Efficient Modeling of Latent Information in Supervised Learning using Gaussian Processes

Zhenwen Dai Mauricio A. Álvarez Neil D. Lawrence

Gaussian Process Approximation Workshop, 2017

Motivation

- Machine learning has been very successful in providing tools for learning a function mapping from an input to an output.

$$y = f(x) + \epsilon$$

- The modeling in terms of function mapping assumes a one/many to one mapping between input and output.
- In other words, ideally the input should contain sufficient information to uniquely determine/disambiguise the output apart from some sensory noise.

Data: a Combination of Multiple Scenarios

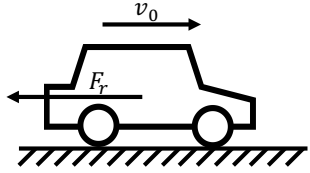
- In most of cases, this assumption does not hold.
- We often collect data as a combination of multiple scenarios, e.g., the voice recording of multiple persons, the images taken from different models of cameras.
- We only have *some labels* to identify these scenarios in our data, e.g., we can have the names of the speakers and the specifications of the used cameras.
- These labels are represented as *categorical* data in some database.

How to model these labels?

- A common practice in this case would be to ignore the difference of scenarios, but fails to model the corresponding variations.
- Model each scenario separately.
- Use a one-hot encoding.
- In both of these cases, generalization/transfer to new scenario is not possible.
- **Any better solutions?** Latent variable models!

A Toy Problem: The Braking Distance of a Car

- To model the braking distance of a car in a *completely data-driven* way.
- Input: the speed when starting to brake
- Output: the distance that the car moves before fully stopped
- We know that the braking distance depends on the friction coefficient.
- We can conduct experiments with a set of different tyre and road conditions, each associated with a condition *ID*.
- How can we model the relation between the speed and distance in a data-driven way, so that we can extrapolate to a new condition with *only one experiment*?

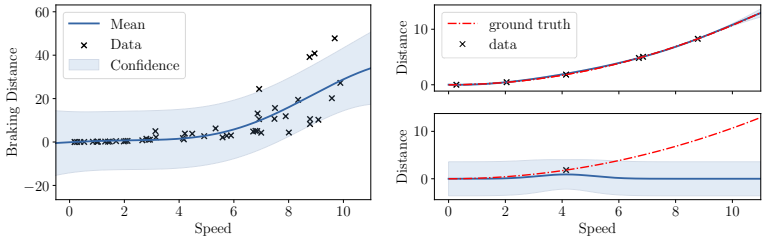


Common Modeling Choices with Non-parametric Regression

- A straight-forward modeling choice to ignore the difference in conditions. The relation between the speed and distance can be modeled as

$$y = f(x) + \epsilon, \quad f \sim GP,$$

- Alternatively, we can model each condition separately, i.e., $f_d \sim GP, d = 1, \ldots, D$.

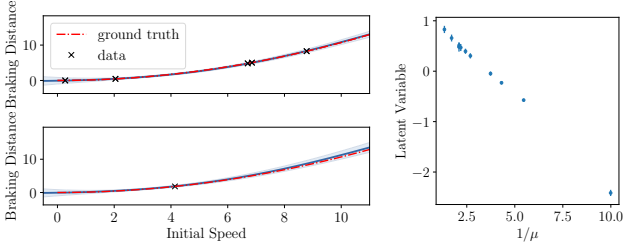


Modeling the Conditions Jointly

- A probabilistic approach is to assume a latent variable.
- With a latent variable $\mathbf{h}_d$, the relation between speed and distance for the condition d is, then, modeled as

$$y = f(x, \mathbf{h}_d) + \epsilon, \quad f \sim GP, \quad \mathbf{h}_d \sim \mathcal{N}(0, \mathbf{I}). \tag{1}$$

- A special Bayesian GPLVM?
  - Efficiency, $O(N^3D^3)$ or $O(NDM^2)$.
  - The balance among different conditions in inference.



