Sufficient Dimension Reduction for ABC

Marco Banterle

Supervised by C. Robert - Université Paris Dauphine

09 / 04 / 2014

Likelihood

Statistics in most of its flavors is based on the likelihood function

$$f\left(x|\theta\right)$$

Likelihood

Statistics in most of its flavors is based on the likelihood function

$$f(x|\theta)$$

Some will maximize it wrt θ,
some will combine it with prior information $\pi(\theta)$ into

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{m(x)}$$

Likelihood

Statistics in most of its flavors is based on the likelihood function

$$f(x|\theta)$$

Some will maximize it wrt θ,
some will combine it with prior information $\pi(\theta)$ into

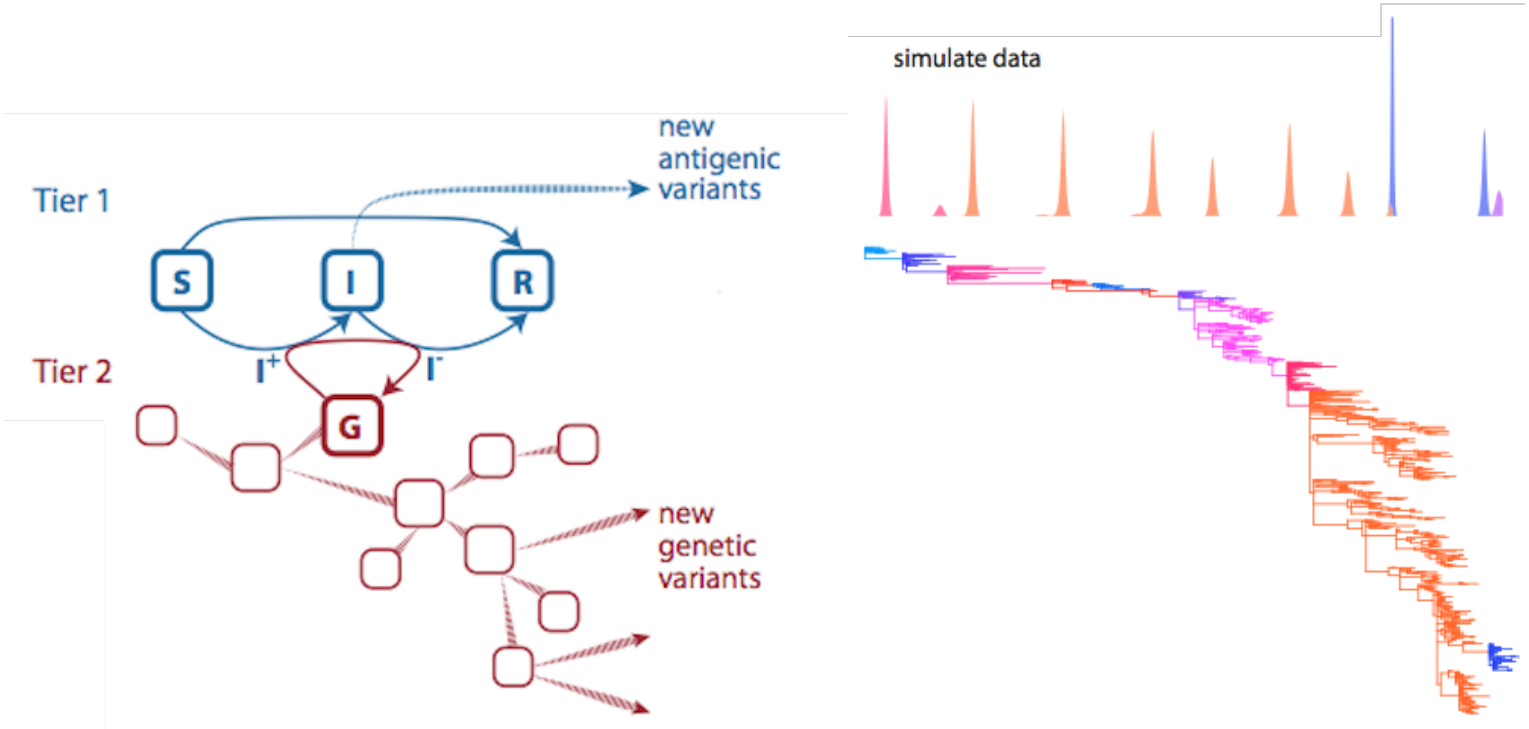
$$\pi(\theta|x) \propto \pi(\theta) f(x|\theta)$$

Likelihood

Statistics in most of its flavors is based on the likelihood function

$$f(x|\theta)$$

In complex models however



this could be impossible to compute for a number or reasons

Likelihood

Statistics in most of its flavors is based on the likelihood function

$$f(x|\theta)$$

In complex models however
this could be impossible to compute for a number or reasons

- Give up?
- Simplify our model?
- Approximate Inference!

