STAT 518
Update Student Presentation

Wen Wei Loh

April 30, 2013

What's the paper about again?

- Gaussian Processes (GP)
  - Focus on covariance functions between outcomes

  $\rightarrow$ how two outcomes are correlated, based on values of their predictors

  $\rightarrow$ letting the data speak for itself
- Prediction
  - Objective is to obtain a predictive distribution for the outcome of a future observation
- Radford M. Neal [1999]

Yes, there was some math ...

- Simple Linear Regression Example
- Observed data : $\left(x^{(1)}, t^{(1)}\right), \ldots, \left(x^{(n)}, t^{(n)}\right)$
- $t^{(i)}$ = outcome (target) for case i
- $x^{(i)}$ = fixed input (predictor) for case i

$\rightarrow$ univariate for this example, but may be p-dimensional

- Fit a familiar model:

$$t^{(i)} = \alpha + x^{(i)}\beta + \epsilon^{(i)}, \quad \epsilon^{(i)} \underset{iid}{\sim} N\left(0, \sigma_\epsilon^2\right)$$

- Put independent priors on unknown parameters α and β:

$$\alpha \sim N\left(0, \sigma_\alpha^2\right), \quad \beta \underset{iid}{\sim} N\left(0, \sigma_\beta^2\right)$$

Yes, there was some math ...

$\Rightarrow$ Prior joint multivariate Gaussian distribution for t:

$$\boxed{t = \left[t^{(1)}, \ldots, t^{(n)}\right]^T \sim N_n\left(\mathbf{0}, C\right)}$$

Covariance matrix $C = \left\{\mathrm{Cov}\left[t^{(i)}, t^{(j)}\right] \;\; i, j \in [1, n]\right\}$

$$\mathrm{Cov}\left[t^{(i)}, t^{(j)}\right] = \mathrm{E}\left[\left(\alpha + x^{(i)}\beta + \epsilon^{(i)}\right)\left(\alpha + x^{(j)}\beta + \epsilon^{(j)}\right)\right]$$

$$= \sigma_\alpha^2 + x^{(i)}x^{(j)}\sigma_\beta^2 + \delta_{ij}\sigma_\epsilon^2,$$

$\delta_{ij} = 1 \quad$ if $i = j$ and 0 otherwise i.e. Kronecker delta

So why the fuss over the covariance function?

$$C_{ij} = \text{Cov}\left[t^{(i)}, t^{(j)}\right] = \underbrace{\sigma_\alpha^2}_{Constant} + \underbrace{x^{(i)}x^{(j)}\sigma_\beta^2}_{Linear} + \underbrace{\delta_{ij}\sigma_\epsilon^2}_{Noise}$$

- Covariance between the outcomes is fully described by the relationship between the predictors

e.g. Linear function of predictors: restrictive?
Suppose the covariance depends on differences between predictor values, rather than just the values themselves?

$\Rightarrow$ Are there other ways to describe the covariance function?

So why the fuss over the covariance function?

- How about an exponential term instead?

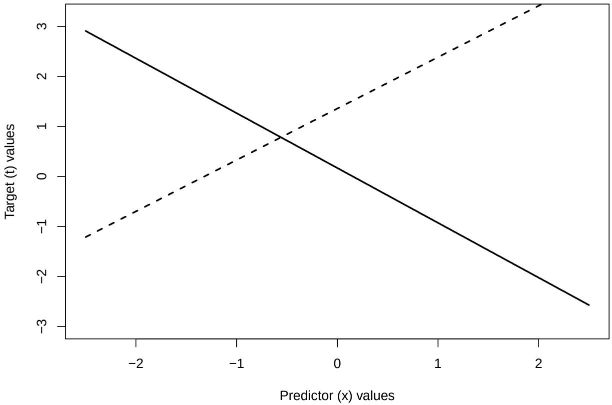
$$C_{ij} = \mathrm{Cov}\left[t^{(i)}, t^{(j)}\right] = \underbrace{\eta^2 \exp\left(-\rho^2 \left(x^{(i)} - x^{(j)}\right)^2\right)}_{\textit{Exponential}} + \underbrace{\delta_{ij}\sigma_\epsilon^2}_{\textit{Noise}}$$

- Observations with predictor values that are "far apart" have much smaller covariances
- η = magnitude
- ρ = scale

$\Rightarrow$ Gaussian kernel: $\exp\left(-\frac{\left(x^{(i)}-x^{(j)}\right)^2}{1/\rho^2}\right)$

