Variational Bayesian inference for mixture of Gaussian process models under Dirichlet process prior

Presenter: Peng Chen
Materials Process Design and Control Laboratory
Sibley School of Mechanical and Aerospace Engineering
101 Frank H. T. Rhodes Hall, Cornell University
Ithaca, NY 14853-3801
Reference

- “Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow prediction” – by Shiliang Sun, Xin Xu, IEEE transactions on intelligent transportation systems, Vol. 12, No. 2, 2011.


Outline

- Motivation
- Problem definition
- Variational inference
  - Mean field variational method
  - Nonparametric variational method
- Gaussian Process
- Dirichlet Process
Motivation

Multiscale deformation problem:

Microstructure

Uncertain deformation process

mechanical responses

Porous media flow problem:

permeability

Uncertain boundary conditions

physical response

More problem: Robotic learning, financial modeling, dynamic model and etc.
## Motivation

- Dealing with
  - Non-stationary covariance function
  - Discontinuities
  - Multimodality
  - Overlapping output signals

### Why variational method?
- Provide analytical approximation of the posterior
- Super fast compared to MCMC or Gibbs sampling and etc.

### Why Gaussian Process?
- Flexible nonparametric nature
- Computational simplicity

### Why Dirichlet process prior?
- Automatically determine the number of mixture components
- Conjugacy to multinomial distribution
Problem definition

Let’s use $\mathbf{X}$ to denote the input, $\mathbf{Y}$ to denote the output.

In general, $\mathbf{X}$ is in high dimension, so we consider a lower dimension representation of the input:

$$\mathbf{x} = C(\xi) = \sum_{i=1}^{M} \xi_i \psi_i + \mu = \psi \xi + \mu$$

In this work, we consider a mixture of input model:

$$\xi_i \sim \sum_{k=1}^{K} \omega_k p_k(\xi_i)$$
Variational calculus lends itself nicely to approximate Bayesian inference

\[
\ln p(y) = \ln \frac{p(y, \theta)}{p(\theta|y)} \\
= \int q(\theta) \ln \frac{p(y, \theta)}{p(\theta|y)} \, d\theta \\
= \int q(\theta) \ln \frac{p(y, \theta)}{p(\theta|y)} \frac{q(\theta)}{q(\theta)} \, d\theta \\
= \int q(\theta) \left( \ln \frac{q(\theta)}{p(\theta|y)} + \ln \frac{p(y, \theta)}{q(\theta)} \right) \, d\theta \\
= \int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} \, d\theta + \int q(\theta) \ln \frac{p(y, \theta)}{q(\theta)} \, d\theta \\
\text{KL}[q||p] \quad \text{divergence between } q(\theta) \text{ and } p(\theta|y) \\
F(q, y) \quad \text{free energy}
\]
Variational inference

In summary, the log model evidence can be expressed as:

\[ \ln p(y) = \text{KL}[q||p] + F(q, y) \]

- divergence: \( \geq 0 \)
- free energy: (easy to evaluate for a given \( q \))

Maximizing \( F(q, y) \) is equivalent to:

- minimizing \( \text{KL}[q||p] \)
- tightening \( F(q, y) \) as a lower bound to the log model evidence
The mean-field assumption

When inverting models with several parameters, a common way of restricting the class of approximate posteriors $q(\theta)$ is to consider those posteriors that factorize into independent partitions,

$$q(\theta) = \prod_i q_i(\theta_i),$$

where $q_i(\theta_i)$ is the approximate posterior for the $i^{th}$ subset of parameters.
**Variational inference under mean-field assumption**

