2021 SISCER: Age-Period-Cohort Modeling and Analysis
Lecture 3: Splines, Smoothing, Bayes and INLA

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

In the `Epi` package implementation of APC models, there is an emphasis on spline modeling.

In the `BAPC` package, Bayesian fitting is carried out using random walk of order 2 (RW2) models, and the INLA method for summarizing posterior distributions.

Spline Models

Spline Modeling

- Factor models have lots of parameters and do not impose any form of smoothing that respects the ordering of age, period or cohort.
- Spline models are based on piecewise polynomial fitting and are extremely popular.
- The following description is taken from Chapter 11 of **?**).
- Within the `Epi` package, linear, natural and *B*-splines may be fitted.

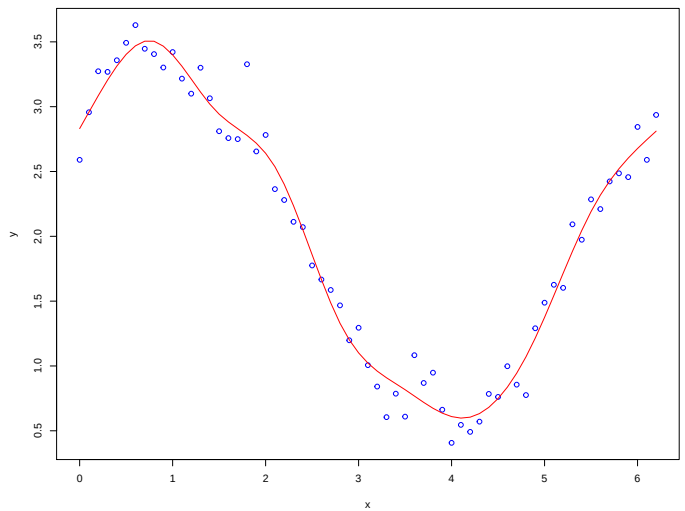


Figure 1: Simulated data: natural spline fit with 10 knots.

Piecewise Polynomials and Splines

For data that we referred to as continuous responses, splines are simply linear models, but with an enhanced basis set that provides flexibility.

Let $h_j(x) : \mathbb{R} \to \mathbb{R}$ denote the m-th function of x, for $j = 1, ..., J$.

A generic linear model consists of the linear basis expansion in x:

$$f(x) = \sum_{j=1}^{J} \beta_j h_j(x).$$

An obvious choice of basis is a polynomial of degree $J - 1$, but the global behavior of such a choice can be poor.

However, local behavior can be well represented by relatively low order polynomials.

Light Detection and Ranging Example

We illustrate various spline models using data, taken from **?**), from a light detection and ranging (LIDAR) experiment.

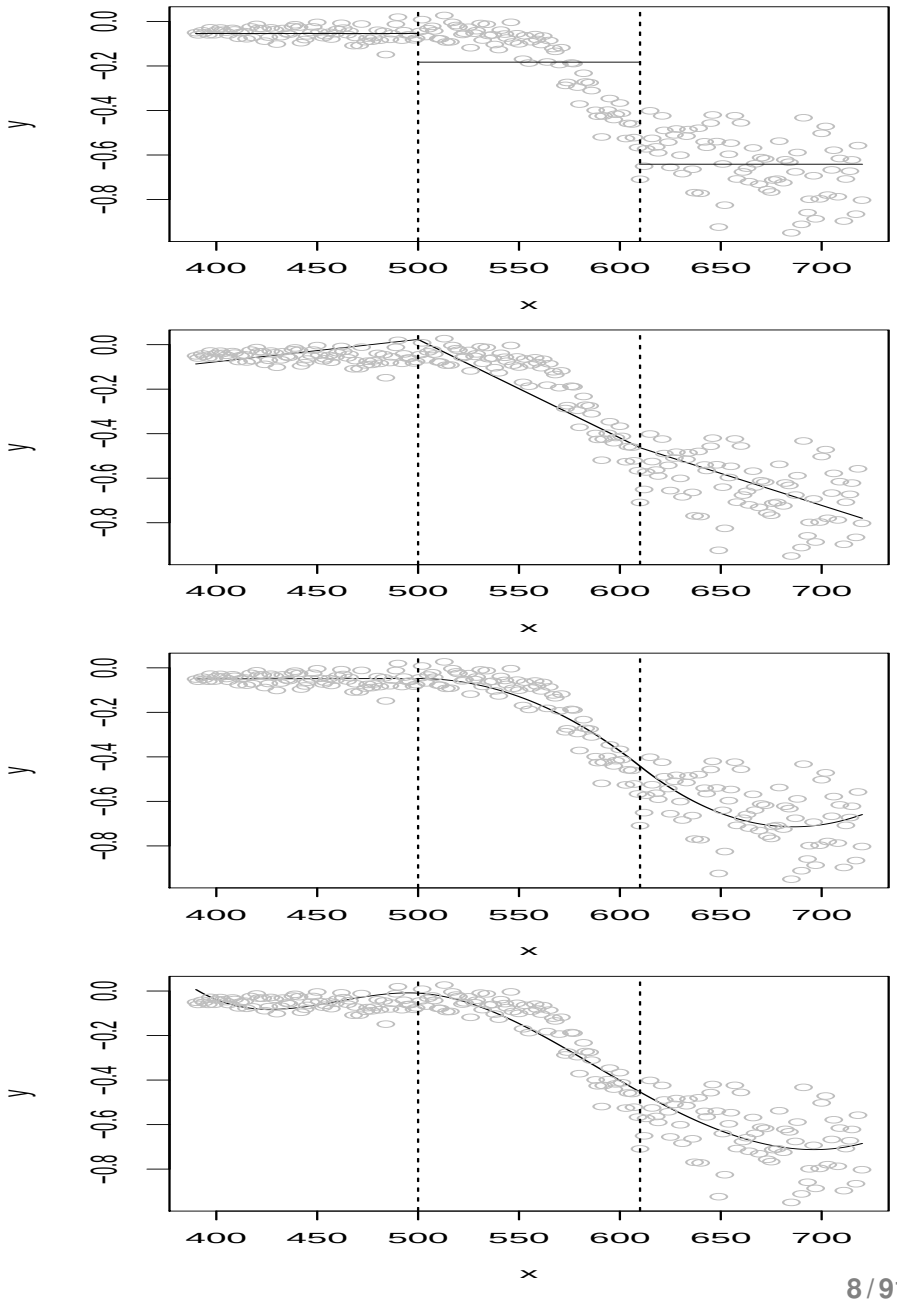
The LIDAR technique (which is similar to radar technology) uses the reflection of laser-emitted light to monitor the distribution of atmospheric pollutants.

The data we consider concern mercury. The x axis measures distance traveled before light is reflected back to its source (and is referred to as the range), and the y axis is the logarithm of the ratio of distance measured for two laser sources: one source has a frequency equal to the resonant frequency of mercury, and the other has a frequency off this resonant frequency.

For these data, point and interval estimates for the association between the log ratio and range are of interest.

Light Detection and Ranging Example

- To motivate spline models, we fit piecewise constant, linear, quadratic and cubic models using least squares, with three pieces in each case.
- The piecewise linear model is shown at the top: By forcing the curve to be continuous but only allowing linear segments we see that the fit is not good (particularly in the first segment). The lack of smoothness is also undesirable.
- The quadratic and cubic fits in panels 2 and 3 are far more visually appealing, though neither provide satisfactory fits, because we have only allowed three piecewise polynomials. In particular, in panel 4 the cubic fit is still poor at the left endpoint.



Piecewise Polynomials and Splines

We now start the description of spline models by introducing some notation.

Let $\xi_1 < \xi_2 < ... < \xi_K$ be a set of ordered points, called knots, contained in some interval (a, b).

An M-th order spline is a piecewise $M - 1$ degree polynomial with $M - 2$ continuous derivatives at the knots.

Splines are very popular in nonparametric modeling though, as we shall see, care is required in choosing the degree of smoothing.

The latter depends on a variety of factors including the order of the spline, and the number and position of the knots.

Piecewise Polynomials and Splines

We begin with a discussion of the order of the spline. The most basic piecewise polynomial is a piecewise constant function, which is an order-1 spline.

With two knots, ξ_1 and ξ_2, there are three basis functions:

$$h_1(x) = I(x < \xi_1), \quad h_2(x) = I(\xi_1 \leq x < \xi_2), \quad h_3(x) = I(\xi_2 \leq x)$$

where $I(\cdot)$ is the indicator function. Note that there are no continuous derivatives at the knots.

To obtain linear models in each of the intervals we may introduce three additional bases

$$h_{3+j} = h_j(x)x, \quad j = 1, 2, 3$$

to give the model

$$f(x) = I(x < \xi_1)(\beta_1 + \beta_4 x) + I(\xi_1 \leq x < \xi_2)(\beta_2 + \beta_5 x) + I(\xi_2 \leq x)(\beta_3 + \beta_6 x),$$

which contains six parameters.

Piecewise Polynomials and Splines

The lack of continuity is a problem with this model, but we can impose two constraints to enforce

$$f(\xi_1^-) = f(\xi_1^+)$$

and

$$f(\xi_2^-) = f(\xi_2^+),$$

which implies

$$\begin{aligned} \beta_1 + \xi_1\beta_4 &= \beta_2 + \xi_1\beta_5 \\ \beta_2 + \xi_2\beta_5 &= \beta_3 + \xi_2\beta_6 \end{aligned}$$

to give four parameters in total.

Piecewise Polynomials and Splines

A neater way of incorporating these constraints is with the basis:

$$h_1(x) = 1, \quad h_2(x) = x, \quad h_3(x) = (x - \xi_1)_+, \quad h_4(x) = (x - \xi_2)_+ \quad (1)$$

where t_+ denotes the positive part.

The generic basis $(x - \xi)_+$ is sometimes referred to as a truncated line.

The resultant function

$$f(x) = \beta_0 + \beta_1 x + \beta_2(x - \xi_1)_+ + \beta_3(x - \xi_2)_+$$

is continuous at the knots, since all prior basis functions are contributing to the fit up to any particular x value.

Piecewise Polynomials and Splines

The model defined by the basis (1) is an order-2 spline and the first derivative is discontinuous.

Figure 2 shows the basis functions for this representation.

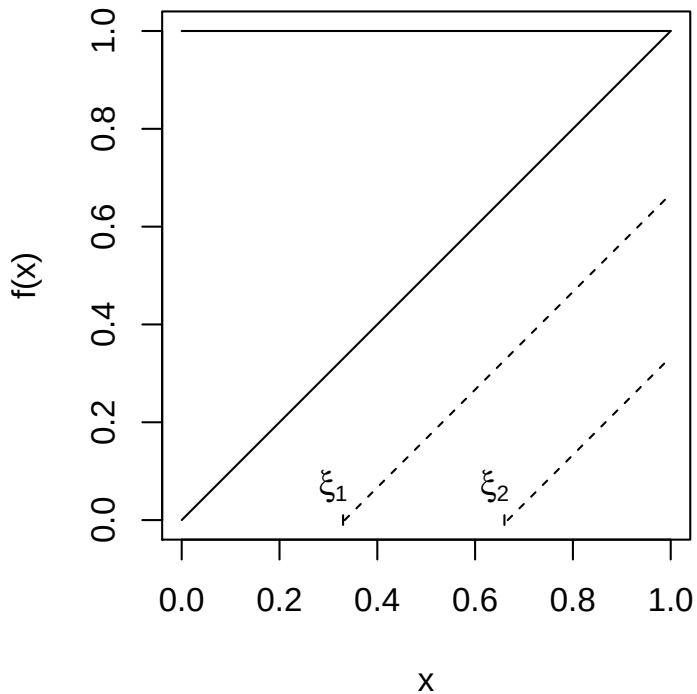


Figure 2: Basis functions for piecewise linear model with two knots at ξ_1 and ξ_2. The solid lines are the bases 1 and x, and the dashed lines are the bases $(x - \xi_1)_+$ and $(x - \xi_2)_+$.

