Séminaire INED-SMS "Notions d'âge, de période et de cohorte"

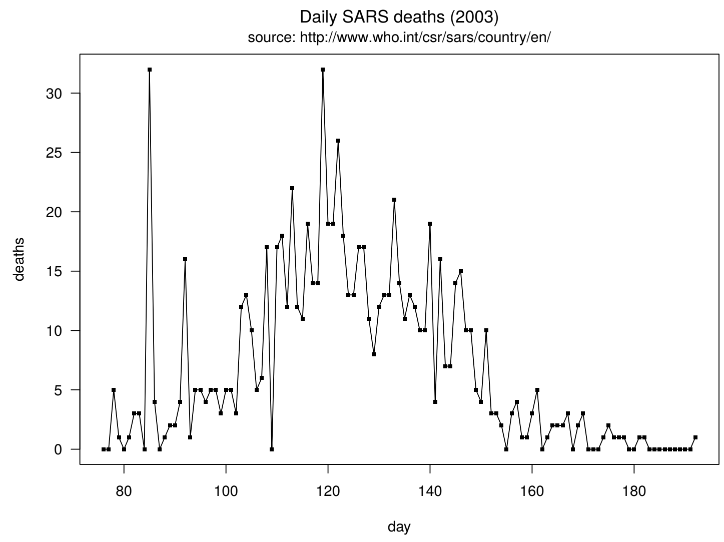
# Modelling mortality over age and time, and age and cohort: a nonparametric approach

Carlo Giovanni Camarda
Institut National d'Etudes Démographiques
carlo-giovanni.camarda@ined.fr

September 2015, Paris



## Direct smoothing: an example








## Direct smoothing: how does it work?

- We have an $n$-vector: $\mathbf{y}$
- We aim to have a fitted $n$-vector, $\boldsymbol{\eta}$ with (balanced)
  - accuracy to the data (fit)
  - smooth behavior (roughness)
- We measure the fit: $\sum_n (y_i - \eta_i)^2$
- $\boldsymbol{\eta}$ is rough when $\eta_i - \eta_{i-1} = \Delta\eta_i$ is big
- We measure the roughness as $\sum_n (\Delta\eta_i)^2$
- We minimize (Whittaker, 1923):

$$S = \sum_n (y_i - \eta_i)^2 + \lambda \sum_n (\Delta\eta_i)^2$$

- Higher $\lambda \Rightarrow$ smoother $\boldsymbol{\eta}$





## Direct smoothing in matrix

- Minimize:

$$S = |\mathbf{y} - \boldsymbol{\eta}|^2 + \lambda|\mathbf{D}\boldsymbol{\eta}|^2$$

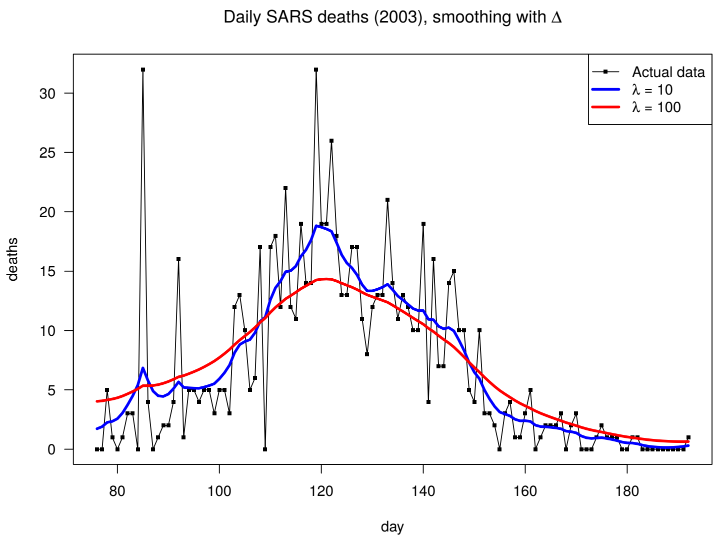
- Example of $\mathbf{D}$:

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

- This is a linear system:

$$\hat{\boldsymbol{\eta}} = (\mathbf{I} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{y}$$

# Direct smoothing: fitted data with $d = 1$








# Direct smoothing with higher differences

- We can measure roughness by second differences: $\sum_{i=3}^{n}(\eta_i - 2\eta_{i-1} + \eta_{i-2})^2$
- In general we minimize:

$$S = \sum_{i=1}^{n}(y_i - \eta_i)^2 + \lambda \sum_{i=d+1}^{n}(\Delta^d \eta_i)^2$$

with explicit solution:

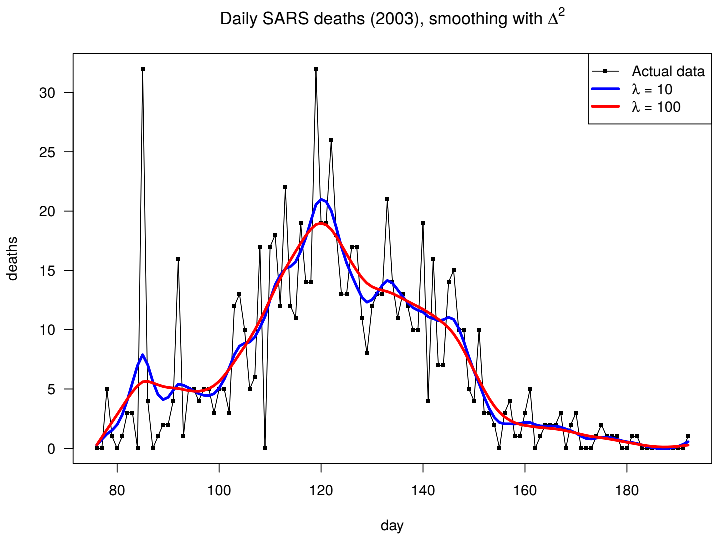
$$\hat{\boldsymbol{\eta}} = (\mathbf{I} + \lambda \mathbf{D}_d' \mathbf{D}_d)^{-1}\mathbf{y}$$

