Exploiting Geometry to Design MCMC Methods for Nonlinear Dynamic Systems

Julien Cornebise, Mark Girolami
Department of Statistics
University College London

Workshop on Geometric and Algebraic Statistics 3
Apr 5-7 2011
CRiSM, Warwick, UK

Outline

We are here → •

Components: for parameters $\theta = (u, v, \text{vec}(\Sigma))$

1. Dynamic system, hidden: $x(u,t)$, solution of ODE system $\dot{x}(u,t) = f(x(u,t,),u,t)$
2. Response Function: $r(x)$ – e.g. partial measurement
3. Observation Model: $m(r,v)$ – e.g. unknown scaling constants (cytography)
4. Error Model: noise $\mathbf{z}$ and $h(m)$ – transformation, e.g. log-scale

Observations $h(y_j) \sim h(m(r(x(u,t)),v) + \sqrt{\Sigma}\mathbf{z})$

Typical problems:

- Nonlinearity
- High number of parameters
- Lack of identifiability or Weak identifiability
- Partial observation / Heteroscedastic variance

Expected FIM

$$\mathbf{G}(\theta) = -\mathbb{E}\left[H_\theta \log p(y,\theta)\right] = \sum_{i=1}^{n} \mathbf{j}_i^T(\theta)\Sigma^{-1}\mathbf{j}_i(\theta) + H_\theta \log p(\theta)$$
$$= \mathbf{j}^T(\theta)\left(I_n \otimes \Sigma^{-1}\right)\mathbf{j}(\theta) + H_\theta \log p(\theta)$$

$$\mathbf{j}(\theta) = [\mathbf{j}_1(\theta) \dots \mathbf{j}_n(\theta)]$$
$$\mathbf{j}_i(\theta) = J_\theta h \circ m \circ r \circ x|_{t_i,u,v}$$
$$= \begin{bmatrix} J_u h \circ m \circ r \circ x|_{t_i,u,v} & J_v h \circ m|_{r\circ x(u,t_i),v} & I_d \end{bmatrix}$$

Basic chain rule: $J_u h \circ m \circ r \circ x = J_m h \cdot J_r m \cdot J_x r \cdot J_u x$.

Sensitivities of Dynamic Systems

Sensitivities $J_u x$ simply computed by extending the ODE

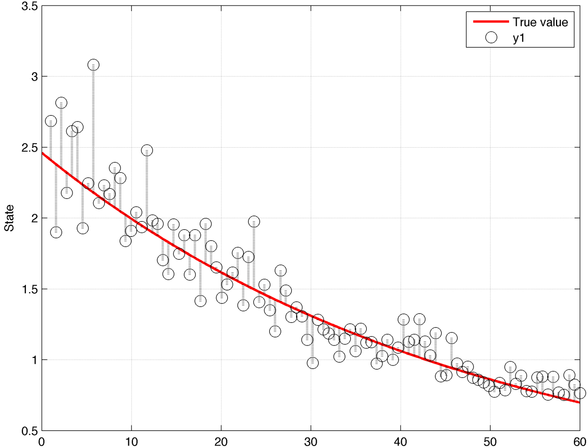
$$\begin{cases} \dot{x}(u,t) = f(x(u,t),u,t) \\ \dot{J_u x}\Big|_{u,t} = J_x f|_{x(u,t),u,t} \cdot J_u x|_{u,t} + J_u f|_{x(u,t),u,t} \end{cases}$$

Same holds for $H_u x$ appearing in $\partial \mathbf{G}/\partial\theta_k$ for connexions and Christoffel symbols.

Even **simplest** model illustrates those problems:

$$\begin{cases} \dot{x}(t) = -k_1 x(t) \\ \log y_j = \log x(t_j) + \sigma \epsilon_j \end{cases}$$

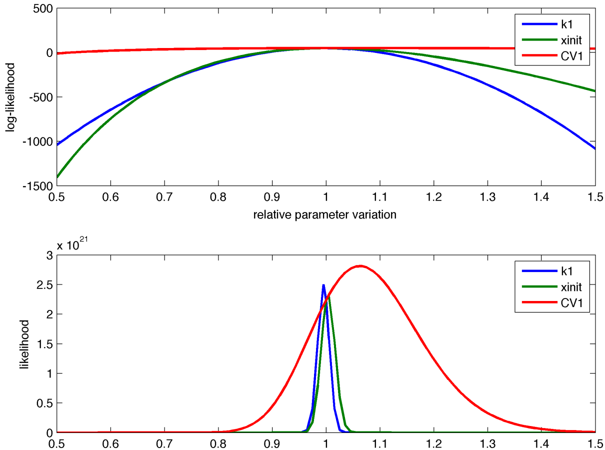
unknown initial condition $x(0) = x_{\text{init}}$



Even **simplest** model illustrates those problems:

$$\begin{cases} \dot{x}(t) = -k_1 x(t) \\ \log y_j = \log x(t_j) + \sigma \epsilon_j \end{cases}$$

