PLS-models in Practice: sparse and sparse group extensions

Lecture 2/3

MSc Health Data Analytics – Computational epidemiology – February 11, 2021

Marc Chadeau-Hyam

m.chadeau@imperial.ac.uk

Imperial College London

Penalisation: Sparse PLS

- Aim: identify a sparse set of variables to be included in each component
- Defining a penalised objective function:

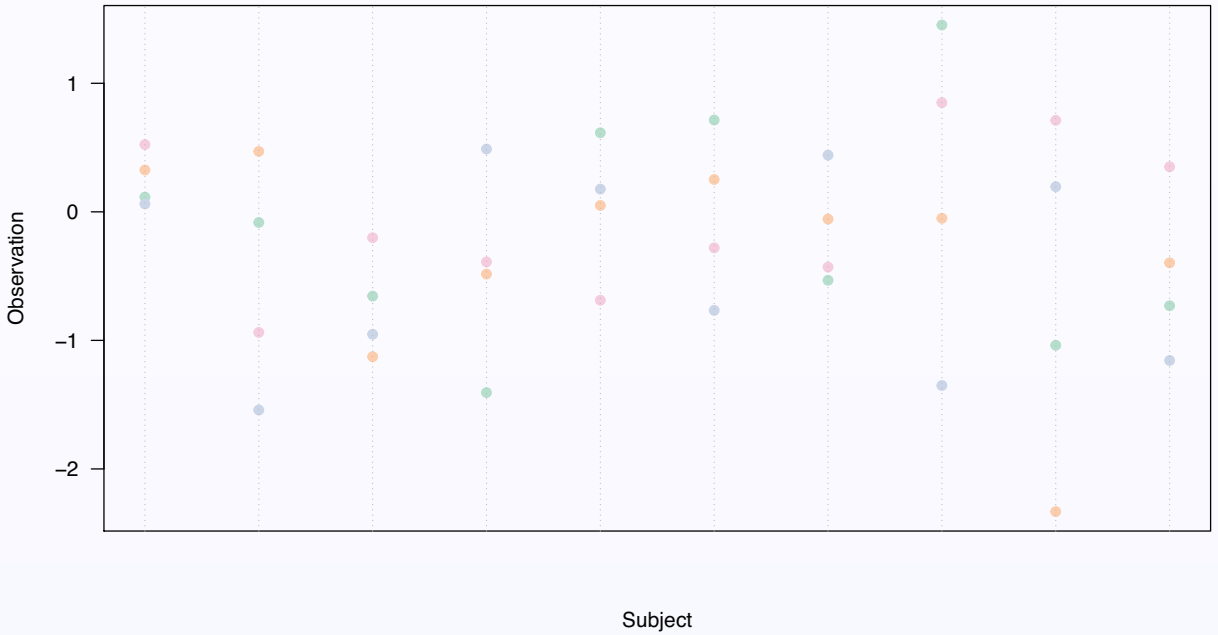
$$\min_{||u_h||=1,\ ||v_h||=1} || \underbrace{X_h^T Y_h}_{\text{observed covariance}} - \underbrace{u_h v_h^T}_{\text{covariance low rank approximation}} ||_F^2 + \underbrace{P_{\lambda_1}(u_h)}_{\text{sparsity on X}} + \underbrace{P_{\lambda_2}(v_h)}_{\text{sparsity on Y}}$$

- Classically the ℓ_1 norm penalty is used: $P_\lambda(u) = \lambda ||u||_{\ell_1} = \lambda \sum_{j=1}^{p} |u_j|$
- Estimation: Soft thresholding approximation to estimate the ℓ_1 norm:
  1. Estimate u and v without penalisation
  2. Update coordinates of u:
     $$\text{sign}(u)(max(0, |u| - \frac{\lambda_1}{2})) \text{ and estimate } v \text{ (step 1)}$$
  3. Update coordinates of v:
     $$\text{sign}(v)(max(0, |v| - \frac{\lambda_2}{2})) \text{ and estimate } u \text{ (step 1)}$$
  4. Repeat 1-3 until convergence
- In practice in the `mixOmics R` package λ is not calibrate, but instead the number of variables to be selected

