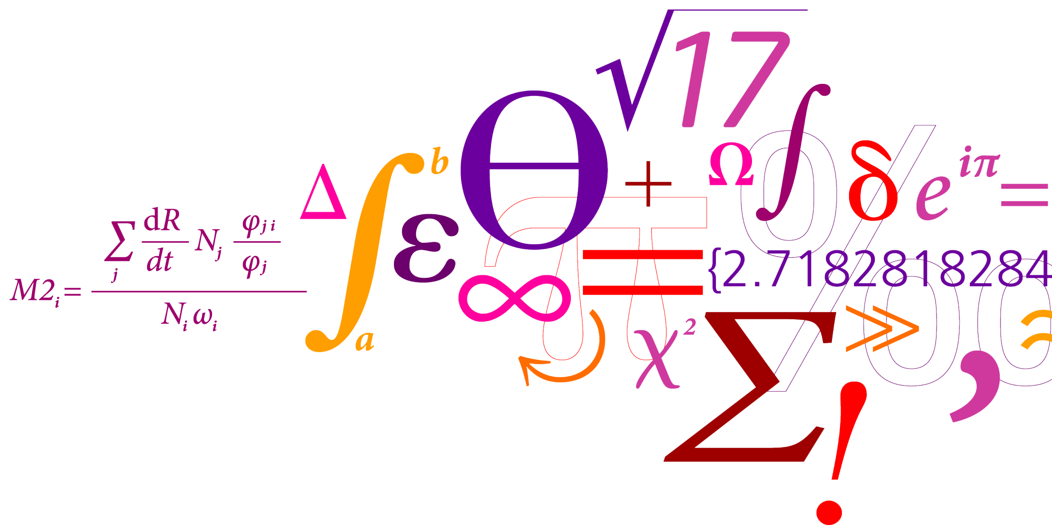
State–space fish stock assessment model as simple alternative to (semi-) deterministic approaches and full parametric stochastic models

Anders Nielsen & Casper W. Berg
an@aqua.dtu.dk

DTU Aqua
National Institute of Aquatic Resources



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:

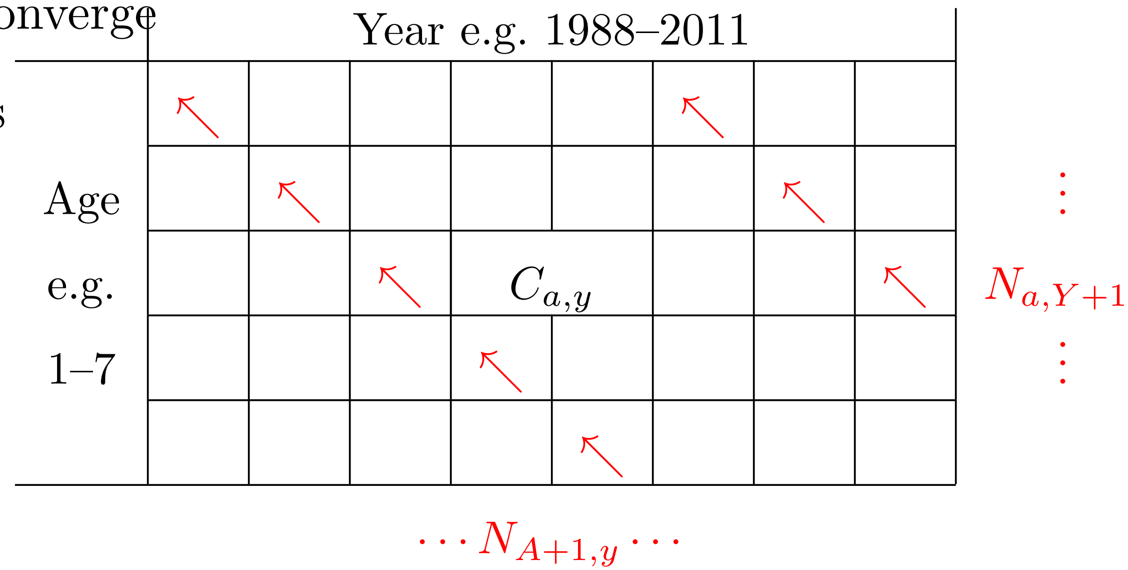
  0: Guess the number of survivors $N_{A+1,y}$ and $N_{a,Y+1}$

  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

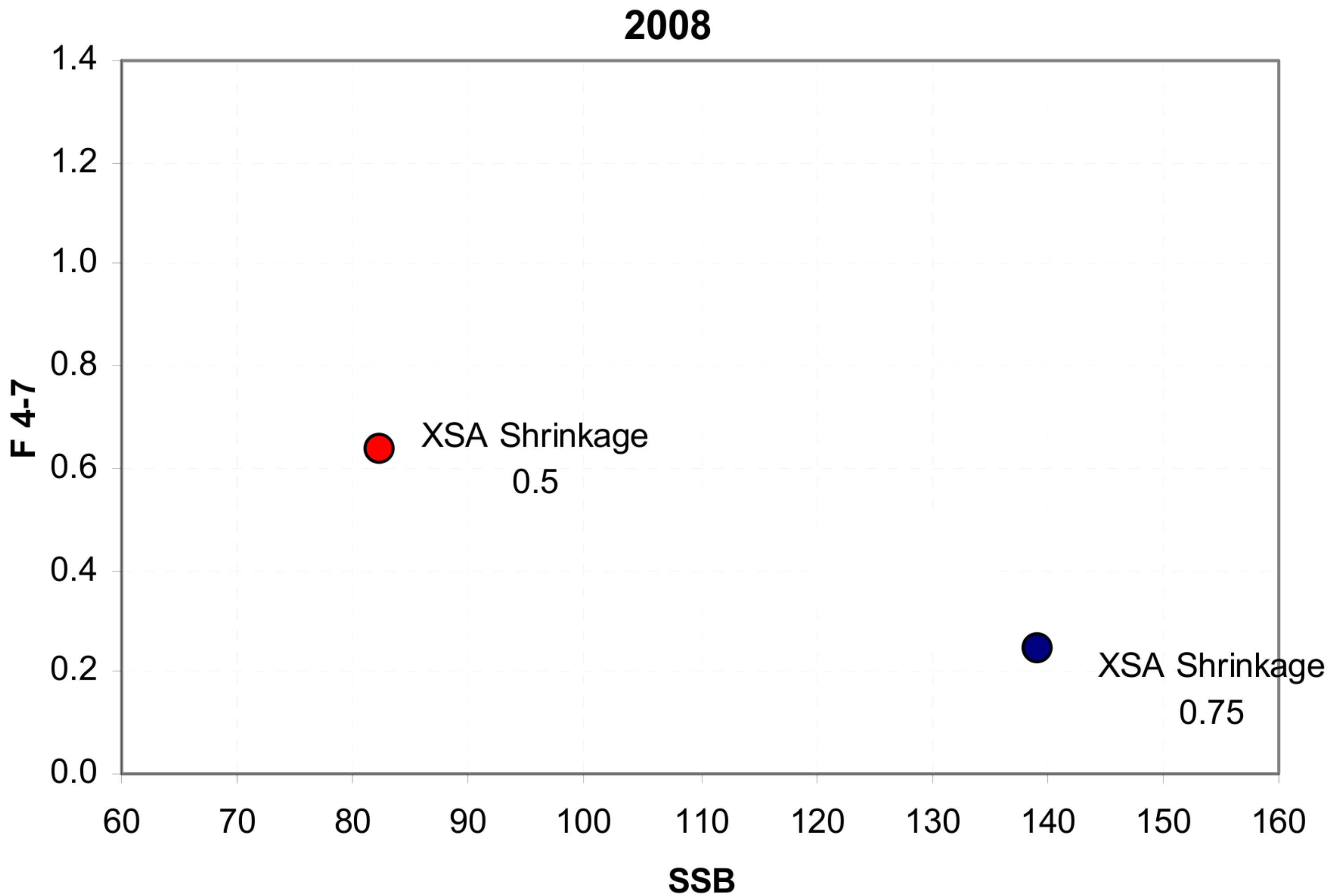
– These algorithms contain many ad-hoc settings (shrinkage, tapered time weights, ...) that makes them less objective

