Gaussian Process Regression and Emulation

STAT8810, Fall 2017

M.T. Pratola

August 28, 2017

Today

More on GP Regression

Constructing Correlation Functions

- How can we systematically construct valid correlation functions?

Theorem: (Bochner)

If $f(\omega)$ is any p.d.f. on $\mathbb{R}^d$ that is symmetric about the origin (zero), then,

$$R(h) = \int_{\omega} cos(h^T \omega) f(\omega) d\omega$$

is a valid correlation function.

Example

- $d = 1, \chi \subseteq \mathbb{R}^1$.
- $f(\omega)$ is taken to be the Uniform$(-\frac{1}{\theta}, \frac{1}{\theta})$ density.
- then

$$R(h) = \int_{-1/\theta}^{1/\theta} \frac{\theta}{2} cos(h\omega) d\omega = \begin{cases} sin(h/\theta), & h \neq 0 \\ 1, & h = 0 \end{cases}$$

Note: $R(-h) = \frac{sin(-h)/\theta}{-h/\theta} = \frac{-\theta sin(h/\theta)}{-h/\theta} = R(h)$ as required, since $sin(\cdot)$ is an odd function.

Example: Gaussian Correlation

- $d = 1, \chi \subseteq \mathbb{R}^1$.
- $f(\omega)$ is taken to be $N(0, \frac{2}{\theta^2}), \theta > 0$.

$$
\begin{aligned}
R(h) &= \int_{-\infty}^{\infty} cos(h\omega) \frac{\theta}{\sqrt{2\pi}\sqrt{2}} exp\left(-\frac{\theta^2}{2(2)}\omega^2\right) d\omega \\
&= exp(\frac{-h^2}{\theta})
\end{aligned}
$$

(Abromowitz and Stegun, 1972, pg.302 eq.7.4.6).

- θ is a *scale* or *length* parameter
- as $\theta \to \infty$ then $exp(-\frac{h^2}{\theta^2}) \to 1$ which implies highly correlated or smoother paths.
- This is called the Gaussian, or squared exponential, correlation function.

Alternative form of Gaussian Correlation

- An alternative parameterization is $exp(-\theta h^2)$ where now θ is intepreted as a *roughness* parameter since $\theta \to \infty$ implies $exp(-\theta h^2) \to 0$.
- ρ^{h^2} (i.e. $\rho = exp(-\theta)$) where one thinks of ρ as correlation scale since $0 \leq \rho \leq 1$ and $h^2 = 1$ implies $Cor(Z(x), Z(x + h)) = \rho$.
- A GP with Gaussian correlation function is a continuous and infinitely differentiable process.

Example: Power Exponential Correlation

- $d = 1, \chi \subseteq \mathbb{R}^1$.
- $R(h) = exp(-\theta||h||^p), 0 < p \leq 2$.
- if $p = 2$: Gaussian correlation function
- if $p = 1$: $Z(x)$ is Ohrenstein-Uhlenbeck process - continuous, nowhere differentiable
- $R(0) = 1, R(-h) = R(h)$ (easy)
- Harder to show non-negative definite property
- For $0 < p < 2$, $R(h)$ is continuous at $h = 0$ but is not differentiable at $h = 0$, and process is continuous but nowhere differentiable:

$$R'(h) = \begin{cases} -\frac{\theta h^p pexp(-\theta h^p)}{h}, h > 0 \\ +\frac{\theta h^p pexp(-\theta h^p)}{h}, h < 0 \end{cases}$$

(where $lim_{h \to 0^-} R'(h) \neq lim_{h \to 0^+} R'(h)$).

Example: Matern Correlation

- $d = 1, \chi \subseteq \mathbb{R}^1$.
- $f(\omega|\nu, \theta)$ is taken to be $t_{\nu/\theta}$, $\theta > 0, \nu \in \{1, 2, 3, \ldots\}$.

$$R(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|h|}{\theta} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}|h|}{\theta} \right), h \in \mathbb{R}^1.$$

- K_ν is called the modified Bessel function of order ν.
- $K_\nu(x)$ is the solution of $x^2 y''(x) + xy'(x) - (x^2 + \nu^2)y(x) = 0$.

