Lecture 12
Kernel Methods and Poisson Processes

Dennis Sun
Stanford University
Stats 253

July 29, 2015

Kernel Methods

Kernel methods are popular in machine learning.

The idea is that many methods only depend on inner products between observations.

Example (ridge regression): Like OLS, but coefficients are shrunk towards zero:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{ridge} &= \text{argmin}_{\boldsymbol{\beta}} \, ||\mathbf{y} - X\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||^2. \\ &= (X^T X + \lambda I)^{-1} X^T \mathbf{y}\end{aligned}$$

Suppose I want to predict y_0 at a new point $\mathbf{x}_0$. Ridge prediction is

$$\begin{aligned}\hat{y}_0 &= \mathbf{x}_0^T (X^T X + \lambda I_p)^{-1} X^T \mathbf{y} \\ &= \mathbf{x}_0^T X^T (X X^T + \lambda I_n)^{-1} \mathbf{y} \\ &= K_{01} (K_{11} + \lambda I)^{-1} \mathbf{y}\end{aligned}$$

where $K_{01} = \begin{pmatrix} \langle \mathbf{x}_0, \mathbf{x}_1 \rangle & \cdots & \langle \mathbf{x}_0, \mathbf{x}_n \rangle \end{pmatrix}$ and $(K_{11})_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$.

Kernel Methods

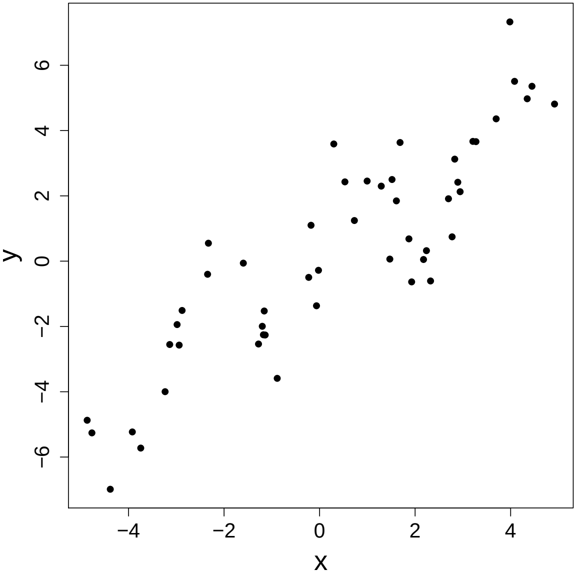
For now, the **kernel** is the *linear kernel*: $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$.

We can replace the linear kernel by any positive-semidefinite kernel, such as

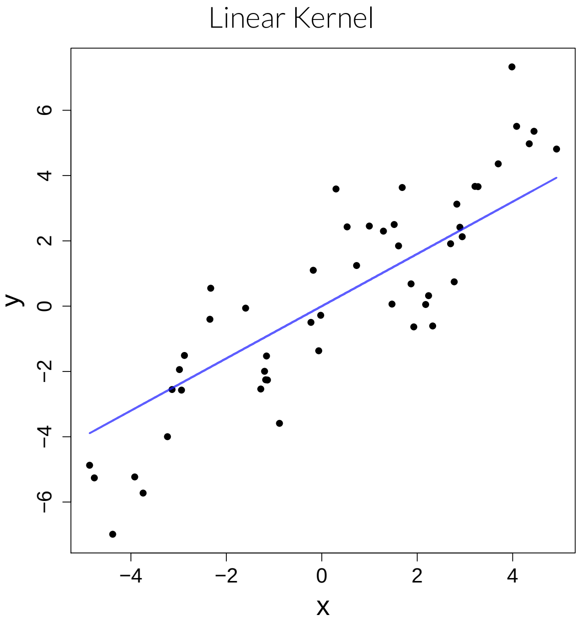
$$K(\mathbf{x}, \mathbf{x}') = \theta_1 e^{-\theta_2 ||\mathbf{x} - \mathbf{x}'||^2}.$$

What will this do?