Latent Variable Multiple Output Gaussian Processes (LVMOGP)

- We propose a new model which assumes the covariance matrix can be decomposed as a Kronecker product of the covariance matrix of the latent variables $\mathbf{K}^H$ and the covariance matrix of the inputs $\mathbf{K}^X$.
- The probabilistic distributions of LVMOGP is defined as

$$p(\mathbf{Y}_{:}|\mathbf{F}_{:}) = \mathcal{N}\left(\mathbf{Y}_{:}|\mathbf{F}_{:}, \sigma^2\mathbf{I}\right), \quad p(\mathbf{F}_{:}|\mathbf{X}, \mathbf{H}) = \mathcal{N}\left(\mathbf{F}_{:}|0, \mathbf{K}^H \otimes \mathbf{K}^X\right), \tag{2}$$

  where the latent variables $\mathbf{H}$ have unit Gaussian priors, $\mathbf{h}_d \sim \mathcal{N}(0, \mathbf{I})$
- This is a special case of the model in (1).

Scalable Variational Inference

- Sparse GP approximation with $\mathbf{U} \in \mathbb{R}^{M_X \times M_H}$:

$$\log p(\mathbf{Y}|\mathbf{X}, \mathbf{H}) \geq \langle \log p(\mathbf{Y}_{:}|\mathbf{F}_{:}) \rangle_{q(\mathbf{F}|\mathbf{U})q(\mathbf{U})} + \left\langle \log \frac{p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{H})p(\mathbf{U})}{q(\mathbf{F}|\mathbf{U})q(\mathbf{U})} \right\rangle_{q(\mathbf{F}|\mathbf{U})q(\mathbf{U})}$$

- Lower bounding the marginal likelihood

$$\log p(\mathbf{Y}|\mathbf{X}) \geq \mathcal{F} - \mathrm{KL}\left(q(\mathbf{U}) \,\|\, p(\mathbf{U})\right) - \mathrm{KL}\left(q(\mathbf{H}) \,\|\, p(\mathbf{H})\right), \tag{3}$$

Closed-form Variational Lower Bound (SVI-GP)

- It is known that the optimal posterior distribution of $q(\mathbf{U})$ is a Gaussian distribution [Titsias, 2009, Matthews et al., 2016]. With an explicit Gaussian definition of $q(\mathbf{U}) = \mathcal{N}\left(\mathbf{U}|\mathbf{M}, \mathbf{\Sigma}^U\right)$, the integral in $\mathcal{F}$ has a closed-form solution:

$$\begin{aligned}\mathcal{F} = &-\frac{ND}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\mathbf{Y}_{:}^{\top}\mathbf{Y}_{:} - \frac{1}{2\sigma^2}\mathrm{Tr}\left(\mathbf{K}_{uu}^{-1}\Phi\mathbf{K}_{uu}^{-1}(\mathbf{M}_{:}\mathbf{M}_{:}^{\top} + \mathbf{\Sigma}^U)\right) \\ &+ \frac{1}{\sigma^2}\mathbf{Y}_{:}^{\top}\Psi\mathbf{K}_{uu}^{-1}\mathbf{M}_{:} - \frac{1}{2\sigma^2}\left(\psi - \mathrm{tr}\left(\mathbf{K}_{uu}^{-1}\Phi\right)\right)\end{aligned}$$

  where $\psi = \langle \mathrm{tr}\left(\mathbf{K}_{ff}\right)\rangle_{q(\mathbf{H})}$, $\Psi = \langle \mathbf{K}_{fu}\rangle_{q(\mathbf{H})}$ and $\Phi = \left\langle \mathbf{K}_{fu}^{\top}\mathbf{K}_{fu}\right\rangle_{q(\mathbf{H})}$

- The computational complexity of the closed-form solution is $O(NDM_X^2M_H^2)$.

