Reconstruction of Large-Scale Gene Regulatory Networks Using Regression-based Models

Faridah Hani Mohamed Salleh
Department of Software Engineering
College of Computer Science & IT
University of Tenaga Nasional
Jalan IKRAM-UNITEN
43000 Kajang, MALAYSIA

faridahh@uniten.edu.my

Suhaila Zainudin
Centre of Artificial Intelligence
Faculty of Information Sciences & Technology
Universiti Kebangsaan Malaysia
43650 UKM Bangi,
MALAYSIA

suhaila.zainudin @ukm.edu.my

Mohd Firdaus Raih
School of Biosciences & Biotechnology
Faculty of Science & Technology
Universiti Kebangsaan Malaysia
43600, Bangi, Selangor,
MALAYSIA
firdaus@mfrlab.org

***Abstract* -Gene regulatory networks (GRN) reconstruction is the process of identifying gene regulatory interactions from experimental data through computational analysis. GRN reconstruction-related works have boosted many major discoveries in finding drug targets for the treatment of human diseases, including cancer. However, reconstructing GRNs from gene expression data is a challenging problem due to high-dimensionality and very limited number of observations data, severe multicollinearity and the tendency of generating cascade errors. These problems lead to the reduced performance of GRN inference methods, hence resulting in the method being unreliable for scientific usage. We propose a method called P-CALS (Principal Component Analysis and Partial Least Squares) that is derived from the combination of PCA (Principal Component Analysis) with PLS (Partial Least Squares). The performance of P-CALS is assessed to the genome-scale GRN of E. coli, S. cerevisiae and an in-silico datasets. We discovered that P-CALS achieved satisfactory results as all of the sub-networks from diverse datasets achieved AUROC values above 0.5 and gene relationships were discovered at the most complex network tested in the experiments.**

Keywords: Principal Component Analysis; Partial Least Squares; gene regulatory networks; multivariate analysis

I. INTRODUCTION

The GRN inference-related works have boosted many major breakthroughs in finding drug targets for the treatment of human diseases. Due to the complexity of GRN, computational methods are used to identify and predict the interactions between a transcription factor and its target genes. However, some of the problems pertaining to these computational methods have not been adequately addressed. Firstly, most of the false positives errors made by the previous researches had been due to cascade errors. The accuracy of any GRN inference method will be tremendously improved if the cascade errors can be avoided. On top of that, the performance of some of the current prediction methods are affected by severe multicollinearity. Severe multicollinearity is a problem because it can increase the variance of the *coefficient estimates,* hence resulting the estimation become very sensitive to minor changes in the computational method. The result is that the coefficient estimates are unstable and difficult to interpret. Multicollinearity weakens the statistical power of the analysis, can cause the coefficients to switch signs, and makes it more difficult to specify the correct computational method. Next, an appropriate dimension reduction method is needed because less dimensions leads to less computing, hence resulting the usage of computationally expensive algorithm such as *NIPALS (NonLinear Iterative Partial Least Squares)* is possible in P-CALS. The name P-CALS is derived from **P**rincipal **C**omponent **A**nalysis and **P**artial **L**east **S**quares. We provide recommendations to solve all the above mentioned problems which are supported by several experiments that will be described in the subsequent sections. Next, most expression data contains the total number of observations very less compared to the number of genes ($n \leq p$, where n = observations and p = variables/genes) and developing a GRN inference method that can be applied to $n \leq p$ data is of great practical importance. The solutions to this problem are beneficial not only for gene expression data but also for datasets from other domain. We also used Martens' Uncertainty Test (MUT), which is a unique method based on "Jack-knifing" [1], [2]. MUT shows which variables are significant or not, the uncertainty estimates for the variables and the model robustness. We perform the computation of P-CALS using The Unscrambler X and the performance assessment was

978-1-5386-7128-3/18/$31.00 ©2018 IEEE

performed using MATLAB programs. We hope that our work will contribute to the study of data analytics-related research in general, and GRN inference in specific.