Kernel Regression: A Simple Example

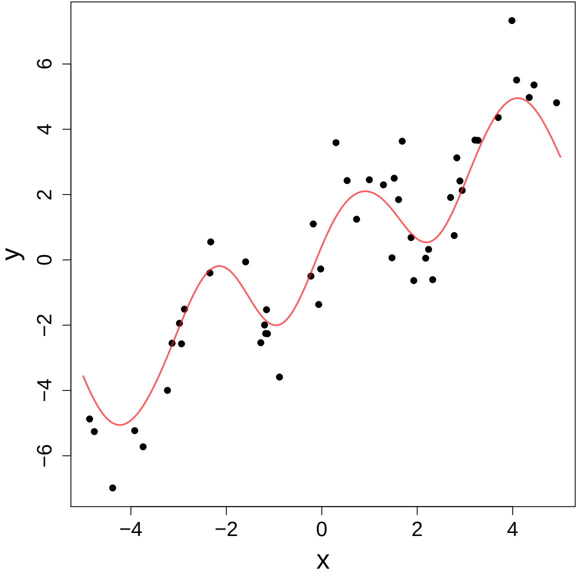


Kernel Regression: A Simple Example



Kernel Regression: A Simple Example

$$K(x, x') = e^{-(x-x')^2}$$



Why does kernel regression work?

TWO INTERPRETATIONS:

1. **Implicit basis expansion:** Since the kernel is positive semidefinite, we can do an eigendecomposition:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{\infty} \lambda_i \varphi_i(\mathbf{x}) \varphi_i(\mathbf{x}').$$

   We can think of this informally as:

$$K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle,$$

   where $\Phi(\mathbf{x}) \stackrel{def}{=} \left(\sqrt{\lambda_1}\varphi_1(\mathbf{x}) \quad \sqrt{\lambda_2}\varphi_2(\mathbf{x}) \quad \cdots \right)$.

2. **Representer theorem:** The solution $\hat{f}$ satisfies

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

   for some coefficients α_i.

