An Alternative Infinite Mixture Of Gaussian Process Experts

Edward Meeds and Simon Osindero

Presented by Paul Kernfeld

POTENTIAL STUMBLING BLOCK

- Rasmussen and Ghahramani 2002
- Sample conditional vs. joint

Problem Setup

- Pairs of (x, y) data points
- Using GP's for regression

Advantages

Overall

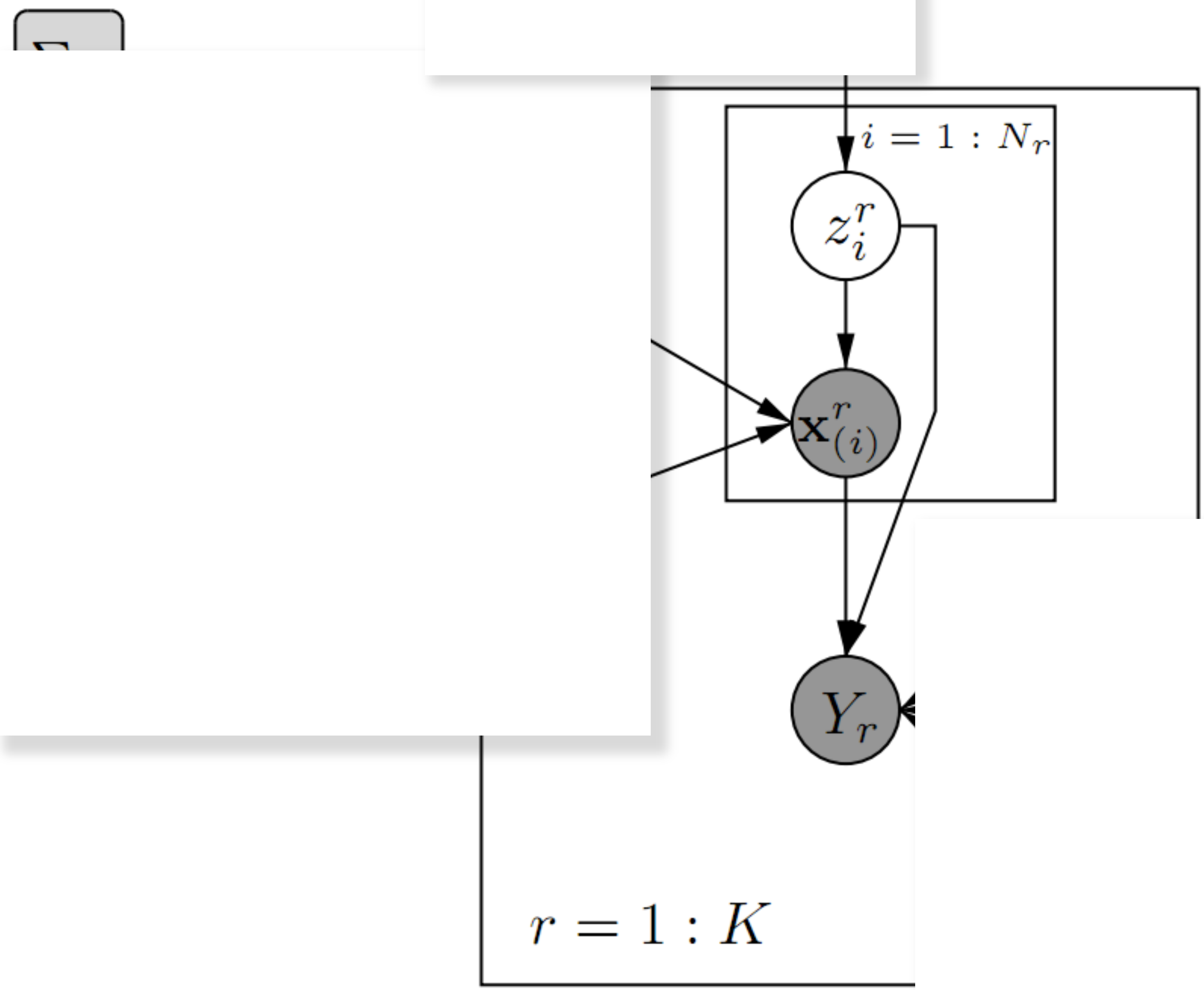
- Computation (n x n inversion)
- Multi-modal data
- Non-stationary covariance

wrt Rasmussen 2002

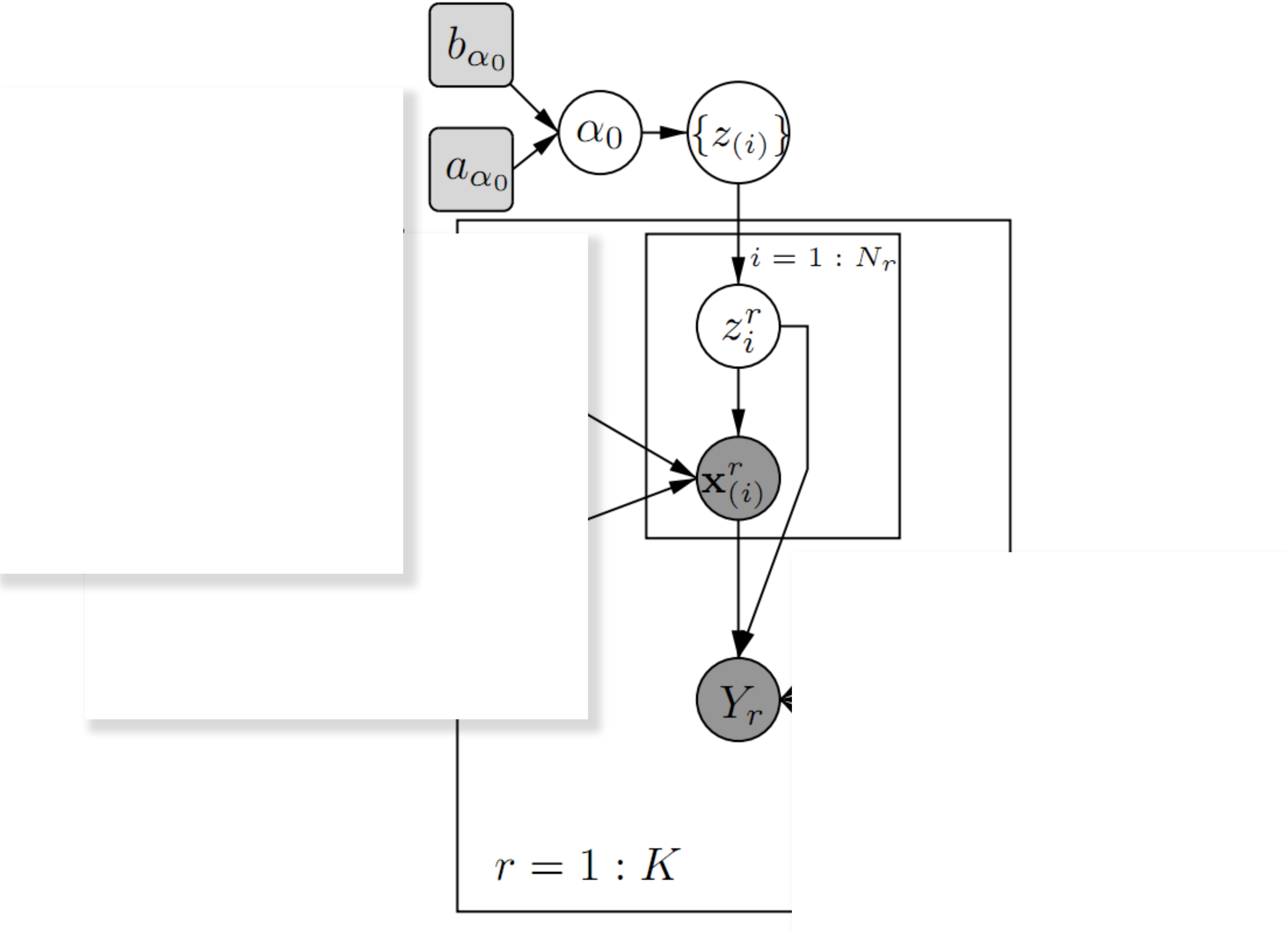
- Full generative model
- Missing inputs OK
- Slightly more Bayesian

