Aster Models and Fitness Landscapes

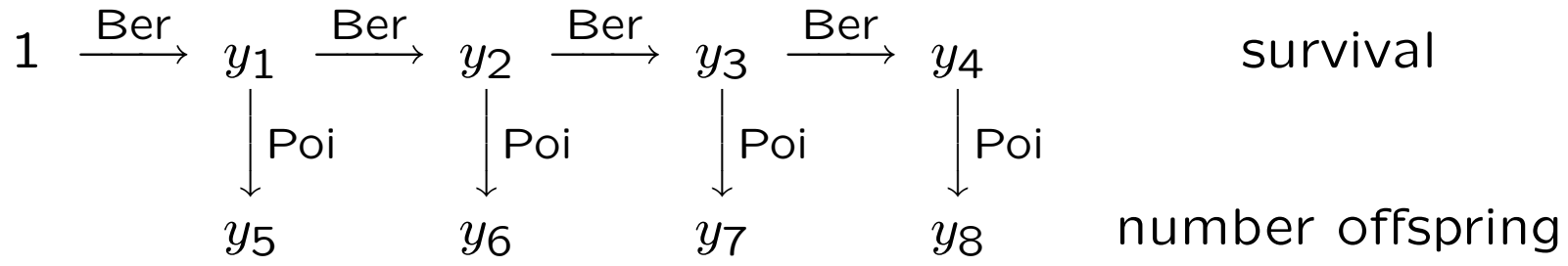
Charles J. Geyer
School of Statistics
University of Minnesota

Ruth G. Shaw
Department of Ecology, Evolution, and Behavior
University of Minnesota

http://www.stat.umn.edu/geyer/aster/

Why Aster Models?

Complex biological data require it.



y_j are components of response for one individual

Arrows indicate conditional distributions (Ber = Bernoulli and Poi = Poisson)

Each Bernoulli (zero-or-one-valued variable) is survival in one year. Each Poisson is number of offspring in that year.

Virtues of Aster Models

Joint analysis of all variables much better than separate analyses.

Joint statistical model for all the data means many issues that are problematic in separate analyses — such as how to treat individuals that die before the time period the separate analysis deals with — are no problem for joint analysis.

Maximum likelihood just works.

The Alternative

Lande, R. and Arnold, S. J. (1983).
The measurement of selection on correlated characters.
Evolution, **37**, 1210–1226.
Cited by 1192 says Google Scholar (Apr, 6, 2008)

Estimates best quadratic approximation (BQA) to fitness landscape.

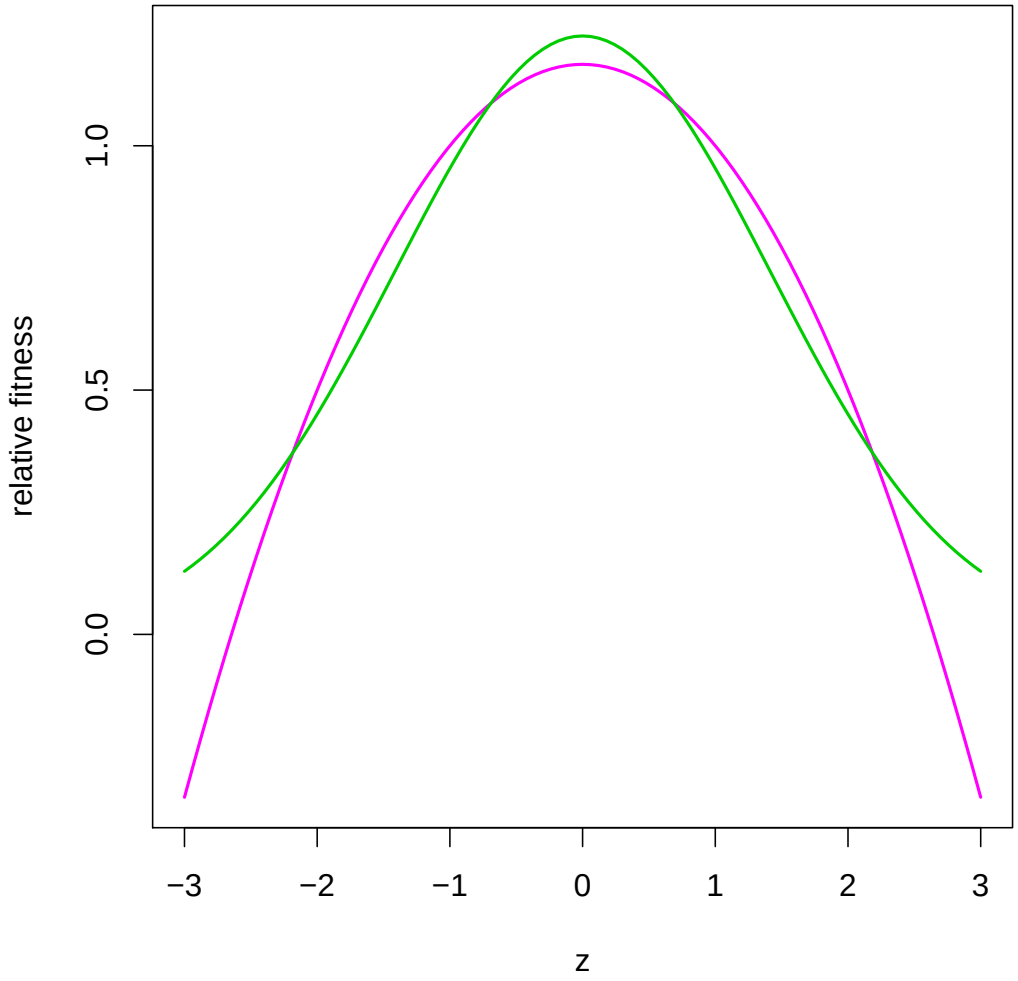
Ordinary least squares (OLS) estimate is best linear unbiased estimator (BLUE) of BQA surface.

