RAND HEALTH

Demographic Forecasting and the Role of Priors

Federico Girosi

The RAND Corporation

Santa Monica, CA, USA

Reference

All the material for this lecture can be found at http://gking.harvard.edu/files/smooth/

Plan of the Lecture

- Demographic forecasting is a machine learning problem
- Solving the problem in the Bayesian/ regularization framework
- A closer look at one dimensional priors
- A closer look at the smoothness parameter
- Examples/Demos

Forecasting Mortality and Disease Burden Has Important Applications

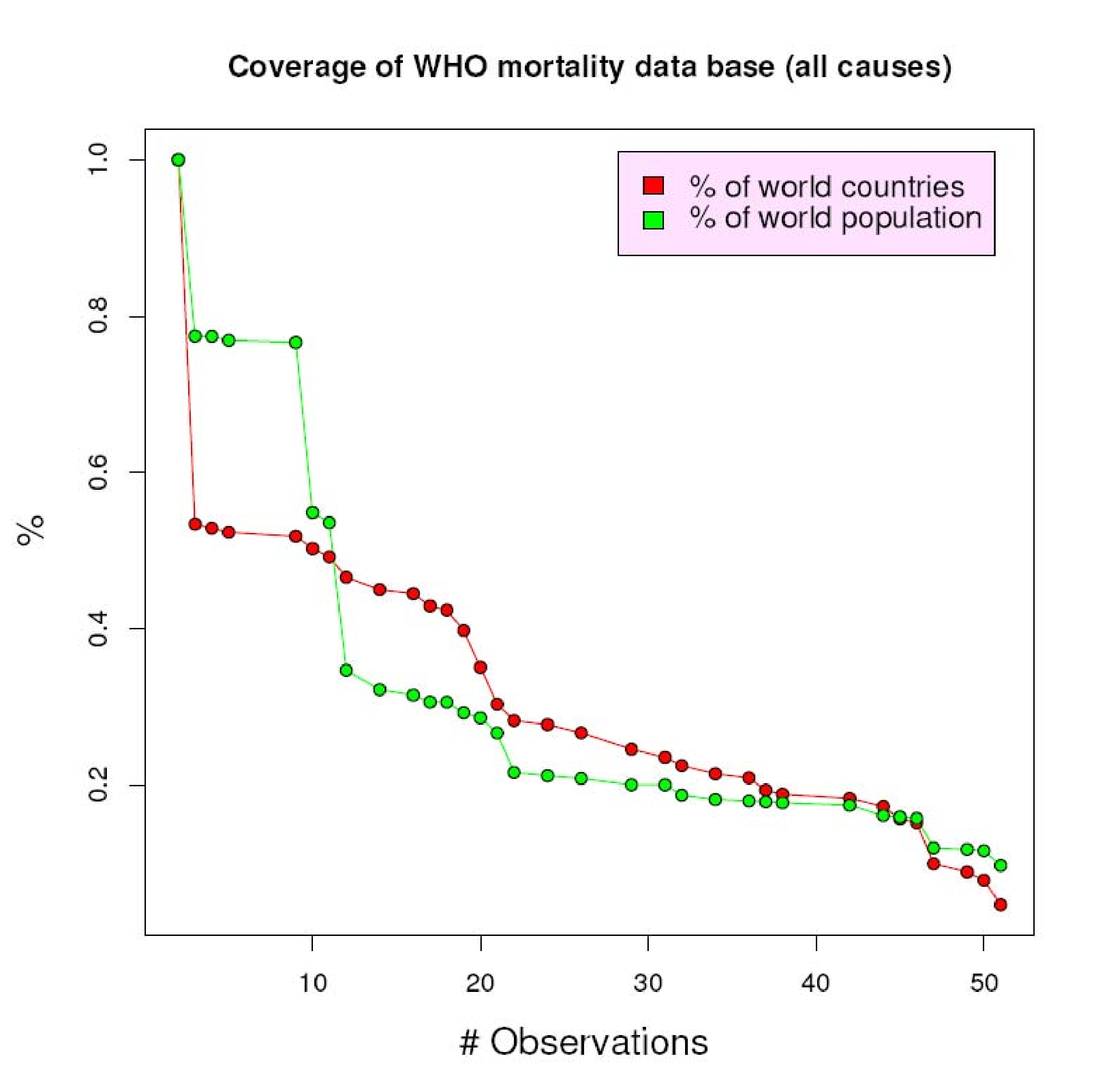
Pension planning

Allocation of public health resources

