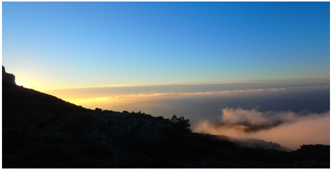
Inference in generative models using the Wasserstein distance

Christian P. Robert
(Paris Dauphine PSL & Warwick U.)
joint work with E. Bernton (Harvard), P.E. Jacob (Harvard), and M. Gerber (Bristol)



Big Bayes, CIRM, Nov. 2018

Outline

Outline

Generative model

Assumption of a data-generating distribution $\mu_\star^{(n)}$ for data

$$y_{1:n} = y_1, \ldots, y_n \in \mathcal{Y}^n$$

Parametric generative model

$$\mathcal{M} = \{\mu_\theta^{(n)} : \theta \in \mathcal{H}\}$$

such that sampling (generating) $z_{1:n}$ from $\mu_\theta^{(n)}$ is feasible

Prior distribution $\pi(\theta)$ available as density *and* generative model

Goal: inference on parameters θ given observations $y_{1:n}$

Generic ABC

Basic (summary-less) ABC posterior with density

$$(\theta, z_{1:n}) \sim \pi(\theta) \frac{\mu_\theta^{(n)} \mathbb{1}\left(\|y_{1:n} - z_{1:n}\| < \varepsilon\right)}{\int_{\mathcal{Y}^n} \mathbb{1}\left(\|y_{1:n} - z_{1:n}\| < \varepsilon\right) dz_{1:n}}.$$

and ABC marginal

$$q^\varepsilon(\theta) = \frac{\int_{\mathcal{Y}^n} \prod_{i=1}^n \mu(dz_i|\theta) \mathbb{1}\left(\|y_{1:n} - z_{1:n}\| < \varepsilon\right)}{\int_{\mathcal{Y}^n} \mathbb{1}\left(\|y_{1:n} - z_{1:n}\| < \varepsilon\right) dz_{1:n}}$$

unbiasedly estimated by $\pi(\theta)\mathbb{1}\left(\|y_{1:n} - z_{1:n}\| < \varepsilon\right)$ where $z_{1:n} \sim \mu_\theta^{(n)}$

Reminder: ABC-posterior goes to posterior as $\varepsilon \to 0$

Summary statistics

Since random variable $\|y_{1:n} - z_{1:n}\|$ may have large variance,

$$\{\|y_{1:n} - z_{1:n}\| < \varepsilon\}$$

gets rare as $\varepsilon \to 0$ and rarer when $d \uparrow$

When using

$$|\eta(y_{1:n}) - \eta(z_{1:n})\| < \varepsilon$$

based on (insufficient) summary statistic η, variance and dimension decrease but q-likelihood differs from likelihood

Arbitrariness and impact of summaries, incl. curse of dimensionality
[X et al., 2011; Fearnhead & Prangle, 2012; Li & Fearnhead, 2016]

Distances between samples

Aim: Ressort to alternate distances $\mathfrak{D}$ between samples $y_{1:n}$ and $z_{1:n}$ such that

$$\mathfrak{D}(y_{1:n}, z_{1:n})$$

has smaller variance than

$$\|y_{1:n} - z_{1:n}\|$$

while induced ABC-posterior still converges to posterior when $\varepsilon \to 0$

ABC with order statistics

Recall that, for univariate i.i.d. data, **order statistics** are sufficient

1. sort observed and generated samples $y_{1:n}$ and $z_{1:n}$
2. compute

$$\|y_{\sigma_y(1:n)} - z_{\sigma_z(1:n)}\|_p = \left(\sum_{i=1}^n |y_{(i)} - z_{(i)}|^p\right)^{1/p}$$

for order p (e.g. 1 or 2) instead of

$$\|y_{1:n} - z_{1:n}\| = \left(\sum_{i=1}^n |y_i - z_i|^p\right)^{1/p}$$

Toy example

- Data-generating process given by

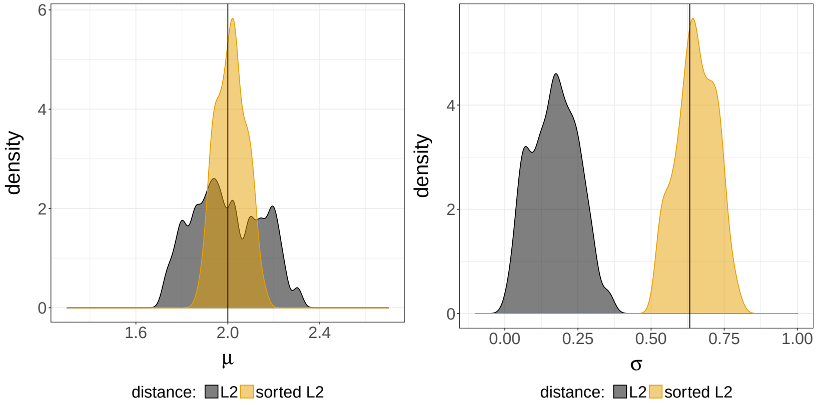
$$Y_{1:1000} \sim \text{Gamma}(10, 5)$$

- Hypothesised model:

$$\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}$$

- Prior $\mu \sim \mathcal{N}(0, 1)$ and $\sigma \sim \text{Gamma}(2, 1)$
- ABC-Rejection sampling: 10^5 draws, using Euclidean distance, on sorted vs. unsorted samples and keeping 10^2 draws with smallest distances

