Some Links between Variational Approximation and Composite Likelihoods?

S. Robin

UMR 518 AgroParisTech / INRA Applied Math & Comput. Sc.

MSTGA, Paris, November 22-23, 2012

Main references for this talk

- Minka, T. (2005), Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd.
- Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. Statistica Sinica. 21 5–42. [1]
- Lyu, S. (2011). Unifying non-maximum likelihood learning objectives with minimum KL contraction. In NIPS, (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, ed.), 64–72.

[1] Opening paper of a special issue on composite likelihoods

Variational Approximations [*Minka (2005)*]

Variational Approximations [*Minka (2005)*]

Aim: Approximate a 'complex' distribution p with a simpler one q .

Variational Approximations [*Minka (2005)*]

Aim: Approximate a 'complex' distribution p with a simpler one q .

General principle: Choose q within a certain class, such that it minimizes a divergence measure wrt p.

Variational Approximations [*Minka (2005)*]

Aim: Approximate a 'complex' distribution p with a simpler one q .

General principle: Choose q within a certain class, such that it minimizes a divergence measure wrt p.

Examples:

- Mean-field approximation: $q^* = \arg\min_q KL(q||p)$ where

$$KL(q||p) := \int q(y) \log \frac{q(y)}{p(y)} \mathrm{d}y - \int [q(y) - p(y)] \mathrm{d}y;$$

Variational Approximations [*Minka (2005)*]

Aim: Approximate a 'complex' distribution p with a simpler one q .

General principle: Choose q within a certain class, such that it minimizes a divergence measure wrt p.

Examples:

- Mean-field approximation: $q^* = \arg\min_q KL(q||p)$ where

$$KL(q||p) := \int q(y) \log \frac{q(y)}{p(y)} \mathrm{d}y - \int [q(y) - p(y)] \mathrm{d}y;$$

- Expectation propagation (EP): $q^* = \arg\min_q KL(p||q)$;

Variational Approximations [*Minka (2005)*]

Aim: Approximate a 'complex' distribution p with a simpler one q .

General principle: Choose q within a certain class, such that it minimizes a divergence measure wrt p.

Examples:

- Mean-field approximation: $q^* = \arg\min_q KL(q||p)$ where

$$KL(q||p) := \int q(y) \log \frac{q(y)}{p(y)} \mathrm{d}y - \int [q(y) - p(y)] \mathrm{d}y;$$

- Expectation propagation (EP): $q^* = \arg\min_q KL(p||q)$;
- Power EP: $q^* = \arg\min_q D_\alpha(p||q)$ where

$$D_\alpha(p||q) := \frac{\int \alpha p(y) + (1-\alpha) q(y) - p(y)^\alpha q(y)^{1-\alpha} \mathrm{d}y}{\alpha(1-\alpha)}.$$

Two main uses

Variational approximation are generally used for two purposes (possibly combined).

Two main uses

Variational approximation are generally used for two purposes (possibly combined).

Shape approximation. To, e.g., access close form estimates:

$$p(y) = \prod_k p_k(y) \quad \approx \quad q(y) = \prod_k q_k(y)$$

where each q_k belongs to, say, the exponential family.

Two main uses

Variational approximation are generally used for two purposes (possibly combined).

Shape approximation. To, e.g., access close form estimates:

$$p(y) = \prod_k p_k(y) \quad \approx \quad q(y) = \prod_k q_k(y)$$

where each q_k belongs to, say, the exponential family.

Break down dependencies.

$$p(y) \text{ not factorisable} \quad \approx \quad q(y) = \prod_k q_k(y).$$

Properties of variational estimates

'Approximate' likelihood inference. Standard MLE are often replaced by

$$\widehat{\theta}_L = \arg\max_{\theta} \log p(Y;\theta) \quad \rightarrow \quad \widehat{\theta}_{VL} = \arg\max_{\theta} \log q(Y;\theta)$$

Properties of variational estimates

'Approximate' likelihood inference. Standard MLE are often replaced by

$$\widehat{\theta}_L = \arg\max_\theta \log p(Y;\theta) \quad \rightarrow \quad \widehat{\theta}_{VL} = \arg\max_\theta \log q(Y;\theta)$$

but the statistical properties of the resulting estimates are not known in general:

- Consistency: $\widehat{\theta}^n_{VL} \xrightarrow{p} \theta$?
  Except for some special cases (e.g. [*Wang and Titterington (2006)*], [*Celisse* et al. *(2012)*], ...).
- Asymptotic distribution: $\sqrt{n}(\widehat{\theta}^n_{VL} - \theta) \xrightarrow{d} \mathcal{N}$?

