Consistency of ELBO maximization for model selection

Badr-Eddine Chérief-Abdellatif

CREST
CENTER FOR RESEARCH IN ECONOMICS AND STATISTICS

ENSAE ParisTech
École nationale de la statistique et de l'administration économique
université PARIS-SACLAY

Symposium on Advances in Approximate Bayesian Inference
Montréal, Canada
December 2, 2018

Outline of the talk

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P^0 in a model $\mathcal{M}_K = \{P_\theta, \theta \in \Theta_K\}$ associated with a likelihood L_n. We define a prior Π_K on Θ_K.

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P^0 in a model $\mathcal{M}_K = \{P_\theta, \theta \in \Theta_K\}$ associated with a likelihood L_n. We define a prior Π_K on Θ_K.

The posterior

$$\pi_n^K(\mathrm{d}\theta) \propto L_n(\theta)\Pi_K(\mathrm{d}\theta).$$

Notations

Assume that we observe $X_1, \ldots, X_n$ i.i.d from P^0 in a model $\mathcal{M}_K = \{P_\theta, \theta \in \Theta_K\}$ associated with a likelihood L_n. We define a prior Π_K on Θ_K.

The posterior

$$\pi_n^K(\mathrm{d}\theta) \propto L_n(\theta)\Pi_K(\mathrm{d}\theta).$$

The tempered posterior - $0 < \alpha < 1$

$$\pi_{n,\alpha}^K(\mathrm{d}\theta) \propto [L_n(\theta)]^\alpha \Pi_K(\mathrm{d}\theta).$$

Various reasons to use a tempered posterior

- Easier to sample from

G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

Various reasons to use a tempered posterior

- Easier to start from

G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

- Robust to model misspecification

P. Grünwald and T. Van Ommen (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*.

Various reasons to use a tempered posterior

- Easier to sample from

G. Behrens, N. Friel & M. Hurn. (2012). Tuning tempered transitions. *Statistics and Computing*.

- Robust to model misspecification

P. Grünwald and T. Van Ommen (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*.

- Theoretical analysis easier

A. Bhattacharya, D. Pati & Y. Yang (2016). Bayesian fractional posteriors. *Preprint arxiv :1611.01125*.

Variational Bayes

$$\begin{aligned}\tilde{\pi}^{K}_{n,\alpha} &= \arg\min_{\rho \in \mathcal{F}_K} \mathcal{K}(\rho, \pi^{K}_{n,\alpha}) \\ &= \arg\max_{\rho \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho, \Pi_K) \right\}.\end{aligned}$$

Variational Bayes

$$\begin{aligned}\tilde{\pi}_{n,\alpha}^{K} &= \arg\min_{\rho\in\mathcal{F}_K} \mathcal{K}(\rho, \pi_{n,\alpha}^{K}) \\ &= \arg\max_{\rho\in\mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho, \Pi_K) \right\}.\end{aligned}$$

Examples :

Variational Bayes

$$
\begin{aligned}
\tilde{\pi}_{n,\alpha}^{K} &= \arg\min_{\rho \in \mathcal{F}_K} \mathcal{K}(\rho, \pi_{n,\alpha}^{K}) \\
&= \arg\max_{\rho \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho, \Pi_K) \right\}.
\end{aligned}
$$

Examples :

- parametric approximation

$$
\mathcal{F}_K = \left\{ \mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+ \right\}.
$$

Variational Bayes

$$\begin{aligned}\tilde{\pi}^K_{n,\alpha} &= \arg\min_{\rho \in \mathcal{F}_K} \mathcal{K}(\rho, \pi^K_{n,\alpha}) \\ &= \arg\max_{\rho \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta)\rho(\mathrm{d}\theta) - \mathcal{K}(\rho, \Pi_K) \right\}.\end{aligned}$$

Examples :

- parametric approximation

$$\mathcal{F}_K = \left\{ \mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+ \right\}.$$

- mean-field approximation, $\Theta = \Theta_1 \times \Theta_2$ and

$$\mathcal{F}_K = \{ \rho : \rho(\mathrm{d}\theta) = \rho_1(\mathrm{d}\theta_1) \times \rho_2(\mathrm{d}\theta_2) \}.$$