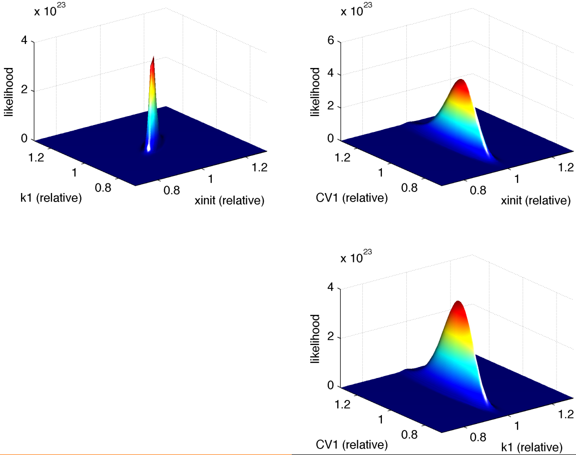
unknown initial condition $x(0) = x_{\text{init}}$



Even **simplest** model illustrates those problems:

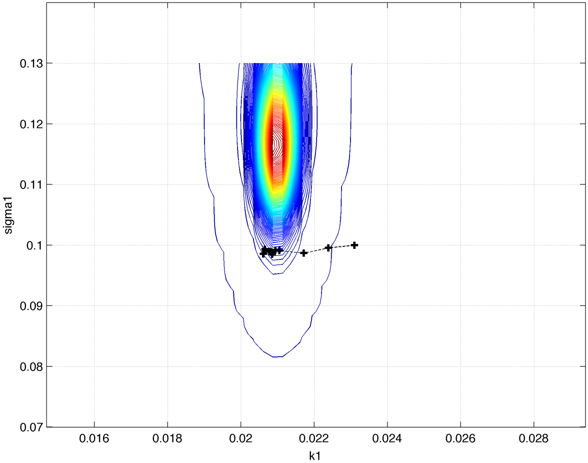
$$\begin{cases} \dot{x}(t) = -k_1 x(t) \\ \log y_j = \log x(t_j) + \sigma \epsilon_j \end{cases}$$

unknown initial condition $x(0) = x_{\text{init}}$



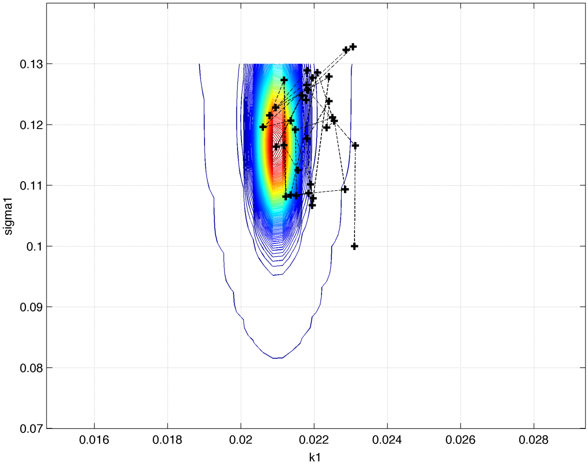
Extremely narrow ridges: impossible for run-off-the-mill MCMC

Tuned Spherical Proposal



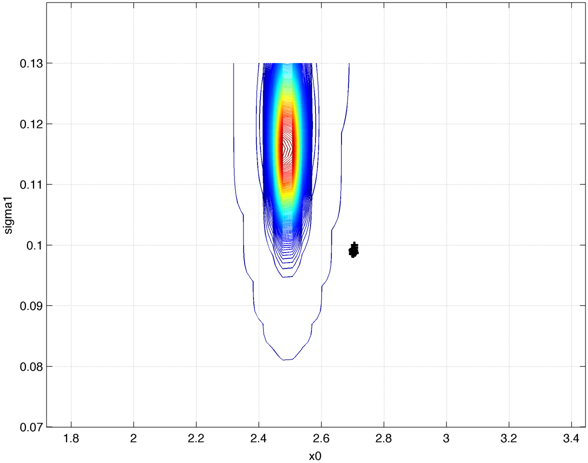
Extremely narrow ridges: impossible for run-off-the-mill MCMC

inv(FIM) as Proposal



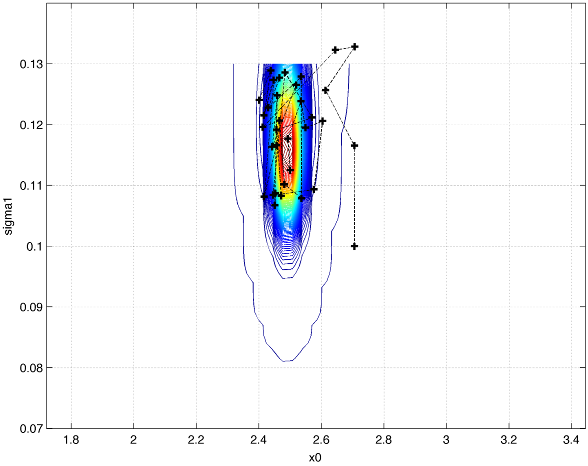
Extremely narrow ridges: impossible for run-off-the-mill MCMC

