4 - Beyond variational inference

S. Robin

INRAE / AgroParisTech / univ. Paris-Saclay
Muséum National d’Histoire Naturelle

Winter School on Mathematical Statistics, Luxembourg, Dec’20
1 – Models with latent variables in ecology (statistical ecology)

2 – Variational inference for incomplete data models (statistics)

3 – Variational inference for species abundances and network models (statistical ecology)

4 – Beyond variational inference (statistics)
Part 4

Algorithmic improvements

Guaranties about variational estimates

Combining variational inference with ...
  Frequentist inference
  Bayesian inference

Conclusion (?)
Outline

Algorithmic improvements

Guarantees about variational estimates

Combining variational inference with ...
  - Frequentist inference
  - Bayesian inference

Conclusion (?)
Algorithmic improvements

Borrowed from many fields.

► Optimization: generic stochastic gradient descent (#21) or more dedicated approaches [HBWP13]

► Bayesian inference: Variational tempering [MMA⁺16]

► Machine learning: Variational autoencoders [KW14,KW19]
  → use neural networks to learn the variational parameters with more flexibility
Outline

Algorithmic improvements

Guarantees about variational estimates

Combining variational inference with ...
  Frequentist inference
  Bayesian inference

Conclusion (?)
Statistical guarantees: *no big picture*

Accuracy of variational estimates.

- Most often assessed empirically (numerical simulations) see e.g. #22
Statistical guarantees: *no big picture*

Accuracy of variational estimates.

- Most often assessed empirically (numerical simulations) see e.g. #22

'Negative' results.

- VEM estimates $\neq$ stationary point of the likelihood function \[GB05\]
- Too small posterior variance provided by variational Bayes \[WT05,MT07,CM07\]
Statistical guarantees: *no big picture*

Accuracy of variational estimates.

- Most often assessed empirically (numerical simulations) see e.g. #22

'Negative' results.

- VEM estimates $\neq$ stationary point of the likelihood function [GB05]
- Too small posterior variance provided by variational Bayes [WT05, MT07, CM07]

Balanced results.

- Mean-field approximation provides consistent estimates (binary SBM affiliation: [ZZ20])
- Naive implementation may yield instabilities [GJM19, ZZ20]
Statistical guarantees: no big picture

Accuracy of variational estimates.
- Most often assessed empirically (numerical simulations) see e.g. #22

'Negative' results.
- VEM estimates $\neq$ stationary point of the likelihood function [GB05]
- Too small posterior variance provided by variational Bayes [WT05,MT07,CM07]

Balanced results.
- Mean-field approximation provides consistent estimates (binary SBM affiliation: [ZZ20])
- Naive implementation may yield instabilities [GJM19,ZZ20]

Positive results.
- Some results for specific models [HOW11]
- Some attempts for a general theory via $M$-estimation [WM19]
- Most studied case: mean-field VEM binary stochastic block-model (see next)
Binary stochastic block-model

A series of results: [CDP12, BCCZ13, MM15, ZZ20]

- Consistency of variational estimates
- Asymptotic normality of variational estimates
- Class recovery (node classification, including LBM)

Why does it work? Theorem 3.1 in [CDP12] states that

\[
\sum_{z \neq z^*} p_\theta(Z = z | Y) \leq O(n^{-\kappa})
\]

uniformly in $z^*$, with $\kappa = \kappa(\theta)$.

- Intuition: $p_\theta(Z | Y)$ is asymptotically Dirac, which belongs to $Q = Q_{\text{fact}}$.

