Variational approximations & Composite likelihoods: Some links?

S. Robin

RESeau Statistiques pour données Spatio-TEmporelles

Paris, Nov. 2015

Dealing with complex models

Models with complex dependency structure (spatial, temporal, network-shaped) yield in complex likelihoods or conditional distributions.

2 main approaches.

- Stochastic algorithms (Monte-Carlo, MCMCM, SMC, ...): sample in the distribution of interest
- Deterministic algorithms: try to optimize or compute a surrogate of the distribution of interest

Deterministic approaches all need to break down dependencies

- Composite likelihood: statistical guaranties [31] but dedicated algorithms need to be designed
- Variational approximation: efficient algorithms [23,33] but no general statistical properties

General setting

Notations.

- $y =$ observed data;
- $\theta =$ parameter to be inferred;
- $z =$ latent variable.

Typical (conditional) distributions of interest.

	Frequentist	Bayesian
Fully observed	$p_\theta(y)$	$p(\theta \mid y)$
Incomplete data	$p_\theta(z \mid y)$	$p(\theta, z \mid y)$

Outline

Composite Likelihoods

General form. *Varin & al, Statistica Sinica, 2011* [31]

$$CL(y;\theta) = \prod_a p_a(y;\theta)^{w_a}, \qquad p_a = p(y \in \mathcal{A}_a;\theta)$$

where $\{\mathcal{A}_1, \dots, \mathcal{A}_A\}$ = set of marginal or conditional events.

Composite Likelihoods

General form. *Varin & al, Statistica Sinica, 2011* [31]

$$CL(y;\theta) = \prod_a p_a(y;\theta)^{w_a}, \qquad p_a = p(y \in \mathcal{A}_a;\theta)$$

where $\{\mathcal{A}_1, \dots, \mathcal{A}_A\}$ = set of marginal or conditional events.

Composite conditional likelihood.

$$\prod_i p(y_i | y_{\setminus i};\theta) \qquad \text{or} \qquad \prod_{i \neq j} p(y_i | y_j;\theta)$$

Composite Likelihoods

General form. *Varin & al, Statistica Sinica, 2011* [31]

$$CL(y;\theta) = \prod_a p_a(y;\theta)^{w_a}, \qquad p_a = p(y \in \mathcal{A}_a;\theta)$$

where $\{\mathcal{A}_1, \dots, \mathcal{A}_A\}$ = set of marginal or conditional events.

Composite conditional likelihood.

$$\prod_i p(y_i | y_{\setminus i};\theta) \qquad \text{or} \qquad \prod_{i \neq j} p(y_i | y_j;\theta)$$

Composite marginal likelihood.

$$\prod_i p(y_i;\theta), \qquad \prod_{i \neq j} p(y_i, y_j;\theta), \qquad \prod_{i \neq j} p(y_i - y_j;\theta).$$

General properties

MCLE. Maximum composite likelihood estimate:

$$\widehat{\theta}_{CL} = \arg\max_{\theta} CL(y;\theta).$$

Asymptotic normality. Under regularity conditions

$$\sqrt{n}\left(\widehat{\theta}_{CL} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, G(\theta)^{-1}\right), \qquad G = \text{Gotambe matrix.}$$

Relative efficiency. Measured by comparing $G(\theta)$ with Fisher $I(\theta)$.

Tests. CL versions of Wald or likelihood ratio test exist but 'suffer from practical limitations'.

Asymptotic variance

Reminder on likelihood:

$$I(\theta) = -\mathbb{E}_\theta[\nabla^2_\theta \log L(y;\theta)] = \mathbb{V}_\theta[\nabla_\theta \log L(y;\theta)]$$

Asymptotic variance

Reminder on likelihood:

$$I(\theta) = -\mathbb{E}_\theta[\bigtriangledown^2_\theta \log L(y;\theta)] = \mathbb{V}_\theta[\bigtriangledown_\theta \log L(y;\theta)]$$

Sensitivity matrix: – mean second derivative

$$H(\theta) = -\mathbb{E}_\theta[\bigtriangledown^2_\theta \log CL(y;\theta)]$$

Variability matrix: score variance

$$J(\theta) = \mathbb{V}_\theta[\bigtriangledown_\theta \log CL(y;\theta)] \qquad \neq H(\theta)$$

