

Carnegie Mellon

School of Computer Science

Deep Reinforcement Learning and Control

# Bayesian Optimization / Experiment Design with Gaussian Processes

Spring 2020, CMU 10-403

Katerina Fragkiadaki



# Used Materials

- **Disclaimer:** Some material and slides for this lecture were borrowed from Nando de Freitas lecture of Gaussian processes and Bayesian Optimization, from Richard Turner's lecture on Gaussian process, and from Kirthevasan Kandasamy's lecture on Bayesian optimization.

# This lecture - Motivation

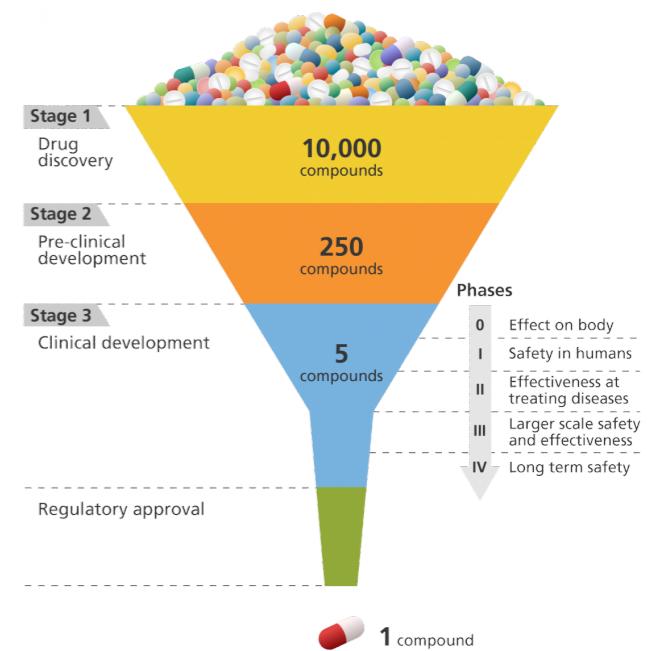
Learning to act in a non-sequential setup with continuous actions:

- Each action returns an immediate reward.
- We want to choose actions that maximize our expected (immediate) reward.

Example: drug discovery

Actions: the compounds to mix

Rewards: drug effectiveness/safety (e.g., as measured in mice).



# This lecture - Motivation

Learning to act in a non-sequential setup:

- Each action returns an immediate reward.
- We want to choose actions that maximize our expected (immediate) reward.

Example: drilling for oil

Actions: where to drill next

Rewards: how much oil I found



# This lecture - Motivation

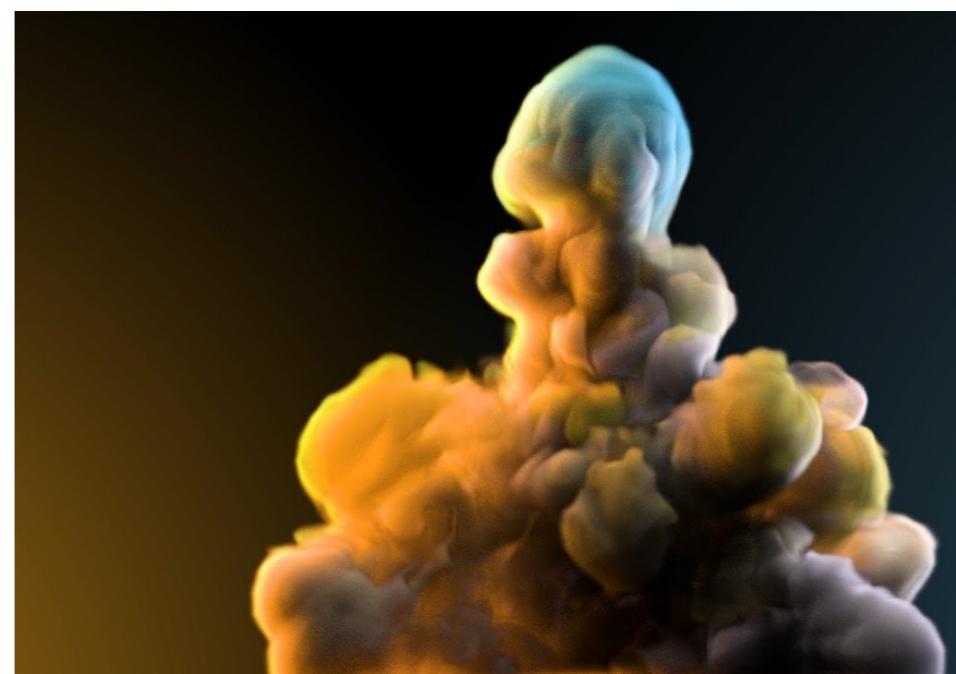
Learning to act in a non-sequential setup:

- Each action returns an immediate reward.
- We want to choose actions that maximize our expected (immediate) reward.

Example: simulating smoke

Actions: what simulation parameters to use

Rewards: how realistic the resulting smoke looks



# This lecture - Motivation

Learning to act in a non-sequential setup:

- Each action returns an immediate reward.
- We want to choose actions that maximize our expected (immediate) reward.

Example: walking after breaking your ankle

Actions: what walking style to use

Rewards: how (non) painful it is (more in the next lecture)



# This lecture - Motivation

Learning to act in a non-sequential setup:

- Each action returns an immediate reward.
- We want to choose actions that maximize our expected (immediate) reward.

It turns out, this is equivalent to maximizing a function for which:

- We do not have an explicit parametric form, e.g., we do not know the mapping from smoke simulation parameters to realism/human pleasure from watching the smoke  
unknown relationship between actions vs immediate rewards??
- We may have a parametric form but function evaluation is very expensive.

In both cases, we cannot use gradient information.

# This lecture - Motivation

Learning to act in a non-sequential setup:

- Each action returns an immediate reward.
- We want to choose actions that maximize our expected (immediate) reward.

It turns out, this is equivalent to **black-box (no gradients) optimization** of functions.

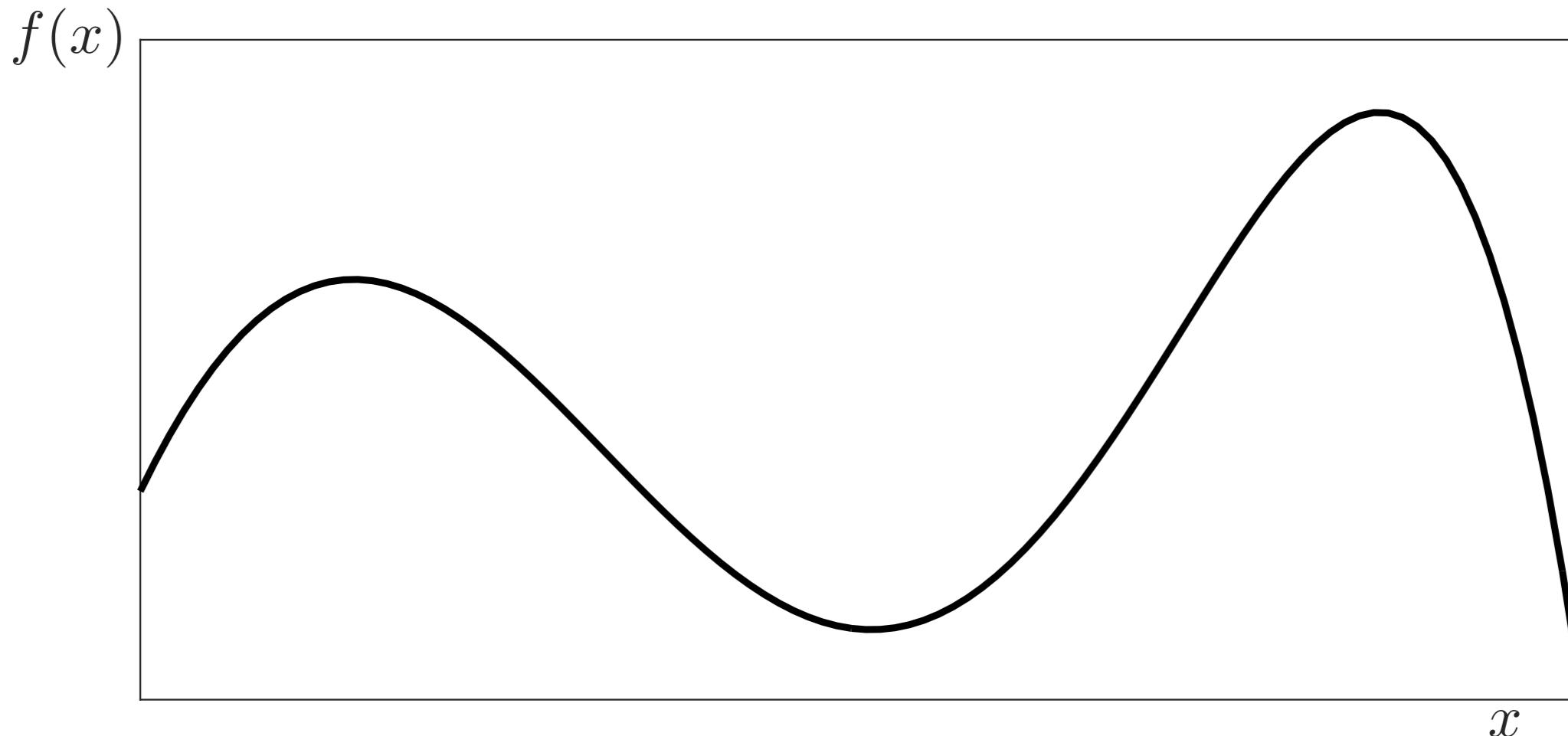
too complicated model - no explicit gradient is available - like cooling system in buildings

Actions: places to evaluate the function.

Rewards: the value of the function.

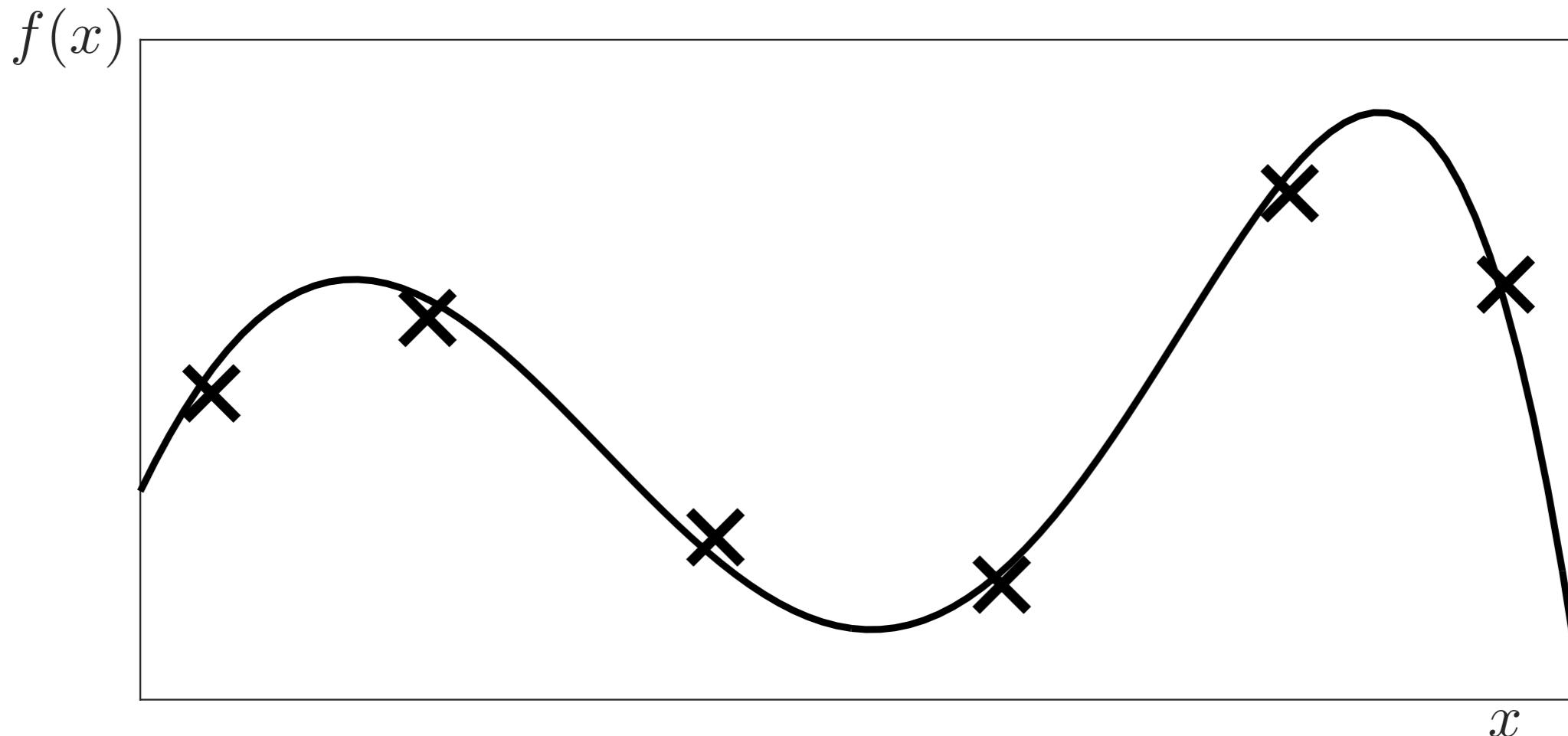
# Black-box Optimisation

$f: \mathcal{X} \rightarrow \mathbb{R}$  is an expensive black-box function, accessible only via noisy evaluations.



# Black-box Optimisation

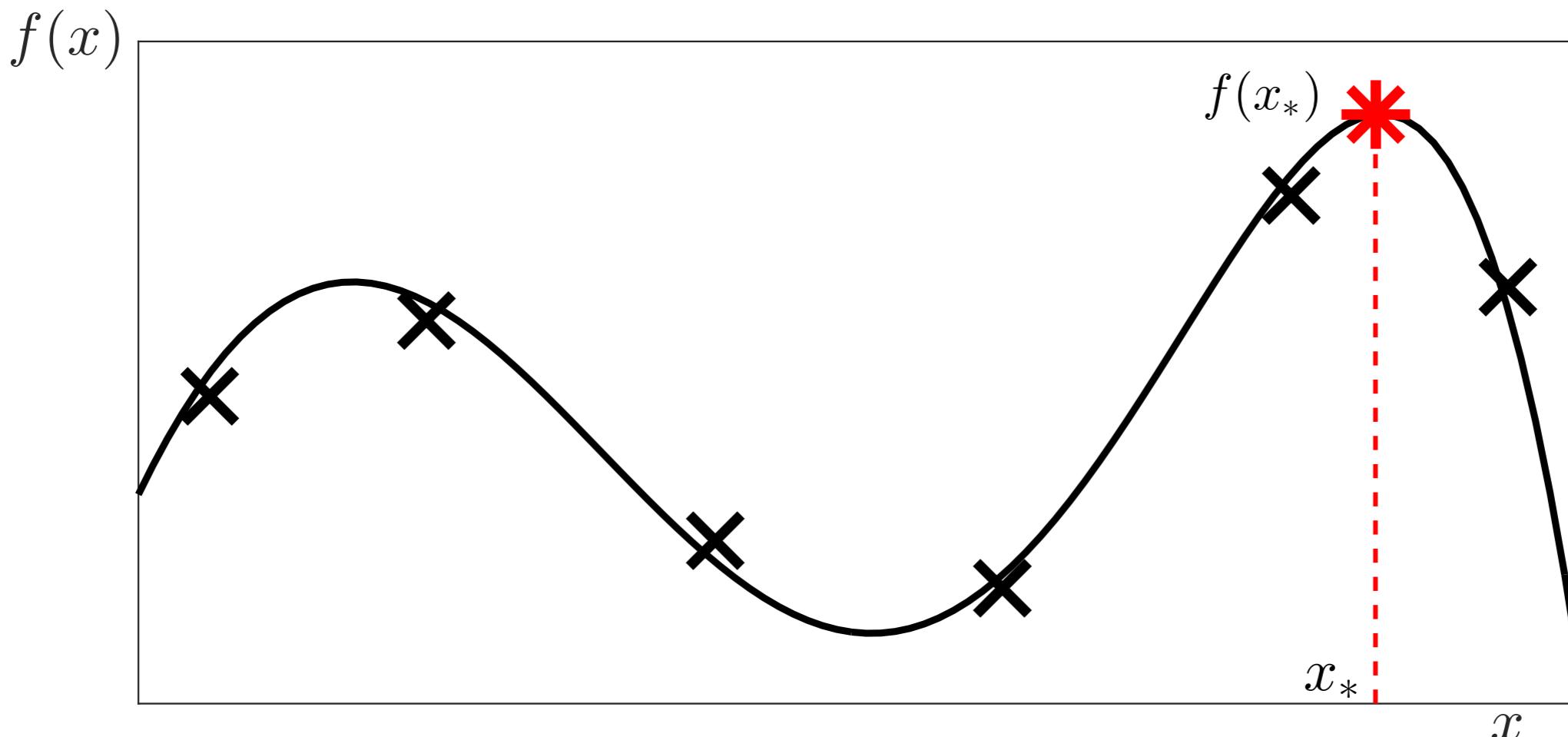
$f: \mathcal{X} \rightarrow \mathbb{R}$  is an expensive black-box function, accessible only via noisy evaluations.



# Black-box Optimisation

$f: \mathcal{X} \rightarrow \mathbb{R}$  is an expensive black-box function, accessible only via noisy evaluations.

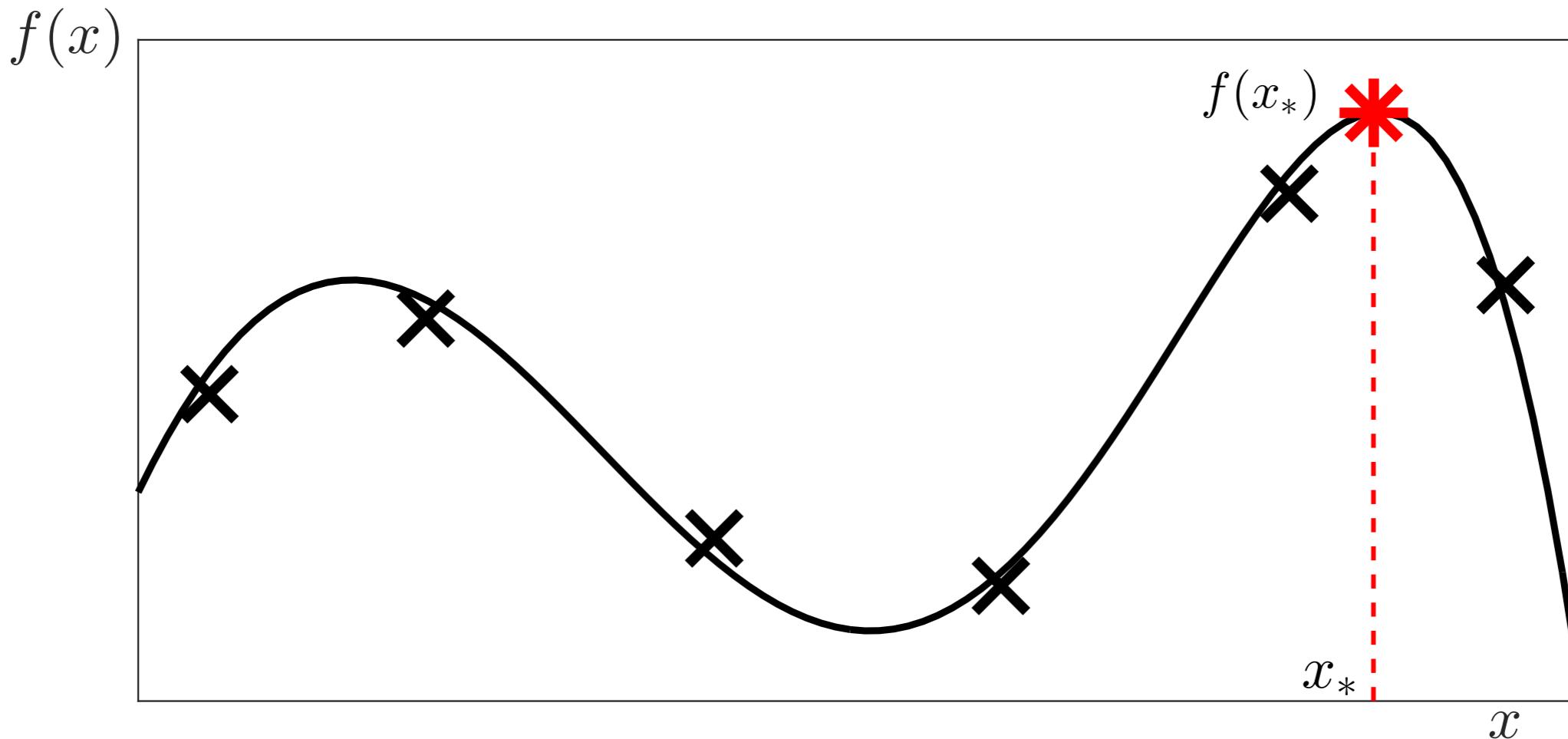
Let  $x_* = \operatorname{argmax}_x f(x)$



# Black-box Optimisation

$f: \mathcal{X} \rightarrow \mathbb{R}$  is an expensive black-box function, accessible only via noisy evaluations.

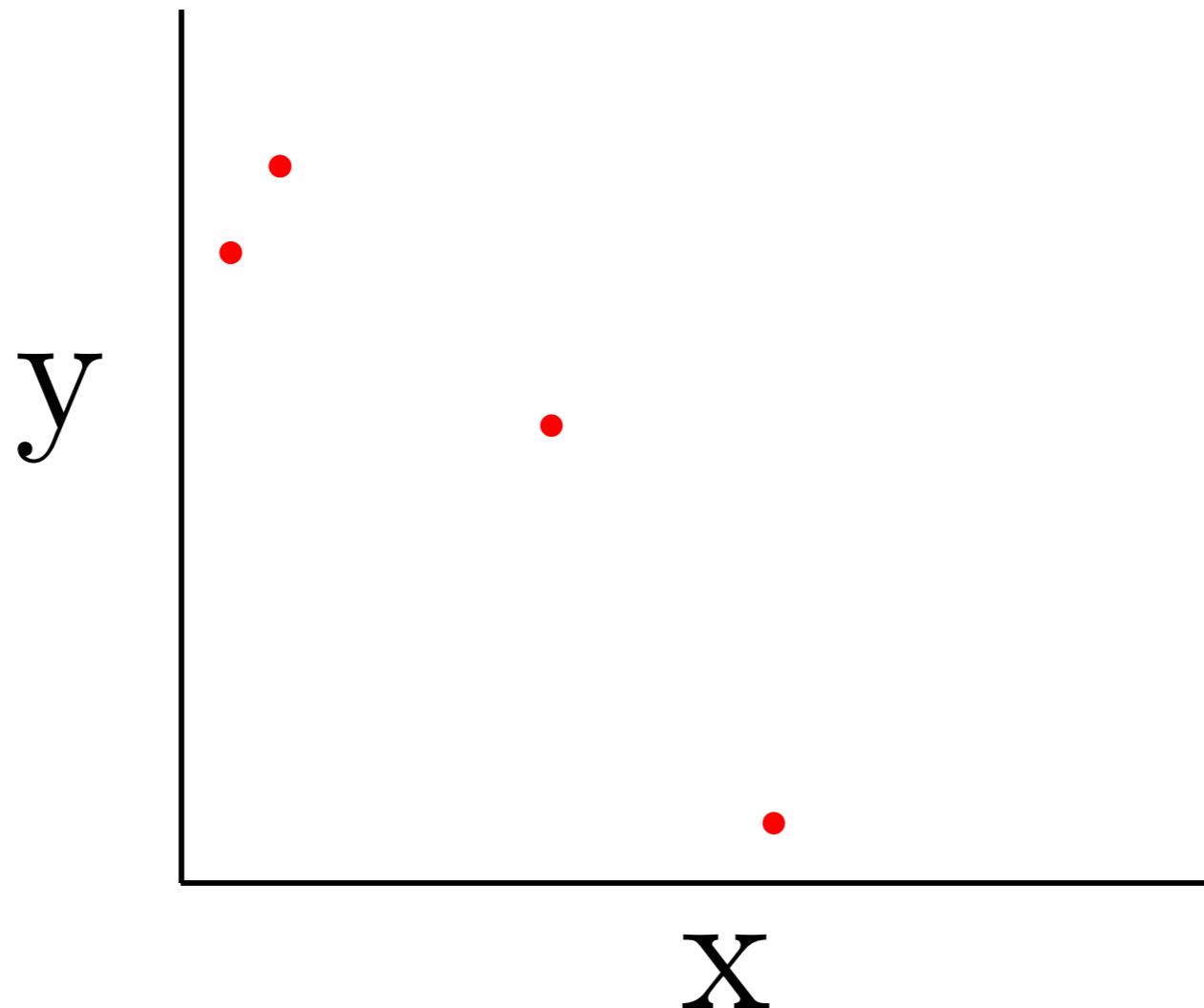
Let  $x_* = \operatorname{argmax}_x f(x)$



We want to **find the point  $x^*$  with as few function evaluations as possible.**

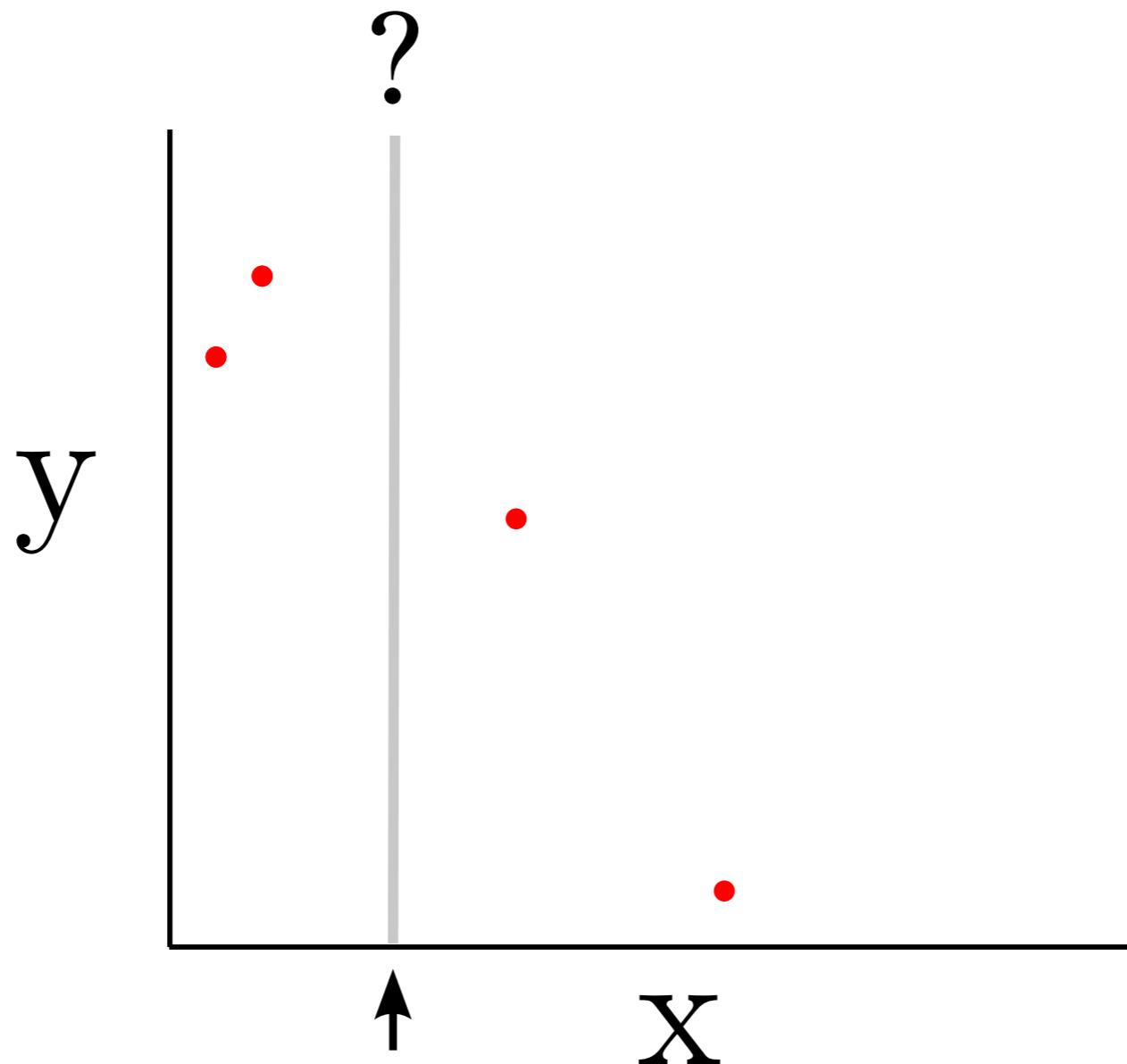
My action space: where in the  $x$  axis I should evaluate the function next.

# Non-linear regression

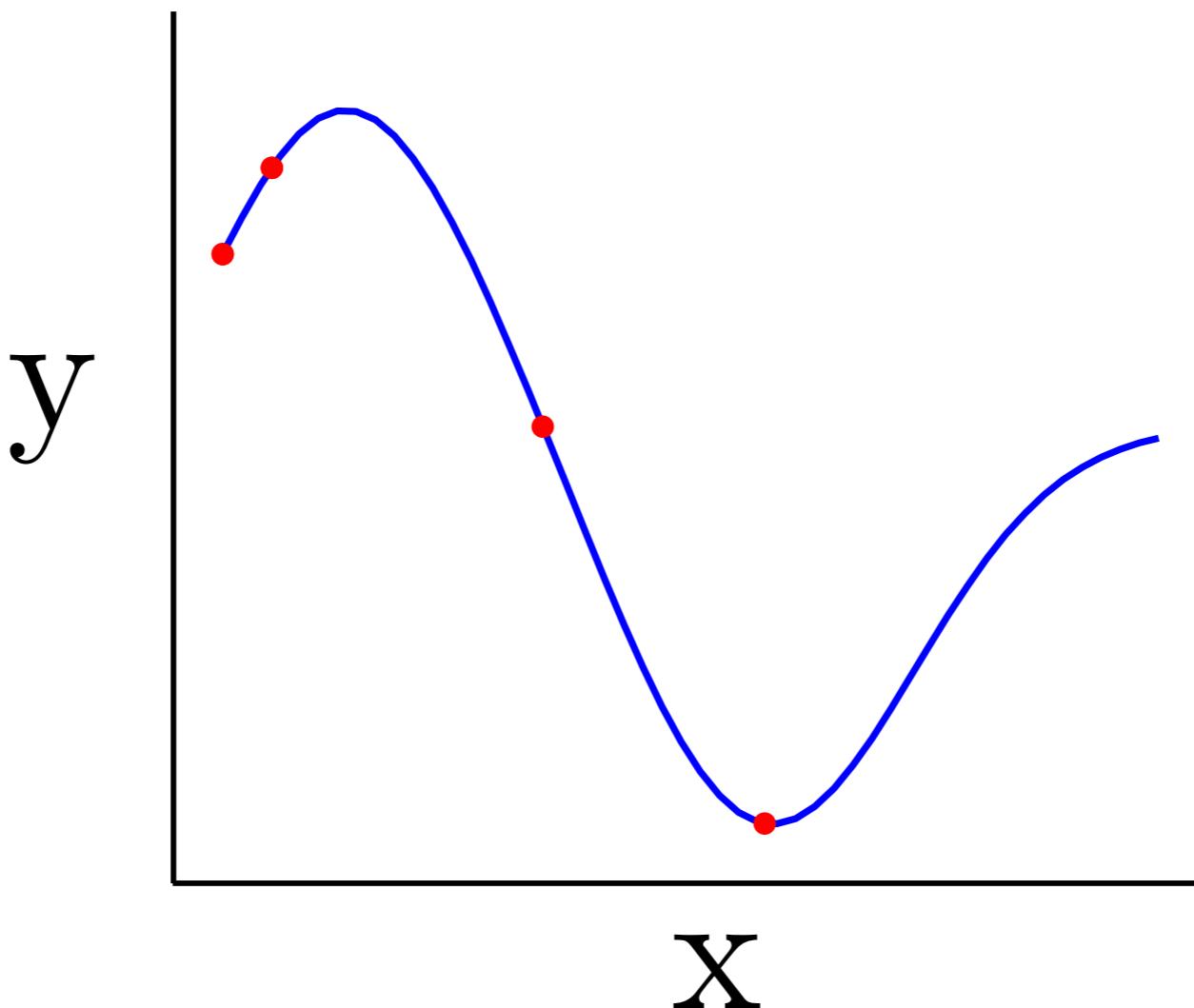


Which x location would you select next?

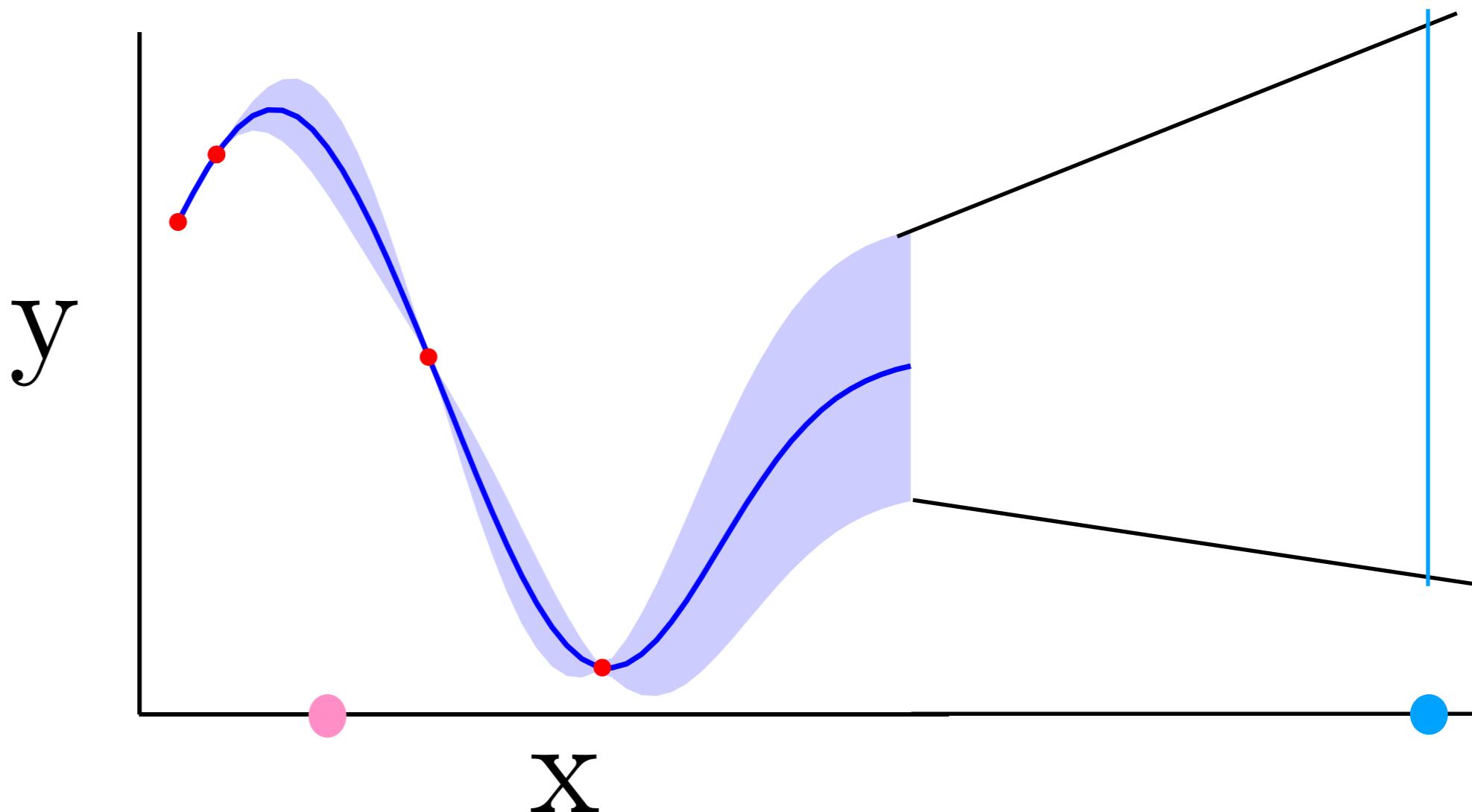
# Non-linear regression



# Non-linear regression



# Non-linear regression with uncertainty



- This point seems the most promising from what I know so far (exploit)
- This point seems the point I am most uncertain about (explore)

Next: Non-linear regression **with error bars** using Gaussian processes

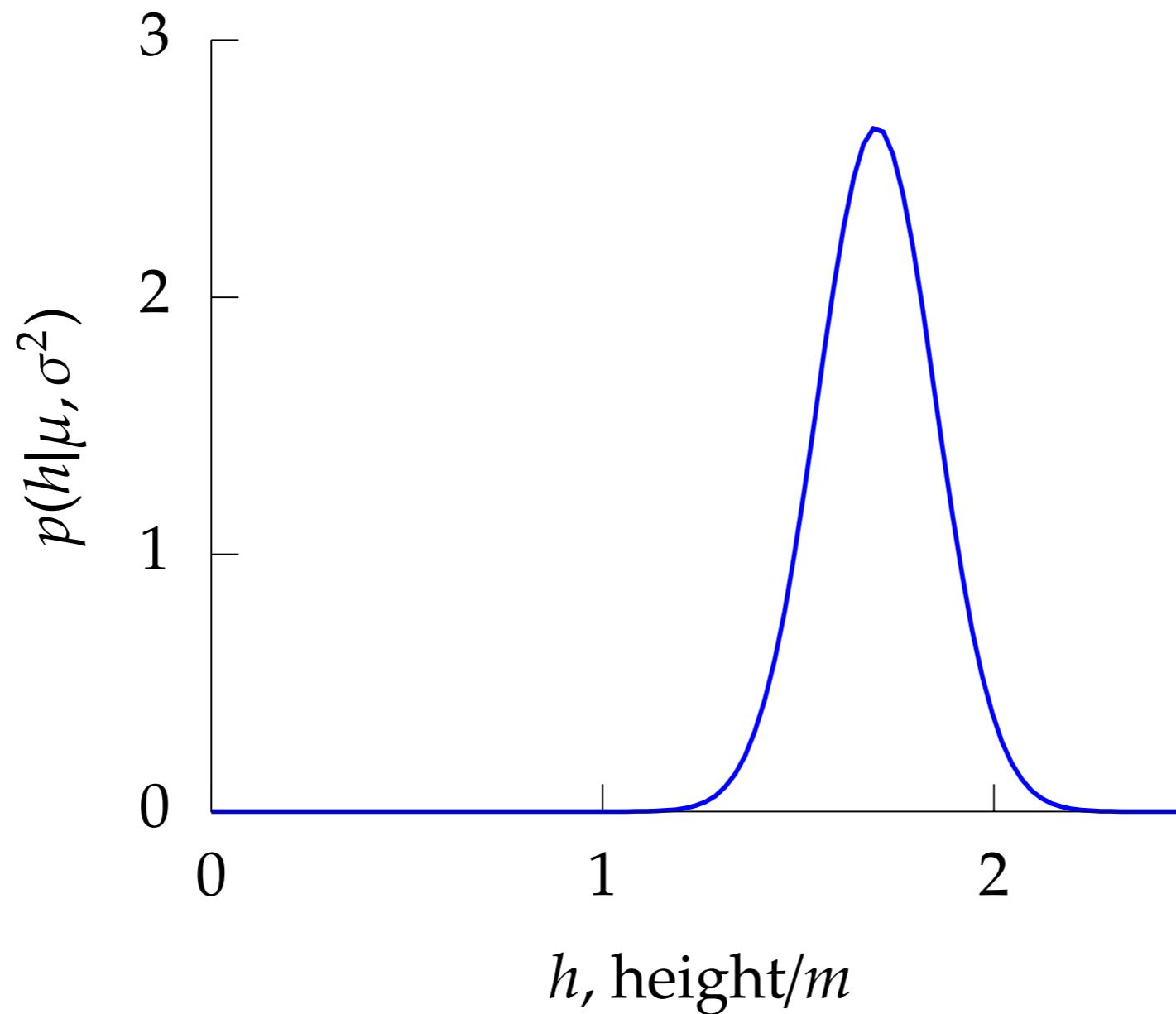
# Gaussian Density

Perhaps the most common probability density

$$\mathcal{N}(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right)$$

$\sigma^2$  is the variance of the density and  $\mu$  is the mean.

# Gaussian Density



Population of students distributed based on their height.

# Two Important Gaussian Properties

## Sum of Gaussians

- ▶ Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

# Two Important Gaussian Properties

## Sum of Gaussians

- ▶ Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

And the sum is distributed as

$$\sum_{i=1}^n y_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

# Two Important Gaussian Properties

## Sum of Gaussians

- ▶ Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

And the sum is distributed as

$$\sum_{i=1}^n y_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

(*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

# Two Important Gaussian Properties

## Scaling a Gaussian

- ▶ Scaling a Gaussian leads to a Gaussian.

# Two Important Gaussian Properties

## Scaling a Gaussian

- ▶ Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}(\mu, \sigma^2)$$

# Two Important Gaussian Properties

## Scaling a Gaussian

- ▶ Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}(\mu, \sigma^2)$$

And the scaled density is distributed as

$$wy \sim \mathcal{N}(w\mu, w^2\sigma^2)$$

# Multivariate Consequence

- ▶ If

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

# Multivariate Consequence

- ▶ If

$$\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$$

- ▶ And

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

# Multivariate Consequence

- ▶ If

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- ▶ And

$$\mathbf{y} = \mathbf{W}\mathbf{x}$$

- ▶ Then

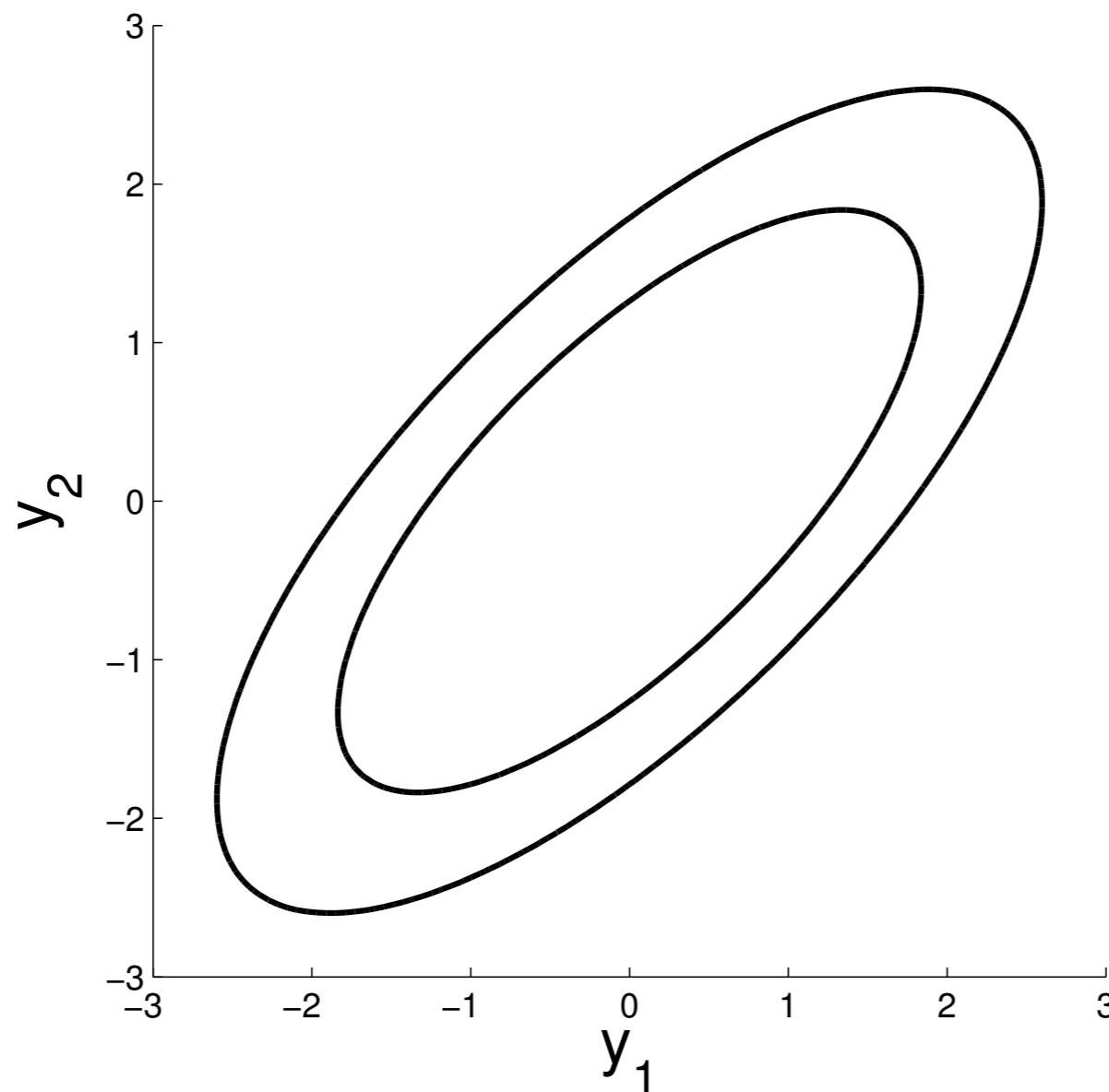
$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}\boldsymbol{\mu}, \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^\top)$$

# Gaussian Distribution

$$\Sigma_{i,j} = \mathbb{E} [(y_i - \mathbb{E}(y_i))(y_j - \mathbb{E}(y_j))]$$

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

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



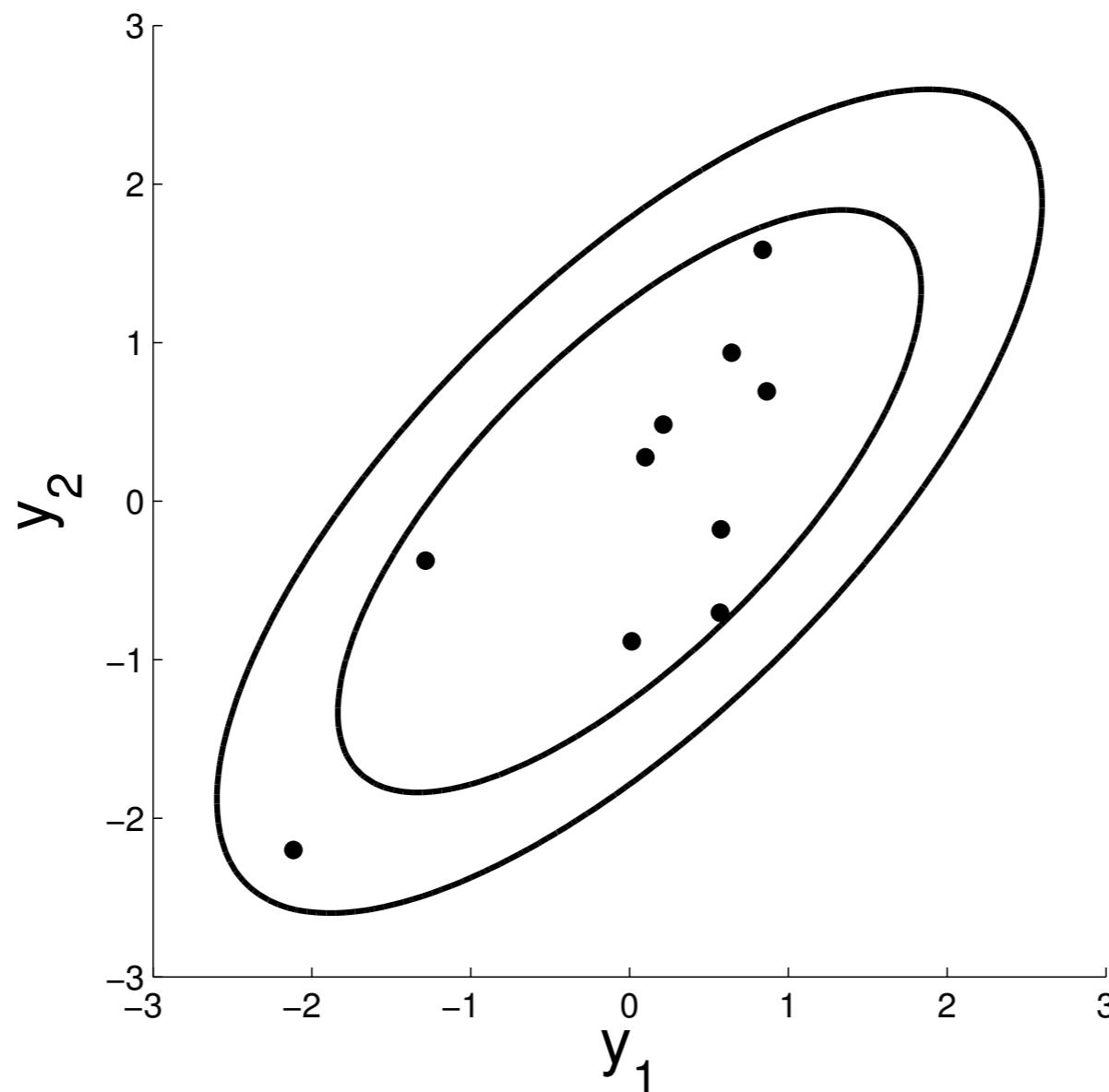
$y_i$  : scalar random variable  
 $\mathbf{y}$  : vector random variable