- For a given λ, outcomes are less smooth





# Direct smoothing: fitted data with $d = 2$







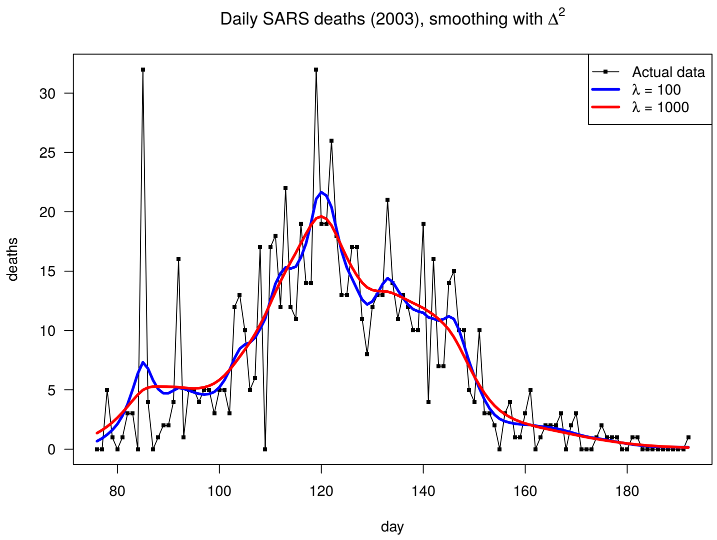

# Direct smoothing of counts

- We have counts (we also need to force positive fitted values)
- We assume Poisson distribution: $P(y_i|\mu_i) = \mu_i^{y_i}\exp(-\mu_i)/y_i!$
- We work on the expectation: $E(y_i) = \mu_i$
- ... and "linear predictor": $\boldsymbol{\eta} = \ln(\boldsymbol{\mu})$
- We have a non-linear system of equations, in an iterative process, at step $t+1$:

$$\begin{aligned} \boldsymbol{\eta}_{t+1} &= (\mathbf{I}'\mathbf{W}_t\mathbf{I} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{I}'\mathbf{W}_t\left[\frac{\mathbf{y} - \boldsymbol{\mu}_t}{\boldsymbol{\mu}_t} + \boldsymbol{\eta}_t\right] \\ \boldsymbol{\eta}_{t+1} &= (\mathbf{W}_t + \lambda\mathbf{D}'\mathbf{D})^{-1}\left[\mathbf{y} - \boldsymbol{\mu}_t + \boldsymbol{\eta}_t\mathbf{W}_t\right] \end{aligned}$$

where $\mathbf{W}_t = diag(\boldsymbol{\mu}_t)$ and $\boldsymbol{\eta}_1 = \ln(\mathbf{y}+1)$

# Generalized linear smoothing








# Direct smoothing: features

- Advantages:
  - easy to explain and to use
  - easy to generalize
  - automatic interpolation
  - fast (with sparse matrices)
- Disadvantages:
  - equispaced **y**
  - no easy diagnostic
  - the size of the problem increases proportionally with $n$





# Simple least-squares

- We have an $n$-vector $\mathbf{y}$ and covariate(s) $\mathbf{x}$
- We assume $E(y_i) = \mu_i = \alpha_0 + \alpha_1 x_i$
- We minimize:

$$S = \sum_{i=1}^{n} [y_i - (\alpha_0 + \alpha_1 x_i)]^2$$

- In matrix:

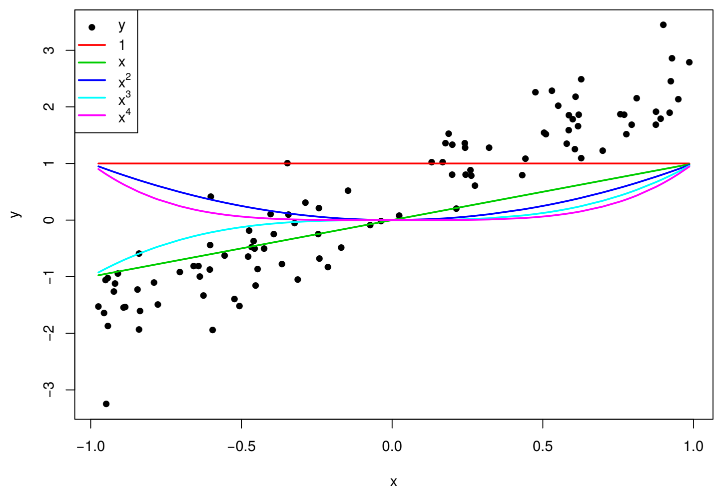
$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^4 & x_n^4 \end{bmatrix} \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} \quad , \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\alpha}$$

- We minimize: $|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}|^2 \Rightarrow \mathbf{X}'\mathbf{X}\boldsymbol{\alpha} = \mathbf{X}\mathbf{y} \Rightarrow \hat{\alpha} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$





# Model matrices



# Can we do better?

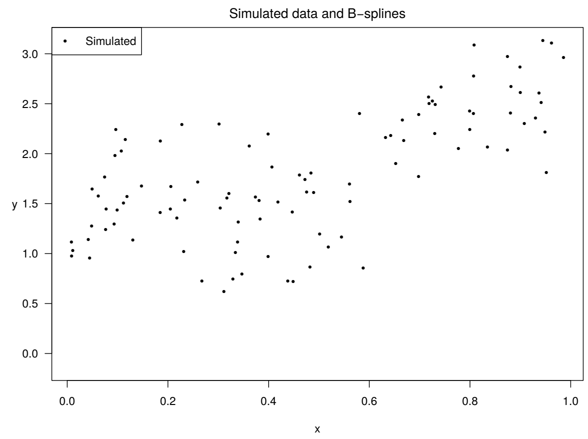
- Simple basis is good for simple example
- Basis function (powers of $x$) are global
- Moving one end moves the other end too
- Unexpected wiggles
- The higher the degree the more is sensitive

- We seek for local basis
- Useful for more complex data
- No assumptions on the trend (let the data speak by themselves!)
- Smooth outcomes





# Slightly more complex example



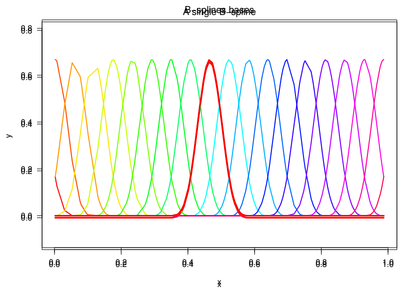





# Introducing *B*-splines

- Create a suitable basis $\Rightarrow$ (equidistant) *B*-splines:

$$\mathbf{B} = \begin{bmatrix} B_1(x_1) & B_2(x_1) & B_3(x_1) & \dots & B_k(x_1) \\ B_1(x_2) & B_2(x_2) & B_3(x_2) & \dots & B_k(x_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ B_1(x_n) & B_2(x_n) & B_3(x_n) & \dots & B_k(x_n) \end{bmatrix}$$