Godambe information matrix:

$$G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)$$

Application: Stochastic Block Model

Stochastic block model (SBM) [14,24]. n nodes, edges $y = (y_{ij})$, z_i = group of node i

$$P(z_i = k) = \pi_k, \qquad y_{ij} | z_i, z_j \sim f(\cdot; \gamma_{z_i z_j}), \qquad \theta = (\pi, \gamma)$$

Likelihood.

$$p(y; \theta) = \sum_z p(y, z; \theta)$$

Application: Stochastic Block Model

Stochastic block model (SBM) [14,24]. n nodes, edges $y = (y_{ij})$, z_i = group of node i

$$P(z_i = k) = \pi_k, \qquad y_{ij} | z_i, z_j \sim f(\cdot; \gamma_{z_i z_j}), \qquad \theta = (\pi, \gamma)$$

Likelihood.

$$p(y; \theta) = \sum_z p(y, z; \theta)$$

Composite log-likelihood [1].

$$CL(y; \theta) = \prod_{i \neq j \neq k} p(y_{ij}, y_{jk}, y_{ik}; \theta).$$

triplets of edges are required to guaranty identifiability.

Application: Paired HMM

Model. M series, K hidden states, $\pi : (K^M) \times (K^M)$,

$$\{z_t = (z_{it})\}_t \sim MC(\pi), \qquad \{y_{it}\} \text{ indep.} | z, \quad (y_{it} | z_{it} = k) \sim f(\gamma_k).$$

Composite likelihood [10]. $\theta = (\pi, \gamma)$

$$CL(y;\theta) = \prod_{i \neq j} p(y_i, y_j; \theta)$$

1

Application: Paired HMM

Model. M series, K hidden states, $\pi : (K^M) \times (K^M)$,

$$\{z_t = (z_{it})\}_t \sim MC(\pi), \qquad \{y_{it}\} \text{ indep.}|z, \quad (y_{it}|z_{it} = k) \sim f(\gamma_k).$$

Composite likelihood [10]. $\theta = (\pi, \gamma)$

$$CL(y; \theta) = \prod_{i \neq j} p(y_i, y_j; \theta)$$

$\rightarrow$ CL-EM algorithm

- E-step: compute via forward-backward with K^2 hidden states[1]

$$p(z_i, z_j | y_i, y_j; \theta);$$

- M-step: update

$$\widehat{\theta} = \arg\max_\theta \sum_{i \neq j} \mathbb{E}\left[\log p(y_i, y_j, z_i, z_j; \theta) | y_i, y_j\right]$$

[1] but $\{(z_{it}, z_{jt})\}_t$ is not a Markov chain in general...

Outline

Variational techniques

Origin. Mostly arise from the machine learning community:

- optimization techniques
- efficient algorithms
- related to graphical models [18]

A huge literature.

- Plenty of tutorials: [17,15]
- Plenty of reviews: [23,29]
- A joint AIGM work: [27]
- An *opus magnus*:

Wainwright & Jordan, Found. Trends Mach. Learn, 2008 [33]

(Very) general principle

Aim: For some 'hidden' $h = \theta$ or z or (θ, z), find

$$q(h) \simeq p(h|y)$$

taking

$$q \in \mathcal{Q}$$

(Very) general principle

Aim: For some 'hidden' $h = \theta$ or z or (θ, z), find

$$q(h) \simeq p(h|y)$$

taking

$$q \in \mathcal{Q}$$

$\mathcal{Q}$ = class of 'nice' distributions:

(*i*) Provided by some (efficient) algorithm.

(*ii*) Parametric family

$$\mathcal{Q} = \{\mathcal{N}(\mu, \Sigma)\}$$

(*iii*) Breaking down some dependencies

$$\mathcal{Q} = \{q(h) = \prod_a q_a(h^a)\}, \qquad h^a = (h_j)_{j \in a}$$

(i): Belief propagation

Exact algorithms allow to compute the conditional distribution $p(z|y)$ for some specific dependency structures:

- Forward-Backward for hidden Markov models;
- Upward-Downward for tree-shaped graphical models.

(i): Belief propagation

Exact algorithms allow to compute the conditional distribution $p(z|y)$ for some specific dependency structures:

- Forward-Backward for hidden Markov models;
- Upward-Downward for tree-shaped graphical models.

