Combining Synthetic Likelihood and Variational Bayes for Performing High Dimensional Likelihood-Free Bayesian Inference

Dr Christopher Drovandi

School of Mathematical Sciences
ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)
Queensland University of Technology

Collaborators: Victor Ong, David Nott, Minh-Ngoc Tran, Scott Sisson

Likelihood-Free Methods

Here we are interested in models where the likelihood is intractable, but simulation of data x from the model is feasible.

Likelihood-free methods simulate data and compare x with y based on some data summary $S(\cdot)$.

Bayesian methods target $p(\theta|s_y) \propto p(s_y|\theta)p(\theta)$ where $s_y = S(y)$.

Approximate Bayesian Computation

Approximate Bayesian computation (ABC, e.g. Sisson and Fan (2010)) is current state-of-the-art likelihood-free Bayesian method.

Choice of summary function $S(\cdot)$ trade-off between information loss and dimensionality.

Compares s_y to s_x non-parametrically (Blum 2010):

$$\hat{p}_\epsilon(s_y|\theta) = \frac{1}{n}\sum_{i=1}^{n} K_\epsilon\left(\rho(s_y, s_i)\right).$$

where $s_i \sim p(s_y|\theta)$, ρ is distance function, K_ϵ is kernel weighting function with bandwidth ϵ.

Approximate Bayesian Computation

Disadvantages

- Highly sensitive to choice of tuning parameter ϵ, $\rho(\cdot)$ and to a lesser extent $K_\epsilon(\cdot)$
- No standard way to select ϵ or $\rho(\cdot)$.
- Suffers from curse of dimensionality with respect to size of summary statistic

Bayesian Synthetic Likelihood

The synthetic likelihood (SL) method of Wood (2010) uses a multivariate normal approximation: $p(s_y|\theta) \approx \mathcal{N}(s_y; \mu(\theta), \Sigma(\theta))$.

- Suitable when summary statistics are subject to the central limit theorem
- Transformations to multivariate normality of summary statistics
- Summary statistics from indirect inference
- Popular & convenient choice

The BSL replacement likelihood is

$$\mathcal{N}(s_y; \mu_n(\theta), \Sigma_n(\theta)).$$

sample mean μ_n, sample covariance matrix Σ_n. Can put into MCMC.

Background

- We have seen that ABC does not scale with the dimension of the summary statistic
- Synthetic Likelihood (SL) can help with handling high dimensional statistic

But...

- SL remains simulation intensive
- MCMC versions of ABC and SL are not well equipped to handle high-dimensional parameter space

How can we handle high-dimensional summary and parameter in the likelihood-free setting? Here we propose to combine Variational Bayes methods with synthetic likelihood.

Variational approximation

Variational Bayes (VB) assumes that the posterior can be well approximated by a parameteric probability distribution (e.g. multivariate normal)

Denote the parametric approximation by $q_\lambda(\theta)$. VB sets out to minimize the KLD between $q_\lambda(\theta)$ and $p(\theta|y)$

$$\hat{\lambda} = \arg\min_{\lambda} KL(q_\lambda(\theta)||p(\theta|y))$$

where

$$KL(q_\lambda(\theta)||p(\theta|y)) = \int \log \frac{q_\lambda(\theta)}{p(\theta|y)} q_\lambda(\theta)\, d\theta$$

Variational approximation

By using Bayes' rule we can show that

$$KL(q_\lambda(\theta)||p(\theta|y)) = \log p(y) - \int \log \frac{p(\theta)p(y|\theta)}{q_\lambda(\theta)} q_\lambda(\theta)\, d\theta$$

If we set $\mathcal{L}(\lambda) = \int \log \frac{p(\theta)p(y|\theta)}{q_\lambda(\theta)} q_\lambda(\theta)\, d\theta$ we see that minimising $KL(q_\lambda(\theta)||p(\theta|y))$ is the same as maximising $\mathcal{L}(\lambda)$

$$\hat{\lambda} = \arg\max_\lambda \mathcal{L}(\lambda)$$

$\mathcal{L}(\lambda)$ is called the variational lower bound because of the way it lower bounds the log of the marginal likelihood.