Example: Matern Correlation

- $K_{1/2}(x) = exp(-x)\sqrt{\pi}\frac{1}{\sqrt{2x}} \Rightarrow R(h) = exp(-\frac{|h|}{\theta})(p = 1)$.
- For $n \in \{1, 2, \ldots\}$,

$$K_{n+1/2}(x) = exp(-x)\sqrt{\frac{\pi}{2x}}\sum_{k=0}^{\infty}\frac{(n+k)!}{k!(n-k)!}\left(\frac{1}{2x}\right)^k$$

- Fact: $R(h|\nu, \theta) \to exp(-\frac{|h|^2}{2\theta^2})$ as $\nu \to \infty$. That is, Matern correlation becomes the Gaussian correlation in the limit.

Example: Matern Correlation

- Typically the Matern is used with specific settings of ν which greatly simplify it's computation:
- $\nu = \frac{1}{2} : R(h) = exp(-\frac{|h|}{\theta})$
- $\nu = \frac{3}{2} : R(h) = \left(1 + \frac{\sqrt{3}|h|}{\theta}\right) exp(-\frac{\sqrt{3}|h|}{\theta})$
- $\nu = \frac{5}{2} : R(h) = \left(1 + \frac{\sqrt{5}|h|}{\theta} + \frac{5|h|^2}{3\theta^2}\right) exp(-\frac{\sqrt{5}|h|}{\theta})$
- Realizations are almost surely $\lceil \nu \rceil - 1$ times differentiable

Example: Cubic Correlation

- $d = 1, \chi \in \mathbb{R}^1$. Fix $\theta > 0$.

$$R(h|\theta) = \begin{cases} 1 - 6(\frac{h}{\theta})^2 + 6(\frac{|h|}{\theta})^3, |h| \leq \frac{\theta}{2}\dagger \\ 2(1 - \frac{|h|}{\theta})^3, \frac{\theta}{2} < |h| < \theta \\ 0, |h| > \theta. \end{cases}$$

$\dagger$ (i.e. $-\frac{1}{2} \leq \frac{h}{\theta} \leq \frac{1}{2}$)

- This means that when x_1, x_2 are a distance greater than θ apart, $Z(x_1), Z(x_2)$ are uncorrelated. Indeed, since we are using GP's, they are independent.
- Realizations are continuous and differentiable.

Simulating Draws from a GP

- Suppose $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ are i.i.d. Normal with mean 0 and variance 1.
- Suppose $\mathbf{L}$ is an $n \times n$ lower-triangular matrix of real numbers of full rank and μ is an $n \times 1$ vector of real numbers.
- Then $\mathbf{Y} = (Y_1, \ldots, Y_n)^T = \mathbf{LZ} + \mu$ has a MVN distribution with mean μ and covariance matrix $\Sigma = \mathbf{LL}^T$.
- Check:

$$
\begin{aligned}
E[\mathbf{Y}] &= E[\mathbf{LZ} + \mu] = \mu \\
Cov(\mathbf{Y}) &= E[(\mathbf{Lz} + \mu - \mu)(\mathbf{LZ} + \mu - \mu)] \\
&= E[\mathbf{LZZ}^T\mathbf{L}^T] \\
&= \mathbf{LI}_n\mathbf{L}^\mathbf{T} \\
&= \mathbf{LL}^T
\end{aligned}
$$

Simulating Draws from a GP

To generate samples from a realization of a GP, we work backwards:

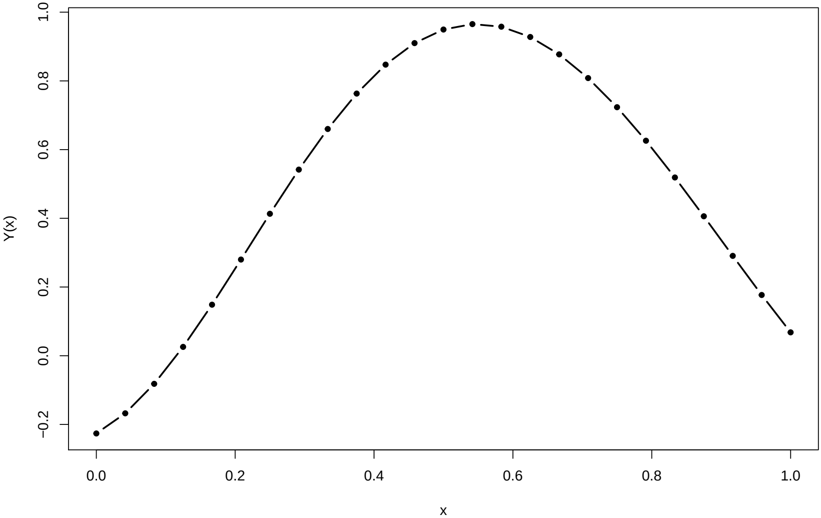
1. Form the $n \times n$ covariance matrix $\Sigma = cov(\mathbf{Y})$ according to your desired variance and desired correlation function $c(\cdot)$.
2. Find $\mathbf{L}$. $\mathbf{L}\mathbf{L}^T = \Sigma^{1/2}\Sigma^{1/2} = \Sigma$ so take $\mathbf{L} = \text{chol}(\Sigma)$.
3. Generate $\mathbf{Z} \sim N(0, \mathbf{I}_n)$ from a random number generator.
4. Calculate $\mathbf{Y} = \mathbf{L}\mathbf{Z} + \mu$. Then $\mathbf{Y}$ is a vector of observations taken from a realization of a GP with the desired (constant) mean function μ and desired correlation function $c(\cdot)$.

1D Example

```
set.seed(88)
n=25
x=seq(0,1,length=n)
X=abs(outer(x,x,"-"))
rho=0.1
R=rho^(X^2)
L=t(chol(R+diag(n)*.Machine$double.eps*100))
mu=0
Z=rnorm(n,mean=0,sd=1)
Y=L%*%Z+mu
```

1D Example

```
plot(x,Y,xlab="x",ylab="Y(x)",type='b',lwd=2,pch=20)
```

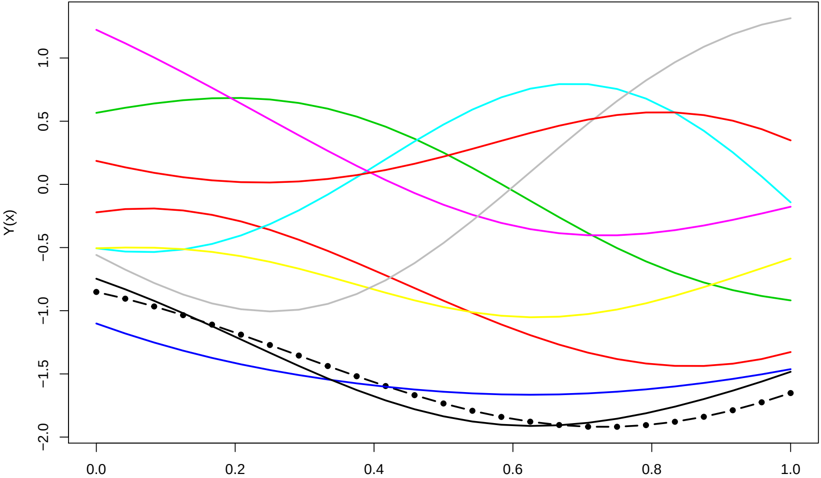


1D Example

```r
m=10
Ymat=matrix(0,nrow=m,ncol=n)
for(i in 1:m) {
    Z=rnorm(n,mean=0,sd=1)
    Ymat[i,]=L%*%Z+mu
}
```

