mathematics

MDPI

Article

Missing Data Imputation in Balanced Construction for Incomplete Block Designs

Haiyan Yu [1,†], Bing Han [2,*,†], Nicholas Rios [3,†] and Jianbin Chen [4,*,†]

[1] Center for Data and Decision Sciences, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; yuhy@cqupt.edu.cn
[2] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[3] Department of Statistics, George Mason University, Fairfax, VA 22031, USA; nrios4@gmu.edu
[4] School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081, China
* Correspondence: hanbing841208@163.com (B.H.); chenjianbin@bit.edu.cn (J.C.)
† These authors contributed equally to this work.

Abstract: Observational data with massive sample sizes are often distributed on many local machines. From an experimental design perspective, investigators often desire to identify the effect of new treatments (even ML algorithms) on many blocks of experimental data. With time requirements or budget constraints, assigning all treatments to each block is not always feasible. This creates incomplete responses with respect to a randomized complete block design (RCBD). These incomplete responses are missing by design. However, whether they can be estimated with missing imputation methods is not well understood. Thus, it is challenging to correctly identify the treatment effects with missing data. To this end, this paper provides a method for imputation and analysis of the responses with missing data. The proposed method consists of three steps: Reconstruction, Imputation, and 'Complete'-data Analysis (RICA). The incomplete responses are imputed with the expectation-maximization (EM) algorithm. The RCBD model is then fitted by the resulting dataset. The identifiability result suggests that the missing may be nonignorable for each block, but the whole data of an incomplete design are missing by design when the design is balanced. Theoretical results on relative efficiency also inform us when the missingness should be imputed for incomplete designs with the role of balanced variance. Applications on real-world data verify the efficacy of this method.

Keywords: distributed data; expectation-maximization; incomplete block design; missing by design; machine learning

MSC: 62K10; 62D10; 62D20

check for updates

Citation: Yu, H.; Han, B.; Rios, N.; Chen, J. Missing Data Imputation in Balanced Construction for Incomplete Block Designs. *Mathematics* **2024**, *12*, 3419. https://doi.org/10.3390/math12213419

Academic Editor: Florin Leon

Received: 20 September 2024
Revised: 29 October 2024
Accepted: 29 October 2024
Published: 31 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Observational data with massive sample sizes are often distributed on local machines. For example, Google AI proposed a new machine learning (ML) method called "Federated Learning" [1], in which the training datasets remain distributed over many machines. The distributed datasets can be employed for distributed estimation [2]. From an experimental design perspective, the investigator desires to identify the effect of many new methods (e.g., new computation algorithms or ghost data methods [3,4]) on those blocks of massive datasets. The treatments for such experiments are the available methods for certain tasks, e.g., identifying optimal orders [5–8], dimensional analysis [9], constrained optimization for stratified treatment rules [10,11], or causal inference [12]. The responses are the measurements of the task's accuracy. However, within a certain time requirement or budget [12,13], running all methods on every block is impossible. Thus, it is still challenging to correctly identify or compare all the methods' treatment effects while accounting for the blocks.

The balanced incomplete block design (BIBD) [14] is useful for such a problem. Some experimental treatments, such as computation algorithms, are absent from certain blocks in a BIBD. A BIBD is called "balanced" when each pair of treatments appears together in a block with the same frequency. This situation complicates the identification and estimation of treatment effects, as opposed to the randomized complete block design (RCBD), where all of the responses are observed [15,16]. Due to the randomization in the RCBD, the treatments' presence or assignment can be considered independent of their potential responses or expected values. If a treatment is absent from a particular block, it may not be independent of its expected value. The designer needs to randomize treatment assignment [17].

In general, there are drawbacks in analyzing the missing data of incomplete block designs (IBDs). For example, a BIBD is not orthogonal because the adjusted sum of squares of treatments (SST) and the adjusted sum of squares of blocks (SSB) are not orthogonal. In the conventional methods of analyzing BIBDs, the adjusted SST, the unadjusted SSB, and the corresponding sum of squares of error (SSE) are orthogonal. The adjusted SSB, the unadjusted SST, and the corresponding SSE are orthogonal. The orthogonalizing expectation maximization [18] is a design-based least squares algorithm that aims to orthogonalize a design matrix by adding new rows and then solve the original problem by embedding the augmented design in a missing data framework. Moreover, the imputation process may make the unbalanced data of IBD constructs "balanced", where the sample sizes are equal across all treatment combinations. The balanced constructs often reduce the estimation bias of the treatment effects or in ANOVA. How to reduce such bias is an essential issue of effect estimation in unbalanced designs.

From the perspective of missing data [19], the maximum likelihood estimates (MLEs) of treatment ratings provided a simple solution to the analysis of IBDs [20]. The modified MLE was presented as an organized solution to the problem of computing inter- and intra-block analyses of IBDs [21]. The residual (restricted) MLE [22] was applied to the analysis of the partial BIBDs. Regularized EM algorithms [23] and regularized MLEs [24] were implemented to handle the constrained parameter inference and high-dimensional problems. In software, SAS PROC MIXED [25] uses matrix notations to find the variance estimates and the combined treatment means. To estimate components of variance in unbalanced data [26], it also uses MLEs [27] and restricted MLEs. In Amelia2 [28], the expectation-maximum (EM) algorithm was integrated with a bootstrap approach for multiple missing data imputations. In R package lem3 [29], the restricted MLE model was developed to estimate item response models.

To the best of our knowledge, there are few algorithms for obtaining MLEs for a BIBD or general IBD from a missing data perspective. We have three main research questions:

(a) When is such an MLE identifiable and estimatable? (b) Will balance play any role in the estimation process? (c) How do the MLEs on an imputed dataset compare to the parameter estimates from the classical BIBDs? To solve these issues, this paper presents a new estimation method for IBDs. This algorithm contains three steps: reconstruction, imputation, and complete-case analysis (RICA). First, the IBD is treated as an RCBD with missing observations. Second, the missing observations are imputed via the MLE. Finally, the complete responses are analyzed with the statistical analysis model of the RCBD. The complete responses include the existing data from the IBD and the imputed responses from RICA.

Our contributions mainly lie in three aspects. First, the novel method RICA is proposed for the estimation of parameters in IBDs. It is the first to analyze the IBD from a new perspective on nonignorable missing data [30]. Second, theoretical supports are also provided to validate its identifiability and estimatability. This method provides a better way to understand the role of balance in IBDs. Third, this method provides new tools for comparing ML algorithms with a number of datasets (blocks). The results of those comparisons based on BIBDs can provide efficient estimation while maintaining small total runs of the algorithms. The effects of treatment with RICA are much easier to interpret for the ML algorithms than those of IBDs.

The rest of this paper is organized as follows. Section 2 introduces the BIBD and its corresponding model. Section 3 demonstrates the proposed method, which treats a BIBD as an RCBD with missing data. In Section 4, theoretical support is given for identifiability and estimatability. Section 5 presents applications of the RICA method to real-world datasets. Section 6 provides concluding remarks with more discussion.

2. Notations and Statistical Model of BIBDs

Before describing the RICA algorithm, we briefly overview balanced incomplete block (BIB) designs and their conventional analysis. The details can be found in [31].

2.1. Balanced Incomplete Block Designs

A randomized complete block design (RCBD) is one in which there are t treatments per block b, and the treatments are randomized within each of the blocks. In other words, each block contains a complete set of $k = t$ treatments. A BIBD is a set of t treatments and a collection of b blocks such that the following conditions are satisfied: (i) each block consists of exactly k treatments, $t > k > 0$; (ii) each treatment appears in exactly r blocks, $b > r > 0$; and (iii) each pair of treatments appears simultaneously in exactly λ blocks, for an integer $\lambda > 0$. BIBDs are often called (t, b, r, k, λ)-designs. Here, balance refers to the constancy of the parameter λ. When $\lambda = 1$, the design is called a simple design; otherwise, the design has repeated blocks. The word "incomplete" in BIBDs refers to the condition $t > k$, which implies that no block contains all the treatments.

The following necessary conditions demonstrate the constraints of parameters of a BIBD. Given a (t, b, r, k, λ)-design, it must hold that $bk = tr$ and $r(k-1) = \lambda(t-1)$. Therefore, the five parameters are not independent. BIBDs are usually written as (t, k, λ)-designs or (t, b, k)-designs, where b and r may be found using t, k, and λ. With Fisher's inequality [32], in a (t, k, λ)-design, there exists $t \leq b$.

Table 1 shows an example of a (4,4,3,3,2)-design. Here, four treatments are denoted by A, B, C, and D. There are also four blocks, each containing three treatments. Both the size, k, and the number of replicates, r, are 3. Each pair of treatments appears together in two blocks, so $\lambda = 2$. Since none of the blocks contain all the treatments ($k < t$), missingness exists in the responses for such blocks.

Table 1. Design table of BIBDs: (4,4,3,3,2)-design.

A. Design Table			
Block	**Treatments**		
$i = 1$	D	C	A
$i = 2$	A	D	B
$i = 3$	A	B	C
$i = 4$	C	B	D

B. Incidence Matrix: $(\mathbf{R}^*_{\mathbf{t}\times\mathbf{b}})$				
	Block (i)			
Treatment	**1**	**2**	**3**	**4**
A (j = 1)	1	1	1	0
B (j = 2)	0	1	1	1
C (j = 3)	1	0	1	1
D (j = 4)	1	1	0	1

Note: A, B, C, and D are symbols of treatments; 1 indicates the block is related with the treatment, otherwise 0.

2.2. Statistical Model for BIBDs

If both treatments and blocks are fixed, the parameters in the design model can be estimated by least squares. The linear statistical model is

$$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}, \tag{1}$$

where y_{ij} is the response, β_i and τ_j are the effects of block i and treatment j, and $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$; $i = 1, 2, \ldots, b$; and $j = 1, 2, \ldots, t$. The set of parameters of interest is $\theta = \{\mu, \beta, \tau\}$, as the variance ($\sigma^2$) is assumed to be fixed. The coefficients β and τ are subject to identifiability constraints $\sum_{i=1}^{b} \beta_i = 0$ and $\sum_{j=1}^{t} \tau_j = 0$.

Due to the incompleteness, the analysis of BIBDs is slightly different from that with an RCBD. When analyzing the data from an IBD, the marginal treatment means are not unbiased estimators of the estimatable effects $(\mu + \tau)$. For analysis of BIBDs, the conventional methods often utilize adjusted sums of squares of treatments to account for incompleteness [33].