Piecewise Polynomials and Splines

We now consider how the piecewise linear model may be extended. Naively, we might assume the quadratic form:

$$f(x) = \beta_0+\beta_1 x+\beta_2 x^2+\beta_3(x-\xi_1)_+ +\beta_4(x-\xi_1)_+^2+\beta_5(x-\xi_2)_+ +\beta_6(x-\xi_2)_+^2,$$

which is continuous, but has first derivative

$$f'(x) = \beta_1+2\beta_2 x+\beta_3 I(x>\xi_1)+2\beta_4(x-\xi_1)_+ +\beta_5 I(x>\xi_2)+2\beta_6(x-\xi_2)_+,$$

which is discontinuous at the knot points ξ_1 and ξ_2, and is undesirable because of the lack of smoothness. Hence, we drop the truncated linear bases to give

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3(x-\xi_1)_+^2 + \beta_4(x-\xi_2)_+^2$$

which has continuous first derivative,

$$f'(x) = \beta_1 + 2\beta_2 x + 2\beta_3(x-\xi_1)_+ + 2\beta_4(x-\xi_2)_+.$$

The second derivative is discontinuous, however, which may also be undesirable.

Piecewise Polynomials and Splines

Hence, a popular form is a cubic spline.

We will concentrate on cubic splines in some detail and so we introduce a slight change of notation, with respect to the truncated cubic parameters. With two knots the function and first three derivatives are

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + b_1 (x - \xi_1)_+^3 + b_2 (x - \xi_2)_+^3 \\ f'(x) &= \beta_1 + 2\beta_2 x + 3\beta_3 x^2 + 3b_1 (x - \xi_1)_+^2 + 3b_2 (x - \xi_2)_+^2 \\ f''(x) &= 2\beta_2 + 6\beta_3 x + 6b_1 (x - \xi_1)_+ + 6b_2 (x - \xi_2)_+ \\ f'''(x) &= 6\beta_3 + 6b_1 I(x > \xi_1) + 6b_2 I(x > \xi_2). \end{aligned}$$

The latter is discontinuous, with a jump at the knots.

Figure 3 shows the basis function for the cubic spline, with two knots, and Figure 4 the fit to the LIDAR data.

Piecewise Polynomials and Splines

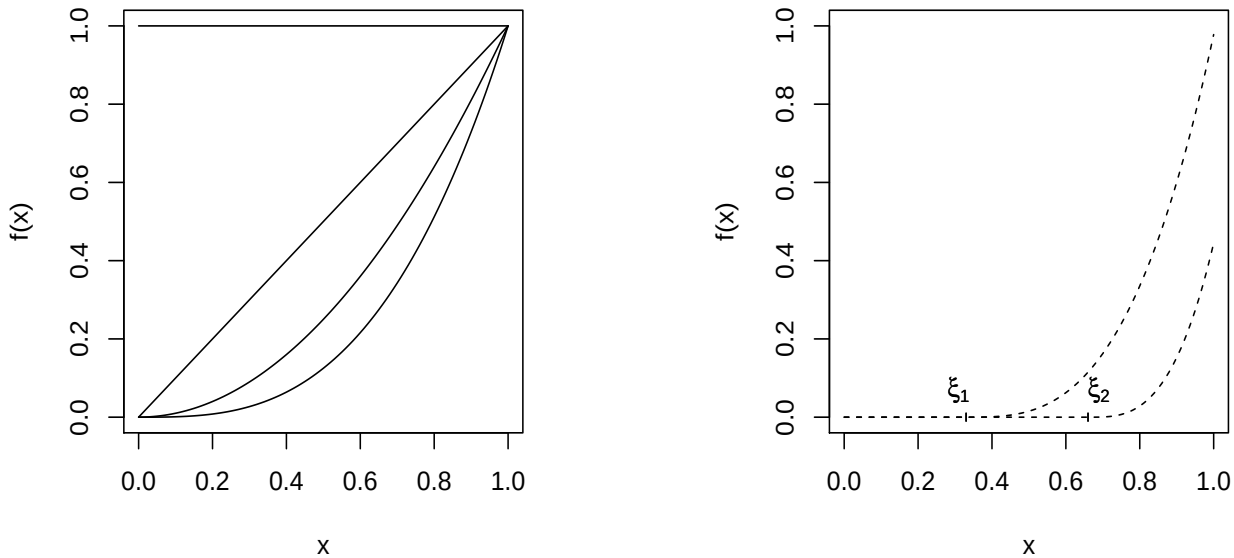


Figure 3: Basis functions for a piecewise cubic spline model, with two knots at ξ_1 and ξ_2. Panel (a) shows the bases 1, x, x^2, x^3, and panel (b) the bases $(x-\xi_1)^3_+$ and $(x-\xi_2)^3_+$.

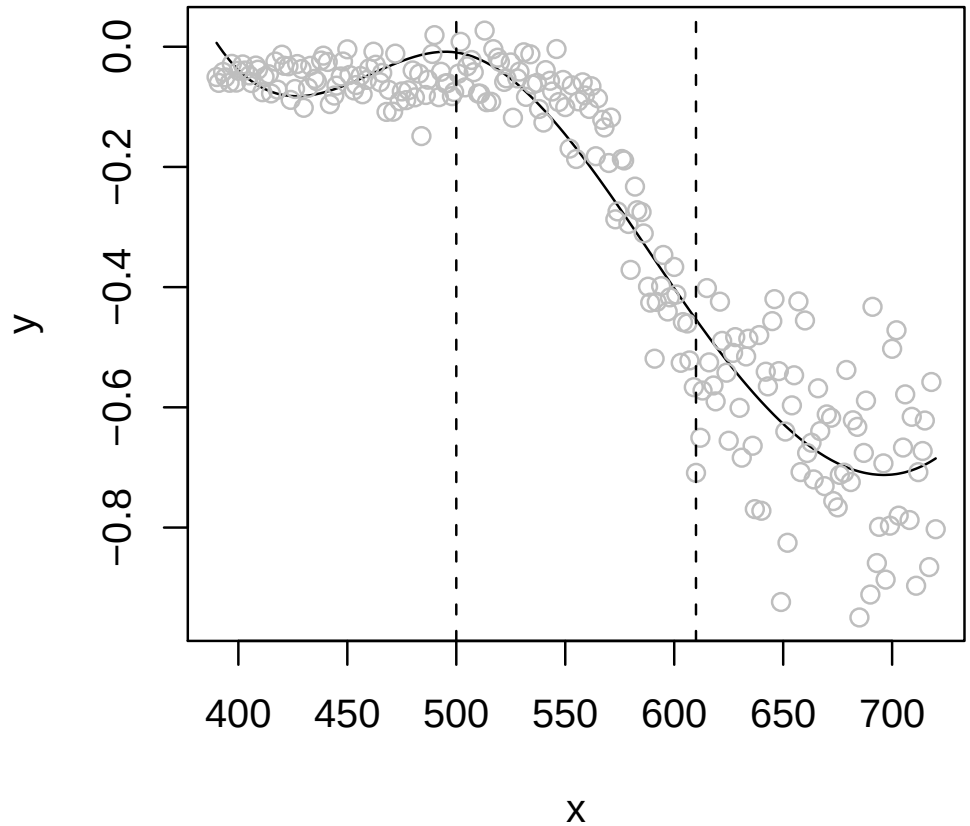


Figure 4: Piecewise cubic fit to LIDAR data.

Cubic Splines

For K knots we write the cubic spline function as

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} b_k (x - \xi_k)_+^3, \qquad (2)$$

so that we have $K + 4$ coefficients.

We simply have a linear model, $f(x) = \mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{c}] = \boldsymbol{c}\gamma$, where

$$\boldsymbol{c} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & (x_1 - \xi_1)_+^3 & \dots & (x_1 - \xi_K)_+^3 \\ 1 & x_2 & x_2^2 & x_2^3 & (x_2 - \xi_1)_+^3 & \dots & (x_2 - \xi_K)_+^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & (x_n - \xi_1)_+^3 & \dots & (x_n - \xi_K)_+^3 \end{bmatrix}, \quad \gamma = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ b_1 \\ \vdots \\ b_K \end{bmatrix}.$$

Estimator: $\widehat{\gamma} = (\boldsymbol{c}^\top \boldsymbol{c})^{-1} \boldsymbol{c}^\top \boldsymbol{Y}$. Linear smoother: $\widehat{\boldsymbol{Y}} = \boldsymbol{S}\boldsymbol{Y}$, $\boldsymbol{S} = \boldsymbol{c}(\boldsymbol{c}^\top \boldsymbol{c})^{-1} \boldsymbol{c}^\top$.

Natural Cubic Splines

Spline models such as (2) can produce erratic behavior beyond the extreme knots.

A natural spline enforces linearity beyond the boundary knots, i.e.

$$\begin{aligned} f(x) &= a_1 + a_2 x \quad \text{for} \quad x \leq \xi_1 \\ f(x) &= a_3 + a_4 x \quad \text{for} \quad x \geq \xi_K . \end{aligned}$$

The first condition only considers values of x before the knots, and therefore the b_k parameters in (2) are irrelevant.

Natural Cubic Splines

It is straightforward to see that for linear before $x \leq \xi_1$ we require

$$\beta_2 = \beta_3 = 0. \tag{3}$$

For $x \geq \xi_K$:

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \sum_{k=1}^{K} b_k (x - \xi_k)^3 \\ &= \beta_0 + \beta_1 x + \sum_{k=1}^{K} b_k (x^3 - 3x^2\xi_k + 3x\xi_k^2 - \xi_k^3), \end{aligned}$$

and so for linearity

$$\sum_{k=1}^{K} b_k = \sum_{k=1}^{K} b_k \xi_k = 0, \tag{4}$$

to get rid of the x^3 and x^2 terms.

Hence, we have four additional constraints in total, so that the basis for a natural cubic spline has K elements.

Cubic Smoothing Splines

We now present a formal derivation of the natural cubic spline.

Result: Consider the penalized least squares criterion

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx, \tag{5}$$

where the second term penalizes the *roughness* of the curve, and λ controls the degree of roughness.

It is clear that without penalization we could choose an infinite number of curves that interpolate the data (in the case of unique x values at least), with arbitrary behavior in between.

The $f(\cdot)$ that minimizes (5) is the natural cubic spline with knots at the unique data points, we call this function $g(x)$.

Proof is in **?**).

Cubic Smoothing Splines

We stress that the fitted natural cubic smoothing spline will not typically interpolate the data, and the level of smoothness will be determined by the value of λ chosen.

Low values of λ (large effective degrees of freedom), impose little smoothness and bring the fit closer to interpolation.

In terms of interpretation, if a thin piece of flexible wood (a mechanical spline) is placed over the points (x_i, y_i), $i = 1, ..., n$, then the position taken up by the piece of wood will be of minimum energy, and will describe a curve that is approximately minimizes $\int f''^2$, over curves that interpolate the data.

Example: Light Detection and Ranging

- For a natural cubic spline to the LIDAR data, the top figure shows the ordinary and generalized cross-validation scores, respectively) versus the effective degrees of freedom.
- The curves are very similar with well-defined minima.
- The OCV and GCV scores are minimized at 9.3 and 9.4 effective degrees of freedom, respectively.
- The lower plot fit (using the GCV minimum, which corresponds to $\widehat{\lambda} = 959$), appears good. In particular we note that the boundary behavior is reasonable.

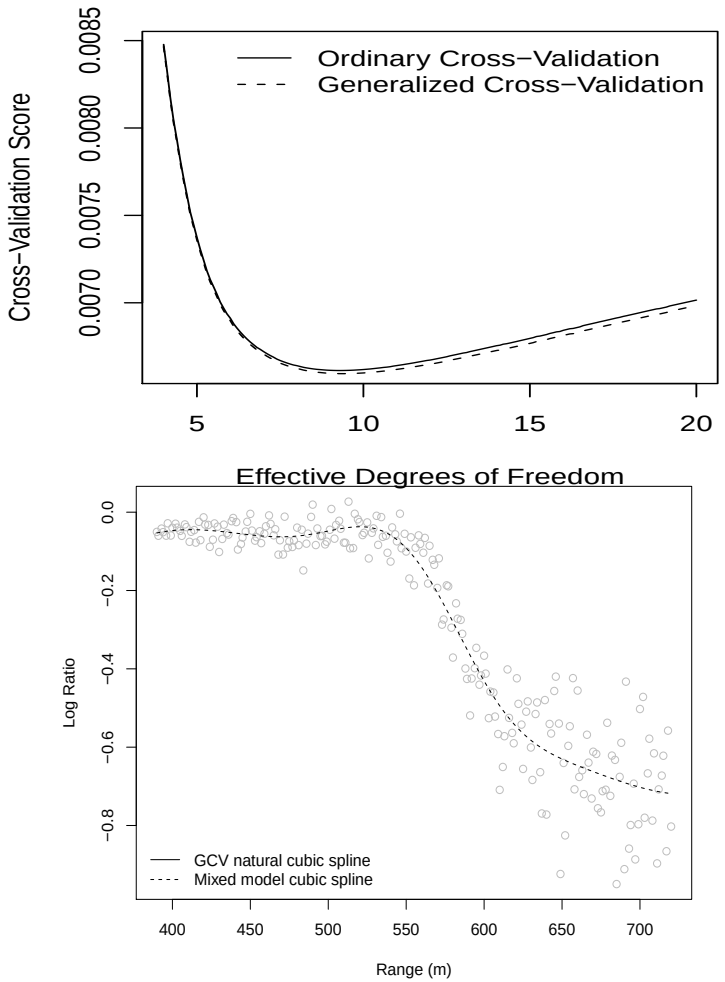


Figure 5: OCV and GCV scores vs effective degrees of freedom, for natural cubic spline and fit to LIDAR data.

