Robust and Efficient ABC.

David T. Frazier

October 15, 2020

MONASH University

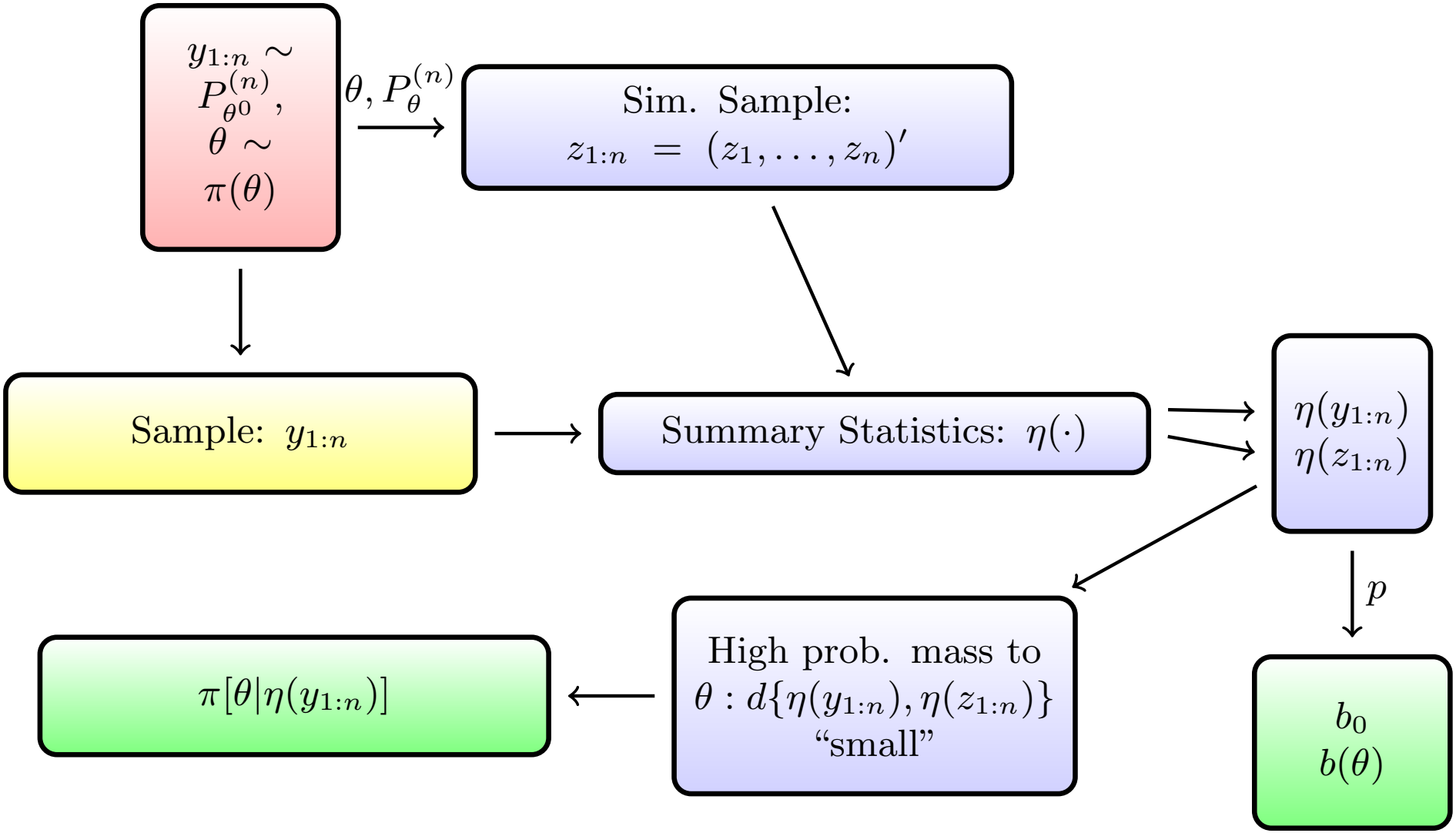
Motivation

- **Goal:** Bayesian inference on $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$
- **Beliefs:** (i) observed data $y_{1:n} = (y_1, ..., y_n)' \sim \mathcal{P} = \{P_\theta : \theta \in \Theta\}$; (ii) and $\theta \sim \Pi(\theta)$.
- Likelihood $dP_\theta / d\mu$ intractable.
- Conduct inference on θ via Approximate Bayesian Computational methods.

Motivation

- **Goal:** Bayesian inference on $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$
- **Beliefs:** (i) observed data $y_{1:n} = (y_1, ..., y_n)' \sim \mathcal{P} = \{P_\theta : \theta \in \Theta\}$; (ii) and $\theta \sim \Pi(\theta)$.
- Likelihood $dP_\theta / d\mu$ intractable.
- Conduct inference on θ via **A**pproximate **B**ayesian **C**omputational methods.

Canonical Idea



Two Summary-based Approaches

1. ABC: select θ such that $d\{\eta(y_{1:n}), \eta(z_{1:n})\} \leq \epsilon$; ϵ is the tolerance.

2. BSL:
   - Simulate m iid data sets and calculate $\{\eta(z^j_{1:n})\}_{j=1}^m$;
   - Calculate sample mean $\mu_m(\theta)$ and covariance $\Sigma_m(\theta)$ of simulated summaries;
   - $d\{\mu_m(\theta), \eta(y_{1:n})\} := [\eta(y_{1:n}) - \mu_m(\theta)]' \, \Sigma_m^{-1}(\theta) \, [\eta(y_{1:n}) - \mu_m(\theta)]$;
   - Use $|\Sigma_m(\theta)|^{-1/2} \exp\left[-\frac{1}{2} d\{\mu_m(\theta), \eta(y_{1:n})\}\right]$ as likelihood in MCMC.

