
Probabilistic structure discovery in time series data

David Janz

University of Cambridge
dj343@cam.ac.uk

Brooks Paige

University of Oxford
brooks@robots.ox.ac.uk

Tom Rainforth

University of Oxford
twgr@robots.ox.ac.uk

Jan-Willem van de Meent

Northeastern University
j.vandemeent@northeastern.edu

Frank Wood

University of Oxford
fwood@robots.ox.ac.uk

Abstract

Existing methods for structure discovery in time series data construct interpretable, compositional kernels for Gaussian process regression models. While the learned Gaussian process model provides posterior mean and variance estimates, typically the structure is learned via a greedy optimization procedure. This restricts the space of possible solutions and leads to over-confident uncertainty estimates. We introduce a fully Bayesian approach, inferring a full posterior over structures, which more reliably captures the uncertainty of the model.

1 Introduction

How much of the data science process can we automate? Many techniques such as random forests [1], neural networks [12] and support vector machines [4] can deliver outstanding predictive performance when used in an out-of-the-box fashion [7, 15], while the rise of so-called automated machine learning [8] is providing increased automation of the full data-science pipeline. However, such methods tend to focus only on predictive performance, giving crude estimates for uncertainty and providing negligible human insight into the data generation process. Such approaches are not, in isolation, sufficient in the strife for automated data science.

This paper focuses on automatic methods for model construction and selection in the context of models built up compositionally from simple atomic parts [6, 10]. These models can be expressive and powerful while still maintaining human interpretability, a key characteristic for automatic ML systems that are to be trusted by end users. Specifically, our method builds upon automatic Bayesian covariance discovery [13] (ABCD), which employs search over the kernel structure and parameters of a Gaussian process [16] (GP) for regressing time series data.

We suppose that some time series data $\mathcal{D} = \{(y_n, x_n) : n = 1, \dots, N\}$ was generated by a latent function f with additive Gaussian noise. We consider a GP prior $f \sim \mathcal{GP}(0, k_\theta)$ for some positive semi-definite kernel function $k_\theta(x, x')$ parametrised by a vector of hyperparameters $\theta = (\theta_1, \dots, \theta_P)^\top$. Noting that the set of positive semi-definite functions \mathbf{S}^+ has the property $z, z' \in \mathbf{S}^+$ implies $z + z' \in \mathbf{S}^+$ and $z \times z' \in \mathbf{S}^+$ [16], we define a space of kernels as the space of hierarchical products and sums of base kernels $b \in \mathcal{B}$. Here \mathcal{B} consists of the linear (LIN), squared exponential (SE), periodic (PER) and rational quadratic (RQ) kernels. ABCD selects a single k_θ using a greedy optimisation based search that maximises the marginal likelihood of the model.

We expand on this by integrating out k_θ . In particular, we explore inferring the distribution over possible kernel compositions, to provide better a better uncertainty estimate and allow exploration of the multiple modes in the model space. Integrating out k_θ has previously been attempted by Schaechtle et al. [17] in a restricted setting, where set of possible kernel combinations is finite and heavily limited; we extend this to use a true stochastic context-free grammar as a prior on kernel structure. Our approach is implemented within the probabilistic programming framework of Anglican [18], allowing integration with other systems in a probabilistic modelling pipeline.

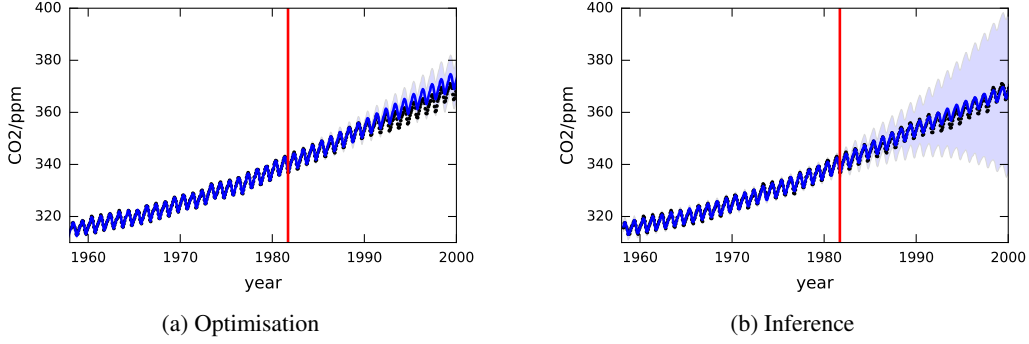


Figure 1: Regression of the Maunu Lao CO₂ dataset [16]. Models are trained on data from 1957 through 1983 (red line); we compare the extrapolation to collected data from 1984 through 2000. Median plotted, 10-90 percentile region shaded. (1a) The learned best-fit model and model parameters using ABCD fits the training data closely, but yields overconfident predictions in the extrapolation. (1b) Averaging over multiple plausible models yields better-calibrated uncertainty estimates.

