On Feature Learning of Recursive Feature Machines and Automatic Relevance Determination

Daniel Gedon*, Amirhesam Abedsoltan†, Thomas B. Schön*, Mikhail Belkin†

*Uppsala University, †UC San Diego

UPPSALA UNIVERSITET

UC San Diego

Problem definition

Setup: Regression with tabular data

Point estimate → RMSE, ...

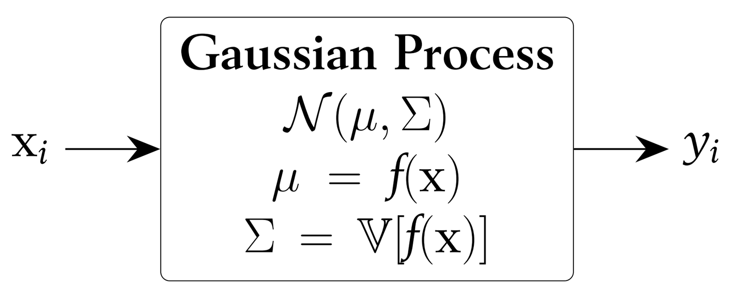
\+ **Uncertainty quantification** → NLL, coverage, avg. interval length, ...

SOTA: Boosting-based methods vs Classic approach: Gaussian Processes

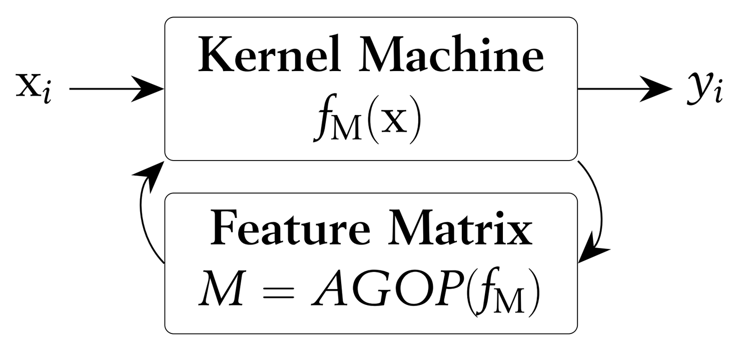
Question: Can we include feature learning in GPs?

Background

Gaussian Process (GP)



Recursive Feature Machine (RFM)



Model parametrization

pred. function $f(x) = k(x, X)\alpha$

RBF (or Laplace) kernel
$K_M(x, z) = \exp(-\gamma\|x - z\|^2_M)$
with Auto. Relevance Det. (ARD)
$M^{-1} = \mathrm{diag}([\ell_1^2, \dots, \ell_d^2])$