Belief propagation: apply such an algorithm to a structure for which it is not exact [23,9].

(i): Belief propagation

Exact algorithms allow to compute the conditional distribution $p(z|y)$ for some specific dependency structures:

- Forward-Backward for hidden Markov models;
- Upward-Downward for tree-shaped graphical models.

Belief propagation: apply such an algorithm to a structure for which it is not exact [23,9].

Alternative = reduction: Merge some $(h_{j_1}, \dots h_{j_m})$ into multivariate h^a so that the dependency structure of $p(\{h^a\}|y)$ is tree-shaped [27].

(ii): Approximate Gaussian posteriors

Bayesian logistic regression: covariates $y_i \in \mathbb{R}^d$, response $y_i \in \{0, 1\}$:

$$\theta \sim \mathcal{N}(\mu, \Sigma), \qquad y|\theta \sim \mathcal{B}(g(y_i^\intercal \theta)) \quad \text{with } g(u) = (1 + e^{-u})^{-1}$$

$\rightarrow p(\theta|y) =?$

2

(ii): Approximate Gaussian posteriors

Bayesian logistic regression: covariates $y_i \in \mathbb{R}^d$, response $y_i \in \{0, 1\}$:

$$\theta \sim \mathcal{N}(\mu, \Sigma), \qquad y|\theta \sim \mathcal{B}(g(y_i^\intercal \theta)) \quad \text{with } g(u) = (1 + e^{-u})^{-1}$$

$\rightarrow p(\theta|y) =?$

Variational Gaussian posterior [16]: No conjugacy arises but, because of ([2]),

$$\log p(\theta, y) \geq \text{quadratic form on } \theta$$

$\widetilde{\mu}(y) \leftarrow$ first order terms, $\widetilde{\Sigma}(y) \leftarrow$ quadratic terms, so

$$p(\theta|y) \simeq \mathcal{N}(\widetilde{\mu}(y), \widetilde{\Sigma}(y)).$$

See also [26,30] for GLMM.

[2] $-\log(1 + e^{-u}) = \frac{u}{2} - \log(e^{u/2}e^{-u/2}) \geq \log g(u_0) + \frac{1}{2}(u - u_0) + \frac{1}{4u_0} \tanh\left(\frac{u_0}{2}\right)(u^2 - u_0^2)$

(*ii*) and (*iii*): More explicit principle

Aim: Find

$$q(h) \simeq p(h|y)$$

taking

$$\arg\min_{q \in \mathcal{Q}} D[q||p]$$

(ii) and (iii): More explicit principle

Aim: Find

$$q(h) \simeq p(h|y)$$

taking

$$\arg\min_{q \in \mathcal{Q}} D[q||p]$$

$\rightarrow$ Need for

- $D[\cdot||\cdot]$: a measure of divergence between distributions:

$$KL[q||p], \qquad KL[p||q], \qquad Hellinger[q,p], \quad D_\alpha[q||p]$$

  see [23] for a review and a comparison of respective merits.

- $\mathcal{Q}$: a class of 'nice' distributions:

$$\mathcal{Q} = \{\mathcal{N}(\mu, \Sigma)\}, \qquad \mathcal{Q} = \{q(h) = \prod_a q_a(h^a)\}$$

Lower bound of the likelihood

Lower bound: Two equivalent problems

$$\arg\min_q \; D[q||p(\cdot|y)] \quad = \quad \arg\max_q \; \log p(y) - D[q||p(\cdot|y)]$$

Lower bound of the likelihood

Lower bound: Two equivalent problems

$$\arg\min_q \; D[q||p(\cdot|y)] \quad = \quad \arg\max_q \; \log p(y) - D[q||p(\cdot|y)]$$

Küllback-Leibler divergence: $KL[q||p] = \mathbb{E}_q \log(q/p)$

$$\begin{aligned} \log p(y) - KL[q(\cdot)||p(\cdot|y)] &= \log p(y) - \int q(h) \log \frac{q(h)p(y)}{p(h,y)} \mathrm{d}h \\ &= -KL[q||p(\cdot, y)] \end{aligned}$$

Lower bound of the likelihood

Lower bound: Two equivalent problems

$$\arg\min_q \; D[q||p(\cdot|y)] \quad = \quad \arg\max_q \; \log p(y) - D[q||p(\cdot|y)]$$