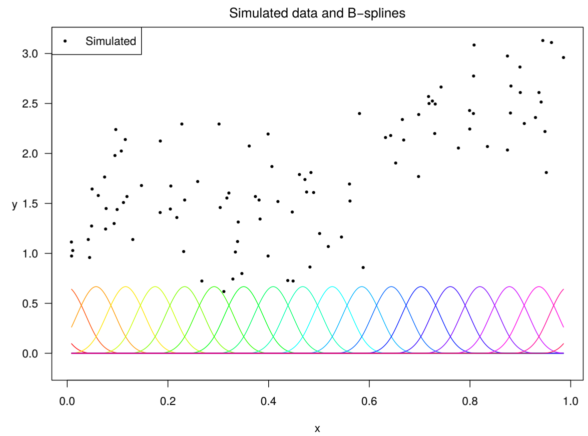



- $E(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{B}\,\boldsymbol{\alpha} \quad \Rightarrow \quad \hat{\boldsymbol{\alpha}} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{y}$

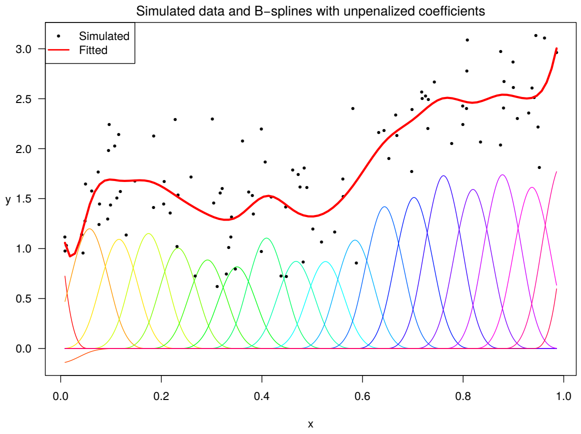




# Fitting with 20 *B*-splines



# Fitting with 20 *B*-splines








# Penalizing the coefficients: *P*-splines

- Outcomes are not smooth, we could:
  - take less *B*-splines
  - place each *B*-splines in specific positions
  - set a double goal:

1. good fit to the data, i.e. low least-squares: $S = |\mathbf{y} - \mathbf{B}\alpha|^2$
2. smooth curve, i.e. low roughness: $R = |\mathbf{D}\alpha|^2$

- We balance this two object-functions:

$$S^* = S + \lambda R = |\mathbf{y} - \mathbf{B}\alpha|^2 + \lambda|\mathbf{D}\alpha|^2$$

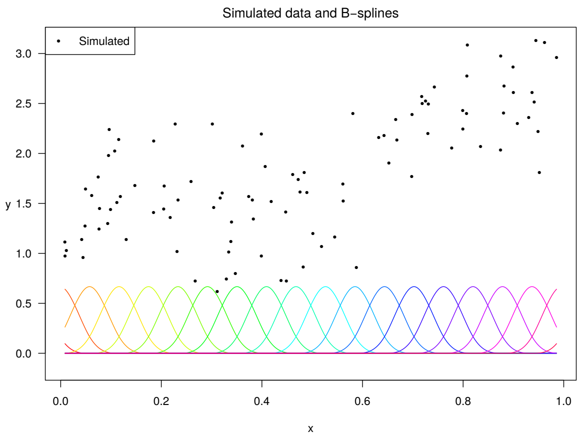
- Given a $\lambda$, this is again a linear system of equation with explicit solution:

$$\hat{\alpha} = (\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'\mathbf{y}$$





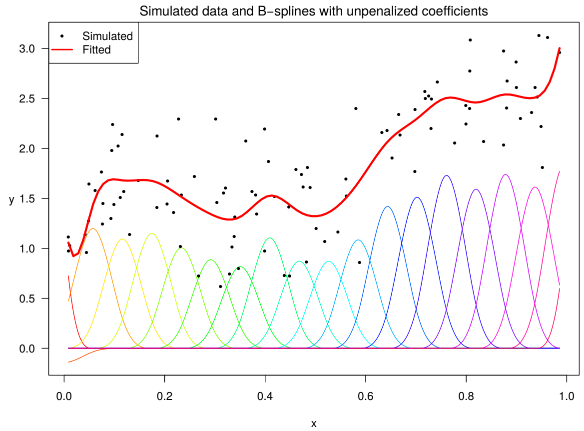
# Including a penalty, $d = 2$



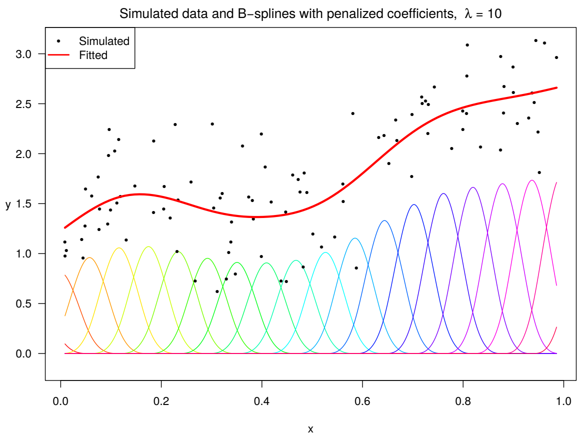





# Including a penalty, $d = 2$



# Including a penalty, $d = 2$








# *P*-splines for Poisson data

- We assume again a Poisson distribution: $y_i \sim Poi(\mu_i)$
- We use again a "linear predictor": $\boldsymbol{\eta} = \ln(\boldsymbol{\mu}) = \mathbf{B}\boldsymbol{\alpha}$
- We have a non-linear system of equations, in an iterative process, at step $t+1$:

$$\tilde{\alpha}_{t+1} = (\mathbf{B}'\tilde{\mathbf{W}}_t\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'\hat{\mathbf{W}}_t\tilde{\mathbf{z}}_t\,,$$

where
$\tilde{\mathbf{z}} = (\mathbf{y} - \tilde{\boldsymbol{\mu}})/\tilde{\boldsymbol{\mu}} + \mathbf{B}\tilde{\boldsymbol{\alpha}}$ (working dependent variable)
$\tilde{\mathbf{W}} = diag(\tilde{\boldsymbol{\mu}})$ (weight matrix)

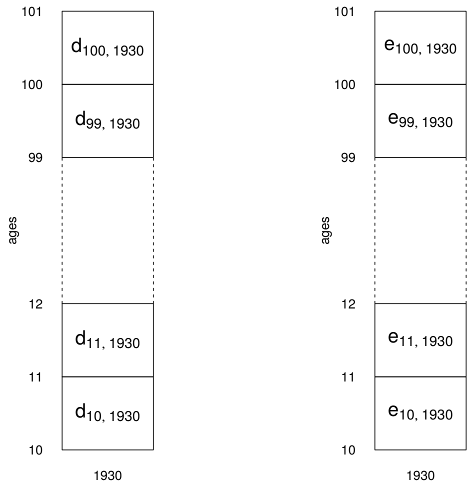
- We can include an offset (exposures for mortality data) just changing:

$$\boldsymbol{\mu} = \mathbf{e}\,\exp(\boldsymbol{\eta}) = \exp(\mathbf{B}\boldsymbol{\alpha} + \ln(\mathbf{e}))$$





