Creating Non-Gaussian Processes from Gaussian Processes by the Log-Sum-Exp Approach

Radford M. Neal, 28 February 2005

A Very Brief Review of Gaussian Processes

A Gaussian process is a distribution over functions, $X(t)$. Here, t could be anything, but is typically a vector in R^p. We can use such a distribution as the prior for a Bayesian regression model.

The defining feature of a Gaussian process is that the joint distribution of the function values at *any finite number* of points, $t_1, \ldots, t_n$, is multivariate Gaussian.

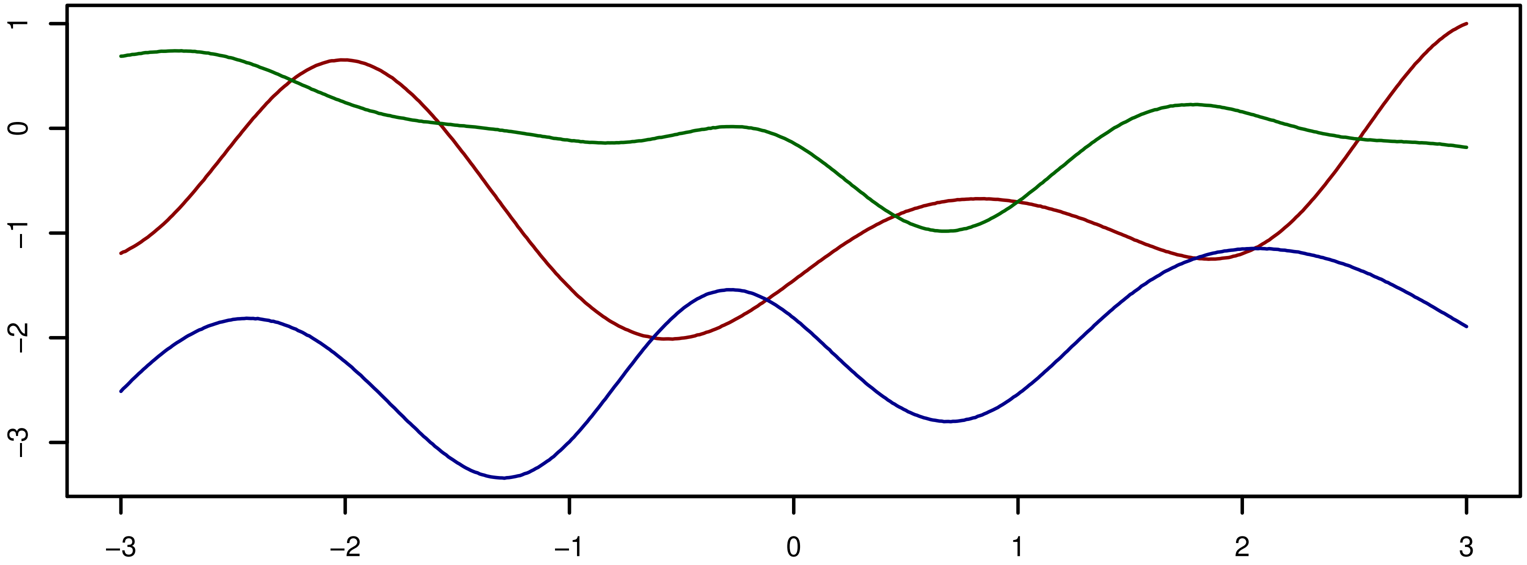
Consequently, a Gaussian process can be completely specified by

- its *mean function*: $\mu(t) = E[X(t)]$, and
- its *covariance function*: $C(t, t') = E[(X(t) - \mu(t))(X(t') - \mu(t'))]$.

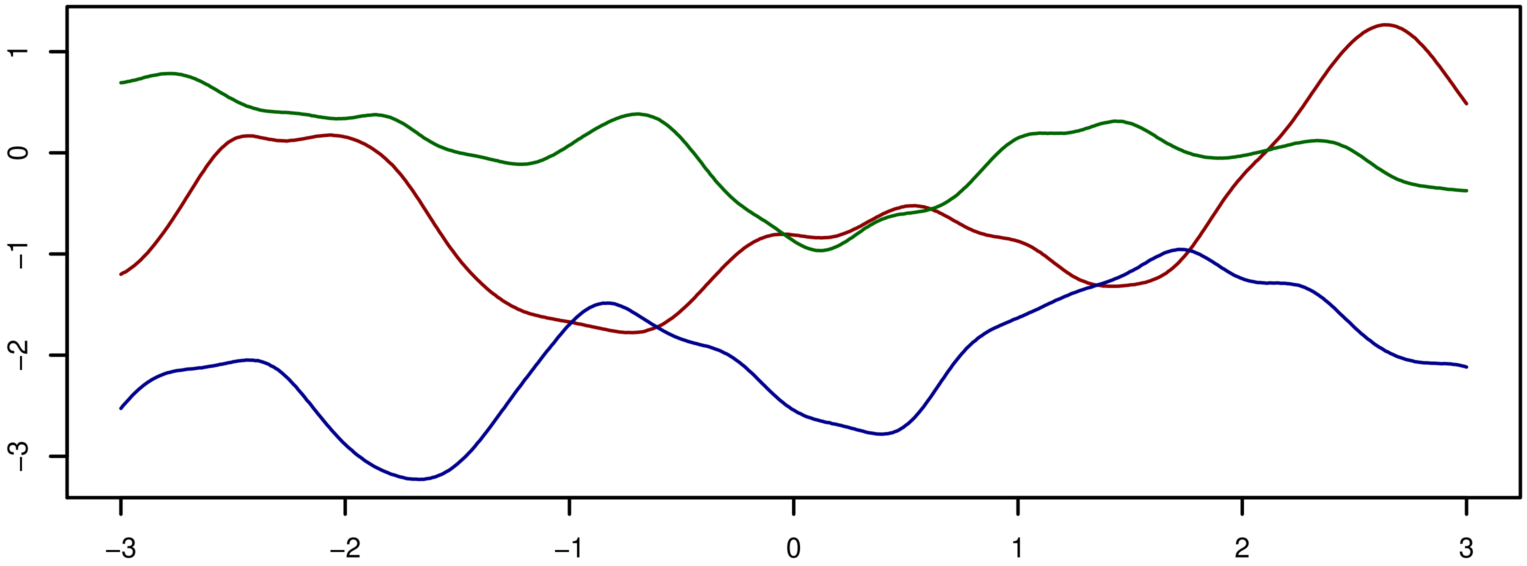
From these, we can compute the mean vector and covariance matrix for the joint distribution of $X(t_1), \ldots, X(t_n)$. Of course, the covariance function must be such that the resulting covariance matrix is always positive semi-definite.

