Srrr-cluster: Using Sparse Reduced-Rank Regression to Optimize iCluster

Shu-Guang Ge[1], Jun-Feng Xia[2,3], Pi-Jing Wei[1], and Chun-Hou Zheng[1(✉)]

[1] College of Electrical Engineering and Automation, Anhui University, Hefei 230601, Anhui, China
zhengch99@126.com
[2] Institute of Health Sciences, Anhui University, Hefei 230601, Anhui, China
[3] Center of Information Support and Assurance Technology, Anhui University, Hefei 230601, Anhui, China

Abstract. Cancer genome projects can provide different types of data on the genetic level, which is significant for cancer research and biological processes in computational methods. Thus, computational methods used to identify cancer subtypes should fully focus on integrating these multidimensional data (e.g., DNA methylation data, mRNA expression data, etc.). Sparse reduced-rank regression (Srrr) method, a state-of-the-art multiple response linear regression method, can easily deal with high dimensional statistical data. In this paper, we introduced Srrr method combining iCluster (Srrr-cluster) to discovery cancer subtypes. Firstly, we used Srrr to estimate the coefficient matrix and then cancer subtypes were clustered by iCluster. Finally, we used our Srrr-cluster method to analyze glioblastoma and breast cancer data. The results show that our Srrr-cluster method is effective for cancer subtype identification.

Keywords: Cancer subtypes · Clustering · iCluster · Sparse reduced-rank regression · High dimensional statistical data

1 Introduction

The genomes of cancer often contain a large number of somatical aberrations information, e.g., DNA copy number aberrations are closely relation to tumor gene by gene amplification or tumor suppressor loss because of genomic instability and deregulation [1, 2]. Other cases, epigenetic aberrations also result in oncogene such as genomic methylation [3]. DNA sequence change will directly affect the mRNA expression levels even other non-coding microRNA, and then change the outcome of the transcriptome, eventually produce individual heterogeneity and lead to distortion of cancer cells. The same cancer may have diverse somatic mutation and transcriptional level, so that the formation of different kinds of subtypes has diverse heterogeneity of biological progresses and phenotypes [4]. For example, glioblastoma (GBM) can be defined as the Classical, Mesenchymal, and Proneual subtypes by aberrations and gene expression of EGFR, NF1 and PDGFRA/IDH1 [5].

D.-S. Huang et al. (Eds.): ICIC 2016, Part III, LNAI 9773, pp. 99–106, 2016.
DOI: 10.1007/978-3-319-42297-8_10

Recently, many cancer genome projects are established and amassed a large number of various types of data. For example, The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/) contains genome, transcriptome and expression information for over 20 cancers from thousands of patients, which produced several types of data, such as methylation data, mRNA expression data, DNA copy number data and so on. Currently, some integrative methods have been proposed which combine different biological data for cancer subtype classification. For example, iCluster is a integrating probability model of multiple data based on Gaussian latent variable model. Which first structures an optimizing penalized log-likelihood function to estimate using Expectation-Maximization algorithm with lasso-type sparse [6, 7], then using K-means to get subtypes. However, bulky datasets also bring about many challenges for subtypes classification. Firstly, the key of data-integration clustering method tend to construct a variance - covariance structure within data types, namely coefficient matrix solving, which is equivalent to a feature selection process. The coefficient matrix is a projection matrix that projects the original data onto an eigengene-eigenarray subspace. Secondly, high-dimensional datasets have a common feature that the number of samples small yet the number of genes is large, so the dimension reduction of coefficient matrix is essential. iCluster used PCA method to estimate the coefficient matrix that defined the first k–1 eigenvectors by a pivoted QR decomposition [8, 9]. However, in the Gaussian latent variable models, PCA has many deficiencies: (i) significant features can't completely be extracted when facing high dimensional statistical data. (ii) eigenvalue of the first principal component is much larger than the eigenvalues of the other main components.

