Estimation With Variational Bayesian Methods

Silèye Ba and Laurent Girin

Perception, INRIA Grenoble Rhône-Alpes

EARS Technical Meeting
Aldebaran
27/10/2015

Talk outline

1. Bayesian inference
2. Variational Bayesian principles
3. Variational offline learning of a Gaussian mixture model
4. Variational online estimation of a linear dynamical system
5. Summary and conclusions

Outline

1. Bayesian inference
2. Variational Bayesian principles
3. Variational offline learning of a Gaussian mixture model
4. Variational online estimation of a linear dynamical system
5. Summary and conclusions

Being a Bayesian

We are concerned with data Y and learning about its generating process which is driven by unknown parameters θ.

Bayesian modelling

- Use probabilities to quantify our belief about uncertainties
- Observation model $p(Y|\theta)$: belief about data
- Model prior $p(\theta)$: prior belief before data is available
- A posteriori model belief: Bayesian inversion $p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{\int_\theta p(Y|\theta)p(\theta)d\theta}$

Estimation

- Parameter marginalization: $p(y) = \int_\theta p(y|\theta)p(\theta)d\theta$
- Problem: usually involved intractable integrals
- Solution: Laplace approximation, Monte Carlo integration, variational Bayesian approximation

Being Bayesian: linear regression example

Linear regression model

$y_n = x_n^\top \theta + \epsilon_n$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$
$x = (1, x^{(1)}, ..., x^{(d)})^\top, \theta = (\theta^{(0)}, \theta^{(1)}, ..., \theta^{(d)})^\top$

Data and prior models

Data : $p(y|x, \theta) = g(y; x_n^\top \theta, \sigma^2)$
Prior : $p(\theta) = g(\theta; m_0, P_0)$

Parameter distribution given data $p(\theta|y_{1:N}, x_{1:N}) = g(\theta; m, P)$

$$P^{-1} = P_0^{-1} + \frac{1}{\sigma^2} X^\top X$$

$$m = P^{-1}(P_0^{-1} m_0 + \sigma^2 X^\top y)$$

$$X = \begin{pmatrix} 1 & x_1^{(1)} & \dots & x_1^{(d)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N^{(N)} & \dots & x_N^{(d)} \end{pmatrix} \text{ and } Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

Bayesian solution

$$p(y|x) = \int g(y; x^\top \theta, \sigma^2) g(\theta; m; P) \mathrm{d}\theta$$

Talk outline

1 Bayesian inference

2 Variational Bayesian principles

3 Variational offline learning of a Gaussian mixture model

4 Variational online estimation of a linear dynamical system

5 Summary and conclusions

Variational Bayesian estimation

Principles

- Approximate the a posteriori parameter distribution in factorized form

$$p(\theta|Y) \approx q(\theta) = q(\theta_1)q(\theta_2)$$

- Optimality is with respect to the Kullback Liebler divergence

$$\hat{q}(\theta) = \underset{q(\theta_1)q(\theta_2)}{\arg\min}\, d_{KL}\left(q(\theta_1)q(\theta_2), p(\theta|Y)\right)$$

- Kullback-Liebler divergence:

$$d_{KL}(q(\theta), p(\theta|Y)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta|Y)} d\theta$$

Remarks

- KL divergence is not symmetric: input ordering matters
- θ has to be multivariate in order VB approximation to be applicable

The variational Bayesian theorem

Let $p(\theta|Y)$ be the posterior distribution of a multivariate parameter θ. The latter is partitioned into K sub-vectors of parameters:

$$\theta = (\theta_1, \theta_2, ..., \theta_K).$$

Let $q(\theta)$ be an approximate distribution restricted to the set of conditionally independent distribution for θ_1,θ_2,...,θ_K:

$$q(\theta) = q(\theta_1, \theta_2, ..., \theta_K) = \prod_{k=1}^{K} q(\theta_k).$$

Then the minimum of the variational Bayesian approximation problem is reached for

$$\hat{q}(\theta_k) \propto \exp\left(\mathbb{E}_{\hat{q}(\theta_{1:K/k})} \log p(\theta|Y)\right),$$

where $\theta_{1:K/k}$ denotes the complement of θ_k in the parameter vector θ, and $\hat{q}(\theta/k) = \prod_{l \neq k} \hat{q}(\theta_l)$.

Iterative variational Bayesian algorithm

IVB Algorithm

1. Choose a partition of $\theta = (\theta_1, \theta_2, ..., \theta_K)$
2. For $k = 1, ..., K$ "choose" parametric forms for $q(\theta_k)$
3. For $k = 1, ..., K$ compute

$$\hat{q}(\theta_k) \propto \exp\left(\int \hat{q}(\theta_{1:K/k}) \log p(\theta | Y) d\theta_{1:K/k} \right)$$

4. Loop to step 3 until convergence

Remarks

- The variational Bayesian approximation is deterministic;
- Cross-correlation between θ_k are not present in the approximation, the variables partition has to be chosen judiciously;
- IVB algorithm is reminiscent of varational E.M. algorithm however for IVB only expectations are computed. However, in variational E.M., maximization steps can applied to hyper-parameters which have no prior distribution.

