Chapter 1

Introduction

> "I only work on intractable nonparametrics - Gaussian processes don't count."
>
> Sinead Williamson, personal communication

1.1 Regression

The general problem of regression consists of learning a function f mapping from some input space $\mathcal{X}$ to some output space $\mathcal{Y}$. We would like an expressive language which can represent both simple parametric forms of f such as linear, polynomial, etc. and also complex nonparametric functions specified in terms of properties such as smoothness, periodicity, etc. Fortunately, Gaussian processes (GPs) provide a very general and analytically tractable way of capturing both simple and complex functions.

1.2 Gaussian process models

Gaussian processes are distributions over functions such that any finite subset of function evaluations, $(f(x_1), f(x_2), \ldots f(x_N))$, have a joint Gaussian distribution (**?**). A GP is completely specified by its mean function, $\mu(x) = \mathbb{E}(f(x))$ and kernel (or covariance) function $k(x, x') = \text{Cov}(f(x), f(x'))$. It is common practice to assume zero mean, since marginalizing over an unknown mean function can be equivalently expressed as a zero-mean GP with a new kernel. The structure of the kernel captures high-level properties of the unknown function, f, which in turn determines how the model generalizes or extrapolates to new data. We can therefore define a language of regression models by specifying a language of kernels.

1.2.1 Useful properties of Gaussian process models

- **Tractable inference** Given a kernel function, the posterior distribution can be computed exactly in closed form. This is a rare property for nonparametric models to have.

- **Expressivity** by choosing different covariance functions, we can express a very wide range of modeling assumptions.

- **Integration over hypotheses** the fact that a GP posterior lets us exactly integrate over a wide range of hypotheses means that overfitting is less of an issue than in comparable model classes - for example, neural nets.

- **Marginal likelihood** A side benefit of being able to integrate over all hyoptheses is that we compute the *marginal likelihood* of the data given the model. This gives us a principled way of comparing different Gaussian process models.

- **Closed-form posterior** The posterior predictive distribution of a GP is another GP. This means that GPs can easily be composed with other models or decision procedures. For example, (*) Carl's reinforcement learning work.

Figure 1.1 shows a Gaussian process posterior. Typically, it's rendered with the mean and +- 2SD, but there's nothing special about mean.

1.3 Latent Variable Models

Besides being useful for modeling functions, a simple extension allows GPs to be useful for general density modeling. Unfortunately, this extension causes many of the useful properties of the GP not to hold.

1.4 Structure through additivity

A theme throughout this thesis is exploring the idea that a lot of the expressivity of GP models comes from the fact that these models can be combined and decomposed additively.

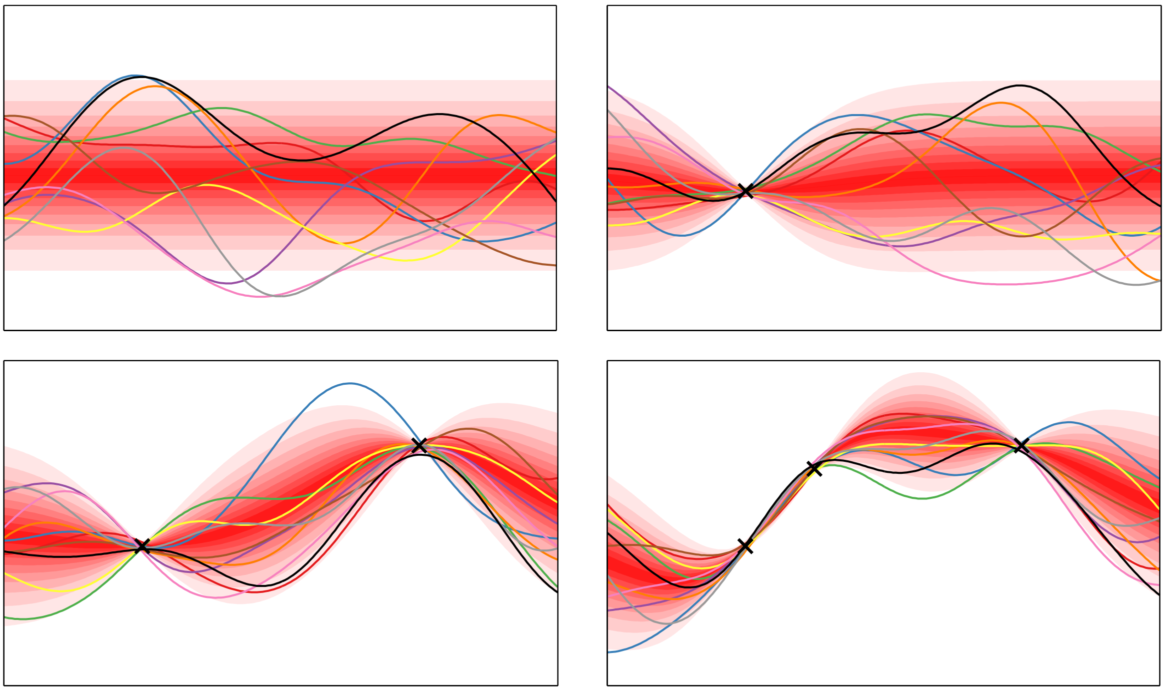

Fig. 1.1 A visual representation of a one-dimensional Gaussian process posterior. Red isocountours show the marginal density at each input location. Coloured lines are samples from the posterior.

