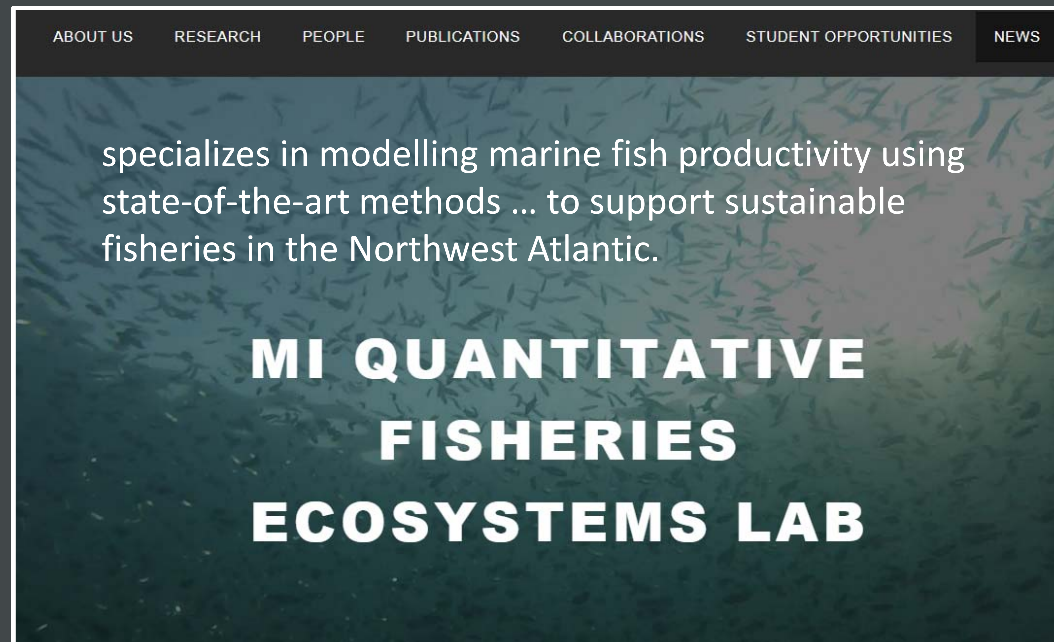
Approaches for modelling landings and catch age comps in state-space assessment models

Noel Cadigan

OCEAN CHOICE INTERNATIONAL (OCI) INDUSTRIAL RESEARCH CHAIR IN FISH STOCK ASSESSMENT ….



CAPAM Workshop. November 7 2019

MARINE INSTITUTE

UNIVERSITY

Catch “Data”

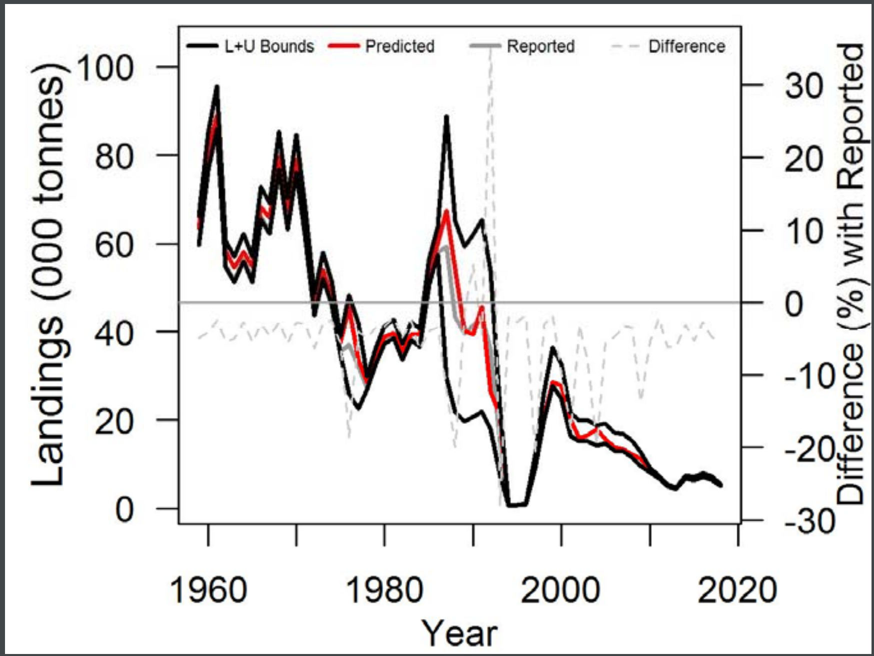
- Catch-at-age <= landings, proportions-at-age, and weights-at-age
- All of these have some uncertainty
- The integrated assessment philosophy suggests we should include different likelihoods (equiv. nll’s) for different data sources
- This allows us to treat the uncertainty in different data sources differently
- I focus on landings and age-comps (weights for another day)

Landings

- Usually have little information about the accuracy of landings (uncertain uncertainty)
- But we may suspect or know they are under-reported
- Or there is some information on discards, area mis-reporting, …
- Information will change over the time-series
- Some early NAFO landings <u>negotiated</u>???
- Optimal solutions may be case-specific
- I have been using a censored likelihood approach …

Censored (i.e. partial) landings

- Landings bounds from ***expert opinion***
- bounds used via a censored likelihood
- Tell the model how good landings are
- Models may not determine this well.



Hammond, T. R., & Trenkel, V. M. Censored catch data in fisheries stock assessment.

Detecting and correcting underreported catches in fish stock assessment: trial of a new method

Nicolas Bousquet, Noel Cadigan, Thierry Duchesne, and Louis-Paul Rivest

A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates[1]

Noel G. Cadigan

How catch underreporting can bias stock assessment of and advice for northwest Atlantic mackerel and a possible resolution using censored catch

Elisabeth Van Beveren[a,*], Daniel Duplisea[a], Martin Castonguay[a], Thomas Doniol-Valcroze[a],

MARINE INSTITUTE UNIVERSITY

Censored log-likelihood for Landings

- We only know a range of values for possible landings….
- L_y is the total model predicted landings in year y
- B_{ly} and B_{uy} are the lower and upper bounds
- assume lognormal measurement error