B-Splines

There are many ways of choosing a basis to represent a cubic spline; the so-called *B*-spline basis functions are popular, a primary reason being that they are non-zero over a limited range, which aids in computation.

B-splines also form the building blocks for other spline models.

B-splines are available for splines of general order, which we again denote by M (so that for a cubic spline, $M = 4$).

The number of bases functions is $K + M$ since we have an $M - 1$ degree polynomial (giving M bases), and one basis for each knot.

The original set of knots are denoted ξ_k, $k = 1, ..., K$, and we let $\xi_0 < \xi_1$ and $\xi_K < \xi_{K+1}$ represent two boundary knots.

B-Splines

We define an augmented set of knots, τ_j, $j = 1, ..., K + 2M$, with

$$\begin{aligned} \tau_1 \leq \tau_2 \leq ... \leq \tau_M &\leq \xi_0 \\ \tau_{j+M} &= \xi_j, \; j = 1, ..., K \\ \xi_{K+1} &\leq \tau_{K+M+1} \leq \tau_{K+M+2} \leq \cdots \leq \tau_{K+2M} \end{aligned}$$

where the choice of the additional knots is arbitrary and so we may, for example, set

$$\tau_1 = ... = \tau_M = \xi_0$$

and

$$\xi_{K+1} = \tau_{K+M+1} = ... = \tau_{K+2M}.$$

These additional knots ensure the bases functions detailed below are defined close to the boundaries.

B-Splines

To construct the bases, first define

$$B_j^1(x) = \begin{cases} 1 & \text{if } \tau_j \leq x < \tau_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

for $j = 2, ..., K + 2M - 1$. For $1 < m \leq M$ define

$$B_j^m(x) = \frac{x - \tau_j}{\tau_{j+m-1} - \tau_j} B_j^{m-1} + \frac{\tau_{j+m} - x}{\tau_{j+m} - \tau_{j+1}} B_{j+1}^{m-1} \tag{7}$$

for $j = 1, ..., K + 2M - m$. If we divide by zero then we define the relevant basis element to be zero.

The *B*-spline bases are non-zero over a domain spanned by at most $M + 1$ knots.

For example, the support of cubic *B*-splines ($M = 4$) is at most five knots. At any x, M of the *B*-splines are non-zero.

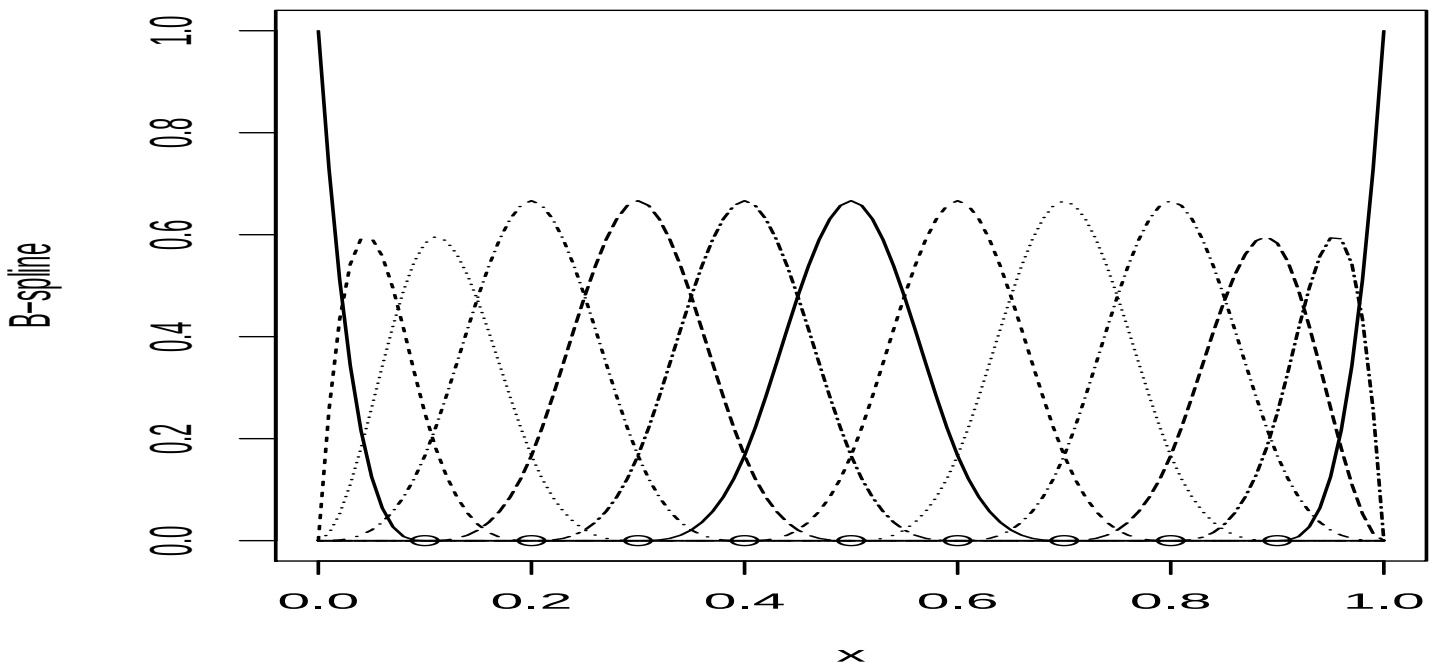
B-Splines

The cubic *B*-spline model is

$$f(x) = \sum_{j=1}^{K+4} B_j^4(x)\beta_j. \tag{8}$$

For further details on computation, see Hastie et al. (2009, p.186).

Figure 6 shows the cubic *B*-spline basis (excluding the intercept) with $K = 9$ knots.



Penalized Regression Splines

Although the result on natural cubic splines is of theoretical interest, in general we would like to have a functional form that has less parameters than data points.

Regresssion splines are defined with respect to a reduced set of $K < n$ knots.

Deciding on the number and location of knots automatically is difficult.

For example, starting with n knots and then selecting via stepwise methods is fraught with difficulties since there are 2^n models to choose from (assuming the intercept and linear terms are always present).

Penalized Regression Splines

An alternative penalized regression spline approach, with $K < n$ knots is to choose sufficient knots for flexibility, but then to penalize the parameters associated with the knot bases.

If this approach is followed the number and selection of knots is far less important than the choice of smoothing parameter.

An obvious choice is to place an L_2 penalty on the coefficients, i.e. $\lambda \sum_{k=1}^{K} b_k^2$.

The resultant low rank smoothers use considerably less than n basis functions.

A Brief Spline Summary

The terminology associated with splines can be confusing, so we provide a brief summary.

For simplicity we assume that the covariate x is univariate, and that $x_1, \ldots, x_n$ are unique.

- A smoothing spline contains n knots, and
- a cubic smoothing spline is piecewise cubic.
- A natural spline is linear beyond the boundary knots.
- If there are $K < n$ knots we have a regression spline.
- A penalized regression spline imposes a penalty on the coefficients associated with the coefficients of the piecewise polynomial. The penalty terms may take a variety of forms.

The number of bases functions that define the spline depends on the number of knots and the degree of the polynomial, with a reduced number of bases if a natural spline.

Spline models may be parameterized in many different ways.

Parameterization of the Spline Model

?) is a strong advocate for the use of spline models in age-period-cohort modeling:

- Fewer parameters than factor models.
- Smooth functions of time variables.
- Can be used with unequally-spaced data.

But which type of splines to use, and how to choose knots/smooth?

Parameterization of the Spline Model

The following is based on Section 6.2 of **?**).

Recommendations are:

1. The age function should be interpretable as log age-specific rates in cohort c_0 (a reference cohort) after adjustment for the period effect.
2. The cohort function is 0 at a reference cohort c_0, and so is interpretable as the log relative rate, relative to cohort c_0.
3. 3.1 The period function is 0 on average with 0 slope, and so is interpretable as the log relative rate, relative to the age-cohort prediction (the residual log relative rate).
   3.2 Alternatively, the period function could be constrained to be 0 at a reference date, p_0. In this case the age-effects at $a_0 = A + p_0 - c_0$ would equal the fitted rate for period p_0 (and cohort c_0), and the period effects would be residual log relative rates relative to p_0.

Parameterization of the Spline Model

The second choice fixes one point on the curve(0 at c_0), and the third fixes a level (0 on average or 0 at p_0) and a slope (0 slope for the period function).

The inclusion of the slope (drift) with the cohort effect makes the age-effects interpretable as cohort-specific rates of disease (longitudinal rates).

Depending on the subject matter, the role of cohort and period could be interchanged, in which case the age-effects would be cross-sectional rates for the reference period.

Spline Model

Table 1 gives summaries from the spline model which was fitted in the `Epi` packge with the call:

```
apc.fit( dfEpi,npar=5,model="ns",dr.extr="Holford",parm="ACP")
```

This fits a natural spline model with 5 degrees of freedom.

	Resid. Df	Resid. Dev	Df	Deviance	p-value
Age	105	15242.0			
Age-drift	104	6564.0	1	8678.0	$< 2.2 \times 10^{-16}$
Age-Cohort	101	1016.4	3	5547.6	$< 2.2 \times 10^{-16}$
Age-Period-Cohort	98	419.3	3	597.1	$< 2.2 \times 10^{-16}$
Age-Period	101	2910.5	-3	-2491.3	$< 2.2 \times 10^{-16}$
Age-drift	104	6564.0	-3	-3653.5	$< 2.2 \times 10^{-16}$

Table 1: Spline models for Danish male lung cancer data.

The conclusions are similar to the factor model, though the fitted curves are smoother in the extremes of the data.

The overall fit is not good (419 on 98 df) for the APC model.

