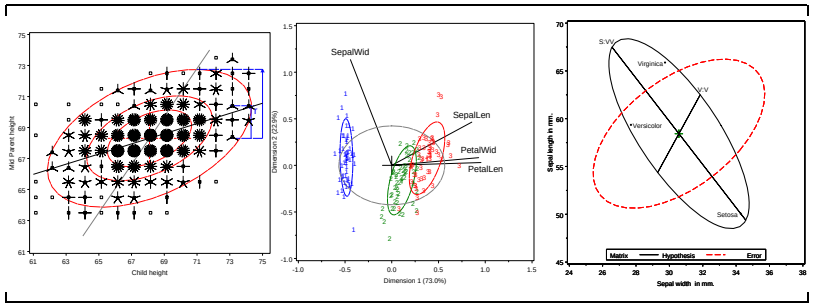
Data ellipses, HE plots and Reduced-rank displays for MLMs: SAS software & examples



Michael Friendly

York University, <friendly@yorku.ca>

November, 2006

Outline

- **Introduction**
- **Bivariate plots for multivariate data**
  - Data ellipse
  - Simpson's paradox: marginal vs. conditional relations
  - Partial plots
  - Robust data ellipse and outlier detection
- **Biplots**: reduced-rank displays
- **HE plots**
  - HE plot matrices
  - Showing contrasts
  - HE plots for MMRA
- **Canonical discriminant plots**: Reduced-rank HE plots

Introduction: The LM family and friends

- **LM family**
  - Classical univariate models: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$,
  - $\longrightarrow$ regression, ANOVA, ANCOVA, response surface models, ...
  - Many graphical methods: effect plots, spread-level plots, influence plots, ...
- **GLM family**
  - Generalized univariate models: $g(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with general variance function
  - $\longrightarrow$ poisson regression, logistic regression, loglinear models
  - Some graphical methods: effect plots, fourfold plots, mosaic plots, half-normal plots, ...
- **MLM family**
  - Classical multivariate models: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$, with $\boldsymbol{E} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I} \otimes \boldsymbol{\Sigma})$,
  - $\longrightarrow$ MANOVA, MMRA, MANCOVA, ...
  - Graphical methods: ???
- **MGLM family**: ??

SAS macros

Available at `http://www.math.yorku.ca/SCS/sasmac/`:

`biplot`	Generalized biplot display of variables and observations
`canplot`	Canonical discriminant structure plots
`ellipses`	Plot data ellipses (renamed from `contour` macro)
`heplot`	Plot H and E matrices for a bivariate MANOVA effect
`hemat`	HE plots for all pairs of response variables
`hemreg`	Extract H and E matrices for multivariate regression
`panels`	Display a set of plots in a rectangular layout
`outlier`	Robust multivariate outlier detection
`robcov`	Calculate robust covariance matrix via MCD or MVE
`scatmat`	Scatterplot matrices

Macros and named examples: `http://euclid.psych.yorku.ca/SCS/Papers/Private/hesoft-sas.zip`

The Data Ellipse

- **Visual summary for bivariate marginal relations**
  - Shows: means, standard deviations, correlation, regression line(s)
  - Defined: set of points whose squared Mahalanobis distance $\leq c^2$,

$$D^2(\boldsymbol{y}) \equiv (\boldsymbol{y} - \bar{\boldsymbol{y}})^{\mathsf{T}} \, \boldsymbol{S}^{-1} \, (\boldsymbol{y} - \bar{\boldsymbol{y}}) \leq c^2$$

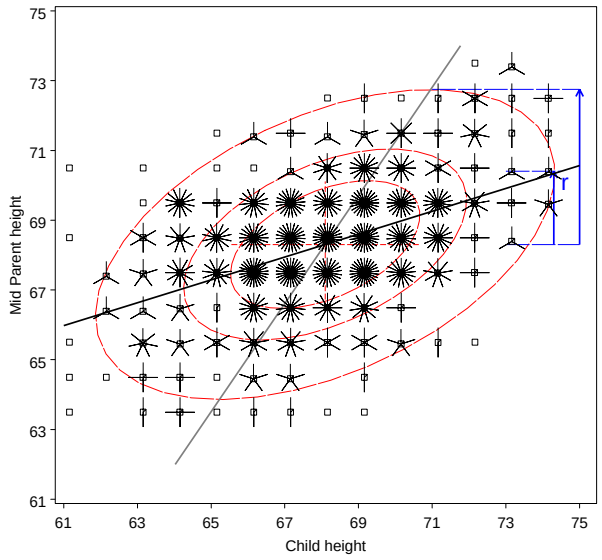
$\boldsymbol{S} =$ sample variance-covariance matrix