Examples of Functions Drawn from Gaussian Processes

Three functions from a GP with $\mu(t) = 0$, $C(t, t') \; = \; 1.5^2 \; + \; \exp(-(t - t')^2)$:



Three with $\mu(t) = 0$, $C(t, t') \; = \; 1.5^2 \; + \; \exp(-(t - t')^2) \; + \; 0.2^2 \exp(-10(t - t')^2)$:



Non-Gaussian Distributions Over Functions

The choice of covariance function allows considerable flexibility in GP models, but some distributions cannot be obtained.

Consider the following class of functions:

$$X(t) \;=\; \begin{cases} b & \text{if } t < a \\ -b & \text{if } t \ge a \end{cases}$$

where b is randomly chosen to be $+1$ or -1 with equal probabilities, and a is randomly chosen uniformly over some broad range.

We can easily see that $\mu(t) = E[X(t)] = 0$. The covariance function is

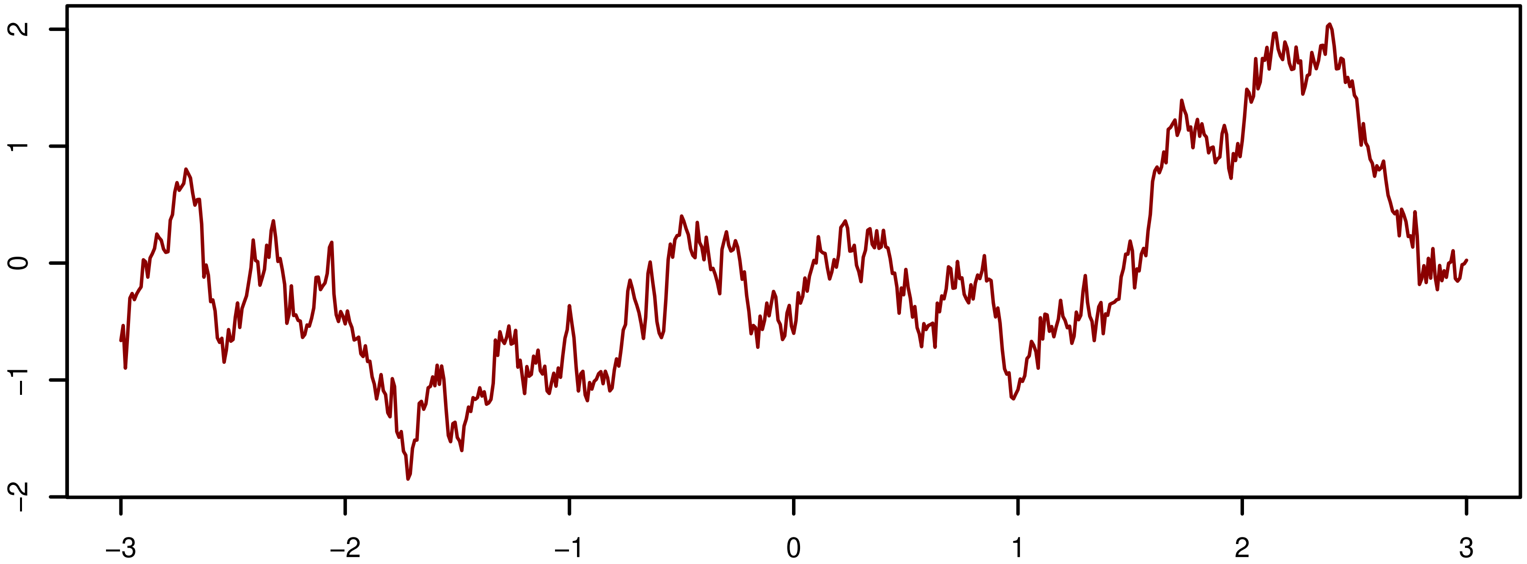
$$C(t,t') \;=\; E[X(t)X(t')] \;=\; P(a \notin (t,t')) \;-\; P(a \in (t,t'))$$

Over the broad range for a, we see that for nearby t and t', $C(t,t') \approx 1 - c|t-t'|$, for some c, which is locally similar to $\exp(-c|t-t'|)$.