# Data and fitting in one dimension

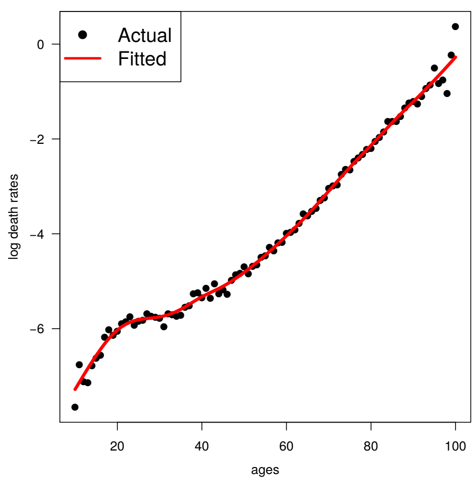



A graphical representation of the data set in 1D over ages (10-100) for a given year (1930).





# Data and fitting in one dimension




Actual and smooth death rates in log-scale over ages (10-100) for a given year (1930). Denmark, females.

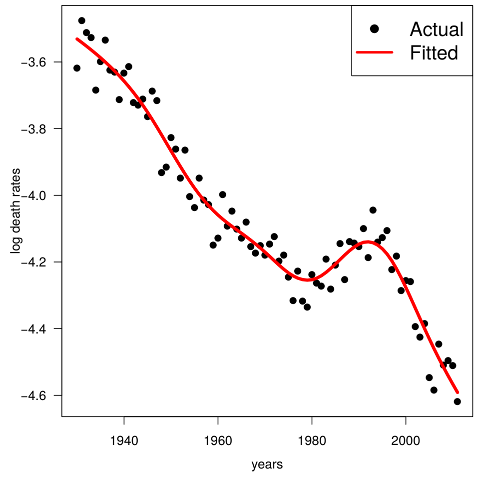
# Data and fitting in one dimension


A graphical representation of the data set in 1D over years (1930-2011) for a given age (65).





# Data and fitting in one dimension

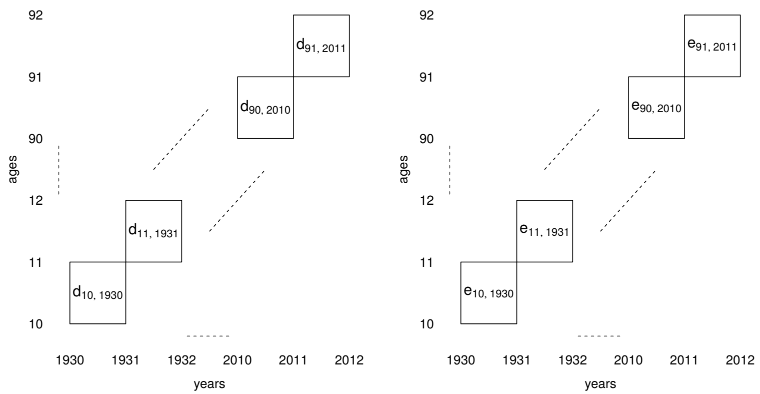



Actual and smooth death rates in log-scale over years (1930-2011) for a given age (65).
Denmark, females.





# Data and fitting in one dimension

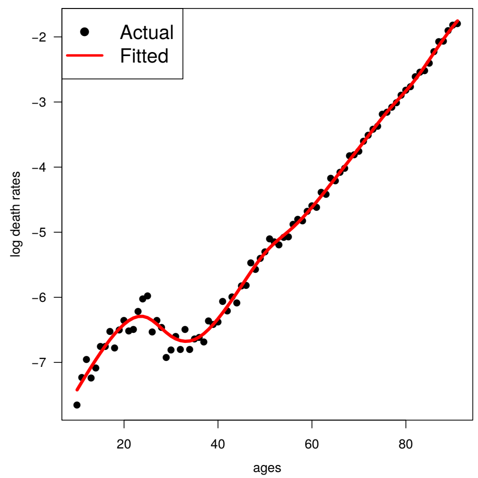



A graphical representation of the data set in 1D over ages (10-91) for a given cohort (1920).



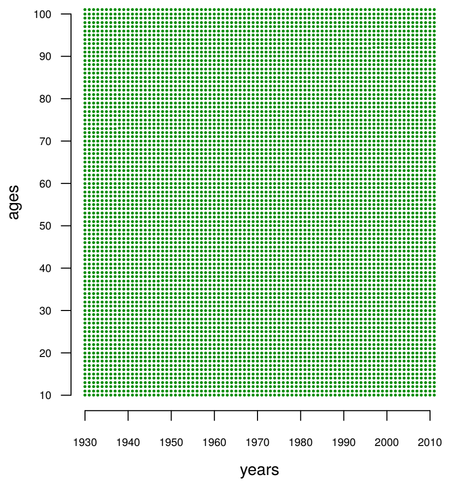


# Data and fitting in one dimension




Actual and smooth death rates in log-scale over ages (10-91) for the cohort 1920.
Denmark, females.

# Looking at mortality data over age and years

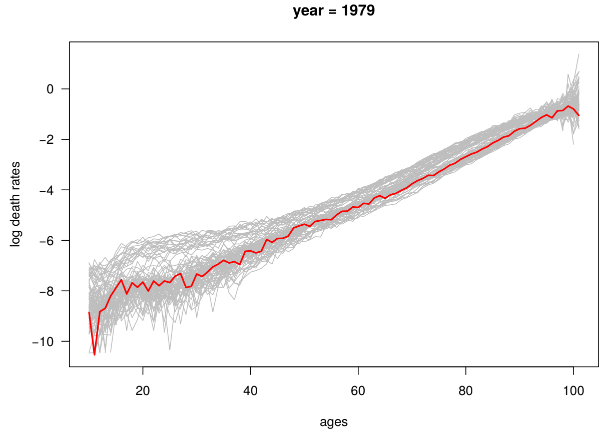



A graphical representation of the available data over ages (10-100) and years (1930-2011). Denmark, females.





# Looking at mortality data over age and years

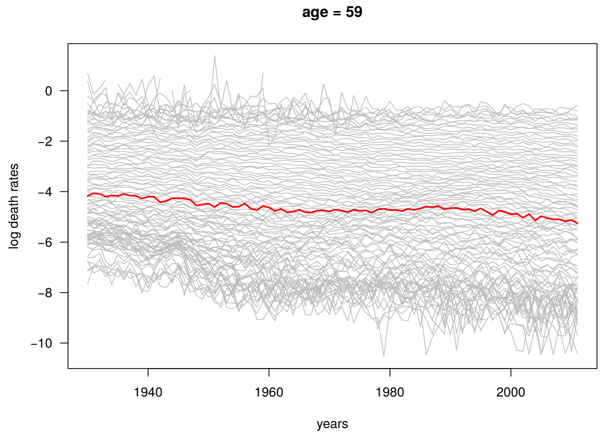



Death rates in log scale over ages for different years.
Ages from 10 to 100. Denmark, females, 1930-2011.