Composite Likelihoods [*Varin* et al. *(2011)*]

Composite Likelihoods [*Varin* et al. *(2011)*]

General form.

$$CL(Y;\theta) = \prod_k L_k(Y;\theta)^{w_k}, \qquad L_k = p(Y \in \mathcal{A}_k;\theta)$$

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

Composite Likelihoods [*Varin* et al. *(2011)*]

General form.

$$CL(Y;\theta) = \prod_k L_k(Y;\theta)^{w_k}, \qquad L_k = p(Y \in \mathcal{A}_k;\theta)$$

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

Composite conditional likelihood.

$$\prod_t p(Y_t | Y^{-t};\theta) \qquad \text{or} \qquad \prod_{t \neq s} p(Y_t | Y_s;\theta)$$

Composite Likelihoods [*Varin* et al. *(2011)*]

General form.

$$CL(Y;\theta) = \prod_k L_k(Y;\theta)^{w_k}, \qquad L_k = p(Y \in \mathcal{A}_k;\theta)$$

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

Composite conditional likelihood.

$$\prod_t p(Y_t | Y^{-t};\theta) \qquad \text{or} \qquad \prod_{t \neq s} p(Y_t | Y_s;\theta)$$

Composite marginal likelihood.

$$\prod_t p(Y_t;\theta), \qquad \prod_{t \neq s} p(Y_t, Y_s;\theta), \qquad \prod_{t \neq s} p(Y_t - Y_s;\theta).$$

General properties

MCLE. Maximum composite likelihood estimate:

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

General properties

MCLE. Maximum composite likelihood estimate:

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

Asymptotic normality. 'Under regularity conditions' to be checked

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

General properties

MCLE. Maximum composite likelihood estimate:

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

Asymptotic normality. 'Under regularity conditions' to be checked

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

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

General properties

MCLE. Maximum composite likelihood estimate:

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

Asymptotic normality. 'Under regularity conditions' to be checked

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

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

Tests. Composite likelihood versions of the Wald test or the likelihood ratio test can be derived but 'suffer from practical limitations' and may involve non-standard distributions.

Asymptotic variance

Reminder on likelihood:

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

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:

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

Asymptotic variance (cont'd)

Reminder on likelihood. Denoting $L'(y;\theta) = \bigtriangledown_\theta L(y;\theta)$:

$$-\mathbb{E}_\theta[\bigtriangledown^2_\theta \log L(Y;\theta)] = \mathbb{E}_\theta\left[\frac{L'(Y;\theta)^2}{L(Y;\theta)^2}\right] - \int L''(y;\theta)\mathrm{d}y$$

Asymptotic variance (cont'd)

Reminder on likelihood. Denoting $L'(y;\theta) = \nabla_\theta L(y;\theta)$:

$$
\begin{aligned}
-\mathbb{E}_\theta[\nabla^2_\theta \log L(Y;\theta)] &= \mathbb{E}_\theta\left[\frac{L'(Y;\theta)^2}{L(Y;\theta)^2}\right] - \int L''(y;\theta)\mathrm{d}y \\
\mathbb{V}_\theta[\nabla_\theta \log L(Y;\theta)] &= \mathbb{E}_\theta\left[\frac{L'(Y;\theta)^2}{L(Y;\theta)^2}\right] - \left[\int L'(y;\theta)\mathrm{d}y\right]^2
\end{aligned}
$$

Asymptotic variance (cont'd)

Reminder on likelihood. Denoting $L'(y;\theta) = \nabla_\theta L(y;\theta)$:

$$-\mathbb{E}_\theta[\nabla^2_\theta \log L(Y;\theta)] = \mathbb{E}_\theta\left[\frac{L'(Y;\theta)^2}{L(Y;\theta)^2}\right] - \int L''(y;\theta)\mathrm{d}y$$

$$\mathbb{V}_\theta[\nabla_\theta \log L(Y;\theta)] = \mathbb{E}_\theta\left[\frac{L'(Y;\theta)^2}{L(Y;\theta)^2}\right] - \left[\int L'(y;\theta)\mathrm{d}y\right]^2$$

