

Quantification of Curves' Variation and Simplicity to Identify Genetic Constraints

Travis L. Gaydos

Joel G. Kingsolver

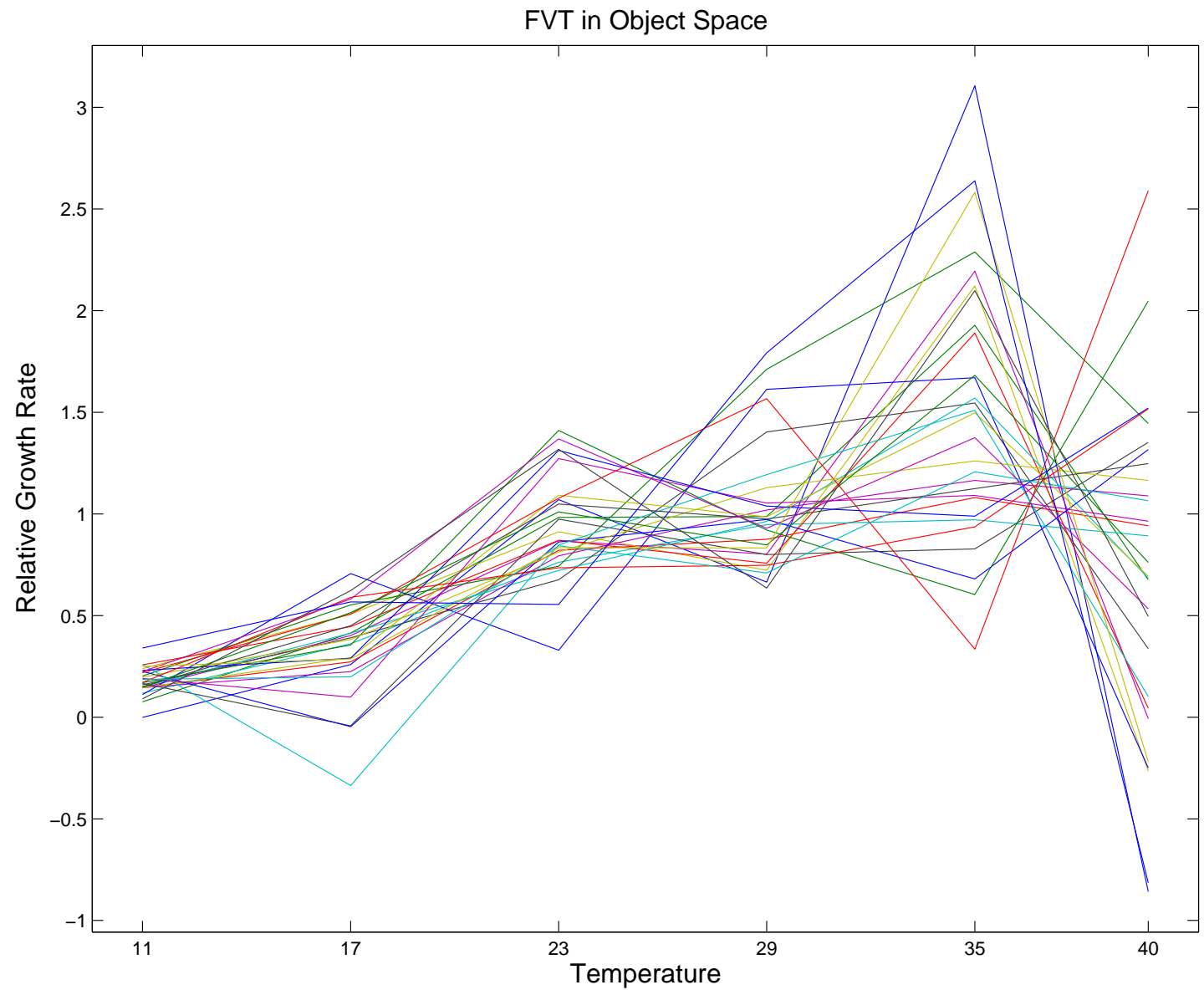
J. S. Marron

UNC - Chapel Hill

Phenotype and FVT

- Phenotype – an observable characteristic of an individual
- Measure phenotype over environment levels
- FVT – Phenotype is continuous w.r.t. environment levels
- Examples
 - Caterpillar growth rate as funct. of temp.
 - height of plant as function of age

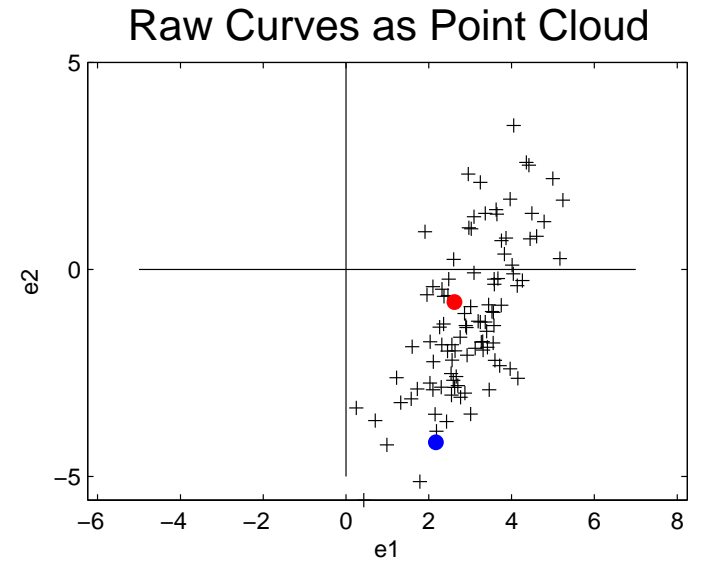
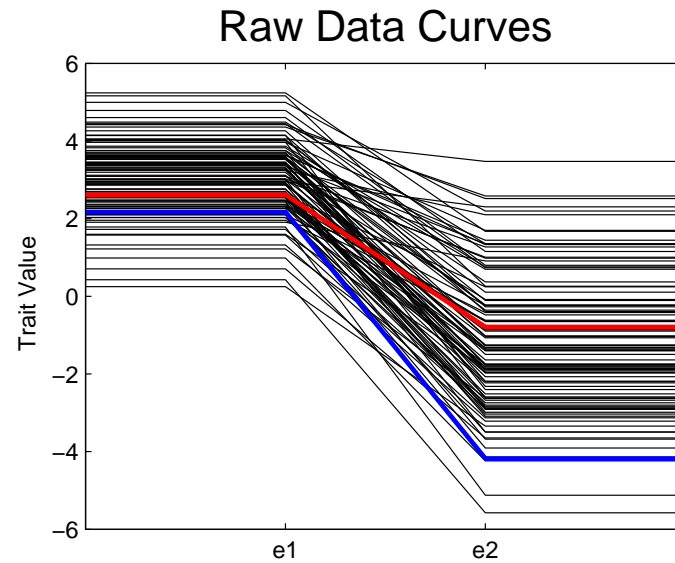
Pieris rapae Caterpillar FVT