Talk outline

1. Bayesian inference
2. Variational Bayesian principles
3. Variational offline learning of a Gaussian mixture model
4. Variational online estimation of a linear dynamical system
5. Summary and conclusions

Variational learning of a Gaussian mixture model (GMM): problem 1/2

The problem

- Data points $y_{1:N} = \{y_1, y_2, ..., y_N\}$ assumed to be drawn from a K component GMM: $y_n \sim \sum_{k=1}^{K} \pi_k g(y_n; \mu_k, \Sigma_k)$;
- Goal: estimate parameter vector $\theta = (\pi_k, \mu_k, k = 1, ..., K)$;
- Introduce latent variables $Z = \{z_1, z_2, ..., z_N\}$ assigning every data point y_n to one of the K mixtures component: $p(z_n = k) = \pi_k$.

A posteriori distribution $p(\theta, z_{1:N}|y_{1:N})$

$$
\begin{aligned}
p(\theta, z_{1:N}|y_{1:N}) &\propto p(y_{1:N}|\theta, z_{1:N}) p(z_{1:N}|\theta) p(\theta) \\
&\propto p(\theta) \prod_{n=1}^{N} p(y_n|z_n, \mu_{1:K}, \Sigma_{1:K}) p(z_n|\pi_{1:K}) \\
&\propto p(\theta) \prod_{n=1}^{N} \prod_{k=1}^{K} (p(y_n|z_n = k, \mu_k, \Sigma_k) p(z_n = k|\pi_{1:K}))^{\delta_k(z_n)}
\end{aligned}
$$

Variational learning of a GMM: problem 2/2

Parameters prior distribution $p(\theta)$

- Parameters prior distribution: $p(\theta) = p(\pi) \prod_{k=1}^{K} p(\mu_k)$.
- Dirichlet over mixture weights: $p(\pi) = c(\alpha) \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$
- Gaussian over Gaussian means: $p(\mu_k) = g(\mu_k; \nu_k, \Lambda_k)$

Variational approximating distribution $q(\theta, z_{1:N})$

$$p(\theta, z_{1:N}|y_{1:N}) \approx q(\theta, z_{1:N}) = q(\pi) \prod_{k=1}^{K} q(\mu_k) \prod_{n=1}^{N} q(z_n)$$

Using the IVB procedure we iteratively compute $q(z_n)$, $q(\pi)$, $q(\mu_k)$.

Variational learning of a GMM: computing $\hat{q}(z_n)$

$$\log p(\theta, z_{1:N}|y_{1:N}) = \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k) + \sum_{n=1}^{N} \sum_{k=1}^{K} \delta_k(z_n)(\log g(y_n; \mu_k, \Sigma_k) + \log \pi_k))$$

According to the VB theorem

$$\begin{aligned} \log \hat{q}(z_n) &= \mathbb{E}_{q(\theta, z_{1:N}/z_n)}[\log p(\theta, z_{1:N}|y_{1:N})] \\ &= \sum_{k=1}^{K} \delta_k(z_n)(\mathbb{E}[\log g(y_n; \mu_k, \Sigma_k)] + \log \pi_k)) + C \end{aligned}$$

If we denote $\rho_{nk} = \mathbb{E}[\log g(y_n; \mu_k, \Sigma_k)] + \log \pi_k$ then:

$$\hat{q}(z_n) \propto \prod_{k=1}^{K} \rho_{nk}^{\delta_k(z_n)}$$

$$\hat{q}(z_n = k) = a_{nk} = \frac{\rho_{nk}}{\sum_{l=1}^{K} \rho_{nl}}$$

Variational learning of a GMM: computing $\hat{q}(\pi)$

According to the VB theorem:

$$\log \hat{q}(\pi) = \mathbb{E}_{q(\theta, z_{1:N}/\pi)}[\log p(\theta, z_{1:N}|y_{1:N})]$$

which gives:

$$\log \hat{q}(\pi) = \sum_{k=1}^{K}(\alpha_k - 1)\log \pi_k + \sum_{n=1}^{N}\sum_{k=1}^{K} a_{nk} \log \pi_k + C$$

This implies that

$$\hat{q}(\pi) = c(\beta)\prod_{k=1}^{K} \pi_k^{\beta_k - 1} \text{ with } \beta_k = \sum_{n=1}^{N} a_{nk} + \alpha_k$$

$\hat{q}(\pi)$ is the density of a Dirichlet distribution $\mathcal{D}(\beta)$

Variational learning of a GMM: computing $\hat{q}(\mu_k)$

According to the VB theorem:

$$\log \hat{q}(\mu_k) = \mathbb{E}_{q(\theta, z_{1:N}/\mu_k)}[\log p(\theta, z_{1:N}|y_{1:N})]$$

which, denoting $N_k = \sum_{n=1}^{N} a_{nk}$, gives:

$$\log \hat{q}(\mu_k) = \mu_k^\top (\Lambda_k^{-1} + N_k \Sigma_k^{-1})\mu_k - 2\mu_k^\top (\Lambda_k^{-1}\nu_k + \sum_{n=1}^{N} a_{nk}\Sigma_k^{-1} y_n) + C$$

If we denote by

$$V_k^{-1} = \Lambda_k^{-1} + N_k \Sigma_k^{-1}$$

$$m_k = V_k(\Lambda_k^{-1}\nu_k + \sum_{n=1}^{N} a_{nk}\Sigma_k^{-1} y_n)$$

$\hat{q}(\mu_k)$ is a Gaussian density $g(\mu_k; m_k, V_k)$.