Stochastic gradient variational Bayes

We will use stochastic gradient ascent methods for optimizing the lower bound.

Suppose that $\nabla_\lambda \mathcal{L}(\lambda)$ is the gradient of the objective $\mathcal{L}(\lambda)$ and that $\widehat{\nabla_\lambda \mathcal{L}(\lambda)}$ is an unbiased estimate of it.

Stochastic gradient ascent

- Initialize $\lambda^{(0)}$
- for $t = 0, 1, \ldots$ and until some stopping rule is satisfied

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho_t \widehat{\nabla_\lambda \mathcal{L}(\lambda^{(t)})}$$

Typically the learning rate sequence ρ_t, $t \geq 0$ satisfies the Robbins-Monro conditions (Robbins and Monro, 1951), $\sum_t \rho_t = \infty$, $\sum_t \rho_t^2 < \infty$.

Stochastic gradient variational Bayes

The lower bound is

$$\mathcal{L}(\lambda) = \int \log \frac{p(\theta)p(y|\theta)}{q_\lambda(\theta)} q_\lambda(\theta)\, d\theta.$$

Differentiating under the integral sign, and using $E(\nabla_\lambda \log q_\lambda(\theta)) = 0$ (the "log derivative trick") some algebra gives

$$\nabla_\lambda \mathcal{L}(\lambda) = \int \{\log p(\theta)p(y|\theta) - \log q_\lambda(\theta)\} \nabla_\lambda \log q_\lambda(\theta) q_\lambda(\theta)\, d\theta.$$

An expectation with respect to $q_\lambda(\theta)$. Unbiased estimator:

$$\frac{1}{S}\sum_{i=1}^{S} \nabla_\lambda \log q_\lambda(\theta^{(i)}) \left\{\log p(\theta^{(i)}) + \log p(y|\theta^{(i)}) - \log q_\lambda(\theta^{(i)})\right\}$$

This can be combined with unbiased estimation of the log likelihood itself (subsampling a large data set for example).

Likelihood free inference

How can we use these VB methods in the likelihood-free setting?

Tran et al 2015 use unbiased estimate of ABC likelihood (not log-likelihood) in stochastic gradient variational Bayes.

Ong et al 2016 use an unbiased estimator of log synthetic likelihood due to Ripley 1996 (call this VBSL)

$$\hat{l}_N^U(s|\theta) = -\frac{d}{2}\log 2\pi - \frac{1}{2}\left\{\log|\hat{\Sigma}(\theta)| + d\log\left(\frac{N-1}{2}\right) - \sum_{i=1}^{d}\psi\left(\frac{N-i}{2}\right)\right\}$$
$$- \frac{1}{2}\left\{\frac{N-d-2}{N-1}(s-\hat{\mu}(\theta))^T\hat{\Sigma}(\theta)^{-1}(s-\hat{\mu}(\theta)) - \frac{d}{N}\right\}$$

Handling High-Dimensional Statistic

Can we gain even more efficiency in the presence of a high-dimensional statistic?

We can improve the efficiency further by using shrinkage estimators of covariance matrix $\Sigma(\theta)$ (similar to An et al 2017). Here we use the estimator of Warton 2008:

$$\hat{\Sigma}_\gamma = \hat{D}^{1/2}(\gamma \hat{C} + (1-\gamma) I)\hat{D}^{1/2}$$

where $\hat{C} = \hat{D}^{-1/2}\hat{\Sigma}\hat{D}^{-1/2}$ (sample correlation matrix), $\hat{D}$ is the diagonal matrix with diagonal entries equal to those of $\hat{\Sigma}$ and λ is the shrinkage parameter (γ is chosen by cross validation and updated at various times during stochastic gradient optimisation).

Call this method VBSL_S.

