Recent advances on RGCCA

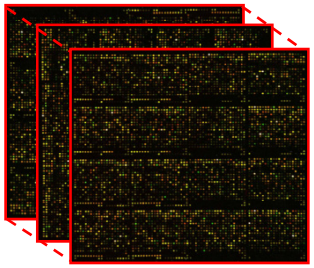
Arthur Tenenhaus, Supelec

Beijing, 28/10/2011

Glioma Cancer Data

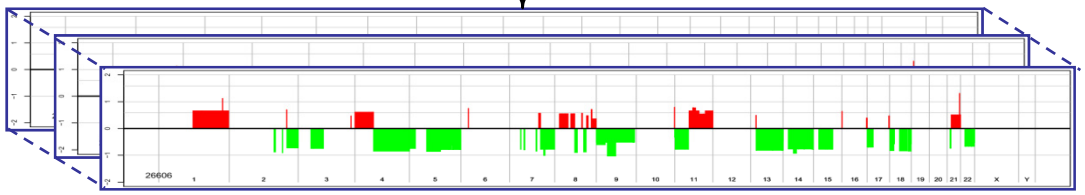
(Department of Pediatric Oncology of the Gustave Roussy Institute)

Transcriptomic data (X_1)



outcome (X_3)

	Gene 1	Gene 2	…	Gene 15201	CGH1	…	CGH 15201	Localization
Patient 1	0.18	-0.21		-0.73	0.00		-0.55	Hemisphere
Patient 2	1.15	-0.45		0.27	-0.30		0.00	Midline
Patient 3	1.35	0.17		0.22	0.33		0.64	DIPG
⋮								
⋮								
Patient 56	1.39	0.18	…	-0.17	0.00	…	0.43	Hemisphere



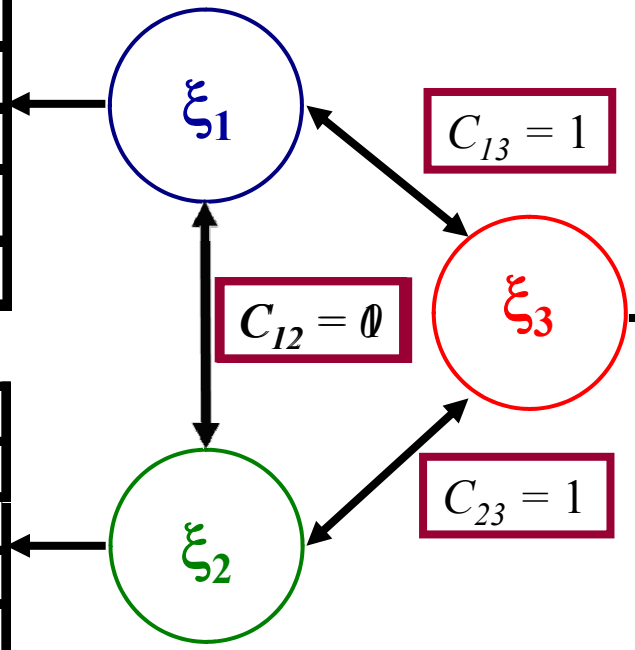
CGH data (X_2)

Glioma Cancer Data: from a multi-block viewpoint

(Department of Pediatric Oncology of the Gustave Roussy Institute)

	Gene 1	…	Gene 15201
Patient 1	0.18		-0.73
Patient 2	1.15		0.27
Patient 3	1.35		0.22
⋮			
Patient 56	1.39		-0.17

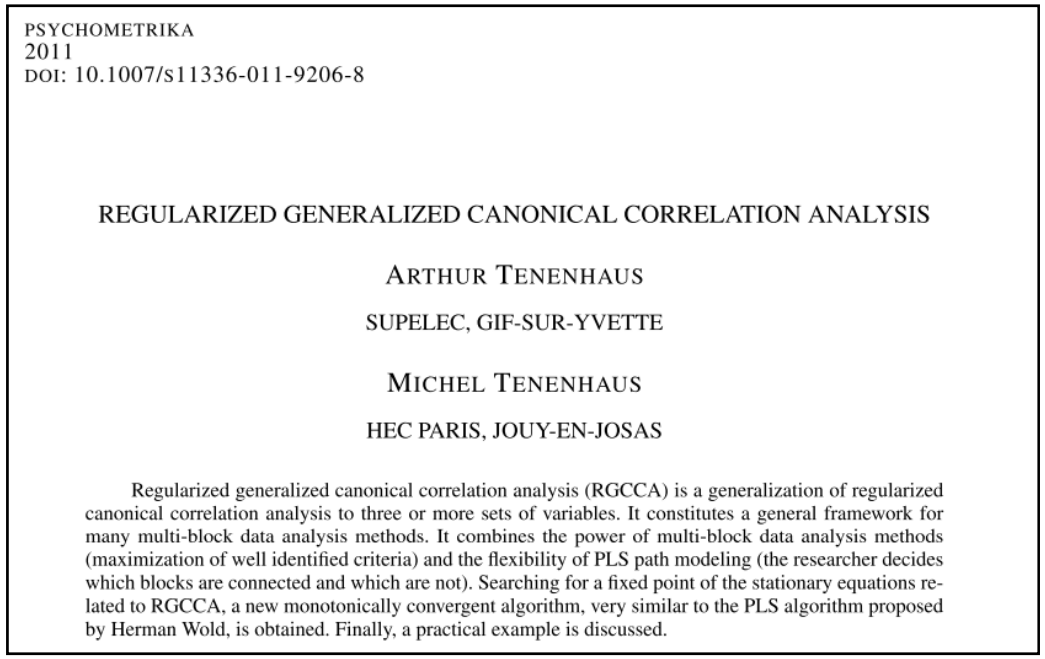
	CGH1	…	CGH 15201
Patient 1	0.00		-0.55
Patient 2	-0.30		0.00
Patient 3	0.33		0.64
⋮			
Patient 56	0.00		0.43



