

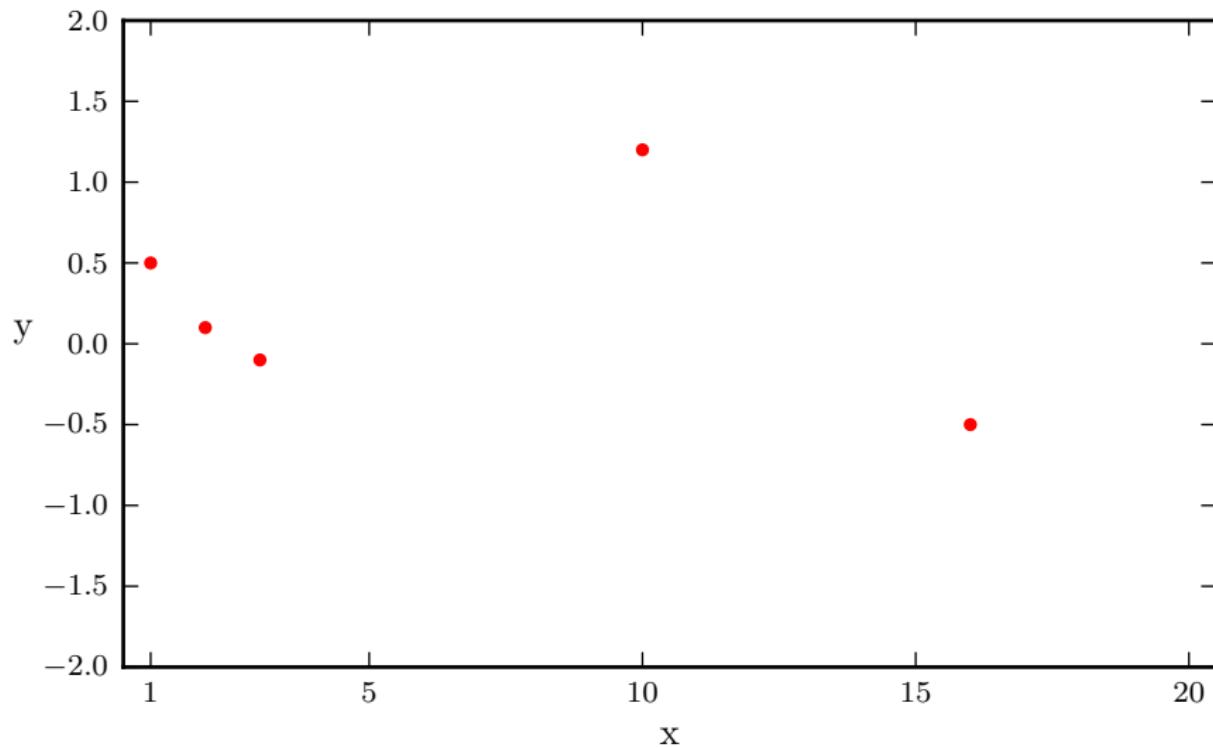
Gaussian Processes: From the Basics to the State-of-the-Art

Dr. Richard E. Turner (ret26@cam.ac.uk)
Computational and Biological Learning Lab, Department of
Engineering, University of Cambridge

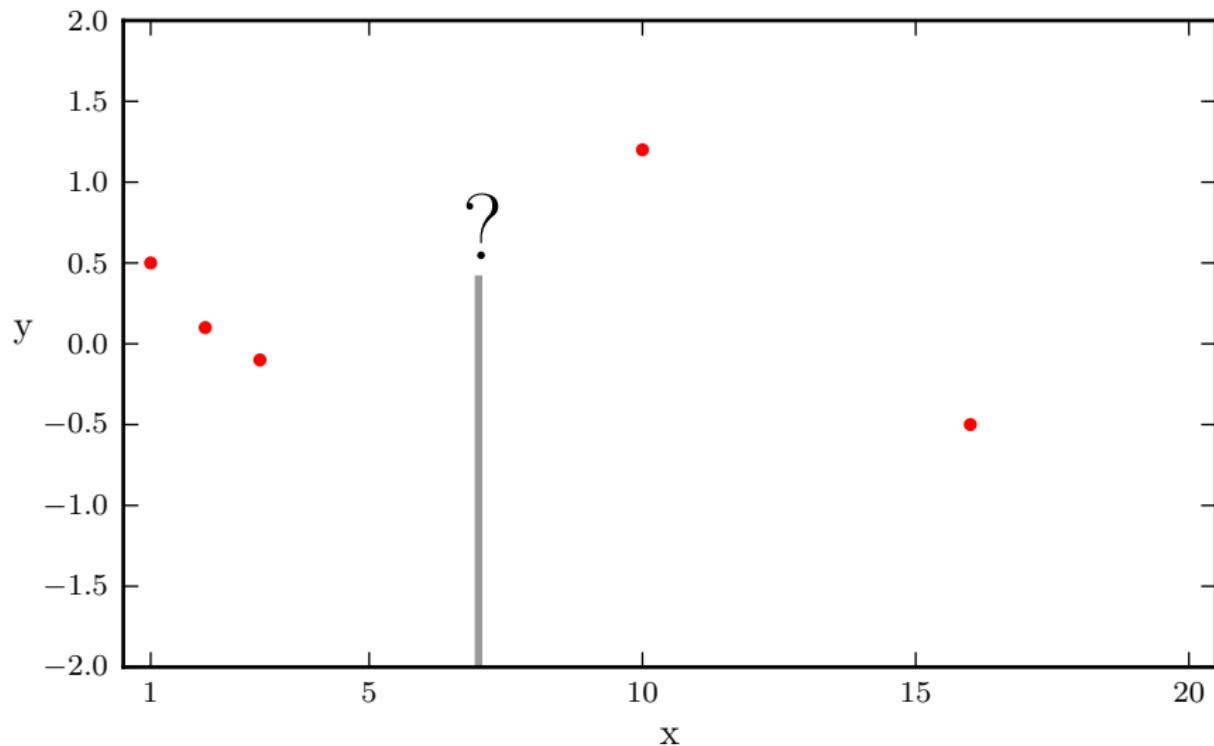
with Thang Bui, Yingzhen Li, José Miguel Hernández Lobato,
Daniel Hernández Lobato, Josiah Jan, Alex Navarro,
Felipe Tobar, and Maneesh Sahani



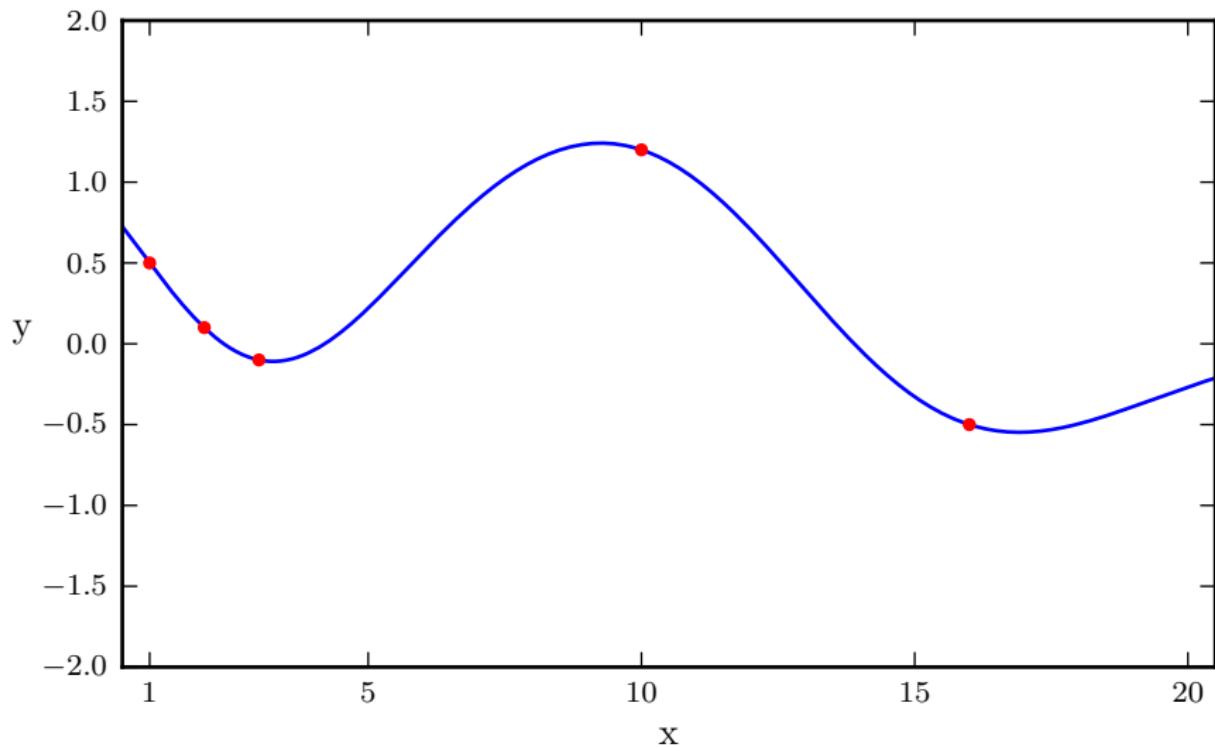
Motivation: non-linear regression



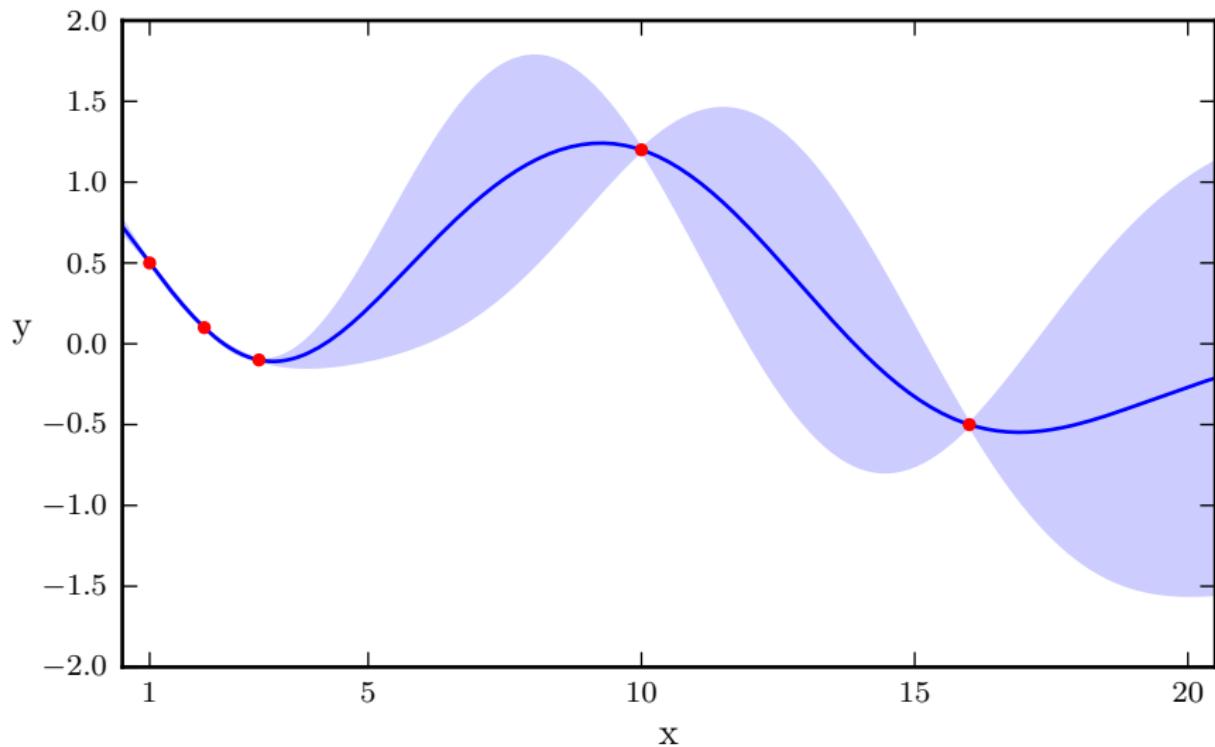
Motivation: non-linear regression



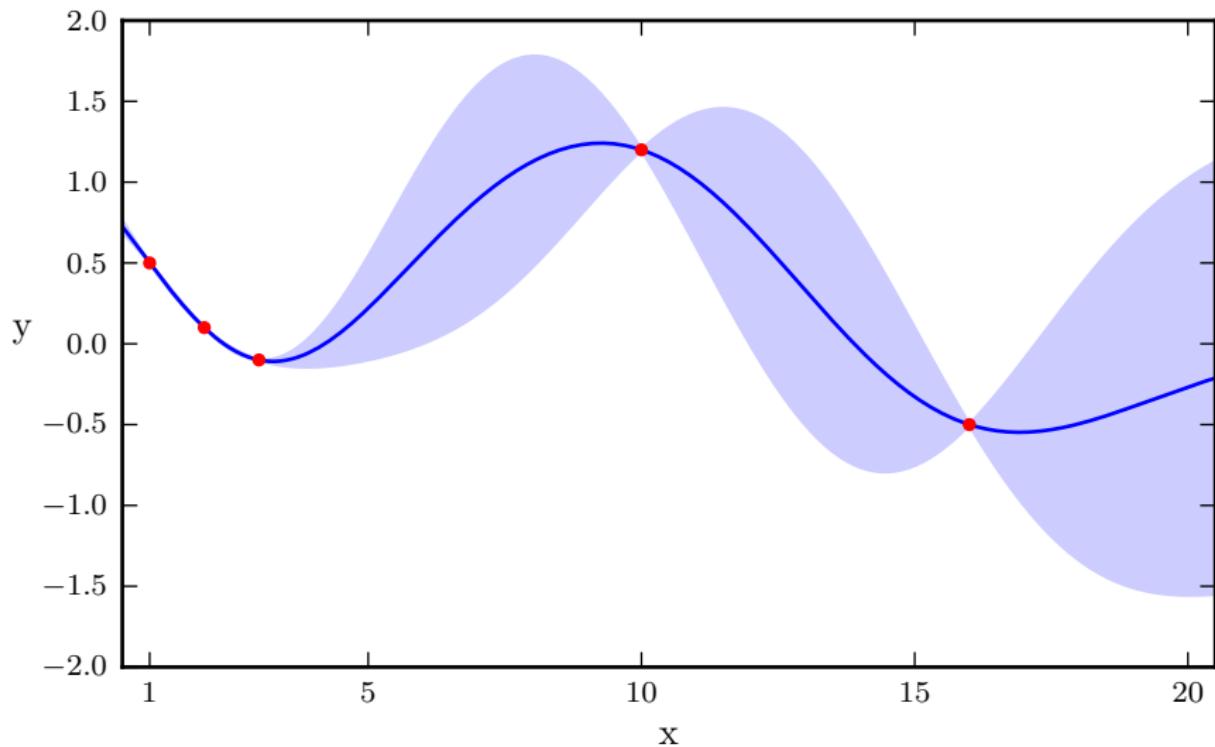
Motivation: non-linear regression



Motivation: non-linear regression



Motivation: non-linear regression

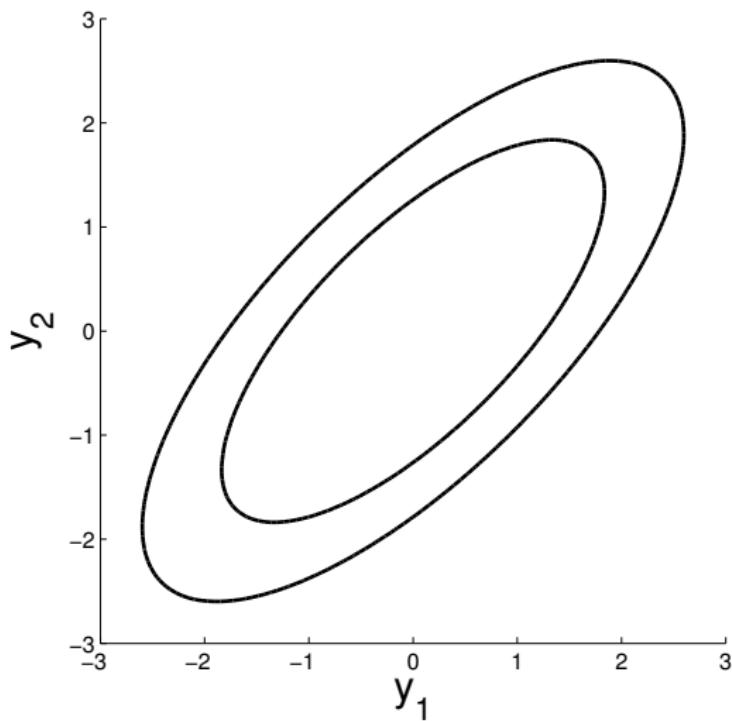


Can we do this with a plain old Gaussian?

Gaussian distribution

$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1} \mathbf{y}\right)$$

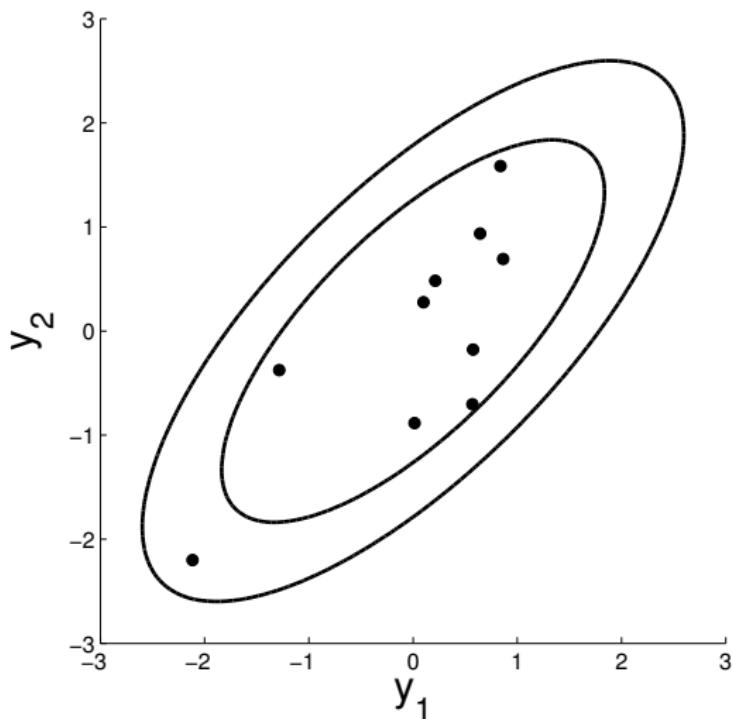
$$\Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$$



Gaussian distribution

$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1} \mathbf{y}\right)$$

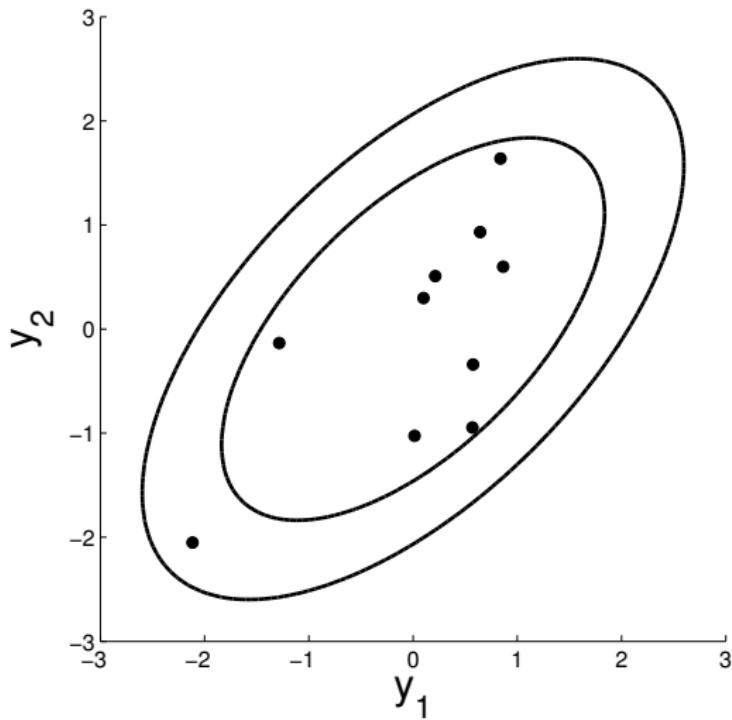
$$\Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$$



Gaussian distribution

$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1} \mathbf{y}\right)$$

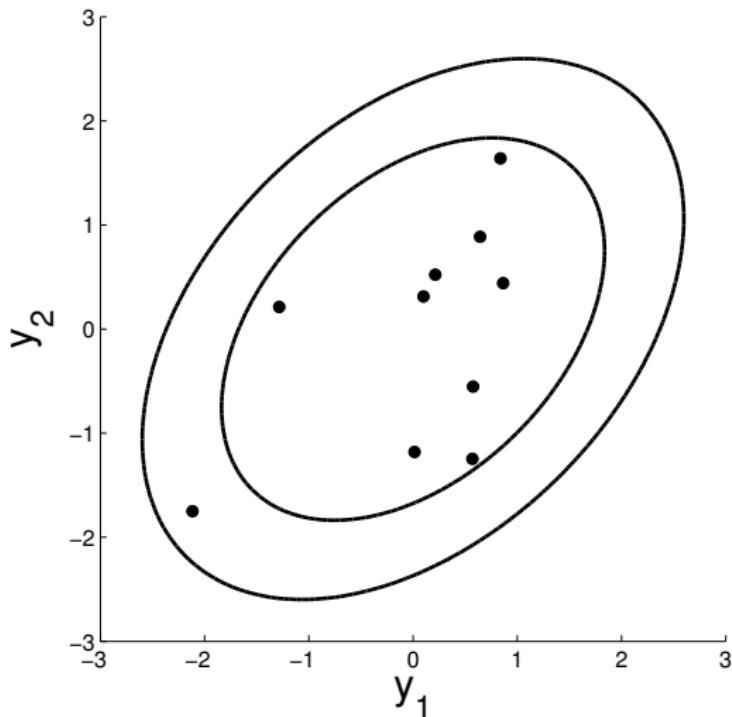
$$\Sigma = \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}$$



Gaussian distribution

$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1} \mathbf{y}\right)$$

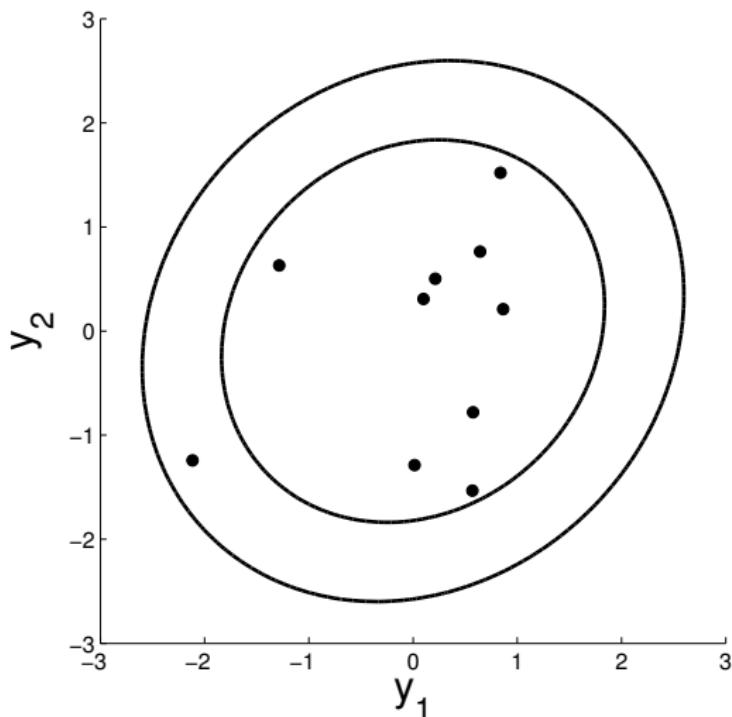
$$\Sigma = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$$



Gaussian distribution

$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1} \mathbf{y}\right)$$

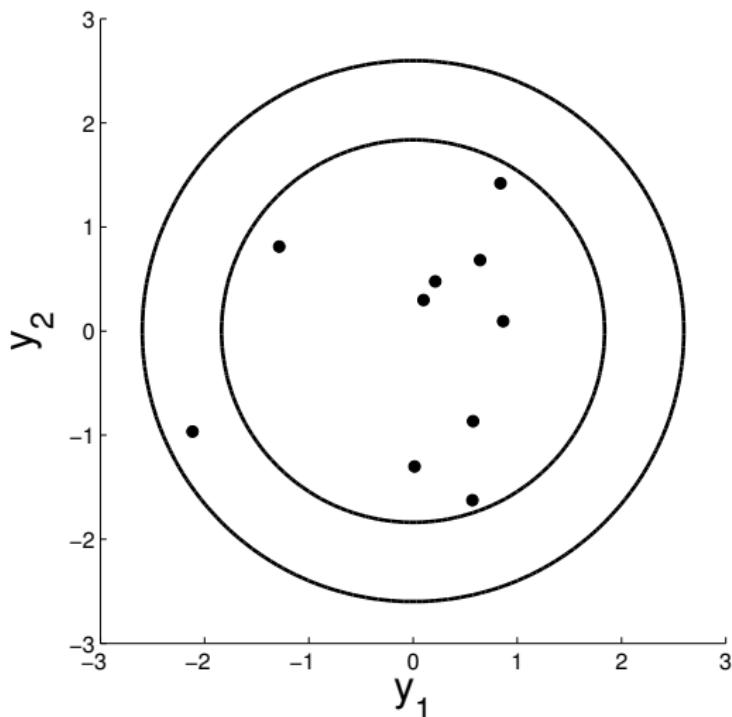
$$\Sigma = \begin{bmatrix} 1 & .1 \\ .1 & 1 \end{bmatrix}$$



Gaussian distribution

$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1} \mathbf{y}\right)$$

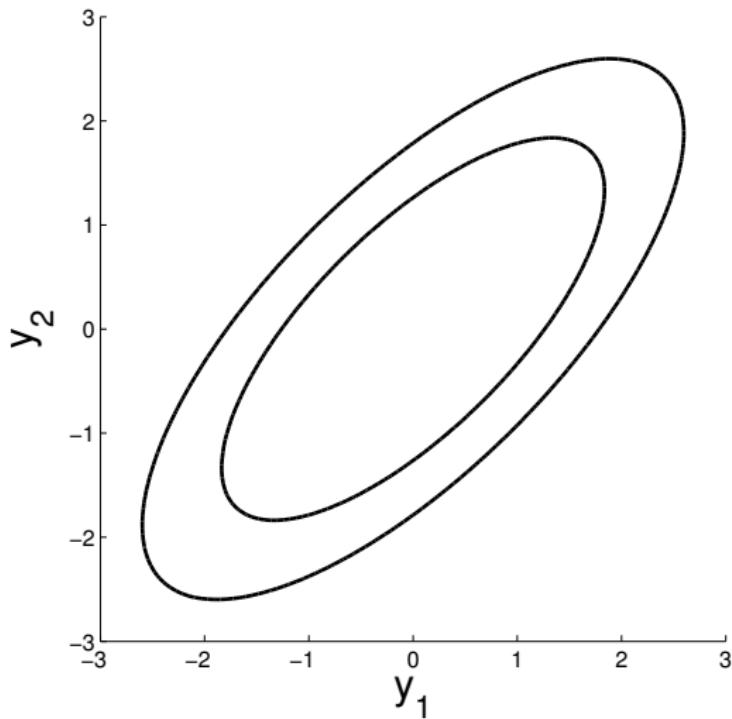
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Gaussian distribution

$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1} \mathbf{y}\right)$$

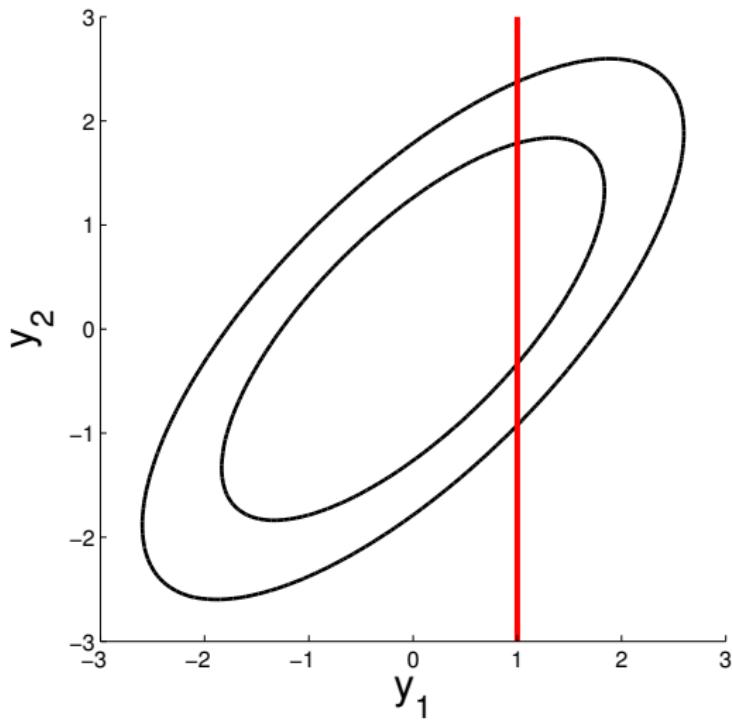
$$\Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$$



Gaussian distribution

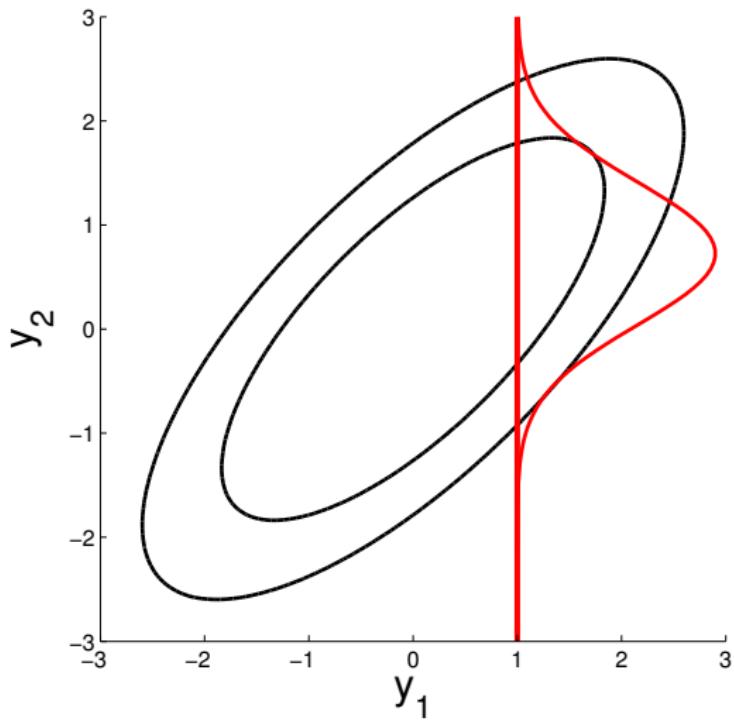
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1} \mathbf{y}\right)$$

$$\Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$$



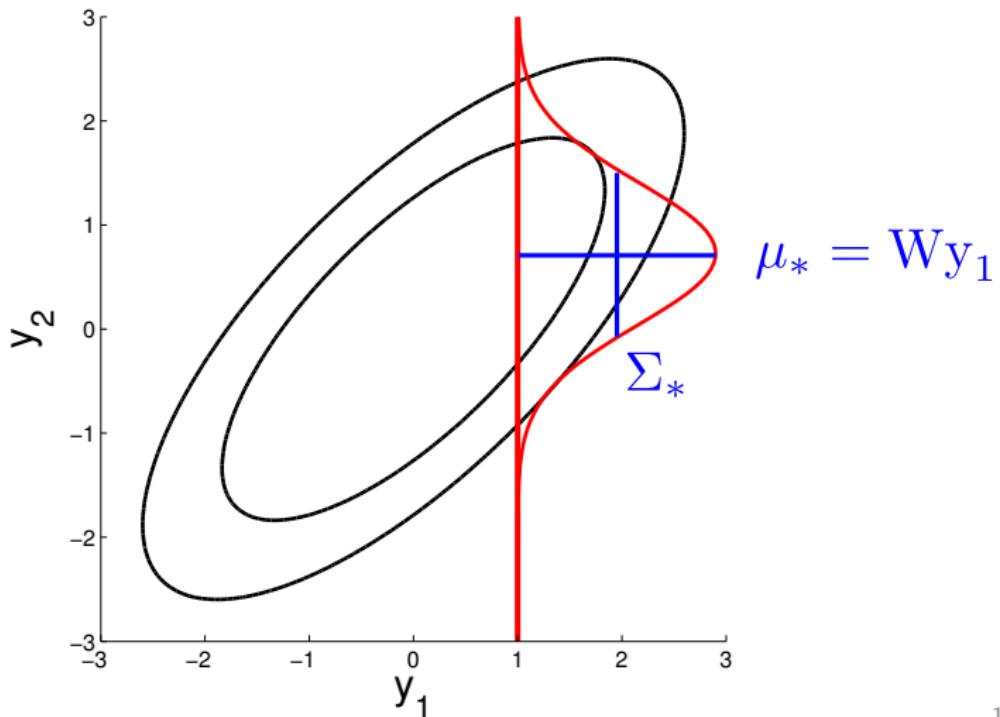
Gaussian distribution

$$p(y_2|y_1, \Sigma) \propto \exp\left(-\frac{1}{2}(y_2 - \mu_*)\Sigma_*^{-1}(y_2 - \mu_*)\right)$$



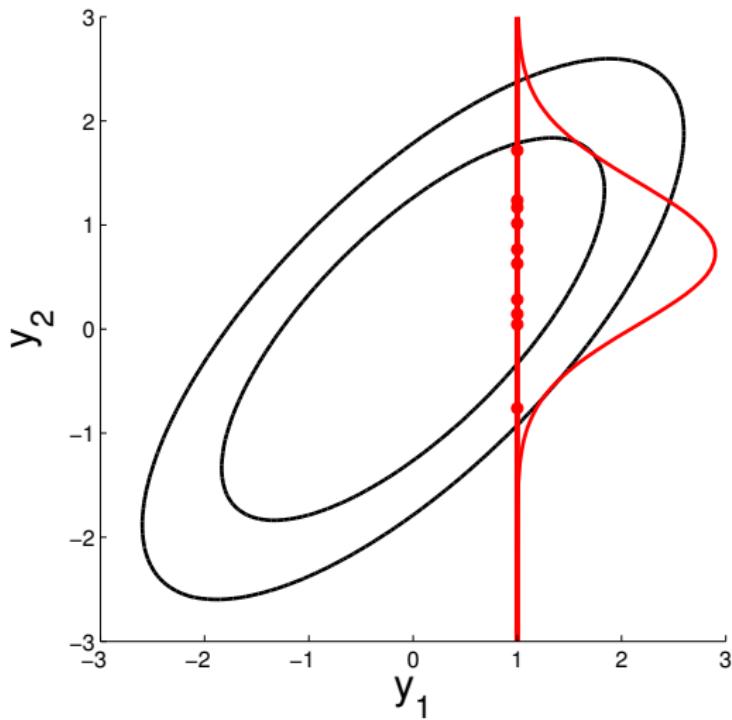
Gaussian distribution

$$p(y_2|y_1, \Sigma) \propto \exp\left(-\frac{1}{2}(y_2 - \mu_*)\Sigma_*^{-1}(y_2 - \mu_*)\right)$$



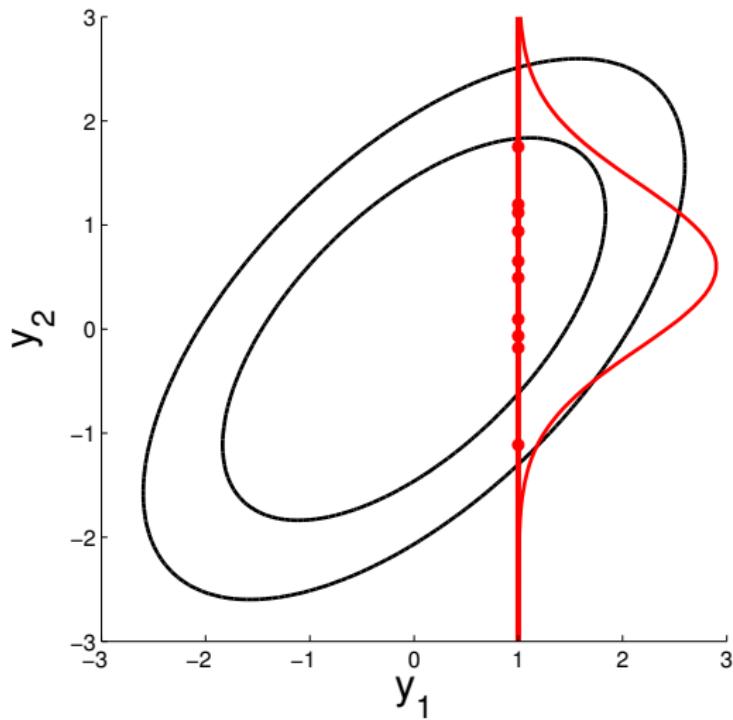
Gaussian distribution

$$p(y_2|y_1, \Sigma) \propto \exp\left(-\frac{1}{2}(y_2 - \mu_*)\Sigma_*^{-1}(y_2 - \mu_*)\right)$$



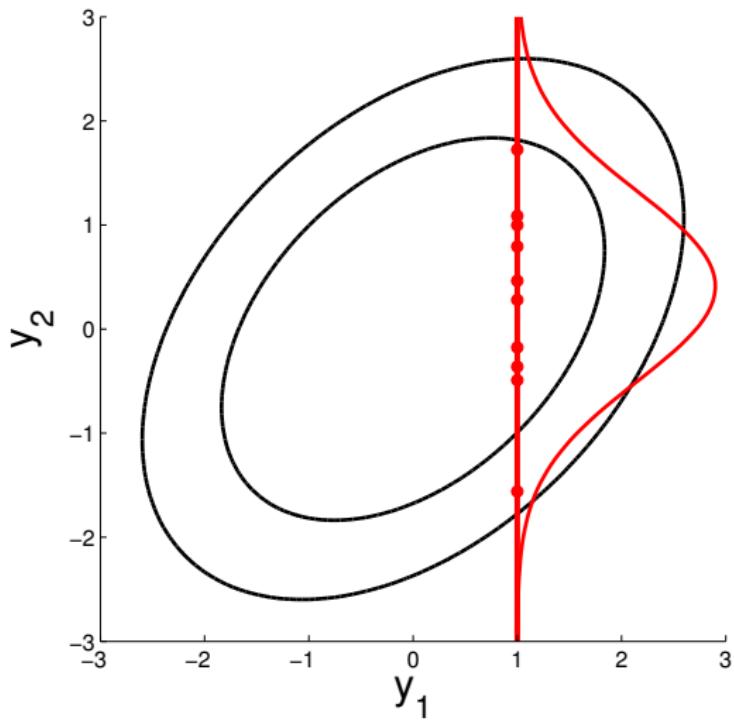
Gaussian distribution

$$p(y_2|y_1, \Sigma) \propto \exp\left(-\frac{1}{2}(y_2 - \mu_*)\Sigma_*^{-1}(y_2 - \mu_*)\right)$$



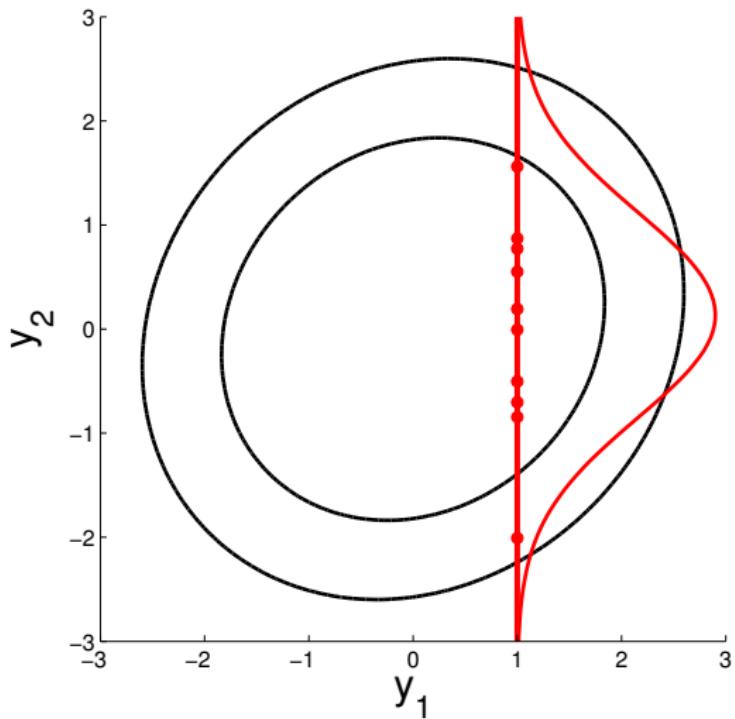
Gaussian distribution

$$p(y_2|y_1, \Sigma) \propto \exp\left(-\frac{1}{2}(y_2 - \mu_*)\Sigma_*^{-1}(y_2 - \mu_*)\right)$$



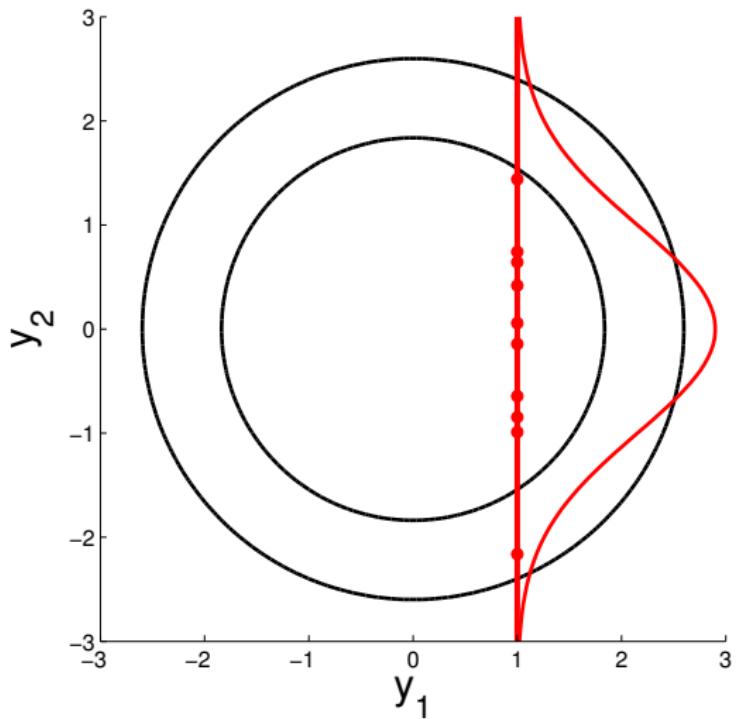
Gaussian distribution

$$p(y_2|y_1, \Sigma) \propto \exp\left(-\frac{1}{2}(y_2 - \mu_*)\Sigma_*^{-1}(y_2 - \mu_*)\right)$$

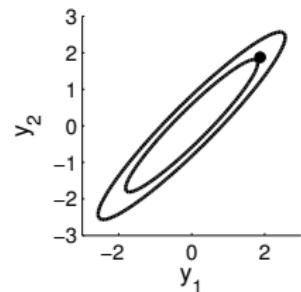


Gaussian distribution

$$p(y_2|y_1, \Sigma) \propto \exp\left(-\frac{1}{2}(y_2 - \mu_*)\Sigma_*^{-1}(y_2 - \mu_*)\right)$$

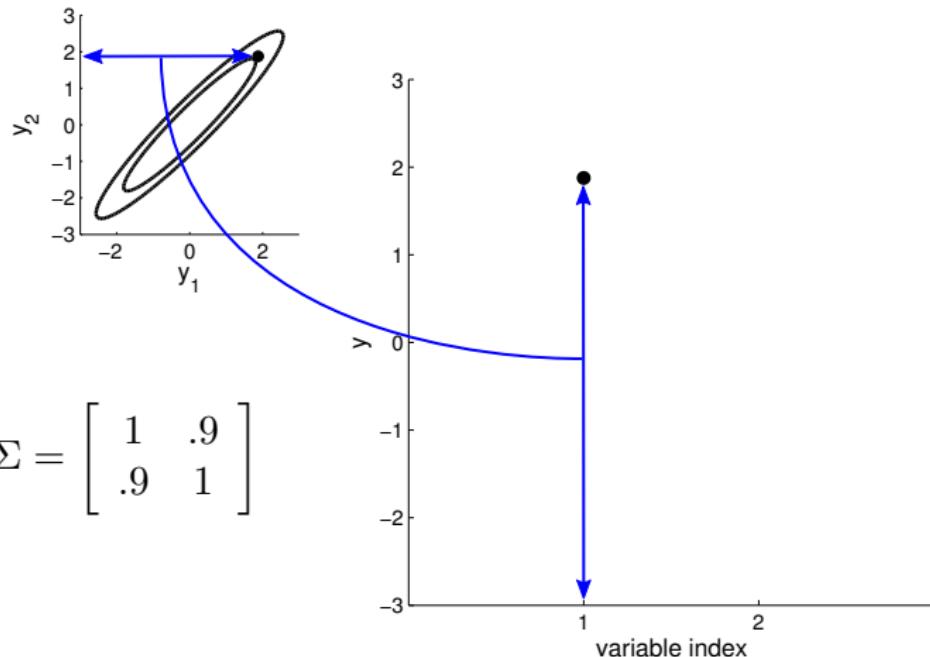


New visualisation

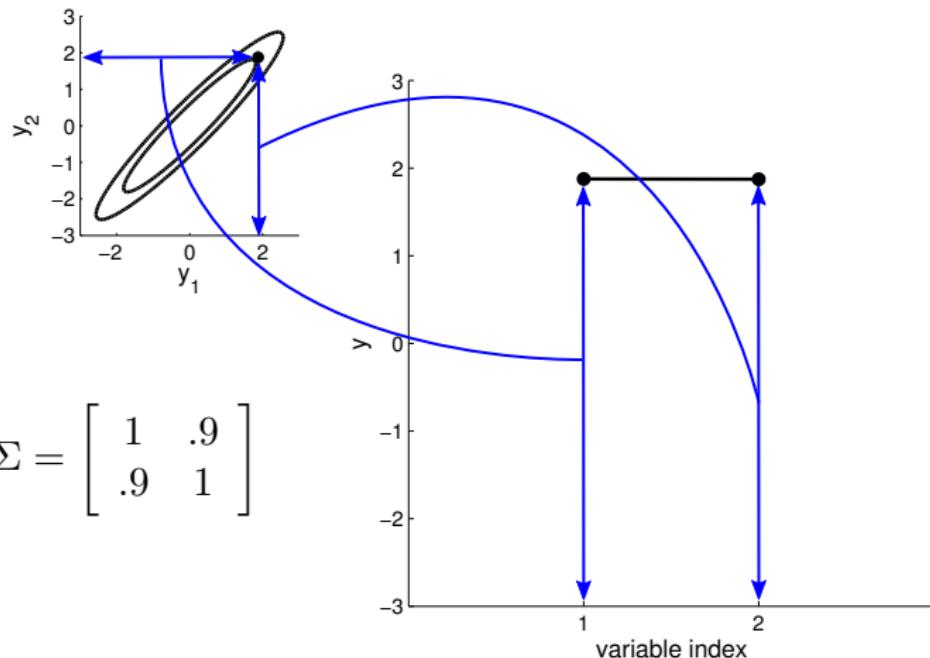


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

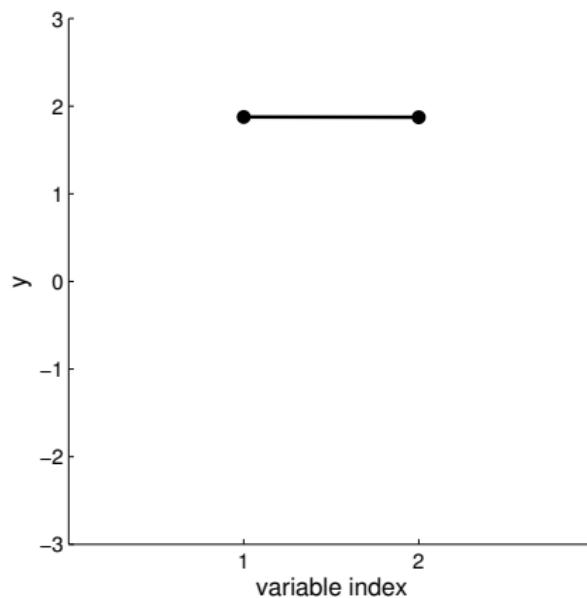
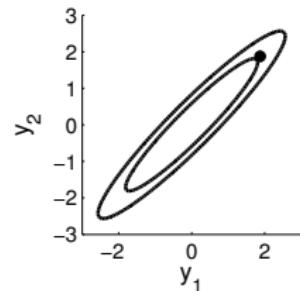
New visualisation



New visualisation

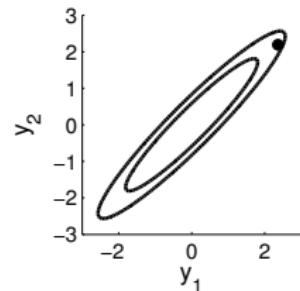


New visualisation

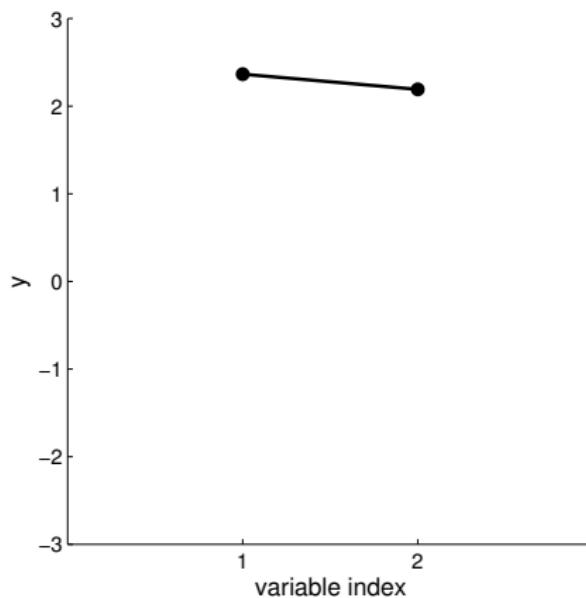


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

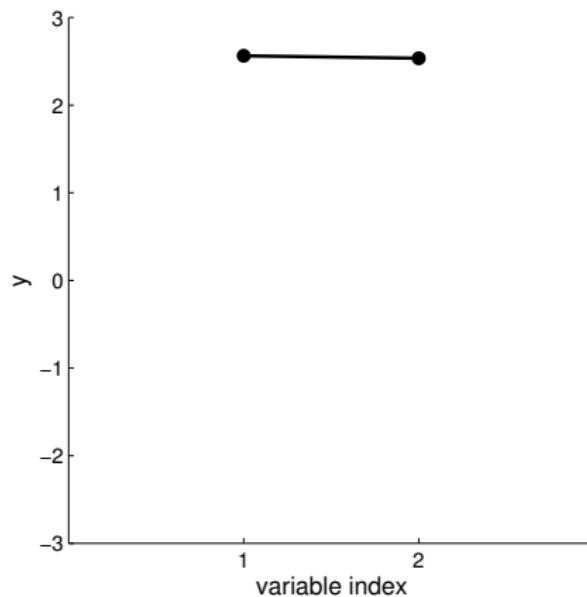
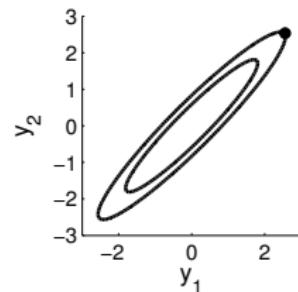
New visualisation



$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

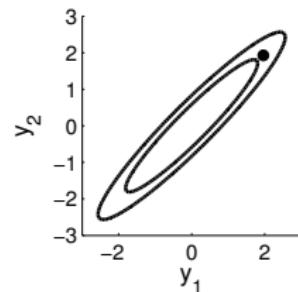


New visualisation

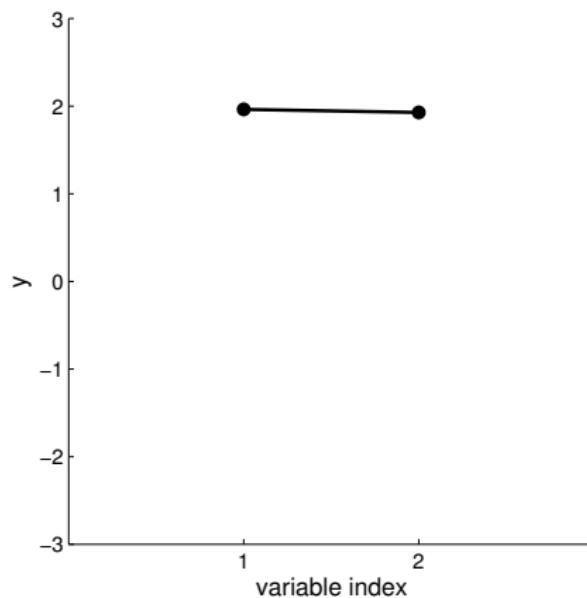


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

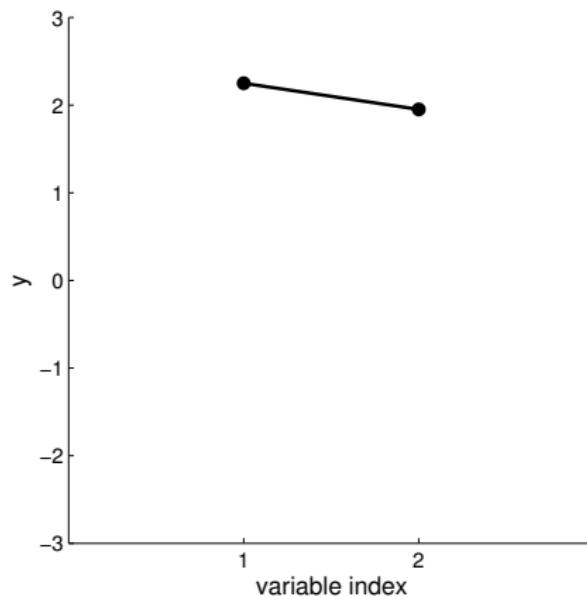
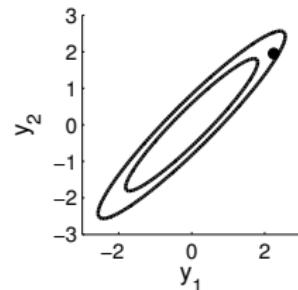
New visualisation



$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

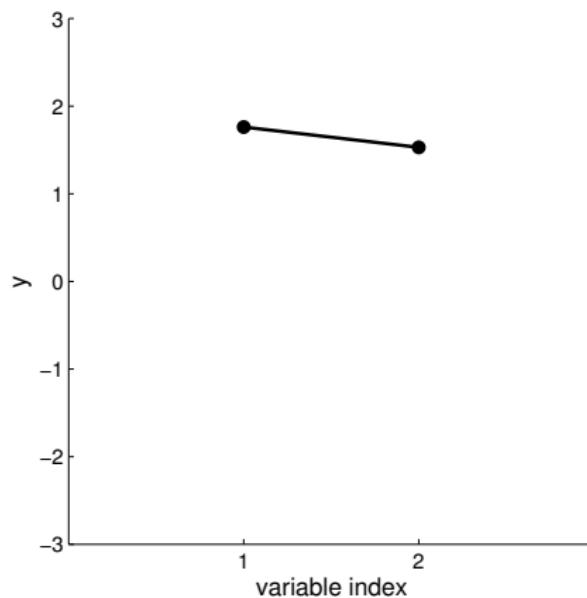
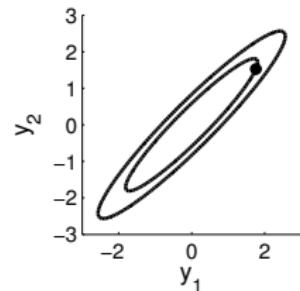


New visualisation



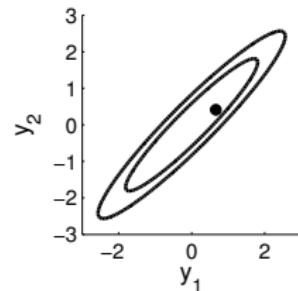
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation

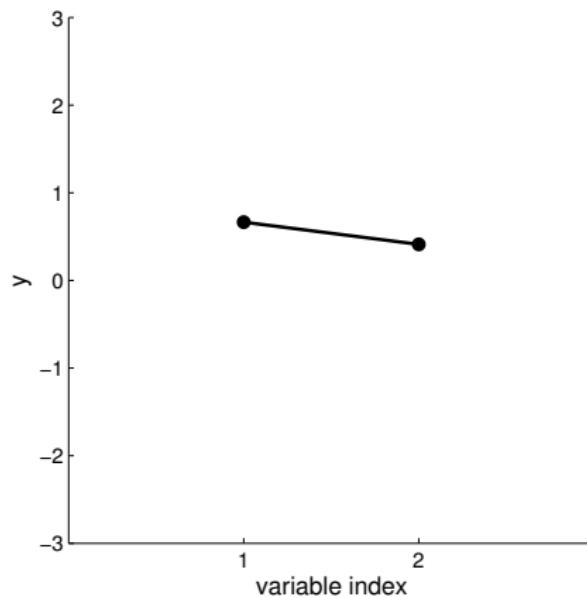


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

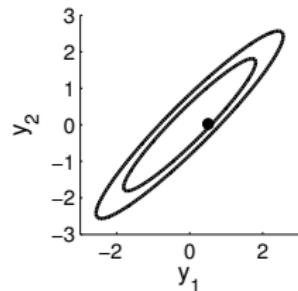
New visualisation



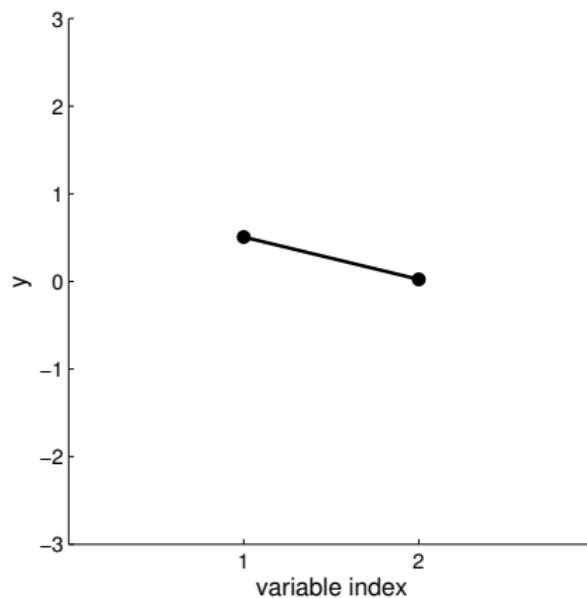
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



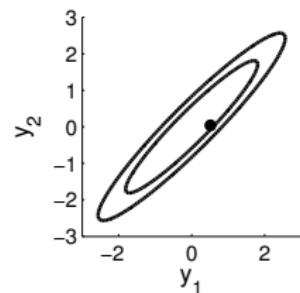
New visualisation



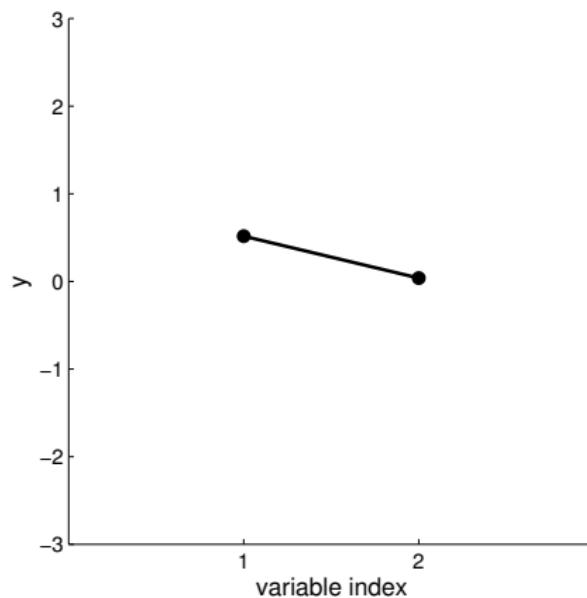
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



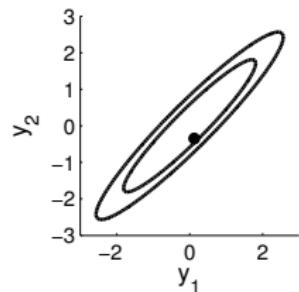
New visualisation



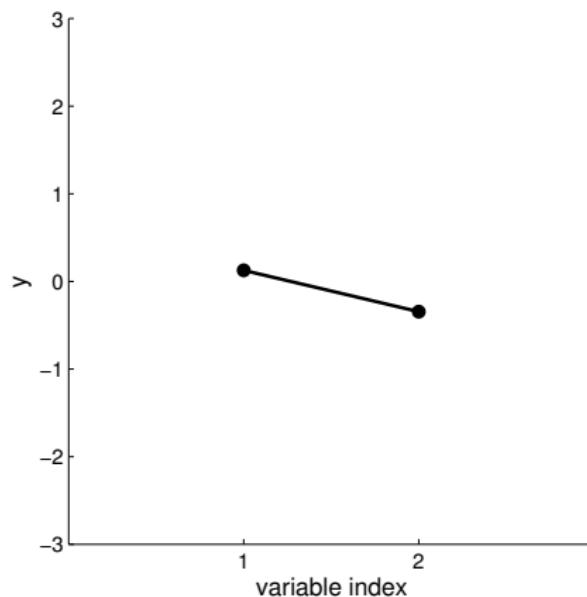
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



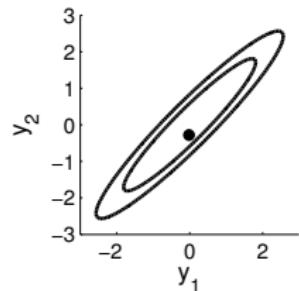
New visualisation



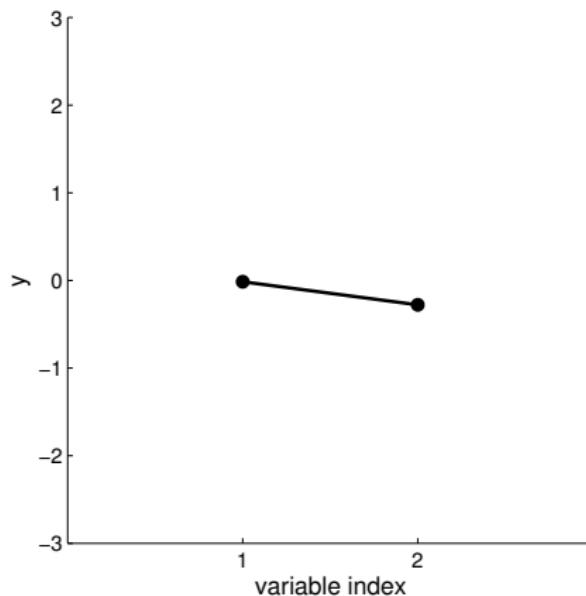
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



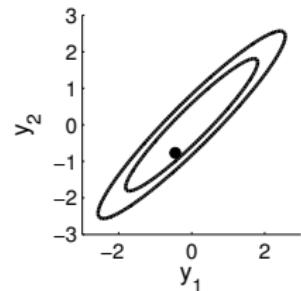
New visualisation



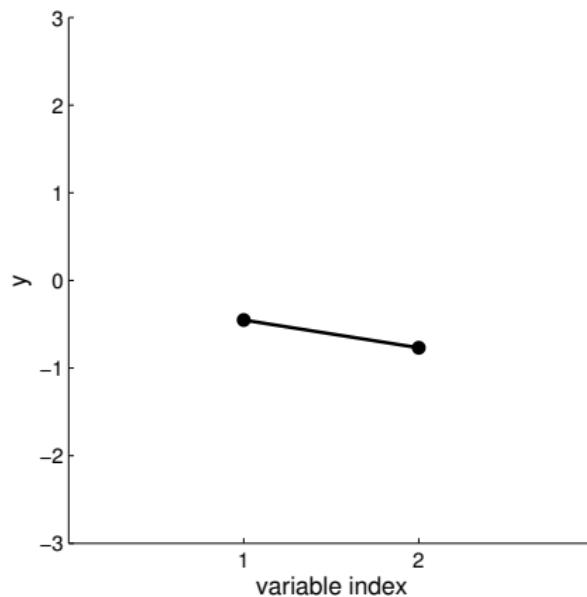
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



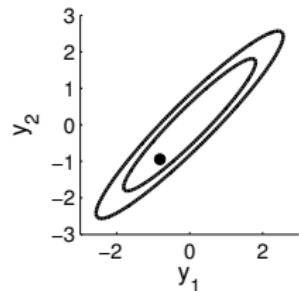
New visualisation



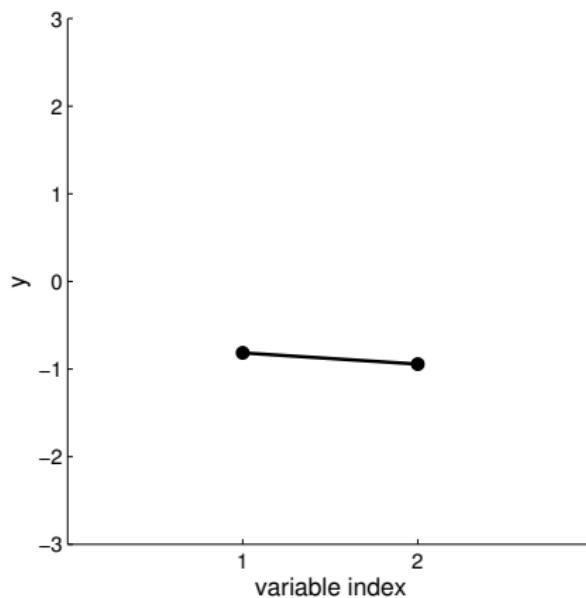
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



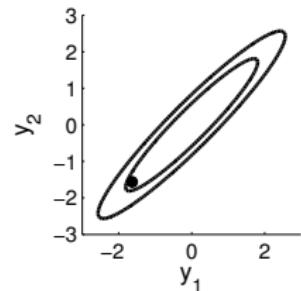
New visualisation



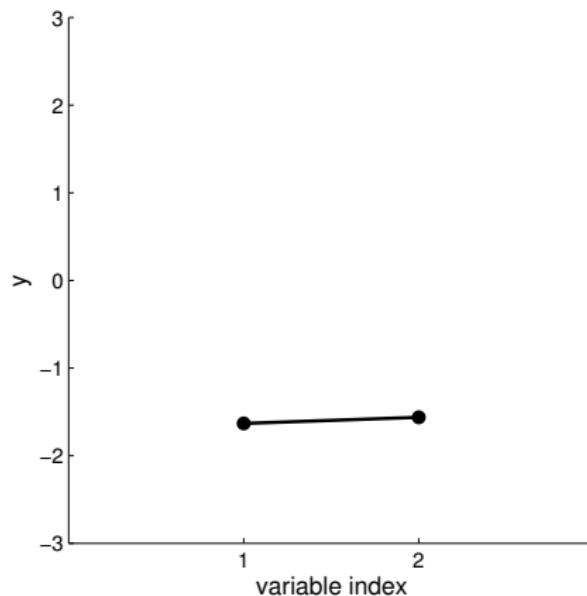
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



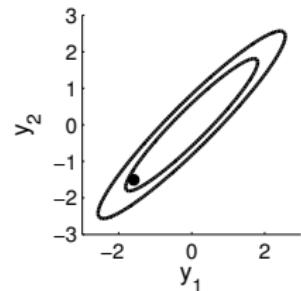
New visualisation



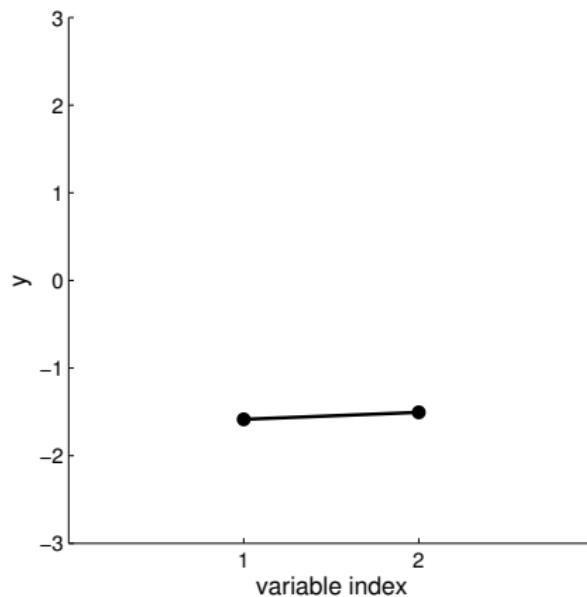
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



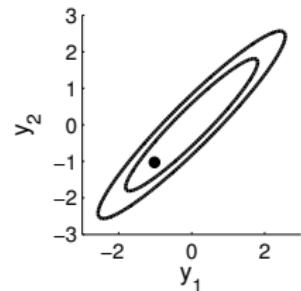
New visualisation



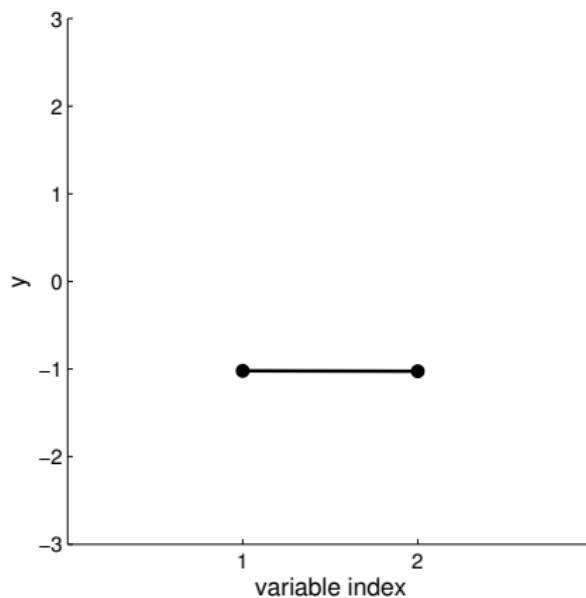
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



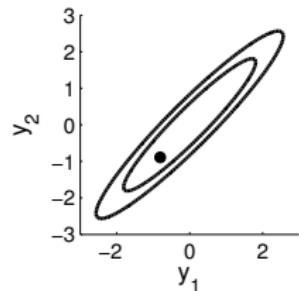
New visualisation



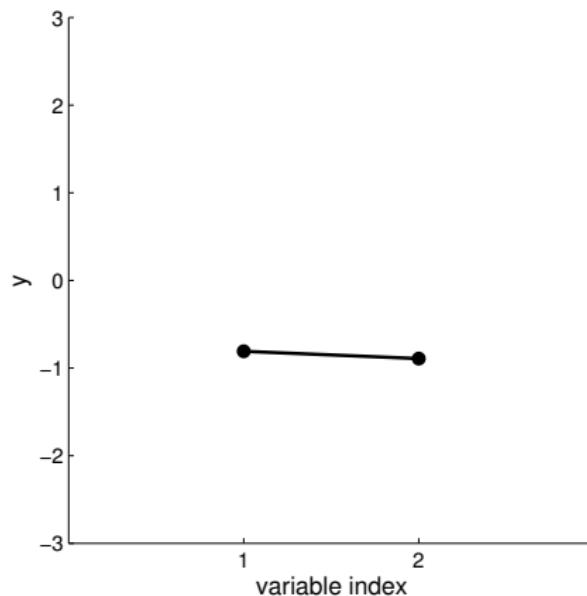
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



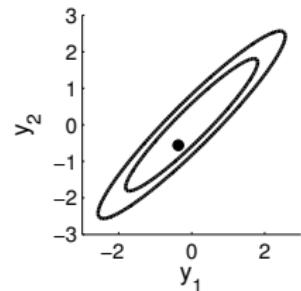
New visualisation



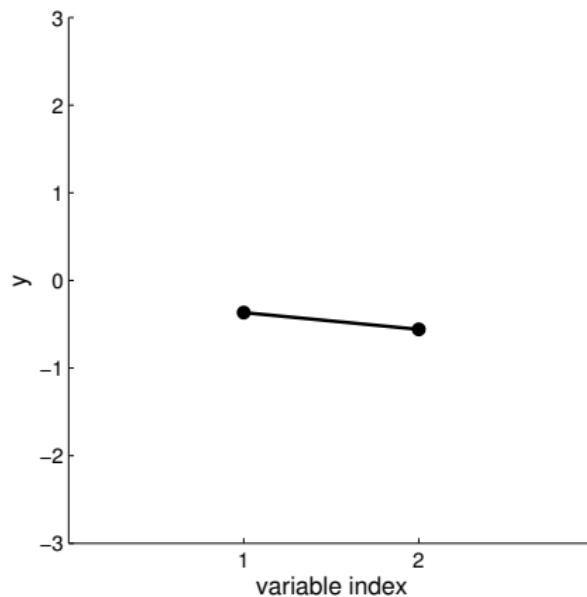
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



