4 - Beyond variational inference

S. Robin

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

Winter School on Mathematical Statistics, Luxembourg, Dec'20

Outline

Part 4

Algorithmic improvements

Guaranties about variational estimates

Combining variational inference with ...

- Frequentist inference
- Bayesian inference

Conclusion (?)

Outline

Algorithmic improvements

Guaranties 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]

  $\rightarrow$ use neural networks to learn the variational parameters with more flexibility

Outline

Algorithmic improvements

Guaranties 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)

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

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

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 $\mathcal{Q} = \mathcal{Q}_{fact}$.
- The 'largest gap' algorithm [CDR12] takes advantage of a similar concentration #23
- The proofs do not easily adapt to other VEM

Outline

Frequentist inference

Frequentist inference

Maximum likelihood inference.

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; \log p_\theta(Y)$$

is intractable because the likelihood involves an integration over the latent Z

$$\text{PLN:} \qquad \log p_\theta(Y) = \sum_i \log \left(\int_{\mathbb{R}^p} p_\Sigma(Z_i) \prod_j p_\beta(Y_{ij} \mid Z_{ij}) \; \mathrm{d}Z_i \right)$$

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

Frequentist inference

Maximum likelihood inference.

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \log p_{\theta}(Y)$$

is intractable because the likelihood involves an integration over the latent Z

$$\text{PLN:} \qquad \log p_{\theta}(Y) = \sum_i \log \left(\int_{\mathbb{R}^p} p_{\Sigma}(Z_i) \prod_j p_{\beta}(Y_{ij} \mid Z_{ij}) \, \mathrm{d}Z_i \right)$$

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

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

$\rightarrow$ think, e.g., of M-estimation

Composite likelihood

Sum of partial likelihoods:

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

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

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

Composite likelihood

Sum of partial likelihoods:

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

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

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

Practical implementation.

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

Bayesian inference

Bayesian inference

Reminder.

- Prior: $p(\theta)$ $\qquad \theta_{PLN} = (\beta, \Sigma), \quad \theta_{SBM} = (\pi, \alpha, \beta)$
- Latent: $p(Z \mid \theta)$
- Observed: $p(Y \mid Z, \theta)$
- Posterior:

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

Bayesian inference

Reminder.

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

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 $(\theta^b, Z^b) \overset{\text{iid}}{\sim} p(\theta, Z \mid Y)$

Bayesian inference

Reminder.

- Prior: $p(\theta)$ $\qquad \theta_{PLN} = (\beta, \Sigma), \quad \theta_{SBM} = (\pi, \alpha, \beta)$
- 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 $(\theta^b, Z^b) \overset{\text{iid}}{\sim} p(\theta, Z \mid Y)$
- MCMC: construct a Markov chain with $p(\theta, Z \mid Y)$ as a stationary distribution

Bayesian inference

Reminder.

- Prior: $p(\theta)$ $\qquad \theta_{PLN} = (\beta, \Sigma), \quad \theta_{SBM} = (\pi, \alpha, \beta)$
- 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 $(\theta^b, Z^b) \overset{\text{iid}}{\sim} p(\theta, Z \mid Y)$
- MCMC: construct a Markov chain with $p(\theta, Z \mid Y)$ as a stationary distribution
- Importance sampling: $(\theta^b, Z^b) \overset{\text{iid}}{\sim} q(\theta, Z)$ and reweight each draw with weight

$$w^b = \frac{p(\theta^b, Z^b \mid Y)}{q(\theta^b, Z^b)}$$

Bayesian inference

Reminder.

