DISCOVERING ASSOCIATIONS IN HIGH DIMENSIONAL IMAGING-GENETICS DATA: A COMPARISON STUDY OF DIMENSION REDUCTION AND REGULARISATION STRATEGIES COMBINED WITH PARTIAL LEAST SQUARES

Le Floch E.[1,2] *Pinel P.*[3] *Tenenhaus A.*[4] *Trinchera L.*[5] *Poline J.B.* [1,6] *Frouin V.*[1] *Duchesnay E.*[1,2]

[1]CEA, Neurospin, LNAO, Gif-sur-Yvette, France
[2]INSERM-CEA U1000, Neuroimaging & Psychiatry Unit, SHFJ, Orsay, France
[3]INSERM U992, Cognitive Neuroimaging Unit, Neurospin, Gif-sur-Yvette, France
[4]Supélec, Department of Signal Processing and Electronic Systems, Gif-sur-Yvette, France
[5]AgroParisTech, UMR518 MIA, Paris, France
[6]INRIA, Parietal, Saclay, France

ABSTRACT

Brain imaging is increasingly recognised as an intermediate phenotype in the understanding of the complex path between genetics and behavioural or clinical phenotypes. In this context, a first goal is to propose methods to identify the part of genetic variability that explains some neuroimaging variability. Here, we investigate multivariate methods, Partial Least Squares (PLS) regression and Canonical Correlation Analysis (CCA), in order to identify a set of Single Nucleotide Polymorphisms (SNPs) covarying with a set of neuroimaging phenotypes derived from functional Magnetic Resonance Imaging (fMRI). Because in such high-dimensional settings multivariate methods overfit the data, we propose a comparison study of several dimension reduction and regularisation strategies combined with PLS or CCA. We demonstrate that the combination of univariate filtering and sparse PLS outperforms all other strategies and is able to extract a significant link between a set of SNPs and a set of brain regions activated during a reading task.

Index Terms— Imaging Genetics, Multivariate Analysis, Partial Least Squares, Regularisation, Dimension Reduction

1. INTRODUCTION

Imaging genetics studies that require the joint analysis of a large amount of data in both their imaging and genetic components are facing statistical challenges for which there is no standard answer.

Classical approaches rely on massive univariate pairwise linear analysis [1], also called Mass-Univariate Linear Modelling (MULM), ignoring the potential interactions between genes or between brain regions.

To overcome this limitation, we investigate exploratory multivariate methods to identify a functional brain network covarying with a set of genetic polymorphisms. Partial Least Squares (PLS) regression [2] and Canonical Correlation Analysis (CCA) [3] appear to be good candidates in order to look for associations between two blocks of data, as they extract pairs of covarying/correlated latent variables (one from each block). Another approach based on parallel Independent Component Analysis has also been proposed in order to combine functional MRI data and SNPs from candidate regions [4].

Nevertheless, all these multivariate methods encounter critical overfitting issues in exploratory imaging genetics studies, due to the very high dimensionality of the data. To face these issues, dimension reduction or regularisation techniques can be used.

In this paper, we propose a comparison study of various dimension reduction and regularisation strategies combined with PLS regression (in its symmetrical mode relative to the two blocks of data) or CCA to detect associations between genome wide SNP data and fMRI data corresponding to two specific tasks: reading and speech comprehension. The choice of the two tasks was motivated by previous work on dyslexia [5], which is increasingly acknowledged to be a disorder of genetic origin with a basis in the brain [6]. Here we broaden this question and study the impact of common genetic polymorphisms on inter-individual variability in brain activation of language areas for healthy subjects.

To face the dimensionality of the data, we first investigate two types of preliminary dimension reduction: Principal Components Analysis (PCA), which performs feature extraction by looking for a low-dimensional representation of the data that explains most of its variability, and univariate filtering which performs feature selection based on univariate tests.

