Curve Smoothing via Latent Gaussian Models

Alex Stringer

June 17th 2019

University of Toronto and Centre for Global Health Research, St. Michael's Hospital

Plan of talk

This talk is about curve smoothing, how to fit $y_i = f(x_i) + \epsilon_i$ given **structural assumptions** about unknown f.

Want a general framework that covers commonly seen data types and is easy to extend.

Need uncertainty estimates (we're Statisticians!)

We will use Latent Gaussian Models and Bayesian inference.

I'll go through a detailed example, then briefly mention current and future work.

What is a Latent Gaussian Model?

In order to model **linear** associations between covariates and a response, take a Gaussian distribution and decompose its mean ("linear regression"):

$$y_i \overset{ind}{\sim} \text{Normal}(x_i^T \beta, \sigma^2) \iff y_i = x_i^T \beta + \epsilon_i \tag{1}$$

In order to model **non-linear** associations, it's tempting to mess around with how the mean is decomposed ("non-linear regression").

It's much easier to stick with the above, but mess around with the **residuals**:

$$y|U = X\beta + AU + \epsilon;\ U \sim \text{Normal}\left(0, \Sigma_U\right) \tag{2}$$

The **latent** quantity U is **Gaussian** $\implies$ "Latent Gaussian Model".

Motivating Example: US opioid mortality epidemic

Opioid mortality is increasing *really fast*. Check out PA and OH:

