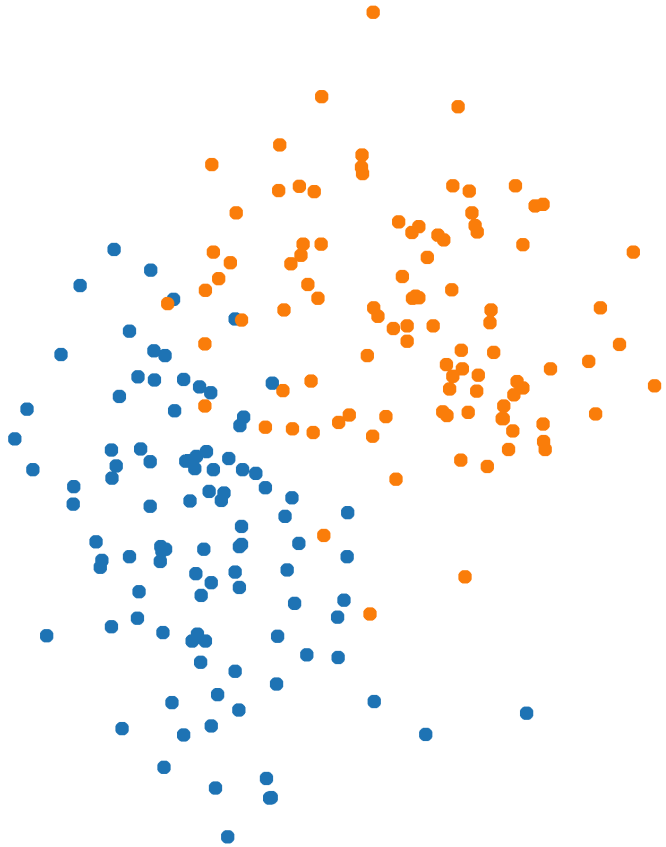
A Scalable Version of MADD for Classification Problems

Adrija Saha, Roll No: MD2203

Supervisor: Dr. Soham Sarkar (SMU, ISI Delhi)

23/05/2024



Last time...

Introduction

- We were interested in **Classification** problems.
- A popular choice is to consider **K-Nearest Neighbor Classifier** based on the Euclidean distance.
- But in High Dimensional problems, KNN based on the Euclidean distance performs poorly.
- If location difference is dominated by scale difference, NN classifier assigns all observations to the population with smaller dispersion!

Mean Absolute Difference of Distances

- MADD (Sarkar and Ghosh 2019) is a semi-metric based on available data cloud, defined as:

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{1}{n-2} \sum_{\mathbf{z} \in \mathrm{X} \setminus \{\mathbf{x}, \mathbf{y}\}} \left| \|\mathbf{x} - \mathbf{z}\| - \|\mathbf{y} - \mathbf{z}\| \right|$$

- Roy et. al. 2022 used MADD for high dimension, low sample size classification problems.
- Computation of MADD between two points requires $\mathrm{O}(nd)$ operations.
- Complexity becomes $\mathrm{O}(n^2 d)$ for classifying a single observation $\implies$ Quadratic in n.

High Dimensional Behavior of MADD

- If $\mathbf{X} \sim F_1, \mathbf{Y} \sim F_2$ are two independent observations, then under certain conditions

$$d^{-1/2}||\mathbf{X} - \mathbf{Y}|| \xrightarrow{P} \sqrt{\nu_{12}^2 + \sigma_1^2 + \sigma_2^2} \text{ as } d \to \infty$$

- Let us look at the expression,

$$\rho_0(\mathbf{X}, \mathbf{Y}) = d^{-1/2} \left[\frac{1}{n-2} \sum_{\mathbf{Z} \in \mathcal{X} \setminus \{\mathbf{X}, \mathbf{Y}\}} |\|\mathbf{X} - \mathbf{Z}\| - \|\mathbf{Y} - \mathbf{Z}\|| \right]$$

$$= \frac{1}{n-2} \left\{ \sum_{\mathbf{Z} \in \mathcal{X}_1 \setminus \{\mathbf{X}\}} \underbrace{d^{-1/2} |\|\mathbf{X} - \mathbf{Z}\| - \|\mathbf{Y} - \mathbf{Z}\||}_{} + \sum_{\mathbf{Z} \in \mathcal{X}_2 \setminus \{\mathbf{Y}\}} d^{-1/2} |\|\mathbf{X} - \mathbf{Z}\| - \|\mathbf{Y} - \mathbf{Z}\|| \right\}$$

High Dimensional Behavior of MADD

- If $\mathbf{X} \sim F_1, \mathbf{Y} \sim F_2$ are two independent observations, then under certain conditions

$$d^{-1/2}||\mathbf{X} - \mathbf{Y}|| \xrightarrow{P} \sqrt{\nu_{12}^2 + \sigma_1^2 + \sigma_2^2} \text{ as } d \to \infty$$

- Let us look at the expression,

$$\rho_0(\mathbf{X}, \mathbf{Y}) = d^{-1/2} \left[\frac{1}{n-2} \sum_{\mathbf{Z} \in \mathcal{X} \setminus \{\mathbf{X}, \mathbf{Y}\}} |\|\mathbf{X} - \mathbf{Z}\| - \|\mathbf{Y} - \mathbf{Z}\|| \right]$$

$$= \frac{1}{n-2} \left\{ \sum_{\mathbf{Z} \in \mathcal{X}_1 \setminus \{\mathbf{X}\}} d^{-1/2} |\|\mathbf{X} - \mathbf{Z}\| - \|\mathbf{Y} - \mathbf{Z}\|| + \sum_{\mathbf{Z} \in \mathcal{X}_2 \setminus \{\mathbf{Y}\}} \underbrace{d^{-1/2} |\|\mathbf{X} - \mathbf{Z}\| - \|\mathbf{Y} - \mathbf{Z}\||}_{} \right\}$$

Modified Version of MADD

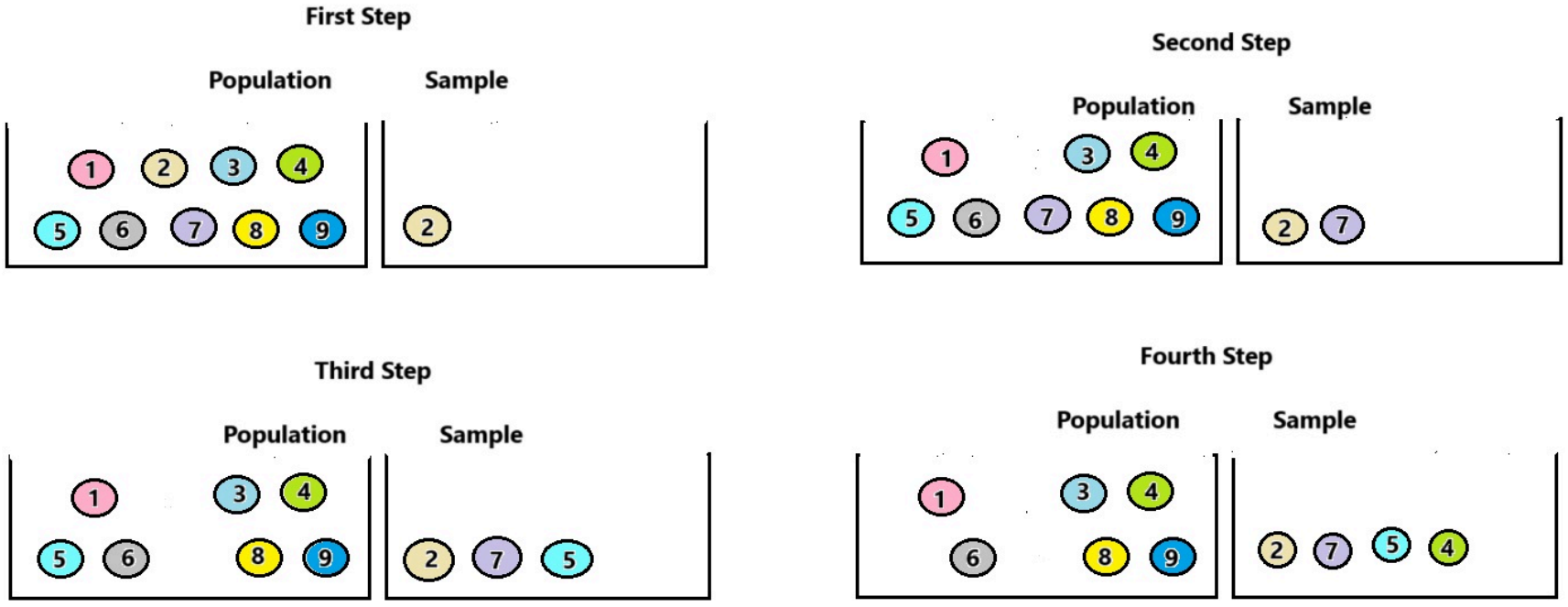
The modified version of MADD, will be of the form:

$$\rho_{Mod}(\mathbf{x},\mathbf{y}) = \frac{1}{|\mathcal{X}^* \setminus \{\mathbf{x},\mathbf{y}\}|} \sum_{\mathbf{z}\in\mathcal{X}^*\setminus\{\mathbf{x},\mathbf{y}\}} \big|\|\mathbf{x}-\mathbf{z}\| - \|\mathbf{y}-\mathbf{z}\|\big|$$

Where,

- $|\mathcal{X}^* \setminus \{\mathbf{x},\mathbf{y}\}|$ denotes the cardinality of $\mathcal{X}^* \setminus \{\mathbf{x},\mathbf{y}\}$.
- $\mathcal{X}^* \subset \mathcal{X}$.
- $\mathcal{X}^* \cap \mathcal{X}_j \neq \phi, \quad \text{for } j = 1, 2.$

Strategy 01: SRSWOR



- Straightforward choice
- Does not consider diversity in the data structure $\Longrightarrow$ May lack in representing the whole sample.

Strategy 02: Determinantal Point Process

DPP | L-Ensemble | k-DPP

- DPP creates a repulsion between points leading to the selection of a diverse set.
- A point process $\mathcal{P}$ on a discrete set $\mathcal{Y} = \{1, \ldots, N\}$ is a probability measure on $2^{\mathcal{Y}}$, the set of all subsets of $\mathcal{Y}$.
- It is said to be a DPP, if for a random set $\mathbf{Y}$ drawn as per $\mathcal{P}$, $\mathcal{P}(A \subseteq \boldsymbol{Y}) = det(K_A)$ for every subset $A \subseteq \mathcal{Y}$.

$$\mathcal{P}(i \in \boldsymbol{Y}) = K_{ii}, \quad \mathcal{P}(i \in \boldsymbol{Y}, j \in \boldsymbol{Y}) = K_{ii}K_{jj} - K_{ij}^2, \quad \text{for all } i, j \in \mathcal{Y}$$

- The general definition of DPP gives the probability of **inclusion** of a subset.

Strategy 02: Determinantal Point Process

DPP | L-Ensemble | k-DPP

- Defines a DPP through a positive semi-definite matrix L indexed by elements of $\mathcal{Y}$.

$$\mathcal{P}_L(\boldsymbol{Y} = A) = \det(L_A)/\det(L + I)$$

- Directly represent the probabilities of observing each subset of $\mathcal{Y}$.

Strategy 02: Determinantal Point Process

DPP | L-Ensemble | k-DPP

- A k-DPP selects exactly k points according to a DPP.
- DPP conditioned on the cardinality of the selected subset.

Strategy 02: Determinantal Point Process

- Considers the diversity in the data points.
- Depends on the choice of Kernel Matrix.

DPP-1

- Let us take $L = XX'$, where $X_{n \times p}$ is the data matrix.
- Denote the rows of X by $\{\mathbf{X_i}\}_{i=1}^{n}$, $\mathcal{P}_L(\mathbf{Y}) \propto det(L_{\mathbf{Y}}) = Vol^2(\{\mathbf{X_i}\}_{i \in \mathbf{Y}})$.

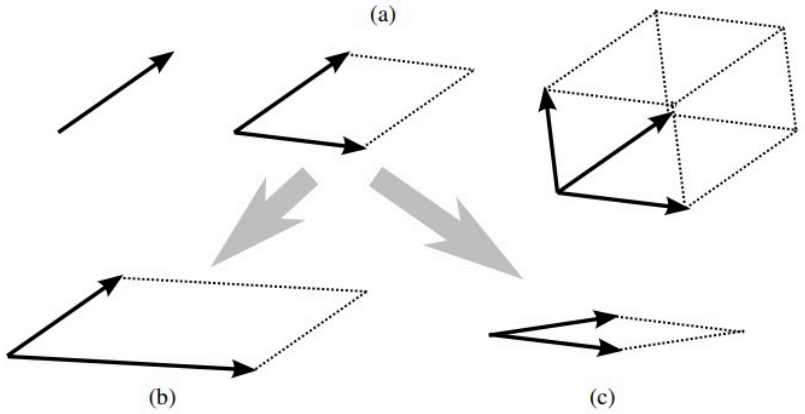


Figure: Geometrical View of DPPs (Source: Kulesza and Taskar 2012)

DPP-2

- In a 2-DPP problem, where $L = ((L_{i,j}))$, with $L_{i,j} = e^{-\frac{||\mathbf{x_i}-\mathbf{x_j}||^2}{d}}$.
- For the set $A = \{\mathbf{x_i}, \mathbf{x_j}\}$,

$$\mathcal{P}_L(A) \propto det(L_A) = 1 - e^{-\frac{2||\mathbf{x_i}-\mathbf{x_j}||^2}{d}}$$

- As the distance between $\mathbf{x_i}$ and $\mathbf{x_j}$ increases, the probability of their selection also increases.
- Thus we consider *Radial-Basis Function Kernel* between the data points present in each population.
- For two data points $\mathbf{x_i}$ and $\mathbf{x_j}$, it is defined as

$$L(\mathbf{x_i}, \mathbf{x_j}) = e^{-\frac{||\mathbf{x_i}-\mathbf{x_j}||^2}{d}}$$

- Being a kernel it will result in a positive semi-definite symmetric matrix (Lanckriet et al. 2002).