Tuned Spherical Proposal



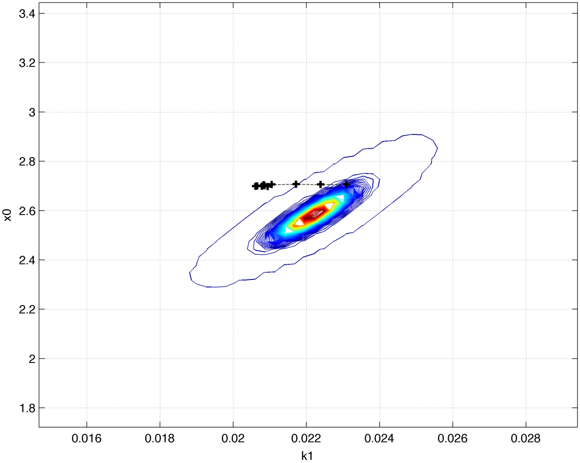
Extremely narrow ridges: impossible for run-off-the-mill MCMC

inv(FIM) as Proposal



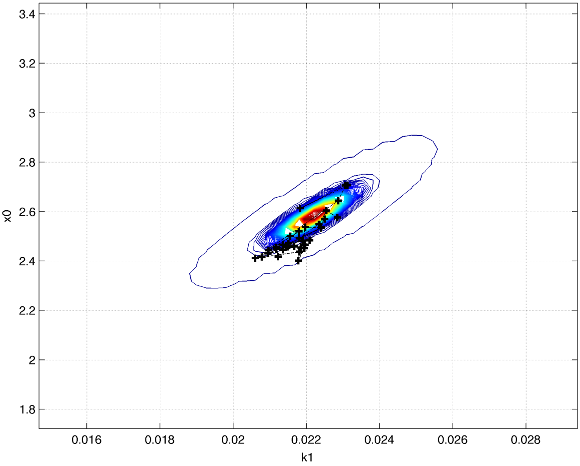
Extremely narrow ridges: impossible for run-off-the-mill MCMC

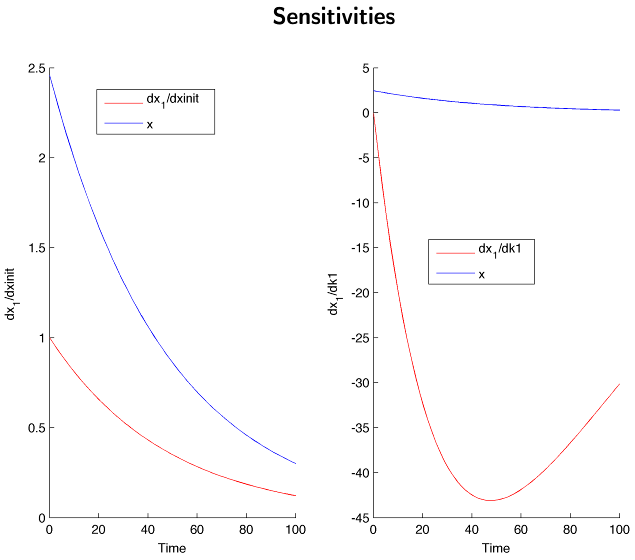
Tuned Spherical Proposal



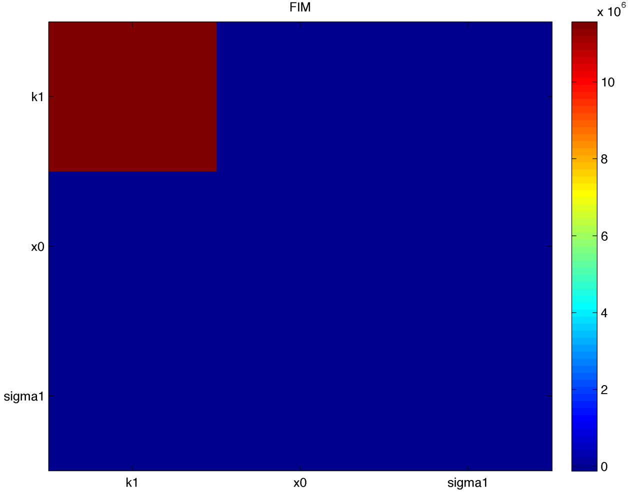
Extremely narrow ridges: impossible for run-off-the-mill MCMC

inv(FIM) as Proposal

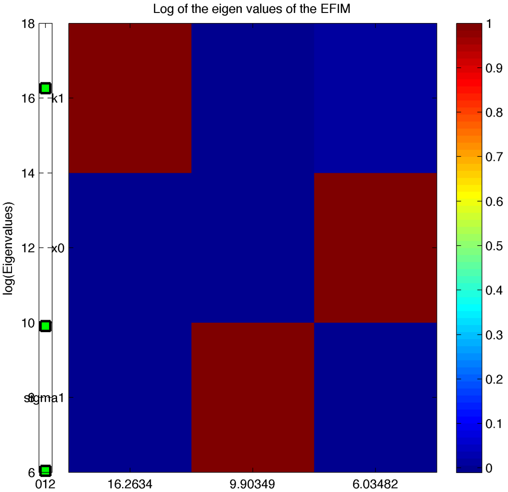



Sensitivities
2.5
2
1.5
1
0.5
0
0
20
40
60
80
100
Time
dx1/dxinit
x
5
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
dx1/dk1
Time