2.3. Schema of the Proposed RICA Method

Under the missingness scheme, this diagram of the proposed RICA method presents a structured framework for addressing the challenges of missing data in block design analysis (Figure 1). The primary goal is to identify and analyze treatment effects within the RCBD framework. Applications to real-world data evaluate the performance of AUC or risk assessment using machine learning algorithms.

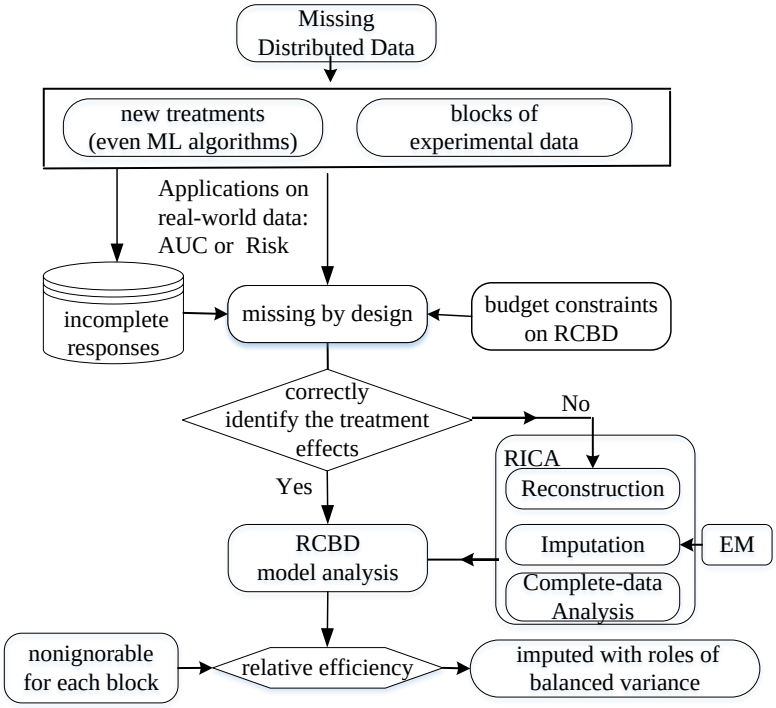


Figure 1. Schema of the proposed RICA method.

First, an RCBD model is used to account for the experimental design structure, including both blocks and treatments. However, specific research questions or practical constraints can lead to the omission of response data, resulting in missing data by design.

Second, the research question is whether the investigator can accurately identify the effects of the treatment. If so, the analysis of the RCBD model can proceed directly; otherwise, missing values are filled with estimates using the RICA method.

Third, in RICA, reconstruction explicitly models the missing data mechanism, treating it as a nonignorable missing data process. It reconstructs missing data based on available

information and assumptions about the data-generating process. Subsequently, the RICA method fills in the missing values using an imputation method, such as the EM algorithm, enabling a 'complete' data analysis with the RCBD model. RICA combines the strengths of reconstruction, to model the missing data mechanism, and imputation, to fill in the missing values.

Fourth, since there is uncertainty about whether missingness should be imputed for incomplete block designs, the relative efficiency of RICA is evaluated within the context of block designs. In practice, it is useful to determine when RICA analysis is preferable. To address this, the relative efficiency of RICA is derived, demonstrating the advantages of RICA through balanced variance.

The diagram highlights the importance of selecting an appropriate method for handling missing data in RCBD analysis. RICA provides a valuable framework by combining reconstruction and imputation while explicitly accounting for the missing data mechanism, resulting in more accurate and reliable outcomes. It suggests that RICA is well-suited for application to real-world data, especially in cases with incomplete or missing data within RCBDs.

3. BIBD as RCBD with Missing Data

This section provides a new framework for analyzing BIBDs from the perspective of missing data. The method of RICA is provided for imputing the incompleteness of the responses of BIBDs, then analyzing the data with the statistical model.

3.1. Incompleteness in BIBDs

Let $\mathbf{Y}^*$ be a vector of the true responses, and ϵ be a vector of imputation errors. Then, the imputed responses in the BIBDs are modeled by

$$\hat{y}_{ij} = y^*_{ij} + \epsilon_{ij}, \quad \mathbf{E}(\epsilon|\mathbf{Y}^*) = 0, \quad i = 1, 2, \ldots, b; \quad j = 1, 2, \ldots, t,$$

where $\hat{y}_{ij}$ is the imputed response for treatment t in block b, ϵ_{ij} is the corresponding errors, and $y^*_{ij} \in \mathbf{Y}^*$.

In BIBDs, the number of responses is bk or rt. Given the known responses of a (t, b, r, k, λ)-design (BIBDs), the fraction of incompleteness of the potential responses of (t, b)-RCBD is ρ, $\rho = 1 - \frac{k}{t} = 1 - \frac{r}{b}$.

Let $\mathbf{X}_C$ be a known $bt \times 2$ complete model matrix for the RCBD, and $\mathbf{X}$ is a $bk \times 2$ model matrix with BIBDs. Then it follows that

$$\mathbf{X}_{\mathrm{C}} = \left(\mathbf{X}^\top, \mathbf{X}^\top_{mis}\right)^\top,$$

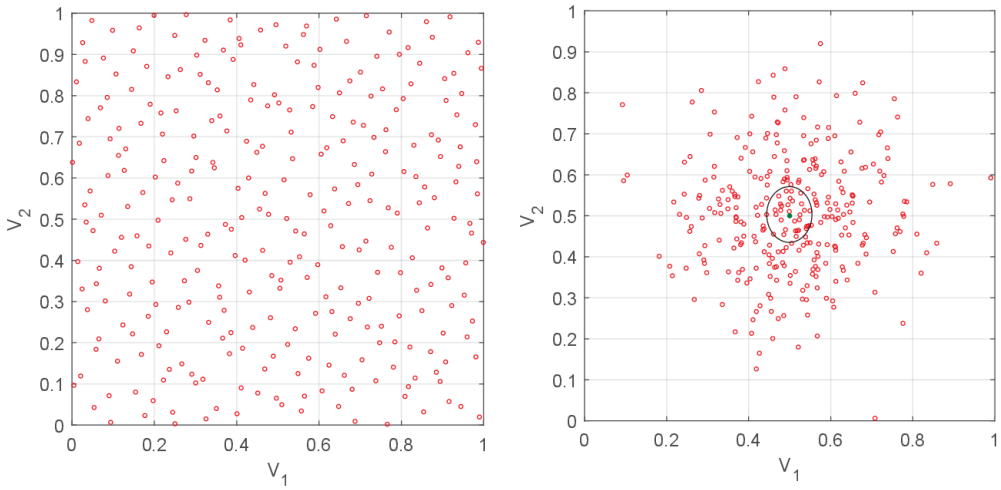
where $\mathbf{X}_{mis}$ is a $b(t-k) \times 2$ model matrix whose entries correspond to the block-treatment cells with missing responses. Let $\mathbf{Y}_{\mathrm{C}} = \left(\mathbf{Y}^\top, \mathbf{Y}^\top{}_{mis}\right)^\top$ be the vector of complete responses with missing data $\mathbf{Y}_{mis}$ of length $b(t-k)$. Let i^* and j^* be the cell indexes with missing responses.

3.2. Mechanisms of Missingness

Let $\mathbf{R}^*$ be the indicator matrix of the block designs' responses (see Table 1), and $\mathbf{R}^*_{ij}$ be the element of $\mathbf{R}^*$ at block i with treatment j. For simplicity, we define an indicator variable R, i.e., $R = 1$ indicates completeness, and $R = 0$ indicates a missing (incomplete) response. Thus, we have the sets $I(R = 1) = \{(i, j) \mid \mathbf{R}^*_{ij} = 1\}$ and $I(R = 0) = \{(i, j) \mid \mathbf{R}^*_{ij} = 0\}$. Let $\Pr(R = 1)$ be the probability of being observed due to the treatment assignment and $\Pr(R = 0)$ of being missing in block $i = 1, 2, ..., b$. For block i in a design, define the bias as $\Gamma_i = \mathbf{E}(\bar{y}_{io}|R = 1) - \mathbf{E}(\bar{y}_{io}|R = 0)$, $i = 1, \cdots, b$, where $\bar{y}_{io}$ is the mean of the responses in block i. Assume that the full data of block i^*, $(\mathbf{R}^*_{i^*j}, Y_{i^*j})$, are iid with $\mathbf{E}(Y_{i^*j}) = \mu_{i^*}$, $j = 1, 2, \ldots, t$.

Proposition 1 (Nonignorable missing). *If* $\Gamma_i \neq 0$, *then the missing data mechanism in block i is nonignorable, i.e.,* $\mathbf{E}(\bar{y}_{io}) \neq \mathbf{E}(\bar{y}_{io}|R = 1)$.

With the definition of nonignorable missingness [34], the conditional distribution of R at $R = 1$ (or $R = 0$) and $\bar{y}_{io}$ varies with the possible values of the missing data in each block. To illustrate this, Figure 2 shows examples of the possible values of the quasi-random data points in the context of random distributions and Gaussian (0.5, 0.02). Random dropouts in the random distribution can be illustrated as missing at random (MAR), while random dropouts in Gaussian (0.5, 0.02) can be represented as responses for missing not at random (MNAR). In the scatter of MAR data points (on the left), the conditional distributions in each cell are close to each other, while those in the scatter of MNAR data points (on the right) vary with possible values of the missing values.



(**a**) (**b**)

Figure 2. Illustrative examples of possible values with 300 points in two dimensions (V_1, V_2). (**a**) Quasi-random data points; (**b**) responses of a design sourcing form Gaussian (0.5, 0.02). The conditional distributions of the possible data values in the circle (**b**) are much denser than those outside the circle.

