Novel Statistical Techniques for Complex Data Structures

Li, S.

DOI
10.4233/uuid:9b1a2648-4408-4e1e-b172-f842eea1dcd1

Publication date
2025

Document Version
Final published version

Citation (APA)
Li, S. (2025). *Novel Statistical Techniques for Complex Data Structures*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:9b1a2648-4408-4e1e-b172-f842eea1dcd1

Novel Statistical Techniques for Complex Data Structures

Shixiang Li

NOVEL STATISTICAL TECHNIQUES FOR COMPLEX DATA STRUCTURES

NOVEL STATISTICAL TECHNIQUES FOR COMPLEX DATA STRUCTURES

Dissertation

for the purpose of obtaining the degree of doctor

at Delft University of Technology

by the authority of the Rector Magnificus,

prof. dr. ir. T.H.J.J. van der Hagen,

chair of the Board for Doctorates

to be defended publicly on

Monday, 22 September 2025 at 17:30 o'clock

by

Shixiang LI

Bachelor of Science in Statistics,
Beijing Institute of Technology, China
born in Hebei, China

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. G. Jongbloed,	Delft University of Technology, *promotor*
Prof. dr. Y.B. Tian,	Beijing Institute of Technology, *promotor*
Dr. P. Chen,	TU Delft / Zhejiang University, China, *copromotor*
Independent members:	
Prof. dr. ir. H.X. Lin,	Delft University of Technology / Leiden University
Prof. dr. A.J. Cabo,	Delft University of Technology
Dr. A.F.F. Derumigny,	Delft University of Technology
Dr. S. Kapodistria,	Eindhoven University of Technology
Prof.dr. H.M. Schuttelaars,	Delft University of Technology, *reserve member*

The doctoral research has been carried out in the context of an agreement on joint doctoral supervision between the Beijing Institute of Technology, China and the Delft University of Technology, the Netherlands.

Keywords: Complex data, Gaussian process, Efficient global optimization, Spatiotemporal model, Process capability index, Multivariate Poisson distribution.

To my parents and my brother

CONTENTS

SUMMARY

The main concern of this thesis is applying statistical methods to tackle some issues that arise from specific cases where data are hard to obtain, spatiotemporal, or non-normal. The range of such potential applications is huge. These hard-to-solve problems often necessitate innovative strategies, prompting the development of novel statistical models and methods. While the subjects covered in this thesis are diverse, they consistently revolve around the modeling methods in Statistics for complicated data structures, where traditional methods fail to provide viable solutions. Among the topics presented, we have the optimization for a computationally expensive black-box function based on limited data; predicting a spatiotemporal flow at a new input using the data stored as a tensor; the evaluation of a manufacturing process considering non-normal data and asymmetric tolerances; and, in the presence of dependence structures, defining the joint distribution of multivariate Poisson counts.

In this thesis, we separate all topics into two kinds of issues and propose several tailored statistical methods. First, we consider the case where the process of acquiring data is intricate. Two specific examples are considered, including the optimization for black-box functions and the prediction for spatiotemporal flows. In these two cases, data are both expensive and time-consuming to obtain. Therefore, the key is to achieve specific goals while minimizing collection costs. We design proper inputs to feed and apply the Gaussian process model as a surrogate model. Then, some customized methods are employed. For black-box functions, where no explicit expression is available, the optimal approach is to select inputs that are as close as possible to the true optimum. Thus, a sequential design based on Gaussian process models is conducted. It will explore the whole feasible region and find a local area close to the true optimum, only based on limited inputs. Then, more points will be chosen over this region to find the optimum. When predicting flows, the scenario is notably distinct due to the multi-scale nature of the limited data in both spatial and temporal dimensions, which are typically collected and represented as a tensor. Directly applying surrogate models is not feasible because there are plenty of data observations for training, leading to extremely high computational complexity. To handle such data, a tailored multi-output Gaussian process via a compression method is used to extract spatial and temporal basis functions, along with low-dimensional input-dependent coefficients, respectively. Then, surrogate modelling is implemented based on input-dependent features, allowing

for accurate and fast predictions of unobserved flows.

Next, we address the challenge of analysing non-normal data, which includes both continuous and discrete scenarios. The former occurs when assessing the manufacturing processes, and the latter involves analysis of multiple dependent Poisson counts. We handle non-normal distributions based on the following fact: applying the cumulative distribution function (CDF) of a random variable to that random variable yields a uniformly distributed random variable on $[0, 1]$, assuming that this CDF is continuous. When evaluating a manufacturing process, some data on a characteristic of interest are collected first. Then, the process capability indices (PCIs) can be calculated for the quality control. We focus on the case where there are asymmetric tolerances to define PCIs, and the characteristic follows an unknown continuous univariate distribution. Two tailored PCIs are proposed, with an inverse transformation introduced to handle non-normal data. We estimate the underlying CDF of the data via B-splines, transforming the data to be normal. Then, our PCIs can be used to assess the capability of an in-control production process in manufacturing conforming products based on non-normal data. Furthermore, when dealing with multiple Poisson counts, it is essential for both the marginal and joint model specifications to accurately capture the random behavior of the variables. Although there has been a growing interest in this type of data across various scientific fields, models that account for multivariate Poisson distributions remain relatively uncommon. To address that, we introduce a novel multivariate Poisson distribution, leveraging multivariate reduction technology and copula methods, which offers unprecedented flexibility in modeling joint distributions and dependence structures. Various probabilistic properties and several estimation techniques are also explored.

In summary, this thesis makes significant contributions to the field by advancing statistical modeling techniques, particularly in addressing complex and challenging data scenarios.

SAMENVATTING

De kern van dit proefschrift is het toepassen van statistische methoden om problemen aan te pakken die zich voordoen in specifieke situaties waarin gegevens moeilijk te verkrijgen zijn, ruimte-tijdgebonden zijn of geen normale verdeling volgen. De reikwijdte potentiële toepassingen is aanzienlijk. Deze moeilijk oplosbare problemen vereisen vaak innovatieve strategieën, wat aanleiding geeft tot de ontwikkeling van nieuwe statistische modellen en methoden. Hoewel de behandelde onderwerpen in dit proefschrift divers zijn, draait het steeds om statistische modelleringsmethoden voor complexe gegevensstructuren, waarbij traditionele methoden tekortschieten. Onder de besproken onderwerpen vallen onder andere: de optimalisatie van een computationeel dure 'black-box'-functie op basis van beperkte gegevens; het voorspellen van een ruimte-tijdgebonden gegevensstroom bij een nieuwe invoer met behulp van als tensor opgeslagen data; de evaluatie van een productieproces waarbij geen normale verdeling en asymmetrische toleranties in acht worden genomen; en het definiëren van de gezamenlijke verdeling van multivariate Poisson-aantallen in aanwezigheid van afhankelijkheidsstructuren.

In dit proefschrift worden deze onderwerpen onderverdeeld in twee soorten problemen, waarvoor telkens op maat gemaakte statistische methoden worden voorgesteld. Als eerste beschouwen we het geval waarin het verzamelen van gegevens complex is. Twee specifieke voorbeelden worden behandeld: de optimalisatie van 'black-box'-functies en de voorspelling van ruimte-tijdgebonden gegevensstromen. In beide gevallen zijn gegevens duur en tijdrovend om te verkrijgen. Het belangrijkste doel is dan om gewenste resultaten te behalen met minimale verzamelkosten. Hiervoor ontwerpen we geschikte invoerwaarden en passen we het Gaussian Process-model toe als surrogaatmodel. Vervolgens worden aangepaste methoden ingezet. Voor 'black-box'-functies, waarbij geen expliciete vorm beschikbaar is, is de optimale aanpak om invoerwaarden te selecteren die zo dicht mogelijk bij het werkelijke optimum liggen. Daarom wordt een sequentieel ontwerp uitgevoerd op basis van Gaussian Process-modellen. Dit verkent de volledige realiseerbare ruimte en identificeert een lokaal gebied dicht bij het ware optimum, enkel gebaseerd op beperkte invoer. Vervolgens worden extra punten binnen dit gebied geselecteerd om het optimum te nauwkeuriger te benaderen. Bij het voorspellen van stromen is het scenario wezenlijk anders vanwege de multi-scale aard van de beperkte gegevens in zowel

ruimtelijke als temporele dimensies, die doorgaans als een tensor worden opgeslagen. Het direct toepassen van surrogaatmodellen is hier niet haalbaar vanwege de grote hoeveelheid waarnemingen, wat leidt tot een zeer hoge computationele complexiteit. Om dergelijke gegevens te verwerken, gebruiken we een aangepast multi-output Gaussian Process via een compressiemethode om ruimtelijke en temporele basisfuncties te extraheren, samen met invoerafhankelijke coëfficiënten in een lage dimensie. Daarna wordt het surrogaatmodel toegepast op basis van deze invoerafhankelijke kenmerken, wat nauwkeurige en snelle voorspellingen van niet-geobserveerde stromen mogelijk maakt.

Vervolgens behandelen we de uitdaging van het analyseren van niet-normale gegevens, waaronder zowel continue als discrete scenario's vallen. Het eerste geval doet zich voor bij de evaluatie van productieprocessen; het tweede bij de analyse van meerdere afhankelijke Poisson-aantallen. Niet-normale verdelingen worden aangepakt op basis van het volgende principe: het toepassen van de cumulatieve verdelingsfunctie (CDF) van een stochastische variabele op zichzelf resulteert in een uniforme verdeling op het interval $[0, 1]$, mits we aannemen dat de CDF continu is. Bij de evaluatie van een productieproces worden eerst gegevens verzameld over een belangrijke karakteristiek. Vervolgens kunnen procescapaciteitsindices (PCIs) worden berekend ten behoeve van kwaliteitscontrole. Wij richten ons op de situatie met asymmetrische toleranties, waarbij de karakteristiek een onbekende continue univariate verdeling volgt. Hiervoor worden twee aangepaste PCIs voorgesteld, met een inverse transformatie om met niet-normale gegevens om te gaan. We schatten de onderliggende CDF via B-splines en transformeren de gegevens naar een normale verdeling. De voorgestelde PCIs kunnen vervolgens worden gebruikt om de geschiktheid van een stabiel productieproces te beoordelen voor het produceren van conforme producten op basis van niet-normale gegevens. Bij de analyse van meerdere Poisson-aantallen is het van essentieel belang dat zowel de marginale als gezamenlijke modelaannames het willekeurige gedrag van de variabelen accuraat weergeven. Ondanks de toenemende interesse in dit type gegevens binnen diverse wetenschappelijke domeinen, blijven modellen voor multivariate Poisson-verdelingen relatief zeldzaam. Om hieraan tegemoet te komen, introduceren we een nieuw multivariaat Poisson-model dat gebruik maakt van multivariate reductietechnieken en copula-methoden. Dit model biedt ongeëvenaarde flexibiliteit bij het modelleren van gezamenlijke verdelingen en afhankelijkheidsstructuren. Verschillende probabilistische eigenschappen en schattingstechnieken worden eveneens onderzocht.

Samenvattend levert dit proefschrift een belangrijke bijdrage aan het vakgebied door de ontwikkeling van geavanceerde statistische modelleringsmethoden, in het bijzonder voor het analyseren van complexe en uitdagende gegevenssituaties.

1

INTRODUCTION

The use of statistical methods to face the complicated nature of data is a major challenge. To name a few pertinent examples, complex issues may occur when data are costly to gather, rendering the analysis based on comprehensive data unaffordable or extremely difficult to conduct. The data may also encompass spatiotemporal characteristics inherent to fluid flow dynamics, exhibit non-normal distributions, or comprise correlated Poisson random variables. Traditional models may lack efficiency when analyzing them. This thesis aims to address specific types of complex data by developing tailored statistical models and methods.

1.1. LIMITED DATA SETS

Although a vast number of measurements for various variables can now be obtained effectively and at a lower cost than in the past, data collection can still be challenging in some situations. This section examines two issues that are motivated by such a case.

1.1.1. OPTIMIZATION FOR BLACK-BOX FUNCTIONS

In numerous engineering optimization scenarios, a black-box function refers to a mathematical relationship whose explicit analytical form remains inaccessible. Unlike parametric functions defined by tractable expressions, these systems are often characterized implicitly through complex simulation frameworks, experimental measurements, or heuristic rules. A critical challenge arises from the inherent constraints on function evaluations, which are frequently limited by time or cost, thereby restricting the available data for analysis. It presents a significant challenge for global optimization, as traditional machine learning-based approaches heavily rely on extensive function evaluations, leading to prohibitive computational overhead. Additionally, gradient-based methods are infeasible due to the unavailability of analytical derivatives. To address these limitations, Bayesian optimization frameworks have emerged as a principled solution. These approaches assign a prior stochastic model, capturing prior beliefs on the objective function, which is usually a Gaussian process model (also known as a kriging model). The posterior distribution is iteratively refined through sequential new points using observed data points, enabling the construction of a surrogate response surface. Such points are selected by some sampling criteria-acquisition functions derived from the posterior distribution—that balance exploration of the unknown feasible space and exploitation of regions with high potential. As the cycle goes on, the fitted model is more and more accurate, providing a reliable estimated optimum of the black-box function.

Among the well-established acquisition strategies, the expected

improvement criterion remains the most widely adopted due to its closed-form expression and computational tractability. However, its greedy exploitation bias tends to concentrate sampling efforts near the current surrogate optimum, potentially neglecting comprehensive exploration of less promising regions. A further limitation is its sequential nature: traditional acquisition functions suggest only one query point per iteration. This approach may not fully leverage modern computational architectures that support parallel simulations. Given the increasing availability of high-performance computing resources, there is a pressing need to develop parallel Bayesian optimization strategies capable of identifying multiple query points simultaneously, thereby reducing computational overhead. Such advancements would significantly enhance the optimization of black-box systems in real-world engineering applications.

1.1.2. PREDICTION FOR SPATIOTEMPORAL OUTPUTS

In engineering and scientific research, physical experiments often encounter challenges when characterizing complex systems or extreme operating conditions. To address these limitations, high-fidelity numerical simulations have emerged as critical tools for quantifying spatiotemporal dynamics. For instance, in computational fluid dynamics, parameterized time-dependent partial differential equations solved via finite element or finite volume methods are widely used to model transient flow phenomena across multiple scales. However, these high-fidelity simulations impose significant computational burdens, particularly for multi-query applications such as design optimization, uncertainty quantification, and control systems development, which require repeated evaluations across broad parameter spaces.

To address this, Gaussian process models have proven effective as surrogate modelling tools for interpolating complex parametric relationships, providing fast and reliable predictions even with limited data. However, traditional Gaussian process frameworks face inherent limitations when applied to spatiotemporal flows. This is because the parameter space contains only a few outputs that are high-dimensional in both space and time, exceeding the computational tractability of standard Gaussian process implementations. For example, simulating ten parameter configurations could produce millions of spatiotemporal data points, rendering direct application of Gaussian process models infeasible. To mitigate these challenges, reduced-order modelling techniques have been introduced to compress spatiotemporal dynamics into low-dimensional representations. By projecting full-order solutions onto orthogonal basis functions (e.g., via proper orthogonal decomposition), they enable computationally efficient approximations while maintaining physical fidelity. However, existing frameworks struggle to address two

critical challenges: the multiscale nature of fluid dynamics and the sparse sampling of parameter space due to costly simulations. This gap highlights the need for a novel framework that combines reduced-order models to handle spatiotemporal features with Gaussian process models for fast and accurate predictions, enabling data-driven analysis of complex dynamic flows under resource constraints.

1.2. DATA WITHOUT NORMALITY

In real-world applications such as epidemiology, environmental monitoring, and industrial systems, observed datasets often exhibit distributions that significantly deviate from normal distribution assumptions. A prime example arises in discrete count data, where outcomes represent non-negative integers such as patient admissions in healthcare systems or rainfall events in hydrological studies. While Bayesian hierarchical models offer greater flexibility in modelling non-normal data, their implementation often demands substantial computational resources and careful calibration of prior distributions, which hinders their widespread adoption. To address these gaps, this research focuses on two problems: non-normal continuous variables in process capability evaluation and multivariate Poisson counts.

1.2.1. CONTINUOUS UNIVARIATE DISTRIBUTIONS

A motivating example of non-normally distributed continuous data arises in process capability analysis, a critical aspect of statistical process control. In manufacturing and service industries, quality specifications are typically defined by engineering tolerances or customer requirements, which establish acceptable ranges for critical product characteristics. A capable production process must ensure that measured quality attributes consistently align with these specifications. Consequently, after collecting data on these features, the process capability index has been introduced to provide objective measurements of quality control standards. It provides quantitative assessments of how well a process meets predefined tolerances, offering objective benchmarks for evaluating conformance to specifications. For enterprises seeking to optimize operational efficiency and mitigate financial risks, such indices are essential.

Early developments in the process capability index assumed that quality characteristics follow a normal distribution. However, modern manufacturing systems often generate data with non-normal features such as skewness due to complex interactions between process parameters, material variations, or environmental factors. Still using traditional methods based on the normality assumption may cause a misleading assessment as well as potential losses. Existing methods

for non-normal data either aren't interpretable enough or lack flexibility for various kinds of distributions. Meanwhile, it seems unfair to assign equal costs to differences solely based on their equal values, as these differences may occur in different directions, resulting in asymmetric tolerances. To address these issues simultaneously, flexible process capability indices and efficient methods for non-normal data need to be developed.

1.2.2. MULTIVARIATE POISSON DISTRIBUTIONS

Multivariate count data arise across diverse scientific disciplines, including demography, epidemiology, and sports analytics. However, unlike normal distributions, constructing joint distributions for multiple Poisson counts is more challenging, as it requires both marginal and joint specifications to accurately represent the stochastic behaviour of count variables. Additionally, the dependence structures among these variables must be precisely captured.

To address these challenges, several techniques have been proposed for constructing multivariate models capable of accommodating Poisson margins. A frequently adopted strategy involves multivariate reduction frameworks, where shared latent variables are introduced to define dependence. This approach is based on the fact that a Poisson random variable can be decomposed into sums of independent Poisson variables. The primary limitation of such a model is that it can only capture some positive dependence structures. Subsequent advancements, including comonotonic shock models and copula-based approaches, have expanded modelling capabilities but still lack the flexibility to characterize diverse joint distributions. These limitations underscore the critical need to develop an interpretable and flexible multivariate Poisson modelling framework that can accurately capture dependence structures across a broader range of joint distributions than current state-of-the-art approaches.

1.3. THESIS OUTLINE

In this thesis, several tailored statistical models and methods are proposed for different kinds of complicated data sets. The thesis is organized as follows. In Chapter 1, a brief introduction to the subject is given.

Chapter 2 explores the optimization of expensive black-box functions when data acquisition is difficult and budget-limited. Traditional Bayesian optimization methods for this problem, such as expected improvement, evaluate only one point at a time, which is not efficient enough and can easily be trapped in a local optimum. This chapter introduces a novel parallel Bayesian framework based on the minimum energy

criterion to enhance single-point strategies. The proposed framework aims to generate multiple points, thereby reducing the number of iterations and avoiding the risk of becoming trapped in local optima through parallel exploration of the feasible space. Additionally, a shrinkage-augment strategy is introduced to refine the surrogate model of the black-box function, further enhancing optimization efficiency. Numerical experiments and a real example are presented to showcase the superiority of the proposed method over existing Bayesian approaches. These experiments demonstrate the framework's ability to effectively manage limited data and achieve superior performance over existing methods in global optimization tasks.

Chapter 3 focuses on analyzing dynamic datasets from flow fields, represented as tensors, to make spatiotemporal predictions. These datasets typically originate from high-fidelity simulators, which have become crucial for understanding complex systems as alternatives to physical experiments. This chapter proposes a tailored multi-output Gaussian process surrogate model for spatiotemporal emulation, incorporating uncertainty quantification. Our approach leverages singular value decomposition to project high-dimensional spatiotemporal data into a low-dimensional latent space, with explicit analysis of decomposition errors. This reformulates the multi-output Gaussian process as a linear combination of independent Gaussian processes, each modeling the relationship between projection coefficients and simulator inputs. The resulting surrogate achieves reliable prediction of unobserved fluid dynamics while significantly accelerating computation, facilitating efficient design exploration. Validation through numerical experiments and real-data case studies demonstrates superior accuracy and efficiency over existing methods.

The next two chapters focus on the analysis of non-normal data. Chapter 4 addresses the challenge of process capability analysis using process capability indices for asymmetric tolerances and non-normal data. To the best of our knowledge, there is no method considering these two issues together, which are often involved in real-world scenarios. We introduce two new classes of process capability indices designed specifically for asymmetric tolerances, discuss practical estimation methods, and establish confidence intervals based on asymptotic variance. To address the non-normal distributions of data, we propose constrained B-spline regression to estimate the cumulative distribution function directly, transforming the data to be normal via a tailored inverse transformation procedure. Some asymptotic results for the estimator are also studied. Through numerical simulations and a real-world case study, we illustrate the effectiveness of these indices in accurately handling asymmetric tolerances and non-normal data sets.

Chapter 5 proposes a multivariate model for multiple dependent Poisson count data. In this chapter, we define an intuitive and interpretable

multivariate Poisson distribution, allowing for more flexibility than existing models in terms of joint distributions and dependence structures. Our proposed model generalizes existing methods by reducing itself to them when some of our model's parameters are preset. We establish several essential probabilistic properties and present a set of propositions that underscore our model's novelty. We also provide an extensive discussion on maximum likelihood estimation, inference function for margins, and the expectation-maximization algorithm. To validate the performance of our model, we conduct simulations and illustrate its practical utility through a real-world example.

Chapter 6 concludes the thesis, and Chapter 7 consists of some supplemental materials containing technical details and additional simulation results.

2

A PARALLEL FRAMEWORK FOR BLACK-BOX OPTIMIZATION

Parts of this chapter have been published in Journal of Statistical Computation and Simulation, 93(17): 3104-3125, 2023.

2.1. INTRODUCTION

In many applications, it is difficult to solve the following black-box optimization problem:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\Omega} G(\boldsymbol{x}),$$

where $G(\boldsymbol{x})$ is a single-objective black-box function and it is assumed to have only one optimum, $\boldsymbol{x} \in \mathbb{R}^p$ is a p-dimensional input variables, and $\Omega \subseteq \mathbb{R}^p$ is the related feasible domain. The term "black-box function" is commonly defined as a complex, large-scale system that involves numerical simulations or experimental measurements. Unlike traditional functions, these systems lack explicit mathematical expressions for their input-output relationships or gradients, making the use of classical gradient-based algorithms impractical. Moreover, evaluating black-box functions is computationally expensive due to the significant time required for simulations or measurements. For instance, models used in wind plant control strategies [1] or high-resolution reservoir engineering models [2] can take hours or even days to evaluate once. Thus, evolutionary algorithms and swarm intelligence algorithms, which typically require a large number of function evaluations to produce reliable results, are often not well-suited. This inherent complexity and computational cost make black-box functions a critical challenge in optimization problems.

To deal with this problem, some Bayesian optimization (BO) methods have been proposed. These methods regard the objective function G as a realisation of a stochastic process. An initial surrogate model $\hat{G}(\boldsymbol{x})$, which is usually easy and quick to evaluate, is first built via some previously evaluated samples to approximate $G(\boldsymbol{x})$. In each iteration, an infill sampling criterion or an acquisition function based on $\hat{G}$ is used to select a candidate point, maximally reducing the uncertainty about the location of the optimum of the objective function. The point is then evaluated using G and used to update the surrogate model. The sequential sampling process goes on until some stopping condition is met. Queried points will improve the accuracy of $\hat{G}(\boldsymbol{x})$, especially over the domain around $\boldsymbol{x}^*$. Thus, the optimum of the surrogate model $\hat{\boldsymbol{x}}^*$ can be used as an estimate of the global optimum. In the literature, many surrogate models and acquisition functions have been proposed. About the former, the most popular one is the Gaussian process (GP) [3]. Meanwhile, the typically used acquisition function is the expected utility of a new point given the observed data. [4] proposed the expected improvement (EI) method, which is one of the most popular BO approaches, by introducing a simple closed-form acquisition function. [5] studied the convergence of the EI method under the assumption that the covariance function of the surrogate model is non-degenerate.

Despite the popularity of EI, there are still some drawbacks. [6] noted that conclusions of [5] are misleading in some applications. They

extended the result to more general settings, showing that there exist some smooth functions for which EI can not find the global optimum. [7] further pointed out for which EI over-exploits the fitted model on $G(\cdot)$ and under-explores optimization space Ω. To address that, some studies have been proposed in the literature. [8] improved the EI by setting the infill criterion as the expected value of $I^g(\boldsymbol{x})$ with an integer-valued g, where $I(\boldsymbol{x})$ denotes the improvement at $\boldsymbol{x}$. The larger the value of g is, the more exploration of the feasible region that the algorithm will encourage. [9] introduced a weight into EI to adjust the trade-off between exploration and exploitation. [7] proposed the hierarchical expected improvement (HEI) with a hierarchical Bayesian approach to encourage global exploration. Other than EI, BO methods have also considered other acquisition functions, including upper confidence bound [10–12], Thompson sampling [13], and predictive entropy search [14]. Most of them are one-point methods, only querying one point in each iteration. It may not be efficient enough, especially when parallel calculations are feasible.

To address that, some parallel frameworks have been established in the literature. For example, [15] proposed the q-EI infill criterion, generating multiple points in each iteration. However, the criterion is extremely time-consuming to calculate due to the high-dimensional integrations involved. Instead of solving q-EI exactly, [16] recommended using two heuristic strategies: Kriging Believer (KB) and Constant Liar (CL). Subsequently, several studies have explored parallel optimization approaches, including [17–21]. However, these methods primarily focus on selecting multiple points for evaluation rather than balancing exploitation and exploration. This limitation reduces their applicability and efficiency, as they may not achieve optimal performance in complex and computationally expensive optimization problems.

For Bayesian optimization, we find that although numerous methods have been developed, there remains an outstanding question of proposing a parallel BO framework. The method should work well for considering parallel simulation, exploration-encouraging, and cost-saving simultaneously. In this chapter, we propose a new parallel method using the minimum energy criterion to overcome this challenge. The new method is able to query multiple points using a black-box function near the global optimum. At the same time, it will explore the optimization space better due to the characteristics of the minimum energy criterion. Since the new method considers the EI or HEI criterion as the charge function, it preserves the desirable properties of the EI or HEI. Additionally, the parallel strategy can reduce the number of iterations. It makes a better trade-off between exploitation and exploration. To improve the prediction of the surrogate model for the black-box function, a heuristic shrinkage-augment strategy is also presented. It can place more points near $\boldsymbol{x}^*$ and fit a more accurate local surrogate model. We integrate

the new parallel method and the shrink-augment strategy as a new framework. It can encourage exploration and find a better estimation of the global optimum. Finally, some numerical and illustrative examples are presented to demonstrate the superiority of the proposed framework.

The remainder of this chapter is organised as follows. Section 2.2 reviews the GP model and two existing one-point infill acquisition functions, EI and HEI. The parallel method based on the minimum energy criterion is presented in Section 2.3. This section also presents the proposed heuristic shrink-augment strategy and the new efficient optimization algorithm. Section 2.4 performs a comparative study for benchmarking purposes. Some numerical simulations are conducted to compare the new method with some existing methods. Further, an application to the optimization problem of a wind plant is presented. Finally, the conclusions are drawn in Section 2.5. Additional simulation results are shown in Section 7.1.

2.2. REVIEW ABOUT THE GP MODEL AND EI/HEI CRITERIA

2.2.1. GP MODEL

We first use the following GP model [3] as the surrogate model for the black-box objective function G:

$$\tilde{G}(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x}) = \boldsymbol{f}^{\top}(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}), \tag{2.1}$$

where $\mu(\boldsymbol{x})$ is the mean function of the process, $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}) \dots f_d(\boldsymbol{x}))^{\top}$ are d known basis functions and $\boldsymbol{\beta} \in \mathbb{R}^d$ are corresponding unknown coefficients. $Z(\boldsymbol{x}) \sim GP(0, \sigma^2 K)$ is a stationary Gaussian process. $K(\cdot,\cdot)$ denotes the correlation function of GP, and σ^2 is the process variance. Model (2.1) is called the universal kriging model in geostatistics [22]. It reduces to the ordinary kriging model if there is a constant mean, i.e., $\boldsymbol{f}^{\top}(\boldsymbol{x}) = \boldsymbol{1}$.

Let $\mathcal{D}_n = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i$ is the input of the black-box function and y_i is the corresponding response. Let $\boldsymbol{y}_n = (y_1 \cdots, y_n)^{\top}$ be the vector of observed responses. According to the GP model, correlation between $Z(\boldsymbol{x}_i)$ and $Z(\boldsymbol{x}_j)$ only depends on the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, which means that $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = R(H_{ij})$ with some unknown parameters, where $H_{ij} = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2$ with the Euclidean norm $||\cdot||_2$. There are several families of commonly used correlation functions, such as the Gaussian family and Matérn family. In this chapter, we use the latter, which is defined as

$$R(H) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu}H)^{\nu}B_{\nu}(2\sqrt{\nu}H)$$

where B_{ν} is the modified Bessel function of the second kind, and ν represents a smoothness parameter. Let $\boldsymbol{K}_n(\boldsymbol{x}) = (K(\boldsymbol{x}, \boldsymbol{x}_i))_{i=1}^n$ be the

correlation column vector between the new query point and observed points, $\mathbf{f}_n = (\boldsymbol{f}(\boldsymbol{x}_1), \cdots, \boldsymbol{f}(\boldsymbol{x}_n))^\top$ be values of the basis functions at the observed points, and $\mathbf{R}_n = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1}^n$ be the correlation matrix of the observed points. Denote the parameters in correlation function $K(\cdot,\cdot)$ and σ^2 as $\boldsymbol{\kappa}$. [3] showed that the posterior distribution of $G(\boldsymbol{x})$ at an unobserved input $\boldsymbol{x}$, given $\boldsymbol{\kappa}$, can be derived as

$$[G(\boldsymbol{x})|\mathcal{D}_n] \sim N\big(g(\boldsymbol{x}), s^2(\boldsymbol{x})\big),$$

where $g(\boldsymbol{x}) = \boldsymbol{f}^\top(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{K}_n^\top(\boldsymbol{x})\mathbf{R}_n^{-1}(\boldsymbol{y}_n - \mathbf{f}_n^\top\boldsymbol{\beta})$ is the best linear unbiased prediction (BLUP) of $G(\boldsymbol{x})$ and $\boldsymbol{\beta} = (\mathbf{f}_n^\top\mathbf{R}_n^{-1}\mathbf{f}_n)^{-1}\mathbf{f}_n^\top\mathbf{R}_n^{-1}\boldsymbol{y}_n$. The posterior variance at an unobserved point is derived as

$$s^2(\boldsymbol{x}) = \sigma^2\Big[1 - \boldsymbol{K}_n^\top(\boldsymbol{x})\mathbf{R}_n^{-1}\boldsymbol{K}_n(\boldsymbol{x}) + \boldsymbol{h}^\top\mathbf{Q}^{-1}\boldsymbol{h}\Big],$$

where $\boldsymbol{h} = \boldsymbol{f}(\boldsymbol{x}) - \mathbf{f}_n^\top\mathbf{R}_n^{-1}\boldsymbol{K}_n(\boldsymbol{x})$ and $\mathbf{Q} = \mathbf{f}_n^\top\boldsymbol{K}_n^{-1}\mathbf{f}_n$. Note that $\boldsymbol{\kappa}$ needs to be estimated from the observed data. The most popular method is maximum likelihood estimation (MLE) (see [3] for more details). Via estimated $\hat{\boldsymbol{\kappa}}$, the corresponding estimated posterior mean is $\hat{G}$, the empirical best linear unbiased prediction, and estimated posterior variance is $\hat{s}^2$.

2.2.2. EI AND HEI

When a surrogate model is constructed, the following important step in Bayesian optimization methods is querying new points according to an appropriate acquisition function. [4] proposed the most popular EI infill criterion, which can be described as follows. Let $g_{\min}$ be the minimum function value at the observed points, i.e., $g_{\min} = \min_{\boldsymbol{x}\in\mathcal{X}_n}\{G(\boldsymbol{x})\}$, where $\mathcal{X}_n = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$. The improvement at the point $\boldsymbol{x}$ can be defined as $I(\boldsymbol{x}) = \max\{g_{\min} - G(\boldsymbol{x}), 0\}$. Taking the expected value with respect to the posterior distribution of $G(\boldsymbol{x})$, we can obtain the EI acquisition function as follows

$$EI(\boldsymbol{x}) = E[I(\boldsymbol{x})] = (g_{\min} - \hat{G}(\boldsymbol{x}))\Phi\left(\frac{g_{\min} - \hat{G}(\boldsymbol{x})}{\hat{s}(\boldsymbol{x})}\right) + \hat{s}(\boldsymbol{x})\phi\left(\frac{g_{\min} - \hat{G}(\boldsymbol{x})}{\hat{s}(\boldsymbol{x})}\right), \quad (2.2)$$

where Φ and ϕ are the cumulative distribution functions and probability density function of the standard normal distribution, respectively. Then, $\widetilde{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\Omega}{arg\ max}\, EI(\boldsymbol{x})$ is the best next query point. We can see that in (2.2), there are two parts that indicate the trade-off between exploitation and exploration. For points $\boldsymbol{x}$ with a smaller posterior means $\hat{G}(\boldsymbol{x})$, the first term in (2.2) will be large, which implies the exploitation. For $\boldsymbol{x}$ with larger posterior variances $\hat{s}(\boldsymbol{x})$, the second term in (2.2) will be large, which implies the exploration. Several studies have noted that EI shows over-greediness because $\phi(\cdot)$ in the second term of (2.2) converges to 0

more quickly than $\Phi(\cdot)$. As mentioned before, there exist some functions such that EI may fail to find any global optimum [7].

To overcome this drawback of EI, [7] proposed an HEI acquisition function via a hierarchical GP model. In [7], the unknown coefficients $\boldsymbol{\beta}$ are assigned non-informative priors. The process variance σ^2 is assigned a conjugate inverse-gamma prior, which is

$$[\boldsymbol{\beta}] \propto \mathbf{1}, \tag{2.3a}$$

$$[\sigma^2] \sim IG(a, b). \tag{2.3b}$$

Under model assumption (2.3), [7] showed that the posterior distribution of $G(\boldsymbol{x})$ is

$$[G(\boldsymbol{x})|\mathcal{D}_n] \sim T_1(2a + n - d, \hat{G}(\boldsymbol{x}), \tilde{s}^2(\boldsymbol{x})),$$

where $T_1(\eta, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$ is a 1-dimensional non-standardized t-distribution with degrees of freedom η,location vector $\boldsymbol{\gamma}$ and scale matrix $\boldsymbol{\Sigma}$, $\tilde{s}^2(\boldsymbol{x}) = \tilde{\sigma}^2\left[1 - \boldsymbol{K}_n^\top(\boldsymbol{x})\mathbf{R}_n^{-1}\boldsymbol{K}_n(\boldsymbol{x}) + \mathbf{h}^\top\mathbf{Q}^{-1}\mathbf{h}\right]$, $\tilde{\sigma}^2 = (2b + n\hat{\sigma}^2)/(2a + (n-d))$, and $\hat{\sigma}^2$ is the MLE of σ^2. Then, deriving the expectation of $I(\boldsymbol{x})$ under hierarchical prior, we have the HEI criterion as

$$HEI(\boldsymbol{x}) = (g_{\min} - \hat{G}(\boldsymbol{x}))\Phi_{\nu_n}\left(\frac{g_{\min} - \hat{G}(\boldsymbol{x})}{\tilde{s}(\boldsymbol{x})}\right) + m_n\tilde{s}(\boldsymbol{x})\phi_{\nu_n - 2}\left(\frac{g_{\min} - \hat{G}(\boldsymbol{x})}{m_n\tilde{s}(\boldsymbol{x})}\right),$$

where m_n and ν_n are corresponding parameters in [7]. When $a < d/2$, $\tilde{s}^2(\boldsymbol{x})$ is larger than $\hat{s}^2(\boldsymbol{x})$, which implies that the HEI indeed encourages exploration. It is worth noting that these two methods are both non-parallel and require many iterations to obtain a good solution. In the next section, a parallel framework will be proposed for improving such one-point acquisition criteria.

2.3. A PARALLEL BAYESIAN OPTIMIZATION FRAMEWORK

In this section, one novel parallel framework will be proposed based on the minimum energy criterion. Because of the advancement of supercomputers, it is now possible to run multiple complex models simultaneously with no increased costs. When the objective function is a time-consuming black-box function, selecting a batch of points once is more appropriate, as noted by [17].

To address that, we propose a new sampling method for Bayesian optimization. We show that it is a parallel version of the one-point acquisition criteria. The sampling function is straightforward to understand and implement. And it can handle a high number of simultaneous evaluations. By involving one-point criteria such as EI and HEI, the proposed method could preserve the desirable properties of the original EI or HEI. Moreover, it makes an interpretable trade-off between

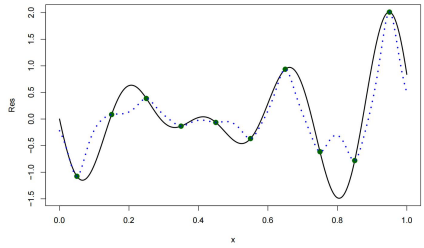
exploration and exploitation. Then, we apply a shrink-augment strategy to shrink the feasible area sequentially, encouraging the exploitation. A final two-stage parallel framework is established. It can find a better optimal estimation solution without too many greedy calculations.

2.3.1. A PARALLEL METHOD BY USING MINIMUM ENERGY CRITERION

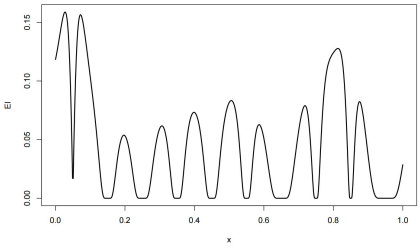
By maximising the EI or HEI function, we acquire the most valuable point, update the surrogate model, and iterate. Finally, we are going to get an estimation of the minimum. From (2.2) and (2.3), we can see that there are two parts in the EI or HEI acquisition function: exploitation and exploration. According to the first term, the points with smaller predicted values will have larger EI or HEI values. On the other hand, the latter term assigns greater values for points from the area in which there are fewer queried points. The points with smaller responses and high prediction variances should be considered at the same time to reduce iterations. Therefore, there is a strong need to query multiple points at each iteration based on the current fitted surrogate model. A natural idea is to query points that have higher EI or HEI criteria. However, these points with larger EI or HEI values usually concentrate in a local area. When the surface of the EI or HEI criterion has a very sharp peak, it doesn't balance exploitation and exploration very well. This is a waste in terms of objective function evaluations.

Figure 2.1 gives an illustrative toy example. We construct a Gaussian process model, as depicted in Figure 2.1(a), based on ten space-filling points. These points are arranged to be uniformly distributed within the design region. Such uniformity can be defined by certain distance-based or discrepancy-related criteria. Then we need to select a batch of points based on the EI curve, which is presented in Figure 2.1(b). One natural idea is to choose peaks of the EI curve. However, there are still some drawbacks. For instance, the first and second peaks will be selected according to Figure 2.1(b) if we need two points in each iteration, They will be too close to each other, which means the information provided by them is similar, being unfavourable for searching the whole feasible region. Meanwhile, these two points may result in a singular correlation matrix and an unstable GP model. To select separated points with larger EI values, we develop a method to choose any number of space-filling points with relatively high criterion values.

For the EI or the HEI curve, the region we are interested in is the whole feasible domain. Meanwhile, points with high acquisition function values should also be considered. Inspired by [23], we apply the minimum energy criterion to query points of particular interest in the feasible space. A parallel sampling method is then proposed. Based on this



(a) The solid line represents an objective function that has been sampled at ten points shown as dark green circles. The dotted line is the fitted Gaussian process model.



(b) The expected improvement function when only ten points have been sampled.

Figure 2.1: A toy example

method, the points can be obtained by minimising the following criterion

$$E(\boldsymbol{x}_q) = \max_{i \neq j} \frac{q(\boldsymbol{x}_i) q(\boldsymbol{x}_j)}{d(\boldsymbol{x}_i, \boldsymbol{x}_j)},$$

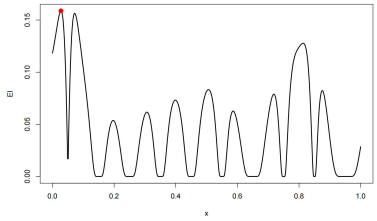
where $q(\boldsymbol{x})$ is called a charge function and $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the distance measure between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We take the charge function $q(\boldsymbol{x}) = 1/\gamma(\boldsymbol{x})$, where $\gamma(\boldsymbol{x})$ denotes the value of the acquisition function at $\boldsymbol{x}$. Then, it is equivalent to query $\boldsymbol{x}_q$ by maximizing the following criterion

$$\Psi(\boldsymbol{x}_q) = \min_{i \neq j} \gamma(\boldsymbol{x}_i) \gamma(\boldsymbol{x}_j) d(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{2.4}$$

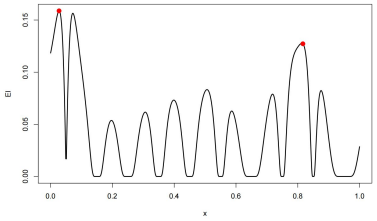
The intuitive explanation about (2.4) can be described as follows. We assume that the related optimization region is a bounded box and regard every point as a charged particle. We set these points with the same sign; each point will repel the other as much as possible. Moreover, these particles will occupy positions inside the box to minimise the total potential energy. Consequently, the desirable queried points consist of the positions occupied by particles. This is the so-called 'repulsion principle'. It requires all particles to have charges with the same sign, which is easy to accomplish because EI or HEI itself is always non-negative. By maximising the product of the charge function and distance, points with high acquisition values will be located in the whole feasible region. We can not only find the global maximum of EI or HEI but also explore regions with high uncertainty. Unlike choosing peaks, this criterion takes into account the distance. Also, it spreads out the queried points in the region of interest to avoid a local optimum. This physical analogy helps to improve sequential designs. It balances the criteria values and distance in a physical and interpretable way.

Figure 2.2 shows some batches of points selected (red circles) with different numbers by the proposed sampling method. As shown in Figure 2.2(a), the highest peak is selected as the first point. By integrating the EI function into the parallel method, the first point selected will be the same as the one selected by EI. It suggests that our sampling method won't be worse than EI or other one-point criteria. According to the first peak, the next points are sequentially selected based on the criterion (2.4). Figure 2.2(b) shows the second point to query. We can see that even if the second peak is higher, it is not chosen to query because it is so close to the first peak. By contrast, the third-highest peak is regarded as the second point due to a greater distance. The proposed method first chooses the points with the highest EI value in the sense of considering the best possible performance. Then it will put more weight on areas with high uncertainties. As a result, a highly valued point that is far away from the first point will be chosen. Therefore, as a parallel method, it indeed considers exploitation and exploration at the same time. For the case where five points are queried during each iteration, the chosen samples are shown in Figure 2.2(c). The second peak is still not queried to make sure the points selected are spread out, which encourages exploration a lot. Meanwhile, all points are located near different peaks. Corresponding larger EI values result in efficient exploitation. To sum up, the sampling method measures the average performance and the uncertainty in a comprehensive way. Figure 2.2(d) shows the flexibility with respect to the number of points. Not limited by the peaks, we can query any number of points during each iteration, depending on the resources we have to make parallel simulations. As the number of points q increases, the samples via the minimum energy criterion constitute a space-filling design, which is also noted by [23].

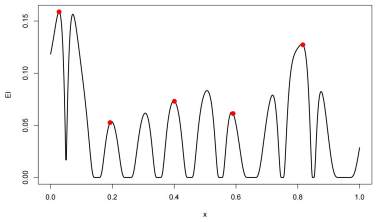
As introduced before, it is clear that our proposed sampling criterion will place queries as far apart as possible. At the same time, it selects more points from the area with high acquisition function values. In other words, the query points will either have smaller objective function values or depart from the observed points. These points are calculated using G in parallel to enrich $\mathcal{D}_n$ and update the surrogate model. Note that both the EI and HEI acquisition functions are closed-form. Therefore, choosing the next query point by maximising criterion (2.4) is not difficult. Furthermore, [24] presented a fast algorithm to optimise (2.4). An R package called `mined` is also provided. By integrating EI or HEI, we still try to generate points with these infill criteria. This suggests that the new method still has its desirable properties. Beginning with the point that has the maximal criteria value, our parallel framework won't be worse than EI or HEI. In addition, the minimum energy criterion can not only guide sampling multiple points at the same time but also make sure that points will be apart to encourage exploration.



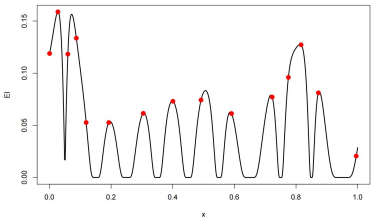
(a) Only one point needs to be queried.



(b) Two points need to be queried.



(c) Five points need to be queried.



(d) Fifteen points need to be queried.

Figure 2.2: Different batches of points to be queried (solid line: EI curve, red circles: selected points).

2.3.2. A SHRINK-AUGMENT STRATEGY

As more and more points are queried, the region where the EI or HEI acquisition functions are positive narrows because the EI or HEI values at the observed points are 0. The surrogate model could be unstable, owing to the singularity of the correlation matrix. In this case, the parallel method can only find the same point as the iteration progresses, which is inefficient for updating $\hat{G}(\boldsymbol{x})$. Recall that there is only one global optimum of G. In the optimization problems, our major interest centres around the area that includes the global optima. Therefore, we need a local method to further explore the feasible region and acquire a better estimated optimum $\hat{\boldsymbol{x}}^*$.

To perform a local optimization, we draw inspiration from a trust region Bayesian optimization in [25]. Unlike choosing the trust region at the beginning, we decide to scale the feasible domain when the optimum of the surrogate model converges or our framework can't find multiple points. Benefiting from the proper balance made by our new sampling algorithm, we have confidence in the final surrogate model to some extent. That is why we apply the information inside $\hat{G}(\boldsymbol{x})$ to develop a shrink-augment strategy. In this strategy, we first determine the best point based on the current observations and the surrogate model. Next, a hyperrectangle centred on the best solution is created. Then, we rescale it and query new points in the new feasible sub-domain.

For the size of this hyperrectangle, it should be sufficiently large to contain good solutions and be small enough to find the true optimum as well. Here, we let it contain enough points of $\mathcal{D}_n$ to recreate a local surrogate model, usually between $5p$ and $10p$. With these experimental points, we don't need to make calculations at new points. It saves the time a lot, especially for an expensive black-box function. Then, we rescale this local rectangle to be a larger one. In this way, the region around the best solution will be less sharp and easier to explore. The parallel sampling algorithm is again applied for a further search. If the updated $\hat{\boldsymbol{x}}^*$ makes some progress, it suggests that the local area around the current optimal solution is proper. If not, the hyperrectangle may be too small to improve optimal solutions. Then we expand it to contain double the number of experimental points in $\mathcal{D}_n$. So far, we have detailed a single local shrink-augment strategy. For the case where the objective function has multiple minima, we can create multiple regions based on $\mathcal{D}_n$ and $\hat{G}$. In this chapter, we only focus on single-objective problems. The details of the strategy are as follows (take the feasible region $[0, 1]^p$ as an example).

1. The point with the minimum posterior mean value is searched based on the current surrogate model, that is $\hat{\boldsymbol{x}}^* = \underset{\boldsymbol{x}\in[0,1]^p}{\arg\min}\, \hat{G}(\boldsymbol{x})$.

2. Determine a hyperrectangle region Ω^* that includes m already

chosen points $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$ centring around $\hat{\boldsymbol{x}}^*$ in $\mathcal{D}_n$.

3. Update the observation data as $\mathcal{D}_m = \{(\boldsymbol{x}_i, G(\boldsymbol{x}_i))\}_{i=1}^m$.

4. Rescale Ω^* to $[0,1]^p$ and define it as a new feasible domain; at the same time, transform $\mathcal{D}_m$ identically and update the surrogate model.

5. Get the updated optimum via parallel searching. If it is not better, let $m = 2m$, back to Step 2.

2.3.3. THE TWO-STAGE PARALLEL FRAMEWORK

In this subsection, the parallel sampling method is integrated with the shrink-augment strategy. Finally, we establish a new parallel efficient global optimization framework. Without loss of generality, we present it by choosing the EI acquisition function, which is referred to as PMEI. The details of the framework are presented in the following Algorithm 1. During the optimization, the feasible domain is assumed to be a unit hyperrectangle, $[0,1]^p$. In Algorithm 1, Steps 6–15 query new points sequentially in a parallel sampling way to enrich observation data $\mathcal{D}_n$. Steps 5 and 16-24 adjust the feasible domain according to the shrink-augment strategy. The steps for the PMHEI, which chooses the HEI acquisition function as $\gamma(\cdot)$ in (2.4), are similar.

2.4. NUMERICAL EXPERIMENTS

In this section, we perform a comparative study to present the performance of our new parallel framework. To illustrate the benefits, we implement EI and qEI. The latter is based on two heuristic ways. As noted by [16], CL can perform comparably with q-EI. Therefore, only CL with the minimum and maximum of the calculated responses are chosen, denoted as CLmin and CLmax. Also, the parallel algorithm involving HEI is considered to present the flexibility of our framework. Finally, a practical optimal example of the windmill is presented. All implementations have been done in the `R` environment [26]. We use the packages `DiceOptim` for EI and CL, `mined` for minimum energy criterion, `MaxPro` for initial design, and `DiceKriging` for the surrogate model.

2.4.1. SIMULATION-BASED BENCHMARKING

We consider three benchmark problems to show some visualised results. The three optimization functions are described as [27]

- **Three-hump camel** (function of two input variables on domain $\Omega = [-2,2]^2$)

Algorithm 1: Two-stage Parallel Framework PMEI.

Input:
G:p-dimensional real black-box function;
n: initial sampling size;
q: number of points chosen during each iteration;
m: number of existing points in the shrunk region.
Output: Predicted optimum $\hat{\boldsymbol{x}}^*$.