- Prior: $p(\theta)$ $\qquad \theta_{PLN} = (\beta, \Sigma), \quad \theta_{SBM} = (\pi, \alpha, \beta)$
- 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 $(\theta^b, Z^b) \overset{\text{iid}}{\sim} p(\theta, Z \mid Y)$
- MCMC: construct a Markov chain with $p(\theta, Z \mid Y)$ as a stationary distribution
- Importance sampling: $(\theta^b, Z^b) \overset{\text{iid}}{\sim} q(\theta, Z)$ 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 $p_{start}(U)$

Sequential Monte-Carlo sampling

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

- given $p_{start}(U)$
- aiming at $p_{target}(U) = p(U \mid Y)$

Sequential Monte-Carlo sampling

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

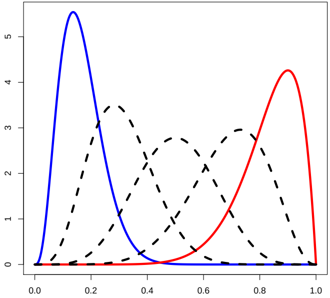
- given $p_{start}(U)$
- aiming at $p_{target}(U) = p(U \mid Y)$
- sample from a sequence of distributions

$$p_{start} = p_0, \; p_1, \; \dots, \; p_{H-1}, \; p_H = p_{target}$$

with

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

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



see #24 for tuning of the ρ_h

Sequential Monte-Carlo sampling

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

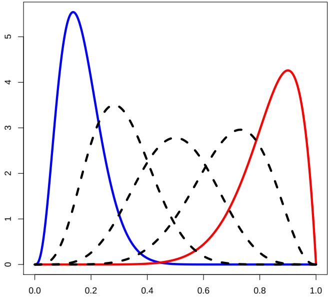
- given $p_{start}(U)$
- aiming at $p_{target}(U) = p(U \mid Y)$
- sample from a sequence of distributions

$$p_{start} = p_0,\ p_1,\ \dots,\ p_{H-1},\ p_H = p_{target}$$

with

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

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



see #24 for tuning of the ρ_h

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

VBEM: directly use $p_{start} = p_{VBEM}$

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

Back to the tree interaction network

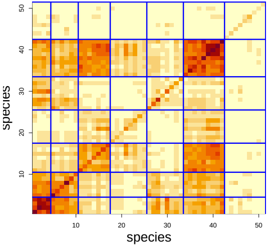
$$Y_{ij} = \text{number of shared parasites}$$
$$x_{ij} = \text{taxonomic distance}$$
$$Y_{ij} \sim \mathcal{P}(\exp(x_{ij}^\intercal \beta + \alpha_{Z_i Z_j}))$$

Estimates:

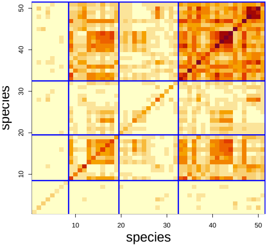
$$\widehat{K}_{ICL} = 4 \qquad \widehat{\beta} = -.317$$

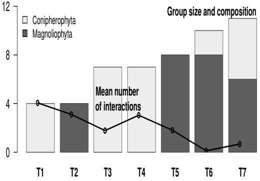
- Taxonomy (partially) explains the links (smaller $\widehat{K}$)
- Distant species share less parasites ($\widehat{\beta} < 0$)
- The remaining structure is not related to taxonomy

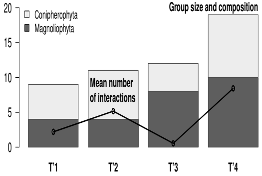
No covariate: $\widehat{K}_{ICL} = 7$



Taxonomic dist.: $\widehat{K}_{ICL} = 4$







Tree network: model selection

Model selection.

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

Tree network: model selection

Model selection.

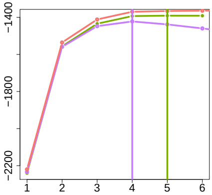
- Number of groups K
- Set S of relevent covariates: $S \subset \{\text{taxonomy, geography, phylogeny}\}$

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



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

$J_{\widehat{\theta}, \widehat{q}}$

vICL

Tree network: model selection

Model selection.