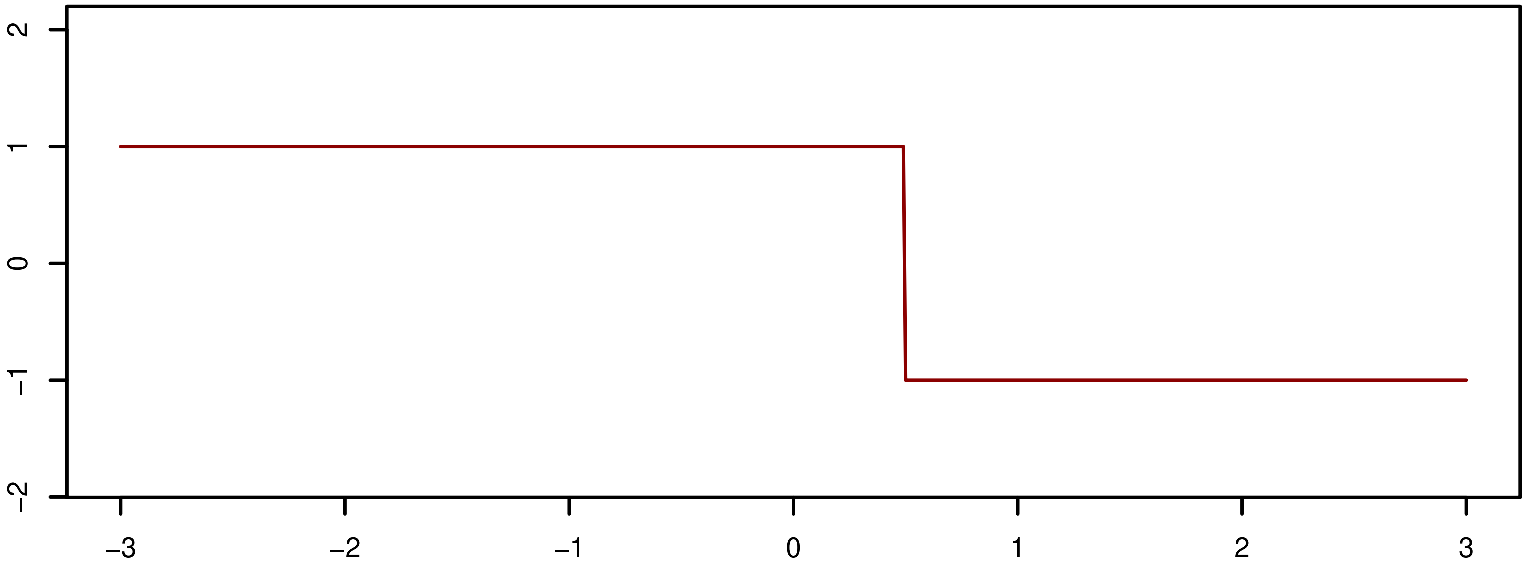
But what do functions drawn from the Gaussian process with this mean and covariance look like?

A Function Drawn From This Gaussian Process

Here's a function drawn from the GP with $\mu(t) = 0$, $C(t, t') = \exp(-|t - t'|)$:



This is nothing like what we were looking for, which is something like this:



Warped Gaussian Processes

One way to get non-Gaussian processes from Gaussian processes is to "warp" the function by applying some non-linear transformation.

This is often done manually — eg, we may decide to model the logs of data points rather than the original values.