POTENTIAL STUMBLING BLOCK

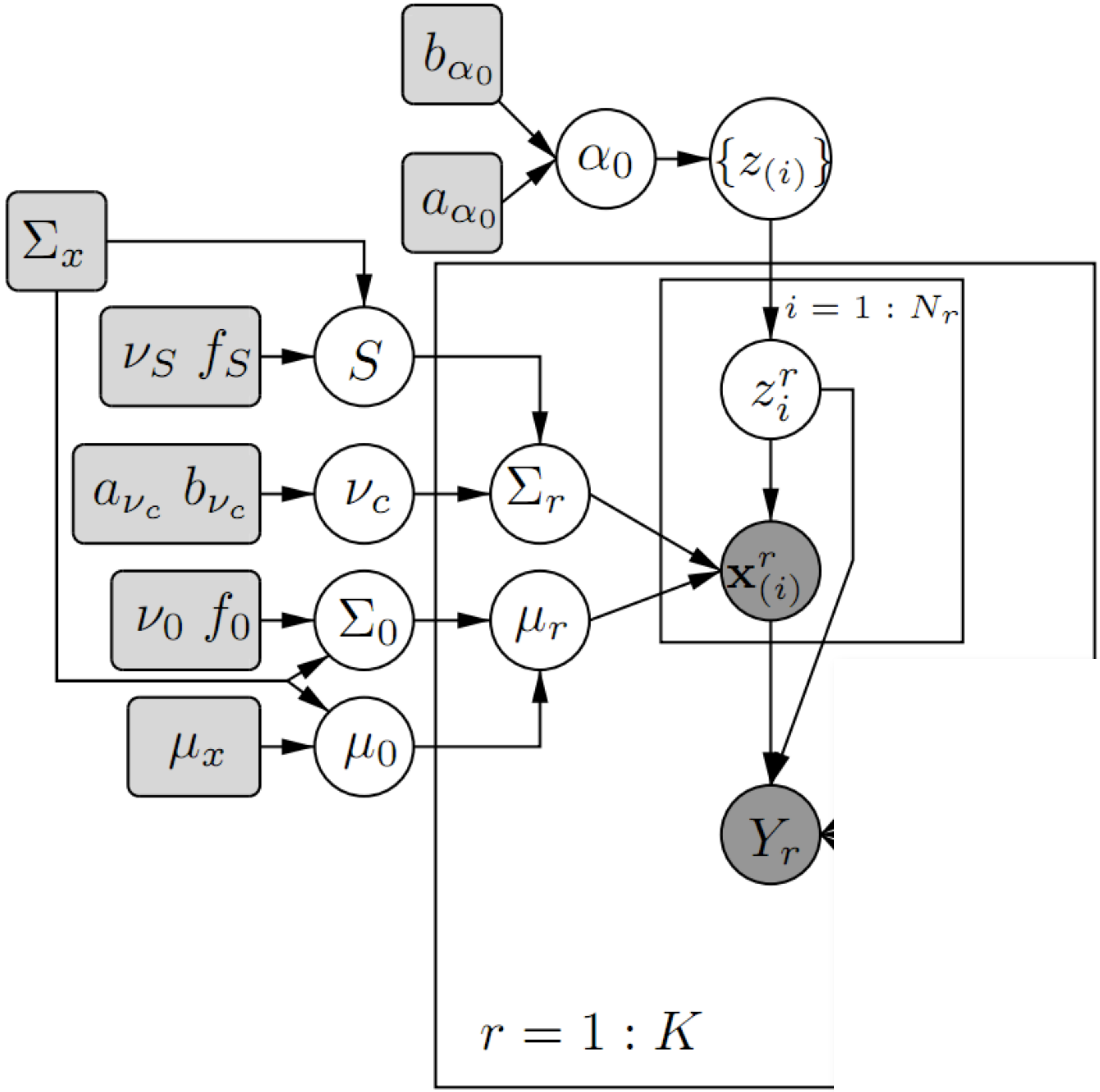
- "Experts" are individual GP's
- "Gating network" gives latent assignments



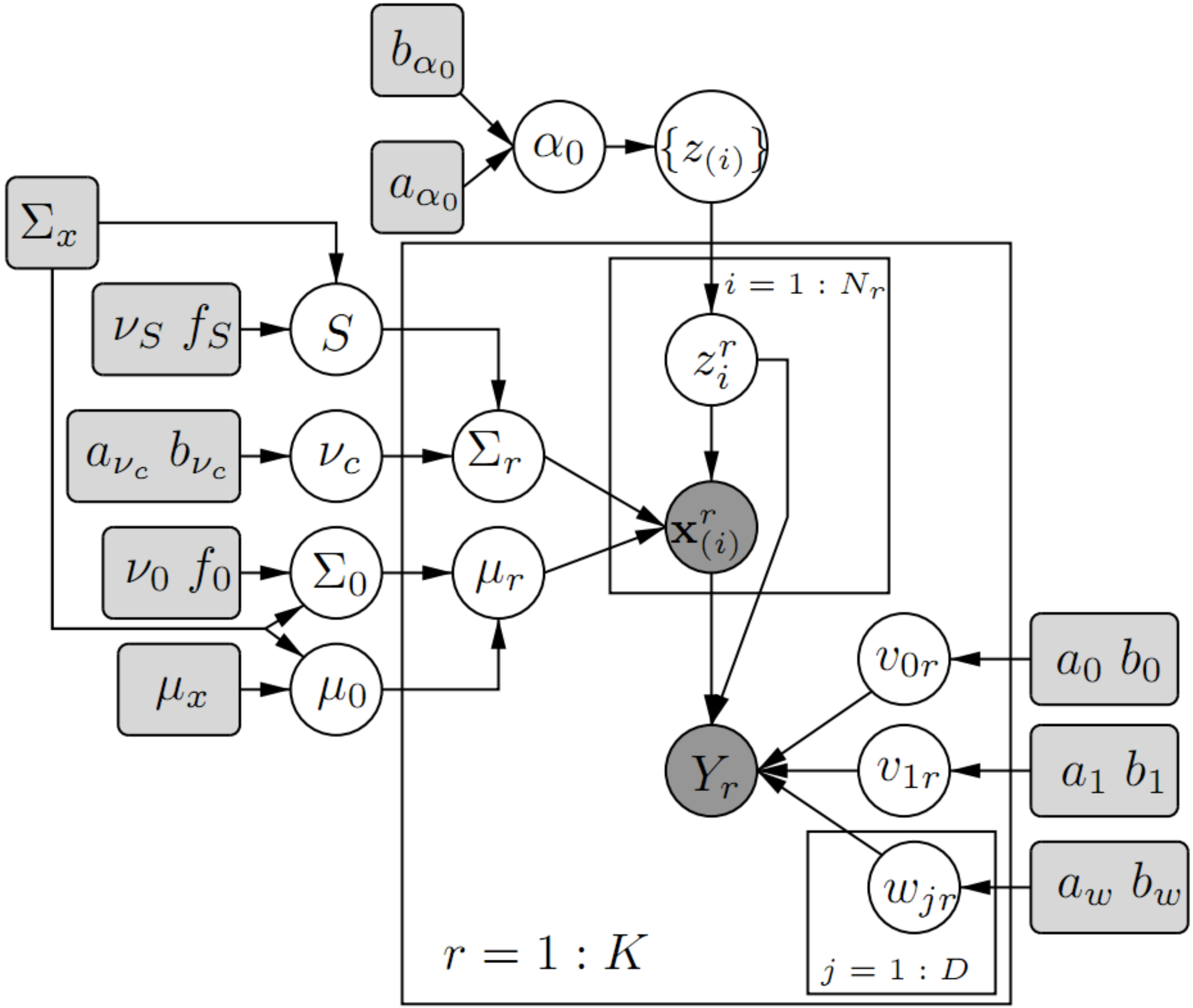
$i = 1 : N_r$
z_i^r
$\mathbf{x}_{(i)}^r$
Y_r
$r = 1 : K$



b_{α_0}
a_{α_0}
α_0
$\{z_{(i)}\}$
$i = 1 : N_r$
z_i^r
$\mathbf{x}_{(i)}^r$
Y_r
$r = 1 : K$

