Assessing the impact of variable scaling on projection methods in the analysis of spectral data

Jean-Pierre Labuschagne*
University of South-Africa, Roodepoort, South-Africa - jp.labuschagne@gmail.com

René Pellissier
University of Massachusetts, United States of America - renepellissier@gmail.com

Abstract

Data pre-processing performs a core function in the analysis of spectral data. The choice of the pre-processing method not only impacts the final model, but also affects the extent of influence that variables play in estimating the final model. In this research the impact that different variable scaling methods have on principal component analysis (PCA) and variable importance on projection (VIP) was investigated. Auto, Pareto, level, power, vast and range scaling was applied to eight spectral datasets. Results from PCA indicated that centred and log scaled data consistently extracted the largest amount of variation with the lowest number of components throughout all the datasets. VIP scores were heavily affected by the use of different scaling methods, furthermore, patterns pertaining to the ranking of the VIP scores between the different scaling methods were observed. This research has confirmed to what extent data pre-processing influences analyses when using projection methods, and care should therefore be taken in the selection of the data pre-treatment method. This result could change the way previous results were interpreted and it is hoped that it will contribute to setting a standard for pre-processing data in a set way as well as enhance comparisons in studies were different scaling methods are employed.
Keywords: data pre-processing; variable scaling; PCA; VIP, PLS; variable scaling; spectral data.

1. Introduction

Spectral data is used in a wide array of fields that include metabonomics, proteomics and metabolomics. Each of these fields is heavily reliant on data analysis to reach its particular research objectives, some of which include novel discoveries for drug development Holzgrabe et al. (2005), biomarker identification Cloarec et al. (2005) and determining differences between different plant cultivars Ward, Baker & Beale (2007) - to name a few. The purpose of data pre-processing in the context of analysing spectral data is to remove as much unwanted noise (variation) as possible, so that the targeted biological signals are depicted clearly Kohl et al. (2012) and to correct for heteroscedasticity in the data. According to Zhang et al. (2009), these methods can be separated into two distinct groups: a) methods that remove unwanted sample-to-sample variation and b) methods aimed at transforming data towards homoscedasticity. The first group consists of functions applied to the rows of the dataset,aiming to reduce technical variation. This group will be referred to as data normalisation. The second group is known as variable scaling methods, where the function is applied to the each of the column vectors within the data matrix. Since data normalisation is row operation and has the same effect on all of the variables, the impact thereof on projection methods is assumed to be less significant than the impact that variable scaling would have, seeing as variable scaling purposefully alters the variance structure of the variables. The adjustment for heteroscedasticity is, however, not the only function of these variable scaling methods. The scaling effect also creates an equal footing for small differences between different spectra in relation to large differences between spectra Eriksson & Umetrics AB (2006). In principal component analysis (PCA), without the equal footing provided by variable scaling methods, centred data would favour large differences in spectra above small differences in spectra, primarily because of the larger variances associated with it. This is an important characteristic in the context of spectral data analysis, because in relative terms, small differences in spectra may be just as important as large differences. The focus of this research is therefore on the statistical analysis of spectral data, although more specifically the impact that variable scaling methods have on the projection methods used in the data analysis routine. This analysis routine, commonly found in chemometric data analysis, is visualised in figure 1 below.

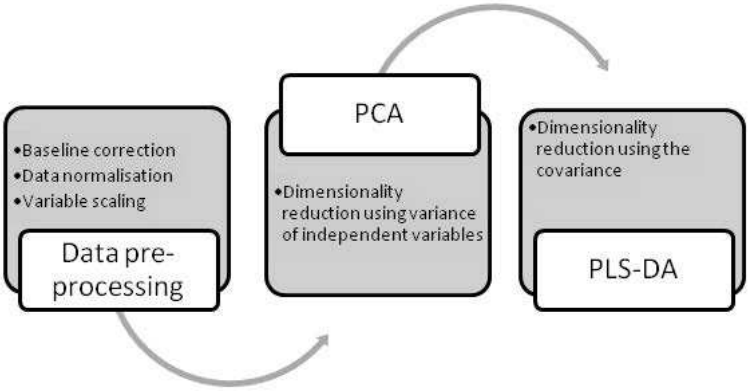

Figure 1: Process flow of 1H NMR data analysis

2. Scaling methods

The variable scaling methods for data pre-treatment can be separated into two sections van den Berg et al. (2006); the first being methods that use the measure of data dispersion as a scaling factor. These methods include auto scaling, Pareto scaling, range scaling and vast scaling. The second section comprises of methods that use a size measure as scaling factor, for example, level scaling and mean-centring. A potential pitfall when relying purely on either level or mean-centring, however, is in scenarios where heteroscedasticity is present; this is because of the overshadowing effect that variables with larger variances have over variables with smaller variances in the use of projection methods. For ease of reference, mean-centring will henceforth be referred to as centring and mean-centred data as centred data. A list of the different scaling methods that are used in this research, together with the corresponding formulae, is given in table 1.

