Correcting Missingness in Passively-Generated Mobile Data using Multi-Task Gaussian Processes

Ekin Ugurel, Xiangyang Guan,
Yanchao Wang, Shuai Huang,
Qi Wang, Cynthia Chen

TRB Standing Committee on Travel Survey Methods (AEP25)

January 9th, 2023

UNIVERSITY of WASHINGTON

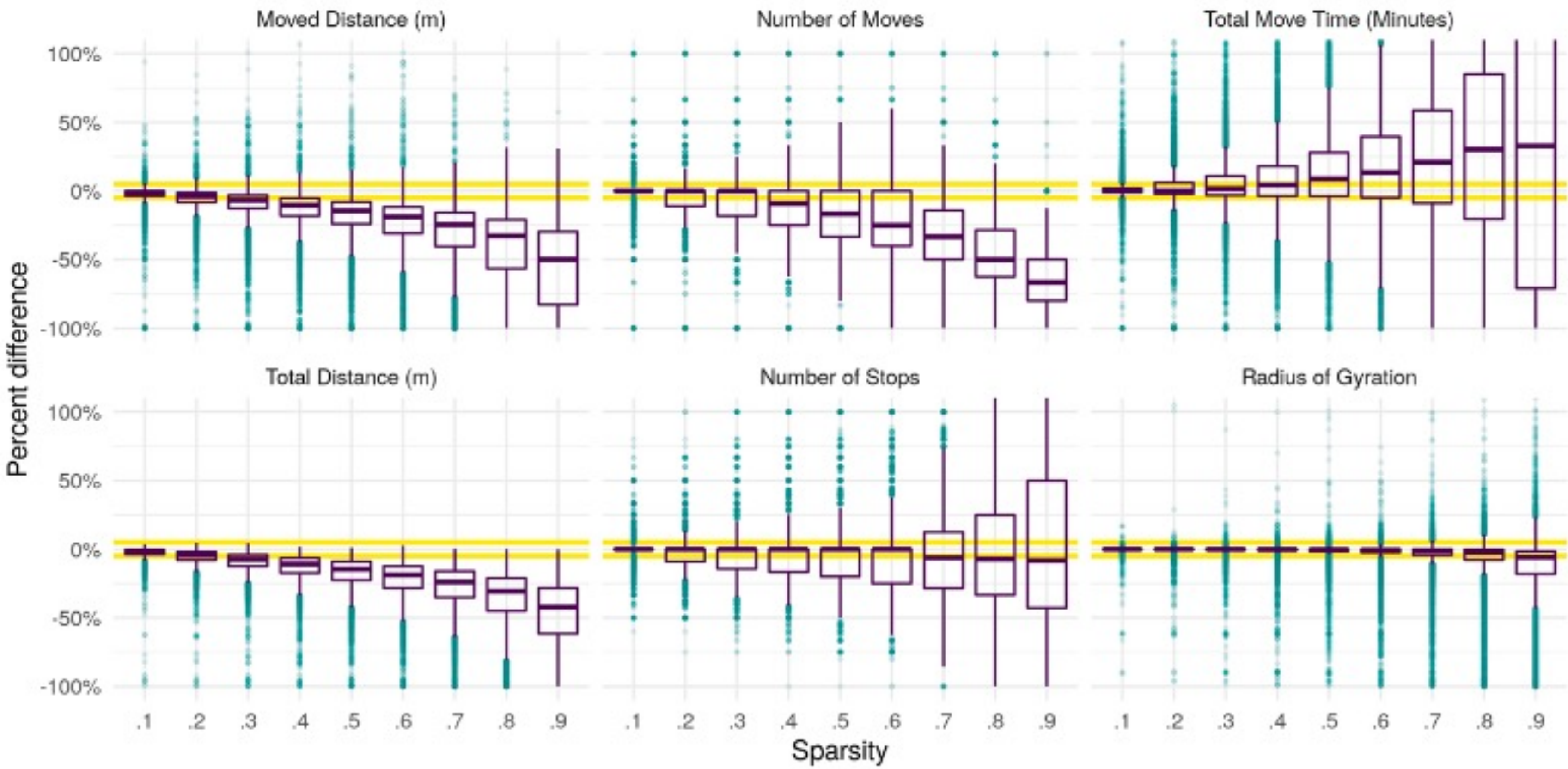
Motivation

- **The past:** active solicitation (i.e., travel surveys)
  - Low sample sizes
  - Mixed reporting accuracy
  - Demographic info available
- **The present (and future):** passively-generated mobile data
  - Massive sample sizes
  - Found “in the wild”; data points are not generated due to any research-related processes
  - Prevalence of sparsity (large chunks of missing data)