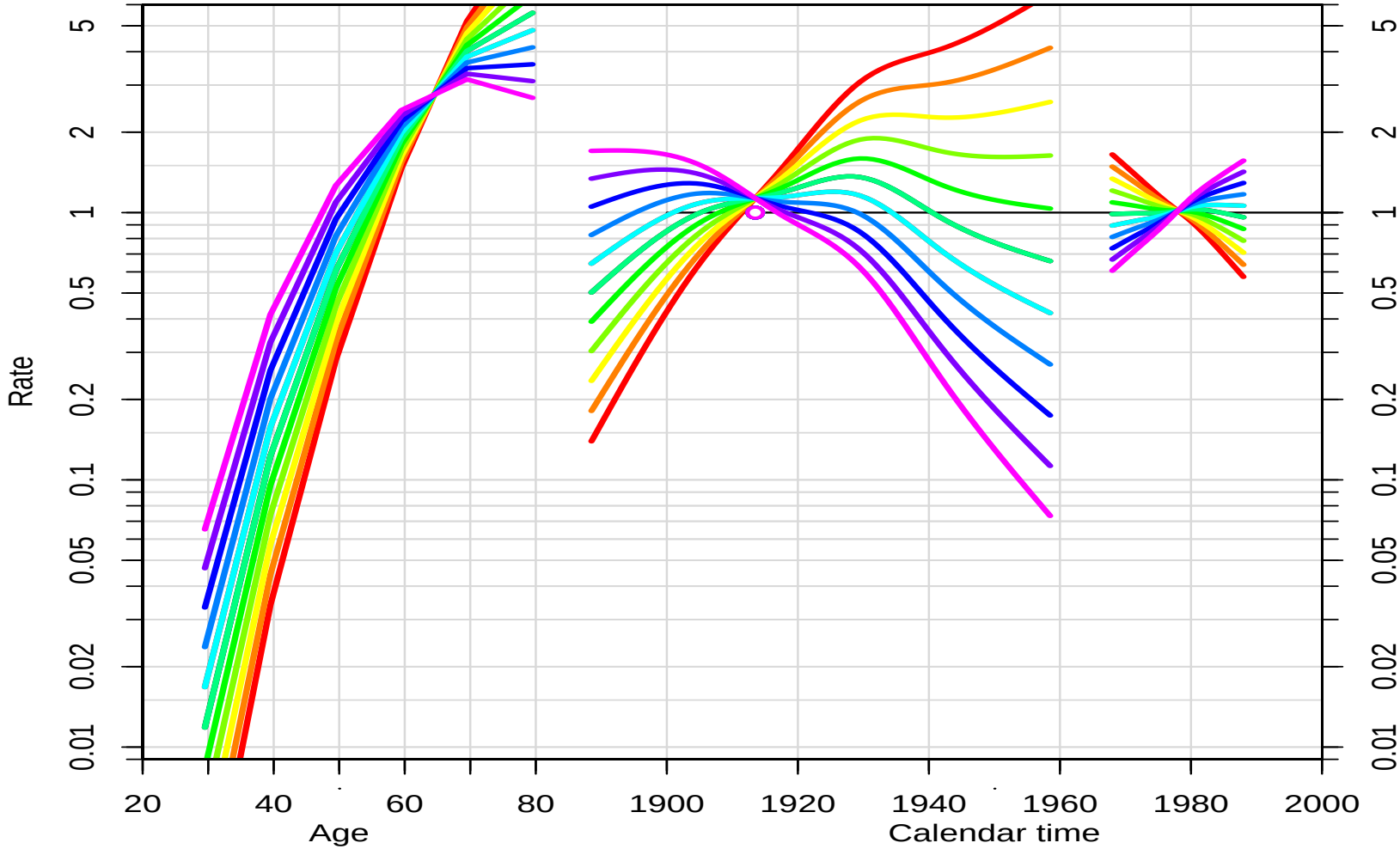


Figure 7: Age-period-cohort estimates from the spline model. Curves with added annual period drifts of -4%, -3%,. . . ,4% are also shown. The rates predicted from curves of like colors are the same.

Smoothing Models

Smoothing/Penalization

- We will first generically talk about Bayesian smoothing models
- In general, when looking at estimates over time, we want to know if the differences we see are “real”, or simply reflecting sampling variability.
- In data sparse situations, when one expects similarity smoothing local patterns (in time, space, or both) can be highly beneficial.
- This can equivalently be thought of penalization, in which large deviations from “neighbors”, suitably defined, are discouraged.
- In this section we will generically think of modeling prevalence.

Motivation for Smoothing: Temporal Case

- Temporal setting (assume period only): Even if the underlying prevalence is the same over time, we will see differences in the empirical estimates.
- Figure 8 demonstrates: I simulated binomial data with $n = 10, 20, 200$ and $p = 0.2$ (shown in blue) in all cases.
- In the top plot in particular, we might conclude large temporal variation, but all we are seeing is sampling variation.
- Figure 9 summarizes estimates from a second simulation in which there is a real temporal pattern – here we would not want to oversmooth and remove the trend.
- Later we will apply temporal smoothing models to these two sets of data.

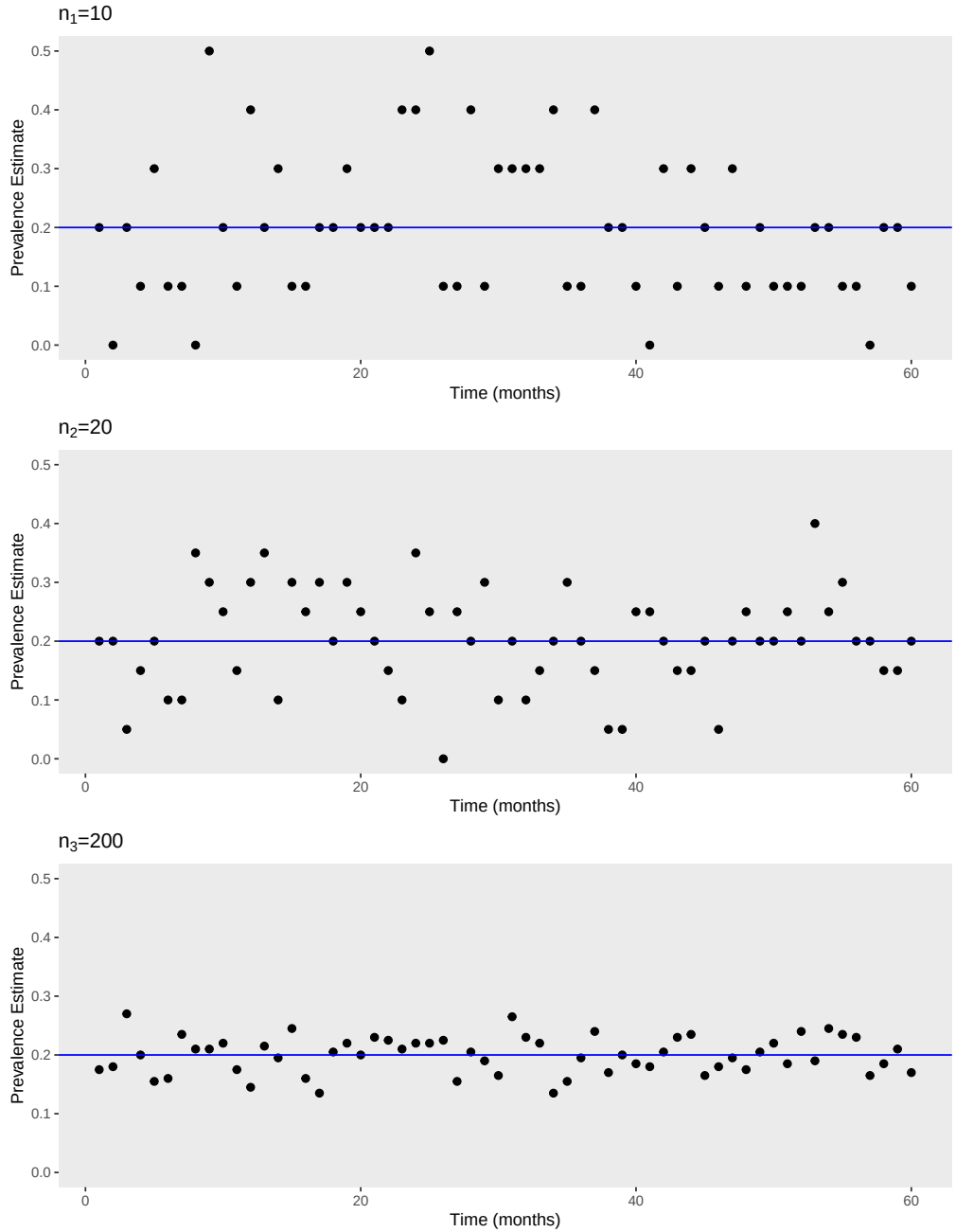


Figure 8: Prevalence estimates over time from simulated data with true prevalence of $p = 0.2$ (blue solid lines).

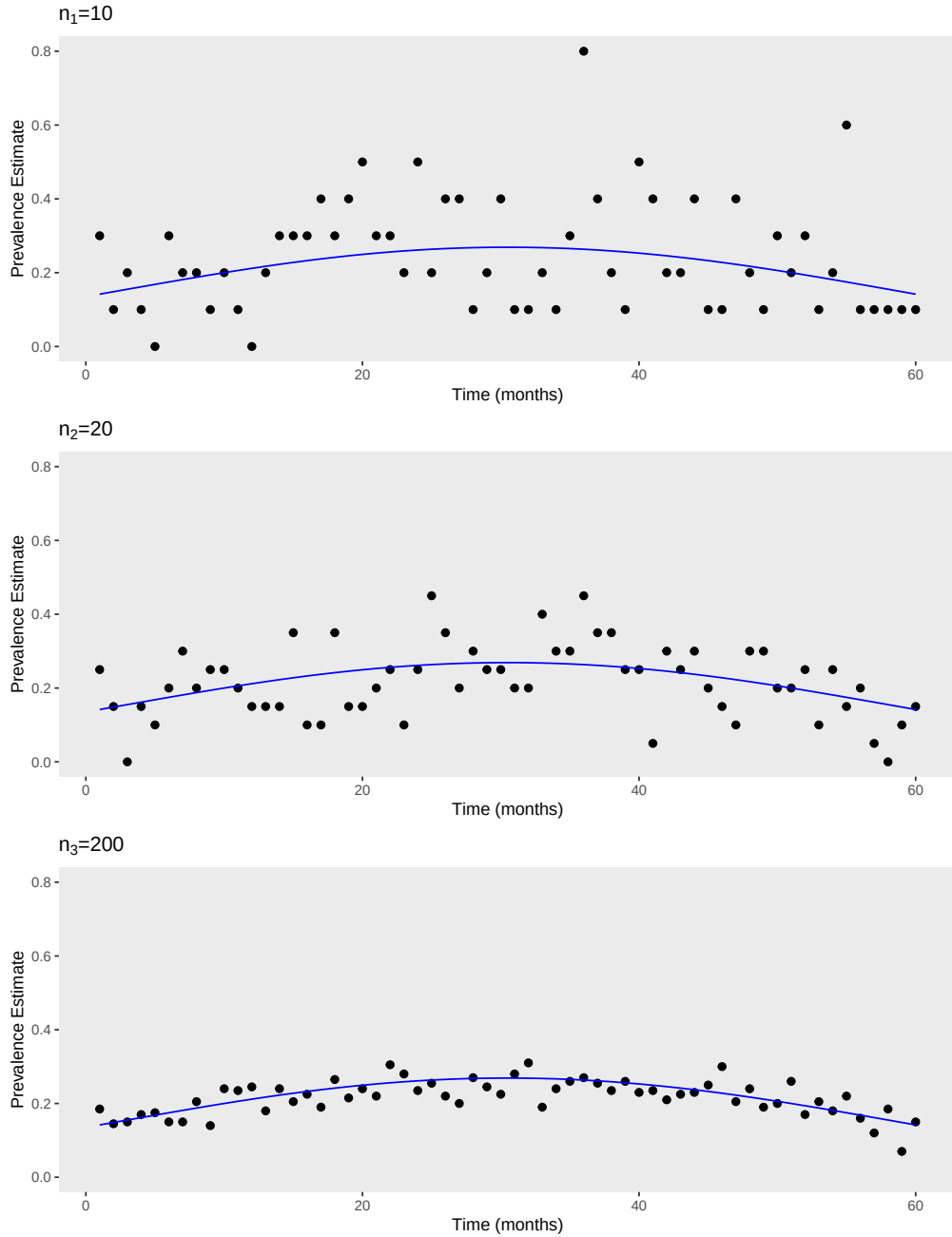


Figure 9: Prevalence estimates over time from simulated data, true prevalence corresponds to curved blue solid line.

Smoothing

When faced with estimation *n* different quantities of the **prevalence** under different conditions, there are three model choices:

- The true underlying prevalence risks are **ALL THE SAME**.
- The true underlying prevalence risks are **DISTINCT** but not linked (like a factor model in APC context).
- The true underlying prevalence risks are **SIMILAR IN SOME SENSE**.

The third option seems plausible when the conditions are related, but how do we model “similarity”?

Smoothing

There are a number of possibilities for **SMOOTHING** models:

- The prevalences are drawn from some **COMMON** probability distribution, but are not ordered in any way. We refer this as the independent and identically distributed, or **IID** model. We could think of this as saying we think the prevalences are likely to be of the same order of magnitude.
- The prevalences display **DEPENDENCE** over time.

These are both examples of **HIERARCHICAL** or **RANDOM EFFECTS MODELS** — a key element is estimating the **SMOOTHING PARAMETER**.

Smoothing over Time

Rationale and overview of models for temporal smoothing:

- We often expect that the true underlying prevalence in an area will exhibit some degree of smoothness over time.
- A linear trend in time is unlikely to be suitable for more than a small number of years, and higher degree polynomials can produce erratic fits.
- Hence, local smoothing is preferred.
- Splines and random walk models have proved successful as local smoothers.
- And to emphasize again, in either approach, the choice of smoothing parameter is crucial.

Random Walk Models

We use random walk models which encourage the mean responses (e.g., prevalences) across time to not deviate too greatly from their neighbors.