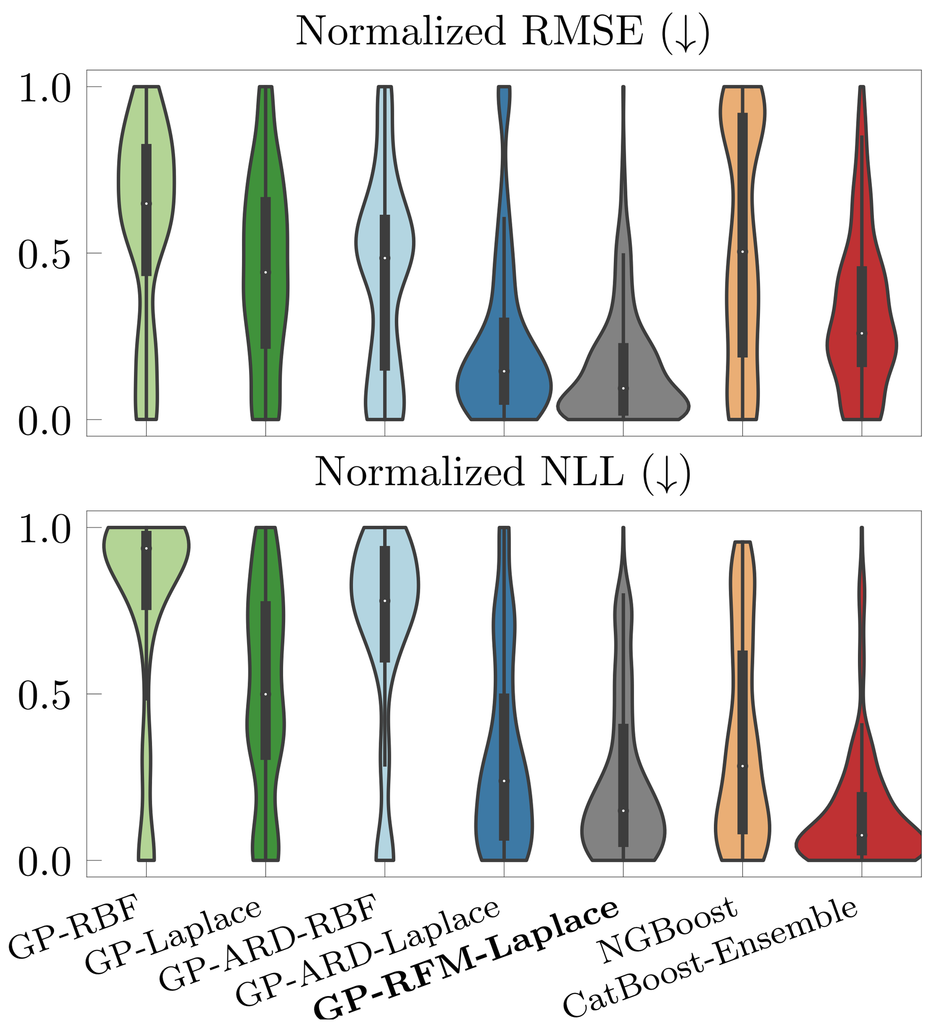
Laplace kernel
$K_M(x, z) = \exp(-\gamma\|x - z\|_M)$
with Mahalanobis distance
$\|x - z\|_M = \sqrt{(x - z)^\top M (x - z)}$

Training procedure

Maximum Likelihood Estimation
$\arg\min_{\theta} - \log p(y \mid X, \theta)$
with $\theta = \{\ell_1, \dots, \ell_d\}$

Kernel weights
$\alpha = (k_M(x, X) + \lambda_\alpha I_n)^{-1} y$
Average gradient outer product (AGOP)
$M = \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_M(x_i) \nabla_x f_M(x_i)^\top$

Tabular datasets



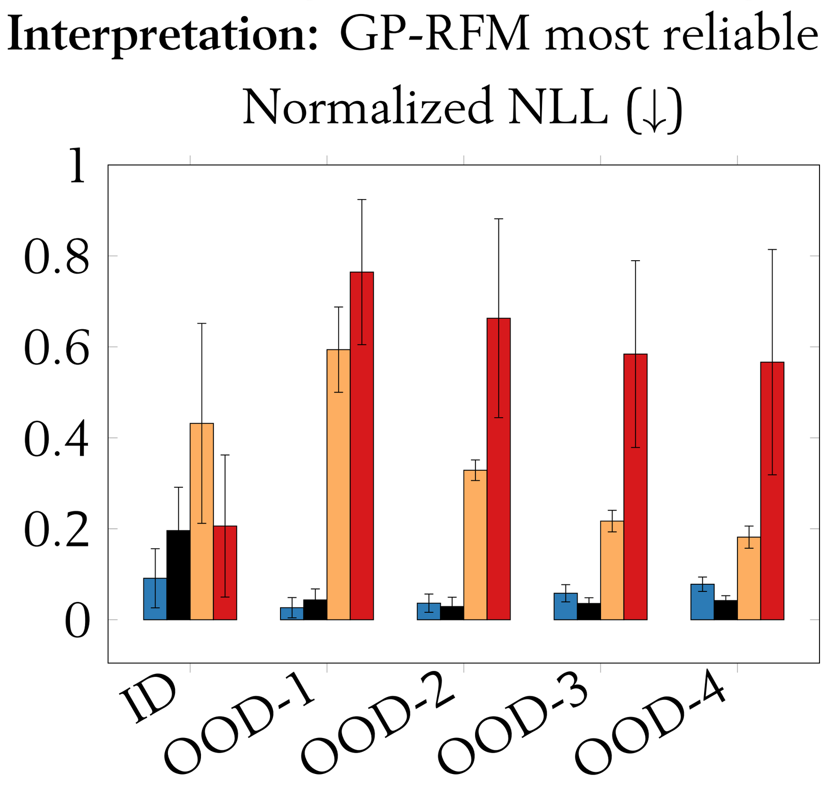
Data:
Tabular benchmark; 16 datasets; 5 – 613 features; 6 497 – 22 784 samples.