Toy example



density
6
4
2
0
1.6
2.0
2.4
μ
distance: L2 sorted L2
density
4
2
0
0.00
0.25
0.50
0.75
1.00
σ
distance: L2 sorted L2

ABC with transport distances

Distance

$$\mathfrak{D}(y_{1:n}, z_{1:n}) = \left(\frac{1}{n} \sum_{i=1}^{n} |y_{(i)} - z_{(i)}|^p \right)^{1/p}$$

is **p-Wasserstein distance** between empirical cdfs

$$\hat{\mu}_n(dy) = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}(dy) \quad \text{and} \quad \hat{\nu}_n(dy) = \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}(dy)$$

Rather than comparing samples as vectors, alternative representation as empirical distributions

© Novel ABC method, which does not require summary statistics, available with multivariate or dependent data

Outline

Wasserstein distance

Ground distance $\rho\ (x, y) \mapsto \rho(x, y)$ on $\mathcal{Y}$ along with order $p \geq 1$ leads to Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathcal{Y}), p \geq 1$:

$$\mathfrak{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{Y} \times \mathcal{Y}} \rho(x, y)^p d\gamma(x, y) \right)^{1/p}$$

where $\Gamma(\mu, \nu)$ set of joints with marginals μ, ν and $\mathcal{P}_p(\mathcal{Y})$ set of distributions μ for which $\mathbb{E}_\mu[\rho(Y, y_0)^p] < \infty$ for one y_0

Wasserstein distance: univariate case

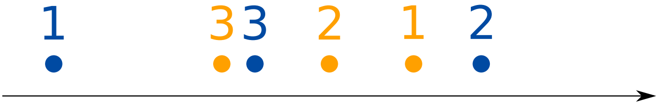
Two empirical distributions on $\mathbb{R}$ with 3 atoms:

$$\frac{1}{3}\sum_{i=1}^{3}\delta_{y_i} \quad \text{and} \quad \frac{1}{3}\sum_{j=1}^{3}\delta_{z_j}$$

Matrix of pair-wise costs:

$$\begin{pmatrix} \rho(y_1,z_1)^p & \rho(y_1,z_2)^p & \rho(y_1,z_3)^p \\ \rho(y_2,z_1)^p & \rho(y_2,z_2)^p & \rho(y_2,z_3)^p \\ \rho(y_3,z_1)^p & \rho(y_3,z_2)^p & \rho(y_3,z_3)^p \end{pmatrix}$$

Wasserstein distance: univariate case



Joint distribution

$$\gamma = \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} & \gamma_{1,3} \\ \gamma_{2,1} & \gamma_{2,2} & \gamma_{2,3} \\ \gamma_{3,1} & \gamma_{3,2} & \gamma_{3,3} \end{pmatrix},$$

with marginals $(1/3 \quad 1/3 \quad 1/3)$, corresponds to a transport cost of

$$\sum_{i,j=1}^{3} \gamma_{i,j} \rho(y_i, z_j)^p$$

Wasserstein distance: univariate case

Optimal assignment:

$$\begin{aligned} y_1 &\longleftrightarrow z_3 \\ y_2 &\longleftrightarrow z_1 \\ y_3 &\longleftrightarrow z_2 \end{aligned}$$

corresponds to choice of joint distribution γ

$$\gamma = \begin{pmatrix} 0 & 0 & 1/3 \\ 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \end{pmatrix},$$

with marginals $(1/3 \quad 1/3 \quad 1/3)$ and cost $\sum_{i=1}^{3} \rho(y_{(i)}, z_{(i)})^p$

Wasserstein distance

Two samples $y_1, \ldots, y_n$ and $z_1, \ldots, z_m$

$$\mathcal{W}_p(\hat{\mu}_n, \hat{\nu}_m) = \frac{1}{nm} \sum_{i,j} \rho(y_, z_j)$$

Important special case when $n = m$, for which solution to the optimization problem $\gamma^\star$ corresponds to an assignment matrix, with only one non-zero entry per row and column, equal to n^{-1}.

[Villani, 2003]

Wasserstein distance

Two samples $y_1, \ldots, y_n$ and $z_1, \ldots, z_m$

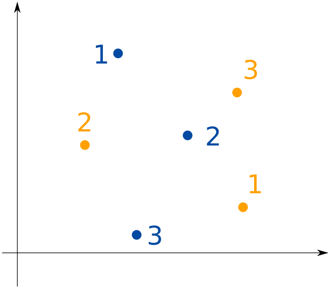
$$\mathcal{W}_p(\hat{\mu}_n, \hat{\nu}_m) = \frac{1}{nm} \sum_{i,j} \rho(y, z_j)$$

Wasserstein distance thus represented as

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

Computing Wasserstein distance between two samples of same size equivalent to optimal matching problem.

