Supplementary Material:
Covariate-adjusted heatmaps for visualizing biological data 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 Statistics, National Taipei University, New Taipei City 23741, 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 Heatmaps in the framework of matrix visualization

We use the GAP (Wu, Tien, and Chen, 2010) approach to illustrate basic principles of MV using the crab data as an example. The Leptograpsus variegatus crab data set (Campbell and Mahon, 1974) contains five morphological measurements: frontal lobe size (FL), rear width (RW), carapace length (CL), carapace width (CW), and body depth (BD) on 50 crabs of each of two species, blue (B) and orange (O), and of each sex. These 200 crabs, collected at Fremantle, Western Australia, have been used to study morphological variation in the species by genders using multivariate approaches.

GAP integrates proximity matrix maps for rows (subjects) and columns (variables) with a data matrix map for a complete MV of a given data matrix. In our presentations of the crab data in Figure 1 we first standardized each of the five morphological measurements. Figure 1(a) displays their matrix visualizations for the 200 crabs (randomly permuted within each species-gender combination): (i) the standardized data matrix map with a green-black-red spectrum representing negative-zero-positive standardized measurements; (ii) the Pearson product moment correlation matrix map for the five morphological variables; and (iii) the Euclidean distance matrix map among the 200 crabs calculated from the five standardized morphological variables. Color coding is used to index two covariates of the crabs with blue/orange for two species and magenta/cyan for female/male. From

*to whom correspondence should be addressed

Figure S 1(a:i), one can roughly identify that the female-blue crabs have relatively smaller morphological measures, while orange crabs are relatively larger in size. We do not have the sampling (capturing) mechanism information, and so are unsure about sampling bias on the species (such as birth season).

The rank-two elliptical seriation (R2E) (Chen, 2002; Tien *et al.*, 2008) is then applied to permute both columns (morphological variables) and rows (crabs), with the resulting matrix visualizations displayed in Figure 1(b). A very smooth trend can be seen in Figure 1(b:iii) for the crab distance matrix map. A similar trend, small (green) to large (red) in size, can also be observed in Figure 1(b:i) for all five sorted morphological variables except RW, where some disturbances can be spotted. The sorted correlation map in Figure S 1(b:ii) also reveals that the correlation pattern of RW to the other four variables is not as coherent as those among the four variables alone. From the corresponding covariate bars in Figure S 1(b), we see that the two covariates, species and gender, have low correlation with the smooth trend identified in the permuted data matrix (Figure 1(b:i)) and the distance matrix (Figure 1(b:iii)), although a weak pattern of blue to orange still exists in the species covariate bar.

1.1 Morphological measurements on leptograpsus variegatus crabs revisited

In this section we use the crab data set with five morphological measurements to illustrate the proposed covariate-adjusted heatmap with a continuous covariate and a discrete one. Principal component analysis (PCA) plays an important role in general morphological growth studies, and we also display the R2E sorted first principal component (PC1) of all five morphological measurements as a covariate in Figure 1(b). An almost perfect (smooth) rainbow spectrum shows that PC1 has extremely high correlation ($r = 0.999$) to the R2E ordering of the Euclidean distance matrix of the 200 crabs.

We then treated PC1 as a latent continuous covariate for adjusting the observed data matrix of five morphological measurements, and Figure 1(c) has the visualizations of three matrices (data, variable correlation, and sample distance) of the residual data (after adjusting for PC1) of five morphological measurements. The green-black-red spectrum for color coding raw morphological measurements has been replaced by a blue-white-red spectrum in representing the after-adjustment residuals. These three adjusted matrix maps have again been sorted using the R2E algorithm. The disorderly pattern of PC1 is now uncorrelated to the overall residual data pattern. Instead the four-group covariate pattern (orange/female, blue/female, blue/male, and orange/male) now corresponds to the four-block pattern in the distance and residual data matrices of five adjusted morphological measurements. All five adjusted morphological variables have a strong correlation with the two covariates, among which the adjusted RW and CW variables have strongest correlation to gender and species, respectively. This analysis demonstrates the effectiveness of matrix visualization with adjustment of PC1 in the classification of gender and species

in the Leptograpsus variegatus crab; this can be carried out in similar morphological and comparative studies.

