STAT 518
Final Student Presentation

Wen Wei Loh

May 23, 2013

Recap: what is the paper about?

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

  → how two outcomes are correlated, based on values of their predictors
  - Parameters describe the relationships between predictors, rather than the predictors directly
- Prediction
  - Objective is to obtain a predictive distribution for the outcome of a future observation
  - Take a Bayesian approach to integrate the parameters out of the predictive distribution
- Radford M. Neal [1999]

Recap: Covariance functions

- Observed data : $\left(x^{(1)}, t^{(1)}\right), \ldots, \left(x^{(n)}, t^{(n)}\right)$
- Assume joint multivariate Gaussian distribution for t:

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

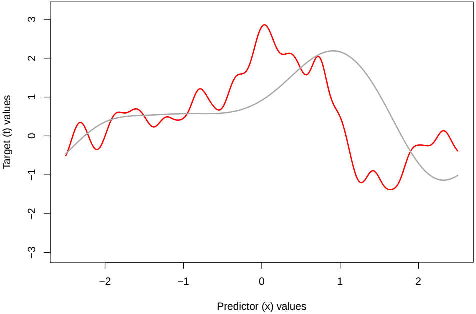
$$\text{Covariance } \textbf{matrix } C = \left\{ \underbrace{\text{cov}\left[t^{(i)}, t^{(j)}\right]}_{\text{Covariance } \textbf{function}} \qquad i, j \in [1, n] \right\}$$

$$C_{ij} = \text{cov}\left[t^{(i)}, t^{(j)}\right] = f\left(x^{(i)}, x^{(j)}\right)$$

- Covariance between the outcomes is fully described by the relationship between the predictors
- Covariance **function** f can take infinitely many forms ...

Recap: Covariance functions

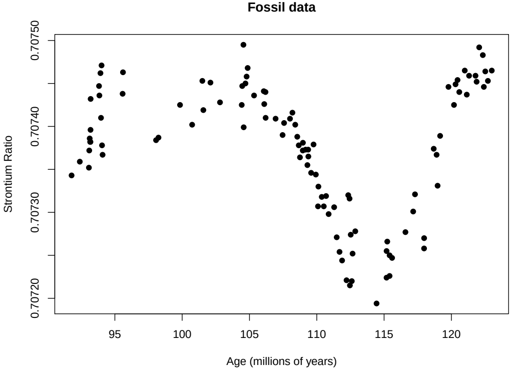
Smooth regression models: $C_{ij} = \underbrace{\eta^2 \exp\left(-\rho^2 \left(x^{(i)} - x^{(j)}\right)^2\right)}_{\textit{Exponential}} + \underbrace{\delta_{ij}\sigma_\epsilon^2}_{\textit{Noise}}$



e.g. $C_{ij} = \exp\left(-\left(x^{(i)} - x^{(j)}\right)^2\right)$

$$C_{ij} = \exp\left(-\left(x^{(i)} - x^{(j)}\right)^2\right) + \mathbf{0.5^2} \exp\left(-\mathbf{5^2}\left(x^{(i)} - x^{(j)}\right)^2\right)$$

An “old” dataset



- 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]

A “parameter-free” predictive distribution

- Assume the prior covariance function:

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

- σ_ϵ^2 assumed to be fixed at 10^{-9}

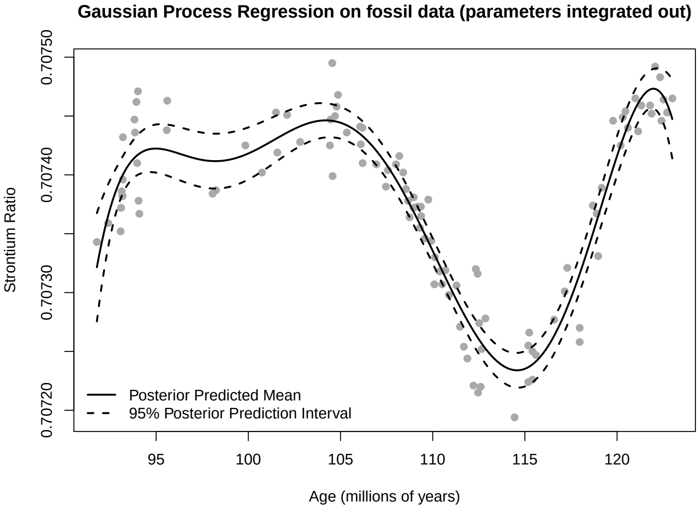
$\Rightarrow$ Integrate η, ρ out to get “parameter-free” predictions:

