RGC Ref. No.:
UGC/FDS14/P02/18
(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

Model Selection with High Dimensional Incomplete Data

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

Research Team	Name / Post	Unit / Department / Institution
Principal Investigator	TANG, Man-Lai/Professor	Mathematics, Statistics and Insurance/The Hang Seng University of Hong Kong
Co-Investigator(s)	Nil	Nil
Others	Nil	Nil

3. Project Duration

	Original	Revised	Date of RGC / Institution Approval *(must be quoted)*
Project Start Date	01/01/2019		
Project Completion Date	31/12/2020	30/06/2021	24/11/2020
Duration *(in month)*	24 months	30 months	24/11/2020
Deadline for Submission of Completion Report	31/12/2021	30/06/2022	24/11/2020

Part B: The Final Report

5. Project Objectives

5.1 Objectives as per original application

1. **(Model selection based on information criteria)** A method, which minimizes an observed information criteria, is developed for high dimensional incomplete data. The proposed method is applied to variable selection in regression, generalized linear models and graphical model selection. Its numerical convergence is proved.

2. **(Graphical model selection based on multiple imputation)** Multiple imputation combined with stability selection is proposed for graphical model selection.

3. Simulation studies and real data analysis are conducted to evaluate the performance of the proposed method..

5.2 Revised objectives

Date of approval from the RGC: NA

Reasons for the change: NA

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)

Together with my collaborators, we have successfully developed the expectation model selection (EMS) algorithm to minimize the Bayesian information criterion (BIC) for latent variable selection in multidimensional item response theory models with a known number of latent traits. The result has been published in *British Journal of Mathematical and Statistical Psychology* (2022, **72**, 363–394) (i.e., fulfilled Objectives 1-3). We also develop variable selection procedures based on penalized estimating equations for competing risks quantile regression. The result has been published in *Statistics in Medicine* (2019;**38**:4670–4685) (i.e., fulfilled Objectives 1-3).

.

5.4 Summary of objectives addressed to date

Objectives *(as per 5.1/5.2 above)*	**Addressed** *(please tick)*	**Percentage Achieved** *(please estimate)*
1. **(Model selection based on information criteria)** A method, which minimizes an observed information criteria, is developed for high dimensional incomplete data. The proposed method is applied to variable selection in regression, generalized linear models and graphical model selection. Its numerical convergence is proved.	✓	100%
2. **(Graphical model selection based on multiple imputation)** Multiple imputation combined with stability selection is proposed for graphical model selection.	✓	100%
3. Simulation studies and real data analysis are conducted to evaluate the performance of the proposed method.	✓	100%
4.		

Research Outcome

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

The aim of latent variable selection in multidimensional item response theory (MIRT) models is to identify latent traits probed by test items of a multidimensional test. In this paper the expectation model selection (EMS) algorithm proposed by Jiang et al. (2015) is applied to minimize the Bayesian information criterion (BIC) for latent variable selection in MIRT models with a known number of latent traits. Under mild assumptions, we prove the numerical convergence of the EMS algorithm for model selection by minimizing the BIC of observed data in the presence of missing data. For the identification of MIRT models, we assume that the variances of all latent traits are unity and each latent trait has an item that is only related to it. Under this identifiability assumption, the convergence of the EMS algorithm for latent variable selection in the multidimensional two-parameter logistic (M2PL) models can be verified.We give an efficient implementation of the EMS for the M2PL models. Simulation studies show that the EMS outperforms the EM-based L1 regularization in terms of correctly selected latent variables and computation time. These results have been published in *British Journal of Mathematical and Statistical Psychology* (2022, **72**, 363–394).

The proportional subdistribution hazard regression model has been widely used by clinical researchers for analyzing competing risks data. It is well known that quantile regression provides a more comprehensive alternative to model how covariates influence not only the location but also the entire conditional distribution. In this paper, we develop variable selection procedures based on penalized estimating equations for competing risks quantile regression. Asymptotic properties of the proposed estimators including consistency and oracle properties are established. Monte Carlo simulation studies are conducted, confirming that the proposed methods are efficient. These results have been published in *Statistics in Medicine* (2019, **38**, 4670–4685).

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

To model count data with excess zeros, ones and twos, we introduce a so-called *zero-one-two- inflated Poisson* (ZOTIP) distribution, including the *zero-inflated Poisson* (ZIP) and the *zero and-one-inflated Poisson* (ZOIP) distributions as two special cases. We establish three equivalent stochastic representations for the ZOTIP random variable to develop important distributional properties of the ZOTIP distribution. The Fisher scoring and *expectation–maximization* (EM) algorithms are derived to obtain the maximum likelihood estimates of parameters of interest. The corresponding results may be extended to high dimensional incomplete counting data. These results have been published in Journal of Statistical Computation and Simulation (2021, DOI: 10.1080/00949655.2021.1970162)

7. Layman's Summary

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

High dimensional data analysis has become increasingly frequent and important in diverse fields; for example, genomics, health sciences, economics and machine learning. Model selection plays a pivotal role in contemporary scientific discoveries. There have been a large body of works on model selection for complete data. However, complete data are often not available for every subject due to many reasons, including the unavailability of covariate measurements and loss of data. The literature on model selection for high dimensional data in the presence of missing or incomplete values is relatively sparse. Therefore, efficient methods and algorithms for model selection with incomplete data are of great research interest and practical demand.

Part C: Research Output

8. 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 correspond-ing 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)_						
2019				Li, E, Tian, M. & **Tang, M. L.***	Variable selection in competing risks models based on quantile regression. *Statistics in Medicine*, 38(23), 4670–4685. https://doi.org/10.1002/sim.8326		Yes (Annex I)	Yes	Yes (https://research hsu.edu hk/project/?project_title=Model%20Selection%20with%20High%20Dimensional%20Incomplete%20Data)
2021				Sun, Y*, Zhao, S., Tian, G. L., **Tang, M. L.** & Li, T.	Likelihood-based methods for the zero-one-two inflated Poisson model with applications to biomedicine. *Journal of Statistical Computation and Simulation*. DOI: 10.1080/00949655.2021.1970162		Yes (Annex II)	Yes	Yes (https://research hsu.edu hk/project/?project_title=Model%20Selection%20with%20High%20Dimensional%20Incomplete%20Data)
2022				Xu, P. F., Shang, L., Zheng, Q. Z, Shan, N.* & **Tang, M. L.**	Latent variable selection in multidimensional item response theory models using the expectation		Yes (Annex III)	Yes	Yes (https://research hsu.edu hk/project/?project_title=Model%20Selection%20with%20High%20Dimensional%20Incomplete%20Data)

					model selection algorithm. *British Journal of Mathematical and Statistical Psychology*, 72(2), 363–394. https://doi.org/10.1111/bmsp.12261				

9. 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)*
N.A.						

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

(Please elaborate)

No. The results that have been developed are too technical to students in my University.

11. 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.			

12. Other Impact

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

N.A.

13. 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	3	0	0	0	Type	No.

14. 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
N.A.	