Recent Advances in Bayesian Synthetic Likelihood

Dr Christopher Drovandi

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

Collaborators: Leah Price (South), Ziwen An, Victor Ong, David Nott, Minh-Ngoc Tran, Scott Sisson

Bayesian Statistics

In Bayesian statistics we are interested in sampling from the posterior:

$$p(\theta|y) \propto p(y|\theta)p(\theta),$$

where $p(y|\theta)$ is the likelihood and $p(\theta)$ is the prior.

Markov Chain Monte Carlo

Construct ergodic Markov chain with invariant distribution $p(\theta|y)$ (Metropolis et al., 1953)

A common MCMC algorithm is Metropolis Hastings (MH) MCMC, where proposals θ^* are accepted with probability

$$min\left(1, \frac{p(y|\theta^*)p(\theta^*)q(\theta|\theta^*)}{p(y|\theta)p(\theta)q(\theta^*|\theta)}\right),$$

where $q(\cdot)$ is the proposal density.

For complex models, an analytic form for $p(y|\theta)$ may not be available and/or may not be computable.

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 (2011)) is current state-of-the-art likelihood-free Bayesian method.

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

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

Approximate Bayesian Computation

ABC Approximation of likelihood $p(s_y|\theta)$

- Simulate n iid datasets, denoted $x_{1:n} = (x_1, ..., x_n)$, from the model based on θ.
- Calculate n sets of summary statistics, $s_{1:n} = (s_1, ..., s_n)$
- The intractable $p(s_y|\theta)$ is replaced with the estimated ABC likelihood,

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

  - $\rho(\cdot)$ is called the discrepancy function
  - $K_\epsilon(\cdot)$ is a kernel weighting function with bandwidth ϵ
  - ϵ is called the ABC tolerance (bias/variance trade-off)

Approximate Bayesian Computation

Disadvantages

- Highly sensitive to choice of tuning parameter ϵ, $\rho(\cdot)$ and to a lesser extent $K_\epsilon(\cdot)$ (Marin et al., 2012)
- No standard way to select ϵ or $\rho(\cdot)$.
- Suffers from curse of dimensionality with respect to size of summary statistic

Cell Biology Example - High Dimensional Summary Statistic

Cell motility and proliferation are important parts of many biological processes (e.g. skin cancer growth, wound healing).

One way to investigate this is through a scratch assay. A 'scratch' is made which separates the cells. Images of the cells are taken at regular time intervals until the cells are once again in contact.

Images taken every 5 minutes for 12 hours (145 images)

Approximately map cells onto a rectangular lattice (binary matrix where a 1 indicates presence of a cell a particular location).

Cell Biology Example - High Dimensional Summary Statistic

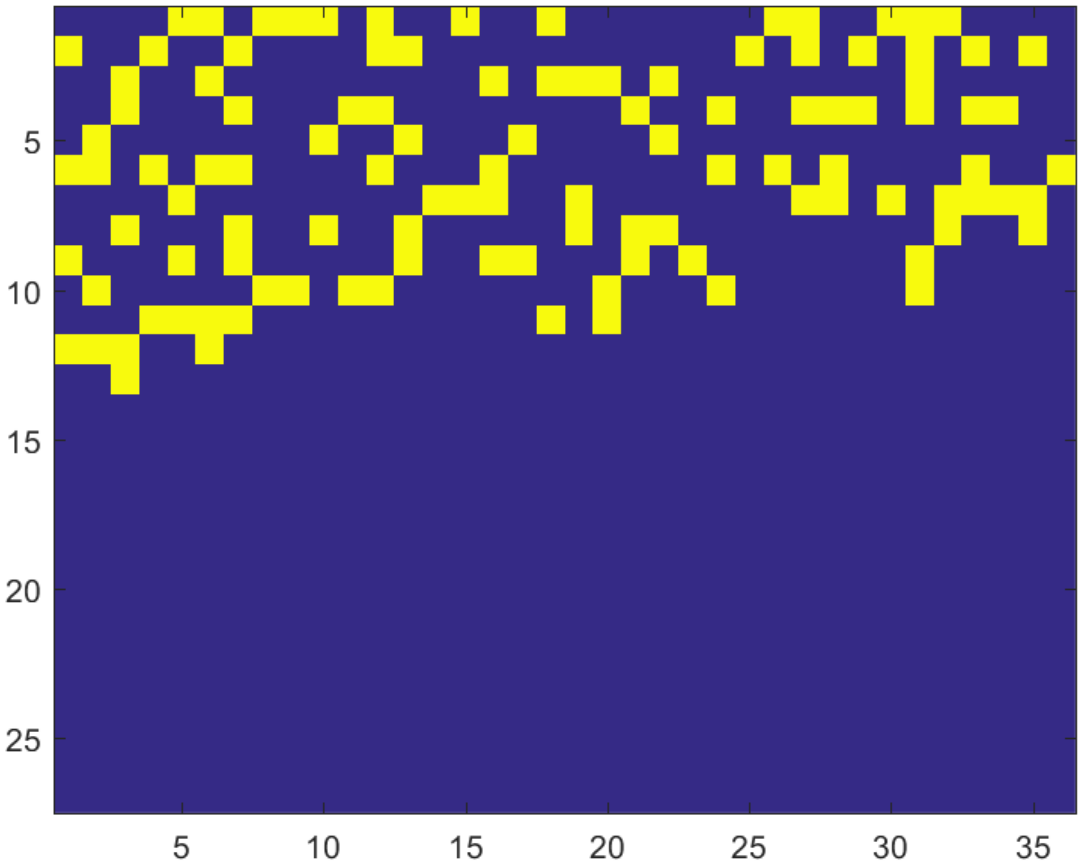
Stochastic Model (see Johnston et al 2014)

In time step τ cells given chance to move to neighbouring location with probability P_m.

During time step cells can give 'birth' with probability P_p and place new cell at neighbouring location.

Cell Biology Example - High Dimensional Summary Statistic

Simulated data from model with $P_m = 0.35$ and $P_p = 0.001$



Cell Biology Example - High Dimensional Summary Statistic

Aim is to obtain posterior distribution for P_m and P_p (can convert these to biologically relevant diffusivity D and the proliferation rate λ).

Model has no tractable likelihood function, but is easily simulated.

Summary statistics are hamming distance between adjacent lattices (images) and total number of cells at end of experiment.

145 summary statistics too large for conventional ABC.

Parametric Alternatives

It might be possible to overcome some drawbacks of ABC by using a parametric approximation to $p(s_y|\theta)$ instead of the non-parametric approximation used by ABC.

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

We refer to a Bayesian version of SL as BSL (Price et al 2018).

Bayesian Synthetic Likelihood

Basic method

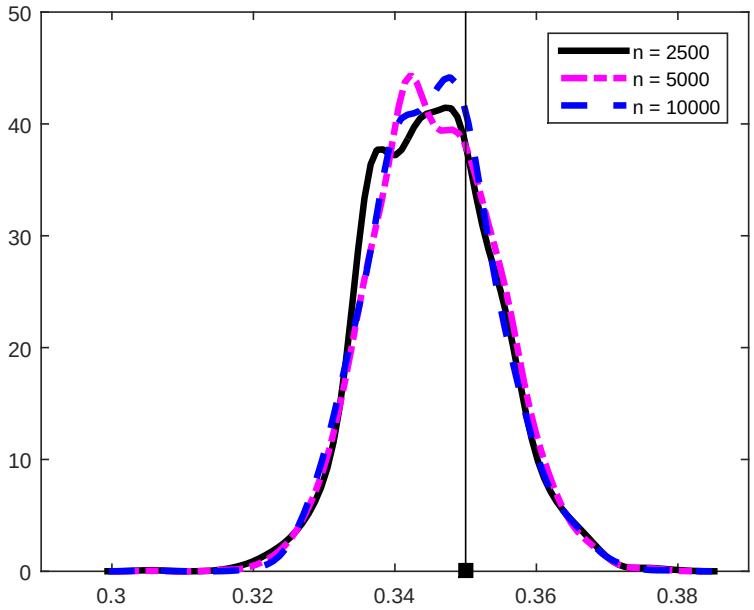
- Simulate n iid datasets from the model based on θ
- Calculate the n sets of summary statistics
- Calculate the sample mean, μ_n, and sample covariance matrix, Σ_n, of the set of simulated summary statistics
- The BSL replacement likelihood is

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

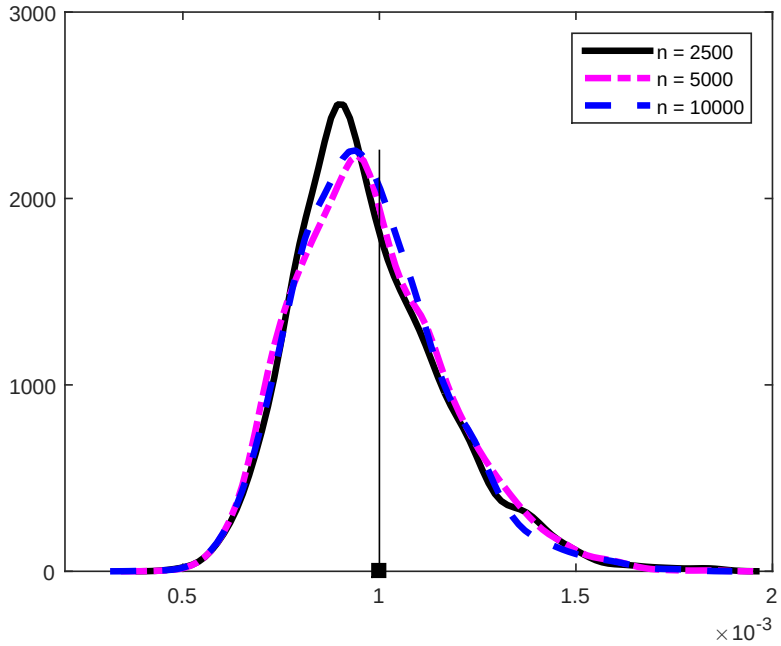
- Only tuning parameter is n.
- Target depends on multivariate normal assumption and on n because $\mathcal{N}(s_y; \mu_n(\theta), \Sigma_n(\theta))$ is not an unbiased estimator for $\mathcal{N}(s_y; \mu(\theta), \Sigma(\theta))$.

