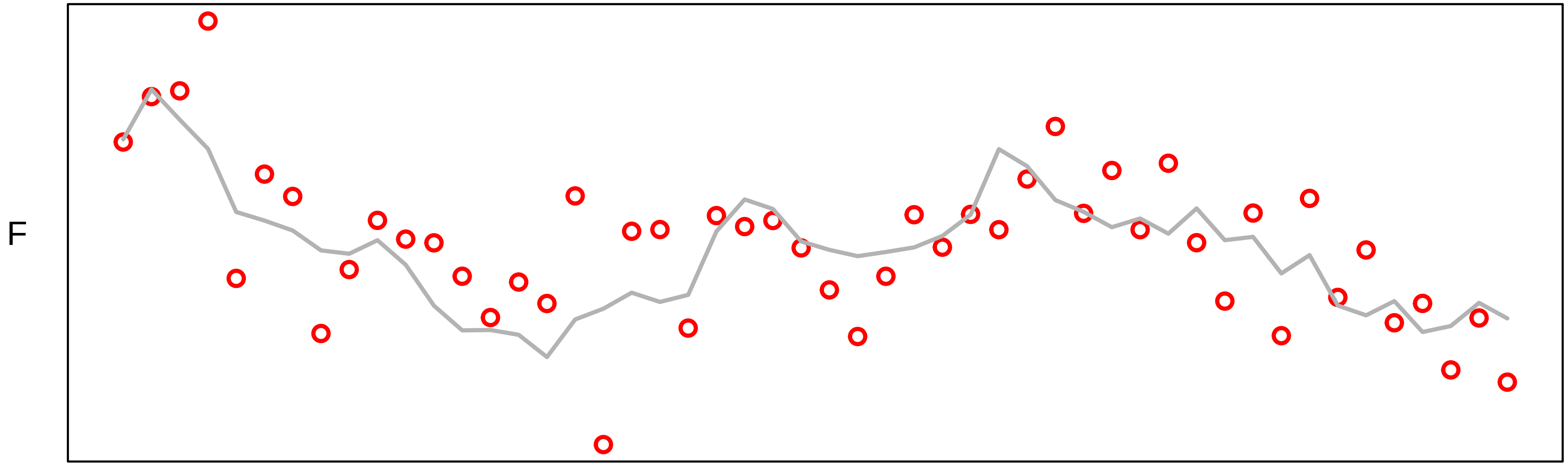
A basic statistical assessment model

Anders Nielsen, Ethan Lawler, & Sean Anderson

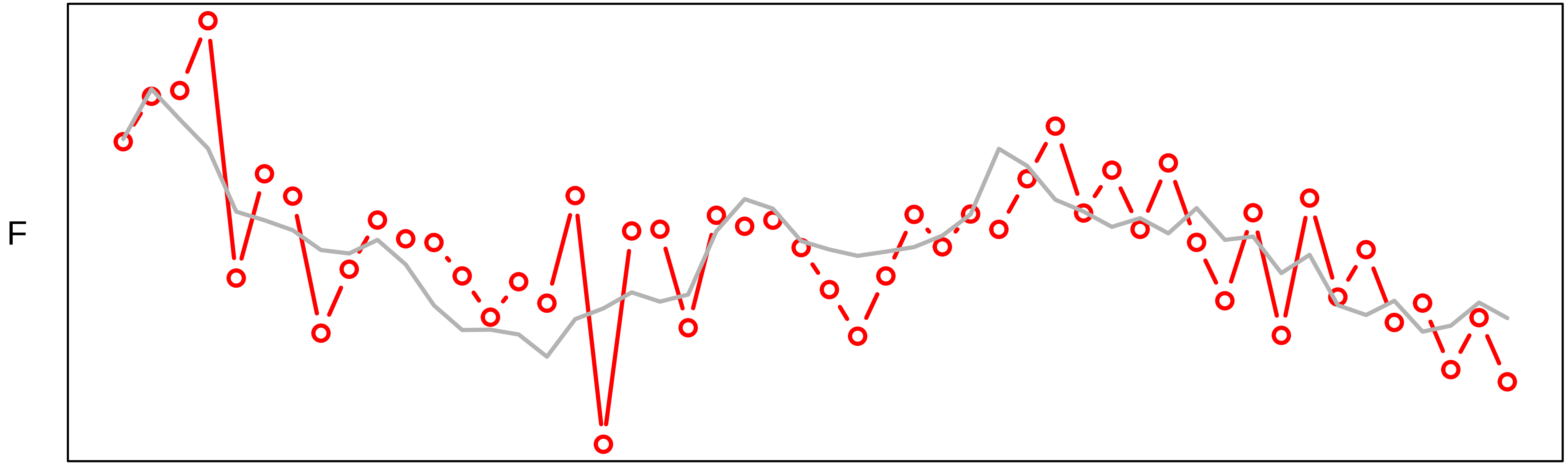
an@aqua.dtu.dk

The problem in a nutshell



- Consider this example:
  - The true underlying fishing mortality (here grey) we can never observe
  - We only observe Y (here red circles) which is $F+$'noise'
  - The key question is: How do we estimate/predict the gray process?

