Robustness, Inference and Gradient Matching

Giles Hooker

BIRS, July 2014

Problem Statement

Assume an ODE model for system dynamics

$$D\mathbf{x} = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}), \ \mathbf{x}(t_0) = \mathbf{x}_0$$

along with an observation model

$$\mathbf{y}_i = \mathbf{x}(t_i) + \boldsymbol{\epsilon}_i, \ i = 1, \ldots, n$$

and we wish to estimate $\boldsymbol{\theta}$.

Gradient Matching

Also two-stage least-squares

1. Estimate $\hat{\mathbf{x}}(t)$, $D\hat{\mathbf{x}}$ non-parametric smooth.
2. Choose $\boldsymbol{\theta}$ to minimize

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin} \sum_{q=1}^{Q} w_q \left|\left| D\hat{\mathbf{x}}(s_q) - \mathbf{f}(\hat{\mathbf{x}}(s_q), \boldsymbol{\theta}) \right|\right|^2$$

- Various smoothers suggested: polynomials (Bellman and Roth '71), basis expansions (Varah '82), Neural Networks (Ellner et. al. '00), local polynomials (Liang and Wu, '08)
- Requires set of evaluation points s_q to grow faster than observations (Xue et. al. '10)
- Can demonstrate consistency (Brunel '08), rates (Gugushvili '10), CLT (Wu et. al. '12)

Reasons for Two-Stage Least Squares

Usually justified over "trajectory matching" by

1. numerical expense of solving ODEs (not really more costly than evaluating GM objective)
2. less difficult objective functions $\Rightarrow$ at least good starting values

I think these miss the most important aspect of gradient matching:

- Allows for lack of fit due to system disturbances:

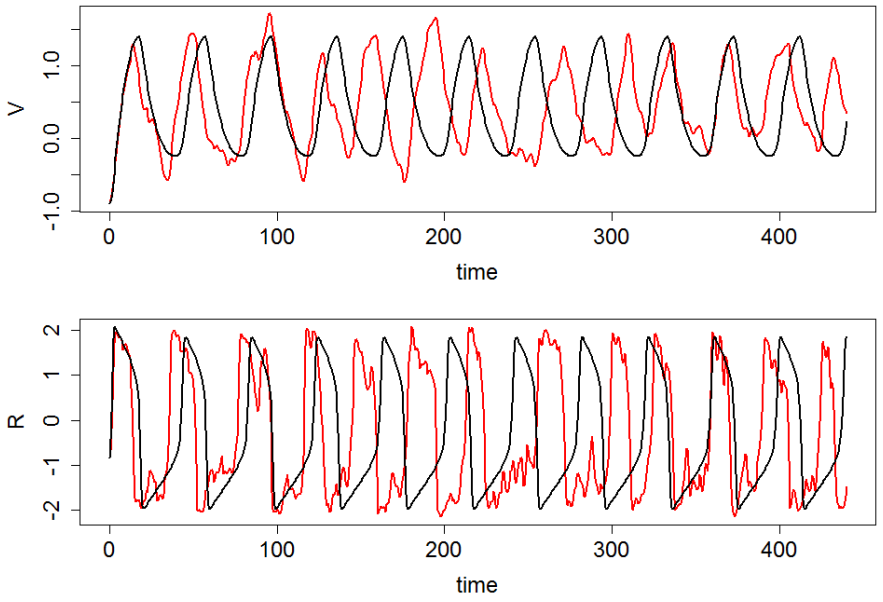
$$D\mathbf{x} = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) + \mathbf{g}(t)$$

We can think of this as a form of robustness to system mis-specification, or as allowing stochastic effects.

If disturbances random, must account for them in variance estimates.

