

A Brief Tutorial of Mixtures of Gaussian Processes

Jing Zhao
jzhao@cs.ecnu.edu.cn

Outline

- Mixtures of Gaussian Processes
 - Finite mixture
 - Infinite mixture
 - Generative mixture
 - Multivariate mixture
- Inference
 - Variational Inference
 - Monto Carlo

Finite mixture

- Volker Tresp (NIPS, 2001)
 - “Mixtures of Gaussian Processes”
- Motivation
 - General conditional probability densities
 - Fast to train
 - Different length scales
- Application
 - Fitting data and making prediction

Mixtures of Gaussian Processes

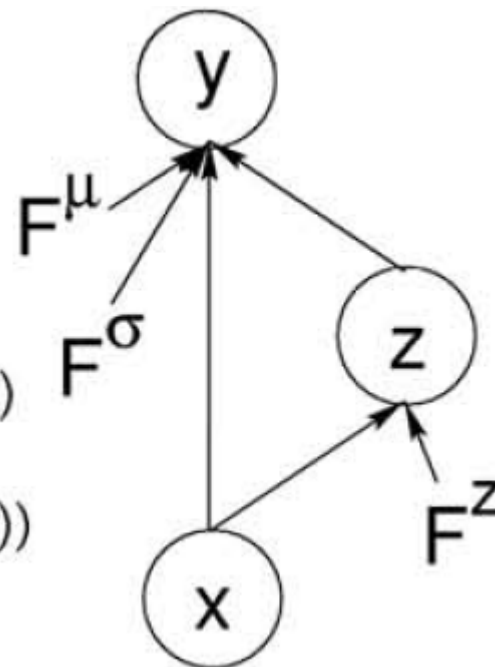
- Indicator

- $$P(z = i | F^z(x)) = \frac{\exp(f_i^z(x))}{\sum_{j=1}^M \exp(f_j^z(x))}$$

- Likelihood

- $$P(y|z, F^\mu(x), F^\sigma(x)) = G(y; f_z^\mu(x), \exp(2f_z^\sigma(x)))$$

- $$P(y|x) = \sum_{i=1}^M P(z = i|x) G(y; f_i^\mu(x), \exp(2f_i^\sigma(x)))$$

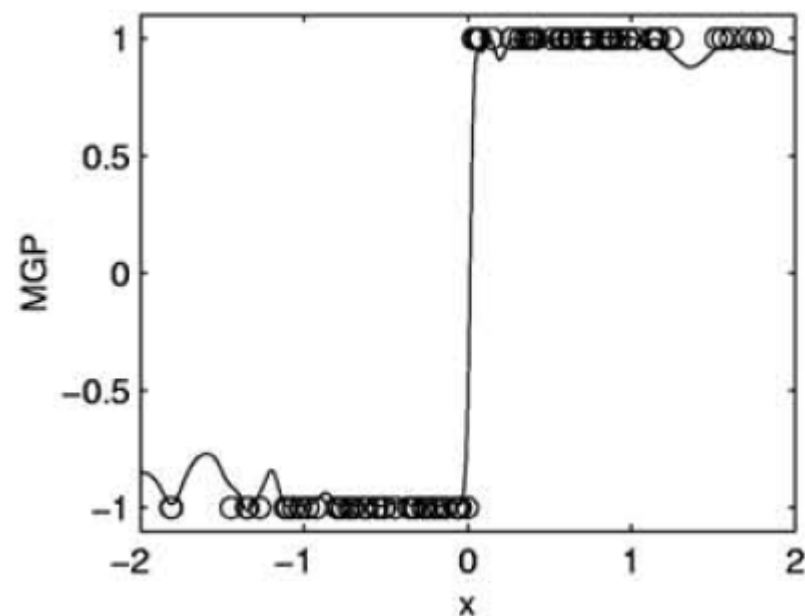
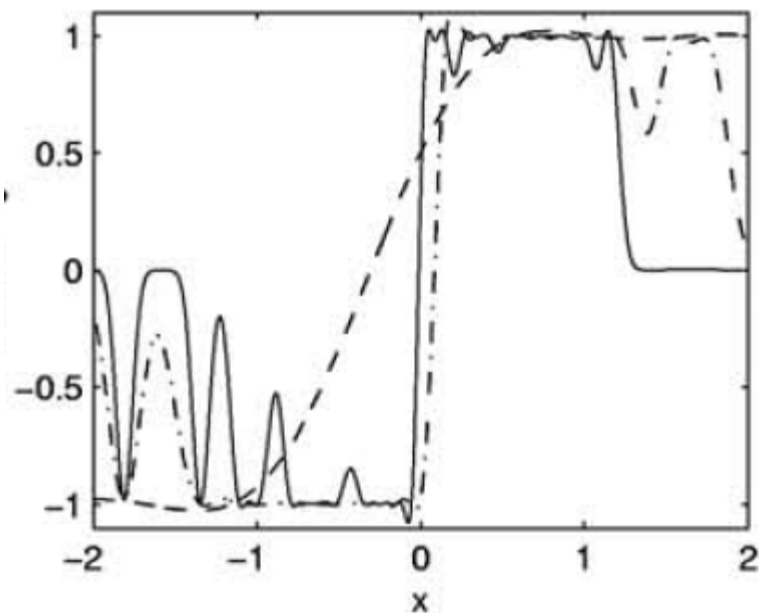


- EM for MAP

- $$\hat{P}(z = i | x_k, y_k) = \frac{\hat{P}(z = i | x_k) G(y_k; \hat{f}_i^\mu(x_k), \exp(2\hat{f}_i^\sigma(x_k)))}{\sum_{j=1}^M \hat{P}(z = j | x_k) G(y_k; \hat{f}_j^\mu(x_k), \exp(2\hat{f}_j^\sigma(x_k)))}$$

