Molecular BioSystems

Supplementary Material

Oblique rotation of factors: A novel pattern recognition strategy to classify fluorescence excitation-emission matrices of human blood plasma for early diagnosis of colorectal cancer

Mohammad Shahbazy[a], Mahdi Vasighi*[b], Mohsen Kompany-Zareh*[a] and Davide Ballabio[c]

[a] *Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), 45137-66731, Zanjan, Iran.*

[b] *Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences (IASBS), 45137-66731, Zanjan, Iran.*

[c] *Milano Chemometrics and QSAR Research Group, Department of Environmental and Earth Sciences, University of Milano- Bicocca, P.za della Scienza, 1-20126 Milano, Italy.*

**Corresponding authors: Phone: (+98) 24 3315 3123, E-mail: kompanym@iasbs.ac.ir* (M. Kompany-Zareh); *Phone:* (+98) *24 3315 3378, E-mail: vasighi@iasbs.ac.ir* (M. Vasighi).

- **SOM algorithm**

Firstly, a two dimensional array is made up then weights of each unit vector are randomized (initialization between 0 and 1). Subsequently, the samples (input vectors) are projected (mapped) on all of the neurons space. There are several stages in the SOM algorithm:

(1) In order to identify the winner (central) neuron which shows most similarity to input vector $\mathbf{x_i}$ pattern, the similarity of neurons and input sample will be obtained via Euclidian distance calculation between them. Hence, there is a competition between neurons for winning (competitive learning).

$$\left(c \,\middle|\, out_c\right) = \left(c \,\middle|\, min_j\left(\|x_i - w_j\|^2\right)\right) \quad \textbf{(1)}$$

(2) Weights vector of winner neuron and its neighbors are modified (updated) so that their similarities to the input vector are gradually enhanced (based on distance with winner neuron). The weights correction defined as follow:

$$w_{ji}^{new} - w_{ji}^{old} = \Delta w_{ij} = \eta(t)a(d_r,t)(x_i - w_{ji}) \quad \textbf{(2)}$$

where;

Molecular BioSystems

Supplementary Material

$$\eta(t) = \left(\eta^{start} - \eta^{end}\right)\frac{n_{epoch} - t}{n_{epoch} - 1} + \eta^{end} \quad \textbf{(3)}$$

How do the weights correction, depends on a coefficient termed learning rate (η) that is a function of learning epoch (iteration) (t), neighborhood (window) function coefficient (a) that is function of topological distance with the winner neuron (c), the neuron number (j) and the d_r is the number of neurons separating a neuron from the winner ($d_r = d_c - d_j$)).

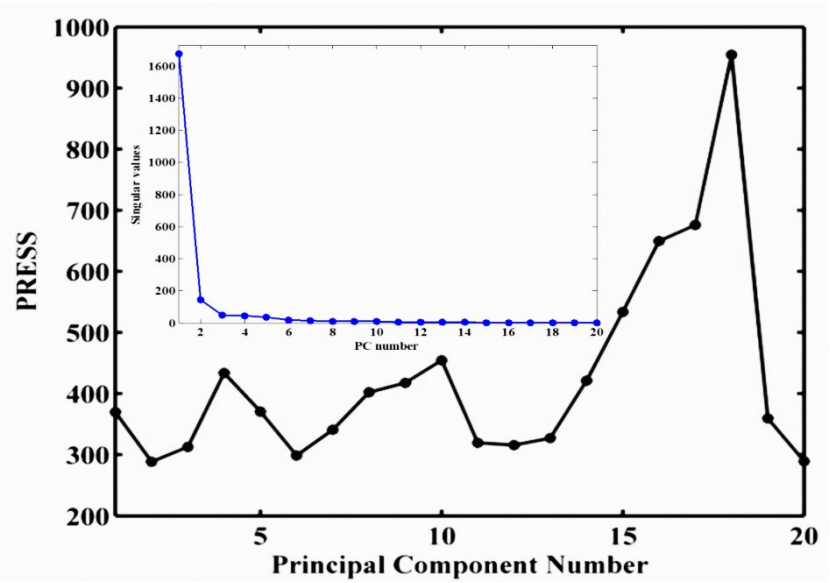
$$a(d_r, t) = (1 - \frac{d_r}{p+1}) \quad \textbf{(4)}$$

$$p = \frac{(n_{epoch} - t)}{(n_{epoch} - 1)} N_{net} \quad \textbf{(5)}$$

The p denotes the size of neighborhood which at the beginning of learning (t=1), p covers the entire network (p=N_{net}). Via correction (updating) of neurons, the weight vectors of neurons in an especial area (locally) on the map will be somewhat similar to the input vector. Thus, if the next input vector presents to the network which is more similar to the first input vector, the winning probability of mentioned area will enhance.