Problems with Lande-Arnold Analysis

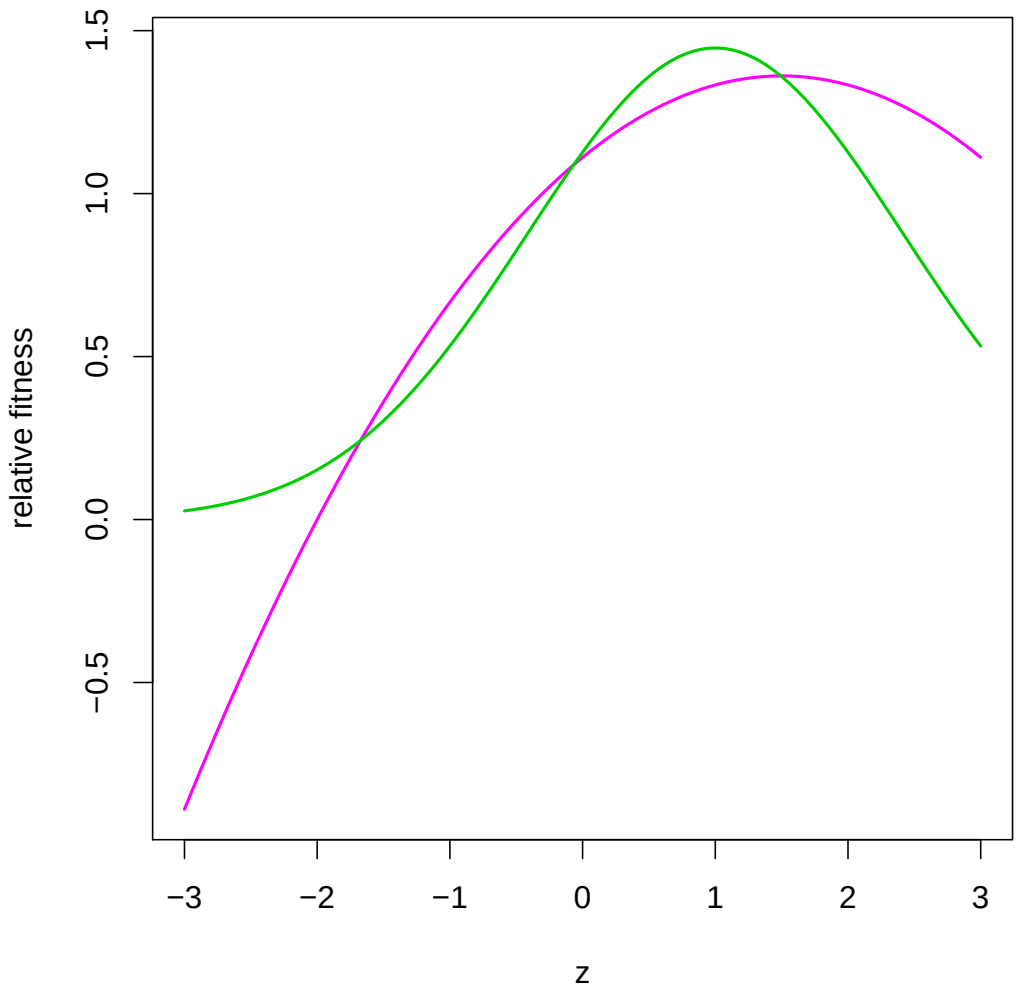
Fitness landscape is never close to quadratic — well, hardly ever — so quadratic approximation is bad approximation.

