Dave Campbell and Russell Steele

Smooth Functional Tempering for Nonlinear Differential Equation Models

Functional Data Methods for Bayesian Parameter Estimation in DE Models

$$\frac{d\mathbf{x}(t)}{dt} = D\mathbf{x}(t) = f\Big(\mathbf{x}(t), \theta\Big)$$

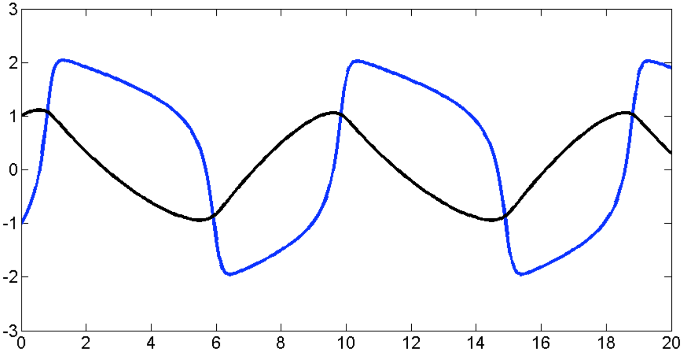
$$\Rightarrow \mathbf{x}(t) = ??$$

- The goal is to estimate θ
- 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

FitzHugh-Nagumo system

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



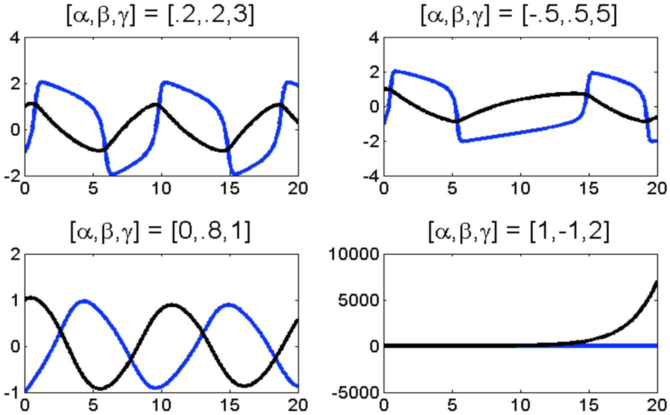
Numerical Solution to the ODE using:

$\boldsymbol{\theta} = [\alpha, \beta, \gamma] = [0.2, 0.2, 3]$ and $[V_0, R_0] = [-1, 1]$

FitzHugh-Nagumo system

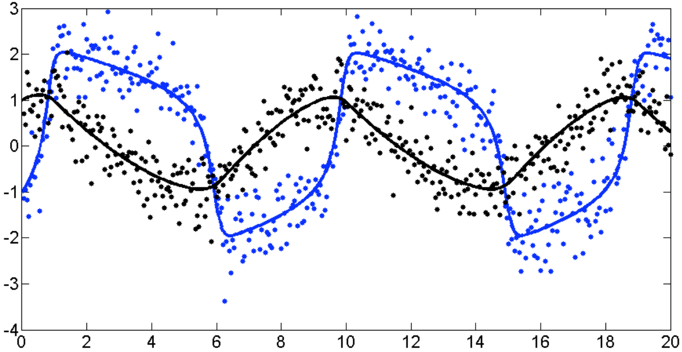
$$DV = \gamma\Big(V - V^3/3 + R\Big), \quad DR = -(\beta R + \alpha - V)/\gamma$$

The behaviour modeled changes with α, β, γ, V_0, and R_0



FitzHugh-Nagumo system

$$DV = \gamma\left(V - V^3/3 + R\right), \qquad DR = -\left(\beta R + \alpha - V\right)/\gamma$$



401 evenly spaced points with noise $N(0, .5^2)$ and $N(0, .4^2)$.
$\boldsymbol{\theta} = [\alpha, \beta, \gamma] = [0.2, 0.2, 3]$ 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 θ but we need $\mathbf{x}_0$ to produce a numerical solution.