Consequences of Missingness

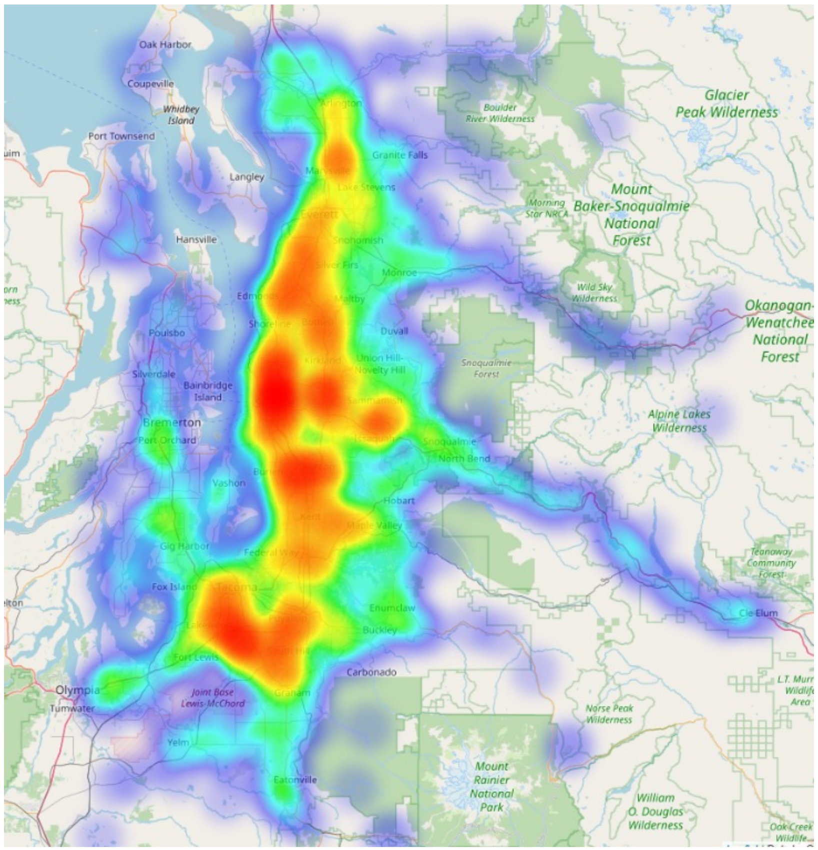
> Mobility patterns observed from sparse mobile data are vulnerable downward bias
>
> – Fewer trips are inferred from sparse mobile data compared to household survey data (Wang et al., 2019)
>
> – Sparse mobile data underestimates the maximum distance covered by an individual in a given time period (Guan et al., 2022; McCool et al., 2022)

Inferred travel behavior is a function of sparsity



Source: McCool et al., 2022

Spectus Dataset



Observations per user per day	
Mean	135
Standard Deviation	162
Min	1
25%	40
50%	98
75%	181
Max	9,159

(left) Heat map of a random sample of 20,000 GPS traces in the Greater Seattle Area; (right) summary GPS trace count statistics of the entire sample of 2,000 users

Notation

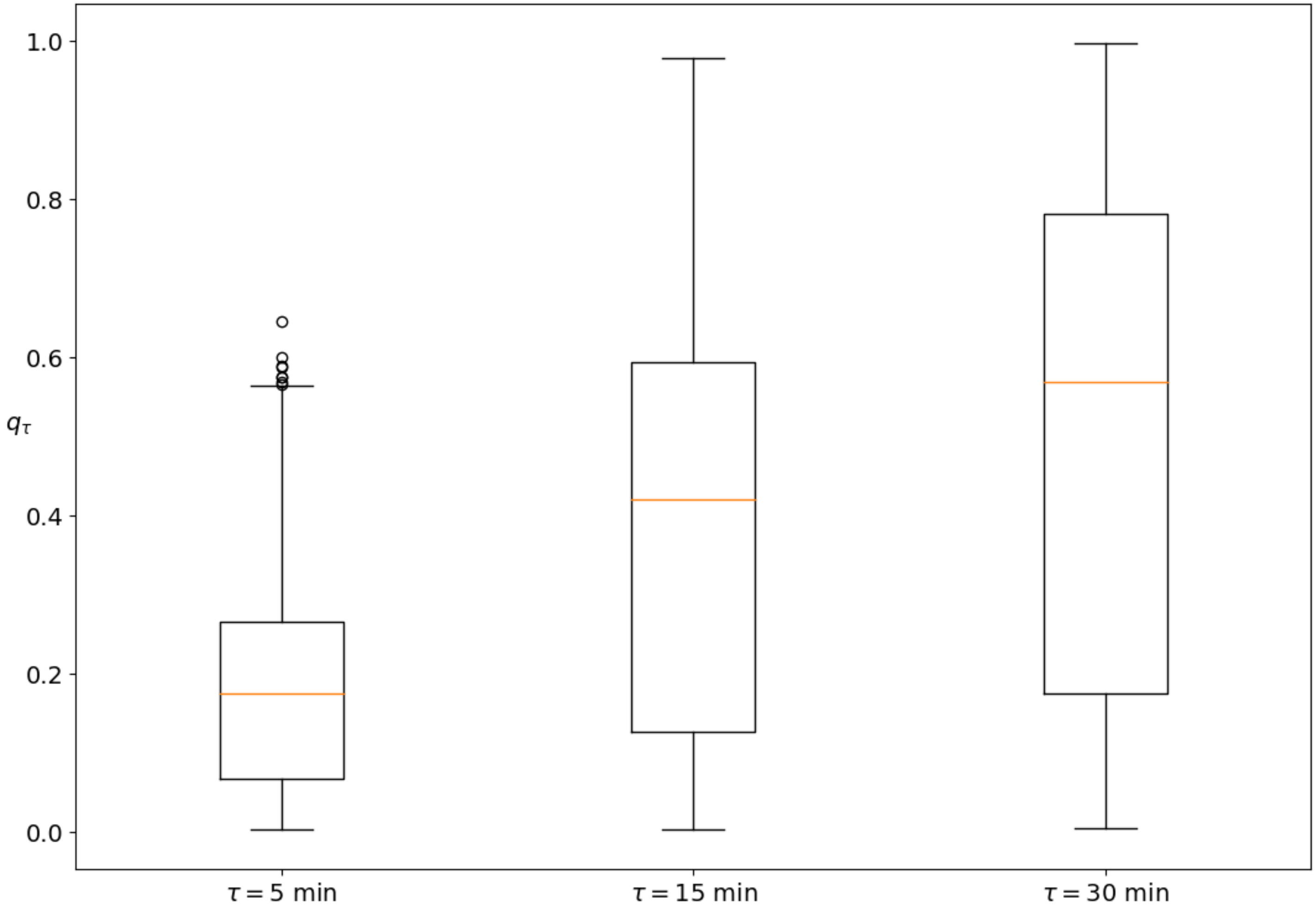
We discretize a user's total available data time $\mathcal{T}$ into P intervals $(1, \dots, P)$ of length τ, which we refer to as the "temporal resolution." The choice of τ is important—it decides the sparseness of a user's observed trajectory, in which each interval is assigned an indicator variable

