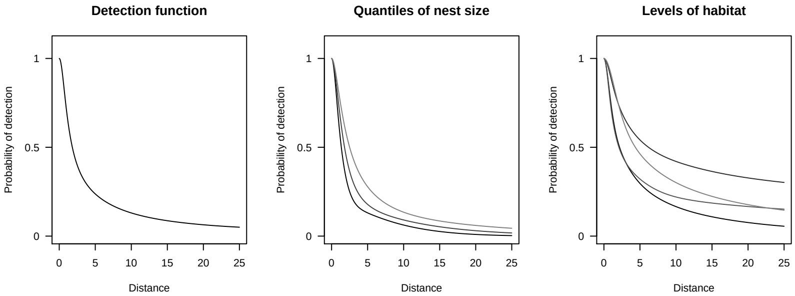
Strategies for correlated covariates in distance sampling

David Lawrence Miller

CREEM, University of St Andrews

Covariates in distance sampling

- CDS: $\mathbb{P}$(observing an object) depends on distance
- MCDS: what about other factors?
  - per animal (sex, size, . . .)
  - environmental effects (weather, time of day, habitat, . . .)
  - observer effects (individual, team, pilot, . . .)
  - (group size – not addressed here)



Detection functions

- Models of the form

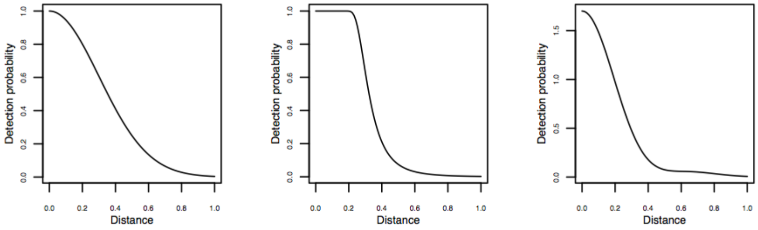
$$g(x; \boldsymbol{\beta}, z_1, \ldots, z_J) = \mathbb{P}(\text{detected}|\text{observed } x, z_1, \ldots, z_J)$$

- distances x
- estimate parameters $\boldsymbol{\beta}$
- covariates $z_1, \ldots, z_J$, that affect detection
- covariates enter model via scale parameter:

$$\sigma(z_1, \ldots, z_J) = \exp(\beta_0 + \sum_j z_j \beta_j)$$

Constraints and particulars

- g has fixed functional form
- usually <5 covariates
- covariates independent from distance (in population)
- inference on likelihood *conditional* on observed covariates



Motivating example: AK black bears

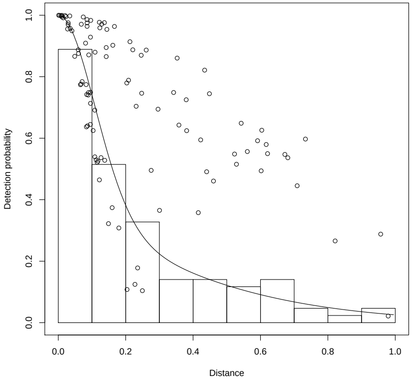
- Black bear data from Alaska
- 301 aerial observations
- 3 covariates (low correlation):
  - search distance
  - % foliage cover
  - % snow cover

What can go wrong?

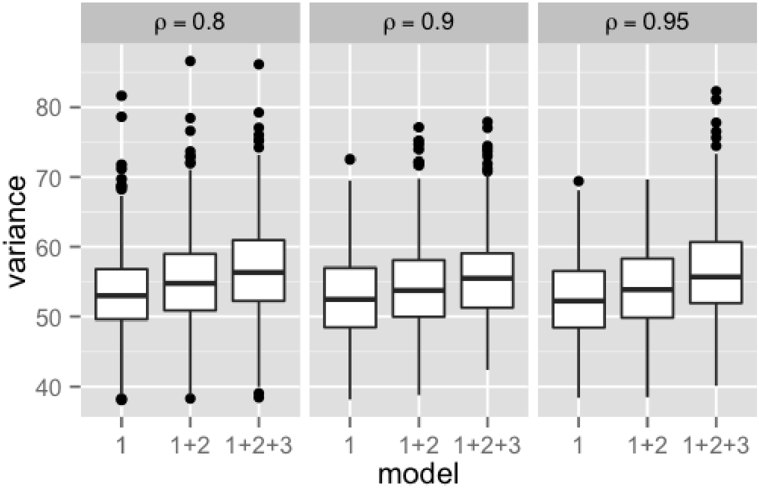
- from linear model literature:
  - fitting problems
  - high(er) variance
  - non-interpretable covariates
  - prediction usually fine
- important for DS:
  - fitted values ($\hat{p}_i(z_1, \ldots, z_J)$) important
  - rarely "predict"
  - variance important
  - covariates are nuisance