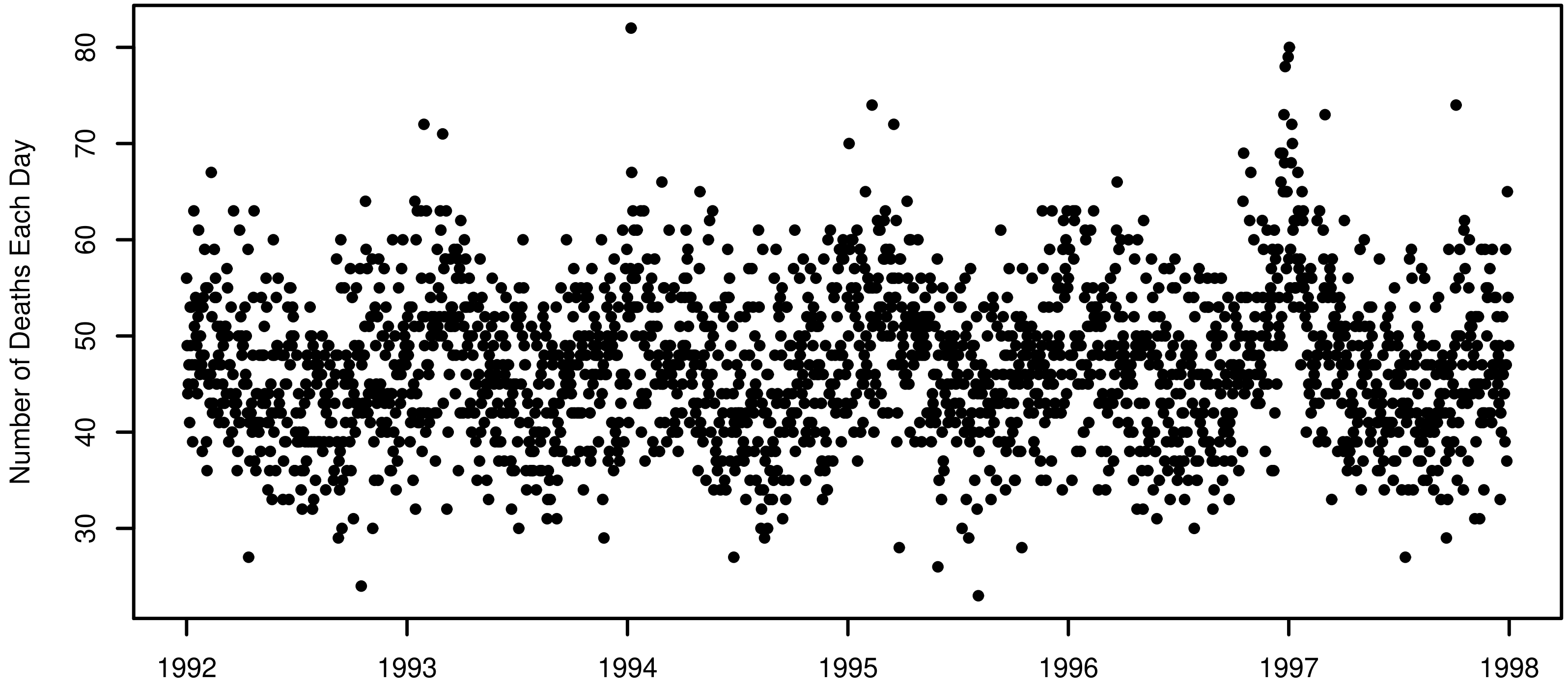
Snelson, Rasmussen, and Ghahramani (NIPS 2003) find a monotonic warping function automatically — eg, from the class $f(t) = t + \sum_{i=1}^{I} a_i \tanh(b_i(t + c_i))$. The hyperparameters a_i b_i, and c_i can be set to their maximum *a posteriori* probability values (as SRG do) or their posterior can be sampled using MCMC.

Importantly, fitting warped Gaussian processes to data is not much harder than fitting ordinary Gaussian processes. In both cases, the crucial computations are done with matrix operations. Doing the warping requires only that we be able to compute f, f', and f^{-1}.

However, warping gets us only a small extension beyond Gaussian processes. For instance, we can't get processes whose smoothness characteristics vary over the input space, in *a priori* unknown ways.

Modeling Functions as Sums of Other Functions

Here is data on daily mortality in Toronto from 1992 to 1997:



We can model these counts a being Poisson distributed, with the mean being a function of time. This function can incorporate yearly and weekly cycles, an overall trend, as well as short-term dependencies, such as could result from flu epidemics.

People die of many causes, however. Rather than model the mean directly, perhaps we should use a model that represents the mean number of deaths as a sum of several terms, corresponding to various causes of death.

How About Using a Sum of Gaussian Processes?

At first, we might think this is easy to do with a Gaussian process, since if $X(t) = A(t) + B(t) + C(t)$, and $A(t)$, $B(t)$, and $C(t)$ are independent Gaussian processes, then $X(t)$ will also be a Gaussian process. Its mean and covariance functions will just be the sums of those for $A(t)$, $B(t)$, and $C(t)$.

However: The mean number of deaths is non-negative, whereas Gaussian processes range over the whole real line. This motivates modeling the log of the mean with a GP, not the mean itself.

In any case: If the mean number of deaths is a Gaussian process, it will have the limitations of a Gaussian process. Consequently, we won't be able to model events like sudden flu epidemics very well.

The Log-Sum-Exp Approach

Let $X(t)$ be the log of the mean number of deaths on day t. If we want to model the mean as the sum of contributions from various causes of death, we are led to models such as the following:

$$X(t) \;=\; \log(\exp(A(t)) \,+\, \exp(B(t)) \,+\, \exp(C(t)))$$

where $A(t)$, $B(t)$, and $C(t)$ represent the logs of the mean numbers of deaths from three causes. We give Gaussian process priors to $A(t)$, $B(t)$, and $C(t)$.