Planning manpower needs

Guidance for epidemiological studies

Problem: forecasting very short time series



The forecasting problem is set as a regression problem

$$\mu_{cat} \equiv \mathrm{E}[\log m_{cat}] = \beta'_{ca} x_{ca,t-T}$$

m_{cat} : Mortality in country c, age a and time t

β_{ca} : Regression coefficient

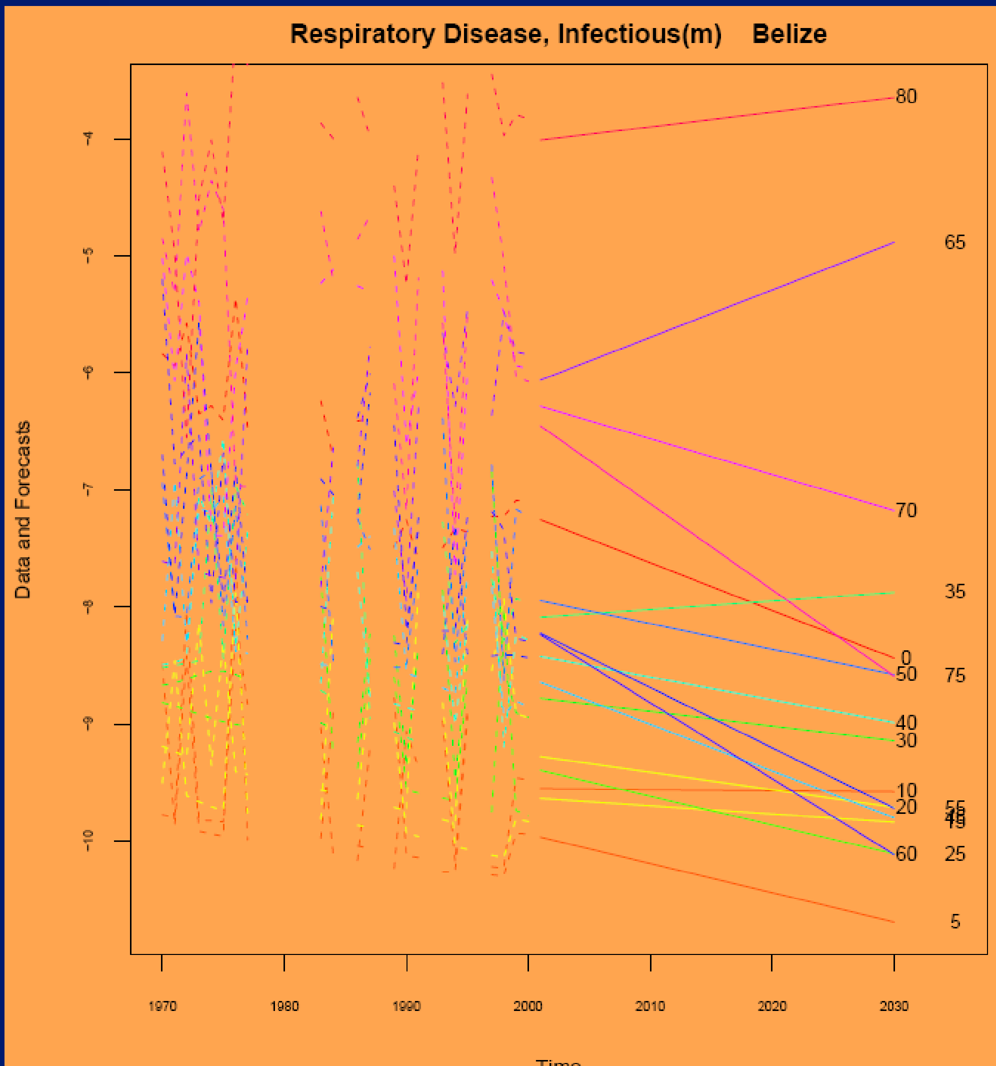
$x_{ca,t-T}$: Lagged covariates

Typical Lagged Covariates

$x_{ca,t-T}$: Lagged covariates

- GDP
- Human capital
- Fat consumption
- Water quality
- Cigarette consumption

In most cases some "pooling" is necessary



Regressions cannot be estimated separately across age groups or countries.