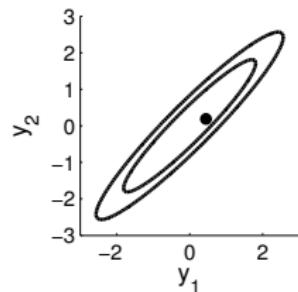
New visualisation



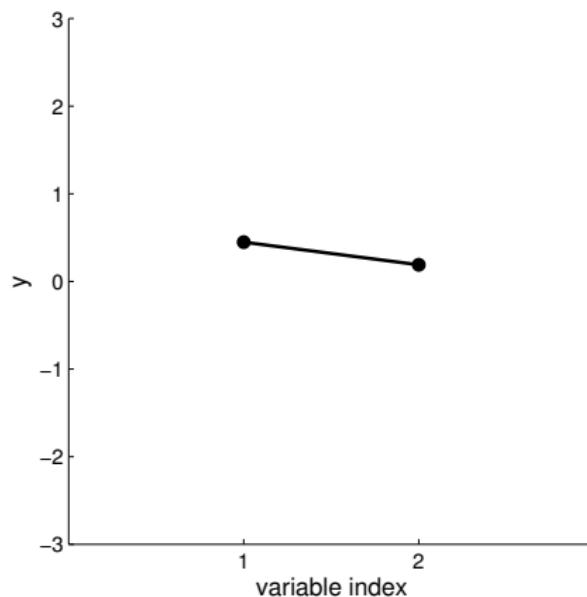
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



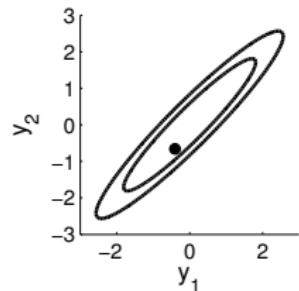
New visualisation



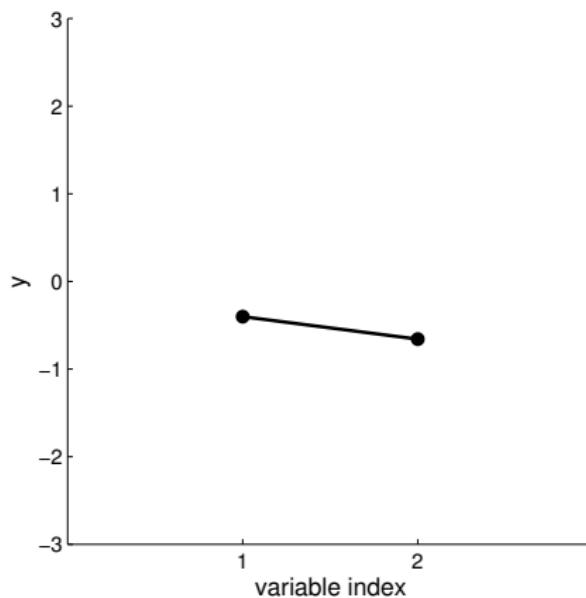
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



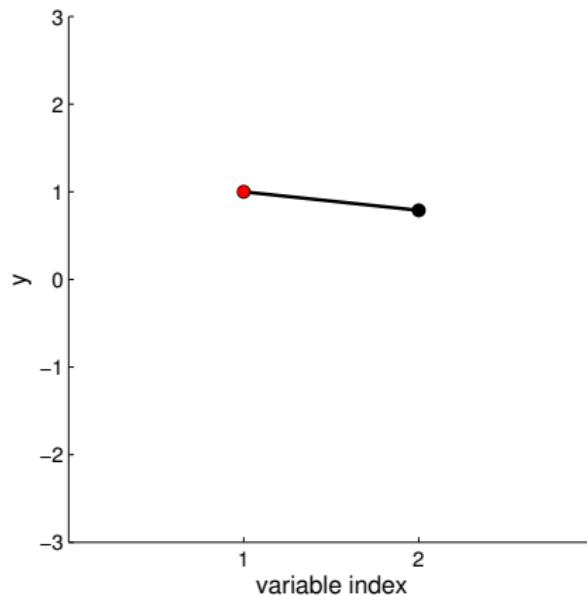
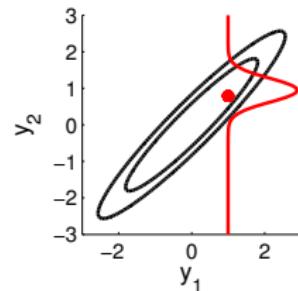
New visualisation



$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

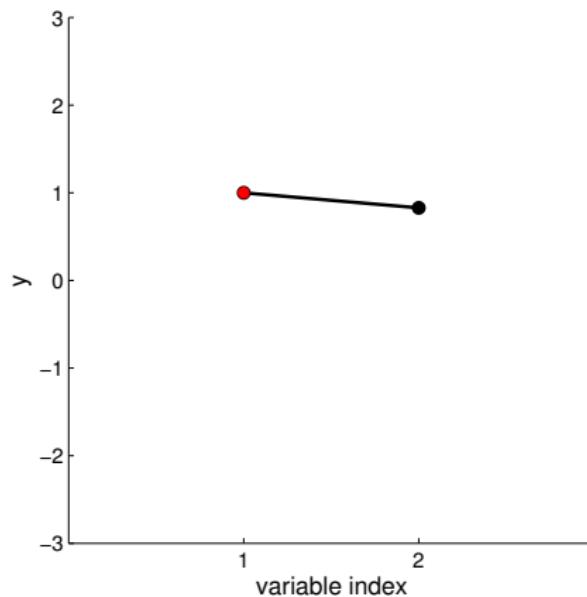
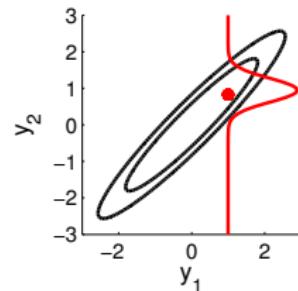


New visualisation



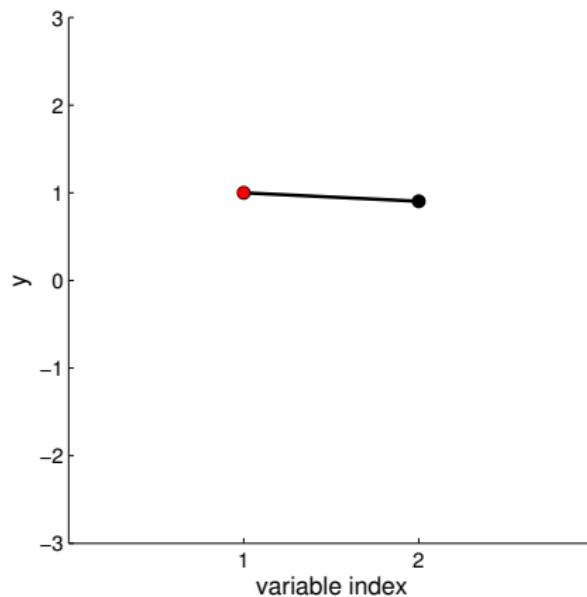
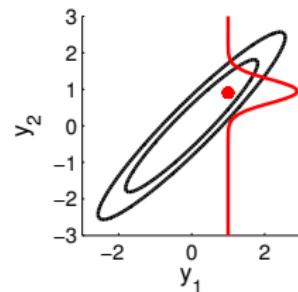
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



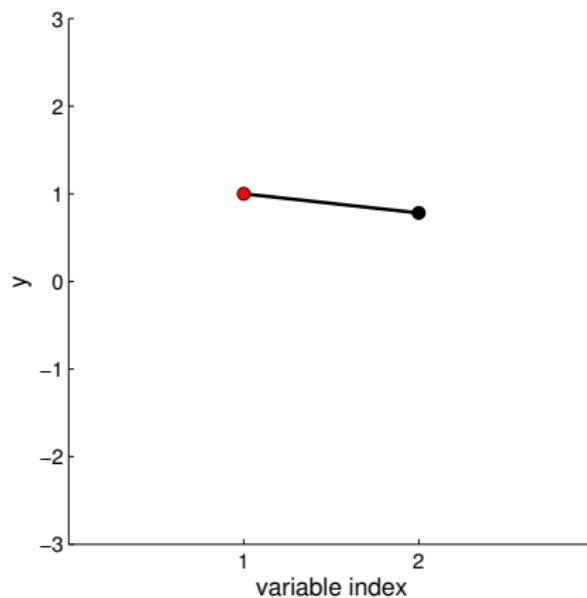
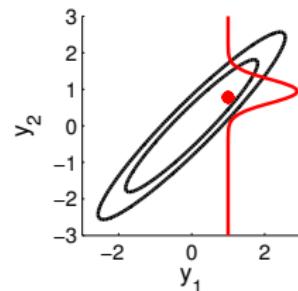
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



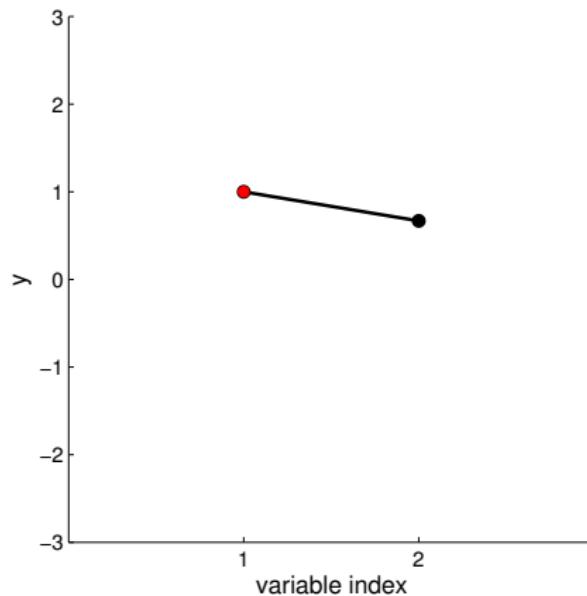
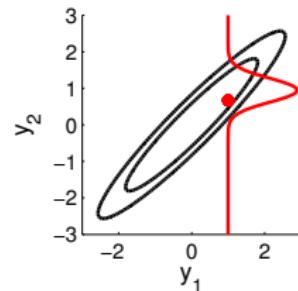
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



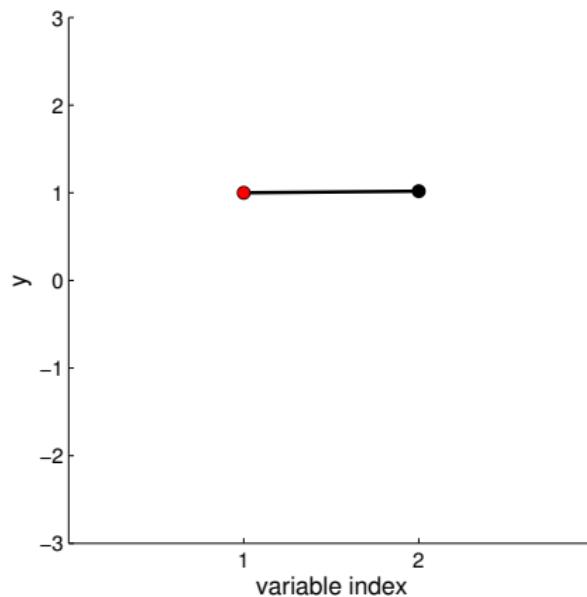
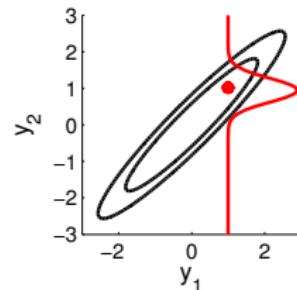
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



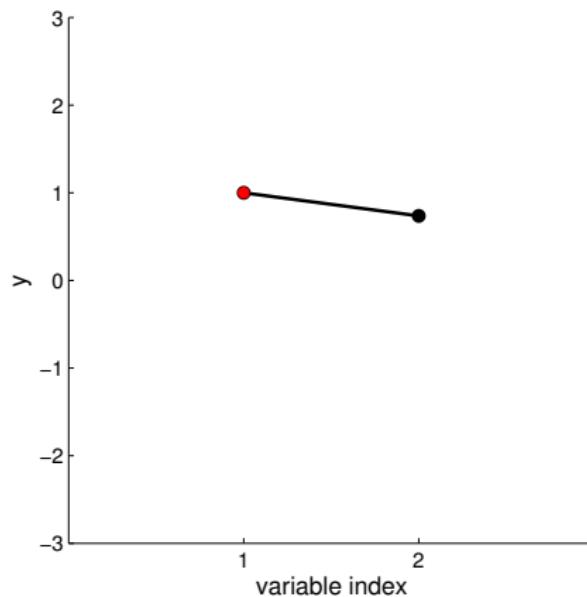
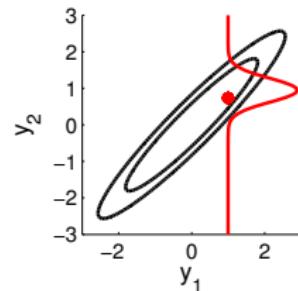
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



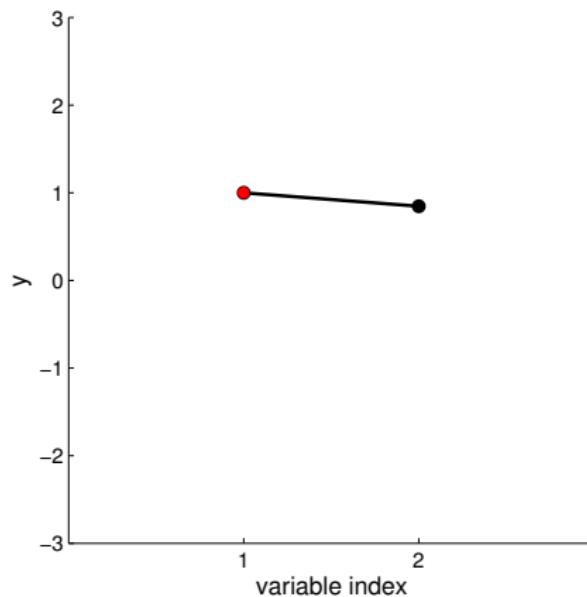
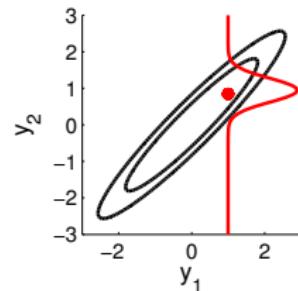
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



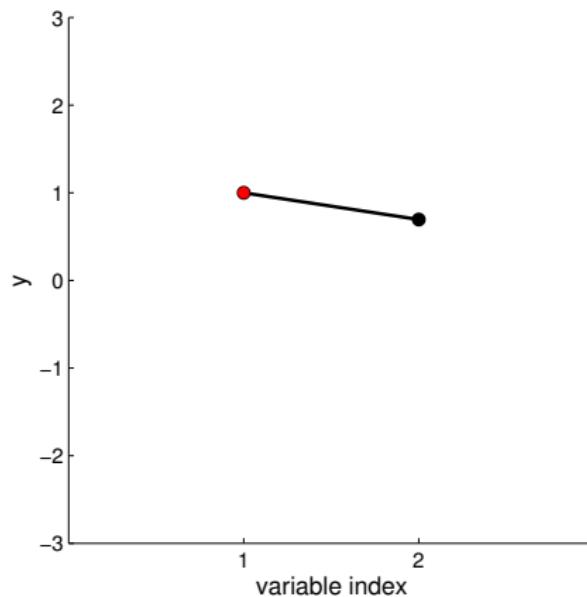
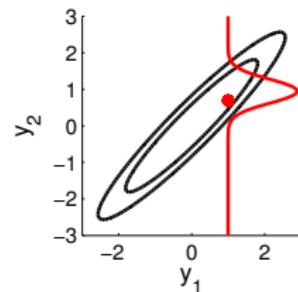
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



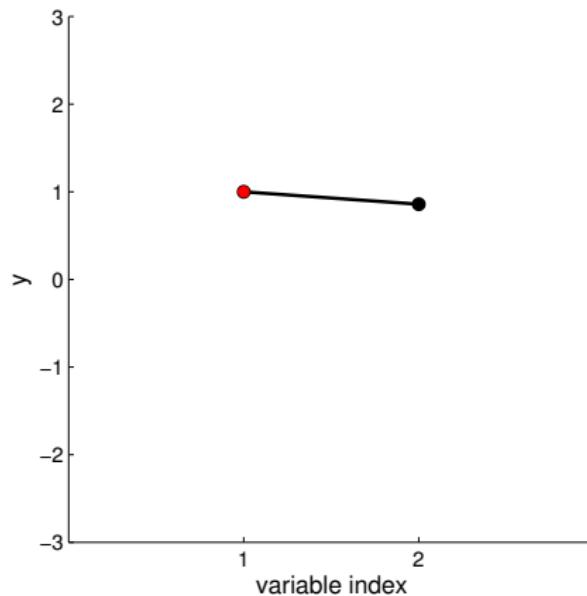
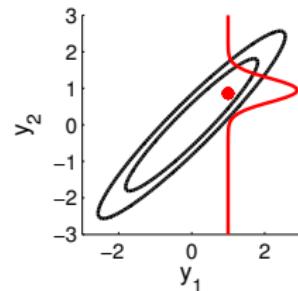
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



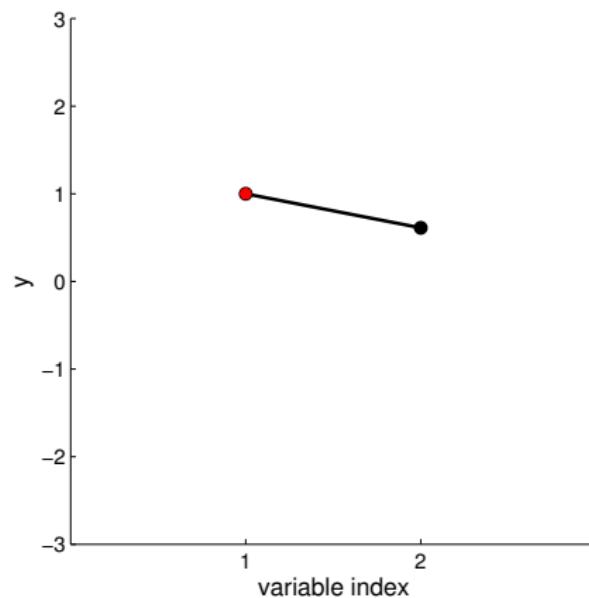
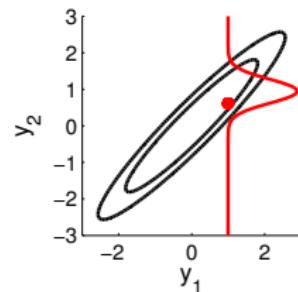
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



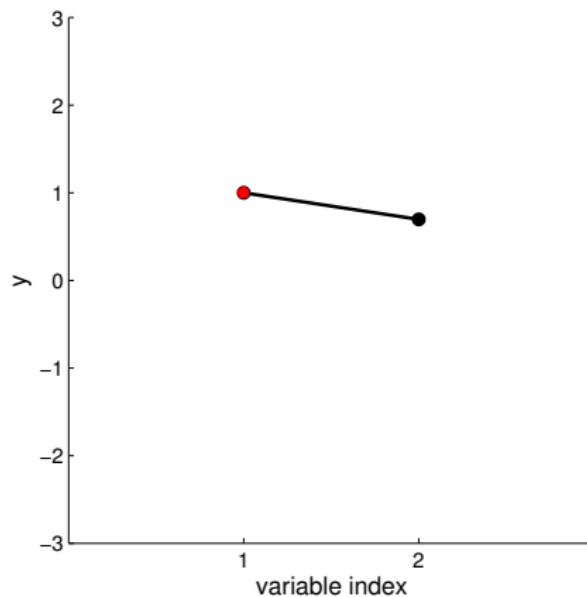
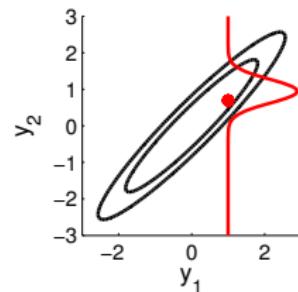
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



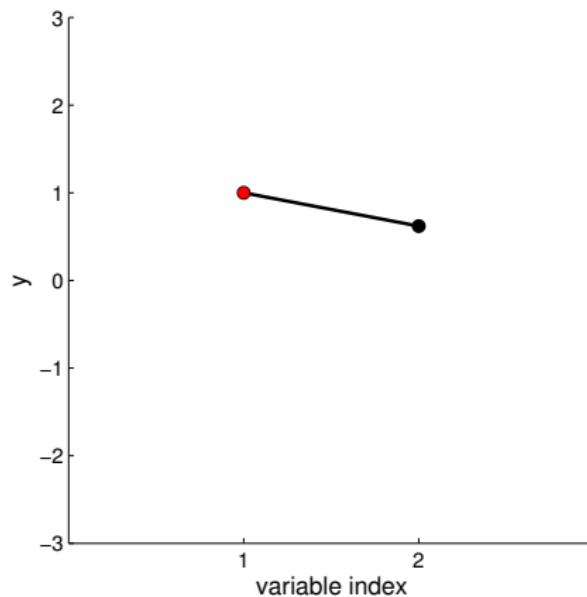
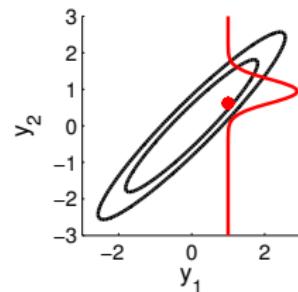
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



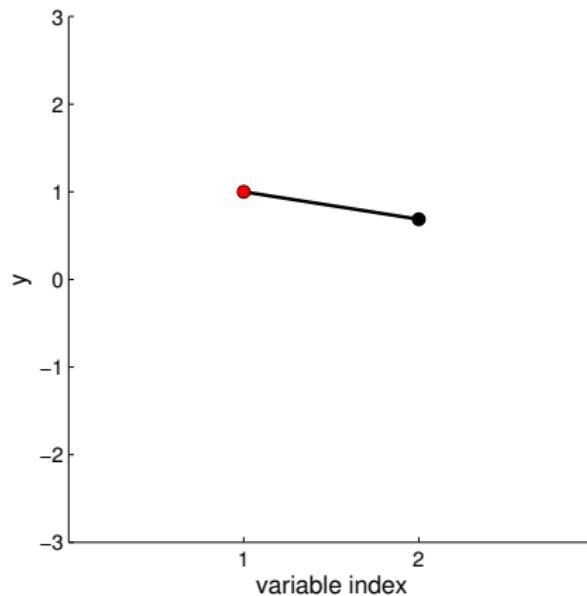
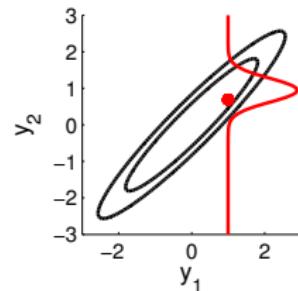
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



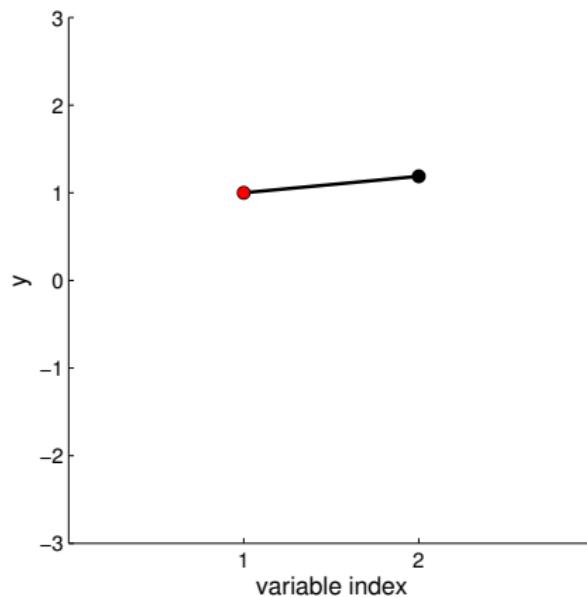
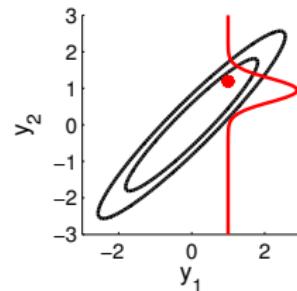
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



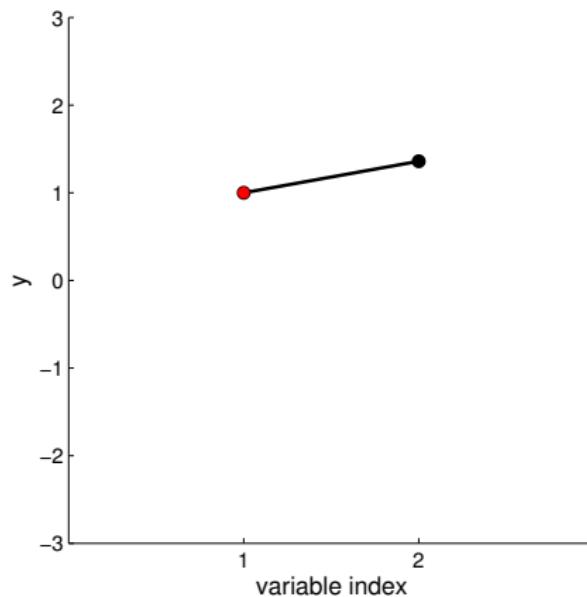
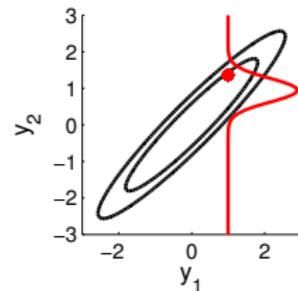
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



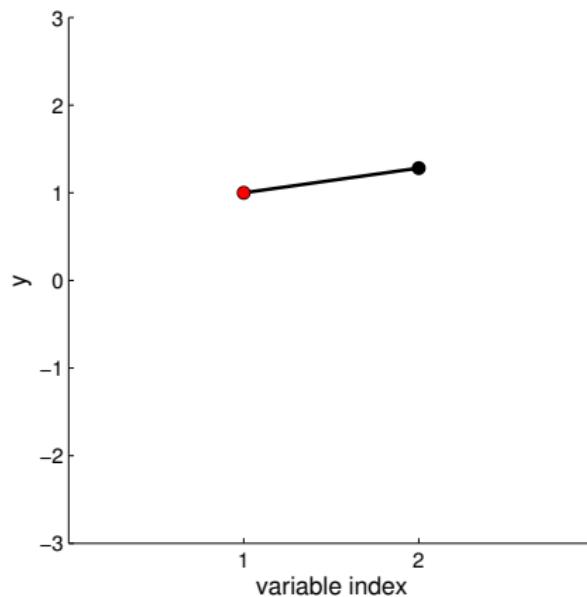
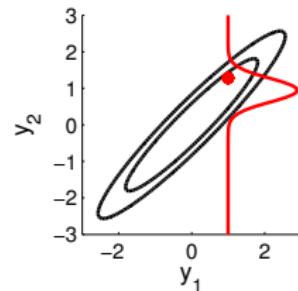
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



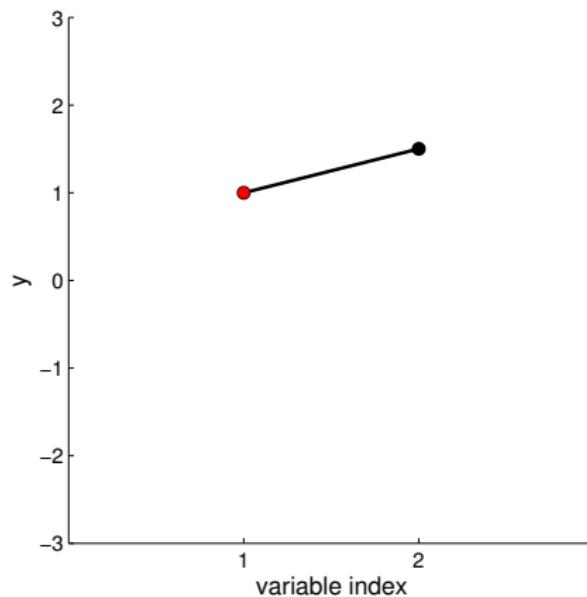
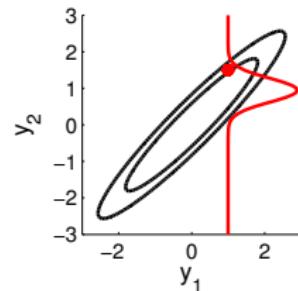
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



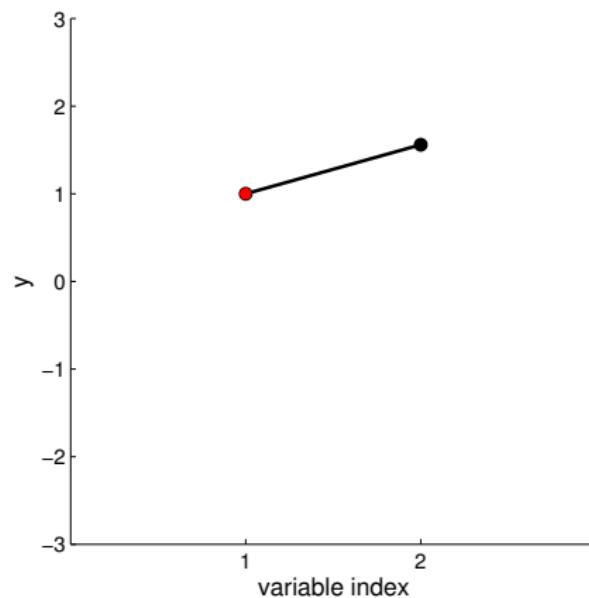
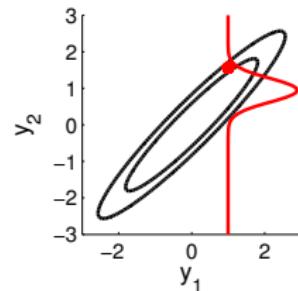
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



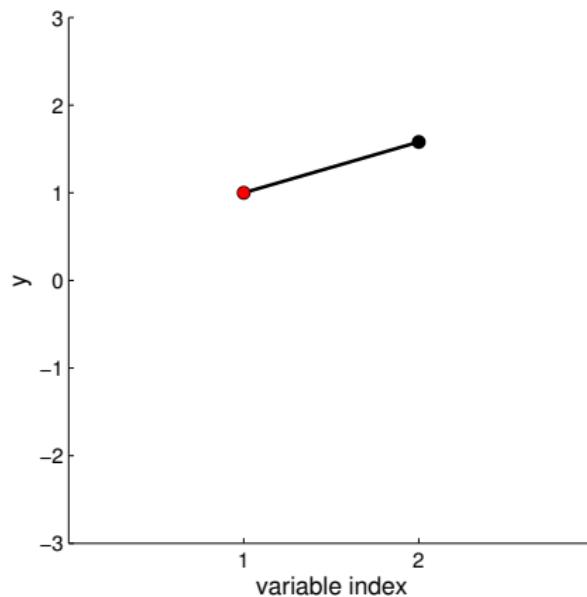
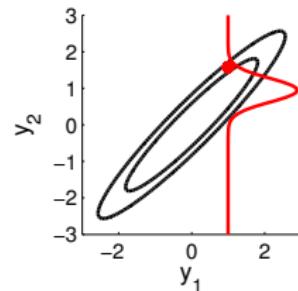
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



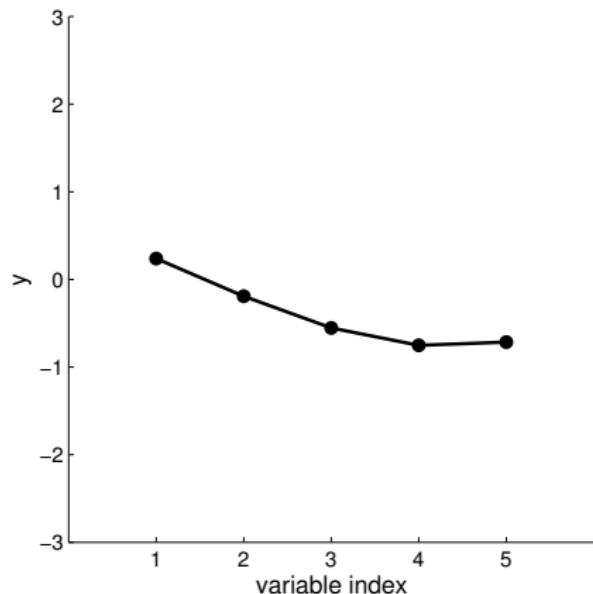
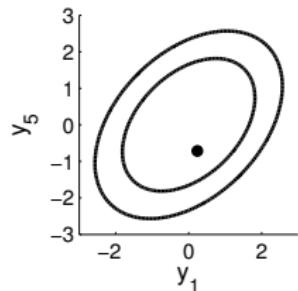
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



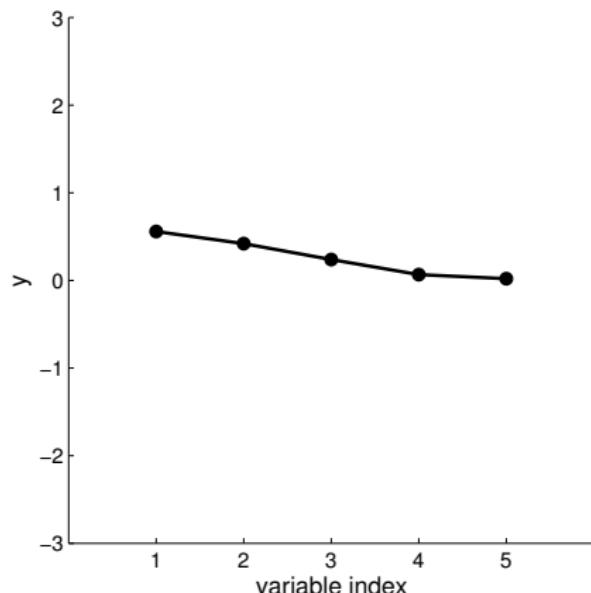
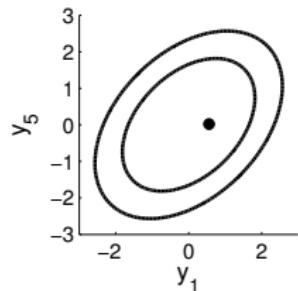
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation



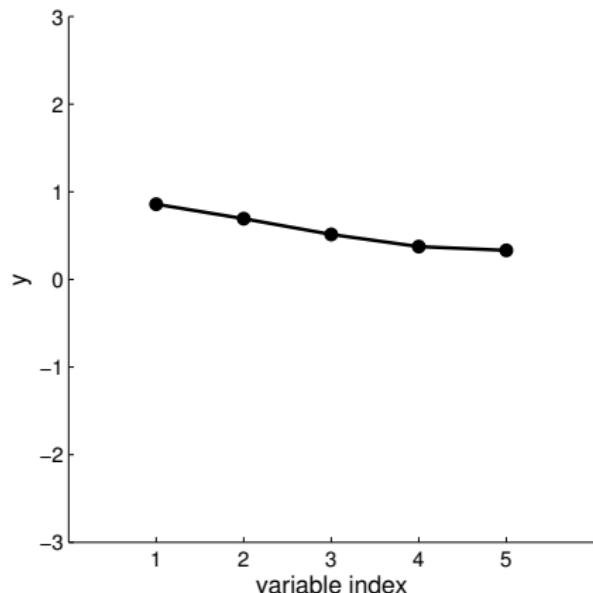
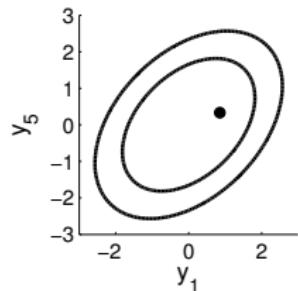
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



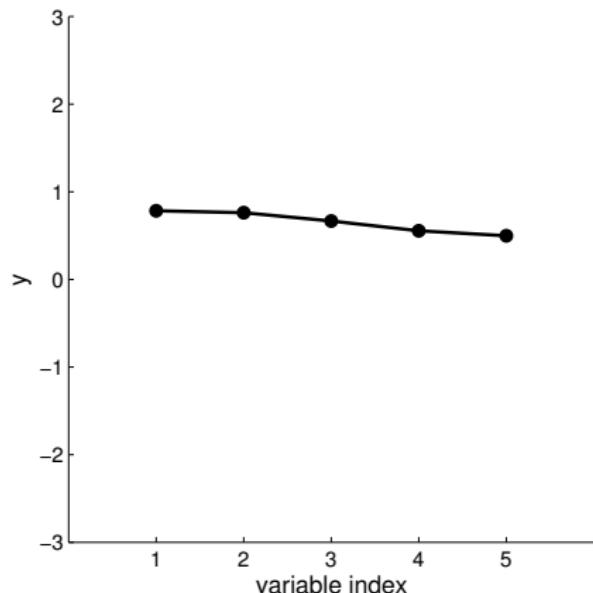
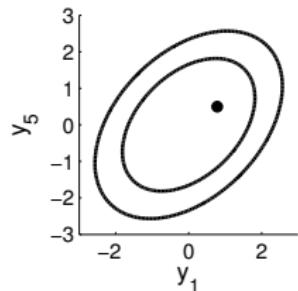
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



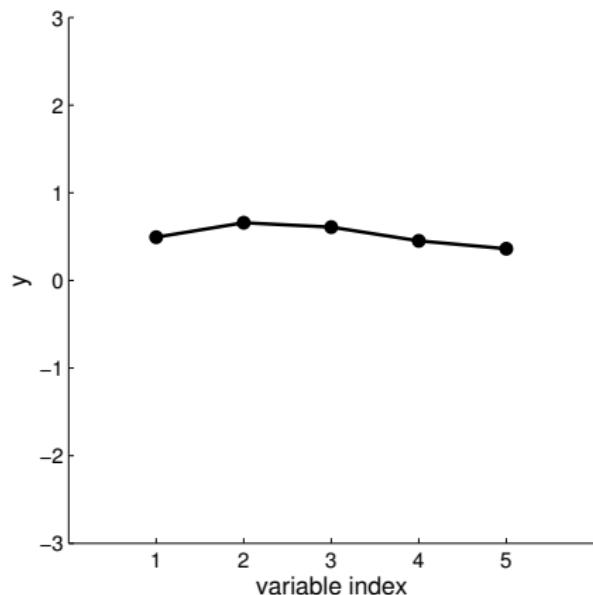
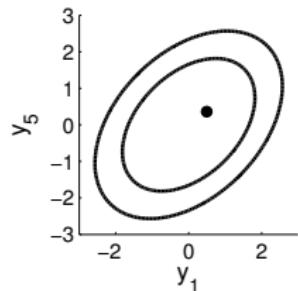
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



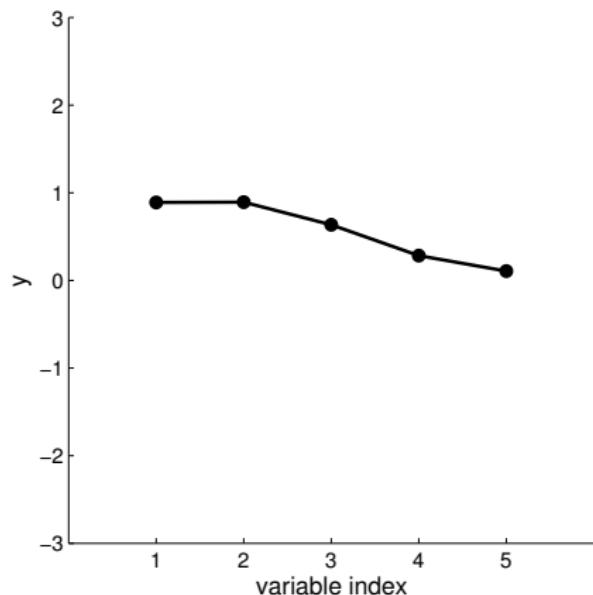
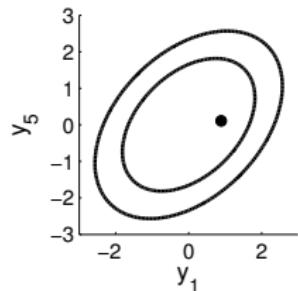
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



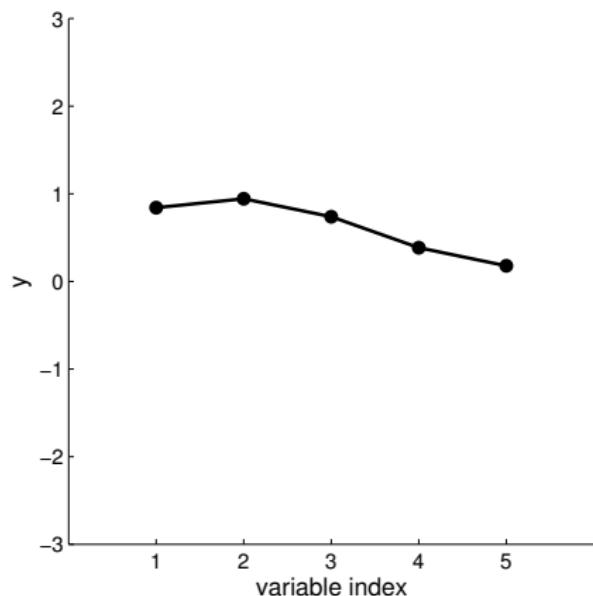
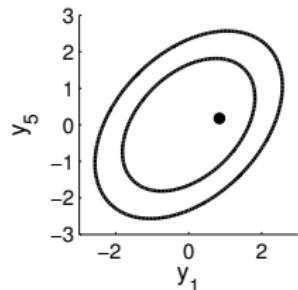
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



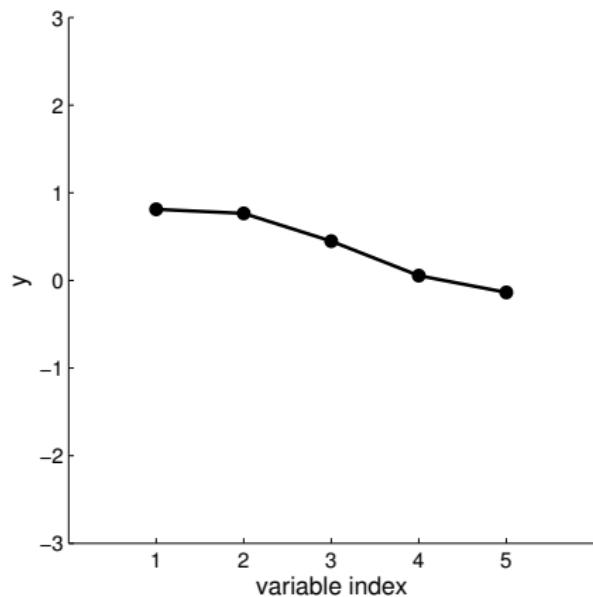
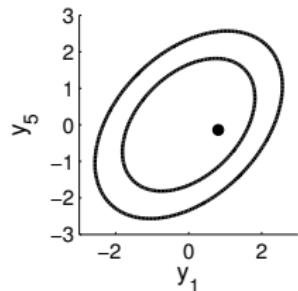
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



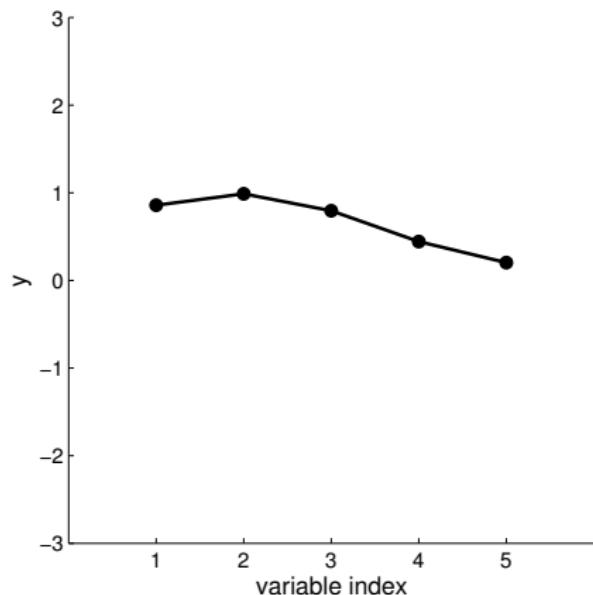
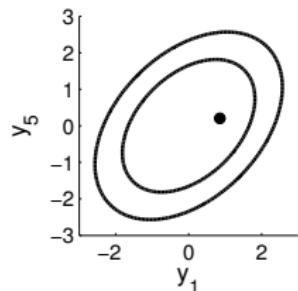
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



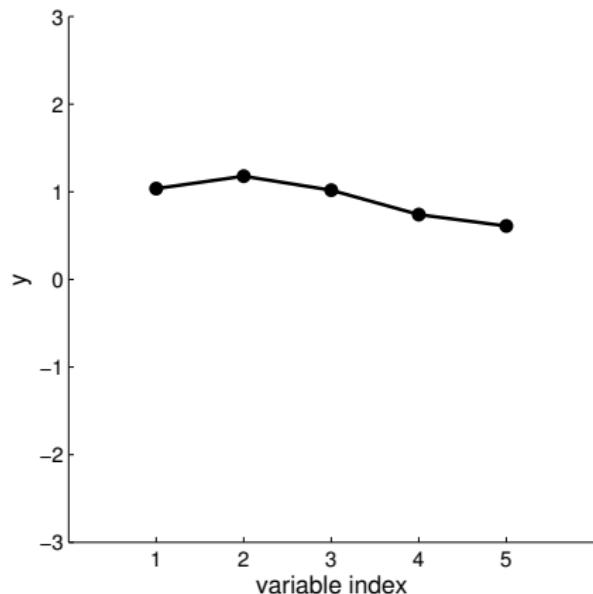
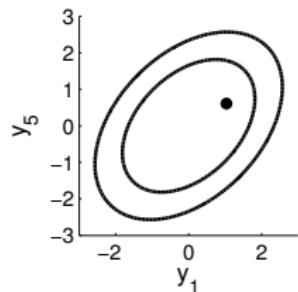
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



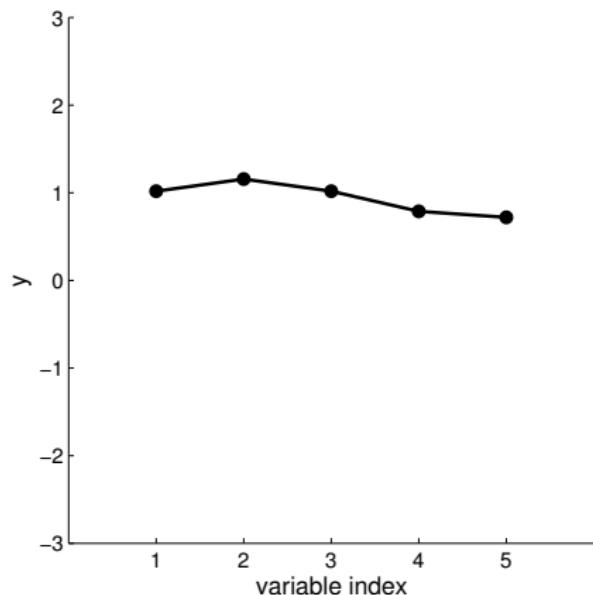
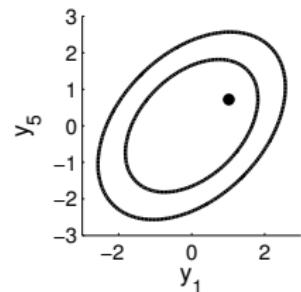
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



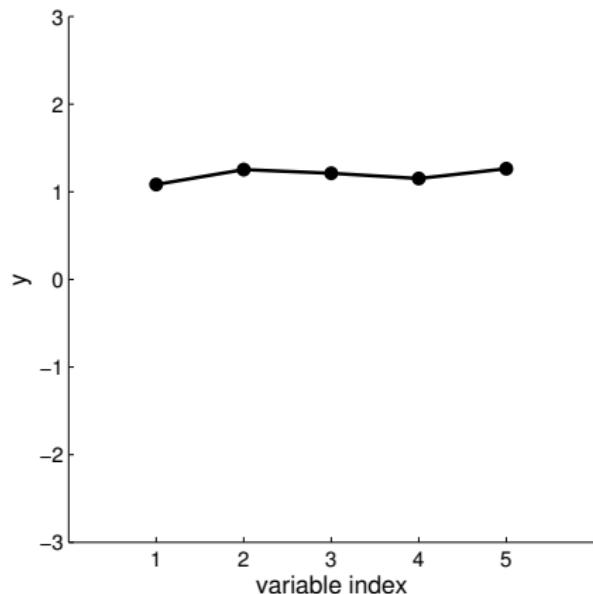
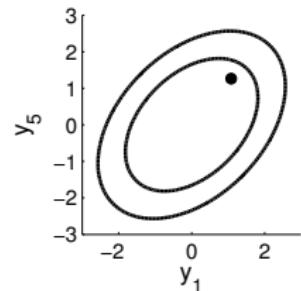
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



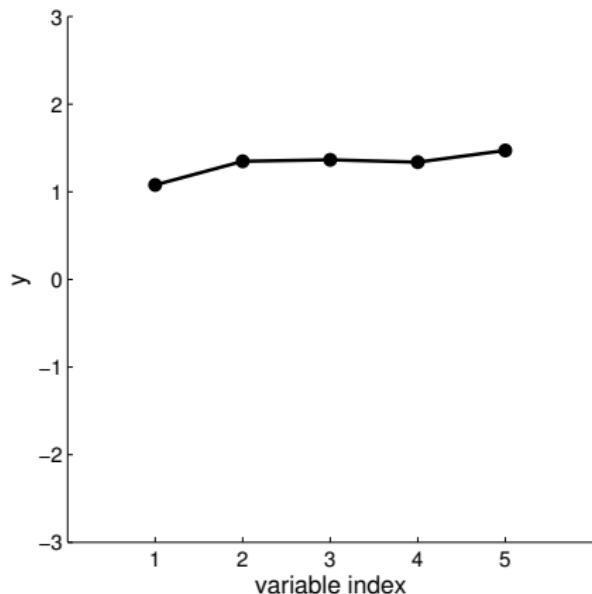
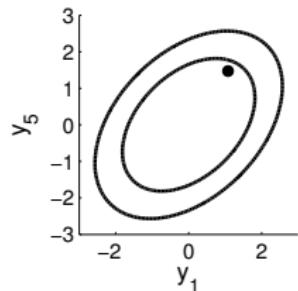
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



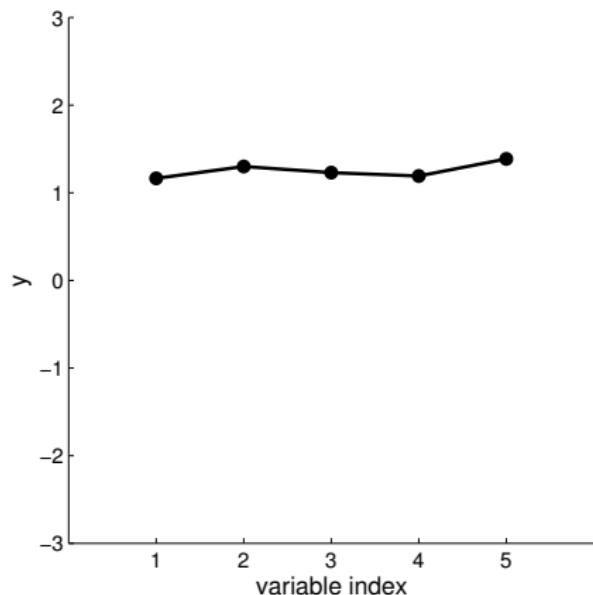
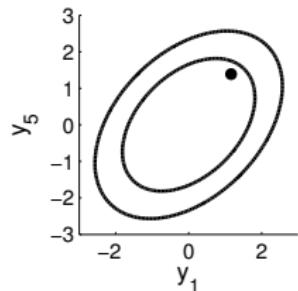
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



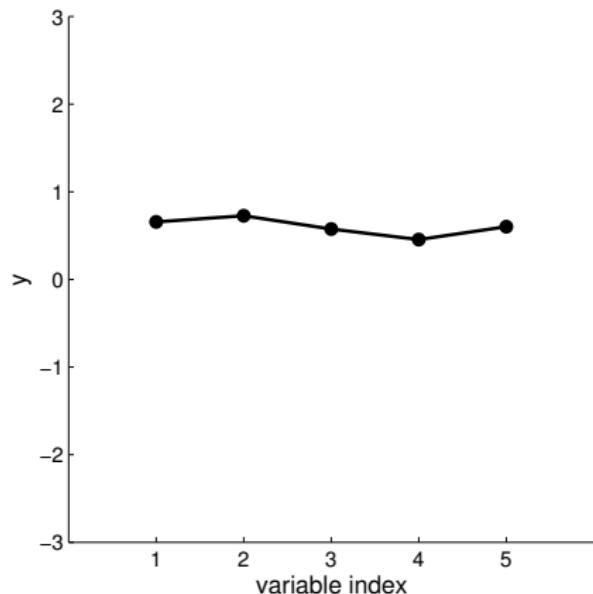
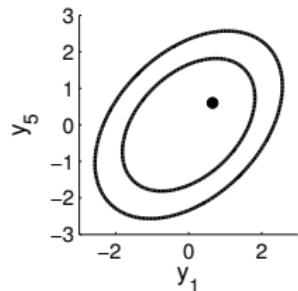
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



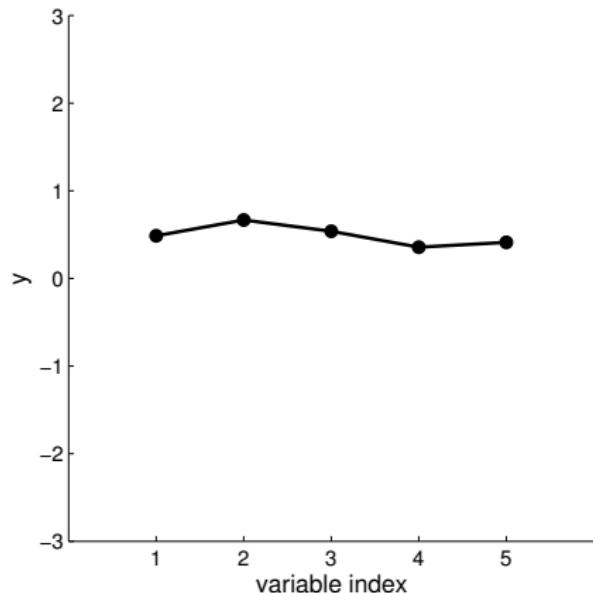
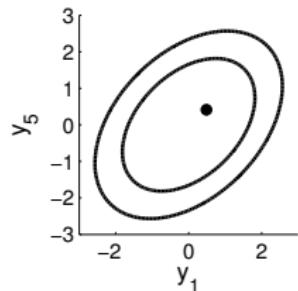
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



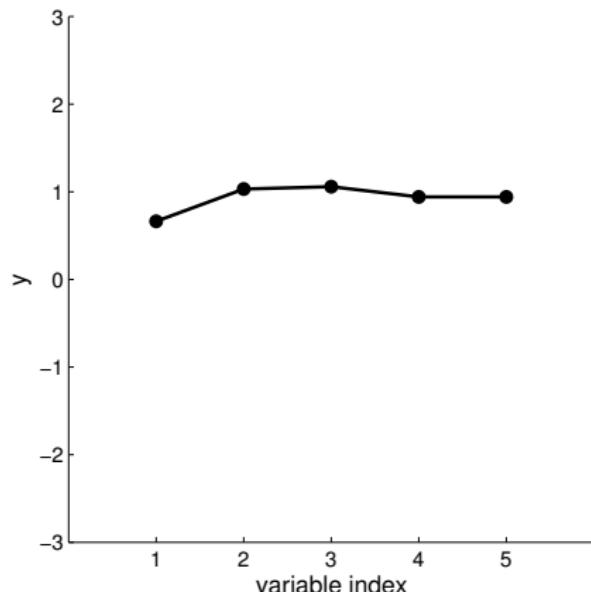
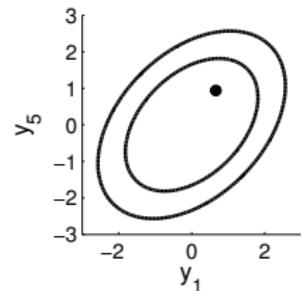
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



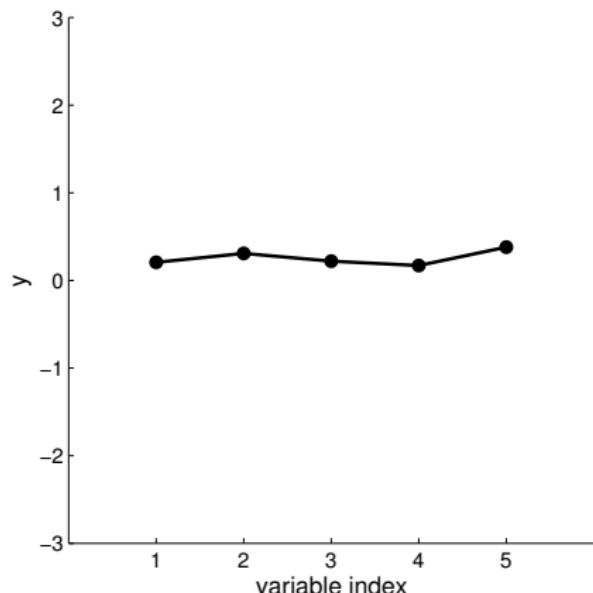
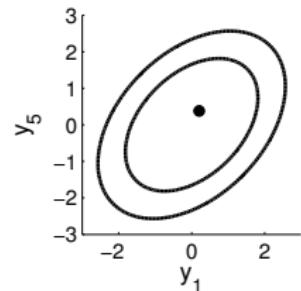
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



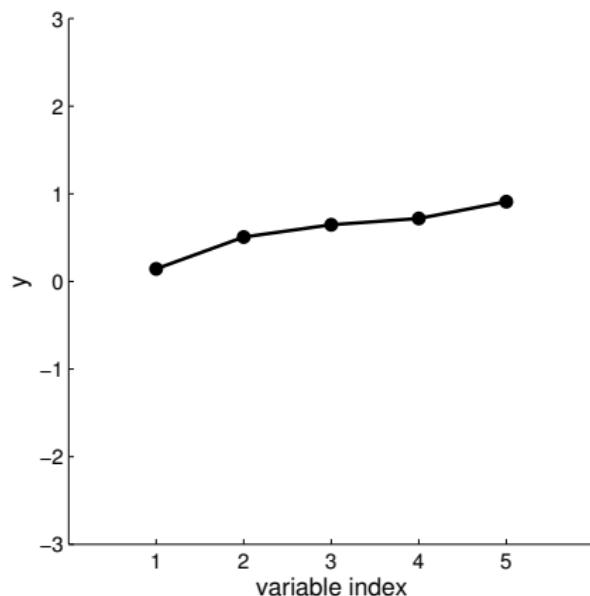
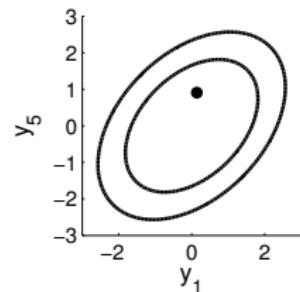
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



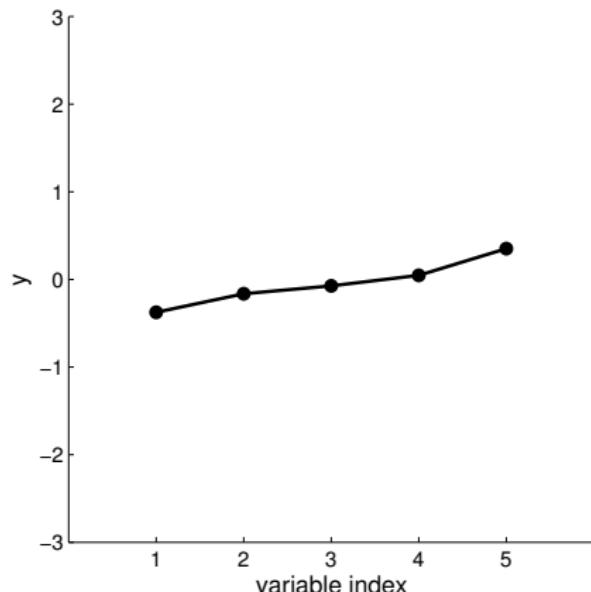
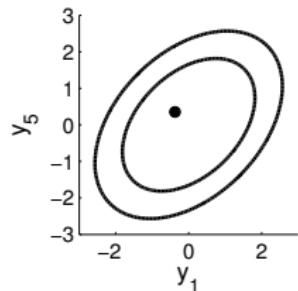
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



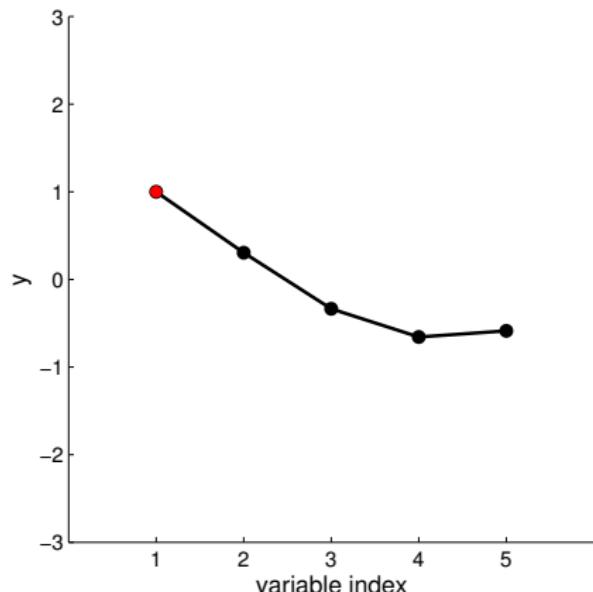
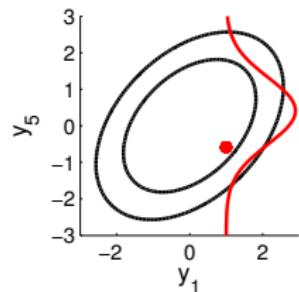
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



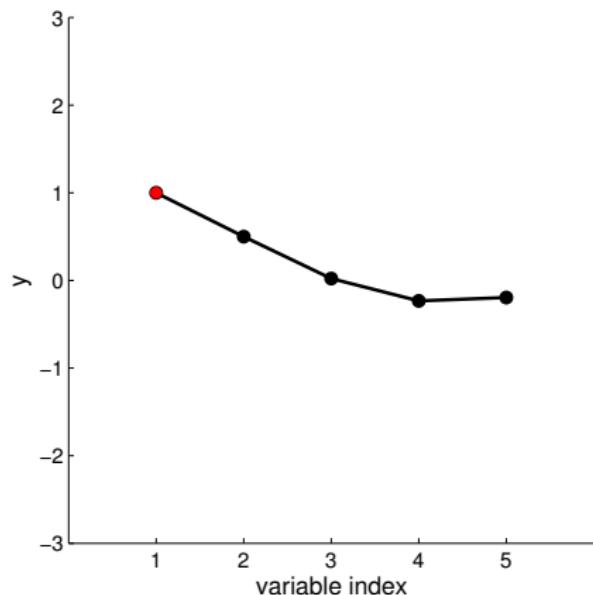
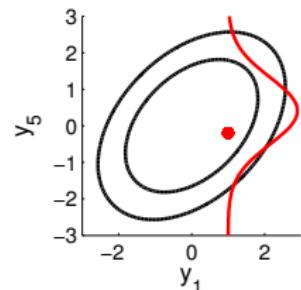
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



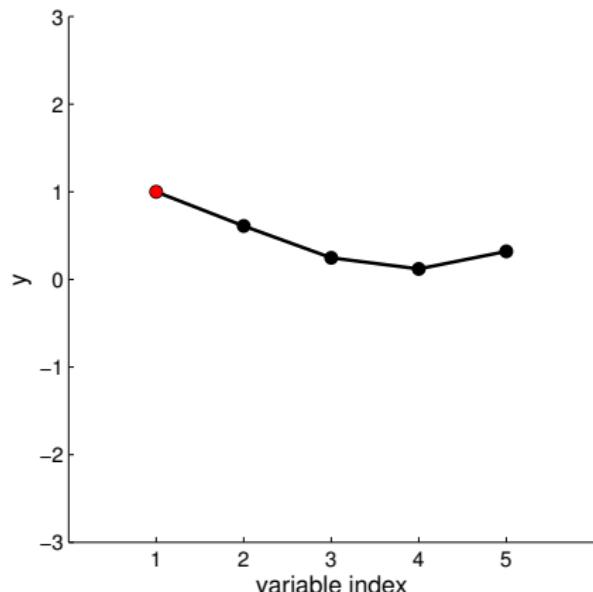
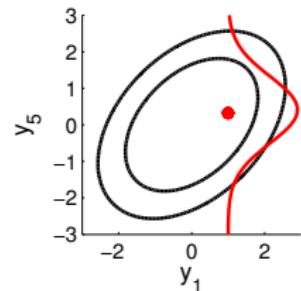
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



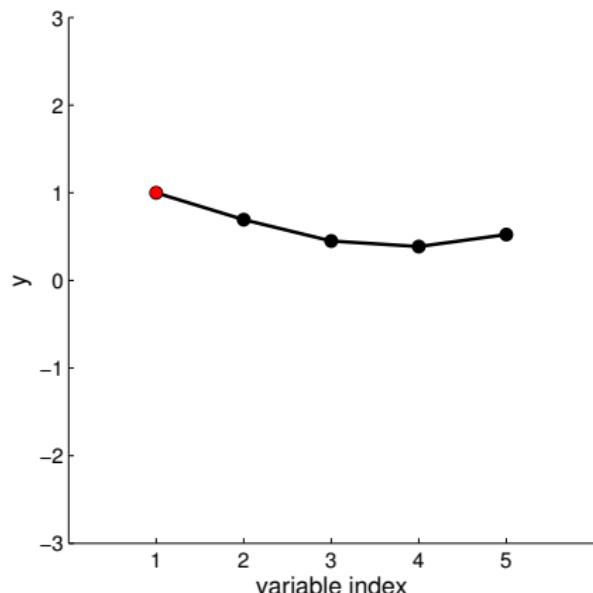
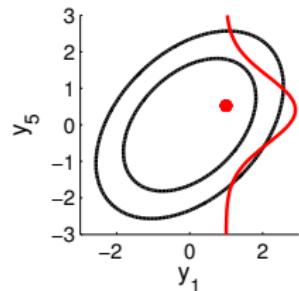
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



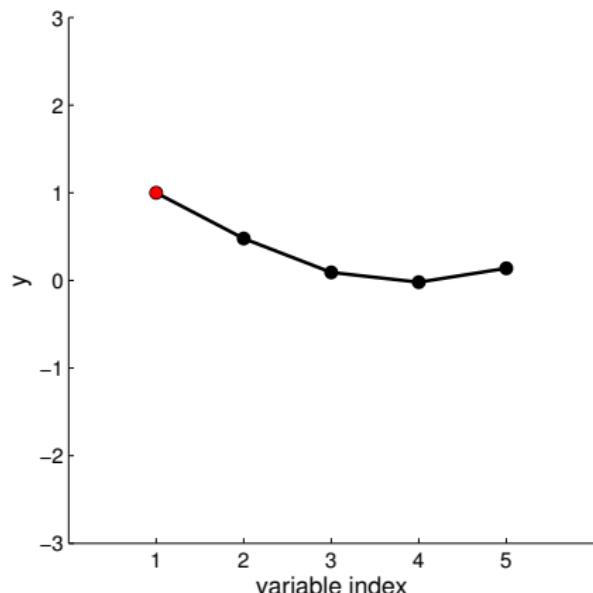
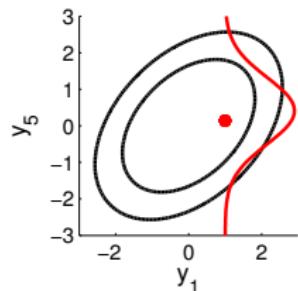
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



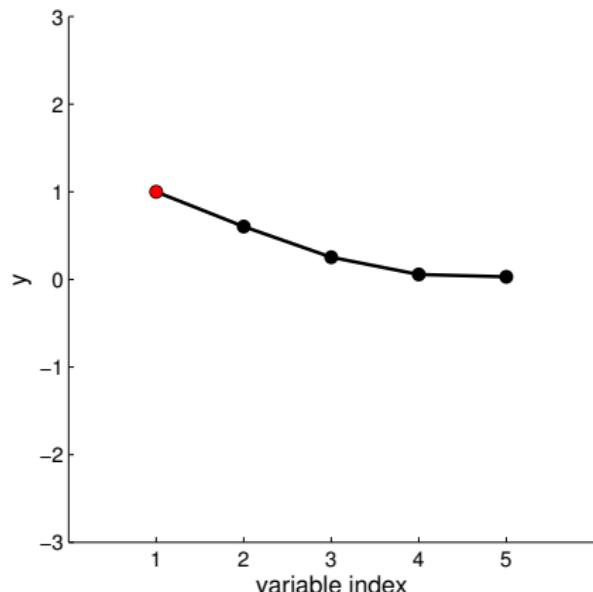
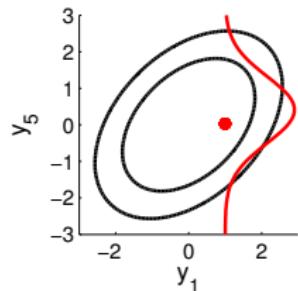
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



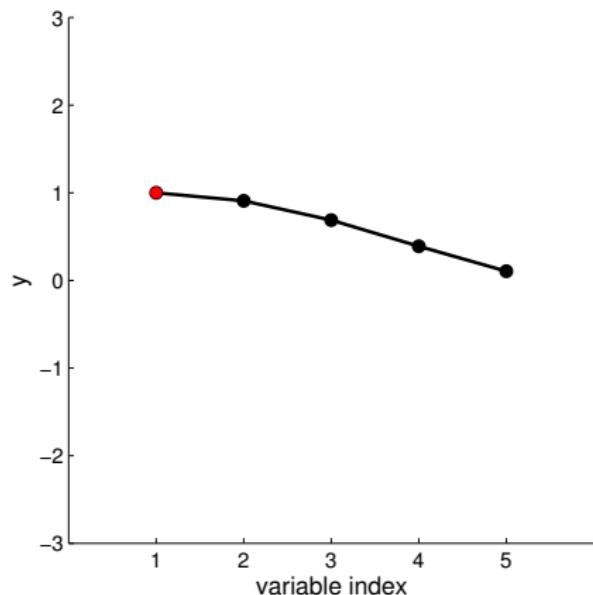
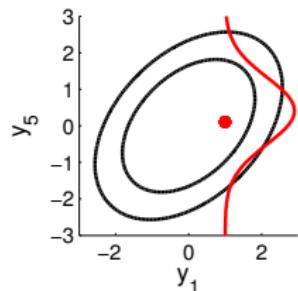
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



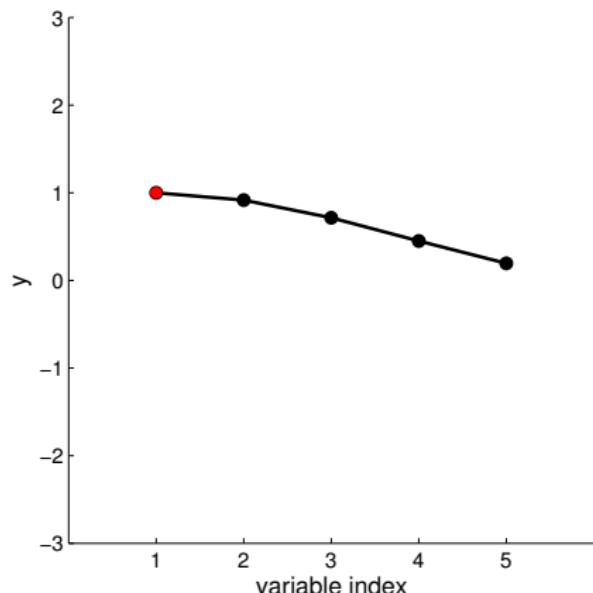
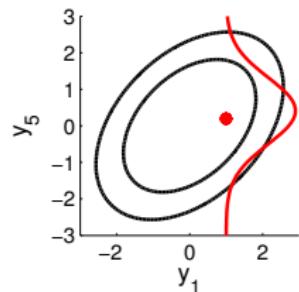
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