Simulation Studies

Experiment Specifications:

> - 5-nearest neighbor classifier
> - Test set size: 500 (250 observations from each class)
> - Sample sizes: $n_1 = n_2 = n = 50, 100$
> - Dimensionality: $d = 20, 50, 100$
> - No. of points Selected from each population: $k = 2, 4, 8, 16$
> - 100 replications considered

Simulation 01: A Pure Location Problem

- Population 1 $\equiv N_d(\mathbf{0}, I_d)$ & Population 2 $\equiv N_d(0.5\mathbf{1_d}, I_d)$

d= 20 d= 50 d= 100

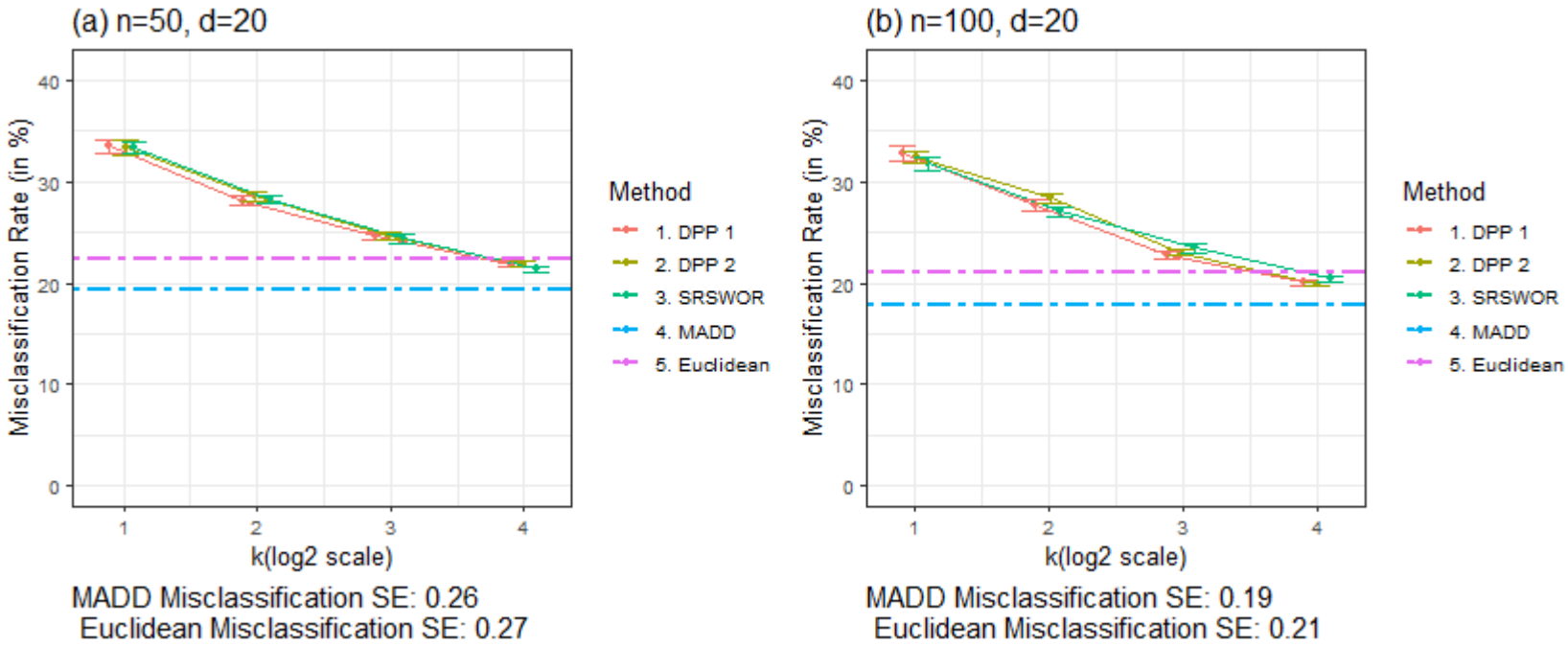


Figure: Misclassification rates (in %) in a Pure Location Problem. The reported numbers are averages `±` SE based on 100 replications.

Simulation 01: A Pure Location Problem

- Population 1 $\equiv N_d(\mathbf{0}, I_d)$ & Population 2 $\equiv N_d(0.5\mathbf{1_d}, I_d)$

d= 20 | **d= 50** | d= 100

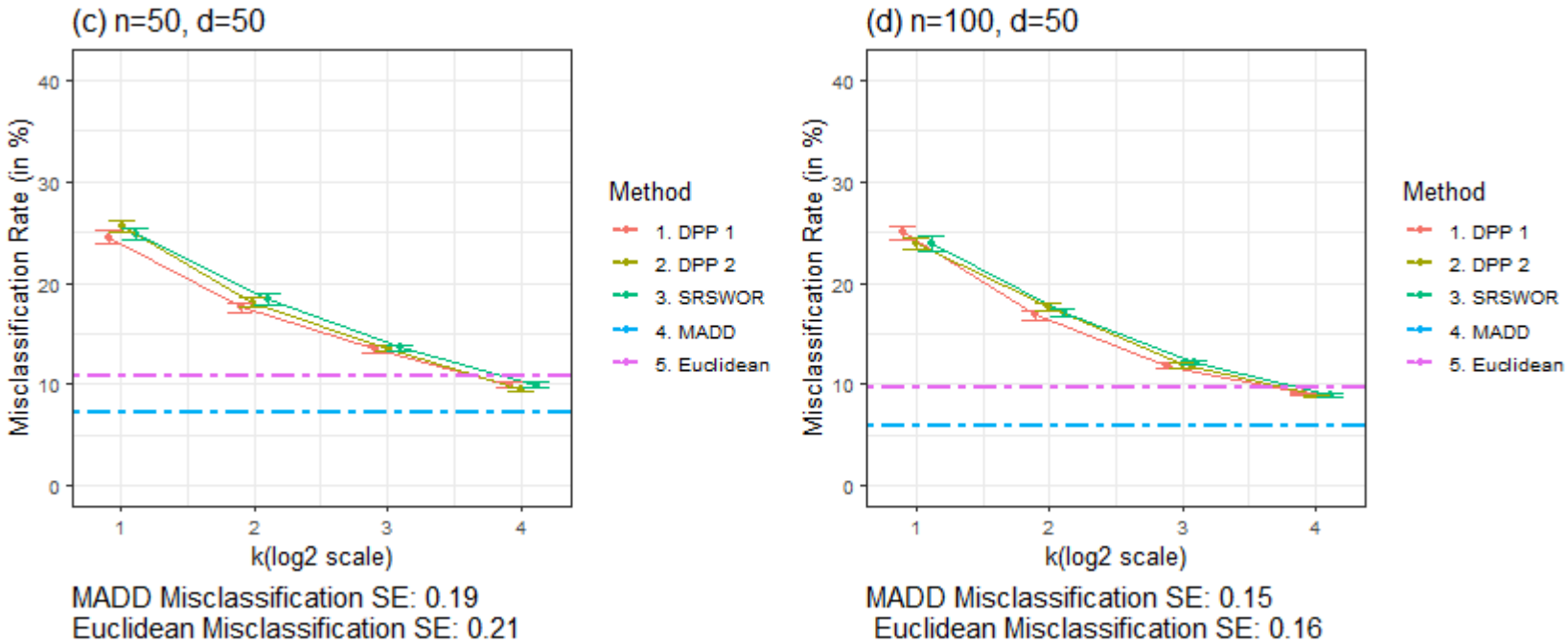


Figure: Misclassification rates (in %) in a Pure Location Problem. The reported numbers are averages `±` SE based on 100 replications.

Simulation 01: A Pure Location Problem

- Population 1 $\equiv N_d(\mathbf{0}, I_d)$ & Population 2 $\equiv N_d(0.5\mathbf{1_d}, I_d)$

d= 20 d= 50 **d= 100**

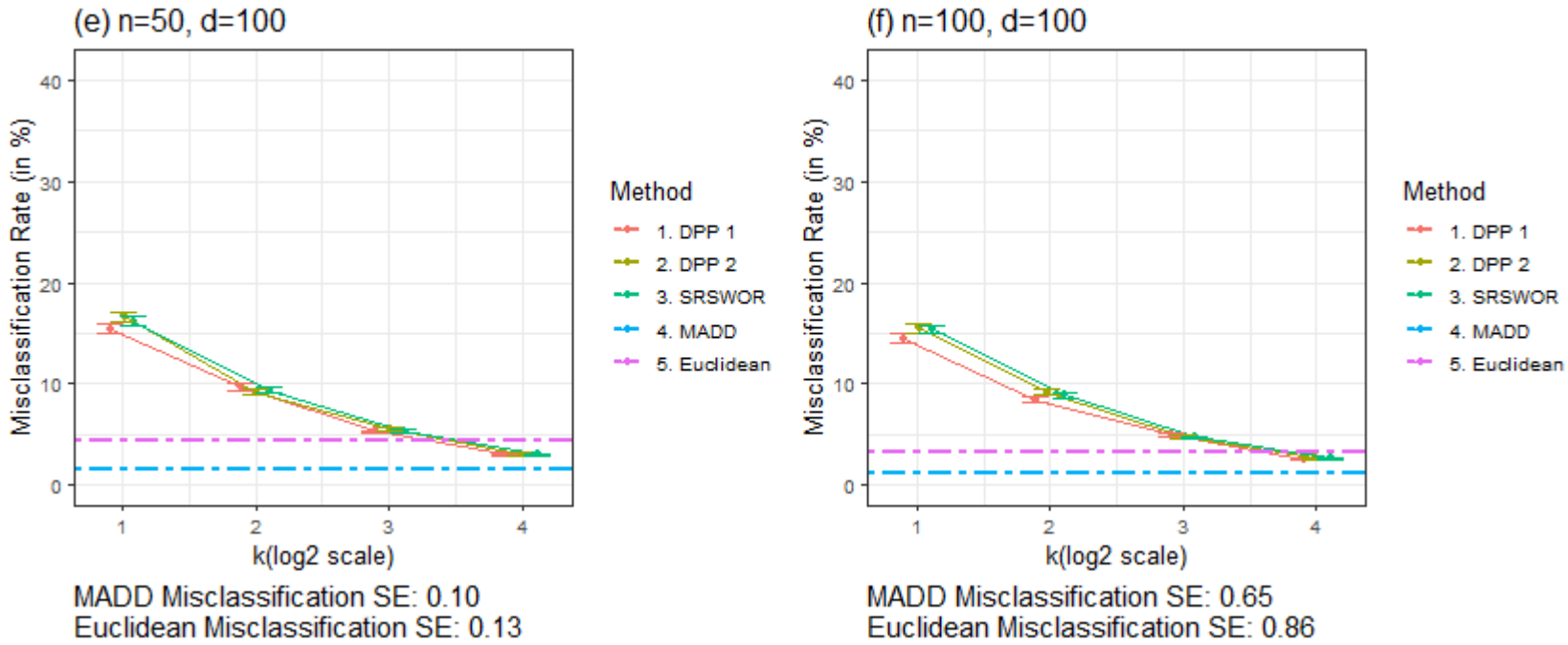


Figure: Misclassification rates (in %) in a Pure Location Problem. The reported numbers are averages `±` SE based on 100 replications.

Simulation 02: A Pure Scale Problem

- Population 1 $\equiv N_d(\mathbf{0}, I_d)$ & Population 2 $\equiv N_d(\mathbf{0}, 2I_d)$

d= 20 | d= 50 | d= 100

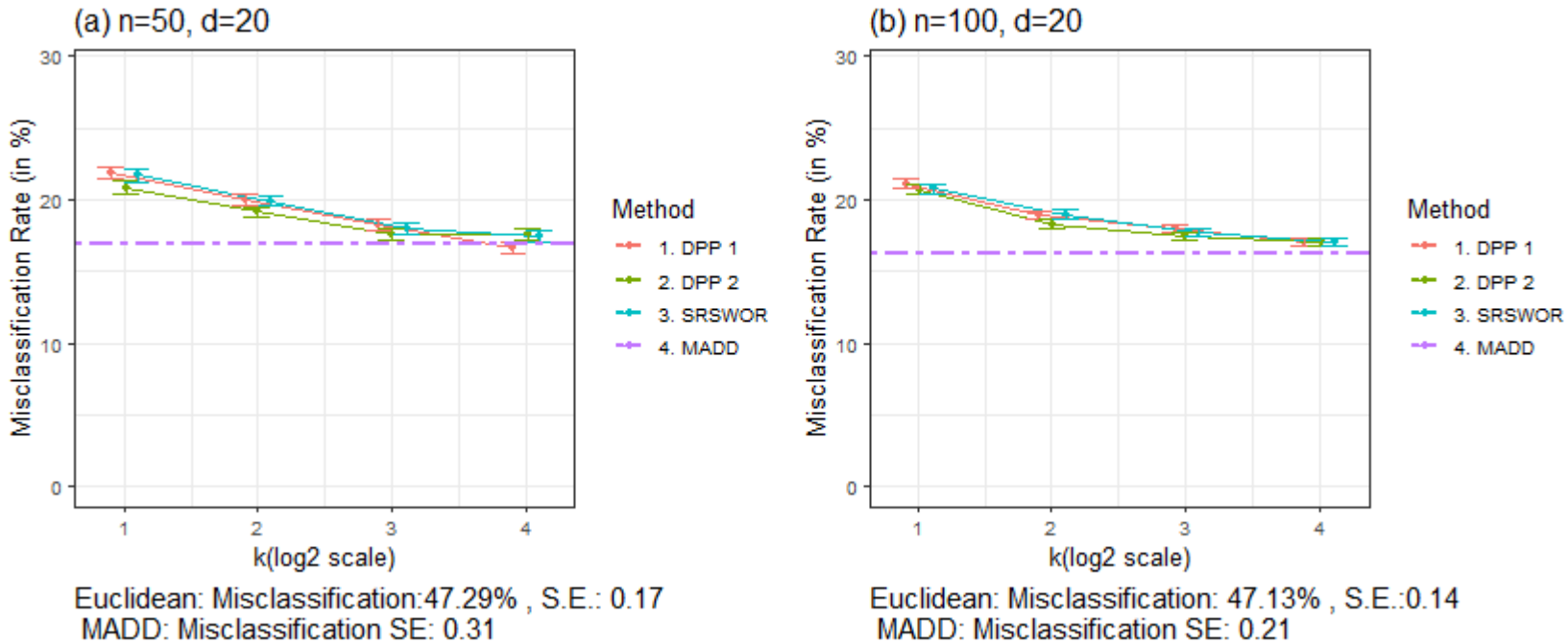


Figure: Misclassification rates (in %) in a Pure Scale Problem. The reported numbers are averages `±` SE based on 100 replications.