A Maturing Literature

- What inference issues remain?
  1. **Statistical Efficiency.**
  2. **Robustness to Misspecification.**
- ABC/BSL: must choose $\eta \mapsto \eta(y_{1:n})$.
  1. Loss of information.
  2. No (feasible) "best" choice.
  3. Robustness is η-dependent.
- **Our Goal:** propose partial solution.

A Maturing Literature

- What inference issues remain?
  1. **Statistical Efficiency.**
  2. **Robustness to Misspecification.**
- ABC/BSL: must choose $\eta \mapsto \eta(y_{1:n})$.
  1. Loss of information.
  2. No (feasible) "best" choice.
  3. Robustness is η-dependent.
- **Our Goal:** propose partial solution.

An alternative: no summaries?

- **Above issues:** stem from choice of $\eta \mapsto \eta(y_{1:n})$.
- Avoid summarizing $y_{1:n}$? **Yes! Bernton et al., (2019, JRSS:B)**
- Wasserstein-ABC (W-ABC):
  - Replace $d\{\eta(y_{1:n}), \eta(z_{1:n})\}$, by the p- Wasserstein distance

$$\mathcal{W}_p(y_{1:n}, z_{1:n})^p = \inf_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^{n} \gamma(y_i, z_{\sigma(i)})^p, \tag{1}$$

  - $\mathcal{S}_n$ is the set of permutations of $\{1, \ldots, n\}$.
  - Common approach: Take $\gamma(y_i, z_{\sigma(i)})^p = |y_i - z_j|$;
  - $\implies$ match all n order statistics.

An alternative: no summaries?

- **Above issues:** stem from choice of $\eta \mapsto \eta(y_{1:n})$.
- Avoid summarizing $y_{1:n}$? **Yes! Bernton et al., (2019, JRSS:B)**
- Wasserstein-ABC (W-ABC):
  - Replace $d\{\eta(y_{1:n}), \eta(z_{1:n})\}$, by the p- Wasserstein distance

$$\mathcal{W}_p(y_{1:n}, z_{1:n})^p = \inf_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^{n} \gamma(y_i, z_{\sigma(i)})^p, \tag{1}$$

  - $\mathcal{S}_n$ is the set of permutations of $\{1, \ldots, n\}$.
  - Common approach: Take $\gamma(y_i, z_{\sigma(i)})^p = |y_i - z_j|$;
  - $\implies$ match all n order statistics.

An alternative: no summaries?

- **Above issues:** stem from choice of $\eta \mapsto \eta(y_{1:n})$.
- Avoid summarizing $y_{1:n}$? **Yes! Bernton et al., (2019, JRSS:B)**
- Wasserstein-ABC (W-ABC):
  - Replace $d\{\eta(y_{1:n}), \eta(z_{1:n})\}$, by the p- Wasserstein distance

$$\mathcal{W}_p(y_{1:n}, z_{1:n})^p = \inf_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^{n} \gamma(y_i, z_{\sigma(i)})^p, \tag{1}$$

  - $\mathcal{S}_n$ is the set of permutations of $\{1, \ldots, n\}$.
  - Common approach: Take $\gamma(y_i, z_{\sigma(i)})^p = |y_i - z_j|$;
  - $\implies$ match all n order statistics.

W-ABC

- W-ABC posterior: $A \subset \Theta$, $1[\cdot]$ indicator function.

$$\Pi_\epsilon(\theta \in A | y_{1:n}) = \int_A \frac{\mathrm{d}\Pi(\theta) \int_{\mathcal{Y}^n} 1\left[\mathcal{W}_p(y_{1:n}, z_{1:n})^p \leq \epsilon\right] \mathrm{d}P_\theta^{(n)}(z_{1:n})}{\int_\Theta \mathrm{d}\Pi(\theta) \int_{\mathcal{Y}^n} 1\left[\mathcal{W}_p(y_{1:n}, z_{1:n})^p \leq \epsilon\right] \mathrm{d}P_\theta^{(n)}(z_{1:n})},$$

- Benefits:
  - No explicit choice for $\eta(\cdot)$.✓
  - Using full sample (possibly more efficient). ✓
- Questions:
  - Theoretical guarantees? (Posterior concentration ✓)
  - Efficiency?
  - Robustness?

W-ABC

- W-ABC posterior: $A \subset \Theta$, $1[\cdot]$ indicator function.

$$\Pi_\epsilon(\theta \in A | y_{1:n}) = \int_A \frac{\mathrm{d}\Pi(\theta) \int_{\mathcal{Y}^n} 1\left[\mathcal{W}_p(y_{1:n}, z_{1:n})^p \leq \epsilon\right] \mathrm{d}P_\theta^{(n)}(z_{1:n})}{\int_\Theta \mathrm{d}\Pi(\theta) \int_{\mathcal{Y}^n} 1\left[\mathcal{W}_p(y_{1:n}, z_{1:n})^p \leq \epsilon\right] \mathrm{d}P_\theta^{(n)}(z_{1:n})},$$

- Benefits:
  - No explicit choice for $\eta(\cdot)$.✓
  - Using full sample (possibly more efficient). ✓
- Questions:
  - Theoretical guarantees? (Posterior concentration ✓)
  - Efficiency?
  - Robustness?