OLS estimate is BLUE of BQA, but is quite biased estimate of actual fitness landscape.

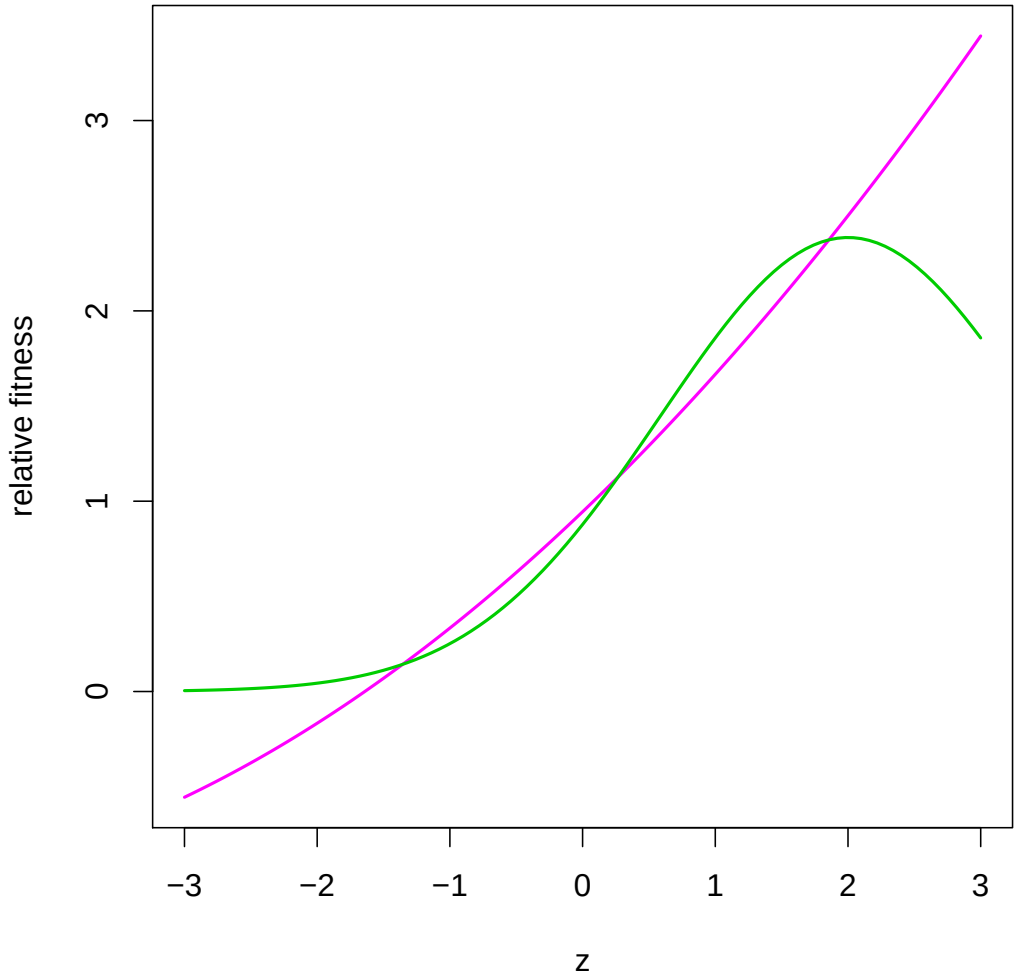
Fitness is never normal, not even close. Standard errors, etc. from OLS not valid.