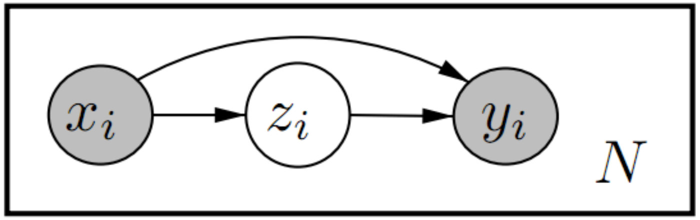


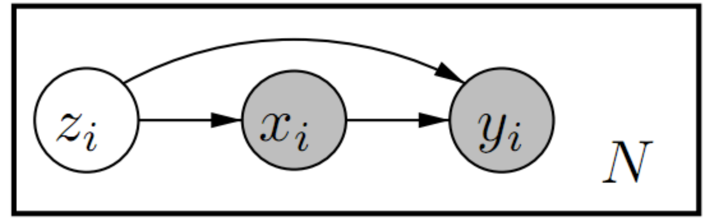
b_{α_0}
a_{α_0}
α_0
$\{z_{(i)}\}$
Σ_x
$\nu_S\ f_S$
S
$a_{\nu_c}\ b_{\nu_c}$
ν_c
$\nu_0\ f_0$
Σ_0
μ_x
μ_0
Σ_r
μ_r
$i = 1 : N_r$
z_i^r
$\mathbf{x}_{(i)}^r$
Y_r
$r = 1 : K$



b_{α_0}
a_{α_0}
α_0
$\{z_{(i)}\}$
Σ_x
S
$\nu_S \; f_S$
$a_{\nu_c} \; b_{\nu_c}$
ν_c
Σ_r
$\nu_0 \; f_0$
Σ_0
μ_r
μ_x
μ_0
$i = 1 : N_r$
z_i^r
$\mathbf{x}_{(i)}^r$
Y_r
v_{0r}
$a_0 \; b_0$
v_{1r}
$a_1 \; b_1$
w_{jr}
$a_w \; b_w$
$r = 1 : K$
$j = 1 : D$