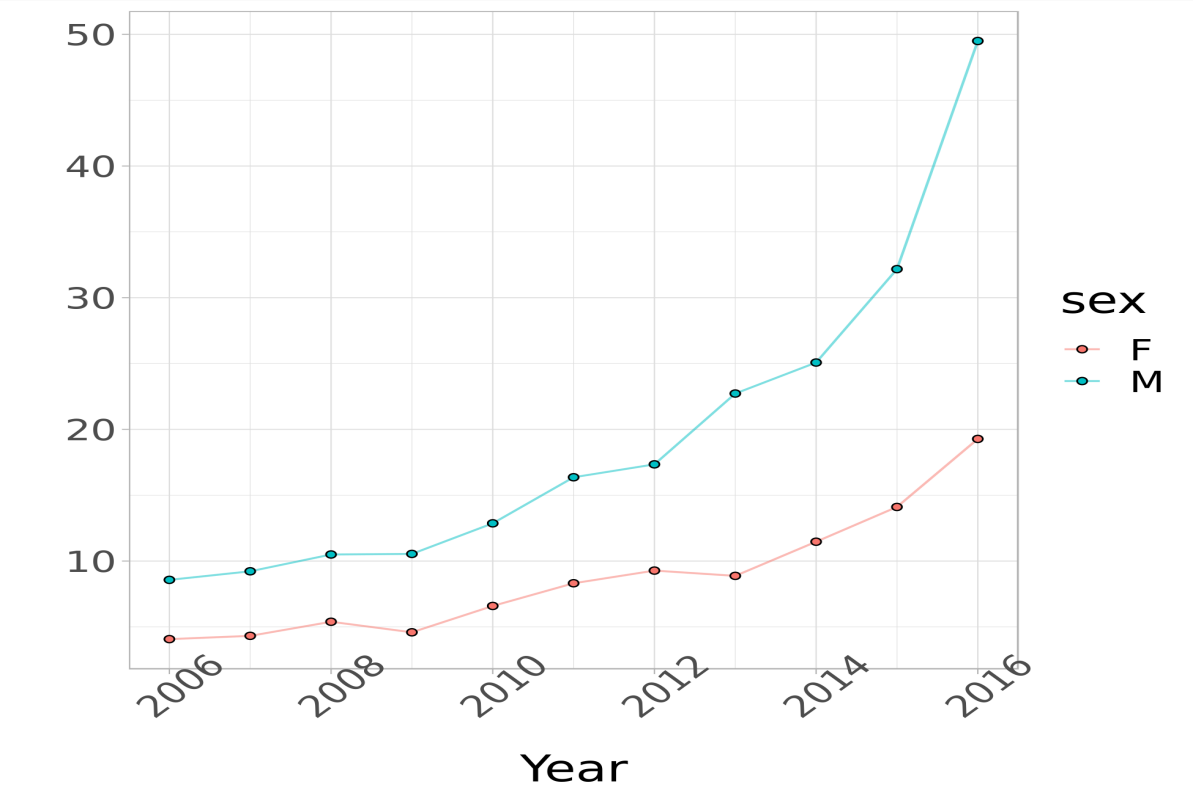
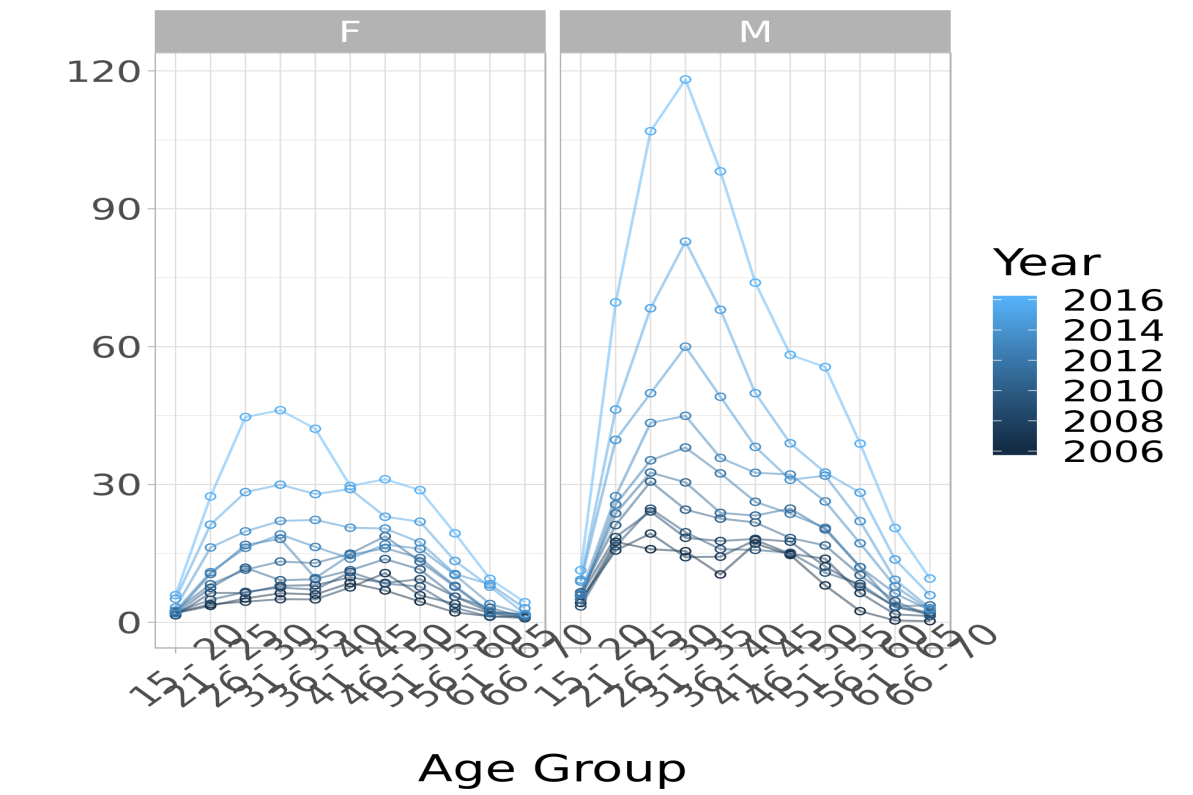


Figure 1: Opioid mortality per 100,000 residents, non-hispanic whites

Motivating Example: US opioid mortality epidemic

The trend is increasing with time, and affects different age groups differently, and this effect appears to *depend* on time:



How do we model these potentially non-linear effects of age, time, and cohort (interaction)? For each gender separately?

