Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Imputing missing values for genetic interaction data

Yishu Wang [a], Lin Wang [a], Dejie Yang [b], Minghua Deng [a,c,d,*]

[a] Center for Quantitative Biology, Peking University, Beijing 100871, China
[b] Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China
[c] School of Mathematical Sciences, Peking University, Beijing 100871, China
[d] Center for Statistical Sciences, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 2 October 2013
Accepted 27 March 2014
Available online xxxx

Keywords:
Soft-SVD
Imputation
EMAP
Genetic interaction

ABSTRACT

Background: Epistatic Miniarray Profiles (EMAP) enable the research of genetic interaction as an important method to construct large-scale genetic interaction networks. However, a high proportion of missing values frequently poses problems in EMAP data analysis since such missing values hinder downstream analysis. While some imputation approaches have been available to EMAP data, we adopted an improved SVD modeling procedure to impute the missing values in EMAP data which has resulted in a higher accuracy rate compared with existing methods.
Results: The improved SVD imputation method adopts an effective soft-threshold to the SVD approach which has been shown to be the best model to impute genetic interaction data when compared with a number of advanced imputation methods. Imputation methods also improve the clustering results of EMAP datasets. Thus, after applying our imputation method on the EMAP dataset, more meaningful modules, known pathways and protein complexes could be detected.
Conclusion: While the phenomenon of missing data unavoidably complicates EMAP data, our results showed that we could complete the original dataset by the Soft-SVD approach to accurately recover genetic interactions.

1. Introduction

Genetic interactions refer to the phenomenon whereby the mutation phenotype of two genes differs to the superimposition effect of two single mutations [1]. In budding yeast and fission yeast, genetic interactions can be acquired using the high-throughput technology known as the Epistatic Miniarray Profile (EMAP) platform [2]. EMAP can construct double deletion strains systematically by crossing query strains with a library of test strains. Afterwards, the colony size of the double mutant strains is measured to get the S score, which can indicate the genetic interaction, either synthetic sick/lethal or alleviating [3]. In EMAP datasets, each gene in the query, or library, has its genetic interaction spectrum constructed by the genetic interaction S score with other genes in the library, or query. Researchers can exploit biological pathways and reveal protein complexes by clustering the S score matrix and, as a result, find cellular organization and gene functions.

However, one common characteristic of running EMAP is the significantly high proportion of missing values, even up to 35%, which can reduce the effectiveness of such data analysis techniques as cluster analysis and even prevent the use of some matrix factorization techniques, such as SVD or PCA. This phenomenon of missing entries could be explained by the inability of high-throughput technologies to measure genetic interaction strengths. Also, some genetic interactions could be subsequently filtered as a result of unreliability.

The problem of missing values in genetic interaction datasets has been discussed before, but few technologies are used to impute quantitative epistasis values in EMAP datasets [4]. Some previous papers reported improvements in some techniques used in gene expression datasets, subsequently applying them to EMAP data. Four general strategies are considered in EMAP data, including three nearest-neighbor-based: k-Nearest Neighbors imputation (kNN) [5], Local Least Squares imputation (LLS) [6], and Iterated Local Least Square imputation (iLLS) [7]; and one global method known as Bayesian Principal Component Analysis imputation (BPCA) [8]. The term "local" used here represents those algorithms that impute missing values through local information around the missing value. Previously, in order to improve the accuracy of

* Corresponding author at: School of Mathematical Sciences, Peking University, Beijing 100871, China.
E-mail address: dengmh@math.pku.edu.cn (M. Deng).

http://dx.doi.org/10.1016/j.ymeth.2014.03.032
1046-2023/

missing value predictions, these original imputing techniques have incorporated the symmetric character of datasets [4,9]. However, with the recent development of EMAP technology and the need for practical application, most EMAP datasets are asymmetric[10–12]. Therefore, the symmetric characteristic is not appropriate for use in predicting missing entries on an increasing number of new EMAP experimental datasets. We extended the original SVD method by giving it a soft threshold, changing some optimization functions and restricting conditions [13]. This method which can be called Soft-SVD has been used in "Netflix" competition [13], image recovery [13] and eQTL [14], and it has been demonstrated as the most efficient algorithm in these fields. The soft-SVD algorithm is not restricted by symmetry. As such, it can be used in a wide range of EMAP datasets. We introduced this methodology to impute missing values in EMAP datasets, and, hereinafter, we call it Soft-Impute.

The Soft-Impute methodology adopts the soft threshold to SVD algorithm and proposes that the nuclear norm results in a convex optimization problem. It takes advantage of the relevance in a given dataset to impute missing entries by finding a low-rank matrix that is close to the original one on the observed data. This algorithm is suitable for datasets in which modules are found with highly correlated entries. In Part 4, we constructed a synthetic dataset with low-rank matrix to test the effect of the Soft-Impute algorithm and compared it with the imputation accuracy of different imputation methods. EMAP datasets store genetic interaction spectra, where genes in the same protein complex or biological pathway tend to have similar genetic interaction spectra. Accordingly, EMAP data matrices contain several highly correlated modules. As a whole, the matrix is low-rank, which satisfies the request of the Soft-Impute algorithm for datasets.