Küllback-Leibler divergence: $KL[q||p] = \mathbb{E}_q \log(q/p)$

$$\begin{aligned} \log p(y) - KL[q(\cdot)||p(\cdot|y)] &= \log p(y) - \int q(h) \log \frac{q(h)p(y)}{p(h, y)} \mathrm{d}h \\ &= -KL[q||p(\cdot, y)] \end{aligned}$$

- $\neq$ MLE which minimizes $KL[\widehat{p}||q]$
- Only deals with the joint (or complete) distribution $p(h, y)$ [15]
- Can be used for (variational) Bayes model selection or averaging [32]

A functional optimization problem [4]

Optimal q: must satisfy for any function (direction) r:

$$\left.\frac{\partial}{\partial t}\right|_{t=0} D\left[q + tr || p\right] = 0.$$

3

A functional optimization problem [4]

Optimal q: must satisfy for any function (direction) r:

$$\left.\frac{\partial}{\partial t}\right|_{t=0} D\left[q + tr||p\right] = 0.$$

One often has

$$D[q||p] = \int F[q(h), p(h)] \mathrm{d}h$$

so, under regularity conditions,

$$\begin{aligned}\left.\frac{\partial}{\partial t}\right|_{t=0} D\left[q + tr||p\right] &= \int \left.\frac{\partial}{\partial t}\right|_{t=0} F[q(h) + tr(h), p(h)] \mathrm{d}h \\ &= \int r(h) F'[q(h), p(h)] \mathrm{d}h \qquad (^3)\end{aligned}$$

[3] $F'(\cdot, \cdot)$ stands for the derivative wrt the first argument of F.

A functional optimization problem [4]

Optimal q: must satisfy for any function (direction) r:

$$\left.\frac{\partial}{\partial t}\right|_{t=0} D\left[q + tr||p\right] = 0.$$

One often has

$$D[q||p] = \int F[q(h), p(h)] \mathrm{d}h$$

so, under regularity conditions,

$$\begin{aligned} \left.\frac{\partial}{\partial t}\right|_{t=0} D\left[q + tr||p\right] &= \int \left.\frac{\partial}{\partial t}\right|_{t=0} F[q(h) + tr(h), p(h)] \mathrm{d}h \\ &= \int r(h) F'[q(h), p(h)] \mathrm{d}h \qquad (^3) \end{aligned}$$

which must hold for any r, so the optimal q satisfies (see also Thm 3 in [23])

$$F'[q(h), p(h)] = 0.$$

---

[3] $F'(\cdot, \cdot)$ stands for the derivative wrt the first argument of F.

Mean-field approximation

Most popular case: $D[q||p] = KL[q||p]$, $q(h) = \prod_a q_a(h^a)$ gives

$$q_a(h^a) \propto \exp\left(\mathbb{E}_{q_{\setminus a}} \log p(h, y)\right)$$

Remind that

$$p_a(h_a) = \mathbb{E}_{p_{\setminus a}} p(h, y | h^{\setminus a})$$

4

Mean-field approximation

Most popular case: $D[q||p] = KL[q||p]$, $q(h) = \prod_a q_a(h^a)$ gives

$$q_a(h^a) \propto \exp\left(\mathbb{E}_{q_{\setminus a}} \log p(h, y)\right)$$

Remind that

$$p_a(h_a) = \mathbb{E}_{p_{\setminus a}} p(h, y | h^{\setminus a})$$

Stochastic block model (SBM): Conditional distribution ($z_{ik} = \mathbb{I}\{z_i = k\}$)

$$P(z_i = k | y, z_{\setminus i}) \propto \pi_k \prod_j \prod_\ell f(y_{ij}; \gamma_{k\ell})^{z_{j\ell}} \qquad ([4])$$

[4] suggest Gibbs sampling as used in [24]

Mean-field approximation

Most popular case: $D[q||p] = KL[q||p]$, $q(h) = \prod_a q_a(h^a)$ gives

$$q_a(h^a) \propto \exp\left(\mathbb{E}_{q_{\setminus a}} \log p(h, y)\right)$$

Remind that

$$p_a(h_a) = \mathbb{E}_{p_{\setminus a}} p(h, y | h^{\setminus a})$$

Stochastic block model (SBM): Conditional distribution ($z_{ik} = \mathbb{I}\{z_i = k\}$)

$$P(z_i = k | y, z_{\setminus i}) \propto \pi_k \prod_j \prod_\ell f(y_{ij}; \gamma_{k\ell})^{z_{j\ell}} \qquad ([4])$$