	Hemisphere	DIPG
Patient 1	1	0
Patient 2	0	0
Patient 3	0	1
⋮		
Patient 56	1	0

References

- **Paper**



PSYCHOMETRIKA
2011
DOI: 10.1007/s11336-011-9206-8

REGULARIZED GENERALIZED CANONICAL CORRELATION ANALYSIS

ARTHUR TENENHAUS

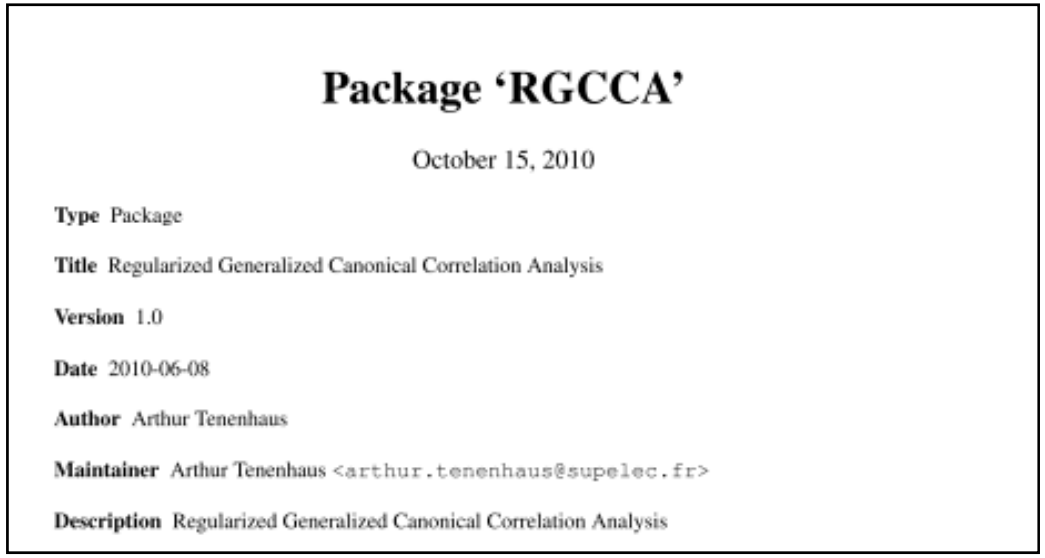
SUPELEC, GIF-SUR-YVETTE

MICHEL TENENHAUS

HEC PARIS, JOUY-EN-JOSAS

Regularized generalized canonical correlation analysis (RGCCA) is a generalization of regularized canonical correlation analysis to three or more sets of variables. It constitutes a general framework for many multi-block data analysis methods. It combines the power of multi-block data analysis methods (maximization of well identified criteria) and the flexibility of PLS path modeling (the researcher decides which blocks are connected and which are not). Searching for a fixed point of the stationary equations related to RGCCA, a new monotonically convergent algorithm, very similar to the PLS algorithm proposed by Herman Wold, is obtained. Finally, a practical example is discussed.

- **Package**



Package 'RGCCA'

October 15, 2010

Type Package

Title Regularized Generalized Canonical Correlation Analysis

Version 1.0

Date 2010-06-08

Author Arthur Tenenhaus

Maintainer Arthur Tenenhaus <arthur.tenenhaus@supelec.fr>

Description Regularized Generalized Canonical Correlation Analysis

Block components

$$\mathbf{y_1} = \mathbf{X_1 a_1} = a_{11}\mathbf{Gene}_1 + \cdots + a_{1,15201}\mathbf{Gene}_{15201}$$

$$\mathbf{y_2} = \mathbf{X_2 a_2} = a_{21}\mathbf{CGH}_1 + \cdots + a_{2,15201}\mathbf{CGH}_{15201}$$

$$\mathbf{y_3} = \mathbf{X_3 a_3} = a_{31}\mathbf{Hemisphere} + a_{32}\mathbf{DIPG}$$

Some modified multi-block methods

SUMCOR (Horst, 1961) $\text{maximize} \sum_{j,k} \text{cor}(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k)$

GENERALIZED CANONICAL CORRELATIC ANALYSIS

SABSCOR (Mathes, 1993 ; Hanafi, 2004) $\text{maximize} \sum_{j,k} |\text{cor}(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k)|$

SUMCOV (Van de Geer, 1984) $\underset{\text{all } \|a_j\|=1}{\text{maximize}} \sum_{j,k} \text{cov}(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k)$

GENERALIZED CANONICAL COVARIANC ANALYSIS

SABSCOV (Krämer, 2006) $\underset{\text{all } \|a_j\|=1}{\text{maximize}} \sum_{j,k} |\text{cov}(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k)|$

Covariance-based criteria

$c_{jk} = 1$ if blocks are linked, 0 otherwise and $c_{jj} = 0$

SUMCOR: $\underset{\text{all var}(\mathbf{X_j a_j})=1}{\text{maximize}} \sum_{j,k} c_{jk}\,\text{cov}(\mathbf{X_j a_j}, \mathbf{X_k a_k})$

SSQCOR: $\underset{\text{all var}(\mathbf{X_j a_j})=1}{\text{maximize}} \sum_{j,k} c_{jk}\,\text{cov}^2(\mathbf{X_j a_j}, \mathbf{X_k a_k})$

SABSCOR: $\underset{\text{all var}(\mathbf{X_j a_j})=1}{\text{maximize}} \sum_{j,k} c_{jk}\,|\text{cov}(\mathbf{X_j a_j}, \mathbf{X_k a_k})|$

SUMCOV: $\underset{\text{all } \|a_j\|=1}{\text{maximize}} \sum_{j,k} c_{jk}\,\text{cov}(\mathbf{X_j a_j}, \mathbf{X_k a_k})$

SSQCOV: $\underset{\text{all } \|a_j\|=1}{\text{maximize}} \sum_{j,k} c_{jk}\,\text{cov}^2(\mathbf{X_j a_j}, \mathbf{X_k a_k})$

SABSCOV: $\underset{\text{all } \|a_j\|=1}{\text{maximize}} \sum_{j,k} c_{jk}\,|\text{cov}(\mathbf{X_j a_j}, \mathbf{X_k a_k})|$

RGGCA optimization problem

$$\underset{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_J}{\operatorname{argmax}} \sum_{j\neq k}^{J} c_{jk}\, g\left(\operatorname{cov}\left(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k\right)\right)$$

Subject to the constraints $(1-\tau_j)\operatorname{var}(\mathbf{X}_j\mathbf{a}_j) + \tau_j \|\mathbf{a}_j\|^2 = 1, j = 1,\ldots,J$

A monotone convergent algorithm related to this optimization problem will be described.

wher … cted

$$g = \begin{cases} \text{identity} & \text{(Horst sheme)} \\ \text{square} & \text{(Factorial scheme)} \end{cases}$$

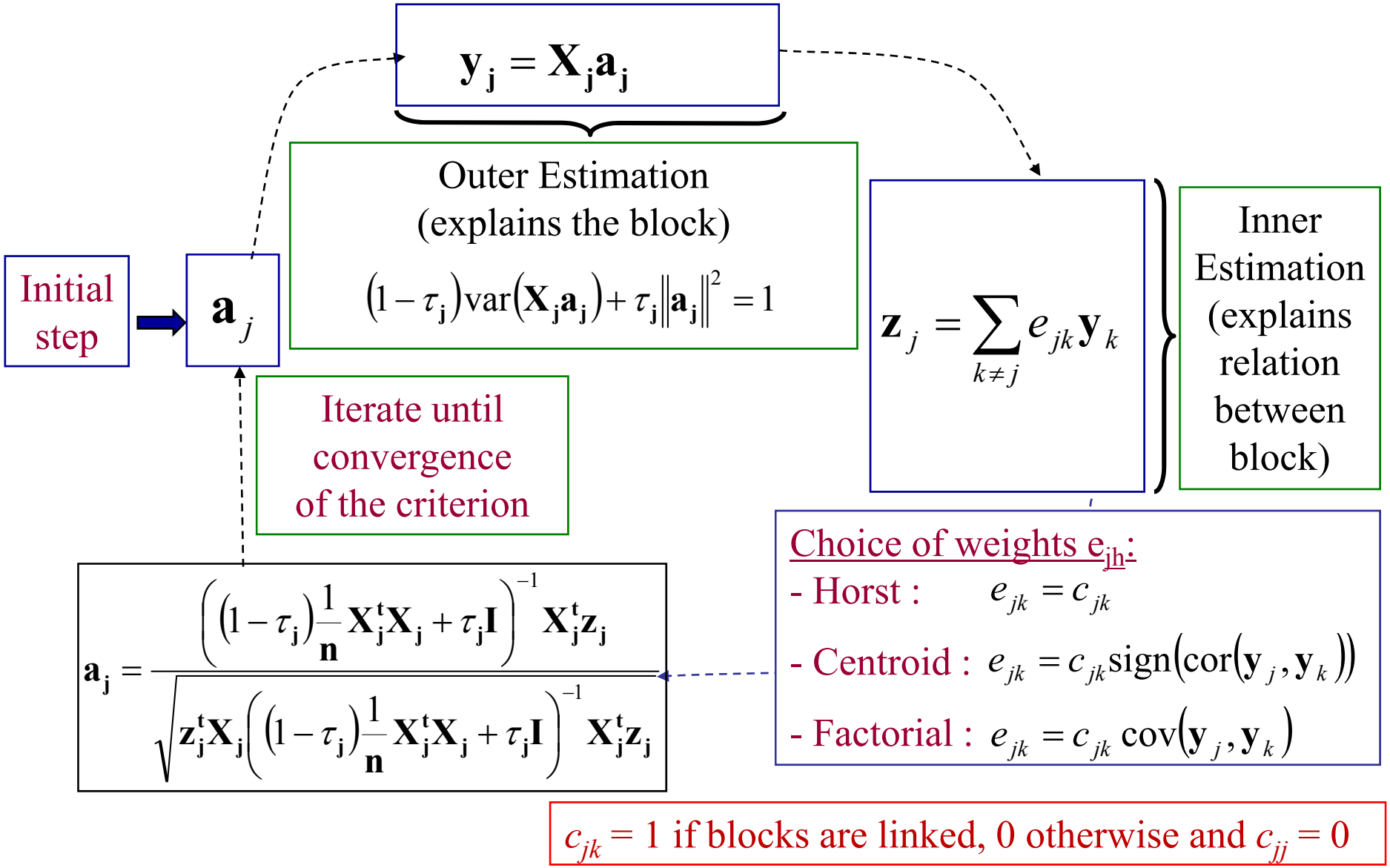
Schäfer and Strimmer formula can be used for an optimal determination of the shrinkage constants

