Chapter 1

Conclusions and Discussion

1.1 Summary of Contributions

The main contribution of this thesis is to show how to automate the discovery and explanation of structure in functions, simply by searching an open-ended language of regression models. This thesis also includes a set of related results showing how Gaussian processes can be extended, or composed with other models.

Chapter 1.3 is a tutorial showing how to build a wide variety of structured models of functions by constructing appropriate covariance functions. We'll also show how GPs can produce nonparametric models of manifolds with diverse topological structures, such as cylinders, toruses and Möbius strips.

Chapter 1.3 shows how to search over a general, open-ended language of models, built by composing together different kernels. Since we can evaluate each model by the marginal likelihood, we can automatically construct custom models for each dataset by a breadth-first search. The nature of GPs allow the resulting models to be visualized by decomposing them into diverse, interpretable components, each capturing a different type of structure. Capturing this high-level structure sometimes even allows us to extrapolate beyond the range of the data.

One benefit of using a compositional model class is that the resulting models are interpretable. **Chapter 1.3** demonstrates a system which automatically describes the structure implied by a given kernel on a given dataset, generating reports with graphs and English-language text describing the resulting model. We'll show several automatic analyses of time-series. Combined with the automatic model search developed in chapter 1.3, this system represents the beginnings of an "automatic statistician".

Chapter 1.3 analyzes deep neural network models by characterizing the prior over

functions obtained by composing GP priors to form *deep Gaussian processes*. We show that, as the number of layers increase, the amount of information retained about the original input diminishes to a single degree of freedom. A simple change to the network architecture fixes this pathology. We relate these models to neural networks, and as a side effect derive several forms of *infinitely deep kernels*.

Chapter 1.3 examines a more limited, but much faster way of discovering structure using GPs. Specifying a kernel with many different types of structure, we use kernel parameters to discard whichever types of structure *aren't* found in the current dataset. The model class we examine is called *additive Gaussian processes*, a model summing over exponentially-many GPs, each depending on a different subset of the input variables. We give a polynomial-time inference algorithm for this model class, and relate it to other model classes. For example, additive GPs are shown to have the same covariance as a GP that uses *dropout*, a recently discovered regularization technique for neural networks.

Chapter 1.3 develops a Bayesian clustering model in which the clusters have non-parametric shapes - the infinite Warped Mixture Model. The density manifolds learned by this model follow the contours of the data density, and have interpretable, parametric forms in the latent space. The marginal likelihood lets us infer the effective dimension and shape of each cluster separately, as well as the number of clusters.

Section 1.3 contains a detailed tutorial on the many sorts of structure that can be computed by kernels. Most of the material there is known by experts, but has not been compiled together into one place before.

1.2 Structured versus Unstructured GP Models

One question left unanswered by this thesis is when to prefer the highly structured models of sections 1.3 to 1.3 to the relatively unstructured models of section 1.3.

The warped mixture model of section 1.3 represents a compromise between these two approaches, combining a discrete clustering model with an unstructured warping function. However, the results of (**?**) suggest that clustering can be automatically accomplished by a sufficiently deep, unstructured GP.

Difficulty of Optimization

The discrete nature of the search over composite kernel structures can be seen as a blessing and a curse. Certainly, a mixed discrete and continuous optimization requires

more complex procedures compared to the continuous-only optimization possible in deep GPs.

However, the discrete nature of the space of composite kernels offers the possibility of learning heuristics to suggest which types of structure to add. For example, finding periodic structure or growing variance in the residuals of a model suggests adding periodic or linear components to the kernel, respectively. It is not clear whether such heuristics can easily be found for optimizing the variational parameters of a deep GP.

Extrapolation

Another question is whether, and how, an equally rich inductive bias can be encoded into relatively unstructured models such as deep GPs. As an example, consider the problem of extrapolating a periodic function. A deep GP could learn a latent representation similar to that of the periodic kernel, projecting into a basis eqeuivalent to $[\sin(x), \cos(x)]$ in the first hidden layer. However, to extrapolate a periodic function, the sin and cos functions would have to continue to repeat beyond the range of the training data, which would not happen if each layer assumed only local smoothness.

One obvious possibility is to marry the two approaches, learning deep GPs with structured kernels. However, we may lose some of the advantages of interpretability by this approach.

Another point to consider is that, in high dimensions, the line between interpolation and extrapolation is blurred, and that learning a suitable representation of the data manifold may be sufficient for most purposes.

Ease of Interpretation

For summarizing the learned structure on low-dimensional datasets, section 1.3 showed that the composite kernels allow a simple recipe for visualizing and desribing the learned structure. On the other hand, **?** showed that deep GP-LVMs allow summarization of the learned structure through sampling from the posterior, examining the dimension of the different latent layers, visualizing the latent coordinates, or examining how the predictive distribution changes as one moves in different directions in the latent space.

1.3 Automating Machine Learning and Statistics

It seems clear that, one way or another, large parts of the existing practice of model-building will be eventually automated. The machine learning community has so far mostly focused on producing efficient inference strategies for powerful model classes, which is sufficient for improving predictive performance. Historically, the statistics community has put much more emphasis on the interpretabilty and meaning of models. To begin to automate the practice of statistics, developing more sophisticated model-description procedures seems like the direction with the most low-hanging fruit.