$$\begin{aligned} & p\left(t^{(n+1)} \mid t^{(1)}, \ldots, t^{(n)}\right) \\ &= \int\int \underbrace{p\left(t^{(n+1)} \mid t^{(1)}, \ldots, t^{(n)}, \eta, \rho\right)}_{\text{univariate Gaussian}} \underbrace{p\left(\eta, \rho \mid t^{(1)}, \ldots, t^{(n)}\right)}_{\textit{posterior}} \, d\eta \, d\rho \\ &\approx \frac{1}{S}\sum_{s=1}^{S} p\left(t^{(n+1)} \mid t^{(1)}, \ldots, t^{(n)}, \eta^{(s)}, \rho^{(s)}\right), \end{aligned}$$

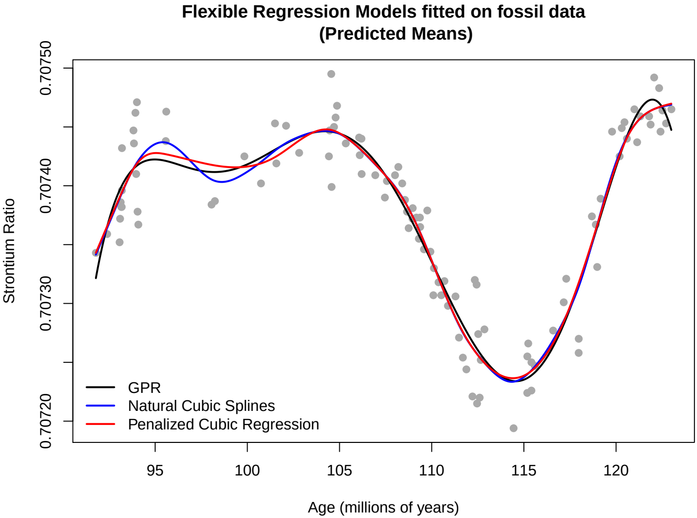
$$\eta^{(s)}, \rho^{(s)} = \text{posterior draws from } p\left(\eta, \rho \mid t^{(1)}, \ldots, t^{(n)}\right)$$

- Thanks to Jon Wakefield

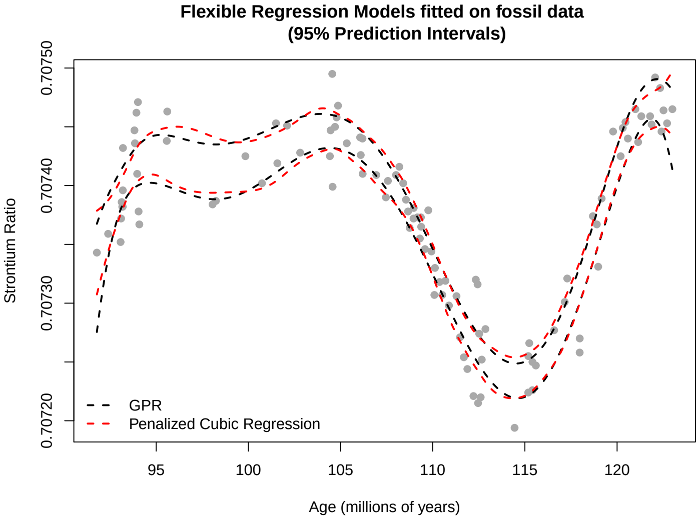
Gaussian Process Regression on Fossil data



