Aggregating Dependent Gaussian Experts in Local Approximation

Hamed Jalali, Gjergji Kasneci

January 13, 2021

Problem Set-up

Gaussian Process (GP):

- Regression problem $y = f(x) + \epsilon$, where $x \in R^D$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
- Objective: learn f from a training set $\mathcal{D} = \{X, y\}_{i=1}^{n}$
  - $f \sim GP\left(m(x), k(x, x^{'})\right)$
- Training: maximise the log-marginal likelihood

$$\log p(y|X) = -\frac{1}{2} y^T \mathcal{C}^{-1} y - \frac{1}{2} \log |\mathcal{C}| - \frac{n}{2} \log(2\pi)$$

  where $\mathcal{C} = K + \sigma^2 I$.
- Cost: $\mathcal{O}(n^3)$ (because of the inversion and determinant of $\mathcal{C}$)

Distributed Gaussian Process

Distributed GP reduces the cost of the standard GP by distributing the training process.

- $\mathcal{D}$ is divided into M partitions $\mathcal{D}_1, \ldots, \mathcal{D}_M$, (called experts)
- Partitions are called experts
- The predictive distribution of the i'th expert $\mathcal{M}_i$ and test input x^* is $p_i(y^*|\mathcal{D}_i, x^*) \sim \mathcal{N}(\mu_i^*, \Sigma_i^*)$,

$$\mu_i^* = k_{i*}^T (K_i + \sigma^2 I)^{-1} y_i,$$

$$\Sigma_i^* = k_{**} - k_{i*}^T (K_i + \sigma^2 I)^{-1} k_{i*}.$$

Aggregation in Local Approximation

- The local Gaussian experts are conditionally independent.
  - CI assumption
- The predictive distribution of DGP is

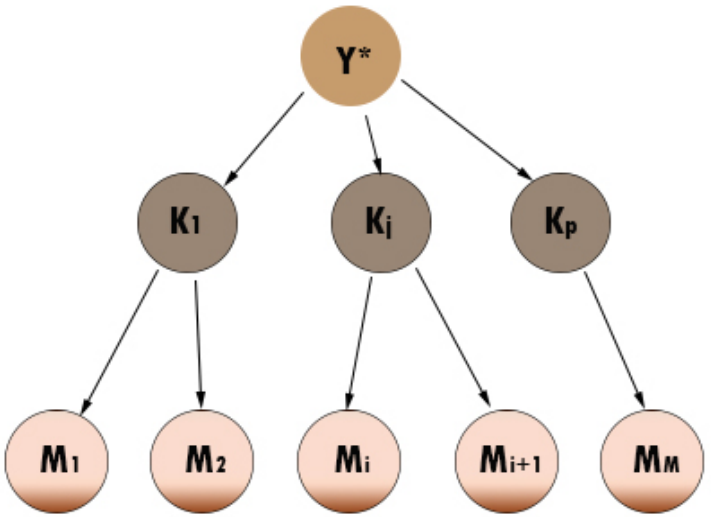
$$p(y^*|\mathcal{D}, x^*) \propto \prod_{i=1}^{M} p_i^{\beta_i}(y^*|\mathcal{D}_i, x^*).$$

- The weights $\beta = \{\beta_1, \ldots, \beta_M\}$ describe the importance of the experts.

The ensembles based on CI return sub-optimal solutions.

Aggregating Dependent Experts

- Instead of the Gaussian experts $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$, we define clusters of correlated experts, $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_P\}$, $P \ll M$.
- Aggregating the experts at each cluster leads to a new layer of experts, $\mathcal{K} = \{\mathcal{K}_1, \ldots, \mathcal{K}_P\}$, which are conditionally independent given y^*.



Dependency Detection

- let $\mu^*_{\mathcal{M}} = [\mu^*_1, \ldots, \mu^*_M]^T$ be a $n_t \times M$ matrix that contains the local predictions of M experts at n_t test points.
- Assumption: The joint distribution of the experts predictions is multivariate normal.
- Gaussian graphical model:

$$p(\mu^*_{\mathcal{M}}|h, \Omega) \propto \exp\left\{-\frac{1}{2}(\mu^*_{\mathcal{M}})^T \, \Omega \, \mu^*_{\mathcal{M}} + h^T \mu^*_{\mathcal{M}}\right\}.$$