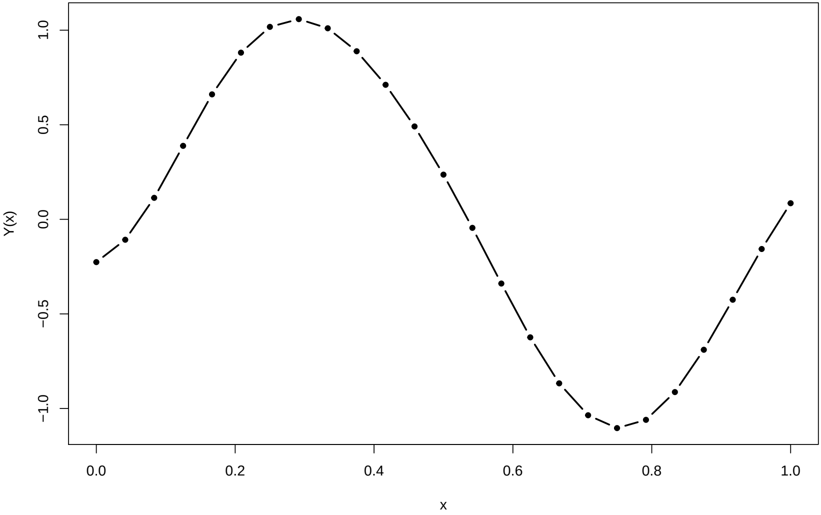
1D Example

```
plot(x,Ymat[1,],xlab="x",ylab="Y(x)",type='b',
     lwd=2,pch=20,ylim=range(Ymat))
for(i in 2:m) lines(x,Ymat[i,],col=i,lwd=2)
```



1D Example

```
set.seed(88)
rho=1e-4
```



Forming Valid Covariance or Correlation Functions

- Assume $c_i(h), R_i(h)$ are valid correlation functions (symmetric, non-negative definite, $R(0) = 1$.)

1. $c(h) = c_1(h) + c_2(h)$ is a valid covariance function

Eg: if $Z_1 \sim N(0, c_1(h))$ and $Z_2 \sim N(0, c_2(h))$ and $Z_1 \perp Z_2$ then for $Z = Z_1 + Z_2$, $Cov(Z)$ is $c_1(h) + c_2(h)$.

Forming Valid Covariance or Correlation Functions

- Assume $c_i(h), R_i(h)$ are valid correlation functions (symmetric, non-negative definite, $R(0) = 1$.)

2. $c(h) = c_1(h)c_2(h)$ is a valid covariance function and $R(h) = R_1(h)R_2(h)$ is a valid correlation function.

Eg: if Z_1, Z_2 are independent with mean 0 and variance σ^2 then $Z(x) = Z_1(x)Z_2(x)$ has mean $E[Z_1(x)Z_2(x)] = E[Z_1(x)]E[Z_2(x)] = 0$ and

$$\begin{aligned}
Cov(Z(x), Z(x+h)) &= Cov(Z_1(x)Z_2(x), Z_1(x+h)Z_2(x+h)) \\
&= E[Z_1(x)Z_1(x+h)Z_2(x)Z_2(x+h) - 0] \\
&= E[Z_1(x)Z_1(x+h)]E[Z_2(x)Z_2(x+h)](indep.) \\
&= c_1(h)c_2(h)
\end{aligned}$$

Forming Valid Covariance or Correlation Functions

- Assume $c_i(h), R_i(h)$ are valid covariance or correlation functions (symmetric, non-negative definite, $R(0) = 1$.)

3. If $0 < \alpha < 1$, $c(h) = \alpha c_1(h) + (1 - \alpha) c_2(h)$ is a valid covariance function, and $R(h) = \alpha R(h) + (1 - \alpha) R(h)$ is a valid covariance function.

Similarly, for $\alpha_1, \ldots, \alpha_n$ where $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$ then $c(h) = \sum_i \alpha_i c_i(h)$ is a valid covariance function and $R(h) = \sum_i \alpha_i R_i(h)$ is a valid correlation function.