Composite likelihood. $\log CL(Y;\theta) = \sum_k \log L_k(Y;\theta)$:

$$-\mathbb{E}_\theta[\nabla^2_\theta \log CL(Y;\theta)] = \mathbb{E}_\theta\left[\sum_k \frac{L'_k(Y;\theta)^2}{L_k(Y;\theta)^2}\right] - \sum_k \int L''_k(y;\theta)\mathrm{d}y$$

Asymptotic variance (cont'd)

Reminder on likelihood. Denoting $L'(y;\theta) = \nabla_\theta L(y;\theta)$:

$$-\mathbb{E}_\theta[\nabla^2_\theta \log L(Y;\theta)] = \mathbb{E}_\theta\left[\frac{L'(Y;\theta)^2}{L(Y;\theta)^2}\right] - \int L''(y;\theta)\mathrm{d}y$$

$$\mathbb{V}_\theta[\nabla_\theta \log L(Y;\theta)] = \mathbb{E}_\theta\left[\frac{L'(Y;\theta)^2}{L(Y;\theta)^2}\right] - \left[\int L'(y;\theta)\mathrm{d}y\right]^2$$

Composite likelihood. $\log CL(Y;\theta) = \sum_k \log L_k(Y;\theta)$:

$$-\mathbb{E}_\theta[\nabla^2_\theta \log CL(Y;\theta)] = \mathbb{E}_\theta\left[\sum_k \frac{L'_k(Y;\theta)^2}{L_k(Y;\theta)^2}\right] - \sum_k \int L''_k(y;\theta)\mathrm{d}y$$

$$\mathbb{V}_\theta[\nabla_\theta \log CL(Y;\theta)] = \mathbb{E}_\theta\left[\sum_k \frac{L'_k(Y;\theta)}{L_k(Y;\theta)}\right]^2 - \left[\sum_k \int L'_k(y;\theta)\mathrm{d}y\right]^2$$

Exercise: AR(1)

Model. With $\{E_t\}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$:

$$(Y_t - \mu) = \phi(Y_{t-1} - \mu) + E_t.$$

Aim. Estimate μ with ϕ and σ^2 known.

Exercise: AR(1)

Model. With $\{E_t\}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$:

$$(Y_t - \mu) = \phi(Y_{t-1} - \mu) + E_t.$$

Aim. Estimate μ with ϕ and σ^2 known.

Log-likelihood. $\log L(Y; \mu) = \log p(Y; \mu) =$

$$\sum_t \log p(Y_t | Y_{t-1}; \mu) \simeq \text{cst} - \frac{1}{2\sigma^2} \sum_t [(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2$$

Exercise: AR(1)

Model. With $\{E_t\}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$:

$$(Y_t - \mu) = \phi(Y_{t-1} - \mu) + E_t.$$

Aim. Estimate μ with ϕ and σ^2 known.

Log-likelihood. $\log L(Y; \mu) = \log p(Y; \mu) =$

$$\sum_t \log p(Y_t | Y_{t-1}; \mu) \simeq \text{cst} - \frac{1}{2\sigma^2} \sum_t [(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2$$

Composite log-likelihood. The stationary variance is $\sigma^2/(1 - \phi^2)$:

$$\log CL(Y; \mu) = \sum_t \log p(Y_t; \mu) = \text{cst} - \frac{1 - \phi^2}{2\sigma^2} \sum_t (Y_t - \mu)^2$$

Exercise: AR(1) (cont'd)

Estimate.

$$\widehat{\mu} = \arg\max_{\mu} L(Y;\mu) = \arg\max_{\mu} CL(Y;\mu) = \frac{1}{n}\sum_t Y_t$$

Exercise: AR(1) (cont'd)

Estimate.

$$\widehat{\mu} = \arg\max_{\mu} L(Y;\mu) = \arg\max_{\mu} CL(Y;\mu) = \frac{1}{n}\sum_t Y_t$$

Likelihood-based variance: $-\mathbb{E}_\mu[-\nabla^2_\mu \log L(Y;\mu)]$

$$\mathbb{V}_L(\widehat{\mu}) = \frac{\sigma^2}{(1-\phi)^2 n} = \mathbb{V}(\widehat{\mu})$$