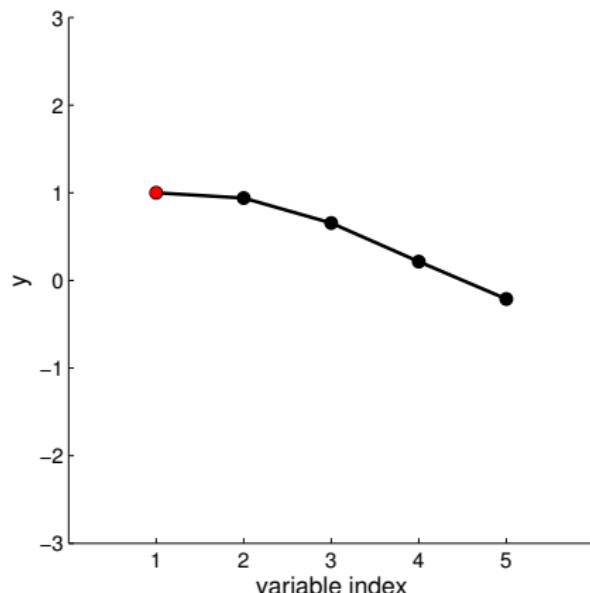
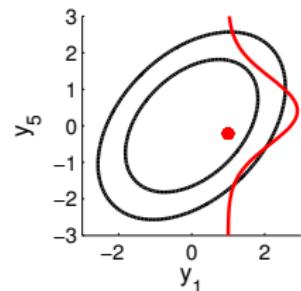
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



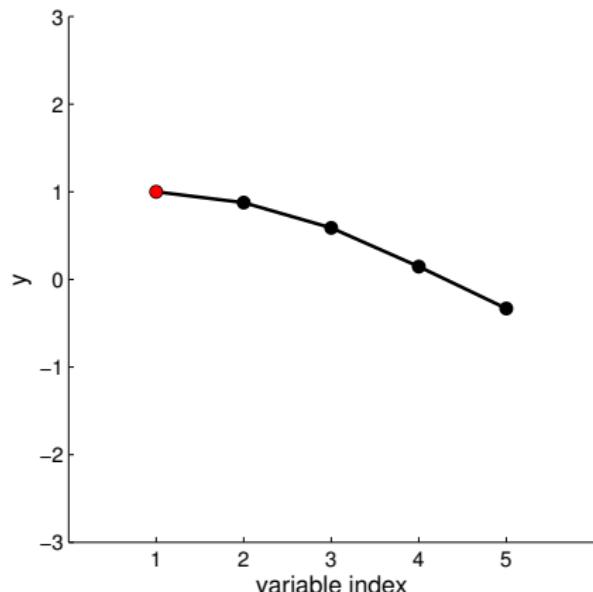
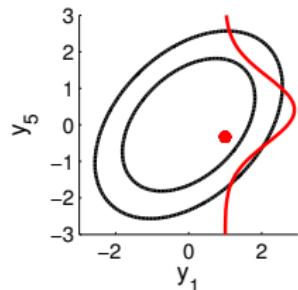
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



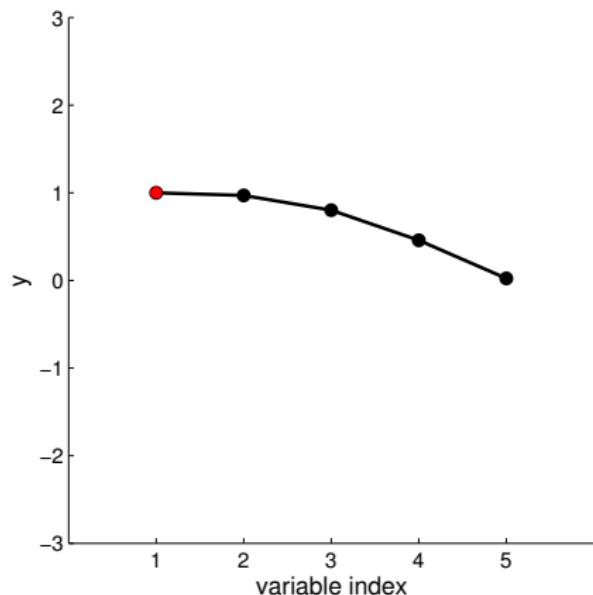
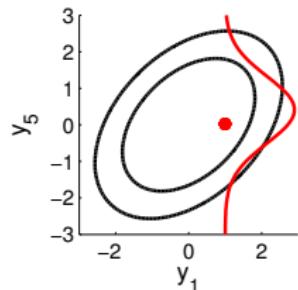
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



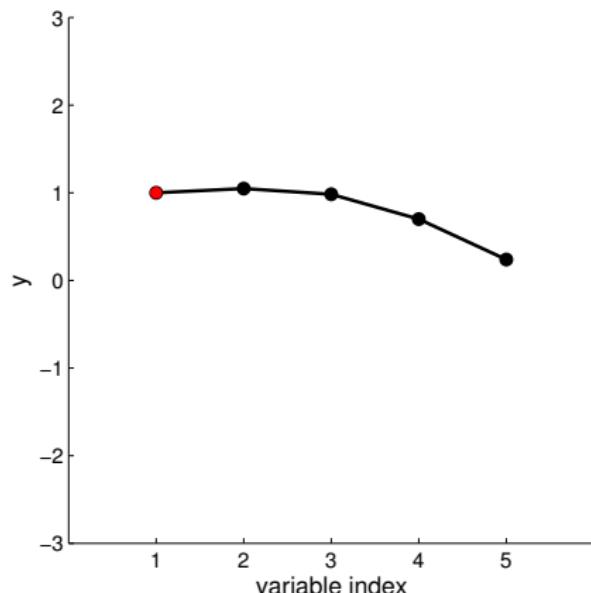
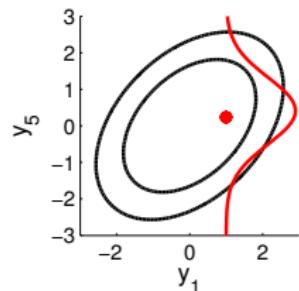
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



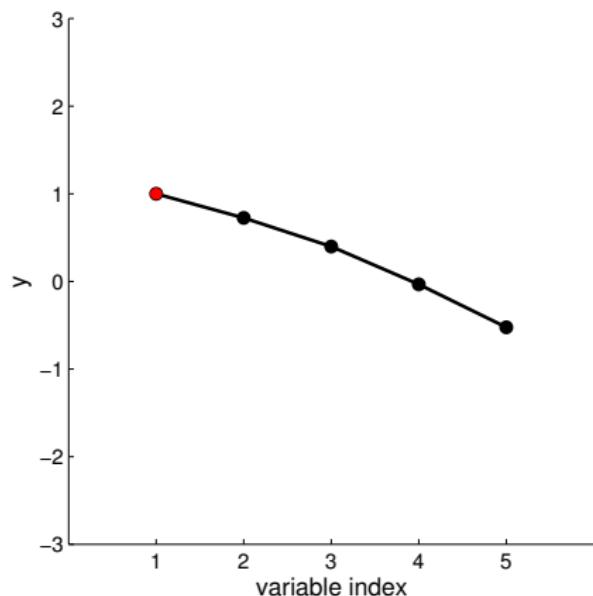
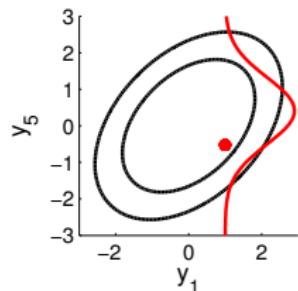
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



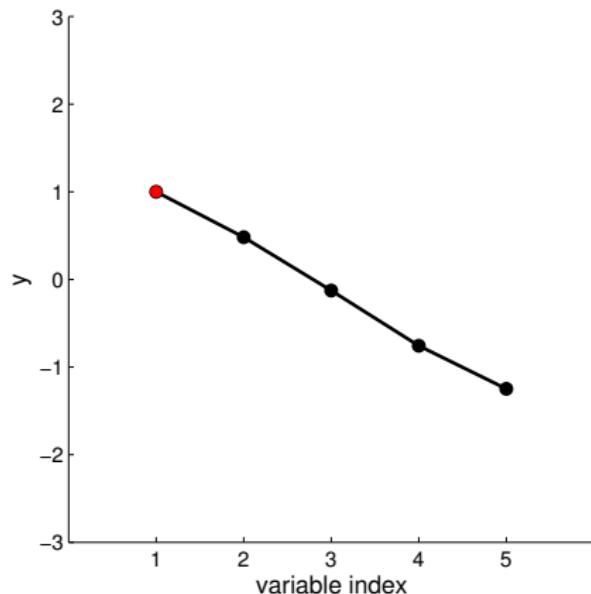
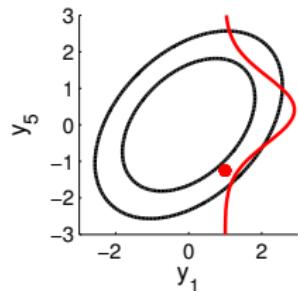
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



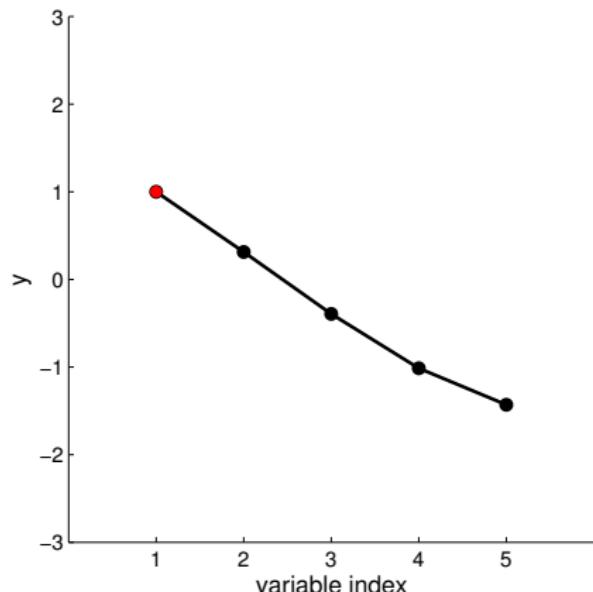
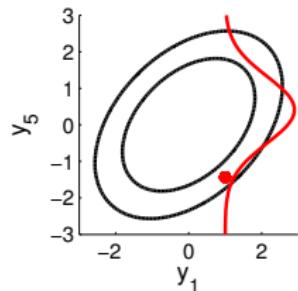
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



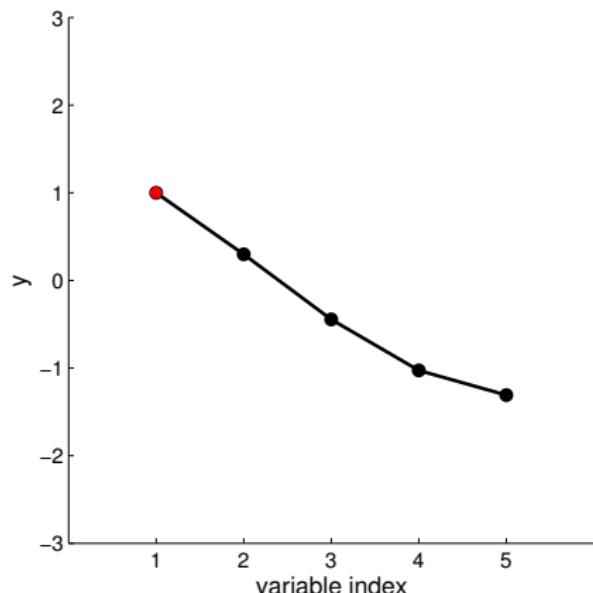
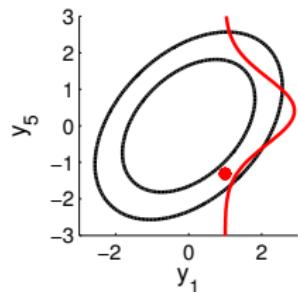
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



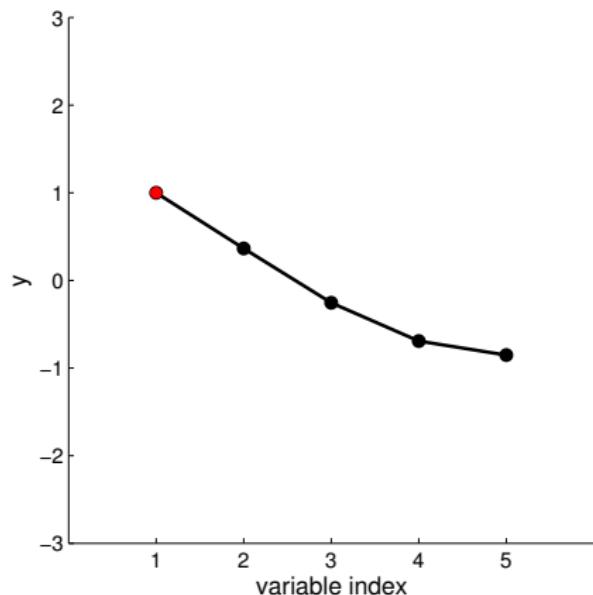
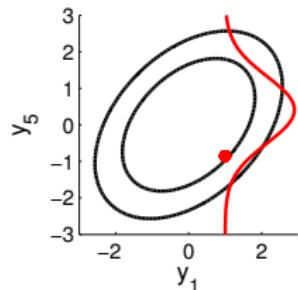
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



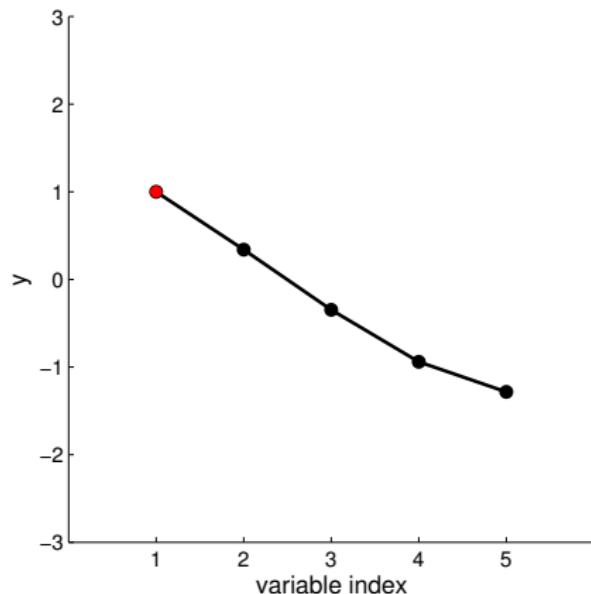
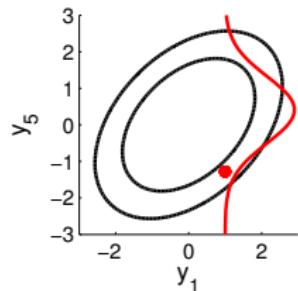
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



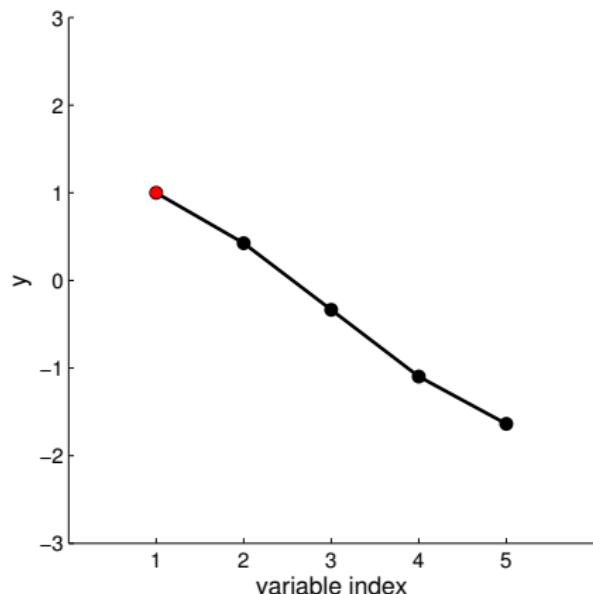
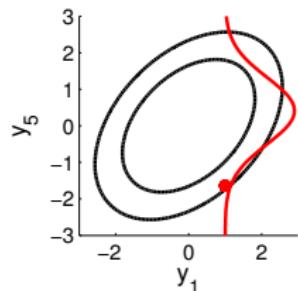
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



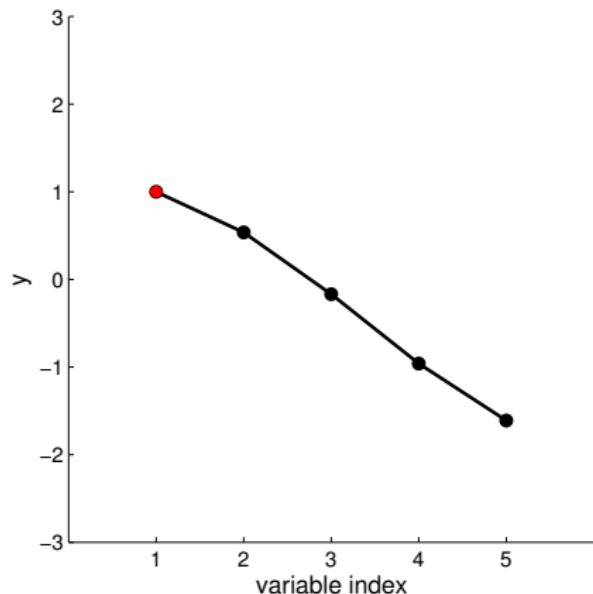
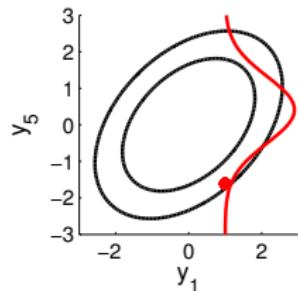
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation



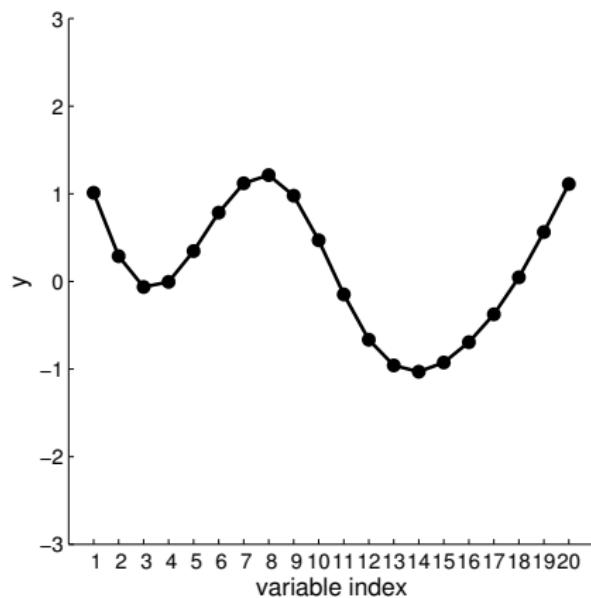
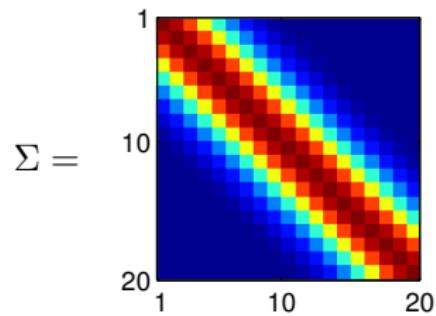
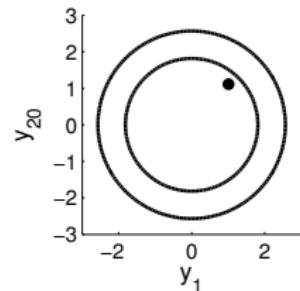
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

New visualisation

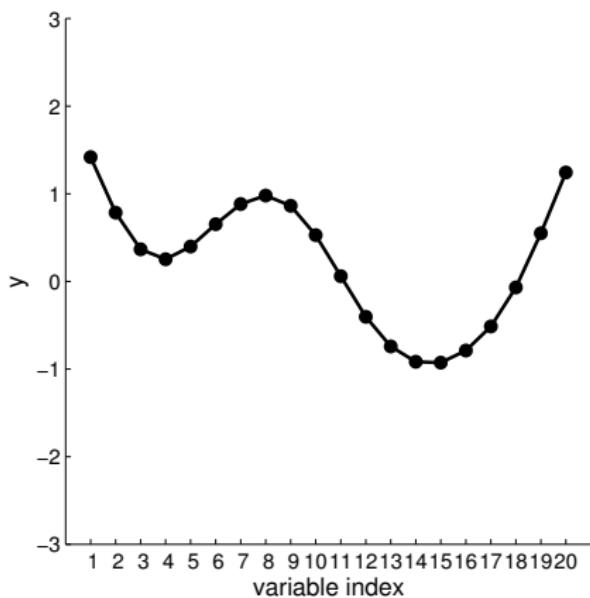
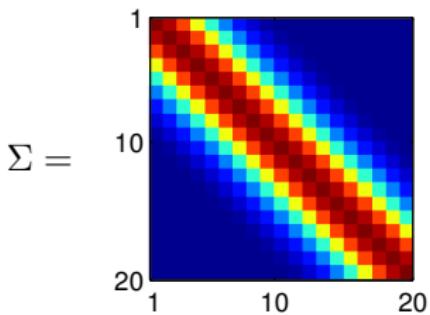
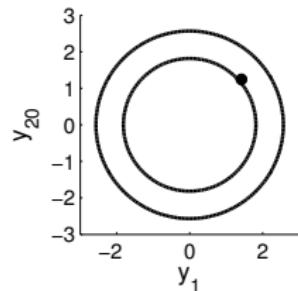


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

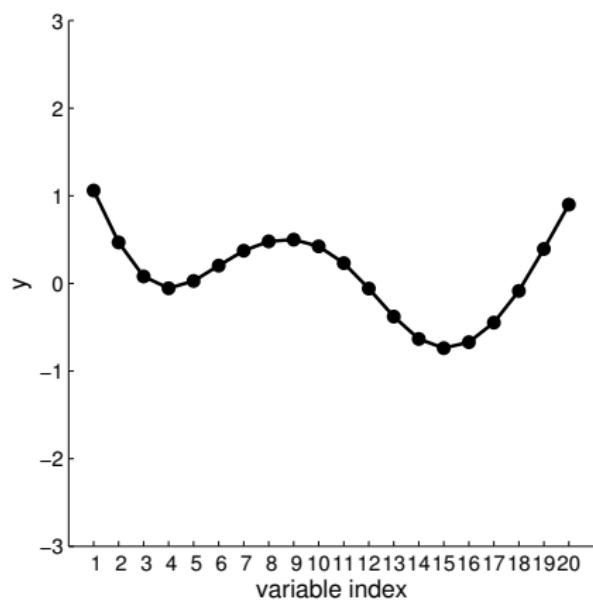
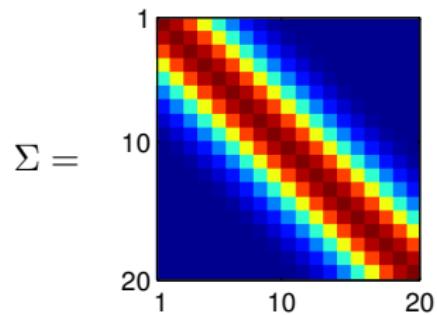
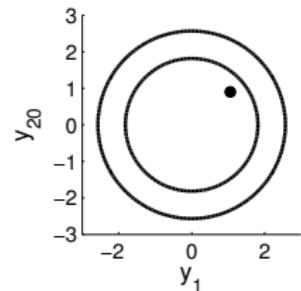
New visualisation



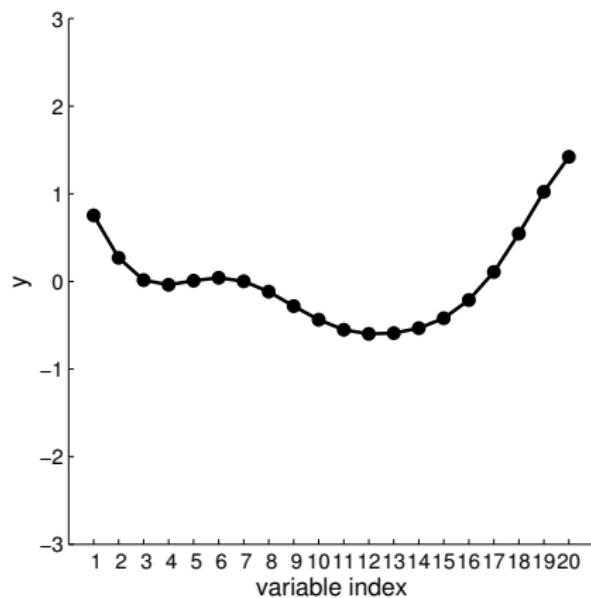
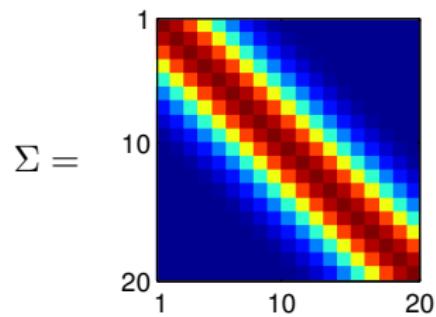
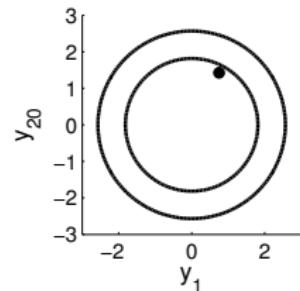
New visualisation



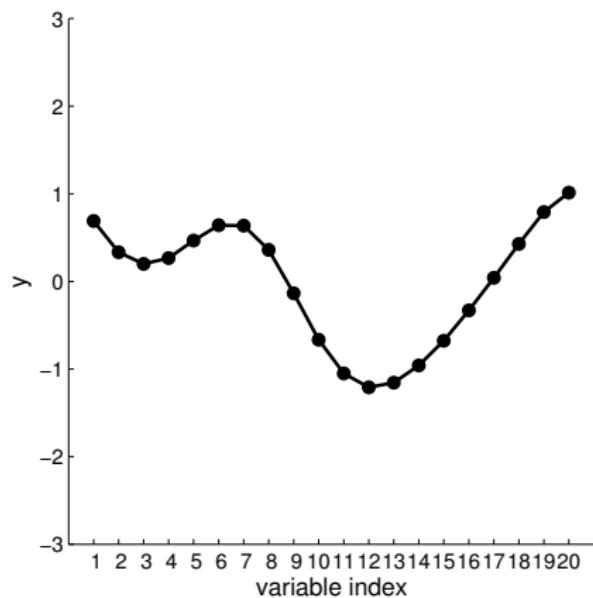
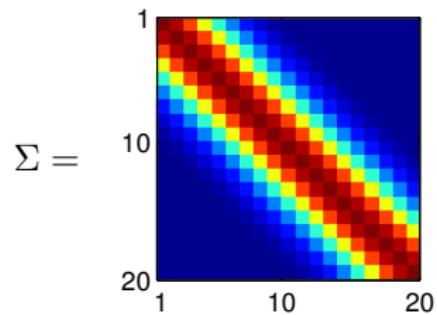
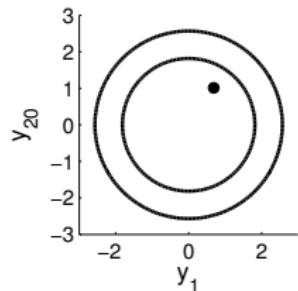
New visualisation



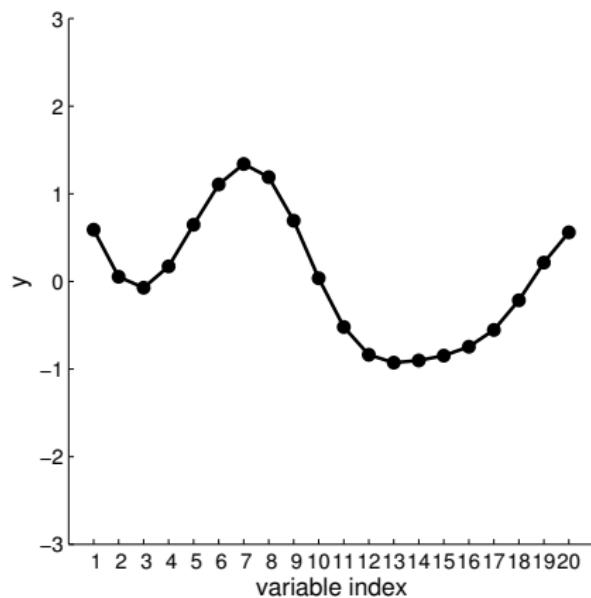
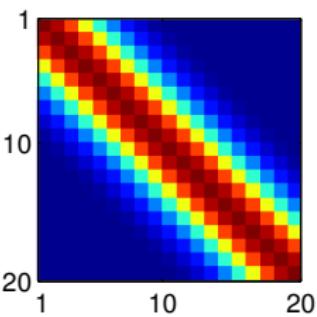
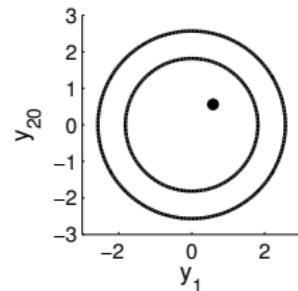
New visualisation



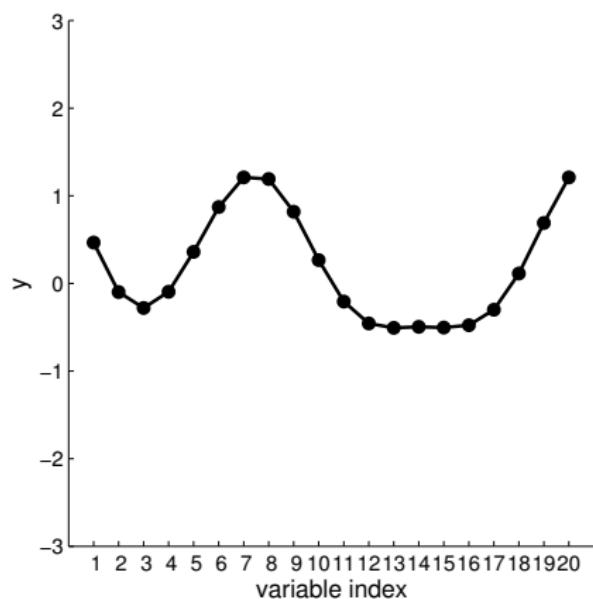
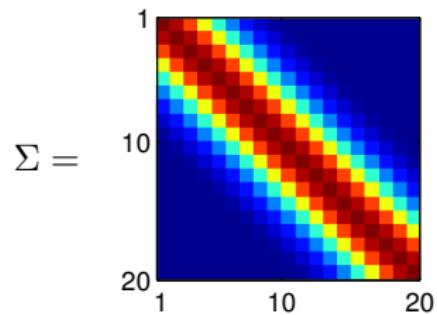
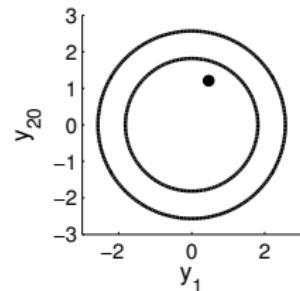
New visualisation



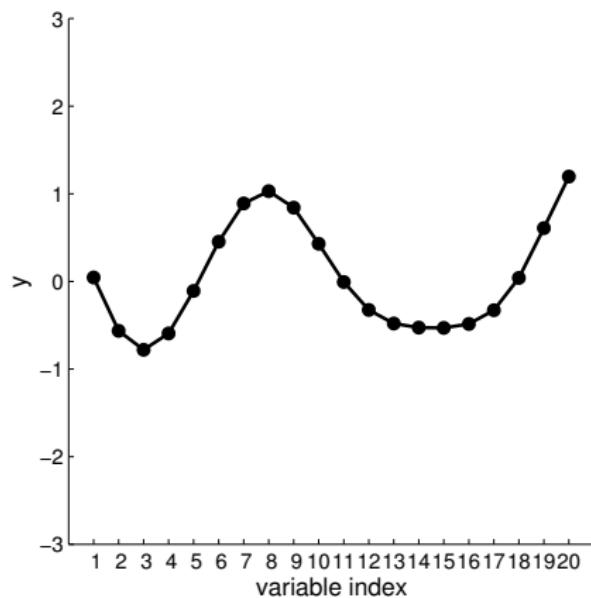
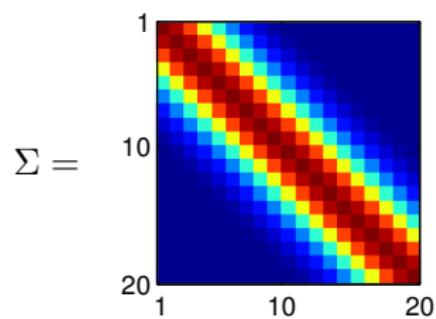
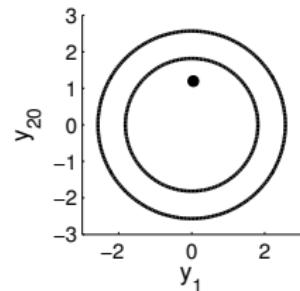
New visualisation



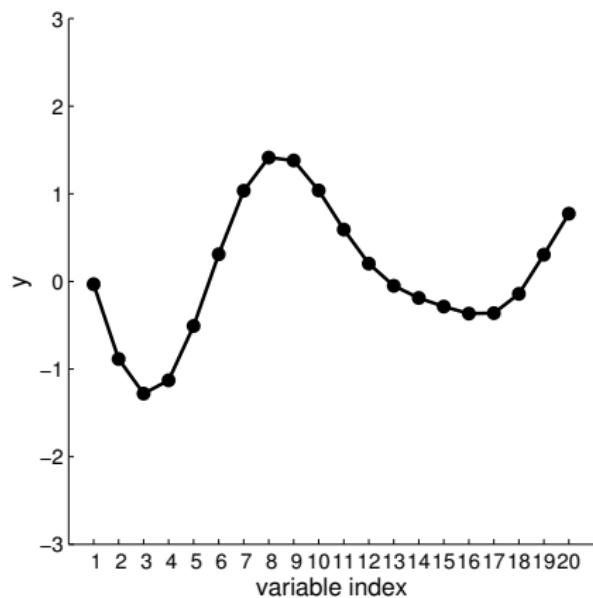
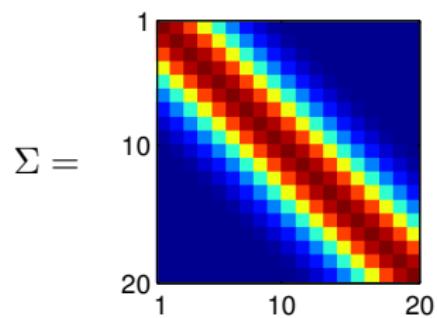
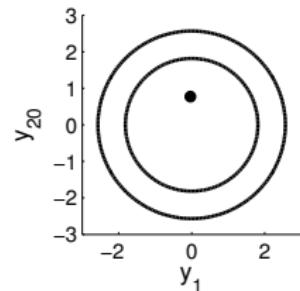
New visualisation



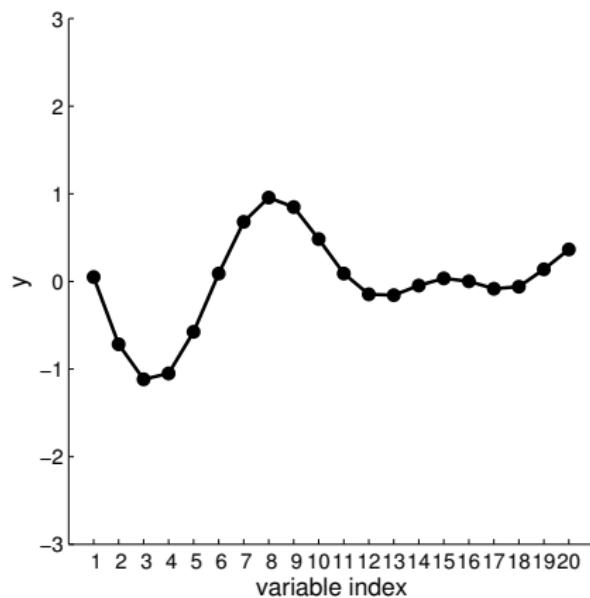
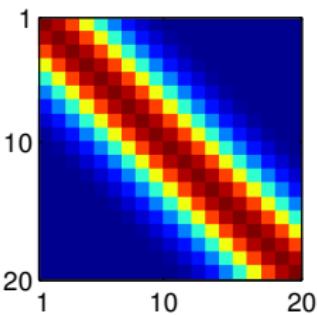
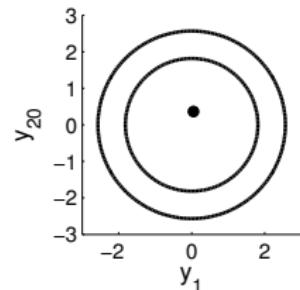
New visualisation



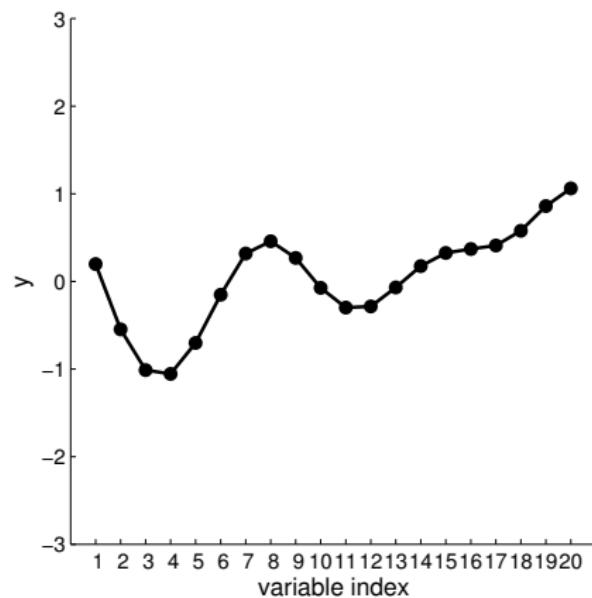
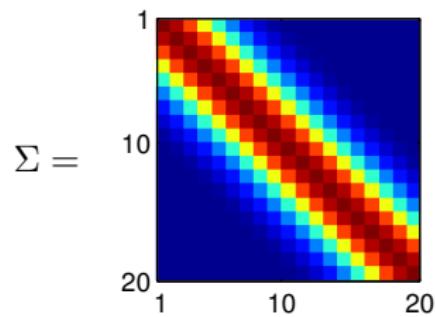
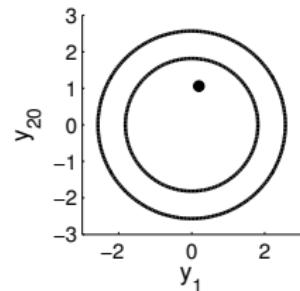
New visualisation



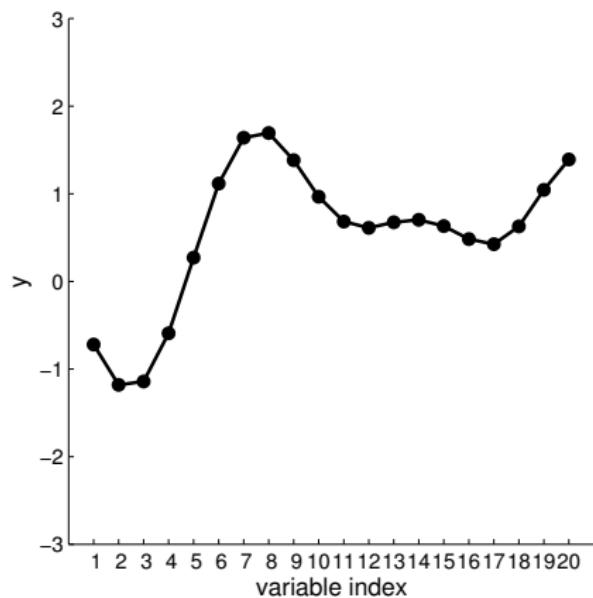
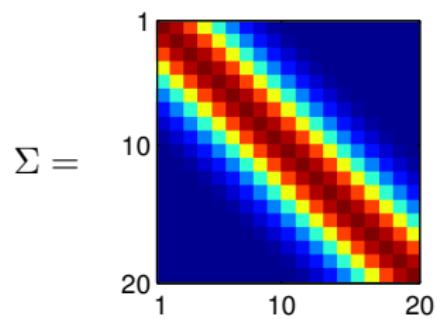
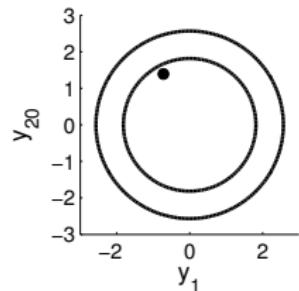
New visualisation



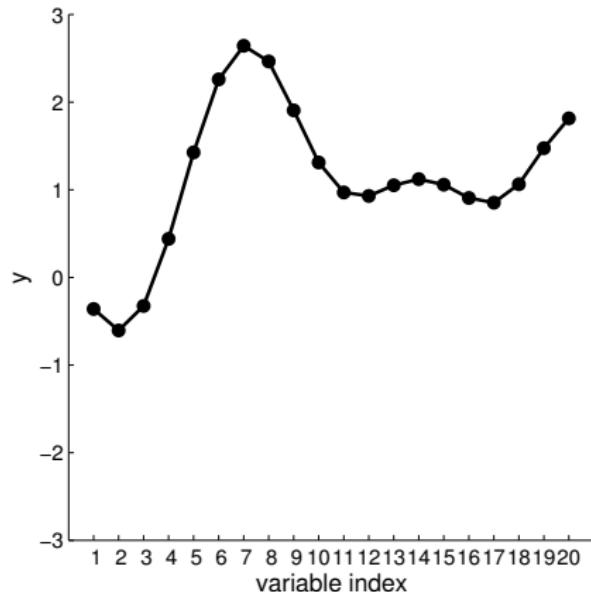
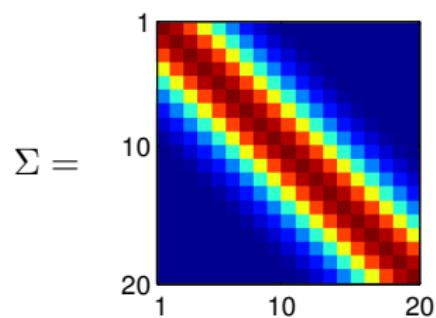
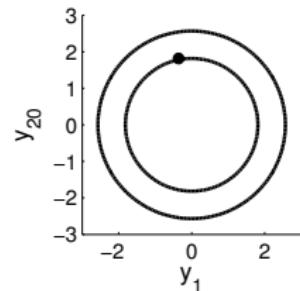
New visualisation



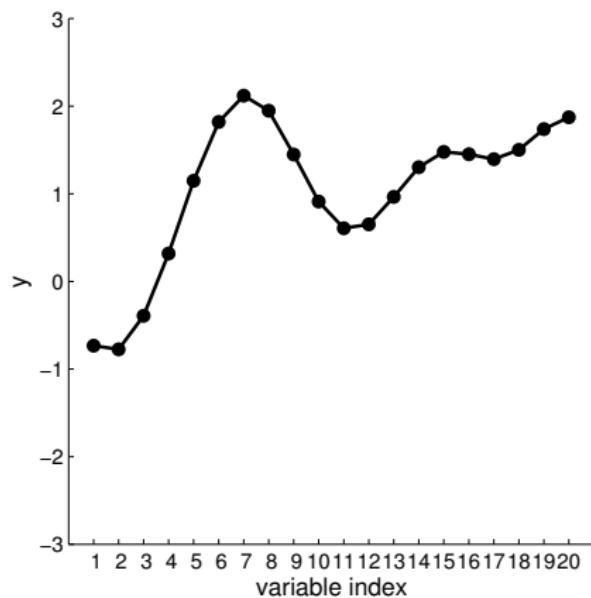
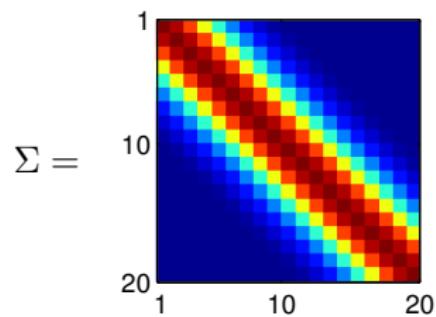
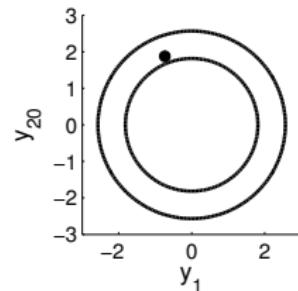
New visualisation



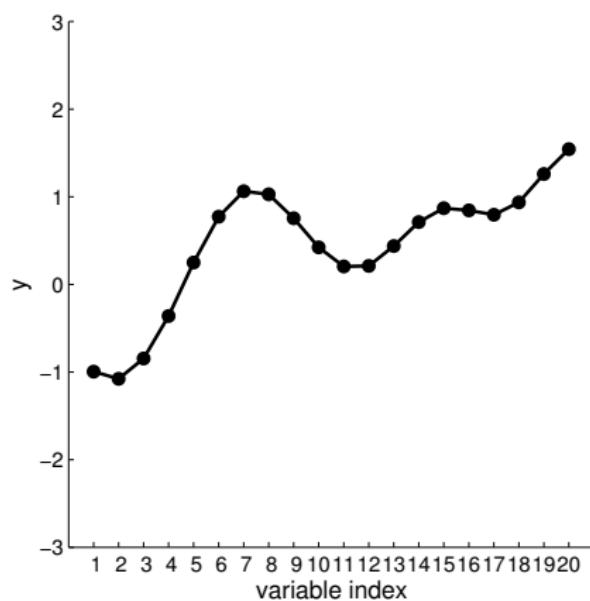
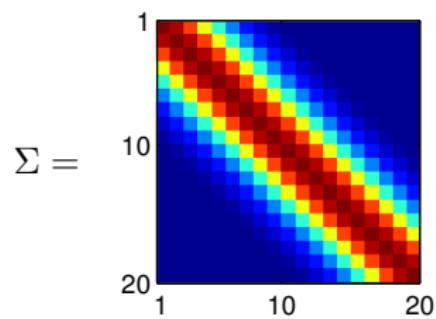
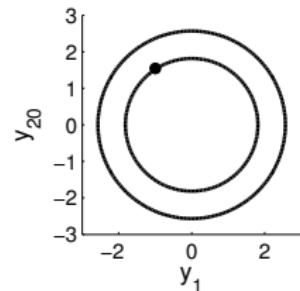
New visualisation



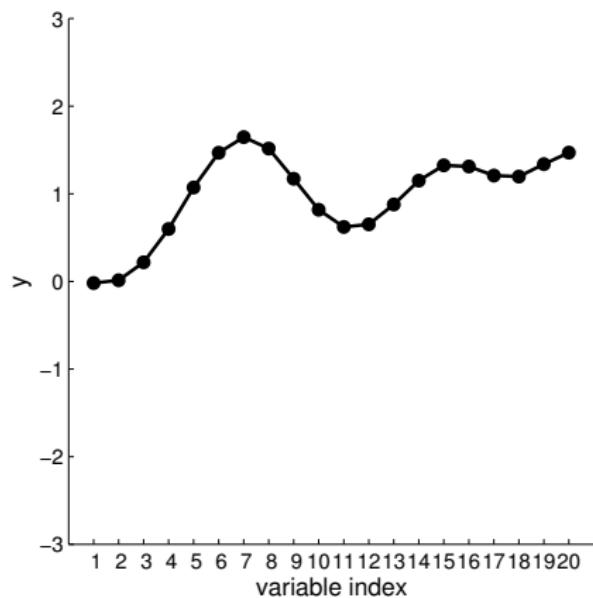
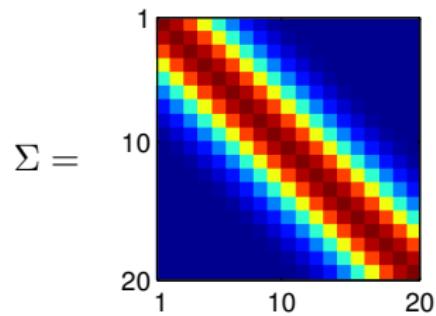
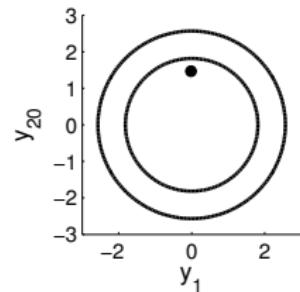
New visualisation



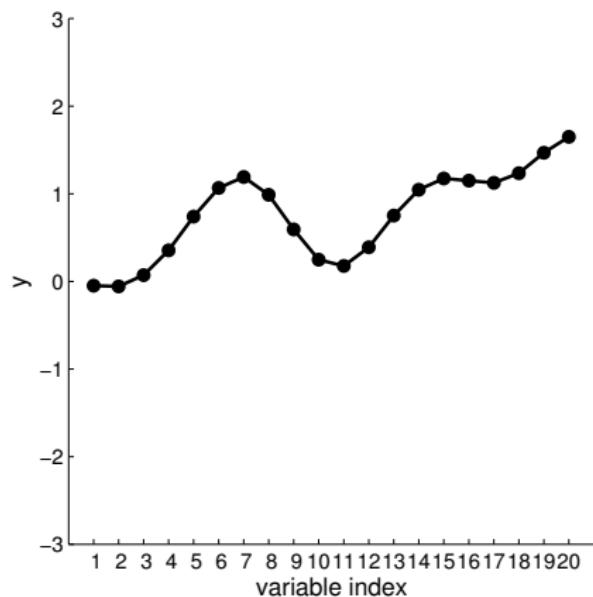
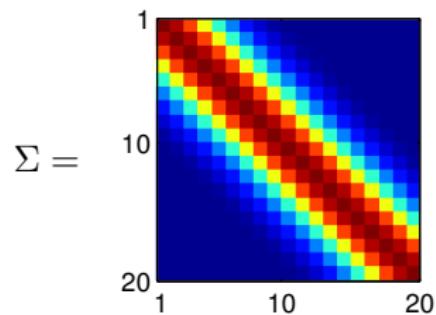
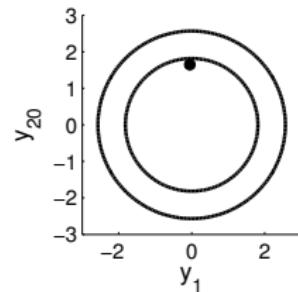
New visualisation



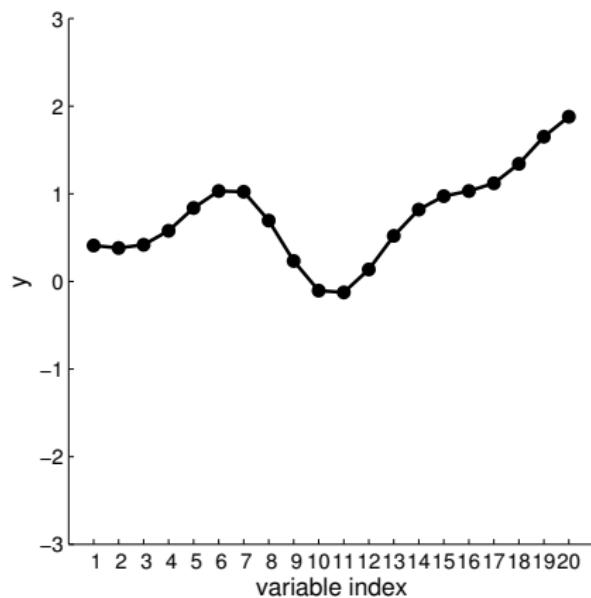
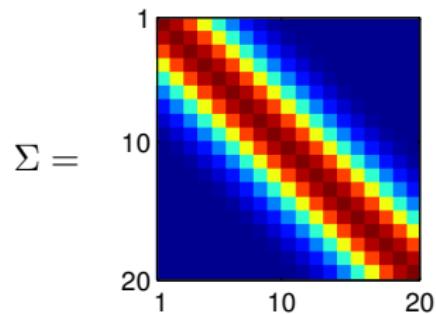
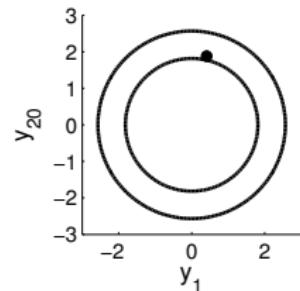
New visualisation



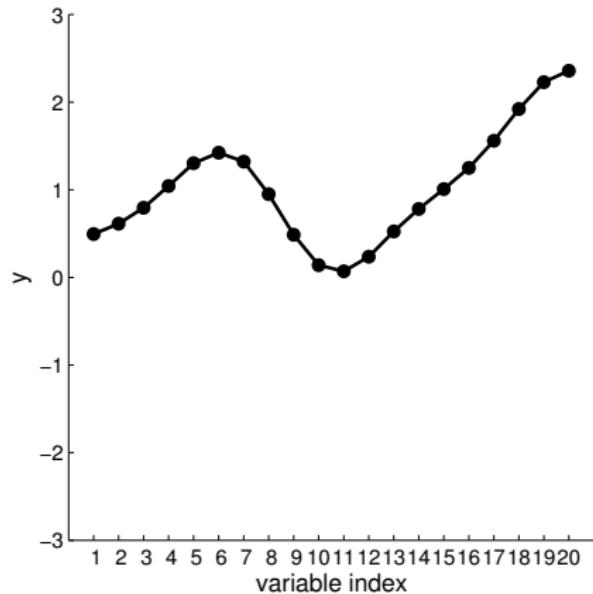
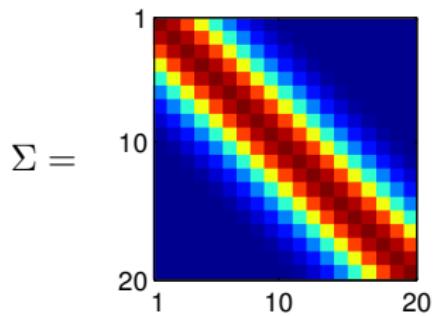
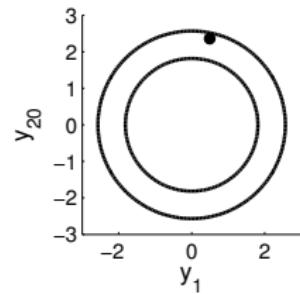
New visualisation



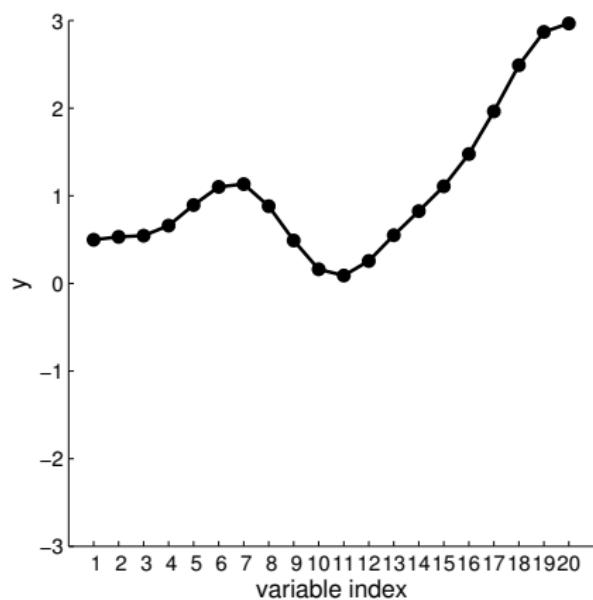
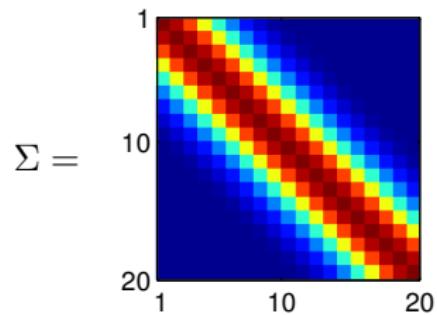
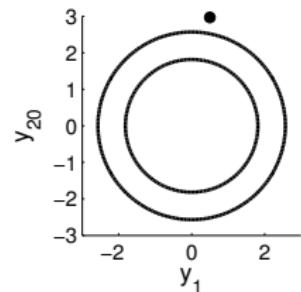
New visualisation



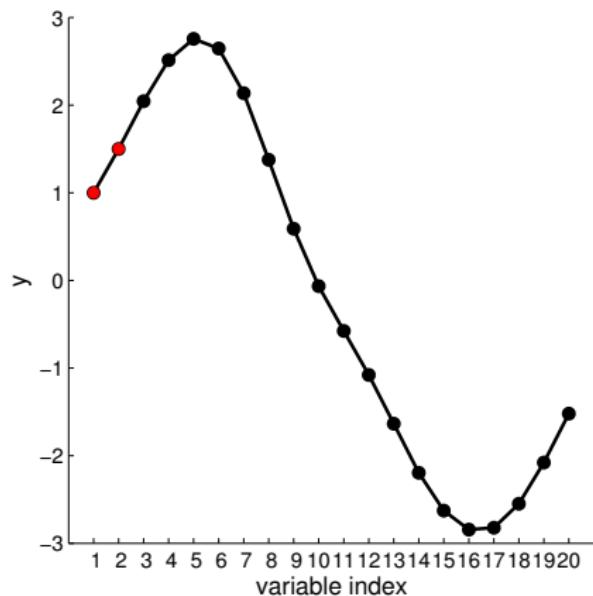
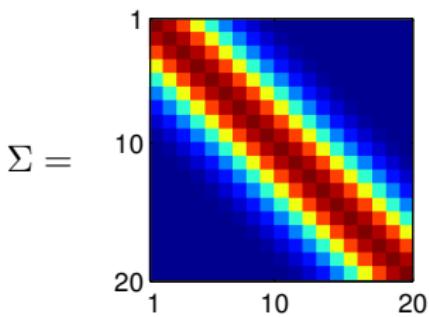
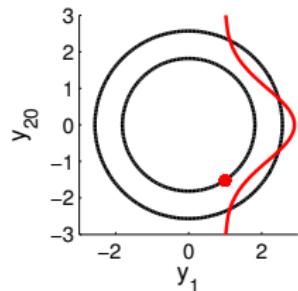
New visualisation



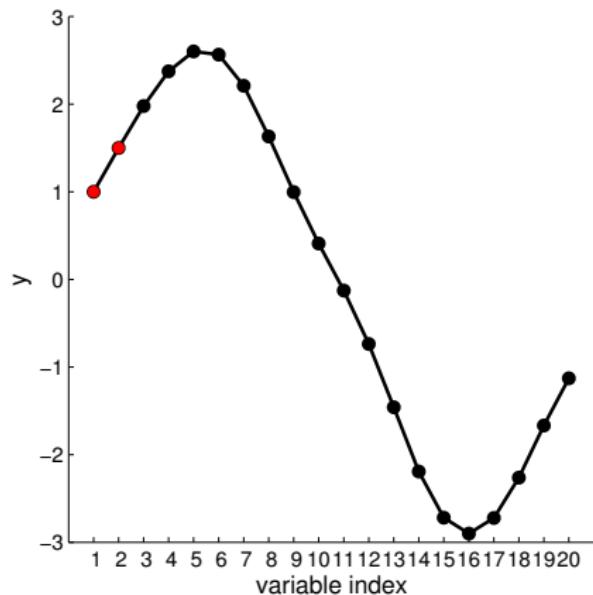
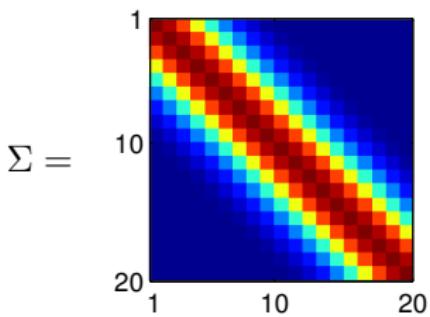
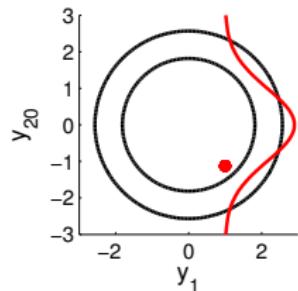
New visualisation



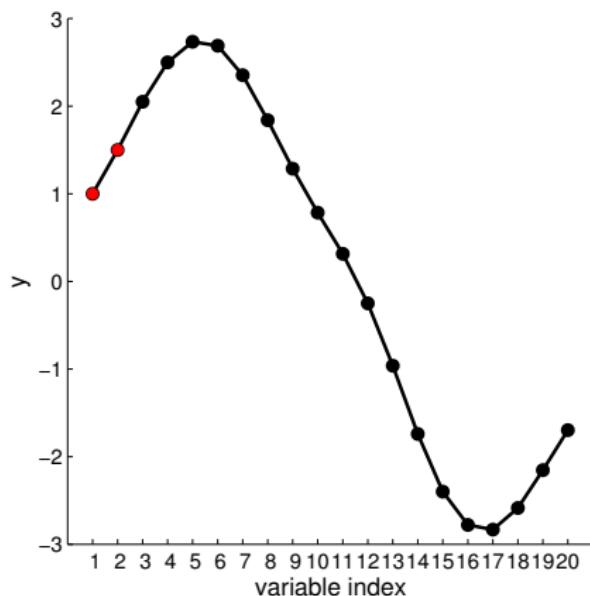
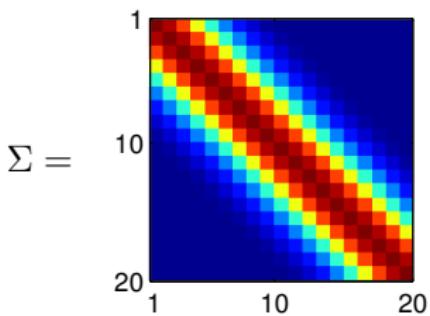
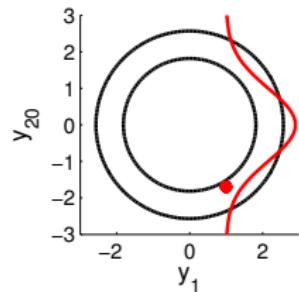
New visualisation



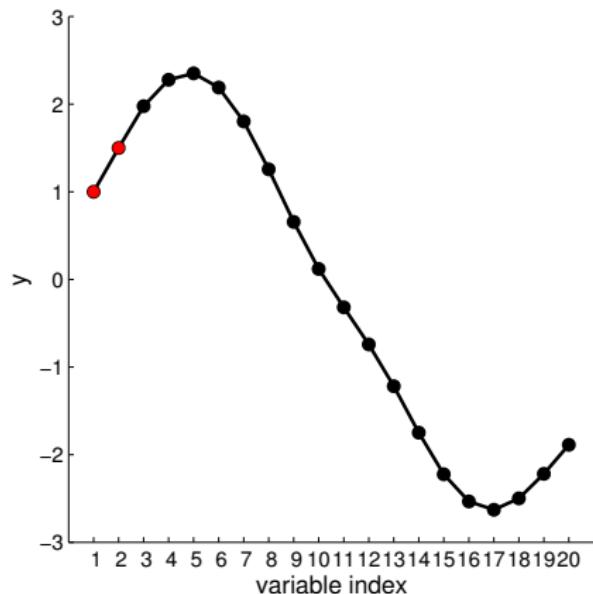
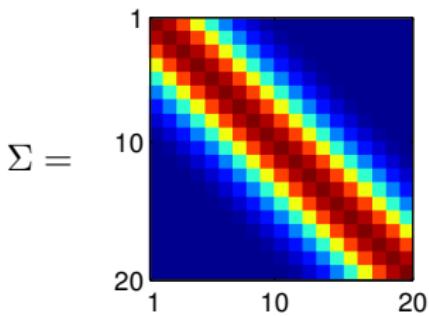
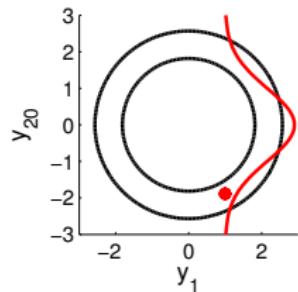
New visualisation



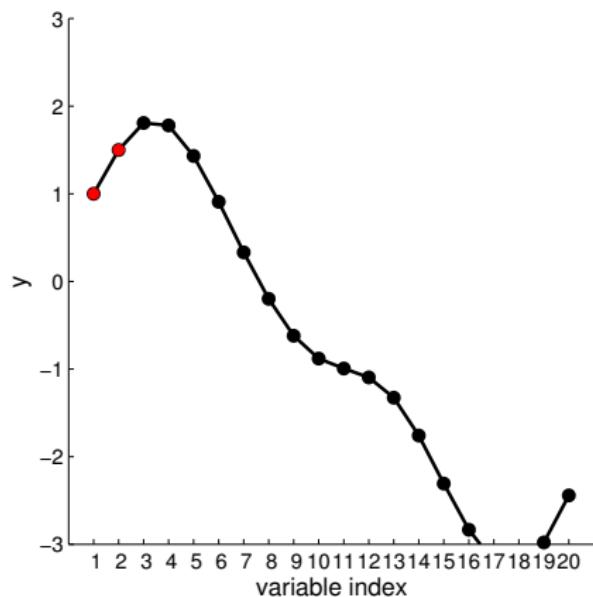
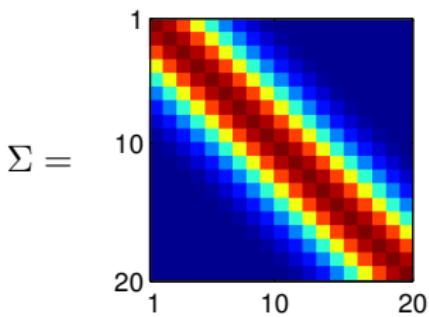
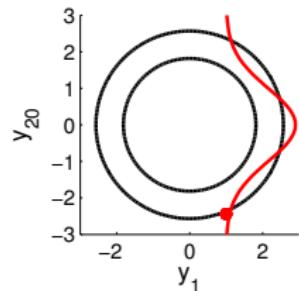
New visualisation



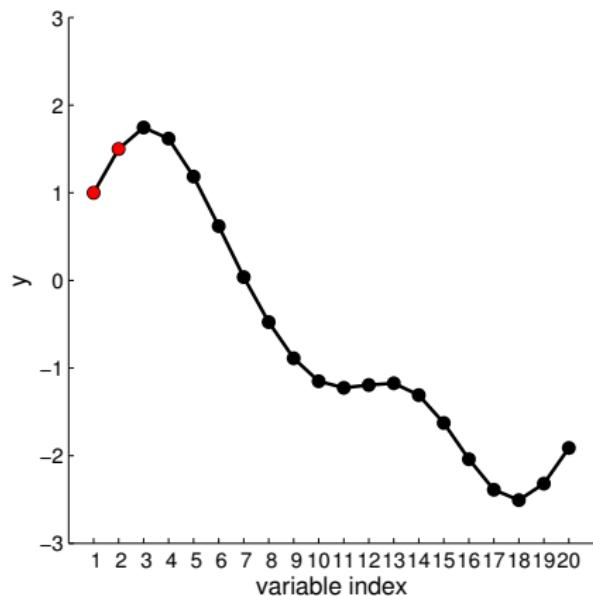
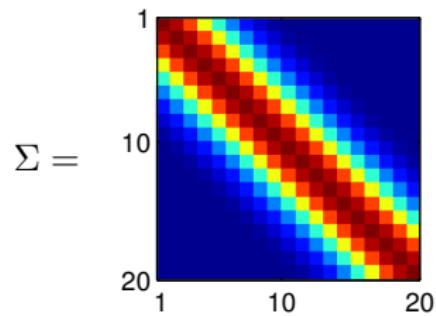
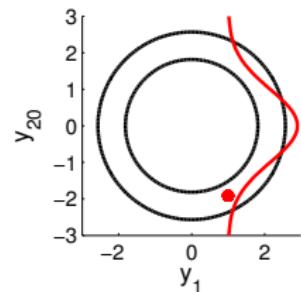
New visualisation



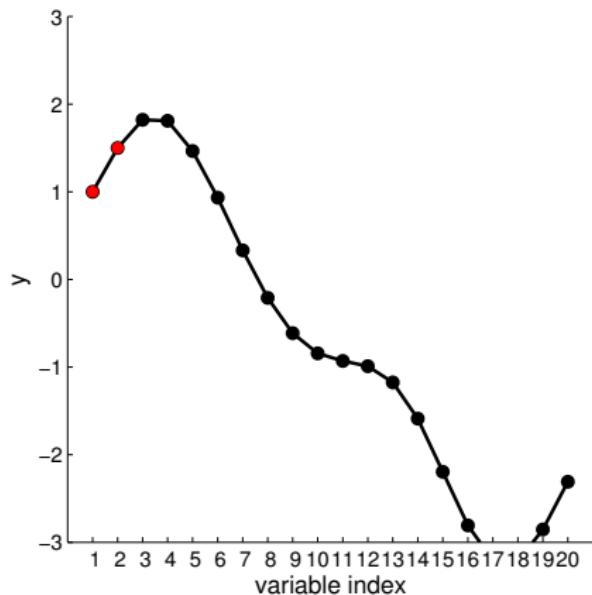
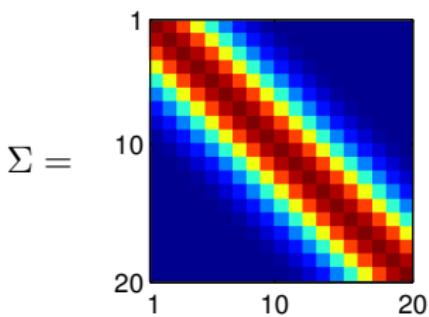
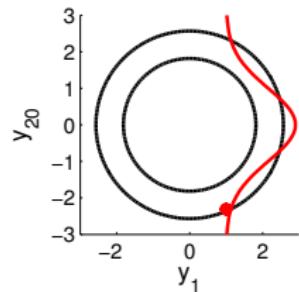
New visualisation



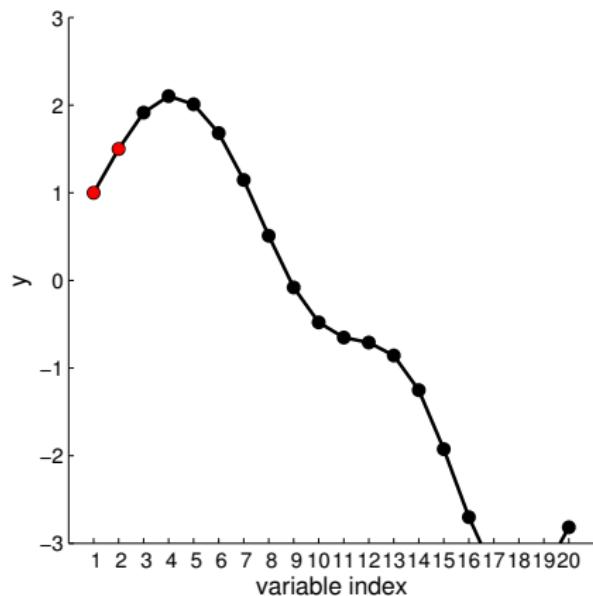
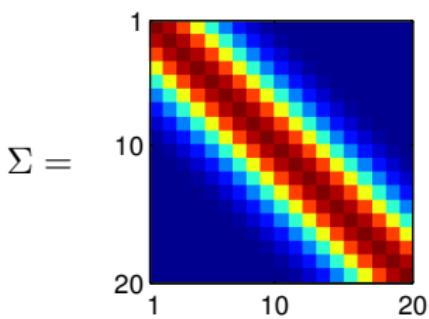
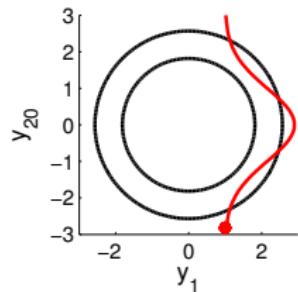
New visualisation



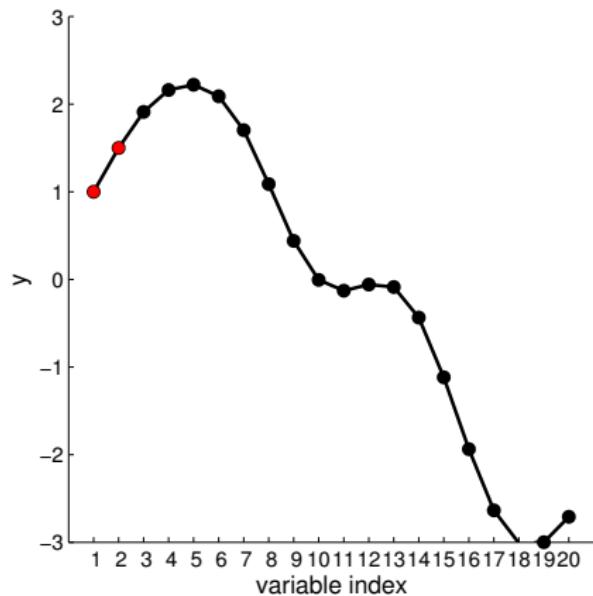
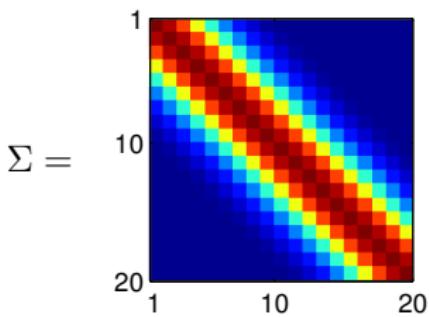
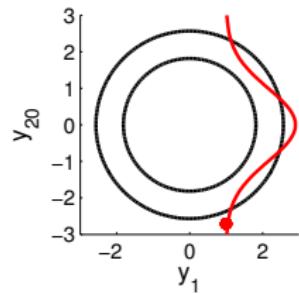
New visualisation