As for regularisation, L1-regularised PLS, called sparse PLS [7], and L2-regularised CCA have recently been shown to provide good results in correlating two blocks of data such as transcriptomic and metabolomic data, gene expression levels and gene copy numbers, or gene expression levels and SNP data. Here we propose to transpose these techniques to the case of SNP and fMRI data. A very similar method, called sparse Reduced-Rank Regression (sRRR) [8] and based on L1 regularisation, has also been proposed to analyse SNP and brain imaging data, but only on simulated data and with a limited number of SNPs. This method is in fact equivalent to sparse PLS, under the approximation made in high dimensional settings that the covariance matrix of each block is diagonal.

We compare the performances of these different dimension reduction and regularisation strategies combined with PLS or CCA, by assessing the generalisability of the link they find between the two blocks of data, using a cross-validation procedure. Finally, the statistical significance of the link obtained on "test" subjects is assessed with randomisation techniques.

In the next sections we first detail the dataset, then introduce the multivariate methods and the performance evaluation techniques that we used, and finally illustrate the results obtained.

2. DATASET

This study is based on $N = 94$ subjects who were genotyped for 1,054,068 SNPs and participated in a general cognitive assessment fMRI task described in [9]. The study (both imaging and genetics components) was approved by the local ethics committee and all subjects gave their informed consent. The task consisted of a short 5 min BOLD acquisition during which subjects were reading or listening to sentences, asked to perform a motor response (button click), subtract numbers, or were shown visual checkerboard. The functional images were acquired either on a 3T Bruker scanner or a 3T Siemens trio scanner using an EPI sequence (TR=2400 ms, TE=60 ms, matrix size=64×64, FOV=19,2 cm$\times$19,2 cm). T1 anatomical images were acquired during the same acquisition session with a resolution of $(1.1 \times 1.1 \times 1.2)$ mm^3. Standard pre-processings were applied to the fMRI data, which consisted of slice-timing correction, motion estimation, spatial normalisation (with a resampling of the functional images at 3mm resolution) and smoothing (FWHM=10 mm). The preprocessings and first level model analyses were performed with SPM5 (`www.fil.ucl.ac.uk/spm`).

In our study, we focused only on two activation contrasts: *reading minus checkerboard viewing* and *speech comprehension minus rest*. We used a first level, subject-specific, General Linear Model (GLM), to obtain parametric estimates of the BOLD activity at each voxel in each subject; the analysis was performed using SPM5, with standard parameters (frequency cut=128s, AR(1) temporal noise model). For each subject s in $\{1, \ldots, n\}$ and each voxel v of the normalised volume, we obtained a map $\hat{\beta}_s(v)$ that represents the amount of BOLD signal associated with the activation contrast, normalised by the average signal. We defined a global brain mask for the group by considering all the voxels that belong to at least half of the individual brain masks (the individual masks were estimated using the standard SPM5 procedure). Then we selected thirty four brain locations of interest (19 from the "reading" contrast and 15 from the "speech comprehension" contrast): most of them were peaks of local maxima on the activation maps, while the others had been reported to be atypically activated during reading in dyslexia [5]. Each contrast map was locally averaged within 4 voxel-radius spheres centred on these peaks, keeping only active clusters of voxels (T$\geq$1 and cluster size$\geq$10 voxels) [10]. This yielded 34 average values corresponding to 34 regions of interest (ROI) and we computed the average values for the 34 mirror ROIs by symmetry with respect to the inter-hemispheric plane. Finally, lateralisation indexes were derived from those regions since lateralisation is thought to play a role in dyslexia, with a reduced activity in regions of the left hemisphere in dyslexics [5]. For each pair of ROIs in the normalised volume and in each subject, an index was computed as follows:

$$\text{Index}_s = \frac{\bar{\hat{\beta}}_s^{\text{right}} - \bar{\hat{\beta}}_s^{\text{left}}}{\sqrt{\left(\bar{\hat{\beta}}_s^{\text{right}}\right)^2 + \left(\bar{\hat{\beta}}_s^{\text{left}}\right)^2}}. \quad (1)$$

The distribution of these indexes spanned the range of $[-1.5; 1.5]$, and variances were homogeneous across regions of interest. The term "phenotypes" will now refer to the lateralisation indexes thus obtained in the different regions.

