Supplementary Material: Covariate-adjusted matrix visualization via correlation decomposition

Han-Ming Wu [1], Yin-Jing Tien [2], Meng-Ru Ho [3,4,5], Hai-Gwo Hwu [6], Wen-chang Lin [5], Mi-Hua Tao [5], and Chun-Houh Chen [2,*]

[1]Department of Mathematics, Tamkang University, Taipei County 25137, Taiwan, R.O.C.

[2] Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C.

[3] Institute of Biomedical Informatics, National Yang-Ming University, Taipei 112, Taiwan, R.O.C.

[4] Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei 115, Taiwan, R.O.C.

[5] Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan, R.O.C.

[6] Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taipei 100, Taiwan, R.O.C. and Department of Psychology, College of Public Health, Neurobiology and Cognitive Science Center, Taipei 100, Taiwan, R.O.C.

1 Simulation Study

We simulated a data matrix $\mathbf{X}_n$ with $N = 200$ rows and $p = 18$ columns associated with the underlying patterns induced by a continuous covariate, a discrete covariate, and a random noise, so that each pattern of the data structure was completely described by a covariate. The aim was to explore the clustering structure of the data points after a covariate adjustment.

Pattern with a continuous covariate. A continuous model data matrix $\mathbf{C}_{200\times 18}$ was constructed using the model $c_{ij} = \beta_{0j} + \beta_{1j} y_i$, where $y_i = i/N$, $\beta_{0j} = j/p$, and $\beta_{1j} \sim a \times \text{uniform}(0.8, 1.0)$, and where $a = 1$ when $j = 1, \cdots, p/2$, and $a = -1$, when $j = p/2 + 1, \cdots, p$, $i = 1, \cdots, N, j = 1, \cdots, p$. The c_{ij}'s were scaled into the range $(-2.8, 2.8)$ (see Figure 1(a), first row).

Pattern with a discrete covariate. A discrete model data matrix $\mathbf{D}_{200\times 18}$ was constructed using the model $d_{ij} = \alpha_{ij} + y_i$, where $y_i \in (0, 1, 2, 3, 4)$, and α_{ij} were constants according to [(2.80, -0.48, -2.32), (-2.16, -3.19, 1.20), (-3.54, -1.23, -2.18), (-4.98, -4.93, -2.62), (-2.11, -4.98, -1.63)], $i = 1, \cdots, N, j = 1, \cdots, p$. The sample size n_k for each category $k = 1, \cdots, 5$ was $\{50, 20, 60, 30, 40\}$. The variable size p_l for each group $l = 1, 2, 3$ was $\{4, 6, 8\}$. The d_{ij}'s were also scaled into the range $(-2.8, 2.8)$

*to whom correspondence should be addressed

(see Figure 1(a), third row).

Random noise. A random noise data matrix $\mathbf{N}_{200\times 18}$ consisted of all data points n_{ij}'s generated from the standard normal (see Figure 1(a), second row).

The data matrix $\mathbf{X}_n$ with mixed patterns was then produced using $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{N}$, shown in the second row of Figure 1(b). We also constructed two simpler noisy data sets with only a single pattern, for comparison. The noisy data set $\mathbf{C}_n$, with continuous pattern, was obtained by summing $\mathbf{C}$ and $\mathbf{N}$ (Figure 1(b), first row), and the noisy data set $\mathbf{D}_n$, with discrete pattern, was obtained by summing $\mathbf{D}$ and $\mathbf{N}$ (Figure 1(b), third row). The meaning of *noisy* here has two aspects: one is the Gaussian noise for numerical values, the other is the random permutation for visualization, as shown in Figure 1(b).

The Pearson correlation matrix for columns and the Euclidean distance matrix for rows were calculated for three model data sets and the three noisy data sets. The matrix maps are shown in Figure 1(a) and (b). As can be seen, the structure of row proximity for model data with a continuous covariate has a smooth transition pattern, while there is a blocking effect with the discrete covariate. This is due to the rows of data being ordered by the corresponding covariates. However, without suitable permutation, no meaningful patterns can be observed in the noisy data sets.