Table 1: Formulae of the different scaling methods adopted from van den Berg et al. (2006,p.4 (page number not for citation purposes))

Method	Formula
Centering	$\bar{x}_{ij} = \bar{x}_{ij} - \bar{x}_i$
Auto scaling	$\bar{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$
Range scaling	$\bar{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{x_{\max i} - x_{\min i}}$
Pareto scaling	$\bar{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$
Vast scaling	$\bar{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i} \times \frac{\bar{x}_i}{s_i}$
Level scaling	$\bar{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$
Log transformation	$\bar{x}_{ij} = \log x_{ij}$
Power transformation	$\bar{x}_{ij} = \sqrt{x_{ij}}$

Pareto scaling is advocated in the majority of metabolomics analysis guides. The function uses the square root of the standard deviation and variable mean to standardise the data. This is effective in reducing the effect of the larger variance variables compared to the smaller variance variables whilst partially maintaining the structure of the data. Another advantage is that the data does not become dimensionless, as after auto scaling van den Berg et al. (2006).Variable stability scaling (Vast) can be regarded as an extension of auto or unit variance scaling. The primary motive is the assumption that stable variables (variables that do not show strong variation) provide better discrimination properties that unstable variables Keun et al. (2003). In

this research, vast scaling is performed using the coefficient of variation (CV) as a stability parameter, but is flexible in that the parameter can be added in accordance with the data needs and requirements. The influence of a priori information regarding the variables could also lead to using alternative parameters other than the CV. Auto scaling is a commonly applied method using the standard deviation to standardise the data. After this type of scaling, the variables have a standard deviation of one, and the correlations, as opposed to covariances, are then used as a base for the analysis van den Berg et al. (2006). Auto scaling is stated to be the most objective of the scaling methods, given that there is no prior information about the characteristics of the data Kowalski (1984). The use of unit variance scaling is recommended by *FAQ | Umetrics* (n.d.) in scenarios where data are on different scales. Range scaling is recommened in situations where keeping the biological structure intact, to keep interference from the variance within the data, range scaling is of importance. Instead of using a function of the standard deviation, range scaling - as the name suggests - uses a range from the smallest value to the largest value within a variable as a method of standardisation. Level scaling uses size as a method of scaling, similar to centring, in that rather than using rows or observation means, the standardisation is done with column means. When comparing level scaling to the more common methods using standard deviation, there is a very noticeable difference. This difference is in the amount of size reduction relative to the variables. From the formula given in table 1, it is evident that a variable that has a larger variance will keep the ratio to a greater extent when using level scaling. In this way, it is opposite to Pareto scaling, where the relative difference between two variables with a greater difference in size is reduced after scaling is performed. Transformations (log transform) are primarily employed to account for, or remedy, the effect of heteroscedasticity. The assumptions made for the analysis of spectral data (NMR) include normality and homoscedasticity and whilst the former can be accounted for with a larger sample size (central limit theorem), the latter requires a multiplicative form of scaling Kvalheim, Brakstad & Liang (1994). Log transform has the benefit of being able to reduce the effect of variables with larger variances greatly, but at the cost of creating a pseudo scaling effect. Another disadvantage is that data entries with a zero value would need to be adjusted for to allow log transformation to be used. As a general observation, Eriksson & Umetrics AB (2006) state that variables that vary more than tenfold are often logarithmically transformed. Power transformation is another method of correcting for heteroscedasticity. With this method there is, however, no problem with smaller values, as is the case with the log transform. A drawback is that this method cannot make multiplicative effects additive and the choice of the square root parameter is arbitrary van den Berg et al. (2006).

