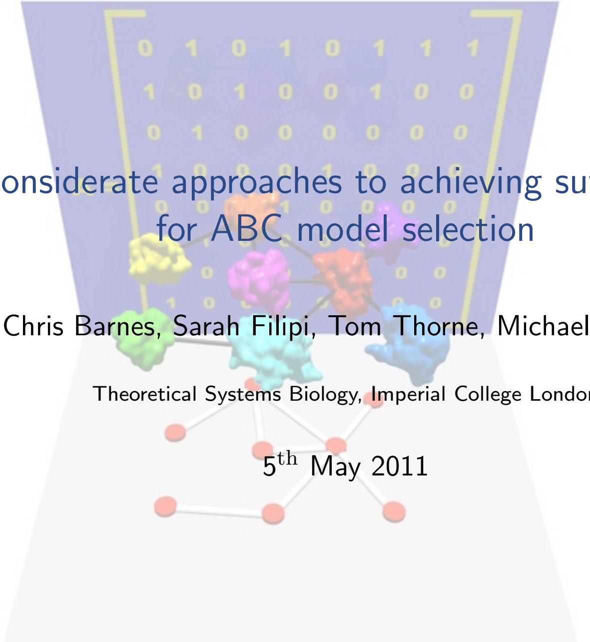
ePrecedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May



Considerate approaches to achieving sufficiency for ABC model selection

Chris Barnes, Sarah Filipi, Tom Thorne, Michael Stumpf

Theoretical Systems Biology, Imperial College London

5$^{\text{th}}$ May 2011

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Sufficient statistics

The *Likelihood principle* states that all the information about parameter θ is contained in the likelihood function $f(x|\theta)$. This principle is complemented by the *sufficiency principle*. Here a summary statistic of the general form

$$S : \mathbb{R}^D \longrightarrow \mathbb{R}^w, S(x) = s$$

with $w \ll D$ typically, is called sufficient if

$$f(X|S(x) = s, \theta) = f(x|S(x) = s)$$

ie the likelihood is independent of the parameter conditional on the value of the summary statistic. The likelihood can then generally be written in the Neyman-Fisher factorized form

$$f(X|\theta) = g(X)h(S(X)|\theta)$$

where $g(X)$ is independent of the parameter θ. Thus $h(S(X)|\theta)$ carries all the information about the parameter

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

ABC, sufficient statistics and model selection

Consider a finite set of models $\mathcal{M} = \{M_1, \ldots, M_q\}$, each of which has an associated parameter vector θ_m, $1 \leq m \leq q$. We aim to perform inference on the *joint space* over models and parameters, (m, θ_m).

$$p(M = m|x) = \frac{\int_{\Theta_m} f(x|\theta_m)\pi(\theta_m)d\theta_m\pi(m)}{\sum_{i=1}^{q}\int_{\Theta_i} p(x|\theta_i)\pi(\theta_i)d\theta_i\pi(i)}.$$

We can apply ABC by replacing evaluation of the likelihood in favour of comparing simulated and real data for different parameters drawn from the posterior, whence we obtain

$$p(M = m|x) \approx \frac{\int_{\Theta_m}\int_{\Omega} \mathbb{1}(\Delta(x,y) \leq \epsilon) f(y|\theta_m)\pi(\theta_m)d\theta_m dy\ \pi(m)}{\sum_{i=1}^{q}\int_{\Theta_i}\int_{\Omega} \mathbb{1}(\Delta(x,y) \leq \epsilon) f(y|\theta_i)\pi(\theta_i)d\theta_i dy\ \pi(i)},$$

which is exact once $\epsilon \longrightarrow 0$.

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

ABC, sufficient statistics and model selection (2)

The same is no longer true, however, once the complete data have been replaced by summary statistics. So in general

$$p(M=m|x) \neq$$
$$\frac{\int_{\Theta_m}\int_{\Omega} \mathbb{1}(\Delta(S_m(x), S_m(y) \leq \epsilon) h(S_m(y)|\theta_m)\pi(\theta_m) d\theta_m dy \ \pi(m)}{\sum_{i=1}^{q}\int_{\Theta_i}\int_{\Omega} \mathbb{1}(\Delta(S_i(x), S_i(y)) \leq \epsilon) h(S_i(y)|\theta_i)\pi(\theta_i) d\theta_i dy \ \pi(i)}.$$

An equality can only hold if the factors $g_i(x)$, $1 \leq i \leq q$ are all identical. Otherwise the different levels of data-compression are lost and unbiased model selection is no longer possible.