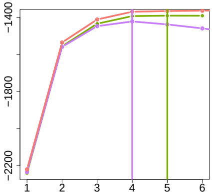
- Number of groups K
- Set S of relevent covariates: $S \subset \{\text{taxonomy, geography, phylogeny}\}$

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



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

$J_{\widehat{\theta}, \widehat{q}}$

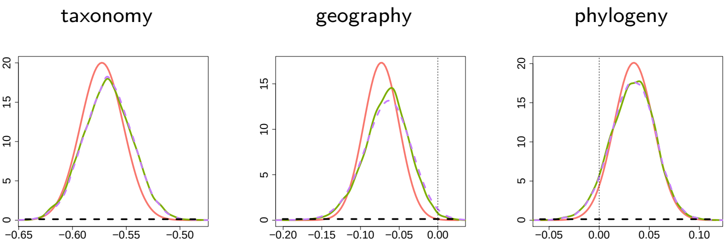
$vICL$

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

$$P\{x = (\text{taxo., geo.}) \mid Y\} \simeq 70\%, \qquad P\{x = (\text{taxo.}) \mid Y\} \simeq 30\%$$

Tree network: significance

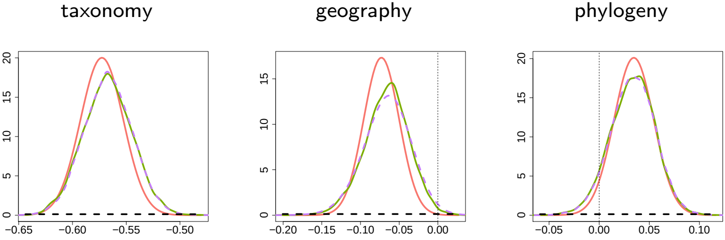
Parameter posterior distribution for $S = $ (taxonomy, geography, phylogeny):



Legend: $q_{VEM}(\beta_j)$, $p(\beta_j \mid S, \widehat{K}(S), Y)$, $p(\beta_j \mid S, Y)$

Tree network: significance

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



Legend: $q_{VEM}(\beta_j)$, $p(\beta_j \mid S, \widehat{K}(S), Y)$, $p(\beta_j \mid S, Y)$

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

Correlation between estimates.

	(β_1, β_2)	(β_1, β_3)	(β_2, β_3)
$p_{VEM}(\beta)$	-0.012	0.021	0.318
$p(\beta \mid Y)$	-0.274	-0.079	-0.088

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

Outline

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

C. Ambroise and C. Matias. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B*, 74(1):3–35, 2012.

P. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, pages 1922–1943, 2013.

A. Celisse, J.-J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Statis.*, 6:1847–99, 2012.

A. Channarond, J.-J. Daudin, and S. Robin. Classification and estimation in the stochastic block model based on the empirical degrees. *Electron. J. Statis.*, 6:2574–601, 2012.

G. Consonni and J.-M. Marin. Mean-field variational approximate Bayesian inference for latent variable models. *Computational Statistics & Data Analysis*, 52(2):790–798, 2007.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436, 2006.

S. Donnet and S. Robin. Bayesian inference for network Poisson models. Technical Report 1907.09771, arXiv, 2019.

A. Gunawardana and W. Byrne. Convergence theorems for generalized alternating minimization procedures. *J. Mach. Learn. Res.*, 6:2049–73, 2005.

S. Gazal, J.-J. Daudin, and S. Robin. Accuracy of variational estimates for random graph mixture models. *Journal of Statistical Computation and Simulation*, 82(6):849–862, 2012.

B. Ghorbani, H. Javadi, and A. Montanari. An instability in variational inference for topic models. In *International conference on machine learning*, pages 2221–2231. PMLR, 2019.

M. D Hoffman, D. M Blei, C. Wang, and J. W. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

P. Hall, J. T Ormerod, and MP Wand. Theory of gaussian variational approximation for a Poisson mixed model. *Statistica Sinica*, pages 369–389, 2011.

References II

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. Technical Report 1312.6114, arXiv, 2014.

D. P. Kingma and M. Welling. An introduction to variational autoencoders. *Foundations and TrendsÂ® in Machine Learning*, 12(4):307–392, 2019.