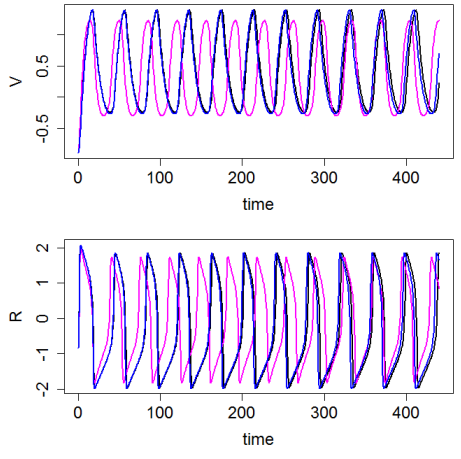
Simulated Example

FitzHugh-Nagumo equations forced by smooth random noise



Parameter Estimates

Truth: 0.080 0.056 0.064 0.500
GradM: 0.079 0.054 0.057 0.479
TrajM: 0.117 0.073 0.144 0.435



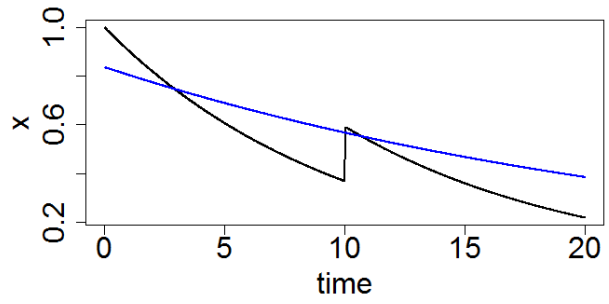
Robustness to System Disturbances

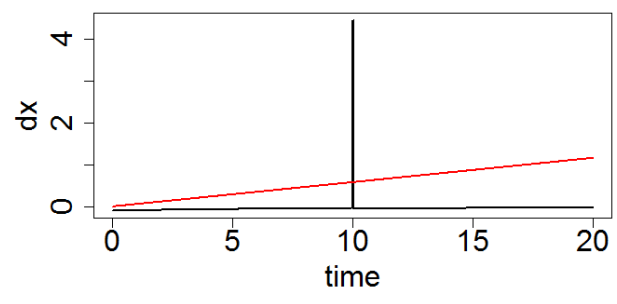
Imagine a model

$$D\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}) + \alpha U_\tau(t), U = I(\tau < t < \tau + \delta)$$

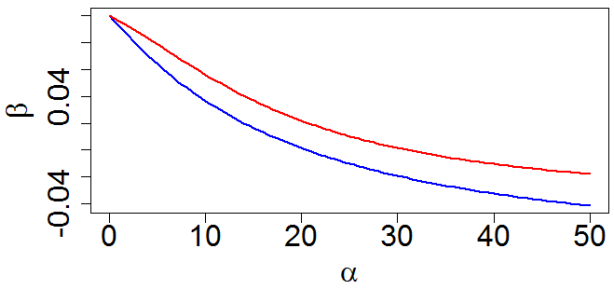
we might hope that $\hat{\boldsymbol{\theta}}$ from gradient matching is less sensitive for $\alpha \to \infty$ than from trajectory matching.

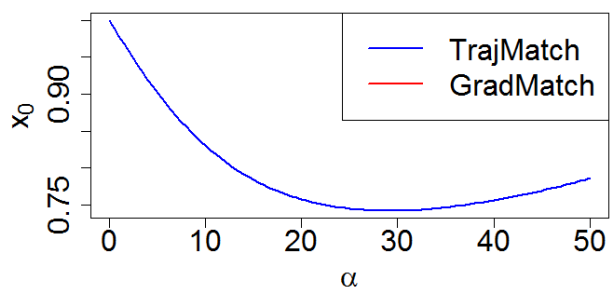
$Dx = -\beta x + \alpha U_\tau(t)$





Parameter Estimates





For squared error; don't distinguish rate of displacement with α.