Gaussian Process Regression on Fossil data



Gaussian Process Regression on Fossil data



Sensitivity to parameterization

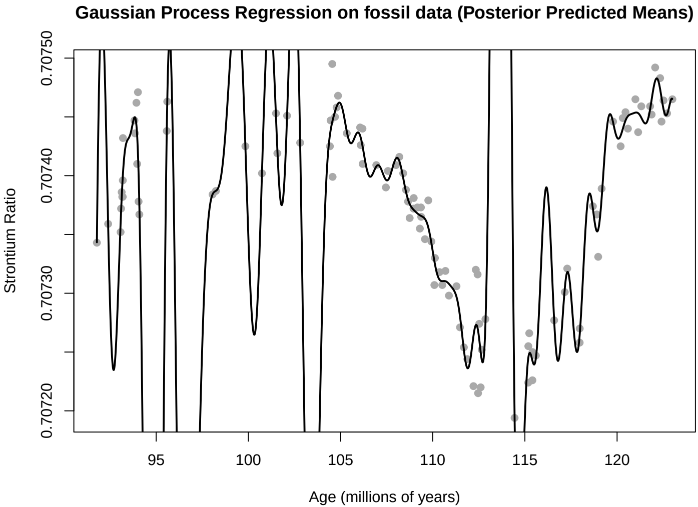
- Could we include another term to capture the smaller differences in x?[1]
- Suppose we assume this prior covariance function instead:

$$\operatorname{cov}\left[t^{(i)}, t^{(j)}\right] = \underbrace{\eta_1^2 \exp\left(-\left(x^{(i)} - x^{(j)}\right)^2\right)}_{\text{force a smooth short-term trend in } x} + \eta_2^2 \exp\left(-\rho^2\left(x^{(i)} - x^{(j)}\right)^2\right) + \delta_{ij}\sigma_\epsilon^2$$

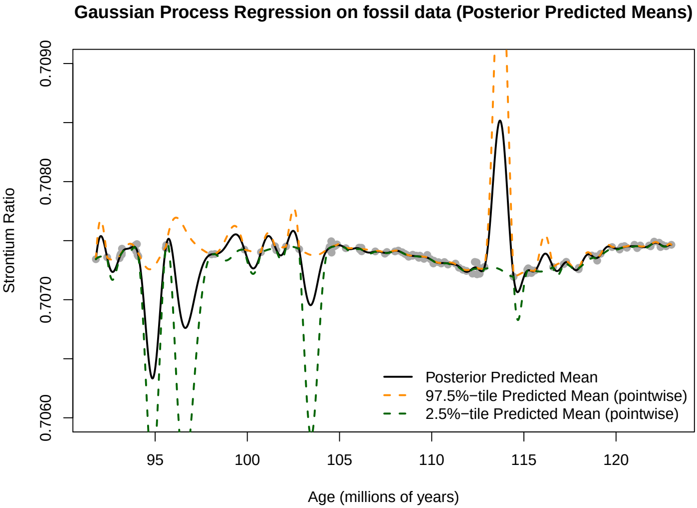
- If there is no short-term trend (on the scale of x), then posterior distribution of η_1 will be concentrated at 0, and have little influence on the covariance?