It will now be shown that the missingness mechanism within each block is nonignorable. It follows that

$$\begin{aligned}
&\mathbf{E}(\bar{y}_{io}) = \mathbf{E}(\bar{y}_{io}|R = 1) \cdot \Pr(R = 1) + \mathbf{E}(\bar{y}_{io}|R = \mathbf{0}) \cdot \Pr(R = 0) \\
&= \mathbf{E}(\bar{y}_{io}|R = 1) - [\mathbf{E}(\bar{y}_{io}|R = 1) - \mathbf{E}(\bar{y}_{io}|R = 0)] \cdot \Pr(R = 0) \\
&= \mathbf{E}(\bar{y}_{io}|R = 1) - \tfrac{k}{t}[\mathbf{E}(\bar{y}_{io}|R = 1) - \mathbf{E}(\bar{y}_{io}|R = 0)].
\end{aligned}$$

Then, the bias between the complete case model and the true model for block i is

$$\mathbf{E}(\bar{y}_{io}|R = 1) - \mathbf{E}(\bar{y}_{io}) = \tfrac{k}{t}[\mathbf{E}(\bar{y}_{io}|R = 1) - \mathbf{E}(\bar{y}_{io}|R = 0)] = \tfrac{k}{t}\Gamma_i.$$

Recall that the cell (i^*, j^*) with $R = 0$ is assigned with different treatments of the other cells with $R = 1$ in block i. When the treatments among those cells have the same effect on the responses, the above expression may be equal to zero. However, in general, $\Gamma_i \neq 0$. Thus, the bias between the complete case model and the true model is not always 0; therefore, $\mathbf{E}(\bar{y}_{io}|R = 1) \neq \mathbf{E}(\bar{y}_{io})$. As a result, every block's outcomes are nonignorable missing, namely, missing not at random (MNAR).

3.3. Expectation Maximization Algorithm for BIBD

To develop the EM algorithm, the likelihood function of the data under this missingness scheme must be derived. The key idea for doing this is summarized in the following result.

Proposition 2 (Independence)**.** *The incompleteness is assigned at the design stage in BIBDs, thus R is independent of Y given X, $R \perp\!\!\!\perp Y|X$, i.e.,*

$$f(\boldsymbol{Y}|R,\boldsymbol{X}) = f(\boldsymbol{Y} \mid \boldsymbol{X}).$$

Essentially, since the treatments are randomly assigned to blocks, i.e., "missing by design", then the probability of a particular response in a BIBD being missing does not depend on the value of the response itself. In general, the density function for an individual cell in a BIBD under this missingness scheme is

$$\begin{aligned} f(R,\mathbf{Y} \mid \mathbf{X}) &= f(\mathbf{Y} \mid R,\mathbf{X}) f(R \mid \mathbf{X}) \\ &= \prod_{i=1}^{b}\prod_{j=1}^{t} f(\mathbf{Y} \mid R=1,\mathbf{X})^{R} P(R=1)^{R} P(R=0)^{1-R} \\ &= \prod_{i=1}^{b}\prod_{j=1}^{t} \Big[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(y_{ij}-\mu-\beta_i-\tau_j\right)^2}{2\sigma^2}} P(R=1)\Big]^{R} P(R=0)^{1-R} \end{aligned}$$

where $R = 1$ if block j contains treatment i, and $R = 0$ otherwise. Since a BIBD is used, $P(R = 1)$ is constant across i and j.

Only the observed cases are required to find the MLE for θ. Thus, the likelihood function of the observed responses of the (b, t, k, r, λ)-BIBD is,

$$L(\theta) = \prod_{i=1}^{b}\prod_{j=1}^{t} f\left(\varepsilon_{ij}\right) \propto \prod_{i=1}^{b}\prod_{j=1}^{t} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(y_{ij}-\mu-\beta_i-\tau_j\right)^2}{2\sigma^2}}\right]^{R}, \tag{2}$$

where each block consists of exactly k treatments, and each treatment appears in exactly r blocks. Therefore, the MLE for θ in BIBDs by maximizing $\hat{\theta} := \arg\max_{\theta} L(\theta)$. The derivation of the MLE is the same as in the RCBD. More details are provided in Section 4.

To estimate and analyze the data from a BIBD, the proposed RICA algorithm is as follows.

The proposed Algorithm 1 estimates the incompleteness with respect to the RCBD. After imputation, the usual ANOVA can be performed using the imputed data (see Algorithm 2). The error degree of freedom [31] is $(b-1)(t-1) - \sum_i \sum_j (1 - \mathbf{R}^*_{ij})$, which implies that $MSE = \frac{SSE}{(b-1)(t-1)-\sum_i \sum_j (1-\mathbf{R}^*_{ij})}$. Although the EM and other missing estimation methods are useful imputation methods, model identifiability may not be guaranteed for BIBDs even if the missing mechanism is parametric. As a fundamental problem, valid statistical inference of these estimation methods relies on their identifiability. The following section provides more theoretical support for this method.

Algorithm 1: RICA Algorithm

Step 1: Reconstruction

Construct the matrix $\mathbf{X}$ of (t, b, r, k, λ)-BIBD with b blocks and t treatments

Obtain the responses $\mathbf{Y}$ of the BIBDs

Reconstruct the new 'complete' design matrix $\mathbf{X_C}$ of the RCBD from $\mathbf{X}$

Generate the matrix $\mathbf{Y}_C$ with $\mathbf{Y}$, with 'NA' values for missing responses.

Algorithm 1: *Cont.*

Step 2: Imputation

Initialize $\hat{\theta}^{(m)} = \left\{\mu^{(0)}, \beta_i^{(0)}, \tau_j^{(0)}\right\}$

Find the target cells (i^*, j^*) with missing values in $\mathbf{Y}_C$

Step 2a: Expectation

Imputing the missing values with their expected values generated by $\hat{\theta}^{(m)}$

Step 2b: Maximization

Obtain the MLEs $\hat{\theta}^{(m+1)}$

Update $\hat{\theta}^{(m)} = \hat{\theta}^{(m+1)}$;

Repeat Step 2a and Step 2b until convergence

Output a complete matrix of responses $\hat{\mathbf{Y}}_C$.

Step 3: 'Complete'-case Analysis

Conduct statistical analysis of the complete matrix $\hat{\mathbf{Y}}_C$ for the design $\mathbf{X}_C$.

Output estimation of treatment effects and block effects.

Algorithm 2: Pseudo code of RICA Algorithm

function RICA $((t, b, r, k, \lambda), \mathbf{Y}, \hat{\theta}^{(0)})$

Input : (t, b, r, k, λ)-design, $\mathbf{Y}$, $\hat{\theta}^{(0)} = \left\{\mu^{(0)}, \beta_i^{(0)}, \tau_j^{(0)}\right\}$, $\mu^{(0)} = \beta_i^{(0)} = \tau_j^{(0)} = 0.5$

Output: $\hat{\mathbf{Y}}_C, \hat{\theta}$.

Construct the matrix $\mathbf{X}$ of (t, b, r, k, λ)-BIBD with b blocks and t treatments.

Obtain the responses $\mathbf{Y}$ of the BIBD.

Generate the new 'complete' design matrix $\mathbf{X}_C$ of the RCBD from $\mathbf{X}$.

Generating the matrix $\mathbf{Y}_C$, including those from the responses of the BIBD, and the others as 'NA'.

Calculate the MLE estimation $\hat{\theta}$ with the responses of the BIBD.

while $(\theta^{(m+1)} - \hat{\theta}^{(m)} < \delta^* | m < 50)$ **do**

  $\hat{\theta}^{(m)} \rightarrow \hat{\theta}^{(m+1)}$ Find the target cells (i^*, j^*) with missing values in $\mathbf{Y}_C$, Impute the missing values with their expected values generated by $\hat{\theta}^{(m)}$, Obtain $\hat{\theta}^{(m+1)}$ with the MLE estimation of the 'complete' responses $\hat{\mathbf{Y}}_C$ using Equation (8).

end

return $\hat{\mathbf{Y}}_C$ as the imputed matrix, $\hat{\theta}$ as the estimated parameters.

4. Theoretical Supports

When is such an imputation method identifiable and estimatable? The following results show that, although identification is not available for each treatment or block under this missingness scheme, it is available for the full data distribution of BIBDs with an additional assumption. The sufficient parametric assumption is that the model in Equation (1) is correctly specified for (t, b, k, r, λ)-BIBD; that is, the parameters θ are identifiable. Classical results for estimatability and the derivation of MLEs are reviewed in Section 4.1. Moreover, the new results on the relative efficiency of RICA (to the classical BIBD methods) are derived in Section 4.3.

4.1. Identifiability

Identification is generally not available when the data are missing not at random (MNAR) without an additional assumption [35]. A possible approach is to make sufficient parametric assumptions (i.e., BIBDs) about the full data distribution for identification.

Recall that the responses are MNAR for each block or treatment. When the responses are collected according to the BIBD, the efficient parametric assumption about the full data distribution can be formulated in the response model (1).

Let $\mathbf{X}_B$ be an $n \times b$ incidence matrix whose (ω, i)th entry equals 1 if run ω is in block i, and equals zero otherwise. Similarly, let $\mathbf{X}_T$ be an $n \times t$ matrix whose (ω, j)th entry is equal to 1 if run ω contains treatment j, and 0 otherwise. The following result summarizes a traditional result about identifiability in this context.

Proposition 3 (Identifiability of the parameters with a construction)**.** *Given the linear regression model for a (b, t, k, r, λ)-BIBD, the parameters θ are identifiable if and only if the matrices $\mathbf{X_T}$ and $\mathbf{X_B}$ are full rank.*

When the matrices $\mathbf{X_T}$ and $\mathbf{X_B}$ are full rank, the parameters θ are identifiable in the linear regression model for a (b, t, k, r, λ)-BIBD.

Identifiability is a desirable property and it is theoretically possible to learn the true effects of this model after obtaining a sufficiently large number of observations. When the linear model (1) is correctly specified, the expectations of the missing data with BIBDs are also identifiable. Let $\hat{\mu}(y)$ be the estimator of $\mathbf{E}(y|R)$, and $\hat{\mu}(y|R=1)$ and $\hat{\mu}(y|R=0)$ be the estimators of $\mathbf{E}(y|R=1)$ and $\mathbf{E}(y|R=0)$. To estimate $\mathbf{E}(y|R=1)$, we have

$$\begin{aligned} \hat{\mu}(y|R=1,\theta) &= \tfrac{1}{bk} \sum_{i=1}^{b} \sum_{j=1}^{t} y_{ij}^{R} \\ &= \tfrac{1}{bk} \left(bk\mu + k \sum_{i=1}^{b} \beta_i + r \sum_{j=1}^{t} \tau_j \right) = \mu + \tfrac{1}{b} \sum_{i=1}^{b} \beta_i + \tfrac{1}{t} \sum_{j=1}^{t} \tau_j. \end{aligned} \tag{3}$$

Since the constraints are $\sum_{i=1}^{b} \beta_i = 0, \sum_{j=1}^{t} \tau_j = 0$, we have $\mathbf{E}(y|R=1) = \mu$.