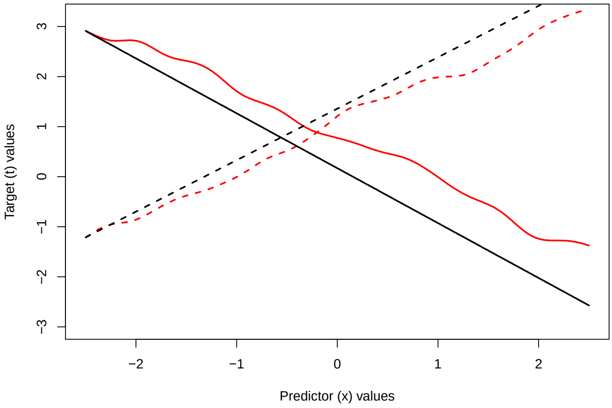
Show me some graphs instead

$$\text{Cov}\left[t^{(i)}, t^{(j)}\right] = 1 + x^{(i)}x^{(j)}$$



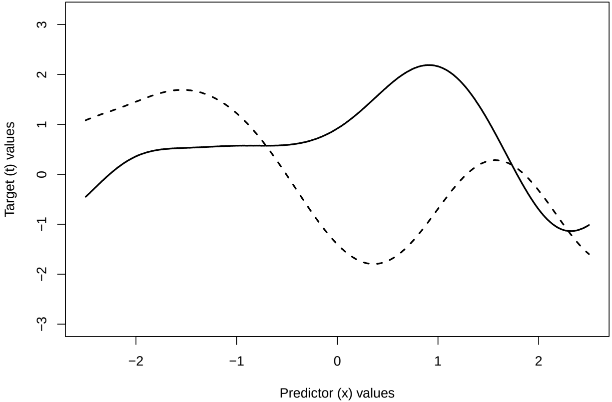
Show me some graphs instead

$$\mathrm{Cov}\left[t^{(i)}, t^{(j)}\right] = 1 + x^{(i)}x^{(j)} + {\color{red}0.1^2 \exp\left(-3^2\left(x^{(i)} - x^{(j)}\right)^2\right)}$$
$$\mathrm{Cov}\left[t^{(i)}, t^{(j)}\right] = 1 + x^{(i)}x^{(j)}$$



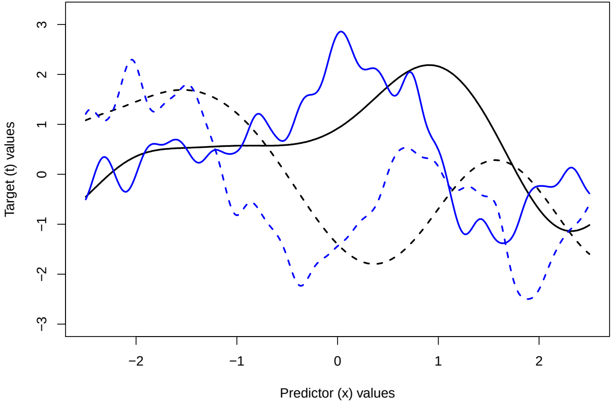
Show me some graphs instead

$$\mathrm{Cov}\left[t^{(i)}, t^{(j)}\right] = \exp\left(-\left(x^{(i)} - x^{(j)}\right)^2\right)$$



Show me some graphs instead

$$\text{Cov}\left[t^{(i)}, t^{(j)}\right] = \exp\left(-\left(x^{(i)} - x^{(j)}\right)^2\right) + \mathbf{0.5^2} \exp\left(-\mathbf{5^2}\left(x^{(i)} - x^{(j)}\right)^2\right)$$
$$\text{Cov}\left[t^{(i)}, t^{(j)}\right] = \exp\left(-\left(x^{(i)} - x^{(j)}\right)^2\right)$$



Enough with the covariance functions ...

- May be constructed with various components e.g. constant, linear, exponential.
- Components of the covariance function may reflect different plausible features of the underlying structure
- Valid covariance function must always result in a positive definite covariance matrix for the targets.
- Different forms of the covariance function $\mathrm{Cov}\left[t^{(i)}, t^{(j)}\right]$ define infinitely many flexible regression models

Step back to prediction

- Targets still have a multivariate Gaussian distribution

$$t = \left[t^{(1)}, \ldots, t^{(n)}\right]^T \sim N_n(\mathbf{0}, C)$$

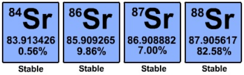
- $C_{ij} = \mathrm{Cov}\left[t^{(i)}, t^{(j)}\right]$
- Given the inputs for a new case $x^{(n+1)}$, the predictive distribution for the new outcome $t^{(n+1)}$ is Gaussian:

$$\boxed{\begin{aligned} \mathrm{E}\left[t^{(n+1)} \mid t^{(1)}, \ldots, t^{(n)}\right] &= k^T C^{-1} t \\ \mathrm{var}\left[t^{(n+1)} \mid t^{(1)}, \ldots, t^{(n)}\right] &= V - k^T C^{-1} k \end{aligned}}$$