We move one step further to adjust the discrete covariate effect of the combination of species and gender (orange/female, blue/female, blue/male, and orange/male). The R2E permuted visualization of three matrices of the residual data (after adjusting for species×gender) of five morphological measurements is illustrated in Figure 1(d), where an even smoother trend is seen in the distance and residual data matrices than was found in the unadjusted data in Figure 1(b). The species and gender covariates are uncorrelated with the residual data, as expected. On the other hand, PC1 is highly correlated with the new smooth trend, but not as much as that seen with the unadjusted data in Figure 1(b). We conjecture that this smooth trend in the adjusted residual data of the five morphological measurements is "age of crab". This age-related variable is PC1 (size in general) related, but not identical. The matter requires further data.

2 The psychosis disorder data

The between-component map **B** shown in Figure S3(a) is permuted by an average-linkage hierarchical clustering tree. By comparing **B** and **R** in Figure S 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 Figure S3(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.

Figure S3(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.

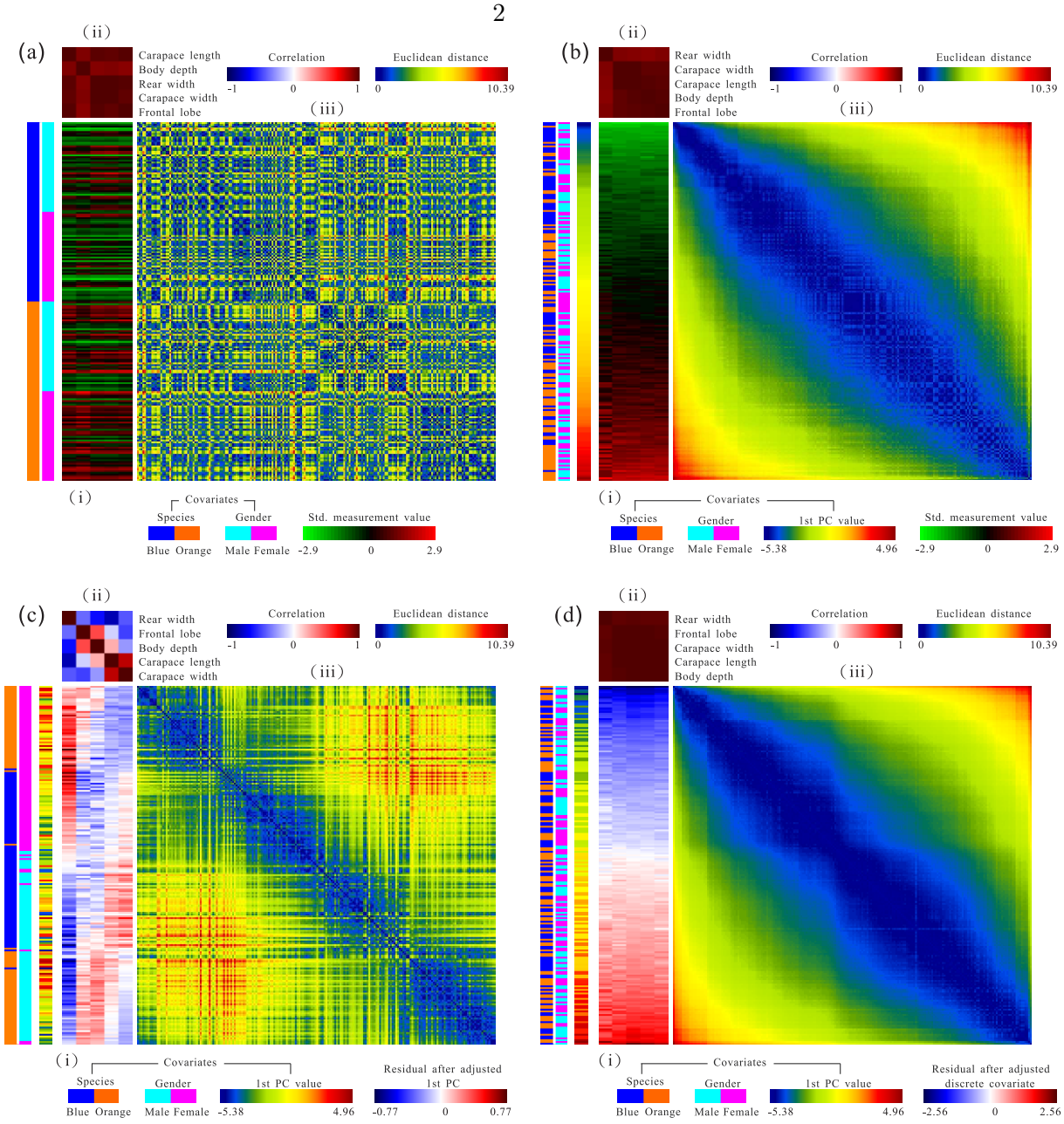

Figure S1: Matrix visualization for the crab data: (a) Original morphological measurements sorted by two covariates. (b) Original morphological measurements sorted by rank-two elliptical seriations. (c) Residuals for morphological measurements after adjustment by the 1st principal component, with covariates and after rank-two elliptical seriation. (d) Residuals for morphological measurements after adjustment for discrete covariates (species, sex), with covariates and after rank-two elliptical seriation. For sub-figures: (i) MV for the standardized morphological measurements (a,b) or for the residuals after covariate-adjustment (c,d); (ii) MV for the correlation matrix of morphological measurements or residuals; (iii) MV for the Euclidean distance matrix of samples.

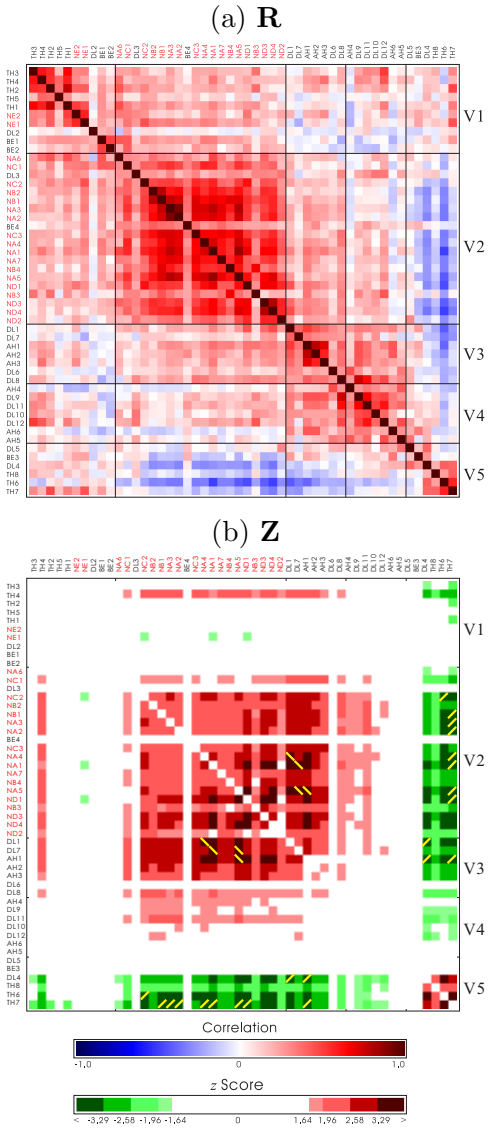

Figure S2: Adjustment for patients' subtype in the psychosis disorder data: (a) the sorted total correlation map **R** by the ellipse seriation for the 50 symptoms, (b) the z-score map with slashes superimposed for reversed correlations in the most significant pairs.