Mixtures of Gaussian Processes

- Better fitting data



Infinite Mixture

- Carl E. Rasmussen and Z. Ghahramani (NIPS,2002)
 - “Infinite Mixtures of Gaussian Process Experts”
- Motivation
 - General conditional probability densities
 - Fast to train
 - Different length scales
 - Infinite number of experts
- Fitting data

- Gating network

- $$p(\pi_1, \dots, \pi_k | \alpha) \sim \text{Dirichlet}(\alpha/k) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_j \pi_j^{\alpha/k-1}$$

components where $n_{-i,j} > 0$:
$$p(c_i = j | \mathbf{c}_{-i}, \alpha) = \frac{n_{-i,j}}{n - 1 + \alpha},$$

all other components combined:
$$p(c_i \neq c_{i'} \text{ for all } i' \neq i | \mathbf{c}_{-i}, \alpha) = \frac{\alpha}{n - 1 + \alpha},$$

$$n_{-i,j} = (n - 1) \frac{\sum_{i' \neq i} K_\phi(x_i, x_{i'}) \delta(c_{i'}, j)}{\sum_{i' \neq i} K_\phi(x_i, x_{i'})}$$

$$K_\phi(x_i, x_{i'}) = \exp \left(-\frac{1}{2} \sum_d (x_{id} - x_{i'd})^2 / \phi_d^2 \right)$$

- Likelihood

- $$\begin{aligned} p(\mathbf{y}|\mathbf{x}, \theta) &= \sum_{\mathbf{c}} p(\mathbf{y}|\mathbf{c}, \mathbf{x}, \theta) p(\mathbf{c}|\mathbf{x}, \phi) \\ &= \sum_{\mathbf{c}} \left[\prod_j p(\{y_i : c_i = j\} | \{x_i : c_i = j\}, \theta_j) \right] p(\mathbf{c}|\mathbf{x}, \phi) \end{aligned}$$

- Gibbs sampling

- $p(c_i = j | \mathbf{c}_{-i}, \mathbf{x}, \mathbf{y}, \theta, \phi) \propto p(\mathbf{y} | c_i = j, \mathbf{c}_{-i}, \mathbf{x}, \theta) p(c_i = j | \mathbf{c}_{-i}, \mathbf{x}, \phi).$
- hyper-parameters of GP kernel
- DP parameter α
- hyper-parameters of gating kernel

iMGPE

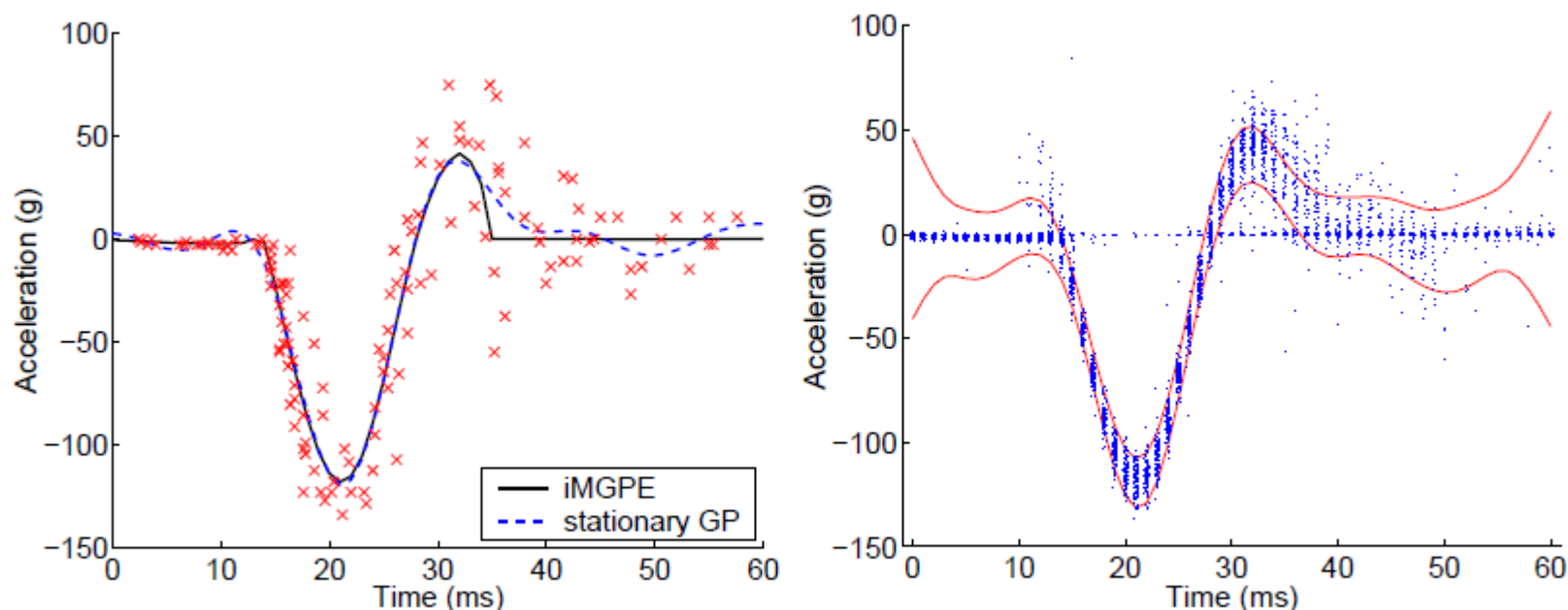


Figure 1: The left hand plot shows the motorcycle impact data (133 points) together with the median of the model's predictive distribution, and for comparison the mean of a stationary covariance GP model (with optimized hyperparameters). On the right hand plot we show 100 samples from the posterior distribution for the iMGPE of the (noise free) function evaluated intervals of 1 ms. We have jittered the points in the plot along the time dimension by adding uniform ± 0.2 ms noise, so that the density can be seen more easily. Also, the ± 2 std error (95%) confidence interval for the (noise free) function predicted by a stationary GP is plotted (thin lines).