1.4.1 Derivation of Component Marginal Variance

In this section, we derive the posterior marginal variance and covariance of the additive components of a GP. These formulas let us plot the marginal variance of each component separately. These formulas can also be used to examine the posterior covariance between pairs of components.

Let us assume that our function $\mathbf{f}$ is a sum of two functions, $\mathbf{f}_1$ and $\mathbf{f}_2$, where $\mathbf{f} = \mathbf{f}_1 + \mathbf{f}_2$. If $\mathbf{f}_1$ and $\mathbf{f}_2$ are a priori independent, and $\mathbf{f}_1 \sim \mathrm{GP}(\mu_1, k_1)$ and $\mathbf{f}_2 \sim \mathrm{GP}(\mu_2, k_2)$, then

$$\begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_1^\star \\ \mathbf{f}_2 \\ \mathbf{f}_2^\star \\ \mathbf{f} \\ \mathbf{f}^\star \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_1^\star \\ \mu_2 \\ \mu_2^\star \\ \mu_1 + \mu_2 \\ \mu_1^\star + \mu_2^\star \end{bmatrix}, \begin{bmatrix} \mathbf{k}_1 & \mathbf{k}_1^\star & 0 & 0 & \mathbf{k}_1 & \mathbf{k}_1^\star \\ \mathbf{k}_1^\star & \mathbf{k}_1^{\star\star} & 0 & 0 & \mathbf{k}_1^\star & \mathbf{k}_1^{\star\star} \\ 0 & 0 & \mathbf{k}_2 & \mathbf{k}_2^\star & \mathbf{k}_2 & \mathbf{k}_2^\star \\ 0 & 0 & \mathbf{k}_2^\star & \mathbf{k}_2^{\star\star} & \mathbf{k}_2^\star & \mathbf{k}_2^{\star\star} \\ \mathbf{k}_1 & \mathbf{k}_1^\star & \mathbf{k}_2 & \mathbf{k}_2^\star & \mathbf{k}_1 + \mathbf{k}_2 & \mathbf{k}_1^\star + \mathbf{k}_2^\star \\ \mathbf{k}_1^\star & \mathbf{k}_1^{\star\star} & \mathbf{k}_2^\star & \mathbf{k}_2^{\star\star} & \mathbf{k}_1^\star + \mathbf{k}_2^\star & \mathbf{k}_1^{\star\star} + \mathbf{k}_2^{\star\star} \end{bmatrix} \right) \tag{1.1}$$

where $\mathbf{k}_1 = k_1(\mathbf{X}, \mathbf{X})$ and $\mathbf{k}_1^\star = k_1(\mathbf{X}^\star, \mathbf{X})$.

