methods

April 1, 2020

1 Overview

CurveFit is an extendable nonlinear mixed effects model or fitting curves. The main application in this development is COVID-19 forecasting, so that the curves we consider are variants of logistic models. However the interface allows any user-specified parametrized family.

Parametrized curves have several key features that make them useful for forecating:

- We can capture key signals from noisy data.
- Parameters are interpretable, and can be modeled using covariates in a transparent way.
- Parametric forms allow for more stable inversion approaches, for current and future work.
- Parametric functions impose rigid assumptions that make forecasting more stable.

1.1 COVID-19 functional forms

We considered two functional forms so far when modeling the COVID-19 epidemic.

- **Generalized Logistic:**
$$f(t;\alpha,\beta,p) = \frac{p}{1+\exp(-\alpha(t-\beta))}$$
- **Generalized Gaussian Error Function**
$$f(t;\alpha,\beta,p) = \frac{p}{2}\left(\Psi(\alpha(t-\beta)\right) = \frac{p}{2}\left(1+\frac{2}{\sqrt{\pi}}\int_0^{\alpha(t-\beta)}\exp\left(-\tau^2\right)d\tau\right)$$

Each form has comparable fundamental parameters:

- **Level** p**:** Controls the ultimate level.
- **Slope** α: Controls speed of infection.
- **Inflection** β: Time at which the rate of change is maximal.

We can fit these parameters to data, but this by itself does not account for covariates, and cannot connect different locations together. The next section therefore specifies statistical models that do this.

1.2 Statistical Model

Statistical assumptions link covariates across locations. Key aspects are the following:

- Parameters may be influenced by covariates, e.g. those that reflect social distancing
- Parameters may be modeled in a different space, e.g. p, α are non-negative

- Parameters and covariate multipliers may be location-specific, with assumptions placed on their variation.

CurveFit specification is tailored to these three requirements. Every parameter in any functional form can be specified through a link function, covariates, fixed, and random effects. The final estimation problem is a nonlinear mixed effects model, with user-specified priors on fixed and random effects.

For example, consider the ERF functional form with covariates α, β, p. Assume we are fitting data in log-cumulative-death-rate space. Input data are:

- S_j: social distancing covariate value at location j
- y_j^t: cumulative death rate in location j at time t

We specify the statistical model as follows:

- Measurement model:

$$\log(y_j^t) = \frac{p_j}{2}\left(1 + \frac{2}{\sqrt{\pi}}\int_0^{\alpha_j(t-\beta_j)} \exp\left(-\tau^2\right) d\tau\right) + \epsilon_{t,j}$$
$$\epsilon_{t,j} \sim N(0, V_t)$$

- β-model specification:

$$\beta_j = \beta + \gamma_j S_j + \epsilon_j^\beta$$
$$\gamma_j \sim N(\overline{\gamma}, V_\gamma)$$
$$\epsilon_j^\beta \sim N(0, V_\beta)$$

- α-model specification:

$$\alpha_j = \exp(\alpha + u_j^\alpha)$$
$$u_{\alpha,j} \sim N(0, V_\alpha)$$

- p-model specification:

$$p_j = \exp(p + u_j^p)$$
$$u_{p,j} \sim N(0, V_p)$$

In this example, the user specifies

- prior mean $\overline{\gamma}$
- variance parameters $V_t, V_\gamma, V_\beta, V_\alpha, V_p$.

CurveFit estimates:

- fixed effects α, β, p
- random effects $\{\gamma_j, u_j^\alpha, u_j^\beta, u_j^p\}$

Exponential link functions are used to model non-negative parameters α, p.

1.3 Constraints

Simple bound constraints on parameters can be used to make the model more robust. For any fixed or random effect, the user can enter simple bound constraints of the form

$$L \leq \theta \leq U.$$

The parameters returned by CurveFit are guaranteed to satisfy these simple bounds.

1.4 Optimization Procedure

The optimization problem we obtain from specifying functional forms, priors, and constraints on all parameters is a bound-constrained nonlinear least squares problem. We explain the solver, derivative computation, and initialization procedure below.

1.4.1 Solver

We solve the problem using L-BFGS-B. The L-BFGS-B algorithm uses gradients to build a Hessian approximation, and efficiently uses that approximation and projected gradient method onto the bound constraints to identify parameter spaces over which solutions can be efficiently found, see the paper. It is a standard and robust algorithm that's well suited to the task.

1.4.2 Derivatives

We do not explicitly compute derivatives of the nonlinear least squares objective induced from the problem specification. Instead, we use the complex step method to do this. The complex step method is a simple example of Automatic Differentiation, that is, it can provide machine precision derivatives at the cost of a function evaluation. This is very useful given the flexibility on functional forms.

1.5 Uncertainty

Currently CurveFit uses model-based uncertainty, with out-of-sample approaches under development.

1.5.1 Model-Based Uncertainty

We partition model-based uncertainty into estimates coming from fixed and random components. Fixed effects capture the variation of the mean effects, and random effects uncertainty captures the variation across locations.

- Fixed Effects

For any estimator obtained by solving a nonlinear least squares problem, we can use the Fisher information matrix to get an asymptotic approximation to the uncertainty. Let

$$\hat{\theta} = \arg\min_{\theta} := \frac{1}{2}\theta^T W^{-1}\theta + \frac{1}{2\sigma^2}\|\Sigma^{-1/2}(y - f(\theta; X))\|^2$$

where W is any prior variance and Σ is the variance of observations. Then our approximation for the variance matrix of the estimate is given by

$$V(\hat{\theta}) = \mathcal{I}(\theta)^{-1} = \left(J_{\hat{\theta}}^T \Sigma^{-1} J_{\hat{\theta}} + W^{-1}\right)^{-1}$$

where

$$J_{\hat{\theta}} := \nabla_\theta f(\theta; X)$$

is the Jacobian matrix evaluated at $\theta = \hat{\theta}$. The Jacobian is also computed using the complex step method.

- Random effects

To obtain the variance of the random effects, we derive an empirical variance matrix across locations. Given a set of zero mean random effect estimates $\{v_j\}$, with each v_j a vector of k of random effect types, we get an empirical matrix $V_0 \in \mathbb{R}^{k \times k}$ by

$$V_0 = \frac{1}{n} \sum_{j=1}^{N} v_j v_j^T$$

To obtain posterior uncertainty for each specific location, we use the empirical V_0 as a prior, and any data at the location as the measurement model, and re-fit the location:

$$\hat{\theta}_i = \arg\min_\theta := \frac{1}{2} \theta_i^T V_0^{-1} \theta_i + \frac{1}{2\sigma^2} \|\Sigma_i^{-1/2} (y_i - f_i(\theta_i; X_i))\|^2$$

Within each location, this is analogous to the fixed effects analysis. The location-specific uncertainty is then estimated from the same Fisher information analysis:

$$V_i(\hat{\theta}) = (J_i^T \Sigma_i^{-1} J_i + V_0^{-1})^{-1}.$$