Generative Mixture

- Literatures

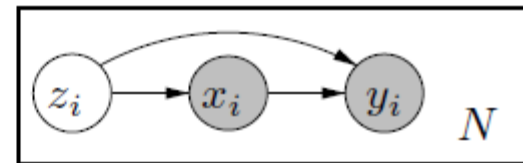
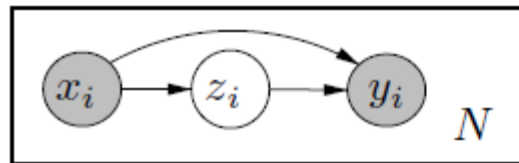
- “An Alternative Infinite Mixture of Gaussian Process Experts”, *E. Meeds and S. Osindero (NIPS, 2006)*
- “Variational Mixture of Gaussian Process Experts”, *C. Yuan and C. Neubauer (NIPS, 2009)*
- “Variational Inference for Infinite Mixtures of Gaussian Processes With Applications to Traffic Flow Prediction”, *S. Sun and X. Xu (TITS, 2013)*

- Motivation

- Deal with partially specified data
- Infer inverse functional mappings
- Consistent prior of outputs and inputs

AiMoGPE

- Likelihood



$$p(y|\mathbf{x}, z = k, \mathbf{w}_k, r_k) = \mathcal{N}(y|\mathbf{w}_k^\top \phi_k(\mathbf{x}), r_k^{-1})$$

$$\phi_k(\mathbf{x}) = \left[\kappa_k(\mathbf{x}, \mathbf{x}_1^{\mathbf{I}_k}), \kappa_k(\mathbf{x}, \mathbf{x}_2^{\mathbf{I}_k}), \dots, \kappa_k(\mathbf{x}, \mathbf{x}_{|\mathbf{I}_k|}^{\mathbf{I}_k}) \right]^\top$$

- Gate network

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

- $\pi_i = \nu_i \prod_{j=1}^{i-1} (1 - \nu_j)$

- $z_n \sim \text{multinomial} \{ \pi_1, \dots, \pi_\infty \}$

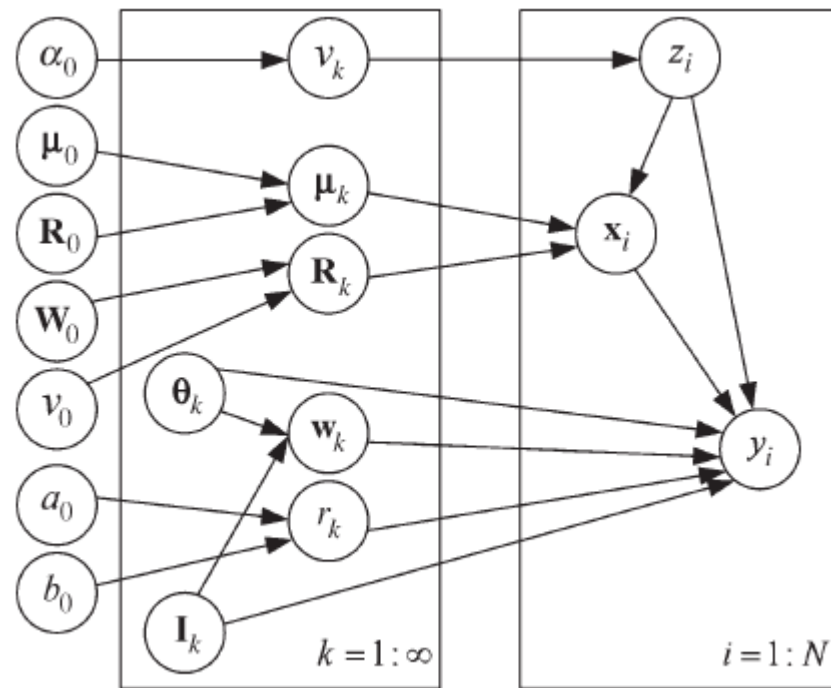
- Prior

$$\mathcal{N}(\mathbf{w}_k | \mathbf{0}, \mathbf{U}_k^{-1}) \quad \mathbf{U}_k \text{ is set to } \mathbf{K}_k + \sigma_{kb}^2 \mathbf{I}$$

$$\Gamma(r_k | a_0, b_0) \propto r_k^{a_0-1} e^{-b_0 r_k}$$

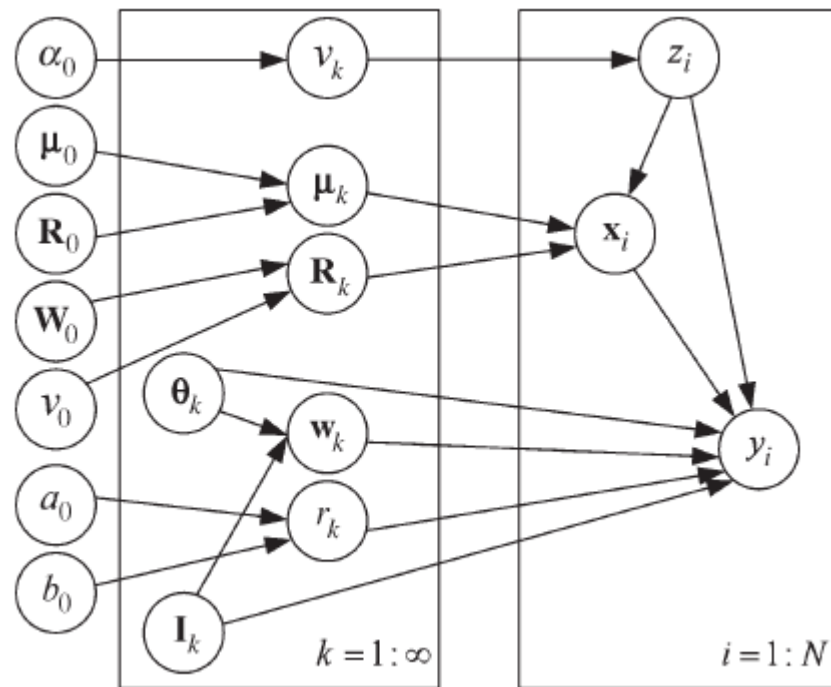
$$p(\mathbf{x} | z = k, \mu_k, \mathbf{R}_k) = \mathcal{N}(\mathbf{x} | \mu_k, \mathbf{R}_k^{-1})$$

