Supplementary material for ’Integrating multi-OMICS data through sparse Canonical Correlation Analysis in predicting outcomes: A comparison study’

Theodoulos Rodosthenous [1], Vahid Shahrezaei [1] and Marina Evangelou [1]

[1]Department of Mathematics, Imperial College London, London, SW7 2AZ, United Kingdom

S1. Selection of tuning parameters

As with all penalisation methods, selecting the "right" tuning parameters is vital in the performance of the algorithm. In sCCA methods, it is necessary to select the optimal choices for the parameters (τ_w) of all input data-sets, which are likely to differ from each other and influence one another.

To achieve this, k-fold cross-validation is performed on different values for those parameters. The aim of sCCA methods is to maximise the canonical correlation and thus the selection of tuning parameters is based on that. Suppose we have X_1 and X_2, then the following measure is taken for every choice of τ_{w_1} and τ_{w_2}, usually within a range of values in $(0, 0.3)$:

$$\Delta_{cor} = \frac{1}{k}\sum_{l=1}^{k} |cor(X_{1_l}\hat{w}^{(1)^{-l}}, X_{2_l}\hat{w}^{(2)^{-l}})| \tag{1}$$

where X_{1_l} and Y_{2_l} represents the testing sets of X_1 and X_2 for fold l, respectively, and $\hat{w}^{(1)^{-l}}, \hat{w}^{(2)^{-l}}$ the estimated canonical variate pair based on the training sets.

In order to determine the optimal tuning parameters at each run of the algorithms, one must first compute Δ_{cor} for all choices of τ_{w_1}, τ_{w_2} for all k folds. The values of τ_{w_1}, τ_{w_2} that maximise Δ_{cor} are then taken as optimal.

Due to the iterative nature of the algorithms, the choice of τ_{w_1} will influence the final outcome of X_2 as well. Hence, selecting the optimal tuning parameters in a multiple-data setting is more complicated and computationally heavy. With two data-sets, $|cor(X_{1_l}\hat{w}^{(1)^{-l}}, X_{2_l}\hat{w}^{(2)^{-l}})| = |cor(X_{2_l}\hat{w}^{(2)^{-l}}, X_{1_l}\hat{w}^{(1)^{-l}})|$, and so we can compute Δ_{cor} once for every combination of τ_{w_1}, τ_{w_2} in each fold. With M data-sets, Δ_{cor} is computed M times. In multiple sCCA, eq. 1, is replaced by:

$$\Delta_{cor} = \frac{1}{Mk}\sum_{l=1}^{k}\sum_{m=1}^{M} |cor(X_{m_l}\hat{w}^{(m)^{-l}}, \sum_{j\neq m} X_{j_l}\hat{w}^{(j)^{-l}})| \tag{2}$$

The time complexity of selecting tuning parameter in multiple data-sets is notably high. To reduce it, a threshold in the correlation values was used. Even though the optimal selection may not be guaranteed, well-performed tuning parameters are selected based on the threshold.

S2. Computing the additional canonical vectors

In computing the additional canonical vectors, we argued (see eq. 14 from the main body of the paper) that by fixing $\mathbf{w}^{(2)}$ and letting $\tilde{X} = \begin{bmatrix} X_1 \\ W_1^T X_1^T X_1 \\ W_2^T X_2^T X_1 \end{bmatrix}$, and $\tilde{Y} = \begin{bmatrix} Y \\ W_2^T X_2^T X_2 \\ W_1^T X_1^T X_2 \end{bmatrix}$ we have that

$$-\mathbf{w}^{(1)^T} X_1^T X_2 \mathbf{w}^{(2)} = -\mathbf{w}^{(1)^T} \tilde{X}_1^T \tilde{X}_2 \mathbf{w}^{(2)} \tag{3}$$

This is true since the the constraints $W_1^T X_1^T X_1 \mathbf{w}^{(1)} = \mathbf{0}_{r-1}$ and $W_2^T X_2^T X_2 \mathbf{w}^{(2)} = \mathbf{0}_{r-1}$ hold and:

$$\begin{align}
-\mathbf{w}^{(1)^T} \tilde{X}_1^T \tilde{X}_2 \mathbf{w}^{(2)} &= -\mathbf{w}^{(1)^T} \begin{bmatrix} X_1 \\ W_1^T X_1^T X_1 \\ W_2^T X_2^T X_1 \end{bmatrix}^T \begin{bmatrix} X_2 \\ W_2^T X_2^T X_2 \\ W_1^T X_1^T X_2 \end{bmatrix} \mathbf{w}^{(2)} \tag{4} \\
&= -\mathbf{w}^{(1)^T} (X_1^T X_2 + (W_1^T X_1^T X_1)^T W_2^T X_2^T X_2 + (W_2^T X_2^T X_1)^T W_1^T X_1^T X_2) \mathbf{w}^{(2)} \tag{5} \\
&= -\mathbf{w}^{(1)^T} X_1^T X_2 \mathbf{w}^{(2)} - \mathbf{w}^{(1)^T} (X_1^T X_1 W_1 W_2^T X_2^T X_2) \mathbf{w}^{(2)} \tag{6} \\
&\quad - \mathbf{w}^{(1)^T} (X_1^T X_2 W_2 W_1^T X_1^T X_2) \mathbf{w}^{(2)} \tag{7} \\
&= -\mathbf{w}^{(1)^T} X_1^T X_2 \mathbf{w}^{(2)} \tag{8}
\end{align}$$

We can use the respective algorithms for RelPMDCCA on $\tilde{X}_1$ and $\tilde{X}_2$ to obtain the remaining canonical variate pairs.

S3. Evaluation measures and Simulation Scenarios

The closer the estimated canonical variate pairs are to the true pairs, the better the performance of sCCA methods. We took similar measures as Bonner and Beyene [2017] used in evaluating sparse PCA approaches.