W-ABC

- W-ABC posterior: $A \subset \Theta$, $1[\cdot]$ indicator function.

$$\Pi_\epsilon(\theta \in A | y_{1:n}) = \int_A \frac{\mathrm{d}\Pi(\theta) \int_{\mathcal{Y}^n} 1\left[\mathcal{W}_p(y_{1:n}, z_{1:n})^p \leq \epsilon\right] \mathrm{d}P_\theta^{(n)}(z_{1:n})}{\int_\Theta \mathrm{d}\Pi(\theta) \int_{\mathcal{Y}^n} 1\left[\mathcal{W}_p(y_{1:n}, z_{1:n})^p \leq \epsilon\right] \mathrm{d}P_\theta^{(n)}(z_{1:n})},$$

- Benefits:
  - No explicit choice for $\eta(\cdot)$.✓
  - Using full sample (possibly more efficient). ✓

- Questions:
  - Theoretical guarantees? (Posterior concentration ✓)
  - Efficiency?
  - Robustness?

W-ABC

- W-ABC posterior: $A \subset \Theta$, $1[\cdot]$ indicator function.

$$\Pi_\epsilon(\theta \in A | y_{1:n}) = \int_A \frac{\mathrm{d}\Pi(\theta) \int_{\mathcal{Y}^n} 1\left[\mathcal{W}_p(y_{1:n}, z_{1:n})^p \leq \epsilon\right] \mathrm{d}P_\theta^{(n)}(z_{1:n})}{\int_\Theta \mathrm{d}\Pi(\theta) \int_{\mathcal{Y}^n} 1\left[\mathcal{W}_p(y_{1:n}, z_{1:n})^p \leq \epsilon\right] \mathrm{d}P_\theta^{(n)}(z_{1:n})},$$

- Benefits:
  - No explicit choice for $\eta(\cdot)$.✓
  - Using full sample (possibly more efficient). ✓

- Questions:
  - Theoretical guarantees? (Posterior concentration ✓)
  - Efficiency?
  - Robustness?

Goal of this Paper

- **ABC** point estimators generally statistically inefficient.
- Lack stability to deviations from modelling assumptions.
- **Goal: efficient and stable inference.**
  - **Stable:** minor deviations from model $\Longrightarrow$ minor inferential changes.
- Want both...

Outline

1. **Alternative approach**
   1. Can yield efficient point inference.
   2. Robust to model misspecification.
   3. Compares favourably to exact Bayes.
2. Comparison.
3. Robustness to model misspecification

ABCDF: ABC based on the (empirical) cdf/pdf

- W-ABC like matching order statistics.

- Match empirical cdf/pdf (in a chosen norm).
  - Define: $\|\cdot\|_{\mathcal{P}} : \mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}_+$.
  - Empirical measure: $\hat{\mu}_n(\cdot) = n^{-1} \sum_{i=1}^{n} \delta_{y_i}(\cdot)$,
  - δ_y - Dirac measure.
  - Simulated measure: $\hat{\mu}_{\theta,n}(\cdot) = n^{-1} \sum_{i=1}^{n} \delta_{z_i}(\cdot)$.
  - Replace $\mathcal{W}_p$ by $\|\hat{\mu}_{\theta,n} - \hat{\mu}_n\|_{\mathcal{P}}$:

$$\Pi_\epsilon(A|y_{1:n}) = \frac{\int_A \mathrm{d}\Pi(\theta) \int_{\mathcal{Y}^n} 1\left[\|\hat{\mu}_{\theta,n} - \hat{\mu}_n\|_{\mathcal{P}} \leq \epsilon\right] \mathrm{d}P_\theta^{(n)}(z_{1:n})}{\int_\Theta \mathrm{d}\Pi(\theta) \int_{\mathcal{Y}^n} 1\left[\|\hat{\mu}_{\theta,n} - \hat{\mu}_n\|_{\mathcal{P}} \leq \epsilon\right] \mathrm{d}P_\theta^{(n)}(z_{1:n})}.$$

The Choice of Norm?

- Choice of $\|\cdot\|_{\mathcal{P}}$: **1) efficient, 2) stable** inferences.
- Natural choice for $\|\cdot\|_{\mathcal{P}}$: **robust distances**!
  1. Hellinger (Hell)
  2. Cramer-von Mises (CvM) (projected-averaged for multivariate data).
  3. L_p-variants.
- $\|\cdot\|_{\mathcal{P}}$ = Hell, efficient (CvM if $\theta \in \mathbb{R}$).
- Hell and CvM known to deliver stable inferences.

The Choice of Norm?

- Choice of $\|\cdot\|_{\mathcal{P}}$: **1) efficient, 2) stable** inferences.
- Natural choice for $\|\cdot\|_{\mathcal{P}}$: **robust distances**!
  1. Hellinger (Hell)
  2. Cramer-von Mises (CvM) (projected-averaged for multivariate data).
  3. L_p-variants.
- $\|\cdot\|_{\mathcal{P}}$ = Hell, efficient (CvM if $\theta \in \mathbb{R}$).
- Hell and CvM known to deliver stable inferences.