$$\mu_k \sim \mathcal{N}(\mu_0, \mathbf{R}_0^{-1}), \quad \mathbf{R}_k \sim \mathcal{W}(\mathbf{W}_0, \nu_0)$$



AiMoGPE

- Joint distribution



$$\begin{aligned}
 p(\mathbf{D}, \Omega) &= p(\bar{\nu})p(\bar{\mu})p(\bar{\mathbf{R}})p(\bar{\mathbf{w}})p(\bar{r}) \\
 &\quad \times \prod_{i=1}^N p(z_i | \bar{\nu})p(\mathbf{x}_i | z_i, \bar{\mu}, \bar{\mathbf{R}})p(y_i | \mathbf{x}_i, z_i, \bar{\mathbf{w}}, \bar{r}) \\
 &= \prod_{k=1}^{\infty} p(\nu_k)p(\mu_k)p(\mathbf{R}_k)p(\mathbf{w}_k)p(r_k) \\
 &\quad \times \prod_{i=1}^N p(z_i | \bar{\nu})p(\mathbf{x}_i | z_i, \bar{\mu}, \bar{\mathbf{R}})p(y_i | \mathbf{x}_i, z_i, \bar{\mathbf{w}}, \bar{r})
 \end{aligned}$$

VI for AiMoGPE

$$\ln p(\mathbf{D}) = \mathcal{L}(q) + \text{KL}(q\|p) \quad (10)$$

where $\mathcal{L}(q) = \int q(\Omega) \ln\{p(\mathbf{D}, \Omega)/q(\Omega)\} d\Omega$, and $\text{KL}(q\|p) = \int q(\Omega) \ln\{q(\Omega)/p(\Omega|\mathbf{D})\} d\Omega$. The Kullback–Leibler divergence $\text{KL}(q\|p)$ is nonnegative and is zero if and only if $q(\Omega) = p(\Omega|\mathbf{D})$ [19]. $\mathcal{L}(q)$ is the lower bound of $\ln p(\mathbf{D})$.

$$q(\Omega) = \prod_{t=1}^{T-1} q(\nu_t) \prod_{k=1}^T q(\mu_k) q(\mathbf{R}_k) q(\mathbf{w}_k) q(r_k) \prod_{n=1}^N q(z_n)$$

$$\ln q(\omega) = \mathbb{E}_{\Omega \setminus \omega} [\ln p(\mathbf{D}, \Omega)] + \text{const}$$

VI for AiMoGPE

$$p(\mathbf{D}, \Omega) = \prod_{k=1}^{\infty} p(\nu_k) p(\mu_k) p(\mathbf{R}_k) p(\mathbf{w}_k) p(r_k) \\ \times \prod_{i=1}^N p(z_i | \bar{\nu}) p(\mathbf{x}_i | z_i, \bar{\mu}, \bar{\mathbf{R}}) p(y_i | \mathbf{x}_i, z_i, \bar{\mathbf{w}}, \bar{r})$$

$$\ln q(\nu_t) = \ln p(\nu_t) + \sum_{n=1}^N \mathbb{E}_{\Omega \setminus \nu_t} [\ln p(z_n | \bar{\nu})] + \text{const.}$$

$$\ln q(\mu_k) = \ln p(\mu_k) + \sum_{n=1} \mathbb{E}_{\Omega \setminus \mu_k} [\ln p(\mathbf{x}_n | z_n, \bar{\mu}, \bar{\mathbf{R}})] + \text{const}$$

$$\ln q(\mathbf{R}_k) = \ln p(\mathbf{R}_k) + \sum_{n=1} \mathbb{E}_{\Omega \setminus \mathbf{R}_k} [\ln p(\mathbf{x}_n | z_n, \bar{\mu}, \bar{\mathbf{R}})] + \text{const.}$$

$$\ln q(\mathbf{w}_k) = \ln p(\mathbf{w}_k) \\ + \sum_{n=1}^N \mathbb{E}_{\Omega \setminus \mathbf{w}_k} [\ln p(y_n | \mathbf{x}_n, z_n, \bar{\mathbf{w}}, \bar{r})] + \text{const.}$$

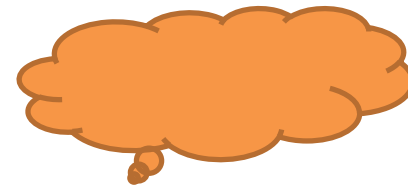
$$\ln q(r_k) = \ln p(r_k) + \sum_{n=1} \mathbb{E}_{\Omega \setminus r_k} [\ln p(y_n | \mathbf{x}_n, z_n, \bar{\mathbf{w}}, \bar{r})] + \text{const.}$$

$$\ln q(z_n) + \text{const} = \mathbb{E}_{\Omega \setminus z_n} [\ln p(z_n | \bar{\nu}) + \ln p(\mathbf{x}_n | z_n, \bar{\mu}, \bar{\mathbf{R}}) \\ + \ln p(y_n | \mathbf{x}_n, z_n, \bar{\mathbf{w}}, \bar{r})].$$