PLS with structured observations: multi-level extension

- Variation of interest: intra-individual variation (before, after swimming in chlorinated pool)
- Decomposition of the k^{th} observation of variable j for individual i

$$x_{ik}^{j} = \underbrace{x_{..}^{j}}_{\text{offset}} + \underbrace{(x_{i.}^{j} - x_{..}^{j})}_{\text{between subject variation}} + \underbrace{(x_{ik}^{j} - x_{i.}^{j})}_{\text{within subject variation}}$$

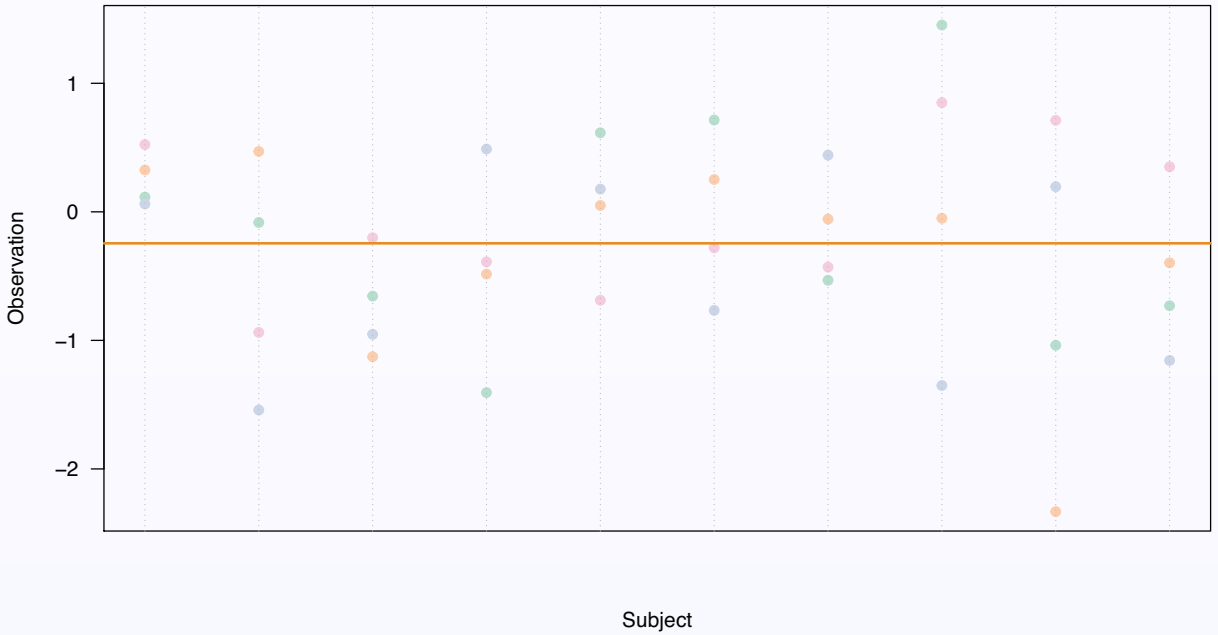


$\Rightarrow$ 10 subjects, 4 observations per subject

PLS with structured observations: multi-level extension

- Variation of interest: intra-individual variation (before, after swimming in chlorinated pool)
- Decomposition of the k^{th} observation of variable j for individual i