Outline

- Alternative Likelihood-free approach
  1. Yields efficient inference.
  2. Robust to model misspecification.
  3. Compares favourably with exact Bayesian inference when feasible.

- **Comparison.**

- Robust to model misspecification

Example: Mixture Models

- Two-component mixture model.

- For $\omega \in [0,1]$, $\varphi(\cdot;\mu,\sigma^2)$ normal density, mean μ and variance σ^2,

$$f_\theta(\cdot) := (1-\omega)\varphi(\cdot;\mu,\sigma_1^2) + \omega\varphi(\cdot;-\mu,\sigma_2^2), \quad \theta := (\mu,\omega,\sigma_1,\sigma_2)'.$$

- Generate $n = 100$ observations: $\theta = (-2, 0.5, 1, 1)'$.

- Priors

$$\mu \sim \mathcal{N}(0,1),\ \omega \sim \mathcal{U}[0,1],\ \sigma_1 \sim \mathcal{U}[0,10],\ \sigma_2 \sim \mathcal{U}[0,10].$$

- Compare: exact Bayes (Exact), W-ABC, H-ABC and CvM-ABC.

Example: Mixture Models; $\theta = (-2, 0.5, 1, 1)'$

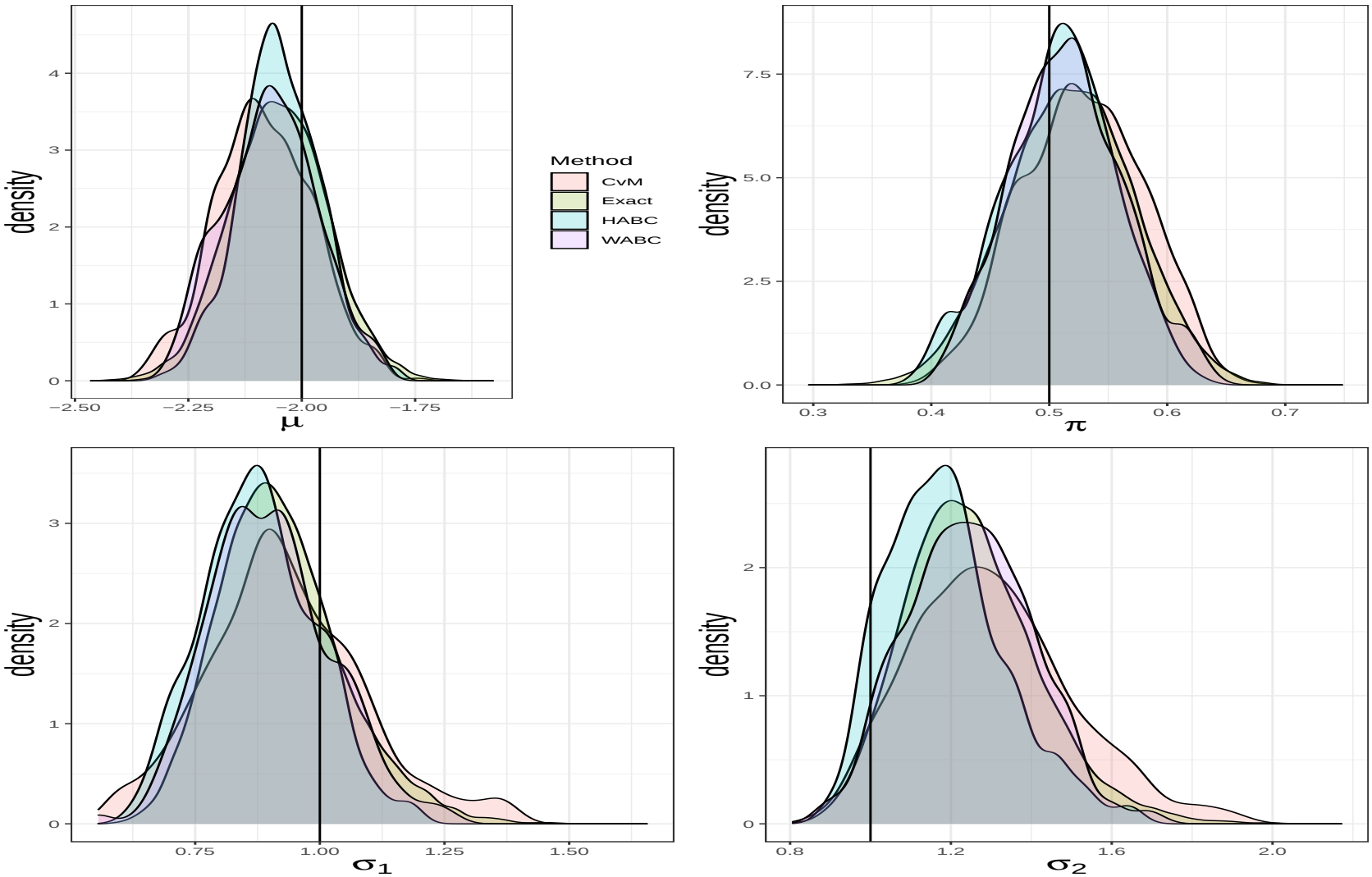


Figure: Posterior marginals in the mixture model example.

Example: Mixture Models; $\theta = (-2, 0.5, 1, 1)'$

Posterior mean (Means), standard deviations (Std), credible set length (Len) and Monte Carlo coverage (Cov). 100 replications.