Forming Valid Covariance or Correlation Functions

- Assume $c_i(h), R_i(h)$ are valid correlation functions (symmetric, non-negative definite, $R(0) = 1$.)

4. If $\{R(h;\theta)\}_{\theta\in\Theta}$ are valid, or $\{c(h;\theta)\}_{\theta\in\Theta}$ are valid and $\pi(\theta)$ is a p.d.f., then

$$c(h) = \int_\theta c(h;\theta)\pi(\theta)d\theta$$

and

$$R(h) = \int_\theta R(h;\theta)\pi(\theta)d\theta$$

are valid.

Forming Valid Covariance or Correlation Functions

- Assume $c_i(h), R_i(h)$ are valid correlation functions (symmetric, non-negative definite, $R(0) = 1$.)

5. A correlation function is said to be *separable* if

$$R(h) = \prod_{i=1}^{d} R_i(h).$$

A popular choice is the separable Gaussian model,

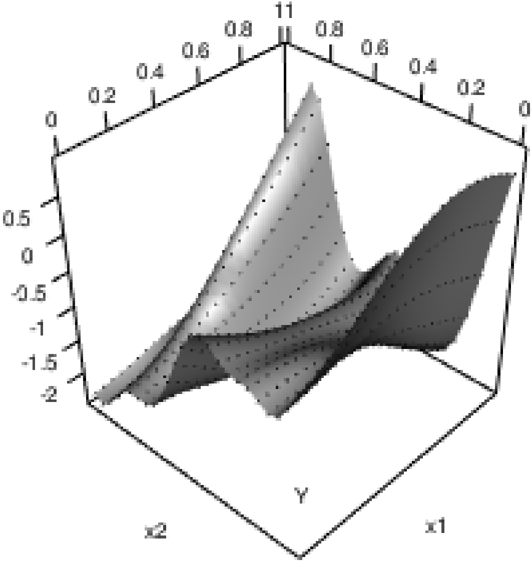
$$R(h) = \prod_{i=1}^{d} exp(-\theta_i h_i^2)$$

where $h_i = ||x_i - x_i'||$ and $\mathbf{x} = (x_1, \ldots, x_d)$.

2D Example

```
set.seed(88)
n=25
x=as.matrix(expand.grid(seq(0,1,length=n),
            seq(0,1,length=n)))
X=abs(outer(x[,1],x[,1],"-"))
rho=0.3
R=rho^(X^2)
X=abs(outer(x[,2],x[,2],"-"))
rho=1e-15
R=R*rho^(X^2)
L=t(chol(R+diag(n^2)*.Machine$double.eps*100))
mu=0
Z=rnorm(n^2,mean=0,sd=1)
Y=L%*%Z+mu
```

2D Example



Kronecker Product Covariances

- Often in emulation problems, the computer code output may be calculated on a regular grid:

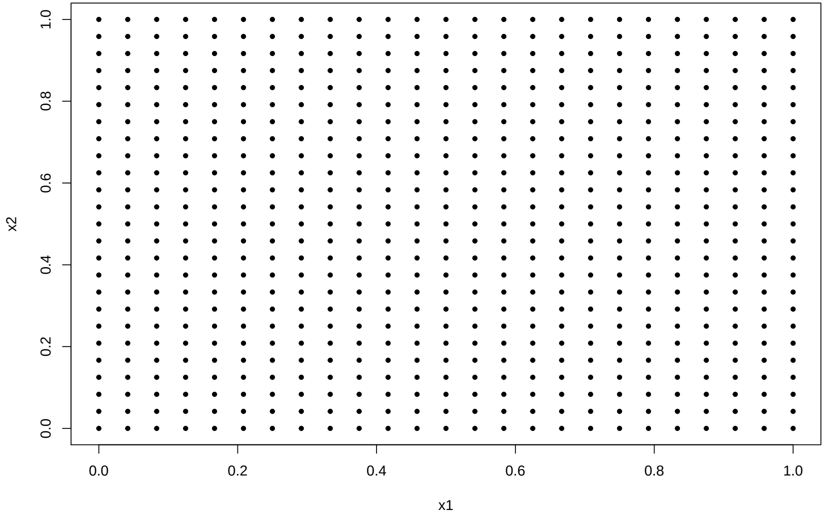
```
X=as.matrix(expand.grid(seq(0,1,length=25),
           seq(0,1,length=25)))
dim(X)
```

```
## [1] 625   2
```