The primary criteria in the evaluation will be the classification of zero-valued and non-zero-valued elements of the canonical vectors, since these would signify the grouping structure. In addition to the structure of the estimated $\hat{\mathbf{w}}^{(1)}$ and $\hat{\mathbf{w}}^{(2)}$, it is important to estimate values close to the true ones. Hence, we also measured numerical differences between true and estimated values of the canonical vectors.

In particular, we examined the performance of the simulations based on the following measures:

1. **NZ**: The number of non-zero values remaining in the estimated pairs. Expecting a sparse representation
2. **TRUENZ**: The number of correctly classified **non-zero** values
3. **TRUEZ**: The number of correctly classified **zero** values
4. **ANGLE**: A measure of the distance between true and estimated canonical variate pairs. A value between 0 (perfect) and 1(worst) It is calculated for each $\mathbf{w}^{(i)}$, separately as follows: $\mathbf{ANGLE}(\hat{\mathbf{w}}, \mathbf{w}) = dist(\hat{\mathbf{w}}, \mathbf{w}) = \sqrt{1 - (\mathbf{w}^T \hat{\mathbf{w}})^2}$, where $\mathbf{w}$ is the true canonical variate and $\hat{\mathbf{w}}$ is the estimated one.

5. **LOSS** (X. Suo *et al.*, 2017): A loss value between the true and the estimated canonical correlation pairs. For each element of the pair, $\mathbf{w}$, the loss is computed as follows:

$$\mathbf{LOSS}(\hat{\mathbf{w}},\mathbf{w}) = \min(||\hat{\mathbf{w}}-\mathbf{w}||_2^2, ||\hat{\mathbf{w}}+\mathbf{w}||_2^2) \tag{9}$$

6. **CORR**: The estimated canonical correlation $Cor(X_1\hat{\mathbf{w}}^{(1)}, X_2\hat{\mathbf{w}}^{(2)})$

Note that all measures except **CORR**, were computed separately for all $\hat{\mathbf{w}}^{(i)}, \quad i=1,2$.

The measures **TRUENZ** and **TRUEZ** are not very intuitive on their own, especially since simulations were conducted with different parameters. That is, a different true number of non-zero values are to be identified, depending on the scenario performed. Hence, a confusion matrix of the zero and non-zero elements found, against the simulated truth, was computed. True Positive Rate (**TPR**), False Positive Rate (**FPR**), Positive Predictive Value (**PPV**), Negative Predictive Value (**NPV**) and Accuracy (**ACC**) were then obtained.

Overall, the evaluation measures cover the entire performance of the methods: the correct identification of non-zero (and zero) values, the exact values of the canonical vectors (through **ANGLE** and **LOSS**) and the correlation within the canonical pairs.

The covariance-based data generating model, used in the simulation studies to generate two datasets is as follows. Suppose, $X_1 \in \mathbb{R}^{n\times p_1}$ and $X_2 \in \mathbb{R}^{n\times p_2}$ with data being generated by:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{X_1X_1} & \Sigma_{X_1X_2} \\ \Sigma_{X_2X_1} & \Sigma_{X_2X_2} \end{pmatrix}\right]$$

where $\Sigma_{X_1X_2} = \rho\Sigma_{X_1X_1}\mathbf{w}^{(1)}\mathbf{w}^{(2)^T}\Sigma_{X_2X_2}$, with $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ being the true canonical vectors and ρ the true canonical correlation. The covariance matrices $\Sigma_{X_1X_1}$ and $\Sigma_{X_2X_2}$ are explicitly defined based on the type of data generating model.

S4. Algorithms

```
Data: $M$ datasets, $X_m \in \mathbb{R}^{n \times p_m}, \quad \forall m \in \{1, \cdots, M\}$
Tuning parameters: $\tau_{w_m}, \quad \forall m \in \{1, \cdots, M\}$
Result: Canonical vectors, $\mathbf{w}^{(m)}, \quad \forall m \in \{1, \cdots, M\}$
begin
    Select tuning parameters $\tau_{w_m}, \quad \forall m \in \{1, \cdots, M\}$ via cross-validation
    Initialise canonical vector $(\mathbf{w}^{(m)})^0, \quad \forall m \in \{1, \cdots, M\}$ and set $k = 0$
    while not converged do
        for m=1 to M do
            Compute $K_{mj} = \Sigma_{X_m X_m}^{-\frac{1}{2}} \Sigma_{X_m X_j} \Sigma_{X_j X_j}^{-\frac{1}{2}}, \quad \forall j \neq m$
            $(\mathbf{w}^{(m)})^{k+1} \leftarrow \sum_{j \neq m} K_{mj} (\mathbf{w}^{(m)})^k$
            Normalise $(\mathbf{w}^{(m)})^{k+1} = \frac{(\mathbf{w}^{(m)})^{k+1}}{||(\mathbf{w}^{(m)})^{k+1}||}$
            Apply soft-thresholding: $(w_l^{(m)})^{k+1} = S((w_l^{(m)})^{k+1}, \frac{1}{2}\tau_{w_m}), \quad l = 1, \cdots, p_m$
            Normalise $(\mathbf{w}^{(m)})^{k+1} = \frac{(\mathbf{w}^{(m)})^{k+1}}{||(\mathbf{w}^{(m)})^{k+1}||}$
        end
        $k \leftarrow k + 1$
    end
end
```