Wasserstein distance: bivariate case



there exists a joint distribution γ minimizing cost

$$\sum_{i,j=1}^{3} \gamma_{i,j}\rho(y_i, z_j)^p$$

with various algorithms to compute/approximate it

Wasserstein distance

- also called Vaseršteĭn, Earth Mover, Gini, Mallows, Kantorovich, Rubinstein, &tc.
- can be defined between arbitrary distributions
- actual distance
- statistically sound:

$$\hat{\theta}_n = \operatorname*{arginf}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}, \mu_\theta) \rightarrow \theta_\star = \operatorname*{arginf}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta),$$

at rate $\sqrt{n}$, plus asymptotic distribution

[Bassetti & al., 2006]

Optimal transport to Parliement

Législatives 2017

POLITIQUE | ELECTIONS LÉGISLATIVES 2017

Législatives 2017 : Cédric Villani est élu député de la 5e circonscription de l'Essonne

Le mathématicien est élu face à la candidate Les Républicains Laure Darcos.

LE MONDE | 18.06.2017 à 22h35 • Mis à jour le 19.06.2017 à 00h02

Abonnez vous à partir de 1 € | Réagir | Ajouter | Partager (8 355) | Tweeter

Short bio

Leonid Vaseršteĭn is a Russian-American mathematician, currently Professor of Mathematics at Penn State University. His research is focused on algebra and dynamical systems. He is well known for providing a simple proof of the Quillen–Suslin theorem, a result in commutative algebra, first conjectured by Jean-Pierre Serre in 1955, and then proved by Daniel Quillen and Andrei Suslin in 1976.

Vaseršteĭn got his Master's degree and doctorate in Moscow State University, where he was until 1978. He then moved to Europe and United States.

The Wasserstein metric was named after him by R.L. Dobrushin in 1970.

Outline

Computing Wasserstein distances

- when $\mathcal{Y} = \mathbb{R}$, computing $\mathfrak{W}_p(\mu_n, \nu_n)$ costs $\mathcal{O}(n \log n)$
- when $\mathcal{Y} = \mathbb{R}^d$, exact calculation is $\mathcal{O}(n^3)$ [Hungarian]or $\mathcal{O}(n^{2.5} \log n)$ [short-list]

For entropic regularization, with $\delta > 0$

$$\mathfrak{W}_{p,\delta}(\hat{\mu}_n, \hat{\nu}_n)^p = \inf_{\gamma \in \Gamma(\hat{\mu}_n, \hat{\nu}_n)} \left\{ \int_{\mathcal{Y} \times \mathcal{Y}} \rho(x,y)^p d\gamma(x,y) - \delta \mathrm{H}(\gamma) \right\},$$

where $\mathrm{H}(\gamma) = -\sum_{ij} \gamma_{ij} \log \gamma_{ij}$ entropy of γ, existence of Sinkhorn's algorithm that yields cost $\mathcal{O}(n^2)$

[Genevay et al., 2016]

Computing Wasserstein distances

- other approximations, like Ye et al. (2016) using Simulated Annealing
- regularized Wasserstein not a distance, but as δ goes to zero,
$$\mathfrak{W}_{p,\delta}(\hat{\mu}_n, \hat{\nu}_n) \to \mathfrak{W}_p(\hat{\mu}_n, \hat{\nu}_n)$$
- for δ small enough, $\mathfrak{W}_{p,\delta}(\hat{\mu}_n, \hat{\nu}_n) = \mathfrak{W}_p(\hat{\mu}_n, \hat{\nu}_n)$ (exact)
- in practice, δ 5% of median$(\rho(y_i, z_j)^p)_{i,j}$

[Cuturi, 2013]

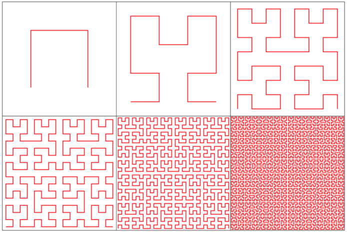
Computing Wasserstein distances

- cost linear in the dimension of observations
- distance calculations model-independent
- other transport distances calculated in $\mathcal{O}(n \log n)$, based on different generalizations of "sorting" (swapping, Hilbert) [Gerber & Chopin, 2015]
- acceleration by combination of distances and subsampling

[source: Wikipedia]

Computing Wasserstein distances

- cost linear in the dimension of observations
- distance calculations model-independent
- other transport distances calculated in $\mathcal{O}(n \log n)$, based on different generalizations of “sorting” (swapping, Hilbert)
  [Gerber & Chopin, 2015]
- acceleration by combination of distances and subsampling



[source: Wikipedia]

Transport distance via Hilbert curve

Sort multivariate data via space-filling curves, like Hilbert space-filling curve

$$H : [0,1] \to [0,1]^d$$

continuous mapping, with pseudo-inverse

$$h : [0,1]^d \to [0,1]$$

Compute order $\sigma \in \mathfrak{S}$ of projected points, and compute

$$\mathfrak{h}_p(y_{1:n}, z_{1:n}) = \left(\frac{1}{n} \sum_{i=1}^{n} \rho(y_{\sigma_y(i)}, z_{\sigma_z(i)})^p \right)^{1/p},$$

called Hilbert ordering transport distance

[Gerber & Chopin, 2015]

Transport distance via Hilbert curve

Fact: $\mathfrak{h}_p(y_{1:n}, z_{1:n})$ is a distance between empirical distributions with n atoms, for all $p \geq 1$

Hence, $\mathfrak{h}_p(y_{1:n}, z_{1:n}) = 0$ if and only if $y_{1:n} = z_{\sigma(1:n)}$, for a permutation σ, with hope to retrieve posterior as $\varepsilon \to 0$

Cost $\mathcal{O}(n \log n)$ *per calculation*, but encompassing sampler might be more costly than with regularized or exact Wasserstein distances

