Constrained Principal Component Analysis (CPCA) and Sensometrics

Yoshio Takane
Department of Psychology, University of Victoria, Canada

Sensometrics-2014, Chicago
July 30 to August 2

University of Victoria

Data Analytic Situation

- **Z**: The main data matrix (the sort we may be tempted to apply PCA to). Rows pertain to subjects, and columns to variables.
- **G** and/or **H**: The matrices of external information.

  G: The row (left-hand) side information matrix (e.g., demographic information about the subjects).

  H: The column (right-hand) side information matrix (e.g., stimulus design).

University of Victoria

Constrained Principal Component Analysis (CPCA)

- Two Phases: External Analysis and Internal Analysis.
- External Analysis: Decompose **Z** into additive components according to the external information **G** and/or **H**. This is done by row and/or column regressions.
- Internal Analysis: Apply PCA to the decomposed matrices to explore structures within the components.
- Each term in the decomposition obtained in the External Analysis has specific meaning. PCA results tend to be more interpretable.

University of Victoria

External Analysis: Decompositions of $\mathbf{Z}$

- $\mathbf{Z} = \mathbf{P}_G\mathbf{Z} + \mathbf{Q}_G\mathbf{Z}$ (Column Regression).
- $\mathbf{Z} = \mathbf{Z}\mathbf{P}_H + \mathbf{Z}\mathbf{Q}_H$ (Row Regression).
- $\mathbf{Z} = \mathbf{P}_G\mathbf{Z}\mathbf{P}_H + \mathbf{Q}_G\mathbf{Z}\mathbf{P}_H + \mathbf{P}_G\mathbf{Z}\mathbf{Q}_H + \mathbf{Q}_G\mathbf{Z}\mathbf{Q}_H$ (Row and Column Regressions).
- $\mathbf{P}_G = \mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'$ and $\mathbf{Q}_G = \mathbf{I} - \mathbf{P}_G$ are the orthogonal projectors.
- $\mathbf{P}_G^2 = \mathbf{P}_G = \mathbf{P}_G'$, $\mathbf{Q}_G^2 = \mathbf{Q}_G = \mathbf{Q}_G'$ (symmetric and idempotent), $\mathbf{P}_G\mathbf{Q}_G = \mathbf{Q}_G\mathbf{P}_G = \mathbf{O}$ (orthogonal).
- $\mathbf{P}_H$ and $\mathbf{Q}_H$ are similar.
- $\mathbf{P}_G\mathbf{Z}$ represents the portion of $\mathbf{Z}$ that can be explained by $\mathbf{G}$, and $\mathbf{Q}_G$ that cannot.

University of Victoria

Internal Analysis: SVD

- Singular Value Decomposition (SVD)
- Used for obtaining the best low rank approximation to a matrix.
- Look for the subspace that captures the largest variability in the original space.
- PCA (Principal Component Analysis)

University of Victoria

Example 1: A Vector Preference Model

- 100 subjects made pairwise preference judgments on 9 stimuli.
- 9 celebrities in three distinct groups, 3 politicians (1. Brian Mulroney, 2. Ronald Reagan, 3. Margaret Thatcher), 3 athletes (4. Jacqueline Gareau, 5. Wayne Gretzky, 6. Steve Podborski), and 3 entertainers (7. Paul Anka, 8. Tommy Hunter, 9. Anne Murray).
- Stimuli are presented in pairs to the subjects who were asked to rate the degree to which they prefer one to the other on 25-point rating scales.
- $\mathbf{Z}$ is a 100 subjects by 36 stimulus pairs matrix.