Exercise: AR(1) (cont'd)

Estimate.

$$\widehat{\mu} = \arg\max_{\mu} L(Y;\mu) = \arg\max_{\mu} CL(Y;\mu) = \frac{1}{n}\sum_t Y_t$$

Likelihood-based variance: $-\mathbb{E}_{\mu}[-\nabla^2_{\mu} \log L(Y;\mu)]$

$$\mathbb{V}_L(\widehat{\mu}) = \frac{\sigma^2}{(1-\phi)^2 n} = \mathbb{V}(\widehat{\mu})$$

Naive composite likelihood-based variance: $-\mathbb{E}_{\mu}[-\nabla^2_{\mu} \log CL(Y;\mu)]$

$$\mathbb{V}^{naive}_{CL}(\widehat{\mu}) = \frac{\sigma^2}{(1-\phi^2)n} \qquad \Rightarrow \qquad \frac{\mathbb{V}^{naive}_{CL}(\widehat{\mu})}{\mathbb{V}_L(\widehat{\mu})} = \frac{1-\phi}{1+\phi}$$

Exercise: AR(1) (cont'd)

Composite likelihood-based variance:

$$\begin{aligned} H(\mu) &= \frac{n(1-\phi^2)}{\sigma^2}, & J(\mu) &= \frac{n^2(1-\phi^2)^2\sigma^2}{(1-\phi)^2} \\ G(\mu) &= \frac{n(1-\phi)^2}{\sigma^2} & \Rightarrow \quad \mathbb{V}_{CL}(\widehat{\mu}) &= G(\mu)^{-1} = \frac{\sigma^2}{n(1-\phi)^2} = \mathbb{V}(\widehat{\mu}) \end{aligned}$$

Exercise: AR(1) (cont'd)

Composite likelihood-based variance:

$$H(\mu) = \frac{n(1-\phi^2)}{\sigma^2}, \qquad J(\mu) = \frac{n^2(1-\phi^2)^2\sigma^2}{(1-\phi)^2}$$

$$G(\mu) = \frac{n(1-\phi)^2}{\sigma^2} \quad \Rightarrow \quad \mathbb{V}_{CL}(\widehat{\mu}) = G(\mu)^{-1} = \frac{\sigma^2}{n(1-\phi)^2} = \mathbb{V}(\widehat{\mu})$$

Remarks.

- Because $\widehat{\mu}_L = \widehat{\mu}_{CL}$ the relative efficiency is $\mathbb{V}(\widehat{\mu}_L)/\mathbb{V}(\widehat{\mu}_{CL}) = 1$.
- $\gamma^2 = \sigma^2/(1-\phi^2)$ could have been estimated as well, but not (σ^2, ϕ).

Exercise: Symmetric multivariate Gaussian

Model. Uniform correlation ρ

$$\{Y_i\} \text{ iid } \sim \mathcal{N}_p(0, \mathbf{R}), \qquad \mathbf{R} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$$

Log-likelihood.

$$\log L(Y;\rho) = \sum_i \log p(Y_i, \rho) = -\frac{n}{2}\log|\mathbf{R}| - \frac{1}{2}\sum_i Y_i'\mathbf{R}^{-1}Y_i$$

Pairwise marginal composite log-likelihood.

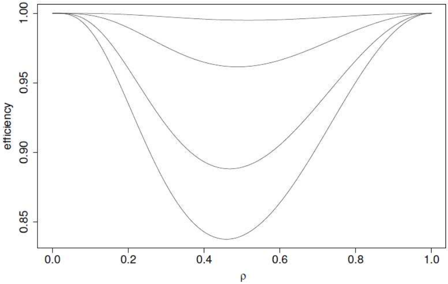
$$\log CL(Y;\rho) = \sum_{j,k}\sum_i \log p(Y_{ij}, Y_{ik};\rho)$$

Exercise: Multivariate Gaussian (cont'd)

Relative efficiency. [*Cox and Reid (2004)*]

$$\frac{\mathbb{V}_\infty(\widehat{\rho}_L)}{\mathbb{V}_\infty(\widehat{\rho}_{CL})} = \frac{[1 + (p-1)\rho]^2[1+\rho^2]^2}{[1+(p-1)\rho^2]C(p,\rho)}$$