Estimate $f(x|\theta_0)$?

Approximate the likelihood with a MC estimator like

$$\hat{f}(x|\theta_0) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}_{\{y_i = x\}}(y_i)$$

where y_i are simulated from $f(\cdot|\theta_0)$.

Estimate $f(x|\theta_0)$?

Approximate the likelihood with a MC estimator like

$$\hat{f}(x|\theta_0) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}_{\{y_i = x\}}(y_i)$$

where y_i are simulated from $f(\cdot|\theta_0)$.

Still this is a point-wise (very naive) estimation...
Bayesian statisticians are better with samplers!

Bayesian Computation

Importance Sampling

For a LOT of js

- Sample from the prior $\theta_j \sim \pi(\cdot)$
- Weight with $f(x|\theta_j)$

The resulting weighted sample has posterior law $\pi(\theta|x)$

Bayesian Computation

Importance Sampling

For a LOT of js

- Sample from the prior $\theta_j \sim \pi(\cdot)$
- Weight with $\hat{f}(x|\theta_j)$

The resulting weighted sample has posterior law $\pi(\theta|x)$?

not-yet-Approximate BC

For a LOT of js

- Sample from the prior $\theta_j \sim \pi(\cdot)$
- Weight with $\hat{f}(x|\theta_j)$

The resulting weighted sample has posterior law $\pi(\theta|x)$?

The answer is (surprisingly?) YES!

Even in the "limit" with $M = 1$ s.t. $\hat{f}(x|\theta) = \mathbb{I}_{\{y=x\}}(y)$!

$$\int_{\{y=x\}} \pi(\theta) f(y|\theta) dy = \pi(\theta) f(x|\theta) \propto \pi(\theta|x)$$

not-yet-Approximate BC

For a LOT of js

- Sample from the prior $\theta_j \sim \pi(\cdot)$
- Weight with $\hat{f}(x|\theta_j)$

The resulting weighted sample has posterior law $\pi(\theta|x)$?

The answer is (surprisingly?) YES!

Even in the "limit" with $M = 1$ s.t. $\hat{f}(x|\theta) = \mathbb{I}_{\{y=x\}}(y)$!

The "only" problem being the event
$\mathbb{I}_{\{y=x\}}(y)$ having null probability in general.

Approximate Bayesian Computation

Tavaré et al. (1997)

In the 90s some population geneticists decided
not to be let down from this triviality and introduced:

- a tolerance level ε s.t. $\hat{f}(x|\theta) = \mathbb{I}_{\{\rho(S(y),S(x))\leq\varepsilon\}}(y)$
- where $S(x)$ as a summary statistics of the data x

finally introducing the approximations we are interested in!