# Gaussian Distribution

$$\Sigma_{i,j} = \mathbb{E} [(y_i - \mathbb{E}(y_i))(y_j - \mathbb{E}(y_j))]$$

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

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

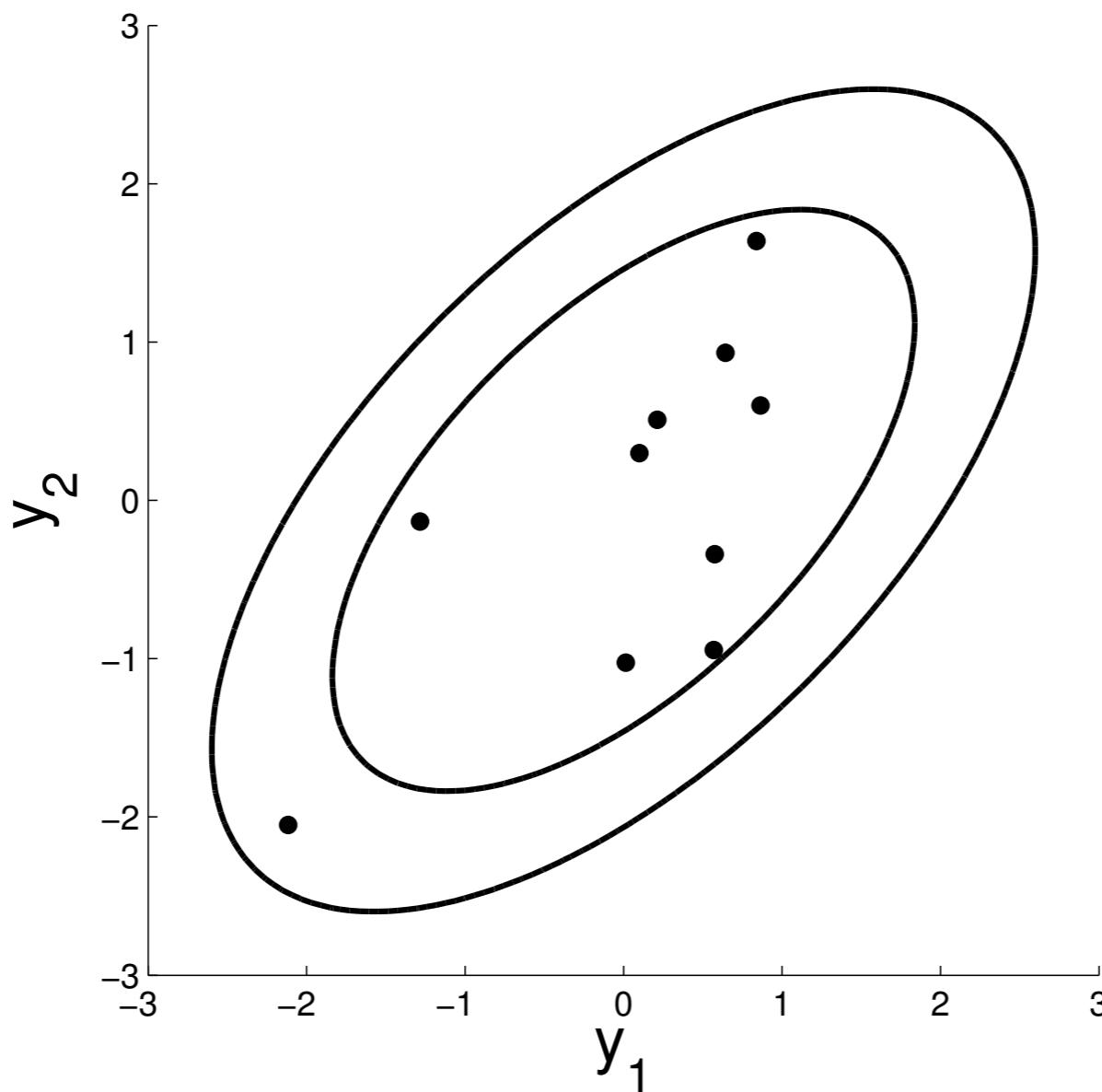


$y_i$  : scalar random variable  
 $\mathbf{y}$  : vector random variable

# Gaussian Distribution

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

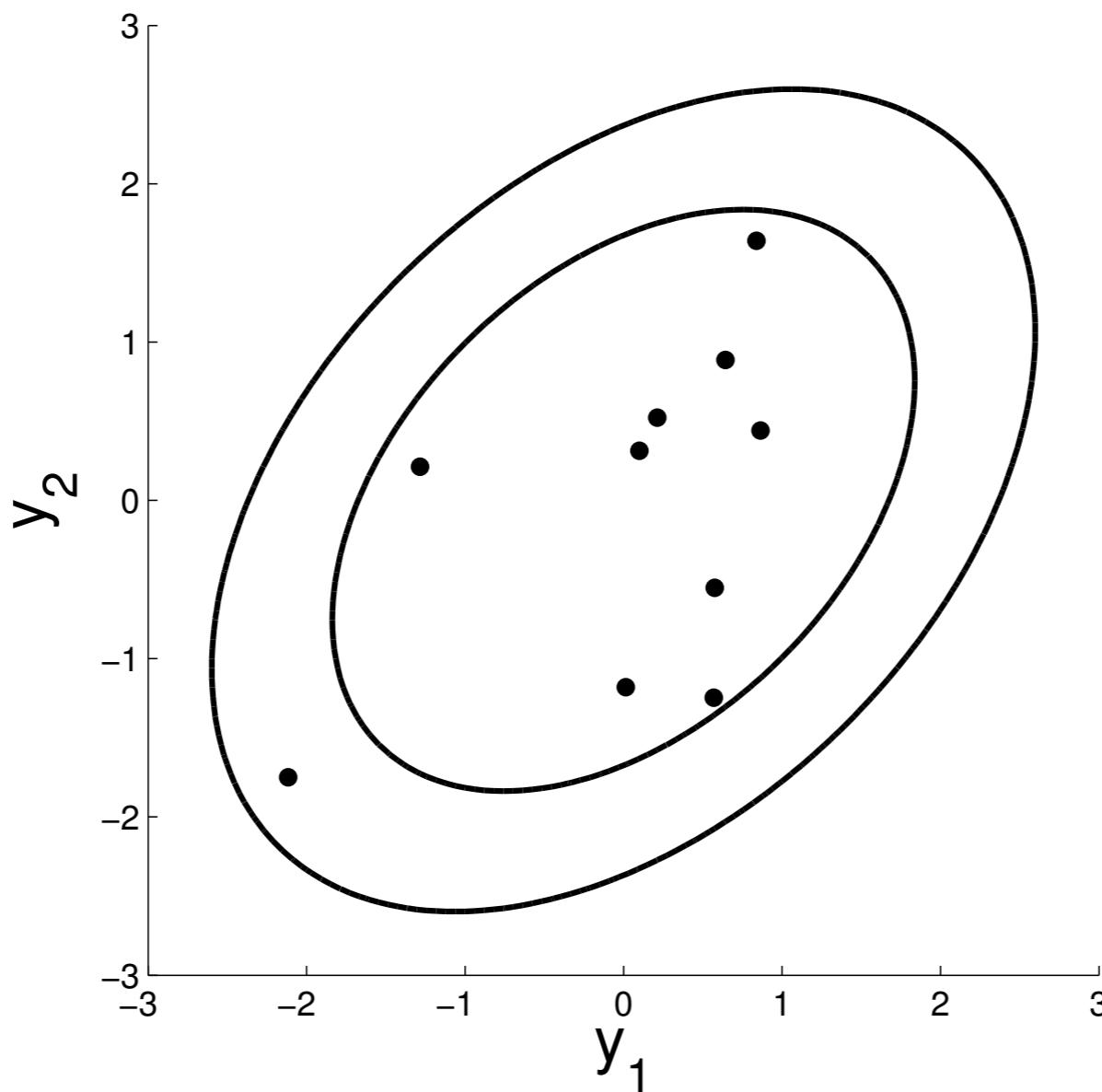
$$\Sigma = \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}$$



# Gaussian Distribution

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

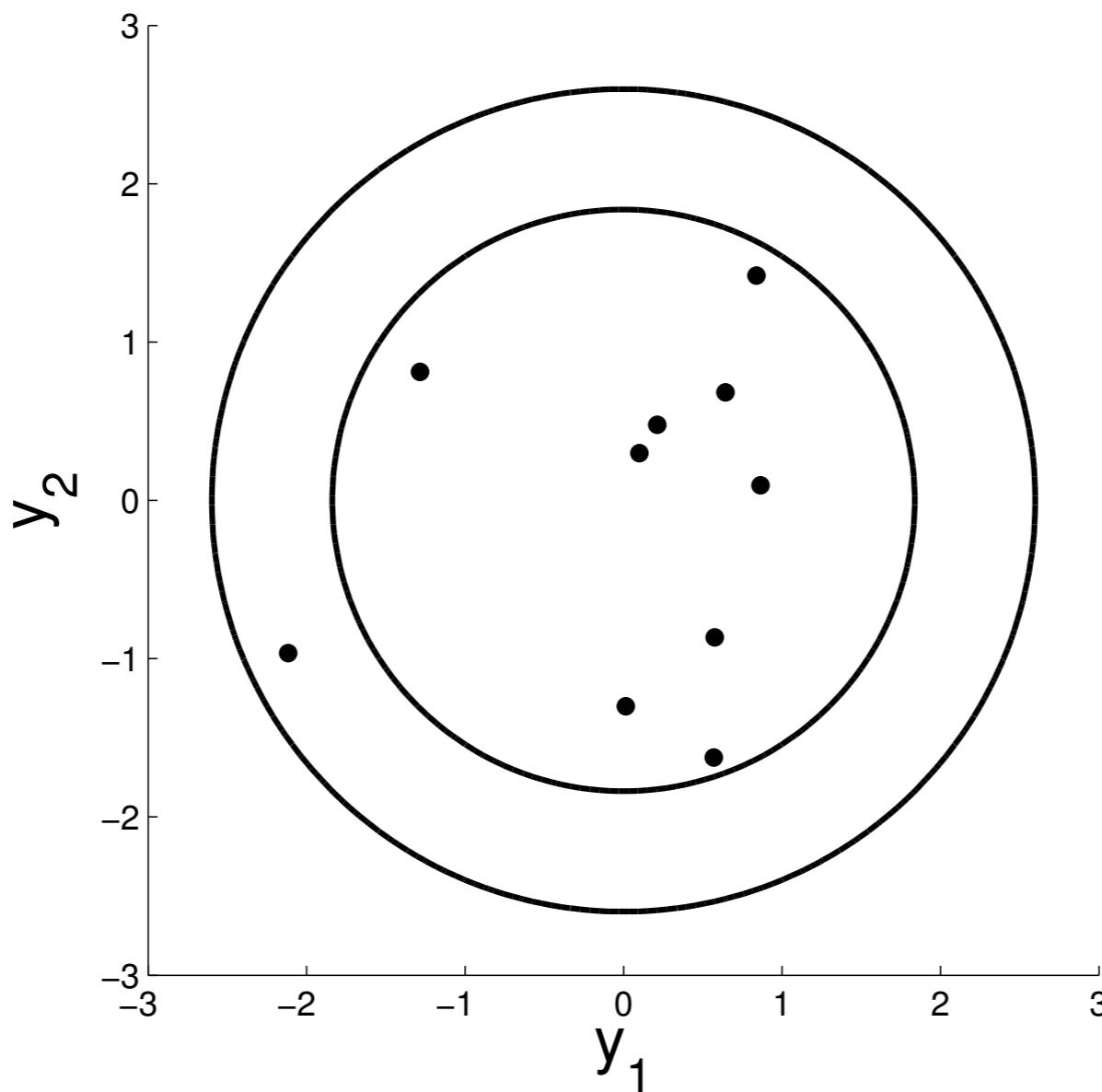
$$\Sigma = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$$



# Gaussian Distribution

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

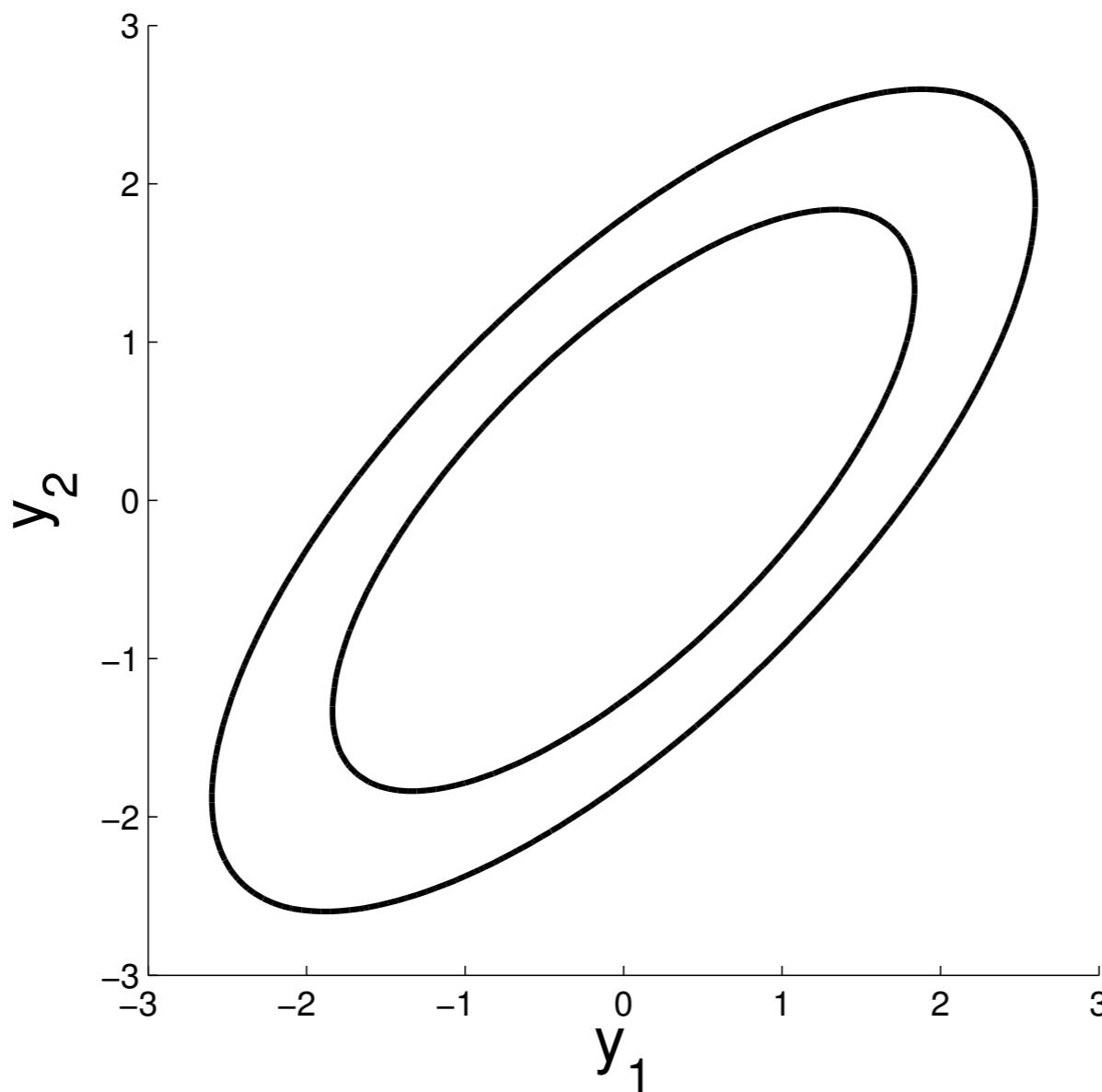
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



# Gaussian Distribution

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

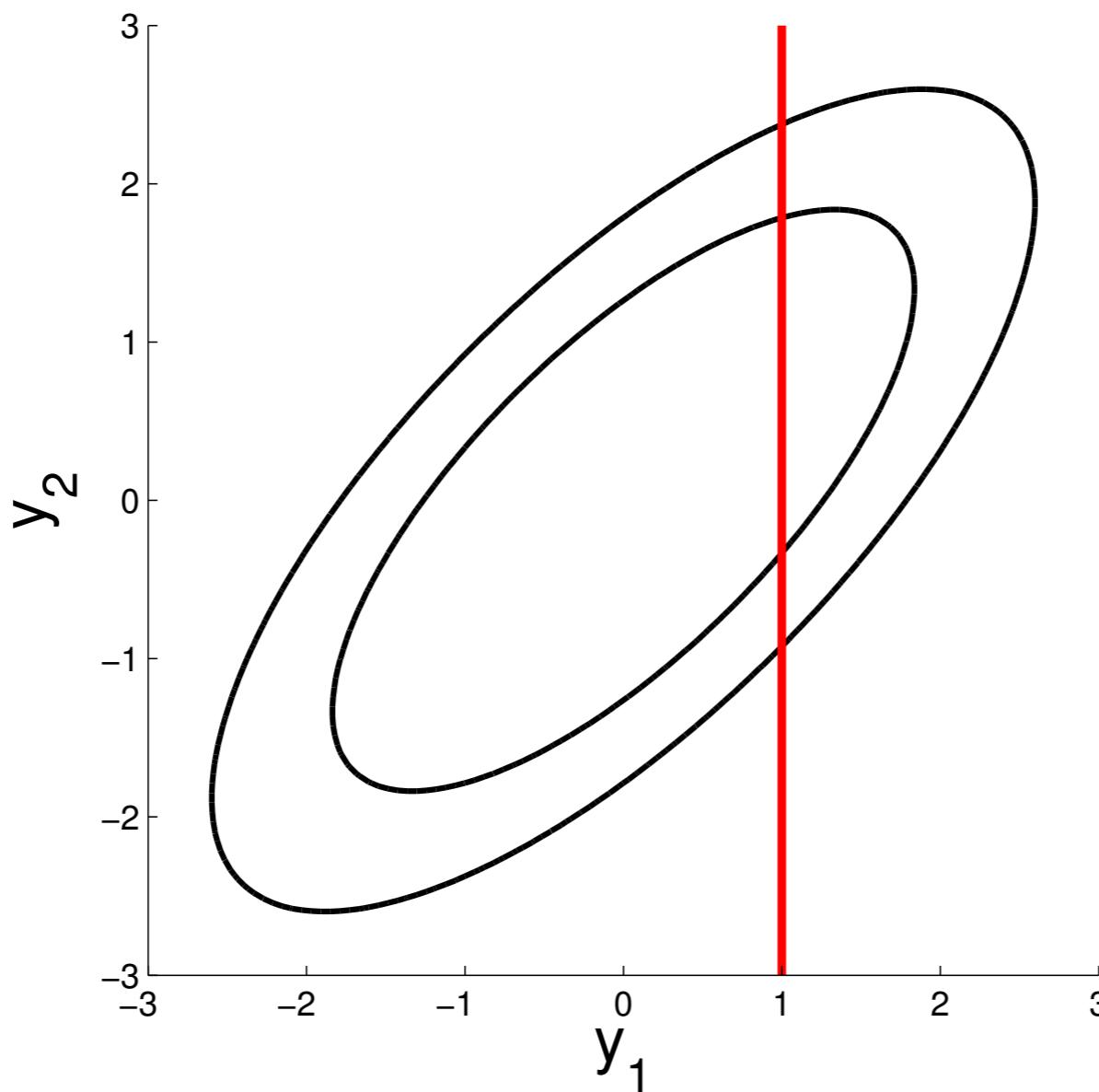
$$\Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$$



# Gaussian distribution - Conditioning

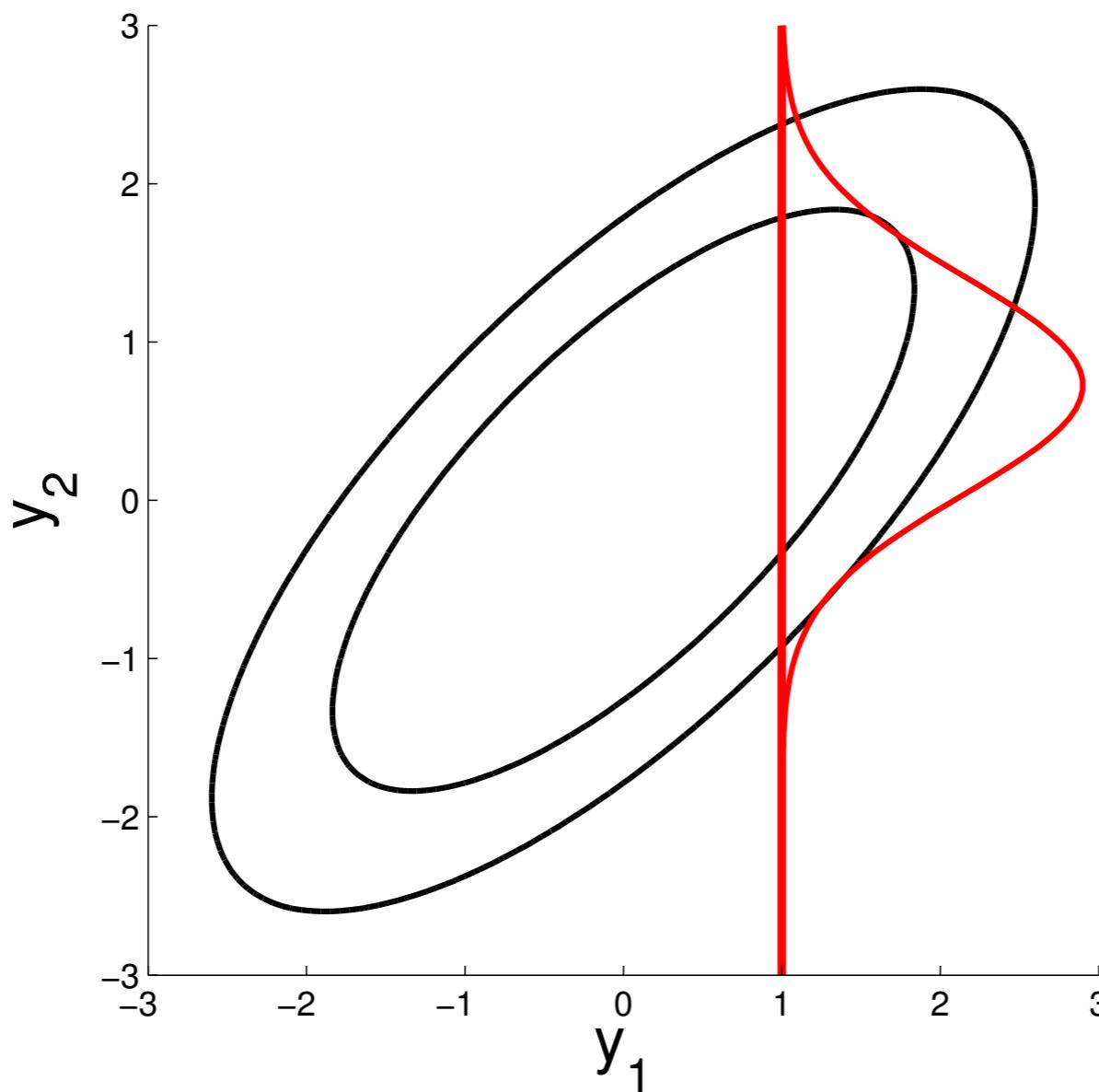
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^\top \Sigma^{-1} \mathbf{y}\right)$$

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



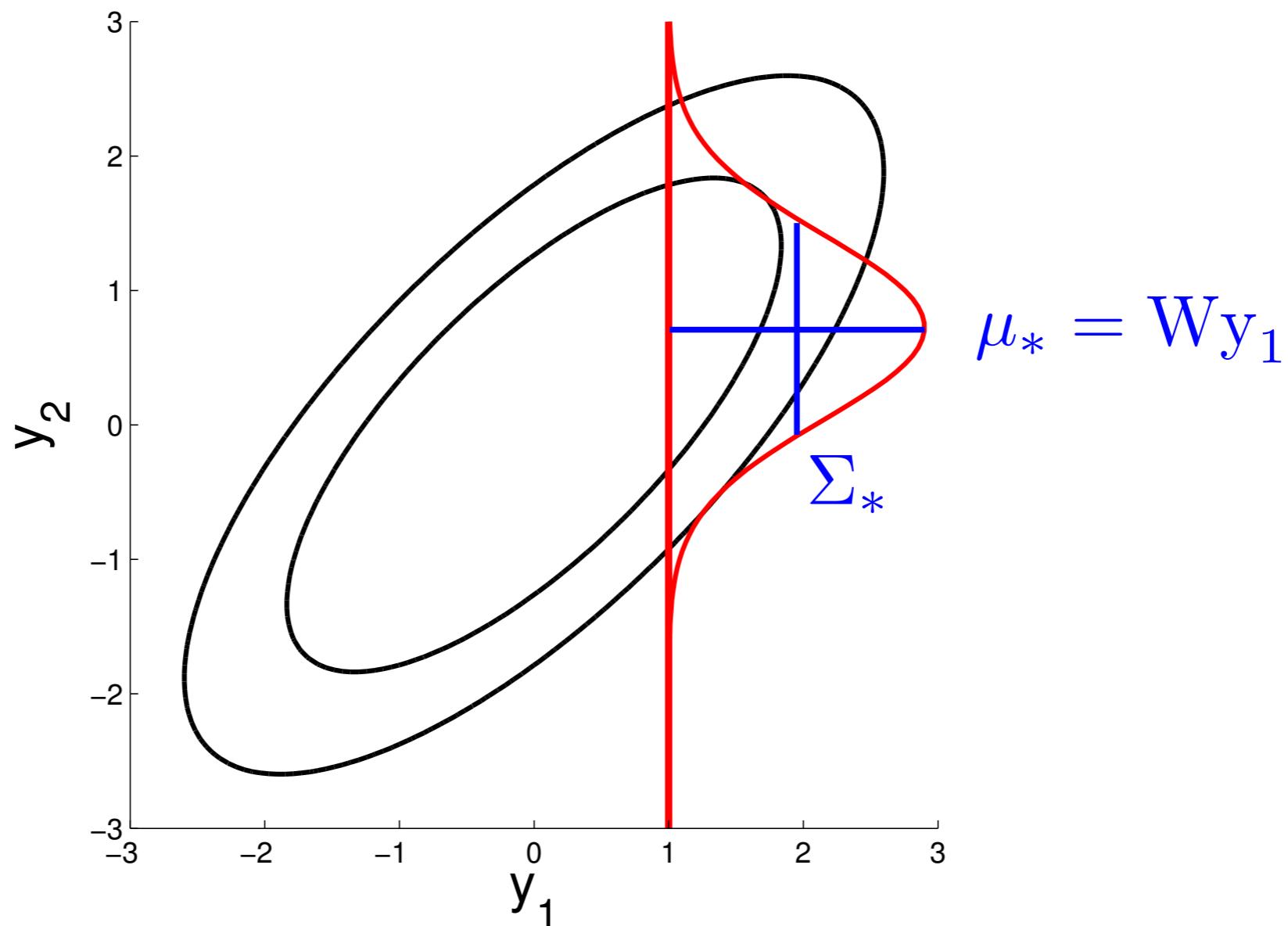
# Gaussian distribution - Conditioning

$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\frac{1}{2}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\right)$$



# Gaussian distribution - Conditioning

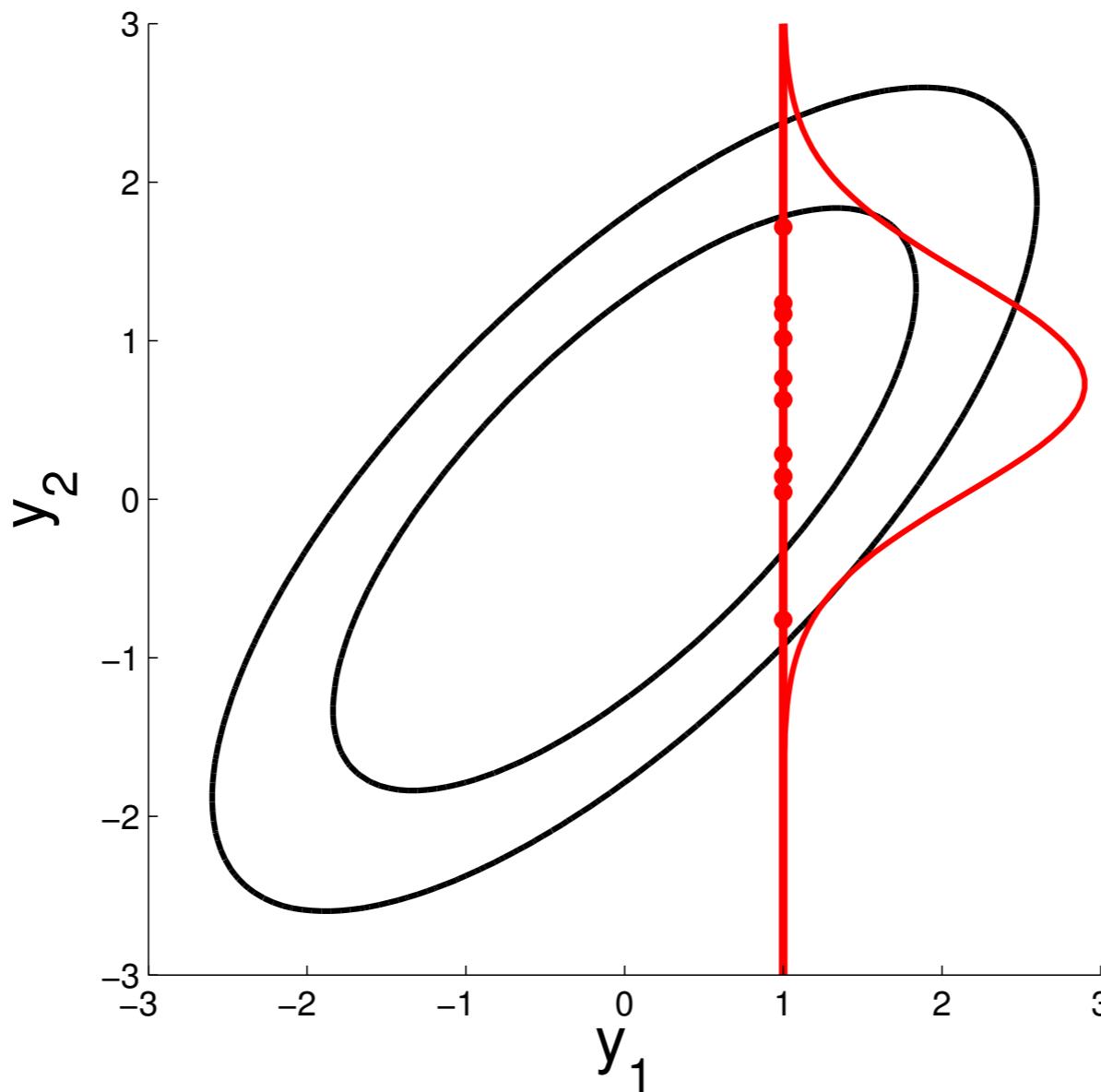
$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\frac{1}{2}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\right)$$



There are closed form solutions for  $\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*$

# Gaussian distribution - Conditioning

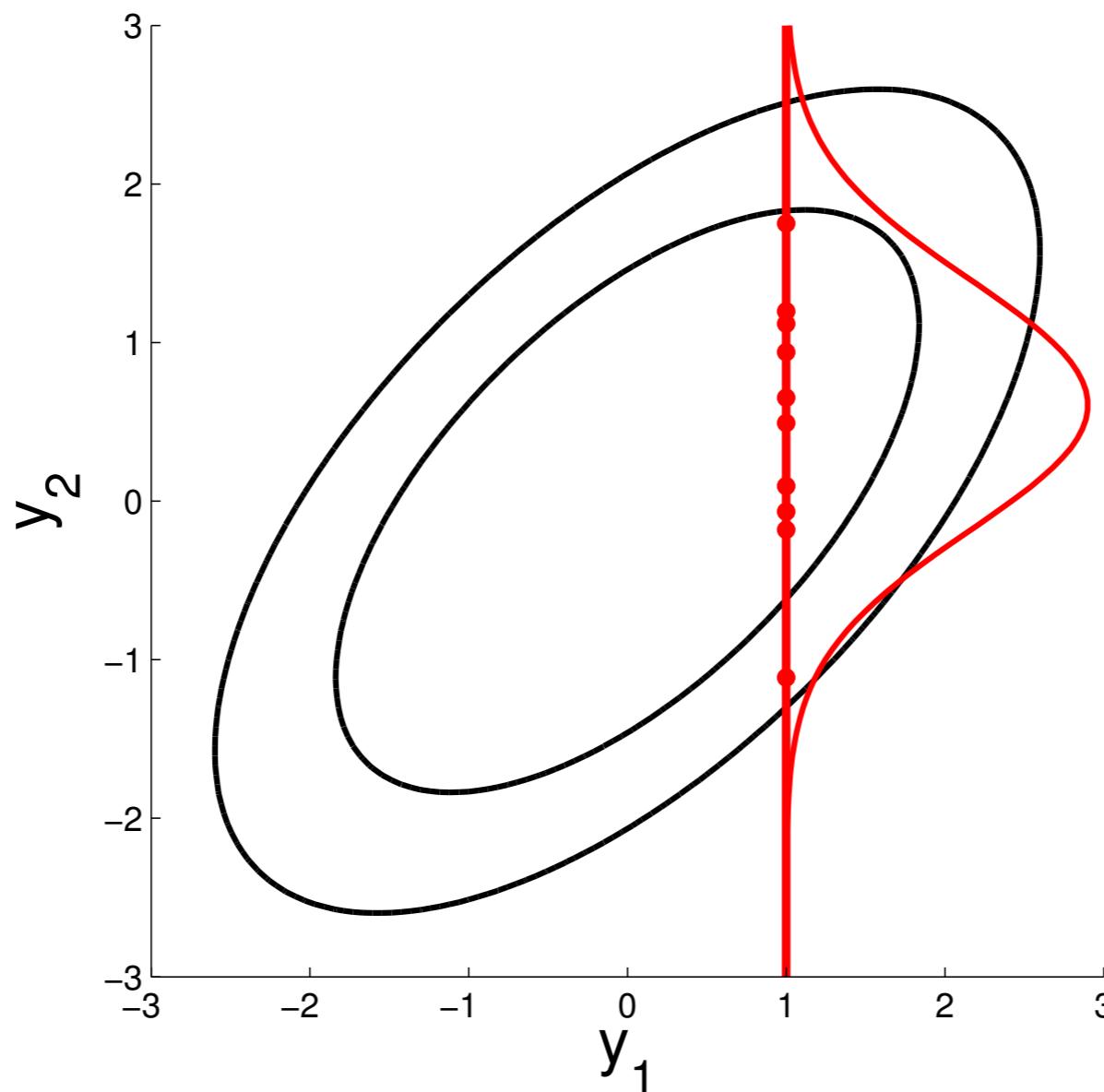
$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\frac{1}{2}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\right)$$



There are closed form solutions for  $\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*$

# Gaussian distribution - Conditioning

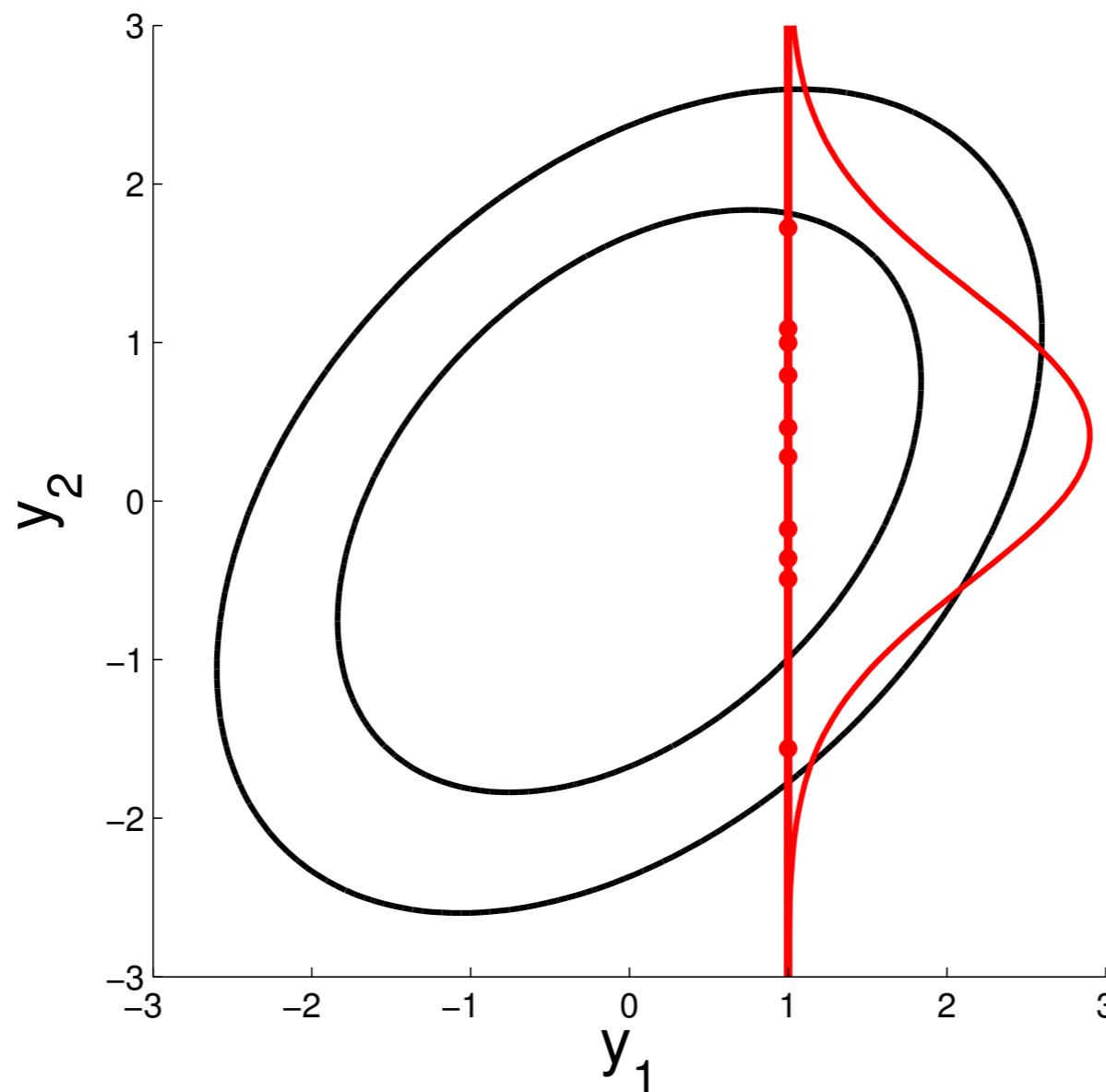
$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\frac{1}{2}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\right)$$



There are closed form solutions for  $\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*$

# Gaussian distribution - Conditioning

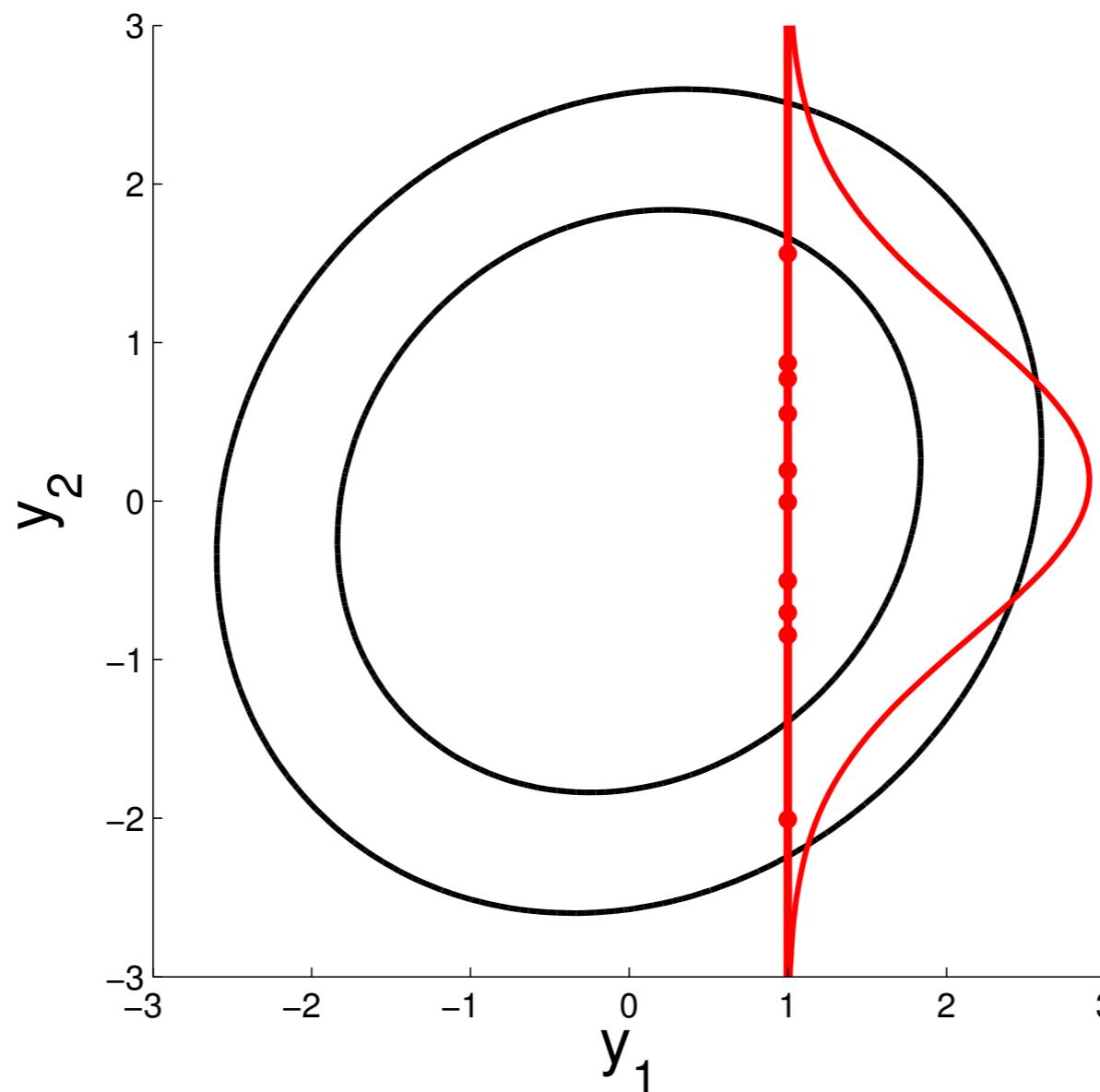
$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\frac{1}{2}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\right)$$



There are closed form solutions for  $\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*$

# Gaussian distribution - Conditioning

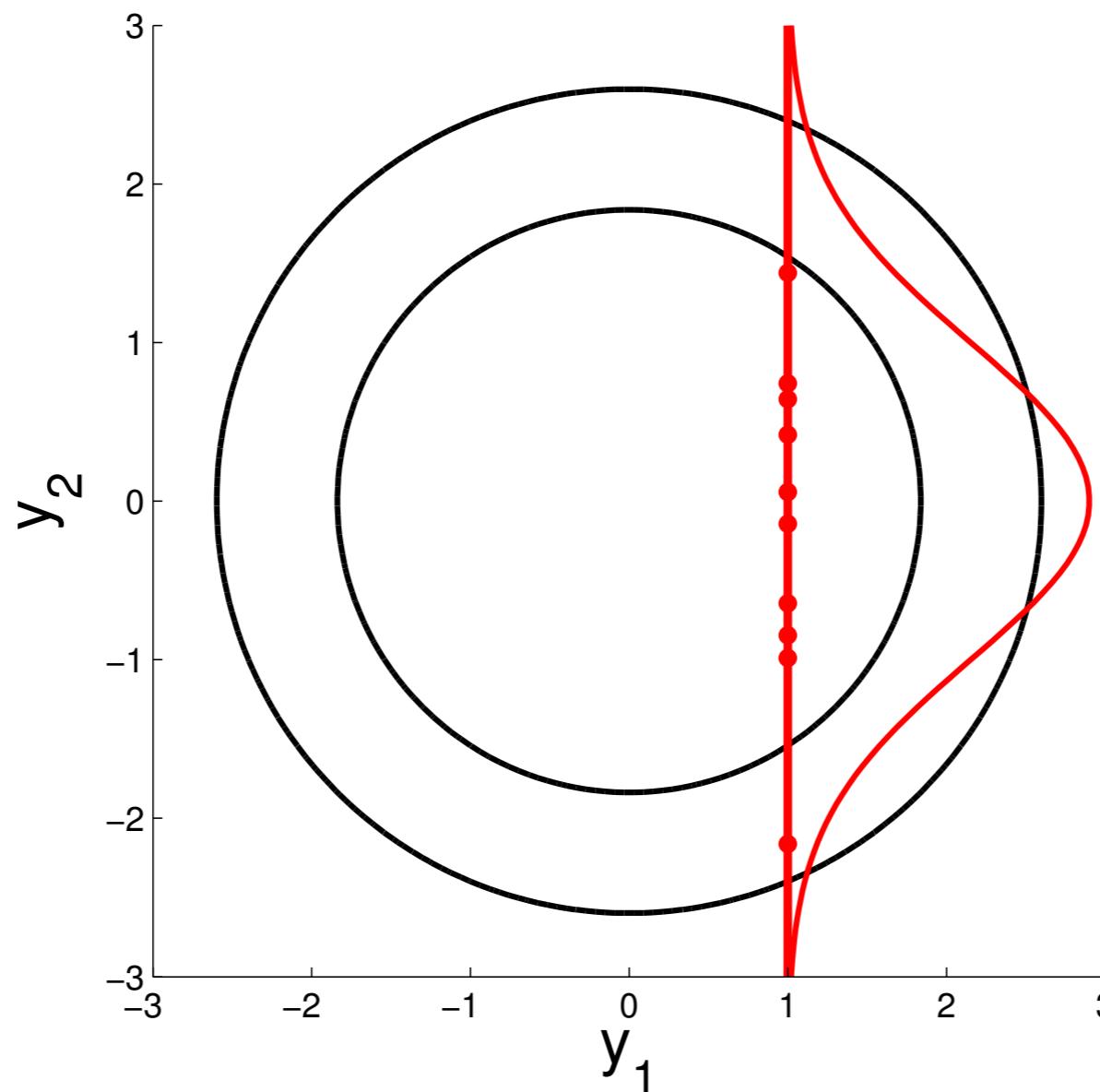
$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\frac{1}{2}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\right)$$



There are closed form solutions for  $\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*$

# Gaussian distribution - Conditioning

$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\frac{1}{2}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\boldsymbol{\Sigma}_*^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_*)\right)$$



There are closed form solutions for  $\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*$

# Multivariate Gaussian Theorem

**Theorem 4.2.1** (Marginals and conditionals of an MVN). Suppose  $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$  is jointly Gaussian with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} \quad (4.12)$$

