RGC Ref. No.: UGC/FDS14/P04/17
(please insert ref. above)

RESEARCH GRANTS COUNCIL
COMPETITIVE RESEARCH FUNDING SCHEMES FOR THE LOCAL SELF-FINANCING DEGREE SECTOR

FACULTY DEVELOPMENT SCHEME (FDS)

Completion Report
(for completed projects only)

Submission Deadlines:
1. *Auditor's report with unspent balance, if any: within **six** months of the approved project completion date.*
2. *Completion report: within **12** months of the approved project completion date.*

Part A: The Project and Investigator(s)

1. Project Title

Supervised Dimensionality Reduction with Unsupervised Learning: Theory and Applications

2. Investigator(s) and Academic Department(s) / Unit(s) Involved

Research Team	Name / Post	Unit / Department / Institution
Principal Investigator	Dr. LAM Shu Yan/ Associate Professor	Department of Mathematics, Statistics and Insurance/The Hang Seng University of Hong Kong
Co-Investigator(s)	Prof. LIEW Wee Chung/ Professor	School of Information and Communication Technology/ Griffith University
Others		

3. Project Duration

	Original	Revised	Date of RGC / Institution Approval *(must be quoted)*
Project Start Date	01/01/2018	N/A	
Project Completion Date	31/12/2020	30/06/2021	17th, Aug, 2020
Duration *(in month)*	36	42	17th, Aug, 2020

Deadline for Submission of Completion Report	31/12/2021	30/06/2022	17th, Aug, 2020

Part B: The Final Report

5. Project Objectives

5.1 Objectives as per original application

1. Develop dimensionality reduction techniques for labelled data based on a novel clustering formulation that can overcome the major limitations of existing LDA methods.
2. Develop new algorithms that can provide robust estimations of class centroids, which are usually defined as the arithmetic means of the class data.
3. Develop new formulations that can perform sensitivity analyses of learnt projection vectors.
4. Develop new algorithms that incorporate robust estimators into dimensionality reduction formulations.
5. Develop efficient algorithms to apply the proposed methods to real-world recognition problems, including face recognition and object recognition.
6. Extend the proposed formulations to other types of dimensionality reduction problems, including, but not limited to, 2D, tensor and kernel formulations.

5.2 Revised objectives

Date of approval from the RGC: N/A

Reasons for the change:

1.

2.

3. ....

5.3 Realisation of the objectives

(Maximum 1 page; please state how and to what extent the project objectives have been achieved; give reasons for under-achievements and outline attempts to overcome problems, if any)

In the paper published in 2022 (**Benson S. Y. Lam** and Alan W. C. Liew, "A Fast Binary Quadratic Programming Solver based on Stochastic Neighborhood Search," *IEEE Transactions on Pattern Analysis and Machine Intelligence* [Impact factor: 17.73; rank 1/266 in Engineering, Electrical & Electronic], Vol. 44, Issue 1, pp. 32-49, 2022), we proposed a fast method to solve a class of binary quadratic programming problems. The method is faster than many state-of-the-art methods, while the solution qualities remain unchanged. Dimensionality reduction is one case of the quadratic programming problems. The proposed method adopts a two-stage iterative approach to find a local optimal solution. It first uses a gradient descent method to find a local optimal solution. Then, different weighted arithmetic means around the local optimal solution are obtained by random sampling (Objectives 1, 2 and 4). The gradient descent method is then applied to all the means and to find local optimal solutions. Only the best solution is kept. Then different weighted arithmetic means around this new solution are obtained by random sampling again. The above procedure is repeated many times. We applied the proposed method to different real-world image processing and pattern recognition problems (Objective 5). These are different types of dimensionality reduction problems (Objective 6). As the method involves a randomized procedure, we also conducted a sensitivity analysis and examined the robustness of the proposed method (Objective 3).

In the paper published in 2021 (**Benson S. Y. Lam**, M. F. Cheung and W. H. Lo, "What Matters Most in the Responses to Political Campaign Posts on Social Media: The Candidate, Message Frame, or Message Format?" *Computers in Human Behavior* [Impact factor: 5.003, rank 4/87 in (Psychology-Experimental), rank 12/138 in Psychology-Multidisciplinary], Vol. 121, 106800, 2021.), we proposed dimensionality reduction techniques to extract useful information from Facebook pages and to analyze an election campaign. We used partial least squares (PLS) to find the connections between Facebook emoji and the campaign posts. We applied cross-validation techniques to perform dimensionality reduction and selected the posts that are most relevant to the emoji. This paper addresses objectives 1, 3, 4 and 6.

