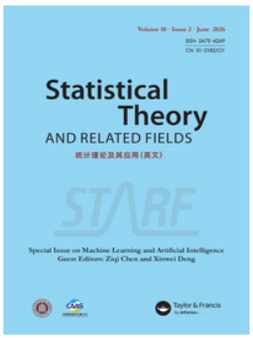

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: www.tandfonline.com/journals/tstf20

CorrDA: correlation-matrix driven discriminant analysis

Feifei Yan, Yingjie Zhang, Jing Ning, Hai Shu & Ziqi Chen

To cite this article: Feifei Yan, Yingjie Zhang, Jing Ning, Hai Shu & Ziqi Chen (2026) CorrDA: correlation-matrix driven discriminant analysis, Statistical Theory and Related Fields, 10:2, 167-183, DOI: 10.1080/24754269.2026.2652551

To link to this article: https://doi.org/10.1080/24754269.2026.2652551

STATISTICAL THEORY AND RELATED FIELDS
2026, VOL. 10, NO. 2, 167–183
https://doi.org/10.1080/24754269.2026.2652551

CorrDA: correlation-matrix driven discriminant analysis

Feifei Yan[a], Yingjie Zhang[b], Jing Ning[c], Hai Shu[d] and Ziqi Chen[b]

[a]School of Mathematics and Statistics, Zhoukou Normal University, Zhoukou, People's Republic of China ;
[b]School of Statistics, KLATASDS-MOE, East China Normal University, Shanghai, People's Republic of China;
[c]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA;
[d]Department of Biostatistics, School of Global Public Health, New York University, New York, NY, USA

ABSTRACT

This article introduces a novel approach to integrating correlation matrix information from training samples to construct a classification rule for testing samples. Traditional discriminant analysis methods that rely solely on mean vectors tend to perform poorly when the mean of the training samples is not indicative of the testing samples. To address this limitation, we propose a new discriminant analysis method called Correlation-matrix driven Discriminant Analysis (CorrDA). By considering the correlation matrices of different classes in the training samples, we can capture the unique patterns among the classes. CorrDA utilizes the Bayes classifier and mixture models to effectively incorporate the correlation matrix information derived from the training samples, thereby improving the discriminant analysis performance on the testing data. Through the analysis of COVID-19 datasets and extensive simulation studies, we provide empirical evidence demonstrating the superior performance of CorrDA.

ARTICLE HISTORY
Received 17 February 2025
Revised 26 November 2025
Accepted 25 March 2026

KEYWORDS
Bayes classifier; correlation matrix; classification; EM algorithm; mixture model; pseudo-likelihood

1. Introduction

Discriminant analysis aims to employ appropriate classification rules to effectively classify observations into distinct classes (Bensmail & Celeux, 1996; Fraley & Raftery, 2002; Mclachlan, 1992). Traditional discriminant analysis methods such as linear or quadratic discriminant analysis (LDA or QDA) classify observations based on the mean vectors of different classes (Anderson, 2003; Bickel & Levina, 2004; Hastie et al., 2009; Jiang et al., 2020; Sohil et al., 2022; Vovan, 2018; Witten & Tibshirani, 2011). There is also a wide range of LDA/QDA variants, including Regularized Discriminant Analysis (Friedman, 1989), High-Dimensional Discriminant Analysis (HDDA) (Bouveyron et al., 2007), Cellwise Robust Regularized Discriminant Analysis (Aerts & Wilms, 2017), Wasserstein Discriminant Analysis (Flamary et al., 2018), and Weighted LDA (L. Xu et al., 2018). These methods account for high dimensionality, robustness, distance metrics, or class imbalance but assume that the training and testing observations share the same distribution, as in traditional LDA and QDA.

CONTACT Ziqi Chen zqchen@fem.ecnu.edu.cn East China Normal University No. 3663 North Zhongshan Road, Putuo District Shanghai, 200062, People's Republic of China
*Equal contributions.

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/24754269.2026.2652551

However, in cases where the mean of the training observations is not informative to the testing dataset, the traditional discriminant analysis methods based on the mean vectors may not perform well. Using the COVID-19 epidemic as a case study, we can observe two clearly distinguishable phases in China subsequent to the implementation of mandatory interventions by the government. These phases include the serious phase, which took place from January 24 to February 6, 2020, and the improved phase, which extended from February 7 to February 20, 2020 (M. Chen et al., 2023; S. Chen et al., 2020; Lauer et al., 2020; Liu et al., 2025; H. Ma et al., 2023; Tian et al., 2020; J. Xu & Tang, 2021). The number of COVID-19 infections varies significantly across regions due to differences in population size, intervention strategies, and epidemic duration. It is therefore challenging to apply the classic discriminant analysis to determine the epidemic phase of other regions of interest by using China's COVID-19 data as the training set, since the magnitude of COVID-19 infections from China is not informative for comparative inference for other regions. After analysing incidence data from multiple cities in China, we discover noteworthy variations in correlation matrices between the two distinct phases as shown in Figure 1. During the serious phase, we observe a progressively stronger trend, while the improved phase shows a gradually diminishing trend.

To enhance the classification accuracy when the mean vectors of the training data fail to provide information for the testing data, we propose integrating correlation information from different classes in the training set. This can be achieved by leveraging the distinct patterns observed in the correlation matrices among various classes. By incorporating this information into discriminant analysis, we can improve the accuracy of determining the labels of observations in the testing set. Our proposed method utilizes mixture models (Fu et al., 2020; McLachlan et al., 2019) to incorporate the distinctive patterns in the correlation matrices among different classes. It utilizes the maximum a posteriori (MAP) rule (Mclachlan, 1992) and employs the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for implementation. Initially, we estimate the unknown parameters using the EM algorithm, and subsequently assign each subject in the testing data to the class that yields the highest pseudo-posterior probability. This method offers computational efficiency due to the closed-form calculation of the conditional expectation in the E-step and most parameters in the M-step.

Prior works that directly utilize correlation information for discriminant analysis, such as those proposed by Y. Ma et al. (2007) and Lei et al. (2009), focus on identifying a linear transformation that maximizes either the difference or the ratio between the between-class and within-class correlation matrices. However, these methods assume that the mean of the training data aligns with that of the testing data, and they do not explicitly model class-specific correlation matrices. In contrast, our approach directly models class-specific correlation matrices and applies a MAP-based discriminant rule combined with the EM algorithm, offering greater flexibility and applicability, especially when the training sample means are not informative for the testing dataset.

The remaining sections of this article are structured as follows. In Section 2, we introduce the proposed CorrDA method. In Section 3, we demonstrate the effectiveness of our method through real data analysis. The performance of CorrDA is further evaluated through simulation studies in Section 4. Finally, we present our concluding remarks in Section 5. All technical formula derivations, together with additional results from the analysis of real data and simulation studies, are deferred to the Supplementary Materials.