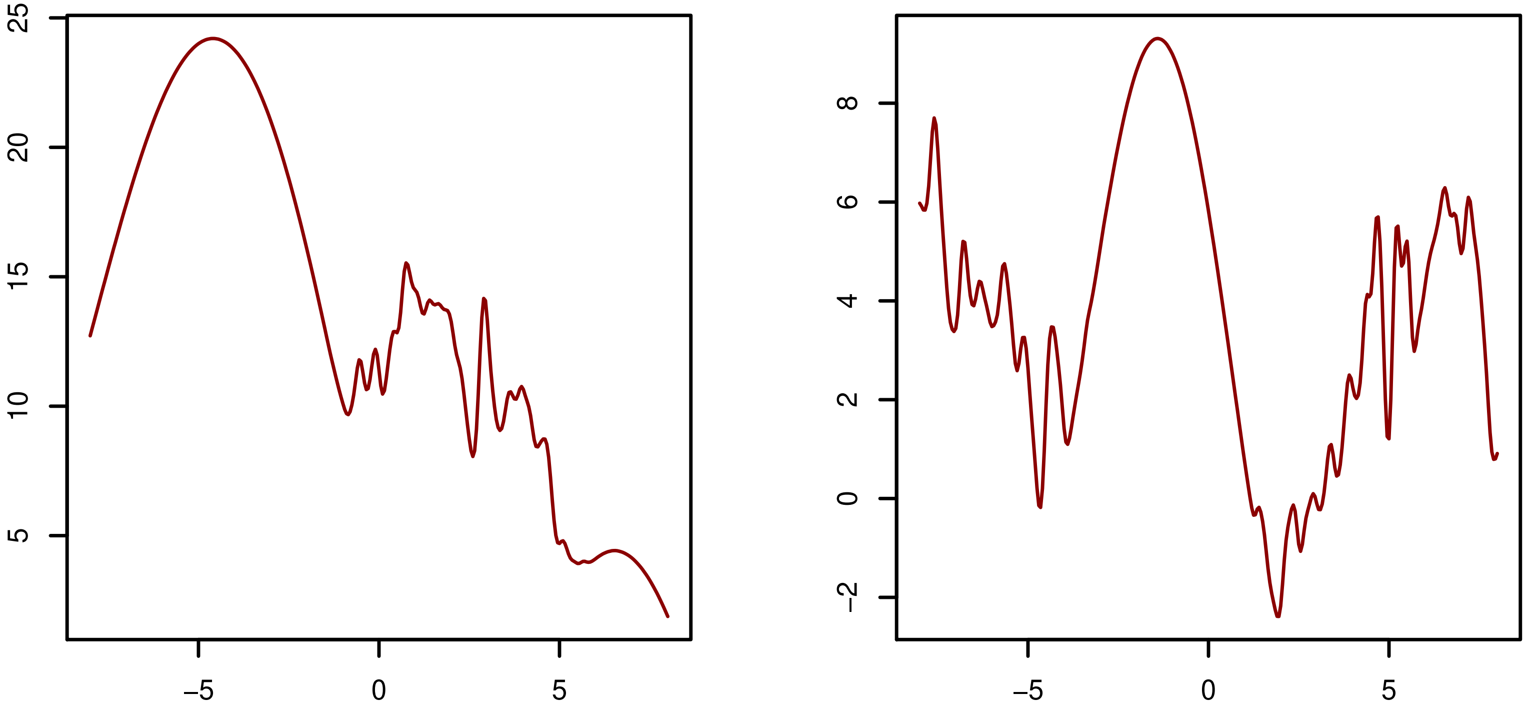
I assume here that we observe *only* the total number of deaths. $A(t)$, $B(t)$, and $C(t)$ are “latent” processes, which will not be observed directly (even noisily). The three causes of death that they correspond to would be “discovered” by the model; they aren’t fixed beforehand.

We don’t observe $X(t)$ either, but we do observe the number of deaths on day t, $N(t) \sim \text{Poisson}(\exp(X(t)))$.

Functions Randomly Drawn from a Log-Sum-Exp GP Model

Here are two functions randomly drawn from a log-sum-exp GP model with $X(t) = \log(\exp(A(t)) + \exp(B(t)))$, with $A(t)$ and $B(t)$ having zero means, and the following covariance functions:

$$\begin{aligned}
C_A(t,t') &= 10^2 \exp(-0.04(t-t')^2) \\
C_B(t,t') &= 10^2 \exp(-0.04(t-t')^2) \ + \ \exp(-25(t-t')^2)
\end{aligned}$$



Computations for Log-Sum-Exp GP Models

For regression with Gaussian noise, computations for a plain GP model involve:

- matrix operations to find the likelihood
- MCMC updates of the hyperparameters, based on the likelihood
- matrix operations to make predictions.

For a log-sum-exp model, we need to explicitly represent $A(t)$, $B(t)$, etc. for all training points. Computations involve:

- MCMC updates for the latent $A(t)$, $B(t)$, etc.
- MCMC updates for hyperparameters, based on current $A(t)$, $B(t)$, etc.
- matrix operations for predictions, with the results combined by log-sum-exp.

MCMC updates for both the latent values and the hyperparameters require matrix operations. Updates of the latent values look at the observations; updates for the hyperparameters look only at $A(t)$, $B(t)$, etc.

Keeping around latent values slows MCMC convergence. For models with non-Gaussian observations (eg, Poisson counts), a latent process needs to be represented explicitly anyway, but a log-sum-exp model requires slower, more complex MCMC updates (more latent values, distributions not log concave).