17 separate regressions (one for each age group)

Bad idea!

Those who have knowledge do not predict. Those who predict do not have knowledge

Lao Tzu, 6th century BC

The Standard Bayesian Approach

Likelihood: $$\mathrm{P}(y \mid \beta) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{cat} \left(\mu_{cat} - \beta' x_{cat} \right)^2 \right)$$

Prior $$\mathrm{P}(\beta) \propto \exp(-\lambda \mathrm{H}[\beta])$$

Posterior $$\mathrm{P}(\beta \mid y) \propto \mathrm{P}(y \mid \beta)\mathrm{P}(\beta)$$

$$\mathrm{P}(\beta \mid y) \propto \exp\left(-\left[\frac{1}{2\sigma^2} \sum_{cat} \left(\mu_{cat} - \beta' x_{cat} \right)^2 + \lambda \mathrm{H}[\beta] \right] \right)$$

A Way Out

- We do need some sort of prior on the β ...
- but we do not really have prior knowledge on β ...
- BUT we do have knowledge on μ!
- AND μ is related to β: μ = X β

$$\mu_{cat} \equiv \mathbf{E}[\log m_{cat}] = \beta'_{ca} x_{ca,t-T}$$

Strategy to build a prior

- Define a non-parametric prior for μ, as a function of the cross-sectional index (age, for example)

$$P(\mu) \propto \exp(-\lambda H[\mu])$$

- Use the relationship between μ and β (μ = X β) to change variables and obtain a prior for β

$$P(\beta) \propto \exp(-\lambda H[X\beta])$$

What type of prior knowledge?

- Mortality age profiles are smooth deformations of <u>well known shapes</u>
- Mortality varies smoothly <u>across countries</u>
- Mortality varies smoothly over time

A Good Prior on μ

$$H[\mu] = \int dt\,da \left(\frac{\partial^n [\mu(a,t) - \overline{\mu}(a)]}{\partial a^n} \right)^2$$

Discretizing age on a grid:

$$H[\mu] = \sum_{taa'} (\mu_{at} - \overline{\mu_a}) W^{(n)}_{aa'} (\mu_{a't} - \overline{\mu_{a'}})$$

Only a Step Away from Prior on β …

- The matrix *W* is fully determined by the order of the derivative n
- The "template" age profile $\bar{\mu}$ can be made disappear by subtracting if from the data
- Just need to substitute the specification $\mu = X\beta$

And the Prior for β is:

$$\mathrm{P}(\beta) \propto \exp\left(-\lambda \sum_{aa'} W_{aa'}^{(n)} \beta_a^{'} \mathrm{C}_{aa'} \beta_{a'} \right)$$

where

$$\mathrm{C}_{aa'} = \frac{1}{T} \mathrm{X}_a^{'} \mathrm{X}_{a'}$$

But What Does the Prior Really Mean?

$$\mathrm{P}(\mu) \propto \exp(-\lambda \mathrm{H}[\mu])$$

$$\mathrm{H}[\mu] = \int dt\,da \left(\frac{\partial^n [\mu(a,t) - \overline{\mu}(a)]}{\partial a^n} \right)^2$$

But What Does the Prior Really Mean?

Discretizing over age and fixing one year in time μ is simply a vector of random variables

$$P(\mu) \propto \exp\left(-\lambda \sum_{aa'} \mu_a W^{(n)}_{aa'} \mu_{a'} \right)$$

How do the samples from this prior look like?

Demos

- Samples from prior with zero mean
- Samples from prior with non zero mean

And what is the role of λ?

