Smooth Functional Tempering for Nonlinear Differential Equation Models

Functional Data Methods for Bayesian Parameter Estimation in DE Models

Dave Campbell and Russell Steele
The goal is to estimate $\theta$

We observe $\mathbf{x}(t)$ but often there is no analytic solution to our model.

If the initial state $\mathbf{x}(0)$ is known then we can numerically produce a solution $\mathbf{S}(\mathbf{x}(0), \theta, t) = \mathbf{x}(t)$
Outline

1. Neurophysiology Example
2. Standard Bayesian Tools
3. Smooth Functional Tempering
FitzHugh-Nagumo system

\[ DV = \frac{dV}{dt} = \gamma (V - \frac{V^3}{3} + R) \]
\[ DR = \frac{dR}{dt} = - (\beta R + \alpha - V) / \gamma \]

Numerical Solution to the ODE using:

\[ \theta = [\alpha, \beta, \gamma] = [0.2, 0.2, 3] \text{ and } [V_0, R_0] = [-1, 1] \]
FitzHugh-Nagumo system

\[ DV = \gamma \left( V - \frac{V^3}{3} + R \right), \quad DR = -\left( \beta R + \alpha - V \right) / \gamma \]

The behaviour modeled changes with \( \alpha, \beta, \gamma, V_0, \) and \( R_0 \)

- \([\alpha, \beta, \gamma] = [.2, .2, .3]\)
- \([\alpha, \beta, \gamma] = [-.5, .5, 5]\)
- \([\alpha, \beta, \gamma] = [0, .8, 1]\)
- \([\alpha, \beta, \gamma] = [1, -1, 2]\)
FitzHugh-Nagumo system

\[
DV = \gamma (V - V^3/3 + R), \quad DR = - (\beta R + \alpha - V) / \gamma
\]

401 evenly spaced points with noise \( N(0, .5^2) \) and \( N(0, .4^2) \).
\( \theta = [\alpha, \beta, \gamma] = [0.2, 0.2, 3] \) and \( [V_0, R_0] = [-1, 1] \).
FitzHugh-Nagumo Challenges

- Model behaviour changes drastically with parameter values.
- There is no closed form solution for the likelihood.
- The goal is to estimate $\theta$ but we need $x_0$ to produce a numerical solution.
Outline

1. Neurophysiology Example
2. Standard Bayesian Tools
3. Smooth Functional Tempering
The Model Set Up

For numerical solution \( S(\theta, V_0, R_0, t) \) to equations:

\[
DV = \gamma \left( V - V^3/3 + R \right), \quad DR = -(\beta R + \alpha - V)/\gamma,
\]

use a likelihood of the form:

\[
y(t) \mid \theta, V_0, R_0, \Sigma \sim N \{ S(\theta, V_0, R_0, t), \Sigma \}.
\]

- Place priors on parameters \( P(\theta, V_0, R_0, \Sigma) \) with the goal of making inference on \( P(\theta, V_0, R_0, \Sigma \mid y\{t\}) \).
- Lack of analytical solution implies there is no closed form for the likelihood.

Topological challenges

- Peaks correspond to (partial) data fits.
- Valleys imply that the fit deteriorates before it can improve.
Parallel Tempering$^2$:

Use the sequence of $M$ approximations to the posterior density:

\[
P_1(\theta \mid y\{t\}) = P(y\{t\} \mid \theta)^{T_1} P(\theta)
\]
\[
\vdots
\]
\[
P_M(\theta \mid y\{t\}) = P(y\{t\} \mid \theta)^{T_M} P(\theta)
\]

Where
\[
0 \leq T_1 < \ldots < T_M = 1
\]

- Run all $M$ parallel MCMC chains.
- Allow parameters to swap between chains.
- Only draws from $P_M$ are of interest.

Parallel Tempering:

Use the sequence of $M$ approximations to the target posterior density:

\[
P_1(\theta \mid y\{t\}) = P(y\{t\} \mid \theta)^{T_1} P(\theta)
\]

\[
\vdots
\]

\[
P_M(\theta \mid y\{t\}) = P(y\{t\} \mid \theta)^{T_M} P(\theta)
\]

Where

\[
0 \leq T_1 < \ldots < T_M = 1
\]

- Run all $M$ parallel MCMC chains.
- Allow parameters to swap between chains.
- Only draws from $P_M$ are of interest.
Parallel Tempering

Advantages:
- ‘Flatter’ chains search the posterior space.
- ‘Better’ parameter values are easily passed onto less ‘flat’ chains.
- Enables steps across low probability regions.
But:

- Flatter chains allow parameters to step into trouble
- If the prior is bad, then tempering is bad
Outline

1. Neurophysiology Example
2. Standard Bayesian Tools
3. Smooth Functional Tempering
Smooth Functional Tempering

Combine parallel tempering with insights from functional data analysis (FDA)

- Run $M$ parallel MCMC chains.
- Each chain uses an approximation of the posterior $P(\theta \mid y\{t\})$.
- Use a basis expansion (collocation) $x(t) = c'\phi(t)$ to smooth the data.
The idea:

- Approximate the numerical solution with a data smooth using coefficients $c$

$$s(\theta, t) \approx x(t) = c' \phi(t)$$

- Use a model based smoothing penalty to ensure fidelity to the DE model
Smooth Functional Tempering

The idea:

- Approximate the numerical solution with a data smooth using coefficients $c$

$$s(\theta, t) \approx x(t) = c^t \phi(t)$$

- Use a model based smoothing penalty to ensure fidelity to the DE model

Now define a tempering strategy based on a sequence of smoothing parameters
Build a sequence of $M$ models with $\lambda_1 < \ldots < \lambda_M \leq \infty$. 

$$y(t) \mid x(t), \sigma^2 \sim N(x(t), \sigma^2)$$

$$\pi(\theta) \propto \exp \left\{ -\lambda_m \int_t (Dx(v) - f(x(v), \theta))^2 dv \right\} p_1(\theta)$$
Build a sequence of $M$ models with $\lambda_1 < \ldots < \lambda_M \leq \infty$.

$$y(t) \mid x(t), \sigma^2 \sim N(x(t), \sigma^2)$$

$$\pi(\theta) \propto \exp \left\{ -\lambda_m \int_t \left( D x(v) - f(x(v), \theta) \right)^2 dv \right\} \rho_1(\theta)$$

This induces a density on \(x(t)\) without requiring us to sample \(c\).

The induced density on \(x(t)\) decreases as \(x(t)\) strays from The DE solution.

The rate of decrease depends on \(\lambda_m\).
Build a sequence of $M$ models with $\lambda_1 < \ldots < \lambda_M \leq \infty$.

$$y(t) \mid x(t), \sigma^2 \sim N(x(t), \sigma^2)$$

$$\pi(\theta) \propto \exp \left\{ -\lambda_m \int_t (Dx(v) - f(x(v), \theta))^2 \, dv \right\} \, p_1(\theta)$$

Using big $\lambda_M$ makes $x(t)$ arbitrarily close to the DE solution.

But:

- We avoid numerically solving the DE.
- And we remove dependence on $x_0$. 
Posterior Cross Section

left (Parallel Tempering), right (Smooth Functional Tempering)
When \( x(0) \) is of interest.

Sometimes we want inference on \( \theta \) and \( x_0 \).

- In that case use the sequence \( \lambda_1 < \ldots < \lambda_M = \infty \).

\[
    y(t) \mid x(x_0, t), \sigma^2 \sim N\left(x(x_0, t), \sigma^2\right)
\]

\[
    \pi(\theta, x_0) \propto \exp\left\{ -\lambda_m \int_t \left[ D_x(x_0, v) - f(x(x_0, v), \theta) \right]^2 dv \right\} p_1(\theta)p_2(x_0)
\]

- Include \( x_0 \) in the mode
When $\mathbf{x}(0)$ is of interest.

Sometimes we want inference on $\theta$ and $\mathbf{x}_0$.

- In that case use the sequence $\lambda_1 < \ldots < \lambda_M = \infty$.

\[
y(t) \mid \mathbf{x}(\mathbf{x}_0, t), \sigma^2 \sim N\left(\mathbf{x}(\mathbf{x}_0, t), \sigma^2\right)
\]

\[
\pi(\theta, \mathbf{x}_0) \propto \exp \left\{ -\lambda_m \int_t^t \left[ D\mathbf{x}(\mathbf{x}_0, v) - f(\mathbf{x}(\mathbf{x}_0, v), \theta) \right]^2 dv \right\} p_1(\theta) p_2(\mathbf{x}_0)
\]

- Include $\mathbf{x}_0$ in the mode
- as $\lambda \to \infty$ using a b-spline basis,
- $\mathbf{x}(\mathbf{x}_0, t) \mid \theta \to s(\mathbf{x}_0, \theta, t)$ using a Runga-Kutta numerical solver
When \( x(0) \) is of interest.

Sometimes we want inference on \( \theta \) and \( x_0 \).