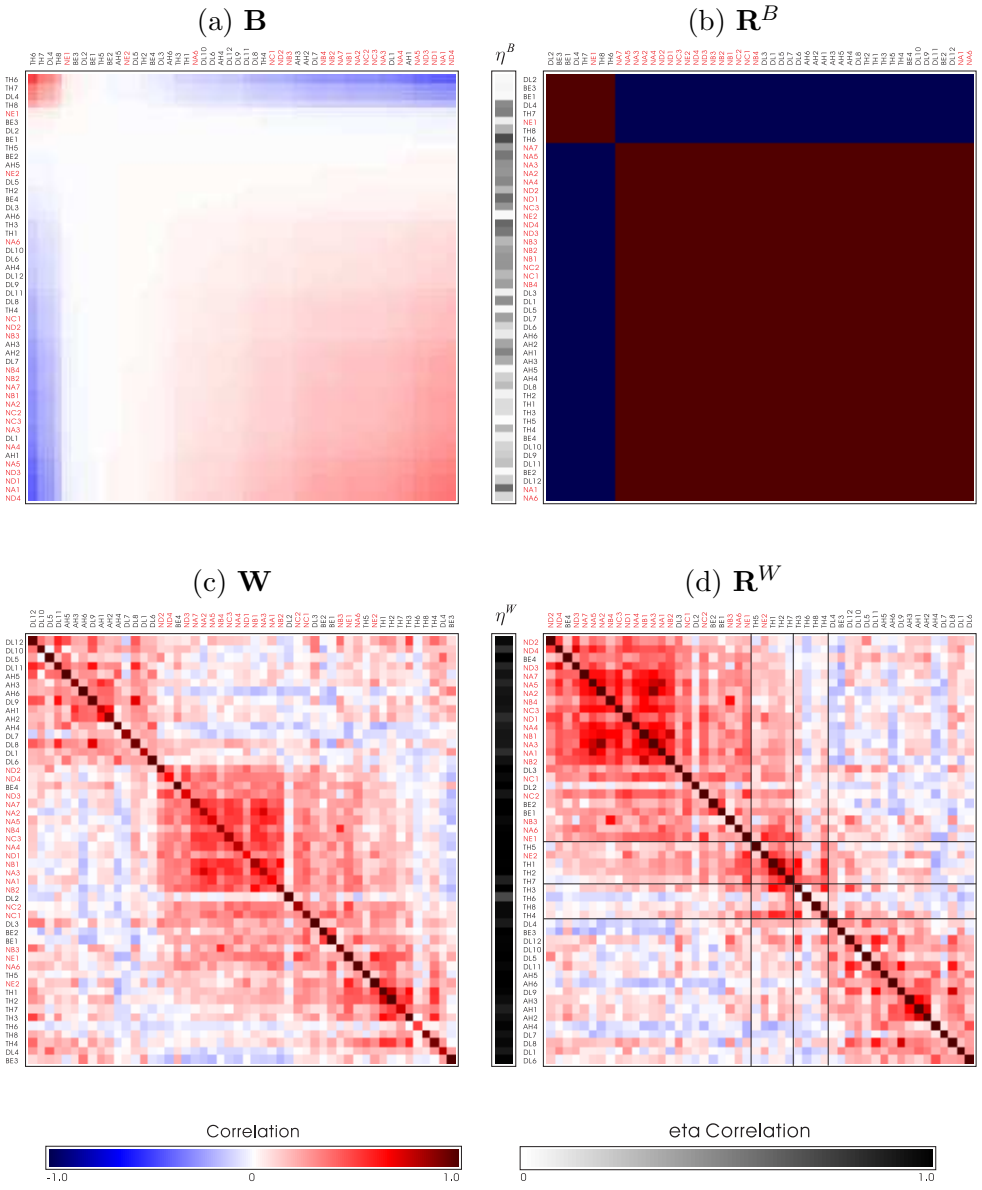

Figure S3: Decomposition of the Pearson correlation matrix for patients' subtype in the psychosis disorder data: (a) the sorted between-component map $\mathbf{B}$, (b) the sorted between-group correlation map $\mathbf{R}^B$, (c) the sorted within-component map $\mathbf{W}$, and (d) the sorted within-group correlation map $\mathbf{R}^W$.