Then the marginals are given by

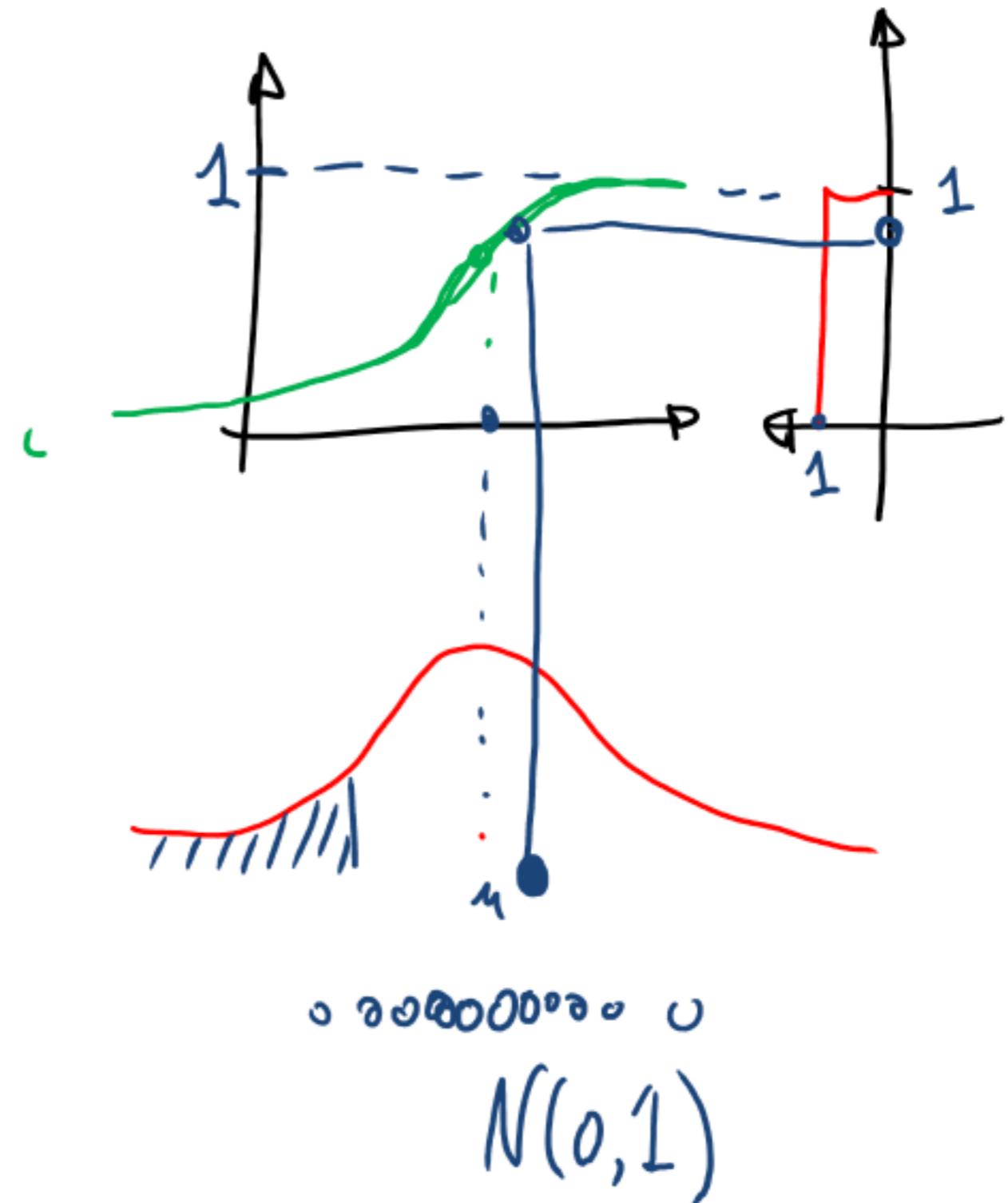
$$\begin{aligned} \rightarrow p(\mathbf{x}_1) &= \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ p(\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \end{aligned}$$

and the posterior conditional is given by

$$\begin{aligned} p(\mathbf{x}_1 | \mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1} \boldsymbol{\Lambda}_{12} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\Sigma}_{1|2} (\boldsymbol{\Lambda}_{11} \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12} (\mathbf{x}_2 - \boldsymbol{\mu}_2)) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1} \end{aligned}$$

# Sampling from a Gaussian density

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

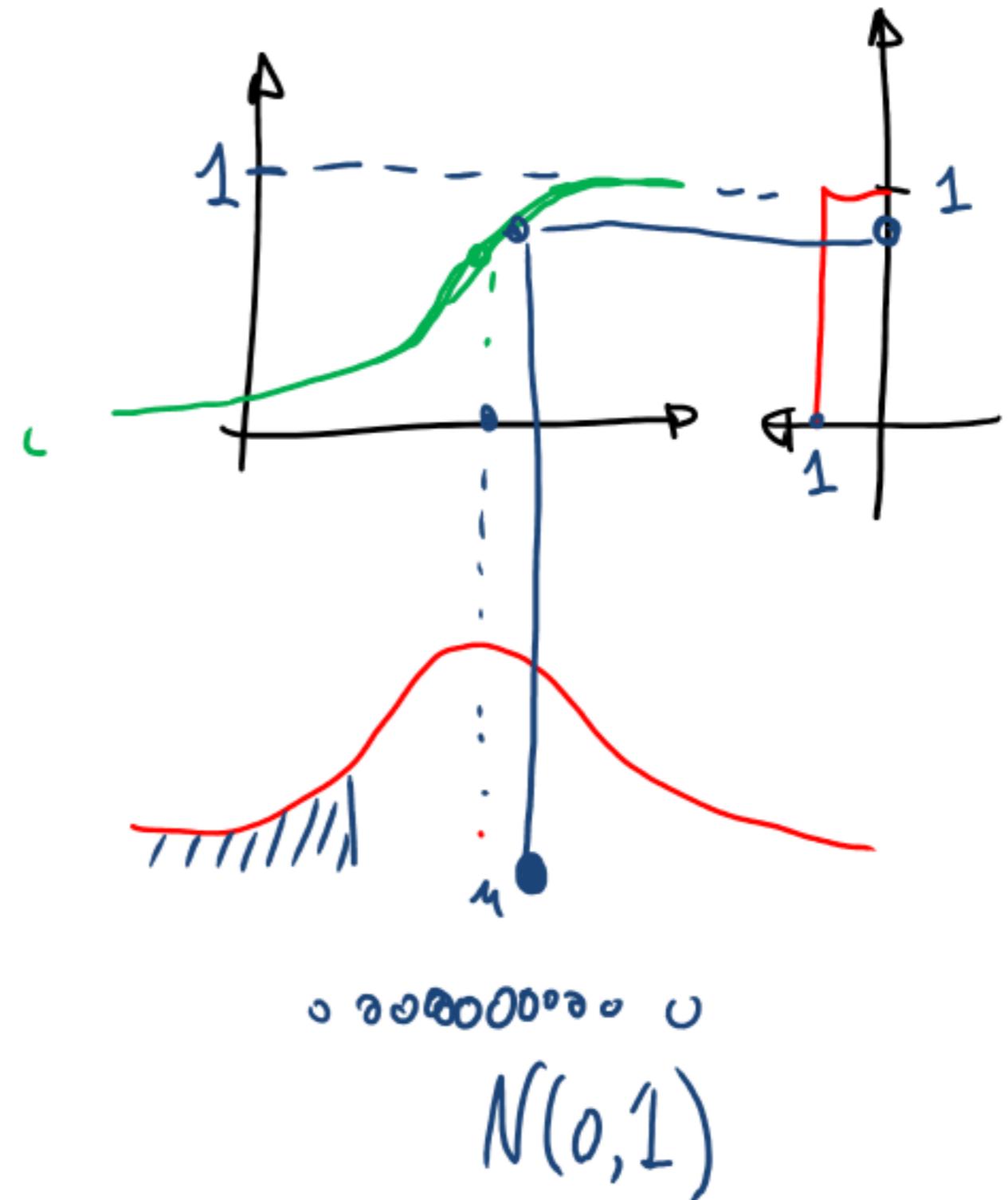


# Sampling from a Gaussian density

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

$x_i \sim \mathcal{N}(\mu, \sigma^2)$

$\sim \mu + \sigma \mathcal{N}(0,1)$



# Sampling from a Gaussian density

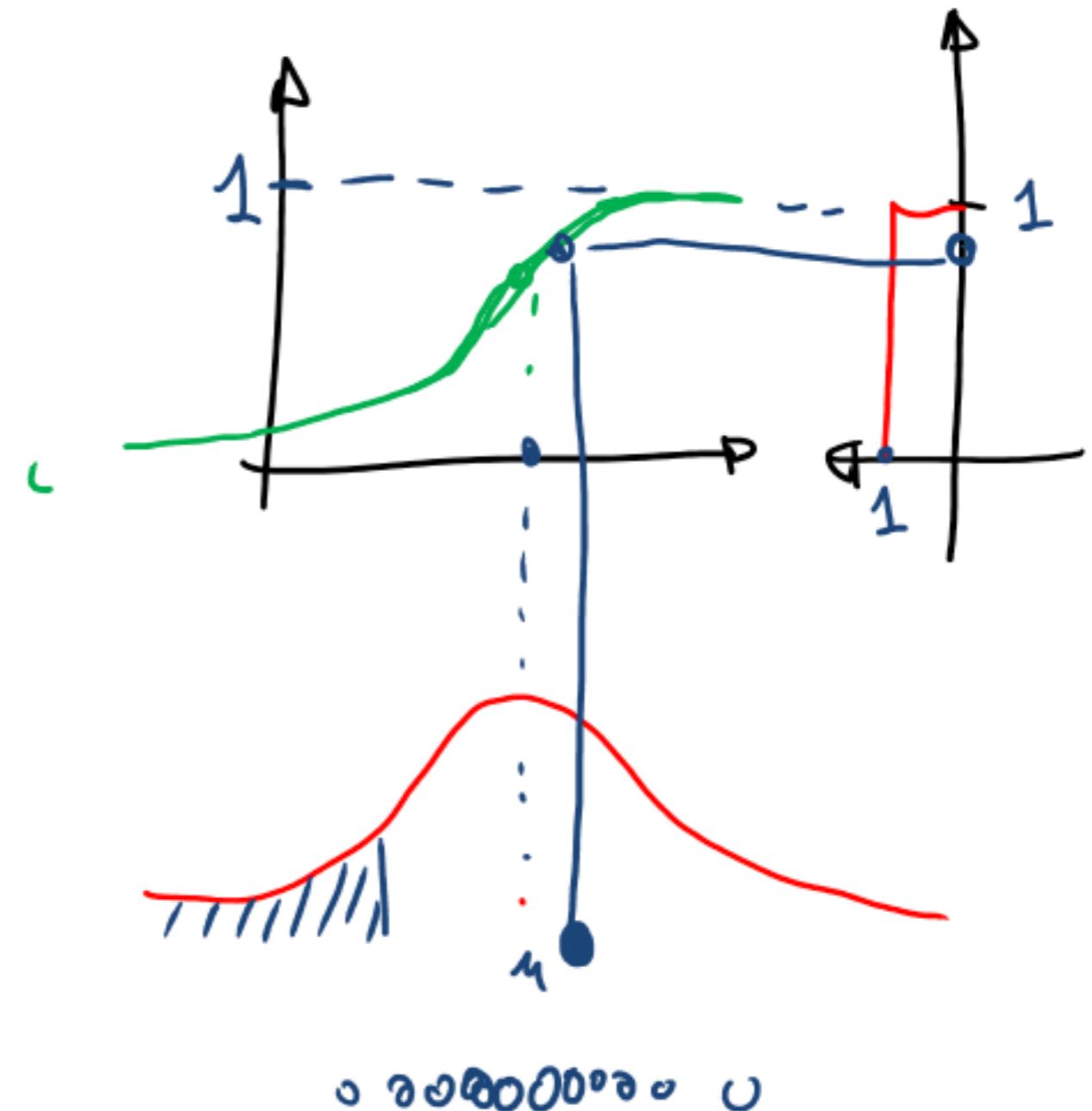
$$x_i \sim \mathcal{N}(0,1)$$

$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$\sim \mu + \sigma \mathcal{N}(0,1)$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_i \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_i \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$



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

# Sampling from a Gaussian density

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

$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$

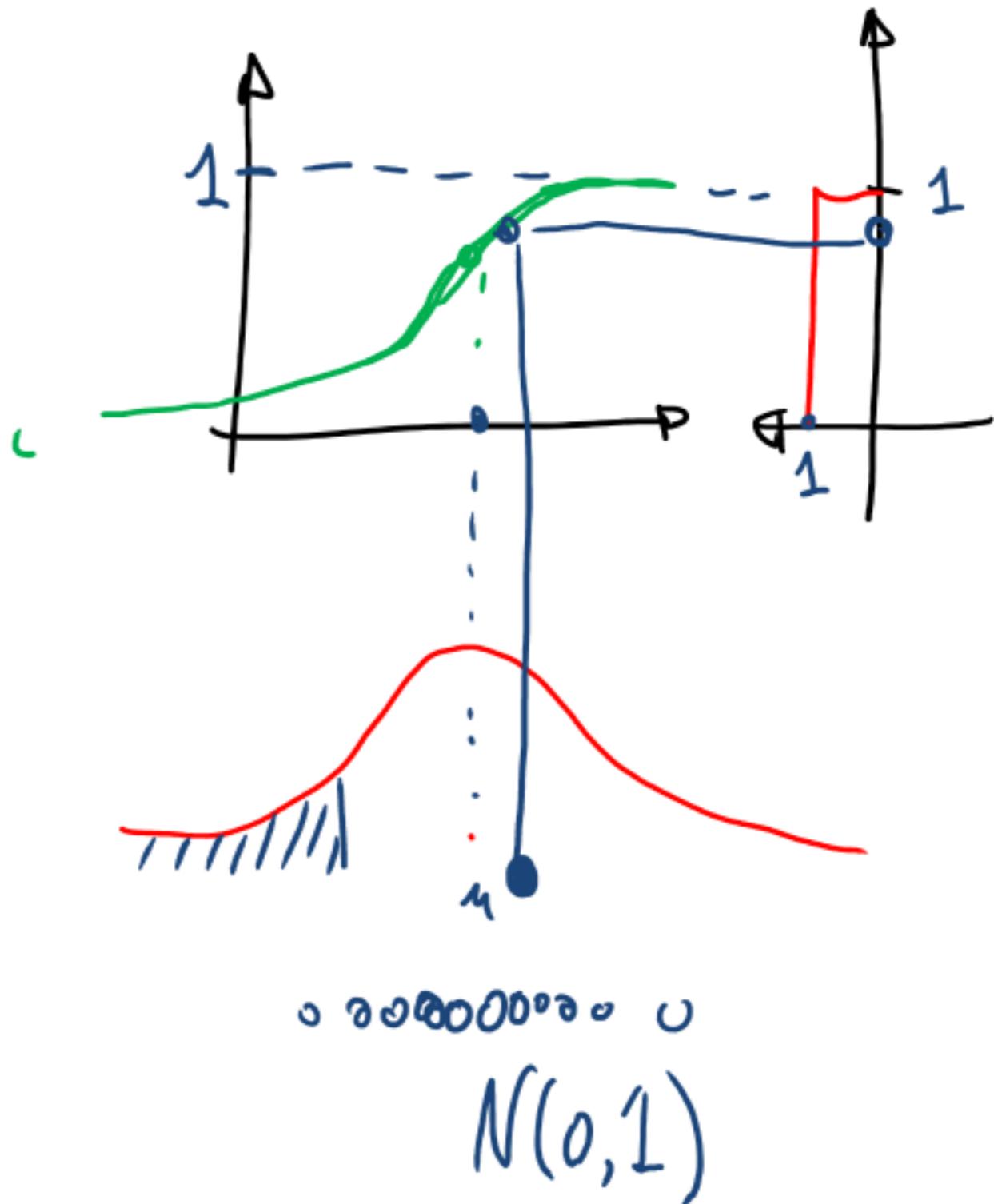
$$\sim \mu + \sigma \mathcal{N}(0,1)$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_i \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_i \sim \mathcal{N} \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

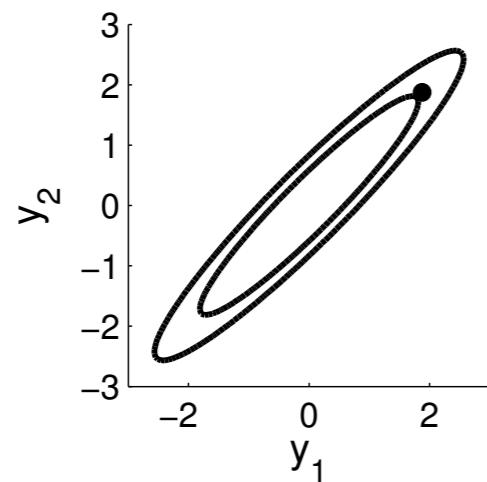
$$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$$

$$x \sim \mu + L\mathcal{N}(0,I)$$



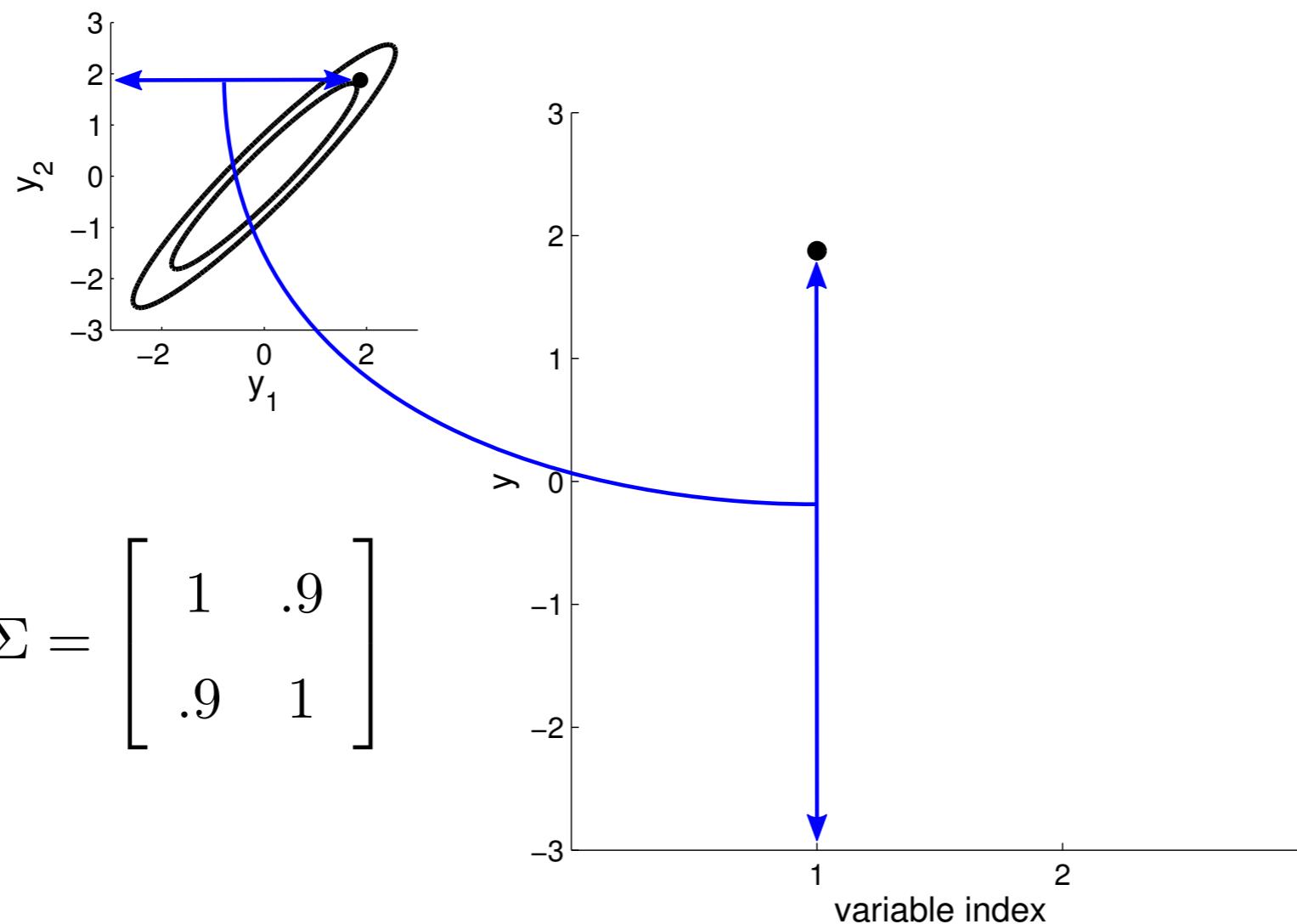
# Cholesky decomposition $\Sigma = LL^T$

# Towards higher dimensional Gaussians - New Visualisation

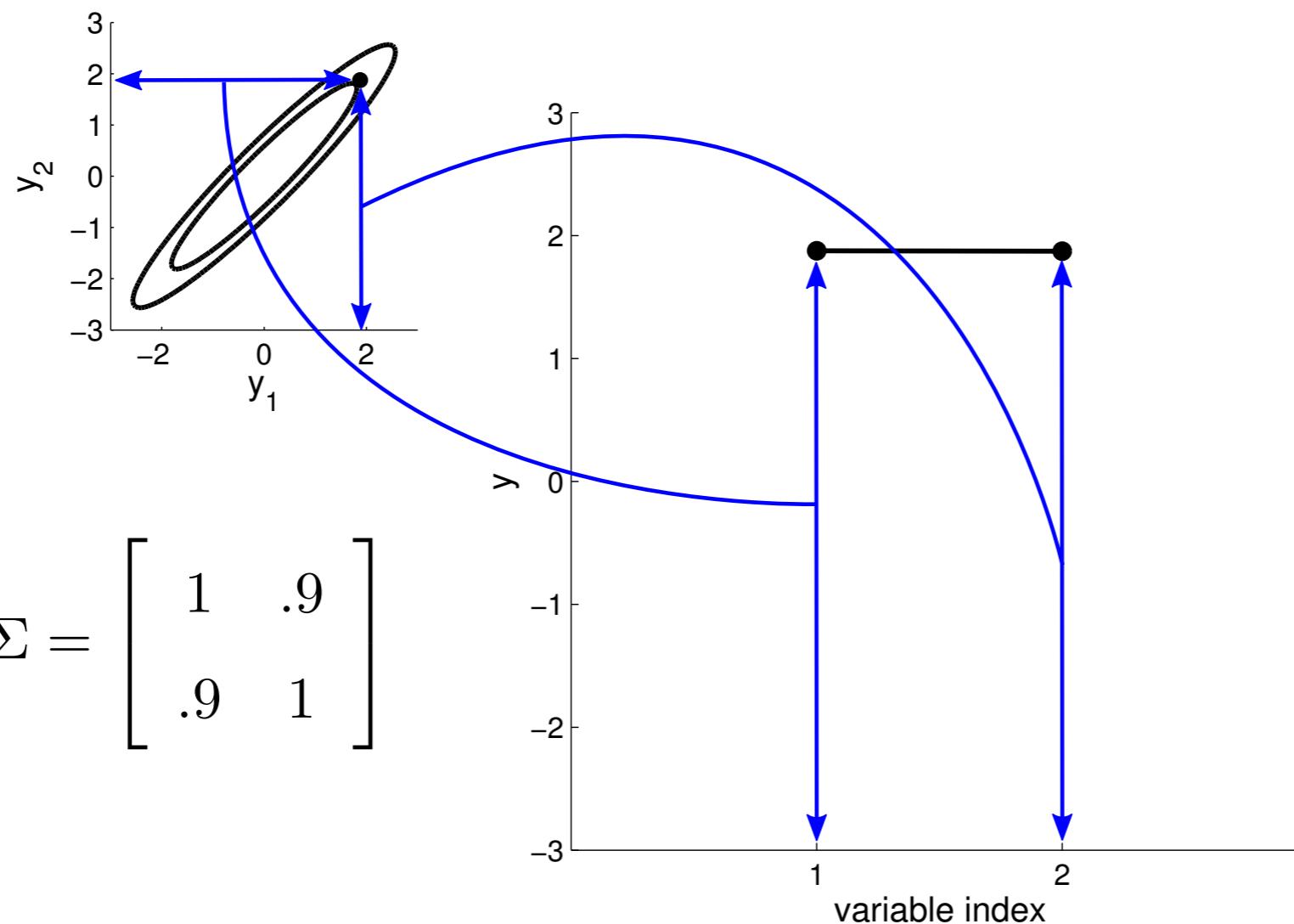


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

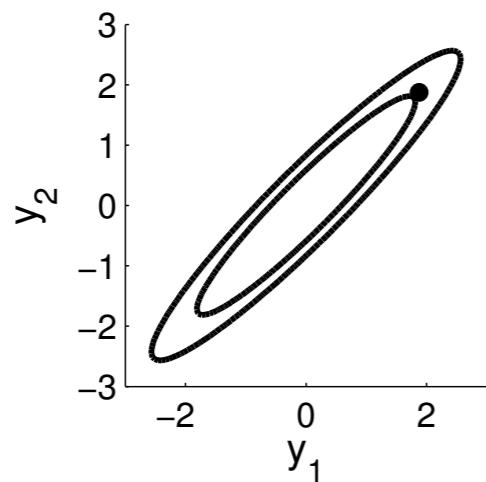
# New Visualisation



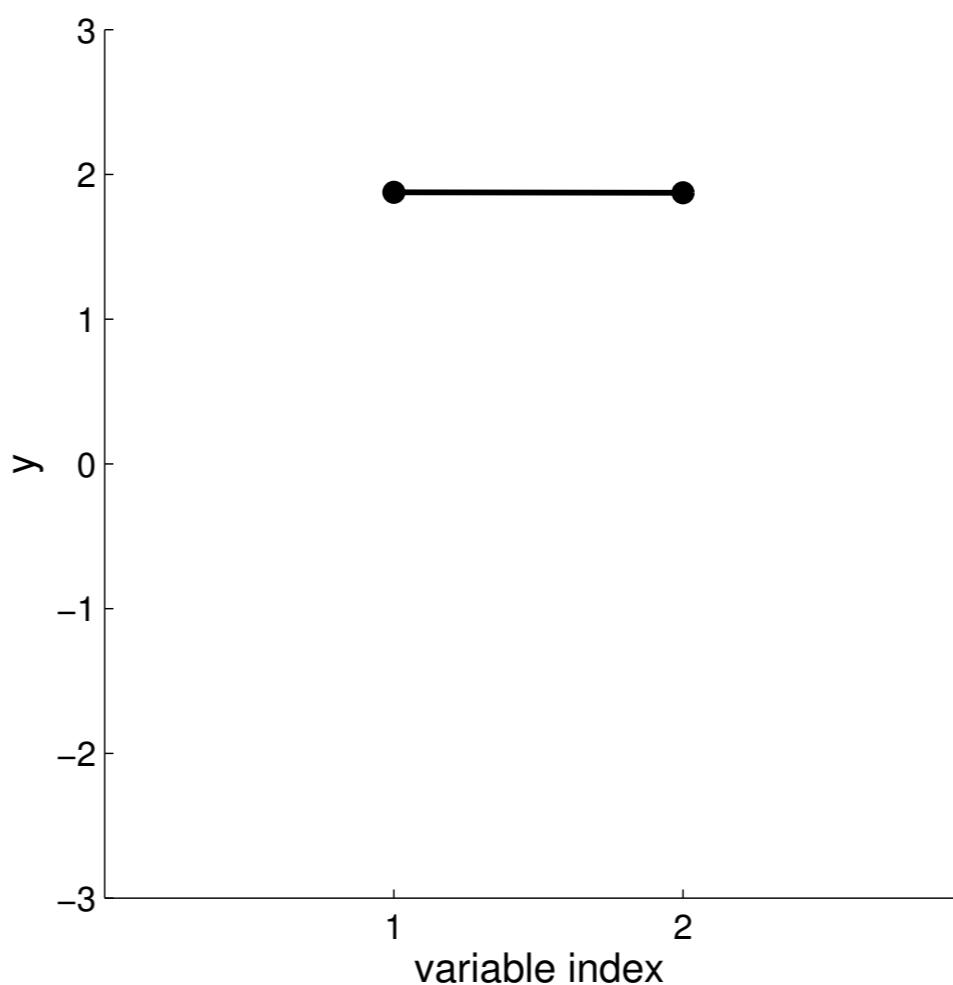
# New Visualisation



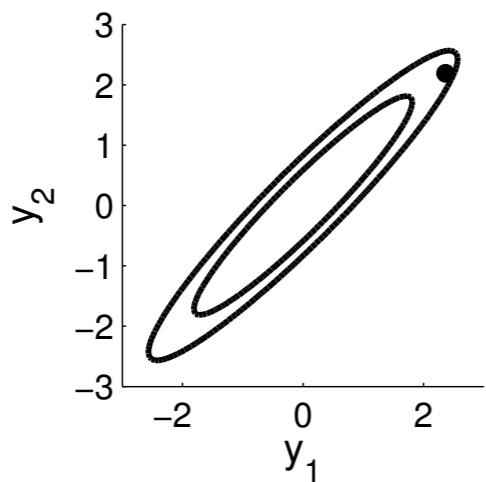
# New Visualisation



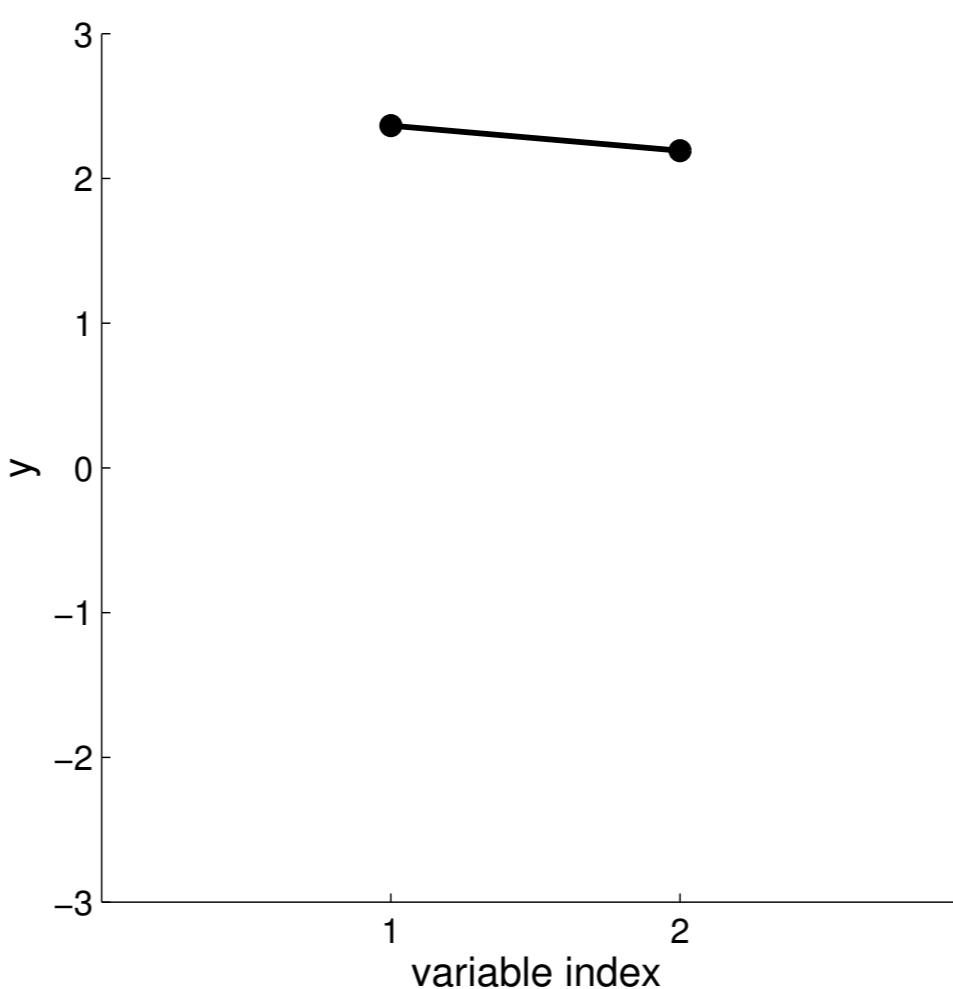
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



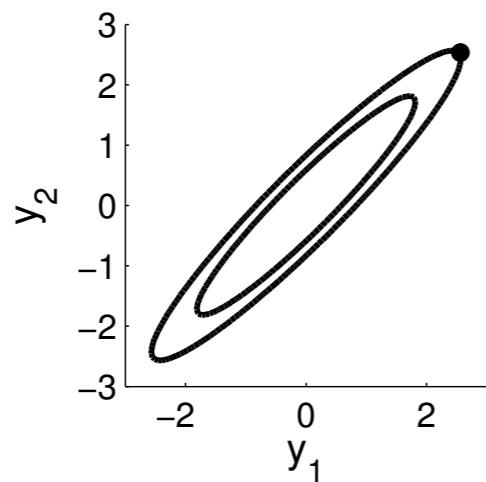
# New Visualisation



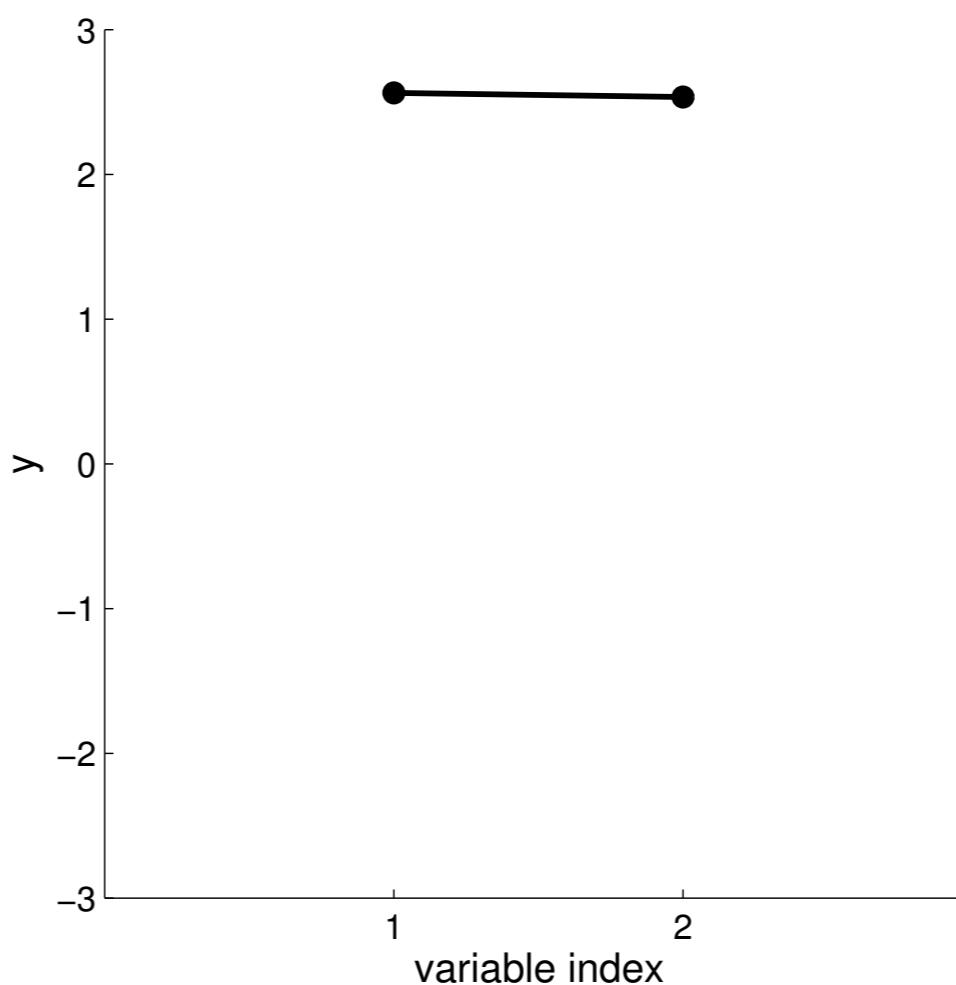
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



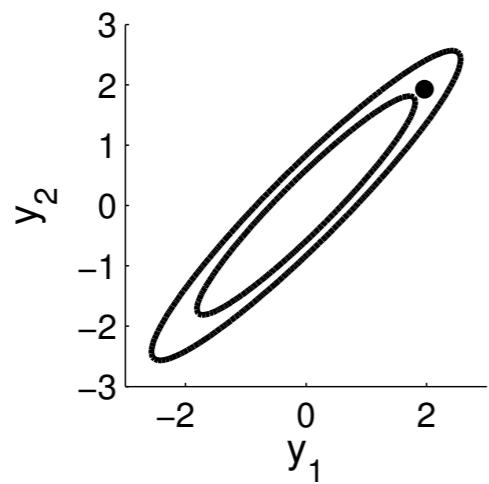
# New Visualisation



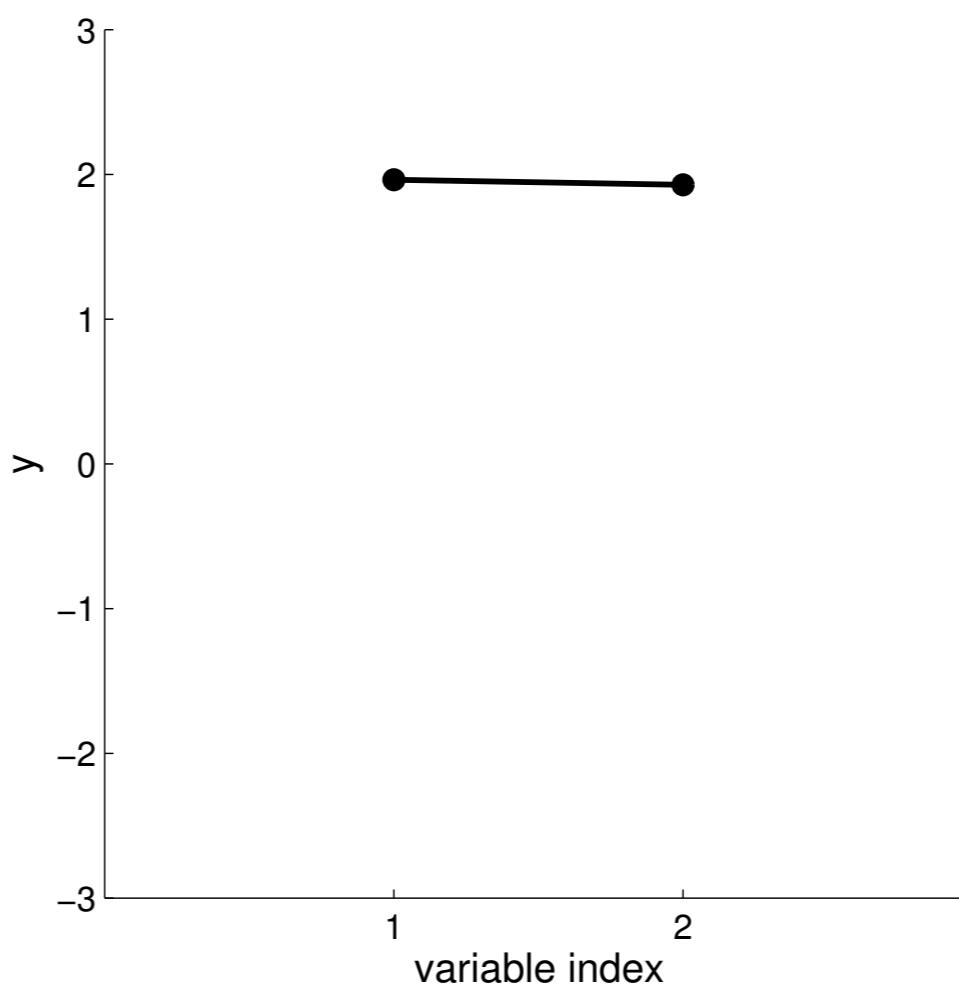
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



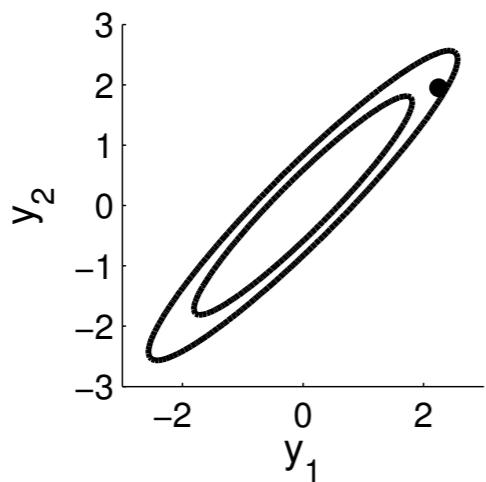
# New Visualisation



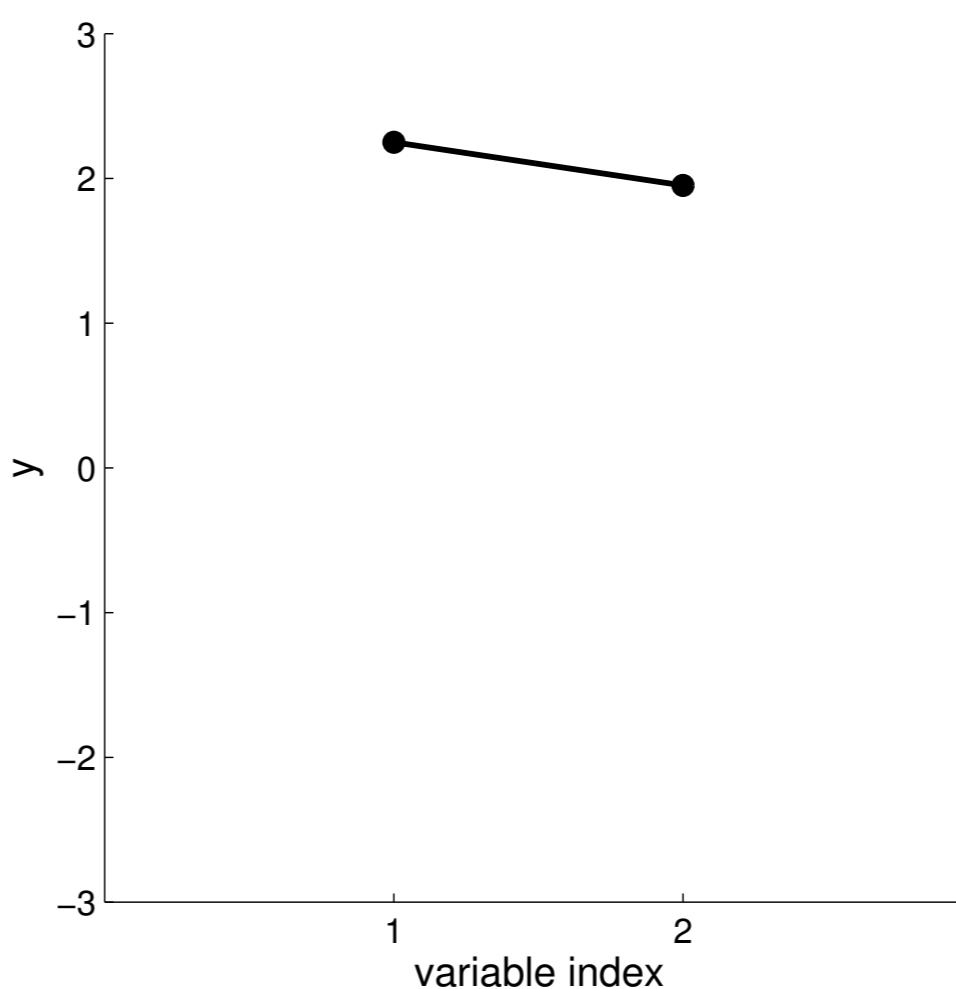
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



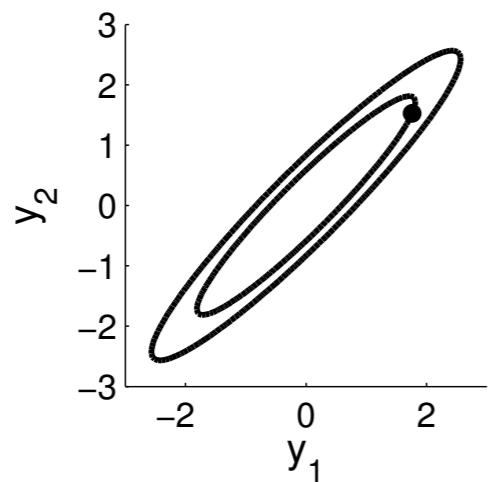
# New Visualisation



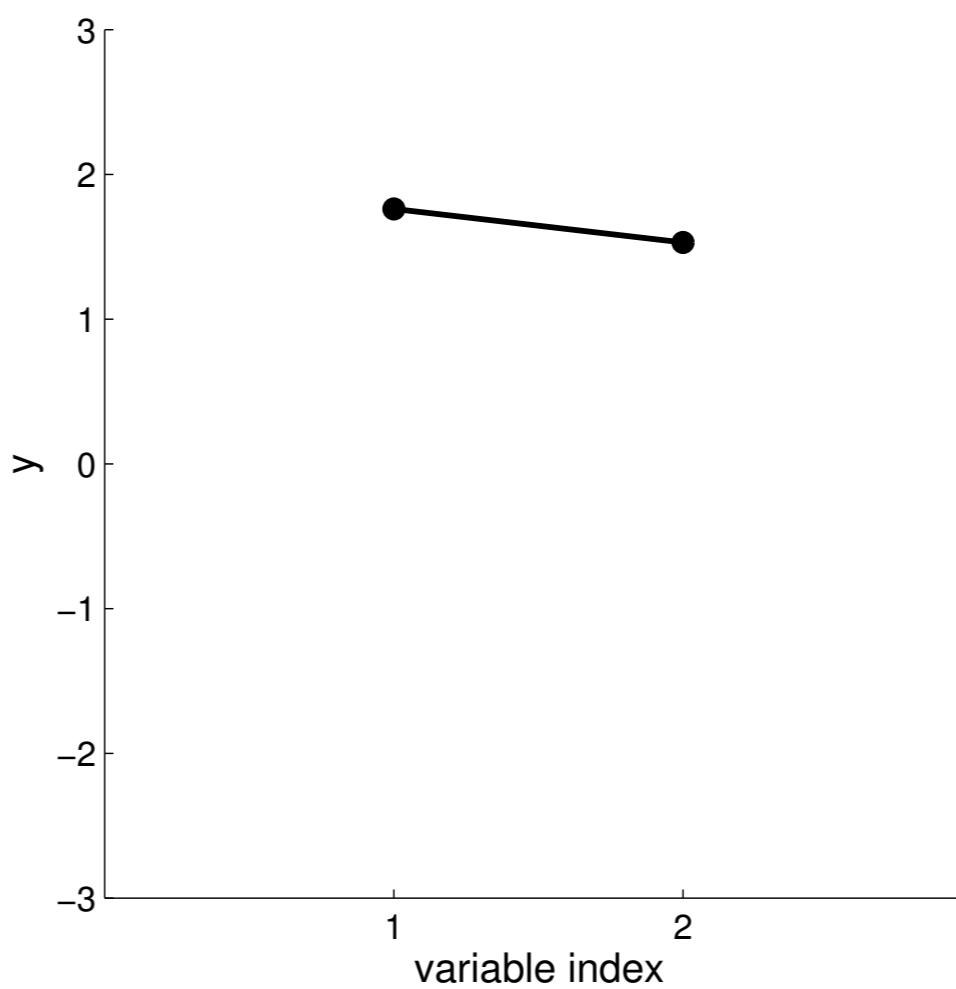
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



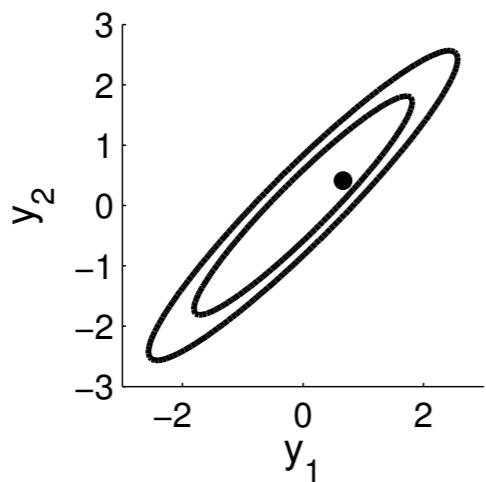
# New Visualisation



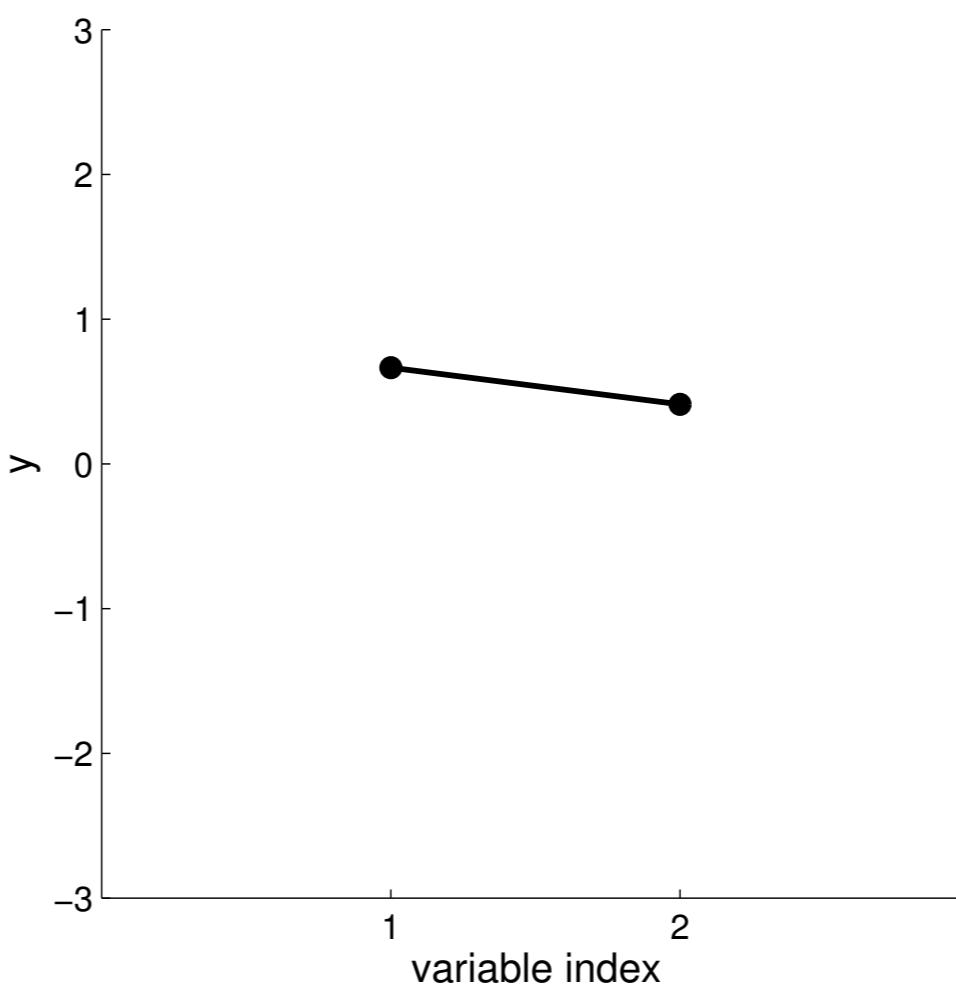
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