VI for AiMoGPE

- Optimizing the Hyper-parameters

- maximizing $\mathbb{E}_{\Omega} \ln p(\mathbf{D}, \Omega | \theta_{1:T})$



- $$\mathbb{E} \left\{ \sum_{k=1}^T \ln p(\mathbf{w}_k) + \sum_{n=1}^N \ln(y_n | \mathbf{x}_n, z_n, \bar{\mathbf{w}}, \bar{r}) \right\}$$

- Optimizing the support set

- maximize the probability density of $q(\mathbf{w}_k)$
-

Prediction

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{D}, \Theta) &= \int p(y|\mathbf{x}, \Omega, \Theta) p(\Omega|\mathbf{D}, \Theta) d\Omega \\ &\simeq \int p(y|\mathbf{x}, \Omega, \Theta) q(\Omega) d\Omega \\ &\simeq p(y|\mathbf{x}, \hat{\Omega}, \Theta) \end{aligned}$$

$$\begin{aligned} p(y|\mathbf{x}, \hat{\Omega}, \Theta) &= \sum_{k=1}^T p(z = k, y|\mathbf{x}, \hat{\Omega}, \Theta) \\ &= \sum_{k=1}^T p(z = k|\mathbf{x}, \hat{\Omega}) p(y|\mathbf{x}, z = k, \hat{\Omega}, \Theta) \\ &= \sum_{k=1}^T \frac{p(z = k|\hat{\Omega}) p(\mathbf{x}|z = k, \hat{\Omega})}{\sum_{i=1}^T p(z = i|\hat{\Omega}) p(\mathbf{x}|z = i, \hat{\Omega})} \\ &\quad \times p(y|\mathbf{x}, z = k, \hat{\Omega}, \Theta). \end{aligned}$$

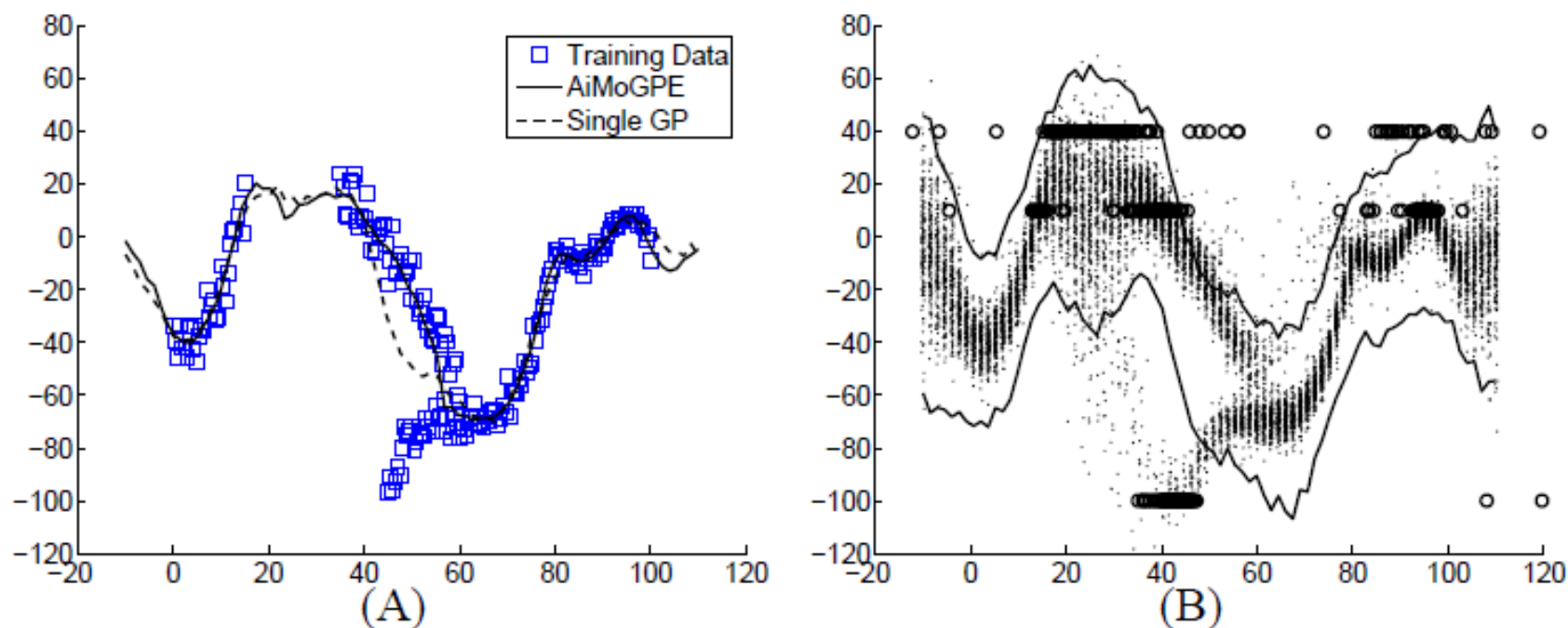


Figure 4: Results on a toy dataset. (A) The training data is shown along with the predictive mean of a stationary covariance GP and the median of the predictive distribution of our model. (B) The small dots are samples from the model (160 samples per location) evaluated at 80 equally spaced locations across the range (but plotted with a small amount of jitter to aid visualisation). These illustrate the predictive density from our model. The solid lines show the ± 2 SD interval from a regular GP. The circular markers at ordinates of 40, 10 and -100 show samples from 'reverse-conditioning' where we sample likely abscissa locations given the test ordinate and the set of training data.