Algorithm 1: Multiple ConvCCA

```
Data: $X_1 \in \mathbb{R}^{n \times p_1}, \quad X_2 \in \mathbb{R}^{n \times p_2}$
Result: $R$ canonical vectors combined in matrices $W_1 \in \mathbb{R}^{n \times R}, \quad W_2 \in \mathbb{R}^{n \times R}$
begin
    Compute first canonical vector via an sCCA method
    for r=2 to R do
        Let $\tilde{X} = \begin{bmatrix} X_1 \\ W_1^T X_1^T X_1 \\ W_2^T X_2^T X_1 \end{bmatrix}$, and $\tilde{Y} = \begin{bmatrix} Y \\ W_2^T X_2^T X_2 \\ W_1^T X_1^T X_2 \end{bmatrix}$
        Compute the $r^{th}$ canonical vector by applying an sCCA method on $\tilde{X}$ and $\tilde{Y}$ to obtain
        $\mathbf{w}_r^{(j)}, \quad j = 1, 2$
        Update $W_j \leftarrow \left[ W_j, \mathbf{w}_r^{(j)} \right], \quad j = 1, 2$
    end
end
```

Algorithm 2: Computing the additional canonical vectors

S5. Additional Null Simulation Model

In the main body of the paper, a null simulation model is described in which a low true canonical correlation was assumed, $\rho = 0.1$. An additional null model was implemented where two independent and uncorrelated datasets were simulated. Two independent normally distributed multivariate datasets with 80 and 60 features respectively were generated ($X_1 \sim \mathcal{N}(0, \mathbf{I}_{80})$ and $X_2 \sim \mathcal{N}(0, \mathbf{I}_{60})$). The results of this simulation study were in agreement with the results presented in Section 3.2.2. Here, we present the canonical correlations obtained in this study.

Canonical correlation on Null simulation model

Sample size	**PMDCCA LASSO**	**ConvCCA LASSO**	**ConvCCA SCAD**	**RelPMDCCA LASSO**	**RelPMDCCA SCAD**
$n = 100$	0.51 (0.08)	0.88 (0.05)	0.87 (0.07)	0.98 (0.02)	0.98 (0.02)
$n = 1000$	0.22 (0.08)	0.47 (0.08)	0.48 (0.07)	0.51 (0.03)	0.49 (0.02)
$n = 10000$	0.07 (0.06)	0.13 (0.05)	0.15 (0.07)	0.16 (0.03)	0.20 (0.04)

Table 1: **Null Simulation Model.** Canonical correlations of PMDCCA, ConvCCA and RelPMDCCA averaged across 100 runs on the null scenarios.

As in the simulation model with low true canonical correlation, the correlation obtained by the three methods decreases as the sample size increases.

S6. Computation Time

Computation Time (minutes)

	PMDCCA LASSO	**ConvCCA LASSO**	**ConvCCA SCAD**	**RelPMDCCA LASSO**	**RelPMDCCA SCAD**
Null, $n = 100$	0.02	0.08	0.08	1.19	2.43
Null, $n = 1000$	0.34	1.89	2.14	8.23	9.46
Null, $n = 10000$	2.02	5.23	5.67	15.65	17.43
Scenario 1	0.02	0.09	0.08	1.64	2.96
Scenario 2	0.37	1.08	3.03	10.05	10.61
Scenario 3	0.41	3.03	3.35	12.45	13.45
Scenario 4	1.23	3.26	2.79	14.56	13.76
Scenario 5	0.55	2.19	2.13	6.57	6.28
Scenario 6	0.44	1.12	2.12	6.89	5.29

Table 2: **Computation time.** Averaged time (in minutes) taken to run a single iteration for each simulation scenario.