Simulation 02: A Pure Scale Problem

- Population 1 $\equiv N_d(\mathbf{0}, I_d)$ & Population 2 $\equiv N_d(\mathbf{0}, 2I_d)$

d= 20 | **d= 50** | d= 100

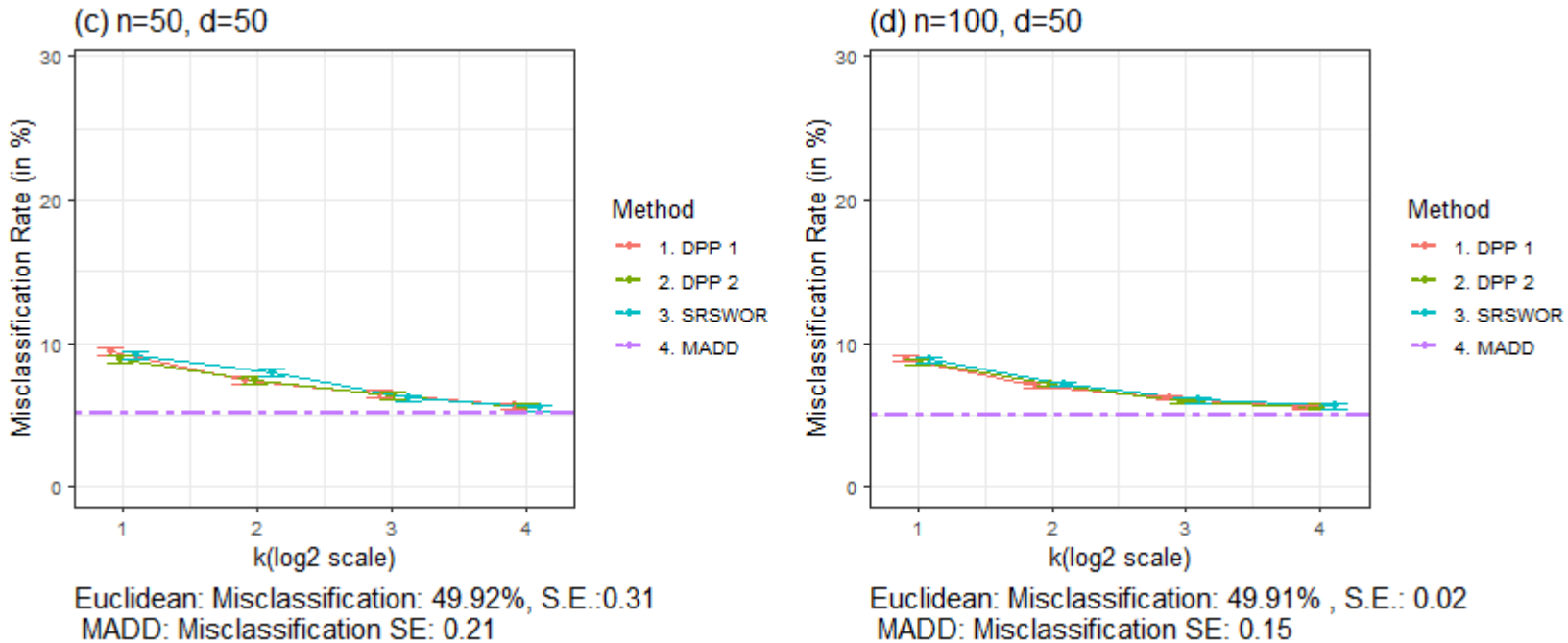


Figure: Misclassification rates (in %) in a Pure Scale Problem. The reported numbers are averages `±` SE based on 100 replications.

Simulation 02: A Pure Scale Problem

- Population 1 $\equiv N_d(\mathbf{0}, I_d)$ & Population 2 $\equiv N_d(\mathbf{0}, 2I_d)$

d= 20 d= 50 **d= 100**

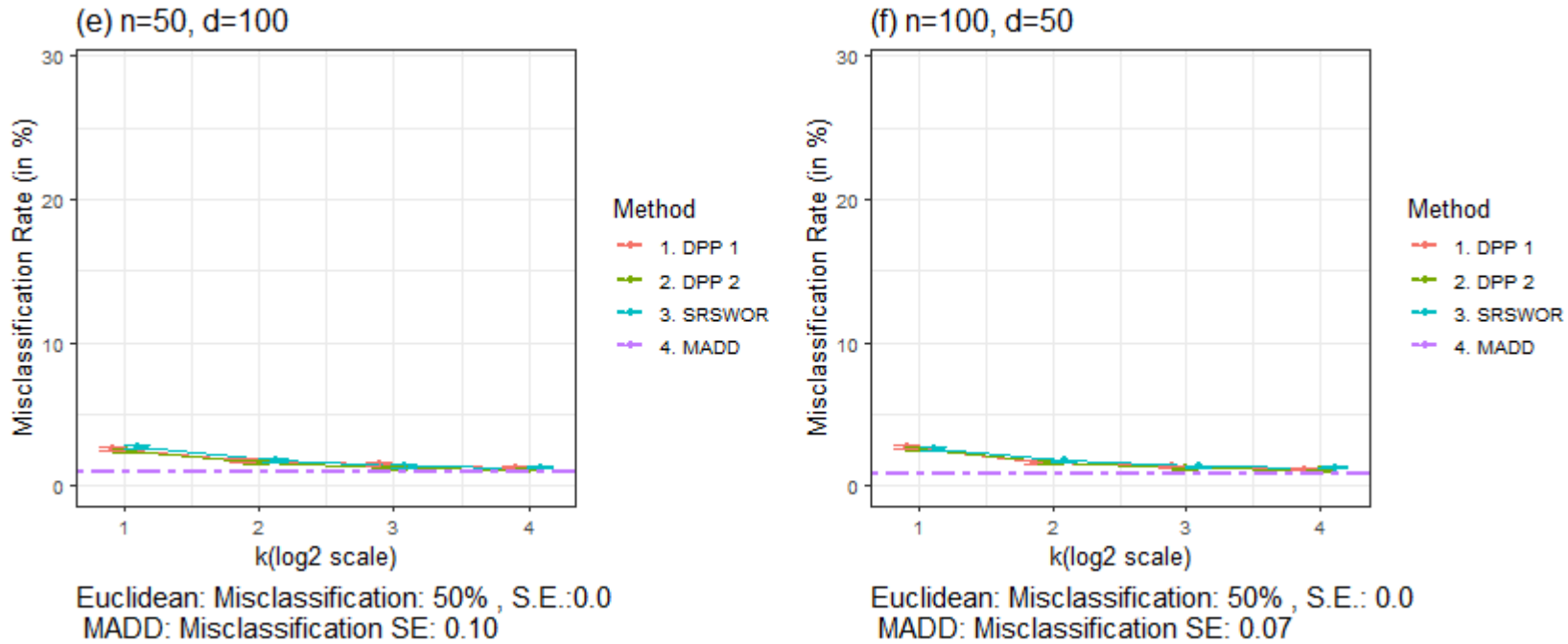


Figure: Misclassification rates (in %) in a Pure Scale Problem. The reported numbers are averages `±` SE based on 100 replications.

Simulation 03: Location Problem with Autocorrelated Features

- Population 1 $\equiv Y_j = Y_{j-1} + \epsilon_j, \epsilon_j \sim^{ind} N(0,1)$ & Population 2 $\equiv Y_j - 0.5 = 0.5(Y_{j-1} - 0.5) + \epsilon_j^*, \epsilon_j^* \sim^{ind} N(0,1)$

d= 20 d= 50 d= 100

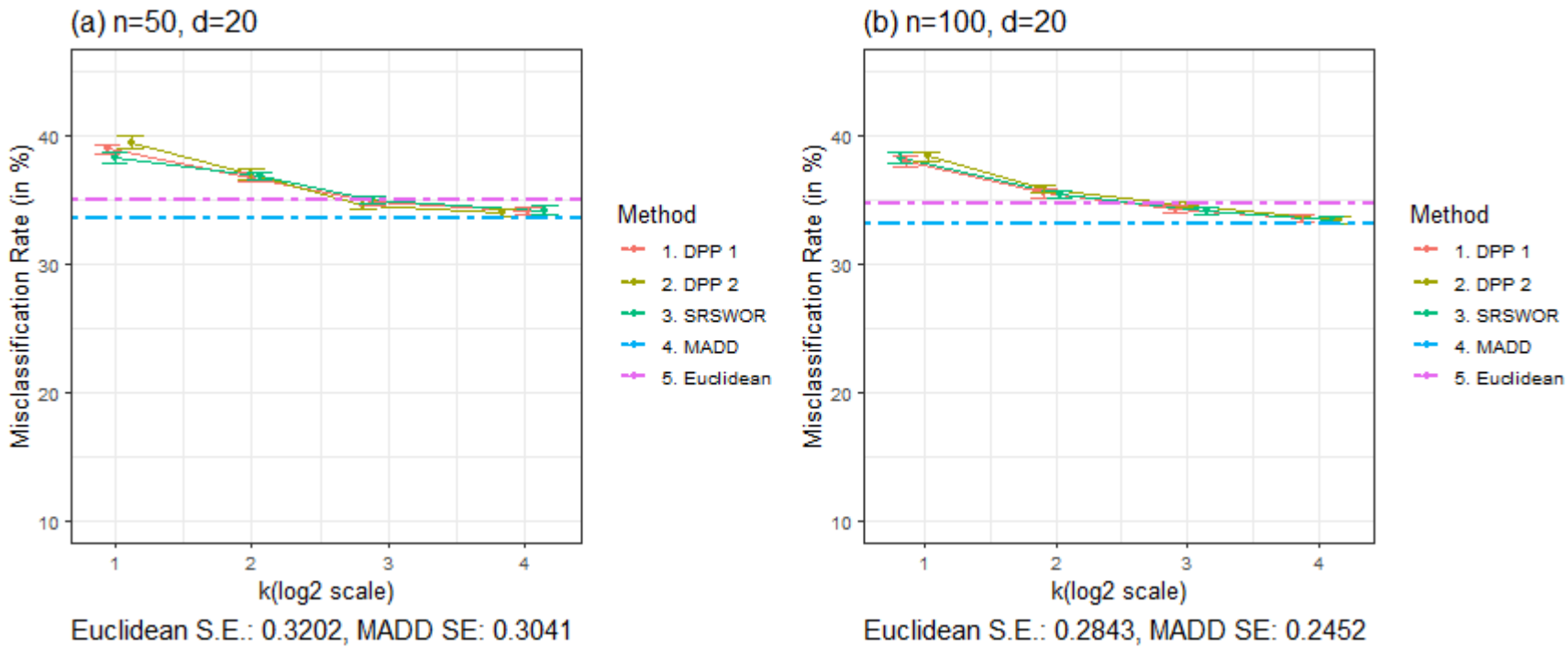


Figure: Misclassification rates (in %) in a location Problem with autocorrelated features. The reported numbers are averages `±` SE based on 100 replications.

Simulation 03: Location Problem with Autocorrelated Features

- Population 1 $\equiv Y_j = Y_{j-1} + \epsilon_j, \epsilon_j \sim^{ind} N(0,1)$ & Population 2 $\equiv Y_j - 0.5 = 0.5(Y_{j-1} - 0.5) + \epsilon_j^*, \epsilon_j^* \sim^{ind} N(0,1)$

d= 20 | **d= 50** | d= 100

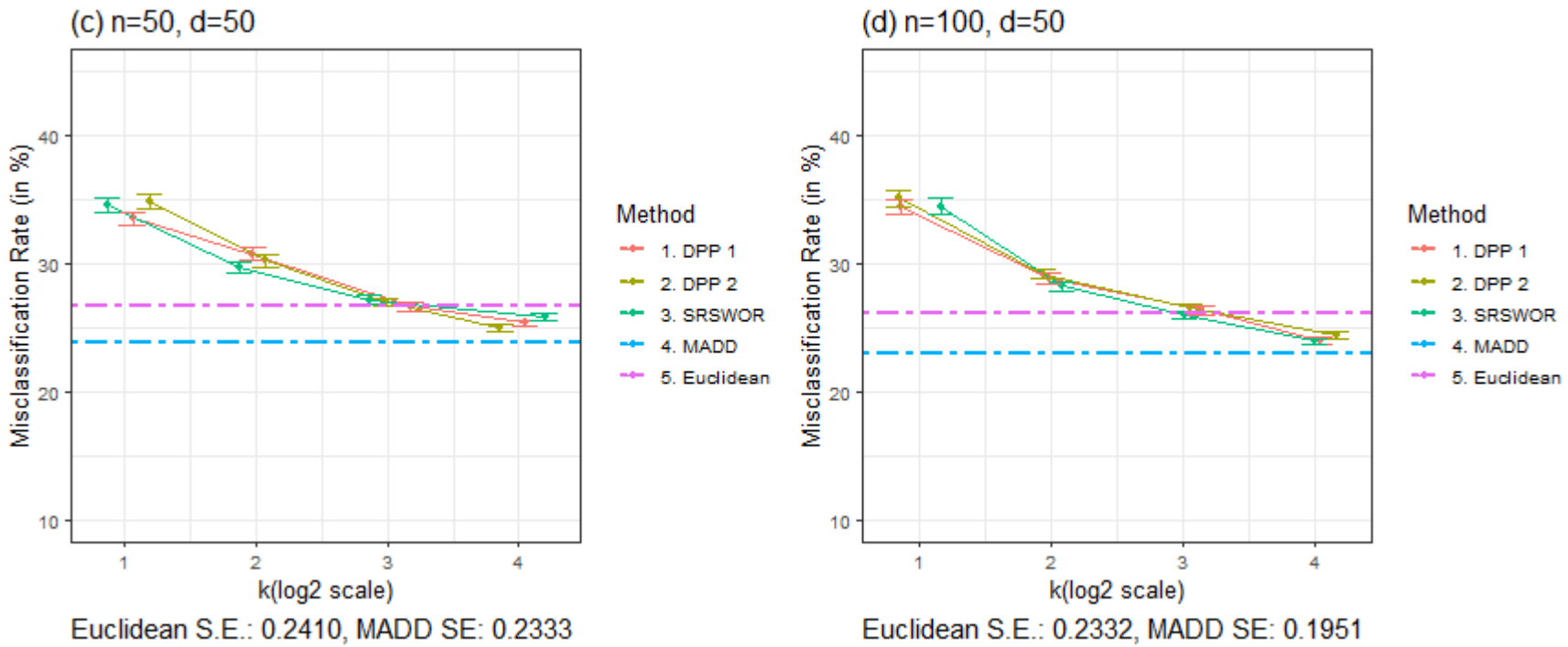


Figure: Misclassification rates (in %) in a location Problem with autocorrelated features. The reported numbers are averages `±` SE based on 100 replications.

Simulation 03: Location Problem with Autocorrelated Features

- Population 1 $\equiv Y_j = Y_{j-1} + \epsilon_j, \epsilon_j \sim^{ind} N(0,1)$ & Population 2 $\equiv Y_j - 0.5 = 0.5(Y_{j-1} - 0.5) + \epsilon_j^*, \epsilon_j^* \sim^{ind} N(0,1)$

d= 20 d= 50 **d= 100**

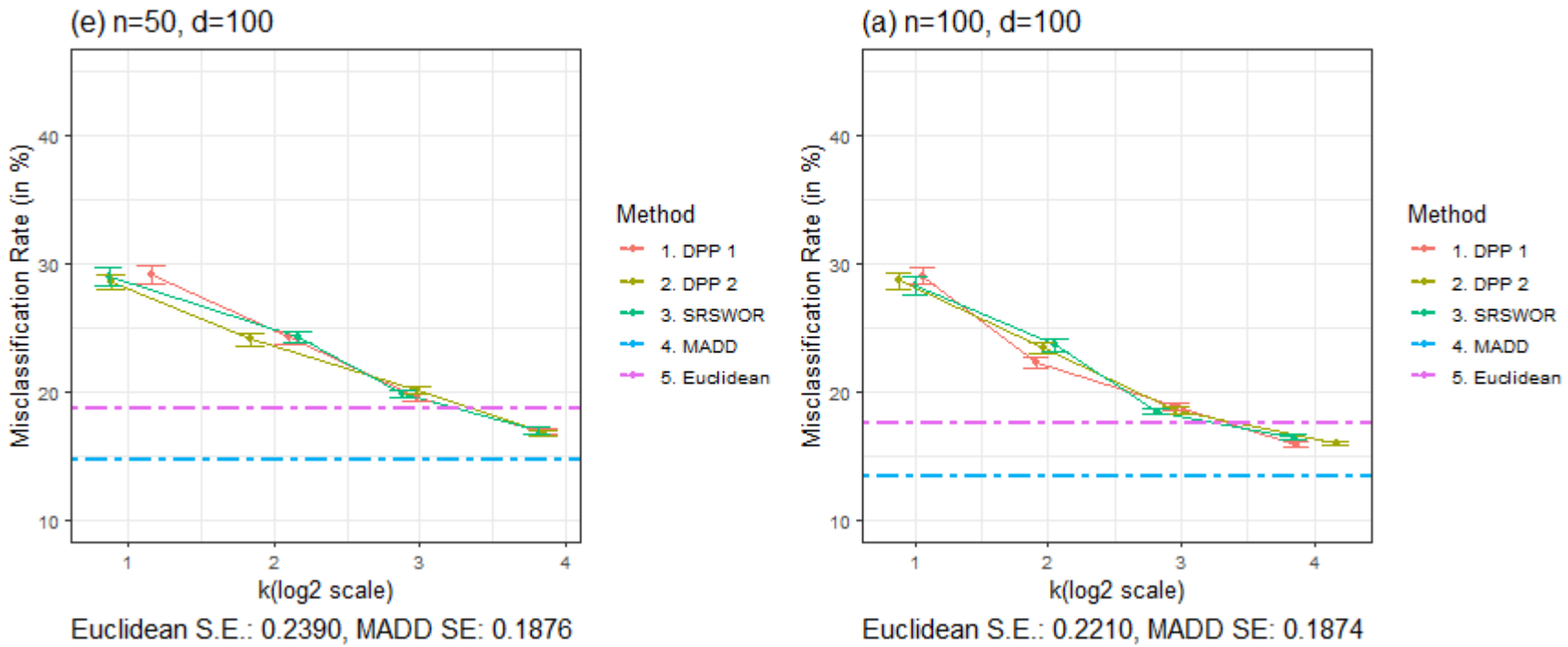


Figure: Misclassification rates (in %) in a location Problem with autocorrelated features. The reported numbers are averages `±` SE based on 100 replications.