  - Radius: when $\boldsymbol{y}$ is approx. bivariate normal, $D^2(\boldsymbol{y})$ has a large-sample χ^2_2 distribution with 2 degrees of freedom.
    - $c^2 = \chi^2_2(0.40) \approx 1$: 1 std. dev univariate ellipse– 1D shadows: $\bar{y} \pm 1s$
    - $c^2 = \chi^2_2(0.68) = 2.28$: 1 std. dev bivariate ellipse
    - $c^2 = \chi^2_2(0.95) \approx 6$: 95% data ellipse, 1D shadows: Scheffé intervals
    - Small samples: $c^2 \approx 2F_{2,n-2}(1-\alpha)$
  - Construction: Transform the unit circle, $\mathcal{U} = (\sin\boldsymbol{\theta}, \cos\boldsymbol{\theta})$,

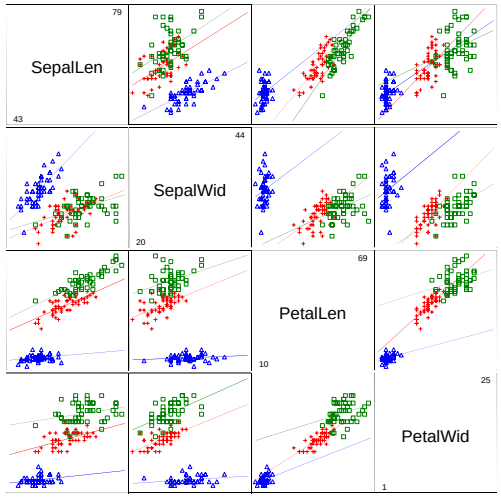
$$\mathcal{E}_c = \bar{\boldsymbol{y}} + c\boldsymbol{S}^{1/2}\mathcal{U}$$

$\boldsymbol{S}^{1/2} =$ any "square root" of $\boldsymbol{S}$ (e.g., Cholesky)

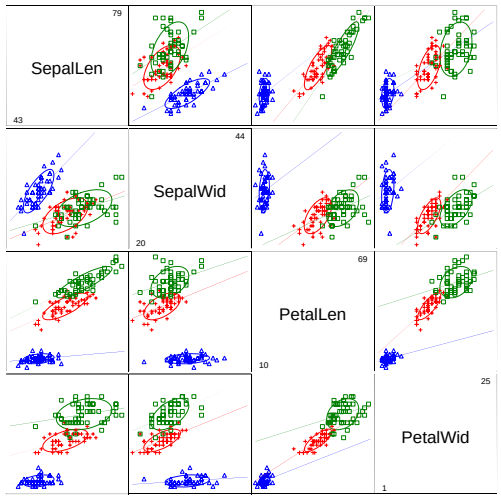
Galton's data: Data Ellipses



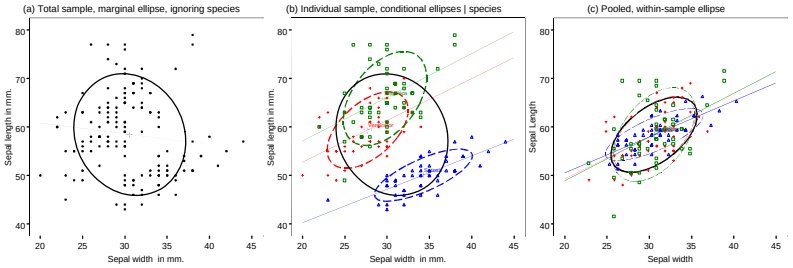
Iris data: scatterplot matrix



Iris data: scatterplot matrix+ data ellipses

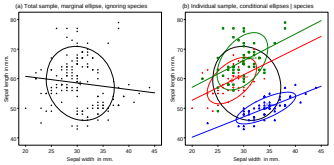


Didactic displays: Simpson's paradox



- Marginal relations: ignore other factors, covariates (species)
- Conditional relations: control or adjust for other factors, covariates
- Simpson's paradox: when these differ in direction.
- Pooled within-sample scatter: $\boldsymbol{S}_{\text{pooled}} = (N-g)^{-1} \sum_{i=1}^{g} (n_i - 1)\boldsymbol{S}_i$
- Visual assessment of equal covariance matrices, $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_g$

Using the ellipses macro



[contiris3.sas ...]