Functional Data Analysis

- FDA: Curve = Statistical Atom
- Object Space = Curves
- Discretized Curves = Vector
- Vector $\subseteq \mathbb{R}^d$ = Point Cloud Space

Object – Pt Cloud (Highlight)

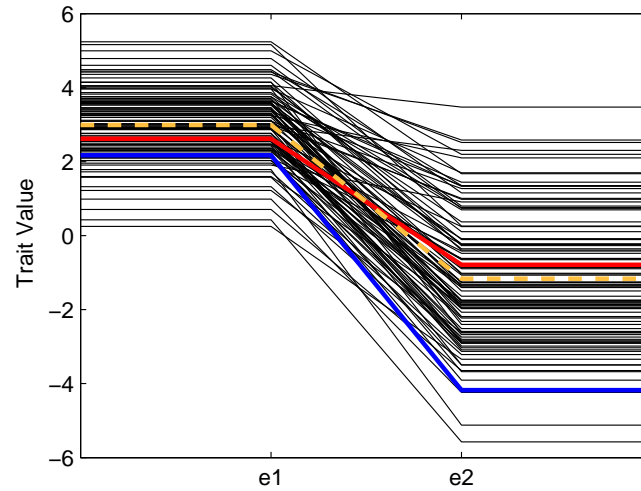


Pt Cloud Space

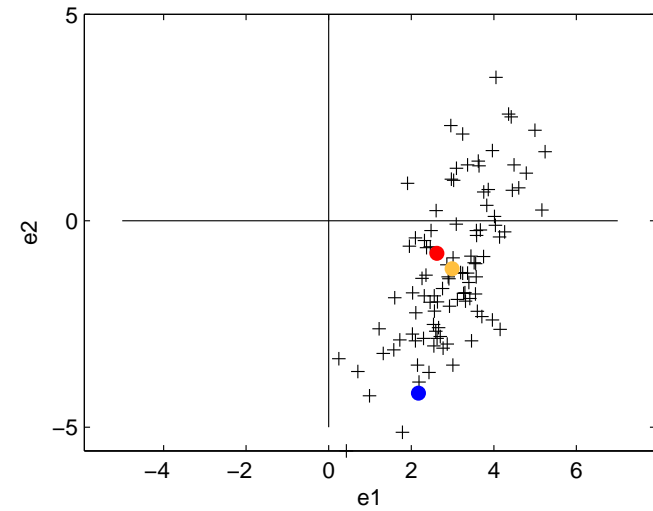
- Point Cloud Space – Statistical Analysis
- Object Space – Understand Results
- Any point can be viewed in Object Space
- Due to Vector Structure
- Example Analysis: Mean and Centering

Object – Pt Cloud (Centered)

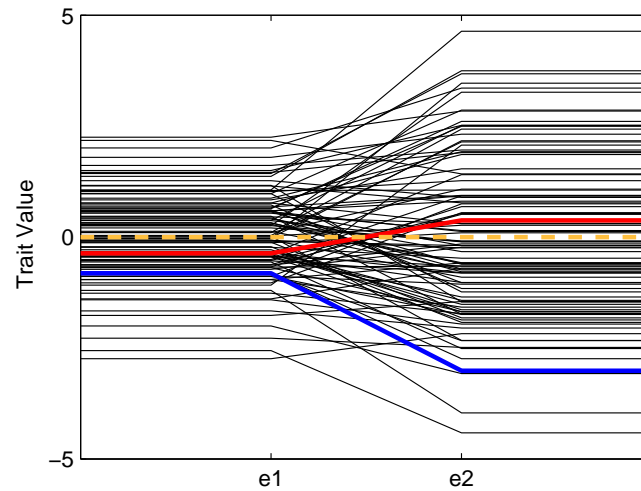
Raw Data Curves



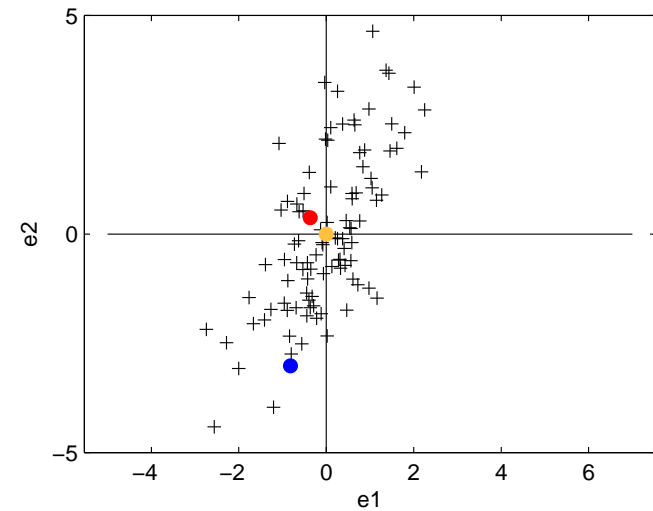
Raw Curves as Point Cloud



Centered Raw Data Curves



Centered Raw Curves as Point Cloud

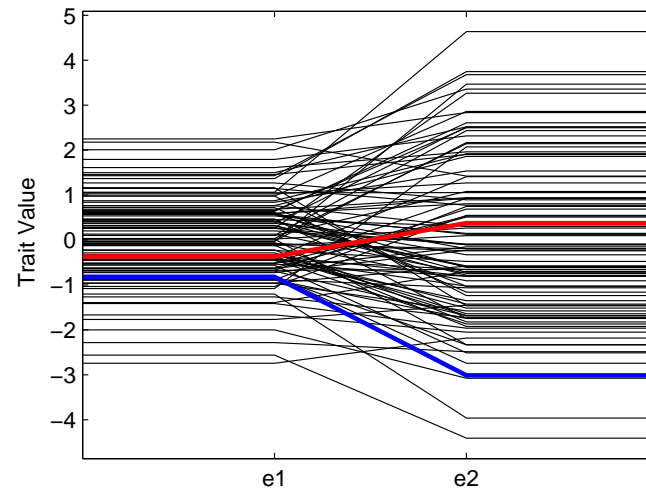


Principal Components Analysis

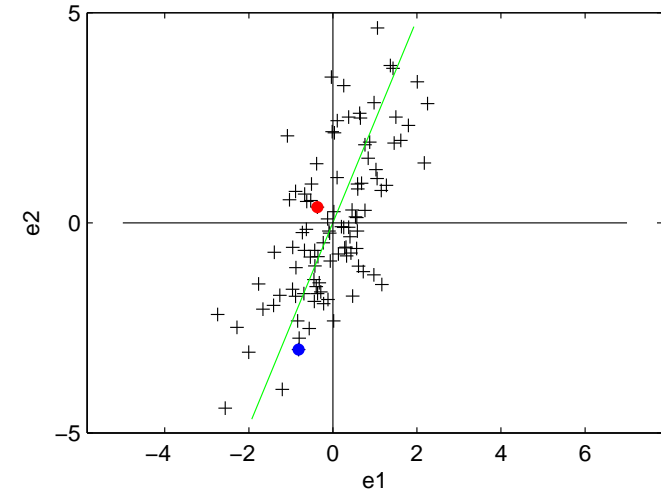
- After Center – Variation from Mean
- PCA: Useful Tool
- \perp directions of maximal variation
- PC 1 = Best 1-d Representation
- PC 2 = 2nd Best 1-d Representation
- etc
- Found by Eigendecomposition of Σ