University of Victoria

$\mathbf{Z} = \mathbf{Z}\mathbf{P}_H + \mathbf{Z}\mathbf{Q}_H$

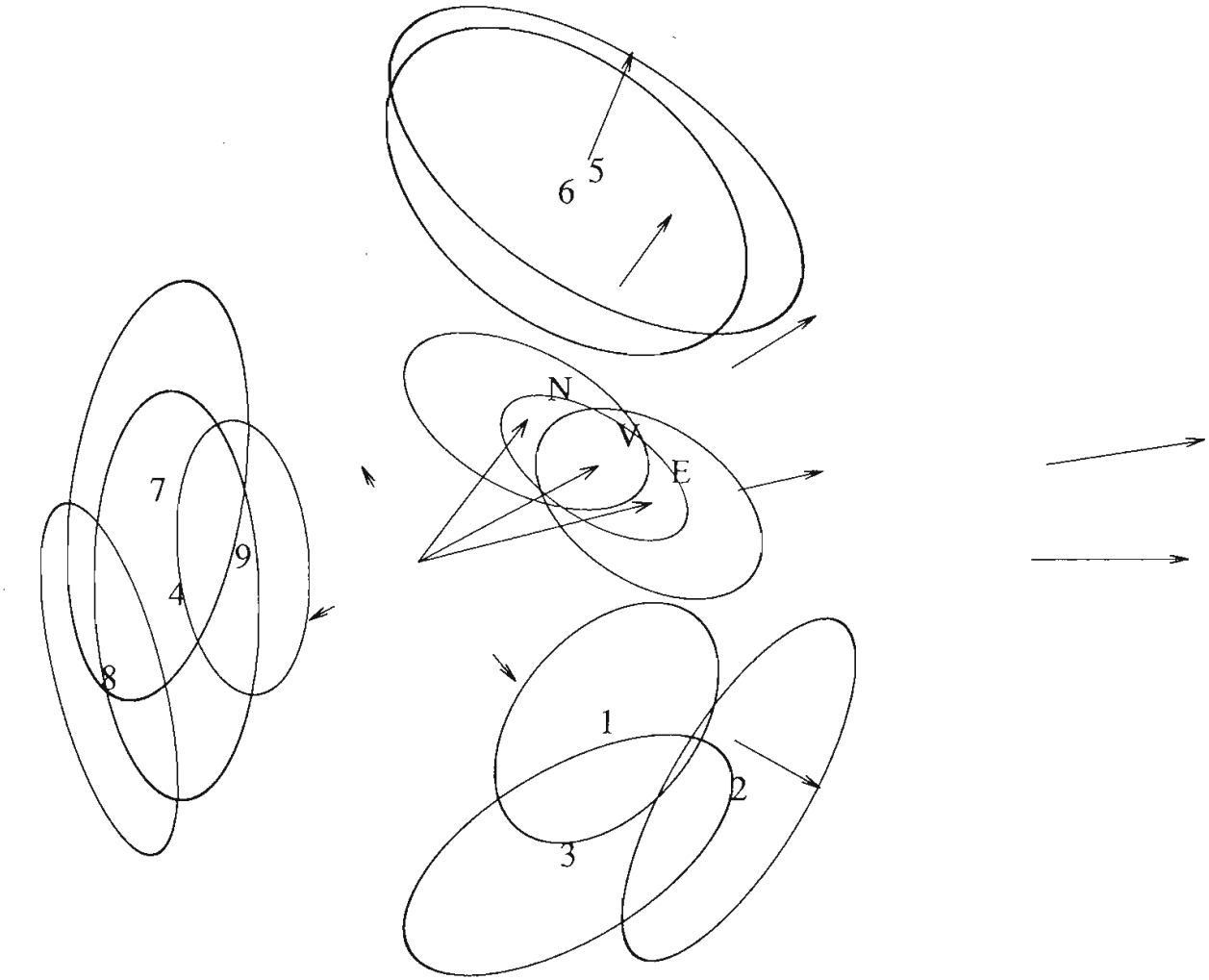
- A Vector Preference Model. Stimuli represented as points and subjects as vectors in a multidimensional space, and subjects' preferences are represented by projections of the stimulus points on the subject vectors. A special case of CPCA.
- $\mathbf{H}$ is a design matrix for pair comparison, e.g.,

$$\mathbf{H} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

- No $\mathbf{G}$.

University of Victoria

Derived Configuration



University of Victoria

Similarity Effects

- Stimulus points, and subject vectors (only for the first 10 subjects).
- Bootstrap confidence ellipsoids.
- Similar stimuli are easier to compare due to higher covariances, tending to cluster together.
- Post-hoc mapping of subject information (anglophones, non-anglophones).

