4 – Beyond variational inference

S. Robin

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

M2 Mathématiques pour les sciences du vivant, Saclay, 2021
1 – Some models with latent variables in biology and ecology

2 – Maximum likelihood inference for incomplete data models

3 – Variational inference for incomplete data models

4 – Beyond variational inference
Outline

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 \( ([#22]) \) or more dedicated approaches [HBWP13]

- Bayesian inference: Variational tempering [MMA16+

- Machine learning: Variational autoencoders [KW14,KW19]

  → 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 (?)
Guaranties about variational estimates

Statistical guarantees: *no big picture*

Accuracy of variational estimates.
- Most often assessed empirically (numerical simulations) [see e.g. #23]

'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.
- Consistency of mean-field estimates for some models (binary SBM affiliation: [ZZ20])
- Naive implementation may yield instabilities [GJM19, ZZ20]

Positive results.
- Some results for specific models (Poisson mixed model: [HOW11])
- Some attempts for a general theory via $M$-estimation [WM19]
- Most studied case: mean-field VEM for 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

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

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 [#24]
- The proofs do not easily adapt to other models
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 \( Z \)

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

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

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

→ think, e.g., of \( M \)-estimation [vdV98]
Composite likelihood

Sum of partial likelihoods:

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

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

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

Practical implementation.

▶ EM algorithms can be designed to maximize composite likelihoods [See #25]

▶ 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

Reminder.

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

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

Posterior inference.

- Posterior mean:

\[
\hat{\theta} = \mathbb{E}(\theta \mid Y)
\]

- Credibility set (with level \( 1 - \alpha \)): Find a set \( \Theta \) such that

\[
\mathbb{P}\{\theta \in \Theta \mid Y\} = \mathbb{E}[\mathbb{I}\{\theta \in \Theta\} \mid Y] = 1 - \alpha
\]
Importance sampling

Monte-Carlo estimate: Sample \( \{\theta^b\}_{1 \leq b \leq B} \) iid from \( p(\cdot \mid Y) \)

\[
\hat{E}[f(\theta) \mid Y] = \frac{1}{B} \sum_{b} f(\theta^b) \quad \text{is an unbiased estimate of } E[f(\theta) \mid Y]
\]

Importance sampling: For any distribution \( q(\cdot) \gg p(\cdot \mid Y) \), we have

\[
E[f(\theta) \mid Y] = \int f(\theta)p(\theta \mid Y) \, d\theta = \int f(\theta)q(\theta) \frac{p(\theta \mid Y)}{q(\theta)} \, d\theta
\]

Sample \( \{\theta^b\}_{1 \leq b \leq B} \) iid from the proposal distribution \( q(\cdot) \)

\[
\hat{E}[f(\theta) \mid Y] = \frac{1}{B} \sum_{b} w(\theta^b)f(\theta^b) \quad \text{is unbiased with } \quad w(\theta) = \frac{p(\theta \mid Y)}{q(\theta)}
\]

\[
\hat{E}[f(\theta) \mid Y] = \frac{\sum_{b} W(\theta^b)f(\theta^b)}{\sum_{b} W(\theta^b)} \quad \text{is (slightly) biased with } \quad W(\theta) = \frac{p(Y, \theta)}{q(\theta)}
\]
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 #26 for tuning of the $\rho_h$]

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]
Back to the tree interaction network

\[ Y_{ij} = \text{number of shared parasites} \]
\[ x_{ij} = \text{taxonomic distance} \]
\[ Y_{ij} \sim P(\exp(x_{ij}^T \beta + \alpha z_i z_j)) \]

Estimates:

\[ \widehat{K}_{ICL} = 4 \quad \widehat{\beta} = -0.317 \]

- Taxonomy (partially) explains the links (smaller \( \widehat{K} \))
- Distant species share less parasites (\( \widehat{\beta} < 0 \))
- The remaining structure is not related to taxonomy
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 \( K \): #28]

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

\[
P\{ x = (\text{taxo., geo.}) \mid Y \} \simeq 52\%, \quad P\{ x = (\text{taxo.}) \mid Y \} \simeq 47\%
\]
Tree network: significance

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

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

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

Correlation between estimates.