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Reviving ABC model selection

While we do agree that problems arise when using inadequate (or *insufficient*) statistics for model selection

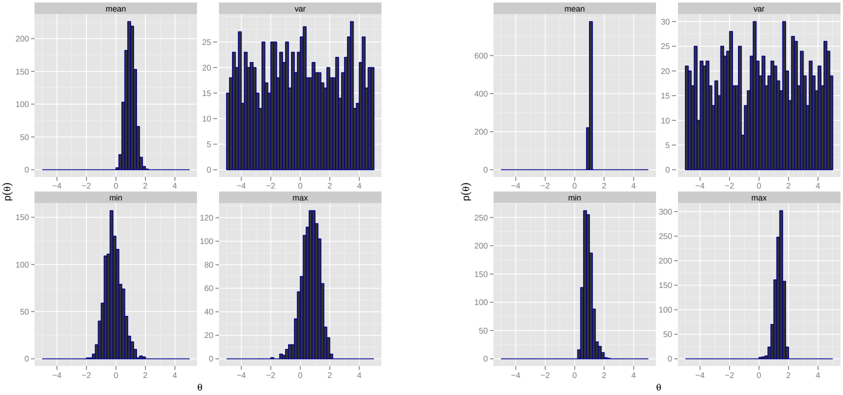
- this mirrors problems that can also be observed in the parameter estimation context.

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Reviving ABC model selection

ABC parameter inference for θ where $\mathbf{y}_{1,\dots,m} \sim N(\theta = 1, 1)$
Use as summary statistics: mean, variance, min and max.

$m = 10$ $\qquad\qquad$ $m = 10000$



Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Reviving ABC model selection

While we do agree that problems arise when using inadequate (or *insufficient*) statistics for model selection

- this mirrors problems that can also be observed in the parameter estimation context.
- for many important applications of ABC this problem can be elegantly avoided by using the whole data rather than summary statistics.

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Reviving ABC model selection

While we do agree that problems arise when using inadequate (or *insufficient*) statistics for model selection

- this mirrors problems that can also be observed in the parameter estimation context.
- for many important applications of ABC this problem can be elegantly avoided by using the whole data rather than summary statistics.
- in cases where summary statistics are required we argue that we can construct approximately sufficient statistics in a disciplined manner using notions from information theory.

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Entropy, conditional entropy and mutual information

The entropy of X, denoted by H, measures the uncertainty of X and is defined as follows

$$H(X) = -\sum_{x} p(x) \log p(x) = -E_{p(X)} [\log p(X)] \geq 0 \ .$$

The conditional entropy $H(Y|X)$ is defined as

$$H(Y|X) = -E_{p(X,Y)} [\log p(Y|X)] \ .$$

The mutual information $I(X;Y)$ measures the amount of information that Y contains about X. It can be seen as the reduction of the uncertainty of X due to the knowledge of Y :

$$\begin{aligned} I(X;Y) = H(X) - H(X|Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= KL(p(X,Y)||p(X)p(Y)) \geq 0 \ . \end{aligned}$$

$I(X;Y) = 0$ if and only if X and Y are independent.

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Data processing inequality and sufficient statistics

The DPE states that for random variables X, Y, and Z such that $X \rightarrow Y \rightarrow Z$, (i.e. Y depends, deterministically or randomly, on X and Z depends on Y)

$$I(X;Y) \geq I(X;Z),$$

with equality only if $X \rightarrow Y \rightarrow Z$ forms a Markov Chain which means that $p(X,Z|Y) = p(X|Y)p(Z|Y)$.
Now consider a family of distributions $\{f_\theta(.)\}$ indexed via θ and let X be a sample from a distribution in this family. Let S be a deterministic statistic of X then $\theta \rightarrow X \rightarrow S$. By the DPE

$$I(\theta;S) \leq I(\theta;X) \; .$$

A statistic S is said to be *sufficient with underlying parameter* θ if and only if S contains all the information in X about θ that is

$$I(\theta;S) = I(\theta;X) \; .$$

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Results for sufficient statistics (1)

The conditional mutual information of discrete random variables X, Y and Z is defined as

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \, .$$

It is the reduction in uncertainty of X due to knowledge of Y when Z is given. This quantity is null if and only if X and Y are conditionally independent given Z, which means that Z contains all the information about X in Y.