$\mathcal{X}_n \leftarrow n$ initial data in $[0,1]^p$;
$\boldsymbol{y}_n \leftarrow$ parallelly calculate $G(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}_n$;
$\Omega \leftarrow [0,1]^p$;
repeat
  Rescale Ω to $[0,1]^p$ and transform $\mathcal{X}_n$ identically;
  repeat
    $\mathcal{D}_n \leftarrow \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \in \mathcal{X}_n, y_i \in \boldsymbol{y}_n, i = 1, \cdots, n\}$;
    $g_{\min} \leftarrow$ get min $\boldsymbol{y}_n$;
    $\hat{G}(\cdot) \leftarrow$ create a surrogate model based on $\mathcal{D}_n$;
    $\mathcal{X}_q \leftarrow$ generate q points by choosing $\gamma(\cdot)$ in (2.4) as EI;
    $\boldsymbol{y}_q \leftarrow$ parallelly calculate $G(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}_q$;
    $\mathcal{X}_n \leftarrow \mathcal{X}_n \cup \mathcal{X}_q$;
    $\boldsymbol{y}_n \leftarrow \boldsymbol{y}_n \cup \boldsymbol{y}_q$;
    $n \leftarrow n + q$;
  until *PMEI finds the same point*;
  $\hat{\boldsymbol{x}}^* \leftarrow \arg\min_{\boldsymbol{x} \in \Omega} \hat{G}(\boldsymbol{x})$;
  repeat
    $\Omega^* \leftarrow$ a rectangular centering at $\hat{\boldsymbol{x}}^*$ including m observed points;
    $n \leftarrow m$;
    $\mathcal{X}_n \leftarrow m$ existing points in Ω^*;
    $\boldsymbol{y}_n \leftarrow$ corresponding responses of $\mathcal{X}_n$;
    $m \leftarrow 2m$;
    $\hat{\boldsymbol{x}}^u \leftarrow$ update $\hat{\boldsymbol{x}}^*$
  until $\hat{\boldsymbol{x}}^u$ *is better than* $\hat{\boldsymbol{x}}^*$;
until *Budget reaches or* $\hat{\boldsymbol{x}}^*$ *converges*;

$$G(\boldsymbol{x}) = 2x_1^2 - 1.05x_1^4 + \frac{x_1^6}{6} + x_1x_2 + x_2^2,$$
$$\boldsymbol{x}^* = (0,0), G(\boldsymbol{x}^*) = 0.$$

- **Griewank function** (function of two input variables on domain $\Omega = [-10, 10]^2$)

$$G(\boldsymbol{x}) = \sum_{i=1}^{2} \frac{x_i^2}{4000} - \prod_{i=1}^{2} \cos(\frac{x_i}{\sqrt{i}}) + 1,$$
$$\boldsymbol{x}^* = (0,0), G(\boldsymbol{x}^*) = 0.$$

- **Hart6 function** (function of six input variables on domain $\Omega = (0,1)^6$)

$$G(\boldsymbol{x}) = -\sum_{i=1}^{4} \alpha_i \exp(-\sum_{j=1}^{6} A_{ij}(x_j - P_{ij})^2),$$
$$\alpha = (1, 1.2, 3, 3.2)^T,$$
$$\mathbf{A} = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix},$$
$$\mathbf{P} = 10^{-4} \begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix},$$
$$\boldsymbol{x}^* = (0.20169, 0.150011, 0.476874, 0.275332, 0.311652, 0.6573),$$
$$G(\boldsymbol{x}^*) = -3.32237.$$

Corresponding visualisations are available at the virtual library. The surface of the Three-hump function is easy to explore. It is proper to demonstrate the efficiency of our shrink-augment strategy. The Griewank function has many widespread local minima, which are used to test whether optimization methods can find the global optimum. The Hart6 function produces a high-dimensional surface with some local optima, which denotes a more complicated feasible region. They assess how optimization methods perform when the problem is tractable, has many local solutions, or involves high-dimensional feasible regions. All feasible regions will be rescaled to be $[0,1]^p$ in simulations.

As the dimension of functions increases, the objective function surface becomes more complicated. In practice, a complex function usually costs

a lot to evaluate once. It is not trivial to calculate it many times. To find the optimal solution quickly, more points at one iteration should be queried. Thus, we set different q and iteration times for all benchmark functions. Let $q = 5,\ 10$ for 2-dimensional cases, and $q = 20,\ 30$ for the 6-dimensional case. The total number of iterations is set as $40,\ 40,\ 20$, respectively. We also compare PMHEI, PMEI, EI, and HEI to show the flexibility of our proposed method. We are interested in both the optimum provided by surrogate models and observed data. Hence, we calculate $\log_{10}(|G(\boldsymbol{x}^*) - G(\hat{\boldsymbol{x}}^*)|)$ to show the accuracy of $\hat{G}$ using different methods, and $\log_{10}(|G(\boldsymbol{x}^*) - g_{\min}|)$ to indicate the quality of the query points. When considering the computational cost, given that the time required for calculating the benchmark function is nearly negligible, the number of evaluations of G can be utilised as a measure. Note that one-point methods share the same number of evaluations in each iteration as methods that select multiple points due to the parallel calculation. Thus, the number of evaluations is directly proportional to the number of iterations, and we employ the latter for our comparison. Given an acceptable error bound, more iterations mean more time costs. All average simulation results over 20 replications are presented.

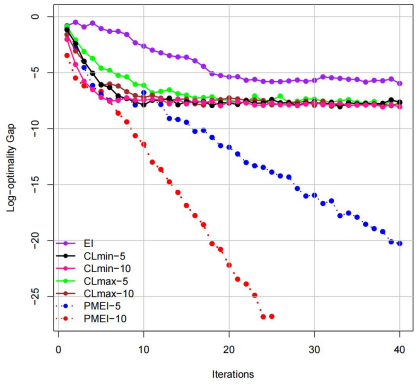
Based on 40 iterations, numerical results for 2-d benchmark functions are shown in Figure 2.3. For the Three-Hump function, the logarithm gaps based on $\hat{G}$ and $\mathcal{D}_n$ suggest that PMEI outperforms other methods. Meanwhile, it indicates that the proposed framework considers the current optima very well. Furthermore, due to multiple points queried at one time, PMEI finds a better solution than EI. It can improve EI in terms of accuracy. In the first few steps, CL and PMEI both output similar results. As the iteration proceeds, PMEI begins to shrink the feasible region. The error for PMEI drops sharply while other methods begin to stagnate. This shows that the heuristic strategy appears to benefit the optimization. In addition, it seems that PMEI-10 converges more rapidly because it stops at the 18th step with the lowest gap. Moreover, given an error tolerance of about 10^{-5}, PMEI and CL both spend about three iterations. However, EI needs four times the costs to get similar solutions, which suggests that PMEI can save time and costs by reducing iterations. For a friendly search landscape, each method can provide an appropriate solution. Because of the shrinkage-augment strategy, PMEI sets more points around $\boldsymbol{x}^*$ and provides a better result.

Figures 2.3(c) and 2.3(d) demonstrate the performance in dealing with the situation where there are lots of local optima. PMEI could explore the whole region during each iteration. It still queries more points with higher qualities while EI gets trapped in a local optimum. This suggests that PMEI can correct over-greediness and enhance EI in terms of exploration-encouraging. With the same q, PMEI performs better in creating an accurate surrogate model. CLmin queries a closer point to the true optimum, which PMEI also selects the point with comparable quality.

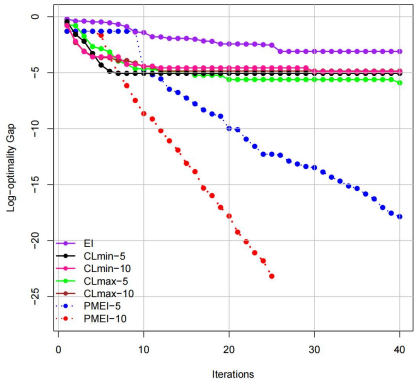
Figure 2.4(a) presents the points produced by the CLmin, CLmax, and PMEI. We can see that there are many local optima near $\boldsymbol{x}^* = (0.5, 0.5)$. CLmax and CLmin are both attracted by local optima and query fewer points around $\boldsymbol{x}^*$. By contrast, PMEI-10 tends to explore the optimization space better. According to the minimum energy criterion, it queries the most points that cluster around $\boldsymbol{x}^*$. As a result, a better estimation of the optimum is provided. Additionally, the points by PMEI appear to be space-filling. It means that our framework balances exploitation and exploration well. Since the solution based on $\hat{G}$ of PMEI is better, our method outperforms the other three methods. Regarding 10^{-5} as the error tolerance again, it takes PMEI about 21 iterations to satisfy the tolerance, while other methods almost fail. For error tolerance 10^{-4}, PMEI spends about 20 iterations while others need at least 30. Particularly, PMEI appears to get similar solutions via 5 iterations while EI needs 40. It suggests that PMEI can improve the one-point method by saving time and finding a better solution. For an objective function involving many local optima, our framework can encourage exploration and avoid the local optimization trap.

For a more complicated feasible space, Figure 2.5 demonstrates the logarithm gap for the Hart6 function. It suggests that PMEI provides an efficient and stable decrease in error with fewer iterations. In Figure 2.5(a), CL and EI seem to keep searching the landscape. Their errors still fluctuate, while the error for PMEI reaches approximately $10^{-0.6}$. From Figure 2.5(b), all methods appear to find points with similar quality except EI. CL queries better points than PMEI at the beginning. However, it may focus more on exploration, leading to a not accurate enough surrogate model. On the other hand, PMEI makes a better trade-off, exploring the feasible region in the early stage and quickly querying points locally. It is worth noting that CL and PMEI perform similarly after 20 iterations, but there are huge fluctuations in the gaps of CL during the first 15 iterations. It may indicate that CL is not efficient if there are limited calculations because it sets a minimal or maximal value of $\boldsymbol{y}_n$ as the virtual response, which is not accurate enough. In other words, CL can provide good results at last. But it seems to need lots of iterations, which is not applicable when the objective function is very expensive to calculate. As a comparison, PMEI finds an estimation of the optimum more efficiently. Considering time costs again, CL and PMEI perform similarly. PMEI appears to reach a smaller error with a step advance than CL. Additionally, it indeed saves lots of steps than EI.

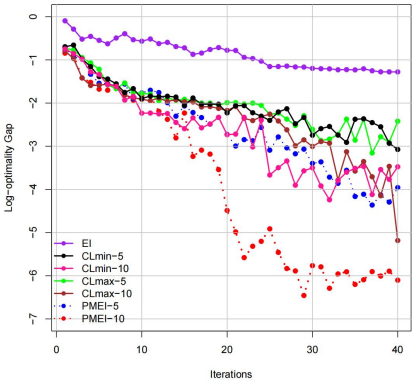
We also compare the PMHEI, PMEI, HEI, and EI with the Griewank and Hart6 benchmark functions. The results are shown in Figure 2.6. There are similar trends in PMEI and PMHEI. The accuracy indeed improves due to the minimum energy criterion. PMEI and PMHEI both provide a better solution for these two functions. Additionally, HEI seems to perform comparably for the Hart6 function. But PMEI and PMHEI converge more



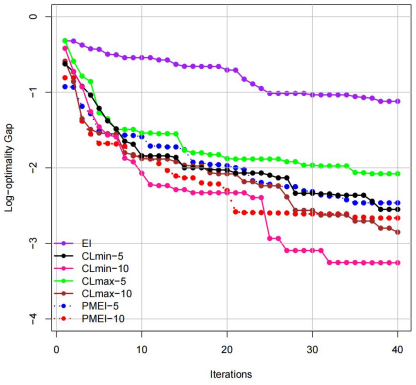
(a) Logarithm optimal gaps based on $\hat{G}$ for the Three-hump function



(b) Logarithm optimal gaps based on $\mathcal{D}_n$ for the Three-hump function

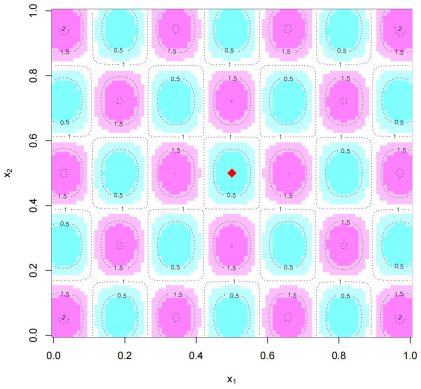


(c) Logarithm optimal gaps based on $\hat{G}$ for the Griewank function

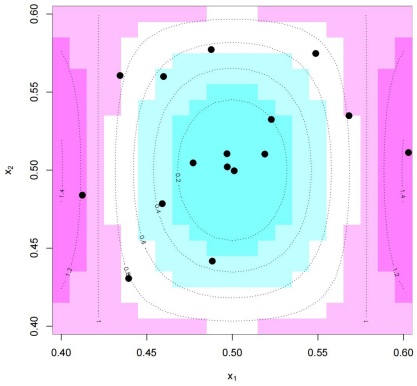


(d) Logarithm optimal gaps based on $\mathcal{D}_n$ for the Griewank function

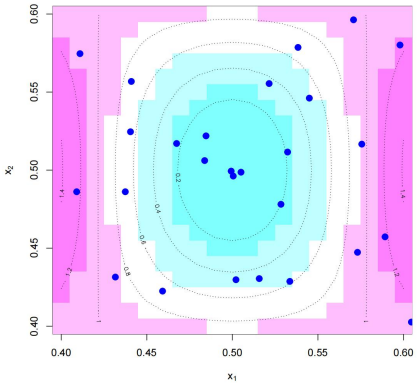
Figure 2.3: Logarithm optimal gaps for 2-D benchmark functions via EI, CLmin, CLmax and PMEI.



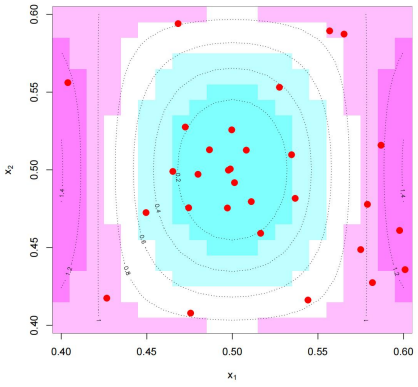
(a) Contour of the Griewank function



(b) Queried points by CLmax-10 around $\boldsymbol{x}^*$

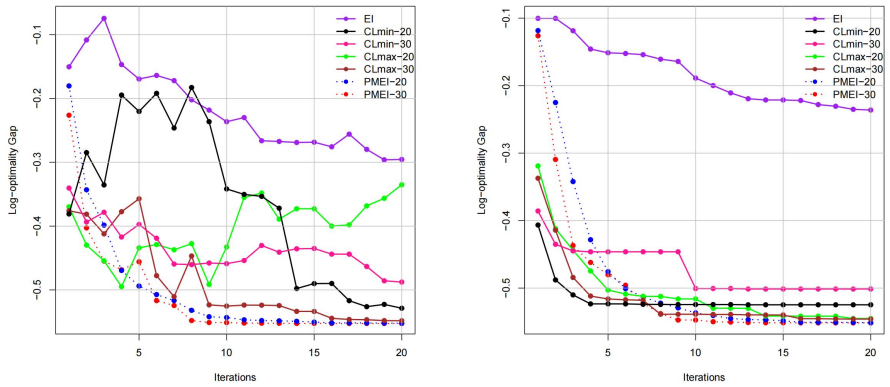


(c) Queried points by CLmin-10 around $\boldsymbol{x}^*$



(d) Queried points by PMEI-10 around $\boldsymbol{x}^*$

Figure 2.4: A visualisation of the queried points on the contour of the Griewank function (red rhombus: the true optimum $\boldsymbol{x}^* = (0.5, 0.5)$, black circles: points via CLmax-10, blue circles: points via CLmin-10, red circles: points via PMEI-10).



(a) Logarithm optimal gaps based on $\hat{G}$ for the Hart6 function

(b) Logarithm optimal gaps based on $\mathcal{D}_n$ for the Hart6 function

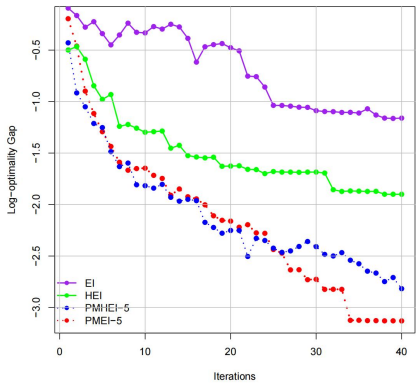
Figure 2.5: Logarithm optimal gaps for high-dimension benchmark functions via EI, CLmin, CLmax and PMEI.

quickly. This suggests that PMEI and PMHEI can find similar solutions with HEI and save 3/4 time. In summary, our proposed framework is suitable for both infill criteria. With a parallel query, it explores the optimization space better and can also reduce the number of iterations.

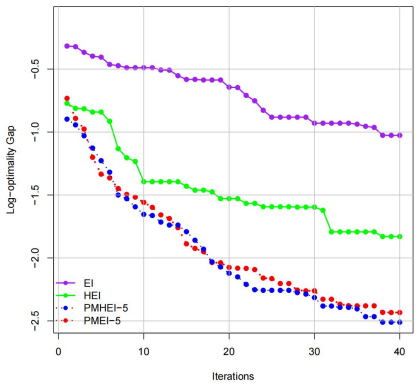
Further, we consider more benchmark functions introduced by [27], with visualisations available at the virtual library. $\log_{10}(|G(\boldsymbol{x}^*) - g_{\min}|)$ is calculated here to evaluate the methods. Most of them have many local minima and only one global optimum. For low-dimensional problems, 40 iterations are implemented to find the optimal solution with $q = 5, 10, 15$. Three kinds of high-dimensional problems are conducted to compare PMEI with CL. A total of 300 points are queried with $q = 10, 10, 15, 15$ for $p = 4, 6, 8, 10$. All results are shown in Section 7.1, suggesting that the proposed PMEI method yields better or comparable results than the others. In the next subsection, the whole framework is applied to deal with a more complicated benchmark application.

2.4.2. A REAL-CASE EXAMPLE

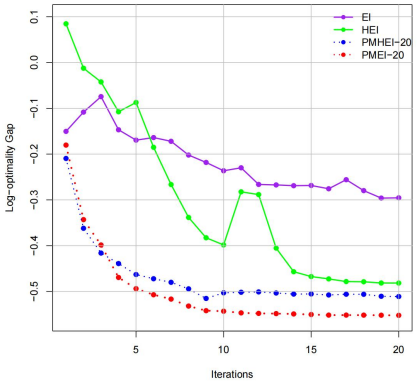
In this subsection, we investigate the performance of PMEI through an optimization problem of wind plant annual energy production (AEP). In wind plant modelling, the main aim is to maximise AEP through layout changes. In this problem, wind turbines can affect other turbines because of their wakes. This interaction lowers the plant's overall energy output on a wind farm. By locating the turbines properly under the most



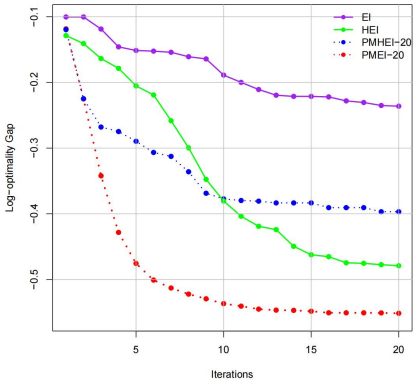
(a) Logarithm optimal gaps based on $\hat{G}$ for the Griewank function



(b) Logarithm optimal gaps based on $\mathcal{D}_n$ for the Griewank function



(c) Logarithm optimal gaps based on $\hat{G}$ for the Hart6 function



(d) Logarithm optimal gaps based on $\mathcal{D}_n$ for the Hart6 function

Figure 2.6: Logarithm optimal gaps via EI, HEI, PMEI and PMHEI.

common wind conditions, the adverse consequences of this aerodynamic interaction can be reduced. The key idea of wind plant wake control was initially introduced in [28]. By coordinating the control operations across the wind turbines, wake losses can be minimised. Also, the performance of the wind plant as a whole can be enhanced. A review of the literature on various wake control strategies can be found in [29]. According to [30], it's a proper benchmark to show how our framework holds up against a more complicated and time-consuming problem.

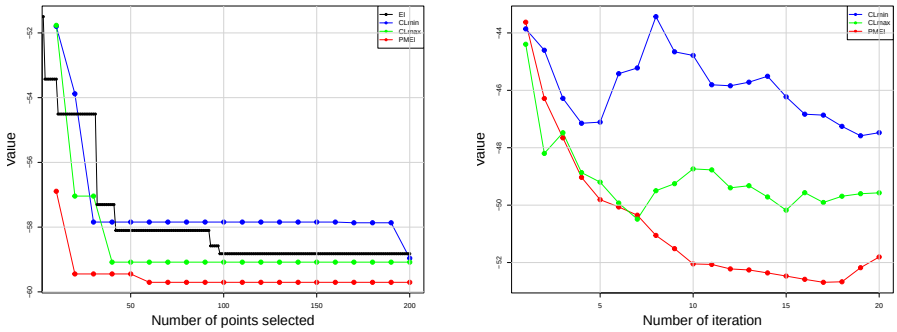
We use a wake simulator called FLORIS to acquire AEP. Given the wind farm layout, it outputs AEP based on some initial wind conditions over a year. Two important factors are wind speed and wind direction. During a whole year, the simulator assumes that the direction condition is a vector containing kinds of angles from 0 to 360. For the speed, there is a preset value based on measurements. Because of the measurement error, it is introduced by randomly generating from a normal distribution. As noted by [30], it looks at productions in different combinations of wind directions and speeds. The final AEP is outputted as the average power over them. More combinations mean that the whole simulation process will be more expensive to run. In addition, the cost depends on the number of wind turbines, taking more time to search best locations for more turbines. For simplification, we try to find the best locations for two turbines. We implement the Python script via `R`, and experiments take about 2 minutes per run. The final objective problem is to minimize

$$(s_{11}^*, s_{12}^*, s_{21}^*, s_{22}^*) = \underset{s_{ij}\in[0,1], i,j=1,2}{\arg\min} -\text{AEP}(s_{11}, s_{12}, s_{21}, s_{22}),$$

where (s_{i1}, s_{i2}) is the location of one turbine. It is worth noting that there are some constraints in this problem. Similar to [30], we set the response to be 0 if two turbines are too close to each other. We begin with a space-filling design with 30 samples and compare PMEI with CL and EI. Totally 200 points are queried during 20 iterations.

The application results based on the same initial design are shown in Figure 2.7. We are interested in querying points by different methods, so here we do not consider initial points. In Figure 2.7(a), the best points found by CL and PMEI are comparable. Furthermore, it suggests that PMEI is more efficient and can find a better solution via fewer iterations. Regarding the time costs, EI needs about 100 iterations and takes about 3.5 hours to converge and find an optimal solution. By contrast, PMEI needs about 2 iterations to find a similar one and spends about 5 minutes. This means that PMEI can save about 3 hours more than EI by reducing iterations. Meanwhile, PMEI finally provides a better solution, which suggests that PMEI can improve EI to some extent. The right panel shows the mean responses of queried points using CL and PMEI as the iteration progresses. In the beginning, three algorithms all select points with comparable qualities. Then, there is a peak at the eighth step

of CLmin, which suggests that CLmin queries some points with higher values. It seems that CLmin is attracted by a local minimum while the other two are still searching the region. By setting the virtual value as the maximum of calculated responses, CLmax has more potential to encourage exploration. PMEI tends to continue querying better points when CLmax begins to stagnate. It suggests that PMEI can find some better points than other methods.



(a) Minimal responses of queried points by EI, CLmin, CLmax and PMEI.

(b) Mean responses of queried points by CLmin, CLmax and PMEI.

Figure 2.7: Queried results by EI, CLmin, CLmax and PMEI for windmill optimization problems.

2.5. CONCLUSIONS

In this chapter, we proposed a parallel Bayesian framework using the minimum energy criterion for the optimization of an expensive black-box objective function. It improves existing one-point acquisition criteria, such as EI and HEI, in terms of time-saving, parallel simulation, and exploration-encouraging. This novel framework is flexible for various one-point infill criteria and also preserves their desirable properties. Regarding these criteria as charge functions, we proposed the efficient global optimization methods—PMEI and PMHEI. These methods can balance the exploitation and the exploration well. Also, they can enhance efficiency by generating more points of high quality at one time. To place more queries in the local area around the global optima, we tried to apply the information provided by the surrogate model. Then we presented a shrinkage-augment strategy that could overcome the unstable surrogate model problem during the optimization. We measured the progress resulting from the strategy to make sure the size of the local area centred at the best solution ever found was suitable. Two

two-stage frameworks, PMEI and PMHEI, were finally established. Some benchmark test functions were presented to evaluate their performance. The numerical results suggested the superiority of our frameworks over some existing Bayesian optimization methods, such as EI, HEI, CL-min, and CL-max. Considering the acquisition function values and space-filling property, our methods presented an efficient and stable decrease of gaps with fewer iterations. An application about windmills was also presented. In practice, the proposed method found a better solution via fewer steps. It can save more time and costs than one-point methods. In the future, the global convergence rates of our methods should be studied.

3

SURROGATE MODELING FOR SPATIOTEMPORAL FLOWS VIA DECOMPOSITION

3.1. INTRODUCTION

In industrial applications, various complex systems and physical phenomena, including aircraft-UAV collisions [31] and reservoir fluid dynamics [2], are extensively studied to understand their dynamic behaviour and inform scientific decisions. For these systems, conducting traditional physical experiments is challenging because of constraints related to environmental, safety, and economic factors. As an alternative, a series of partial differential equations (PDEs), which serve as the underlying governing equations, are often utilized as simulators. Computing high-fidelity spatiotemporal solutions for different simulator parameters is computationally demanding, as it depends on capturing the complex flow field information spanning a broad range of length and time scales. Furthermore, exploring this simulator parameter space (which typically includes physical properties, geometric configuration, initial conditions, and boundary conditions [32]) requires numerous simulations, amplifying this computational burden. For example, the large-eddy simulation (LES) technique requires approximately 500,000 CPU hours on a grid comprising 4 million mesh points [32].

In the literature, researchers have proposed various surrogate models as a more efficient approach to address these issues, including the radial basis functions [33], neural network [34], splines [35], and Gaussian process (GP) [36, 37]. Such methods often struggle with high-dimensional fluid dynamics problems due to unaffordable computational costs. Some studies have employed local approximation techniques [38, 39] to handle large-scale datasets. These methods construct local Gaussian process (GP) models that select relevant data points near unobserved inputs. However, this approach presents two key limitations for spatiotemporal flow prediction. First, fitting numerous independent GPs for each unique combination of spatial coordinates and time points incurs prohibitive computational costs, especially in design exploration tasks [40]. Second, local approximations fail to leverage global spatiotemporal correlations across length and time scales, which remains a critical challenge for dynamic systems. [41] proposed a multi-output surrogate model relying on the eigendecomposition of the covariance matrix to handle spatiotemporal data. However, this approach becomes computationally infeasible for problems with a large number of observations, rendering it unsuitable for our case.

Reduced-order models (ROMs) are a widely used approach for handling spatiotemporal data by projecting the full-order system onto a reduced space. This enables the reconstruction of high-dimensional solutions through linear combinations of basis functions. The relationship between projection coefficients and input parameters is then approximated, allowing the prediction of unobserved flows for new parameters via interpolation or regression. ROMs can be categorized as intrusive or non-intrusive. Intrusive ROMs rely on access to the underlying

PDEs to enforce physical constraints and preserve dynamic structures [42]. Examples include projection-based reduction [43, 44], empirical interpolation [45], and subspace prediction [46]. However, they require intrusive modifications to the simulator code and significant expertise, making them impractical when governing equations are unavailable or inaccessible. Non-intrusive ROMs, driven by data availability and the frequent absence of known governing equations, construct reduced bases directly from simulation data. Representative techniques include the eigensystem realization algorithm [47], dynamic mode decomposition [48], and Koopman operator methods [49], though these are often limited to linear systems. In addition, some studies have attempted to establish an optimal subspace by the tensor compression, as described in [50]. However, such an approximation does not always exist [51], and determining the rank is also a significant challenge [52].

Proper Orthogonal Decomposition (POD) is one of the most commonly used methods for extracting basis functions and building surrogate models [53]. For instance, [32, 54] proposed the common POD (CPOD) for the LES of the injector. They extracted spatial information using POD modes and modelled the corresponding parameter-dependent time-varying coefficients with GPs. [55] and [56] enhanced CPOD using kriging-based weight functions and the Hadamard product. These methods rely on the assumption of independence between time steps, leading to plenty of GP models to fit and encountering challenges in predicting flows accurately. Other POD variants [57–62] often analyze flows for different simulator inputs separately, hindering the identification of a unified low-rank subspace and robust quantification of projection errors.

Machine learning methods, such as self-attention mechanisms with autoencoders [63] and recurrent neural networks with stochastic processes [64], show promise for spatiotemporal forecasting. Nevertheless, they face persistent training challenges like vanishing gradients and convergence degradation [65], limiting their efficiency for tasks requiring rapid predictions or uncertainty quantification. Additionally, Physics-Informed Neural Networks (PINNs) [66] have also been proposed to solve the PDEs, which incorporate physical principles into surrogate models. But such methods remain computationally inefficient for applications requiring repeated training (e.g., optimal control, inverse problems, uncertainty quantification, and design optimization).

Despite these advances, there remains a critical need for computationally efficient statistical emulators that accurately approximate spatiotemporal flows in the absence of expert knowledge about underlying PDEs. To address this gap, we propose a multi-output surrogate model designed for rapid and reliable prediction. Our key contributions are: (1) We propose a tailored multi-output Gaussian process (MOGP) for expensive spatiotemporal simulators, reformulated

via a non-intrusive ROM approach using Nested Decomposition (ND). This leverages two-stage Singular Value Decomposition (SVD) to project data into a low-dimensional latent space. (2) The proposed model operates without knowledge of governing PDEs, making it suitable for diverse dynamics. (3) Unlike existing methods, the new method simultaneously captures correlations among spatial coordinates, temporal steps, and simulator parameters, improving prediction reliability. (4) Our method avoids fitting numerous GPs, significantly reducing model training time and making it ideal for design exploration (e.g., sequential design). (5) We establish bounds for ND errors and provide comprehensive uncertainty quantification for flow predictions.

The rest of this chapter is organized as follows. In Section 3.2, we propose a tailored surrogate model, defined as the MOGP via the ROM based on ND (NDROM), for expensive spatiotemporal simulators. We also discuss the compression error of NDROM and the uncertainty quantification of the predicted flow. In Section 3.3, we conduct several numerical experiments to evaluate the performance of the proposed model. Section 3.4 further presents the applicability of our model using two real examples. Finally, Section 3.5 concludes this study with a summary and directions for future work. All technical details and additional simulation results are provided in the supplementary material.

3.2. A NDROM-BASED SURROGATE MODEL

In this section, we introduce a new multi-output surrogate model based on nested SVD. Let $w(\cdot)$ represent the complex and expensive PDE-based simulator, whose input is a p-dimensional vector $\boldsymbol{u} \in \Omega$ and output is a spatiotemporal flow. Let $\mathcal{S}_n = \{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n\}$ represent the set of design points at which high-fidelity solutions will be obtained using finite element methods. Additionally, let $\{x_i\}_{i=1}^{n_x} \subset \mathcal{X}$ represent n_x spatial coordinates, and $\{t_j\}_{j=1}^{n_t} \subset \mathcal{T}$ represent n_t time points. The response of the expensive simulator at $\boldsymbol{u}_k$ is $\mathbf{Y}_k = \left[w(x_i, t_j; \boldsymbol{u}_k)\right]_{1\le i\le n_x,\ 1\le j\le n_t}$, $k = 1, \cdots, n$. By combining the outputs from all design points over parameter space, we can obtain a tensor dataset denoted as $\mathbb{Y} \in \mathbb{R}^{n\times n_x\times n_t}$.

As mentioned before, such high-fidelity solutions require plenty of time and computational resources. GPs serve as surrogate models, providing the best linear unbiased estimators (BLUE) and enabling uncertainty quantification through appropriate posterior distributions. While applying independent GPs per spatial-temporal point (x, t) with input $\boldsymbol{u}$ is natural [67, 68], it becomes computationally prohibitive for large n_x and n_t. Alternatively, the MOGP leverages inherent spatiotemporal correlations [69]. It applies multivariate normal distributions to account for the correlations among vector-valued outputs of Gaussian processes. However, under this framework, spatiotemporal data across all locations

and time points are treated as vector-valued outputs, leading to an extremely large correlation matrix, which is computationally infeasible for our study.

In this work, we model the spatiotemporal flows as

$$\mathbf{Y}(\boldsymbol{u}) = \begin{bmatrix} G_{1,1}(x_1, t_1, \boldsymbol{u}) & \cdots & G_{1,n_t}(x_1, t_{n_t}, \boldsymbol{u}) \\ \vdots & \ddots & \vdots \\ G_{n_x,1}(x_{n_x}, t_1, \boldsymbol{u}) & \cdots & G_{n_x,n_t}(x_{n_x}, t_{n_t}, \boldsymbol{u}) \end{bmatrix}_{n_x \times n_t}, \tag{3.1}$$

where $G_{i,j}$ are GPs, $i = 1, \cdots, n_x$, $j = 1, \cdots, n_t$. Then, inspired by the work of [70] and [71], we aim to reformulate model (3.1) using fewer independent GPs, i.e., $\mathbf{Y}(\boldsymbol{u}) = \mathbf{A}\mathbf{N}(\boldsymbol{u})$, where $\mathbf{A}$ is a low-rank transformation matrix, and $\mathbf{N}(\boldsymbol{u})$ is a GP matrix that incorporates a reduced set of independent GPs to maintain computational feasibility.

3.2.1. A NDROM-BASED MOGP

To find $\mathbf{A}$ and the corresponding GP matrix $\mathbf{N}(\boldsymbol{u})$, we begin by translating the tensor $\mathcal{Y}$ into a matrix using a tensor unfolding tool. A general definition of unfolding can be found in [72]. More specifically, the tensor is transformed by concatenating all $\mathbf{Y}_i$ column-wise into a unified matrix, which is denoted as $\mathbb{Y} = \left[\mathbf{Y}_1 | \cdots | \mathbf{Y}_n\right]_{n_x \times n_t n}$. The unfolding strategy depends on the relative sizes of n_x, n_t, and n. If $\max\{n_x, n\} < n_t$, row-wise concatenation improves efficiency. This study assumes $\max\{n_t, n\} < n_x$.

We perform singular value decomposition (SVD) on $\mathbb{Y}$ for reformulating model (3.1),

$$\mathbb{Y} = \mathbf{U}_{n_x \times n_x} \begin{bmatrix} \boldsymbol{\Sigma}_{R_1 \times R_1} & \mathbf{0}_{R_1 \times (n_t n - R_1)} \\ \mathbf{0}_{(n_x - R_1) \times R_1} & \mathbf{0}_{(n_x - R_1) \times (n_t n - R_1)} \end{bmatrix}_{n_x \times n_t n} \mathbf{V}^{\top}_{n_t n \times n_t n},$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\boldsymbol{\Sigma} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{R_1})$ contains singular values $\lambda_1 \geq \cdots \geq \lambda_{R_1} > 0$. For dimensionality reduction, we retain only $r_1 \ll R_1$ singular values, approximating $\mathbb{Y}$ as $\tilde{\mathbb{Y}} = \mathbf{U}_{n_x \times r_1} \boldsymbol{\Sigma}_{r_1 \times r_1} \mathbf{V}^{\top}_{r_1 \times n_t n}$. By the Schmidt-Eckart-Young theorem [73], this minimizes the projection error

$$\|\mathbb{Y} - \tilde{\mathbb{Y}}\|_F^2 = \min_{W \in \mathcal{W}} \|\mathbb{Y} - WW^{\top}\mathbb{Y}\|_F^2 = \sum_{i=r_1+1}^{R_1} \lambda_i^2, \tag{3.2}$$

with $\mathcal{W} = \{W \in \mathbb{R}^{n_x \times r_1} : WW^{\top} = \mathbf{I}\}$. The rank r_1 is selected via a threshold ε_1 on the retained energy ratio $\sum_{i=1}^{r_1} \lambda_i / \sum_{i=1}^{R_1} \lambda_i$. Let $\boldsymbol{\psi}_i = \mathbf{U}_i$ be the basis vectors, then each column $\mathbf{Y}_i^j$ of $\mathbf{Y}_i$ can be approximated as a linear combination. Specifically, $\mathbf{Y}_i^j \approx \sum_{k=1}^{r_1} a_k(t_j, \boldsymbol{u}_i)\boldsymbol{\psi}_k$, $i = 1, \ldots, n, j =$

$1, \ldots, n_t$, where $\boldsymbol{a}_k(t_i, \boldsymbol{u}_j) = [a_k(t_1, \boldsymbol{u}_1), \cdots, a_k(t_{n_t}, \boldsymbol{u}_n)]^\top$ is the ith column of $\boldsymbol{\Sigma}_{r_1 \times r_1} \mathbf{V}^\top_{r_1 \times n_t n}$. Note that $\boldsymbol{\Sigma}$ indicates the inherent structure of the data via singular values, not depending on locations and time points. The MOGP model can then be reformulated as

$$\mathbf{Y} = \mathbf{A} \begin{bmatrix} g_{1,1}(t_1, \boldsymbol{u}) & \cdots & g_{1,n_t}(t_{n_t}, \boldsymbol{u}) \\ \vdots & \ddots & \vdots \\ g_{r_1,1}(t_1, \boldsymbol{u}) & \cdots & g_{r_1,n_t}(t_{n_t}, \boldsymbol{u}) \end{bmatrix}, \quad \mathbf{A} = \mathbf{U}_{n_x \times r_1} \boldsymbol{\Sigma}_{r_1 \times r_1}, \tag{3.3}$$

where $g_{i,j}$ are GPs. However, n_t may still remain very large in practical applications; for example, $n_t = 6500$ in the simulation of flow past an infinite rectangular prism [74]. While Model (3.3) provides a reduced representation, it still requires an excessive number of GPs. Following the same approach used to determine $\mathbf{A}$, we reorganize each column of $\mathbf{V}$ into matrix-valued outputs and apply SVD again to achieve further simplification of the GP matrix.

Thus, the ith column of $\mathbf{V}_{n_t n \times r_1}$ is rearranged into the matrix by aligning elements according to subscripts, i.e.,

$$\mathbb{B}^i = \begin{bmatrix} V_i(t_1, \boldsymbol{u}_1) & V_i(t_1, \boldsymbol{u}_2) & \cdots & V_i(t_1, \boldsymbol{u}_n) \\ \vdots & \vdots & \cdots & \vdots \\ V_i(t_{n_t}, \boldsymbol{u}_1) & V_i(t_{n_t}, \boldsymbol{u}_2) & \cdots & V_i(t_{n_t}, \boldsymbol{u}_n) \end{bmatrix}_{n_t \times n},$$

for $i = 1, \cdots, r_1$. SVD is then applied for each $\mathbb{B}^i_{n_t \times n}$ as

$$\mathbb{B}^i_{n_t \times n} \approx \mathbf{P}^i_{n_t \times r_2^i} \mathbf{D}^i_{r_2^i \times r_2^i} \mathbf{Q}^{i^\top}_{r_2^i \times n}, \ i = 1, \cdots, r_1.$$

Each number of modes during the second step, r_2^i, is selected similarly to r_1 with a given ε_2^i. As mentioned before, the Schmidt-Eckart-Young theorem ensures that the columns of $\mathbf{P}^i_{n_x \times r_2^i}$ can define a subspace that achieves the minimal approximation error for $\mathbb{B}^i_{n_t \times n}$. The columns of $\mathbf{Q}^i_{n \times r_2^i}$ represent the projection coefficients that depend on $\boldsymbol{u}$.

Leveraging tensor unfolding and nested decomposition with the SVD tool, we extract the $\mathbf{U}$ and $\mathbf{P}^i$, obtaining parameter-dependent projection coefficients $\mathbf{Q}^i$ as well. Finally, we establish an upper bound for the proposed ND-based ROM with the following proposition.

Proposition 1. *According to* (3.2)*, the decomposition error of the nested SVD framework is not greater than*

$$\sum_{i=r_1+1}^{R_1} \lambda_i^2 + \sum_{i=1}^{r_1} \sum_{j=r_2^i+1}^{R_2} \lambda_i^2 (\lambda_j^i)^2 U^m_{:,i},$$

where $U^m_{\cdot,i}$ is the maximal value of ith column of $\mathbf{U}$, and λ^i_j is the singular values generated during the decomposition of $\mathbb{B}^i$. Therefore, the accuracy of the proposed ND-based ROM framework can be controlled.

Finally, the surrogate model for the spatiotemporal simulator can be formulated as

$$\begin{aligned}\mathbf{Y}(\boldsymbol{u})_{n_x\times n_t} &= \mathbf{U}_{n_x\times r_1}\Sigma_{r_1\times r_1}\mathbf{N}(\boldsymbol{u}),\\ \mathbf{N}(\boldsymbol{u}) &= \begin{bmatrix} \left[\mathbf{P}^1_{n_t\times r_2^1}\mathbf{D}^1_{r_2^1\times r_2^1}\tilde{\boldsymbol{Q}}(\boldsymbol{u})^{1^\top}_{r_2^1\times 1}\right]^\top \\ \vdots \\ \left[\mathbf{P}^{r_1}_{n_t\times r_2^{r_1}}\mathbf{D}^1_{r_2^{r_1}\times r_2^{r_1}}\tilde{\boldsymbol{Q}}(\boldsymbol{u})^{r_1^\top}_{r_2^{r_1}\times 1}\right]^\top \end{bmatrix}_{r_1\times n_t},\end{aligned} \tag{3.4}$$

where $\tilde{\boldsymbol{Q}}$s are assumed as independent GPs. Note that training independent GPs in this study, instead of the MOGP, reduces computation time with minimal accuracy loss, as widely adopted [32, 55, 62, 75].

3.2.2. PREDICT THE PROJECTION PARAMETERS BASED ON GP

The relationships between the projection coefficients and simulator parameters need to be fitted. Let $\boldsymbol{q}^{i,j}$ be the jth column of $\mathbf{Q}^i_{n\times r_2^i}$, assume that

$$q^{i,j}(\boldsymbol{u}) = \mu^{i,j} + Z^{i,j}(\boldsymbol{u}), i = 1,\cdots,r_1, j = 1,\cdots,r_2^i,$$

where $\mu^{i,j}$ is the unknown mean parameter, $Z^{i,j}(\boldsymbol{u}) \sim GP(\mathbf{0}, \sigma^2_{i,j}K^{i,j}(\cdot,\cdot))$ is a zero-mean Gaussian process with process variance $\sigma^2_{i,j}$ and the correlation function $K^{i,j}(\cdot,\cdot)$. Specifically, $K(\boldsymbol{u}_i,\boldsymbol{u}_j) = C(H_{i,j})$ with some unknown parameters, where $H_{i,j} = ||\boldsymbol{u}_i - \boldsymbol{u}_j||_2$ with the Euclidean norm $||\cdot||_2$. There are several families of commonly used correlation functions, such as the Radial Basis Function (RBF) family and Matérn family.

According to [3], under a Gaussian prior, the posterior distribution of $q^{i,j}(\boldsymbol{u})$ at an unobserved input $\boldsymbol{u}$ can be derived as $[q^{i,j}(\boldsymbol{u})|\mathcal{D}_n] \sim N\left(g^{i,j}(\boldsymbol{u}), s^2_{i,j}(\boldsymbol{u})\right)$, where $\mathcal{D}_n = \{(\boldsymbol{u}^{i,j}_k, q^{i,j}_k)\}^n_{k=1}$, $g^{i,j}(\boldsymbol{u}) = \mu^{i,j} + \boldsymbol{K}^{i,j}(\boldsymbol{u})^\top\mathbf{R}^{ij^{-1}}(\boldsymbol{q}^{i,j} - \mu^{i,j}\mathbf{1})$, $\boldsymbol{q}^{i,j} = [q^{i,j}_1\cdots,q^{i,j}_n]^\top$ is the vector of responses, $\boldsymbol{K}^{i,j}(\boldsymbol{u}) = (K^{i,j}(\boldsymbol{u},\boldsymbol{u}_k))^n_{k=1}$ is the correlation vector between the prediction and $\boldsymbol{q}^{i,j}$, $\mathbf{R}^{i,j} = (K^{i,j}(\boldsymbol{u}_k,\boldsymbol{u}_l))^n_{i,j=1}$ is the correlation matrix of the observed points, $\mathbf{1}$ is the n-dimensional column vector of ones and $s^2_{i,j}(\boldsymbol{u})$ is corresponding posterior variance,

$$s^2_{i,j}(\boldsymbol{u}) = \sigma^2_{i,j}\left[1 - \boldsymbol{K}^{i,j}(\boldsymbol{u})^\top\mathbf{R}^{ij^{-1}}\boldsymbol{K}^{i,j}(\boldsymbol{u}) + \frac{(1-\mathbf{1}^\top\mathbf{R}^{ij^{-1}}\boldsymbol{K}^{i,j}(\boldsymbol{u}))^2}{\mathbf{1}^\top\mathbf{R}^{ij^{-1}}\mathbf{1}}\right]. \tag{3.5}$$

For GP models, there are some unknown parameters to estimate, including $\mu^{i,j}$, $\sigma^2_{i,j}$, and the hyperparameters associated with the correlation function $K^{i,j}(\cdot,\cdot)$. Such parameters can be estimated by likelihood-based or Bayesian approaches (see [3] and [36] for more details). One can then plug in these estimates to obtain the posterior mean $\hat{g}^{i,j}(\boldsymbol{u})$ as the prediction $\hat{q}^{i,j}(\boldsymbol{u})$, along with the prediction variance $\hat{s}^2_{i,j}(\boldsymbol{u})$.

For every $i = 1, \cdots, r_1$, the parameter-dependent coefficients can be predicted as $\widehat{\boldsymbol{Q}}^i_{1\times r_2^i} = [\hat{q}^{i1}(\boldsymbol{u}), \cdots, \hat{q}^{ir_2^i}(\boldsymbol{u})]$, and then we have $\widehat{\mathbb{B}}^i_{n_t\times 1} = \mathbf{P}^i_{n_t\times r_2^i}\mathbf{D}^i_{r_2^i\times r_2^i}\widehat{\boldsymbol{Q}}^{i\top}_{r_2^i\times 1}$. Consequently, the discrete solution at a new point $\boldsymbol{u}$ can be approximated as $\widehat{\mathbb{Y}}_{n_x\times n_t}(\boldsymbol{u}) = \mathbf{U}_{n_x\times r_1}\boldsymbol{\Sigma}_{r_1\times r_1}\widehat{\mathbf{V}}^\top_{r_1\times n_t}$, where $\widehat{\mathbf{V}}_{n_t\times r_1} = [\widehat{\mathbb{B}}^1|\cdots|\widehat{\mathbb{B}}^{r_1}]$. The corresponding 95% credible interval is derived in Proposition 2.

Proposition 2. *Given a Gaussian prior and estimated parameters, the prediction at the kth location and the lth snapshot $\hat{y}_{kl}$ at a new input $\boldsymbol{u}$ follows a Gaussian distribution. Specifically, $[y_{k,l}|\mathcal{D}_n] \sim N\left(\hat{y}_{k,l}(\boldsymbol{u}), \tilde{s}^2_{k,l}(\boldsymbol{u})\right)$, where $\hat{y}_{k,l}(\boldsymbol{u}) = \mathbf{U}_{k,\cdot} \times \boldsymbol{\Sigma} \times (\widehat{\mathbf{V}}_{l,\cdot})^\top$ and*

$$\tilde{s}_{k,l}(\boldsymbol{u}) = \sqrt{\sum_{i=1}^{r_1}\sum_{j=1}^{r_2^i}\hat{s}^2_{i,j}(\boldsymbol{u})(U_{k,i}\Sigma_{i,i}P^i_{l,j}D^i_{j,j})^2},$$

with $\hat{s}_{i,j}(\boldsymbol{u})$, the estimated posterior standard variance based on (3.5). Accordingly, the 95% confidence interval is $[\hat{y}_{k,l}(\boldsymbol{u}) - 1.96\tilde{s}_{k,l}(\boldsymbol{u}), \hat{y}_{k,l}(\boldsymbol{u}) + 1.96\tilde{s}_{k,l}(\boldsymbol{u})]$.

Remark 1. *$\widehat{\mathbf{V}}$ captures parameter-dependent dynamics. Combined with $\mathbf{U}$ and $\mathbf{P}^i$, the MOGP (3.4) reconstructs solutions at new points via basis functions, quantifying $\boldsymbol{u}$'s effects on spatial and temporal flow patterns. During SVD, $\mathbf{U}$, $\mathbf{P}^i$, and $\mathbf{Q}^i$ are orthogonal, ensuring small r_1 and r_2^i for simplicity, efficiency, and intuitiveness. However, GP predictions as expansion coefficients need not be orthogonal.*

The whole framework based on the proposed MOGP, denoted as D-SVD, is presented in Algorithm 2, which includes tensor unfolding, ND-based decomposition, and multi-output surrogate modelling. A total of $\sum_{i=1}^{r_1} r_2^i$ independent GP models are trained using parameter-dependent expansion coefficients as responses. For new simulator parameters, high-fidelity solutions are recovered by multiplying basis functions with predicted projection coefficients.

Algorithm 2: Proposed framework for spatiotemporal predictions (D-SVD).

Input: $w(\cdot)$: p-dimensional spatiotemporal computer model; Ω^p: design space; $\boldsymbol{u}$: a new design point.

Output: Predicted flows $\hat{\mathbf{Y}}$ at $\boldsymbol{u}$.

$n \leftarrow 10p$;

Generate n space-filling design points over Ω^p as initial inputs $\mathcal{S}_n = \{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n\}$;

Run $w(\cdot)$ at these inputs, collecting spatiotemporal matrices as $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$;

$n_t \leftarrow$ the number of snapshots when outputs are simulated;

$n_x \leftarrow$ the number of locations where outputs are simulated;

$\mathbb{Y} \leftarrow [\mathbf{Y}_1 | \cdots | \mathbf{Y}_n]$;

$r_1 \leftarrow$ minimal number of modes with given threshold ε_1;

$\mathbb{Y} \approx \mathbf{U}_{n_x \times r_1} \mathbf{\Sigma}_{r_1 \times r_1} \mathbf{V}^{\top}_{r_1 \times n n_t}$ via SVD;

for $i \leftarrow 1$ **to** r_1 **do**

  $\boldsymbol{V}^i_{n n_t \times 1} \leftarrow$ the ith column of $\mathbf{V}_{n n_t \times r_1}$;

  $\mathbb{B}^i_{n_t \times n} \leftarrow$ matrix via transformation of such vector with the alignment criteria;

  $r_2^i \leftarrow$ minimal number of modes with given ε_2;

  $\mathbb{B}^i_{n_t \times n} \approx \mathbf{P}^i_{n_t \times r_2^i} \mathbf{D}^i_{r_2^i \times r_2^i} \mathbf{Q}^{i^{\top}}_{r_2^i \times n}$ by SVD;

  for $j \leftarrow 1$ **to** r_2^i **do**

    $\boldsymbol{q}^{i,j} \leftarrow$ the jth column of $\mathbf{Q}^i_{n \times r_2^i}$;

    $\hat{g}^{i,j}(\cdot) \leftarrow$ a Gaussian process model created with $\mathcal{S}_n$ and $\boldsymbol{q}^{i,j}$;

    $\hat{q}^{i,j} = \hat{g}^{i,j}(\boldsymbol{u})$;

  end

  $\hat{\boldsymbol{Q}}^i_{1 \times r_2^i} \leftarrow [\hat{q}^{i,1}, \cdots, \hat{q}^{i,r_2^i}]$;

  $\hat{\mathbb{B}}^i_{n_t \times 1} \leftarrow \mathbf{P}^i_{n_t \times r_2^i} \mathbf{D}^i_{r_2^i \times r_2^i} \hat{\boldsymbol{Q}}^{i^{\top}}_{r_2^i \times 1}$;

end

$\hat{\mathbf{V}}_{n_t \times r_1} \leftarrow [\hat{\mathbb{B}}^1 | \cdots | \hat{\mathbb{B}}^{r_1}]$;

$\hat{\mathbf{Y}} = \mathbf{U}_{n_x \times r_1} \mathbf{\Sigma}_{r_1 \times r_1} \hat{\mathbf{V}}^{\top}_{r_1 \times n_t}$.

3.3. NUMERICAL EXPERIMENTS

In this section, several numerical experiments are conducted to evaluate the performance of the proposed model. To evaluate our model's accuracy and efficiency, we compare it with existing order-reduced methods: IGP, LaGP [38], CPOD [32], KPOD [55], and TPOD [76]. Thresholds are set as follows: for DSVD, $\epsilon_1 = 0.99$ and $\epsilon_2^i = \epsilon_2 = 0.999$; for CPOD and KPOD, the threshold is 0.99; and for TPOD, thresholds are 0.99, 0.999, and 0.9999.