Fitness landscape (green). Best quadratic approx. (magenta).



Fitness landscape (green). Best quadratic approx. (magenta).



Fitness landscape (green). Best quadratic approx. (magenta).

Statistical Model Hierarchy

- linear models (multiple regression and ANOVA)
  - responses are independent from normal distribution
  - means are linear function of regression coefficients
- generalized linear models (logistic and Poisson regression)
  - responses are independent from same distribution
  - means are monotone function of regression coefficients
- aster models (life history analysis)
  - responses are dependent from different distributions
  - means are monotone function of regression coefficients

Predecessor is Sample Size

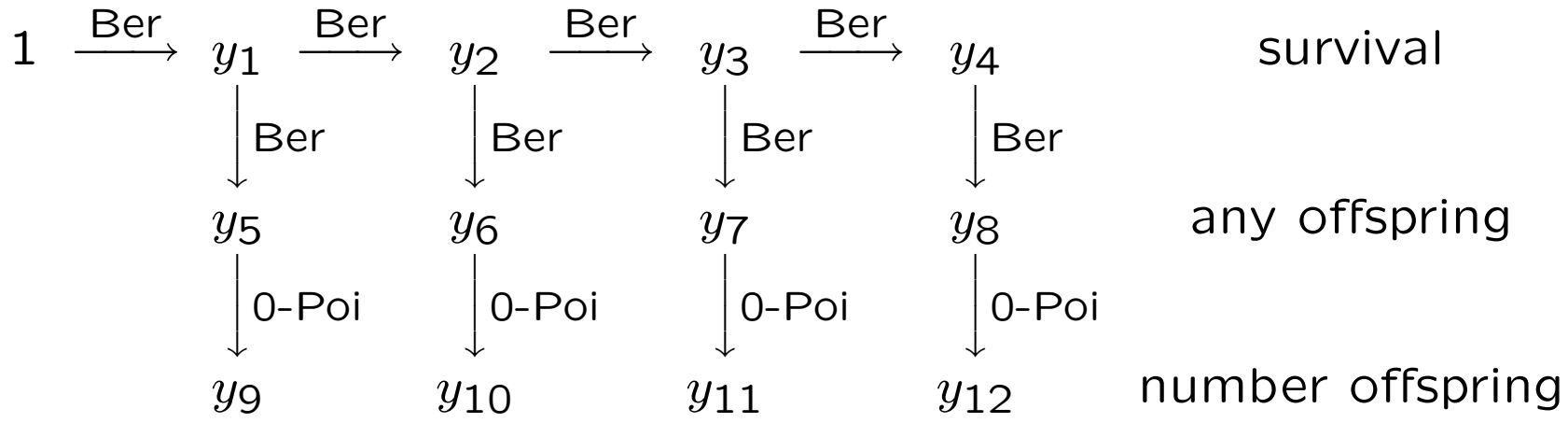
$$\underset{\text{predecessor}}{y_{p(j)}} \longrightarrow \underset{\text{successor}}{y_j}$$

$y_{p(j)}$ is sample size for y_j. Only form of dependence allowed.

$$1 \xrightarrow{\text{Ber}} y_1 \xrightarrow{\text{Ber}} y_2 \xrightarrow{\text{Poi}} y_6$$