– 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

A full parametric statistical model

- The log catches are assumed to follow:

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

$$F_{a,y} = f_y S_a, \text{ with } S_{a=5} = S_{a=6} = S_{a=7} = 1, \text{ and } Z_{a,y} = F_{a,y} + M_{a,y}$$

- 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, S_a, Q_a, σ_c, and σ_s

Fully parametrized statistical assessment models

- A statistical[a] model acknowledges observation noise
- The error structure is part of the model description
- To find the quantities of interest (e.g. $N_{a,y}$, $F_{a,y}$, and observation uncertainties) the likelihood of the actual observations is optimized w.r.t. the model parameters.
- Parametrized statistical assessment models have a number of benefits:
  + All model assumptions are transparent
  + Different model assumptions can be tested against each other (e.g. is $F_5 = F_6$?)
  + Different data sources can be included and correctly and objectively weighted
  + Estimation of uncertainties are an integrated part of the model
- But also a few difficulties:
  - Trade-off between the number model parameters and flexibility of the model (e.g. $F_{a,y}$ vs. $F_{a,y} = S_a f_y$)
  - More advanced software needed

---

[a] a.k.a. stochastic

Problems we wish to solve

- Deterministic approaches
  - Catch at age assumed known without error
  - Procedures not models
  - Convergence of a deterministic procedure
  - Ad-hoc adjustments
- Full parametric statistical models
  - Parametric F–structure (e.g. multiplicative)
  - Trade off between flexible with (too) many parameters and rigid with tractable number of parameters
  - Number of parameters increase with every new year of data added