# Looking at mortality data over age and years

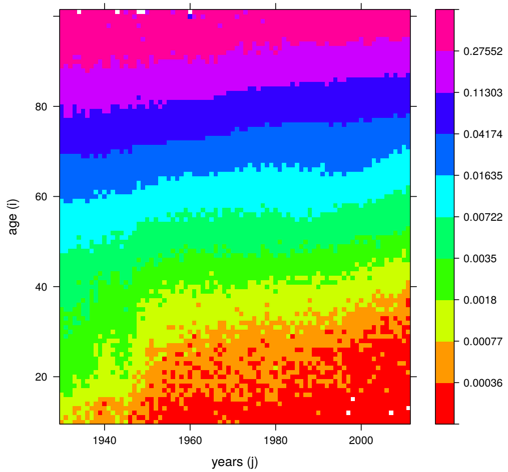



Death rates in log scale over years for different ages.
Ages from 10 to 100. Denmark, females, 1930-2011.





# Looking at mortality data over age and years




Death rates. Ages from 10 to 100. Denmark, females, 1930–2011.

## *P*-splines for mortality on rectangular grid

- We arrange both death and exposures matrices in columns
- The regression matrix for our two-dimensional model is the Kronecker product $\mathbf{B} = \mathbf{B}_y \otimes \mathbf{B}_a$, where $a$ and $y$ stand for age and year dimensions
- $\mathbf{B}$ has an associated vector of regression coefficients $\boldsymbol{\alpha}$, which can be arranged in a matrix $\mathbf{A}$
- We penalized the coefficients to each of the columns and rows of $\mathbf{A}$:

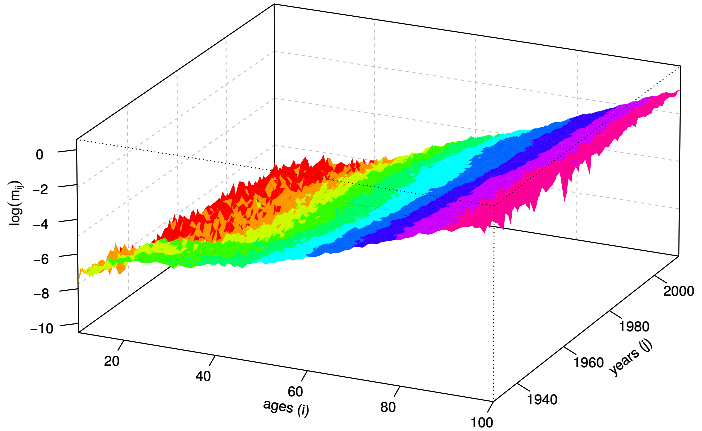
$$\mathbf{P} = \lambda_a \mathbf{I}_{c_y} \otimes \mathbf{D}_a'\mathbf{D}_a + \lambda_y \mathbf{D}_y'\mathbf{D}_y \otimes \mathbf{I}_{c_a}$$

  $\lambda_a$ and $\lambda_y$ are the smoothing parameters used for age and year, respectively (Currie et al. 2006).