Briefly, we can see that the estimator of coefficient matrix is very important. Considering Gaussian latent variable model, sparse reduced-rank regression (Srrr) is a useful parsimony model when facing a large number of data for multiple response regression [10–14]. Generally, Srrr with the purpose of solving an indicator matrix can be divided into three steps in different algorithm: (i) Working out reduced-rank matrix that can reduce the noise of the model and improve the robustness. (ii) Constructing sparse group lasso, group bridge or group MCP term, which can solve the problem that the sample volume is pretty smaller than the gene volume [10, 11]. (iii) Establishing minimum optimization function to solve the coefficient matrix. Until now, Srrr method has been applied in several research area. E.g., Lin et al. (2013) used it to detect genetic networks associated with brain functional networks in schizophrenia [12]. Chen et al. (2012) proposed a weighted rank-constrained group lasso approach with two heuristic numerical algorithms and studied its large sample asymptotics [13].

In this paper, we used subspace assisted regression with row sparsity (SARRS) algorithm that proposed by Ma et al. (2014) [14], combining with iCluster (Srrr-cluster) to discovery cancer subtypes. Srrr-cluster can be regarded as a data-integration clustering method which first estimating the coefficient matrix of the latent variable model using the Srrr method, and then solving the estimator of the desigen matrix through optimizing a penalized complete-data log-likelihood with sparse term using the Expectation-Maximization (EM) algorithm.

2 Srrr-cluster Methods

2.1 Data Types Integration and a Gaussian Latent Variable Model Representation

Mo et al. summarized the different data types to adapt to different mathematical probability models [15]. For example, mutation status is defined binary variable that is suit for logistic regression model; copy number loss, gain, and normal status are defined multicategory variable that are suit for multilogit regression; NDA copy number data, DNA methylation data, mRNA expression data and so on are defined continuous variable that are suit for Gaussian latent model. In this paper, different types of continuous data are regressed using Srrr model to discovery cancer subtypes. We can fuse the same samples with different types of continuous data into a multiple genomic data. Therefore, we employ an integrating genomic data that harbor different levels of expression and transcriptome information to search subtypes.

Firstly, we establish a Gaussian latent variable model:

$$X = ZW + \varepsilon \tag{1}$$

here $X = \{X_1, \ldots X_P\}$ is the original integration data of dimension $n \times m$, where X_1 can denote DNA methylation data of dimension $n \times m_1$, X_2 can denote DNA copy number data of dimension $n \times m_2$, X_p can denote mRNA expression of dimension $n \times m_p$ and so forth. Z is the design matrix of dimension $\mathrm{n} \times l$, W is the coefficient matrix of dimension $l \times m$, ε is the error term and make the additional assumption that $Z \sim N(0, I)$ and $\varepsilon \sim N(0, \psi)$. p is the number of genomic data types, n is the number of samples, m is the number of the genes, l is the number of predictors. Ding et al. (2004) noted that the K-means solution of Z can directly be selected using the first k–1 eigenvectores that span a low-dimensional latent space where the original data are projected onto each of the first K–1principal directions such that the total variance is maximized by PCA. So, Z is the design matrix of dimension $\mathrm{n} \times (\mathrm{k} - 1)$ that is finally clusters latent tumor subtypes and the initial value of Z is the first k–1 eigenvectores by PCA, where k is the number of clusters [9].

2.2 An Adaptive Srrr Method and Srrr-cluster

Following Eq. (1), we can afford to estimate the solution of the coefficient matrix W using an adaptive Srrr method. The goal is to reduce the rank r of W under the Gaussian latent variable model. Firstly, two error parameters, i.e., a noise level σ expressed as $\sigma = median(\sigma(X))/\sqrt{\min(n, m)}$, where $\sigma(X)$ is the collection of all nonzero singular values of X, and a noise rank level η, expressed as $\eta = \sqrt{2m} + \sqrt{2(\min(n, k)}$, are estimated to work out the reduced-rank r and an orthonormal matrix $V_{(0)}$ that is non-orthogonal to the right singular subspace of W. The estimator of r is computed by:

$$r = \max\{j : \sigma_j(Z(Z'Z)^{-}Z'X) \geq \sigma\eta\} \tag{2}$$

Where $(ZZ')^{-}$ is Moore-Penrose pseudo-inverse. So, the Srrr method use the first r-th right vector of $Z(Z'Z)^{-}Z'X$ to estimate the orthonormal matrix $V_{(0)}$:

$$V_{(0)} = (V_1^{(0)}, \ldots, V_r^{(0)}) \tag{3}$$

Depending on characters of the orthonormal matrix $V_{(0)}$, such as $W = WV_{(0)}V'_{(0)}$, the reduced-rank matrix B can be expressed as:

$$B = WV_{(0)} \tag{4}$$

with dimension $(k-1) \times r$ which columns being the estimator of rank. What more, VV' is a projection matrix that approximatively maps onto the right singular subspace of W.

