Principal Bicorrelation Analysis:
Unraveling Associations Between Three Data Sources.

Supplementary Material

FEDERICO MATTIELLO¹, OLIVIER THAS^{1,2}, AND BIE VERBIST¹

Members of the QSTAR Consortium

¹*Department of Mathematical Modelling, Statistics and Bioinformatics,
Ghent University*

²*National Institute for Applied Statistics Research Australia (NIASRA),
School of Mathematics and Applied Statistics, University of Wollongong,
Australia*

March 2, 2015

Federico Mattiello is a PhD student (e-mail: Federico.Mattiello@UGent.be); Olivier Thas is a tenured professor; and Bie Verbist is a postdoctoral researcher. All three work at the Department of Mathematical Modelling, Statistics and Bioinformatics of Ghent University, Coupure Links 653, 9000 Ghent, Belgium and they are members of the QSTAR consortium.

A SIMULATIONS SETTINGS

Table 1 shows, for each type of FF, the mean correlations (averaged over the 50 simulations and over the 10 FF of the FF-group) with BA, GSS_1^* , GSS_2^* and the first 10 individual genes that do not belong to GSS_1^* or GSS_2^* .

Table 1: Relevant correlations between the main elements of the three simulated data sets (averaged over the 100 simulation runs and averaged over the 10 FF within a FF-group).

	FF group		
	1	2	3
BA	0.513	-0.016	0.492
GSS_1^*	0.675	-0.016	0.218
GSS_2^*	-0.011	0.845	-0.002
gene 1	-0.339	0.019	-0.121
gene 2	0.015	-0.001	-0.008
gene 3	-0.432	0.017	-0.146
gene 4	0.073	-0.015	0.037
gene 5	0.360	-0.004	0.127
gene 6	-0.010	-0.213	-0.000
gene 7	-0.024	0.216	-0.002
gene 8	-0.015	-0.348	-0.002
gene 9	0.012	0.240	0.000
gene 10	0.015	-0.117	-0.001

B TABLE OF SIMULATION RESULTS

Table 2: Simulation results. FF called at the FDR level of 20% for the four groups of FF and, among the called FF, the number of genes selected from GSS_1^* , GSS_2^* , and from the *noisy genes*. For GSS_1^* and GSS_2^* the last column is = 5 (number of genes composing the GSS^* s) while for the noisy genes there can be more than 5 of them selected, hence ≥ 5 .

FF group	FF calls		GSS*	Number of GSS* genes selected: % (count)					
	count	%		0	1	2	3	4	≥ 5
1st	981	56.1	1	0.4 (4)	1.4 (14)	33.9 (333)	53.5 (525)	8.9 (87)	1.8 (18)
			2	75.3 (739)	13.9 (136)	5.7 (56)	3.4 (33)	1.4 (14)	0.3 (3)
2nd	73	4.2	1	1.4 (1)	9.6 (7)	34.2 (25)	42.5 (31)	11 (8)	1.4 (1)
			2	57.5 (42)	17.8 (13)	9.6 (7)	11 (8)	2.7 (2)	1.4 (1)
3rd	438	25.1	1	0.7 (3)	4.3 (19)	30.4 (133)	53.7 (235)	9.1 (40)	1.8 (8)
			2	68.9 (302)	10.7 (47)	9.6 (42)	7.8 (34)	1.8 (8)	1.1 (5)
4th	256	14.6	1	0.8 (2)	5.1 (13)	33.2 (85)	54.7 (140)	6.2 (16)	0 (0)
			2	64.8 (166)	14.5 (37)	11.7 (30)	5.5 (14)	2.7 (7)	0.8 (2)
Noisy Genes									
1st	981	56.1		0.7 (7)	3.4 (33)	29.3 (287)	16.5 (162)	7.8 (77)	42.3 (415)
2nd	73	4.2		1.4 (1)	11 (8)	27.4 (20)	23.3 (17)	12.3 (9)	24.7 (18)
3rd	438	25.1		2.7 (12)	4.3 (19)	25.6 (112)	14.6 (64)	7.3 (32)	45.4 (199)
4th	256	14.6		3.9 (10)	3.5 (9)	29.7 (76)	16 (41)	6.6 (17)	40.2 (103)