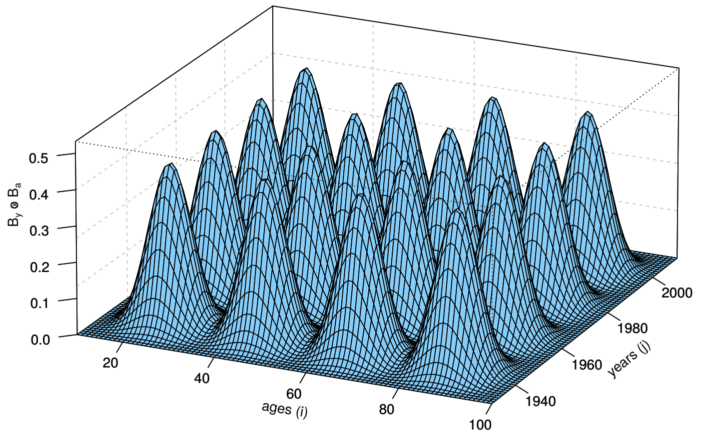
## *P*-splines on rectangular grid







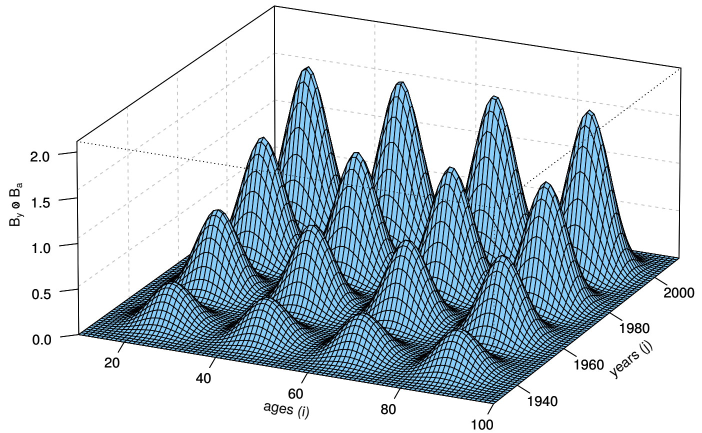

## *P*-splines on rectangular grid



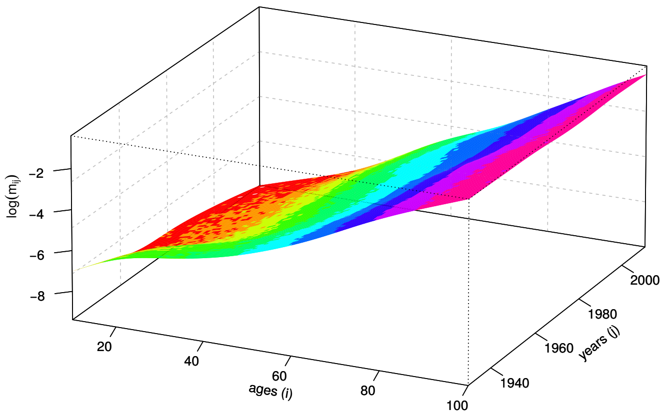





## *P*-splines on rectangular grid



## Fitted mortality over ages and years

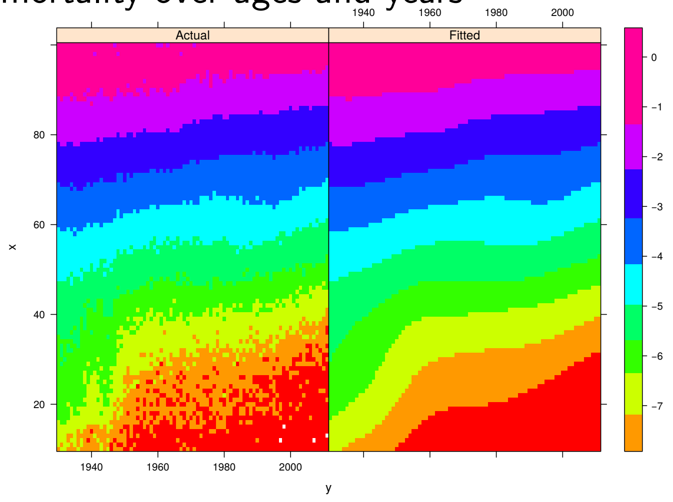


Death rates. Ages from 10 to 100. Denmark, females, 1930–2011. Colors based on actual level and third dimensions.





## Fitted mortality over ages and years

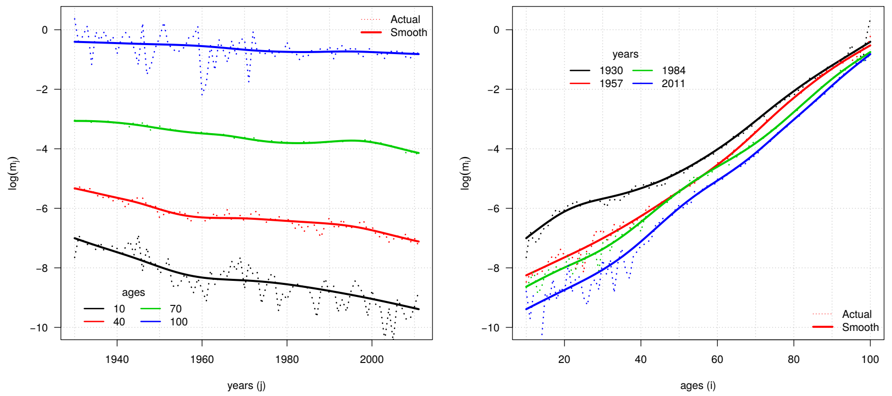


Death rates. Ages from 10 to 100. Denmark, females, 1930–2011. Actual (left) smooth (right).





## Fitted mortality over ages and years

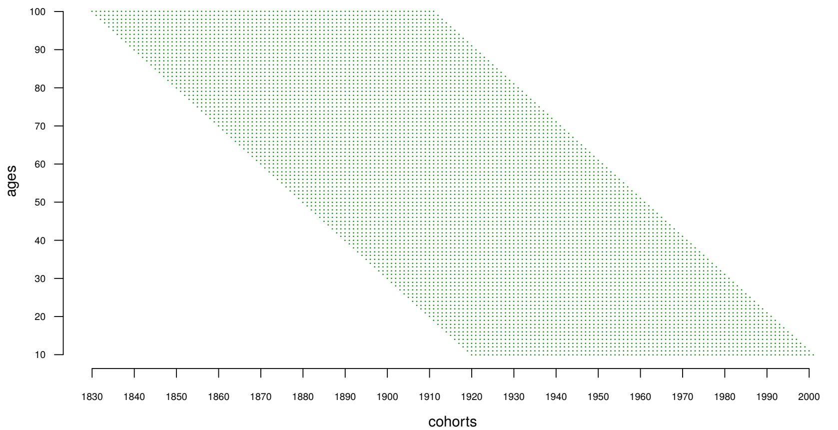


Actual and smooth death rates in log scale.
Ages from 10 to 100. Denmark, females, 1930–2011.



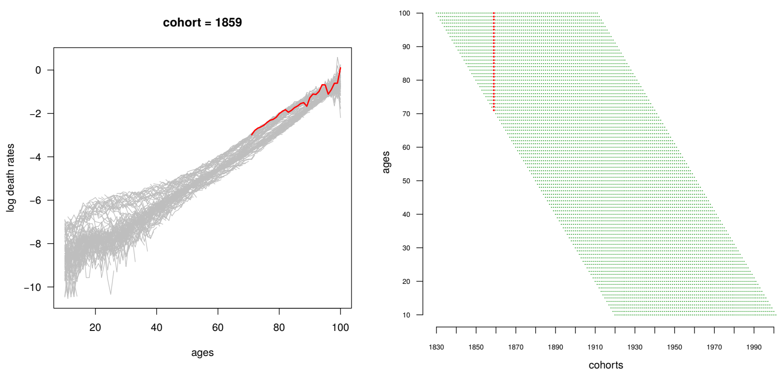


## Looking at mortality data over age and cohorts



A graphical representation of the available data over ages (10-100) and cohorts (1920-2001). Denmark, females.

## Looking at mortality data over age and cohorts

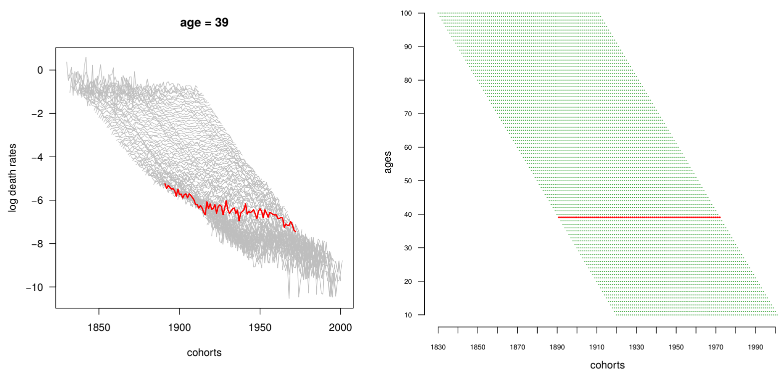



Death rates in log scale over ages for different cohorts.
Ages from 10 to 100. Denmark, females, cohorts 1920-2001.





