Supplementary Material for

“Improved Quality Control Processing

# of Peptide-Centric LC-MS Proteomics Data”

Melissa M. Matzke[1], Katrina M. Waters[1], Thomas O. Metz[1], Jon M. Jacobs[1],
Amy C. Sims[2], Ralph S. Baric[2], Joel G. Pounds[1] and Bobbie-Jo M. Webb-Robertson[1*]

[1]Pacific Northwest National Laboratory, P.O. BOX 999, Richland, WA 99352 and
[2]University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

1 SIMULATIONS

Simulations of size 500 based on the p-variate standard normal distribution *Np(*0,I*),* and an empirically influenced p-variate normal distribution *Np(*μ,Σ*)* were performed to examine a range of outlier configurations. In addition, we assess the performance of the multi-dimensional outlier detection method against the conventional method of using a correlation coefficient (previously described in the manuscript section 2.1 as metric 1 – Eq. 1) to ascertain whether a LC-MS run is an outlier.

1.1 Np(0,1) SIMULATION

Methods

We employed the method of Maronna and Zamar (2002), as implemented in Filzmoser et al. (2008), to generate correlated multivariate standard normal data for the purpose of exploring the performance of the rMd-PAV outlier detection method on LC-MS based proteomics data. Our simulation study is based on 100 simulate runs, with α outlier runs and (100 – α) non-outlier runs, with the simulation component repeated 500 times.

We start by generating the (100 – α) non-outliers from a p-variate normal distribution $Np(\mathbf{0},\boldsymbol{I})$, where p = q = 5; μ_{1x5} = average metric value from the metric matrix of the real LC-MS data set; and, Σ_{5x5} = covariance matrix dervived the (n x 5) metric matrix. The α outliers are generated from $Np(\boldsymbol{z_0}, \boldsymbol{kI})$ such that k is a scalar which determines the scatter of the outlier values from the rest of the data, and $\boldsymbol{z_0} = c\boldsymbol{a_0}$ where $\boldsymbol{a_0} = \left(b_1 - \bar{b},\ldots,b_p - \bar{b}\right) \Big/ \sqrt{\sum_{j=1}^{p}\left(b_j - \bar{b}\right)^2}$ such that $\boldsymbol{b} = (b_1, \ldots, b_p)$ consists of random draws from a U(0,1) and $\bar{b}$ is the arithmetic mean of $\boldsymbol{b}$**.** We explored the space for α = 1, 5 and 10; c = 0, 5 and 10; and, k = 0.1, 2 and 5. We then combine the non-outliers and outliers in a single data set ***X*** and introduce correlation by multiplying by ***R,*** which is defined as a (5 x 5) matrix with 1's on the diagonal and $\rho = 0.5$ on the off diagonal. And, to be able to make a comparison of performance to use of the correlation metric alone, we randomly selected a column from the 100 x 5 simulated metric matrix to represent the values of the correrlation metric.

For each simulation of 100 runs we collected the true positive rate (TPR) and false positive rate (FPR) for the results from the outlier detection method and using the correlation coefficient alone. The TPR and FPR were used to generate a ROC curve. The area under the ROC curve was calculated for each iteration of the simulation. This was repeated 500 times for each combination of number of outliers, location and scatter values, and the results averaged.