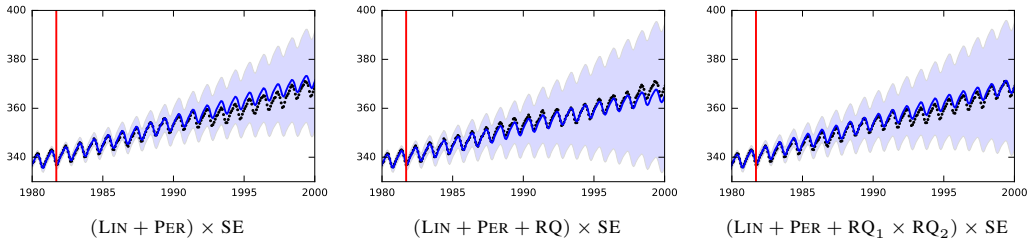


Figure 2: Three particles generated in our approach demonstrate how competing explanations, differing both in kernel structure and kernel parameters, combine to produce posterior estimates.

2 Model & inference

Given a prior distribution over the kernel $p(k, \theta) = p(k)p(\theta|k)$, where k denotes the composite kernel structure and θ the kernel parameters, and assuming a Gaussian process prior $p(f|\mathcal{D}, k, \theta)$, the posterior for predictions y^* at x^* can be written as

$$p(y^*|x^*, \mathcal{D}) = \iint p(k, \theta) \left[\int p(y^*|x^*, f)p(f|\mathcal{D}, k, \theta) df \right] dk d\theta. \quad (1)$$

Taking $p(y^*|x^*, f)$ to be a Gaussian likelihood, the integral over f can be computed in closed form. The key task is thus to define a suitable kernel prior and perform integration over k and θ . Using a sampling based system for this integration produces a weighted mixture of GPs posterior, retaining an analytic predictive distribution.

We define $p(k)$ using a probabilistic context free grammar [14] with production rules $R = \{s \rightarrow s + s | s \times s | b \in \mathcal{B}\}$, where s is a non-terminal symbol and b is a base kernel, and fixed context free rule probabilities $\{p_r : r \in R\}$. We use a separable prior on θ , $p(\theta|k) = \prod_{i=1}^P p(\theta_i|k)$, with each term determined by type of the hyperparameter θ_i and its corresponding base kernel in k . Jointly, these define a generative model for k_θ that we can sample from.

We marginalise over k and θ using a population Monte Carlo method [2] inspired by the use of Monte Carlo methods in decision tree inference [3, 11]. We sample initial particles (k^j, θ^j) for $j = 1, \dots, M$ from the joint prior and repeatedly mutate them using a ‘resample-move’ algorithm [9]. At each iteration, we propose a mutation to each particle, using both

1. changing the kernel structure by sampling an existing component base kernel b from the composite kernel k and replacing it using one of: $b + b'$, $b \times b'$, b or \emptyset , where b' is a new randomly chosen base kernel and \emptyset indicates removing b from k ; and
2. applying multiple Hamiltonian Monte Carlo (HMC) transitions to θ , given the composite kernel k .

This produces a new, modified set of particles drawn from a proposal distribution which can be described as a mixture of Markov kernels; with appropriate weights [5], we can perform a resampling step to construct a new particle set which again approximates the posterior. Posterior predictive comparisons to ABCD are shown in figure 1 and example individual particles are shown in figure 2.

3 Discussion

We have developed a fully probabilistic approach to structure discovery in time series and implemented it in Anglican. Through capturing the uncertainty in both the structure of the kernel and the hyperparameters, the system provides better calibrated uncertainty estimates than previous greedy optimisation approaches. The system allows for the posterior distribution over models to be inspected to get an understanding of how and why the system has made its predictions; one can think of the produced samples as providing different possible explanations for the data generation. Thus, the system provides not only a powerful regressor in a rich class of models, but also interpretable output.

References

- [1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [2] O. Cappé, A. Guillin, J.-M. Marin, and C. P. Robert. Population monte carlo. *Journal of Computational and Graphical Statistics*, 2012.
- [3] H. A. Chipman, E. I. George, and R. E. McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.
- [4] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [5] P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.
- [6] D. Duvenaud, J. Lloyd, R. Grosse, J. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. *ICML*, 30:1166–1174, 2013.
- [7] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems. *JMLR*, 15(1):3133–3181, 2014.
- [8] M. Feurer, A. Klein, K. Eggenberger, J. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. In *NIPS*, 2015.
- [9] W. R. Gilks and C. Berzuini. Following a moving target—monte carlo inference for dynamic bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):127–146, 2001.
- [10] R. B. Grosse. *Model selection in compositional spaces*. PhD thesis, Massachusetts Institute of Technology, 2014.
- [11] B. Lakshminarayanan, D. Roy, and Y. W. Teh. Top-down particle filtering for Bayesian decision trees. In *ICML*, 2013.
- [12] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [13] J. R. Lloyd, D. Duvenaud, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *AAAI*, 2014.
- [14] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.
- [15] T. Rainforth and F. Wood. Canonical correlation forests. *arXiv preprint arXiv:1507.05444*, 2015.
- [16] C. E. Rasmussen. *Gaussian processes for machine learning*. MIT Press, 2006.
- [17] U. Schaehtle, B. Zinberg, A. Radul, K. Stathis, and V. K. Mansinghka. Probabilistic Programming with Gaussian Process Memoization. *arXiv preprint arXiv:1512.05665*, 2015.
- [18] F. Wood, J. W. van de Meent, and V. Mansinghka. A new approach to probabilistic programming inference. In *AISTATS*, pages 1024–1032, 2014.