Remark 2. *Given that the prior knowledge of the underlying PDEs is unknown, intrusive methods are excluded. As noted in Section 3.1, machine learning methods are omitted due to inefficiency. Additionally, the approach in [41] fails for extremely high-dimensional outputs (e.g., 80,000 for the Burgers equation), while [39] addresses time series responses rather than spatiotemporal flows, rendering SVD implementation infeasible. Thus, both methods are excluded.*

IGP builds independent GPs per location-time combination instead of MOGP, following [67]. LaGP fits local GPs for spatiotemporal prediction, training on subsets of $(\boldsymbol{u}, x, t)$ *data. Among POD-based ROMs, we compare with TPOD, which applies POD to each* $\mathbf{Y}_i$*, decomposes bases twice for subspace identification, and models coefficients via a third POD. The method in [56] needs extra modes due to non-orthogonal bases, so only CPOD and KPOD are retained. CPOD derives common bases and models coefficients with GPs, while KPOD performs POD on each* $\mathbf{Y}_i$ *and predicts basis functions and coefficients via GPs together. Both CPOD and KPOD assume temporal independence; see [32] and [55] for more details.*

The performance of various methods is evaluated using two metrics: prediction accuracy and computational cost. Prediction accuracy is assessed via the average mean relative error (AMRE) and proper scoring rule (PSR) $S(\mathbf{u})$ [39]. Specifically,

$$\mathrm{AMRE}(t) = \frac{1}{n_t n_x} \sum_{i=1}^{n_x} \sum_{j=1}^{n_t} \frac{|w(x_i, t_j; \boldsymbol{u}) - \hat{w}(x_i, t_j; \boldsymbol{u})|}{|w(x_i, t_j; \boldsymbol{u})|} \times 100\%$$

and

$$S(\boldsymbol{u}) = \frac{1}{n_x n_t} \sum_{i=1}^{n_x} \sum_{j=1}^{n_t} \frac{(w(x_i, t_j; \boldsymbol{u}) - \hat{w}(x_i, t_j; \boldsymbol{u}))^2}{\tilde{s}_{i,j}^2(\boldsymbol{u})} + \frac{1}{n_x n_t} \sum_{i=1}^{n_x} \sum_{j=1}^{n_t} \log \tilde{s}_{i,j}^2(\boldsymbol{u}),$$

where lower values indicate superior performance. Note that PSR assumes normally distributed predictions and is thus unsuitable for KPOD.

For the computational cost, the total time for the entire procedure, including decomposition and GP model training, is recorded, and the number of GPs trained is also tracked. Two types of PDEs are considered

in the numerical experiments: the Burgers equation and the nonlinear Schrödinger equation. Additional benchmarking PDEs and simulation results are presented in Chapter 7.2. All implementations and analyses are conducted in the R environment [77].

3.3.1. THE BURGERS EQUATION

Firstly, we consider the Burgers equation

$$\frac{\partial w}{\partial t} + w\frac{\partial w}{\partial x} = 0,$$
$$w(t=0,x) = A\sin(\pi x),\ \ w(t,x=0) = w(t,x=2),$$

where $t \in [0,2]$, $x \in [0,2]$ and $A \in [0,1]$. By combining nonlinear wave motion with linear diffusion, the model analyzes the combined effect of nonlinear advection and diffusion. Parameter A controls the amplitude at the initial time. There is only one input in this benchmark problem, which simplifies the predictions. We choose 10 space-filling design points, $(0.1, \cdots, 1)$, for training with $n_t = 200$ and $n_x = 400$. Flows with another 100 inputs are used for testing.

Taking the spatial extraction as an illustration, the approximation results of the ten training data sets and the spatial modes based on SVD are shown in Figure 3.1. This indicates that, with only four modes, SVD provides reliable approximation surfaces for the training data, resulting in relatively small AMREs. As seen, columns of **U** share variation patterns concerning x with flows presented in Figure 3.1. They are all smooth but have a jump at $x = 1$. Therefore, it is reasonable to assume that **U** consists of basis functions of x.

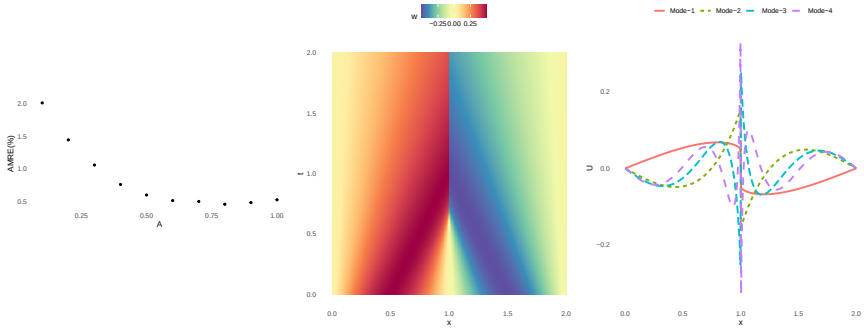


Figure 3.1: The visualization examples of SVD. Left: the AMREs of low-rank approximations for 10 data sets; middle: the density map of the solution at $A = 0.5$; right: columns of **U** varying with x.

Furthermore, Figure 3.2 presents prediction performances with various

$K(\cdot,\cdot)$, including RBF, exponential, Matérn (5/2) and Matérn (3/2), taking $\epsilon_1 = 0.99$ and $\epsilon_2 = 0.999$ as an example, and different thresholds given Matérn (5/2) correlation function. It suggests that for the Burgers equation, such four correlation functions yield comparable predictions while GP based on Matérn (5/2) is relatively better. For the threshold, all AMRE values are acceptable. As such two thresholds increase, the AMRE tends to be smaller.

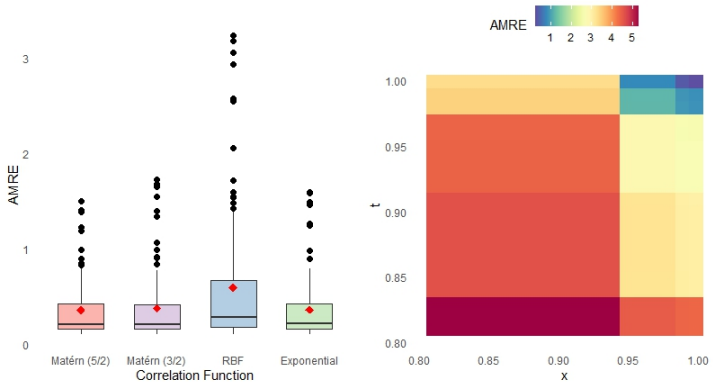


Figure 3.2: AMRE values of D-SVD based on 100 testing data with different correlation functions (left) and thresholds (right) for the Burgers equation, respectively. Red rhombus denotes the average AMRE.

Table 3.1 compares methods across 100 test datasets. IGP and LaGP incur high computational costs despite competitive AMRE values, requiring 80,000 independent GP models. This burden escalates with spatiotemporal resolution, making them impractical for high-fidelity analysis. In contrast, CPOD, KPOD, and TPOD achieve AMRE values of 1.77%, 4.79%, and 1.03% in 3.36 seconds, 3.27 seconds, and 1.04 seconds, respectively, while significantly reducing GP model counts. D-SVD delivers superior accuracy (0.36% AMRE) in about 1 second due to its multi-output GP formulation. For PSR, D-SVD outperforms TPOD, while LaGP yields the lowest score (15.61). IGP and CPOD fail to provide meaningful PSR in 98/100 and 3/100 cases, respectively, due to zero-variance predictions at certain coordinates, reflecting oversimplified modelling of x-t correlations. Notably, solving the Burgers equation requires 2.43 seconds versus D-SVD's prediction time of 0.02 seconds. D-SVD demonstrates overall superiority in accuracy, efficiency, and uncertainty quantification.

Table 3.1: Computation time (in seconds), number of GPs, average AMRE with standard deviation in parentheses, and average PSR over 100 test data sets with six methods for the Burgers equation.

	Time	Number	Average AMRE	Average PSR
D-SVD	1.11	37	0.36% (0.32%)	592.03
CPOD	3.36	600	1.77% (0.14%)	NA
KPOD	3.27	420	4.79% (4.19%)	—
IGP	475.42	80000	0.14% (0.30%)	NA
LaGP	1040.12	80000	5.87% (5.68%)	15.61
TPOD	1.04	18	1.03% (0.69%)	3628.71

NA: the result is not available; —: the result is not suitable.

3.3.2. THE NONLINEAR SCHRÖDINGER EQUATION

Next, the Schrödinger equation is used to evaluate how various methods handle the nonlinear case. Such equations with default zero-gradient boundary conditions [78] are

$$\frac{\partial w}{\partial t} = i\frac{\partial^2 w}{\partial x^2} + i|w|^2 w,$$
$$w(t=0,x) = f(x,a) + f(x-25,b),$$

where $t \in [0,40]$, $x \in [-20,80]$, $a, b \in [0,1]$, $i = \sqrt{-1}$, w is complex and

$$g(x) = \frac{2}{\exp(x)+\exp(-x)},\ f(x,\theta) = \exp(0.5\theta \times i \times x) \times g(\sqrt{0.5}x).$$

Here we focus on the absolute value. This equation is one of the fundamental equations of quantum mechanics, modelling solitons in optical fiber pulse propagation. As an initial condition, a and b are used to estimate the profile for one soliton.

We set up two scenarios to assess the performance of these methods with different n_x and n_t. The first scenario (N1) involves 400 snapshots and 200 locations, while the second scenario (N2) has 400 snapshots and 400 locations. A space-filling design involving 20 points is used for training, and 1000 data sets are used for testing. Obtaining a solution required approximately 0.97 seconds for N1 and 1.18 seconds for N2, exceeding the D-SVD prediction time (0.02 seconds). Performance metrics for D-SVD under varying correlation functions and thresholds

are detailed in the supplementary material. Despite achieving high prediction accuracy, IGP and LaGP were excluded from further analysis due to prohibitive computational costs.

Results are summarized in Table 3.2. The increased complexity of this dynamic system substantially elevated computational demands and the number of required GP models. Specifically, CPOD and KPOD incurred significant computational burdens due to their time-independence assumptions, necessitating large numbers of GP models. In contrast, our framework effectively reduced computational costs while maintaining accurate and reliable predictions. This advantage became more pronounced with increasing n_t: at $n_t = n_x = 400$, CPOD and KPOD required 38 and 78 seconds, respectively, whereas D-SVD required only 17 seconds. Furthermore, D-SVD achieved the lowest average AMRE and PSR under both scenarios. Although TPOD was faster than D-SVD, its inferior prediction accuracy and inadequate uncertainty quantification result in higher PSR values. Consequently, for complex non-smooth flows, our model delivers substantial computational savings with effective uncertainty quantification while ensuring reliable predictions.

Table 3.2: Computation time (in seconds), number of GPs, average AMRE with standard deviation in parentheses, and average PSR over 1000 test data sets with four methods for the nonlinear Schrödinger equation.

Scenario		Time	Number	Average AMRE	Average PSR
N1	D-SVD	16.43	1123	56.60% (15.31%)	3.98
	CPOD	24.31	1600	125.35% (9.79%)	6.93×10^{7}
	KPOD	44.57	3080	93.52% (160.64%)	—
	TPOD	12.54	825	59.98% (15.26%)	10.48
N2	D-SVD	17.25	1134	54.90% (15.72%)	0.74
	CPOD	38.13	3200	124.83% (9.36%)	1.11×10^{8}
	KPOD	78.55	6300	81.51% (32.31%)	—
	TPOD	13.49	826	58.53% (15.59%)	9.52

—: the result is not suitable.

3.4. REAL DATA ANALYSIS

In this section, we extend our analysis by comparing existing methods with our proposed approach using two different types of real data sets.

3.4.1. DAM DATA

A dam is a structure constructed across a river to obstruct, redirect, or slow down its movement. To analyse the water flow over a dam, [79] modelled it as the flow of water over a vertical obstacle using the Navier-Stokes equations. At low Reynolds numbers, the fluid's motion is primarily influenced by viscous forces, causing it to flow vertically down the dam. As the flow speed increases, inertial forces become more significant, leading to the formation of a jet. Gravity then causes the fluid to fall towards the boundary, where collisions create a reverse flow that impacts the dam with a velocity greater than that of the incoming flow. Consequently, the dam flow data set is valuable for studying how well models can learn to handle flows with varying viscous and inertial forces. For additional details about the data, please refer to [79]. In such a data set, there are 50 flows with different parameters $h \in \{0.11, 0.12, 0.13, 0.14, 0.15\}$ and $w \in \{0.01, 0.02, \cdots, 0.1\}$, involving 100 snapshots ($t \in [0, 10]$) and 4096 2-dimensional locations ($x \in [0, 1.5]$, $y \in [0, 0.4]$). We select 30 flows as training data and the rest 20 for testing.

Outputs with $h = 0.13$ and $w = 0.04$ at four snapshots are presented in Figure 3.3, showing how flood varies with time. Table 3.3 demonstrates the superior performance of our model. Specifically, KPOD incurs high computational costs (40 minutes) due to per-dataset ROM construction. CPOD improves speed but sacrifices accuracy. Our NDROM-based MOGP achieves both the smallest AMRE values (5.86%) and PSR scores (1.26) in relatively small costs. It suggests that our model can handle the spatiotemporal structure well for dam data based on the Navier-Stokes equations, yielding rapid and reliable predictions.

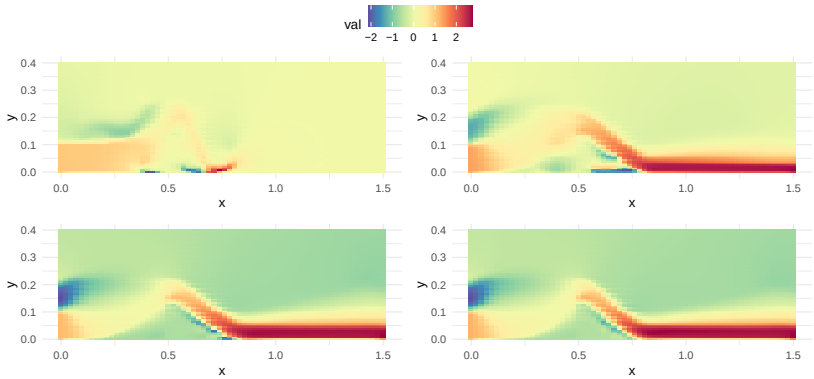


Figure 3.3: Time evolution of the dam flow with $t = 0.1$ (top-left), $t = 1$ (top-right), $t = 5$ (bottom-left) and $t = 10$ (bottom-right). $h = 0.13$, $w = 0.04$.

Table 3.3: Computation time (in seconds), number of GPs, average AMRE, and PSR with four methods for the dam data.

	Time	Number	Average AMRE	Average PSR
D-SVD	47.88	3787	5.86%	1.26
CPOD	139.83	700	10.70%	25.87
KPOD	2667.39	1300	12.59%	—
TPOD	47.80	3341	9.93%	7.46

—: the result is not suitable.

3.4.2. PETROLEUM INDUSTRY DATA

The second case is about the petroleum industry, which has a fundamental research area called reservoir simulation. Generally, a reservoir model with higher resolution is considered more realistic, especially when it is built upon a solid foundation of geological knowledge. This level of detail allows for more credible and accurate calculation results. Nevertheless, computer models based on reservoir geology can be computationally intensive, involving millions of blocks and taking hours, or even days, to complete simulations. The Delft Advanced Research Terra Simulator (DARTS) in [80] is employed as a simulator. Our investigation centres on two pivotal parameters: permeability, spanning a range from 200 to 800, and porosity, characterized by a range within $[0, 0.5]$. A total of 21 experiments are collected, each taking about 20 seconds to run. Outputs encompass 60 snapshots across varying time intervals ($t = (30, 60, \cdots, 1800)$), and span 1800 2-dimensional locations ($x \in [0, 60]$, $y \in [0, 30]$). We combine these spatial locations into a unified index and use the leave-one-out cross-validation for the case study.

Take the case where permeability is 500 and porosity is 0.3 as an example, Figure 3.4 illustrates the surface at four distinct time points. As time goes on, there is a larger and larger flow. Via the leave-one-out cross-validation, computation time, average number of GPs, average AMRE, and PSR based on 21 validations are shown in Table 3.4. D-SVD achieves the most accurate predictions with an average AMRE of approximately 9% while maintaining computational efficiency (7.8 seconds). It significantly outperforms other methods with the lowest AMRE and PSR values and enables rapid predictions in 0.02 seconds, orders of magnitude faster than DARTs. This demonstrates our framework's capability to efficiently capture spatiotemporal structures.

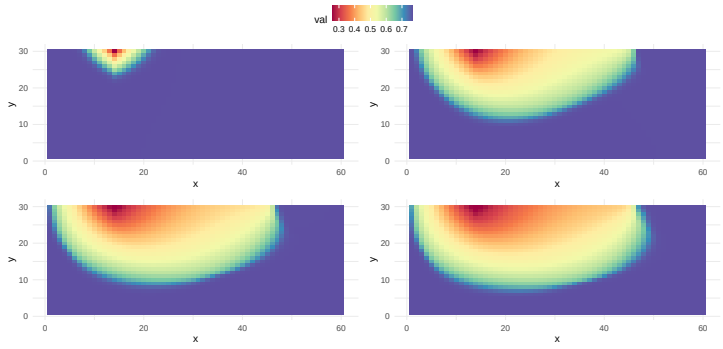


Figure 3.4: Time evolution of the reservoir simulation with $t = 30$ (top-left), $t = 600$ (top-right), $t = 1180$ (bottom-left) and $t = 1740$ (bottom-right). Permeability is 500, and porosity is 0.3.

Table 3.4: Computation time (in seconds), average number of GPs, average AMRE, and PSR with four methods for the petroleum industry data.

	Time	Average number	Average AMRE	Average PSR
D-SVD	7.80	35.10	9.05%	1.03
CPOD	10.12	120	10.11%	3.23
KPOD	140.60	160	10.57%	—
TPOD	6.78	35.09	10.18%	2.08

—: the result is not suitable.

3.5. CONCLUSION

In this work, we presented a new multi-output surrogate modelling framework for expensive spatiotemporal simulators to emulate spatiotemporal flows. An innovative non-intrusive NDROM approach via two-stage SVD was employed to reformulate such a model with a series of independent GPs. This approach achieves accurate predictions of unobserved flows while maintaining computational efficiency. Furthermore, we developed a theoretical error analysis for the NDROM and provided uncertainty quantification results in the prediction of our model. Six benchmark PDEs across diverse scenarios demonstrate the superior performance of our method compared to existing methods. Two practical applications further illustrate the model's capability to deliver accurate spatiotemporal predictions while simultaneously reducing both the number of required surrogate models and computational time.

Future research directions present several promising opportunities, including the integration of boundary and initial conditions, along with the incorporation of alternative surrogate models beyond GPs, which will require the development of advanced surrogate techniques specifically designed for physical dynamic flows. These advancements will further enhance our framework and contribute to the field's progression.

4

PCIS FOR ASYMMETRIC TOLERANCES AND NON-NORMAL DATA

Parts of this chapter have been published in Journal of Quality Technology, 57(3):200–219, 2025.

4.1. INTRODUCTION

4.1.1. BACKGROUND

The advent of innovative technologies has consistently driven the expansion of the industrial sector, fostering a highly competitive landscape that underscores the imperative of delivering superior-quality products and services. For organisations striving to enhance their manufacturing efficiency and mitigate potential operational risks, the accurate assessment of process capability indices has become a crucial prerequisite for maintaining competitive advantage and ensuring sustainable production performance. Process capability refers to the capability of an in-control production process in manufacturing conforming products. For a given product, there are usually various requirements on its quality determined by customers' needs and/or engineering tolerances, and they are called specification limits. A capable production process should have the distribution of its quality characteristic lie almost completely within the specification limits. The process capability analysis has been performed by most manufacturing companies worldwide, and its importance is significantly increasing due to the role of product quality in the modern competitive market. For example, consider a furniture manufacturer that produces tables. To ensure a consistent level of comfort and customer satisfaction, the company sets a specification limit for the height of its tables between 100 and 105 centimetres. This range is determined based on ergonomic studies and customer feedback. To select a production process that can consistently meet these requirements, the company conducts a process capability analysis. This analysis helps them determine whether the variation in table heights falls within the acceptable range and whether the process is capable of producing tables that meet the desired specifications.

In practice, process capability is often quantified by using process capability indices (PCIs). Customers rely on them to make informed decisions about accepting or rejecting supplier products, while engineers use them to continuously improve production processes to meet consumer expectations. They have been widely applied in a variety of industries to evaluate the machining process quantitatively [81–84]. One of the key advantages of the PCI lies in its simplicity, both in terms of computation and interpretation. It is important to emphasise that a low PCI value indicates that the process fails to meet customer requirements, reflecting inadequate capability, and a high PCI value signifies superior process quality. For a univariate quality characteristic, such as the height of the tables in the previous example, it is often assumed to follow a normal distribution $N(\mu, \sigma^2)$. Based on customer requirements, such as the height needing to fall within the range $[100, 105]$, we define the lower and upper specification limits as LSL and USL, respectively. The

earliest PCIs are C_p and C_{pk}, which are defined as

$$C_p = \frac{USL - LSL}{6\sigma},$$
$$C_{pk} = \min\left\{\frac{USL-\mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right\}.$$

Generally speaking, C_p measures the manufacturing tolerance relative to the spread of the distribution as a measure of process precision. Assuming that the mean of X is centred at $(LSL + USL)/2$, a larger PCI indicates reduced variability and, consequently, an improved production process. However, such an assumption is not common in practice. C_{pk} is then defined to take both the centre and spread of the distribution into account. On the other hand, customers may have an expectation of the quality characteristics of the products. The manufacturing process must attempt to obtain the required mean as well as maintain reduced variability to be within the tolerances. Following the previous example, one may prefer the height of the table to be 102.5 centimetres. Thus, with the same variability, the process whose products' heights are closer to 102.5 is better. To take the target value T into account, C_{pm} and C_{pmk} were further developed and they are given by

$$C_{pm} = \frac{C_p}{\sqrt{1 + (\frac{\mu - T}{\sigma})^2}} = \frac{USL - LSL}{6\sqrt{\sigma^2 + (\mu - T)^2}},$$
$$C_{pmk} = \frac{C_{pk}}{\sqrt{1 + (\frac{\mu - T}{\sigma})^2}} = \min\left\{\frac{USL-\mu}{3\sqrt{\sigma^2 + (\mu - T)^2}}, \frac{\mu - LSL}{3\sqrt{\sigma^2 + (\mu - T)^2}}\right\}.$$

It is clear that these two PCIs are adapted from C_p and C_{pk}, but they further reflect the extent to which the process mean differs from the target value.

It is important to note that the above-mentioned classical PCIs rely on two assumptions. Firstly, the quality characteristic X is assumed to be normal. Although this assumption is widely adopted in the PCI literature, it may not be appropriate for a variety of quality characteristics [85, 86]. Some examples include the measurements of metallic impurities in silicon wafers [87], the weights of rubber edges [88], and the capacitances of non-polarised capacitors [89]. When the normality assumption is violated, the classical PCIs cannot reflect the process capability. This is because these indices are based on the idea that the range $[\mu - 3\sigma, \mu + 3\sigma]$ covers approximately 99.73% of the data, resulting in a long-term defect rate of fewer than 3.4 defects per million opportunities. However, if the underlying distribution deviates from normality—for example, if it is heavy-tailed or skewed—the 6σ range may no longer correspond to the expected defect rate. In such cases, directly applying the 6σ rule can be

misleading, as the actual proportion of defects may significantly differ from the theoretical value derived under the normality assumption.

Secondly, the target value T is implicitly assumed to be in the middle point of the specification limits, i.e., $T = (USL + LSL)/2$. In the presence of asymmetric tolerances about the target value, C_{pm} and C_{pmk} will produce unsatisfactory results as they cannot reflect process performance relative to the process characteristic staying within the tolerances. The asymmetric tolerance is, however, not uncommon in real applications when customers have different preferences in the two directions from the target value [90]. Using the table production example again, even though the height of the tables should fall within the range $[100, 105]$ centimetres, the company prefers the height to be closer to the upper limit (e.g., 104 centimetres) for easier maintenance. In such cases, it is unfair to assign equal costs to deviations from the target value solely based on their magnitude, as the direction of the deviation (whether above or below the target) may have different implications. For instance, a height below the target might lead to usability issues, while a height above the target might only slightly affect maintenance efforts. Moreover, asymmetric tolerance naturally arises when data transformation is performed to achieve approximate normality.

There have been a variety of PCIs proposed in the literature to deal with non-normal data and asymmetric tolerances, but few studies have considered these two important issues simultaneously. In the next subsection, we give a literature review of the existing PCIs.

4.1.2. LITERATURE REVIEW

For concise presentation, we first introduce the following superstructure proposed by [91], which unifies the four classical PCIs

$$C_p(u, v) = \frac{d - u|\mu - M|}{3\sqrt{\sigma^2 + v(\mu - T)^2}}, \tag{4.1}$$

where $u, v \geq 0$, $d = (USL - LSL)/2$, and $M = (USL + LSL)/2$. It is easy to see that $C_p(0, 0) = C_p$, $C_p(1, 0) = C_{pk}$, $C_p(0, 1) = C_{pm}$, and $C_p(1, 1) = C_{pmk}$. In the following two subsections, we review the existing PCIs for asymmetric tolerances and non-normal process data, respectively.

PCIS FOR ASYMMETRIC TOLERANCES

The determination of target values is typically established at the midpoint value within the specification limits. This conventional approach is predominantly adopted to optimise manufacturing efficiency by minimising the probability of non-conforming product generation beyond the prescribed tolerance boundaries. Nevertheless, in specific engineering applications, non-central target values are deliberately

implemented to enhance assembly compatibility or mitigate potential challenges in subsequent manufacturing operations. Consequently, this practice leads to asymmetric tolerances. Under the normal distribution, adjusting asymmetric specification limits and then using classical PCIs on the new specification limits is one of the earliest ways to deal with asymmetric tolerances. [92] considered shifting one of the two specification limits and applied $(T-d^*, T+d^*)$ instead of the true limits $(T-D_l, T+D_u)$, where $D_l = T-LSL$, $D_u = USL-T$ and $d^* = \min\{D_l, D_u\}$. The PCIs can then be obtained by replacing d by d^* in the superstructure (4.1). By confining the process to a subset of the real specification range, these indices provide an approximate quantification of the process capability. For instance, the limits are 100 and 105, and the target value is 104. One can use $[103, 105]$, $[100, 108]$ or $[101.5, 106.5]$ to introduce a symmetric case. However, depending on such virtual limits, those PCIs can not evaluate the manufacturing process exactly according to customers' needs.

The other category of methods is to propose tailored PCIs to address the asymmetric tolerances [93–98]. For example, [94] considered the effects of the direction between μ and T, and proposed the PCIs by adding the term $|\mu-T|$ to the numerator of (4.1). The so-obtained PCIs cannot reach the maximum at $\mu = T$, which is not practical since a process that meets the exact target value should be the best. [95] further proposed another class of PCIs as

$$C_p^F(u,v) = \frac{d^* - uF^*}{3\sqrt{\sigma^2 + vF^2}}, \tag{4.2}$$

where

$$F = \max\left\{\frac{d(\mu-T)}{D_u}, \frac{d(T-\mu)}{D_l}\right\},$$
$$F^* = \max\left\{\frac{d^*(\mu-T)}{D_u}, \frac{d^*(T-\mu)}{D_l}\right\}.$$

This superstructure obtains its maximum at $\mu = T$ and is not symmetric around T, which reflects the customers' preference towards different directions from the target value. More recently, [98] noted that $C_p^F(u,v)$ is zero at both specification limits, which may not be reasonable as PCIs at the closer limit to T should have a higher value. Toward this end, the authors proposed C_p^A, replacing F^* by

$$A^* = \frac{(\mu-T)^2}{D_u} I\{\mu > T\} + \frac{(T-\mu)^2}{D_l} I\{\mu \le T\},$$

and F by

$$A = \frac{d(\mu-T)}{D_u} I\{\mu > T\} + \frac{d(T-\mu)}{D_l} I\{\mu \le T\}.$$

To the best of our knowledge, these PCIs are by far commonly used when there are different preferences on the direction of the process mean from the target value. However, as we will demonstrate, there are still some drawbacks when using them.

PCIS FOR NON-NORMAL DATA

There are broadly two approaches to developing PCIs when the quality characteristic is not normally distributed. The first approach is the percentile-based PCIs, originally proposed by [99]. The 6σ range can be replaced with distribution percentiles. Specifically, after fitting the data to an appropriate probability distribution, the discrepancy between the 0.135% percentile and the 99.865% percentile can be used as an alternative measure. Subsequent research has expanded this framework in multiple directions. For instance, [100] enhanced its applicability by incorporating fuzzy sets to better handle data uncertainty and imprecision. Additionally, adaptations to specific probability distributions, such as the Rayleigh [101], gamma [102], logistic-exponential [103], and inverse Gaussian [104], have further improved its flexibility and accuracy within defined distribution families. Furthermore, recent studies have emphasised the construction of confidence limits for percentile-based PCIs, contributing to their robustness and reliability [105–107]. Despite these advancements, the effectiveness of this approach remains contingent on the accurate selection and estimation of an appropriate distribution, which can be challenging in practical applications.

The second approach is the data-transformation method, which has gained increasing attention in recent years as an alternative to the percentile-based approach. Instead of fitting the data to a specific distribution, this method transforms the original data into a normally distributed form, allowing existing PCIs to be directly applied. Commonly used transformations include the logarithmic transformation [108, 109], the Box-Cox transformation [110, 111], and the inverse normal transformation [112, 113]. However, a key challenge of this approach is the difficulty in interpreting the results in the original data scale. To address this issue, [114] proposed a transformation based on the cumulative distribution function (CDF), and demonstrated its applicability to any continuous process data while preserving a quantitative interpretation of process capability.

4.1.3. AIMS AND OUTLINE

Despite extensive research on PCIs, a unified and versatile framework that accommodates both different tolerance structures (symmetric or asymmetric) and data distributions (normal or non-normal) remains lacking. In particular, as will be discussed later, existing PCIs for

asymmetric tolerances often lack key desirable properties, while kernel estimation methods for the inverse transformation in non-normal data are prone to substantial bias near the boundaries of the data range. To address these shortcomings, this study proposes new capability indices together with a novel inverse transformation method based on constrained B-spline regression. This approach not only resolves the limitations in handling asymmetric tolerances but also offers a more robust solution for non-normal data. Importantly, our unified framework remains applicable even when tolerances are symmetric and data are normally distributed, ensuring consistent performance across various settings. Consequently, it simplifies decision-making and enhances process capability assessment in a wide range of manufacturing scenarios.

The rest of this chapter is organised as follows. In Section 4.2, we first give an in-depth discussion on the desirable properties of PCIs for the asymmetric tolerances, based on which two novel superstructures are proposed. Moreover, their relationships with the process centring and with the process yield are established. Section 4.3 provides comparison studies between the proposed PCIs and the existing ones through extensive simulations. Building on the idea of inverse CDF transforming, the non-normal issue is considered in Section 4.4. We propose a constrained B-spline regression model to obtain a smooth and non-decreasing estimator of the CDF and compare it with some existing methods. A real example is used for illustration in Section 4.5, and concluding remarks are made in Section 4.6. All technical proofs and additional simulation results are shown in Section 7.3.

4.2. NEW PCIS FOR ASYMMETRIC TOLERANCES

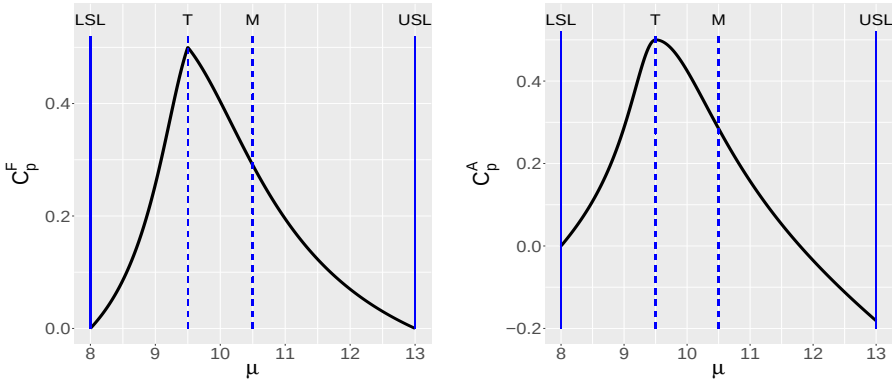
In this section, we first give a thorough analysis of the existing PCIs for asymmetric tolerances and elucidate the essential properties that effective PCIs should embody. Subsequently, we introduce two new PCIs that encompass all discernible properties identified during our analysis. Throughout this section, we assume that the quality characteristic X follows a normal distribution $N(\mu, \sigma^2)$.

4.2.1. EXISTING PCIS FOR ASYMMETRIC TOLERANCES

As reviewed in Section 4.1.2, two useful PCIs for asymmetric tolerances were developed in [95] and [98], and their motivation is based on two deficiencies of the previous PCIs. Firstly, these indices are not maximised when $\mu = T$ for a given σ, which is undesirable. This is because a process whose mean (μ) aligns with the target value (T)—meaning the products fully meet customers' needs—should be considered better than other processes with the same variability (σ). Therefore, the corresponding PCI

values should be the largest in such cases. Consequently, the application of conventional PCIs may lead to suboptimal process selection decisions, potentially resulting in significant quality-related losses and compromised operational efficiency. Secondly, they do not reflect the directions of μ towards T, consequently limiting their applicability in assessing the process's ability to precisely satisfy customer-specific requirements and preferences. For illustration, consider a manufacturing scenario where $LSL = 100$, $USL = 105$, and $T = 103$. In this case, two distinct processes - one with $\mu = 102$ and another with $\mu = 104$ - yield identical $C_p(u, v)$ values. Nevertheless, from a quality control perspective, the process with $\mu = 102$ demonstrates superior performance as it generates a statistically lower expected proportion of non-conforming units compared to its counterpart.

In contrast, Figure 4.1 shows how these two PCIs change with μ when $(LSL,\ T,\ USL)$ are set as (8, 9.5, 13), $\sigma = 1$ and $u = v = 1$. As seen, both PCIs are maximised at $\mu = T$ and they are not symmetric about T.



(a) Values of C_p^F for different μ.

(b) Values of C_p^A for different μ.

Figure 4.1: C_p^F, C_p^A values with $\sigma = 1$, $T = 9.5$ and μ ranging between the specification limits (8, 13).

One major problem with C_p^F, as noted by [98], is that it does not credibly quantify the shifts from the target value. Take Figure 4.1 as an example, the PCI at LSL should be larger than the PCI at USL since T is closer to LSL and the process has the same yield at $\mu = USL$ and $\mu = LSL$. This deficiency is remedied by C_p^A, which yields different values at USL and LSL. However, C_p^A will be negative for a certain range of μ, which brings difficulty in interpreting the process yield. Moreover, unlike other PCIs, C_p^A does not (4.1) when the process is symmetric, i.e., $T = M$, which further restricts its applications in practice.

Based on the preceding analysis, we systematically summarise the essential properties of PCIs that are particularly relevant for asymmetric tolerance scenarios. These properties encapsulate the collective advantages of existing PCIs when applied to non-symmetric processes, as outlined below.

(i) PCIs should be maximised at $\mu = T$. This reflects the fundamental principle that a process is at its best when it consistently meets the desired target, assuming the same level of variability σ.

(ii) PCIs should decline more steeply as μ shifts from T to the closer specification limit. This design captures the preference for focusing more on the direction that is more critical to customer requirements.

(iii) PCIs should be smaller on the further specification limit to T than those on the closer specification limit.

(iv) PCIs should be non-negative since negative values lack practical interpretation.

(v) As a broadening version of the traditional case, PCIs should degenerate to the classical ones when the tolerance is symmetric. In other words, PCIs can be utilised for both kinds of tolerances.

Table 4.1 summarizes the existing PCIs for asymmetric tolerance and their properties, where $S(x, y) = \Phi^{-1}(\Phi(x)/2 + \Phi(y)/2)$ with the cumulative distribution function of the standard normal distribution Φ. As seen, none of them satisfy all five properties, which motivates us to propose new PCIs.

4.2.2. PROPOSAL I: $C_{pn1}(u, v)$

From Table 4.1, we can see that $C_p^F(u, v)$ meets all the desired properties except for the third one. Based on the formula of $C_p^F(u, v)$ in (4.2), it is readily seen that F^* is equal to d^* when $u = 1$ and $\mu = LSL$ or USL, making $C_p^F(1, v) = 0$ at both specification limits. To make Property (iii) hold, we add an additional term $d^+ = \max\{D_l, D_u\}$ to the denominator of F^*, and the corresponding PCIs are

$$C_{pn1}(u, v) = \frac{d^* - uF^+}{3\sqrt{\sigma^2 + vF^2}},$$

where

$$F = \max\left\{\frac{d(\mu - T)}{D_u}, \frac{d(T - \mu)}{D_l}\right\},$$

$$F^+ = \max\left\{\frac{2d^*(\mu - T)}{d^+ + D_u}, \frac{2d^*(T - \mu)}{d^+ + D_l}\right\}.$$

Table 4.1: Summary of existing PCIs for asymmetric tolerances; x denotes violation of a certain property.

PCIs	Form	(i)	(ii)	(iii)	(iv)	(v)
[92]	$\min\{\frac{USL-T}{3\sigma}, \frac{T-LSL}{3\sigma}\}$	x	x	x		
[93]	$S(\frac{USL-\mu}{\sigma}, \frac{\mu-LSL}{\sigma})$		x	x	x	
[94]	$\frac{d-\lvert\mu-M\rvert-u\lvert\mu-T\rvert}{3\sqrt{\sigma^2+v(\mu-T)^2}}$	x	x	x		
[95]	C_p^F			x		
[96]	$[1-\max\{\frac{\mu-T}{USL-T}, \frac{T-\mu}{T-LSL}\}]\frac{d}{3\sigma}$			x		
[97]	$1-\frac{\max\{d^*(\mu-T)/D_u, d^*(T-\mu)/D_l\}}{d^*}$			x		
[98]	C_p^A				x	x

Because F and F^+ are always non-negative, $C_{pn1}(u,v)$ reaches its maximum when $F = F^+ = 0$, which corresponds to the case of $\mu = T$. Secondly, by retaining D_l and D_u in F^+, Property (ii) holds for $C_{pn1}(u,v)$ when $u, v > 0$. Thirdly, F^+ is equal to d^* on the further limit, and it is less than d^* on the closer limit. As a result, $C_{pn1}(u,v)$ remains 0 on the further limit and is strictly positive on the closer limit, which makes Property (iii) hold. Based on the above discussion, Property (iv) also holds for $C_{pn1}(u,v)$ when $0 \le u \le 1$. Lastly, we have $D_l = D_u = d^* = d^+ = d$ and $\max\{\mu - M, M - \mu\} = |\mu - M|$ when $T = M$. Therefore, $C_{pn1}(u,v)$ degenerates to the classical superstructure $C_p(u,v)$ in (4.1). In summary, $C_{pn1}(u,v)$ meets all the five properties for any $u \in [0,1]$, $v \ge 0$.

RELATION TO PROCESS CENTRING

Process centring is the ability of processes to congregate around the target value T, which can be measured as the distance between μ and T, i.e., $|\mu - T|$. It can reflect the variation from the target value, which is also a guiding principle in measuring quality improvement. For instance, a higher PCI value coupled with a greater $|\mu - T|$ suggests that there may be more products exceeding the specified limits, or that the PCI might overestimate the process capability. For the tailored PCIs for asymmetric tolerances, it is important to link them with the process centring. Given $C_{pn1}(u,v) = c$, where c is often determined by engineers or customers in their purchasing contract, it is readily seen that

$$c = \frac{d^* - uF^+}{3\sqrt{\sigma^2 + vF^2}} \le \frac{d^* - uF^+}{3\sqrt{vF^2}}.$$

When $\mu > T$, we have

$$c \le \frac{d^* - u\frac{2d^*}{d^+ + D_u}(\mu - T)}{3\sqrt{v}\frac{d}{D_u}(\mu - T)},$$

and hence

$$\mu - T \le \frac{d^*}{3c\sqrt{v}\frac{d}{D_u} + u\frac{2d^*}{d^+ + D_u}} \triangleq \lambda_1.$$

Similarly, when $\mu \le T$, we have

$$T - \mu \le \frac{d^*}{3c\sqrt{v}\frac{d}{D_l} + u\frac{2d^*}{d^+ + D_l}} \triangleq \lambda_1'.$$

Thus, the mean is bounded by

$$T - \lambda_1' \le \mu \le T + \lambda_1.$$

Note that we assume that $u \neq 0$ in the derivations, which is also adopted in [98].

RELATION TO PROCESS YIELD

Another major role of PCIs is to infer the process yield, which is defined as the percentage of products that pass inspections, i.e., the probability of the quality characteristic that is within the specification limits. In practice, the process yield is often quantified by using the percentage of non-conforming (NC) products, which is given by $\text{NC} = P(X < LSL) + P(X > USL)$. We aim to build the relation between $C_{pn1}(u, v)$ and NC here.

Recall that $T - \lambda_1' \le \mu \le T + \lambda_1$. If the target is closer to the upper limit (i.e., $T > M$), we have more non-conforming products when $\mu > T$. Therefore, given $C_{pn1}(u, v) = c$, we have

$$\begin{aligned} \text{NC} \le 2P(X > USL) &= 2\left[1 - P(X < USL)\right], \\ &= 2\left[1 - \Phi\left(\frac{USL - \mu}{\sigma}\right)\right], \\ &\le 2\left[1 - \Phi\left(\frac{D_u - \lambda_1}{\sigma}\right)\right]. \end{aligned}$$

Note that such an upper bound can also be used as a metric of quality assurance for customers. From (**??**) we have $c \le \frac{d^*}{3\sigma}$, and hence σ is bounded by $\frac{d^*}{3c}$. As a result, the upper bound of NC is

$$\text{NC} \le 2\left[1 - \Phi\left(\frac{3c(D_u - \lambda_1)}{d^*}\right)\right].$$

Similarly, if the target is closer to the lower limit, i.e., $T \le M$, then we have more non-conforming products when $\mu \le T$.

$$\begin{aligned} \text{NC} \le 2P(X < LSL) &= 2\left[\Phi\left(\frac{LSL - \mu}{\sigma}\right)\right], \\ &= 2\left[1 - \Phi\left(\frac{\mu - LSL}{\sigma}\right)\right], \\ &\le 2\left[1 - \Phi\left(\frac{3c(D_l - \lambda_1')}{d^*}\right)\right]. \end{aligned} \tag{4.3}$$

4.2.3. PROPOSAL II: $C_{pn2}(u, v)$

Our second proposal of the PCIs for asymmetric tolerances is based on the classical superstructure (4.1). The major difficulty of the construction lies in Property (ii). To make the PCIs asymmetric about T, we introduce a ratio $R_1 = \min\{T/\mu,\ \mu/T\}$ as a function of μ. It is easy to check that $R_1(T + \epsilon)$ is larger than $R_1(T - \epsilon)$ for any $\epsilon > 0$. Therefore, R_1 is asymmetric about T, and it is suitable for the case of $T < M$. Moreover, to

make PCIs unequal at the two specification limits, another ratio function $R_2 = D_l/D_u$ is further introduced. Taken together, the second proposal $C_{pn2}(u, v)$ for $T < M$ is constructed as

$$C_{pn2}(u, v) = \frac{D_l - uR_2|\mu - T|}{3\sqrt{\sigma^2 + v(\mu - T)^2}} \times \mathcal{I}, \tag{4.4}$$

where $\mathcal{I} = I\{T < M\} \times R_1 + I\{T = M\}$ and I is the indicator function. When $T > M$, given the sample $X_1, \cdots, X_n$ with the specification tolerances $(LSL,\ T,\ USL)$, we can apply the transformation $L - X_1, \cdots, L - X_n$ with a constant $L > \max\{X_1, \cdots, X_n\}$. Then, the transformed target value and middle point satisfy $L - T < L - M$, making (4.4) applicable. Without loss of generality, we only consider the case of $T < M$ for $C_{pn2}(u, v)$ in this study.

It remains to verify that C_{pn2} satisfies Property (i), (iv) and (v). Firstly, when $\mu = T$, R_2 is a constant and R_1 reaches its maximal value. Thus, C_{pn2} is maximized at $\mu = T$. Secondly, D_l is always larger than R_2 with $u \leq 1$, and $\mathcal{I}$ is always positive. As a result, C_{pn2} is always non-negative. Lastly, when $T = M$, we have $D_l = D_u = d$, $R_2 = 1$ and $\mathcal{I} = 1$, and hence C_{pn2} degenerates to (4.1).

RELATION TO PROCESS CENTRING

Given $C_{pn2}(u, v) = c$ and $T < M$, we have

$$c \leq \frac{D_l - u\frac{D_l}{D_u} \times |\mu - T|}{3\sqrt{\sigma^2 + v(\mu - T)^2}} \times \frac{T}{\mu} \leq \frac{D_l - u\frac{D_l}{D_u} \times |\mu - T|}{3\sqrt{v(\mu - T)^2}},$$

which indicates that

$$|\mu - T| \leq \frac{D_l}{u\frac{D_l}{D_u} + 3c\sqrt{v}}.$$

Therefore, the range of μ is symmetric about T, which is

$$T - \frac{D_l}{u\frac{D_l}{D_u} + 3c\sqrt{v}} \leq \mu \leq T + \frac{D_l}{u\frac{D_l}{D_u} + 3c\sqrt{v}}. \tag{4.5}$$

Based on the derivation in Section 4.2.2, the range of μ given $C_{pn1} = c$ and $T < M$ is

$$T - \frac{D_l}{3c\sqrt{v}\frac{d}{D_l} + u\frac{2D_l}{D_u + D_l}} \leq \mu \leq T + \frac{D_l}{3c\sqrt{v}\frac{d}{D_u} + u\frac{D_l}{D_u}}.$$

Because $d/D_u < 1$, the upper bound of μ based on C_{pn2} is smaller than that based on C_{pn1}. On the other hand, C_{pn1} gives tighter lower bound than C_{pn2} as $2D_l/(D_u + D_l) > D_l/D_u$ and $d/D_l > 1$. The detailed comparison is given in Section 4.2.4.

RELATION TO PROCESS YIELD

Similar to C_{pn1}, we can also derive the relation between C_{pn2} and NC. Under the assumption of normal distributions and $T < M$, we have

$$\begin{aligned} \text{NC} \leq 2P(X < LSL) &= 2\left[\Phi\left(\frac{LSL-\mu}{\sigma}\right)\right], \\ &= 2\left[1-\Phi\left(\frac{\mu-LSL}{\sigma}\right)\right]. \end{aligned}$$

Based on the lower bounds of μ in (4.5) and the similar upper bound of σ as in C_{pn1}, we have

$$\text{NC} \leq 2\left[1-\Phi\left(3c\left(1-\frac{D_u}{uD_l+3D_uc\sqrt{v}}\right)\right)\right].$$

Compared with (4.3), this upper bound on NC is more conservative as the lower bound of μ based on C_{pn2} tends to be smaller.

4.2.4. COMPARISON STUDIES

In this subsection, we demonstrate the properties of the two proposed PCIs by comparing them with C_p^F and C_p^A. We use the same settings as in Section 4.2.1, i.e., $(LSL,\ T,\ USL) = (8,\ 9.5,\ 13)$ and $\sigma = 1$. All implementations and analyses are carried out in the R environment [77]. Figure 4.2 shows how the four PCIs vary when $\mu \in [8, 13]$ at $(u, v) = (0, 0), (0, 1), (1, 0)$ and $(1, 1)$. When $(u, v) = (0, 0)$, it is observed that C_{pn1}, C_p^F and C_p^A are the same constant because they all reduce to $d^*/(3\sigma)$, which does not depend on μ. This is undesirable because the difference between T and μ cannot be reflected. On the other hand, because of the term $\mathcal{I}$, C_{pn2} is able to take the direction of the departure from the target into account, and it is maximised at $\mu = T$. Therefore, C_{pn2} is the only suitable PCI for asymmetric tolerances when $u = v = 0$.

Under the other three settings of (u, v), the four PCIs are all maximised at $\mu = T$ and asymmetric around T. Because of the intimate relation between C_{pn1} and C_p^F, it is not surprising to find that they perform similarly in all the cases. In fact, they are identical when $\mu > T$, but C_{pn1} improves C_p^F by imposing the positive value at $\mu = LSL$ as T is closer to LSL than to USL. It is worth mentioning that C_p^A is also able to reflect the differences at the two specification limits, but it is the only one that yields negative values, violating Property (iv).

We then compare these four PCIs in terms of their relations with process centring and process yield. Because the bounds based on C_p^A are not well defined when $u = 0$ [98] and the bounds based on C_{pn1} and C_{pn2} are invariant to index values when $v = 0$, we only consider the

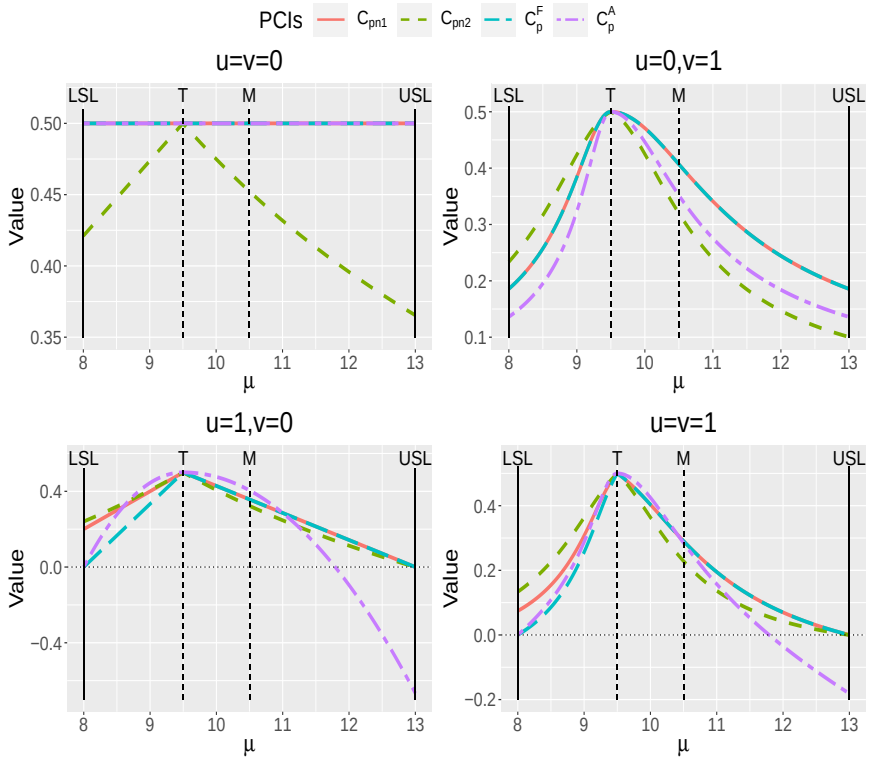


Figure 4.2: Numerical comparisons of C_{pn1}, C_{pn2}, C_p^F and C_p^A under different combinations of u and v, where $(LSL,\ T,\ USL) = (8,\ 9.5,\ 13)$ and $\sigma = 1$.

case $(u, v) = (1, 1)$ here. Figure 4.3 shows the bounds of μ and NC given different PCIs when $u = v = 1$. As seen from the left panel, the trends of the bounds from all four PCIs are consistent: both the upper and lower bounds get tight first and then become loose, and the smallest gap is achieved when $\mu = T$ as the PCIs are maximised at $\mu = T$. Among the four PCIs, the lower bound based on C_{pn2} seems to be the most conservative, but the upper bound is the tightest. On the other hand, C_{pn1} performs more similarly as C_p^F and C_p^A. Notably, the bounds based on C_p^A fail to contain the true mean when μ is large. This is because C_p^A becomes negative at large μ (see Figure 4.2), which further indicates the importance of Property (iv). On the other hand, the upper bounds of NC are conservative based on all the considered PCIs. Between the proposed C_{pn1} and C_{pn2}, C_{pn1} performs better and it is comparable to C_p^F. The bound based on C_p^A again has an irregular trend when μ is large because of the negative values of C_p^A.