- The 'largest gap' algorithm [CDR12] takes advantage of a similar concentration.
Binary stochastic block-model

A series of results: [CDP12, BCCZ13, MM15, ZZ20]

▶ Consistency of variational estimates
▶ Asymptotic normality of variational estimates
▶ Class recovery (node classification, including LBM)

Why does it work? Theorem 3.1 in [CDP12] states that

\[
P \left( \sum_{z \neq z^*} \frac{p_\theta(Z = z \mid Y)}{p_\theta(Z = z^* \mid Y)} > t \right) = O \left( n e^{-\kappa nt} \right)
\]

uniformly in \( z^* \), with \( \kappa = \kappa(\theta) \).
Binary stochastic block-model

A series of results: [CDP12,BCCZ13,MM15,ZZ20]

- Consistency of variational estimates
- Asymptotic normality of variational estimates
- Class recovery (node classification, including LBM)

Why does it work? Theorem 3.1 in [CDP12] states that

$$P\left( \sum_{z \neq z^*} \frac{p_\theta(Z = z \mid Y)}{p_\theta(Z = z^* \mid Y)} > t \right) = O\left(ne^{-\kappa nt}\right)$$

uniformly in $z^*$, with $\kappa = \kappa(\theta)$.

- Intuition: $p_\theta(Z \mid Y)$ is asymptotically Dirac, which belongs to $Q = Q_{fact}$.
- The 'largest gap' algorithm [CDR12] takes advantage of a similar concentration #23
- The proofs do not easily adapt to other VEM
Outline

Algorithmic improvements

Guaranties about variational estimates

Combining variational inference with ...
  Frequentist inference
  Bayesian inference

Conclusion (?)
Frequentist inference

Maximum likelihood inference.

\[ \hat{\theta} \text{MLE} = \arg \max_{\theta} \log p_{\theta}(Y) \]

is intractable because the likelihood involves an integration over the latent variables:

SBM: \[ \log p_{\theta}(Y) = \log \left( \sum_{Z \in [K]} \prod_{i} p_{\pi}(Z_i) \prod_{i,j} p_{\alpha,\beta}(Y_{ij}|Z_i, Z_j) \right) \]

The (log-)likelihood is far from being the only admissible estimation function; for example, think of M-estimation.
Frequentist inference

Maximum likelihood inference.

\[ \hat{\theta}_{\text{MLE}} = \text{arg max}_\theta \log p_\theta (Y) \]

is intractable because the likelihood involves an integration over the latent \(Z\):

**PLN:**

\[
\log p_\theta (Y) = \sum_i \log \left( \int_{R^p} p_\Sigma (Z_i) \prod_j p_\beta (Y_{ij} | Z_{ij}) \, dZ_i \right)
\]

**SBM:**

\[
\log p_\theta (Y) = \log \left( \sum_{Z \in [K]^n} \prod_i p_\pi (Z_i) \prod_{i,j} p_\alpha,\beta (Y_{ij} | Z_i, Z_j) \right)
\]
Frequentist inference

Maximum likelihood inference.

\[ \hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \log p_{\theta}(Y) \]

is intractable because the likelihood involves an integration over the latent \( Z \)

\[
\text{PLN: } \log p_{\theta}(Y) = \sum_{i} \log \left( \int_{\mathbb{R}^{p}} p_{\Sigma}(Z_{i}) \prod_{j} p_{\beta}(Y_{ij} | Z_{ij}) \, dZ_{i} \right)
\]

\[
\text{SBM: } \log p_{\theta}(Y) = \log \left( \sum_{Z \in [K]^{n}} \prod_{i} p_{\pi}(Z_{i}) \prod_{i,j} p_{\alpha,\beta}(Y_{ij} | Z_{i}, Z_{j}) \right)
\]

The (log-)likelihood is far from being the only admissible estimation function

→ think, e.g., of \( M \)-estimation
Composite likelihood

Sum of partial likelihoods:

PLN: \[ \hat{\theta}_{CL} = \arg \max_{\theta} \sum_{i} \sum_{j,k} \log p_{\theta}(Y_{ij}, Y_{ik}) \] only requires \( \int_{\mathbb{R}^2} \)

SBM: \[ \hat{\theta}_{CL} = \arg \max_{\theta} \sum_{i,j,k} \log p_{\theta}(Y_{ij}, Y_{ik}, Y_{jk}) \] only requires \( \sum_{Z \in [K]^3} \)

→ Generic results (consistency, asymptotic normality) exist for \( \hat{\theta}_{CL} \) [VRF11] + see [AM12] for binary SBM
Composite likelihood