The true underlying mean of the prevalence at time t is modeled as a function of its neighbors:

$$\alpha_t \mid \mu_{\text{NE}(t)} \sim \text{N}(m_t, v_t),$$

where

- α_t is the mean prevalence (or some function of it such as the logit) at time t.
- $\alpha_{\text{NE}(t)}$ is the set of neighboring means – with the number of neighbors chosen depending on the model used – typically 2 or 4.
- m_t is the mean of some set of neighbors – for a **first order random walk** or **RW1** it is simply $\frac{1}{2}(\alpha_{t-1} + \alpha_{t+1})$.
- v_t is the variance, and depends on the number of neighbors – for the RW1 model it is $\sigma^2/2$, where σ^2 is a smoothing parameter – small values give large smoothing.

Random Walk Models

- The smoothing parameter σ^2 is estimated from the data, and determines the extent deviations from the mean are penalized.
- The penalty term for the RW1 model is:

$$p(\alpha_t \mid \alpha_{t-1}, \alpha_{t+1}, \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\alpha_t - \tfrac{1}{2}\left(\alpha_{t-1} + \alpha_{t+1}\right)\right]^2\right\}.$$

- Hence:
  - Values of α_t that are close to $\frac{1}{2}(\alpha_{t-1} + \alpha_{t+1})$ are favored (higher density).
  - The relative favorability is governed by σ^2 – if this variance is small, then α_t can't stray too far from its neighbors.
- Predictions from the RW1 are

$$\alpha_{n+s} | \alpha_1, \ldots, \alpha_n, \sigma^2 \sim \mathrm{N}(\alpha_n, \sigma^2 \times s).$$

Figure 10: Illustration of the RW1 model for smoothing at time 3. The mean of the smoother is the average of the two adjacent points (and is highlighted as •), and deviations from this mean are penalized via the normal distribution shown in **red**.

RW1 Model

- Form of the prior density is:

$$\begin{aligned}\pi(\boldsymbol{\alpha}|\sigma^2) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T-1}(\alpha_{t+1}-\alpha_t)^2\right)\\ &= \exp\left(-\frac{1}{2\sigma^2}\sum_{t\sim t'}(\alpha_t-\alpha_{t'})^2\right) = \exp\left(-\frac{1}{2}\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\alpha}\right)\end{aligned}$$

where $t \sim t'$ indicates t is a neighbor of t' and the precision is $\boldsymbol{Q} = \boldsymbol{R}/\sigma^2$ with

$$\boldsymbol{R} = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix}$$

and zeroes everywhere else.

- This sparsity leads to big gains in computational efficiency.

RW1 Model

We might consider the model:

Stage 1: $Y_t|\lambda_t \sim \text{Poisson}(N_t\lambda_t)$, $t = 1, \ldots, n$.

Stage 2: $\log \lambda_t = \alpha_t + \epsilon_t$, with $\epsilon_t \sim \text{N}(0, \sigma^2_\epsilon)$, $t = 1, \ldots, n$.

The RW1 prior is not proper – informally, the collection $\alpha_1, \ldots, \alpha_n$ has a multivariate normal distribution with rank deficiency 1.

More precisely, the precision matrix[1] implied by the conditional distributions is of rank $n - 1$.

[1] inverse of the variance-covariance matrix

RW1 Model

This class of prior is often called an intrinsic model and, the overall level is not identified, but if the Stage 1 data model is identifiable, then the posterior is identifiable.

If there is an intercept in the model, then a constraint is required, and typically a sum-to-zero is specified, $\sum_{t=1}^{n} \alpha_t = 0$.

But the Stage 1 model is not identifiable for APC data when we have all three variables!

RW2 Model

- The second order RW (RW2) model produces smoother trajectories than the RW1, and has more reasonable short term predictions, which is desirable for modeling child prevalence.
- In terms of second differences:

$$(\alpha_t - \alpha_{t-1}) - (\alpha_{t-1} - \alpha_{t-2}) \sim \mathrm{N}(\,0, \sigma^2\,),$$

  showing that deviations from linearity are discouraged.
- Forecasts S steps ahead have a normal distribution with mean:

$$\mathrm{E}[\alpha_{n+S} \mid \alpha_1, \ldots, \alpha_n] = \alpha_t + S(\alpha_t - \alpha_{t-1})$$

  which is a linear function of the values at the last two time points.
- The variance is

$$\mathrm{var}(\alpha_{n+S} \mid \alpha_1, \ldots, \alpha_n) = \frac{\sigma^2}{6} \times S(S+1)(2S+1)$$

  which is cubic in the number of periods S, so blows up very quickly.

RW2 Model

- Form of the prior density is:

$$\begin{aligned}\pi(\boldsymbol{\alpha}|\sigma^2) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{n-2}(\alpha_{t+2}-2\alpha_{t+1}+\alpha_t)^2\right)\\ &= \exp\left(-\frac{1}{2}\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\alpha}\right)\end{aligned}$$

where the precision is $\boldsymbol{Q} = \boldsymbol{R}/\sigma^2$ with

$$\boldsymbol{R} = \begin{bmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & 1 & -4 & 6 & -4 & 1 & \\ & & . & . & . & . & . \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{bmatrix}$$

and zeroes everywhere else.

?) showed that RW2 models as priors, lead to estimators that are smoothing splines.

RW2 Model

Like the RW1 prior, the RW2 prior is not a proper multivariate normal distribution: the precision matrix implied by the full conditionals is of rank $n - 2$.

Again an intrinsic GMRF, and when there is an intercept and slopes in the RW2 model, the impropriety is usually addressed by imposing two constraints.

Specifically, for RW2 models there is a sum-to-zero constraint and a zero slope constraint.

For example, for the age effects, this is equivalent to $\boldsymbol{L}\boldsymbol{\alpha} = \mathbf{0}$ where the a-th column of $\boldsymbol{L}$ is $\{1, a\}$, $a = 1, \ldots, A$.

In the APC context, these constraints give a model which is not over-parameterized but do not yield interpretable intercepts or slopes, since the data cannot inform on these.

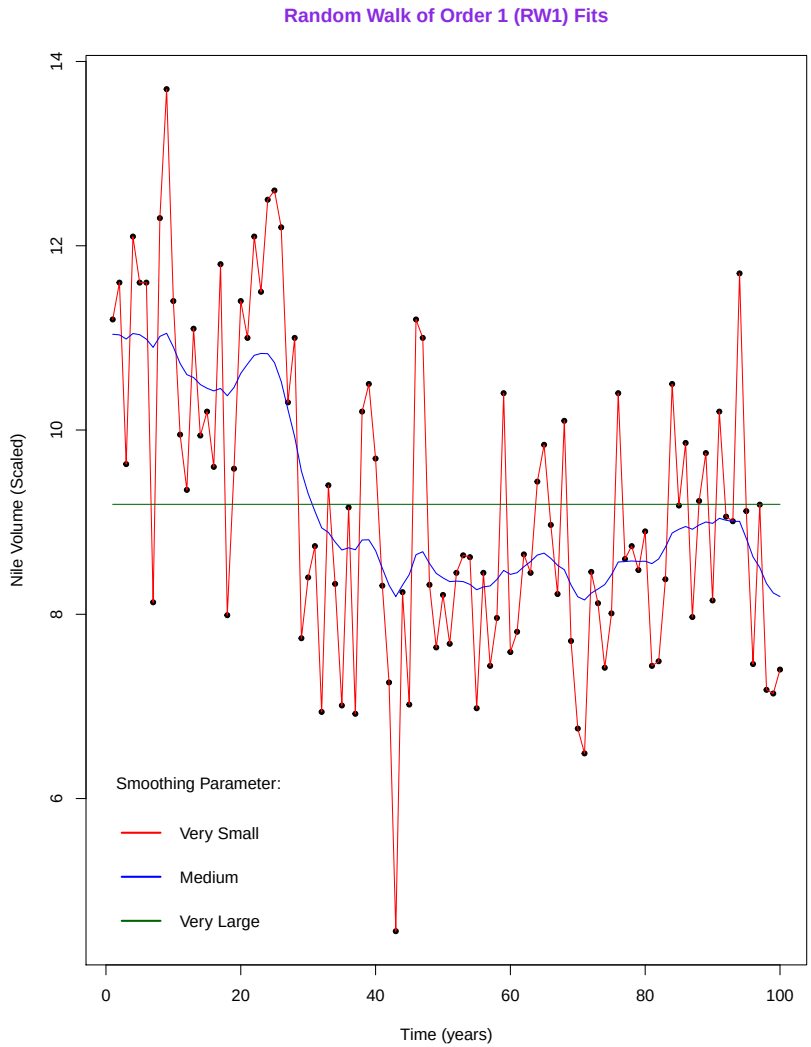


Figure 11: Nile data with RW1 fits under different priors for smoothing parameter σ^{-2}.

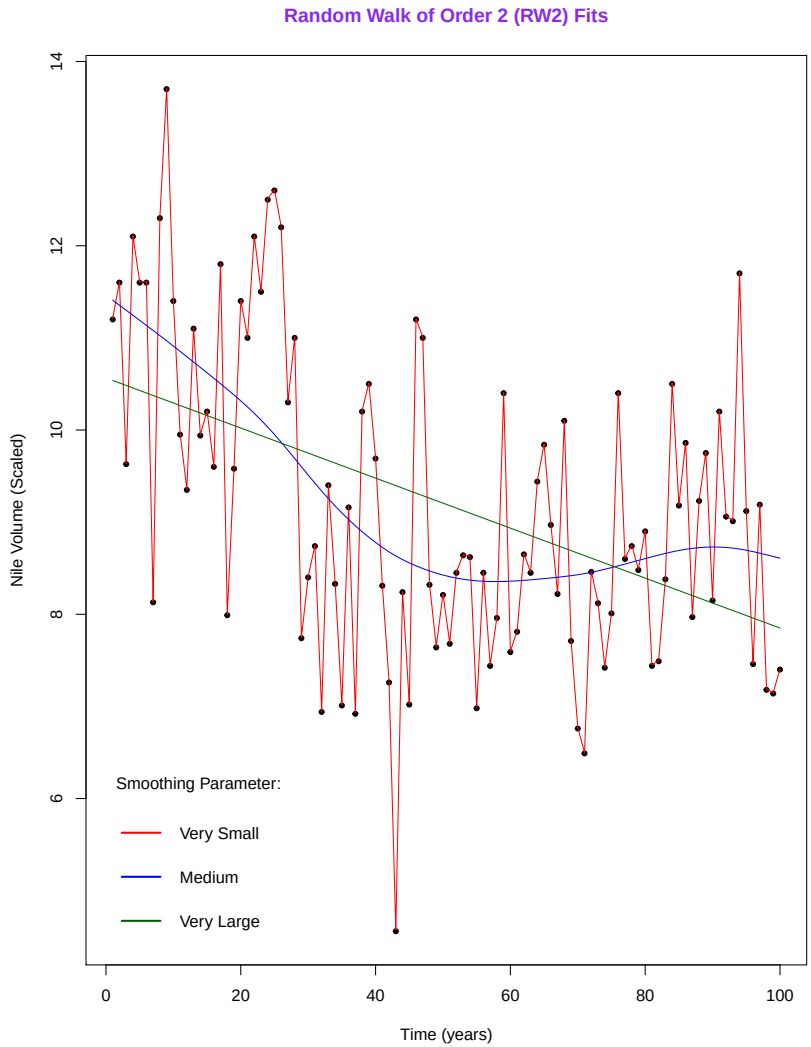


Figure 12: Nile data with RW2 fits under different priors for smoothing parameter σ^{-2}.

Temporal Smoothing Model Summary

We have three models:

IID MODEL:

$$\alpha_t \sim \mathrm{N}(0, \sigma^2),$$

smooth towards zero.

RW1 MODEL:

$$\alpha_t - \alpha_{t-1} \sim \mathrm{N}(0, \sigma^2),$$

smooth towards the previous value.

RW2 MODEL:

$$(\alpha_t - \alpha_{t-1}) - (\alpha_{t-1} - \alpha_{t-2}) \sim \mathrm{N}(0, \sigma^2),$$

smooth towards the previous slope.

RW Fitting to Simulated Data

- We illustrate fitting with the RW2 model, using the simulated data seen earlier.
- The model is:

$$
\begin{aligned}
Y_t | p_t &\sim \text{Binomial}(n_t, p_t), \qquad t = 1, \ldots, n \\
\frac{p_t}{1 - p_t} &= \exp(\delta + \alpha_t) \\
(\alpha_1, \ldots, \alpha_n) &\sim \text{RW2}(\sigma^2) \\
\sigma^2 &\sim \text{Prior on Smoothing Parameter} \\
\delta &\sim \text{Prior on Intercept}
\end{aligned}
$$

- Fit using `R-INLA`.
- On Figures 13 and 14 the fitted values are shown in red – in both the constant prevalence and curved prevalence cases, the reconstruction is reasonable.