University of Victoria

Finer Decompositions

- $\mathbf{G} = [\mathbf{X}, \mathbf{Y}]$.
- (1) $\mathbf{P}_G = \mathbf{P}_X + \mathbf{P}_Y \Longleftrightarrow \mathbf{X}'\mathbf{Y} = \mathbf{O}$.
- (2) $\mathbf{P}_G = \mathbf{P}_X + \mathbf{P}_Y - \mathbf{P}_X\mathbf{P}_Y \Longleftrightarrow \mathbf{P}_X\mathbf{P}_Y = \mathbf{P}_Y\mathbf{P}_X$.
- (3) $\mathbf{P}_G = \mathbf{P}_X + \mathbf{P}_{Q_X Y} = \mathbf{P}_Y + \mathbf{P}_{Q_Y X}$.
- (4) $\mathbf{P}_G = \mathbf{P}_{GA} + \mathbf{P}_{G(G'G)^{-1}B}$, where $\mathbf{B}$ is such that $\mathrm{Ker}(\mathbf{B}') = \mathrm{Sp}(\mathbf{A})$. (Let $\mathbf{C}$ represent the matix of regression coefficients. we impose constraints of the form $\mathbf{B}'\mathbf{C} = \mathbf{O}$, or equivalently $\mathbf{C} = \mathbf{A}\mathbf{C}^*$ for some $\mathbf{C}^*$, where $\mathbf{B}$ and/or $\mathbf{A}$ are known constraint matrices.)
- Terms on the right are mutually orthogonal.

University of Victoria

Example 2: Decomposition (2)

- $N = 100$ rectangles (10 levels of height and 10 levels of width).
- 4 groups (1st, 3rd, 5th and 7th graders) of approximately 40 children in each group were asked to judge if the rectangles looked large or small, and the numbers of times the rectangles were judged large were used as "ordinal" measures of the subjective areas of the rectangles.
- $\mathbf{Z}$ is a 100 rectangles by 4 age group matrix. The "ordinal" measures mean that the best monotonic transformations are sought for and applied to columns of the data matrix.
- Centering of the transformed data eliminates the need for the third term in the model.