Smoothly Stochastic Systems

An alternative model for inexact ODE's is with smooth, stochastic stationary perturbations:

$$D\hat{\mathbf{x}}(t) = \mathbf{f}(\hat{\mathbf{x}}(t), \boldsymbol{\theta}) + \boldsymbol{\eta}(t)$$

where $\boldsymbol{\eta}(t)$ is random with

$$E\boldsymbol{\eta}(t) = 0, \ \text{cov}(\boldsymbol{\eta}(t), \boldsymbol{\eta}(t+s)) = \mathbf{R}(s)$$

(Note lack of fit includes smoothing)

Here we must account for two sources of variation

1. Observational noise $\boldsymbol{\epsilon}_i$
2. Process noise $\boldsymbol{\eta}(t)$

This is easy to do in gradient matching (much harder in trajectory matching).

But no papers examine it.

Accounting for Variance Components

Write out first order expansion of our estimate as functions of smooth condition on $\boldsymbol{\eta}$:

$$\hat{\boldsymbol{\theta}}(\hat{X}, \widehat{DX}) = \boldsymbol{\theta}(E\hat{X}, E\widehat{DX}) - \left[\sum_{j=1}^{d} F_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_0)^T W_j F_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right]^{-1} \sum_{j=1}^{d} F_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_0)^T W_j \left(\widehat{DX}_j - F_j(\boldsymbol{\theta}_0)\right)$$

- $\hat{X}$, $\widehat{DX}$ = matrix of evaluations of smooths at quadrature points
- $F(\boldsymbol{\theta}_0)$ = matrix containing $\mathbf{f}(E\hat{X}, \boldsymbol{\theta}_0)$
- $F_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_0) = \nabla_{\boldsymbol{\theta}} f_j(E\hat{X}, \boldsymbol{\theta}_0)$.
- W_j = diagonal matrix of weights.

Accounting for Variance Components

Re-writing

$$\hat{\boldsymbol{\theta}}(\hat{X}, \widehat{DX}) = \boldsymbol{\theta}(E\hat{X}, E\widehat{DX}) - H(\boldsymbol{\theta}_0)^{-1} \sum_{j=1}^{d} F_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_0)^T W_j \left(\widehat{DX}_j - F_j(\boldsymbol{\theta}_0) \right)$$

Variance comes from

$$\begin{aligned} \widehat{DX}_j - F_j(\boldsymbol{\theta}_0) \approx \widehat{DX}_j - E\widehat{DX}_j \\ + \nabla_{\mathbf{x}} f_j(\hat{X}; \boldsymbol{\theta}_0)(\hat{X} - E\hat{X}) \\ + E\widehat{DX}_j - f_j(E\hat{X}; \boldsymbol{\theta}_0). \end{aligned}$$

after tailor expansion, (later write $F_{j\mathbf{x}} = \nabla_{\mathbf{x}} f_j(\hat{X}; \boldsymbol{\theta}_0)$).

First two terms = variance associated with ϵ, last = variance from η.

Accounting for Measurement Variance

Assume that $\hat{X}$, $\widehat{DX}$ obtained by linear smoothers:

$$\begin{bmatrix} \hat{X} \\ \widehat{DX} \end{bmatrix} = \begin{bmatrix} S_0 \\ S_1 \end{bmatrix} Y$$

then we can estimate a covariance Σ of rows of Y and write

$$J_j(\boldsymbol{\theta}) = H(\boldsymbol{\theta})^{-1} \left(\sum_{k=1}^{d} F_{k\boldsymbol{\theta}}(\boldsymbol{\theta})^T W_k \begin{bmatrix} \mathrm{diag}(F_{kx_j}(\boldsymbol{\theta})) \\ \mathrm{diag}(1_{j=k}) \end{bmatrix} \right)$$

and for observation-only variances

$$\mathrm{var}(\boldsymbol{\theta}_0) = V = \sum_{j=1}^{d} \sum_{k=1}^{d} \sigma_{jk} J_j(\boldsymbol{\theta}) \begin{bmatrix} S_0 \\ S_1 \end{bmatrix} \begin{bmatrix} S_0^T & S_1^T \end{bmatrix} J_k(\boldsymbol{\theta})^T.$$

Other literature bootstraps empirical residuals.

Accounting for Process Variance

Assume $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are independent $\Rightarrow$ can separate variance contributions.