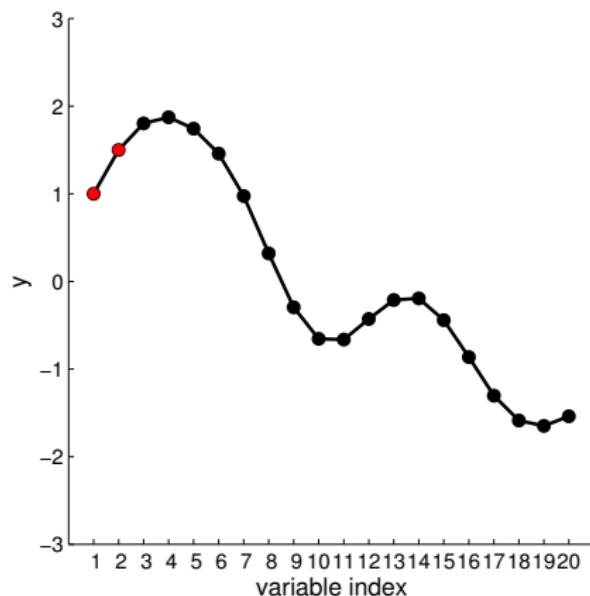
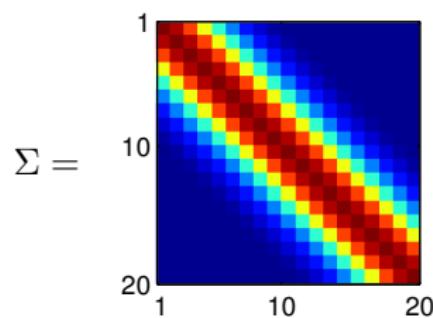
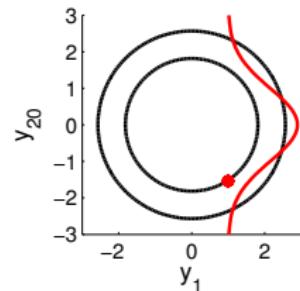
New visualisation



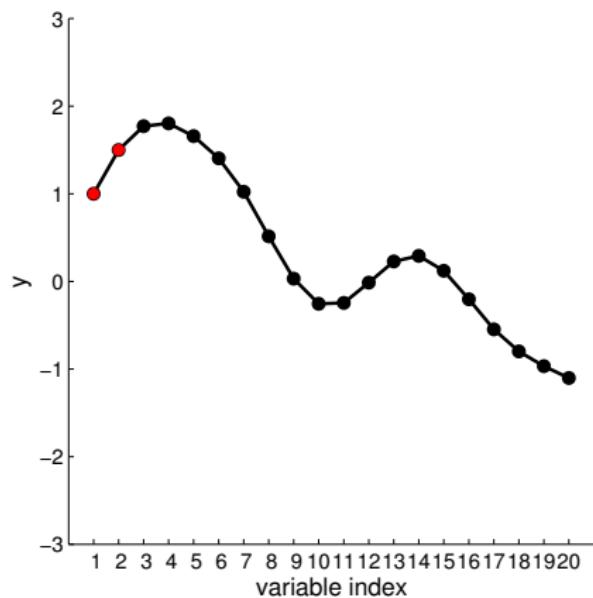
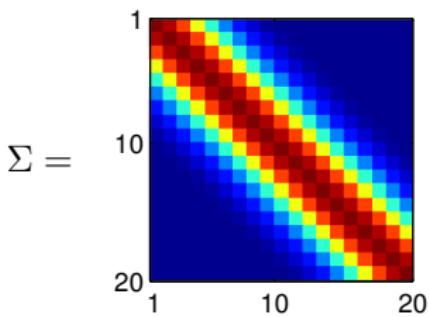
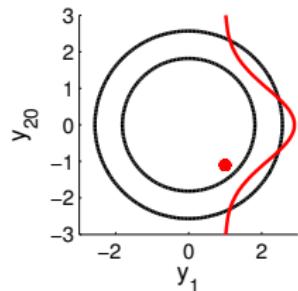
New visualisation



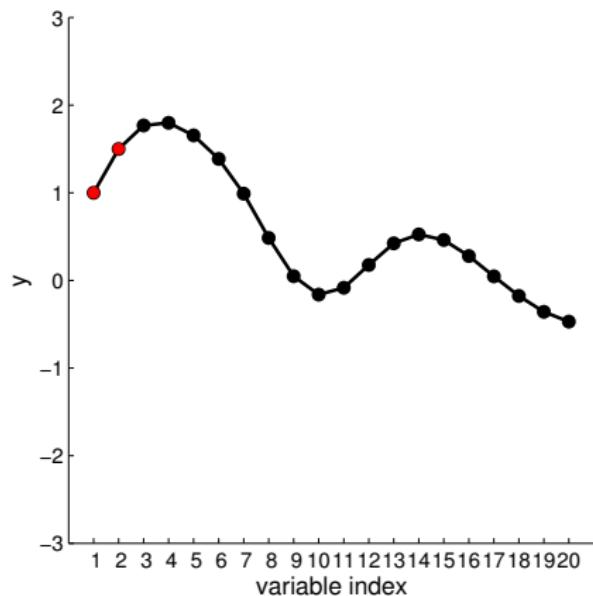
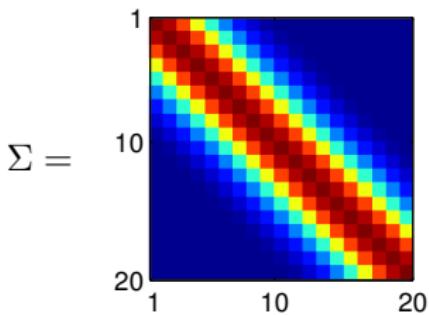
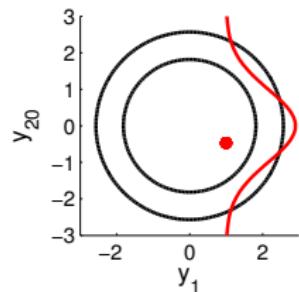
New visualisation



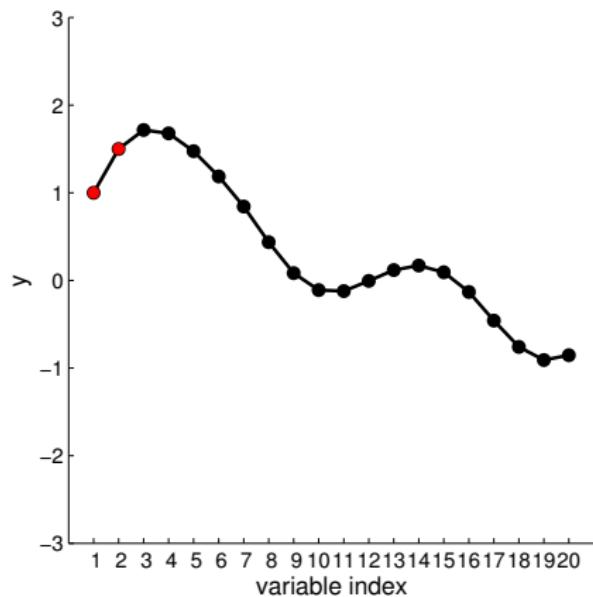
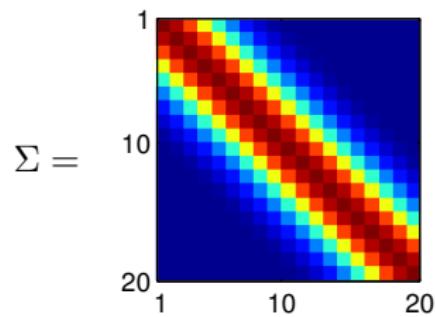
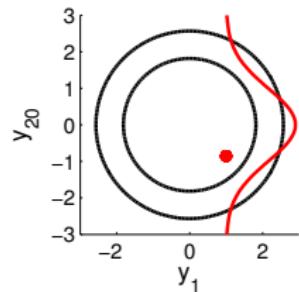
New visualisation



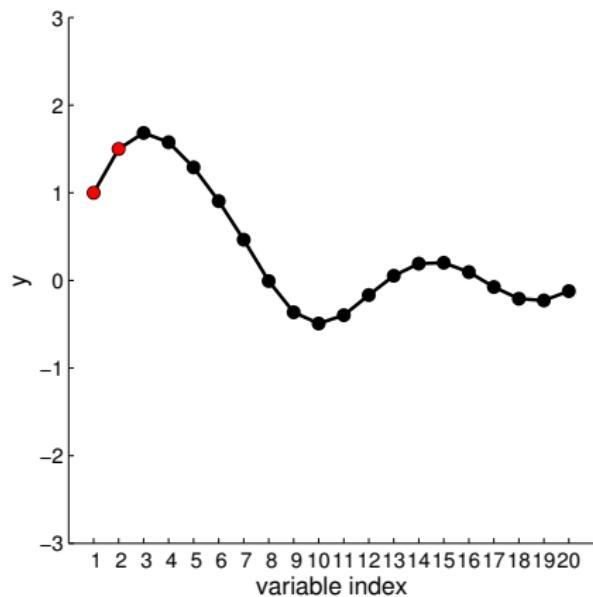
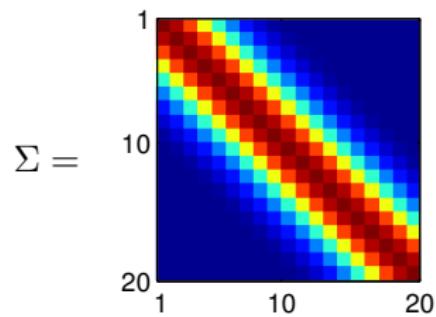
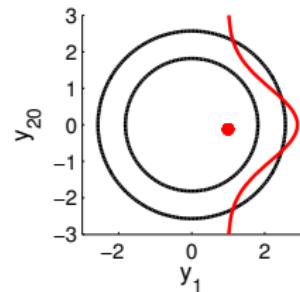
New visualisation



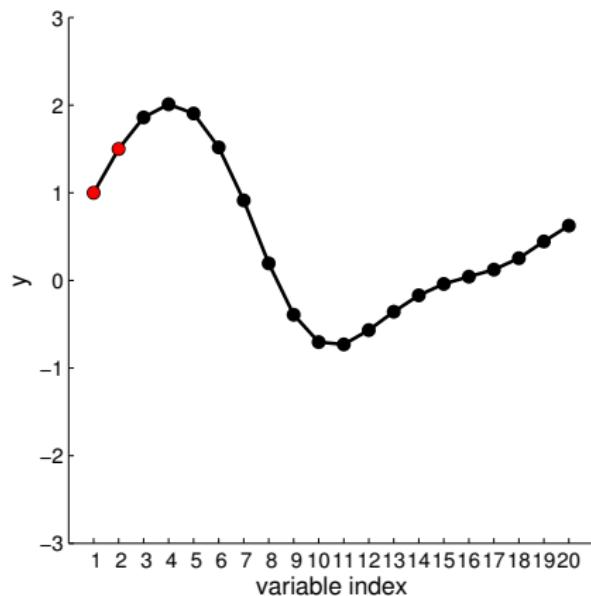
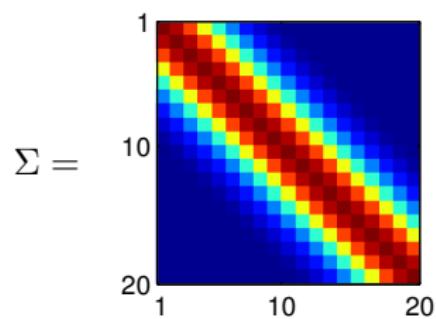
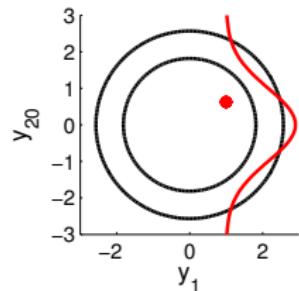
New visualisation



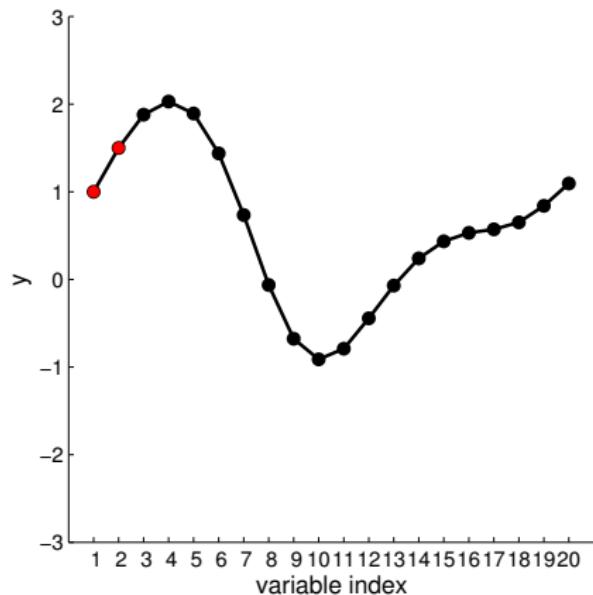
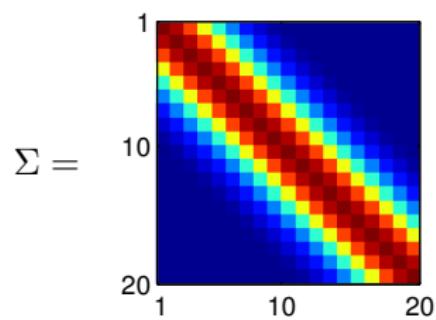
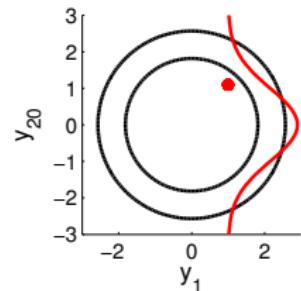
New visualisation



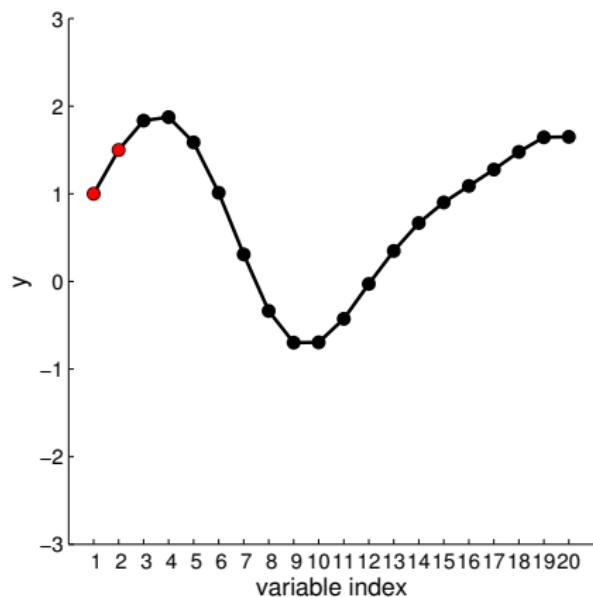
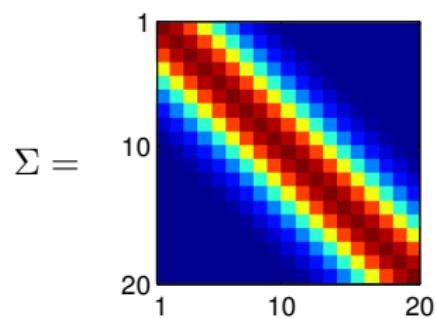
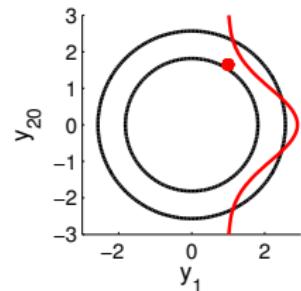
New visualisation



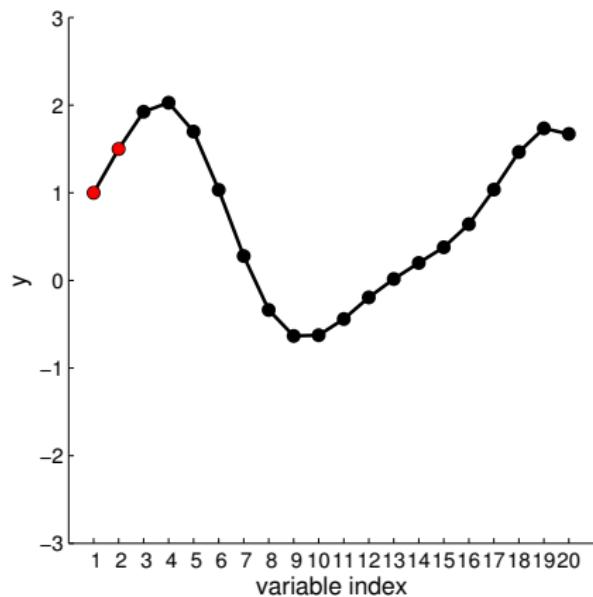
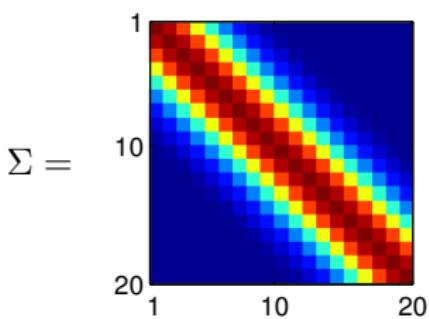
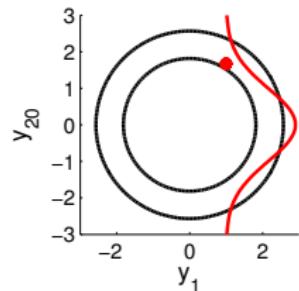
New visualisation



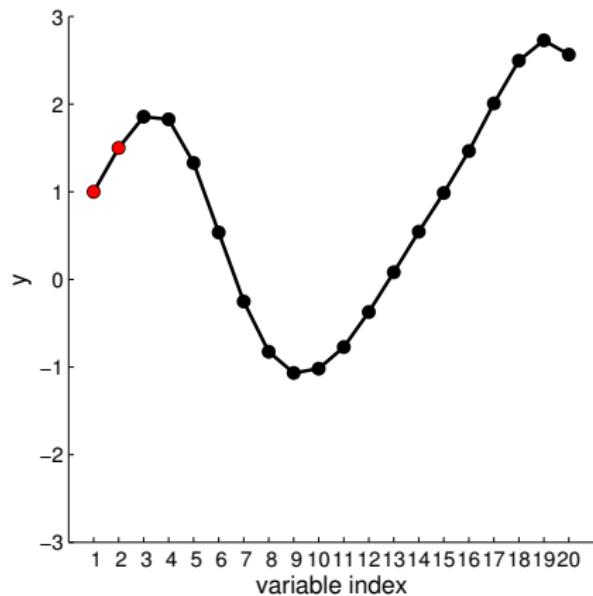
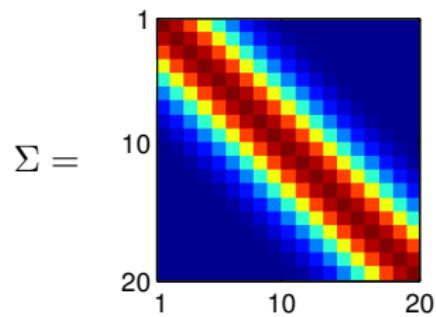
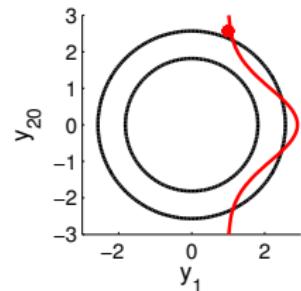
New visualisation



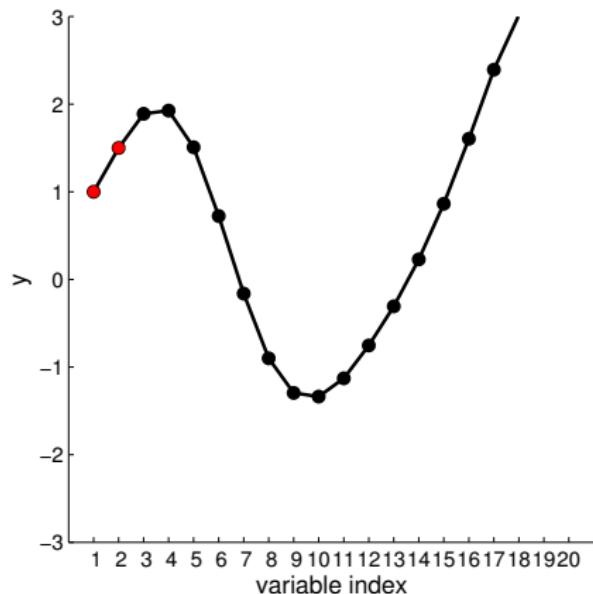
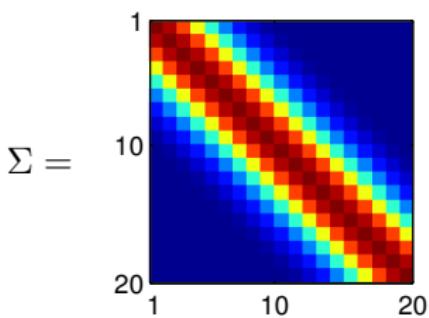
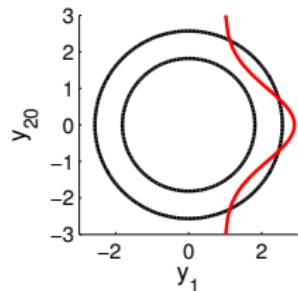
New visualisation



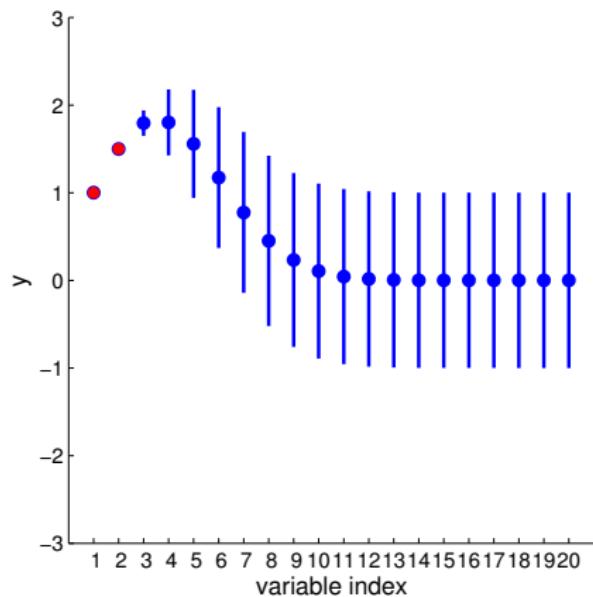
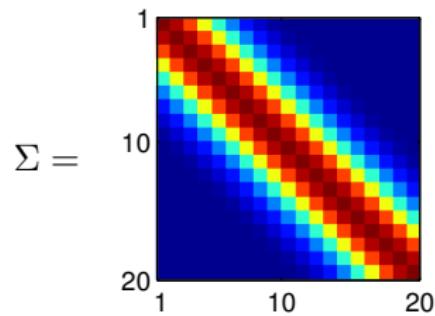
New visualisation



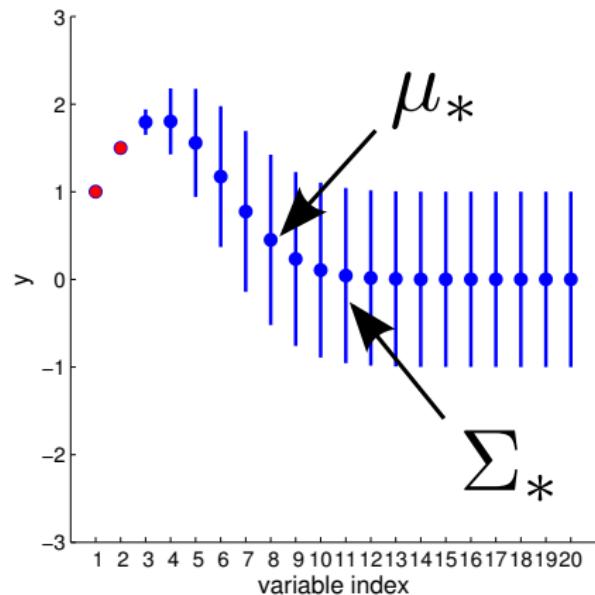
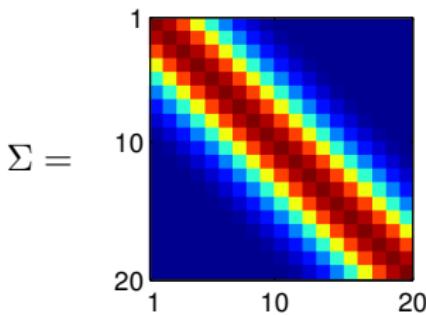
New visualisation



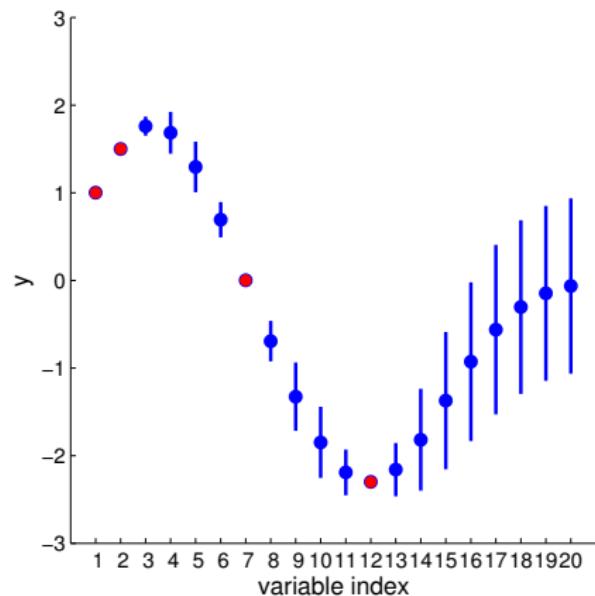
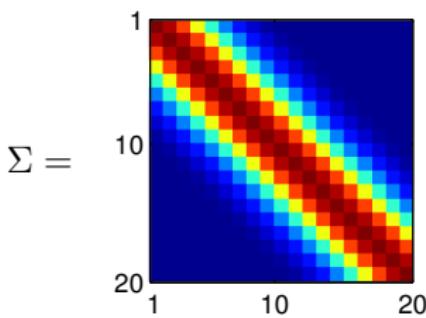
Regression using Gaussians



Regression using Gaussians



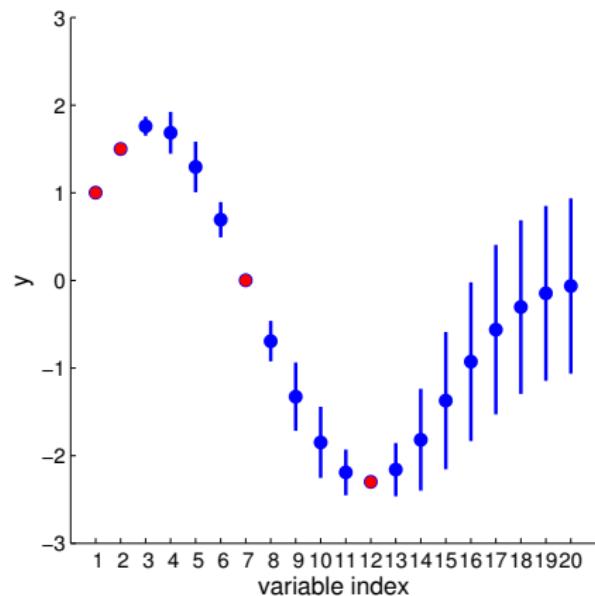
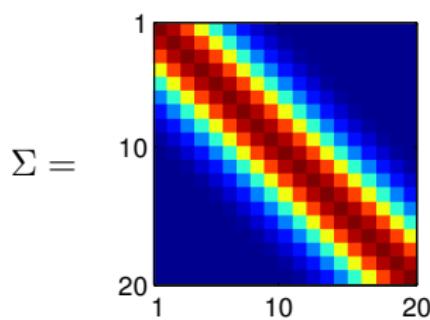
Regression using Gaussians



Regression using Gaussians

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

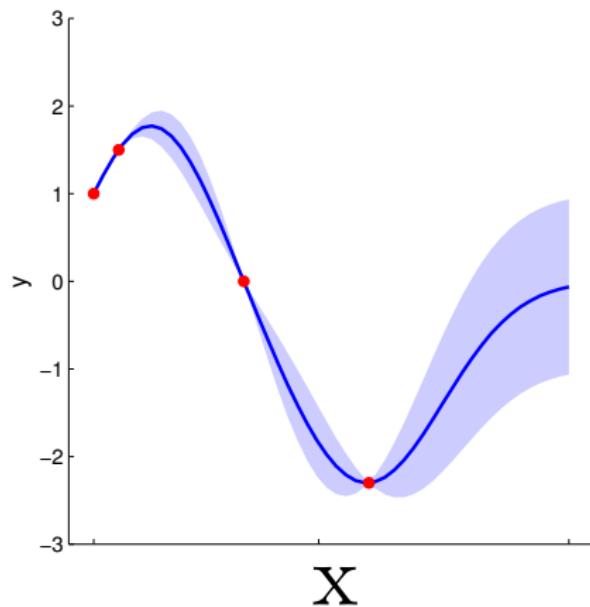
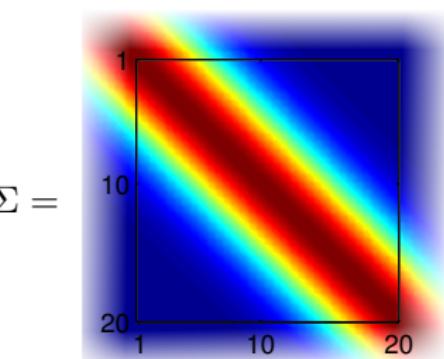
$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



Regression: probabilistic inference in function space

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



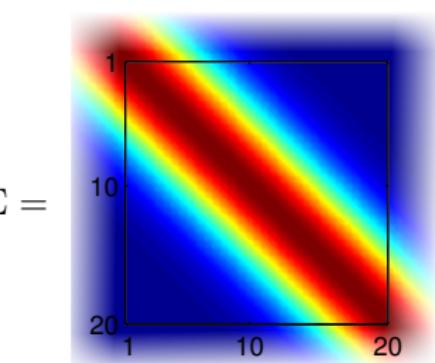
Regression: probabilistic inference in function space

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

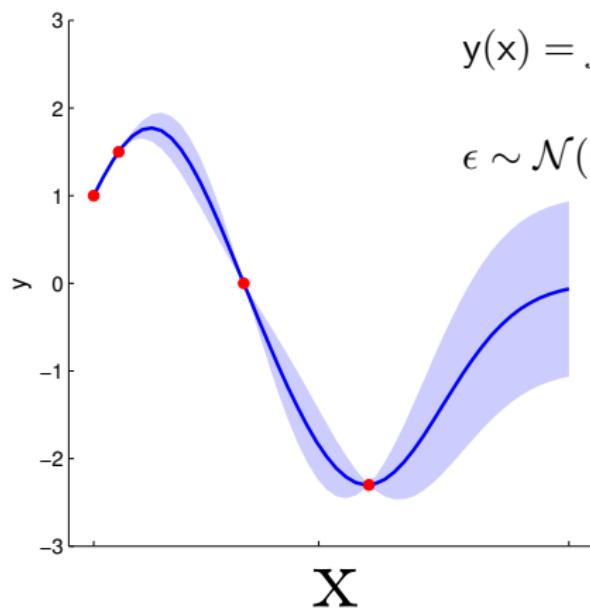
$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



$$\Sigma =$$

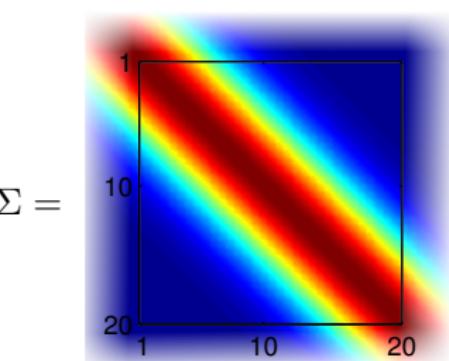
Regression: probabilistic inference in function space

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

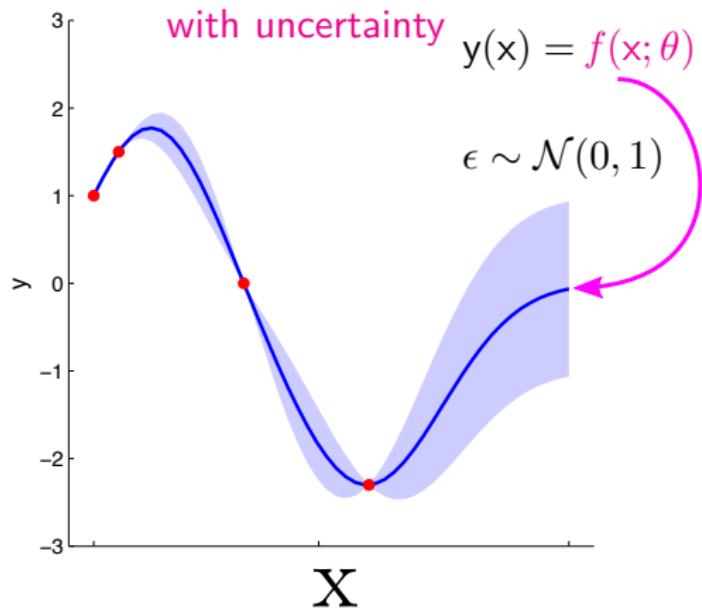


Parametric model

function estimate
with uncertainty

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



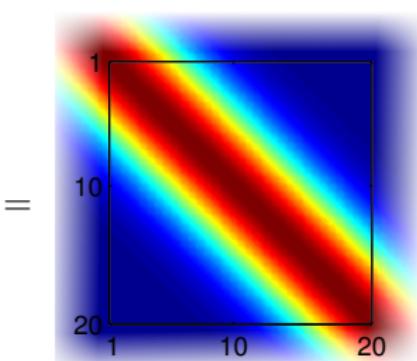
Regression: probabilistic inference in function space

Non-parametric (∞ -parametric)

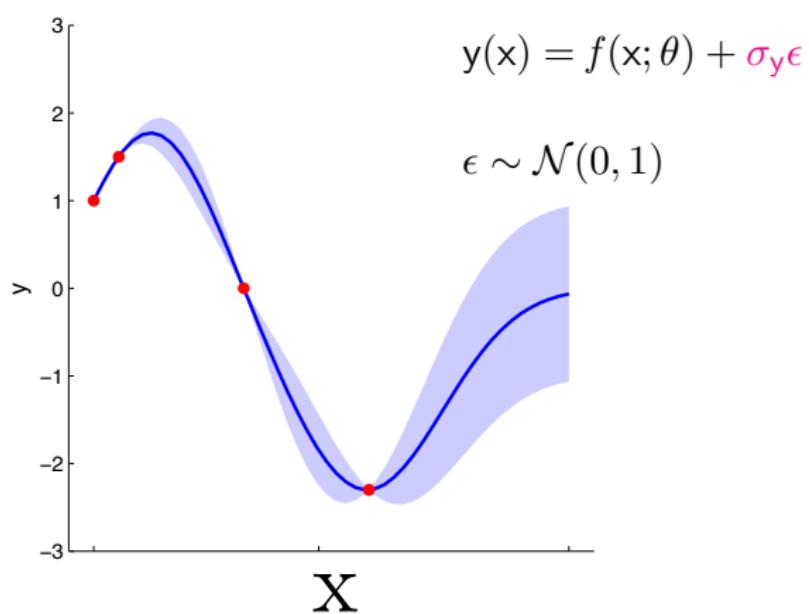
$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

noise
↓
 $\Sigma(x_1, x_2) = K(x_1, x_2) + \sigma_y^2$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



Parametric model



Regression: probabilistic inference in function space

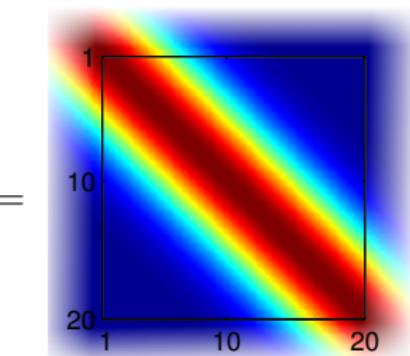
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

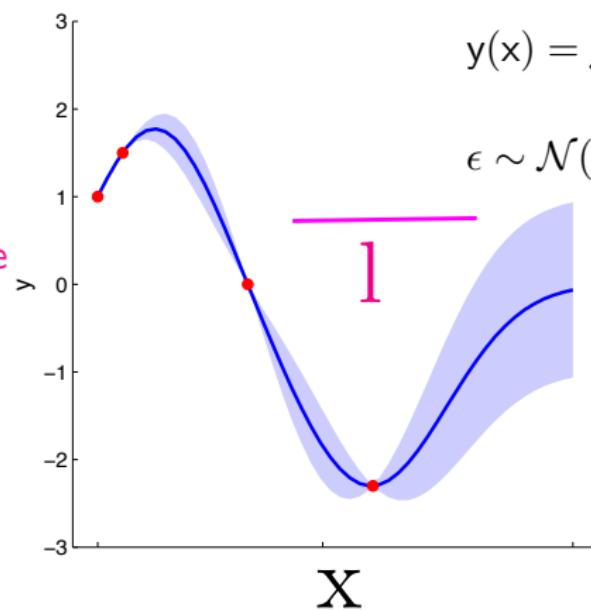
↑
horizontal-scale



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



Regression: probabilistic inference in function space

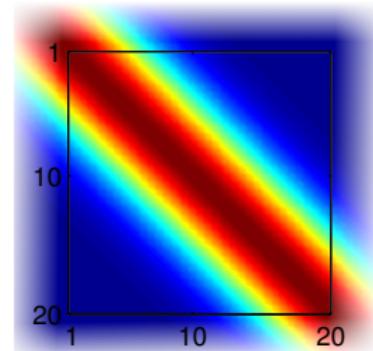
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

vertical-scale horizontal-scale

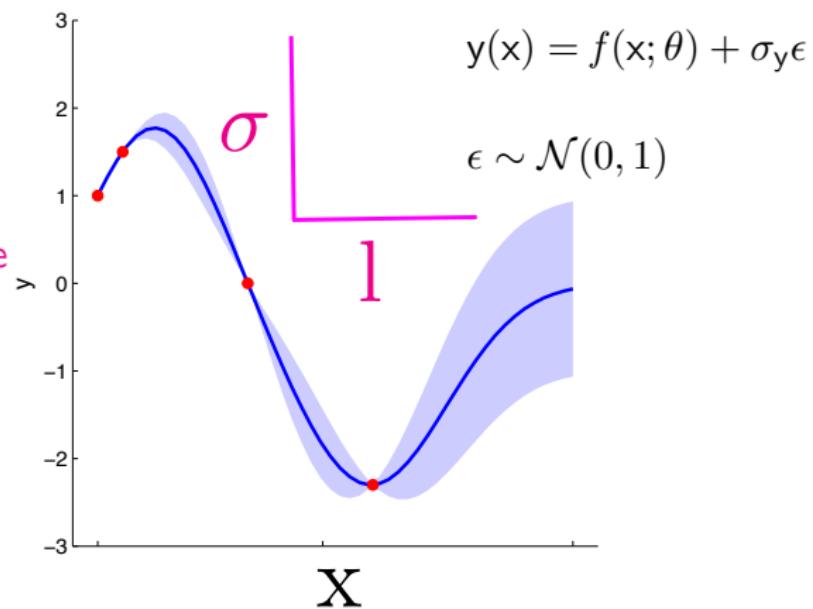


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



Mathematical Foundations: Definition

Gaussian process = generalisation of multivariate Gaussian distribution to infinitely many variables.

Definition: a Gaussian process is a collection of random variables, any finite number of which have (consistent) Gaussian distributions.

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

$$\mathbf{f} = (f_1, \dots, f_n) \sim \mathcal{N}(\mu, \Sigma), \quad \text{indices } i = 1, \dots, n$$

A Gaussian process is fully specified by a mean function $m(\mathbf{x})$ and covariance function $K(\mathbf{x}, \mathbf{x}')$:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')), \quad \text{indices } \mathbf{x}$$

Mathematical Foundations: Regression

Q1. What's the formal justification for how we were using GPs for regression?

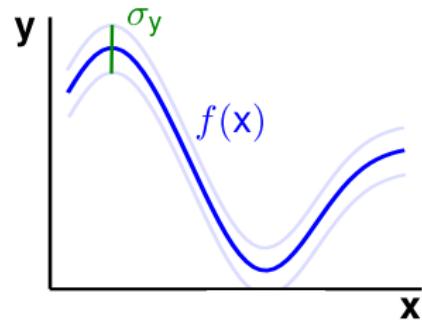
Mathematical Foundations: Regression

Q1. What's the formal justification for how we were using GPs for regression?

generative model (like non-linear regression)

$$y(x) = f(x) + \epsilon\sigma_y$$

$$p(\epsilon) = \mathcal{N}(0, 1)$$



Mathematical Foundations: Regression

Q1. What's the formal justification for how we were using GPs for regression?

generative model (like non-linear regression)

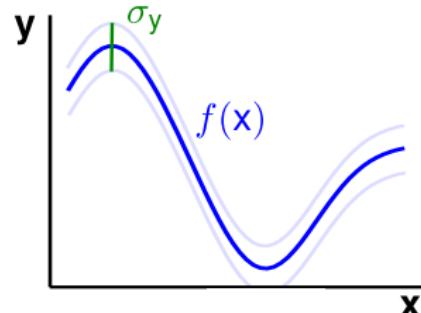
$$y(x) = f(x) + \epsilon \sigma_y$$

$$p(\epsilon) = \mathcal{N}(0, 1)$$

place GP prior over the non-linear function

$$p(f(x)|\theta) = \mathcal{GP}(0, K(x, x'))$$

$$K(x, x') = \sigma^2 \exp\left(-\frac{1}{2l^2}(x - x')^2\right) \quad (\text{smoothly wiggling functions expected})$$



Mathematical Foundations: Regression

Q1. What's the formal justification for how we were using GPs for regression?

generative model (like non-linear regression)

$$y(x) = f(x) + \epsilon \sigma_y$$

$$p(\epsilon) = \mathcal{N}(0, 1)$$

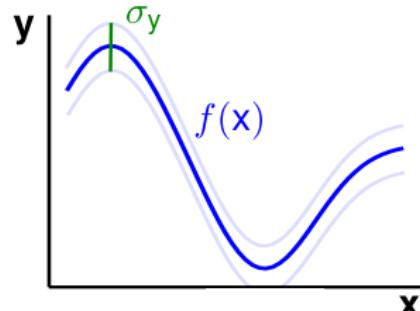
place GP prior over the non-linear function

$$p(f(x)|\theta) = \mathcal{GP}(0, K(x, x'))$$

$$K(x, x') = \sigma^2 \exp\left(-\frac{1}{2l^2}(x - x')^2\right) \quad (\text{smoothly wiggling functions expected})$$

sum of Gaussian variables = Gaussian: induces a GP over $y(x)$

$$p(y(x)|\theta) = \mathcal{GP}(0, K(x, x') + I\sigma_y^2)$$



Mathematical Foundations: Marginalisation

Q2. A GP is "like" a Gaussian distribution with an infinitely long mean vector and an "infinite by infinite" covariance matrix, so how do we represent it on a computer?

Mathematical Foundations: Marginalisation

Q2. A GP is "like" a Gaussian distribution with an infinitely long mean vector and an "infinite by infinite" covariance matrix, so how do we represent it on a computer?

We are saved by the marginalisation property:

$$p(\mathbf{y}_1) = \int p(\mathbf{y}_1, \mathbf{y}_2) d\mathbf{y}_2$$

$$p(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right)$$

Mathematical Foundations: Marginalisation

Q2. A GP is "like" a Gaussian distribution with an infinitely long mean vector and an "infinite by infinite" covariance matrix, so how do we represent it on a computer?

We are saved by the marginalisation property:

$$p(\mathbf{y}_1) = \int p(\mathbf{y}_1, \mathbf{y}_2) d\mathbf{y}_2$$

$$p(\mathbf{y}_1 | \mathbf{y}_2) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{bmatrix} \right) \implies p(\mathbf{y}_1) = \mathcal{N}(\mathbf{a}, \mathbf{A})$$

Mathematical Foundations: Marginalisation

Q2. A GP is "like" a Gaussian distribution with an infinitely long mean vector and an "infinite by infinite" covariance matrix, so how do we represent it on a computer?

We are saved by the marginalisation property:

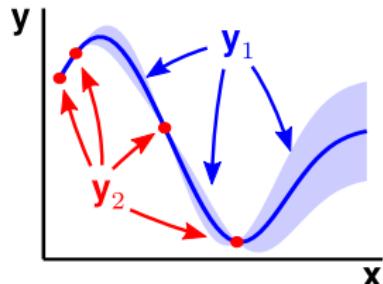
$$p(\mathbf{y}_1) = \int p(\mathbf{y}_1, \mathbf{y}_2) d\mathbf{y}_2$$

$$p(\mathbf{y}_1 | \mathbf{y}_2) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right) \implies p(\mathbf{y}_1) = \mathcal{N}(\mathbf{a}, \mathbf{A})$$

⇒ Only need to represent finite dimensional projections of GPs on computer.

Mathematical Foundations: Prediction

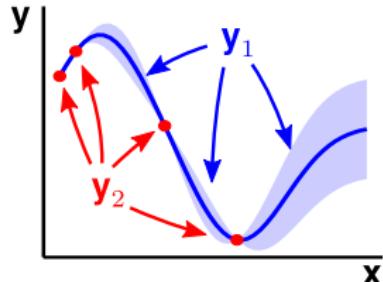
Q3. How do we make predictions?



Mathematical Foundations: Prediction

Q3. How do we make predictions?

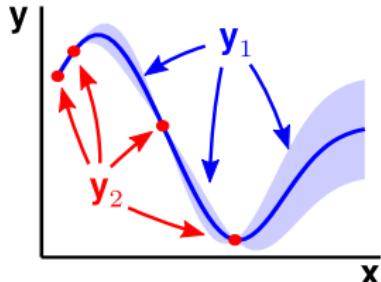
$$p(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right)$$



Mathematical Foundations: Prediction

Q3. How do we make predictions?

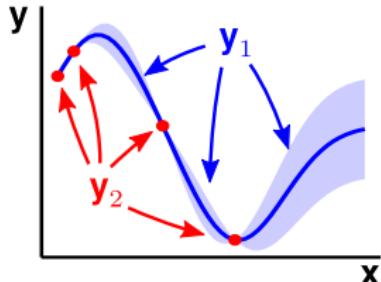
$$p(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right)$$
$$p(\mathbf{y}_1 | \mathbf{y}_2) = \frac{p(\mathbf{y}_1, \mathbf{y}_2)}{p(\mathbf{y}_2)}$$



Mathematical Foundations: Prediction

Q3. How do we make predictions?

$$p(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right)$$
$$p(\mathbf{y}_1 | \mathbf{y}_2) = \frac{p(\mathbf{y}_1, \mathbf{y}_2)}{p(\mathbf{y}_2)}$$



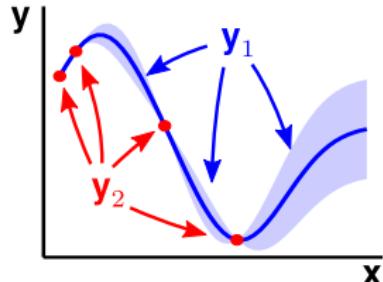
$$\implies p(\mathbf{y}_1 | \mathbf{y}_2) = \mathcal{N}(\mathbf{a} + \mathbf{B}\mathbf{C}^{-1}(\mathbf{y}_2 - \mathbf{b}), \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T)$$

Mathematical Foundations: Prediction

Q3. How do we make predictions?

$$p(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right)$$

$$p(\mathbf{y}_1 | \mathbf{y}_2) = \frac{p(\mathbf{y}_1, \mathbf{y}_2)}{p(\mathbf{y}_2)}$$



$$\implies p(\mathbf{y}_1 | \mathbf{y}_2) = \mathcal{N}(\mathbf{a} + \mathbf{B}\mathbf{C}^{-1}(\mathbf{y}_2 - \mathbf{b}), \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T)$$

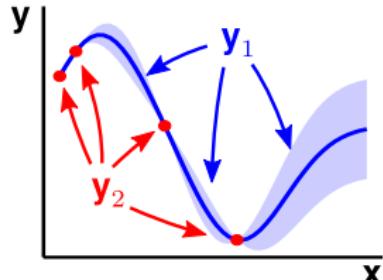
predictive mean

$$\mu_{\mathbf{y}_1 | \mathbf{y}_2} = \mathbf{a} + \mathbf{B}\mathbf{C}^{-1}(\mathbf{y}_2 - \mathbf{b})$$

Mathematical Foundations: Prediction

Q3. How do we make predictions?

$$p(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right)$$
$$p(\mathbf{y}_1 | \mathbf{y}_2) = \frac{p(\mathbf{y}_1, \mathbf{y}_2)}{p(\mathbf{y}_2)}$$



$$\implies p(\mathbf{y}_1 | \mathbf{y}_2) = \mathcal{N}(\mathbf{a} + \mathbf{B}\mathbf{C}^{-1}(\mathbf{y}_2 - \mathbf{b}), \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T)$$

predictive mean

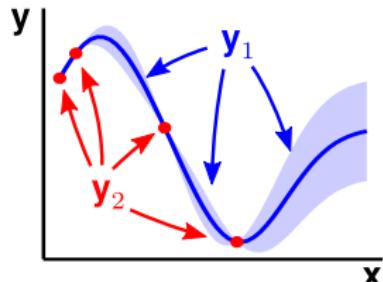
$$\begin{aligned} \mu_{\mathbf{y}_1 | \mathbf{y}_2} &= \mathbf{a} + \mathbf{B}\mathbf{C}^{-1}(\mathbf{y}_2 - \mathbf{b}) \\ &= \mathbf{B}\mathbf{C}^{-1}\mathbf{y}_2 \\ &= \mathbf{W}\mathbf{y}_2 \end{aligned}$$

linear in the data

Mathematical Foundations: Prediction

Q3. How do we make predictions?

$$p(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{N} \left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \right)$$
$$p(\mathbf{y}_1 | \mathbf{y}_2) = \frac{p(\mathbf{y}_1, \mathbf{y}_2)}{p(\mathbf{y}_2)}$$



$$\implies p(\mathbf{y}_1 | \mathbf{y}_2) = \mathcal{N}(\mathbf{a} + \mathbf{B}\mathbf{C}^{-1}(\mathbf{y}_2 - \mathbf{b}), \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T)$$

predictive mean

$$\begin{aligned}\mu_{\mathbf{y}_1 | \mathbf{y}_2} &= \mathbf{a} + \mathbf{B}\mathbf{C}^{-1}(\mathbf{y}_2 - \mathbf{b}) \\ &= \mathbf{B}\mathbf{C}^{-1}\mathbf{y}_2 \\ &= \mathbf{W}\mathbf{y}_2\end{aligned}$$

linear in the data

predictive covariance

$$\Sigma_{\mathbf{y}_1 | \mathbf{y}_2} = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T$$

predictive uncertainty = prior uncertainty - reduction in uncertainty

predictions more confident than prior

Regression: probabilistic inference in function space

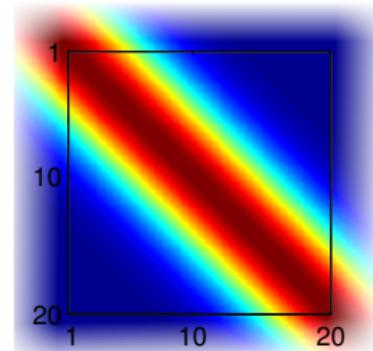
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

vertical-scale horizontal-scale

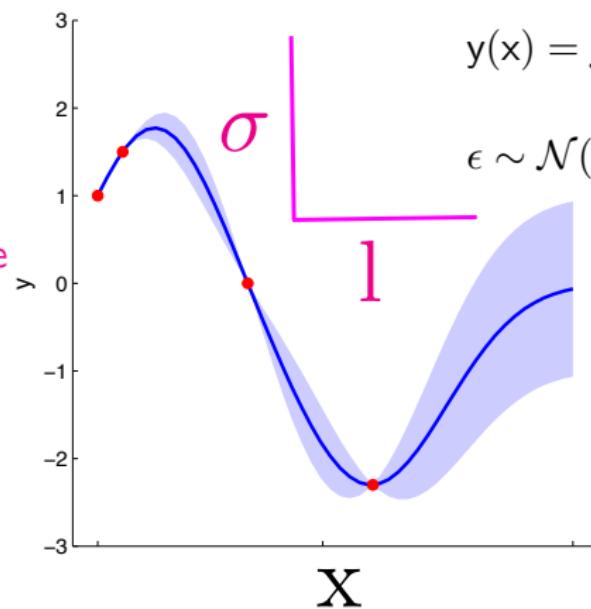


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

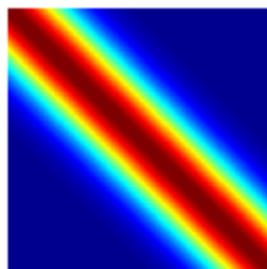
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

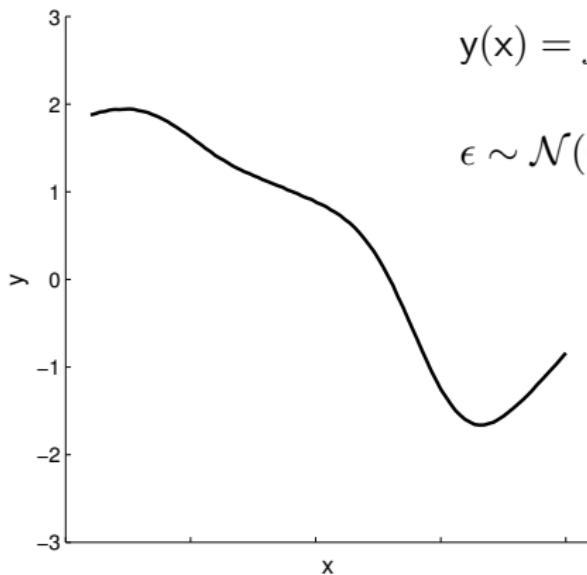
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



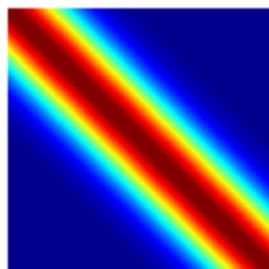
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

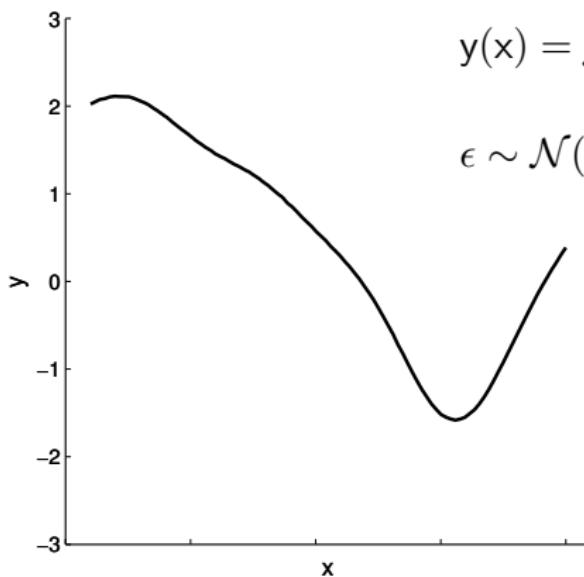


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

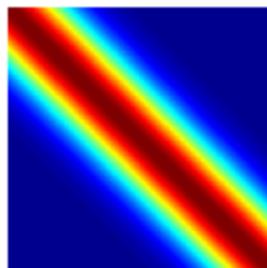
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

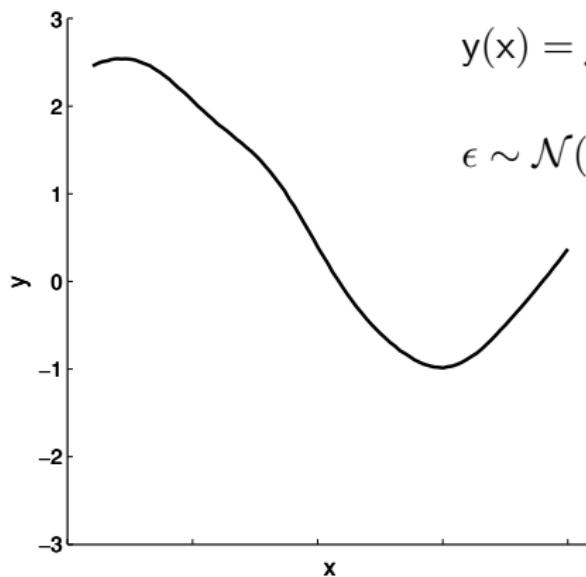
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

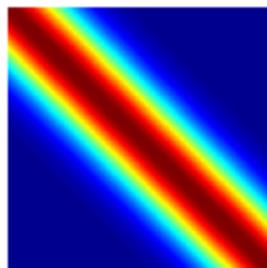
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

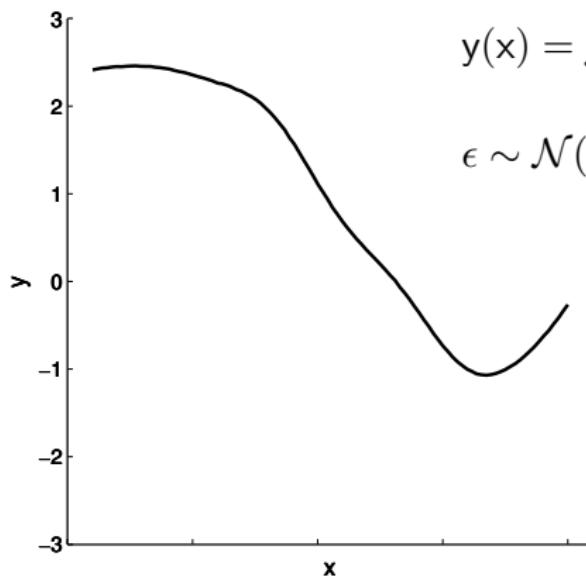
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

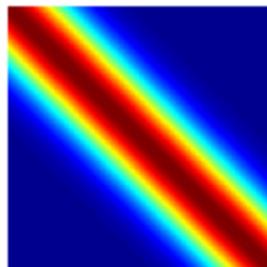
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

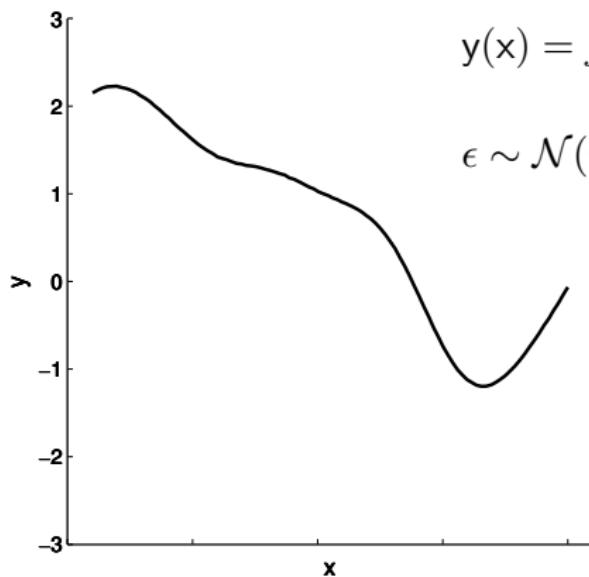
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

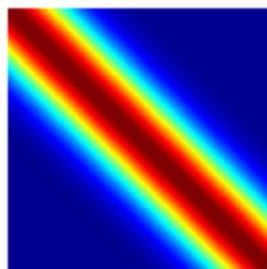
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

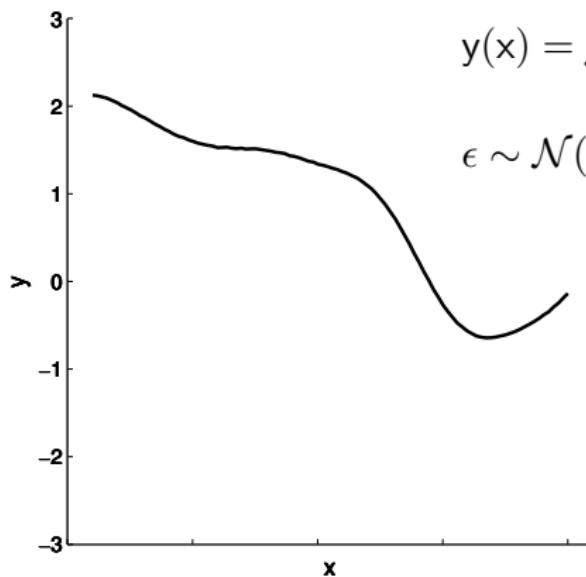
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

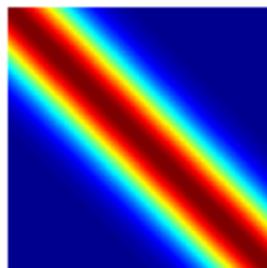
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

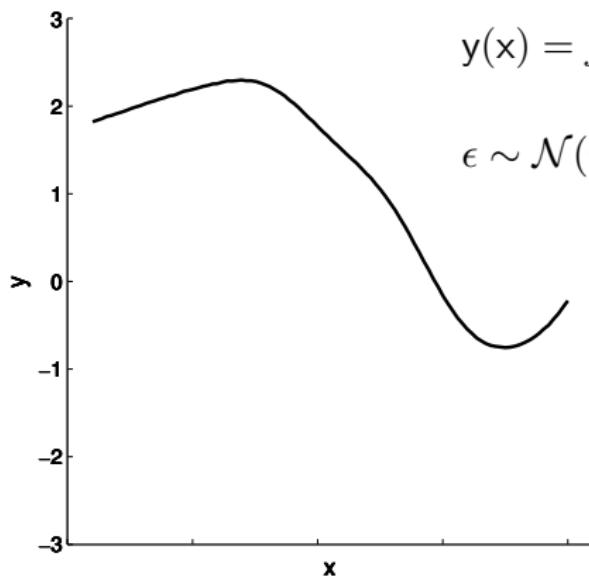
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

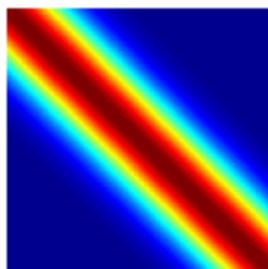
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

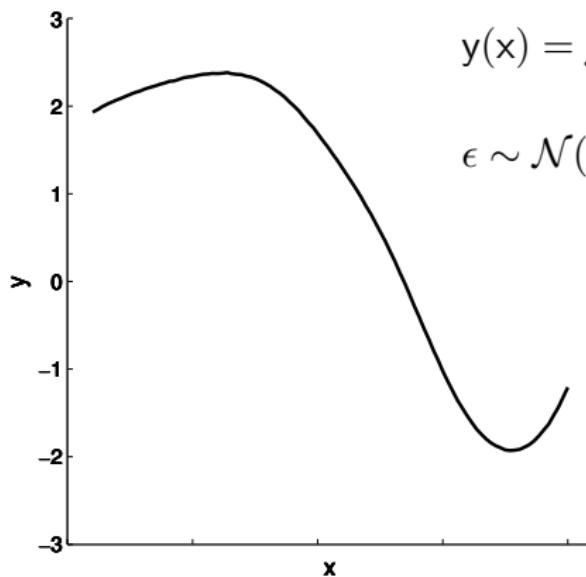
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

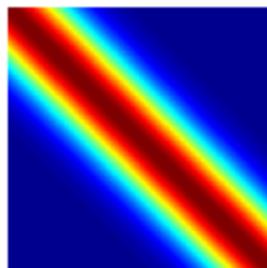
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

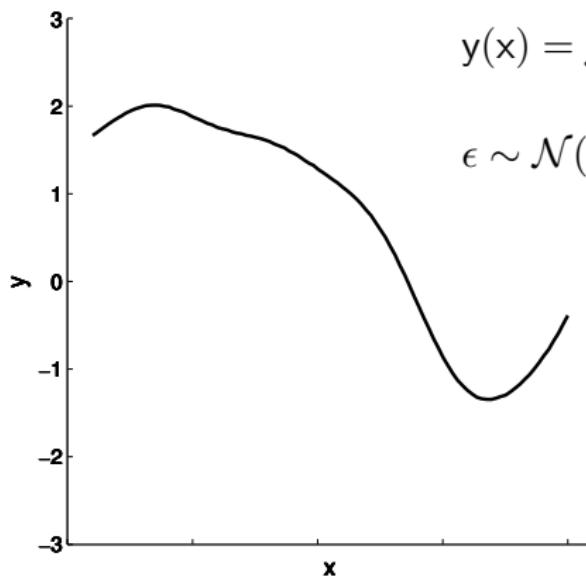
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



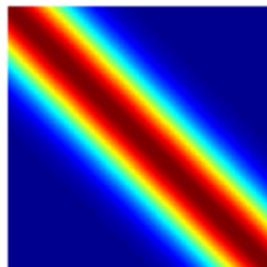
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

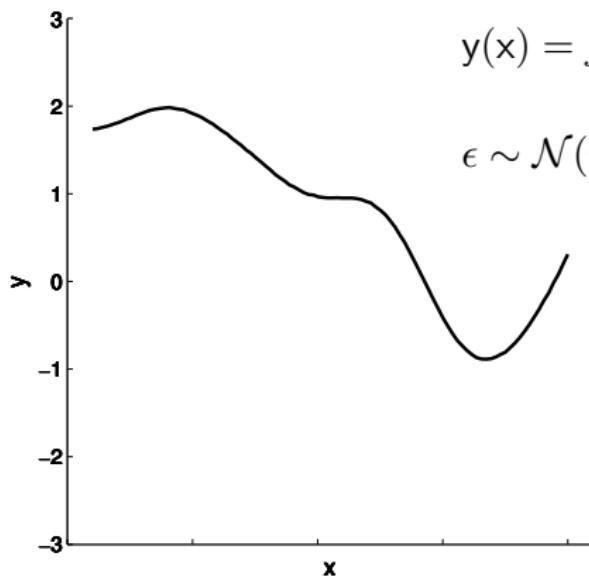


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



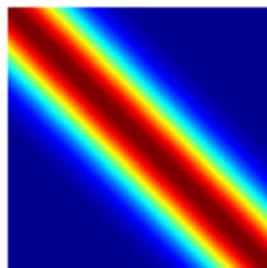
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

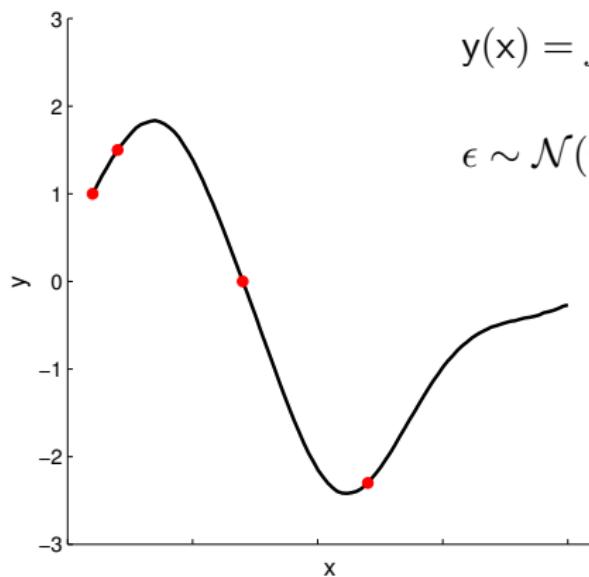


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



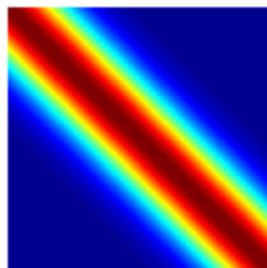
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

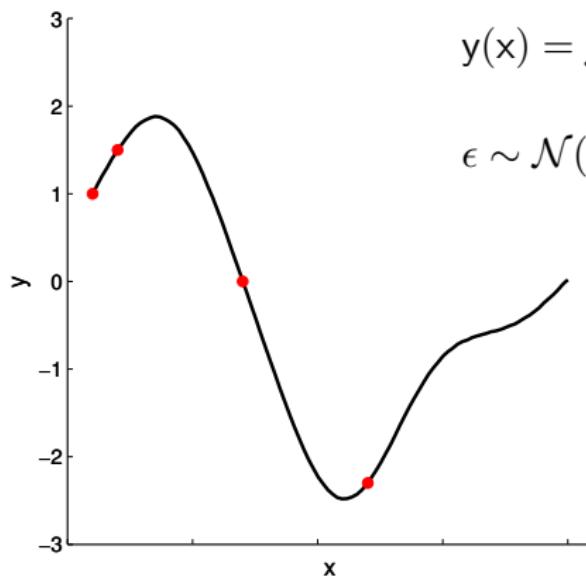


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



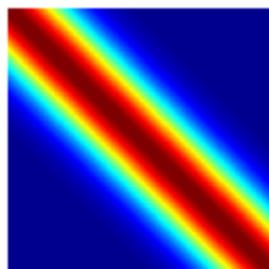
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

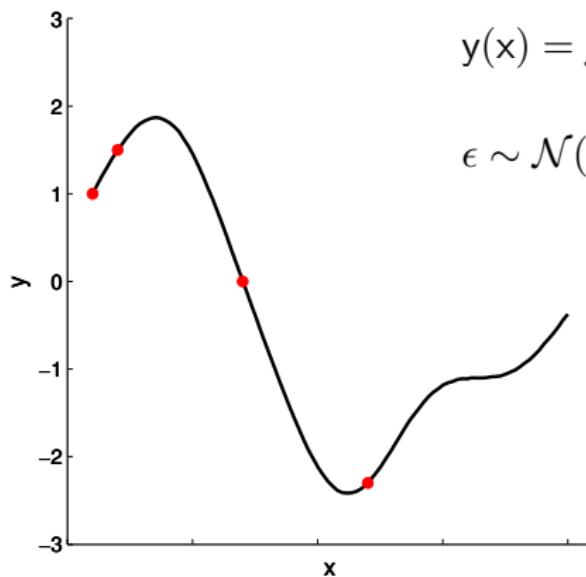


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



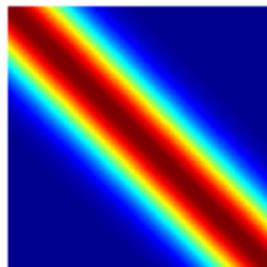
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

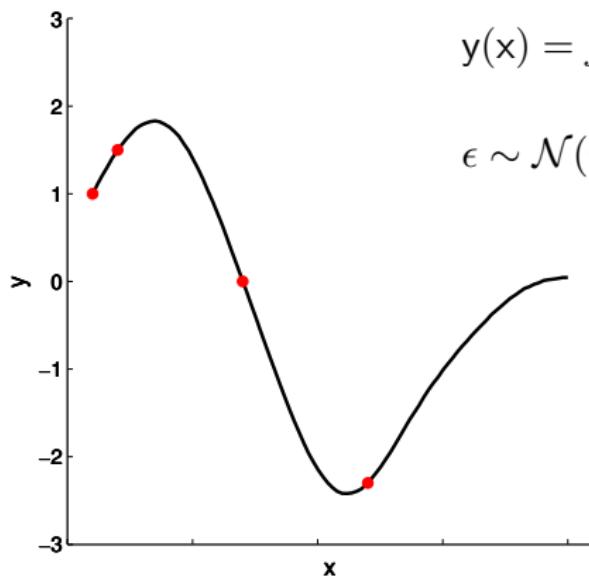


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



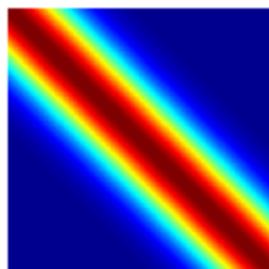
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

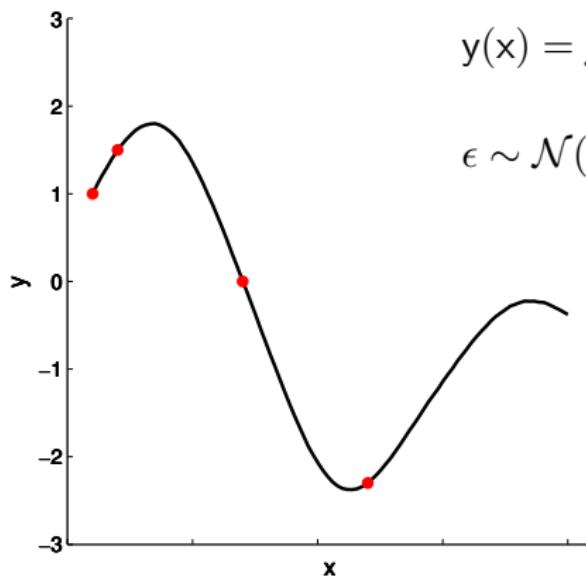


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



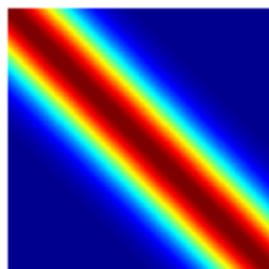
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

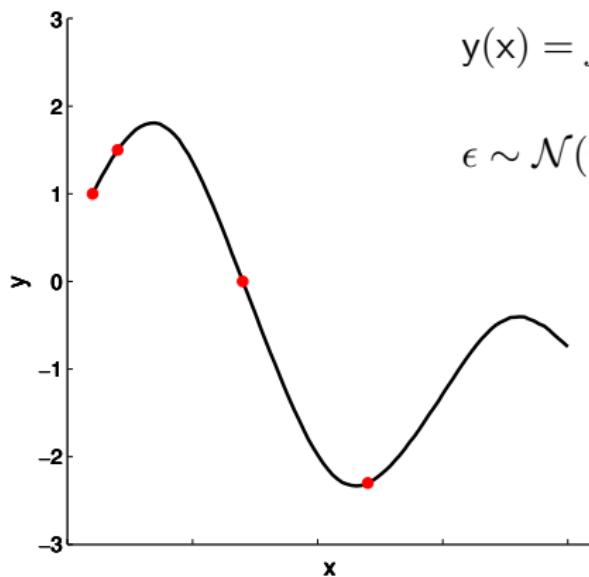


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



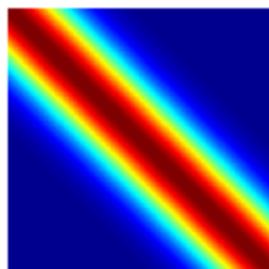
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

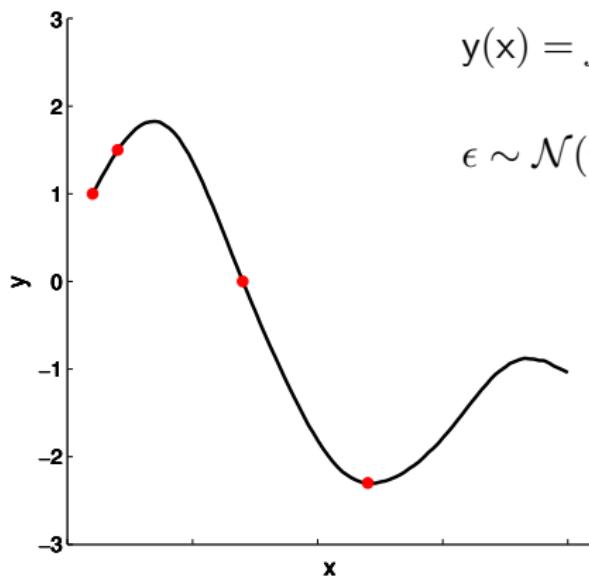


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



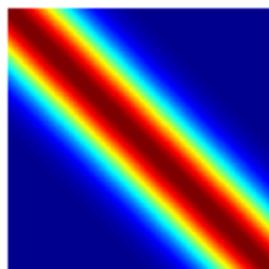
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

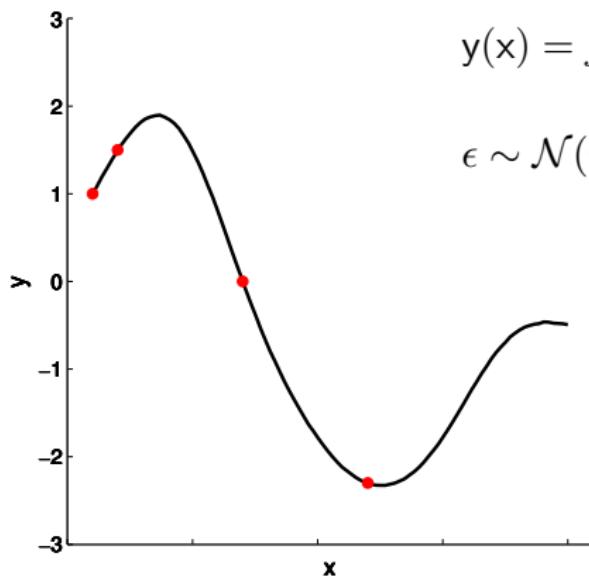


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



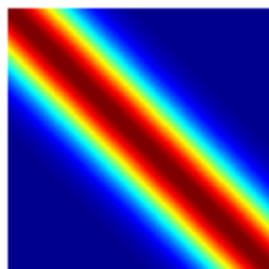
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

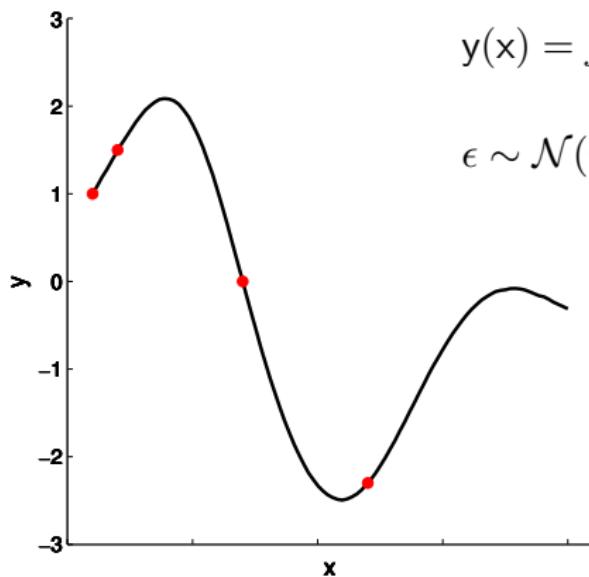


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



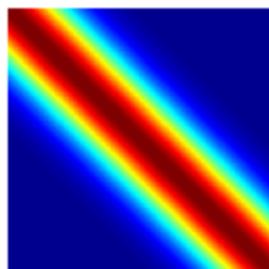
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

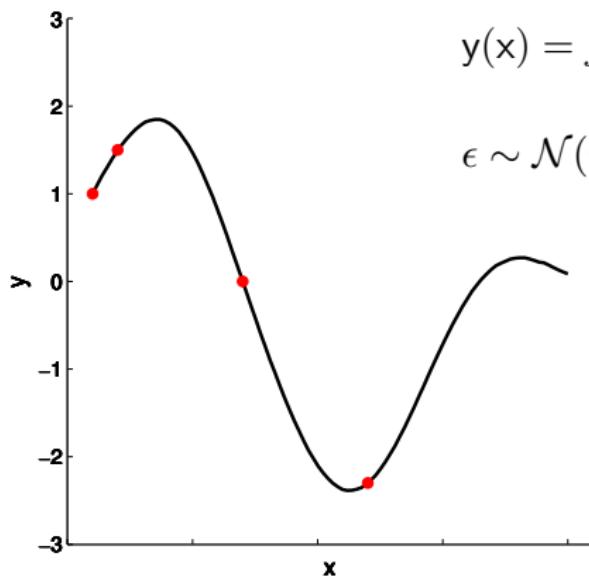


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



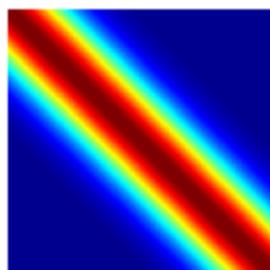
What effect do the hyper-parameters have?