Conditional vs. Joint



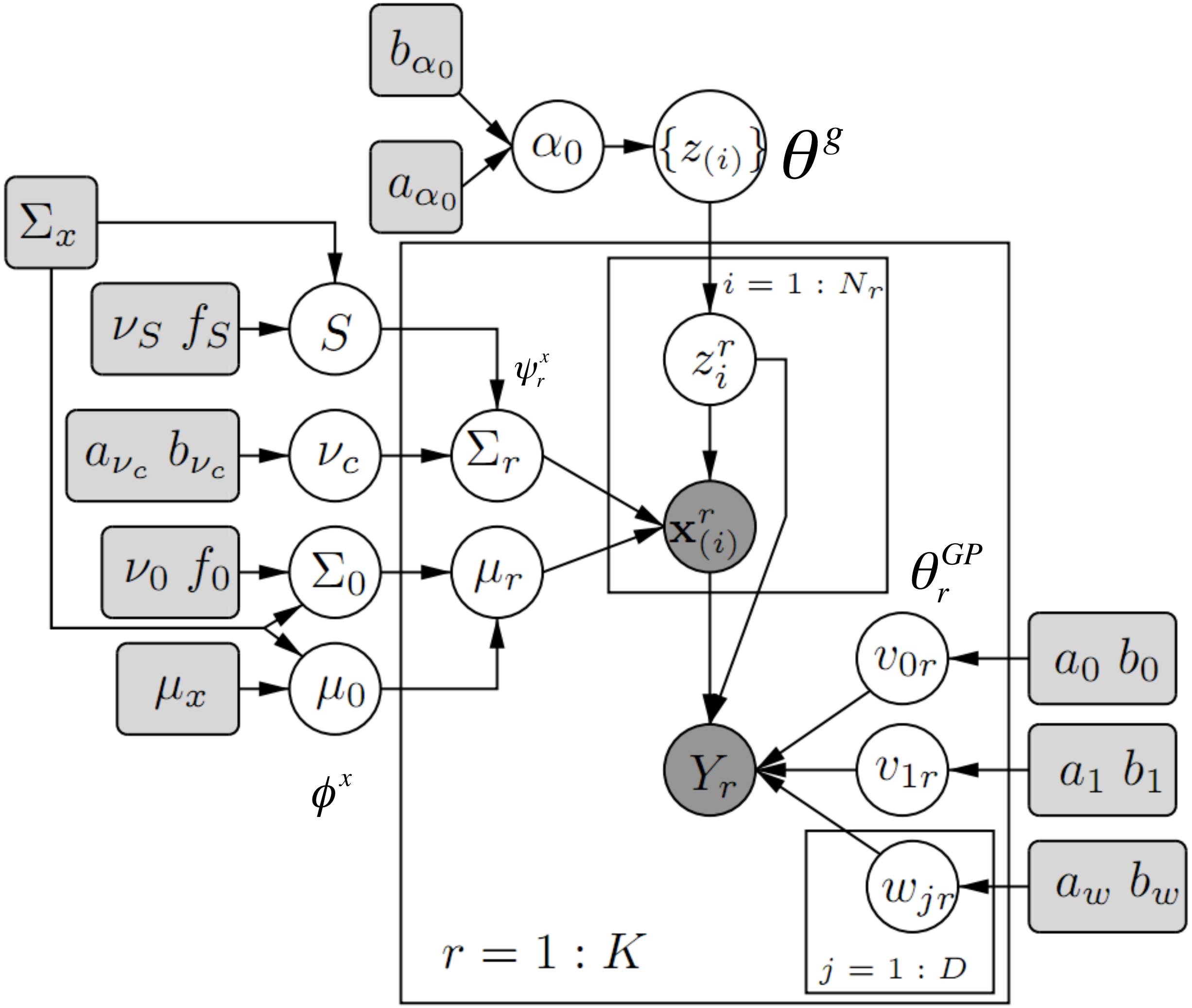


Rasmussen and Ghahramani 2002

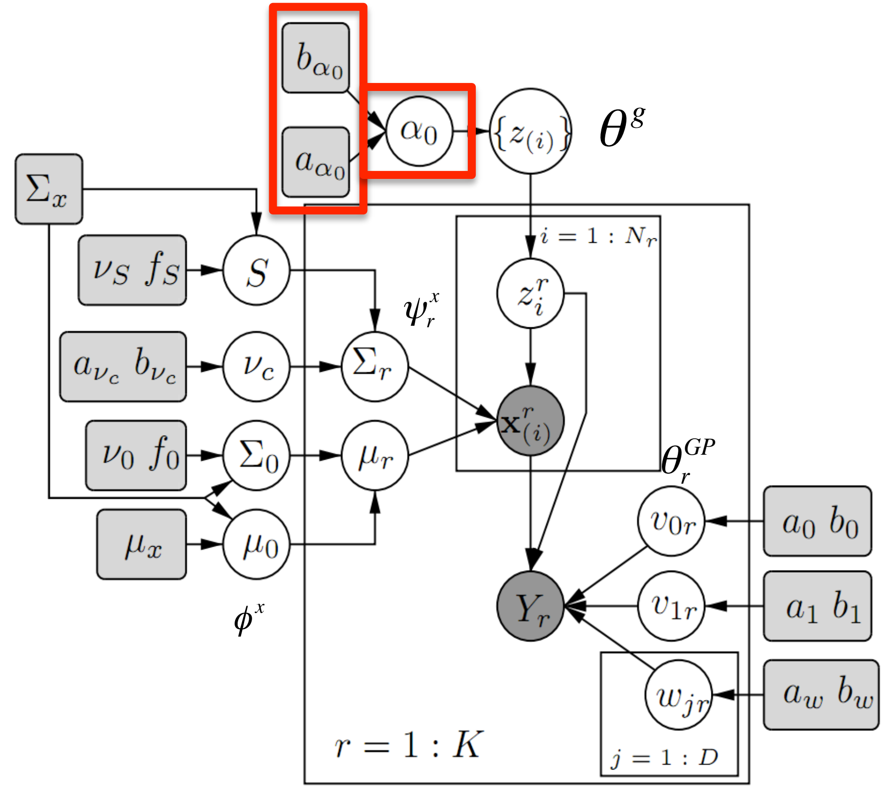
$$P(\{y_{(i)}\}|\{\mathbf{x}_{(i)}\},\theta) = \sum_{\mathcal{Z}} P(\{z_{(i)}\}|\{\mathbf{x}_{(i)}\},\theta^g) \prod_r P(\{y_{(i)} : z_{(i)} = r\}|\{\mathbf{x}_{(i)} : z_{(i)} = r\},\theta_r^{\mathrm{GP}})$$

This

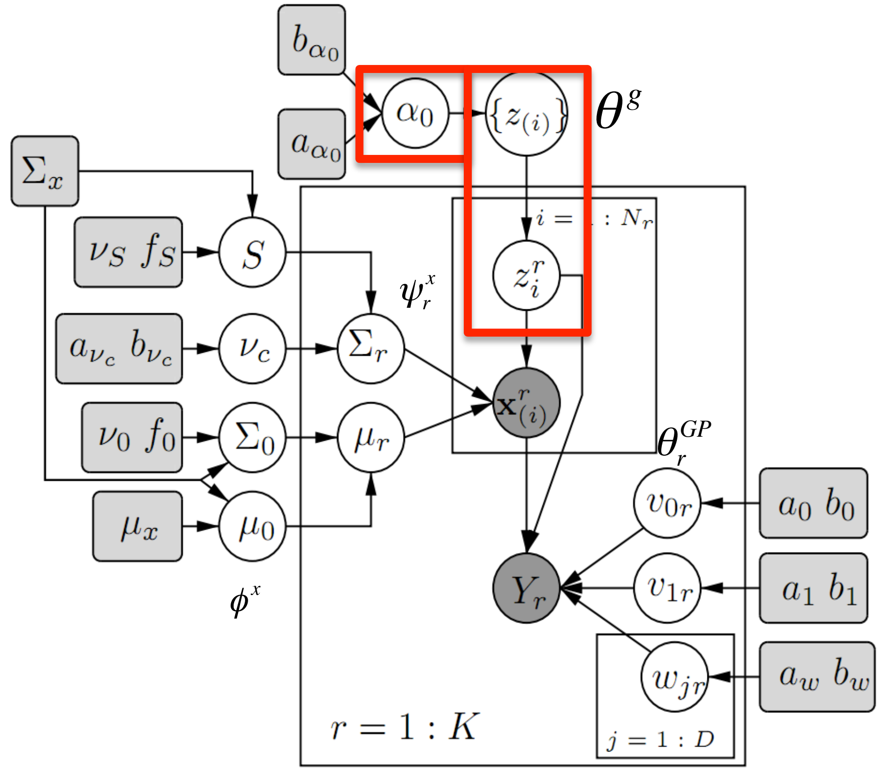
$$P(\{\mathbf{x}_{(i)}\},\{y_{(i)}\}|\theta) = \sum_{\mathcal{Z}} P(\{z_{(i)}\}|\theta^g) \times \prod_r P(\{y_{(i)} : z_{(i)} = r\}|\{\mathbf{x}_{(i)} : z_{(i)} = r\},\theta_r^{\mathrm{GP}}) P(\{\mathbf{x}_{(i)} : z_{(i)} = r\}|\theta^g)$$



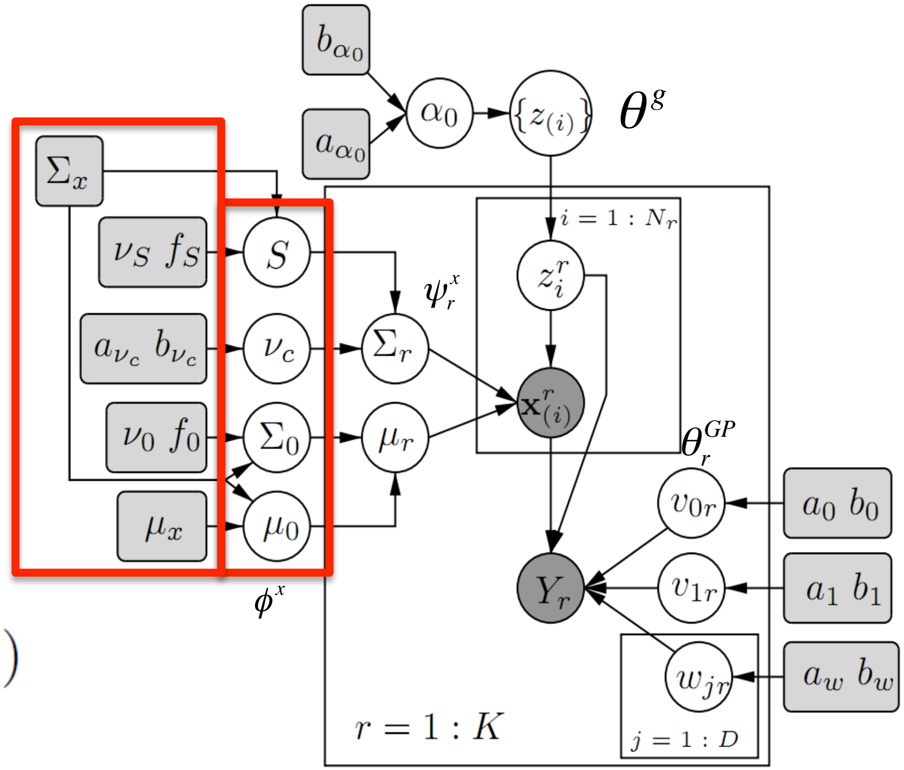
$$
\begin{aligned}
&P(\{\mathbf{x}_{(i)}, y_{(i)}\}_{i=1}^{N}, \{z_{(i)}\}_{i=1}^{N}, \{\psi_r^{\mathbf{x}}\}_{r=1}^{N}, \{\theta_r^{\mathrm{GP}}\}_{r=1}^{N}, \alpha_0, \phi^{\mathbf{x}} | N, \Omega) = \\
&\quad \prod_{r=1}^{N} \left[H_r^N P(\psi_r^{\mathbf{x}} | \phi^{\mathbf{x}}) P(X_r | \psi_r^{\mathbf{x}}) P(\theta_r^{\mathrm{GP}} | \Omega) P(Y_r | X_r, \theta_r^{\mathrm{GP}}) + (1 - H_r^N) D_0(\psi_r^{\mathbf{x}}, \theta_r^{\mathrm{GP}}) \right] \\
&\qquad \times P(\{z_{(i)}\}_{i=1}^{N} | N, \alpha_0) P(\alpha_0 | \Omega) P(\phi^{\mathbf{x}} | \Omega)
\end{aligned}
\tag{3}
$$