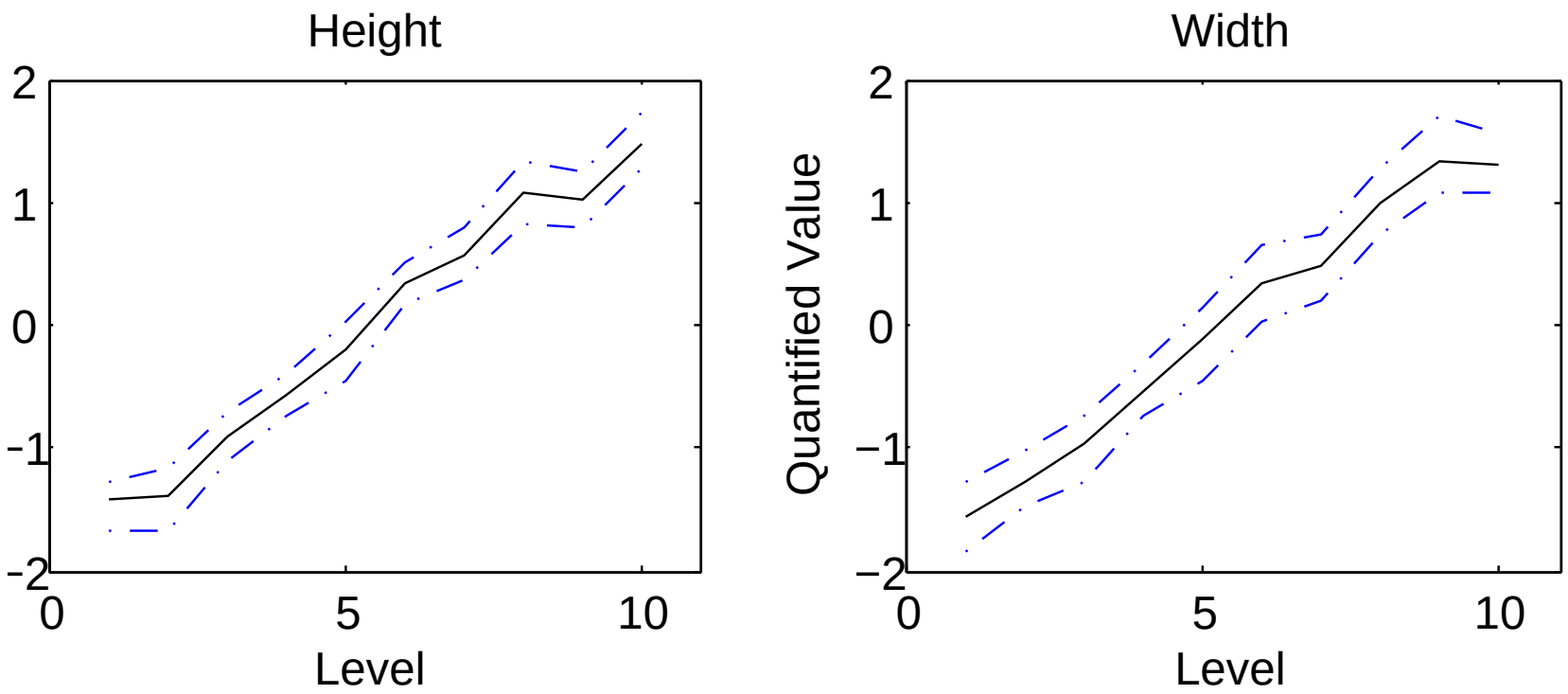
University of Victoria

Weighted Additive Model (WAM)

- $\mathbf{Z}^*$: Monotonically transformed data.
- $\mathbf{G}_H$, $\mathbf{G}_W$: Design matrices indicating the levels of height and width, respectively, of the rectangles.
- $\mathbf{u}_H$, $\mathbf{u}_W$: Effects of the levels of the height and width factors on area judgments.
- $\mathbf{v}_H$, $\mathbf{v}_W$: Weights attached to the height and width dimensions by the 4 age groups.
- Model: $\mathbf{Z}^* = \mathbf{G}_H\mathbf{u}_H\mathbf{v}'_H + \mathbf{G}_W\mathbf{u}_W\mathbf{v}'_W + \mathbf{E}$.
- A special case of the Finer Decompositions: (2). Two-way ANOVA without interactions. Centering the data eliminates the need for the third term in (2).