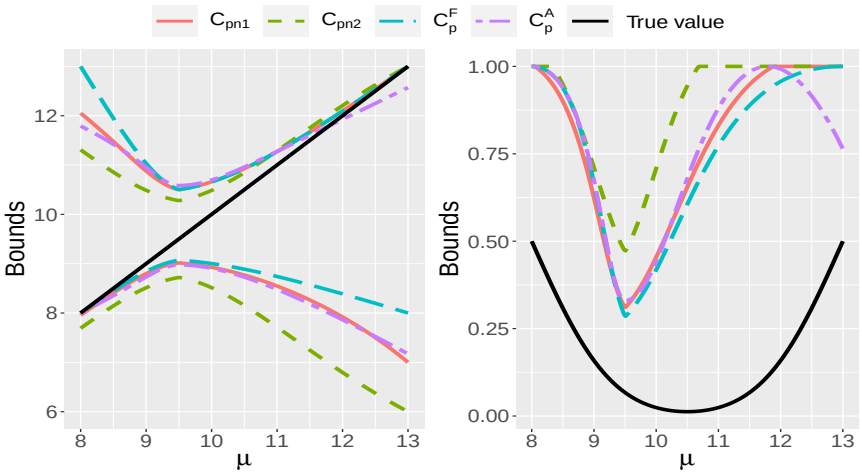


Figure 4.3: Upper and lower bounds of μ (left) and upper bounds of NC (right) given different PCIs. $LSL = 8$, $USL = 13$, $T = 9.5$, $\sigma = 1$ and $u = v = 1$.

Our comparative analysis demonstrates that the proposed C_{pn1} and C_{pn2} effectively address the limitations of C_p^F and C_p^A. Both indices evaluate whether manufactured items meet asymmetric specifications by credibly capturing process variability, yet they exhibit distinct behaviours, particularly in terms of process centring and yield. As highlighted by [115], selecting an appropriate PCI is crucial in process capability studies. To guide this selection, we provide practical recommendations based on specific scenarios. For instance, as shown in Figure 4.2, C_{pn2}

yields higher values when μ is closer to the LSL, thus offering a more conservative overall assessment. Consequently, C_{pn2} is recommended when asymmetric tolerances exist and the cost implications of deviations in either direction are similar; it is also the sole meaningful index when $u = v = 0$. Conversely, if the primary concern is process centring and yield, C_{pn1} is preferable. In summary, the two proposed PCIs are designed for different application contexts, which provide a comprehensive and flexible framework for assessing process capability across diverse manufacturing scenarios.

4.3. STATISTICAL INFERENCE

In practice, the proposed PCIs have to be estimated based on the observed data. Let $\boldsymbol{x} = [x_1, \ldots, x_n]^\top$ be a vector of random copies of the quality characteristic X. This section proposes point and interval estimation procedures for C_{pn1} and C_{pn2}.

4.3.1. MAXIMUM LIKELIHOOD ESTIMATION

Recall that $X \sim N(\mu, \sigma^2)$. The maximum likelihood (ML) estimators of μ and σ are $\hat{\mu} = \bar{x} \equiv \sum_{i=1}^{n} x_i/n$ and $\hat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$. Based on the invariance property of ML estimators and the plug-in method, the point estimators of C_{pn1} and C_{pn2} are

$$\begin{aligned} \hat{C}_{pn1}(u, v) &= \frac{d^* - u\hat{F}^+}{3\sqrt{\hat{\sigma}^2 + v\hat{F}^2}}, \\ \hat{C}_{pn2}(u, v) &= \frac{D_l - uR_2|\hat{\mu} - T|}{3\sqrt{\hat{\sigma}^2 + v(\hat{\mu} - T)^2}} \times \hat{\mathcal{I}}, \end{aligned} \tag{4.6}$$

where

$$\begin{aligned} \hat{F} &= \max\left\{\frac{d(\hat{\mu} - T)}{D_u}, \frac{d(T - \hat{\mu})}{D_l}\right\}, \\ \hat{F}^+ &= \max\left\{\frac{2d^*(\hat{\mu} - T)}{d^+ + D_u}, \frac{2d^*(T - \hat{\mu})}{d^+ + D_l}\right\}, \end{aligned}$$

$\hat{R_1} = \min\{T/\hat{\mu},\ \hat{\mu}/T\}$, $R_2 = D_l/D_u$, and $\hat{\mathcal{I}} = I\{T < M\} \times \hat{R_1} + I\{T = M\}$.

On the other hand, the confidence limits of the PCIs are often of more interest in practice. There has been a multitude of studies addressing interval estimation in the literature [116–118], but few studies have investigated PCIs for asymmetric tolerances. Because of the complex forms in (4.6), the exact variance of $\hat{C}_{pn1}(u, v)$ and $\hat{C}_{pn2}(u, v)$ cannot be analytically derived. However, their asymptotic variance can be readily obtained by using the asymptotic variance of $\hat{\mu}$ and $\hat{\sigma}^2$. Note that $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2)$ and $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$, where $\xrightarrow{d}$

denotes convergence in distribution. Meanwhile, such two estimators are independent, which means the vector $\sqrt{n}[(\hat{\mu}-\mu), \hat{\sigma}^2-\sigma^2]^\top \xrightarrow{d} N(0,\Sigma)$, where

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

Given a function $g(x,y)$ with the continuous first partial derivatives, the delta method yields that [119]

$$\sqrt{n}\left[g(\bar{x},\hat{\sigma}^2)-g(\mu,\sigma^2)\right] \xrightarrow{d} N([0,0]^\top, \nabla g(\mu,\ \sigma^2)\Sigma(\nabla g(\mu,\ \sigma^2))^\top),$$

with

$$g(x,y) = \left(\frac{\partial g}{\partial x},\ \frac{\partial g}{\partial y}\right).$$

In terms of the proposed PCIs, we focus on the $T < M$ case, and confidence limits for the $T > M$ case can be constructed similarly. To meet the conditions on the derivatives, we derive the function $g_1(x,y)$ for $C_{pn1}(u,v)$ and $g_2(x,y)$ for $C_{pn2}(u,v)$ by considering three cases: $x > T$, $x = T$ and $x < T$. More specifically, we have

$$g_1(x,y) = \begin{cases} g_{11}(x,y) = \dfrac{D_l - u\frac{2d^*(x-T)}{d^+ + D_u}}{3\sqrt{y+v(\frac{d}{D_u}(x-T))^2}} & x > T, \\ g_{12}(x,y) = \dfrac{D_l}{3\sqrt{y}} & x = T, \\ g_{13}(x,y) = \dfrac{D_l - u\frac{2d^*(T-x)}{d^+ + D_l}}{3\sqrt{y+v(\frac{d}{D_l}(T-x))^2}} & x < T. \end{cases}$$

And

$$g_2(x,y) = \begin{cases} g_{21}(x,y) = \dfrac{D_l - u\frac{D_l}{D_u}(x-T)}{3\sqrt{y+v(x-T)^2}} \times \dfrac{T}{x} & x > T, \\ g_{22}(x,y) = \dfrac{D_l}{3\sqrt{y}} & x = T, \\ g_{23}(x,y) = \dfrac{D_l - u\frac{D_l}{D_u}(T-x)}{3\sqrt{y+v(x-T)^2}} \times \dfrac{x}{T} & x < T. \end{cases}$$

The delta method can then be invoked in each scenario, and the unknown parameters can be replaced by their ML estimators. In summary, the asymptotic variance of $C_{pni}(u,v)$, $i = 1, 2$, can be consistently estimated

by

$$\hat{V}_i = \begin{cases} \nabla g_{i1}(\hat{\mu}, \hat{\sigma}^2)\hat{\Sigma}(\nabla g_{i1}(\hat{\mu}, \hat{\sigma}^2))^\top & \hat{\mu} > T, \\ \dfrac{D_l^2}{18\hat{\sigma}^2} & \hat{\mu} = T, \\ \nabla g_{i3}(\hat{\mu}, \hat{\sigma}^2)\hat{\Sigma}(\nabla g_{i3}(\hat{\mu}, \hat{\sigma}^2))^\top & \hat{\mu} < T. \end{cases} \tag{4.7}$$

Although the delta method is primarily tailored for smooth and differentiable functions, its potential applicability in this context merits consideration. Considering various samples, corresponding estimators and functions g_1 and g_2 emerge, each associated with distinct scenarios, all maintaining continuity and differentiability. For example, when $\hat{\mu} > T$, the resulting index values manifest as $g_{11}(\hat{\mu}, \hat{\sigma}^2)$ and $g_{21}(\hat{\mu}, \hat{\sigma}^2)$. Notably, both previously introduced functions, g_{11} and g_{21}, demonstrate suitability for the delta method. Given a confidence level $1-\alpha$, the asymptotic confidence interval can be obtained as $\hat{C}_{pni}(u, v) \pm z_{\alpha/2}\sqrt{\hat{V}_i/n}$, $i = 1, 2$, where $z_{\alpha/2}$ is the critical value of the standard normal distribution. Note that such an interval will contain the true parameter value with a probability of $1-\alpha$ as $n \to +\infty$.

4.3.2. SIMULATION STUDY

A simulation is conducted to assess the performance of the proposed confidence intervals. We consider the asymmetric tolerance $(LSL,\ T,\ USL) = (20,\ 26,\ 40)$ and $u = v = 1$. Based on (4.7), three scenarios $\mu < T$, $\mu = T$ and $\mu > T$ are simulated. Under $\mu < T$ ($\mu > T$), we fix $\sigma = 1$ and construct 95% confidence intervals with $\mu \in [LSL, T)$ ($\mu \in (T, USL]$). For $\mu = T$, a sequence of σ involving 30 different values from 0.3 to 9 is considered. The coverage probabilities based on 10000 replications for the three scenarios are shown in Figures 4.4-4.6, where $n = 40, 80, 120, 160, 200$ are considered. As seen, the coverage probabilities are generally close to the nominal level in all the cases, and confidence limits are asymptotic with the sample size n. These results indicate that the proposed methods can be credibly used to construct confidence intervals of $C_{pn1}(u, v)$ and $C_{pn2}(u, v)$.

4.4. THE NON-NORMAL CASE

The proposed PCIs are under the assumption that the process characteristic X follows a normal distribution. In this section, we aim to extend the proposed PCIs to the non-normal case using transformation methods. As previously discussed, one effective way involves using the inverse transformation with the CDF, offering a versatile means to convert the data into a normal distribution. Assume that X is continuous,

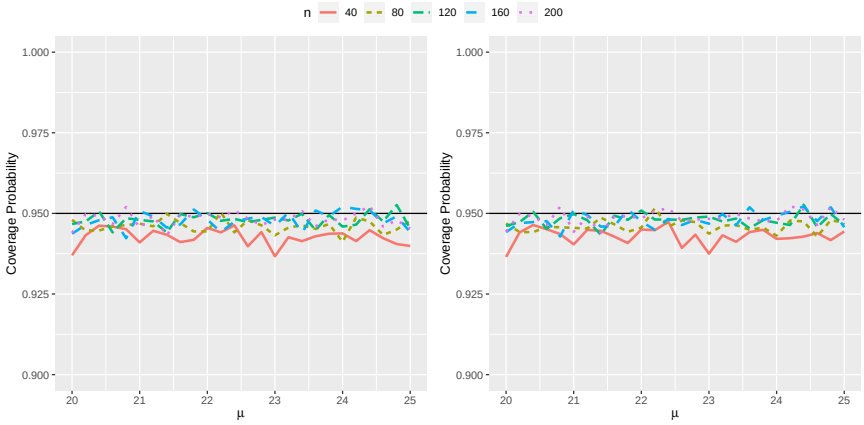


Figure 4.4: Coverage probabilities for 95% confidence interval with $LSL = 20$, $USL = 40$, $T = 26$, $\sigma = 1$ and different $n \in \{40, 80, 120, 160, 200\}$ when $\mu < T$. (left:C_{pn1}, right:C_{pn2})

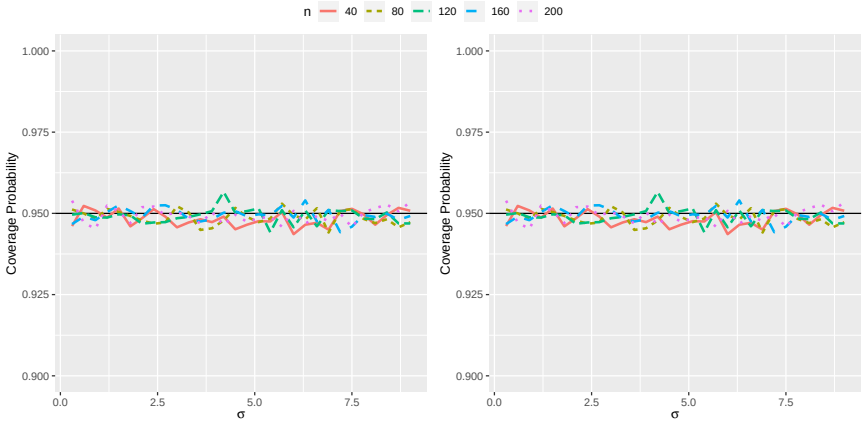


Figure 4.5: Coverage probabilities for 95% confidence interval with $LSL = 20$, $USL = 40$, $T = \mu = 26$ and different σ. $n \in \{40, 80, 120, 160, 200\}$. (left:$C_{pn1}$, right:$C_{pn2}$)

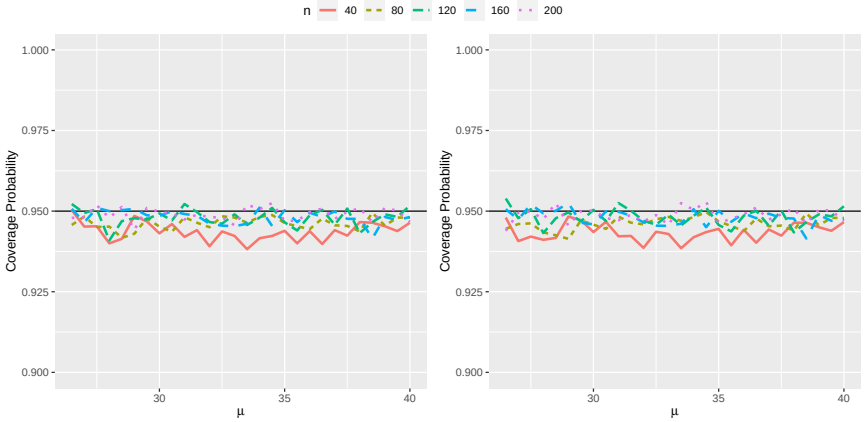


Figure 4.6: Coverage probabilities for 95% confidence interval with $LSL = 20$, $USL = 40$, $T = 26$, $\sigma = 1$ and different $n \in \{40, 80, 120, 160, 200\}$ when $\mu > T$. (left:C_{pn1}, right:C_{pn2})

and let $F(\cdot)$ be its CDF. It is well known that $F(X)$ follows the standard uniform distribution $U(0,1)$, and $\Phi^{-1}(F(X)) \sim N(0,1)$, where $\Phi^{-1}(\cdot)$ is the quantile function of $N(0,1)$. Consequently, our focus shifts to the estimator of $F(\cdot)$ based on the available data.

One well-known nonparametric way to estimate F is to use kernel density estimation. More specifically, once the kernel density estimator $\hat{f}_h(x)$ is obtained with a suitable bandwidth h, the kernel distribution estimator is simply $\hat{F}_h(x) = \int_{-\infty}^{x} \hat{f}_h(t)dt$. This idea has been employed to construct PCIs for non-normal continuous data [114]. Despite its widespread use, kernel estimation is not robust for the inverse transformation because it may exhibit substantial bias near the boundaries of the data range [120]. This bias can result in invalid transformed data, as will be demonstrated later, which limits the applicability of the method proposed by [114] for non-normal data. To overcome these shortcomings, we propose an alternative approach to estimate F using constrained B-spline regression.

4.4.1. THE PROPOSED METHOD

Suppose there is a set of random samples denoted as $\boldsymbol{x} = [x_1, \cdots, x_n]^\top$, representing the quality characteristic X which is drawn from an unknown distribution. To estimate F, a practical nonparametric approach is to use the empirical cumulative distribution function (ECDF) denoted as $F_n(\cdot)$. Specifically, $F_n(x)$ is defined as $\sum_{i=1}^{n} I(x_i \le x)/(n+1)$ [121], where I represents the indicator function. According to the Glivenko-Cantelli theorem, F_n converges to F uniformly with probability 1. However, it is important to note that F_n is a step function, jumping with size $1/(n+1)$

at each x_i. Directly using F_n is impractical and not accurate enough since F is typically continuous. To overcome this limitation, we opt for a continuous approximation of F and introduce a smooth estimator denoted as $\hat{F}_p$.

Regarding $\{F_n(x_i)\}_{i=1}^n$ as responses from a regression model, we have

$$F_n(x_i) = F(x_i) + \epsilon_i, \qquad i = 1, \cdots, n, \tag{4.8}$$

where ϵ_i's are zero-mean error terms. Such a formulation naturally motivates the use of regression techniques to estimate F and perform the inverse transformation. [122] employed spline regression due to its superior flexibility over conventional polynomial or linear approximations [123]. Spline functions strike an optimal balance between local adaptability and global smoothness by utilising carefully constructed basis systems. Although various basis functions, such as I-splines [124], are available, their scope is relatively limited. In contrast, B-splines (BSs) offer the significant advantage of compact local support, meaning that their nonzero values are confined to specific intervals. This localisation not only improves local fitting accuracy but also enables the capture of global functional behaviour through coefficient patterns. As a result, $F(x)$ can be expressed as

$$F(x) = \sum_{l=1}^{L} \beta_l B_l(x), \tag{4.9}$$

where $B_l(\cdot)$ are B-splines (BSs), L is the total number of BSs and $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_L]^\top$ is a vector of unknown parameters. Then, a smooth and monotone estimator of $F(\cdot)$ is produced by employing constrained regression with BSs (CBS) on n pairs $\{(x_i, F_n(x_i))\}_{i=1}^n$.

However, the estimator suffers from three limitations. Firstly, it relies on the unweighted least squares method, which may not be appropriate due to the heteroscedasticity in the regression model. Specifically, as indicated by the central limit theorem, the error term asymptotically follows a normal distribution, namely

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x))). \tag{4.10}$$

Since the variance varies with x, the weighted least squares (WLS) estimation should be employed [125]. Secondly, the estimator deviates from the desired range of $[0, 1]$, which can result in failures during the inverse transformation, leading to infinite values. Thirdly, BSs are generally unsuitable for extrapolation. When the specification limits extend beyond the range of available samples, the estimates of $F(LSL)$ and $F(USL)$ may lack accuracy.

To address these issues and ensure reliable estimators, we propose incorporating ECDF values of the specification limits and transforming $F_n(x_i)$ before applying WLS. Instead of regarding $\{F_n(x_i)\}_{i=1}^n$ as

responses, we propose to use transformed pseudo-observations $y_n(x_i) = \ln[F_n(x_i)/(1-F_n(x_i))]$, $i = 1, \cdots, n$, to estimate $y(x) = \ln[F(x)/(1-F(x))]$. Due to the definition of $F_n(x)$, $y_n(x_i)$ for $i = 1, \cdots, n$ are always well defined. This one-to-one transformation allows us to obtain the estimated CDF $\hat{F}_p(x)$ as follows

$$\hat{F}_p(x) = \frac{e^{\hat{y}(x)}}{e^{\hat{y}(x)} + 1},$$

where $\hat{y}(\cdot)$ represents a smooth estimator of $y(\cdot)$. By employing this approach, the estimate will always fall within the range of $[0, 1]$, ensuring that $\hat{F}_p(\cdot)$ is a valid CDF estimator. Subsequently, the corresponding regression model is given by

$$y_n(x_i) = \ln\left[\frac{F_n(x_i)}{1-F_n(x_i)}\right] = y(x_i) + \epsilon_i, \qquad i = 1, \cdots, n. \tag{4.11}$$

For simplicity, we still use ϵ_i to denote a zero-mean term.

Similar to (4.8), we employ the BSs to approximate $y(x)$ in (4.11). Defining BSs involves dividing the interval over which F is approximated into $m+1$ subintervals using a sequence of knots or breakpoints denoted as τ_m. Specifically,

$$\tau_m = \{\min\{\boldsymbol{x}\} = s_0 < s_1 \cdots < s_{m+1} = \max\{\boldsymbol{x}\}\}.$$

Within each interval, polynomial functions of a specified order k (with a degree of $k-1$) are employed to generate BSs. However, such basis functions have a limitation: they contribute only within the range of $[s_0, s_{m+1}]$ and are zero elsewhere, which becomes an issue when $LSL < s_0$ or $USL > s_{m+1}$. To overcome this, we define a new sequence over $[\min\{LSL, s_0\}, \max\{USL, s_{m+1}\}]$ instead of $[s_0, s_{m+1}]$, which is also denoted as τ_m for simplicity. Two additional corresponding transformed pseudo-observations, $y_n(LSL)$ and $y_n(USL)$, are introduced for regression. Note that $F_n(LSL)$ will be 0 if $LSL < s_0$, and hence we set it as $F_n(s_0)$ to ensure $y_n(LSL)$ will not be $-\infty$. Afterwards, estimates of $F(LSL)$ and $F(USL)$ can be obtained even if LSL and USL are beyond the range of samples.

Regarding the choice of the spline order k, [126] observed that for $k \geq 4$, the resulting expansions tend to be similar, with cubic B-splines (i.e., $k = 4$) being a conventional choice. Therefore, we adopt $k = 4$, which results in a total of $L = m + k$ B-spline basis functions. The parameter m is selected either to achieve the desired smoothness or based on model selection criteria, such as the AIC. Using these B-splines, the regression model is formulated as

$$y(x) = \sum_{l=1}^{L} \beta_l B_l(x).$$

We then apply WLS with the weight function defined by the variance function $G(x)$. By employing (4.10) and the delta method, it can be shown that the errors ϵ_i in (4.11) are independent and asymptotically normal with mean zero and variance $1/\{nF(x_i)[1-F(x_i)]\}$. Consequently, the weight function is set as

$$G(x) = \frac{1}{nF_n(x)[1-F_n(x)]},$$

where $F_n(x)$ denotes the ECDF. The variance-covariance matrix of the error term Σ_e is defined as a diagonal matrix with the principal diagonal elements given by

$$[G(x_1), \cdots, G(x_n), G(LSL), G(USL)].$$

As previously mentioned, if $LSL < s_0$, $F_n(LSL)$ is set as $F_n(s_0)$ to prevent $-\infty$. Additionally, we enforce a monotonicity constraint to ensure a reliable estimator, given that F, as a CDF, is inherently non-decreasing, akin to the approach in [122]. Following [127], a sufficient condition for $\sum_{l=1}^{L} \beta_l B_l(x)$ to be non-decreasing is $\beta_{l-1} \le \beta_l$ for $l = 2, \cdots, L$. Denoting $\mathcal{B}^L$ as the set of all $[\beta_1, \cdots, \beta_L] \in \mathcal{R}^L$ satisfying $\beta_{l-1} \le \beta_l$, the estimates of $\boldsymbol{\beta}$ via WLS can be obtained by

$$\begin{aligned}\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}^L} & \sum_{i=1}^{n} \frac{(y_n(x_i) - \sum_{l=1}^{L} \beta_l B_l(x_i))^2}{G(x_i)} \\ & + \frac{(y_n(USL) - \sum_{l=1}^{L} \beta_l B_l(USL))^2}{G(USL)} \\ & + \frac{(y_n(LSL) - \sum_{l=1}^{L} \beta_l B_l(LSL))^2}{G(LSL)}.\end{aligned} \tag{4.12}$$

The final smooth and monotone estimator $\hat{F}_p(\cdot)$ is then given by

$$\hat{F}_p(x) = \frac{\exp(\sum_{l=1}^{L} \hat{\beta}_l B_l(x))}{\exp(\sum_{l=1}^{L} \hat{\beta}_l B_l(x)) + 1}. \tag{4.13}$$

Based on $\Phi^{-1}(\hat{F}_p(x_i))$, $i = 1, \cdots, n$, $\Phi^{-1}(\hat{F}_p(LSL))$, $\Phi^{-1}(\hat{F}_p(T))$ and $\Phi^{-1}(\hat{F}_p(USL))$, the estimates of two proposed PCIs can be readily obtained.

It is worth noting that [112] also applied BSs and inverse transformation for non-normal data. Their method sought a transformation $s(\cdot)$, with a form similar to (4.9) and involving BSs, such that $s(X)$ follows a standard normal distribution. Specifically, they considered the relationship $q_z = s(q_x)$, where q_x and q_z denote the percentiles of the distribution

F and the standard normal distribution, respectively, with the ECDF F_n substituting for F when it is unknown. However, because $s(\cdot)$ should be non-decreasing and its range must lie within $[0, 1]$, the lack of such constraints in their model leads to potential issues. In contrast, our method explicitly incorporates these constraints, thereby ensuring a more reliable and appropriate inverse transformation.

To analyse the efficiency of the proposed estimator, asymptotic properties are studied here. This analysis becomes particularly insightful as the sample size grows, offering a deeper understanding of the estimator's behaviour and facilitating the identification of corresponding confidence regions [128]. Here, we consider the asymptotic normality of $\hat{\boldsymbol{\beta}}$. Assuming that the true underlying function is fully represented by the chosen B-spline basis, we derive the asymptotic distribution of $\hat{y}$. Define $||\phi||^2 = \mathbb{E}[\phi(x)^2]$, the rate of convergence of our estimator is also shown in the following theorem and corollary. Assumptions and proofs are presented in Section 7.3.

Theorem 1. *Under assumptions (1)-(5) in Section 7.3, it follows that*

$$\frac{\hat{y}(x) - y(x)}{\sigma(x)} \xrightarrow{d} N(0, 1),$$

$$||\hat{y} - y|| = O_p(m^{-k} + \sqrt{m/n}),$$

where $\sigma^2(x) = \boldsymbol{B}(x)^\top (\boldsymbol{b}^\top \mathbf{V} \boldsymbol{b})^{-1} \boldsymbol{B}(x)$, $\boldsymbol{B}(x) = [B_1(x), \cdots, B_L(x)]^\top$, $\boldsymbol{b} = [\mathbf{b}_1, \cdots, \mathbf{b}_L]$, $\mathbf{b}_l = [B_l(x_1), \cdots, B_l(x_n), B_l(LSL), B_l(USL)]^\top$, *and* $\mathbf{V} = \Sigma_e^{-1}$. *Here,* $W_n = O_p(a_n)$ *means given a sequence of positive number* a_n,

$$\lim_{c \to \infty} \limsup_n P(|W_n| \geq c a_n) = 0,$$

holds for all $c > 0$.

Corollary 2. *Under assumptions (1)-(5) in Section 7.3, it follows that*

$$\frac{\hat{F}_p(x) - F(x)}{\sigma_F(x)} \xrightarrow{d} N(0, 1),$$

where

$$\sigma_F^2(x) = F(x)(1 - F(x))\sigma^2(x).$$

In addition, the convergence rate of $\hat{F}_p$ *is not slower than* $\hat{y}$*'s convergence rate when* n *is sufficiently large.*

4.4.2. COMPARISON STUDIES

In this subsection, we conduct some simulation studies to assess the performance of our proposed estimator, denoted as $\hat{F}_p$. Furthermore,

we incorporate several existing methods into our comparative studies, including the kernel estimator $\hat{F}_h$, ECDF estimator $\hat{F}_e$, and the estimator based on constrained regression with BSs in [122] $\hat{F}_b$. We utilize specific `R` packages, namely `splines2` for BSs [129], `Rsolnp` for estimating $\boldsymbol{\beta}$ [130], and `ks` for the kernel estimator of CDF [131].

We choose three different distributions for X, including the Weibull distribution $WB(k, 1)$ with shape k and scale 1, the log-normal distribution $LN(\mu, 1)$ with location μ and scale 1, and the gamma distribution $GA(\theta, 2)$ with shape θ and rate 2. These three distributions are commonly used and recognised as applicable models for the process data [115]. The USL and LSL are set as the 0.99 and 0.01 percentiles of the true distribution, and the target value T is set as $T = M = (LSL + USL)/2$. Hence, the inverse transformation will lead to asymmetric tolerances. Taking C_{pn1} as an example, the corresponding index values based on estimated CDF are denoted as $\hat{C}^h_{pn1}$, $\hat{C}^e_{pn1}$, $\hat{C}^b_{pn1}$ and $\hat{C}^p_{pn1}$. We repeat the simulation 1000 times by independently regenerating the data for two different scenarios. Various estimators are evaluated by two kinds of log mean squared error (LMSE), which are given by

$$\text{LMSE}_{\text{CDF}} = \frac{1}{1000}\sum_{i=1}^{1000}\log_{10}\left[\frac{1}{n}\sum_{j=1}^{n}(\hat{F}(x_{ij}) - F(x_{ij}))^2\right],$$

$$\text{LMSE}_{\text{PCI}} = \frac{1}{1000}\sum_{i=1}^{1000}\log_{10}\left[(\hat{C}^i_{pn1} - C^i_{pn1})^2\right],$$

where $\hat{F}(\cdot)$ is the estimator, $F(\cdot)$ is the true distribution function, and $\hat{C}_{pn1}$, C_{pn1} are corresponding estimated and true PCI values based on them.

In the initial phase of our comparison studies, we set $k = 1$, $\mu = 0$, and $\theta = 1$, estimating F across 20 different sample sizes ranging from 50 to 1000. The average LMSE_{CDF} is illustrated in Figure 4.7, where our method stands out as the most accurate among all examined approaches. This superiority can be attributed to two pivotal factors. Firstly, our estimation directly targets the CDF using B-splines via ECDF values. Secondly, our utilisation of transformed pseudo-observations ensures a performance advantage over the estimator based on CBS. Specifically, the CBS-based $\hat{F}_b$ occasionally yields estimates outside the valid range of $[0, 1]$, resulting in elevated LMSE values. In contrast, our proposed estimator guarantees that the estimated CDF remains within the valid range. Additionally, despite $\hat{F}_p$ and $\hat{F}_e$ both being based on the ECDF, our method, employing the WLS method, provides a continuous estimator. This contributes to higher accuracy compared to using the ECDF directly, as evident in the lower LMSE values for our proposed estimator.

We extend our evaluation to assess the performance of estimators concerning PCIs, and the results are illustrated in Figure 4.8. Notably,

$\hat{C}^b_{pn1}$ encounters challenges across all three distributions considered. CBS, being insensitive to the range of the CDF, leads to infinite values after data transformation, rendering the method impractical for PCI calculations. This issue is particularly pronounced in the case of the gamma distribution $GA(1, 2)$, as detailed in Table 4.2, where CBS fails in the majority of replications across various sample sizes n. The results underscore the significance of employing transformed pseudo-observations in our proposed method. Consequently, for LMSE evaluation, we only consider $\hat{C}^h_{pn1}$, $\hat{C}^e_{pn1}$, and $\hat{C}^p_{pn1}$. Notably, $\hat{C}^h_{pn1}$ exhibits some missing points, indicating potential limitations in small sample sizes, particularly when USL surpasses the sample range, leading to $\hat{F}_k(USL) = 1$ and transformed USL being $+\infty$, rendering it unsuitable for evaluation. As a comparison, our proposed $\hat{C}^p_{pn1}$ demonstrates robustness, with estimates based on (4.13) consistently falling within the range of $(0, 1)$. Furthermore, the results suggest that our method provides reliable estimates at specification limits, even if they are beyond the range of available data.

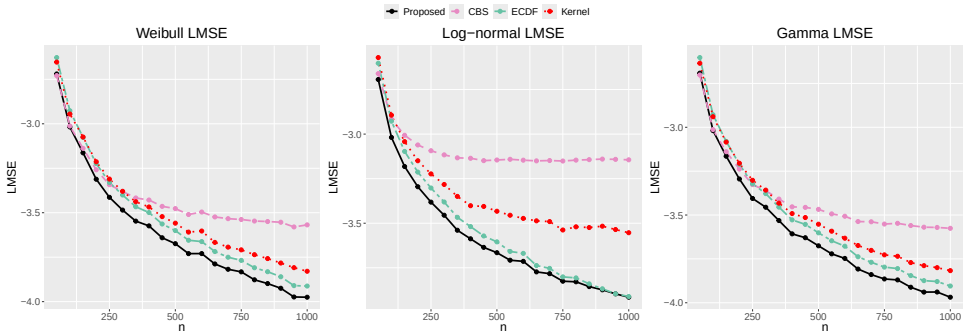


Figure 4.7: The average LMSE_{CDF} of different methods based on 20 different sample sizes $n \in \{50, 100, \cdots, 1000\}$ for $WB(1, 1)$, $LN(0, 1)$ and $GA(1, 2)$.

In the second scenario, we set the sample size to 100 and vary the values of k, μ, θ as $0.5,\ 1,\ \cdots,\ 5$. As shown previously, our method can estimate the CDF accurately. Thus, four methods are only evaluated in terms of PCIs. Similar to the first scenario, the LMSEs based on PCI values of all methods except CBS are presented in Figure 4.9. The results suggest comparable performances. However, the kernel estimation, like CBS, also encounters difficulties in dealing with the inverse transformation. Taking $GA(\theta, 2)$ as an example, the number of failures when calculating PCIs in 1000 replications for different θ are presented in Table 4.3.

While the kernel and CBS methods can provide good estimators of the

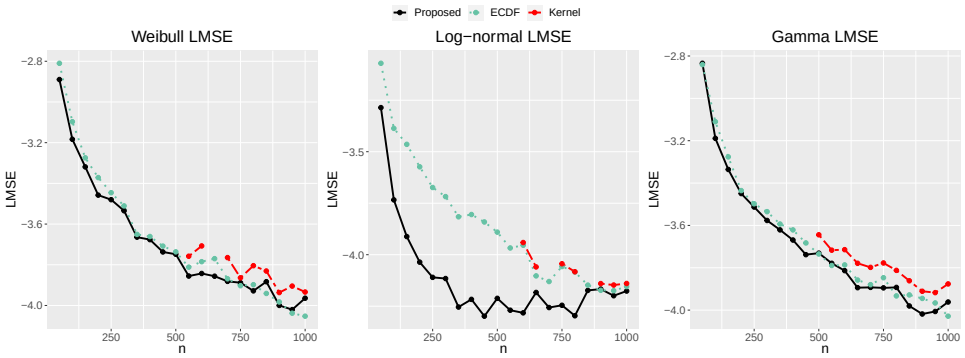


Figure 4.8: The average LMSE_{PCI} of different methods based on 20 different sample sizes $n \in \{50, 100, \cdots, 1000\}$ for $WB(1, 1)$, $LN(0, 1)$ and $GA(1, 2)$.

Table 4.2: Number of failures using $\hat{F}_b$ to calculate PCI values with different n for $GA(1, 2)$.

The number of failures of CBS for $GA(1, 2)$										
n	50	100	150	200	250	300	350	400	450	500
Number	383	624	714	796	851	866	921	941	961	979
n	550	600	650	700	750	800	850	900	950	1000
Number	983	982	992	987	992	998	1000	999	996	1000

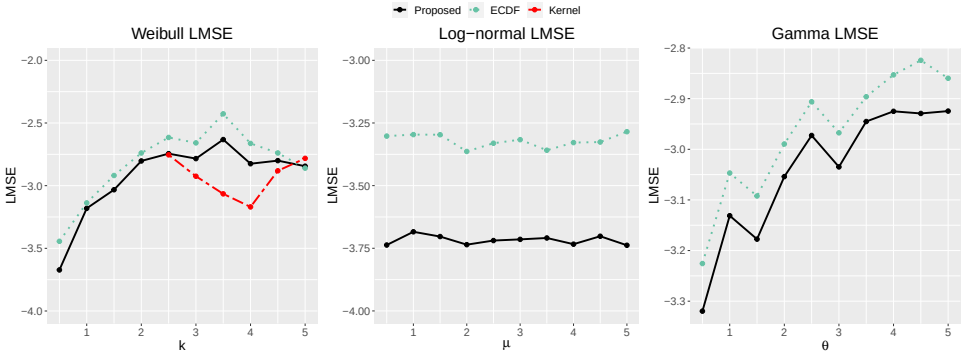


Figure 4.9: The average LMSE_{PCI} based on different methods for $WB(k, 1)$, $LN(\mu, 1)$ and $GA(\theta, 2)$ with different parameters ranging from 0.5 to 5 and the same sample size of 100.

Table 4.3: Numbers of failures using $\hat{F}_k$ and $\hat{F}_b$ to calculate PCI values for $GA(\theta, 2)$ with $n = 100$.

Numbers of failures of kernel and CBS for $GA(\theta, 2)$										
θ	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
kernel	259	133	103	56	36	22	11	14	7	7
CBS	254	269	229	227	224	206	236	241	204	217

CDF, it seems that sometimes they yield undesirable estimates for PCIs at some $\boldsymbol{x}$ or the specification limits. The reasons for the failure of these two methods appear to be different. The kernel estimator fails when the sample size is small or the distribution has a heavy tail. In such cases, the specification limits often fall beyond the range of the available data, causing the kernel estimation to yield 0 or 1. On the other hand, the number of failures for $\hat{F}_b$ increases with larger sample sizes. In contrast, our proposed method incorporates transformed pseudo-observations and is based on a new sequence of knots that encompasses the specification limits, making it more suitable for the inverse transformation.

4.5. ILLUSTRATIVE EXAMPLE

In this section, we apply the proposed two PCIs along with the normalised method to the weight (in grams) of powder juice packages. According to the standards established by the National Institute of Metrology, Quality and Technology in Brazil, *LSL* and *USL* are assumed as 18 and 22, respectively [102]. The target value is the midpoint between the two specification limits, which is 20. A total of 50 samples from two different processes are presented in Table 4.4.

We commence our analysis by examining the data from the first process, denoted as P_1. The left histogram in Figure 4.10 shows that the original samples follow a skewed distribution, which motivates the use of the transformation method. We start by computing the ECDF from the data of process P_1. The transformed pseudo-observations, defined in (4.11), are then used as the responses in the spline regression model. Consistent with our comparison study, knots are selected at the $1/11, \cdots, 10/11$ quantiles of the P_1 data to construct the BS basis functions. The WLS approach, as described in (4.12), is used to estimate the coefficients $\hat{\boldsymbol{\beta}}$. The corresponding estimated CDF, $\hat{F}_p$, given in (4.13) and shown in Figure 4.11, closely aligns with the ECDF points, capturing the overall trend effectively. Finally, the data from P_1 is transformed using $\Phi^{-1}(\hat{F}_p(\cdot))$ to achieve normality.

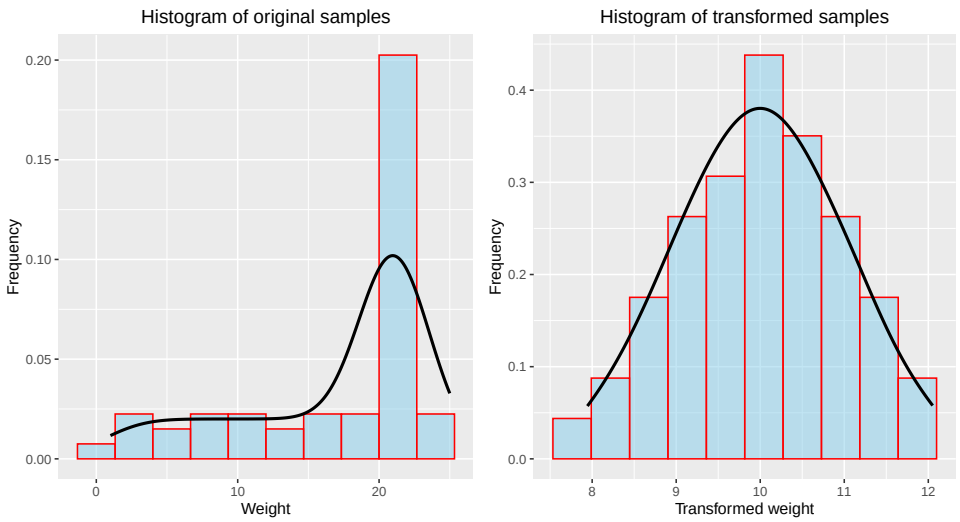


Figure 4.10: Histograms of original samples from P_1 (left) and transformed samples (right).

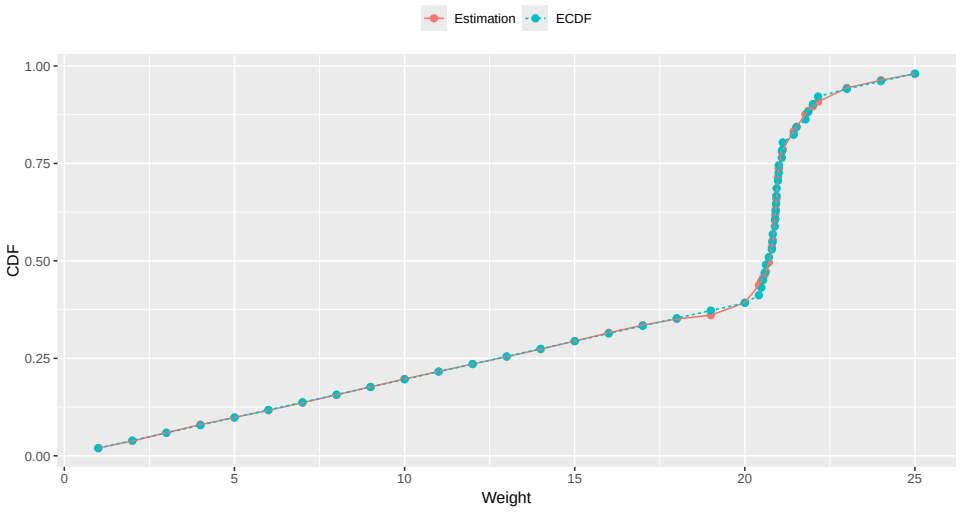


Figure 4.11: Estimated CDF via the proposed method and the ECDF based on data from P_1.

Table 4.4: Weight (in grams) of powder juice packages from two different processes.

P_1	20.587	21.784	21.088	20.997	21.100
	22.155	21.116	20.707	20.413	20.822
	20.883	20.930	20.908	20.897	20.486
	20.935	21.867	20.814	20.795	21.520
	20.537	21.438	20.621	20.975	20.919
P_2	21.583	21.813	22.025	20.892	20.241
	21.816	21.232	21.730	20.529	21.435
	21.106	20.519	21.263	20.684	21.233
	19.624	21.150	20.962	21.024	20.316
	21.942	21.495	20.819	20.973	21.115

The right panel of Figure 4.10 suggests that the transformed data approximately follows a normal distribution. In addition, we perform a Shapiro-Wilk test, obtaining $p = 5.518 \times 10^{-7}$ before transformation and $p = 0.999$ afterwards. Subsequently, the new specification limits LSL, T, and USL are determined as 9.617, 9.728, and 11.264, respectively, indicating that the new target is not centred between the new limits. Therefore, we apply the proposed PCI measures to evaluate process P_1. Similarly, we transform 25 observations from the second process, P_2, leading to asymmetric tolerances. The Shapiro-Wilk test yields a p-value of 7.682×10^{-7} before transformation, which increases to 0.995 afterwards.

The results of applying the proposed PCIs to the two processes are summarised in Table 4.5. These findings indicate minimal differences between the two PCI measures, both consistently suggesting that the first process exhibits higher quality. It is worth noting that all upper bounds of NC based on our PCIs are 100% for both processes. Meanwhile, PCI values are very close to 0, indicating relatively poor quality products. As highlighted by [102], neither process meets the requirements, aligning with the conclusions drawn from the proposed PCIs. In summary, our proposed PCIs, along with the normalised method, effectively account for both non-normal data and asymmetric tolerances. These indices serve as valuable tools for evaluating production processes in quality.

4.6. CONCLUSIONS

This study addressed two critical issues simultaneously: asymmetric tolerances and a non-normal quality characteristic, offering a compre-

Table 4.5: Points estimates and 95% confidence intervals of two proposed PCIs, along with the corresponding upper bounds of non-conforming (NC) percentage.

	PCIs	Estimate	95% confidence interval	Upper bound of NC
P_1	C_{pn1}	0.032	[0.023, 0.040]	100%
	C_{pn2}	0.030	[0.024, 0.037]	100%
P_2	C_{pn1}	0.050	[0.035, 0.065]	100%
	C_{pn2}	0.048	[0.037, 0.059]	100%

hensive evaluation of an actual manufacturing process. While most existing articles focus solely on one of these issues, we handle them at the same time. Initially, we summarised the essential properties of PCIs in the presence of asymmetric tolerances and normal data. Notably, we found that none of the currently available PCIs satisfy all these desired properties. To bridge this gap, we introduced two novel classes of PCIs, complete with parametric estimation procedures and asymptotic confidence limits. A comparative analysis with existing PCIs was then presented, encompassing estimators and their relationships to process centring and yield. Transitioning to the realm of non-normal distributed characteristics, we employed constrained B-spline regression on transformed pseudo-observations to estimate the CDF. Additionally, we proposed a tailored inverse transformation to normalise the data. Demonstrating the efficacy of our approach, we conducted simulation studies that underscored its superiority. In the final phase of our investigation, we successfully applied the proposed PCIs and the transformation method to a real data set, providing practical illustrations of their applicability.

5

MODELING JOINT DISTRIBUTIONS FOR DEPENDENT POISSON DATA

5.1. INTRODUCTION

Poisson distributions are commonly used to model event occurrences over a fixed period, such as random arrival streams [132] and customer counts [133]. In situations involving multiple dependent variables, however, the multivariate Poisson distribution remains under-explored and inadequately defined in the literature. This is surprising given the practical significance of multivariate Poisson variables, which arise in various real-world applications. For example, the monthly number of notifications of hepatitis C [134], the crash data [135], and the surveillance of legionellosis [136] all involve multiple dependent Poisson variables. A key challenge in specifying these models lies in accurately defining the joint distribution while incorporating marginal distributions and adequately capturing the dependence structures among the variables. Existing models often fail to address this challenge with sufficient flexibility. In the following, we review some common models for multivariate Poisson variables, focusing on their dependence structures, which are typically measured by the Pearson correlation coefficient.

Suppose we have d dependent Poisson variables, $X_1, \cdots, X_d$, with positive marginal parameters $\lambda_1, \cdots, \lambda_d$. A widely used approach to model the joint distribution is the multivariate reduction (MR) technique, which relies on the convolution property of independent Poisson random variables. In this framework, each X_i can be expressed as $X_i = Y_i + Y$ for $i = 1, \cdots, d$, where Y_i and Y are mutually independent Poisson random variables with positive parameters α_i, $i = 1 \cdots, d$, and β. Because the sum of independent Poisson random variables is again Poisson distributed [137], X_i follows a Poisson distribution with the marginal mean $\lambda_i = \alpha_i + \beta$. When $d = 2$, the model simplifies to the classical bivariate Poisson distribution described in [138], commonly referred to as the trivariate reduction (TR) model. Various researchers have further developed the multivariate extension of this model for Poisson variables, including [139] and [140]. In the MR model, the correlation coefficient $\rho^X_{i,j}$ between X_i and X_j is given by $\rho^X_{i,j} = \beta/\sqrt{\lambda_i \lambda_j}$. As seen, one notable limitation is that it imposes a structural constraint on the joint distribution, restricting the correlation between the Poisson variables to be non-negative, which may limit its applicability in scenarios where negative correlations are observed [141].

To address this limitation, [142] and [143] extended the TR model to allow for joint distributions with negative correlation coefficients. This extension relies on the comonotonic random variables, which are represented by the same uniform random variable defined over the interval $[0, 1]$. Specifically, let $F^{-1}_\lambda(y) := \inf\{x \in \mathbb{N} : F_\lambda(x) > y\}$ for $y \in [0, 1]$, where F_λ denotes the cumulative distribution function (CDF) of a Poisson random variable with parameter λ and $\mathbb{N}$ is the set of non-negative integers. The trivariate reduction extension (TREx) model

proposed for two dependent Poisson variables is given by

$$\begin{aligned} X_1 &= Y_1 + F^{-1}_{\gamma\lambda_1}(U), \quad X_2 = Y_2 + F^{-1}_{\gamma\lambda_2}(U) \quad \text{if } \rho^X_{1,2} > 0, \\ X_1 &= Y_1 + F^{-1}_{\gamma\lambda_1}(U), \quad X_2 = Y_2 + F^{-1}_{\gamma\lambda_2}(1-U) \quad \text{if } \rho^X_{1,2} < 0, \end{aligned} \tag{5.1}$$

where U is uniformly distributed over $[0, 1]$, Y_1 and Y_2 are independent Poisson variables with marginal expectations $(1-\gamma)\lambda_1$ and $(1-\gamma)\lambda_2$, and $\gamma \in [0, 1]$ controls the strength of correlation. As noted by [144], the range of attainable correlation coefficients given λ_1 and λ_2 between two Poisson variables is

$$[\text{Corr}(F^{-1}_{\lambda_1}(U), F^{-1}_{\lambda_2}(1-U)), \text{Corr}(F^{-1}_{\lambda_1}(U), F^{-1}_{\lambda_2}(U))].$$

Thus, the joint CDF defined by Model (5.1) is capable of capturing the full range of attainable correlation coefficients as γ varies from 0 to 1. Recent studies, such as [145] and [146], have extended this model to the multivariate setting. However, limitations remain when employing comonotonic variables. In particular, when modelling more than two Poisson variables, the limited number of parameters defining the CDF restricts the range of dependence that can be captured. Moreover, this approach does not accommodate negative correlations effectively, rendering it unsuitable for scenarios where multiple Poisson variables exhibit negative dependence.

Several alternative methods have been proposed for modelling multivariate Poisson variables, among which copula-based approaches are particularly flexible for constructing joint distributions from specified marginals [147]. Copula methods have attracted significant attention in the literature, as demonstrated by recent studies such as [148], [149], [150], and [151]. According to Sklar's theorem [152], any joint distribution can be decomposed into its marginal distributions and a copula function that captures the dependence structure. Among various copula families, the Gaussian copula is frequently employed due to its versatility in modelling dependencies among Poisson variables [153]. However, the Gaussian copula may not adequately capture certain dependence structures, such as those characterized by high upper-tail dependence, and identifying a suitable parametric family can be challenging. Moreover, while the copula is unique when all marginal distributions $F_1, \ldots, F_d$ are continuous [e.g., 154, Theorem 2.4.1], this uniqueness does not extend to cases with discrete marginals. Alternatively, multivariate Poisson distributions can be constructed using mixture models, as discussed in [155] and [156]. For a comprehensive review of multivariate Poisson models, see [157] and [158].