ABC

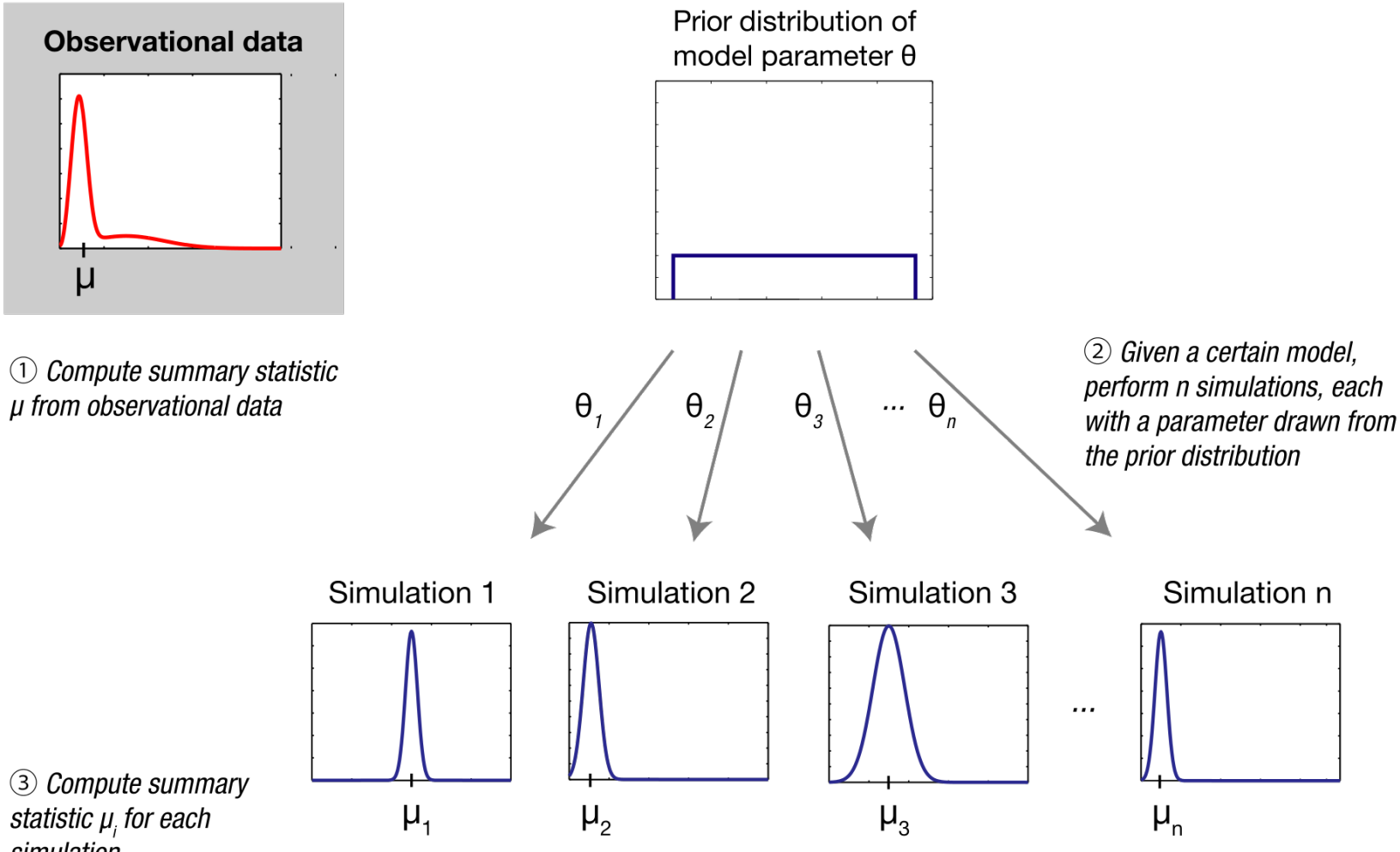
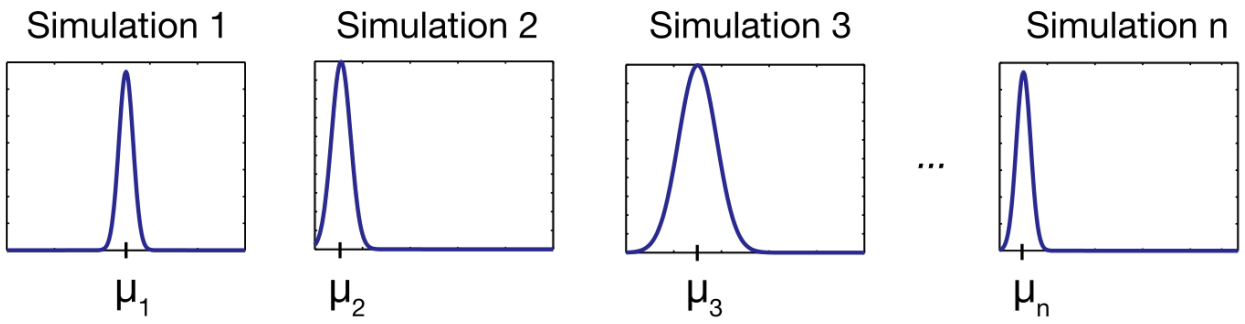