$$\mathrm{I}_p = \begin{cases} 1 & \textbf{if } p \textbf{ has at least one observation} \\ 0 & \textbf{otherwise} \end{cases}$$

We thus define temporal occupancy (or the inverse of sparsity) as

$$q_\tau = \frac{1}{P}\sum_{p=1}^{P} \mathrm{I}_p$$

“Missingness” is a function of how we define sparsity



Challenges

> **Mode changes**
> – Can occur intra- or inter-trip

> **Heterogeneous human mobility behavior**
> – Varying tendencies to explore and exploit

Any method to correct missingness need to be flexible enough to capture these individual-level complexities

Research Question

- **To what extent is a Gaussian Process-based framework a suitable method for correcting missingness in mobile data?**
  - What are some factors that affect imputation accuracy?
  - How do model parameters change as a function of trip characteristics?

Methodology: Gaussian Process (GP)

- Generalization of the Gaussian probability distribution
  - Probability distribution → scalars or vectors (if multivariate)
  - Stochastic *process* → properties of functions
- Fully specified by a mean function $m(\mathbf{x})$ and a covariance function $K(\mathbf{x}, \mathbf{x}')$. Formally,

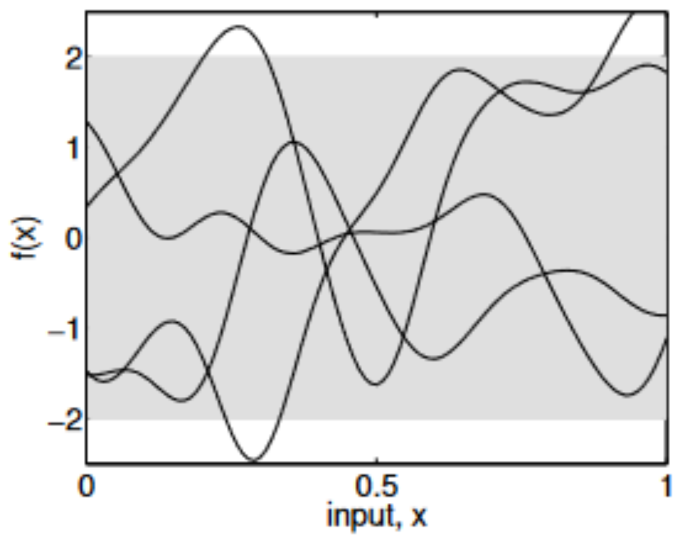
$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})],$$

$$K(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))],$$

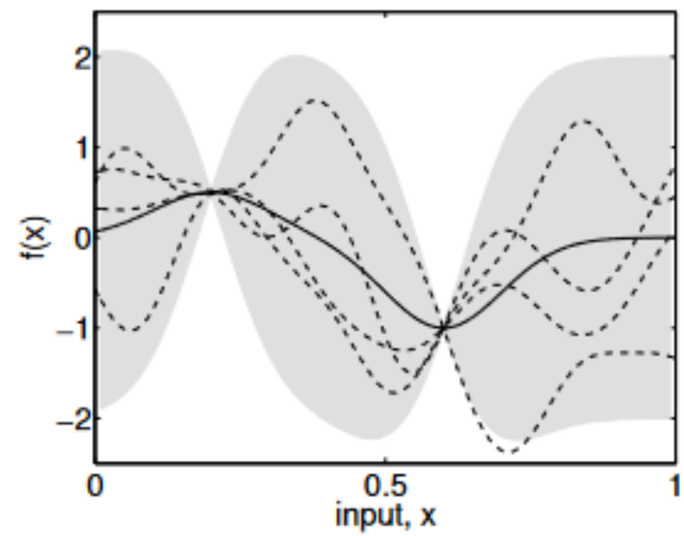
$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}'))$$

- The covariance function is how we deal with the aforementioned challenges of mode changes and heterogeneous travel behavior

GPs consider the space of all possible models and output the most likely given your training data



(a), prior



(b), posterior

Panel (a) shows four samples drawn from the prior distribution. Panel (b) shows the situation after two datapoints have been observed. The mean prediction is shown as the solid line and four samples from the posterior are shown as dashed lines. Shaded region denotes twice the standard deviation at each input value x