Upper bound on corresponding Wasserstein distance, only accurate for small dimension

Adaptive SMC with r-hit moves

Start with $\varepsilon_0 = \infty$

1. $\forall k \in 1:N$, sample $\theta_0^k \sim \pi(\theta)$ (prior)
2. $\forall k \in 1:N$, sample $z_{1:n}^k$ from $\mu_{\theta^k}^{(n)}$
3. $\forall k \in 1:N$, compute the distance $d_0^k = \mathfrak{D}(y_{1:n}, z_{1:n}^k)$
4. based on $(\theta_0^k)_{k=1}^N$ and $(d_0^k)_{k=1}^N$, compute ε_1, s.t. resampled particles have at least 50% unique values

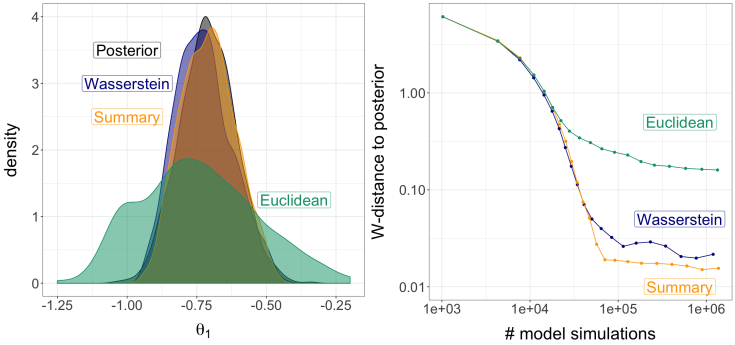
At step $t \geq 1$, weight $w_t^k \propto \mathbb{1}(d_{t-1}^k \leq \varepsilon_t)$, resample, and perform r-hit MCMC with adaptive independent proposals

[Lee, 2012; Lee and Łatuszyński, 2014]

Toy example: bivariate Normal

100 observations from bivariate Normal with variance 1 and covariance 0.55

Compare WABC with ABC versions based on raw Euclidean distance and Euclidean distance between (sufficient) sample means on 10^6 model simulations.



Toy example: bivariate Normal

100 observations from bivariate Normal with variance 1 and covariance 0.55
Compare WABC with ABC versions based on raw Euclidean distance and Euclidean distance between (sufficient) sample means on 10^6 model simulations.
In terms of computing time, based on our R implementation on an Intel Core i7-5820K (3.30GHz), each Euclidean distance calculation takes an average 2.2 x 10^4 s while each Wasserstein distance calculation takes an average 8:2 x 10^3s, i.e. 40 times greater

Quantile "g-and-k" distribution

bivariate extension of the g-and-k distribution with quantile functions

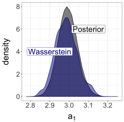
$$a_i + b_i \left(1 + 0.8 \frac{1 - \exp(-g_i z_i(r))}{1 + \exp(-g_i z(r))}\right) \left(1 + z_i(r)^2\right)^k z_i(r) \qquad (1)$$

and correlation ρ

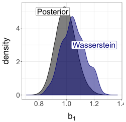
Intractable density that can be numerically approximated [Rayner and MacGillivray, 2002; Prangle, 2017]

Simulation by MCMC and W-ABC (sequential tolerance exploration)

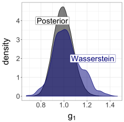
Quantile “g-and-k” distribution



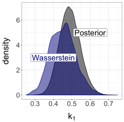
(a) a_1.



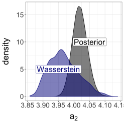
(b) b_1.



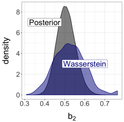
(c) g_1.



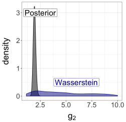
(d) k_1.



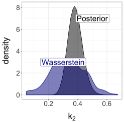
(e) a_2.



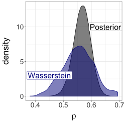
(f) b_2.



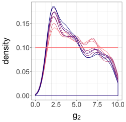
(g) g_2.



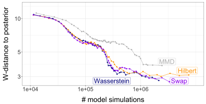
(h) k_2.



(i) ρ.



(j) g_2 last 10 steps.



(k) $\mathfrak{W}_1$ to posterior, vs. simulations.

Outline

Minimum Wasserstein estimator

Under some assumptions, $\hat{\theta}_n$ exists and

$$\limsup_{n\to\infty} \operatorname*{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta) \subset \operatorname*{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta),$$

almost surely

In particular, if $\theta_\star = \operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$ is unique, then

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_\star$$

Minimum Wasserstein estimator

Under stronger assumptions, incl. well-specification, $\dim(\mathcal{Y}) = 1$, and $p = 1$

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \xrightarrow{w} \operatorname*{argmin}_{u \in \mathcal{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_\star(t) \rangle| dt,$$

where $G_\star$ is a $\mu_\star$-Brownian bridge, and $D_\star \in (L_1(\mathbb{R}))^{d_\theta}$ satisfies

$$\int_{\mathbb{R}} |F_\theta(t) - F_\star(t) - \langle \theta - \theta_\star, D_\star(t) \rangle| dt = o(\|\theta - \theta_\star\|_{\mathcal{H}})$$

[Pollard, 1980; del Barrio et al., 1999, 2005]

Hard to use for confidence intervals, but the bootstrap is an intersting alternative.