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

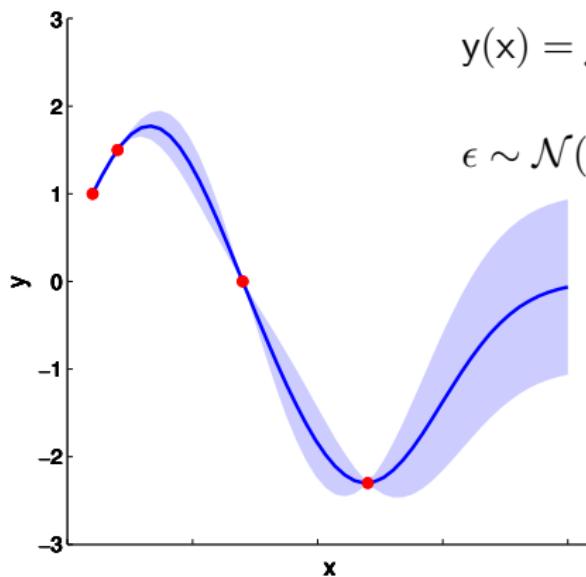


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

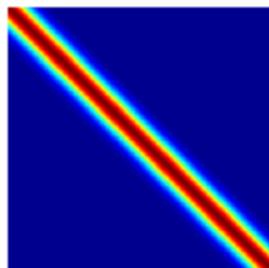
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

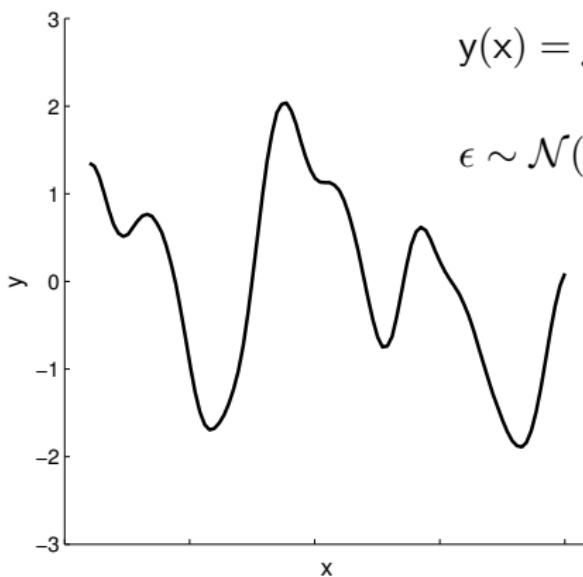


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

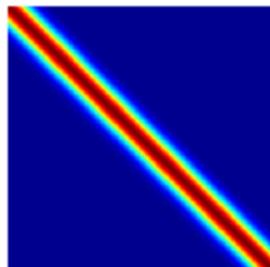
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

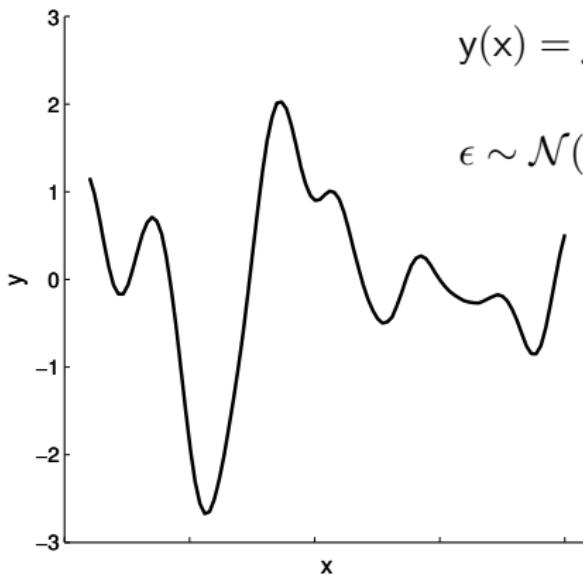


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

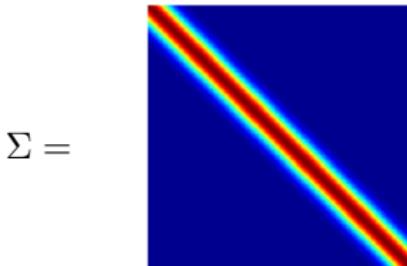
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

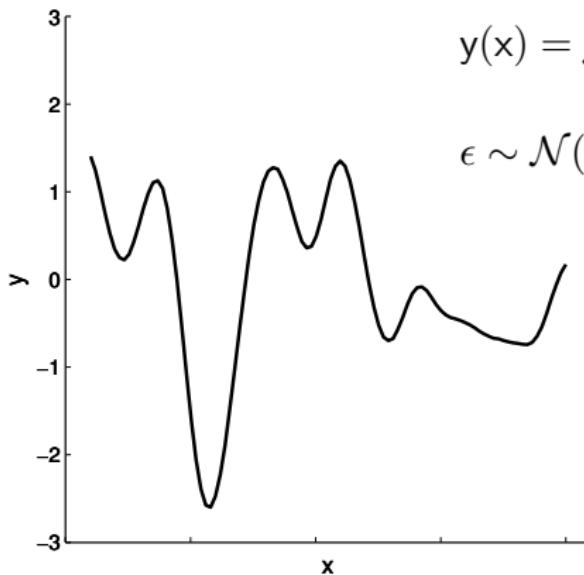


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

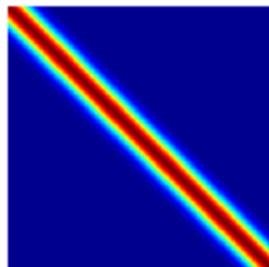
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

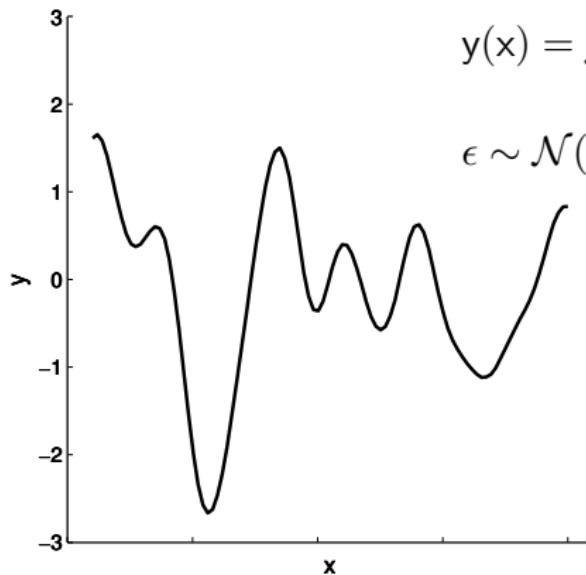


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

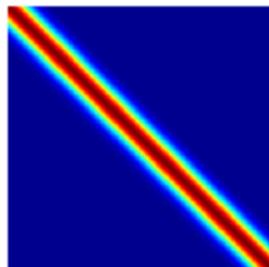
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

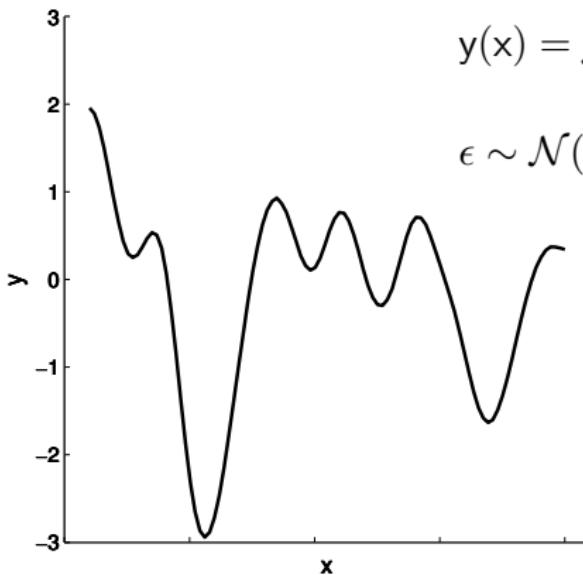


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

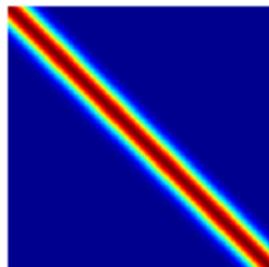
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

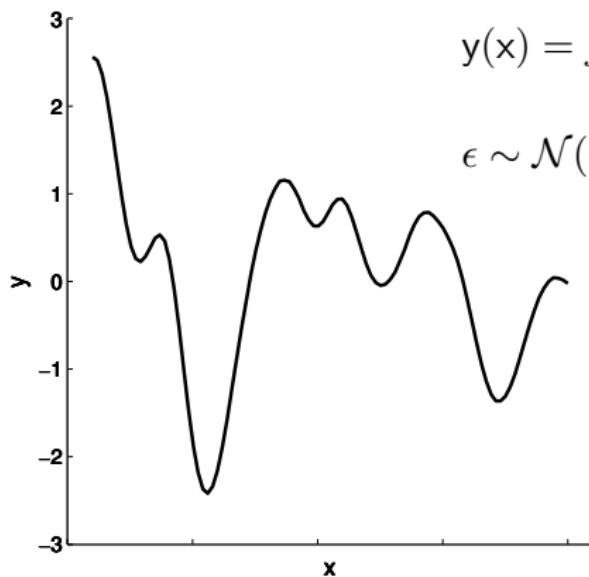


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

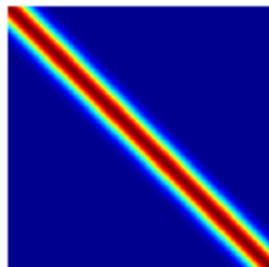
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

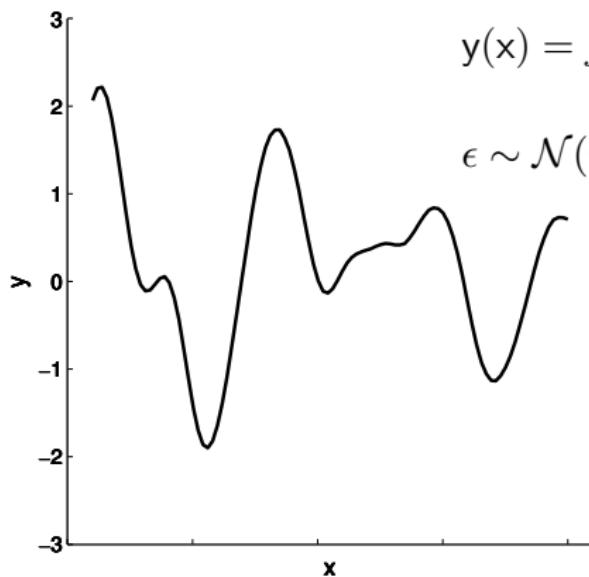


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

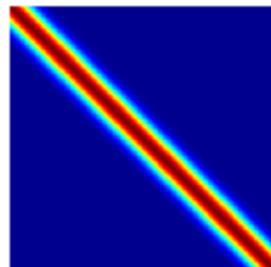
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

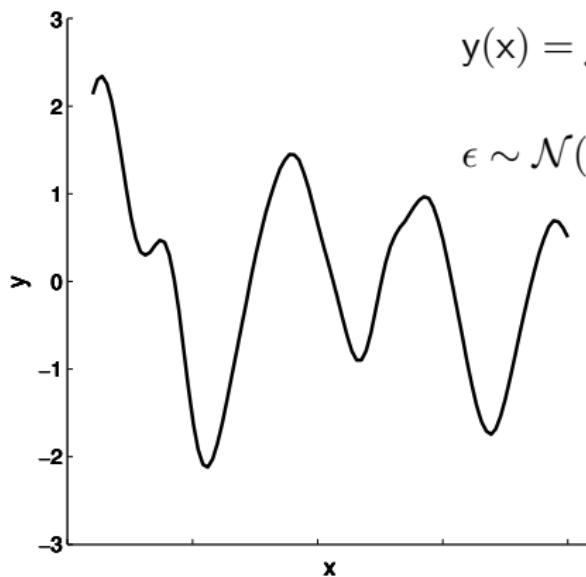


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

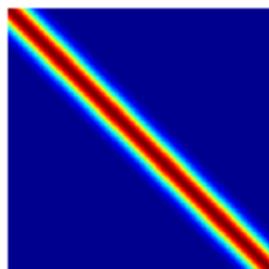
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

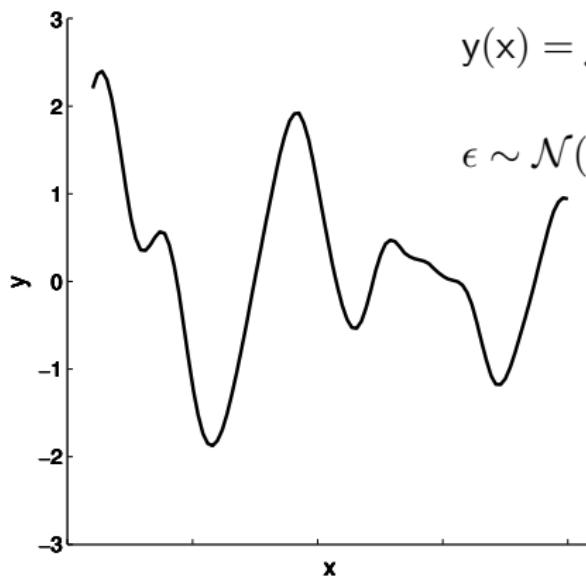


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

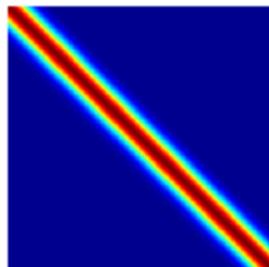
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

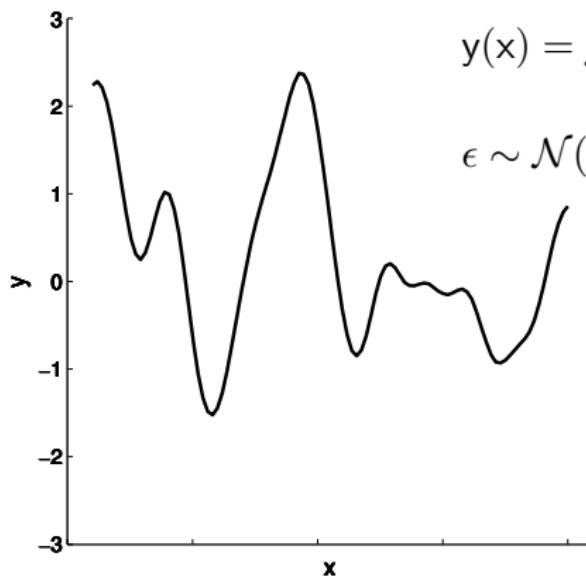


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

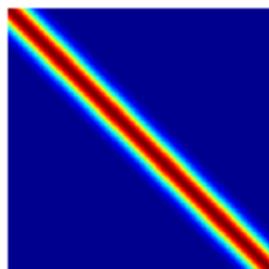
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

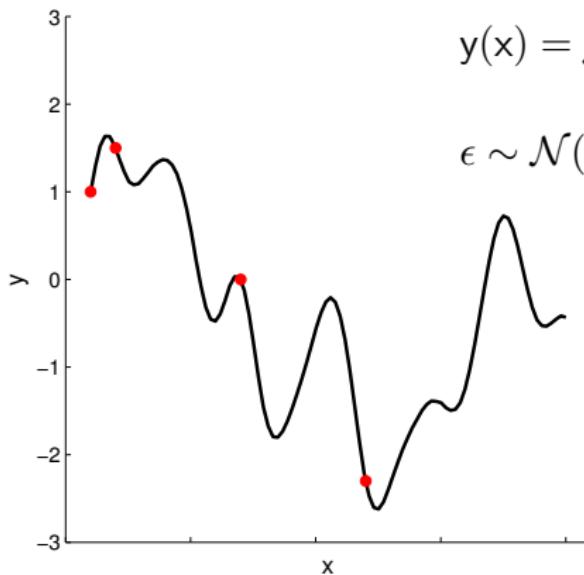


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

short horizontal length-scale

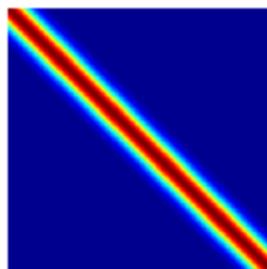
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

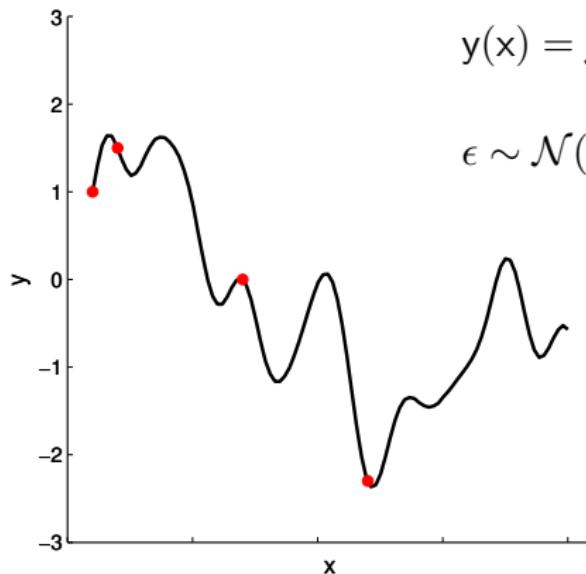
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

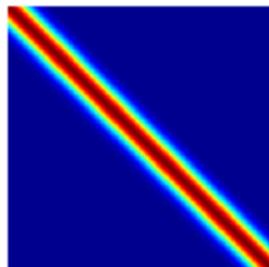
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

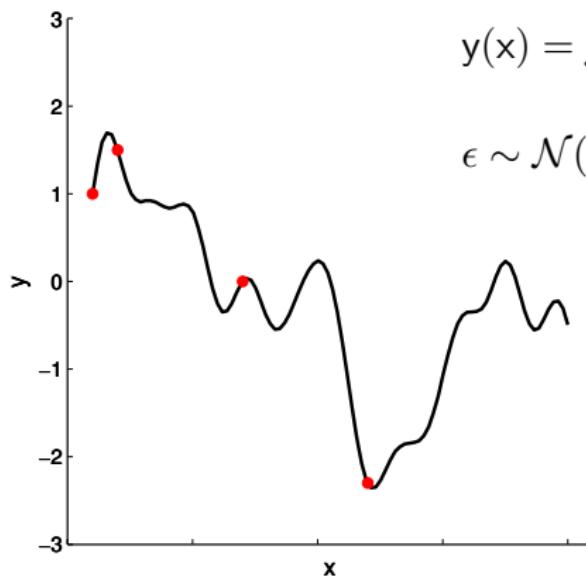


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

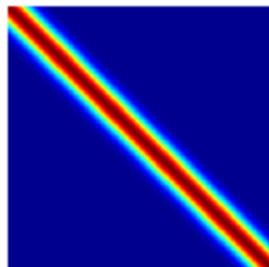
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

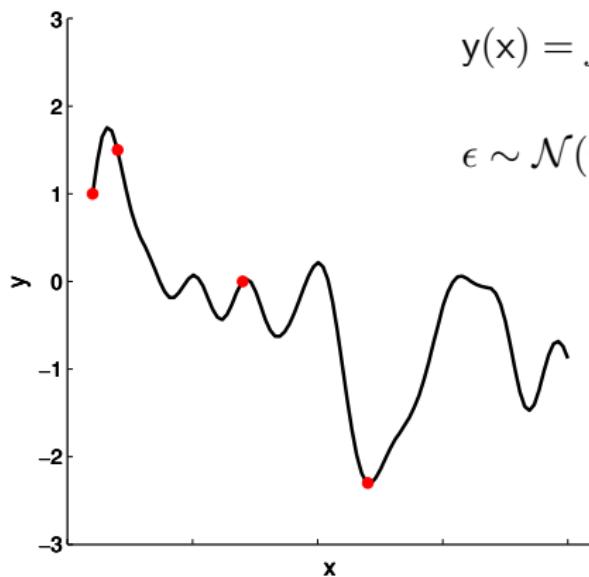


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

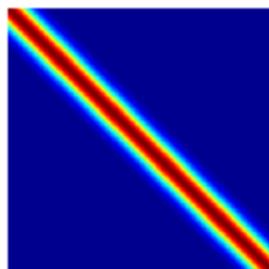
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

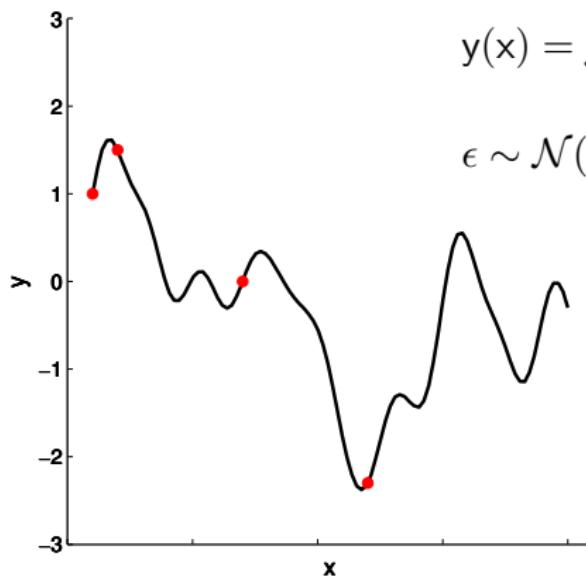


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

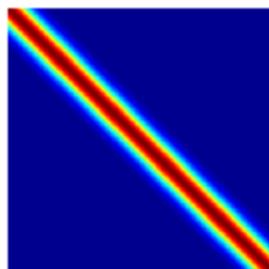
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

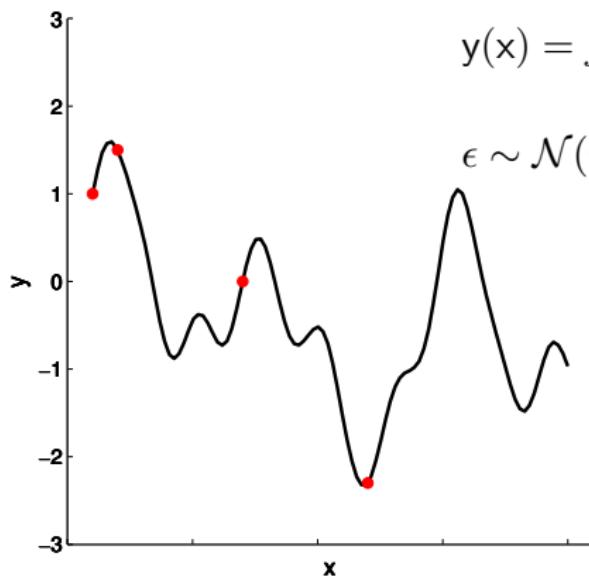


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

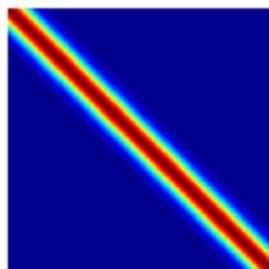
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

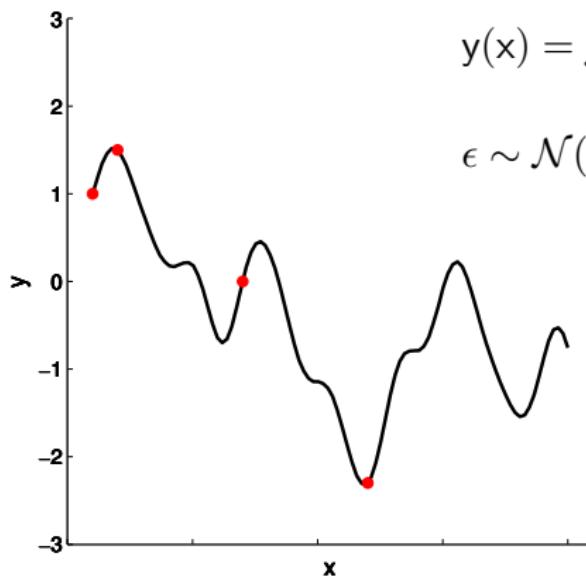


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

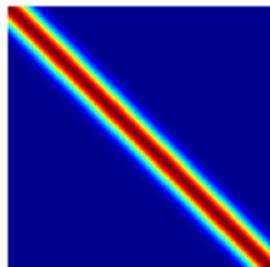
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

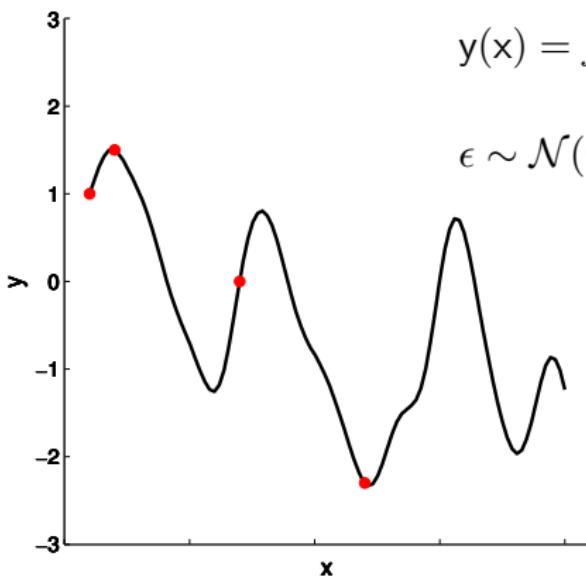


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

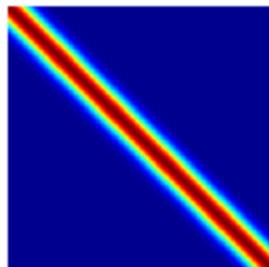
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

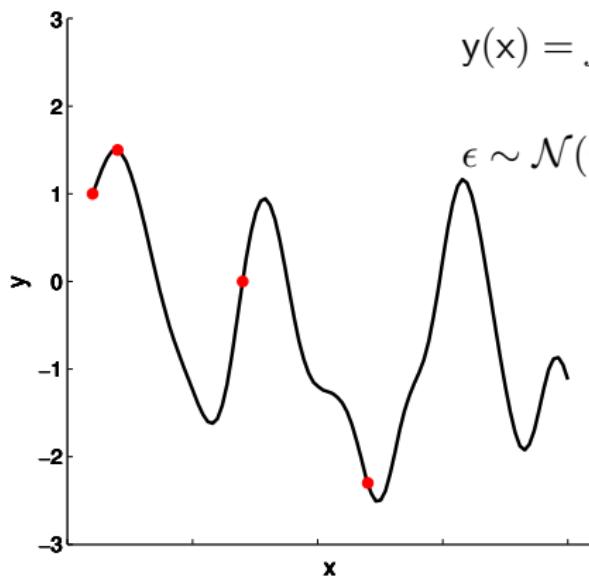


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

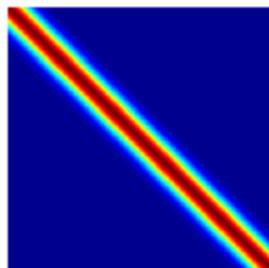
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

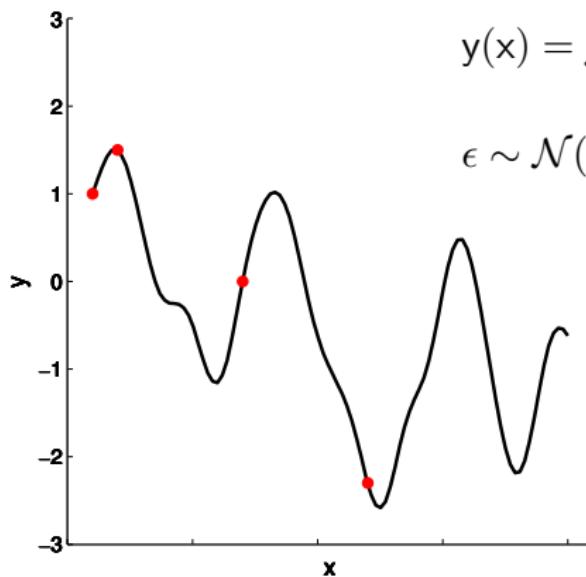


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

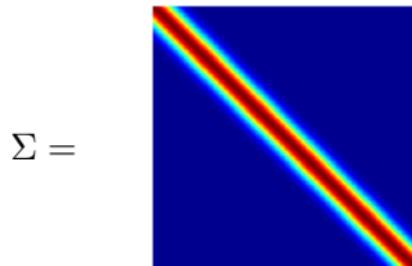
short horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

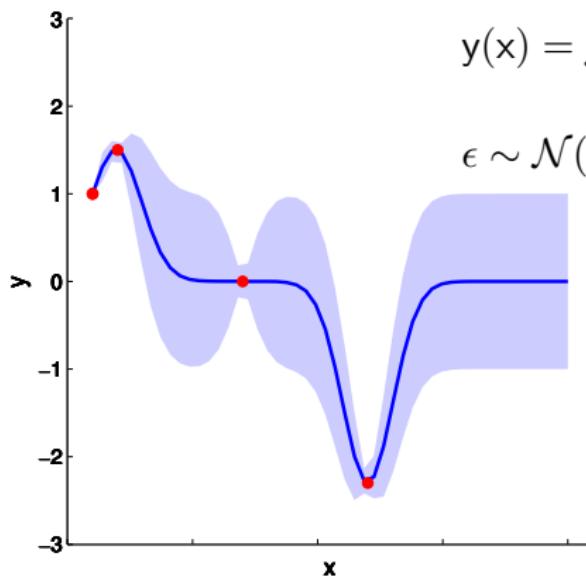


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

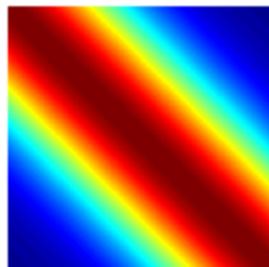
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

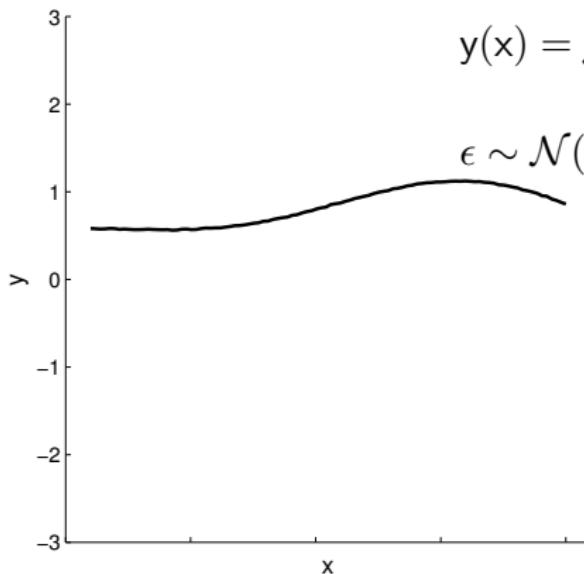


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

long horizontal length-scale

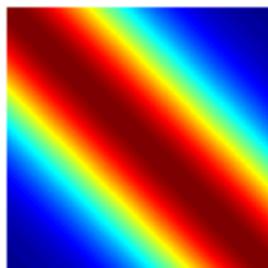
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

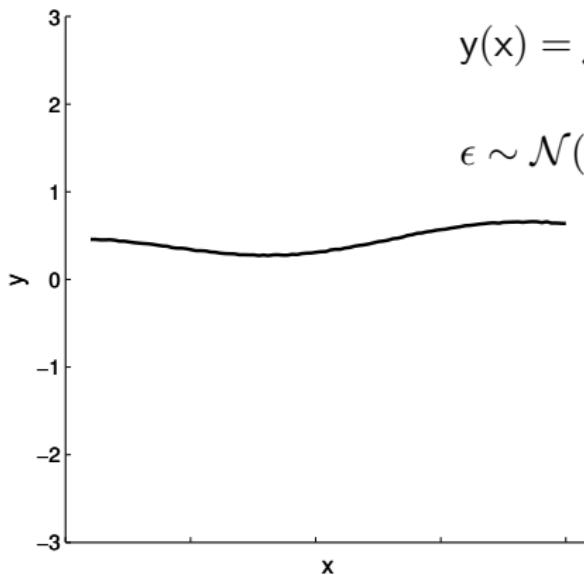
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

long horizontal length-scale

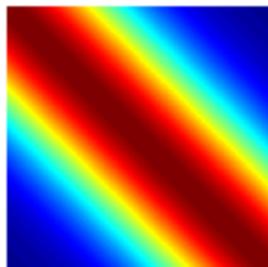
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

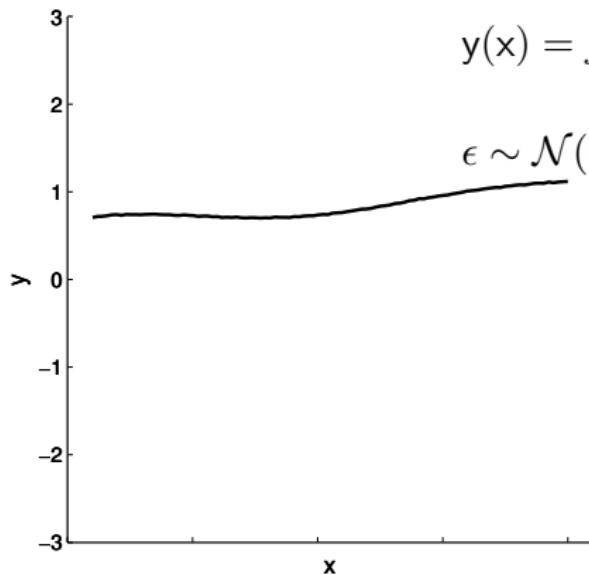
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

long horizontal length-scale

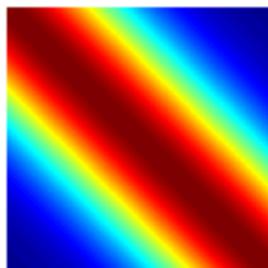
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

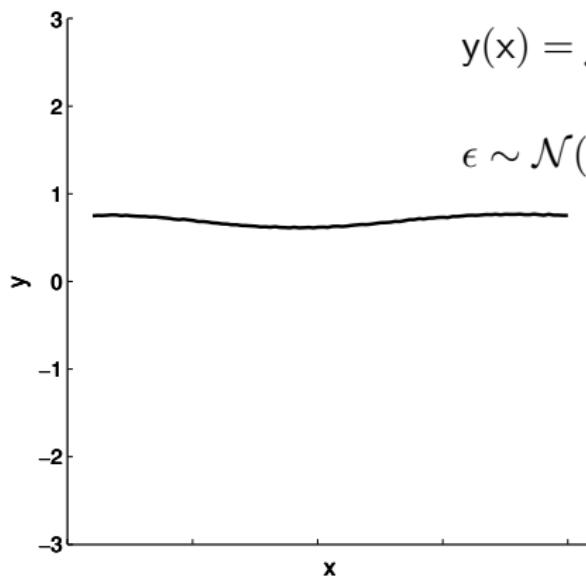
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

long horizontal length-scale

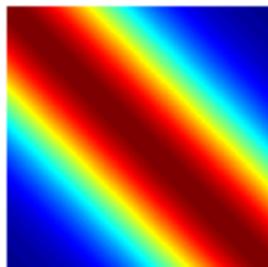
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

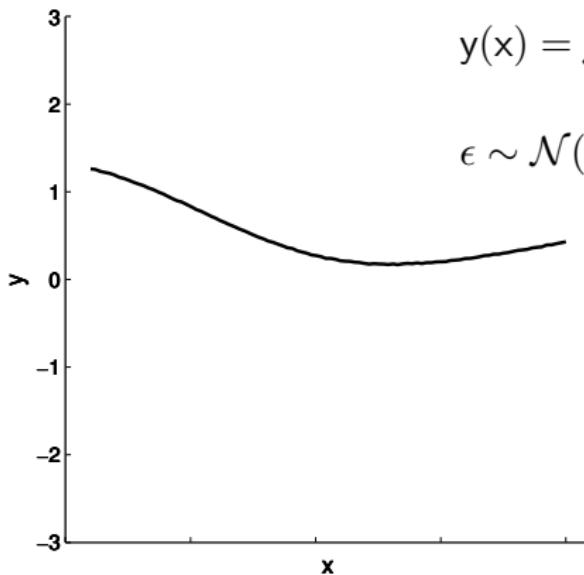
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

long horizontal length-scale

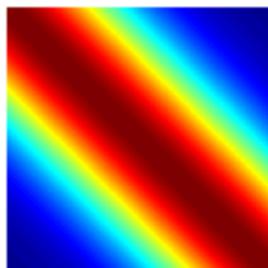
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

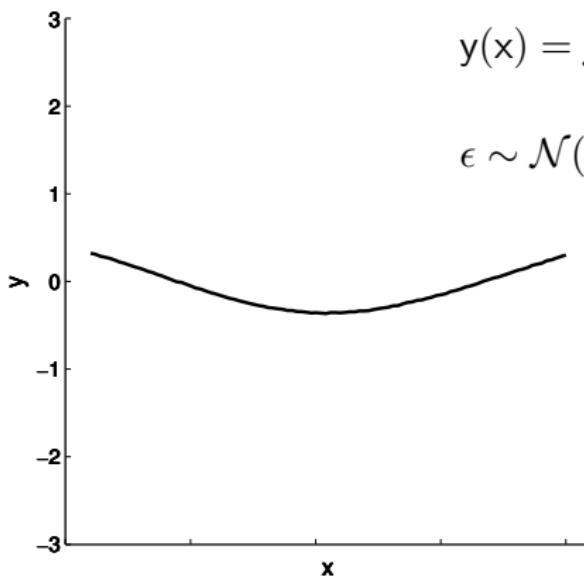
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

long horizontal length-scale

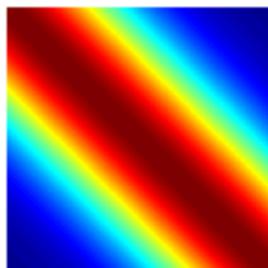
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

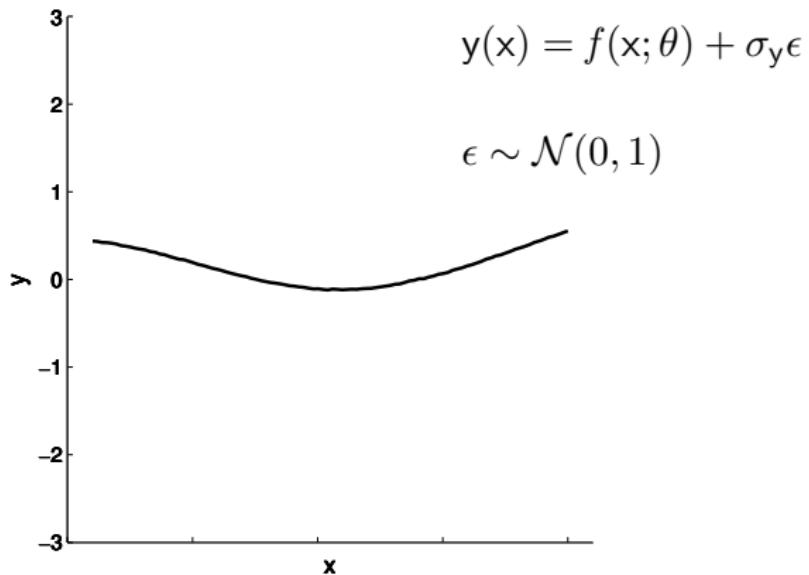
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

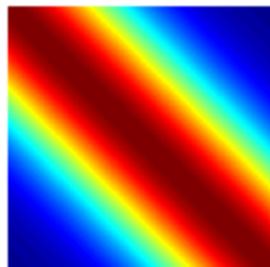
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

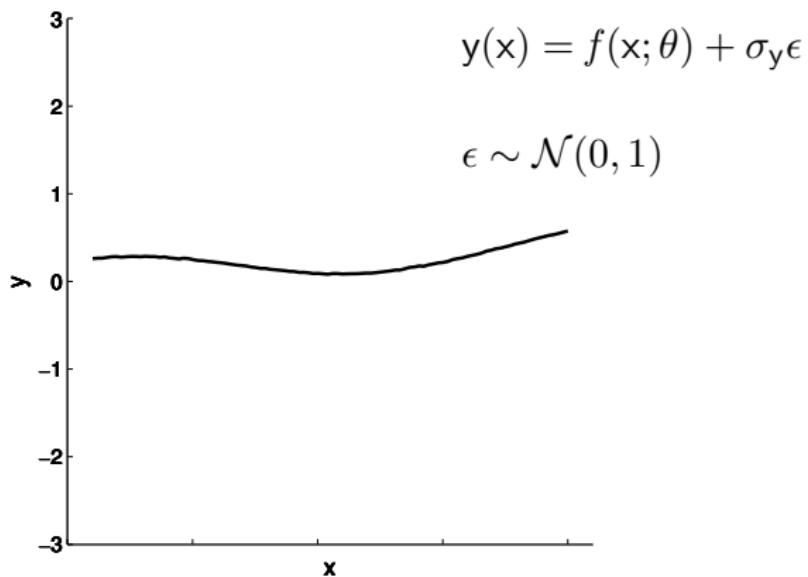


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

long horizontal length-scale

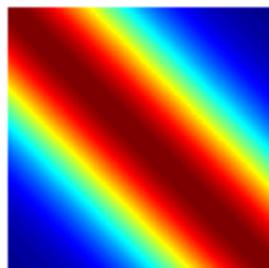
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

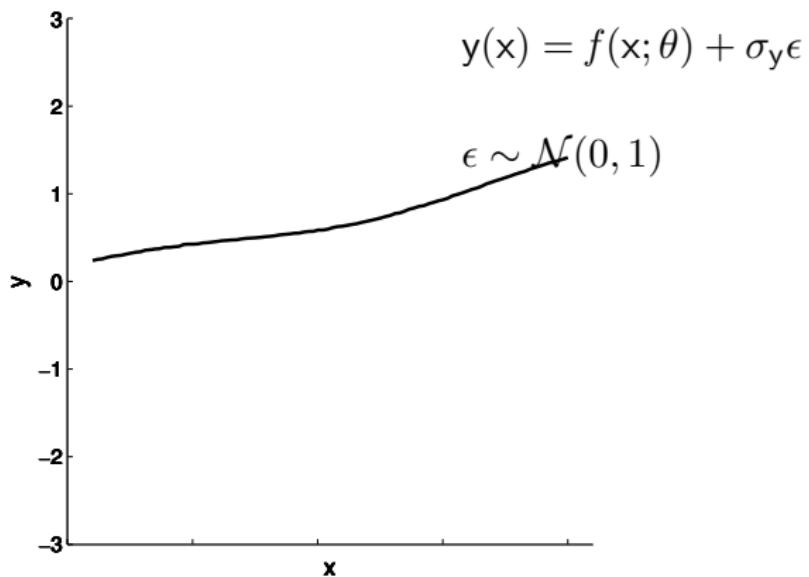
$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

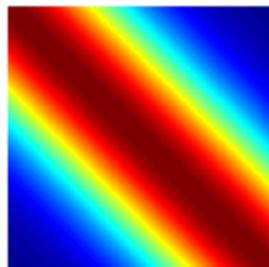
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

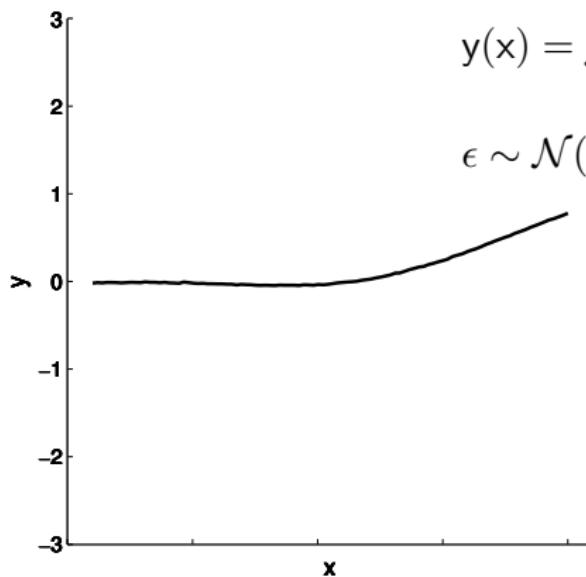


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

long horizontal length-scale

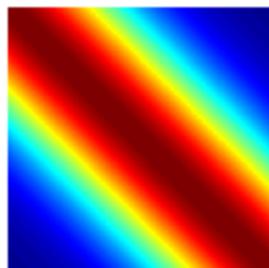
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

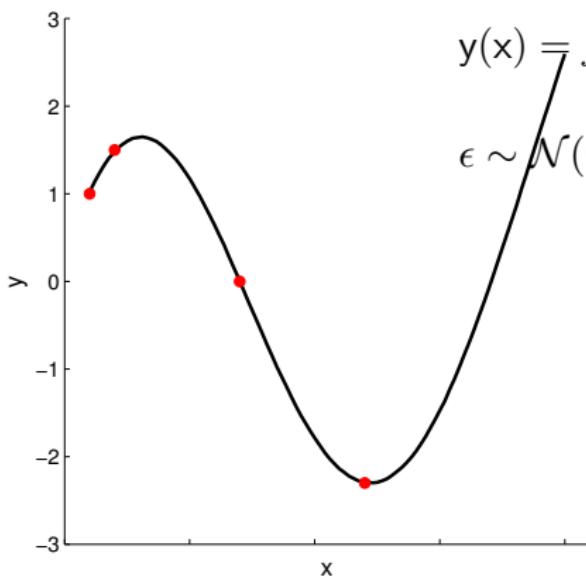
$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$
$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

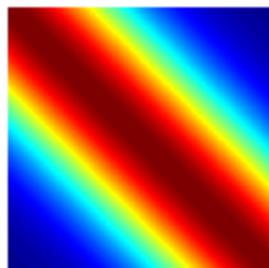
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

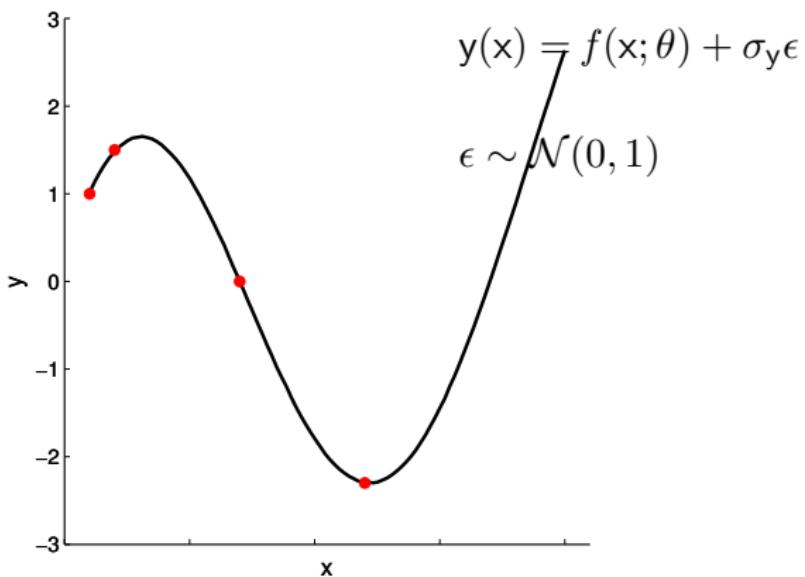
$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



$$\Sigma =$$

Parametric model



What effect do the hyper-parameters have?

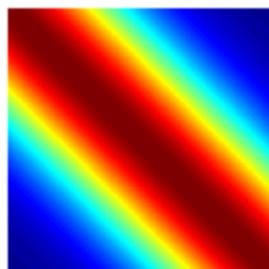
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

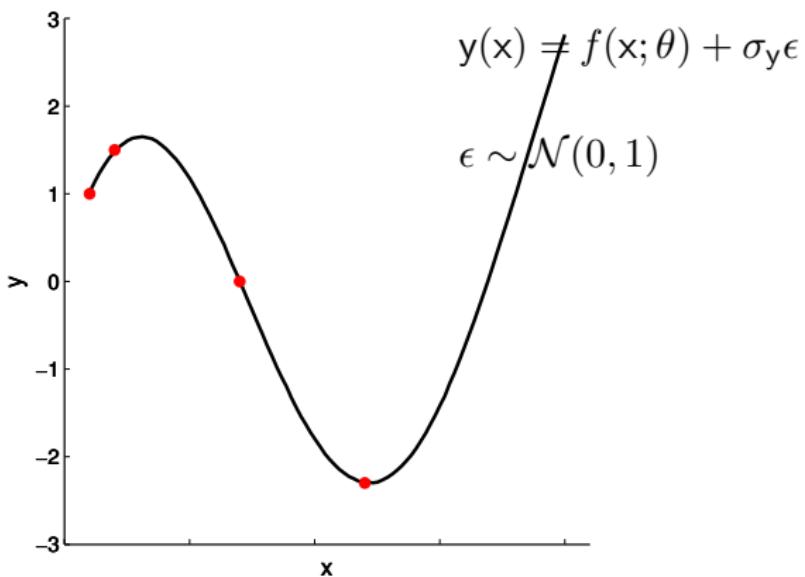
$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



$$\Sigma =$$

Parametric model



What effect do the hyper-parameters have?

long horizontal length-scale

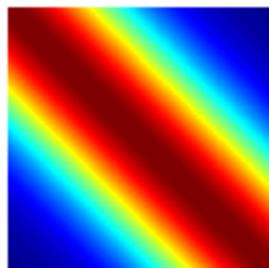
Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

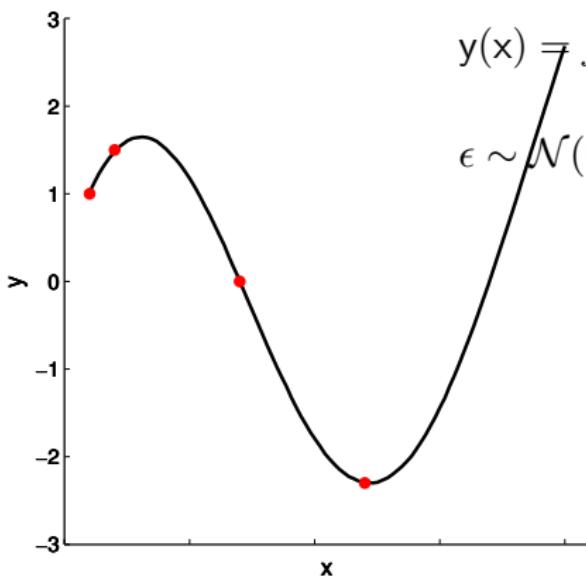
$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

$$\Sigma =$$



Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$
$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

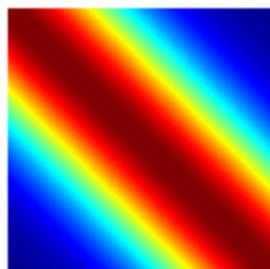
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

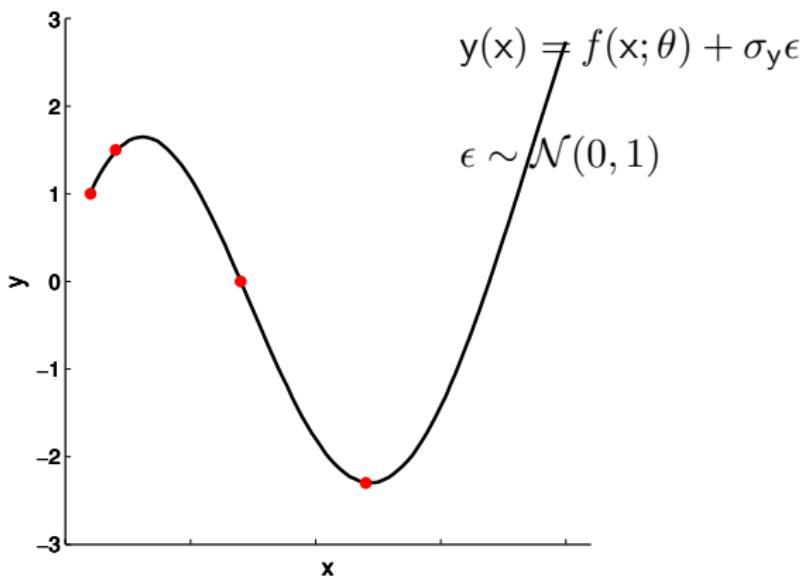
$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



$$\Sigma =$$

Parametric model



What effect do the hyper-parameters have?

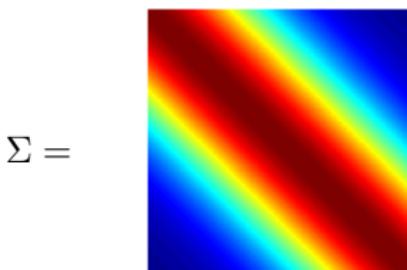
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

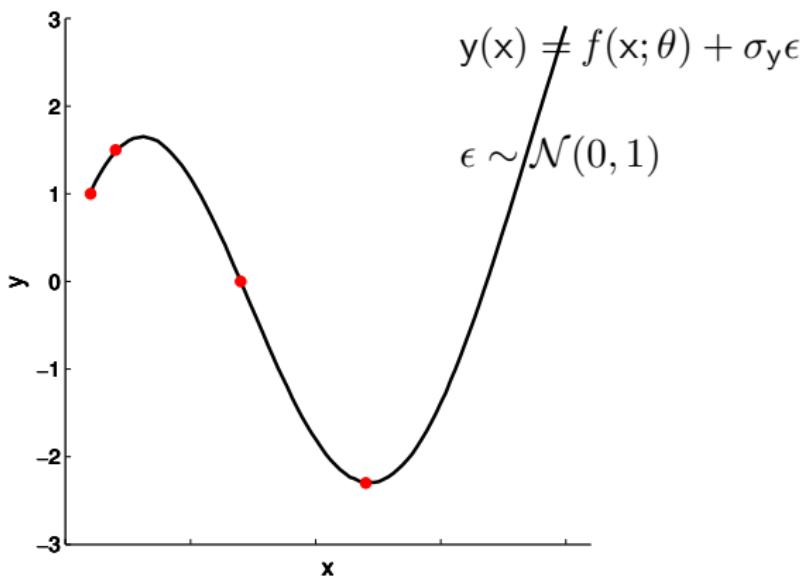
$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



$$\Sigma =$$

Parametric model



What effect do the hyper-parameters have?

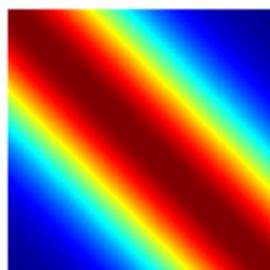
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

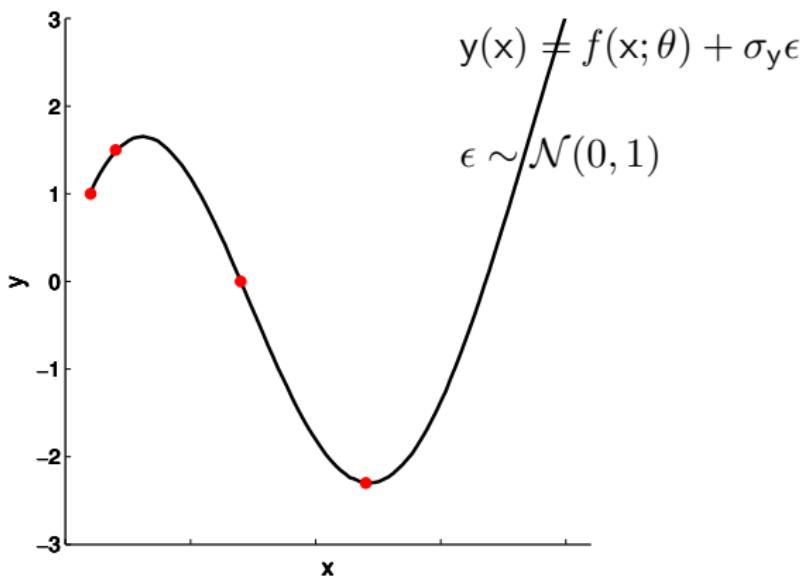
$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



$$\Sigma =$$

Parametric model



What effect do the hyper-parameters have?

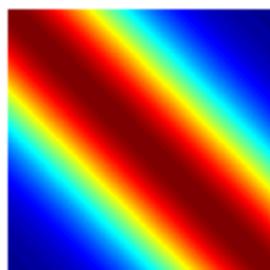
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

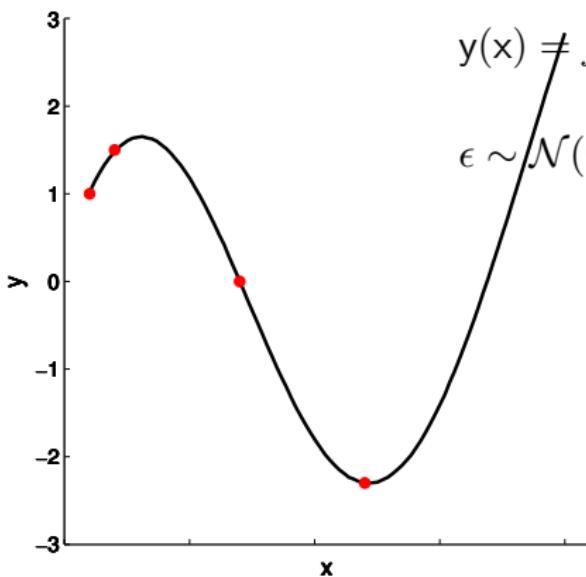
$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$
$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

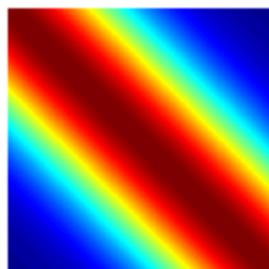
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

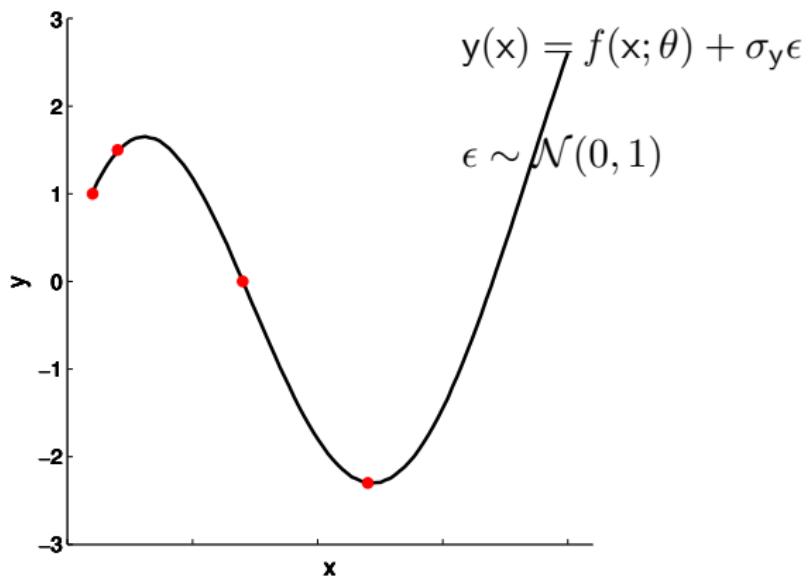
$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



$$\Sigma =$$

Parametric model



$$y(x) = f(x; \theta) + \sigma_y \epsilon$$
$$\epsilon \sim \mathcal{N}(0, 1)$$

What effect do the hyper-parameters have?

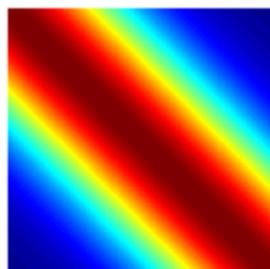
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

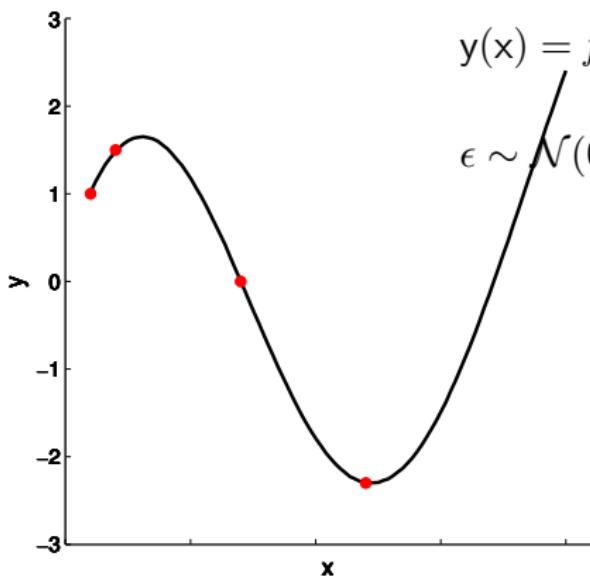
$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$



$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$
$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

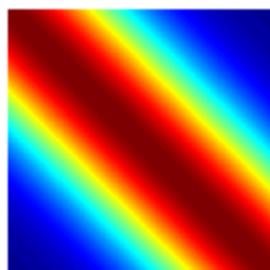
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

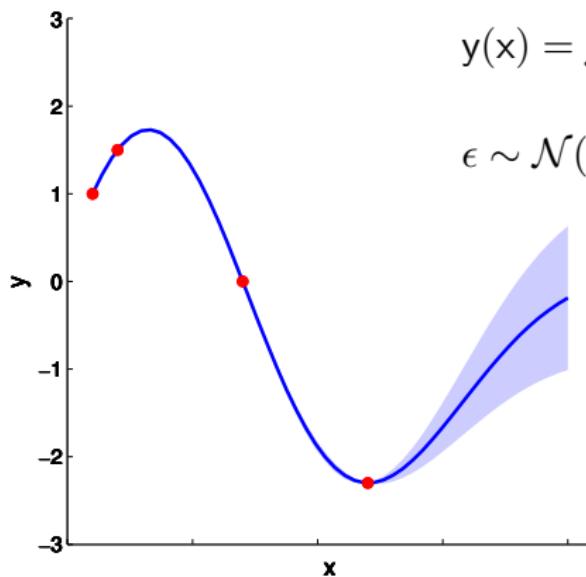


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

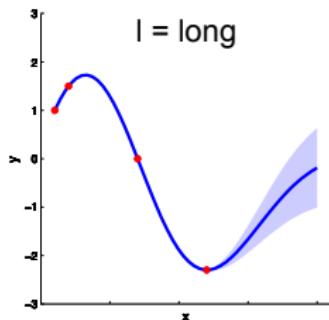
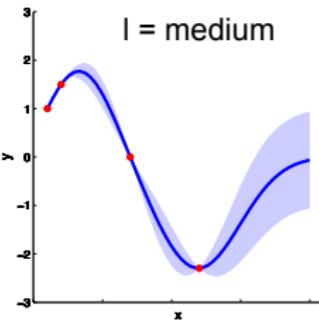
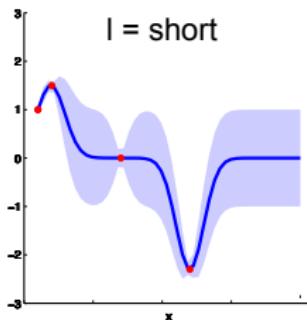
$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect do the hyper-parameters have?

$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

- Hyper-parameters have a strong effect
 - ▶ l controls the horizontal length-scale
 - ▶ σ^2 controls the vertical scale of the data
- \Rightarrow need automatic learning of hyper-parameters from data

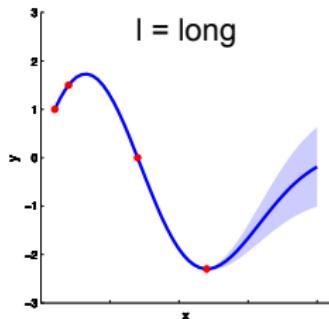
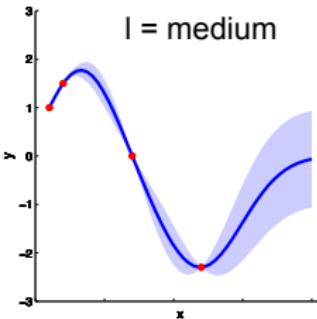
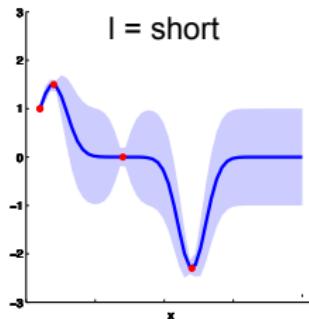


What effect do the hyper-parameters have?