[1] Suggestion from Patrick Heagerty

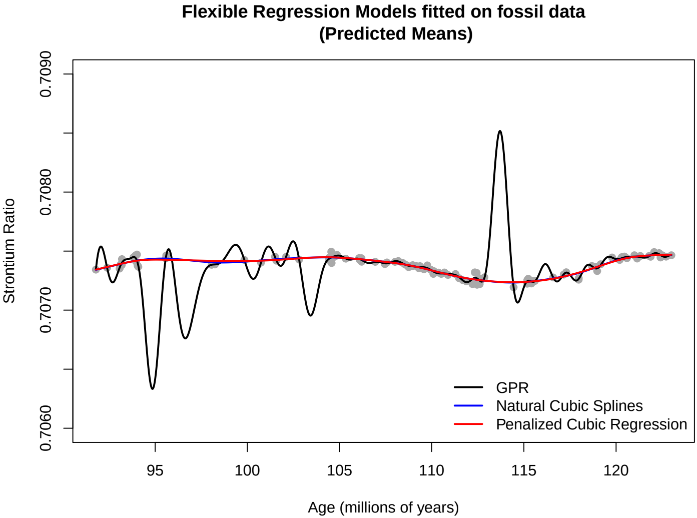
Sensitivity to discontinuities



Sensitivity to discontinuities



Sensitivity to discontinuities



Classification / Discrimination

- Suppose the targets $t^{(i)}$ are from the set $\{0, \ldots, K-1\}$.
- The distribution of $t^{(i)}$ would then be in terms of unobserved real-valued "latent" variables $y_1^{(i)}, \ldots, y_{K-1}^{(i)}$ for each case i
- Class probabilities for K classes:

$$\Pr\left(t^{(i)} = k\right) = \frac{\exp\left(y_k^{(i)}\right)}{\sum_{h=0}^{K-1} \exp\left(y_h^{(i)}\right)}$$

- Latent variables $y_1^{(i)}, \ldots, y_{K-1}^{(i)}$ modelled with GP regression

Simulation example

- For each observation i, generate four variables:

$$\tilde{x}_1^{(i)}, \ldots \tilde{x}_4^{(i)} \underset{iid}{\sim} Unif(0, 1)$$

- Use only $\tilde{x}_1^{(i)}$ and $\tilde{x}_2^{(i)}$ in determining the class $t^{(i)}$:

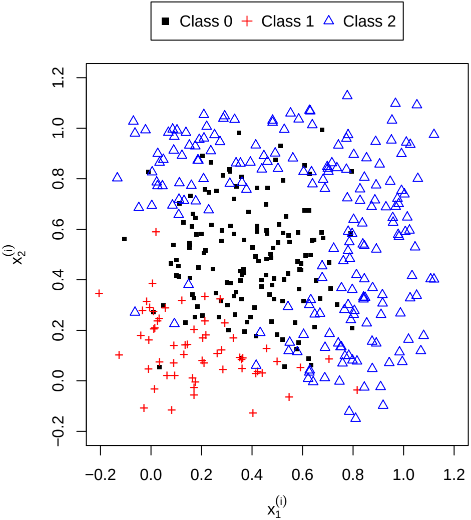
$$t^{(i)} \in \{0, 1, 2\}$$

- $\tilde{x}_3^{(i)}$ and $\tilde{x}_4^{(i)}$ have no effect on $t^{(i)}$
- Assume the following covariance function for the *latent* variables $y_k^{(i)},\ k \in \{0, 1, 2\}$:

$$\mathrm{cov}\left[y_k^{(i)}, y_k^{(j)}\right] = \sigma_\alpha^2 + \eta^2 \exp\left(-\sum_{u=1}^{4} \rho_u^2 \left(x_u^{(i)} - x_u^{(j)}\right)^2\right) + \delta_{ij}\sigma_\epsilon^2$$

- Assume $\sigma_\alpha^2 = \sigma_\epsilon^2 = 10^2$

Simulation example



Simulated observations from [Neal, 1999];

covariates $x_u^{(i)} = \tilde{x}_u^{(i)} + z, \; z \underset{iid}{\sim} N(0, 0.1)$

Simulation example

- Fit the model with 400 labelled training cases
- Find misclassification error rate with 600 test cases
- Posterior distribution for $x_3^{(i)}$ and $x_4^{(i)}$ will be concentrated near zero
- Results with other classification methods:

Method	Misclassification error rate (%)
Neal [1999] (from paper)	13
Classification Tree	19
Multinomial Logistic Regression	31