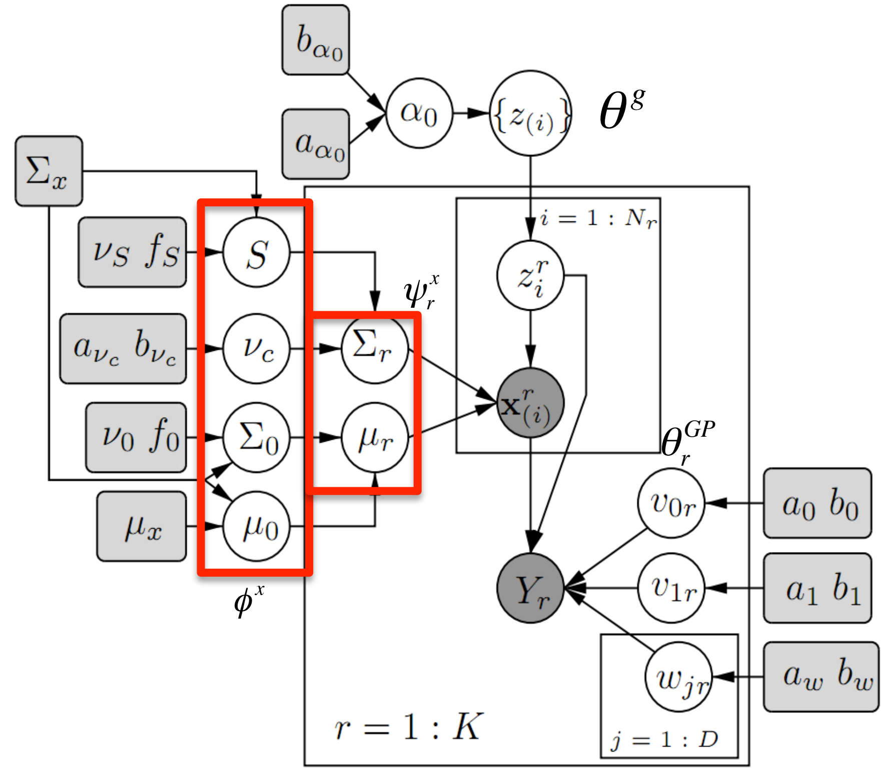
$$
\begin{aligned}
P(\alpha_0|\Omega) &= \mathcal{G}(\alpha_0;\ a_{\alpha_0}, b_{\alpha_0}) \\
P(\{z_{(i)}\}_{i=1}^{N}|N,\Omega) &= \mathcal{PU}(\alpha_0, N) \\
P(\phi^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_0;\ \mu_x, \Sigma_x/f_0)\mathcal{W}(\Sigma_0^{-1};\ \nu_0, f_0\Sigma_x^{-1}/\nu_0) \\
&\quad \mathcal{G}(\nu_c;\ a_{\nu_c}, b_{\nu_c})\mathcal{W}(S^{-1};\ \nu_S, f_S\Sigma_x/\nu_S) \\
P(\psi_r^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_r;\ \mu_0, \Sigma_0)\mathcal{W}(\Sigma_r^{-1};\ \nu_c, S/\nu_c) \\
P(X_r|\psi_r^{\mathbf{x}}) &= \mathcal{N}(X_r;\ \mu_r,\ \Sigma_r) \\
P(\theta_r^{\mathrm{GP}}|\Omega) &= \mathcal{G}(v_{0r};\ a_0, b_0)\mathcal{G}(v_{1r};\ a_1, b_1)\prod_{j=1}^{D}\mathcal{LN}(w_{jr};\ a_w, b_w) \\
P(Y_r|X_r, \theta_r^{\mathrm{GP}}) &= \mathcal{N}(Y_r;\ \mu_{Q_r}, \sigma_{Q_r}^2)
\end{aligned}
$$



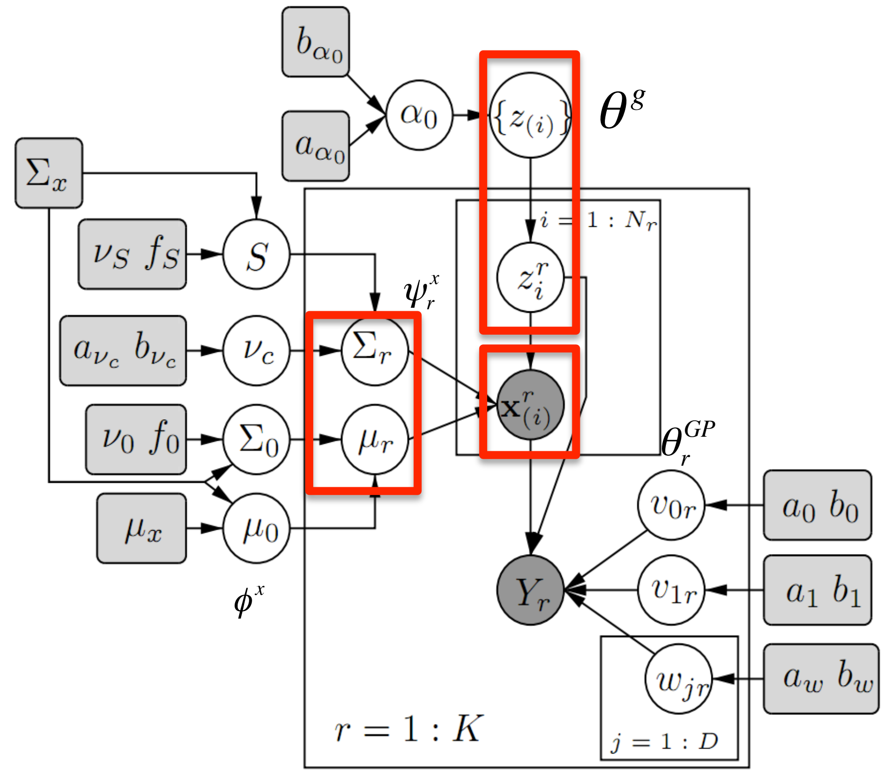
$$
\begin{aligned}
P(\alpha_0|\Omega) &= \mathcal{G}(\alpha_0;\ a_{\alpha_0}, b_{\alpha_0}) \\
P(\{z_{(i)}\}_{i=1}^{N}|N,\Omega) &= \mathcal{PU}(\alpha_0, N) \\
P(\phi^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_0;\ \mu_x, \Sigma_x/f_0)\mathcal{W}(\Sigma_0^{-1};\ \nu_0, f_0\Sigma_x^{-1}/\nu_0) \\
&\quad \mathcal{G}(\nu_c;\ a_{\nu_c}, b_{\nu_c})\mathcal{W}(S^{-1};\ \nu_S, f_S\Sigma_x/\nu_S) \\
P(\psi_r^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_r;\ \mu_0, \Sigma_0)\mathcal{W}(\Sigma_r^{-1};\ \nu_c, S/\nu_c) \\
P(X_r|\psi_r^{\mathbf{x}}) &= \mathcal{N}(X_r;\ \mu_r,\ \Sigma_r) \\
P(\theta_r^{\mathrm{GP}}|\Omega) &= \mathcal{G}(v_{0r};\ a_0, b_0)\mathcal{G}(v_{1r};\ a_1, b_1)\prod_{j=1}^{D}\mathcal{LN}(w_{jr};\ a_w, b_w) \\
P(Y_r|X_r, \theta_r^{\mathrm{GP}}) &= \mathcal{N}(Y_r;\ \mu_{Q_r}, \sigma_{Q_r}^2)
\end{aligned}
$$