Sum of partial likelihoods:

\[
\text{PLN: } \hat{\theta}_{CL} = \arg \max_{\theta} \sum_i \sum_{j,k} \log p_{\theta}(Y_{ij}, Y_{ik}) \quad \text{only requires } \int_{\mathbb{R}^2}
\]

\[
\text{SBM: } \hat{\theta}_{CL} = \arg \max_{\theta} \sum_{i,j,k} \log p_{\theta}(Y_{ij}, Y_{ik}, Y_{jk}) \quad \text{only requires } \sum_{Z \in [K]^3}
\]

\[\rightarrow \text{ Generic results (consistency, asymptotic normality) exist for } \hat{\theta}_{CL} \quad [\text{VRF11}] + \text{ see } [\text{AM12}] \text{ for binary SBM}\]

Practical implementation.

- EM algorithms can be designed to maximize composite likelihoods
- Getting \( \hat{\theta}_{CL} \) is still demanding (many terms in the sum: \( np^2 \) for PLN, \( n^3 \) for SBM)
- \( \hat{\theta}_{VEM} \) usually provides a (very) good starting point
Bayesian inference
Bayesian inference

Reminder.

- **Prior:** $p(\theta)$
- **Latent:** $p(Z \mid \theta)$
- **Observed:** $p(Y \mid Z, \theta)$
- **Posterior:**

$$p(\theta, Z \mid Y) = \frac{p(\theta) \ p(Z \mid \theta) \ p(Y \mid \theta, Z)}{p(Y)}$$

\[
\begin{align*}
\theta_{PLN} &= (\beta, \Sigma), \\
\theta_{SBM} &= (\pi, \alpha, \beta)
\end{align*}
\]
Bayesian inference

Reminder.
- Prior: $p(\theta)$
- Latent: $p(Z \mid \theta)$
- Observed: $p(Y \mid Z, \theta)$
- Posterior:

$$p(\theta, Z \mid Y) = \frac{p(\theta) \ p(Z \mid \theta) \ p(Y \mid \theta, Z)}{p(Y)}$$

Sampling methods.

$\theta_{PLN} = (\beta, \Sigma), \quad \theta_{SBM} = (\pi, \alpha, \beta)$
Bayesian inference

Reminder.

- **Prior**: $p(\theta)$
  - $\theta_{PLN} = (\beta, \Sigma)$, $\theta_{SBM} = (\pi, \alpha, \beta)$
- **Latent**: $p(Z | \theta)$
- **Observed**: $p(Y | Z, \theta)$
- **Posterior**:
  \[
p(\theta, Z | Y) = \frac{p(\theta) \ p(Z | \theta) \ p(Y | \theta, Z)}{p(Y)}
  \]

Sampling methods.

- **Monte-Carlo**: sample $(\theta^b, Z^b) \overset{iid}{\sim} p(\theta, Z | Y)$
Bayesian inference

Reminder.

- Prior: $p(\theta)$
- Latent: $p(Z \mid \theta)$
- Observed: $p(Y \mid Z, \theta)$
- Posterior:

$$p(\theta, Z \mid Y) = \frac{p(\theta) p(Z \mid \theta) p(Y \mid \theta, Z)}{p(Y)}$$

Sampling methods.

- Monte-Carlo: sample $(\theta^b, Z^b) \overset{iid}{\sim} p(\theta, Z \mid Y)$
- MCMC: construct a Markov chain with $p(\theta, Z \mid Y)$ as a stationary distribution
Bayesian inference

Reminder.

- Prior: $p(\theta)$
- Latent: $p(Z | \theta)$
- Observed: $p(Y | Z, \theta)$
- Posterior:

$$p(\theta, Z | Y) = \frac{p(\theta) p(Z | \theta) p(Y | \theta, Z)}{p(Y)}$$

Sampling methods.

- Monte-Carlo: sample $(\theta^b, Z^b) \sim p(\theta, Z | Y)$
- MCMC: construct a Markov chain with $p(\theta, Z | Y)$ as a stationary distribution
- Importance sampling: $(\theta^b, Z^b) \sim q(\theta, Z)$ and reweight each draw with weight

$$w^b = \frac{p(\theta^b, Z^b | Y)}{q(\theta^b, Z^b)}$$
Combining variational inference with ... Bayesian inference

Bayesian inference

Reminder.