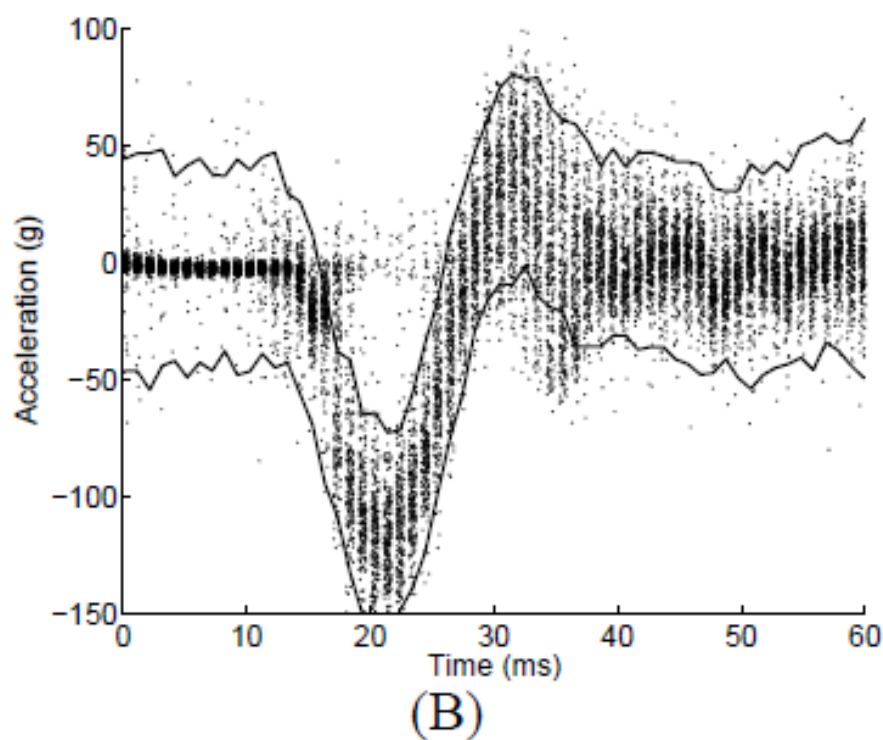
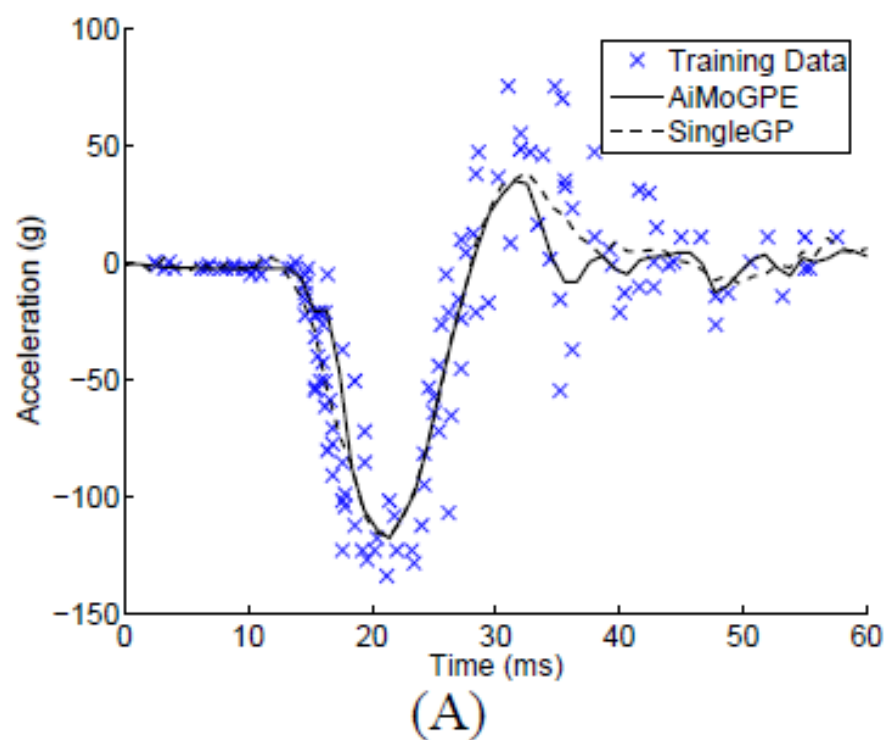
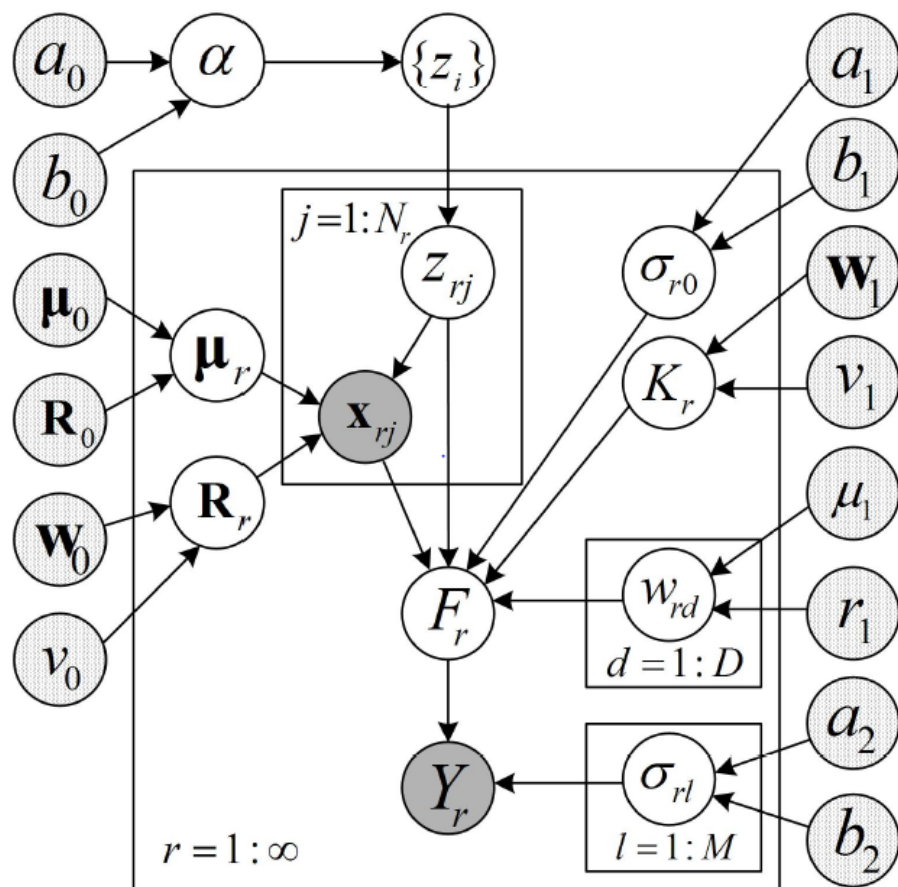