(3) Other input vectors present to the network and the operations to finding winning neuron, training and subsequently correction of network weights repeat as before. When all input vectors once introduce to network, one training cycle (epoch) will be completed.

Finally, a neurons layer with corrected weights will be obtained. If once again, all input vectors present to the network, the winning neuron position in the table as same as dimensions with the network are just registered by a labeling action, then a map will be achieved named “top-map”. Through examination of trained Kohonen network layers, the weight maps can be achieved that play an important role to determine importance of each variable in forming of top-map. Therefore, weight maps can be fruitful to identify the discriminant and most informative variables.[1,2]



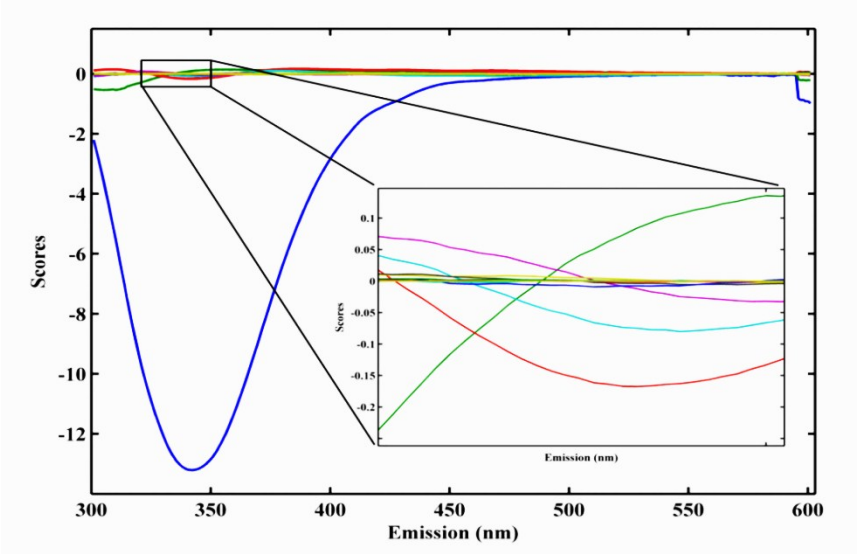


Fig. S1. (a) Eigenvector based cross validation.[3] Calculated *PRESS* values for building PCA models obtained from **D** using different number of PCs. The internal Figure relates to the obtained singular values from SVD that exhibits uncertainly to determine optimal chemical rank. (b) The obtained scores of the first EEM from PCA. The internal Figure displays the overlapping between some components including amino acids and proteins which impresses on accuracy to find out the optimal number of components.

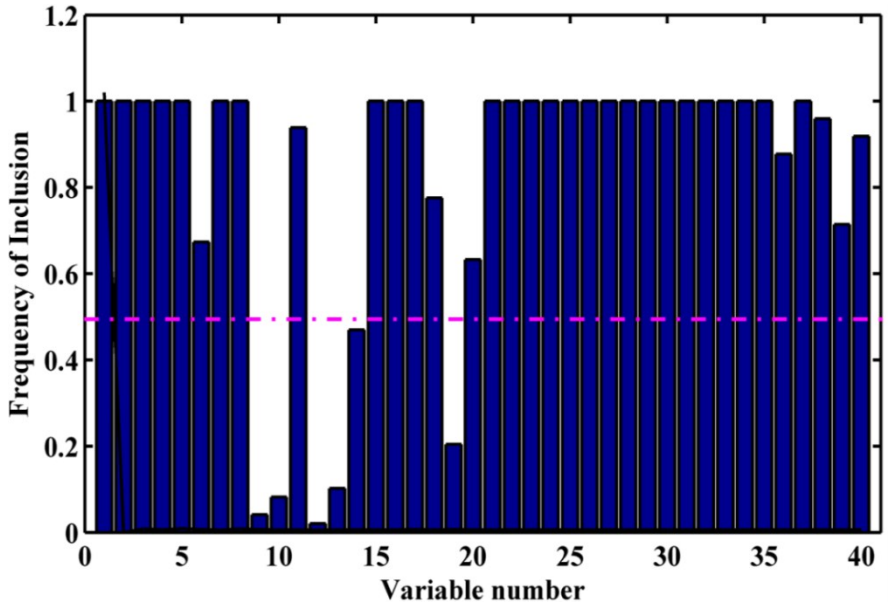


Fig. S2. Factor selection among 40 PLS latent variables by using GA, the frequency of latent variables usage in constructed models (the dashed pink line approximately relates to the assignation threshold).