1 out of 100 (1%) outliers

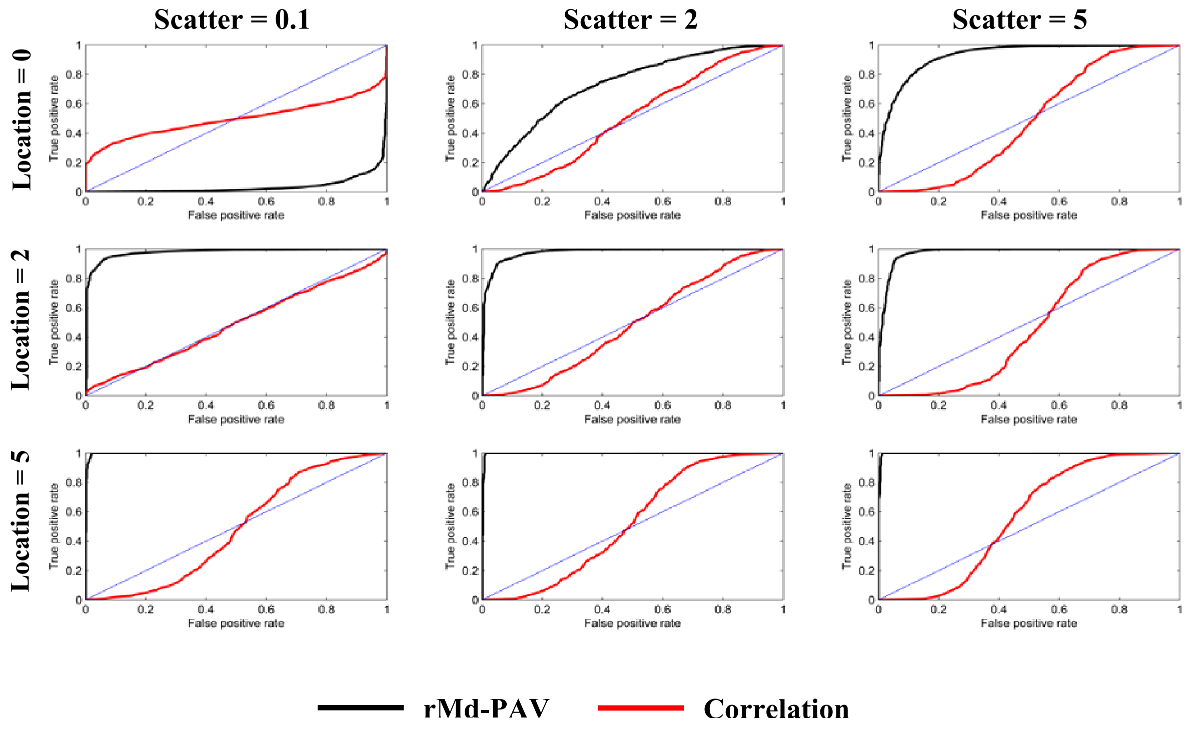

Figure S1. The average ROC curve from 500 simulations of a (100 x 5) metric matrix in which 1 out of 100 runs is a statistical outlier (black line), and the corresponding ROC curve for the simulated correlation vector (red line). Across the matrix of location and scatter values, with the exception of location = 0 and scatter = 0.1, using the $N_p(0,1)$ simulated metric matrix to calculate rMd-PAV scores to identify outliers significantly outperforms using the simulated correlation coefficient alone for identifying the outlier values (Wilcoxon sign rank p-value << 0.0001).

