



**Table S2.** Datasets used to compare for the evaluation of different factors.

Factor to evaluate	Datasets to compare
Number of subjects	1, 2, 3, 4
Number of variables	2, 14, 15, 16
Inter-individual variability	2, 8, 9, 10
Intra-individual variability	2, 11, 12, 13
Number of time points	2, 5, 6, 7

**Table S3.** Performance of five PLS models evaluated on simulated dataset with same number of latent variables.

Model	# LV	Q <sup>2</sup>	AUC	AUVSC	# Varsel <sup>a</sup>	# TP <sup>b</sup>
1	2	0.58 (0.02)	0.86 (0.05)	0.89 (0.02)	249.1 (9.1)	55.2 (4.7)
2	2	1 (0)	0.6 (0.08)	0.72 (0.01)	607.8 (10.9)	14.3 (4.7)
3	2	0.77 (0.02)	0.74 (0.02)	0.99 (0)	192.7 (10)	76 (1.4)
4	2	0.6 (0.01)	1 (0)	0.88 (0.02)	171.4 (11.3)	52.4 (3.6)
5	2	0.6 (0.01)	0.75 (0)	0.87 (0.02)	174.3 (11)	52.2 (4.7)

<sup>a</sup> # LV, number of latent variables; <sup>b</sup> # Varsel, number of selected variables; <sup>c</sup> # TP, number of true positives.

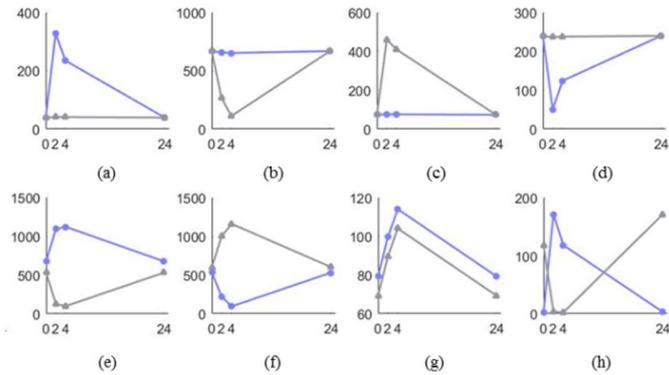
**Table S4.** Parameters used for pre-processing the onion study data in MZmine2.

Batch step	Parameters
Raw data import	
Mass detection	Noise level: 15
Chromatogram builder	Min time span (min): 0.02; Min height: 5.0E1; m/z tolerance: 0.04 m/z or 30 ppm
Chromatogram deconvolution	Chromatographic threshold: 5%; Search minimum in RT range (min): 0.01; Minimum relative height: 5%; Minimum absolute height: 5.0E1; Min ratio of peak/top edge: 1.5; Peak duration range (min): 0.01-0.2
Isotopic pattern	m/z tolerance: 0.08 or 60 ppm; Retention time tolerance: 0.02; Monotonic shape; maximum charge: 1
Join aligner	m/z tolerance: 0.06 or 30 ppm; Absolute retention time tolerance: 0.06; Weight for both m/z tolerance and retention time tolerance: 10
Duplicate peak filter	m/z tolerance: 0.2 or 200 ppm; RT tolerance: 0.01

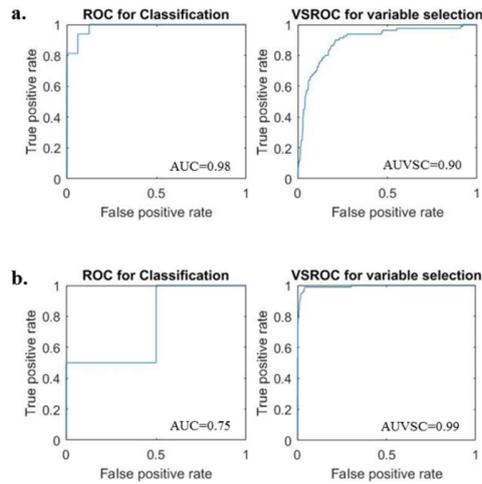
Peak list rows filter	Min peaks in a row: 5 Minimum peaks in an isotope pattern: 1; m/z range: 50-1000; RT range: 0-7;
Peak finder	Intensity tolerance: 1%; m/z tolerance: 0.04 or 30 ppm; Absolute retention time tolerance: 0.02