Two important, related identities

$$\mathrm{E}[\mathrm{H}[\mu]] = \frac{\operatorname{rank}(W^{(n)})}{\lambda}$$

$$\frac{1}{A}\sum_a \mathrm{E}[\mu_a^2] = \frac{\operatorname{tr}\left(\left(W^{(n)}\right)^+\right)}{A\lambda}$$

The role of λ

- λ determines the size of the smoothness functional
- λ determines the average standard deviation of the prior

Demos

- Standard deviation of the prior
- Samples from prior with non zero mean: varying the smoothness parameter

Other Types of Priors

- **Time**

$$H[\mu] = \int dtda \left(\frac{\partial^n \mu(a,t)}{\partial t^n} \right)^2$$

- **Time trends over age**

$$H[\mu] = \int dtda \left(\frac{\partial^{n+m} \mu(a,t)}{\partial t^n \partial^m a} \right)^2$$

Dealing with Multiple Smoothness Parameters

- **Writing the priors is easy …**

- **Estimating the 3 smoothing parameters is very difficult**

  - **Cross validation is hard to do with very short time series**

- **Some prior knowledge on the smoothing parameters is needed**

Estimating the smoothness parameters

- **Key observation:** the smoothness parameters control ALL expected values of the prior

$$\mathrm{E}[F_1(\mu)] = g_1(\lambda_1, \lambda_2, \lambda_3)$$

$$\mathrm{E}[F_2(\mu)] = g_2(\lambda_1, \lambda_2, \lambda_3)$$

$$\mathrm{E}[F_3(\mu)] = g_3(\lambda_1, \lambda_2, \lambda_3)$$

Estimating the smoothness parameters

- Sometimes we do have other forms of prior knowledge

  - How much the dependent variables changes from one cross section (or year) to the next

$$F_1(\mu) = \sum_{at} | \mu_{at} - \mu_{a+1,t} |$$

Estimating the smoothness parameters

- Expected values of any function of μ can be estimated empirically, by sampling the prior
- The following equations can be solved numerically:

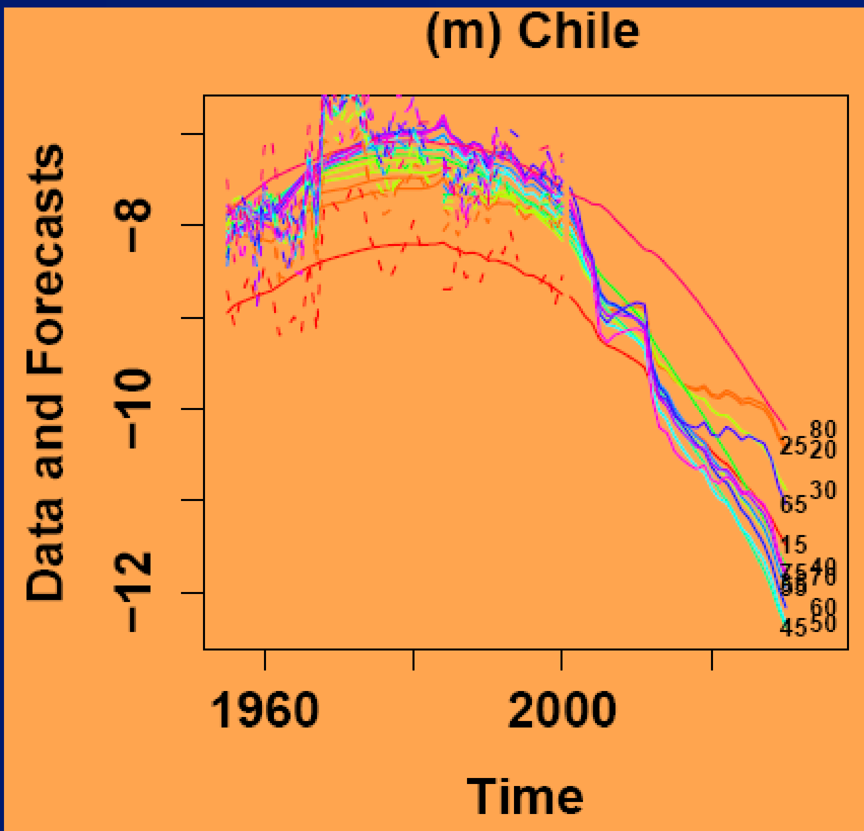
$$\mathrm{E}[F_1(\mu)] = g_1(\lambda_1, \lambda_2, \lambda_3)$$