- This is obviously problematic: here the number of “pixels” making up our output are growing like 25^d.

Kronecker Product Covariances

```
plot(X,pch=20,xlab="x1",ylab="x2")
```



Kronecker Product Covariances

- Such cases can be simplified using the Kronecker product.

```
set.seed(88)
n=25
x1=seq(0,1,length=n)
x2=seq(0,1,length=n)

X1=abs(outer(x1,x1,"-"))
rho=0.3
R1=rho^(X1^2)
X2=abs(outer(x2,x2,"-"))
rho=1e-15
R2=rho^(X2^2)
RR=R2%x%R1 #kronecker product
sum(abs(RR-R))
```

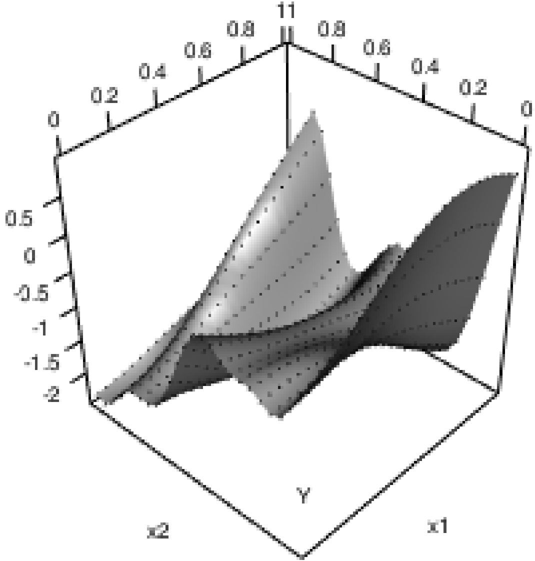
```
## [1] 0
```

Kronecker Product Covariances

```
LL=t(chol(R2+diag(n)*.Machine$double.eps*100))%x%
    t(chol(R1+diag(n)*.Machine$double.eps*100))
mu=0
Z=rnorm(n^2,mean=0,sd=1)
Y=LL%*%Z+mu

par3d(cex=0.5)
persp3d(matrix(Y,n,n),col="grey",xlab="x1",ylab="x2",
        zlab="Y",box=FALSE)
plot3d(x[,1],x[,2],Y,col="black",type='s',radius=0.01,
        add=TRUE)
rgl.snapshot("kronecker.png")
```

Kronecker Product Covariances



Kronecker Product Covariances

- Other properites that may be useful:
  - $A \otimes (B + C) = A \otimes B + A \otimes C$
  - $A \otimes B \neq B \otimes A$ (in general)
  - $A \otimes (B \otimes C) = (A \otimes B) \otimes C$
  - $\alpha A \otimes \beta B = \alpha\beta(A \otimes B)$
  - $(A \otimes B)^T = A^T \otimes B^T$
  - $(A \otimes B)(C \otimes D) = AC \otimes BD$
  - $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$
  - $\text{rank}(A \otimes B) = \text{rank}(A)\text{rank}(B)$
  - $\det(A \otimes B) = \det(A)^{\text{rank}(B)}\det(B)^{\text{rank}(A)}$

Kronecker Product Covariances

- Using this trick is one way to getting around manipulating and storing large correlation matrices so that we can use the GP model on moderately sized datasets.
- We will see some other tricks later.
- These tricks really only get us so far.