- **Prior:** $p(\theta)$
  \[ \theta_{PLN} = (\beta, \Sigma), \quad \theta_{SBM} = (\pi, \alpha, \beta) \]

- **Latent:** $p(Z \mid \theta)$

- **Observed:** $p(Y \mid Z, \theta)$

- **Posterior:**
  \[ p(\theta, Z \mid Y) = \frac{p(\theta) p(Z \mid \theta) p(Y \mid \theta, Z)}{p(Y)} \]

Sampling methods.

- **Monte-Carlo:** sample $(\theta^b, Z^b) \overset{iid}{\sim} p(\theta, Z \mid Y)$

- **MCMC:** construct a Markov chain with $p(\theta, Z \mid Y)$ as a stationary distribution

- **Importance sampling:** $(\theta^b, Z^b) \overset{iid}{\sim} q(\theta, Z)$ and reweight each draw with weight
  \[ w^b = \frac{p(\theta^b, Z^b \mid Y)}{q(\theta^b, Z^b)} \]

- **Sequential Monte-Carlo:** construct a sequence of distribution going from $q(\theta, Z)$ to $p(\theta, Z \mid Y)$
Sequential Monte-Carlo sampling

Principle. [DDJ06] $U = (\theta, Z)$
Sequential Monte-Carlo sampling

Principle. [DDJ06] \( U = (\theta, Z) \)

- given \( p_{\text{start}}(U) \)
Sequential Monte-Carlo sampling

Principle. [DDJ06] \( U = (\theta, Z) \)

- given \( p_{\text{start}}(U) \)
- aiming at \( p_{\text{target}}(U) = p(U \mid Y) \)
Sequential Monte-Carlo sampling

Principle. [DDJ06] $U = (\theta, Z)$

- given $p_{\text{start}}(U)$
- aiming at $p_{\text{target}}(U) = p(U \mid Y)$
- sample from a sequence of distributions $p_{\text{start}} = p_0, p_1, \ldots, p_{H-1}, p_H = p_{\text{target}}$

with

$$p_h(U) \propto p_{\text{start}}(U)^{1-\rho_h} p_{\text{target}}(U)^{\rho_h}$$

and $0 = \rho_0 < \rho_1 < \cdots < \rho_H = 1$

see #24 for tuning of the $\rho_h$
Sequential Monte-Carlo sampling

Principle. \[ DDJ06 \] \( U = (\theta, Z) \)

- given \( p_{\text{start}}(U) \)
- aiming at \( p_{\text{target}}(U) = p(U \mid Y) \)
- sample from a sequence of distributions

\[ p_{\text{start}} = p_0, p_1, \ldots, p_{H-1}, p_H = p_{\text{target}} \]

with

\[ p_h(U) \propto p_{\text{start}}(U)^{1-\rho_h} p_{\text{target}}(U)^{\rho_h} \]

and \( 0 = \rho_0 < \rho_1 < \cdots < \rho_H = 1 \)

Most often: \( p_{\text{start}} = p_{\text{prior}} \) (long way to the posterior)

**VBEM**: directly use \( p_{\text{start}} = p_{\text{VBEM}} \)

**VEM**: use (approximate) Louis formulas [Lou82] to derive \( p_{\text{start}} = p_{\text{VEM}} \) [DR19]

see #24 for tuning of the \( \rho_h \)
Back to the tree interaction network

$Y_{ij} =$ number of shared parasites
$x_{ij} =$ taxonomic distance
$Y_{ij} \sim \mathcal{P}(\exp(x_{ij}^\top \beta + \alpha z_i z_j))$

Estimates:

$\hat{K}_{ICL} = 4 \quad \hat{\beta} = -0.317$

- Taxonomy (partially) explains the links (smaller $\hat{K}$)
- Distant species share less parasites ($\hat{\beta} < 0$)
- The remaining structure is not related to taxonomy
Tree network: model selection

Model selection.
- Number of groups $K$
- Set $S$ of relevant covariates: $S \subset \{\text{taxonomy, geography, phylogeny}\}$
Combining variational inference with Bayesian inference

Tree network: model selection

Model selection.