To estimate $\mathbf{E}(y|R=0)$, we have

$$\begin{aligned} \hat{\mu}(y|R=0,\theta) &= \tfrac{1}{b(t-k)} \sum_{i=1}^{b} \sum_{j=1}^{t} y_{ij}^{1-R} \\ &= \tfrac{1}{b(t-k)} \left[b(t-k)\hat{\mu} + (t-k) \sum_{i=1}^{b} \hat{\beta}_i + (b-r) \sum_{j=1}^{t} \hat{\tau}_j \right] \\ &= \hat{\mu} + \tfrac{1}{b} \sum_{i=1}^{b} \hat{\beta}_i + \tfrac{1}{t} \sum_{j=1}^{t} \hat{\tau}_j. \end{aligned} \tag{4}$$

If the parameters are correctly estimated with $\mu = \hat{\mu}$, $\beta = \hat{\beta}$, $\tau = \hat{\tau}$, then we also have $\mathbf{E}(y|R=0) = \mu$.

For a design, let $[\mathbf{E}(y|R=1) - \mathbf{E}(y|R=0)] = \Pi$. For all the responses in the experiments,

$$\begin{aligned} \mathbf{E}(y) &= \mathbf{E}(y|R=1)Pr(R=1) + \mathbf{E}(y|R=0)Pr(R=0) \\ &= \mathbf{E}(y|R=1) - \mathbf{E}(y|R=1)Pr(R=0) + \mathbf{E}(y|R=0)Pr(R=0) \\ &= \mathbf{E}(y|R=1) - [\mathbf{E}(y|R=1) - \mathbf{E}(y|R=0)]Pr(R=0) \\ &= \mathbf{E}(y|R=1) - \Pi Pr(R=0). \end{aligned} \tag{5}$$

Thus, the bias between the complete case model $\mathbf{E}(y|R=1)$ and the true $\mathbf{E}(y)$ is

$$\begin{aligned} \mathbf{E}(y|R=1) - \mathbf{E}(y) &= [\mathbf{E}(y|R=1) - \mathbf{E}(y|R=0)]Pr(R=0) \\ &= \Pi Pr(R=0), \end{aligned} \tag{6}$$

which vanishes if either $\Pi = 0$ or if $Pr(R=0) = 0$, then there are no missing data.

In the presence of the BIBD responses, the second condition will definitely not hold. However, it was shown that for the BIBD, $\Pi = 0$, namely, $\mathbf{E}(y) = \mathbf{E}(y|R=1) = \mathbf{E}(y|R=0)$. This result suggests that the missingness may be nonignorable for each block, but the whole

data of an incomplete design are missing randomly when the design is balanced. Thus, the expectations of the missing data with BIBDs are identifiable.

4.2. Estimatability with the Parameters

In the previous section, we asked when such an MLE is "estimatable". This section reviews classical results that support the argument that the responses' missing values are estimatable in a design if and only if the design is connected. Furthermore, formulas for the MLEs of μ, β, τ are provided for readers who are unfamiliar. Here, we first introduce the concept of disconnectedness of a design [36].

Definition 1 (Disconnectedness)**.** *A design is considered disconnected if we can partition it into two separate groups that do not share any common treatments; that is, no block in one group contains treatments from the other group. Mathematically, let $G = (V, E)$ be an undirected graph with vertex set V and edge set E. For a given block design, let $V = (1, 2, \ldots, t)$, and define an edge (j, k) to be in E if and only if treatments j and k appear together in at least one block of the design. A block design is termed connected if there is at least one path between any two distinct vertices in G. Conversely, a block design that lacks this connectivity is referred to as disconnected.*

In incomplete block designs, there is a result in the estimation of contrasts. All contrasts in the treatment effects are estimatable in a design if and only if the design is connected.

The concept of connectivity provides a simple way of checking estimatability. Taking the (4,4,3,3,2)-design, for example, the incidence matrix of BIBDs (in Table 1B) can be used to construct a graph to check connectivity. With the observations of BIBDs, incompleteness is viewed as missing responses from the RCBD. Based on Equation (2), the log-likelihood $L(\beta, \tau)$ is,

$$\ell(\beta, \tau) = \log L(\beta, \tau) = -\frac{1}{2} bk \cdot \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{b} \sum_{j=1}^{t} \Big[\big(y_{ij} - \mu - \beta_i - \tau_j\big)^2 \Big]^R. \quad (7)$$

To obtain the parameters $\theta = \{\mu, \beta, \tau\}$, we can solve $\frac{\partial \ell}{\partial \theta} = 0$ to find the MLE parameter, $\hat{\theta} = \{\hat{\mu}, \hat{\beta}, \hat{\tau}\}$.

Let $\theta^{(m+1)}$ be the parameters of RICA at an iteration, $\theta^{(m+1)} \in \{\mu^{(m+1)}, \beta^{(m+1)}, \tau^{(m+1)}\}$. The imputed dataset, $Y^{(m)}$, contains the responses of BIBDs and the completely imputed dataset $\{\hat{y}_{ij}^{(m)}\}$, $y_{ij}^{(m)} \in Y^{(m)}$. Its estimation of MLE is $\hat{\theta}^{(m+1)}$ with the solution of $\frac{\partial \ell}{\partial \theta} = 0$. Thus, we have $\hat{\mu}^{(m+1)} = \bar{y}_{oo}^{(m+1)}$, and

$$\hat{\beta}_i^{(m+1)} = \bar{y}_{io}^{(m+1)} - \bar{y}_{oo}^{(m+1)}, \ \hat{\tau}_j^{(m+1)} = \bar{y}_{oj}^{(m+1)} - \bar{y}_{oo}^{(m+1)}, \quad (8)$$

where $\bar{y}_{io}^{(m+1)} = \frac{\sum_{j=1}^{t} y_{ij}^{(m)}}{t}$, $\bar{y}_{oj}^{(m+1)} = \frac{\sum_{i=1}^{t} y_{ij}^{(m)}}{b}$, $\bar{y}_{oo}^{(m+1)} = \frac{\sum_{i=1}^{b} \sum_{j=1}^{t} y_{ij}^{(m)}}{bt}$, $i = 1, 2, \cdots, b$, $j = 1, 2, \cdots, t$.

The result suggests that yields the same MLEs as the RCBD. It is also important to note that estimatability of these treatment effects also extends to contrasts.

If the contrasts in the treatment effects are estimatable in a design, then the missing values of the responses in this design are also estimatable. The potential responses can be estimated after estimating those parameters with the maximum likelihood method.

The missing values of the responses are estimatable in a design if and only if the design is connected. The proposed RICA method can estimate the missing responses in BIBDs. Then, the resulting 'complete' matrix of responses can be analyzed via the RCBD. The missing responses can also be estimated for connected incomplete block designs, even if they are not BIBDs. Why are balanced designs preferred? The next section will investigate more results on the balance property for this method.

4.3. The Role of Balance

Balance is often determined by examining the information matrix of incomplete block designs. Incomplete designs provide us with a new perspective to check the variance of the parameter estimation. Let $V(\hat{\tau}_j - \hat{\tau}_l)$ be the variance of the best linear unbiased estimator (BLUE) [37] of the simple difference $(\hat{\tau}_j - \hat{\tau}_l)$, where τ_j and τ_l are two treatment effects. Balanced designs keep the variance of pairwise contrasts equally small. Another way of thinking about this is that the information, or the inverse of the variance, is equally large. To investigate the role of balance, we have to introduce a definition of the information matrix [32].

Let $\mathbf{Q}$ be an $n \times n$ matrix, which gives an orthogonal projection onto the orthogonal complement of $\mathbf{X}_B$'s column space. The information matrix, $\mathbf{C}$, of the design is a $t \times t$ matrix defined by

$$\mathbf{C} = \mathbf{X}_T^\top \mathbf{Q} \mathbf{X}_T, \ \mathbf{Q} = \mathbf{I}_n - \mathbf{X}_B \mathbf{K}^{-1} \mathbf{X}_B^\top, \tag{9}$$

where $\mathbf{K}$ is a $b \times b$ diagonal matrix whose entries are the block sizes. It can be shown that $\mathbf{Q}\mathbf{X}_B = \mathbf{0}$, $\mathbf{Q}^\top = \mathbf{Q}$, and $\mathbf{Q}^2 = \mathbf{Q}$. The rows of $\mathbf{C}$ all sum to zero. This implies that a vector of all ones is in the kernel of $\mathbf{C}$. Thus, $\mathbf{C}$ is not invertible.

An incomplete block design is variance balanced if the information matrix, $\mathbf{C}$, has an eigenvalue 0 with multiplicity 1; the other eigenvalues are all equal. The multiplicity of 1 indicates that the design is connected. In the case of variance balance, since the vector of ones is a null vector of $\mathbf{C}$, the design is balanced if and only if $\mathbf{C}$ is completely symmetric.

When balanced, the variance can be estimated for an elementary contrast under the intrablock analysis. The following results compare the variance of the RICA method with the variance in classical RCBDs and BIBDs. Theoretical results are also given on when the RICA is more efficient than direct analysis with the BIBDs.

Lemma 1 (Relative efficiency of BIBDs)**.** *Let $\hat{\tau}_j - \hat{\tau}_{j'}$ be an elementary contrast under the intrablock analysis, $j \neq j'$. The relative efficiency of (b, t, k, r, λ)-BIBD relative to (b,t)-RCBD is*

$$\frac{Var(\hat{\tau}_j - \hat{\tau}_{j'})_{BIB}}{Var(\hat{\tau}_j - \hat{\tau}_{j'})_C} = \frac{rk}{\lambda t}\left(\frac{\sigma^2}{\sigma_C^2}\right),$$

where $Var(\hat{\tau}_j - \hat{\tau}_{j'})_C$ and $Var(\hat{\tau}_j - \hat{\tau}_{j'})_{BIB}$ are the variances of an elementary contrast under an RCBD and BIBD, $Var(y_{ij})_C = \sigma_C^2$, and $Var(y_{ij})_{BIB} = \sigma^2$, which is unknown and one estimate from the given data.

$$\hat{\sigma}^2 = \frac{SS_{Error(t)}}{bk - b - t + 1}, \tag{10}$$

where the residual sum of squares is obtained by $SS_{Error(t)} = SS_{Total} - SS_{Block(unadj)} - SS_{Treat(adj)}$, SS_{Total} is the total sum of squares, $SS_{Block(unadj)}$ the unadjusted block sum of squares, and $SS_{Treat(adj)}$ the adjusted treatment sum of squares in the intrablock analysis of BIBDs.