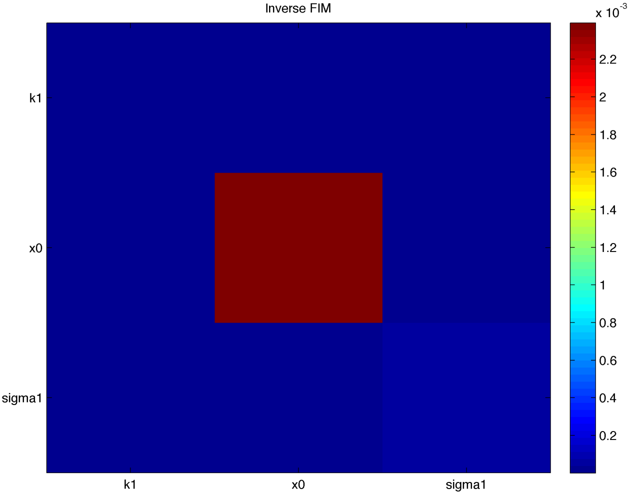
Expected FIM at Generative (Parameter) Value



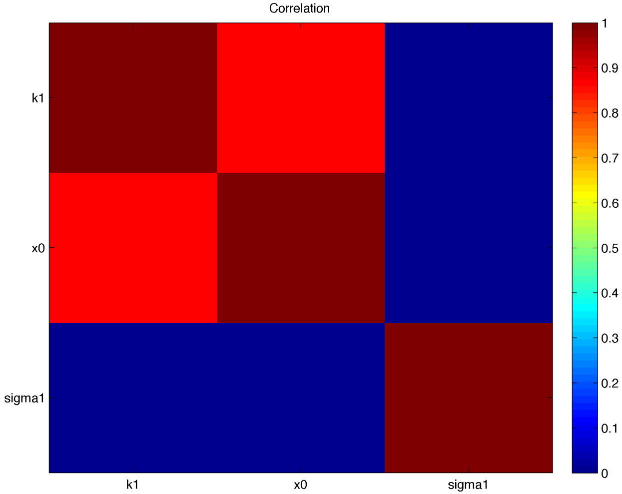
Eigenvalues and Eigenvectors of Expected FIM at Generative Value



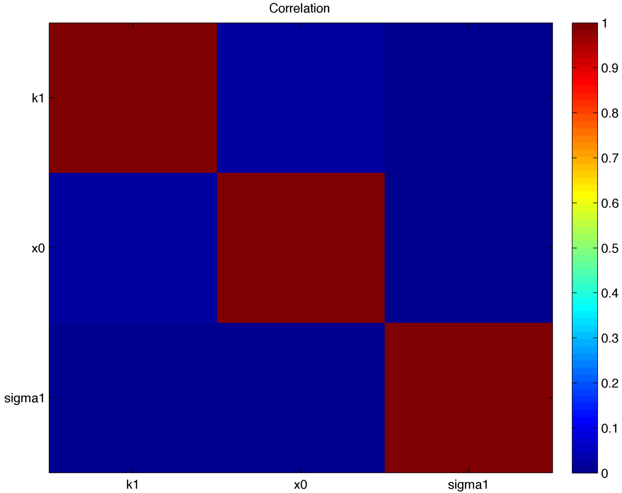
Inverse of FIM at Generative Value

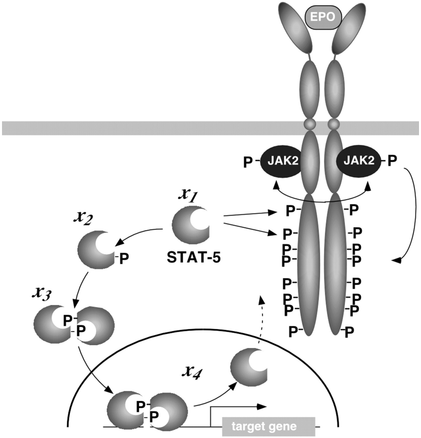


Normalised Inverse of FIM (Correlation Matrix) at Generative Value



Normalised Inverse of FIM Far from Generative Value





Courtesy of Swameye et al. (2003, PNAS)

Hidden process:

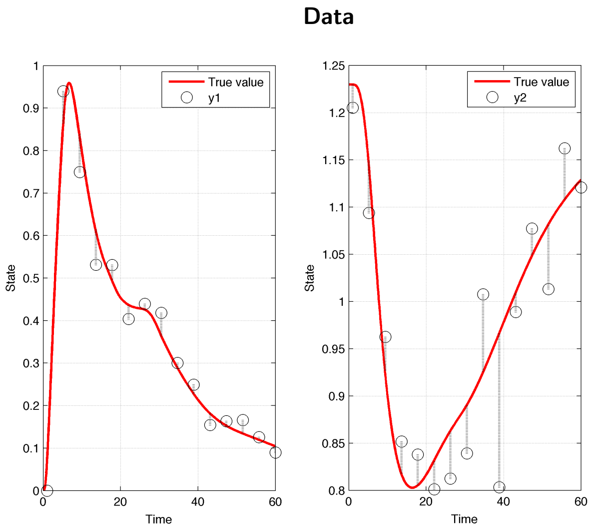
$$
\begin{cases}
\dot{x}_1(t) = -k_1 x_1(t)\mathsf{EpoR_A}(t) + k_5 x_5(t) \\
\dot{x}_2(t) = -x_2^2(t) + k_1 x_1(t)\mathsf{EpoR_A}(t) \\
\dot{x}_3(t) = -k_3 x_3(t) + \frac{1}{2} x_2^2(t) \\
\dot{x}_4(t) = -k_4 x_4(t) + k_3 x_3(t) \\
\dot{x}_5(t) = -k_5 x_5(t) + 2k_4 x_4(t)
\end{cases}
$$

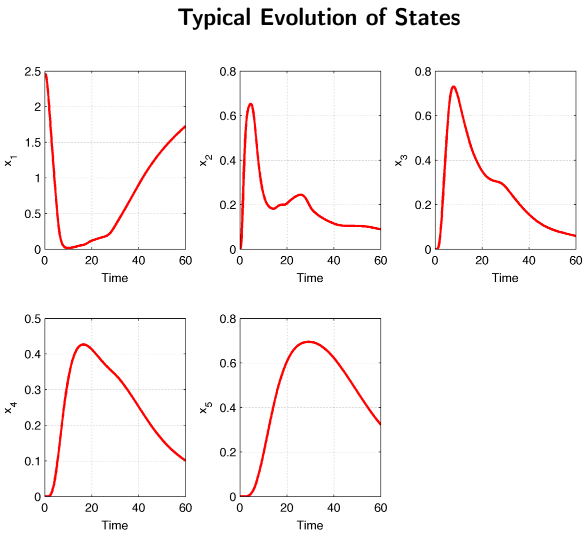
Initial condition:

$x(0) = (x_{\mathsf{init}}, 0, 0, 0)$.

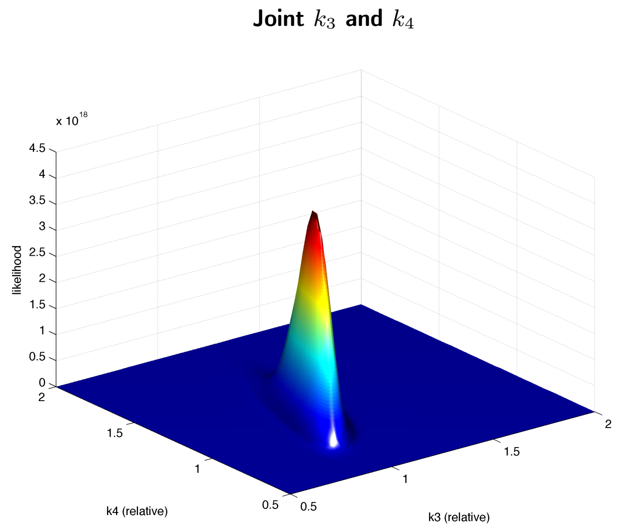
Observation:

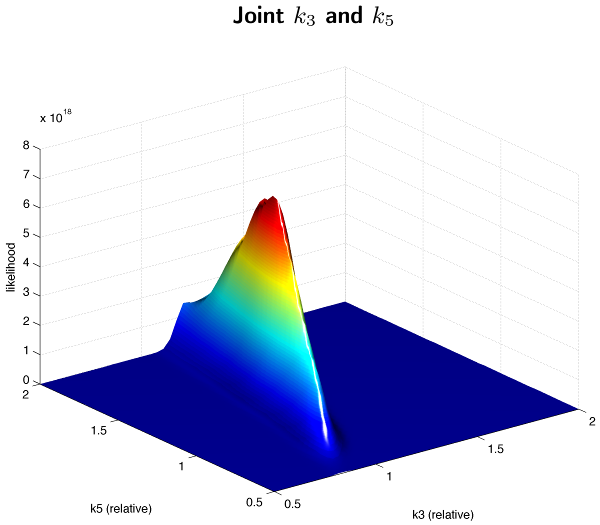
$$
\log y_{j,1} = \log\left[v_1(x_2(t_j) + 2x_3(t_j))\right] + \sigma_1 \epsilon_{j,1}
$$
$$
\log y_{j,2} = \log\left[v_2(x_1(t_j) + x_2(t_j) + 2x_3(t_j))\right] + \sigma_2 \epsilon_{j,2}
$$