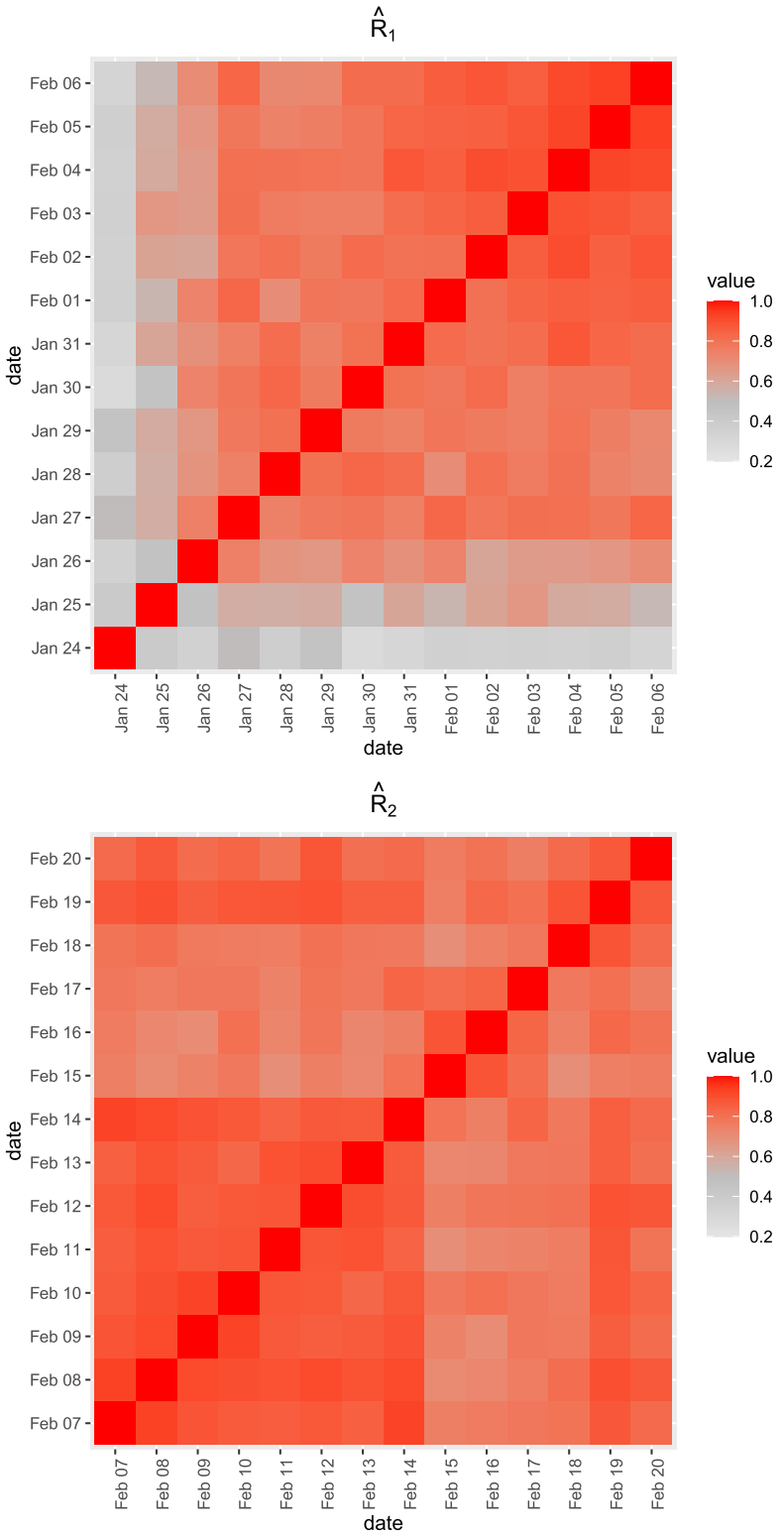

Figure 1. The heatmap of sample correlation matrices $\widehat{\boldsymbol{R}}_1$ (January 24 to February 6, 2020) and $\widehat{\boldsymbol{R}}_2$ (February 7 to February 20, 2020) for China's COVID-19 data.

2. Methodology

In the training samples, we denote the p-dimensional sample mean vector and the $p \times p$ sample correlation matrix of Class h as $\tilde{\boldsymbol{\mu}}_h$ and $\widetilde{\boldsymbol{R}}_h$, respectively, where h ranges from 1 to K with K being the number of classes. Assume that $\{\widetilde{\boldsymbol{R}}_h\}_{h=1}^K$ characterize the distinctive pattern among the K classes. Given the ith observation in the testing dataset, $\boldsymbol{V}_i \in \mathbb{R}^p$ $(i = 1, \ldots, n)$, we aim to classify it into one of the K classes. Denote the true class membership of $\boldsymbol{V}_i$ as $Z_i \in \{1, \ldots, K\}$. Let the mean vector of $\boldsymbol{V}_i$ be $\boldsymbol{\mu}$, and let the covariance matrix of $\boldsymbol{V}_i$ conditional on $Z_i = h$ have a variance-correlation decomposition $\boldsymbol{A}^{\frac{1}{2}}\widetilde{\mathbf{R}}_h\boldsymbol{A}^{\frac{1}{2}}$. The probability density function conditional on $Z_i = h$ is assumed to be $f_h(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \widetilde{\mathbf{R}}_h)$. Let π_h be the prior probability of $Z_i = h$ with $\sum_{l=1}^K \pi_l = 1$. The Bayes' theorem indicates that the posterior probability that an observation $\boldsymbol{V}_i$ belongs to Class h is

$$P(Z_i = h \mid \boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi}) = \frac{\pi_h f_h(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \widetilde{\mathbf{R}}_h)}{\sum_{l=1}^K \pi_l f_l(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \widetilde{\mathbf{R}}_l)},$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. The Bayes classifier uses the MAP rule to assign $\boldsymbol{V}_i$ to the class with the highest posterior probability and can minimize the expected misclassification rate (Bouveyron, 2014; Fraley & Raftery, 2002). Specifically, the class assignment is determined by

$$\widetilde{Z}_i = \arg\max_h \frac{\pi_h f_h(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \widetilde{\mathbf{R}}_h)}{\sum_{l=1}^K \pi_l f_l(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \widetilde{\mathbf{R}}_l)}.$$

We define the probability of misclassification of a classifier C as $R(C) = P(C(\boldsymbol{V}_i) \neq Z_i)$. We have $\widetilde{Z}_i = \arg\min_{C=1,\ldots,K} R(C)$; a detailed derivation of this result is provided in Section 1 of the Supplementary Material. We propose to assign $\boldsymbol{V}_i$ to the class with the highest posterior probability under the MAP framework.

In order to adopt the MAP rule, we need to estimate $(\boldsymbol{\mu}, \boldsymbol{A})$ and the prior probabilities $\{\pi_h\}_{h=1}^K$, as well as specify the distribution of $\boldsymbol{V}_i$ given $Z_i = h$. However, the exact distribution of $\boldsymbol{V}_i$ is unknown. To address this, we employ a pseudo-likelihood approach, as proposed in L. Xu et al. (2012) and H. Ma and Jiang (2023). For $i = 1, \ldots, n$, given $Z_i = h$, we specify the pseudo-density of $\boldsymbol{V}_i$ as

$$g_h(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \widetilde{\mathbf{R}}_h) \propto |\boldsymbol{A}^{\frac{1}{2}}\widetilde{\mathbf{R}}_h\boldsymbol{A}^{\frac{1}{2}}|^{-\frac{1}{2}} \exp\left\{-(\boldsymbol{V}_i - \boldsymbol{\mu})^\top (\boldsymbol{A}^{\frac{1}{2}}\widetilde{\mathbf{R}}_h\boldsymbol{A}^{\frac{1}{2}})^{-1}(\boldsymbol{V}_i - \boldsymbol{\mu})/2\right\}.$$

The mixture pseudo-density of $\boldsymbol{V}_i$ is then expressed as

$$g(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi}) = \sum_{h=1}^K \pi_h g_h(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \widetilde{\mathbf{R}}_h).$$

Denote the maximum pseudo-likelihood estimators of $\boldsymbol{\mu}$, $\boldsymbol{A}$ and $\boldsymbol{\pi}$ as

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\pi}}) = \arg\max_{\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi}} \sum_{i=1}^n \log g(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi}).$$

Assume $\boldsymbol{V}_i$ has a distribution function with true density $q(\cdot)$. Let $\mathrm{KL}(g, q) = \int g \log(g/q)$ be Kullback-Leibler distance between g and q. Suppose $(\boldsymbol{\mu}^*, \boldsymbol{A}^*, \boldsymbol{\pi}^*)$ be the unique minimum of $\mathrm{KL}(g, q)$. By Lemma A.1 in Appendix, we have that $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{\pi}}$ converge to $\boldsymbol{\mu}^*$, $\boldsymbol{A}^*$

and $\boldsymbol{\pi}^*$, respectively, almost surely (a.s.), as $n \to \infty$. If $q(\boldsymbol{v}) = \sum_{h=1}^{K} \pi_{h0} g_h(\boldsymbol{v}; \boldsymbol{\mu}_0, \boldsymbol{A}_0, \widetilde{\mathbf{R}}_h)$, we have $\boldsymbol{\mu}^* = \boldsymbol{\mu}_0$, $\boldsymbol{A}^* = \boldsymbol{A}_0$ and $\boldsymbol{\pi}^* = \boldsymbol{\pi}_0$ with $\boldsymbol{\pi}_0 = (\pi_{10}, \ldots, \pi_{K0})$. Then, $\widehat{\boldsymbol{\mu}} \xrightarrow{\text{a.s.}} \boldsymbol{\mu}_0$, $\widehat{\boldsymbol{A}} \xrightarrow{\text{a.s.}} \boldsymbol{A}_0$ and $\widehat{\boldsymbol{\pi}} \xrightarrow{\text{a.s.}} \boldsymbol{\pi}_0$, as $n \to \infty$. That is, provided that we accurately define the density of $\boldsymbol{V}_i$, the estimators $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{\pi}}$ exhibit consistency.