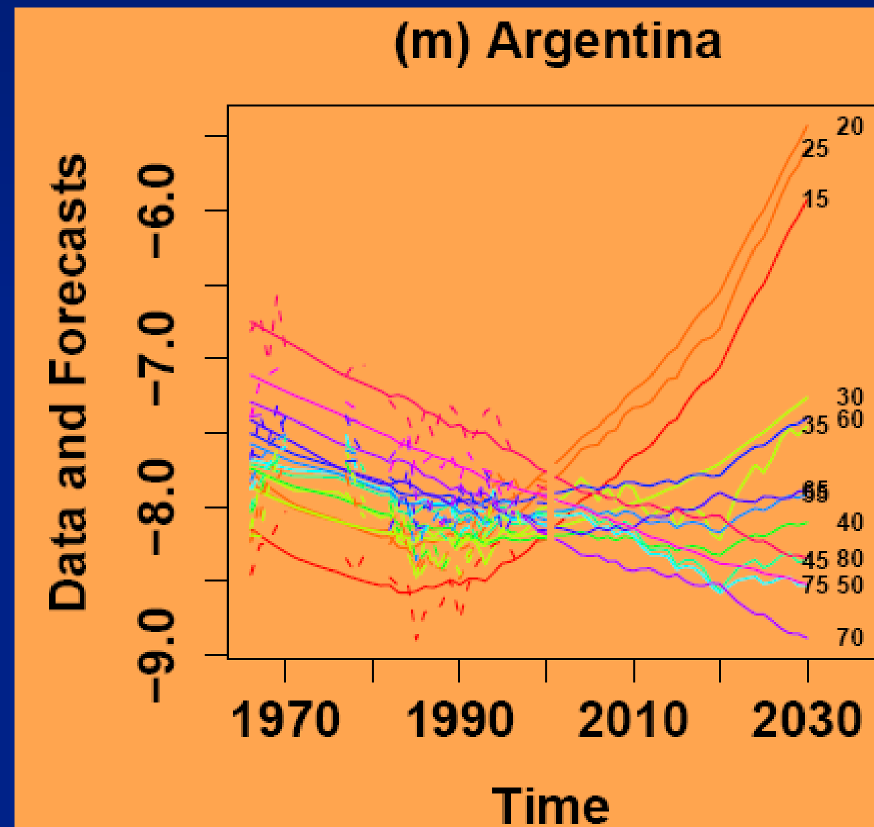
$$\mathrm{E}[F_2(\mu)] = g_2(\lambda_1, \lambda_2, \lambda_3)$$