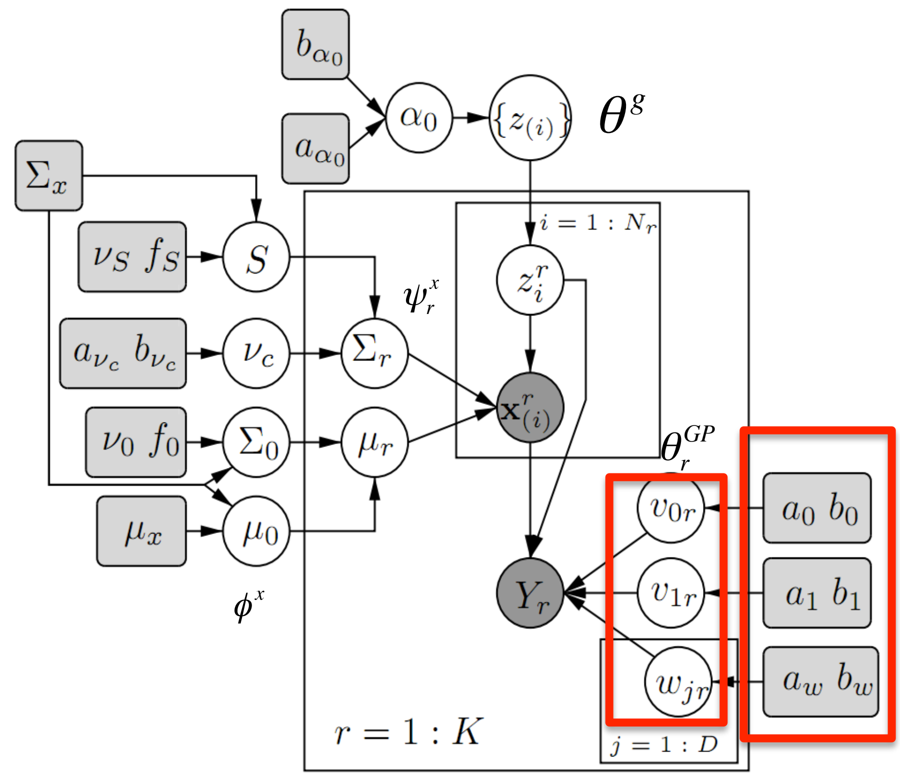
$$
\begin{aligned}
P(\alpha_0|\Omega) &= \mathcal{G}(\alpha_0;\ a_{\alpha_0}, b_{\alpha_0}) \\
P(\{z_{(i)}\}_{i=1}^{N}|N,\Omega) &= \mathcal{PU}(\alpha_0, N) \\
P(\phi^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_0;\ \mu_x, \Sigma_x/f_0)\mathcal{W}(\Sigma_0^{-1};\ \nu_0, f_0\Sigma_x^{-1}/\nu_0) \\
&\quad \mathcal{G}(\nu_c;\ a_{\nu_c}, b_{\nu_c})\mathcal{W}(S^{-1};\ \nu_S, f_S\Sigma_x/\nu_S) \\
P(\psi_r^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_r;\ \mu_0, \Sigma_0)\mathcal{W}(\Sigma_r^{-1};\ \nu_c, S/\nu_c) \\
P(X_r|\psi_r^{\mathbf{x}}) &= \mathcal{N}(X_r;\ \mu_r,\ \Sigma_r) \\
P(\theta_r^{\mathrm{GP}}|\Omega) &= \mathcal{G}(v_{0r};\ a_0, b_0)\mathcal{G}(v_{1r};\ a_1, b_1)\prod_{j=1}^{D}\mathcal{LN}(w_{jr};\ a_w, b_w) \\
P(Y_r|X_r, \theta_r^{\mathrm{GP}}) &= \mathcal{N}(Y_r;\ \mu_{Q_r}, \sigma_{Q_r}^2)
\end{aligned}
$$



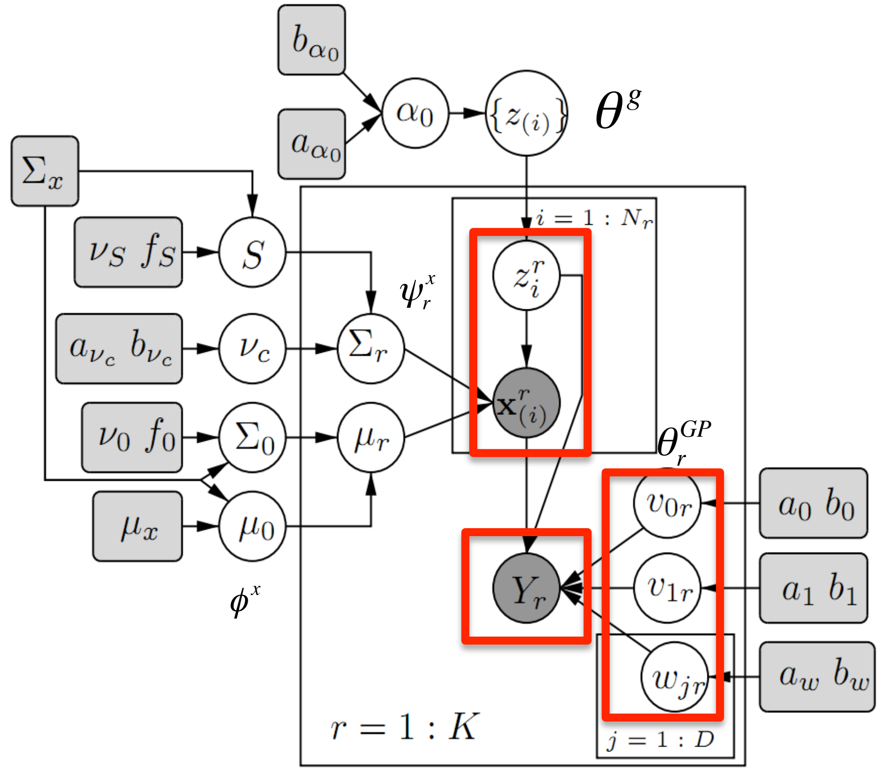
$$
\begin{aligned}
P(\alpha_0|\Omega) &= \mathcal{G}(\alpha_0;\ a_{\alpha_0}, b_{\alpha_0}) \\
P(\{z_{(i)}\}_{i=1}^{N}|N,\Omega) &= \mathcal{PU}(\alpha_0, N) \\
P(\phi^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_0;\ \mu_x, \Sigma_x/f_0)\mathcal{W}(\Sigma_0^{-1};\ \nu_0, f_0\Sigma_x^{-1}/\nu_0) \\
&\quad \mathcal{G}(\nu_c;\ a_{\nu_c}, b_{\nu_c})\mathcal{W}(S^{-1};\ \nu_S, f_S\Sigma_x/\nu_S) \\
P(\psi_r^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_r;\ \mu_0, \Sigma_0)\mathcal{W}(\Sigma_r^{-1};\ \nu_c, S/\nu_c) \\
P(X_r|\psi_r^{\mathbf{x}}) &= \mathcal{N}(X_r;\ \mu_r,\ \Sigma_r) \\
P(\theta_r^{\text{GP}}|\Omega) &= \mathcal{G}(v_{0r};\ a_0, b_0)\mathcal{G}(v_{1r};\ a_1, b_1)\prod_{j=1}^{D}\mathcal{LN}(w_{jr};\ a_w, b_w) \\
P(Y_r|X_r,\theta_r^{\text{GP}}) &= \mathcal{N}(Y_r;\ \mu_{Q_r}, \sigma_{Q_r}^2)
\end{aligned}
$$