and: τ_j = Shrinkage constant between 0 and 1

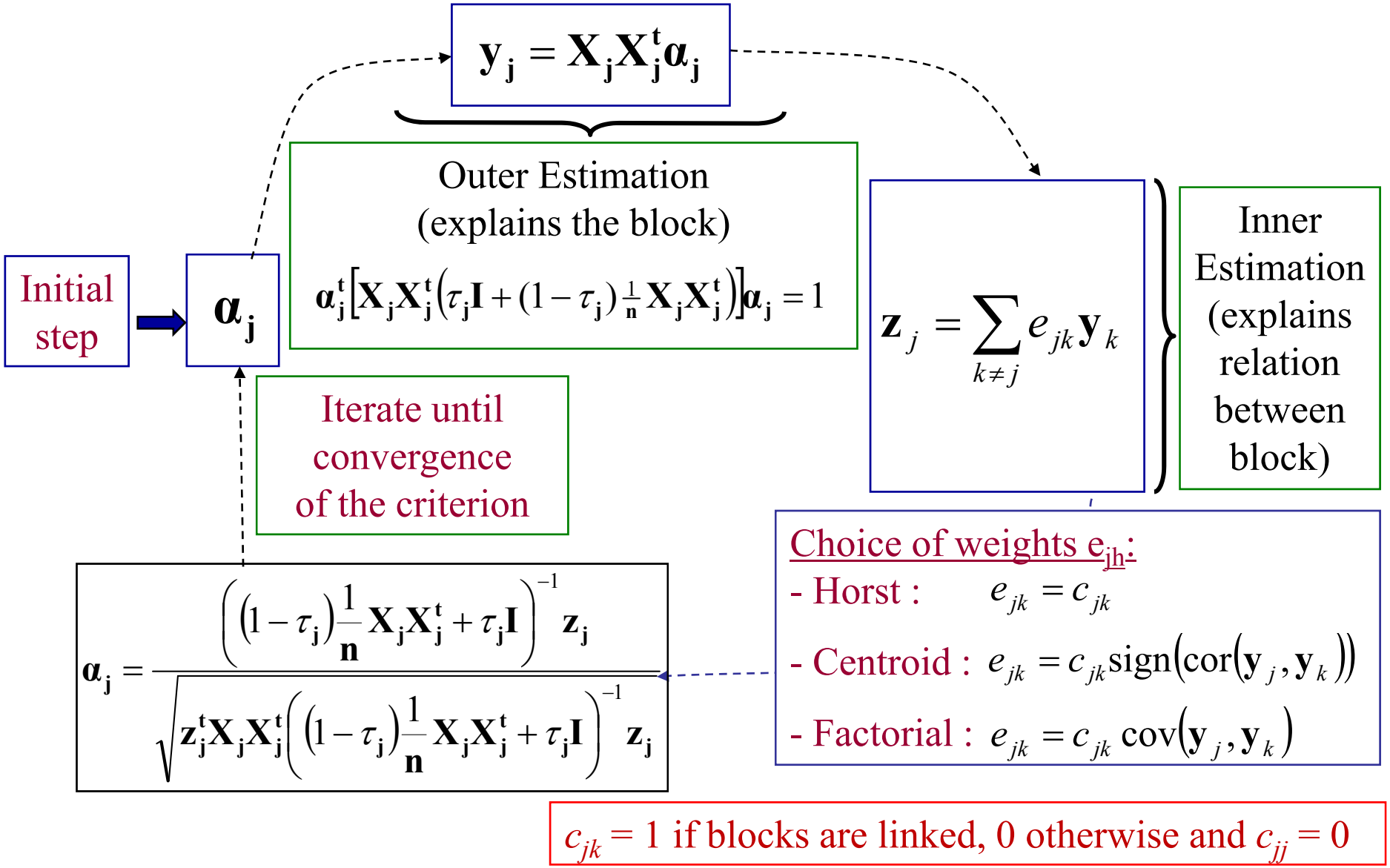
Construction of monotone convergent algorithms for these criteria

- Construct the Lagrangian function related to the optimization problem.
- Cancel the derivative of the Lagrangian function with respect to each $\mathbf{a}_j$.
- Use the Wold's procedure to solve the stationary equations (≈ Gauss-Seidel algorithm).
- This procedure is monotonically convergent: the criterion increases at each step of the algorithm.

The RGCCA algorithm (primal version)



The RGCCA algorithm (dual version)