$$\mathrm{E}[F_3(\mu)] = g_3(\lambda_1, \lambda_2, \lambda_3)$$

Demo: Deaths by Transportation Accidents in Chile

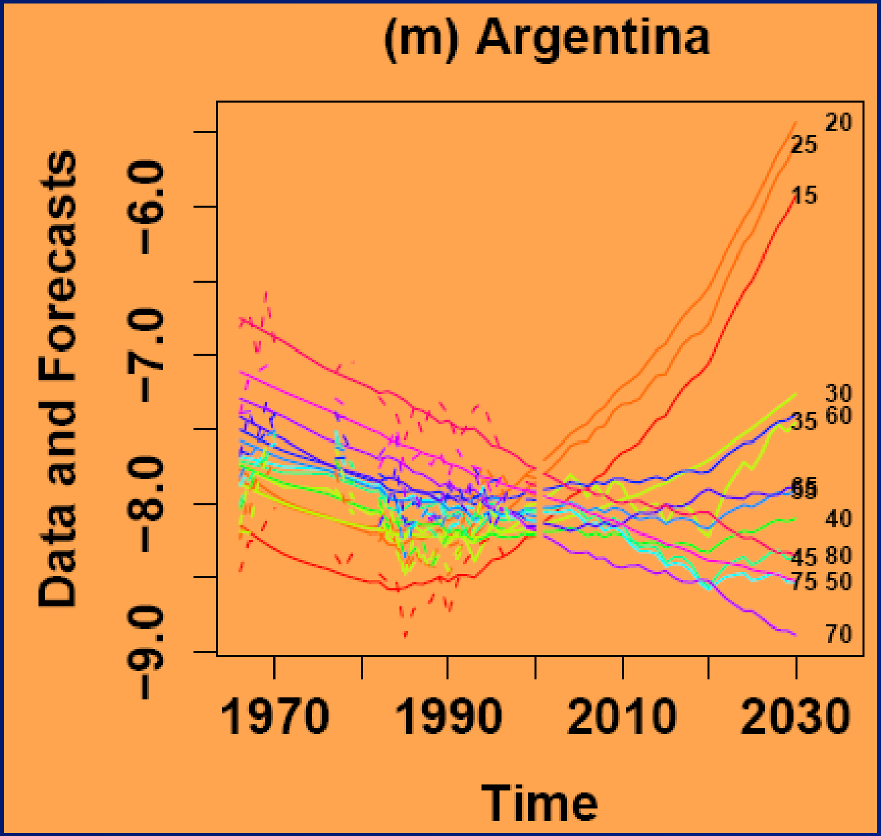
Transportation Accidents: no pooling



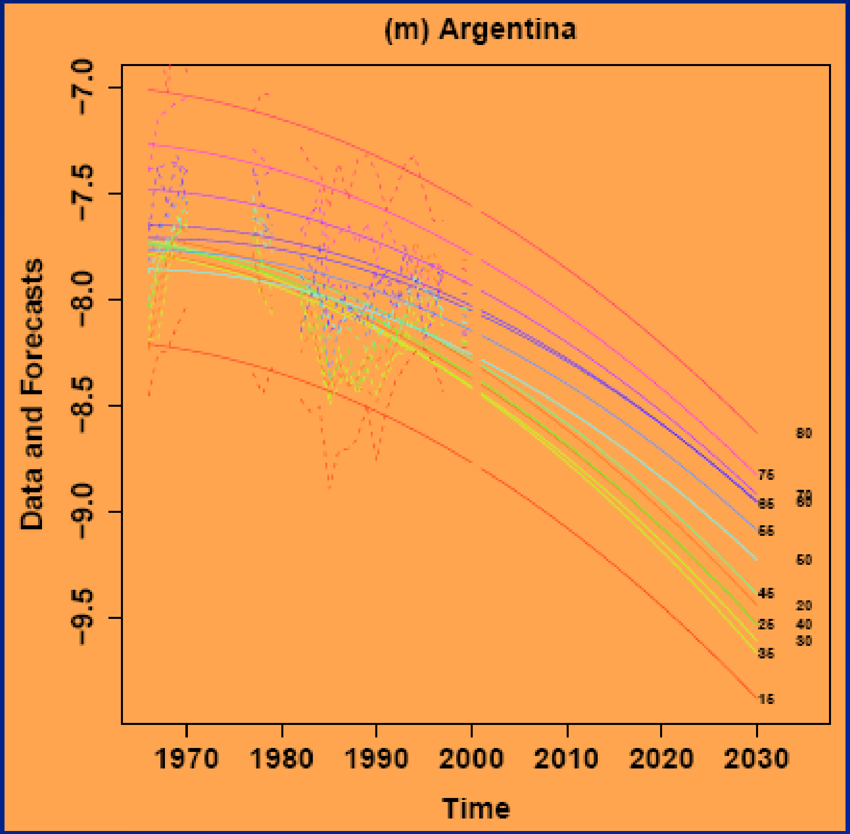


Pooling Over Countries: Transportation Accidents in Argentina

No Pooling



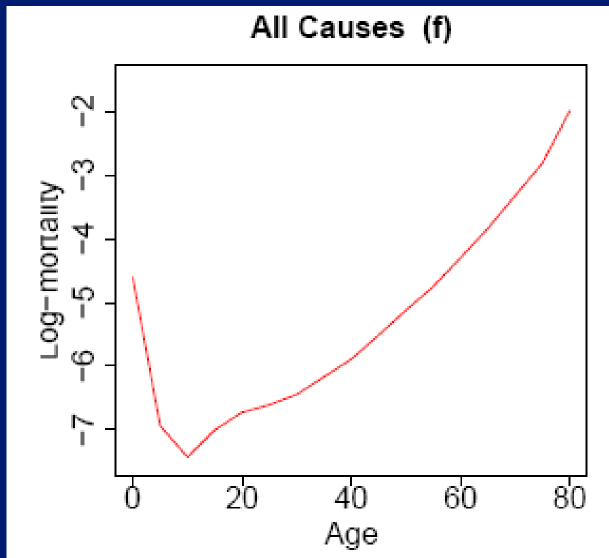
Pooling