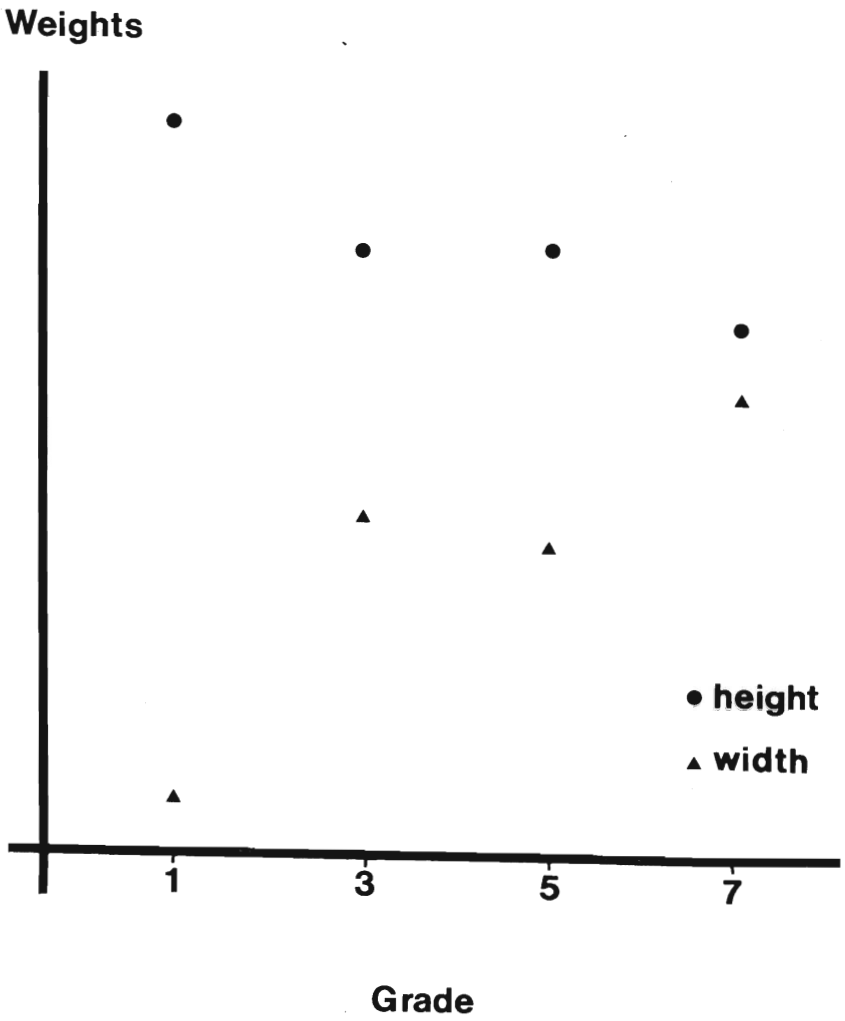
University of Victoria

Plot of Quantified Effects: $\mathbf{u}_H$ and $\mathbf{u}_W$



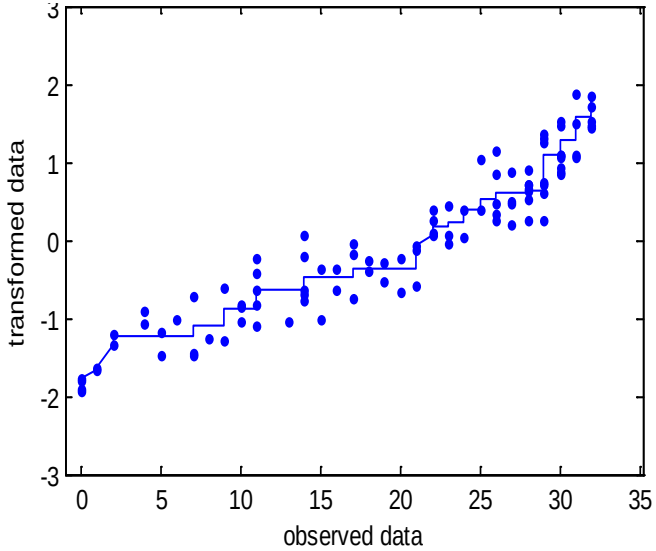
University of Victoria

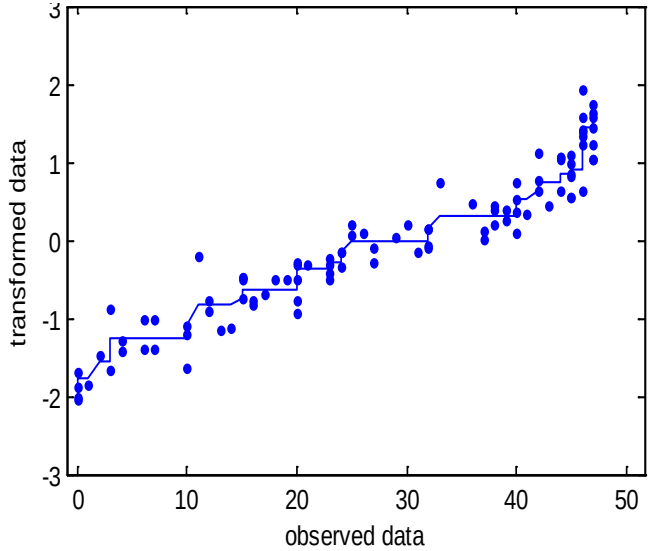
Plot of Weights: $\mathbf{v}_H$ and $\mathbf{v}_W$