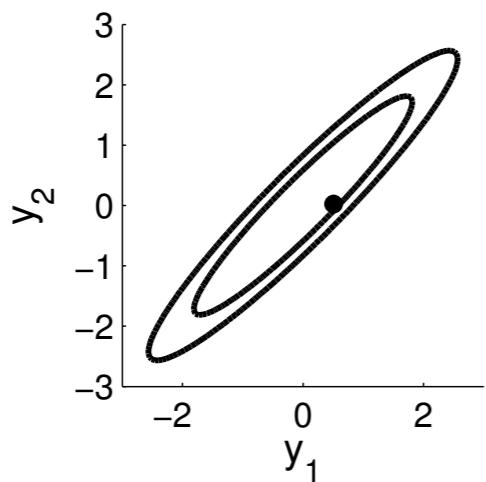
# New Visualisation



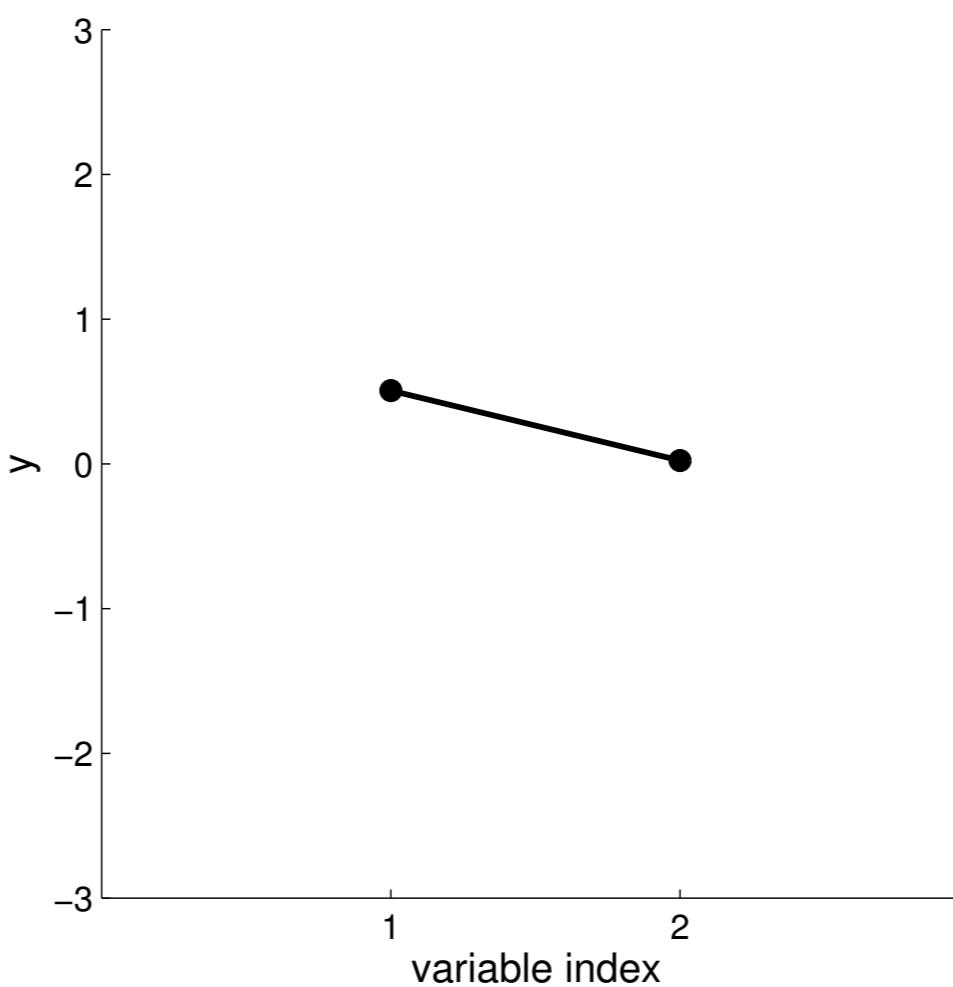
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



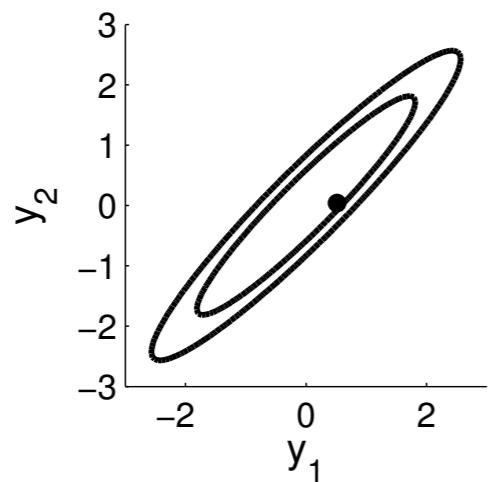
# New Visualisation



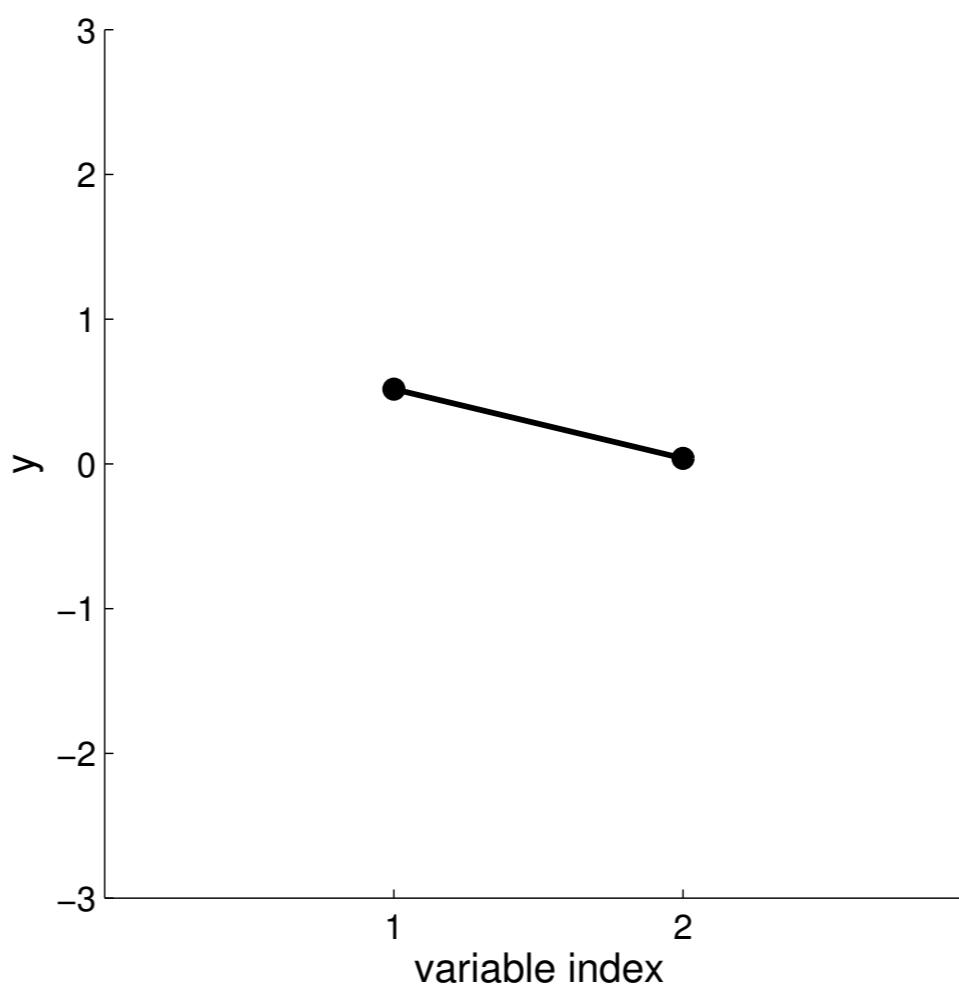
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



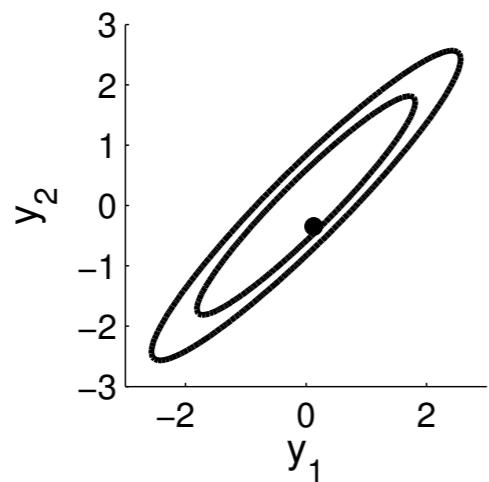
# New Visualisation



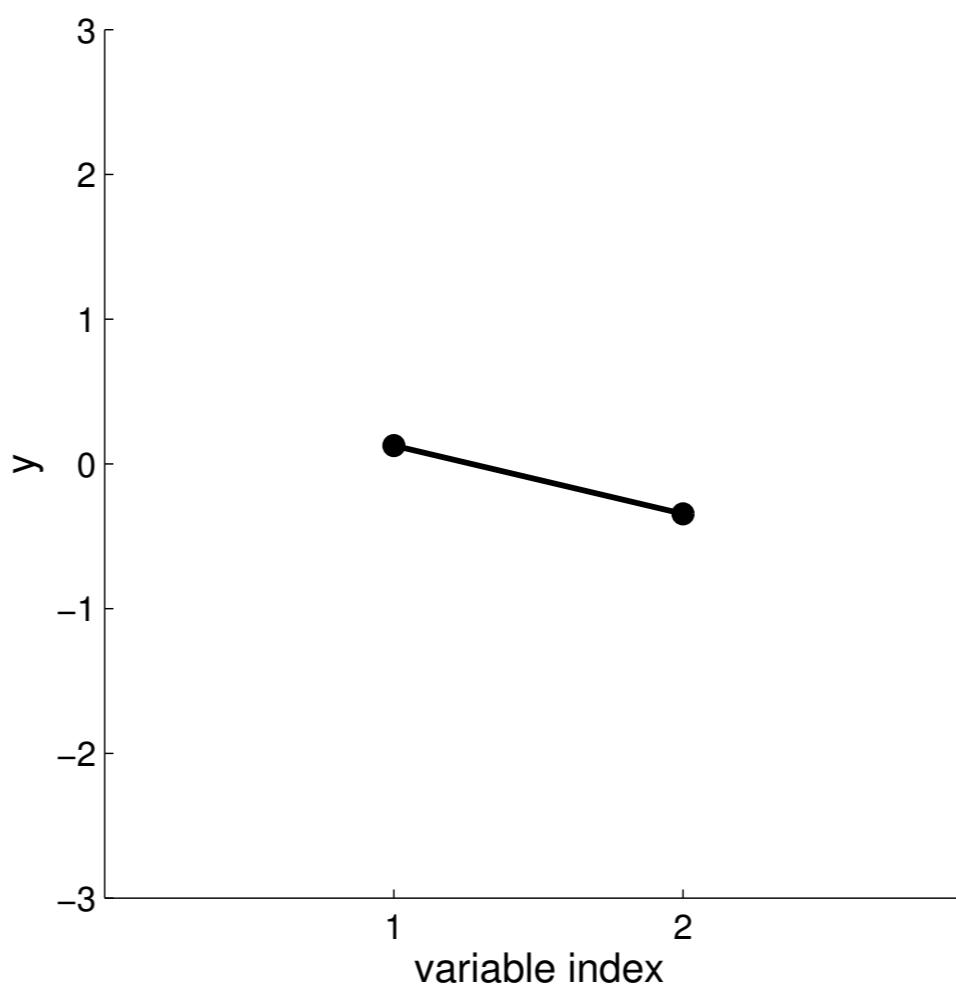
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



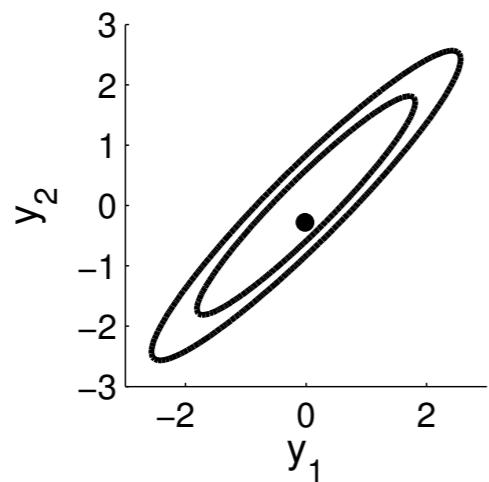
# New Visualisation



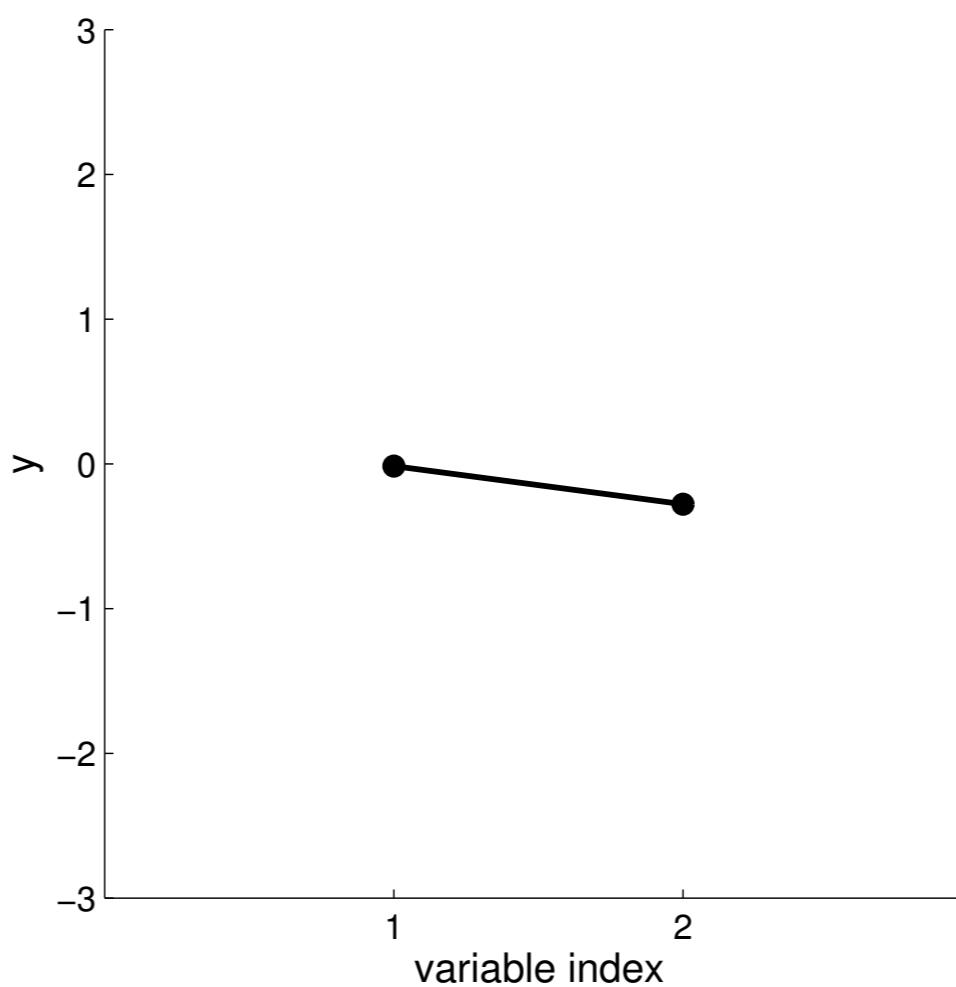
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



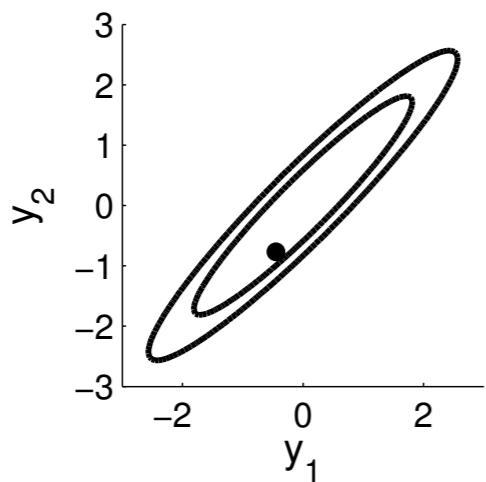
# New Visualisation



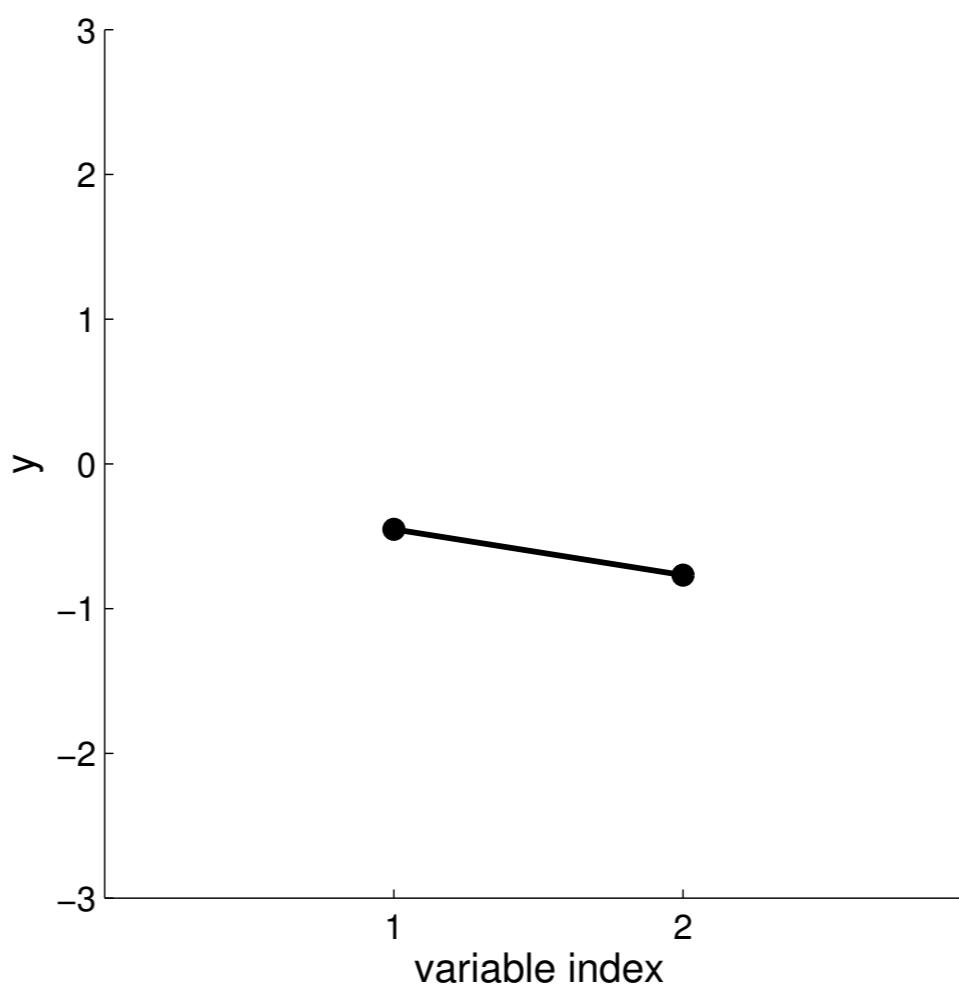
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



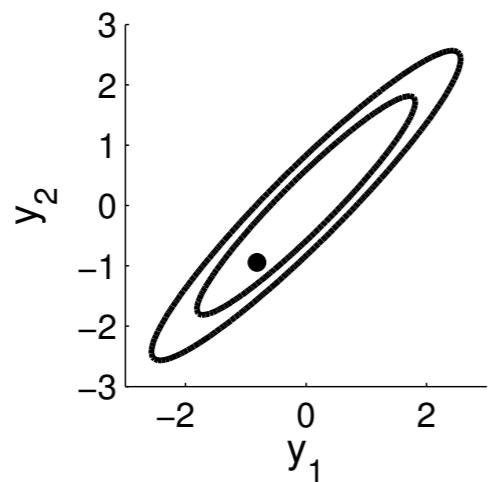
# New Visualisation



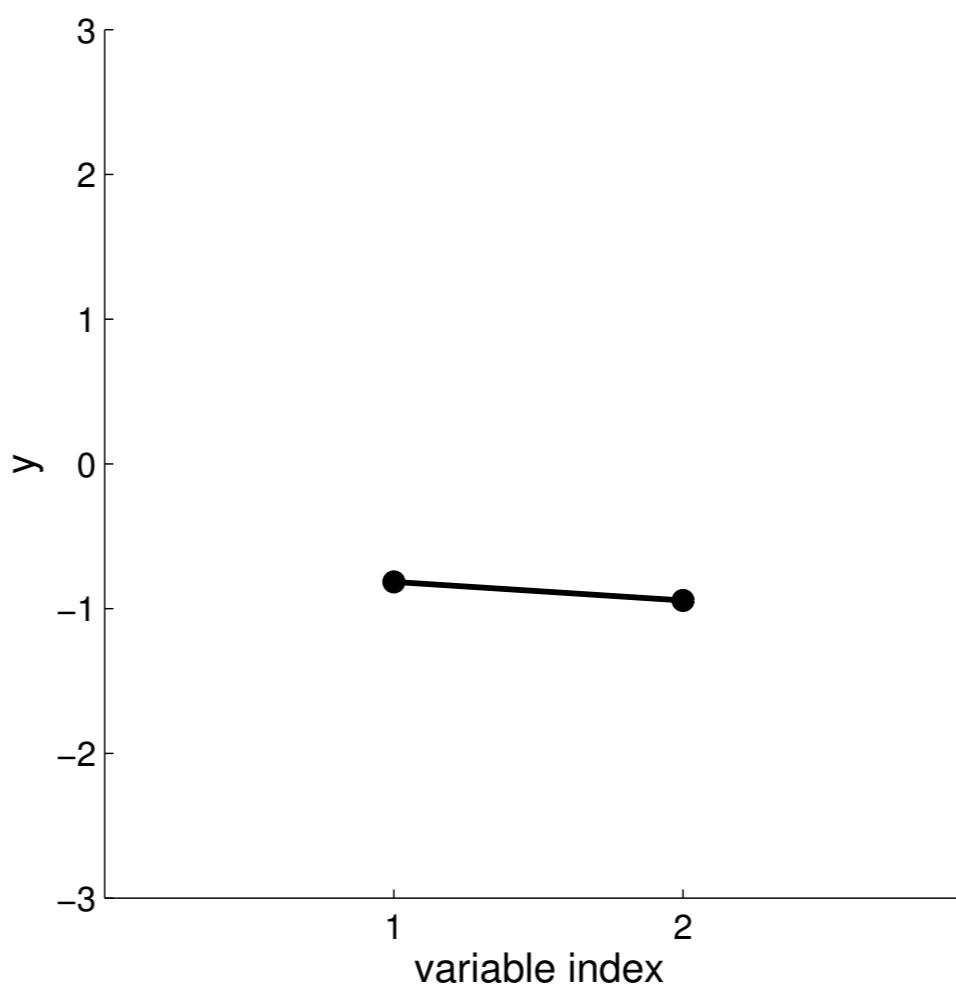
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



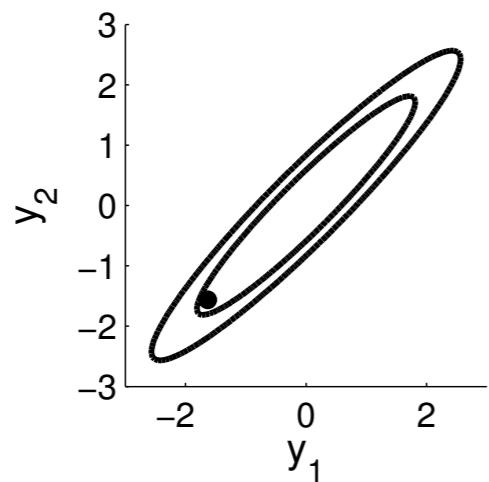
# New Visualisation



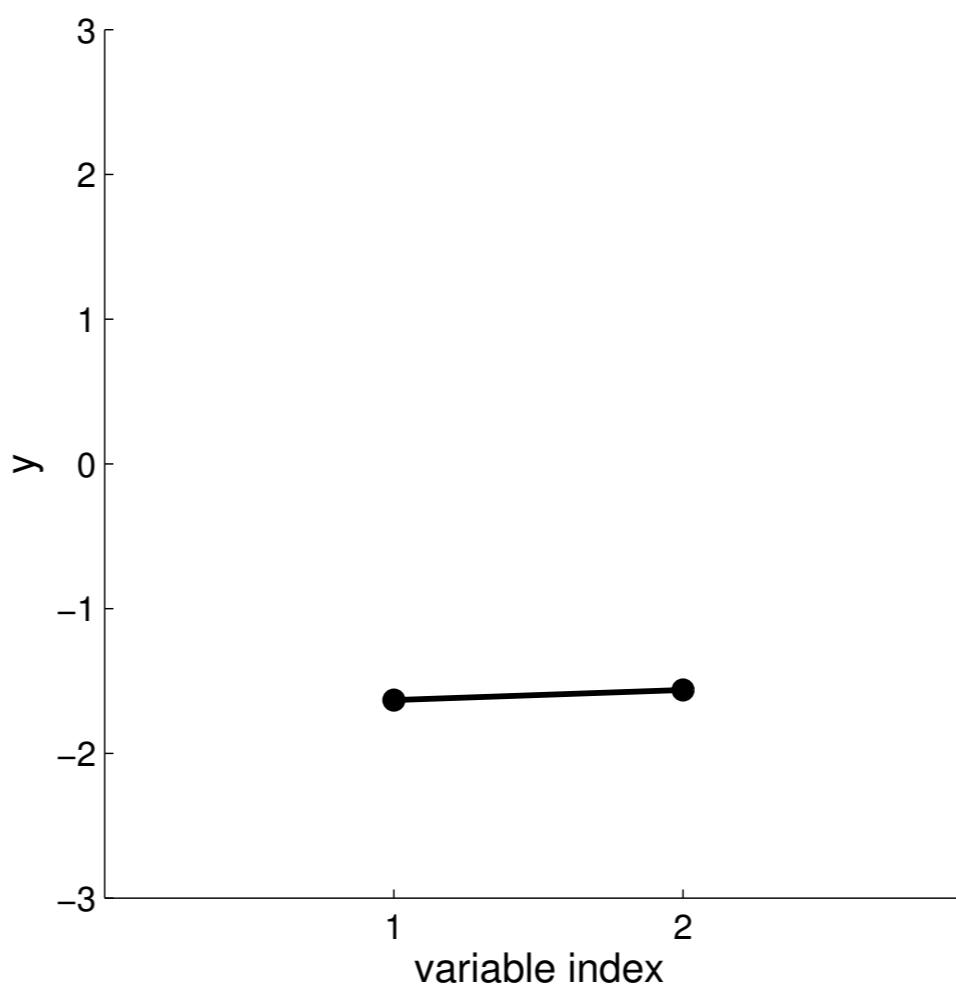
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



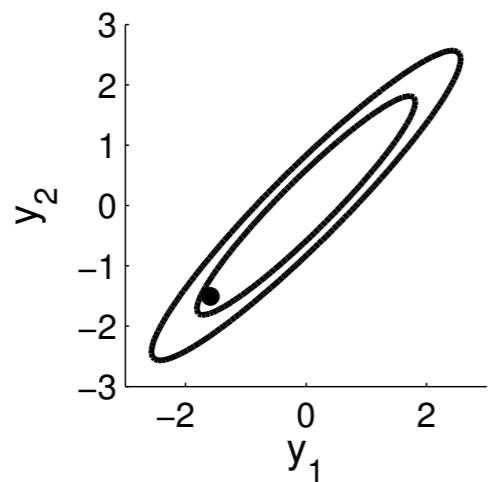
# New Visualisation



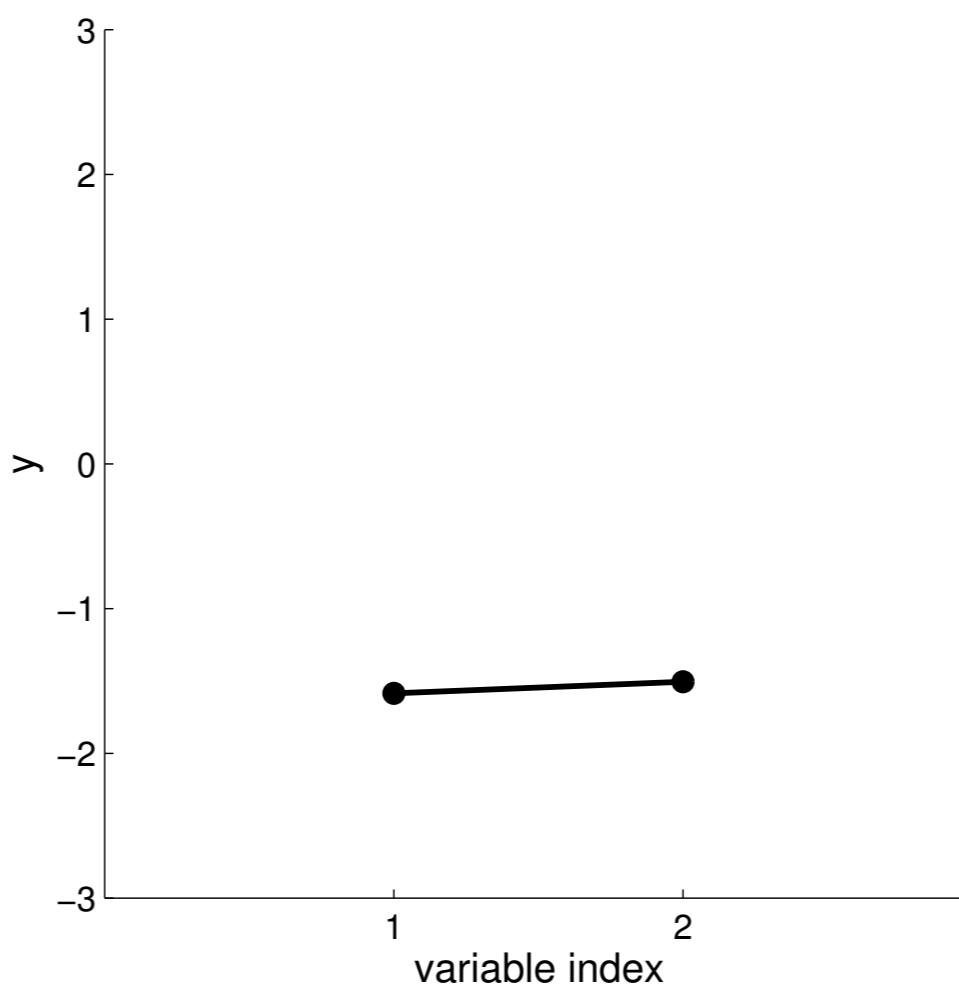
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



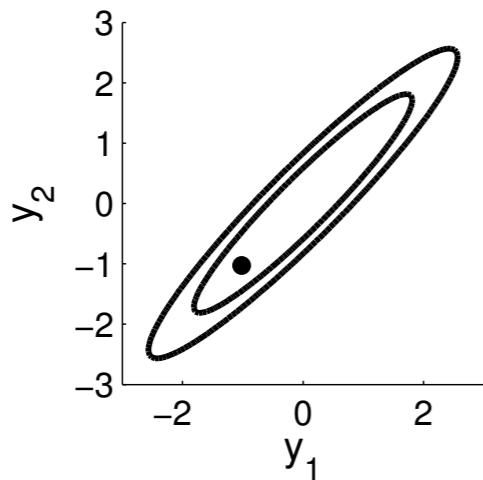
# New Visualisation



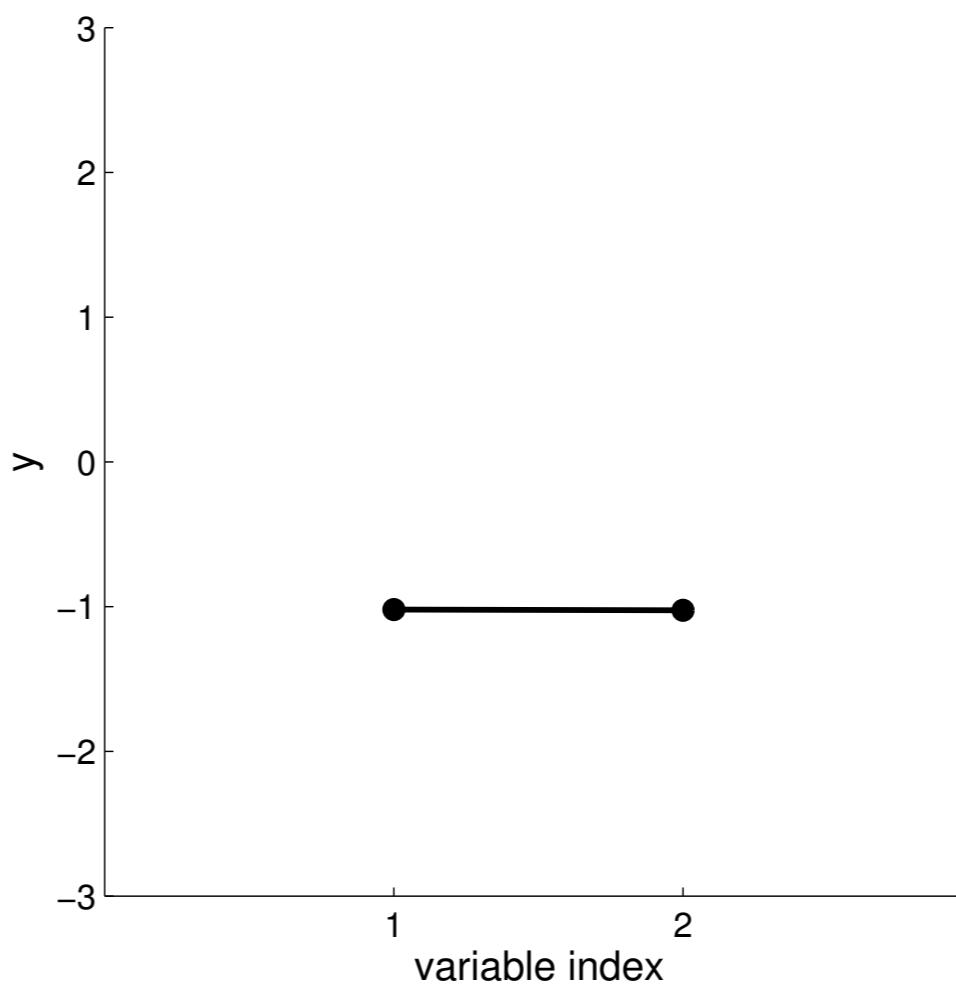
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



# New Visualisation

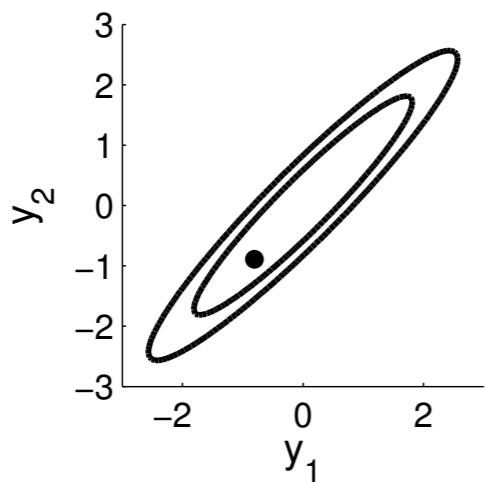


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

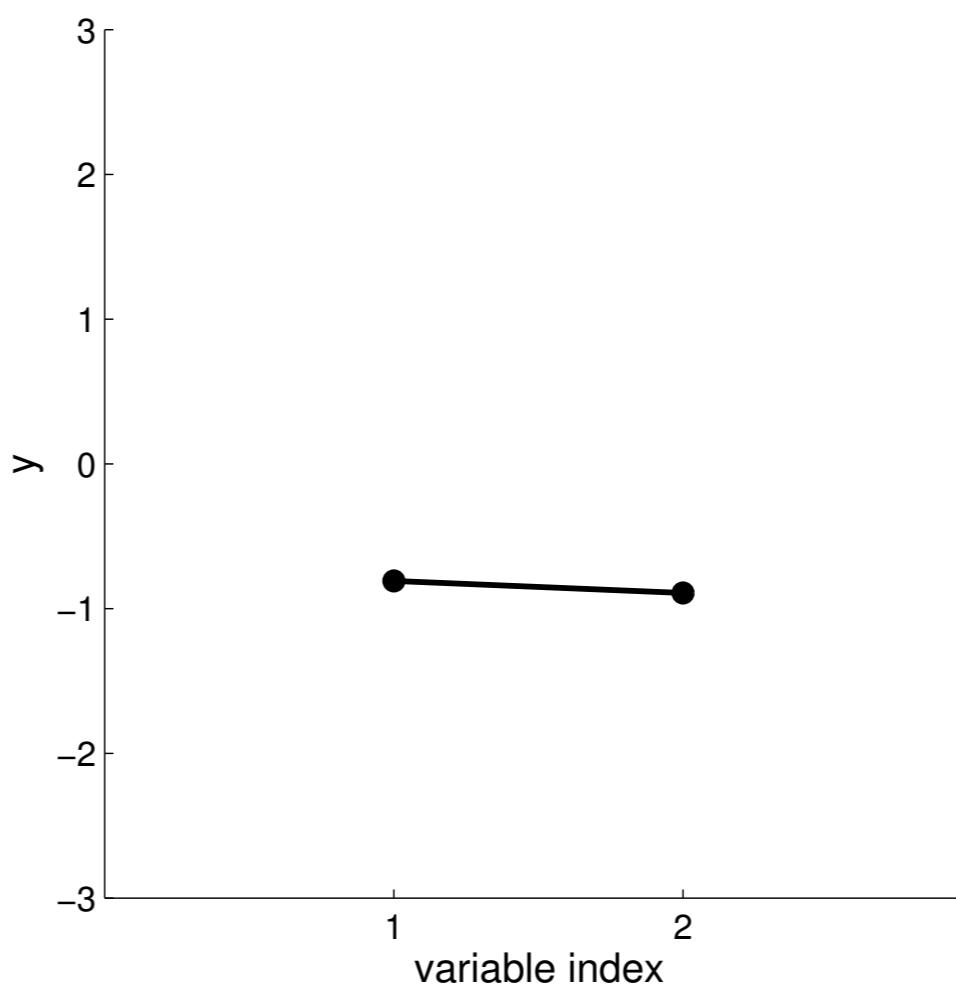


Q: How would the bar look like if the cross-correlation of  $y_1, y_2$  was 1?

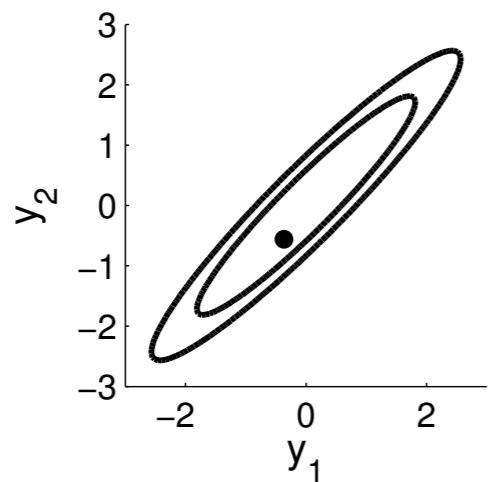
# New Visualisation



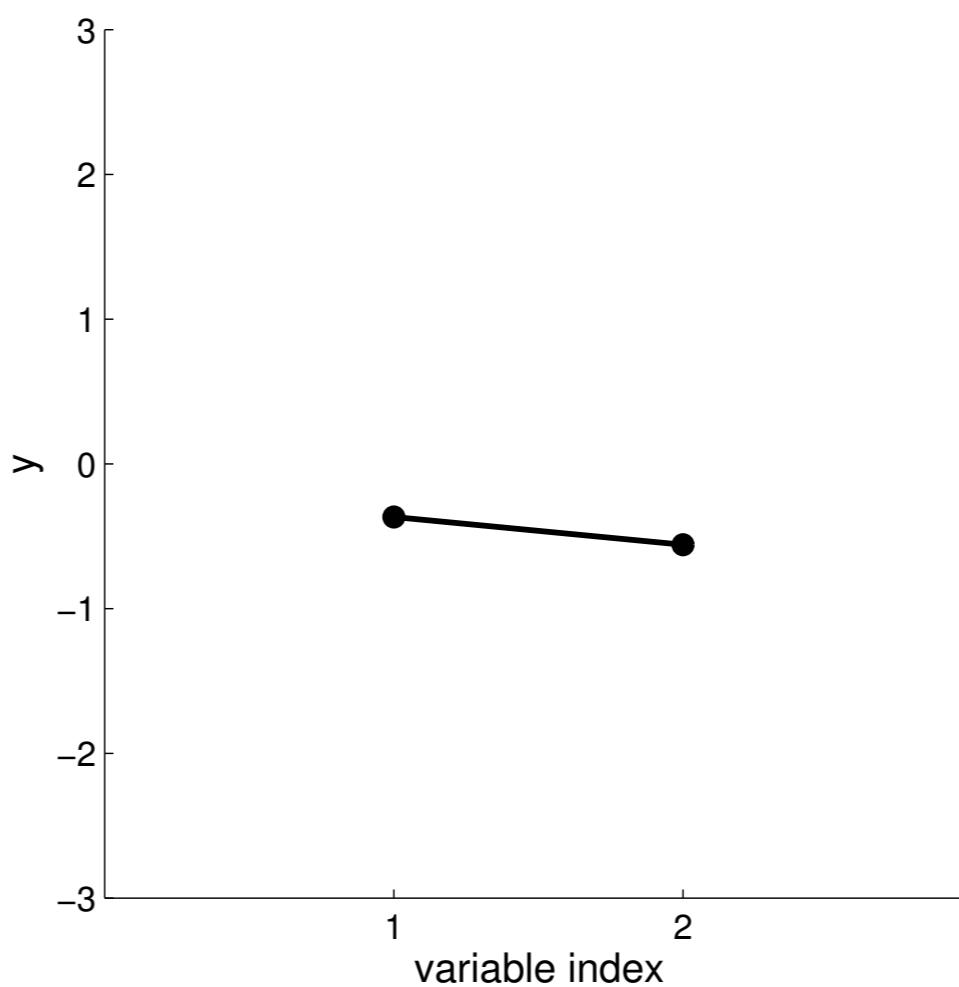
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



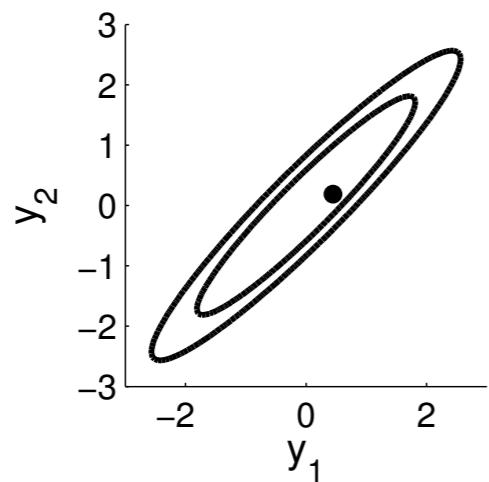
# New Visualisation



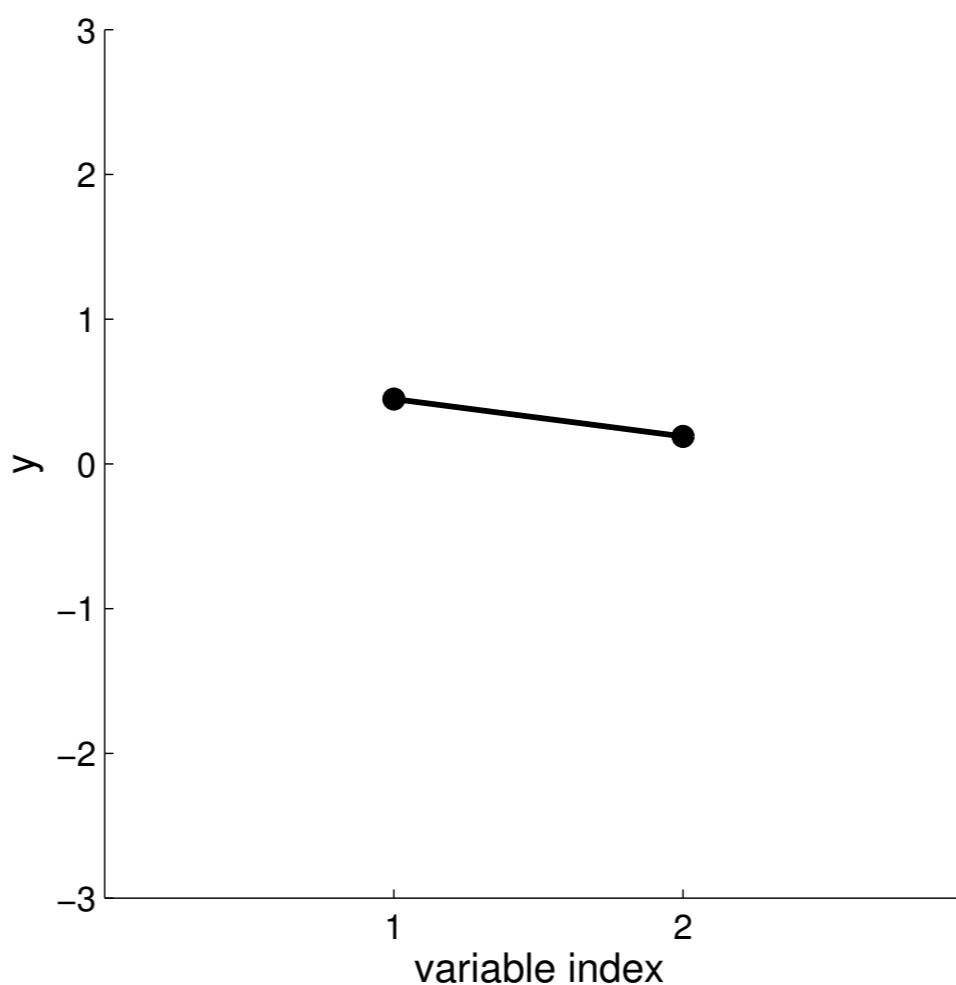
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