II. THE METHOD: P-CALS

GRN can be simplified as the complex interactions of components in an organism. A GRN consists of a set of DNA, RNA, proteins, and other molecules that describe regulatory mechanisms among these components. GRN represents the scenario where the predictor variables are likely to be correlated with each other and they could all influence the response variables. P-CALS works by producing the new variables from the common information of the original variables, where these variables represent the latent variables. These latent variables are assumed as a linear function for both the original variables and observations. Both matrices involved in the calculation of P-CALS are represented by X and Y, where X is the predictor and Y is a dependent variable.

PCA helps to identify how a sample does differs from the other, where the variables contribute to the majority of the difference and whether the variable correlated or independent from each other. The idea of implementation of PCA is to replace a complex multidimensional data to a simpler data, which involves less number of dimensions but at the same time fitting the original data to the best possible approximation [3]. The number of components run in our PCA model is 7. PCA calculates the amount of useful information and disregard the noise or insignificant variation contained in the datasets [4]. There are a few techniques can be applied to improve PCA model. In this experiment, we compare the methods to improve our PCA model by (1) deleting and (2) weighting the non-significant variables. Weighting allows the multiplication of selected variables by a very small number, such that the variables do not participate in the model calculation, but their correlation structure can still be observed in the scores and loadings plots particularly in the correlation loadings plot. PLS was tested at the beginning of this work because it suited the nature of GRN inference that requires the models to involve both the predictors and responses matrices simultaneously. PLS models both the predictors and responses to find the latent variables in predictors that will best predict the latent variables in responses [5]. The other selected recent works used PLS such as [6],[7] and [8]. The modified Jack-knifing method named Martens' Uncertainty Test (MUT) was used in this work has been invented by Harald Martens, and was published in [2]. MUT works with PLS with cross validation by assessing the significance of predictors and remove unimportant predictors from PLS model. There are several methods available to determine the weightage of variables, such as (1) setting the weight to a constant value independently from standard deviation, (2) standardize the variance of variables so that all variables will have equal level of influence for components estimation, (3) not setting any weightage to all variables, which means all computation is based on the raw data, (4) setting a low constant value as a weightage to all variables in such a way that all variables will not have any influence to the model [9],[10]. We eliminated the process called standardization, which weights A/(SDev + B) with A = 1.0 and B = 0.0 that used to give all the variables the same variance. This is because our data are noisy variables with small standard deviation that if they are standardized, their influence is exaggerated and will make the model less reliable. Knowing the nature of our data that contain irrelevant variables, we avoided either using all variables as they are or setting the weight to a constant number independently of the standard deviation. We discovered that using the *residual variance* to identity irrelevant variables was not efficient because all variables seemed to demonstrate equal level of importance to the calculation of our model. Thus, we propose to weight the variables with loadings ($-0.2 < PC\ 1 < 0.2$) $\cap$ ($0.2 > PC2 > -0.2$) because it had increased the performance when compared to the method that includes all variables into the model. In short, P-CALS is derived from two steps:

1) Starting from identifying the significant predictor variables performed by PCA ($X=T.P^T + E$) and
2) Then, the predictor variables are included into PLS together with the predictors identified by MUT in PLS model.

Variables that are non-significant display non-structured variation such as noise. If a suitable method is applied to eliminate or reduce the influence of the non-significant variables, the resulting model will be more stable and robust thus less sensitive to noise [11] and decreases the prediction error. We conducted several experiments before to identify the best method that can be performed to the identified non-significant variables such as weighting, removing and simply include all variables into the calculation. We found that weighting the variables is the most efficient way