Let $s_{ih} = 1$ if $\boldsymbol{V}_i$ belongs to the hth class (i.e., $Z_i = h$) and $s_{ih} = 0$ otherwise. Define $\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)$ and $\boldsymbol{S} = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n)$ with $\boldsymbol{S}_i = (s_{i1}, \ldots, s_{iK})^\top$. Let $b_1(\boldsymbol{s}; \boldsymbol{\pi})$ and $b_2(\boldsymbol{v}; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi} \mid \boldsymbol{S}_i)$ be the probability densities of $\boldsymbol{S}_i$ and $\boldsymbol{V}_i \mid \boldsymbol{S}_i$, respectively, for $i = 1, \ldots, n$. The class memberships $\boldsymbol{S}$ are unobserved, and we treat $(\boldsymbol{V}, \boldsymbol{S})$ as the complete data following the principle of the EM algorithm. The log-pseudo-likelihood for the complete data can be written as

$$
\begin{aligned}
l(\boldsymbol{V}, \boldsymbol{S}; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi}) &= \log \left\{ \prod_{i=1}^{n} b_1(\boldsymbol{S}_i; \boldsymbol{\pi}) b_2(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi} \mid \boldsymbol{S}_i) \right\} \\
&= \sum_{h=1}^{K} \sum_{i=1}^{n} s_{ih} \left\{ \log \pi_h + \log g_h(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \widetilde{\mathbf{R}}_h) \right\}.
\end{aligned}
$$

Note that the class-specific correlation matrices are obtained from the training data and are treated as known parameters in the likelihood.

We then employ the EM algorithm to obtain the estimators of $(\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi})$ using the testing data. In the E-step of the $(k+1)$th iteration of the algorithm, denote the current parameter estimators as $(\boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)})$. We compute the expectation of $l(\boldsymbol{V}, \boldsymbol{S}; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi})$ conditional on the testing data $\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)$ and $(\boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)})$, denoted as $Q(\boldsymbol{V}; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi} \mid \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)})$:

$$
\begin{aligned}
&Q(\boldsymbol{V}; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi} \mid \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)}) \\
&\quad = E_{\boldsymbol{S} \mid \boldsymbol{V}; \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)}} \{ l(\boldsymbol{V}, \boldsymbol{S}; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi}) \} \\
&\quad = \sum_{h=1}^{K} \sum_{i=1}^{n} \tau_h(\boldsymbol{V}_i; \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)}) \left\{ \log \pi_h + \log g_h(\boldsymbol{V}_i; \boldsymbol{\mu}, \boldsymbol{A}, \widetilde{\mathbf{R}}_h) \right\}, \qquad (1)
\end{aligned}
$$

where

$$
\begin{aligned}
\tau_h(\boldsymbol{V}_i; \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)}) &= \mathbf{E} \left\{ s_{ih} \mid \boldsymbol{V}_i; \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)} \right\} \\
&= \frac{\pi_h^{(k)} g_h(\boldsymbol{V}_i; \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \widetilde{\mathbf{R}}_h)}{\sum_{l=1}^{K} \pi_l^{(k)} g_l(\boldsymbol{V}_i; \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \widetilde{\mathbf{R}}_l)},
\end{aligned}
$$

which is the estimated posterior probability of observation $\boldsymbol{V}_i$ belonging to the hth class in the kth iteration. In the M-step, we maximize $Q(\boldsymbol{V}; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi} \mid \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)})$ to update estimators of $\boldsymbol{\mu}$, $\boldsymbol{A}$ and $\boldsymbol{\pi}$. Specifically,

$$
\begin{aligned}
\boldsymbol{\mu}^{(k+1)} = &\left[\sum_{h=1}^{K} \sum_{i=1}^{n} \tau_h(\boldsymbol{V}_i; \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)}) \left\{ (\boldsymbol{A}^{(k)})^{\frac{1}{2}} \widetilde{\mathbf{R}}_h (\boldsymbol{A}^{(k)})^{\frac{1}{2}} \right\}^{-1} \right]^{-1} \\
&\times \left[\sum_{h=1}^{K} \sum_{i=1}^{n} \tau_h(\boldsymbol{V}_i; \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)}) \left\{ (\boldsymbol{A}^{(k)})^{\frac{1}{2}} \widetilde{\mathbf{R}}_h (\boldsymbol{A}^{(k)})^{\frac{1}{2}} \right\}^{-1} \boldsymbol{V}_i \right], \qquad (2)
\end{aligned}
$$

$$\pi_h^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_h(\boldsymbol{V}_i; \boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)})}{n}, \tag{3}$$

and $\boldsymbol{A}^{(k+1)}$ can be obtained using the R software function `nleqslv(·)`. A detailed derivation of these update formulas is provided in Section 2 of the Supplementary Material.

We repeat the E-step and M-step until convergence is achieved and obtain the estimators $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{\pi}}$. For $h = 1, \ldots, K$, the pseudo-posterior probability estimator of $\boldsymbol{V}_i$ belonging to the hth class is

$$\tau_h(\boldsymbol{V}_i; \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\pi}}) = \frac{\hat{\pi}_h g_h(\boldsymbol{V}_i; \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{A}}, \widetilde{\mathbf{R}}_h)}{\sum_{l=1}^{K} \hat{\pi}_l g_l(\boldsymbol{V}_i; \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{A}}, \widetilde{\mathbf{R}}_l)}. \tag{4}$$

We assign $\boldsymbol{V}_i$ to the class with the highest pseudo-posterior probability estimator. The summary of the proposed CorrDA is provided in Algorithm 1. When the mean vector and variances do not provide informative cues, instead of relying on the proposal of Celeux and Govaert (1995) to use mixture models for classification based on mean vectors and covariance matrices, we suggest utilizing the correlation matrix as the basis for classification. As mentioned in Section 1, the mean information from China's epidemic data is not generalizable to other regions, and hence we alternatively propose to integrate the correlation matrices from the training dataset (i.e., China's COVID-19 data) to build a classification rule for the testing dataset (i.e., US COVID-19 data) in the real data analysis in Section 3.

An extension of CorrDA to testing data with class-specific mean vectors is provided in Section 3 of the Supplementary Material.

Remark 2.1: In both real data analysis and simulation studies, we set the tolerance as 10^{-5} in Algorithm 1. We initialize $\boldsymbol{\pi}^{(0)} = (0.5, 0.5)$, $\boldsymbol{\mu}^{(0)} = \mathbf{0}_p$, and $\boldsymbol{A}^{(0)} = \boldsymbol{I}_p$; a sensitivity analysis of these initial values is provided in Section 4.2 of the Supplementary Material.

Algorithm 1 CorrDA Algorithm

1: Initialize $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{A}^{(0)}$, and $\boldsymbol{\pi}^{(0)}$.
2: E-step: calculate the expectation of the complete data log-likelihood $l(\boldsymbol{V}, \boldsymbol{S}; \boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\pi})$ conditional on $\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)$ and $(\boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)})$ to obtain (1).
3: M-step: maximize (1) with respect to $\boldsymbol{\mu}$, $\boldsymbol{A}$, and $\boldsymbol{\pi}$, and update $\boldsymbol{\mu}^{(k+1)}$, $\boldsymbol{A}^{(k+1)}$, and $\boldsymbol{\pi}^{(k+1)}$ using (2), R function nleqslv(·), and (3), respectively.
4: Repeat Steps 2 and 3 until the difference between $(\boldsymbol{\mu}^{(k)}, \boldsymbol{A}^{(k)}, \boldsymbol{\pi}^{(k)})$ and $(\boldsymbol{\mu}^{(k+1)}, \boldsymbol{A}^{(k+1)}, \boldsymbol{\pi}^{(k+1)})$ reaches the pre-specified tolerance, yielding the estimators $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{\pi}}$.
5: Compute (4) for $i = 1, \ldots, n$ and $h = 1, \ldots, K$.
6: $\boldsymbol{V}_i$ is assigned to class with the highest pseudo-posterior probability estimator, for $i = 1, \ldots, n$.