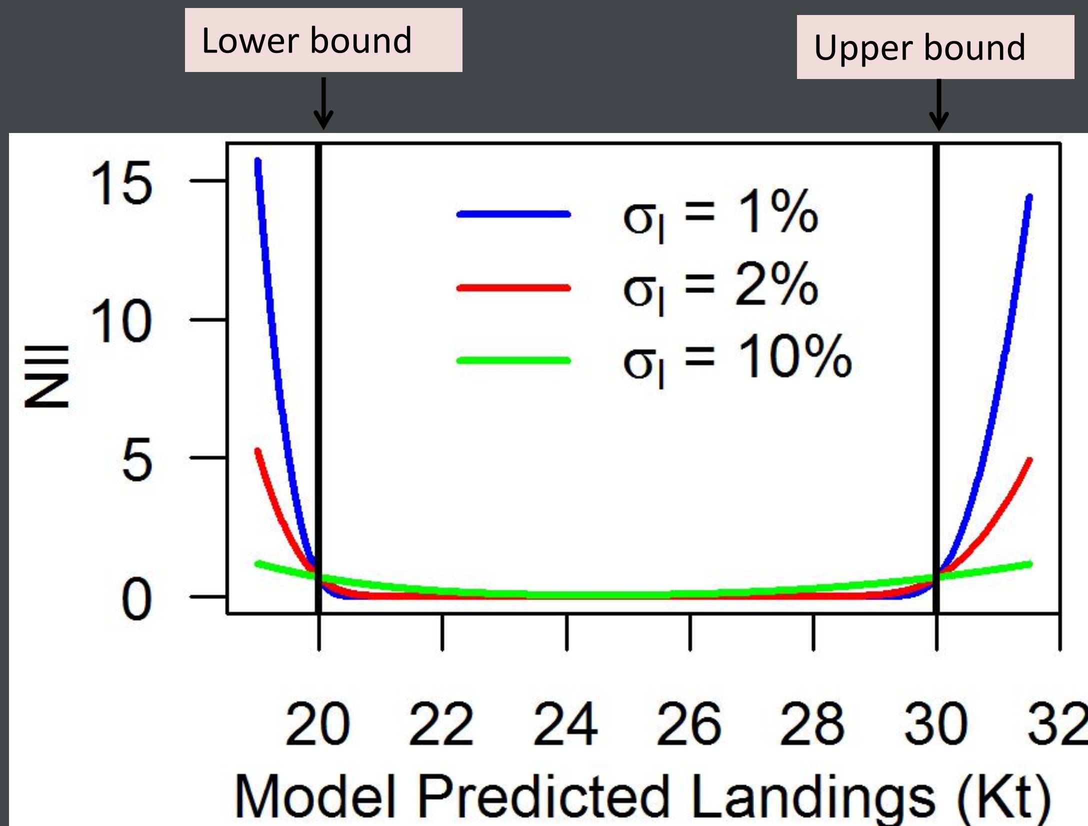
Assessment model predicted

$$l(\theta|\{B_l, B_u\}|) = \sum_{y=1}^{Y} log\left[\Phi\left\{\frac{\log(B_{uy}/L_y)}{\sigma_C}\right\} - \Phi\left\{\frac{\log(B_{ly}/L_y)}{\sigma_C}\right\}\right]$$

$[\cdot] = \log\{\Pr(B_{ly} \le L_y \le B_{uy} \,|\, \theta)\}$

Assumed measurement error in range

Censored nll for landings, illustration



Likelihood for catch-at-age proportions

- Often difficult to evaluate from first principles the statistical properties of catch-at-age :: very complex sampling
- Often poorly documented over time
- Francis (2014) concluded that the **logistic normal multinomial (LNM)** distribution showed great promise ….
- Aitchison (2003) referred to this as the additive logistic transformation.
- He argued that for ordered compositional data, such as for ages and lengths, the multiplicative logistic transformation is more appropriate

Multiplicative LMVN distribution

- Let $\pi_b = P_b/(P_b + \ldots + P_B) = Prob(State = b \mid State \geq b)$

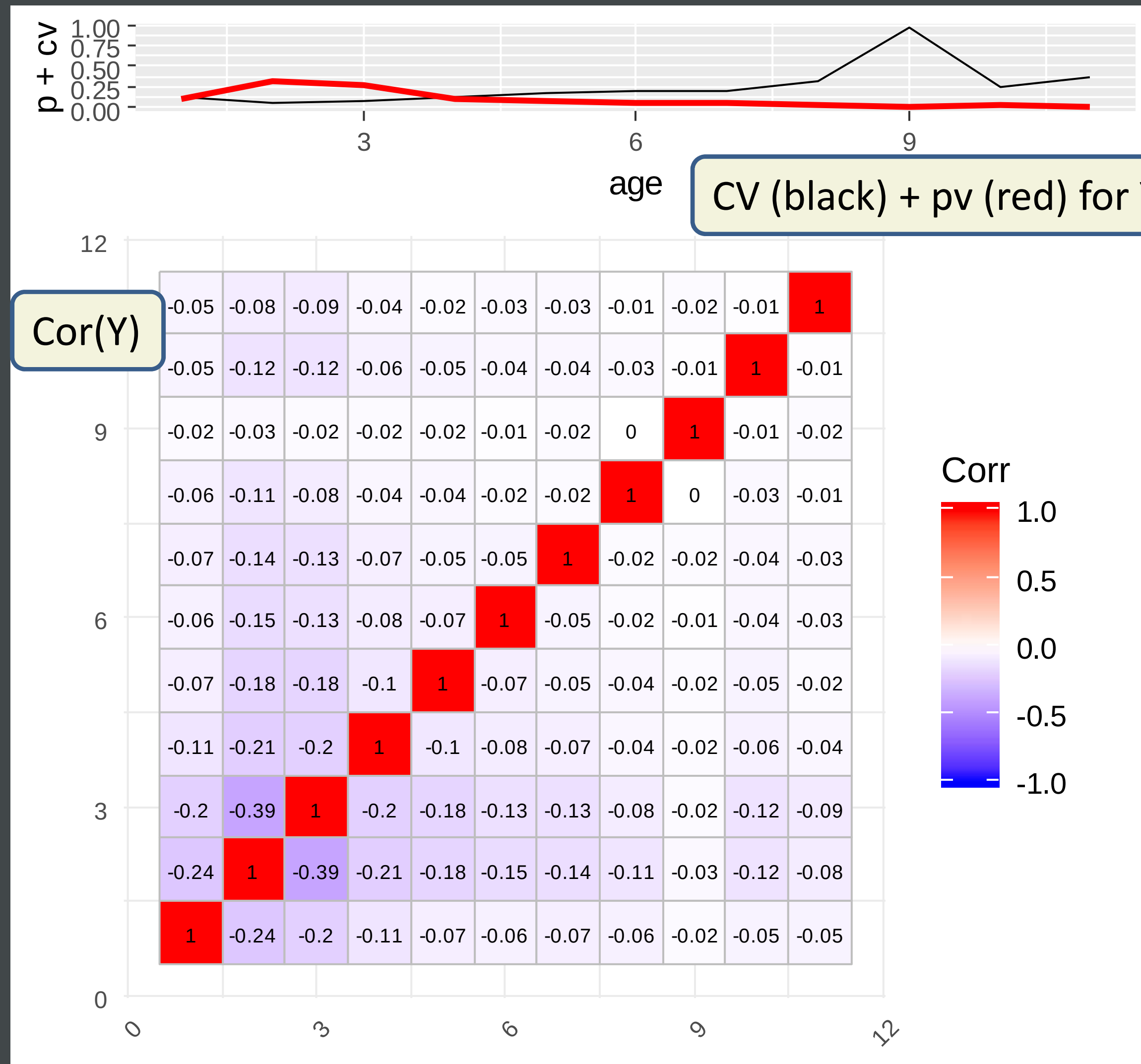
$$X_b = log\left(\frac{\pi_b}{1-\pi_b}\right) = log\left(\frac{P_b}{P_{b+1}+, \ldots, +P_B}\right), b = 1, \ldots, B-1$$