Table S1. The setting of GA parameters that was used to select the discriminant factors.

Parameter	**Population size**	**Window width**	**Initial terms %**	**Penalty slope**	**Max generation**	**% convergence**	**Mutation rate**
Setting	256	1	30	0	200	80	0.005
Parameter	**Cross over**	**Model type**	**Cross validation**	**Number of splits**	**Number of Iteration**	**Replicate run**	**Fitness function**
Setting	Double	PLS-DA	Contiguous	5	10	5	*NERcv*

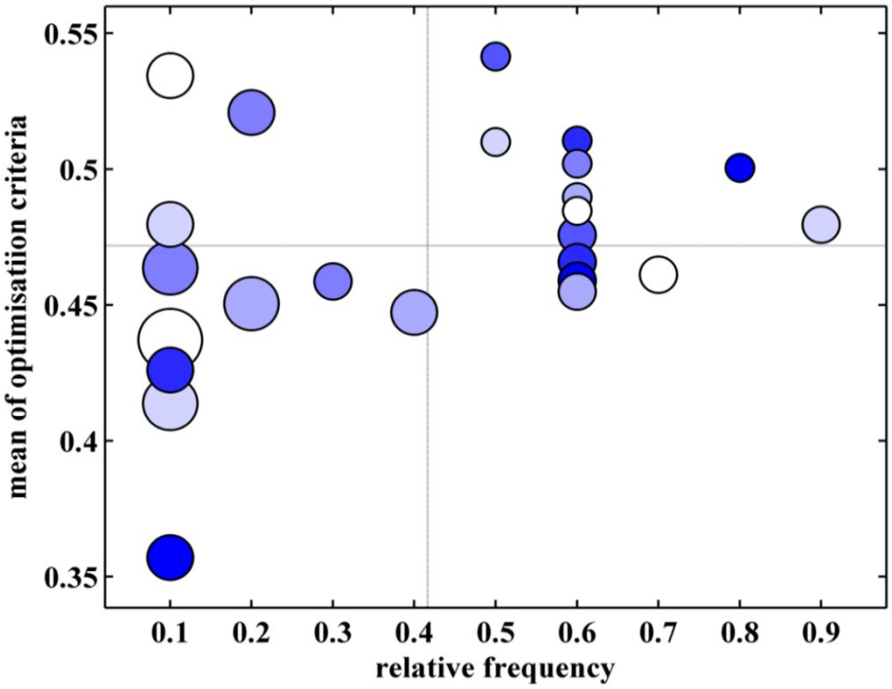


Fig. S3. Plot of relative frequency of selection and mean value of fitness function (*NERcv*). The dimension of each bubble is proportional to the network size; the color of the bubbles is proportional to the number of epochs. The dotted line represents the overall means of frequencies and fitness function.[4]

Table S2. The optimized setting of SKN for building of classification model.

Parameter	**Model type**	**SKN class scaling factor**	**No. of neurons**	**Training epochs**	**Network topology**
Setting	SKN	1	6×6 (4×4)*	150	hexagonal
Parameter	**Training algorithm**	**Boundary condition**	**Initialization of weights**	**Initial learning rate**	**Final learning rate**
Setting	batch	normal	random	0.5	0.01

*About the two-class classification modeling (i.e., CRC and healthy cases).