In the under-review paper (Title: Bayes-optimal Linear Discriminant Analysis Based on Laplace Approximation and a Selective Sampling Scheme), we performed dimensionality reduction by minimizing a lower bound of the Bayes error, which is the theoretical error rate of a general classification problem. Laplace approximation is employed to estimate the density functions of the lower bound. Laplace approximation can be used to approximate any density function that has high concentration in a small domain. This implicitly formulates the class centroids (Objectives 1 and 2). The bootstrapping method together with a selection scheme is proposed to remove low-quality information (Objective 3). After that, the between-class scatter and within-class scatter matrices are computed and the projection matrix is found (Objective 4). We applied the method to different real-world problems and different types of dimensionality reduction problems (Objectives 5 and 6).

In the under-preparation paper (Title: A Novel Linear Discriminant Analysis based on a Between-class Distribution Modeling), we performed dimensionality reduction by minimizing the dominated part of the Bayes error. We formulated the Bayes error rate as a coupling distribution. We estimate the dominated part of the coupling distribution by the between-class distribution with high-density regions. We employed the sampling distributions of the sample median to model the dominated part (Objectives 1 and 2). The sampling distribution is constructed using bootstrapping (Objective 3). It is well-known that median is a robust estimator. Based on the median distributions, we compute the between-class scatter and within-class scatter matrices. Then, the projection matrix is obtained by solving another modified version of eigenvalue problem, which can perform heavy noise reduction and retain the discriminant information of the matrices (Objective 4). We applied the method to different real-world problems and different types of dimensionality reduction problems (Objectives 5 and 6).

5.4 Summary of objectives addressed to date

Objectives *(as per 5.1/5.2 above)*	**Addressed** *(please tick)*	**Percentage Achieved** *(please estimate)*
1. Develop dimensionality reduction techniques for labelled data based on a novel clustering formulation that can overcome the major limitations of existing LDA methods.	√	100%
2. Develop new algorithms that can provide robust estimations of class centroids, which are usually defined as the arithmetic means of the class data.	√	100%
3. Develop new formulations that can perform sensitivity analyses of learnt projection vectors.	√	100%
4. Develop new algorithms that incorporate robust estimators into dimensionality reduction formulations.	√	100%
5. Develop efficient algorithms to apply the proposed methods to real-world recognition problems, including face recognition and object recognition.	√	100%
6. Extend the proposed formulations to other types of dimensionality reduction problems, including, but not limited to, 2D, tensor and kernel formulations.	√	100%

6 Research Outcome

6.1 Major findings and research outcome *(Maximum 1 page; please make reference to Part C where necessary)*

Research outcome 1: **Benson S. Y. Lam** and Alan W. C. Liew, "A Fast Binary Quadratic Programming Solver based on Stochastic Neighborhood Search," *IEEE Transactions on Pattern Analysis and Machine Intelligence* [Impact factor: 17.73; rank 1/266 in Engineering, Electrical & Electronic], Vol. 44, Issue 1, pp. 32-49, 2022
Findings: The use of randomized arithmetic means can effectively find a local optimal solution of a class of the binary quadratic programming problems. The proposed method is 10 times faster than state-of-the-art methods. The solution qualities remain unchanged.

Research outcome 2: **Benson S. Y. Lam**, M. F. Cheung and W. H. Lo, "What Matters Most in the Responses to Political Campaign Posts on Social Media: The Candidate, Message Frame, or Message Format?" *Computers in Human Behavior* [Impact factor: 5.003, rank 4/87 in (Psychology-Experimental), rank 12/138 in Psychology-Multidisciplinary], Vol. 121, 106800, 2021.
Findings: The dimensionality reduction techniques can find the associations between the Facebook emoji and Facebook posts.

Research outcome 3: (Under-review work) Title: Bayes-optimal Linear Discriminant Analysis Based on Laplace Approximation and a Selective Sampling Scheme
Findings: Bayes error rate is the theoretical error rate of a general classification problem. The projection matrix that minimizing a lower bound of the rate performs better than state-of-the-art methods.

Research outcome 4: (Under-preparation work) Title: A Novel Linear Discriminant Analysis based on a Between-class Distribution Modeling
Findings: Bayes error rate is the theoretical error rate of a general classification problem. The projection matrix that minimizing the dominating part of the rate performs better than state-of-the-art methods.

6.2 Potential for further development of the research and the proposed course of action (Maximum *half a page)*

In this project, two of the works are to perform dimensionality reduction by minimizing the Bayes error rate. One is the lower bound of the Bayes error while the other one is a dominated structure of the Bayes error. This project only shows two of the ways to estimate the Bayes error and find the optimal projection matrix. It is possible to find more different ways to estimate Bayes error. Besides, we can incorporate different constraints to the Bayes error rate to deal with different situations. One example is the sparsity constraint. It can select useful features that can enhance classification rates.