T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B*, pages 226–233, 1982.

P. Latouche and S. Robin. Variational Bayes model averaging for graphon functions and motif frequencies inference in W-graph models. *Statistics and Computing*, 26(6):1173–1185, 2016.

M. Mariadassou and C. Matias. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573, 2015.

S. Mandt, J. McInerney, F. Abrol, R. Ranganath, and D. Blei. Variational tempering. In *Artificial Intelligence and Statistics*, pages 704–712, 2016.

C. A. McGrory and D. M. Titterington. Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, 51:5332–67, 2007.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

T. Westling and T. H McCormick. Beyond prediction: A framework for inference with variational approximations in mixture models. *Journal of Computational and Graphical Statistics*, 28(4):778–789, 2019.

B. Wang and D. M Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *AISTATS*, 2005.

A. Y. Zhang and H. H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *Annals of Statistics*, 48(5):2575–2598, 2020.

Reparametrization trick

Denoting by ψ 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 $Z^b \overset{\text{iid}}{\sim} q_\psi$ to estimate $\mathbb{E}_{q_\psi}$.

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

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

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

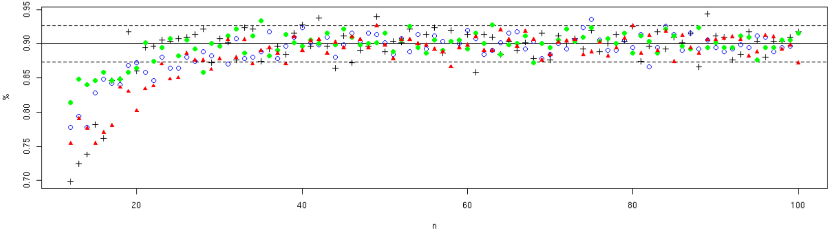
$$\nabla_\psi \, \mathbb{E}_{q_\psi} \log \frac{q_\psi(Z)}{p_\theta(Z \mid Y)} \simeq \nabla_\psi \left(\frac{1}{B} \sum_b \log \frac{q_\psi(f(\epsilon^b, \psi))}{p_\theta(f(\epsilon^b, \psi) \mid Y)} \right)$$

Back to #5

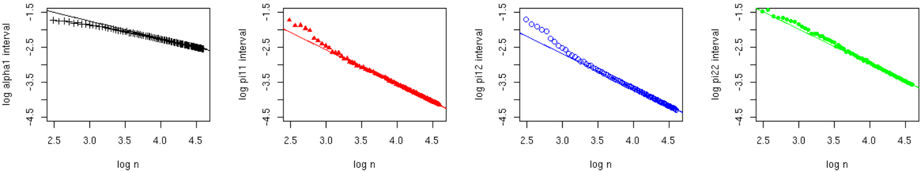
[1] Think of $q^0 = \mathcal{N}(0, I)$, $\psi = (\mu, \Sigma)$, $q_\psi = \mathcal{N}(\mu, \Sigma)$.

VBEM for binary SBM

Posterio credibility intervals (CI) [GDR12]: Actual level for π_1 ($+$), γ_{11} ($\triangle$), γ_{12} ($\circ$), γ_{22} ($\bullet$)



Width of the posterior CI. π_1, γ_{11}, γ_{12}, γ_{22}



$\rightarrow$ Width $\approx 1/\sqrt{n}$ for π_1 and $\approx 1/n = 1/\sqrt{n^2}$ for γ_{11}, γ_{12} and γ_{22}.