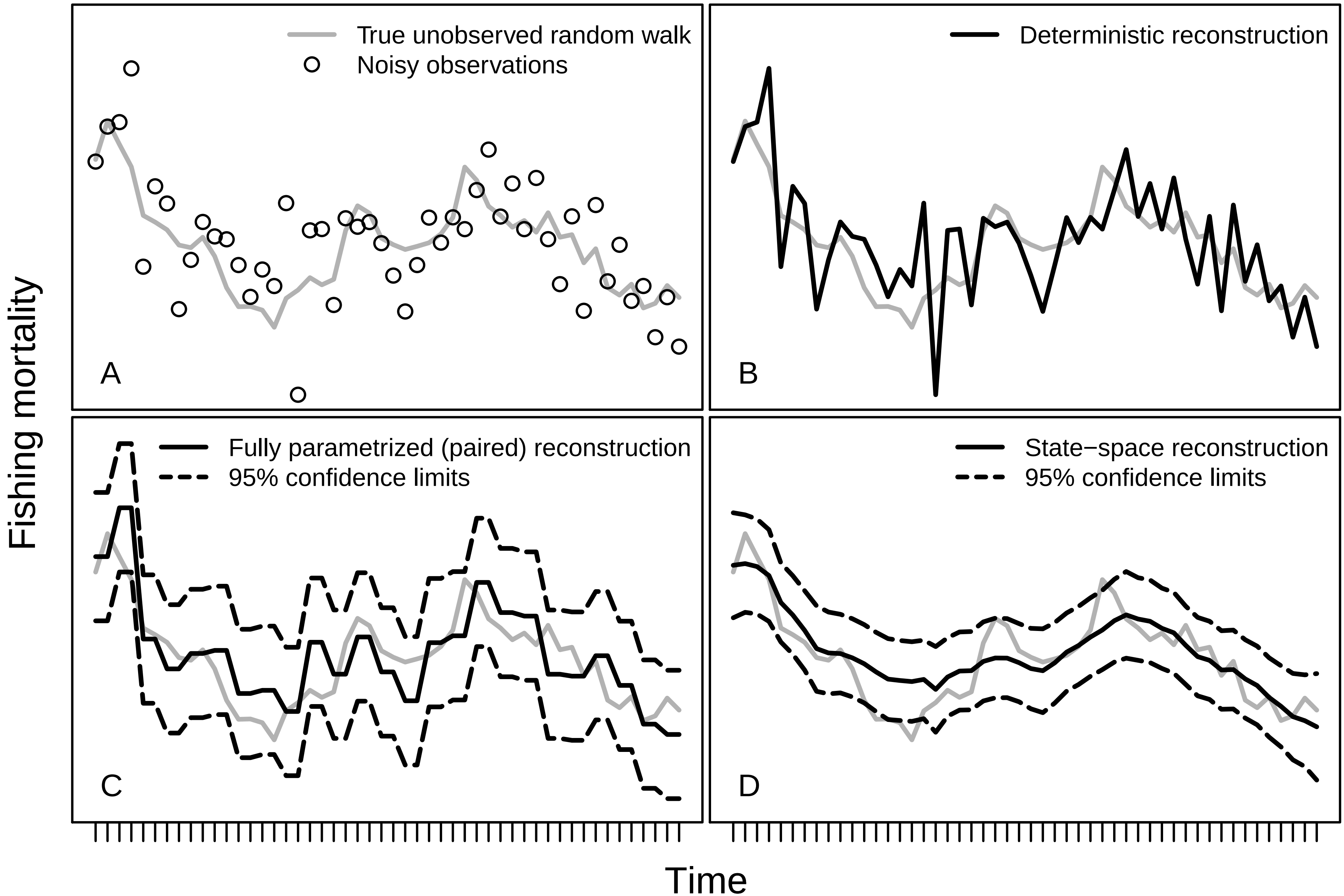
State-space assessment models

- This model class[a] is used in most other quantitative fields
- It is a very useful extension to full parametric statistical models.
- Introduced for stock assessment by Gudmundsson (1987,1994) and Fryer (2001)
- The reason state-space models have not been more frequently used in stock assessment is that software to easily handle these models has not been available
- Can give very flexible models with low number of model parameters
- For instance we can include things like:

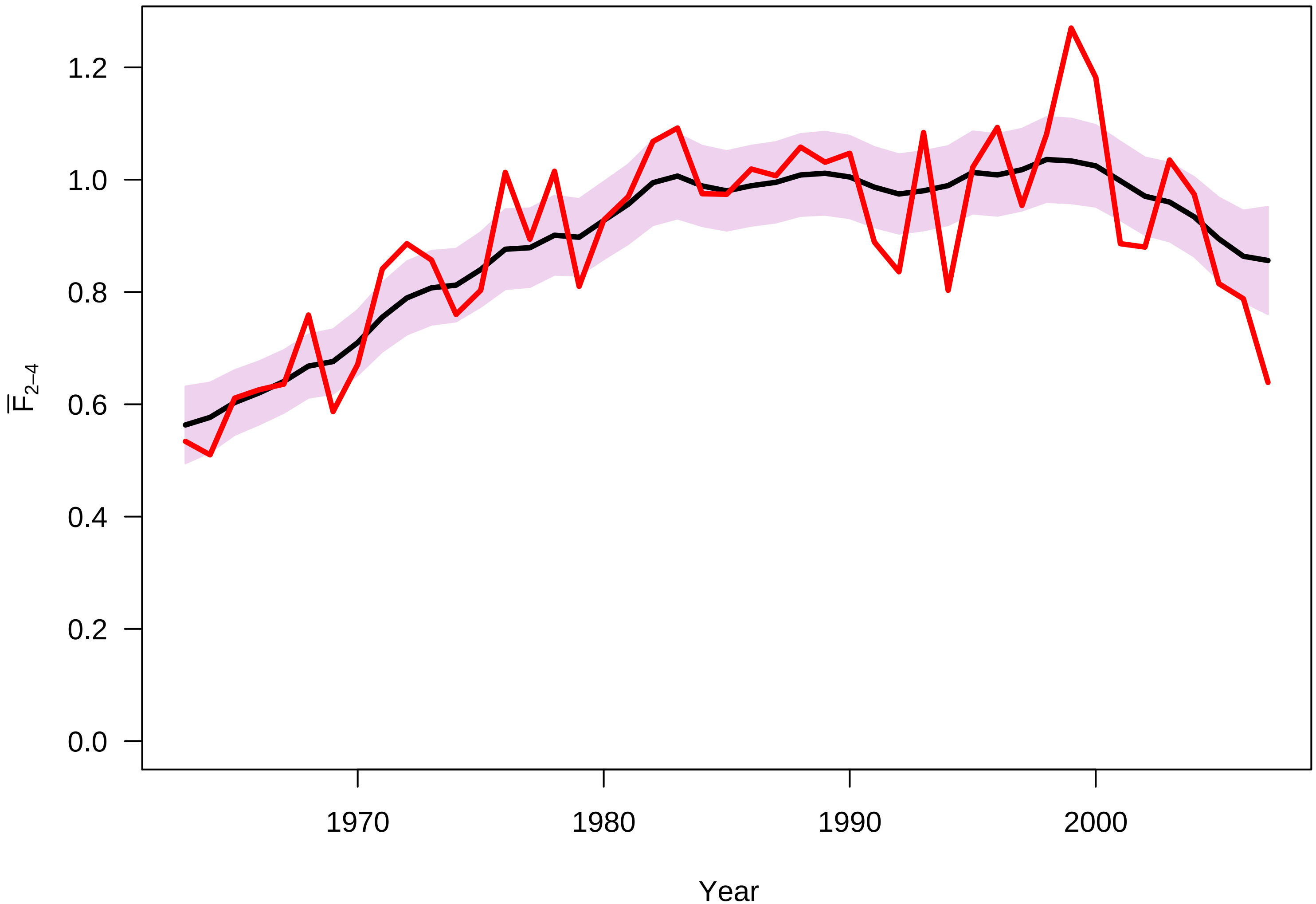
$$F_{3,y} \text{ is a random walk with yearly variance } \sigma^2$$

[a] a.k.a. random effects models, mixed models, latent variable models, hierarchical models, ...

Illustration of the three types of models



Example: $\overline{F}_{2-4}$ for North Sea Cod



Model

States are the random variables that we don't observe ($N_{a,y}$, $F_{a,y}$)

$$\begin{pmatrix} \log(N_y) \\ \log(F_y) \end{pmatrix} = T \begin{pmatrix} \log(N_{y-1}) \\ \log(F_{y-1}) \end{pmatrix} + \eta_y$$

Observations are the random variables that we do observe ($C_{a,y}$, $I_{a,y}^{(s)}$)