Methodology

- **Multi-task learning**
  - Imputing latitudes/longitudes simultaneously, while leveraging the correlations between them
    - Caused by the built environment + people's relationship to it
- **Nonlinear optimization**
  - Marginal log-likelihood maximization
  - Adaptive Moment Estimation (Kingma and Ba, 2017)
  - Initialization is a prerequisite to avoid model misspecification
- **Uncertainty quantification**

Implementation

- **GPyTorch (Gardner et al., 2018)**
  - Extensive documentation, familiarity to ML researchers
- **Downside: computational complexity**
  - GPs run in $O(n^3)$ due to the inversion of the $n \times n$ covariance matrix

Table 1: Temporal dimensions used in our experiments

Variable	Notation	Type	Model Inputs
Unix time (normalized)	t_u	Continuous	[0, 1, …, $\mathcal{T}$]
Seconds after midnight	t_s	Continuous	[0, 1, …, 86400]
Day of week	t_d	Categorical	[0, 1, 2, 3, 4, 5, 6]
Week of the month	t_{wk}	Categorical	[0, 1, 2, 3, 4]
Public holiday	t_h	Binary	[0, 1]
Weekend or not	t_{we}	Binary	[0, 1]
AM peak	t_{am}	Binary	[0, 1]
PM peak	t_{pm}	Binary	[0, 1]

Kernels for Modeling Mobile Data

> **Squared Exponential (SE)**

$$K_{SE}(\mathbf{x}-\mathbf{x}') = \sigma^2 \exp\left(-\frac{1}{2\ell^2}|\mathbf{x}-\mathbf{x}'|^{\mathbf{2}}\right)$$

> **Periodic (PER)**

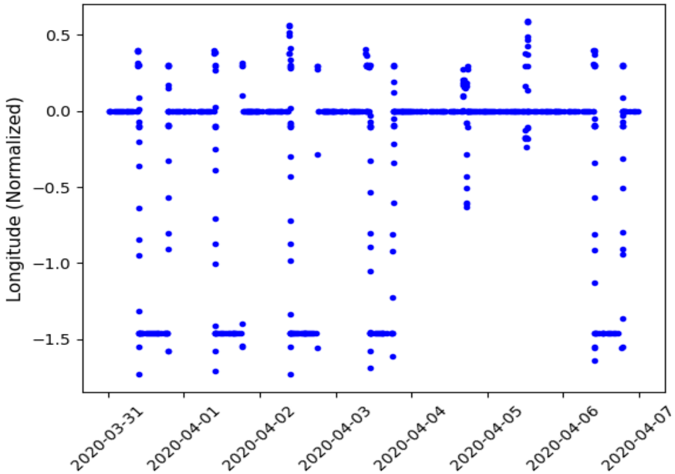
$$K_{PER}(\mathbf{x}-\mathbf{x}') = \sigma^2 \exp\left(-\frac{2\sin^2(\pi|\mathbf{x}-\mathbf{x}'|/p)}{\ell^2}\right)$$

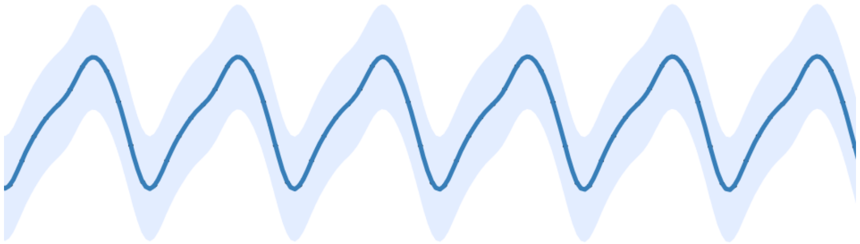
> **Rational Quadratic (RQ)**

$$K_{RQ}(\mathbf{x},\mathbf{x}') = \sigma^2\left(1+\frac{(\mathbf{x}-\mathbf{x}')^2}{2\alpha\ell^2}\right)^{-\alpha}$$

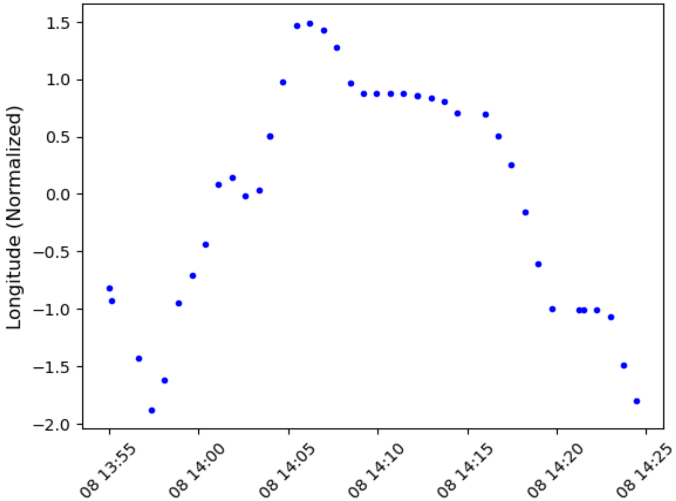
Where ℓ is a lengthscale (smoothing) parameter, σ^2 is the output variance, p is the period length, and α is the scale mixture (i.e., the relative weight of large- and small-scale variances)