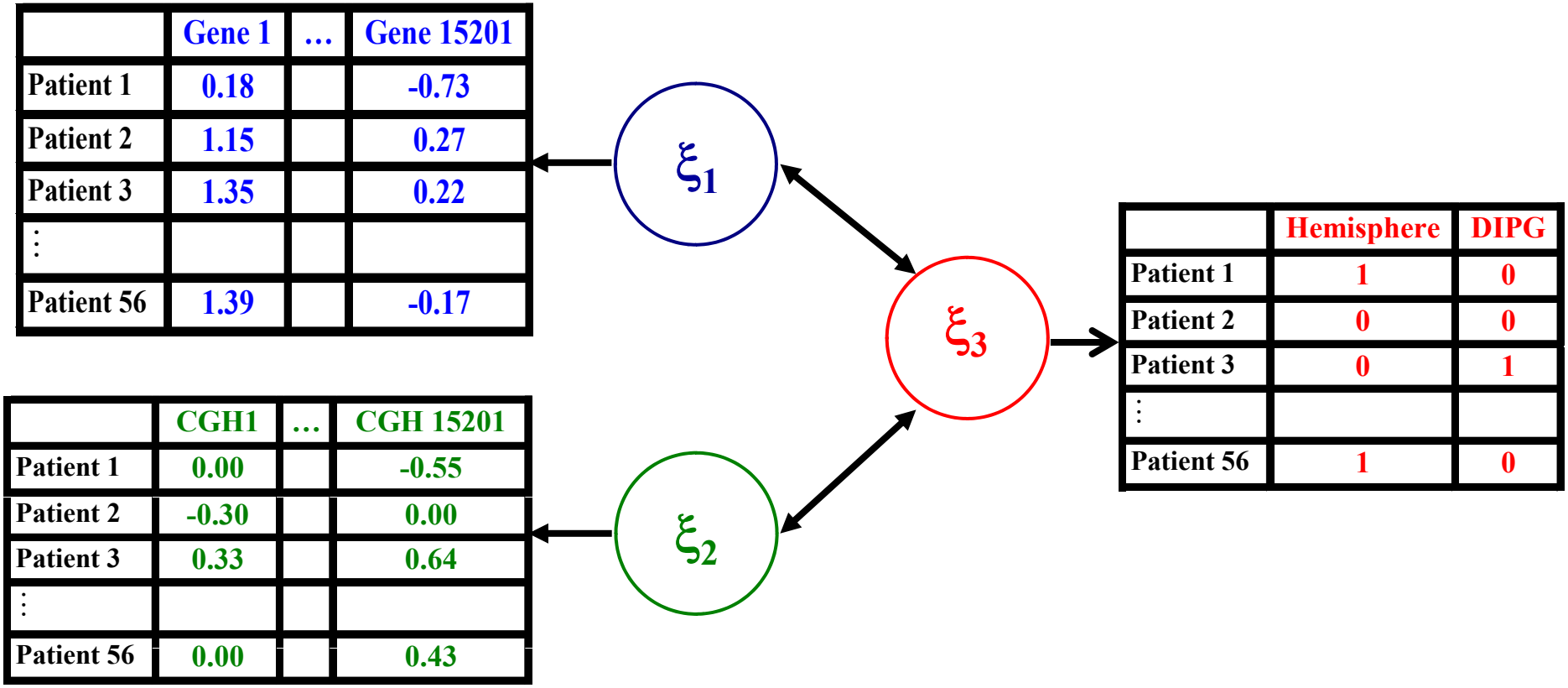


special cases of RGCCA (among others)

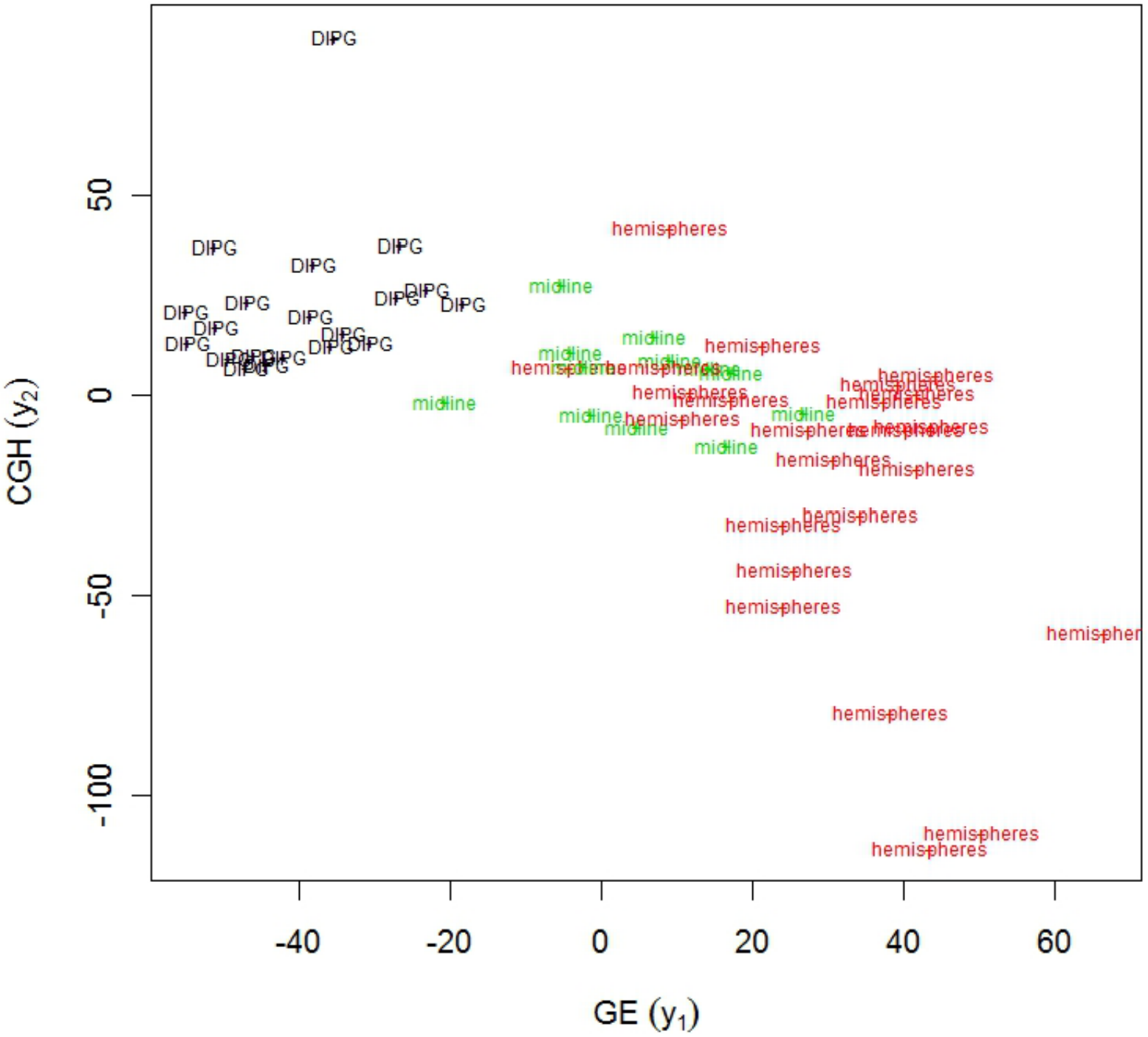
PLS Regression	Wold S., Martens & Wold H. (1983): The multivariate calibration problem in chemistry solved by the PLS method. In Proc. Conf. Matrix Pencils, Ruhe A. & Kåstrøm B. (Eds), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, p. 286-293.
Redundancy analysis	Barker M. & Rayens W. (2003): Partial least squares for discrimination, *Journal of Chemometrics*, 17, 166-173.
Regularized CCA	Vinod H. D. (1976): Canonical ridge and econometrics of joint production. *Journal of Econometrics,* 4, 147–166.
Inter-battery factor analysis	Tucker L.R. (1958): An inter-battery method of factor analysis, *Psychometrika*, vol. 23, n° 2, pp. 111-136.
MCOA	Chessel D. and Hanafi M. (1996): Analyse de la co-inertie de *K* nuages de points. *Revue de Statistique Appliquée*, 44, 35-60
SSQCOV	Hanafi M. & Kiers H.A.L. (2006): Analysis of K sets of data, with differential emphasis on agreement between and within sets, *Computational Statistics & Data Analysis*, 51, 1491-1508.
SUMCOR	Horst P. (1961): Relations among m sets of variables, *Psychometrika*, vol. 26, pp. 126-149.
SSQCOR	Kettenring J.R. (1971): Canonical analysis of several sets of variables, *Biometrika*, 58, 433-451
MAXDIFF	Van de Geer J. P. (1984): Linear relations among k sets of variables. *Psychometrik*a, 49, 70-94.
PLS path modeling (mode B)	Tenenhaus M., Esposito Vinzi V., Chatelin Y.-M., Lauro C. (2005): PLS path modeling. *Computational Statistics and Data Analysis,* 48, 159-205.
Generalized Orthogonal MCOA	Vivien M. & Sabatier R. (2003): Generalized orthogonal multiple co-inertia analysis (-PLS): new multiblock component and regression methods, *Journal of Chemometrics*, 17, 287-301.
Caroll's GCCA	Carroll, J.D. (1968): A generalization of canonical correlation analysis to three or more sets of variables, *Proc. 76th Conv. Am. Psych. Assoc*., pp. 227-228.