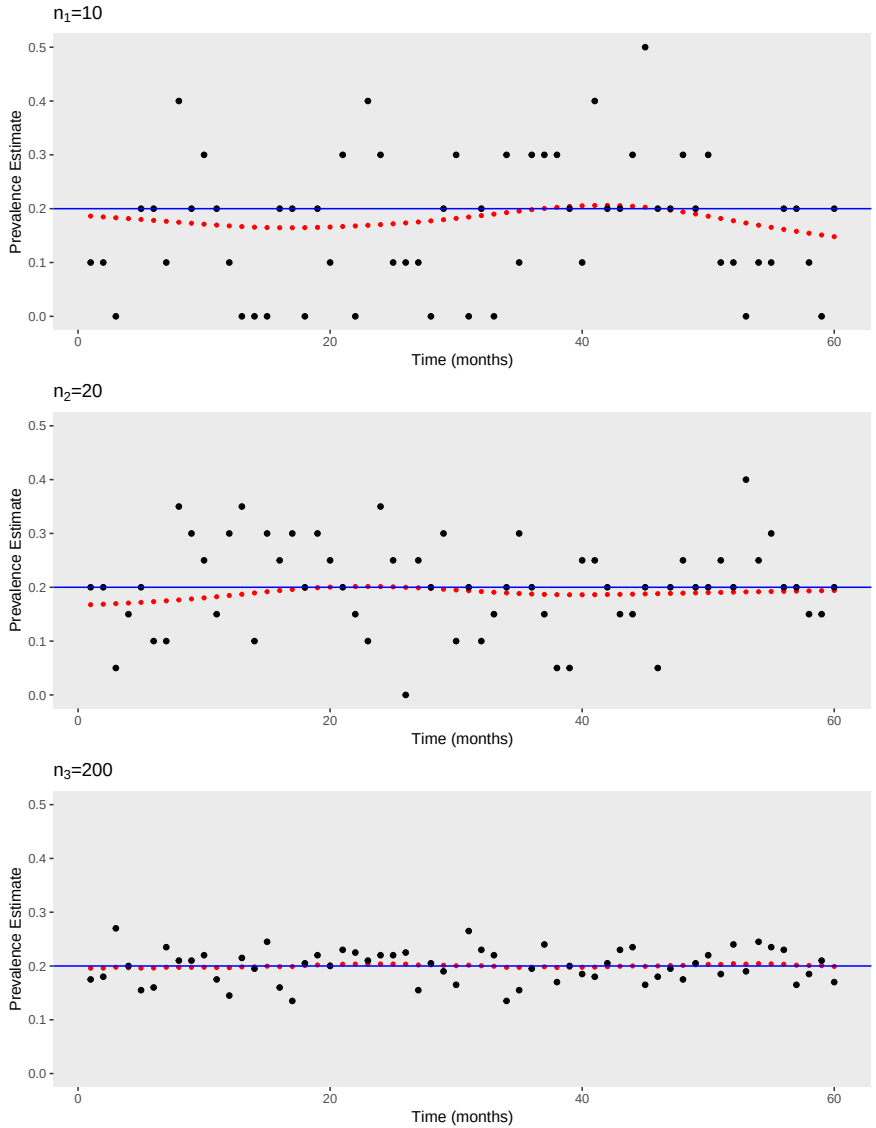


Figure 13: Prevalence estimates over time from simulated data, true prevalence $p = 0.2$ (blue solid lines). Smoothed random walk estimates in **red**.

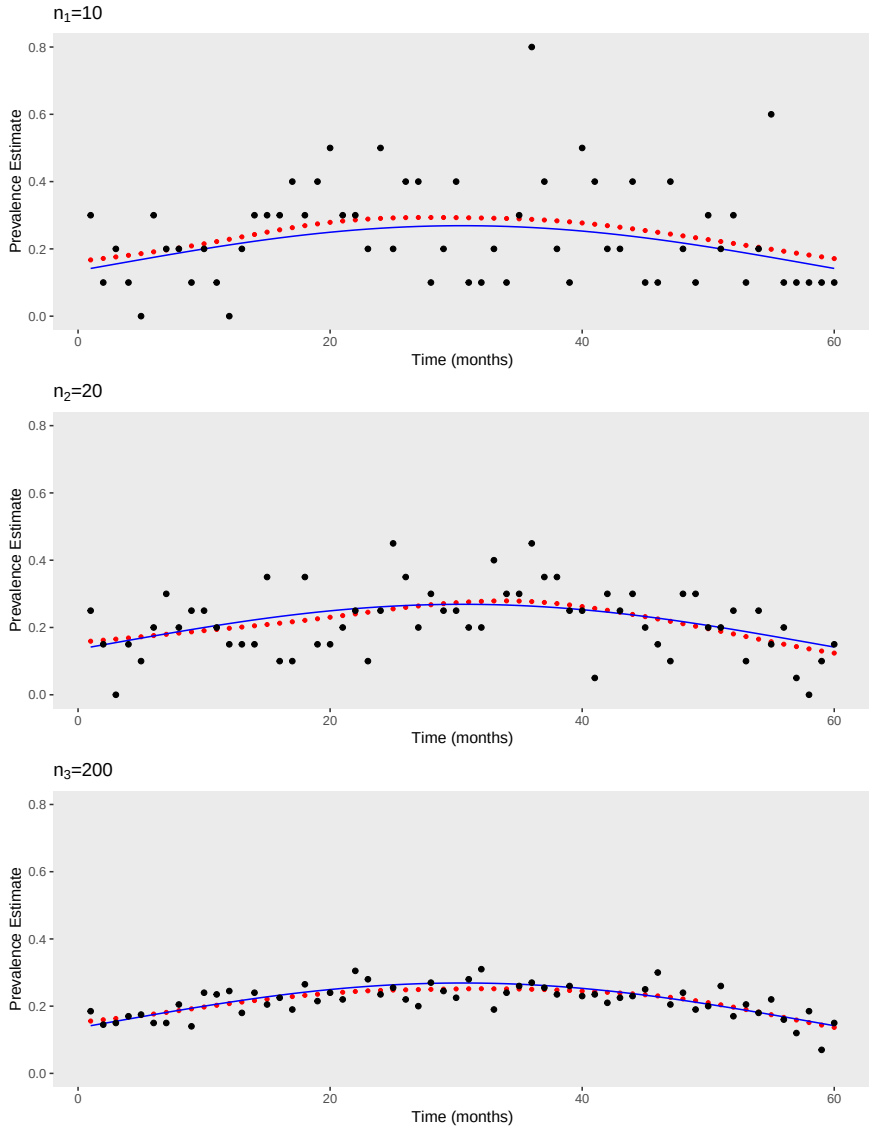


Figure 14: Prevalence estimates over time from simulated data, true prevalence corresponds to curved blue solid line. Smoothed random walk estimates in **red**.

Bayesian Inference

Bayesian Inference

Bayesian inference is a convenient framework within which to implement smoothing models.

- A Data Model (Likelihood) is probabilistically combined with
- A Penalization (Prior) that expresses beliefs about the parameters θ encoding the model.
- Combination occurs via **Bayes Theorem:**

$$\underbrace{p(\theta|y)}_{\textbf{Posterior}} \propto \underbrace{L(\theta)}_{\textbf{Likelihood}} \times \underbrace{\pi(\theta)}_{\textbf{Prior}}.$$

- On the log scale:

$$\underbrace{\log p(\theta|y)}_{\textbf{Updated Beliefs}} = \underbrace{\log L(\theta)}_{\textbf{Data Model}} + \underbrace{\log \pi(\theta)}_{\textbf{Penalization}}.$$

Bayesian Inference

- In a Bayesian analysis the complete set of unknowns (parameters) is summarized via the **multivariate posterior distribution**.
- The marginal distribution for each parameter may be summarized via its mean, standard deviation, or quantiles.
- It is common to report the **posterior median** and a **90% or 95% posterior range** for parameters of interest.
- The range that is reported is known as a credible interval.
- The computations required for Bayesian inference (integrals) is often not trivial and many be carried out using a variety of analytic, numeric and simulation based techniques.
- We use the integrated nested Laplace approximation (INLA), introduced by **?**).
- Book-length treatments:
  - **?**) – space-time models.
  - **?**) – general models.
  - **?**) – advanced space-time models.

Bayes Example

- Imagine the data model is normal with an unknown mean μ:

$$\overline{y} \mid \mu \sim \mathsf{N}(\mu, \sigma^2/n),$$

  where σ^2/n is assumed known ($\sigma/\sqrt{n}$ is the standard error).
- We also imagine the prior is normal:

$$\mu \sim \mathsf{N}(m, v),$$

  so that values of the mean μ that are (relatively) far from m are penalized.
- The log posterior is:

$$\underbrace{\log p(\mu \mid y\ )}_{\text{Updated Beliefs}} = \underbrace{-\frac{n}{2\sigma^2}(\overline{y} - \mu)^2}_{\text{Data Model}} \underbrace{-\frac{1}{2v}(\mu - m)^2}_{\text{Penalization}}.$$

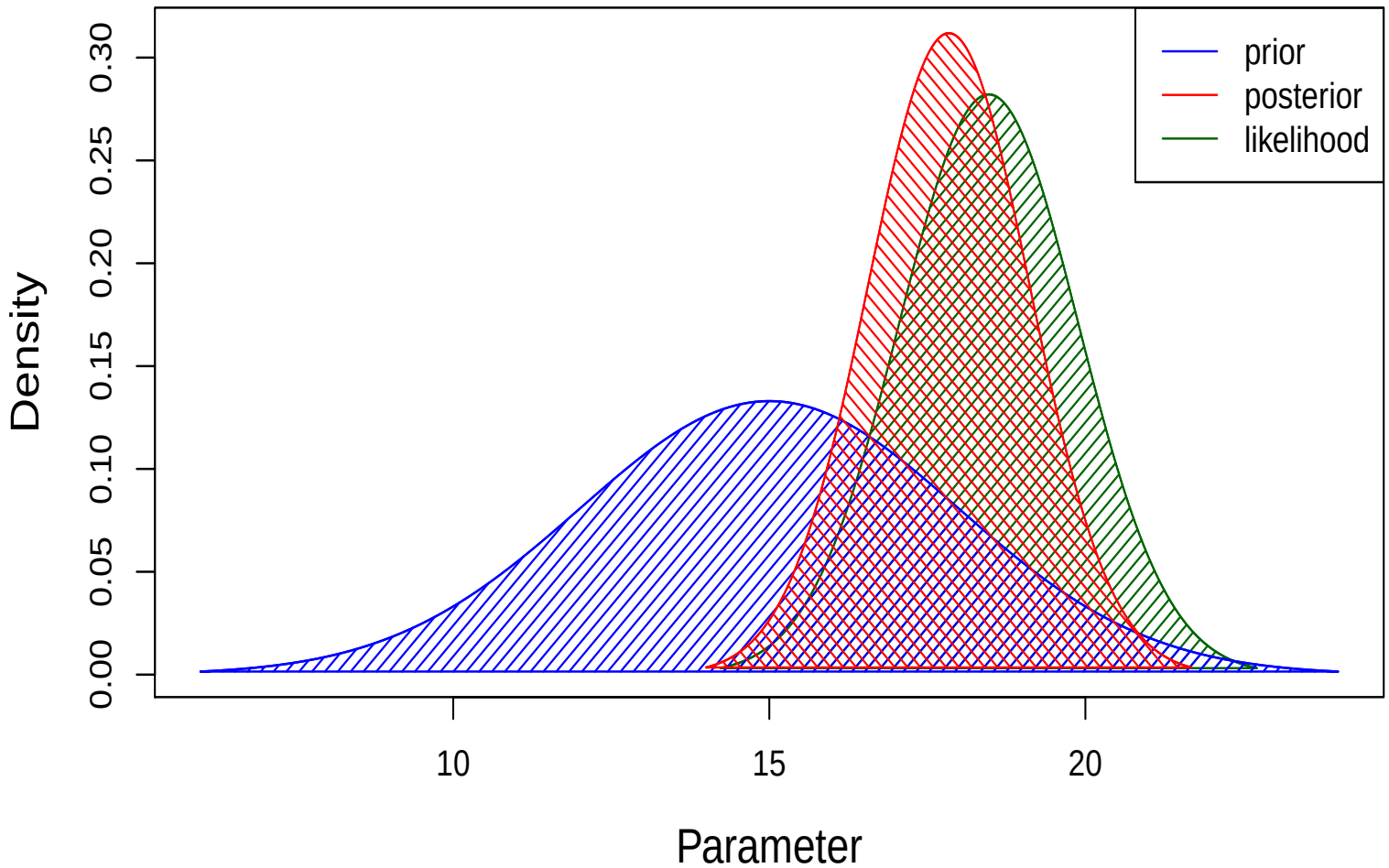


Figure 15: Normal data model with $n = 10$, $\overline{y} = 19.3$ and standard error 1.41. The prior for μ has mean m =15 and $v = 3^2$. The posterior for the parameter μ is a compromise between the two sources of information: the posterior mean is 18.5 and the posterior standard deviation is 1.28.