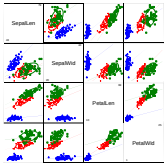
```
title '(a) Total sample, marginal ellipse, ignoring species';
axis2 order=(20 to 45 by 5) offset=(3);
%ellipses(data=iris,
   x=SepalWid, y=SepalLen,
   colors=black,
   interp=rl,      /* add regression line */
   width=3,        /* line width          */
   symbols=dot,    /* plotting symbol     */
   haxis=axis2,    /* horizontal axis     */
   pvalue=.68);    /* ellipse size(s)     */
```


Using the ellipses macro

[... contiris3.sas ...]

```
title '(b) Individual sample, conditional ellipses | species';
%ellipses(data=iris,
   x=SepalWid, y=SepalLen,
   group=species,   /* grouping variable           */
   colors=blue red green black,
   interp=rl,
   all=YES,         /* include total sample ellipse */
   symbols=triangle plus square,
   line = 5 5 5 1,  /* line styles                 */
   width= 3 3 3 1,  /* line widths                 */
   haxis=axis2,
   pvalue=.68);
```

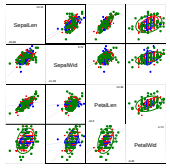

Using the scatmat macro



[scatirisd.sas ...]

```
%include data(iris);
%include macros(scatmat);    *-- or place scatmat.sas in SASAUTOS;
%scatmat(data=iris,
   var=SepalLen SepalWid PetalLen PetalWid,
   group=Species,
   symbols=triangle plus square,
   colors= blue     red  green,
   hsym=4, htitle=9,
   interp=rl,               /* draw regression lines */
   anno=ellipse);           /* and add data ellipses */
```

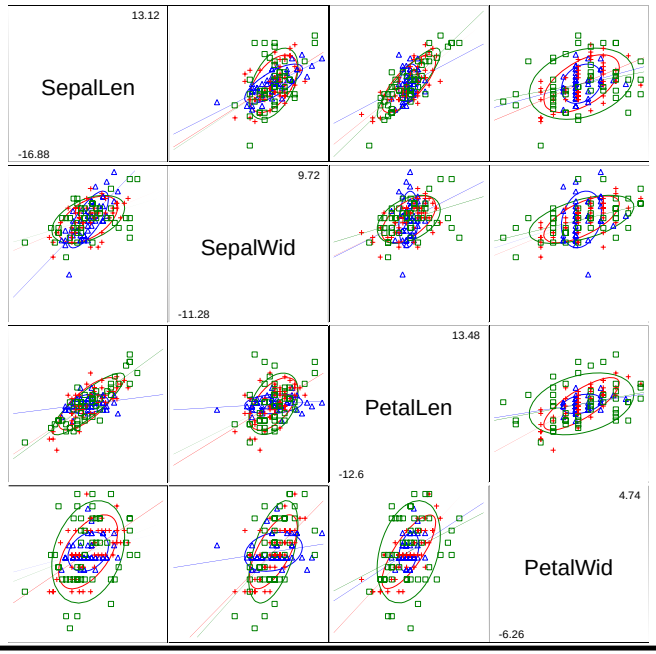
Partial plots



- View covariation after some effects "adjusted for," $\text{VAR}(\boldsymbol{Y} \mid \boldsymbol{X}) = \text{VAR}(\boldsymbol{U})$.
- `model SepalLen--PetalWid = Species` → within-cell error ($\boldsymbol{E}$ matrix)

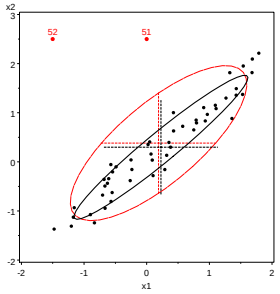
[... scatirisd.sas]

```
*-- remove group means to view within-cell relations;
proc glm data=iris noprint;
   class species;
   model SepalLen SepalWid PetalLen PetalWid = Species /nouni;
   output out=resids
      r=seplen sepwid petlen petwid;
%scatmat(data=resids,
   var=SepLen SepWid PetLen PetWid, group=Species,
   names=SepalLen SepalWid PetalLen PetalWid,
   symbols=triangle plus square,
   colors= blue     red  green,
   hsym=4, htitle=9,
   interp=rl,
   anno=ellipse);
```

Robust data ellipse and outlier detection

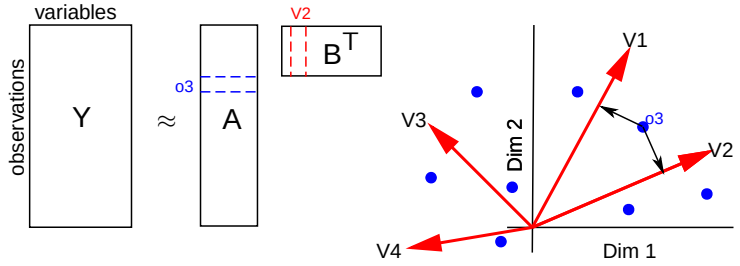
- Multivariate outliers:
  - Shift the mean vector
  - Inflate $\boldsymbol{S}$ → makes them look less extreme