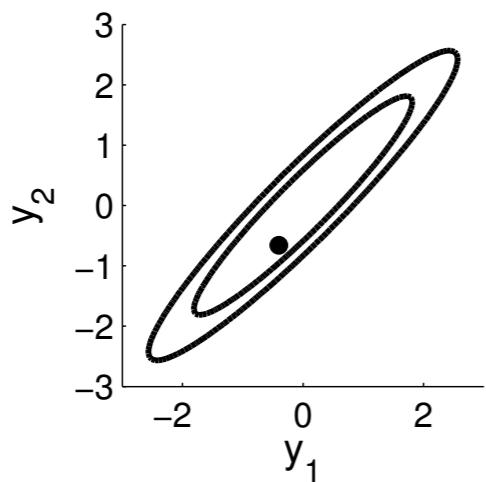
# New Visualisation



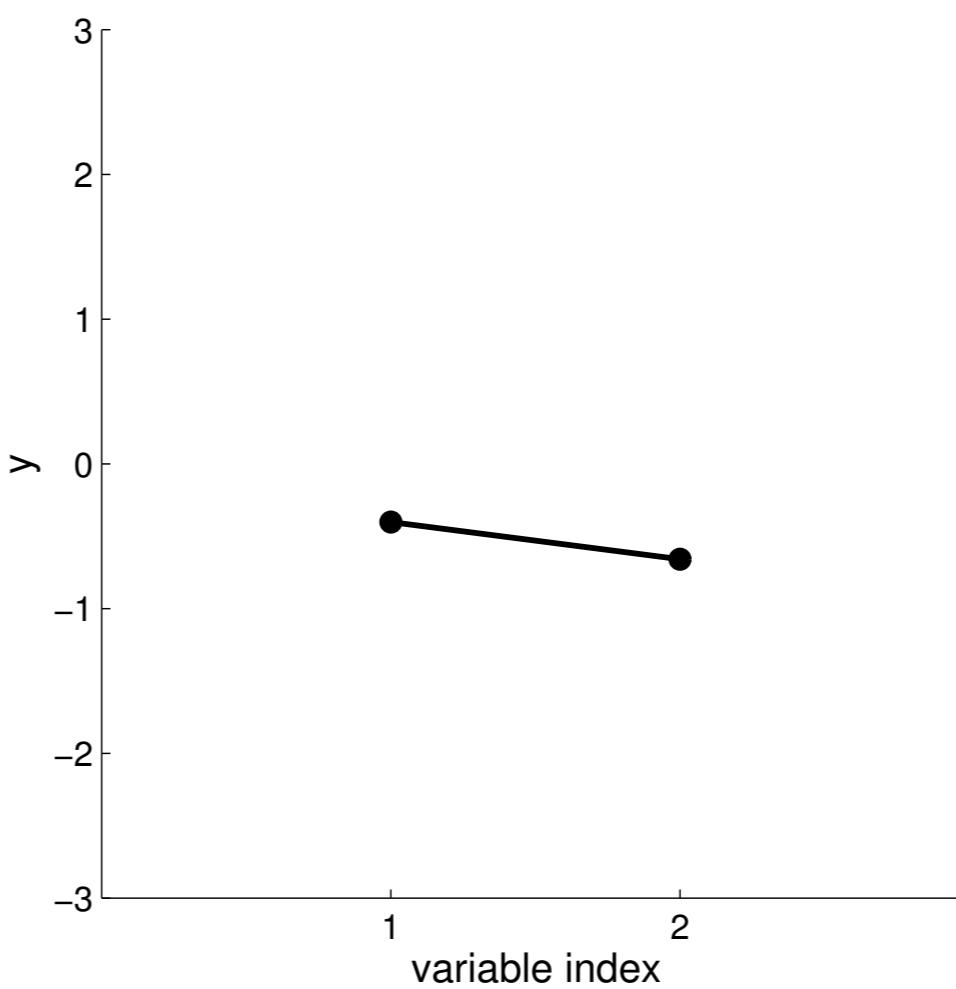
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



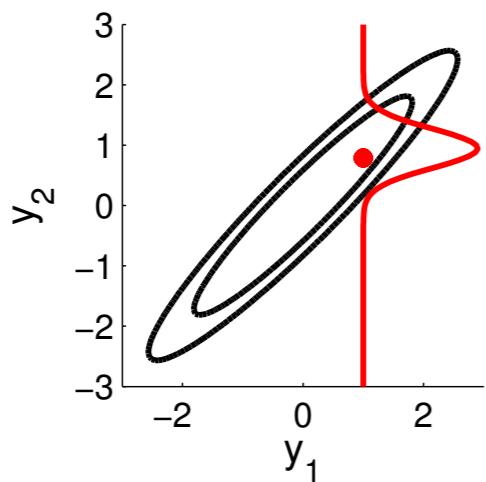
# New Visualisation



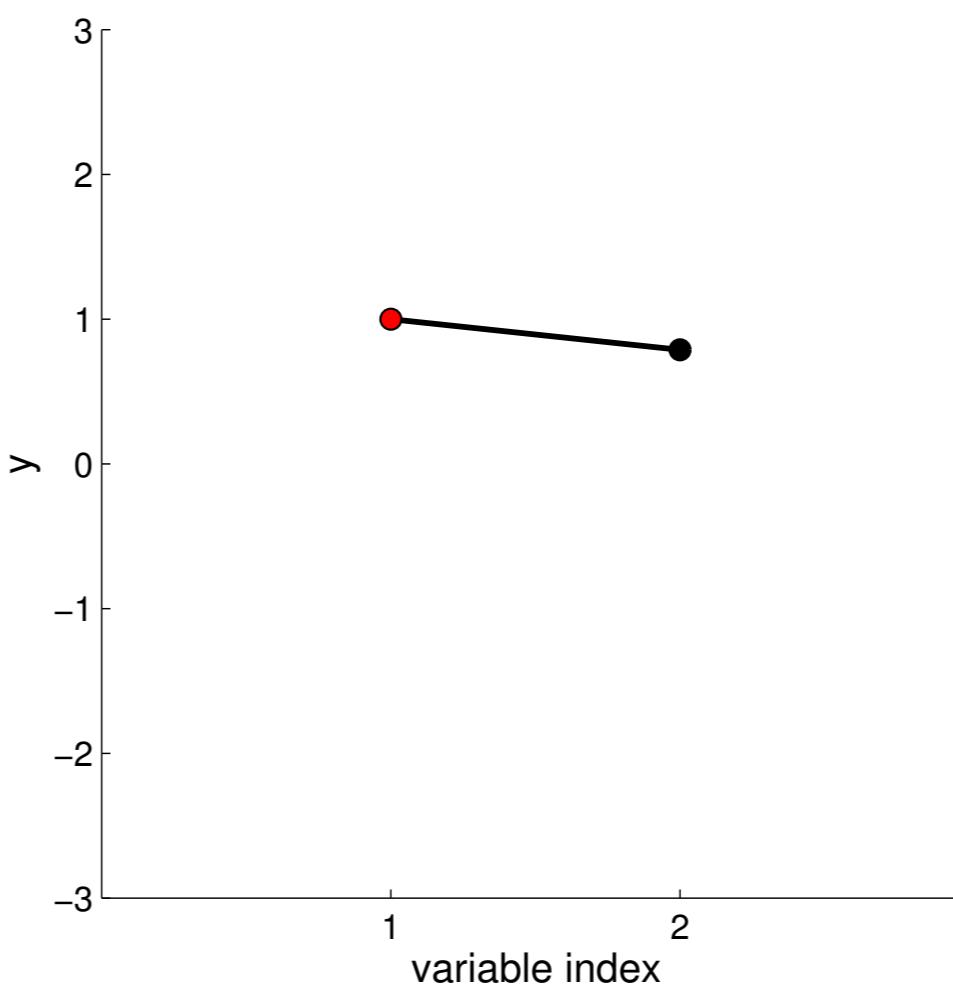
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



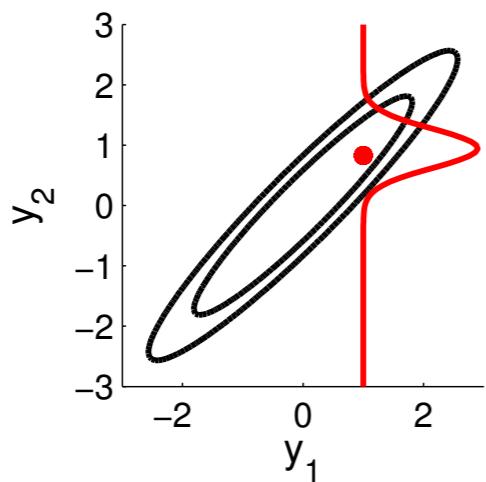
# New Visualisation



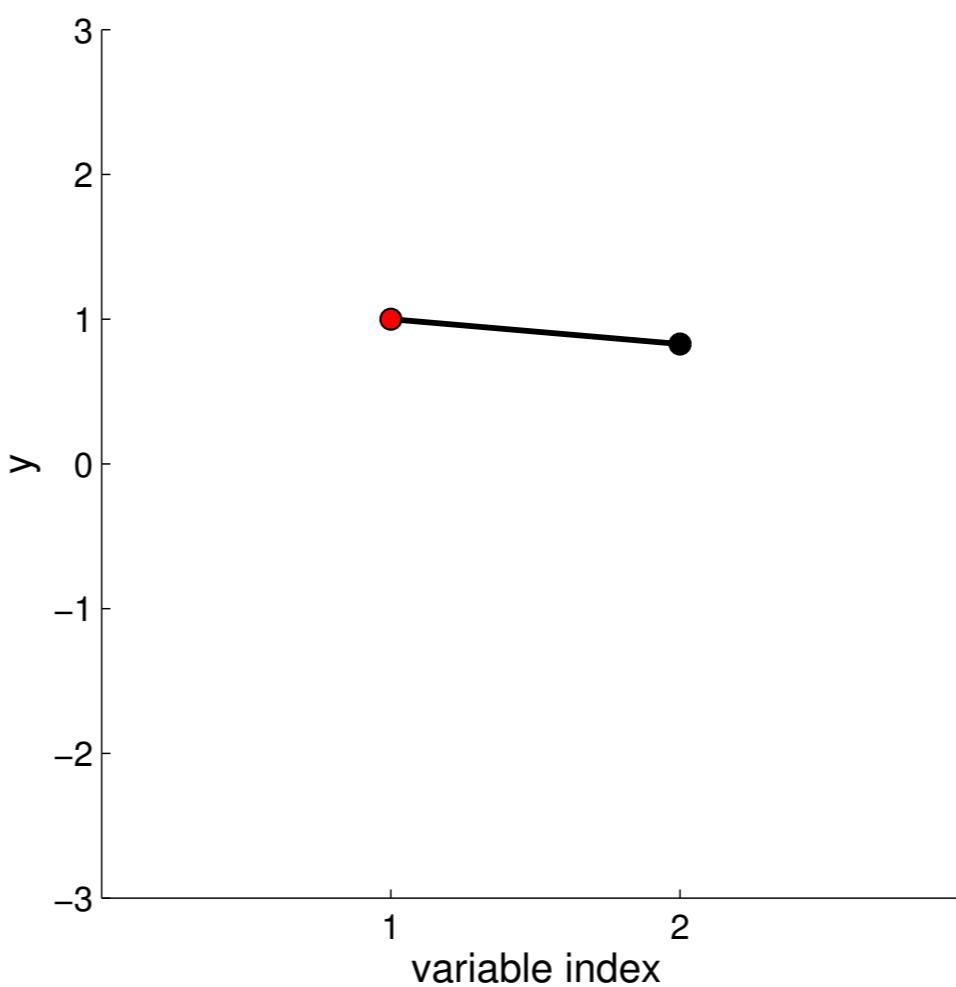
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



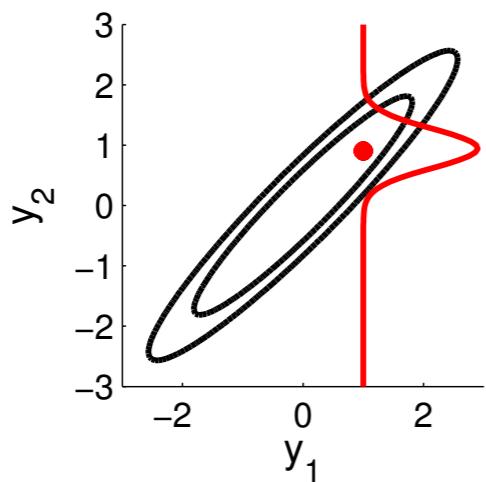
# New Visualisation



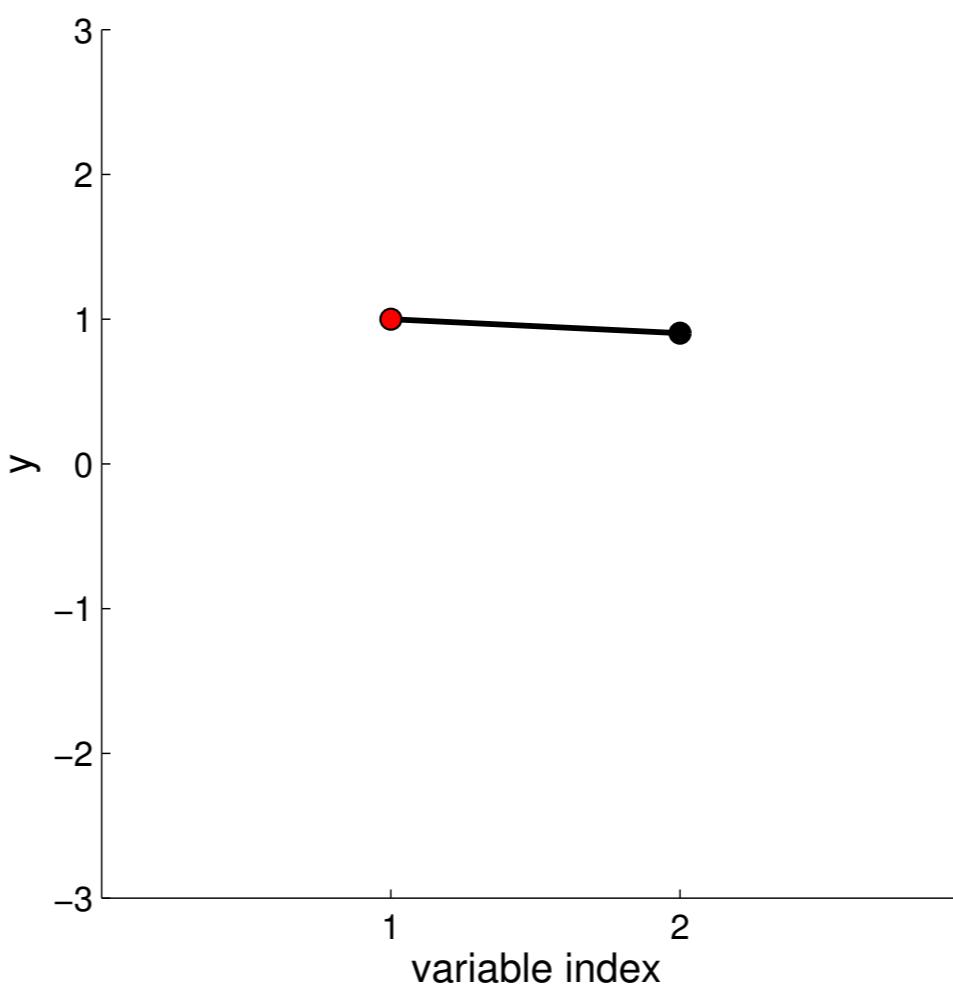
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



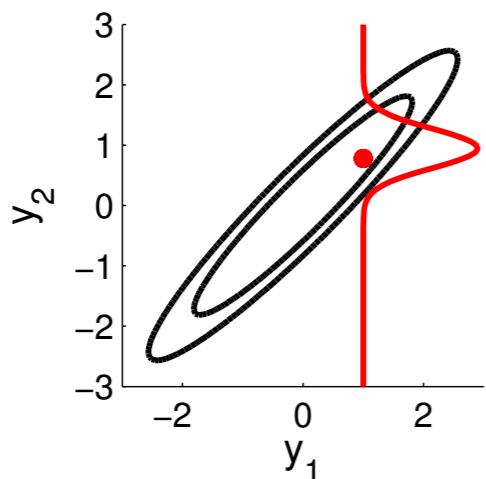
# New Visualisation



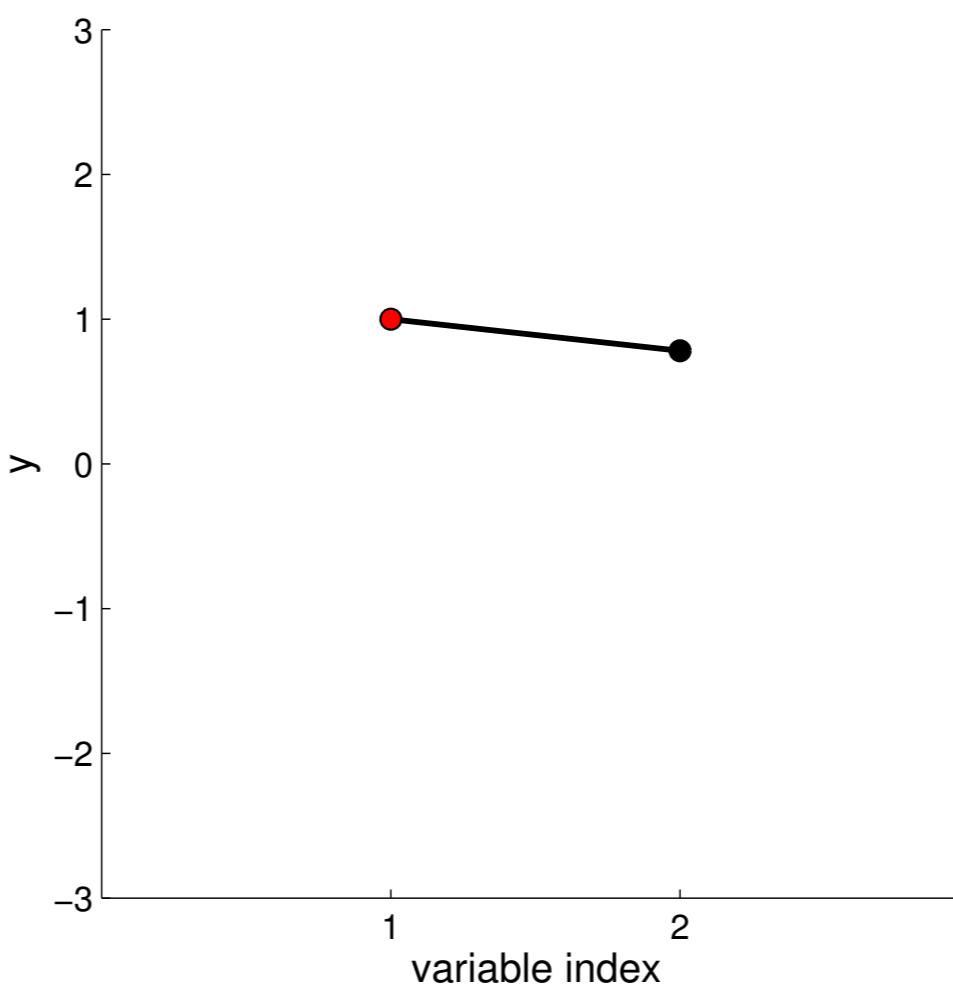
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



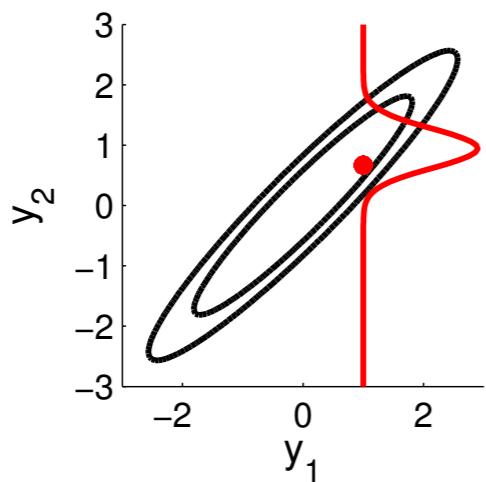
# New Visualisation



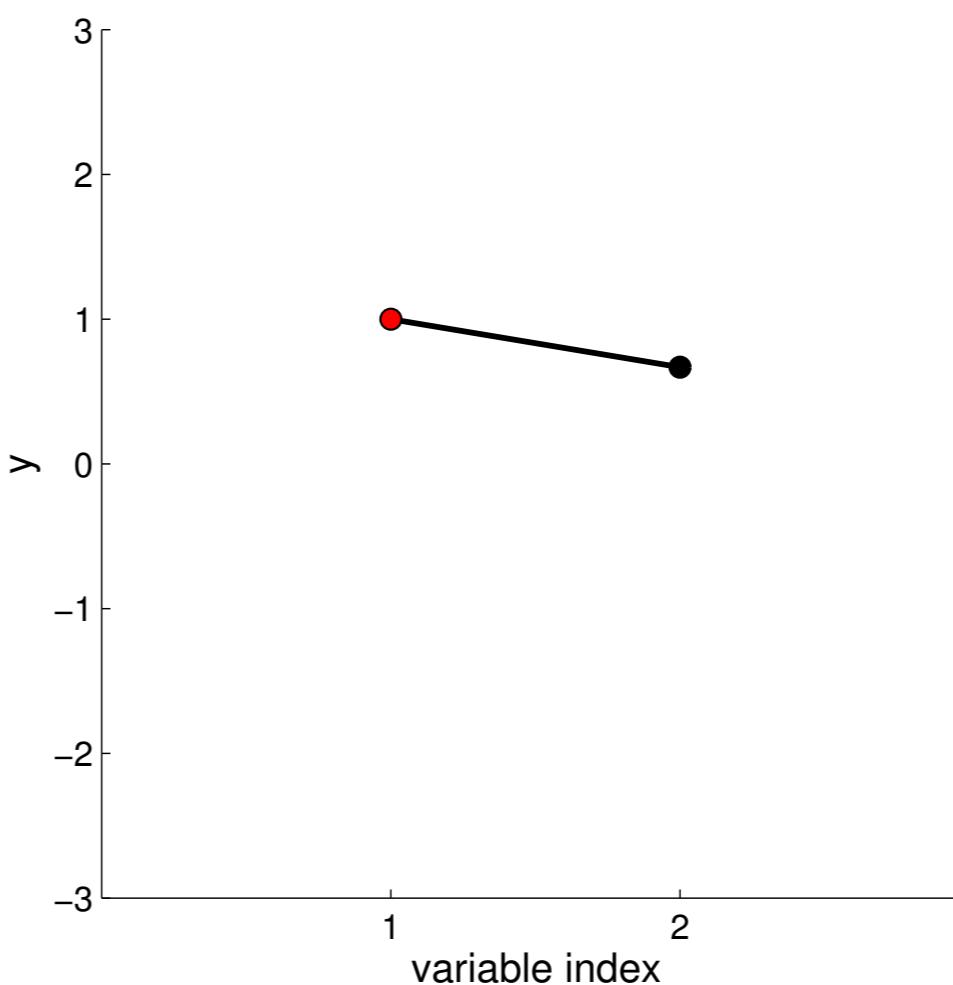
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



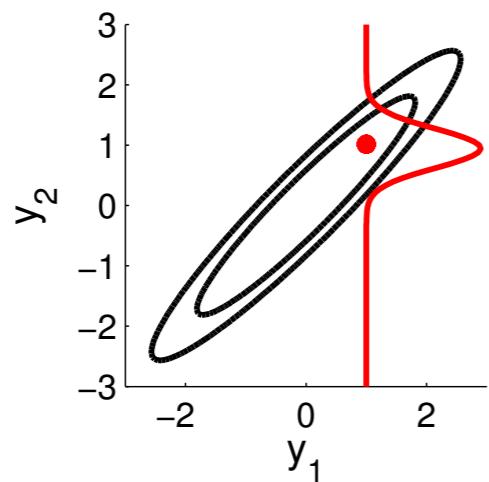
# New Visualisation



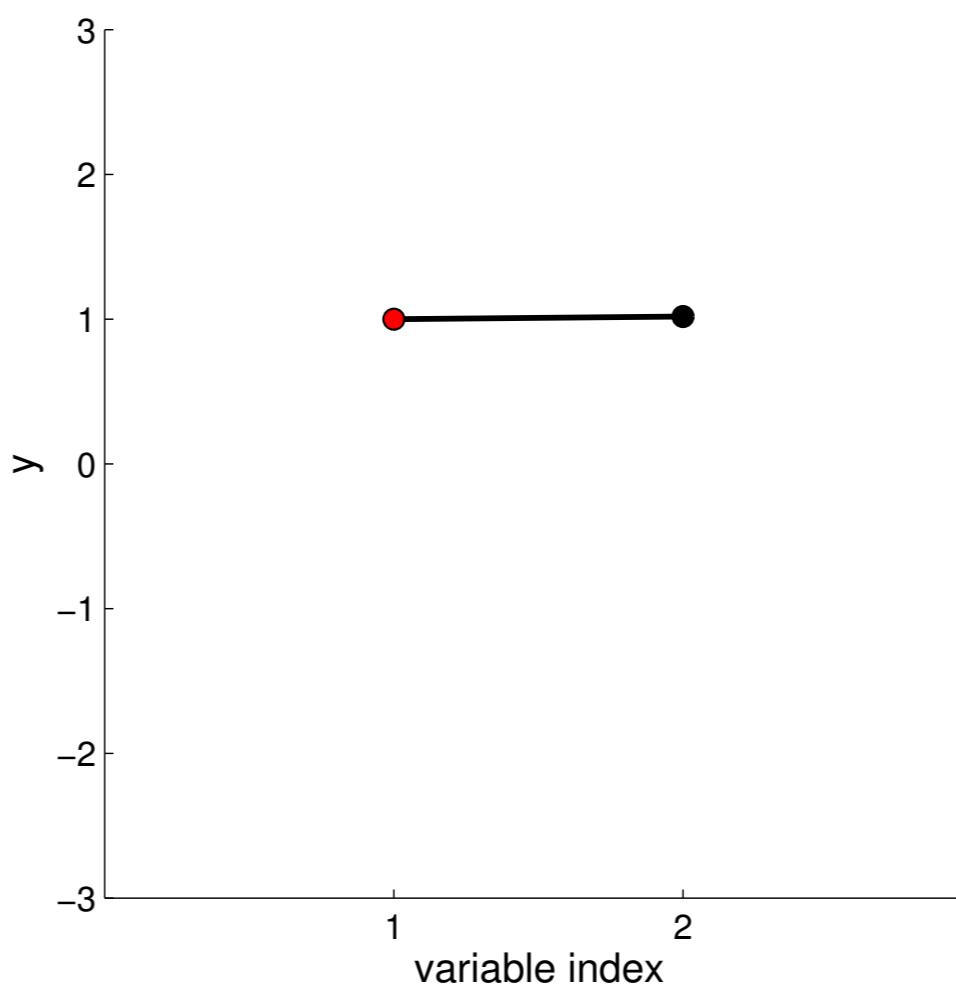
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



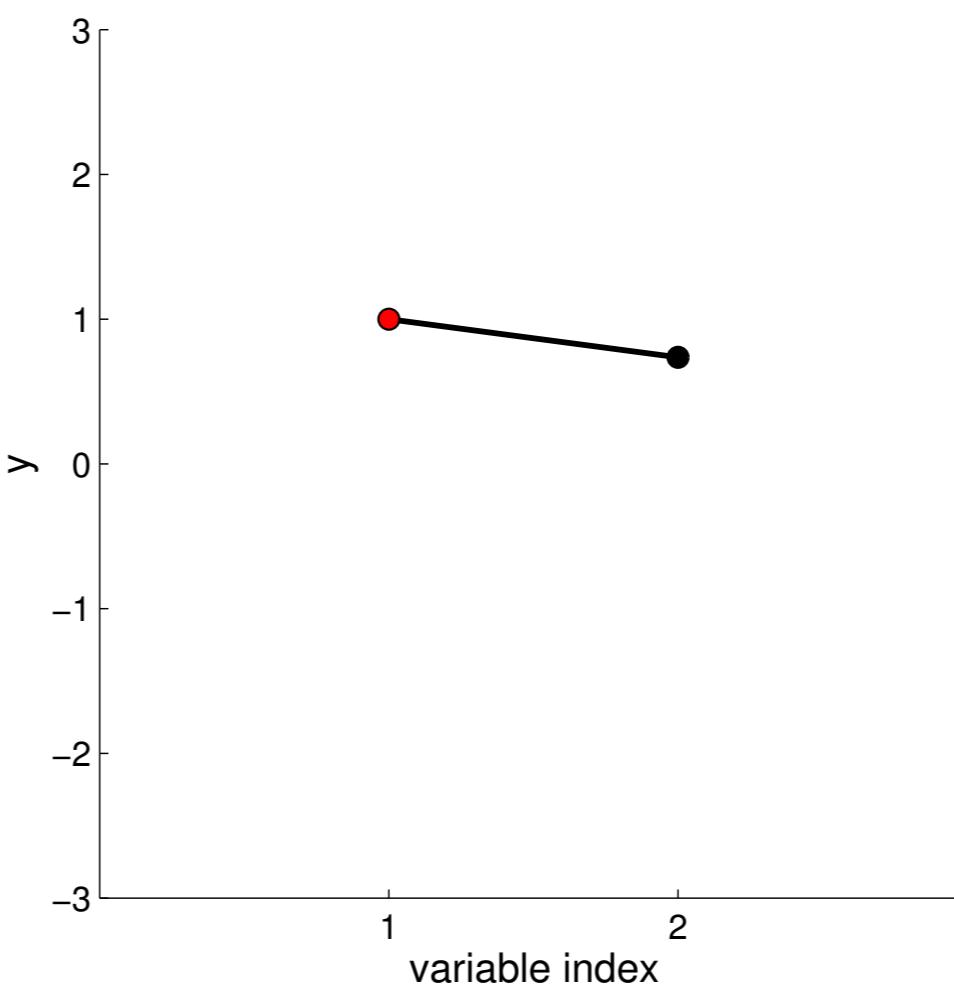
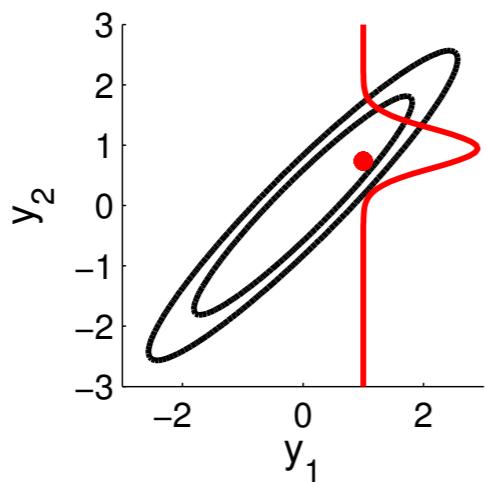
# New Visualisation



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

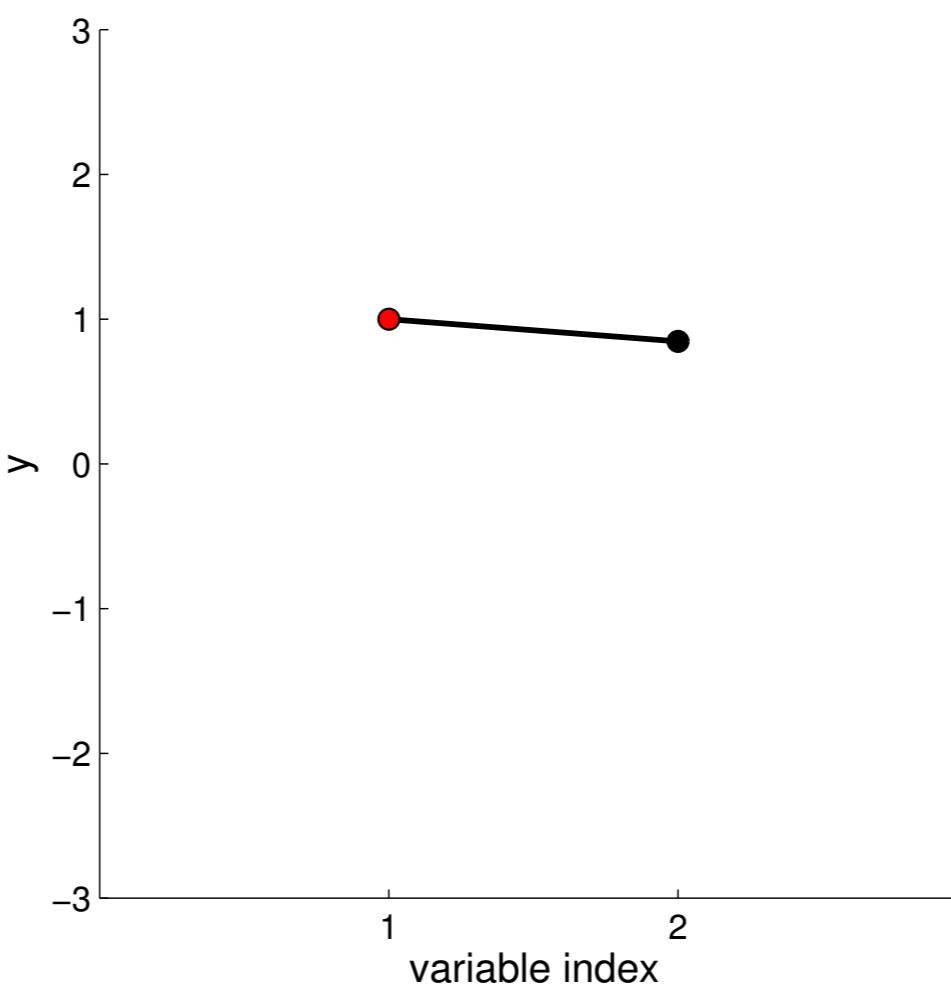
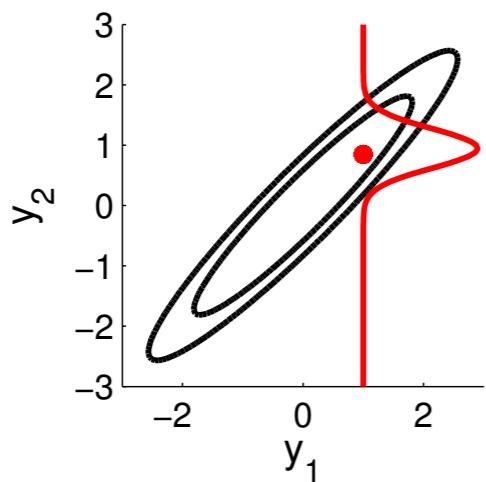


# New Visualisation



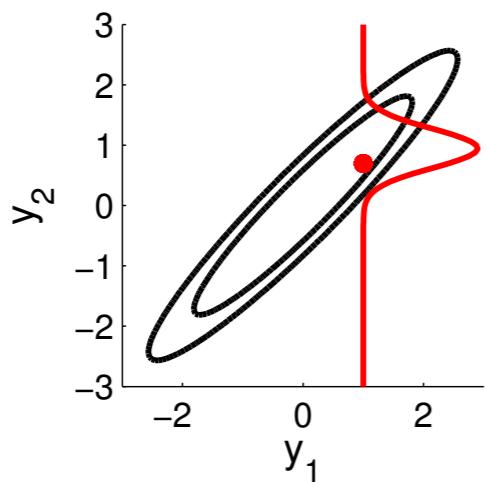
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

# New Visualisation

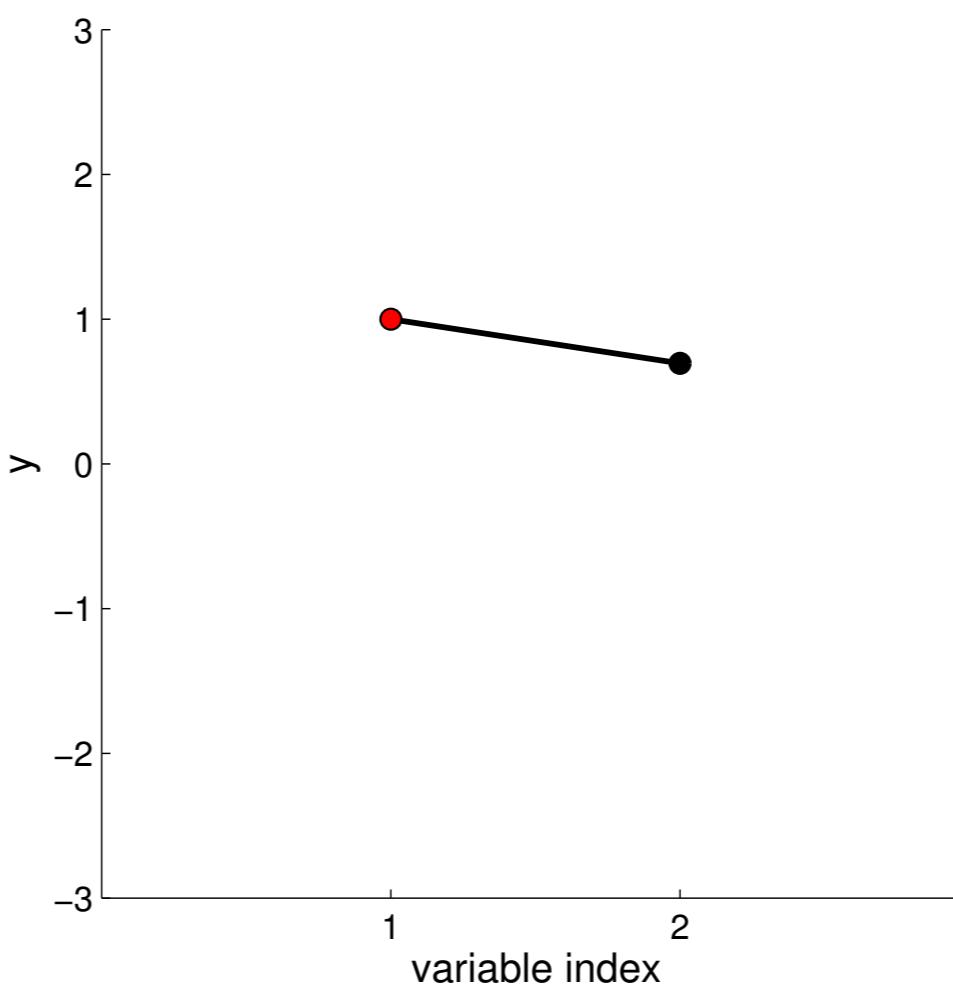


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

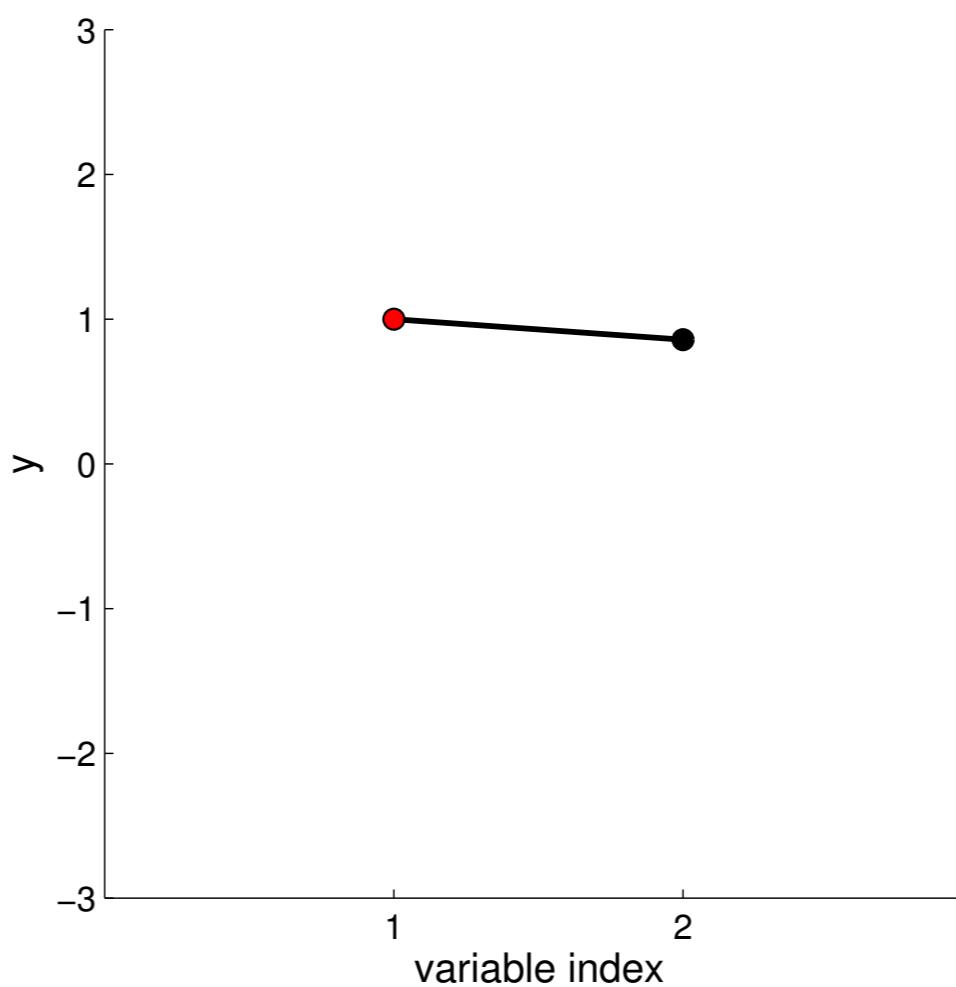
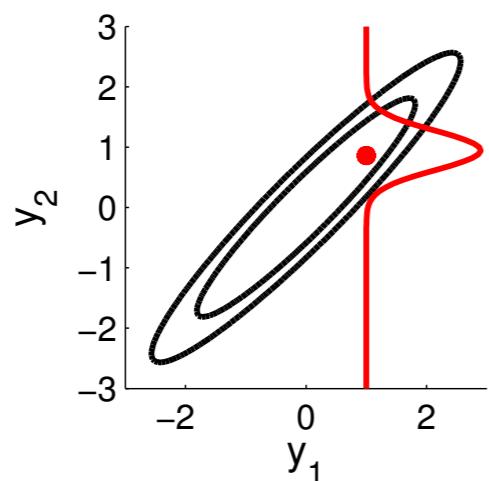
# New Visualisation



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

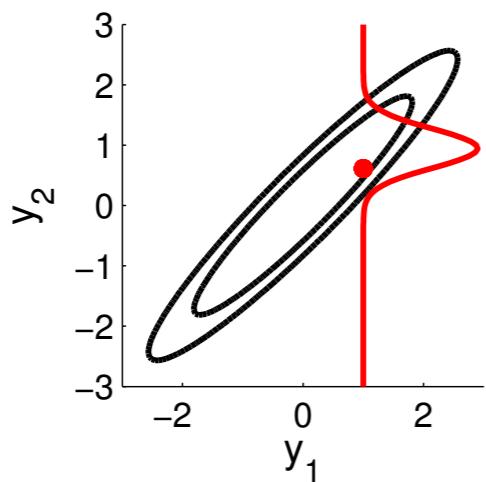


# New Visualisation

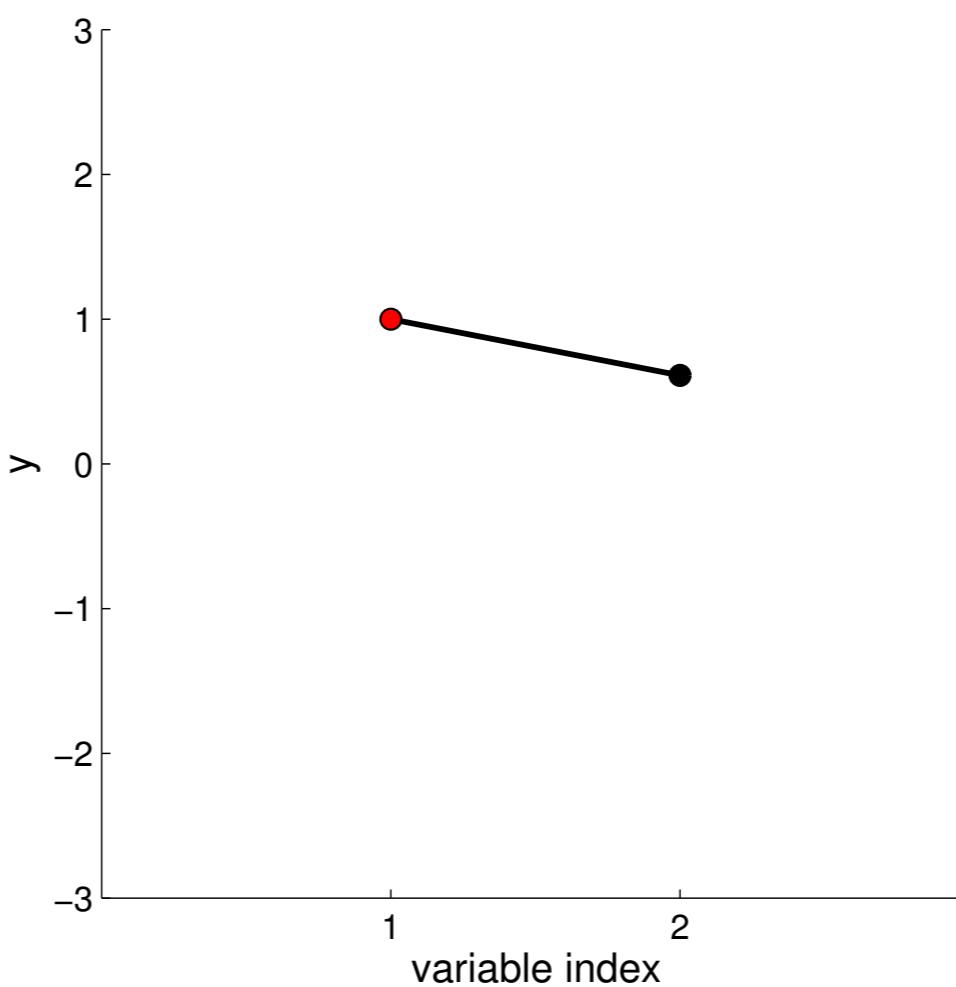


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

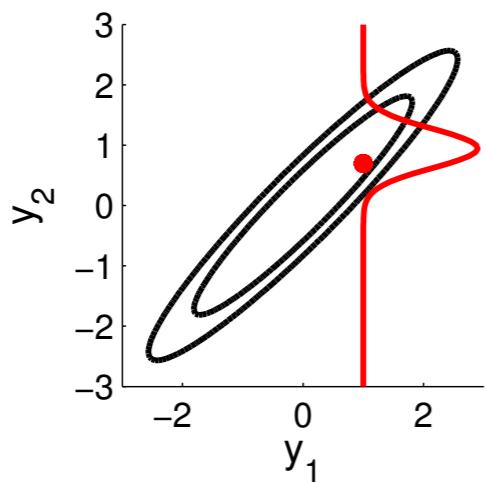
# New Visualisation



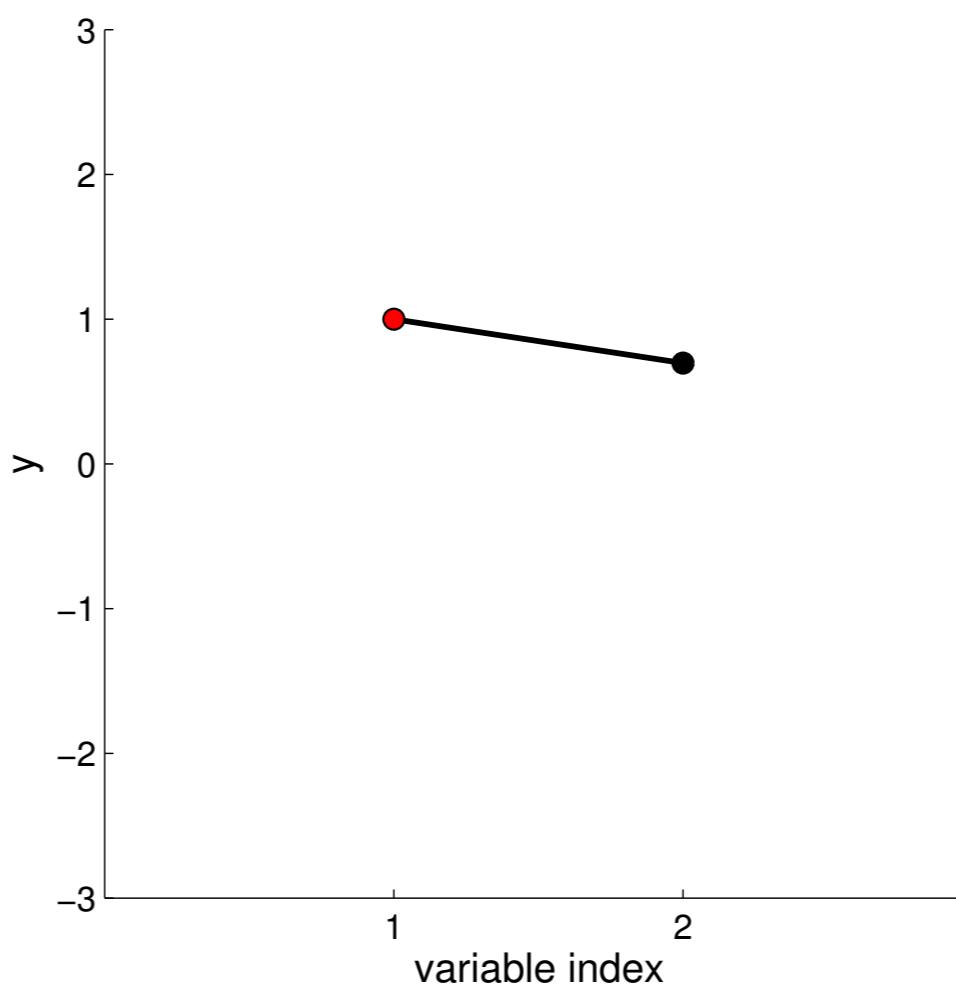
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