PC in Object – Pt Cloud

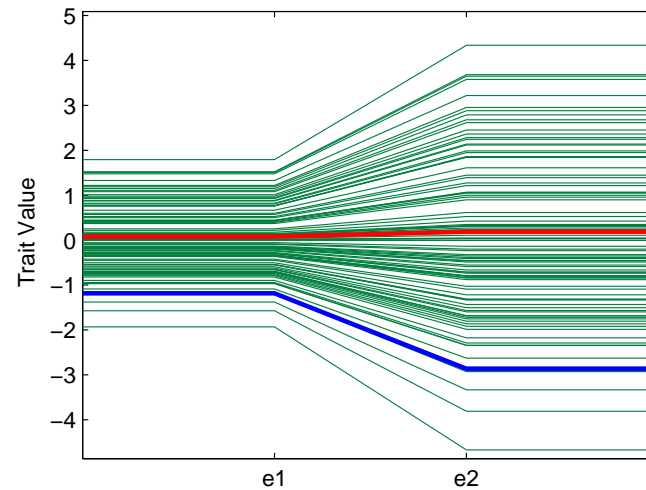
Centered Raw Data Curves



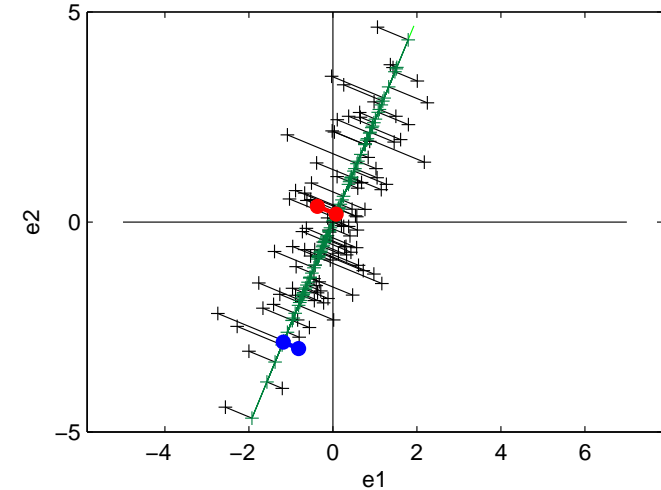
Centered Raw Curves as Point Cloud



Projection of Centered Curves on PC1



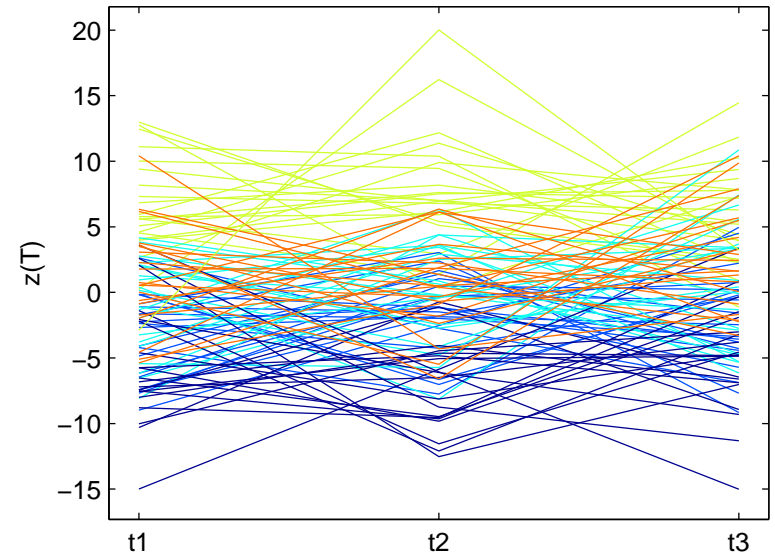
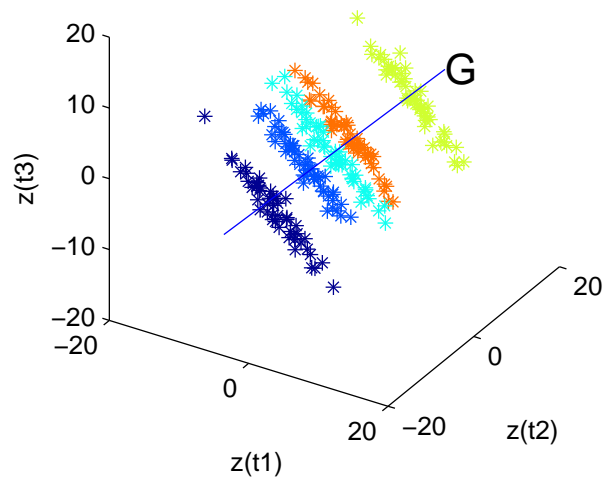
Projection of Points onto PC1



Evolutionary Biology

- Phenotypic Variation – Total
- Genetic Variation – Between Clones
- Environmental Variation – Within Clones
- $P = G + E$
- Random Effects ANOVA analysis

3-D Toy G



Evol. Response and G

- Evol Response = change from one gen to next
- Evol Response = Amount of Genetic Var.
- $\Delta\bar{z} = G\beta$
- 1st PC of G = direction of most evol response