For each subject, an Illumina platform was used to genotype 1,054,068 SNPs and processed with the standard platform software. Considering all genotyped data available, we successively applied the following filters on all SNPs: (1) Minor Allele Frequency at least 10%, (2) call rate at least 95%, and (3) Hardy-Weinberg test not significant at the 0.005 level. Genetic data were recoded as the number of minor alleles (denoted as A), $\{0, 1, 2\}$, hence a value of 0 means homozygous wild-type individuals (BB). The frequency of homozygous individuals for the minor allele (AA) was 0.03–0.13 in 75% of the cases. Missing SNP data were imputed with their corresponding median value across subjects. These analyses were carried out using the open-source R software and the storage facilities for genetic data provided in the package `snpMatrix` [11].

After these preprocessing steps, our analyses were performed on two blocks of data $\mathbf{Y}$ (fMRI) and $\mathbf{X}$ (genetics) of size 94×34 and $94 \times 622,534$ respectively.

3. METHODS

3.1. Partial Least Squares methods and Canonical Correlation Analysis

Partial Least Squares regression or Canonical Correlation Analysis are used to model the associations between two blocks of variables ($\mathbf{X}$ and $\mathbf{Y}$) hypothesising that they are linked through unobserved latent variables. The latent variables (or components) are linear combinations of the observed variables of each block, obtained by finding two weight vectors ($\mathbf{u}$ and $\mathbf{v}$).

PLS regression builds successive and orthogonal latent variables for each block such that at each step h the covariance between the pair of latent variables is maximal:

$$\max_{\|\mathbf{u}_h\|_2=\|\mathbf{v}_h\|_2=1} \mathbf{u}_h' \mathbf{X}_{h-1}' \mathbf{Y}_{h-1} \mathbf{v}_h \quad (2)$$

where $\mathbf{X}_0$=$\mathbf{X}$ and $\mathbf{Y}_0$=$\mathbf{Y}$ (whose columns have been standardised). This optimisation problem is solved using the iterative algorithm called NIPALS. Then, using the symmetrical deflation mode of PLS regression between the two blocks (called canonical mode), the effects of the two latent variables obtained at step h are subtracted from the blocks $\mathbf{X}_{h-1}$ and $\mathbf{Y}_{h-1}$ respectively, leading to new blocks $\mathbf{X}_h$ and $\mathbf{Y}_h$ for the next iteration.

CCA differs from PLS in that it maximises the correlation (instead of the covariance) between latent variables at each step h:

$$\max_{\|\mathbf{u}_h\|_2=\|\mathbf{v}_h\|_2=1} \frac{\mathbf{u}_h' \mathbf{X}' \mathbf{Y} \mathbf{v}_h}{\sqrt{\mathbf{u}_h' \mathbf{X}' \mathbf{X} \mathbf{u}_h} \sqrt{\mathbf{v}_h' \mathbf{Y}' \mathbf{Y} \mathbf{v}_h}} \quad (3)$$

The solution may be obtained by computing the SVD of $\mathbf{X}'\mathbf{X}^{-1/2}\ \mathbf{X}'\mathbf{Y}\ \mathbf{Y}'\mathbf{Y}^{-1/2}$. For numerical issues, we use the dual formulation of CCA based on a linear kernel: Kernel CCA (KCCA).

3.2. Regularisation techniques

L2 regularisation of CCA: In order to first reduce the overfitting issue when dealing with CCA, a regularised version of KCCA (rKCCA) based on L2 penalisation may be used, by replacing the matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$ by $\mathbf{X}'\mathbf{X} + \lambda_2 I$ and $\mathbf{Y}'\mathbf{Y} + \lambda_2 I$ respectively. However, in high-dimensional settings the approximation is often made that the scatter matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$ may be replaced by identity matrices, which may be viewed as an extreme case of L2 regularisation. In fact, this approximation makes CCA equivalent to PLS, which implies that PLS may be viewed as an extreme case of L2-regularised CCA and is thus less prone to overfitting than (regularised) CCA.

