Assessing the performance of sparse PLS regression and CCA with high-dimensional two-block data structure

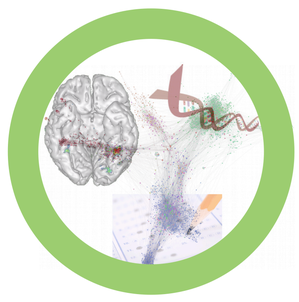
See also Posters # 1218 MT PM, 1090 MT PM 1443 WTh AM, 962 WTh PM

Christophe Lalanne[1,2], Édouard Duchesnay*[1,3,4], Vincent Frouin[1], Benjamin Thyreau[1], Arthur Tenenhaus[5], Bertrand Thirion[1,6], Jean-Baptiste Poline*[1,3,4,6]

(1) CEA, I2BM, Neurospin, Gif-sur-Yvette; (2) INSERM U669, Univ Paris Sud and Univ Paris Descartes, UMR-SO669, Paris; (3) IFR49, Institut d'Imagerie Neurofonctionnelle, Paris; (4) INSERM-CEA U1000, Neuroimaging & Psychiatry Unit, SHFJ, Orsay; (5) Supélec, Department of Signal Processing and Electronic Systems, Gif-sur-Yvette; (6) INRIA Saclay-Ile-de-France, Parietal project, France
* edouard.duchesnay@cea.fr; jbpoline@cea.fr

Background

The Imaging genetics paradigm. Imaging genetic studies aim at mining robust associations between DNA sequence polymorphisms or gene expression level, and activity recorded in the brain, e.g. using fMRI techniques, during cognitive tasks. From a statistical perspective, this raises interesting challenges since a large amount of partly collinear predictors generally entails poor model performance (noisy parameters estimates, overfitting and lack of generalizability). Multivariate methods, like PLS regression or CCA, have been proposed to cope with such high-dimensional data sets ($N \ll P + Q$), with appropriate regularization scheme to overcome the curse of dimensionality[1,3,4]. With PLS, we seek latent (unobserved) variables that account for the maximum of linear information contained in the X block while allowing us to predict the Y block with minimal error, which would likely remain unseen by single-marker analysis



Aims. (1) Validate the use of **sparse PLS regression with univariate feature selection** for extracting covarying networks of variables in two-block structure, and (2) Apply this computational framework to **candidate SNPs, BOLD signals, and measurement on personality scales**.

Sparse PLS regression benchmarking

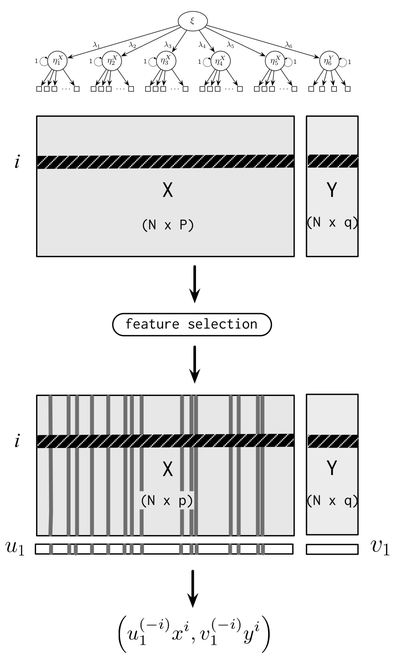
Simulation setup

We generate a two-block data set with a hierarchical model, where factor loadings F ($P \times k$) stand for intra-block correlations with k blocks of varying size, and G ($k \times 1$) reflects inter-block correlations. Individuals scores then follow an MVN$(0_k; \Sigma)$, where $\Sigma = F(GG')F'$, with N subjects. We then applied sparse PLS and CCA models proposed by Lê Cao et al.[3] and Parkhomenko et al.[4] to extract relevant X features for explaining/summarizing $X - Y$ links.