$p = 3, 5, 8, 10$

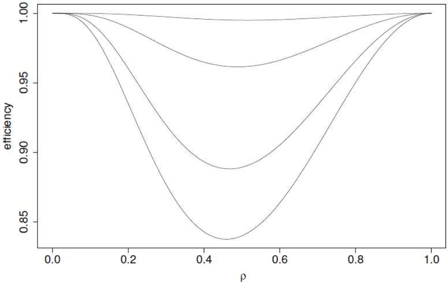


Exercise: Multivariate Gaussian (cont'd)

Relative efficiency. [*Cox and Reid (2004)*]

$$\frac{\mathbb{V}_\infty(\widehat{\rho}_L)}{\mathbb{V}_\infty(\widehat{\rho}_{CL})} = \frac{[1+(p-1)\rho]^2[1+\rho^2]^2}{[1+(p-1)\rho^2]C(p,\rho)}$$

$p = 3, 5, 8, 10$



Remark. $\mathbb{V}_\infty(\widehat{\rho}_L)/\mathbb{V}_\infty(\widehat{\rho}_{CL}) = 1$ for $p = 2$.

Application: Stochastic Block Model

Model.

$$\{Z_i\} \text{ iid } \mathcal{M}(1;\pi), \qquad \{Y_{ij}\} \text{ indep.}|Z, \quad (Y_{ij}|Z_i = k, Z_j = \ell) \sim \mathcal{B}(\gamma_{k\ell}).$$

Likelihood. $\theta = (\pi, \gamma)$

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

$\rightarrow$ Variational EM inference

Application: Stochastic Block Model

Model.

$$\{Z_i\} \text{ iid } \mathcal{M}(1;\pi), \qquad \{Y_{ij}\} \text{ indep.}|Z, \quad (Y_{ij}|Z_i = k, Z_j = \ell) \sim \mathcal{B}(\gamma_{k\ell}).$$

Likelihood. $\theta = (\pi, \gamma)$

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

$\rightarrow$ Variational EM inference

Composite log-likelihood. [*Ambroise and Matias (2011)*]

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

(Triplets of edges are required to guaranty identifiability.)

Application: Multivariate HMM

Model.

$$\{Z_t = (Z_{it})\}_t \sim MC(\pi), \qquad \{Y_{it}\} \text{ indep.}|Z, \quad (Y_{it}|Z_{it} = k) \sim \mathcal{F}(\theta_k).$$

Composite likelihood. [*Gao and Song (2011)*]

$$CL(Y;\theta) = \prod_{i \neq j} p(Y_i, Y_j; \theta)$$

[2] But is $\{(Z_{it}, Z_{jt})\}_t$ a Markov chain?

Application: Multivariate HMM

Model.

$$\{Z_t = (Z_{it})\}_t \sim MC(\pi), \qquad \{Y_{it}\} \text{ indep.}|Z, \quad (Y_{it}|Z_{it} = k) \sim \mathcal{F}(\theta_k).$$

Composite likelihood. [*Gao and Song (2011)*]

$$CL(Y; \theta) = \prod_{i \neq j} p(Y_i, Y_j; \theta)$$

$\rightarrow$ CL-EM algorithm

- E-step: compute via forward-backward[2]

$$p(Z_i, Z_j | Y_i, Y_j; \theta);$$

- M-step: update

$$\widehat{\theta} = \arg\max_\theta \sum_{i \neq j} \mathbb{E}\left[\log p(Y_i, Y_j, Z_i, Z_j; \theta) | Y_i, Y_j\right]$$

[2] But is $\{(Z_{it}, Z_{jt})\}_t$ a Markov chain?

Some Links? [*Lyu (2011)*]

Some Links? [*Lyu (2011)*]

Variational methods allow to deal with complex dependency structures by breaking down dependencies and provide efficient algorithms, but with almost no guaranty as for the parameter estimates.

Some Links? [*Lyu (2011)*]

Variational methods allow to deal with complex dependency structures by breaking down dependencies and provide efficient algorithms, but with almost no guaranty as for the parameter estimates.

Composite likelihood methods allow to deal with complex dependency structures by breaking down dependencies and provide guaranties as for the parameter estimates.