Structured Additive Regression Models

Start with a **generalized additive model**:

$$
\begin{aligned}
Y_i(t) &\sim \text{Poisson}(O_i(t)\lambda_i(t)) \\
\log \lambda_i(t) &= x_i(t)^T\beta + \gamma(z_{1i}) + \alpha(z_2(t)) + \gamma \circ \alpha(z_{1i} \times z_2(t))
\end{aligned}
$$

Already pretty complicated, but it captures the terms we want.

How to infer values of *smooth* functions $\gamma(\cdot), \alpha(\cdot)$ at observed ages z_{1i} and times $z_2(t)$?

This slide is really important

The key is to use a **random effects model**, but with added **structure** on the residuals.

For a single covariate $z_i(t)$ that we want to model smoothly, collect all *evaluations* of γ into one big *latent vector*:

$$U = (U_1(1), \ldots, U_n(T)) = (\gamma(z_1(1)), \ldots, \gamma(z_n(T)))$$

Put a **distribution** on U that reflects known **structure** of the function γ. Common choice is saying "γ is first-order smooth", and putting a **random walk** "distribution" on U:

$$\Delta U_i(t) \equiv \begin{pmatrix} U_2 - U_1 \\ \vdots \\ U_n - U_{n-1} \end{pmatrix} \overset{iid}{\sim} \text{Normal}(0, \sigma_U^2)$$

(Can also do with second differences $\implies$ smoother)

Heirarchical Models

With this discretization and suitable priors, we get a **heirarchical model**. With W a giant vector collecting all latent quantities:

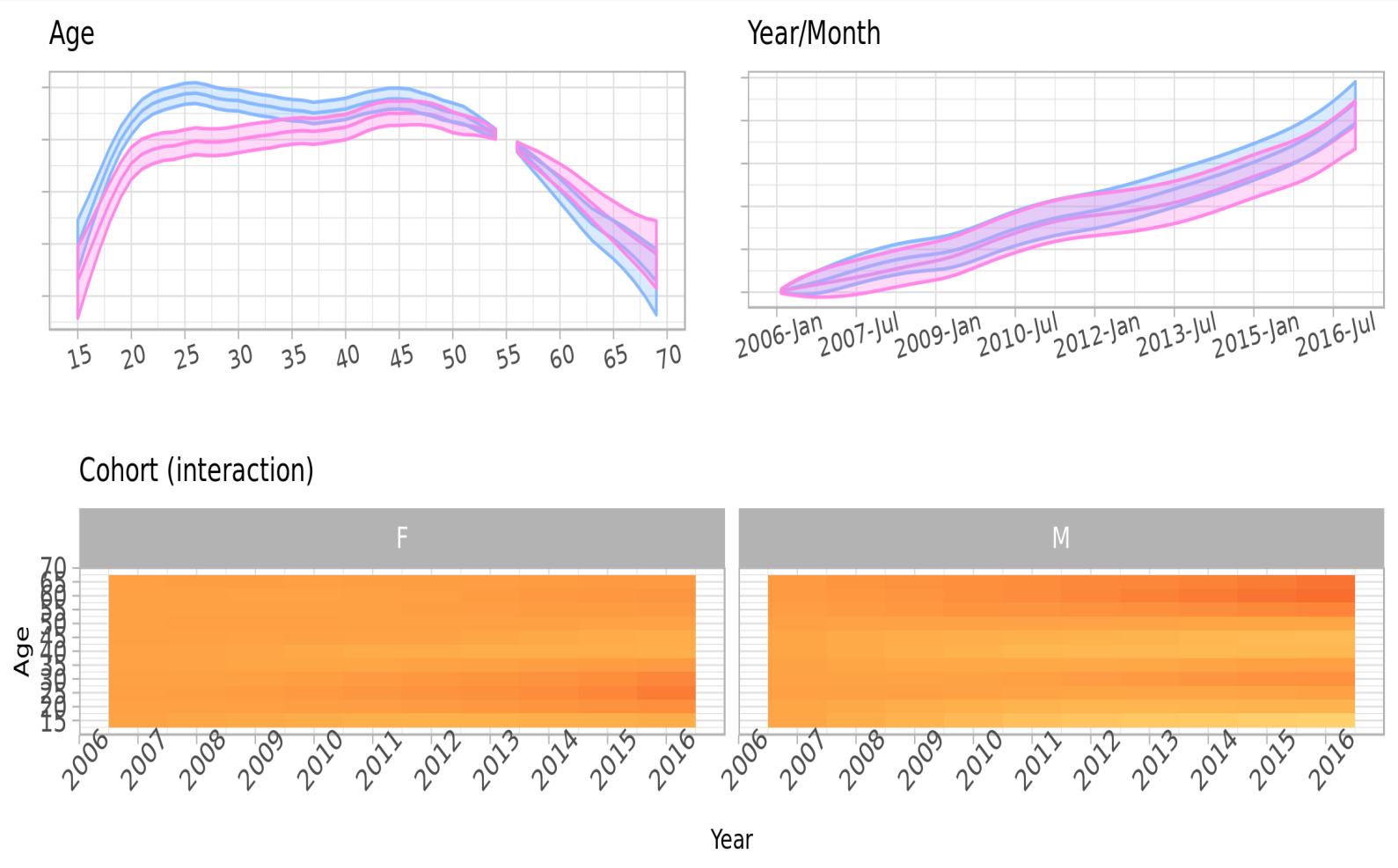
$$
\begin{aligned}
Y_j(it)|W,\theta &\sim \text{Poisson}(\lambda_j(it)O_j(it)) \\
\log \lambda_i(t) &= \beta_0 1(\text{Fem}) + T(t) \\
&\quad + A_i + AT_i(t) \\
W|\theta &\sim \text{N}\left(0, Q^{-1}(\theta)\right)
\end{aligned}
\qquad
\begin{aligned}
\Delta^2 A_i &\sim \text{N}\left(0, \sigma_a^2\right) \\
\Delta^2 T_j(t) &\sim \text{N}\left(0, \sigma_t^2\right) \\
\Delta^2 AT_i(t) &\sim \text{N}\left(0, \sigma_{at}^2\right)
\end{aligned}
$$