More Efficient Formulation

- The Kronecker product decomposition of covariance matrices are not exploited.
- Firstly, the expectation computation can be decomposed,

$$\psi = \psi^H \text{tr}\left(\mathbf{K}_{ff}^X\right), \quad \Psi = \Psi^H \otimes \mathbf{K}_{fu}^X, \quad \Phi = \Phi^H \otimes \left((\mathbf{K}_{fu}^X)^\top \mathbf{K}_{fu}^X\right), \tag{4}$$

where $\psi^H = \left\langle \text{tr}\left(\mathbf{K}_{ff}^H\right)\right\rangle_{q(\mathbf{H})}$, $\Psi^H = \left\langle \mathbf{K}_{fu}^H \right\rangle_{q(\mathbf{H})}$ and $\Phi^H = \left\langle (\mathbf{K}_{fu}^H)^\top \mathbf{K}_{fu}^H \right\rangle_{q(\mathbf{H})}$.

More Efficient Formulation

- Secondly, we assume a Kronecker product decomposition of the covariance matrix of $q(\mathbf{U})$, i.e., $\Sigma^U = \Sigma^H \otimes \Sigma^X$.
- The number of variational parameters in the covariance matrix from $M_X^2 M_H^2$ to $M_X^2 + M_H^2$.
- The direct computation of Kronecker products is completely avoided.

$$
\begin{aligned}
\mathcal{F} = & -\frac{ND}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\mathbf{Y}_{:}^{\top}\mathbf{Y}_{:} \\
& - \frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{M}^{\top}((\mathbf{K}_{uu}^X)^{-1}\Phi^C(\mathbf{K}_{uu}^X)^{-1})\mathbf{M}(\mathbf{K}_{uu}^H)^{-1}\Phi^H(\mathbf{K}_{uu}^H)^{-1}\right) \\
& - \frac{1}{2\sigma^2}\mathrm{tr}\left((\mathbf{K}_{uu}^H)^{-1}\Phi^H(\mathbf{K}_{uu}^H)^{-1}\mathbf{\Sigma}^H\right)\mathrm{tr}\left((\mathbf{K}_{uu}^X)^{-1}\Phi^X(\mathbf{K}_{uu}^X)^{-1}\mathbf{\Sigma}^X\right) \\
& \ldots
\end{aligned}
$$

Prediction

- Given both a set of new inputs $\mathbf{X}^*$ with a set of new scenarios $\mathbf{H}^*$, the prediction of noiseless observation $\mathbf{F}^*$ can be computed in closed-form.

$$
\begin{aligned}
q(\mathbf{F}^*_{:}|\mathbf{X}^*, \mathbf{H}^*) &= \int p(\mathbf{F}^*_{:}|\mathbf{U}_{:}, \mathbf{X}^*, \mathbf{H}^*) q(\mathbf{U}_{:}) \mathrm{d}\mathbf{U}_{:} \\
&= \mathcal{N}\left(\mathbf{F}^*_{:}|\mathbf{K}_{f^*u}\mathbf{K}_{uu}^{-1}\mathbf{M}_{:}, \mathbf{K}_{f^*f^*} - \mathbf{K}_{f^*u}\mathbf{K}_{uu}^{-1}\mathbf{K}_{f^*u}^{\top} + \mathbf{K}_{f^*u}\mathbf{K}_{uu}^{-1}\mathbf{\Sigma}^{U}\mathbf{K}_{uu}^{-1}\mathbf{K}_{f^*u}^{\top}\right),
\end{aligned}
$$

- For a regression problem, we are often more interested in predicting for the existing condition from the training data. We can approximate the prediction by integrating the above prediction equation with $q(\mathbf{H})$,

$$
q(\mathbf{F}^*_{:}|\mathbf{X}^*) = \int q(\mathbf{F}^*_{:}|\mathbf{X}^*, \mathbf{H}) q(\mathbf{H}) \mathrm{d}\mathbf{H}.
$$