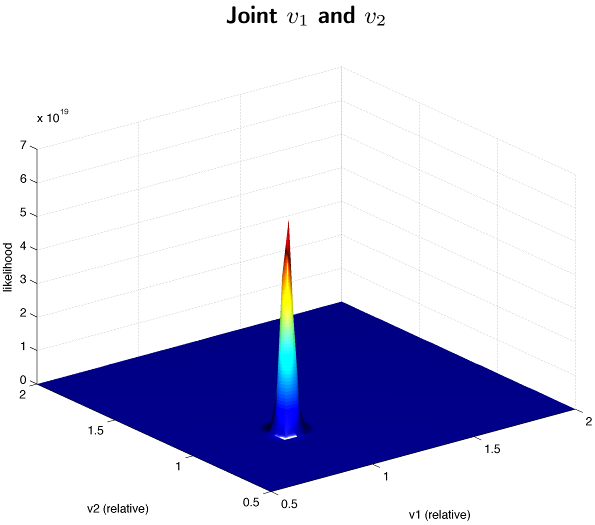

Data
State
Time
True value
y1
y2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
40
60
1.25
1.2
1.15
1.1
1.05
1
0.95
0.9
0.85
0.8

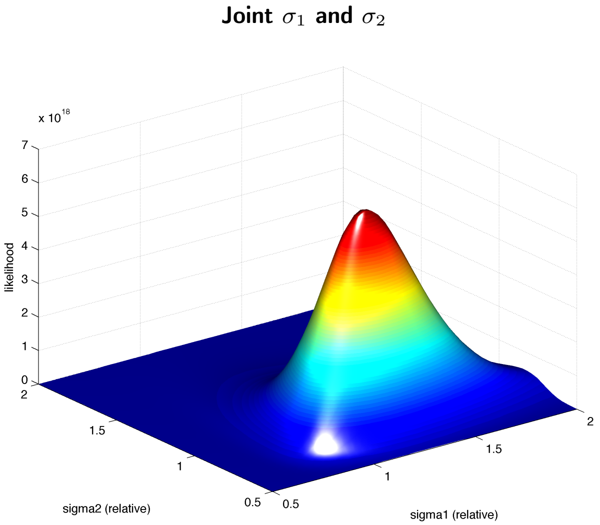

Typical Evolution of States
x1
x2
x3
x4
x5
Time
0
20
40
60
2.5
2
1.5
1
0.5
0.8
0.6
0.4
0.2
0.5
0.3
0.1

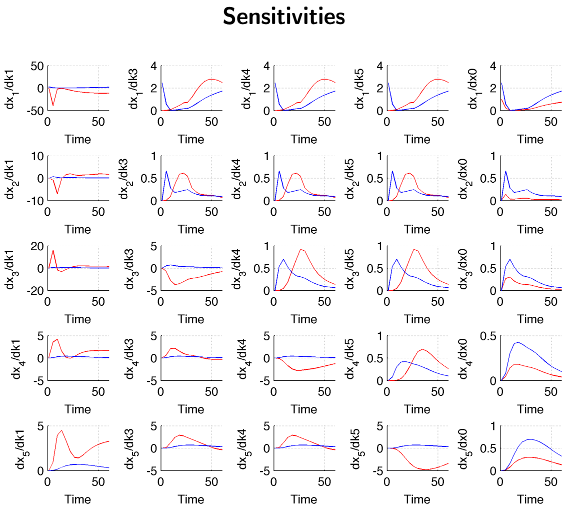
Metric at the Generative Value



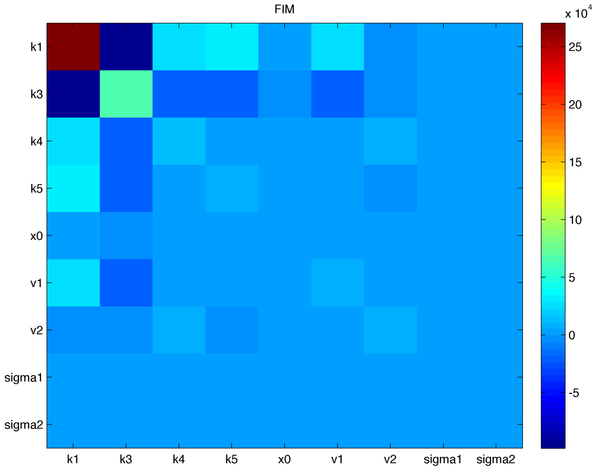

Joint k_3 and k_5
x 10^18
likelihood
8
7
6
5
4
3
2
1
0
2
1.5
1
0.5
k5 (relative)
0.5
1
1.5
2
k3 (relative)

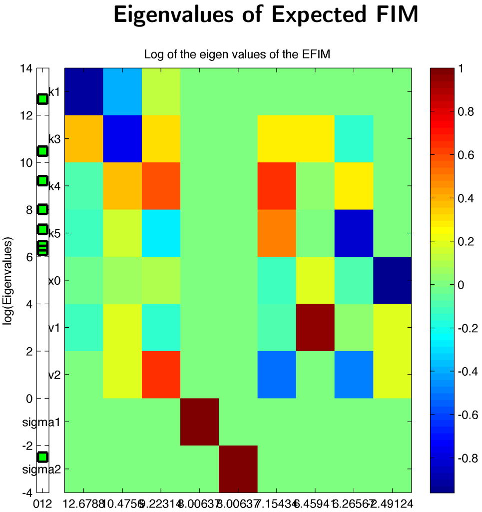

Joint v_1 and v_2
x 10^19
7
6
5
4
3
2
1
0
likelihood
2
1.5
1
0.5
v2 (relative)
0.5
1
1.5
2
v1 (relative)