Toy Gamma example

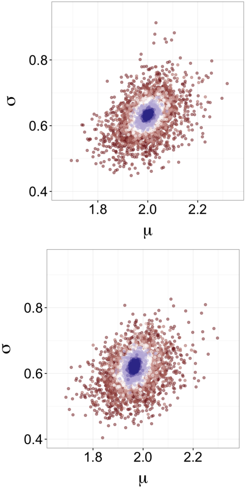
Data-generating process:

$$y_{1:n} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(10, 5)$$

Model:

$$\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}$$

MLE converges to
$\text{argmin}_{\theta \in \mathcal{H}} \text{KL}(\mu_\star, \mu_\theta)$



MLE top, MWE bottom

Toy Gamma example

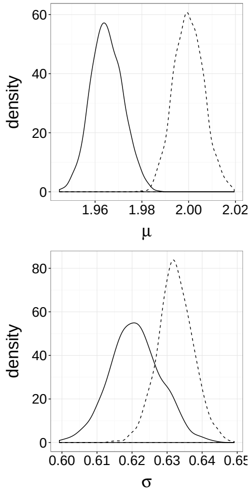
Data-generating process:

$$y_{1:n} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(10, 5)$$

Model:

$$\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}$$

MLE converges to
$\text{argmin}_{\theta \in \mathcal{H}} \text{KL}(\mu_\star, \mu_\theta)$



$n = 10,000$, MLE dashed,
MWE solid

Minimum expected Wasserstein estimator

$$\hat{\theta}_{n,m} = \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathbb{E}[\mathcal{D}(\hat{\mu}_n, \hat{\mu}_{\theta,m})]$$

with expectation under distribution of sample $z_{1:m} \sim \mu_\theta^{(m)}$ giving rise to $\hat{\mu}_{\theta,m} = m^{-1} \sum_{i=1}^m \delta_{z_i}$.

Minimum expected Wasserstein estimator

$$\hat{\theta}_{n,m} = \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathbb{E}[\mathcal{D}(\hat{\mu}_n, \hat{\mu}_{\theta,m})]$$

with expectation under distribution of sample $z_{1:m} \sim \mu_\theta^{(m)}$ giving rise to $\hat{\mu}_{\theta,m} = m^{-1} \sum_{i=1}^m \delta_{z_i}$.

Under further assumptions, incl. $m(n) \to \infty$ with n,

$$\inf_{\theta \in \mathcal{H}} \mathbb{E}\mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \to \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta)$$

and

$$\limsup_{n \to \infty} \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathbb{E}\mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \subset \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta).$$

Minimum expected Wasserstein estimator

$$\hat{\theta}_{n,m} = \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathbb{E}[\mathcal{D}(\hat{\mu}_n, \hat{\mu}_{\theta,m})]$$

with expectation under distribution of sample $z_{1:m} \sim \mu_\theta^{(m)}$ giving rise to $\hat{\mu}_{\theta,m} = m^{-1} \sum_{i=1}^m \delta_{z_i}$.

Further, for n fixed,

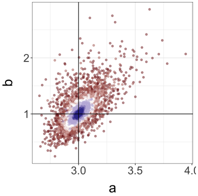
$$\inf_{\theta \in \mathcal{H}} \mathbb{E}\mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \to \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta)$$

as $m \to \infty$ and

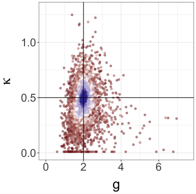
$$\limsup_{m \to \infty} \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathbb{E}\mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \subset \operatorname*{argmin}_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n, \mu_\theta).$$

Quantile “g-and-k” distribution

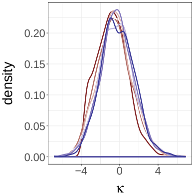
Sampling achieved by plugging standard Normal variables into (1) in place of $z(r)$.
MEWE with large m can be computed to high precision



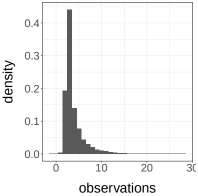
(a) MEWE: a vs b.



(b) MEWE: g vs κ.



(c) $\sqrt{n}$-scaled estim. of κ.



(d) Histogram of data.

Asymptotics of WABC-posterior

- convergence to true posterior as $\epsilon \to 0$
- convergence to non-Dirac as $n \to \infty$ for fixed ϵ
- Bayesian consistency if $\epsilon_n \downarrow \epsilon^\star$ at proper speed

[Frazier, X & Rousseau, 2017]

WARNING: Theoretical conditions extremely rarely open checks in practice

Asymptotics of WABC-posterior

- convergence to true posterior as $\epsilon \to 0$
- convergence to non-Dirac as $n \to \infty$ for fixed ϵ
- Bayesian consistency if $\epsilon_n \downarrow \epsilon^\star$ at proper speed

[Frazier, X & Rousseau, 2017]

WARNING: Theoretical conditions extremely rarely open checks in practice

Asymptotics of WABC-posterior

For fixed n and $\varepsilon \to 0$, for i.i.d. data, assuming

$$\sup_{y,\theta} \mu_\theta(y) < \infty$$

$y \mapsto \mu_\theta(y)$ continuous, the Wasserstein ABC-posterior converges to the posterior irrespective of the choice of ρ and p

