Aggregating Dependent Gaussian Experts in Local Approximation

Hamed Jalali, Gjergji Kasneci
Data Science and Analytics,
University of Tuebingen, Germany

Introduction

- Distributed Gaussian process (GP) is a prominent method to scale Gaussian Processes (GPs) to big data based on the divide-and-conquer approach, they train local experts by dividing the training set into subsets, thus reducing the time complexity.
- This strategy is based on the conditional independence assumption (CI), which basically means that there is a perfect diversity between the local experts.
- The CI assumption is often violated in practice and the aggregation of experts leads to sub-optimal and inconsistent solutions. The proposed models to cope with the consistency problem suffer from high computational costs and poor predictions, in particular when the data set is randomly divided into subsets.
- The key contribution of our work lies in considering the dependency between Gaussian experts and improve the prediction quality in an efficient way. First, we develop an approach to detect the conditional correlation between experts, and then we modify the aggregation using this knowledge.

Problem Set-up

Consider the regression problem $y = f(x) + \epsilon$, where the objective is to learn the latent function f from a training set $\mathcal{D} = \{X, y\}_{i=1}^{n}$.

Distributed GP involves dividing the $\mathcal{D}$ into M partitions $\mathcal{D}_1, \dots, \mathcal{D}_M$, and training the standard GP at each partition with predictive distribution $p_i(y^*|\mathcal{D}_i, x^*) \sim \mathcal{N}(\mu_i^*, \Sigma_i^*)$ for test point x^*.

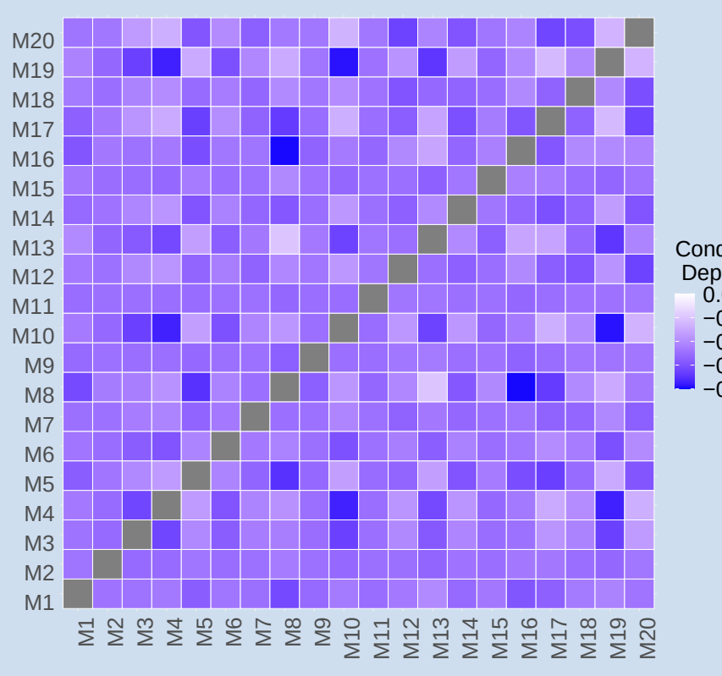
The predictive distribution is given as the product of multiple densities. For independent experts the predictive distribution is