	Means				Std			
	μ	π	σ_1	σ_2	μ	π	σ_1	σ_2
CvM	-1.96	0.49	1.10	1.11	0.13	0.05	0.21	0.23
Exact	-1.96	0.51	1.08	1.07	0.12	0.06	0.16	0.15
Hell	-1.98	0.49	1.01	1.01	0.09	0.05	0.12	0.12
WABC	-1.93	0.49	1.09	1.09	0.19	0.06	0.27	0.28
	COV				RMSE			
	μ	π	σ_1	σ_2	μ	π	σ_1	σ_2
CvM	88%	92%	92%	96%	0.21	0.06	0.24	0.29
Exact	90%	98%	90%	92%	0.15	0.06	0.20	0.17
Hell	92%	92%	90%	96%	0.10	0.06	0.13	0.12
WABC	100%	96%	100%	100%	0.22	0.06	0.23	0.22

g-and-k Model

- Common test example.
- Quantile function

$$\tau \in (0,1) \mapsto a + b\left(1 + 0.8\frac{1-\exp\{-gz(\tau)\}}{1+\exp\{-gz(\tau)\}}\right)\left\{1+z(\tau)^2\right\}^k z(\tau),$$

  - $z(\tau)$: standard normal quantile.
  - a location; b scale; g skewness; k kurtosis
  - $\theta = (a, b, g, k)$

- Priors

$$a \sim \mathcal{U}[0,10],\ b \sim \mathcal{U}[0,10],\ \ g \sim \mathcal{U}[0,10],\ k \sim \mathcal{U}[0,10],$$

g-and-k Example

- Compare: exact Bayes (Exact), W-ABC and CvM-ABC.
- Generate $n = 100$: $\theta = (3, 1, 2, 0.5)$
- 100 replications.
- ABC: SMC-ABC, $N = 1,024$ particles, 5×10^5 total simulations.
- Exact: MH with numerical estimator of density.

Results

Table: Repeated sampling result for marginal posteriors in the g-and-k model, with $\theta_\star = (3, 1, 2, 0.5)'$.

	Means				Std			
	a	b	g	k	a	b	g	k
CvM	2.99	0.99	2.29	0.55	0.09	0.20	0.71	0.18
Exact	2.89	0.83	2.27	0.54	0.09	0.21	0.32	0.16
WABC	2.89	1.04	5.04	0.45	0.13	0.26	2.62	0.19
	COV				RMSE			
	a	b	g	k	a	b	g	k
CvM	88%	96%	96%	100%	0.12	0.22	0.79	0.15
Exact	100%	100%	100%	100%	0.11	0.18	0.18	0.04
WABC	90%	98%	98%	100%	0.15	0.24	3.06	0.16

Outline

- Alternative Likelihood-free approach
  1. Yields efficient inference.
  2. Robust to model misspecification.
  3. Compares favourably with exact Bayesian inference when feasible.

- Comparison.

- **Robust to model misspecification**
  1. Mixture model.

ABC Under Model Misspecification

1. All models are flawed.

   "Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration."

2. ABC: **heroic** belief that $y_{1:n} \sim P_{\theta^0}$, for $\theta^0 \in \Theta$. **Impact?**

3. ABC: Frazier, Robert and Rousseau (2020, JRSS:B)
   1. Posterior concentration: $(\eta, \|\cdot\|)$-dependent.
   2. Can display **non-standard** asymptotic behavior.

4. BSL: Frazier and Drovandi (2020); Frazier, Drovandi and Nott (2020).
   1. Situation arguably worse for BSL.
   2. Possibly **NO** posterior concentration...

Example: Mixture Model

- Assumed model:

$$f_\theta(\cdot) := (1-\omega)\varphi(\cdot;\mu,\sigma_1^2) + \omega\varphi(\cdot;-\mu,\sigma_2^2), \quad \theta := (\mu,\omega,\sigma_1^2,\sigma_2^2)'.$$