In experiments, we assume that $Var(y_{ij})_C = Var(y_{ij})_{BIB} = \sigma^2$, then the relative efficiency of (b, t, k, r, λ)-BIBD relative to (b, t)-RCBD is $\frac{rk}{\lambda t} = \frac{k(t-1)}{t(k-1)}$. For RICA, its variance is derived as follows. The proof is displayed in the Appendix A.

Theorem 1 (Variance of RICA)**.** *Given (b, t, k, r, λ)-BIBD, the variance of RICA is*

$$\sigma_{RICA}^2 = \hat{\sigma}^2 + Var(\hat{y}_{ij}), \tag{11}$$

where $\hat{\sigma}^2$ *is the estimate from the given data of BIBDs,* $\hat{\sigma}^2 = \frac{SS_{Error(t)}}{bk-b-t+1}$, $SS_{Error(t)}$ *is the residual sum of squares of BIBDs,*

$$Var(\hat{y}_{ij}) = \mathbf{E}\left(\frac{\sum_{i=1}^{b}\sum_{j=1}^{t}\left(e_{ij}+\bar{e}_i+\bar{e}_j\right)^2}{(b-1)(t-1)}\right),$$

where $e_{ij} = y^*_{ij} - y_{ij}$, $\bar{e}_i = \hat{\beta}_i - \hat{\beta}_{i^*}$, $\bar{e}_j = \hat{\tau}_j - \hat{\tau}_{j^*}$, *and*

$$y^*_{ij} = \begin{cases} y_{ij}, & R = 1 \\ \hat{y}_{ij}, & R = 0, \ for\ imputation. \end{cases}$$

In practice, it is helpful to know when RICA or traditional BIBD analysis is preferred. The relative efficiency of RICA to BIBDs is derived to address this question.

Corollary 1 (Relative efficiency of RICA)**.** *Let* $\hat{\tau}_j - \hat{\tau}_{j'}$ *be an elementary contrast under the intrablock analysis,* $j \neq j'$. *The relative efficiency of RICA relative to* (b, t, k, r, λ)*-BIBD is*

$$\frac{Var(\hat{\tau}_j - \hat{\tau}_{j'})_{RICA}}{Var(\hat{\tau}_j - \hat{\tau}_{j'})_{BIB}} = \frac{\lambda t}{rk}\left(1 + \frac{Var(\hat{y}_{ij})}{\hat{\sigma}^2}\right),$$

where $Var(y_{ij})_{BIB} = \hat{\sigma}^2$ *is the estimate from the given data of BIBDs, and*

$$Var(\hat{y}_{ij}) = \mathbf{E}\left(\frac{\sum_{i=1}^{b}\sum_{j=1}^{t}\left(e_{ij}+\bar{e}_i+\bar{e}_j\right)^2}{(b-1)(t-1)}\right).$$

The actual efficiency of RICA over BIBDs not only depends on efficiency factor ($\frac{\lambda t}{rk}$), but also on the ratio of variances. As $Var(\hat{y}_{ij}) \geq 0$, $\sigma^2_{RICA}(= \hat{\sigma} + Var(\hat{y}_{ij}))$ can be larger than σ^2.

This result suggests that RICA is more efficient than the traditional analysis for (b, t, k, r, λ)-BIBDs if

$$Var(\hat{y}_{ij}) < \left(\frac{rk}{\lambda t} - 1\right)\hat{\sigma}^2, \tag{12}$$

where $\hat{\sigma}^2$ is the estimate of $Var(y_{ij})_{BIB}$ from the given data of BIBDs.

As $\lambda(t-1) = r(k-1)$, the efficient condition of RICA to BIBDs can also be written as $Var(\hat{y}_{ij}) < \frac{t-k}{t(k-1)}\hat{\sigma}^2$. If this holds, then the result of RICA is more efficient than BIBDs. Therefore, these theoretical results on relative efficiency inform us when the missingness should be imputed for BIBDs. Practitioners should choose an imputation method with small errors and prediction variances, such as the EM algorithm.

5. Applications on Real-World Datasets

Here, we employ data from ML and agricultural experiments to validate the proposed RICA method.

5.1. Application to ML Datasets

To illustrate the proposed approach, we apply RICA to a binary classification application with four datasets as blocks. The statistical information of those four blocks of datasets is demonstrated in Table 2. The Heart-c and Pima datasets are from UC Irvine's ML database [38,39]. The Framingham dataset is from the heart study [40]. The SHNHS dataset is from the clinical data of the Shanghai nutrition health study [41]. The objectives of these datasets are to diagnostically predict whether or not a patient has heart disease, diabetes, or unhealthy obesity. The classification tasks were conducted based on certain diagnostic attributes of measurements included in the dataset.

Table 2. Statistics of the four datasets as blocks in the design of experiments.

Block	No. Instances	No. Attributes	Classes	Proportion (0:1)
Heart-c	271	13	Absent/Present	150:120
Pima	532	7	No/Yes	355:177
Framingham	4240	14	Absent/Present	3596:644
SHNHS	15,151	15	Absent/Present	13,534:1617

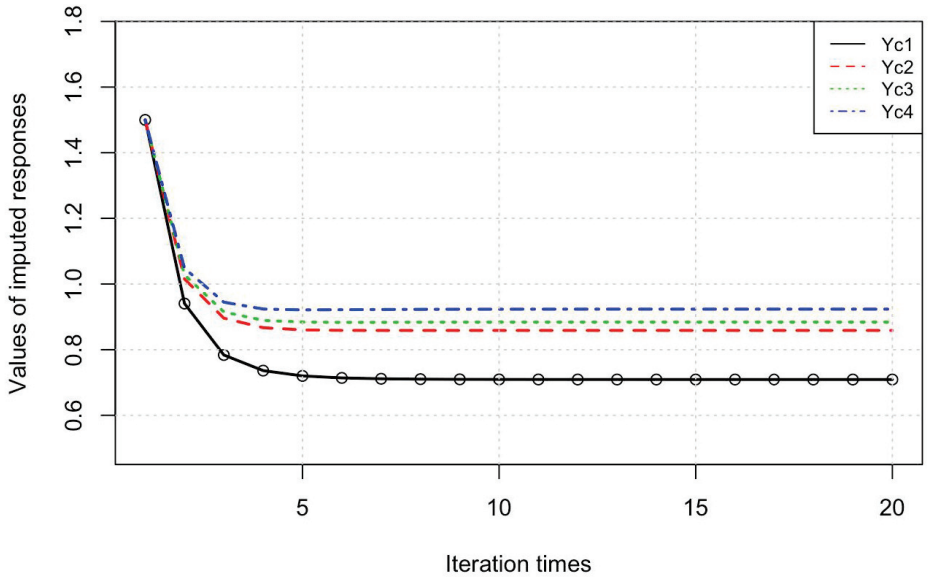
Note: 0:1: Absent/Present, or No/Yes.

From an experimental design perspective, the experiments were designed to compare four methods on the four blocks of datasets. The treatments in the experiment are four ML algorithms (MLAs) for the classification task. These four algorithms are the zero rule (zeroR) algorithm (a benchmark procedure for classification), Ranger (a fast implementation of random forests) [42], XGBoost (an optimized distributed gradient boosting) [43], and support vector machine (SVM) [44]. The parameters of those algorithms are mainly set as default. Each dataset was divided into training and testing sets. The accuracy was calculated using five-fold cross-validation on each dataset.

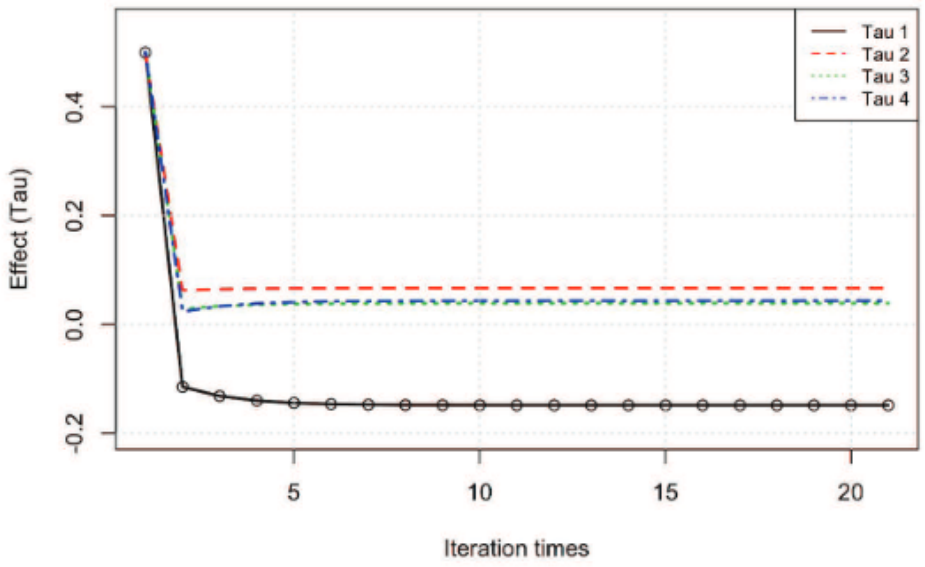
The responses are the measurements of classification accuracy during training and testing. The classification accuracy is the ratio of the number of correct predictions to the total number of total samples. To validate the efficacy of RICA, only three of the four algorithms were applied to each dataset (block). The responses of BIBDs in training and testing are demonstrated in Table 3. The response is measured by the classification accuracy [45], which is the ratio of the number of correct predictions to the total number of total samples.