We then applied R2E to the row and column proximities of the three noisy data sets, as shown in Figure 1(c). Clearly, the permuted row proximity data map of the noisy data with continuous or discrete pattern is close to that of the model data, and both covariates have a good ordering. However, the sorted row proximity data map for the noisy data with mixed patterns shows different structures than that with purely continuous or discrete covariate. In this case, the continuous and discrete covariate cannot be ordered in a meaningful way since the noisy data with mixed patterns mimics the real world data set we observed. Without any covariate adjustment on the mixed noisy data set, the underly pattern cannot be recovered. Applying WABA on X_n, the effect of the discrete pattern is removed, and the pattern left is similar to that of the sorted noisy data with continuous pattern (the left hand side of Figure 1(d)). On the other hand, applying adjustment for a continuous covariate, the masked discrete pattern of the noisy data can be revealed as in the right hand side of Figure 1(d). This clustering pattern is close to that of noisy data with only a discrete pattern.

2 The psychosis disorder data

The between-component map $\mathbf{B}$ shown in Fig. 3(a) is permuted by an average-linkage hierarchical clustering tree. By comparing $\mathbf{B}$ and $\mathbf{R}$ in Fig. 2, the negative correlations of the mania symptoms (DL4, TH6-8) with the negative symptoms (NC1-ND4) and the delusion/hallucination symptoms are mostly due to the patients' diagnostic categories.

The between-correlation map $\mathbf{R}^B$ shown in Fig. 3(b) is permuted also by an average-linkage hierarchical clustering tree. All correlations are either positive one or negative one since there are only two diagnostic categories for patients. Two clusters (DL2-TH6) and (NA7-NA6) are formed and are negatively correlated. For 50 between-eta correlations, symptom TH6, with the darkest between-eta, η^B, has the most significant difference between schizophrenic and bipolar disorders.

Fig. 3(c) and (d) show the within-component and within-group correlation (also the total correlations for the adjusted (residual) data) maps as sorted by the rank-two ellipse ordering. Four new symptom groups are identified: (ND2-NE1), (TH5-TH7), (TH3-TH4), and (DL4-DL6). Four symptoms NE1, DL2, BE1, and BE2 were grouped into the original negative symptoms group. The symptoms in the TH (thought disorder) were grouped into two highly correlated subgroups (TH3-TH4, Th5-TH7). All hallucination symptoms (AH1-6) and most of the delusion symptoms (except DL2, DL3) were clustered together after adjusting for patients' diagnostic categories.