Talk outline

1 Bayesian inference

2 Variational Bayesian principles

3 Variational offline learning of a Gaussian mixture model

4 Variational online estimation of a linear dynamical system

5 Summary and conclusions

Variational online estimation of a linear dynamical system (LDS): pedagogical example

Problem

- Temporal observations $y_1,...,y_t$ emitted by a hidden state X_t;
- Observation model: $y_t = HX_t + \eta_t$, $\eta_t \sim \mathcal{N}(0, \Sigma)$;
- Dynamical model: $X_t = GX_{t-1} + \omega_t$, $\omega_t \sim \mathcal{N}(0, \Lambda)$;
- Goal: estimate joint a-posteriori distribution $p(X_t, X_{t-1}|y_1, ..., y_t)$.

A posteriori distribution

Previous time state filtering distribution approximation is available $p(X_{t-1}|y_{1:t-1}) \approx q(X_{t-1}) = g(X_{t-1}; m_{t-1}, V_{t-1})$ then

$$\begin{aligned} p(X_t, X_{t-1}|y_1, ..., y_t) &\propto p(y_t|X_t)p(X_t|X_{t-1})p(X_{t-1}|y_{1:t-1}) \\ &\propto p(y_t|X_t)p(X_t|X_{t-1})q(X_{t-1}) \end{aligned}$$

Variational approximating distribution $q(X_t, X_{t-1})$

$$p(X_t, X_{t-1}|y_{1:t}) \approx q(X_t, X_{t-1}) = q(X_t)q_p(X_{t-1}) \text{ where}$$
$$q_p(X_{t-1}) \approx p(X_{t-1}|y_{1:t})$$

Variational online estimation of an LDS: computing $\hat{q}(X_{t-1})$

Log of aposteriori distribution

$$\log p(X_t, X_{t-1}|y_{1:t}) \approx \log g(y_t; HX_t, \Sigma) + \log g(X_t; GX_{t-1}, \Lambda) + \log g(X_{t-1}; m_{t-1}, V_{t-1})$$

According to the VB theorem:

$$\begin{aligned}\log \hat{q}_p(X_{t-1}) &= \mathbb{E}_{q(X_t)}[\log p(X_t, X_{t-1}|y_{1:t})] \\ &= X_{t-1}^\top (G^\top \Lambda^{-1} G + V_{t-1}^{-1}) X_{t-1} - 2X_{t-1}^\top (G^\top \Lambda^{-1} m_t + V_{t-1}^{-1} m_{t-1})\end{aligned}$$

where $m_t = \mathbb{E}_{q(X_t)}[X_t]$.

This means $\hat{q}_p(X_{t-1}) = g(X_{t-1}; m_{t-1|t}, V_{t-1|t})$where:

$$\begin{aligned}V_{t-1|t} &= (G^\top \Lambda^{-1} G + V_{t-1}^{-1})^{-1} \\ m_{t-1|t} &= V_{t-1|t}(G^\top \Lambda^{-1} m_t + V_{t-1}^{-1} m_{t-1})\end{aligned}$$

Variational online estimation of an LDS: computing $\hat{q}(X_t)$

According to the VB theorem:

$$\begin{aligned}\log \hat{q}(X_t) &= \mathbb{E}_{q_p(X_{t-1})}[\log p(X_t, X_{t-1}|y_{1:t})] \\ &= X_t^\top(\Lambda^{-1} + H^\top \Sigma^{-1} H)X_t - 2X_t^\top(\Lambda^{-1} G m_{t-1|t} + H^\top \Sigma^{-1} y_t)\end{aligned}$$

This means $\hat{q}(X_t) = g(X_t; m_t, V_t)$ where:

$$\begin{aligned}V_t &= (\Lambda^{-1} + H^\top \Sigma^{-1} H)^{-1} \\ m_t &= V_t(\Lambda^{-1} G m_{t-1|t} + H^\top \Sigma^{-1} y_t)\end{aligned}$$

Talk outline

1. Bayesian inference
2. Variational Bayesian principles
3. Variational offline learning of a Gaussian mixture model
4. Variational online estimation of a linear dynamical system
5. Summary and conclusions

Summary and conclusions

Summary

- Bayesian philosophy for estimation
- Variational Bayesian as solution
- Alternative: Monte Carlo sampling
- Two examples: offline GMM and online LDS estimation
- Two building blocks can be used for more complex models: switching Kalman filters, locally linear affine mapping

Following: variational Bayesian in EARS Project

- Laurent Girin, variational Bayesian for sound source separation
- Sileye Ba, variational Bayesian for multi-person tracking