3. Methods

In order to promote the transferability of the research, datasets were purposefully chosen. The objective for selection was to obtain a diverse collection of spectral data. A total of seven datasets were used for the analysis. These datasets include: two chemical mixtures of three compounds using a high performance liquid chromatography equipped with ultraviolet detector (HPLC-UV), each matrix consisting of 73 wavelengths and 40 time points Tauler, Lacorte & Barcel (1996), Wehrens (2011); a near infrared reflectance (NIR) dataset consisting of 654 mass spectra Wehrens (2011); a mass spectrometry (MS) dataset consisting of 654 mass spectra Qu et al. (2002), Adam et al. (2002), Wehrens (2011); and four nuclear magnetic resonance (NMR) datsets, consisting of 30, 93, 60 and 34 samples respectively. In a study evaluating data pre-processing of 1H NMR spectroscopic data, Craig et al. (2006) concluded that analogous considerations would be needed for other biofluids, other analytical approaches (e.g. HPLCMS), and indeed for other omics techniques (i.e., transcriptomics or proteomics) and for integrated studies with fused data sets. The scaling methods chosen for inclusion in the research were selected according to prominence in literature van den Berg et al. (2006), Craig et al. (2006), Keun et al. (2003).In all of the datasets, negative values were set to 10^{-10}; this was done in order to make the use of log scaling feasible. The method of setting negative values to approximately zero is supported by Halouska & Powers (2006) and van den Berg et al. (2006). Once the data had been normalised, the baseline corrected, and the negative values transformed, the different variable scaling methods, as referred to in table 1 were applied using R Core Team (2014) pacakge muma Gaude et al. (2012) together with in-house R scripts . PCA was performed using Kucheryavskiy (2014). PLS-DA models and VIP scores were calculated using *FAQ | Umetrics* (n.d.). To enable a comparison between the different VIP scores of the differently scaled models, the five highest VIP scores from auto scaling were used as the reference variables.

4. Results

Results from PCA indicated that centred and log scaled data consistently extracted the largest amount of variation with the lowest number of components throughout all the datasets. Auto and range scaling extracted significantly lower amounts of variation, with range scaling found in the lowest two methods in all the datasets (table 2).

Table 2: PCA: Average explained variance per component

	Scaling	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
1	Centred	82.47	9.49	3.49	1.30	0.81	0.56	0.26
2	Log	82.30	4.39	2.24	1.28	1.06	0.82	0.70
3	Power	61.55	14.27	7.02	3.76	2.50	1.74	1.25
4	Vast	50.20	10.57	7.42	4.41	3.47	2.84	2.41
5	Pareto	49.39	20.36	9.38	4.64	3.15	2.20	1.39
6	Level	46.44	14.50	9.17	6.25	4.09	3.35	2.81
7	Range	41.64	11.98	6.88	5.00	4.07	3.14	2.80
8	Auto	40.34	13.51	7.00	5.13	4.02	3.08	2.84

The VIP scores of the PLS-DA models are shown in table 3. The scaling methods are ranked according to the respective median VIP scores in descending order in each of the different datasets. Centred data showed extreme rankings in each of the datasets, indicating that VIP scores generated from centred data share little similarity with VIP scores generated from variable scaled data. A pattern of scaling methods with higher scores and those with lower scores was observed. Centred, log, vast and level scaled data grouped in the higher scoring group, whilst power and Pareto scaling are found at the lower end of the group. Auto and range scaling obtained similar VIP scores.

Table 3: Summary of VIP scores in ranks using PLS-DA.

	Sutherlandia	*L. fruticosus*	*Maroela*	Transxl	HPLC	Prostate	Shootout
1	Centred	Log	Vast	Level	Level	Centred	Log
2	Vast	Centred	Auto	Vast	Log	Log	Auto
3	Pareto	Vast	Log	Centred	Range	Level	Vast
4	Power	Range	Centred	Log	Auto	Vast	Pareto
5	Level	Pareto	Range	Range	Vast	Pareto	Range
6	Auto	Auto	Level	Auto	Power	Power	Level
7	Range	Power		Power	Pareto	Auto	Power
8	Log	Level		Pareto	Centred	Range	Centred

Conclusion.

Scaling has an undeniable impact on the resulting analysis. The effect of scaling on dimension reduction techniques such as PCA and PLS-DA has been substantiated in this research. The number of components extracted to explain a certain amount of variation differs greatly between scaling methods. This fact alone provides a good basis for the selection of the most effective scaling method for a given dataset, where model simplicity is an important factor. Following PLS-DA and OPLS-DA model estimation, it was observed that both contribution scores and VIP scores were heavily affected by the use of different scaling methods. Furthermore patterns pertaining to the proximity of the VIP scores between the different scaling methods were observed. Considering that the variables in spectral data are at times representative of metabolites, and that the VIP scores of these metabolites could be used as evidence for biomarker discovery, the importance of an understanding of the impact of variable scaling in spectral data analysis is critical. This result could possibly change the way previous articles where evaluated and it is hoped that it will contribute to setting

a standard for both pre-processing data in a set way, and enhancing comparisons in studies were different scaling methods were employed. In addition to relying purely on the predictive ability of the estimated model, input from researchers and scientists in the respective fields would greatly contribute to ensuring that biological significance remains intact throughout the analysis procedure.