$$
\begin{aligned}
P(\alpha_0|\Omega) &= \mathcal{G}(\alpha_0;\ a_{\alpha_0}, b_{\alpha_0}) \\
P(\{z_{(i)}\}_{i=1}^{N}|N,\Omega) &= \mathcal{PU}(\alpha_0, N) \\
P(\phi^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_0;\ \mu_x, \Sigma_x/f_0)\mathcal{W}(\Sigma_0^{-1};\ \nu_0, f_0\Sigma_x^{-1}/\nu_0) \\
&\quad \mathcal{G}(\nu_c;\ a_{\nu_c}, b_{\nu_c})\mathcal{W}(S^{-1};\ \nu_S, f_S\Sigma_x/\nu_S) \\
P(\psi_r^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_r;\ \mu_0, \Sigma_0)\mathcal{W}(\Sigma_r^{-1};\ \nu_c, S/\nu_c) \\
P(X_r|\psi_r^{\mathbf{x}}) &= \mathcal{N}(X_r;\ \mu_r,\ \Sigma_r) \\
P(\theta_r^{\text{GP}}|\Omega) &= \mathcal{G}(v_{0r};\ a_0, b_0)\mathcal{G}(v_{1r};\ a_1, b_1)\prod_{j=1}^{D}\mathcal{LN}(w_{jr};\ a_w, b_w) \\
P(Y_r|X_r, \theta_r^{\text{GP}}) &= \mathcal{N}(Y_r;\ \mu_{Q_r}, \sigma_{Q_r}^2)
\end{aligned}
$$



$$
\begin{aligned}
P(\alpha_0|\Omega) &= \mathcal{G}(\alpha_0;\ a_{\alpha_0}, b_{\alpha_0}) \\
P(\{z_{(i)}\}_{i=1}^{N}|N,\Omega) &= \mathcal{PU}(\alpha_0, N) \\
P(\phi^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_0;\ \mu_x, \Sigma_x/f_0)\mathcal{W}(\Sigma_0^{-1};\ \nu_0, f_0\Sigma_x^{-1}/\nu_0) \\
&\quad \mathcal{G}(\nu_c;\ a_{\nu_c}, b_{\nu_c})\mathcal{W}(S^{-1};\ \nu_S, f_S\Sigma_x/\nu_S) \\
P(\psi_r^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_r;\ \mu_0, \Sigma_0)\mathcal{W}(\Sigma_r^{-1};\ \nu_c, S/\nu_c) \\
P(X_r|\psi_r^{\mathbf{x}}) &= \mathcal{N}(X_r;\ \mu_r,\ \Sigma_r) \\
P(\theta_r^{\text{GP}}|\Omega) &= \mathcal{G}(v_{0r};\ a_0, b_0)\mathcal{G}(v_{1r};\ a_1, b_1)\prod_{j=1}^{D}\mathcal{LN}(w_{jr};\ a_w, b_w) \\
P(Y_r|X_r, \theta_r^{\text{GP}}) &= \mathcal{N}(Y_r;\ \mu_{Q_r}, \sigma_{Q_r}^2)
\end{aligned}
$$

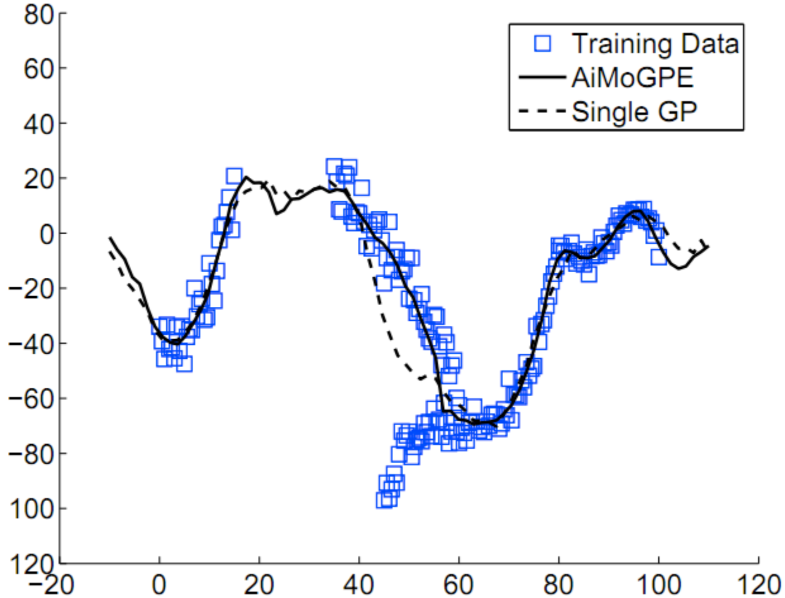


$$
\begin{aligned}
P(\alpha_0|\Omega) &= \mathcal{G}(\alpha_0;\ a_{\alpha_0}, b_{\alpha_0}) \\
P(\{z_{(i)}\}_{i=1}^{N}|N,\Omega) &= \mathcal{PU}(\alpha_0, N) \\
P(\phi^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_0;\ \mu_x, \Sigma_x/f_0)\mathcal{W}(\Sigma_0^{-1};\ \nu_0, f_0\Sigma_x^{-1}/\nu_0) \\
&\quad \mathcal{G}(\nu_c;\ a_{\nu_c}, b_{\nu_c})\mathcal{W}(S^{-1};\ \nu_S, f_S\Sigma_x/\nu_S) \\
P(\psi_r^{\mathbf{x}}|\Omega) &= \mathcal{N}(\mu_r;\ \mu_0, \Sigma_0)\mathcal{W}(\Sigma_r^{-1};\ \nu_c, S/\nu_c) \\
P(X_r|\psi_r^{\mathbf{x}}) &= \mathcal{N}(X_r;\ \mu_r,\ \Sigma_r) \\
P(\theta_r^{\mathrm{GP}}|\Omega) &= \mathcal{G}(v_{0r};\ a_0, b_0)\mathcal{G}(v_{1r};\ a_1, b_1)\prod_{j=1}^{D}\mathcal{LN}(w_{jr};\ a_w, b_w) \\
P(Y_r|X_r, \theta_r^{\mathrm{GP}}) &= \mathcal{N}(Y_r;\ \mu_{Q_r}, \sigma_{Q_r}^2)
\end{aligned}
$$