Cell Biology Example

BSL results - Sensitivity to n



BSL, P_m

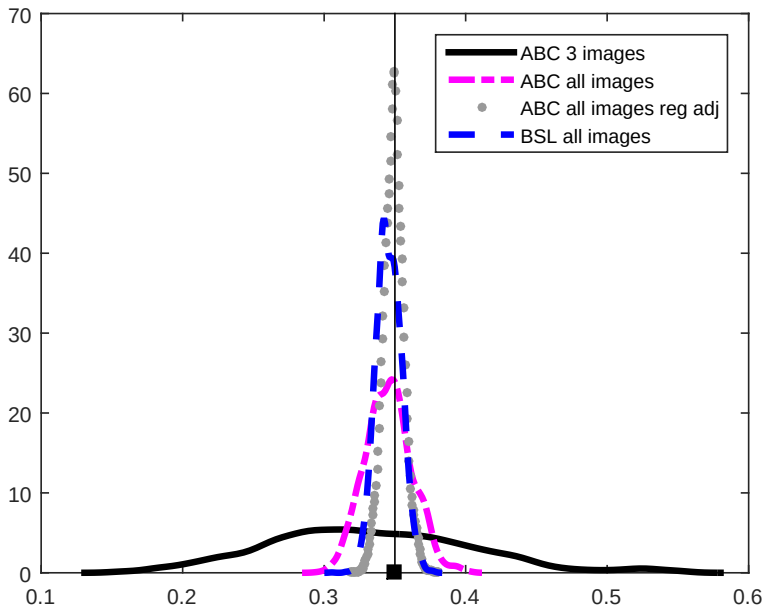


BSL, P_p

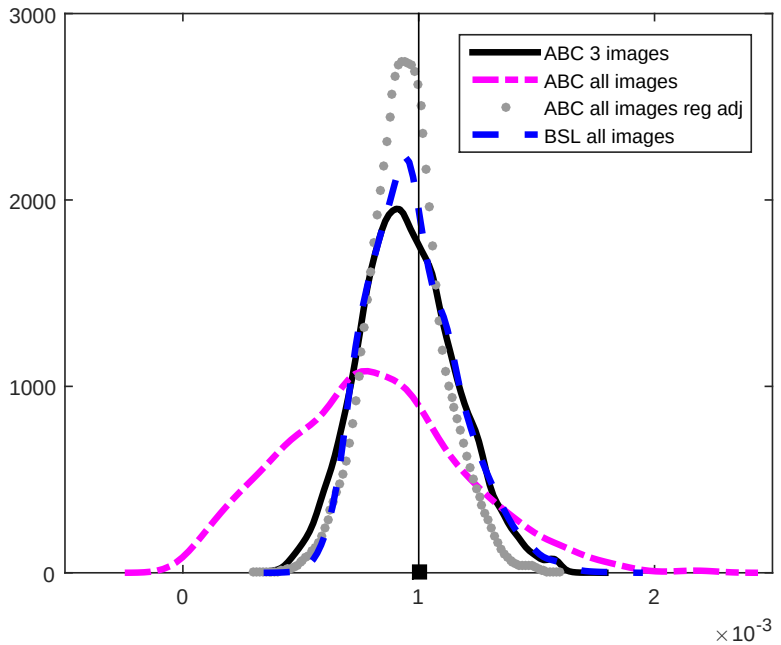
$n = 5000$ most efficient but 2500-10000 also efficient.

Cell Biology Example

Comparison to ABC



P_m



P_p

Using 16 core CPU average model simulations per second: 75 for ABC and 820 for BSL.

Drawbacks of BSL

- The distribution of the summary statistic must be roughly normal.
- The number of simulations per iteration, n, needs to be large depending on the size of the summary statistic to obtain a good sample covariance estimate.
- Reliance on MCMC to explore parameter space (not ideal in high dimensions)

Shrinkage Covariance Estimation

When sample size n is small, the sample covariance matrix can have poor properties.

We propose to use shrinkage covariance matrix estimation to improve efficiency of BSL.

We have considered the graphical lasso (Friedman et al. (2008)) and shrinkage estimator of Warton (2008).

BSL with shrinkage estimator

Replace sample covariance $\Sigma_n(\theta)$ with shrinkage estimator.

Put estimated synthetic likelihood in BSL algorithm.

Amount of shrinkage trades-off the accuracy of the posterior distribution (relative to BSL) against computational efficiency.

Note: See Everitt (2017) for a bootstrapping procedure for estimating covariance matrix.

Warton 2008 Shrinkage Estimator

The estimator of Warton 2008 is given by:

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

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 γ is the shrinkage parameter.

Advantages: γ easy to interpret and estimator has analytical form.

Graphical Lasso

Graphical lasso (Friedman et al. (2008)) estimates sparse precision matrix, $\Omega = \Sigma^{-1}$. The goal is to maximise the following penalised log-likelihood

$$\log p(s_{1:n}|\Omega) = K + \log|\Omega| - \text{tr}(\Omega S) - \lambda||\Omega||_1,$$

S is the sample covariance of $s_{1:n}$ and $||\cdot||_1$ denotes the L_1-norm.

The tuning parameter λ controls the sparsity of Ω.

Penalty in Graphical Lasso

Penalty, λ, controls the sparsity of the graphical lasso precision matrix.

- Smaller λ leads to posterior that is closer to BSL posterior.
- The value of λ is tied in with choice of n. To get smaller λ need larger n.

Standard approaches to selecting λ in the literature are BIC and cross validation error.

Select penalty within MCMC BSLasso

- Literature (eg Doucet et al. (2015)) that log likelihood estimators in MCMC algorithms should have standard deviation σ roughly 1.
- Price et al (2018) find that in the context of BSL σ between 1-2 gives a good trade-off between mixing and computational cost.
- Here we aim for $\sigma \approx 1.5$.

Procedure to select penalty within MCMC BSLasso

for $m = 1$ *to* M **do**

  Generate a collection of summary statistics $s_{1:n} \overset{\text{iid}}{\sim} p(s|\theta_0)$

  Compute the sample mean $\mu_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} s_i$

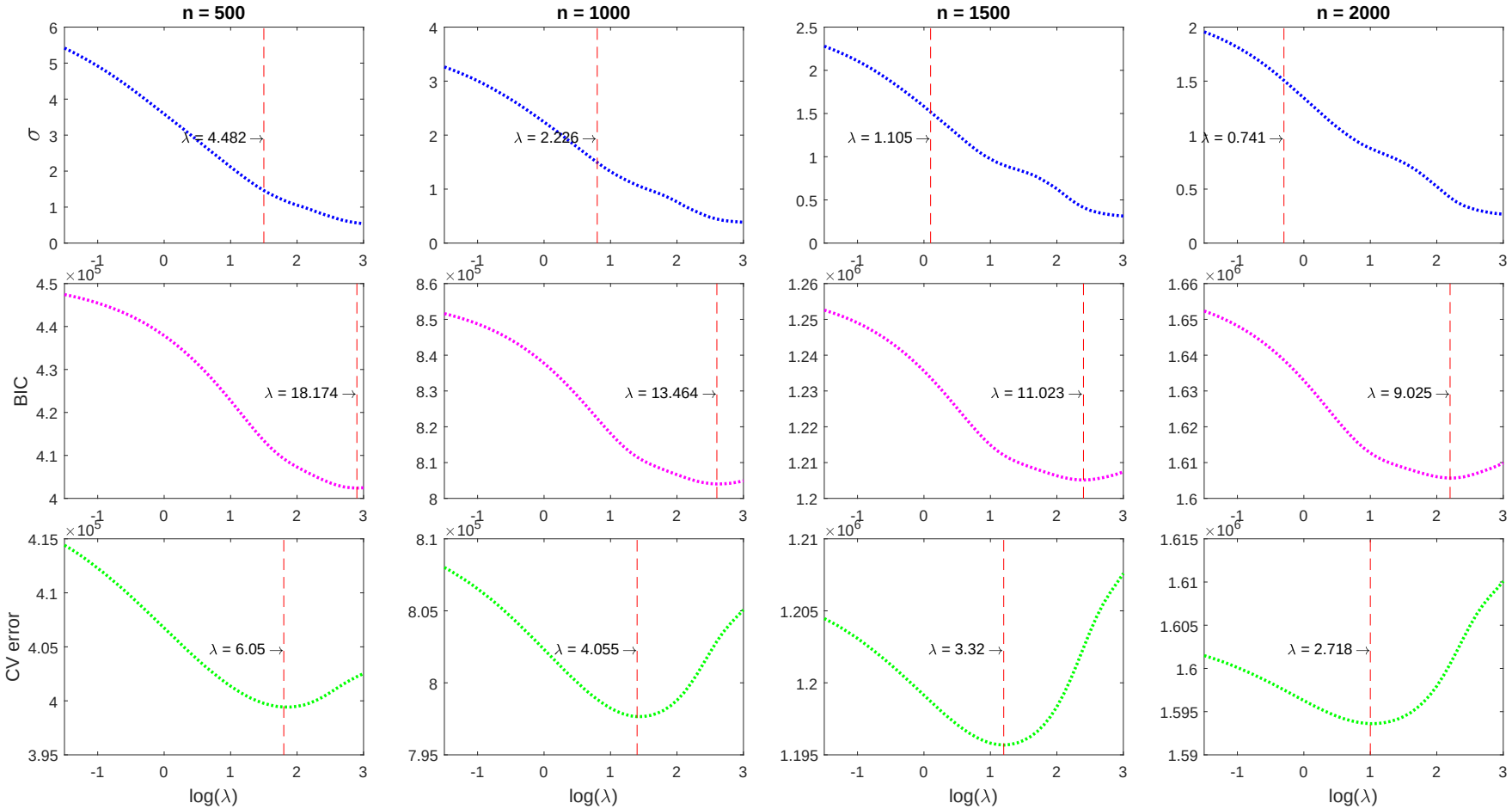
  Use the graphical lasso to obtain $\Sigma_n^{\lambda_k}(\theta_0)$ for each $k = 1, \ldots, K$ based on the same simulations $\{s_i\}_{i=1}^{n}$

  Use the estimated mean $\mu_n(\theta_0)$ and the collection of covariance matrix estimates $\{\Sigma_n^{\lambda_k}(\theta_0)\}_{k=1}^{K}$ to estimate the log SL $\log\{p_A^{n,\lambda_k}(s_y|\theta_0)\}_{k=1}^{K}$

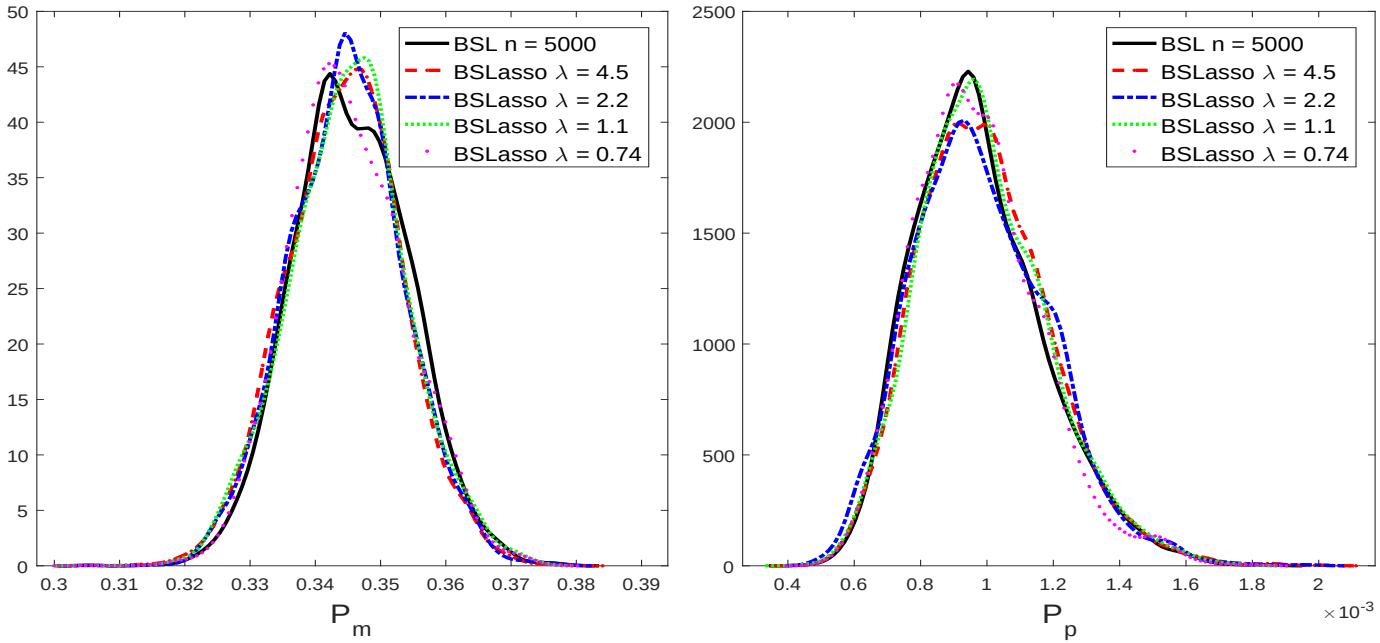
end

Find λ that leads to $\sigma \approx 1.5$.