- The multiplicative LMVN distribution is based on assuming $X_1, \ldots, X_{B-1}$ is MVN with *some correlation structure*.
- I examine **Dirichlet** samples $Y_1, \ldots, Y_A \sim D(\lambda P_1, \ldots, \lambda P_A)$ with
  - $E(Y_i) = P_i$
  - $Var(Y_i) = P_i (1-P_i)/(1+\lambda)$
  - $Cov(Y_i, Y_j) = -P_i P_j/(1+\lambda)$

Case 1:

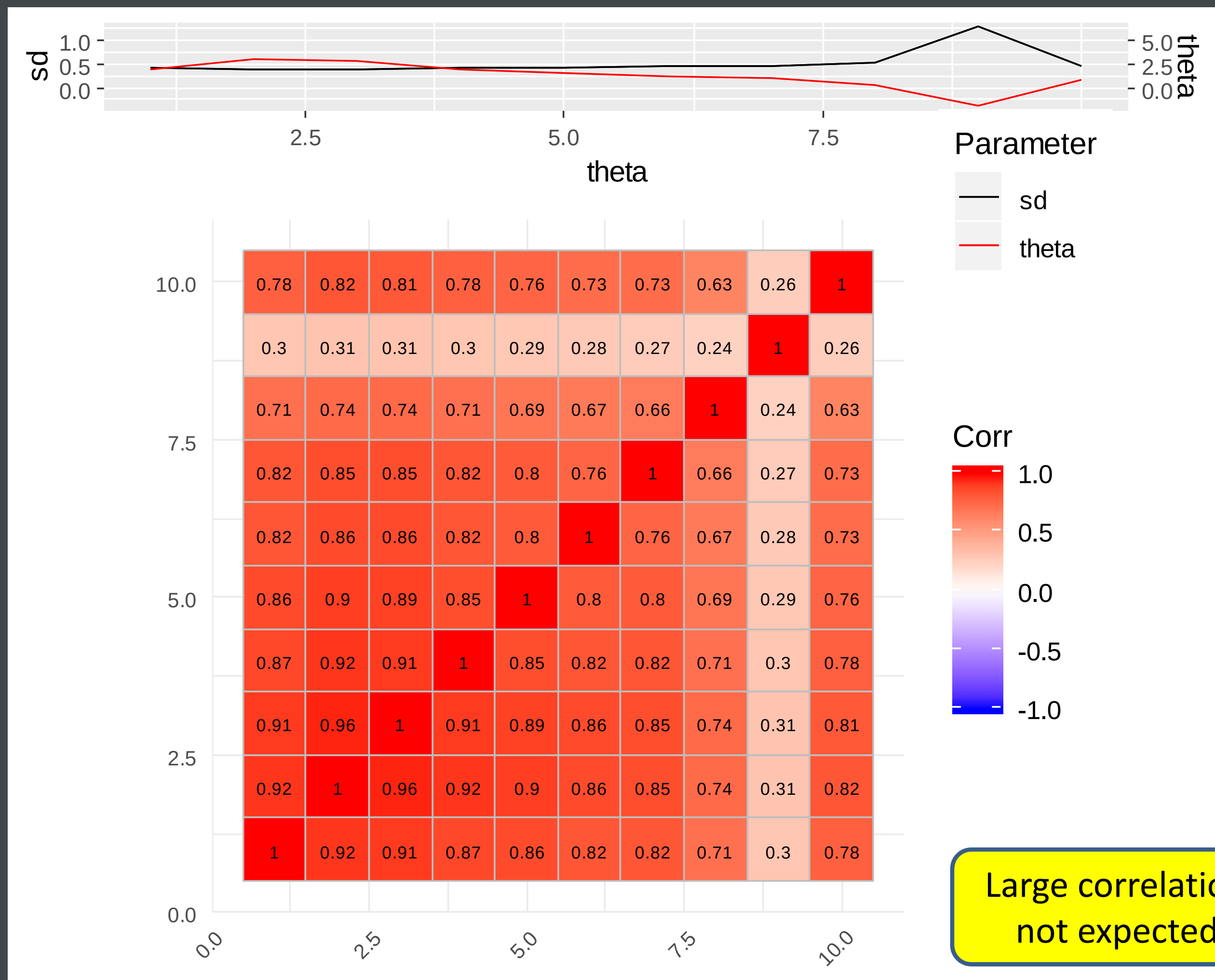
```
c=c(45,147,124,44,32,21,20,9,1,15,7)
pv=c/sum(c)
lambda = 500;phi = lambda*pv
n.year=10000
Y = matrix(NA,nrow=n.year,ncol=n.age)
```

```
for(i in 1:n.year){
  Y[i,]=rdirichlet(1,phi)
}
```



All as expected!