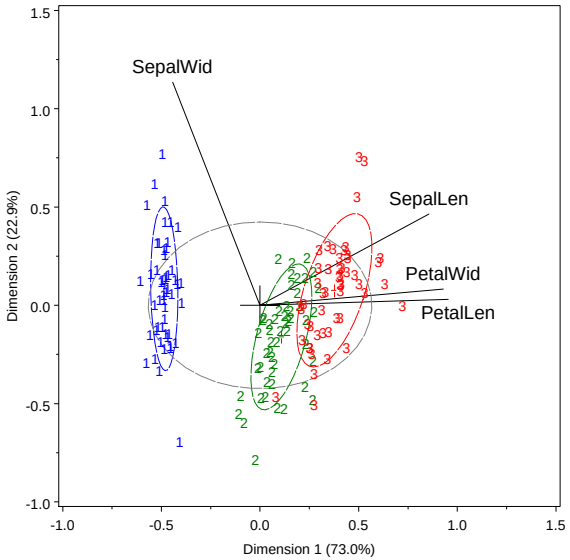
- Robust covariance estimation
  - Multivariate trimming : ignore observations where $\Pr(D^2(\boldsymbol{y}_i) > \chi^2_p) < \alpha$. [implemented in `outlier` macro.]
  - MVE and MCD methods minimize volume or determinant of $\boldsymbol{S}$ containing a fraction γ of points. [implemented in `robcov` macro.]
  - → use data ellipse based on robust $\boldsymbol{S}$ [implemented via `WEIGHT=` in `ellipses` macro.].

Biplots: Low-D views of multivariate data

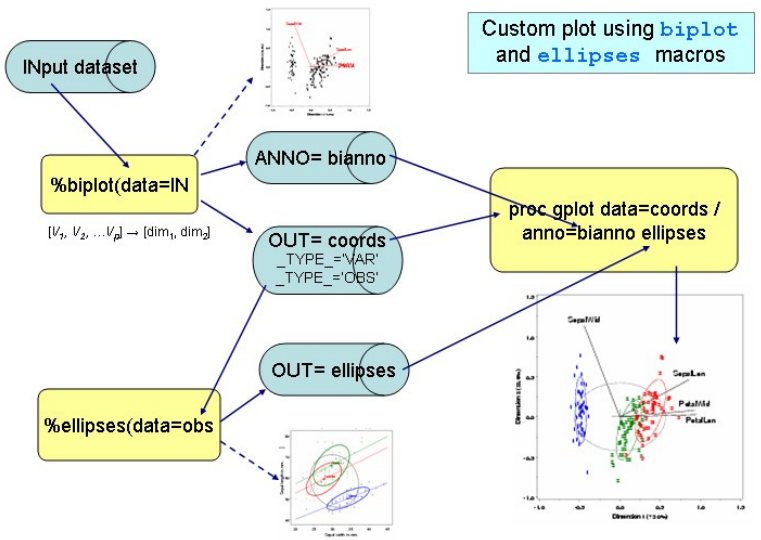
- Display variables *and* observations in a reduced-rank space of d (=2 or 3) dimensions,



- Biplot properties:
  - Plot observations as points, variables as vectors from origin (mean)
  - Angles between vectors show correlations ($r \approx \cos(\theta)$)
  - $y_{ij} \approx \boldsymbol{a}_i^{\mathsf{T}} \boldsymbol{b}_j$: projection of observation on variable vector
  - Observations are uncorrelated overall (but not necessarily within group)



Biplot of Iris data: *setosa*: blue (1); *versicolor*: green (2); *virginca*: red (3)



Custom plot using biplot macro and ellipses macro

- Use biplot macro to obtain biplots scores and variable vectors
- Use ellipses macro to obtain 68% data ellipses
- SAS "Annotate" programming to produce custom display

[bipliris.sas ...]

```
*-- Get biplot scores (_type_='OBS') and var vectors (_type_='VAR');
%biplot(data=iris,
   var=SepalLen SepalWid PetalLen PetalWid,
   id=id,            /* observation ID, here, species number */
   std=std,          /* standardize to mean=0, var=1         */
   scale=0.36,       /* Scale factor for variable vectors    */
   htext=1.5 2,      /* text heights for obs. and var labels */
   xextra=0 1,       /* extra tick mark for labels           */
   gplot=no,         /* suppress the plot                    */
   colors=black,     /* we change these later                */
   out=biplot,       /* output coordinates data set          */
   anno=bianno);     /* output Annotate data set             */
```