Variational approximation ($\tau_{ik} = \mathbb{E}_{q_i}(z_{ik})$) ($\rightarrow$ Variational EM = VEM [7])

$$\tau_{ik} \propto \pi_k \prod_j \prod_\ell f(y_{ij}; \gamma_{k\ell})^{\tau_{j\ell}}$$

[4] suggest Gibbs sampling as used in [24]

Variational Bayes inference

Bayesian model with latent variable defined by

$$\text{prior } p(\theta), \qquad p(z|\theta), \qquad p(y|\theta, z) \qquad \Rightarrow \qquad p(\theta, z|y) = ?$$

Variational Bayes inference

Bayesian model with latent variable defined by

$$\text{prior } p(\theta), \qquad p(z|\theta), \qquad p(y|\theta, z) \qquad \Rightarrow \qquad p(\theta, z|y) = ?$$

Variational Bayes EM (VBEM): taking $h^1 = z$ and $h^2 = \theta$ gives

- Variational E-step:

$$q_1(z) \propto \exp\left[\mathbb{E}_{q_2} \log p(z, y|\theta)\right]$$

- Variational M-step:

$$q_2(\theta) \propto \exp\left[\mathbb{E}_{q_1} \log p(\theta, z, y)\right]$$

Variational Bayes inference

Bayesian model with latent variable defined by

$$\text{prior } p(\theta), \qquad p(z|\theta), \qquad p(y|\theta, z) \qquad \Rightarrow \qquad p(\theta, z|y) = ?$$

Variational Bayes EM (VBEM): taking $h^1 = z$ and $h^2 = \theta$ gives

- Variational E-step:

$$q_1(z) \propto \exp \left[\mathbb{E}_{q_2} \log p(z, y|\theta)\right]$$

- Variational M-step:

$$q_2(\theta) \propto \exp \left[\mathbb{E}_{q_1} \log p(\theta, z, y)\right]$$

All updates are explicit if $p(z, y|\theta)$ belongs to the exponential family and a conjugate prior $p(\theta)$ is used [3].

Statistical properties of variational approximations

Negative.

- VEM algorithm optimum $\neq$ ML in general [12]
- VBEM posterior variance too small [6]
- Precise analysis for mixture and hidden Markov models [34,22]

Statistical properties of variational approximations

Negative.

- VEM algorithm optimum $\neq$ ML in general [12]
- VBEM posterior variance too small [6]
- Precise analysis for mixture and hidden Markov models [34,22]

Positive.

- Mean field approximations are asymptotically exact for models with 'infinite range dependency' [25]
- Consistency of the parameters of the approximate Gaussian posterior for generalized linear mixed model [26], special case of Poisson regression [13]
- Consistency of VEM estimates for SBM [5,21] + empirical accuracy of the VBEM posterior for SBM [11]

Outline

Some Links?

In presence of a complex dependency structure:

- Variational methods break dependencies down and apply efficient algorithms, with few statistical guaranties;
- Composite likelihood methods break dependencies down with statistical guaranties but not always efficient algorithms

Some Links?

In presence of a complex dependency structure:

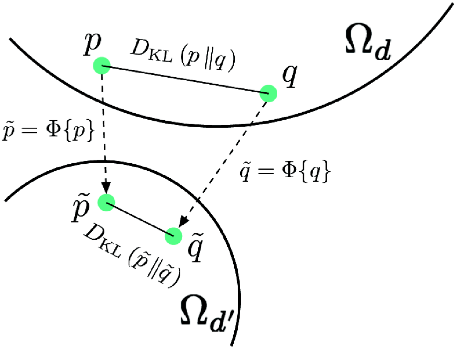
- Variational methods break dependencies down and apply efficient algorithms, with few statistical guaranties;
- Composite likelihood methods break dependencies down with statistical guaranties but not always efficient algorithms

Question. Are variational methods like Mr Jourdain for composite likelihoods?

Main reference: *Luy, NIPS, 2011* [20] ... very few citations since then.