INLA

The Context

As a running example, consider the mixed effects model:

$$\begin{aligned} y_i|\boldsymbol{\beta}, u_i, \boldsymbol{\theta}_1 &\sim p(y_i|\boldsymbol{\beta}, u_i, \boldsymbol{\theta}_1) \\ \boldsymbol{u}|\boldsymbol{\theta}_2 &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)) \end{aligned}$$

for $i = 1, \ldots, n$, where

- $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_J]^\intercal$ are fixed effects,
- $\boldsymbol{u} = [u_1, \ldots, u_n]$ are random effects following a zero mean multivariate normal distribution, with $\boldsymbol{Q}$ the precision matrix,
- $\boldsymbol{\theta}_1$ are scale parameters in the likelihood, $p(y_i|\boldsymbol{\beta}, u_i, \boldsymbol{\theta}_1)$,
- $\boldsymbol{\theta}_2$ are variance-covariance parameters in the random effects distribution.
- We write $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]^\intercal$ to represent all variance parameters.

Computation, from either a frequentist or Bayesian perspective, is not straightforward for this model.

Bayesian inference for the mixed model

Bayesian analysis adds a hyperprior, with independence often assumed,

$$\pi(\boldsymbol{\beta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \pi(\boldsymbol{\beta}) \times \pi(\boldsymbol{\theta}_1) \times \pi(\boldsymbol{\theta}_2).$$

Penalized complexity (PC) priors are recommended **?**).

The posterior is,

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{y}) = p(\boldsymbol{y} | \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\theta}_1) \times p(\boldsymbol{u} | \boldsymbol{\theta}_2) \times \pi(\boldsymbol{\beta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) / p(\boldsymbol{y}),$$

where

$$p(\boldsymbol{y}) = \int_{\beta} \int_{u} \int_{\theta_1} \int_{\theta_2} p(\boldsymbol{y} | \boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\theta}_1) \times p(\boldsymbol{u} | \boldsymbol{\theta}_2) \times \pi(\boldsymbol{\beta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \, d\boldsymbol{\beta} d\boldsymbol{u} d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2,$$

a typically high-dimensional integral.

Integrals, integrals, integrals all around

- Posterior marginal distributions, e.g.,

$$p(\beta_j|\boldsymbol{y}) = \int_{\beta_{-j}} \int_{u} \int_{\theta_1} \int_{\theta_2} p(\boldsymbol{\beta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{y})\, d\boldsymbol{\beta}_{-j} d\boldsymbol{u} d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2.$$

  To reconstruct a density we need to do this for multiple values of β_j.
- The posterior mean is,

$$\mathrm{E}[\boldsymbol{\beta}_j|\boldsymbol{y}] \quad = \quad \int_{\beta_j} \beta_j p(\beta_j|\boldsymbol{y})\, d\beta_j$$

  with the variance requiring $\mathrm{E}[\beta_j^2|\boldsymbol{y}]$.
- The posterior median $\widetilde{\beta}_j = \mathrm{Median}(\boldsymbol{\beta}_j|\boldsymbol{y})$ is that value that solves

$$\int_{-\infty}^{\widetilde{\beta}_j} p(\beta_j|\boldsymbol{y})\, d\beta_j = 0.5,$$

  with posterior quantiles found, similarly.
- Predictive distributions:

$$p(z|\boldsymbol{y}) = \int_{\beta} \int_{u^\star} \int_{\theta} p(z|\boldsymbol{\beta}, u^\star, \boldsymbol{\theta}) p(\boldsymbol{\beta}, u^\star, \boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\beta} du^\star d\boldsymbol{\theta}.$$

Laplace Approximations

Integrals may be calculated using analytical approximations, numerical integration and Monte Carlo methods – we describe an example of the first of these, Laplace's method.

Let $g(u)$, be a one-dimensional function and

$$\mathcal{I} = \int_{-\infty}^{\infty} \exp[g(u)]du,$$

denote a generic integral of interest and suppose $\widetilde{u}$ is the maximum.

By Taylor's theorem,

$$g(u) = \sum_{k=0}^{\infty} \frac{(u-\widetilde{u})^k}{k!} g^{(k)}(\widetilde{u}),$$

where $g^{(k)}(\widetilde{u})$ represents the k-th derivative of $g(\cdot)$ evaluated at $\widetilde{u}$.

Laplace Approximations

Hence,

$$\begin{aligned}\mathcal{I} &= \int_{-\infty}^{\infty} \exp\left[\sum_{k=0}^{\infty} \frac{(u-\widetilde{u})^k}{k!} g^{(k)}(\widetilde{u})\right] du \\ &= \exp[g(\widetilde{u})] \int_{-\infty}^{\infty} \exp\left[\frac{g^{(2)}(\widetilde{u})}{2}(u-\widetilde{u})^2\right] \exp\left[\sum_{k=3}^{\infty} \frac{(u-\widetilde{u})^k}{k!} g^{(k)}(\widetilde{u})\right] du\end{aligned}$$

Taking the approximation to the second term of the Taylor series and letting

$$v = -1/[g^{(2)}(\widetilde{u})]$$

gives

$$\begin{aligned}\widehat{\mathcal{I}} &= \exp\left[g(\widetilde{u})\right] \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2v}(u-\widetilde{u})^2\right\} du \\ &= \exp\left[g(\widetilde{u})\right] (2\pi)^{1/2} v^{1/2}.\end{aligned}$$

Laplace in a Bayesian context

Laplace approximations have a long history in Bayesian computation, and **?**) is a key reference.

Suppose we wish to evaluate the posterior expectation of a positive function of interest $\phi(u)$, i.e.

$$\begin{aligned} \mathrm{E}[\phi(u)|\boldsymbol{y}] &= \frac{\int \exp[\log \phi(u) + \log p(\boldsymbol{y}|u) + \log \pi(u)]\, du}{\int \exp[\log p(\boldsymbol{y}|u) + \log \pi(u)]\, du} \\ &= \frac{\int \exp[g_1(u)]\, du}{\int \exp[g_2(u)]\, du}. \end{aligned}$$

Application of Laplace's method to numerator and denominator gives

$$\widehat{\mathrm{E}}[\phi(u) \mid \boldsymbol{y}] = \frac{\widetilde{v}_1}{\widetilde{v}_2} \frac{\exp[g_1(\widetilde{u}_1)]}{\exp[g_2(\widetilde{u}_2)]}$$

where $\widetilde{u}_j$ is the maximum of $g_j(\cdot)$ and $\widetilde{v}_j = -1/g_j^{(2)}(\widetilde{u}_j)$, $j = 1, 2$.

Laplace in a Bayesian context

In asymptotic terms, Laplace's method typically has an error of order $O(n^{-1})$.

For the above calculation, however, it may be shown that (**?**),

$$\widehat{\mathrm{E}}[\phi(u) \mid \boldsymbol{y}] = \mathrm{E}[\phi(u) \mid \boldsymbol{y}](1 + O(n^{-2})),$$

since errors in the numerator and denominator cancel.

If ϕ is not positive then a simple solution is to add a large constant to ϕ; Laplace's method may then be applied with the constant subtracted at the end.

See **?**) for more details in a Bayesian context.

Multivariate Laplace

Now consider multivariate $\boldsymbol{u}$ with $\dim(\boldsymbol{u}) = p$ and with required integral

$$\mathcal{I} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp[g(\boldsymbol{u})]\, du_1 \ldots du_p.$$

the above argument may be generalized to give the Laplace approximation

$$\widehat{\mathcal{I}} = \exp\left[g(\widetilde{\boldsymbol{u}})\right] (2\pi)^{p/2} \mid \widetilde{\boldsymbol{v}} \mid^{1/2}, \tag{9}$$

where $\widetilde{\boldsymbol{u}}$ is the maximum of $g(\cdot)$ and $\widetilde{\boldsymbol{v}}$ is the $p \times p$ matrix whose (i, j)-th element is

$$- \left. \frac{\partial^2 g}{\partial u_i \partial u_j} \right|_{\widetilde{\boldsymbol{u}}}.$$

So for implementation, we need to maximize functions, and we need second derivatives – the latter can be a big pain to calculate analytically, so a numerical approach is desirable.

Laplace approximations in practice

In general, the Laplace approximation works well when the integrand, with respect to whatever is being integrated over, is “normal-like” – this is heavily dependent on the parameterization adopted.

In a Bayesian setting, where we want to integrate over all parameters, we must identify parameters that are not normal-like, and either reparameterize, or treat differently.

Variance components in particular, require special attention.

Laplace in a Bayesian context

Tierney and Kadane (1986, Section 4.1) discuss how to approximate the marginal posterior density, and this is explicitly used in the INLA method.

Simplify by assuming a single parameter set $\boldsymbol{u} = [u_1, \ldots, u_p]$ and suppose $\widetilde{\boldsymbol{u}} = [\widetilde{u}_1, \widetilde{\boldsymbol{u}}_2]$ maximizes the posterior, which is proportional to

$$p(\boldsymbol{y}|\boldsymbol{u})\pi(\boldsymbol{u}),$$

so that $\widetilde{\boldsymbol{u}}$ is the posterior mode.

Let $\boldsymbol{u} = [u_1, \boldsymbol{u}_2]$ with $\boldsymbol{u}_2 = [u_2, \ldots, u_p]$ and define $\widetilde{\boldsymbol{v}}$ to be the $p \times p$ matrix corresponding to the inverse of the Hessian of $p(\boldsymbol{y}|\boldsymbol{u})\pi(\boldsymbol{u})$.

For fixed u_1, let $\widetilde{\boldsymbol{u}}_2^\star = \widetilde{\boldsymbol{u}}_2^\star(u_1)$ maximize $p(\boldsymbol{y}|u_1, \boldsymbol{u}_2)\pi(u_1, \boldsymbol{u}_2)$, and let $\widetilde{\boldsymbol{v}}^\star = \widetilde{\boldsymbol{v}}^\star(u_1)$ be the $(p-1) \times (p-1)$ matrix corresponding to the inverse of the Hessian of $p(\boldsymbol{y}|u_1, \boldsymbol{u}_2)\pi(u_1, \boldsymbol{u}_2)$ (i.e., the second derivates with respect to the elements of $\boldsymbol{u}_2$).

Laplace in a Bayesian context

Now apply Laplace's method to the numerator and denominator of,

$$p(u_1|\boldsymbol{y}) = \frac{\int p(\boldsymbol{y}|u_1, \boldsymbol{u}_2)\pi(u_1, \boldsymbol{u}_2)\, d\boldsymbol{u}_2}{\int p(\boldsymbol{y}|\boldsymbol{u})\pi(\boldsymbol{u})\, d\boldsymbol{u}},$$

which is the marginal density at the point u_1.

This gives the approximation,

$$\widehat{p}(u_1|\boldsymbol{y}) = (2\pi)^{-1/2} \left(\frac{|\widetilde{\boldsymbol{v}}^{\star}(u_1)|}{|\widetilde{\boldsymbol{v}}|} \right)^{1/2} \frac{\pi(u_1, \widetilde{\boldsymbol{u}}_2^{\star}) p(\boldsymbol{y}|u_1, \widetilde{\boldsymbol{u}}_2^{\star})}{p(\boldsymbol{y}|\widetilde{\boldsymbol{u}})\pi(\widetilde{\boldsymbol{u}})} \tag{10}$$

It can be shown (**?**) that the error in the approximation is of order $O(n^{-3/2})$ in $n^{-1/2}$ neighborhoods of $\widetilde{u}_1$.

LGMs

The integrated nested Laplace approximation (INLA) for Bayes computation was introduced by **?**).