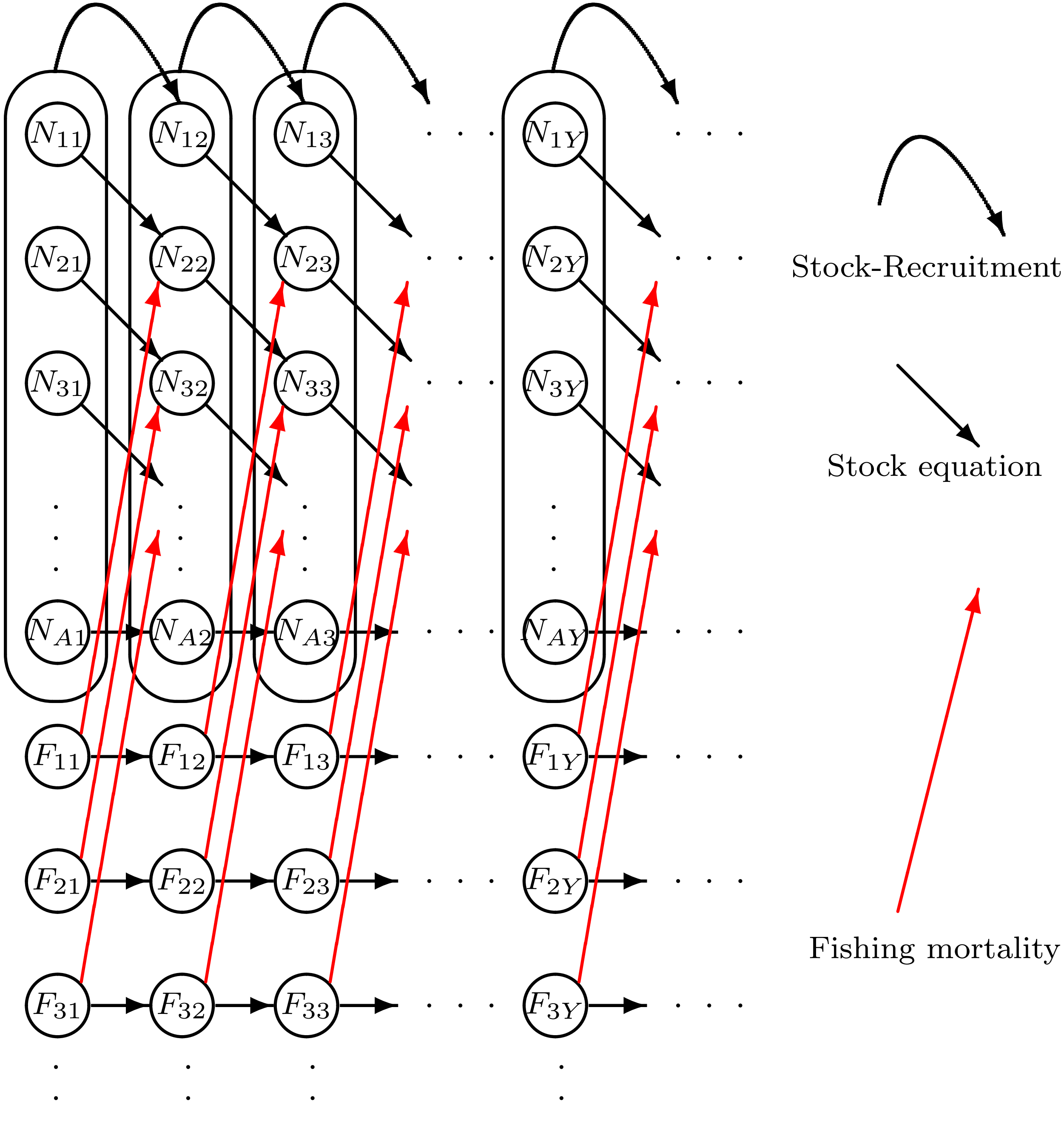
$$\begin{pmatrix} \log(C_y) \\ \log(I_y^{(s)}) \end{pmatrix} = O \begin{pmatrix} N_y \\ F_y \end{pmatrix} + \varepsilon_y$$

Model and parameters are what describes the distribution of states and observations through T, O, η_y, and ε_y.

Parameters: Survey catchabilities, S-R parameters, process and observation variances.

All model equation are as expected:

- Standard stock equation
- Standard stock recruitment (B-H, Ricker, or RW)
- Standard equations for total landings and survey indices



N_{11}
N_{12}
N_{13}
N_{1Y}
N_{21}
N_{22}
N_{23}
N_{2Y}
N_{31}
N_{32}
N_{33}
N_{3Y}
N_{A1}
N_{A2}
N_{A3}
N_{AY}
F_{11}
F_{12}
F_{13}
F_{1Y}
F_{21}
F_{22}
F_{23}
F_{2Y}
F_{31}
F_{32}
F_{33}
F_{3Y}
Stock-Recruitment
Stock equation
Fishing mortality