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Results for sufficient statistics (1)

The conditional mutual information of discrete random variables X, Y and Z is defined as

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \; .$$

It is the reduction in uncertainty of X due to knowledge of Y when Z is given. This quantity is null if and only if X and Y are conditionally independent given Z, which means that Z contains all the information about X in Y.

Result 1

S is a sufficient statistic with underlying parameter θ if and only if

$$I(\theta;X|S) = 0 \; .$$

This states that conditional on S there is no further information in X on θ.

Results for sufficient statistics (2)

Suppose that we have a finite set of deterministic statistics $S = \{S_1, \ldots, S_n\}$ and assume that S is a sufficient statistic. We aim to identify a subset U of S which is sufficient for θ. The following result characterizes such a subset.

Result 2

Let S be a finite set of deterministic statistics of X and assume that S is a sufficient statistic. Let U be a vector composed of elements of S. The following statements hold

$$
\begin{aligned}
& U \text{ is a sufficient statistic} \\
\Leftrightarrow\quad & I(\theta; S|U) = 0 \\
\Leftrightarrow\quad & E_{p(X)}\left[KL(p(\theta|S)||p(\theta|U)) = 0\right] \ .
\end{aligned}
$$

ePrecedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

e Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Algorithm 1: Minimization of $I(\theta; X|S)$

input: a sufficient set of deterministic statistics whose values on dataset is $s^* = \{s_1^*, \dots, s_n^*\}$
output: a subset U^* of s^*
for all $u^* \subset s^*$ **do**
    perform ABC to obtain $\hat{p}(\theta|u^*)$
end for
let $T^* = \{u^* \subset s^* \text{ such that } KL\left(\hat{p}(\theta|s^*)||\hat{p}(\theta|u^*)\right) = 0\}$
return $U^* = \text{argmin}_{u^* \in T^*} |u^*|$

Algorithm 2: Greedy minimization of $I(\theta; X|S)$
Algorithm 3: Stochastic minimization of $I(\theta; X|S)$

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Addition of statistics in stochastic minimization

In practice we can use different measures. We add the statistic $s^*_{(k)}$

- if $KL(p(\theta|s^*_{(1)},\dots,s^*_{(k)})||p(\theta|s^*_{(1)},\dots,s^*_{(k-1)})) \geq \delta_k$ where δ_k is a threshold (which could in theory be computed by bootstrapping the data)
- if the Hellinger distance between $\hat{p}(\theta|s^*_{(1)},\dots,s^*_{(k)})$ and $\hat{p}(\theta|s^*_{(1)},\dots,s^*_{(k-1)})$ is larger than ϵ.

$$Hd(\hat{p}_1,\hat{p}_2) \leq \sqrt{\log(2)\frac{N}{n}\log\left(\frac{2N}{\delta}\right)}$$

  with probability $1-\delta$. We denote by n the size of the sample and N the number of the bins used to compute the empirical distributions.
- tests for independence (KS, Pearson) enable us to compare $p(\theta|s^*_{(1)},\dots,s^*_{(k-1)})$ and $p(\theta|s^*_{(1)},\dots,s^*_{(k)})$ and the statistic $s^*_{(k)}$ is added if the test has a significant p-value

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Relation to previous work

Joyce and Marjoram (2008)

Developed a notion of approximate sufficiency for parameter inference and a sequential algorithm to score statistics according to whether their inclusion will improve inference.

Nunes and Balding (2010)

Proposed a heuristic algorithm to minimise the entropy of the posterior wrt sets of summary statistics. Additionally proposed a second step where the posterior mean squared error is minimised over simulated datasets 'close' to the true data.