of eliminating the irrelevant variables, without inadvertently removing the variables completely from the model. The non-significant variables can be identified by extracting the variables that have large residual variance [10]. Variables with large residual variance for all components or for the 3-4 first components have a small or moderate relationship with the other variables [12]. These variables have to be kept out from the calculation. However, our PCA model has less than 0.1% number of variables that have much larger residual variance than all the other variables in the model. All variables seem to demonstrate equal level of importance to the calculation of PCA, which is illogical. We identified the other way of extracting the non-significant variables, which is by focusing on the loadings. Weighting the variables make the process of studying the relationships between variables become possible and at the same time able to limit or reduce the influence of certain variables in the model. Down-weighting the variables cause the variables to have low influence to the model. After identifying the insignificant variables, model will rebuilt with only the significant variables will be taken into account to produce a better model. Apart from identifying the relationships using quadrant-based approach, we applied the other way of identifying the genes relationships, which is by focusing on the variables with high loadings along the same PC. Among all results produced from PCA, the loadings contribute the most to the analysis as it describes the data structure in terms of variable correlations. Each variable has a loading on each PC. The loading on PC indicates how much the variable contributed to that PC and how well that PC takes into account the variation of that variable over the data points [13]. In geometrical terms, a loading is the cosine of the angle between the variable and the current PC. The smaller the angle (the higher the link between variable and PC), the larger the loading [14]. The loadings range between –1 and +1 [15]. We identify the relevant variables by selecting variables with high loadings; (0.2< PC 1<-0.2) U (-0.2> PC2 >0.2). The experiments that present the comparison of performance between the quadrant-based and high loadings along the same PC is shown in TABLE II.

III. THE DATASETS

We perform our predictions on the data provided by the DREAM5 challenge [16] and [17]. The structures are shown in TABLE I. M3D provided manually curated metadata for their chip measurements. The expression data can be obtained from http://m3d.mssm.edu/. As for the DREAM5, this is one of the challenges in DREAM project, which is a framework to enable such an assessment through standardized performance metrics and common benchmarks (http://www.the-dream-project.org/). DREAM5 datasets consist of various organisms with different level of complexity. The predicted networks were evaluated based on AUROC and AUPR. The experiment that assessed the performance of P-CALS in predicting GRN was extended to assess the ability of P-CALS in avoiding cascade error by using a novel experimental procedure that we proposed in our previous work [18].

TABLE 1. THE PROPERTIES OF DATASETS USED IN EXPERIMENTS

Network label/ Organism	Transcription Factors	Number of Genes	Number of Chips/ Samples
Net1(in-silico)	195	1643	805
Net3 (E. coli)	334	4511	805
Net4 (S. cerevisiae)	333	5950	536
M3D	Not given	4297	907

IV. RESULTS AND ANALYSIS OF THE EXPERIMENTS

There are 3 main methods (not including the proposed method) that were tested in our work. PCA has various ways of determining the gene relationships such as identifying the variables with high loadings along the same PC and quadrant-based method. The techniques to reduce dimensionality are varied as well and we chosen weighted the insignificant variables with loadings (0.2< PC 1<-0.2) ∩ (-0.2> PC2 >0.2). Due to the variations in methods of implementing PCA and dimension reduction techniques, we conducted 3 PCA experiments to assess the performance of each of these methods. As for PLS, we aim to compare the performance when PLS was used with MUT and without MUT. On the other hand, experiment using MLR had been described in our previous publication [18] on M3D datasets. As seen in TABLE II, the initial results obtained from the experiment performed using PCA were not promising with most of the AUROCs are less than 0.5. In contrast with the other experiment performed using PLS, PLS with MUT had proved to be better than the PLS alone. On the other hand, the combination of PCA and PLS that we called as P-CALS yielded the highest AUROC and AUPR values in all sub-networks compared to the previous experiment. The highest and almost perfect

identification was obtained for Net4, which is the most complex sub-networks among all. However, there were no significant difference at AUPR. The positive results obtained by PLS-MUT, particularly at AUROC of Net4, has motivated us to further explore the potential of PLS-MUT by focusing on the numbers of significant variables that was included in the PLS model calculation. As seen in TABLE II, the performance of PCA and PLS were slightly lower compared to P-CALS. The results proved that PLS showed better performance than PCA. With an intention to fully utilized the identified significant variables, we had tried to re- run PLS using the significant variables identified from both PLS and PCA but the obtained results were very poor. We found that P-CALS worked well if the predictions of PCA were combined with predictions done by PLS, with each of these models performed its calculations separately. Next, we conducted an experiment to assess the performance of P-CALS when dealing with cascade motifs.

