

## Supplementary information

### The influence of scaling metabolomics data on model classification accuracy

Piotr S. Gromski<sup>1</sup>, Yun Xu<sup>1</sup>, Katherine A. Hollywood<sup>1</sup>, Michael L. Turner<sup>2</sup> and Royston Goodacre<sup>1,\*</sup>

<sup>1</sup> School of Chemistry, Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.

<sup>2</sup> School of Chemistry, Brunswick Street, The University of Manchester, Manchester, M13 9PL, UK.

\* Correspondence to Prof Roy Goodacre: [roy.goodacre@manchester.ac.uk](mailto:roy.goodacre@manchester.ac.uk),

Tel: +44 (0) 161 306-4480

### Additional information on the algorithms employed in this study

**Table S1** General list of R packages and functions employed in this study.

Method	Package	Function	References
PC-DFA	'MASS' 'stats'	<i>lda</i> <i>prcomp</i>	Venables and Ripley, 2002 R Core Team, 2013
SVM	'e1071'	<i>svm</i>	Karatzoglou <i>et al.</i> , 2006
RF	'randomForest'	<i>randomForest</i>	Liaw and Wiener, 2002
KNN	'class'	<i>knn</i>	Venables and Ripley, 2002
For prediction	'stats'	<i>predict</i>	R Core Team, 2013

## Data structure and statistics used

For each data matrix  $x$  (illustrated in Figure S1) there are  $i$  rows (observation, samples) and  $j$  columns (variables, frequencies, metabolites), the mean  $\bar{x}_i$  and standard deviation  $SD_i$  are calculated thus:

$$\bar{x}_i = \frac{1}{J} \sum_{j=1}^J x_{ij}$$

$$SD_i = \sqrt{\frac{\sum_{j=1}^J (x_{ij} - \bar{x}_i)^2}{J - 1}}$$

$1$	$2$	$3$	.	.	.	.	.	.	$j$
$2$									
$3$									
.									
.									
.									
.									
.									
.									
.									
.									
$i$									$x_{ij}$

**Figure S1** Graphical representation of data matrix  $x$

Based on the above statistical features the data pre-treatment approaches are estimated based on the following equations:

$$\text{Autoscaling} = \frac{x_{ij} - \bar{x}_i}{SD_i}$$

$$\text{Range scaling} = \frac{x_{ij} - \bar{x}_i}{(x_{i_{max}} - x_{i_{min}})}$$

$$\text{Pareto scaling} = \frac{x_{ij} - \bar{x}_i}{\sqrt{SD_i}}$$

$$\text{Vast scaling} = \frac{(x_{ij} - \bar{x}_i)}{SD_i} \times \frac{\bar{x}_i}{SD_i}$$

$$\text{Level scaling} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$$

**Table S2** The average minimum and maximum classification rates (%) across 100 bootstrapped iterations run 10 times based on different pre-treatment approaches for NMR type 2 diabetes data (T2DM vs. Control)

	<i>Autoscaling</i>	<i>Range</i>	<i>Level</i>	<i>Pareto</i>	<i>Vast</i>	<i>None</i>
<b>PC-DFA</b>	(89.50;91.36)	(89.78;91.38)	(83.99;86.24)	(89.39;90.73)	(90.52;92.20)	(88.01;89.90)
<b>SVM</b>	(88.79;90.76)	(88.47;90.87)	(81.25;84.34)	(89.20;91.90)	(91.19;92.76)	(89.16;90.48)
<b>RF</b>	(89.50;91.43)	(89.42;90.64)	(89.56;91.37)	(89.10;91.53)	(89.29;91.25)	(89.81;91.28)
<b>KNN</b>	(84.83;86.81)	(86.05;87.67)	(78.76;80.55)	(88.15;89.14)	(83.52;85.42)	(80.83;82.65)

**Table S3** The average minimum and maximum classification rates (%) across 100 bootstrapped iterations run 10 times based on different pre-treatment approaches for GC-TOF MS *Arabidopsis Thaliana* data (*wild type, tt4* and *mt0 1*)

	<i>Autoscaling</i>	<i>Range</i>	<i>Level</i>	<i>Pareto</i>	<i>Vast</i>	<i>None</i>
<b>PC-DFA</b>	(95.74;97.89)	(95.49;97.99)	(94.49;96.62)	(94.19;96.52)	(97.16;98.19)	(93.44;96.64)
<b>SVM</b>	(95.23;97.18)	(95.06;97.11)	(94.01;95.12)	(91.56;94.99)	(95.42;97.21)	(90.54;91.75)
<b>RF</b>	(86.73;90.50)	(87.28;90.37)	(86.66;90.14)	(87.28;90.61)	(87.33;90.13)	(86.93;90.27)
<b>KNN</b>	(83.38;87.20)	(84.20;87.50)	(81.39;85.26)	(82.11;85.23)	(76.73;78.97)	(69.26;72.65)

**Table S4** The average minimum and maximum classification rates (%) across 100 bootstrapped iterations run for 10 times based on different pre-treatment approaches for NMR *Acute Kidney injury* data (normal vs. injury)

	<i>Autoscaling</i>	<i>Range</i>	<i>Level</i>	<i>Pareto</i>	<i>Vast</i>	<i>None</i>
<b>PC-DFA</b>	(73.58;75.83)	(73.85;75.60)	(67.59;70.49)	(72.88;74.07)	(75.06;76.60)	(68.56;71.52)
<b>SVM</b>	(75.71;77.21)	(74.93;76.95)	(68.29;71.98)	(71.62;73.99)	(74.38;76.69)	(70.70;73.68)
<b>RF</b>	(74.66;76.01)	(74.69;75.52)	(74.17;75.87)	(74.39;75.91)	(74.48;75.76)	(74.38;75.74)
<b>KNN</b>	(67.59;71.00)	(69.87;72.23)	(65.58;67.75)	(68.15;72.28)	(68.93;71.68)	(69.15;71.15)

**Table S5** The average classification rates (%) across 100 bootstrapped iterations with the 95 % CI in parentheses based on log<sub>10</sub> transformation for GC-TOF MS *Arabidopsis Thaliana* data (*wild type, tt4* and *mt0 1*)

	<i>log<sub>10</sub> transformation</i>
<b>PC-DFA</b>	96.52 (96.36-96.68)
<b>SVM</b>	95.45 (95.28-95.62)
<b>RF</b>	89.26 (89.10-89.42)
<b>KNN</b>	87.17 (86.89-87.45)

## References

- Karatzoglou, A., Meyer, D. and Hornik, K. (2006) Support Vector Machines in R. *J STAT SOFTW* **15**.
- Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *R News* **2**, 18-22.
- R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org/>
- Venables, W.N. and Ripley, B.D. (2002) *Modern Applied Statistics with S*, Fourth edn. Springer, New York.