Select Penalty

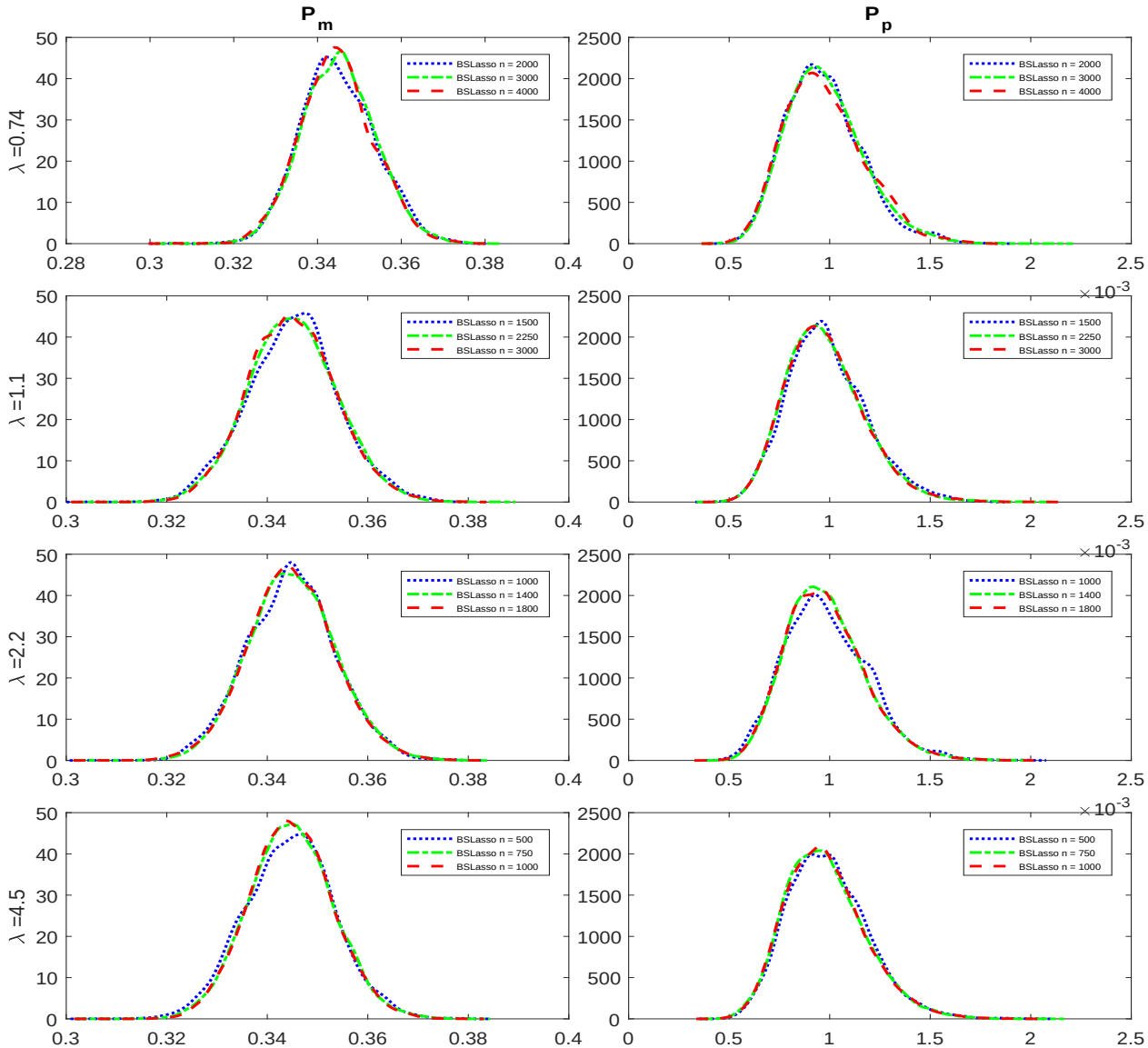


Posterior Distributions and Results Table



n	λ	acc. rate (%)	ESS P_m	ESS P_p
5000	-	21	8	8
500	4.50	16	55	63
1000	2.20	17	36	22
1500	1.10	16	16	19
2000	0.74	16	14	16

Sensitivity to n



Summary of the Cell Biology Example

- Very little accuracy is lost (relative to standard BSL) even for relatively large λ values.
- The posterior distributions are not sensitive to n.
- Our method to select penalty is robust to the above ns in terms of acceptance rate.

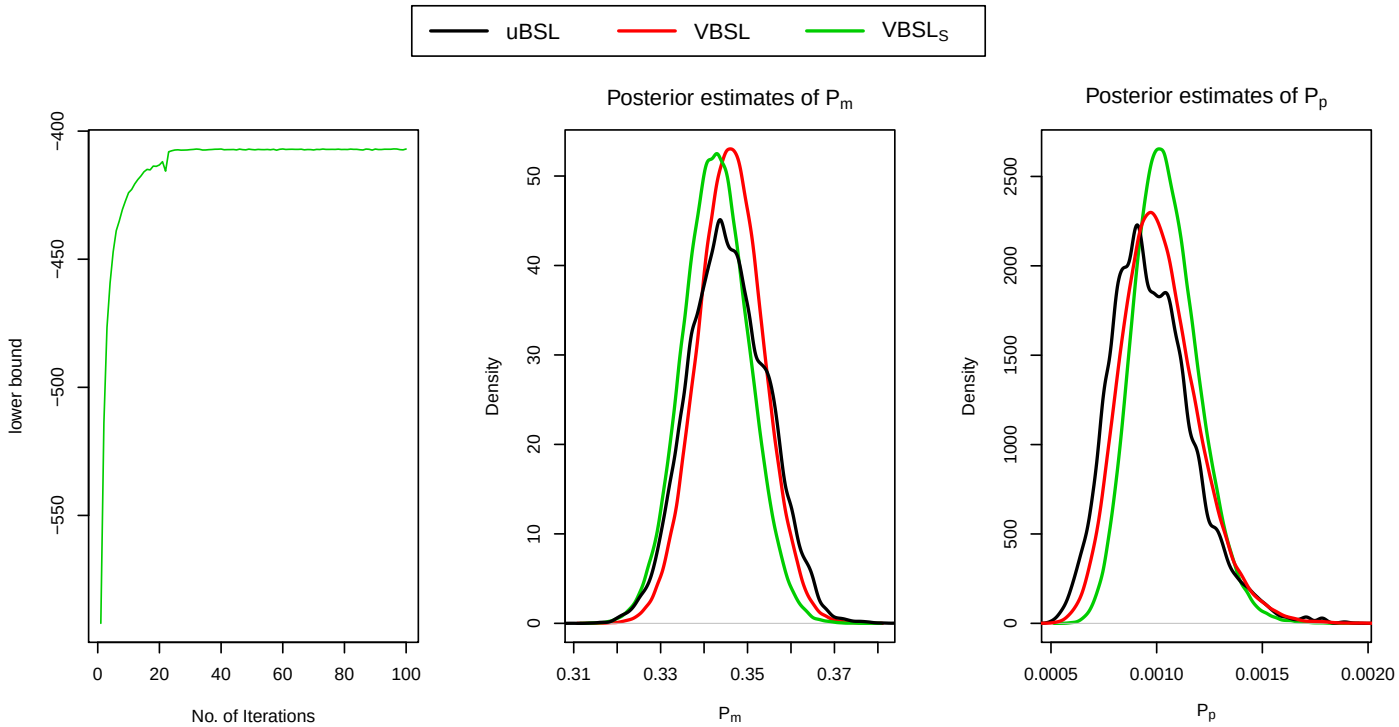
Variational Bayes BSL

When the posterior is roughly normal, our variational Bayes approaches to BSL are very efficient.

Ong et al 2018 (Statistics and Computing) first VBSL approach.

Ong et al 2018 (Submitted) extends above work to high-dimensional summary statistic and/or parameter.

VBSL Cell Biology Results



Lower bound and estimated posterior densities for the cell motility example using VBSL$_S$ with $N = 300$ and $S = 100$. Standard MCMC BSL ($N = 5000$) in Price et al 2018 and VBSL ($N = 1000$, $S = 100$) in Ong et al 2018.

Robust BSL

BSL relies on the multivariate normal assumption of model summary statistics.

Interested in how robust BSL to lack of normality.

Propose a more robust BSL via a semi-parametric estimator (see also Fasiolo et al 2018 and Dutta et al 2017 for other robust synthetic likelihood procedures).

Semi-parametric BSL

We improve flexibility of synthetic likelihood by using a semi-parametric Gaussian Copula model.

We use a kernel density estimate for the marginals:

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i),$$

Use fitted KDE to transform marginals to roughly normal.
$u_i = F_i(x_i), i = 1, \ldots, d$ and $\eta_i = \Phi^{-1}(u_i), i = 1, \ldots, d$.

Semi-parametric BSL

Capture dependence with a Gaussian Copula:

$$c_{\boldsymbol{R}}(\boldsymbol{u}) = \frac{1}{\sqrt{2\pi|\boldsymbol{R}|}} \exp\left\{ -\frac{1}{2}\eta^{\top}(\boldsymbol{R} - \mathcal{I})^{-1}\eta \right\},$$

where $\boldsymbol{R}$ is a correlation matrix. We estimate $\boldsymbol{R}$ using the Gaussian rank correlation.

Overall semi-parametric estimator given by

$$g(\boldsymbol{s_y}|\theta) = \frac{1}{\sqrt{|\hat{\boldsymbol{\Sigma}}|}} \exp\left\{ -\frac{1}{2}\hat{\eta}_{\boldsymbol{s_y}}^{\top}(\hat{\boldsymbol{\Sigma}} - \mathcal{I})\hat{\eta}_{\boldsymbol{s_y}} \right\} \prod_{i=1}^{d} \hat{f}_i(s_{x_i}),$$

MA2 Example

We consider an MA(2) time series model given by

$$y_t = \theta_1 z_t + \theta_2 z_{t-1},$$

for $t = 1, \ldots, T$, where $z_t \sim N(0, 1)$ for $t = 0, \ldots, T$, $\theta = (\theta_1, \theta_2)^T$ and $y = (y_1, \ldots, y_T)^T$.

Likelihood is multivariate normal with $\text{Var}[y_t] = 1 + \theta_1^2 + \theta_2^2$, $\text{Cov}(y_t, y_{t-1}) = \theta_1 + \theta_1\theta_2$, $\text{Cov}(y_t, y_{t-2}) = \theta_2$.

Prior distribution is uniform over:
$\Theta \equiv \{\mathbb{R}^2 : -1 < \theta_2 < 1, \theta_1 + \theta_2 > -1, \theta_1 - \theta_2 < 1\}$.

Summary statistic is all data.