Back to #7

Largest gap algorithm

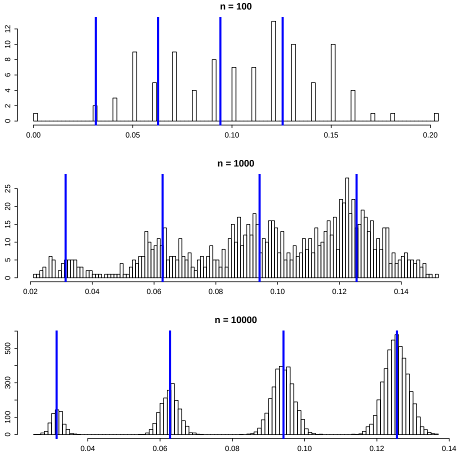
- Degree of a node: $D_i = \sum_{j \neq i} Y_{ij}$
- Mean connection from group k:

$$\overline{\gamma}_k = \sum_\ell \pi_\ell \gamma_{k\ell}$$

- Degree distribution[2]

$$(D_i \mid Z_i = k) \sim \mathcal{B}(n-1, \overline{\gamma}_k)$$

- Concentration of $D_i/(n-1)$ around $\overline{\gamma}_{Z_i}$ at exponential rate



$\rightarrow$ Ensures consistency [CDR12] (including sparse regime)

Back to #8

[2] Balanced affiliation model = nasty case: $\pi_k \equiv 1/K, \gamma_{kk} = \gamma_{in}, \gamma_{k\ell} = \gamma_{out} \quad \Rightarrow \quad \overline{\gamma}_k \equiv (\gamma_{in} + (K-1)\gamma_{out})/K$

Sequential importance sampling scheme

Consider $U = (\theta, Z)$

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

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

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

Sequential importance sampling scheme

Consider $U = (\theta, Z)$

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

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

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

Sequential sampling. At each step h, provides

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

Sequential importance sampling scheme

Consider $U = (\theta, Z)$

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

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

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

Sequential sampling. At each step h, provides

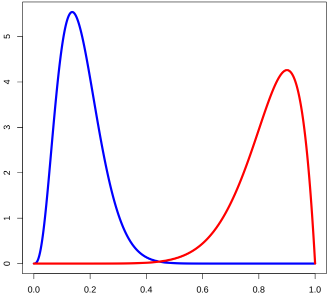
$$\mathcal{E}_h = \{(U_h^m, w_h^m)\}_m = \text{ weighted sample of } p_h$$

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

$\rightarrow$ Doable because $r(U)$ does not depend on ρ.

Sequential sampling: in pictures

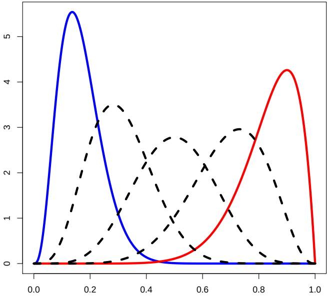
- p_{start} = proposal, p_{target} = target



Back to #13

Sequential sampling: in pictures

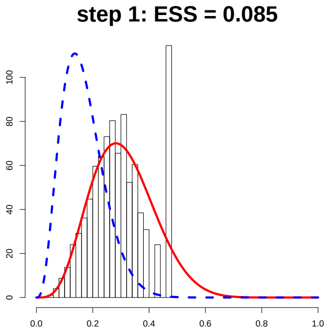
- p_{start} = proposal, p_{target} = target
- Intermediate distributions $p_{\text{start}} = p_0, p_1, \dots, p_H = p_{\text{target}}$



Back to #13

Sequential sampling: in pictures

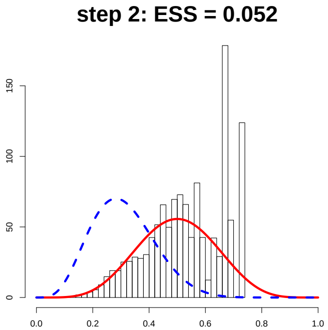
- p_{start} = proposal, p_{target} = target
- Intermediate distributions $p_{\text{start}} = p_0, \; p_1, \; \dots, \; p_H = p_{\text{target}}$
- Iteratively:
  use p_h to get a sample from p_{h+1}