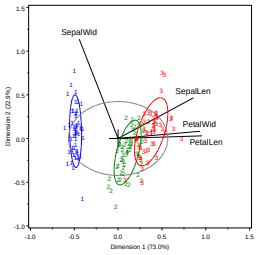

Custom plot using biplot macro and ellipses macro

[... bipliris.sas ...]

```
*-- Customize the Annotate data set;
data bianno;
   set bianno;
   if _type_='OBS' then do;
      *-- change colors;
      select(_name_);
         when ('3') color='RED';
         when ('2') color='GREEN';
         when ('1') color='BLUE';
         otherwise;
       end;
      end;
   *-- adjust label position to avoid overplotting;
   else do;  /*_type_='VAR' */
      if _name_='PetalLen' and function='LABEL' then position='E';
      end;
```

[... bipliris.sas ...]

```
*-- Select just the observation scores;
data biplobs;
   set biplot;
   where (_type_='OBS');

*-- Obtain data ellipses for each group and total sample;
%ellipses(data=biplobs,
   x=dim1, y=dim2,     /* data ellipses for biplot dimension */
   group=_name_,
   all=yes,            /* include total sample ellipse */
   colors=blue green red gray,
   plot=no,            /* suppress the plot */
   pvalue=0.68,
   vaxis=axis98,       /* use AXIS statements generated by %biplot */
   haxis=axis99,
   out=ellipses        /* output Annotate data set */
   );
```

[... bipliris.sas]

```
*-- Join the biplot annotations and the data ellipses;
data bianno;
   set bianno ellipses;

*-- Plot the observations along with annnotations;
symbol1 v=none;
proc gplot data=biplot;
   plot dim2 * dim1 /
      anno=bianno frame
      vaxis=axis98 haxis=axis99
      vminor=1 hminor=1;
```


HE plots: Visualization for the MLM

- **LM**: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$
  - Test *any* hypothesis by *General Linear Test*: $H_0 : \boldsymbol{C\beta} = \boldsymbol{0}$, where $\boldsymbol{C} =$ matrix of constants; rows specify h linear combinations or contrasts of parameters.
  - e.g., Test of $H_0 : \beta_1 = \beta_2 = 0$ in model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$

$$\boldsymbol{C\beta} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

  - All $\longrightarrow$ F-test: How big is SS_H relative to SS_E?

$$F = \frac{SS_H/\text{df}_h}{SS_E/\text{df}_e} = \frac{MS_H}{MS_E} \longrightarrow (MS_H - F\, MS_E) = 0$$

HE plots: Visualization for the MLM

- **MLM**: $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}$, for p responses, $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p)$
  - *General Linear Test*: $H_0 : \boldsymbol{CB} = \boldsymbol{0}$
  - Analogs of sums of squares, SS_H and SS_E are $(p \times p)$ matrices, $\boldsymbol{H}$ and $\boldsymbol{E}$,

$$\boldsymbol{H} = (\boldsymbol{C}\widehat{\boldsymbol{B}})^{\mathsf{T}} [\boldsymbol{C}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-}\boldsymbol{C}^{\mathsf{T}}]^{-1} (\boldsymbol{C}\widehat{\boldsymbol{B}}) \ ,$$

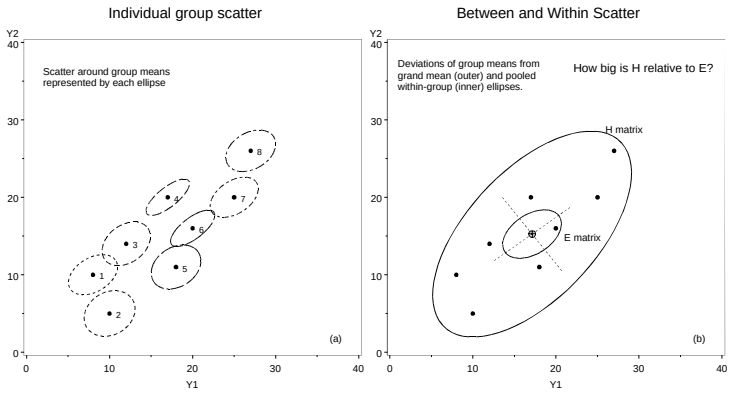
$$\boldsymbol{E} = \boldsymbol{Y}^{\mathsf{T}}[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y} \ .$$