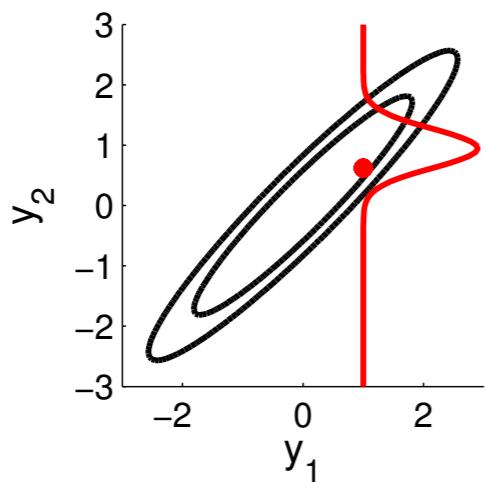
# New Visualisation



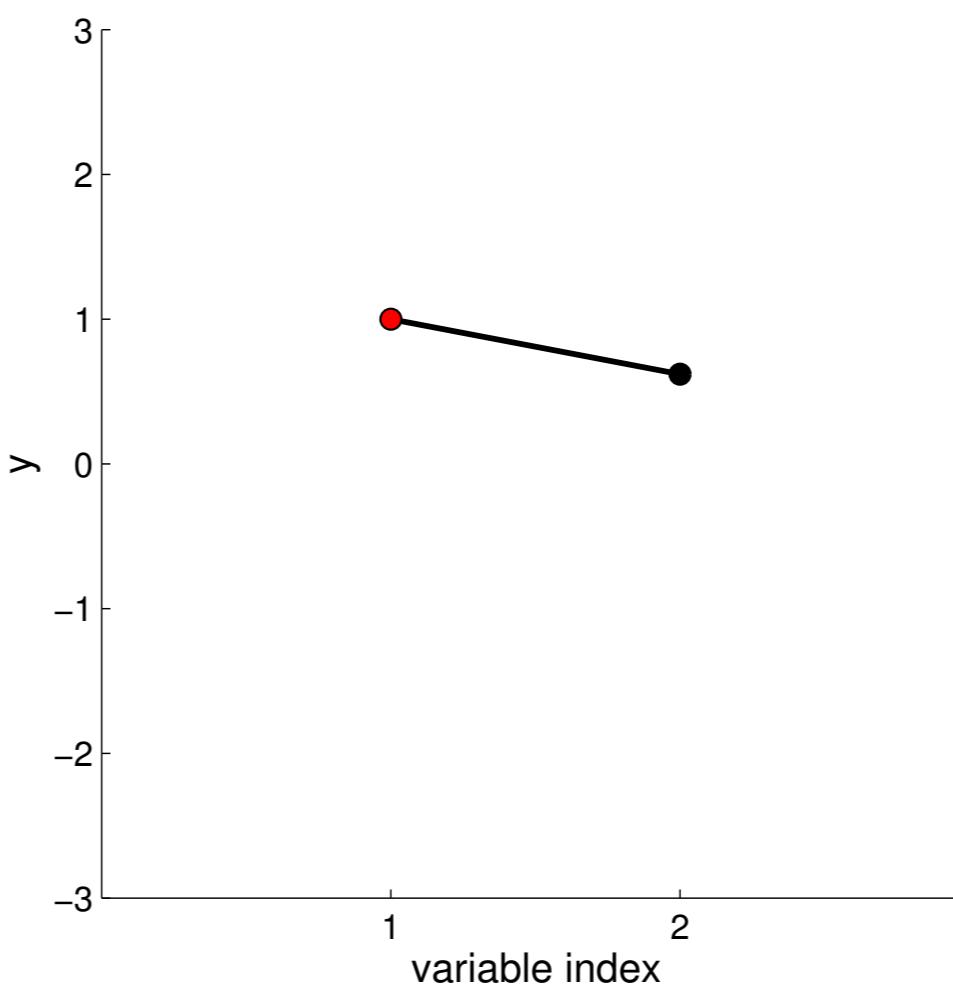
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



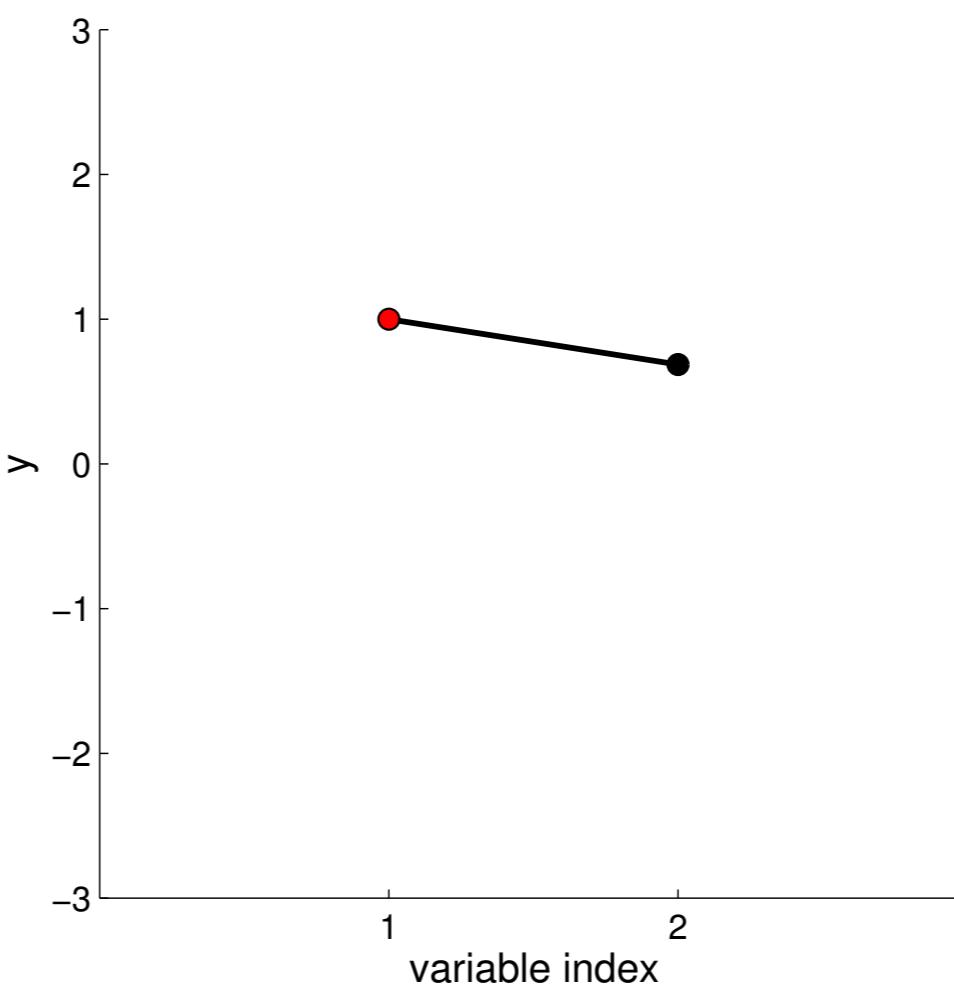
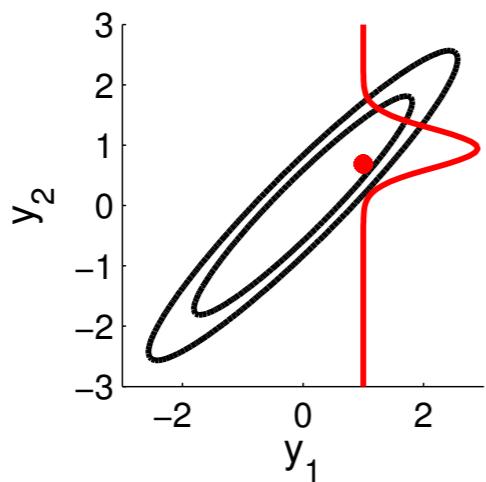
# New Visualisation



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

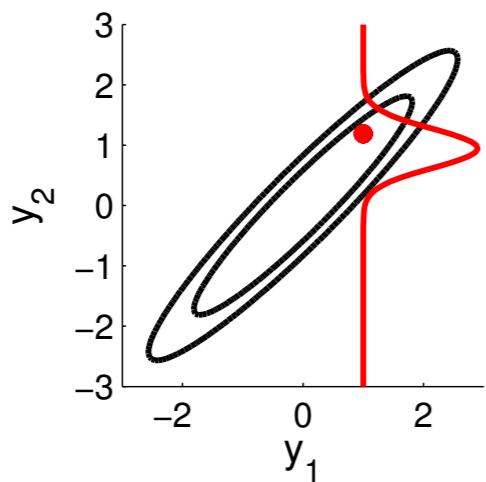


# New Visualisation

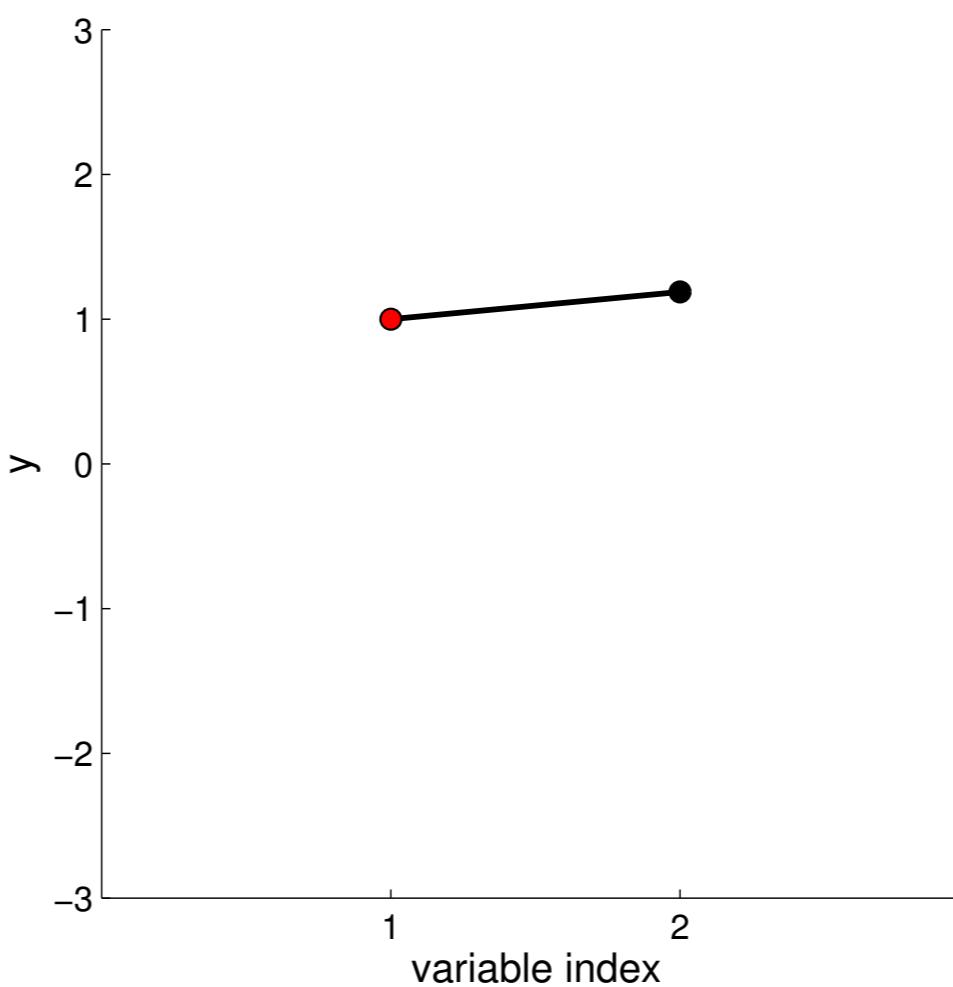


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

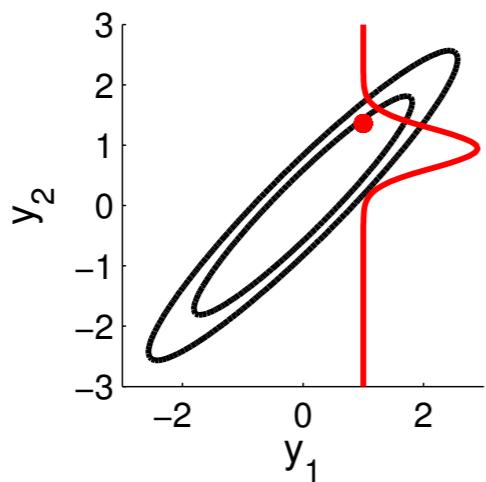
# New Visualisation



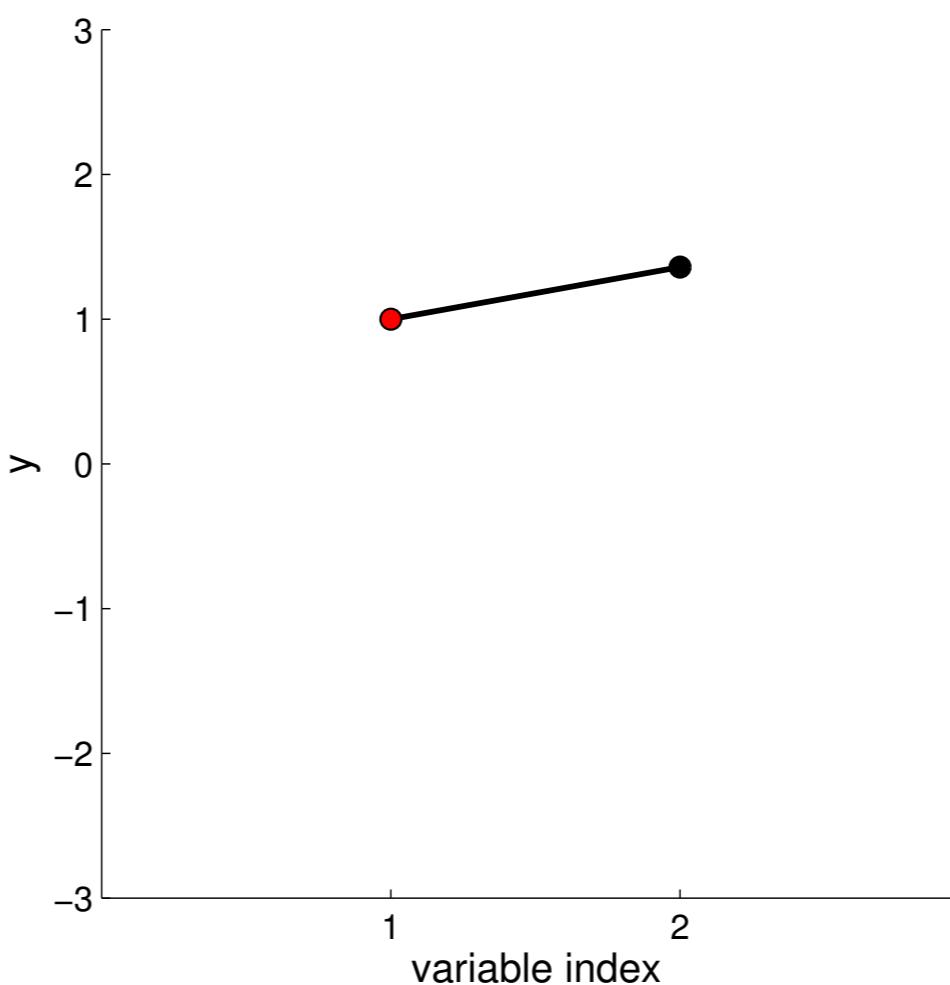
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



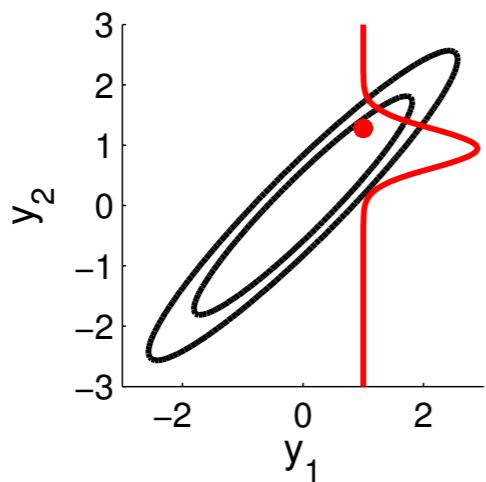
# New Visualisation



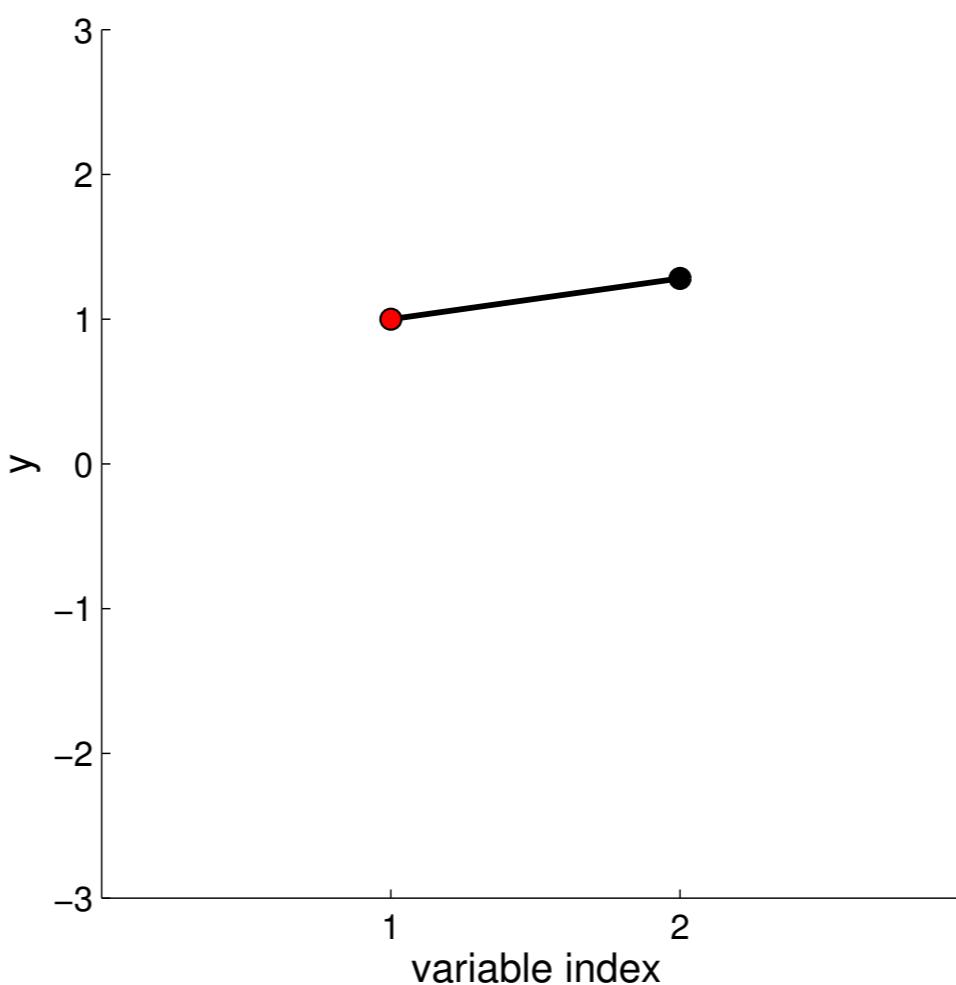
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



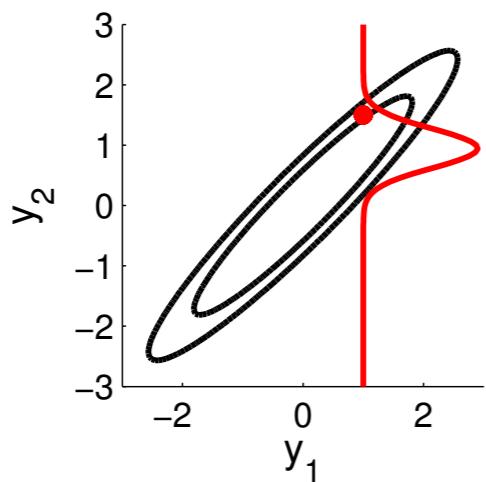
# New Visualisation



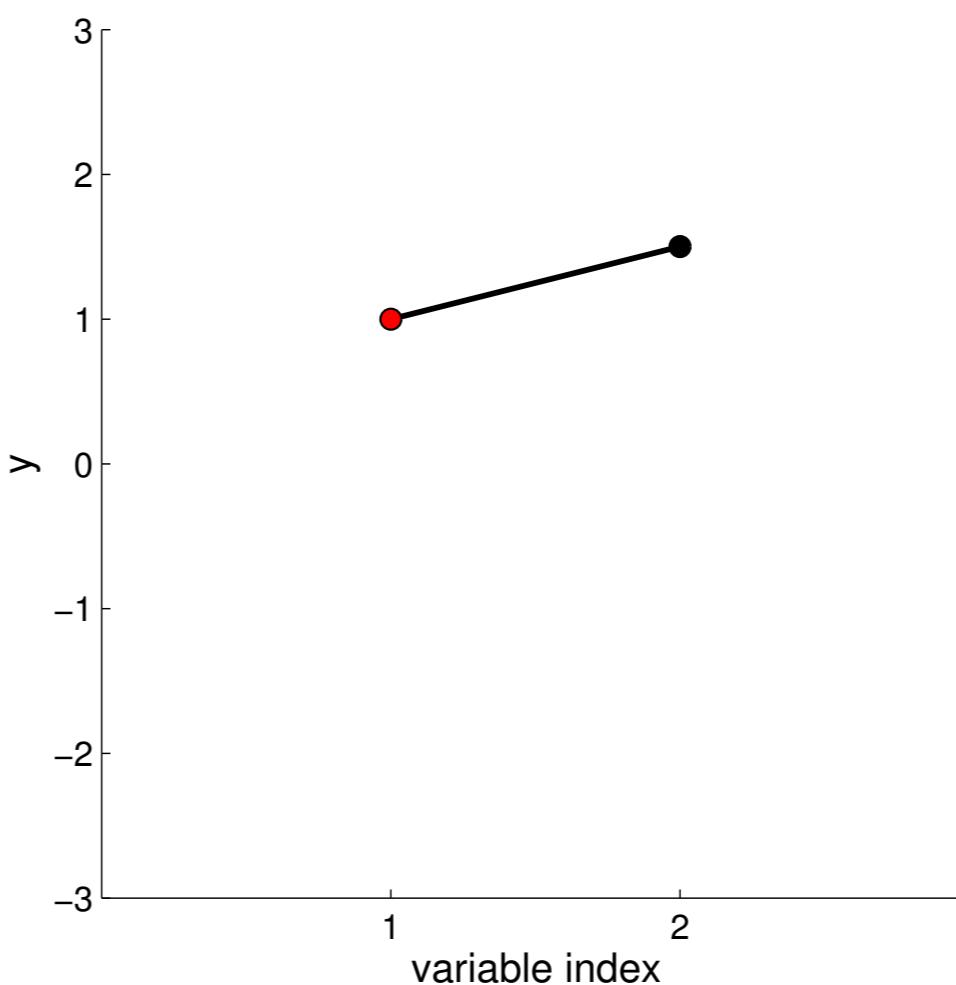
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



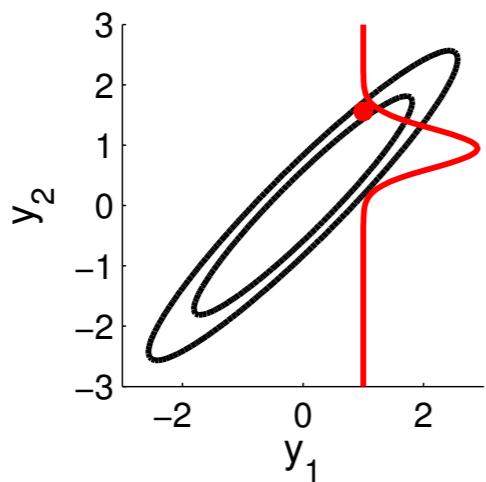
# New Visualisation



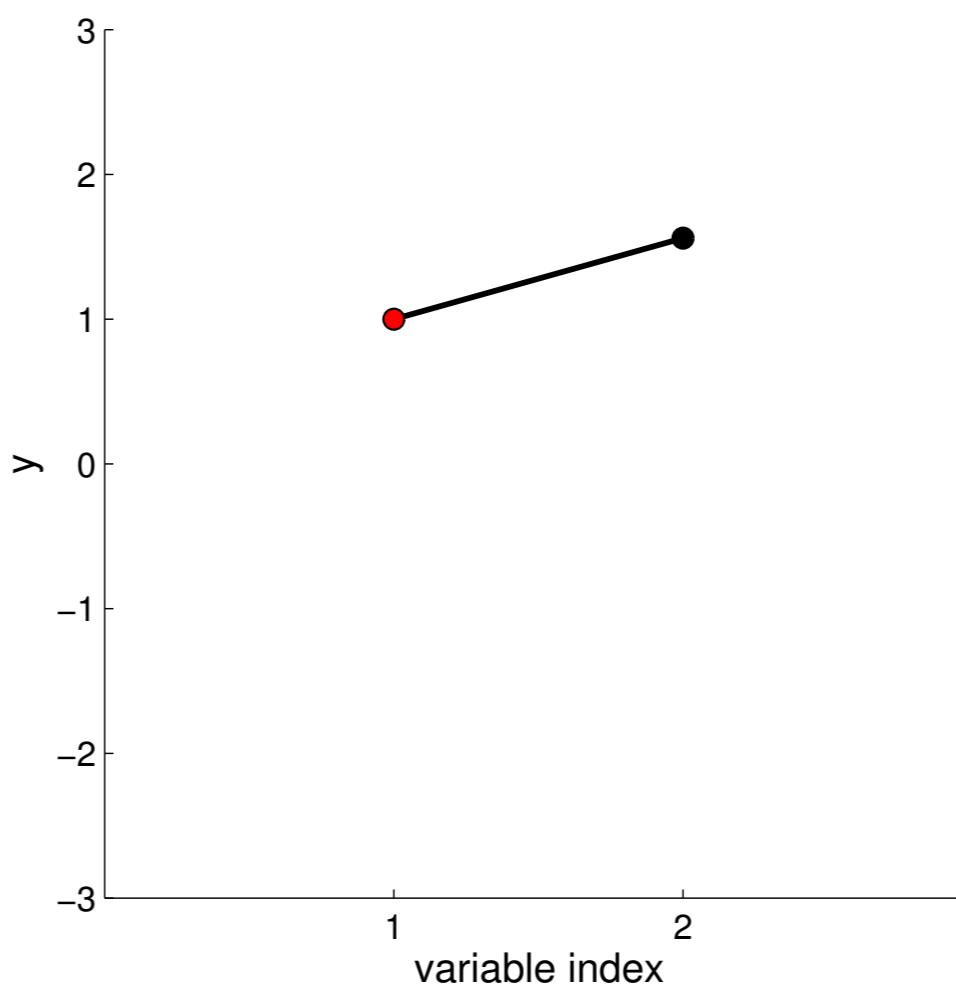
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



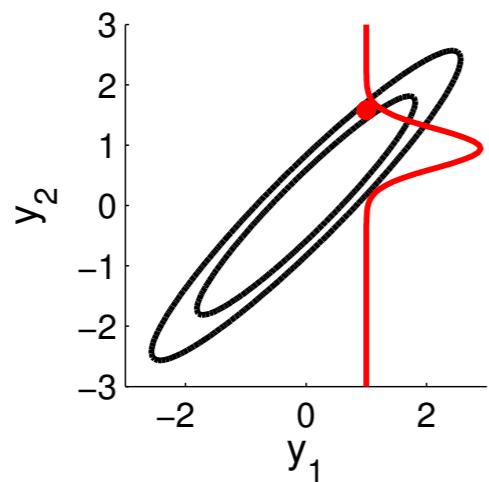
# New Visualisation



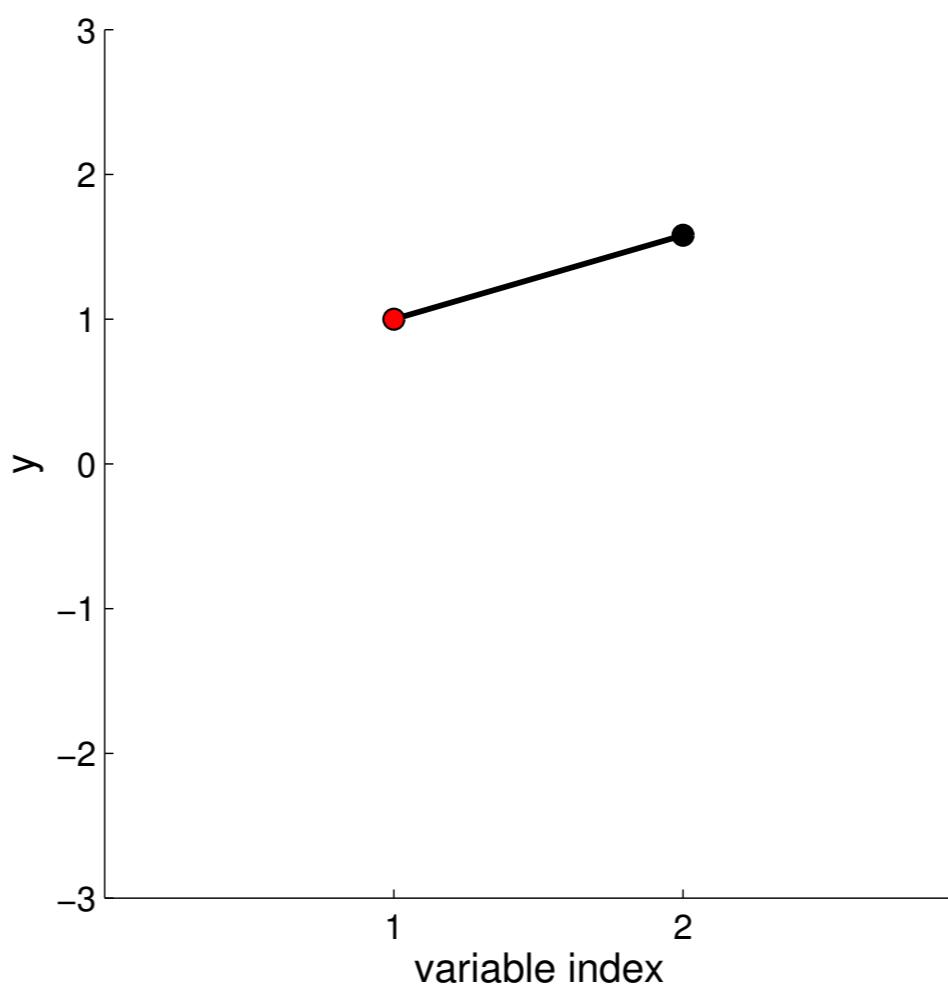
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



# New Visualisation

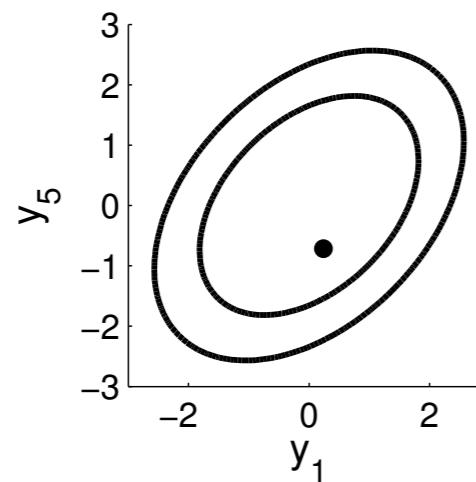


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

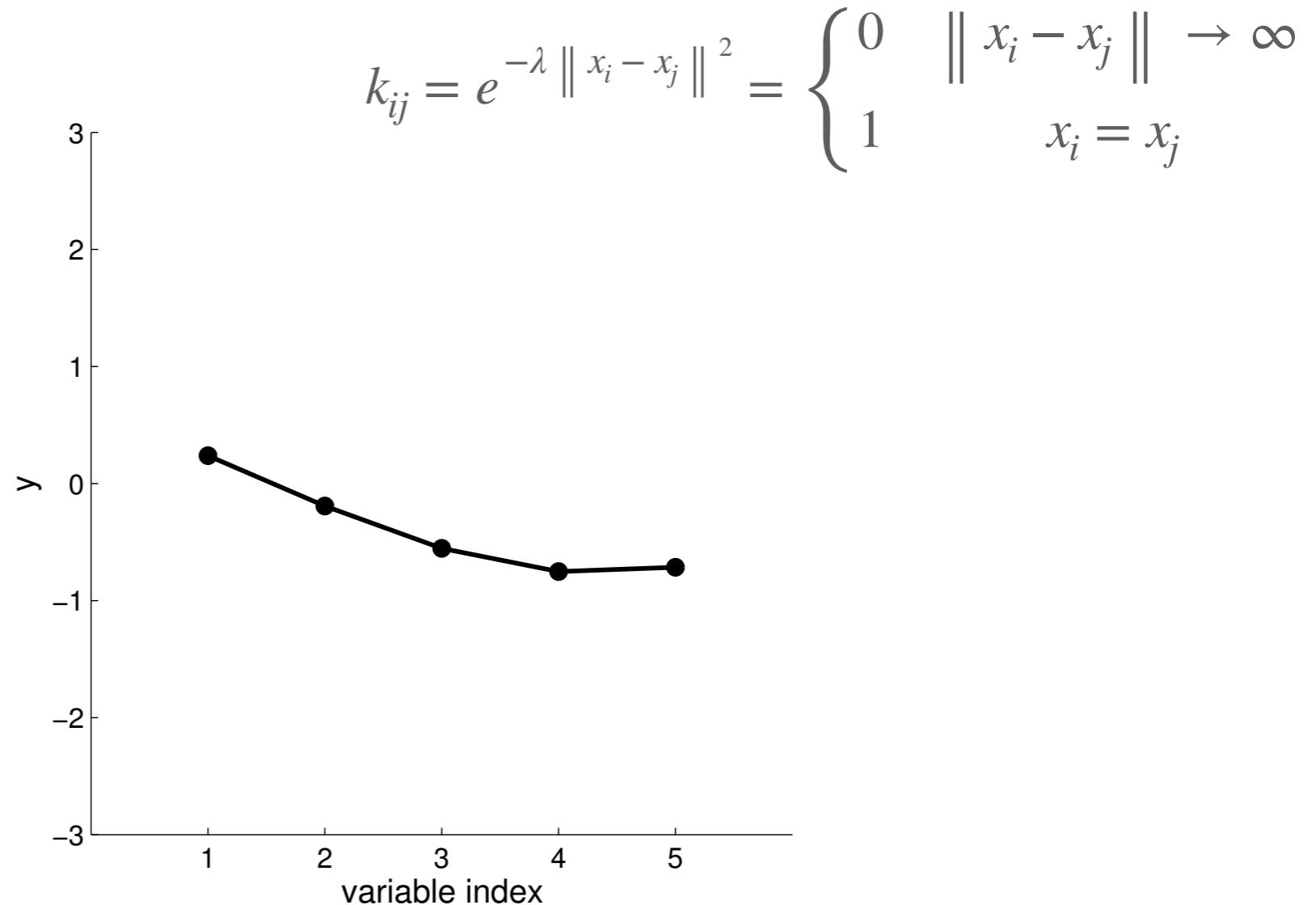


# Special covariance matrix

► Correlations fall off the further the indices of the variables!

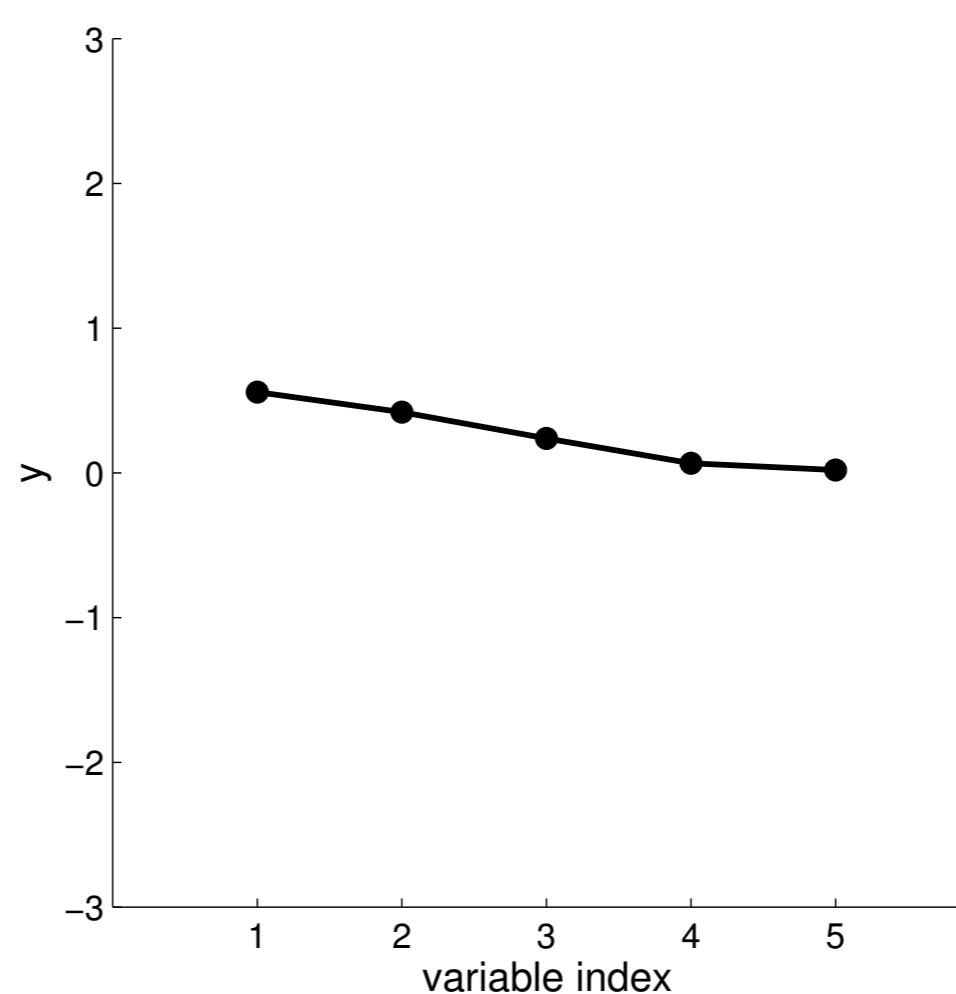
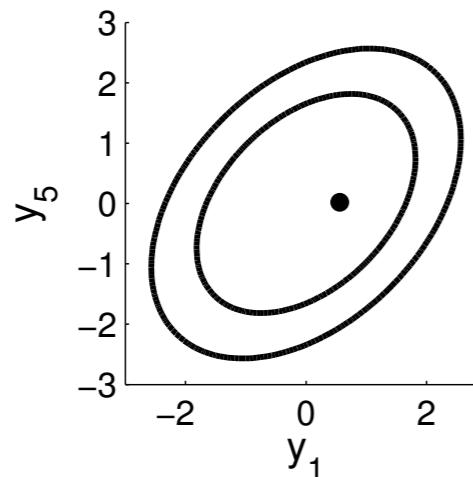


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$



# Special covariance matrix

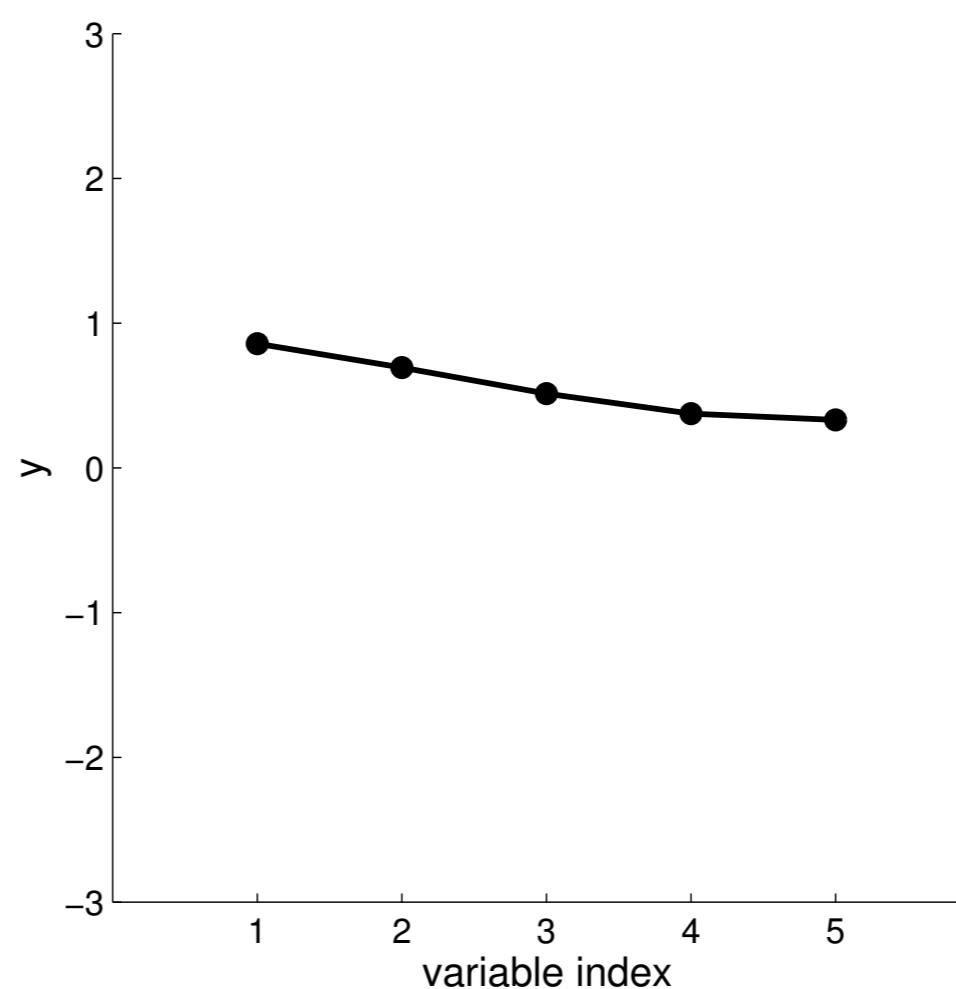
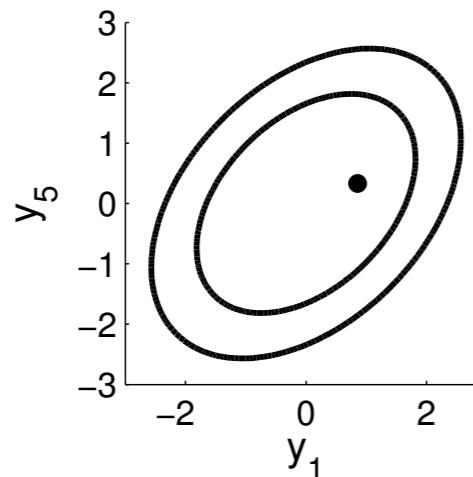
► Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

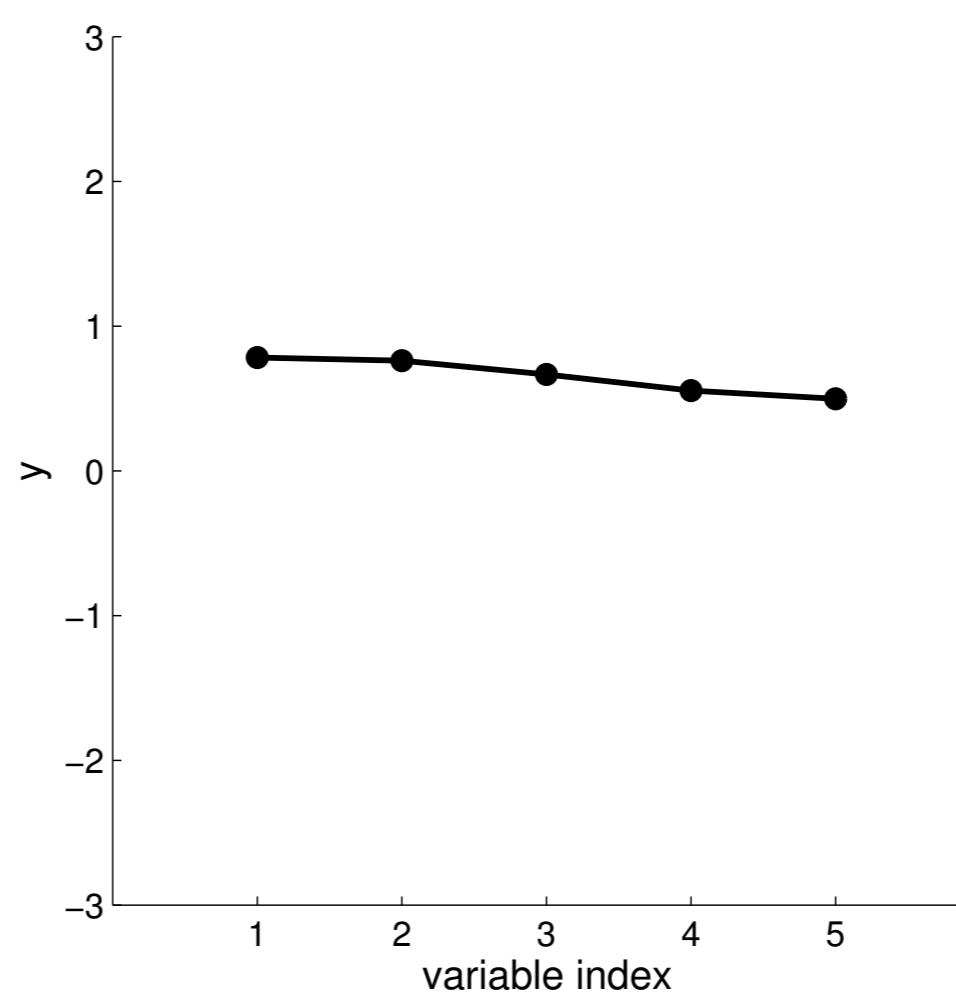
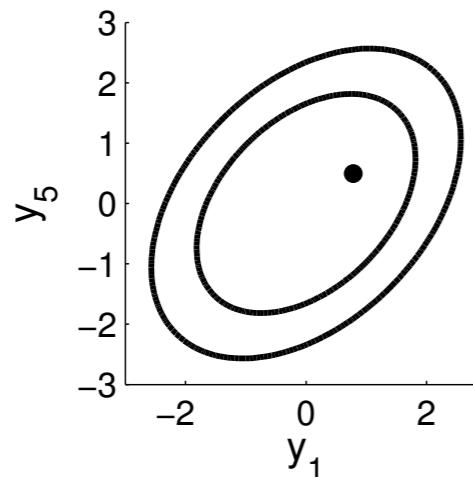
► Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