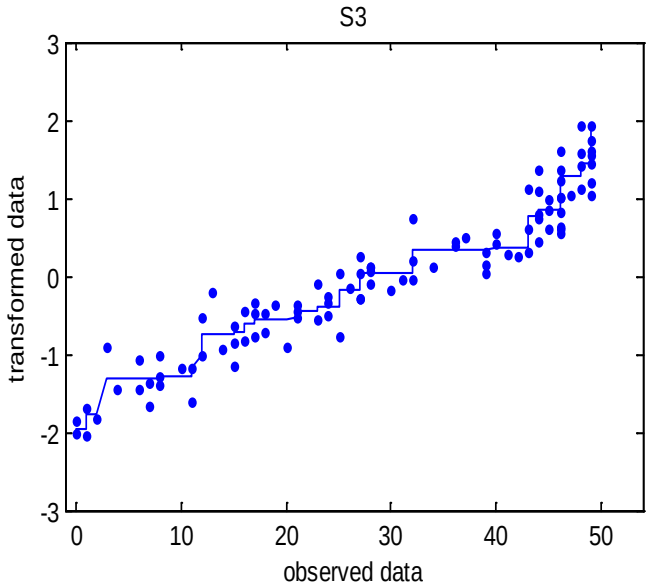


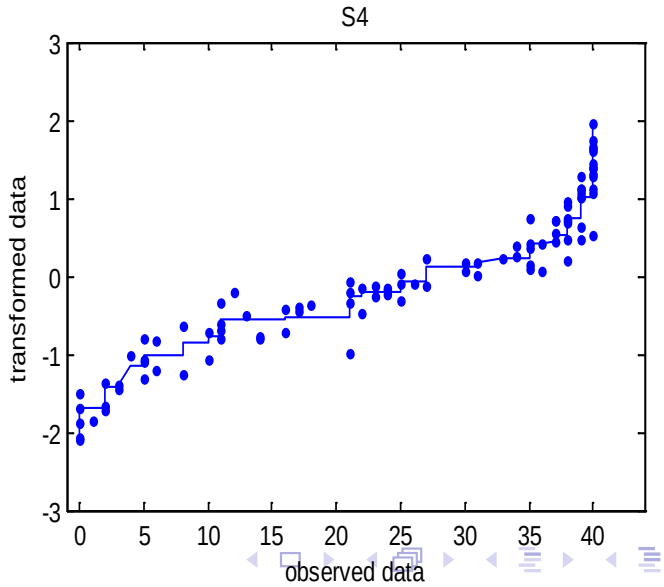
University of Victoria

Plot of Optimal Monotonic Transformations of Data









University of Victoria

Some Extensions (1)

- Nonidentity metric matrices ($\Leftarrow$ Weighted least squares estimation)
- A nonnegative definite matrix $\mathbf{K}$ such that $\text{rank}(\mathbf{KG}) = \text{rank}(\mathbf{G})$.
- $\mathbf{P}_{G/K} = \mathbf{G}(\mathbf{G}'\mathbf{KG})^{-1}\mathbf{G}'\mathbf{K}$, and $\mathbf{Q}_{G/K} = \mathbf{I} - \mathbf{P}_{G/K}$.
- $\mathbf{P}_{G/K}^2 = \mathbf{P}_{G/K}$, $\mathbf{Q}_{G/K}^2 = \mathbf{Q}_{G/K}$ (idempotent)
- $(\mathbf{KP}_{G/K})' = \mathbf{KP}_{G/K}$, $(\mathbf{KQ}_{G/K})' = \mathbf{KQ}_{G/K}$ (K-symmetric).
- $\mathbf{P}'_{G/K}\mathbf{KQ}_{G/K} = \mathbf{Q}'_{G/K}\mathbf{KP}_{G/K} = \mathbf{O}$ (K-orthogonal).
- Analogous for the $\mathbf{H}$ side projectors.

University of Victoria

Some Extensions (2)

- Oblique projectors ($\Leftarrow$ Instrumental variable estimation).
- Ridge operators ($\Leftarrow$ Ridge least squares estimation).
- Components restricted to be inside Sp($\mathbf{Z}$). Set $\mathbf{H} = \mathbf{Z}'\mathbf{G}$.
- Nonorthogonal decompositions $\Rightarrow$ DCDD (Different Constraints on Different Dimensions), GSCA (Generalized Structured Component Analysis).