Comparison of Computing times

d= 20 | d= 50 | d= 100

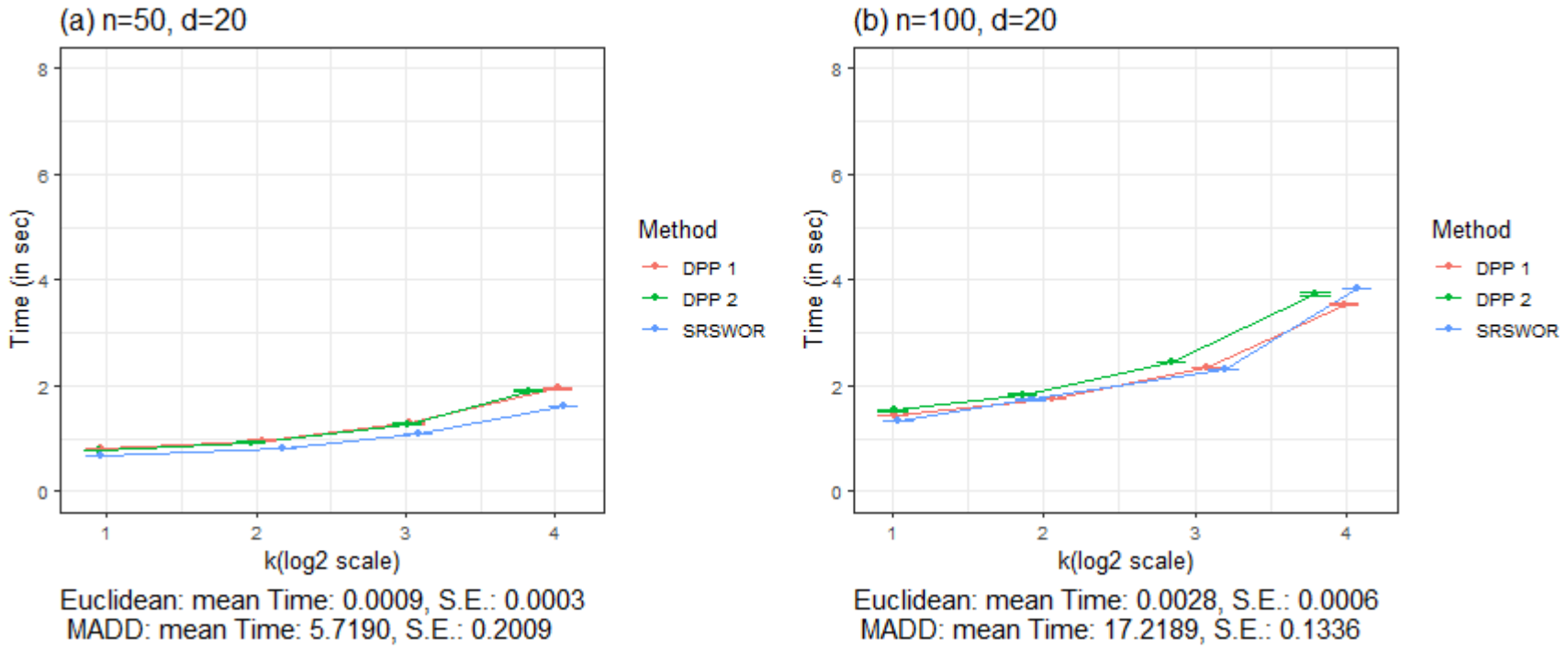


Figure: Computing Times taken by different Methods (in seconds). The reported numbers are averages `±` SE based on 100 replications.

Comparison of Computing times

d= 20 | **d= 50** | d= 100

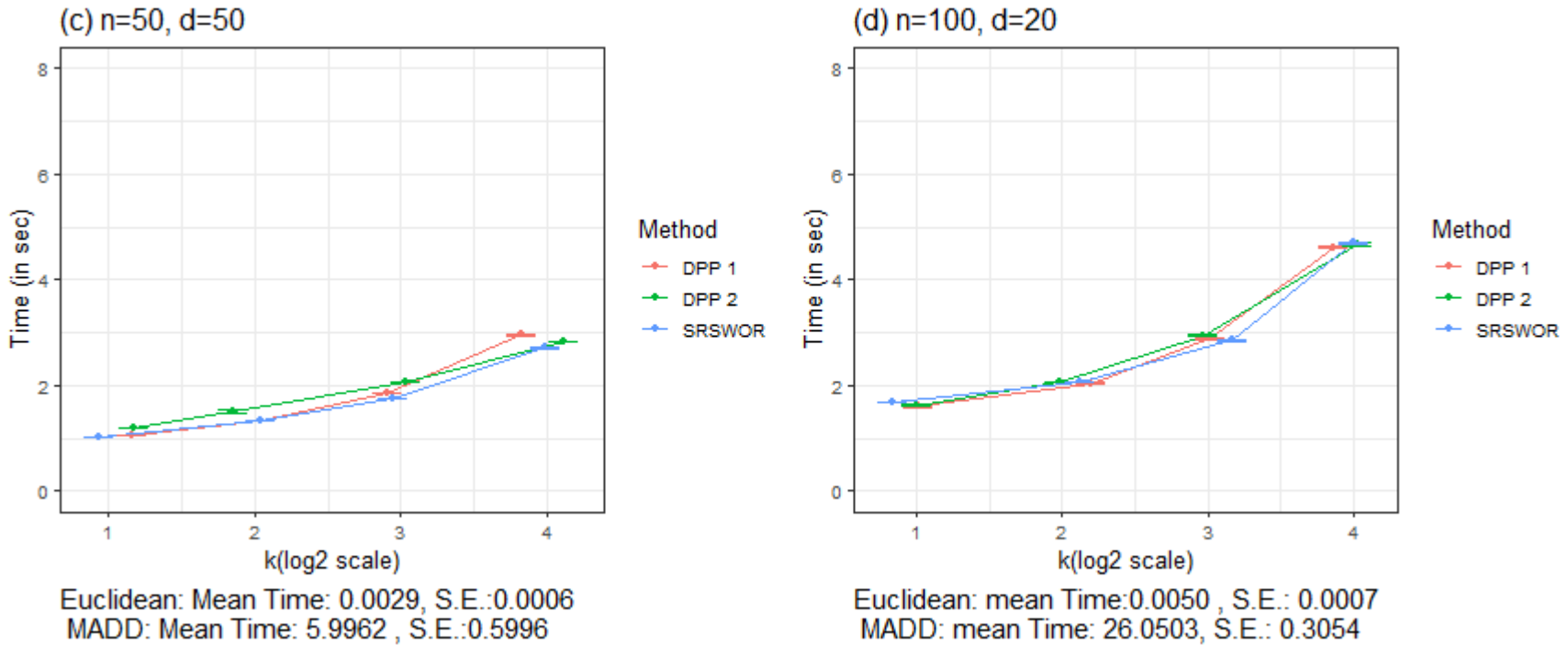


Figure: Computing Times taken by different Methods (in seconds). The reported numbers are averages `±` SE based on 100 replications.

Comparison of Computing times

d= 20 d= 50 **d= 100**

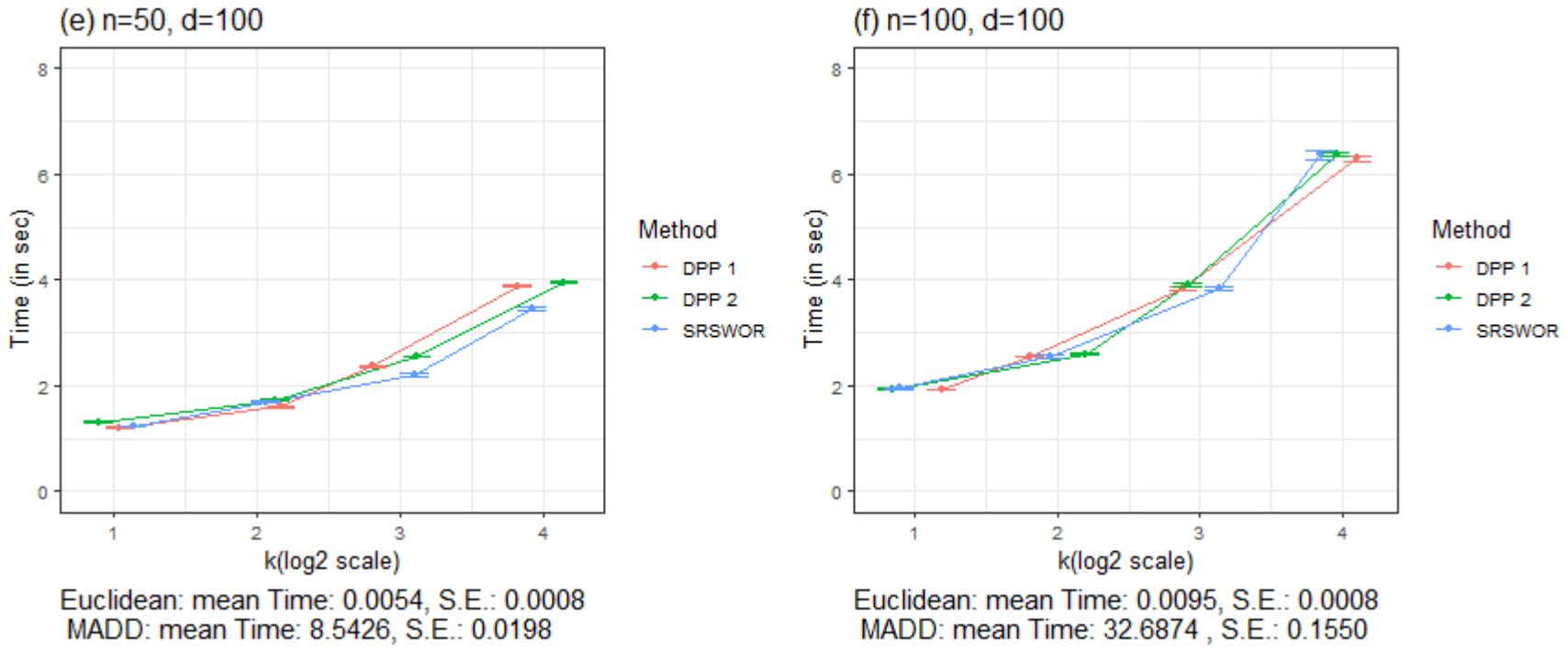


Figure: Computing Times taken by different Methods (in seconds). The reported numbers are averages `±` SE based on 100 replications.

Extension to Multi-Class Classification Problem

Scalable Version of MADD

$$\rho_{Mod}(\mathbf{x}, \mathbf{y}) = \frac{1}{|\mathcal{X}^* \setminus \{\mathbf{x}, \mathbf{y}\}|} \sum_{\mathbf{z} \in \mathcal{X}^* \setminus \{\mathbf{x}, \mathbf{y}\}} \left| \|\mathbf{x} - \mathbf{z}\| - \|\mathbf{y} - \mathbf{z}\| \right|$$

Where,

- $|\mathcal{X}^* \setminus \{\mathbf{x}, \mathbf{y}\}|$ denotes the cardinality of $\mathcal{X}^* \setminus \{\mathbf{x}, \mathbf{y}\}$.
- $\mathcal{X}^* \subset \mathcal{X}$.
- $\mathcal{X}^* \cap \mathcal{X}_j \neq \phi, \quad \text{for } j = 1, 2, \ldots, J.$

Misclassification Rate for a Three-Class Pure Location Problem

- Population 1 $\equiv N_d(-0.5\mathbf{1_d}, I_d)$ & Population 2 $\equiv N_d(\mathbf{0}, I_d)$ & Population 3 $\equiv N_d(0.5\mathbf{1_d}, I_d)$

d= 20 d= 50

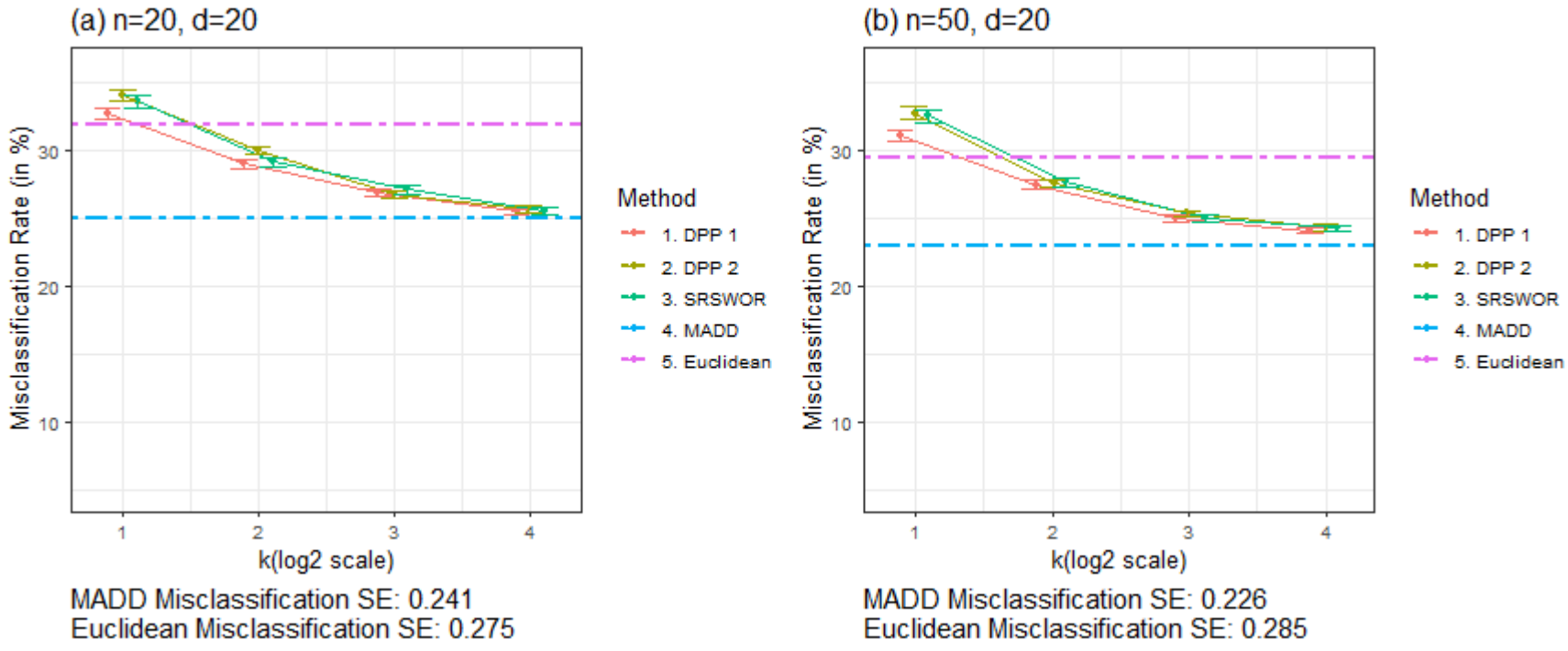


Figure: Misclassification rates (in %) in a Pure Location Problem for a 3-class classification problem. The reported numbers are averages `±` SE based on 100 replications.