Image taken from www.wikipedia.com

ABC

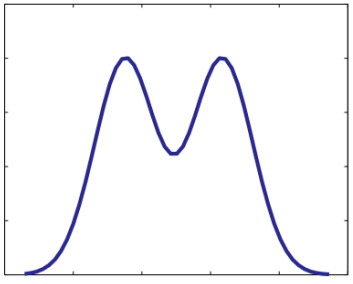


③ *Compute summary statistic μ_i for each simulation*

$$\rho(\mu_i, \mu) \overset{?}{\leq} \varepsilon$$

④ *Based on a distance $\rho(\cdot,\cdot)$ and a tolerance ε, decide for each simulation whether its summary statistic is sufficiently close to that of the observed data.*

Posterior distribution of model parameter θ



⑤ *Approximate the posterior distribution of θ from the distribution of parameter values θ_i associated with accepted simulations.*

Image taken from www.wikipedia.com

Trade-Off

$$S(\cdot) \longleftarrow - \longrightarrow \varepsilon$$

In an efficient sampler:

- $S(\cdot)$ needs to be high dimensional to retain informations
- this push ε to be fairly large to retain some of the samples

Trade-Off

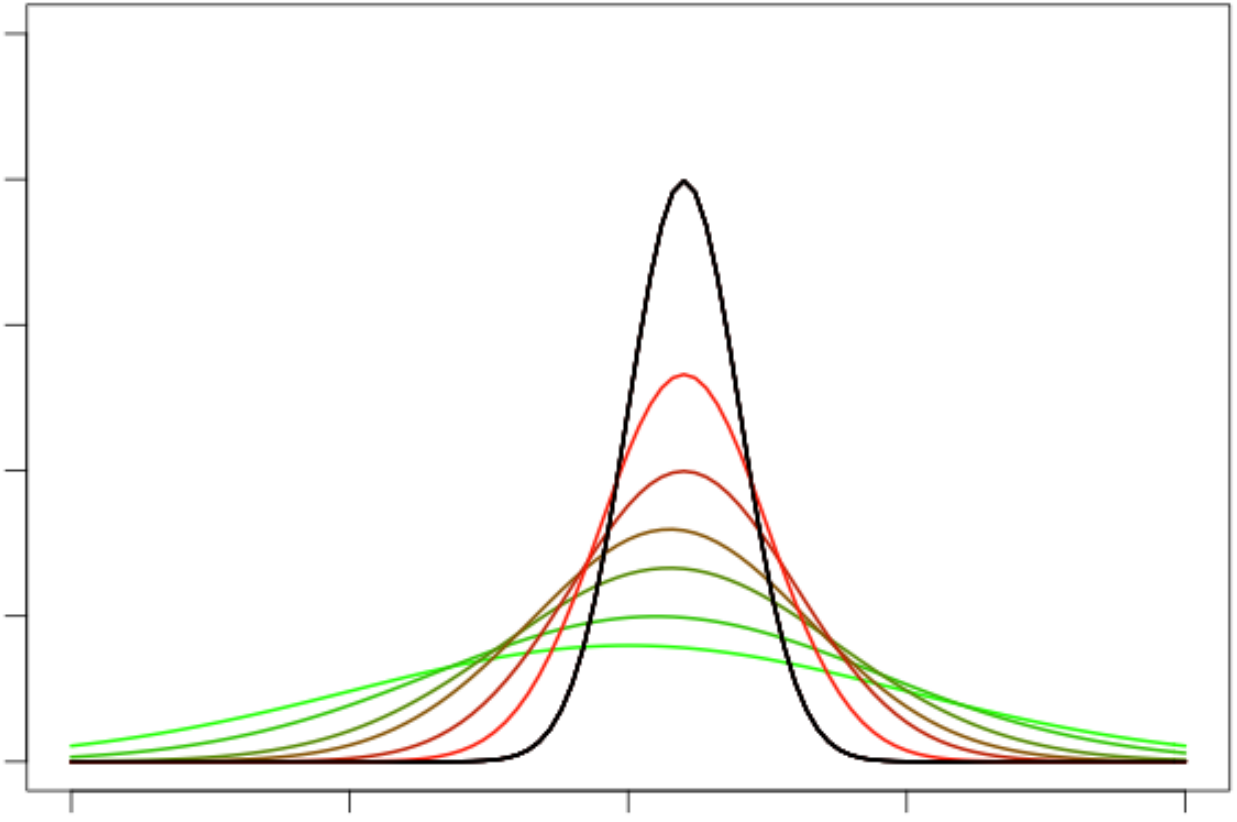
$$S(\cdot) \longleftarrow - \longrightarrow \varepsilon$$

In an efficient sampler:

- we want $\varepsilon \to 0$ to be not quite far from the truth
- $S(\cdot)$ needs to be low dimensional and hence we're going to lose some information

Approximating better

The problems associated with ε are being currently talked by relying on more efficient samplers, especially sequential ones.