$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

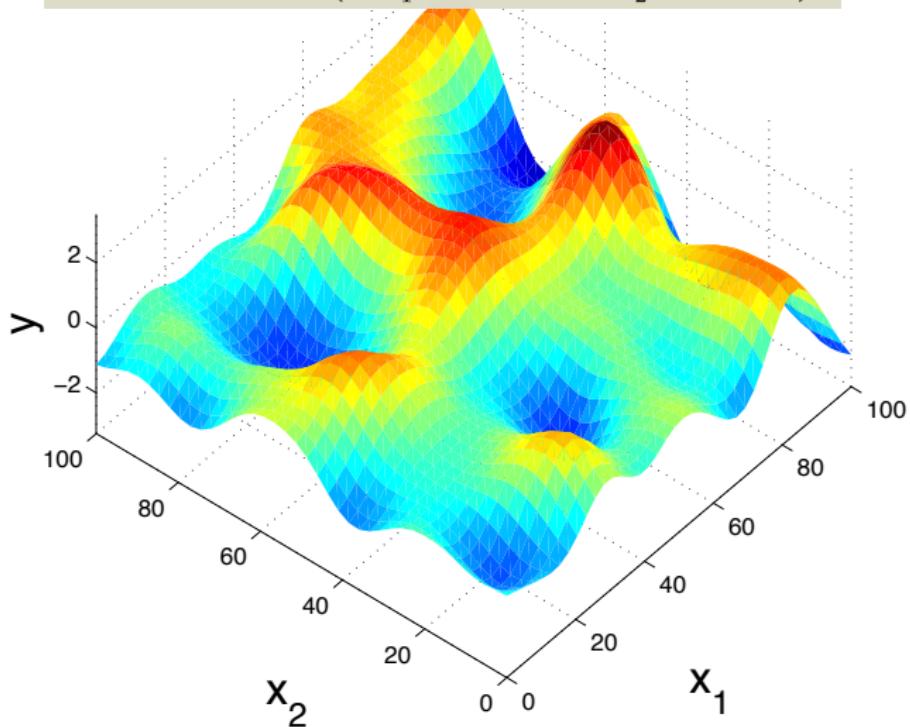
- Hyper-parameters have a strong effect
 - l controls the horizontal length-scale
 - σ^2 controls the vertical scale of the data
- ⇒ need automatic learning of hyper-parameters from data e.g.

$$\arg \max_{l, \sigma^2} \log p(\mathbf{y} | \theta)$$

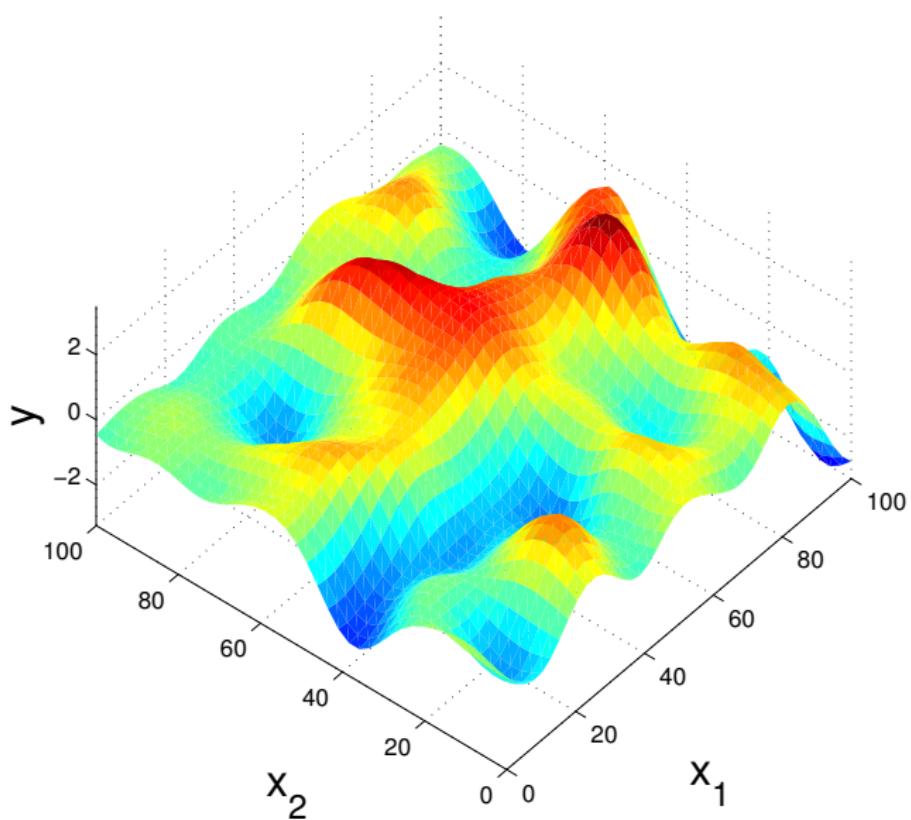


Higher dimensional input spaces

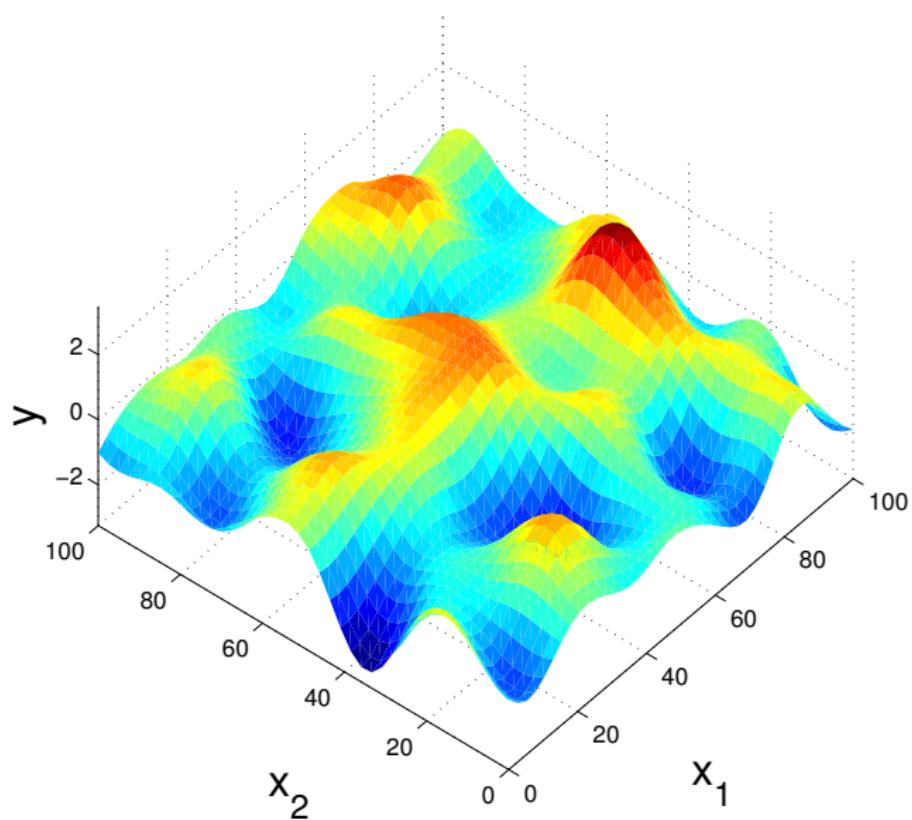
$$K(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp \left(-\frac{1}{2l_1^2}(\mathbf{x}_1 - \mathbf{x}'_1)^2 - \frac{1}{2l_2^2}(\mathbf{x}_2 - \mathbf{x}'_2)^2 \right)$$



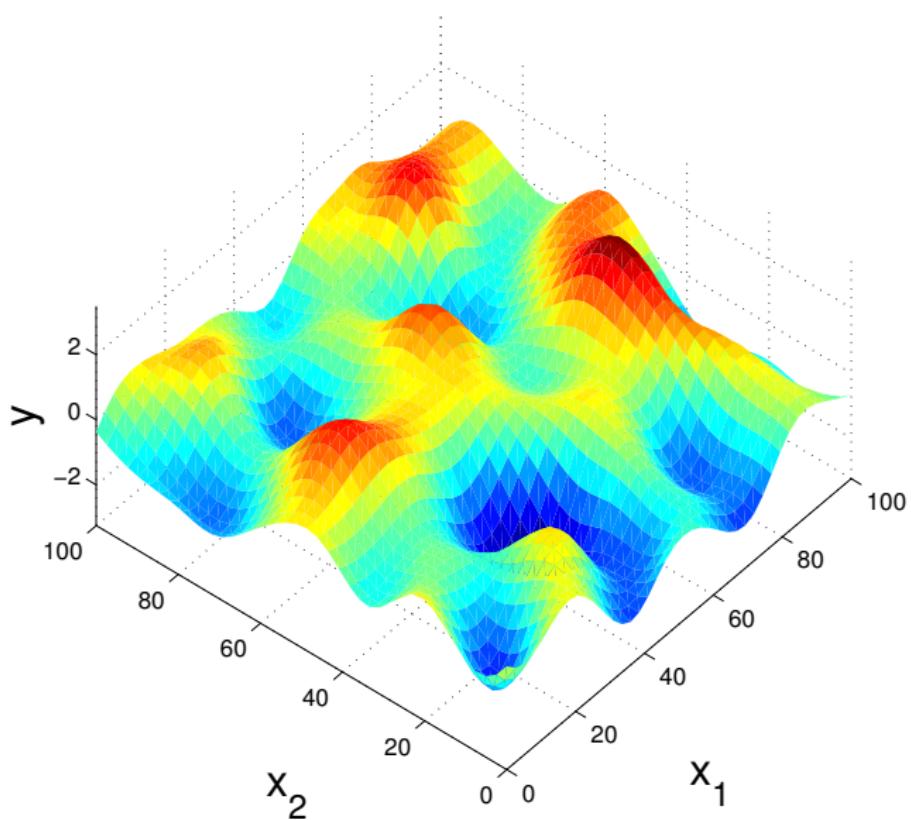
Higher dimensional input spaces



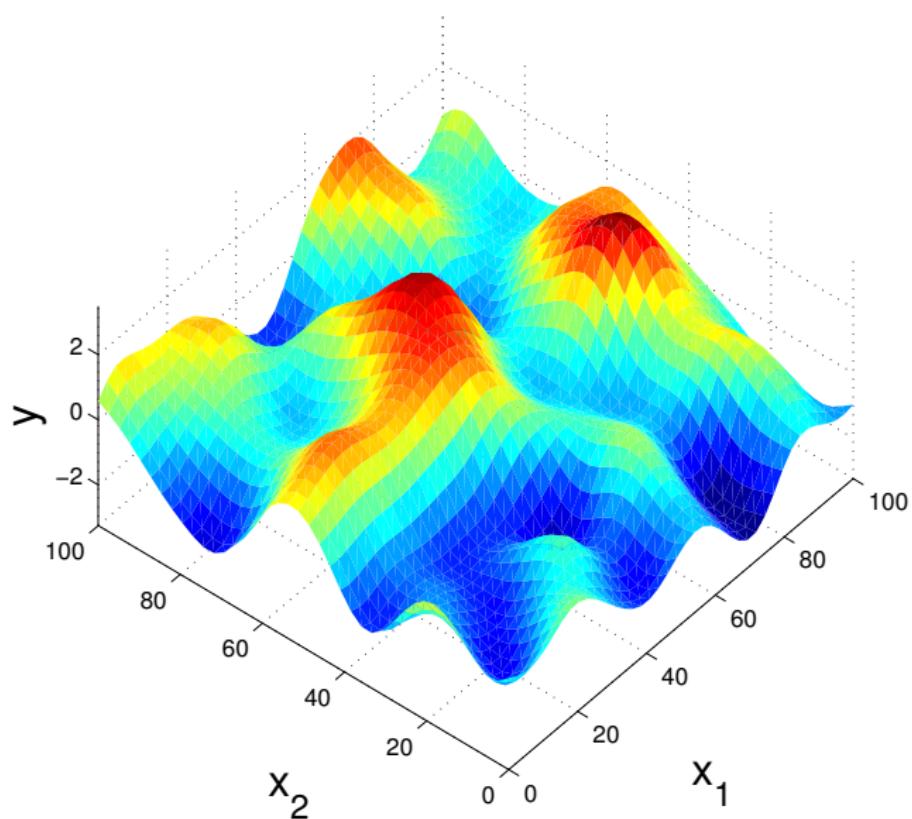
Higher dimensional input spaces



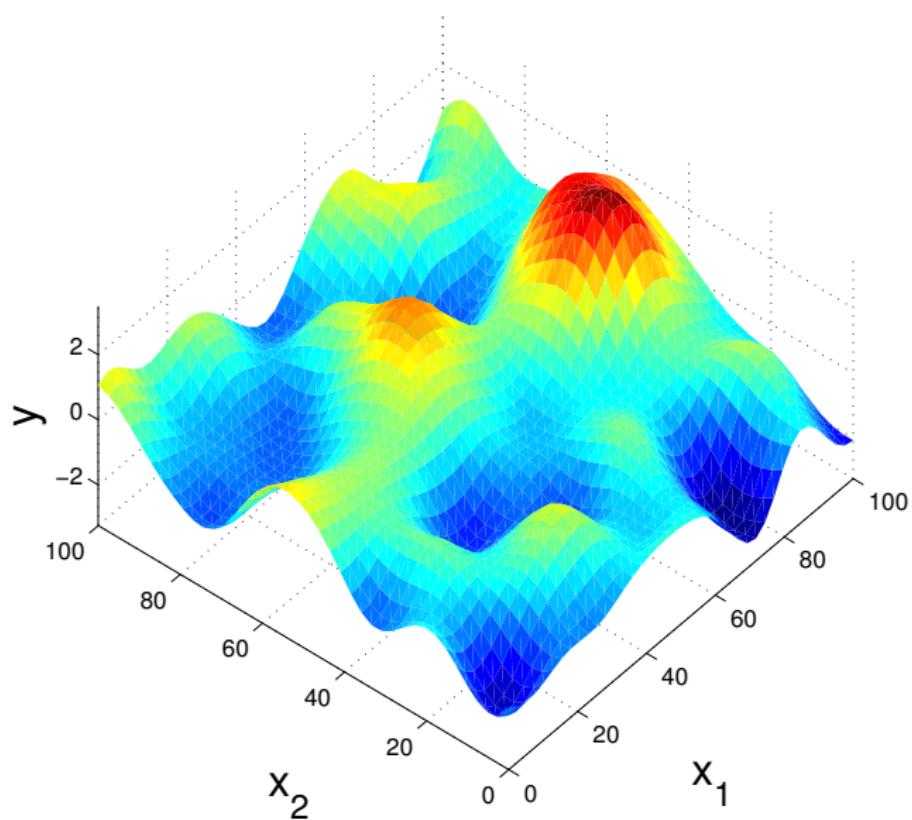
Higher dimensional input spaces



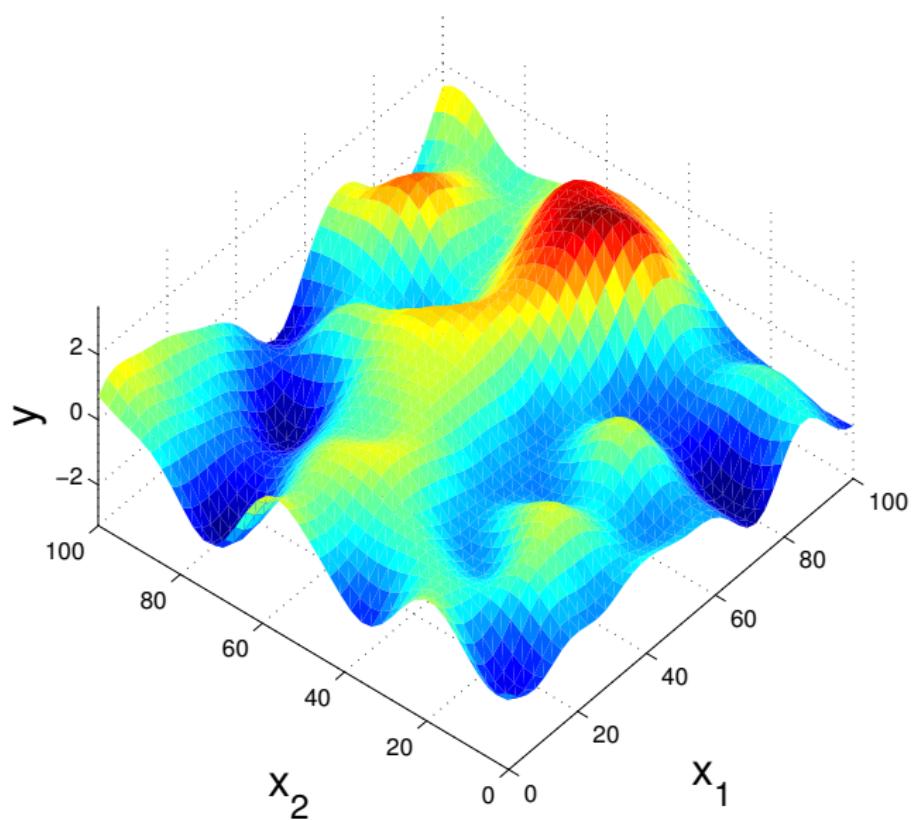
Higher dimensional input spaces



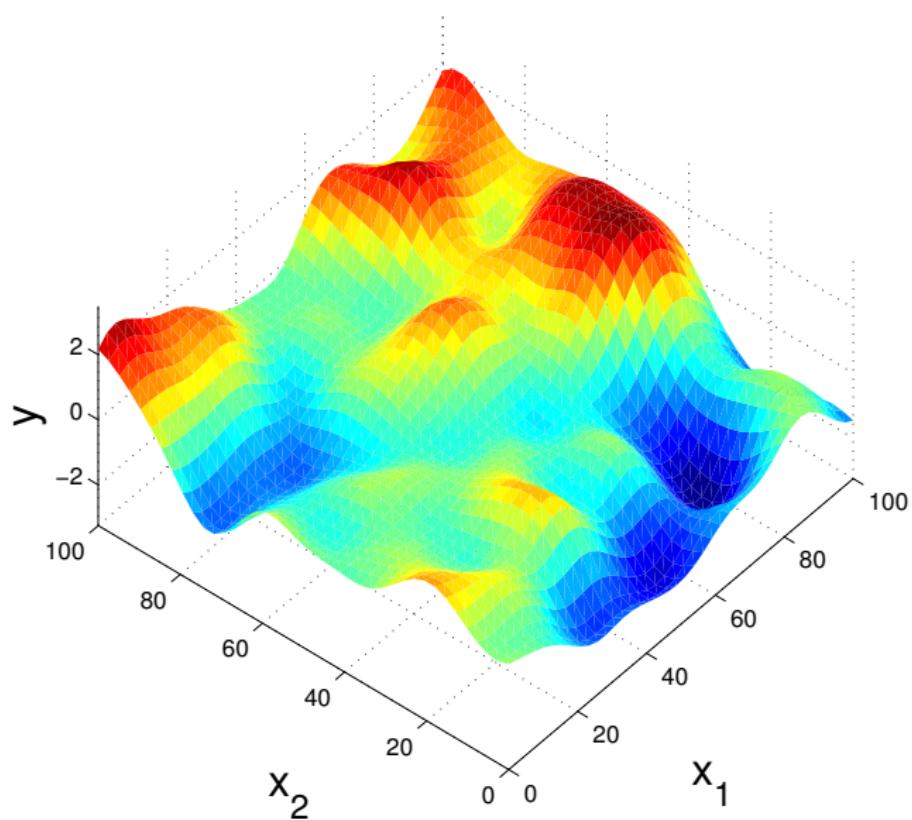
Higher dimensional input spaces



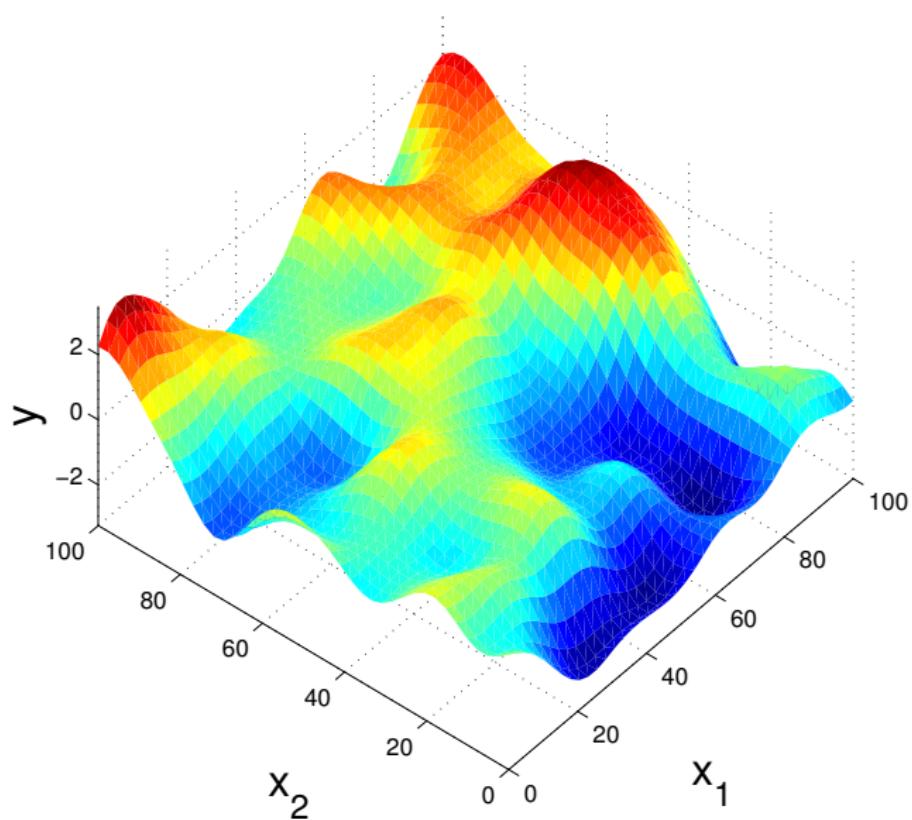
Higher dimensional input spaces



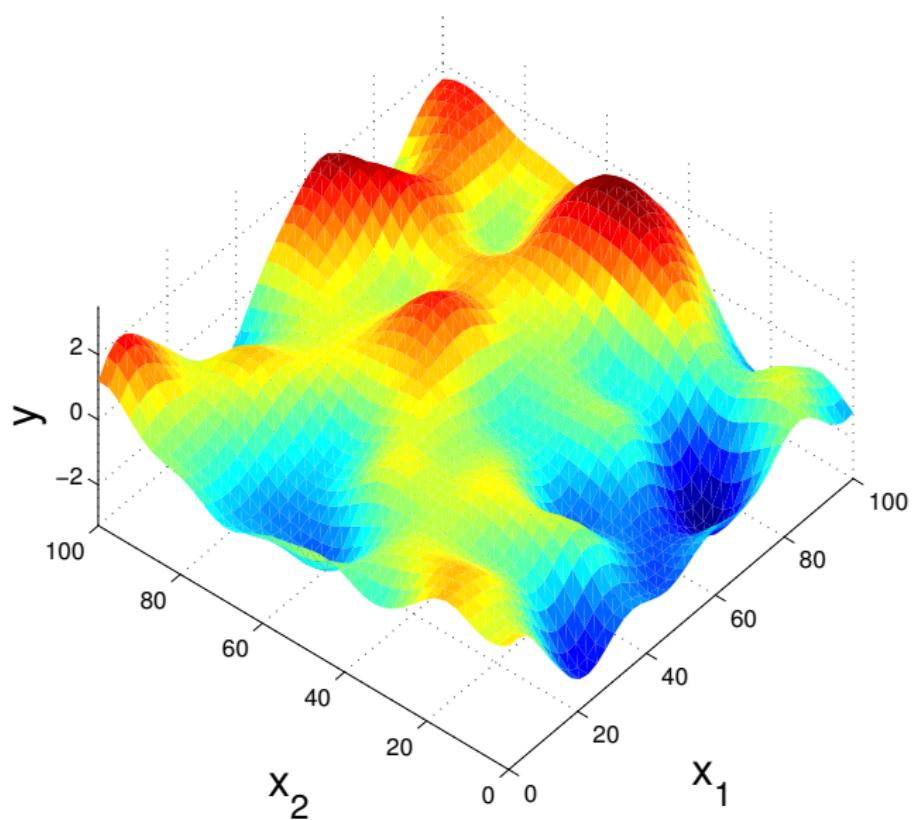
Higher dimensional input spaces



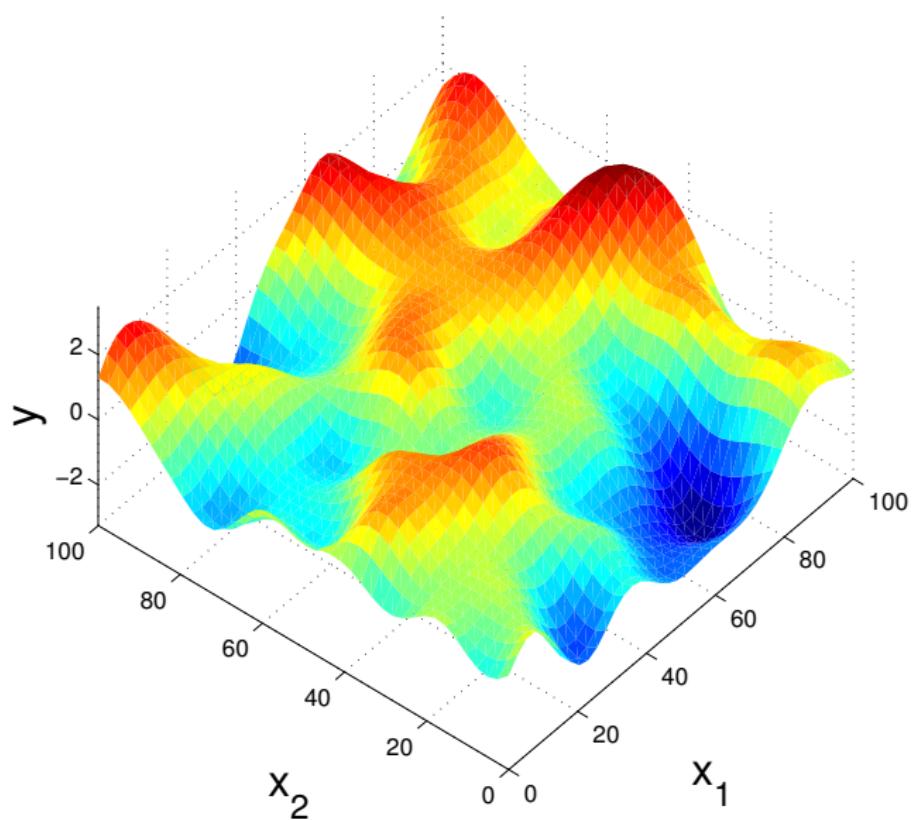
Higher dimensional input spaces



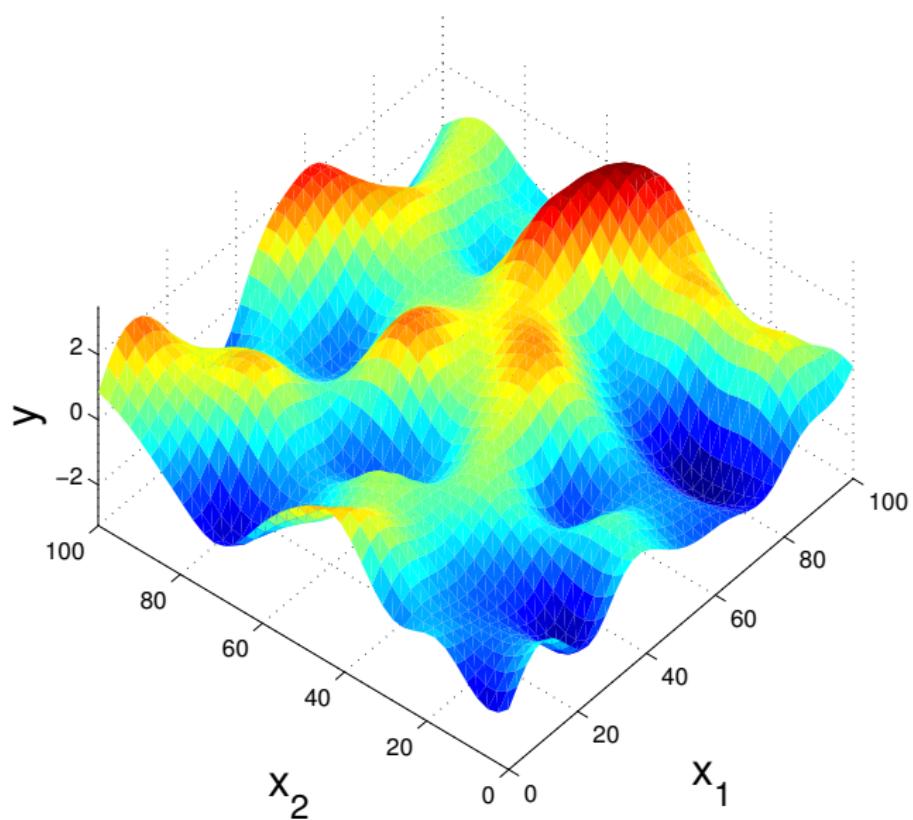
Higher dimensional input spaces



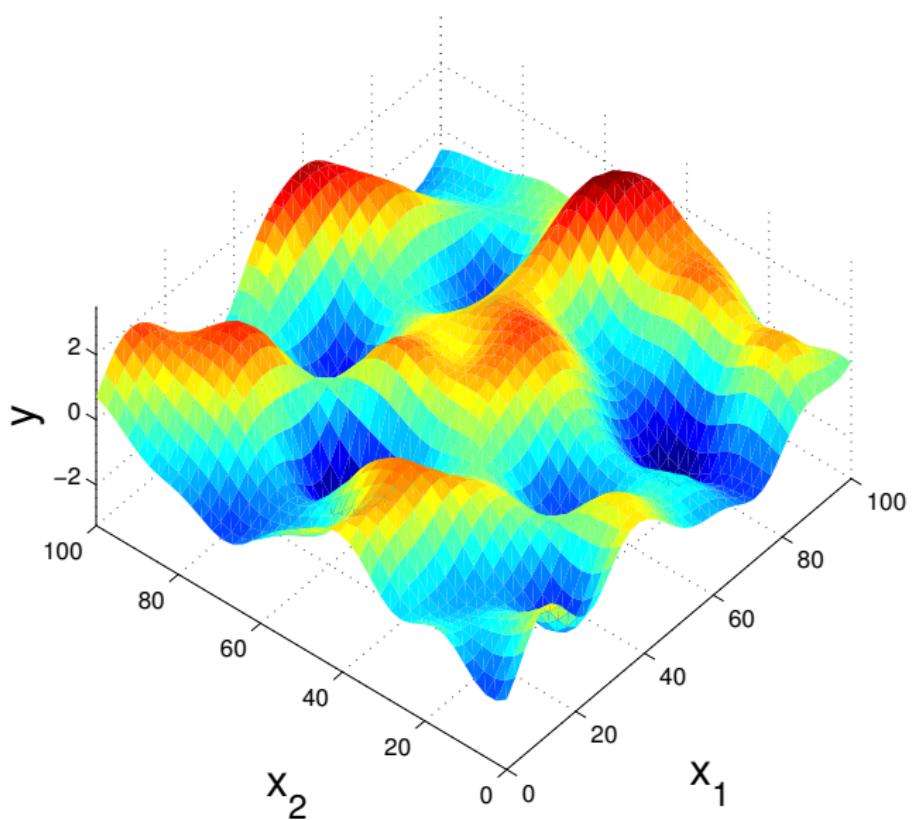
Higher dimensional input spaces



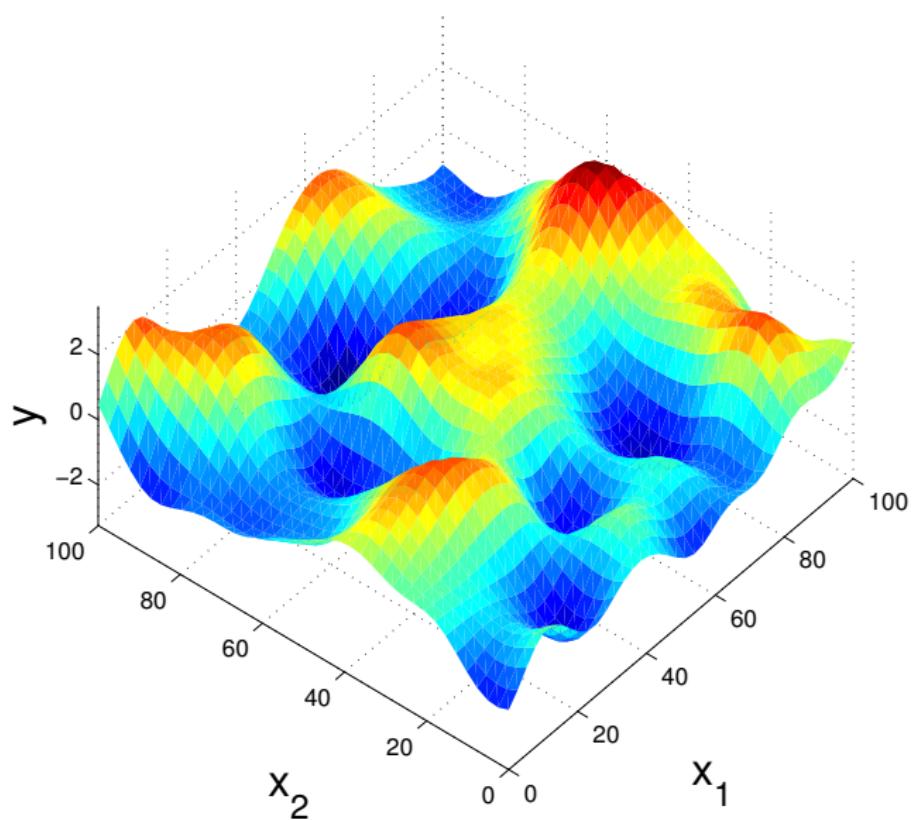
Higher dimensional input spaces



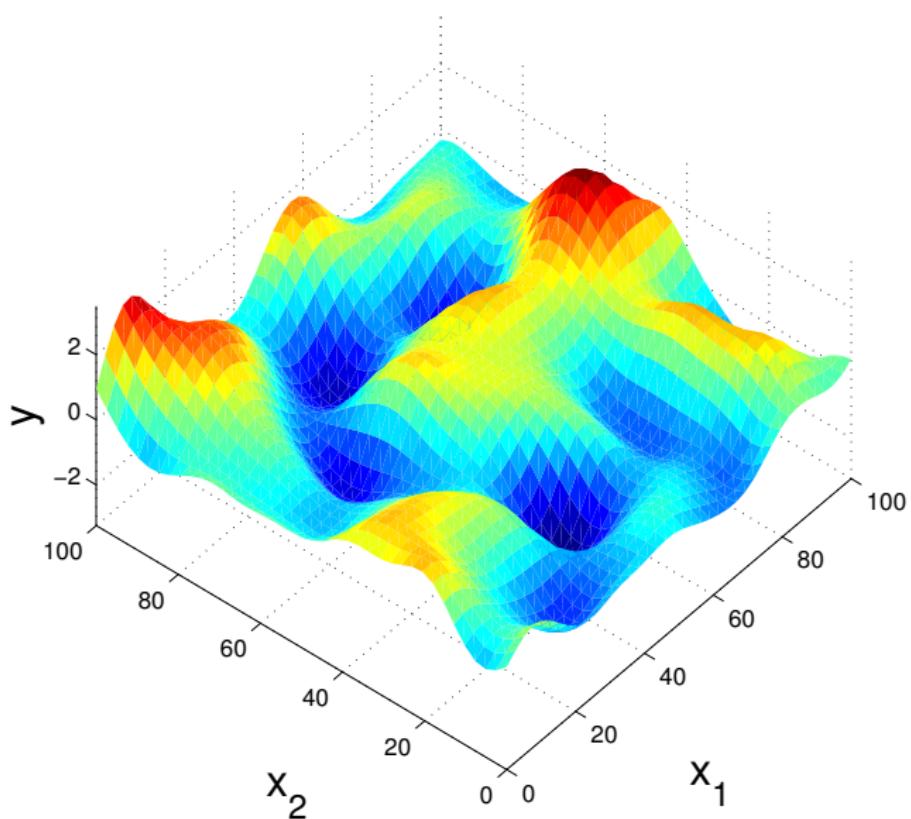
Higher dimensional input spaces



Higher dimensional input spaces



Higher dimensional input spaces



What effect do the hyper-parameters have?

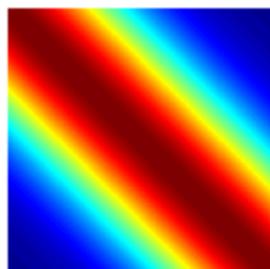
long horizontal length-scale

Non-parametric (∞ -parametric)

$$p(y|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 e^{-\frac{1}{2l^2}(x_1 - x_2)^2}$$

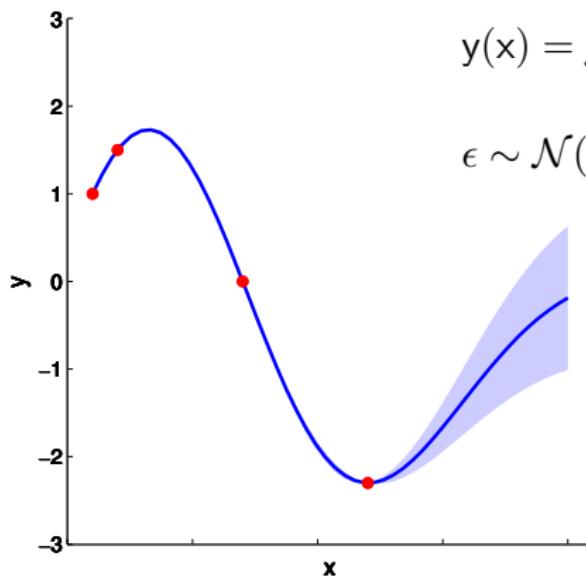


$$\Sigma =$$

Parametric model

$$y(x) = f(x; \theta) + \sigma_y \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$



What effect does the form of the covariance function have?

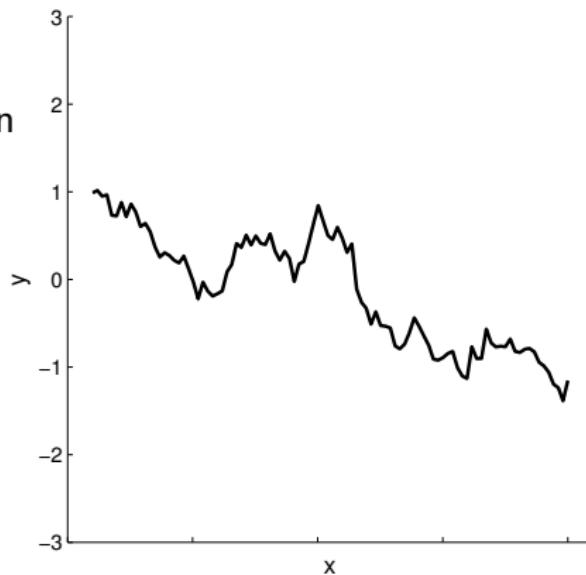
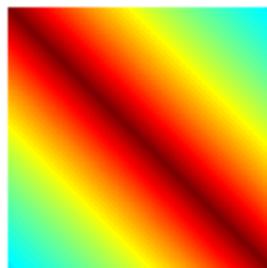
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

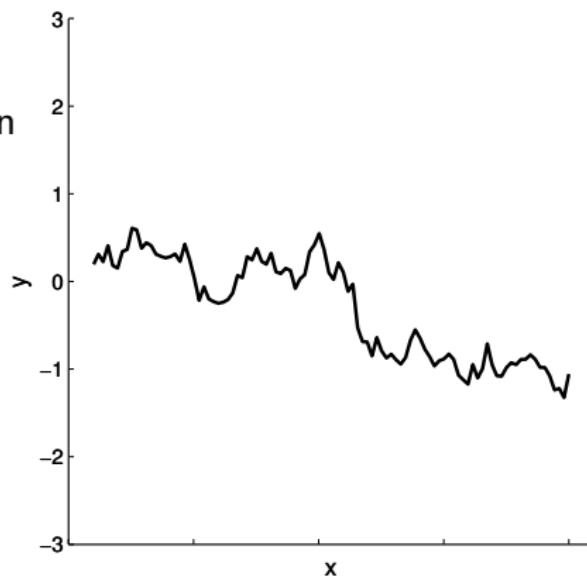
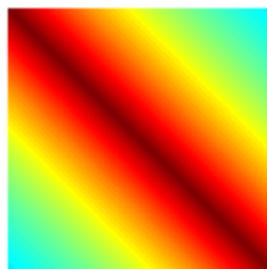
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

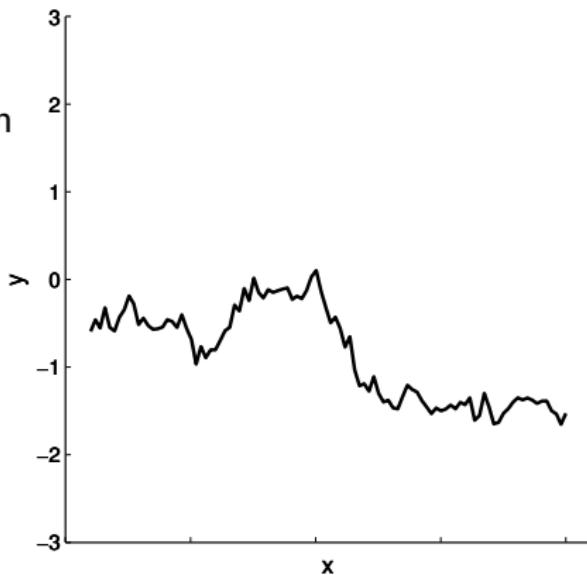
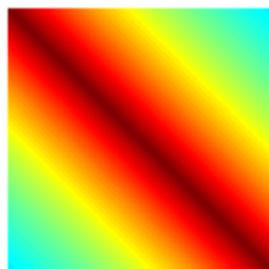
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

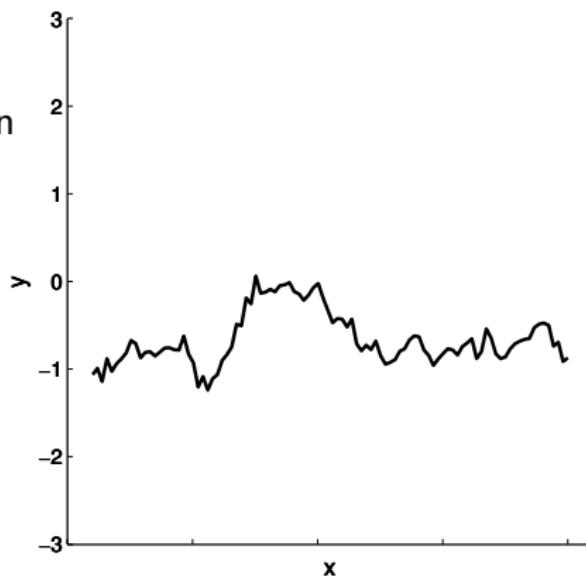
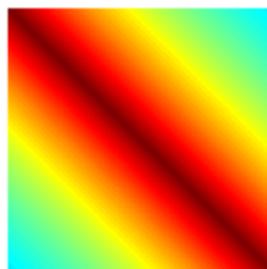
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

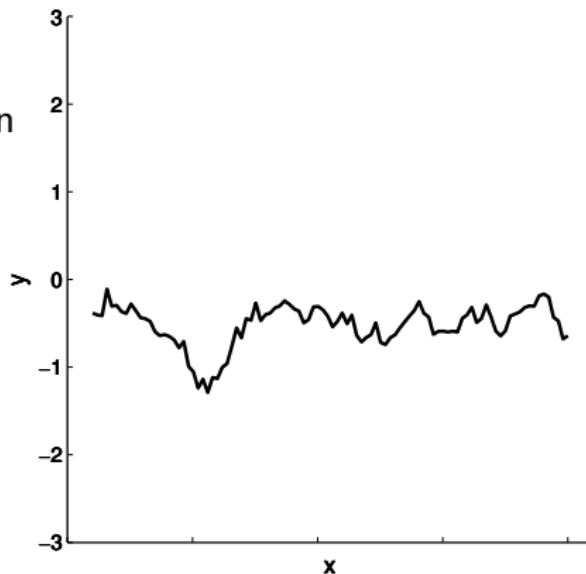
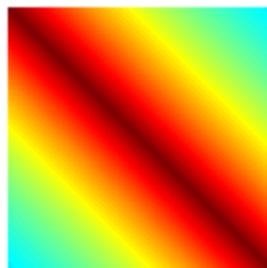
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

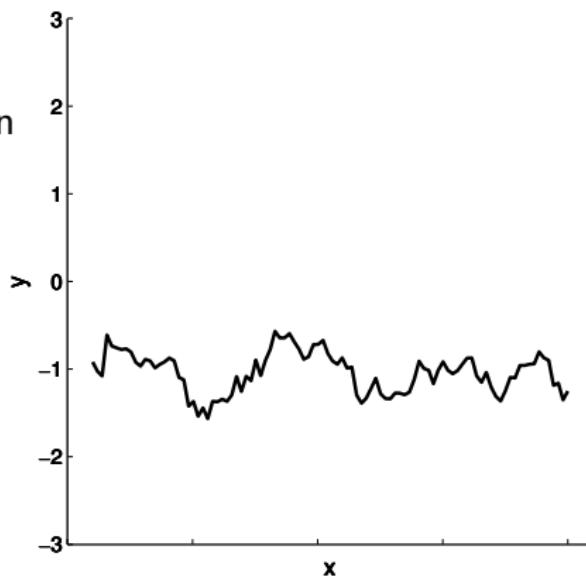
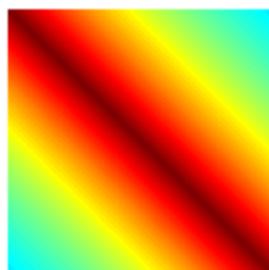
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

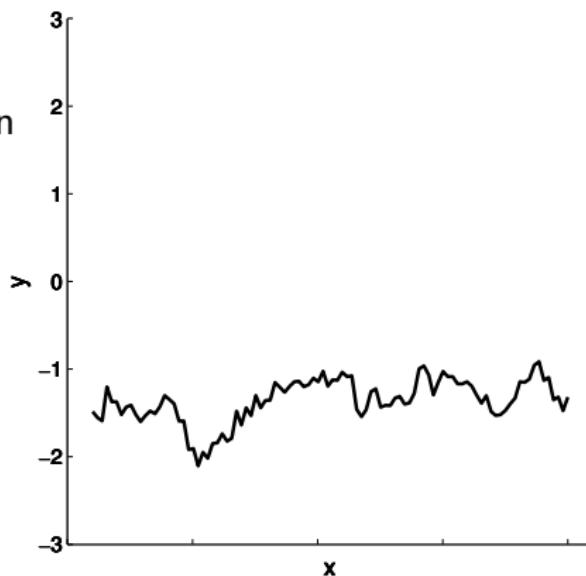
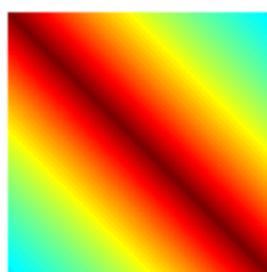
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

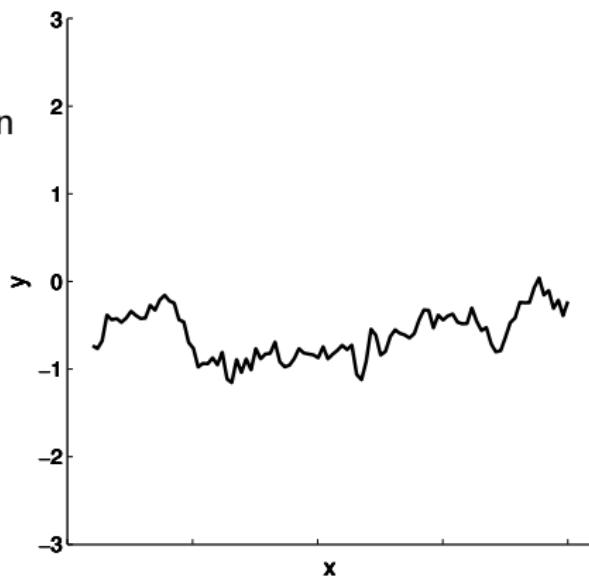
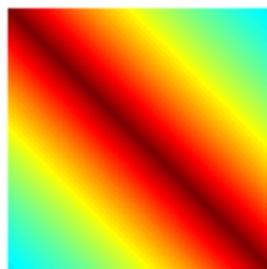
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

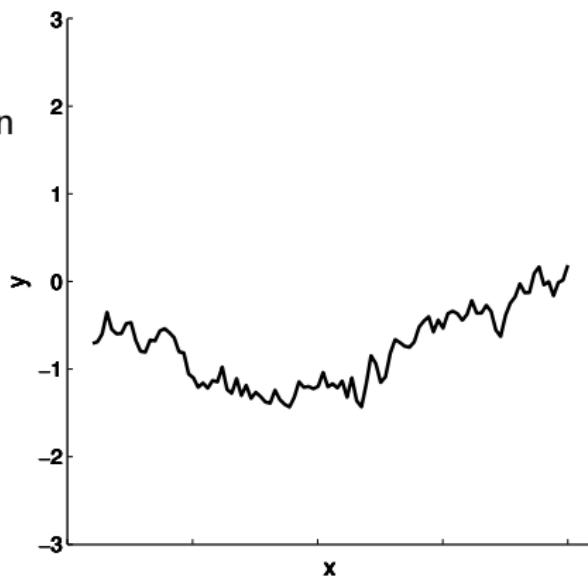
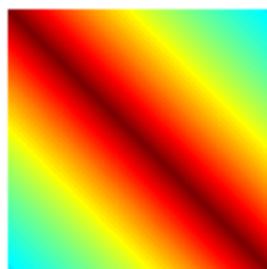
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

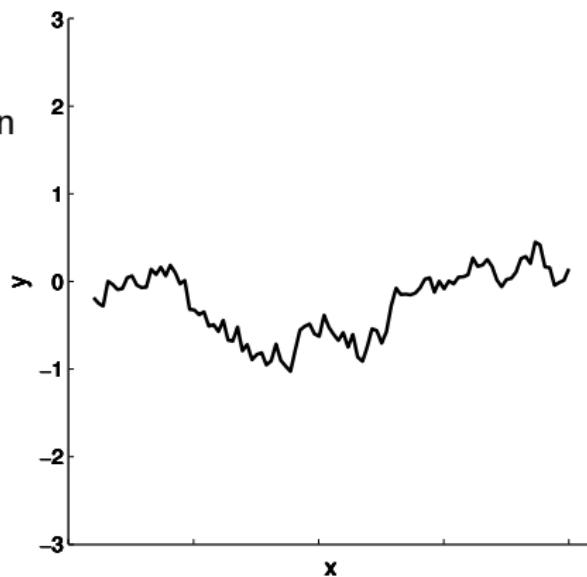
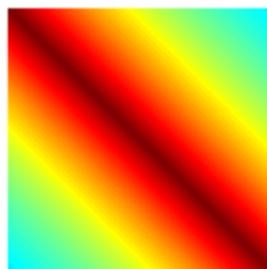
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

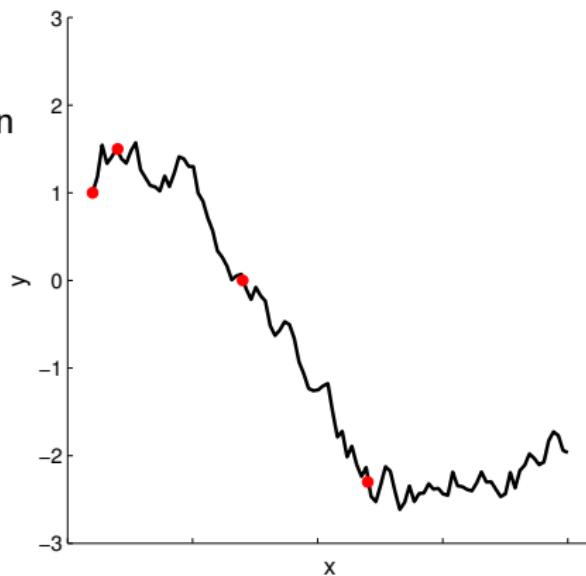
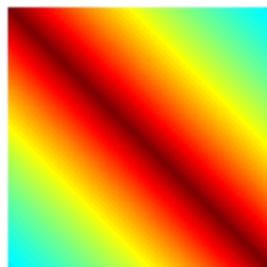
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

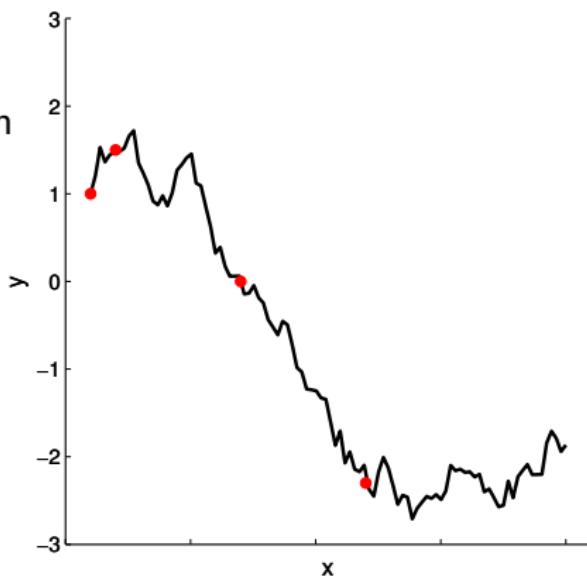
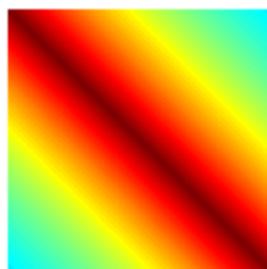
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

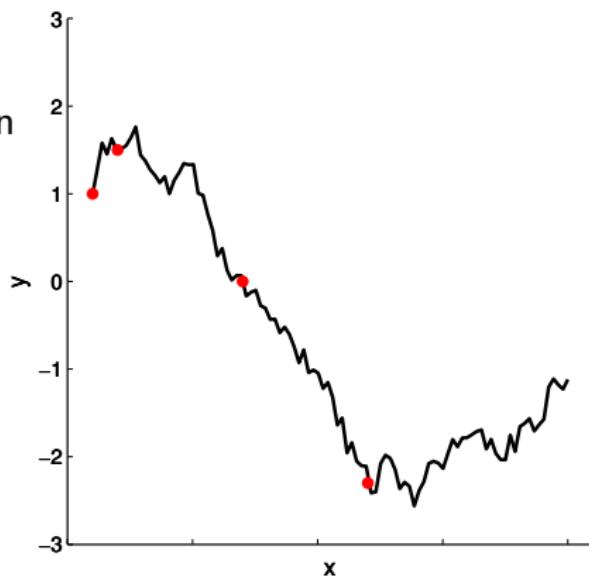
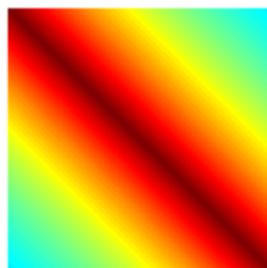
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

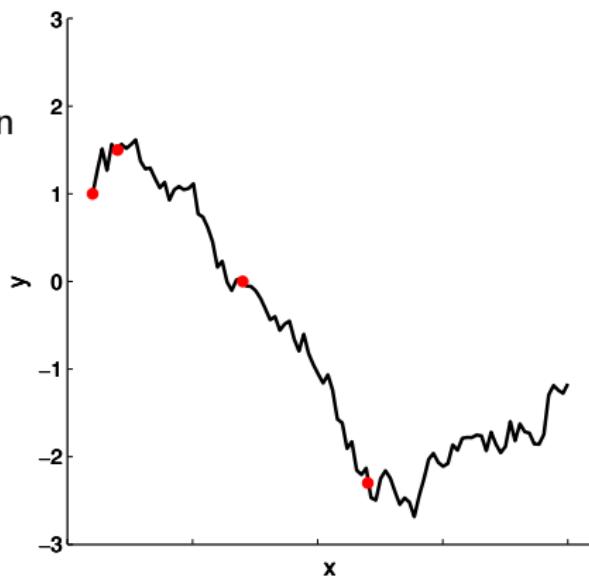
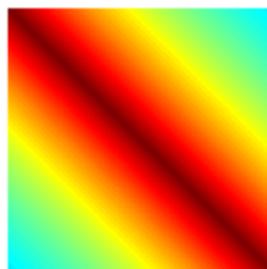
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

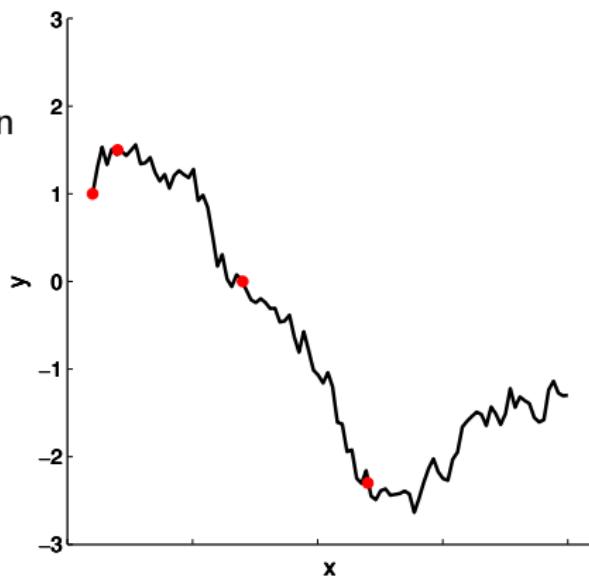
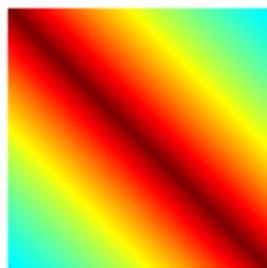
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

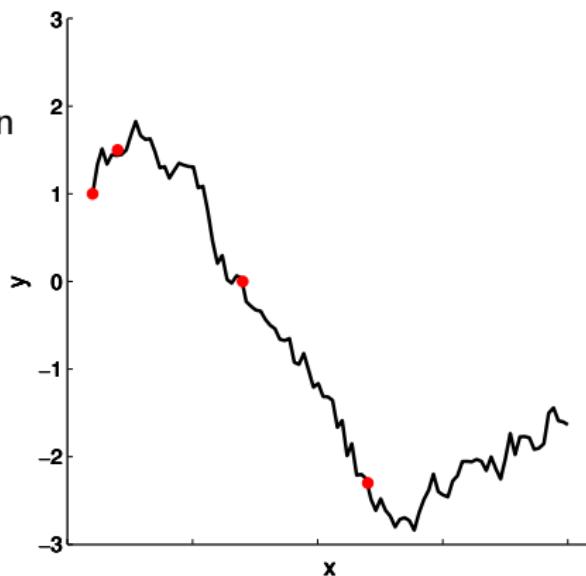
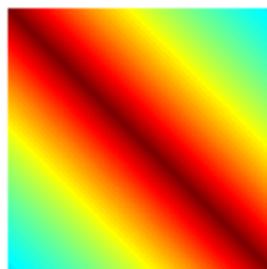
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

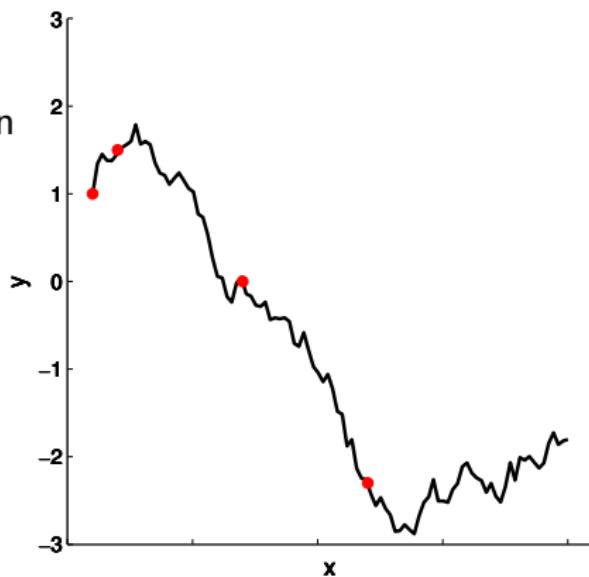
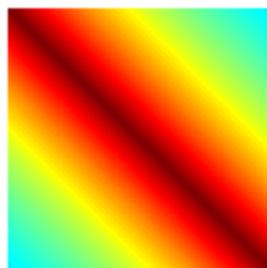
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

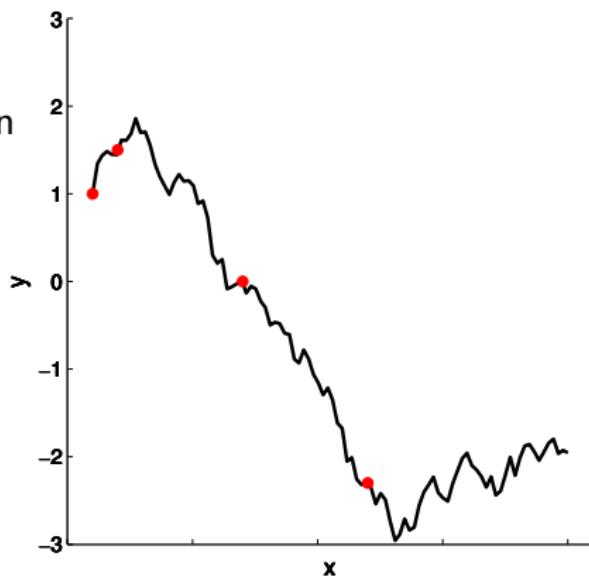
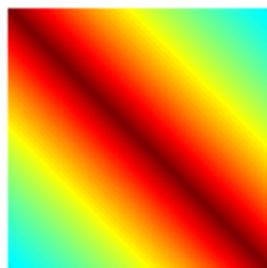
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

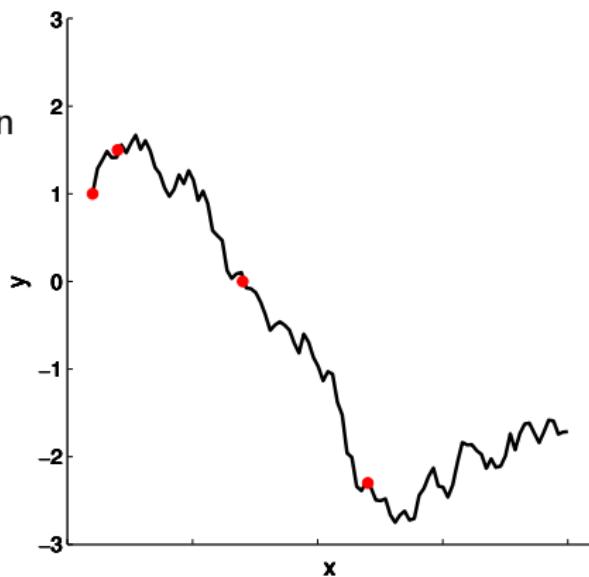
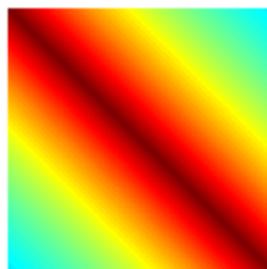
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

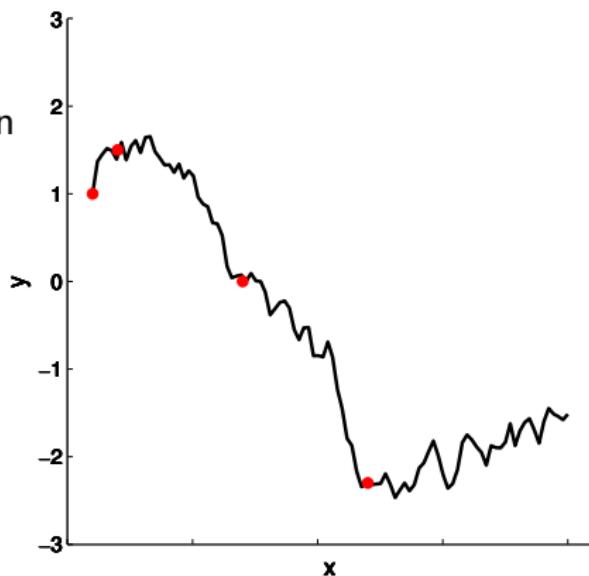
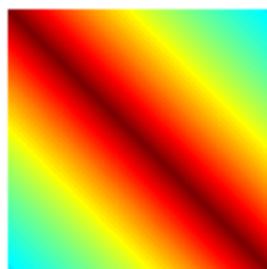
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

$\Sigma =$



What effect does the form of the covariance function have?