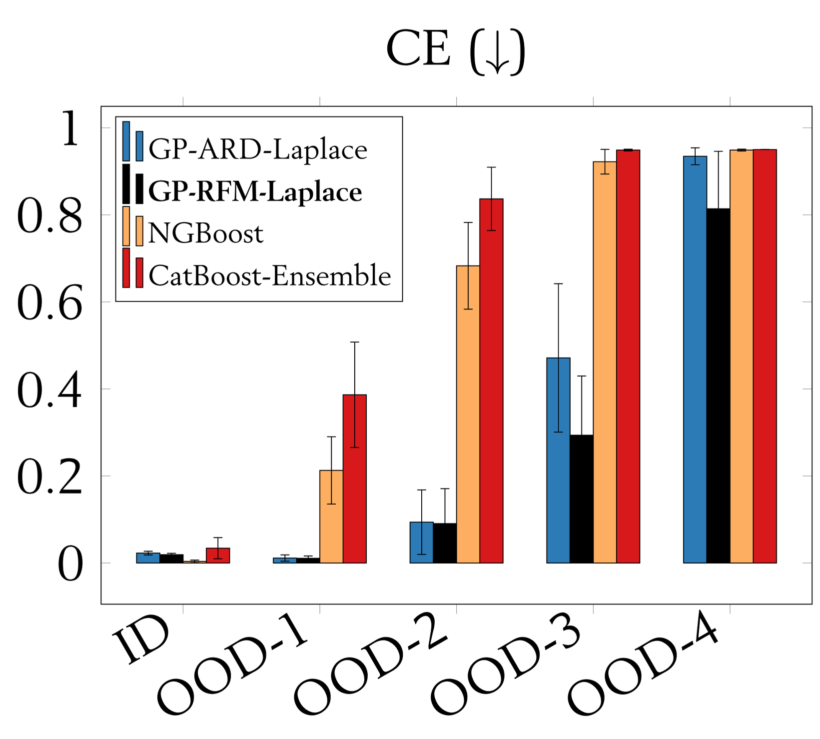
Setup:
Hyperparameter tuning over 20 seeds; normalize metrics for each dataset.

Interpretation:
ARD-Laplace and RFM-Laplace
- can outperform/match boosting methods.
- yield similar performance.