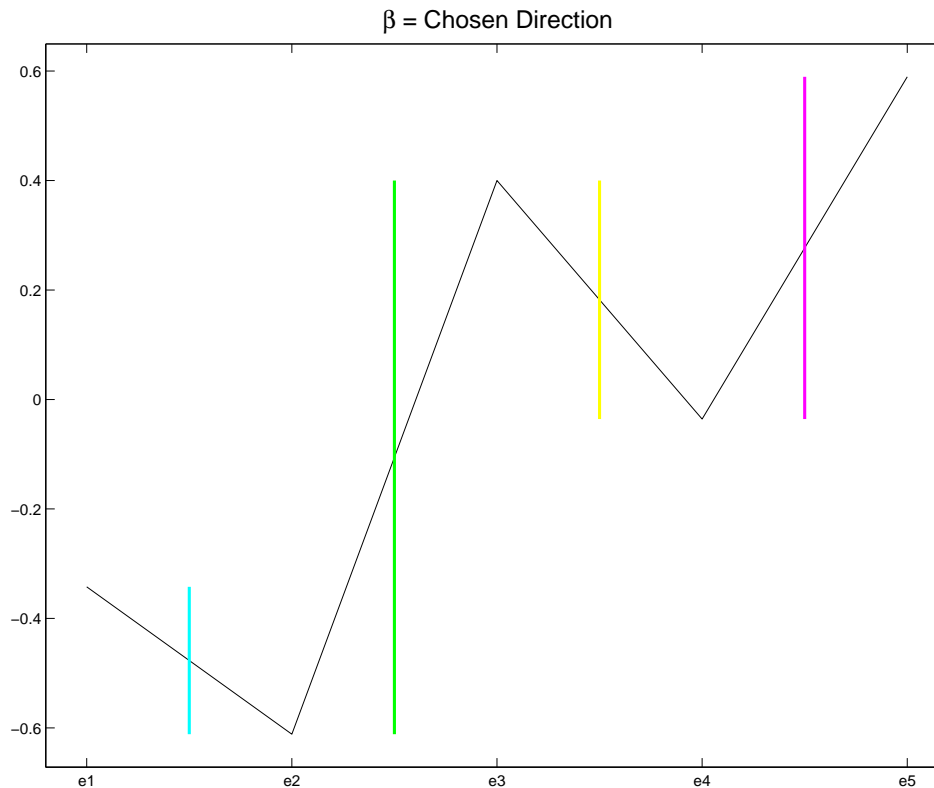
Genetic Constraints

- Dir. of small Evol Response
- Dir. of with little genetic var
- Corresponds to lower PCs

Interesting Genetic Constraints

- Dir of Low Gen Var.
- Dir with Interpretable Curves
- Interpretability = Simplicity
- small change in trait value for adjacent environment levels

Measuring Simplicity - Object Space



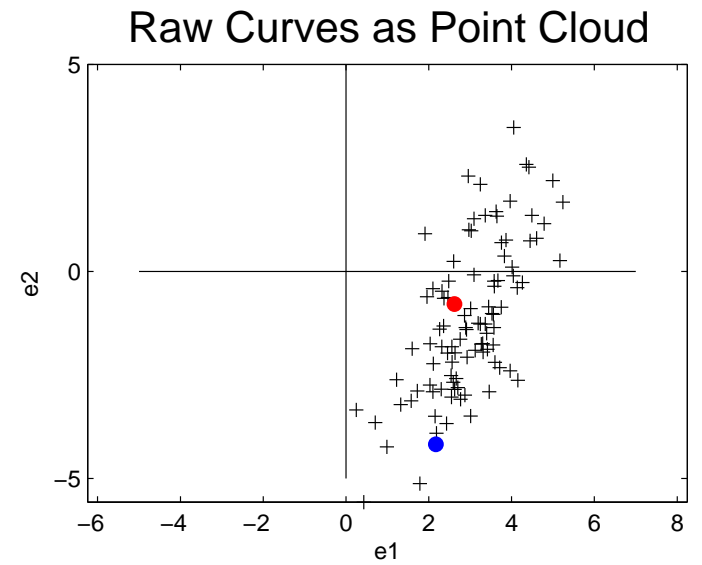
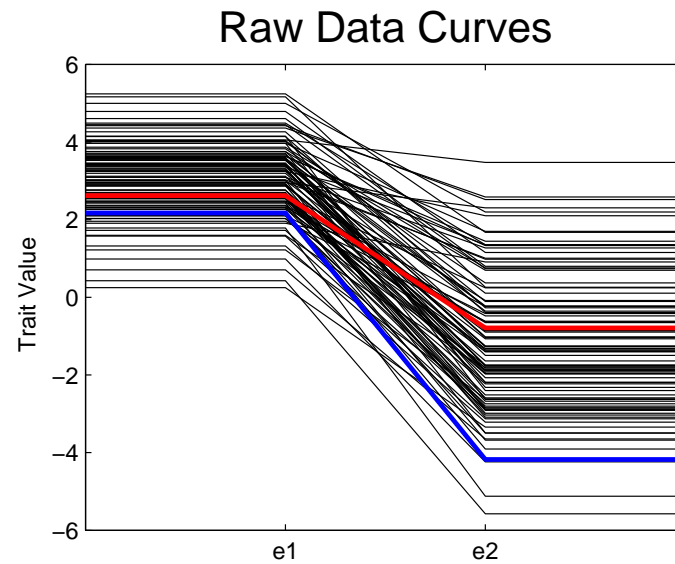
- $(\beta(t_2) - \beta(t_1))^2 + \dots + (\beta(t_5) - \beta(t_4))^2$
- $(\beta^T D)(\beta^T D)^T = \beta^T (DD^T)\beta$

Understanding D

$$D = \begin{pmatrix} -1 & 0 & 0 & \dots & 0 \\ 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & \ddots & & \vdots \\ 0 & \dots & 0 & -1 & 0 \\ 0 & \dots & 0 & 1 & -1 \\ 0 & \dots & 0 & 0 & 1 \end{pmatrix}$$

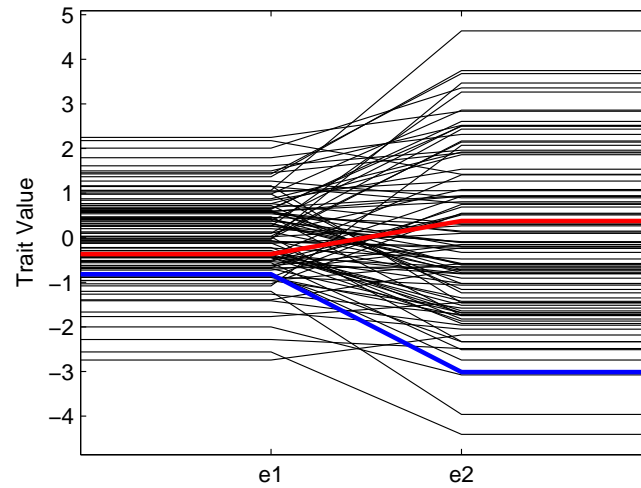
- $d \times (d - 1)$ difference matrix

Object – Pt Cloud (Highlight)

