Invariant Densities of Stationary Processes: Method and Simulation". Journal of Science, Technology and Engineering, Mien Tay Construction University, 9:26-32, 2024.

[16] J. Fan, "On the optimal rates of convergence for nonparametric deconvolution problems", The Annals of Statistics, 19(3):1257-1272, 1991. Doi: 10.1214/aos/1176348248