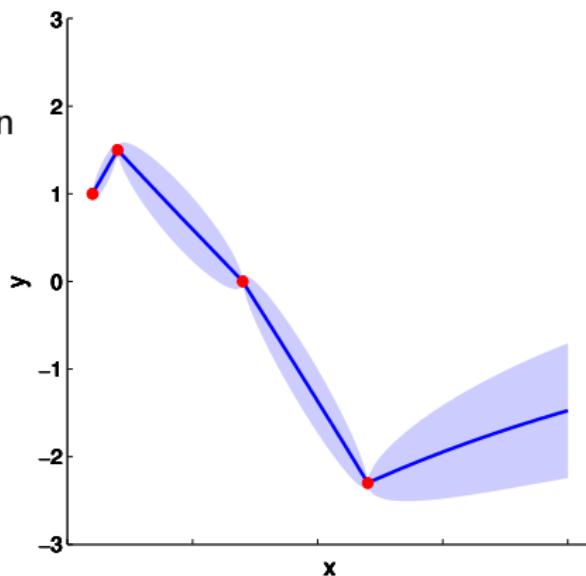
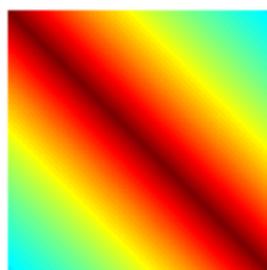
$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2} |x_1 - x_2|\right)$$

Laplacian covariance function

Browninan motion

Ornstein-Uhlenbeck

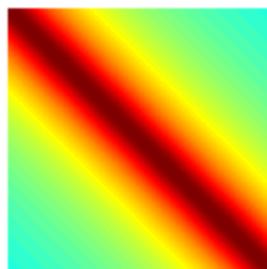
$\Sigma =$



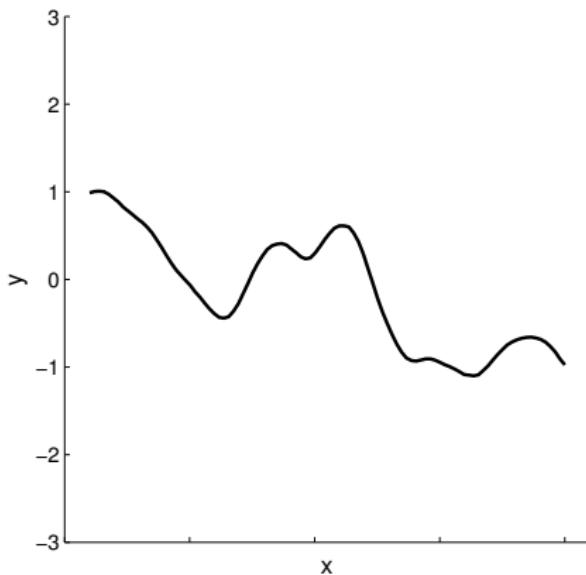
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



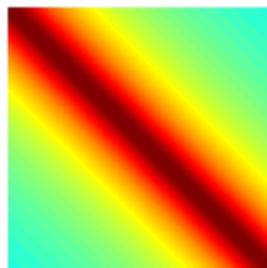
$\Sigma =$



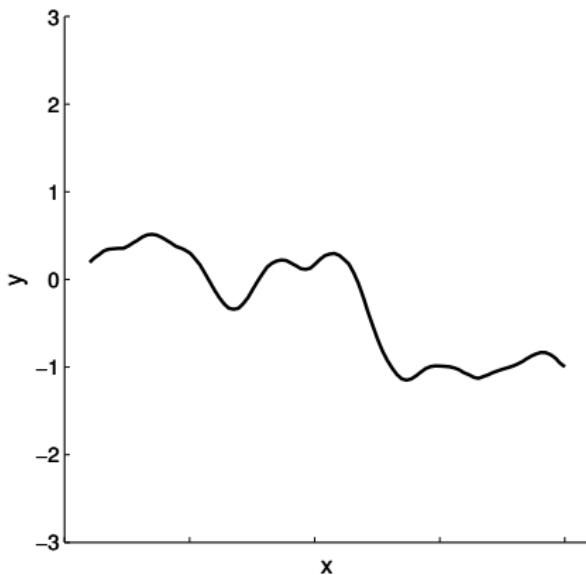
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



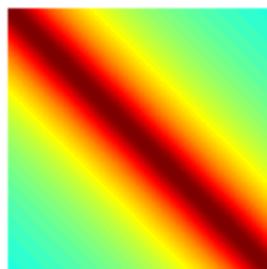
$\Sigma =$



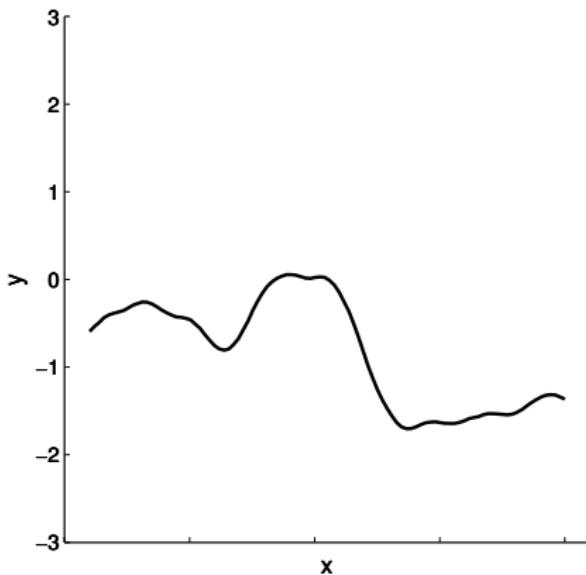
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



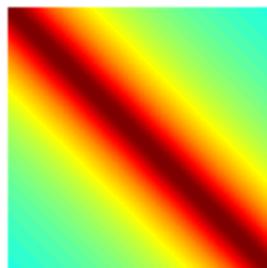
$\Sigma =$



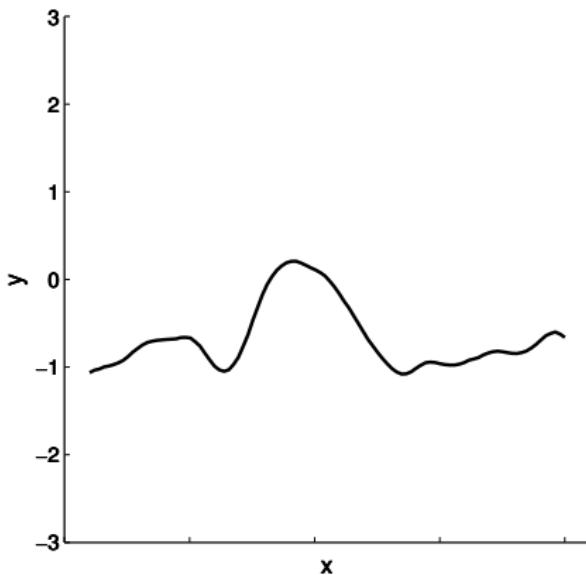
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



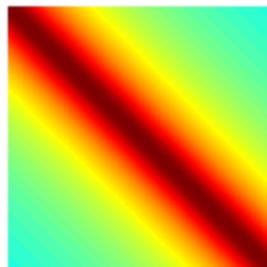
$\Sigma =$



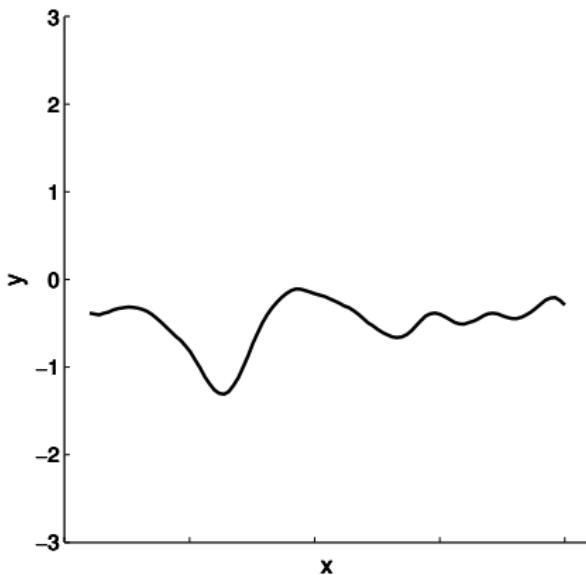
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



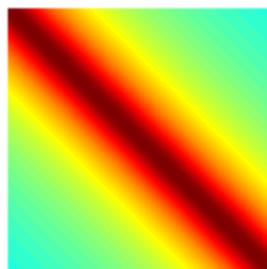
$\Sigma =$



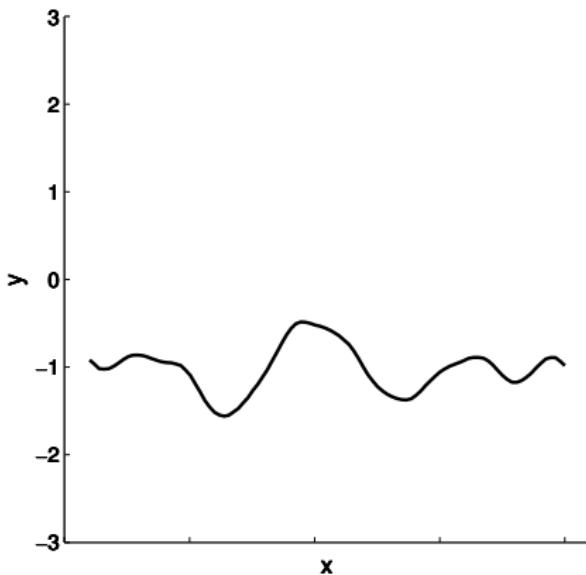
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



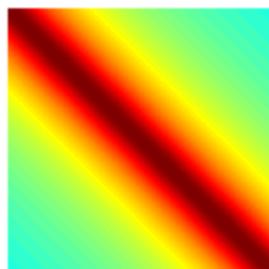
$\Sigma =$



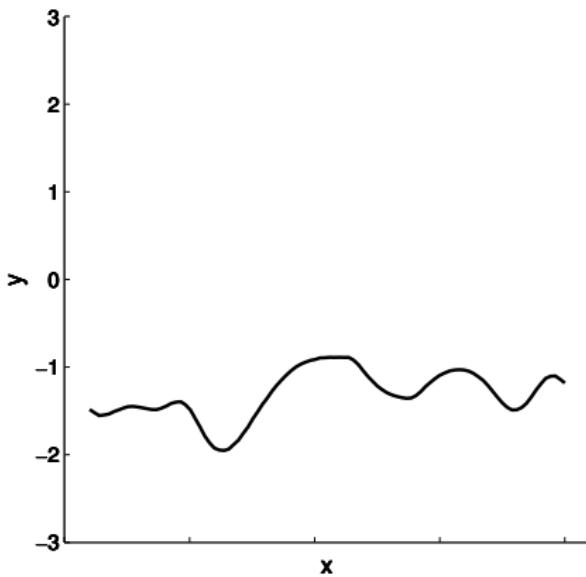
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



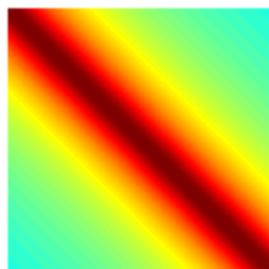
$\Sigma =$



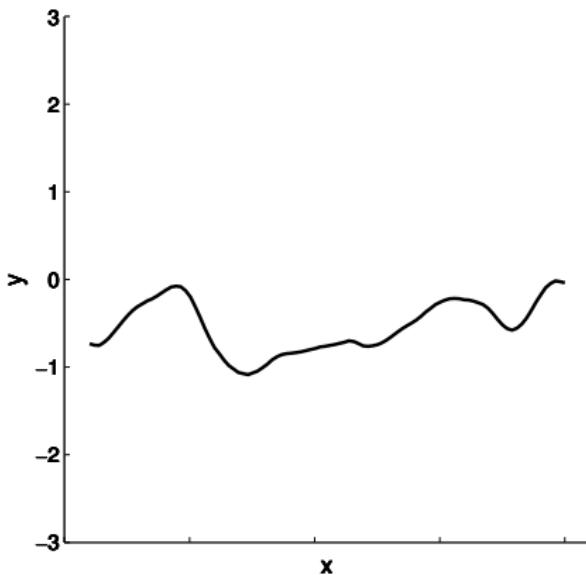
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



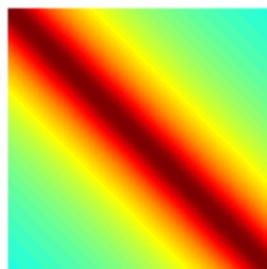
$\Sigma =$



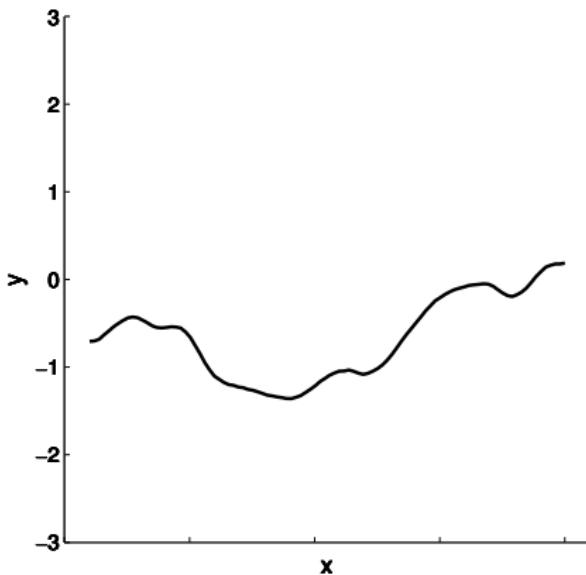
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



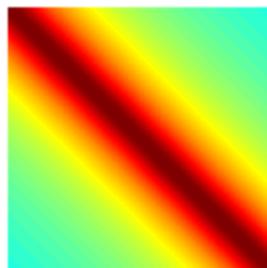
$\Sigma =$



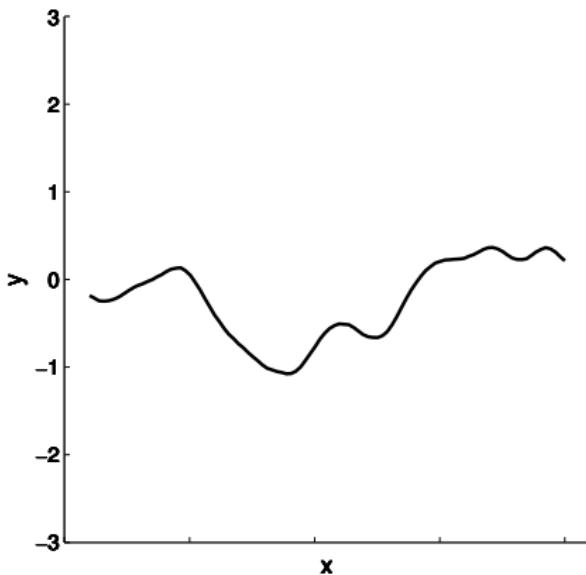
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



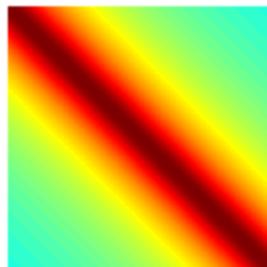
$\Sigma =$



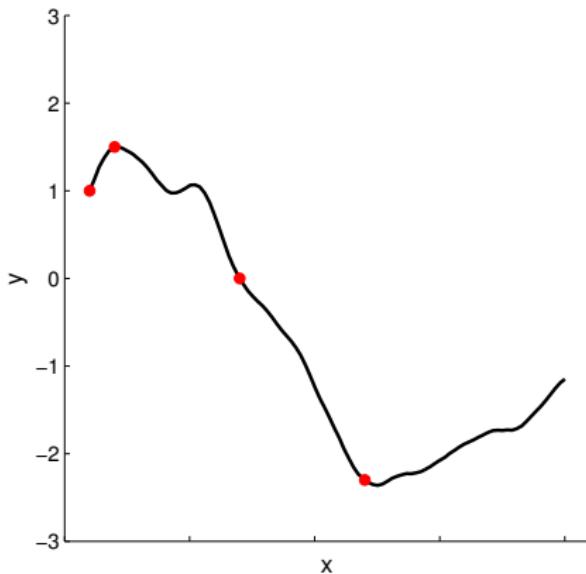
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



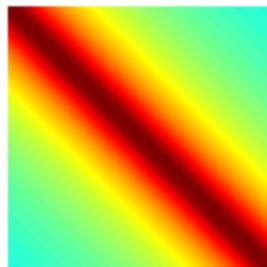
$\Sigma =$



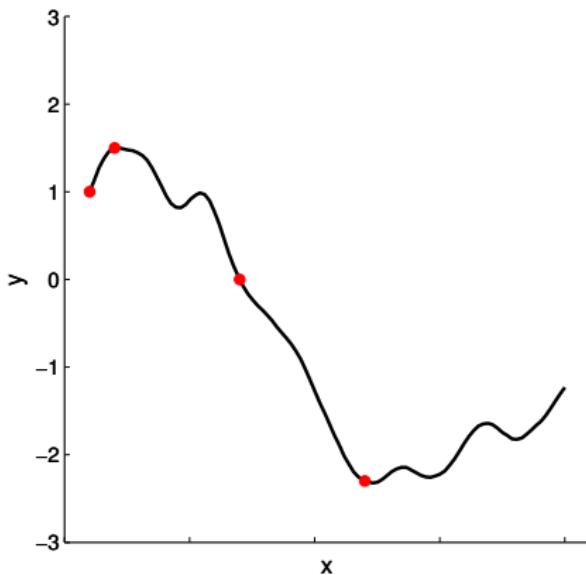
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



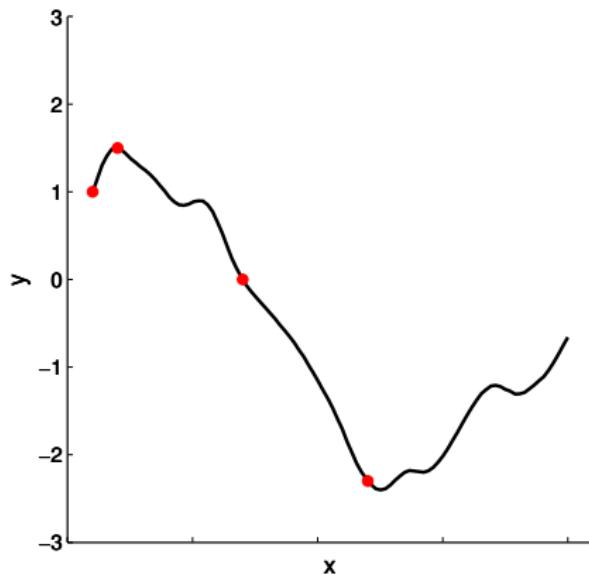
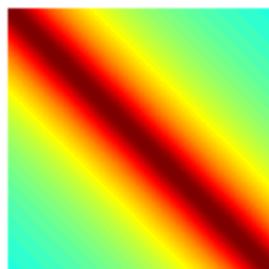
$\Sigma =$



What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

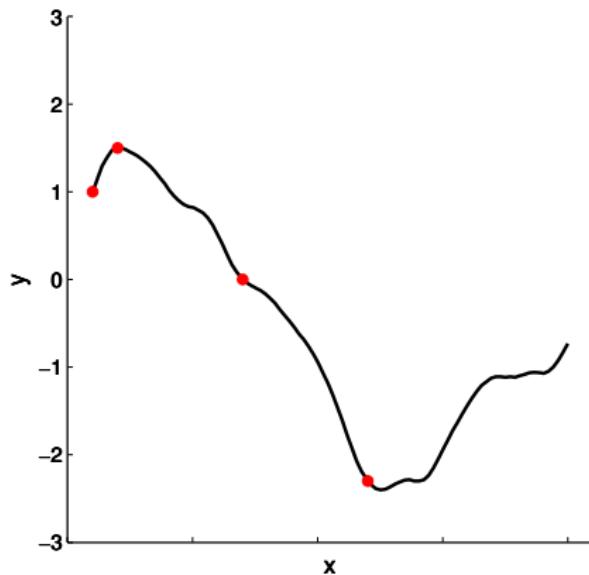
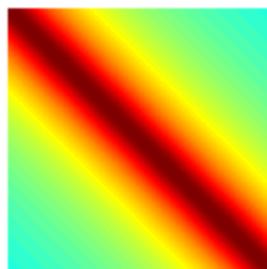
Rational Quadratic



What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

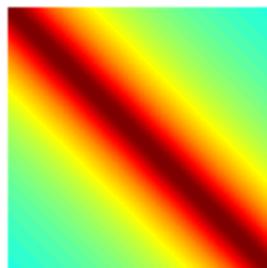
Rational Quadratic



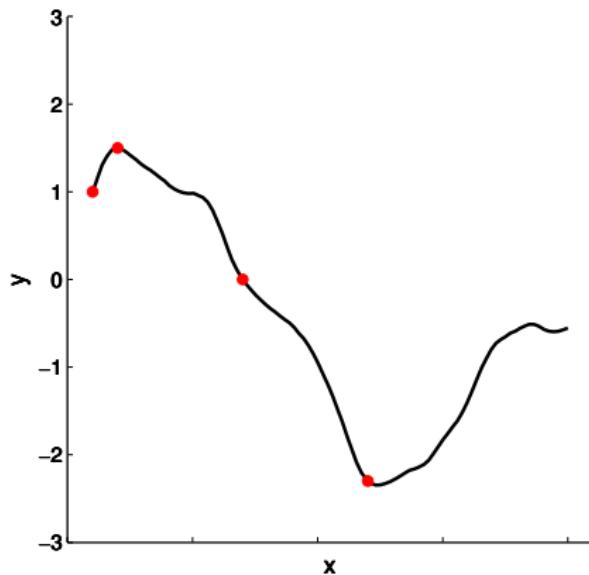
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



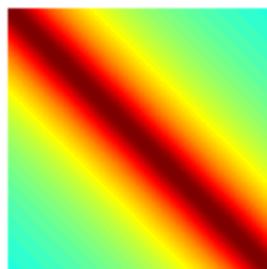
$\Sigma =$



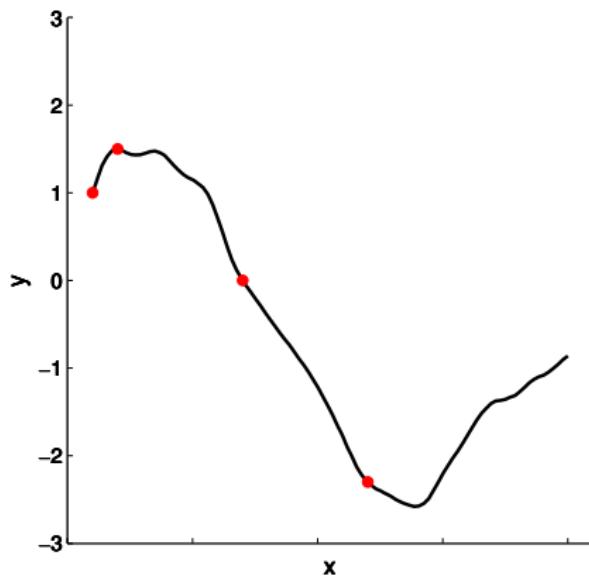
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



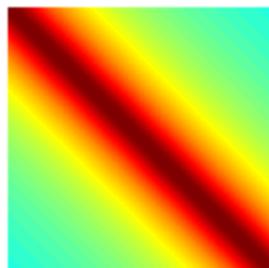
$\Sigma =$



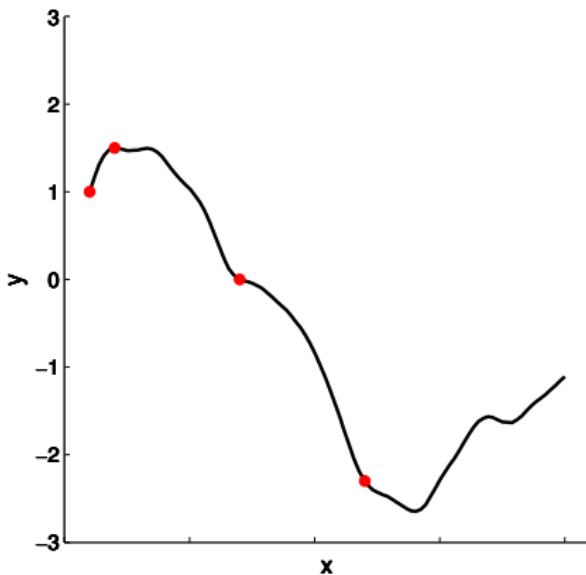
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



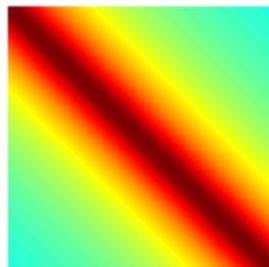
$\Sigma =$



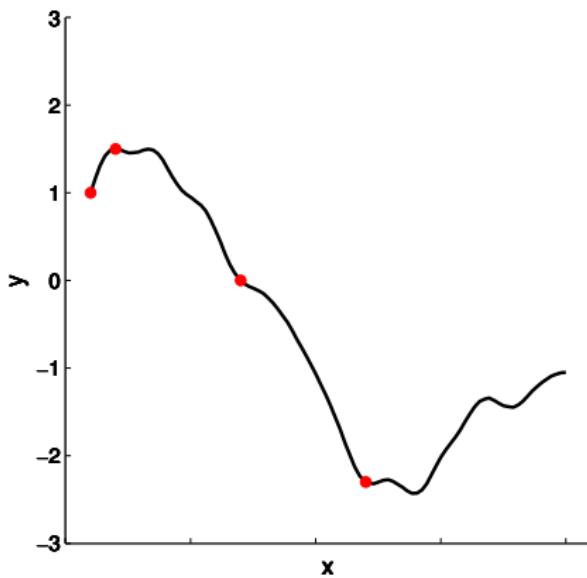
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



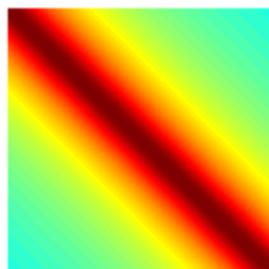
$\Sigma =$



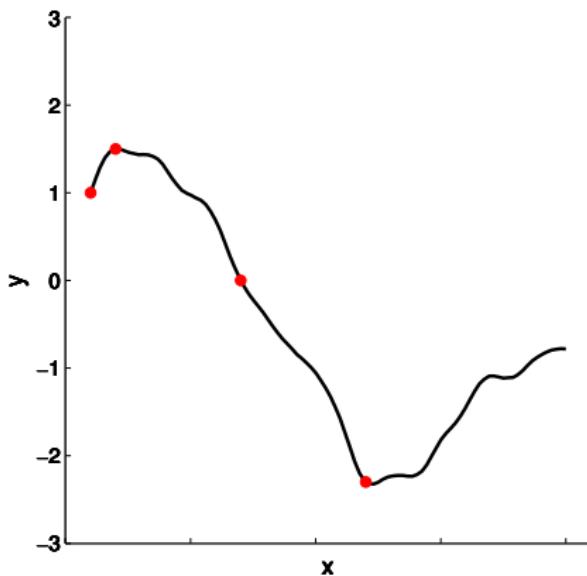
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



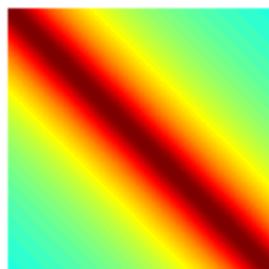
$\Sigma =$



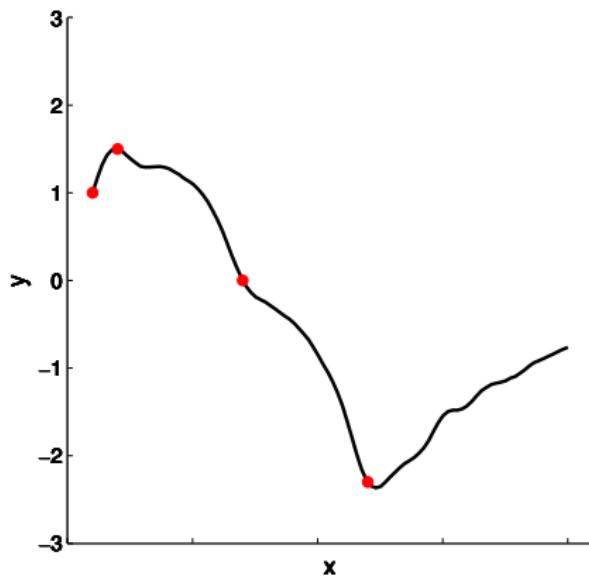
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



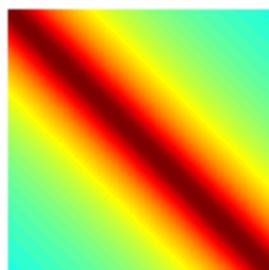
$\Sigma =$



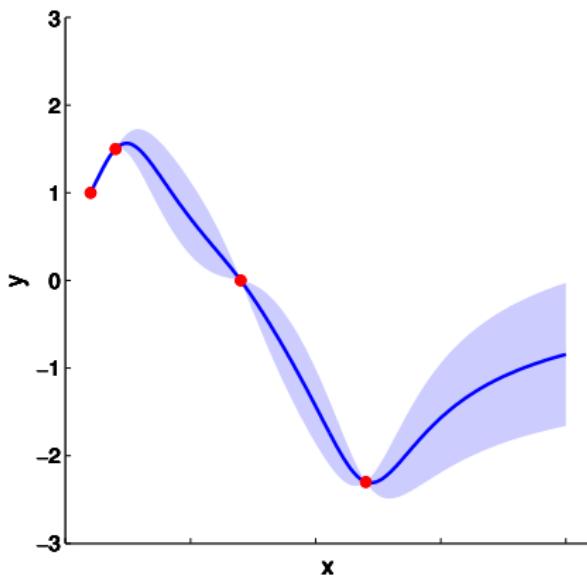
What effect does the form of the covariance function have?

$$K(x_1, x_2) = \sigma^2 \left(1 + \frac{1}{2\alpha l^2} |x_1 - x_2|\right)^{-\alpha}$$

Rational Quadratic



$\Sigma =$

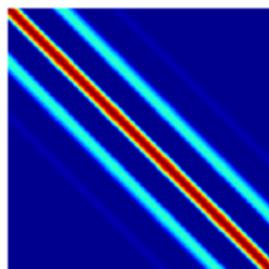


What effect does the form of the covariance function have?

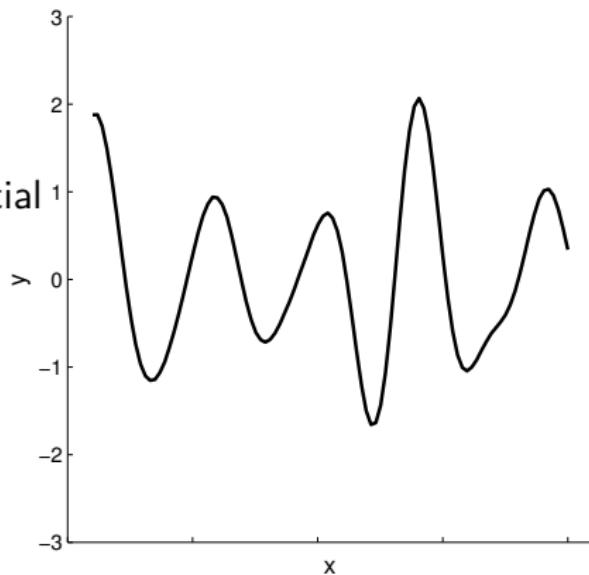
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

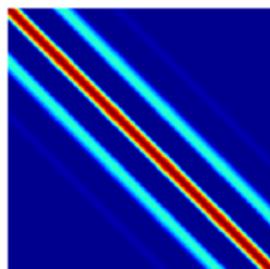


What effect does the form of the covariance function have?

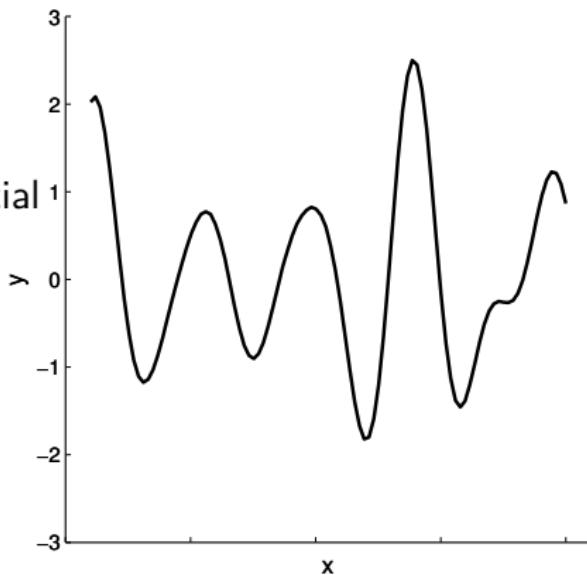
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

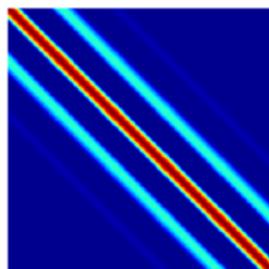


What effect does the form of the covariance function have?

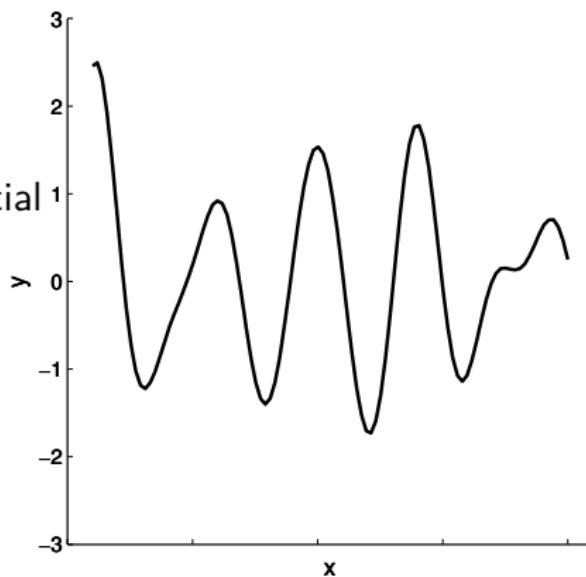
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

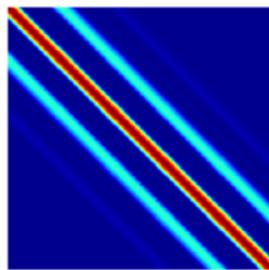


What effect does the form of the covariance function have?

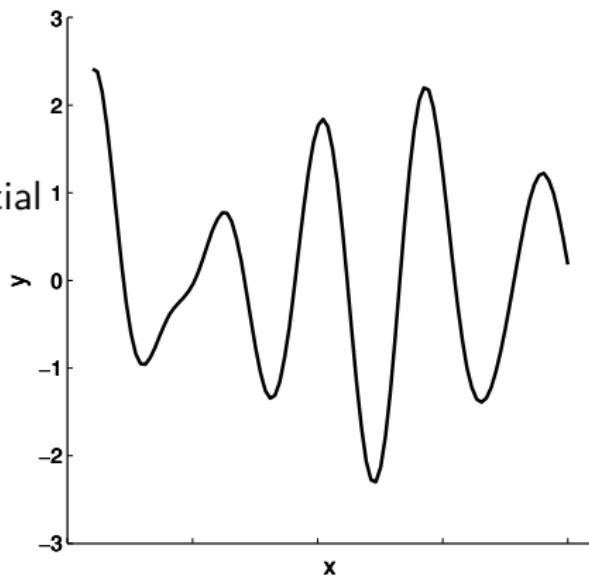
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

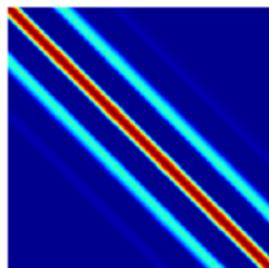


What effect does the form of the covariance function have?

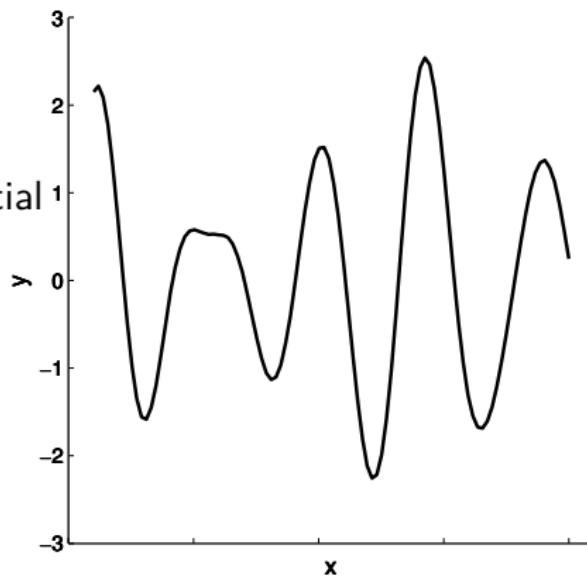
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

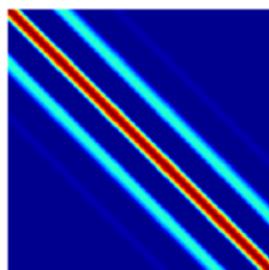


What effect does the form of the covariance function have?

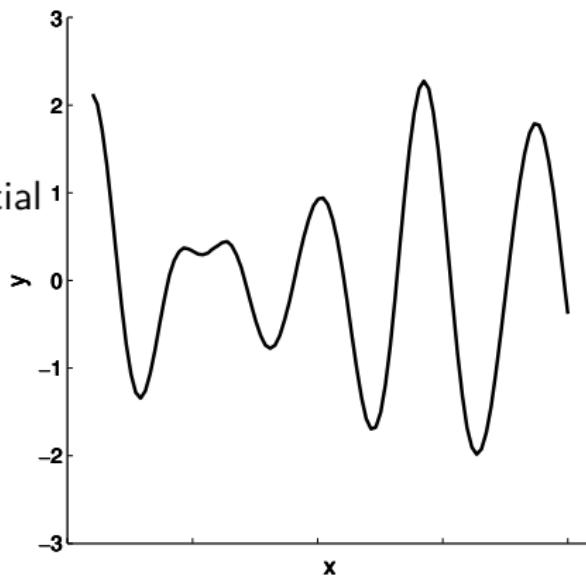
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

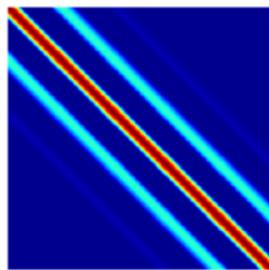


What effect does the form of the covariance function have?

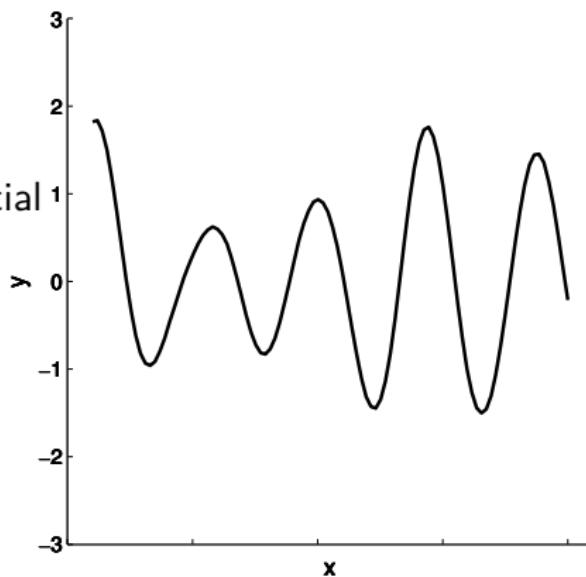
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

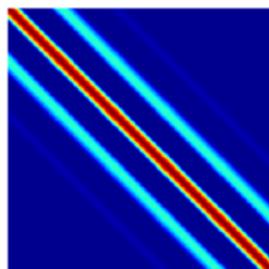


What effect does the form of the covariance function have?

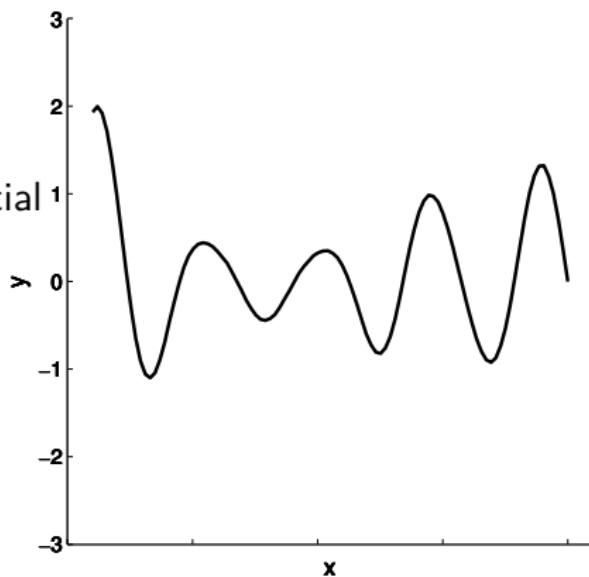
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

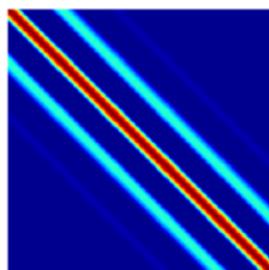


What effect does the form of the covariance function have?

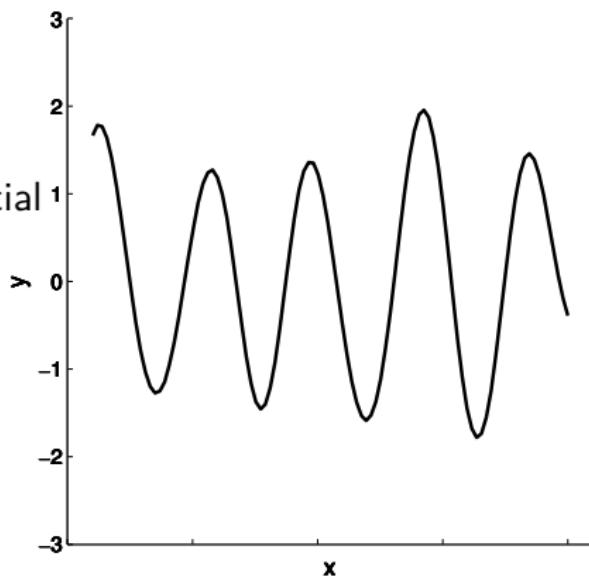
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

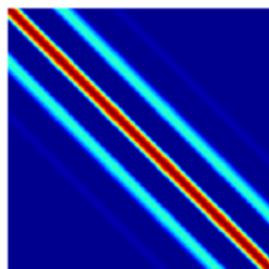


What effect does the form of the covariance function have?

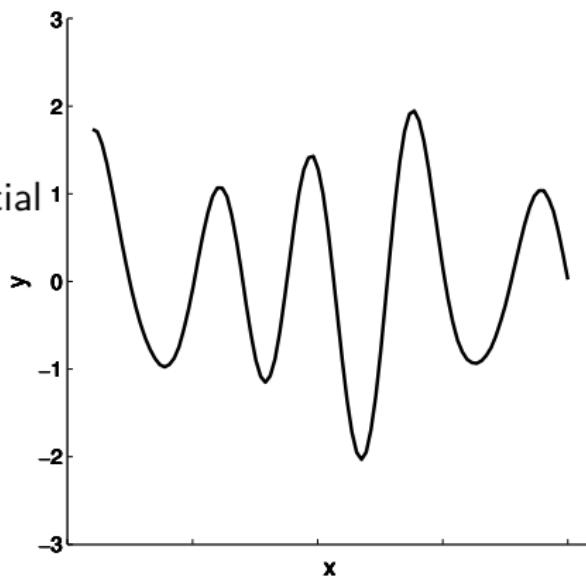
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

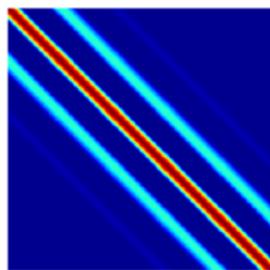


What effect does the form of the covariance function have?

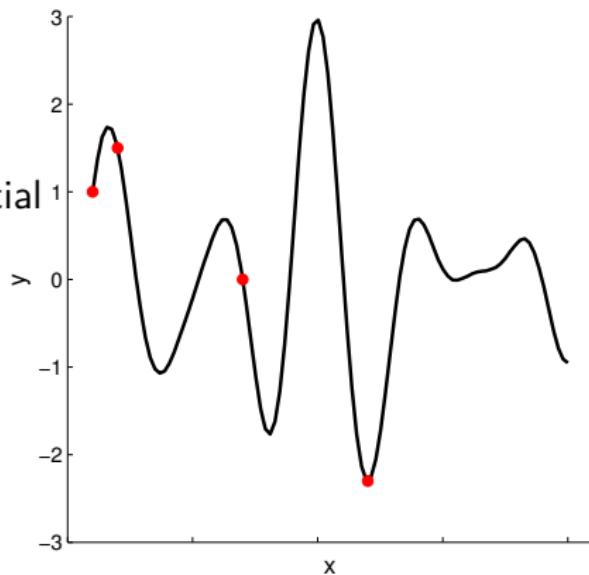
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

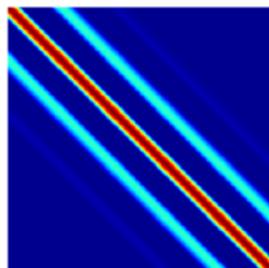


What effect does the form of the covariance function have?

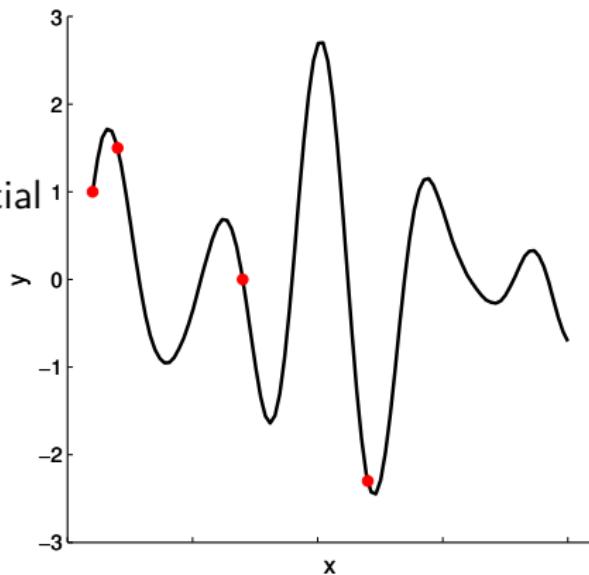
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

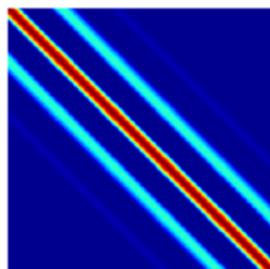


What effect does the form of the covariance function have?

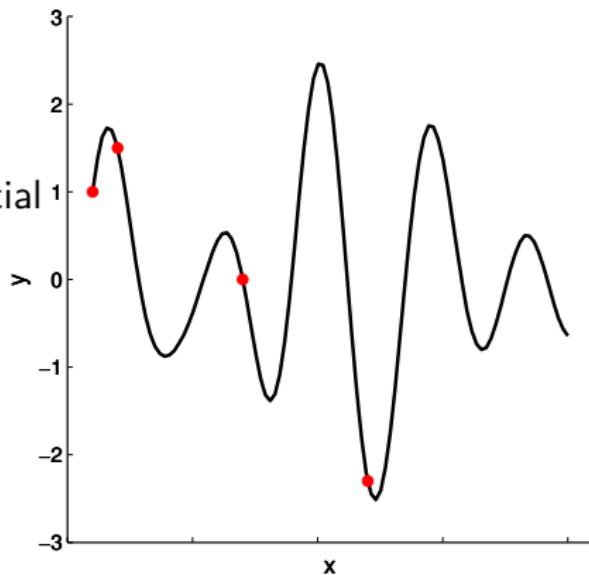
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

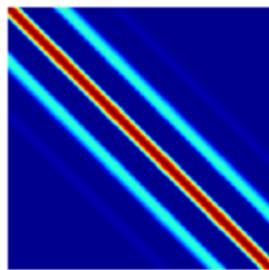


What effect does the form of the covariance function have?

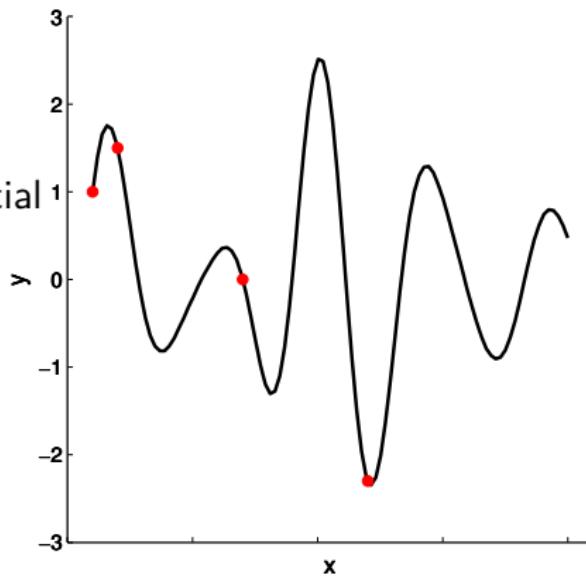
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

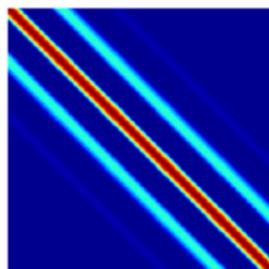


What effect does the form of the covariance function have?

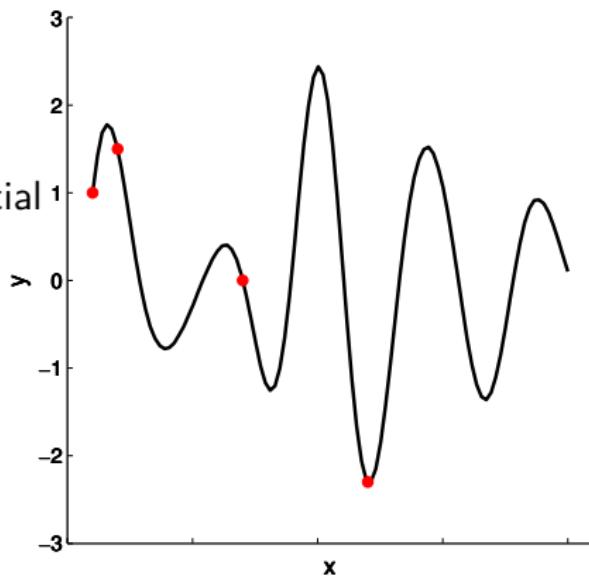
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

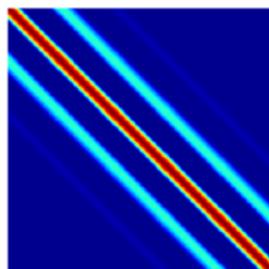


What effect does the form of the covariance function have?

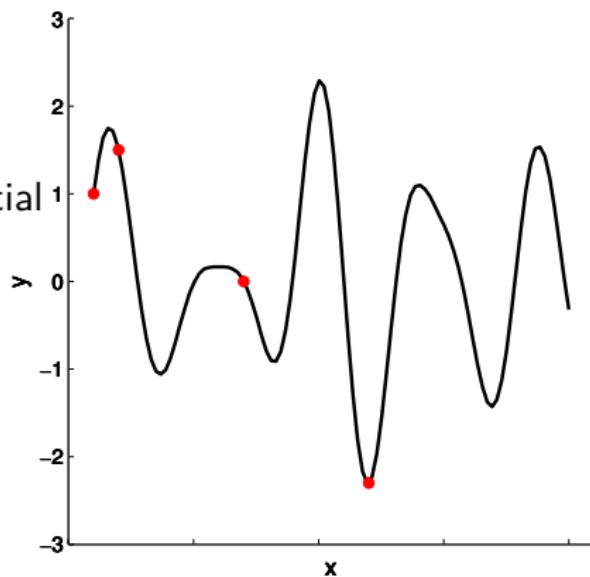
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

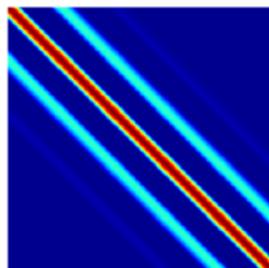


What effect does the form of the covariance function have?

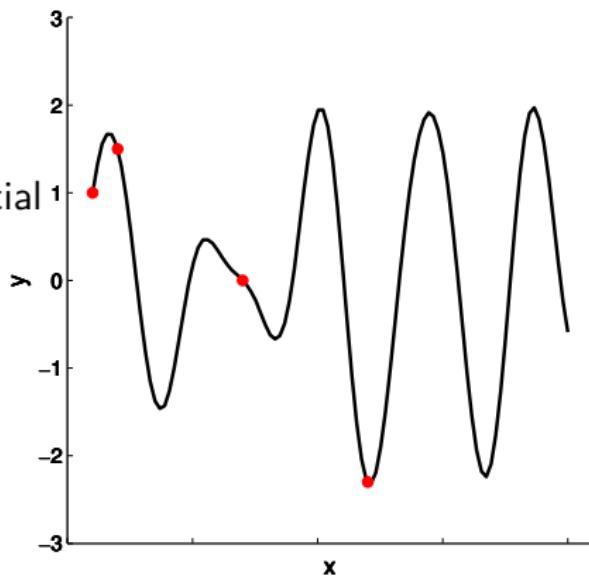
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

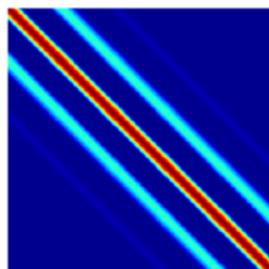


What effect does the form of the covariance function have?

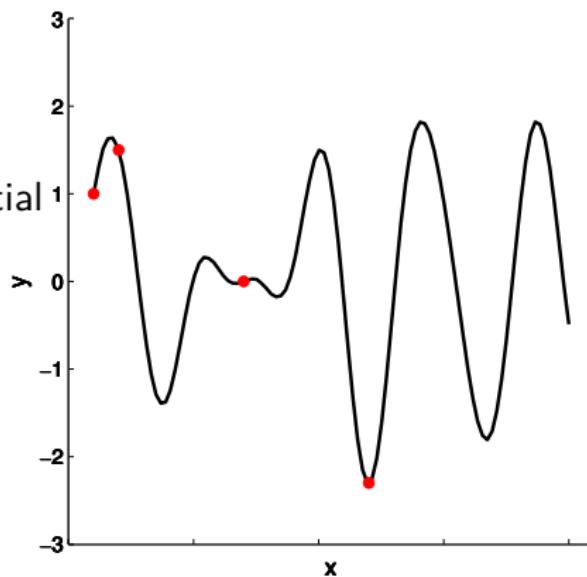
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

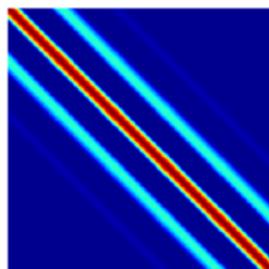


What effect does the form of the covariance function have?

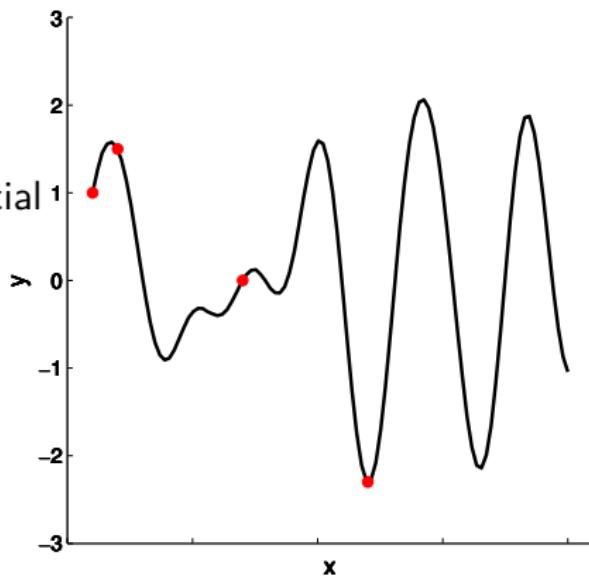
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

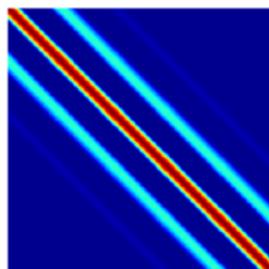


What effect does the form of the covariance function have?

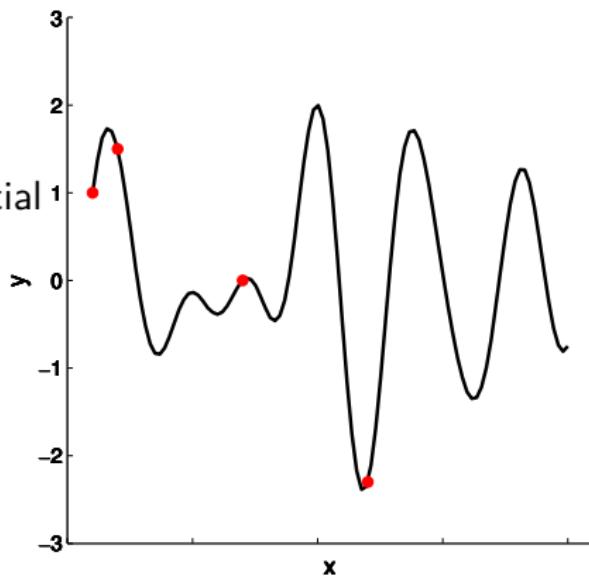
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential



$\Sigma =$

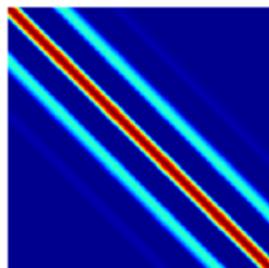


What effect does the form of the covariance function have?

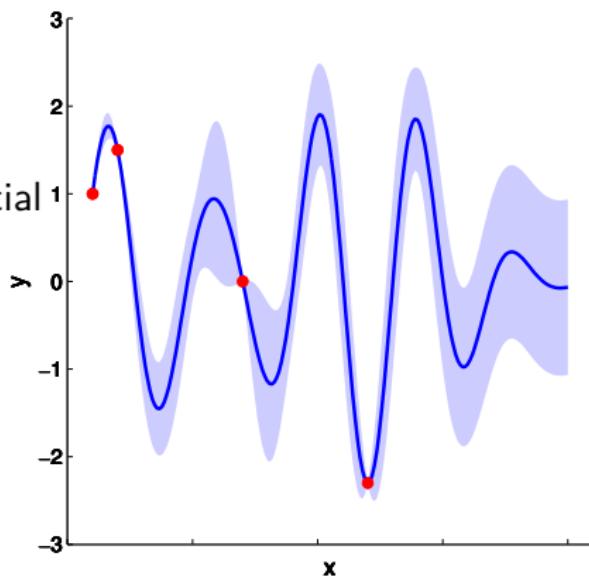
$$K(x_1, x_2) = \sigma^2 \cos(\omega(x_1 - x_2)) \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

Periodic

sinusoid \times squared exponential

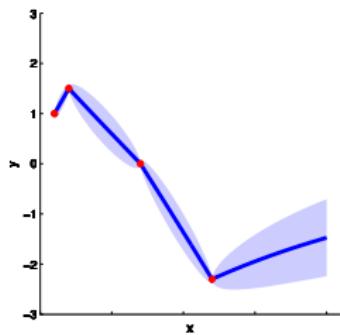


$\Sigma =$

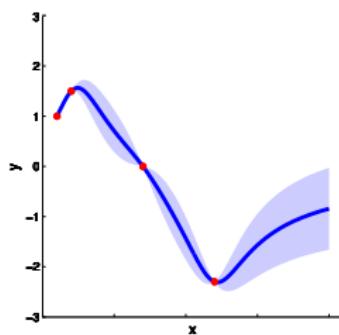


The covariance function has a large effect

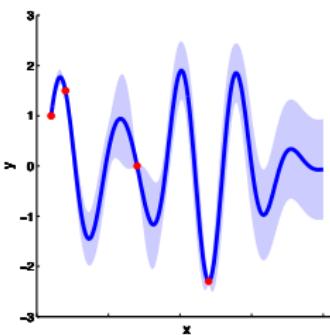
OU



RQ

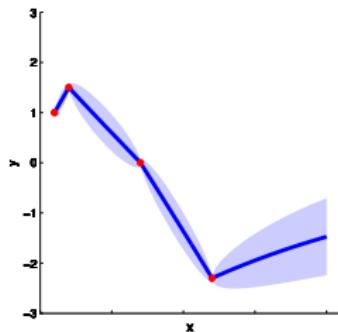


periodic

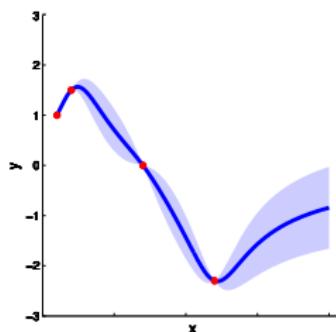


The covariance function has a large effect

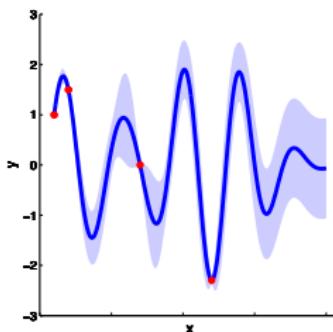
OU



RQ



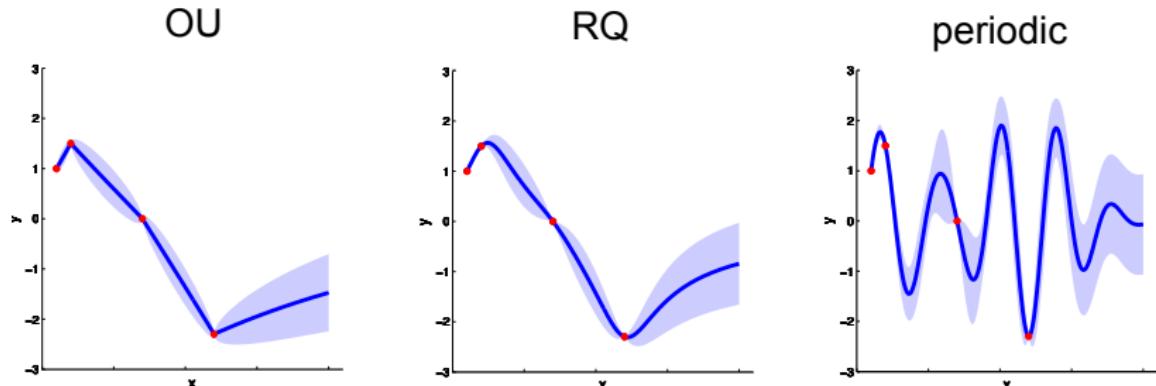
periodic



Bayesian model comparison:

$$p(M|\mathbf{y}_{1:N}) = \frac{p(\mathbf{y}_{1:N}|M)p(M)}{\sum_{M'} p(\mathbf{y}_{1:N}|M')p(M')}$$

The covariance function has a large effect

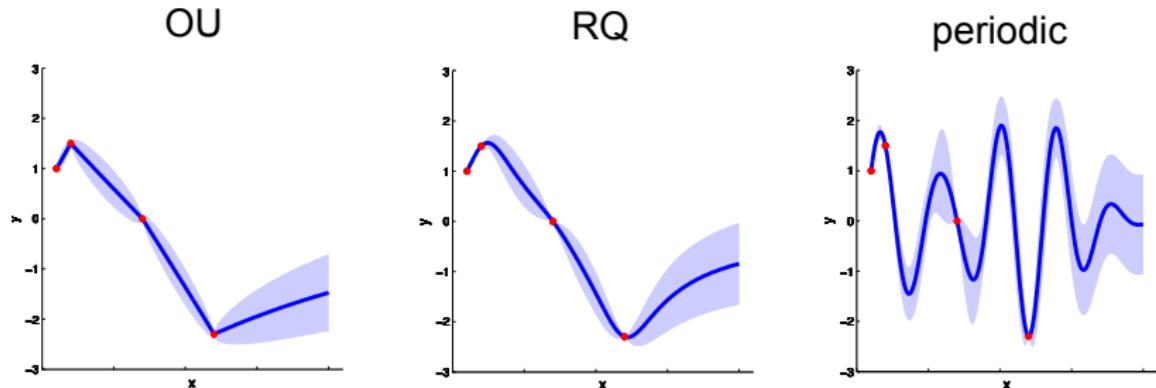


Bayesian model comparison:

$$p(M|\mathbf{y}_{1:N}) = \frac{p(\mathbf{y}_{1:N}|M)p(M)}{\sum_{M'} p(\mathbf{y}_{1:N}|M')p(M')}$$

← prior over models

The covariance function has a large effect



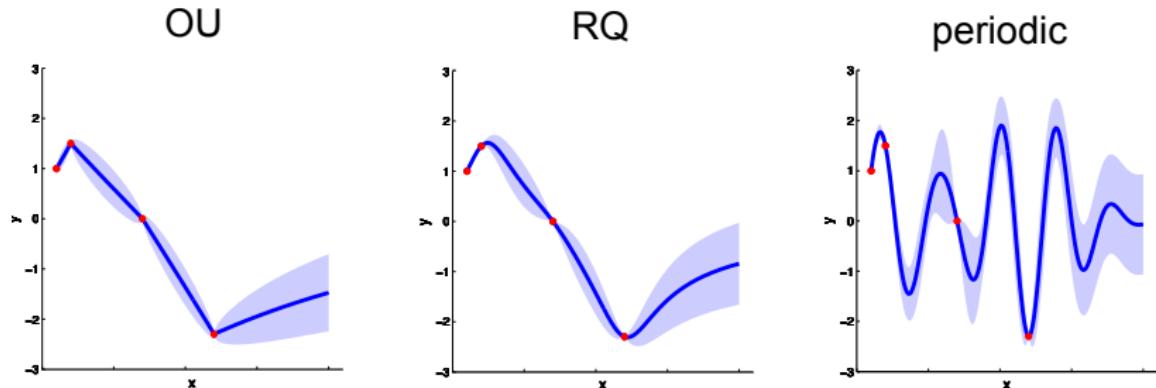
Bayesian model comparison:

$$p(M|\mathbf{y}_{1:N}) = \frac{p(\mathbf{y}_{1:N}|M)p(M)}{\sum_{M'} p(\mathbf{y}_{1:N}|M')p(M')}$$

prior over models

marginal likelihood $p(\mathbf{y}_{1:N}|M) = \int d\theta p(\mathbf{y}_{1:N}|\theta, M)p(\theta|M)$

The covariance function has a large effect



Bayesian model comparison:

$$p(M|\mathbf{y}_{1:N}) = \frac{p(\mathbf{y}_{1:N}|M)p(M)}{\sum_{M'} p(\mathbf{y}_{1:N}|M')p(M')}$$

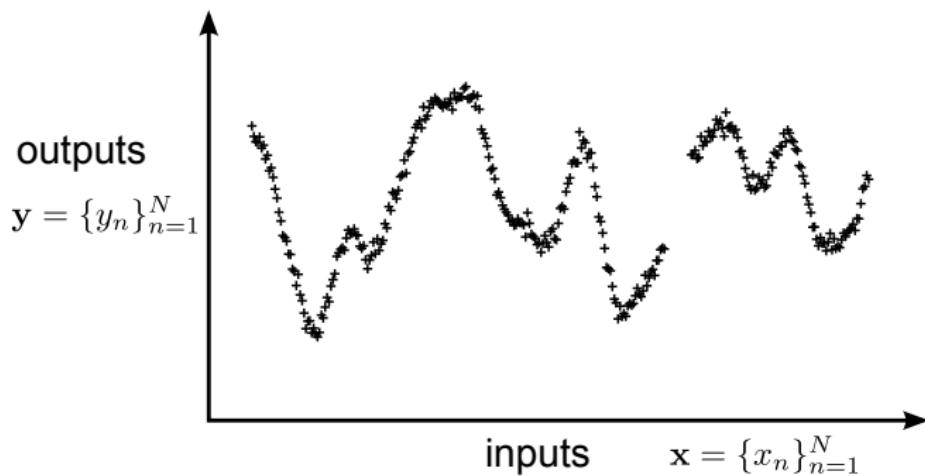
marginal likelihood $p(\mathbf{y}_{1:N}|M) = \int d\theta p(\mathbf{y}_{1:N}|\theta, M)p(\theta|M)$

prior over models

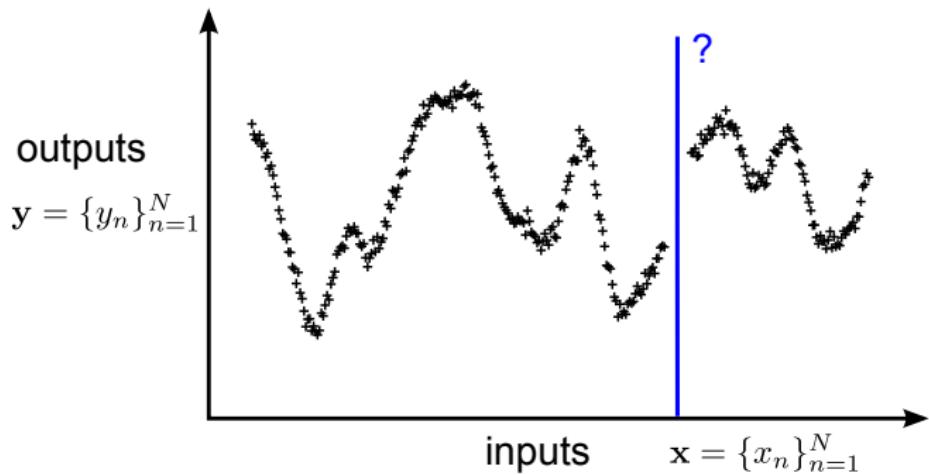
Health warnings: Hard to compute (need approximations)
Often results are very sensitive to the priors $p(\theta|M)$

Scaling Gaussian Processes to Large Datasets

Motivation: Gaussian Process Regression



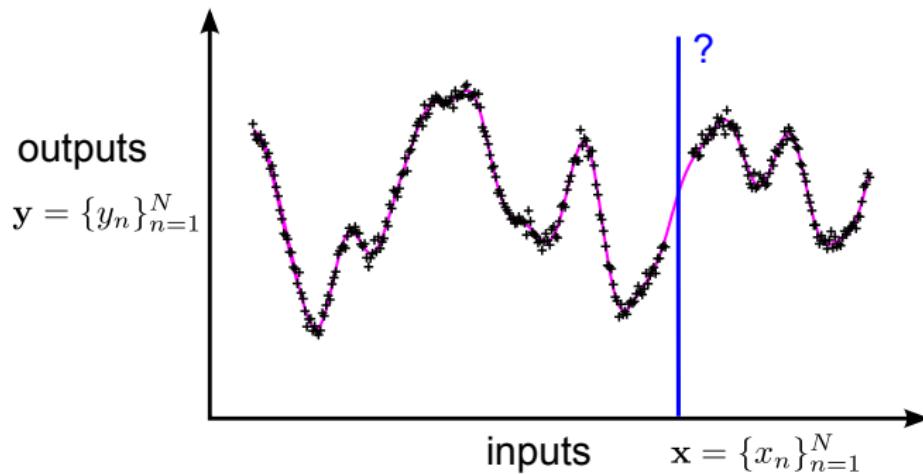
Motivation: Gaussian Process Regression



Motivation: Gaussian Process Regression

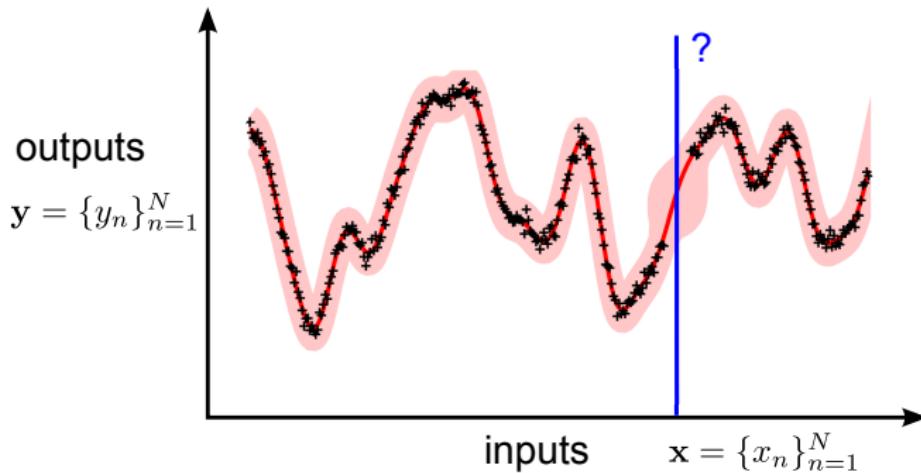
$$p(f|\theta) = \mathcal{GP}(\textcolor{magenta}{f}; 0, K_\theta)$$

$$p(y_n | \textcolor{magenta}{f}, x_n, \theta)$$



Motivation: Gaussian Process Regression

$$\begin{array}{ccc} p(f|\theta) = \mathcal{GP}(f; 0, K_\theta) & \xrightarrow{\text{inference \& learning}} & p(f|\mathbf{y}, \mathbf{x}, \theta) \\ p(y_n|f, x_n, \theta) & & p(\mathbf{y}|\mathbf{x}, \theta) \end{array}$$



Motivation: Gaussian Process Regression

$$p(f|\theta) = \mathcal{GP}(f; 0, K_\theta)$$

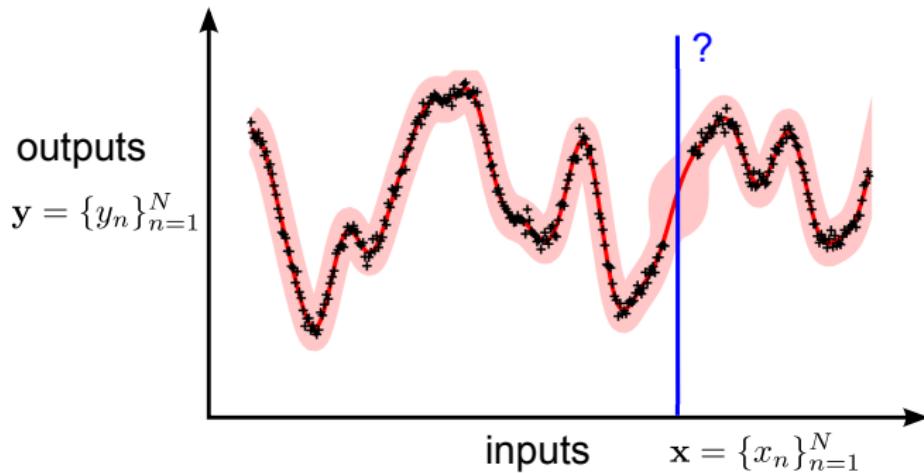
$$p(y_n|f, x_n, \theta)$$

inference & learning

intractabilities
computational $\mathcal{O}(N^3)$
analytic

$$p(f|\mathbf{y}, \mathbf{x}, \theta)$$

$$p(\mathbf{y}|\mathbf{x}, \theta)$$



Motivation: Gaussian Process Regression

$$p(f|\theta) = \mathcal{GP}(f; 0, K_\theta)$$

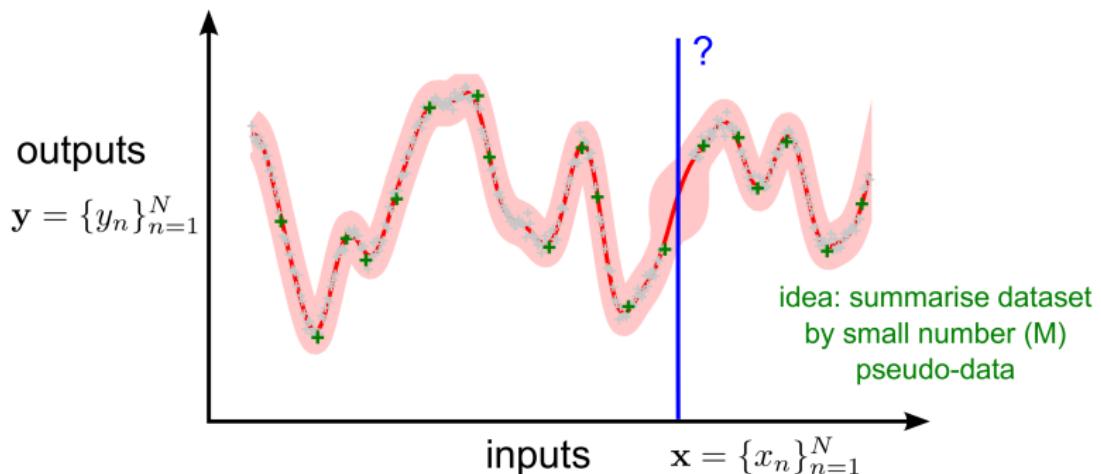
$$p(y_n|f, x_n, \theta)$$

inference & learning

intractabilities
computational $\mathcal{O}(N^3)$
analytic

$$p(f|\mathbf{y}, \mathbf{x}, \theta)$$

$$p(\mathbf{y}|\mathbf{x}, \theta)$$



A Brief History of Gaussian Process Approximations

FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

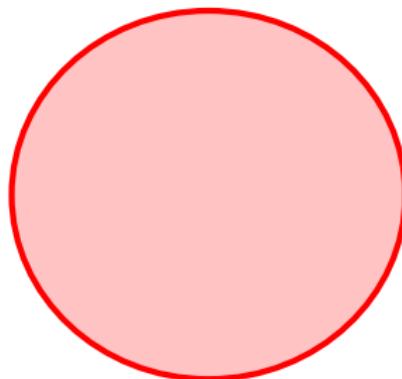
VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

$$\text{div}[p(\mathbf{f}, \mathbf{y}) || q(\mathbf{f}, \mathbf{y})]$$



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

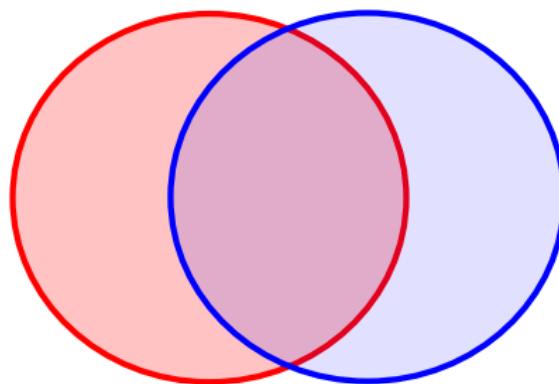
DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

methods employing
pseudo-data

$$\text{div}[p(\mathbf{f}, \mathbf{y}) || q(\mathbf{f}, \mathbf{y})]$$



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

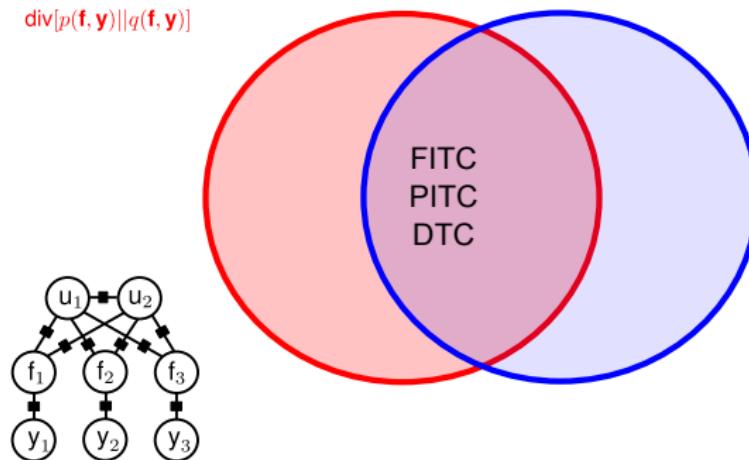
DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

methods employing
pseudo-data

$$\text{div}[p(\mathbf{f}, \mathbf{y}) || q(\mathbf{f}, \mathbf{y})]$$



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

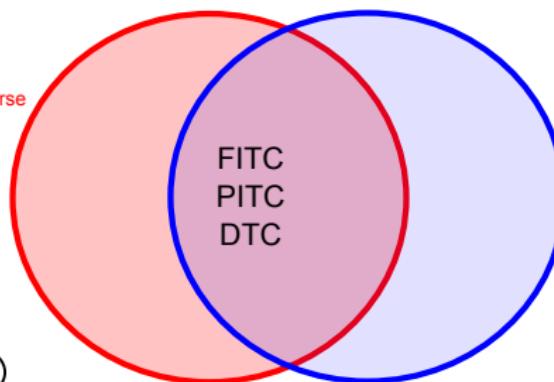
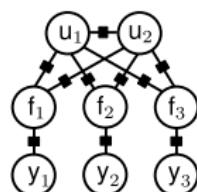
A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

methods employing
pseudo-data

$$\text{div}[p(\mathbf{f}, \mathbf{y}) || q(\mathbf{f}, \mathbf{y})]$$

A Unifying View of Sparse
Approximate Gaussian
Process Regression
Quinonero-Candela &
Rasmussen, 2005
(FITC, PITC, DTC)