\[
F(q, y) = \int q(\theta) \ln \frac{p(y, \theta)}{q(\theta)} \, d\theta \\
= \int \prod_{i} q_i \times \left( \ln p(y, \theta) - \sum_{i} \ln q_i \right) \, d\theta \\
= \int q_j \prod_{i \neq j} q_i \left( \ln p(y, \theta) - \ln q_j \right) \, d\theta - \int q_j \prod_{i \neq j} q_i \sum_{i \neq j} \ln q_i \, d\theta \\
= \int q_j \left( \int \prod_{i \neq j} q_i \ln p(y, \theta) \, d\theta_{\neq j} - \ln q_j \right) \, d\theta_j - \int q_j \int \prod_{i \neq j} q_i \ln \prod_{i \neq j} q_i \, d\theta_{\neq j} \, d\theta_j \\
= \int q_j \ln \frac{\exp \left( \langle \ln p(y, \theta) \rangle_{q_{\neq j}} \right)}{q_j} \, d\theta_j + c \\
= -KL \left[ q_j \mid \mid \exp \left( \langle \ln p(y, \theta) \rangle_{q_{\neq j}} \right) \right] + c
\]
Variational inference under mean-field assumption

In summary:
\[ F(q, y) = -\text{KL} \left[ q_j || \exp \left( \langle \ln p(y, \theta) \rangle_{q_{\backslash j}} \right) \right] + c \]

Suppose the densities \( q_{\backslash j} \equiv q(\theta_{\backslash j}) \) are kept fixed. Then the approximate posterior \( q(\theta_j) \) that maximizes \( F(q, y) \) is given by:

\[
q_j^* = \arg \max_{q_j} F(q, y)
= \frac{1}{Z} \exp \left( \langle \ln p(y, \theta) \rangle_{q_{\backslash j}} \right)
\]

Therefore:
\[
\ln q_j^* = \langle \ln p(y, \theta) \rangle_{q_{\backslash j}} - \ln Z
= I(\theta_j)
\]

This implies a straightforward algorithm for variational inference:

1. Initialize all approximate posteriors \( q(\theta_i) \), e.g., by setting them to their priors.

2. Cycle over the parameters, revising each given the current estimates of the others.

3. Loop until convergence.
A standard Gaussian process is defined as

\[ p(z \mid x, \theta) = \mathcal{N}\left(0, K(x, x; s^2, r) + g\|I\right) \]

Where the covariance function is selected as Square Exponential

\[ k(x_1, x_2; s^2, r) = s^2 \exp\left\{-\frac{1}{2} \sum_{i=1}^{d} \left(\frac{x_{1,i} - x_{2,i}}{r_i^2}\right)^2\right\} \]

\[ s^2, r_i, g: \text{hyper-parameters} \]

The log-likelihood is

\[ L(\theta) = -\frac{1}{2} \log |K| - \frac{1}{2} y^T K^{-1} y - \frac{N}{2} \log 2\pi \]

Two widely-used approach to obtain the hyper-parameters, (1) point estimator by stochastic gradient descent, (2) Sampling method
Nonparametric variational inference

Remember:
\[
\ln p(y) = \text{KL}[q||p] + F(q, y)
\]

The lower bound (free energy) can be written as:
\[
F(q, y) = \mathbb{E}_q \left[ \ln \frac{p(y, \theta)}{q(\theta)} \right] = \mathcal{H}[q] + \mathbb{E}_q \left[ \ln p(y, \theta) \right] = \mathcal{H}[q] + \mathbb{E}_q \left[ f(\theta) \right]
\]

Nonparametric variational inference choose:
\[
q(\theta) = \frac{1}{N} \sum_{l=1}^{L} \mathcal{N}\left( \theta; \mu_l, \sigma_l^2 \mathbb{I} \right)
\]

Lower bound the entropy
\[
\mathcal{H}[q] = -\int_{\theta} q(\theta) \ln q(\theta) d\theta \geq -\frac{1}{N} \sum_{l=1}^{L} \ln q_l
\]

where
\[
q_l = \frac{1}{N} \sum_{j=1}^{L} \mathcal{N}\left( \mu_l; \mu_j, \left( \sigma_l^2 + \sigma_j^2 \right) \mathbb{I} \right)
\]
Nonparametric variational inference

\[ \mathbb{E}_q \left[ f(\theta) \right] = \frac{1}{N} \sum_{l=1}^{L} \int_\theta \mathcal{N}(\theta; \mu_l, \sigma_l^2 \mathbb{I}) f(\theta) d\theta \]