Deterministic model estimates



- If we assume no observation error the estimate of F is Y
- Too fluctuating
- No quantification of uncertainties
- Does this remind you of anything?

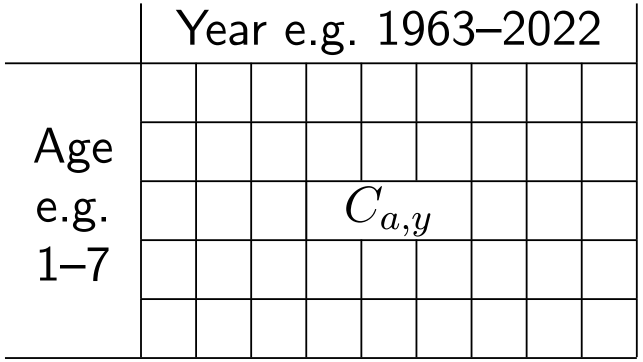
Fish Stock Assessment

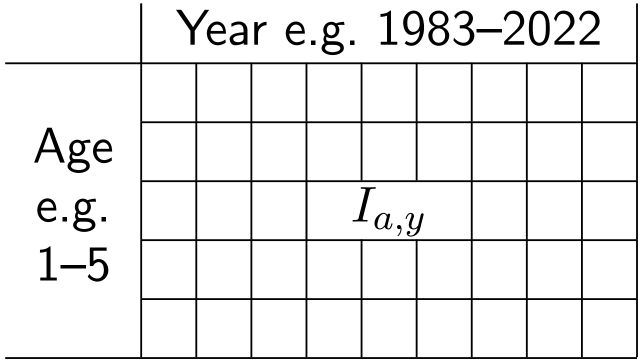
Problem: How many fish (relative or absolute) are left in the ocean?

Data:

$C_{a,y}$: Yearly catches (divided into age–classes)