Misclassification Rate for a Three-Class Pure Location Problem

- Population 1 $\equiv N_d(-0.5\mathbf{1_d}, I_d)$ & Population 2 $\equiv N_d(\mathbf{0}, I_d)$ & Population 3 $\equiv N_d(0.5\mathbf{1_d}, I_d)$

d= 20 | d= 50

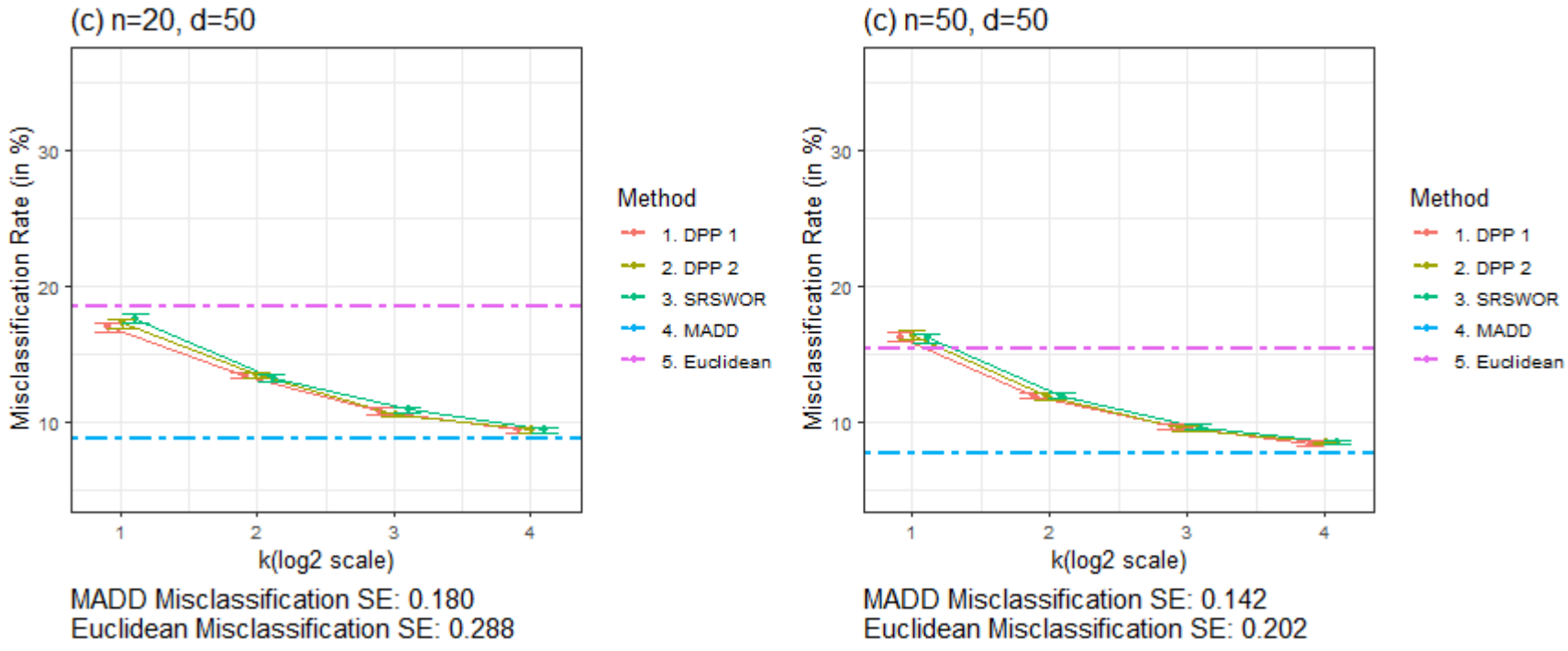


Figure: Misclassification rates (in %) in a Pure Location Problem for a 3-class classification problem. The reported numbers are averages `±` SE based on 100 replications.

Comparison of Computing Times

d= 20 | d= 50

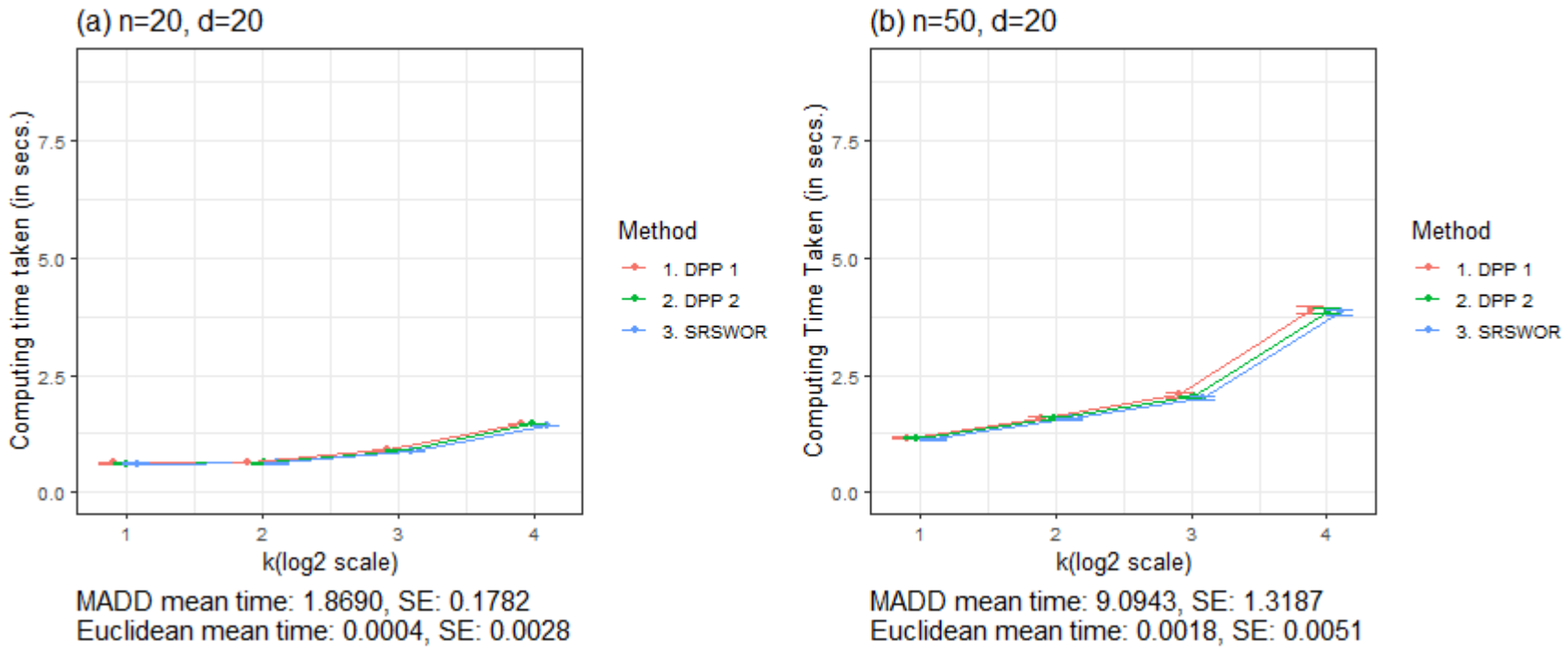


Figure: Computing times taken (in secs.) in a 3-class classification problem with only location difference. The reported numbers are averages `±` SE based on 100 replications.

Comparison of Computing Times

d= 20 | d= 50

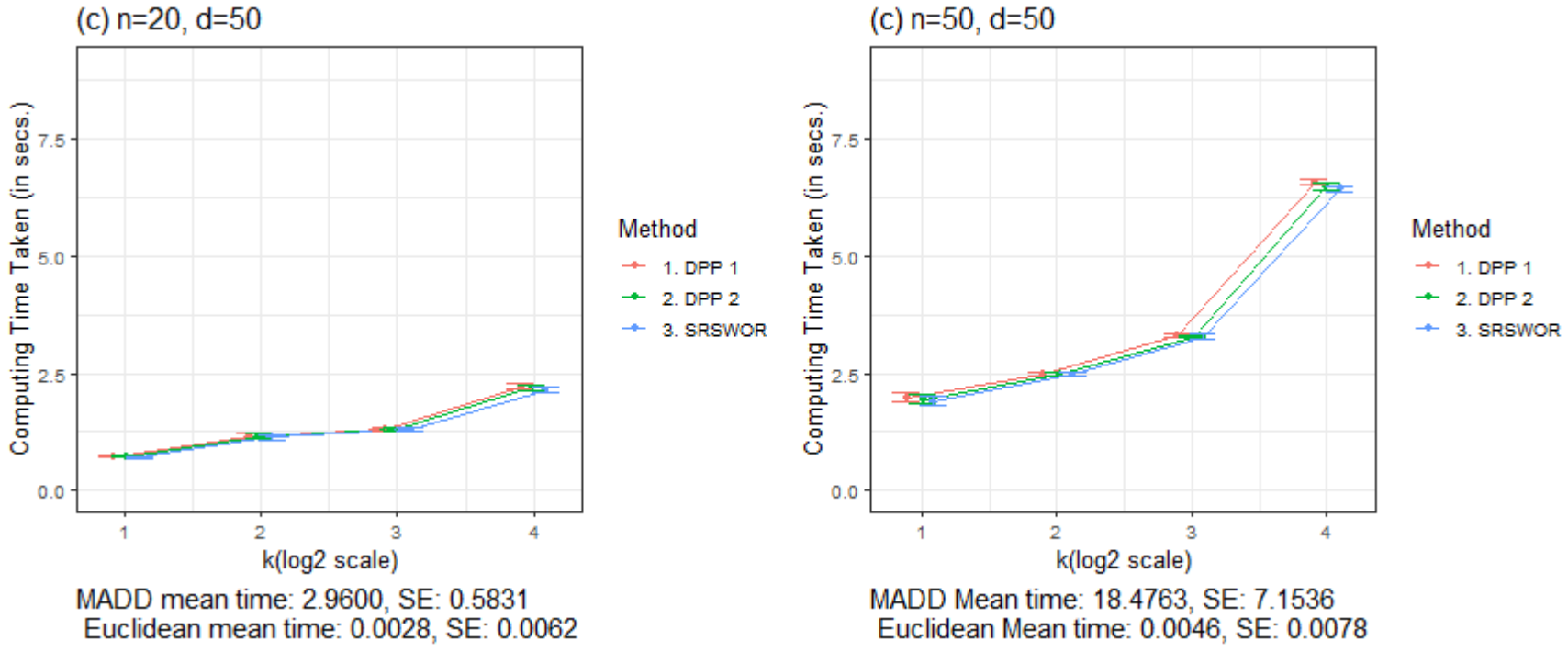


Figure: Computing times taken (in secs.) in a 3-class classification problem with only location difference. The reported numbers are averages `±` SE based on 100 replications.

Scalable Version of g-MADD

Generalized MADD: An Overview

- Usual MADD is only confined to the cases where populations either differ in their location or in their total variance.
- For two vectors $\mathbf{x}$ and $\mathbf{y}$, define,

$$\rho_{h,\psi}(\mathbf{x},\mathbf{y}) = \frac{1}{n-2} \sum_{\mathbf{z}\in\mathcal{X}\setminus\{\mathbf{x},\mathbf{y}\}} \left|\phi_{h,\psi}(\mathbf{x},\mathbf{z}) - \phi_{h,\psi}(\mathbf{y},\mathbf{z})\right|$$

Where,

- $h:\mathbb{R}^+\to\mathbb{R}^+$ and $\psi:\mathbb{R}^+\to\mathbb{R}^+$ continuous, monotonically increasing, $h(0)=\psi(0)=0$ such that $\phi_{h,\psi}(\mathbf{x},\mathbf{y}) = h\left(\frac{1}{d}\sum_{q=1}^{d}\psi\left(|x^{(q)}-y^{(q)}|\right)\right)$.
- Here, we will take $h(t)=t$, $\psi(t)=1-e^{-t}$.

Computational Challenges with g-MADD

- Again, the computational complexity becomes $O(n^2 d)$ for classifying single observation.
- We will propose a scalable version of g-MADD of the form:

$$\rho_{h,\psi}^{Mod}(\mathbf{x}, \mathbf{y}) = \frac{1}{|\mathcal{X}^* \setminus \{\mathbf{x}, \mathbf{y}\}|} \sum_{\mathbf{z} \in \mathcal{X}^* \setminus \{\mathbf{x}, \mathbf{y}\}} \left| \phi_{h,\psi}(\mathbf{x}, \mathbf{z}) - \phi_{h,\psi}(\mathbf{y}, \mathbf{z}) \right|$$

where $\mathcal{X}^*$ is a subset of $\mathcal{X}$ such that $\mathcal{X}^* \cap \mathcal{X}_j \neq \phi, \quad \text{for } j = 1, 2.$

Choice of L-ensemble Matrix

Generalization of DPP-1

- Euclidean distance is directly related to the volume of the parallelepiped.
- No such direct connection for a general distance.

Generalization of DPP-2

- Recall DPP-2: Used $L(\mathbf{x_i}, \mathbf{x_j}) = e^{-\frac{||\mathbf{x_i}-\mathbf{x_j}||^2}{d}}$.
- For g-MADD, propose: $L(\mathbf{x_i}, \mathbf{x_j}) = e^{-\frac{1}{d}\sum_{q=1}^{d}\psi(|x_i^{(q)}-x_j^{(q)}|)}$.
- Replace Euclidean distance with a general distance function.