Outline

The Model Set Up[1]

For numerical solution $S(\boldsymbol{\theta}, V_0, R_0, t)$ to equations:

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

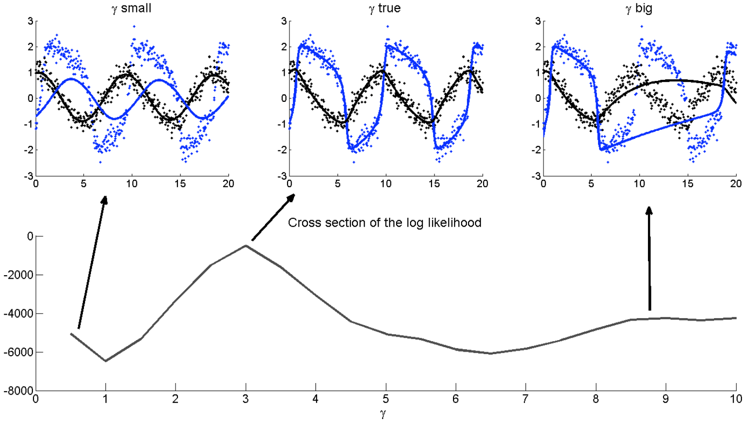
use a likelihood of the form:

$$\mathbf{y}(t) \mid \boldsymbol{\theta}, V_0, R_0, \Sigma \sim N\left\{S(\boldsymbol{\theta}, V_0, R_0, t), \Sigma\right\}.$$

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

[1] Gelman, Bois and Jiang, (1996), JASA, 91, 1400–1412. Huang and Wu (2006), Jo. of Ap. Stat., 33, 155-174.

Topological challenges



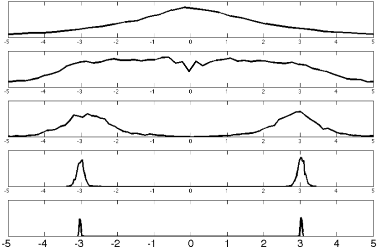
- Peaks correspond to (partial) data fits.
- Valleys imply that the fit deteriorates before it can improve.

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.

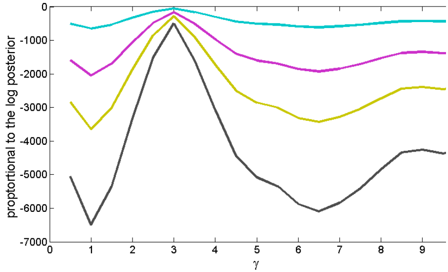
[2] Geyer, 1991, "Markov Chain Monte Carlo Maximum Likelihood", in Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface.

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.

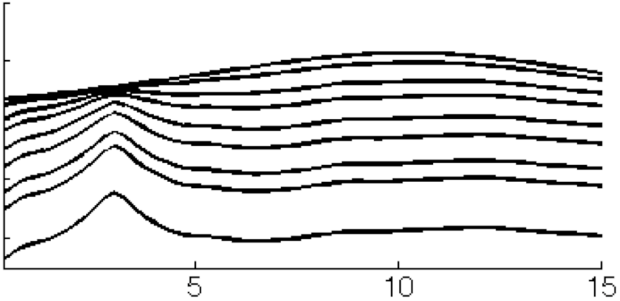
Parallel Tempering

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

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) $\mathbf{x}(t) = \mathbf{c}'\phi(t)$ to smooth the data.

Smooth Functional Tempering

The idea:

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

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

- 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**

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

- 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 $\lambda_1 < \ldots < \lambda_M \leq \infty$.

$$y(t) \mid \mathbf{x}(t), \sigma^2 \sim N\Big(\mathbf{x}(t), \sigma^2\Big)$$

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

- Build a sequence of M models with $\lambda_1 < \ldots < \lambda_M \leq \infty$.

$$y(t) \mid \mathbf{x}(t), \sigma^2 \sim N\Big(\mathbf{x}(t), \sigma^2\Big)$$

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

- This induces a density on $\mathbf{x}(t)$ without requiring us to sample $\mathbf{c}$.
- The induced density on $\mathbf{x}(t)$ decreases as $\mathbf{x}(t)$ strays from The DE solution.
- The rate of decrease depends on λ_m.