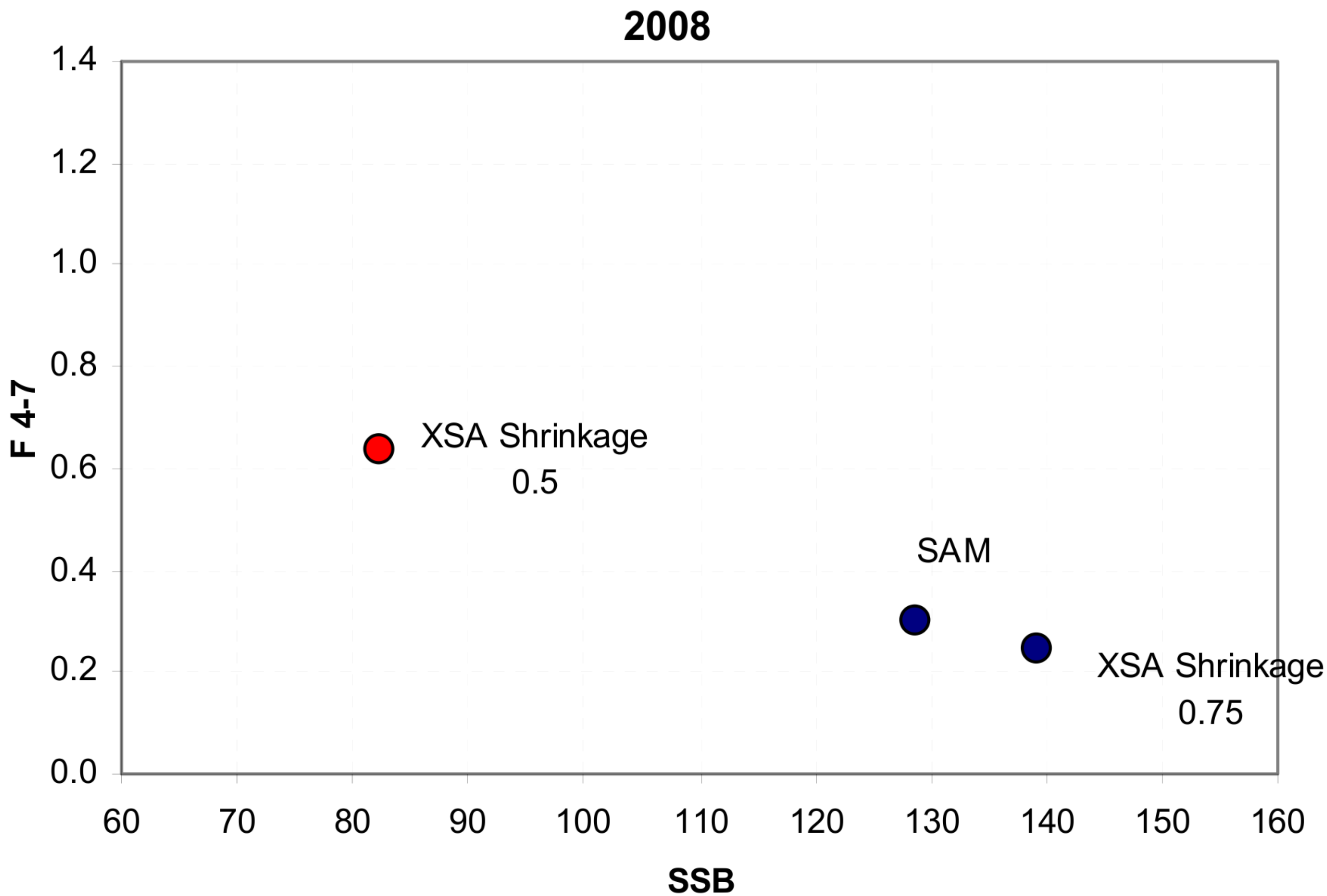
Numerical Methods

- Unscented Kalman Filter ✓
- Laplace approximation ✓
- Sampling based methods ✓

(Numerical methods are needed to calculate the marginal distribution)

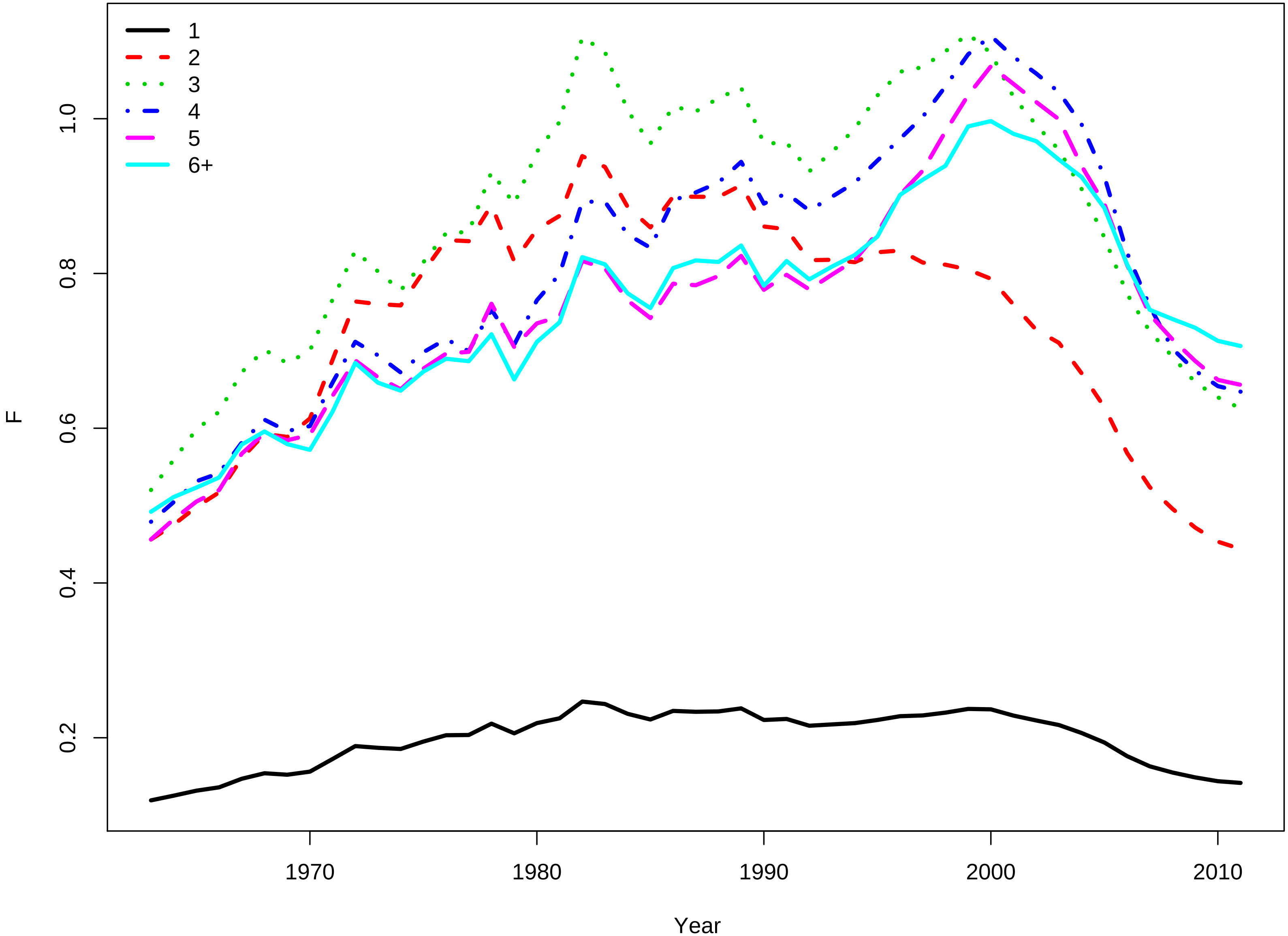
Optimization is done using AD Model Builder