5 out of 100 (5%) outliers

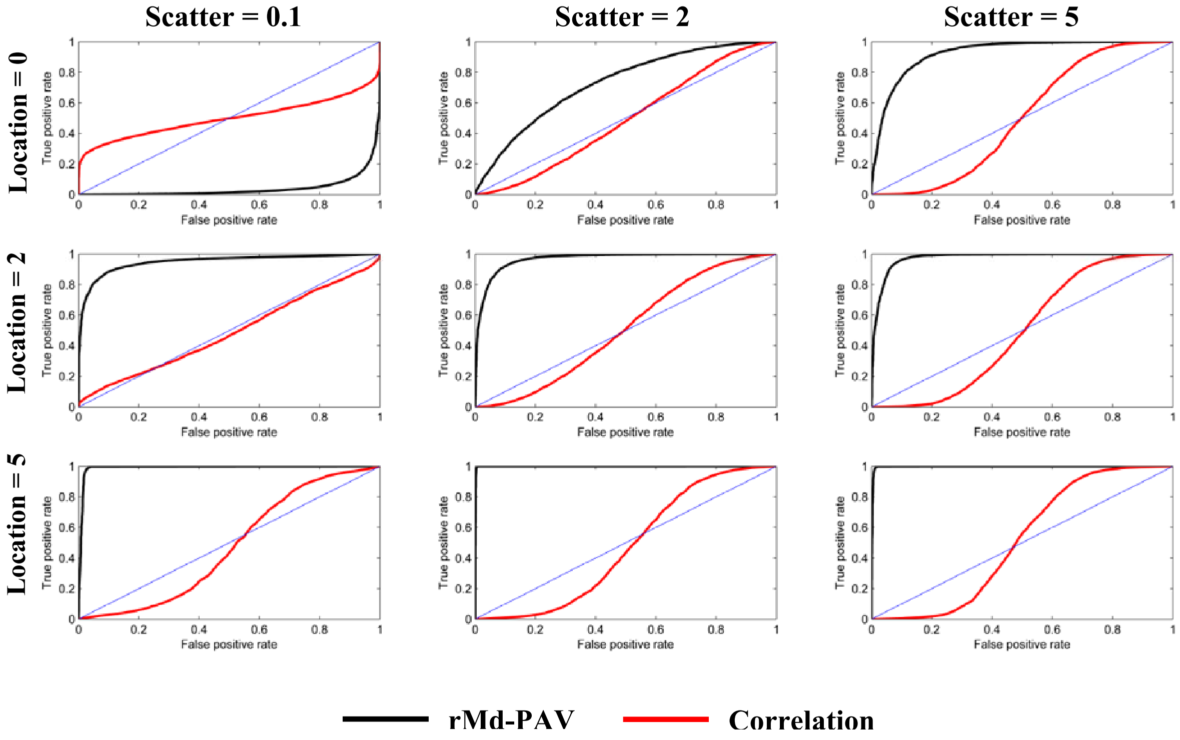

Figure S2. The average ROC curve from 500 simulations of a (100 x 5) metric matrix in which 5 out of 100 runs is a statistical outlier (black line), and the corresponding ROC curve for the simulated correlation vector (red line). Across the matrix of location and scatter values, with the exception of location = 0 and scatter = 0.1, using the $N_p(0,1)$ simulated metric matrix to calculate rMd-PAV scores to identify outliers significantly outperforms using the simulated correlation coefficient alone for identifying the outlier values (Wilcoxon sign rank p-value << 0.0001).

10 out of 100 (5%) outliers

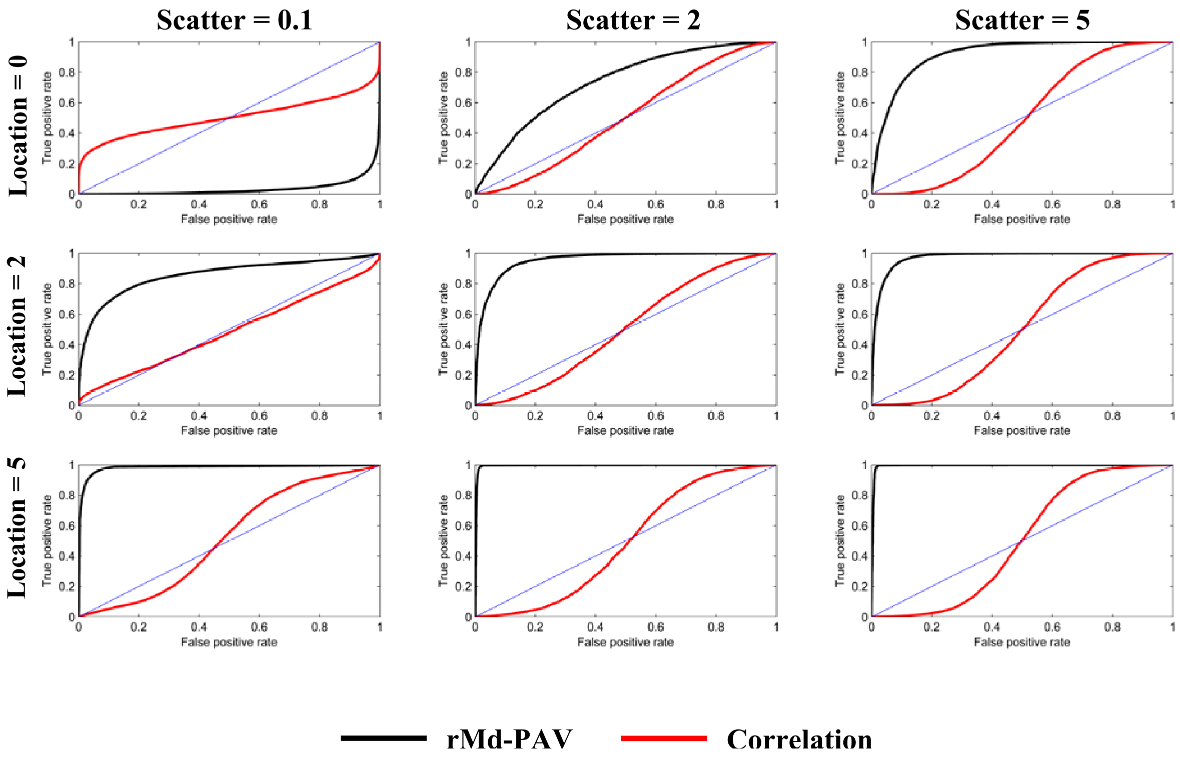

Figure S3. The average ROC curve from 500 simulations of a (100 x 5) metric matrix in which 10 out of 100 runs is a statistical outlier (black line), and the corresponding ROC curve for the simulated correlation vector (red line). Across the matrix of location and scatter values, with the exception of location = 0 and scatter = 0.1, using the $N_p(0,1)$ simulated metric matrix to calculate rMd-PAV scores to identify outliers significantly outperforms using the simulated correlation coefficient alone for identifying the outlier values (Wilcoxon sign rank p-value < 0.0001).