3. Application to the COVID-19 pandemic

3.1. Correlation matrices from the training data

The training data used in our study are sourced from China's COVID-19 data. China's COVID-19 data are selected as the training dataset since China went through the complete

process of the COVID-19 pandemic, starting from the outbreak and eventually achieving successful control. Specifically, there are roughly two phases during the epidemic in China after implementing the mandatory interventions by the government: a serious phase from January 24 to February 6, 2020 and an improved phase from February 7 to February 20, 2020 (S. Chen et al., 2020; Lauer et al., 2020; Tian et al., 2020). The daily numbers of newly confirmed COVID-19 cases were officially reported by the Health Committees of each province in China. We collect data from January 24, 2020, to February 20, 2020, for a total of 44 cities in China. These cities are chosen based on the criterion that each city had accumulated more than 100 confirmed COVID-19 cases during the specified period; for details see Figures 1 and 2 in Section 5 of Supplementary Material. Let $\tilde{x}_{i,j}$ be the number of daily confirmed cases in the jth city for the ith day, ranging from January 24, 2020 to February 20, 2020. We define $y_{i,j}$ as $y_{i,j} = \log(\tilde{x}_{i,j} + 1)$. This transformation stabilizes the variance, makes the distribution closer to Gaussian, and is a widely used preprocessing step for count data in epidemiological and genomic studies (Bosse et al., 2023; Delatola et al., 2017). Then $\boldsymbol{Y}_j^1 = (y_{1,j}, \ldots, y_{14,j})^\top$ and $\boldsymbol{Y}_j^2 = (y_{15,j}, \ldots, y_{28,j})^\top$ respectively represent the data from the jth city during the serious phase and the improved phase. The sample correlation matrices $\widehat{\boldsymbol{R}}_1$ and $\widehat{\boldsymbol{R}}_2$ based on $(\boldsymbol{Y}_1^1, \ldots, \boldsymbol{Y}_{44}^1)$ and $(\boldsymbol{Y}_1^2, \ldots, \boldsymbol{Y}_{44}^2)$, respectively, are plotted in Figure 1. The correlation matrices display noteworthy disparities between the two phases. In the serious phase, a progressively stronger trend is observed, whereas in the improved phase, a gradually milder trend is evident.

3.2. COVID-19 epidemic classification of counties in the United States

The Center for Systems Science and Engineering at Johns Hopkins University has collected daily information, including the number of confirmed COVID-19 cases, for counties across the United States. The dataset is available to the public through the GitHub repository: https://github.com/CSSEGISandData/COVID-19. Our analysis focuses on the 131 U.S. counties that recorded over 1000 cumulative confirmed cases between July 12, 2020, and July 25, 2020. Our goal is to determine the status of these counties, whether they are in the serious phase or the improved phase of the pandemic. Several public-health studies indicate that the U.S. COVID-19 epidemic in mid-2020 can likewise be broadly divided into a serious phase followed by an improved phase. According to Bergquist et al. (2020), Zhang and Warner (2020), and Truong and Truong (2021), the daily number of newly confirmed cases in the U.S. rose sharply and reached a peak around mid-July 2020, corresponding to a serious phase. As daily new cases subsequently began to decline from late July 2020 onward, the epidemic entered a sustained lower-incidence period, which we refer to as the improved phase. Moreover, the cross-date state ranking correlation matrix reported by Jalal et al. (2024) for the 48 continental U.S. states during mid-2020 displays an early diagonal block in which rank correlations between days become progressively stronger, followed by a later block in which these correlations gradually weaken, mirroring the phase-specific correlation patterns we observe for Chinese cities.

Upon comparing the data between China and the United States, we observe that the average number of daily confirmed cases in U.S. counties is significantly higher than that in Chinese cities. This raises concerns about the suitability of utilizing LDA or QDA to effectively integrate information from China's data for discriminant analysis. However, despite this limitation, the estimated correlation matrices $\widehat{\boldsymbol{R}}_1$ and $\widehat{\boldsymbol{R}}_2$ from China's COVID-19 data can effectively capture the distinctive pattern between the serious phase and the improved phase.

The close agreement between these correlation structures and those in U.S. data motivates the use of correlation-based features as a transferable representation of epidemic phase, even when average daily case counts differ markedly across populations.

Let x_{ij} denote the number of daily confirmed cases for the jth county on the ith day, starting from July 12, 2020, to July 25, 2020. We define v_{ij} as the log-transformed value of x_{ij} by adding 1, i.e., $v_{ij} = \log(x_{ij} + 1)$. The outcome is denoted as $\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n) = (v_{ij}) \in \mathbb{R}^{p \times n}$, where $p = 14$ represents the number of days, and $n = 131$ denotes the total number of counties. We utilize the proposed CorrDA method to determine the epidemic phase for each county within the designated period. Out of the 131 counties studied, the CorrDA classifies 79 counties as being in the improved phase and 52 counties as still being in the serious phase between July 12, 2020 and July 25, 2020. For a comprehensive list of the 79 counties identified in the improved phase and the 52 counties remaining in the serious phase, please refer to Section 6 of the Supplementary Material. In addition to CorrDA, we also utilize the LDA, QDA, Multilayer Perceptron (MLP) (Venables & Ripley, 2013), Random Forest (RF) (Breiman, 2001), and Support Vector Machine (SVM) (Cortes & Vapnik, 1995) for the same purpose.

In order to evaluate the performance of the six methods, we consider the effective reproduction number (R_t) (Felizola Diniz-Filho et al., 2020; Gostic et al., 2020). R_t is a measure of the average number of secondary infectious cases generated by a primary infectious case.

The website http://metrics.COVID19-analysis.orgprovides R_t values for cities and counties worldwide. If the R_t values for a particular county remain consistently below 1 throughout the studied period, we conclude that the county is in an improved phase based on R_t. Conversely, if the R_t values exceed 1, the county is in a serious phase. We treat the serious phase as the positive class and the improved phase as the negative class. To provide a comprehensive evaluation of performance, we consider the following metrics:

$$\text{Concordant rate} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \quad \text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}.$$

Here, TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Table 1 summarizes the classification performance of CorrDA and the five competing methods (LDA, QDA, MLP, RF, and SVM) on the United States data in terms of these metrics. For the MLP classifier, all counties are classified as the serious

Table 1. Classification performance of CorrDA, LDA, QDA, MLP, RF, and SVM on the United States data in terms of concordant rate, sensitivity, specificity, MCC, precision, and F1.

	CorrDA	LDA	QDA	MLP	RF	SVM
Concordant rate	0.779	0.496	0.603	0.466	0.374	0.466
Sensitivity	0.689	0.869	0.542	1.000	0.414	0.869
Specificity	0.857	0.171	0.266	0.000	0.328	0.114
MCC	0.556	0.056	−0.196	NaN	−0.258	−0.026
Precision	0.808	0.477	0.500	0.466	0.414	0.461
F1	0.743	0.616	0.520	0.635	0.414	0.602

phase (sensitivity = 1 and specificity = 0), which implies TN = FN = 0. In this degenerate case, the MCC formula reduces to 0/0, so the MCC value is undefined and is therefore reported as NaN. As shown in Table 1, our method achieves the highest concordant rate, specificity, MCC, precision, and F1, as well as nearly the highest sensitivity. These metrics demonstrate that CorrDA provides the most balanced and reliable classification performance across the serious and improved phases.

4. Simulation studies

We conduct simulation studies to evaluate the performance of the proposed CorrDA and then compare it with LDA, QDA, MLP, RF, and SVM. We conduct 1000 random replications for each simulation setting.

4.1. Study 1

In the testing set, the observations of the two classes are generated from p-dimensional Gaussian distributions. The observations of Class 1 follow a Gaussian distribution $N(\mathbf{1}_p, \boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\Sigma}_1\boldsymbol{A}^{\frac{1}{2}})$ with a proportion of π_1, while those of Class 2 follow a Gaussian distribution $N(\mathbf{1}_p, \boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\Sigma}_2\boldsymbol{A}^{\frac{1}{2}})$ with a proportion of $\pi_2 = 1 - \pi_1$, where $\mathbf{1}_p$ is a p-dimensional vector with all elements 1 and $\boldsymbol{A} = 1.5\boldsymbol{I}_p$ with $\boldsymbol{I}_p$ being a $p \times p$ unit matrix. The covariance matrix $\boldsymbol{\Sigma}_1$ has an autoregressive AR(1) structure characterized by a parameter α_1, i.e., $(\boldsymbol{\Sigma}_1)_{ij} = \alpha_1^{|i-j|}$ and the covariance matrix $\boldsymbol{\Sigma}_2$ has an exchangeable structure characterized by a parameter α_2, i.e., $(\boldsymbol{\Sigma}_2)_{ij} = 1_{\{i=j\}} + \alpha_2 1_{\{i\neq j\}}$, where $1_{\{\cdot\}}$ is an indicator function. The values being considered for α_1 and α_2 are 0.3, 0.5, and 0.7. We consider various values of p: 14, 20, 30, and 40, as well as proportions π_1: 1/2, 1/3, and 2/3, and sample sizes n of either 90 or 150. For the AR(1) correlation structure, the sample size is $n\pi_1$, and for the exchangeable correlation structure, the sample size is $n - n\pi_1$. In this study, we characterize the distinctive pattern between the two classes by utilizing the true correlation matrices, rather than relying on estimated correlation matrices obtained from the training dataset.