Avoiding ad-hoc choices — Eastern Baltic Cod

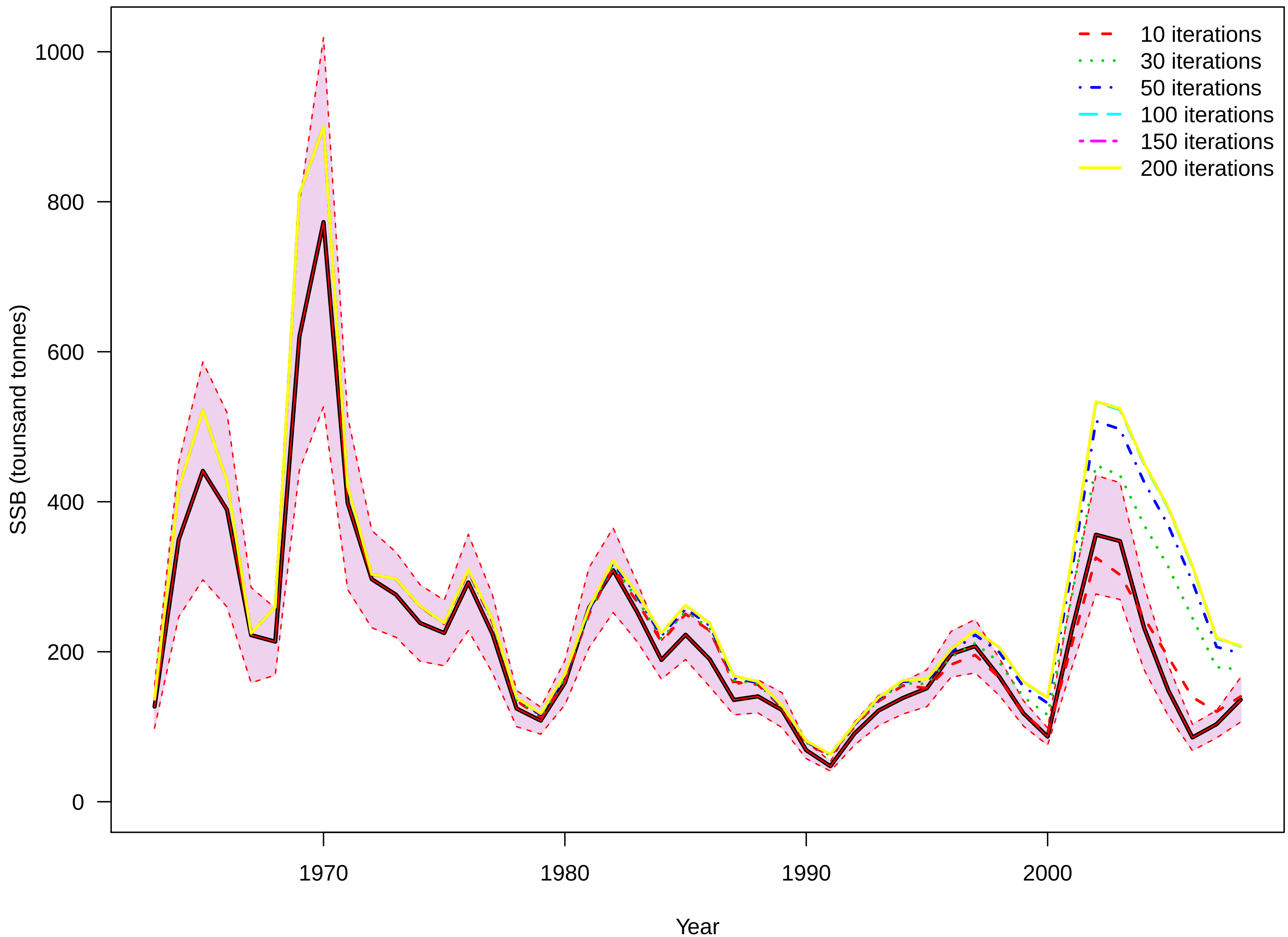


- Using the State-space Assessment Model (SAM) gives us an objective criteria

Evolving selectivity — North Sea Cod



Convergence issues — North Sea Haddock



Features of the State-space assessment model

- Statistical model
  - Maximum likelihood estimation of model parameters
  - Estimation of uncertainties are an integrated part of the model
  - Prediction is straight-forward
- Consistent treatment of all $N_{a,y}$
- Allows selectivity to evolve
- Built-in (objective!) 'F-shrinkage' and 'tapered time weights'
- Nicely handles missing observations
- Room for additional features

Summary

- State-space assessment model is a valid alternative when:
  - Catches cannot be considered know without error
  - Quantification of uncertainties are needed
  - Ad-hoc specifications are problematic
  - Parametric structures are considered too rigid
  - The number of model parameters are worrying

Appendix

Status

- **Primary model in ICES for:**
  - Western Baltic Cod
  - NE Atlantic Blue Whiting
  - Kattegat Cod
  - North Sea Cod
  - Skagerrak Sole
  - North Sea Herring
  - Bothnian Sea Herring
- **Exploratory model in ICES for:**
  Eastern Baltic Cod, North Sea Sole, North Sea Haddock, Skagerrak Plaice
- **Quick tests for some other stocks:**
  Western Baltic herring, 3PS Cod, 4VWX Herring, Greenland Halibut SA2+3KLMNO, American Plaice, Namibian Hake, Georges Bank Yellowtail Flounder, ...

Random effects in AD Model Builder

- In random effects models we have
  - Random variables we observe: x
  - Random variables we do not observe: z
  - Model parameters we want to estimate: θ
- If we had observed x and z we would have a joint likelihood $L(x, z, \theta)$
- but z is unobserved so we have to estimate θ in the marginal likelihood:

$$L(x, \theta) = \int L(x, z, \theta) dz$$

- This requires a high dimensional integral — which is difficult
- This is (part of) the reason MCMC methods are so widely used
- MCMC can be slow, difficult to judge convergence, and in tools like winBugs a prior must be assigned to everything — even when you have no prior information.
- AD Model Builder has a better solution

Laplace approximation

- Want to compute the marginal likelihood for a given θ value:

$$L(x,\theta) = \int L(x,z,\theta)dz$$

- First the joint likelihood $L(x,z,\theta)$ is optimized w.r.t. z.
- This optimization yields an estimate $\hat{z}$, and an estimated hessian $\mathcal{H}(\hat{z})$.
- Next a Gaussian approximation is assumed and the result (apart from a constant) is:

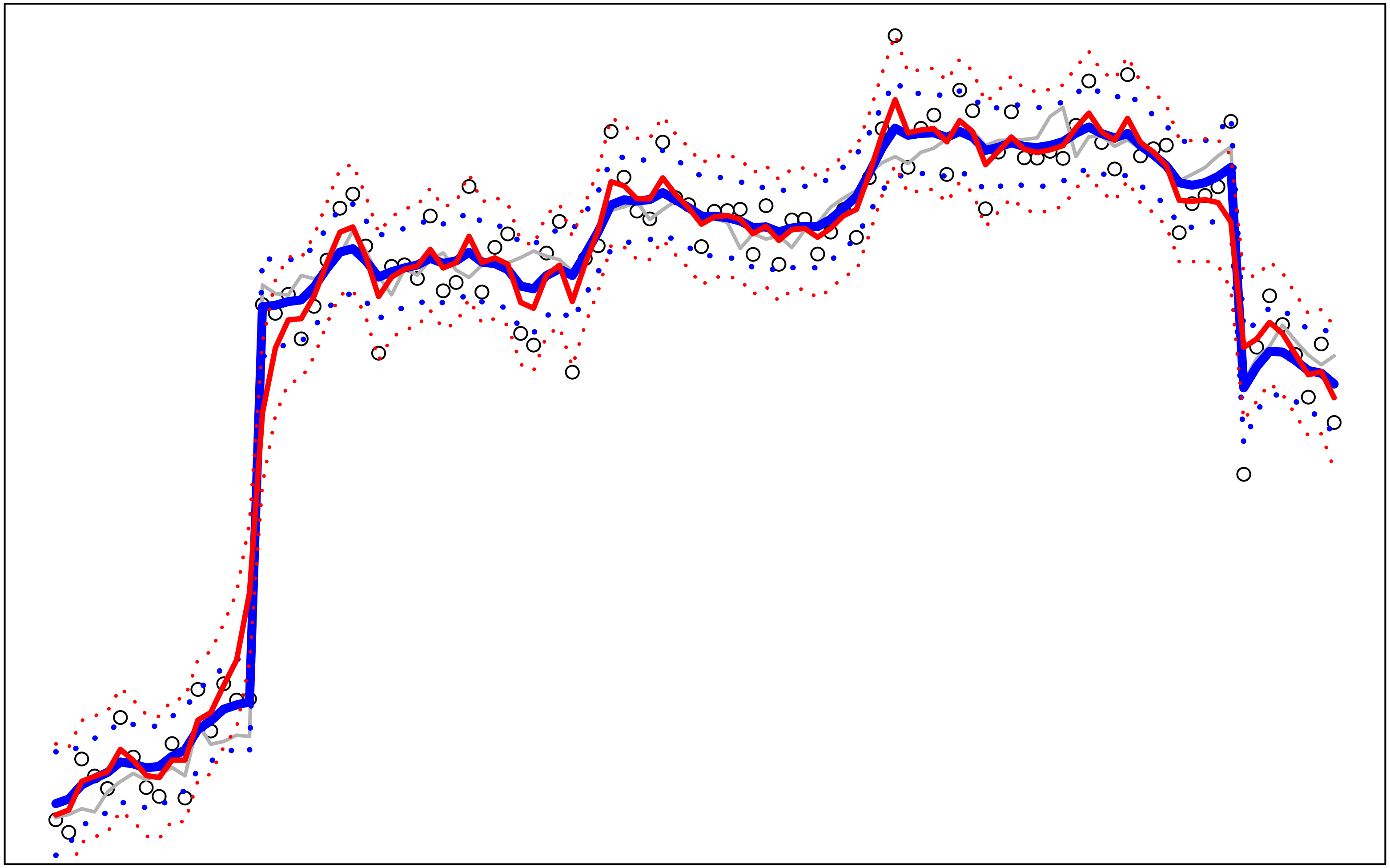
$$L(x,\theta) \approx |\det(\mathcal{H}(\hat{z}))|^{-0.5} L(x,\hat{z},\theta)$$

- Notice that when defined in this way $\hat{z}$ and $\mathcal{H}(\hat{z})$ and also depend on θ, which makes AD of this pretty difficult, but all solved for us in AD Model Builder.
- Actually this is all very simple to use. All we have to do is:
  - Code up the joint negative log likelihood
  - declare as `random_effects_vector z(1,n);`

From Fryer's listed disadvantages

- Requires normally distributed errors. **No, but they are still convenient.**
- Requires linear approximation of non-linear equations. **Not anymore.**
- There is some arbitrariness in the starting values. **Not anymore.**
- The likelihood can be very flat. **No change.**
- Maximum likelihood estimation can take a long time. **1-2 minutes on my laptop.**
- Initial coding is hard. **ADMB makes it easier**
- Favours status quo so struggles to pick up a collapsing stock.