- Number of groups $K$
- Set $S$ of relevant covariates: $S \subset \{\text{taxonomy, geography, phylogeny}\}$

Choosing $K$ for a given $S$:

$$p(K \mid Y, S) \propto p(Y \mid S, K)$$

here: $S = (\text{taxonomy, geography})$

Averaging over $K$: #26
Tree network: model selection

Model selection.
- Number of groups $K$
- Set $S$ of relevant covariates: $S \subset \{\text{taxonomy, geography, phylogeny}\}$

Choosing $K$ for a given $S$:

$$p(K \mid Y, S) \propto p(Y \mid S, K)$$

here: $S = (\text{taxonomy, geography})$

Averaging over $K$: #26

Variable selection. $p(S \mid Y) = \sum_K p(S, K \mid Y)$

$$P\{x = (\text{taxo., geo.}) \mid Y\} \simeq 70\%, \quad P\{x = (\text{taxo.}) \mid Y\} \simeq 30\%$$
Tree network: significance

Parameter posterior distribution for \( S = \) (taxonomy, geography, phylogeny):

- **taxonomy**
- **geography**
- **phylogeny**

Legend: \( q_{VEM}(\beta_j) \), \( p(\beta_j | S, \hat{K}(S), Y) \), \( p(\beta_j | S, Y) \)
Tree network: significance

Parameter posterior distribution for $S = (\text{taxonomy, geography, phylogeny})$:

<table>
<thead>
<tr>
<th></th>
<th>taxonomy</th>
<th>geography</th>
<th>phylogeny</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{q}_{VEM}(\beta_j)$</td>
<td>p($\beta_j \mid S, \hat{K}(S), Y$)</td>
<td>p($\beta_j \mid S, Y$)</td>
<td></td>
</tr>
</tbody>
</table>

Legend: $\text{q}_{VEM}(\beta_j), p(\beta_j \mid S, \hat{K}(S), Y), p(\beta_j \mid S, Y)$

Why so many steps to go from $\text{q}_{VEM}(\beta_j)$ to $p(\beta_j \mid Y)$?

Correlation between estimates.

<table>
<thead>
<tr>
<th></th>
<th>($\beta_1, \beta_2$)</th>
<th>($\beta_1, \beta_3$)</th>
<th>($\beta_2, \beta_3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{VEM}(\beta)$</td>
<td>-0.012</td>
<td>0.021</td>
<td>0.318</td>
</tr>
<tr>
<td>$p(\beta \mid Y)$</td>
<td>-0.274</td>
<td>-0.079</td>
<td>-0.088</td>
</tr>
</tbody>
</table>

+ $p(Z \mid Y)$ in #27
Outline

Algorithmic improvements

Guarantees about variational estimates

Combining variational inference with ...
  Frequentist inference
  Bayesian inference

Conclusion (?)
Conclusion

Latent variable models (in ecology).
  ▶ Very useful (hope you’re convinced)

Variational inference (computational side).
  ▶ Computationally efficient
  ▶ Reasonably easy to implement (hope you’re convinced too)

Variational inference (theoretical side).
  ▶ Generic analysis of variational estimation still to do
  ▶ Alternatively: combine with other inference methods to combine computational efficiency with pre-existing statistical guarantees
References I


References II


Reparametrization trick

Denoting by $\psi$ the variational parameter, the VE step aims at minimizing

$$KL[q_\psi(Z)\|p_\theta(Z \mid Y)] = \mathbb{E}_{q_\psi} \log \frac{q_\psi(Z)}{p_\theta(Z \mid Y)}$$

Stochastic gradient descent requires an unbiased estimate of the gradient $\nabla_\psi \mathbb{E}_{q_\psi} (\cdot)$ ... which is not provided by sampling $Z^b \overset{iid}{\sim} q_\psi$ to estimate $\mathbb{E}_{q_\psi}$.