Approximating better

The problems associated with ε are being currently talked by relying on more efficient samplers, especially sequential ones.

When the high complexity of the data prevents comparison directly between raw data, choosing the correct set of statistics is on the other hand still an open and highly debated subject.

Sufficient Statistics

While a sufficient statistics clearly solves our problems

$$\pi(\theta|x) = \pi(\theta|S(x))$$

Sufficiency is *HARD* to obtain and test.

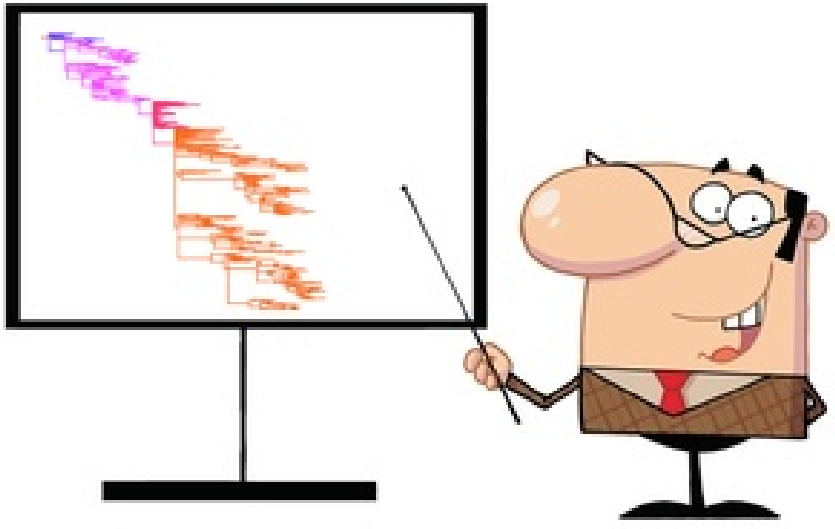
(outside the exponential family)

Sufficient Statistics

While a sufficient statistics clearly solves our problems

$$\pi(\theta|x) = \pi(\theta|S(x))$$

Sufficiency is *HARD* to obtain and test.



Typically an expert of the field chooses some statistics which are *likely* to contain most of the informations about the given data.

Sufficient Statistics

While a sufficient statistics clearly solves our problems

$$\pi(\theta|x) = \pi(\theta|S(x))$$

Sufficiency is *HARD* to obtain and test.

Typically an expert of the field chooses some statistics which are likely to contain most of the information about the given data but could be high dimensional, redundant and still un-sufficient.

(bref: not an easy task!)

Dimension Reduction

The better technique is to choose the set conservatively and then trim it down to the most efficient subset.

Dimension Reduction

Let s be the set of summaries and define $u \subseteq s$ the minimal subset that it contains the *whole information* in s

Dimension Reduction

Let s be the set of summaries and define $u \subseteq s$ the minimal subset that it contains the *whole information* in s

Information is sadly not a universal concept..

State of the Art

Some efforts in this direction are described in Blum et al. (2013):

- Best subset selection techniques, which select a subset u based on some criterion
- Projection techniques that aim at reducing the dimension by combining statistics into a new (orthogonal) set in order to reduce collinearity, like PCA or PLS.

State of the Art

Some efforts in this direction are described in Blum et al. (2013):

- Best subset selection techniques, which select a subset u based on some criterion
- Projection techniques that aim at reducing the dimension by combining statistics into a new (orthogonal) set in order to reduce collinearity, like PCA or PLS.

Both these methods have their drawbacks,
BSS are generally based on an *arbitrary* criterion while
PLS and PCA rely on searching just for *linear* relations between variables.

Let's combine them

Define information as Thomas would:

Let's define u s.t.

$$\pi(\theta|u) = \pi(\theta|s)$$

Let's combine them

Define information as Thomas would:

$$\pi(\theta|u) = \pi(\theta|s) \leftarrow \pi(\theta|u) \perp\!\!\!\perp p(s \setminus u)$$

So PLS isn't a bad idea! But still relies on the normal assumption.