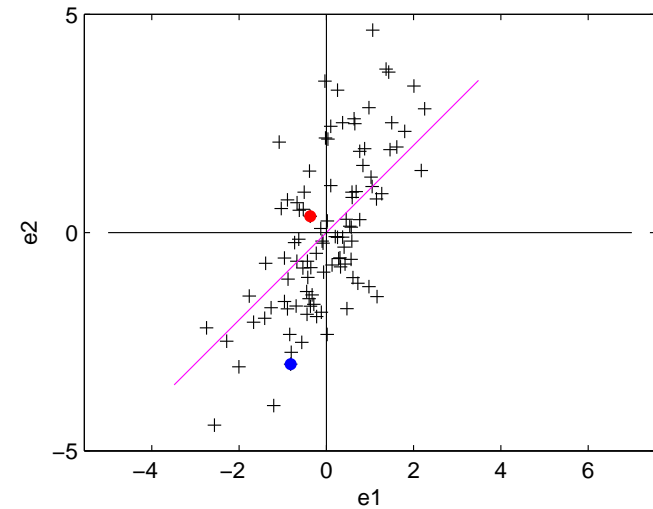


Toy Example – Simple Direction

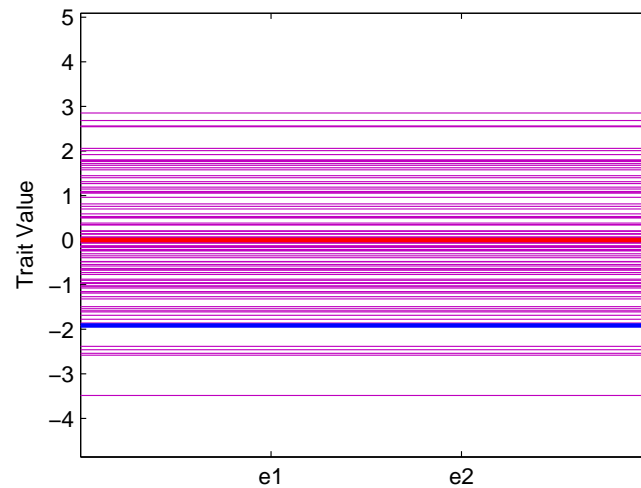
Centered Raw Data Curves



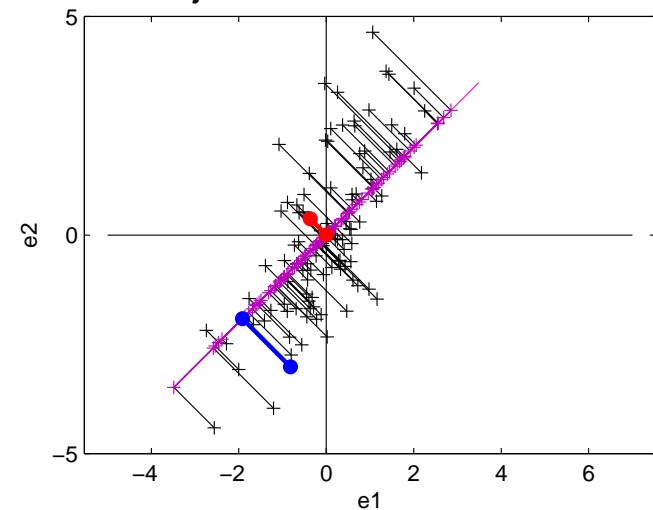
Centered Raw Curves as Point Cloud



Projection of Centered Curves on S1



Projection of Points onto S1



Var-Simplicity View of Candidate Dir.

- $\beta =$ Candidate direction
- $\beta^T \beta = 1$
- $\beta^T \mathbf{G} \beta =$ Evolutionary Response (var.)
- $4 - \beta^T (DD^T) \beta = \beta^T (4I_d - DD^T) \beta =$ simplicity score (Interp.)
- Interesting Gen Const
 - Small Var
 - Large Simp Score

Nearly Null Space

- Find Subspace of Gen Constraints
- PCA of G = Orders dir by Amount of Gen. Var.
- Large PC – Large Evol Response (Model)
- Small PC – genetic constraints (Null)
- Find Basis of Nearly Null Space

PCA basis of Nearly Null Space

- Weak biology signal
- Biology Signal Mixes With Noise
- Hard to Interpret

Simple Curve (SC) Basis

- Null Space PCA basis – Hard to Interpret
- Choose basis based on interpretability
- Order \perp dir by Simplicity Score
- Nancy Heckman, Mark Kirkpatrick

Method to Find SC basis

- B = basis of Null Space
- D = Difference Matrix
- $F = B^T(4I_d - DD^T)B$ = Simplicity Matrix
- Eigenanalysis of F leads to Simple Curve Basis

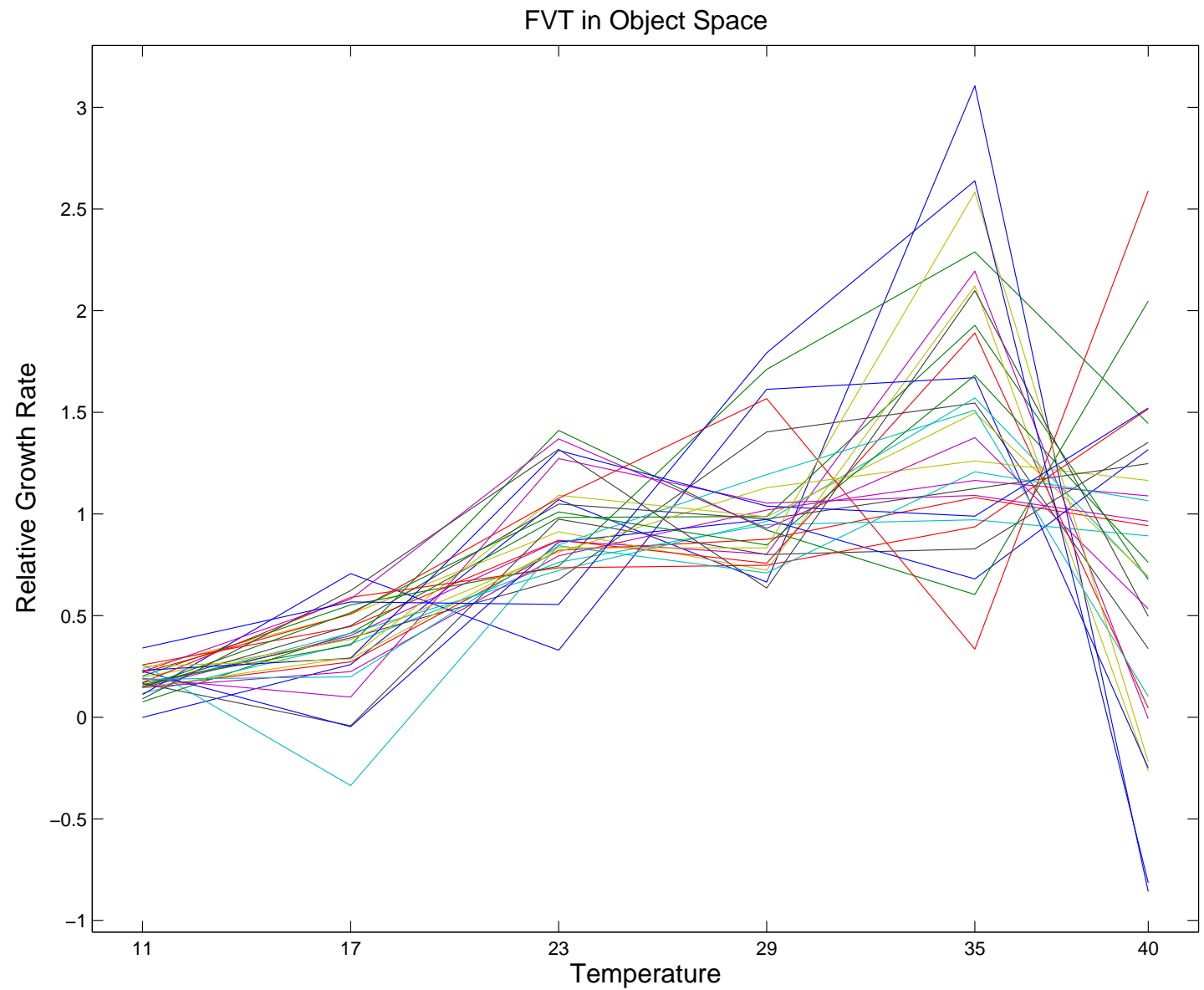
Understanding F

- $B = [\beta_1, \beta_2]$
- $D = \text{Difference Matrix}$

$$F = \begin{pmatrix} \beta_1^T (4I_d - DD^T)\beta_1 & \beta_1^T (4I_d - DD^T)\beta_2 \\ \beta_2^T (4I_d - DD^T)\beta_1 & \beta_2^T (4I_d - DD^T)\beta_2 \end{pmatrix}$$