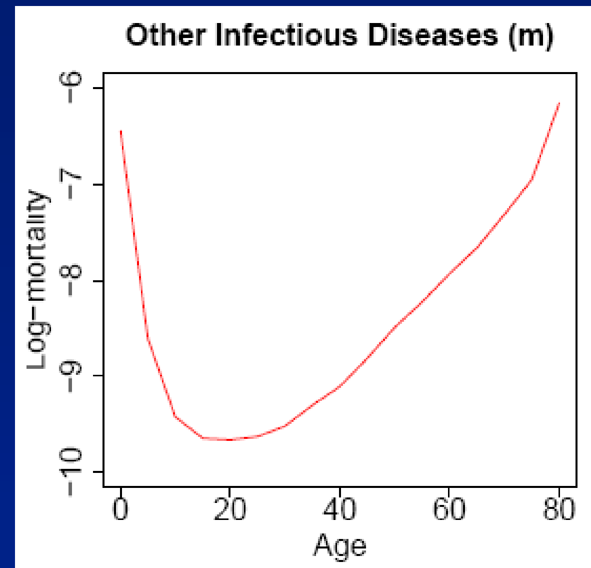


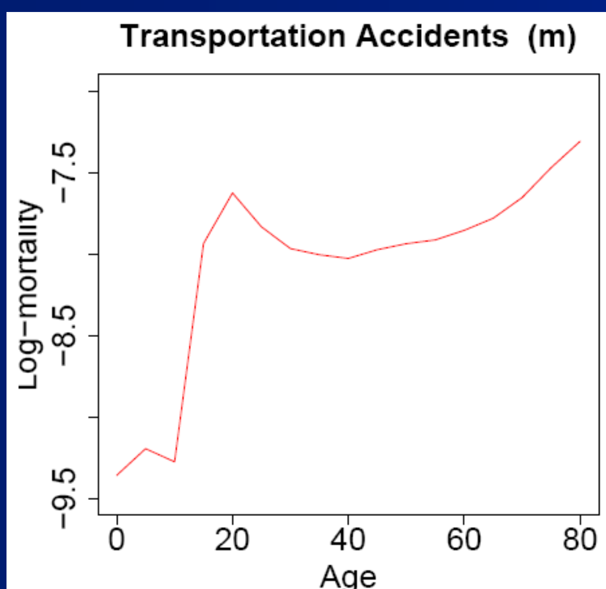
Summary

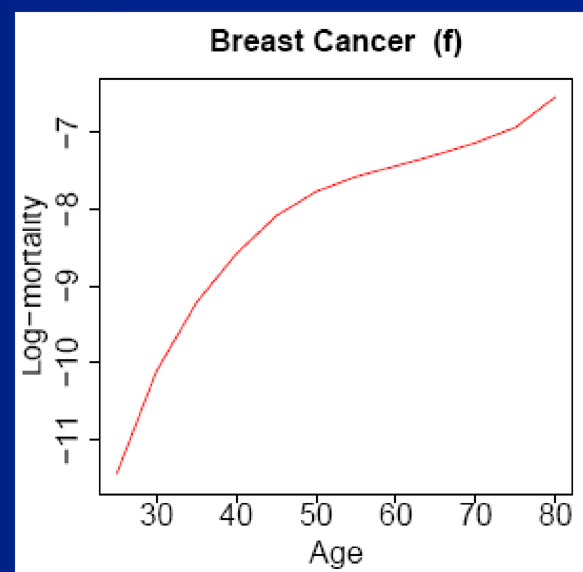
- Regularization theory is a powerful framework that reaches beyond standard pattern recognition
- In some application it is important to pay attention to the precise nature of the prior
- Prior knowledge applies to the smoothness parameter too