“The end has no end”

$$\operatorname{cov}\left[y_k^{(i)}, y_k^{(j)}\right] = \sigma_\alpha^2 + \eta^2 \exp\left(-\sum_{u=1}^{4} \rho_u^2 \left(x_u^{(i)} - x_u^{(j)}\right)^2\right) + \delta_{ij}\sigma_\epsilon^2$$

- How to conduct classification with multiple latent variables per observation?
    - Integrate out latent variables to get predictive probabilities

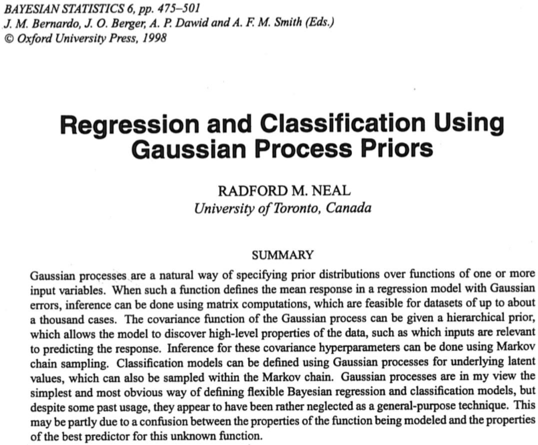
$$\begin{aligned}\Pr\left(t = k \mid x, \theta\right) &= \int \dots \int \Pr\left(t = k, y_0, \dots, y_{K-1} \mid x, \theta\right)\; dy_0 \cdots dy_{K-1} \\ &= \int \dots \int \Pr\left(t = k \mid y_0, \dots, y_{K-1}\right) \underbrace{\Pr\left(y_0, \dots, y_{K-1} \mid x, \theta\right)}_{posterior}\; dy_0 \cdots dy_{K-1}\end{aligned}$$

-
    - Integrate out parameter θ to get “parameter-free” predictive distribution

“The time to hesitate is through”

- + Gaussian Process Regression and Classification is a flexible tool for predictions
- + Parameters in the assumed covariance function may be used for inference on the underlying data structure
- + Bayesian approach allows parameter-free posterior predictions
- - Family of possible regression surfaces may be sensitive to (or limited by) the assumed covariance function
- - Balance has to be made between modelling complex covariance functions and obtaining interpretable, positive-definite covariance matrices
- - May be computationally more expensive than other nonparametric or machine-learning methods

Bayesian Statistics, 6: 475-501, 1999



BAYESIAN STATISTICS 6, pp. 475–501
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)
© Oxford University Press, 1998

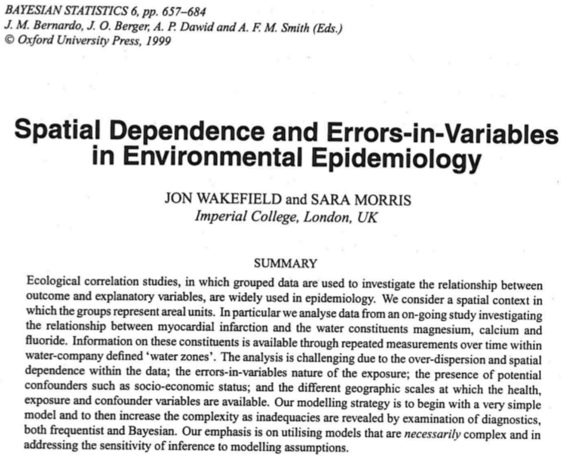
Regression and Classification Using Gaussian Process Priors

RADFORD M. NEAL
University of Toronto, Canada

SUMMARY

Gaussian processes are a natural way of specifying prior distributions over functions of one or more input variables. When such a function defines the mean response in a regression model with Gaussian errors, inference can be done using matrix computations, which are feasible for datasets of up to about a thousand cases. The covariance function of the Gaussian process can be given a hierarchical prior, which allows the model to discover high-level properties of the data, such as which inputs are relevant to predicting the response. Inference for these covariance hyperparameters can be done using Markov chain sampling. Classification models can be defined using Gaussian processes for underlying latent values, which can also be sampled within the Markov chain. Gaussian processes are in my view the simplest and most obvious way of defining flexible Bayesian regression and classification models, but despite some past usage, they appear to have been rather neglected as a general-purpose technique. This may be partly due to a confusion between the properties of the function being modeled and the properties of the best predictor for this unknown function.