Concentration as both $n \to \infty$ and $\varepsilon \to \varepsilon_\star = \inf \mathfrak{W}_p(\mu_\star, \mu_\theta)$ [Frazier et al., 2018]

Concentration on neighborhoods of $\theta_\star = \text{arginf}\ \mathfrak{W}_p(\mu_\star, \mu_\theta)$, whereas posterior concentrates on arginf $\text{KL}(\mu_\star, \mu_\theta)$

Asymptotics of WABC-posterior

Rate of posterior concentration (and choice of ε_n) relates to rate of convergence of the distance, e.g.

$$\mu_\theta^{(n)}\left(\mathfrak{W}_p\left(\mu_\theta, \frac{1}{n}\sum_{i=1}^{n}\delta_{z_i}\right) > u\right) \leq c(\theta) f_n(u),$$

[Fournier & Guillin, 2015]

Rate of convergence decays with the dimension of $\mathcal{Y}$, fast or slow, depending on moments of μ_θ and choice of p

Asymptotics of WABC-posterior

Rate of posterior concentration (and choice of ε_n) relates to rate of convergence of the distance, e.g.

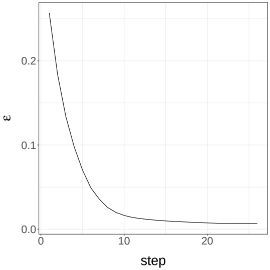
$$\mu_\theta^{(n)}\left(\mathfrak{W}_p\left(\mu_\theta, \frac{1}{n}\sum_{i=1}^{n}\delta_{z_i}\right) > u\right) \leq c(\theta) f_n(u),$$

[Fournier & Guillin, 2015]

Rate of convergence decays with the dimension of $\mathcal{Y}$, fast or slow, depending on moments of μ_θ and choice of p

Toy example: univariate

Data-generating process:
$Y_{1:n} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(10, 5)$, $n = 100$, with mean 2 and standard deviation ≈ 0.63

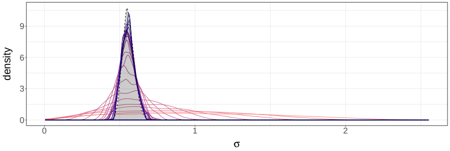


Evolution of ε_t against t, the step index in the adaptive SMC sampler

Theoretical model:

$$\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}$$

Prior: $\mu \sim \mathcal{N}(0, 1)$ and $\sigma \sim \text{Gamma}(2, 1)$



Convergence to posterior

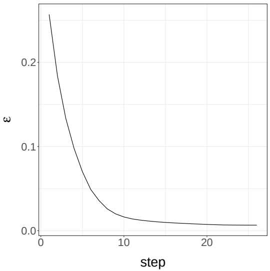
Toy example: univariate

Data-generating process:
$Y_{1:n} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(10, 5)$, $n = 100$, with mean 2 and standard deviation ≈ 0.63

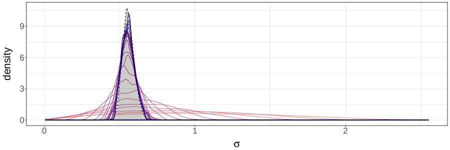
Theoretical model:

$$\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2) : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}$$

Prior: $\mu \sim \mathcal{N}(0, 1)$ and $\sigma \sim \text{Gamma}(2, 1)$



Evolution of ε_t against t, the step index in the adaptive SMC sampler



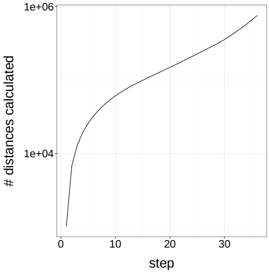
Convergence to posterior

Toy example: multivariate

Observation space: $\mathcal{Y} = \mathbb{R}^{10}$
Model: $Y_i \sim \mathcal{N}_{10}(\theta, S)$, for $i \in 1:100$, where $S_{kj} = 0.5^{|k-j|}$ for $k, j \in 1:10$
Data generated with $\theta_\star$ defined as a 10-vector, chosen by drawing standard Normal variables
Prior: $\theta_i \sim \mathcal{N}(0,1)$ for all $i \in 1:10$

Toy example: multivariate

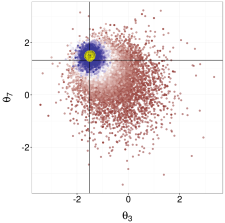
Observation space: $\mathcal{Y} = \mathbb{R}^{10}$
Model: $Y_i \sim \mathcal{N}_{10}(\theta, S)$, for $i \in 1:100$, where $S_{kj} = 0.5^{|k-j|}$ for $k, j \in 1:10$
Data generated with $\theta_\star$ defined as a 10-vector, chosen by drawing standard Normal variables
Prior: $\theta_i \sim \mathcal{N}(0, 1)$ for all $i \in 1:10$



Evolution of number of distances calculated up to t, step index in adaptive SMC sampler

Toy example: multivariate

Observation space: $\mathcal{Y} = \mathbb{R}^{10}$
Model: $Y_i \sim \mathcal{N}_{10}(\theta, S)$, for $i \in 1:100$, where $S_{kj} = 0.5^{|k-j|}$ for $k, j \in 1:10$
Data generated with $\theta_\star$ defined as a 10-vector, chosen by drawing standard Normal variables
Prior: $\theta_i \sim \mathcal{N}(0,1)$ for all $i \in 1:10$