In this paper, we systematically described how the Soft-Impute method is applied to EMAP dataset imputation and then applied this method, along with four other imputation methods, to three public EMAP datasets. We then conducted a detailed comparison among these imputation techniques, showing the marked improvement of the Soft-Impute algorithm in the performance of missing entries estimation. Beyond imputation accuracy, we also evaluated these methods in terms of their ability to detect genetic interaction modules in which genes have similar interaction profiles, are involved in the same physical complex or pathway, and are also enriched in GO terms. After imputing missing entries in the EMAP score matrix, we demonstrated the downstream analysis in which hierarchical clustering results were highly improved, and more significant genetic interaction modules, which are enriched in the known discoveries, could be exploited.

2. Materials

2.1. Epistatic Miniarray Profile (EMAP)

The genetic interaction datasets used here are from EMAP analysis. Three EMAP datasets are used in our analysis, including the early secretory pathway (ESP) [15], chromosome biology (CHR) [16] for budding yeast, and a genome-wide EMAP profile for fission yeast [17]. The first two datasets are symmetric matrices for which query genes are the same as library genes. There are 424 genes with about 80,000 pairwise measurements in the ESP dataset, containing about 7.5% missing entries, while there are 743 genes with about 200,000 pairwise measurements in the CHR dataset, containing about 34% missing entries. On the contrary, the genome-scale genetic interaction matrix of fission yeast is not symmetric and contains 953 alleles of 876 genes against a mutant library of more than 2000 deletions, resulting in an EMAP profile with 1.6 million genetic interactions and 16% missing entries.

2.2. Synthetic data

We created a synthetic dataset with low-rank to realize the Soft-Impute algorithm and compared it with other imputation methods. We assumed k modules in the synthetic dataset, in which entries in the same module have higher relevance. In contrast, entries not in the same module have lower relevance. We constructed a dataset of 250 elements representing query genes in 500 dimensions standing for library genes as follows:

1. We first constructed a vector of 500 elements randomly chosen from ESP dataset. as Denoted by $\overrightarrow{A_1} = \{a_1, a_2, \ldots, a_{500}\}$
   (a) Multiply by Gaussian noise to every element a_i of $\overrightarrow{A_1}$, which is randomly chosen from $N(1, |a_i|)$. and denoted by $\overrightarrow{A_2}$.
   (b) Repeat (a) for k times which results in k vectors $\{\overrightarrow{A}_1, \ldots \overrightarrow{A}_k\}$
2. Generate a matrix with rank k using the above k vectors.
   (a) Generate k random numbers $n_1, n_2, \ldots, n_k$, ranging from 3 to 30, such that their sum is 250.
   (b) Generate a matrix with 250 vectors by repeating vector $\overrightarrow{A_1}$ for n_1 times, vector $\overrightarrow{A_2}$ for n_2 times,..., vector $\overrightarrow{A_k}$ for n_k times. Apparently, such matrix is of rank k.
3. Add a Gaussian noise drawn from $N(0, 0.5)$ to each entry of the above matrix.
4. Now we construct a matrix of 250 vectors with 500 dimensions.

3. Model and algorithm

3.1. Soft-Impute model

EMAP data can be represented by a matrix $\mathbf{X}_{m\times n}$, where m and n represent the number of query genes and library genes. To represent the existence of missing entries in EMAP dataset matrix, $\Omega \subset \{1, \ldots, m\} \times \{1, \ldots, n\}$ denotes the indices of observed entries. Therefore, $\mathbf{X}$ is the original data with observed entries denoted by Ω and missing values denoted by $\Omega^{\perp}$. To impute this matrix, we aim to find a complete matrix $\mathbf{Z}$, which is close to $\mathbf{X}$ on the observed entries Ω and has low rank. Here, the low- rank assumption is based on the consideration that the genetic interaction profile for cofunctional genes is shown to have highly correlated relationships [15,16]. Srebro et al. have studied generalization error bounds for learning the low-rank matrices [18], Their work also showed, theoretically, that the true underlying matrix could be recovered with very high accuracy under certain assumptions based on the entries of the matrix, locations, and proportion of unobserved entries [19–21]. Mazumder et al. formulated the above problem as the following optimization problem [13]:

$$\begin{aligned} &\text{minimize } \operatorname{rank}(\mathbf{Z}) \\ &\text{subject to } \sum_{(i,j)\in\Omega} (X_{ij} - Z_{ij})^2 \leqslant \delta \end{aligned} \tag{1}$$

where $\delta \geqslant 0$ is a regularization parameter to control the error tolerance.

However, in the above optimization, the rank constraint makes the problem combinatorially hard for general Ω [22]. One small modification to 1 is [13]:

$$\begin{aligned} &\text{minimize } \|\mathbf{Z}\|_* \\ &\text{subject to } \sum_{(i,j)\in\Omega} (X_{ij} - Z_{ij})^2 \leqslant \delta \end{aligned} \tag{2}$$

where $\|\mathbf{Z}\|_*$ is the nuclear norm of $\mathbf{Z}$ ($\|\mathbf{Z}\|_* = \sum_{i=1}^{r} \sigma_i$, where $\sigma_1, \ldots, \sigma_r$ are the singular values of $\mathbf{Z}$ and r is the rank of $\mathbf{Z}$). This modification constitutes a problem in convex optimization[23]. We can be reformulated 2 to the Lagrange form [13]:

$$\underset{Z}{\text{minimize}} \frac{1}{2} \sum_{(i,j)\in\Omega} (X_{ij} - Z_{ij})^2 + \lambda \|\mathbf{Z}\|_* \tag{3}$$