Model selection

Several models

$$\{\mathcal{M}_K / K \in \mathbb{N}^*\}$$

Model selection

Several models

$$\{\mathcal{M}_K / K \in \mathbb{N}^*\}$$

Prior

$$\pi = \sum_{K \in \mathbb{N}^*} \pi_K \Pi_K$$

Model selection

Several models

$$\{\mathcal{M}_K / K \in \mathbb{N}^*\}$$

Prior

$$\pi = \sum_{K \in \mathbb{N}^*} \pi_K \Pi_K$$

Tempered posteriors & their VBs

$$\pi^K_{n,\alpha}(d\theta_K) \propto L_n(\theta_K)^\alpha \Pi_K(d\theta_K)$$

and

$$\tilde{\pi}^K_{n,\alpha} = \arg\min_{\rho_K \in \mathcal{F}_K} \mathcal{K}(\rho_K, \pi^K_{n,\alpha})$$

ELBO criterion

ELBO maximization program

$$\tilde{\pi}_{n,\alpha}^{K} = \arg\max_{\rho_K \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta_K) \rho_K(d\theta_K) - \mathcal{K}(\rho_K, \Pi_K) \right\}$$

ELBO criterion

ELBO maximization program

$$\tilde{\pi}_{n,\alpha}^{K} = \arg\max_{\rho_K \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta_K) \rho_K(d\theta_K) - \mathcal{K}(\rho_K, \Pi_K) \right\}$$

ELBO

$$\mathcal{L}(K) = \alpha \int \ell_n(\theta_K) \tilde{\pi}_{n,\alpha}^{K}(d\theta_K) - \mathcal{K}(\tilde{\pi}_{n,\alpha}^{K}, \Pi_K)$$

ELBO criterion

ELBO maximization program

$$\tilde{\pi}_{n,\alpha}^{K} = \arg\max_{\rho_K \in \mathcal{F}_K} \left\{ \alpha \int \ell_n(\theta_K) \rho_K(d\theta_K) - \mathcal{K}(\rho_K, \Pi_K) \right\}$$

ELBO

$$\mathcal{L}(K) = \alpha \int \ell_n(\theta_K) \tilde{\pi}_{n,\alpha}^{K}(d\theta_K) - \mathcal{K}(\tilde{\pi}_{n,\alpha}^{K}, \Pi_K)$$

Model selection criterion

$$\hat{K} = \arg\max_{K \geq 1} \left\{ \mathcal{L}(K) - \log\left(\frac{1}{\pi_K}\right) \right\}$$

Technical condition for posterior concentration

If there is a true model ($\exists K_0$, $\exists \theta^0 \in \Theta_{K_0}$, $P_{\theta^0} = P^0$) :

Technical condition for posterior concentration

If there is a true model $(\exists K_0,\ \exists \theta^0 \in \Theta_{K_0},\ P_{\theta^0} = P^0)$:

Prior mass condition for concentration of tempered posteriors

The rate (r_n) is such that

$$\Pi_{K_0}[\mathcal{B}(r_n)] \geq e^{-nr_n}$$

where $\mathcal{B}(r) = \{\theta \in \Theta_{K_0} : \mathcal{K}(P_{\theta^0}, P_\theta) \leq r\}$.

Technical condition for posterior concentration

If there is a true model $(\exists K_0,\ \exists \theta^0 \in \Theta_{K_0},\ P_{\theta^0} = P^0)$:

Prior mass condition for concentration of tempered posteriors

The rate (r_n) is such that

$$\Pi_{K_0}[\mathcal{B}(r_n)] \geq e^{-nr_n}$$

where $\mathcal{B}(r) = \{\theta \in \Theta_{K_0} : \mathcal{K}(P_{\theta^0}, P_\theta) \leq r\}$.