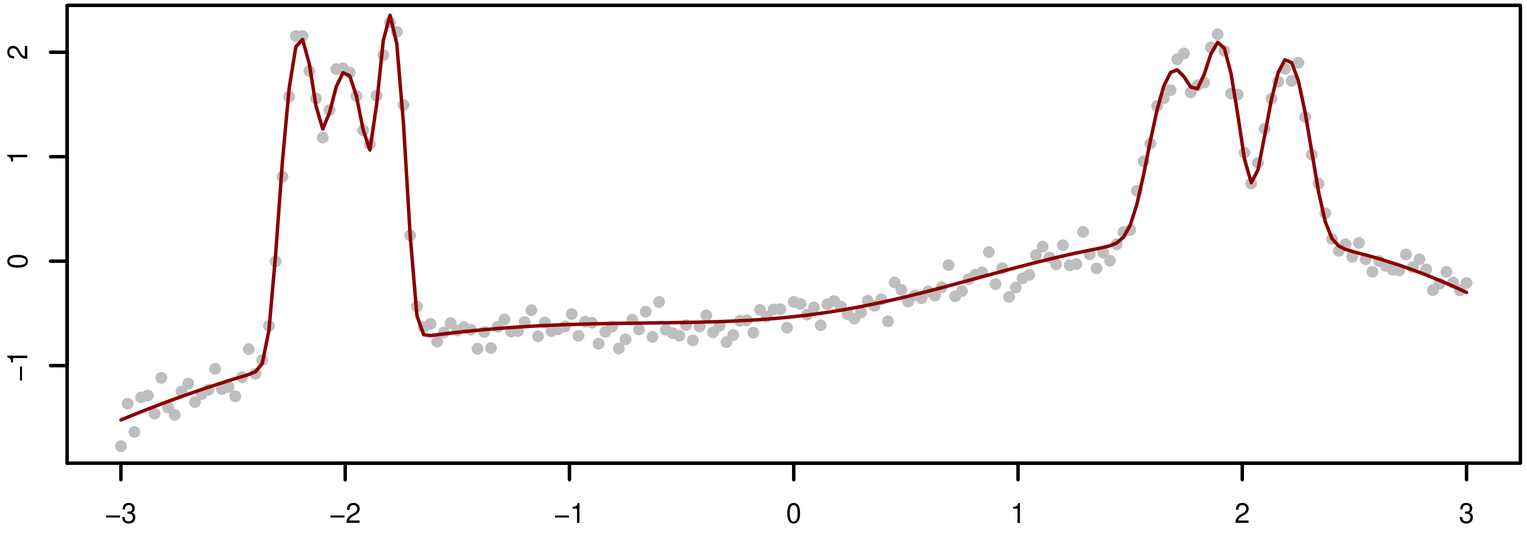
A Test Problem

I tested a log-sum-exp model on data from a sum of Gaussian density functions:

$$\begin{aligned} f(t) \;=\;& 3N(t,+2,1) \;+\; 2N(t,-1,1.5) \;+ \\ & N(t,-2.2,0.05) \;+\; N(t,-1.8,0.04) \;+\; N(t,-2.0,0.07) + \\ & N(t,+1.7,0.08) \;+\; N(t,+1.9,0.06) \;+\; N(t,+2.2,0.07) \end{aligned}$$

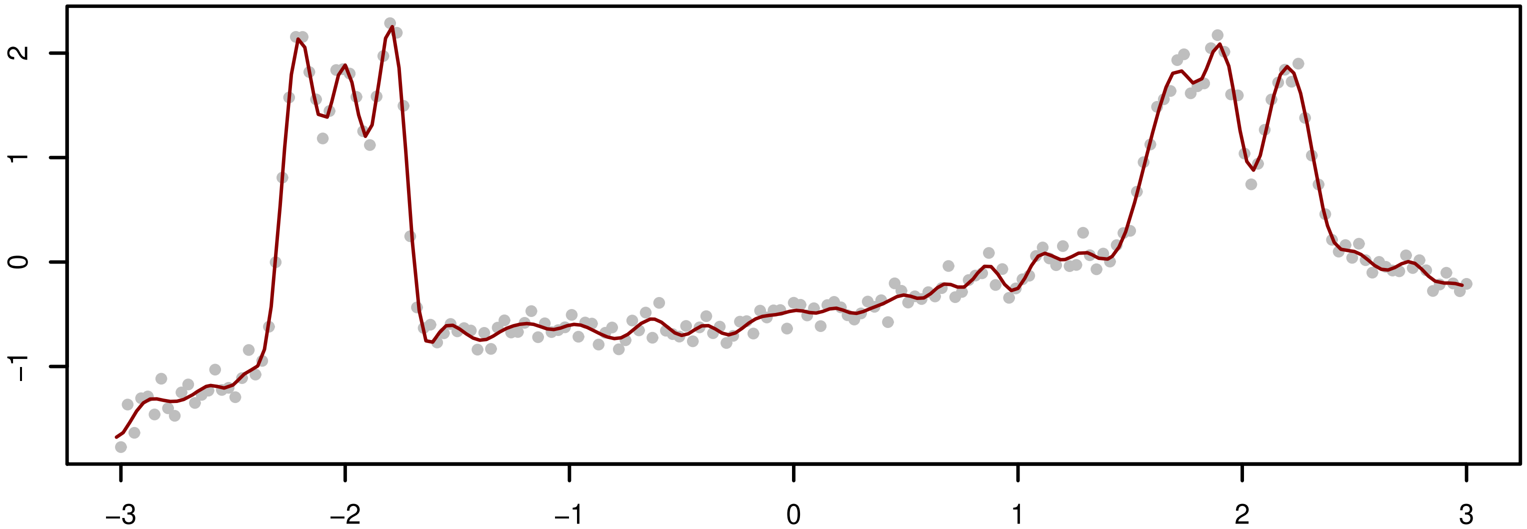
where $N(x, m, s)$ is the Gaussian density function with mean m and standard deviation s. I then took the log of this function and added Gaussian noise with standard deviation 0.1.