Handling High-Dimensional Parameter

If θ had dimension 100 then the number of variational parameters would be $100 + (100 \times 101)/2 = 5150$.

We propose to parameterise the posterior covariance matrix using Factor Analysis (e.g. Bartholomew et al 2011)

$$\Sigma = BB^{\top} + D^2$$

where D is a diagonal matrix and B is a $p \times f$ matrix. If $f << p$ then significant reduction in dimension of λ.

Need to place some identifiability constraints on B (upper triangular elements of B are zero, $B_{ij} = 0$ if $j > i$).

Call the method that uses the shrinkage (summary) covariance estimator and factor analysis representation of posterior covariance, VBSL_{SP}

Multivariate *g*-and-*k* example

- The *g*-and-*k* distribution (eg Rayner and McGillivray 2002) is defined through its quantile function, $Q(p)$, $p \in (0,1)$ say, where writing $z(p) = \Phi^{-1}(p)$

$$Q(p) = A + B\left[1 + c\frac{1-\exp(-gz(p))}{1+\exp(-gz(p))}\right](1+z(p)^2)^k z(p).$$

- Parameters A, $B > 0$, g and $k > -0.5$ controlling respectively the location, scale, skewness and kurtosis. (c is conventionally fixed at 0.8).
- Simulation from the *g*-and-*k* model is easy, $Q(U)$ for $U \sim U[0,1]$. This makes ABC methods for inference attractive.

Multivariate g-and-k example

- Multivariate extension (Drovandi and Pettitt 2011).
- Independent and identically distributed multivariate observations $y_1, \ldots, y_n$ where $y_i = (y_{i1}, \ldots, y_{iq})^T$.
- Each y_{ir} follows a g-and-k distribution marginally, with parameters $\theta_r = (A_r, B_r, g_r, k_r)$, $r = 1, \ldots, q$.
- Dependence between components of y_i will be modelled using a Gaussian copula. The copula is specified through $q(q-1)/2$ correlation parameters.
- A q component model has $4q + q(q-1)/2$ parameters.

Multivariate *g*-and-*k* example

Summary Statistics

- Robust estimates of location, scale, skewness and kurtosis for each marginal (Drovandi and Pettitt 2011).
- Pairwise Gaussian rank correlations (eg Boudt et al 2012) between all components + Fisher transformation to improve normality.
- Effectively one summary statistic per parameter.

Multivariate *g*-and-*k* example

Data

- Data on foreign currency exchange log daily returns against the Australian dollar (AUD) for 200 trading days between January 2, 2007 and 19 October, 2007.
- We consider data for 5 foreign currencies, the US dollar (USD), Japanese Yen (JY), the Euro (EUR), the United Kingdom Pound sterling (UKPS) and the Swiss Franc (CHF)