Bayesian Statistics, 6: 657-684, 1999



BAYESIAN STATISTICS 6, pp. 657–684
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)
© Oxford University Press, 1999

Spatial Dependence and Errors-in-Variables in Environmental Epidemiology

JON WAKEFIELD and SARA MORRIS
Imperial College, London, UK

SUMMARY

Ecological correlation studies, in which grouped data are used to investigate the relationship between outcome and explanatory variables, are widely used in epidemiology. We consider a spatial context in which the groups represent areal units. In particular we analyse data from an on-going study investigating the relationship between myocardial infarction and the water constituents magnesium, calcium and fluoride. Information on these constituents is available through repeated measurements over time within water-company defined 'water zones'. The analysis is challenging due to the over-dispersion and spatial dependence within the data; the errors-in-variables nature of the exposure; the presence of potential confounders such as socio-economic status; and the different geographic scales at which the health, exposure and confounder variables are available. Our modelling strategy is to begin with a very simple model and to then increase the complexity as inadequacies are revealed by examination of diagnostics, both frequentist and Bayesian. Our emphasis is on utilising models that are *necessarily* complex and in addressing the sensitivity of inference to modelling assumptions.

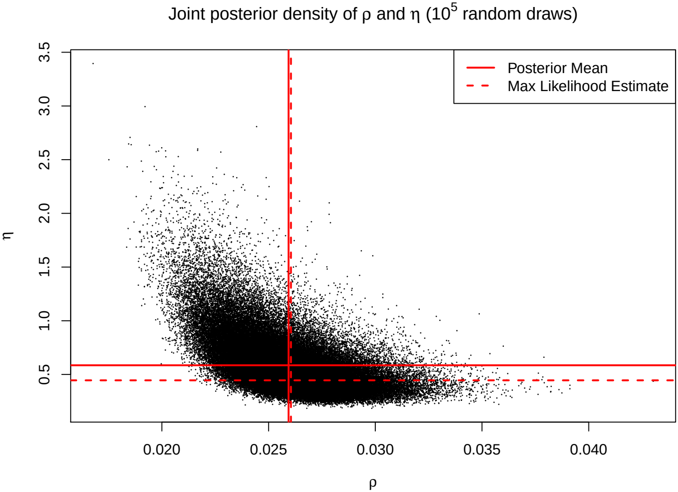
References

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.

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

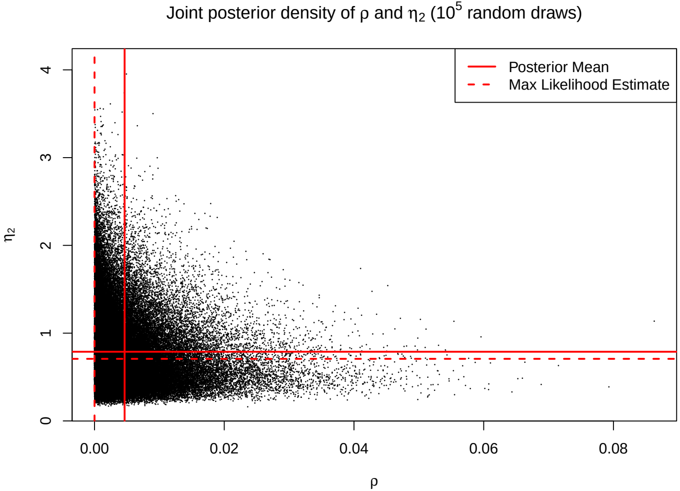
David Ruppert, Matthew P Wand, and Raymond J Carroll. *Semiparametric regression*, volume 12. Cambridge University Press, 2003. ISBN 0521785162.