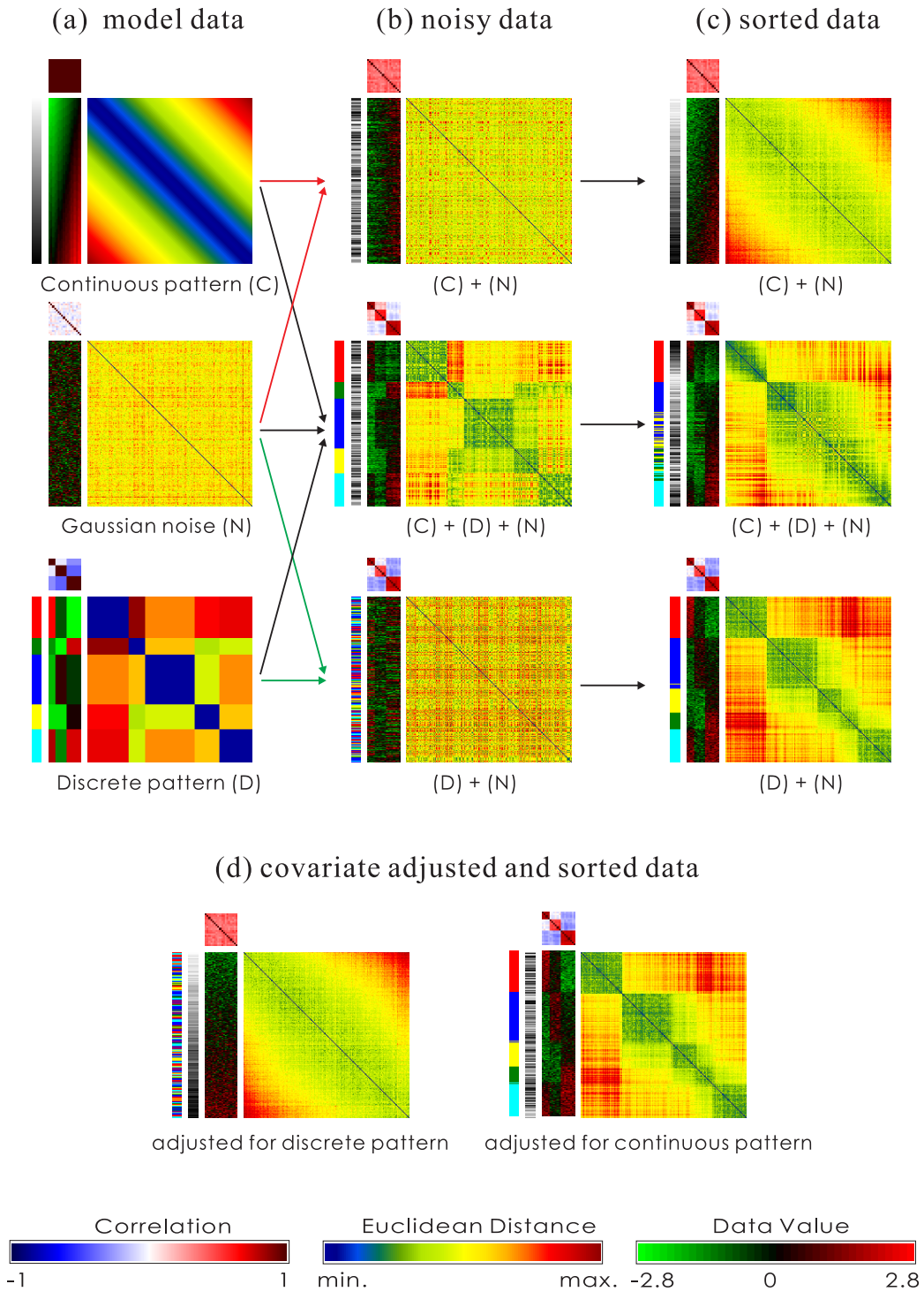

Figure 1: The GAP approach for a covariate adjustment using the simulation data sets: (a) the model data sets, (b) noisy data sets, (c) sorted data sets, and (d) covariate adjusted and sorted data sets.

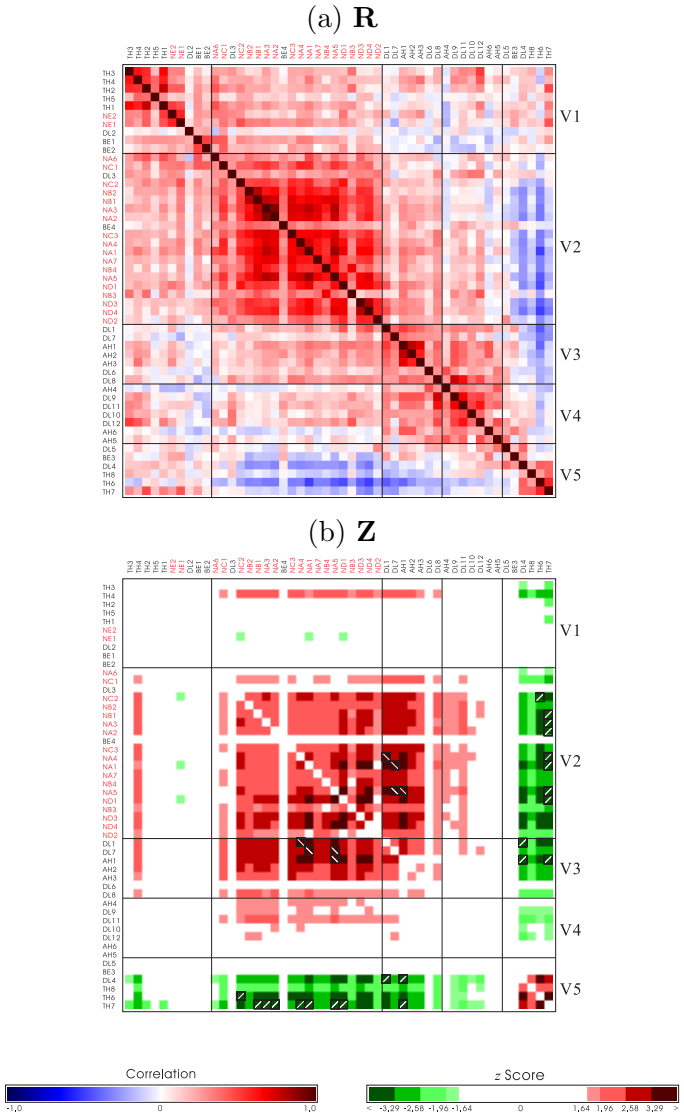

Figure 2: Adjustment for patients' subtype in the psychosis disorder data: (a) the sorted total correlation map **R** by the ellipse seriation for the 50 symptoms, (b) the z-score map with slashes superimposed for reversed correlations in the most significant pairs.