## Looking at mortality data over age and cohorts

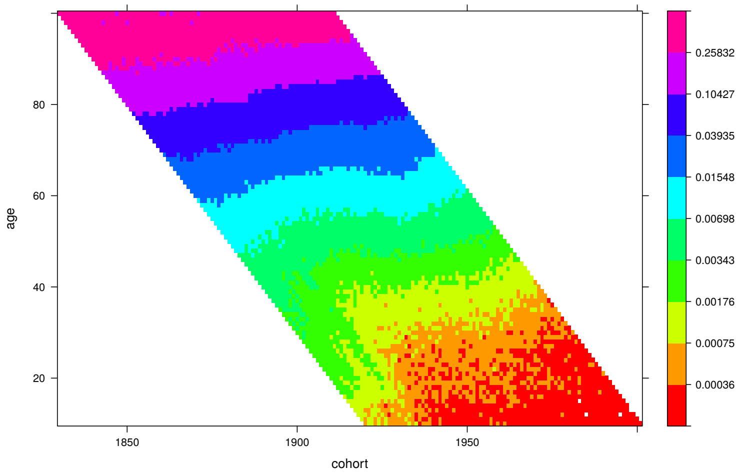



Death rates in log scale over cohorts for different ages.
Ages from 10 to 100. Denmark, females, cohorts 1920-2001.





## Looking at mortality data over age and cohorts




Death rates. Ages from 10 to 100. Denmark, females, cohorts 1830–2001.





## Estimating uncompleted cohorts

- We treat the uncompleted cohorts as a missing value problem
- The model works when data are on a rectangular grid

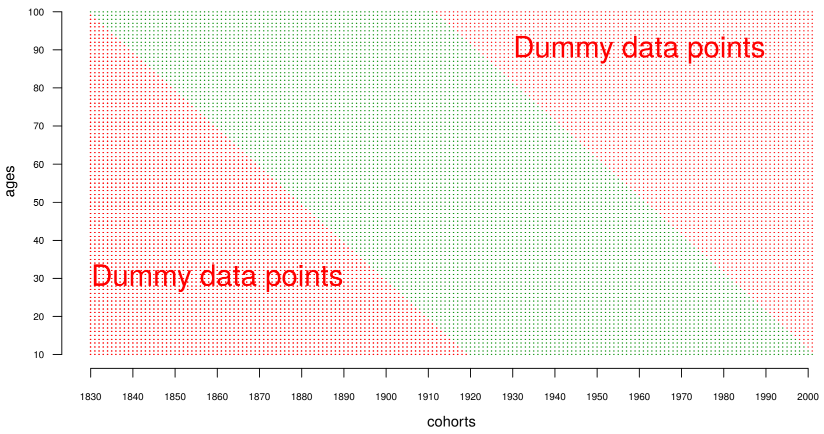
  $$\Downarrow$$

  re-arrange parallelogram to obtain a rectangular grid by
  - placing dummy variables for death and exposures over the uncompleted cohorts ($\breve{\mathbf{Y}}$, $\breve{\mathbf{E}}$) and augmenting the *B*-spline basis ($\breve{\mathbf{B}}$)
  - adding in the algorithm a weight matrix ($\mathbf{V}$) with zero weights over the uncompleted cohorts

  $$(\breve{\mathbf{B}}'\mathbf{V}\tilde{\mathbf{W}}\breve{\mathbf{B}} + \mathbf{P})\tilde{\alpha} = \breve{\mathbf{B}}'\mathbf{V}\tilde{\mathbf{W}}\tilde{\mathbf{z}}$$

  with $\tilde{\mathbf{z}} = \mathbf{V}(\breve{\mathbf{y}} - \breve{\mathbf{e}}\tilde{\mu})/(\breve{\mathbf{e}}\tilde{\mu}) + \breve{\mathbf{B}}\tilde{\alpha}$.
- We estimate the fitted and "extrapolated" values simultaneously
- This approach is useful to forecast uncompleted cohorts

# Fitted mortality over ages and cohorts

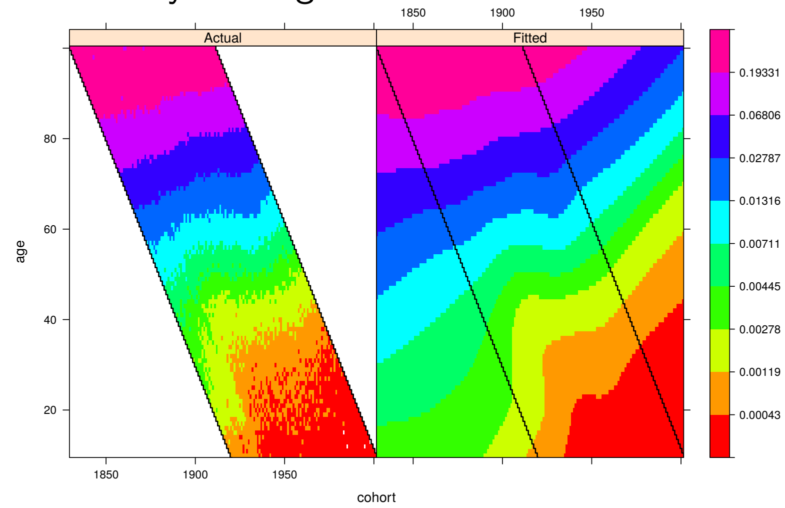



A graphical representation of the available data over ages (10-100) and cohorts (1920-2001). Denmark, females.





# Fitted mortality over ages and cohorts

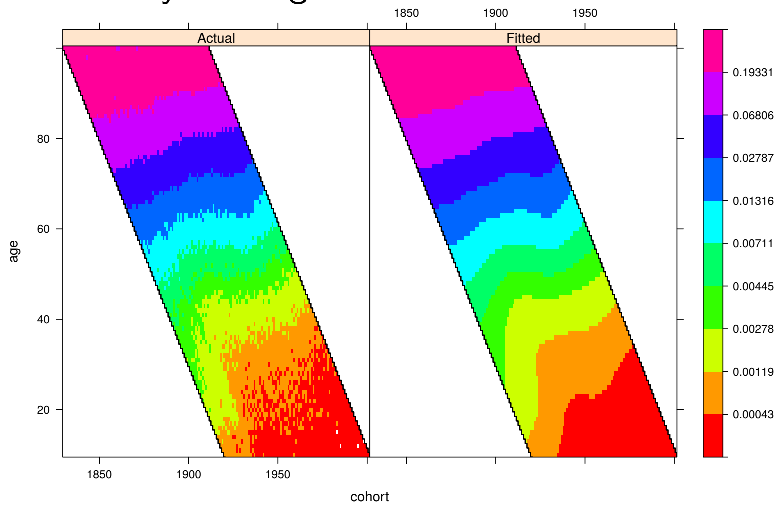



Actual and smooth death rates in a log scale.
Ages from 10 to 100. Denmark, females, cohorts 1830-2001.