y_2 is successor of y_1 and predecessor of y_6

Another Graphical Model

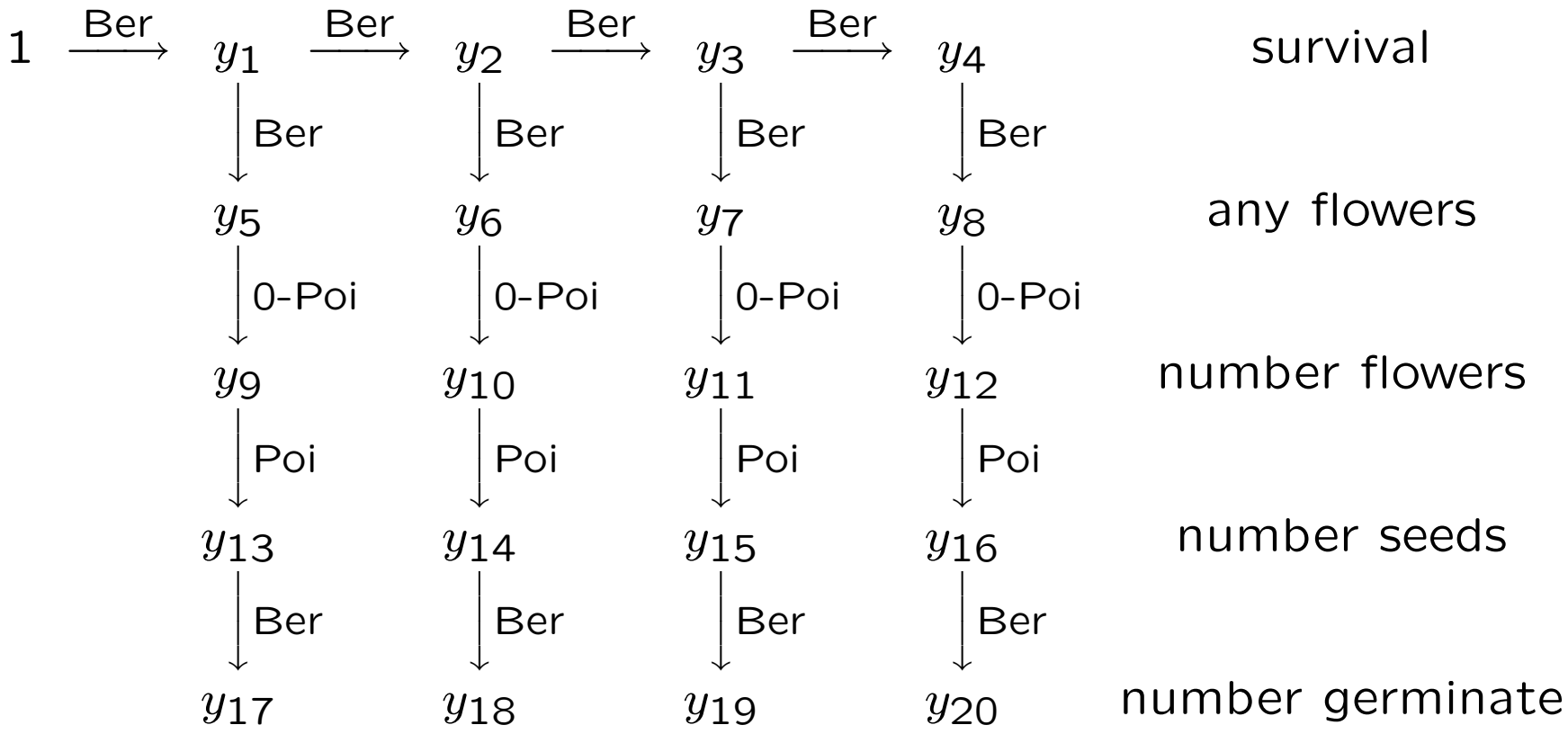


0-Poi = zero-truncated Poisson

$$y_1 \xrightarrow{\text{Ber}} y_5 \xrightarrow{\text{0-Poi}} y_9$$

Conditional distribution of y_9 given y_1 is zero-inflated Poisson.

Graphical Model for Simulated Data



Model Fitting

Also simulate two covariates $\mathbf{z}_1$ and $\mathbf{z}_2$ phenotypic variables. As with linear and generalized linear models, covariates treated as nonrandom, distribution not modeled.

