PSEUDO INPUTS FOR PAIRWISE LEARNING WITH GAUSSIAN PROCESSES

Jens Brehm Nielsen[1,2], Bjørn Sand Jensen[1] & Jan Larsen[1]
[1]{jenb,bjje,jl}@imm.dtu.dk, Technical University of Denmark, [2]jeb@widex.com, Widex A/S

Abstract

We consider learning and prediction of pairwise comparisons between instances. The problem is motivated from a perceptual view point, where pairwise comparisons serve as an effective and extensively used paradigm. A state-of-the-art method for modeling pairwise data in high dimensional domains is based on a classical pairwise probit likelihood imposed with a Gaussian process prior. While extremely flexible, this non-parametric method struggles with an inconvenient $\mathcal{O}\left(n^3\right)$ scaling in terms of the n input instances which limits the method only to smaller problems. To overcome this, we derive a specific sparse extension of the classical pairwise likelihood using the pseudo-input formulation. The behavior of the proposed extension is demonstrated on a toy example and on two real-world data sets which outlines the potential gain and pitfalls of the approach. Finally, we discuss the relation to other similar approximations that have been applied in standard Gaussian process regression and classification problems such as fully independent (training) conditional and partially independent (training) conditional.

Methods

Introduction

We focus on pairwise comparisons modeled by the likelihood function considered in [1, 2], extended with Gaussian processes (GP) priors in [3]. GP based models struggle with an inconvenient $\mathcal{O}\left(n^3\right)$ scaling. Different suggestions have been proposed to remedy this issue for standard GP regression [4, 5, 6, 7, 8, 9, 10] with the most well-known methods being the FI(T)C or PI(T)C, resulting in a (sparse) conditional GP prior. These methods can probably be adapted to the pairwise model, however,

Our contribution is to introduce sparsity in the pairwise model [3] starting from the original pseudo-input formulation [6] using a set of l inducing points, where $l << n$, resulting in an $\mathcal{O}\left(\min\left\{mn^2, l^3\right\}\right)$ scaling.

Sparse Pairwise GP Setup

- Inputs: $\mathcal{X} = \{\mathbf{x}_i | i = 1, \ldots, n\}$, where $\mathbf{x}_i \in \mathbb{R}^d$
- Data: $\mathcal{Y} = \{y_k; u_k, v_k | k = 1, ..., m\}$, where $u_k \neq v_k$ and $\mathbf{x}_{u_k}, \mathbf{x}_{v_k} \in \mathcal{X}$
- Latent function: $f(\mathbf{x}_i) \sim \mathcal{GP}\left(0, k(\mathbf{x}_i, \cdot)\right)$
- Likelihood: $p\left(y_k | \mathbf{f}_k, \boldsymbol{\theta}_{\mathcal{L}}\right) = \Phi\left(y_k \frac{f(\mathbf{x}_{u_k}) - f(\mathbf{x}_{v_k})}{\sqrt{2}\sigma}\right)$, $\mathbf{f}_k = \left[f\left(\mathbf{x}_{u_k}\right), f\left(\mathbf{x}_{v_k}\right)\right]^\top$

Pairwise likelihood function transformation using the l pseudo inputs, $\bar{\mathbf{X}}$

$$p\left(\bar{\mathbf{f}} | \bar{\mathbf{X}}\right) = \mathcal{N}\left(\bar{\mathbf{f}} | \mathbf{0}, \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}\right) \rightarrow \begin{bmatrix} \mathbf{f}_k \\ \bar{\mathbf{f}} \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\mathbf{x}_k\mathbf{x}_k} & \left[\mathbf{k}_{u_k}, \mathbf{k}_{v_k}\right]^\top \\ \left[\mathbf{k}_{u_k}, \mathbf{k}_{v_k}\right] & \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}} \end{bmatrix}\right)$$

$$p\left(y_k | \mathbf{x}_{u_k}, \mathbf{x}_{v_k}, \bar{\mathbf{X}}, \bar{\mathbf{f}}\right) = \int p\left(y_k | \mathbf{f}_k\right) p\left(\mathbf{f}_k | \bar{\mathbf{f}}, \bar{\mathbf{X}}\right) d\mathbf{f}_k = \Phi\left(\frac{y_k \left(\mathbf{k}_{u_k}^\top - \mathbf{k}_{v_k}^\top\right) \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1} \bar{\mathbf{f}}}{\sigma_k^*}\right)$$