- Build a sequence of M models with $\lambda_1 < \ldots < \lambda_M \leq \infty$.

$$y(t) \mid \mathbf{x}(t), \sigma^2 \sim N\Big(\mathbf{x}(t), \sigma^2\Big)$$

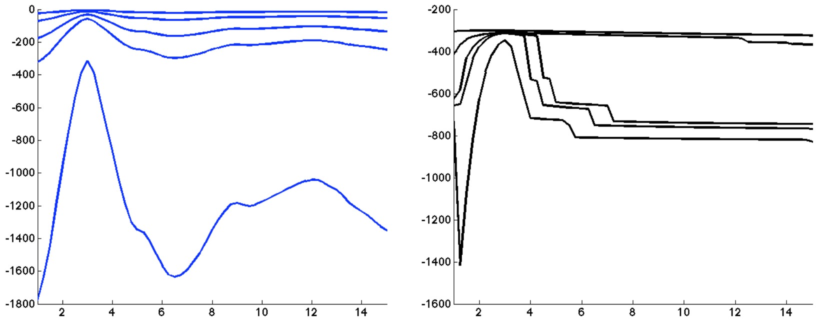
$$\pi(\boldsymbol{\theta}) \propto \exp\left\{-\lambda_m \int_t \left(D\mathbf{x}(v) - f(\mathbf{x}(v), \boldsymbol{\theta})\right)^2 dv\right\} p_1(\boldsymbol{\theta})$$

- Using big λ_M makes $\mathbf{x}(t)$ arbitrarily close to the DE solution.

But:

- We avoid numerically solving the DE.
- And we remove dependence on $\mathbf{x}_0$.

Posterior Cross Section



left (Parallel Tempering), right (Smooth Functional Tempering)

When $\mathbf{x}(0)$ is of interest.

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

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

$$y(t) \mid \mathbf{x}(\mathbf{x}_0, t), \sigma^2 \sim N\Big(\mathbf{x}(\mathbf{x}_0, t), \sigma^2\Big)$$

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

- Include $\mathbf{x}_0$ in the mode

When $\mathbf{x}(0)$ is of interest.

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

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

$$y(t) \mid \mathbf{x}(\mathbf{x}_0, t), \sigma^2 \sim N\Big(\mathbf{x}(\mathbf{x}_0, t), \sigma^2\Big)$$

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

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

When $\mathbf{x}(0)$ is of interest.

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

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

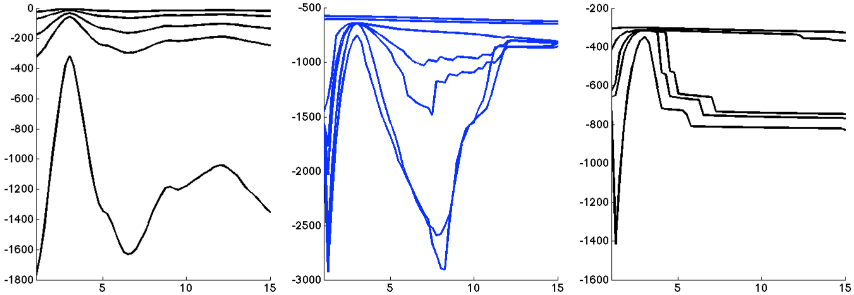
$$y(t) \mid \mathbf{x}(\mathbf{x}_0, t), \sigma^2 \sim N\Big(\mathbf{x}(\mathbf{x}_0, t), \sigma^2\Big)$$