Have linear model on linear predictor scale

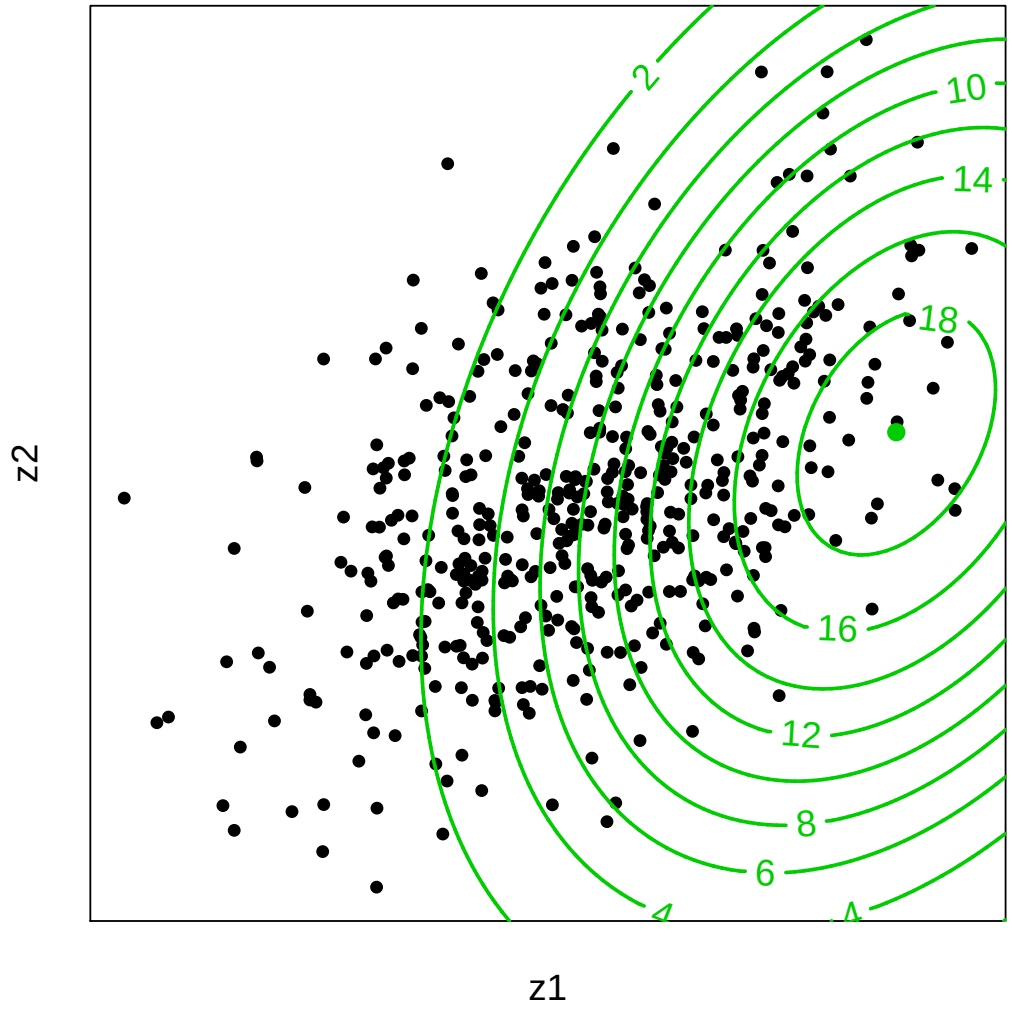
$$\boldsymbol{\eta} = \beta_1 + \beta_2 \mathbf{d}_1 + \cdots + \beta_k \mathbf{d}_{k-1} + \beta_{k+1} \mathbf{z}_1 + \beta_{k+2} \mathbf{z}_2 + \beta_{k+3} \mathbf{z}_1^2 + \beta_{k+4} \mathbf{z}_1 \mathbf{z}_2 + \beta_{k+5} \mathbf{z}_2^2$$

$\mathbf{d}_1$, ..., $\mathbf{d}_k$ dummy variables indicating which node of graph each component of response goes with.

Model Fitting (cont.)

$\mathbf{z}_1$ and $\mathbf{z}_2$ only nonzero for responses in the bottom layer of the graph (counting number of seeds that germinate).