- **Classification parameters**

ROC curve is the specificity of model (number of predicted samples which are not in the target class divided by actual number that is not in the same class) versus the model sensitivity (number of predicted samples which are in the target class divided by actual number that is in the same class). By means of the ROC curves, the model

ability can be interpreted. If model's sensitivity and specificity are close to one, the obtained model is more predictive and valid.[5-7]

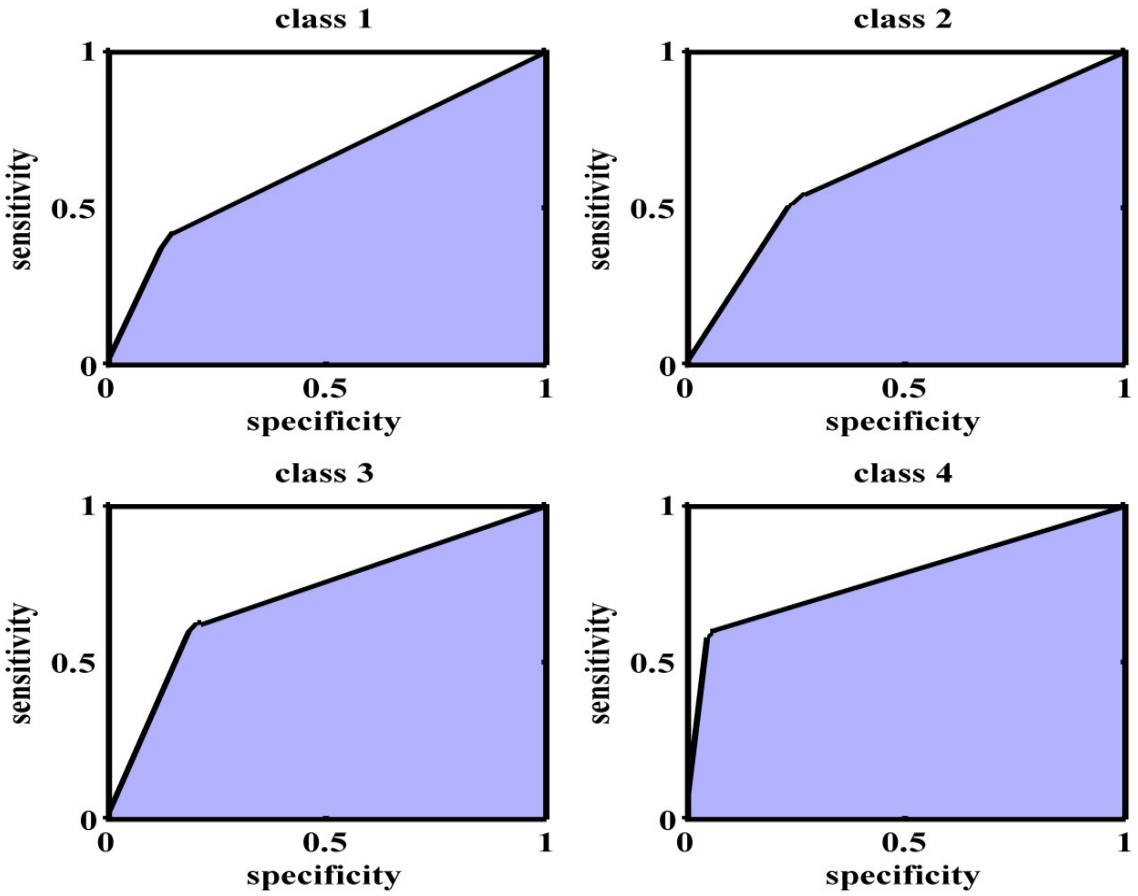


Fig. S4. The ROC curves, model sensitivity versus specificity for various classes in the SKN model. The class numbers as 1, 2, 3 and 4 related to CRC, Adenomas, Onf and Healthy samples, respectively.