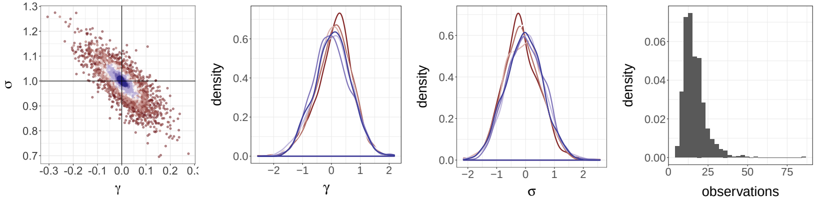


Bivariate marginal of (θ_3, θ_7) approximated by SMC sampler *(posterior contours in yellow, $\theta_\star$ indicated by black lines)*

sum of log-Normals

Distribution of the sum of log-Normal random variables intractable but easy to simulate

$$x_1, \ldots, x_L \sim \mathcal{N}(\gamma, \sigma^2) \qquad y = \sum_{\ell=1}^{L} \exp(x_\ell)$$



(a) MEWE of (γ, σ).

(b) $\sqrt{n}$-scaled estim. of γ.

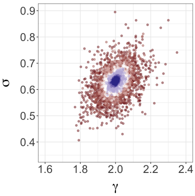
(c) $\sqrt{n}$-scaled estim. of σ.

(d) Histogram of data.

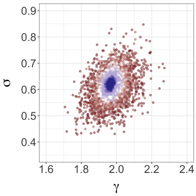
misspecified model

Gamma Gamma$(10, 5)$ data fitted with a Normal model $\mathcal{N}(\gamma, \sigma^2)$
approximate MEWE by sampling $k = 20$ independent $u^{(i)}$ and minimize

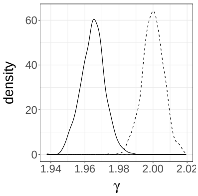
$$\theta \mapsto k^{-1} \sum_{i=1}^{k} \mathcal{W}_p(y_{1:n}, g_m(u^{(i)}, \theta))$$



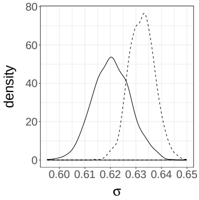
(a) MLE of (γ, σ).



(b) MEWE of (γ, σ).



(c) Estimators of γ.



(d) Estimators of σ.

Outline

Method 1 (0?): ignoring dependencies

Consider only marginal distribution

AR(1) example:

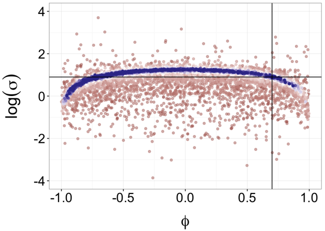
$$y_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right), \quad y_{t+1} \sim \mathcal{N}\left(\phi y_t, \sigma^2\right)$$

Marginally

$$y_t \sim \mathcal{N}\left(0, \sigma^2 / \left(1-\phi^2\right)\right)$$

which identifies $\sigma^2/\left(1-\phi^2\right)$ but not (ϕ, σ)

Produces a region of plausible parameters



For $n = 1,000$, generated with $\phi_\star = 0.7$ and $\log \sigma_\star = 0.9$

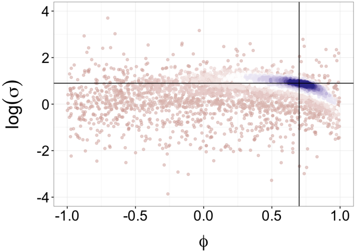
Method 2: delay reconstruction

Introduce $\tilde{y}_t = (y_t, y_{t-1}, \ldots, y_{t-k})$ for lag k, and treat $\tilde{y}_t$ as data

AR(1) example: $\tilde{y}_t = (y_t, y_{t-1})$ with marginal distribution

$$\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \frac{\sigma^2}{1-\phi^2}\begin{pmatrix}1 & \phi\\ \phi & 1\end{pmatrix}\right),$$

identifies both ϕ and σ
Related to Takens' theorem in dynamical systems literature



For $n = 1,000$, generated with $\phi_\star = 0.7$ and $\log \sigma_\star = 0.9$.

Method 3: residual reconstruction

Time series $y_{1:n}$ deterministic transform of θ and $w_{1:n}$

Given $y_{1:n}$ and θ, reconstruct $w_{1:n}$

Cosine example:

$$\begin{aligned} y_t &= A\cos(2\pi\omega t + \phi) + \sigma w_t \\ w_t &\sim \mathcal{N}(0,1) \\ w_t &= (y_t - A\cos(2\pi\omega t + \phi))/\sigma \end{aligned}$$

and calculate distance between reconstructed $w_{1:n}$ and Normal sample

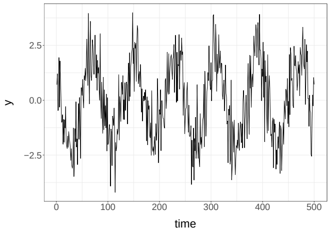
[Mengersen et al., 2013]