Although several existing methods offer considerable flexibility in modelling dependencies, the resulting distributions are often complex and challenging to interpret in practice. A key unresolved issue in the

literature is the development of an intuitive and interpretable multivariate Poisson model that flexibly captures the underlying dependence structure. In response to this challenge, we revisit traditional MR techniques and propose an innovative approach that integrates the MR model with a Gaussian copula, rather than relying solely on the convolution of independent Poisson variables. As we will demonstrate, the proposed model not only provides enhanced flexibility in modelling joint distributions and capturing complex dependencies but also yields a clear and intuitive interpretation of both the marginal distributions and the dependence structure. This dual advantage significantly improves the theoretical robustness and practical applicability of our approach. We rigorously investigate the probability properties of the model and introduce a series of propositions to highlight the novelty and contributions of our approach. Some methods, including maximum likelihood (ML) estimation and inference function for margins [159], are explored, and a tailored expectation-maximization (EM) algorithm is proposed to obtain estimators.

The remainder of this chapter is organized as follows. In Section 5.2, we present the formulation of the proposed model and provide a detailed exploration of its probabilistic properties. Section 5.3 extends this discussion and demonstrates the novelty of our approach by comparing it with existing models. In Section 5.4, we discuss the methods for parameter estimation and the corresponding asymptotic analysis. Section 5.5 provides numerical simulations to illustrate the advantages of our model, while Section 5.6 presents an application of the model using bike rental data. Finally, concluding remarks are presented in Section 5.7. Technical proofs and additional simulation results are given in Section 7.4.

5.2. PROPOSED MODEL

Consider d dependent Poisson random variables, $X_1, \ldots, X_d$, each with positive marginal parameters $\lambda_1, \ldots, \lambda_d$. As previously mentioned, while the TR and TREx models rely on independent Poisson variables to model dependence, they struggle to accommodate negative correlations. Although the Gaussian copula can capture a broader range of dependence structures, it is ineffective to model tail dependence and extreme co-movements, which may result in an inadequate representation of joint distributions. To address these shortcomings, we propose a generalized model that integrates the strengths of both approaches. Our model incorporates multiple dependent normal variables along with additional parameters, and hence accommodates a wider range of distributions and captures a more comprehensive dependence structure.

Let $\mathcal{P}(\lambda)$ denote the univariate Poisson distribution with parameter $\lambda > 0$. By extension, we define $\mathcal{P}(0)$ to be the Dirac distribution at 0, i.e., the distribution of a random variable that is constant with the value 0.

The corresponding CDF is denoted by F_0, which is such that $F_0(x) = \mathbb{1}_{x \geq 0}$, with an indicator function $\mathbb{1}$. Let Φ denote the CDF of the standard normal distribution. Let $\boldsymbol{Z} = [Z_1, \cdots, Z_d]^\top$ follow a multivariate normal distribution $\mathcal{N}_d(\mathbf{0}, \mathbf{R})$ with a correlation matrix $\mathbf{R} = [\rho_{i,j}]_{d \times d}$, assumed to be invertible. This means that $Z_1, \ldots, Z_d$ are standard normal random variables and $\rho_{i,j} \in [-1, 1]$ are the correlations between Z_i and Z_j. We propose our multivariate Poisson model as

$$\boldsymbol{X} = [X_1, \cdots, X_d]^\top = [Y_1 + V_1, \cdots, Y_d + V_d]^\top, \tag{5.2}$$

where $Y_1, \ldots, Y_d$ are mutually independent Poisson variables with parameters $\alpha_i \geq 0$, independent of $\boldsymbol{Z}$, and $V_i := F^{-1}_{\beta_i}(\Phi(Z_i))$ with $\beta_i \geq 0$. As a particular case, remark that with our conventions, $Y_i := 0$ if $\alpha_i = 0$ and $V_i := 0$ if $\beta_i = 0$. In this sense, our proposed model introduces latent Poisson $(Y_1, \ldots, Y_d)$ and Normal $(Z_1, \ldots, Z_d)$ variables, allowing us to define the joint distribution based on a Gaussian copula while preserving the probabilistic structure. This enables the accurate modelling of discrete Poisson variables and hence extends the copula framework effectively to the discrete setting.

It is well known that $\Phi(Z_i)$ follows the standard uniform distribution $\mathcal{U}(0, 1)$ and $F^{-1}_{\beta_i}(\Phi(Z_i)) \sim \mathcal{P}(\beta_i)$. Since Y_i and Z_i are independent, the marginal distribution of $\boldsymbol{X}$ is also Poisson with rate $\alpha_i + \beta_i$, i.e. $X_i \sim \mathcal{P}(\lambda_i)$ where $\lambda_i := \alpha_i + \beta_i$. Let $\rho^X_{i,j}$ be the correlation coefficient between X_i and X_j, given by

$$\rho^X_{i,j} = \frac{\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]}{\sqrt{(\alpha_i + \beta_i)(\alpha_j + \beta_j)}} = \frac{\mathbb{E}[F^{-1}_{\beta_i}(\Phi(Z_i))F^{-1}_{\beta_j}(\Phi(Z_j))] - \beta_i\beta_j}{\sqrt{(\alpha_i + \beta_i)(\alpha_j + \beta_j)}}. \tag{5.3}$$

The presence of normal variables in our model allows for pairwise correlations spanning the entire attainable range. When $\beta_i = \beta_j = 0$ or $\rho_{i,j} = 0$, X_i and X_j are independent, whereas the condition $\alpha_i = \alpha_j = 0$ and $\rho_{i,j} = 1$ produces the maximal correlation. Between these two extremes, varying these five parameters results in a wide range of attainable dependence structures, which we explore in greater detail in Section 5.3.

The proposed model reduces to the TREx model when the following two conditions are satisfied: $\rho_{i,j} \in \{-1, 1\}$ for all i, j; and there exist $\gamma \in [0, 1]$ such that $\alpha_i = (1 - \gamma)\lambda_i$ and $\beta_i = \gamma\lambda_i$. Additionally, it encompasses the original MR models when $\beta_1 = \beta_2 = \cdots = \beta_d$ and $\boldsymbol{R}$ is the identity matrix, ensuring the generalization of traditional methods. Furthermore, in the special case where $\alpha_1 = \alpha_2 = \cdots = \alpha_d = 0$, the model simplifies to a Gaussian copula. This flexibility highlights the strength of our approach, as it not only subsumes existing models but also extends them by allowing for more complex dependence structures with the same marginal distributions. In this way, the proposed model provides greater generality without sacrificing the features of established convolution models. More detailed comparisons and discussions are provided in Section 5.3.

5.2.1. DISTRIBUTION FUNCTION AND PROBABILITY MASS FUNCTION

The CDF of $\boldsymbol{V} = [V_1, \cdots, V_d]^\top$ can be expressed as

$$\begin{aligned}
F_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) &= \Pr(V_1 \le v_1, \cdots, V_d \le v_d) \\
&= \Pr(F_{\beta_1}^{-1}(\Phi(Z_1)) \le v_1, \cdots, F_{\beta_d}^{-1}(\Phi(Z_d)) \le v_d) \\
&= \Pr(Z_1 \le \Phi^{-1}(F_{\beta_1}(v_1)), \cdots, Z_d \le \Phi^{-1}(F_{\beta_d}(v_d))) \\
&= \Phi_{\mathbf{R}}(\Phi^{-1}(F_{\beta_1}(v_1)), \cdots, \Phi^{-1}(F_{\beta_d}(v_d))) \\
&= \int_{-\infty}^{\Phi^{-1}(F_{\beta_1}(v_1))} \cdots \int_{-\infty}^{\Phi^{-1}(F_{\beta_d}(v_d))} \frac{1}{(2\pi)^{d/2}|\mathbf{R}|^{1/2}} \exp\Big(-\frac{\boldsymbol{z}^\top \mathbf{R}^{-1} \boldsymbol{z}}{2}\Big) d\boldsymbol{z},
\end{aligned}$$

where $\Phi_{\mathbf{R}}$ denotes the CDF of $\mathcal{N}_d(\mathbf{0}, \mathbf{R})$, $\boldsymbol{z} = [z_1, \cdots, z_d]^\top$, $\boldsymbol{v} = [v_1, \cdots, v_d]^\top \in \mathbb{N}^d$ and $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_d]^\top$. Note that this equation also holds whenever some of the β_i are zero, since $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = +\infty$, with the convention that $\int_{-\infty}^{-\infty} g = 0$ for any function g. Moreover, the probability mass function (PMF) of $\boldsymbol{V}$ is

$$\begin{aligned}
f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) &= \Pr(V_1 = v_1, \cdots, V_d = v_d) \\
&= \Pr(F_{\beta_1}^{-1}(\Phi(Z_1)) = v_1, \cdots, F_{\beta_d}^{-1}(\Phi(Z_d)) = v_d) \\
&= \Pr(F_{\beta_i}(v_i - 1) < \Phi(Z_i) \le F_{\beta_i}(v_i), i = 1, \cdots, d), \\
&= \Pr(\Phi^{-1}(F_{\beta_i}(v_i - 1)) < Z_i \le \Phi^{-1}(F_{\beta_i}(v_i)), i = 1, \cdots, d) \\
&= \int_{\Phi^{-1}(F_{\beta_1}(v_1 - 1))}^{\Phi^{-1}(F_{\beta_1}(v_1))} \cdots \int_{\Phi^{-1}(F_{\beta_d}(v_d - 1))}^{\Phi^{-1}(F_{\beta_d}(v_d))} \frac{1}{(2\pi)^{d/2}|\mathbf{R}|^{1/2}} \exp\Big(-\frac{\boldsymbol{z}^\top \mathbf{R}^{-1} \boldsymbol{z}}{2}\Big) d\boldsymbol{z}.
\end{aligned}$$

Note that $F_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R})$ is piecewise constant in $\boldsymbol{v}$ since it is the CDF of a multivariate discrete random vector. Therefore, some additional formulas have to be adapted to deal with the discontinuities. Denote $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho})$ as all the unknown parameters, where $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_d]^\top$ and $\boldsymbol{\rho} = [\rho_{1,2}, \rho_{1,3}, \cdots, \rho_{d-1,d}]^\top$. Based on the above expressions, the CDF of the proposed model for any $\boldsymbol{x} = [x_1, \ldots, x_d]^\top \in \mathbb{N}^d$ can be derived by conditioning on $\boldsymbol{V}$ as follows

$$\begin{aligned}
F_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \Pr(X_1 \le x_1, \cdots, X_d \le x_d) \\
&= \Pr(Y_1 + V_1 \le x_1, \cdots, Y_d + V_d \le x_d) \\
&= \sum_{v_1=0}^{x_1} \cdots \sum_{v_d=0}^{x_d} \Pr(Y_1 + v_1 \le x_1, \cdots, Y_d + v_d \le x_d) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) \\
&= \sum_{v_1=0}^{x_1} \cdots \sum_{v_d=0}^{x_d} \prod_{i=1}^{d} F_{\alpha_i}(x_i - v_i) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) \\
&= \sum_{v_1=0}^{x_1} \cdots \sum_{v_d=0}^{x_d} \prod_{i=1}^{d} f_{\alpha_i}(x_i - v_i) F_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}),
\end{aligned} \tag{5.4}$$

where f_λ is the univariate Poisson PMF given by $f_\lambda(x) = e^{-\lambda}\lambda^x/x!$ for $x \in \mathbb{N}$ if $\lambda > 0$ and $f_\lambda(x) = \mathbb{I}_{x=0}$ if $\lambda = 0$. In addition, the PMF of the proposed model is

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{v_1=0}^{x_1} \cdots \sum_{v_d=0}^{x_d} \prod_{i=1}^{d} f_{\alpha_i}(x_i - v_i) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}). \tag{5.5}$$

Recurrence relations can be derived to simplify the calculation of the PMF. In classical MR models, the joint distribution is constructed using a single variable Y, which allows for conditional independence and facilitates the derivation of recurrence relations. However, the proposed model introduces more complex structures to increase flexibility, making the computation of recurrence relations less straightforward. Recall that the univariate Poisson distribution has the simple recurrence relation $f_\lambda(x) = (\lambda/x) f_\lambda(x-1)$, for $x \in \{1, 2, \cdots\}$, and $f_\lambda(0) = e^{-\lambda}$. The multivariate PMF of $\boldsymbol{X}$ satisfies a similar relationship given by

$$\begin{aligned}
&f_{\boldsymbol{\theta}}(x_1 - 1, x_2, \cdots, x_d) \\
&= \sum_{v_1=0}^{x_1} \cdots \sum_{v_d=0}^{x_d} f_{\alpha_1}(x_1 - 1 - v_1) \prod_{i=2}^{d} f_{\alpha_i}(x_i - v_i) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) \\
&= \sum_{v_1=0}^{x_1} \cdots \sum_{v_d=0}^{x_d} \frac{x_1 - v_1}{\alpha_1} f_{\alpha_1}(x_1 - v_1) \prod_{i=2}^{d} f_{\alpha_i}(x_i - v_i) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) \\
&= \frac{f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\alpha_1} \sum_{v_1=0}^{x_1} \cdots \sum_{v_d=0}^{x_d} (x_1 - v_1) \frac{\prod_{i=1}^{d} f_{\alpha_i}(x_i - v_i) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R})}{f_{\boldsymbol{\theta}}(\boldsymbol{x})} \\
&= \frac{f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\alpha_1} \sum_{v_1=0}^{x_1} \cdots \sum_{v_d=0}^{x_d} (x_1 - v_1) \frac{\Pr(Y_i = x_i - v_i, V_i = v_i,\ i = 1, \cdots d)}{\Pr(X_i = x_i,\ i = 1, \cdots d)} \\
&= \frac{f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\alpha_1} \mathbb{E}[Y_1 | X_i = x_i,\ \forall i = 1, \cdots, d],
\end{aligned} \tag{5.6}$$

assuming $\alpha_1 > 0$. Other recurrence relations can be derived similarly.

5.2.2. MOMENTS AND CORRELATIONS

Using the proposed model, the marginal means and variances are given by $\alpha_i + \beta_i$ for $i = 1, \cdots, d$. To analyze the correlation between X_i and X_j, the key is to compute the term $\mathbb{E}[V_i V_j]$. This can be formalized through the following theorem.

Theorem 3. *Assume that Model (5.2) holds. Then $\mathbb{E}[V_i V_j]$ is given by*

$$\sum_{l=0}^{+\infty} \sum_{k=0}^{+\infty} \Pr(Z_i > \Phi^{-1}(F_{\beta_i}(l)), Z_j > \Phi^{-1}(F_{\beta_j}(k))) = \sum_{l=0}^{+\infty} \sum_{k=0}^{+\infty} \Pr(V_i \geq l+1, V_j \geq k+1).$$

This theorem, combined with Equation (5.3), gives

$$\rho_{i,j}^{X} := \frac{\sum_{l=0}^{+\infty}\sum_{k=0}^{+\infty}\bar{\Phi}_{\rho_{i,j}}(\Phi^{-1}(F_{\beta_i}(l)),\Phi^{-1}(F_{\beta_j}(k)))-\beta_i\beta_j}{\sqrt{(\alpha_i+\beta_i)(\alpha_j+\beta_j)}}, \tag{5.7}$$

where $\bar{\Phi}_{\rho_{i,j}}(x,y) := \int_x^{+\infty}\int_y^{+\infty}(2\pi(1-\rho_{i,j}^2))^{-1}\exp(-(z_i^2-2\rho_{i,j}z_iz_j+z_j^2)/(2-2\rho_{i,j}^2))dz_idz_j$. This expression involves summing over an infinite number of terms. To make computations feasible, it is beneficial to truncate this summation while ensuring that the sum remains sufficiently accurate. In this chapter, we achieve this by using some upper bounds in the indices, denoted by L and K, typically chosen as large (e.g., 99%) quantiles of $\mathcal{P}(\beta_i)$ and $\mathcal{P}(\beta_j)$, respectively. Finally, the correlation coefficient $\rho_{i,j}^{X}$ is approximated by $\rho_{i,j}^{X,L,K}$, defined by

$$\rho_{i,j}^{X,L,K} := \frac{\sum_{l=0}^{L}\sum_{k=0}^{K}\bar{\Phi}_{\rho_{i,j}}(\Phi^{-1}(F_{\beta_i}(l)),\Phi^{-1}(F_{\beta_j}(k)))-\beta_i\beta_j}{\sqrt{(\alpha_i+\beta_i)(\alpha_j+\beta_j)}}.$$

Recall that $\rho_{i,j}$ denotes the correlation between two standard normal variables Z_i and Z_j. Some properties related to $\rho_{i,j}^{X}$ are summarized in the following propositions.

Proposition 3. *The function $\rho_{i,j}^{X} = \rho_{i,j}^{X}(\alpha_i,\alpha_j,\beta_i,\beta_j,\rho_{i,j})$ is continuous on the domain $(0,+\infty)^4\times(-1,1)$.*

Proposition 4. *Assume that α_i, α_j, β_i and β_j are all fixed.*

- *The function $\rho_{i,j}^{X}$ is nondecreasing in $\rho_{i,j}\in[-1,1]$.*
- *When $\rho_{i,j}\le(\ge)0$, $\mathrm{Corr}(V_i,V_j)\le(\ge)0$. In particular, $\mathrm{Corr}(V_i,V_j)$ is minimal, maximal and 0 when $\rho_{i,j}$ is -1, 1, and 0.*

These results illustrate how $\rho_{i,j}$ affects the correlation structure of Poisson variables. For parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, analogous results are given in the following proposition.

Proposition 5. *$\rho_{i,j}^{X}$ is a strictly decreasing (increasing) function in $\alpha_i\ge 0$ or $\alpha_j\ge 0$ if other parameters are fixed and $\rho_{i,j}>0$ ($\rho_{i,j}<0$). In addition, $\mathbb{E}[V_iV_j]$ is a strictly increasing function for $\beta_i>0$ or $\beta_j>0$ if other parameters are fixed.*

The proposition illustrates how $\boldsymbol{\alpha}$ captures the independent component, while $\boldsymbol{\beta}$ indicates the significance of V_i. For example, setting $\boldsymbol{\alpha}=\mathbf{0}$ and all $\rho_{i,j}=1$ achieves the maximum correlation. There is numerical evidence suggesting that $\rho_{i,j}^{X}$ is a non-decreasing (non-increasing) function of β_i or β_j when the other parameters are fixed and $\rho_{i,j}>0$ ($\rho_{i,j}<0$). Although theoretical proof remains elusive, this behaviour appears intuitive, as increasing the dependent component will enhance the correlation.

5.3. PROPERTIES AND COMPARISONS

In this section, we present a comprehensive comparative analysis of our proposed model and existing approaches, emphasizing the key properties that distinguish our framework from conventional methods.

A key property of our proposed model is its ability to achieve the full range of pairwise correlations $\rho^X_{i,j}$, given the marginal means λ_i and λ_j. By setting $\alpha_i = \alpha_j = 0$, $\beta_i = \lambda_i$, and $\beta_j = \lambda_j$, our model can attain any feasible $\rho^X_{i,j}$ by varying $\rho_{i,j}$ over the interval $[-1, 1]$. This capability is guaranteed by Propositions 3 and 4, which show that $\rho^X_{i,j}$ is a continuous and non-decreasing function of $\rho_{i,j}$ on $[-1, 1]$, ensuring that a solution exists for any given pairwise correlation. On the other hand, TREx and the Gaussian copula approach can also accommodate the full range of $\rho^X_{i,j}$ by adjusting parameters γ and $\rho_{i,j}$, respectively, while the TR model is limited and cannot capture the complete range, particularly the negative correlations.

Secondly, our proposed model can capture a wider range of dependence structures than TREx, particularly in high-dimensional settings. Since our approach generalizes TREx, it is capable of representing all dependence structures achievable by TREx while overcoming some of its limitations. To illustrate this, consider the case with $d = 3$. We define

$$\bar{\rho}^X_{i,j} := \mathrm{Corr}\big(F^{-1}_{\lambda_i}(U), F^{-1}_{\lambda_j}(U)\big) \quad \text{and} \quad \underline{\rho}^X_{i,j} := \mathrm{Corr}\big(F^{-1}_{\lambda_i}(U), F^{-1}_{\lambda_j}(1-U)\big),$$

which represent the maximum and minimum attainable correlations between X_i and X_j, respectively. For example, our model can yield correlation vectors such as:

- $[\rho^X_{1,2}, \rho^X_{1,3}, \rho^X_{2,3}] = [0, 0, \bar{\rho}^X_{2,3}]$, $[0, \bar{\rho}^X_{1,3}, 0]$, or $[\bar{\rho}^X_{1,2}, 0, 0]$;
- $[\rho^X_{1,2}, \rho^X_{1,3}, \rho^X_{2,3}] = [0, 0, \underline{\rho}^X_{2,3}]$, $[0, \underline{\rho}^X_{1,3}, 0]$, or $[\underline{\rho}^X_{1,2}, 0, 0]$.

To see why TREx fails in such cases, we present the following proposition:

Proposition 6. *Let U be uniformly distributed on $[0, 1]$, and let $X_i \sim \mathcal{P}(\lambda_i)$ and $X_j \sim \mathcal{P}(\lambda_j)$ for $i, j = 1, \ldots, d$, $i \neq j$. Then,*

$$\mathrm{Corr}(X_i, X_j) = \mathrm{Corr}\Big(F^{-1}_{\lambda_i}(U), F^{-1}_{\lambda_j}(U)\Big)$$

if and only if (X_i, X_j) has the same distribution as $\Big(F^{-1}_{\lambda_i}(U), F^{-1}_{\lambda_j}(U)\Big)$, i.e., the maximum possible correlation between X_i and X_j is attained. A similar result holds for the minimal attainable correlation.

According to Proposition 6, if $\rho^X_{1,3}$ is maximal, then the parameter γ in the TREx model must be set to 1, forcing $X_1 = F^{-1}_{\lambda_1}(U)$ and

$X_3 = F^{-1}_{\lambda_3}(U)$. Since all pairwise correlations are governed by γ in (5.1), this constraint forces the remaining correlations to be either maximal or minimal, limiting the flexibility of TREx. In contrast, our model, with its additional parameters, can readily accommodate a broader spectrum of dependence structures. For instance, by setting all α's to 0 and letting $\beta_i = \lambda_i$ for $i = 1, 2, 3$, if we specify $[\rho_{1,2}, \rho_{1,3}, \rho_{2,3}] = [0, 1, 0]$ in our model, the resulting correlations $[\rho^X_{1,2}, \rho^X_{1,3}, \rho^X_{2,3}]$ become $[0, \bar{\rho}^X_{1,3}, 0]$. This example clearly demonstrates the superior flexibility of our approach in capturing a full range of dependence structures compared to TREx.

Notably, while the Gaussian copula shares some fundamental properties with our model, its one-to-one mapping between $\rho_{i,j}$ and $\rho^X_{i,j}$ yields a unique joint distribution for any given set of marginals and dependence structure. In contrast, our model can generate infinitely many joint Poisson distributions for fixed marginals and a prescribed correlation matrix. To illustrate this novelty, we present the following proposition.

5

Proposition 7. *For $d = 3$, there exist parameters $\lambda_1, \lambda_2, \lambda_3 \in (0, +\infty)$ and target correlations $\rho^X_{1,2}, \rho^X_{1,3}, \rho^X_{2,3} \in (0, 1)$ such that there are infinitely many joint Poisson distributions with marginals $\mathcal{P}(\lambda_1)$, $\mathcal{P}(\lambda_2)$, $\mathcal{P}(\lambda_3)$ and correlation matrix*

$$\begin{pmatrix} 1 & \rho^X_{1,2} & \rho^X_{1,3} \\ \rho^X_{1,2} & 1 & \rho^X_{2,3} \\ \rho^X_{1,3} & \rho^X_{2,3} & 1 \end{pmatrix}.$$

This proposition demonstrates that, even when the marginal distributions and the correlation matrix are fixed, the joint Poisson distribution is not unique. In our model, given the marginal means λ_i and λ_j and the target correlations $\rho^X_{i,j}$ for $i, j = 1, \ldots, d$, we determine parameters α_i, α_j, β_i, and β_j such that $\alpha_i + \beta_i = \lambda_i$ and $\alpha_j + \beta_j = \lambda_j$. Propositions 3 and 4 ensure that the corresponding correlations $\rho_{i,j}$ are uniquely determined by these parameters. Thus, by varying the values of α_i or β_i while keeping their sum fixed, our model can generate infinitely many joint Poisson distributions with identical marginals and dependence structures. This capability enables our approach to capture a wider range of joint distributions than the Gaussian copula.

To further illustrate this property, consider a two-dimensional case with correlated Poisson variables having marginal means $\lambda_1 = 2$ and $\lambda_2 = 1$. By varying the parameter vector $\boldsymbol{\theta}$ within our model, we generate a diverse set of Poisson distributions. We then fit a Gaussian copula to match the preset marginals and correlations, and compare the resulting PMFs. As shown in Figure 5.1, although the joint distribution defined by the Gaussian copula may appear similar to that produced by our approach, the two distributions are distinct. This example highlights that our model can effectively capture joint distributions that the Gaussian copula is unable to describe.

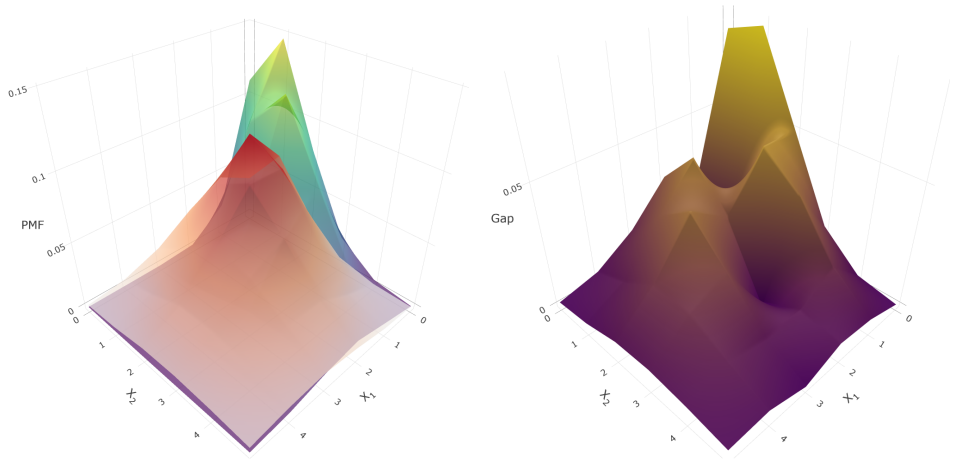


Figure 5.1: Comparison study between our model and the Gaussian copula. Left: the PMF based on our model (cool colors) with $\boldsymbol{\alpha} = [0.42, 0.41]^\top$, $\boldsymbol{\beta} = [1.58, 0.59]^\top$ and $\rho_{1,2} = 0.99$, alongside the PMF defined by Gaussian copula (warm colours) with $\lambda_1 = 2$ and $\lambda_2 = 1$; Right: the absolute gap between such two PMFs.

In conclusion, our proposed model offers significant improvements over existing methods across several key dimensions. It is capable of attaining the full range of pairwise correlations, including negative values—a feature that traditional models such as TR cannot accommodate. Moreover, the model demonstrates superior flexibility in high-dimensional settings, effectively capturing diverse dependence structures that may be missed by TREx. Unlike conventional approaches, our model can generate infinitely many multivariate Poisson distributions for a given set of marginals and dependence structure, thus overcoming the limitations of both TR-based methods and the Gaussian copula. Collectively, these features underscore the model's versatility and broad applicability in modelling complex multivariate Poisson distributions.

5.4. ESTIMATION

Let $\mathbf{x} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^\top$ represent n random samples, each of dimension d, where $\boldsymbol{x}_i = [x_{i1}, \cdots, x_{id}]^\top$ for $i = 1, \cdots, n$. The goal of this section is to estimate the model parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho})$ within the parameter set $\Theta := (0, +\infty)^{2d} \times \mathcal{C}$, where $\mathcal{C}$ represents the set of distinct elements in d-dimensional correlation matrices, i.e., square matrices of size d which are positive semidefinite with ones along the diagonal. Note that

$\mathbf{R} = [\rho_{j,k}]_{j,k=1,\cdots d}$ with $\rho_{j,k} = \rho_{k,j}$ and $\rho_{j,j} = 1$. Since this chapter focuses on multivariate Poisson variables, we assume both $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ are nonzero. In this section, various estimation methods that provide consistent estimators for $\boldsymbol{\theta}$ are explored, including ML estimation and inference function for margins. A tailored EM algorithm is further proposed to ease the computational difficulties.

5.4.1. MAXIMUM LIKELIHOOD ESTIMATION

The likelihood function is $L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(x_{i1}, \cdots, x_{id})$, and the corresponding log-likelihood is, according to (5.5),

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{x}) &= \sum_{i=1}^{n} \ln f_{\boldsymbol{\theta}}(x_{i1}, \cdots, x_{id}) \\ &= \sum_{i=1}^{n} \ln \left\{ \sum_{v_1=0}^{x_{i1}} \cdots \sum_{v_d=0}^{x_{id}} \prod_{j=1}^{d} f_{\alpha_j}(x_{i,j} - v_j) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) \right\}. \end{aligned} \tag{5.8}$$

The ML estimate of $\boldsymbol{\theta}$ can be derived by setting the derivative of (5.8), which is known as the score function, to 0. Since the PMF based on our model can be viewed as a weighted sum of the probability density function (PDF) of the underlying multivariate normal distribution, it is easy to derive all derivatives. Beginning with $\boldsymbol{\alpha}$, the derivative of $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ with respect to α_j, $j = 1, \cdots, d$, is given by

$$\frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \alpha_j} = \sum_{v_1=0}^{x_{i1}} \cdots \sum_{v_d=0}^{x_{id}} \frac{\partial f_{\alpha_j}(x_{i,j} - v_j)}{\partial \alpha_j} \prod_{k=1, k\neq j}^{d} f_{\alpha_k}(x_{i,k} - v_k) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}).$$

Since

$$\frac{\partial f_{\alpha_j}(x_{i,j} - v_j)}{\partial \alpha_j} = \frac{\partial}{\partial \alpha_j} \frac{\alpha_j^{x_{i,j} - v_j} e^{-\alpha_j}}{[x_{i,j} - v_j]!} = f_{\alpha_j}(x_{i,j} - v_j - 1) - f_{\alpha_j}(x_{i,j} - v_j),$$

one can proceed to differentiate $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ by

$$\begin{aligned} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \alpha_j} &= \sum_{v_1=0}^{x_{i1}} \cdots \sum_{v_d=0}^{x_{id}} f_{\alpha_j}(x_{i,j} - v_j - 1) \prod_{k=1, k\neq j}^{d} f_{\alpha_k}(x_{i,k} - v_k) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) \\ &\quad - \sum_{v_1=0}^{x_{i1}} \cdots \sum_{v_d=0}^{x_{id}} f_{\alpha_j}(x_{i,j} - v_j) \prod_{k=1, k\neq j}^{d} f_{\alpha_k}(x_{i,k} - v_k) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) \\ &= f_{\boldsymbol{\theta}}(x_{i1}, \cdots, x_{i,j} - 1, \cdots, x_{id}) - f_{\boldsymbol{\theta}}(x_{i1}, \cdots, x_{id}). \end{aligned}$$

The score functions based on (5.6) are

$$\begin{aligned}\frac{\partial l(\boldsymbol{\theta};\mathbf{x})}{\partial \alpha_j} &= \sum_{i=1}^{n} \frac{1}{f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \alpha_j} \\ &= \sum_{i=1}^{n} \frac{f_{\boldsymbol{\theta}}(x_{i1}, \cdots, x_{i,j}-1, \cdots, x_{id}) - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \\ &= \sum_{i=1}^{n} \frac{\mathbb{E}[Y_j | x_{i,j},\ j = 1, \cdots, d]}{\alpha_j} - n.\end{aligned}$$

In addition, the derivative with respect to β_j is

$$\frac{\partial l(\boldsymbol{\theta};\mathbf{x})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \sum_{v_1=0}^{x_{i1}} \cdots \sum_{v_d=0}^{x_{id}} \prod_{j=1}^{d} f_{\alpha_j}(x_{i,j} - v_j) \frac{\partial f_{\boldsymbol{\beta}}(\boldsymbol{v};\mathbf{R})}{\partial \beta_j},$$

where

$$\frac{\partial f_{\boldsymbol{\beta}}(\boldsymbol{v};\mathbf{R})}{\partial \beta_j} = \frac{\partial}{\beta_j} \int_{\Phi^{-1}(F_{\beta_1}(v_1-1))}^{\Phi^{-1}(F_{\beta_1}(v_1))} \cdots \int_{\Phi^{-1}(F_{\beta_d}(v_d-1))}^{\Phi^{-1}(F_{\beta_d}(v_d))} \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{R}|^{\frac{1}{2}}} \exp(-\frac{\boldsymbol{z}^\top \mathbf{R}^{-1} \boldsymbol{z}}{2}) d\boldsymbol{z}.$$

In principle, such derivatives can be derived explicitly, using the derivative of an integral with two variable limits and the derivative of inverse functions. In addition, the derivative with respect to $\mathbf{R}$, which is invertible, is also a multiple integration of $\phi_{\mathbf{R}}$, which is the joint PDF of $\boldsymbol{Z}$. According to the matrix differential methods [160], we have

$$\frac{\partial \phi_{\mathbf{R}}}{\partial \mathbf{R}} = \frac{\partial \phi_{\mathbf{R}}}{\partial \ln \phi_{\mathbf{R}}} \frac{\partial \ln \phi_{\mathbf{R}}}{\partial \mathbf{R}} = \phi_{\mathbf{R}} \frac{\partial \ln \phi_{\mathbf{R}}}{\partial \mathbf{R}} = \phi_{\mathbf{R}} \cdot -\frac{1}{2} \mathbf{R}^{-1} (\mathbf{R} - \boldsymbol{z}\boldsymbol{z}^\top) \mathbf{R}^{-1},$$

where $\boldsymbol{z}$ is a column vector containing $z_1, \cdots, z_d$. Via the Leibniz integral rule, it follows that

$$\begin{aligned}\frac{\partial l(\boldsymbol{\theta};\mathbf{x})}{\partial \mathbf{R}} &= \sum_{i=1}^{n} \frac{1}{f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \sum_{v_1=0}^{x_{i1}} \cdots \sum_{v_d=0}^{x_{id}} \prod_{j=1}^{d} f_{\alpha_j}(x_{i,j} - v_j) \frac{\partial f_{\boldsymbol{\beta}}(\boldsymbol{v};\mathbf{R})}{\partial \mathbf{R}} \\ &= \sum_{i=1}^{n} \frac{1}{f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \sum_{v_1=0}^{x_{i1}} \cdots \sum_{v_d=0}^{x_{id}} \prod_{j=1}^{d} f_{\alpha_j}(x_{i,j} - v_j) \\ &\quad \int_{\Phi^{-1}(F_{\beta_1}(v_1-1))}^{\Phi^{-1}(F_{\beta_1}(v_1))} \cdots \int_{\Phi^{-1}(F_{\beta_1}(v_d-1))}^{\Phi^{-1}(F_{\beta_1}(v_d))} \phi_{\mathbf{R}} \cdot -\frac{1}{2} \mathbf{R}^{-1} (\mathbf{R} - \boldsymbol{z}\boldsymbol{z}^\top) \mathbf{R}^{-1} d\boldsymbol{z},\end{aligned}$$

which is a function of $\boldsymbol{\rho}$. As demonstrated, the computation of all derivatives is feasible but requires multidimensional integration. This can

be accomplished using numerical integration methods, such as Monte Carlo integration [161].

Let $[\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}}]^\top$ represent the ML estimators. Under standard regularity conditions, these estimators are consistent and asymptotically normal, expressed as

$$\sqrt{n}([\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}}]^\top - [\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}]^\top) \rightsquigarrow N(0, \mathcal{I}^{-1}(\boldsymbol{\theta})), \quad n \to \infty,$$

where $\mathcal{I}(\boldsymbol{\theta})$ denotes the Fisher information matrix and "$\rightsquigarrow$" denotes convergence in distribution. The asymptotic variance can be estimated in practice using the bootstrap method. Despite this, such multi-parameter numerical optimization remains challenging due to the repeated need for optimization during the bootstrap process. To address this issue, the inference function for marginals (IFM) is introduced in the next subsection.

5

5.4.2. INFERENCE FUNCTION FOR MARGINS

IFM, as a two-stage estimation method, originates from the general framework in [162] and [163]. Some recent studies include [164] and [165]. The procedure first estimates the d marginal distributions, and in the second stage, it maximizes the likelihood with respect to the remaining parameters. More specifically, by rewriting the joint PMF in a slightly different manner, we have

$$\begin{aligned} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) &= \sum_{v_1=0}^{x_{i1}} \cdots \sum_{v_d=0}^{x_{id}} \prod_{j=1}^{d} f_{\alpha_j}(x_{i,j} - v_j) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) \\ &= \prod_{j=1}^{d} f_{\alpha_j+\beta_j}(x_{i,j}) \times \sum_{v_1=0}^{x_{i1}} \cdots \sum_{v_d=0}^{x_{id}} \prod_{j=1}^{d} \frac{f_{\alpha_j}(x_{i,j} - v_j) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R})}{f_{\alpha_j+\beta_j}(x_{i,j})} \\ &= \prod_{j=1}^{d} f_{\alpha_j+\beta_j}(x_{i,j}) \times \omega(\boldsymbol{x}_i; \boldsymbol{\theta}), \ i = 1, \cdots, n, \end{aligned}$$

where $\omega(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{v_1=0}^{x_{i1}} \cdots \sum_{v_d=0}^{x_{id}} \prod_{j=1}^{d} f_{\alpha_j}(x_{i,j} - v_j) f_{\boldsymbol{\beta}}(\boldsymbol{v}; \mathbf{R}) / f_{\alpha_j+\beta_j}(x_{i,j})$. Remark that the log-likelihood function can be rewritten as

$$l(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{d} \ln f_{\lambda_j}(x_{i,j}) + \sum_{i=1}^{n} \ln \omega(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{j=1}^{d} l_j(\lambda_j; \mathbf{x}) + l_{d+1}(\mathbf{x}; \boldsymbol{\theta}),$$

where $l_j(\lambda_j; \mathbf{x})$ denotes the marginal likelihood and $l_{d+1}(\mathbf{x}; \boldsymbol{\theta})$ encompasses the correlations. The IFM approach is outlined as follows.

(S1). Estimate $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_d)$ by maximizing marginal likelihood $l_j(\lambda_j; \mathbf{x})$ for $j = 1, \cdots, d$, resulting in the vector of sample means $\bar{\boldsymbol{x}} = (\bar{x}_1, \cdots, \bar{x}_d)$.

(S2). Estimate the remaining parameters by maximizing $l_\omega(\mathbf{x};\boldsymbol{\theta})$ with the constraint $\boldsymbol{\alpha}+\boldsymbol{\beta}=\bar{\boldsymbol{x}}$.

This procedure is easier to implement than the ML estimation due to the reduced complexity of the search space, which benefits from the incorporation of marginal means. Denote the IFM estimator as $[\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\rho}}]^\top$. According to Page 404-405 in [159], the limiting distribution is normal,

$$\sqrt{n}([\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\rho}}]^\top-[\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\rho}]^\top)\rightsquigarrow N(\mathbf{0},\mathcal{W}),\ n\to\infty, \tag{5.9}$$

with the asymptotic covariance matrix $\mathcal{W}$ given in Section 7.4.

5.4.3. EXPECTATION-MAXIMIZATION METHOD

Due to the large number of parameters, directly maximizing the full log-likelihood (see Section 5.4.1) or even the likelihood after marginal estimation (see Section 5.4.2) becomes impractical. Therefore, in this subsection, we employ the EM algorithm to obtain the estimators. The EM algorithm is widely used for estimation in the presence of missing data [e.g., 166–168]. In our framework, we treat the latent variables $V_j:=F_{\beta_j}^{-1}(\Phi(Z_j)), j=1,\cdots,d$, as unobserved components. The corresponding unobserved data can be represented as an $n\times d$ matrix $\mathbf{v}=[\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n]^\top$, where $\boldsymbol{v}_i=[v_{i,1},\cdots,v_{i,d}]^\top$. The complete dataset is $\{(x_{i,j},v_{i,j})\}_{i=1,\cdots,n,j=1,\cdots,d}$, of which only the $x_{i,j}$ are directly observed.

To obtain estimates via such data, two steps, including the E-step and the M-step, are employed iteratively. In the kth E-step, we compute the conditional expected value $Q(\boldsymbol{\theta};\check{\boldsymbol{\theta}}^{(k)})$ of the complete log-likelihood function $l(\boldsymbol{\theta};\mathbf{x},\mathbf{v})$ given the observed dataset $\mathbf{x}$ under the distribution corresponding to the current estimate $\check{\boldsymbol{\theta}}^{(k)}$ of the parameters. In the kth M-step, the algorithm updates the parameter estimates to maximize the expected complete log-likelihood function $Q(\boldsymbol{\theta};\check{\boldsymbol{\theta}}^{(k)})$ computed in the E-step over $\boldsymbol{\theta}$. This involves solving an optimization problem, often done efficiently using standard optimization techniques. In other words, starting from an initial value $\check{\boldsymbol{\theta}}^{(0)}$, the algorithm iterates $\check{\boldsymbol{\theta}}^{(k+1)}=\arg\max_{\boldsymbol{\theta}\in\Theta}Q(\boldsymbol{\theta};\check{\boldsymbol{\theta}}^{(k)})$.

Following the above general procedures, the complete density function is given by

$$\begin{aligned}
&f_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{v})\\
&=\prod_{i=1}^{d}f_{\alpha_i}(x_i-v_i)f_{\boldsymbol{\beta}}(\boldsymbol{v};\mathbf{R})\\
&=\prod_{i=1}^{d}f_{\alpha_i}(x_i-v_i)\int_{\Phi^{-1}(F_{\beta_1}(v_1-1))}^{\Phi^{-1}(F_{\beta_1}(v_1))}\cdots\int_{\Phi^{-1}(F_{\beta_d}(v_d-1))}^{\Phi^{-1}(F_{\beta_d}(v_d))}\frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{R}|^{\frac{1}{2}}}\exp(-\frac{\boldsymbol{z}\mathbf{R}^{-1}\boldsymbol{z}^\top}{2})d\boldsymbol{z},
\end{aligned}$$

and the conditional density of $\boldsymbol{v}$ given $\boldsymbol{x}$ is

$$\begin{aligned} f_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{x}) &= \Pr(V_1 = v_1, \cdots, V_d = v_d | X_1 = x_1, \cdots, X_d = x_d) \\ &= \frac{\Pr(V_1 = v_1, \cdots, V_d = v_d, X_1 = x_1, \cdots, X_d = x_d)}{\Pr(X_1 = x_1, \cdots, X_d = x_d)} \\ &= \frac{\Pr(Y_1 = x_1 - v_1, \cdots, Y_d = x_d - v_d, V_1 = v_1, \cdots, V_d = v_d)}{\Pr(X_1 = x_1, \cdots, X_d = x_d)} \\ &= \frac{\prod_{j=1}^{d} \Pr(Y_j = x_j - v_j) \Pr(V_1 = v_1, \cdots, V_d = v_d)}{\Pr(X_1 = x_1, \cdots, X_d = x_d)}. \end{aligned}$$

This conditional density can be obtained by calculating the numerator and rescaling it by the corresponding sum. Since the samples are i.i.d., the complete log-likelihood is given by

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{x}, \mathbf{v}) &= \sum_{i=1}^{n} \ln f_{\boldsymbol{\beta}}(\boldsymbol{v}_i; \mathbf{R}) + \sum_{i=1}^{n} \sum_{j=1}^{d} \ln f_{\alpha_j}(x_{i,j} - v_{i,j}) \\ &= \sum_{i=1}^{n} \ln f_{\boldsymbol{\beta}}(\boldsymbol{v}_i; \mathbf{R}) + \sum_{i=1}^{n} \sum_{j=1}^{d} [-\alpha_j + (x_{i,j} - v_{i,j}) \ln \alpha_j - \ln((x_{i,j} - v_{i,j})!)]. \end{aligned}$$

Thus, the function Q can be computed by

$$\begin{aligned} Q(\boldsymbol{\theta}; \check{\boldsymbol{\theta}}^{(k)}) &= \mathbb{E}_{f(\boldsymbol{v}|\boldsymbol{x})}\left[l(\boldsymbol{\theta}; \mathbf{x}, \mathbf{v}) | \check{\boldsymbol{\theta}}^{(k)} \right] \\ &= \sum_{i=1}^{n} \sum_{v_{i1}}^{x_{i1}} \cdots \sum_{v_{id}}^{x_{id}} \ln f_{\boldsymbol{\beta}}(\boldsymbol{v}_i; \mathbf{R}) f(\boldsymbol{v}_i | \mathbf{x}; \check{\boldsymbol{\theta}}^{(k)}) \\ &+ \sum_{i=1}^{n} \sum_{j=1}^{d} \left\{ -\alpha_j + (x_{i,j} - \mathbb{E}[v_{i,j} | \mathbf{x}; \check{\boldsymbol{\theta}}^{(k)}]) \ln \alpha_j - \mathbb{E}[\ln((x_{i,j} - v_{i,j}))! | \mathbf{x}; \check{\boldsymbol{\theta}}^{(k)}] \right\}. \end{aligned}$$

In order to compute $\check{\boldsymbol{\theta}}^{(k+1)}$, which is the maximizer of $Q(\cdot; \check{\boldsymbol{\theta}}^{(k)})$, we start by maximizing in $\boldsymbol{\alpha}$. This computation is straightforward and yields $\check{\boldsymbol{\alpha}}^{(k+1)} = \bar{\boldsymbol{x}} - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\boldsymbol{v}_i | \mathbf{x}; \check{\boldsymbol{\theta}}^{(k)}]$, where $\bar{\boldsymbol{x}}$ is the vector of coordinate-wise averages of $X_{i,j}$ and $(\check{\boldsymbol{\beta}}^{(k+1)}, \check{\mathbf{R}}^{(k+1)})$ is the maximizer of

$$\begin{aligned} \check{Q}(\boldsymbol{\beta}, \mathbf{R}; \check{\boldsymbol{\theta}}^{(k)}) &:= \max_{\boldsymbol{\alpha} \in (0, +\infty)^d} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{R}; \check{\boldsymbol{\theta}}^{(k)}) \\ &= \sum_{i=1}^{n} \sum_{v_{i1}}^{x_{i1}} \cdots \sum_{v_{id}}^{x_{id}} \ln f_{\boldsymbol{\beta}}(\boldsymbol{v}_i; \mathbf{R}) f(\boldsymbol{v}_i | \mathbf{x}; \check{\boldsymbol{\theta}}^{(k)}) + \sum_{i=1}^{n} \sum_{j=1}^{d} \left\{ -\left(\bar{x}_j - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[v_{i,j} | \mathbf{x}; \check{\boldsymbol{\theta}}^{(k)}] \right) \right. \\ &\left. + (x_{i,j} - \mathbb{E}[v_{i,j} | \mathbf{x}; \check{\boldsymbol{\theta}}^{(k)}]) \ln \left(\bar{x}_j - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[v_{i,j} | \mathbf{x}; \check{\boldsymbol{\theta}}^{(k)}] \right) - \mathbb{E}[\ln((x_{i,j} - v_{i,j}))! | \mathbf{x}; \check{\boldsymbol{\theta}}^{(k)}] \right\}. \end{aligned}$$

Note that the expectations involved in the previous equations can be computed easily as

$$\begin{aligned}\mathbb{E}[v_{i,j}|\mathbf{x};\check{\boldsymbol{\theta}}^{(k)}] &= \sum v_{i,j}\Pr(V_{i,j}=v_{i,j}|\mathbf{x};\check{\boldsymbol{\theta}}^{(k)}) \\ &= \sum v_{i,j} \sum_{v_{i,i'},i'\neq j} \Pr(V_{i,1}=v_{i,1},\cdots,V_{i,d}=v_{i,d}|\mathbf{x};\check{\boldsymbol{\theta}}^{(k)}) \\ &= \sum v_{i,j} \sum_{v_{i,i'},i'\neq j} f(v_{i1},\cdots,v_{id}|\mathbf{x};\check{\boldsymbol{\theta}}^{(k)}).\end{aligned}$$

In this chapter, we apply derivative-free optimization methods for the maximization of $\check{Q}(\cdot;\check{\boldsymbol{\theta}}^{(k)})$, i.e., for the computation of $(\check{\boldsymbol{\beta}}^{(k+1)},\check{\mathbf{R}}^{(k+1)})$. The EM algorithm can then be iteratively executed until convergence is achieved based on a specified criterion, such as when the L_1 distance between the parameter estimates in two successive iterations falls below a predefined threshold. The following result shows the convergence of the EM algorithm to stationary points of the log-likelihood function $l(\boldsymbol{\theta};\mathbf{x})$. The proof can be found in Section 7.4.

Proposition 8. *All limit points of the sequence* $\{\check{\boldsymbol{\theta}}^{(k)}\}$ *are stationary points of the log-likelihood function* $l(\boldsymbol{\theta};\mathbf{x})$ *defined in* (5.8).

5.5. NUMERICAL EXPERIMENTS

In this section, we present a comprehensive series of numerical experiments, all of which are based on artificial data sets. These experiments have been carefully designed to illustrate the novelty of our proposed model. We incorporate several well-established methods into our comparative studies, including TR, TREx, Gaussian copula (GC), and a t-copula with four degrees of freedom (TC-4). To ensure a thorough evaluation, we set different marginal means and correlation matrices with diverse dimensions. All of the implementations and analyses are carried out within the `R` environment [77]. A few specific R packages have been utilized in our study: `Rsolnp` [130] for the optimization, and `mvtnorm` [169] for the computations involving multivariate normal and t distributions. The performances of various methods are assessed and compared using Kullback–Leibler (KL) divergence between estimated PMF $\hat{f}$ and true f. Specifically, $D(\hat{f}||f)=\sum_{\boldsymbol{x}\in\mathbb{N}^2}\hat{f}(\boldsymbol{x})\ln\hat{f}(\boldsymbol{x})/f(\boldsymbol{x})$. Take two dependent Poisson variables as an example, let testing points be a Cartesian product of two integer sets $\{0,\cdots,M\}$ and $\{0,\cdots,N\}$,

$$\mathcal{X}_t=\begin{bmatrix}0 & 0 & \cdots & 0 & 1 & 1 & 1 & \cdots & M\\ 0 & 1 & \cdots & N & 0 & 1 & 2 & \cdots & N\end{bmatrix}^{\top},$$

where M and N are 99% percentiles of two marginal Poisson distributions during the simulation, respectively. Then, the KL divergence can be approximated as $D(\hat{f}||f)=\sum_{\boldsymbol{x}\in\mathcal{X}_t}\hat{f}(\boldsymbol{x})\ln\hat{f}(\boldsymbol{x})/f(\boldsymbol{x})$.

5.5.1. POSITIVELY-CORRELATED BIVARIATE POISSON RANDOM VARIABLES

We begin by considering two positively correlated Poisson random variables, X_1 and X_2, to demonstrate our model's performance in settings where the data are well-suited to all methods. To further showcase the model's flexibility in capturing various distributions, we generate additional datasets using a range of copulas, including Clayton, Frank, Gumbel, Ali-Mikhail-Haq, and Joe, via the `copula` package. For each copula-based data generation mechanism, we define four scenarios by selecting four distinct sets of parameter values (detailed in Section 7.4), each representing a different data type. In each scenario, we conduct 100 independent simulation replications, with each replication generating 100 two-dimensional Poisson observations. We then compute the average KL divergence to compare performance. In every replication, the models TR, TREx, GC, TC-4, and our proposed method are applied to capture the joint distribution and dependence structure of the data.