The performance of the proposed CorrDA method is evaluated using the classification accuracy (CA), which represents the concordance rate between the true class memberships and the estimated class memberships. Tables 2 and 3 present the average and standard deviation (SD) of the CA values obtained by CorrDA across different scenarios for sample sizes of $n = 90$ and $n = 150$, respectively. CorrDA demonstrates satisfactory performance in accurately classifying observations across all scenarios considered. The average CA values consistently exceed 68%, and in some scenarios, reach as high as 99%. As expected, the average CA value tends to increase as the dimension grows, owing to the greater amount of information provided by the correlation matrices with larger dimensions. Additionally, when there is a larger distinction between the two correlation matrices (indicated by greater differences between α_1 and α_2 values), CorrDA exhibits better performance. On the other hand, the relative size of the two classes, π_1/π_2, has a minor impact on the performance of CorrDA.

4.2. Study 2

This study is identical to Study 1, except for the covariance matrix $\boldsymbol{\Sigma}_2$, which has an independence structure instead of an exchangeable structure, i.e., $\boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$.

Table 2. The average and standard deviation (in parentheses) of classification accuracy by CorrDA over 1000 simulation replications in Study 1 with $n = 90$.

(α_1, α_2)	$p = 14$	$p = 20$	$p = 30$	$p = 40$
$\pi_1 : \pi_2 = 1 : 1$				
(0.3, 0.3)	0.712(0.050)	0.774(0.040)	0.836(0.040)	0.875(0.034)
(0.3, 0.5)	0.809(0.043)	0.866(0.029)	0.918(0.029)	0.947(0.024)
(0.3, 0.7)	0.921(0.030)	0.958(0.014)	0.984(0.014)	0.993(0.009)
(0.5, 0.3)	0.773(0.044)	0.839(0.033)	0.900(0.033)	0.938(0.025)
(0.5, 0.5)	0.811(0.041)	0.872(0.027)	0.930(0.027)	0.959(0.022)
(0.5, 0.7)	0.899(0.033)	0.944(0.015)	0.979(0.015)	0.991(0.010)
(0.7, 0.3)	0.875(0.035)	0.928(0.018)	0.969(0.018)	0.986(0.013)
(0.7, 0.5)	0.850(0.038)	0.912(0.021)	0.960(0.021)	0.982(0.015)
(0.7, 0.7)	0.875(0.034)	0.931(0.017)	0.974(0.017)	0.990(0.011)
$\pi_1 : \pi_2 = 1 : 2$				
(0.3, 0.3)	0.689(0.057)	0.753(0.049)	0.820(0.049)	0.862(0.043)
(0.3, 0.5)	0.796(0.054)	0.855(0.036)	0.913(0.036)	0.943(0.029)
(0.3, 0.7)	0.915(0.036)	0.956(0.017)	0.983(0.017)	0.993(0.010)
(0.5, 0.3)	0.749(0.058)	0.825(0.037)	0.892(0.037)	0.932(0.028)
(0.5, 0.5)	0.798(0.051)	0.864(0.034)	0.925(0.034)	0.958(0.024)
(0.5, 0.7)	0.895(0.038)	0.943(0.018)	0.977(0.018)	0.990(0.013)
(0.7, 0.3)	0.863(0.045)	0.919(0.022)	0.966(0.022)	0.985(0.015)
(0.7, 0.5)	0.840(0.046)	0.904(0.024)	0.957(0.024)	0.981(0.017)
(0.7, 0.7)	0.872(0.041)	0.931(0.021)	0.972(0.021)	0.989(0.012)
$\pi_1 : \pi_2 = 2 : 1$				
(0.3, 0.3)	0.682(0.063)	0.748(0.049)	0.821(0.049)	0.863(0.044)
(0.3, 0.5)	0.792(0.057)	0.857(0.037)	0.911(0.037)	0.945(0.028)
(0.3, 0.7)	0.913(0.037)	0.955(0.017)	0.981(0.017)	0.993(0.010)
(0.5, 0.3)	0.751(0.054)	0.821(0.038)	0.890(0.038)	0.931(0.030)
(0.5, 0.5)	0.789(0.054)	0.857(0.033)	0.924(0.033)	0.955(0.025)
(0.5, 0.7)	0.889(0.041)	0.938(0.019)	0.976(0.019)	0.990(0.013)
(0.7, 0.3)	0.864(0.042)	0.921(0.022)	0.967(0.022)	0.986(0.015)
(0.7, 0.5)	0.832(0.046)	0.902(0.024)	0.957(0.024)	0.980(0.017)
(0.7, 0.7)	0.857(0.044)	0.922(0.022)	0.969(0.022)	0.986(0.015)

4.3. Study 3

This study is similar to Study 2, with the only difference being that the covariance matrix $\mathbf{\Sigma}_1$ has an exchangeable structure instead of an autoregressive structure. Studies 2 and 3 yield comparably favourable results to those obtained in Study 1. Detailed results will be presented in the Supplementary Material.

4.4. Study 4

The aim of this study is to assess the robustness of the proposed method to the data distribution. This study replicates Study 1 with one key difference: the observations are generated from a sub-Gaussian distribution (Janková & van de Geer, 2017; Maurer & Pontil, 2021). To generate the data, we use a uniform distribution on the interval $[-\sqrt{3}, \sqrt{3}]$ to generate independent entries for a p-dimensional vector $\boldsymbol{U}$. The vector is then transformed to $\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{\Sigma}^{1/2}\boldsymbol{U}$. Note that the expectation of $\boldsymbol{X}$ is $\boldsymbol{\mu}$ and the covariance matrix of $\boldsymbol{X}$ is $\mathbf{\Sigma}$. It follows by Hoeffding's inequality that $\boldsymbol{X}$ defined as described above is sub-Gaussian. The results are summarized in Tables 4 and 5, demonstrating similarly promising findings as those obtained in Study 1. This illustrates the robustness of our proposed method.

We perform two additional simulation studies. The first study is similar to setting described in the previous paragraph, with the only difference being that the covariance matrix $\mathbf{\Sigma}_2$ has an independence structure instead of an exchangeable structure. The second one is

Table 3. The average and standard deviation (in parentheses) of classification accuracy values by CorrDA over 1000 simulation replications in Study 1 with $n = 150$.

(α_1, α_2)	$p = 14$	$p = 20$	$p = 30$	$p = 40$
$\pi_1 : \pi_2 = 1 : 1$				
(0.3, 0.3)	0.721(0.037)	0.777(0.029)	0.838(0.029)	0.878(0.027)
(0.3, 0.5)	0.816(0.032)	0.868(0.023)	0.922(0.023)	0.950(0.018)
(0.3, 0.7)	0.926(0.023)	0.961(0.010)	0.986(0.010)	0.994(0.006)
(0.5, 0.3)	0.778(0.034)	0.840(0.023)	0.905(0.023)	0.939(0.020)
(0.5, 0.5)	0.815(0.033)	0.876(0.020)	0.933(0.020)	0.960(0.016)
(0.5, 0.7)	0.904(0.024)	0.950(0.012)	0.980(0.012)	0.992(0.007)
(0.7, 0.3)	0.878(0.026)	0.929(0.014)	0.970(0.014)	0.987(0.009)
(0.7, 0.5)	0.853(0.029)	0.912(0.016)	0.962(0.016)	0.983(0.011)
(0.7, 0.7)	0.878(0.027)	0.935(0.012)	0.975(0.012)	0.990(0.008)
$\pi_1 : \pi_2 = 1 : 2$				
(0.3, 0.3)	0.695(0.046)	0.759(0.035)	0.826(0.035)	0.867(0.032)
(0.3, 0.5)	0.801(0.038)	0.858(0.027)	0.916(0.027)	0.946(0.022)
(0.3, 0.7)	0.921(0.026)	0.959(0.011)	0.986(0.011)	0.994(0.007)
(0.5, 0.3)	0.758(0.043)	0.825(0.028)	0.897(0.028)	0.935(0.023)
(0.5, 0.5)	0.801(0.038)	0.867(0.025)	0.928(0.025)	0.958(0.019)
(0.5, 0.7)	0.898(0.030)	0.947(0.013)	0.980(0.013)	0.991(0.009)
(0.7, 0.3)	0.867(0.035)	0.923(0.016)	0.968(0.016)	0.986(0.011)
(0.7, 0.5)	0.840(0.036)	0.906(0.019)	0.959(0.019)	0.981(0.013)
(0.7, 0.7)	0.871(0.032)	0.932(0.015)	0.975(0.015)	0.990(0.009)
$\pi_1 : \pi_2 = 2 : 1$				
(0.3, 0.3)	0.685(0.052)	0.751(0.037)	0.825(0.037)	0.867(0.033)
(0.3, 0.5)	0.798(0.043)	0.860(0.029)	0.915(0.029)	0.947(0.022)
(0.3, 0.7)	0.916(0.027)	0.958(0.012)	0.985(0.012)	0.994(0.007)
(0.5, 0.3)	0.753(0.044)	0.827(0.029)	0.897(0.029)	0.933(0.023)
(0.5, 0.5)	0.793(0.041)	0.865(0.025)	0.925(0.025)	0.958(0.018)
(0.5, 0.7)	0.893(0.033)	0.944(0.014)	0.978(0.014)	0.991(0.009)
(0.7, 0.3)	0.870(0.033)	0.924(0.016)	0.969(0.016)	0.986(0.011)
(0.7, 0.5)	0.837(0.036)	0.904(0.019)	0.958(0.019)	0.981(0.012)
(0.7, 0.7)	0.864(0.032)	0.927(0.015)	0.973(0.015)	0.989(0.009)