**Trick** [KW14,KW19]. Suppose there exist a fixed distribution $q^0$ and a function $f$, such that$^1$

$$\epsilon \sim q^0 \quad \Rightarrow \quad Z = f(\epsilon, \psi) \sim q_\psi,$$

Then, sampling $\epsilon^b \overset{iid}{\sim} q^0$ provides an unbiased estimate of the gradient:

$$\nabla_\psi \mathbb{E}_{q_\psi} \log \frac{q_\psi(Z)}{p_\theta(Z \mid Y)} \simeq \nabla_\psi \left( \frac{1}{B} \sum_b \log \frac{q_\psi(f(\epsilon^b, \psi))}{p_\theta(f(\epsilon^b, \psi) \mid Y)} \right)$$

Back to #5

$^1$Think of $q^0 = \mathcal{N}(0, I)$, $\psi = (\mu, \Sigma)$, $q_\psi = \mathcal{N}(\mu, \Sigma)$. 

S. Robin 4 - Beyond variational inference

Luxembourg, Dec’20 21 / 18
**VBEM for binary SBM**

**Posterio credibility intervals (CI)** [GDR12]: Actual level for $\pi_1 (\text{+}), \gamma_{11} (\triangle), \gamma_{12} (\circ), \gamma_{22} (\bullet)$

- **Width of the posterior CI.** $\pi_1, \gamma_{11}, \gamma_{12}, \gamma_{22}$

  $\rightarrow$ Width $\approx 1/\sqrt{n}$ for $\pi_1$ and $\approx 1/n = 1/\sqrt{n^2}$ for $\gamma_{11}, \gamma_{12}$ and $\gamma_{22}$.

Back to #7
Largest gap algorithm

- **Degree of a node**: \( D_i = \sum_{j \neq i} Y_{ij} \)
- **Mean connection from group \( k \)**:
  \[ \overline{\gamma}_k = \sum_{\ell} \pi_\ell \gamma_{k\ell} \]
- **Degree distribution**: \( (D_i \mid Z_i = k) \sim B(n - 1, \overline{\gamma}_k) \)
- **Concentration of** \( D_i/(n - 1) \) **around** \( \overline{\gamma}_{Z_i} \) **at exponential rate**

→ Ensures consistency \([CDR12]\) (including sparse regime)

---

\( ^2 \)Balanced affiliation model = nasty case: \( \pi_k \equiv 1/K, \gamma_{kk} = \gamma_{in}, \gamma_{k\ell} = \gamma_{out} \Rightarrow \overline{\gamma}_k \equiv (\gamma_{in} + (K - 1)\gamma_{out})/K \)
Sequential importance sampling scheme

Consider $U = (\theta, Z)$

Distribution path: set $0 = \rho_0 < \rho_1 < \cdots < \rho_{H-1} < \rho_H = 1$,

$$p_h(U) \propto p_{\text{start}}(U)^{1 - \rho_h} \times p_{\text{target}}(U)^{\rho_h}$$

$$\propto p_{\text{start}}(U) \times r(U)^{\rho_h},$$

$$r(U) = \frac{p(U)p(Y | U)}{p_{\text{start}}(U)}$$
Sequential importance sampling scheme

Consider $U = (\theta, Z)$

**Distribution path:** set $0 = \rho_0 < \rho_1 < \cdots < \rho_{H-1} < \rho_H = 1$, 

$$p_h(U) \propto p_{\text{start}}(U)^{1-\rho_h} \times p_{\text{target}}(U)^{\rho_h}$$

$$\propto p_{\text{start}}(U) \times r(U)^{\rho_h}, \quad r(U) = \frac{p(U)p(Y|U)}{p_{\text{start}}(U)}$$

**Sequential sampling.** At each step $h$, provides

$$\mathcal{E}_h = \{(U_h^m, w_h^m)\}_m = \text{weighted sample of } p_h$$
Sequential importance sampling scheme

Consider $U = (\theta, Z)$

Distribution path: set $0 = \rho_0 < \rho_1 < \cdots < \rho_{H-1} < \rho_H = 1$

$$p_h(U) \propto p_{\text{start}}(U)^{1-\rho_h} \times p_{\text{target}}(U)^{\rho_h}$$

$$\propto p_{\text{start}}(U) \times r(U)^{\rho_h}, \quad r(U) = \frac{p(U)p(Y \mid U)}{p_{\text{start}}(U)}$$

Sequential sampling. At each step $h$, provides

$$\mathcal{E}_h = \{(U_h^m, w_h^m)\}_m = \text{weighted sample of } p_h$$

Tune $\rho_{h+1}$ to keep the efficient sample size sufficiently high at each step.