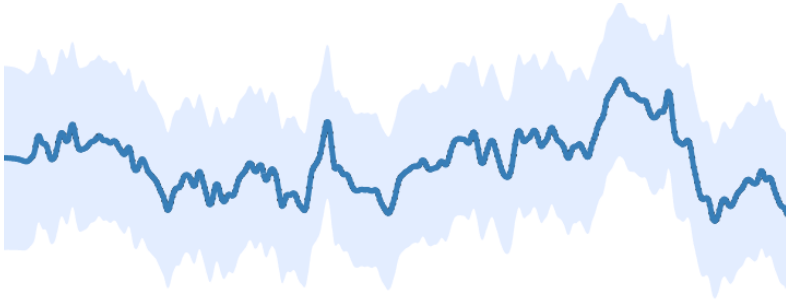
Kernels for Modeling Mobile Data





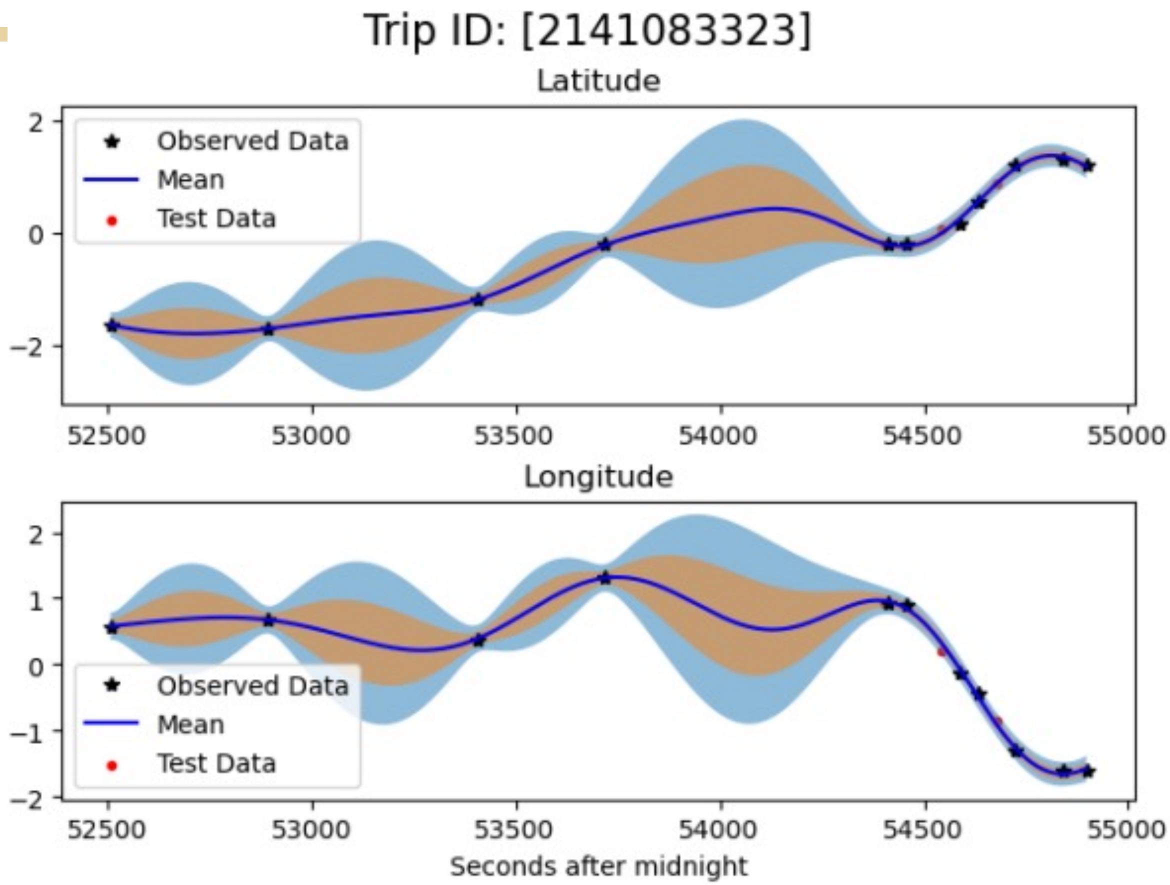
$K_{SE} \times K_{PER}$





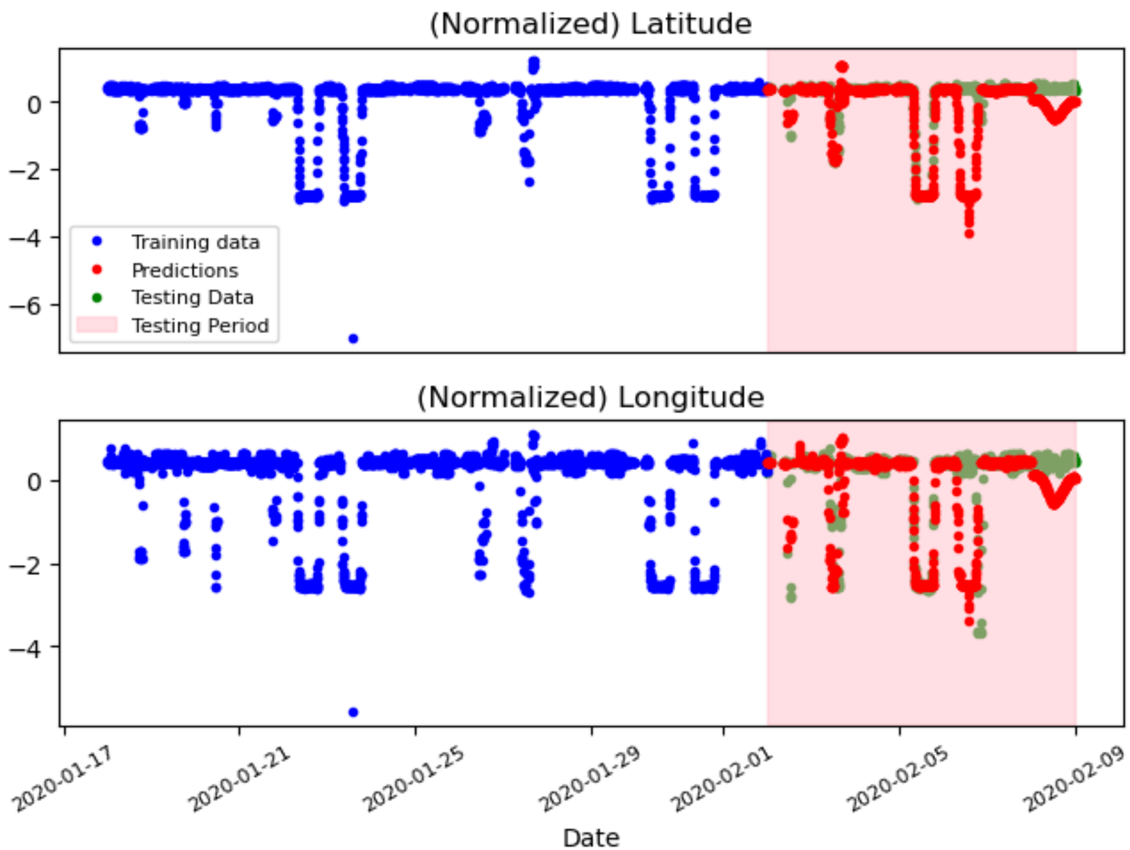
$K_{SE} \times K_{RQ}$

Short Gap Example



$$K = \prod_{d=1}^{n} K_{SE,d}$$

Long Gap Example



$$K = \prod_{d=1}^{n} K_{RQ,d} \times K_{PER,1} + \prod_{d=1}^{n} K_{RQ,d} \times K_{PER,1}$$

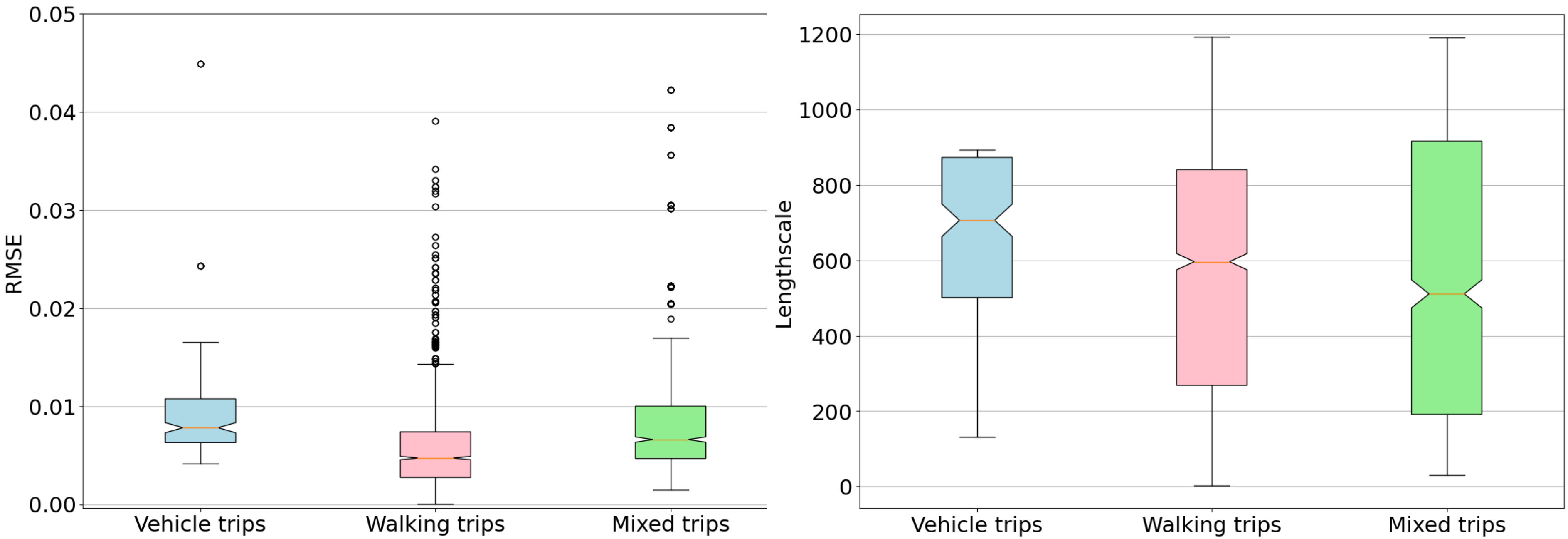
Experiments

> K-means clustering by mobility metrics

Table 2: Summary of trip clusters

Cluster	Avg. Vel. [m/s]	Trip Distance [m]	Trip Duration [s]	Heading Change Rate	Velocity Change Rate	Number of Observations	Stop Rate
Slow, short trips	9.29	8,088	1,062	0.0019	0.0024	22.79	0.0007
Medium speed, medium distance	13.94	29,693	2,362	0.0007	0.0008	49.86	0.0002
Fast, distant trips	17.86	59,299	3,449	0.0005	0.0006	141.8	0.0001