Simulation: Features are Independent Normal vs Features are Independent t

- Both the population have mean $\mathbf{0}$ and dispersion $3I_d$.

d= 20 | d= 50 | d= 100

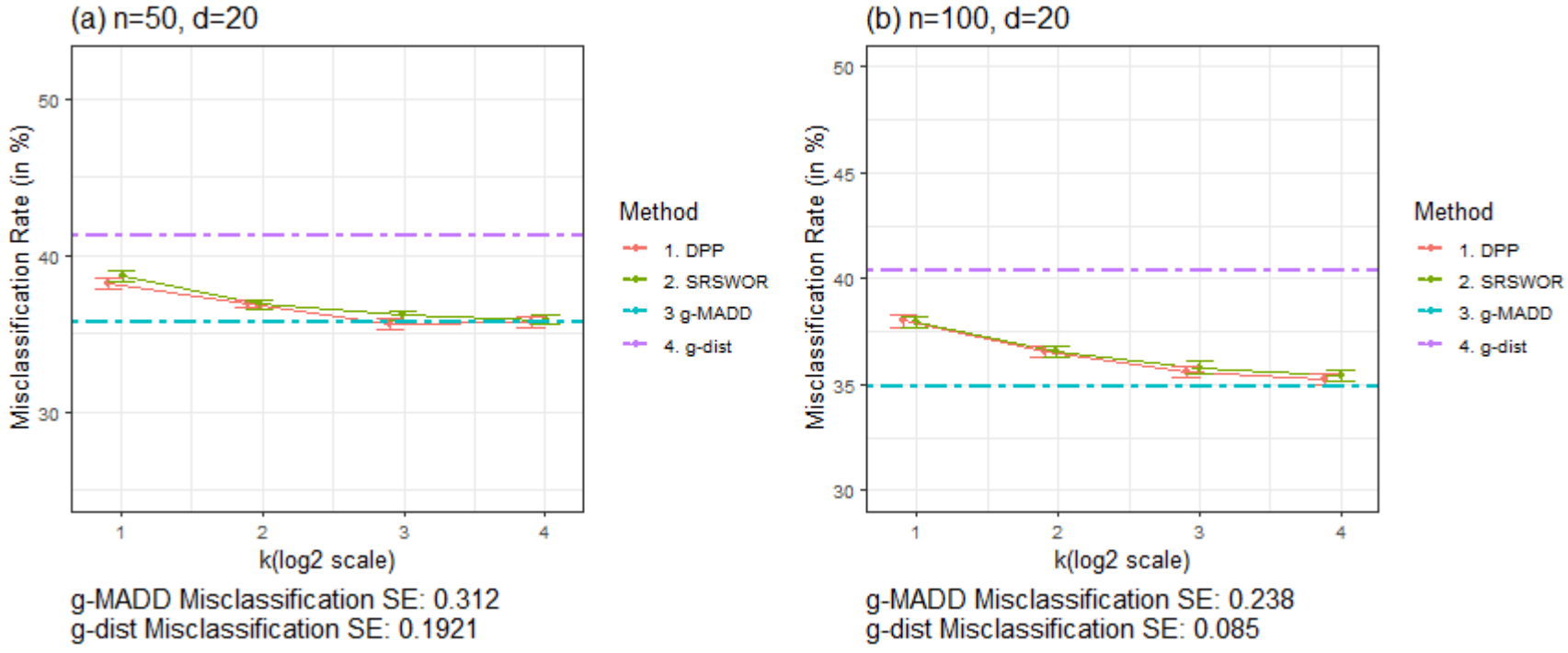


Figure: Misclassification rates (in %) when the features are independent normal vs when the features are independent t. The reported numbers are averages `±` SE based on 100 replications.

Simulation: Features are Independent Normal vs Features are Independent t

- Both the population have mean $\mathbf{0}$ and dispersion $3I_d$.

d= 20 | **d= 50** | d= 100

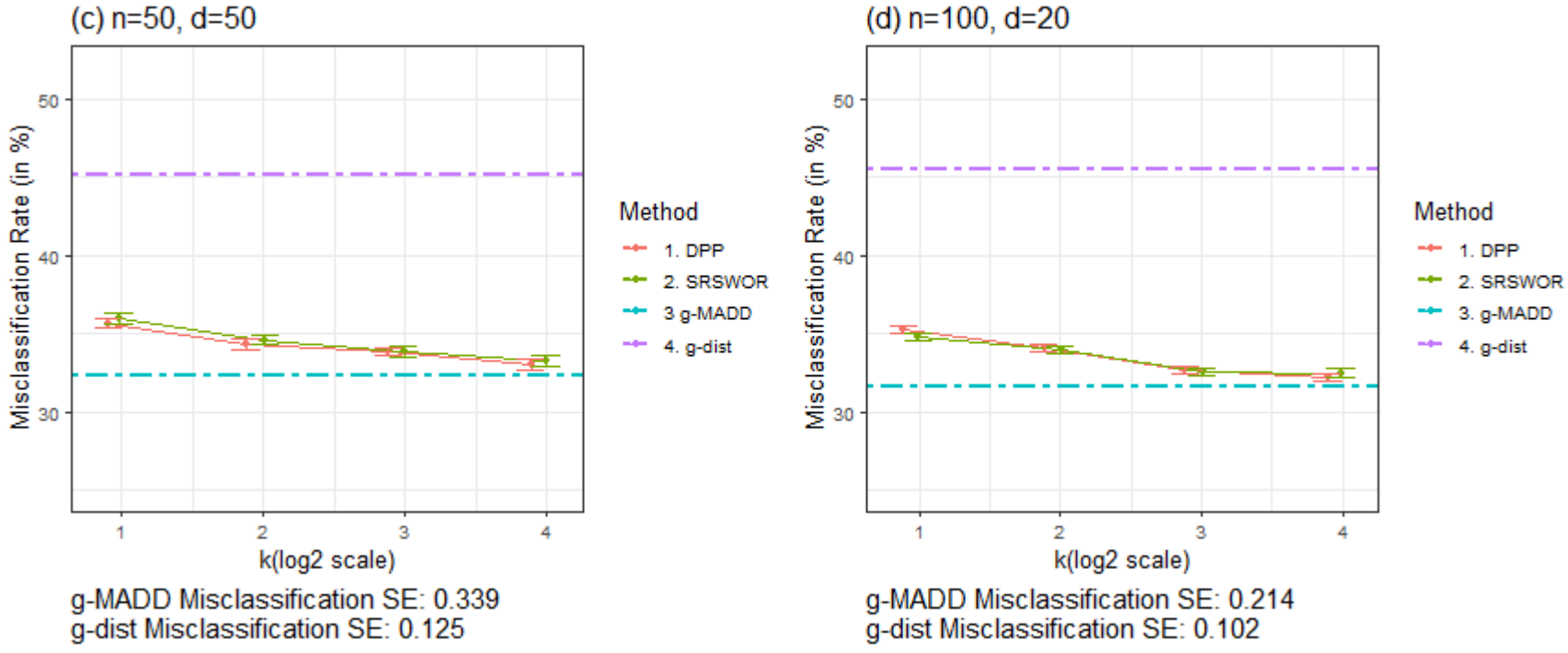


Figure: Misclassification rates (in %) when the features are independent normal vs when the features are independent t. The reported numbers are averages `±` SE based on 100 replications.

Simulation: Features are Independent Normal vs Features are Independent t

- Both the population have mean $\mathbf{0}$ and dispersion $3I_d$.

d= 20 d= 50 **d= 100**

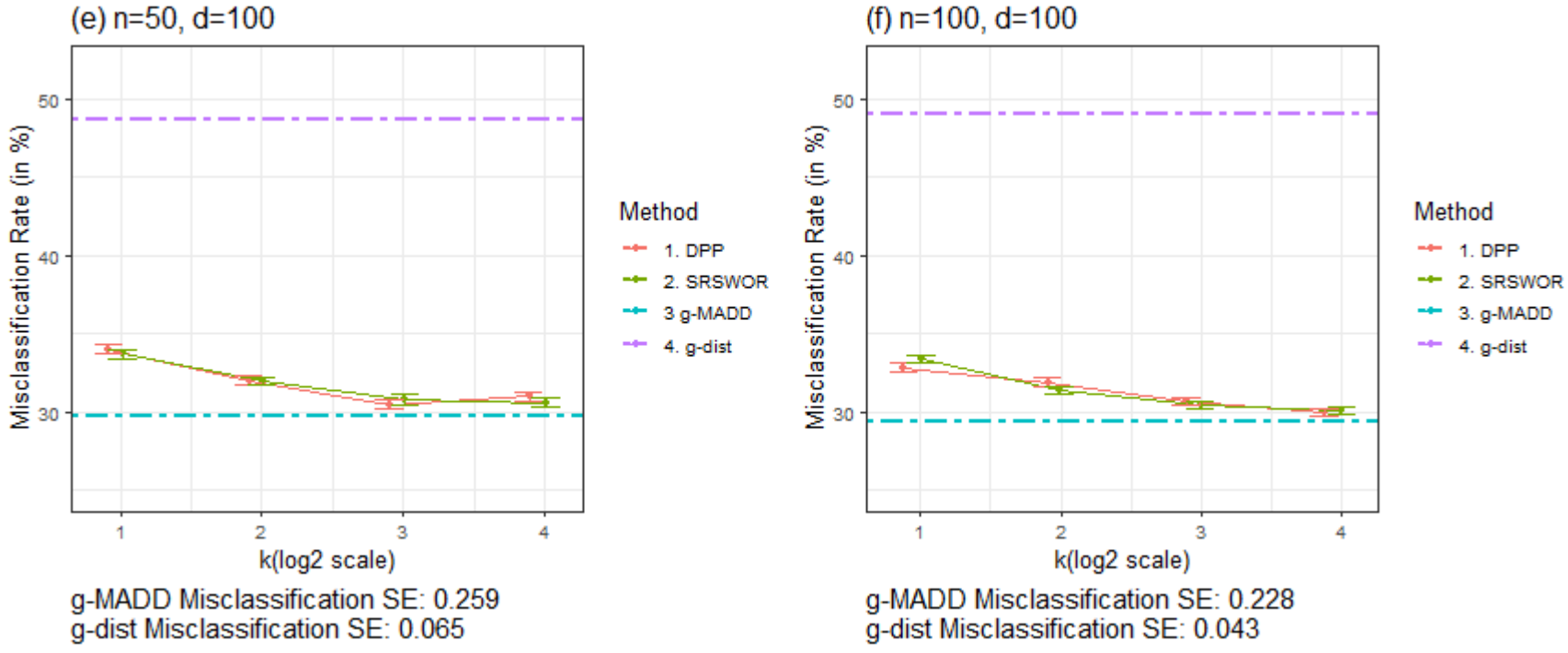


Figure: Misclassification rates (in %) when the features are independent normal vs when the features are independent t. The reported numbers are averages `±` SE based on 100 replications.

Comparison of Computing Times

d= 20 | d= 50 | d= 100

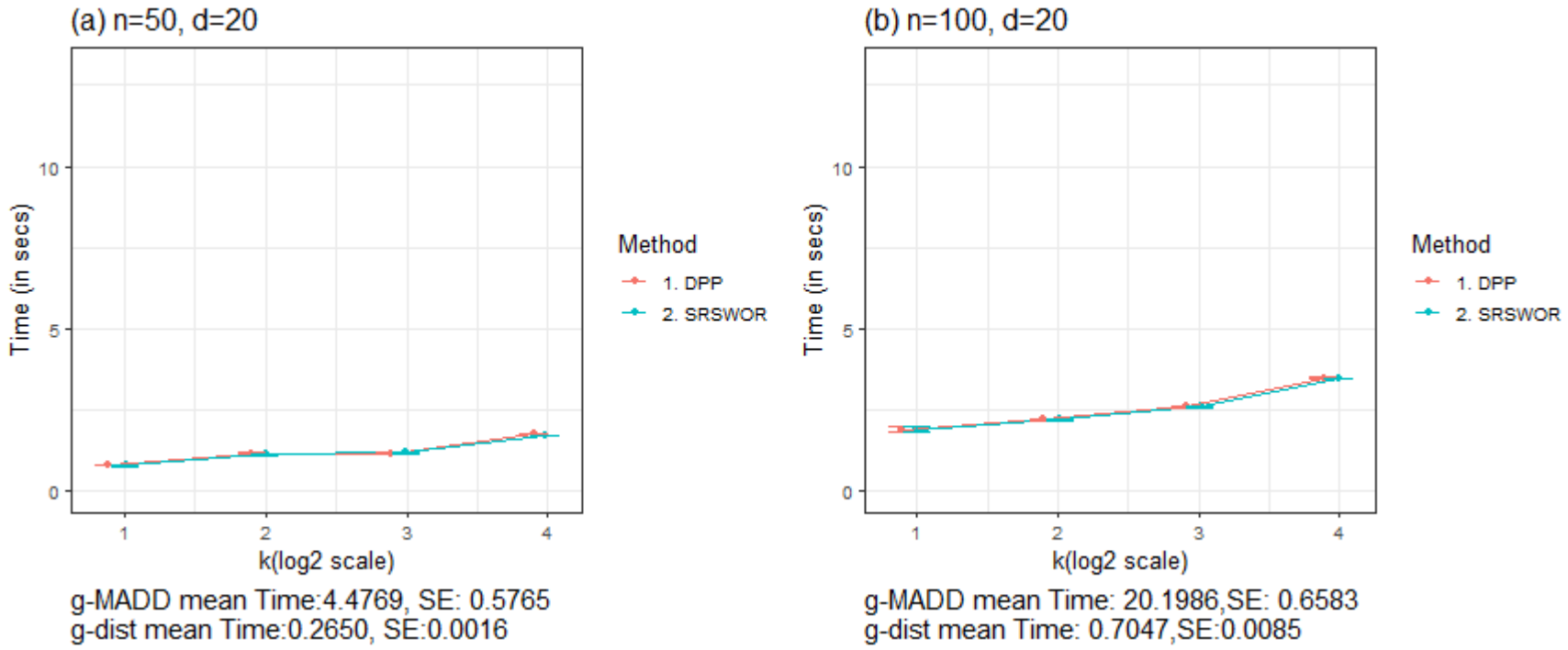


Figure: Comparison of computing times when the features are independent normal vs when the features are independent t (with same mean and dispersion). The reported numbers are averages `±` SE based on 100 replications.