Here are 201 training points (on a grid from -3 to 3), with the noise-free function:

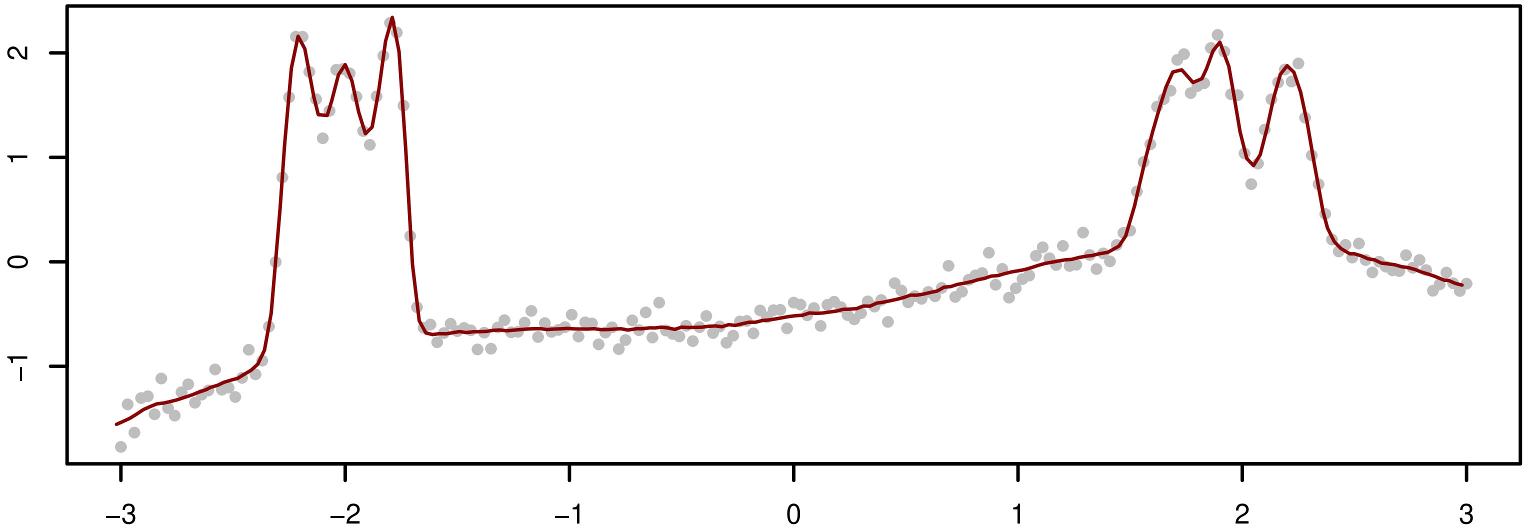


Results on the Test Problem

Fit of a plain GP model with covariance function $\eta \exp(-\rho(t-t')^2)$:



Fit of a log-sum-exp model, two latent processes, covariances $\eta_1 \exp(-\rho_1(t-t')^2)$ and $\eta_2 \exp(-\rho_2(t-t')^2) + \eta_3 \exp(-\rho_3(t-t')^2)$, prior for ρ_3 favouring large values:



A Model for the Toronto Mortality Data

To model the Toronto mortality data, I used as covariates:

– time in days, 10s of days, 100s of days, and 1000s of days

– sine and cosine of "year angle" (for seasonal effects)

– sine and cosine of "week angle" (for day-of-week effects)

