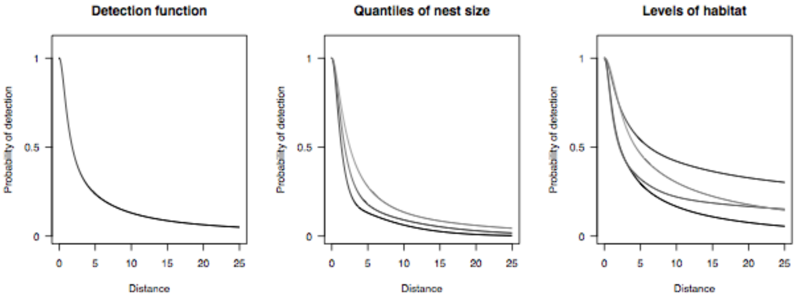
Strategies for correlated covariates in distance sampling

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

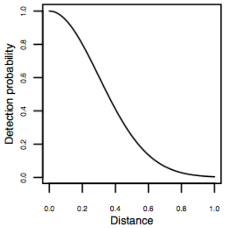
$$g(x; \theta, z) = \mathbb{P}(\text{detected}|\text{observed } x, z_1, \ldots, z_J)$$

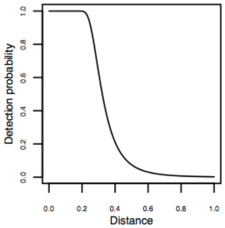
- distances x
- estimate parameters θ
- covariates z, that affect detection
- covariates enter model via scale parameter:

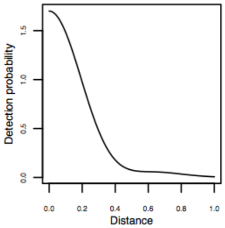
$$\sigma(z_1, \ldots, z_J) = \exp(\beta_0 + \sum_j z_j \beta_j)$$

Constraints and particulars

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





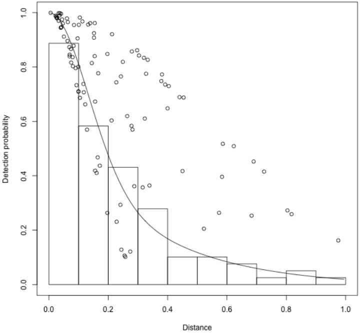


What can go wrong?

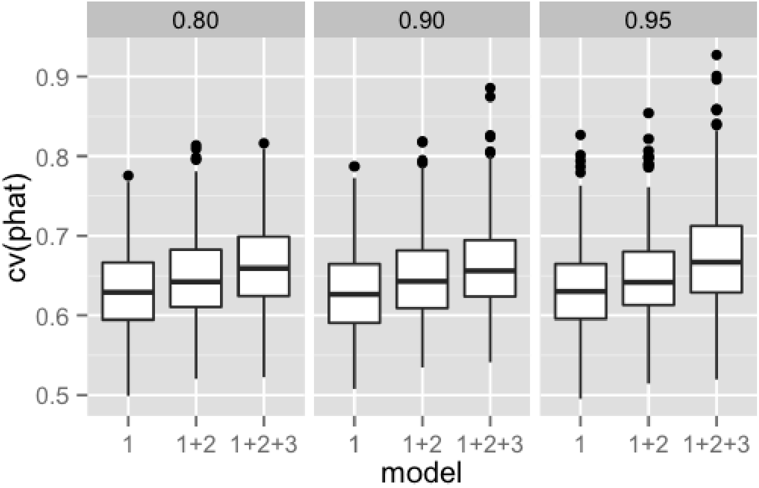
- from linear model literature:
  - fitting problems
  - prediction fine
  - high(er) variance
  - non-interpretable covariates
- important for DS:
  - fitted values ($\hat{p}_i(\mathbf{z}_j)$) important
  - rarely "predict"
  - variance important
  - covariates are nuisance