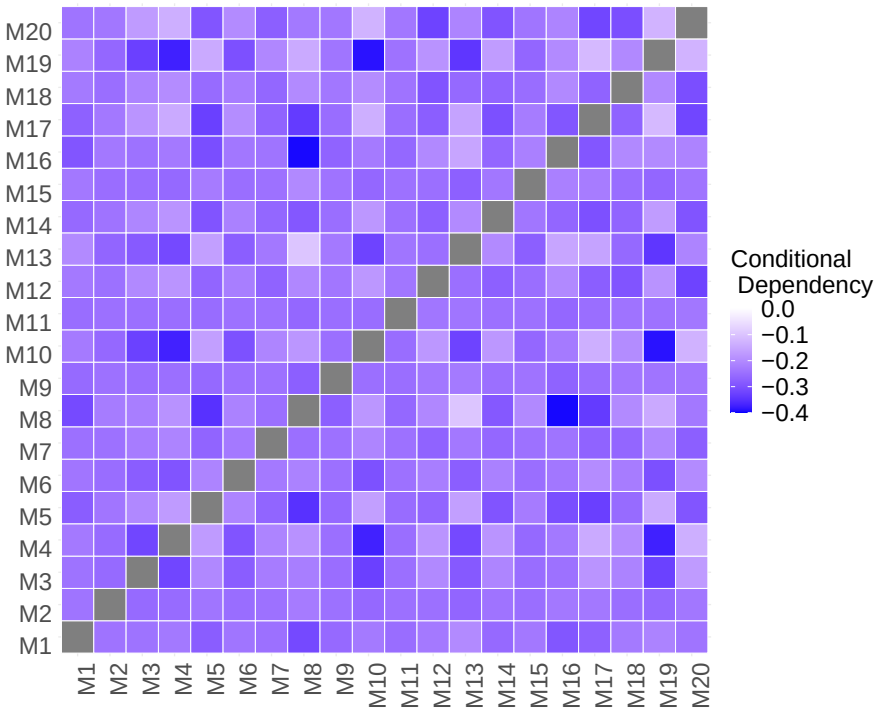
- Ω is the precision matrix and encodes the conditional dependency.
- Ω is calculated using the GLasso method

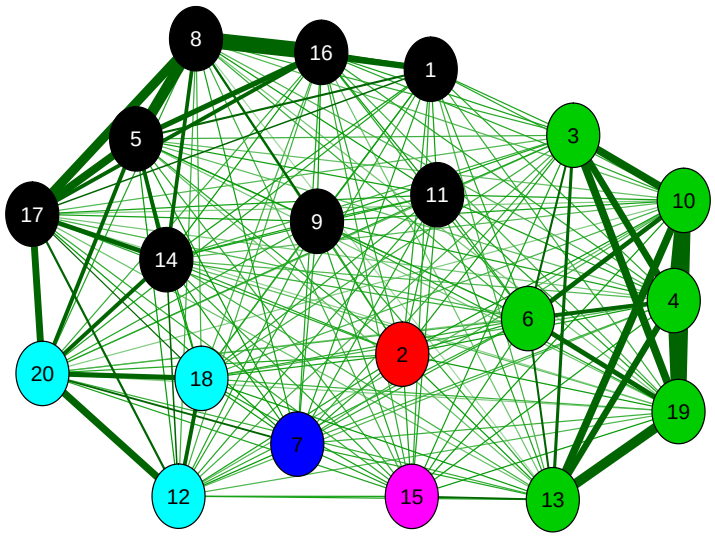
$$\hat{\Omega} = \arg\max_{\Omega} \log|\Omega| - trace(S\Omega) - \lambda \|\Omega\|_1 .$$