L1 regularisation of PLS: Pushing further regularisation to reduce even more the overfitting issue, authors [7, 8] have proposed an approach that includes variable selection in PLS regression, based

on L1 penalisation [12] on the SNP weight vector $\mathbf{u}$ and leading to a sparse PLS (sPLS):

$$\min_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1} -\mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{v} + \lambda_{1X} \parallel \mathbf{u} \parallel_1 \quad (4)$$

where λ_{1X} is the L1-penalisation parameter for the weight vector of block $\mathbf{X}$. The sPLS criterion is bi-convex in $\mathbf{u}$ and $\mathbf{v}$ and may still be solved iteratively for $\mathbf{u}$ fixed or $\mathbf{v}$ fixed, using soft-thresholding of $\mathbf{u}$ within the NIPALS loop. Indeed, for $\mathbf{v}$ fixed,

$$\widehat{\mathbf{u}} = g_{\lambda_{1X}}(\mathbf{X}'\mathbf{Y}\mathbf{v}) \quad (5)$$

where $g_\lambda(y) = \text{sign}(y)(|y| - \lambda)_+$ is the soft-thresholding function. For the sake of simplicity, in the rest of the paper, we replaced the penalisation term by the sPLS selection rate: $s_{\lambda_{1X}}$, as the number of selected SNPs (with non null weights) out of the total number of variables of that block. One may note that sparse methods based on L1 regularisation may also be seen as feature selection methods.

3.3. Dimension reduction methods

Principal Component-based dimension reduction: We first used Principal Component Analysis (PCA) to perform dimension reduction on each block of data before applying PLS (PCPLS) or KCCA (PCKCCA). For each block of data, we kept as many components as necessary to explain 99% of the variance of that block. Regularisation is not necessary anymore in that case, as the number of dimensions has been dramatically reduced and has become lower than the number of subjects.

Univariate SNP filtering: We investigated massive univariate filtering as another way to perform dimension reduction before applying sparse PLS. This step consisted of 1- $p \times q$ linear regressions based on an additive genetic model, 2- ranking the SNPs according to the minimal p-value each SNP gets across all phenotypes, and 3- keeping the set of SNPs with the lowest "minimal" p-values. The two-step method (filter and sparse PLS) is denoted fsPLS. It should be noted that the PLS regression criterion maximises the inter-block correlation term *and* the two intra-block variance terms and that the last two terms weigh much more than the first one in high-dimensional settings. Univariate filtering may help to solve this problem by reducing the number of SNPs and selecting the ones that are more correlated to the imaging phenotypes. Similarly, we applied L2-regularised KCCA after a first step of univariate filtering. We call this method frKCCA.

4. RESULTS

Comparative performance assessment: We evaluated whether the link obtained between the imaging and SNP datasets could generalise to new unseen test subjects, using a 10-fold cross-validation scheme. We used the correlation coefficient between latent variables as a criterion to measure this link because of our symmetrical approach between the two blocks. Training samples were used to learn the two first pairs of weight vectors ($\mathbf{u}_1$, $\mathbf{v}_1$) and ($\mathbf{u}_2$, $\mathbf{v}_2$) that were then applied to the test samples to form the two first pairs of test latent variables. Finally, we computed the test correlation of each pair of variables. This correlation was averaged over folds.

Table 1 summarises the two first average correlation coefficients obtained on "test" samples for the different methods tested, as well as the two first average correlation coefficients obtained on "training" samples. The "optimal" parameters for regularisation and filtering used here are those that gave the best average cross-validated correlation coefficients, among all parameters tested.

	ρ^1_{test}	ρ^2_{test}	$\rho^1_{training}$	$\rho^2_{training}$
MULM	0.036	-0.104	-0.458	-0.451
PLS	-0.092	0.218	0.990	0.984
sPLS	0.008	0.201	0.938	0.922
PCKCCA	0.010	0.008	1.000	1.000
PCPLS	-0.088	0.217	0.990	0.984
frKCCA	0.245	0.324	0.963	0.954
fPLS	0.236	0.268	0.962	0.953
fsPLS	**0.432***	0.210	0.772	0.788

Table 1. The two first average correlation coefficients found on left-out "test" samples (ρ^1_{test} and ρ^2_{test}) and on "training" samples ($\rho^1_{training}$ and $\rho^2_{training}$). * indicates that the correlation coefficient is significant ($p \leq 0.05$) after permutations and correction for multiple comparisons.