Mortality age profiles are well known and consistent across countries and time

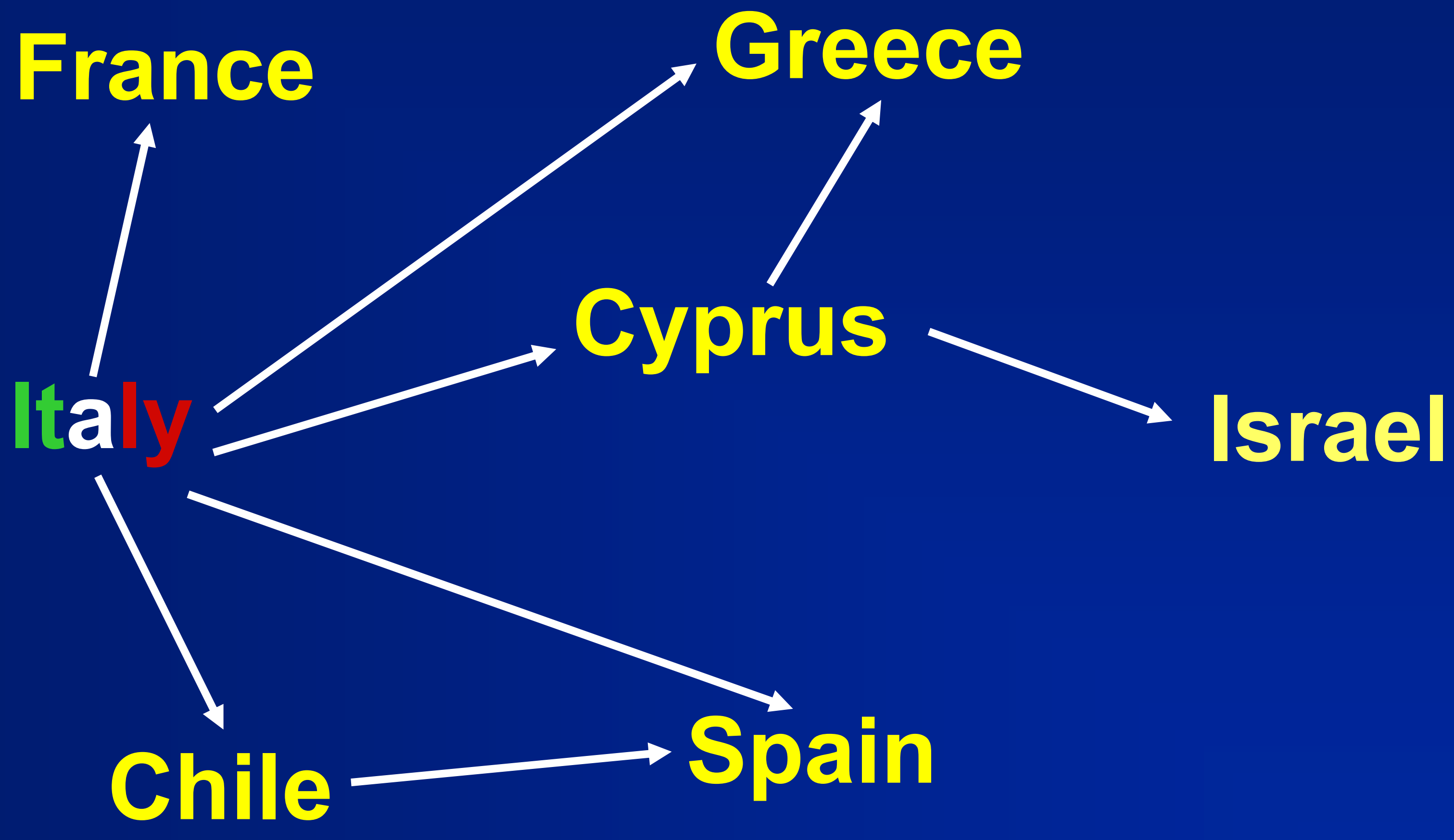








"Similar" countries have similar mortality patterns



Before and After the Cure
Respiratory Infections in Belize