Chapter 1

Discussion: Structured versus Unstructured GP Models

One question left unanswered by this thesis is when to prefer the highly structured models of sections 1.1 to 1.1 to the relatively unstructured models of section 1.1.

The warped mixture model of section 1.1 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 1.1 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.

1.1 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.