The initial values of all the parameters were set to 0.5, and the initialized imputation of all the missing values was 1.5. The algorithm converges to stable values within around ten steps (see Figure 3). To further verify Corollary 1, the variances from the given data of BIBDs are estimated and compared with those estimated via the method of multiple imputations (MIs) [46]. They are 0.0121 and 0.0137 in the training and testing, respectively. For the (4,4,3,3,2)-design, $Var(\hat{y}_{ij}) < \left(\frac{rk}{\lambda t} - 1\right)\hat{\sigma}^2 = \frac{1}{8}\hat{\sigma}^2$ with Equation (12), where $Var(\hat{y}_{ij})$ is obtained with the EM algorithm. With the MI method, the variances of 10 imputations are nearly 0.0060, 0.0129, 0.0180, and 0.0167 for the four missing values of the responses in the training, respectively. All of these variances are larger than the threshold $\frac{1}{8}\hat{\sigma}^2 = \frac{1}{8}0.0121 = 0.0015$. Thus, the MI method is less efficient than traditional BIBD analysis by the condition in Corollary 1. Moreover, the means are also compared among the complete data (in the RCBD), the imputations with MIs, and the imputation with EM (see Figure 4). The diamonds in the boxplots are the estimated mean of the imputed values. The ten imputations of Yc2 take two levels, with the median taking the bottom level, where Yc2 denotes the missing value of Y in column 2 of Table 3b. Although the performance of MIs and EM varies among the four missing values, the total error of EM (0.0440) is less than MIs (0.0514). The total error is measured by $\sum_{i=1}^{b}\sum_{j=1}^{t}(1 - \mathbf{R}^*_{ij})(\hat{y}_{ij} - y_{ij})^2$. Thus, the results favor EM over MIs as an imputation method in this scenario. The performances of the proposed method are showed in Table 4. It can be found that the RICA performs well in training and testing data.

Table 5 shows the maximum likelihood estimates of the treatment and block effects. RICA and BIBDs produced the same parameter estimates. The complete responses were also used to fit the model in Equation (8). Although RICA uses imputed data, its results are easier to interpret and calculate than those of traditional BIBDs. This is because the treatment effect is estimated with $\bar{y}_{oj} - \bar{y}_{oo}$, which illustrates that the effect is the difference between the average accuracy of the j-th ML algorithm and the overall accuracy. For example, the RICA estimate of $\hat{\tau}_2$ = 0.067 means that the average prediction accuracy of the Ranger algorithm is 6.7% higher than the grand mean. This can be easily compared with the RICA estimate $\hat{\tau}_1 = -0.149$, which shows that the first ML (zeroR) has an average accuracy of 14.9% lower than the grand mean.



(**a**)



(**b**)

Figure 3. Iterations of imputed values and effects with ML data in testing. Yc1: the missing value of Y in column 1 of Table 3b. The effects of Tau 3 and Tau 4 are close to each other. (**a**) Imputed values; (**b**) Effects of treatments.

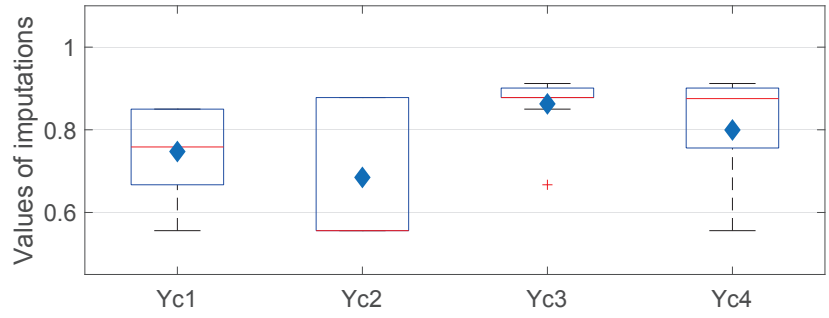
Table 3. BIBDs with ML algorithms and datasets.

(a) Table of BIBDs			
Block	**Assignments of MLA**		
Heart-c	SVM	XGBoost	ZeroR
Pima	ZeroR	SVM	Ranger
Fmh	ZeroR	Ranger	XGBoost
SHNHS	XGBoost	Ranger	SVM

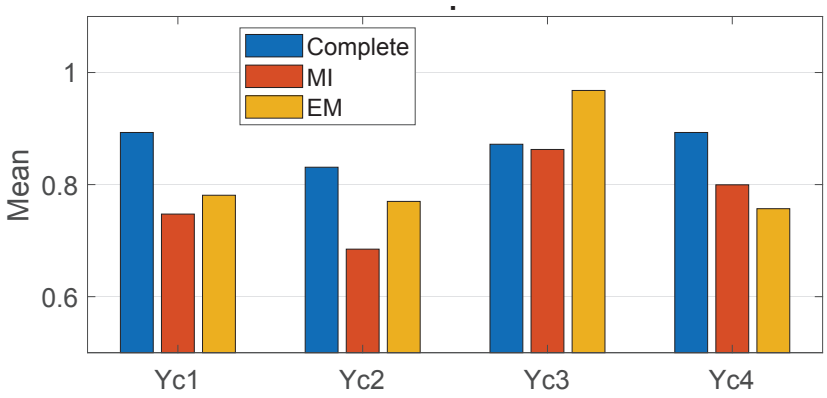
Table 3. *Cont.*

(b) Incidence Matrix of BIBDs				
Block	**ZeroR**	**Ranger**	**XGBoost**	**SVM**
Heart-c	1	0	1	1
Pima	1	1	0	1
Fmh	1	1	1	0
SHNHS	0	1	1	1

Note: Fmh stands for the Framingham dataset.



(**a**)



(**b**)

Figure 4. Boxplots of imputed values from multiple imputations (**a**) and comparisons on their means (**b**). The diamonds in the boxplots are the estimated mean of the imputed values. The response is averaged across ten imputations for MI; the observed value is obtained via the RCBD (in blue), while the response is imputed once using the EM algorithm. Yc1: the missing value of Y in column 1 of Table 3b.

Table 4. Experiments with two designs (block: data subsets; treatment: ML algorithms).

Run	**Block**	**Treatment**	**Training**			**Testing**		
ω	i	j	**RCBD**	**BIBD**	**RICA**	**RCBD**	**BIBD**	**RICA**
1	Heart-c	ZeroR	0.556	0.556	0.556	0.556	0.556	0.556
2	Heart-c	Ranger	0.893	NA	0.781	0.871	NA	0.858
3	Heart-c	XGBoost	0.756	0.756	0.756	0.879	0.879	0.879
4	Heart-c	SVM	0.912	0.912	0.912	0.874	0.874	0.874
5	Pima	ZeroR	0.667	0.667	0.667	0.667	0.667	0.667
6	Pima	Ranger	0.824	0.824	0.824	0.985	0.985	0.985
7	Pima	XGBoost	0.831	NA	0.77	0.828	NA	0.884
8	Pima	SVM	0.818	0.818	0.818	0.845	0.845	0.845
9	Framingham	ZeroR	0.848	0.848	0.848	0.848	0.848	0.848

Table 4. *Cont.*

Run	Block	Treatment	Training			Testing		
ω	i	j	RCBD	BIBD	RICA	RCBD	BIBD	RICA
10	Framingham	Ranger	0.878	0.878	0.878	0.886	0.886	0.886
11	Framingham	XGBoost	0.85	0.85	0.85	0.863	0.863	0.863
12	Framingham	SVM	0.872	NA	0.968	0.876	NA	0.923
13	SHNHS	ZeroR	0.893	NA	0.757	0.893	NA	0.709
14	SHNHS	Ranger	0.912	0.912	0.912	0.912	0.912	0.912
15	SHNHS	XGBoost	0.901	0.901	0.901	0.904	0.904	0.904
16	SHNHS	SVM	0.907	0.907	0.907	0.906	0.906	0.906

Note: 0.893 (NA, 0.709) indicates the responses (ML accuracy) with RCBD, incompleteness with BIBDs, and imputation with RICA.

The results of ANOVA are summarized in Table 6. Using $\alpha = 0.05$, RICA suggests that mean classification accuracy differs among the different ML algorithms (p-value = 0.0041). On the other hand, both the RCBD and BIBDs do not detect significant differences among the MLAs at the $\alpha = 0.05$ level. In practice, the RICA method can be implemented to compare the performance of the ML algorithms with a number of blocks of datasets.

Table 5. MLE estimation of the parameters with MLA and the datasets.

	RCBD		BIBDs		RICA	
Index	**Beta**	**Tau**	**Beta**	**Tau**	**Beta**	**Tau**
Training						
1	−0.053	−0.091	−0.068	−0.112	−0.068	−0.112
2	−0.047	0.044	−0.049	0.030	−0.049	0.030
3	0.030	0.002	0.067	<0.001	0.067	<0.001
4	0.071	0.045	0.050	0.082	0.050	0.082
Testing						
1	−0.055	−0.109	−0.052	−0.149	−0.052	−0.149
2	−0.018	0.064	0.002 *	0.067	0.002 *	0.067
3	0.019	0.019	0.036	0.039	0.036	0.039
4	0.054	0.026	0.014	0.043	0.014	0.043

Note: 0.002 *: close to 0.0015. <0.001: positive but close to 0.

Table 6. ANOVA of the RCBD, BIBDs, and RICA with MLA and the datasets in testing.

Model	Df	Sum Sq	Mean Sq	F Value	Pr (>F)
RCBD					
Treatment (MLA)	3	0.07	0.02	3.39	0.0676 .
Block (datasets)	3	0.03	0.01	1.30	0.3337
Residuals	9	0.06	0.01		
BIBDs					
Treatment (MLA)	3	0.10	0.03	4.21	0.0779 .
Block (datasets)	3	0.01	<0.01	0.49	0.7068
Residuals	5	0.04	0.01		
RICA					
Treatment (MLA)	3	0.12	0.04	9.23	0.0041 ..
Block (datasets)	3	0.02	0.01	1.31	0.3296
Residuals	9 #	0.04	<0.01		

#: the error degree of freedom for RICA is $(b-1)(t-1) - \sum_i \sum_j (1 - \mathbf{R}^*_{ij})$. Signif. codes: $p > 0.001$ '..' $p > 0.01$ '.'.

5.2. Comparison with Related Methods

For further comparison, we represented the machine learning (ML) data as a matrix with three columns: outcome, treatment, and block. We applied SoftImpute 100 times using default configurations in R. The average total imputation error for the training dataset was 0.2478 (see Table 7), which is significantly larger than the errors associated with EM at 0.0440 and multiple imputations (MIs) at 0.0514.