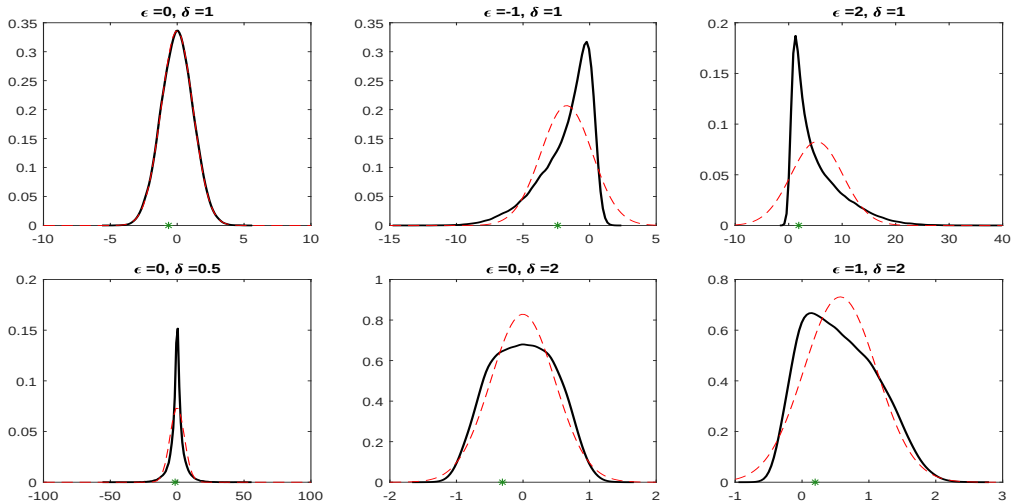
MA2 Example

To disturb normality for the marginals we consider following transformation:

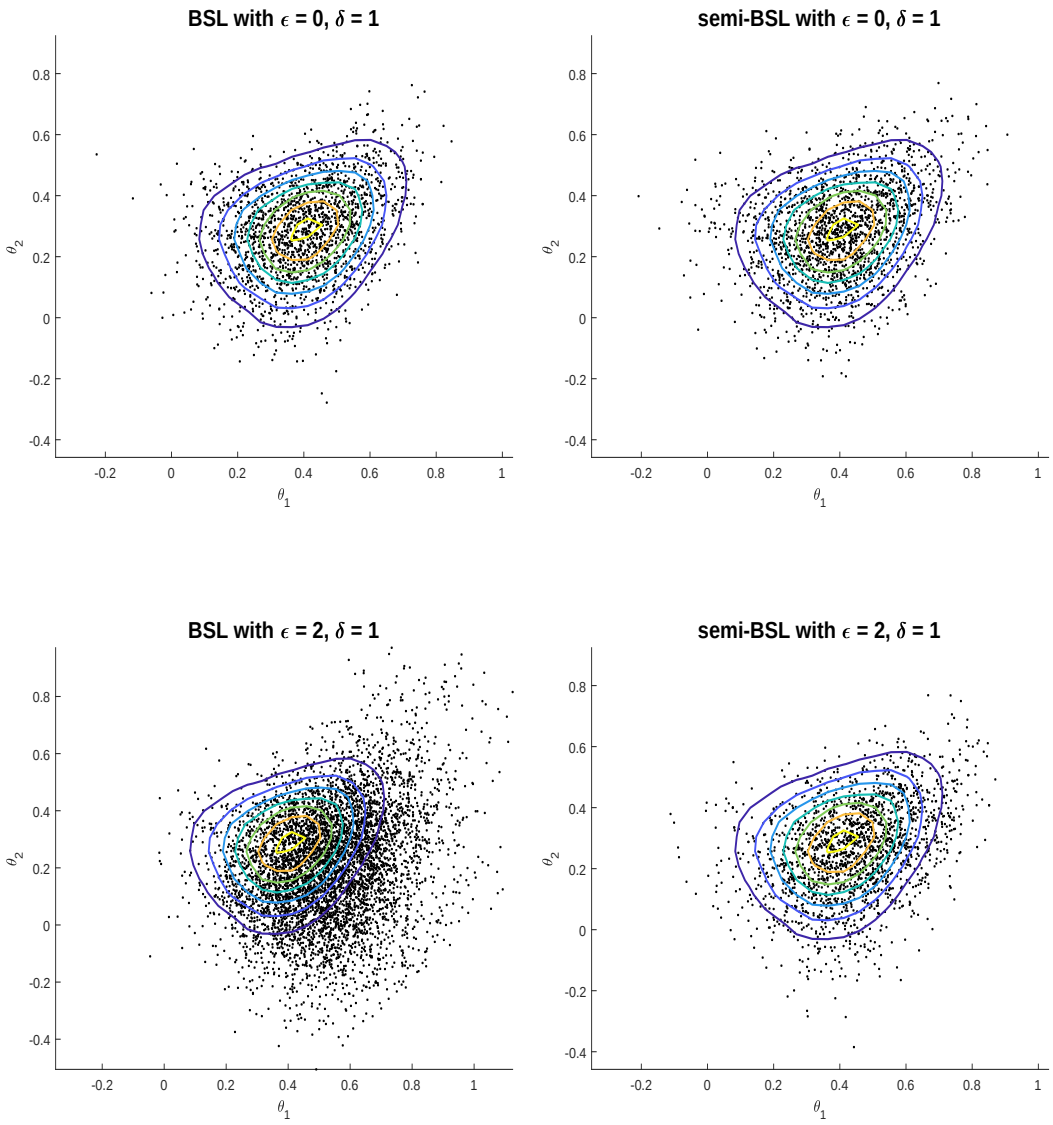
$$s_t = \sinh[\delta_t \sinh^{-1}(y_t) - \epsilon_t],$$

where ϵ_t and δ_t control the skewness and kurtosis, respectively.

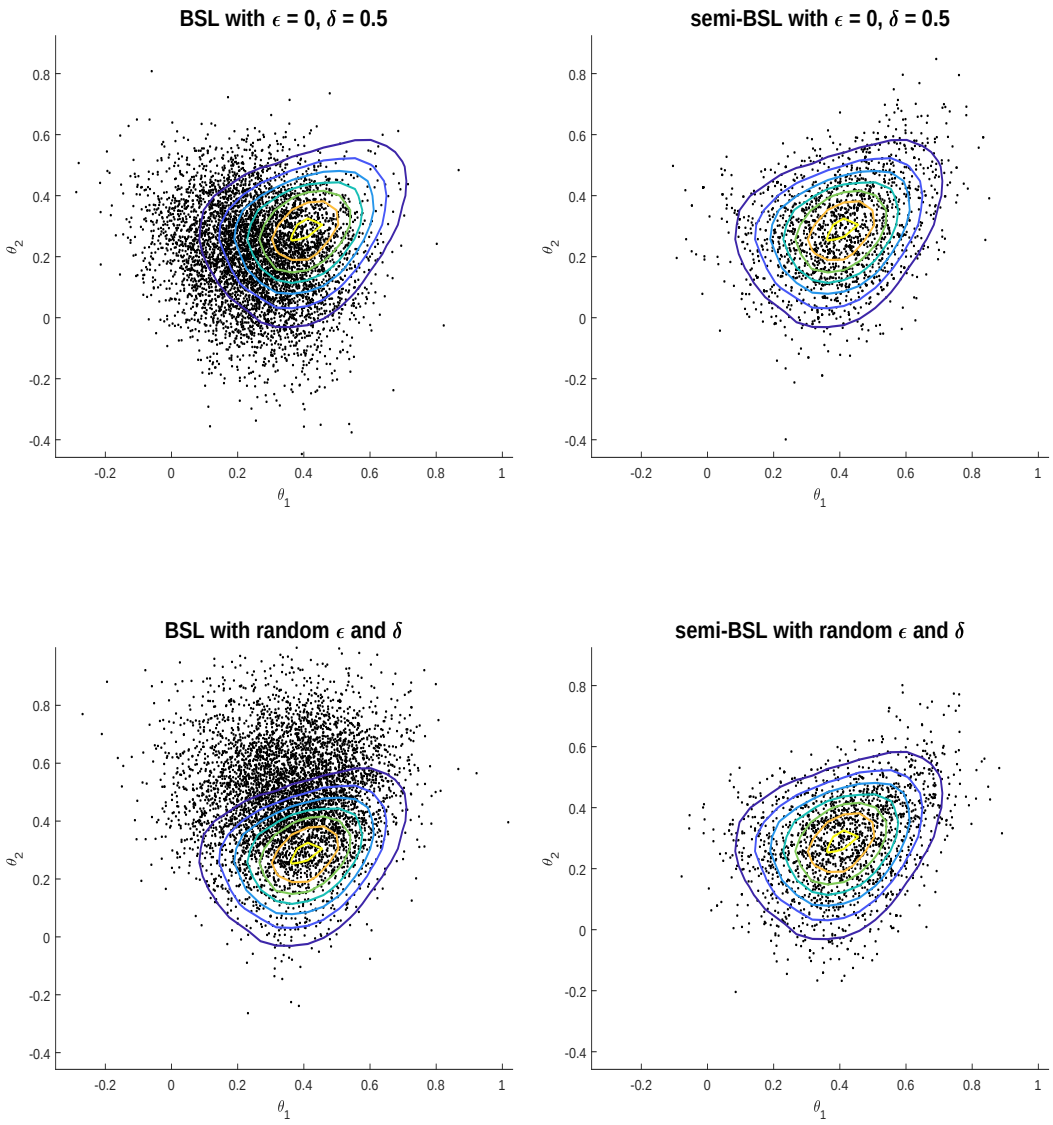
Here we consider different combinations of ϵ and δ.