All results are provided in Section 7.4. Figure 5.2 visually summarizes the average KL divergences computed from data generated by the TR, TREx, GC, and TC-4 models under four different parameter scenarios. Under conditions of positive correlation, all methods perform well in capturing the full range of $\rho^X_{1,2}$, except for TR. TR has the simplest form, and its parameters are easy to estimate. However, it struggles to handle certain types of positive correlations, while the other methods demonstrate better performance in modelling all positive correlation structures. In addition, each existing method produces distinct PMFs, meaning that none of these approaches works universally well except when the data are generated from their respective models. In contrast, our proposed model consistently ranks second in performance across all scenarios. Although our model extends the TR, TREx, and GC frameworks, it does not simply replicate their results; the additional parameters can sometimes yield higher likelihood values at the cost of divergent parameter estimates. Nonetheless, the KL divergences of our model remain comparable to those of the original models.

Table 5.1 provides detailed results via 100 replications for a specific scenario across various data sources. While all methods demonstrate reasonable performance, our proposed model consistently achieves a low KL divergence, comparable to the true model and superior to other approaches. This highlights our model's enhanced flexibility in modelling positively correlated Poisson variables, particularly when the true PMF is too complex for the TR, TREx, GC, and TC-4 models. By incorporating additional parameters, our model exhibits an improved ability to accommodate a wider variety of distributions.

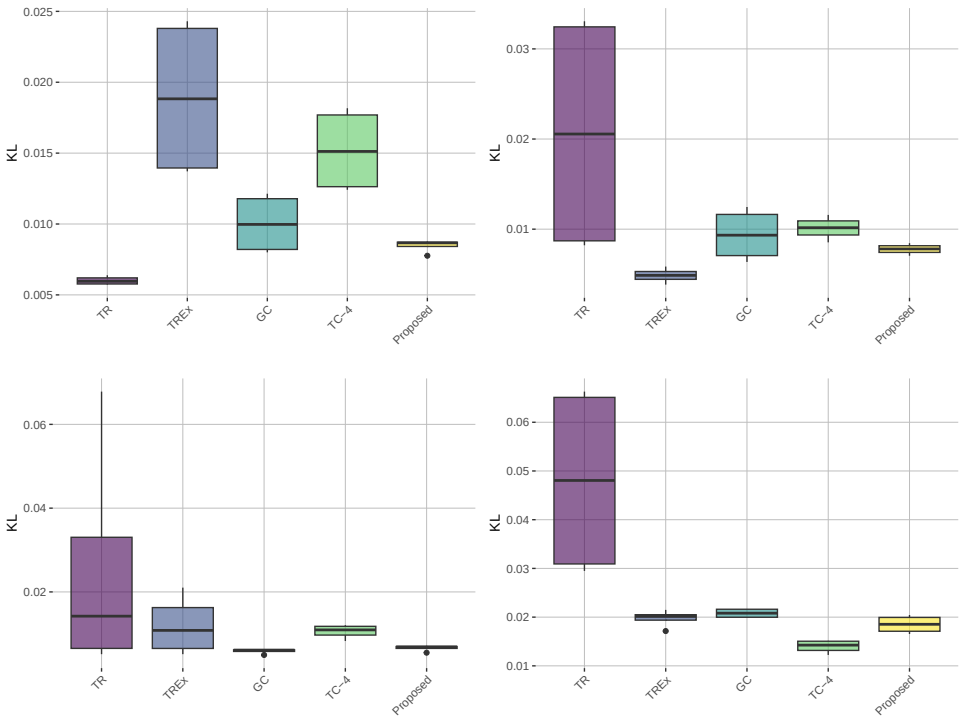


Figure 5.2: Box plots of average KL divergence across different parameter values for five methods. Four groups of two-dimensional positively correlated Poisson data are generated using TR (top-left), TREx (top-right), GC (bottom-left), and TC-4 (bottom-right), each with distinct parameter sets.

Table 5.1: Average KL distances for five methods. Two-dimensional positively correlated Poisson data are generated from nine generation mechanisms.

Data Source	Marginals	Parameter	TR	TREx	GC	TC-4	Proposed
TR	$[2, 1]$	$\beta = 0.9$	0.006	0.024	0.012	0.018	0.009
TREx	$[2, 1]$	$\gamma = 0.8$	0.032	0.005	0.011	0.011	0.008
GC	$[2, 1]$	$\rho = 0.9$	0.068	0.021	0.006	0.012	0.007
TC-4	$[2, 1]$	$\rho = 0.9$	0.065	0.020	0.020	0.013	0.017
Clayton	$[1, 2]$	$\theta = 4.5$	0.064	0.061	0.039	0.045	0.039
Frank	$[1, 2]$	$\theta = 9$	0.041	0.021	0.016	0.020	0.017
Gumbel	$[1, 2]$	$\theta = 5$	0.191	0.021	0.011	0.009	0.012
Amh	$[1, 2]$	$\theta = 0.91$	0.012	0.015	0.011	0.014	0.009
Joe	$[1, 2]$	$\theta = 5$	0.146	0.034	0.036	0.028	0.030

5.5.2. TRIVARIATE POISSON RANDOM VARIABLES

In this subsection, we investigate more complex three-dimensional Poisson variables, X_1, X_2, and X_3, with marginal means fixed at 1, 2, and 3, respectively. For data sets with all positive correlations, we generate samples using the TR, Clayton, Frank, Gumbel, and Joe copulas, whereas those with all negative correlations are produced using the Gaussian copula. In cases where only one pairwise correlation is negative (e.g., $\rho^X_{1,2} < 0$), we employ the t copula with 4 degrees of freedom. Additionally, we use TREx with $1 - U$ to construct joint distributions featuring two negative correlations with only $\rho^X_{1,2} > 0$. Various scenarios (as presented in Section 7.4) are considered by configuring the correlation matrix $\mathbf{R}$ differently. Note that the Ali–Mikhail–Haq copula is bivariate and is therefore not used in this subsection. For each scenario, 100 samples are generated, and each simulation is repeated 100 times. Average KL divergence is used again to evaluate the performance.

As shown in Figure 5.3, when all correlations are positive, all methods provide accurate estimates of the joint distribution, with TR performing best since the data are generated from it. However, in the presence of negative correlations, the TR model fails to capture the joint distribution adequately, leading to higher KL divergence values. Although TREx performs well when two correlations are negative, it struggles when only one or all three correlations are negative. This limitation arises because TREx relies solely on U and $1-U$ to model negative dependence: including one or two $1-U$ terms in (5.1) induces two negative correlations,

while the absence or inclusion of three $1-U$ terms results in only positive correlations. Consequently, TREx is unable to flexibly accommodate mixed dependence structures. In contrast, the GC, TC-4, and our proposed model effectively capture a variety of dependence patterns. Notably, our model demonstrates robust performance across different types of PMFs, consistently achieving low KL divergences and handling a broader range of dependence structures.

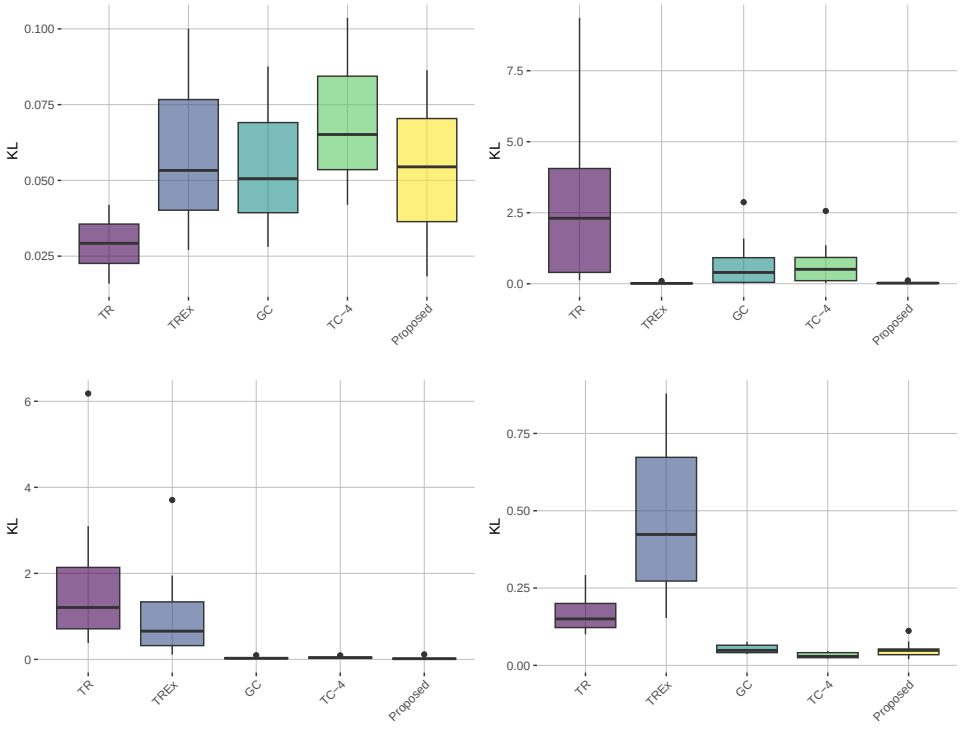


Figure 5.3: Box plots of average KL divergence across different scenarios for five methods. Four different correlation structures are considered, including all positive generated by TR (top-left), one positive $\rho^X_{1,2} > 0$ by TREx (top-right), all negative by GC (bottom-left), and one negative $\rho^X_{1,2} < 0$ by TC-4 (bottom-right).

Table 5.2 summarizes the estimation results for data generated by various methods; additional details are provided in Section 7.4. Overall, both the GC model and our proposed approach yield comparable KL divergence values, and our method generally performs better across a wider range of cases. Notably, when Poisson variables are generated

using the TC-4 mechanism, our method estimates the PMF more accurately than GC-even though our proposed model is based on a Gaussian copula. This finding suggests that the inclusion of the parameter vector $\boldsymbol{\alpha}$ enhances the model's ability to capture diverse distributions, particularly those with heavy-tailed PMFs.

Table 5.2: Average KL divergence for five methods. Three-dimensional correlated Poisson data are generated from eight generation mechanisms.

Data source	Parameter	TR	TREx	GC	TC-4	Proposed
TR	$\beta = 0.4$	0.016	0.027	0.028	0.042	0.018
TREx	$\theta = 0.3$	0.391	0.022	0.056	0.121	0.034
GC	$\boldsymbol{\rho} = [-0.2, -0.3, -0.4]$	0.381	0.112	0.014	0.027	0.016
TC-4	$\boldsymbol{\rho} = [0.3, 0.6, -0.5]$	0.214	0.700	0.065	0.041	0.049
Clayton	$\theta = 0.25$	0.067	0.016	0.016	0.029	0.015
Frank	$\theta = 0.5$	0.101	0.011	0.014	0.032	0.009
Gumbel	$\theta = 1.2$	0.054	0.012	0.021	0.036	0.016
Joe	$\theta = 1.2$	0.084	0.012	0.021	0.032	0.009

5.5.3. BIVARIATE POISSON RANDOM VARIABLES GENERATED VIA THE PROPOSED MODEL

In this subsection, we illustrate the performance of our proposed EM estimation procedure using bivariate Poisson random variables generated by our model. For this experiment, we fix the marginal means at 1 and set the sample size to $n = 100$. The data generation process proceeds as follows. First, we generate $n = 100$ two-dimensional samples from a bivariate normal distribution with mean vector $[0, 0]^\top$, unit variances, and a preset correlation coefficient $\rho_{1,2}$. These samples form an $n \times 2$ matrix $\mathbf{z} = [z_{i,j}]$, where $i = 1, \ldots, n$ and $j = 1, 2$. Second, we generate another $n \times 2$ matrix $\mathbf{y} = [y_{i,j}]$ consisting of independent Poisson random variables with predetermined marginal expectations $\alpha_1 > 0$ and $\alpha_2 > 0$. Finally, we construct the two-dimensional dependent Poisson variables via $x_{i,j} := y_{i,j} + F_{\beta_j}^{-1}\big(\Phi(z_{i,j})\big)$, $i = 1, \ldots, n$, $j = 1, 2$, where $\beta_j > 0$ is predetermined and F_λ denotes the CDF of a Poisson random variable with parameter λ. To generate a diverse set of data, we vary the parameters $\rho_{1,2}$, α_j, and β_j. In particular, to illustrate positive dependence, we let $\rho_{1,2}$ vary over $\{0.1, 0.2, \ldots, 0.9\}$, set α_j to values in $\{0.15, 0.3, \ldots, 0.9\}$,

and define $\beta_j = 1 - \alpha_j$. This configuration yields a total of 324 distinct scenarios. The entire simulation procedure is carried out 100 times.

Table 5.3 summarizes the biases and the mean squared errors (MSEs) of estimates computed in 100 replications for different combinations of αs and $\rho_{1,2}$. Specifically, $[\alpha_1, \alpha_2] = [0.15, 0.15]$, $[0.30, 0.15]$, $[0.45, 0.15]$, with $\beta_j = 1 - \alpha_j$ for $j = 1, 2$ and $\rho \in \{0.8, 0.9\}$. Additional results are provided in Section 7.4. Our analysis indicates that the proposed EM algorithm yields accurate estimates, particularly when the values of β and $\rho_{1,2}$ are larger. Moreover, Figure 5.4 shows the KL divergences for different values of $\rho_{1,2}$, demonstrating that the differences between the true PMF and the estimated PMF remain relatively low. These findings further confirm the effectiveness of the proposed EM approach.

Table 5.3: Average biases and MSEs (in parentheses) of estimates computed from 100 replications. S1-S2: $\alpha_1 = 0.15$, $\alpha_2 = 0.15$, $\rho \in \{0.8, 0.9\}$; S3-S4: $\alpha_1 = 0.30$, $\alpha_2 = 0.15$, $\rho \in \{0.8, 0.9\}$; S5-S6: $\alpha_1 = 0.45$, $\alpha_2 = 0.15$, $\rho \in \{0.8, 0.9\}$ with $\beta_j = 1 - \alpha_j$ for $j = 1, 2$.

	α_1	α_2	β_1	β_2	$\rho_{1,2}$
S1	0.040	0.066	-0.040	-0.073	0.100
	(0.030)	(0.033)	(0.034)	(0.043)	(0.021)
S2	0.007	-0.004	-0.012	-0.023	0.026
	(0.019)	(0.018)	(0.030)	(0.022)	(0.006)
S3	-0.009	0.048	0.016	-0.037	0.060
	(0.047)	(0.040)	(0.046)	(0.050)	(0.022)
S4	-0.050	0.015	0.054	-0.015	0.004
	(0.035)	(0.025)	(0.038)	(0.031)	(0.008)
S5	-0.048	0.023	0.054	-0.020	0.053
	(0.072)	(0.049)	(0.069)	(0.059)	(0.022)
S6	-0.061	-0.017	0.050	0.010	0.012
	(0.044)	(0.027)	(0.048)	(0.033)	(0.010)

5.6. CASE STUDY

In this section, we extend our analysis by performing a comprehensive comparison between existing models and our proposed model, using data from BIXI, which is Montréal's bike-sharing system providing open access

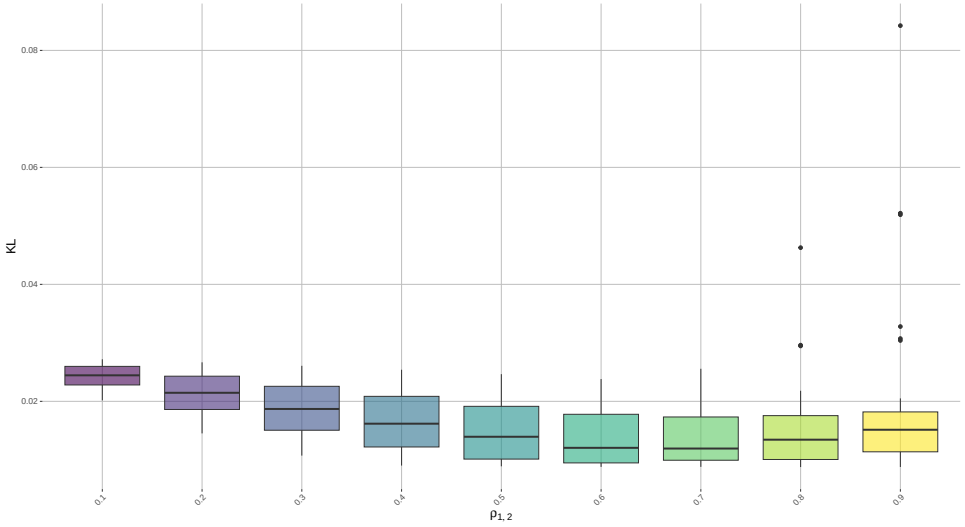


Figure 5.4: Boxplots of average KL divergence across $\alpha_j \in \{0.15, 0.3, \ldots, 0.9\}$ and $\beta_j = 1 - \alpha_j$, $j = 1, 2$, for different values of $\rho_{1,2}$. Poisson data are generated from our model.

to detailed bike usage data. The raw data, available through BIXI, records each bike rental occurrence, including specific details such as the exact date and time of departure, the departure station, the destination station, the arrival date, and time, the total duration of the bike rental, and a binary variable indicating whether the rider was a BIXI member. For this analysis, we focus on daily rentals over three months for BIXI members at three distinct departure stations. These stations are strategically located at the intersection of St. Charles and Charlotte Street (Station 1), the Jewish General Hospital (Station 2), and Place Longueuil (Station 3). During data collection, three Poisson random variables, X_1, X_2, X_3, from three stations on the same day are recorded. Each of them has 56 daily bike rental data $x_{i,k}$, $k = 1 \cdots, 56$, as detailed in Table 5.4.

First, we assess whether the data satisfy the marginal Poisson assumption. Using a chi-squared goodness-of-fit test, the p-values for data from the three stations are 0.780, 0.135, and 0.349, respectively, indicating that there is no evidence to reject the Poisson assumption. Additionally, the QQ plots in Figure 5.5 confirm that X_1, X_2, and X_3 approximately follow Poisson distributions.

Next, we estimate the correlations among the data collected from the

Table 5.4: Real data of bike rental information during 56 days.

Station 1	5	3	3	1	1	2	1	1	2	0	1	0	2	1
	2	2	0	0	3	0	0	1	1	0	1	2	1	2
	2	2	1	1	1	1	5	1	3	0	2	0	1	1
	5	0	4	1	2	0	0	1	2	0	1	2	1	2
Station 2	3	5	6	5	6	10	9	4	4	3	6	8	8	9
	6	4	4	3	8	6	10	6	5	7	9	7	16	10
	9	5	2	5	12	8	6	8	7	4	8	12	3	11
	0	12	11	11	4	4	5	13	13	2	5	3	10	6
Station 3	0	0	1	2	1	1	1	1	5	1	2	1	1	0
	0	0	0	0	0	1	0	0	1	0	1	0	1	0
	0	0	0	0	0	2	0	0	0	4	1	0	1	0
	0	0	2	2	1	1	5	0	1	1	2	0	0	0

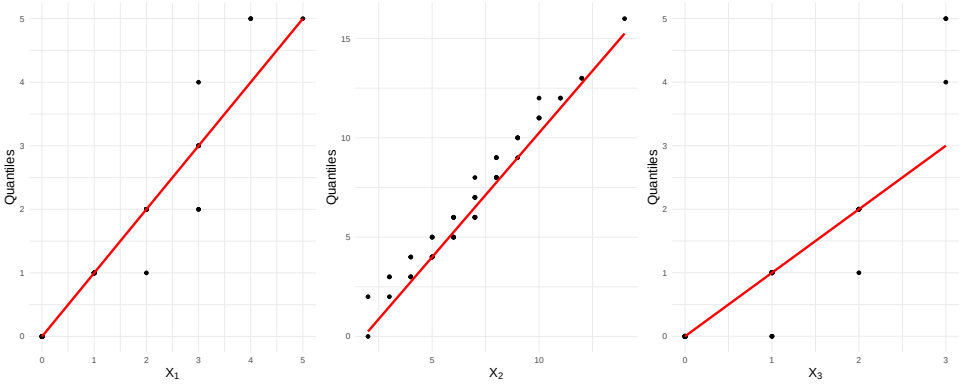


Figure 5.5: QQ plots assessing the marginal Poisson assumption for variables X_1 (left), X_2 (middle) and X_3 (right).

three stations using the sample correlation coefficient, defined as

$$\widehat{\rho}_{i,j}^{X} = \frac{\sum_{k=1}^{56} (x_{i,k} - \bar{x}_i)(x_{j,k} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{56} (x_{i,k} - \bar{x}_i)^2 \sum_{k=1}^{56} (x_{j,k} - \bar{x}_j)^2}}, \quad i, j = 1, 2, 3,$$

where $\bar{x}_i$ denotes the sample mean for station i. The estimated correlations are $\widehat{\rho}_{1,2}^{X} = -0.117$, $\widehat{\rho}_{1,3}^{X} = -0.183$, and $\widehat{\rho}_{2,3}^{X} = -0.124$. As observed, all three variables exhibit negative correlations. Consequently, we focus our further analysis on methods applicable to handle negative dependence, namely GC, TC-4, and our proposed model.

Table 5.5 summarizes our case study results on the BIXI dataset, including the number of estimated parameters, parameter estimates, log-likelihood, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Hannan–Quinn Information Criterion (HQ). Our proposed model achieves a higher log-likelihood and lower AIC, BIC, and HQ values compared to other methods, indicating more accurate modeling of daily bike rental counts across the three departure stations. These improvements enable a more reliable estimation of rental patterns, which play a critical role in operational decision-making. In practice, such enhanced modelling can inform better resource allocation, maintenance scheduling, and service planning in bike-sharing systems, ultimately providing valuable management insights for optimizing urban mobility services.

Table 5.5: Summary of three methods for modelling bike rental data.

	Number	Estimates	Log-likelihood	AIC	BIC	HQ
GC	6	$\hat{\boldsymbol{\lambda}} = [1.428, 6.893, 0.786]^{\mathsf{T}}$ $\hat{\boldsymbol{\rho}} = [-0.117, -0.183, -0.124]^{\mathsf{T}}$	-304.63	621.26	640.01	628.87
TC-4	6	$\hat{\boldsymbol{\lambda}} = [1.428, 6.893, 0.786]^{\mathsf{T}}$ $\hat{\boldsymbol{\rho}} = [-0.117, -0.185, -0.125]^{\mathsf{T}}$	-307.67	627.35	646.10	634.96
Proposed	9	$\check{\boldsymbol{\alpha}} = [1.258, 6.803, 0.063]^{\mathsf{T}}$ $\check{\boldsymbol{\beta}} = [0.171, 0.089, 0.722]^{\mathsf{T}}$ $\check{\boldsymbol{\rho}} = [-0.112, -0.197, -0.789]^{\mathsf{T}}$	-242.28	502.56	530.67	513.97

5.7. CONCLUSION

In this study, we proposed a novel model for handling multivariate Poisson variables that overcomes key limitations of existing methods such

as TR, TREx, and the Gaussian copula. Our model serves as a generalized framework, encompassing these traditional approaches as special cases through specific parameter configurations. By introducing additional parameters, it offers greater flexibility in modeling joint distributions, particularly capturing a broader range of dependencies—including negative correlations—that are challenging for conventional methods. Our theoretical analysis demonstrates that the proposed model can represent a wider variety of joint distributions. We also developed efficient estimation techniques, including a tailored EM algorithm, to facilitate practical implementation. Simulation results validate the effectiveness of our approach, and a case study further shows that our model achieves superior log-likelihood values and information criteria (AIC, BIC, and HQ) compared to established methods.

Future research can build on our work by addressing several key challenges to further enhance the modeling of dependent data. First, while our proposed model demonstrates flexibility in capturing a wide range of dependence structures, its framework can be extended to incorporate additional copula families or alternative regularization schemes. Such extensions could improve the model's ability to handle extreme tail dependence and non-standard correlation patterns often encountered in high-dimensional applications. Second, although our study focuses on multivariate Poisson distributions, many practical problems involve mixed data types (e.g., combinations of continuous and count data). Extending our framework to jointly model such mixed-type data would broaden its applicability and provide deeper insights into complex real-world systems. Finally, given the computational efficiency demonstrated by our approach, developing distributed or parallel algorithms specifically tailored to our model could significantly enhance its scalability, facilitating the analysis of massive datasets common in contemporary applications.

6

CONCLUSION

In this dissertation, we addressed several complex problems arising from data analysis, focusing on scenarios with limited and expensive data, spatiotemporal tensor data, non-normal continuous data, and multivariate Poisson data. Each chapter presented novel methods that not only address specific challenges but also contribute to complex modeling.

Chapter 2 considered an optimization problem of black-box functions whose derivatives are unknown. Such functions are also expensive to run, leading to limited data. Due to budget restrictions, we proposed a parallel sequential design using minimum energy design to estimate the optimal solution by selecting valuable points to calculate. This new approach, combined with GP models, surpasses existing one-point-at-a-time strategies like EI and HEI due to its advantages in several key areas, including time-saving, parallel simulation, and exploration-encouraging. A shrinkage-augment strategy was proposed to select more points located in the local area around the global optimum, ensuring the estimates are more reliable. Simulation and case studies show that our method can estimate the true optimum better with less time, compared with some existing methods.

Chapter 3 focused on limited data with spatiotemporal fluid flows stored as a tensor. Due to the complex nature of the data, including high dimensionality and information on spatial and temporal scales, directly applying GP models for fast predictions is not feasible. To address that, a multi-output Gaussian process via the tensor unfolding and SVD was used for the spatiotemporal data, predicting an unobserved flow with fixed spatiotemporal modes and predicted coefficients based on GP models. Some benchmark problems and two real data sets show that our method outperforms others in terms of more accurate predictions with less computation cost.

Chapter 4 and Chapter 5 considered continuous and discrete data that are not normally distributed, respectively. They both defined the specific distributions via the fact that the CDF follows a uniform distribution over

$[0, 1]$. According to Chapter 4, data are collected to evaluate a manufacturing process. Two issues were addressed together, involving asymmetric tolerances and a non-normal quality characteristic. We summarized some necessary properties of a proper PCI for asymmetric tolerances and normal data. Based on such properties, two novel process capability indices were defined. We also discussed the parametric estimation procedure and asymptotic confidence limits. Note that such proposed PCIs are still based on normal variables. To free the normality assumption, we used B-Splines to estimate the underlying unknown CDF of the data and introduced a tailored inverse transformation to normalize the data. Such a non-parametric estimator is suitable for all kinds of CDFs, with asymptotic properties also studied. Then our PCIs can be utilized to evaluate the manufacturing process regardless of the data distribution. We conducted some simulation studies and employed a real data set to show the novelty and efficiency of our method.

When there are some Poisson counts, Chapter 5 developed a new multivariate Poisson model based on MRT technology and a Gaussian copula. It was demonstrated that via our model, pairwise correlations can vary across the entire spectrum, from complete independence to maximal positive or negative correlation. This flexibility allows for a nuanced representation of different levels of dependence. Some distribution properties, like the PMF and Pearson correlation coefficient, were derived from our model. We also introduced some propositions to further highlight the novelty compared with existing methods. Three estimation procedures were explored, along with asymptotic results for corresponding estimators. The proposed model was shown to perform well through several simulations as well as real data analysis.

In conclusion, this dissertation provides a comprehensive set of tools and frameworks for addressing a variety of complicated data structures. The proposed methods not only improve upon existing approaches but also open new avenues for research in the fields of statistical modeling and data analysis.

In the future, the analytical frameworks developed in this thesis can be extended to investigate more complex problems. For instance, in scenarios involving limited datasets that follow multi-dimensional discrete distributions, a tailored PCI could be defined using joint discrete distributions. This approach, when integrated with GP models that can generate more valuable data, is beneficial for comprehensive process evaluation.

7

SUPPLEMENTARY MATERIAL

7.1. SUPPLEMENT TO CHAPTER 2

Additional simulation results considering more benchmark problems are shown in Tables S1, S2, S3 and S4.

Table S1: Logarithm optimal gaps between true optima and estimated values by different methods for some low-dimensional benchmark problems. $q = 5$ for CLmin, CLmax and PMEI.

Benchmark	p	EI	CLmin-5	CLmax-5	PMEI-5
Drop-wave	2	-0.47	-0.99	-0.99	-1.10
Levy N.13	2	0.25	-0.70	-0.70	-0.95
Schaffer	2	-0.75	-1.11	-1.11	-1.31
Bohachevsky1	2	1.72	-0.77	-0.77	-0.49
Bohachevsky2	2	1.72	-1.00	-1.00	-1.13
Bohachevsky3	2	1.72	-0.64	-0.64	-1.34
Booth	2	-3.72	-3.39	-3.39	-9.23
Matyas	2	-2.58	-5.81	-5.81	-9.48
McCormick	2	-2.31	-3.89	-3.89	-4.11
Six-hump	2	-1.52	-4.01	-4.01	-4.17
Michalewicz	2	-0.62	-3.84	-3.84	-5.42
Sphere	3	-1.43	-3.48	-3.31	-6.24
Hart3	3	-4.32	-4.74	-5.15	-4.72
Rotahted	3	-1.93	-1.00	-1.83	-3.29
Zakharov	3	0.49	0.04	0.49	0.25
Rosenbrock	3	0.15	-1.86	-1.66	-2.35

7.2. SUPPLEMENT TO CHAPTER 3

7.2.1. PROOF OF THE PROPOSITION 1

Denote $\tilde{\mathbf{V}}$ as the approximation during the second step. Given r_1 and r_2^i for $i = 1, \cdots, r_1$, the decomposition error for $\mathcal{Y}$ is defined as $\varepsilon_d = ||\mathcal{Y} - \tilde{\mathcal{Y}}||_F^2$, where $\tilde{\mathcal{Y}}$ is the final low-rank approximation. Via unfolding and transforming the vector to a matrix, the error is identical. Thus, we

Table S2: Logarithm optimal gaps between true optima and estimated values by different methods for some low-dimensional benchmark problems. $q = 10$ for CLmin, CLmax and PMEI.

Benchmark	p	EI	CLmin-10	CLmax-10	PMEI-10
Drop-wave	2	-0.47	-1.07	-1.07	-1.22
Levy N.13	2	0.25	-1.25	-1.25	-1.06
Schaffer	2	-0.75	-1.52	-1.52	-1.33
Bohachevsky1	2	1.72	-0.21	-0.21	-4.55
Bohachevsky2	2	1.72	-0.92	-0.92	-2.27
Bohachevsky3	2	1.72	-1.28	-1.28	-2.33
Booth	2	-3.72	-3.63	-3.63	-14.96
Matyas	2	-2.58	-4.63	-4.63	-13.89
McCormick	2	-2.31	-3.94	-3.94	-4.11
Six-hump	2	-1.52	-4.49	-4.49	-12.61
Michalewicz	2	-0.62	-3.75	-3.75	-5.47
Sphere	3	-1.43	-3.81	-3.49	-7.80
Hart3	3	-4.32	-4.60	-5.25	-6.27
Rotahted	3	-1.93	-1.65	-1.26	-6.05
Zakharov	3	0.49	-1.67	-0.21	-0.92
Rosenbrock	3	0.15	-2.17	-2.35	-3.80

Table S3: Logarithm optimal gaps between true optima and estimated values by different methods for some low-dimensional benchmark problems. $q = 15$ for CLmin, CLmax and PMEI.

Benchmark	p	EI	CLmin-15	CLmax-15	PMEI-15
Drop-wave	2	-0.47	-1.08	-1.08	-1.45
Levy N.13	2	0.25	-1.47	-1.47	-2.27
Schaffer	2	-0.75	-1.49	-1.49	-1.94
Bohachevsky1	2	1.72	-0.50	-0.50	-8.81
Bohachevsky2	2	1.72	-0.77	-0.77	-2.96
Bohachevsky3	2	1.72	-1.04	-1.04	-3.51
Booth	2	-3.72	-2.87	-2.87	-19.54
Matyas	2	-2.58	-3.83	-3.83	-19.29
McCormick	2	-2.31	-3.41	-3.41	-4.11
Six-hump	2	-1.52	-3.91	-3.91	-17.07
Michalewicz	2	-0.62	-3.49	-3.49	-5.40
Sphere	3	-1.43	-3.44	-3.08	-10.67
Hart3	3	-4.32	-4.15	-4.45	-6.61
Rotahted	3	-1.93	-0.38	-1.31	-8.11
Zakharov	3	0.49	-2.24	-0.19	-1.96
Rosenbrock	3	0.15	-1.87	-2.12	-3.36

Table S4: Logarithm optimal gaps between true optima and estimated values by CLmin, CLmax and PMEI for some high-dimensional benchmark problems. $q = 10, 10, 15, 15$ for $p = 4, 6, 8, 10$.

Benchmark	p	CLmin	CLmax	PMEI
Ackley	4	-0.28	-0.25	-0.29
	6	-0.14	-0.12	-0.14
	8	0.09	0.08	-0.06
	10	0.08	0.07	0.04
Griewank	4	-1.06	-1.15	-1.81
	6	-0.72	-0.57	-0.87
	8	-0.36	-0.38	-0.88
	10	-1.02	-1.06	-2.09
Schwefel	4	-2.71	-2.15	-2.62
	6	-2.54	-2.50	-2.61
	8	-1.77	-1.70	-1.88
	10	-1.61	-1.56	-1.64

have the equivalent error given by

$$
\begin{aligned}
\varepsilon_d &= ||\mathbb{Y} - \mathbf{U\Sigma\tilde{V}}^\top||_F^2 \\
&\leq ||\mathbb{Y} - \tilde{\mathbb{Y}}||_F^2 + ||\tilde{\mathbb{Y}} - \mathbf{bU\Sigma\tilde{V}}^\top||_F^2 \\
&\leq \sum_{i=r_1+1}^{R_1} \lambda_i^2 + ||\mathbf{U\Sigma V}^\top - \mathbf{U\Sigma\tilde{V}}^\top||_F^2 \\
&\leq \sum_{i=r_1+1}^{R_1} \lambda_i^2 + ||\mathbf{U\Sigma}(\mathbf{V}^\top - \tilde{\mathbf{V}}^\top)||_F^2 \\
&\leq \sum_{i=r_1+1}^{R_1} \lambda_i^2 + \sum_{i=1}^{r_1} ||\mathbf{U\Sigma}_{i,}(\mathbf{V}_{,i}^\top - \tilde{\mathbf{V}}_{,i}^\top)||_F^2 \\
&\leq \sum_{i=r_1+1}^{R_1} \lambda_i^2 + \sum_{i=1}^{r_1} ||\lambda_i \mathbf{U}_{,i}(\mathbf{V}_{,i}^\top - \tilde{\mathbf{V}}_{,i}^\top)||_F^2 \\
&\leq \sum_{i=r_1+1}^{R_1} \lambda_i^2 + \sum_{i=1}^{r_1} \lambda_i^2 U_{,i}^m ||\mathbf{V}_{,i}^\top - \tilde{\mathbf{V}}_{,i}^\top||_F^2 \\
&\leq \sum_{i=r_1+1}^{R_1} \lambda_i^2 + \sum_{i=1}^{r_1} \lambda_i^2 U_{,i}^m ||\mathbf{B}^i - \mathbf{P}^i\mathbf{D}^i{\mathbf{Q}^i}^\top||_F^2 \\
&\leq \sum_{i=r_1+1}^{R_1} \lambda_i^2 + \sum_{i=1}^{r_1} \sum_{j=r_2^i+1}^{R_2} \lambda_i^2 (\lambda_j^i)^2 U_{,i}^m,
\end{aligned}
$$

with $\tilde{\mathbb{Y}}$, the low-rank structure after the first decomposition.

7.2.2. PROOF OF THE PROPOSITION 2

Denote the ith row of a matrix $\mathbf{A}$ as $\mathbf{A}_{i,.}$, the jth column of a matrix $\mathbf{A}$ as $\mathbf{A}_{.,j}$, and the element of $\mathbf{A}$ in ith row and jth column as $A_{i,j}$. Let the ith element of a vector $\boldsymbol{A}$ be A_i. As introduced before, for a new input $\boldsymbol{u}$, the prediction at the kth location and the lth snapshot $\hat{y}_{kl}$ is

$$\hat{y}_{kl} = \mathbf{U}_{k,.} \times \mathbf{\Sigma} \times (\hat{\mathbf{V}}_{l,.})^\top.$$

Note that $\hat{\mathbb{Y}} = [\hat{y}_{kl}]$ for $k = 1, \cdots, n_x$ and $l = 1, \cdots, n_t$. Specifically, we have

$$\begin{aligned}
\hat{y}_{kl} &= [U_{k,1}, \cdots, U_{k,r_1}] \times \begin{bmatrix} \Sigma_{1,1} & & \\ & \ddots & \\ & & \Sigma_{r_1,r_1} \end{bmatrix} \times \begin{bmatrix} \hat{V}_{l,1} \\ \vdots \\ \hat{V}_{l,r_1} \end{bmatrix} \\
&= \sum_{i=1}^{r_1} U_{ki} \Sigma_{i,i} \hat{V}_{l,i} \\
&= \sum_{i=1}^{r_1} U_{ki} \Sigma_{i,i} \hat{B}_l^i \\
&= \sum_{i=1}^{r_1} U_{ki} \Sigma_{i,i} \sum_{j=1}^{r_2^i} P_{l,j}^i D_{j,j}^k \hat{Q}_j^i(\boldsymbol{u}) \\
&= \sum_{i=1}^{r_1} \sum_{j=1}^{r_2^i} U_{ki} \Sigma_{i,i} P_{l,j}^i D_{j,j}^k \hat{Q}_j^i(\boldsymbol{u}),
\end{aligned}$$

where $\hat{Q}_j^i(\boldsymbol{u})$ is the prediction at the unobserved input of the GP model, which is also $\hat{q}^{ij}(\boldsymbol{u})$ introduced previously, and other quantities are all known according to the SVD. As seen, every element of the final prediction matrix $\hat{Y}$ is a linear combination of predictors. Then the corresponding variance can be expressed as

$$\mathrm{Var}(\hat{y}_{kl}) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2^i} \hat{s}_{ij}^2(\boldsymbol{u}) (U_{k,i} \Sigma_{i,i} P_{l,j}^i D_{j,j}^i)^2,$$

with the estimated posterior standard variance $\hat{s}_{ij}(\boldsymbol{u})$. Thus, the pointwise 95% confidence interval of an unobserved spatiotemporal surface at a new design point can be computed readily.

7.2.3. SIMULATION RESULTS FOR OTHER TYPES OF PDES

FOUR BENCHMARK FUNCTIONS

Four other benchmark PDEs in [78] used here are

- The wave function.

$$\begin{aligned}
&\frac{\partial^2 w}{\partial t^2} = c^2 \frac{\partial^2 w}{\partial x^2}, \\
&\frac{\partial w}{\partial t}(0, x) = 0, \\
&w(0, x) = \exp(-\lambda x^2), \\
&w(t, -\infty) = w(t, \infty) = 0,
\end{aligned}$$

where $t \in [0, 40]$, $x \in [-40, 40]$, $\lambda \in [0, 1]$ and $c \in [0, 1]$. 20 space-filling design points are conducted for training data, along with 400 locations and 200 snapshots. Flows with another 1000 inputs are used as testing data.

- The heat function, $t \in [0, 5]$ and $x \in [0, 1]$, with the input $D \in [0, 1]$,

$$\frac{\partial Y}{\partial t} = D\frac{\partial^2 Y}{\partial x^2},$$
$$Y(t, x = 0) = 0,\ Y(t, x = 1) = 1,$$
$$Y(t = 0, x) = \sin(\pi x).$$

  10 space-filling design points are conducted as training data, with $n_x = 100$ and $n_t = 501$. Another 100 data sets are used for testing.

- The Laplace's equation , $x \in [0, 1]$ and $y \in [0, 1]$, with the input $a \in [0, 1.5]$,

$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = 0,$$
$$w(x = 0, y) = w(x, y = 0) = 0,$$
$$\frac{\partial w}{\partial y}_{y=0} = 0,\ \frac{\partial w}{\partial y}_{y=1} = \sin(a \times \pi x)\sinh(\pi).$$

  10 space-filling design points are conducted for training data, involving 100 xs and ys. w with 100 values of a are used as testing data.

- The one-dimensional Brusselator function from the chemical reactions, $t \in [0, 10]$ and $x \in [0, 1]$, with inputs $a \in [0.9, 1.2]$, $b \in [0.5, 1.5]$ and $c \in [0.5, 2]$,

$$\frac{\partial X_1}{\partial t} = D_{X_1}\delta^2 X_1 + 1 + X_1^2 X_2 - 4X_1,$$
$$\frac{\partial X_2}{\partial t} = D_{X_2}\delta^2 X_2 + 3X_1 - X_1^2 X_2,$$
$$X_1(t = 0, x) = 1 + \sin(2\pi x),\ X_2(t = 0, x) = 3,$$

  where the diffusion coefficient D is taken to be 0.02, and we take for the boundary concentrations a value of 1 and 3 for X_1 and X_2, respectively. Two responses are predicted here, namely X_1 (denoted as Brusselator-1) and X_2 (Brusselator-2). Data at 50 locations and 101 snapshots with 30 space-filling design points are collected for training, while the other 1000 flows are used for testing.

RESULTS

All results, including computation time (in seconds), number of GPs, average AMRE, and proper scoring rule (PSR), are presented in Tables S5 and S6. Overall, D-SVD can balance the costs and accuracy well, and provide valid predictions with less computation time.

Table S5: Computation time (in seconds), number of GPs, average AMRE with standard deviation in parentheses, and average PSR over 1000 test data sets with four methods for the wave function.

	Time	Number	Average AMRE	Average PSR
D-SVD	9.64	626	84.80% (41.99%)	-0.21
CPOD	53.97	4200	100.54% (50.46%)	296.81
KPOD	100.46	7920	150.27% (89.07%)	—
TPOD	6.99	454	86.06% (43.99%)	0.87

NA: the result is not available; —: the result is not suitable.

7.2.4. CREDIBLE INTERVAL RESULTS

In this section, we use the Burgers equation and the wave function as an illustration.

Figure S1 shows the true solution and the predicted results using D-SVD at $A = 0.55$ while Figure S2 presents the point-wise credible intervals at a fixed time snapshot. These intervals exhibit relatively small widths, indicating low posterior variance. These results reveal that the predicted flow can closely approximate the true solution. This success can be attributed to the proposed MOGP, which is reformulated as a series of independent GPs via the nested compression. Further, the cover probabilities (CP) of whether the 95% credible region contains the true value over 100 testing data sets at each x and t are calculated. The average CP over all combinations of x and t is 95.64%. The absolute gap between CP and 95% is shown in Figure S3. Almost all values are close to 0, further interpreting the results presented in Proposition 2.

Figure S4 provides the performance for the wave function. The proposed framework excels in predicting spatiotemporal characteristics, although it encounters challenges when there are sharp changes. This limitation is further illustrated in Figure S5. On the one hand, the worst prediction, associated with the largest MRE, shows that the credible intervals generally encompass the true values, except in regions around the two peaks. On the other hand, our model tends to perform well for

Table S6: Computation time (in seconds), number of GPs, average AMRE with standard deviation in parentheses, and average PSR over test data sets with four methods for different benchmark functions.

		Heat	Laplace	Brusselator-1	Brusselator-2
Time	D-SVD	0.41	0.22	4.73	3.04
	CPOD	4.24	0.82	3.93	4.17
	KPOD	7.31	2.34	8.89	6.21
	TPOD	0.30	0.11	3.11	2.72
Number	D-SVD	15	8	164	105
	CPOD	501	100	202	202
	KPOD	1022	220	393	262
	TPOD	10	7	117	72
Average AMRE	D-SVD	0.13% (0.12%)	0.18% (0.07%)	4.67% (3.21%)	2.59% (1.66%)
	CPOD	3.11% (0.49%)	6.30% (3.99%)	9.12% (3.79%)	32.68% (1.26%)
	KPOD	0.36% (0.24%)	2.37% (2.09%)	15.97% (77.87%)	33.42% (148.08%)
	TPOD	0.56% (0.20%)	1.47% (0.47%)	5.42% (3.21%)	2.62% (1.65%)
Average PSR	D-SVD	119	6339	-0.55	-1.03
	CPOD	NA	1.24×10^{6}	2.36×10^{7}	1.49×10^{23}
	KPOD	—	—	—	—
	TPOD	5318.08	1.61×10^{6}	8.56×10^{-2}	0.54

NA: the result is not available; —: the result is not suitable.

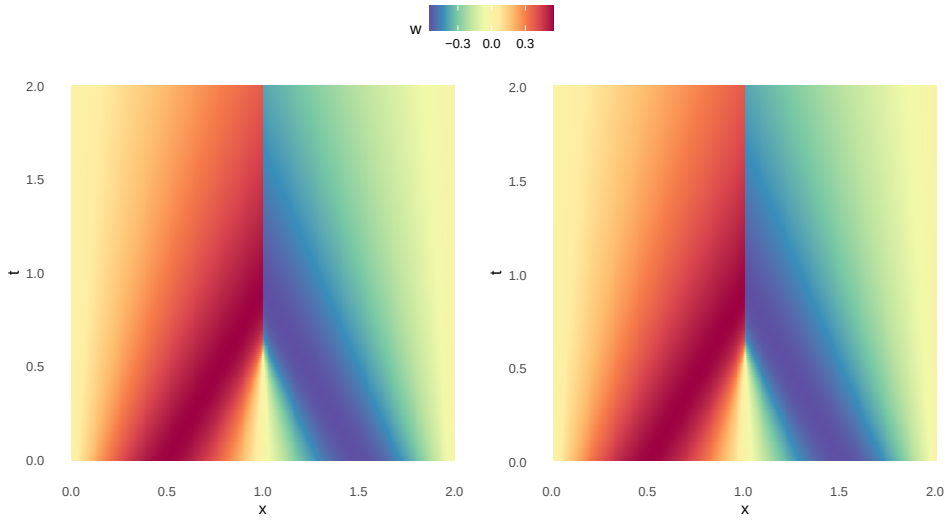


Figure S1: The distribution of the true solution (left) and predicted flows using the D-SVD (right) for the Burgers equation when $A = 0.55$.

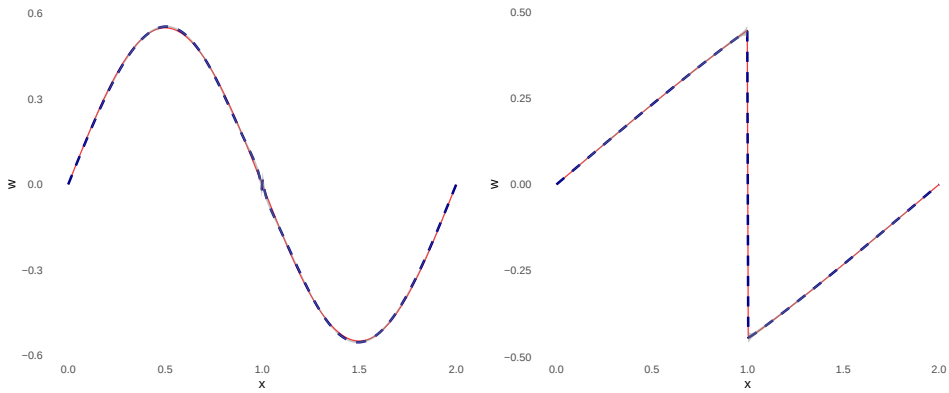


Figure S2: The true curve w (red solid line), prediction (blue dashed line), 95% credible interval (gray shadow) the Burgers equation when $A = 0.55$ at the snapshot with the largest MRE (left) and the smallest one (right).

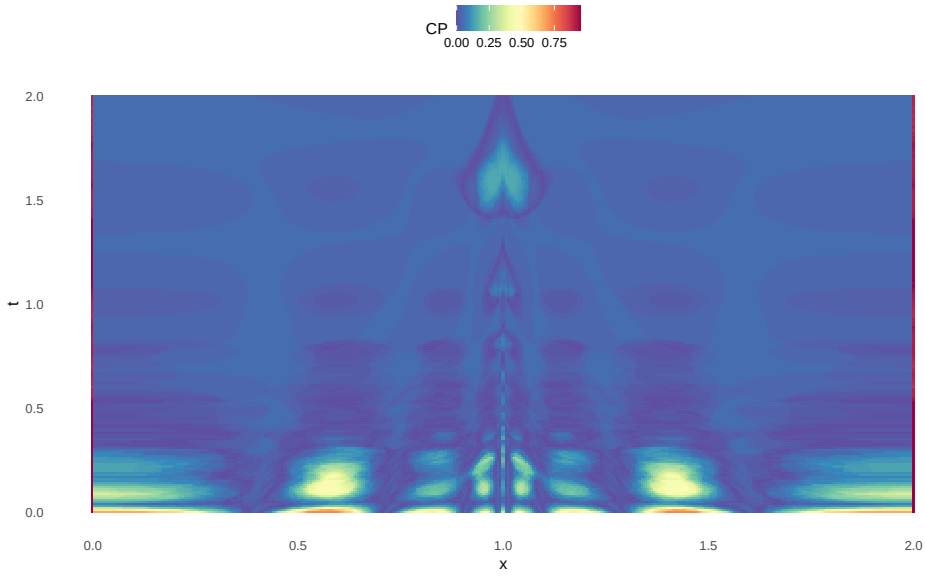


Figure S3: The absolute gap between the cover probabilities (CP) and 95% over 100 data testing sets at each x and t for the Burgers equation.

a single peak. It appears that extracting the main information through SVD may potentially discard some uncommon information, such as sudden spikes at only a few locations. Consequently, predictions and corresponding credible intervals are not accurate enough but still reliable. Similar to the Burgers equation, the average CP is 93.23% and the small absolute gap between CP and 95% over 1000 test data sets in Figure S6, both coinciding with Proposition 2.

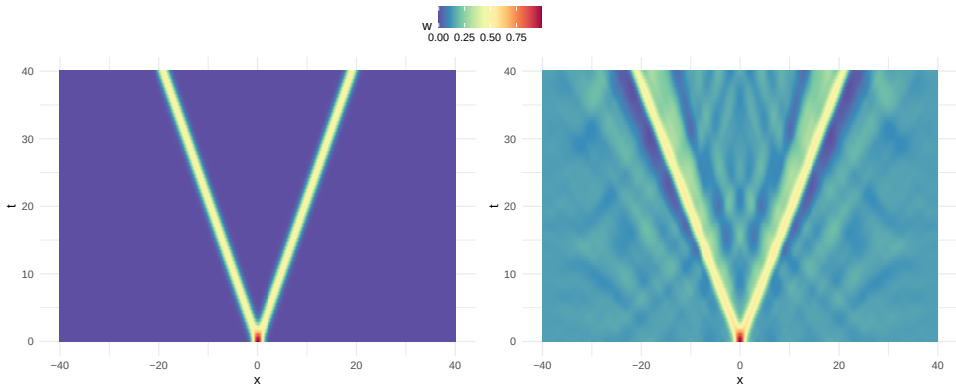


Figure S4: The distribution of the true solution (left) and predicted flows via the D-SVD (right) for the wave function when $\lambda = 0.475$ and $c = 0.825$.

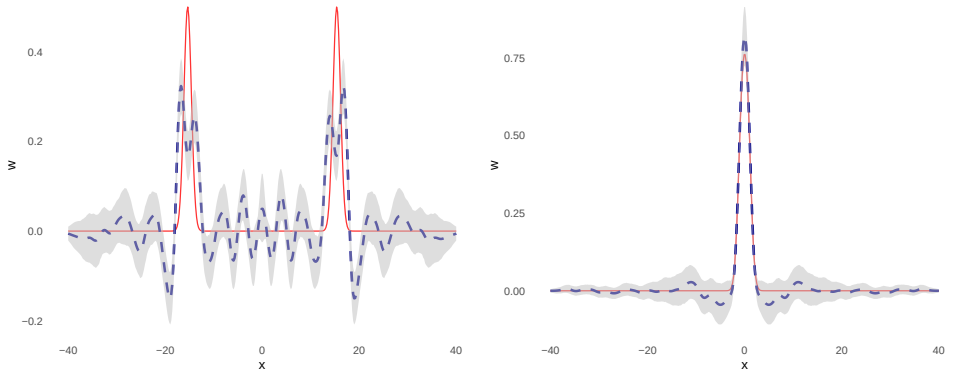


Figure S5: The true curve w (red solid line), prediction (blue dashed line), 95% credible interval (gray shadow) for the wave function when $\lambda = 0.475$ and $c = 0.825$ at the snapshot with the largest MRE (left) and the smallest one (right).