Among the 100 imputation trials using SoftImpute, none yielded p-values below 0.05 (mean = 0.0642, standard deviation = 0.0031), suggesting that this approach fails to detect significant differences in prediction accuracy among the ML algorithms.

Finally, the ANOVA results for the imputed datasets indicated both average and median mean squared errors (MSEs) of approximately 0.0096 (rounded to three decimal places), which exceed the MSE for RICA, approximately 0.0058. Therefore, EM is a superior choice for imputation in RICA compared with SoftImpute and multiple imputation methods.

Table 7. Comparison with related methods.

Methods	**SoftImpute**	**RICA**
Total imputation error (mean)	0.2478	0.044
p-values	0.0642 ± 0.0031	N/A
ANOVA MSE (both mean and median)	0.0096	0.0058

"N/A" denote there is not exist value.

5.3. More Applications on Agricultural Datasets

Datasets from agricultural experiments [47] were also employed to validate the proposed method. These datasets try to determine the impact of different lines (genotypes) of plants on various outcomes. Here, we used the dataset (Dataset_1_EYT) collected in wheat yield trials. In this dataset, four locations/environments were used as blocks ($b = 4$). The dataset has four possible responses: days to heading (DTHD), days to maturity (DTMT), plant height (Height), and grain yield (GY). This example is well-motivated for BIBD and RICA, because it is often expensive and time-consuming to plant many crops.

A subset (renamed Pheno_subset) of the data (Dataset_1_EYT) was taken with the first four lines, that is, $t = 4$ treatments. Note that there are many possible treatments; here we just tried the first four available. Assuming a BIBD with a block size of 3, we had the same missingness scheme as in Table 3b. The grain yield (GY) was taken as the response. After adjusting the DF for residuals in RICA, the results for the RICA and RCBD match (in Table 8), but the BIBD seems to give a false positive.

Table 8. ANOVA of RCBD, BIBDs, and RICA with Pheno_subset.

Model	**Df**	**Sum Sq**	**Mean Sq**	**F Value**	**Pr (>F)**
RCBD					
Treatment (genotype)	3	1.62	0.54	3.04	0.09 .
Block (location)	3	38.16	12.72	71.65	<0.001 ...
Residuals	9	1.60	0.18		
BIBDs					
Treatment (genotype)	3	1.46	0.49	14.04	0.007 ..
Block (location)	3	26.00	8.67	250.86	<0.001 ...
Residuals	5	0.17	0.04		

Table 8. *Cont.*

Model	Df	Sum Sq	Mean Sq	F Value	Pr (>F)
RICA					
Treatment (genotype)	3	1.30	0.44	5.55	0.02 .
Block (location)	3	32.27	10.76	137.35	<0.001 ...
Residuals	9 #	0.70	0.08		

#: the error degree of freedom for RICA is $(b-1)(t-1) - \sum_i \sum_j (1 - \mathbf{R}^*_{ij})$. Signif. codes: $p > 0.001$ '..' $p > 0.01$ '.' $p > 0.05$ '...'.

6. Discussion and Conclusions

In general, whenever missing data arises in a dataset, the question of whether to impute the missing data emerges. This work serves as an important initial step in providing a theoretical guideline for addressing this question in the context of data obtained from a BIBD. This paper provides a novel method for the analysis of BIBDs called RICA. This method imputes missing responses in a BIBD and analyzes the subsequent dataset with the RCBD model. This method has two main contributions: it imputes the missing part of BIBDs and estimates the effects of treatments and blocks. This method provides a new perspective on the analysis of BIBDs, with a link to missing value problems and a better understanding of the balance property. Moreover, the efficiency of this method is demonstrated with an application to ML algorithms. Furthermore, the identifiability result suggests that the missingness scheme may be nonignorable for each block, but the whole data of an incomplete design are missing by design when the design is balanced. Theoretical results on relative efficiency also provide guidelines on when the missingness should be imputed for incomplete designs with the role of balanced variance.

This method provides new insights for comparing ML algorithms on tasks (such as classification) with massive blocks of data. The comparisons based on BIBDs can provide an efficient and interpretable estimation of the treatment effect of these algorithms on the prediction accuracy, while maintaining small total runs of the algorithms. However, according to the missingness mechanism, the complete block means may not be equal to the mean of the observed data in that block because there are unobserved treatment effects in each block. These treatment effects lead to the bias in the estimation of the mean in that block. Although $\bar{y}_{oo}$ can be de-biased with BIBDs, the linear model cannot be used to estimate the value of the incomplete data directly. More assumptions are required to achieve unbiased estimators of the model parameters (β_i, τ_j).

Corollary 1 is significant because it outlines when imputation should be applied, based on the relative variances of treatment contrasts within the design. It demonstrates that this relative efficiency relies on the ratio of estimated variances and an efficiency factor that is influenced by the design choice. This finding is crucial as it indicates that there is no "one-size-fits-all" approach to handling missing data; in some cases, imputation may be advantageous, while in others, it may not be.

Moreover, practitioners should recognize that imputed data do not constitute independent observations, which necessitates modifications to the degrees of freedom in the final ANOVA table. From a theoretical standpoint, it is important to understand how this result derives from the assumptions underlying the experimental design. Specifically, Proposition 2 assumes that the probability of a particular value being missing is independent of its actual value, which follows from the selection of the BIBD from an equiprobable set of BIBDs prior to observing any values. It is essential for both practitioners and theorists to grasp that, if the experimental design is not randomly selected, the missingness mechanism may be nonignorable (as indicated in Proposition 1), potentially causing most imputation methods to fail.

Our results assume that the model in (1) is correct, meaning that the responses are continuous and follow a normal distribution. To extend this work to binary responses, we could employ a Generalized Linear Model (GLM) with a binary response, with logistic regression being the most common example. In this case, we would replace the model (1) with a logistic regression model and substitute the likelihood in Equation (2) with a binomial likelihood function.

One limitation is that the maximum likelihood estimates (MLEs) for the logistic regression model are not available in closed form; they are typically determined using iterative numerical algorithms, such as Newton–Raphson. This prevents us from providing explicit formulas for the MLEs, although they can still be computed. A challenging aspect would be deriving the relative efficiency of RICA compared with BIBD under this model, as the variance of MLEs in a logistic regression model explicitly depends on the true unknown treatment and block effects.

Furthermore, some drawbacks of the RICA method are worth noting. The fitted model is strictly linear and does not allow for a more general nonlinear relationship between the treatments and responses; this limitation is shared with BIBDs and RCBDs. Another drawback of RICA is that running ANOVA on the imputed data can lead to misleading results. There are two main reasons. First, the presented RICA method results in a downward-biased MSE because it does not estimate the variance due to imputation (i.e., prediction variance). If the estimated variance of RICA (i.e., $\hat{\sigma}^2$) is close to that of the RCBD, then the downward bias is not a large problem. Second, the imputed data in the RCBD are not truly independent observations because the imputed values are calculated deterministically from the observed values. It may be worth considering a multiple imputation method to remedy this.

In the future, the RICA method would be extended for generalized incomplete block design (IBD). In general, IBDs may have more missing responses than BIBDs, e.g., partial BIBDs. One possible approach is to impute enough missing responses in a partial BIBD to reach a BIBD. This could be compared to imputing all missing responses in a partial BIBD. Future work will provide more examples of PBIBDs and their real RICA applications.

Author Contributions: Conceptualization, H.Y., B.H. and J.C.; methodology, H.Y., B.H. and J.C.; software, H.Y.; validation, H.Y. and J.C.; formal analysis, B.H. and N.R.; data curation, N.R.; writing—original draft preparation, H.Y.; writing—review and editing, B.H., N.R. and J.C.; supervision, H.Y., B.H. and J.C.; funding acquisition, H.Y. and J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research Yu was supported by Chongqing Municipal Education Commission (No. KJQN202200608, 22SKGH149), Ministry of Education Foundation on Humanities and Social Sciences (No. 23YJC630215, 23XJAZH004) and Chongqing Municipal Science and Technology Bureau (No. 2022TIAD-KPX0155, 2023DBXM008, CSTB2023NSCQ-MSX0073). Chen was supported by Beijing Institute of Technology Research Fund Program for Young Scholars and the National Natural Science Foundation of China via grant number 12301318.

Data Availability Statement: All data included in this article are available upon request by contact with the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proof of Theorem 1

Proof. The expectation of mean squares of errors with RICA is

$$SSE_{RICA} = \sum_{i=1}^{b}\sum_{j=1}^{t}\left(y_{ij}^{*}-\hat{y}_{ij}\right)^2 = \sum_{i=1}^{b}\sum_{j=1}^{t}\left(y_{ij}^{*}-\hat{\mu}-\hat{\beta}_{i^*}-\hat{\tau}_{j^*}\right)^2,$$

where $\hat{\beta}_{i^*}$ and $\hat{\tau}_{j^*}$ are the MLEs of BIBDs and RICA.

Let $\hat{\beta}_i$ and $\hat{\tau}_j$ be the estimators obtained in the MLE solution of the RCBD. Thus, we have

$$\begin{aligned} SSE_{RICA} &= \sum_i \sum_j \left(y_{ij}^* - y_{ij} + y_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j + \bar{e}_i + \bar{e}_j\right)^2 \\ &= \sum_i \sum_j \left[(y_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j) + \left(y_{ij}^* - y_{ij} + \bar{e}_i + \bar{e}_j\right)\right]^2 \\ &= \sum_i \sum_j \left[(y_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j)^2 + (e_{ij} + \bar{e}_i + \bar{e}_j)^2\right] \\ &\quad + 2\sum_i \sum_j \left[(y_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j)(e_{ij} + \bar{e}_i + \bar{e}_j)\right], \end{aligned}$$

where $e_{ij} = y_{ij}^* - y_{ij}$, $\bar{e}_i = \hat{\beta}_i - \hat{\beta}_{i^*}$, $\bar{e}_j = \hat{\tau}_j - \hat{\tau}_{j^*}$. $\sum_i \sum_j (y_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j)(e_{ij} + \bar{e}_i + \bar{e}_j) = 0$.