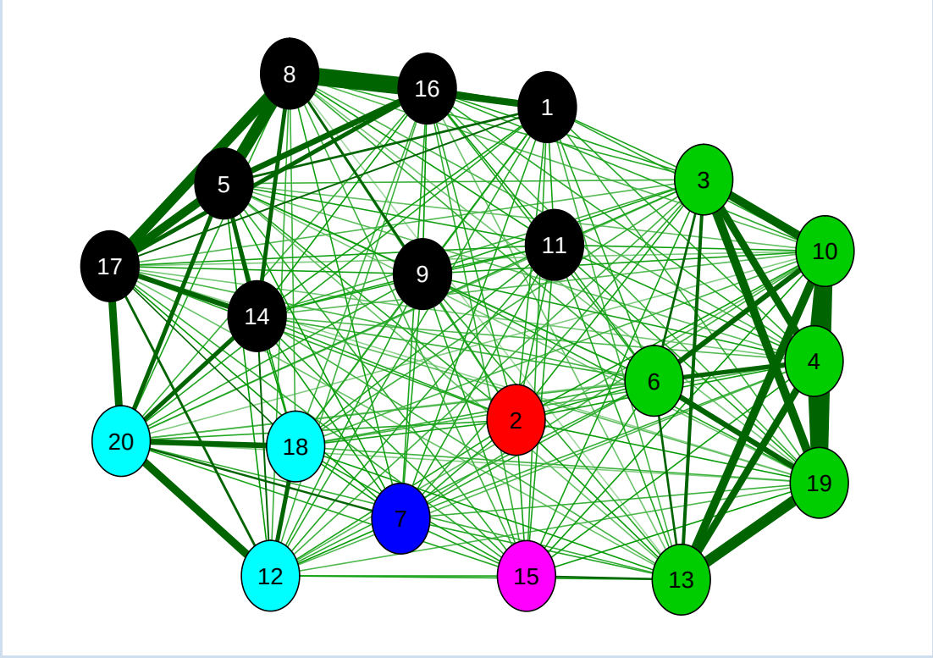
$$p(y^*|\mathcal{D}, x^*) \propto \prod_{i=1}^{M} p_i^{\beta_i}(y^*|\mathcal{D}_i, x^*). \qquad (1)$$

Methodology

Let $\mathcal{M} = \{\mathcal{M}_1, \dots, \mathcal{M}_M\}$ be Gaussian experts. The experts' predictions matrix $\mu_{\mathcal{M}}^*$ is used in order to detect strong dependencies between experts. This step results clusters of correlated experts, $\mathcal{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_P\}$, $P \ll M$.

Aggregating the experts at each cluster leads to a new layer of experts, $\mathcal{K} = \{\mathcal{K}_1, \dots, \mathcal{K}_P\}$, which are conditionally independent given y^*. The final prediction is done by using the $\mathcal{K}$ instead of $\mathcal{M}$.





Dependency Detection: If the joint distribution of Gaussian experts is multivariate normal, then the Gaussian graphical model is used to find the precision matrix Ω:

$$p(\mu_{\mathcal{M}}^*|\Omega) \propto \exp\left\{-\frac{1}{2}(\mu_{\mathcal{M}}^*)^T\, \Omega\, \mu_{\mathcal{M}}^*\right\}, \qquad (2)$$

where Ω encodes the conditional dependency and is calculated using the GLasso method:

$$\hat{\Omega} = \arg\max_{\Omega} \log|\Omega| - trace(S\Omega) - \lambda\,\|\Omega\|_1\,, \quad (3)$$

where $S = cov(\mu_{\mathcal{M}}^*)$, λ is the penalty term, and $\|.\|_1$ is the L_1-norm.