Estimate $\hat{\boldsymbol{\eta}}$ and its autocovariance

$$\hat{\boldsymbol{\eta}}(t) = \widehat{D\mathbf{x}}(t) - f(t, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}), \;\; \hat{\mathbf{R}}(u) = \frac{1}{1-u} \int_0^{t-u} \hat{\boldsymbol{\eta}}(t)\hat{\boldsymbol{\eta}}(t+u)dt$$

Then let

$$\left[R_{ij}^*\right]_{qr} = R_{ij}(|t_q - t_r|)$$

and

$$\text{var}(\hat{\boldsymbol{\theta}}) = V + H(\hat{\boldsymbol{\theta}})^{-1} \left[\sum_{j=1}^{d} \sum_{k=1}^{d} F_{j\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})^T W_j R_{jk}^* W_k F_{k\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) \right] H(\hat{\boldsymbol{\theta}})^{-1}.$$

Also accounts for process variance.

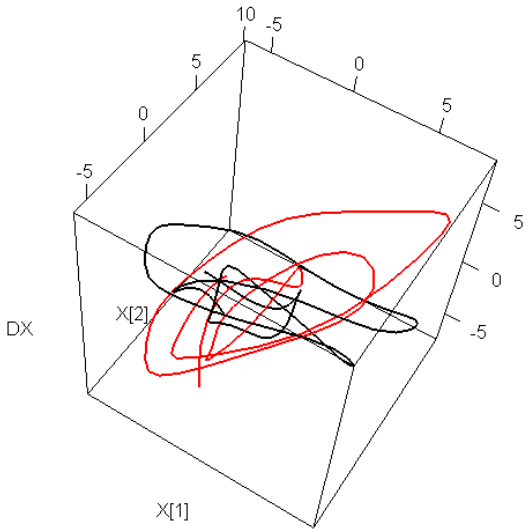
A Simulation Study

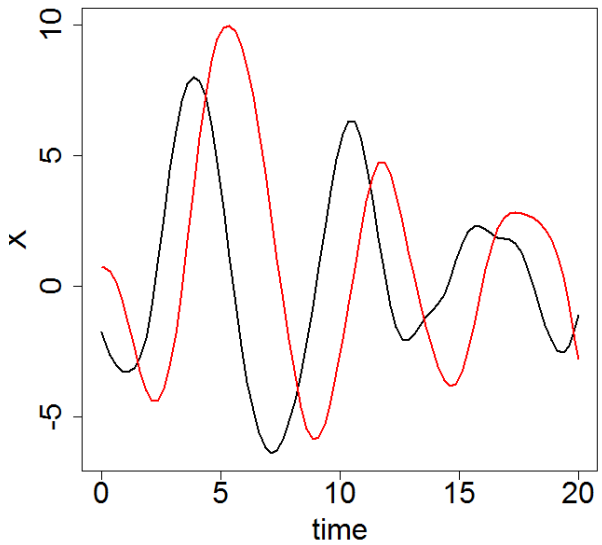
Two-dimensional Linear model

$$D\mathbf{x} = A\mathbf{x} + B\boldsymbol{\eta}(t)$$

with

$$\mathrm{cov}(\boldsymbol{\eta}(t), \boldsymbol{\eta}(t+s)) = e^{-s^2}, \; B = \mathrm{diag}[\sqrt{2}, 1]$$





500 Simulations

Estimates from gradient matching

True	0	-1	1	0
Mean	0.080	-0.920	0.980	0.028
SD	0.088	0.125	0.080	0.073

Coverage of confidence intervals with

Observation Noise Only	0.566	0.364	0.620	0.630
Process Noise	0.950	0.848	0.894	0.888
Newey-West on Scores	0.724	0.476	0.554	0.588

Testing Nonlinear Effects: FitzHugh-Nagumo Example

Estimates from gradient matching

True	0.080	0.056	0.064	0.500
Mean	0.079	0.054	0.059	0.513
SD	0.003	0.006	0.008	0.035

