Structured Additive Regression Models: What, Why and How

...and an analysis of age-specific opioid mortality rates over time for two US states

Alex Stringer [1,2] Patrick Brown [1,2] Monica Alexander [1,3] Jamie Stafford [1]

[1]Department of Statistical Sciences, University of Toronto [2]Centre for Global Health Research, St. Michaels Hospital [3] Department of Sociology, University of Toronto

Many statistical models share common structure

Many statistical models used in common practice can be built from a few key ingredients:

- **Observable quantities** $y_1, \ldots, y_n$, the data, and an assumed **joint distribution** $\pi(y|\cdot)$,
- **Latent quantities** $w_1, \ldots, w_m$ with an assumed **joint distribution** $\pi(w|\cdot)$,
- **Unknown parameters** $\beta_1, \ldots, \beta_p, \sigma_1, \ldots, \sigma_k$, e.g. means and variances,
- A **linear predictor** $\eta_1, \ldots, \eta_n$ which combines the latent quantities and unknown parameters,
- A **link function** $g(\cdot)$, $g(\mathbb{E}(Y_i|W_i, \beta, \sigma)) = \eta_i$.

All of the following models can be built from these ingredients (fun game: try and identify them!):

- **Linear Regression**: $Y_i \sim \text{Normal}\left(x_i^T\beta, \sigma^2\right)$,
- **Generalized Linear Models**: $Y_i \sim$ Exponential Family; $g(\mathbb{E}(Y_i)) = x_i^T\beta$,
- **Mixed Effects Models**: $Y_{ij}|U_i \sim \text{Normal}\left(x_{ij}^T\beta + U_i, \sigma^2\right)$, $U_i \sim \text{Normal}\left(0, \sigma_U^2\right)$,
- **Generalized Additive Models**: $Y_i \sim$ Exponential Family; $g(\mathbb{E}(Y_i)) = \sum_{j=1}^p f_j(x_{ij})$,

...and many others. The goal of any given analysis is to summarize **unknown** and **unobserved** quantities using **point estimates**, and to **quantify uncertainty** in those estimates through the use of a **probability distribution**, or otherwise.

A guiding principle is that in doing this, **the analyst should be constrained only by their data and their imagination**. We need a single **comprehensive, scalable paradigm** for fitting these models and quantifying the uncertainty in the results.

Structured Additive Regression Models

A **structured additive regression model** is a **hierarchical model** of the form

$$\begin{aligned} Y_i|W, \theta_2 &\sim \pi(y|w, \theta_2) \\ W|\theta_1 &\sim \pi(w|\theta_1) \\ \eta_i &= \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \sum_{k=1}^K U_k + \epsilon_i \\ \epsilon_i &\sim \text{Normal}(0, \tau^{-1}) \\ W &= (\beta_0, \ldots, \beta_p, U_1, \ldots, U_K, \eta_1, \ldots, \eta_n) \\ \theta &= (\theta_1, \theta_2) \end{aligned} \tag{1}$$

We have

- A **likelihood** for the **observable quantities**,
- A **distribution** for the **latent quantities**,
- A **linear predictor**, with **measurement error**

The term "structured" comes from the fact that by controlling the distribution placed on the latent quantities, you can recover **random effects**, **random walk smoothing**, **longitudinal models**, **spatial models**, and many more.

The dimension of the **latent field** $(W_1, \ldots, W_N)$ is huge, typically a multiple of the sample size (or worse). How could these models possibly be fit to large, modern datasets?

Efficient Gaussian Approximations

In order to get computational efficiency, it is sufficient (actually, *necessary*; ask me why!) to put a **joint Gaussian distribution** on the latent field,

$$W \sim \text{Normal}\left(0, Q^{-1}\right) \tag{2}$$

The **conditional independence structure** of W will imply a **sparse precision matrix** Q, which is necessary and sufficient for **computational efficiency**. To see this, write the linear predictor in vector form,

$$\eta = X\beta + AU + Z \tag{3}$$

where $Z \sim \text{Normal}(0, \tau^{-1}I)$. The random effect design matrix A is **very sparse**. You can work out the precision matrix for the whole latent field:

$$Q = \begin{pmatrix} \tau I & -\tau A & -\tau X \\ -\tau A^T & \Sigma_U^{-1} + \tau A^T A & \tau A^T X \\ -\tau X^T & \tau X^T A & \Sigma_\beta^{-1} + \tau X^T X \end{pmatrix} \tag{4}$$

- Looks ugly, but it's **very sparse**: most of the size comes from the A and A^TA terms, which are huge and sparse,
- Operations involving **solving linear systems** involving Q can be done quickly,
- Storing and manipulating Q has **low memory cost**.

All of this can be applied to form a **Gaussian approximation to the posterior of the latent field**,

$$\pi_G(W|y, \theta) \sim \text{Normal}\left(\hat{W}(\theta), \left(Q(\theta) + \hat{C}(\theta)\right)^{-1}\right) \tag{5}$$

- $\hat{W}$ is the **mode** of $\log\pi(W|y, \theta) = \text{const} + \log\pi(y|W, \theta) + \log\pi(W|\theta)$,
- $\hat{C}(\theta)$ is the negative hessian of the log-likelihood evaluated at this mode,
- Requires **high-dimensional optimization**, but a single step of a Newton-based procedure only requires **solving a sparse system**, so it's **actually pretty easy**

We use the highly robust `IPOPT` software, through `R`, for optimization.

"But you don't know θ"! "But Gaussian approximations are inaccurate!"! I hear you cry. This is where INLA comes in.