For the sake of simplicity, Ma et al. take sparse group lasso in this model, where each row of the B is regarded as a group and all groups are of the same size r [14]. Each row takes sparse process by the ℓ_2 matrix norm as follows:

$$\rho(B;\lambda) = \lambda \sum_{j=1}^{k-1} \left\|B_{j*}\right\|_2 \tag{5}$$

where λ is the penalty level.

Following these, Srrr method constructs a right bias-variance tradeoff function with reduced-rank term representing the variance part and sparse lasso term representing the bias part using SARRS algorithm:

$$W = \underset{Z \in \Re^{(k-1)\times n}}{\arg\min} \left\{ \left\| XV_{(0)}V'_{(0)} - ZWV_{(0)}V'_{(0)} \right\|_F^2 \Big/ 2 + \rho(WV_{(0)}V'_{(0)};\lambda) \right\} \tag{6}$$

we can further reduce the computation cost by first solving:

$$B_{(1)} = \underset{B \in \Re^{(k-1)\times r}}{\arg\min} \left\{ \left\| XV_{(0)} - ZB \right\|_F^2 \Big/ 2 + \rho(B;\lambda) \right\} \tag{7}$$

However, $B_{(1)}$ is not accurate but close to $WV_{(0)}$ because the columns of $V_{(0)}$ is just approximate to the right singular subspace of W.

It is worth noting that the right singular subspace of W is exactly the same as that of ZW. Next step, we can estimate the left singular subspace $U_{(1)} \in \mathbb{R}^{n\times r}$ of $ZB_{(1)}$. Due to (4), $U_{(1)}$ is exactly the left singular subspace of $ZWV_{(0)}$, which in turn equals the left subspace of ZW. Through the same line of logic, $U_{(1)}U'_{(1)}$ is a projection matrix that accurately maps onto the left singular subspace of WZ. Then, we can easily compute the right singular vectors $V_{(1)} \in \mathbb{R}^{m\times r}$ of $U_{(1)}U'_{(1)}X$, which in turn equals the right

subspace of ZW. Successfully, a pretty accurate right singular vectors of W is estimated. Finally, using $V_{(1)}$ instead of $V_{(0)}$ to solve the equation:

$$B_{(2)} = \underset{B \in \Re^{(k-1)\times r}}{\arg\min} \left\{ \left\| XV_{(1)} - ZB \right\|_F^2 / 2 + \rho(B; \lambda) \right\} \tag{8}$$

Hopefully, we compute the estimated indictor matrix by $W = B_{(2)} V'_{(1)}$.

Given two or more types of data from the same cohort of patients, our Srrr-cluster method first fuse these data into an integrative matrix, and then use the optimized PCA to compute a design matrix for the integrative data. The next step is to use the adaptive Srrr method to calculate the coefficient matrix under the Gaussian latent variable model, which can project sample × gene space of the original data into eigenarray × eigengene subspace. Finally, we use iCluster method to discovery cancer subtypes.

2.3 Evaluation Metric

We use three commonly used metrics to evaluate Srrr-cluster performance by identifying subtypes in these cancers. (i) Silhouette score, a measure of cluster homogeneity, which is defined as $s(i) = (b(i) - a(i))/(\max(a(i), b(i)))$, where $a(i)$ is average dissimilarity between i and all the other points of the same subtypes, $b(i)$ is average dissimilarity between i and all the other points of the different subtypes, i is an arbitrary sample. If sihouette value is close to 1, it means that the data are appropriate [16]. (ii) P value in Cox log-rank test, which is used to assess the significance of the different in survival profiles between subtypes [17]. (iii) The proportion of deviance (POD), which is a score of evaluating cluster degree of separation by a diagonal block structure. We set a matrix $A = Z^T Z, A \in \Re^{n\times n}$. Then the elements of A is defined as $a_{ij} \big/ \sqrt{a_{ii} a_{jj}}$ for i = 1,…, n and j = 1,…,n, and set negative values to zero, which can order cancers belonging to the same clusters into a adjacent structure. If the diagonal block matrices were prefect, all elements of the diagonal blocks would be non-negative and all elements of the off-diagonal blocks would be zero. So, compared A with the prefect diagonal block structure, we define a deviance measure d, which is the sum of quantities that the diagonal blocks' elements of A appear zero and the off-diagonal blocks' elements of A appear non-negative values. POD is defined as d/n^2 so that POD is between 0 and 1. Small values of POD indicate strong cluster separability, and large values indicate of POD indicate poor cluster separability[6].