Coverage of confidence intervals with

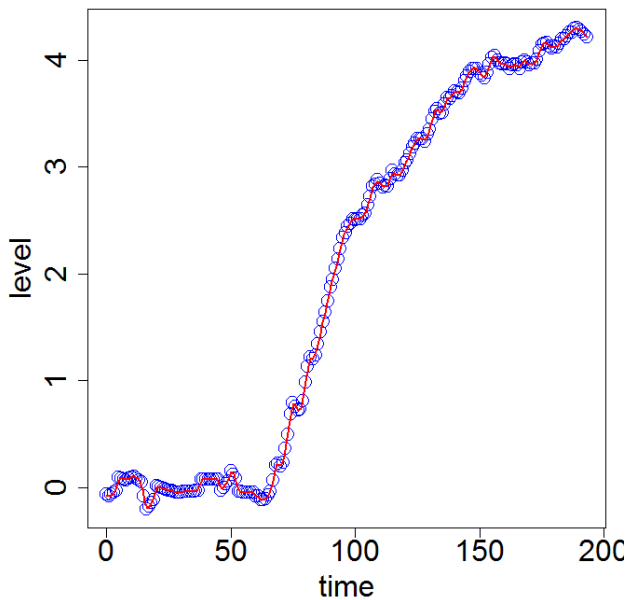
Observation Noise Only	0.316	0.314	0.318	0.668
Process Noise	0.900	0.868	0.822	0.905
Newey-West on Scores	0.446	0.494	0.434	0.668

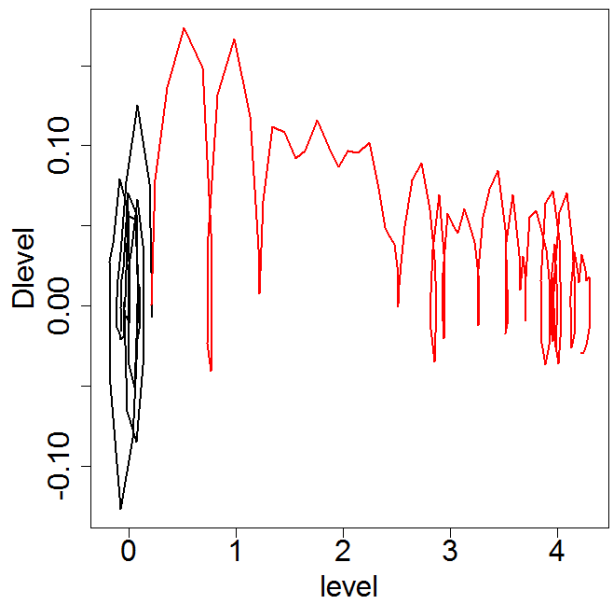
Real World Impacts

Refinery data from Functional Data Analysis

Assumed model

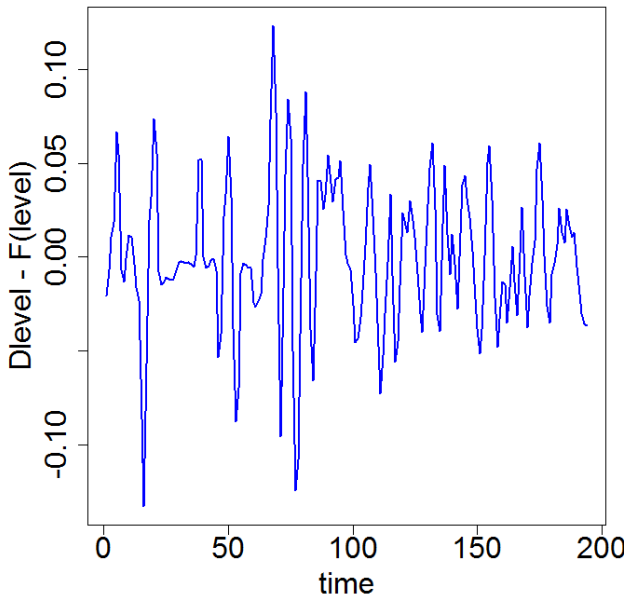
$$D\mathbf{x} = \beta\mathbf{x} + \alpha(t < 69) + \gamma(t > 69)$$