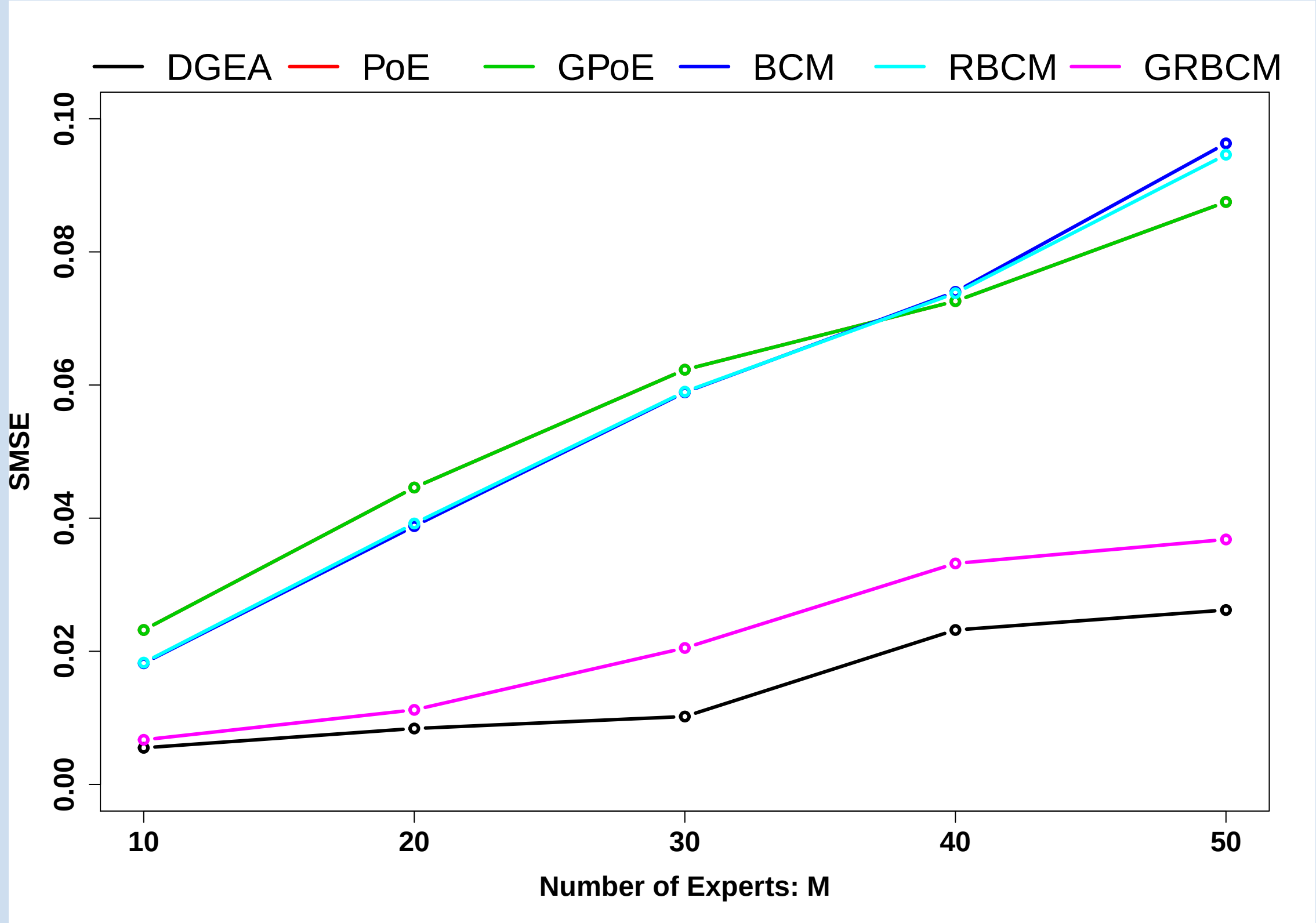
To find the new layer $\mathcal{K}$, we apply Spectral Clustering which makes use of the relevant eigenvectors of the Laplacian matrix of Ω.