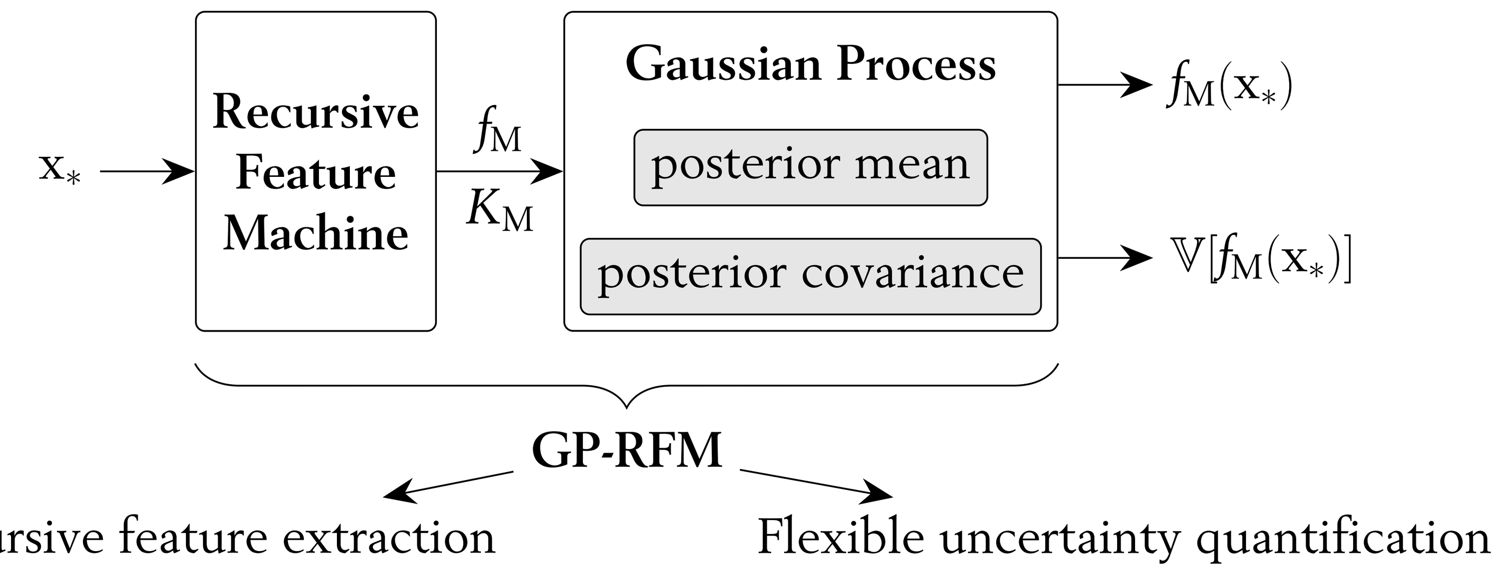
Extension: out-of-distribution data

Data: Housing data with increasing target (price) OOD shift.
Interpretation: GP-RFM most reliable method under OOD shift.



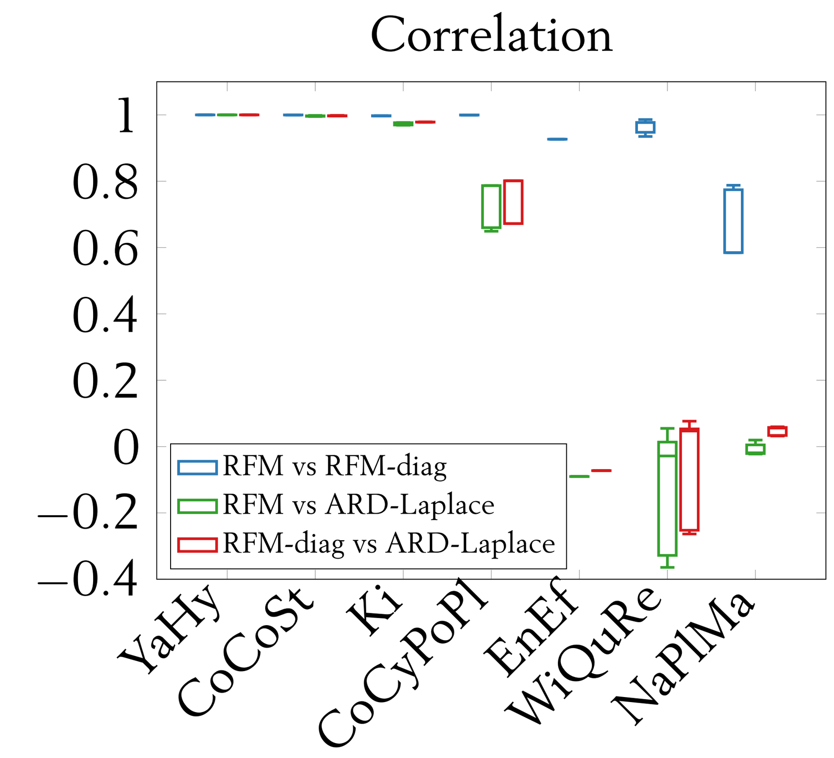


Method



Penalize off-diagonal elements: $M = \frac{1}{n}\sum_{i=1}^{n} \nabla_x f_M(x) \nabla_x f_M(x)^\top + \lambda_M I_d$

Correlation of learnt feature matrix M



Data:
UCI benchmark; 7 datasets;
4 – 16 features; 308 – 11 934 samples.