  - Analog of univariate F is

$$|\boldsymbol{H} - \lambda \boldsymbol{E}| = 0 \ ,$$

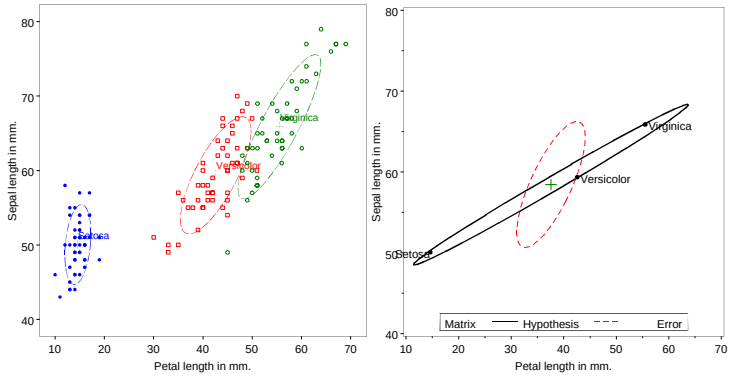
  - How big is $\boldsymbol{H}$ relative to $\boldsymbol{E}$?
    - Latent roots $\lambda_1, \lambda_2, \ldots \lambda_s$ measure the "size" of $\boldsymbol{H}$ relative to $\boldsymbol{E}$ in $s = \min(p, df_h)$ orthogonal directions.
    - Test statistics (Wilks' Λ, Pillai trace criterion, Hotelling-Lawley trace criterion, Roy's maximum root) all combine info across these dimensions
  - HE plot: Shows size, dimensionality, and effect-correlation of $\boldsymbol{H}$ relative to $\boldsymbol{E}$.

Didactic displays for MANOVA tests



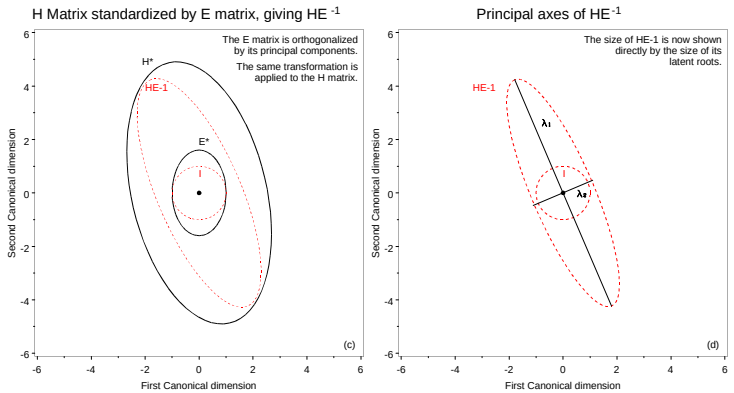
Essential ideas behind multivariate tests: (a) Data ellipses; (b) $\boldsymbol{H}$ and $\boldsymbol{E}$ matrices

HE plot for iris data



(a) Data ellipses and (b) $\boldsymbol{H}$ and $\boldsymbol{E}$ matrices (scaled by $1/df_e$: effect size)

- $\boldsymbol{H}$ ellipse: data ellipse for fitted values, $\hat{\boldsymbol{y}}_{ij} = \bar{\boldsymbol{y}}_j$.
- $\boldsymbol{E}$ ellipse: data ellipse of residuals, $\hat{\boldsymbol{y}}_{ij} - \bar{\boldsymbol{y}}_j$.



Essential ideas behind multivariate tests: latent roots of $\boldsymbol{HE}^{-1}$

HE plots: heplot macro

- Use PROC GLM to obtain outstat= data set w/ $\boldsymbol{H}$ and $\boldsymbol{E}$
- (One $\boldsymbol{H}$ matrix for each effect in model)

[heplot3a.sas ...]

```
proc glm data=iris outstat=stats noprint;
   class species;
   model SepalLen SepalWid PetalLen PetalWid = species / nouni ss3;
```

- heplot macro draws $\boldsymbol{H}$ and $\boldsymbol{E}$ for specified effect

[... heplot3a.sas ...]

```
axis1 label=(a=90) order=(40 to 80 by 10);
%heplot(data=iris,
   stat=stats,            /* Data set containing H & E matrices */
   var=Petallen SepalLen, /* Variables to plot */
   effect=species,        /* Effect to plot */
   vaxis=axis1);
```

PROC GLM outstat= dataset

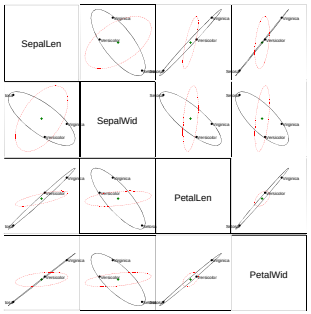
```
proc glm data=iris outstat=stats noprint;
   class species;
   model SepalLen SepalWid PetalLen PetalWid = species / nouni ss3;
*  contrast 'S:VV' species -2  1  1;
*  contrast 'V:V'  species  0 -1  1;
```

The `outstat=` dataset contains:

- $\boldsymbol{E}$ matrix for full model
- One $\boldsymbol{H}$ matrix for each effect in `model`
- One $\boldsymbol{H}$ matrix for each `contrast` statement