TABLE II. THE RESULTS OF THE EXPERIMENTS USING PCA (WITH DIFFERENT WAY OF IMPLEMENTATIONS), PLS, MLR AND P-CALS.

The Methods	AUROC M3D	AUROC Net1	AUROC Net3	AUROC Net4	Average AUROC	AUPR Net1	AUPR Net3	AUPR Net4	Average AUPR
PCA1	-	0.5240	0.4710	0.4820	***0.4923***	0.0170	0.0120	0.0170	***0.0153***
PCA2	-	0.5160	0.4830	0.4820	***0.4937***	0.0150	0.0120	0.0170	***0.0147***
PCA3	-	0.4980	0.4920	0.4980	***0.496***	0.0140	0.0130	0.0170	***0.0147***
PLS with Marten Uncertainty Test (MUT)	-	0.5230	0.5440	0.5860	***0.5510***	0.0160	0.0170	0.0230	***0.0187***
PLS	-	0.5330	0.5200	0.5260	***0.5263***	0.0160	0.0140	0.0190	***0.0163***
MLR (Source:[18])	0.5580	*	*	*	*	*	*	*	*
P-CALS	0.6500	0.522	0.6013	0.6435	***0.5889***	0.016	0.0342	0.0233	***0.0735***

Note:
*** PCA and PLS were performed using various ways of implementation and dimensionality reduction techniques. We chose not to further used MLR in Net1, Net3 and Net4 as MLR had been proven unable to calculate the whole network and less effective when performed on M3D.*
*** PCA1: recalculated after insignificant variables down-weighted, quadrant-based, PCA2: not recalculated, quadrant-based, PCA3: recalculated after insignificant variables down-weighted, high loadings along the same PC.*
*****Experiments marked with asterisks were not conducted because MLR is proven to be unable to work with n>=p data.*

TABLE III presents the number of cascade errors made by each of the tested methods. Even though PCA obtained 0% for cascade errors, this did not reflect the actual performance as the number of predictions made by PCA was too small compared to the other techniques. In addition, AUROC obtained by PCA was lower than PLS. P-CALS obtained the lowest number of cascade errors among all the tested methods. TABLE IV shows the comparison of P-CALS with the other 5 selected methods, where the results were reported by [7]. P-CALS outperformed other methods at Net 4 (the most complex network). From the results, we conclude that when the wrong predictions made due to cascade errors is at a minimal level, there is a high chance for the improved GRN inference method performance. This has been proven by the P-CALS experiments performed on Net 4 that were all correct in its predictions.

V. DISCUSSION

We found that PCA worked well with gene expression data to reduce dimensionality by identifying the significant variables for calculation. As a way to reduce data dimensionality, we used PCA to identify the variables with low loadings and give these variables a very low weight in the analysis, so that they will lose any influence they might have on the model. Removing the variables with low loadings ($0.2< PC 1<-0.2) \cap (-0.2> PC2 >0.2$) from PCA was ineffective as the whole calculation was interrupted, hence giving the imprecise results of prediction. The PCA experiments proved that the variables with high loadings along the same PC were connected to each other's, and using this technique to identify genes relationship is more efficient than quadrant-based method. Even though NIPALS is computationally expensive (in comparison with SVD), we found that it gives more numerically accurate results for calculating the principal components of a data set. In addition, we

are able to use complex algorithm NIPALS because our dimension reduction method was effective to reduce the number of variables to be processed. In addition, the dimension reduction method takes care of multicollinearity that improves the model performance and helpful in noise removal as well. Despite of all advantages of PCA, it cannot predict the value of one variable if the values of other variables are not known while this prediction is possible in regression analysis as it establishes a functional relationship between the variables. While PCA need not imply cause and effect relationship between the variables under study, regression analysis presumes the cause and effect relationship. Therefore, we decided to test PLS that handles multicollinearity better. The important finding discovered from PLS is the usage of MUT that had increased the performance of PLS far much better than PCA. MUT, based on Jack-kniffing was effective in identifying the significant variables for PLS model.