Example

- half-normal detection function
- 1 "real" continuous covariate – beta(0.1,0.4)
- 2 correlated "fake" covariates
- select terms by AIC
- ~95 samples per realisation



Example - CV(p)



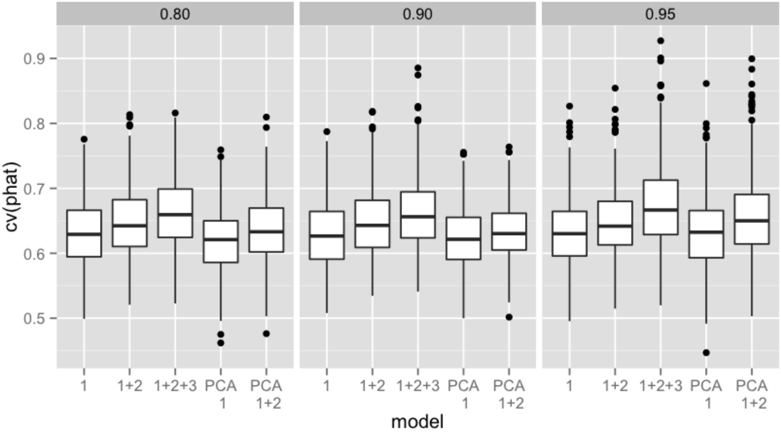
Stealing ideas from regression

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

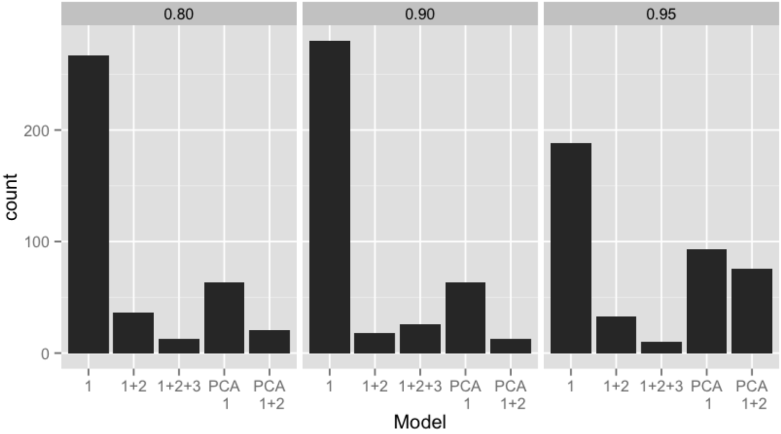
Simple solutions

- 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 covars in order
  - using all gives same fit as no-PCA model

Simulation revisit - CV(p)



Simulation revisit - AIC



Black bears

- Black bear data from Alaska
- 301 observations from Piper Super Cub
- double observer ignored
- heavily left truncated (99m for this analysis)
- 3 covariates:
  - "search distance" (composite measure)
  - % foliage cover
  - % snow cover
- 1 covariate correlated 0.9 search distance

Black bears - results

- fitting 3 covariate model (observed covars)
  - 2 PCs better AIC than 1 PC
  - 3 PC (full model) better than 2 in AIC
  - 2 PC model give $\sim$ 2% saving in var
- fitting 4 covariates (observed + 1 correlated)
  - 3 PC model better AIC than 2 or 1
  - 3 PC model give $\sim$ 10% saving in var

What's going on?

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

$$\begin{aligned}\text{var}(\hat{N}_c) &= w^2 \sum_i \hat{f}(0|\mathbf{z})^2 - \hat{N}_c + \left[\frac{\partial N_c}{\partial \theta}\right]^\mathsf{T} H^{-1} \left[\frac{\partial N_c}{\partial \theta}\right] \\ &= \sum_i \frac{1-\hat{p}_i}{\hat{p}_i^2} + \left[\frac{\partial \hat{N}_c}{\partial \hat{\theta}}\right]^\mathsf{T} H^{-1} \left[\frac{\partial \hat{N}_c}{\partial \hat{\theta}}\right]\end{aligned}$$

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`