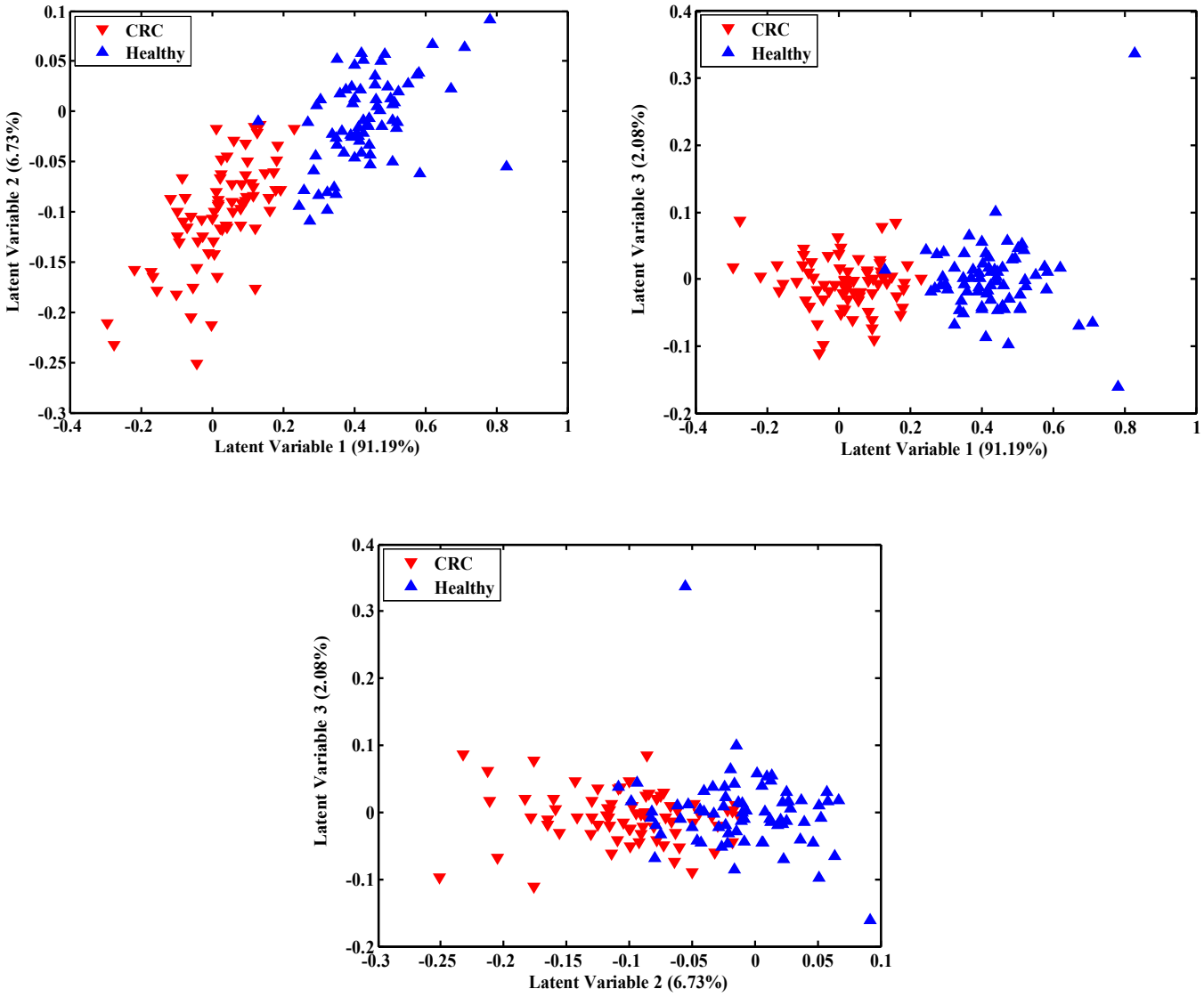


Fig. S5. The scores on the first three PLS-DA latent variables for classification model on CRC and healthy individuals by using the optimal PLS factors as 2D plots for (a) 1st and 2nd LV; (b) 1st and 2nd LV; (c) 2nd and 3rd LV.