- **BUT**: $y_1,\ldots,y_n$ iid from

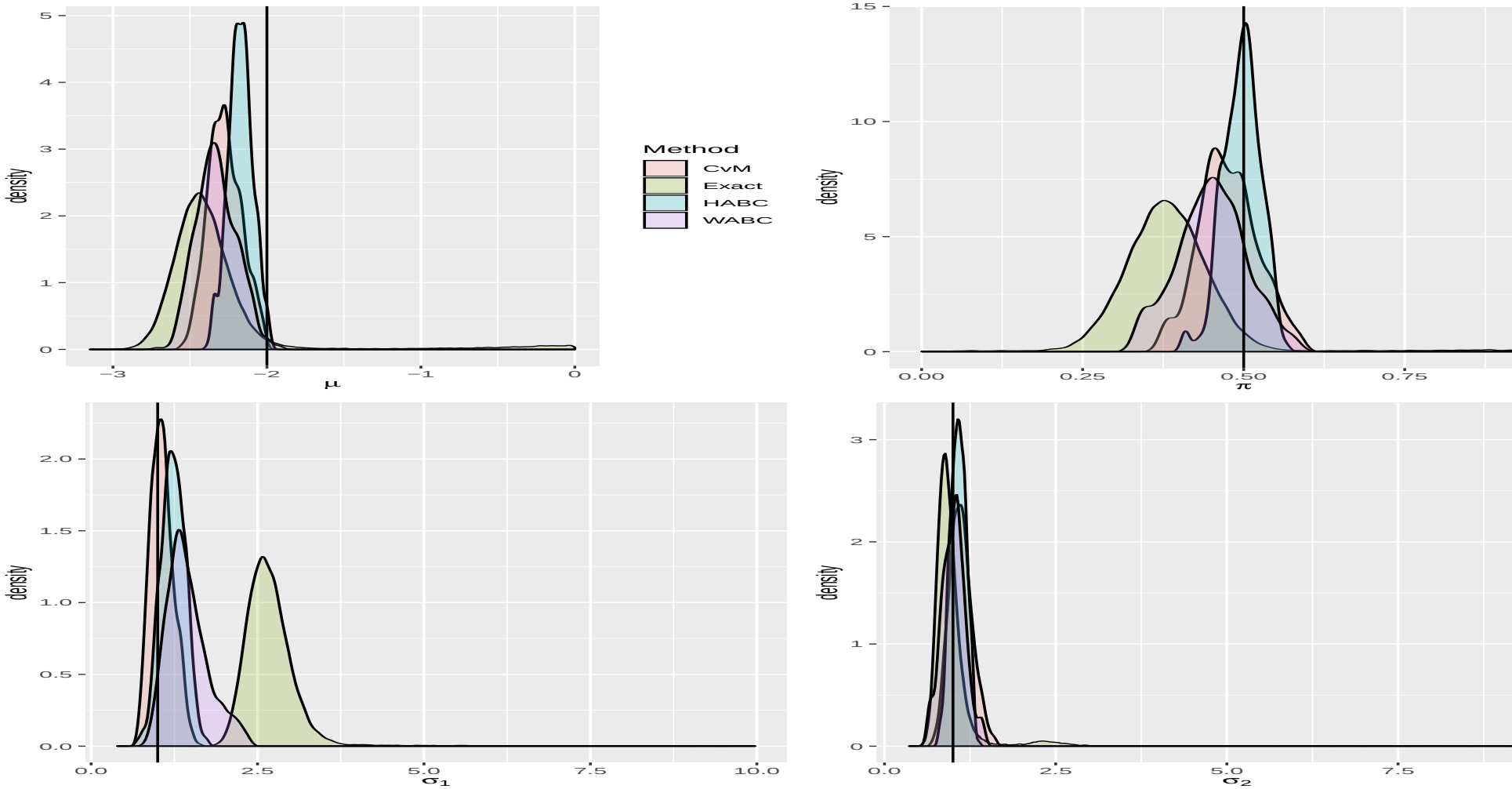
$$f_\star(\cdot) := (1-\alpha)f_\theta(\cdot) + \alpha\varphi(\cdot;\zeta,\nu), \quad \alpha \in [0,1].$$

- α - level of contamination. Impact controlled by - $(\alpha,\zeta,\nu)'$.

Example: Mixture Model

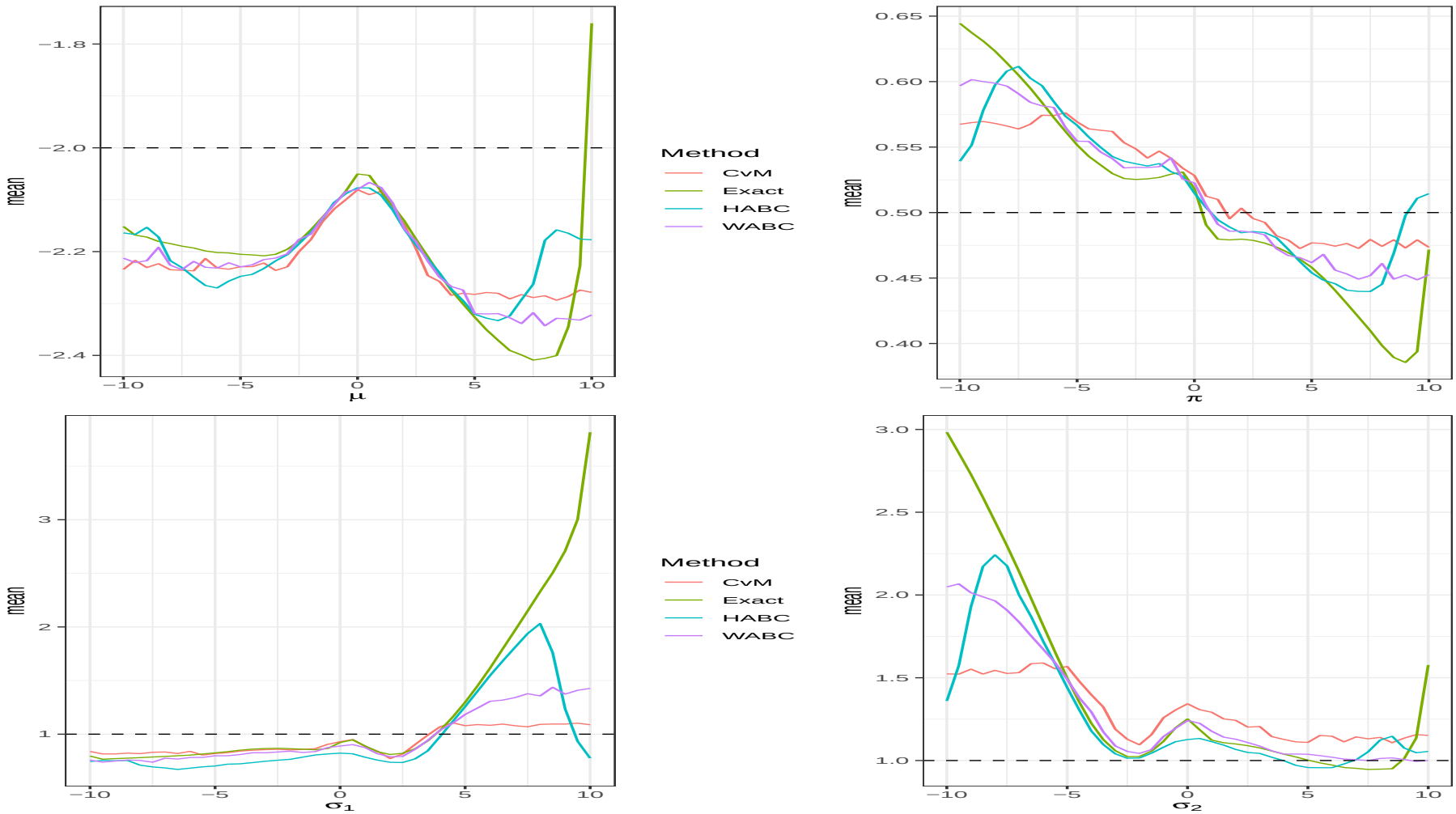
- Compare:
  1. exact Bayes (Exact), W-ABC, H-ABC, CvM-ABC.
  2. Fix $\alpha = 0.05$, $\nu = 0.01$, consider $\zeta \in [-10, 10]$. $\theta = (-2, 0.5, 1, 1)'$.
  3. Sample size $n = 100$,
- First, compare posteriors at $\zeta = 9$.