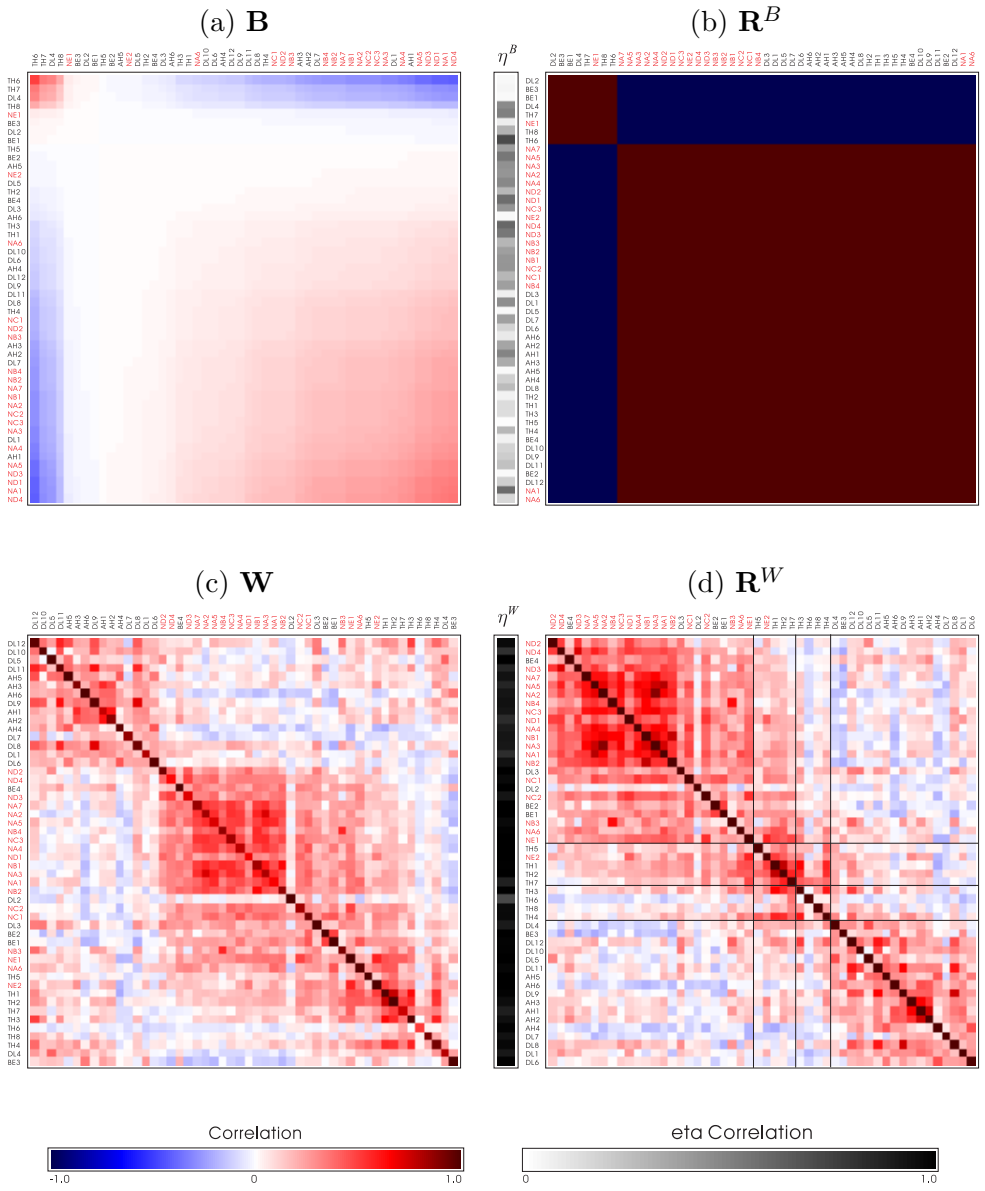

Figure 3: Decomposition of the Pearson correlation matrix for patients' subtype in the psychosis disorder data: (a) the sorted between-component map $\mathbf{B}$, (b) the sorted between-group correlation map $\mathbf{R}^B$, (c) the sorted within-component map $\mathbf{W}$, and (d) the sorted within-group correlation map $\mathbf{R}^W$.