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Sufficiency for the joint space

Consider q models each with an associated set of parameters $\theta_i, i \in \{1, ..., q\}$. We aim to identify a set of sufficient statistics for model selection. Let M be a random variable taking value in $\{1, \ldots, q\}$.

A statistic, S, is sufficient for model selection if and only if it is sufficient for the joint space $\{M, \{\theta_i\}_{1 \leq i \leq q}\}$ i.e. $I(M, \theta_1, ..., \theta_q; X|S) = 0$.

Result 3

For all deterministic statistics S of X,

$$I(M, \theta_1, ..., \theta_q; X|S) = I(M; X|\theta_1, ..., \theta_q, S) + \sum_i I(\theta_i; X|S)$$

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Toy example: Model selection for normals with known variance

We have two models

$$\mathbf{y}_{M_1} \sim N(\theta, \sigma_1^2),\ \mathbf{y}_{M_2} \sim N(\theta, \sigma_2^2)$$

with $\sigma_1 = 0.3$ and $\sigma_2 = 0.6$.
We observe $\mathbf{y} = (y_1, ...y_{15})$ from $M_1(\theta = 0)$ and perform stochastic minimisation of $I(M, \theta_1, \theta_2; X|S)$ with five statistics:

$$S_1 = \bar{y},\ S_2 = \sum(y - \bar{y}),\ S_3 = \text{range}\, y,\ S_4 = \max y,\ S_5 \sim U(0, 2)$$

S_1 is sufficient for parameter estimation and the pair $\{S_1, S_2\}$ is sufficient for model selection.

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

An aside on weighted statistics and distance

In general the distributions of the statistics, $p(S_i|\theta = t)$, can vary by orders of magnitude. Thus they must be weighted appropriately when using Euclidean distance which is impossible *a priori*.

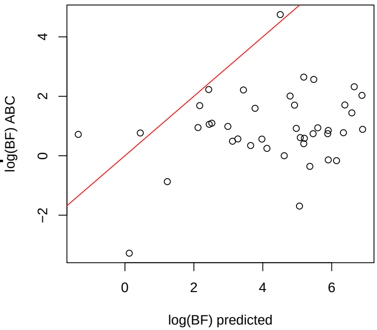
To circumvent this problem we use as a distance function:

$$\Delta(\mathbf{S}(\mathbf{x}), \mathbf{S}(\mathbf{y})) = \sum_i [\log(|S_i(\mathbf{x})|) - \log(|S_i(\mathbf{y})|)]^2$$
$$= \sum_i [\log(|S_i(\mathbf{x})|)/|S_i(\mathbf{y})|)]^2$$

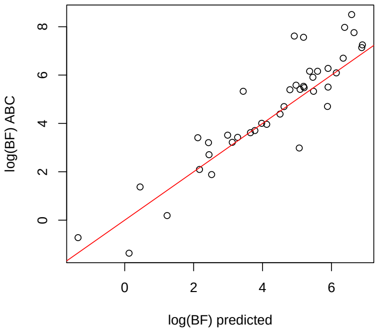
ePrecedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Results: Comparison to true Bayes factor