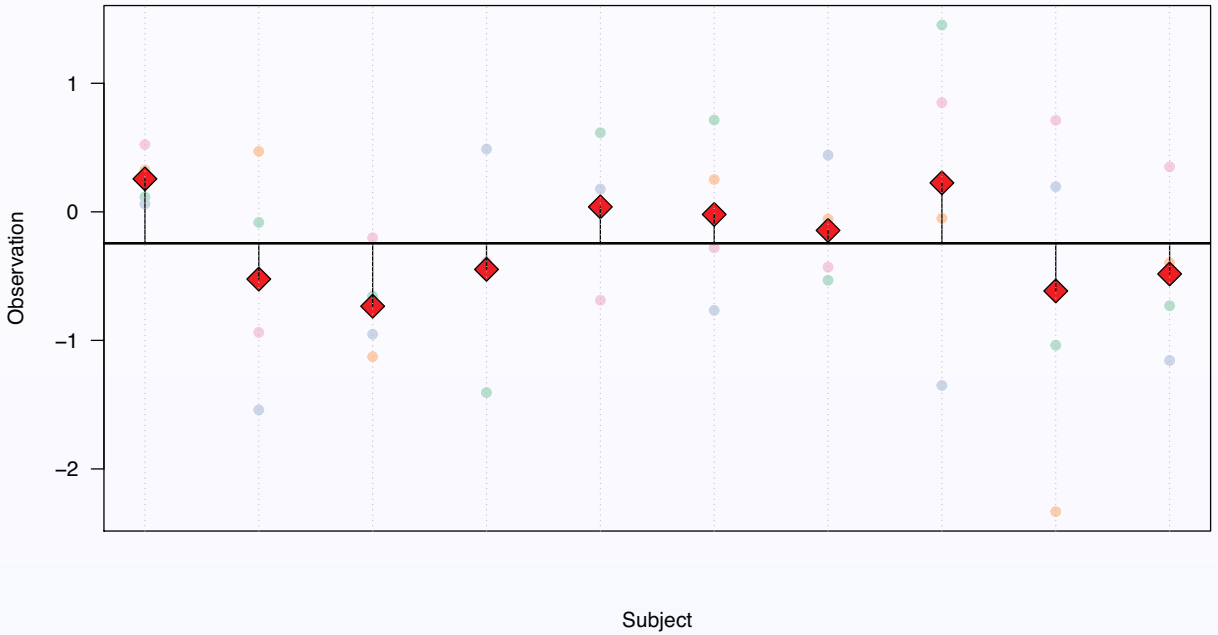
$$x_{ik}^{j} = \underbrace{x_{..}^{j}}_{\text{offset}} + \underbrace{(x_{i.}^{j} - x_{..}^{j})}_{\text{between subject variation}} + \underbrace{(x_{ik}^{j} - x_{i.}^{j})}_{\text{within subject variation}}$$



PLS with structured observations: multi-level extension

- Variation of interest: intra-individual variation (before, after swimming in chlorinated pool)
- Decomposition of the k^{th} observation of variable j for individual i