SOURCE	_TYPE_	_NAME_	DF	sepallen	sepalwid	petallen	petalwid
ERROR	ERROR	sepallen	147	3896	1363	2462	564
		sepalwid	147	1363	1696	812	481
		petallen	147	2462	812	2722	627
		petalwid	147	564	481	627	616
species	SS3	sepallen	2	6321	-1995	16525	7128
		sepalwid	2	-1995	1134	-5724	-2293
		petallen	2	16525	-5724	43710	18677
		petalwid	2	7128	-2293	18677	8041

HE plot matrices: hemat macro



[... hematiris.sas]

```
%hemat(data=iris,
   stat=stats,
   var=SepalLen SepalWid PetalLen PetalWid,
   effect=species);
```


MANOVA: Contrasts

- As in ANOVA, significant effects in MANOVA may be interpreted by using contrasts
- In the GLT, any contrast(s) can be expressed as a $(h_i \times q)$ $\boldsymbol{C}$ matrix, whose rows sum to zero, e.g.,

$$\boldsymbol{C} = \begin{bmatrix} -2 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}$$

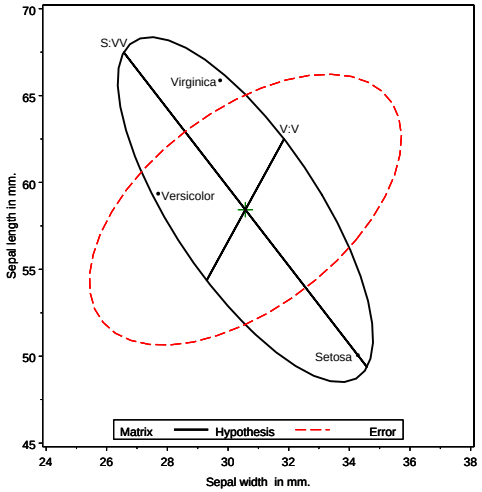
- MANOVA: Pairwise orthogonal 1 df contrasts ($\boldsymbol{C}_i^{\mathsf{T}} \boldsymbol{C}_j = 0$) $\longrightarrow$ rank 1 $\boldsymbol{H}_i$ matrices that decompose the overall hypothesis SSCP matrix,

$$\boldsymbol{H} = \boldsymbol{H}_1 + \boldsymbol{H}_2 + \cdots + \boldsymbol{H}_{\mathsf{df}_h} \ .$$

- Each $\boldsymbol{H}_i$ plots as a vector
  - collection $\longrightarrow$ visual summary of the overall test
  - orthogonal in the $\boldsymbol{H}\boldsymbol{E}^{-1}$ vs. $\boldsymbol{I}$ plot

- **Example**: Iris data
  - $\boldsymbol{C}_1$: *setosa* vs. average of *versicolor* and *virginica* ("S:VV")
  - $\boldsymbol{C}_2$: *versicolor* vs. *virginica* ("V:V")

$$\boldsymbol{C}_1 = \begin{pmatrix} -2 & 1 & 1 \end{pmatrix}$$
$$\boldsymbol{C}_2 = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}$$



Overlaying multiple HE plots with the panels macro

[heplot4.sas ...]

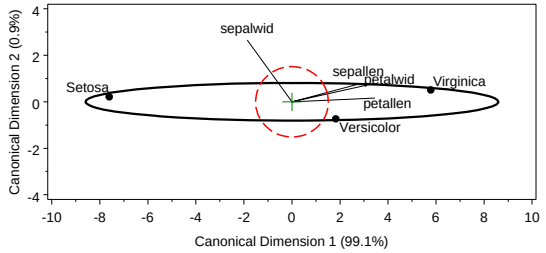
```
proc glm data=iris outstat=stats;
   class species;
   model SepalLen sepalwid PetalLen petalwid = species / nouni ss3;
   contrast 'S:VV' species -2  1  1;
   contrast 'V:V'  species  0 -1  1;
   manova H=species /short summary;
```

[... heplot4.sas]

```
goptions nodisplay;
%heplot(data=iris,stat=stats, var=SepalWid SepalLen, effect=species,
    vaxis=axis1, haxis=axis2);
*-- Contrasts;
%heplot(data=iris,stat=stats, var= SepalWid SepalLen,
   effect=S:VV, ss=contrast, class=, efflab=S:VV,
   vaxis=axis1, haxis=axis2);
%heplot(data=iris,stat=stats, var= SepalWid SepalLen,
   effect=V:V, ss=contrast, class=, efflab=V:V,
   vaxis=axis1, haxis=axis2);
goptions display;
*-- Overlay panels;
%panels(rows=1, cols=1, replay=1:1 1:2 1:3);
```


Reduced-rank HE plots: canplot macro and heplot macro

- HE plot in 2D space that maximally discriminates among groups
- → HE plot of canonical scores: weights given by the latent vectors of $\boldsymbol{HE}^{-1}$



Canonical HE plot for the iris data.

