Active Structure Discovery for Gaussian Processes

Gustavo Malkomes and **Roman Garnett** {LUIZGUSTAVO, GARNETT}@WUSTL.EDU
Department of Computer Science and Engineering
Washington University in St. Louis, St. Louis, MO 63130, United States

Abstract

We introduce a novel information-theoretic approach for active model selection. Although our method can work with arbitrary models, we focus on actively learning the appropriate structure for Gaussian process regression. We then apply this framework to active structure discovery. Our method does not require model retraining to evaluate candidate points, making it more feasible than previous approaches.

Keywords: Gaussian processes, active learning, model selection, structure learning.

1. Introduction

Over the last two decades, there has been much interest in kernel-based nonparametric models, of which Gaussian processes and support vector machines are the two most-famous representatives. An important decision for the effectiveness of these methods is the choice of the kernel structure, which often requires a considerable amount of expertise.

Recent works, however, have developed automatic methods for choosing a model structure (Grosse et al., 2012; Duvenaud et al., 2013). In particular, the latter work proposed an automatic method for finding kernels to explain a fixed dataset. Here, we extend their work to the active setting by designing a novel active-model-selection method based on maximizing the mutual information between the output variable and the model class.

Comparatively, Lloyd et al. (2014) described an *automatic statistician* that is able to automatically infer an appropriate Gaussian process model given a static, fixed dataset to be analyzed; by analogy, our contributions here essentially construct an automatic *experimenter,* who is able to gather new data intelligently so as to learn the appropriate model quickly[1].

2. Bayesian model selection

We consider supervised learning problems defined on an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Suppose we are given a set of observed data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, where $\mathbf{X}$ represents the design matrix of independent variables $\mathbf{x}_i \in \mathcal{X}$ and $\mathbf{y}$ the associated vector of dependent variables $y_i = y(\mathbf{x}_i) \in \mathcal{Y}$.

Let $\mathcal{M}$ be a probabilistic model, and let θ be an element of the parameter space indexing $\mathcal{M}$. Given a set of observations $\mathcal{D}$, we wish to compute the probability of $\mathcal{M}$ being the correct model to explain $\mathcal{D}$, compared to other models. The key quantity of interest to

1. A longer version of this manuscript is currently under review for formal publication.

model selection is the *model evidence*:

$$p(\mathbf{y} \mid \mathbf{X}, \mathcal{M}) = \int p(\mathbf{y} \mid \mathbf{X}, \theta, \mathcal{M}) p(\theta \mid \mathcal{M}) \, \mathrm{d}\theta, \tag{1}$$

which represents the probability of having generating the observed data under the model, marginalized over θ to account for all possible members of that model under a prior $p(\theta \mid \mathcal{M})$ (MacKay, 2003). Given a set of M candidate models $\{\mathcal{M}_i\}_{i=1}^{M}$, and the computed evidence for each, we can apply Bayes' rule to compute the posterior probability of each model given the data:

$$p(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \mathcal{M}) p(\mathcal{M})}{p(\mathbf{y} \mid \mathbf{X})} = \frac{p(\mathbf{y} \mid \mathbf{X}, \mathcal{M}) p(\mathcal{M})}{\sum_i p(\mathbf{y} \mid \mathbf{X}, \mathcal{M}_i) p(\mathcal{M}_i)}, \tag{2}$$

where $p(\mathcal{M})$ represents the prior probability distribution over the models.

2.1. Active Bayesian model selection

Suppose that we have a mechanism for actively selecting new data—choosing $\mathbf{x}^* \in \mathcal{X}$ and observing $y^* = y(\mathbf{x}^*)$—to add to our dataset $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, in order to better distinguish the candidate models $\{\mathcal{M}_i\}$. After making this observation, we will form an augmented dataset $\mathcal{D}' = \mathcal{D} \cup \big\{(\mathbf{x}^*, y^*)\big\}$, from which we can recompute a new model posterior $p(\mathcal{M} \mid \mathcal{D}')$.