3 Results

3.1 Subtypes Discovery in Breast Cancer

Using DNA copy number and mRNA expression on the same cDNA microarrys that contain 6691 genes form Pollack et al. [1] from 37 primary breast cancers and four breast cancer cell lines, we compared the Srrr-cluster results with iCluster. As well known, the expression profiles of the four cell line samples (BT474, T47D, MCF7 and

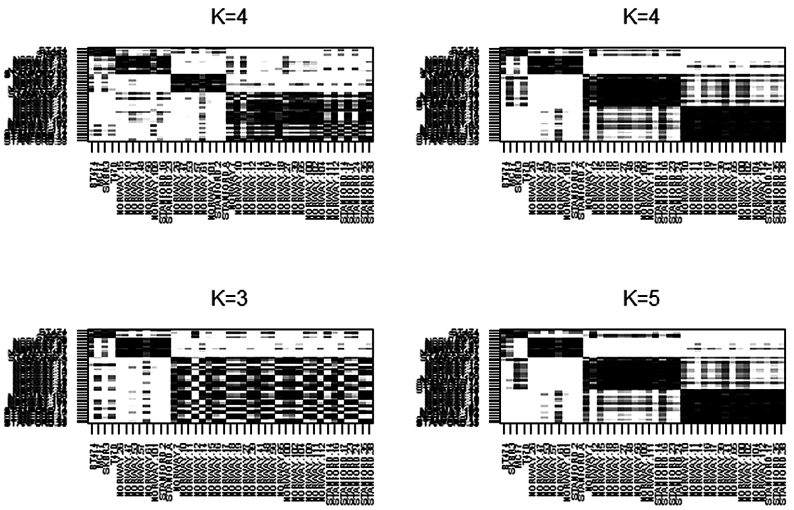

Fig. 1. Diagonal blocks structures obtained using iCluster (k = 4) and Srrr-cluster (k = 4, k = 3 and k = 5) methods.

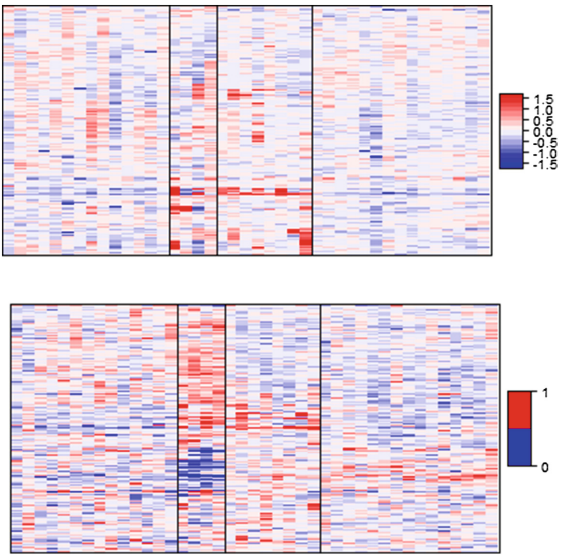

Fig. 2. Heatmaps of copy number (DNA) and gene expression (mRNA) with samples arranged by the integrated cluster assignment under the Srrr-Cluster model. (Color figure online)

SKBR3) should be similar so that they can be considered as a subtype from the rest of the tumor samples. Additional, HER2/ERBB2 is an important prognostic factor for breast cancer near17q12, the clinical features and biological behavior of a special

performance, treatment modalities HER2-positive (overexpression or amplification) of breast cancer also have a great difference with other types of breast cancer.