Reduced-rank HE plots: canplot macro and heplot macro

- canplot macro: get scores can1 and can2 on 1st two canonical dimensions

[hecaniris.sas ...]

```
%canplot(
   data=iris,
   class=species,
   var=SepalLen SepalWid PetalLen PetalWid,
   plot=NO,
   scale=3.5,        /* scale factor for variable vectors         */
   out=canscores,    /* output data set containing discrim scores */
   anno=cananno);    /* output data set containing annotations    */
```


Reduced-rank HE plots: canplot macro and heplot macro

- HE plot of canonical scores

[... hecaniris.sas]

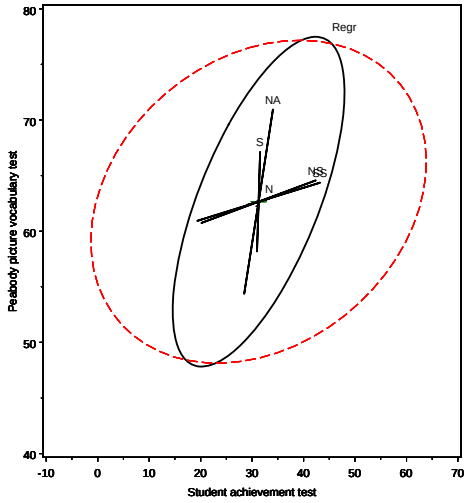
```
*-- Get H and E matrices for canonical scores;
proc glm data=canscores outstat=stats;
   class species;
   model can1 can2  = species / nouni ss3;
   manova h=species;
   run;

*-- Axis statements to equate axis units;
axis1 length=2.6 IN order=(-4 to 4 by 2) label=(a=90);
axis2 length=6.5 IN order=(-10 to 10 by 2);
%heplot(data=canscores, stat=stats,
   x=Can1, y=Can2,
   effect=species,
   haxis=axis2, vaxis=axis1,
   legend=none, hsym=1.6,
   anno=cananno);
```

HE plots for MMRA

- **Overall test**: $H_0 : \boldsymbol{B} = \boldsymbol{0}$ (all coefficients for all responses are zero)
  - $\rightarrow \boldsymbol{C} = \boldsymbol{I}$ in GLT $\rightarrow \boldsymbol{H} = \widehat{\boldsymbol{B}}^{\mathsf{T}} \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1} \widehat{\boldsymbol{B}}$
- **Individual predictors**: $H_0 : \boldsymbol{\beta}_i = \boldsymbol{0}$
  - $\rightarrow \boldsymbol{C} = (0, 0, \ldots, 1, 0, \ldots, 0) \rightarrow \boldsymbol{H}_i = \hat{\boldsymbol{\beta}}_i^{\mathsf{T}} (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1} \hat{\boldsymbol{\beta}}_i$
- **HE plot**
  - Overall $\boldsymbol{H}$ ellipse: how predictors relate collectively to responses
  - Individual $\boldsymbol{H}$ ellipses (vectors):
    - orientation $\rightarrow$ relation of $\boldsymbol{x}_i$ to $\boldsymbol{y}_1, \boldsymbol{y}_2$
    - length $\rightarrow$ strength of relation

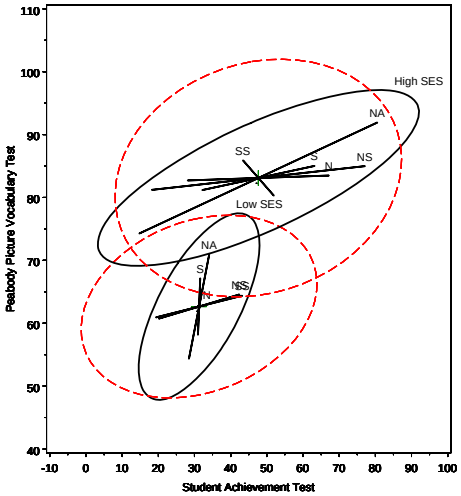
HE plots for MMRA: Example

- Rohwer data on $n = 37$ low SES children, for 5 PA tasks (N, S, NS, NA, SS) predicting intelligence/achievement (PPVT, SAT, Raven): Type III SSCPs



HE plots for MMRA: ANCOVA

- Rohwer data on $n_1 = 37$ low SES children, and $n_2 = 32$ high SES children
  - Fit separate regressions for each group



HE plots for MMRA: ANCOVA

- Rohwer data on $n_1 = 37$ low SES children, and $n_2 = 32$ high SES children
  - Fit ANCOVA model (assuming equal slopes)