Integrated Nested Laplace Approximations

Put a **prior** on $\theta \sim \pi(\theta)$, and use this to compute an approximation to the posterior:

$$\begin{aligned} \log\pi(\theta|y) &= \log\frac{\int \pi(y|W, \theta)\pi(W|\theta)\pi(\theta)dW}{\int \pi(y|W, \theta)\pi(W|\theta)\pi(\theta)dWd\theta} \\ &\approx \text{const} + \log\pi(y|W, \theta) + \log\pi(\hat{W}|\theta) + \log\pi(\theta) - \frac{1}{2}\log|Q(\theta) + \hat{C}(\theta)| \end{aligned} \tag{6}$$

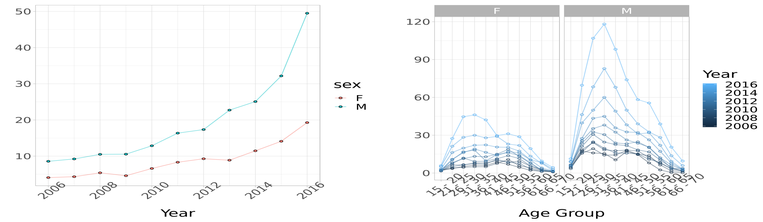
- This is the approximation to marginal posteriors developed by Tierney and Kadane (1986).
- Build an interpolent to this, a set $(\theta_1, \ldots, \theta_k)$ which caputures most of the mass, and
- **Numerically integrate**, to get

$$\pi(W|y) = \int \pi(W, \theta|y)d\theta = \int \pi(W|y, \theta)\pi(\theta|y)d\theta \approx \sum_{k=1}^K \pi_G(W|y, \theta_k)\pi(\theta_k|y)\Delta_k \tag{7}$$

where the Δ_k are weights. In practice INLA replaces the Gaussian approximation with one that also corrects for skewness; the resulting **mixture of skew-normals** is fast to compute, and empirically accurate, though there is no real theory that quantifies this (yet).

Application: modelling opioid mortality rates

The opioid epidemic in North America is a modern public health emergency. Mortality rates have been sharply increasing over the last 10 years, reaching around 10 deaths per 100,000 people per year in the United States in 2016. In certain age and gender groups, the rates are **drastically higher**. Figure 1 shows the yearly trend; opioid-specific mortality rates for 25 - 30 year old non-hispanic whites in particular are of the same order of magnitude as **deaths due to heart disease** in 50+ adult men.



(a) Opioid mortality rates sharply increasing. (b) The spike is highly pronounced for certain age groups.

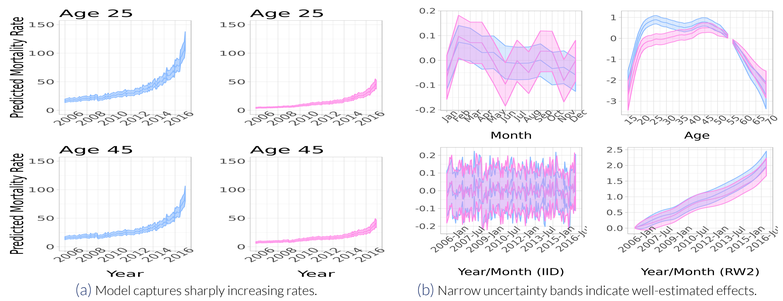
Figure 1. Opioid mortality rates over time, OH and PA

We want to quantify the **effects of age and year**, their interaction, and a possible seasonal effect within year, all using **smooth functions**, split over gender, with uncertainty estimates for everything. Maybe the data has enough structure in it to support such a complicated model, and maybe not—the point is, **this is our only concern**, not whether we can fit such a model using available tools. We model:

$$\begin{aligned} Y_j(it)|W, \theta &\sim \text{Poisson}(\lambda_j(it)O_j(it)) \\ \log\lambda_j(it) &= \beta_0 1(j = F) + M_j(m_t) + \\ &\quad R_j(t) + T_j(t) + A_j(i) + AT_j(it) \\ W|\theta &\sim \text{N}\left(0, Q^{-1}(\theta)\right) \\ M_j(m) &\sim \text{N}\left(0, \tau_m^{-1}\right) \end{aligned} \qquad \begin{aligned} R_j(t) &\sim \text{N}\left(0, \tau_r^{-1}\right) \\ \Delta^2 A_j(i) &\sim \text{N}\left(0, \tau_a^{-1}\right) \\ \Delta^2 T_j(t) &\sim \text{N}\left(0, \tau_t^{-1}\right) \\ \Delta^2 AT_j(it) &\sim \text{N}\left(0, \tau_{at}^{-1}\right) \end{aligned}$$

1. $Y_j(it)$: `death count` at time t for `age` group i, `gender` j,
2. $\lambda_j(it), O_j(it)$: mortality rate, population offset,
3. $M_j(m_t), R_j(t)$: **unstructured** `month` and `year-month` effect,
4. $A_j(i), T_j(t), AT_j(it)$: **structured** `age` and `year-month` effects.

Ooof! That's a complicated model. Fitting using `R-INLA` with around 67,000 observations took about 3.5 hours on an AWS VM with 8 cores and 32 gigs of RAM. The resulting point estimates with 95% posterior credible intervals and mortality rate predictions are shown in Figure 2. The model picks up on the structured age and time trends that we see in Figure 1, and has trouble quantifying a seasonal month effect.



(a) Model captures sharply increasing rates. (b) Narrow uncertainty bands indicate well-estimated effects.

Figure 2. Model predictions and posterior effects.