S7. Supplementary Figures

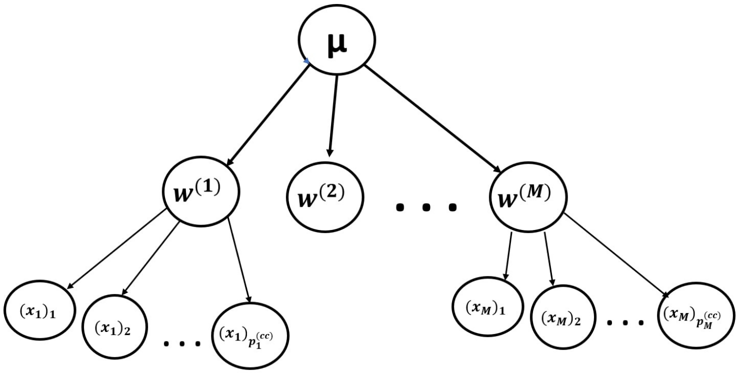

Figure 1: The single latent variable model simulation

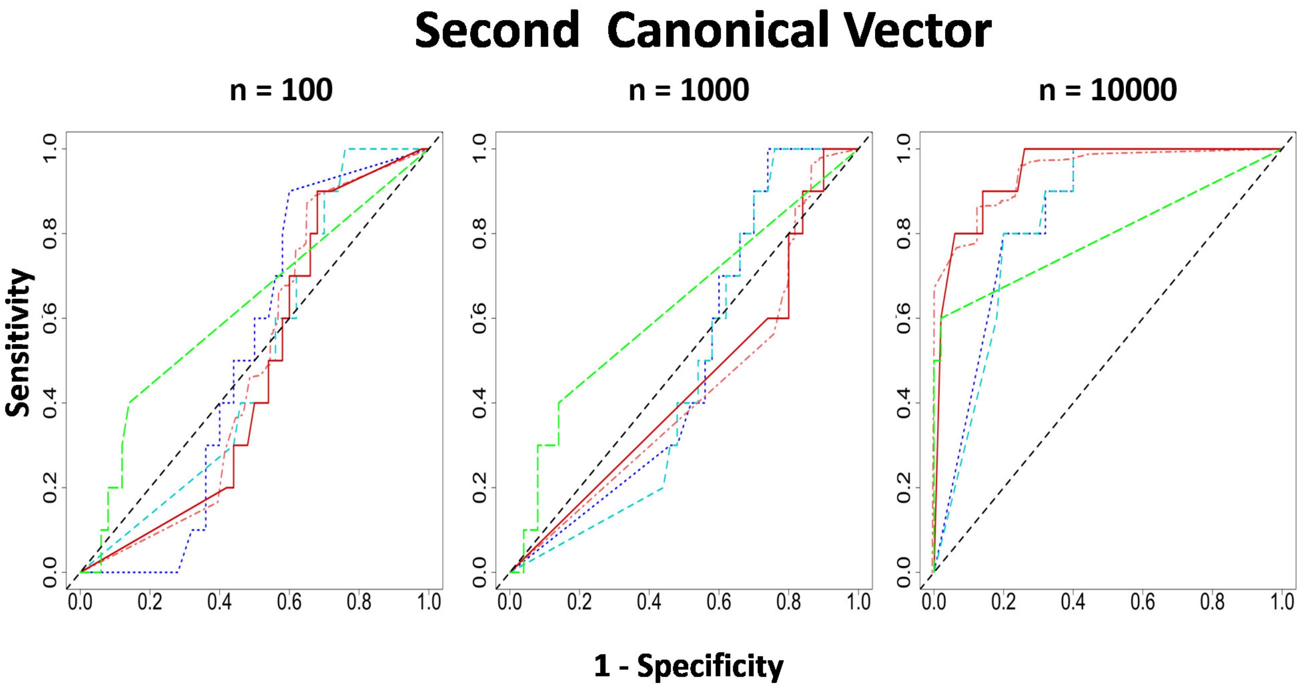

Figure 2: **sCCA performance on Null scenario with low true canonical correlation.** ROC curves of the second canonical vector by all sCCA methods on Null scenario with sample sizes $n = 100, 1000, 10000$

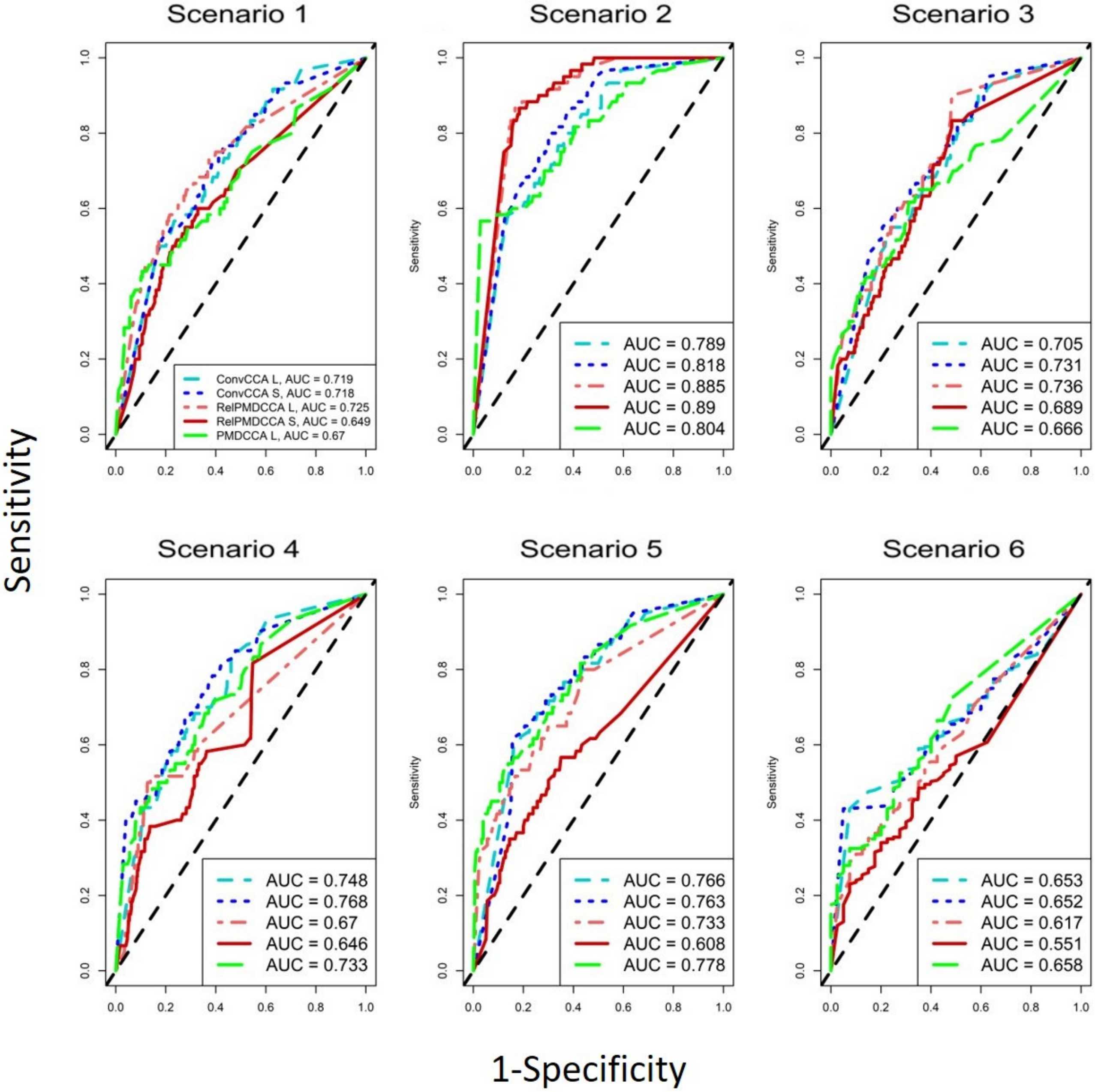

Figure 3: ROC curve plots, showing averaged results (over the models) for each scenario on $X_2\mathbf{w}^{(2)}$.