We approximate each term in this sum with a second order Taylor series expansion of \( f(\theta) \) around \( \mu_l \)

\[ f(\theta) \approx \hat{f}_n(\theta) = f(\mu_l) + \nabla f(\mu_l)(\theta - \mu_l) + \frac{1}{2}(\theta - \mu_l)^T H_n(\theta - \mu_l) \]

Then,

\[ \mathbb{E}_q \left[ f(\theta) \right] = \frac{1}{N} \sum_{l=1}^{L} f(\mu_l) + \frac{\sigma_l^2}{2} \text{Tr}(H_n) \]

This gives the approximate lower bound

\[ F(q, y) \approx \mathcal{L}_2[q] = \frac{1}{N} \sum_{l=1}^{L} f(\mu_l) + \frac{\sigma_l^2}{2} \text{Tr}(H_n) - \ln q_i \]
**Dirichlet Process**

**Definition:** A Dirichlet Process (DP) is a distribution over probability measures.

A DP has two parameters:

- Base distribution $H$, *which is like the mean of the DP*.
- Strength parameter $\alpha$, *which is like an inverse-variance of the DP*.

$$G \sim \text{DP}(\alpha, H)$$

If for any partition of $(A_0, \ldots, A_n)$ of $X$

$$\left( G(A_1), \ldots, G(A_n) \right) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_n))$$
Stick-Breaking representation

\[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \quad \pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j), \]

\[ v_k \sim \text{Beta}(1, \alpha) \quad \theta_k \sim G_0 \]
**Stick-Breaking representation**

\[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \quad \pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j), \]

\[ v_k \sim \text{Beta}(1, \alpha) \quad \theta_k \sim G_0 \]

stick-breaking weights \( \pi \)
Conjugate prior selection

Let introduce a latent variable $z_n$ that assigns the index of the parameter associated with the training data set

Prior selection:

Hyper-parameters for DP: 
\[
\bar{\nu} = \{\nu_1, \ldots, \nu_{\infty}\} \quad \nu_i \sim \text{Beta}(1, \alpha_0)
\]

Hyper-parameters for input:
\[
p(\xi | z = k, \mu_k, \Sigma_k) = \mathcal{N}(\xi | z = k, \mu_k, \Sigma_k^{-1})
\]
\[
\mu_k \sim \mathcal{N}(\mu_0, \Lambda_0^{-1}), \quad \Sigma_k \sim \mathcal{W}(W_0, \nu_0)
\]

Hyper-parameters for Gaussian Process model:
\[
p(s) \propto 1, \quad p(r_k) \propto 1, \quad p(g_k) \sim \Gamma(g_k; a_0, b_0)
\]
Graphical model representation

\[ p(\mathcal{D}, \Omega) = \prod_{t} p(\nu_t) p(\mu_t) p(\Sigma_t) p(g_t) \]

\[ = \prod_{n=1}^{N} p(z_n | \nu) p(\xi_n | z_n, \nu, \Sigma) p(y_n | z_n, \xi_n, \theta_{z_n}) \]
Variational inference

Truncated representation of the Dirichlet process

We fix a value $T$ and let $q\left(\nu_T = 1\right) = 1$

Now the factorized variational approximation to the posterior becomes

$$q(\Omega) = \prod_{t=1}^{T-1} p(\nu_t) \prod_{k=1}^{T} p(\mu_k) p(\Sigma_k) p(\theta_k)$$

Hyper-parameters for DP:

Prior:

$$\nu_t \sim \text{Beta}(1, \alpha_0)$$

Posterior:

$$\nu_t \sim \text{Beta}(\nu_{t1}, \nu_{t1} + \alpha_0)$$

where

$$\nu_{t1} = \sum_{n=1}^{N} q(z_n > t)$$

$$\nu_{t1} = \sum_{n=1}^{N} q(z_n = t)$$
Hyper-parameters for input:

Prior:

\[ \mu_k \sim \mathcal{N}(\mu_0, \Lambda_0^{-1}) \]

Posterior:

\[ \mu_k \sim \mathcal{N}\left((\Lambda_0 + \Sigma_{k2})^{-1}(\Lambda_0 \mu_0 + \Sigma_{k1}), (\Lambda_0 + \Sigma_{k2})^{-1}\right) \]

where

\[ \Sigma_{k1} = \sum_{n=1}^{N} q (z_n = k) \mathbb{E}[\Sigma_k] \xi_n \]

\[ \Sigma_{k2} = \sum_{n=1}^{N} q (z_n = k) \mathbb{E}[\Sigma_k] \]

\[ \Sigma_k \sim \mathcal{W}(W_0, \nu_0) \]

Posterior:

\[ \Sigma_k \sim \mathcal{W}(W^k, \nu^k) \]

where

\[ W^k = \left(W_0^{-1} + \mu_{k2}\right)^{-1} \]

and

\[ \nu^k = \nu_0 + \mu_{k1} \]

\[ \mu_{k1} = \sum_{n=1}^{N} q (z_k = k) \]

\[ \mu_{k2} = \sum_{n=1}^{N} q (z_k = k) \mathbb{E}_{\xi_k} \left[ (\xi_n - \mu_k)(\xi_n - \mu_k)^T \right] \]
Variational inference

Hyper-parameters $\theta_k$ for Gaussian process model:

$$q(\theta_k) = \frac{1}{N} \sum_{l=1}^{L} N(\theta_k; \mu_l, \sigma_l^2 \mathbb{I})$$

Approximated using non-parametric variational inference

Update of $q(z_n)$ is slightly complex

$$\ln q(z_n) = \mathbb{E}_{\Omega \setminus z_n} \left[ \ln p(z_n | \nu) + \ln p(\xi_n | z_n, \mu, \Sigma) + \ln p(y_n | z_n, \xi_n, \theta) \right]$$

$$q(z_n = k) = \tilde{\rho}_{nk} = \frac{\rho_{nk}}{\sum_{i=1}^{T} \rho_{ni}}$$

where

$$\ln \rho_{nk} = \mathbb{E}[\ln \nu_k] + \sum_{i=1}^{k-1} \mathbb{E}[\ln(1 - \nu_i)] + \frac{1}{2} \left[ \mathbb{E} \ln |\Sigma_k| - d_{\xi} \ln 2\pi - \mathbb{E} \left( (\xi_n - \nu_k)^T \Sigma_k (\xi_n - \nu_k) \right) \right]$$

$$-\frac{1}{2} \sum_{d=1}^{d_y} \left[ \mathbb{E} \ln A_n^* | + \ln 2\pi + \mathbb{E} \left( (y - m_n^*)^T (A_n^*)^{-1} (y - m_n^*) \right) \right]$$

$$m_n^* = a_{\xi_n}^T (A_{M_i} + g_i I_{M_i})^{-1} y_i$$ and

$$A_n^* = A_{\xi_n \xi_n} - a_{\xi_n}^T (A_{M_i} + g_i I_{M_i})^{-1} a_{\xi_n} + g_i$$
**Summary algorithm**

Initialization: select the prior 

\[ q(\Omega) = \prod_{t=1}^{T-1} p(v_t) \prod_{k=1}^{T} p(\mu_k) p(\Sigma_k) p(\theta_k) \]

randomly set the hidden variable \( z \)

Do {

update \( q(v_t) \)
update \( q(\mu_k) \)
update \( q(\Sigma_k) \)
Do {

update \( q(\theta_k) \)

} while the approximate local lower bound \( \mathcal{L}_2[q] \) converges

update \( q(z_n) \)

} while the approximate lower bound \( F(q,y) \) converges
Thank you!