$I_{a,y}$: Scientific surveys



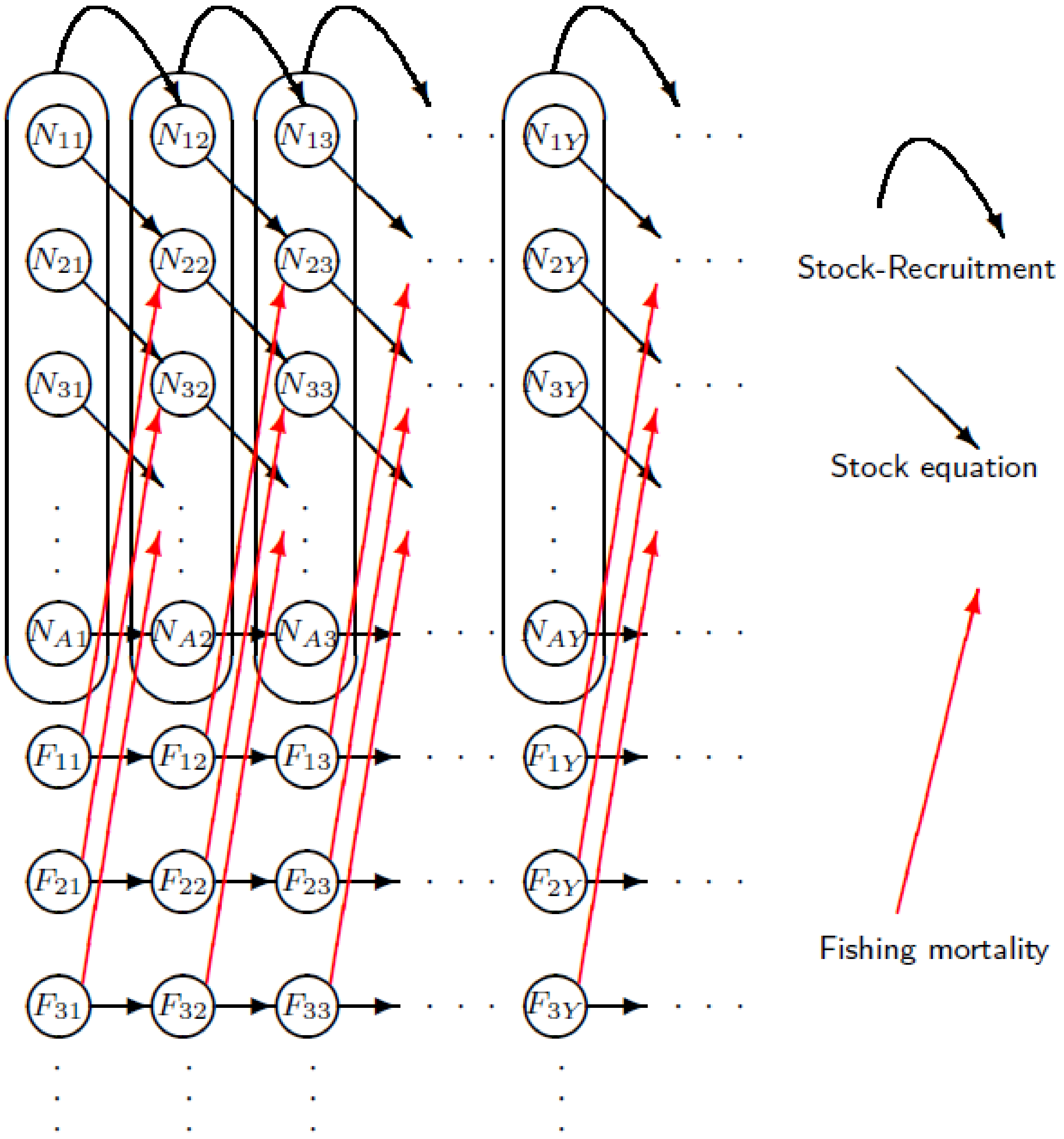


Often we have catches ($C_{f,a,y}$) from more than one fleet and indices ($I_{s,a,y}$) from more than one survey, but here we keep it simple.

Important output: "Reference points" SSB, $\overline{F}$, ...

Assessment Models

- Based on a standard set of equations describing the structure of the system
- Account for selectivity
- Account for different assumptions about 'natural' mortality
- can account for different effort numbers
- Can include different assumptions about recruitment
- All of these things have to be accounted for



Basic equations

Stock equation: The number of fish in a cohort is expected to follow:

$$\frac{dN_t}{dt} = -(F_t + M_t)N_t$$

If F and M are assumed constant within each year we get:

$$N_{a+1,y+1} = N_{a,y}e^{-(F_{a,y}+M_{a,y})}$$

Catch equation: The number of fish in a cohort after one year can be separated into:

$$N_{a,y} = \underbrace{N_{a+1,y+1}}_{\text{survived}} + \underbrace{C_{a,y} + D_{a,y}}_{\text{died}}$$

The part of the cohort that was caught during the year is:

$$C_{a,y} = \frac{F_{a,y}}{F_{a,y} + M_{a,y}}\left(1 - e^{-(F_{a,y}+M_{a,y})}\right)N_{a,y}$$

Stock–recruitment equation: Obviously connected — Different opinions about how

Calculation example

- Imagine that a cohort consists of $N_0 = 1000$ fish at time 0 and we know $M = 0.2$ and $F = 0.1$, then we can calculate

```
> N0 <- 1000; F <- 0.1; M <- 0.2
> N1 <- N0*exp(-F-M)
> C0 <- F/(F+M)*(1-exp(-F-M))*N0
```

- This gives $N_1 \approx 741$ survivors and a catch of 0 year olds of $C_0 \approx 86$
- We can continure the calculation and find the number surviving to age 2 and the catch of 1 year olds:

```
> N2 <- N1*exp(-F-M)
> C1 <- F/(F+M)*(1-exp(-F-M))*N1
```

- These calculations can be gathered in a loop — try it. Start with:

```
> N0<-1000; F<-.1; M<-.2
> N<-c(N0, rep(NA,19))
```

- In addition what is the cohort size at time 1.75?