**Table S5.** Parameters used for pre-processing the coffee study data in MZmine2.

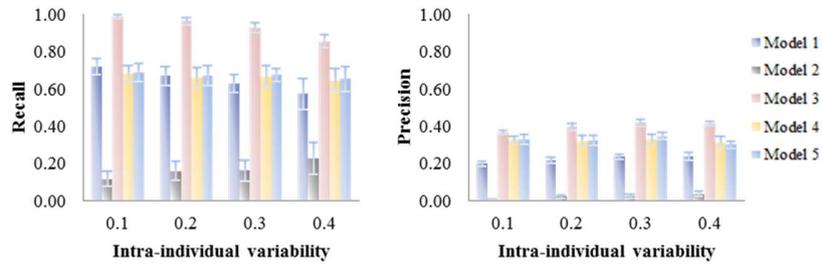
Batch step	Parameters
Raw data import	
Mass detection	Noise level: 20
Chromatogram builder	Min time span (min): 0.02; Min height: 4.0E1; m/z tolerance: 0.04 m/z or 30 ppm
Chromatogram deconvolution	Chromatographic threshold: 5%; Search minimum in RT range (min): 0.01; Minimum relative height: 5%; Minimum absolute height: 4.0E1; Min ratio of peak/top edge: 1.4; Peak duration range (min): 0.01-0.2
Isotopic pattern	m/z tolerance: 0.08 or 60 ppm; Retention time tolerance: 0.02; Monotonic shape; maximum charge: 1
Join aligner	m/z tolerance: 0.06 or 30 ppm; Absolute retention time tolerance: 0.06; Weight for both m/z tolerance and retention time tolerance: 10
Duplicate peak filter	m/z tolerance: 0.2 or 200 ppm; RT tolerance: 0.01
Peak list rows filter	Min peaks in a row: 7 Minimum peaks in an isotope pattern: 1; m/z range: 50-1000; RT range: 0-7;
Peak finder	Intensity tolerance: 1%; m/z tolerance: 0.04 or 30 ppm; Absolute retention time tolerance: 0.02



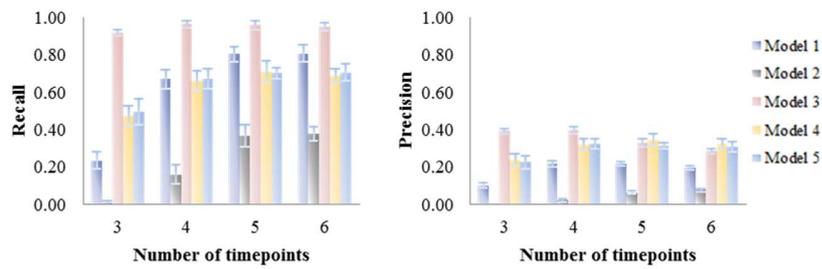
**Figure S1.** Temporal profiles of metabolites in simulated data with a time-series design. (a)-(h) are temporal profiles of eight metabolites simulated for two subject in control (grey) and intervention (purple) group, respectively. For each type of metabolites in each group,  $c$ ,  $a$ ,  $\alpha$ ,  $\beta$  are adjusted to create temporal profiles shown in Fig 1 (a)-(h) correspondingly. For instance, to simulate metabolite (a) in intervention group, the parameters are generated as follows:  $c \sim U(1,100)$ ,  $a \sim U(100,500)$ ,  $\alpha \sim U(5,20)$ ,  $\beta \sim U(40,50)$ .



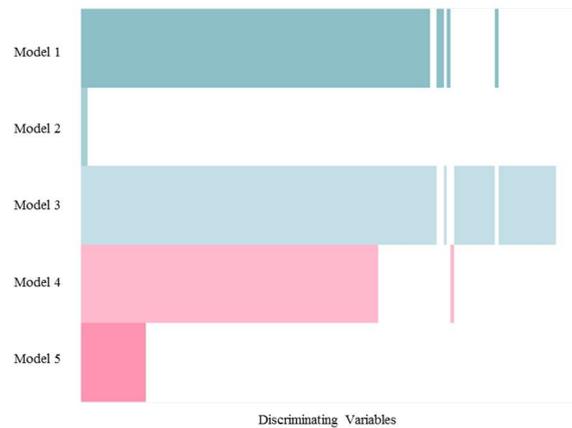
**Figure S2.** Convention ROC curve and Variable Selection ROC curve (VSROC) for a. model 1 and b. model 3 in simulated Dataset 3. Convention ROC and VSROC were used to evaluate the classification and variable selection performance of the model, respectively.