Comparison of Computing Times

d= 20 | **d= 50** | d= 100

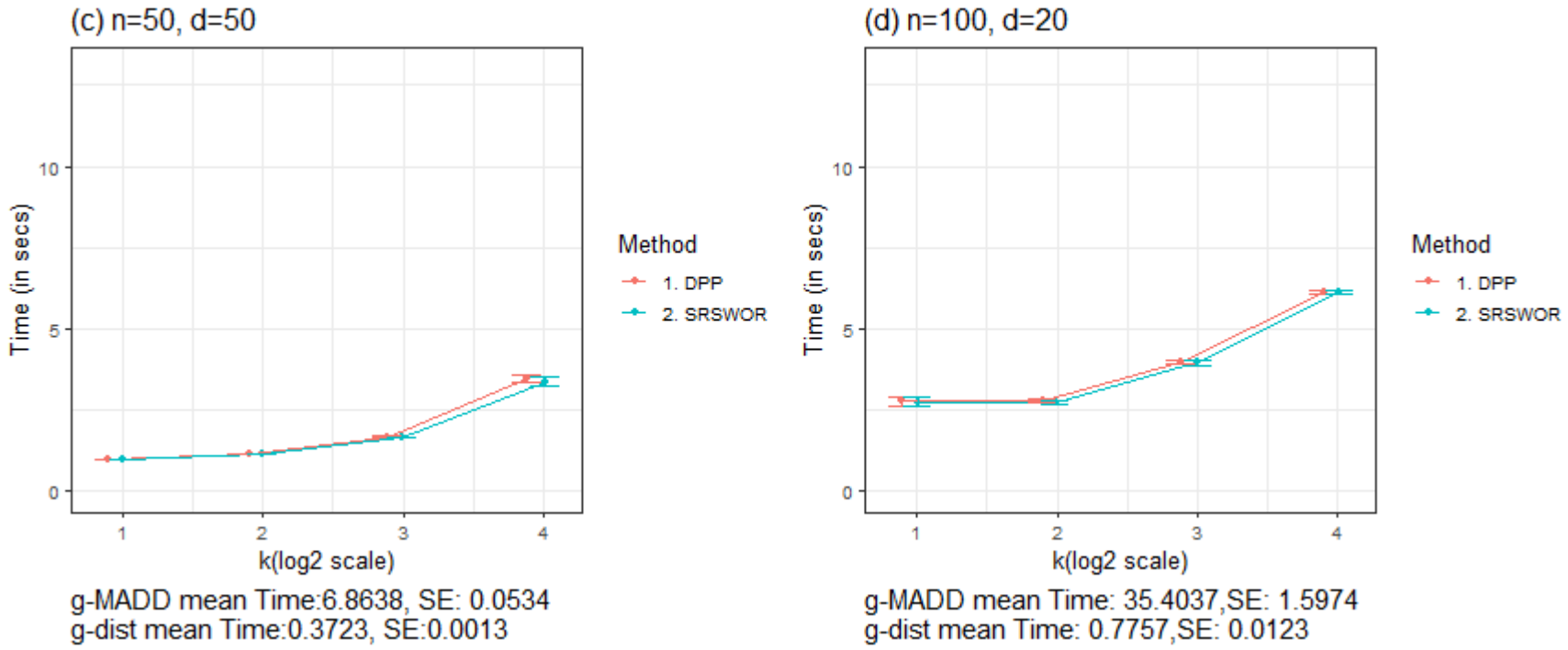


Figure: Comparison of computing times when the features are independent normal vs when the features are independent t (with same mean and dispersion). The reported numbers are averages `±` SE based on 100 replications.

Comparison of Computing Times

d= 20 d= 50 **d= 100**

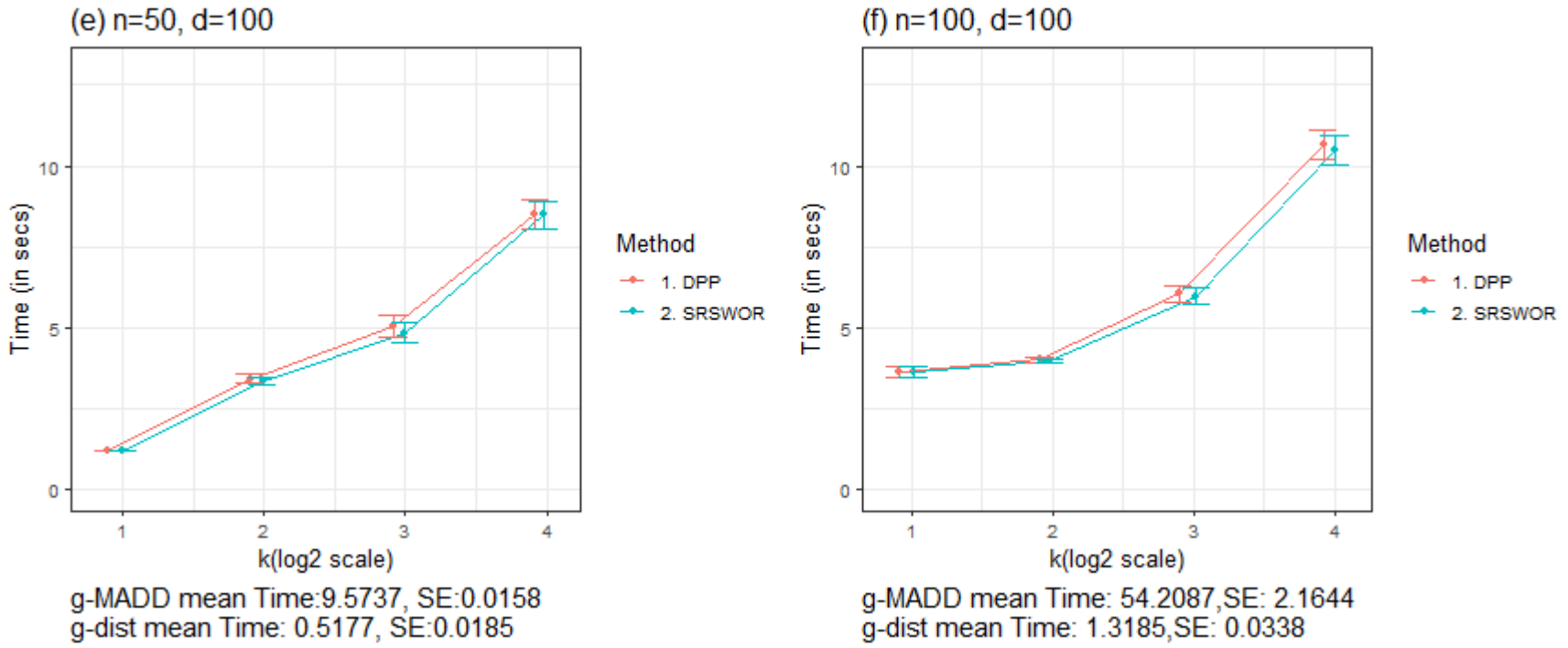
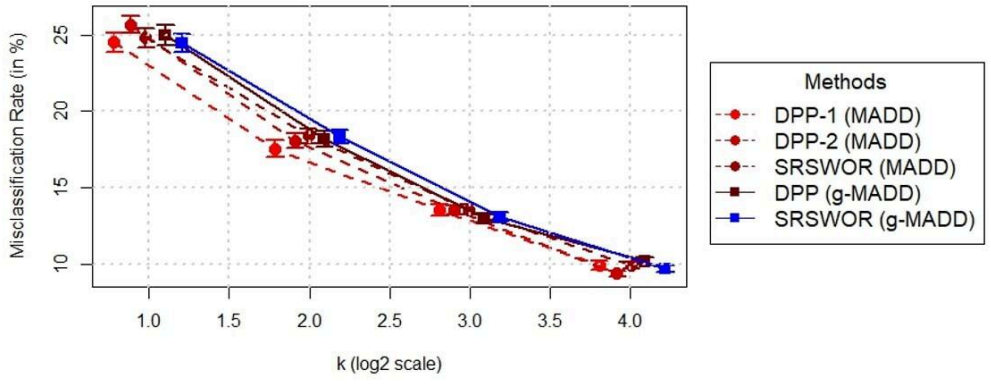
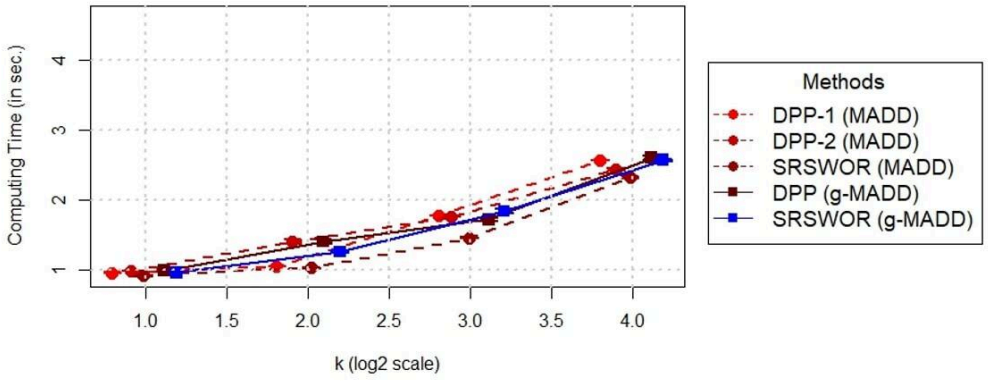


Figure: Comparison of computing times when the features are independent normal vs when the features are independent t (with same mean and dispersion). The reported numbers are averages `±` SE based on 100 replications.

Comparison of Scalable Version of MADD and g-MADD: Misclassification Rate and Computing Time

- Population 1 $\equiv N_d(\mathbf{0}, I_d)$ & Population 2 $\equiv N_d(0.5\mathbf{1_d}, I_d)$ ($n = 50$, $d = 50$)

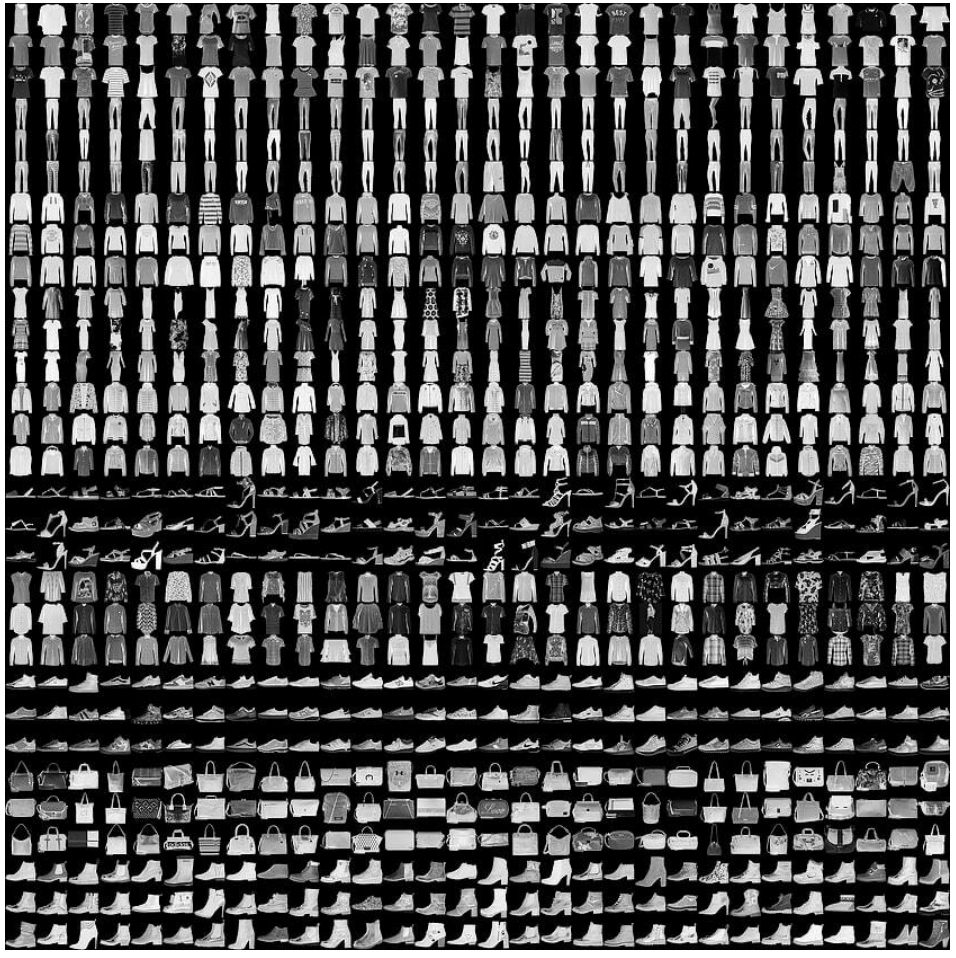




	Misclassification Rate (in %)	SE	Computing Time (in sec.)	SE
Traditional MADD	7.21	0.1521	5.9966	0.5592
Traditional g-MADD	7.26	0.1305	6.9581	0.7411

Application on Benchmark Dataset: Fashion MNIST

- Contains $60,000$ grayscale images of size 28×28 of the 10 fashion article classes.
- Also contains a test set of $10,000$ images.



Defining Distance Metric, MADD and g-MADD for Image Data

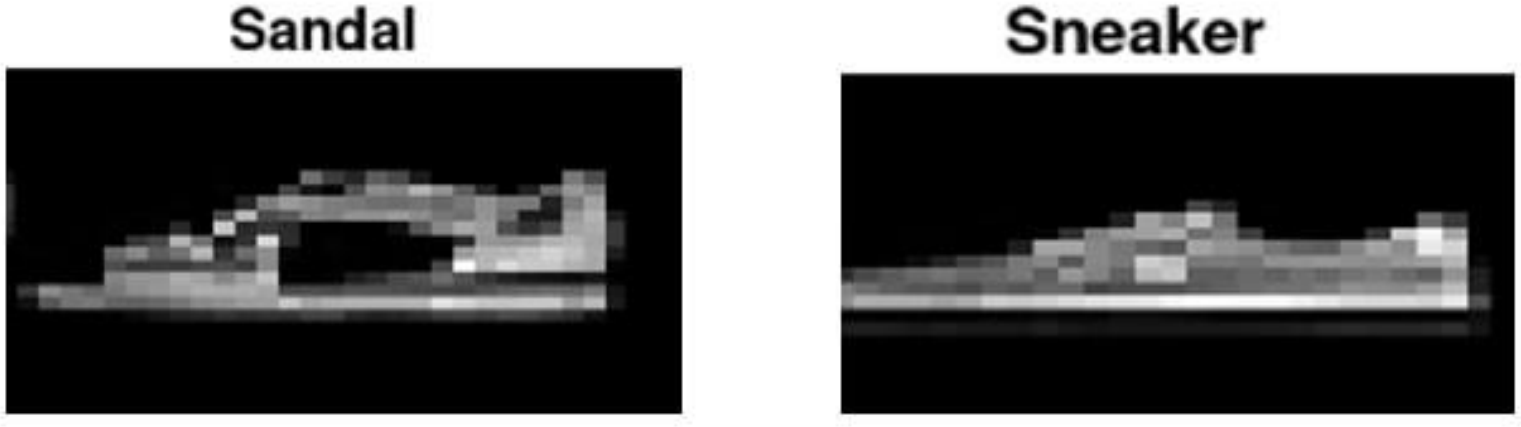
- Suppose each grayscale image has $p \times p$ pixels, totaling $D = p^2$ pixels.
- The distance between two images I_1 and I_2 is calculated as:

$$d(I_1, I_2) = \sqrt{\sum_{d=1}^{D} (I_1^d - I_2^d)^2}$$

- This formula treats pixel values as feature values for each image.
- It is indeed the usual distance after vectorizing the images.
- Following the same concept, MADD and g-MADD and their Scalable versions are also consistent with the usual ones.

Two-Class Classification Problem

- Randomly chosen 200 observations from `Sandal` and 200 observations from class `Sneaker` as training observations.