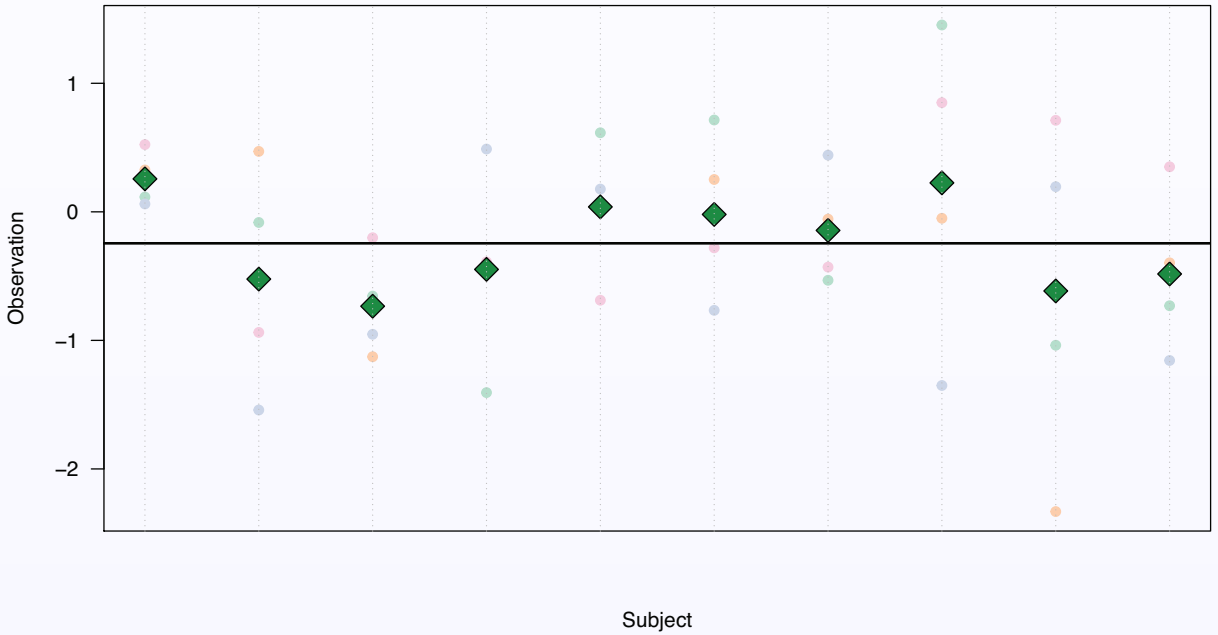
$$x_{ik}^{j} = \underbrace{x_{..}^{j}}_{\text{offset}} + \underbrace{(x_{i.}^{j} - x_{..}^{j})}_{\text{between subject variation}} + \underbrace{(x_{ik}^{j} - x_{i.}^{j})}_{\text{within subject variation}}$$



PLS with structured observations: multi-level extension

- Variation of interest: intra-individual variation (before, after swimming in chlorinated pool)
- Decomposition of the k^{th} observation of variable j for individual i

$$x_{ik}^{j} = \underbrace{x_{..}^{j}}_{\text{offset}} + \underbrace{(x_{i.}^{j} - x_{..}^{j})}_{\text{between subject variation}} + \underbrace{(x_{ik}^{j} - x_{i.}^{j})}_{\text{within subject variation}}$$



Multilevel PLS: formalism

- Variation of interest: intra-individual variation (before, after swimming in chlorinated pool)
- Decomposition of the k^{th} observation of variable j for individual i

$$x_{ik}^{j} = \underbrace{x_{..}^{j}}_{\text{offset}} + \underbrace{(x_{i.}^{j} - x_{..}^{j})}_{\text{between subject variation}} + \underbrace{(x_{ik}^{j} - x_{i.}^{j})}_{\text{within subject variation}}$$

- Can be rewritten as:

$$\underbrace{(x_{ik}^{j} - x_{i.}^{j})}_{\text{within subject variation}} = \underbrace{(x_{.k}^{j} - x_{..}^{j})}_{\text{experimental effect}} + \underbrace{(x_{ik}^{j} - x_{i.}^{j} - x_{.k}^{j} + x_{..}^{j})}_{\text{residual variability}}$$

- Multilevel in Practice:
  1. Calculate the within individual variability
  2. Run (s)PLS on the within individual variability

sparse PLS: Calibration

- Calibration of the number of components based on Q^2 criterion
  - R^2: proportion of variance explained by the model

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{k \in training} (y_k - \hat{y}_k)^2}{\sum_{k \in training} (y_k - \bar{y})^2}$$