Model

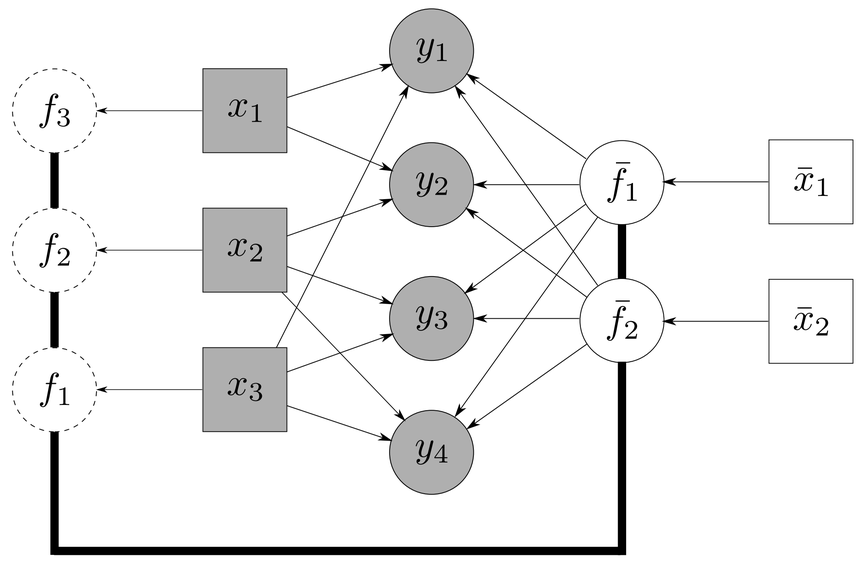
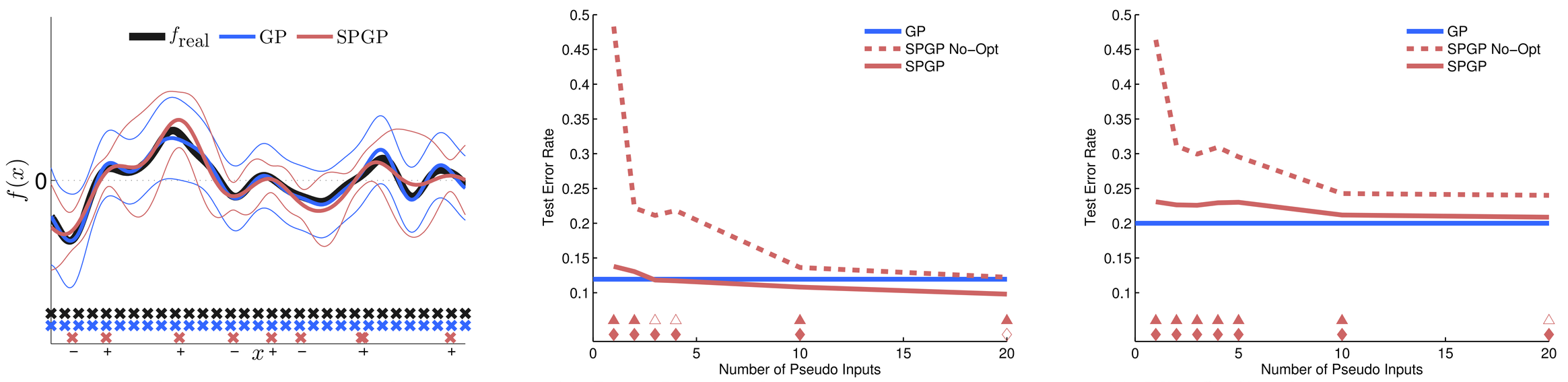


Fig. 1: *Graphical chain model illustration of the sparse, pairwise GP model. Grey solid: observed variables. White solid: unknown variables to be inferred. White dashed: unknown variables not to be inferred directly. Square solid: inputs*

Inference

As for the standard pairwise likelihood model, exact inference is intractable. Instead, the Laplace approximation is used for posterior approximation and predictions. The hyperparameters ($\{\boldsymbol{\theta}, \bar{\mathbf{X}}\}$) are optimized using gradient based MAP-II maximization using a BFGS method with k-means initialization.