Deterministic models

- A deterministic model is a model where observation noise is ignored
- Typically catches are assumed known without error
- Most commonly applied fish stock assessment models are (semi-)deterministic
- These algorithms work (very simply put) by:

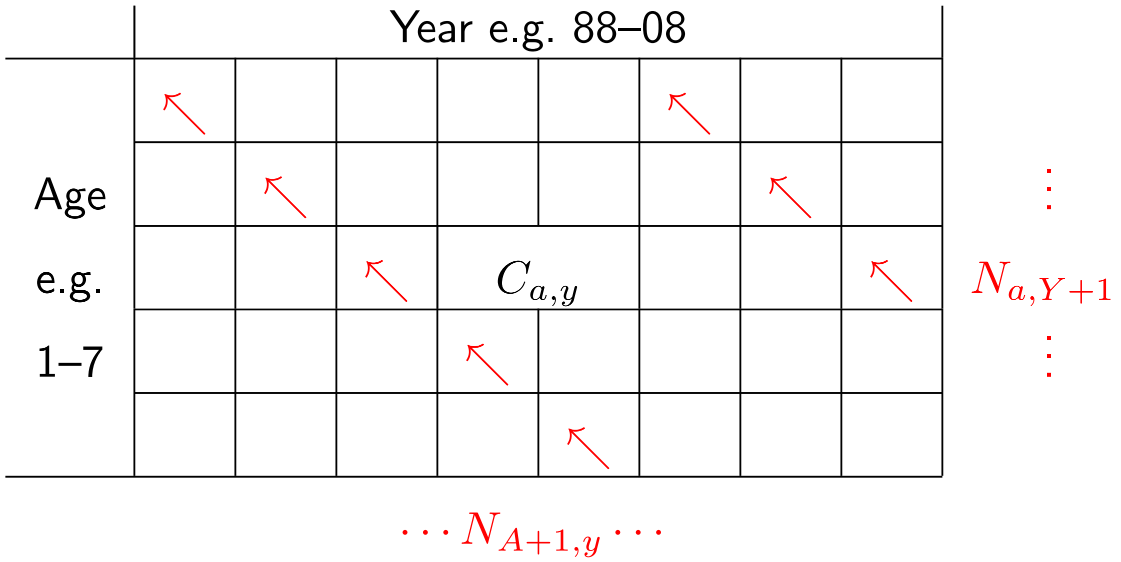
  0: Guess the number of survivors $N_{A+1,y}$ and $N_{a,Y+1}$

  1: Back calculate ($\nwarrow$) all $N_{a,y}$ by subtracting catch and natural mortality