Glioma Cancer Data: from an RGCCA viewpoint

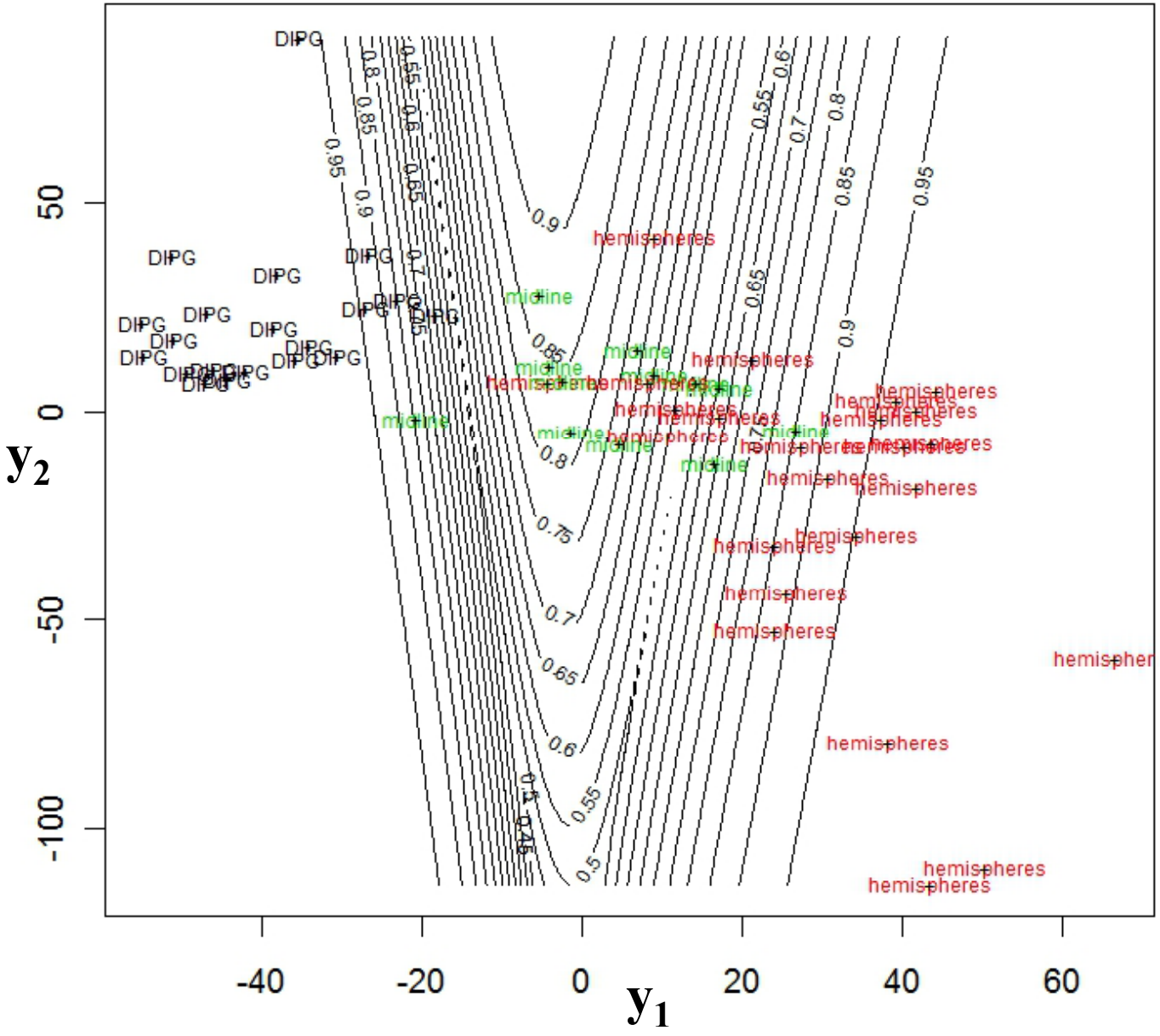
(Department of Pediatric Oncology of the Gustave Roussy Institute)



High dimensional block settings ⟹ dual algorithm for RGCCA


DIPG
hemispheres
midline
CGH (y_2)
GE (y_1)
50
0
-50
-100
-40
-20
0
20
40
60

Bayesian Discriminant Analysis of localization on y_1 and y_2



Predictive performance

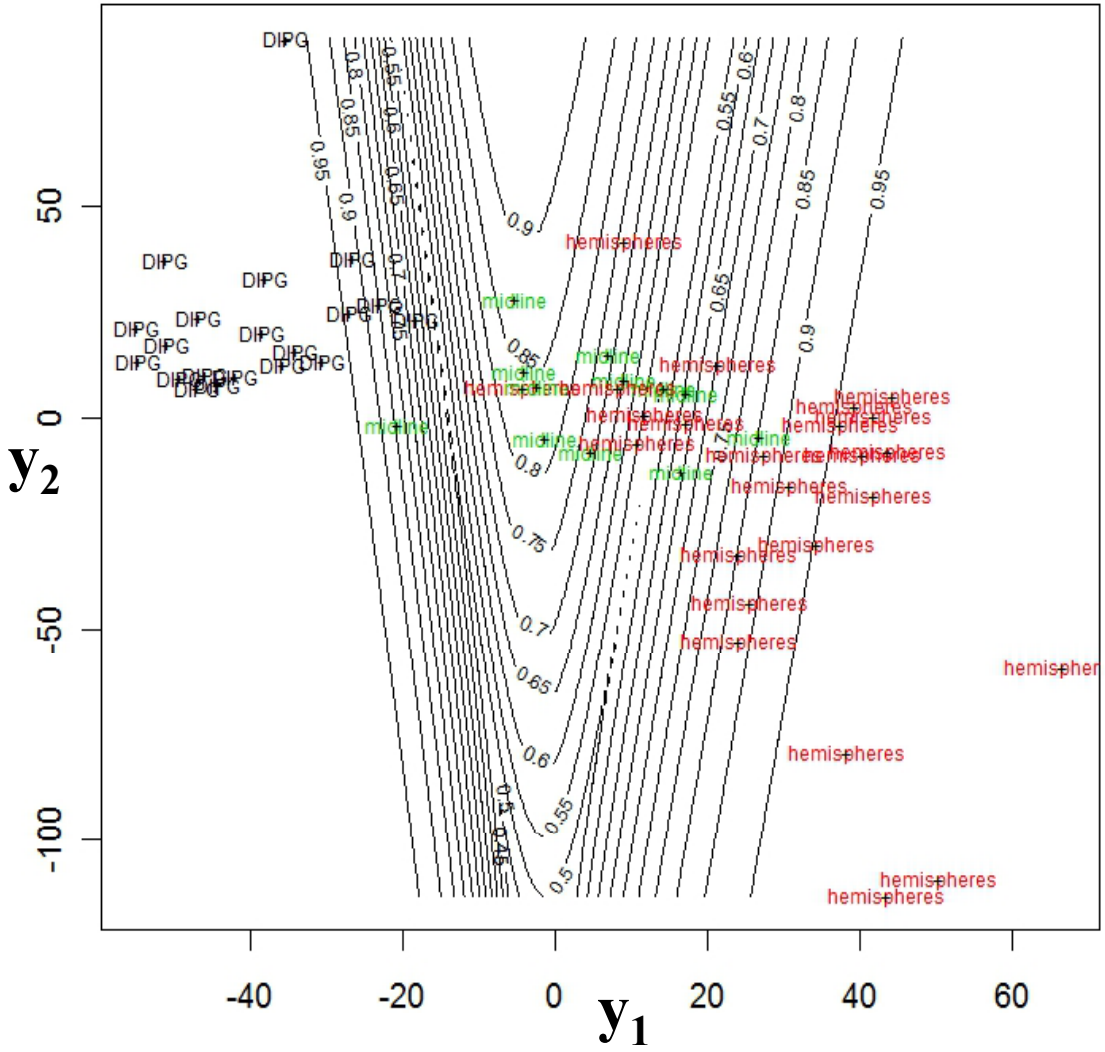


Table 1. Learning phase

Observed / Predicted	DIPG	Hemispheres	Midline
DIPG	20	0	1
Hemispheres	0	19	4
Midline	0	5	7