An approach motivated by information theory is to select the location maximizing the *mutual information* between the observation value y^* and the unknown model:

$$I(y^*; \mathcal{M} \mid \mathbf{x}^*, \mathcal{D}) = H[\mathcal{M} \mid \mathcal{D}] - \mathbb{E}_{y^*}\big[H[\mathcal{M} \mid \mathcal{D}']\big] \tag{3}$$

$$= H[y^* \mid \mathbf{x}^*, \mathcal{D}] - \mathbb{E}_{\mathcal{M}}\big[H[y^* \mid \mathbf{x}^*, \mathcal{D}, \mathcal{M}]\big], \tag{4}$$

where H indicates (differential) entropy. While Equation (3) is computationally problematic (involving costly model retraining), the equivalent expression (4) is typically more tractable, has been applied fruitfully in various active-learning settings (Houlsby et al., 2011; Garnett et al., 2014; Gardner et al., 2015; Hernández-Lobato et al., 2014; Houlsby et al., 2014), and requires only computing the differential entropy of the model-marginal predictive distribution:

$$p(y^* \mid \mathbf{x}^*, \mathcal{D}) = \sum_{i=1}^{M} p(y^* \mid \mathbf{x}^*, \mathcal{D}, \mathcal{M}_i) p(\mathcal{M}_i \mid \mathcal{D}) \tag{5}$$

and the model-conditional predictive distributions $\big\{p(y^* \mid \mathbf{x}^*, \mathcal{D}, \mathcal{M}_i)\big\}$ with all models trained with the currently available data. In contrast to (3), this does not involve any retraining cost. Although computing the entropy in (5) might be problematic, we note that this is a one-dimensional integral that can easily be resolved with quadrature. Our proposed approach, which we call *Bayesian active model selection* (BAMS) is then to compute, for each candidate location $\mathbf{x}^*$, the mutual information between y^* and the unknown model, and query where this is maximized:

$$\arg\max_{\mathbf{x}^*} I(y^*; \mathcal{M} \mid \mathbf{x}^*, \mathcal{D}). \tag{6}$$

Although active learning and model selection have been widely investigated, active model selection has received comparatively less attention. Ali et al. (2014) proposed active learning

model selection. Their method requires leave-two-out cross validation when evaluating each candidate $\mathbf{x}^*$, requiring $\mathcal{O}\big(B^2 M|\mathbf{X}^*|\big)$ model updates per iteration, where B is the total budget. Kulick et al. (2014) also considered an information-theoretic approach to active model selection, suggesting maximizing the expected cross entropy between the current model posterior $p(\mathcal{M} \mid \mathcal{D})$ and the updated distribution $p(\mathcal{M} \mid \mathcal{D}')$. This approach also requires extensive model retraining, with $\mathcal{O}\big(M|\mathbf{X}^*|\big)$ model updates per iteration, to estimate this expectation for each candidate. These approaches become prohibitively expensive for real-time applications with large number of candidates.

3. Active model selection for Gaussian processes

In the previous section, we proposed a general framework for performing sequential active Bayesian model selection, without making any assumptions about the forms of the models $\{\mathcal{M}_i\}$. Here we will discuss specific details of our proposal when these models represent alternative structures for Gaussian process priors on a latent function.

