Non-negative Principal Component Analysis Based Scaling: Application on NMR Spectroscopic Metabolomics

Lingli Deng[1,2], Jiyang Dong[1], and Zhong Chen[1,2]

[1]*Department of Electronic Science, Xiamen University, Xiamen, Fujian, China,* [2]*Department of Communication Engineering, Xiamen University, Xiamen, Fujian, China*

Target audience

The target audience is basic scientists and medical statistical scientists who are interested in NMR spectroscopic metabolomics and relevant data processing.

Purpose

Scaling is an important data preprocessing procedure prior to multivariate statistical analysis for nuclear magnetic resonance (NMR) spectroscopic metabolomics to ensure low and high concentration metabolites contribute equally to the multivariate model. The commonly used methods, such as unit variance (UV) scaling, Pareto scaling and variable stability (VAST) scaling, scale each variable of the data independently, which ignores the chemical meaning of the spectra (and hence the natural correlates) and may make the subsequent analysis be hard to interpret. A new scaling method based on non-negative principal component analysis (NPCA) [1] is proposed in this paper. The new method aims to perform scaling on the concentration of the metabolites rather than on the variables of the data. So the scaling results would have a better physical approximation to the data.

Methods

NPCA is used to factorize the data matrix $\mathbf{X} = \mathbf{C}\mathbf{S}^t + \mathbf{E}$, where $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_K)$ and $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_K)$ can be regarded as concentration and spectral matrix of the K metabolites respectively, and $\mathbf{E}$ is the residual matrix corresponding to noise and inter-individual difference. Data scaling is performed on the matrix $\mathbf{C}$ rather on the matrix $\mathbf{X}$, resulting in each metabolites has equal potential to influence the model, *i.e.* with unit variance, and the scaled data matrix would be, $\mathbf{X}^s = \mathbf{C}^s\mathbf{S}^t + \mathbf{E} = \mathbf{C}\mathbf{A}\mathbf{S}^t + \mathbf{E}$, where $\mathbf{A} = diag(\alpha_1, \alpha_2, \cdots, \alpha_K)$, $\alpha_k \in \mathbb{R}^+$ is the scaling factor of metabolite k. At last multivariate statistical analysis can be employed on $\mathbf{X}^s$, where low and high concentration metabolites has equal weights.

Results and discussion

A simulated metabolomic dataset contains 100 synthetic ^{1}H NMR spectra with 11 metabolites were simulated by MetAssimulo package [2], and the concentrations of metabolites reference the example available in MetAssimulo. Scaling results of simulated dataset are given in Fig1. After scaling by common used variable scaling method, i.e. UV, the weights of noise variables were greatly increased, and chemical meaning of variables were lost (Fig. 1b). While the chemical meaning of the spectra in the scaled data was kept as far as possible by NPCA-based scaling, showing in Fig. 1c.

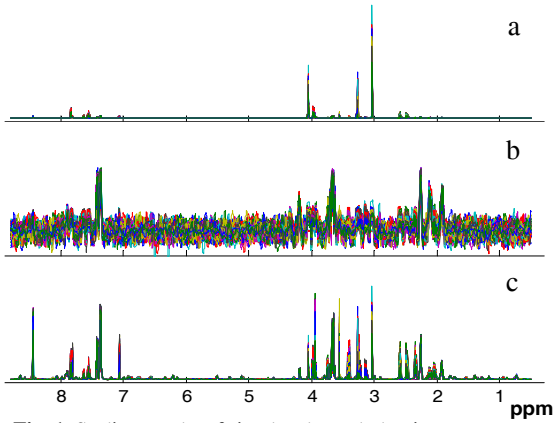

Fig. 1. Scaling results of simulated metabolomic spectra. (a) Stacked spectra of original dataset; (b) scaled by UV; and (c) scaled by mew method.

A real metabolomic dataset consisted of ^{1}H NMR spectra obtained from literature [3]. These spectra were phased, baseline corrected, aligned manually to correct for peak shift, and bucked into 1340 bins with fix-width interval (0.005 ppm). Dataset is scaled by UV and new scaling method respectively, and then employed by OPLS-DA. The results of multivariable analysis are shown in Fig. 2. Obviously the loading of OPLS-DA model by new method scaled dataset is more interpretable than UV method.

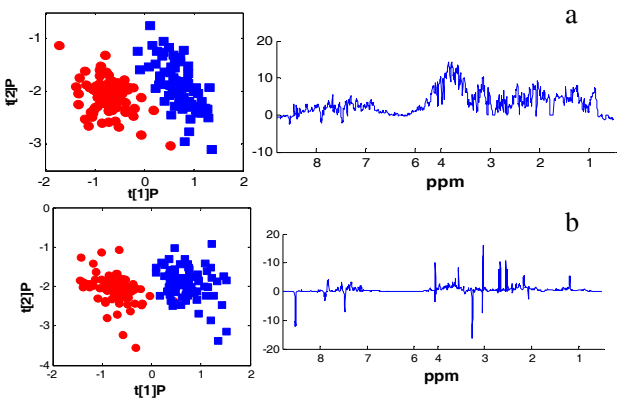

Fig. 2. OPLS-DA scores (left) and loading plots (right) of (a) UV method scaled dataset; and (b) new method scaled dataset.

Conclusion

NPCA based scaling method, scaling be performed on the concentration of metabolites, improved interpretability of both scaled dataset and the results of subsequent multivariable analysis. It offers a new angle of view for the data scaling of NMR spectroscopic metabolomics.

Reference

1. Deng LL, Cheng KK, Dong JY, et al. Non-negative principal component analysis for NMR-based metabolomic data analysis. Chemometrics Intell Lab Syst. 2012; 118:51-61.
2. Muncey HJ, Jones R, De Iorio M, et al. MetAssimulo: simulation of realistic NMR metabolic profiles. BMC bioinformatics. 2010; 11:496-506.
3. Xu JJ, Yang SY, Cai SH, et al. Identification of biochemical changes in lactovegetarian urine using ^{1}H NMR spectroscopy and pattern recognition. Anal Bioanal Chem. 2009; 396:1451-1463.