Joint σ_1 and σ_2
x 10^18
7
6
5
4
3
2
1
0
likelihood
2
1.5
1
0.5
sigma2 (relative)
0.5
1
1.5
2
sigma1 (relative)

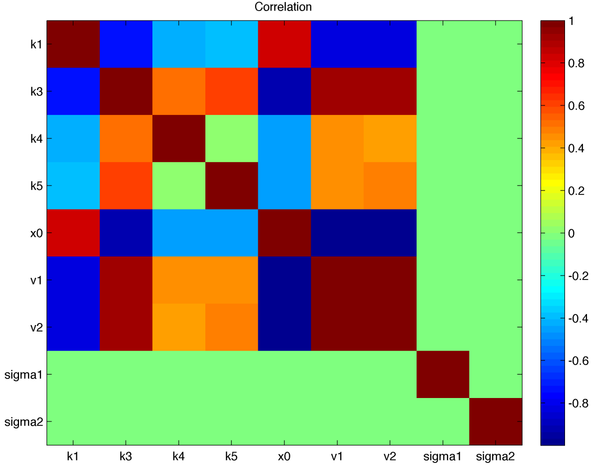

Sensitivities
dx_1/dk1
dx_1/dk3
dx_1/dk4
dx_1/dk5
dx_1/dx0
dx_2/dk1
dx_2/dk3
dx_2/dk4
dx_2/dk5
dx_2/dx0
dx_3/dk1
dx_3/dk3
dx_3/dk4
dx_3/dk5
dx_3/dx0
dx_4/dk1
dx_4/dk3
dx_4/dk4
dx_4/dk5
dx_4/dx0
dx_5/dk1
dx_5/dk3
dx_5/dk4
dx_5/dk5
dx_5/dx0
Time

Expected FIM



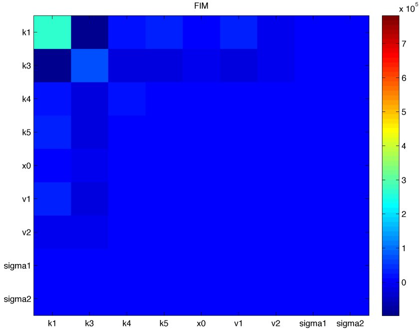

Eigenvalues of Expected FIM
Log of the eigen values of the EFIM
log(Eigenvalues)
14
12
10
8
6
4
2
0
-2
-4
012
k1
k3
k4
k5
x0
v1
v2
sigma1
sigma2
12.6788
10.4756
9.22314
8.00637
8.00637
7.15434
6.45941
6.26567
2.49124
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8

Normalised Inverse of FIM

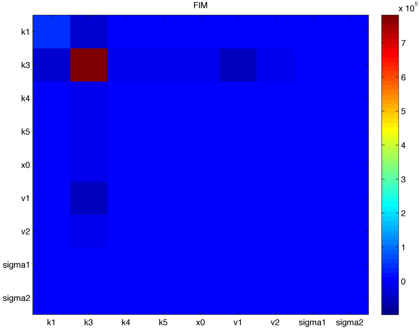


Expected FIM

Generative

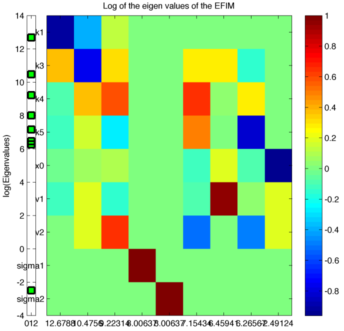


Modified k_3

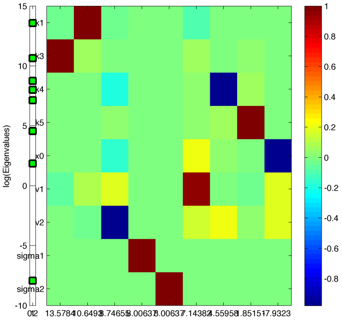


Eigenvalues of Expected FIM

Generative

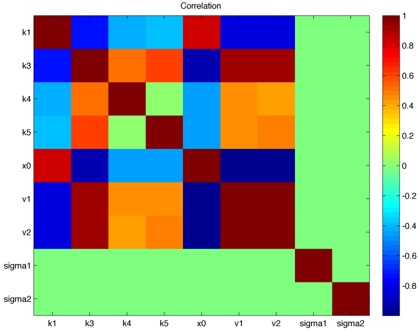


Modified k_3

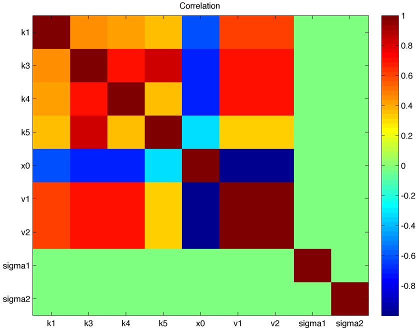


Normalised Inverse of FIM

Generative



Modified k_3



We are here ⟶ •

1. Nonlinear Dynamic Systems: Why We Need Manifold MCMC
   - Specificities of Nonlinear Dynamic Systems
   - Expected FIM and Sensitivities
   - Toy Example: Exponential Decay
   - Systems Biology: JAK-STAT Signaling Pathway