Some Links? [*Lyu (2011)*]

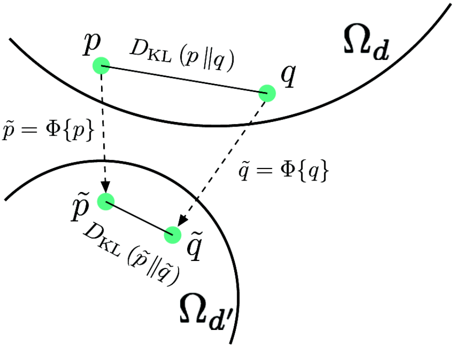
Variational methods allow to deal with complex dependency structures by breaking down dependencies and provide efficient algorithms, but with almost no guaranty as for the parameter estimates.

Composite likelihood methods allow to deal with complex dependency structures by breaking down dependencies and provide guaranties as for the parameter estimates.

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

KL contraction

Definition [*Lyu (2011)*]. Denote Ω_d the set of all distributions over $\mathbb{R}^d$.



$\Phi : \Omega_d \mapsto \Omega_{d'}$ 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

For a given distribution $t(y|x)$

Examples of KL contraction

For a given distribution $t(y|x)$

- Marginal distribution: $\Phi^m_A\{p\}(x) = \int p(x)\mathrm{d}x_{\setminus A}$.

Examples of KL contraction

For a given distribution $t(y|x)$

- Marginal distribution: $\Phi_A^m\{p\}(x) = \int p(x)\mathrm{d}x_{\setminus A}$.
- Conditional distribution: $\Phi_t^c\{p\}(y) = \int p(x)t(y|x)\mathrm{d}x$.

Examples of KL contraction

For a given distribution $t(y|x)$

- Marginal distribution: $\Phi_A^m\{p\}(x) = \int p(x)\mathrm{d}x_{\setminus A}$.
- Conditional distribution: $\Phi_t^c\{p\}(y) = \int p(x)t(y|x)\mathrm{d}x$.
- Marginal grafting: (replaces $p_A(x_A)$ with $t_A(x_A)$)

$$\Phi_{t,A}^g\{p\}(x) = p(x)\frac{t_A(x_A)}{p_A(x_A)} = t_A(x_A)p_{\setminus A|A}(x_{\setminus A}|x_A).$$

Examples of KL contraction

For a given distribution $t(y|x)$

- Marginal distribution: $\Phi_A^m\{p\}(x) = \int p(x)\mathrm{d}x_{\setminus A}$.
- Conditional distribution: $\Phi_t^c\{p\}(y) = \int p(x)t(y|x)\mathrm{d}x$.
- Marginal grafting: (replaces $p_A(x_A)$ with $t_A(x_A)$)

$$\Phi_{t,A}^g\{p\}(x) = p(x)\frac{t_A(x_A)}{p_A(x_A)} = t_A(x_A)p_{\setminus A|A}(x_{\setminus A}|x_A).$$

- Binary mixture: $\Phi_t^b\{p\}(x) = \pi t(x) + (1-\pi)p(x)$.

Examples of KL contraction

For a given distribution $t(y|x)$

- Marginal distribution: $\Phi_A^m\{p\}(x) = \int p(x)\mathrm{d}x_{\setminus A}$.
- Conditional distribution: $\Phi_t^c\{p\}(y) = \int p(x)t(y|x)\mathrm{d}x$.
- Marginal grafting: (replaces $p_A(x_A)$ with $t_A(x_A)$)
$$\Phi_{t,A}^g\{p\}(x) = p(x)\frac{t_A(x_A)}{p_A(x_A)} = t_A(x_A)p_{\setminus A|A}(x_{\setminus A}|x_A).$$
- Binary mixture: $\Phi_t^b\{p\}(x) = \pi t(x) + (1-\pi)p(x)$.
- Lumping (= discretization): $\mathcal{S} = (S_1, \dots S_m)$ a partition of $\mathbb{R}^d$,
$$\Phi_{\mathcal{S}}^\ell\{p\}(i) = \int_{S_i} p(x)\mathrm{d}x.$$

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

$$KL(p||q_\theta) - \beta\; KL(\Phi\{p\}||\Phi\{q_\theta\}) \qquad (1)$$

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

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

$$\arg\min_\theta \sum_k w_k \left[KL(p||q_\theta) - \beta_k KL(\Phi_k\{p\}||\Phi_k\{q_\theta\})\right].$$