Illustration of the Representer Theorem

$$K(x, x_{30}) = e^{-(x-x_{30})^2}$$

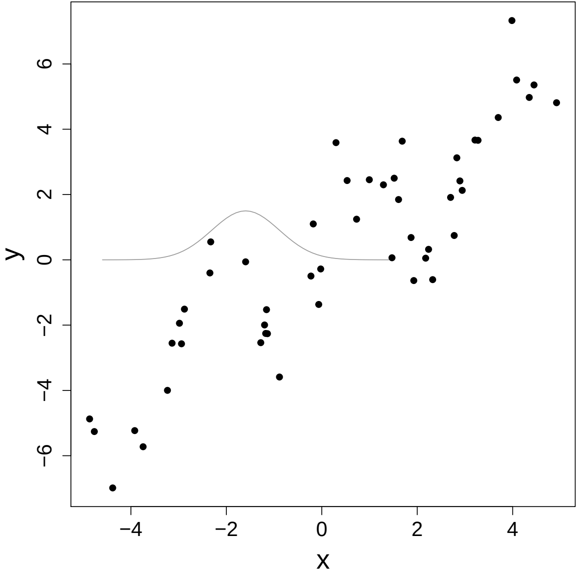


Illustration of the Representer Theorem

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i e^{-(x-x_i)^2}$$

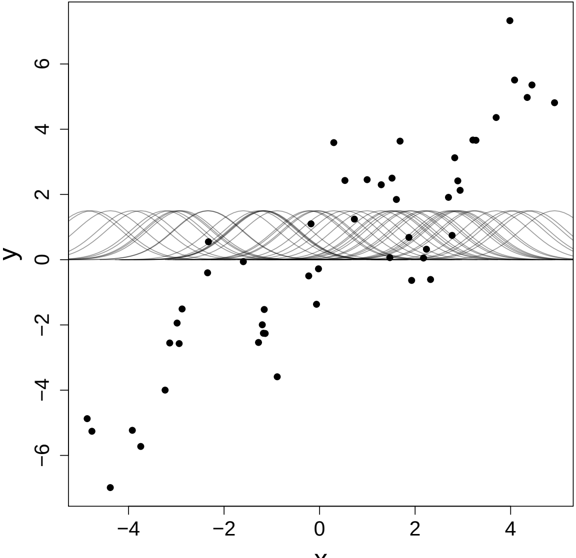
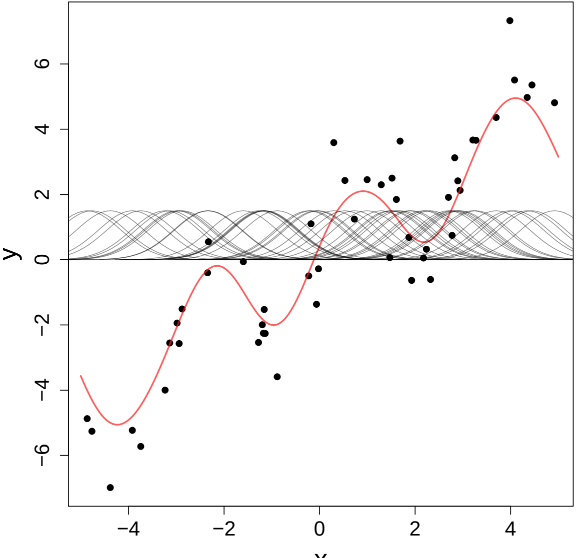


Illustration of the Representer Theorem

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i e^{-(x-x_i)^2}$$



Kernel Methods and Kriging

Hopefully, you've realized by now that a positive semidefinite "kernel" is just a covariance function.

The prediction equation

$$\hat{y}_0 = K_{01}(K_{11} + \lambda I)^{-1}\mathbf{y}$$

should remind you of the kriging prediction

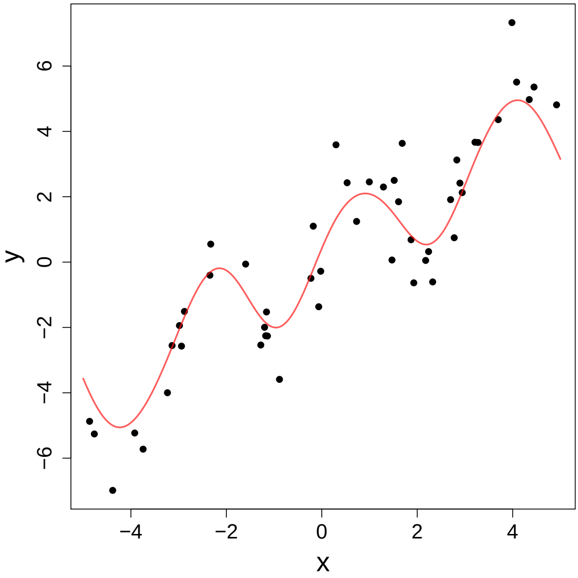
$$\hat{y}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}^{GLS} + \Sigma_{01}\Sigma_{11}^{-1}(\mathbf{y} - X\hat{\boldsymbol{\beta}}^{GLS}).$$

If we do not have any predictors (the only information is the spatial coordinates), then this reduces to the kernel predictions (without the λI.

$$\hat{y}_0 = \Sigma_{01}\Sigma_{11}^{-1}\mathbf{y}.$$

Spatial Interpretation of Kernel Regression

$$\Sigma(x, x') = e^{-(x-x')^2}$$



1. Kernel Methods
2. Gaussian Processes
3. Point Processes
4. Course Logistics

Gaussian Process Model

Suppose the function f is random, sampled from a Gaussian process with mean 0 and covariance function $\Sigma(\mathbf{x}, \mathbf{x}')$.