KL contraction

Definition. Denote Ω_d the set of all distributions over $\mathbb{R}^d$.



[20]

$\Phi : \Omega_d \mapsto \Omega_{d'}$ is KL-contactant iff, $\exists \beta \geq 1$, $\forall p, q \in \Omega_d$:

$$KL[p||q] - \beta\ KL[\Phi\{p\}||\Phi\{q\}] \geq 0.$$

Examples of KL contraction

Examples of KL contraction

- Marginal distribution:

$$\Phi_a^m\{p\}(x) = \int p(x)\mathrm{d}y_{\setminus a}.$$

Examples of KL contraction

- Marginal distribution:

$$\Phi_a^m\{p\}(x) = \int p(x) \mathrm{d}y_{\setminus a}.$$

- Conditional distribution: for a given distribution $t(y|x)$

$$\Phi_t^c\{p\}(y) = \int p(x) t(y|x) \mathrm{d}x.$$

Examples of KL contraction

- Marginal distribution:

$$\Phi_a^m\{p\}(x) = \int p(x) \mathrm{d}y_{\setminus a}.$$

- Conditional distribution: for a given distribution $t(y|x)$

$$\Phi_t^c\{p\}(y) = \int p(x) t(y|x) \mathrm{d}x.$$

- Marginal grafting: replace $p_a(y^a)$ with $t_a(y^a)$

$$\Phi_{t,a}^g\{p\}(x) = p(x) \frac{t_a(y^a)}{p_a(y^a)} = t_a(y^a) p_{\setminus a|a}(y^{\setminus a}|y^a).$$

Examples of KL contraction

- Marginal distribution:

$$\Phi_a^m\{p\}(x) = \int p(x) \mathrm{d}y_{\setminus a}.$$

- Conditional distribution: for a given distribution $t(y|x)$

$$\Phi_t^c\{p\}(y) = \int p(x) t(y|x) \mathrm{d}x.$$

- Marginal grafting: replace $p_a(y^a)$ with $t_a(y^a)$

$$\Phi_{t,a}^g\{p\}(x) = p(x)\frac{t_a(y^a)}{p_a(y^a)} = t_a(y^a) p_{\setminus a|a}(y^{\setminus a}|y^a).$$

- $+$ binary mixture ($\approx$ shrinkage), lumping ($=$ discretization), ...

Possible use for inference

Type I: Avoid to compute normalizing constants, which can vanish in the difference

$$KL[p||q_\theta] - \beta \; KL[\Phi\{p\}||\Phi\{q_\theta\}] \tag{1}$$

Possible use for inference

Type I: Avoid to compute normalizing constants, which can vanish in the difference

$$KL[p||q_\theta] - \beta \, KL[\Phi\{p\}||\Phi\{q_\theta\}] \tag{1}$$

Type II: Define a easy-to-handle objective function based on a Taylor expansion of (1).

Possible use for inference

Type I: Avoid to compute normalizing constants, which can vanish in the difference

$$KL[p||q_\theta] - \beta \; KL[\Phi\{p\}||\Phi\{q_\theta\}] \tag{1}$$

Type II: Define a easy-to-handle objective function based on a Taylor expansion of (1).

Type III: Use a set of contractions $(\Phi_1, \dots \Phi_A)$ to infer θ with

$$\arg\min_\theta \sum_a w_a \left[KL[p||q_\theta] - \beta_a KL[\Phi_a\{p\}||\Phi_a\{q_\theta\}]\right].$$

Links with composite likelihoods

Maximum likelihood: Taking $p =$ empirical distribution and $q_\theta =$ parametric model,

$$\widehat{\theta}_{ML} = \arg\min_\theta KL[p||q_\theta]$$

Links with composite likelihoods

Maximum likelihood: Taking $p =$ empirical distribution and $q_\theta =$ parametric model,

$$\widehat{\theta}_{ML} = \arg\min_\theta KL[p||q_\theta]$$

Type III with marginal contraction = Conditional composite likelihood: For subsets $a_1, a_2, \dots,$ using $\Phi_a^m \rightarrow$

$$\arg\max_\theta \sum_a w_a \log q_{\setminus a|a}(y^{\setminus a}|y^a;\theta)$$