Accuracy = 82%

Table 2. Testing phase (leave-one-out)

Observed / Predicted	DIPG	Hemispheres	Midline
DIPG	18	1	1
Hemispheres	0	17	4
Midline	2	6	7

Accuracy = 75%

Variable selection for RGCCA

$$\underset{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_J}{\operatorname{argmax}} \sum_{j \neq k}^{J} c_{jk}\, g\left(\operatorname{cov}\left(\mathbf{X}_j \mathbf{a}_j, \mathbf{X}_k \mathbf{a}_k\right)\right)$$

Subject to the constraints

$$\begin{cases} \|\mathbf{a}_j\|_2^2 = 1, j = 1, \ldots, J \\ \|\mathbf{a}_j\|_1 \leq c_j, j = 1, \ldots, J \end{cases}$$

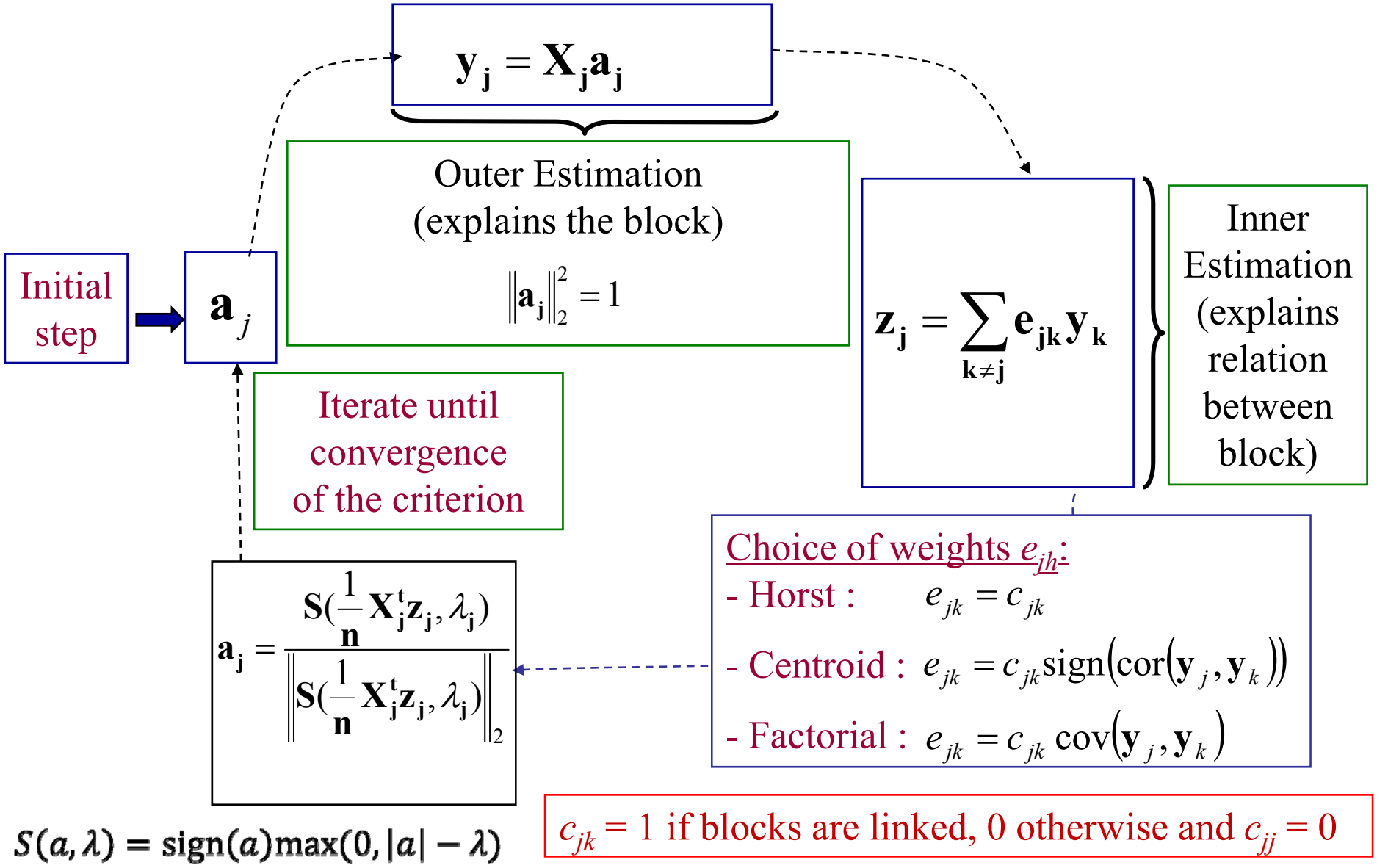
where:

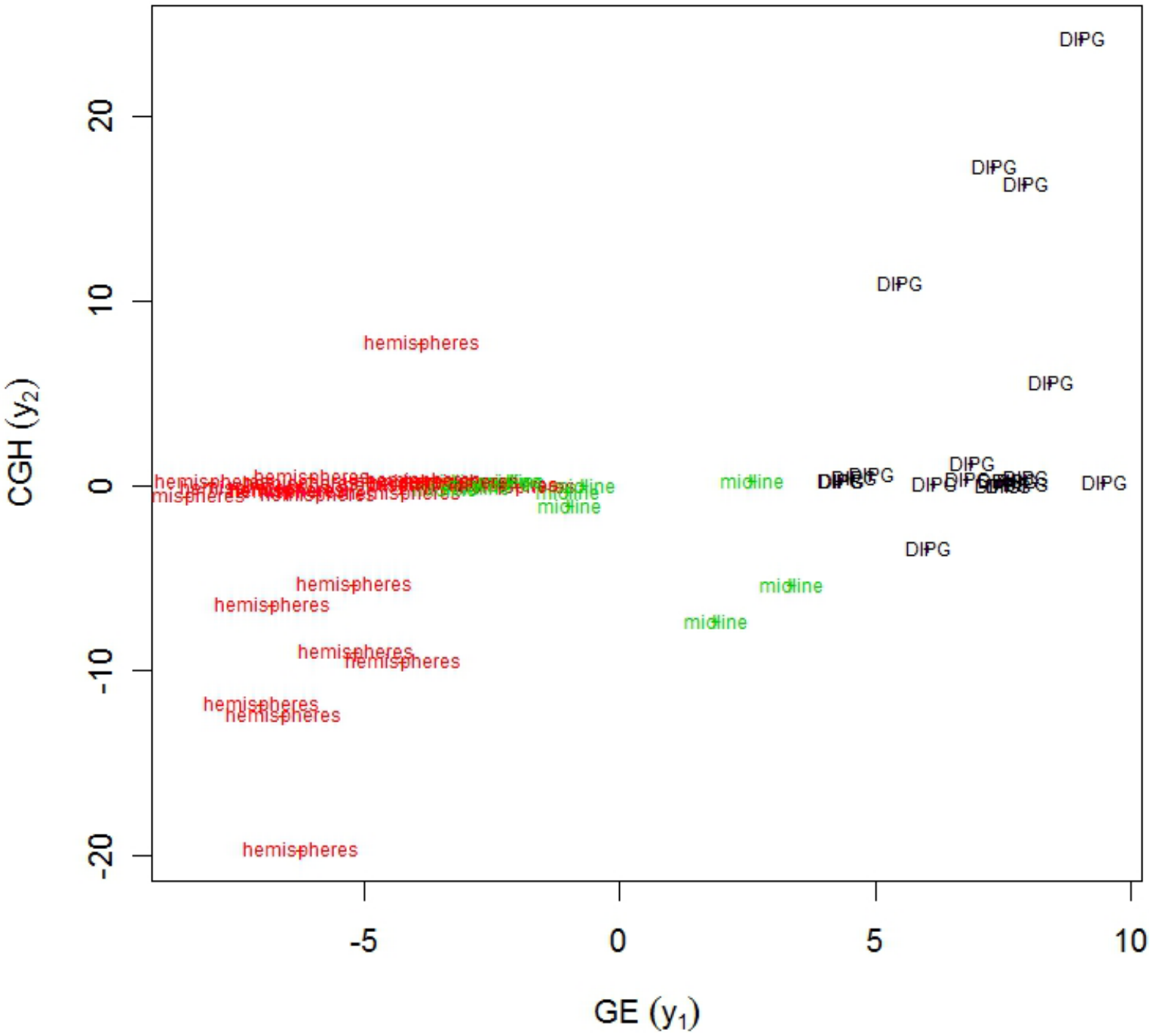
$$c_{jk} = \begin{cases} 1 & \text{if } \mathbf{X_j} \text{ and } \mathbf{X_k} \text{ is connected} \\ 0 & \text{otherwise} \end{cases}$$

$$g = \begin{cases} \text{identity} & \text{(Horst sheme)} \\ \text{square} & \text{(Factorial scheme)} \\ \text{abolute value} & \text{(Centroid scheme)} \end{cases}$$

and: $\tau_j =$ Shrinkage constant between 0 and 1

Sparse GCCA


$\mathbf{y_j} = \mathbf{X_j a_j}$
Outer Estimation
(explains the block)
$\|\mathbf{a_j}\|_2^2 = 1$
Initial step
$\mathbf{a}_j$
$\mathbf{z_j} = \sum_{\mathbf{k \neq j}} \mathbf{e_{jk} y_k}$
Inner Estimation (explains relation between block)
Iterate until convergence of the criterion
$\mathbf{a_j} = \frac{\mathbf{S}(\frac{1}{\mathbf{n}}\mathbf{X_j^t z_j}, \lambda_j)}{\left\|\mathbf{S}(\frac{1}{\mathbf{n}}\mathbf{X_j^t z_j}, \lambda_j)\right\|_2}$
Choice of weights e_{jh}:
- Horst : $e_{jk} = c_{jk}$
- Centroid : $e_{jk} = c_{jk}\,\mathrm{sign}(\mathrm{cor}(\mathbf{y}_j, \mathbf{y}_k))$
- Factorial : $e_{jk} = c_{jk}\,\mathrm{cov}(\mathbf{y}_j, \mathbf{y}_k)$
$S(a, \lambda) = \mathrm{sign}(a)\max(0, |a| - \lambda)$
c_{jk} = 1 if blocks are linked, 0 otherwise and c_{jj} = 0