1.2 RESTRICTED SIMULATION

Methods

We followed the method of Penny and Jolliffe (2001) using real LC-MS data as the basis of a restricted p-variate normal distribution $Np(\mu,\Sigma)$. Although restricted, this study reflects a covariance structure observed in the (n x 5) metric matrix of a real LC-MS data set (human plasma samples with mass analysis performed on a LTQ-Orbitrap^{TM}) which is believed to have no outlier runs (a detailed study description is listed below). Our simulation study is based on 100 simulated runs, with α outlier runs and (100 – α) non-outlier runs, with the simulation component repeated 500 times.

We start by generating the (100 – α) non-outliers from a p-variate normal distribution $Np(\mu,\Sigma)$, where p = q = 5; μ_{1x5} = average metric value from the metric matrix of the real LC-MS data set; and, Σ_{5x5} = covariance matrix derived the (n x 5) metric matrix. The α outliers are generated from a $Np(\mu,k\Sigma)$ such that k is a scalar which determines the scatter of the outlier values from the rest of the data. We explored the space for α = 1, 5, and 10; and, k = 12.25 , 16, 20.25, and 25, which represents 3.5, 4, 4.5 and 5 standard deviations (SD).

For each simulation of 100 runs, we collected the true positive rate (TPR) and false positive rate (FPR) for the results using both the multivariate outlier detection method and the correlation coefficient, R_i, alone. The TPR and FPR were used to generate a Receiver Operating Characteristic (ROC) curve. The area under the ROC curve (AUC) was calculated for each iteration of the simulation. This was repeated 500 times for each α and k combination, and the results were averaged.

Study Description

Human plasma samples were analyzed using a LTQ-Orbitrap^{TM} mass spectrometer (Thermo Electron Corp., Waltham, MA). Nanoelectrospray ionization was used in the analysis of all samples. Spectra were collected at 400-2000 *m/z* with a resolution of 100k and analyzed using the accurate mass and elution time (AMT) tag approach. The mass deisotoping process was performed using Decon2LS, and the matching process was performed using VIPER. Features from the LC-MS analyses were matched to AMT tags to identify peptides, using an initial tolerance of +/- 3 ppm for mass and 0.025% for the LC normalized elution time (NET). The peptide datasets were further processed to remove peptides identified with low confidence, using the uniqueness filter Statistical Likelihood Confidence (SLiC) score of 0.5 and a DelSLiC of 0.2. In circumstances where a peptide was identified in some runs, but not others, the missing data were coded as 'NaN'. All peptide abundance values were transformed to the $\log_{10}$ scale. Minimum occurrence data filters were used to identify those peptides for which the amount of data present was not adequate for differential abundance analysis.

Plasma samples of 28 representative individuals from a cohort of 500 tobacco smokers or non-smokers, determined to be either obese or non-obese based on body mass index (BMI), were selected for quantitative proteome analysis. Each plasma sample was analyzed in duplicate or triplicate technical runs resulting in a total of 59 runs. The number of samples for the 2-factor (Smoking Status x BMI) study is provided in Supplementary Table S1. A total of 4,686 peptides were retained in the final dataset based on the minimum occurrence filter.

Results

A comparison of the ROC curves for the rMd-PAV scores and correlation alone by a Wilcoxon sign rank test results in no significant differences between the curves for 5% outliers and scatter of 3.5 SD (k=12.25), however rMd-PAV significantly outperforms correlation alone in the identification of outlier LC-MS runs for 5% outliers and scatter of 5 SD (k=25) (one-sided p-value < 0.0001). These results are consistent across α values. Results for α = 5 and k = 12.25 and 25, are in Figure S4.