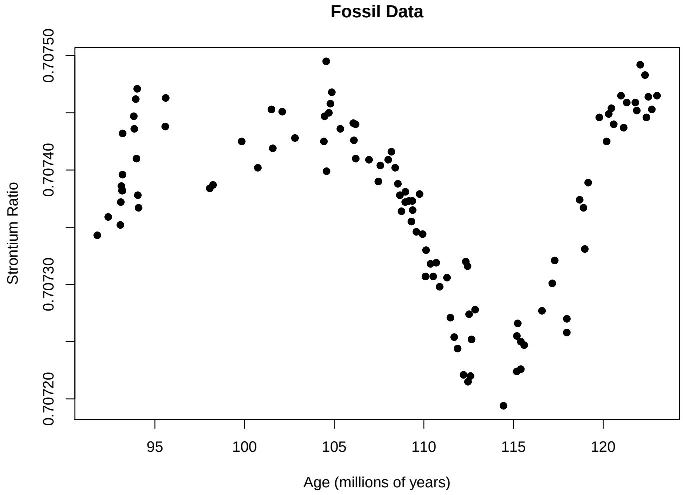
- $k = \left(\mathrm{Cov}\left[t^{(n+1)}, t^{(1)}\right], \ldots, \mathrm{Cov}\left[t^{(n+1)}, t^{(n)}\right]\right)^T$
- $V = \mathrm{Cov}\left[t^{(n+1)}, t^{(n+1)}\right] = \text{prior var}\left[t^{(n+1)}\right]$

Finally . . . some data

- 106 observations of fossil shells
- Age (in millions of years) and Ratios of strontium isotopes
- Previous example from STAT 527
- Data from `SemiPar` package in `R` [Ruppert et al., 2003], who got it from Bralower et al. [1997]



An “old” dataset



How to fit a Gaussian Process Regression Model

- Assume the prior covariance function:

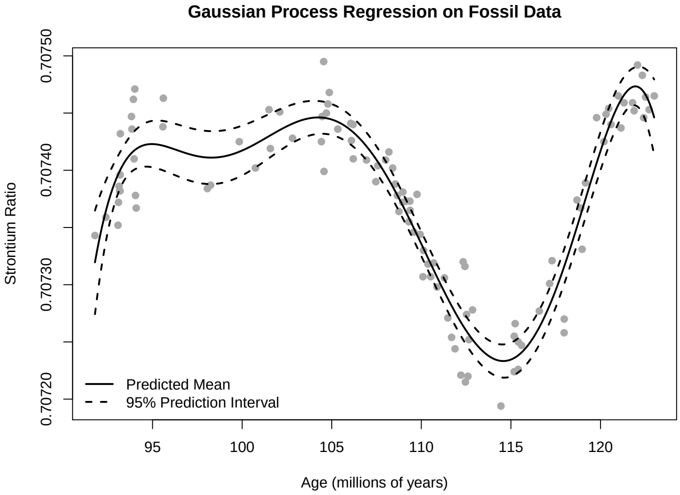
$$C_{ij} = \mathrm{Cov}\left[t^{(i)}, t^{(j)}\right] = \eta^2 \exp\left(-\rho^2 \left(x^{(i)} - x^{(j)}\right)^2\right) + \delta_{ij}\sigma_\epsilon^2$$

- Based on the multivariate Gaussian distribution of the outcomes, the log-likelihood is:

$$\log p(t \mid x, \eta, \rho) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|C| - \frac{1}{2}t^T C^{-1} t$$

- Find maximum likelihood esimates $\hat{\eta}, \hat{\rho}$
- σ_ϵ^2 assumed to be fixed at 10^{-9}

Gaussian Process Regression on Fossil data



What's Next?

$$C_{ij} = \mathrm{Cov}\left[t^{(i)}, t^{(j)}\right] = \eta^2 \exp\left(-\rho^2 \left(x^{(i)} - x^{(j)}\right)^2\right) + \delta_{ij}\sigma_\epsilon^2$$

- How to implement a Bayesian approach?
  - Integrate out parameters to get "parameter-free" marginal distribution $p\left(t^{(n+1)} \mid t^{(1)}, \ldots, t^{(n)}\right)$ for prediction?
- More to come ...
  - Compare against other semi-parametric procedures
  - p-dimension covariate regression example ($p > 1$)
  - How to implement a three-way classification / discrimination procedure?

References

T J Bralower, P D Fullagar, C K Paull, G S Dwyer, and R M Leckie. Mid-Cretaceous strontium-isotope stratigraphy of deep-sea sections. *Geological Society of America Bulletin*, 109:1421–1442, 1997. doi: 10.1130/0016-7606(1997)109⟨1421.

R M Neal. Regression and classification using Gaussian process priors (with discussion). *Bayesian Statistics*, 6: 475–501, 1999.

David Ruppert, Matthew P Wand, and Raymond J Carroll. *Semiparametric regression*, volume 12. Cambridge University Press, 2003. ISBN 0521785162.