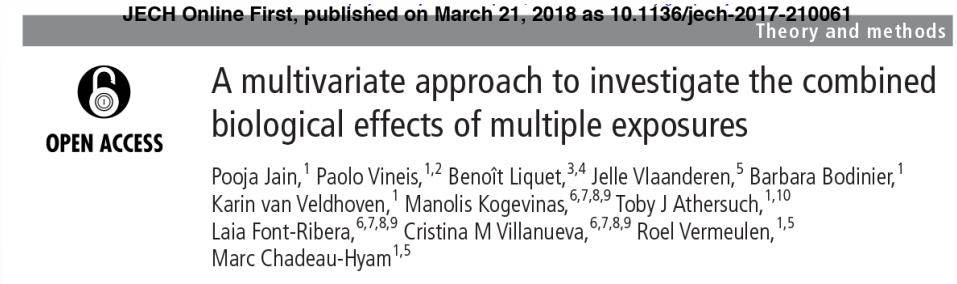
  - Q^2 analogous to R^2 on the testing set:

$$Q^2 = 1 - \frac{\sum_{k \in test} (y_k - \hat{y}_k)^2}{\sum_{k \in test} (y_k - \bar{y})^2}$$

- Calibration of the number of variables to be included in each component (sPLS)

$$MSEP = \frac{1}{n} \sum_{k \in test} (y_k - \hat{y}_k)^2$$

Application to PISCINA data


JECH Online First, published on March 21, 2018 as 10.1136/jech-2017-210061

OPEN ACCESS

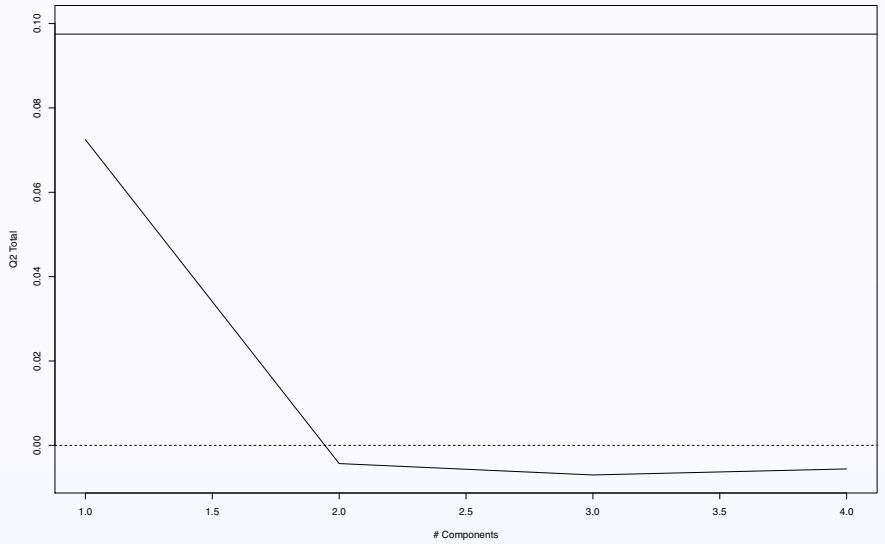
A multivariate approach to investigate the combined biological effects of multiple exposures

Pooja Jain,[1] Paolo Vineis,[1,2] Benoît Liquet,[3,4] Jelle Vlaanderen,[5] Barbara Bodinier,[1] Karin van Veldhoven,[1] Manolis Kogevinas,[6,7,8,9] Toby J Athersuch,[1,10] Laia Font-Ribera,[6,7,8,9] Cristina M Villanueva,[6,7,8,9] Roel Vermeulen,[1,5] Marc Chadeau-Hyam[1,5]

- Question: due to exposure co-occurence, are all exposures needed to explain the inflammatory response?
- Method: Multi-level (sparse) PLS model of (N=4) exposures $vs.$ (N=13) proteins
- Research questions and corresponding models:
  1. how does exposure jointly affect protein levels (PLS)
  2. which (sets of) exposure are affecting proteins level (sPLS on X)
  3. which (sets of) proteins are affected by exposures (sPLS on Y)
  4. what set of exposures most affect a subset of the proteins (sPLS on X& Y)