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

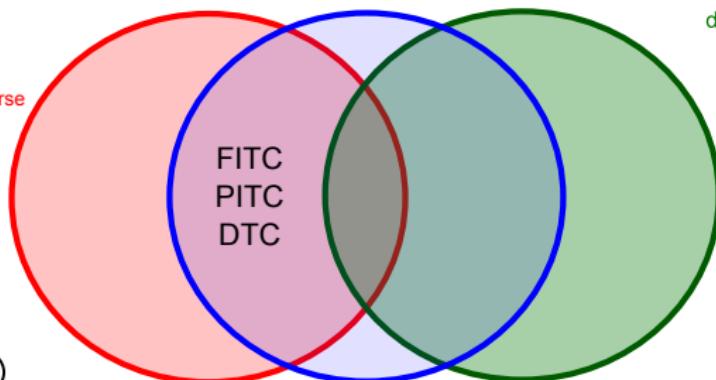
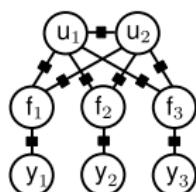
A Brief History of Gaussian Process Approximations

approximate generative model
exact inference methods employing
pseudo-data exact generative model
approximate inference

$$\text{div}[p(\mathbf{f}, \mathbf{y}) || q(\mathbf{f}, \mathbf{y})]$$

$$\text{div}[p(\mathbf{f} | \mathbf{y}) || q(\mathbf{f})]$$

A Unifying View of Sparse
Approximate Gaussian
Process Regression
Quinonero-Candela &
Rasmussen, 2005
(FITC, PITC, DTC)



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

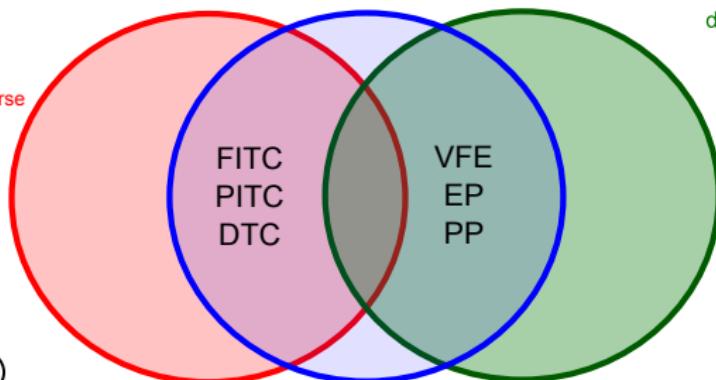
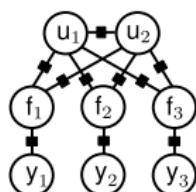
A Brief History of Gaussian Process Approximations

approximate generative model
exact inference methods employing
pseudo-data exact generative model
approximate inference

$$\text{div}[p(\mathbf{f}, \mathbf{y}) || q(\mathbf{f}, \mathbf{y})]$$

$$\text{div}[p(\mathbf{f}|\mathbf{y}) || q(\mathbf{f})]$$

A Unifying View of Sparse
Approximate Gaussian
Process Regression
Quinonero-Candela &
Rasmussen, 2005
(FITC, PITC, DTC)



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

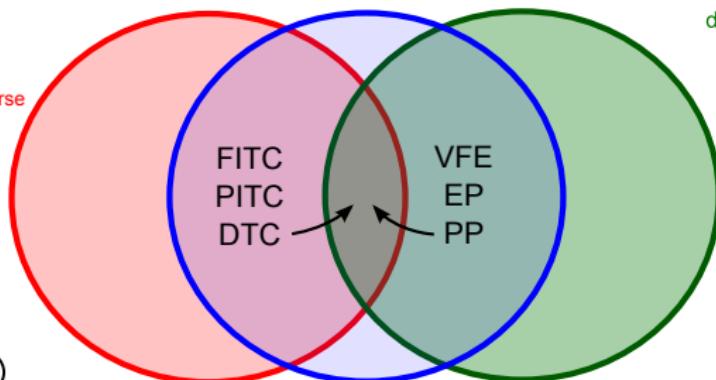
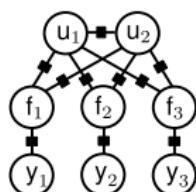
A Brief History of Gaussian Process Approximations

approximate generative model
exact inference methods employing
pseudo-data exact generative model
approximate inference

$$\text{div}[p(\mathbf{f}, \mathbf{y}) || q(\mathbf{f}, \mathbf{y})]$$

$$\text{div}[p(\mathbf{f} | \mathbf{y}) || q(\mathbf{f})]$$

A Unifying View of Sparse
Approximate Gaussian
Process Regression
Quinonero-Candela &
Rasmussen, 2005
(FITC, PITC, DTC)



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

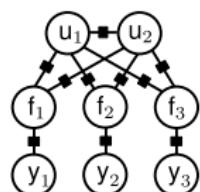
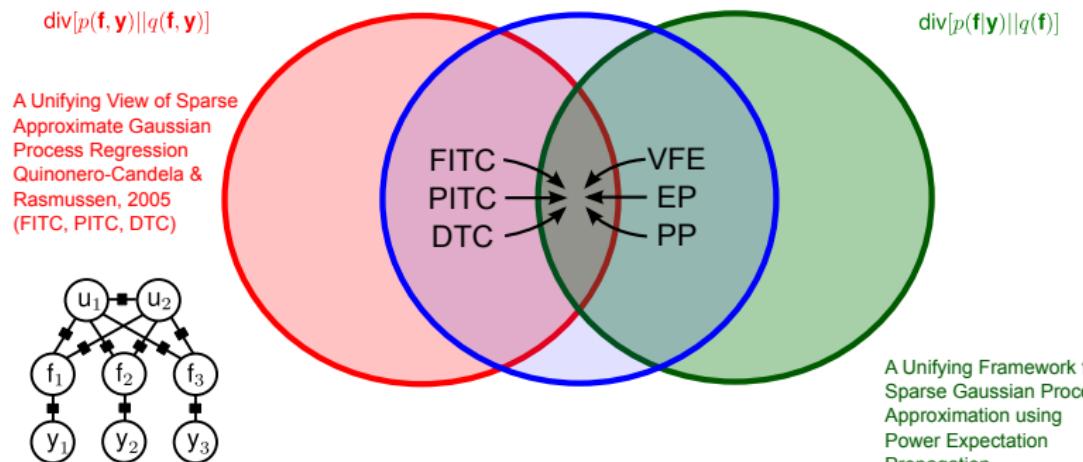
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference methods employing
pseudo-data exact generative model
approximate inference



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

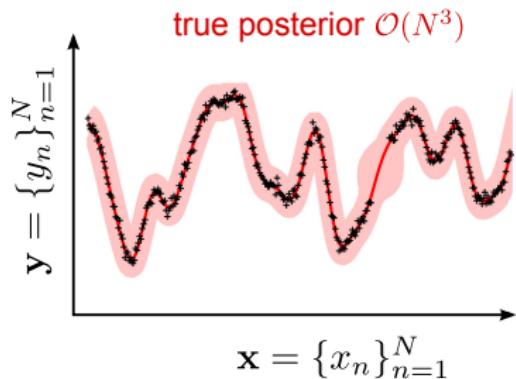
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

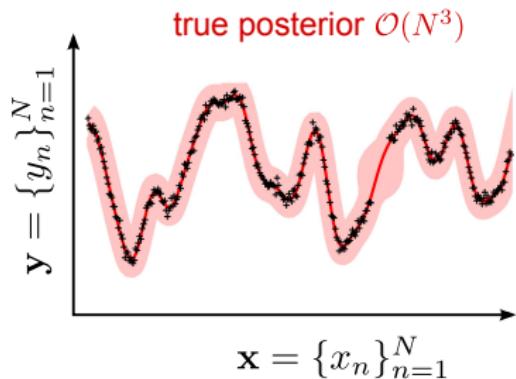
EP pseudo-point approximation

$$p^*(f) = p(f, \mathbf{y} | \mathbf{x}, \theta)$$



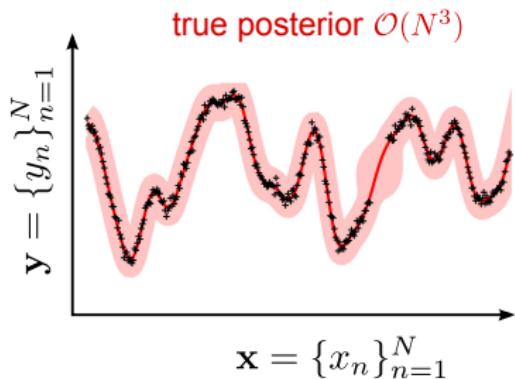
EP pseudo-point approximation

$$\begin{aligned} p^*(f) &= p(f, \mathbf{y} | \mathbf{x}, \theta) \\ &= p(f | \theta) \prod_{n=1}^N \underline{p(y_n | f, x_n, \theta)} \end{aligned}$$



EP pseudo-point approximation

$$\begin{aligned} p^*(f) &= p(f, \mathbf{y} | \mathbf{x}, \theta) \\ &= p(f | \theta) \prod_{n=1}^N \underline{p(y_n | f, x_n, \theta)} \\ &= \underbrace{p(\mathbf{y} | \mathbf{x}, \theta)}_{\text{marginal likelihood}} \underbrace{p(f | \mathbf{y}, \mathbf{x}, \theta)}_{\text{posterior}} \end{aligned}$$



EP pseudo-point approximation

$$p^*(f) = p(f, \mathbf{y} | \mathbf{x}, \theta)$$

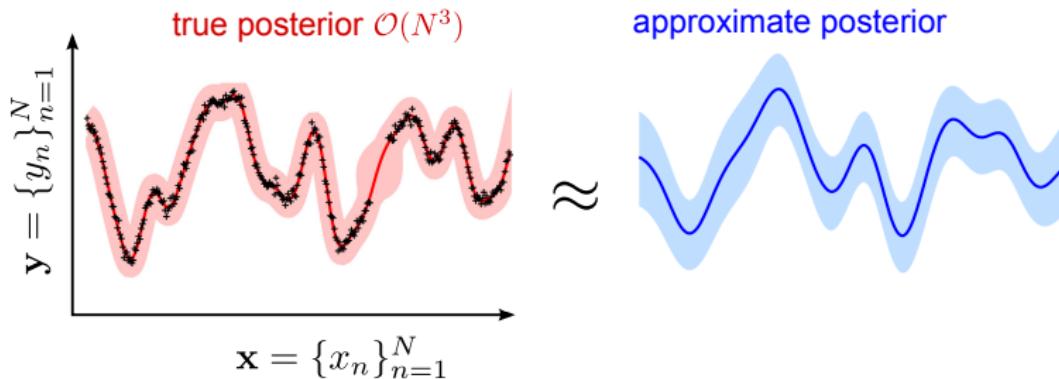
$$= p(f|\theta) \prod_{n=1}^N \underline{p(y_n|f, x_n, \theta)}$$

$$= \underline{p(\mathbf{y}|\mathbf{x}, \theta)} \underline{p(f|\mathbf{y}, \mathbf{x}, \theta)}$$

marginal
likelihood

posterior

$$q^*(f) = p(f|\theta) \prod_{n=1}^N \underline{t_n(f)}$$



EP pseudo-point approximation

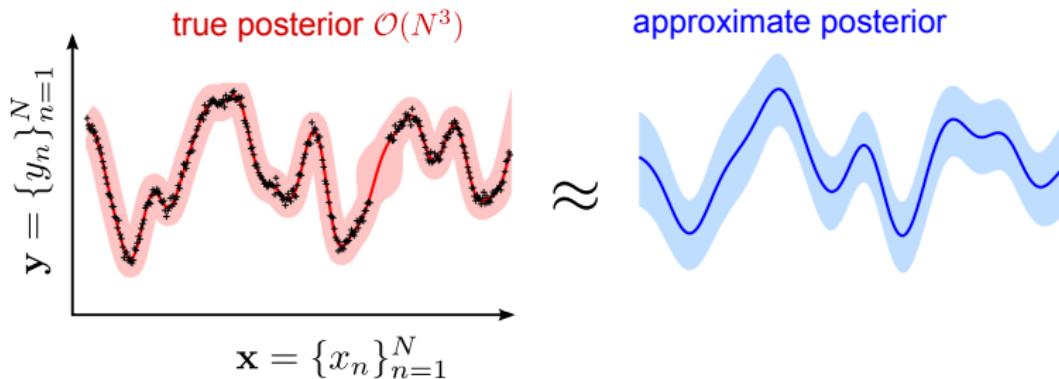
$$p^*(f) = p(f, \mathbf{y} | \mathbf{x}, \theta)$$

$$= p(f|\theta) \prod_{n=1}^N \underline{p(y_n|f, x_n, \theta)}$$

$$= \underbrace{p(\mathbf{y}|\mathbf{x}, \theta)}_{\text{marginal likelihood}} \underbrace{p(f|\mathbf{y}, \mathbf{x}, \theta)}_{\text{posterior}}$$

$$q^*(f) = p(f|\theta) \prod_{n=1}^N \underline{t_n(f)}$$

$$= \underbrace{Z_{\text{EP}}}_{\text{ }} \underbrace{q(f)}_{\text{ }}$$



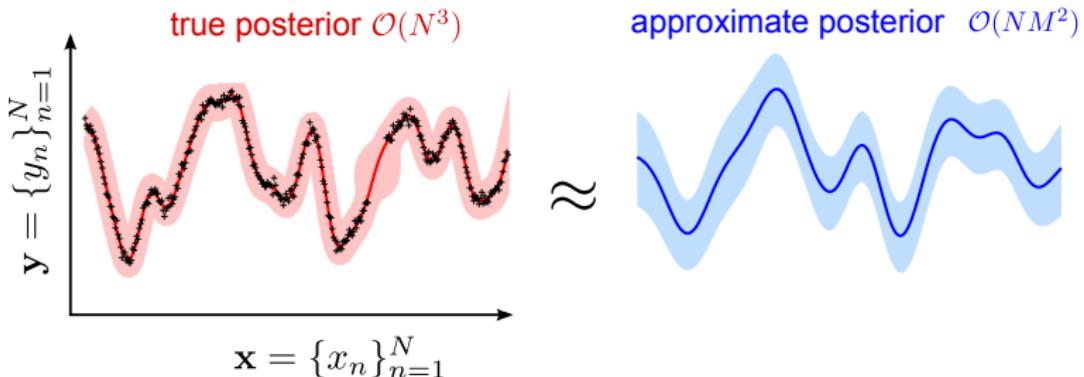
EP pseudo-point approximation

$$p^*(f) = p(f, \mathbf{y} | \mathbf{x}, \theta)$$

$$= p(f|\theta) \prod_{n=1}^N \underline{p(y_n|f, x_n, \theta)}$$

$$= \underbrace{p(\mathbf{y}|\mathbf{x}, \theta)}_{\text{marginal likelihood}} \underbrace{p(f|\mathbf{y}, \mathbf{x}, \theta)}_{\text{posterior}}$$

$$\begin{aligned} q^*(f) &= p(f|\theta) \prod_{n=1}^N \underline{t_n(f)} \\ &= \underline{Z_{\text{EP}}} \underline{q(f)} \\ t_n(f) &= \mathcal{N}(\mathbf{u}; \mu_n, \Sigma_n) \\ \dim(\mathbf{u}) &= M \quad f = \{\mathbf{u}, f_{\neq \mathbf{u}}\} \end{aligned}$$

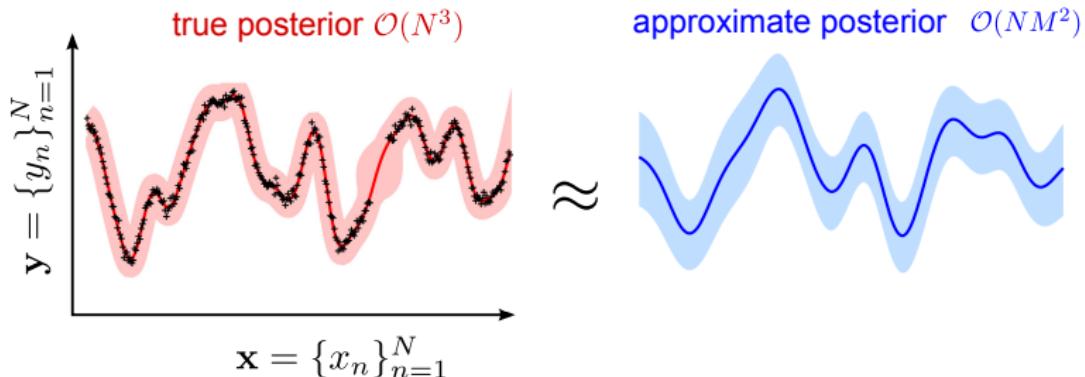


EP pseudo-point approximation

$$\begin{aligned} p^*(f) &= p(f, \mathbf{y} | \mathbf{x}, \theta) \\ &= p(f | \theta) \prod_{n=1}^N \underline{p(y_n | f, x_n, \theta)} \\ &= \underline{p(\mathbf{y} | \mathbf{x}, \theta)} \underline{p(f | \mathbf{y}, \mathbf{x}, \theta)} \end{aligned}$$

marginal likelihood posterior

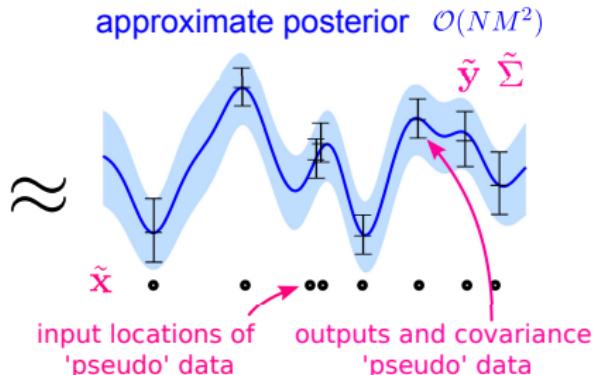
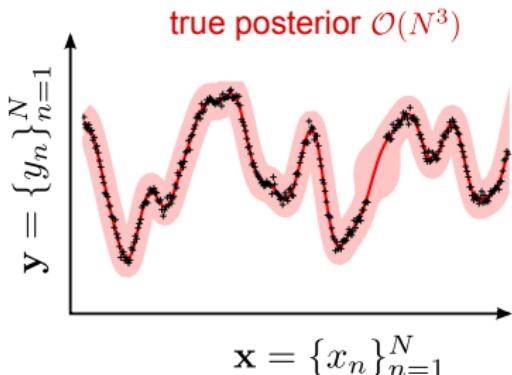
$$\begin{aligned} q^*(f) &= p(f | \theta) p(\tilde{\mathbf{y}} | \mathbf{u}, \tilde{\Sigma}) \\ &= p(f | \theta) \prod_{n=1}^N \underline{t_n(f)} \\ &= \underline{Z_{\text{EP}}} \underline{q(f)} \\ t_n(f) &= \mathcal{N}(\mathbf{u}; \mu_n, \Sigma_n) \\ \dim(\mathbf{u}) = M \quad f &= \{\mathbf{u}, f_{\neq \mathbf{u}}\} \end{aligned}$$



EP pseudo-point approximation

$$\begin{aligned} p^*(f) &= p(f, \mathbf{y} | \mathbf{x}, \theta) \\ &= p(f | \theta) \prod_{n=1}^N p(y_n | f, x_n, \theta) \\ &= \underbrace{p(\mathbf{y} | \mathbf{x}, \theta)}_{\text{marginal likelihood}} \underbrace{p(f | \mathbf{y}, \mathbf{x}, \theta)}_{\text{posterior}} \end{aligned}$$

$$\begin{aligned} q^*(f) &= p(f | \theta) p(\tilde{\mathbf{y}} | \mathbf{u}, \tilde{\Sigma}) \\ &= p(f | \theta) \prod_{n=1}^N t_n(f) \\ &= \underbrace{Z_{\text{EP}}}_{\text{exact joint of new GP regression model}} \underbrace{q(f)}_{t_n(f) = \mathcal{N}(\mathbf{u}; \mu_n, \Sigma_n)} \\ \dim(\mathbf{u}) &= M \quad f = \{\mathbf{u}, f_{\neq \mathbf{u}}\} \end{aligned}$$



EP algorithm

EP algorithm

1. remove

$$q^{\setminus n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one
pseudo-observation
likelihood

EP algorithm

1. remove

$$q^{\setminus n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one
pseudo-observation
likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\setminus n}(f)p(y_n|f, x_n, \theta)$$

tilted

add in one
true observation
likelihood

EP algorithm

1. remove

$$q^{\setminus n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one
pseudo-observation
likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\setminus n}(f)p(y_n|f, x_n, \theta)$$

tilted

KL between unnormalised
stochastic processes

add in one
true observation
likelihood

3. project

$$q^*(f) = \operatorname{argmin}_{q^*(f)} \text{KL} [p_n^{\text{tilt}}(f) || q^*(f)]$$

project onto
approximating
family

EP algorithm

1. remove

$$q^{\setminus n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one
pseudo-observation
likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\setminus n}(f)p(y_n|f, x_n, \theta)$$

tilted

KL between unnormalised
stochastic processes

add in one
true observation
likelihood

3. project

$$q^*(f) = \operatorname{argmin}_{q^*(f)} \text{KL} [p_n^{\text{tilt}}(f) || q^*(f)]$$

project onto
approximating
family

4. update

$$t_n(\mathbf{u}) = \frac{q^*(f)}{q^{\setminus n}(f)}$$

update
pseudo-observation
likelihood

EP algorithm

1. remove

$$q^{\setminus n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one
pseudo-observation
likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\setminus n}(f)p(y_n|f, x_n, \theta)$$

tilted

KL between unnormalised
stochastic processes

add in one
true observation
likelihood

3. project

$$q^*(f) = \underset{q^*(f)}{\operatorname{argmin}} \text{KL} [p_n^{\text{tilt}}(f) || q^*(f)]$$

project onto
approximating
family

1. minimum: moments matched at pseudo-inputs $\mathcal{O}(NM^2)$
2. Gaussian regression: matches moments everywhere

4. update

$$t_n(\mathbf{u}) = \frac{q^*(f)}{q^{\setminus n}(f)}$$

update
pseudo-observation
likelihood

EP algorithm

1. remove

$$q^{\setminus n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one
pseudo-observation
likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\setminus n}(f)p(y_n|f, x_n, \theta)$$

tilted

KL between unnormalised
stochastic processes

add in one
true observation
likelihood

3. project

$$q^*(f) = \underset{q^*(f)}{\operatorname{argmin}} \text{KL} [p_n^{\text{tilt}}(f) || q^*(f)]$$

project onto
approximating
family

1. minimum: moments matched at pseudo-inputs $\mathcal{O}(NM^2)$
2. Gaussian regression: matches moments everywhere

4. update

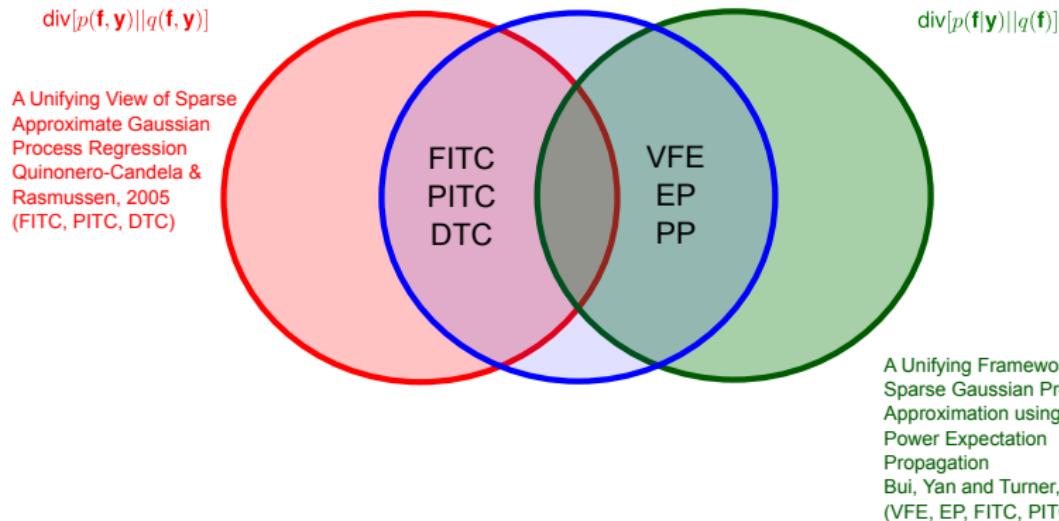
$$\begin{aligned} t_n(\mathbf{u}) &= \frac{q^*(f)}{q^{\setminus n}(f)} \\ &= z_n \mathcal{N}(\mathbf{K}_{f_n \mathbf{u}} \mathbf{K}_{\mathbf{u} \mathbf{u}}^{-1} \mathbf{u}; g_n, v_n) \end{aligned}$$

update
pseudo-observation
likelihood

rank 1

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference methods employing
pseudo-data exact generative model
approximate inference



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

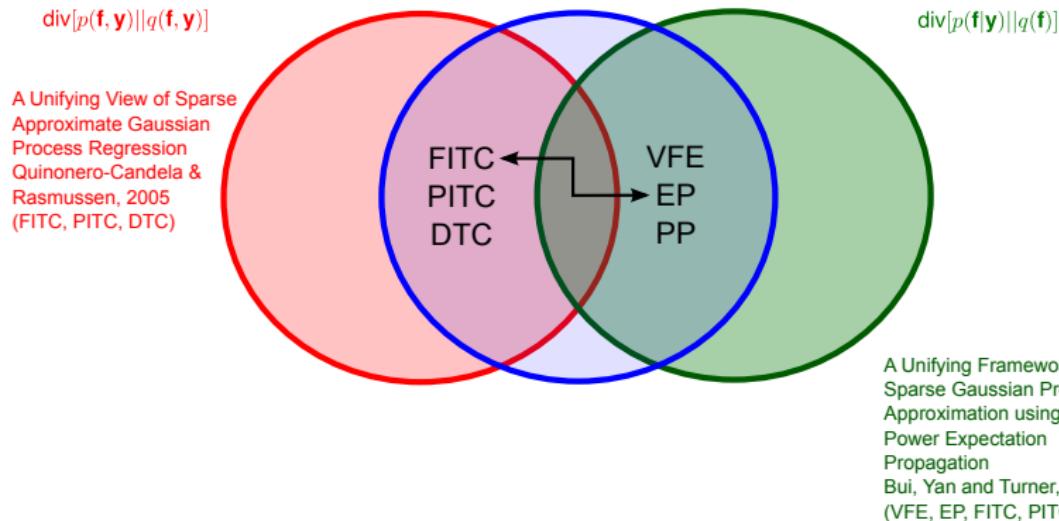
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

Fixed points of EP = FITC approximation

approximate generative model
exact inference methods employing
pseudo-data exact generative model
approximate inference



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

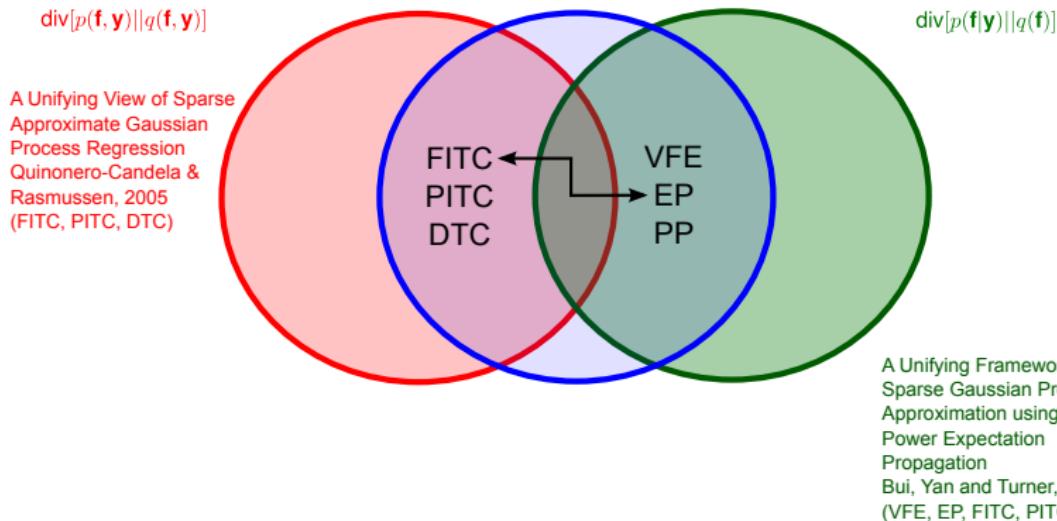
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

Fixed points of EP = FITC approximation

approximate generative model
exact inference methods employing
pseudo-data exact generative model
approximate inference



FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

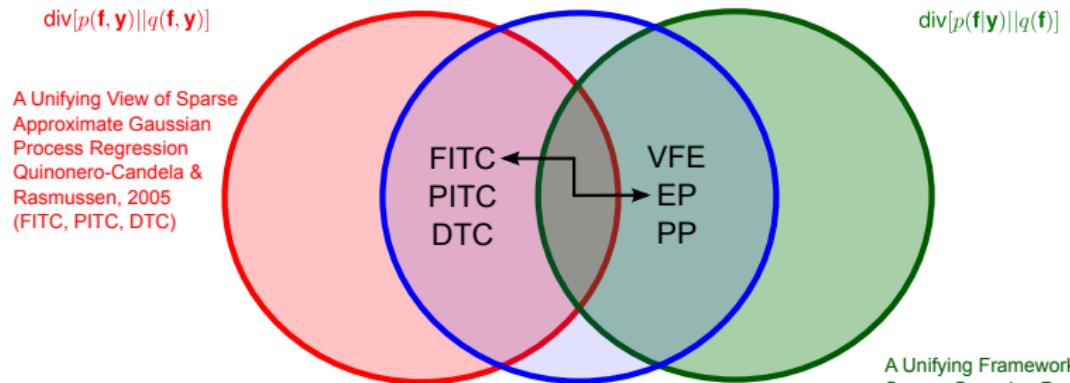
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

Fixed points of EP = FITC approximation

approximate generative model
exact inference methods employing
pseudo-data exact generative model
approximate inference



interpretation resolves issues with FITC:
why does it work so well?
are we allowed to increase M with N

FITC: Snelson et al. "Sparse Gaussian Processes using Pseudo-inputs"

PITC: Snelson et al. "Local and global sparse Gaussian process approximations"

EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."

VFE: Titsias "Variational Learning of Inducing Variables in Sparse Gaussian Processes"

DTC / PP: Seeger et al. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression"

EP algorithm

1. remove

$$q^{\setminus n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one
pseudo-observation
likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\setminus n}(f)p(y_n|f, x_n, \theta)$$

↑
tilted

add in one
true observation
likelihood

KL between unnormalised
stochastic processes

3. project

$$q^*(f) = \underset{q^*(f)}{\operatorname{argmin}} \text{KL} [p_n^{\text{tilt}}(f) || q^*(f)]$$

project onto
approximating
family

1. minimum: moments matched at pseudo-inputs $\mathcal{O}(NM^2)$

2. Gaussian regression: matches moments everywhere

4. update

$$t_n(\mathbf{u}) = \frac{q^*(f)}{q^{\setminus n}(f)}$$

update
pseudo-observation
likelihood

$$= z_n \mathcal{N}(\mathbf{K}_{f_n \mathbf{u}} \mathbf{K}_{\mathbf{u} \mathbf{u}}^{-1} \mathbf{u}; g_n, v_n)$$

rank 1

Power EP algorithm (as tractable as EP)

1. remove

$$q^{\setminus n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})^\alpha}$$

cavity

take out fraction of
pseudo-observation
likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\setminus n}(f)p(y_n|f, x_n, \theta)^\alpha$$

↑
tilted

add in fraction of
true observation
likelihood

KL between unnormalised
stochastic processes

3. project

$$q^*(f) = \underset{q^*(f)}{\operatorname{argmin}} \text{KL} [p_n^{\text{tilt}}(f) || q^*(f)]$$

project onto
approximating
family

1. minimum: moments matched at pseudo-inputs $\mathcal{O}(NM^2)$

2. Gaussian regression: matches moments everywhere

4. update

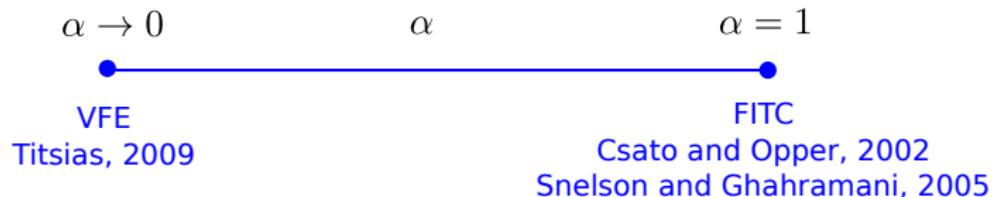
$$t_n(\mathbf{u})^\alpha = \frac{q^*(f)}{q^{\setminus n}(f)}$$

update
pseudo-observation
likelihood

$$t_n(\mathbf{u}) = z_n \mathcal{N}(\mathbf{K}_{f_n \mathbf{u}} \mathbf{K}_{\mathbf{u} \mathbf{u}}^{-1} \mathbf{u}; g_n, v_n)$$

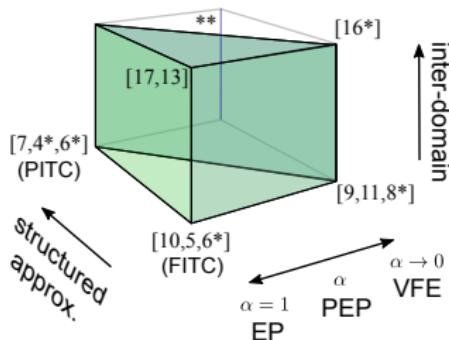
rank 1

Power EP: a unifying framework

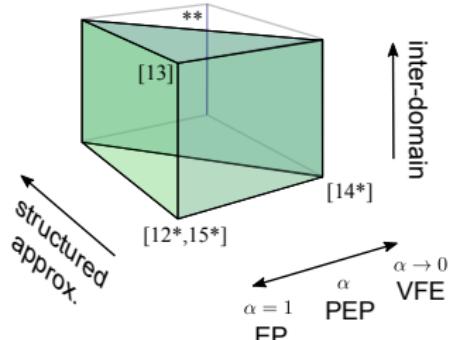


Power EP: a unifying framework

GP Regression



GP Classification



- [4] Quiñonero-Candela et al. 2005
- [5] Snelson et al., 2005
- [6] Snelson, 2006
- [7] Schwaighofer, 2002

* = optimised pseudo-inputs

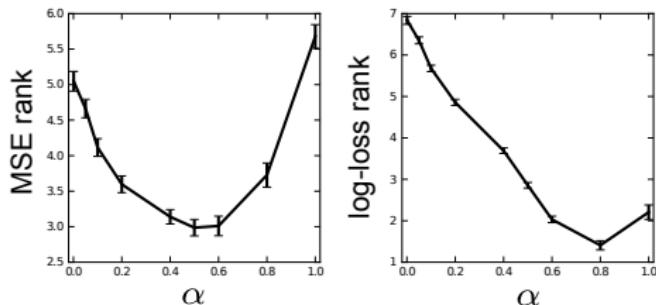
** = structured versions of VFE recover VFE

- [8] Titsias, 2009
- [9] Csató, 2002
- [10] Csató et al., 2002
- [11] Seeger et al., 2003

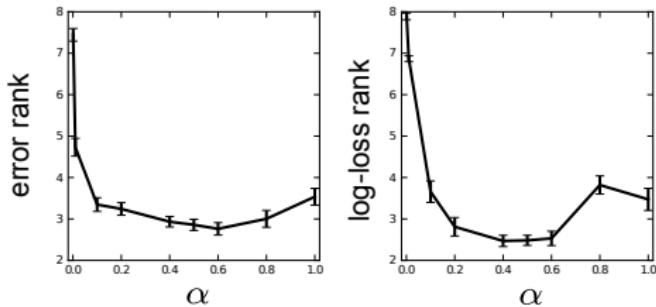
- [12] Naish-Guzman et al, 2007
- [13] Qi et al., 2010
- [14] Hensman et al., 2015
- [15] Hernández-Lobato et al., 2016
- [16] Matthews et al., 2016
- [17] Figueiras-Vidal et al., 2009

How should I set the power parameter α ?

8 UCI regression datasets
20 random splits
 $M = 0 - 200$
hypers and inducing
inputs optimised



6 UCI classification datasets
20 random splits
 $M = 10, 50, 100$
hypers and inducing
inputs optimised



$\alpha = 0.5$ does well on average

Deep Gaussian Processes for Regression

Pros and cons of Gaussian Process Regression

Pros and cons of Gaussian Process Regression

Probabilistic ✓

- need well-calibrated predictive uncertainty (decision making)
- big models (even with small data)
- neural networks do not provide well calibrated uncertainty estimates

Pros and cons of Gaussian Process Regression

Probabilistic ✓

need well-calibrated predictive uncertainty (decision making)

big models (even with small data)

neural networks do not provide well calibrated uncertainty estimates

Non-parametric (∞ - parametric) ✓

model parameters: information bottleneck between training and test data

training data $\rightarrow \theta \rightarrow$ test data

unbounded model (θ grows with data) pruned by data

Pros and cons of Gaussian Process Regression

Probabilistic ✓

need well-calibrated predictive uncertainty (decision making)

big models (even with small data)

neural networks do not provide well calibrated uncertainty estimates

Non-parametric (∞ - parametric) ✓

model parameters: information bottleneck between training and test data

training data $\rightarrow \theta \rightarrow$ test data

unbounded model (θ grows with data) pruned by data

Deep (and wide) ✗

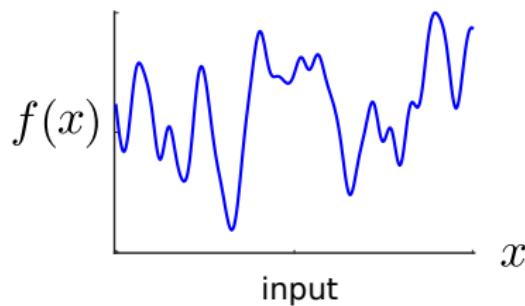
intrinsic hierarchical structure in real world data

simpler to learn a series of weak non-linearities

From Gaussian Processes to Deep Gaussian Processes

$$y(x) = \textcolor{blue}{f}(x) + \sigma_y \epsilon$$

$$f(x) = \mathcal{GP}(0, K_f(x, x'))$$

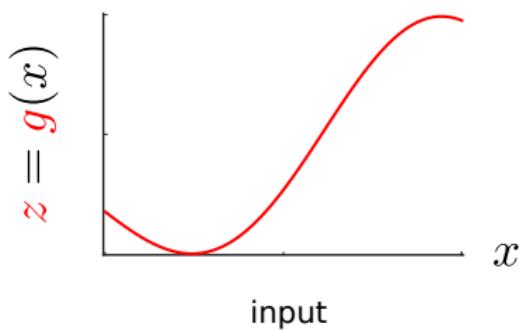
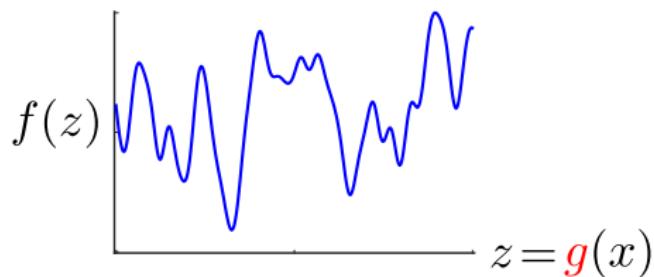


From Gaussian Processes to Deep Gaussian Processes

$$y(x) = \textcolor{blue}{f}(\textcolor{red}{g}(x)) + \sigma_y \epsilon$$

$$f(x) = \mathcal{GP}(0, K_f(x, x'))$$

$$g(x) = \mathcal{GP}(0, K_g(x, x'))$$

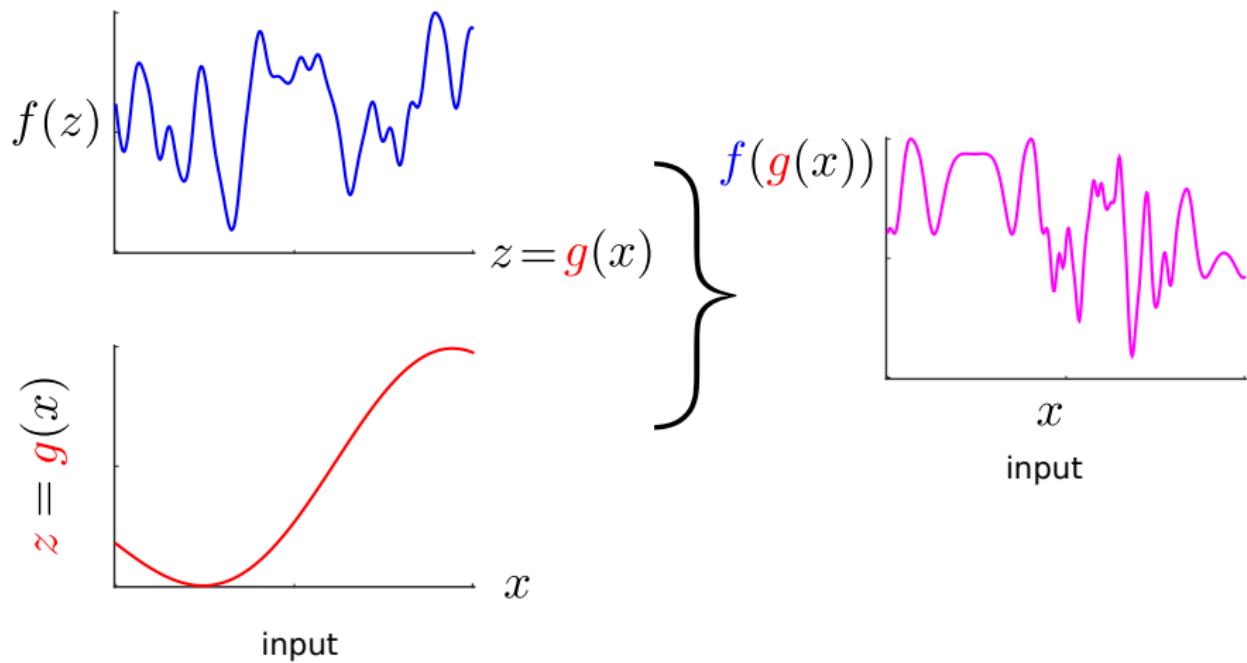


From Gaussian Processes to Deep Gaussian Processes

$$y(x) = \textcolor{blue}{f}(\textcolor{red}{g}(x)) + \sigma_y \epsilon$$

$$f(x) = \mathcal{GP}(0, K_f(x, x'))$$

$$g(x) = \mathcal{GP}(0, K_g(x, x'))$$

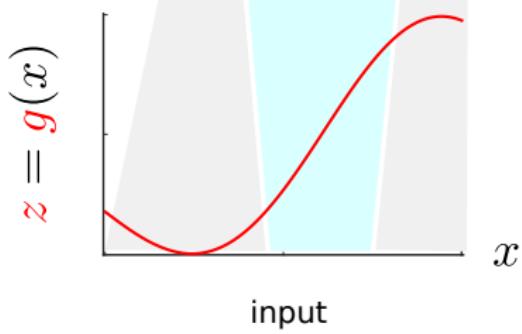
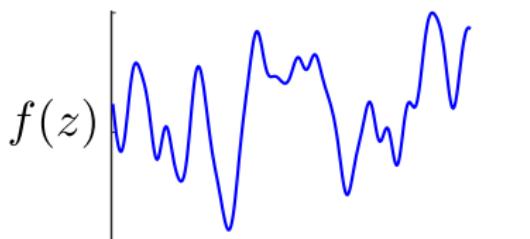


From Gaussian Processes to Deep Gaussian Processes

$$y(x) = \mathbf{f}(\mathbf{g}(x)) + \sigma_y \epsilon$$

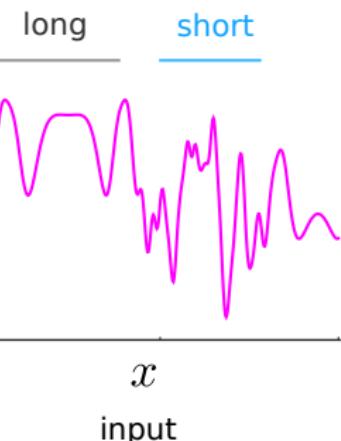
$$f(x) = \mathcal{GP}(0, K_f(x, x'))$$

$$g(x) = \mathcal{GP}(0, K_g(x, x'))$$



$z = \mathbf{g}(x)$

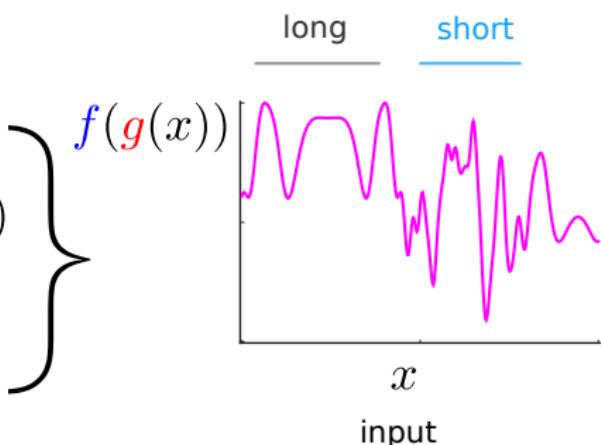
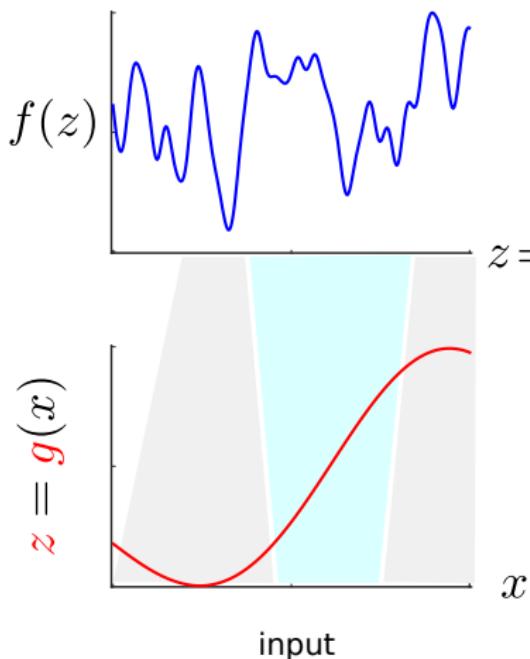
$\mathbf{f}(\mathbf{g}(x))$



From Gaussian Processes to Deep Gaussian Processes

$$y(x) = \textcolor{blue}{f}(\textcolor{red}{g}(x)) + \sigma_y \epsilon$$

$$\begin{aligned} f(x) &= \mathcal{GP}(0, K_f(x, x')) \\ g(x) &= \mathcal{GP}(0, K_g(x, x')) \end{aligned}$$



DGP = multi-layer neural network,
infinitely wide hidden layer

Deep Gaussian Processes

$$f_l \sim \mathcal{GP}(0, k(., .))$$

$$y_n = g(\mathbf{x}_n) = f_L(f_{L-1}(\cdots f_2(f_1(\mathbf{x}_n)))) + \epsilon_n$$

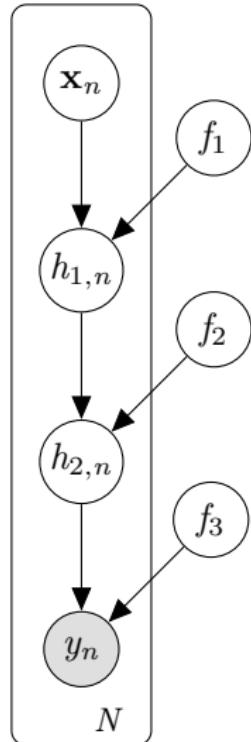
$$h_{L-1,n} := f_{L-1}(\cdots f_1(\mathbf{x}_n)), y_n = f_L(h_{L-1,n}) + \epsilon_n$$

Deep GPs^a are

- multi-layer generalisation of Gaussian processes,
- equivalent to deep neural networks with infinitely wide hidden layers

Questions:

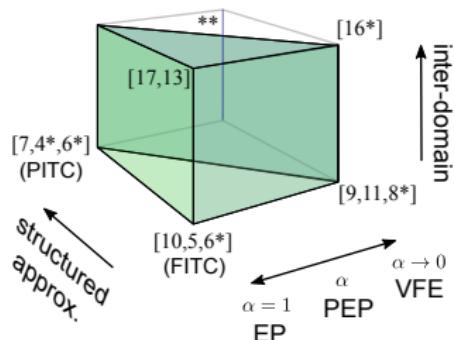
- How to perform inference and learning tractably?
- How Deep GPs compare to alternative,
e.g. Bayesian neural networks?



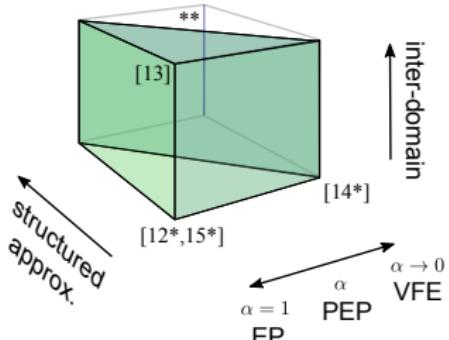
^aDamianou and Lawrence (2013) [unsupervised learning]

Approximate inference for (Deep) Gaussian Processes

GP Regression



GP Classification



[4] Quiñonero-Candela et al. 2005

[5] Snelson et al., 2005

[6] Snelson, 2006

[7] Schwaighofer, 2002

[8] Titsias, 2009

[9] Csató, 2002

[10] Csató et al., 2002

[11] Seeger et al., 2003

[12] Naish-Guzman et al, 2007

[13] Qi et al., 2010

[14] Hensman et al., 2015

[15] Hernández-Lobato et al., 2016

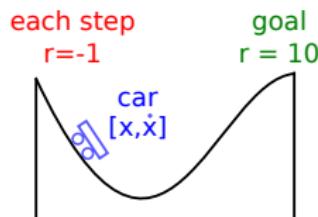
[16] Matthews et al., 2016

[17] Figueiras-Vidal et al., 2009

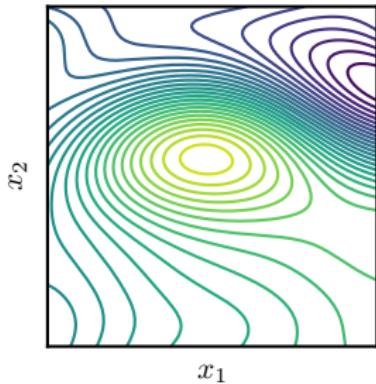
* = optimised pseudo-inputs

** = structured versions of VFE recover VFE

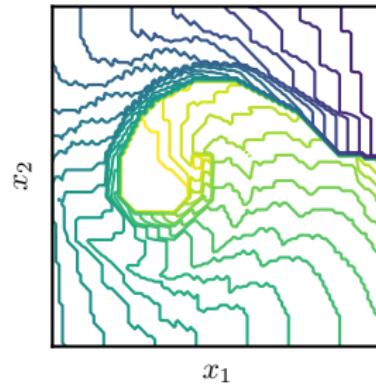
Experiment: Value function of the mountain car problem



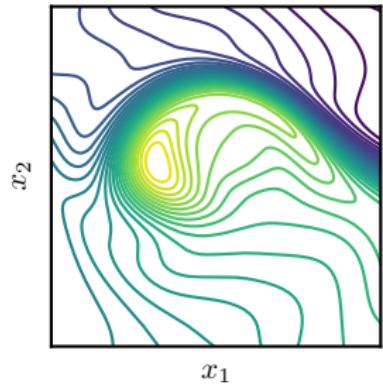
GP fit



Value function



DGP fit



Experiment: Comparison to Bayesian neural networks

Compare **DGPs** with **GPs** and **Bayesian neural networks** with one and two hidden layers using:

VI(G): Graves' VI [diagonal Gaussian, without the reparam. trick]

VI(KW): Kingma and Welling's VI [with the reparam. trick]

PBP: ADF with Probabilistic Backpropagation

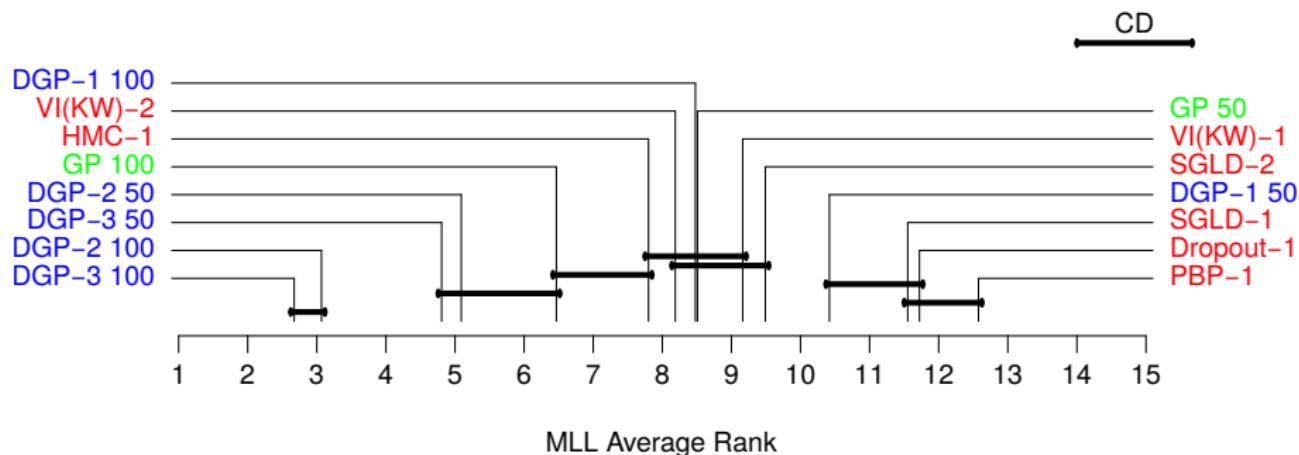
Dropout: Combining dropout predictions at test time

SGLD: Stochastic gradient Langevin dynamics

HMC: Hamiltonian Monte Carlo [only for small networks]

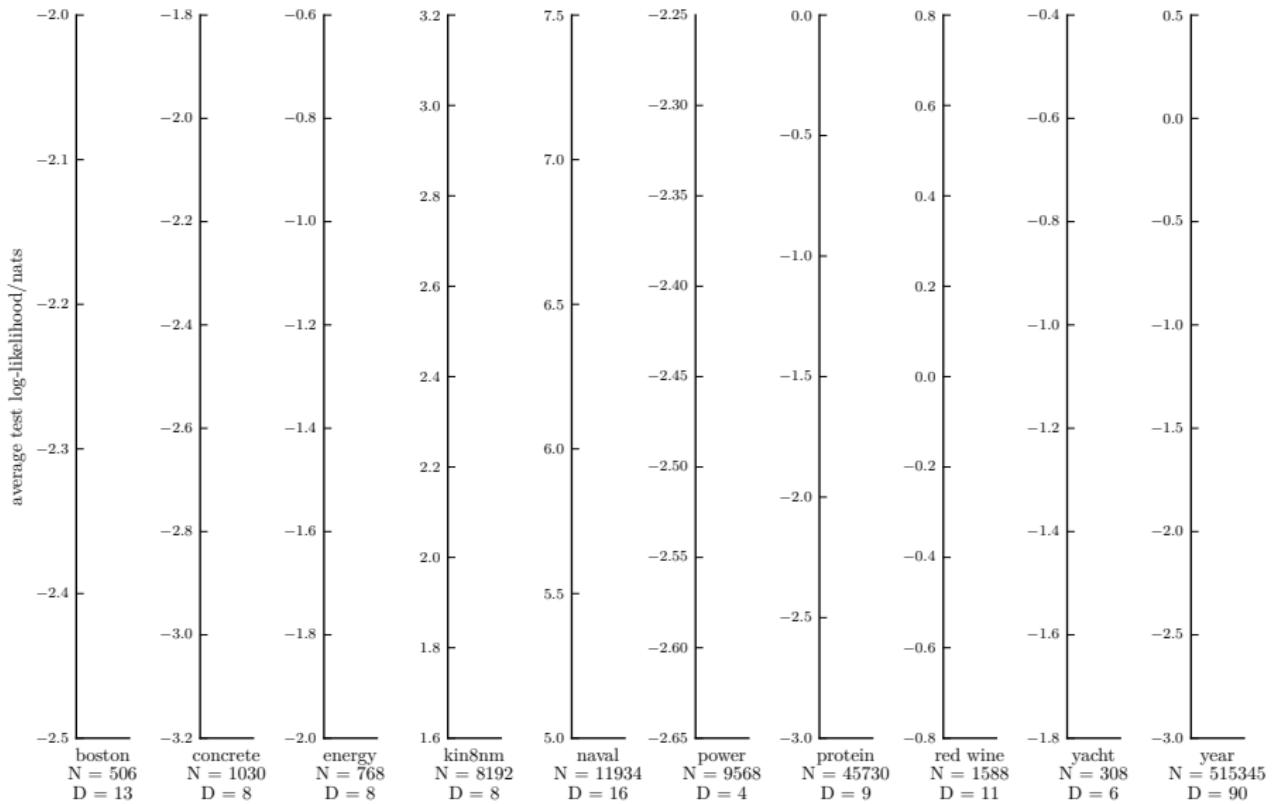
Experiment: Comparison to Bayesian neural networks

Rankings of all methods across all datasets

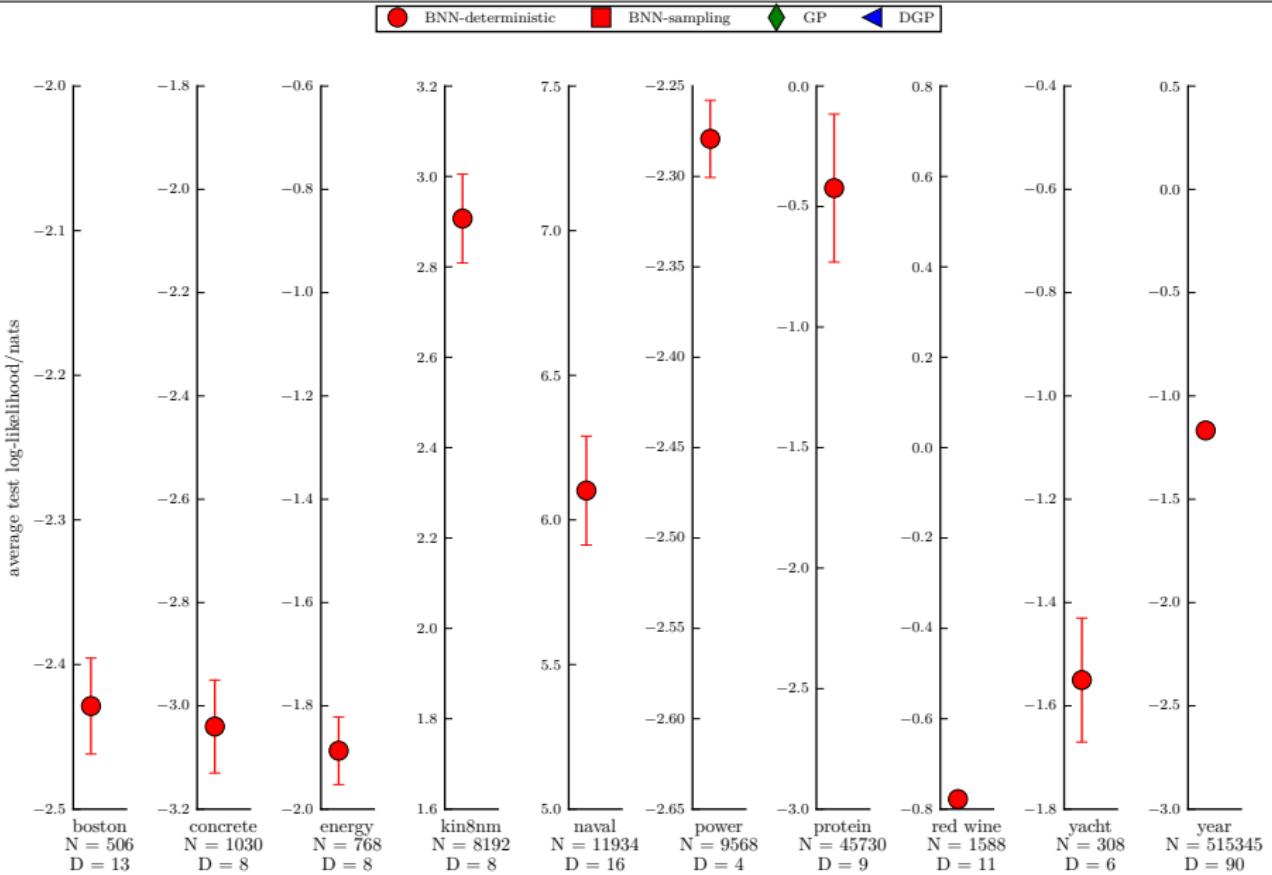


Experiment: Comparison to Bayesian neural networks [Best results]

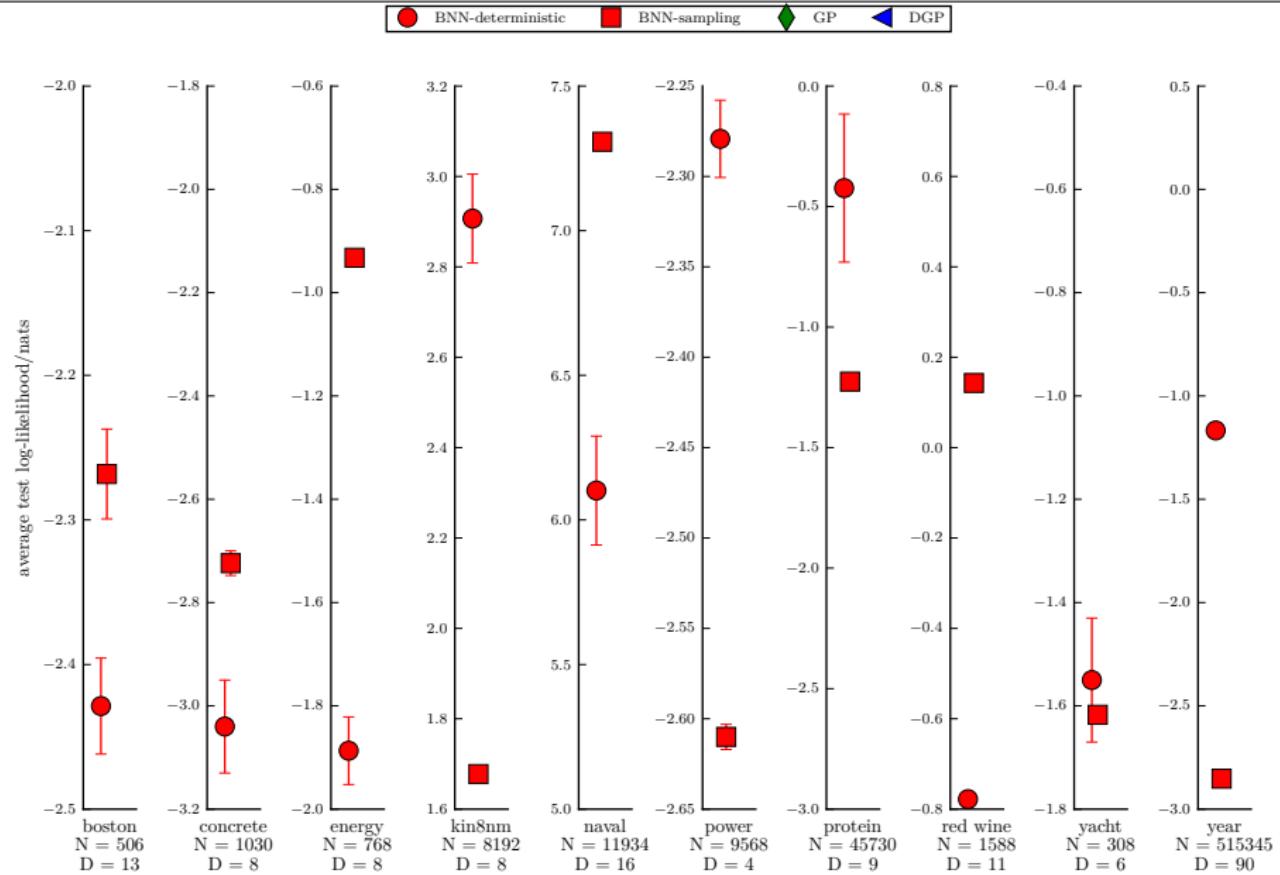
BNN-deterministic BNN-sampling GP DGP



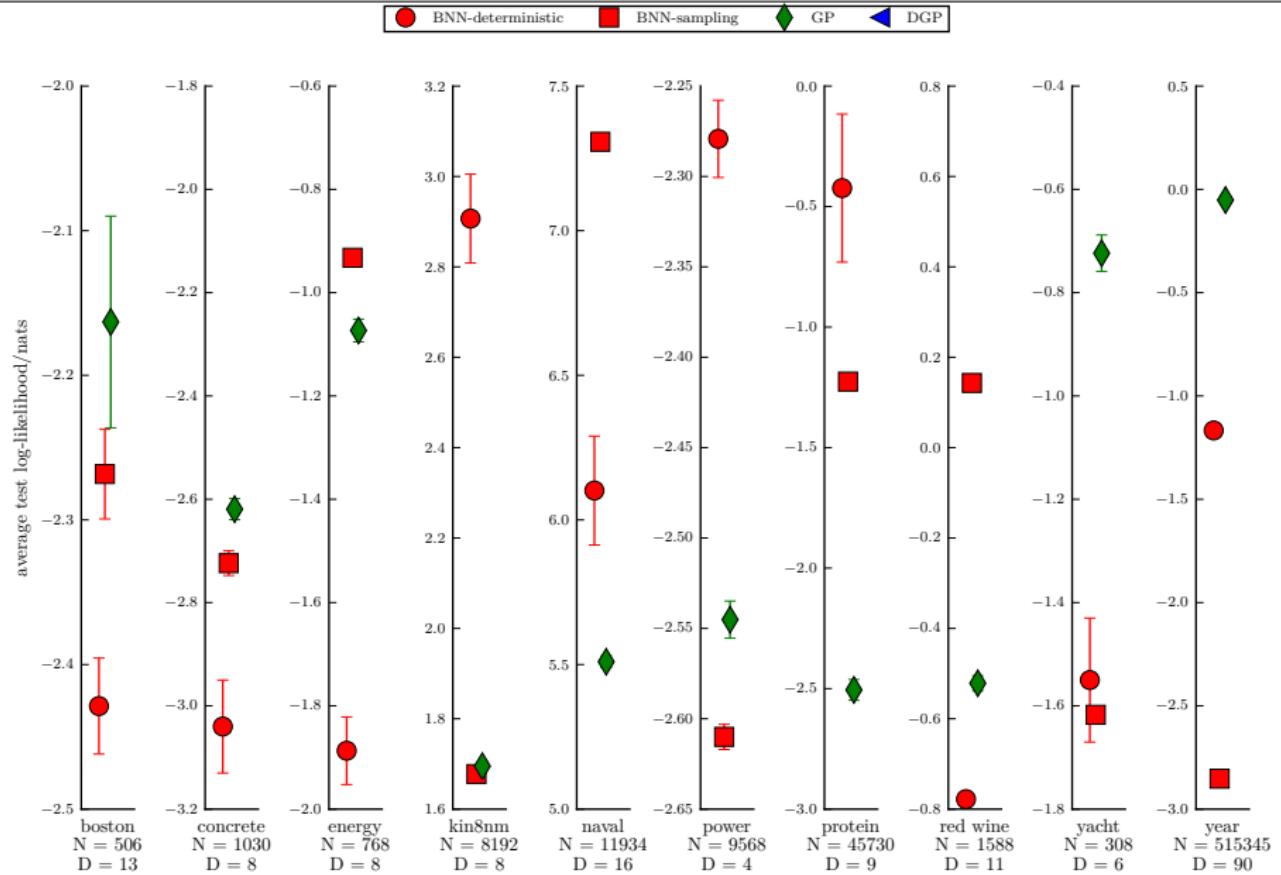
Experiment: Comparison to Bayesian neural networks [Best results]



Experiment: Comparison to Bayesian neural networks [Best results]

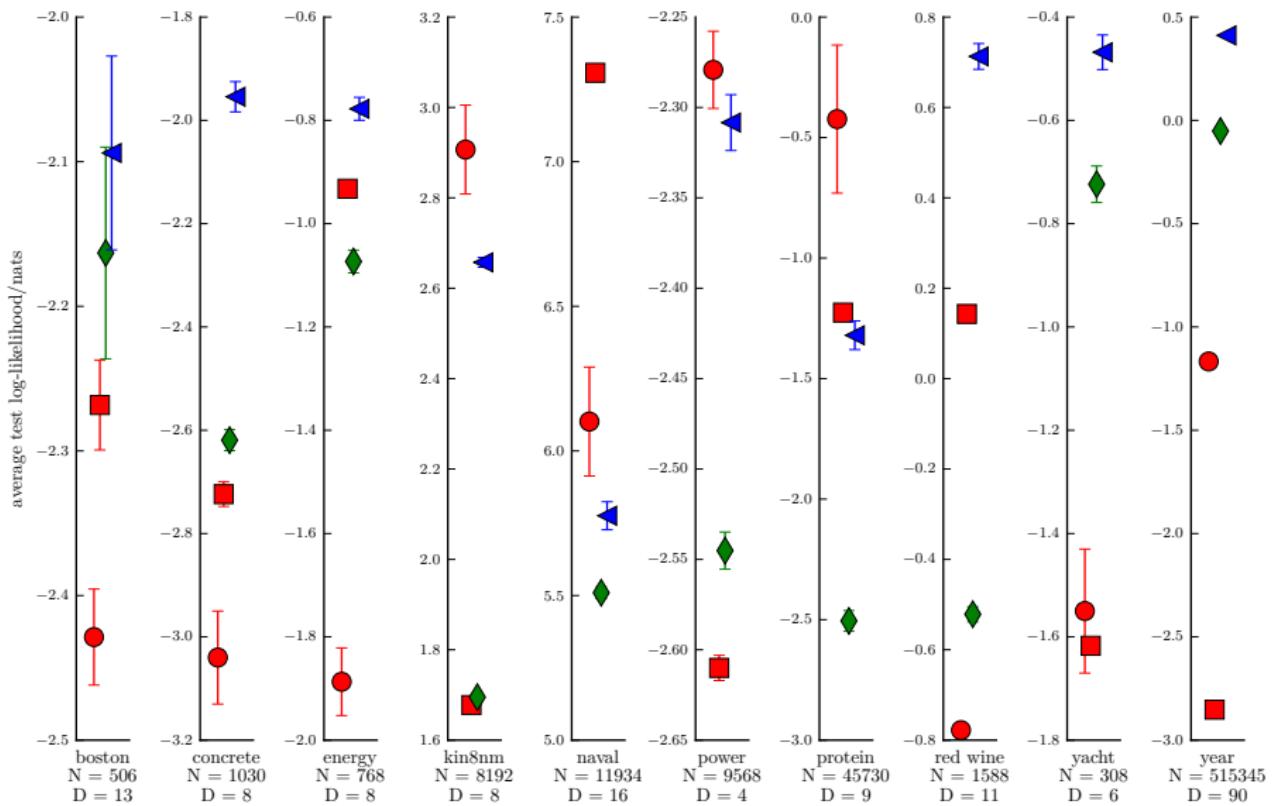


Experiment: Comparison to Bayesian neural networks [Best results]



Experiment: Comparison to Bayesian neural networks [Best results]

BNN-deterministic BNN-sampling GP DGP



Pros and cons of Gaussian Process Regression

Probabilistic ✓

need well-calibrated predictive uncertainty (decision making)

big models (even with small data)

neural networks do not provide well calibrated uncertainty estimates

Non-parametric (∞ - parametric) ✓

model parameters: information bottleneck between training and test data

training data $\rightarrow \theta \rightarrow$ test data

unbounded model (θ grows with data) pruned by data

Deep (and wide) ✓

intrinsic hierarchical structure in real world data

simpler to learn a series of weak non-linearities

References (hyperlinked)

Deep Gaussian Processes:

- Deep Gaussian Processes for Regression using Approximate Expectation Propagation, ICML 2016

Approximate inference in GPs:

- A Unifying Framework for Sparse Gaussian Process Approximation using Power Expectation Propagation, arXiv preprint 2016

Scalable Approximate inference:

- Stochastic Expectation Propagation, NIPS 2015
- Black-box α -divergence Minimization, ICML 2016