  2: Use surveys to adjust all $N_{a,y}$ and update survivors accordingly

  3: Repeat 1-3 until survivors converge

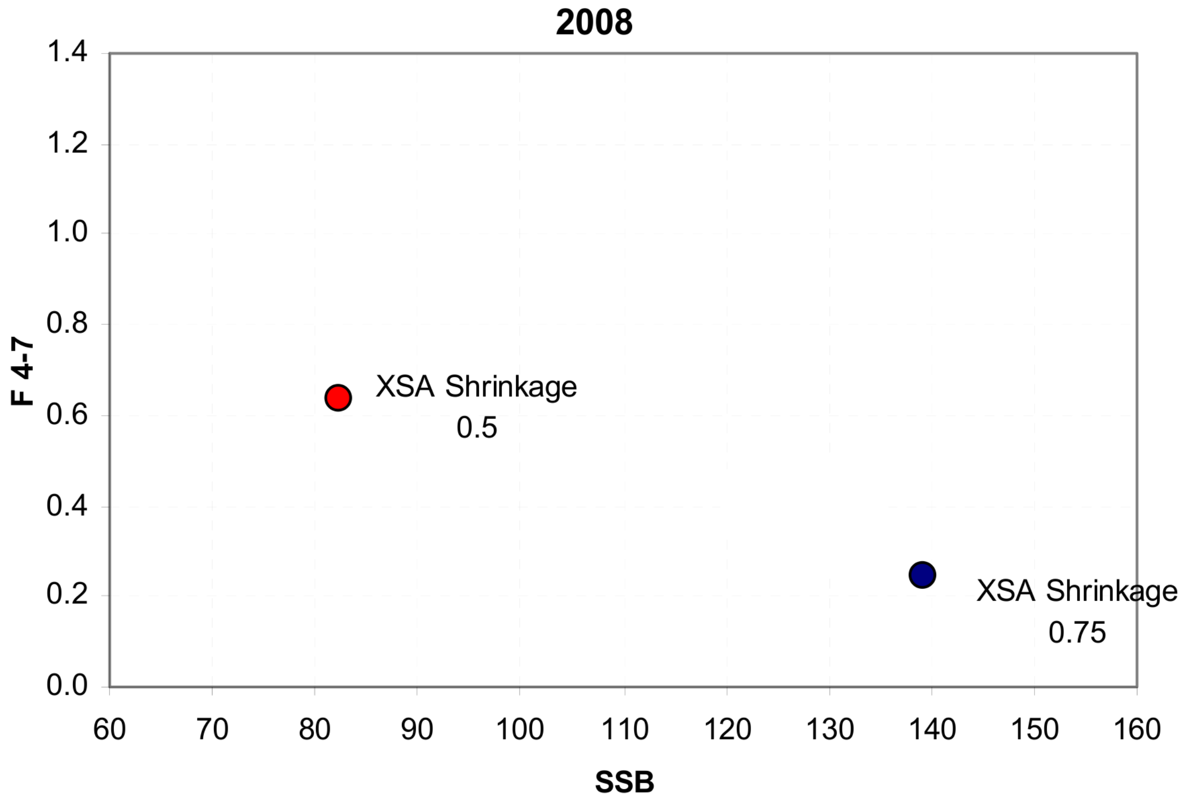
- Doing 0-1 just ones is known as Virtual Population Analysis



Features of deterministic models

+ Super fast to compute
+ Fairly simple to explain the path from data to stock numbers (especially VPA)
– Difficult to explain why it works (converges), and what a solution mean
– These algorithms contain many ad-hoc settings (shrinkage, tapered time weights, ...) that makes them less objective
– No quantification of uncertainties within model
? What exactly is the model
  - The assumptions are difficult to identify and verify
  - With no clearly defined model more ad-hoc methods are needed to make predictions
- No framework for comparing models (different settings)

Example: F-shrinkage for Eastern Baltic Cod



- These differences are not small and theoretical
- There are no objective way to choose between these two deterministic approaches
- Things would be simpler if we had a statistical model

Now we know

- **Assessment** Basic equations:
  - Stock equation: The number of fish in a cohort is expected to follow:

$$N_{a+1,y+1} = N_{a,y} e^{-(F_{a,y}+M_{a,y})}$$

  - and further if we need the stock size some fraction τ into a year

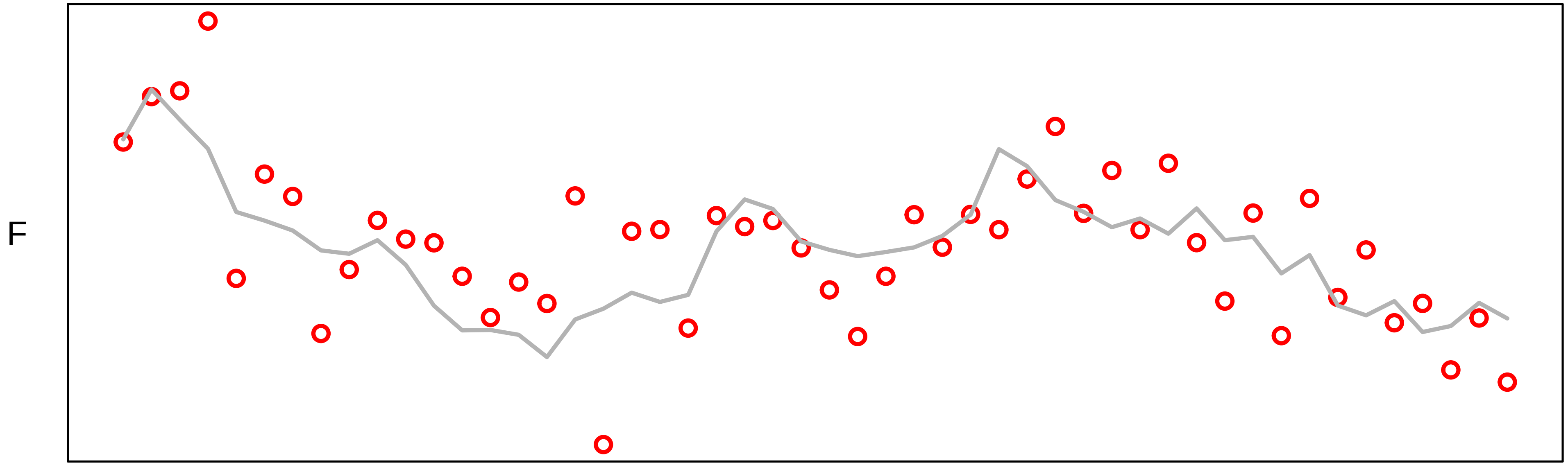
$$N_{a+\tau,y+\tau} = N_{a,y} e^{-(F_{a,y}+M_{a,y})\tau}$$