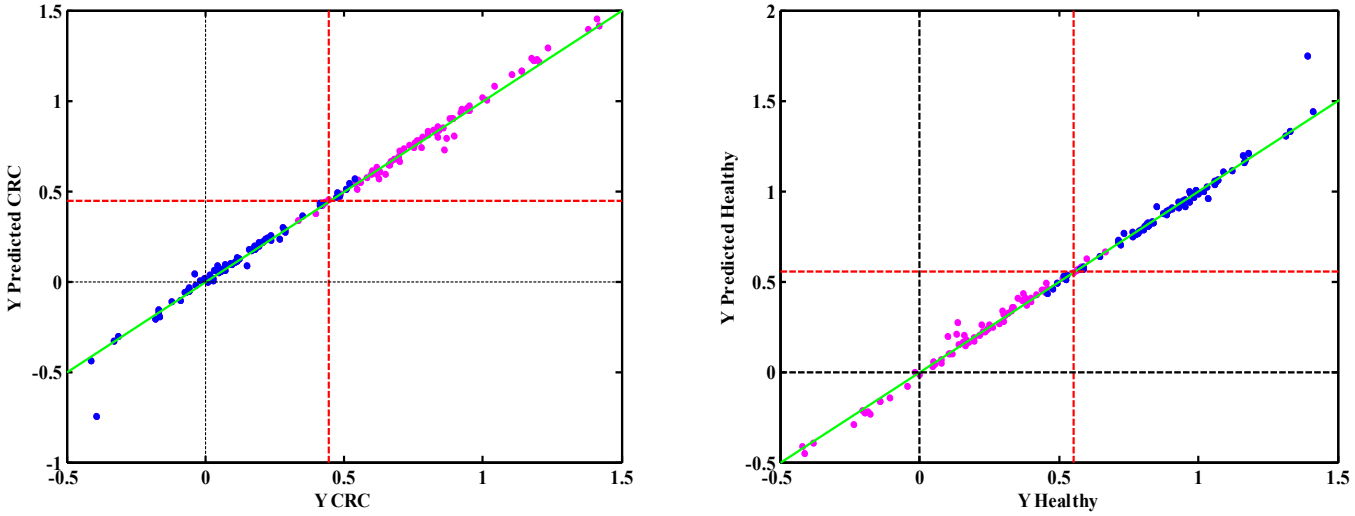


Fig. S6. The predicted Y by using PLS-DA model and the optimal PLS factors versus real Y for the (a) CRC; and (b) healthy classes, respectively. There are two horizontal and vertical red dashed lines as the assignation threshold for the samples to be belonged to one of the CRC and healthy classes.

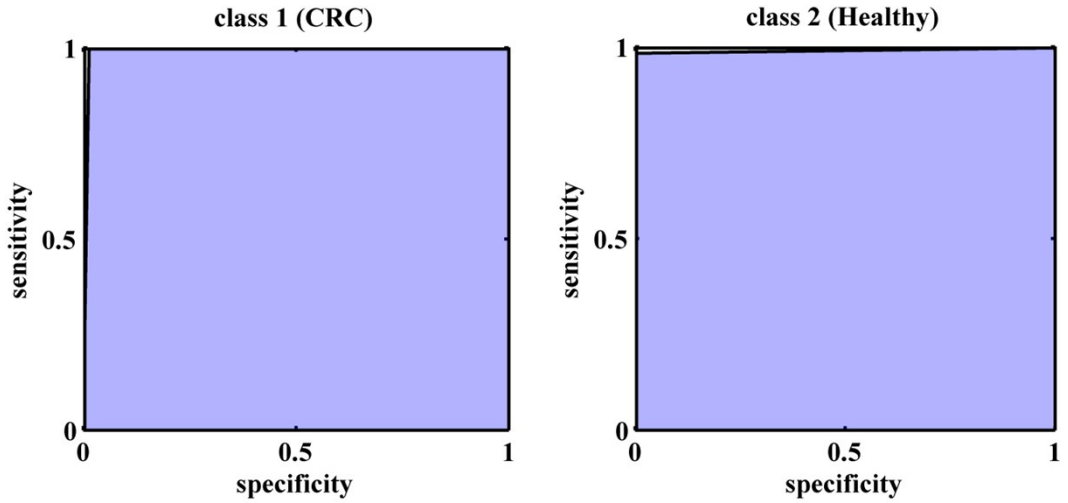


Fig. S7. The ROC curves, model sensitivity versus specificity for the CRC and healthy classes in the SKN model.

References

1. F. Marini, in *Comprehensive Chemometrics*, eds. S. D. Brown, R. Tauler and B. Walczak, Elsevier, Oxford, 2009, pp. 477-505.
2. J. Zupan, M. Novič and I. Ruisánchez, *Chemom. Intell. Lab. Syst.*, 1997, **38**, 1-23.
3. R. Bro, K. Kjeldahl, A. K. Smilde and H. A. L. Kiers, *Anal. Bioanal.Chem.*, 2008, **390**, 1241-1251.
4. D. Ballabio, M. Vasighi, V. Consonni and M. Kompany-Zareh, *Chemom. Intell. Lab. Syst.*, 2011, **105**, 56-64.
5. R. G. Brereton, *Chemometrics for Pattern Recognition*, first edn., John Wiley & Sons, Chichester, 2009.
6. B. K. Lavine and W. S. Rayens, in *Comprehensive Chemometrics*, eds. S. D. Brown, R. Tauler and B. Walczak, Elsevier, Oxford, Editon edn., 2009, pp. 507-515.
7. Wikipedia: the free encyclopedia, *Receiver operating characteristic (ROC) curve*, http://en.wikipedia.org/wiki/Receiver_operating_characteristic.