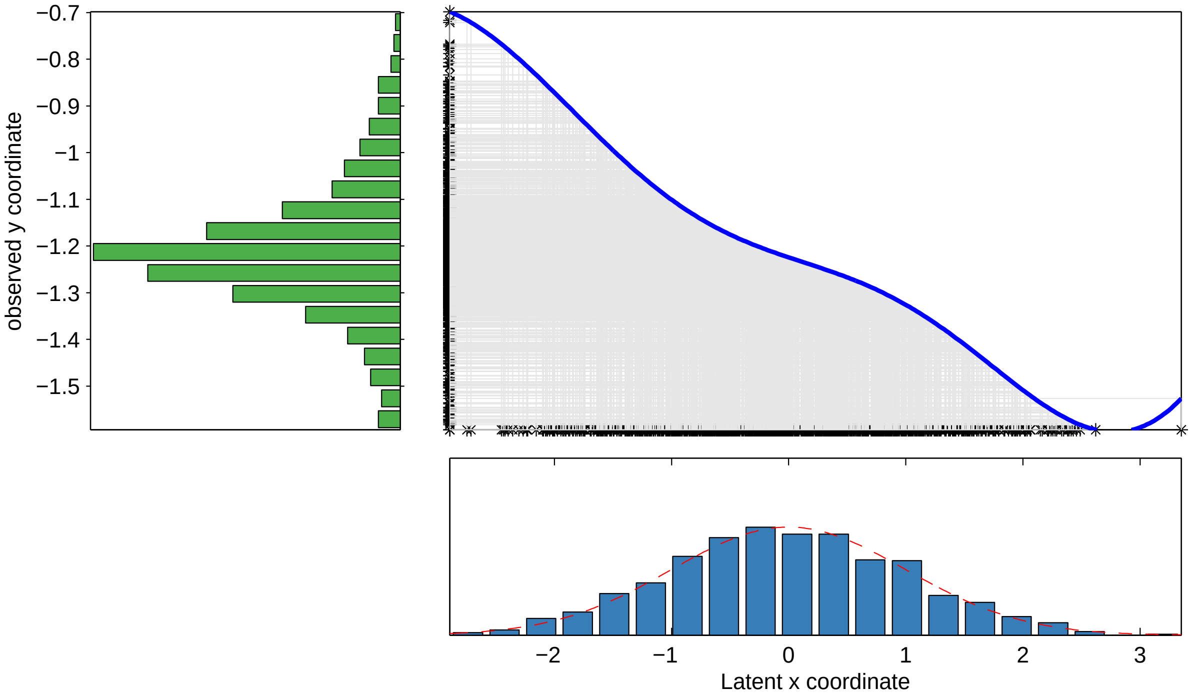

Fig. 1.2 A visual representation of the Gaussian process latent variable model. Bottom: density and samples from a 1D Gaussian, specifying the distribution $p(\mathbf{X})$ in the latent space. Top Right: A function drawn from a GP prior. Left: A nonparametric density defined by warping the latent density through the function drawn from a GP prior.

By the formula for Gaussian conditionals:

$$\mathbf{x}_A | \mathbf{x}_B \sim \mathcal{N}(\mu_A + \mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}(\mathbf{x}_B - \mu_B), \mathbf{\Sigma}_{AA} - \mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}\mathbf{\Sigma}_{BA}), \tag{1.2}$$

we get that the conditional variance of a Gaussian conditioned on its sum with another Gaussian is given by

$$\begin{aligned}\mathbf{f}_1(\mathbf{x}^\star)|\mathbf{f}(\mathbf{x}) \sim \mathcal{N}\Big(&\mu_1(\mathbf{x}^\star) + \mathbf{k}_1(\mathbf{x}^\star, \mathbf{x})\left[\mathbf{K}_1(\mathbf{x}, \mathbf{x}) + \mathbf{K}_2(\mathbf{x}, \mathbf{x})\right]^{-1}(\mathbf{f}(\mathbf{x}) - \mu_1(\mathbf{x}) - \mu_2(\mathbf{x})), \\ &\mathbf{k}_1(\mathbf{x}^\star, \mathbf{x}^\star) - \mathbf{k}_1(\mathbf{x}^\star, \mathbf{x})\left[\mathbf{K}_1(\mathbf{x}, \mathbf{x}) + \mathbf{K}_2(\mathbf{x}, \mathbf{x})\right]^{-1}\mathbf{k}_1(\mathbf{x}, \mathbf{x}^\star)\Big).\end{aligned} \tag{1.3}$$

These formulae express the posterior model uncertainty about different components of the signal, integrating over the possible configurations of the other components.

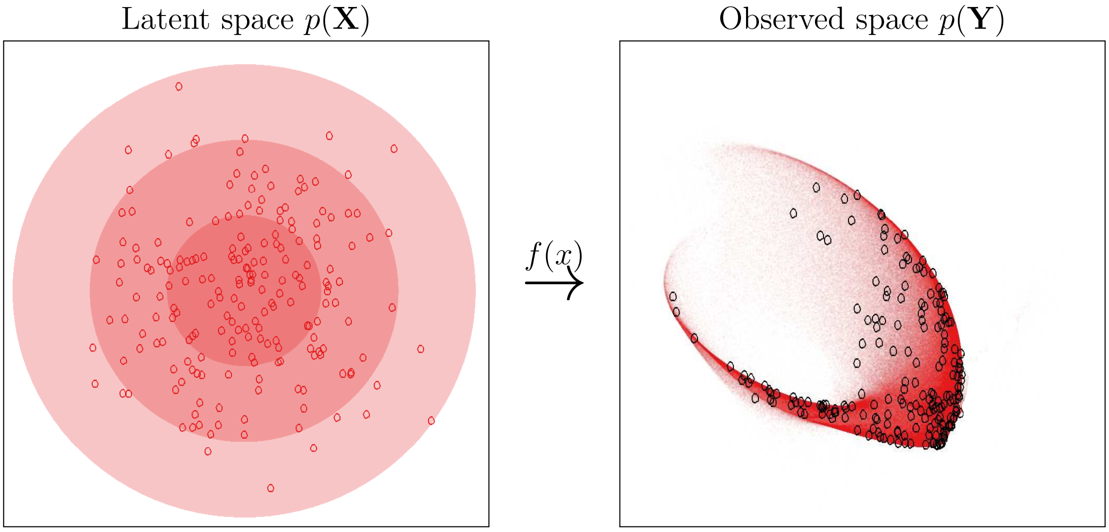

Fig. 1.3 A visual representation of the Gaussian process latent variable model. Left: Isocontours and samples from a 2D Gaussian, specifying the distribution $p(\mathbf{X})$ in the latent space. Right: Density and samples from a nonparametric density defined by warping the latent density through a function drawn from a GP prior.

1.5 Covariance functions

Kernels specify similarity between function values of two objects, not between similarity of objects