University of Victoria

Software

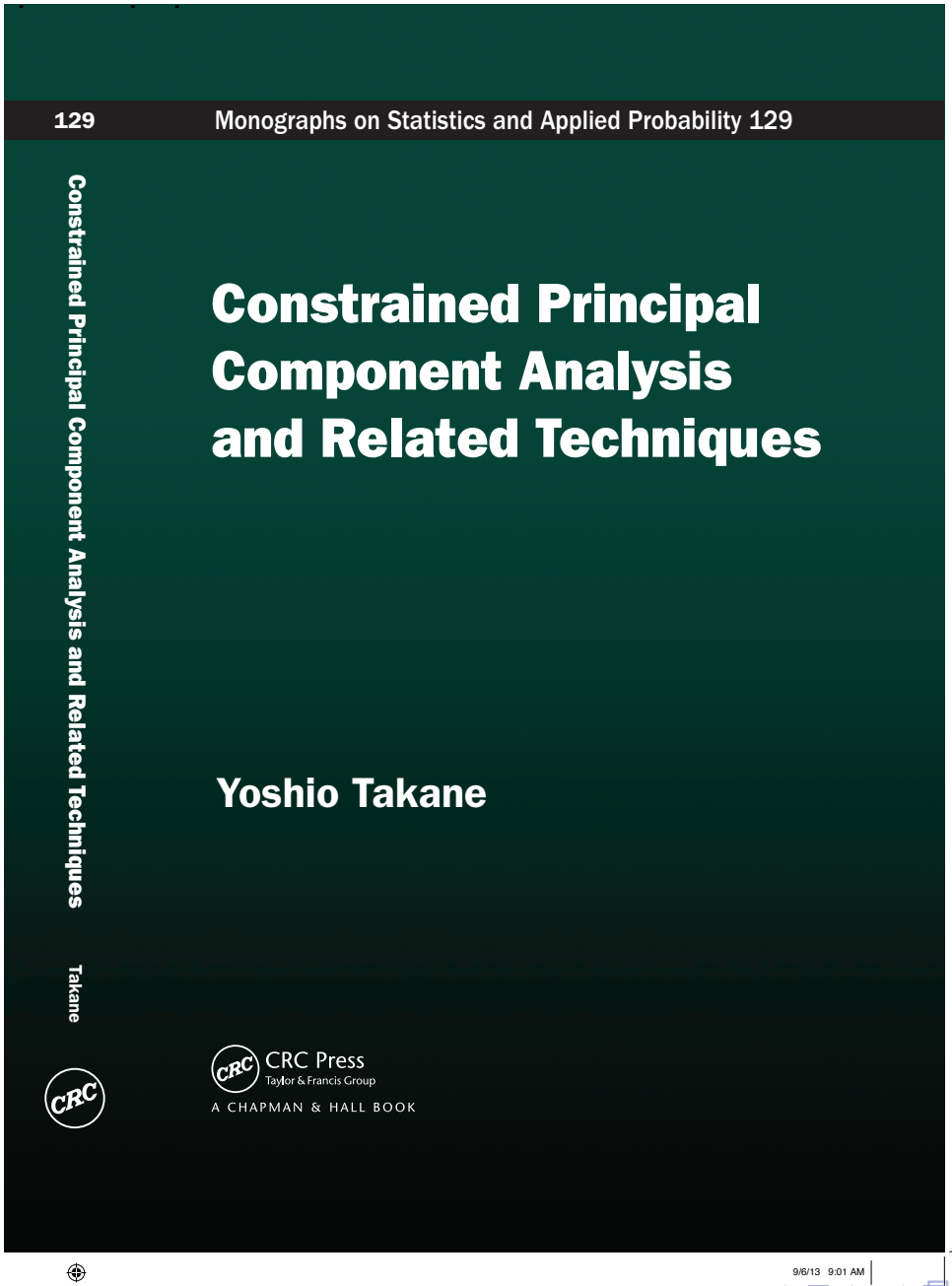
- Construction of my own webpage has been delayed, but is coming; Matlab and Fortran.
- Professor Todd Woodward at UBC (http://www..nitrc.org/projects/fmricpca).
- Professor Heungsun Hwang at McGill (http//www.sem-gsca.org/) for GeSCA.
- Professor Herve Abdi at UT at Dallas. R programs.