Links with composite likelihoods

Maximum likelihood: Taking $p =$ empirical distribution and $q_\theta =$ parametric model,

$$\widehat{\theta}_{ML} = \arg\min_\theta KL[p||q_\theta]$$

Type III with marginal contraction = Conditional composite likelihood: For subsets $a_1, a_2, \dots$, using $\Phi_a^m \rightarrow$

$$\arg\max_\theta \sum_a w_a \log q_{\setminus a|a}(y^{\setminus a}|y^a; \theta)$$

Type III with marginal grafting = Marginal composite likelihood: using $\Phi_{p,a}^g \rightarrow$

$$\arg\max_\theta \sum_a w_a \log q_a(y^a; \theta)$$

$+$ Gaussian approximation using Φ_t^c where $t = \mathcal{N}(x, \sigma^2)$.

Outline

Conclusion: There is no conclusion

Connexions do exist.

- Some variational approximations of the likelihood are actually composite likelihoods [20,35].
- Many authors observe the connexion between composite likelihoods and (contrastive) divergence-based learning, but some end up using MCMC [19,2,8]...

[20]: '*While many non-ML learning methods covered in this work have been shown to be consistent individually, the unification based on the minimum KL contraction may provide a general condition for such asymptotic properties.*' ...

Conclusion: There is no conclusion

Connexions do exist.

- Some variational approximations of the likelihood are actually composite likelihoods [20,35].
- Many authors observe the connexion between composite likelihoods and (contrastive) divergence-based learning, but some end up using MCMC [19,2,8]...

[20]: '*While many non-ML learning methods covered in this work have been shown to be consistent individually, the unification based on the minimum KL contraction may provide a general condition for such asymptotic properties.*' ...

But

- No nice example to show
- Our favorite $KL[q_\theta||p]$ does not involve any contraction.
- No generic way to make the connexion.

Variational/composite posterior as proposals

Approximate posterior. Both variational Bayes and composite likelihood provide approximations of the posterior:

$$p(\theta|y) \simeq q_{VB}(\theta), \qquad p(\theta|y) \simeq q_{CL}(\theta) := p(\theta)e^{CL(y;\theta)}.$$

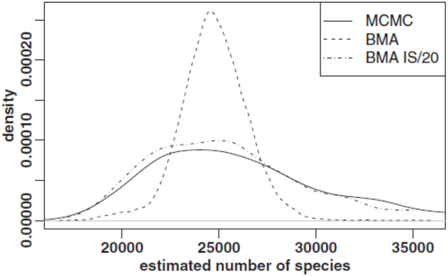
Importance sampling. q_{VB} and q_{CL} can be used as proposal for importance sampling

$$\widehat{\mathbb{E}}_p[f(\theta)|y] = \sum_b \frac{p(\theta^b, y)}{q(\theta^b)} f(\theta^b) \Bigg/ \sum_b \frac{p(\theta^b, y)}{q(\theta^b)}, \qquad \{\theta^b\} \text{ iid } \sim q$$

but often lead too poor efficiency.

Calibrating variational/composite posteriors

Both variational Bayes and composite likelihood posterior need to be improved:



Can we do that in a automated (sequential) way? Calibration: [28], Optimizing some efficiency criterion: [On-going work].

Christophe Ambroise and Catherine Matias.
New consistent and asymptotically normal parameter estimates for random-graph mixture models.
J. R. Statist. Soc. B, 74(1):3–35, 2012.

Arthur U Asuncion, Qiang Liu, Alexander T Ihler, and Padhraic Smyth.
Learning with blocks: Composite likelihood and contrastive divergence.
In *International Conference on Artificial Intelligence and Statistics*, pages 33–40, 2010.

J. Beal, M. and Z. Ghahramani.
The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures.
Bayes. Statist., 7:543–52, 2003.

M.J. Beal.
Variational Algorithms for Approximate Bayesian Inference.
PhD thesis, University College London, 2003.

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.

Guido Consonni and Jean-Michel Marin.
Mean-field variational approximate bayesian inference for latent variable models.
Computational Statistics & Data Analysis, 52(2):790–798, 2007.

J.-J. Daudin, F. Picard, and S. Robin.
A mixture model for random graphs.
Stat. Comput., 18(2):173–83, Jun 2008.