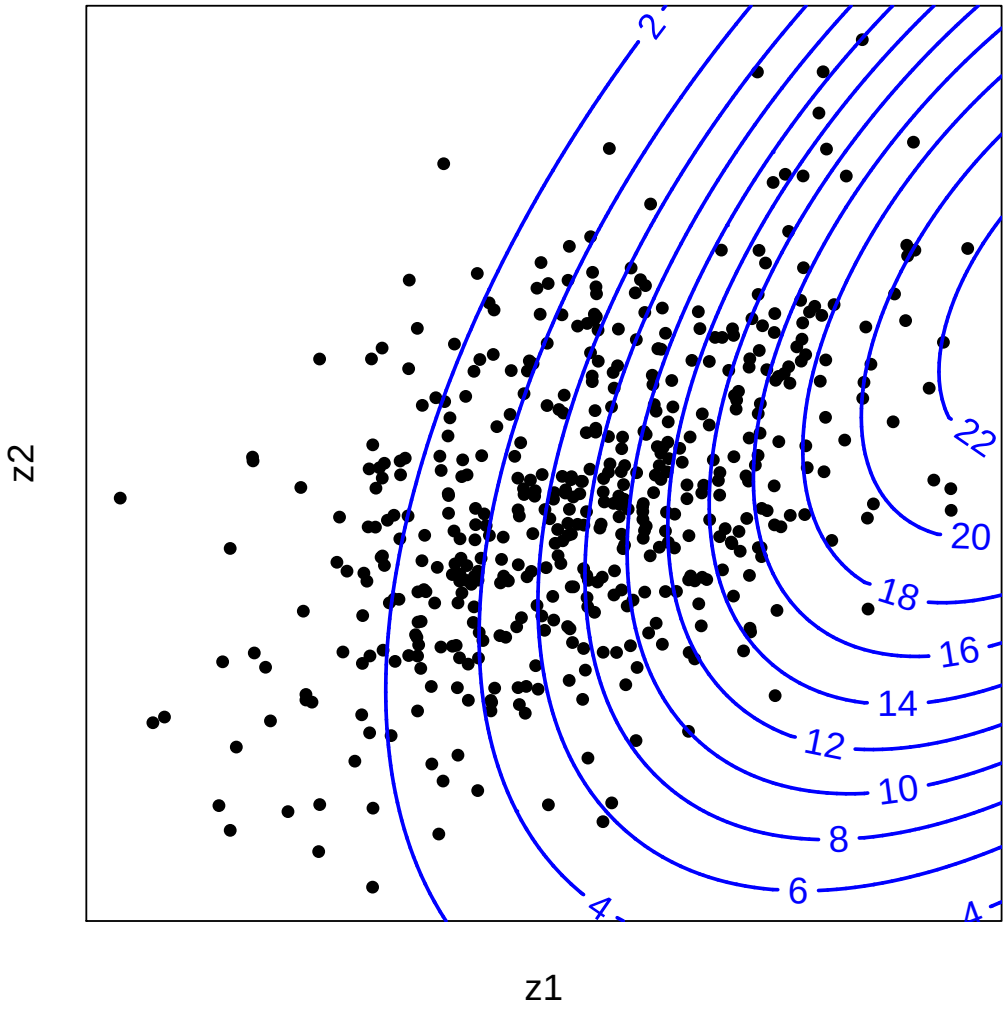
Linear predictor $\boldsymbol{\eta}$ is quadratic function of `z1` and `z2`, and expected fitness (total number germinating seeds) is monotone function of $\boldsymbol{\eta}$.



Simulation Truth Fitness Landscape

Green lines: contours of expected fitness.