Case 1 Additive Logistic: $\theta_b = log\{Y_b/Y_B\}$

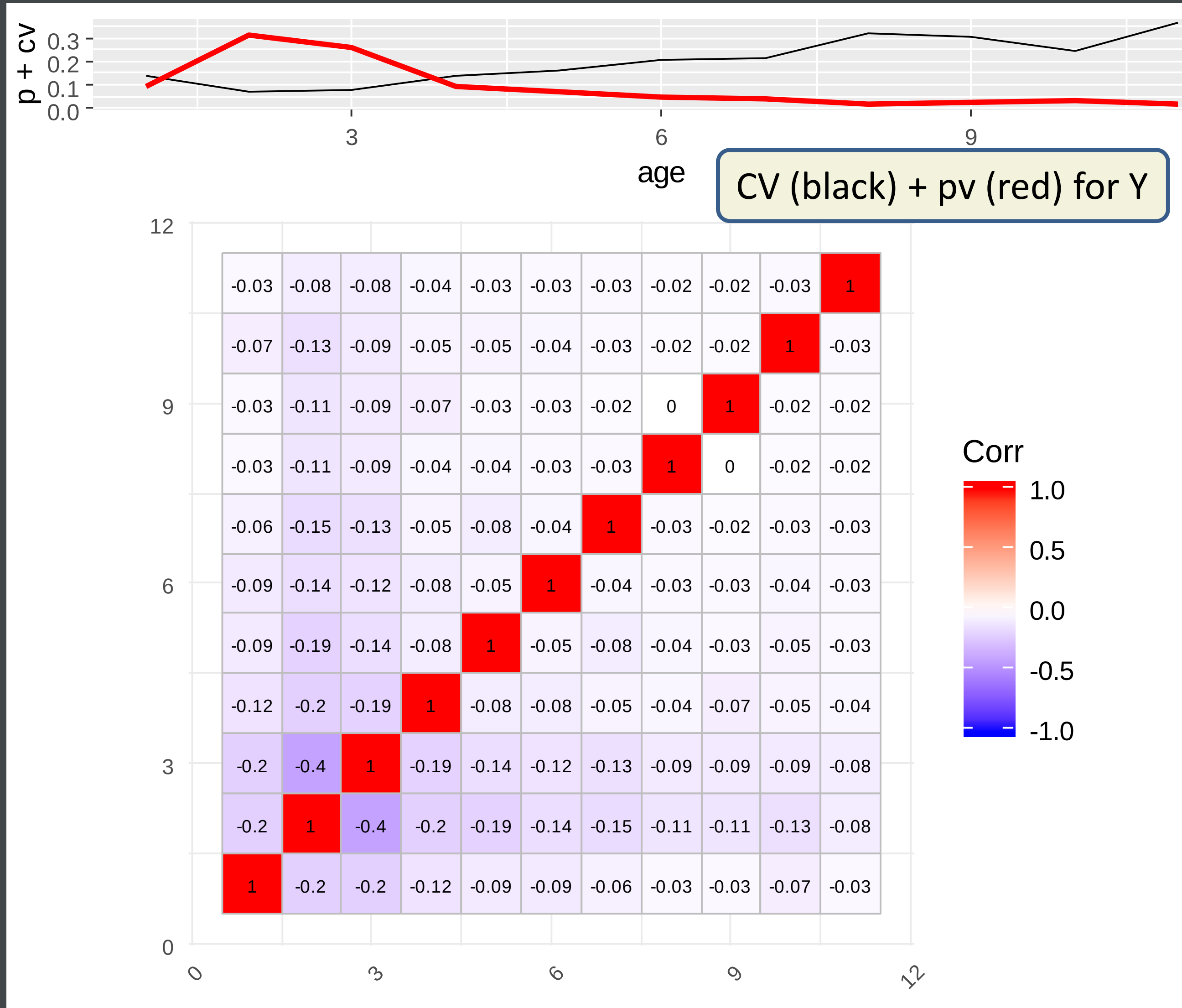


Large correlations
not expected!

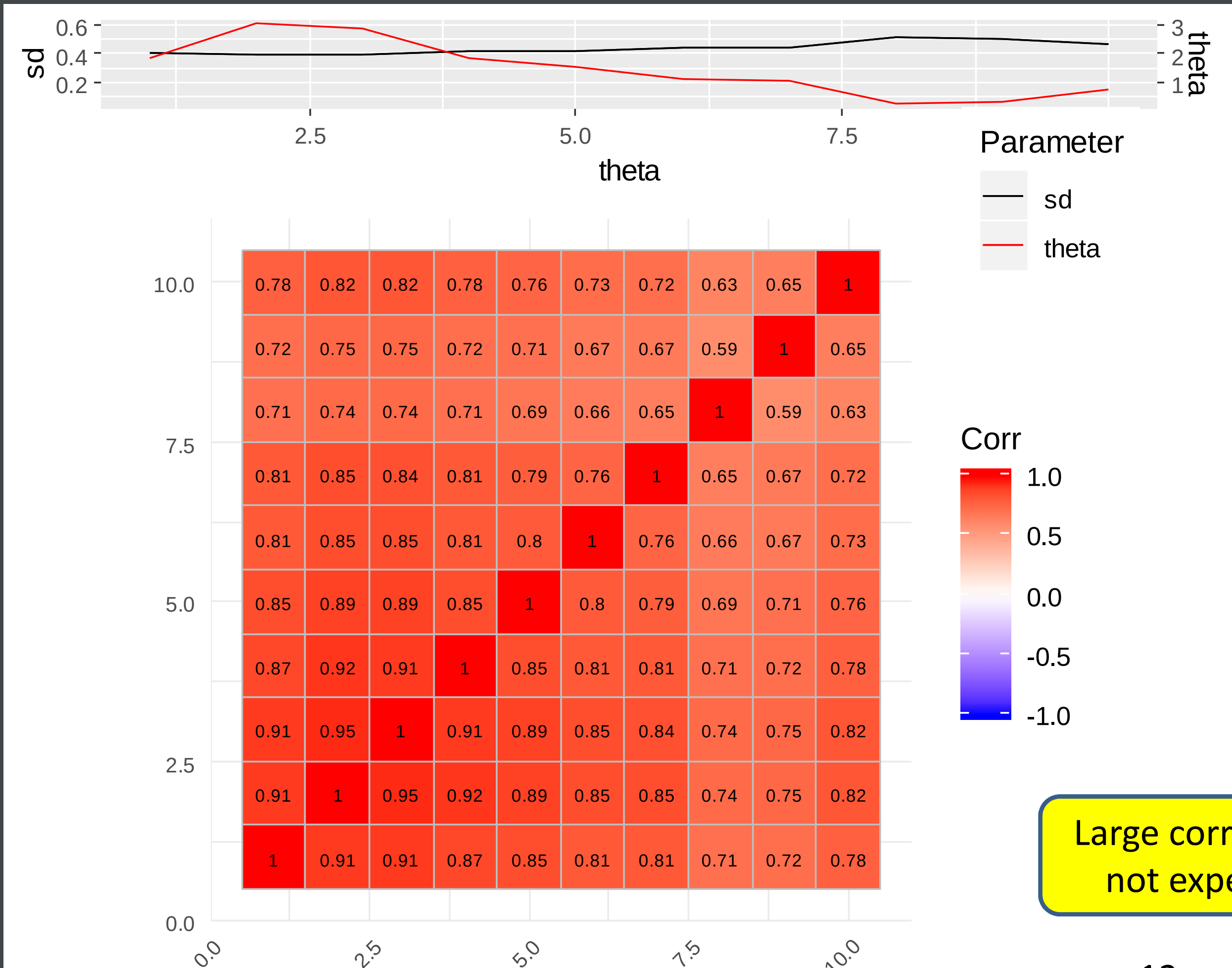
Case 2:

```
c=c(45,147,124,44,32,21,20,9,10,15,7)
pv=c/sum(c)
lambda = 500;phi = lambda*pv
n.year=10000
```

All as expected!