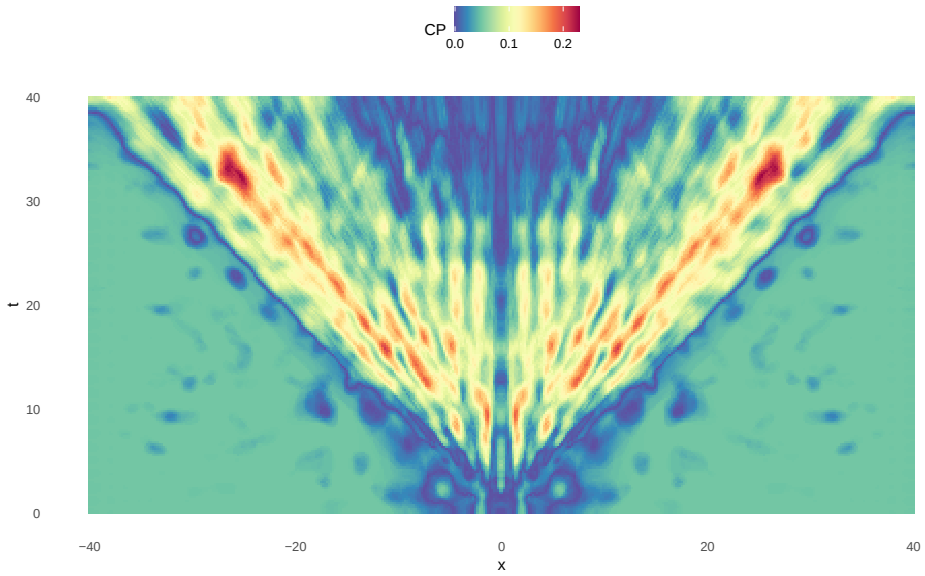


Figure S6: The absolute gap between the cover probabilities (CP) and 95% over 100 data testing sets at each x and t for the wave function.

7.2.5. PREDICTION ACCURACY WITH DIFFERENT THRESHOLDS AND CORRELATION FUNCTIONS

In this section, we present the average mean relative errors (AMRE) of D-SVD for the nonlinear Schrödinger equation. Various $K(\cdot,\cdot)$ are considered, including RBF, exponential, Matérn (5/2) and Matérn (3/2), taking $\epsilon_1 = 0.99$ and $\epsilon_2 = 0.999$ as an example. Meanwhile, different thresholds given Matérn (5/2) correlation function are also used to show how the proposed model performs with different thresholds.

Results are shown in Figures S7 and S8. Note that there seem to be ill-conditioned problems when using RBF and exponential correlation functions, so corresponding AMRE values are not available. Matérn correlations are more robust and provide comparable predictions. Similar to the Burgers equation, all combinations of thresholds can yield reliable predictions.

7.2.6. PREDICTIONS FOR THE BOUNDARY AND INITIAL CONDITIONS

In this section, we apply the Burgers equation to present how the proposed method performs when predicting the boundary and initial

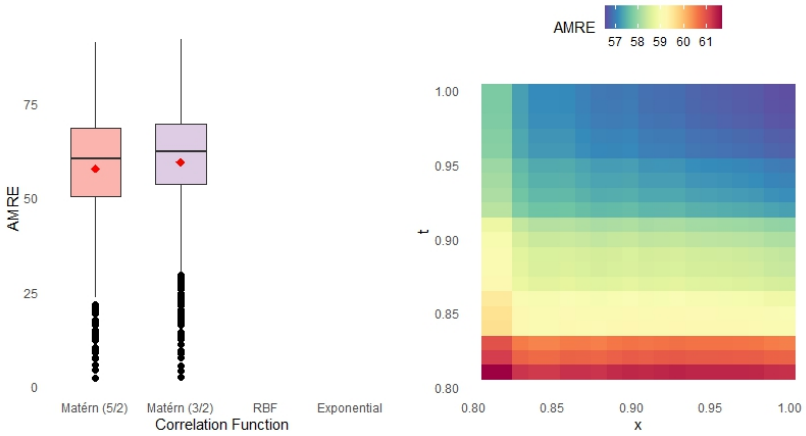


Figure S7: AMRE values of D-SVD based on 1000 testing data with different correlation functions (left) and thresholds (right) for N1, respectively. Red rhombus denotes the average AMRE.

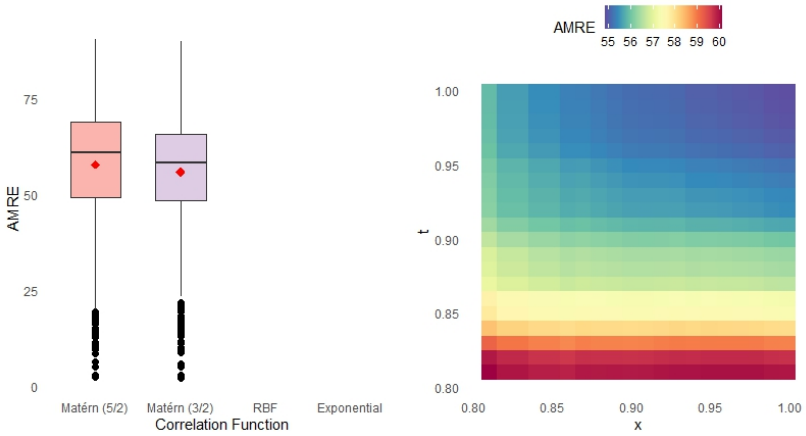


Figure S8: AMRE values of D-SVD based on 1000 testing data with different correlation functions (left) and thresholds (right) for N2, respectively. Red rhombus denotes the average AMRE.

conditions. Firstly, Figure S9 show predictions when $A = 0.55$. As seen, the predictions at $x = 0$ and $x = 1$ are almost the same, coinciding with the initial conditions relatively. Meanwhile, the predicted $w(t = 0, x)$ via our method is very close to the true boundary condition. Secondly, the

mean absolute gap (MAP)

$$\frac{1}{n_t}\sum_{j=1}^{n_t}|\hat{w}(t_j, x=0)-\hat{w}(t_j, x=2)|$$

and the mean absolute error (MAE)

$$\frac{1}{n_x}\sum_{i=1}^{n_x}|\hat{w}(0, x_i)-w(0, x_i)|$$

are calculated for each testing data set. The average MAP and MAE over 100 testing data sets are 1.22×10^{-15} and 0.004, respectively. These findings indicate our data-driven method achieves comparable accuracy for initial and boundary conditions.

7.3. SUPPLEMENT TO CHAPTER 4

7.3.1. NOTATIONS

We assume that the characteristic X ranges over $\mathcal{X}$, a compact subset of some Euclidean space. In what follows, via the sequence of knots τ_m, BSs with degree $k-1$ forms a linear space with dimension $L=m+k$, denoted as $\mathcal{G}_m^k=\mathcal{G}(k,\tau_m)$. For any $g_1, g_2\in\mathcal{G}_m^k$, set $\langle g_1, g_2\rangle=\mathbb{E}[g_1(x)g_2(x)]$ and $\langle g_1, g_2\rangle_n=\sum_{i=1}^n[g_1(x_i)g_2(x_i)]/n$.

7

7.3.2. ASSUMPTIONS

We need the following technical assumptions for our results.

(1) The distribution function y is k times continuously differentiable for some $k\geq2$, and has bounded $(k-1)$th derivative.

(2) The density of X is bounded away from zero and infinity on $\mathcal{X}$.

(3) The knot sequence $\tau_m=\{s_0<s_1\cdots<s_{m+1}\}$ has a bounded mesh ratio. That is, there exists a constant c such that

$$\frac{\max\limits_{0\leq i\leq m} s_{i+1}-s_i}{\min\limits_{0\leq i\leq m} s_{i+1}-s_i}\leq c.$$

And $\max\limits_{1\leq i\leq m}|h_{i+1}-h_i|=o(m^{-1})$ with $h_i=s_i-s_{i-1}$.

(4) $m\geq Cn^{1/(2k+1)}$ for some constant $C>0$.

(5) $m\to\infty$ and as $n\to\infty$, $m/n\to0$, $n^{1/(2k+1)}/L\to0$, and $L\log n/n\to0$.

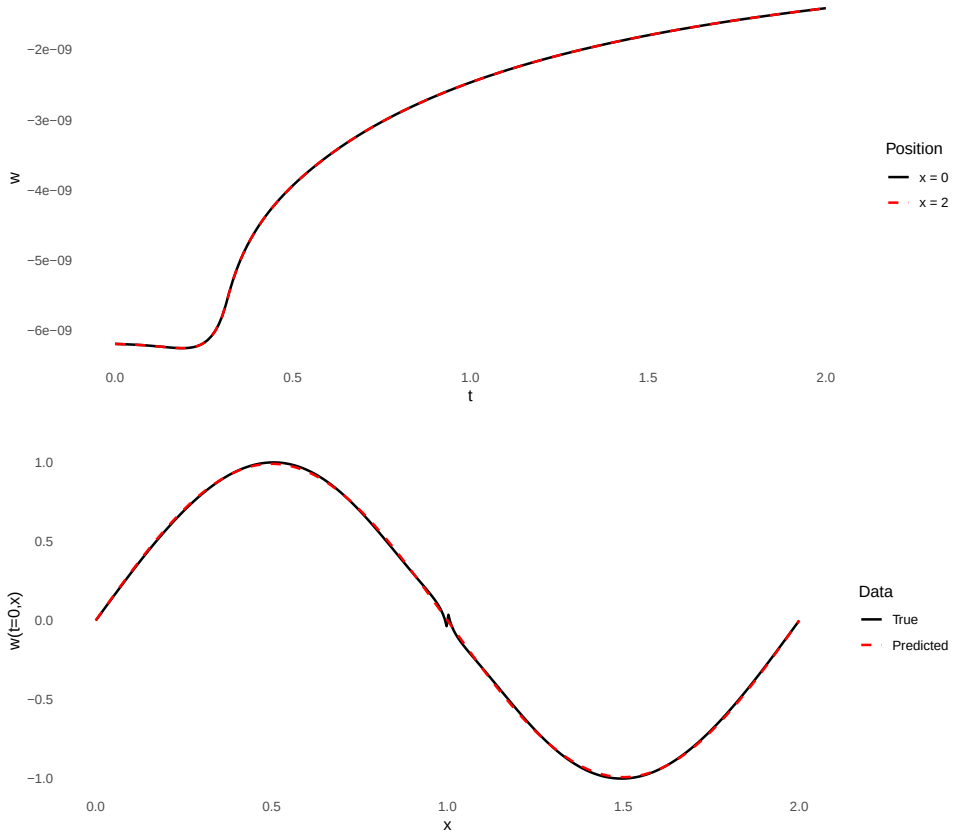


Figure S9: Predicted initial conditions (top) and the comparison between true and predicted boundary conditions (bottom) for the Burgers equation. $A = 0.55$.

7.3.3. PROOF OF THEOREM 1

As noted by [122], asymptotic properties are the same whether the constraints of monotone non-decreasing are applied or not. Therefore, we consider an unconstrained estimator here, which is also denoted as $\hat{y}$. Regarding the transformed ECDF introduced in Section 4.4 as responses, we have

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{b}^\top V \boldsymbol{b})^{-1} \boldsymbol{b}^\top \mathbf{V} \boldsymbol{y}_n,$$

where $\boldsymbol{b} = [\mathbf{b}_1, \cdots, \mathbf{b}_L]$,

$$\mathbf{b}_l = [B_l(x_1), \cdots, B_l(x_n), B_l(LSL), B_l(USL)]^\top,$$

$\mathbf{V} = \Sigma_e^{-1}$, and $\boldsymbol{y}_n = [y_n(x_1), \cdots, y_n(x_n), y_n(LSL), y_n(USL)]^\top$.

First, we consider the asymptotic distributions, the estimator $\hat{y}(\cdot)$ is given by

$$\hat{y}(x) = \sum_{l=1}^{L} \hat{\beta}_l B_l(x) = \hat{\boldsymbol{\beta}}^\top \boldsymbol{B}(x),$$

where $\hat{\boldsymbol{\beta}} = [\hat{\beta_1}, \cdots, \hat{\beta_L}]^\top$ and $\boldsymbol{B}(x) = [B_1(x), \cdots, B_L(x)]^\top$. Denote the true coefficients as $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_L]^\top$, we have

$$\begin{aligned}
\mathbb{E}[\hat{y}(x) - y(x)] &= \mathbb{E}[\hat{\boldsymbol{\beta}}^\top \boldsymbol{B}(x) - \boldsymbol{\beta}^\top \boldsymbol{B}(x)] \\
&= \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{B}(x)] \\
&= ((\boldsymbol{b}^\top \mathbf{V} \boldsymbol{b})^{-1} \boldsymbol{b}^\top \mathbf{V} \mathbb{E}[\boldsymbol{y}_n] - \boldsymbol{\beta})^\top \boldsymbol{B}(x) \\
&= ((\boldsymbol{b}^\top \mathbf{V} \boldsymbol{b})^{-1} \boldsymbol{b}^\top \mathbf{V} \boldsymbol{b} \boldsymbol{\beta} - \boldsymbol{\beta})^\top \boldsymbol{B}(x) \\
&= 0.
\end{aligned}$$

Thus, $\hat{y}(\cdot)$ is an unbiased estimator. It then follows that

$$\begin{aligned}
\mathrm{Var}[\hat{y}(x) - y(x)] &= \mathrm{Var}[\hat{y}(x)] \\
&= \boldsymbol{B}(x)^\top \mathrm{Var}[(\boldsymbol{b}^\top \mathbf{V} \boldsymbol{b})^{-1} \boldsymbol{b}^\top \mathbf{V} \boldsymbol{y}_n] \boldsymbol{B}(x) \\
&= \boldsymbol{B}(x)^\top (\boldsymbol{b}^\top \mathbf{V} \boldsymbol{b})^{-1} \boldsymbol{b}^\top \mathbf{V} \mathrm{Var}[\boldsymbol{y}_n] \mathbf{V} \boldsymbol{b} (\boldsymbol{b}^\top \mathbf{V} \boldsymbol{b})^{-1} \boldsymbol{B}(x) \\
&= \boldsymbol{B}(x)^\top (\boldsymbol{b}^\top \mathbf{V} \boldsymbol{b})^{-1} \boldsymbol{b}^\top \Sigma_e^{-1} \Sigma_e \mathbf{V} \boldsymbol{b} (\boldsymbol{b}^\top \mathbf{V} \boldsymbol{b})^{-1} \boldsymbol{B}(x) \\
&= \boldsymbol{B}(x)^\top (\boldsymbol{b}^\top \mathbf{V} \boldsymbol{b})^{-1} \boldsymbol{B}(x).
\end{aligned}$$

Under assumptions (1)-(5), Theorems 3.1 and 4.1 in [170] indicate that the first conclusion follows.

Secondly, consider the linear space $\mathcal{G}_m^k$ and define $\tilde{y} = \mathbb{E}(\hat{y}|x)$, we have the decomposition

$$\hat{y} - y = [\hat{y} - \tilde{y}] + [\tilde{y} - y],$$

where $\hat{y} - \tilde{y}$ and $\tilde{y} - y$ are referred to as the approximation and estimation errors, respectively. Denote $C^k(\mathcal{X})$ as a set of functions that are k times

continuously differentiable on $\mathcal{X}$ for some $k \geq 2$. For $y \in C^k(\mathcal{X})$, [171] indicates that there exist a $g \in \mathcal{G}_m^k$ and a constant c such that

$$||g - y||_\infty \leq cm^{-k},$$

where $||\phi(x)||_\infty = \sup_{x\in\mathcal{X}} \phi(x)$ for any function $\phi \in C^k(\mathcal{X})$. According to Lemma 5.1 and Theorem 5.1 in [172], it then follows that

$$||\tilde{y} - y||_\infty \leq \inf_{g\in\mathcal{G}_m^k} ||g - y||_\infty \leq cm^{-k}.$$

Therefore, $||\tilde{y} - y||_\infty = O_p(m^{-k})$. Based on Corollary 3.1 in [172], we have $\hat{y}(x) - \tilde{y}(x) = O_p(\sqrt{m/n})$ uniformly in $x \in \mathcal{X}$. Therefore, $||\hat{y} - y||_\infty = O_p(m^{-k} + \sqrt{m/n})$. The result on the L_2 norm follows directly from the one on the supremum norm.

7.3.4. PROOF OF COROLLARY 2

Via the asymptotic normality of $\hat{y}$, one can easily obtain the asymptotic distributions of $\hat{F}_p$ based on the delta method. Even though the convergence rate of $\hat{F}_p$ is more complicated due to the transformation $y(x) = \ln[F(x)/(1-F(x))]$, we can demonstrate that it is not slower than $\hat{y}$ when n is sufficiently large. Let $h(y) = \exp(y)/(1+\exp(y))$, we have

$$h'(y) = \lim_{|y_1-y_2|\to 0} \frac{h(y_1)-h(y_2)}{y_1-y_2} = \frac{\exp(y)}{(\exp(y)+1)^2} < 1.$$

Thus, there exists $\delta > 0$ such that if $|y_1 - y_2| < \delta$, $|h(y_1) - h(y_2)| < |y_1 - y_2|$. According to the proof of Theorem 1, for $\forall \varepsilon_1 > 0$, there exists a sufficient large constant $C > 0$ and $N_1 > 0$, such that $\forall n > N_1$ and $\forall c > C$

$$\sup_{x\in\mathcal{X}} P(|\hat{y}(x) - \tilde{y}(x)| \geq c\sqrt{\frac{m}{n}}) < \varepsilon_1,$$

where X ranges over $\mathcal{X}$. Under the assumption (5), for $\forall \varepsilon_2/c > 0$, there exists a sufficient large constant $N_2 > 0$ such that $\forall n > N_2$, $\sqrt{m/n} < \varepsilon_2/c$. Thus, when $|\hat{y}(x) - \tilde{y}(x)| < \min\{\delta, \varepsilon_2/c\}$, there exists $N > \max\{N_1, N_2\}$ such that

$$\sup_{x\in\mathcal{X}} P(|\hat{y}(x) - \tilde{y}(x)| \geq \min\{\delta, \varepsilon_2\}) < \varepsilon_1,$$

implies

$$\sup_{x\in\mathcal{X}} P(|\hat{F}_p(x) - \tilde{F}(x)| \geq \min\{\delta, \varepsilon_2\}) < \varepsilon_1.$$

This completes the proof.

7.3.5. SIMULATION RESULTS

In this section, three types of distributions are considered.

- **Weibull Distribution:** We evaluate $WB(k, \lambda)$ with shape $k \in \{0.5, 1, 2\}$ and scale $\lambda \in \{1, 2\}$. This selection captures the different PDF shapes that arise when $k \leq 1$ versus $k > 1$.
- **Gamma Distribution:** We consider $GA(\theta, \eta)$ with shape $\theta \in \{0.5, 1, 2\}$ and rate $\eta \in \{1, 2\}$, since the gamma PDF also varies significantly based on whether $\theta \geq 1$.
- **Log-normal Distribution:** We analyze $LN(\mu, \sigma^2)$ with $\mu \in \{0, 1\}$ and scale $\sigma \in \{0.5, 1, 2\}$, recognizing that the scale parameter strongly influences the PDF's shape.

For each of these distributions, we consider 20 sample sizes ($n \in \{50, 100, \ldots, 1000\}$), resulting in a total of 360 simulation scenarios. This comprehensive setup ensures a thorough evaluation of our method under a wide range of conditions. We also consider ECDF, kernel estimation, and CBS in [122] as competitors. The complete set of PCI results is presented in Figures S10, S11, and S12. All results show that kernel estimation can not provide PCIs for three distribution types with a small sample size, while CBS fails for all scenarios. This occurs primarily because they tend to introduce biases near or beyond the data range boundaries, leading to ∞ during the inverse transformation. Although ECDF exhibits greater robustness, its accuracy remains limited for small sample sizes. The reason is that in such cases the tolerances are out of the data range and ECDF can not estimate $F(LSL)$ and $F(USL)$ very well. It is worth noting that ECDF achieves comparable LMSEs to our proposed method for large sample sizes with heavy-tailed distributions, as illustrated in Figure S12. This suggests that ECDF can effectively capture the inherent shape of the CDF given sufficient data. Our method, grounded in sufficient conditions for nonnegativity and monotonicity, may exhibit slight deviations in localised regions such as the tails of the CDF under specific conditions. Nevertheless, it offers distinct advantages in terms of flexibility and robustness, consistently providing reliable PCI values across diverse scenarios. This comprehensive performance establishes our approach as a more versatile and practical choice for real-world applications.

7.4. SUPPLEMENT TO CHAPTER 5

7.4.1. PROOF OF THEOREM 3

Without loss of generality, we can assume that $d = 2$. The main tool of the proof is the probability-generating function (see e.g. [173]). It is defined as $G(s_1, s_2) = \mathbb{E}[s_1^{V_1} s_2^{V_2}] = \sum_{i=0}^{+\infty} \sum_{j=0}^{+\infty} s_1^i s_2^j \Pr(V_1 = i, V_2 = j)$.

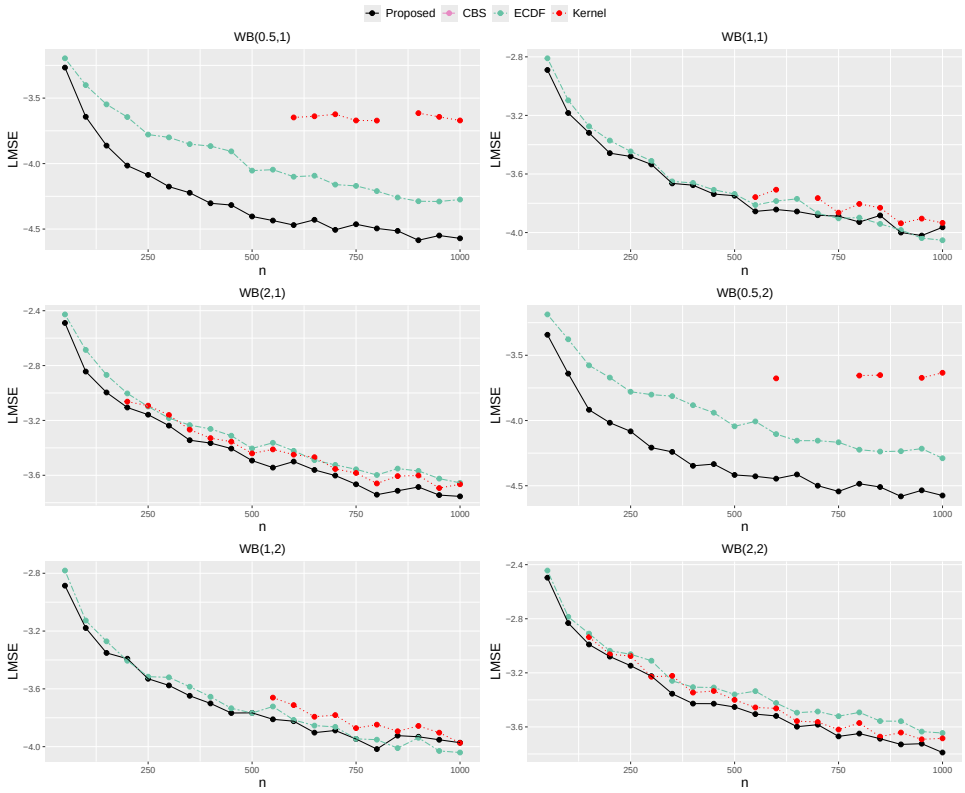


Figure S10: The average LMSE_{PCI} of different methods based on 20 different sample sizes $n \in \{50, 100, \cdots, 1000\}$ for various Weibull distributions.

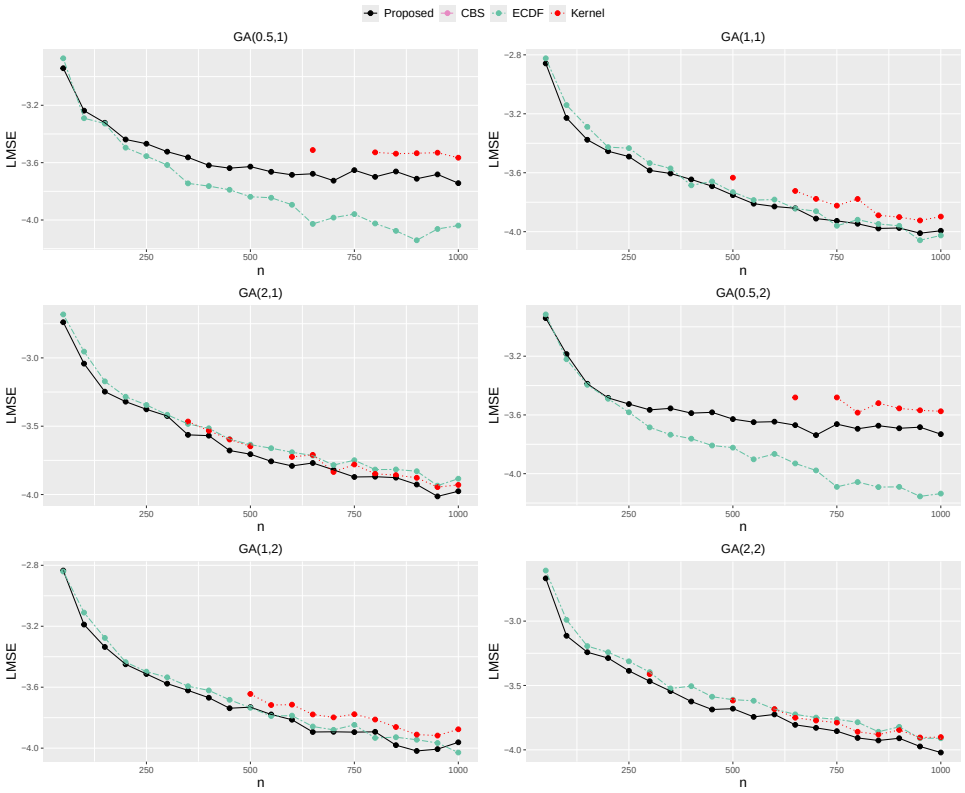


Figure S11: The average LMSE_{PCI} of different methods based on 20 different sample sizes $n \in \{50, 100, \cdots, 1000\}$ for various gamma distributions.

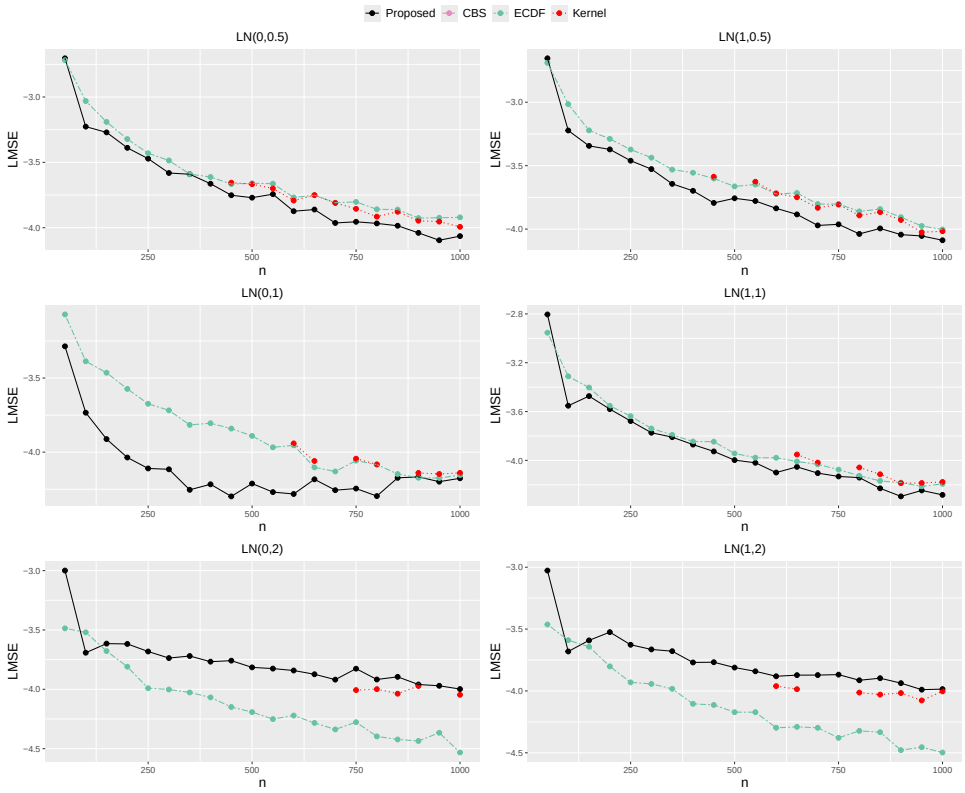


Figure S12: The average LMSE_{PCI} of different methods based on 20 different sample sizes $n \in \{50, 100, \cdots, 1000\}$ for various log-normal distributions.

To calculate the joint probabilities of two Poisson counts $\Pr(V_1 = i, V_2 = j)$, we rewrite them in terms of the complementary CDF. In particular, for any $i, j \in \mathbb{N}$,

$$\begin{aligned}\Pr(V_1 = i, V_2 = j) = \Pr(V_1 \geq i, V_2 \geq j) + \Pr(V_1 \geq i+1, V_2 \geq j+1) \\ - \Pr(V_1 \geq i+1, V_2 \geq j) - \Pr(V_1 \geq i, V_2 \geq j+1).\end{aligned}$$

Recall that $(V_1, V_2) := (F_{\beta_1}^{-1}(\Phi(Z_1)), F_{\beta_2}^{-1}(\Phi(Z_2)))$, it then follows that

$$\begin{aligned}&\sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} s_1^i s_2^j \Pr(V_1 \geq i, V_2 \geq j) \\ &= \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} s_1^i s_2^j \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i-1)), Z_2 > \Phi^{-1}(F_{\beta_2}(j-1))) \\ &= s_1^0 s_2^0 \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(-1)), Z_2 > \Phi^{-1}(F_{\beta_2}(-1))) \\ &\quad + \sum_{i=1}^{+\infty}\sum_{j=1}^{+\infty} s_1^i s_2^j \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i-1)), Z_2 > \Phi^{-1}(F_{\beta_2}(j-1))) \\ &\quad + \sum_{i=1}^{+\infty} s_1^i s_2^0 \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i-1)), Z_2 > \Phi^{-1}(F_{\beta_2}(-1))) \\ &\quad + \sum_{j=1}^{+\infty} s_1^0 s_2^j \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(-1)), Z_2 > \Phi^{-1}(F_{\beta_2}(j-1))) \\ &= 1 + \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} s_1^{i+1} s_2^{j+1} \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i)), Z_2 > \Phi^{-1}(F_{\beta_2}(j))) \\ &\quad + \sum_{i=0}^{+\infty} s_1^{i+1} \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i))) + \sum_{j=0}^{+\infty} s_2^{j+1} \Pr(Z_2 > \Phi^{-1}(F_{\beta_2}(j))).\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} s_1^i s_2^j \Pr(V_1 \geq i+1, V_2 \geq j) \\ &= \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} s_1^i s_2^j \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i)), Z_2 > \Phi^{-1}(F_{\beta_2}(j-1))) \\ &= \sum_{i=0}^{+\infty}\sum_{j=1}^{+\infty} s_1^i s_2^j \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i)), Z_2 > \Phi^{-1}(F_{\beta_2}(j-1))) \\ &\quad + \sum_{i=0}^{+\infty} s_1^i s_2^0 \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i)), Z_2 > \Phi^{-1}(F_{\beta_2}(-1)))\end{aligned}$$

$$= \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} s_1^i s_2^{j+1} \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i)), Z_2 > \Phi^{-1}(F_{\beta_2}(j)))$$
$$+ \sum_{i=0}^{+\infty} s_1^i \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i))).$$

And

$$\sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} s_1^i s_2^j \Pr(V_1 \geq i, V_2 \geq j+1)$$
$$= \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} s_1^i s_2^j \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i)), Z_2 > \Phi^{-1}(F_{\beta_2}(j-1)))$$
$$= \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} s_1^{i+1} s_2^j \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i)), Z_2 > \Phi^{-1}(F_{\beta_2}(j)))$$
$$+ \sum_{j=0}^{+\infty} s_2^j \Pr(Z_2 > \Phi^{-1}(F_{\beta_2}(j))).$$

Combine them all, and it follows that

$$G(s_1, s_2) = \mathbb{E}[s_1^{V_1} s_2^{V_2}]$$
$$= 1 + (s_1 - 1)\sum_{i=0}^{+\infty} s_1^i \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i)))$$
$$+ (s_2 - 1)\sum_{j=0}^{+\infty} s_2^j \Pr(Z_2 > \Phi^{-1}(F_{\beta_2}(j)))$$
$$+ \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty}(s_1 - 1)(s_2 - 1)\Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i)), Z_2 > \Phi^{-1}(F_{\beta_2}(j))).$$

It is straightforward to verify that

$$\mathbb{E}[V_1 V_2] = \frac{\partial^2 G(s_1, s_2)}{\partial s_1 \partial s_2}\Big|_{s_1 = s_2 = 1}$$
$$= \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} \Pr(Z_1 > \Phi^{-1}(F_{\beta_1}(i)), Z_2 > \Phi^{-1}(F_{\beta_2}(j))) \tag{7.1}$$
$$= \sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} \Pr(V_1 \geq i+1, V_2 \geq j+1).$$

7.4.2. PROOF OF PROPOSITION 3

To establish the continuity of the function $\rho^X_{i,j}$ for $i, j = 1, \cdots, d,\ i \le j$, it is sufficient to demonstrate the continuity of the function $\mathbb{E}[V_i V_j]$ with respect to $\rho_{i,j}$, β_i and β_j. According to (7.1), it follows that

$$\begin{aligned}\mathbb{E}[V_i V_j] &= \sum_{l=0}^{+\infty}\sum_{k=0}^{+\infty} \bar{\Phi}_{\rho_{i,j}}(\Phi^{-1}(F_{\beta_i}(l)), \Phi^{-1}(F_{\beta_j}(k))) \\ &= \sum_{l=0}^{+\infty}\sum_{k=0}^{+\infty} \Pr(V_i \ge l+1, V_j \ge k+1).\end{aligned}$$

Firstly, we show that $\bar{\Phi}_{\rho_{i,j}}(\Phi^{-1}(F_{\beta_i}(l)), \Phi^{-1}(F_{\beta_j}(k)))$ is continuous, which is straightforward due to the facts that $\beta \mapsto F_\beta$, Φ^{-1}, and $\rho_{i,j} \mapsto \Phi_{\rho_{i,j}}$ are all continuous. Secondly, we prove that the infinite sum converges to $\mathbb{E}[V_i V_j]$ in a uniform way. Note that for any i and j

$$\begin{aligned}\Pr(V_i \ge l+1, V_j \ge k+1) \le \Pr(V_i \ge l+1) &= \frac{1}{\sqrt{2\pi}}\int_{l+1}^{\infty} 1 \times e^{\frac{-v_i^2}{2}} dv_i \\ &\le \frac{1}{\sqrt{2\pi}}\int_{l+1}^{+\infty} \frac{v_i}{l+1} \times e^{\frac{-v_i^2}{2}} dv_i \\ &= \frac{1}{(l+1)\sqrt{2\pi}} e^{-\frac{(l+1)^2}{2}} \le e^{-\frac{(l+1)^2}{2}} \le e^{-\frac{1}{2}} \times e^{-\frac{l}{2}},\end{aligned}$$

for $l = 0, 1, \cdots$. Since $\sum_{l=0}^{+\infty} \exp(-l/2) = 1/(1-\exp(-1/2))$, convergence of the infinite sum is uniform according to the Weierstrass M-test. Finally, we conclude that $\mathbb{E}[V_i V_j]$ is continuous on the domain $(0, +\infty)^4 \times (-1, 1)$, combining with the dominated convergence theorem.

7.4.3. PROOF OF PROPOSITION 4

Via our model, we have

$$\mathrm{Corr}(X_i, X_j) = \frac{\mathrm{Cov}(V_i, V_j)}{\sqrt{(\alpha_i+\beta_i)(\alpha_j+\beta_j)}} = \mathrm{Corr}(V_i, V_j)\frac{\sqrt{\beta_i\beta_j}}{\sqrt{(\alpha_i+\beta_i)(\alpha_j+\beta_j)}}.$$

Proving the property of $\mathrm{Corr}(X_i, X_j)$ is equivalent to showing Proposition 4 holds for $\mathrm{Corr}(V_i, V_j)$. Adapting Propositions 1 and Theorem 1 in [174] to the case of Poisson marginals, it is straightforward that $\mathrm{Corr}(V_i, V_j)$ is nondecreasing in $\rho_{i,j} \in [-1, 1]$ and $\rho_{i,j} \le (\ge) 0$ if $\mathrm{Corr}(V_i, V_j) \le (\ge) 0$. Via Proposition 2 in [174], $\mathrm{Corr}(V_i, V_j)$ will be minimal, maximal and 0 when $\rho_{i,j}$ is -1, 1, and 0. This completes the proof.

7.4.4. PROOF OF PROPOSITION 5

The first part of the proposition is straightforward since α_i and α_j are both in the denominator. For the second part, we know that $\mathbb{E}[V_iV_j] = \sum_{m=0}^{+\infty}\sum_{n=0}^{+\infty}\Pr(Z_i > \Phi^{-1}(F_{\beta_i}(m)), Z_j > \Phi^{-1}(F_{\beta_j}(n)))$. As seen, $\bar{\Phi}_{\rho_{i,j}}(x,y) > \bar{\Phi}_{\rho_{i,j}}(x',y')$ if $x < x'$ and $y < y'$, and $\Phi^{-1}(x) < \Phi^{-1}(x')$ if $x < x'$. Therefore, it is equivalent to prove that $F_{\beta_i}(m)$ or $F_{\beta_j}(n)$ is strictly decreasing with respect to β_i or β_j. Take the former as an example, $F_{\beta_i}(m) = e^{-\beta_i}\sum_{k=0}^{m}\beta_i^k/k!$. It then follows that

$$\begin{aligned}\frac{\partial F_{\beta_i}(m)}{\partial \beta_i} &= -e^{-\beta_i} + \frac{\partial}{\partial \beta_i}e^{-\beta_i}\sum_{k=1}^{m}\frac{\beta_i^k}{k!}\\ &= -e^{-\beta_i} - e^{-\beta_i}\sum_{k=1}^{m}\frac{\beta_i^k}{k!}e^{-\beta_i} + e^{-\beta_i}\sum_{k=1}^{m}\frac{\beta_i^{k-1}}{(k-1)!}\\ &= -e^{-\beta_i}\frac{\beta_i^m}{m!} < 0,\end{aligned}$$

which completes the proof.

7.4.5. PROOF OF PROPOSITION 6

$\Leftarrow$: If (X_i, X_j) has the same distribution as $(F_{\lambda_i}^{-1}(U), F_{\lambda_j}^{-1}(U))$, i.e., $\Pr(X_i = m, X_j = n) = \Pr(F_{\lambda_i}^{-1}(U) = m, F_{\lambda_j}^{-1}(U) = n)$, $m, n \in \mathbb{N}$, then

$$\begin{aligned}\mathrm{Corr}(X_i, X_j) &= \frac{\mathbb{E}[X_iX_j] - \lambda_i\lambda_j}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}}\\ &= \frac{\sum_m\sum_n mn\Pr(X_i = m, X_j = n) - \lambda_i\lambda_j}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}}\\ &= \frac{\sum_m\sum_n mn\Pr(F_{\lambda_i}^{-1}(U) = m, F_{\lambda_j}^{-1}(U) = n) - \lambda_i\lambda_j}{\sqrt{\mathrm{Var}(F_{\lambda_i}^{-1}(U))\mathrm{Var}(F_{\lambda_j}^{-1}(U))}}\\ &= \mathrm{Corr}[F_{\lambda_i}^{-1}(U), F_{\lambda_j}^{-1}(U)].\end{aligned}$$

$\Rightarrow$: If $\mathrm{Corr}[X_i, X_j] = \mathrm{Corr}[F_{\lambda_i}^{-1}(U), F_{\lambda_j}^{-1}(U)]$, we obtain $\mathbb{E}[X_iX_j] = \mathbb{E}[F_{\lambda_i}^{-1}(U)F_{\lambda_j}^{-1}(U)]$ since these have the same marginal distributions (and therefore the same expectations). Hence, we have

$$\sum_{m=0}^{+\infty}\sum_{n=0}^{+\infty}mn\Pr(X_i = m, X_j = n) = \sum_{m=0}^{+\infty}\sum_{n=0}^{+\infty}mn\Pr(F_{\lambda_i}^{-1}(U) = m, F_{\lambda_j}^{-1}(U) = n). \quad (7.2)$$

Via Lemma 2.2 and Theorem 2.5 in [144], $(F_{\lambda_i}^{-1}(U), F_{\lambda_j}^{-1}(U))$ has the CDF H^* such that $\Pr(X_i = m, X_j = n) \le H^*(m, n)$ for any $m, n = 0, 1, \cdots$.

7

Combining this with (7.2), it follows that $\Pr(X_i = m, X_j = n) = \Pr(F_{\lambda_i}^{-1}(U) = m, F_{\lambda_j}^{-1}(U) = n)$. That completes the proof.

7.4.6. PROOF OF PROPOSITION 7

We begin the proof with a simple model inspired by the TR model. Assuming $d = 3$, let

$$\begin{aligned} X_1 &= Y_1 + Y_{1,2} + Y_{1,3} + Y_{1,2,3}, \\ X_2 &= Y_2 + Y_{1,2} + Y_{2,3} + Y_{1,2,3}, \\ X_3 &= Y_3 + Y_{1,3} + Y_{2,3} + Y_{1,2,3}, \end{aligned} \tag{7.3}$$

where Y_1, Y_2, Y_3, $Y_{1,2}$, $Y_{1,3}$, $Y_{2,3}$ and $Y_{1,2,3}$ are mutually independent Poisson variables with parameters η_1, η_2, η_3, $\eta_{1,2}$, $\eta_{1,3}$, $\eta_{2,3}$ and $\eta_{1,2,3}$. According to the infinite divisibility of Poisson distributions, X_1, X_2 and X_3 are positively correlated Poisson counts with parameters $\eta_1 + \eta_{1,2} + \eta_{1,3} + \eta_{1,2,3}$, $\eta_2 + \eta_{1,2} + \eta_{2,3} + \eta_{1,2,3}$ and $\eta_3 + \eta_{1,3} + \eta_{2,3} + \eta_{1,2,3}$.

Given marginals λ_1, λ_2 and λ_3, and correlations falling in $(0, 1)$, namely $\rho^X_{1,2}$, $\rho^X_{1,3}$ and $\rho^X_{2,3}$, we have

$$\begin{aligned} \eta_1 + \eta_{1,2} + \eta_{1,3} + \eta_{1,2,3} &= \lambda_1, \\ \eta_2 + \eta_{1,2} + \eta_{2,3} + \eta_{1,2,3} &= \lambda_2, \\ \eta_3 + \eta_{1,3} + \eta_{2,3} + \eta_{1,2,3} &= \lambda_3, \\ \eta_{1,2} + \eta_{1,2,3} &= \rho^X_{1,2}\sqrt{\lambda_1\lambda_2}, \\ \eta_{1,3} + \eta_{1,2,3} &= \rho^X_{1,3}\sqrt{\lambda_1\lambda_3}, \\ \eta_{2,3} + \eta_{1,2,3} &= \rho^X_{2,3}\sqrt{\lambda_2\lambda_3}. \end{aligned}$$

Then it is easy to show that solutions are

$$\begin{aligned} \eta_{1,2,3} &= \eta_{1,2,3}, \\ \eta_{1,2} &= \rho^X_{1,2}\sqrt{\lambda_1\lambda_2} - \eta_{1,2,3}, \\ \eta_{1,3} &= \rho^X_{1,3}\sqrt{\lambda_1\lambda_3} - \eta_{1,2,3}, \\ \eta_{2,3} &= \rho^X_{2,3}\sqrt{\lambda_2\lambda_3} - \eta_{1,2,3}, \\ \eta_1 &= \lambda_1 - \rho^X_{1,2}\sqrt{\lambda_1\lambda_2} - \rho^X_{1,3}\sqrt{\lambda_1\lambda_3} + \eta_{1,2,3}, \\ \eta_2 &= \lambda_2 - \rho^X_{1,2}\sqrt{\lambda_1\lambda_2} - \rho^X_{2,3}\sqrt{\lambda_2\lambda_3} + \eta_{1,2,3}, \\ \eta_3 &= \lambda_3 - \rho^X_{1,3}\sqrt{\lambda_1\lambda_3} - \rho^X_{2,3}\sqrt{\lambda_2\lambda_3} + \eta_{1,2,3}. \end{aligned} \tag{7.4}$$

To ensure all parameters are positive, it follows that

$$\begin{aligned}\eta_{1,2,3} \leq \min\{&\rho^X_{1,2}\sqrt{\lambda_1\lambda_2}, \rho^X_{1,3}\sqrt{\lambda_1\lambda_3}, \rho^X_{2,3}\sqrt{\lambda_2\lambda_3},\\ &-\lambda_1 + \rho^X_{1,2}\sqrt{\lambda_1\lambda_2} + \rho^X_{1,3}\sqrt{\lambda_1\lambda_3},\\ &-\lambda_2 + \rho^X_{1,2}\sqrt{\lambda_1\lambda_2} + \rho^X_{2,3}\sqrt{\lambda_2\lambda_3},\\ &-\lambda_3 + \rho^X_{1,3}\sqrt{\lambda_1\lambda_3} + \rho^X_{2,3}\sqrt{\lambda_2\lambda_3}\}.\end{aligned}$$

Under such a constraint, if

$$\begin{aligned}\min\{&\rho^X_{1,2}\sqrt{\lambda_1\lambda_2}, \rho^X_{1,3}\sqrt{\lambda_1\lambda_3}, \rho^X_{2,3}\sqrt{\lambda_2\lambda_3},\\ &-\lambda_1 + \rho^X_{1,2}\sqrt{\lambda_1\lambda_2} + \rho^X_{1,3}\sqrt{\lambda_1\lambda_3},\\ &-\lambda_2 + \rho^X_{1,2}\sqrt{\lambda_1\lambda_2} + \rho^X_{2,3}\sqrt{\lambda_2\lambda_3},\\ &-\lambda_3 + \rho^X_{1,3}\sqrt{\lambda_1\lambda_3} + \rho^X_{2,3}\sqrt{\lambda_2\lambda_3}\} > 0,\end{aligned} \tag{7.5}$$

there will be an infinite number of solutions to match the given marginals and correlations. Note that the inequality (7.5) is not difficult to satisfy, so there are many cases for which this can happen.

To show that these infinite joint distributions are indeed all different, we can prove that $p_{000} = \Pr(X_1 = 0, X_2 = 0, X_3 = 0)$ is different for these different solutions.

$$\begin{aligned}p_{000} &= \Pr(Y_1 = 0, Y_2 = 0, Y_3 = 0, Y_{1,2} = 0, Y_{1,3} = 0, Y_{2,3} = 0, Y_{1,2,3} = 0)\\ &= \Pr(Y_1 = 0)(Y_2 = 0)\Pr(Y_3 = 0)\Pr(Y_{1,2} = 0)\Pr(Y_{1,3} = 0)\Pr(Y_{2,3} = 0)\Pr(Y_{1,2,3} = 0)\\ &= \exp\left[-(\eta_1 + \eta_2 + \eta_3 + \eta_{1,2} + \eta_{1,3} + \eta_{2,3} + \eta_{1,2,3})\right].\end{aligned}$$

According to (7.4), it is straightforward to see that p_{000} is a function of $\eta_{1,2,3}$ given correlations and marginals. Thus, we conclude that in some cases defined by model (7.3), there are several possible joint distributions when marginal distributions and correlations are preset.

7.4.7. DERIVATION OF $\mathcal{W}$ IN (5.9)

For a d-dimensional Poisson random vector, there are a total of $D = (d^2 + 3d)/2$ parameters to estimate. The corresponding Fisher information matrix for $[\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\rho}]^\top$ can be decomposed in blocks as

$$\mathcal{I} := \begin{bmatrix} I_{1,1} & I_{1,2} & \cdots & I_{1,d} & \cdots & I_{1,D} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ I_{d,1} & I_{d,2} & \cdots & I_{d,d} & \cdots & I_{d,D} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ I_{D,1} & I_{D,2} & \cdots & I_{D,d} & \cdots & I_{D,D} \end{bmatrix} = \begin{bmatrix} I_{1,1} & I_{1,2} & \cdots & I_{1d} & \boldsymbol{I}_{1d^*} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ I_{d1} & I_{d2} & \cdots & I_{dd} & \boldsymbol{I}_{dd^*} \\ \boldsymbol{I}_{d^*1} & \boldsymbol{I}_{d^*2} & \cdots & \boldsymbol{I}_{d^*d} & \boldsymbol{I}_{d^*d^*} \end{bmatrix},$$

where $\boldsymbol{I}_{id^*}$ is a $(D-d)$-dimensional row vector for $i = 1\cdots, d$, $\boldsymbol{I}_{d^*j}$ is a $(D-d)$-dimensional column vector for $j = 1, \cdots, d$, $\mathbf{I}_{d^*d^*}$ is a $(D-d)\times(D-d)$ matrix and $d^* = d+1$.