Example: Mixture Model; $\theta = (-2, 0.5, 1, 1)'$



Example: Mixture Model; $\theta = (-2, 0.5, 1, 1)'$

- Posteriors means: $\zeta \in [-10, 10]$.



Example: Mixture

- Exact Bayes is poor. W-ABC, H-ABC stable.
- CvM-ABC "most stable."
- Compare posterior means, standard deviation, credible set length and coverage,
- 100 replications: $\zeta = 9$, $\alpha = 0.05$.

Example: Mixture; $\theta = (-2, 0.5, 1, 1)'$

	Means				Std			
	μ	π	σ_1	σ_2	μ	π	σ_1	σ_2
CvM	-1.88	0.43	1.61	1.29	0.30	0.08	0.42	0.54
Exact	-1.48	0.42	3.19	1.76	0.45	0.16	0.98	1.11
Hell	-1.74	0.50	1.57	1.33	0.27	0.10	0.67	0.57
WABC	-1.49	0.39	2.31	1.61	0.56	0.13	0.64	0.93
	COV				RMSE			
	μ	π	σ_1	σ_2	μ	π	σ_1	σ_2
CvM	88%	86%	58%	96%	0.37	0.10	0.77	0.58
Exact	76%	78%	02%	84%	0.85	0.14	2.34	1.32
Hell	88%	98%	90%	92%	0.52	0.07	1.18	0.75
WABC	100%	82%	34%	100%	0.63	0.14	1.42	0.79

Theoretical Results

- Posterior concentration as in Frazier, Martin, Robert and Rousseau (2018: Biometrika).

- H-ABC and CvM-ABC, point estimators asymptotically equivalent to

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \|\hat{\mu}_n - \hat{\mu}_{\theta,n}\|_{\mathcal{P}}.$$

  1. **Important**: $\hat{\theta}_n$ - minimax robust against DGPs in

$$\mathcal{B}_d\left(\mu_\star, r/\sqrt{n}\right) = \left\{\mu \in \mathcal{M} : d\left(\mu, \mu_\star\right) \leq r/v_n\right\}.$$

  2. CvM-ABC (resp., H-ABC) robust if DGP in $\mathcal{B}_d$.

  3. H-ABC point estimators efficient (Cramer-Rao).

Conclusions

- Goal: Obtain efficient and stable inferences in complex models.
- Reasonable to assume models are misspecified.
- New approach: mitigate impact of model misspecification, without sacrificing efficiency.
- Compares favourably against
  1. Exact Bayes (likelihood-based).
  2. W-ABC (likelihood-free).
- Theoretical behavior (to be completed):
  1. Correct uncertainty quantification?
  2. Efficiency of point estimators?
  3. Provable robustness?

Conclusions

- Goal: Obtain efficient and stable inferences in complex models.
- Reasonable to assume models are misspecified.
- New approach: mitigate impact of model misspecification, without sacrificing efficiency.
- Compares favourably against
  1. Exact Bayes (likelihood-based).
  2. W-ABC (likelihood-free).
- Theoretical behavior (to be completed):
  1. Correct uncertainty quantification?
  2. Efficiency of point estimators?
  3. Provable robustness?

Conclusions

- Goal: Obtain efficient and stable inferences in complex models.
- Reasonable to assume models are misspecified.
- New approach: mitigate impact of model misspecification, without sacrificing efficiency.
- Compares favourably against
  1. Exact Bayes (likelihood-based).
  2. W-ABC (likelihood-free).
- Theoretical behavior (to be completed):
  1. Correct uncertainty quantification?
  2. Efficiency of point estimators?
  3. Provable robustness?