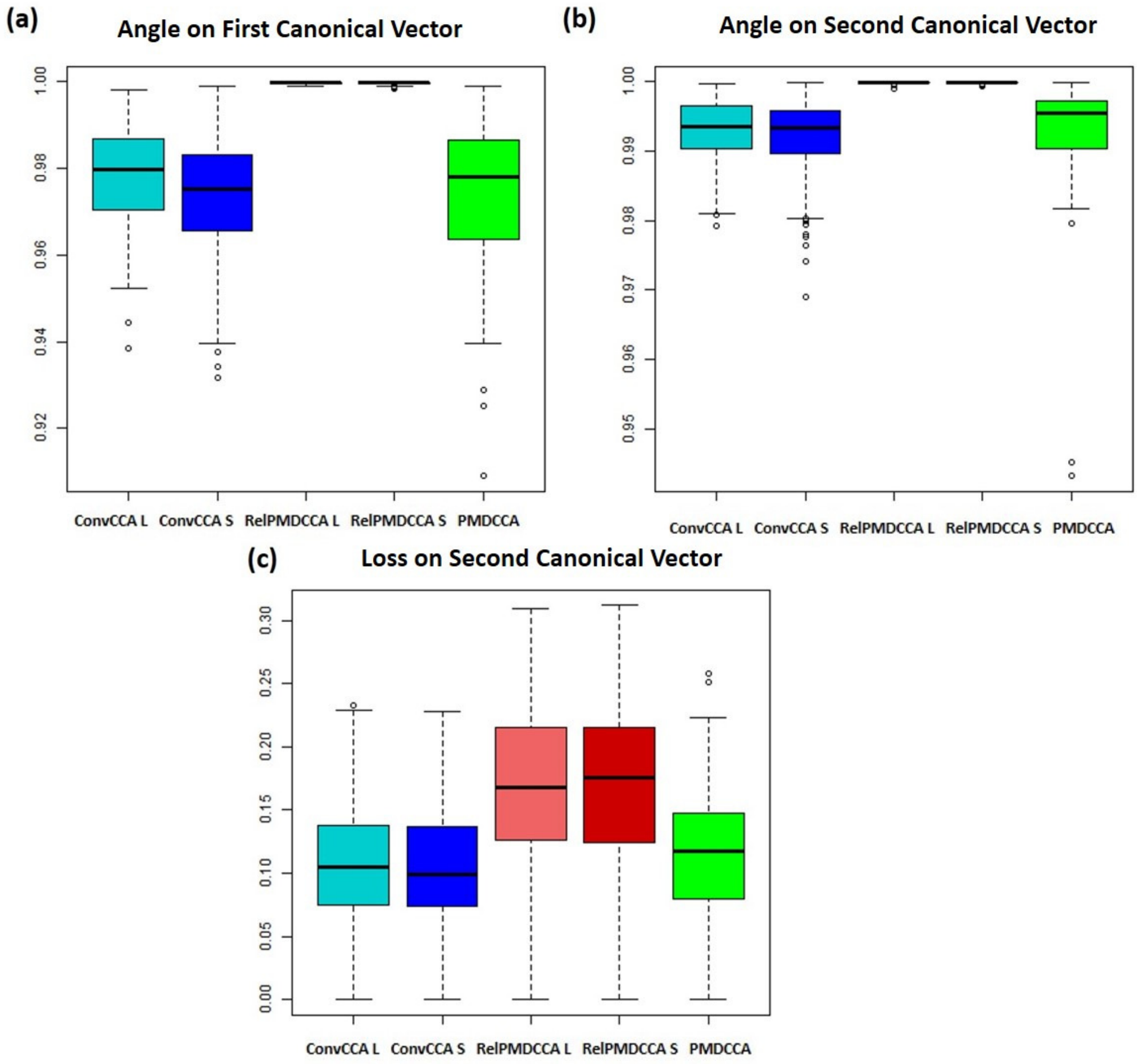

Figure 4: **(a)** Angle evaluation measure on the first canonical vector. **(b)** Angle on the second canonical vector **(c)** Loss on the second canonical vector