→ Doable because $r(U)$ does not depend on $\rho$. 
Sequential sampling: in pictures

$\mathbf{p}_{\text{start}} = \text{proposal}, \quad \mathbf{p}_{\text{target}} = \text{target}$
Sequential sampling: in pictures

- $p_{\text{start}} = \text{proposal}$, $p_{\text{target}} = \text{target}$

- Intermediate distributions $p_{\text{start}} = p_0, p_1, \ldots, p_H = p_{\text{target}}$

Back to #13
Sequential sampling: in pictures

- $p_{\text{start}} = \text{proposal}$, $p_{\text{target}} = \text{target}$

- Intermediate distributions $p_{\text{start}} = p_0, p_1, \ldots, p_H = p_{\text{target}}$

- Iteratively:
  use $p_h$ to get a sample from $p_{h+1}$

Back to #13
Sequential sampling: in pictures

- $p_{\text{start}} = \text{proposal}$, $p_{\text{target}} = \text{target}$

- Intermediate distributions $p_{\text{start}} = p_0, p_1, \ldots, p_H = p_{\text{target}}$

- Iteratively:
  - use $p_h$ to get a sample from $p_{h+1}$

\[
\text{step 2: ESS} = 0.052
\]
Sequential sampling: in pictures

- $p_{\text{start}} = \text{proposal}$, $p_{\text{target}} = \text{target}$

- Intermediate distributions $p_{\text{start}} = p_0, p_1, ..., p_H = p_{\text{target}}$

- Iteratively:
  use $p_h$ to get a sample from $p_{h+1}$

---

step 3: ESS = 0.078
Sequential sampling: in pictures

- \( p_{\text{start}} = \) proposal, \( p_{\text{target}} = \) target
- Intermediate distributions \( p_{\text{start}} = p_0, p_1, \ldots, p_H = p_{\text{target}} \)
- Iteratively:
  use \( p_h \) to get a sample from \( p_{h+1} \)

step 4: ESS = 0.16

Back to #13
Sequential sampling: in pictures

- \( p_{\text{start}} = \) proposal, \( p_{\text{target}} = \) target

- Intermediate distributions \( p_{\text{start}} = p_0, p_1, \ldots, p_H = p_{\text{target}} \)

- Iteratively:
  use \( p_h \) to get a sample from \( p_{h+1} \)

+ resampling/propagation to avoid complete degeneracy [DR19]

Back to #13
Residual 'graphon'
Graphon representation of $(\pi, \alpha)$. [LR16,DR19] 

$$\phi_K : (0, 1) \times (0, 1) \mapsto \mathbb{R} \quad \text{block wise constant}$$

For a given set $S$, averaging over $K$ gives

$$\hat{\phi}(u, v) = \mathbb{E}(\phi_K(u, v) \mid Y, S) = \sum_K p(K \mid Y, S) \mathbb{E}(\phi_K(u, v) \mid Y, S, K)$$
Residual 'graphon'

Graphon representation of $\pi, \alpha$. [LR16,DR19]

$$\phi_K : (0, 1) \times (0, 1) \mapsto \mathbb{R} \quad \text{block wise constant}$$

For a given set $S$, averaging over $K$ gives

$$\hat{\phi}(u, v) = \mathbb{E}(\phi_K(u, v) \mid Y, S) = \sum_K p(K \mid Y, S) \mathbb{E}(\phi_K(u, v) \mid Y, S, K)$$
SMC path

Tree network, \( S = \{\text{taxo.}, \text{geo.}\} \)

Simulations

\[
\rho_h \quad KL \left( p_h(Z) \parallel \prod_i p_h(Z_i) \right)
\]

defined in [DR19]

Back to \#16