TABLE III. CASCADE ERRORS EVALUATION

	Net 1	**Net 3**	**Net 4**	**M3D**
Total cascade motifs	5,955	1,619	4,300	949
Total PCA predictions	381	125,695	187,150	-
Total PCA cascade errors	0	71	321	-
%	**0.00**	**4.39**	**7.47**	-
Total PLS-MUT predictions	596	489,386	1,956,473	-
Total PLS-MUT cascade errors	0	0	3271	-
%	**0.00**	**0.00**	**76.07**	**-**
Total P-CALS predictions	69,197	576,379	1,282,301	3,543,442
Total P-CALS cascade errors	5	430	1070	98
%	**0.08**	**26.7**	**24.88**	**10.33**
Total MLR SET 3 predictions	*	*	*	25,502
Total MLR SET 3 cascade errors	*	*	*	94
%	*	*	*	**4.75**

*** Note = We did not conduct further experiments marked with dashes (-) because the experiments conducted on Net1, Net3 and Net4 had proved PCA and PLS-MUT were less inefficient compared to P-CALS. The experiments marked with asterisks (*) were not conducted because the experiment performed on M3D datasets had proved that MLR was unable to calculate n < p data.*

TABLE IV. THE RESULTS OF THE EXPERIMENTS USING P-CALS AND THE COMPARISON WITH THE OTHER ESTABLISHED METHODS

The Methods	**AUROC M3D**	**AUROC Net1**	**AUROC Net3**	**AUROC Net4**	**AUPR Net1**	**AUPR Net3**	**AUPR Net4**
P-CALS	0.6000	0.5220	0.6013	**0.6435**	0.0160	0.0342	**0.0233**
GENIE3 (Irrthum, Wehenkel et al. 2010)	0.6730	**0.8150**	0.6170	0.5180	0.2910	0.0930	0.0210
ANOVA (Küffner, Petri et al. 2012)	**0.7980**	0.7800	**0.6710**	0.5190	0.2450	**0.1190**	0.0220
CLR (Faith, Hayete et al. 2007)	0.6420	0.7730	0.5900	0.5160	0.2550	0.0750	0.0210
ARACNE (Margolin, Nemenman et al. 2006)	0.6350	0.7630	0.5720	0.5040	0.1870	0.0690	0.0180
TIGRESS (Haury, Mordelet et al. 2012)	Not applicable	0.7890	0.5890	0.5140	**0.3200**	0.0660	0.0200

Note: The bold values indicate the highest in each evaluation matrices.

In comparison with the other established methods, AUROC and AUPR of P-CALS outperformed other methods at Net4, which is the most complex network with the largest total number of genes. P-CALS identifies more than 40% of the correct gene relationships at 2 out of 4 networks and more than 80% correct for the most complex network. However, the AUROC was not really high most probably because of the large number of predictions made by P-CALS that will indirectly increase the number of false positives. Another notable finding is that, the performance of all the tested methods were poor when conducted on synthetic datasets; this indicated a need to further study the method used to generate the synthetic datasets. We discovered that cascade errors in datasets would be evaded if all the variables are calculated in the model simultaneously, with attention is given for both sides; from predictors to responses and vice versa. Despite of the advantages, P-CALS has several weaknesses. A large number of predictions may be good to ensure all possible relationships are taken into account. However, a very large number of predictions

have exposed P-CALS to the occurrences of large number of false positives, which resulting low AUROC.