Parameter posterior distributions



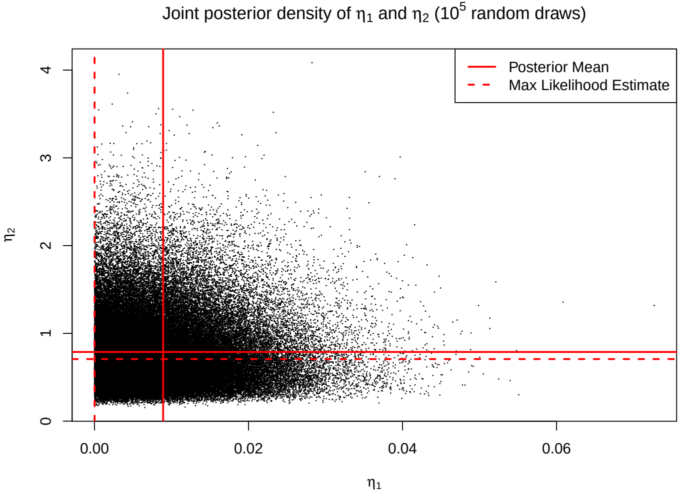
$$\text{cov}\left[t^{(i)}, t^{(j)}\right] = \eta^2 \exp\left(-\rho^2\left(x^{(i)} - x^{(j)}\right)^2\right) + \delta_{ij}\sigma_\epsilon^2$$

Parameter posterior distributions



$$\operatorname{cov}\left[t^{(i)}, t^{(j)}\right] = \eta_1^2 \exp\left(-\left(x^{(i)} - x^{(j)}\right)^2\right) + \eta_2^2 \exp\left(-\rho^2\left(x^{(i)} - x^{(j)}\right)^2\right) + \delta_{ij}\sigma_\epsilon^2$$

Parameter posterior distributions



$$\mathrm{cov}\left[t^{(i)}, t^{(j)}\right] = \eta_1^2 \exp\left(-\left(x^{(i)} - x^{(j)}\right)^2\right) + \eta_2^2 \exp\left(-\rho^2\left(x^{(i)} - x^{(j)}\right)^2\right) + \delta_{ij}\sigma_\epsilon^2$$

How to obtain a predictive distribution

- Recall the *observed* n cases have the joint distribution:

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

⇒ The joint distribution of a *new* outcome $t^{(n+1)}$ with the *observed* n cases is:

$$\left[t^{(1)}, \ldots, t^{(n)}, t^{(n+1)}\right]^T \sim N_{n+1}\left(\mathbf{0}, \begin{bmatrix} C & k \\ k^T & V \end{bmatrix}\right)$$

- $k = \left(\text{cov}\left[t^{(n+1)}, t^{(1)}\right], \ldots, \text{cov}\left[t^{(n+1)}, t^{(n)}\right]\right)^T$

- $V = \text{cov}\left[t^{(n+1)}, t^{(n+1)}\right] = \text{prior var}\left[t^{(n+1)}\right]$

- C, k, V are just defined by ...

How to obtain a predictive distribution

- . . . a Covariance **function**!
- Assume the prior covariance function:

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

- σ_ϵ^2 assumed to be fixed at 10^{-9}

⇒ Given the inputs for a *new* case $x^{(n+1)}$, the predictive distribution for the *new* outcome $t^{(n+1)}$ is *Gaussian*:

$$\mathrm{E}\left[t^{(n+1)} \mid t^{(1)}, \ldots, t^{(n)}, \eta, \rho\right] = k^T C^{-1} t$$
$$\mathrm{var}\left[t^{(n+1)} \mid t^{(1)}, \ldots, t^{(n)}, \eta, \rho\right] = V - k^T C^{-1} k$$

- What shall we do with η, ρ?