MA2 Example – Results

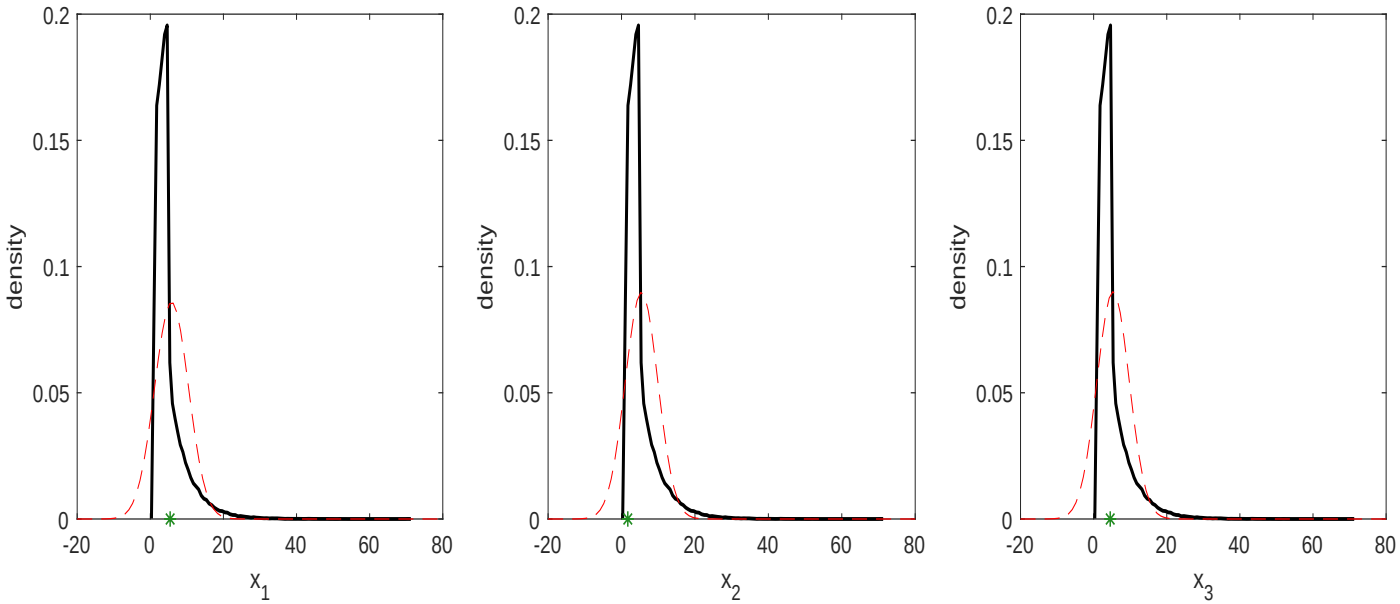


MA2 Example – Results

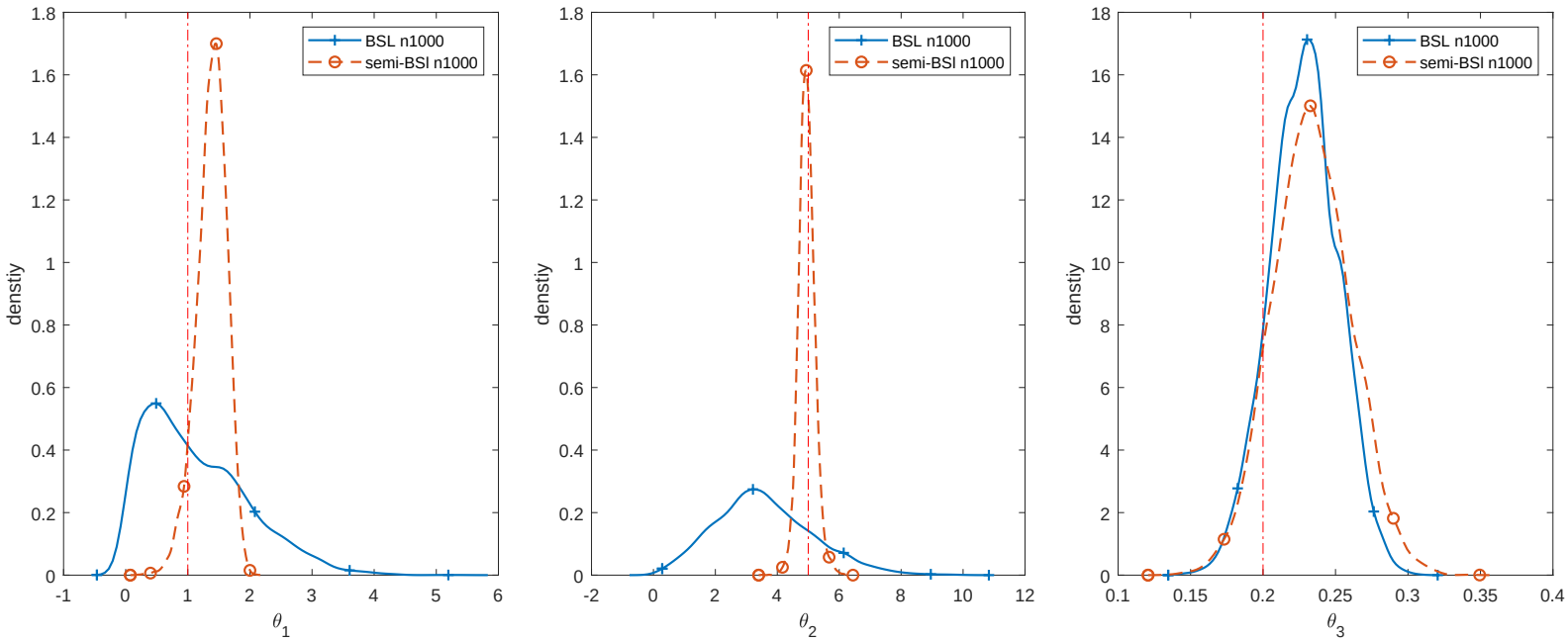


M/G/1 Queuing Model

Three parameters. 50 summary statistics (inter-departure times).
Basically no correlation between summaries.

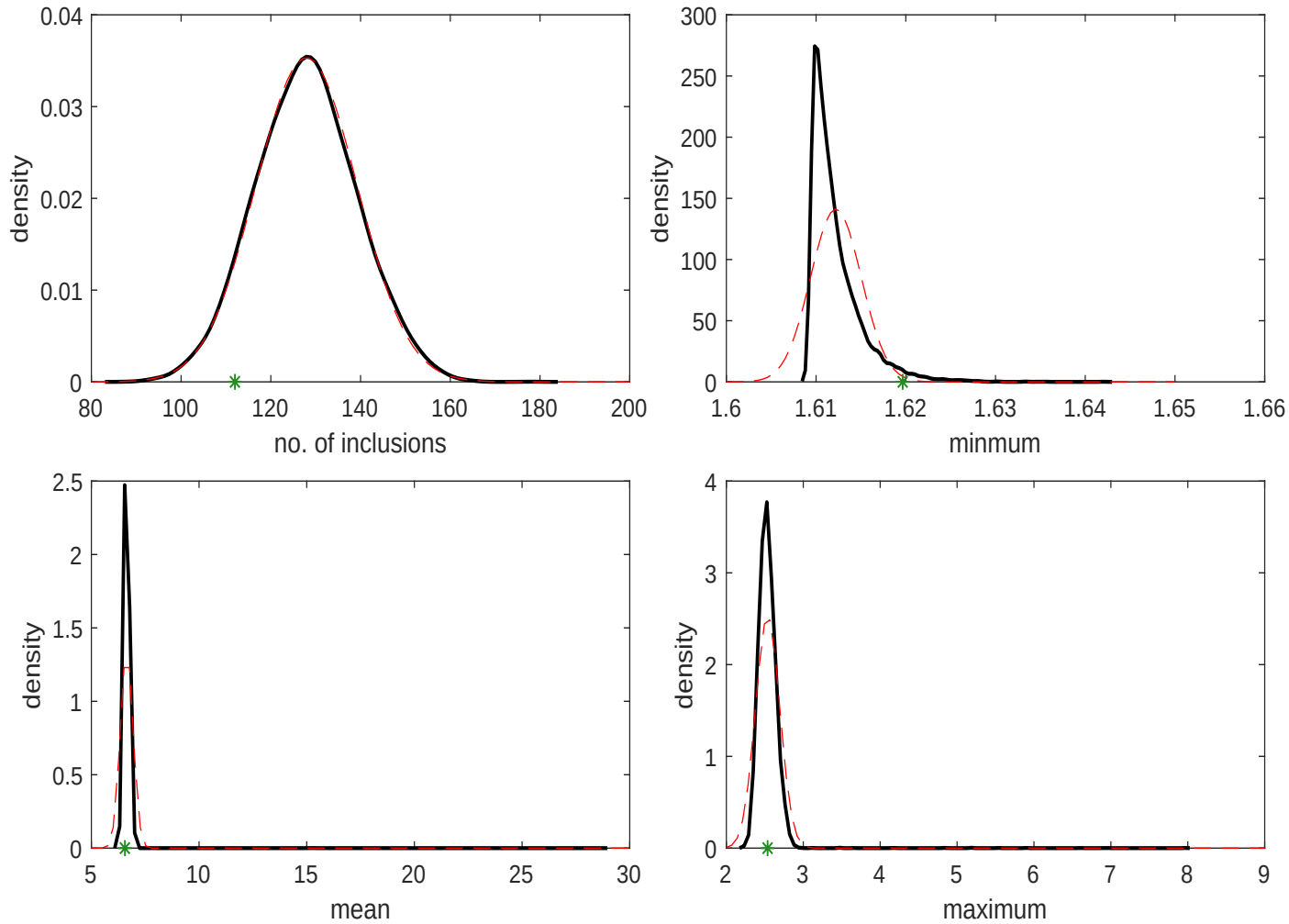


M/G/1 Queuing Model – Results

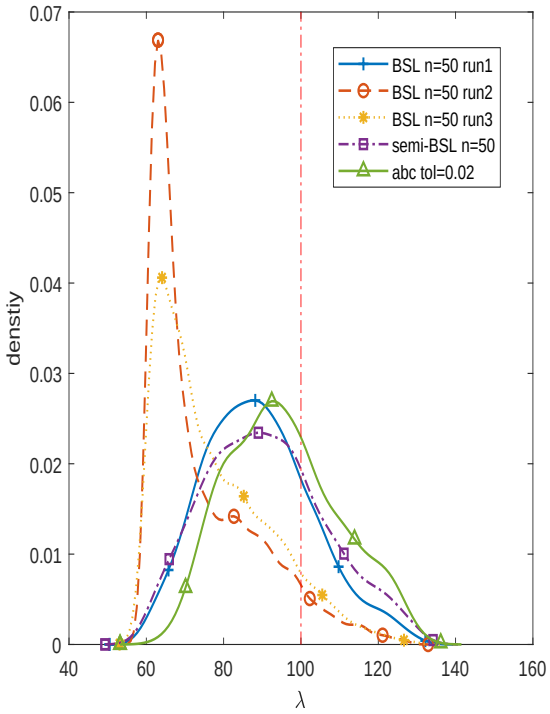


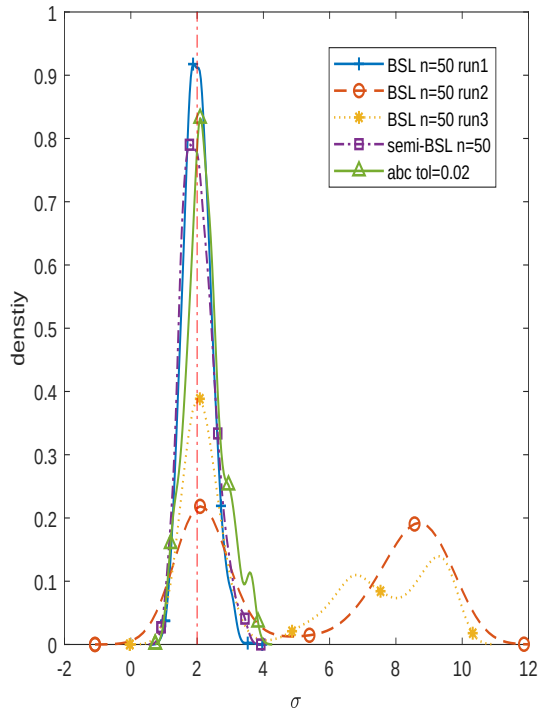
Stereological Extreme Example

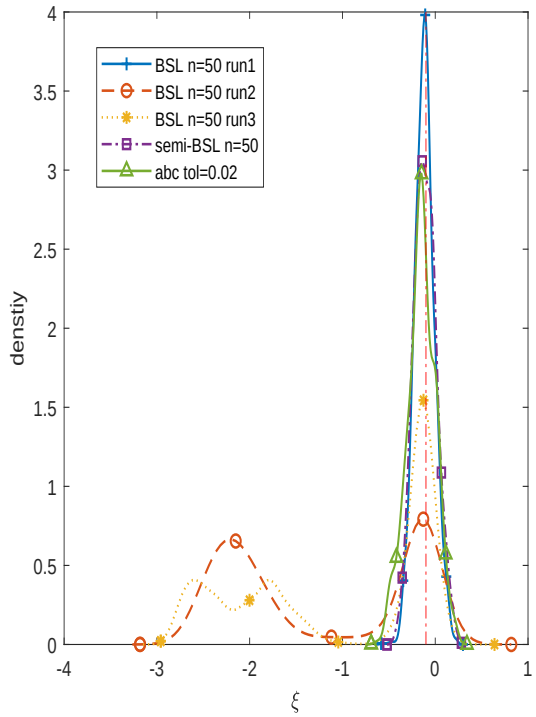
Three parameters and four summary statistics.