7. Layman's Summary

(Describe in layman's language the nature, significance and value of the research project, in no more than 200 words)

In this project, we explored the theory and applications of dimensionality reduction techniques. We developed a new optimization scheme to solve a class of binary quadratic programming problems, which are also a kind of dimensionality reduction problems. The proposed method is 10 times faster than the state-of-the-art methods while the solution quality remains high. The method shows superiority performance in various image processing and pattern recognition problems. Besides, we developed new dimensionality reduction techniques that minimize the Bayes error rate. The Bayes error rate is the theoretical error rate of a general classification problem. We proposed two estimates to the theoretical error rate. One is a lower bound of the Bayes error rate while the other one is the dominated structure of the Bayes error. Experimental results show that the two estimates perform better than many different methods in many face recognition and data mining problems. Other than the above works, we also applied dimensionality reduction techniques to analyze Facebook information and found the associations between the Facebook emoji and Facebook posts.

Part C: Research Output

6 Peer-Reviewed Journal Publication(s) Arising Directly From This Research Project

(Please attach a copy of the publication and/or the letter of acceptance if not yet submitted in the previous progress report(s). All listed publications must acknowledge RGC's funding support by quoting the specific grant reference.)

The Latest Status of Publications				Author(s) *(denote the corresponding author with an asterisk*)*	Title and Journal / Book *(with the volume, pages and other necessary publishing details specified)*	Submitted to RGC *(indicate the year ending of the relevant progress report)*	Attached to this Report *(Yes or No)*	Acknowledged the Support of RGC *(Yes or No)*	Accessible from the Institutional Repository *(Yes or No)*
Year of Publication	**Year of Acceptance** *(For paper accepted but not yet published)*	**Under Review**	**Under Preparation** *(optional)*						
2021				Benson S. Y. Lam*, M. F. Cheung and W. H. Lo	Title: What Matters Most in the Responses to Political Campaign Posts on Social Media: The Candidate, Message Frame, or Message Format? Journal: *Computers in Human Behavior,121, 106800.* [Impact factor: 5.003, rank 4/87 in (Psychology-Experimental), rank 12/138 in Psychology-Multidisciplinary],	**No**	**Yes (Annex I)**	**Yes**	**Yes (https://researchdb.hsu.edu.hk/view/publication/20210100104)**
2022				Benson S. Y. Lam* and Alan W. C. Liew	Title: A Fast Binary Quadratic Programming Solver based on Stochastic Neighborhood Search, Journal: *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(1), 32-49.* doi: 10.1109/TPAMI.2020.3010811. [Impact factor: 17.73; rank 1/266 in Engineering, Electrical & Electronic], 44(1), 32-49.	**No**	**Yes (Annex 1I)**	**Yes**	**Yes (https://researchdb.hsu.edu.hk/view/publication/20210100103)**
		2022			Title: Linear Discriminant Analysis Based on Laplace Approximation and a Selective Sampling Scheme Journal: *Pattern Recognition*	**No**	**Yes (Annex III)**	**Yes**	**No**

			2022		Title: A Novel Linear Discriminant Analysis based on a Between-class Distribution Modeling	**No**	**No**	**Yes**	**No**

7 Recognized International Conference(s) In Which Paper(s) Related To This Research Project Was / Were Delivered

(Please attach a copy of each conference abstract)

Month / Year / Place	**Title**	**Conference Name**	**Submitted to RGC** *(indicate the year ending of the relevant progress report)*	**Attached to this Report** *(Yes or No)*	**Acknowledged the Support of RGC** *(Yes or No)*	**Accessible from the Institutional Repository** *(Yes or No)*
May/ 2020/ USA (Virtual)	Candidates, Message Frames or Forms: What Matters More in Political Expressions in Social Media Campaigns	*The 70th Annual Conference of International Communication Association (ICA)*	No	Yes (Annex IV)	Yes	Yes (https://researchdb.hsu.edu.hk/view/publication/202000251)

8 Whether Research Experience And New Knowledge Has Been Transferred / Has Contributed To Teaching And Learning

(Please elaborate)

We introduced some updates of dimensionality reduction techniques to two of the core modules of the BSc in Data Science and Business Intelligence. The two modules are AMS3640 Data Mining and AMS4640 Machine Learning.

9 Student(s) Trained

(Please attach a copy of the title page of the thesis)

Name	**Degree Registered for**	**Date of Registration**	**Date of Thesis Submission / Graduation**
N/A			

10 **Other Impact**
(e.g. award of patents or prizes, collaboration with other research institutions, technology transfer, teaching enhancement, etc.)

N/A

11 **Statistics on Research Outputs**

	Peer-reviewed Journal Publications	**Conference Papers**	**Scholarly Books, Monographs and Chapters**	**Patents Awarded**	**Other Research Outputs (please specify)**	
No. of outputs arising directly from this research project	2	1			Type	No.

12 **Public Access Of Completion Report**
(Please specify the information, if any, that cannot be provided for public access and give the reasons.)

Information that Cannot Be Provided for Public Access	**Reasons**
Nil	Nil