- In that case use the sequence \( \lambda_1 < \ldots < \lambda_M = \infty \).

\[
y(t) \mid x(x_0, t), \sigma^2 \sim N\left(x(x_0, t), \sigma^2\right)
\]

\[
\pi(\theta, x_0) \propto \exp \left\{ -\lambda_m \int_t \left[ Dx(x_0, v) - f(x(x_0, v), \theta) \right]^2 dv \right\} p_1(\theta)p_2(x_0)
\]

- Include \( x_0 \) in the mode
- as \( \lambda \to \infty \) using a b-spline basis,
- \( x(x_0, t) \mid \theta \to s(x_0, \theta, t) \) using a Runga-Kutta numerical solver
- \( M^{th} \) model is equivalent to:

\[
y(t) \mid x_0, \theta, \sigma^2 \sim N\left(s(x_0, \theta, t), \sigma^2\right)
\]

\[
\pi(\theta) \sim p_1(\theta)
\]
left (Parallel Tempering), mid (Smooth Functional Tempering with $x_0$) right (Smooth Functional Tempering without $x_0$)
Bimodal FitzHugh-Nagumo density

\[ DV = |\gamma| \left( V - \frac{V^3}{3} + R \right), \quad DR = -\left( \beta R + \alpha - V \right) / |\gamma| \]

Assume that all parameters except \( \gamma \) are known and fixed and \( P(\gamma) = \text{Uniform}(-15, 15) \)

Tempering is required to sample from both modes.
Posterior Densities

**SFT1**, **PT SFT2 (no $X_0$).**
Samples from the $m = 1^{st}$ (the flattest) of the parallel chains using largest $\lambda_1$ that enables $\gamma = \pm 3$ modes to be sampled

left (Parallel Tempering), mid (Smooth Functional Tempering (SFT1) with $x_0$) right (Smooth Functional Tempering (SFT2) without $x_0$)
Quality of the $\lambda_M$ chain approximation

Using 50,000 posterior iterations and the metric:

$$D(P_{num}, P_{samp}) = \int \left[ P_{numeric}(\gamma \mid y) - P_{sampled}(\gamma \mid y) \right]^2 d\gamma$$
Quality of the $\lambda_M$ chain approximation

Using 50,000 posterior iterations and the metric:

\[ D(P_{\text{num}}, P_{\text{samp}}) = \int \left[ P_{\text{numeric}}(\gamma | y) - P_{\text{sampled}}(\gamma | y) \right]^2 d\gamma \]

- \[ D(P_{\text{num}}, P_{\text{parallel tempering}}) = .0356 \]
- with $x_0$; \[ D(P_{\text{num}}, P_{SFT}) = .0251 \]
Quality of the $\lambda_M$ chain approximation

Using 50,000 posterior iterations and the metric:

$$D(P_{num}, P_{samp}) = \int \left[ P_{numeric}(\gamma \mid y) - P_{sampled}(\gamma \mid y) \right]^2 d\gamma$$

- $D(P_{num}, P_{parallel \ tempering}) = .0356$
- with $x_0$; $D(P_{num}, P_{SFT}) = .0251$
- without $x_0$; $D(P_{num}, P_{SFT}) = 3.94$

Note: without $x_0$, uses less information than the other methods in this example.
Autocorrelation of Samples from Bimodal problem

Autocorrelation for Uniform and $\chi^2$ based priors, SFT1 —, PT —— and SFT2 (with $x_0$) ...
Autocorrelation of Samples from the negative (L) and positive (R) modes of the Bimodal problem

Autocorrelation for Uniform and $\chi^2$ based priors (top and bottom resp.), SFT1 − , PT −− and SFT2 (with $x_0$) ...
FitzHugh Nagumo with a bad prior

\[ DV = \gamma \left( V - \frac{V^3}{3} + R \right), \quad DR = -\left( \beta R + \alpha - V \right) / \gamma \]

Using a one parameter model with the prior \( N(14,2) \)
FitzHugh Nagumo with a bad prior

\[ DV = \gamma \left( V - \frac{V^3}{3} + R \right), \quad DR = -\frac{(\beta R + \alpha - V)}{\gamma} \]

Using a one parameter model with the prior \( N(14,2) \)
Conclusion

- Faster mixing - less time sampling unimportant minor modes
- Improved basin of attraction by smoothing out the posterior topology.
- Faster convergence.
- Reduces or removes the impact of initial system states.
- Produces Inference on ODE solution and smooth deviations thereof.
- Benefits from feature matching and data fitting.
- Works even when there are unobserved system components