C GENE LOADINGS GRAPHS

Hereafter we report graphs of the gene loadings resulting from the simulations. More in detail, for each group of FFs we report the graph of the gene loadings obtained with those FFs as input, both with our method and for the Lasso regression. The last graph instead refers to the BA~GE Lasso regression, for which we don't have a correspondence with our method as it works with all three data sources together.

Figure 1: Gene loadings in $\hat{\mathbf{w}}_1$ obtained with *PBA*, averaged over the simulation runs for which a FF of the **first group** was called at the FDR level of 20%. For each simulation run only loadings of FFs called at the FDR level of 20% were considered. The “+” symbols represent the average loadings, and the “x” indicate the true loadings with which GSS_1^* and GSS_2^* were built. Note that results are shown only for the first 15 genes: first five are involved in GSS_1^* , second five in GSS_2^* , and last five are just five noisy genes. Our method aims at identifying the genes in GSS_1^* , because these genes are associated with the BA and with FFs 1 to 10 that are in turn associated with the BA. The top panel refers to our method whereas the second panel refers to the Lasso regression of FFs against the GE matrix. Both graphs follow the same logic.

Figure 2: Gene loadings in $\hat{\mathbf{w}}_1$ obtained with the *Lasso regression*, averaged over the simulation runs for which a FF of the **first group** was called at the FDR level of 20%. The “+” symbols represent the average loadings, and the “X” indicate the true loadings with which GSS_1^* and GSS_2^* were built. Results here are basically the same as our method because this group of FFs is linked to the other two data sources.

Figure 3: Gene loadings in $\hat{\mathbf{w}}_1$ obtained with *PBA* (top panel) and with the *Lasso regression* (lower panel), averaged over the simulation runs for which a FF of the **second group** was called at the FDR level of 20%. This graph highlight the big difference between the two methods: Lasso correctly selects the genes linked to the 2nd group FFs (the ones composing GSS_2^*), whereas *PBA* select genes more or less at random because it recognizes the absence of the link between those FFs and the BA.

Figure 4: Gene loadings in $\hat{\mathbf{w}}_1$ obtained with *PBA* (top panel) and with the *Lasso regression* (lower panel), averaged over the simulation runs for which a FF of the **third group** was called at the FDR level of 20%. The information is similar to that of graphs in Figure 1 and 2 as these FFs are similar to those in the first group, although here the association going through the genes is much weaker. This feature is reflected mainly by the magnitude of the loadings values, with slightly better performances of PBA *w.r.t.* Lasso regression.

Figure 5: Gene loadings in $\hat{\mathbf{w}}_1$ obtained with *PBA* (top panel) and with the *Lasso regression* (lower panel), averaged over the simulation runs for which a FF of the **fourth group** (the one under H_0) was called at the FDR level of 20%. The behaviour of the two methods is comparable but loadings values estimated with Lasso regression have more variability, especially on the genes composing the GSS_1^* . This can be interpreted as follows: the method is able to link the structure it finds in the GE matrix to the FF, although the association FF-GE is actually spurious (these FFs are randomly generated).

Figure 6: Gene loadings in $\hat{\mathbf{w}}_1$ obtained with *Lasso regression* of BA over GE, averaged over the simulation runs. This graph is included for sake of completeness, it confirms that Lasso regression is able to detect which genes are linked to the BA, *i.e.* the first 5 ones with which GSS_1^* is built. It is of interest to note that genes 2 and 4 are selected more often than with the PBA in Figure 1, meaning that here Lasso not always selects only one gene out of a set of correlated ones.