Case 2: Additive Logistic



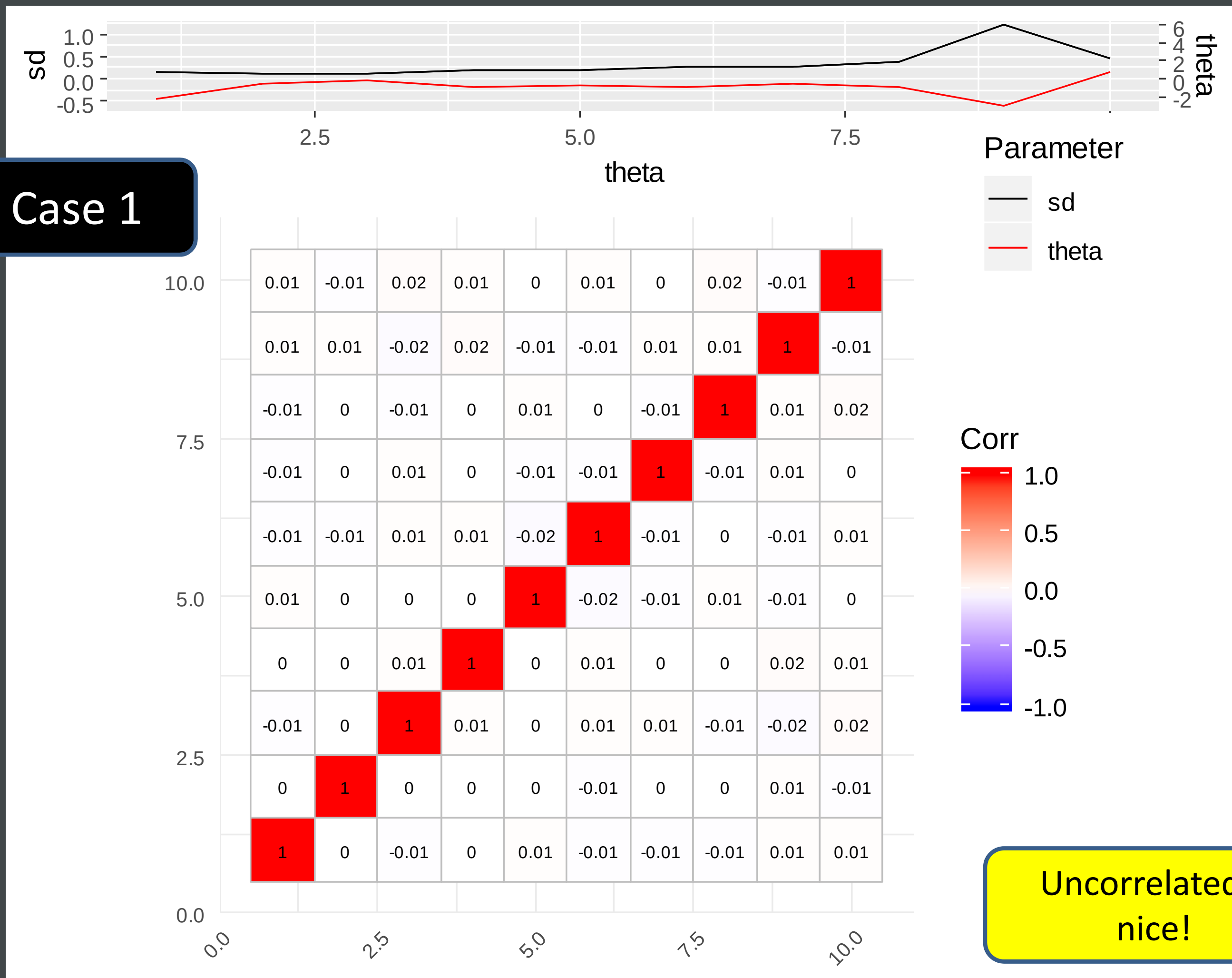
Large correlations not expected!

Conclusions for Additive LMVN distribution

- The additive LMVN approach seems problematic because of the complex correlation structures in the transformed proportions
- We would need to account for this somehow in an observation equation – and this seems messy
- Put in the next-gen trash bin??

Multiplicative Logistic $\theta_b = log\left(\frac{Y_b}{Y_{b+1}+,\ldots,+Y_B}\right), b = 1, \ldots, B-1$