Simulation Results



(a): *Toy:* $d = 1$, $n = 31$, $m = 465$, $l = 9$ **(b):** *Boston Housing:* $d = 10$, $n = 506$, $m = 127765$ **(c):** *Wine Quality:* $d = 11$, $n = 600$, $m = 179700$

Fig. 2: *In general, blue graphs indicate the full model (GP) and red indicate the sparse model (SPGP). In **Fig. (a)** thick graphs indicate means and thin graphs indicate one standard deviation. The black graph indicates the real (deterministic) function used to generate the full pairwise data set between the instances marked with black crosses in the bottom. The two other colors sketch the predictive distribution of the GP and SPGP models using the (pseudo) inputs at the locations marked with the corresponding color in the bottom. **Fig. (b)-(c)** display the performance of the SPGP model evaluated on two real-world data sets as a function of the number of pseudo inputs for the sparse model (red). The performance of the standard model is included as a baseline. The **solid** and **dashed** red graphs show the average test error rate for the optimized and non-optimized SPGP model, respectively. The two rows of markers indicate whether the optimized (triangle) and non-optimized (diamond) SPGP models are significant different from the GP model using the McNemar test. The markers are solid if the null hypothesis that they are equal can be rejected at the 5% significance level.*

References

[1] L. L. Thurstone, "A Law of Comparative Judgement," *Psychological Review*, vol. 34, 1927.
[2] R. D. Bock and J. V. Jones, "The Measurement and Prediction of Judgment and Choice," 1968.
[3] W. Chu and Z. Ghahramani, "Preference Learning with Gaussian Processes," *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 137–144, 2005.
[4] N. Lawrence, M. Seeger, and R. Herbrich, "Fast Sparse Gaussian Process Methods: The Informative Vector Machine," in *Neural Information Processing Systems (NIPS)*, 2002, p. 8.
[5] L. Csato, *Gaussian Processes - Iterative Sparse Approximations*, Ph.D. thesis, Aston University, 2002.
[6] E. Snelson and Z. Ghahramani, "Sparse Gaussian Processes using Pseudo-Inputs," *Advances in neural information processing*, 2006.
[7] C. Walder, K. I. Kim, and B. Schölkopf, "Sparse Multiscale Gaussian Process Regression," in *Proceedings of the 25th international conference on Machine Learning*, 2008, pp. 1112–1119.
[8] M. Lazaro-Gredilla and A. Figueiras-Vidal, "Inter-Domain Gaussian Processes for Sparse Inference using Inducing Features," in *Advances in Neural Information Processing Systems 22*, pp. 1087–1095. 2009.
[9] J. Quiñonero-Candela and C.E. Rasmussen, "A Unifying View of Sparse Approximate Gaussian Process Regression," *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.
[10] E. Snelson and Z. Ghahramani, "Local and Global Sparse Gaussian Process Approximations," in *Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS*, 2007, pp. 524–531.

DTU Informatics
Department of Informatics and Mathematical Modelling

Summary

- We derived a **pairwise** pseudo-input formulation of the pairwise likelihood model with GPs.
- Toy example: Predictive mean is well modeled with $l = 9$ pseudo inputs, but—as expected—the predictive variance differs between the full and sparse model.
- Real-world examples:
  - Few (optimized) pseudo inputs perform only slightly worse than the full model suggesting that the two real-world data sets do not constitute complex pairwise relations. The performance do, however, depend highly on the optimization of the pseudo inputs as seen from the non-optimized models.
  - Further adding pseudo inputs to the sparse model can result in better performance than the full model (Boston Housing). Thus, by using a less complex—but optimized—approximate prior provides better regularization leading to lower generalization error.
- Future work includes comparison with other approximations like FI(T)C and PI(T)C, and other inference methods such as expectation propagation.