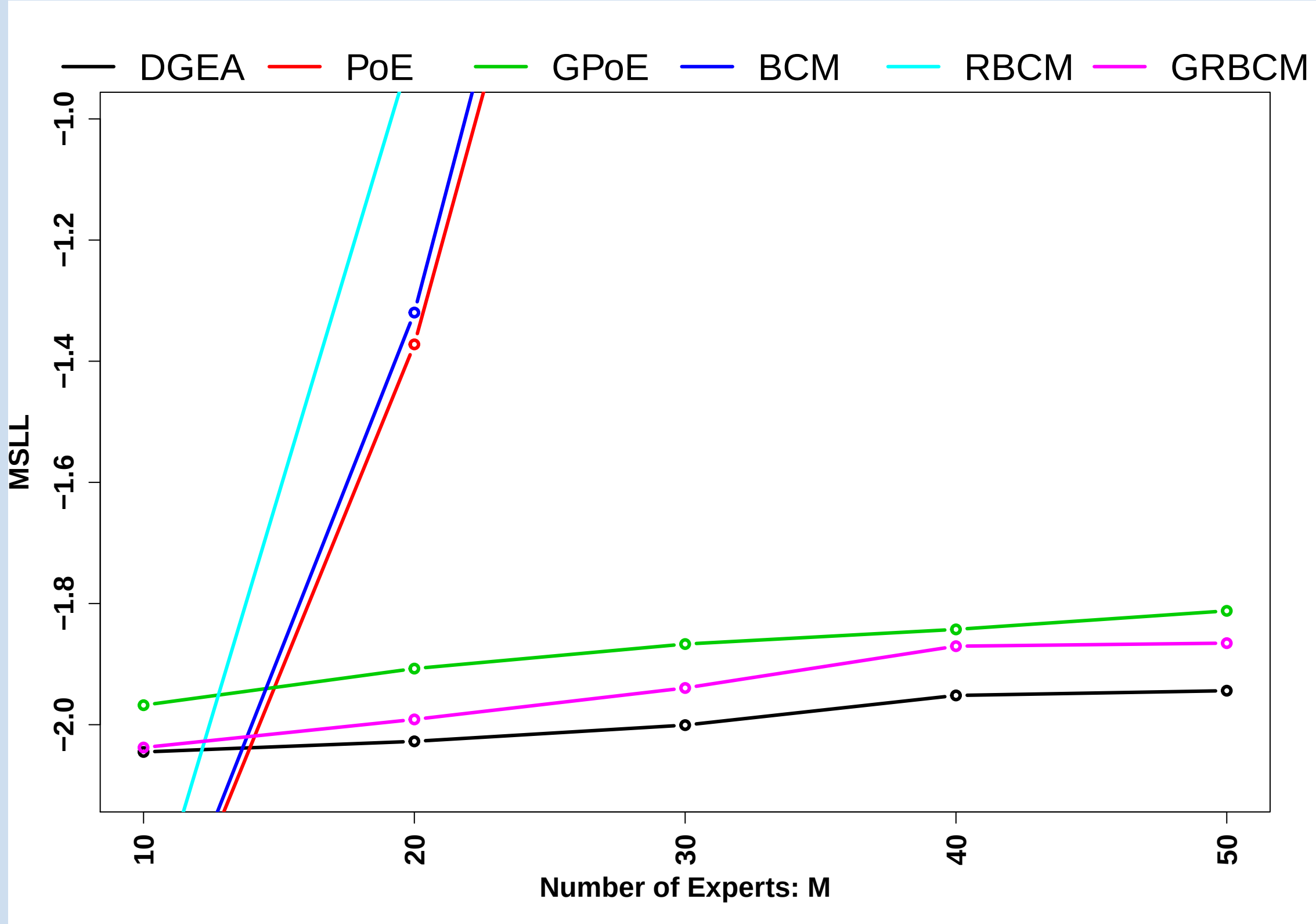
Aggregation

Algorithm 1 Aggregating Dependent Local Gaussian Experts

```
Require: $\mu_M^*$, $\lambda$, P
1: Calculate sample covariance S of experts' predictions
2: Estimate $\hat{\Omega}$ using GLasso
3: Estimate $\mathcal{H}$ by performing spectral clustering $SC(\hat{\Omega}$, P)
4: Obtain new experts $\{\mathcal{K}_1, \dots, \mathcal{K}_P\}$ using GRBCM
5: Aggregate new experts using GPoE or GRBCM
6: return The estimated mean and variance of $p(y^*|\mathcal{D}, x^*, \mathcal{K})$, i.e. $\mu_{\mathcal{K}}^*$ and $\Sigma_{\mathcal{K}}^*$
```

Sensitivity Analysis (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

Conclusion

We proposed DGEA, a novel DGP approach that leverages the dependencies between experts and improves the prediction quality. The DGEA

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

References

[1] H. Liu, J. Cai, Y. Ong, and Y. Wang. Generalized robust bayesian committee machine for large-scale gaussian process regression. *International Conference on Machine Learning*, pages 1–10, 2018.

[2] C. Uhler. Gaussian graphical models: an algebraic and geometric perspective. *arXiv:1707.04345*, 2017.

[3] A. Jaffe, E. Fetaya, B. Nadler, T. Jiang, and Y. Kluger. Unsupervised ensemble learning with dependent classifiers. *In Artificial Intelligence and Statistics*, pages 351–360, 2016.

[4] M. P. Deisenroth and J. W. Ng. Distributed gaussian processes. *International Conference on Machine Learning*, pages 1481–1490, 2015.