Figures 1 and 2 shows the diagonal blocks structures under the sparse solution $\lambda = (0.2, 0.2)$, r corresponding to Cluster (k = 4) and Srrr-cluster (k = 4, k = 3 and k = 5) respectively. POD values of the clustering solutions are 0.1519533, 0.1254317, 0.2478124 and 0.1259145 respectively. Considering the values of POD, the four clusters obtained by Srrr-cluster method should be the best one. Figure shows the heatmaps of the profiles of DNA copy number data and mRNA expression data when samples were splitted four clusters using Srrr-cluster method. Carefully analysing the four clusters combined heatmaps, we can see that cluster 1 is composed of the four cell lines and cluster 2 is amplification in the DNA and overexpression in the mRNA associated with the HER2/ERBB2.

3.2 Subtypes Discovery in GBM

The GBM dataset contains miRNA (534 genes) and mRNA expression (1740 genes) data from 73 patients with GBM [18]. We used three evaluation metrics to evaluate the result of the Srrr-cluster and iCluster Scluster: (i) The sihouette scores. (ii) The P values. (iii) The POD values. The results of these three metrics are listed in Table 1. According to these metrics, we can see that, using Srrr-cluster method, the within-clusters have stronger coherence and the between-clusters have well separability.

Table 1. Three evaluation metrics to evaluate iCluster and Scluster (3 clustering solution)

Evaluation values	iCluster	Srrr-cluster
Sihouette scores	0.42	0.48
P values	0.31	0.04
POD values	0.20	0.17

4 Discussion

Srrr-cluster method can find more suitable coefficient matrix which can project the original data onto an eigengene-eigenarray subspace when analyzing dataset with small sample size and large variables. In this paper, we proposed to use Srrr-cluster method for cancer subtypes discovery. Compared with iCluster method, our method can identify more stable clusters. However, because Srrr-cluster is established on the basis of iCluster, it has a major limitation that it needs a priori gene selection. In future, we will explore how to solve this problem.

Acknowledgments. This work was supported by National Natural Science Foundation of China (31301101 and 61272339), the Anhui Provincial Natural Science Foundation (1408085QF106), the Specialized Research Fund for the Doctoral Program of Higher Education (20133401120011), and the Technology Foundation for Selected Overseas Chinese Scholars from Department of Human Resources and Social Security of Anhui Province (No. [2014]-243).

References

1. Pollack, J.R., et al.: Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc. Natl. Acad. Sci. **99** (20), 12963–12968 (2002)
2. Stratton, M.R., Campbell, P.J., Futreal, P.A.: The cancer genome. Nature **458**(7239), 719–724 (2009)
3. Jones, P.A., Baylin, S.B.: The fundamental role of epigenetic events in cancer. Nat. Rev. Genet. **3**(6), 415–428 (2002)
4. Hoadley, K.A., et al.: Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell **158**(4), 929–944 (2014)
5. Verhaak, R.G., et al.: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell **17**(1), 98–110 (2010)
6. Shen, R., Olshen, A.B., Ladanyi, M.: Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics **25**(22), 2906–2912 (2009)
7. Shen, R., et al.: Integrative subtype discovery in glioblastoma using iCluster. PLoS ONE **7** (4), e35236 (2012)
8. Zha, H., et al.: Spectral relaxation for K-means clustering. Neural Inf. Process. Syst. 1057–1064 (2001)
9. Ding, C., He, X.F.: Cluster structure of K-means clustering via principal component analysis. Adv. Knowl. Discov. Data Min. Proc. **3056**, 414–418 (2004)
10. Simon, N., et al.: A sparse-group lasso. J. Comput. Graph. Stat. **22**(2), 231–245 (2013)
11. Huang, J., Breheny P., Ma S.: A selective review of group selection in high-dimensional models. Stat. Sci. **27**(4) (2012)
12. Lin, D.D., et al.: Network-based investigation of genetic modules associated with functional brain networks in schizophrenia. In: 2013 IEEE International Conference on Bioinformatics and Biomedicine (Bibm) (2013)
13. Chen, L., Huang, J.Z.: Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J. Am. Stat. Assoc. **107**(500), 1533–1545 (2012)
14. Ma, Z., Sun T.: Adaptive sparse reduced-rank regression (2014). arXiv preprint arXiv:1403.1922
15. Mo, Q., et al.: Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc. Natl. Acad. Sci. USA **110**(11), 4245–4250 (2013)
16. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
17. Hosmer Jr, D.W., Lemeshow, S.: Applied Survival Analysis: Regression Modelling of Time to Event Data. European Orthodontic Society (1999)
18. Wang, B., et al.: Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods **11**(3), 333–337 (2014)