I. E Fellows.
Why (and When and How) Contrastive Divergence Works.
Technical report, arXiv:1405.0602, 2014.

Brendan J Frey and David JC MacKay.
A revolution: Belief propagation in graphs with cycles.
Advances in neural information processing systems, pages 479–485, 1998.

Xin Gao and Peter X.-K. Song.

Composite likelihood em algorithm with applications to multivariate hidden markov model.
Statistica Sinica, 21(1):165–185, 2011.

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

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

Peter Hall, JT Ormerod, and MP Wand.
Theory of gaussian variational approximation for a poisson mixed model.
Statistica Sinica, 21:369–389, 2011.

Paul W Holland and Samuel Leinhardt.
Structural sociometry.
Perspectives on social network research, pages 63–83, 1979.

T. Jaakkola.
Advanced mean field methods: theory and practice, chapter Tutorial on variational approximation methods.
MIT Press, 2000.

Tommi S Jaakkola and Michael I Jordan.
Bayesian parameter estimation via variational methods.
Statistics and Computing, 10(1):25–37, 2000.

Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul.
An introduction to variational methods for graphical models.
Machine Learning, 37(2):183–233, 1999.

S.L. Lauritzen.
Graphical Models.
Oxford Statistical Science Series. Clarendon Press, 1996.

Percy Liang and Michael I Jordan.
An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators.

In *Proceedings of the 25th international conference on Machine learning*, pages 584–591. ACM, 2008.

Siwei Lyu.
Unifying non-maximum likelihood learning objectives with minimum KL contraction.
In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *NIPS*, pages 64–72, 2011.

M. Mariadassou, S. Robin, and C. Vacher.
Uncovering latent structure in valued graphs: A variational approach.
Ann. Appl. Stat., 4(2):715–742, 06 2010.

C. A. McGrory and D. M. Titterington.
Variational Bayesian analysis for hidden Markov models.
Austr. & New Zeal. J. Statist., 51(2):227–44, 2009.

Tom Minka.
Divergence measures and message passing.
Technical Report MSR-TR-2005-173, Microsoft Research Ltd, 2005.
`ftp://ftp.research.microsoft.com/pub/tr/TR-2005-173.pdf`.

K. Nowicki and T.A.B. Snijders.
Estimation and prediction for stochastic block-structures.
J. Amer. Statist. Assoc., 96:1077–87, 2001.

M. Opper and O. Winther.
Advanced mean field methods: Theory and practice, chapter From Naive Mean Field Theory to the TAP Equations.
The MIT Press, 2001.

John T Ormerod and MP Wand.
Gaussian variational approximate inference for generalized linear mixed models.
Journal of Computational and Graphical Statistics, 21(1):2–17, 2012.

N. Peyrard, S. de Givry, A. Franc, S. Robin, R. Sabbadin, T. Schiex, and M. Vignes.
Exact and approximate inference in graphical models: variable elimination and beyond.
Technical report, ArXiv:1506.08544, June 2015.

J. Stoehr and N. Friel.
Calibration of conditional composite likelihood for Bayesian inference on Gibbs random fields.

Technical report, 2015.

Shiliang Sun.
A review of deterministic approximate inference techniques for bayesian machine learning.
Neural Computing and Applications, 23(7-8):2039–2050, 2013.

Linda SL Tan and David J Nott.
Variational inference for generalized linear mixed models using partially noncentered parametrizations.
Statistical Science, 28(2):168–188, 2013.

Cristiano Varin, Nancy Reid, and David Firth.
An overview of composite likelihood methods.
Statistica Sinica, 21:5–42, 2011.

Stevenn Volant, Marie-Laure Martin Magniette, and Stéphane Robin.
Variational bayes approach for model aggregation in unsupervised classification with markovian dependency.
Comput. Statis. & Data Analysis, 56(8):2375 – 2387, 2012.

M. J. Wainwright and M. I. Jordan.
Graphical models, exponential families, and variational inference.
Found. Trends Mach. Learn., 1(1–2):1–305, 2008.
http:/dx.doi.org/10.1561/2200000001.

B. Wang and M. Titterington, D.
Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model.
Bayes. Anal., 1(3):625–50, 2006.

Yi Zhang and Jeff Schneider.
A composite likelihood view for multi-label classification.
Journal of Machine Learning Research - Proceedings Track, 22:1407–1415, 2012.