Let s_i be the score function associated with the ith univariate marginal log-likelihood function, and define $J_{i,j} := \mathrm{Cov}(s_i, s_j)$ for $1 \le i,j \le d$. It follows that $J_{i,i}$ is the Fisher information from the marginal log-likelihood of X_i, $i = 1, \cdots, d$. Let $\boldsymbol{\alpha}_i := -\mathbf{I}^{-1}_{d^*d^*}\boldsymbol{I}_{d^*i}J^{-1}_{i,i}$ for $i = 1, \cdots, d$. Similar to $\mathcal{I}$, $\mathcal{W}$ can also be defined as a block matrix, which has (i,j)th element $J^{-1}_{i,i}J_{i,j}J^{-1}_{j,j}$ for $1 \le i,j \le d$; (i,d^*)th element $J^{-1}_{i,i}\sum_{j=1}^{d} J_{i,j}\boldsymbol{\alpha}_j^\top$ for $1 \le i \le d$; (d^*,d^*)th element $\mathbf{I}^{-1}_{d^*d^*} + \sum_{i=1}^{d}\sum_{j=1}^{d}\boldsymbol{\alpha}_i J_{i,j}\boldsymbol{\alpha}_j^\top$. Take two Poisson variables ($d=2$, $D=5$ and $d^*=3$) as an example, we have elements of $\mathcal{W}$ as

$$W_{1,1} = J^{-1}_{1,1} = \lambda_1, W_{2,2} = J^{-1}_{2,2} = \lambda_2,$$

$$W_{1,2} = J^{-1}_{1,1}J_{1,2}J^{-1}_{2,2} = \lambda_1\frac{\mathrm{Cov}(-1+\frac{X_1}{\lambda_1}, -1+\frac{X_2}{\lambda_2})}{\lambda_1\lambda_2}\lambda_2 = \mathrm{Cov}(X_1, X_2),$$

$$\boldsymbol{W}_{1,3} = J^{-1}_{1,1}\sum_{j=1}^{2} J_{1j}\boldsymbol{\alpha}_j^\top = \lambda_1\left[\frac{1}{\lambda_1}(-\mathbf{I}^{-1}_{3,3}\boldsymbol{I}_{3,1})^\top\lambda_1 + \frac{\mathrm{Cov}(X_1,X_2)}{\lambda_1\lambda_2}(-\mathbf{I}^{-1}_{3,3}\boldsymbol{I}_{3,2})^\top\lambda_2\right]$$
$$= -[\lambda_1\boldsymbol{I}_{1,3} + \mathrm{Cov}(X_1,X_2)\boldsymbol{I}_{2,3}]\mathbf{I}^{-1}_{3,3},$$

$$\boldsymbol{W}_{2,3} = J^{-1}_{2,2}\sum_{j=1}^{2} J_{2,j}\boldsymbol{\alpha}_j^\top = \lambda_2\left[\frac{\mathrm{Cov}(X_1,X_2)}{\lambda_1\lambda_2}(-\mathbf{I}^{-1}_{3,3}\boldsymbol{I}_{3,1})^\top\lambda_1 + \frac{1}{\lambda_2}(-\mathbf{I}^{-1}_{3,3}\boldsymbol{I}_{3,2})^\top\lambda_2\right]$$
$$= -[\mathrm{Cov}(X_1,X_2)\boldsymbol{I}_{1,3} + \lambda_1\boldsymbol{I}_{2,3}]\mathbf{I}^{-1}_{3,3},$$

$$\mathbf{W}_{3,3} = \mathbf{I}^{-1}_{3,3} + \sum_{i=1}^{2}\sum_{j=1}^{2}\boldsymbol{\alpha}_i J_{i,j}\boldsymbol{\alpha}_j^\top = \mathbf{I}^{-1}_{3,3} + \mathbf{I}^{-1}_{3,3}[\lambda_1\boldsymbol{I}_{3,1}\boldsymbol{I}_{1,3} + 2\mathrm{Cov}(X_1,X_2)\boldsymbol{I}_{3,1}\boldsymbol{I}_{2,3} + \lambda_2\boldsymbol{I}_{3,2}\boldsymbol{I}_{2,3}]\mathbf{I}^{-1}_{3,3},$$

with two row vectors $\boldsymbol{W}_{1,3}$ and $\boldsymbol{W}_{2,3}$, and a 3×3 matrix $\mathbf{W}_{3,3}$. As a symmetric matrix, $\mathcal{W}$ can be readily obtained.

7.4.8. PROOF OF PROPOSITION 8

Our result is a consequence of Theorem 2 in [175]. We only need to verify that their assumptions are satisfied. Condition (5) in [175] is satisfied since $\Theta \subset \mathbb{R}^{2d+\frac{d(d-1)}{2}}$. For this, we use the canonical embedding $\mathscr{S}_d^+ \subset \mathbb{R}^{\frac{d(d-1)}{2}}$ which identifies a d-dimensional correlation matrix with the (flattened) vector containing the $\frac{d(d-1)}{2}$ correlations. Condition (6) is satisfied since $\mathscr{S}_d^+$ is compact, l is continuous and $l(\boldsymbol{\theta})$ tends to $-\infty$ when $\min(\boldsymbol{\alpha},\boldsymbol{\beta}) \to +\infty$. Condition (7) is straightforwardly satisfied since l is differentiable on Θ as a finite sum of differential functions. Condition (10) is also satisfied since Q is continuous as a finite sum of continuous functions. This means that Theorem 2 of [175] can be applied, which completes the proof.

7.4.9. SIMULATION RESULTS

This section presents additional simulation results. For two positively correlated Poisson counts, we generate 100 samples from different models with varying marginal means, as illustrated in Table S7. Each model is tested across four distinct scenarios. Table S7 displays the average KL divergences between the predicted and actual PMFs over 100 replications. Next, three correlated Poisson counts are generated based on the eight models listed in Table S8. TR, Clayton, Frank, Gumbel, and Joe each define three positive correlations, while TC-4 specifies one positive correlation. TREx outlines two positive correlations, and GC specifies three negative correlations. Marginal means are set at 1 and 2. We consider various scenarios for different dependence structures. Each simulation is executed 100 times, producing 100 samples for each scenario. The results are summarised in Table S8. These tables emphasise our model's ability to capture a range of distributions and dependence structures.

We proceed by employing two-dimensional Poisson data from our model to demonstrate the effectiveness of the EM estimation method. The marginal means are both set to 1. We generate 100 samples with correlation values $\rho_{1,2} = 0.1, 0.2, \cdots, 0.9$, $\alpha_j = 0.15, 0.3, \cdots, 0.9$ and β_j is $1 - \alpha_j$. This entire simulation procedure is carried out 100 times. The Euclidean norm of biases between the estimated parameters $\check{\boldsymbol{\theta}}$ and the true parameters $\boldsymbol{\theta}$ is computed from 100 replications. Corresponding box plots across different αs and βs with different $\rho_{1,2}$ values are presented in Figure S13. The results indicate that the estimates are dependable, with relatively low biases.

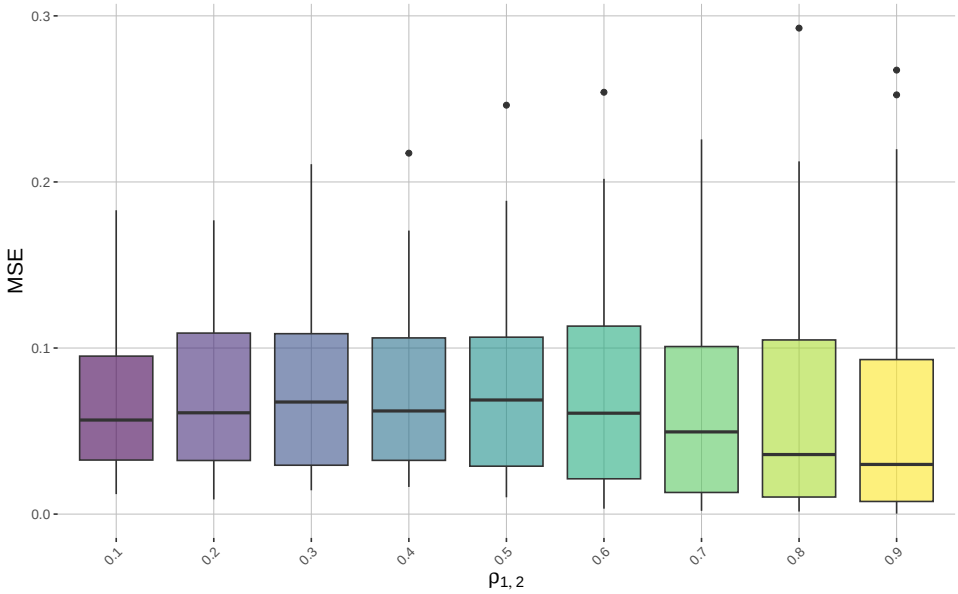
Table S7: Average KL distances based on 100 replications for two positively correlated Poisson counts with various methods.

Data source	Marginals	Parameter	TR	TREx	GC	TC-4	Proposed
TR	$(1,2)$	$\beta = 0.8$	0.006	0.014	0.008	0.012	0.009
	$(1,2)$	$\beta = 0.9$	0.006	0.024	0.012	0.018	0.008
	$(2,1)$	$\beta = 0.8$	0.006	0.014	0.008	0.013	0.009
	$(2,1)$	$\beta = 0.9$	0.006	0.024	0.012	0.018	0.009
TREx	$(1,2)$	$\theta = 0.6$	0.008	0.004	0.006	0.009	0.007
	$(1,2)$	$\theta = 0.8$	0.033	0.006	0.012	0.012	0.008
	$(2,1)$	$\theta = 0.6$	0.009	0.005	0.007	0.010	0.008
	$(2,1)$	$\theta = 0.8$	0.032	0.005	0.011	0.011	0.008
GC	$(1,1)$	$\rho_{1,2} = 0.4$	0.005	0.005	0.005	0.008	0.005
	$(1,1)$	$\rho_{1,2} = 0.6$	0.007	0.007	0.006	0.010	0.007
	$(2,1)$	$\rho_{1,2} = 0.8$	0.021	0.015	0.006	0.012	0.007
	$(2,1)$	$\rho_{1,2} = 0.9$	0.068	0.021	0.006	0.012	0.007
TC-4	$(1,2)$	$\rho_{1,2} = 0.8$	0.031	0.020	0.022	0.015	0.020
	$(1,2)$	$\rho_{1,2} = 0.9$	0.066	0.021	0.022	0.015	0.020
	$(2,1)$	$\rho_{1,2} = 0.8$	0.029	0.017	0.020	0.012	0.017
	$(2,1)$	$\rho_{1,2} = 0.9$	0.065	0.020	0.020	0.013	0.017
Clayton	$(2,1)$	$\theta = 1.5$	0.021	0.031	0.020	0.024	0.021
	$(2,1)$	$\theta = 2.5$	0.031	0.045	0.027	0.032	0.027
	$(2,1)$	$\theta = 3.5$	0.045	0.054	0.034	0.038	0.033
	$(2,1)$	$\theta = 4.5$	0.064	0.061	0.039	0.045	0.039
Frank	$(2,1)$	$\theta = 3$	0.008	0.009	0.008	0.011	0.009
	$(2,1)$	$\theta = 5$	0.013	0.012	0.009	0.013	0.010
	$(2,1)$	$\theta = 7$	0.025	0.016	0.012	0.016	0.013
	$(2,1)$	$\theta = 9$	0.041	0.021	0.016	0.020	0.017
Gumble	$(2,1)$	$\theta = 2$	0.020	0.009	0.012	0.009	0.011
	$(2,1)$	$\theta = 3$	0.069	0.016	0.013	0.011	0.014
	$(2,1)$	$\theta = 4$	0.132	0.020	0.011	0.010	0.011
	$(2,1)$	$\theta = 5$	0.191	0.021	0.011	0.009	0.012
Amh	$(2,1)$	$\theta = 0.28$	0.005	0.005	0.006	0.009	0.005
	$(2,1)$	$\theta = 0.46$	0.006	0.006	0.006	0.009	0.006
	$(2,1)$	$\theta = 0.64$	0.007	0.008	0.006	0.010	0.006
	$(2,1)$	$\theta = 0.91$	0.012	0.015	0.011	0.014	0.009
Joe	$(2,1)$	$\theta = 2$	0.023	0.015	0.023	0.019	0.015
	$(2,1)$	$\theta = 3$	0.059	0.026	0.035	0.029	0.027
	$(2,1)$	$\theta = 4$	0.099	0.030	0.034	0.027	0.028
	$(2,1)$	$\theta = 5$	0.146	0.034	0.036	0.028	0.030

Table S8: Average KL distances based on 100 replications for three correlated Poisson counts with various methods.

	Parameters	TR	TREx	GC	TC-4	Proposed
TR	$\beta = 0.4$	0.016	0.027	0.028	0.042	0.018
	$\beta = 0.6$	0.029	0.053	0.051	0.065	0.054
	$\beta = 0.8$	0.042	0.100	0.088	0.104	0.086
TREx	$\theta = 0.3$	0.391	0.022	0.056	0.121	0.034
	$\theta = 0.4$	0.425	0.013	0.045	0.105	0.026
	$\theta = 0.6$	2.289	0.011	0.325	0.434	0.133
GC	$\boldsymbol{\rho} = [-0.2, -0.3, -0.4]^\top$	0.381	0.112	0.014	0.027	0.016
	$\boldsymbol{\rho} = [-0.2, -0.3, -0.5]^\top$	0.475	0.172	0.012	0.027	0.014
	$\boldsymbol{\rho} = [-0.2, -0.4, -0.5]^\top$	0.598	0.246	0.016	0.030	0.016
TC-4	$\boldsymbol{\rho} = [0.3, 0.3, -0.5]^\top$	0.131	0.394	0.048	0.029	0.040
	$\boldsymbol{\rho} = [0.3, 0.6, -0.5]^\top$	0.214	0.700	0.065	0.041	0.049
	$\boldsymbol{\rho} = [0.4, 0.5, -0.5]^\top$	0.187	0.646	0.065	0.041	0.056
Clayton	$\theta = 0.25$	0.067	0.016	0.016	0.029	0.015
	$\theta = 0.5$	0.047	0.033	0.029	0.051	0.031
	$\theta = 1.75$	0.072	0.113	0.099	0.133	0.070
Frank	$\theta = 0.5$	0.101	0.011	0.014	0.032	0.009
	$\theta = 2$	0.047	0.012	0.019	0.036	0.010
	$\theta = 5$	0.028	0.023	0.038	0.079	0.019
Gumbel	$\theta = 1.2$	0.054	0.012	0.021	0.036	0.016
	$\theta = 3.6$	0.509	0.060	0.092	0.132	0.043
	$\theta = 3.8$	0.696	0.055	0.089	0.122	0.042
Joe	$\theta = 1.2$	0.084	0.012	0.021	0.032	0.009
	$\theta = 2.2$	0.059	0.059	0.113	0.146	0.046
	$\theta = 2.6$	0.103	0.068	0.139	0.166	0.069

Figure S13: Box plots of biases across $\alpha_j \in \{0.15, 0.3, \ldots, 0.9\}$ and $\beta_j = 1 - \alpha_j$, $j = 1, 2$, for different values of $\rho_{1,2}$. Poisson data are generated from our model.



BIBLIOGRAPHY

[1] P. Fleming, P. M. Gebraad, S. Lee, J.-W. van Wingerden, K. Johnson, M. Churchfield, J. Michalakes, P. Spalart, and P. Moriarty. “Simulation comparison of wake mitigation control strategies for a two-turbine case”. In: *Wind Energy* 18.12 (2015), pp. 2135–2143.

[2] X. Tian and D. Voskov. “Efficient application of stochastic Discrete Well Affinity (DiWA) proxy model with adjoint gradients for production forecast”. In: *Journal of Petroleum Science and Engineering* 210 (2022), p. 109911.

[3] T. J. Santner, B. J. Williams, W. I. Notz, and B. J. Williams. *The Design and Analysis of Computer Experiments*. Springer, 2003.

[4] D. R. Jones, M. Schonlau, and W. J. Welch. “Efficient global optimization of expensive black-box functions”. In: *Journal of Global optimization* 13 (1998), pp. 455–492.

[5] E. Vazquez and J. Bect. “Convergence properties of the expected improvement algorithm with fixed mean and covariance functions”. In: *Journal of Statistical Planning and inference* 140.11 (2010), pp. 3088–3095.

[6] A. D. Bull. “Convergence rates of efficient global optimization algorithms.” In: *Journal of Machine Learning Research* 12.10 (2011), pp. 2879–2904.

[7] Z. Chen, S. Mak, and C. J. Wu. “A hierarchical expected improvement method for Bayesian optimization”. In: *Journal of the American Statistical Association* (2023), pp. 1–14.

[8] M. Schonlau, W. J. Welch, and D. R. Jones. “Global versus local search in constrained optimization of computer models”. In: *Lecture Notes-monograph Series* (1998), pp. 11–25.

[9] A. Sóbester, S. J. Leary, and A. J. Keane. “On the design of optimization strategies based on global response surface approximation models”. In: *Journal of Global Optimization* 33 (2005), pp. 31–59.

[10] E. Contal, D. Buffoni, A. Robicquet, and N. Vayatis. “Parallel Gaussian process optimization with upper confidence bound and pure exploration”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2013, pp. 225–240.

[11] T. Desautels, A. Krause, and J. W. Burdick. “Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization.” In: *Journal of Machine Learning Research* (2014), pp. 3873–3923.

[12] T. Kathuria, A. Deshpande, and P. Kohli. “Batched gaussian process bandit optimization via determinantal point processes”. In: *Advances in Neural Information Processing Systems* 29 (2016), pp. 4213–4221.

[13] K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Póczos. “Parallelised Bayesian optimisation via Thompson sampling”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 133–142.

[14] A. Shah and Z. Ghahramani. “Parallel Predictive Entropy Search for Batch Global Optimization of Expensive Objective Functions”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc., 2015.

[15] M. Schonlau. “Computer experiments and global optimization”. PhD thesis. 1997.

[16] D. Ginsbourger, R. Le Riche, and L. Carraro. *A Multi-points Criterion for Deterministic Parallel Global Optimization Based on Gaussian Processes*. Tech. rep. 2008.

[17] J. Azimi, A. Fern, and X. Fern. “Batch Bayesian Optimization via Simulation Matching”. In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. Vol. 23. Curran Associates, Inc., 2010.

[18] F. A. Viana, R. T. Haftka, and L. T. Watson. “Efficient global optimization algorithm assisted by multiple surrogate techniques”. In: *Journal of Global Optimization* 56 (2013), pp. 669–689.

[19] Z. Li, S. Ruan, J. Gu, X. Wang, and C. Shen. “Investigation on parallel algorithms in efficient global optimization based on multiple points infill criterion and domain decomposition”. In: *Structural and Multidisciplinary Optimization* 54 (2016), pp. 747–773.

[20] G. Briffoteaux, M. Gobert, R. Ragonnet, J. Gmys, M. Mezmaz, N. Melab, and D. Tuyttens. “Parallel surrogate-assisted optimization: batched Bayesian neural network-assisted GA versus q-EGO”. In: *Swarm and Evolutionary Computation* 57 (2020), p. 100717.

[21] J. Liu, C. Jiang, and J. Zheng. “Batch Bayesian optimization via adaptive local search”. In: *Applied Intelligence* 51.3 (2021), pp. 1280–1295.

[22] H. Wackernagel. *Multivariate Geostatistics: An Introduction with Applications*. Springer Science & Business Media, 2003.

[23] V. R. Joseph, T. Dasgupta, R. Tuo, and C. J. Wu. “Sequential exploration of complex surfaces using minimum energy designs”. In: *Technometrics* 57.1 (2015), pp. 64–74.

[24] V. R. Joseph, D. Wang, L. Gu, S. Lyu, and R. Tuo. “Deterministic Sampling of Expensive Posteriors Using Minimum Energy Designs”. In: *Technometrics* 61.3 (2019), pp. 297–308.

[25] D. Eriksson, M. Pearce, J. Gardner, R. D. Turner, and M. Poloczek. “Scalable Global Optimization via Local Bayesian Optimization”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[26] R. C. Team *et al.* “R: A language and environment for statistical computing”. In: (2013).

[27] S. Surjanovic and D. Bingham. *Virtual Library of Simulation Experiments: Test Functions and Datasets*. Retrieved August 29, 2025, from http://www.sfu.ca/~ssurjano.

[28] M. Steinbuch, W. De Boer, O. Bosgra, S. Peeters, and J. Ploeg. “Optimal control of wind power plants”. In: *Journal of Wind Engineering and Industrial Aerodynamics* 27.1-3 (1988), pp. 237–246.

[29] T. Knudsen, T. Bak, and M. Svenstrup. “Survey of wind farm control—power and fatigue optimization”. In: *Wind Energy* 18.8 (2015), pp. 1333–1351.

[30] L. Bliek, A. Guijt, R. Karlsson, S. Verwer, and M. De Weerdt. “Benchmarking surrogate-based optimisation algorithms on expensive black-box functions”. In: *Applied Soft Computing* 147 (2023), p. 110744.

[31] X. Liu, X. Liu, T. Kaman, X. Lu, and G. L. and. “Statistical Learning for Nonlinear Dynamical Systems with Applications to Aircraft-UAV Collisions”. In: *Technometrics* 65.4 (2023), pp. 564–578.

[32] S.-T. Yeh, X. Wang, C.-L. Sung, S. Mak, Y.-H. Chang, L. Zhang, C. J. Wu, and V. Yang. “Common proper orthogonal decomposition-based spatiotemporal emulator for design exploration”. In: *AIAA Journal* 56.6 (2018), pp. 2429–2442.

[33] M. D. Buhmann. “Radial basis functions”. In: *Acta Numerica* 9 (2000), pp. 1–38.

[34] S. Haykin. *Neural Networks: a Comprehensive Foundation*. Prentice Hall PTR, 1998.

[35] R.-B. Chen, W. Wang, and C. F. J. W. and. “Sequential Designs Based on Bayesian Uncertainty Quantification in Sparse Representation Surrogate Modeling”. In: *Technometrics* 59.2 (2017), pp. 139–152.

[36] R. B. Gramacy. *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Chapman and Hall/CRC, 2020.

[37] T. Dolski, E. T. Spiller, and S. E. M. and. “Gaussian Process Emulation for High-Dimensional Coupled Systems”. In: *Technometrics* 66.3 (2024), pp. 455–469.

[38] R. B. Gramacy and D. W. A. and. “Local Gaussian Process Approximation for Large Computer Experiments”. In: *Journal of Computational and Graphical Statistics* 24.2 (2015), pp. 561–578.

[39] R. Zhang, C. D. Lin, and P. R. and. “Local Gaussian Process Model for Large-Scale Dynamic Computer Experiments”. In: *Journal of Computational and Graphical Statistics* 27.4 (2018), pp. 798–807.

[40] J. Chen, L. Kang, and G. L. and. “Gaussian Process Assisted Active Learning of Physical Laws”. In: *Technometrics* 63.3 (2021), pp. 329–342.

[41] M. H. Tan. “Gaussian Process Modeling of a Functional Output with Information from Boundary and Initial Conditions and Analytical Approximations”. In: *Technometrics* 60.2 (2018), pp. 209–221.

[42] P. Benner, S. Gugercin, and K. Willcox. “A Survey of Projection-Based Model Reduction Methods for Parametric Dynamical Systems”. In: *SIAM Review* 57.4 (2015), pp. 483–531.

[43] D. Amsallem, J. Cortial, and C. Farhat. “Towards real-time computational-fluid-dynamics-based aeroelastic computations using a database of reduced-order information”. In: *AIAA Journal* 48.9 (2010), pp. 2029–2037.

[44] D. Amsallem and C. Farhat. “Stabilization of projection-based reduced-order models”. In: *International Journal for Numerical Methods in Engineering* 91.4 (2012), pp. 358–377.

[45] A. K. Saibaba. “Randomized discrete empirical interpolation method for nonlinear model reduction”. In: *SIAM Journal on Scientific Computing* 42.3 (2020), A1582–A1608.

[46] R. Zhang, S. Mak, and D. Dunson. “Gaussian Process Subspace Prediction for Model Reduction”. In: *SIAM Journal on Scientific Computing* 44.3 (2022), A1428–A1449.

[47] Z. Ma, S. Ahuja, and C. W. Rowley. “Reduced-order models for control of fluids using the eigensystem realization algorithm”. In: *Theoretical and Computational Fluid Dynamics* 25 (2011), pp. 233–247.

[48] P. J. Schmid. “Dynamic mode decomposition and its variants”. In: *Annual Review of Fluid Mechanics* 54.1 (2022), pp. 225–254.

[49] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. “A data–driven approximation of the koopman operator: Extending dynamic mode decomposition”. In: *Journal of Nonlinear Science* 25 (2015), pp. 1307–1346.

[50] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. “MPCA: multilinear principal component analysis of tensor objects”. In: *IEEE Transactions on Neural Networks* 19.1 (2008), pp. 18–39.

[51] T. G. Kolda and B. W. Bader. “Tensor decompositions and applications”. In: *SIAM Review* 51.3 (2009), pp. 455–500.

[52] L. Cheng, Z. Chen, and Y.-C. Wu. *Bayesian Tensor Decomposition for Signal Processing and Machine Learning: Modeling, Tuning-Free Algorithms, and Applications*. Springer Nature, 2023.

[53] M. Xiao, P. Breitkopf, R. Filomeno Coelho, C. Knopf-Lenoir, M. Sidorkiewicz, and P. Villon. “Model reduction by cpod and kriging: application to the shape optimization of an intake port”. In: *Structural and Multidisciplinary Optimization* 41 (2010), pp. 555–574.

[54] S. Mak, C.-L. Sung, X. Wang, S.-T. Yeh, Y.-H. Chang, V. R. Joseph, V. Yang, and C. J. Wu. “An efficient surrogate model for emulation and physics extraction of large eddy simulations”. In: *Journal of the American Statistical Association* 113.524 (2018), pp. 1443–1456.

[55] Y.-H. Chang, L. Zhang, X. Wang, S.-T. Yeh, S. Mak, C.-L. Sung, C. Jeff Wu, and V. Yang. “Kernel-smoothed proper orthogonal decomposition–based emulation for spatiotemporally evolving flow dynamics prediction”. In: *AIAA Journal* 57.12 (2019), pp. 5269–5280.

[56] Y.-H. Chang, X. Wang, L. Zhang, Y. Li, S. Mak, C.-F. J. Wu, and V. Yang. “Reduced-order modeling for complex flow emulation by common kernel-smoothed proper orthogonal decomposition”. In: *AIAA Journal* 59.9 (2021), pp. 3291–3303.

[57] M. Guo and J. S. Hesthaven. “Reduced order modeling for nonlinear structural analysis using Gaussian process regression”. In: *Computer Methods in Applied Mechanics and Engineering* 341 (2018), pp. 807–826.

[58] X. Sun, X. Pan, and J.-I. Choi. “Non-intrusive framework of reduced-order modeling based on proper orthogonal decomposition and polynomial chaos expansion”. In: *Journal of Computational and Applied Mathematics* 390 (2021), p. 113372.

[59] X.-F. He, L. Li, S. Lanteri, and K. Li. “Model order reduction for parameterized electromagnetic problems using matrix decomposition and deep neural networks”. In: *Journal of Computational and Applied Mathematics* 431 (2023), p. 115271.

[60] L. Cicci, S. Fresca, M. Guo, A. Manzoni, and P. Zunino. “Uncertainty quantification for nonlinear solid mechanics using reduced order models with Gaussian process regression”. In: *Computers & Mathematics with Applications* 149 (2023), pp. 1–23.

[61] M. N. Hossain, K. T. J. Gladwin, and A. B. Narendranath. “A Nonintrusive POD-PCE Scheme for Uncertainty Quantification of Stochastic EM Problems”. In: *IEEE Microwave and Wireless Technology Letters* 33.8 (2023), pp. 1103–1106.

[62] M. Yu and K. Li. “A data-driven reduced-order modeling approach for parameterized time-domain Maxwell’s equations”. In: *Networks and Heterogeneous Media* 19.3 (2024), pp. 1309–1335.

[63] R. Fu, D. Xiao, I. M. Navon, F. Fang, L. Yang, C. Wang, and S. Cheng. “A non-linear non-intrusive reduced order model of fluid flow by auto-encoder and self-attention deep learning methods”. In: *International Journal for Numerical Methods in Engineering* 124.13 (2023), pp. 3087–3111.

[64] W. Koo, E.-Y. Ma, and H. K. and. “Deep Latent Factor Model for Spatio-Temporal Forecasting”. In: *Technometrics* 66.3 (2024), pp. 470–482.

[65] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[66] L. Yang, X. Meng, and G. E. Karniadakis. “B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data”. In: *Journal of Computational Physics* 425 (2021), p. 109913.

[67] E. Spiller, M. Bayarri, J. Berger, E. Calder, A. Patra, E. Pitman, and R. Wolpert. “Automating Emulator Construction for Geophysical Hazard Maps”. In: *SIAM/ASA Journal on Uncertainty Quantification* 2.1 (2014), pp. 126–152.

[68] S. Conti and A. O’Hagan. “Bayesian emulation of complex multi-output and dynamic computer models”. In: *Journal of Statistical Planning and Inference* 140.3 (2010), pp. 640–651.

[69] M. A. Alvarez and N. D. Lawrence. “Computationally efficient convolved multiple output Gaussian processes”. In: *The Journal of Machine Learning Research* 12 (2011), pp. 1459–1500.

[70] J. Svenson and T. Santner. “Multiobjective optimization of expensive-to-evaluate deterministic computer simulator models”. In: *Computational Statistics & Data Analysis* 94 (2016), pp. 250–264.

[71] T. E. Fricker, J. E. Oakley, and N. M. U. and. “Multivariate Gaussian Process Emulators With Nonseparable Covariance Structures”. In: *Technometrics* 55.1 (2013), pp. 47–56.

[72] T. G. Kolda. *Multilinear Operators for Higher-Order Decompositions.* Tech. rep. Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA (United States), 2006.

[73] C. Eckart and G. Young. “The approximation of one matrix by another of lower rank”. In: *Psychometrika* 1.3 (1936), pp. 211–218.

[74] C.-M. Chang, H.-Y. Tsai, T.-Y. Chiu, C.-J. Cheng, W.-A. Lin, R.-B. Chen, and Y.-J. Chou. “Proper-Orthogonal-Decomposition–Based Emulation of Spatiotemporal Evolution of Turbulent Wakes”. In: *AIAA Journal* (2025), pp. 1–13.

[75] K. Li, Y. Li, L. Li, and S. Lanteri. “Surrogate modeling of time-domain electromagnetic wave propagation via dynamic mode decomposition and radial basis function”. In: *Journal of Computational Physics* 491 (2023), p. 112354.

[76] K. Li, T.-Z. Huang, L. Li, Y. Zhao, and S. Lanteri. “A non-intrusive model order reduction approach for parameterized time-domain Maxwell’s equations”. In: *Discrete and Continuous Dynamical Systems-Series B* 28.1 (2022), pp. 449–473.

[77] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2023.

[78] K. Soetaert, J. Cash, and F. Mazzia. *Solving Differential Equations in R*. Springer Berlin Heidelberg, 2012.

[79] Y. Luo, Y. Chen, and Z. Zhang. “CFDBench: A large-scale benchmark for machine learning methods in fluid dynamics”. In: *arXiv preprint arXiv:2310.05963* (2024).

[80] Y. Wang, D. Voskov, M. Khait, and D. Bruhn. “An efficient numerical simulator for geothermal simulation: a benchmark study”. In: *Applied Energy* 264 (2020), p. 114693.

[81] C.-W. Wu, W. Pearn, and S. Kotz. “An overview of theory and practice on process capability indices for quality assurance”. In: *International Journal of Production Economics* 117.2 (2009), pp. 338–359.

[82] A. Otsuka and F. Nagata. “Quality design method using process capability index based on Monte-Carlo method and real-coded genetic algorithm”. In: *International Journal of Production Economics* 204 (2018), pp. 358–364.

[83] C.-H. Wang, K.-S. Chen, and K.-H. Tan. “Lean six sigma applied to process performance and improvement model for the development of electric scooter water-cooling green motor assembly”. In: *Production Planning & Control* 30.5–6 (2019), pp. 400–412.

[84] R. Tomohiro, I. Arizono, and Y. Takemoto. “Economic design of double sampling Cpm control chart for monitoring process capability”. In: *International Journal of Production Economics* 221 (2020), p. 107468.

[85] S. Kumar, A. S. Yadav, S. Dey, and M. Saha. “Parametric inference of generalized process capability index Cpyk for the power Lindley distribution”. In: *Quality Technology & Quantitative Management* 19.2 (2022), pp. 153–186.

[86] J. Shaabani and A. Jafari. “Inference on the lifetime performance index of gamma distribution: point and interval estimation”. In: *Communications in Statistics-Simulation and Computation* 53.3 (2024), pp. 1368–1386.

[87] S. Bittanti, M. Lovera, and L. Moiraghi. “Application of non-normal process capability indices to semiconductor quality control”. In: *IEEE Transactions on Semiconductor Manufacturing* 11.2 (1998), pp. 296–303.

[88] K. S. Chen and W. L. Pearn. “An application of non-normal process capability indices”. In: *Quality and Reliability Engineering International* 13.6 (1997), pp. 355–360.

[89] W. Pearn and K. Chen. “Capability indices for non-normal distributions with an application in electrolytic capacitor manufacturing”. In: *Microelectronics Reliability* 37.12 (1997), pp. 1853–1858.

[90] S. Khlil, H. Al-Khazraji, and Z. Alabacy. “Process capability analysis to assess capacity of a cleaning liquid product with asymmetric tolerances”. In: *Engineering and Technology Journal* 38.6A (2020), pp. 910–916.

[91] K. Vännman. “A unified approach to capability indices”. In: *Statistica Sinica* 5.2 (1995), pp. 805–820.

[92] V. E. Kane. “Process capability indices”. In: *Journal of Quality Technology* 18.1 (1986), pp. 41–52.

[93] R. A. Boyles. “Brocess capability with asymmetric tolerances”. In: *Communications in Statistics - Simulation and Computation* 23.3 (1994), pp. 615–635.

[94] K. Vännman. “A general class of capability indices in the case of asymmetric tolerances”. In: *Communications in Statistics - Theory and Methods* 26.8 (1997), pp. 2049–2072.

[95] K.-S. Chen and W.-L. Pearn. “Capability indices for processes with asymmetric tolerances”. In: *Journal of the Chinese Institute of Engineers* 24.5 (2001), pp. 559–568.

[96] D. Grau. “Algebraic relationship between symmetrical and asymmetrical capability indices”. In: *Quality Engineering* 17.3 (2005), pp. 359–370.

[97] C.-W. Wu, M.-H. Shu, W. L. Pearn, and Y.-C. Tai. “Estimating and testing process accuracy with extension to asymmetric tolerances”. In: *Quality & Quantity* 44.5 (2009), pp. 985–995.

[98] Z. Abbasi Ganji and B. Sadeghpour Gildeh. “A class of process capability indices for asymmetric tolerances”. In: *Quality Engineering* 28.4 (2016), pp. 441–454.

[99] J. A. Clements. “Process capability calculations for non-normal distributions”. In: *Quality Progress* 22 (1989), pp. 95–100.

[100] O. Senvar and C. Kahraman. “Type-2 fuzzy process capability indices for non-normal processes”. In: *Journal of Intelligent & Fuzzy Systems* 27.2 (2014), pp. 769–781.

[101] D. Dariae and B. Sadeghpour Gildeh. “Capability indices for Rayleigh process”. In: *Journal of System Management* 3.2 (2017), pp. 61–76.

[102] M. H. de Almeida, P. L. Ramos, G. S. Rao, and F. A. Moala. “Objective Bayesian inference for the capability index of the Gamma distribution”. In: *Quality and Reliability Engineering International* 37.5 (2021), pp. 2235–2247.

[103] S. Dey, M. Saha, M. Anis, S. S. Maiti, and S. Kumar. “Estimation and confidence intervals of CNp (u, v) for logistic-exponential distribution with application”. In: *International Journal of System Assurance Engineering and Management* 14.Suppl 1 (2023), pp. 431–446.

[104] B. Guo, Q. Xia, Y. Sun, and M. Aslam. “Generalized confidence intervals of quantile-based process capability indices for inverse Gaussian distribution”. In: *Quality Technology & Quantitative Management* 20.3 (2023), pp. 405–417.

[105] C.-H. Wang and K.-S. Chen. “New process yield index of asymmetric tolerances for bootstrap method and six sigma approach”. In: *International Journal of Production Economics* 219 (2020), pp. 216–223.

[106] S. Wang, J.-Y. Chiang, T.-R. Tsai, and Y. Qin. “Robust process capability indices and statistical inference based on model selection”. In: *Computers & Industrial Engineering* 156 (2021), p. 107265.

[107] M. Kashif, S. Ullah, M. Aslam, and M. Z. Iqbal. “Robust process capability indices Cpm and Cpmk using Weibull process”. In: *Scientific Reports* 13.1 (2023), p. 16977.

[108] L. A. Rosas Rivera, N. F. Hubele, and F. P. Lawrence. “Cpk index estimation using data transformation”. In: *Computers & Industrial Engineering* 29.1–4 (1995), pp. 55–58.

[109] R. M. West. “Best practice in statistics: The use of log transformation”. In: *Annals of Clinical Biochemistry* 59.3 (2022), pp. 162–165.

[110] L. C. Tang, S. E. Than, and B. W. Ang. “Computing process capability indices for nonnormal data: a review and comparative study”. In: *Six Sigma* (2006), pp. 107–130.

[111] M. Riani, A. C. Atkinson, and A. Corbellini. “Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression”. In: *Statistical Methods & Applications* 32.1 (2023), pp. 75–102.

[112] H. Wang, J. Yang, and S. Hao. “Two inverse normalizing transformation methods for the process capability analysis of non-normal process data”. In: *Computers & Industrial Engineering* 102 (2016), pp. 88–98.

[113] Z. Zou, Q. Wang, Q. Wu, M. Li, J. Zhen, D. Yuan, M. Zhou, C. Xu, Y. Wang, Y. Zhao, *et al.* “Inversion of heavy metal content in soil using hyperspectral characteristic bands-based machine learning method”. In: *Journal of Environmental Management* 355 (2024), p. 120503.

[114] P. Chen, B. X. Wang, and Z.-S. Ye. “Yield-based process capability indices for nonnormal continuous data”. In: *Journal of Quality Technology* 51.2 (2019), pp. 171–180.

[115] T. P. Ryan. *Statistical Methods for Quality Improvement*. John Wiley & Sons, 2011.

[116] A. Parchami and M. Mashinchi. “Fuzzy estimation for process capability indices”. In: *Information Sciences* 177.6 (2007), pp. 1452–1462.

[117] Y. Cui, J. Yang, and S. Huang. “Interval Estimation of Process Capability Indices Based on the Quality Data of Supplied Products”. In: *2018 12th International Conference on Reliability, Maintainability, and Safety (ICRMS)*. 2018, pp. 400–404.

[118] L. Ouyang, S. Dey, and C. Park. “Robust confidence intervals for the process capability index Cpk with bootstrap improvement”. In: *Quality Engineering* (2023), pp. 1–14.

[119] C. Cox. “Delta method”. In: *Wiley StatsRef: Statistics Reference Online* (2014).

[120] M.-Y. Cheng and L. Peng. “Regression modeling for nonparametric estimation of distribution and quantile functions”. In: *Statistica Sinica* (2002), pp. 1043–1060.

[121] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer London, 2001.

[122] L. Xue and J. Wang. “Distribution function estimation by constrained polynomial spline regression”. In: *Journal of Nonparametric Statistics* 22.4 (2010), pp. 443–457.

[123] L. C. Marsh and D. R. Cormier. *Spline Regression Models*. 137. Sage, 2001.

[124] J. O. Ramsay. “Monotone regression splines in action”. In: *Statistical Science* (1988), pp. 425–441.

[125] Y. M. Kantar. “Generalized least squares and weighted least squares estimation methods for distributional parameters”. In: *REVSTAT-Statistical Journal* 13.3 (2015), pp. 263–282.

[126] A. Perperoglou, W. Sauerbrei, M. Abrahamowicz, and M. Schmid. “A review of spline function procedures in R”. In: *BMC Medical Research Methodology* 19.1 (2019), p. 46.

[127] L. Schumaker. *Spline Functions: Basic Theory*. Cambridge University Press, 2007.

[128] A. W. Van der Vaart. *Asymptotic Statistics*. Vol. 3. Cambridge university press, 2000.

[129] W. Wang and J. Yan. *splines2: Regression Spline Functions and Classes*. R package version 0.4.8. 2023.

[130] A. Ghalanos and S. Theussl. *Rsolnp: General Non-Linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.16. 2015.

[131] T. Duong. *ks: Kernel Smoothing*. R package version 1.14.0. 2022.

[132] R. Liu, M. E. Kuhl, Y. Liu, and J. R. Wilson. “Modeling and simulation of nonstationary non-Poisson arrival processes”. In: *INFORMS Journal on Computing* 31.2 (2019), pp. 347–366.

[133] W.-L. Chen. “Computing the moments of polling models with batch Poisson arrivals by transform inversion”. In: *INFORMS Journal on Computing* 31.3 (2019), pp. 515–526.

[134] F. G. Pascual and S. B. Akhundjanov. “Copula-based control charts for monitoring multivariate Poisson processes with application to hepatitis C counts”. In: *Journal of Quality Technology* 52.2 (2020), pp. 128–144.

[135] K. Wang, T. Bhowmik, S. Zhao, N. Eluru, and E. Jackson. “Highway safety assessment and improvement through crash prediction by injury severity and vehicle damage using multivariate Poisson-lognormal model and joint negative binomial-generalized ordered probit fractional split model”. In: *Journal of Safety Research* 76 (2021), pp. 44–55.

[136] K. Bai, J. Li, and D. Ding. “Two approaches to monitoring multivariate Poisson counts: simple and accurate”. In: *IISE Transactions* 56.1 (2024), pp. 29–42.

[137] J. Zhang, Z.-C. Hu, and W. Sun. “On the measure concentration of infinitely divisible distributions”. In: *arXiv preprint arXiv:2310.03471* (2023).

[138] P. Holgate. “Estimation for the bivariate Poisson distribution”. In: *Biometrika* 51.1-2 (1964), pp. 241–287.

[139] D. Karlis. “An EM algorithm for multivariate Poisson distribution and related models”. In: *Journal of Applied Statistics* 30.1 (2003), pp. 63–77.

[140] D. Karlis and L. Meligkotsidou. “Multivariate Poisson regression with covariance structure”. In: *Statistics and Computing* 15 (2005), pp. 255–265.

[141] J. Muñoz-Pichardo, R. Pino-Mejías, J. García-Heras, F. Ruiz-Muñoz, and M. Luz González-Regalado. “A multivariate Poisson regression model for count data”. In: *Journal of Applied Statistics* 48.13-15 (2021), pp. 2525–2541.

[142] K. Shin and R. Pasupathy. “An algorithm for fast generation of bivariate Poisson random vectors”. In: *INFORMS Journal on Computing* 22.1 (2010), pp. 81–92.

[143] C. Genest, M. Mesfioui, and J. Schulz. “A new bivariate Poisson common shock model covering all possible degrees of dependence”. In: *Statistics & Probability Letters* 140 (2018), pp. 202–209.

[144] W. Whitt. “Bivariate distributions with given marginals”. In: *The Annals of Statistics* 4.6 (1976), pp. 1280–1289.

[145] J. Schulz, C. Genest, and M. Mesfioui. “A multivariate Poisson model based on comonotonic shocks”. In: *International Statistical Review* 89.2 (2021), pp. 323–348.

[146] O. A. Murphy and J. Schulz. “A new multivariate Poisson model”. In: *arXiv preprint arXiv:2407.05896* (2024).

[147] Z. Zhang, C. Charalambous, and P. Foster. “Joint modelling of longitudinal measurements and survival times via a multivariate copula approach”. In: *Journal of Applied Statistics* 50.13 (2023), pp. 2739–2759.

[148] P. Krupskii, F. Harrou, A. S. Hering, and Y. Sun. “Copula-based monitoring schemes for non-Gaussian multivariate processes”. In: *Journal of Quality Technology* 52.3 (2020), pp. 219–234.

[149] Z. Zhao, P. Shi, and X. Feng. “Knowledge learning of insurance risks using dependence models”. In: *INFORMS Journal on Computing* 33.3 (2021), pp. 1177–1196.

[150] A. Dhara, B. Das, and K. Natarajan. “Worst-case expected shortfall with univariate and bivariate marginals”. In: *INFORMS Journal on Computing* 33.1 (2021), pp. 370–389.

[151] C. Eckert and J. Hohberger. “Addressing endogeneity without instrumental variables: an evaluation of the Gaussian copula approach for management research”. In: *Journal of Management* 49.4 (2023), pp. 1460–1495.

[152] A. Sklar. “Random variables, joint distribution functions, and copulas”. In: *Kybernetika* 9.6 (1973), pp. 449–460.

[153] A. K. Nikoloulopoulos. “Copula-based models for multivariate discrete response data”. In: *Copulae in Mathematical and Quantitative Finance: Proceedings of the Workshop Held in Cracow, 10-11 July 2012*. Springer. 2013, pp. 231–249.

[154] R. B. Nelsen. *An introduction to copulas*. Springer, 2006.

[155] G. Tzougas and A. P. di Cerchiara. “Bivariate mixed Poisson regression models with varying dispersion”. In: *North American Actuarial Journal* 27.2 (2023), pp. 211–241.

[156] B. Cekyay, J. B. G. Frenk, and S. Javadi. “On computing the multivariate poisson probability distribution”. In: *Methodology and Computing in Applied Probability* 25.3 (2023), p. 70.

[157] D. I. Inouye, E. Yang, G. I. Allen, and P. Ravikumar. “A review of multivariate distributions for count data derived from the Poisson distribution”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 9.3 (2017), e1398.

[158] J. Schulz. *Multivariate Poisson Models Based on Comonotonic and Counter-monotonic Shocks*. McGill University (Canada), 2019.

[159] H. Joe. “Asymptotic efficiency of the two-stage estimation method for copula-based models”. In: *Journal of Multivariate Analysis* 94.2 (2005), pp. 401–419.

[160] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, 2019.

[161] P. L’Ecuyer, F. Puchhammer, and A. Ben Abdellah. “Monte Carlo and quasi-Monte Carlo density estimation via conditioning”. In: *INFORMS Journal on Computing* 34.3 (2022), pp. 1729–1748.

[162] J. J. Xu. “Statistical modelling and inference for multivariate and longitudinal discrete response data”. PhD thesis. University of British Columbia, 1996.

[163] H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. CRC press, 1997.

[164] R. Zimmerman, R. V. Craiu, and V. Leos-Barajas. “Copula modeling of serially correlated multivariate data with hidden structures”. In: *Journal of the American Statistical Association* (2023), pp. 1–12.

[165] T. F. Ma, Y. Cai, P. Shi, and J. Zhu. “Hierarchical dependence modeling for the analysis of large insurance claims data”. In: *The Annals of Applied Statistics* 18.2 (2024), pp. 1404–1420.

[166] J. Wang and M. R. Taaffe. “Multivariate mixtures of normal distributions: properties, random vector generation, fitting, and as models of market daily changes”. In: *INFORMS Journal on Computing* 27.2 (2015), pp. 193–203.

[167] Q. Zhai and Z.-S. Ye. “A multivariate stochastic degradation model for dependent performance characteristics”. In: *Technometrics* 65.3 (2023), pp. 315–327.

[168] Y. Qiu, Y. Chen, K. Fang, L. Yu, and K. Fang. “Fraud Detection by Integrating Multisource Heterogeneous Presence-Only Data”. In: *INFORMS Journal on Computing* 37.4 (2025), pp. 998–1017.

[169] A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn. *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-3. 2021.

[170] X. Shen, D. Wolfe, and S. Zhou. “Local asymptotics for regression splines and confidence regions”. In: *The Annals of Statistics* 26.5 (1998), pp. 1760–1782.

[171] C. De Boor. *A Practical Guide to Splines*. Springer New York, 1978.

[172] J. Z. Huang. “Local asymptotics for polynomial spline regression”. In: *The Annals of Statistics* 31.5 (2003), pp. 1600–1635.

[173] N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate Discrete Distributions*. Vol. 444. John Wiley & Sons, 2005.

[174] M. C. Cario and B. L. Nelson. “Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix”. In: *Industrial Engineering* (1997).

[175] C. J. Wu. “On the convergence properties of the EM algorithm”. In: *The Annals of statistics* (1983), pp. 95–103.

ACKNOWLEDGEMENTS

First and foremost, I extend my heartfelt gratitude to my promoters, Prof. dr. ir. G. Jongbloed and Prof. dr. Y.B. Tian.

Geurt, thank you for the invaluable opportunity to begin my PhD in Delft. I vividly remember our first online meeting, where you shared your passion for Chinese tea. Your enthusiasm made me feel welcomed and eager for my journey to the Netherlands. At TU Delft, your support and guidance have been instrumental in helping me adapt to Dutch life. When I faced challenges with the registration, your efforts to assist me were deeply appreciated. Your hearty laughter seems to have a magical quality that lifts my spirits and dispels any frustration or discouragement. Your faith in me has been a crucial factor in reaching this milestone.

Yubin, all of my accomplishments today are attributable to your steadfast support. You opened the gateway to scientific research and afforded me opportunities to collaborate with other esteemed scholars. You are like a lighthouse, filling me with confidence and guiding me consistently forward. When I got stuck, you selflessly dedicated your time to provide invaluable guidance, which means a lot to me. I will forever cherish and deeply value your unwavering support and concern.

I would also like to thank my daily mentors, Dr. P. Chen and Dr. D.P. Wang, for their consistent support. Piao, thank you for taking such exceptional care of me while at Delft. You have set an exemplary paragon that motivates me toward continuous advancement and growth. Whenever I faced research-related challenges, your insightful advice was indispensable. Your high standards have encouraged me to strive for excellence and push beyond my limits. Dianpeng, it is truly an honor to be your master's student. Whenever I felt lost, you were always there offering essential support and guidance without hesitation. Pursuing a PhD might not have seemed feasible without your encouragement and trust. Thank you for helping me in both my personal and academic life and for providing me with numerous opportunities to make me better.

I am also grateful to Dr. Alexis Derumigny; collaborating with you has been a tremendous learning experience. Many thanks to all members of my thesis committee for their thorough reading and constructive comments.

I am deeply appreciative of my colleagues and friends: Wenda, Shixuan, Yue, Yanyan, and Chenxi. Your companionship has made my PhD journey, both in Beijing and Delft, far less solitary. I extend my gratitude to Xiaoming, Yuhong, Xiao, Ardjen, Marc, Máté, Naqi, Thomas, Francesco,

Dan, Dominique, Jeffrey, Andreas, Shirong, Jacob, Joris, and everyone else with whom I have shared memorable moments. Meanwhile, I am grateful to Jeffery and Naqi for being my supporters during my PhD defence. I also appreciate the effort of Jeffery for polishing my Dutch Samenvatting. Special thanks to Xiwei, Joffrey, Roniet, and Dorothee for their invaluable support during my time at TU Delft.

Lastly, I am profoundly grateful to my parents and younger brother. I owe everything to their unwavering love and support throughout this journey.

CURRICULUM VITæ

Shixiang LI

28-07-1996	Born in Handan, Hebei, China.

EDUCATION

2002–2008	Experimental Primary School of Handan
2008–2011	Handan Hanguang Middle School
2011–2014	Hebei Jizhou High School
2014–2018	Bachelor of Science in Statistics Beijing Institute of Technology
2016–2018	Bachelor of Economics Beijing Institute of Technology (Dual Degree)
2018–2019	Master of Statistics Beijing Institute of Technology
2019–2025	PhD. Statistics Beijing Institute of Technology
2020–2025	PhD. Statistics Delft University of Technology *Thesis:* Novel Statistical Techniques for Complex Data Structures *Promotor:* Prof. dr. ir. G. Jongbloed Prof. dr. Y.B. Tian Dr.P. Chen

LIST OF PUBLICATIONS

1. **Shixiang Li**, Yubin Tian, and Dianpeng Wang. 2023. "Parallel efficient global optimization by using the minimum energy criterion". *Journal of Statistical Computation and Simulation*, 93(17):3104-3125.

2. **Shixiang Li**, Yubin Tian, and Dianpeng Wang. "A two-stage singular value decomposition-based kriging for spatiotemporally dynamic predictions". *Submitted.*

3. **Shixiang Li**, Sheng Fu, Dianpeng Wang, and Piao Chen. 2025. "A unified class of process capability indices for asymmetric tolerances". *Journal of Quality Technology*, 57(3):200-219.

4. **Shixiang Li**, Geurt Jongbloed, Alexis Derumigny, and Piao Chen. "A multivariate Poisson model with flexible dependence structure". *Submitted.*

5. Sheng Fu, **Shixiang Li**, Kai Yu, Zhisheng Ye, and Piao Chen. "Fast multinomial logistic regression with group sparsity". *Submitted.*

6. Wenda Kang, **Shixiang Li**, Yubin Tian, Ying Yin, Heliang Sui, and Dianpeng Wang. 2024. "A functional data-driven method for modelling degradation of waxy lubrication layer". *Quality and Reliability Engineering International* 40:3751–3773.