Sure, it's really messy looking, but it's **modular**.

It's easier to construct than it is to read ("write only"?)

Results

After accouting for differing baseline opioid mortality rates for each gender, we plot the posterior excess risk over expected for each age group and time point:



Inference Methodology

How did we get there? **Integrated Nested Laplace Approximations** (INLA).

Key observation: the latent quantities in our big model exhibit **conditional independence**, prescribed directly by the structure we put on them.

Jointly Gaussian random vectors have their conditional independence graphs **encoded in their precision matrix**.

The precision matrix of a Gaussian RV with conditional independence is highly **sparse**.

Sparse matrices enable fast, scalable, iterative computation.

Sampling-free! No MCMC!!! **Scales very feasibly with dimension of parameter space.**

Other types of structures

This was a really complicated way to draw 3 lines through some count data.

But now that we're here, we can, with very little effort, impose **any kind of structure we want**.

Longitudinally correlated, spatially correlated...

Even types of structure imposed by differential equations that we know the function must satisfy (more later)!

And we can **change the likelihood to something more complicated** and get all this for harder models than the Poisson.

The distributions for the observable and latent quantities are **separated**.

Future work

How to decide the structure to impose, for a given problem?

Often know/assume $\gamma(\cdot)$ satisfies some (stochastic) **differential equation**. Can lead to certain structures.

The 2^{nd} order random walk was obtained from

$$\frac{d^2}{dt^2} f(t) = \sigma_U \frac{d}{dt} B(t) \qquad \text{“2nd order smooth”} \tag{3}$$

More structure leads to stronger inferences with less data.

Principled regularization.

Future work: how to choose more interesting structure while **retaining computational efficiency**?

Ending with a more interesting example...

I'll end with an example. Suppose you want to model the size of some animal as a function of time $\gamma(t)$ based on a random sample. You want to infer the growth *rate* $\frac{d}{dt}\gamma(t)$.

You know it's monotone (animals don't usually shrink), $\frac{d}{dt}\gamma(t) \geq 0$.

A random walk model will give you a result that *looks* monotone. But when you try to use it to estimate the growth rate...

...you're hosed, because you used a model whose first-differences are **negative half the time**.

Structure matters!

Questions?