Gaussian Process

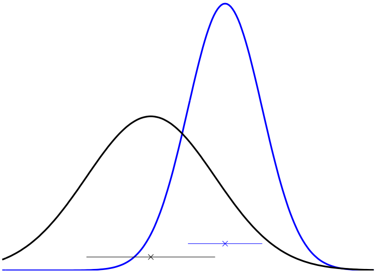
Carl Edward Rasmussen

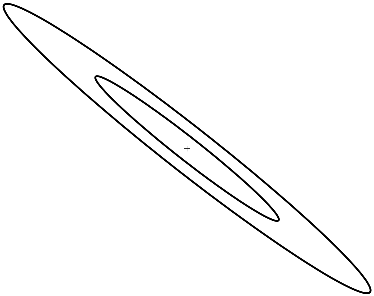
October 10th, 2016

Key concepts

- generalize: scalar Gaussian, multivariate Gaussian, Gaussian process
- Key insight: functions are like infinitely long vectors
- Surprise: Gaussian processes are practical, because of
    - the marginalization property
- generating from Gaussians
    - joint generation
    - sequential generation

The Gaussian Distribution





The univariate Gaussian distribution is given by

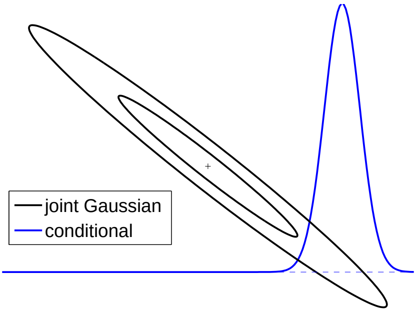
$$p(x|\mu, \sigma^2) \;=\; (2\pi\sigma^2)^{-1/2} \exp\big(-\frac{1}{2\sigma^2}(x-\mu)^2\big)$$

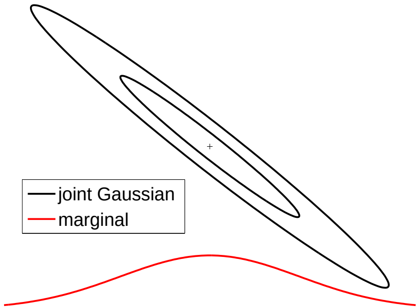
The multivariate Gaussian distribution for D-dimensional vectors is given by

$$p(\mathbf{x}|\mu, \Sigma) \;=\; \mathcal{N}(\mu, \Sigma) \;=\; (2\pi)^{-D/2}|\Sigma|^{-1/2} \exp\big(-\tfrac{1}{2}(\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu)\big)$$

where μ is the mean vector and Σ the covariance matrix.

Conditionals and Marginals of a Gaussian, pictorial





Both the conditionals $p(x|y)$ and the marginals $p(x)$ of a joint Gaussian $p(x, y)$ are again Gaussian.

Conditionals and Marginals of a Gaussian, algebra

If $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian

$$p(\mathbf{x}, \mathbf{y}) \;=\; p\Big(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}\Big) \;=\; \mathcal{N}\Big(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}\Big),$$

we get the marginal distribution of $\mathbf{x}$, $p(\mathbf{x})$ by

$$p(\mathbf{x}, \mathbf{y}) \;=\; \mathcal{N}\Big(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}\Big) \implies p(\mathbf{x}) \;=\; \mathcal{N}(\mathbf{a},\; A),$$

and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ by

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\Big(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}\Big) \implies p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{a} + BC^{-1}(\mathbf{y} - \mathbf{b}),\; A - BC^{-1}B^\top),$$

where $\mathbf{x}$ and $\mathbf{y}$ can be scalars or vectors.

What is a Gaussian Process?

A *Gaussian process* is a generalization of a multivariate Gaussian distribution to infinitely many variables.
Informally: infinitely long vector $\simeq$ function

> **Definition:** *a Gaussian process is a collection of <u>random variables</u>, any finite number of which have (consistent) Gaussian distributions.* $\square$

A Gaussian distribution is fully specified by a mean vector, μ, and covariance matrix Σ:

$$\mathbf{f} = (f_1, \ldots, f_N)^\top \sim \mathcal{N}(\mu, \Sigma), \quad \text{indexes } n = 1, \ldots, N$$