Test procedure (PLS regression):

1000 permutations

1. Sample train/test individuals (10-fold CV)
2. Select the top p best-ranked X features (F-test)
3. Estimate PLS soft-thresholded loadings, u_1 and v_1, on *train* s.t. $\arg \max_{|u_h|=1,|v_h|=1} \text{cov}(X_{h-1}u_h, Yv_h)$
4. Correlate u_1X and v_1Y (factor scores) on *test*

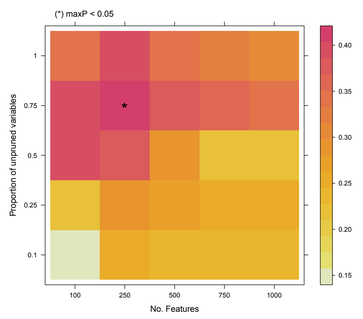


Penalization rate applied on u_1 and v_1 ranged from 0 to 90%. Other parameters were as follows: $N = 90$, $P = 3000$, $k = 4$ blocks in X, of size $b_k \sim \mathcal{U}_{[10;80]}$ for X, and one Y block. F loadings were sampled in the range $[0.2; 0.8]$ and $G = [0.1, 0.0, 0.8, 0.0, 0.7]$ (two blocks in X connected to Y).
For sparse CCA, regularization parameters (λ_i) were choosed so as to be close to % of pruned variables in PLS. In both case, we only compute the first canonical correlation.

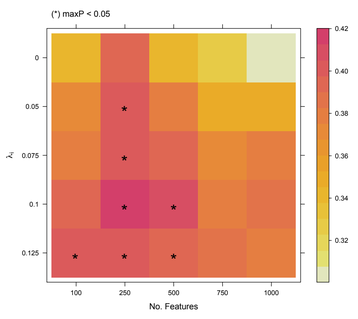
Results

For PLS, the optimal number of X features and Y responses vectors is close to those used to generate the data ($77 + 4$ "true" signals in both blocks), although we might have expected to find the "best" correlation under the condition 100 features without penalization or with 75% of them kept in the model. For CCA, λ_i ranging from 0.05 up to 0.125 yields significant correlation, as well as stronger penalties with 100 and 500 X features.

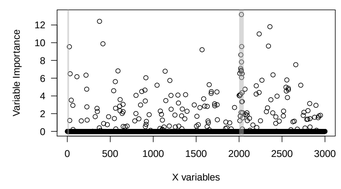
Test correlation for PLS



Test correlation for CCA



Significant test correlations at 5% (max P computed under re-randomization, and correcting for the number of Y variables) are starred in each case. Below is the contribution of each variable in the sparse PLS model ($250 \times 0.75 = 187$ features) applied on all N subjects.



Conclusion

These results suggest that **CCA may be less sensitive than PLS regression**, although values of ℓ_1 penalization do not match exactly soft-thresholding of PLS loadings used therein. Increasing Signal-to-Noise Ratio, or equivalently the reliability of Y measures, emphasizes the role of penalization when seeking for robust correlation (data not shown).

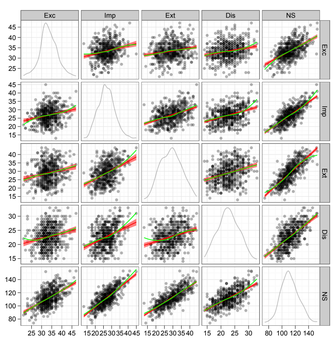
Imaging Genetics of Novelty Seeking trait

Study participants and Data preprocessing

Study sample is composed of $N = 510$ healthy subjects (52% female, 14.4 ± 0.4 y.o.) from the IMAGEN study†, with genetic data acquired on an Illumina QuadChip 660k. We selected 191 SNPs from 17 serotoninergic and dopaminergic genes from HugeNavigator‡ as in Heck et al.[2] Only SNPs with MAF ≥ 0.05, call rate ≥ 0.95, in Hardy-Weinberg equilibrium at $p > 0.0001$, were retained for the analysis (125 SNPs). We considered scores on the TCI Novelty Seeking (NS) scale, and neuroimaging data on a Stop-Signal Task (No-Go responding contrast, see 1090 MT-PM), as phenotypes.

Results

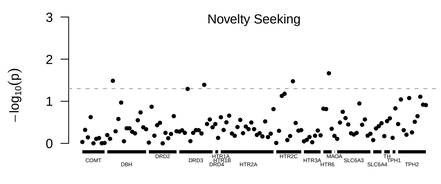
The NS scale is composed of four well-correlated (with $0.194 \leq r_{BP} \leq 0.765$, all $p < 0.001$) subscales (Excitability, Impulsivity, Extraversion, and Disorderness), see below.