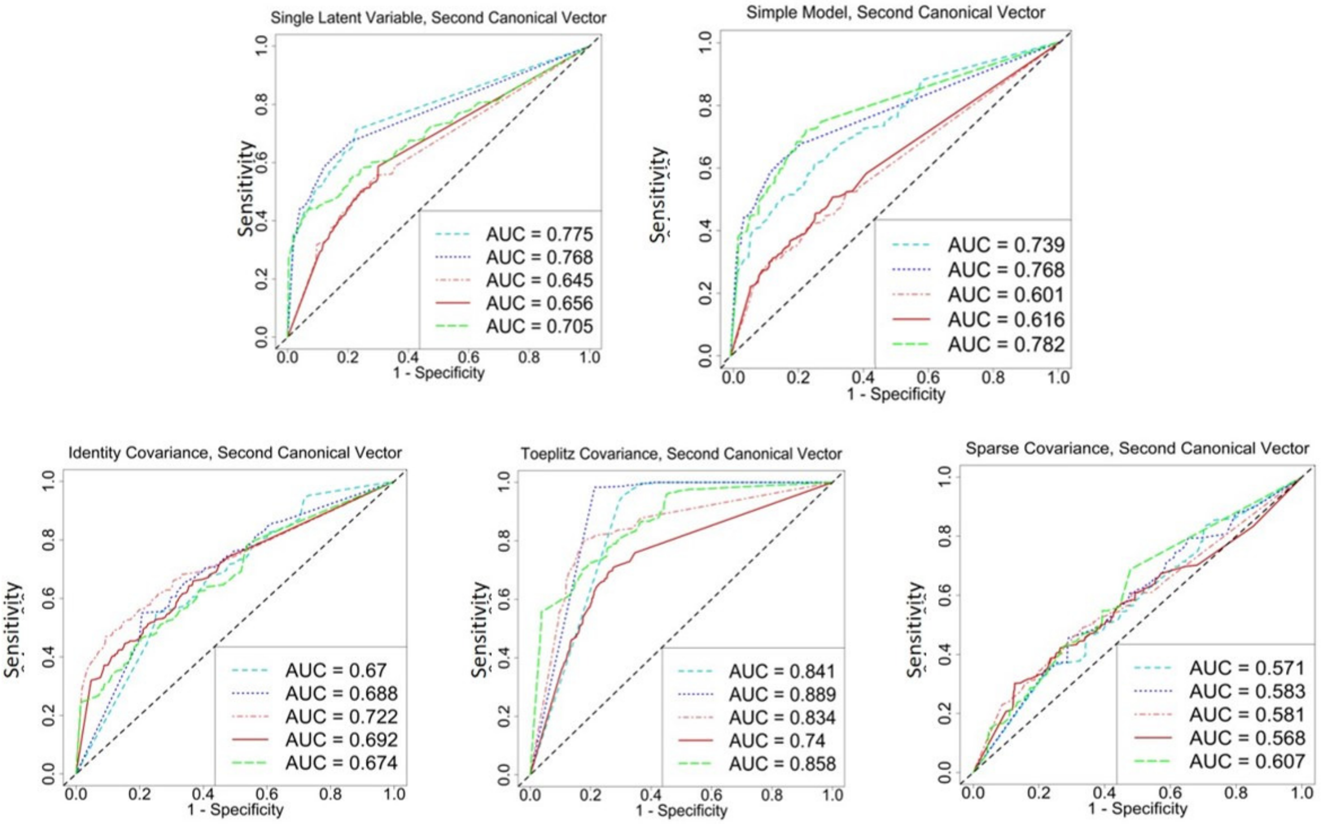

Figure 5: ROC curve plots, showing averaged (over scenarios) results for each model on $X_2\mathbf{w}^{(2)}$.

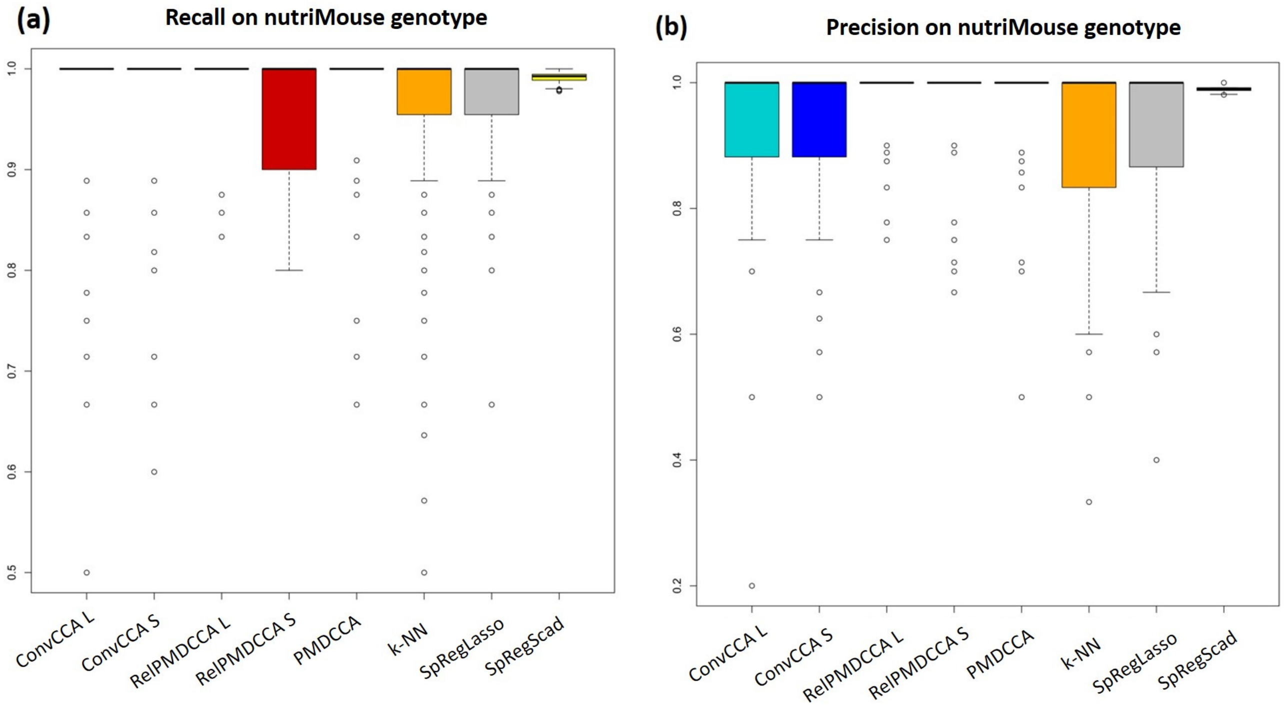

Figure 6: **(a)** Recall and **(b)** precision measures in nutriMouse study in predicting the genotype of mice.