A Gaussian process is fully specified by a mean <u>function</u> $m(x)$ and covariance <u>function</u> $k(x, x')$:

$$f \sim \mathcal{GP}(m, k), \quad \text{indexes: } x \in \mathcal{X}$$

here f and m are functions on $\mathcal{X}$, and k is a function on $\mathcal{X} \times \mathcal{X}$

The marginalization property

Thinking of a GP as a Gaussian distribution with an infinitely long mean vector and an infinite by infinite covariance matrix may seem impractical…

…luckily we are saved by the *marginalization property*:

Recall:

$$p(\mathbf{x}) \;=\; \int p(\mathbf{x},\mathbf{y})d\mathbf{y}.$$

For Gaussians:

$$p(\mathbf{x},\mathbf{y}) \;=\; \mathcal{N}\Big(\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix},\,\begin{bmatrix}A & B\\ B^\top & C\end{bmatrix}\Big) \;\Longrightarrow\; p(\mathbf{x}) \;=\; \mathcal{N}(\mathbf{a},\; A),$$

which works irrespective of the size of $\mathbf{y}$.

Key: only ever ask finite dimensional questions about functions.

Random functions from a Gaussian Process

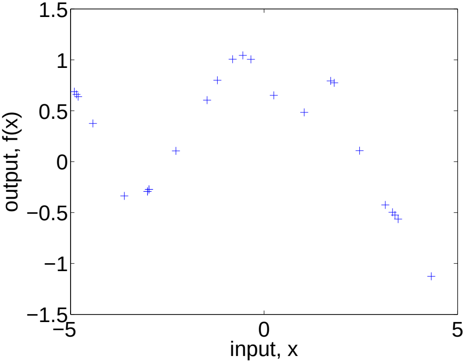
Example one dimensional Gaussian process:

$$p(f) \sim \mathcal{GP}\big(m,\ k\big), \quad \text{where} \quad m(x) = 0, \quad \text{and} \quad k(x, x') = \exp(-\tfrac{1}{2}(x - x')^2).$$

To get an indication of what this distribution over functions looks like, focus on a finite subset of function values $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_N))^\top$, for which

$$\mathbf{f} \sim \mathcal{N}(0, \Sigma), \quad \text{where} \quad \Sigma_{ij} = k(x_i, x_j).$$

Then plot the coordinates of f as a function of the corresponding x values.



Joint Generation

random vector y of dimension D is
Gaussian-distributed as N(y: m, K)

To generate a random sample from a D dimensional joint Gaussian with covariance matrix K and mean vector **m**: (in octave or matlab)

```
z = randn(D,1);
y = chol(K)'*z + m;
```

where `chol` is the Cholesky factor R such that $R^\top R = K$.
Thus, the covariance of **y** is:

$$\mathbb{E}[(\mathbf{y}-\mathbf{m})(\mathbf{y}-\mathbf{m})^\top] \;=\; \mathbb{E}[R^\top \mathbf{z}\mathbf{z}^\top R] \;=\; R^\top \mathbb{E}[\mathbf{z}\mathbf{z}^\top] R \;=\; R^\top I R \;=\; K.$$

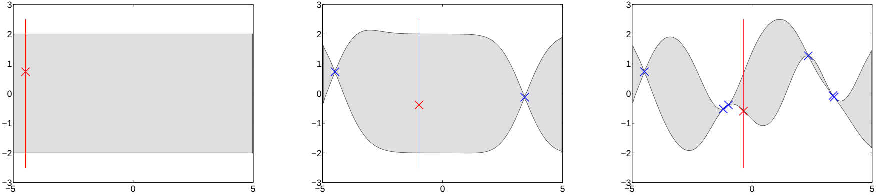
Sequential Generation

Factorize the joint distribution

$$p(f_1, \ldots, f_N | x_1, \ldots x_N) \;=\; \prod_{n=1}^{N} p(f_n | f_{n-1}, \ldots, f_1, x_n, \ldots, x_1),$$

and generate function values sequentially. For Gaussians:

$$\begin{aligned} p(f_n, f_{<n}) \;&=\; \mathcal{N}\Big(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}\Big) \;\Longrightarrow \\ p(f_n | f_{<n}) \;&=\; \mathcal{N}(\mathbf{a} + BC^{-1}(\underline{f_{<n}} - \mathbf{b}),\; A - BC^{-1}B^\top). \end{aligned}$$



Function drawn at random from a Gaussian Process with Gaussian covariance