According to the analysis of variance of the RCBD, the expectation of mean squared errors is

$$\sigma_C^2 = \mathbf{E}(MSE_C) = \mathbf{E}\left(\frac{SSE_C}{(b-1)(t-1)}\right), \tag{A1}$$

where $SSE_C = \sum_i \sum_j \left(y_{ij} - \hat{\mu} - \hat{\beta}_i - \hat{\tau}_j\right)^2$. Thus, we have

$$SSE_{RICA} = SSE_C + \sum_{i=1}^{b} \sum_{j=1}^{t} (e_{ij} + \bar{e}_i + \bar{e}_j)^2.$$

Similarly, for analysis of variance with RICA, the expectation of mean squared errors is

$$\sigma_{RICA}^2 = \mathbf{E}(MSE_{RICA}) = \mathbf{E}\left(\frac{SSE_{RICA}}{(b-1)(t-1)}\right)$$

Thus, the result is obtained as $\sigma_{RICA}^2 = \sigma_C^2 + Var(\hat{y}_{ij})$, where $Var(\hat{y}_{ij}) = \mathbf{E}\left(\frac{\sum_{i=1}^{b} \sum_{j=1}^{t} (e_{ij} + \bar{e}_i + \bar{e}_j)^2}{(b-1)(t-1)}\right)$.

By Lemma 1, we obtain the results as Equation (11). □

References

1. Konečnỳ, J.; McMahan, H.B.; Yu, F.X.; Richtárik, P.; Suresh, A.T.; Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv* **2016**, arXiv:1610.05492.
2. Cai, T.T.; Wei, H. Distributed nonparametric function estimation: Optimal rate of convergence and cost of adaptation. *Ann. Stat.* **2022**, *50*, 698–725. [CrossRef]
3. Yu, H. *Data Quality Management in the Data Age*; Springer Briefs in Service Science: Cham, Switzerland, 2024; Volume 1.
4. Yu, H.; Piri, S.; Qiu, H.; Xu, R.; Miao, H. Personalized algorithmic pricing decision support tool for health insurance: The case of stratifying gestational diabetes mellitus into two groups. *Inf. Manag.* **2024**, *61*, 103945. [CrossRef]
5. Chen, J.; Peng, J.; Lin, D.K. A statistical perspective on non-deterministic polynomial-time hard ordering problems: Making use of design for order-of-addition experiments. *Comput. Ind. Eng.* **2021**, *162*, 107773. [CrossRef]
6. Chen, J.; Han, X.; Lin, D.K.; Yang, L.; Zhou, Y. On ordering problem: A statistical approach. *Stat. Sin.* **2023**, *33*, 1903–1922. [CrossRef]
7. Rios, N.; Winker, P.; Lin, D.K. TA algorithms for D-optimal OofA Mixture designs. *Comput. Stat. Data Anal.* **2022**, *168*, 107411. [CrossRef]
8. Lin, D.K.; Chen, J. Adaptive order-of-addition experiments via the quick-sort algorithm. *Technometrics* **2023**, *65*, 396–405. [CrossRef]
9. Yang, C.C.; Lin, D.K. A note on selection of basis quantities for dimensional analysis. *Qual. Eng.* **2021**, *33*, 240–251. [CrossRef]
10. Huang, S.; Wan, X.; Qiu, H.; Li, L.; Yu, H. Constrained optimization for stratified treatment rules with multiple responses of survival data. *Inf. Sci.* **2022**, *596*, 343–361. [CrossRef]
11. Yu, H.; Chen, J. Treatment effect identification using two-level designs with partially ignorable missing data. *Inf. Sci.* **2022**, *611*, 277–300. [CrossRef]
12. Yu, H.; Zuo, X.; Tang, J.; Fu, Y. Identifying causal effects of the clinical sentiment of patients' nursing notes on anticipated fall risk stratification. *Inf. Process. Manag.* **2023**, *60*, 103481. [CrossRef]
13. Yu, H.; Deng, X.; Tang, J.; Yue, F. Patterns identification using blind source separation with application to neural activities associated with anticipated falls. *Inf. Sci.* **2025**, *689*, 121410. [CrossRef]
14. Box, G.E.; Hunter, J.; Hunter, W. *Statistics for Experimenters: Design, Innovation and Discovery*; John Wiley & Sons: Hoboken, NJ, USA, 2005.

15. Chen, J.; Mukerjee, R.; Lin, D.K. Construction of optimal fractional order-of-addition designs via block designs. *Stat. Probab. Lett.* **2020**, *161*, 108728. [CrossRef]
16. Yu, H.; Chiu, Y.L.; Wang, J.N.; Yu, J.; Hsu, Y.T. Electronic consultation accessibility influence on patient assessments: A case–control study with user-generated tags of physician expertise. *Digit. Health* **2023**, *9*, 20552076231180693. [CrossRef]
17. De Cooman, G.; De Bock, J. Randomness is inherently imprecise. *Int. J. Approx. Reason.* **2022**, *141*, 28–68. [CrossRef]
18. Nie, X.; Huling, J.; Qian, P.Z. Accelerating large-scale statistical computation with the GOEM algorithm. *Technometrics* **2017**, *59*, 416–425. [CrossRef]
19. Yu, H.; Yang, C.C.; Yu, P. Constrained optimization for stratified treatment rules in reducing hospital readmission rates of diabetic patients. *Eur. J. Oper. Res.* **2023**, *308*, 1355–1364. [CrossRef]
20. Bradley, R.A.; Terry, M.E. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **1952**, *39*, 324–345. [CrossRef]
21. Giesbrecht, F. Analysis of data from incomplete block designs. *Biometrics* **1986**, *42*, 437–448. [CrossRef]
22. de Mendiburu, F.; Yaseen, M. Agricolae: Statistical Procedures for Agricultural Research. R Package Version **2020**. Available online: https://cran.r-project.org/package=agricolae (accessed on 20 September 2024).
23. Yi, X.; Caramanis, C. Regularized EM algorithms: A unified framework and statistical guarantees. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 1, pp. 1567–1575.
24. Tan, Z. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika* **2020**, *107*, 137–158. [CrossRef]
25. Wolfinger, R.; Federer, W.T.; Cordero-Brana, O. Recovering information in augmented designs, using SAS PROC GLM and PROC MIXED. *Agron. J.* **1997**, *89*, 856–859. [CrossRef]
26. Patterson, H.D.; Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **1971**, *58*, 545–554. [CrossRef]
27. Federer, W.T. Recovery of interblock, intergradient, and intervariety information in incomplete block and lattice rectangle designed experiments. *Biometrics* **1998**, *54*, 471–481. [CrossRef]
28. Honaker, J.; King, G.; Blackwell, M. Amelia II: A program for missing data. *J. Stat. Softw.* **2011**, *45*, 1–47. [CrossRef]
29. De Boeck, P.; Bakker, M.; Zwitser, R.; Nivard, M.; Hofman, A.; Tuerlinckx, F.; Partchev, I. The estimation of item response models with the lmer function from the lme4 package in R. *J. Stat. Softw.* **2011**, *39*, 1–28. [CrossRef]
30. Manski, C.F. Partial identification with missing data: Concepts and findings. *Int. J. Approx. Reason.* **2005**, *39*, 151–165. [CrossRef]
31. Montgomery, D.C. *Design and Analysis of Experiments*, 10th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2017.
32. Bailey, R.A. *Association Schemes: Designed Experiments, Algebra and Combinatorics*; Cambridge University Press: Cambridge, UK, 2004; Volume 84.
33. Box, G.E.; Draper, N.R. *Response Surfaces, Mixtures, and Ridge Analyses*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2007; Volume 649.
34. Mealli, F.; Rubin, D.B. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **2015**, *102*, 995–1000. [CrossRef]
35. Sun, B.; Liu, L.; Miao, W.; Wirth, K.; Robins, J.; Tchetgen, E.J.T. Semiparametric estimation with data missing not at random using an instrumental variable. *Stat. Sin.* **2018**, *28*, 1965–1983.
36. Dean, A.; Voss, D.; Draguljić, D. *Design and Analysis of Experiments*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2017.
37. Cakiroglu, S.A. Optimal regular graph designs. *Stat. Comput.* **2018**, *28*, 103–112. [CrossRef]
38. Asuncion, A.; Newman, D. UCI Machine Learning Repository. 2007. Available online: http://www.ics.uci.edu/~mlearn/MLRepository.html (accessed on 20 September 2024).
39. Xu, M.; Yu, H.; Shen, J. New approach to eliminate structural redundancy in case resource pools using α mutual information. *J. Syst. Eng. Electron.* **2013**, *24*, 625–633. [CrossRef]
40. Schnabel, R.B.; Sullivan, L.M.; Levy, D.; Pencina, M.J.; Massaro, J.M.; D'Agostino, R.B., Sr.; Newton-Cheh, C.; Yamamoto, J.F.; Magnani, J.W.; Tadros, T.M.; et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): A community-based cohort study. *Lancet* **2009**, *373*, 739–745. [CrossRef] [PubMed]
41. Xu, R.; Zhang, Y.; Gao, X.; Wan, Y.; Fan, Z. High-sensitivity CRP (C-Reactive Protein) is associated with incident carotid artery plaque in Chinese aged adults. *Stroke* **2019**, *50*, 1655–1660. [CrossRef] [PubMed]
42. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17.. [CrossRef]
43. Liu, P.; Fu, B.; Yang, S.X.; Deng, L.; Zhong, X.; Zheng, H. Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer. *IEEE Trans. Biomed. Eng.* **2020**, *68*, 148–160. [CrossRef]
44. Huang, C.M.; Lee, Y.J.; Lin, D.K.; Huang, S.Y. Model selection for support vector machines via uniform design. *Comput. Stat. Data Anal.* **2007**, *52*, 335–346. [CrossRef]
45. Yu, H.; Shen, J.; Xu, M. Temporal case matching with information value maximization for predicting physiological states. *Inf. Sci.* **2016**, *367*, 766–782. [CrossRef]

46. Su, Y.S.; Gelman, A.; Hill, J.; Yajima, M. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *J. Stat. Softw.* **2011**, *45*, 1–31. [CrossRef]
47. Montesinos-Lopez, O.A.; Montesinos-Lopez, A.; Acosta, R.; Varshney, R.K.; Bentley, A.; Crossa, J. Using an incomplete block design to allocate lines to environments improves sparse genome-based prediction in plant breeding. *Plant Genome* **2022**, *15*, e20194. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.