Multivariate *g*-and-*k* Results

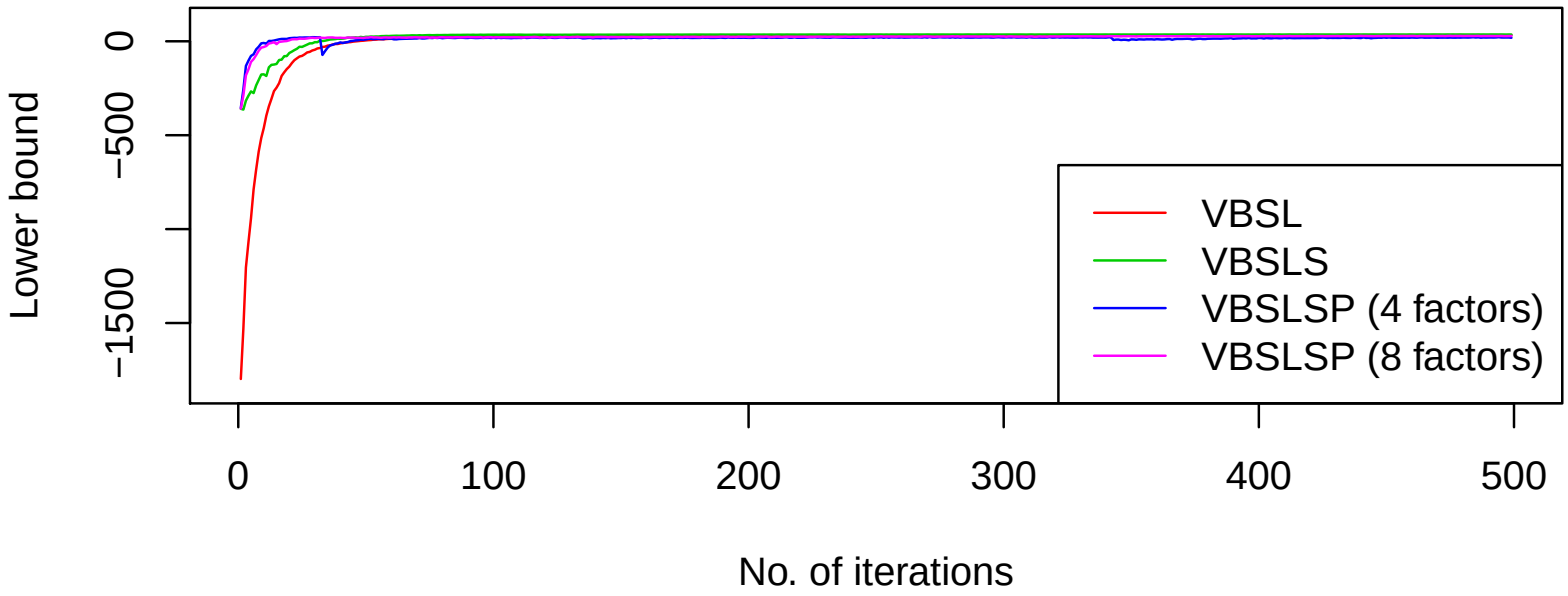