Table 1 shows that, for the first pair of components, L1 regularisation of PLS cannot solve the overfitting issue by itself. Indeed, like PLS, sparse PLS ($s_{\lambda_{1X}}$=0.1%) completely fails to extract a generalisable link in such high dimensions and capture only noise. In such high dimensions, KCCA requires such an extreme L2 regularisation that it is equivalent to PLS.

Therefore a first step of dimension reduction appears to be necessary in order to overcome the overfitting issue. Indeed, even though PCA-based methods do not succeed either, filtering-based methods perform much better. Among these filtering-based methods, fsPLS yields the highest out-of-sample correlation coefficient of 0.43 (when 1000 SNPs are kept after the univariate filter and 5% of those SNPs are kept by sparse PLS). This out-of-sample correlation coefficient happens to be significant (p=3.4%) after 1000 permutations of the row indexes of $\mathbf{Y}$ and correction for the different values of regularisation and filtering parameters tested using a maxT procedure [13]. The second best performance on the first pair of components is obtained with frKCCA with an out-of-sample correlation coefficient of 0.24 (with 1000 SNPs kept after filtering and λ_2=1,000,000). However, with such an extreme L2 regularisation, it is almost equivalent to fPLS (with 1000 SNPs kept after filtering), as can be seen on Table 1.

As for the second component pair, the out-of-sample correlation coefficient obtained by fsPLS is lower than on the first component pair. However for all the other PLS-based methods, the correlation appears to be slightly higher on the second component pair than on the first one. This may be explained by the fact that once the noise leading to overfitting on the first component pair has been removed, some real effects may be observed on further components, while on the opposite, fsPLS prevents from overfitting and can capture some effects on both pairs of components. Finally, MULM and PCKCCA do not seem able to capture any generalisable effects on any of the component pairs.

Imaging genetics findings: In order to obtain the SNPs and the brain phenotypes involved in the link between the two blocks, we then applied fsPLS on all the subjects simultaneously for the pair of parameters giving the best results: 1000 SNPs selected by the univariate filter and a sPLS selection rate $s_{\lambda_{1X}} = 0.05$.

Among the 50 SNPs selected by fsPLS, some of them were located within a gene. Eighteen genes were thus identified, such as PPP2R2B and RBFOX1, which have been reported to be linked with ataxia and a poor coordination of speech and body movements, or also PDS4B which has been associated with schizophrenia.

Figure 1 shows the location of the phenotypes where lateralisation indexes were computed, for both contrast maps of interest

"reading" and "speech comprehension". The weights assigned by fsPLS to the imaging phenotypes on the first component are illustrated according to the colourbar. The phenotypes that obtained the largest weights (in absolute value) mainly come from the "reading" contrast, especially from the temporal lobe, which may be due to the fact that there exist stronger correlations between "reading" imaging phenotypes than between "speech comprehension" phenotypes.

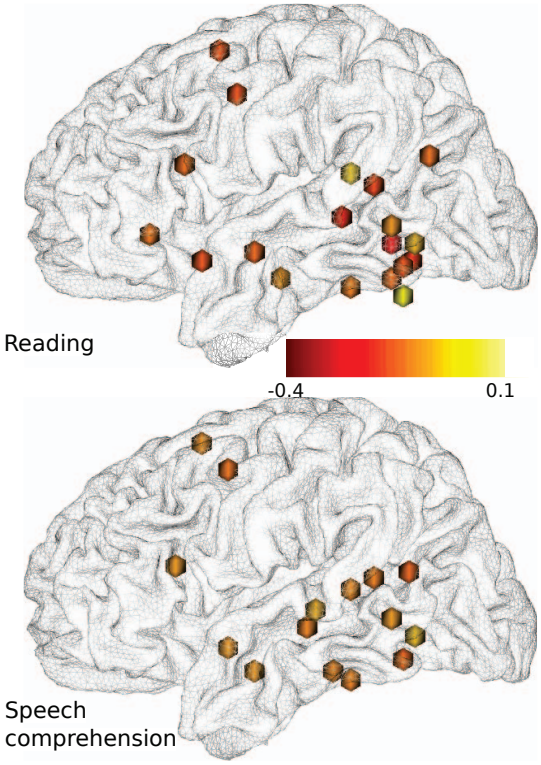


Fig. 1. Weights assigned by fsPLS to imaging phenotypes, with 1000 SNPs kept by filter and $s_{\lambda_{1X}}$=5%.