Stereological Extreme Results



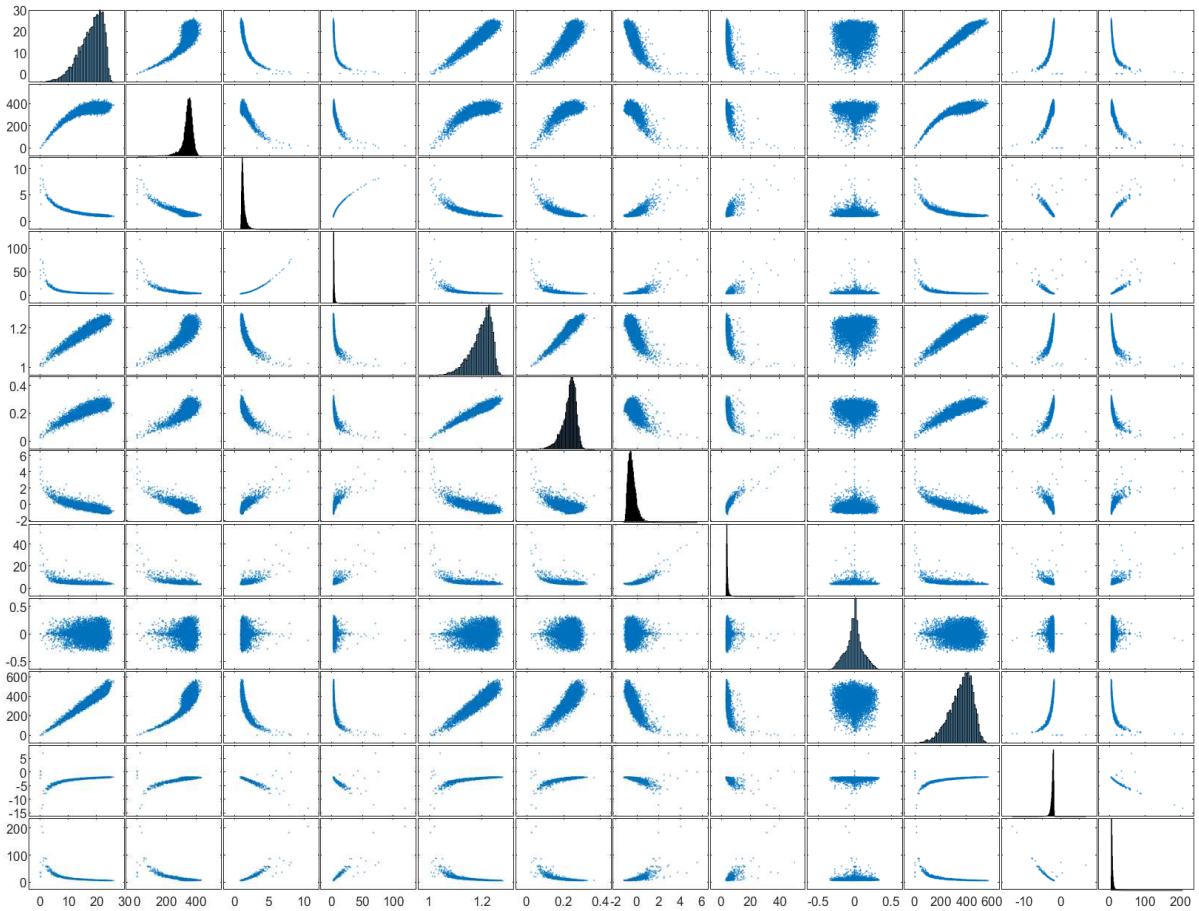




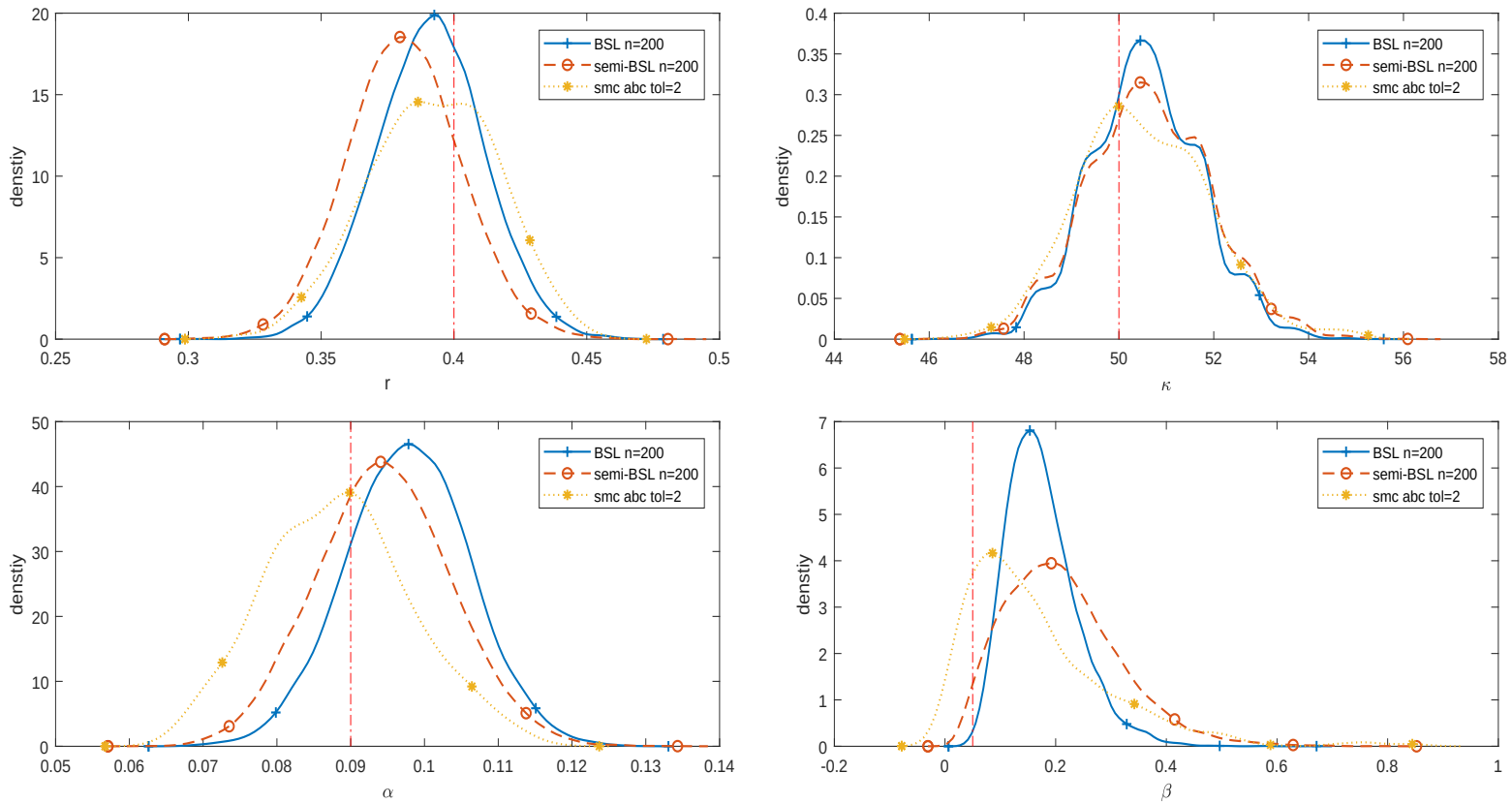
Boom and Bust Model

4 parameters and 12 summary statistics.

Summary statistic distribution very complex.