similar to the first one, except that the covariance matrix $\mathbf{\Sigma}_1$ is structured as exchangeable instead of autoregressive. The two studies show similarly promising results as obtained in Study 1. Detailed results will be provided in the Supplementary Material.

Additional simulation results for modified versions of Studies 1–4 with explicit training data, where CorrDA is compared with LDA, QDA, RF, SVM, and MLP, are presented in Section 4.3 of the Supplementary Material. Further multi-class simulation results under both Gaussian and sub-Gaussian designs are reported in Section 4.4 of the Supplementary Material.

4.5. Study 5

The objective of this study is to compare the performance of the proposed CorrDA with that of LDA, QDA, MLP, RF, and SVM. The training set consists of 100 observations generated from a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{R}_1$ for Class 1, and 100 observations generated from a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{R}_2$ for Class 2. In this context, we set $\boldsymbol{R}_1$ as $\widehat{\boldsymbol{R}}_1$ and $\boldsymbol{R}_2$ as $\widehat{\boldsymbol{R}}_2$, with $\widehat{\boldsymbol{R}}_1$ and $\widehat{\boldsymbol{R}}_2$ representing the two correlation matrices of the two phases in China's COVID-19 data in Section 3.1. We calculate the sample correlation matrices $\widetilde{\boldsymbol{R}}_1$ and $\widetilde{\boldsymbol{R}}_2$ using the simulated training dataset and utilize this information in the discriminant analysis of the testing dataset.

Table 4. The average and standard deviation (in parentheses) of classification accuracy values by CorrDA over 1000 simulation replications in Study 4 with $n = 90$.

(α_1, α_2)	$p = 14$	$p = 20$	$p = 30$	$p = 40$
$\pi_1 : \pi_2 = 1 : 1$				
(0.3, 0.3)	0.709(0.052)	0.773(0.040)	0.841(0.040)	0.882(0.034)
(0.3, 0.5)	0.829(0.047)	0.892(0.025)	0.947(0.025)	0.970(0.019)
(0.3, 0.7)	0.957(0.024)	0.984(0.007)	0.996(0.007)	0.999(0.003)
(0.5, 0.3)	0.768(0.048)	0.831(0.033)	0.897(0.033)	0.934(0.026)
(0.5, 0.5)	0.811(0.042)	0.876(0.027)	0.934(0.027)	0.964(0.020)
(0.5, 0.7)	0.926(0.030)	0.968(0.010)	0.990(0.010)	0.997(0.006)
(0.7, 0.3)	0.882(0.036)	0.934(0.017)	0.974(0.017)	0.988(0.012)
(0.7, 0.5)	0.845(0.038)	0.906(0.021)	0.958(0.021)	0.981(0.014)
(0.7, 0.7)	0.877(0.036)	0.937(0.016)	0.976(0.016)	0.991(0.010)
$\pi_1 : \pi_2 = 1 : 2$				
(0.3, 0.3)	0.698(0.060)	0.764(0.047)	0.832(0.047)	0.877(0.041)
(0.3, 0.5)	0.832(0.050)	0.887(0.030)	0.942(0.030)	0.967(0.023)
(0.3, 0.7)	0.957(0.028)	0.983(0.009)	0.996(0.009)	0.999(0.004)
(0.5, 0.3)	0.746(0.056)	0.819(0.039)	0.890(0.039)	0.929(0.031)
(0.5, 0.5)	0.804(0.048)	0.877(0.030)	0.934(0.030)	0.963(0.024)
(0.5, 0.7)	0.925(0.034)	0.966(0.013)	0.990(0.013)	0.996(0.007)
(0.7, 0.3)	0.854(0.046)	0.921(0.022)	0.968(0.022)	0.986(0.014)
(0.7, 0.5)	0.828(0.047)	0.900(0.025)	0.955(0.025)	0.980(0.017)
(0.7, 0.7)	0.876(0.040)	0.938(0.018)	0.977(0.018)	0.991(0.012)
$\pi_1 : \pi_2 = 2 : 1$				
(0.3, 0.3)	0.671(0.064)	0.736(0.050)	0.817(0.050)	0.864(0.045)
(0.3, 0.5)	0.777(0.070)	0.859(0.037)	0.929(0.037)	0.963(0.027)
(0.3, 0.7)	0.940(0.038)	0.980(0.008)	0.996(0.008)	0.999(0.004)
(0.5, 0.3)	0.754(0.056)	0.822(0.038)	0.892(0.038)	0.930(0.030)
(0.5, 0.5)	0.777(0.056)	0.856(0.033)	0.926(0.033)	0.959(0.023)
(0.5, 0.7)	0.903(0.046)	0.960(0.013)	0.989(0.013)	0.997(0.007)
(0.7, 0.3)	0.888(0.038)	0.937(0.019)	0.973(0.019)	0.989(0.012)
(0.7, 0.5)	0.842(0.044)	0.903(0.024)	0.956(0.024)	0.980(0.016)
(0.7, 0.7)	0.856(0.044)	0.924(0.020)	0.973(0.020)	0.989(0.013)

For the testing set, the data in Class 1 are generated from a multivariate normal distribution $N(\boldsymbol{\mu}_1, \boldsymbol{R}_1)$ with a proportion of π_1, and the data in Class 2 are generated from a multivariate normal distribution $N(\boldsymbol{\mu}_2, \boldsymbol{R}_2)$ with a proportion of $\pi_2 = 1 - \pi_1$. The mean vectors are considered in the following three scenarios:

(a) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 3.0, 3.1, 3.2, 3.3, 3.4)$;
(b) $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1, 5.2, 5.3, 5.4)$;
(c) $\boldsymbol{\mu}_1 = (3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 4.0, 4.1, 4.2, 4.3, 4.4)$,$\boldsymbol{\mu}_2 = (4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1, 5.2, 5.3, 5.4)$.

The proportion of Class 1 is π_1 that takes values from the set $\{1/2, 1/3, 2/3\}$. The sample size n is either 90 or 150, with $n\pi_1$ being the sample size for Class 1 and $n - n\pi_1$ being the sample size for Class 2.

Table 6 presents the average and standard deviation (SD) of classification accuracy (CA) values obtained by CorrDA, LDA, QDA, MLP, RF, and SVM under different settings. The proposed CorrDA demonstrates good performance with CA values ranging from 0.894 to 0.923 on average. In all scenarios, CorrDA consistently outperforms LDA, QDA, MLP, RF, and SVM. Since the training data for the two classes have the same mean, LDA achieves average CA values of approximately 50%.

Table 5. The average and standard deviation (in parentheses) of classification accuracy values by CorrDA over 1000 simulation replications in Study 4 with $n = 150$.