`INLA`, the `R` package implementation is designed for latent Gaussian models (LGMs):

Stage 1: Likelihood $p(y_i|\eta_i, \theta_1)$ where η_i is a linear predictor with a known link function (cf GLMs, though class is more general), and the vector θ_1 contains variance/scale parameters. The linear predictor is of the form

$$\eta_i = \beta_0 + \sum_{j=1}^{J} \beta_j z_{ij} + \sum_{k=1}^{K} f_i^k,$$

where

- $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_J]^\top$ where β_0 is the intercept and β_j are fixed effects associated with observed covariates z_{ij}, $j = 1, \ldots, J$.
- $\{f_i^k, k = 1, \ldots, K\}$ are random effects – these may correspond to smoothers in time and space, among many other choices.

Stage 2: The latent Gaussian field is on $\boldsymbol{x} = [\eta, \beta, \boldsymbol{f}^1, \ldots, \boldsymbol{f}^K]$ with

$$\boldsymbol{x}|\theta_2 \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}(\theta_2)),$$

where $\boldsymbol{Q}(\theta_2)$ is the precision matrix of the latent Gaussian field.

Let $\boldsymbol{\theta} = [\theta_1, \theta_2]^{\mathsf{T}}$.

Stage 3: Hyperpriors: $\pi(\boldsymbol{\theta})$.

The resulting posterior is,

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \exp\left[\sum_{i=1}^{n} \log p(y_i|\eta_i, \boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \log \pi(\boldsymbol{\theta})\right]$$

An Example of a LGM

Consider the Poisson RW2 model for data indexed by age a and period p:

$$\begin{aligned} Y_{a,p}|\eta_{a,p} &\sim \text{Poisson}(n_{a,p}, \exp(\eta_{a,p})) \\ \eta_{a,p} &= \delta + \alpha_a + \beta_p \\ \alpha_a &\sim \text{RW2}(\sigma^2_\alpha), \\ \beta_a &\sim \text{RW2}(\sigma^2_\beta) \end{aligned}$$

with hyperpriors on δ (normal) and $\sigma^2_\alpha, \sigma^2_\beta$.

In the above LGM notation, we have $\boldsymbol{x} = [\delta, \boldsymbol{\alpha}, \boldsymbol{\beta}]$ and $\boldsymbol{\theta}_2 = [\sigma^2_\alpha, \sigma^2_\beta]$.

INLA

INLA calculates the univariate marginals:

$$\begin{aligned}
\pi(\theta_j|\boldsymbol{y}) &= \int\int \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{x} d\boldsymbol{\theta}_{-j} = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}_{-j} \qquad (11)\\
\pi(x_i|\boldsymbol{y}) &= \int\int \pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{x}_{-i} d\boldsymbol{\theta}\\
&= \int \left[\int \pi(x_i,\boldsymbol{x}_{-i}|\boldsymbol{\theta},\boldsymbol{y}) d\boldsymbol{x}_{-i}\right] \pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}\\
&= \int \pi(x_i|\boldsymbol{\theta},\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta} \qquad (12)
\end{aligned}$$

INLA

The latent field $\boldsymbol{x}$ and the variance components $\boldsymbol{\theta}$ are treated quite differently by INLA, because the latter are less normal-like in general, even after reparameterization.

The nested part of INLA reflects that given values of $\boldsymbol{\theta}$ Laplace approximations are carried out for $\boldsymbol{x}$, and these are averaged over using numerical integration techniques.

We now describe the various approximations used in INLA.

INLA

The marginal posterior is, for any value of $\boldsymbol{x}$,

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})} \\ &\propto \frac{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\end{aligned}$$

The numerator is available, while the denominator is in general not.

The approximation is,

$$\widehat{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y}) \propto \frac{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}^k)p(\boldsymbol{x}|\boldsymbol{\theta}^k)\pi(\boldsymbol{\theta}^k)}{\widehat{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}^k,\boldsymbol{y})} \tag{13}$$

where $\widehat{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}^k,\boldsymbol{y})$ is the Gaussian approximation to the conditional which is obtained by matching the mode and the curvature at the mode, and is equivalent to Laplace approximation to the density, i.e., (10).

From **?**),

$$\begin{aligned}\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}) &\propto \exp\left[\sum_{i=1}^{n}\log p(y_i|\eta_i,\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Q}(\boldsymbol{\theta})\right] \\ &\approx (2\pi)^{-n/2}|\boldsymbol{P}(\boldsymbol{\theta})|^{1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))^{\mathsf{T}}\boldsymbol{P}(\boldsymbol{\theta})(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))\right]\end{aligned}$$

where

- $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the location of the mode, and
- $\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{Q}(\boldsymbol{\theta}) + \text{diag}(\boldsymbol{c}(\boldsymbol{\theta}))$ with $\boldsymbol{c}(\boldsymbol{\theta})$ being the negative derivatives of the log-likelihood with respect to x_i, evaluated at the mode.

The Gaussian approximation is likely to be accurate since, relative to the $\text{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$ prior, the log-likelihood terms only shifts the mean, reduces the variance and may introduce some skewness into the marginals – crucially, it doesn't change the dependency structure.

INLA

The marginal (12), i.e., $\pi(x_i|\boldsymbol{y})$, needs to be calculated for a potentially very long vector $\boldsymbol{x}$.

We could take the marginal from $\widehat{\pi}_G(\boldsymbol{x}|\theta^k, \boldsymbol{y})$ but unfortunately this is not generally very accurate.

As an alternative, rewrite as

$$\begin{aligned} \pi(x_i|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \\ &\propto \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{x}, \boldsymbol{\theta})}{\pi(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \end{aligned}$$

and the denominator can again be estimated estimated using the **?**) density approximation.

?) describe a third approximation, the simplified Laplace which corrects the Gaussian approximation for location and skewness using a Taylor's series expansion about the mode.

INLA

The INLA computing scheme therefore consists of (**?**):

1. Explore the $\boldsymbol{\theta}$ space via the approximation $\widehat{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y})$. Specifically, find the mode of $\widehat{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y})$ and identify a set of points $\{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K\}$ in the areas of high density.
2. For these K points, compute $\widehat{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y})$ using (14).
3. Calculate $\widehat{\pi}(x_i|\boldsymbol{\theta}^k, \boldsymbol{y})$ for $k = 1, \ldots, K$ using one of Gaussian, Laplace, simplified Laplace.
4. Use numerical integration to approximate the marginal,

$$\widehat{\pi}(x_i|\boldsymbol{y}) = \sum_{k=1}^{K} \widehat{\pi}(x_i|\boldsymbol{\theta}^k, \boldsymbol{y}) \times \widehat{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y})\Delta_k, \tag{14}$$

using points and weights $\{\boldsymbol{\theta}^k, \Delta_k, k = 1, \ldots, K\}$.

Exploring the θ space

First, a "good" parameterization is found, we assume that θ satisfies this; also let $\dim(\theta) = m$.

Find the mode, $\theta^\star$, and the Hessian matrix $\boldsymbol{H}$; let $\boldsymbol{H}^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}$ be the eigen decomposition, then form the new standardized variable:

$$\boldsymbol{z} = (\boldsymbol{V}\boldsymbol{\Lambda}^{1/2})^{-1}(\theta - \theta^\star),$$

which adjusts for location, scale, and rotation.

Exploring the θ space

?) describe three methods for exploration:

1. *grid*: This approach builds a grid for the standardized variable $\boldsymbol{z}$. Unfortunately the number of points grows exponentially with m; if we use p points in each dimension, p^m are required in total.
2. *empirical Bayes*: just take the posterior mode only, i.e., a single point.
3. *CCD*: use a classical design, specifically the central composite design (CCD) – integration points are placed on spheres.

Grid versus CCD

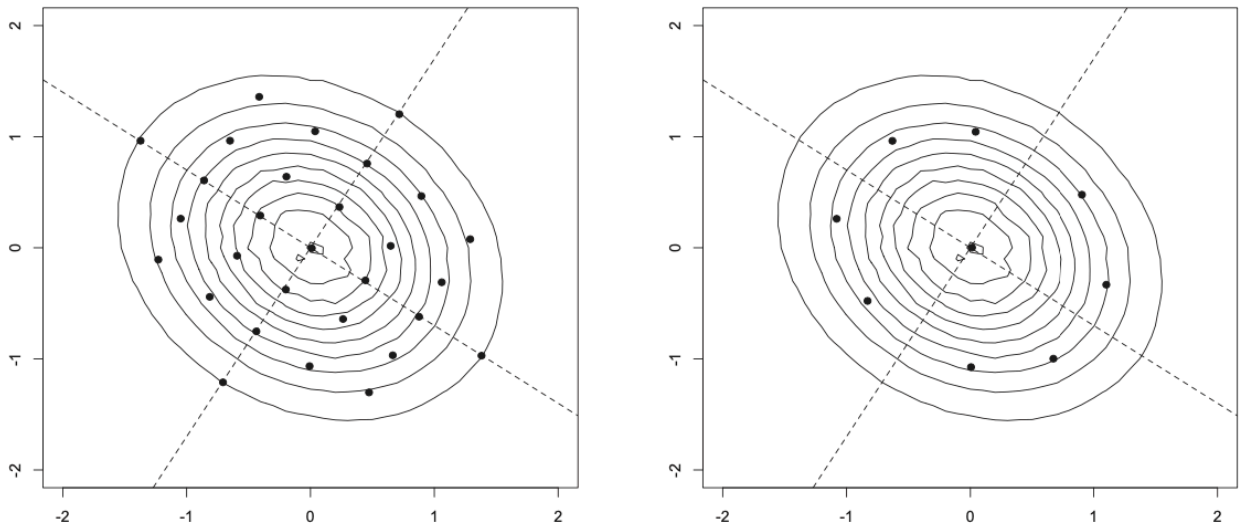


Figure 16: Grid (left) and CCD (right) points for numerical integration, from **?**).

INLA: Posterior sampling

Marginals are the standard output of INLA, but various operations may be carried out using the functions

- `inla.dmarginal` for density values
- `inla.pmarginal` for the CDF
- `inla.qmarginal` for quantiles
- `inla.rmarginal` for random samples
- `inla.hpdmarginal` for HPD regions
- `inla.emarginal` computes the expected values of a function of a parameter
- `inla.tmarginal` calculates the marginal distribution of a transformation of a latent variable or hyperparameter.

INLA: Practical Advice

Some functionals cannot be obtained using these functions, so samples may be drawn, and manipulated:

- `inla.posterior.sample()` draws samples from the approximate posterior distribution of β and θ.
- To make use of this function, use `control.compute = list(config = TRUE)` in the INLA model fit.
- Included in the arguments is `selected` which allows only specific components to be sampled.
- In general, the returned sample contains

```
"hyperpar" "latent" "logdens"
```

Notes on INLA

- A small amount of iid error is added to η_i, to make $\boldsymbol{Q}$ non-singular.
- INLA produces univariate marginals and summaries, by default, but more flexible inference (including multivariate) can be achieved by simulating from an approximation to the posterior.
- For example, for the latent field $\boldsymbol{x}$ we sample from a mixture of multivariate Gaussians, where the weights correspond to the integration weights (for the grid and CCD options).

INLA: Practical Advice

To assess accuracy, one may see how much the results change when different approximation strategies are used.

- Analytic approximation: `inla(...,control.inla=list(strategy="laplace"))`
- Numerical integration strategy: `inla(..., control.inla = list(int.strategy = "grid"), ...)`
- See all the defaults: `inla.set.control.inla.default()`
- For reproducible results, and a better approximation: `inla(..., control.inla = list(strategy = "laplace", int.strategy = "grid", dz=0.1, diff.logdens=20), num.threads=1)` The `diff.logdens` dictates how far we go into the tails when exploring the θ space.
- To make use of multiple cores, INLA uses the OpenMP multiple processing interface, but this produces different results (usually very small) if the same code is rerun – reproducibility is obtained with `num.threads=1`

References

Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. John Wiley and Sons.

Carstensen, B. (2007). Age–period–cohort models for the lexis diagram. *Statistics in Medicine*, **26**, 3018–3045.

Holst, U., Hössjer, O., Björklund, C., Ragnarson, P., and Edner, H. (1996). Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements. *Environmetrics*, **7**, 401–416.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

Martino, S. and Riebler, A. (2019). Integrated nested laplace approximations (INLA). *arXiv preprint arXiv:1907.01248*.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, **4**, 395–421.

Simpson, D., Rue, H., Riebler, A., Martins, T., and Sørbye, S. (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *Statistical Science*, **32**, 1–28.

Speckman, P. L. and Sun, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, **90**, 289–302.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer, New York.

Wang, X., Yue, Y., and Faraway, J. J. (2018). *Bayesian Regression Modeling with INLA*. Chapman and Hall/CRC.