- with eig vec e_i^F corresponding to eig val λ_i^F
- e_i^F linear combination of β_1 and β_2
- $S_1 = Be_1^F = \text{simplest direction in subspace}$
 - e_1^F eig vec corresponding to largest λ_i^F

Pieris rapae Caterpillar FVT



Pieris rapae Caterpillar

- Rel growth rate at 6 temp
 - (11, 17, 23, 29, 35, 40)
- Weight matrix – unevenly spaced index points
- G is estimated from pheno. curves by Restricted Maximum Likelihood

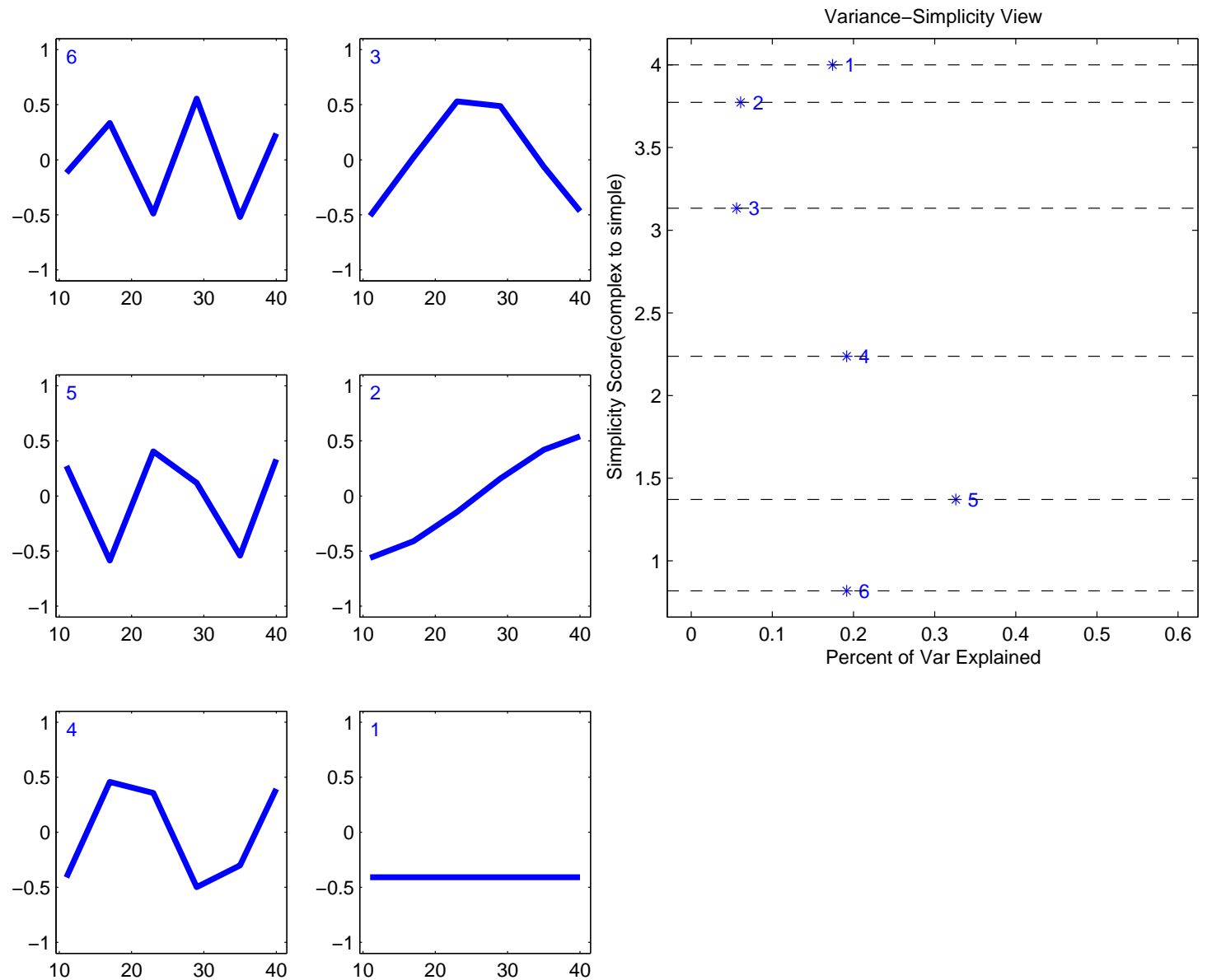
Finding Gen. Constraints

- Goal: Find Interesting Gen. Constraint
- Choose basis
- view genetic var explained
- view simplicity score