  - Catch equation: The part of the cohort that was caught during the year is:

$$C_{a,y} = \frac{F_{a,y}}{F_{a,y}+M_{a,y}} \left(1 - e^{-(F_{a,y}+M_{a,y})}\right) N_{a,y}$$

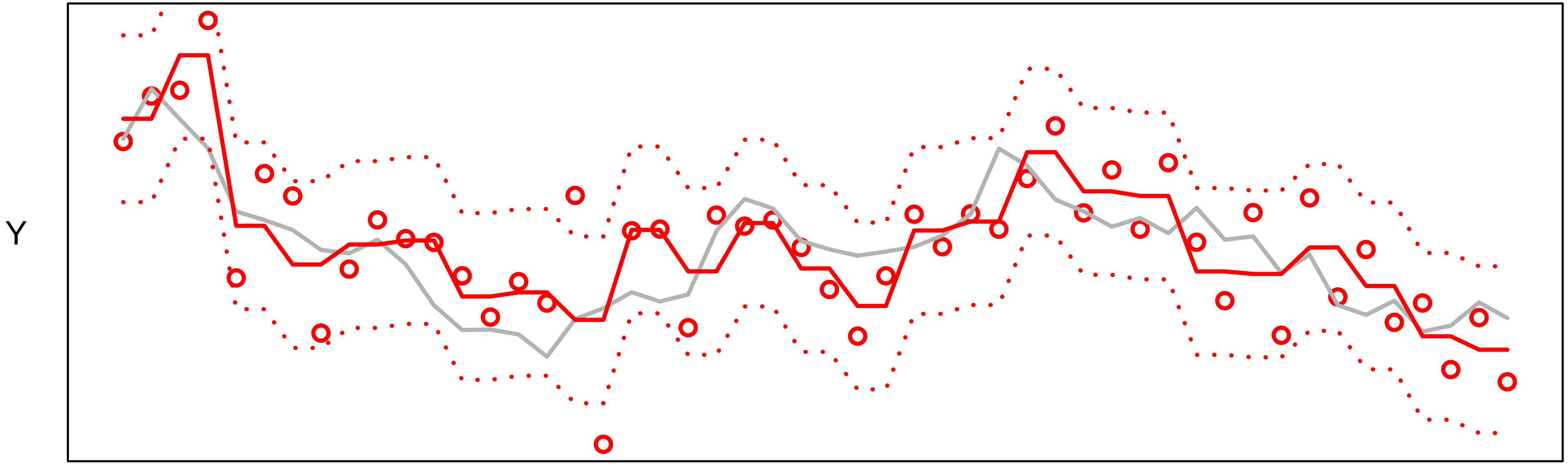
- **Deterministic assessment models** What is the main assumption in a deterministic model?

Allowing observation noise



- How do we:
  - Estimate model parameters?
  - Estimate uncertainties?
  - Compare models?

Model with observation noise



- It allows for observation noise
- We first had to group the observations (here pairs)
- Choice is arbitrary
- The reconstructed track appear OK
- The model contains a lot (26) of model parameters
- Uncertainties are estimated but the confidence interval seems too wide

Now we know

- **Assessment** Basic equations:
  - Stock equation: The number of fish in a cohort is expected to follow:

$$N_{a+1,y+1} = N_{a,y}e^{-(F_{a,y}+M_{a,y})}$$

  - and further if we need the stock size some fraction τ into a year

$$N_{a+\tau,y+\tau} = N_{a,y}e^{-(F_{a,y}+M_{a,y})\tau}$$

  - Catch equation: The part of the cohort that was caught during the year is:

$$C_{a,y} = \frac{F_{a,y}}{F_{a,y}+M_{a,y}}\left(1-e^{-(F_{a,y}+M_{a,y})}\right)N_{a,y}$$