Experts clustering

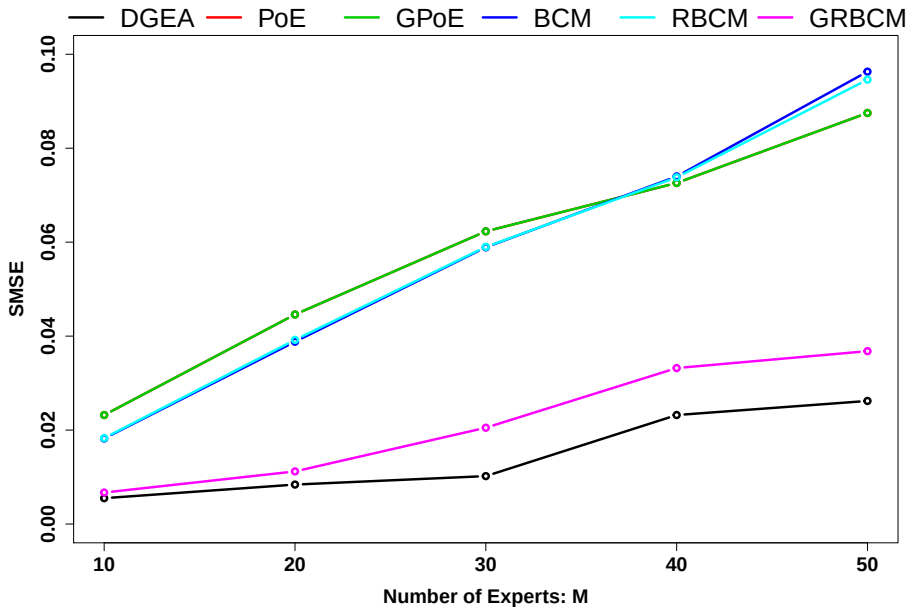
- The precision matrix is used to find the clusters of experts set $\mathcal{C}$.
- Each cluster $\mathcal{C}_i$ contains strongly dependent experts.
- We apply spectral clustering (SC) to find $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_P\}$.
- SC makes use of the similarity matrix (here Ω).
- To aggregate the experts at each cluster, we use GRBCM method that leads to $\mathcal{K} = \{\mathcal{K}_1, \ldots, \mathcal{K}_P\}$.
- New experts are conditionally independent.

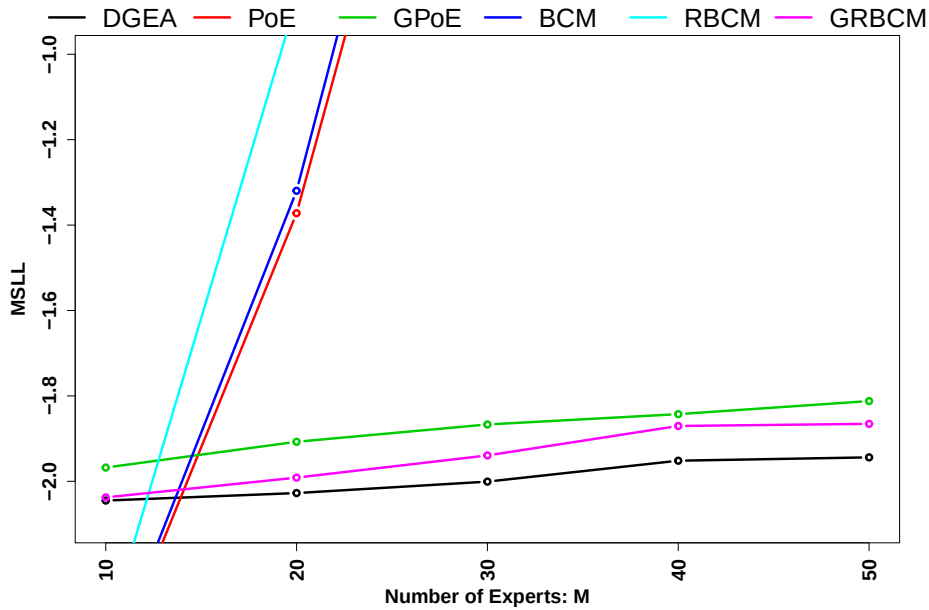
Experts clustering





Experiments-Synthetic Example





Experiments- Realistic Data-sets

	Pumadyn		*Kin40k*		*Sacros*		*Song*	
MODEL	SMSE	MSLL	SMSE	MSLL	SMSE	MSLL	SMSE	MSLL
DGEA (OURS)	**0.0486**	**-1.5133**	**0.0538**	**-1.3025**	**0.0269**	**-1.823**	**0.8084**	**-0.122**
POE	0.0505	4.8725	0.856	2.4153	0.0311	25.2807	0.8169	69.9464
GPOE	0.0505	-1.4936	0.0856	-1.2286	0.0311	-1.7756	0.8169	**-0.123**
BCM	0.0499	4.6688	0.0818	1.6974	0.0308	24.868	10.4291	44.1745
RBCM	0.0498	12.1101	0.0772	2.5256	0.0305	61.5392	5.4373	1.2089
GRBCM	0.0511	-1.488	0.0544	-1.2785	0.0305	-1.4308	0.8268	0.2073

Summary

We proposed a novel approach that leverages the dependencies between experts and improves the prediction quality. It

- uses an undirected graphical model to detect strong dependencies between experts
- defines clusters of interdependent experts
- provides consistent results when $n \to \infty$.