Simulated example

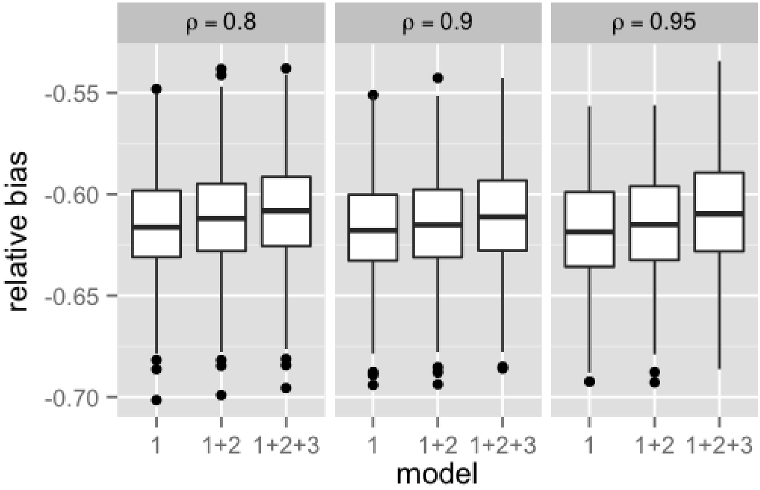
- half-normal detection function
  $\exp\left(\frac{-x^2}{2\sigma(z_1,\ldots,z_J)^2}\right)$
- $z_1 \sim \text{beta}(0.1, 0.4)$
- z_2, z_3 generated to be correlated with z_1
- fitted:
  - $\sigma_1 = \exp(\beta_0 + z_1\beta_1)$
  - $\sigma_{12} = \exp(\beta_0 + z_1\beta_1 + z_2\beta_2 3$
  - $\sigma_{123} = \exp(\beta_0 + z_1\beta_1 + z_2\beta_2 + z_3\beta_3)$
- select model by AIC
- $\sim$90 samples per realisation



Simulated example - Var($\hat{N}_c$)



Simulated example - bias in $\hat{N}_c$



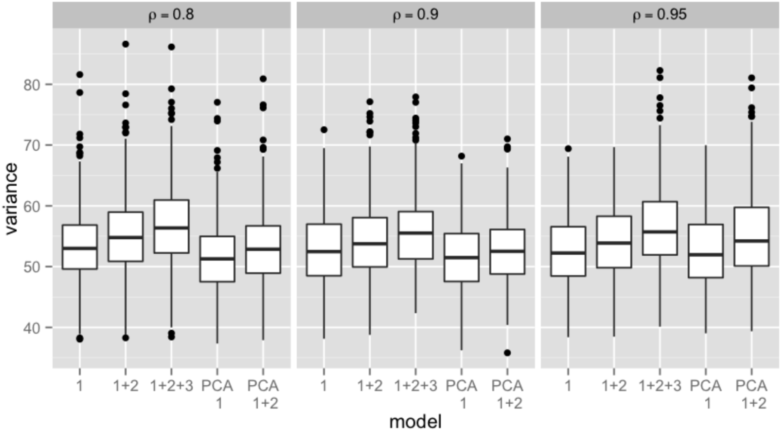
What can we do?

- Obvious possibilities from linear modelling:
  - Ridge regression
  - Lasso
  - PCA
- Shrinkage methods require estimate shrinkage!
  - change in fitting procedure

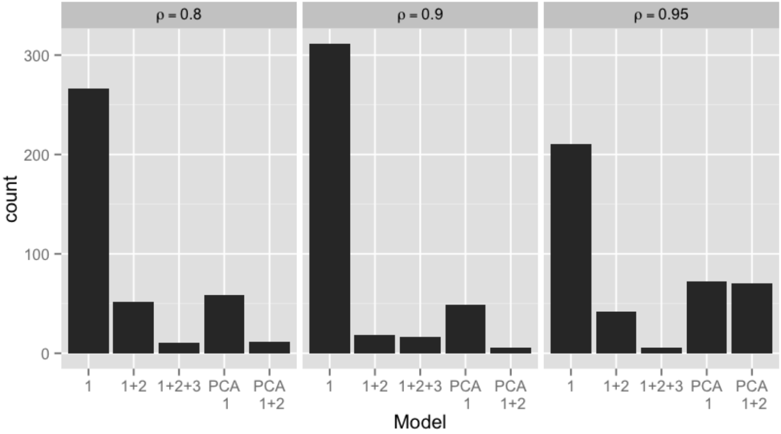
Simple solutions

- Principle components
  - fast, simple, most people know about it
  - "derived" covariates – no change in fitting procedure
  - *only* covariates, **not** distance
- standardise covariates $\Rightarrow \mathcal{Z}$
- take $\mathcal{Z}^\mathsf{T}\mathcal{Z} = U^\mathsf{T}\Lambda U$
- new covariates $z_j^* = \mathcal{Z}\mathbf{u}_j$
- select new PCA covariates in order
  - using all gives same fit as no-PCA model

Simulation revisit - Var($\hat{N}_c$)



Simulation revisit - AIC



Black bears - results

- Full model (3 PC/3 covariates) selected by AIC
- Adding 1 dummy covariate
  - correlated with search distance ($\rho = 0.9$)
  - 3 PC model better AIC than 2 or 1
  - 3 PC model give $\sim 10\%$ saving in variance
- for both models, small changes in $\hat{p}$

What's going on?

In terms of $\hat{N}_c$:

$$\mathrm{var}(\hat{N}_c) = w^2 \sum_i \hat{f}(0|\mathbf{z})^2 - \hat{N}_c + \left[\frac{\partial N_c}{\partial \beta}\right]^{\mathsf{T}} H^{-1} \left[\frac{\partial N_c}{\partial \beta}\right]$$

$$= \sum_i \frac{1-\hat{p}_i}{\hat{p}_i^2} + \left[\frac{\partial \hat{N}_c}{\partial \hat{\beta}}\right]^{\mathsf{T}} H^{-1} \left[\frac{\partial \hat{N}_c}{\partial \hat{\beta}}\right]$$

first term dominates.

Further work

- what about other situations?
  - hazard-rate, etc. detection functions
  - factor covariates
- is it ever "bad" to do this?
- is ridge/lasso more "efficient"?
- is anyone here doing "large" analyses?

Talk available at:

`http://converged.yt/talks/dscorrcovar.pdf`