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

- Include $\mathbf{x}_0$ in the mode
- as $\lambda \to \infty$ using a b-spline basis, $\mathbf{x}(\mathbf{x}_0, t) \mid \boldsymbol{\theta} \to s(\mathbf{x}_0, \boldsymbol{\theta}, t)$ using a Runga-Kutta numerical solver
- M^{th} model is equivalent to:

$$y(t) \mid \mathbf{x}_0, \boldsymbol{\theta}, \sigma^2 \sim N\Big(s(\mathbf{x}_0, \boldsymbol{\theta}, t), \sigma^2\Big)$$

$$\pi(\boldsymbol{\theta}) \sim p_1(\boldsymbol{\theta})$$

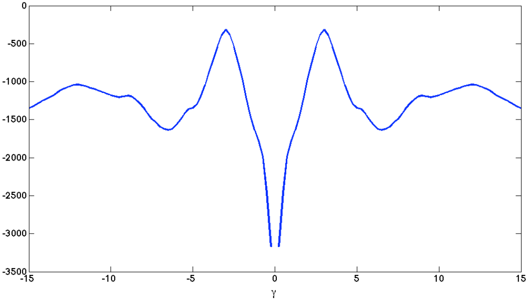


left (Parallel Tempering), mid (Smooth Functional Tempering with $\mathbf{x}_0$) right (Smooth Functional Tempering without $\mathbf{x}_0$)

Bimodal FitzHugh-Nagumo density

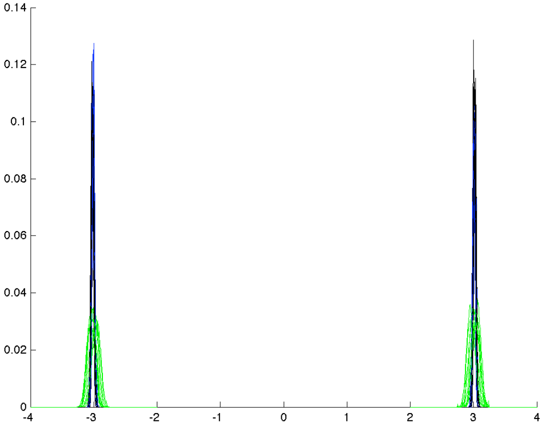
$$DV = |\gamma| \left(V - V^3/3 + R\right), \quad DR = -\left(\beta R + \alpha - V\right) / |\gamma|$$

Assume that all parameters except γ are known and fixed and $P(\gamma) = Uniform(-15, 15)$



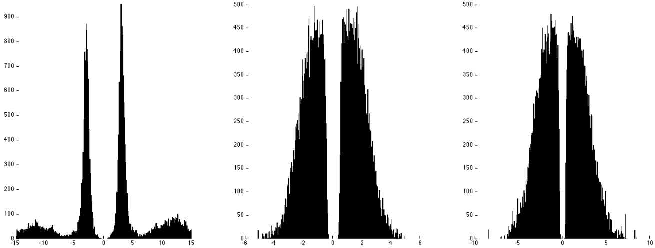
Tempering is required to sample from both modes.

Posterior Densities



SFT1, PT SFT2 (no X_0).

Samples from the $m = 1^{st}$ (the flattest) of the parallel chains using largest λ_1 that enables $\gamma = \pm 3$ modes to be sampled



left (Parallel Tempering), mid (Smooth Functional Tempering (SFT1) with $\mathbf{x}_0$) right (Smooth Functional Tempering (SFT2) without $\mathbf{x}_0$)

Quality of the λ_M chain approximation

Using 50,000 posterior iterations and the metric:

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

Quality of the λ_M chain approximation

Using 50,000 posterior iterations and the metric:

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

- $D(P_{num}, P_{parallel\ tempering}) = .0356$
- with $\mathbf{x}_0$; $D(P_{num}, P_{SFT}) = .0251$