We assume that our observations are generated via a latent function $f\colon \mathcal{X} \to \mathbb{R}$ with a known observation model $p(\mathbf{y} \mid \mathbf{f})$, where $f_i = f(\mathbf{x}_i)$. A standard nonparametric Bayesian approach with such models is to place a Gaussian process (GP) prior distribution on f, $p(f) = \mathcal{GP}(f; \mu, K)$, where $\mu\colon \mathcal{X} \to \mathbb{R}$ is a mean function and $K\colon \mathcal{X}^2 \to R$ is a positive-definite covariance function or kernel (Rasmussen and Williams, 2006). We condition on the observed data to form a posterior distribution $p(f \mid \mathcal{D})$, which is typically an updated Gaussian process (making approximations if necessary). We make predictions at a new input $\mathbf{x}^*$ via the predictive distribution $p(y^* \mid \mathbf{x}^*, \mathcal{D}) = \int p(y^* \mid f^*, \mathcal{D}) p(f^* \mid \mathbf{x}^*, \mathcal{D}) \, \mathrm{d}f^*$, where $f^* = f(\mathbf{x}^*)$. The mean and kernel functions are parameterized by hyperparameters that we concatenate into a vector θ, and different choices of these hyperparameters imply that the functions drawn from the GP will have particular frequency, amplitude, and other properties. Together, μ and K define a model parametrized by the hyperparameters θ. Much attention is paid to learning these hyperparameters in a fixed model class, sometimes under the unfortunate term "model selection."

Note, however, that the *structural* (not hyperparameter) choices made in the mean function μ and covariance function K themselves are typically done by selecting (often blindly!) from several off-the-shelf solutions (see, for example, (Duvenaud, 2014; Rasmussen and Williams, 2006); though also see (Duvenaud et al., 2013; Wilson et al., 2014)), and this choice has substantial bearing on the resulting functions f we can model. Indeed, in many settings, choosing the nature of plausible functions is precisely the problem of model selection; for example, to decide whether the function has periodic structure, exhibits nonstationarity, etc. Our goal is to automatically and actively decide these structural choices during GP modeling through intelligent sampling.

To connect to our active learning formulation, let $\{\mathcal{M}_i\}$ be a set of Gaussian process models for the latent function f. Each model comprises a mean function μ_i, covariance function K_i, and associated hyperparameters θ_i. Our approach outlined in Section 2.1 requires the computation of three quantities that are not typically encountered in GP modeling and inference: the hyperparameter posterior $p(\theta \mid \mathcal{D}, \mathcal{M})$, the model evidence $p(\mathbf{y} \mid \mathbf{X}, \mathcal{M})$, and the predictive distribution $p(y^* \mid \mathbf{x}^*, \mathcal{D}, \mathcal{M})$, where we have marginalized over θ in the latter two quantities. The most-common approaches to GP inference are

maximum likelihood–II (MLE) or maximum *a posteriori*–II (MAP) estimation, where we maximize the hyperparameter posterior Rasmussen and Williams (2006):[2]

$$\hat{\theta} = \arg\max_{\theta} \log p(\theta \mid \mathcal{D}, \mathcal{M}) = \arg\max_{\theta} \log p(\theta \mid \mathcal{M}) + \log(\mathbf{y} \mid \mathbf{X}, \theta, \mathcal{M}). \tag{7}$$

Typically, predictive distributions and other desired quantities are then reported at the MLE/MAP hyperparameters, implicitly making the assumption that $p(\theta \mid \mathcal{D}, \mathcal{M}) \approx \delta(\hat{\theta})$. Although a computationally convenient choice, it does not account for uncertainty in the hyperparameters, which can be nontrivial with small datasets. Furthermore, accounting correctly for model parameter uncertainty is crucial to model selection, where it naturally introduces a model-complexity penalty. We discuss less-drastic approximations to these quantities below.

3.1. Approximating the model evidence and hyperparameter posterior

The model evidence $p(\mathbf{y} \mid \mathbf{X}, \mathcal{M})$ and hyperparameter posterior distribution $p(\theta \mid \mathcal{D}, \mathcal{M})$ are in general intractable for GPs, as there is no conjugate prior distribution $p(\theta \mid \mathcal{M})$ available. Instead, we will use a Laplace approximation, where we make a second-order Taylor expansion of $\log p(\theta \mid \mathcal{D}, \mathcal{M})$ around its mode $\hat{\theta}$ (7). The result is a multivariate Gaussian approximation:

$$p(\theta \mid \mathcal{D}, \mathcal{M}) \approx \mathcal{N}(\theta; \hat{\theta}, \Sigma); \qquad \Sigma^{-1} = -\nabla^2 \log p(\theta \mid \mathcal{D}, \mathcal{M})\big|_{\theta=\hat{\theta}}. \tag{8}$$