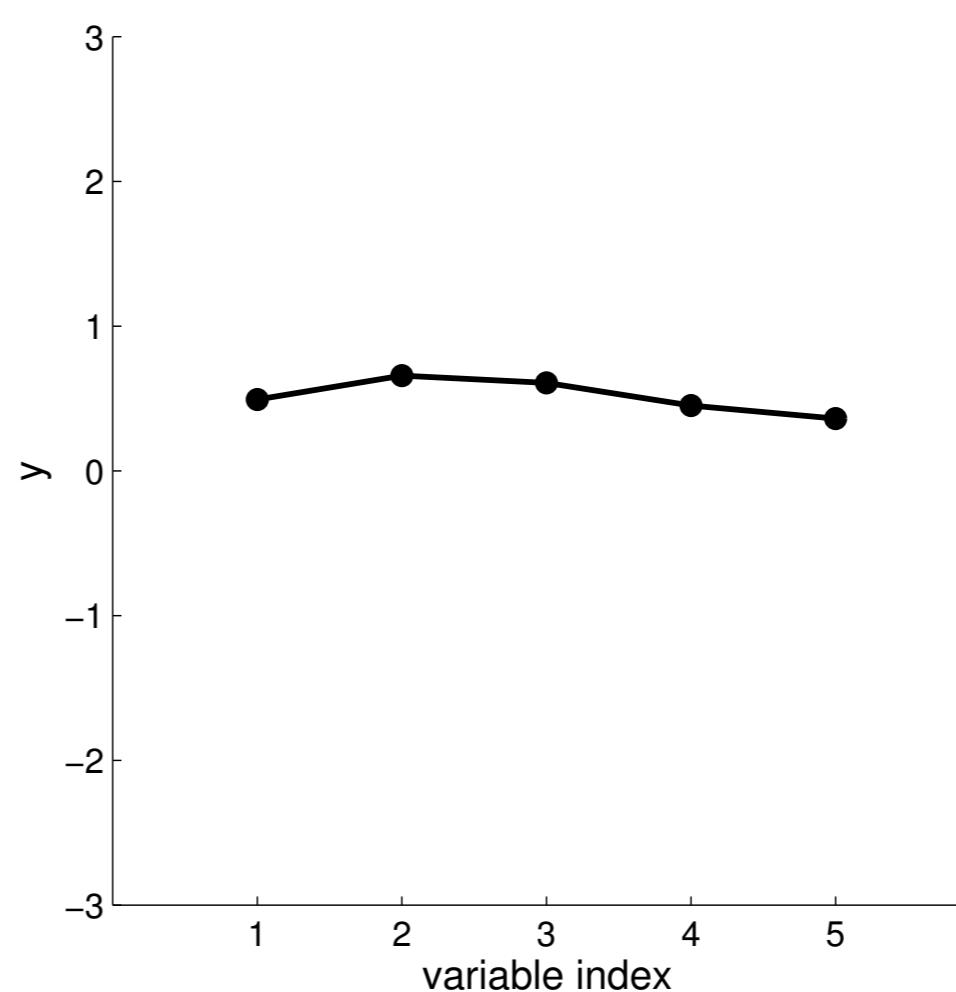
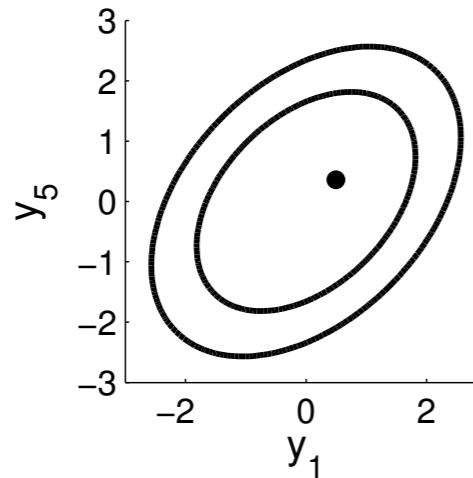
- Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

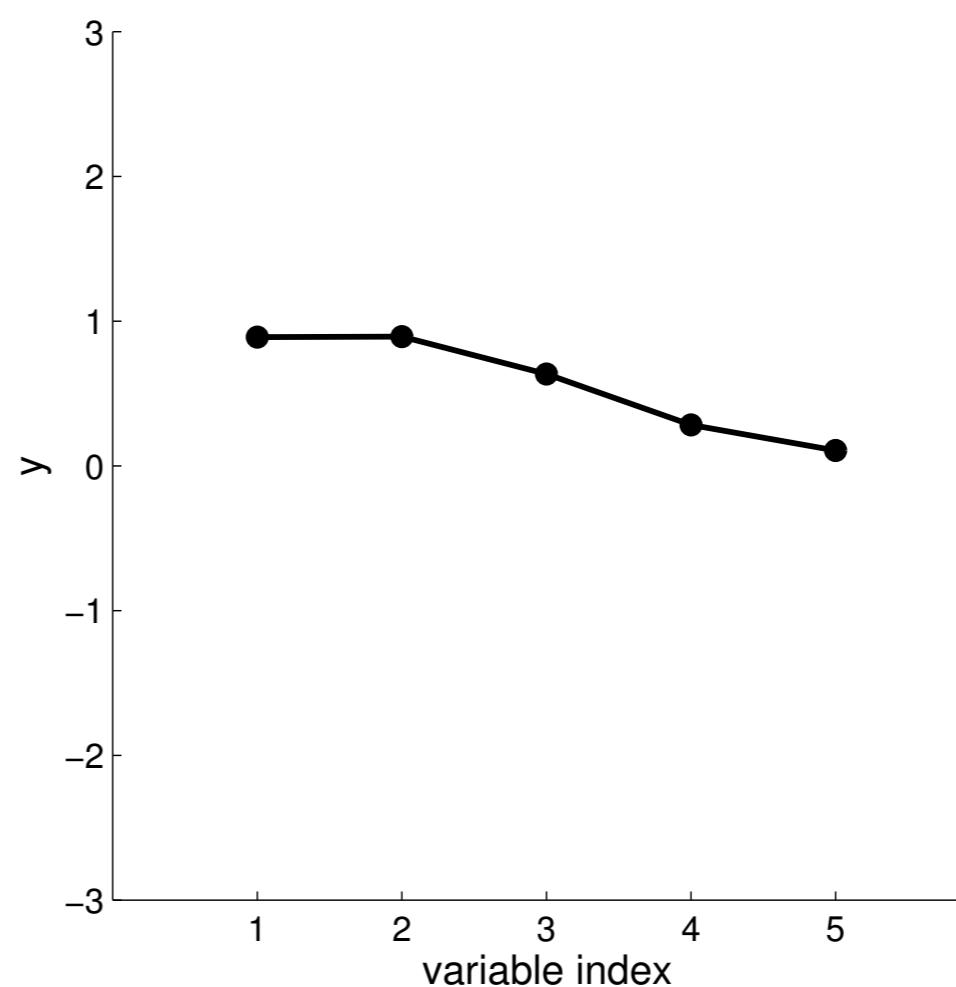
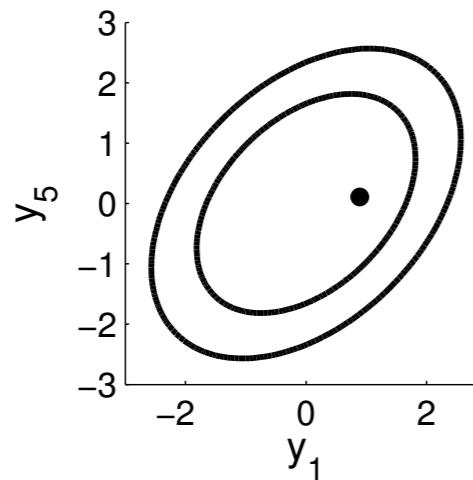
► Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

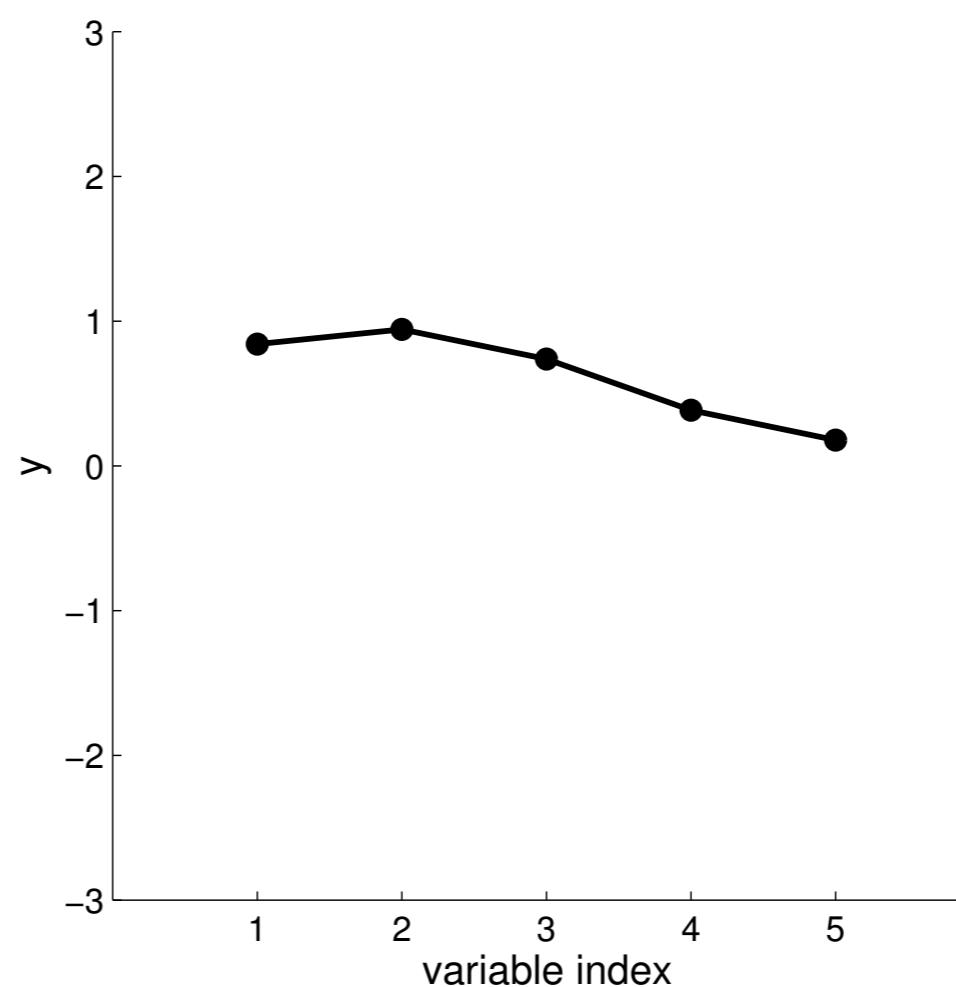
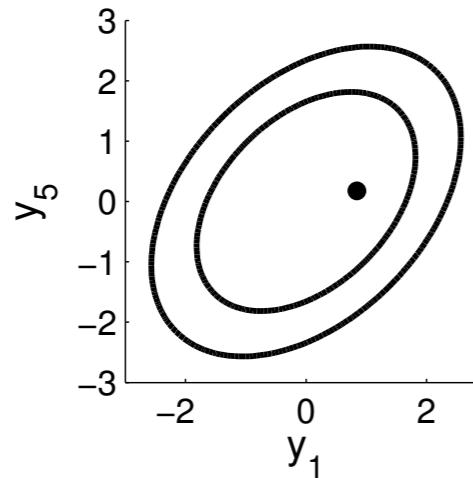
► Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

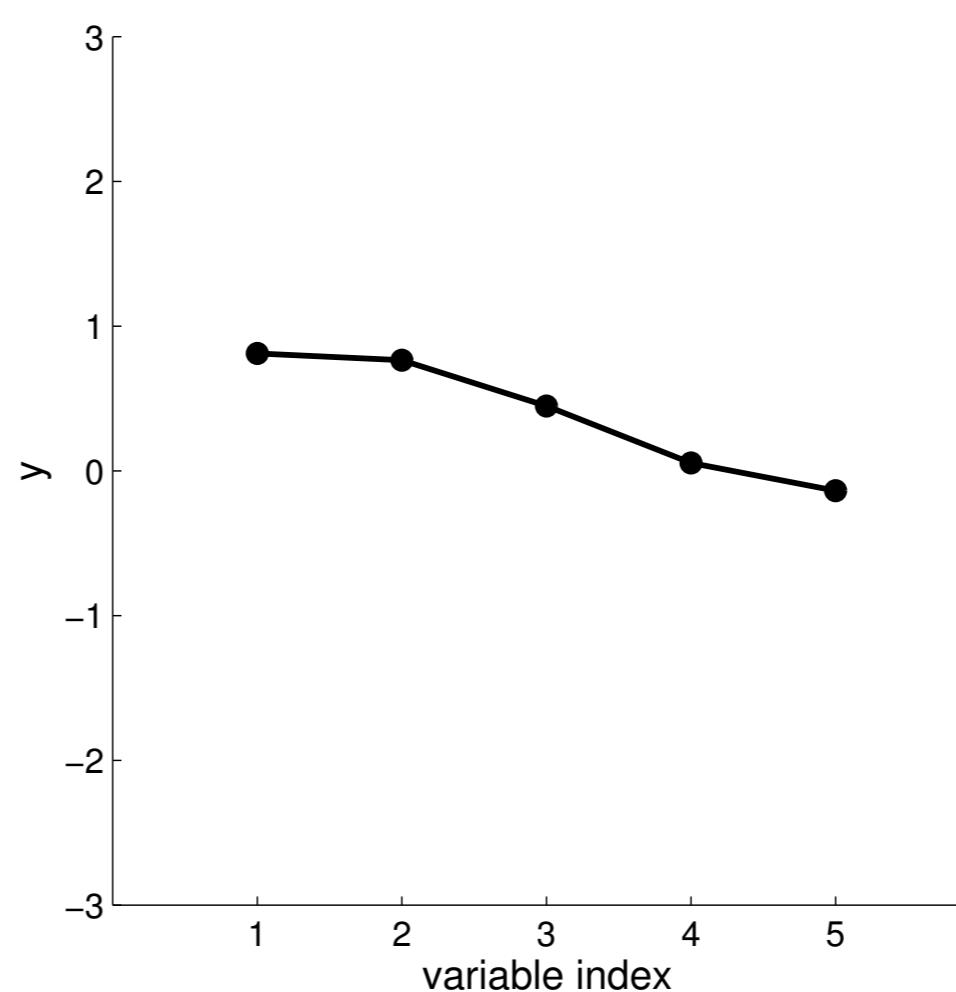
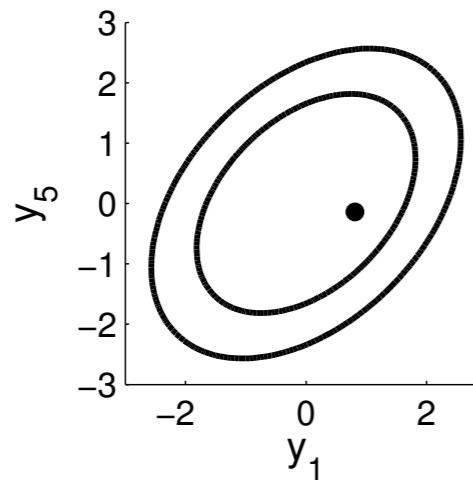
- Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

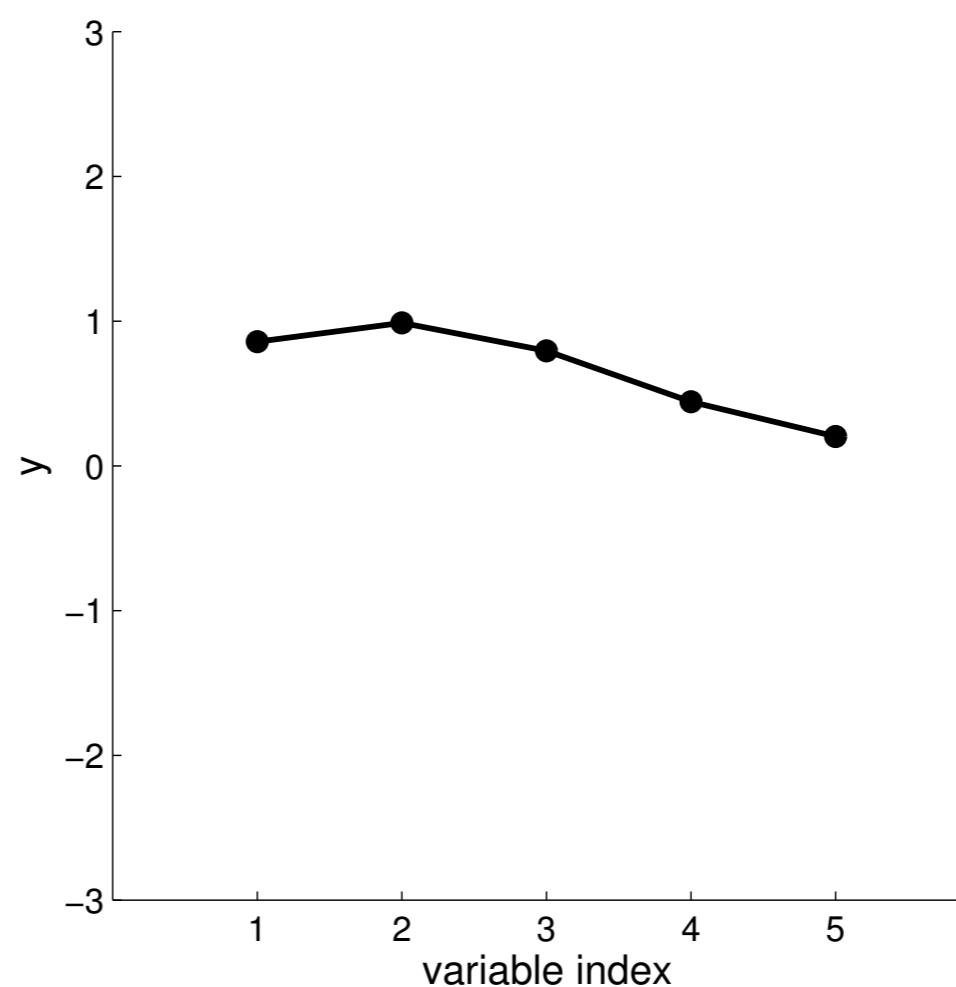
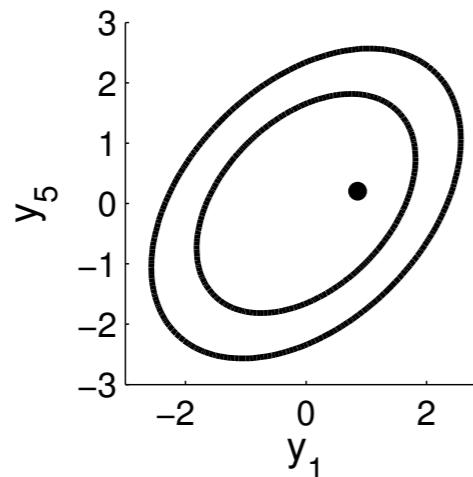
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

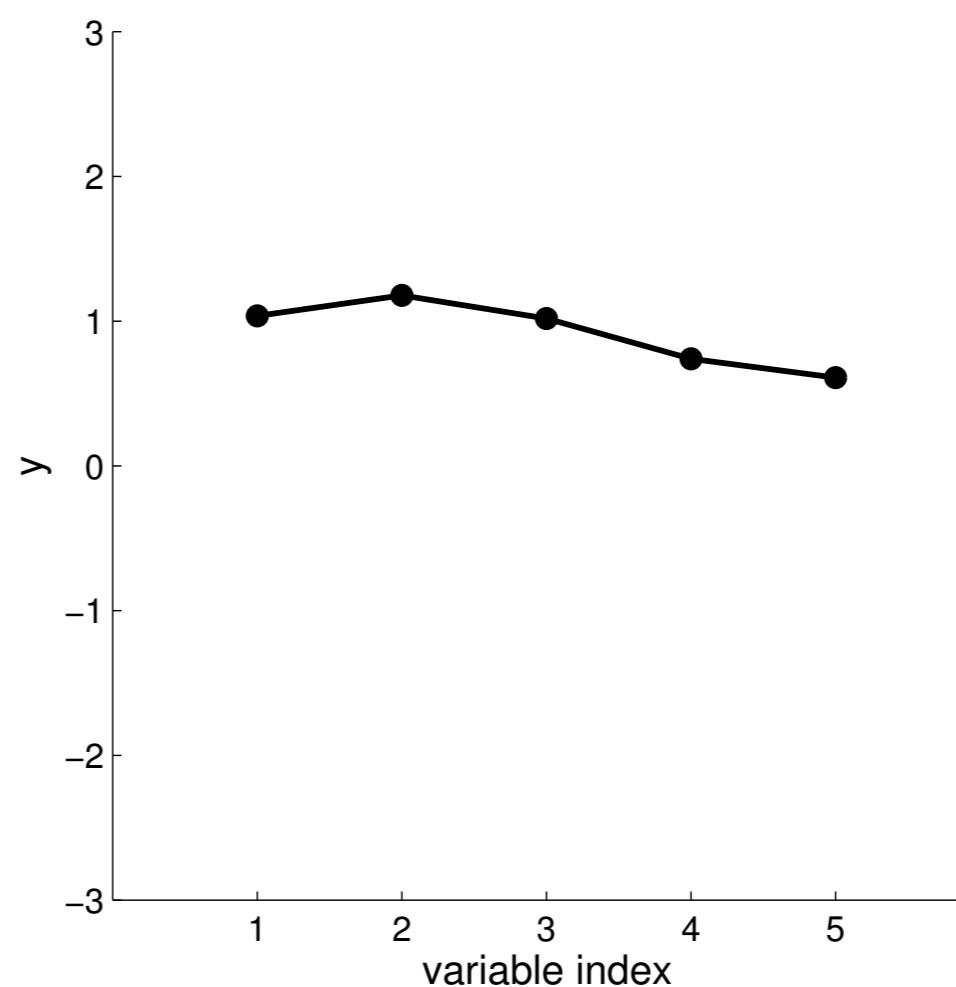
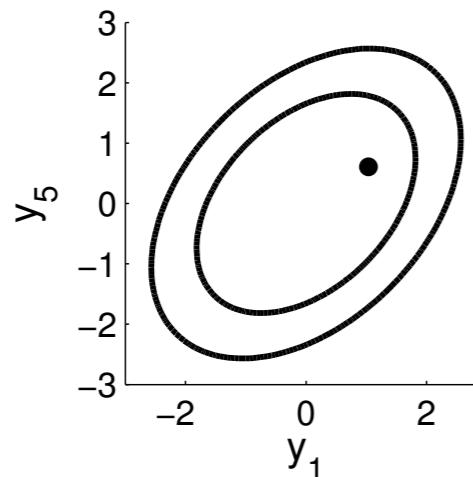
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

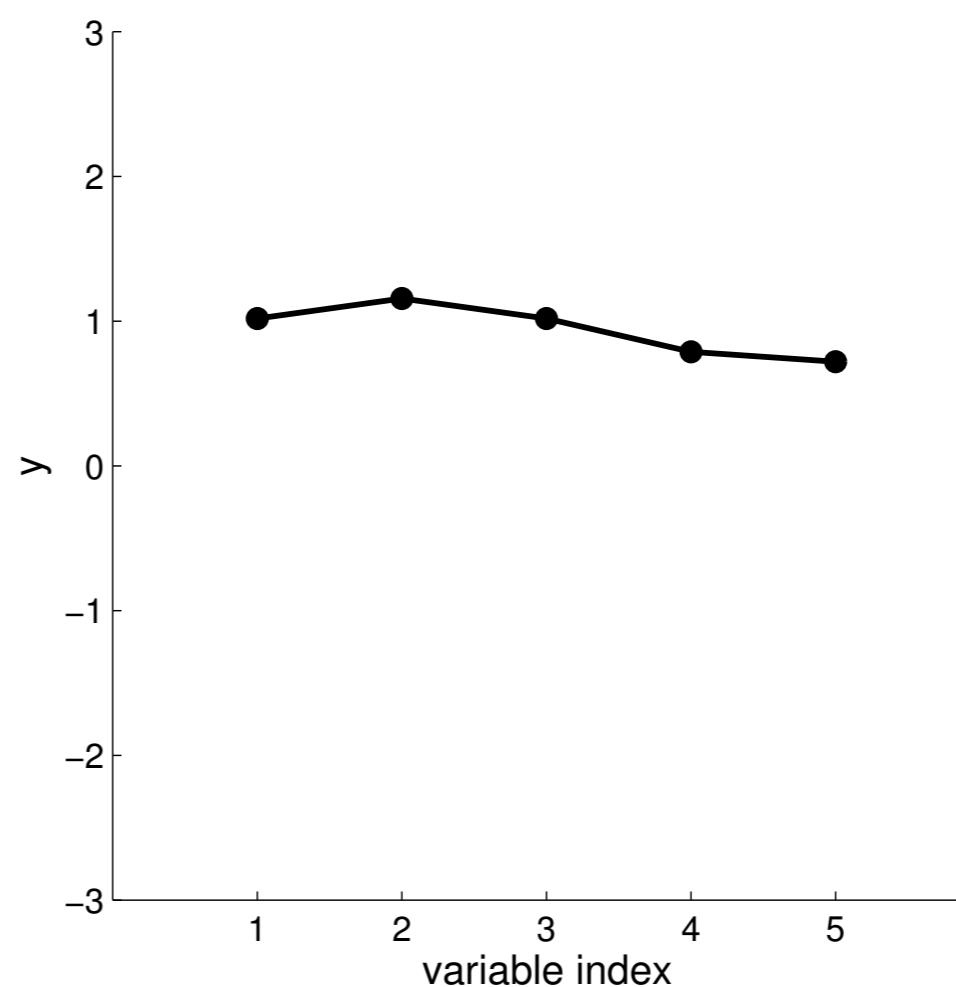
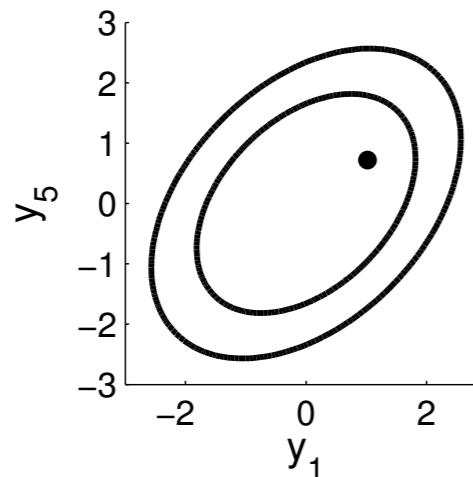
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

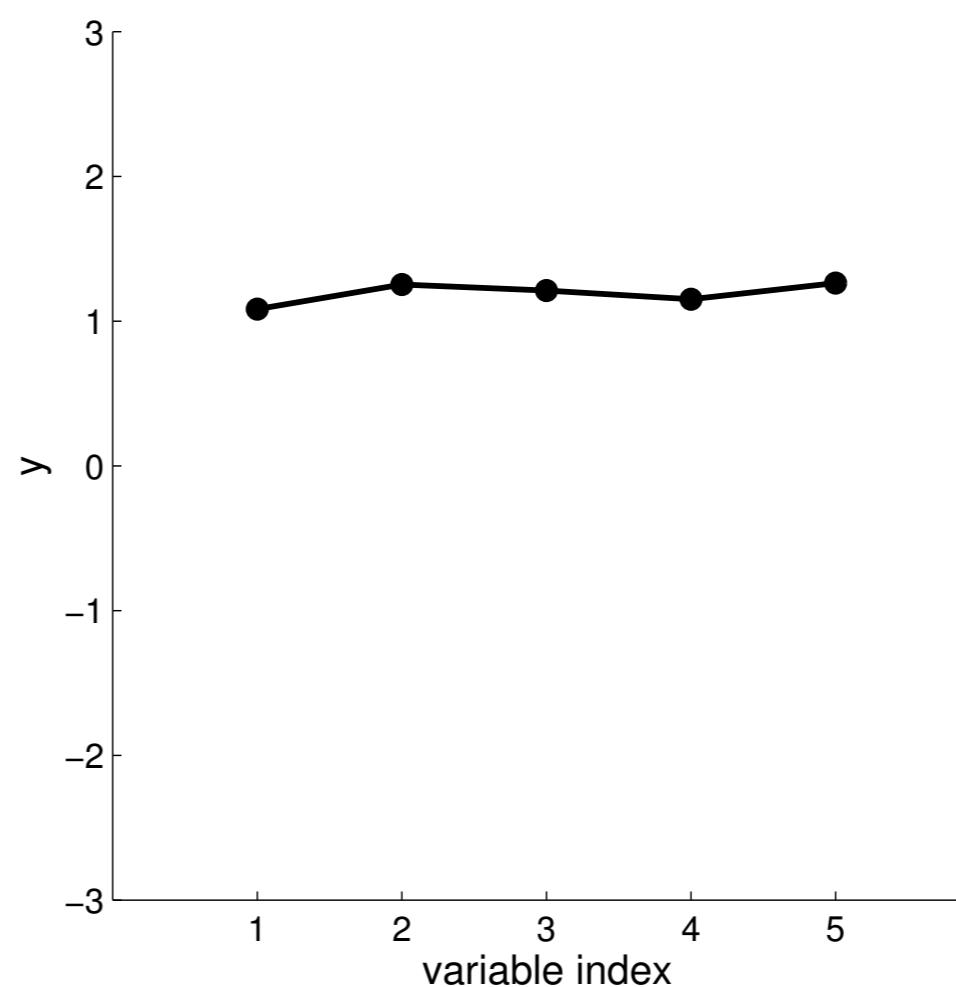
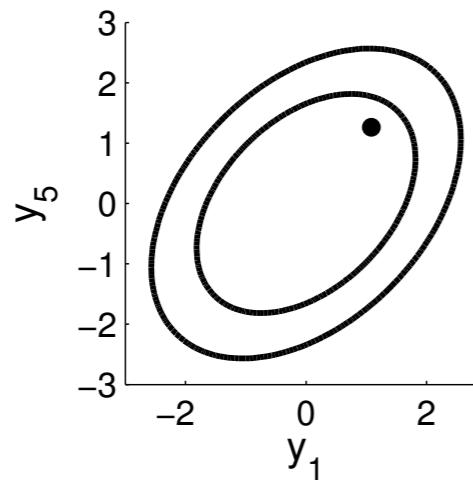
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

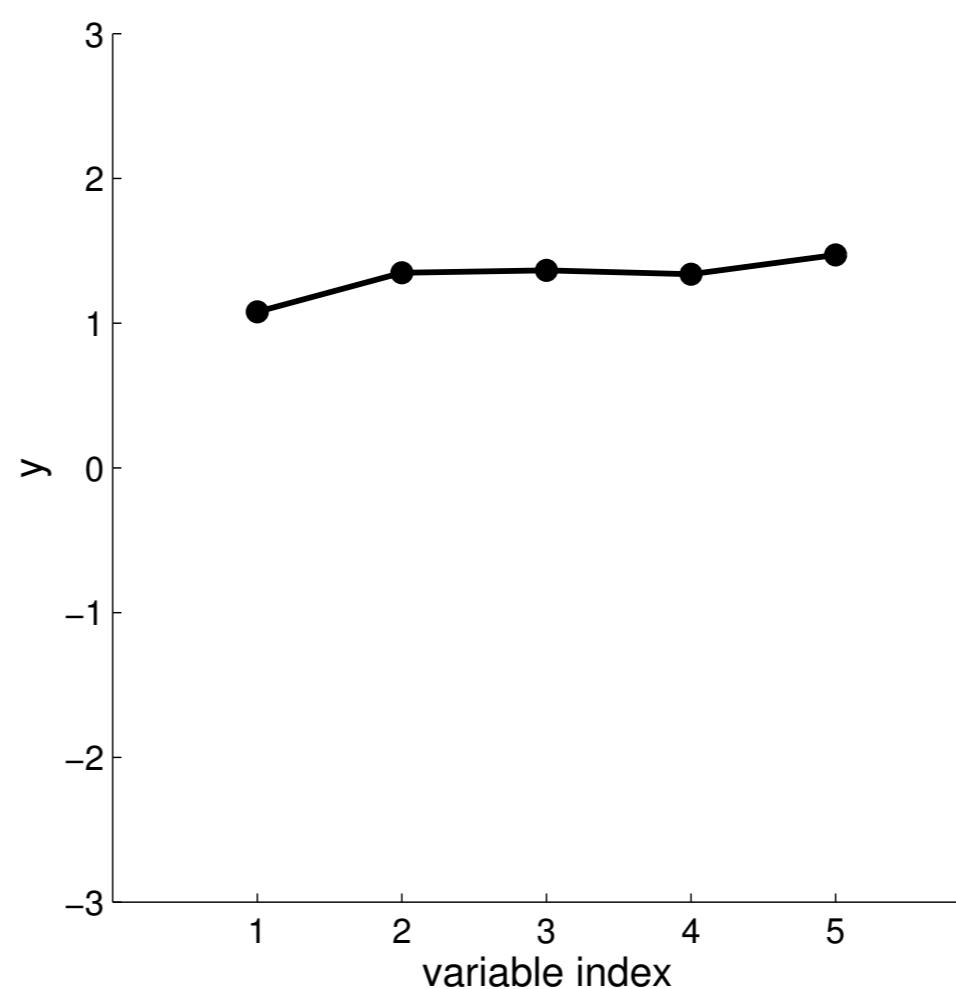
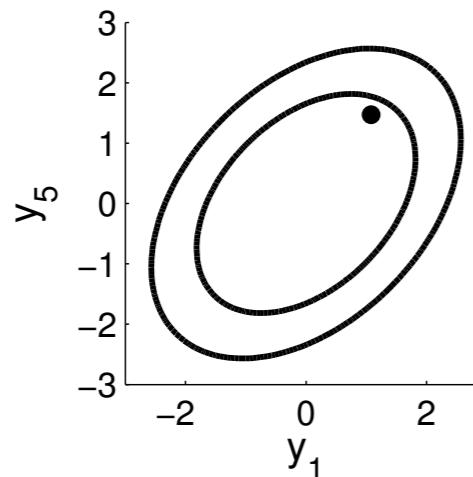
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

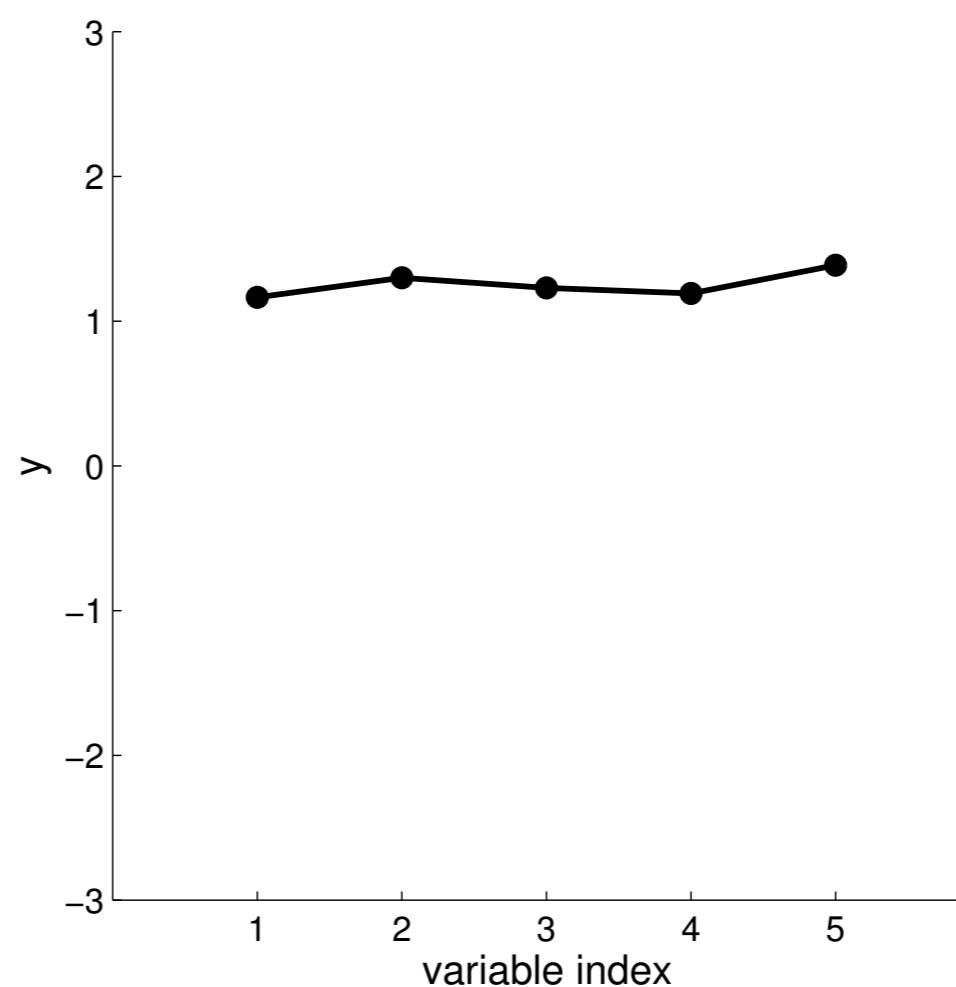
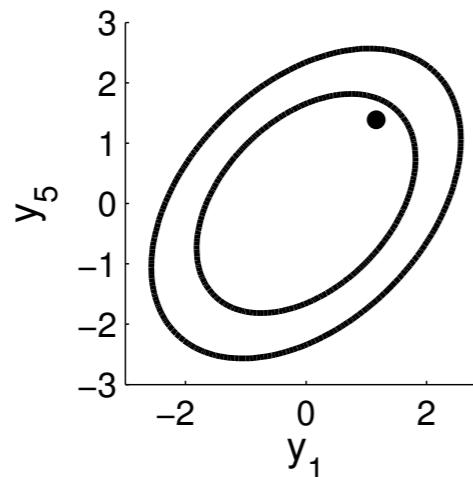
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

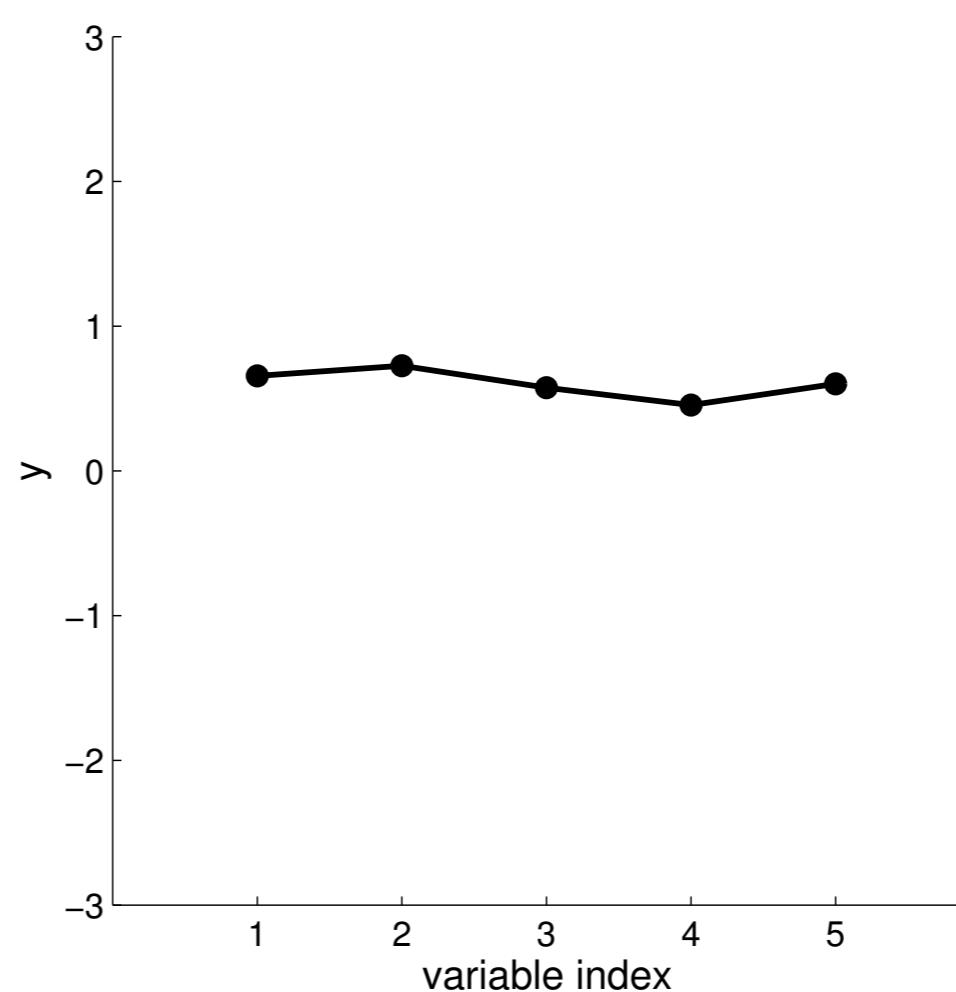
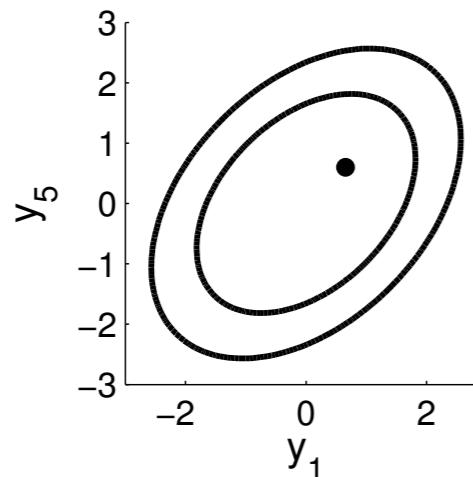
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

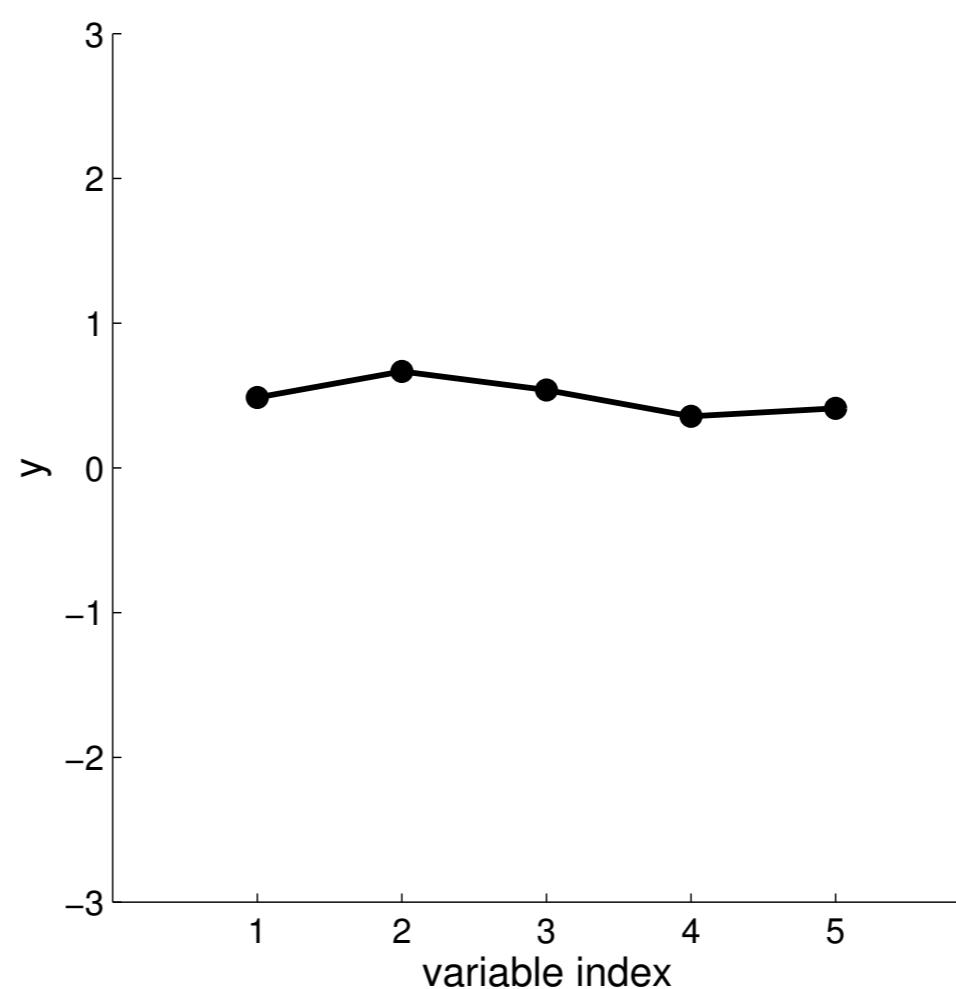
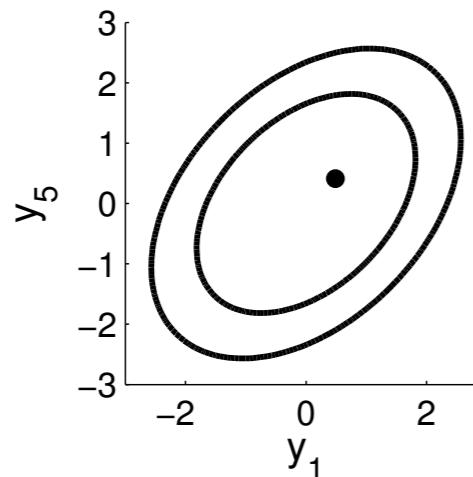
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

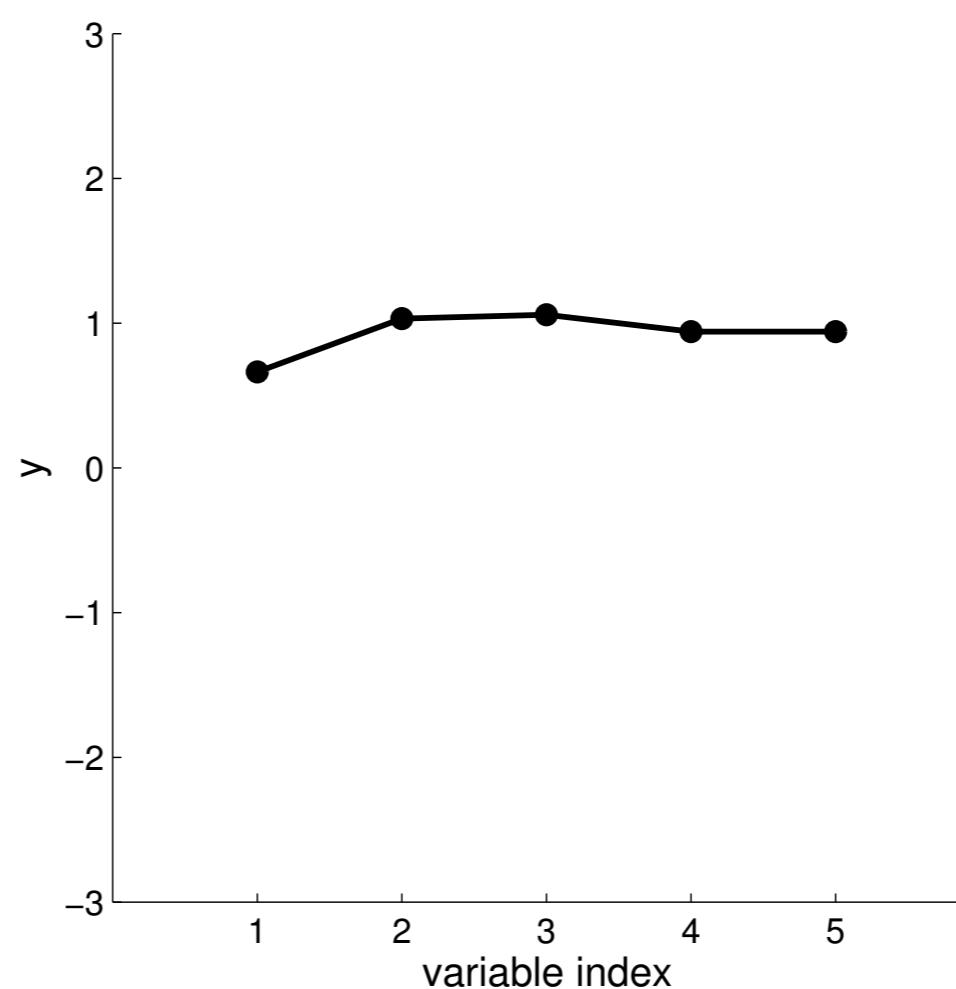
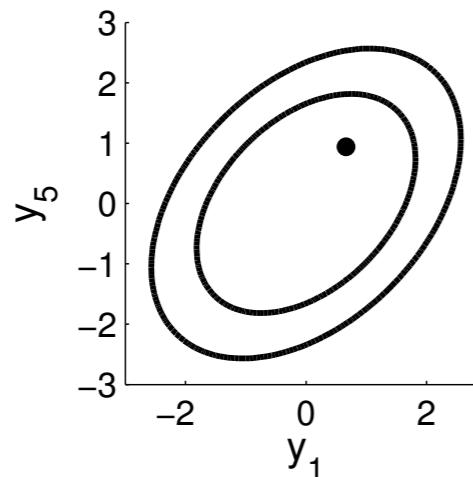
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

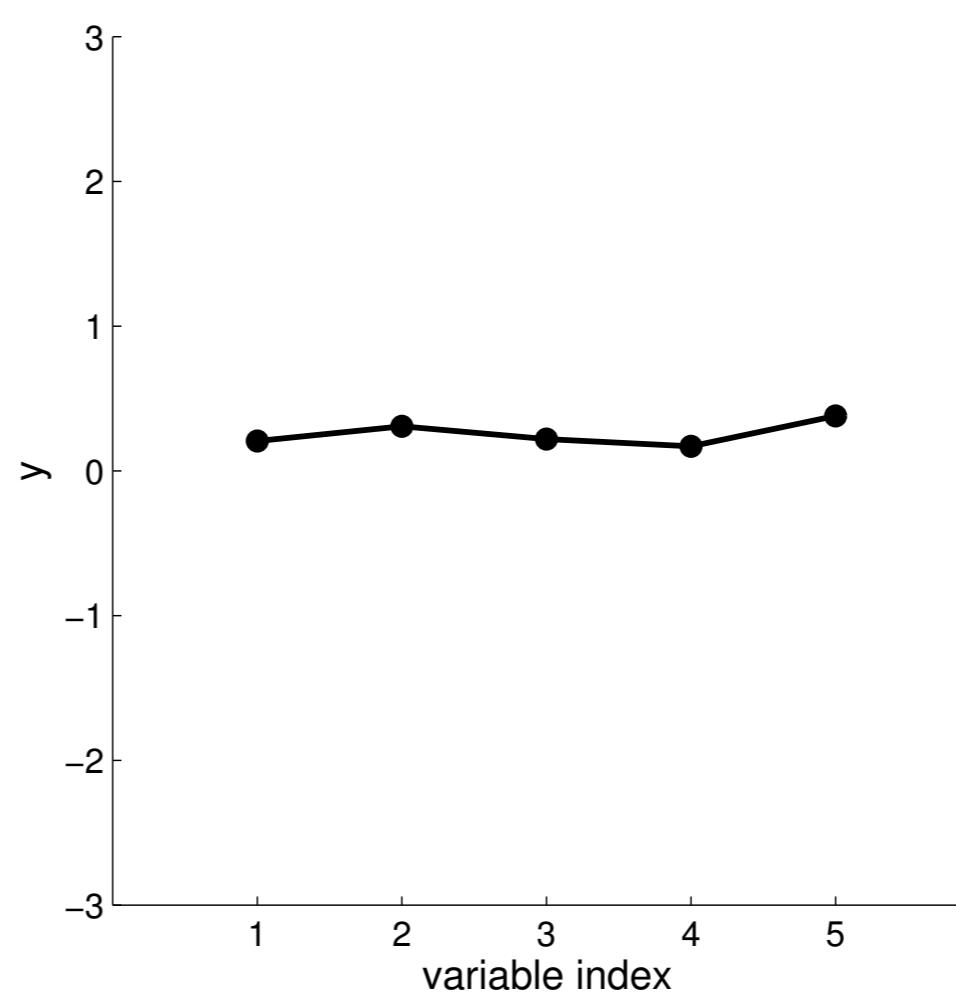
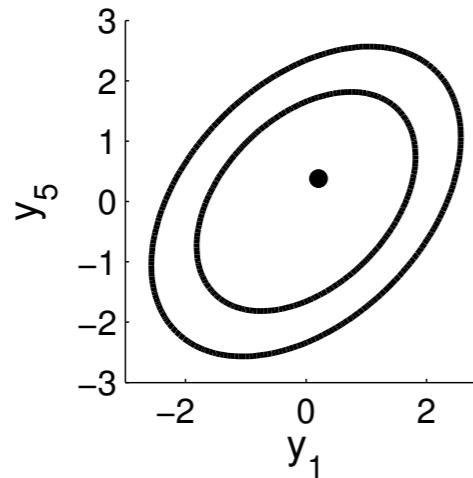
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

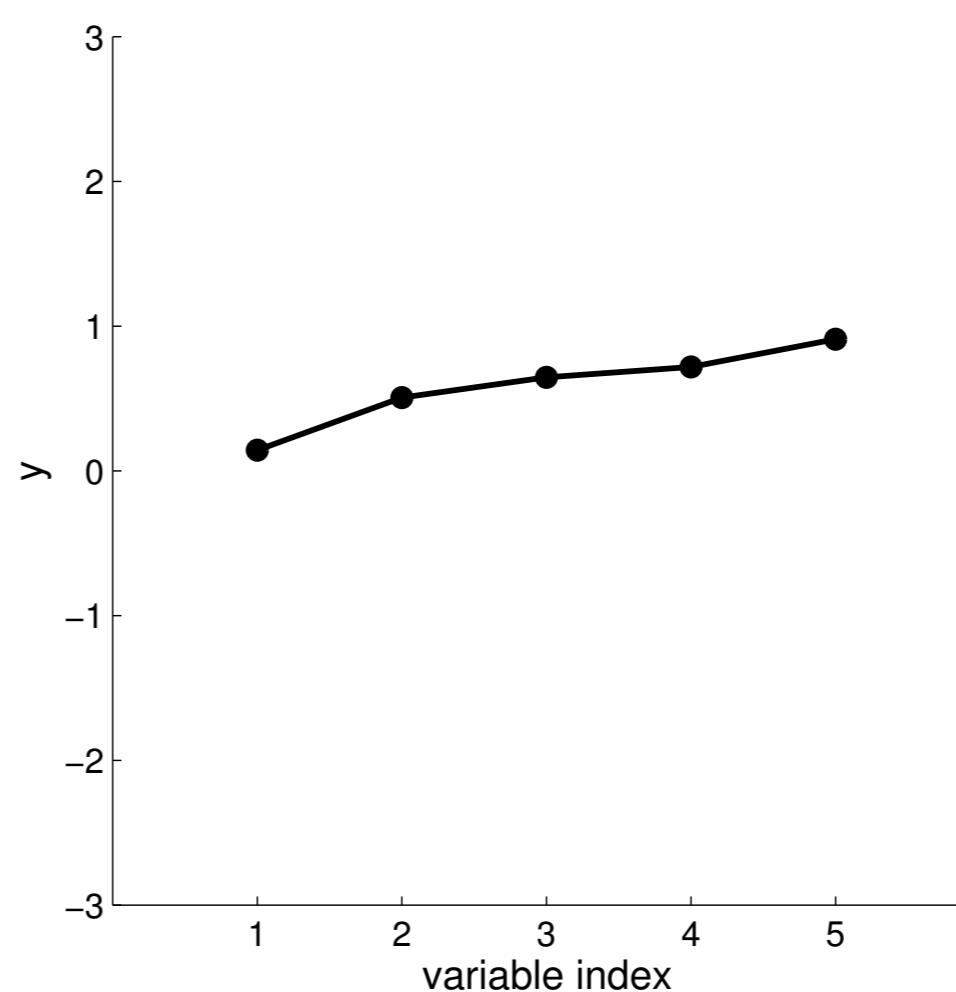
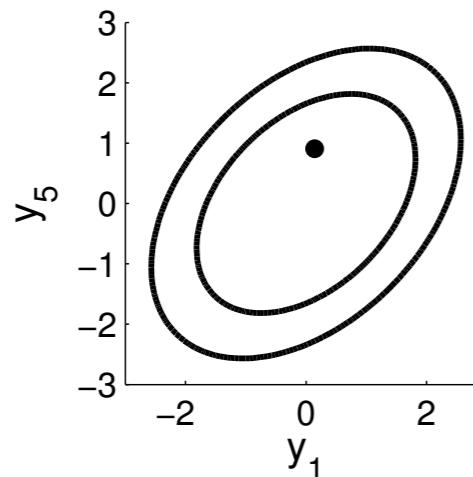
Correlations fall off the further the indices of the variables!