Results



(left) Box plot of total RMSE for trips in different mobility metric clusters; (right) Box plot of optimal lengthscale for trips in different mobility metric clusters

Ongoing Work and Future Direction

> Benchmarking

> Map-matching

> Sparse GPs (to improve scalability)

> Leveraging collective data

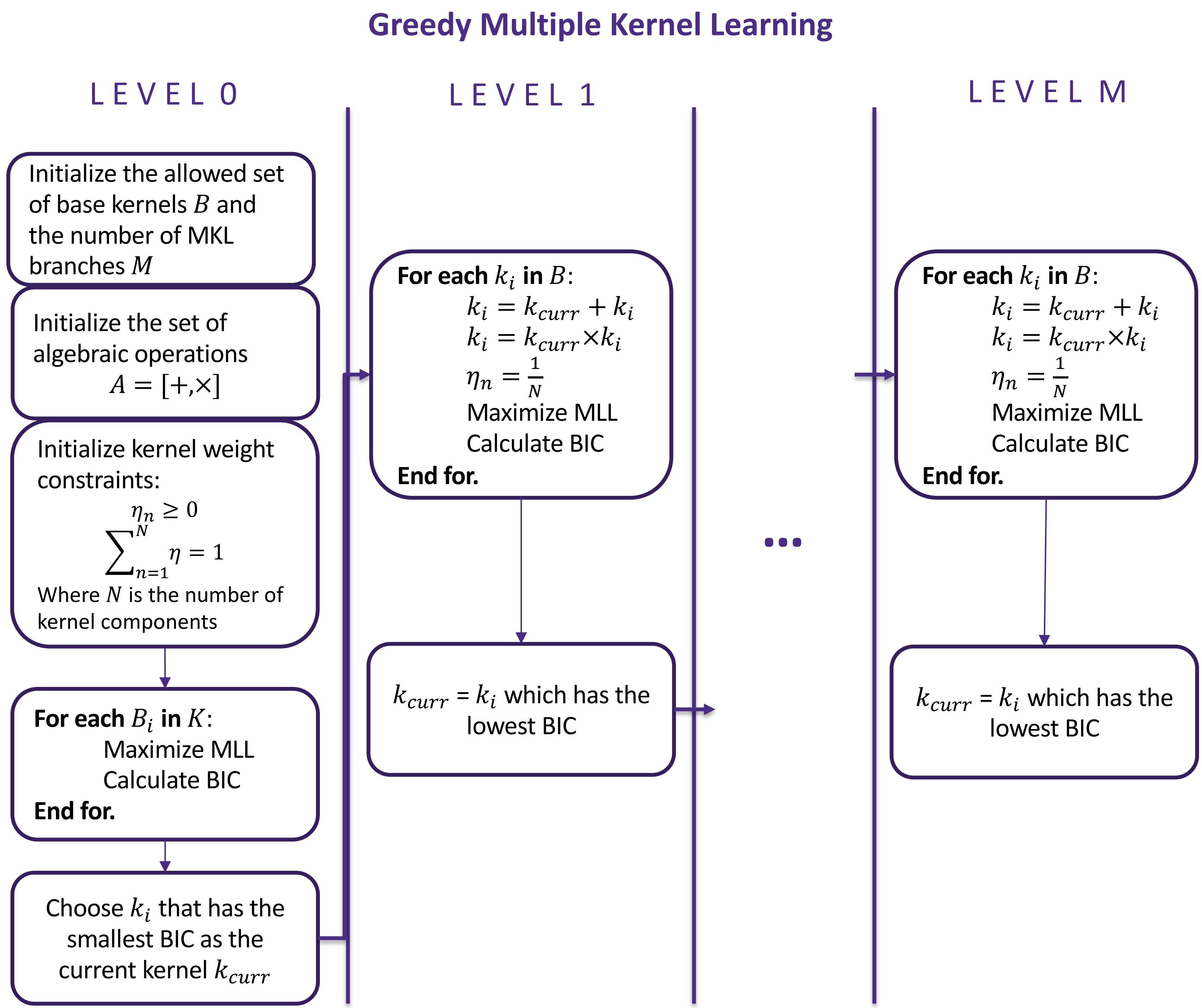
Acknowledgements

The authors are grateful to the funding support from the Center for Teaching Old Models New Tricks (TOMNET), a University Transportation Center sponsored by the US Department of Transportation through Grant No. 69A3551747116 and from the National Science Foundation for the project titled as "A whole-community effort to understand biases and uncertainties in using emerging big data for mobility analysis" (award number 2114260).