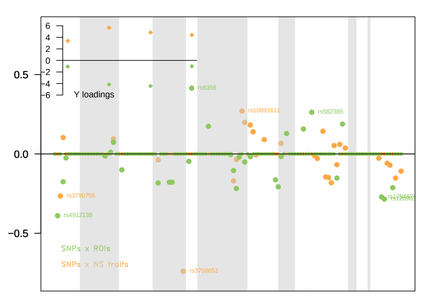
Analysis strategy:

- Apply usual SNP-wise analysis *vs.* PLS on NS traits
- Select the ROIs that best correlate to NS traits
- Regress those ROIs onto the 17 genes

The GWAS view



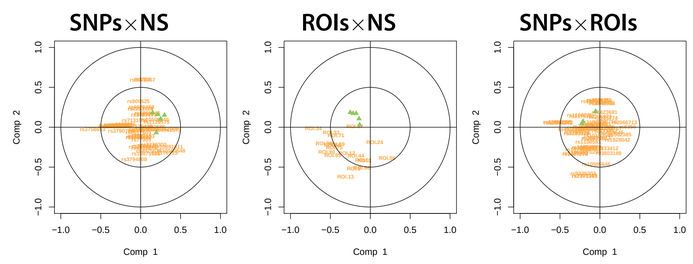
The PLS view



No SNPs survive Bonferroni or BH correction. The "best" genes for NS scale score are: DBH, DRD3, HTR2C, MAOA (uncorrected p-value below 5%).
Sparse PLS regression of SNPs × NS traits yields HTR3A, HTR6, and DRD4 as best associated to a linear combination of the four subscales.
Sparse PLS regression of ROIs × NS traits yields four ROIs as best candidates. Considering the subscale of Impulsivity instead of the summated scale score for NS, gives two additional ROIs (with loadings > 0.25).

ROI coordinates (MNI): 33 (-48,-4,-38), 34 (-60,-43,28), 53 (30,-13,4), 56 (0,-43,-35)
In the case of PLS, NS scale scores were first orthogonalizd w.r.t. Centre and Gender, since GWAS were also adjusted for those covariates.

The four ROIs selected from sparse PLS were then submitted to another sparse PLS regression, together with SNPs ($N = 389$ complete cases). The complete picture for associations unraveled through sparse PLS regression is shown below:



Canonical correlations for the first component were assessed under re-randomization (1000 permutations) for SNPs×NS traits ($N = 443$ c.c.) and SNPs×ROIs ($N = 389$ c.c.) and proved to be significant ($p = 0.001$ and $p = 0.012$, resp.), though in the latest case we may be over-optimistic due to prior selection of ROIs from NS traits.

Conclusion

Sparse PLS regression proves to yield reliable results when combined with univariate feature selection. Compared to ℓ_1 or $\ell_1\ell_2$ regression, it may be more suitable when the Y block included few phenotypes whose linear combinations make sense.

Although we found **no evidence of significant association between DA and NT genes and novelty seeking** using the traditional GWAS approach, the **HTR6 gene** was found to be associated to NS traits but also to the ROIs that best correlate to those traits, when using sparse PLS regression. Further investigations and comparison with whole GWAS analysis need to be carried out in the near future.

This is part of a joint work with Patricia Conrod (Institute of Psychiatry, London) and Hugh Caravan (Trinity College, Dublin).

References

[1] I González, S Déjean, P G P Martin, O Gon¸alves, P Besse, and A Baccini. Highlighting relationships through regularized canonical correlations analysis: Applications to high throughput biology data. *Journal of Biological Systems*, 17(2):173–199, 2009.

[2] A Heck, R Lieb, A Ellgas, H Pfister, S Lucae, D Roeske, B Pütz, B Müller-Myhsok, M Uhr, F Holsboer, and M Ising. Investigation of 17 candidate genes for personality traits confirms effects of the htr2a gene on novelty seeking. *Genes Brain Behav*, 8(4):464–72, 2009.

[3] K-A Lê Cao, P Martin, C Robert-Granié, and Besse P. A sparse pls for variable selection when integrating omics data. *BMC Bioinformatics*, 10(34), 2009.

[4] E Parkhomenko, D Tritchler, and J Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.

Data analysis were done using R 2.11.0, with the snpMatrix and mixOmics packages.
† Information about the IMAGEN project may be found at http://www.imagen-europe.com/.
‡ HugeNavigator is available at http://www.hugenavigator.net/.

Thanks to Édith Le Floch and Laura Trinchera for helpful comments.

IMAGEN receives research funding from the European Community's Sixth Framework Programme (LSHM-CT-2007-037286).