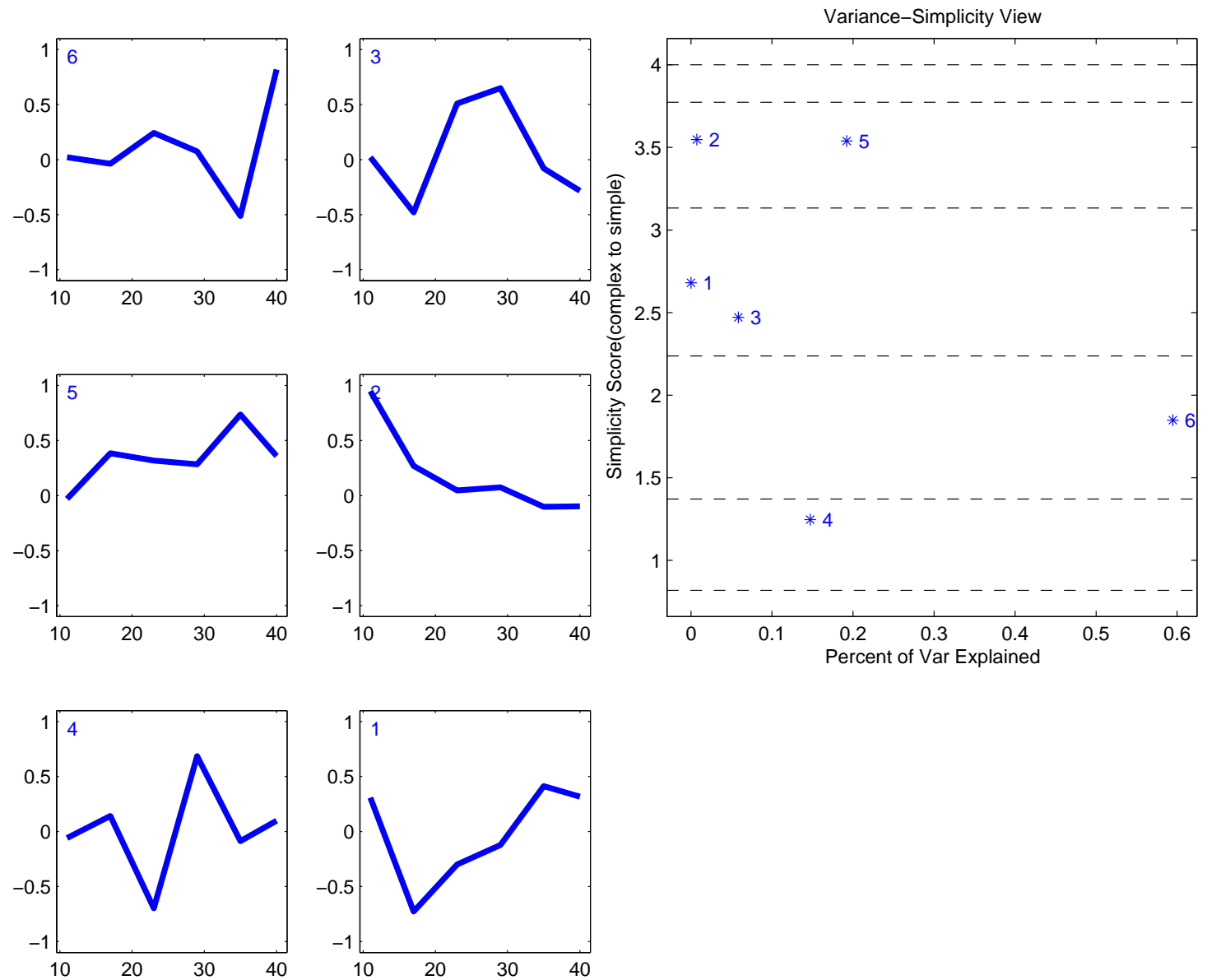
3 Bases

- Basis of simplest directions possible
- Basis of PC directions
- Simple Curve basis of Nearly Null Space