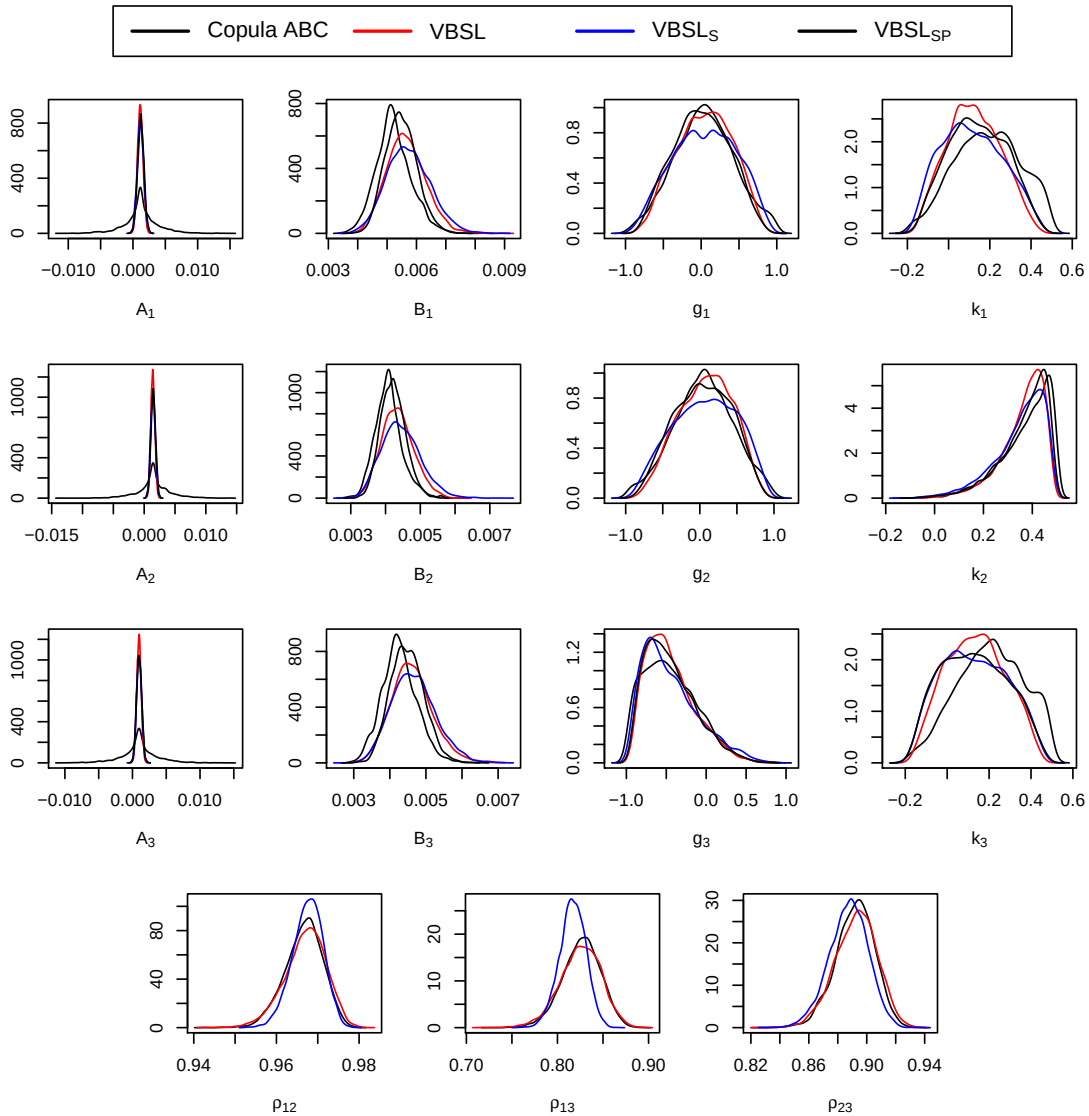
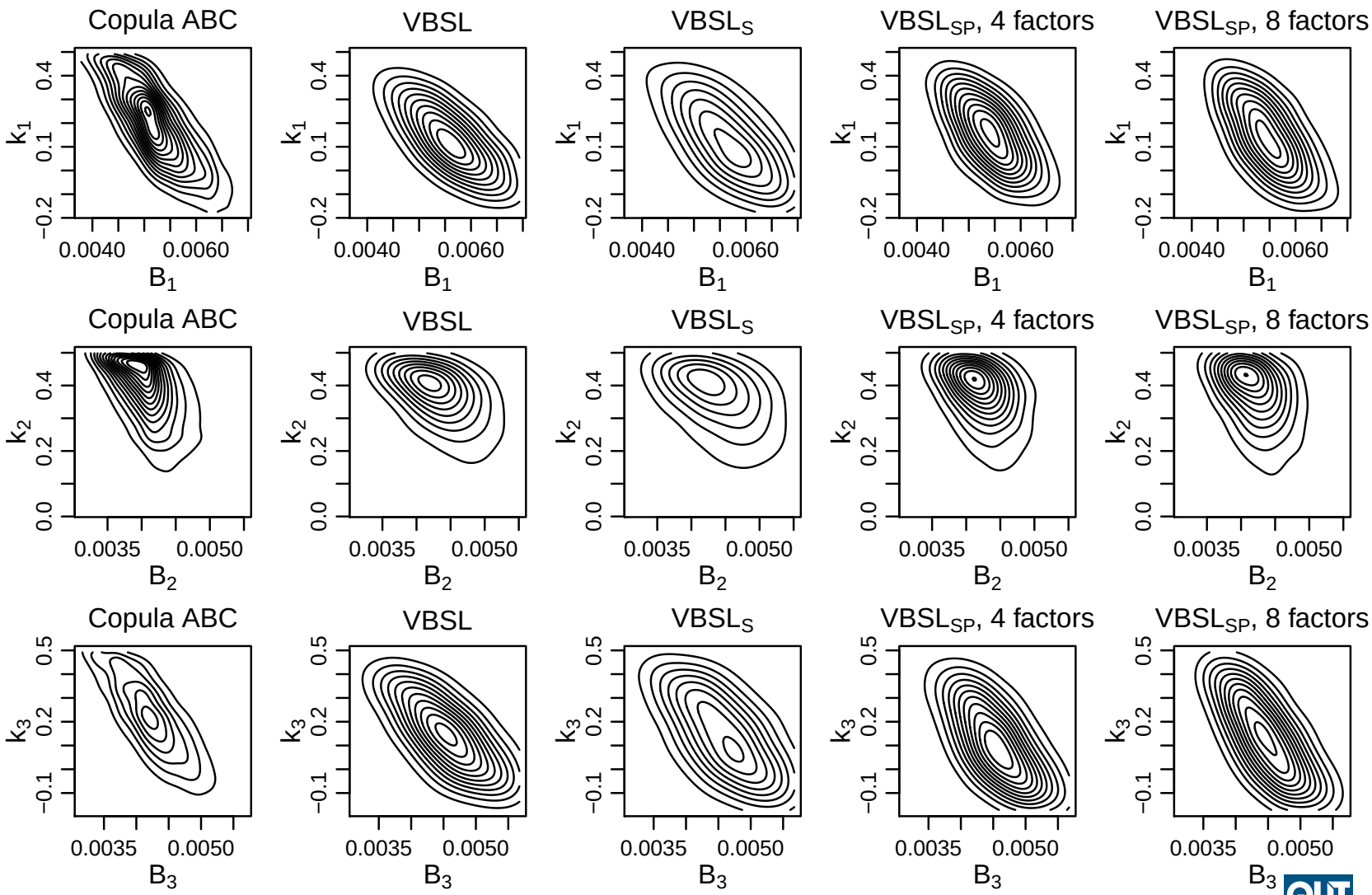