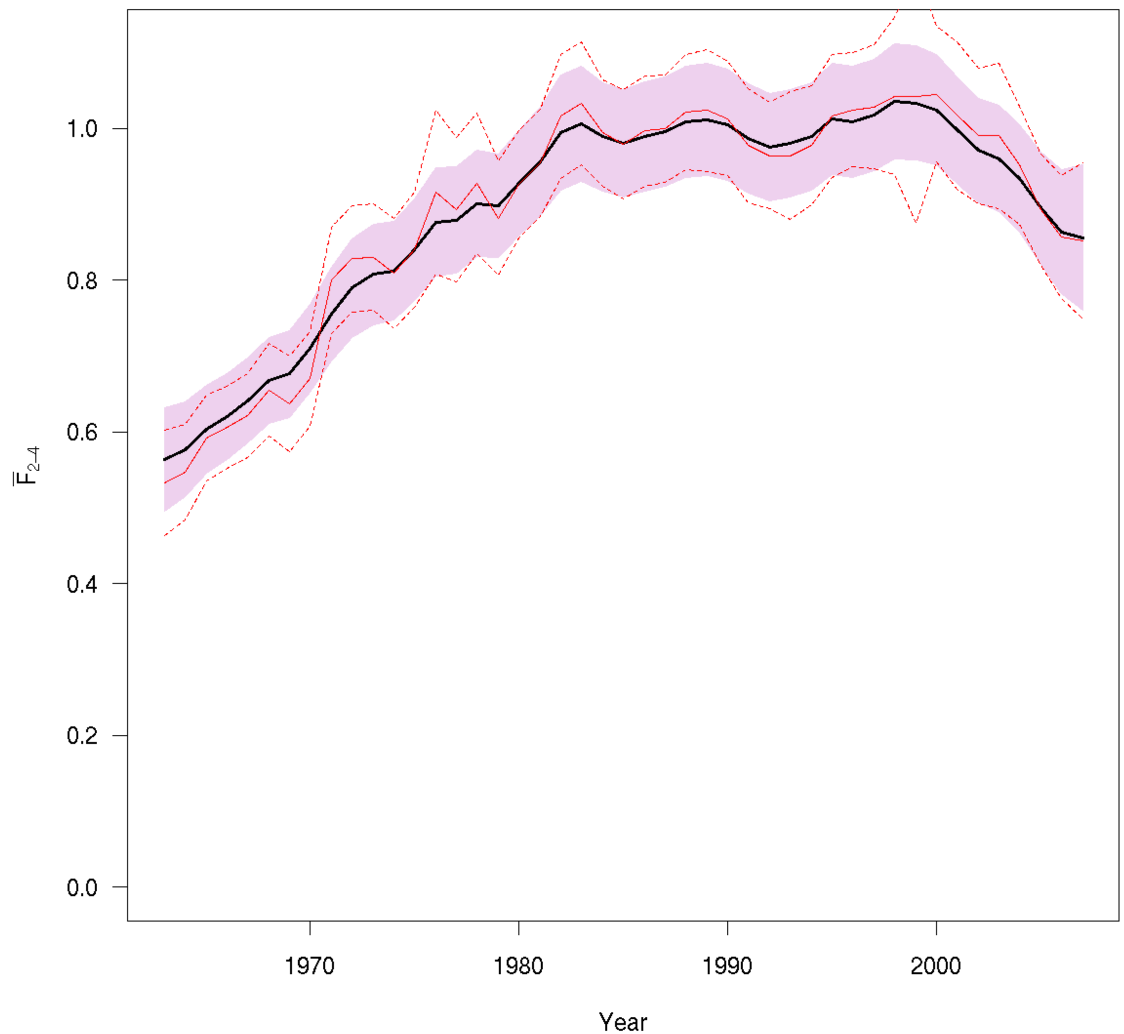
Allow sharp jumps



- In the standard model $\Delta \log F_y = \log F_y - \log F_{y-1}$ is assumed Gaussian
- Instead use a mixture, such as: $\Delta \log F_y \sim (1-p)\mathbf{N}(.,.) + p\mathbf{t}_1(.,.)$

Allow sharp jumps - results

- Allowing the **t**-jump-fraction p to be estimated.
  No change.
- Forcing $p = 10\%$.
  No visible change.
- Forcing $p = 30\%$.
  Visible change,
  but nothing dramatic



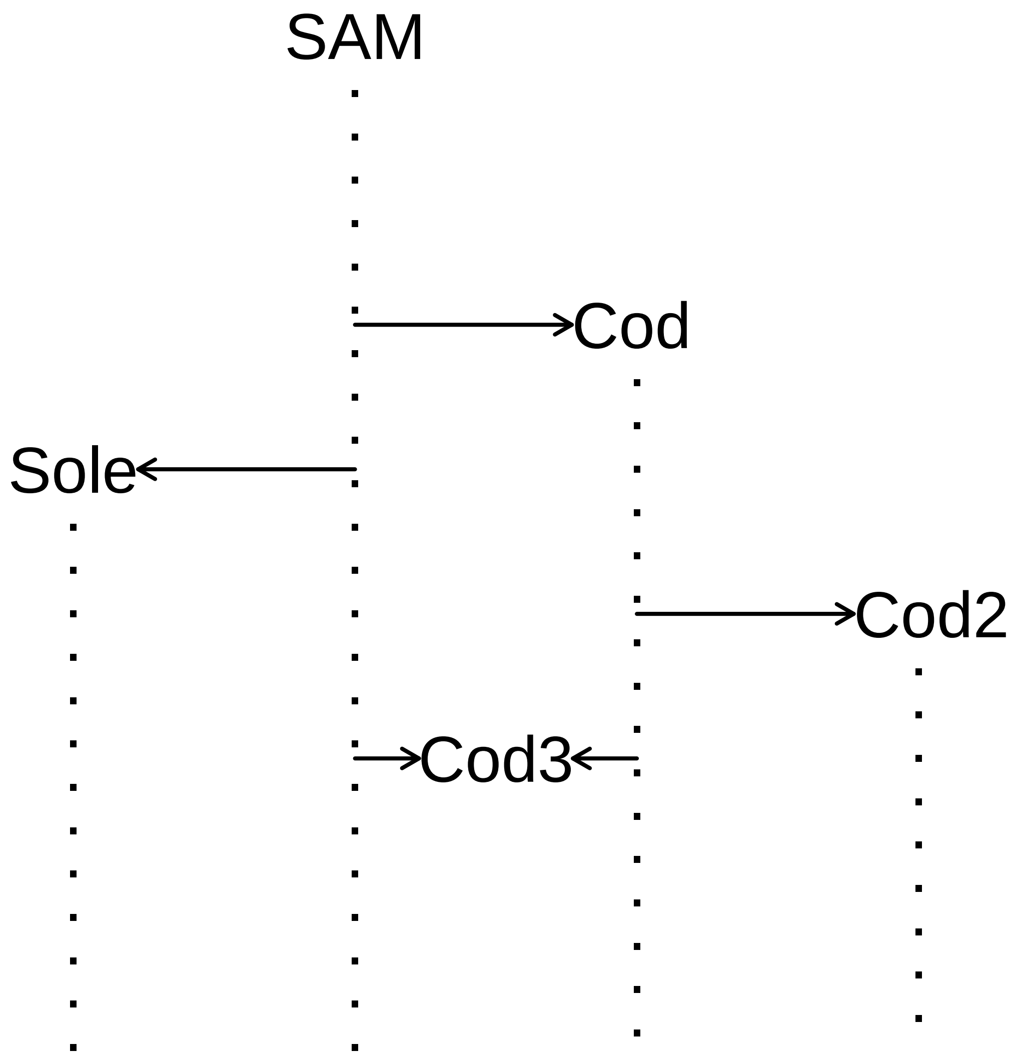
Correlated Random Walks

- Instead of independent random walks for F at different ages, we can allow those random walks to be correlated

$$\Delta \log(F) \sim \mathcal{N}(0, \Sigma)$$

- The covariance matrix Σ is defined via the random walk variances, and the correlation coefficients $\rho_{i,j} = \Sigma_{i,j} / \sqrt{\Sigma_{i,i}\Sigma_{j,j}}$
- We assume the very simple structure

$$\rho_{i,j} = \begin{cases} 1, & \text{for } i = j \\ \rho, & \text{otherwise} \end{cases}$$


SAM
Cod
Sole
Cod2
Cod3

Exercise

- Start with the full parametric catch-at-age model `fsa.tpl` for North Sea Cod.
- Modify the code to make recruitment a random walk.
- Compare the fits
- Discuss pros and cons.