We observe noisy observations $y_i = f(\mathbf{x}_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \lambda)$.

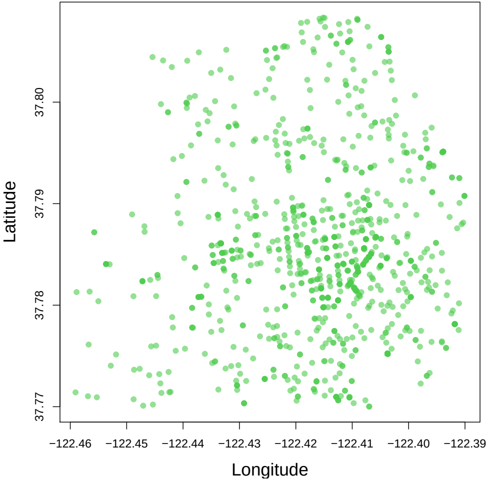
To estimate f, we should use the posterior mean $\hat{f} = \mathrm{E}[f|\mathbf{y}]$.

$$\hat{f}(\mathbf{x}_0) = \Sigma_{01}(\Sigma_{11} + \lambda I)^{-1}\mathbf{y},$$

which again is equivalent to kriging and kernel methods!

Point Processes

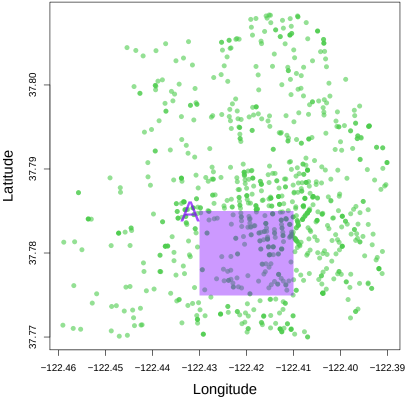
In **point processes**, the locations themselves are random!



Typical Question: Does the process exhibit clustering?

Poisson Point Process

The **(homogeneous) Poisson point process** is a baseline model. It has two defining characteristics:

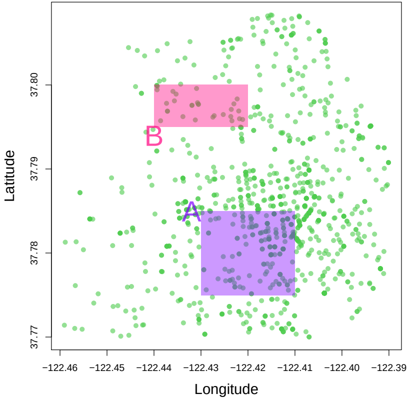


1. The number of events in a given region A is Poisson, with mean proportional to the size of A:

$$N(A) \sim \mathrm{Pois}(\lambda |A|)$$

Poisson Point Process

The **(homogeneous) Poisson point process** is a baseline model. It has two defining characteristics:



2. The number of events in two disjoint regions A and B are independent.

$$N(A) \perp N(B)$$

How would you simulate a Poisson point process?

How to test for homogeneity?

Quadrat test: Divide the domain into disjoint regions. The number in each region should be independent Poisson, so you can calculate

$$X^2 = \sum_r \frac{(O_r - E_r)^2}{E_r}$$

and compare to a χ^2 distribution.

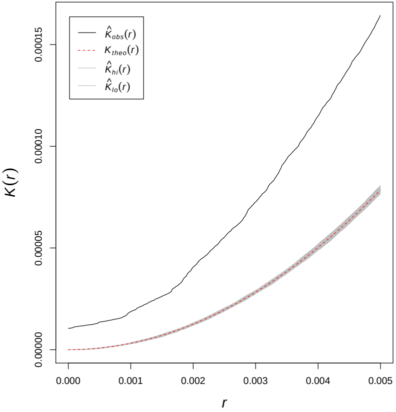
R Code:

```
library(spatstat)
data.ppp <- ppp(data$Longitude, data$Latitude,
           range(data$Longitude), range(data$Latitude))
quadrat.test(data.ppp)

         Chi-squared test of CSR using quadrat counts
         Pearson X2 statistic

 X2 = 863.05, df = 24, p-value < 2.2e-16

 Quadrats: 5 by 5 grid of tiles
```

How to test for homogeneity?