The Laplace approximation also results in an approximation to the model evidence:

$$\log p(\mathbf{y} \mid \mathbf{X}, \mathcal{M}) \approx \log p(\mathbf{y} \mid \mathbf{X}, \hat{\theta}, \mathcal{M}) + \log p(\hat{\theta} \mid \mathcal{M}) - \tfrac{1}{2} \log\det \Sigma^{-1} + \tfrac{d}{2} \log 2\pi, \tag{9}$$

where d is the dimension of θ (Raftery, 1996; Kuha, 2004). The Laplace approximation to the model evidence can be interpreted as rewarding explaining the data well while penalizing model complexity. Note that the *Bayesian information criterion* (BIC), commonly used for model selection, can be seen as an approximation to the Laplace approximation (Murphy, 2012).

3.2. Approximating the predictive distribution

We next consider the predictive distribution:

$$p(y^* \mid \mathbf{x}^*, \mathcal{D}, \mathcal{M}) = \int p(y^* \mid f^*) \underbrace{\int p(f^* \mid \mathbf{x}^*, \mathcal{D}, \theta, \mathcal{M}) p(\theta \mid \mathcal{D}, \mathcal{M}) \, \mathrm{d}\theta}_{p(f^* \mid \mathbf{x}^*, \mathcal{D}, \mathcal{M})} \, \mathrm{d}f^*. \tag{10}$$

The posterior $p(f^* \mid \mathbf{x}^*, \mathcal{D}, \theta, \mathcal{M})$ in (10) is typically a known Gaussian distribution, derived analytically for Gaussian observation likelihoods. However, the integral over θ in (10) is intractible, even with a Gaussian approximation to the hyperparameter posterior as in (8).

Garnett et al. (2014) introduced a mechanism for approximately marginalizing GP hyperparameters (called the MGP), which we will adopt here. The MGP assumes that we have

2. Using a noninformative prior $p(\theta \mid \mathcal{M}) \propto 1$ in the case of maximum likelihood.

made a Gaussian approximation to the hyperparameter posterior, $p(\theta \mid \mathcal{D}, \mathcal{M}) \approx \mathcal{N}(\theta; \hat{\theta}, \Sigma)$.[3] We define the posterior predictive mean and variance functions as

$$\mu^*(\theta) = \mathbb{E}[f^* \mid \mathbf{x}^*, \mathcal{D}, \theta, \mathcal{M}]; \qquad \nu^*(\theta) = \mathrm{Var}[f^* \mid \mathbf{x}^*, \mathcal{D}, \theta, \mathcal{M}].$$

The MGP works by making an expansion of the predictive distribution around the posterior mean hyperparameters $\hat{\theta}$. The nature of this expansion is chosen so as to match various derivatives of the true predictive distribution; see (Garnett et al., 2014) for details. The posterior distribution of f^* is approximated by

$$p(f^* \mid \mathbf{x}^*, \mathcal{D}, \mathcal{M}) \approx \mathcal{N}\big(f^*; \mu^*(\hat{\theta}), \sigma^2_{\text{MGP}}\big), \tag{11}$$

where

$$\sigma^2_{\text{MGP}} = \tfrac{4}{3}\nu^*(\hat{\theta}) + \big[\nabla\mu^*(\hat{\theta})\big]^\top \Sigma \big[\nabla\mu^*(\hat{\theta})\big] + \tfrac{1}{3\nu^*(\hat{\theta})}\big[\nabla\nu^*(\hat{\theta})\big]^\top \Sigma \big[\nabla\nu^*(\hat{\theta})\big]. \tag{12}$$

The MGP thus inflates the predictive variance from the the posterior mean hyperparameters $\hat{\theta}$ by a term that is commensurate with the uncertainty in θ, measured by the posterior covariance Σ, and the dependence of the latent predictive mean and variance on θ, measured by the gradients $\nabla\mu^*$ and $\nabla\nu^*$. With the Gaussian approximation in (11), the integral in (10) now reduces to integrating the observation likelihood against a univariate Gaussian. This integral is an analytic (Rasmussen and Williams, 2006) for Gaussian likelihoods.