Figure 5: (A) Motorcycle impact data together with the median of our model's point-wise predictive distribution and the predictive mean of a stationary covariance GP model. (B) The small dots are samples from our model (160 samples per location) evaluated at 80 equally spaced locations across the range (but plotted with a small amount of jitter to aid visualisation). The solid lines show the ± 2 SD interval from a regular GP.

Multivariate mixture

- “INFINITE MIXTURES OF MULTIVARIATE GAUSSIAN PROCESSES”, S. Sun (ICMLC, 2013)
- Motivation
 - Multi-task
 - Multi-output
 - Multi-modal

MiMGPE



$$\begin{aligned}
 & p(\{\mathbf{x}_i, \mathbf{y}_i\} | \Theta) \\
 &= \sum_Z p(Z | \Theta) \prod_r p(\{\mathbf{y}_i : z_i = r\} | \{\mathbf{x}_i : z_i = r\}, \Theta) \times \\
 & \quad p(\{\mathbf{x}_i : z_i = r\} | \Theta) \\
 &= \sum_Z p(Z | \Theta) \prod_r p(\{\mathbf{y}_i : z_i = r\} | \{\mathbf{x}_i : z_i = r\}, \Theta) \times \\
 & \quad \prod_{j=1}^{N_r} p(\mathbf{x}_{rj} | \mu_r, \mathbf{R}_r).
 \end{aligned}$$

- Likelihood $\mathbf{y}_r = (y_{r1}^1, \dots, y_{r1}^{N_r}, y_{r2}^1, \dots, y_{r2}^{N_r}, \dots, y_{rM}^1, \dots, y_{rM}^{N_r})^\top,$

- $\mathbb{E}(f_{r\ell}(\mathbf{x})f_{rk}(\mathbf{x}')) = \sigma_{r0}K_r(\ell, k)k_r(\mathbf{x}, \mathbf{x}'),$
 $y_{r\ell}(\mathbf{x}) \sim \mathcal{N}(f_{r\ell}(\mathbf{x}), \sigma_{r\ell}),$

- Prior $\mathbf{y}_r \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \sigma_{r0}K_r \otimes K_r^x + D_r \otimes I$

- $p(\mathbf{x}|z=r, \mu_r, \mathbf{R}_r) = \mathcal{N}(\mathbf{x}|\mu_r, \mathbf{R}_r^{-1}),$

$$\mu_r \sim \mathcal{N}(\mu_0, \mathbf{R}_0^{-1}), \quad \mathbf{R}_r \sim \mathcal{W}(\mathbf{W}_0, \nu_0).$$

positive semi-definite matrix K_r is given by a Wishart distribution $\mathcal{W}(\mathbf{W}_1, \nu_1)$. σ_{r0} and $\sigma_{r\ell}$ are given gamma priors $\mathcal{G}(\sigma_{r0}|a_1, b_1)$ and $\mathcal{G}(\sigma_{r\ell}|a_2, b_2)$, respectively. We set

$$k_r(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2} \sum_{d=1}^D w_{rd}^2 (\mathbf{x}_d - \mathbf{x}'_d)^2\right), \quad (5)$$

Sampling methods

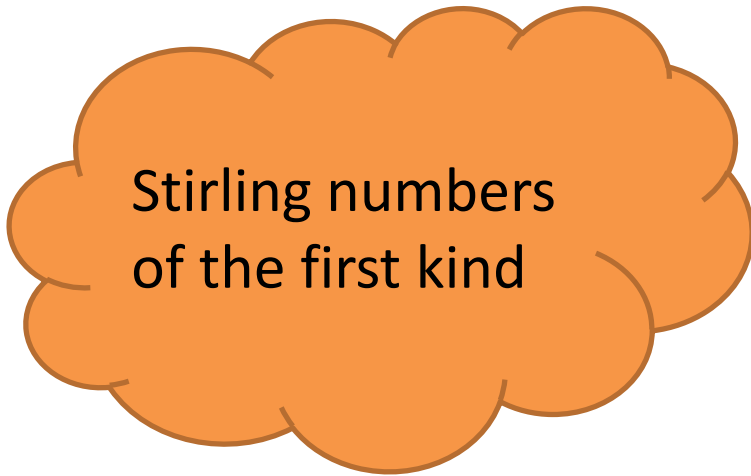
- (1) Update indicator variables $\{z_i\}_{i=1}^N$ one by one, by cycling through the training data.
- (2) Update input and output space Gaussian process parameters $\{\{\boldsymbol{\mu}_r\}, \{\mathbf{R}_r\}, \{\sigma_{r0}\}, \{K_r\}, \{w_{rd}\}, \{\sigma_{r\ell}\}\}$ for each Gaussian process component in turn.
- (3) Update Dirichlet process concentration parameter α .