(α_1, α_2)	$p = 14$	$p = 20$	$p = 30$	$p = 40$
$\pi_1 : \pi_2 = 1 : 1$				
(0.3, 0.3)	0.718(0.040)	0.779(0.031)	0.845(0.031)	0.885(0.025)
(0.3, 0.5)	0.837(0.035)	0.898(0.019)	0.950(0.019)	0.973(0.014)
(0.3, 0.7)	0.963(0.017)	0.987(0.004)	0.997(0.004)	0.999(0.002)
(0.5, 0.3)	0.770(0.035)	0.836(0.025)	0.901(0.025)	0.937(0.020)
(0.5, 0.5)	0.815(0.032)	0.880(0.020)	0.938(0.020)	0.966(0.015)
(0.5, 0.7)	0.931(0.023)	0.970(0.008)	0.992(0.008)	0.998(0.004)
(0.7, 0.3)	0.887(0.027)	0.937(0.013)	0.974(0.013)	0.989(0.009)
(0.7, 0.5)	0.848(0.029)	0.911(0.016)	0.960(0.016)	0.982(0.011)
(0.7, 0.7)	0.881(0.027)	0.939(0.012)	0.979(0.012)	0.992(0.007)
$\pi_1 : \pi_2 = 1 : 2$				
(0.3, 0.3)	0.711(0.046)	0.770(0.035)	0.838(0.035)	0.881(0.032)
(0.3, 0.5)	0.835(0.038)	0.895(0.022)	0.947(0.022)	0.971(0.017)
(0.3, 0.7)	0.961(0.019)	0.985(0.006)	0.997(0.006)	0.999(0.003)
(0.5, 0.3)	0.752(0.042)	0.824(0.030)	0.893(0.030)	0.934(0.024)
(0.5, 0.5)	0.811(0.036)	0.876(0.023)	0.936(0.023)	0.965(0.017)
(0.5, 0.7)	0.929(0.027)	0.968(0.010)	0.990(0.010)	0.997(0.005)
(0.7, 0.3)	0.859(0.036)	0.922(0.016)	0.969(0.016)	0.987(0.011)
(0.7, 0.5)	0.833(0.037)	0.902(0.019)	0.957(0.019)	0.981(0.013)
(0.7, 0.7)	0.881(0.031)	0.939(0.014)	0.978(0.014)	0.992(0.009)
$\pi_1 : \pi_2 = 2 : 1$				
(0.3, 0.3)	0.675(0.052)	0.742(0.038)	0.819(0.038)	0.866(0.034)
(0.3, 0.5)	0.785(0.056)	0.866(0.028)	0.935(0.028)	0.966(0.019)
(0.3, 0.7)	0.945(0.027)	0.983(0.005)	0.997(0.005)	0.999(0.002)
(0.5, 0.3)	0.758(0.040)	0.825(0.029)	0.894(0.029)	0.931(0.023)
(0.5, 0.5)	0.782(0.044)	0.858(0.025)	0.928(0.025)	0.961(0.018)
(0.5, 0.7)	0.912(0.035)	0.965(0.009)	0.991(0.009)	0.998(0.004)
(0.7, 0.3)	0.890(0.030)	0.938(0.015)	0.975(0.015)	0.989(0.009)
(0.7, 0.5)	0.845(0.032)	0.907(0.017)	0.958(0.017)	0.981(0.012)
(0.7, 0.7)	0.860(0.036)	0.927(0.015)	0.975(0.015)	0.991(0.009)

4.6. Study 6

This study is a replication of Study 5, with one key difference: the observations in the testing dataset are generated from the sub-Gaussian distributions described in Study 4. The results of this study are presented in the Supplementary Material, and they demonstrate a similar good performance as observed in Study 5.

A detailed investigation of the robustness of CorrDA to estimation errors in the class-specific correlation matrices, including small training-sample and noisy-data settings, is provided in Section 4.5 of the Supplementary Material.

5. Discussion

In this paper, we introduce a novel discriminant analysis method that utilizes the Bayes classifier and mixture models. Rather than relying on mean vectors of the training set that may not be directly comparable to the testing set, we leverage information from the distinguishable correlation matrices among classes in the training set. This enhances the discriminant analysis performance on the testing samples when the mean vectors of the training data are not informative for comparative inference for the testing data. While our primary focus is on the application to the COVID-19 pandemic, our proposed method can be easily extended to address other epidemic classification problems.

Table 6. The average and standard deviation (in parentheses) of classification accuracy values by CorrDA, LDA, QDA, MLP, RF, and SVM over 1000 simulation replications in Study 5.

	CorrDA	LDA	QDA	MLP	RF	SVM
$n = 90$, Scenario (a)						
$\pi_1 : \pi_2 = 1 : 1$	0.912(0.032)	0.499(0.048)	0.741(0.067)	0.512(0.070)	0.509(0.040)	0.489(0.039)
$\pi_1 : \pi_2 = 1 : 2$	0.918(0.033)	0.500(0.048)	0.742(0.068)	0.513(0.072)	0.508(0.042)	0.490(0.039)
$\pi_1 : \pi_2 = 2 : 1$	0.894(0.039)	0.502(0.050)	0.740(0.071)	0.511(0.069)	0.509(0.044)	0.488(0.040)
$n = 90$, Scenario (b)						
$\pi_1 : \pi_2 = 1 : 1$	0.912(0.032)	0.498(0.041)	0.738(0.072)	0.505(0.063)	0.500(0.009)	0.500(0.008)
$\pi_1 : \pi_2 = 1 : 2$	0.918(0.033)	0.499(0.042)	0.741(0.072)	0.507(0.068)	0.500(0.010)	0.500(0.008)
$\pi_1 : \pi_2 = 2 : 1$	0.897(0.040)	0.499(0.041)	0.738(0.071)	0.505(0.065)	0.500(0.011)	0.500(0.009)
$n = 90$, Scenario (c)						
$\pi_1 : \pi_2 = 1 : 1$	0.916(0.026)	0.504(0.041)	0.737(0.063)	0.506(0.076)	0.502(0.043)	0.476(0.034)
$\pi_1 : \pi_2 = 1 : 2$	0.908(0.037)	0.500(0.054)	0.775(0.080)	0.505(0.079)	0.502(0.044)	0.476(0.035)
$\pi_1 : \pi_2 = 2 : 1$	0.911(0.037)	0.502(0.054)	0.779(0.076)	0.505(0.074)	0.502(0.042)	0.475(0.034)
$n = 150$, Scenario (a)						
$\pi_1 : \pi_2 = 1 : 1$	0.917(0.024)	0.502(0.040)	0.741(0.062)	0.513(0.065)	0.507(0.032)	0.489(0.031)
$\pi_1 : \pi_2 = 1 : 2$	0.923(0.024)	0.501(0.042)	0.738(0.062)	0.510(0.068)	0.507(0.036)	0.490(0.034)
$\pi_1 : \pi_2 = 2 : 1$	0.904(0.032)	0.499(0.042)	0.737(0.065)	0.512(0.067)	0.509(0.036)	0.489(0.033)
$n = 150$, Scenario (b)						
$\pi_1 : \pi_2 = 1 : 1$	0.917(0.025)	0.501(0.037)	0.734(0.065)	0.507(0.061)	0.500(0.008)	0.500(0.007)
$\pi_1 : \pi_2 = 1 : 2$	0.923(0.025)	0.503(0.037)	0.737(0.067)	0.506(0.063)	0.500(0.008)	0.500(0.008)
$\pi_1 : \pi_2 = 2 : 1$	0.904(0.031)	0.501(0.036)	0.739(0.066)	0.506(0.062)	0.500(0.008)	0.500(0.007)
$n = 150$, Scenario (c)						
$\pi_1 : \pi_2 = 1 : 1$	0.920(0.024)	0.501(0.050)	0.772(0.075)	0.506(0.072)	0.502(0.040)	0.476(0.032)
$\pi_1 : \pi_2 = 1 : 2$	0.917(0.027)	0.501(0.049)	0.772(0.076)	0.505(0.074)	0.502(0.041)	0.476(0.033)
$\pi_1 : \pi_2 = 2 : 1$	0.915(0.028)	0.500(0.049)	0.770(0.076)	0.509(0.072)	0.501(0.041)	0.476(0.031)

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was partially supported by the National Natural Science Foundation of China (12271167, 72331005, 12531013 and 12371272), Natural Science Foundation of Shanghai (24ZR1420400) and the scientific research foundation for high - level talents of Zhoukou Normal University (ZKNUC2021005).

Data availability statement

The data that support the findings of this study are openly available through the GitHub repository: https://github.com/yanszx/COVID-19-Dataand https://github.com/CSSEGISandData/COVID-19.

References

Aerts, S., & Wilms, I. (2017). Cellwise robust regularized discriminant analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*(6), 436–447. https://doi.org/10.1002/sam.2017.10.issue-6

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley.