Model calibration

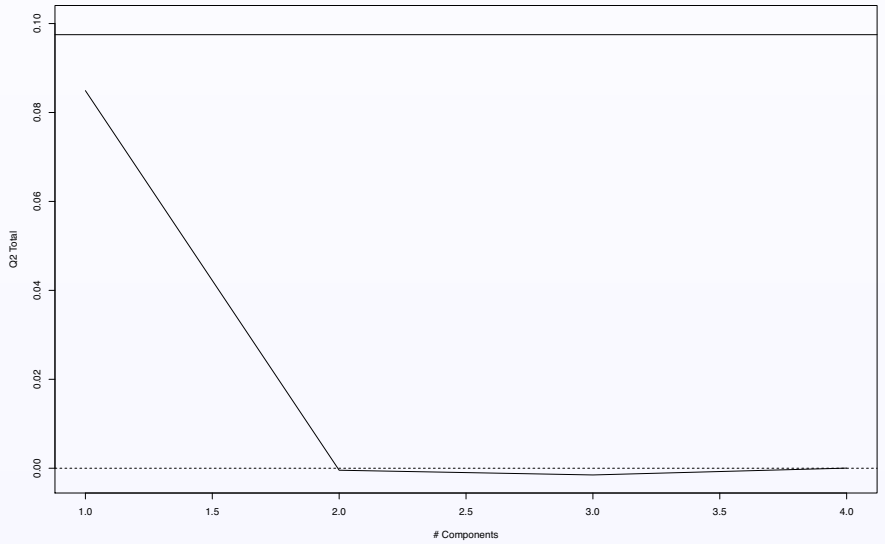
- Number of components:
  1. Q^2 criterion: the maximal the number of components such that adding an additional component would yield a substantive drop in the Q^2 value.
  2. Q^2 by 1,000-5-fold cross-validation: estimate the average Q^2
  3. The number of X and Y components are equal
  4. The maximum number of component is $\min(p = 4, q = 13) = 4$



$\Rightarrow$ for sPLS on X, one component selected.

Model calibration

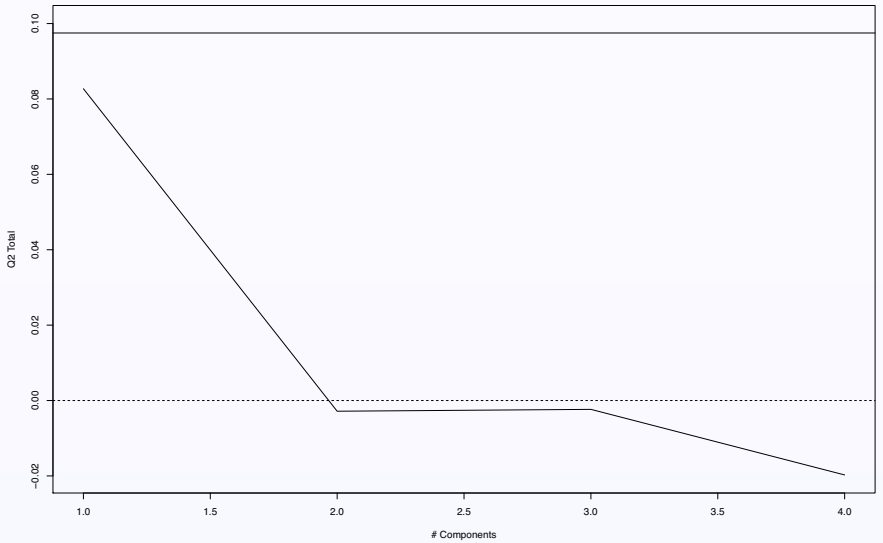
- Number of components:
  1. Q^2 criterion: the maximal the number of components such that adding an additional component would yield a substantive drop in the Q^2 value.
  2. Q^2 by 1,000-5-fold cross-validation: estimate the average Q^2
  3. The number of X and Y components are equal
  4. The maximum number of component is $\min(p = 4, q = 13) = 4$



$\Rightarrow$ for sPLS on Y, one component selected.