VI. CONCLUSIONS

In summary, we proposed P-CALS that combines the predictions of PCA and PLS. PLS calculations in P-CALS summarize the information of the correlated variables where the components are determined depending on their relation with the dependent variable to be studied (relation between responses and predictors). Different from PLS, PCA calculations in P-CALS complement the PLS by determining the components regarding the relation inside the predictors block only. In order to extend the usage of PCA to the fullest capacity, we had made a number of improvements such as down-weighted the identified non-significant variables and determined correlations of genes with high loadings along the same PCs. We applied all these techniques of improvements because P-CALS had demonstrated the increased of performance when applied the identified techniques. We proposed P-CALS with Marten Uncertainty Test (MUT) based on "Jack-knifing" to allow the estimation of the model stability, the identification of perturbing samples or variables, and the selection of significant predictors. Based on the experiment results, P-CALS is able to avoid wrongly inferring of an indirect interaction as a direct interaction. P-CALS solves the $n \leq p$ problem by considering the whole network of all identified relevant variables simultaneously by switching between response variables and predictors iteratively to find the relevant PCs. However, P-CALS generated a very large number of predictions, which have exposed P-CALS to the occurrences of large number of false positives, which resulting low AUROC. Despite the limitation, our proposed method clearly has some worthy advantages. In this work, we examined our developed methods and learned from the experience to improve the performance of existing methods in application to data analysis-related problems. We are hoping to use some of the tools and techniques that we have applied in this work to other large-scale and stochastic data analysis.

REFERENCES

[1] Efron, B. (1982). The jackknife, the bootstrap and other resampling plans, SIAM.

[2] Martens, H. and M. Martens (2000). "Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR)." Food quality and preference **11**(1): 5-16.

[3] Seierstad, T., et al. (2008) "Principal component analysis for the comparison of metabolic profiles from human rectal cancer biopsies and colorectal xenografts using high-resolution magic angle spinning 1H magnetic resonance spectroscopy." Molecular Cancer: 7:33.

[4] Wang, J.,.(2012) "Principal Component Analysis." Geometric Structure of High-Dimensional Data and Dimensionality Reduction: 95-114.

[5] Geladi, P. and kowalski, B.,R., (1986) "Partial least-squares regression: a tutorial". Analytica Chimica Acta: 1-17

[6] Chan, S.-C., et al. (2012). "A new method for preliminary identification of gene regulatory networks from gene microarray cancer data using ridge partial least squares with recursive feature elimination and novel Brier and occurrence probability measures." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on **42**(6): 1514-1528.

[7] Guo, S., et al. (2016). "Gene regulatory network inference using PLS-based methods." BMC Bioinformatics **17**(1): 545.

[8] Chan, S.-C., et al. (2016). "A maximum a posteriori probability and time-varying approach for inferring gene regulatory networks from time course gene microarray data." Computational Biology and Bioinformatics, IEEE/ACM Transactions on **12**(1): 123-135.

[9] Esbensen, K. H., et al. (2002). Multivariate data analysis: in practice: an introduction to multivariate data analysis and experimental design, Multivariate Data Analysis.

[10] Johnson, R. A. and D. W. Wichern (2014). Applied multivariate statistical analysis, Prentice-Hall New Jersey.

[11] Freund, R. and W. Wilson (1998). Regression Analysis: Statistical Modeling of a Response Variable, Academic Press.

[12] CAMO (2003). "The Unscrambler Method References." Retrieved 1/6/2016, 2016.

[13] Martin, M.,Z., (2005) "Analysis of preservative-treated wood by multivariate analysis of laser-induced breakdown spectroscopy spectra." Spectrochimica Acta Part B: Atomic Spectroscopy: 1179-1185.

[14] Ulmschneider, M. and Roggo, Y., (2007), Process Analytical Technology, John Wiley & Sons.

[15] Iranmanesh, V., et al. (2014). "Online handwritten signature verification using neural network classifier based on principal component analysis." The Scientific World Journal **2014**.

[16] Marbach, D., et al. (2012). "Wisdom of crowds for robust gene network inference." Nat Meth **9**(8): 796-804.

[17] Faith, J. J., et al. (2008). "Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata." Nucleic Acids Research **36**(suppl 1): D866-D870.

[18] Salleh, F. H. M., et al. (2017). "Multiple Linear Regression for Reconstruction of Gene Regulatory Networks in Solving Cascade Error Problems." Advances in Bioinformatics **2017**.