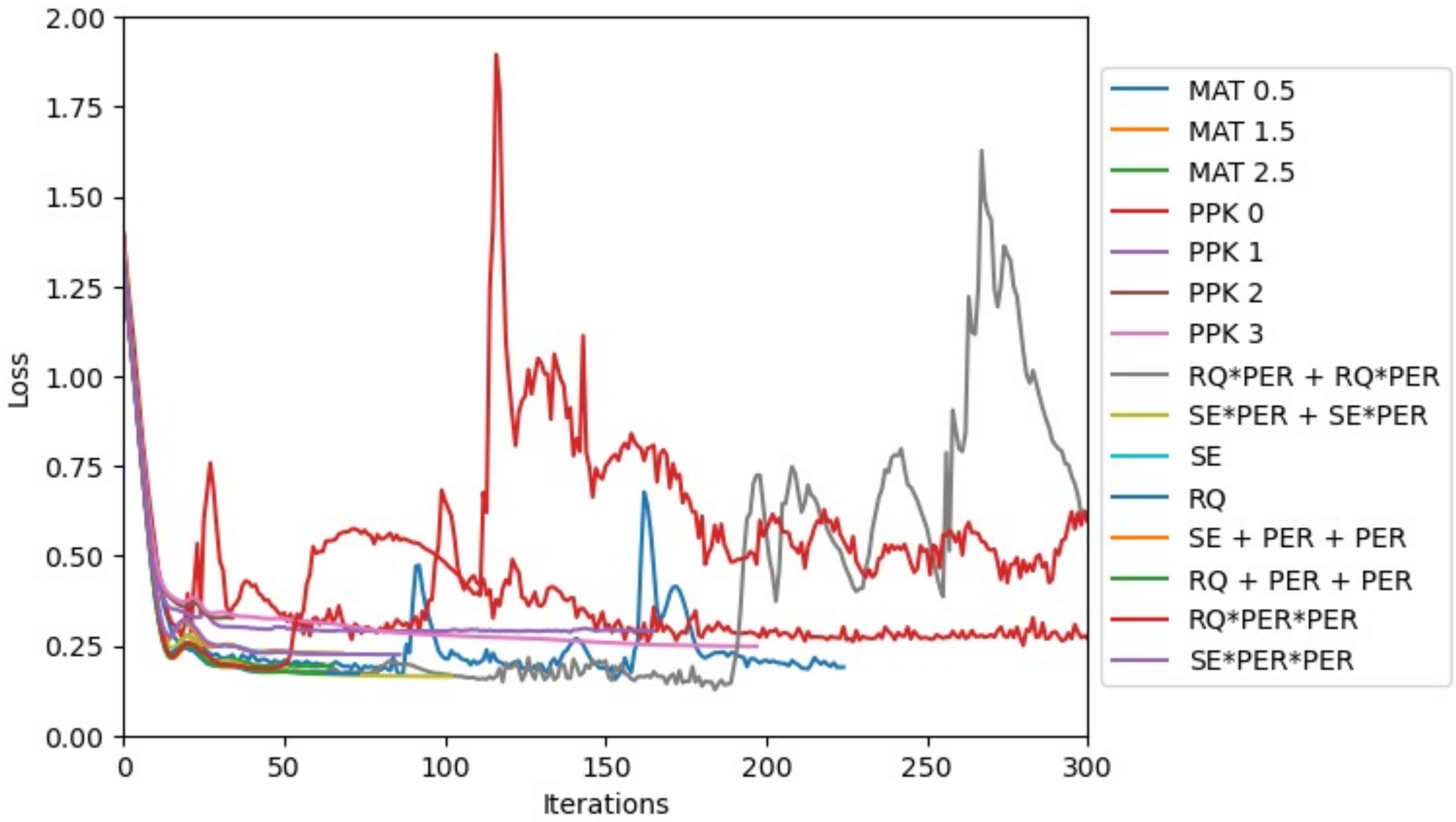
References

- Wang, F., J. Wang, J. Cao, C. Chen, and X. (Jeff) Ban. Extracting Trips from Multi-Sourced Data for Mobility Pattern Analysis: An App-Based Data Example. *Transportation Research Part C: Emerging Technologies*, Vol. 105, 2019, pp. 183–202.https://doi.org/10.1016/j.trc.2019.05.028.
- Guan, X., C. Chen, I. Ren, K. Y. Yeung, L.-H. Hung, and W. J. Lloyd. Mobility Analysis Workflow (MAW): An Accessible, Interoperable, and Reproducible Container System for Processing Raw Mobile Data. *arXiv:2204.09125 [cs, math, stat]*, 2022.
- McCool, D., P. Lugtig, and B. Schouten. Maximum Interpolable Gap Length in Missing Smartphone-Based GPS Mobility Data. *Transportation*, 2022. https://doi.org/10.1007/s11116-022-10328-2.
- Spectus - Data Clean Room for Human Mobility Analysis. Spectus.ai.
- Rasmussen, C. E., & Williams, C. K. I. Gaussian processes for machine learning. MIT Press, 2006.
- Gardner, J., G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. No. 31, 2018.
- Kingma, D. P., and J. Ba. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980. Accessed Jul. 9, 2022.

Appendix:
Multiple Kernel Learning (MKL)


Greedy Multiple Kernel Learning
LEVEL 0
LEVEL 1
LEVEL M
Initialize the allowed set of base kernels B and the number of MKL branches M
Initialize the set of algebraic operations $A = [+,\times]$
Initialize kernel weight constraints: $\eta_n \geq 0$ $\sum_{n=1}^{N} \eta = 1$ Where N is the number of kernel components
For each B_i in K: Maximize MLL Calculate BIC End for.
Choose k_i that has the smallest BIC as the current kernel k_{curr}
For each k_i in B: $k_i = k_{curr} + k_i$ $k_i = k_{curr} \times k_i$ $\eta_n = \frac{1}{N}$ Maximize MLL Calculate BIC End for.
k_{curr} = k_i which has the lowest BIC
...
For each k_i in B: $k_i = k_{curr} + k_i$ $k_i = k_{curr} \times k_i$ $\eta_n = \frac{1}{N}$ Maximize MLL Calculate BIC End for.
k_{curr} = k_i which has the lowest BIC

Different composite kernels showcase varying convergence behavior



Example MKL progression progression