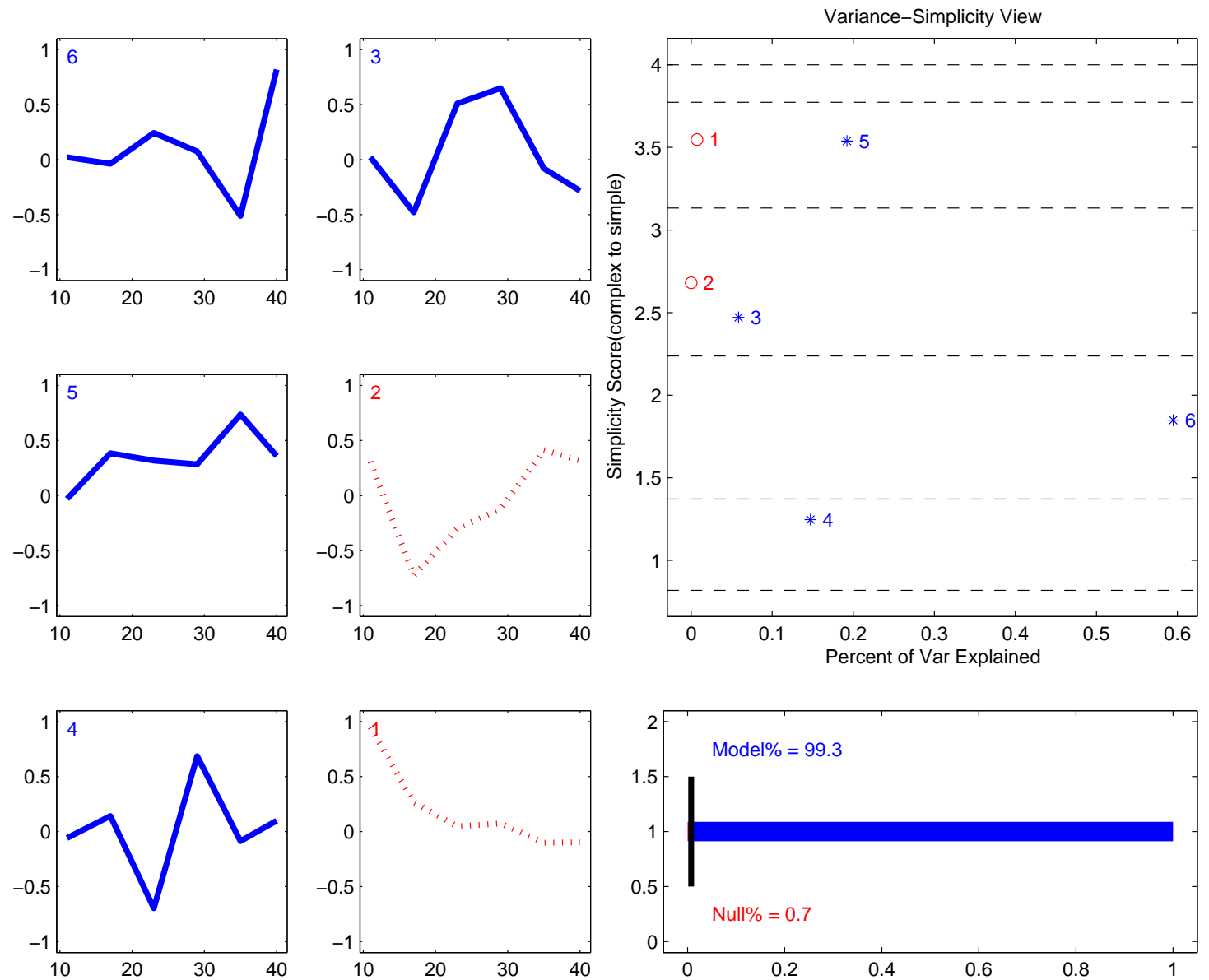
Simple Curve Basis of All Dir's



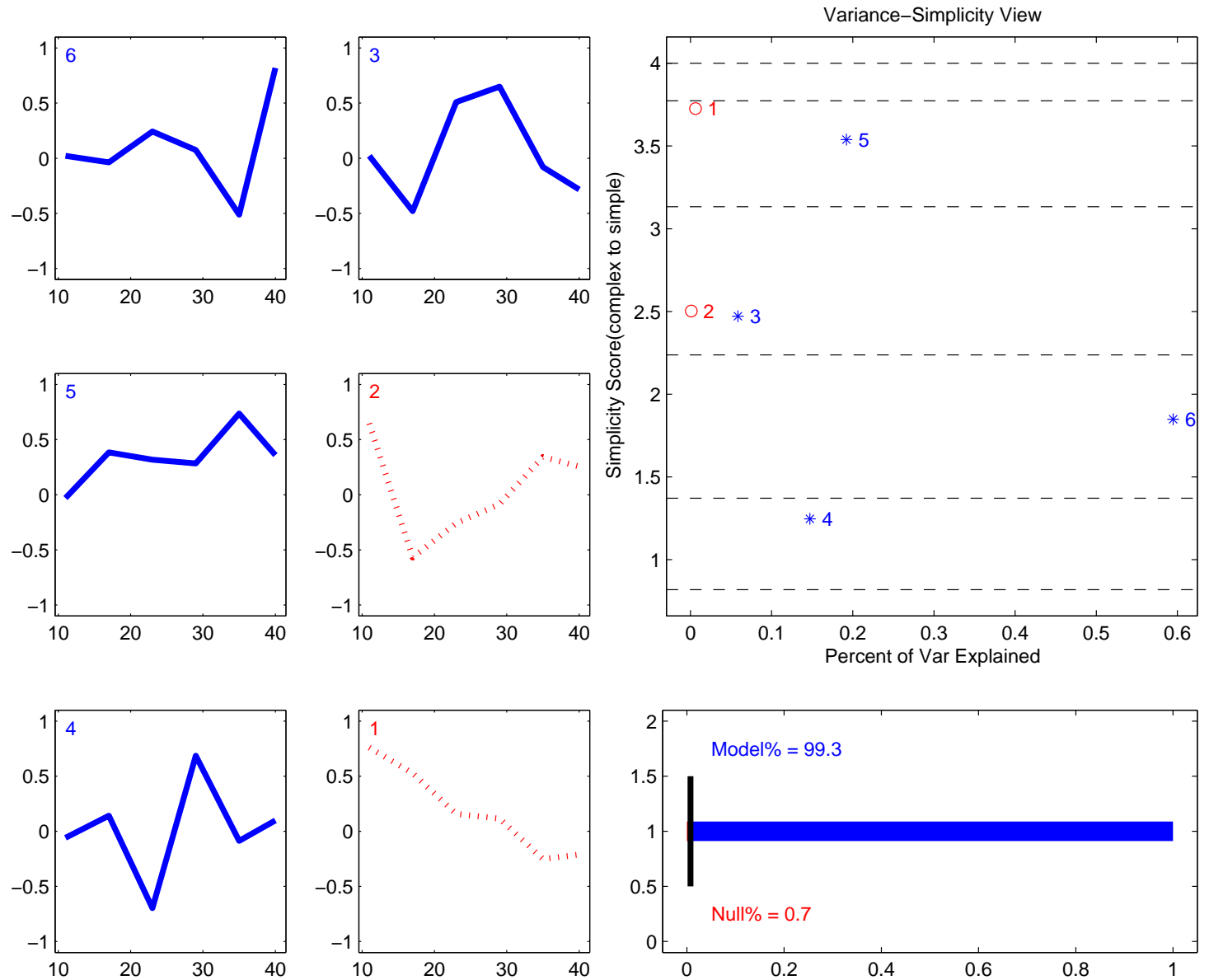
PC basis



Nearly Null Space Defined



Simple Curve Basis of Null Space



Simple Curve Basis – Summary

- Use var and simplicity score
- Two eigenanalyses produce simple curve basis
- Best Chance for Interesting Gen. Constraints
- Better Chance of Being Stable

Questions

- Questions
- Comments

References

- Ramsay J.O. and Silverman B.W. (2002). *Applied Functional Data Analysis: methods and case studies*. Springer Series in Statistics. Springer, New York
- Anderson T.W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York
- Tyler D.E. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics* **9**, 725–736
- Tyler D.E. (1983). A class of asymptotic tests for principal component vectors. *The Annals of Statistics* **11**, 1243–1250

References

- Stewart G.W. and Sun J.g. (1990). *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press, San Diego
- Searle S.R., Casella G. and McCulloch C.E. (1992). *Variance Components*. Wiley Series in Probability and Statistics. Wiley, New York
- van der Vaart A.W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York
- Kingsolver J.G., Ragland G.J. and Shlichta J.G. (2004). Quantitative genetics of continuous reaction norms: Thermal sensitivity of caterpillar growth rates. *Evolution* **58**, 1521–1529

Mathematical Setting

- $X = [X_1, \dots, X_n]$ sequence of i.i.d. Normal d -dim vectors
- X_i is assumed to have a covariance matrix Σ
- Σ has eigenvalues
 $\lambda_1 \geq \dots \geq \lambda_m > \lambda_{m+1} = \dots = \lambda_d$
- $\hat{\Sigma} = \frac{1}{n-1} (X - \bar{X})(X - \bar{X})^T$
 - with estimated eigenvalues $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_d$
 - $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- Study Asymptotic properties as $n \rightarrow \infty$

Definitions of Nearly Null Space

- N_{prop} = subspace generated by eig vec's s.t. $\frac{\sum_{i=m+1}^d \lambda_i}{\sum_{i=1}^d \lambda_i} < C_{prop}$
- \hat{N}_{prop} = subspace generated by est. eig vec's s.t. $\frac{\sum_{i=m+1}^d \hat{\lambda}_i}{\sum_{i=1}^d \hat{\lambda}_i} < C_{prop}$
- N_{thresh} = subspace generated by eig vec's s.t. $\lambda_i < C_{thresh}$
- \hat{N}_{thresh} = subspace generated by est. eig vec's s.t. $\hat{\lambda}_i < C_{thresh}$

Asymptotic properties

- Estimates Converge in Probability to True as $n \rightarrow \infty$
- Convergence shown using Metrics Between Subspaces
- Metrics based on Sines of Canonical Angles
- See Stewart and Sun 1990 for more details

Conditions of Convergence

Theorem 1 Given the definition of N_{thresh} , \hat{N}_{thresh} it follows that for $\forall \epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(D_{sine}(N_{thresh}, \hat{N}_{thresh}) > \epsilon) = 0$$

Theorem 2 Given the definition of N_{thresh} , \hat{N}_{thresh} it follows that for $\forall \epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(D_{gap}(N_{thresh}, \hat{N}_{thresh}) > \epsilon) = 0$$

Conditions of Convergence

Theorem 3 Given the definition of N_{prop} , \hat{N}_{prop} it follows that for $\forall \epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(D_{sine}(N_{prop}, \hat{N}_{prop}) > \epsilon) = 0$$

Theorem 4 Given the definition of N_{prop} , \hat{N}_{prop} it follows that for $\forall \epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(D_{gap}(N_{prop}, \hat{N}_{prop}) > \epsilon) = 0$$

Interesting Genetic Constraint Space

- Subspace Defined by Directions Of SC basis
- Similar Results can be shown
- Based on Eigenvalues and Eigendirections of F