Model calibration

- Number of components:
  1. Q^2 criterion: the maximal the number of components such that adding an additional component would yield a substantive drop in the Q^2 value.
  2. Q^2 by 1,000-5-fold cross-validation : estimate the average Q^2
  3. The number of X and Y components are equal
  4. The maximum number of component is $\min(p = 4, q = 13) = 4$



$\Rightarrow$ for sPLS on both X and Y, one component selected.

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow$ Due to exposure correlation: C_{1X} explains $\sim 95\%$ of the variance in X

$\Rightarrow$ All exposures have negative loadings

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow$ Weaker correlations in proteins: C_{1Y} explains $\sim 20\%$ of the variance in X

$\Rightarrow$ only 4 negative loadings (including CXCL10)

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow C_{1X}$ explains $\sim 10\%$ of the variance in Y

$\Rightarrow$ limited explanatory performances of the exposures

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow C_{2X}, \ldots, C_{4X}$ explain lesse than 5% of the variance in X

$\Rightarrow$ they explain less than 2% of the Y variance

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow$ variable selection on X excludes $BrCH_3$

$\Rightarrow$ explained variance of X and Y by $C_{1X'}$ are not affected

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%

Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow$ variable selection on Y excludes 4 proteins

$\Rightarrow$ resulting improvements in the Y explained variance

Parameter estimates overview:

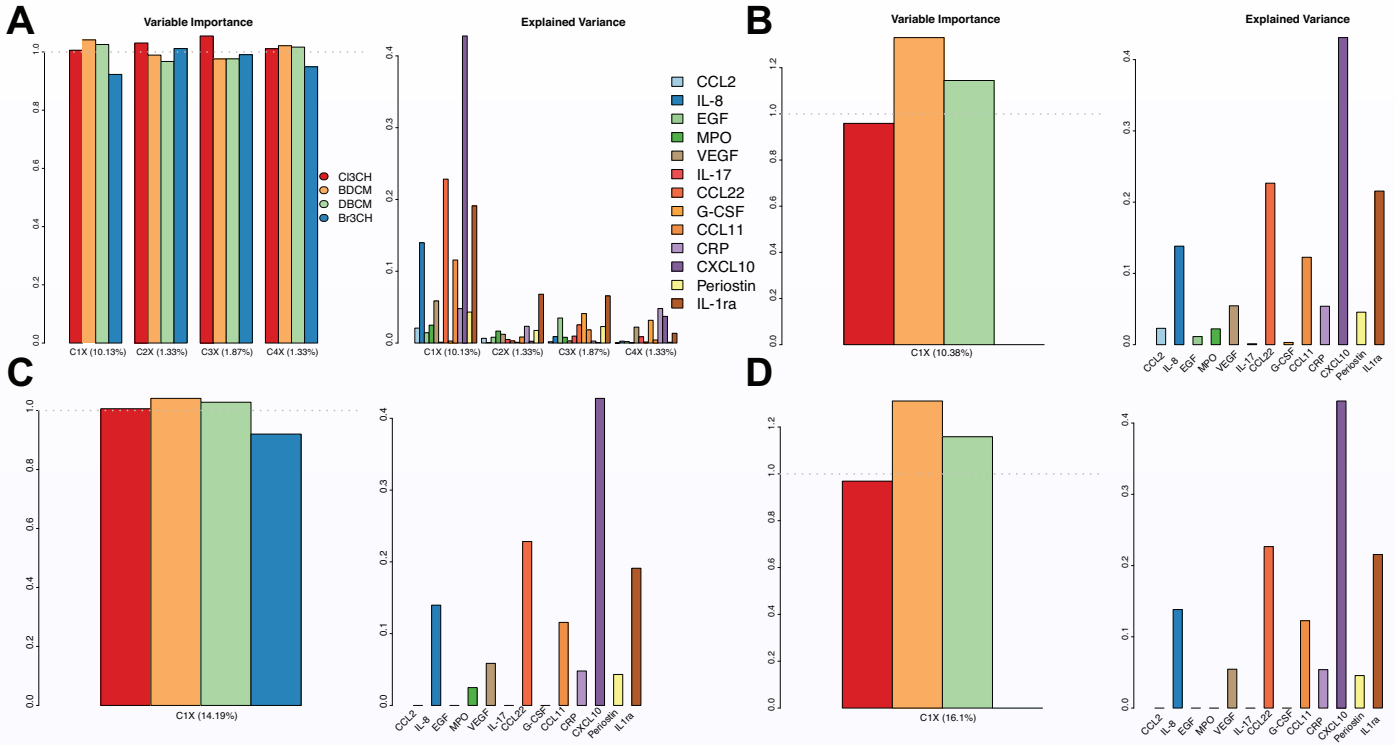
	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%

Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow$ variable selection on X & Y excludes 5 proteins and $BrCH_3$

$\Rightarrow$ yields maximal Y explained variance by C'''_{1X}

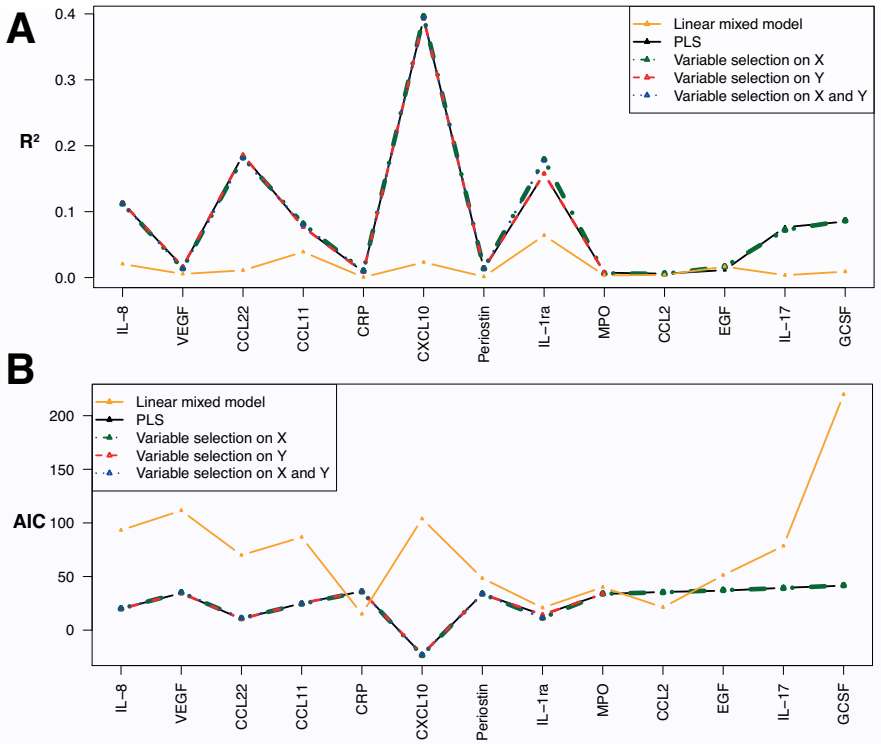
Multi level PLS analyses of PISCINA data



- Despite correlation, 3 exposures selected
  - Bromoform appears less influential (not selected in sPLS-X)
  - Variance is heterogeneously explain across proteins (e.g. IP10, MDCCC, and IL1ra) which are selected in sPLS (on Y)
  - Variable selection on X and Y selects 3 exposures and 8 proteins

Multi level PLS analyses of PISCINA data

- Comparison with mixed models (investigating protein separately):



⇒ accounting for the correlation across proteins (multivariate Y) is improving the model's fit

⇒ All PLS variants have similar performances

⇒ For some proteins LMM and PLS equally mis-perform

⇒ Efficient prioritisation tool