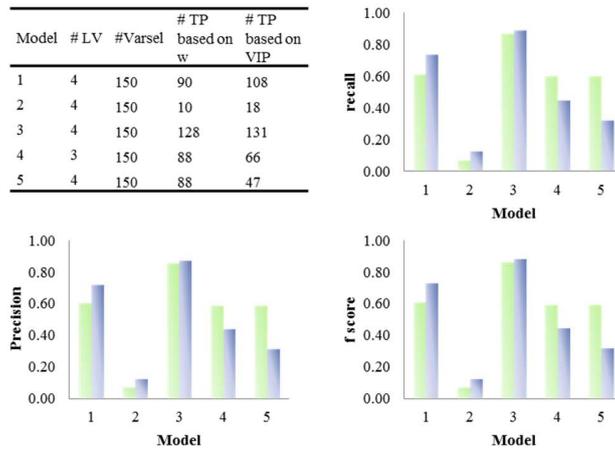
**Figure S3.** Influence of intra-individual variability (0.1-0.4) on the variable selection performance of five PLS models on simulated datasets. Recall and precision were calculated based on the variable selection ROC curve.



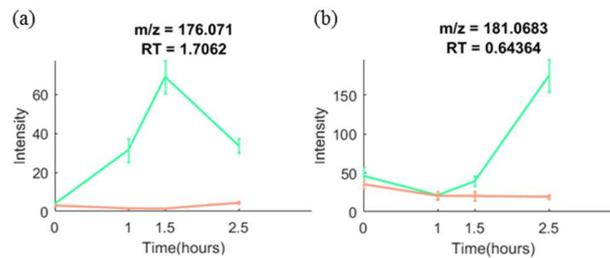
**Figure S4.** Influence of number of time points (3-6) on the variable selection performance of five PLS models on simulated datasets. Recall and precision were calculated based on the variable selection confusion matrix.



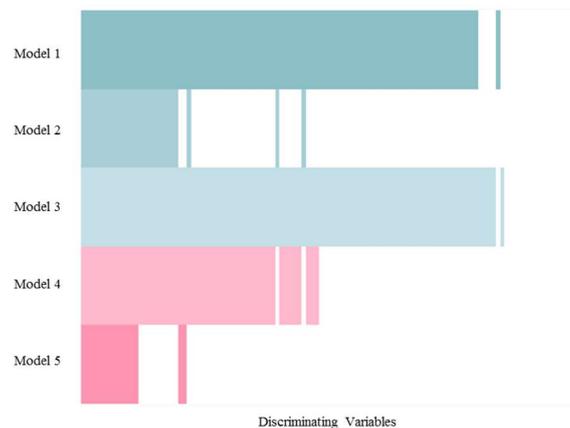
**Figure S5.** Comparison among discriminating variables selected by five PLS models in onion study data. The coloured and white strips represent true positives (selected discriminating variables) and false negatives (unselected discriminating variables), respectively. The discriminating variables were arranged in order so that the variables selected by all five models were on the left side and the variables selected only by one model were on the right side.



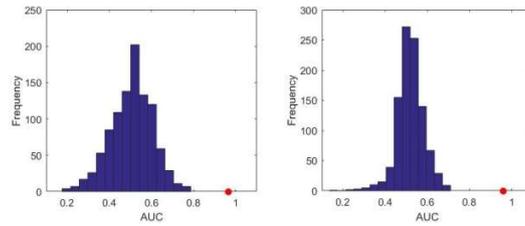
**Figure S6.** Comparison between the variable selection performances based on loading weight (green) and VIP (blue) of five PLS models on onion study data. recall, precision and f-score were calculated based on the variable selection confusion matrix. # LV, number of latent variables; # Varsel, number of selected variables; # TP based on w, true positives selected based on loading weights; # TP based on VIP, true positives selected based on VIP scores. Bootstrap procedure was not used here because when it applied to the variable selection based on loading weights, almost all the variables were selected with the threshold setting to 0 ( $w^* - \sigma_w > 0$ ,  $w^*$  and  $\sigma_w$  are mean and standard deviation of the B w values). 150 variables with highest w or VIP were selected in each model.



**Figure S7.** Temporal profiles of metabolites observed in our coffee data with a time-series design. (a)-(b) are temporal profiles of two metabolites in control (orange) and intervention (green) group, respectively.



**Figure S8.** Comparison among discriminating variables selected by five PLS models in coffee study data. The coloured and white strips represent true positives (selected discriminating variables) and false negatives (unselected discriminating variables), respectively. The discriminating variables were arranged in order so that the variables selected by all five models were on the left side and the variables selected only by one model were on the right side.



**Figure S9.** Area under the ROC curve (AUC) calculated on permuted Y data for model 3 (histogram, 1000 permutations) and original data (red dot) generated on samples obtained from onion (left) and coffee (right) studies. In both cases it is clear that the AUC obtained from the non-permuted Y is in excess of all values obtained through permutation, indicating that the models do not overfit and are significant at an empirical  $p < 0.001$ .