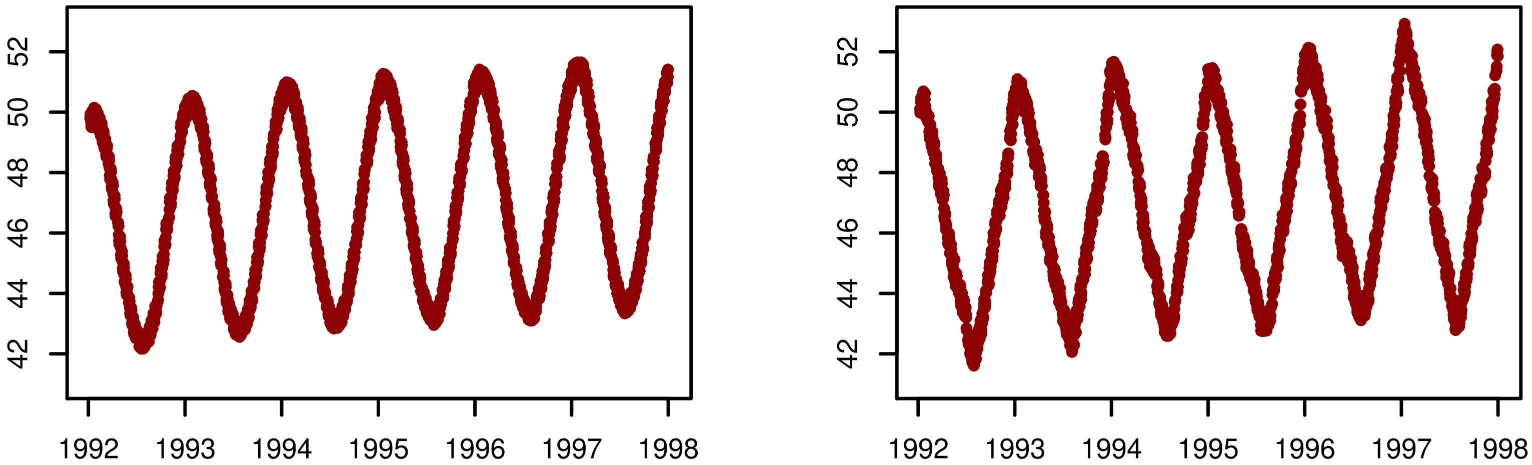
– a binary holiday indicator

Time in 1000s of days and the others were used in both linear and non-linear covariance terms. The four time scales were used in additional terms intended for modeling autocorrelations at various scales.

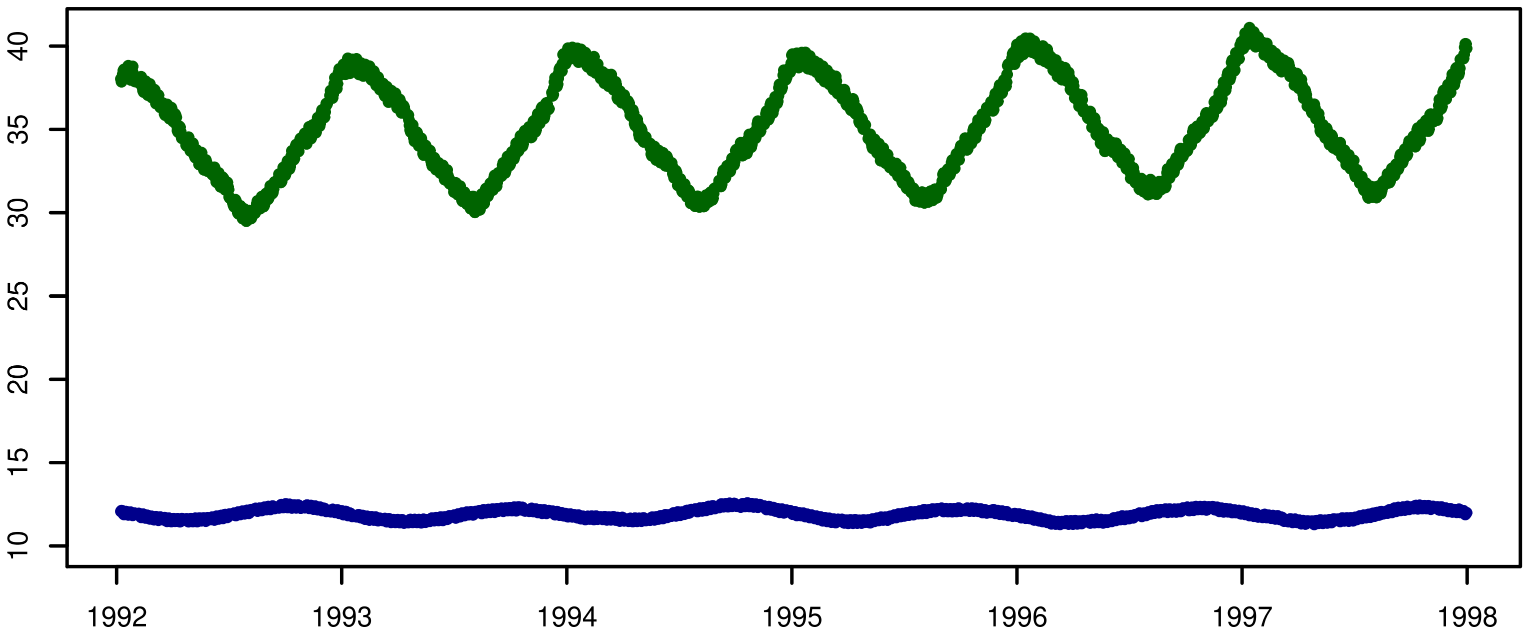
For the log-sum-exp model, two such processes were used, with the same form of covariance and the same priors, but with separate hyperparameters, so they can do different things in the end.

Results on the Toronto Mortality Data

Here are samples from the posterior for the plain GP (left) and a log-sum-exp model (right). The log-sum-exp model does better on a held-out validation set.



Here are the components of this sample from the log-sum-exp model:



Future Research

- Improve MCMC sampling for GP models, especially of the log-sum-exp sort:
  – Metropolis updates while "dragging" latent variables (Andriy is trying this)
  – Hamiltonian Monte Carlo updates of latent variables (while dragging?)

- Try log-sum-exp models for Toronto mortality data with weather and pollution covariates. Hypothesis: These factors affect only a vulnerable subset of the population.

- Try log-sum-exp models for general regression problems. Need to introduce an additional scale factor: $X(t) = \eta \log(\exp(A(t) + \exp(B(t)) + \exp(C(t)))$.

- Investigate more general ways of combining several Gaussian processes. It's much like having a second hidden layer in a multilayer perceptron network.