University of Victoria

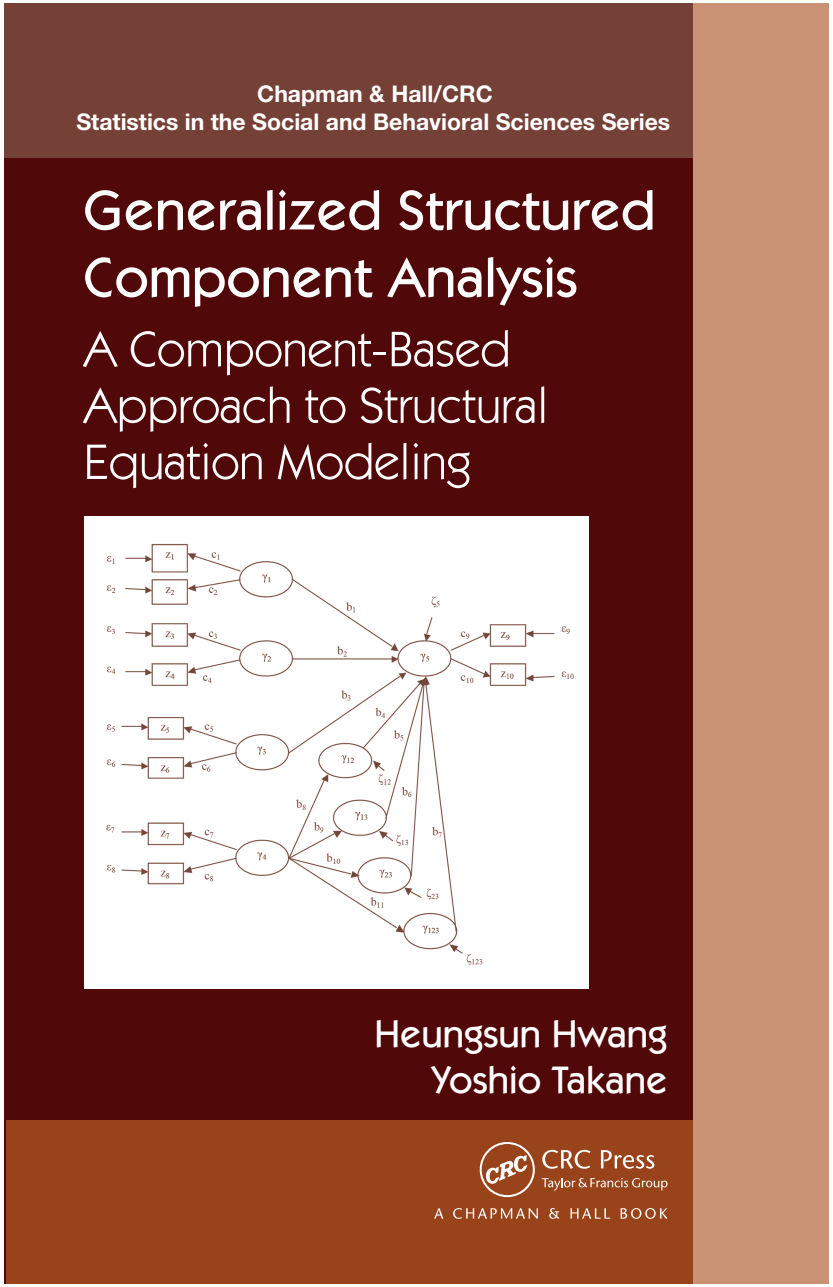
Books

- Yanai, H., Takeuchi, K., and Takane, Y. (2011). *Projection matrices, generalized inverse matrices, and singular value decomposition*. New York: Springer.
- Takane, Y. (2013). *Constrained principal component analysis and related techniques*. Boca Raton, FL: Chapman and Hall/CRC Press. (Promo Code: EZL20 for 20% discount)
- Hwang, H., and Takane, Y. (2014). *Generalized structured component analysis: A component-based approach to structural equation modeling*. Boca Raton, FL: Chapman and Hall/CRC
  Press.

University of Victoria



University of Victoria



University of Victoria