3.3. Implementation

Given the development above, we may now efficiently compute an approximation to the BAMS criterion for active GP model selection. Given currently observed data $\mathcal{D}$, for each of our candidate models $\mathcal{M}_i$, we first find the Laplace approximation to the hyperparameter posterior (8) and model evidence (9). Given the approximations to the model evidence, we may compute an approximation to the model posterior (2). Suppose we have a set of candidate points $\mathbf{X}^*$ from which we may select our next point. For each of our models, we compute the MGP approximation (11) to the latent posteriors $\big\{p(\mathbf{f}^* \mid \mathbf{X}^*, \mathcal{D}, \mathcal{M}_i)\big\}$, from which we compute the predictive distributions $\big\{p(\mathbf{y}^* \mid \mathbf{X}^*, \mathcal{D}, \mathcal{M}_i)\big\}$. Finally, with the ability to compute the differential entropies of these model-conditional predictive distributions, as well as the marginal predictive distribution (5), we may compute the mutual information of each candidate in parallel. For regression with Gaussian observation noise, we must resort to one-dimensional quadrature to evaluate the model-marginal entropy.

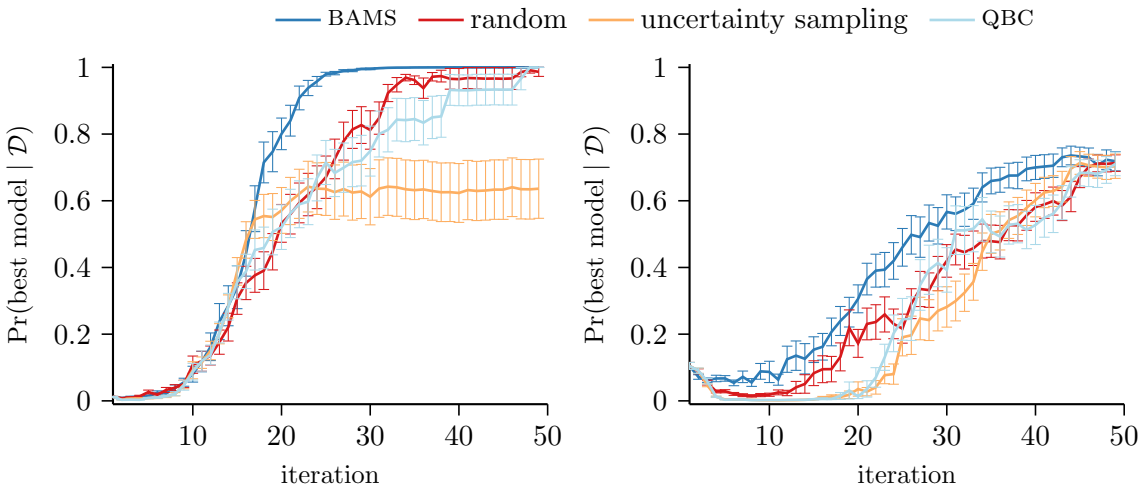
4. Initial results

We evaluated the proposed method using the time series shown in Table 1, also used by Duvenaud et al. (2013). For each dataset, we considered 18 models: four base kernels (squared exponential (SE), periodic (PER), linear (LIN), and rational quadratic (RQ)), 12 kernels obtained by combining these basis kernels with sum and products, an expert model designing by a human (see Table 1) and the kernel found by the compositional kernel search (CKS) (Duvenaud et al., 2013) method. Using all points, we computed the model posterior to determine which model best describes the data, shown in Table 1 (bold).