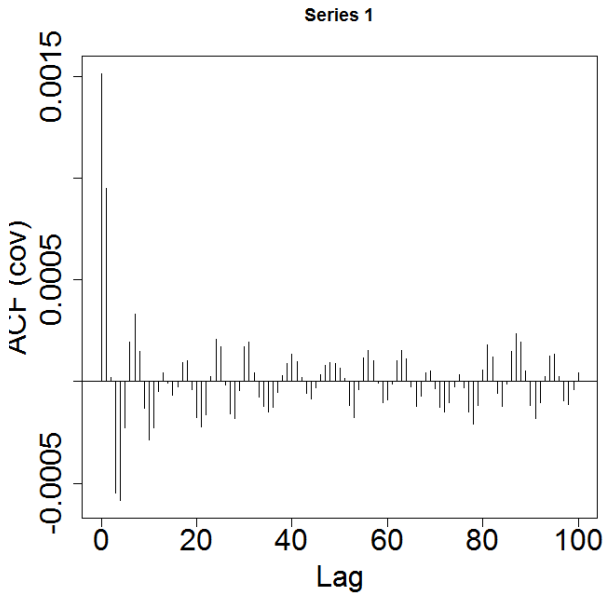




Real World Impacts

Refinery residuals and autocovariance





Estimates

	β	α	γ
Estimate	-0.021	0.004	0.097
Obs Noise SD	0.000173	0.000258	0.000493
All Noise SD	0.002550	0.002955	0.008232

Some Comments

Statistical properties of $\hat{\boldsymbol{\theta}}$ not as clear

- Requires at least expanding domain asymptotics for CLT
- Infill asymptotics to remove bias

Bootstrap estimates useful, too

- Permits bias correction
- Block bootstrap of $\hat{\boldsymbol{\eta}}(t)$
- Requires care to evaluate impact on $\hat{X}$ *and* $\widehat{DX}$.

Extensions 1: Profiling and Other Methods

For other methods can account for $\boldsymbol{\eta}$ in covariance of Y.

Linear noise approximation:

$$\text{cov}(\mathbf{x}(t), \mathbf{x}(t')) \approx e^{\bar{A}(t)+\bar{A}(t')} \int_0^t \int_0^{t'} e^{-\bar{A}(s)-\bar{A}(s')} \mathbf{R}(|s-s'|) ds ds'$$

with

$$\bar{A}(t) = \int_0^t \nabla_{\mathbf{x}} \mathbf{f}(\boldsymbol{\xi}(s), \boldsymbol{\theta}) ds, \ \boldsymbol{\xi} \text{ solves ODE}$$

Add observation variance and can use in

- trajectory matching
- integral matching
- profiling

Profiling gives easy estimate of $\mathbf{R}(s)$. Needs global LNA: not very accurate.

Extensions 2: Covariance Modeling

- Current $\text{cov}(\boldsymbol{\eta}(t), \boldsymbol{\eta}(t+s)) = \mathbf{R}(s)$ rather naive.
- Can incorporate more complex models
  - Scaling variance with $\mathbf{x}$
  - Allowing $\mathbf{R}(s)$ to evolve (slowly)
  - Determining $\mathbf{R}(s)$ from probabilistic model or process, at least partially.
  - May end up with multiplicative noise model; needs new LNA
- Need good ways to specify and estimate a more complex covariance process.
- How to account for smoothing process within this?
- Requires more sophisticated block bootstrap methods.

Conclusion

- Gradient matching can allow for model mis-specification.
- Can account for stochastic effects in confidence intervals.
- Mathematical theory: requires expanding time at least (in-fill asymptotics to control bias).
- Will this account for smoothing over non-smooth systems?
- Profiling: similar cancellation effect to gradient matching; will this make global LNA less bad?
- Many opportunities for more sophisticated covariance modeling.