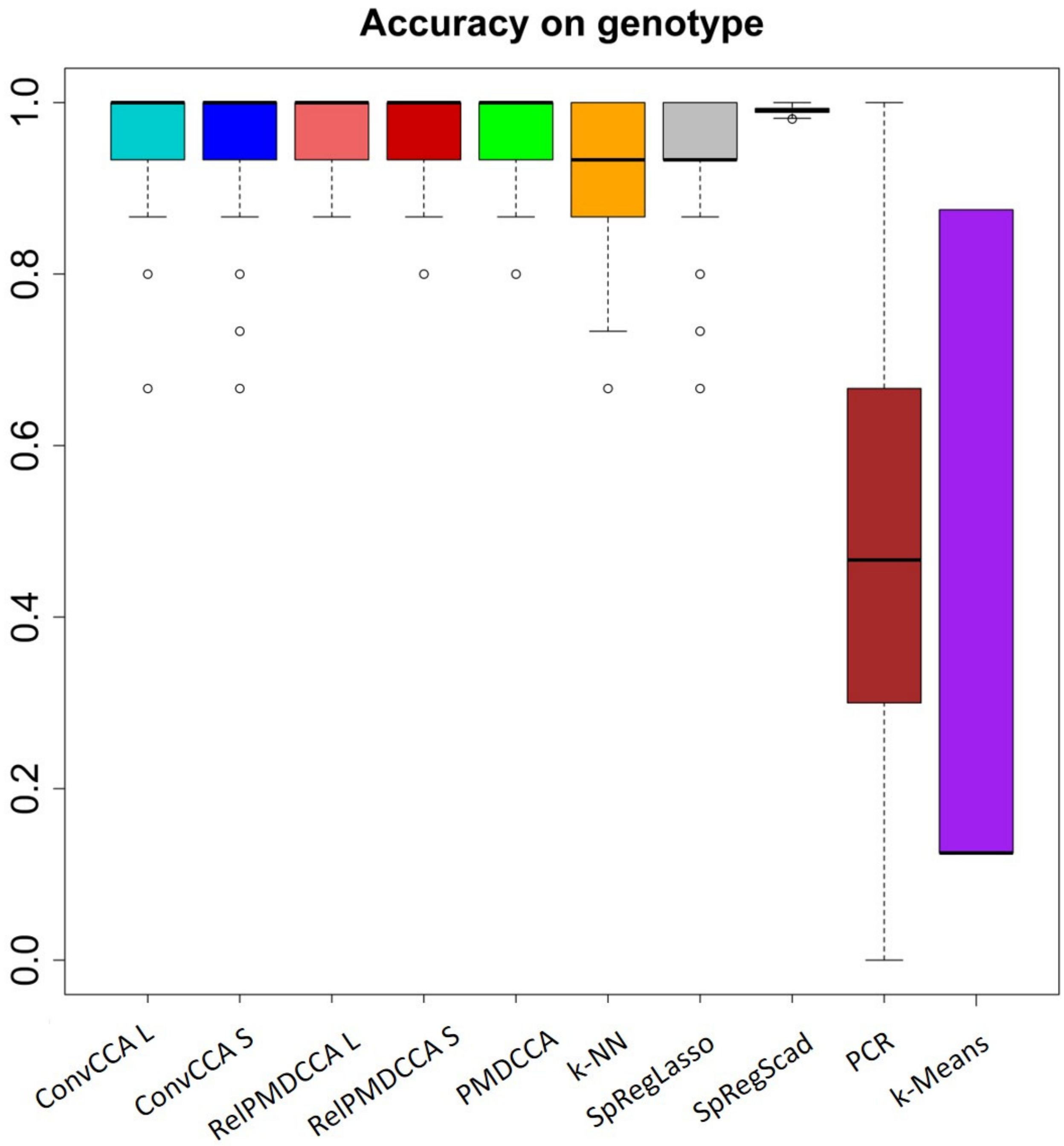

Figure 7: Accuracy of all learning models in predicting the genotype of mice in the nutriMouse study.