# Fitted mortality over ages and cohorts

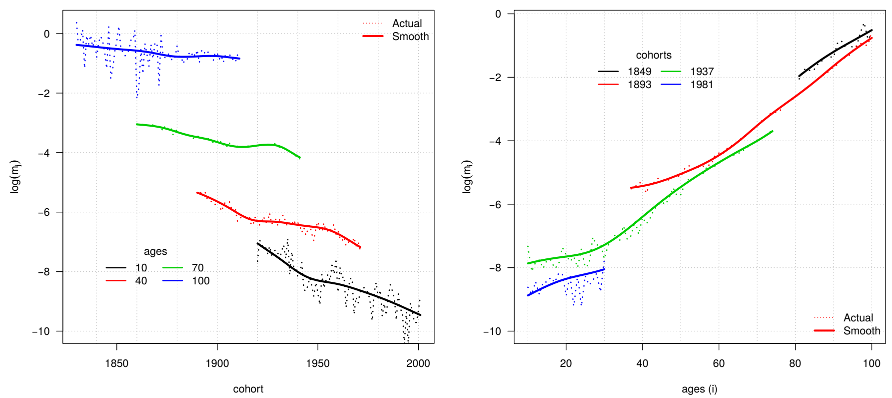



Actual and smooth death rates in a log scale.
Ages from 10 to 100. Denmark, females, cohorts 1830-2001.





# Fitted mortality over ages and cohorts




Actual and smooth death rates in log scale.
Ages from 10 to 100. Denmark, females, cohorts 1830-2001.

## Main bibliography

Camarda, C. G. (2012). `MortalitySmooth`: An `R` Package for Smoothing Poisson Counts with *P*-Splines. *Journal of Statistical Software* **50**, 1–24.

Currie, I. D., M. Durbán and P. H. C. Eilers (2004). Smoothing and forecasting mortality rates. *Statistical Modelling* **4**, 279–298

Currie, I. D., M. Durbán and P. H. C. Eilers (2006). Generalized Linear Array Models with Applications to Multidimentional Smoothing. *Journal of the Royal Statistical Society B* **68**, 259–280.

Eilers, P. H. C. and B. D. Marx (1996). Flexible Smoothing with *B*-splines and Penalties. *Statistical Science* **11**, 89–121.

Whittaker, E. T. (1923). On the New Method of Graduation *Proc. Edinburgh Math. Soc.*





# Appendix





## Optimizing $\lambda$

- A possibility is the Bayesian Information Criterion (BIC):

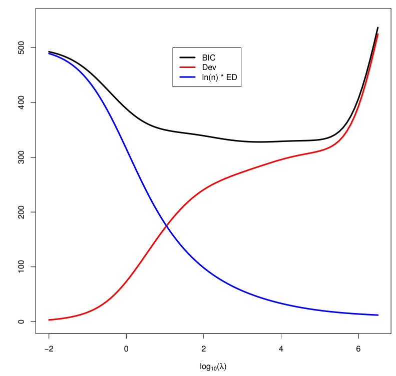
$$BIC = Dev + \ln(n) ED$$

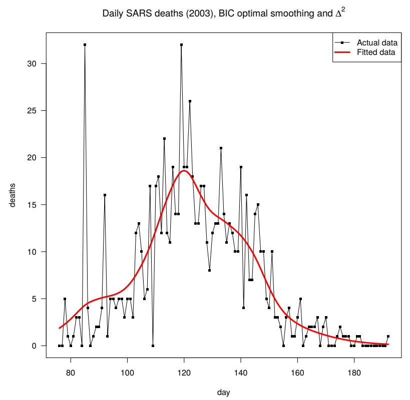
- For Poisson: $Dev = 2 \sum_n y_i log(y_i/\mu_i)$
- Commonly: $ED = trace(\mathbf{H})$
- $\mathbf{H} = (\mathbf{W} + \lambda \mathbf{D}'\mathbf{D})^{-1}\mathbf{W}$
- Other possibilities: AIC, CV, GCV



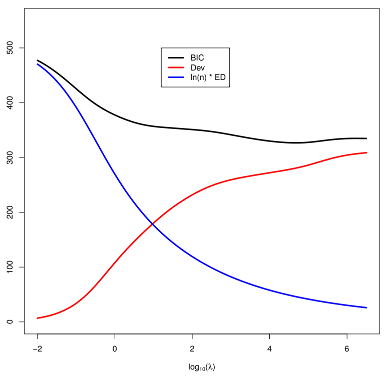


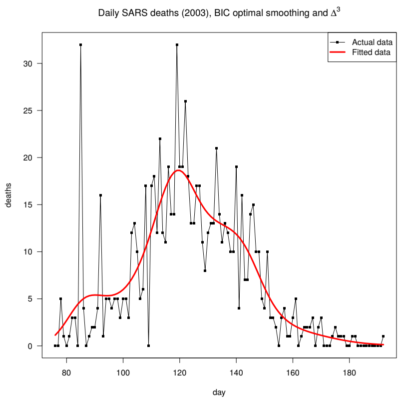
## BIC for the SARS data, $d = 2$



# BIC for the SARS data, $d = 3$

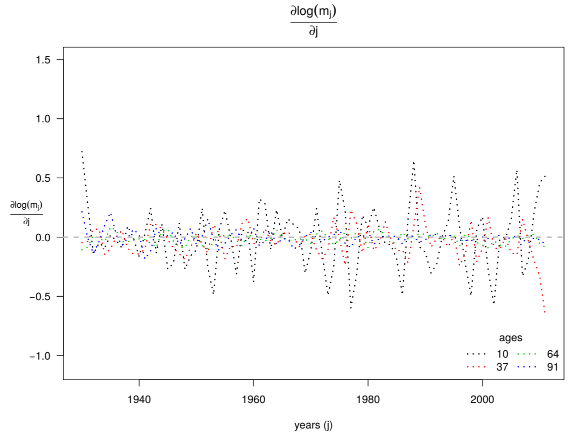







# Looking at selected ages and thier derivatives

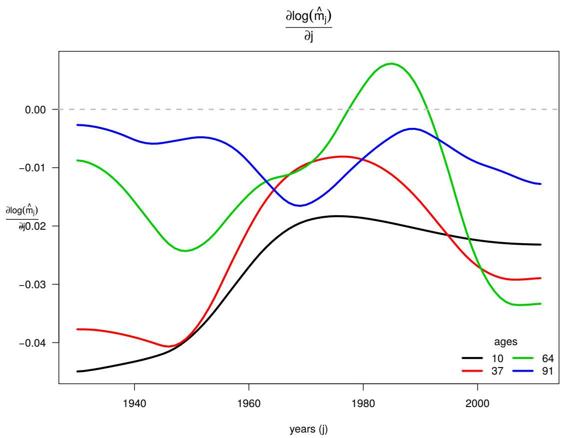



Death rates. Numeric derivatives from actual data for selected ages over years. Denmark, females, 19302011.





# Looking at selected ages and thier derivatives




Death rates. Numeric derivatives from smooth data with 2D P-splines for selected ages over years. Denmark, females, 19302011.