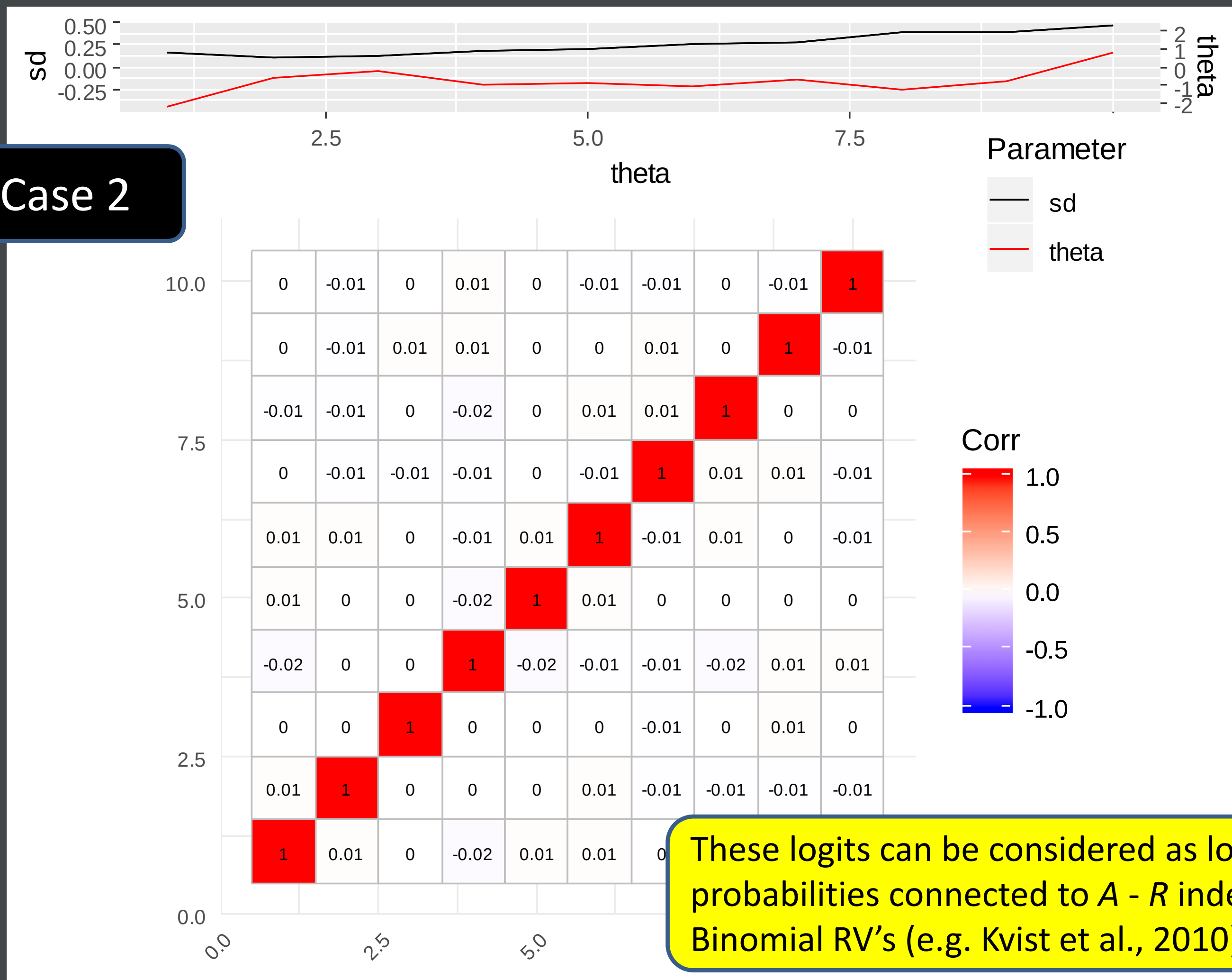
Case 1



Uncorrelated – nice!

Multiplicative LNM $\theta_b = log\left(\frac{Y_b}{Y_{b+1}+,\ldots,+Y_B}\right), b = 1, \ldots, B-1$

Case 2



These logits can be considered as logits of probabilities connected to *A* - *R* independent Binomial RV's (e.g. Kvist et al., 2010)

Conclusions for Multiplicative LMVN distribution

- More simple correlation structure in CRLs
- Cadigan (2016) and Varky et al used CRLs
- Albertsen et al (2016) did not favor this for SSMs…
- What about 0's? A big problem for some stocks
- Mult-LMVN not using age comp sample sizes??
- A possible next-gen option?

Likelihood for catch-at-age counts: Oversimplified Basis

- Assume that the sampled age frequencies are $X_1,\ldots,X_A$ and
- the total sample size is $X_+ = X_1 + \ldots + X_A$.
- If $X_a \overset{iid}{\sim} Poi(\lambda P_a)$ then most statisticians will agree that the relevant distribution for inferences about $P_1,\ldots,P_A$ is based on the conditional distribution of $X_1,\ldots,X_A$ given X_+
- and this is Multinomial with probabilities $P_1,\ldots,P_A$ and total sample size X_+
- Distribution of $X_1,\ldots,X_A|X_+$ does not depend on λ.

Slightly-less Oversimplified Basis

- The Poisson assumption will usually not be appropriate

- If $X_a|\gamma_a \sim Poi(\lambda P_a \gamma_a)$ and γ_a are iid Gamma random variables with mean 1 and variance k^{-1}.

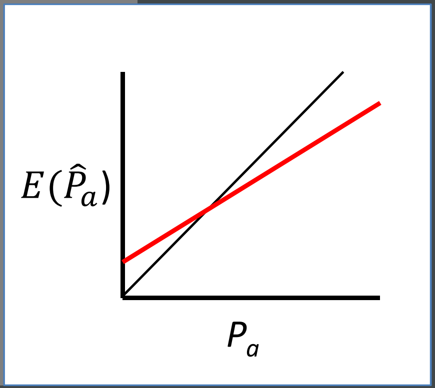
- The <u>marginal</u> distribution of X_a is $\mathrm{NB}(\mu_a, k)$ where $E(X_a) = \mu_a = \lambda P_a$ and $Var(X_a) = \mu_a + \mu_a^2/k$.

- However, the separation of information about $P_1,\ldots,P_A$ in $f(X_1,\ldots,X_A\,|X_+;P_1,\ldots,P_A)$ and λ in $f(X_+;\lambda)$ no longer occurs for NB counts

- The distribution of $f(X_1,\ldots,X_A\,|X_+)$ still depends on λ and k as well as $P_1,\ldots,P_A$. <u>The conditional of the marginal difficult</u>!

NB Counts

- $\hat{P}_a = X_a/X_+$ no longer an unbiased estimate of P_a
- Bias is complicated and depends on k and the difference between X_+ and λ.
- NB over-dispersion acts like covariate measurement error in regression
- There is bias attenuation towards 1/A regardless of P_a values
- Large sample sizes do not remove this bias



More Realistic Basis

- If $\mu_a = \lambda P_a \gamma_a$ then
  - $log(\mu_a) = log(\lambda) + log(P_a) + log(\gamma_a)$ and
  - $log(\gamma_a)$ is like an error term, with $\mathrm{E}\{log(\gamma_a)\} \approx 0$ since $\mathrm{E}(\gamma_a) = 1$.
- The log-gamma distribution ≈ Normal distribution.
- Think of $log(\gamma_a)'s$ as additive error terms.
- and $log(\gamma_1), \ldots, log(\gamma_A)$ ~ MVN with correlation.
- A log-Gaussian Cox Process (LGCP)