3. This is arbitrary and need not be the Laplace approximation in (8), so this is a slight abuse of notation.

dataset	expert model	CKS model
Airline passengers	**LIN+(PER×RQ)**	SE×(LIN+LIN×(PER+RQ))
Mauna Loa atmospheric CO_2	SE+(PER×SE)+RQ+SE	**LIN×SE+SE×(PER+RQ)**

Table 1: Kernels used as experts suggestion and selected by the method of (Duvenaud et al., 2013) for each time series. The kernel that best explains the data using all points is shown in bold. The expert kernel for the airline data was our guess after plotting the data, whereas for Mauna Loa is the one presented in Rasmussen and Williams (2006).



(a) Airline data set

(b) Mauna Loa atmospheric CO_2

Figure 1: Posterior probability of best model as function of iteration number.

Then, we randomly selected two points and applied several techniques to choose a total of 50 points: BAMS, random sampling, uncertainty sampling, and query by committee (QBC)[4] Seung et al. (1992). We repeated this experiment 30 times, using different initial random seeds for each. Figure 2 shows the average posterior probability of the best model as function of iteration number, along with the standard error. The results show that BAMS outperforms all the baselines, more quickly finding the best model.

5. Conclusion

Here we introduced a novel information-theoretic approach for active model selection, BAMS, and present some initial results of its application to structure discovery.

4. We adapted QBC using the entropy of the model-marginal predictive distribution (5) as the disagreement criterion. Furthermore, to consider every committee member's (i.e. each model's) predictions equally important, we compute this quantity assuming a uniform model posterior.

References

A. Ali, R. Caruana, and A. Kapoor. Active Learning with Model Selection. In *AAAI 2014*, pages 1673–1679, 2014.

David Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, Computational and Biological Learning Laboratory, University of Cambridge, 2014.

David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. In *ICML 2013*, pages 1166–1174, 2013.

Jacob R. Gardner, Xinyu Song, Kilian Q. Weinberger, Dennis Barbour, and John P. Cunningham. Psychophysical Detection Testing with Bayesian Active Learning. In *UAI 2015*, 2015.

Roman Garnett, Michael A Osborne, and Philipp Hennig. Active Learning of Linear Embeddings for Gaussian Processes. In *UAI 2014*, pages 230–239, 2014.

Roger B. Grosse, Ruslan Salakhutdinov, William T. Freeman, and Joshua B. Tenenbaum. Exploiting compositionality to explore a large space of model structures. In *UAI 2012*, pages 306–315, 2012.

José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive Entropy Search for Efficient Global Optimization of Black-box Functions. In *NIPS 2014*, pages 918–926, 2014.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. arXiv preprint arXiv:1112.5745 [stat.ML], 2011.

Neil Houlsby, José M Hernández-Lobato, and Zoubin Ghahramani. Cold-start Active Learning with Robust Ordinal Matrix Factorization. In *ICML 2014*, pages 766–774, 2014.

Jouni Kuha. AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods and Research*, 33(2):188–229, 2004.

Johannes Kulick, Robert Lieck, and Marc Toussaint. Active Learning of Hyperparameters: An Expected Cross Entropy Criterion for Active Model Selection. arXiv preprint arXiv:1409.7552 [stat.ML], 2014.

James Robert Lloyd, David Duvenaud, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Automatic Construction and Natural-Language Description of Nonparametric Regression Models. In *AAAI 2014*, pages 1242–1250, 2014.

David JC MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Adrian E Raftery. Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models. *Biometrika*, 83(2):251–266, 1996.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT 92*, pages 287–294, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X.

Andrew G Wilson, Elad Gilboa, Arye Nehorai, and John P Cunningham. Fast Kernel Learning for Multidimensional Pattern Extrapolation. In *NIPS 2014*, pages 3626–3634, 2014.