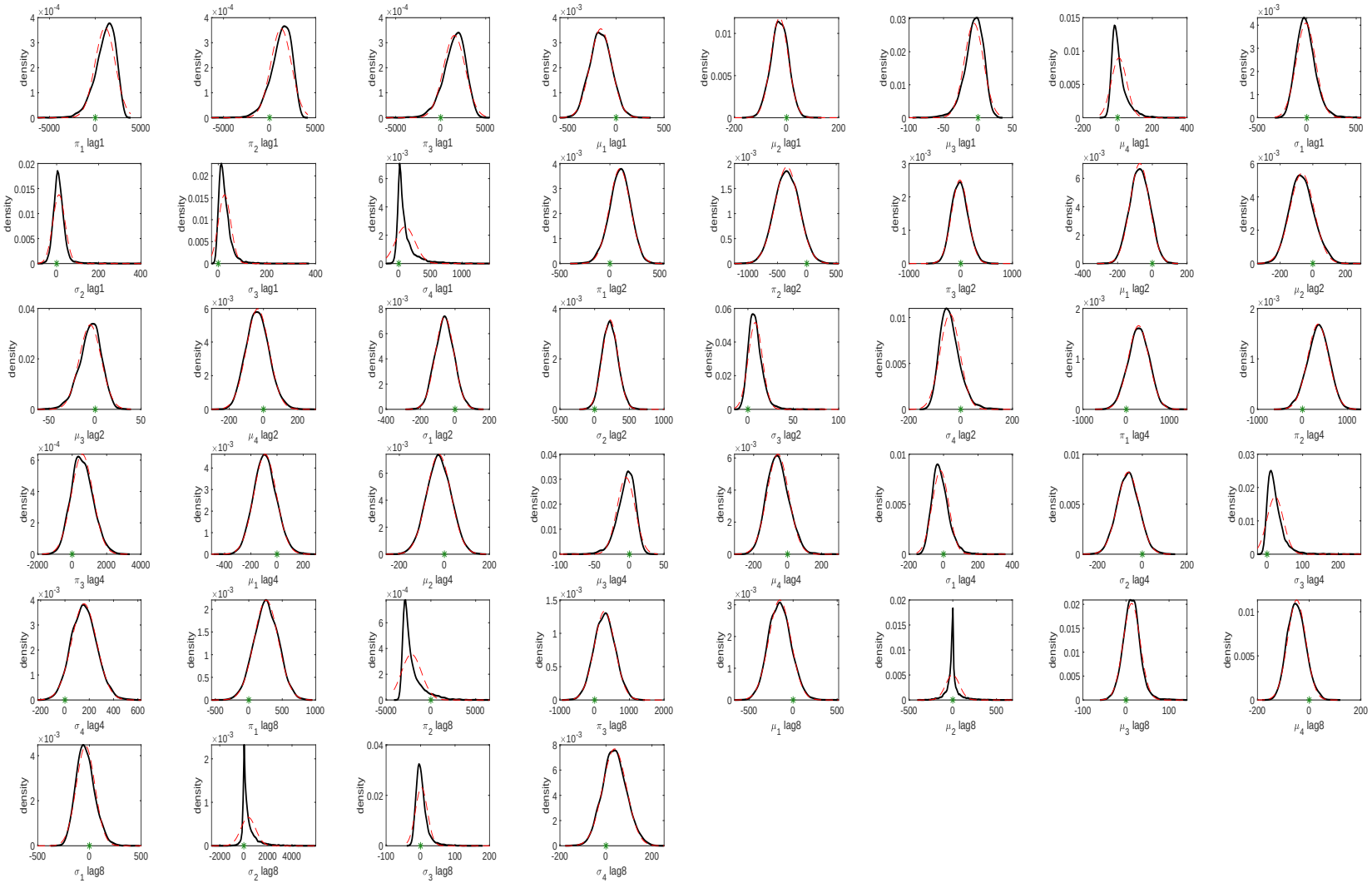


Boom and Bust Model – Results

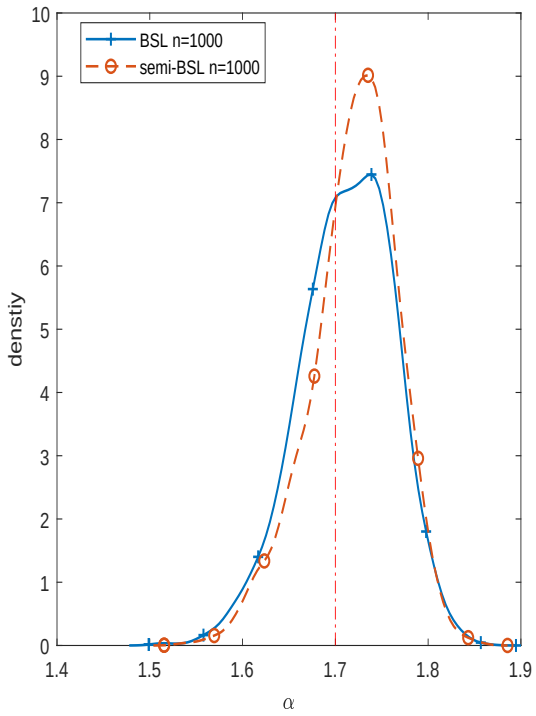


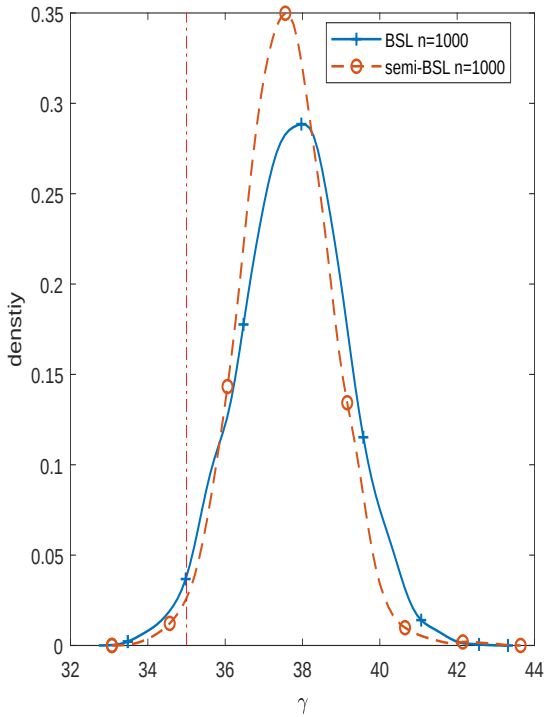
Fowler's Toad Example

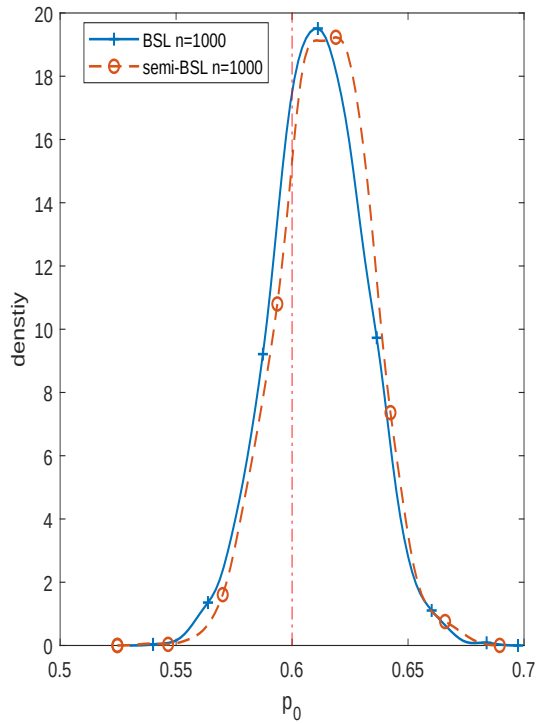
Three parameters, 44 summary statistics



Fowler's Toad Example – Results

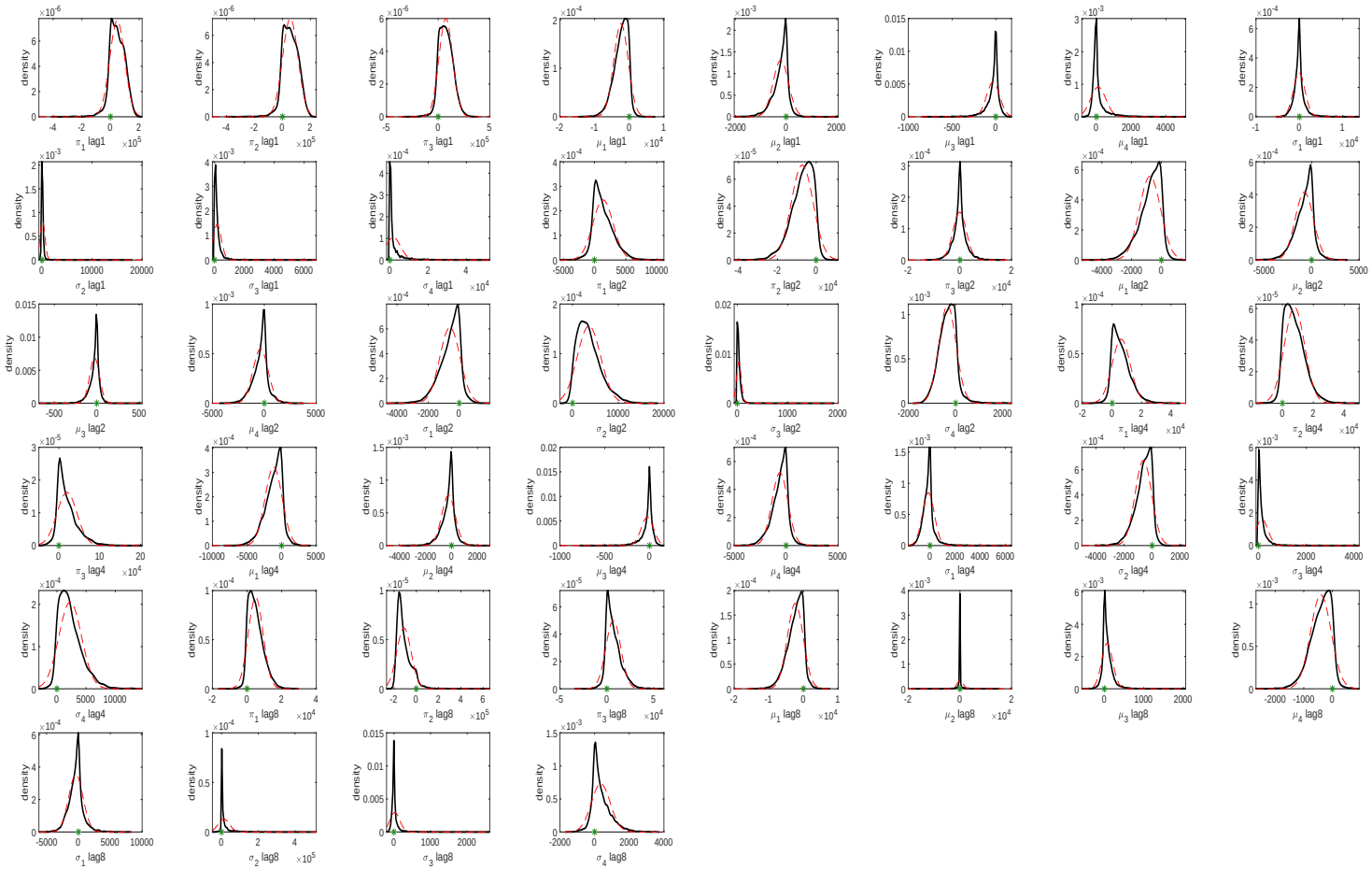




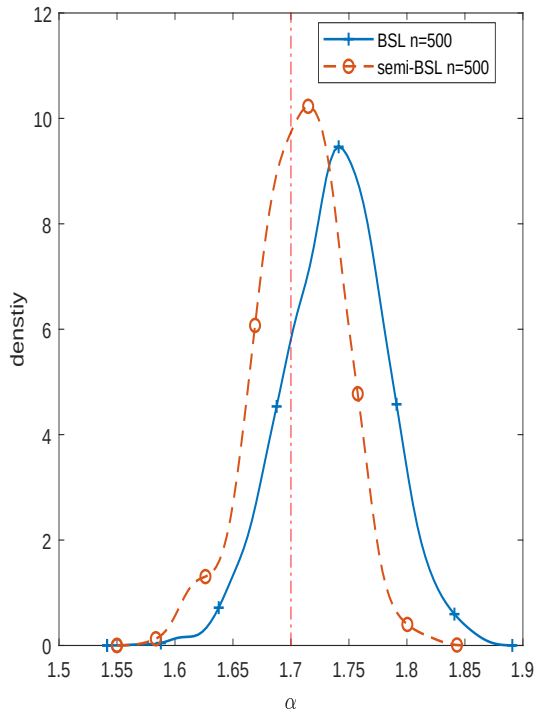


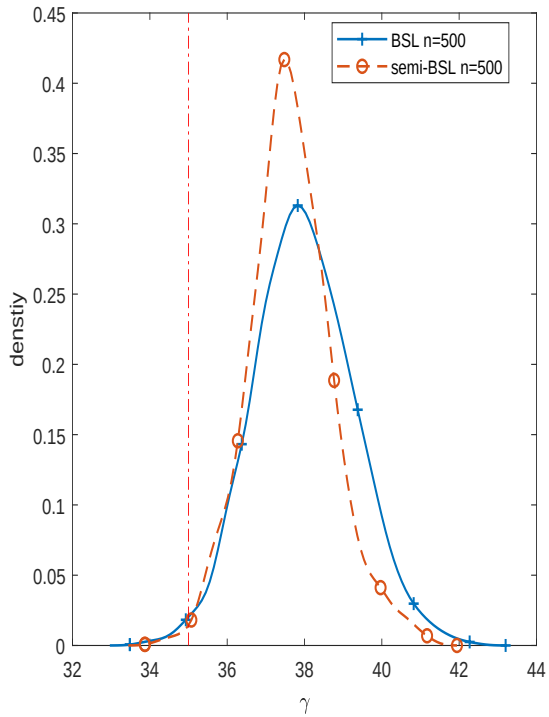
Fowler's Toad Example - Transformed Summaries

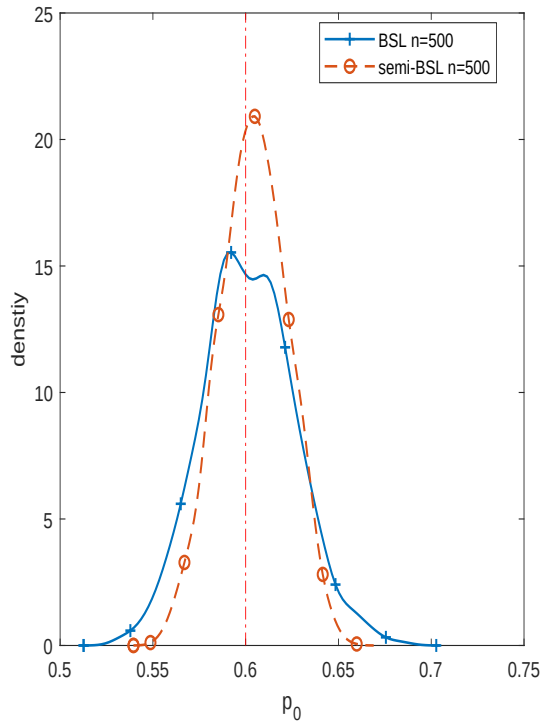
To make the summary statistic distributions less normal, we apply a transformation. $h(\cdot) = \text{sgn}(\cdot) * (|\cdot|)^p$. Here $p = 1.5$.