Back to #13

Sequential sampling: in pictures

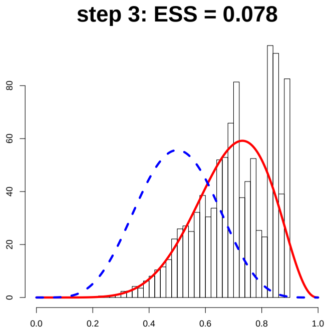
- p_{start} = proposal, p_{target} = target
- Intermediate distributions $p_{\text{start}} = p_0$, p_1, ..., $p_H = p_{\text{target}}$
- Iteratively:
  use p_h to get a sample from p_{h+1}



Back to #13

Sequential sampling: in pictures

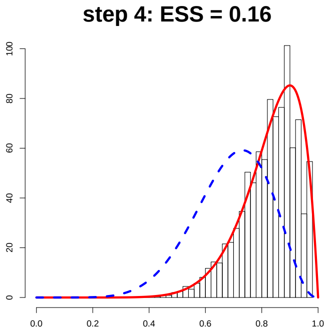
- p_{start} = proposal, p_{target} = target
- Intermediate distributions $p_{\text{start}} = p_0$, p_1, ..., $p_H = p_{\text{target}}$
- Iteratively:
  use p_h to get a sample from p_{h+1}



Back to #13

Sequential sampling: in pictures

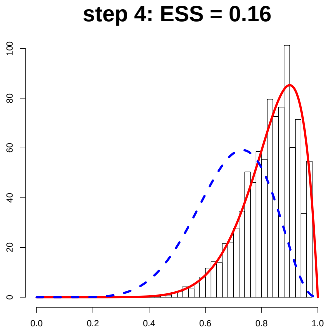
- p_{start} = proposal, p_{target} = target
- Intermediate distributions $p_{\text{start}} = p_0, p_1, \dots, p_H = p_{\text{target}}$
- Iteratively:
  use p_h to get a sample from p_{h+1}



Back to #13

Sequential sampling: in pictures

- p_{start} = proposal, p_{target} = 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}



+ resampling/propagation to avoid complete degeneracy [DR19]

Back to #13

Residual 'graphon'

Graphon representation of (π, α). [LR16,DR19]

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

For a given set S, averaging over K gives

$$\widehat{\phi}(u,v) = \mathbb{E}\left(\phi_K(u,v) \mid Y, S\right) = \sum_K p(K \mid Y, S) \mathbb{E}\left(\phi_K(u,v) \mid Y, S, K\right)$$

Residual 'graphon'

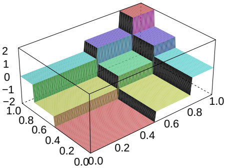
Graphon representation of (π, α). [LR16,DR19]

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

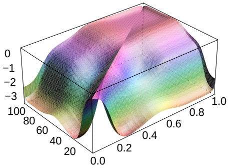
For a given set S, averaging over K gives

$$\widehat{\phi}(u,v) = \mathbb{E}\left(\phi_K(u,v) \mid Y, S\right) = \sum_K p(K \mid Y, S) \mathbb{E}\left(\phi_K(u,v) \mid Y, S, K\right)$$

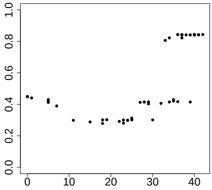
SBM graphon



$\widehat{\phi}$ for the tree network

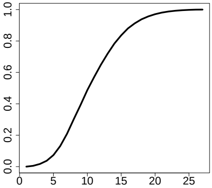
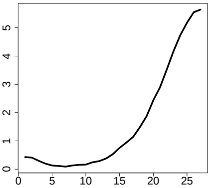
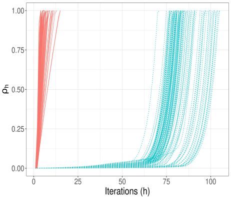


U_i vs nb. neighbors



Back to #15

SMC path

Tree network, $S = \{taxo., geo.\}$		Simulations
		
ρ_h	$KL\left(p_h(Z) \parallel \prod_i p_h(Z_i)\right)$	

from [DR19]

Back to #16