Ripley's K function: Calculate the number of pairs less than a distance r apart for every $r > 0$, giving us a function $K(r)$. Now simulate the distribution of $K^*(r)$ under the null hypothesis and compare it to $K(r)$.

```
plot(envelope(data.ppp, Kest))
```



So you've rejected the null. Now what?

If the process isn't a homogeneous Poisson process, what is it?

It can be an **inhomogeneous Poisson process**. There's some "density" $\lambda(\mathbf{s})$ such that $N(A)$ is Poisson with mean

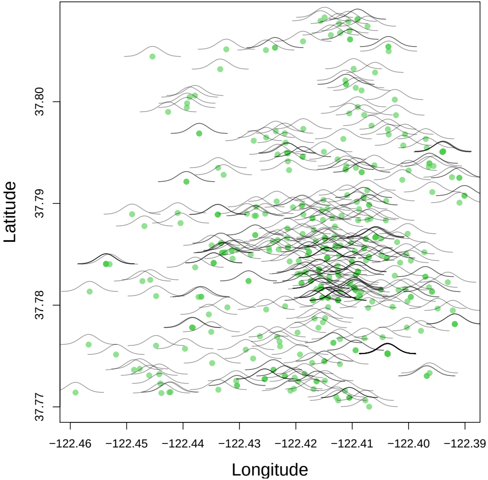
$$\int_A \lambda(\mathbf{s})\, d\mathbf{s}.$$

Note that if $\lambda(\mathbf{s}) \equiv \lambda$, then this reduces to a homogeneous Poisson process.

How do we estimate $\lambda(\mathbf{s})$?

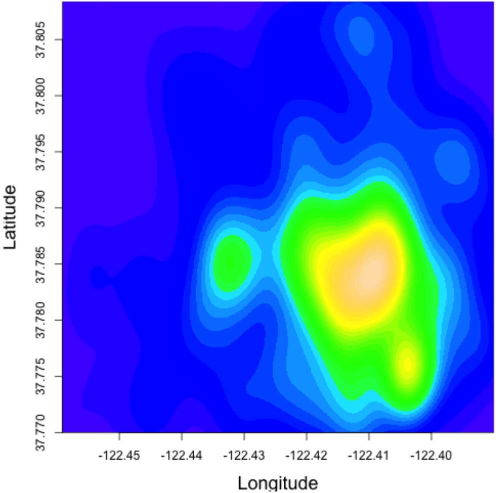
Kernel density estimation! Choose a kernel satisfying

$$\int_D K(\mathbf{s}, \mathbf{s}')\, d\mathbf{s} = 1.$$



How do we estimate $\lambda(s)$?

The estimate is just $\hat{\lambda}(\mathbf{s}) = \sum_{i=1}^{n} K(\mathbf{s}, \mathbf{s}_i)$.



1. Kernel Methods
2. Gaussian Processes
3. Point Processes
4. Course Logistics

Remaining Items

- Homework 3 is due this Friday.
- Please submit project proposals by this Friday as well. I will respond to them over the weekend.
- Please resubmit all quizzes by next Thursday. The graded quizzes can be picked up tomorrow at Jingshu's office hours.
- We may have a guest lecture on Geographic Information Systems (GIS) and mapping next Wednesday. I will keep you posted by e-mail.
- We may or may not have final presentations. If we do, they would be during the scheduled final exam time for this class, which is Friday afternoon.