- **Deterministic assessment models** Assume catches are known without error
- **Statistical/stochastic assessment models** Allow observation noise:
  - Explicit about assumptions w.r.t. observation noise
  - Complete solution for: estimation, uncertainty, model comparison

A full parametric statistical model

- The log catches are assumed to follow:

$$\log(C_{a,y}) \sim \mathcal{N}\left(\log\left(\frac{F_{a,y}}{Z_{a,y}}(1-e^{-Z_{a,y}})N_{a,y}\right), \sigma_c^2\right) \;, \text{ where}$$

$$F_{a,y} = F_y F_a, \text{ with } F_{a=5} = F_{a=6} = F_{a=7} = 1, \text{ and } Z_{a,y} = F_{a,y} + M_{a,y}$$

- The log catches from the survey are assumed to follow:

$$\log(I_{a,y}) \sim \mathcal{N}\left(\log\left(Q_a e^{-Z_{a,y}T} N_{a,y}\right), \sigma_s^2\right) \;, \text{ where}$$

  T is the fraction into the year where the survey is taken, and Q_a is catchability parameter.

- The stock sizes are assumed to follow:

$$N_{a,y} = N_{a-1,y-1} e^{-Za-1,y-1}$$

  Notice that it does not define N in the first year and for the youngest age.

- So the model parameters are the undefined N's, F_y, F_a, Q_a, σ_c, and σ_s

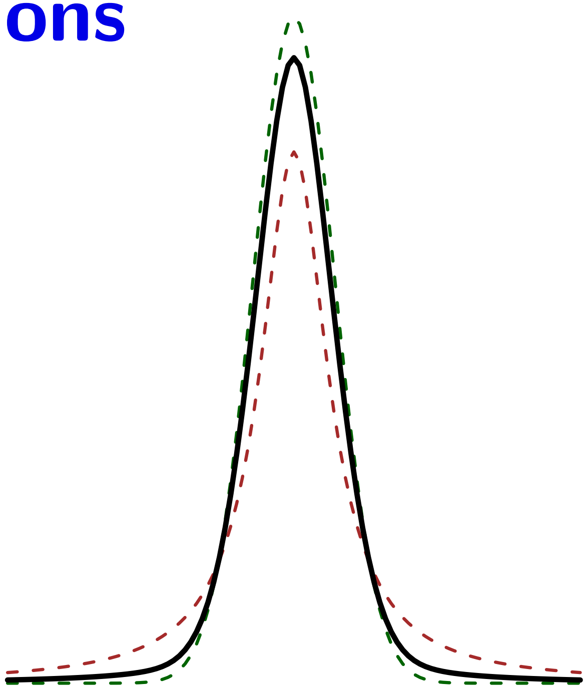
A full parametric statistical model — implementation

- Data is in `fsa.RData` (let's study the content)
- An implementation in $\star$RTMB$\star$ is in `basicfsa.R`
- The code is short, so lets identify the parts about N, F, C, and I

Exercise 3.1: Weighting of information sources

- Weighting to different data sources is a concept often used and often debated in fish stock assessment
- If the likelihood is correct, then the models should be self-weighting
- But imagine that we wanted to give double weight to our survey
- How could we achieve that by changing the code?
- Try two approaches:
  - Double the survey observations
  - Fix the variance to be half of the estimated variance for the survey observations
- Hint: Notice that the parameter is the log of the standard deviation. The variance is the standard deviation squared
- Take home: The weight is the inverse of the variance

Exercise 3.2: Using robust distribution for observations



- The normal distribution is fairly sensitive to outliers
- Standard model uses pdf: $\phi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}$
- Robust model will use pdf:

$$\left((1-p)\phi\left(\frac{x-\mu}{\sigma}\right)+p\psi\left(\frac{x-\mu}{\sigma}\right)\right)\frac{1}{\sigma}$$

  , where the p is an input (not estimated)

- Where ϕ is pdf of N(0,1) and ψ is pdf of heavy-tailed distribution (e.g. t_3).
- **Exercise:** In the simple assessment mode replace 5 (or 10) observations with huge outliers. The run the assessment as is and compare to replacing with the model where the robust density is used for the observations.