Conditional LGCP

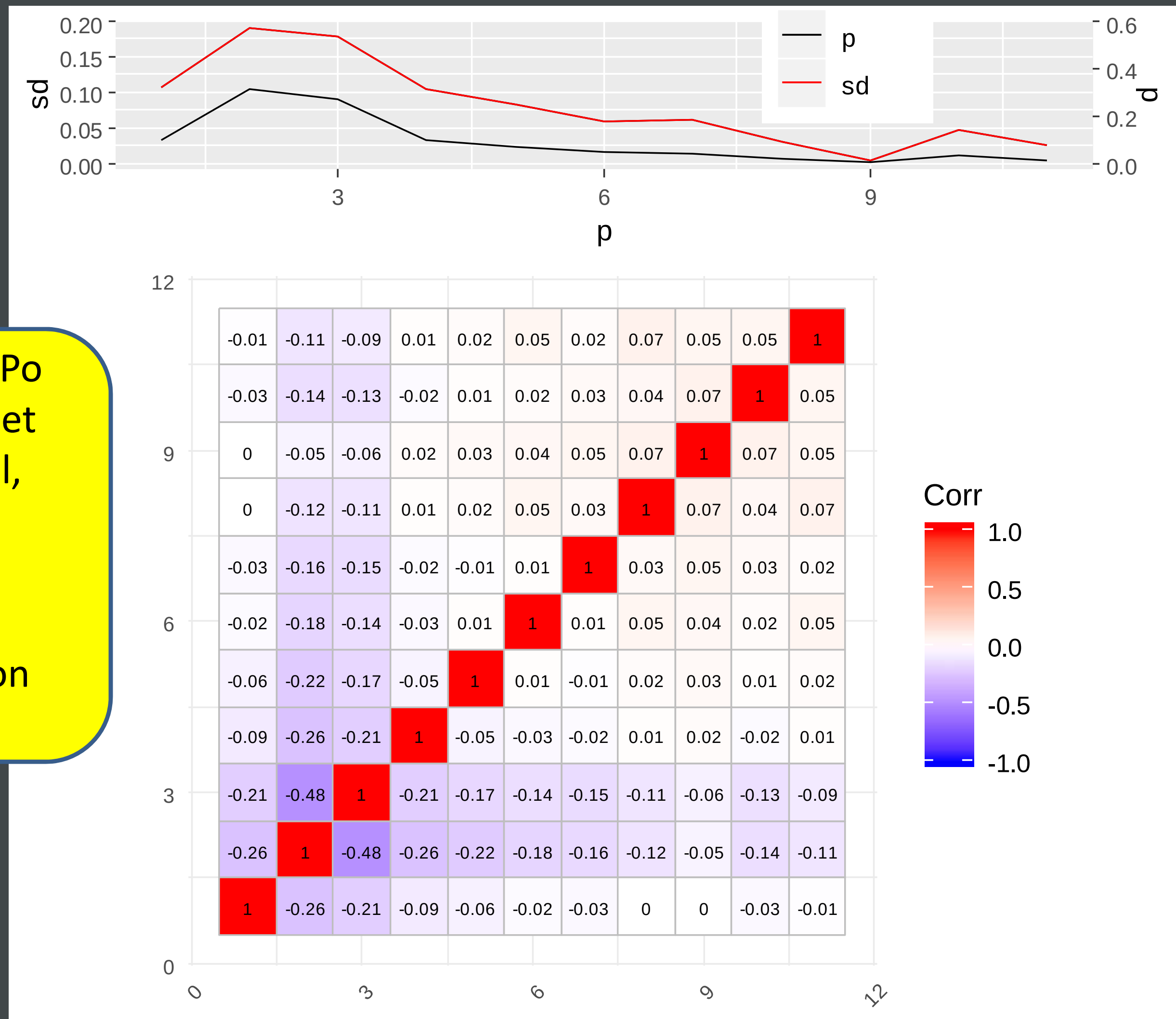
- sample "observations", $P_{oa} = X_{oa}/X_{o+}$.
- $f_1(P_{o1}, \dots, P_{oA} | \{X_{o+}, \gamma_1, \dots, \gamma_A\}, \theta) = \frac{\Gamma(X_{o+}+1)}{\prod_{a=1}^{A} \Gamma(X_{oa}+1)} \prod_{a=1}^{A} \left(\frac{P_a \gamma_a}{\sum_a P_a \gamma_a}\right)^{X_{oa}}$
- $log(\gamma_1), \dots, log(\gamma_A) \sim MVN(0, \Sigma)$
- Condition first, then get marginal $f_1(P_{o1}, \dots, P_{oA} | X_{o+}, \theta)$ and use this as the observation equation (i.e. nll)
- This is technically wrong but useful
- Right: marginal first, then conditional. This is less useful

Conditional LGCP

- Conditional LGCP is different from the Multinomial (MN) and Dirichlet-Multinomial (DMN)
- Both in the variance structure and correlations
- Cond-LGCP does not have all –corr’s like MN and DMN