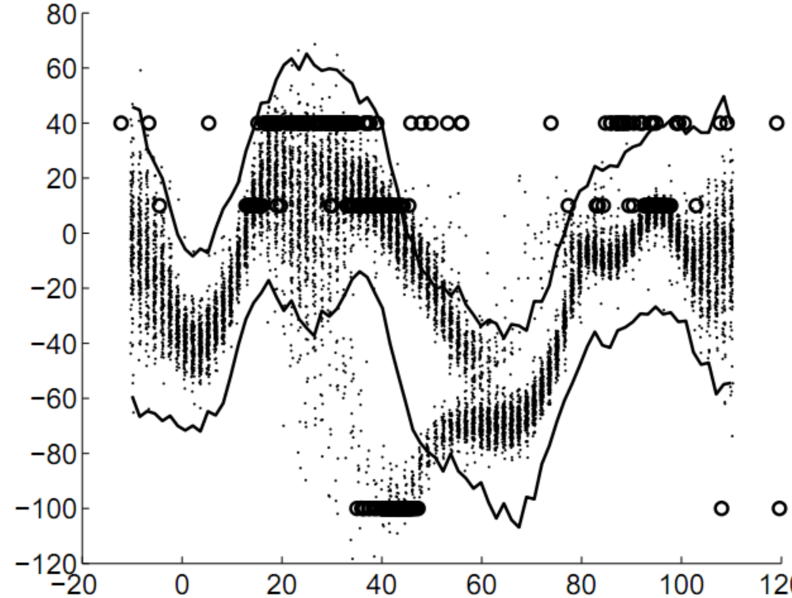
$$
\begin{array}{lrcl}
1. & P(\alpha_0|\Omega) & = & \mathcal{G}(\alpha_0;\ a_{\alpha_0}, b_{\alpha_0}) \\
2. & P(\{z_{(i)}\}_{i=1}^{N}|N,\Omega) & = & \mathcal{PU}(\alpha_0, N) \\
3. & P(\phi^{\mathbf{x}}|\Omega) & = & \mathcal{N}(\mu_0;\ \mu_x, \Sigma_x/f_0)\mathcal{W}(\Sigma_0^{-1};\ \nu_0, f_0\Sigma_x^{-1}/\nu_0) \\
 & & & \mathcal{G}(\nu_c;\ a_{\nu_c}, b_{\nu_c})\mathcal{W}(S^{-1};\ \nu_S, f_S\Sigma_x/\nu_S) \\
4. & P(\psi_r^{\mathbf{x}}|\Omega) & = & \mathcal{N}(\mu_r;\ \mu_0, \Sigma_0)\mathcal{W}(\Sigma_r^{-1};\ \nu_c, S/\nu_c) \\
5. & P(X_r|\psi_r^{\mathbf{x}}) & = & \mathcal{N}(X_r;\ \mu_r,\ \Sigma_r) \\
6. & P(\theta_r^{\mathrm{GP}}|\Omega) & = & \mathcal{G}(v_{0r};\ a_0, b_0)\mathcal{G}(v_{1r};\ a_1, b_1)\prod_{j=1}^{D}\mathcal{LN}(w_{jr};\ a_w, b_w) \\
7. & P(Y_r|X_r, \theta_r^{\mathrm{GP}}) & = & \mathcal{N}(Y_r;\ \mu_{Q_r}, \sigma_{Q_r}^2)
\end{array}
$$

1. Sample Dirichlet process concentration variable α_0 given the top-level hyper-parameters.
2. Construct a partition of N objects into at most N groups using a Dirichlet process. This assignment of objects is denoted by using a set the indicator variables $\{z_{(i)}\}_{i=1}^{N}$.
3. Sample the gate hyperparameters $\phi^{\mathbf{x}}$ given the top-level hyperparameters.
4. For each grouping of indicators $\{z_{(i)} : z_{(i)} = r\}$, sample the input space parameters $\psi_r^{\mathbf{x}}$ conditioned on $\phi^{\mathbf{x}}$. $\psi_r^{\mathbf{x}}$ de□nes the density in input space, in our case a full-covariance Gaussian.
5. Given the parameters $\psi_r^{\mathbf{x}}$ for each group, sample the locations of the input points $X_r \equiv \{\mathbf{x}_{(i)} : z_{(i)} = r\}$.
6. For each group, sample the hyper-parameters for the GP expert associated with that group, θ_r^{GP}.
7. Using the input locations X_r and hyper-parameters θ_r^{GP} for the individual groups, formulate the GP output covariance matrix and sample the set of output values, $Y_r \equiv \{y_{(i)} : z_{(i)} = r\}$ from this joint Gaussian distribution.

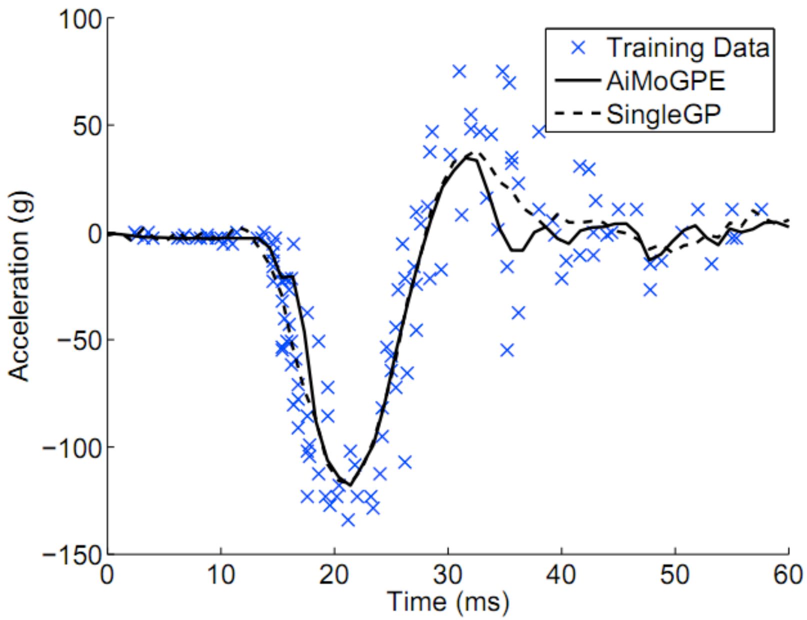
1. Update indicators $\{z_{(i)}\}$ by cycling through the data and sampling one indicator variable at a time. We use algorithm 8 from [9] with $m = 1$ to explore new experts.
2. Update input space parameters.
3. Update GP hyper-params using Hybrid Monte Carlo [10].
4. Update gate hyperparameters. Note that ν_c is updated using slice sampling [11].
5. Update DP hyperparameter α_0 using the data augmentation technique of Escobar and West [12].
6. Resample missing output values by cycling through the experts, and jointly sampling the missing outputs associated with that GP.


Training Data
AiMoGPE
Single GP

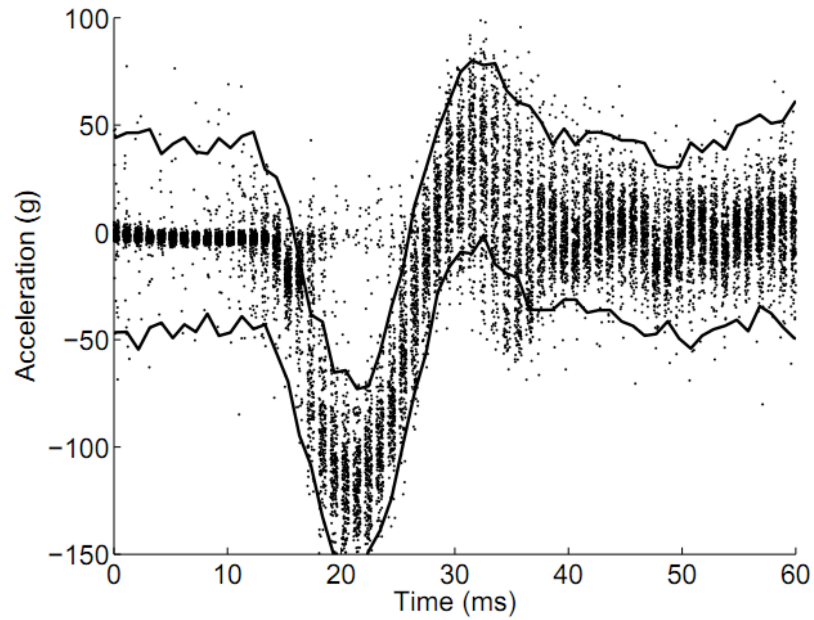
(A)



(B)


Training Data
AiMoGPE
SingleGP
Acceleration (g)
Time (ms)

(A)


Acceleration (g)
Time (ms)

(B)

Extensions

- More powerful input density models
- Explore inferring inverse functional mappings
- Explore perf. on large data sets

Questions?