Results

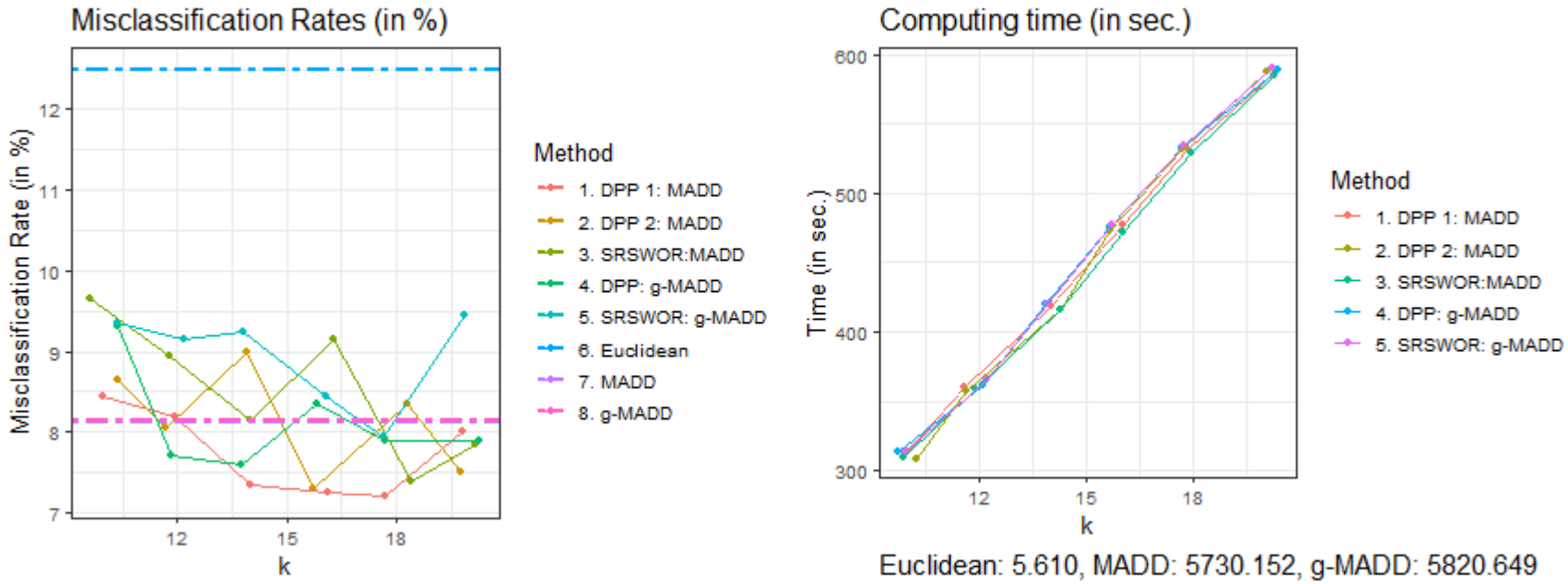
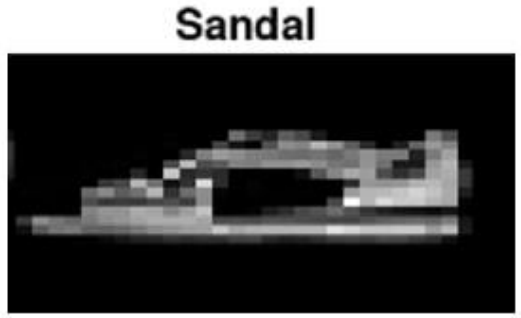
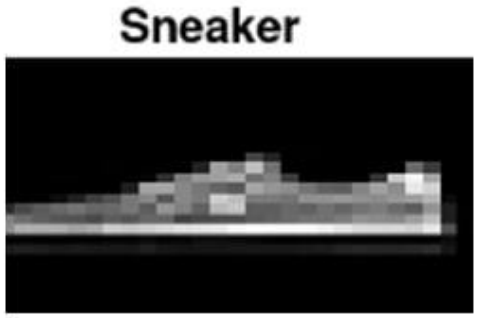


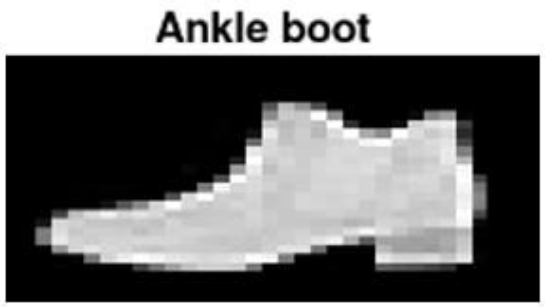
Figure: Misclassification rates (in %) (left) and computing time taken (in sec.) (right) for classification of Sandal and Sneaker using 200 training observations (from each class) from MNIST FASHION dataset.

Three-Class Classification Problem

- Randomly chosen 200 observations to construct the training set.







Results

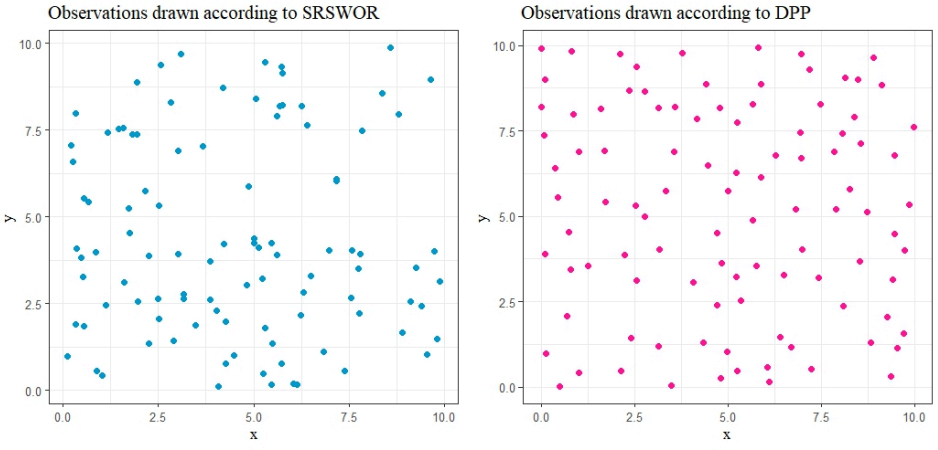
- Here, we have only tabulated the scalable version's performance with $k = 16$.
- We have not used g-MADD here, since previously, we have not seen any better performance using g-MADD.

Table: Misclassification Rate (in %) and Computing Time (in sec.) for classification of Sandal, Sneaker, Ankle boot using 200 training observations (from each class) from MNIST FASHION dataset.

	Misclassification Rate (in %)	Computing Time (in sec.)
DPP 1	11.80	1551.882
DPP 2	12.03	1554.737
SRSWOR	12.93	1543.262
Traditional MADD	12.40	18766.605
Euclidean	14.53	15.577

Concluding Discussions

DPP or SRSWOR?



- One may opt for the scalable version using DPP.
- DPP considers the diversity in the sample.
- DPP showed slight improvement in some cases.
- In real data analysis, DPP performed much better with the same k.
- Computing time for DPP and SRSWOR is very close.

Choice of k

- A crucial part of our Method is deciding the number of observations to draw from each population to balance accuracy and computing time.

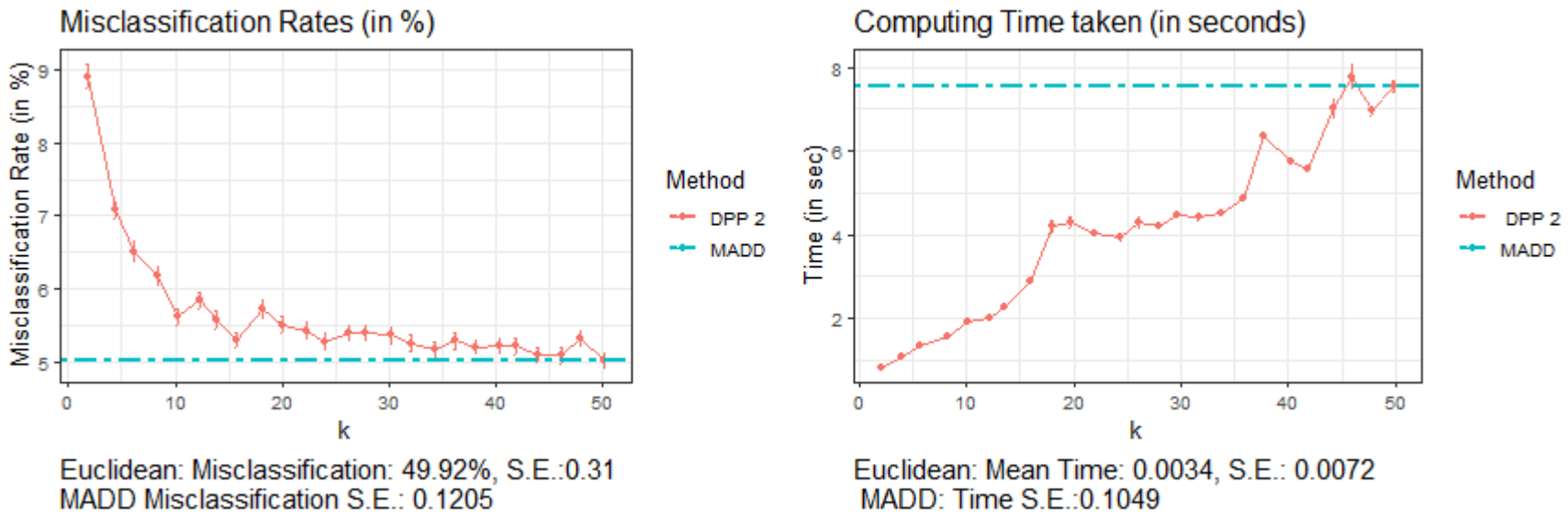


Figure: Misclassification rates (left) and computing time taken (in sec.) (right) for a pure scale problem with scalable version of MADD computed based on DPP-2. The reported numbers are averages `±` SE based on 100 replications.

Comparison with Existing Method for Dealing with Computational Issues

Pal et al. (2016) inspired from Condensed Nearest Neighbor (Hart 1968) or Reduced Nearest Neighbor (Gates 1972) used the following algorithm.

Algorithm CNN-type Algorithm for eliminating bad hubs from the data

Initialize $d_p(w)$ and $d_n(w)$ to 0 for each $w \in \mathcal{W}$
for each $w \in \mathcal{W}$ **do**
  for each $w' \in \mathcal{W}$ **do**
    if $w \neq w'$ and $MADD(w, w')$ indicates w is a neighbor of w'
      if w and w' belong to the same class
        $d_p(w) \leftarrow d_p(w) + 1$
      else
        $d_n(w) \leftarrow d_n(w) + 1$

Initialize $\mathcal{C} \leftarrow \emptyset$
for each $w \in \mathcal{W}$ **do**
  $d_{\text{total}}(w) \leftarrow d_p(w) + d_n(w)$
  if $d_{\text{total}}(w) > 0$ and $d_p(w) - d_n(w) > \eta$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{w\}$
Return $\mathcal{C}$

Comparison with Existing Method for Dealing with Computational Issues

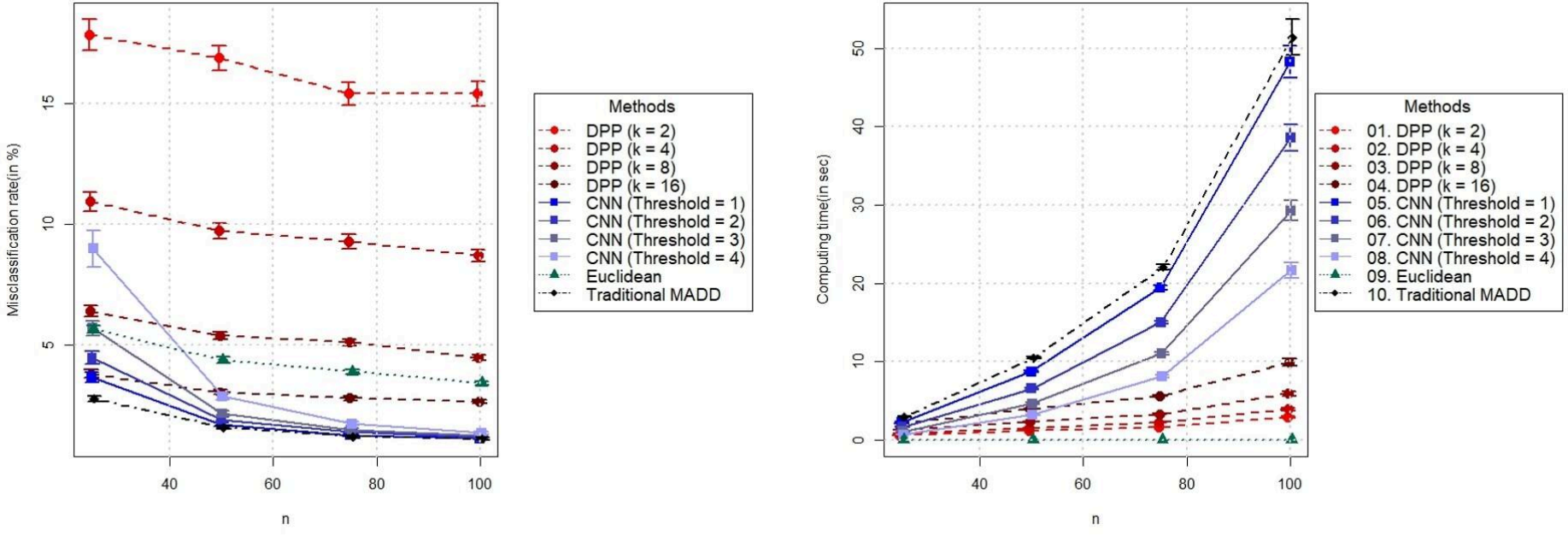


Figure: Misclassification rate (in %) (left) and computing time taken (in sec.) (right) for a pure location problem with scalable version of MADD computed based on DPP-2 and based on CNN with varying threshold.

References

- Gates, G. (1972). The reduced nearest neighbor rule (corresp.). *IEEE Transactions on Information Theory*, 18(3):431–433.
- Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B*, 67(3):427–444.
- Hart, P. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14(3):515–516.
- Pal, A. K., Mondal, P. K., and Ghosh, A. K. (2016). High dimensional nearest neighbor classification based on mean absolute differences of inter-point distances. *Pattern Recognition Letters*, 74:1–8.
- Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., and Jordan, M. (2002). Learning the Kernel Matrix with Semi-Definite Programming. *Journal of Machine Learning Research*, 5:323–330.

- Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B*, 67(3):427–444.

References

- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning. Springer Series in Statistics.* Springer, New York.
- Kulesza, A. and Taskar, B. (2012). Determinantal Point Processes for Machine Learning. *Foundations and Trends® in Machine Learning,* 5(2-3):123–286
- Sarkar, S. and Ghosh, A. (2019). On Perfect Clustering of High Dimension, Low Sample Size Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2257–2272
- Roy, S., Sarkar, S., Dutta, S., and Ghosh, A. K. (2022). On Generalizations of Some Distance Based Classifiers for HDLSS Data. *Journal of Machine Learning Research*, 23(14):1–41.

Thank You

All codes are available in Adrija211200/ScalableMADDR