Bensmail, H., & Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, *91*(436), 1743–1748. https://doi.org/10.1080/01621459.1996.10476746

Bergquist, S., Otten, T., & Sarich, N. (2020). COVID-19 pandemic in the United States. *Health Policy and Technology*, *9*(4), 623–638. https://doi.org/10.1016/j.hlpt.2020.08.007

Bickel, P. J., & Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, *10*(6), 989–1010. https://doi.org/10.3150/bj/1106314847

Bosse, N. I., Abbott, S., Cori, A., Van Leeuwen, E., Bracher, J., & Funk, S. (2023). Scoring epidemiological forecasts on transformed scales. *PLoS Computational Biology*, *19*(8), Article e1011393. https://doi.org/10.1371/journal.pcbi.1011393

Bouveyron, C. (2014). Adaptive mixture discriminant analysis for supervised learning with unobserved classes. *Journal of Classification*, *31*(1), 49–84. https://doi.org/10.1007/s00357-014-9147-x

Bouveyron, C., Girard, S., & Schmid, C. (2007). High-dimensional discriminant analysis. *Communications in Statistics: Theory and Methods*, *36*(14), 2607–2623. https://doi.org/10.1080/03610920701271095

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, *28*(5), 781–793. https://doi.org/10.1016/0031-3203(94)00125-6

Chen, M., Wu, Y., & Jin, B. (2023). Evaluation of the Canadian government policies on controlling the COVID-19 outbreaks. *Statistical Theory and Related Fields*, *7*(3), 223–234. https://doi.org/10.1080/24754269.2023.2201108

Chen, S., Yang, J., Yang, W., Wang, C., & Bärnighausen, T. (2020). COVID-19 control in China during mass population movements at New Year. *The Lancet*, *395*(10226), 764–766. https://doi.org/10.1016/S0140-6736(20)30421-9

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Delatola, E. I., Lebarbier, E., Mary-Huard, T., Radvanyi, F., Robin, S., & Wong, J. (2017). SegCorr a statistical procedure for the detection of genomic regions of correlated expression. *BMC Bioinformatics*, *18*(1), Article 333. https://doi.org/10.1186/s12859-017-1742-5

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Felizola Diniz-Filho, J. A., Jardim, L., Toscano, C. M., & Rangel, T. F. (2020). The effective reproductive number (R_t) of COVID-19 and its relationship with social distancing. *medRxiv*, 2020–07.

Flamary, R., Cuturi, M., Courty, N., & Rakotomamonjy, A. (2018). Wasserstein discriminant analysis. *Machine Learning*, *107*(12), 1923–1945. https://doi.org/10.1007/s10994-018-5717-1

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631. https://doi.org/10.1198/016214502760047131

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*(405), 165–175. https://doi.org/10.1080/01621459.1989.10478752

Fu, Y., Liu, Y., Wang, H.-H., & Wang, X. (2020). Empirical likelihood estimation in multivariate mixture models with repeated measurements. *Statistical Theory and Related Fields*, *4*(2), 152–160. https://doi.org/10.1080/24754269.2019.1630544

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., Hellewell, J., Meakin, S., Munday, J. D., Bosse, N. I., Sherrat, K., Thompson, R. N., White, L. F., Huisman, J. S., Scire, J., … Pitzer, V. E. (2020). Practical considerations for measuring the effective reproductive number, R_t. *PLoS Computational Biology*, *16*(12), Article e1008409. https://doi.org/10.1371/journal.pcbi.1008409

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Jalal, H., Lee, K., & Burke, D. S. (2024). Oscillating spatiotemporal patterns of COVID-19 in the United States. *Scientific Reports*, *14*(1), Article 21562. https://doi.org/10.1038/s41598-024-72517-6

Janková, J., & van de Geer, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test*, *26*(1), 143–162. https://doi.org/10.1007/s11749-016-0503-5

Jiang, B., Chen, Z., & Leng, C. (2020). Dynamic linear discriminant analysis in high dimensional space. *Bernoulli*, *26*(2), 1234–1268. https://doi.org/10.3150/19-BEJ1154

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, *172*(9), 577–582. https://doi.org/10.7326/M20-0504

Lei, Z., Liao, S., & Li, S. Z. (2009). Stepwise correlation metric based discriminant analysis and multi-probe images fusion for face recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops* (pp. 147–153). IEEE.

Liu, X., Peng, J., Wang, D., & Chen, H. (2025). Maximum-likelihood estimation of the Po-MDDRCINAR(p) model with analysis of a COVID-19 data. *Statistical Theory and Related Fields*, *9*(1), 34–58. https://doi.org/10.1080/24754269.2024.2412491

Ma, H., & Jiang, J. (2023). Pseudo-Bayesian classified mixed model prediction. *Journal of the American Statistical Association*, *118*(543), 1747–1759. https://doi.org/10.1080/01621459.2021.2008944

Ma, H., Qin, J., Chen, F., & Zhou, Y. (2023). A novel nonparametric mixture model for the detection pattern of COVID-19 on diamond princess cruise. *Statistical Theory and Related Fields*, *7*(1), 85–96. https://doi.org/10.1080/24754269.2022.2156743

Ma, Y., Lao, S., Takikawa, E., & Kawade, M. (2007). Discriminant analysis in correlation similarity measure space. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 577–584). ACM.

Maurer, A., & Pontil, M. (2021). Concentration inequalities under sub-Gaussian and sub-exponential conditions. In *Advances in Neural Information Processing Systems (NeurIPS)*(pp. 7588–7597). NeurIPS Foundation.

Mclachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, *6*(1), 355–378. https://doi.org/10.1146/statistics.2019.6.issue-1

Sohil, F., Sohali, M. U., & Shabbir, J. (2022). An introduction to statistical learning with applications in R. *Statistical Theory and Related Fields*, *6*(1), 87–87. https://doi.org/10.1080/24754269.2021.1980261

Tian, H., Liu, Y., Li, Y., Wu, C.-H., Chen, B., Kraemer, M. U. G., Li, B., Cai, J., Xu, B., Yang, Q., Wang, B., Yang, P., Cui, Y., Song, Y., Zheng, P., Wang, Q., Bjornstad, O. N., Yang, R., Grenfell, B. T., … Dye, C. (2020). An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science*, *368*(6491), 638–642. https://doi.org/10.1126/science.abb6105

Truong, D., & Truong, M. D. (2021). Projecting daily travel behavior by distance during the pandemic and the spread of COVID-19 infections–Are we in a closed loop scenario? *Transportation Research Interdisciplinary Perspectives*, *9*, Article 100283. https://doi.org/10.1016/j.trip.2020.100283

Venables, W. N., & Ripley, B. D. (2013). *Modern Applied Statistics with S*. Springer Science & Business Media.

Vovan, T. (2018). Some results of classification problem by Bayesian method and application in credit operation. *Statistical Theory and Related Fields*, *2*(2), 150–157. https://doi.org/10.1080/24754269.2018.1528420

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25. https://doi.org/10.2307/1912526

Witten, D. M., & Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(5), 753–772. https://doi.org/10.1111/j.1467-9868.2011.00783.x

Xu, J., & Tang, Y. (2021). An integrated epidemic modelling framework for the real-time forecast of COVID-19 outbreaks in current epicentres. *Statistical Theory and Related Fields*, *5*(3), 200–220. https://doi.org/10.1080/24754269.2021.1872131

Xu, L., Iosifidis, A., & Gabbouj, M. (2018). Weighted linear discriminant analysis based on class saliency information. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 2306–2310). IEEE.

Xu, L., Lin, N., Zhang, B., & Shi, N.-Z. (2012). A finite mixture model for working correlation matrices in generalized estimating equations. *Statistica Sinica*, *22*(2), 755–776. https://doi.org/10.5705/ss.2010.090

Zhang, X., & Warner, M. E. (2020). COVID-19 policy differences across US states: Shutdowns, reopening, and mask mandates. *International Journal of Environmental Research and Public Health*, *17*(24), Article 9520. https://doi.org/10.3390/ijerph17249520

Appendix. Lemma

Lemma A.1 (White, 1982)**:** *Let* $U_1, \ldots, U_N$ *be independently and identically distributed with joint distribution function G and density g.*

(i) *Suppose that a family of distribution functions* $F_\theta(u)$ *has densities* $f_\theta(u)$ *that are measurable in u for every* $\theta \in \Theta$*, and continuous in* θ *for every* $u \in \mathbf{\Omega}$*, with* Θ *a compact subset of space.*

(ii) $E(\log g(U_i))$ *exists and* $|\log f(u,\theta)| \leq M(u)$ *for all* θ *in* Θ*, where M is integrable with respect to G.*

Then, if $\mathrm{KL}(g;f,\theta)$ *has a unique minimum at* θ^* *in* Θ*, we have* $\hat{\theta}_N \to \theta^*$ *almost surely as* $N \to \infty$*, where* $\hat{\theta}_N = \arg\max_{\theta\in\Theta} \sum_{i=1}^{N} \log f(U_i,\theta)$.