Method 3: residual reconstruction

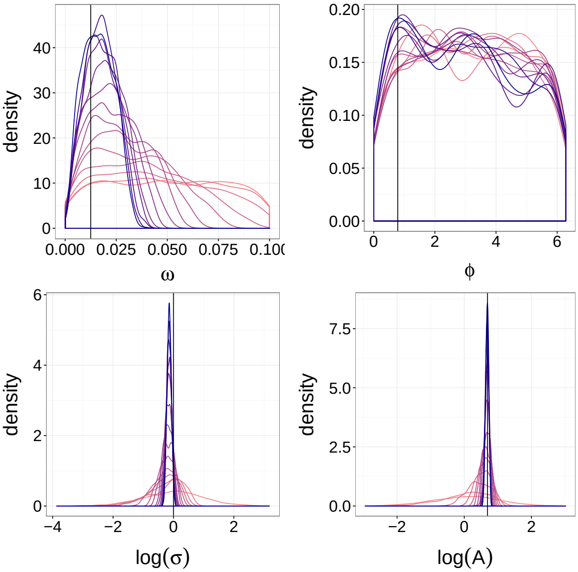
Time series $y_{1:n}$ deterministic transform of θ and $w_{1:n}$

Given $y_{1:n}$ and θ, reconstruct $w_{1:n}$

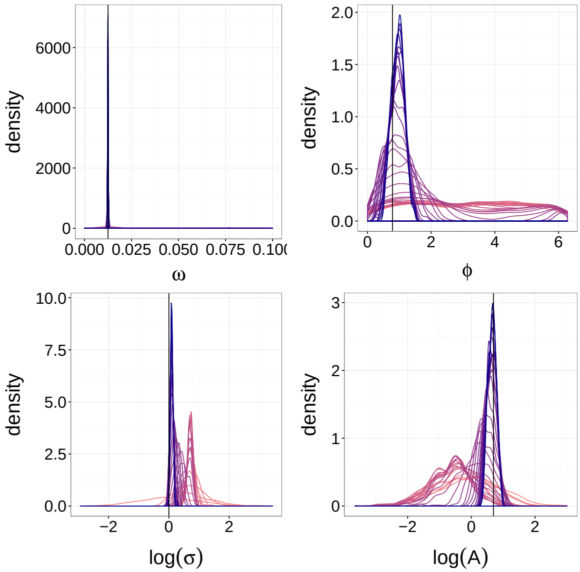
Cosine example:
$y_t = A\cos(2\pi\omega t + \phi) + \sigma w_t$
$n = 500$ observations with
$\omega_\star = 1/80$, $\phi_\star = \pi/4$,
$\sigma_\star = 1, A_\star = 2$, under prior
$\mathcal{U}[0, 0.1]$ and $\mathcal{U}[0, 2\pi]$ for ω and ϕ,
and $\mathcal{N}(0, 1)$ for $\log\sigma, \log A$



Cosine example with delay reconstruction, $k = 3$



and with residual and delay reconstructions, $k = 1$



Method 4: curve matching

Define $\tilde{y}_t = (t, y_t)$ for all $t \in 1:n$.

Define a metric on $\{1, \ldots, T\} \times \mathcal{Y}$. e.g.
$\rho((t, y_t), (s, z_s)) = \lambda|t - s| + |y_t - z_s|$, for some λ

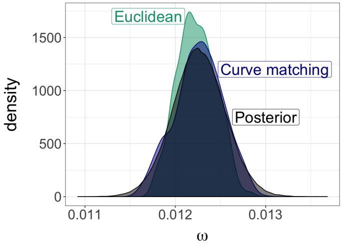
Use distance $\mathcal{D}$ to compare $\tilde{y}_{1:n} = (t, y_t)_{t=1}^n$ and $\tilde{z}_{1:n} = (s, z_s)_{s=1}^n$

If $\lambda \gg 1$, optimal transport will associate each (t, y_t) with (t, z_t)
We get back the "vector" norm $\|y_{1:n} - z_{1:n}\|$.

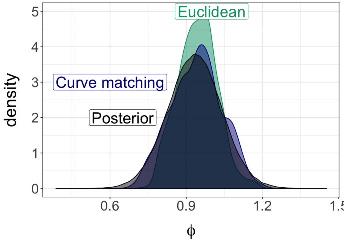
If $\lambda = 0$, time indices are ignored: identical to Method 1

For any $\lambda > 0$, there is hope to retrieve the posterior as $\varepsilon \to 0$

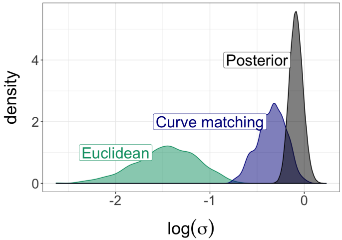
Cosine example



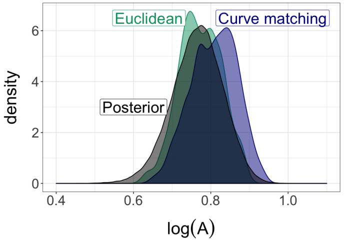
(a) Posteriors of ω.



(b) Posteriors of ϕ.



(c) Posteriors of $\log(\sigma)$.



(d) Posteriors of $\log(A)$.

Discussion

- Transport metrics can be used to compare samples
  Various complexities from $n^3 \log n$ to n^2 to $n \log n$

- Asymptotic guarantees as $\varepsilon \to 0$ for fixed n,
  and as $n \to \infty$ and $\varepsilon \to \varepsilon_\star$

- Various ways of applying these ideas to time series and
  maybe spatial data, maps, images...