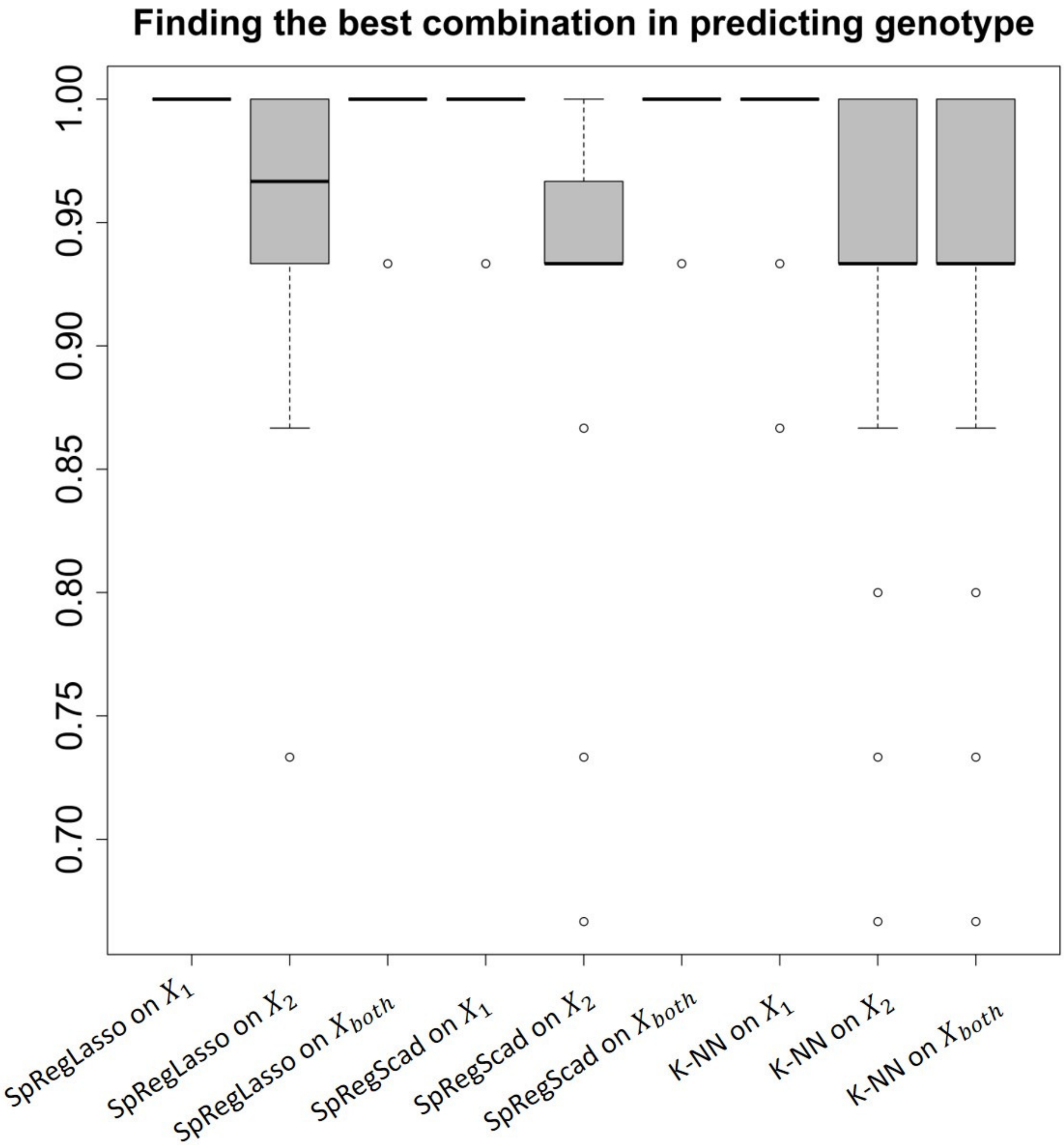

Figure 8: Accuracy measures on SpRegLasso, SpRegScad and k-NN with X_1, X_2 and X_{both} acting as predictors, with the response being the mice genotype from the nutriMouse study.

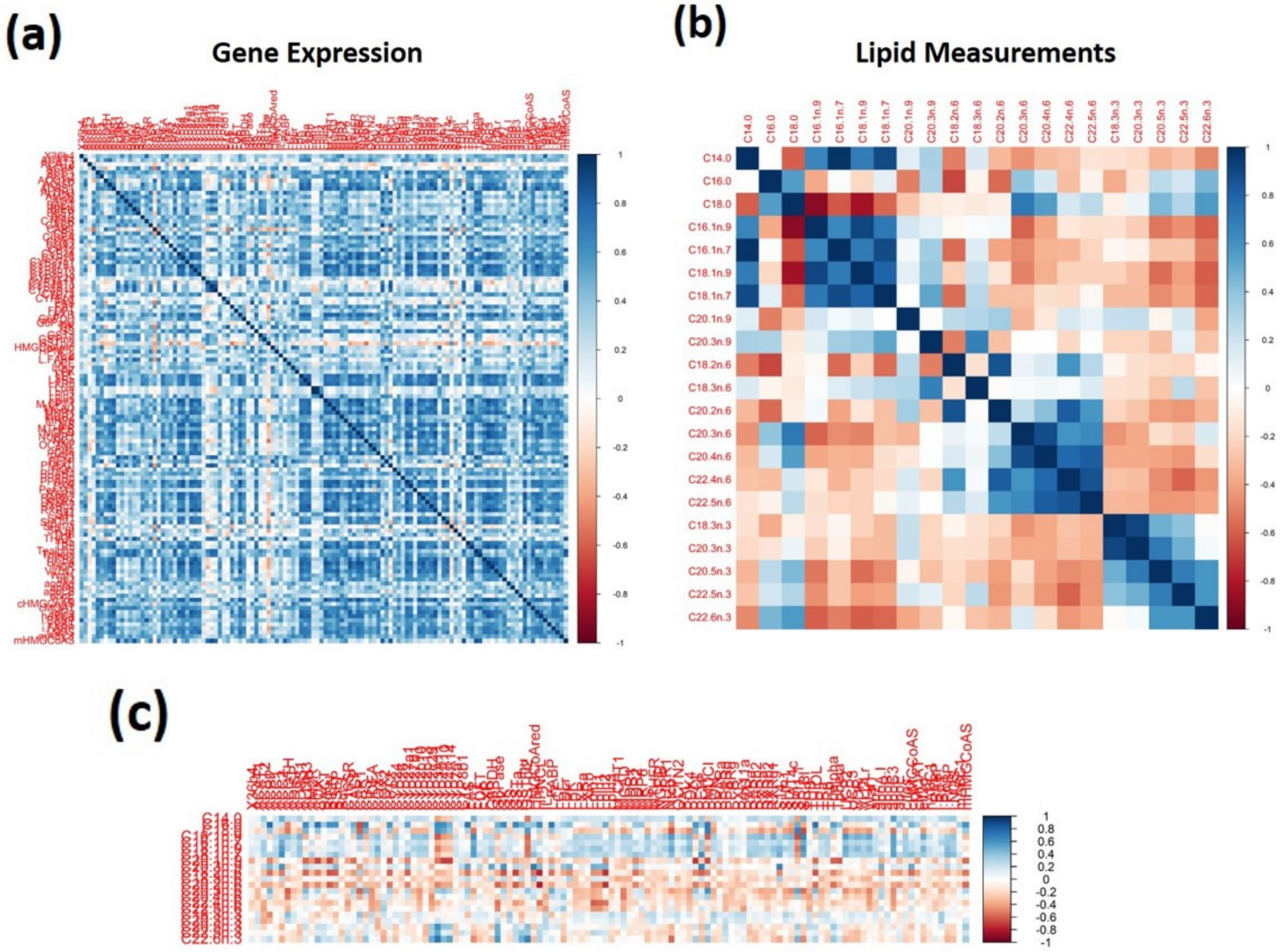

Figure 9: Correlation plots of **(a)** gene expression, and **(b)** lipid measurements in nutriMouse study. **(c)** Cross-correlation of the two data-sets.

References

Bonner A. J. and Beyene J. (2017) Evaluating the performance of sparse principal component analysis methods in high-dimensional data scenarios *Communications in Statistics - Simulation and Computation*, **46**, 3794–3811.