Fowler's Toad Example – Results

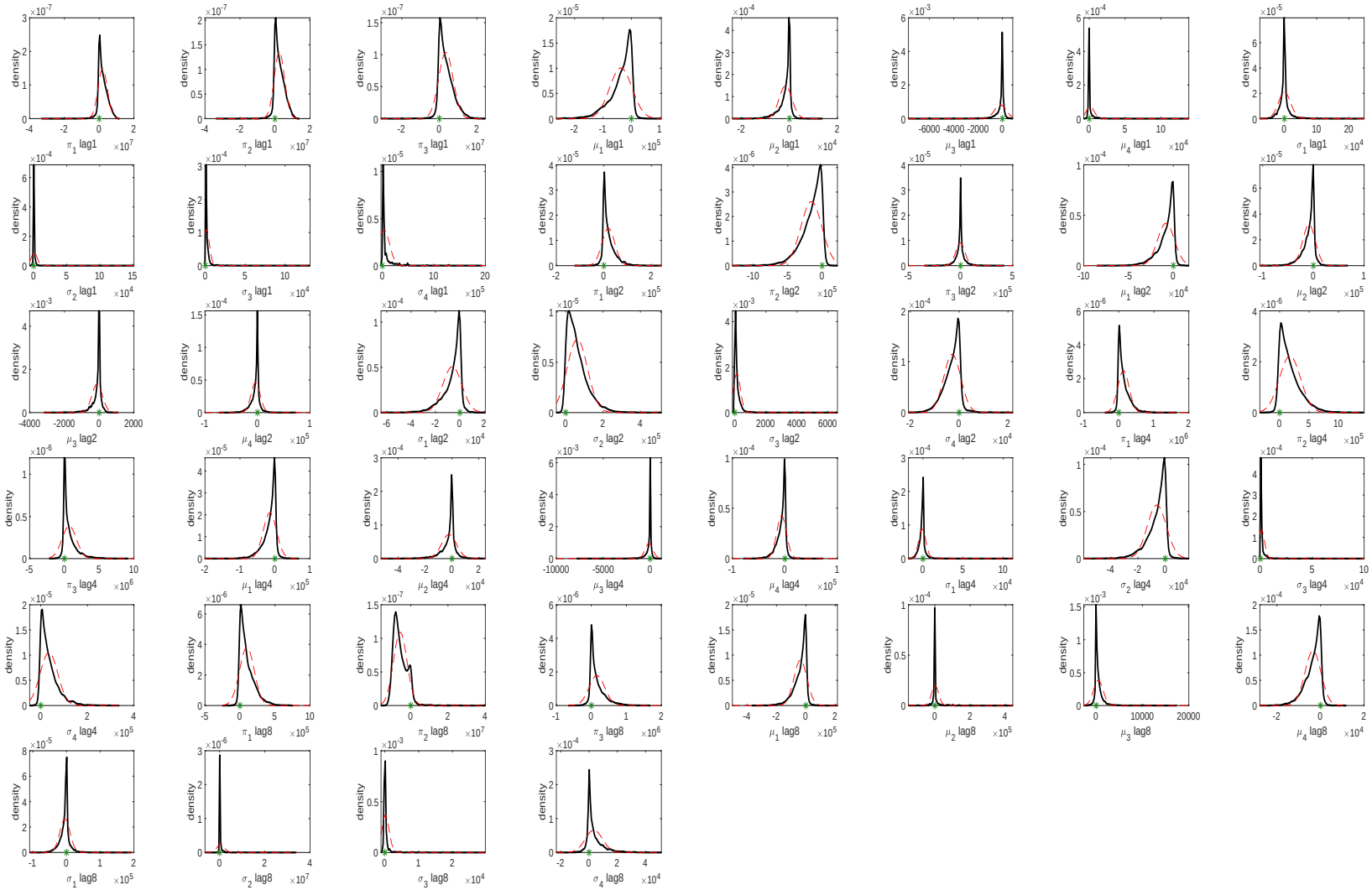






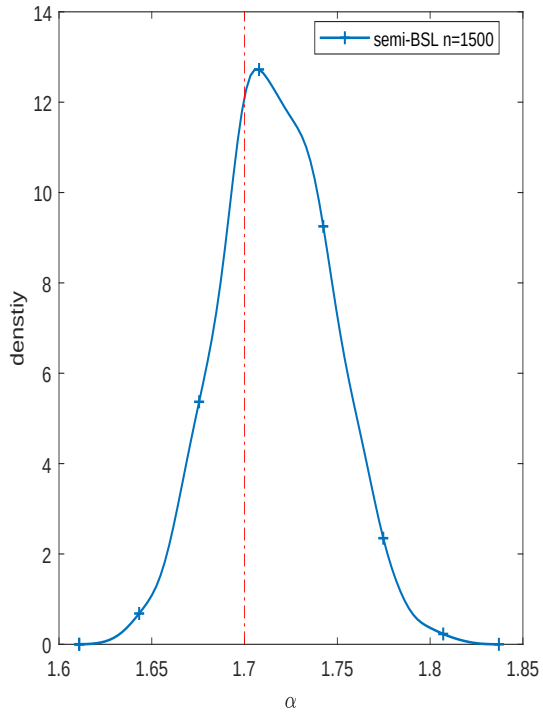
Fowler's Toad Example - Transformed Summaries

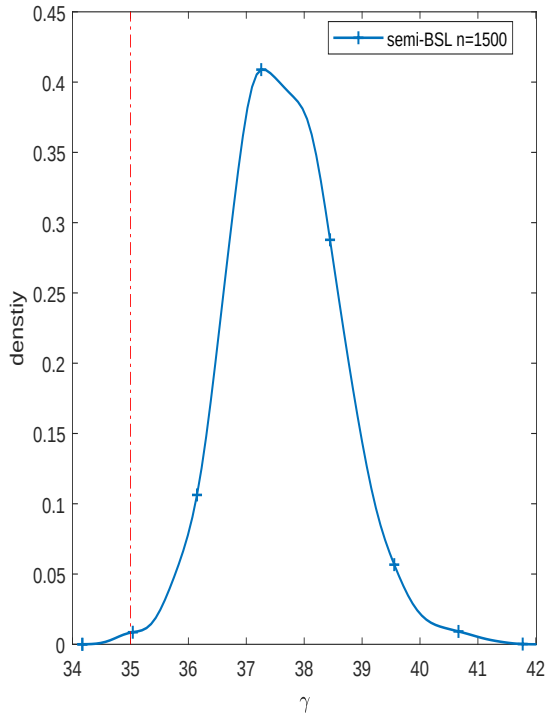
Here $p = 2$.

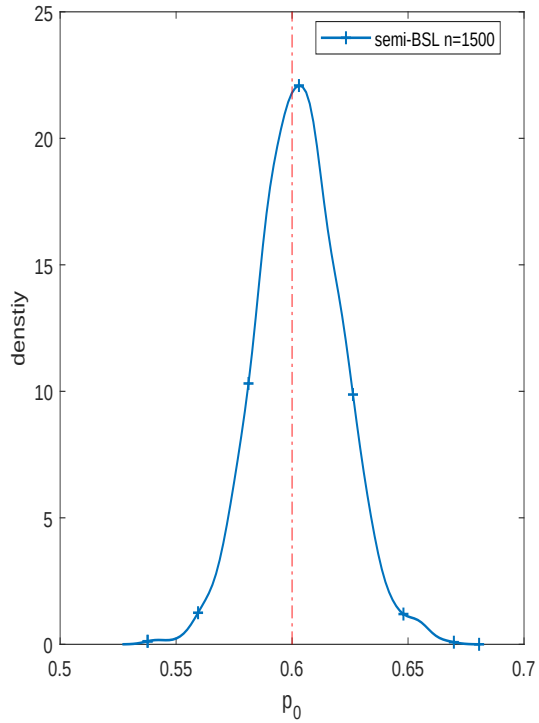


Fowler's Toad Example – Results

Standard BSL fails...

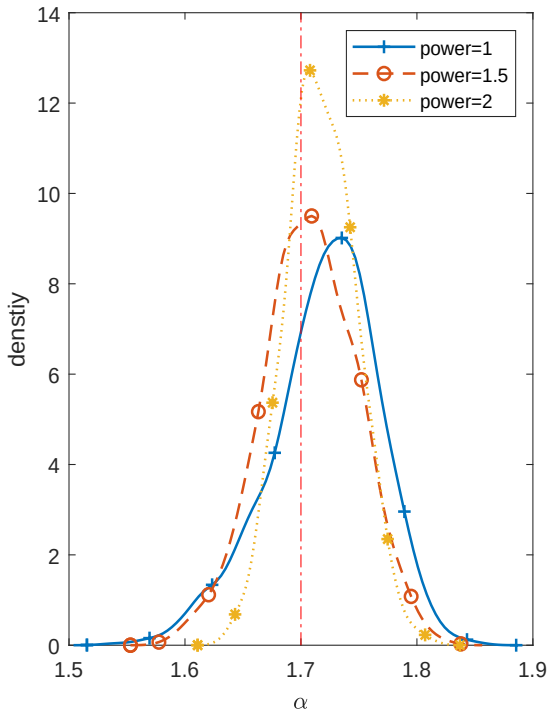


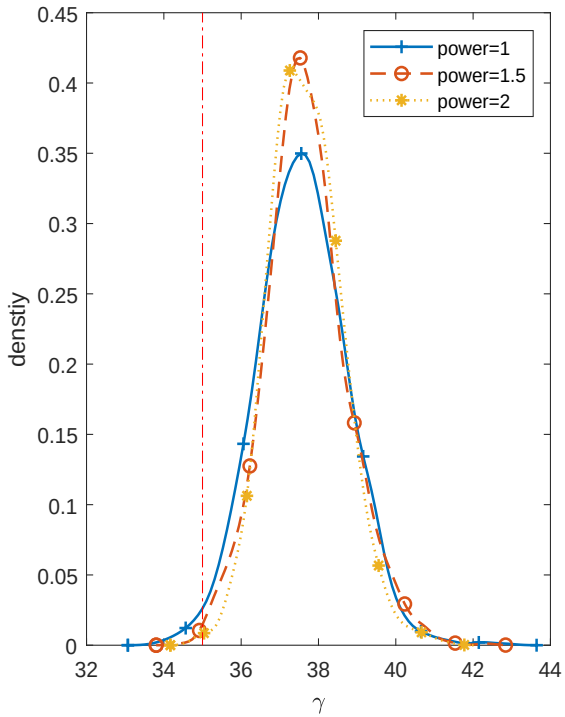


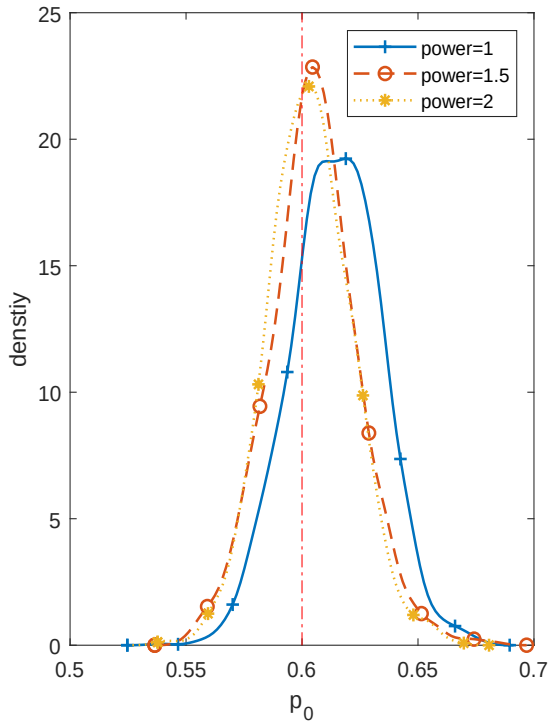


Fowler's Toad Example – Results

semiBSL results for three sets of summaries.







Future Directions

- Other approaches for Robust BSL.
- Extend to more flexible variational distributions (eg mixture of normals or Copula model)
- Asymptotic properties of BSL.

References

Ong, V. M-H., Nott, D. J., Tran, M-N., Sisson, S. A. and Drovandi , C. C. (2018) Variational Bayes with Synthetic Likelihood. Statistics and Computing.

Ong, V. M-H., Nott, D. J., Tran, M-N., Sisson, S. A. and Drovandi, C. C. (2018) Likelihood-Free Inference in High Dimensions with Synthetic Likelihood. In Revision.

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

An, Z., Nott, D. J. and Drovandi, C. C. (2018). 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

Email: c.drovandi@qut.edu.au Web: chrisdrovandi@weebly.com Twitter: @chris_drovandi