References

Adam, B.-L., Qu, Y., Davis, J. W., Ward, M. D., Clements, M. A., Cazares, L. H., Semmes, O. J., Schellhammer, P. F., Yasui, Y., Feng, Z. & Wright, G. L. (2002), 'Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men', *Cancer Res.* **62**(13), 3609–3614.

Cloarec, O., Dumas, M.-E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J. C., Holmes, E. & Nicholson, J. (2005), 'Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic ^{1} h NMR data sets', *Analytical Chemistry* **77**(5), 1282–1289.
URL: *http://pubs.acs.org/doi/abs/10.1021/ac048630x*

Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K. & Lindon, J. C. (2006), 'Scaling and normalization effects in NMR spectroscopic metabonomic data sets', *Analytical Chemistry* **78**(7), 2262–2267.
URL: *http://pubs.acs.org/doi/abs/10.1021/ac0519312*

Eriksson, L. & Umetrics AB (2006), *Multi- and megavariate data analysis*, Umetrics AB;, Umea, Sweden.

FAQ | Umetrics (n.d.).
URL: *http://www.umetrics.com/support/faq*

Gaude, E., Chignola, F., Spiliotopoulos, D., Mari, S., Spitaleri, A. & Ghitti, M. (2012), *muma: Metabolomics Univariate and Multivariate Analysis*. R package version 1.4.
URL: *http://CRAN.R-project.org/package=muma*

Halouska, S. & Powers, R. (2006), 'Negative impact of noise on the principal component analysis of NMR data', *Journal of Magnetic Resonance* **178**(1), 88–95.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S1090780705002958*

Holzgrabe, U., Deubner, R., Schollmayer, C. & Waibel, B. (2005), 'Quantitative NMR spectroscopy: Applications in drug analysis', *Journal of Pharmaceutical and Biomedical Analysis* **38**(5), 806–812.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0731708505001573*

Keun, H. C., Ebbels, T. M., Antti, H., Bollard, M. E., Beckonert, O., Holmes, E., Lindon, J. C. & Nicholson, J. K. (2003), 'Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling', *Analytica Chimica Acta* **490**(1-2), 265–276.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0003267003000941*

Kohl, S. M., Klein, M. S., Hochrein, J., Oefner, P. J., Spang, R. & Gronwald, W. (2012), 'State-of-the art data normalization methods improve NMR-based metabolomic analysis', *Metabolomics* **8**(S1), 146–160.
URL: *http://link.springer.com/10.1007/s11306-011-0350-z*

Kowalski, B. R. (1984), *Chemometrics Mathematics and Statistics in Chemistry*, Springer Netherlands, Dordrecht.
URL: *http://dx.doi.org/10.1007/978-94-017-1026-8*

Kucheryavskiy, S. (2014), *mdatools: Multivariate data analysis for chemometrics*. R package version 0.5.3.
URL: *http://CRAN.R-project.org/package=mdatools*

Kvalheim, O. M., Brakstad, F. & Liang, Y. (1994), 'Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise', *Analytical Chemistry* **66**(1), 43–51.

Qu, Y., Adam, B.-L., Yasui, Y., Ward, M. D., Cazares, L. H., Schellhammer, P. F., Feng, Z., Semmes, O. J. & Wright, G. L. (2002), 'Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients', *Clin. Chem.* **48**(10), 1835–1843.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Tauler, R., Lacorte, S. & Barcel, D. (1996), 'Application of multivariate self-modeling curve resolution to the quantitation of trace levels of organophosphorus pesticides in natural waters from interlaboratory studies', *Journal of Chromatography A* **730**(1-2), 177–183.
URL: *http://linkinghub.elsevier.com/retrieve/pii/0021967395012060*

van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. (2006), 'Centering, scaling, and transformations: improving the biological information content of metabolomics data', *BMC Genomics* **7**, 142.

Ward, J. L., Baker, J. M. & Beale, M. H. (2007), 'Recent applications of NMR spectroscopy in plant metabolomics: NMR spectroscopy in plant metabolomics', *FEBS Journal* **274**(5), 1126–1131.
URL: *http://doi.wiley.com/10.1111/j.1742-4658.2007.05675.x*

Wehrens, R. (2011), *Chemometrics with R multivariate data analysis in the natural sciences and life sciences*, Springer, Heidelberg; New York.
URL: *http://public.eblib.com/choice/publicfullrecord.aspx?p=667105*

Zhang, S., Zheng, C., Lanza, I. R., Nair, K. S., Raftery, D. & Vitek, O. (2009), 'Interdependence of signal processing and analysis of urine ^{1} h NMR spectra for metabolic profiling', *Analytical Chemistry* **81**(15), 6080–6088.
URL: *http://pubs.acs.org/doi/abs/10.1021/ac900424c*