Key steps

- $$\begin{aligned} & p(z_i | Z_{-i}, \Theta, \mathcal{D}) \\ & \propto p(z_i | Z_{-i}, \Theta) p(\mathcal{D} | z_i, Z_{-i}, \Theta) \\ & \propto p(z_i | Z_{-i}, \Theta) p(\mathbf{y}_i | \{\mathbf{y}_j : j \neq i, z_j = z_i\}, \{\mathbf{x}_j : z_j = z_i\}, \Theta) \\ & \quad \times p(\mathbf{x}_i | \boldsymbol{\mu}_{z_i}, \mathbf{R}_{z_i}), \end{aligned} \tag{6}$$

- posterior

- $$p(c | \alpha, N) = \beta_c^N \frac{\alpha^c \Gamma(\alpha)}{\Gamma(N + \alpha)}$$

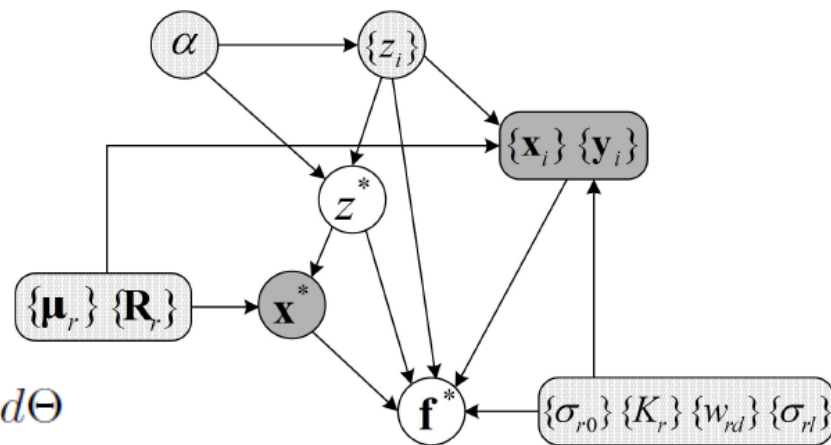


Stirling numbers
of the first kind

$$p(\alpha | c, N) \propto p(\alpha) p(c | \alpha, N) \propto p(\alpha) \frac{\alpha^c \Gamma(\alpha)}{\Gamma(N + \alpha)}$$

Prediction

- Not so easy



$$\begin{aligned}
 p(\mathbf{f}^* | \mathbf{x}^*, \mathcal{D}) &= \int \sum_Z \sum_{z^*} p(\mathbf{f}^*, z^*, Z, \Theta | \mathbf{x}^*, \mathcal{D}) d\Theta \\
 &= \int \sum_Z \sum_{z^*} p(z^*, Z, \Theta | \mathbf{x}^*, \mathcal{D}) p(\mathbf{f}^* | z^*, Z, \Theta, \mathbf{x}^*, \mathcal{D}) d\Theta \\
 &= \int \sum_{Z, z^*} p(z^* | Z, \Theta, \mathbf{x}^*) p(Z, \Theta | \mathbf{x}^*, \mathcal{D}) p(\mathbf{f}^* | z^*, Z, \Theta, \mathbf{x}^*, \mathcal{D}) d\Theta \\
 &\approx \int \sum_{Z, z^*} p(z^* | \mathbf{x}^*, Z, \Theta) p(Z, \Theta | \mathcal{D}) p(\mathbf{f}^* | z^*, Z, \Theta, \mathbf{x}^*, \mathcal{D}) d\Theta \\
 &= \int \sum_Z \left[\sum_{z^*} p(z^* | \mathbf{x}^*, Z, \Theta) p(\mathbf{f}^* | z^*, Z, \Theta, \mathbf{x}^*, \mathcal{D}) \right] \times \\
 &\quad p(Z, \Theta | \mathcal{D}) d\Theta,
 \end{aligned}$$

(18)

$$\begin{aligned}
&= \int \sum_Z \left[\sum_{z^*} p(z^* | \mathbf{x}^*, Z, \Theta) p(\mathbf{f}^* | z^*, Z, \Theta, \mathbf{x}^*, \mathcal{D}) \right] \times \\
&\quad p(Z, \Theta | \mathcal{D}) d\Theta, \\
&\quad p(\mathbf{f}^* | \mathbf{x}^*, \mathcal{D}) \\
&= \frac{p(z^* | \mathbf{x}^*, Z_i, \Theta_i)}{p(\mathbf{x}^* | Z_i, \Theta_i)} \\
&= \frac{p(z^* | Z_i, \Theta_i) p(\mathbf{x}^* | z^*, Z_i, \Theta_i)}{\sum_{z^*} p(z^* | Z_i, \Theta_i) p(\mathbf{x}^* | z^*, Z_i, \Theta_i)} \\
&= \frac{p(z^* | Z_i, \Theta_i) p(\mathbf{x}^* | z^*, \Theta_i)}{\sum_{z^*} p(z^* | Z_i, \Theta_i) p(\mathbf{x}^* | z^*, \Theta_i)}, \\
&\approx \frac{1}{L} \sum_{i=1}^L \left[\sum_{z^*} p(z^* | \mathbf{x}^*, Z_i, \Theta_i) p(\mathbf{f}^* | z^*, Z_i, \Theta_i, \mathbf{x}^*, \mathcal{D}) \right].
\end{aligned}$$

Therefore, the prediction for \mathbf{f}^* is

$$\hat{\mathbf{f}}^* = \frac{1}{L} \sum_{i=1}^L \left[\sum_{z^*} p(z^* | \mathbf{x}^*, Z_i, \Theta_i) \mathbb{E}(\mathbf{f}^* | z^*, Z_i, \Theta_i, \mathbf{x}^*, \mathcal{D}) \right],$$