S st $I(\theta; X|S) = 0$

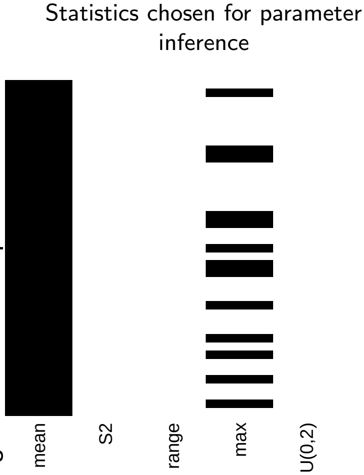


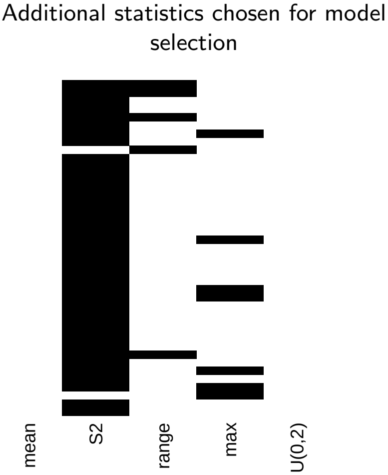
S st $I(M, \theta; X|S) = 0$



Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Results: Statistics chosen



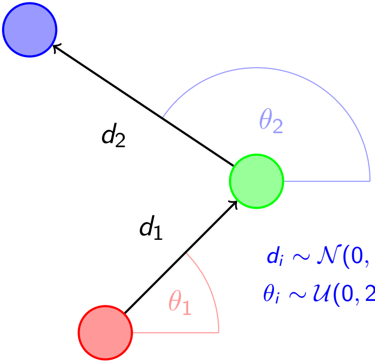


Imperial College London

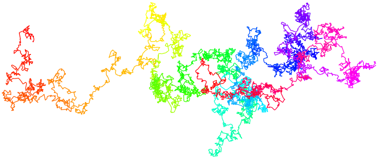
Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Models of random walks

Brownian motion

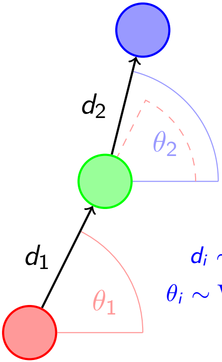


$$d_i \sim \mathcal{N}(0, \sigma)$$
$$\theta_i \sim \mathcal{U}(0, 2\pi)$$

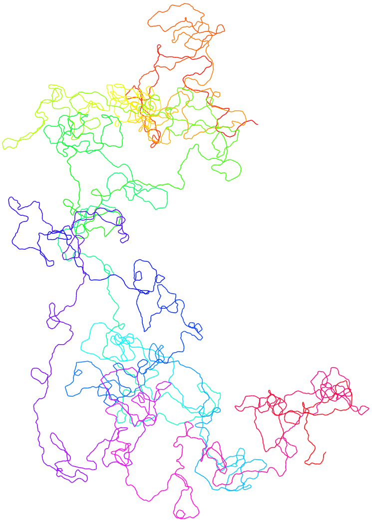


Models of random walks

Persistent random walk



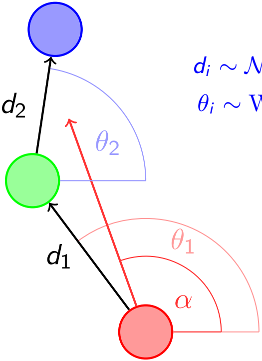
$$d_i \sim \mathcal{N}(0, \sigma)$$
$$\theta_i \sim \mathrm{WrapNorm}(\theta_{i-1}, \rho)$$



Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

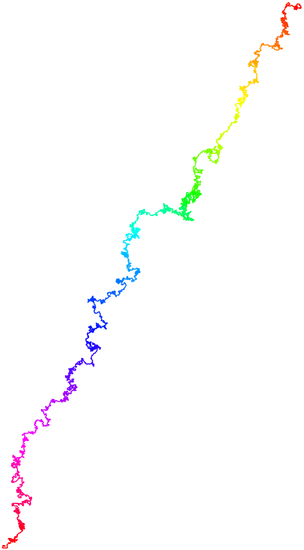
Models of random walks

Biased random walk



$d_i \sim \mathcal{N}(0, \sigma)$

$\theta_i \sim \mathrm{WrapNorm}(\alpha, \rho)$



e Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Summary statistics

- S_1: Mean square displacement

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Summary statistics

- S_1: Mean square displacement
- S_2: Mean x and y displacement

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

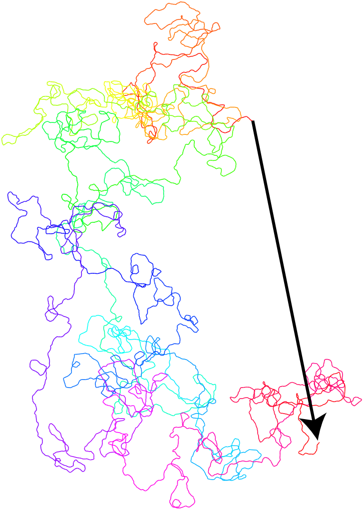
Summary statistics

- S_1: Mean square displacement
- S_2: Mean x and y displacement
- S_3: Mean square x and y displacement