Cov Po

Case 1

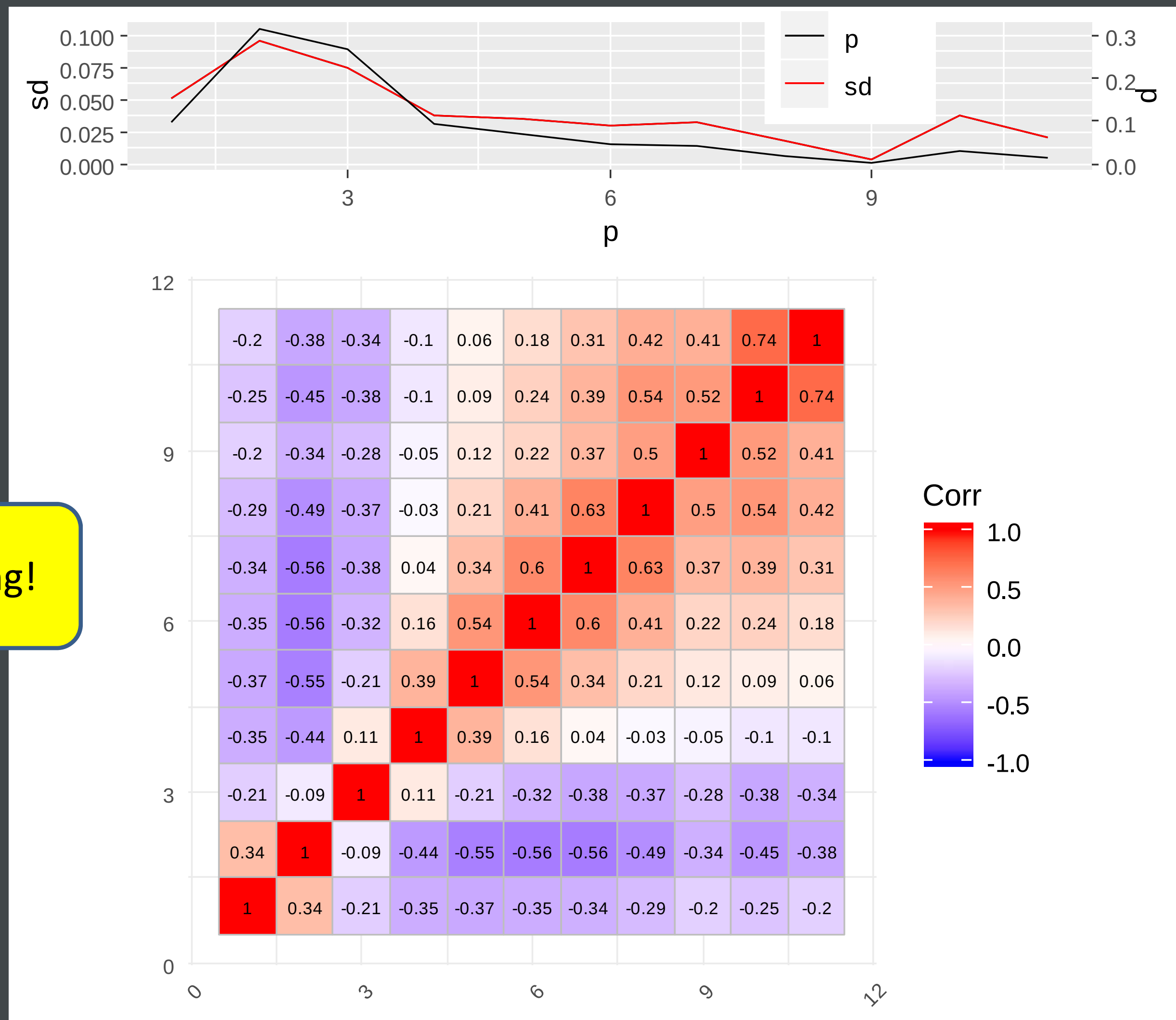


Correlation in Po not like Dirichlet or Multinomial, even when **ar1.phi = 0**

Just the Poisson OD does this

Cov Po

Case 2 ar1.phi = 0.9



Interesting!

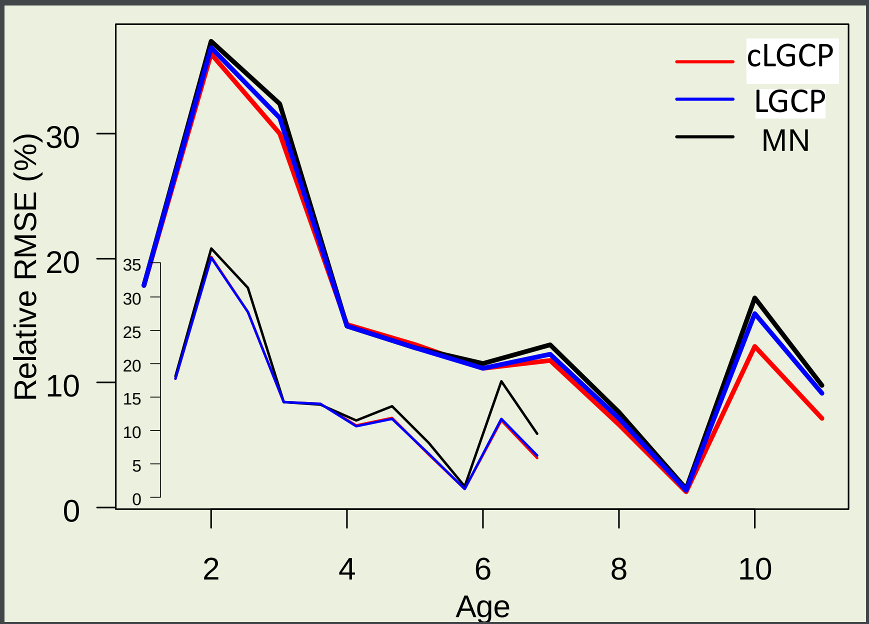
LGCP

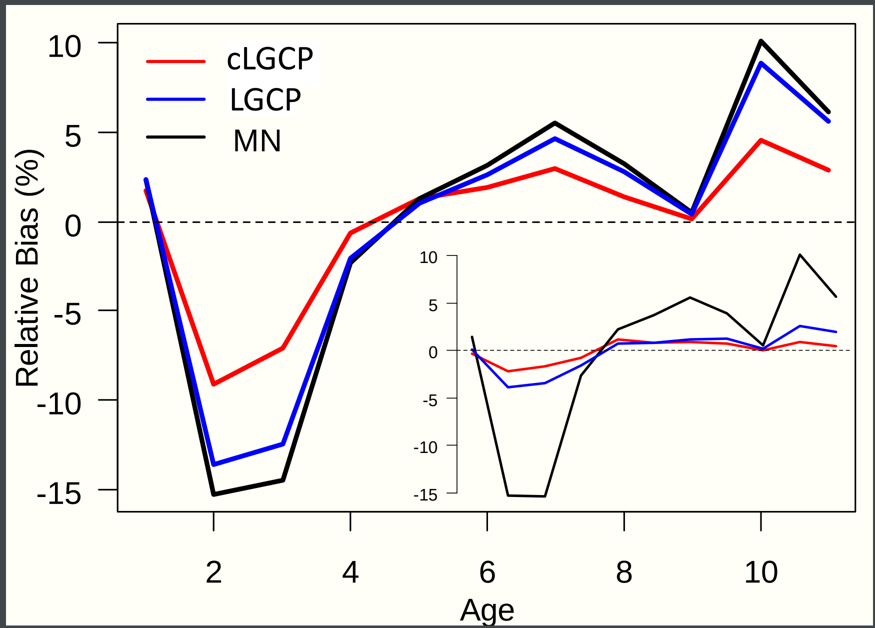
- Could also use the LGCP directly, $X_a|\gamma_a \overset{iid}{\sim} Poi(\lambda P_a \gamma_a)$ (i.e. not conditional) but need a different λ for each sample (year, gear, etc)
- What happens when our guess about sample size is wrong?

Error Low in Sample Size

ar1.phi = 0.9
sigmaP = 1

- **Assumed Xo+=50 (true X+=250), each of 10 years**





Conclusions

- Include uncertainty in landings in next gen
- Seems better to fix age comp sample sizes low with conditional LGCP approach
- multiplicative LMVN did not work well with LGCP counts (*preliminary*)
- Needs testing within stock assessment model framework
- Additive LMVN, MN, DMN => next gen trash bin??