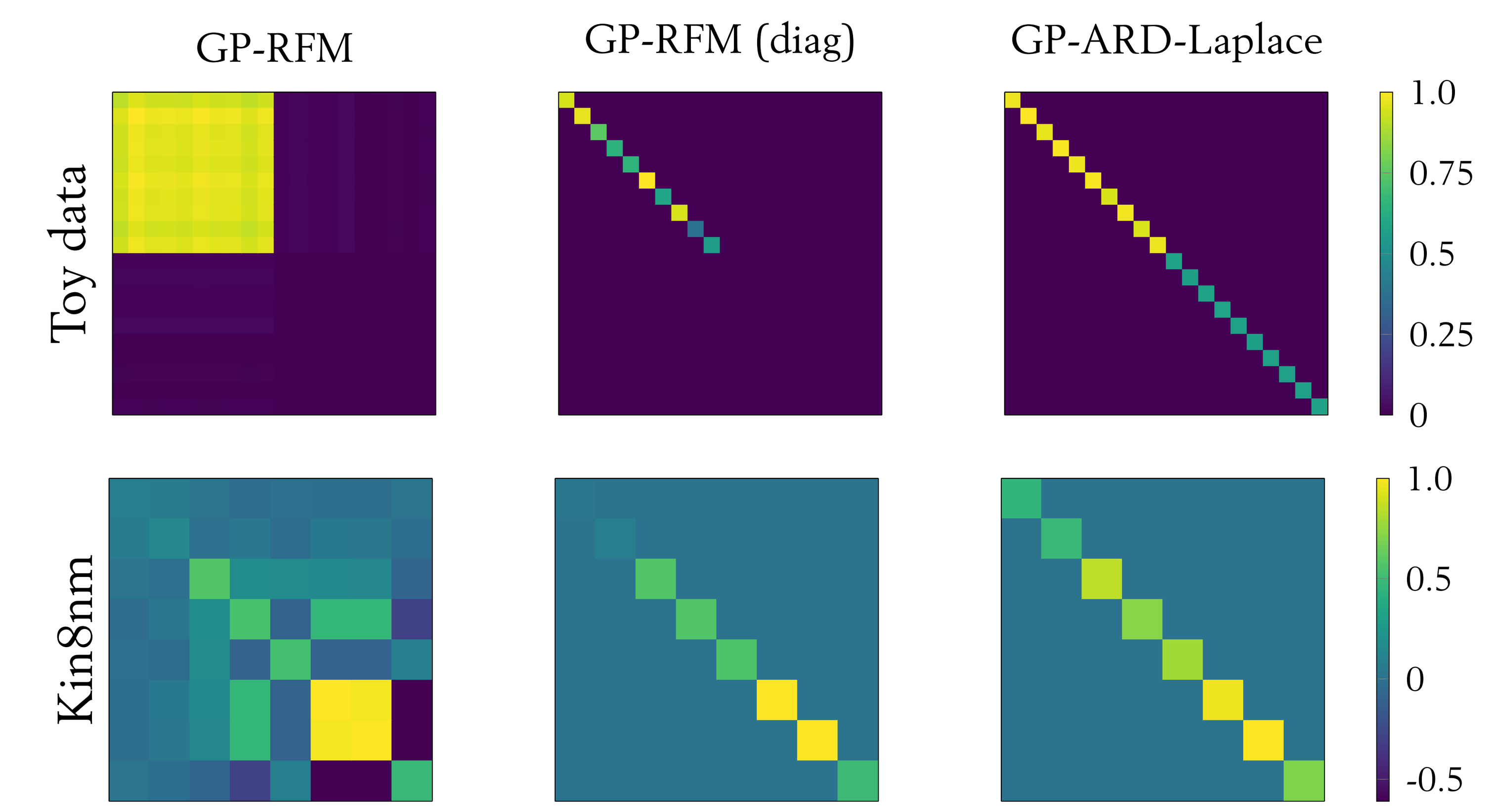
RFM and ARD perform similar on UCI.

Interpretation:
- RFM features sometimes correlate highly with ARD features.
- RFM and ARD can learn different features while performing similarly.

Visualizing feature matrices M

Data: $x \sim \mathcal{N}(0, I)$, $y = (\sum_{i=1}^{10} x_{[i]})^2$ → introduce correlation.

Interpretation:
- Full-RFM learns features to captures correlation.
- Diag-RFM and ARD loose crucial information.



Conclusion

Combining RFMs with GPs → **(1) competitive results** → **(2) partly correlating features**

Main message:
1. RFM and ARD kernels learn sometimes similar features.
2. RFM-Laplace and ARD-Laplace can outperform boosting methods.
3. RFMs are well suited for uncertainty quantification.

Open questions:
- Why do RMFs and ARD sometimes learn different features?
- Is there a theoretical connection between AGOP and MLE?
- Which real-world examples require the full-RFM?

References

Mechanism of feature learning in deep fully connected networks and kernel machines that recursively learn features
Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, Mikhail Belkin
ArXiv preprint arXiv:2212.13881.

daniel.gedon@it.uu.se, aabedsoltan@ucsd.edu