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Summary statistics

- S_1: Mean square displacement
- S_2: Mean x and y displacement
- S_3: Mean square x and y displacement
- S_4: Straightness index $\frac{|\mathbf{u}(1) - \mathbf{u}(N)|}{\sum_i^N l_i}$

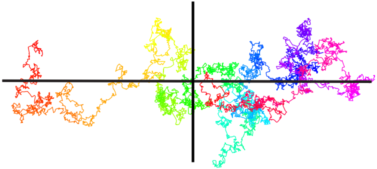


Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Summary statistics

- S_1: Mean square displacement
- S_2: Mean x and y displacement
- S_3: Mean square x and y displacement
- S_4: Straightness index $\frac{|\mathbf{u}(1) - \mathbf{u}(N)|}{\sum_i^N l_i}$
- S_5: Eigenvalues of gyration tensor

$$T_{kl} = \frac{1}{N} \sum_{j=1}^{N} (r_{jk} - < r_k >)(r_{jl} - < r_l >)$$

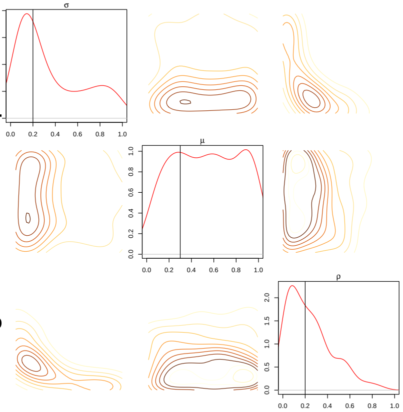


Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

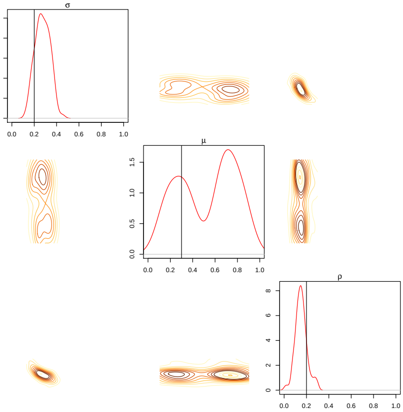
Results

S_5 (eigenvalues of gyration tensor) is consistently chosen as sufficient for Brownian and Persistent walks. The biased walk also requires S_3 (mean square x and y displacement) for sufficiency. The pair $\{S_3, S_5\}$ is also sufficient for the joint space.

Start: $S = \{S_1\}$



End: $S = \{S_3, S_5\}$



Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Conclusions

- Problems of sufficiency pervade both parameter inference and model selection problems.
- For any interesting real world problem there no simple sufficient statistics.
- Information theory allows a disciplined approach to the construction of sets of statistics that together can be (approximately) sufficient.
- We have shown that such an approach works in toy models. It is computationally feasible in more challenging problems.
- If we use the data rather than summary statistics ABC model selection is straightforward.

ePrecedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Thanks!

christopher.barnes@imperial.ac.uk

http://www3.imperial.ac.uk/theoreticalsystemsbiology

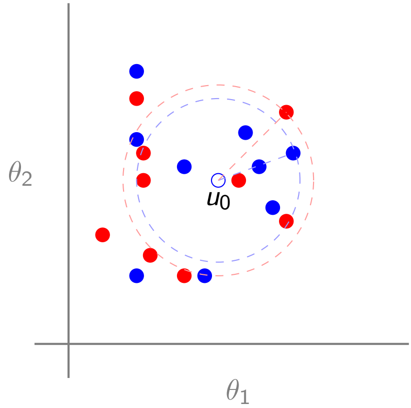
http://abc-sysbio.sourceforge.net/

http://cuda-sim.sourceforge.net/

bbsrc
biotechnology and biological sciences
research council

Nature Precedings : doi:10.1038/npre.2011.5952.1 : Posted 13 May

Calculating KL divergence for multivariate distributions



$$D_{KL(U,V)} \approx \log \frac{N_V}{N_U - 1} + d\mathrm{E}_U[\log \rho_k(\cdot, V)] - d\mathrm{E}_U[\log \rho_k(\cdot, U)]$$

kNN-based high-dimensional Kullback-Leibler distance for tracking, Boltz et al. WIAMIS'07