Missing Data

- The model described previously assumes that for N different inputs, we observe them in all the D different conditions.
- In real world problems, we often collect data at a different set of inputs for each scenario, i.e., for each condition d, $d = 1, \ldots, D$.
- The proposed model can be extended to handle this case by reformulating the $\mathcal{F}$ as

$$\mathcal{F} = \sum_{d=1}^{D} -\frac{N_d}{2}\log 2\pi\sigma_d^2 - \frac{1}{2\sigma_d^2}\mathbf{Y}_d^\top\mathbf{Y}_d - \frac{1}{2\sigma_d^2}\mathrm{Tr}\left(\mathbf{K}_{uu}^{-1}\Phi_d\mathbf{K}_{uu}^{-1}(\mathbf{M}_{:}\mathbf{M}_{:}^\top + \mathbf{\Sigma}^U)\right)$$
$$+ \frac{1}{\sigma_d^2}\mathbf{Y}_d^\top\Psi_d\mathbf{K}_{uu}^{-1}\mathbf{M}_{:} - \frac{1}{2\sigma_d^2}\left(\psi_d - \mathrm{tr}\left(\mathbf{K}_{uu}^{-1}\Phi_d\right)\right),$$

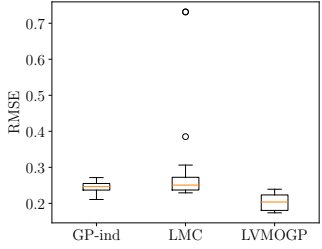
where $\Phi_d = \Phi_d^H \otimes \left((\mathbf{K}_{f_d u}^X)^\top\mathbf{K}_{f_d u}^X\right)$, $\Psi_d = \Psi_d^H \otimes \mathbf{K}_{f_d u}^X$, $\psi_d = \psi_d^H \otimes \mathrm{tr}\left(\mathbf{K}_{f_d f_d}^X\right)$

Related Works

- Multiple Output Gaussian Processes /Multi-task Gaussian proccesses: Ivarez et al. [2012] [Goovaerts, 1997] [Bonilla et al., 2008]
- Our method reduces computationally complexity to $O(\max(N, M_H) \max(D, M_X) \max(M_X, M_H))$ when there are no missing data.
- An additional advantage of our method is that it can easily be parallelized using mini-batches like in [Hensman et al., 2013].
- The idea of modeling latent information about different conditions jointly with the modeling of data points is related to the style and content model by Tenenbaum and Freeman [2000].

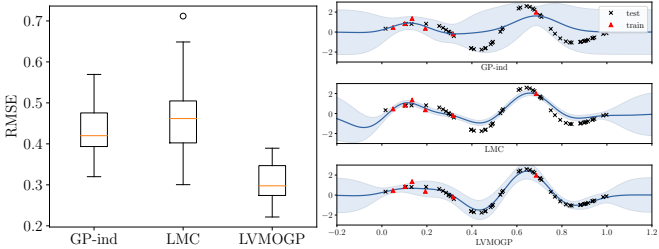
Experiments on Synthetic Data

- 100 different uniformly sampled input locations (50 for training and 50 for testing), where each corresponds to 40 different conditions. An observation noise with variance 0.3 is added onto the training data
- We compare LVMOGP with two other methods: GP with independent output dimensions (GP-ind) and LMC (with a full rank coregionalization matrix).
- First dataset without missing data.



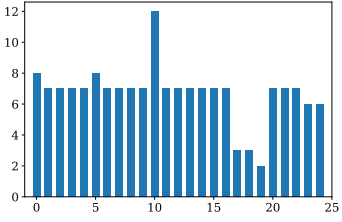
Experiments on Synthetic Data with Missing Data

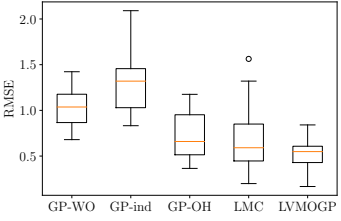
- To generate a dataset with uneven numbers of training data in different conditions, we group the conditions into 10 groups. Within each group, the numbers of training data in four conditions are generated through a three-step stick breaking procedure with a uniform prior distribution (200 data points in total).
- We compare LVMOGP with two other methods: GP with independent output dimensions (GP-ind) and LMC (with a full rank coregionalization matrix).
- GP-ind: 0.43 ± 0.06, LMC:0.47 ± 0.09, LVMOGP 0.30 ± 0.04