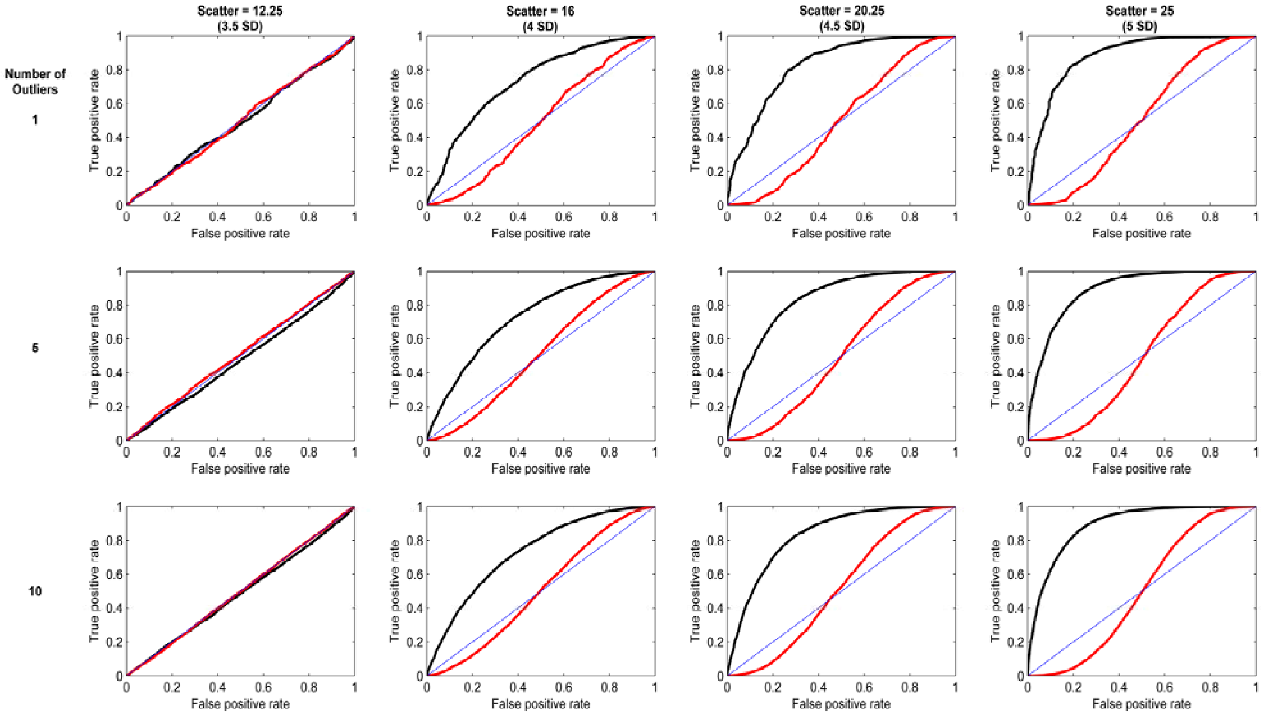

Figure S4. The average ROC curve from 500 restricted simulations of the (100 x 5) metric matrix from *Np*(**μ**,**Σ**) for which the number of statistical outliers is 1, 5 and 10 out of 100 simulated metric vectors and the scatter of the data is 12.25 (3.5 SD), 16 (4 SD), 20.25 (4.5 SD) and 25 (5 SD) (red line), and the corresponding ROC curve for the simulated correlation vector (red line). The *Np*(**μ**,**Σ**) simulated metric matrix to calculate rMd-PAV scors to identify outliers significantly outperforms using the simulated correlation coefficient alone for identifying the outlier values for all α number of outliers and scatter values of 16, 20.25 and 25 (Wilcoxon sign rank one-sided p-value < 0.0001). There is no evidence of statistically significant differences between the curves for scatter value of 12.25 across the number of outliers (α).

Table S1. Human plasma data set summary

Group	BMI	Smoking Category[a]	Group Size	Total Runs	Number of Replicates per Sample						
					1	2	3	4	5	6	7
1	Obese	NS	7	15	2	2	2	2	3	2	2
2	Obese	S	7	16	2	3	2	2	2	3	2
3	Normal	NS	7	14	2	2	2	2	2	2	2
4	Normal	S	7	14	2	2	2	2	2	2	2

[a] NS: Nonsmoking; S: Smoking

Table S2. Human calu-3 cell culture data set information

Group	Hour	Sample Size	Total Runs	Number of Replicates per Sample		
				Sample 1	Sample 2	Sample 3
Sham	0	3	9	3	3	3
	3	3	10	3	3	4
	7	3	9	3	3	3
	12	3	7	2	3	2
	24	3	10	3	4	3
	30	3	8	2	3	3
	36	3	8	3	2	3
	48	3	9	3	3	3
icSARS-CoV	0	3	8	3	3	2
	3	3	9	3	3	3
	7	3	10	4	3	3
	12	3	9	3	3	3
	24	3	9	3	3	3
	30	3	8	2	3	3
	36	3	9	3	3	3
	48	3	9	3	3	3

Table S3. Cigarette smoke exposure data set information

BMI	Inhalation Method	Sample Size	Total Runs	Number of Replicates per Sample							
				S1	S2	S3	S4	S5	S6	S7	S8
Normal	Sham	8	16	2	2	2	2	2	2	2	2
	Side Stream	8	15	2	1	2	2	2	2	2	2
	Main Stream	8	16	2	2	2	2	2	2	2	2
Obese	Sham	8	16	2	2	2	2	2	2	2	2
	Side Stream	8	19	2	2	2	2	3	3	3	2
	Main Stream	8	16	2	2	2	2	2	2	2	2