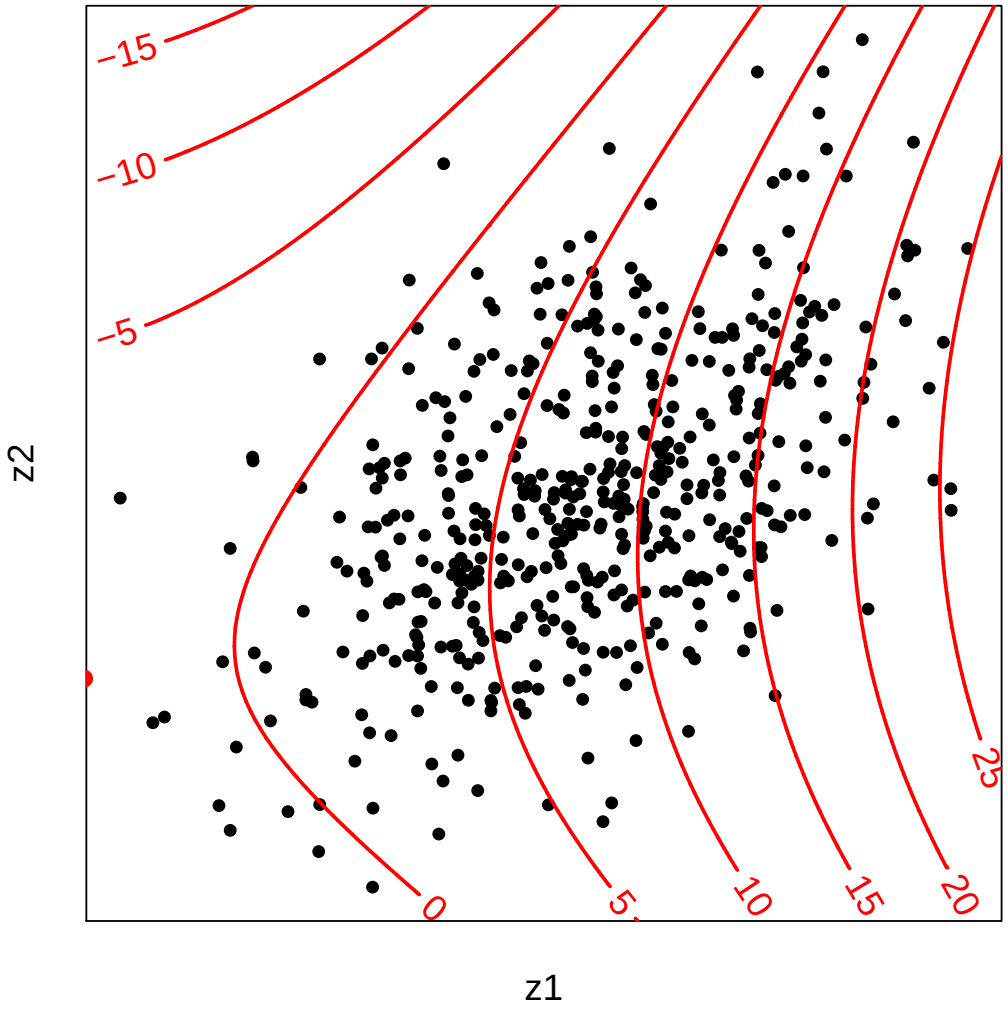
Black dots: simulated phenotype values 500 individuals.



Estimated Fitness Landscape

Blue lines: contours of estimate of expected fitness.

Black dots: simulated phenotype values 500 individuals.



Estimated Best Quadratic Approx. Fitness Landscape

Red lines: contours of estimate of the best quadratic approximation of expected fitness.

Black dots: simulated phenotype values 500 individuals.

Stabilizing or Disruptive Selection?

The aster analysis says selection on z_1 is stabilizing ($P = 0.006$).

The Lande-Arnold analysis suggests selection on z_1 is disruptive ($P = 0.010$, not valid because assumptions for OLS not met).

Much literature about disruptive selection found by Lande-Arnold analysis may wrong — not about biology, only about artifacts of biased statistical analysis.

No way to tell without doing correct statistical analysis.

If you want to take Vienna, take Vienna.

— Napoléon Bonaparte

If you want to estimate the fitness landscape, estimate the fitness landscape.

— us

Geyer, C. J., Wagenius, S. and Shaw, R. G. (2007).
Aster models for life history analysis.
Biometrika, **94** 415–426.

Shaw, R. G., Geyer, C. J., Wagenius, S., Hangelbroek, H. H.,
and Etterson, J. R. (2008).
Unifying life history analysis for inference of fitness and
population growth.
To appear in *American Naturalist*.

All details of all computations given in tech reports at
http://www.stat.umn.edu/geyer/aster/

R contributed package

```
install.packages("aster")
library(aster)
```

Means are Monotone Function of Linear Predictor

In generalized linear model (GLM) or aster model

$$\boldsymbol{\eta} = \mathrm{M}\boldsymbol{\beta}$$

where $\boldsymbol{\eta}$ is "linear predictor" vector, M is model matrix, $\boldsymbol{\beta}$ is vector of regression coefficients.

Vector $\boldsymbol{\mu}$ of response means is multivariate monotone function of linear predictor vector $\boldsymbol{\eta}$

$$(\boldsymbol{\mu} - \boldsymbol{\mu}')^T(\boldsymbol{\eta} - \boldsymbol{\eta}') \geq 0$$