2. Manifold MCMC: a Unifying Framework
   - Adaptive Metropolis Hastings
   - Limitations and Link with Manifold MCMC
   - Conclusion and Future Work

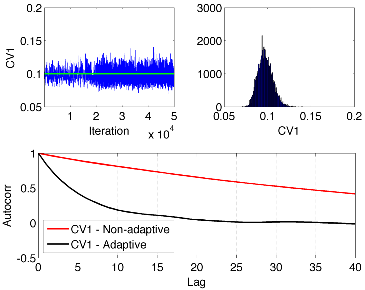
Most important practical innovation in MCMC in 2000s: **Adaptive MCMC**

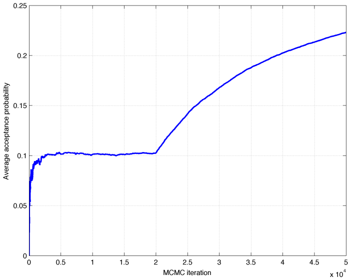
Optimality result for Gaussian target distribution

In Random-Walk Metropolis Hastings

- Optimal proposal covariance = covariance of the target
- Use ergodic estimate of covariance, based on past trajectory, as proposal
- Used as guideline for all cases

And it can work great:





To the Next Step: Link with Manifold

From

$$\theta^* = \theta^{k-1} + \varepsilon\sqrt{\mathbf{M}}z^*$$

to

$$\theta^* = \theta^{k-1} + \varepsilon\sqrt{\widehat{\mathsf{Cov}}(\theta^{0:k-1})}z^*$$

Adaptive MH works great when

- Started near the main mode ...
- ...or long annealing before estimating covariance
- Moderate variation in the curvature

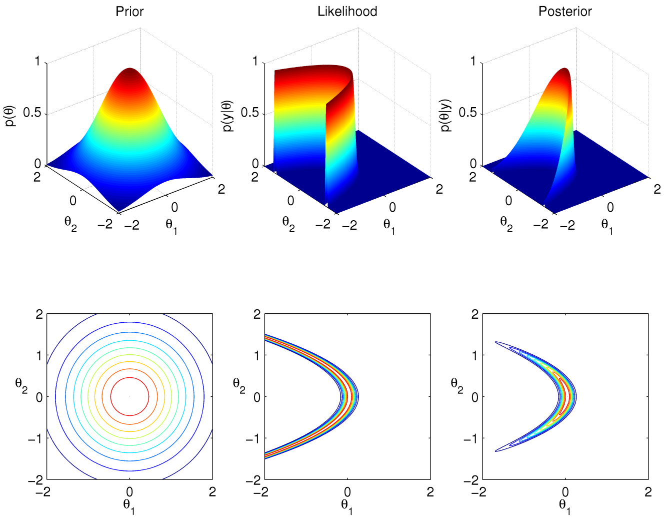
Adaptive MH is Crudely Approximating the Metric

- Constant over the whole space
- or ad-hoc partition of space (RAMA)
- Based on local starting point

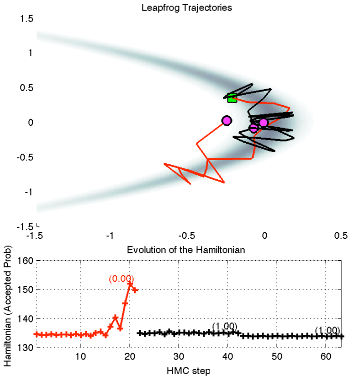
Manifold MCMC gives a geometric rationale, and a sound generalization.

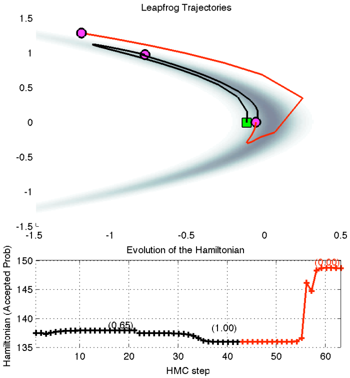
$$\theta^* = \theta^{k-1} + \varepsilon\sqrt{\mathbf{G}^{-1}(\theta^{k-1})}z^*$$

Typical Varying Metric Case



Global (Left, HMC) vs Local (Right, RMHMC)





Take-Home Message

1. Challenging new ODE problems have intricate geometry
2. Information Geometry now used in practical algorithms. . .
3. which unify, justify, and extend cutting-edge developments

Future work / Question to the audience:

- Intrinsic geometry of ODEs: chaos, bifurcation – how to exploit/deal with?
- Observed FIM vs Expected FIM (especially in MMALA/Natural Gradient)?
- Proxy to Observed FIM (when no sufficient statistic)
- Combine adaptation (stochastic approximation) to fit said proxy
- Adaptation of discretization stepsize
- Extension to Sequential Monte Carlo proposal kernel
- Use for Approximate Bayesian Computation Summary Statistics

Thank you for your attention!

julien@stats.ucl.ac.uk