Figure: Variational lower bound for the multivariate *g*-and-*k* model $q = 6$ using the VBSL_{S} and VBSL_{SP} algorithms with $p = 4, 8$.

Multivariate *g*-and-*k* Results



Multivariate *g*-and-*k* Results



Stochastic gradient tricks

Tran, Nott and Kohn, 2015, Ong *et al.*, in progress

- In the variational optimization algorithm it is important to reduce the variance of gradient estimates as much as possible.
  - Variance reduction: Control variates, randomised quasi-Monte Carlo.
  - Parts of the lower bound may be computable analytically (may or may not be beneficial).
  - Use *natural gradient* (Amari, 1998) rather than the ordinary gradient.
  - Reparametrization of the normal variational posterior covariance in terms of the Cholesky factor of the precision matrix.
  - Adaptive step size choices (Ranganath *et al.*, 2013).
- Although I won't talk much about them, these tricks are *very important*.

Future Work

Relaxing normality assumption of summary statistics and posterior approximation.

References

Tran, M.-N., Nott, D. J., and Kohn, R. (2016). Variational Bayes with intractable likelihood. arXiv:1503.08621v2.

Ong, V. M-H., Nott, D. J., Tran, M-N., Sisson, S. A. and Drovandi , C. C. (2016) Variational Bayes with Synthetic Likelihood. Submitted. arXiv:1608.03069.

Ong, V. M-H., Nott, D. J., Tran, M-N., Sisson, S. A. and Drovandi, C. C. (2017) Likelihood-Free Inference in High Dimensions with Synthetic Likelihood. In Preparation.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2017). Bayesian synthetic likelihood. To appear in Journal of Computational and Graphical Statistics.

An, Z., Nott, D. J. and Drovandi, C. C. (2017). Accelerating Bayesian synthetic likelihood with the graphical lasso. Submitted.
`https://eprints.qut.edu.au/102263/`

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. Nature, 466:1102-1107.

Group at QUT