RKHS

Zhang et al. (2012) and Fukumizu et al. (2008) formally derived a *conditional independence test statistics* along with its *asymptotic* distribution under the null using the conditional cross-covariance operator that lies in the RKHS induced by a (characteristic, usually guassian) kernel.

RKHS

Zhang et al. (2012) and Fukumizu et al. (2008) formally derived a *conditional independence test statistics* along with its *asymptotic* distribution under the null using the conditional cross-covariance operator that lies in the RKHS induced by a (characteristic, usually guassian) kernel.

$$H_0 : \pi(\theta|u) \perp\!\!\!\perp p(s \setminus u)$$

s and θ are sampled for ABC!

About the idea

The cross-covariance operator $\Sigma_{Y,X}$ on the RKHS from $\mathcal{H}_X$ to $\mathcal{H}_Y$ is defined by:

$$< g, \Sigma_{Y,X} f >= \mathbb{E}_{X,Y}[f(X)g(Y)] - \mathbb{E}_X[f(X)]\mathbb{E}_Y[g(Y)]$$

for all $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$ and the conditional cross-covariance operator of (X, Y) given Z is then defined as:

$$\Sigma_{X,Y\,|\,Z} = \Sigma_{Y,X} - \Sigma_{Y,Z}\Sigma_{Z,Z}^{-1}\Sigma_{Z,X}$$

The idea is that checking for correlation of functions in the RKHS translate in testing for non-linear correlation (dependence) of the conditional distributions in the original space and hence provides more robust result with respect to PLS.

Computational Burden

The operation involved are dominated by matrix inversions which complexity is $\mathcal{O}(n^3)$
n being the size of the sample used for testing

Proceeding incrementally for single component of θ might lead quick to acceptance and
Greedy procedures have also been developed by minimizing $Tr(\Sigma_{Y|Z})$!

Results in IID

We tested the procedure on IID examples: $\mathcal{P}ois(\lambda)$ and $\mathcal{N}(\mu, \sigma^2)$
$s = (\mu^1, \mu^2, \mu^2, \mu^4, \min, \max, q_{0.25}, q_{0.5}, q_{0.75}, \mathcal{N}(0,1), \mathcal{T}_3)$
(permuted at each repetition)

We repeated 100 times for each model and successfully recovered the sufficient statistics in over 90% of the trials, using a random subsampling with size $B = 800$. In the remaining 10% of the cases the statistics either was often composed of the sufficient set plus a (few) other statistics.

Testing for (unconditioned) independence further reduced the size of u prior to the intensive analysis by removing in 100% of the cases the two random ancillary vectors.

Results in Drug Resistent Tubercolosis

Lastly we examined a Markov processes for epidemiological modeling with 4 parameters and 11 statistics.

Replicated information in the form of (often non-linear) dependence between some statistics is expected. In the original work the authors show in fact that *PSL is outperformed by the non-linear NNet*, that the best performing methods are among the *best subset techniques* and notably that **every** dimension reduction method result in a lower mean RSSE that using the whole s set.

Results in DRT

$$RSSE = \sum_{j=1}^{N} \left(||\theta_j, \theta_{true}||^2 \right)^{1/2}$$

The relative gain in mean RSSE is 20%, as good as the best performing AIC/BIC best subset selection method in Blum et al. (2013).

Results in DRT

$$mean\,RSSE = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{N}\left(||\theta_j, \theta_i||^2\right)^{1/2}$$

The relative gain in mean RSSE is 20%, as good as the best performing AIC/BIC best subset selection method in Blum et al. (2013).

Results in DRT

The relative gain in mean RSSE is 20%, as good as the best performing AIC/BIC best subset selection method in Blum et al. (2013).

"Why bother then?" you may ask, but I assure you AIC/BIC have been proven weak in other counter-examples.

Conclusion and Future Perspectives

We derived a BSS method based on widely accepted notions of conditional independence and "Bayesian sufficiency" which is general enough to have almost no assumptions. The procedure is shown to work well on both synthetic and real data.

We are also investigating further properties of ABC, including:

- How does reducing the dimension of s impact regression adjustment? Can regression **substitute** rejection/weighting? Note that out-of-sample problems here does not (practically) apply!
- Is regression equivalently inducing some ellipsoidal rather than spherical distance? More importantly, is non-linear regression inducing *blobs* around $S(x)$?

Thank you for your attention!