Prior mass condition for concentration of Variational Bayes

The rate (r_n) is such that there exists $\rho_{n,K_0} \in \mathcal{F}_{K_0}$ such that

$$\int \mathcal{K}(P_{\theta^0}, P_\theta)\rho_{n,K_0}(\mathrm{d}\theta) \leq r_n, \ \text{ and } \ \mathcal{K}(\rho_{n,K_0}, \Pi_{K_0}) \leq nr_n.$$

Consistency of the true approximation

P. Alquier & J. Ridgway (2017). Concentration of Tempered Posteriors and of their Variational Approximations. *Preprint arxiv :1706.09293*.

Theorem

If there is a true model ($\exists K_0$, $\exists \theta^0 \in \Theta_{K_0}$, $P_{\theta^0} = P^0$), then under the prior mass condition, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}^{K_0}_{n,\alpha}(d\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n.$$

Consistency of the selected approximation

Theorem

If there is a true model ($\exists K_0,\ \exists \theta^0 \in \Theta_{K_0},\ P_{\theta^0} = P^0$), then under the prior mass condition, for any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta)\right] \leq \frac{1+\alpha}{1-\alpha} r_n + \frac{\log(\frac{1}{\pi_{K_0}})}{n(1-\alpha)}.$$

Robustness to misspecification

Theorem

For any $\alpha \in (0,1)$, for any K, for any $\rho_K \in \mathcal{F}_K$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta)\right] \leq \frac{\alpha}{1-\alpha}\int \mathcal{K}(P^0, P_{\theta_K})\rho_{n,K}(d\theta_K) + \frac{\mathcal{K}(\rho_K, \Pi_K)}{n(1-\alpha)} + \frac{\log(\frac{1}{\pi_K})}{n(1-\alpha)}.$$

Example : Univariate Gaussian mixture models

The true distribution P^0 is such that $\mathbb{E}|X| < +\infty$.
Let $L \geq 1$, $\pi_K = 2^{-K}$, $\Pi_K = \mathcal{D}_K(\alpha_1, \ldots, \alpha_K) \bigotimes \mathcal{N}(0, \mathcal{V}^2)^{\otimes n}$ and

$$r_{n,K} = \left[\frac{8K \log(nK)}{n} \bigvee \left(\frac{8K \log(n\mathcal{V})}{n} + \frac{8KL^2}{n\mathcal{V}^2} \right) \right] + \frac{K \log(2)}{n(1-\alpha)}.$$

Theorem

For any $\alpha \in (0,1)$,

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \tilde{\pi}^{\hat{K}}_{n,\alpha}(d\theta) \right] \leq \inf_{K \geq 0} \left\{ \frac{\alpha}{1-\alpha} \inf_{\theta^* \in \mathcal{S}_K \times [-L,L]^K} \mathcal{K}(P^0, P_{\theta^*}) + \frac{1+\alpha}{1-\alpha} r_{n,K} \right\}.$$

1. Tempered Variational Bayes
   - Tempered posteriors
   - Variational Bayes
2. Model Selection
   - Framework
   - ELBO criterion
3. Consistency result
   - Main result
   - Application

Applications

If there is a true model $(\exists K_0,\, \exists \theta^0 \in \Theta_{K_0},\, P_{\theta^0} = P^0)$:

Applications

If there is a true model $(\exists K_0,\ \exists \theta^0 \in \Theta_{K_0},\ P_{\theta^0} = P^0)$:

Gaussian mixtures : $\theta = (p, (m_1, \sigma_1^2), ..., (m_K, \sigma_K^2))$

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0) \tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta) \right] = \mathcal{O}\left(\frac{K_0 \log(nK_0)}{n} \right)$$

Applications

If there is a true model ($\exists K_0,\ \exists \theta^0 \in \Theta_{K_0},\ P_{\theta^0} = P^0$) :

Gaussian mixtures : $\theta = (p, (m_1, \sigma_1^2), ..., (m_K, \sigma_K^2))$

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta)\right] = \mathcal{O}\left(\frac{K_0 \log(nK_0)}{n}\right)$$

Probabilistic PCA : $\theta \in \mathbb{R}^{d\times K}$

$$\mathbb{E}\left[\int D_\alpha(P_\theta, P^0)\tilde{\pi}_{n,\alpha}^{\hat{K}}(d\theta)\right] = \mathcal{O}\left(\frac{dK_0 \log(dn)}{n}\right)$$

Thank you !