Conclusions

- Goal: Obtain efficient and stable inferences in complex models.
- Reasonable to assume models are misspecified.
- New approach: mitigate impact of model misspecification, without sacrificing efficiency.
- Compares favourably against
  1. Exact Bayes (likelihood-based).
  2. W-ABC (likelihood-free).
- Theoretical behavior (to be completed):
  1. Correct uncertainty quantification?
  2. Efficiency of point estimators?
  3. Provable robustness?

Future Work

- When to summarize?

- Good summaries may provide better finite-sample inference.
  1. Correctly specified models.
  2. Misspecified models.

Future Work: Example MA(1)

- **Belief:** Observed data $y_{1:n} = (y_1, \ldots, y_n)^\top$ is generated according to

$$y_t = e_t + \theta e_{t-1}, \quad t = 1, \ldots, n,$$

- $\theta \in (-1, 1)$, and $\pi(\theta)$ is uniform

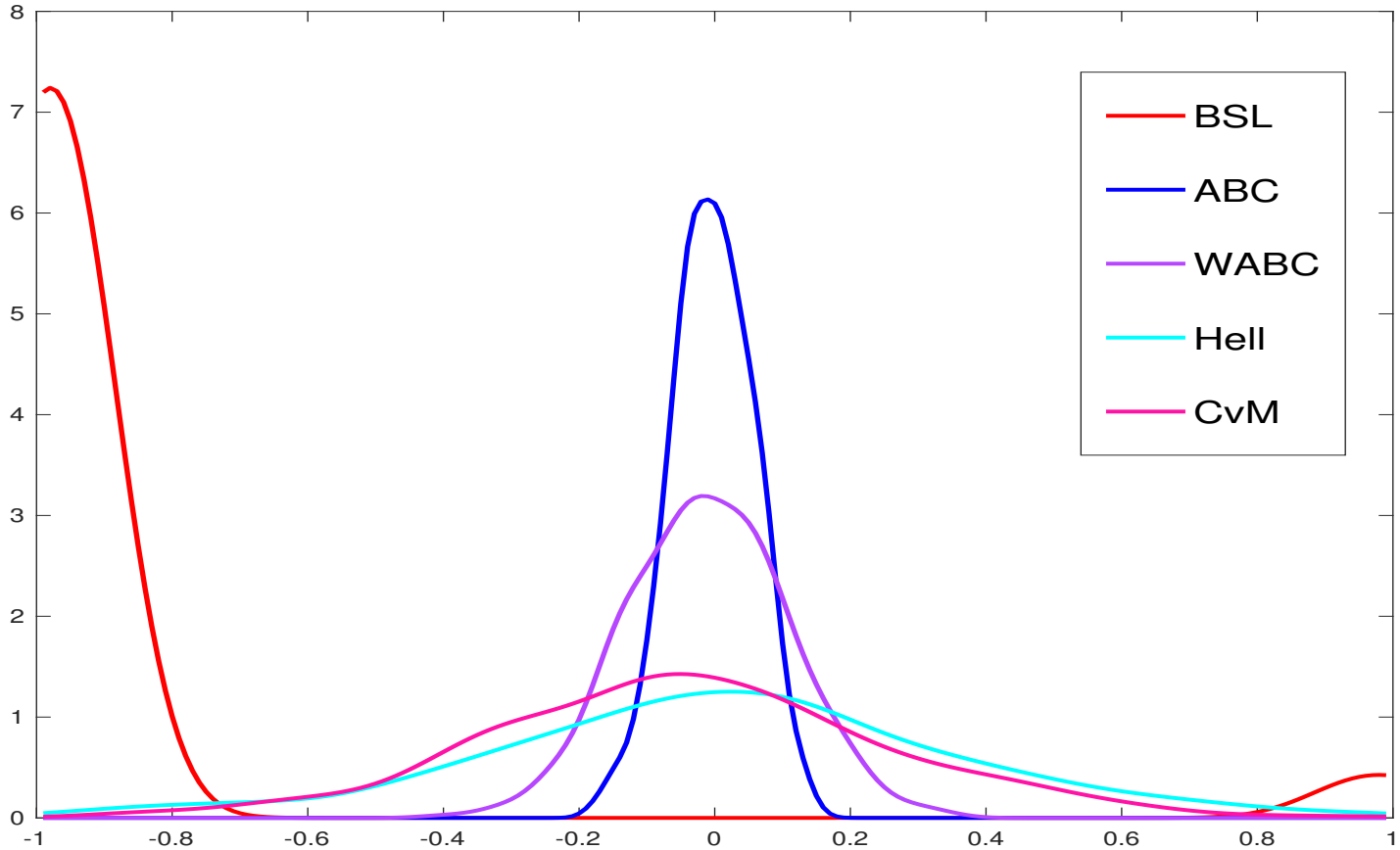
- Summary statistics: sample auto-covariances $\eta_j(y_{1:n}) = \frac{1}{n}\sum_{t=1+j}^{n} y_t y_{t-j}$, for $j \in \{0, 1\}$.

- **Truth:**

$$y_t = \exp(h_t/2)u_t, \quad h_t = \omega + \rho h_{t-1} + v_t \sigma_v,$$

Comparison

$n = 1000$, $\omega = -.736$, $\rho = 0.90$, $\sigma_v = .36$



Thanks!