Quality of the λ_M chain approximation

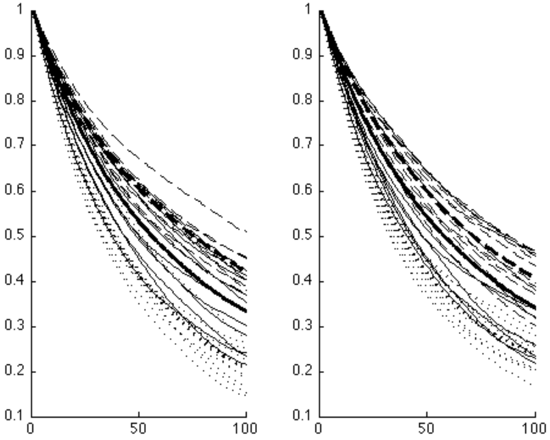
Using 50,000 posterior iterations and the metric:

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

- $D(P_{num}, P_{parallel\ tempering}) = .0356$
- with $\mathbf{x}_0$; $D(P_{num}, P_{SFT}) = .0251$
- without $\mathbf{x}_0$; $D(P_{num}, P_{SFT}) = 3.94$

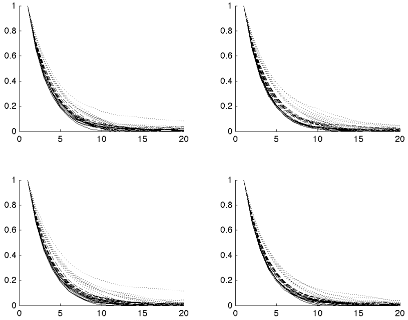
Note: without $\mathbf{x}_0$, uses less information than the other methods in this example.

Autocorrelation of Samples from Bimodal problem



Autocorrelation for Uniform and χ^2 based priors, SFT1 $-$, PT $--$ and SFT2 (with x_0) ...

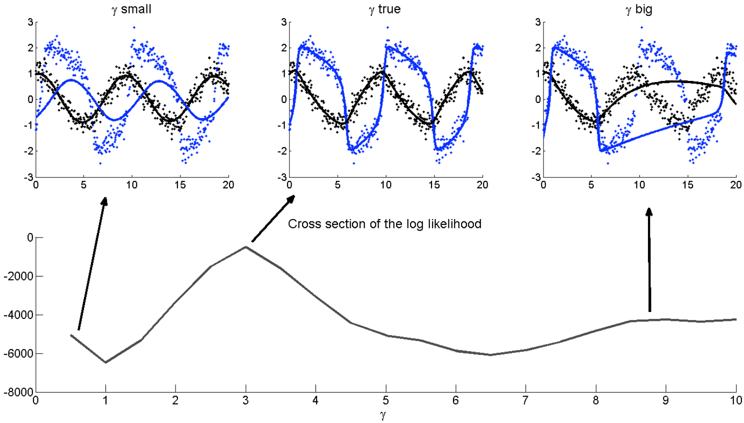
Autocorrelation of Samples from the negative (L) and positive (R) modes of the Bimodal problem



Autocorrelation for Uniform and χ^2 based priors (top and bottom resp.), SFT1 $-$, PT $--$ and SFT2 (with x_0) ...

FitzHugh Nagumo with a bad prior

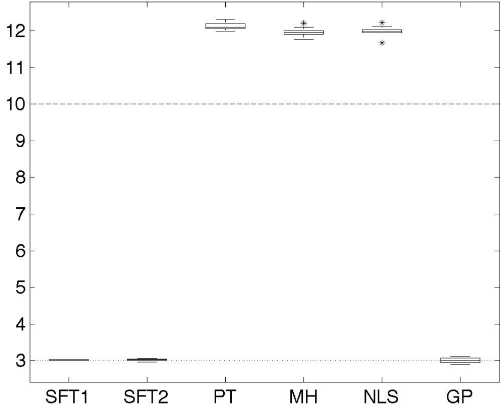
$$DV = \gamma\left(V - V^3/3 + R\right), \quad DR = -\left(\beta R + \alpha - V\right)/\gamma$$



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

FitzHugh Nagumo with a bad prior

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



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

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