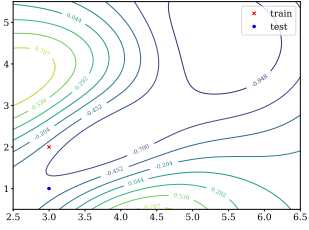


Experiment on Servo Data

- We apply our method to a servo modeling problem, in which the task to predict the rise time of a servomechanism in terms of two (continuous) gain settings and two (discrete) choices of mechanical linkages [Quinlan, 1992].
- The two choices of mechanical linkages: 5 types of motors and 5 types of lead screws.
- We take 70% of the dataset as training data and the rest as test data, and randomly generated 20 partitions.
- GP-WO: 1.03 ± 0.20, GP-ind: 1.30 ± 0.31, GP-OH: 0.73 ± 0.26, LMC:0.69 ± 0.35, LVMOGP 0.52 ± 0.16

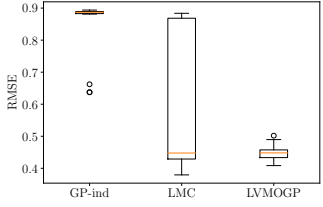






Experiment on Sensor Imputation

- We apply our method to impute multivariate time series data with massive missing data. We take a in-house multi-sensor recordings including a list of sensor measurements such as temperature, carbon dioxide, humidity, etc. [Zamora-Martnez et al., 2014].
- The measurements are recorded every minutes for roughly a month and smoothed with 15 minute means.
- We mimic the scenario of massive missing data by randomly taking out 95% of the data entries and aim at imputing all the missing values.
- GP-ind: 0.85 ± 0.09, LMC:0.59 ± 0.21, LVMOGP 0.45 ± 0.02



Conclusion

- The common practices such as one-hot encoding cannot efficiently model the relation among different conditions and are not able to generalize to a new condition at test time.
- We propose to solve this problem in a principled way, where we learn the latent information of conditions into a latent space as part of the regression model.
- By exploiting the Kronecker product decomposition in the variational posterior, our inference method are able to achieve the same computational complexity as sparse GP with independent observations.
- As shown repeatedly in the experiments, the Bayesian inference of the latent variables in LVMOGP avoids the overfitting problem in LMC.

Reference

Edwin V. Bonilla, Kian Ming Chai, and Christopher K. I. Williams. Multi-task Gaussian process prediction. In John C. Platt, Daphne Koller, Yoram Singer, and Sam Roweis, editors, *NIPS*, volume 20, 2008.

Pierre Goovaerts. *Geostatistics For Natural Resources Evaluation*. Oxford University Press, 1997.

James Hensman, Nicolo Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *UAI*, 2013.

Alexander G. D. G. Matthews, James Hensman, Richard E Turner, and Zoubin Ghahramani. On sparse variational methods and the kullback-leibler divergence between stochastic processes. In *AISTATS*, 2016.

J R Quinlan. Learning with continuous classes. In *Australian Joint Conference on Artificial Intelligence*, pages 343–348, 1992.

JB Tenenbaum and WT Freeman. Separating style and content with bilinear models. *Neural Computation*, 12:1473–83, 2000.

Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *AISTATS*, 2009.

F. Zamora-Martnez, P. Romeu, P. Botella-Rocamora, and J. Pardo. On-line learning of indoor temperature forecasting models towards energy efficiency. *Energy and Buildings*, 83:162–172, 2014.

Mauricio A. Ivarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012. ISSN 1935-8237. doi: 10.1561/2200000036. URL `http://dx.doi.org/10.1561/2200000036`.