5. CONCLUSION

Taken altogether, our results show that fsPLS was sensitive enough to establish a significant link between a subset of SNPs distributed across the genome and a functional brain network activated during a reading task, some of these SNPs being probably indirectly linked to the neuroimaging phenotypes due to linkage disequilibrium. This suggests that individual variability in the entire genome contains predictors of the observed variability in brain activation during language tasks. However the interpretation of the results is still a very difficult issue and the neuroscientific relevance of these findings should be investigated in further research.

From a methodological point of view, these results tend to confirm our previous preliminary study on simulated SNP and fMRI data, where fsPLS showed the best performance as well [14]. In conclusion, even though univariate filtering may seem to contradict the very nature of multivariate methods such as PLS, it still allows sparse PLS to extract multivariate patterns among the remaining variables and may even be necessary to overcome the overfitting issue in very high dimensional settings.

6. REFERENCES

[1] J. Stein, X. Hua, S. Lee, A. Ho, A. Leow, A. Toga, A. Saykin, L. Shen, T. Foroud, N. Pankratz, M. Huentelman, D. Craig, J. Gerber, A. Allen, J. Corneveaux, B. DeChairo, S. Potkin, M. Weiner, and P. Thompson, "Voxelwise genome-wide association study (vGWAS)," *NeuroImage*, vol. 53, pp. 1160–1174, 2010.

[2] S. Wold, H. Martens, and H. Wold, "The multivariate calibration problem in chemistry solved by the PLS method," in *Proceedings Conference Matrix Pencils*, A. Ruhe and B. Kastrøm, Eds. 1983, vol. Lecture Notes in Mathematics, pp. 286–293, Springer-Verlag.

[3] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[4] V.D. Calhoun, J. Liu, and T. Adali, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *NeuroImage*, vol. 45, no. 1(Suppl 1), pp. S163–S172, 2009.

[5] E. Paulesu, J.-F. Demonet, F. Fazio, E. McCrory, V. Chanoine, N. Brunswick, F. Cappa, G. Cossu, M. Habib, C.D. Frith, and U. Frith, "Dyslexia: Cultural diversity and biological unity," *Science*, vol. 291, pp. 2165–2167, 2001.

[6] S.D. Smith, P.M. Kelley, and A.M. Broer, "Molecular approaches to the genetic analysis of specific reading disability," *Human Biology*, vol. 70, no. 2, pp. 239–256, 1998.

[7] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A sparse PLS for variable selection when integrating omics data," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, 2008.

[8] M. Vounou, T.E. Nichols, and G. Montana, "Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank approach," *NeuroImage*, vol. 53, pp. 1147–1159, 2010.

[9] P. Pinel, B. Thirion, S. Meriaux, A. Jobert, J. Serres, D. Le Bihan, J.-B. Poline, and S. Dehaene, "Fast reproducible identification and large-scale databasing of individual functional cognitive networks," *BMC Neuroscience*, vol. 8, no. 91, 2007.

[10] P. Pinel and S. Dehaene, "Beyond hemispheric dominance: Brain regions underlying the joint lateralization of language and arithmetic to the left hemisphere," *Journal of Cognitive Neuroscience.* , 2009, Posted Online January 13, 2009. (doi:10.1162/jocn.2009.21184).

[11] D. Clayton and H.-T. Cheung, "An R package for analysis of whole-genome association studies," *Human Heredity*, vol. 64, pp. 45–51, 2007.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, B*, vol. 58, no. 1, pp. 267–288, 1996.

[13] P.H. Westfall and S.S. Young, Eds., *Resampling-based Multiple Testing*, Wiley, New York, 1993.

[14] E. Le Floch, P. Pinel, A. Tenenhaus, L. Trinchera, J.B. Poline, V. Frouin, and E. Duchesnay, "Dimension reduction and regularisation combined with partial least squares in high dimensional imaging-genetics studies," in *Proceedings International Conference on Partial Least Squares and Related Methods*, 2012.