$\begin{align*}
    \rho_{VEM}(\beta) &\quad (\beta_1, \beta_2) &\quad (\beta_1, \beta_3) &\quad (\beta_2, \beta_3) \\
    p(\beta \mid Y) &\quad -0.012 &\quad 0.021 &\quad 0.318 \\
    p(\beta \mid Y) &\quad -0.274 &\quad -0.079 &\quad -0.088
\end{align*}$

[+ $p(Z \mid Y)$ in #29]
Outline

Algorithmic improvements

Guaranties 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 II


Reparametrization trick

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

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

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

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

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

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

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

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

---

S. Robin  4 – Beyond variational inference  M2 MSV, 2020-21
VBEM for binary SBM

Posterio credibility intervals (CI) [GDR12]: Actual level for $\pi_1$ (+), $\gamma_{11}$ (△), $\gamma_{12}$ (○), $\gamma_{22}$ (●)

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\(^2\)

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

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

→ Ensures consistency [CDR12] (including sparse regime)

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

Decomposition. The EM decomposition holds term by term:

\[
\sum_C \log p_\theta(Y_C) = \sum_C \mathbb{E}_\theta[\log p_\theta(Y_C, Z_C) \mid Y_C] - \mathcal{H}[p_\theta(Z_C \mid Y_C)]
\]

EM main property. Still holds, taking

\[
\theta^{h+1} = \arg \max_\theta \sum_C \mathbb{E}_{\theta^h}[\log p_\theta(Y_C, Z_C) \mid Y_C]
\]

so that

\[
0 \leq \sum_C \left( \mathbb{E}_{\theta^h} \left[ \log \frac{p_{\theta^{h+1}}(Y_C, Z_C)}{p_{\theta^h}(Y_C, Z_C)} \bigg| Y_C \right] \right)
\]

\[
\leq \sum_C \left( \log \mathbb{E}_{\theta^h} \left[ \frac{p_{\theta^{h+1}}(Y_C, Z_C)}{p_{\theta^h}(Y_C, Z_C)} \bigg| Y_C \right] \right)
\]

\[
= \sum_C \log \left( \int \frac{p_{\theta^h}(Y_C, Z_C)}{p_{\theta^h}(Y_C)} \frac{p_{\theta^{h+1}}(Y_C, Z_C)}{p_{\theta^h}(Y_C, Z_C)} \, dZ_C \right)
\]

\[
= \sum_C \log \frac{p_{\theta^{h+1}}(Y_C)}{p_{\theta^h}(Y_C)} = c\ell_{\theta^{h+1}}(Y) - c\ell_{\theta^h}(Y)
\]
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 sampling. At each step $h$, provides

$$\mathcal{E}_h = \{(U^m_h, w^m_h)\}_m = \text{weighted sample of } p_h$$

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

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

Backup
Sequential sampling: in pictures

- $p_{\text{start}} = \text{proposal}$, $p_{\text{target}} = \text{target}$
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 #14]
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 $p_{h+1}$

step 1: ESS = 0.085
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 2: ESS = 0.052
Sequential sampling: in pictures

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

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

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

step 3: ESS = 0.078
Sequential sampling: in pictures

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

- 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 #14]
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 $p_{h+1}$

+ resampling/propagation to avoid complete degeneracy [DR19]

[Back to #14]
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

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

SBM graphon \hspace{1cm} \hat{\phi} \text{ for the tree network} \hspace{1cm} U_i \text{ vs nb. neighbors}
SMC path

Tree network, $S = \{taxo., geo.\}$

Simulations

\[
\begin{align*}
\rho_h & \\
KL \left( p_h(Z) \parallel \prod_i p_h(Z_i) \right)
\end{align*}
\]

from [DR19]

[Back to #17]