DIPG
hemispheres
midline
CGH (y_2)
GE (y_1)
20
10
0
-10
-20
-5
0
5
10

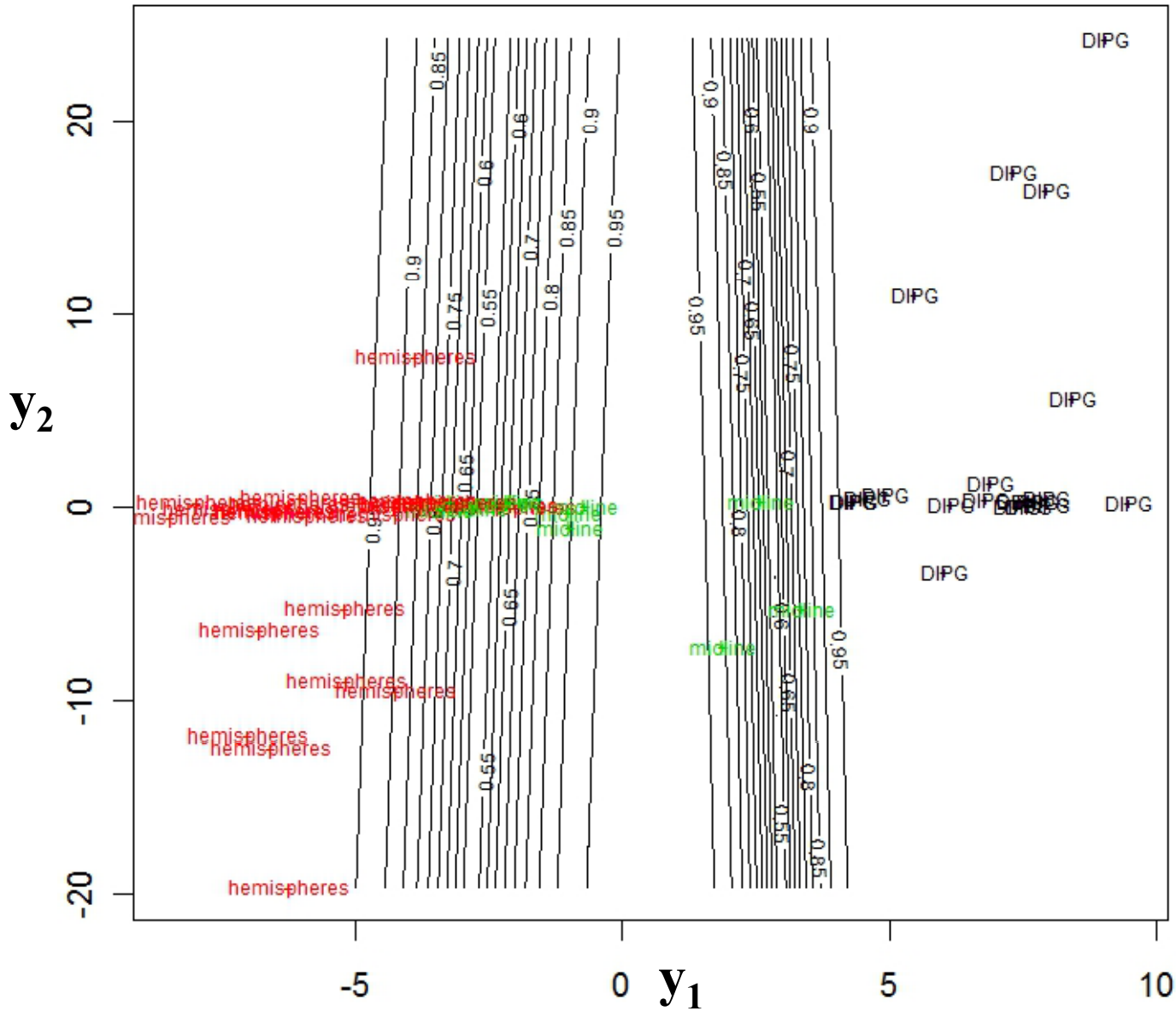
List of selected variables from GE data

FOXG1	PTPN9	CYP4Z1	ARFGAP3
ZFHX4	WNT5A	PI16	PDLIM4
EEPD1	COL10A1	TRIM43	VIPR2
GRID2	PBX3	BTC	ACADL
EMX1	TKTL1	PKNOX2	LAMB3
DLX2	LY6D	SERPINB10	DCAF6
ITM2C	CRYGD	TAAR2	NET1
SEMA3D	HOXA3	ZNF469	ELOVL2
PTHLH	KRTAP9-9	FAM196B	DAAM2
RASL12	LHX1	SLC22A3	CHCHD7
PPAPDC1A	ZNF483	HOXB2	FAIM
HCG4	NLRP7	SLC25A2	HOXA2
TRIM16L	ABI3BP	HES4	SPEF2
NR0B1	MCF2	SYT9	C8orf47
LHX2	SATB2	C2orf88	DLEC1
RNF182	HTR1D	CLDN3	FZD7
KIAA0556	LOXHD1	GLUD2	PLIN4
VAX2	IRX1	OMP	KAL1
ABP1	NRN1	KCND2	LRRC55
SFRP2	C14orf23	C17orf71	FAM89A
HERC3	IRX2	ADAMTS20	RSPH1
SPDEF	C1orf53	SLC1A6	AKR1C3
ONECUT2	GLIS1	SORD	C11orf86
OTX1	HELB	VPS37B	TBX15
OSR1	DLX1	NR2E1	SEMG2

List of selected variables from CGH data

KRAS	STK38L	BBS10	TMEM19
APOLD1	CAPRIN2	TSPAN11	HEBP1
CDKN2B	SOX5	GPRC5D	BHLHE41
CDKN2A	AMN1	GPRC5A	C12orf36
CNOT2	THAP2	DENND5B	RAB21
ABCC9	PYROXD1	NAP1L1	C12orf72
CAPS2	PHLDA1	KLHDC5	GSG1
IAPP	CSRP2	DDX47	C9orf53
PPFIBP1	KRR1	C12orf28	GLIPR1
NAV3	PTPRR	LDHB	PTPRB
SLCO1A2	TM7SF3	FAR2	E2F7
PTHLH	ZFC3H1	ST8SIA1	KIAA0528
ELK3	CCDC91	LRMP	LGR5
KIAA1467	KCNC2	EMP1	ZDHHC17
ETNK1	SLCO1B1	C12orf11	MRPS35
RAB3IP	BCAT1	OSBPL8	C12orf70
TMTC1	LYRM5	KCNJ8	TBC1D15
DDX11	RASSF8	TSPAN8	SSPN
GLIPR1L2	MED21	CASC1	
ITPR2	FGFR1OP2	KCNMB4	

Bayesian Discriminant Analysis of localization on y_1 and y_2



Predictive performance

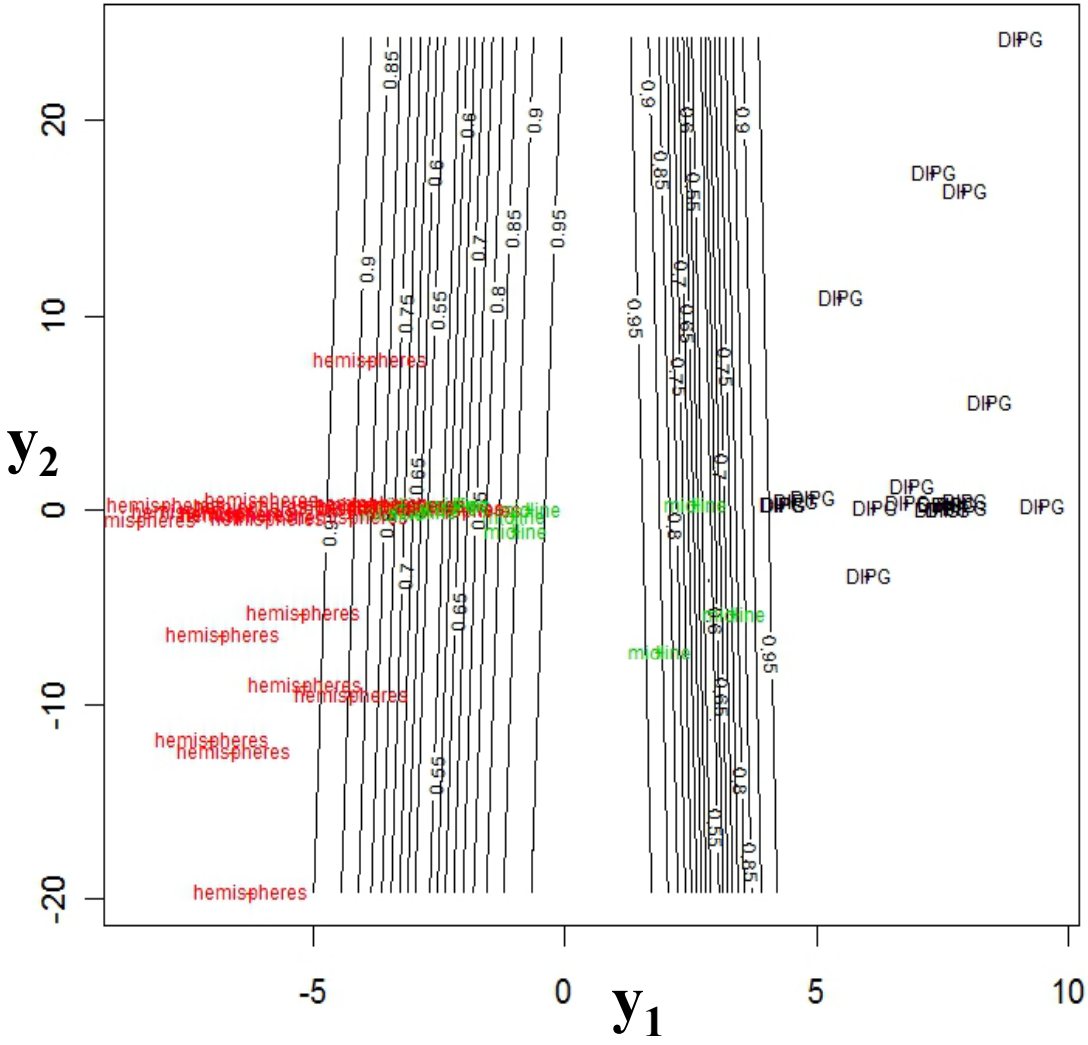


Table 1. Learning phase

Observed / Predicted	DIPG	Hemispheres	Midline
DIPG	20	0	1
Hemispheres	0	22	3
Midline	0	2	8

Accuracy = 89.2%
(82% non sparse)

Table 2. Testing phase (leave-one-out)

Observed / Predicted	DIPG	Hemispheres	Midline
DIPG	20	0	1
Hemispheres	0	20	3
Midline	0	4	8

Accuracy = 85.7%
(75% non sparse)

Conclusion and perspectives

- Depending on the dimension of the blocks, you can use either the primal or the dual algorithm.

- The dual representation of the RGCCA algorithm allows:
    - dealing with symbolic data
    - recovering nonlinear relationship between blocks

- Sparse constraints are useful when the relevant variables are masked by (too many) noisy variables.

- Sparse constraints are useful when we want to select a small number of significant variables which are active in the relationships between blocks.