[*Lyu (2011)*]

Links with composite likelihoods

Marginal contraction $\to$ Conditional composite likelihood: For subsets $A_1, \dots A_K$, p being the true distribution,

$$\arg\min_\theta KL(p||q_\theta) - \sum_k w_k KL(\Phi^m_{A_k}\{p\}||\Phi^m_{A_k}\{q_\theta\})$$

$$= \arg\min_\theta \int p(x) \log \frac{p(x)}{q(x;\theta)} \mathrm{d}x - \sum_k w_k \int p(x) \log \frac{p_{A_k}(x_{A_k})}{q_{A_k}(x_{A_k};\theta)} \mathrm{d}x$$

$$= \arg\max_\theta \sum_k w_k \int p(x) \log \frac{q(x;\theta)}{q_{A_k}(x_{A_k};\theta)} \mathrm{d}x \qquad (p \text{ does not depend on } \theta)$$

$$= \arg\max_\theta \sum_k w_k \int p(x) \log q_{\setminus A_k|A_k}(x_{\setminus A_k}|x_{A_k};\theta) \mathrm{d}x$$

$$\approx \arg\max_\theta \sum_k w_k \frac{1}{n} \sum_i \log q_{\setminus A_k|A_k}(x^i_{\setminus A_k}|x^i_{A_k};\theta) \mathrm{d}x \qquad (p \to \text{empirical dist.})$$

Links with composite likelihoods (cont'd)

Marginal grafting → Marginal composite likelihood:

$$\arg\min_\theta KL(p||q_\theta) - \sum_k w_k KL(\Phi^g_{p,A_k}\{p\}||\Phi^g_{p,A_k}\{q_\theta\})$$

$$= \arg\min_\theta \sum_k w_k KL(\Phi^m_{A_k}\{p\}||\Phi^m_{A_k}\{q_\theta\}) \qquad \text{(cf Lemma 2)}$$

$$= \arg\max_\theta \sum_k w_k \int p_{A_k}(x_{A_k}) \log q_{A_k}(x_{A_k};\theta) \qquad (p \text{ does depend on } \theta)$$

$$\approx \arg\max_\theta \sum_k w_k \frac{1}{n} \sum_i \log q_{A_k}(x^i_{A_k};\theta) \qquad (p \to \text{empirical dist.})$$

Conclusion: There is no conclusion

Connexions do exist. Some variational approximations of the likelihood are actually composite likelihoods.

Conclusion: There is no conclusion

Connexions do exist. Some variational approximations of the likelihood are actually composite likelihoods.

But is it the case of your favorite one?

$$\min KL(q_\theta||p) \neq \min KL(p||q_\theta) \neq KL(p||q_\theta) - \beta KL(\Phi\{p\}||\Phi\{q_\theta\}).$$

Conclusion: There is no conclusion

Connexions do exist. Some variational approximations of the likelihood are actually composite likelihoods.

But is it the case of your favorite one?

$$\min KL(q_\theta || p) \neq \min KL(p || q_\theta) \neq KL(p || q_\theta) - \beta KL(\Phi\{p\} || \Phi\{q_\theta\}).$$

No nice example to show. Not been able to derive the Godambe matrix for a given variational approximation.

$\rightarrow$ Worth trying?

[*Lyu (2011)*]: '*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.*' ...

References

Ambroise, C. and Matias, C. (2011). New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. no–no.

Celisse, A., Daudin, J.-J. and Pierre, L., (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model.

Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*. **91 (3)** 729.

Gao, X. and Song, P. X.-K. (2011). Composite likelihood em algorithm with applications to multivariate hidden markov model. *Statistica Sinica*. **21 (1)** 165–185.

Lyu, S. (2011). Unifying non-maximum likelihood learning objectives with minimum KL contraction. In *NIPS*, (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, ed.), 64–72.

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

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

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

Appendix: Symmetric multivariate Gaussian

Covariance matrix.

$$\mathbf{R} = (1-\rho)\mathbf{I} + \rho\mathbf{J},$$

$$\mathbf{R}^{-1} = (1-\rho)^{-1}\left(\mathbf{I} - \frac{\rho}{1+(p-1)\rho}\mathbf{J}\right),$$

$$|\mathbf{R}| = (1-\rho)^{p-1}[1+(p-1)\rho]$$