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

# Special covariance matrix

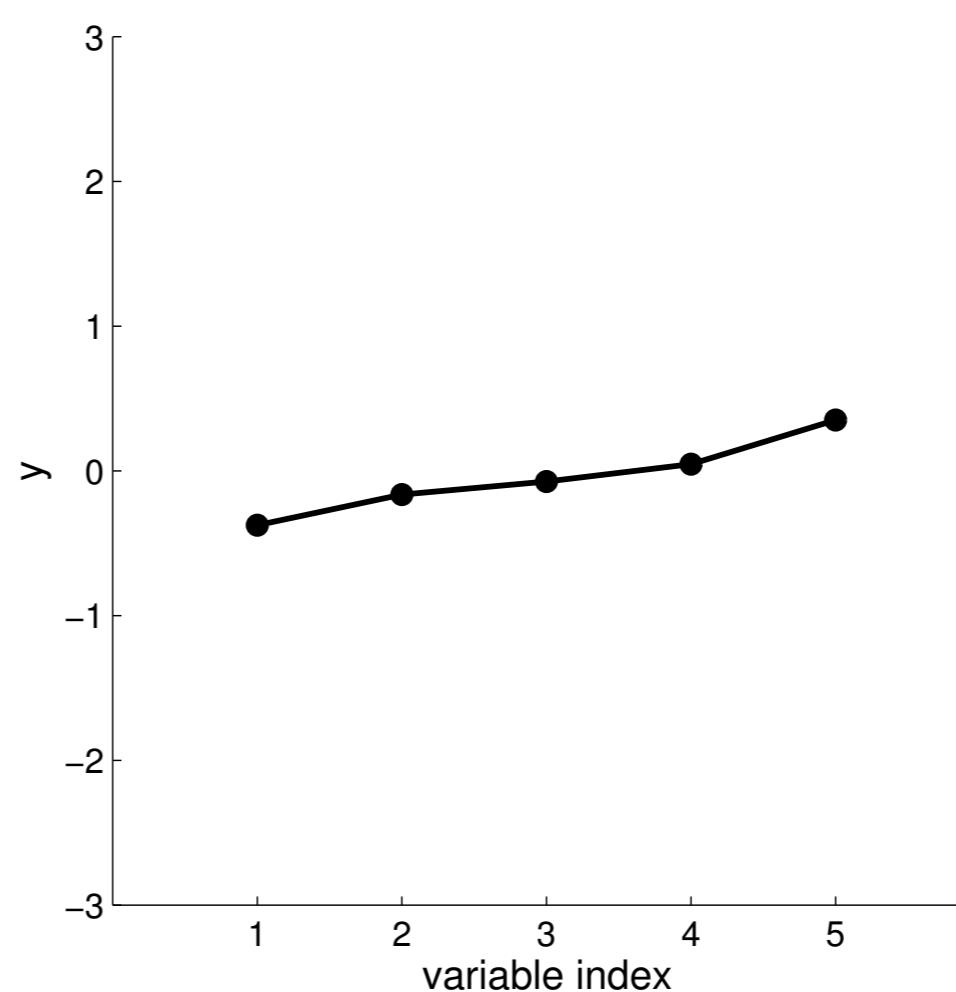
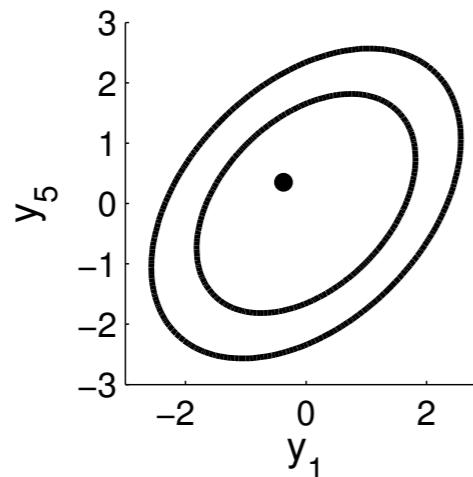
Correlations fall off the further the indices of the variables!



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

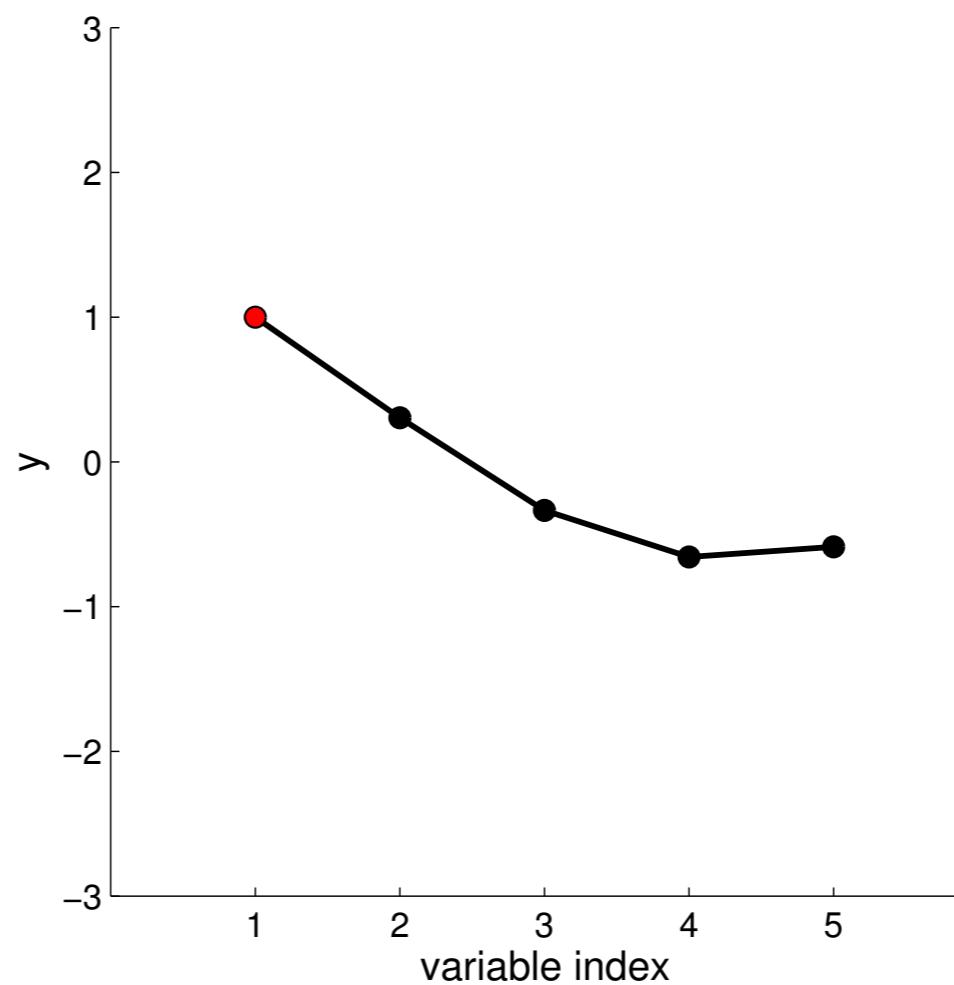
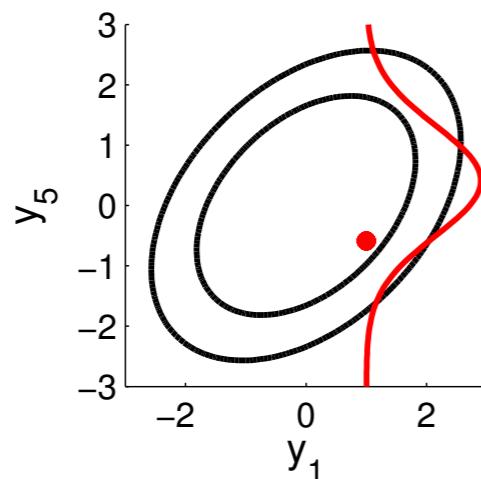
# Special covariance matrix

Correlations fall off the further the indices of the variables!



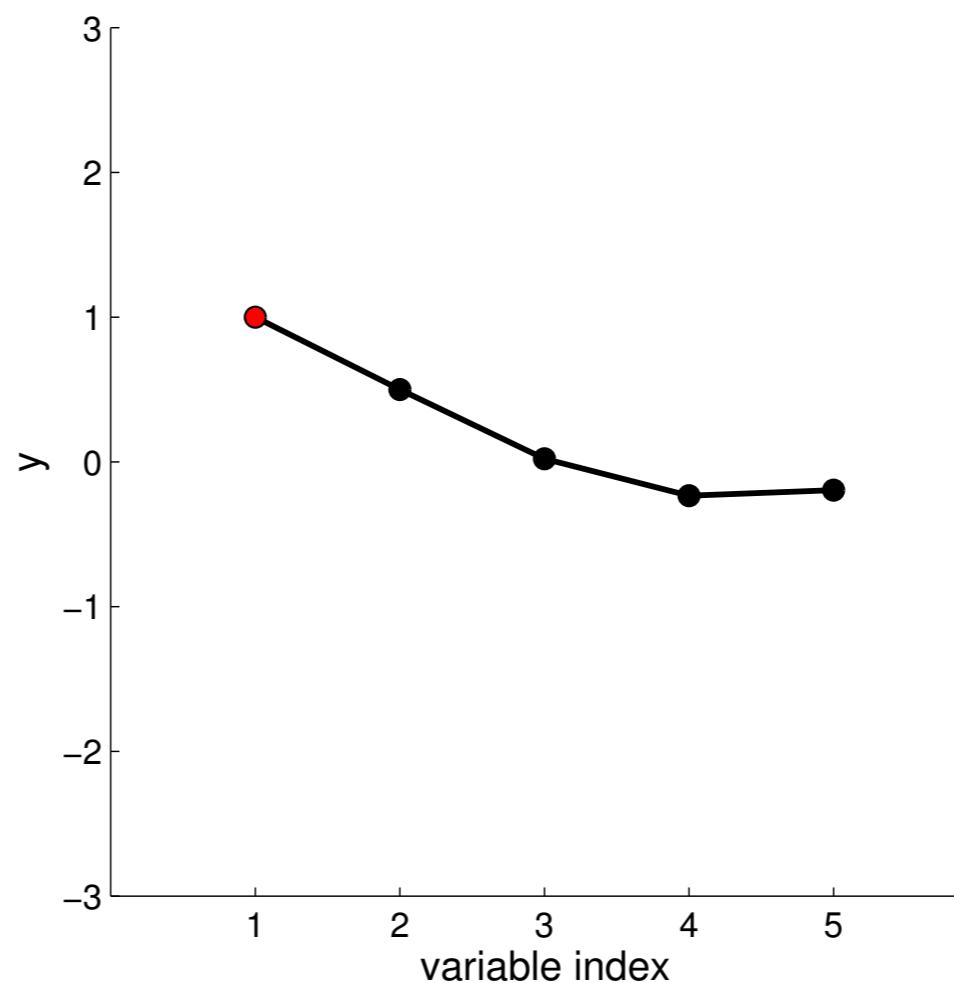
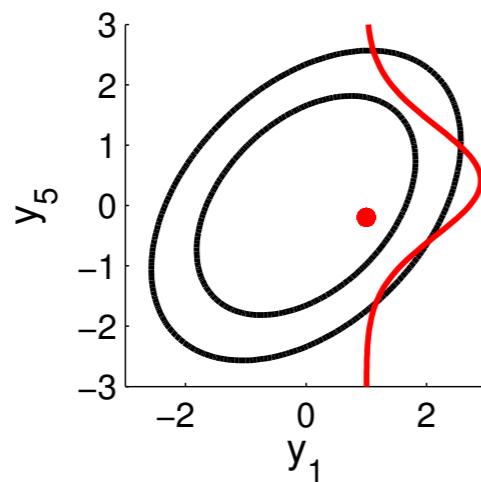
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



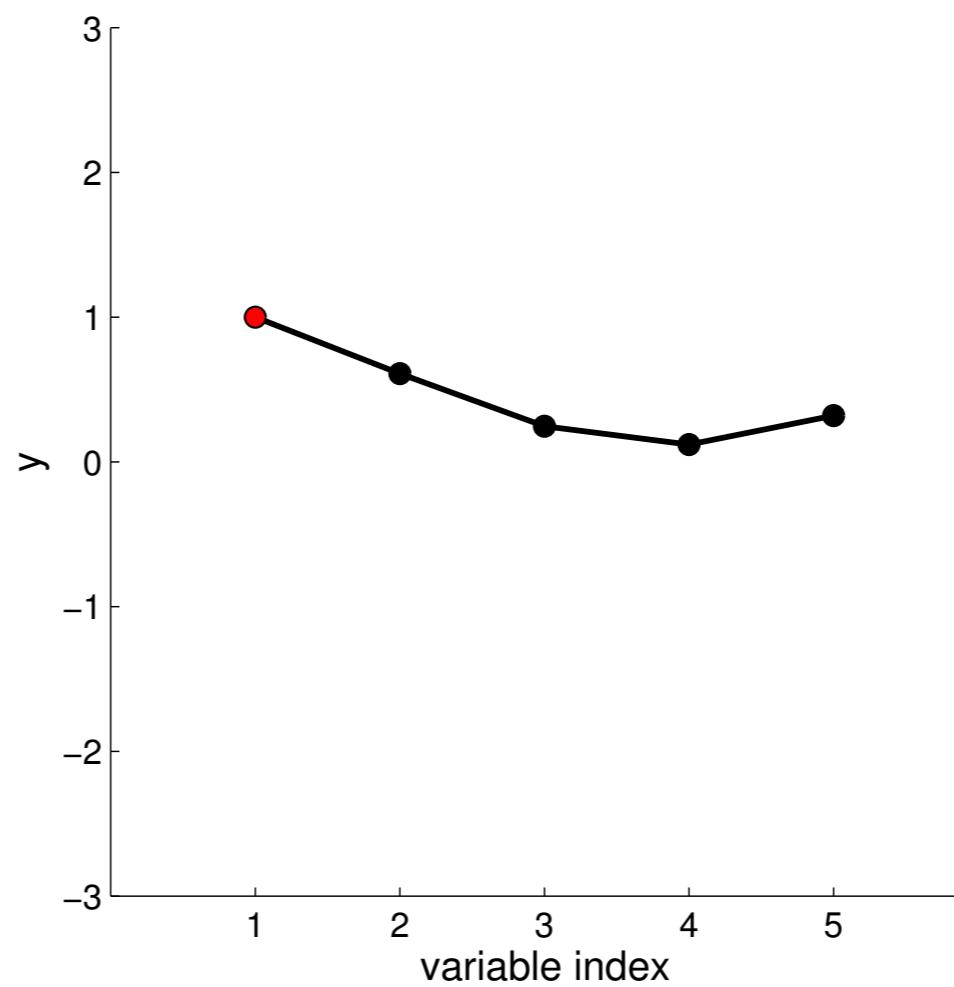
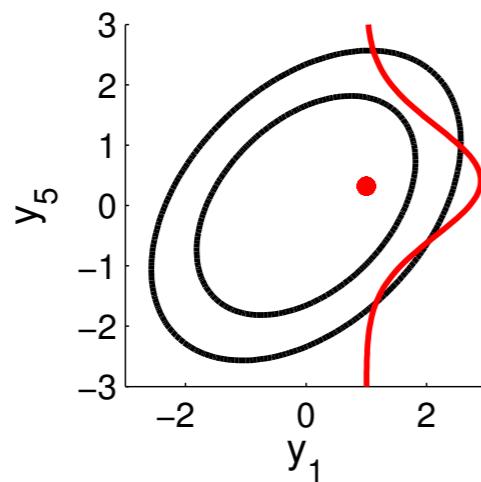
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



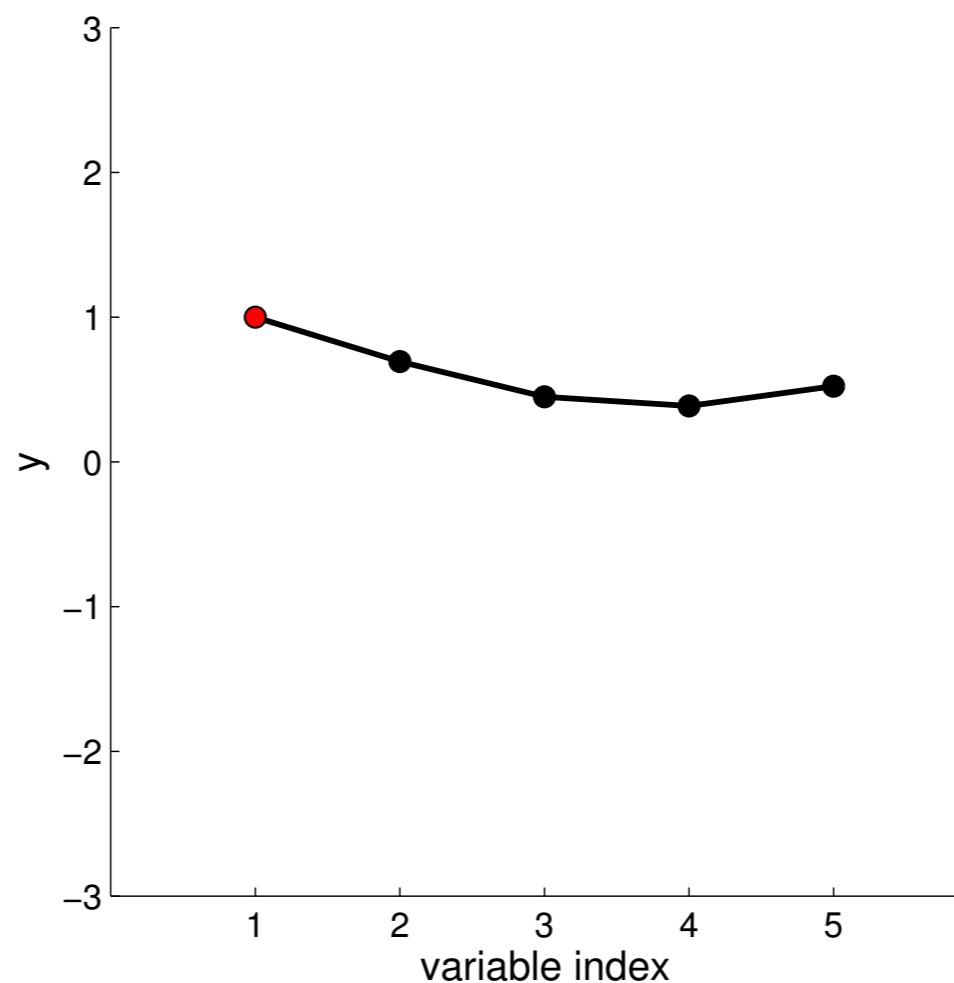
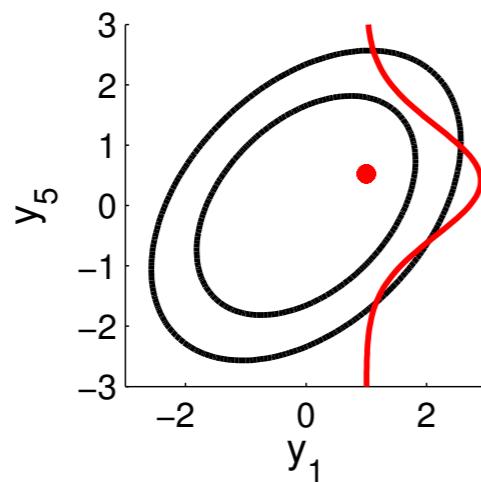
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



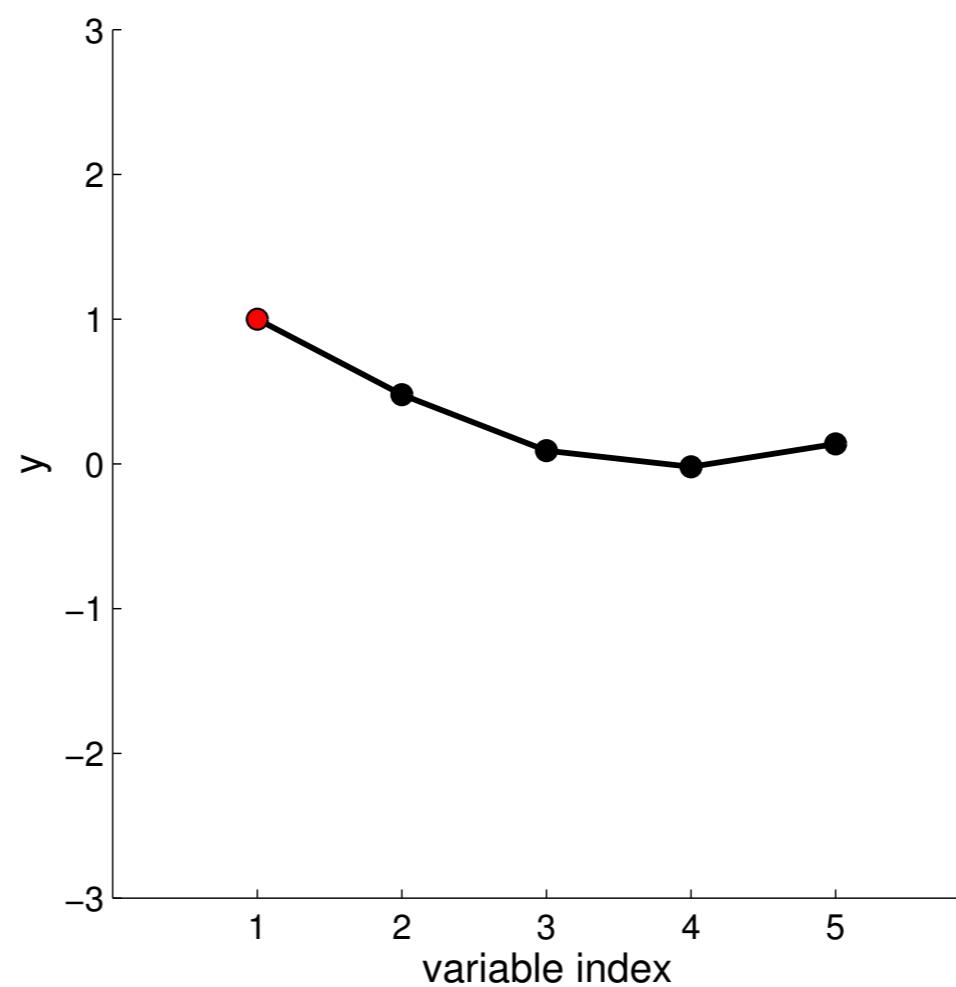
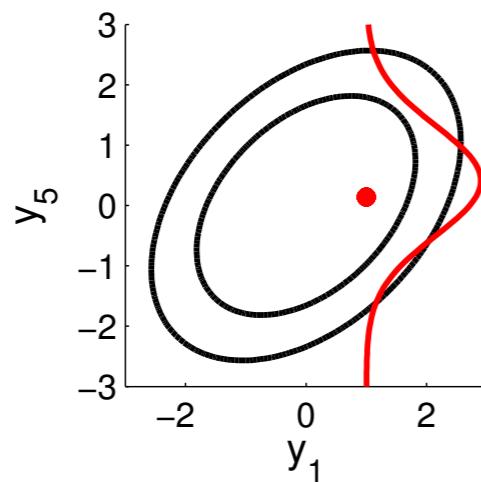
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



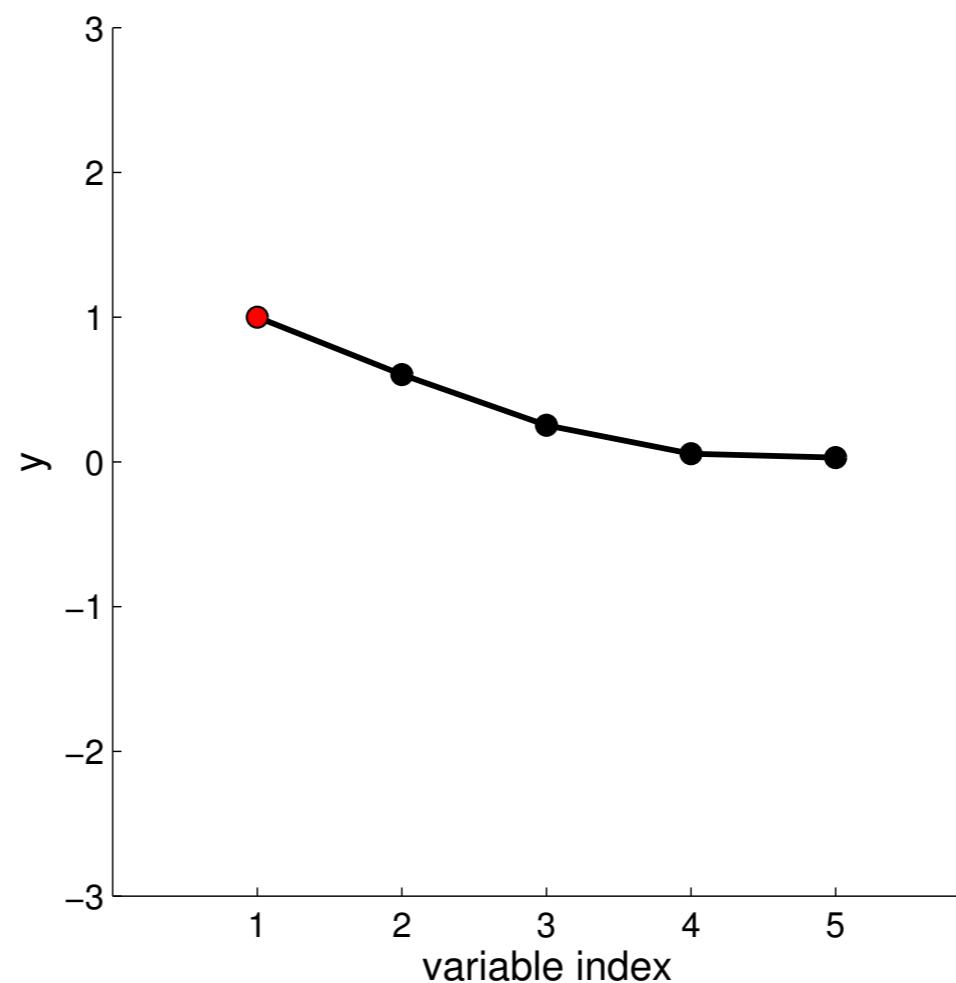
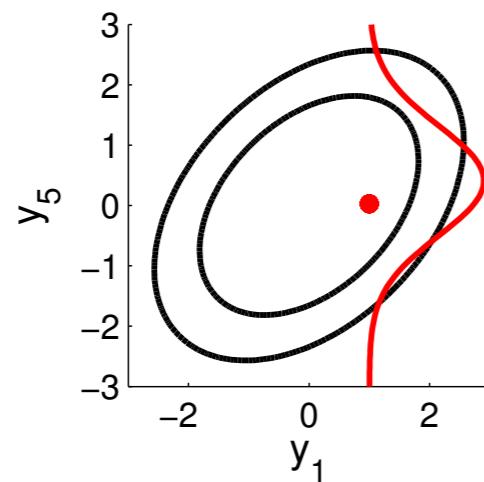
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



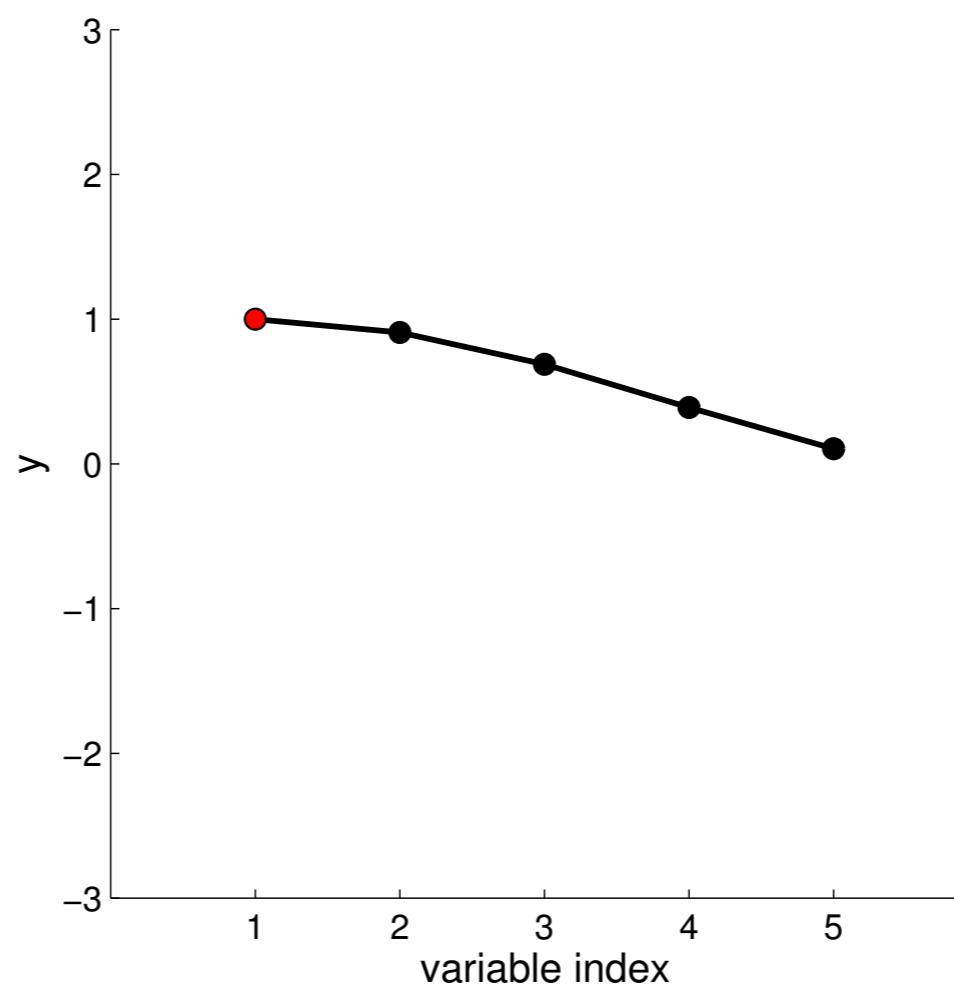
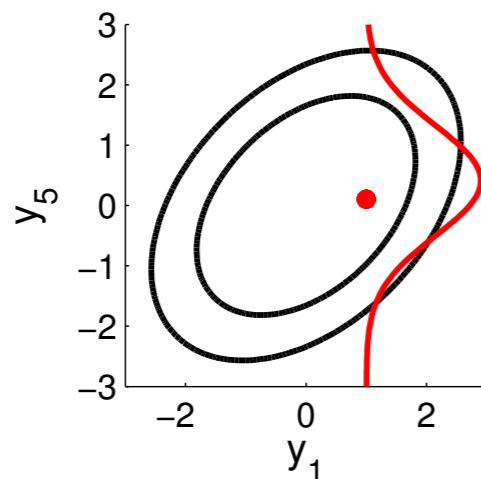
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



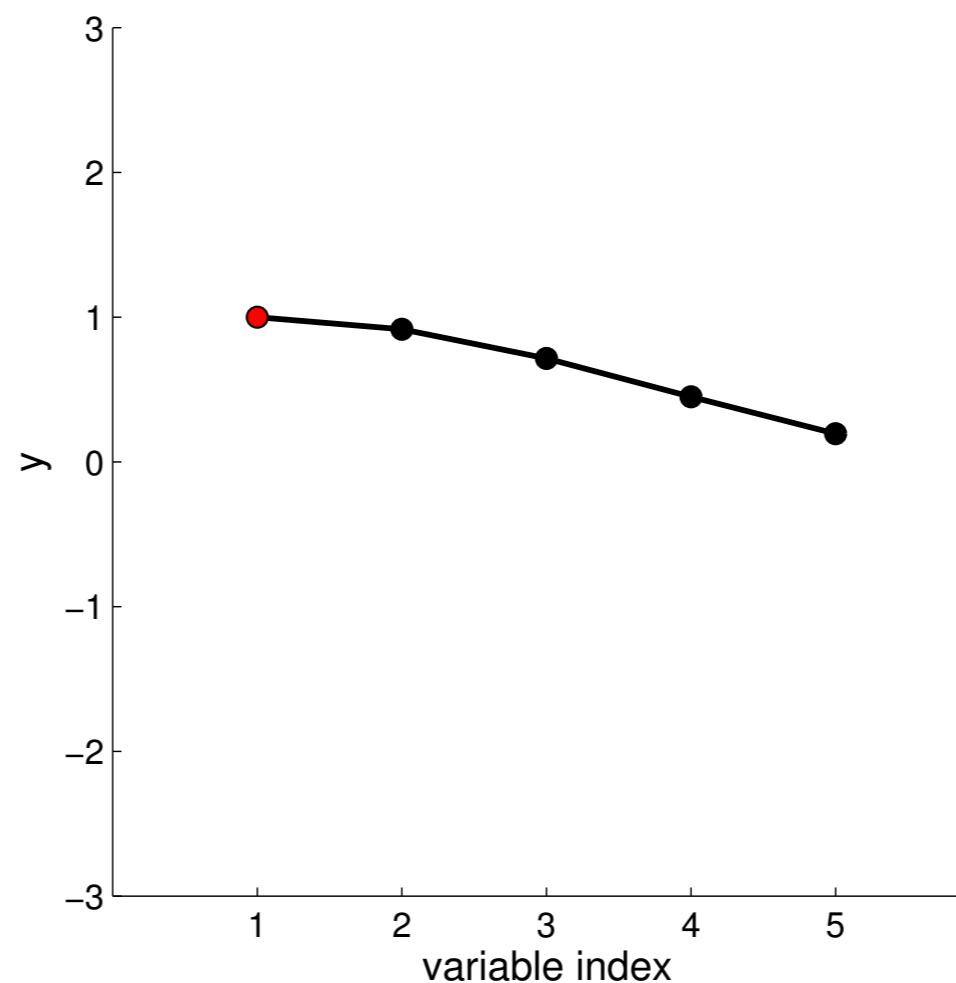
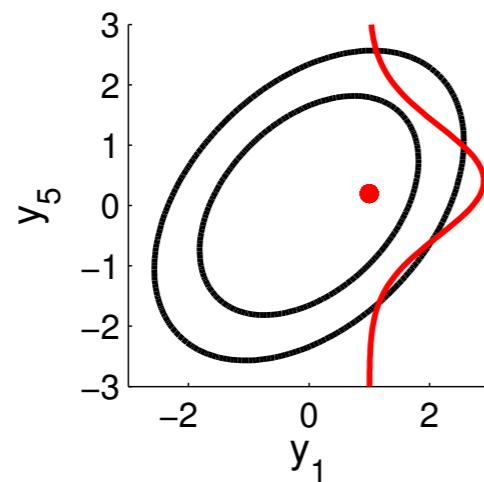
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



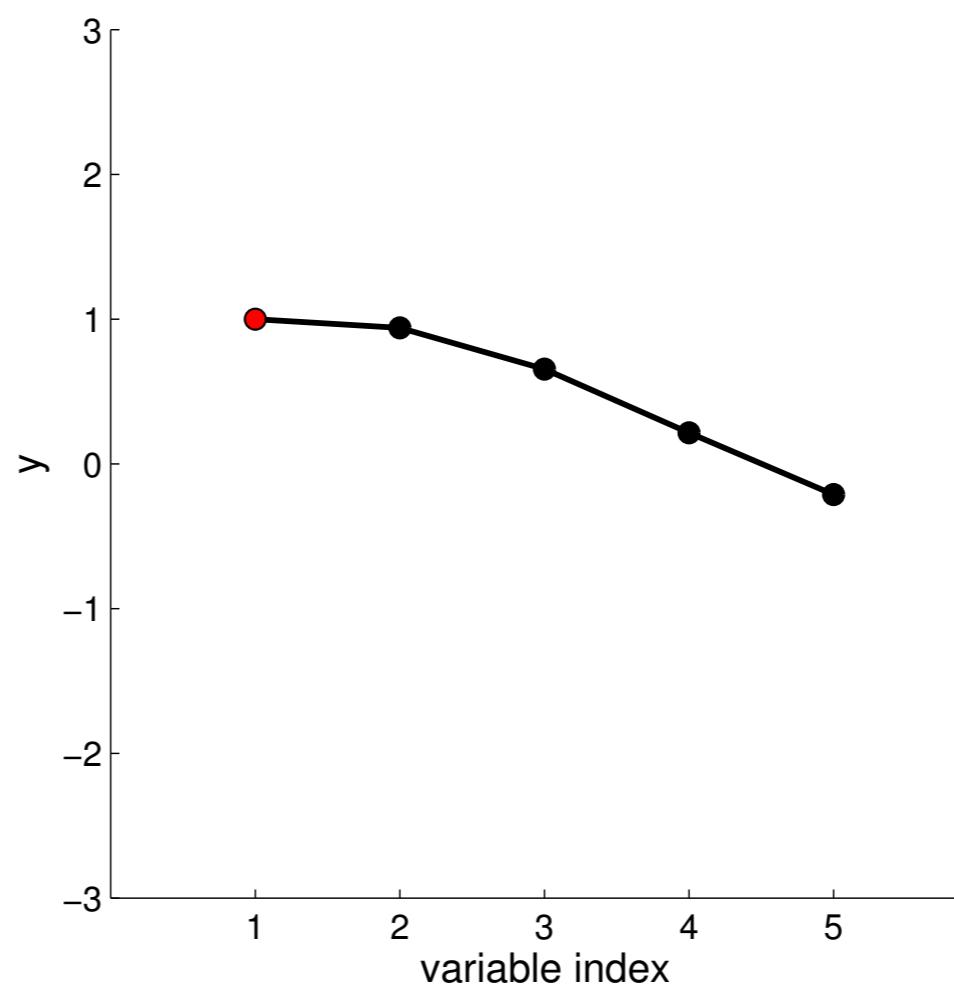
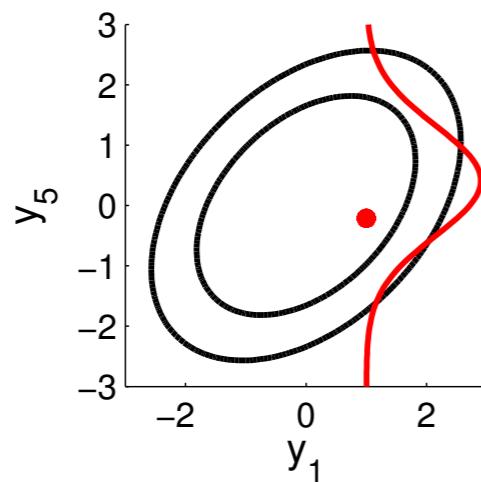
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



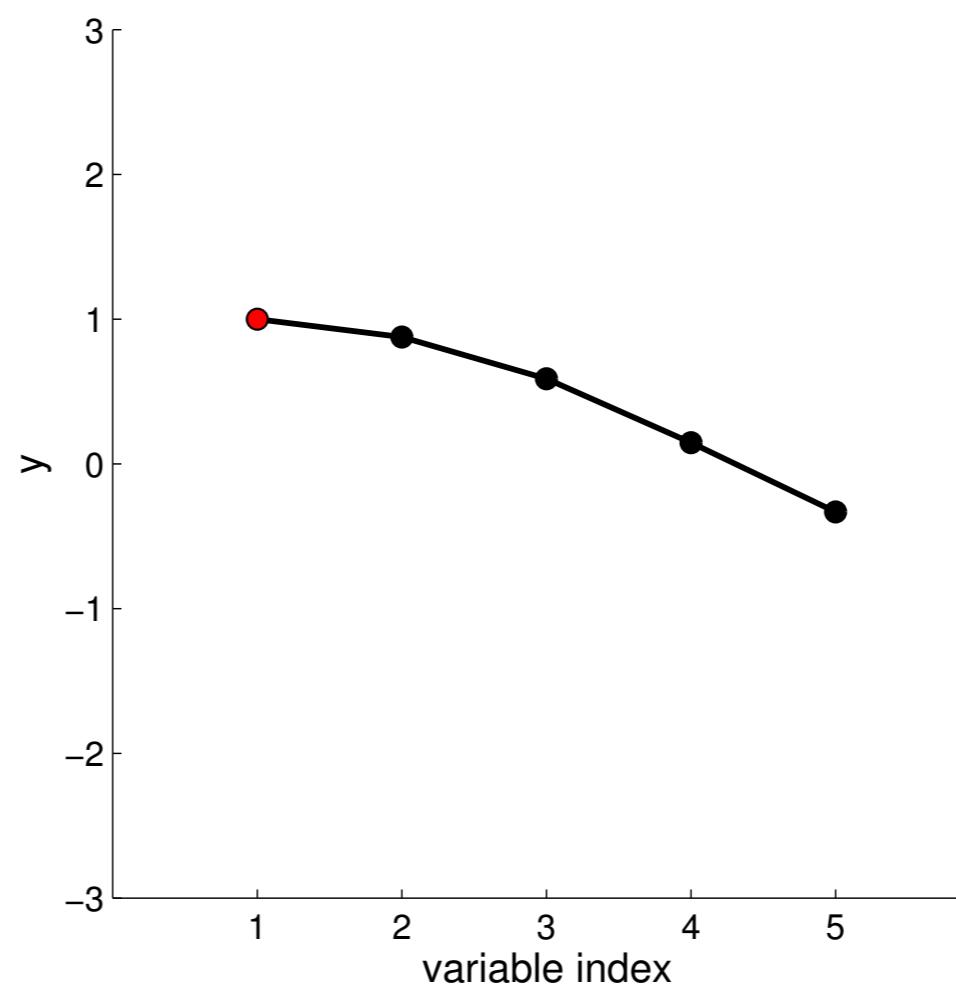
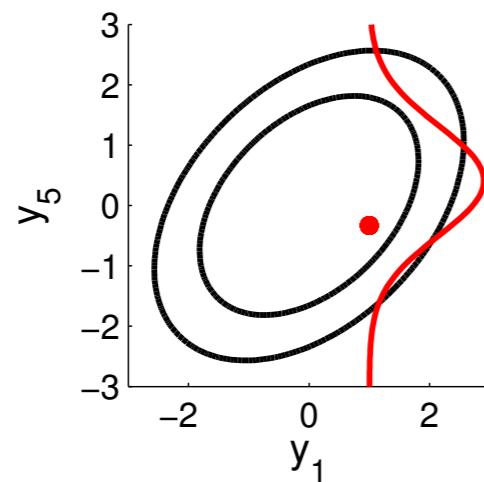
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



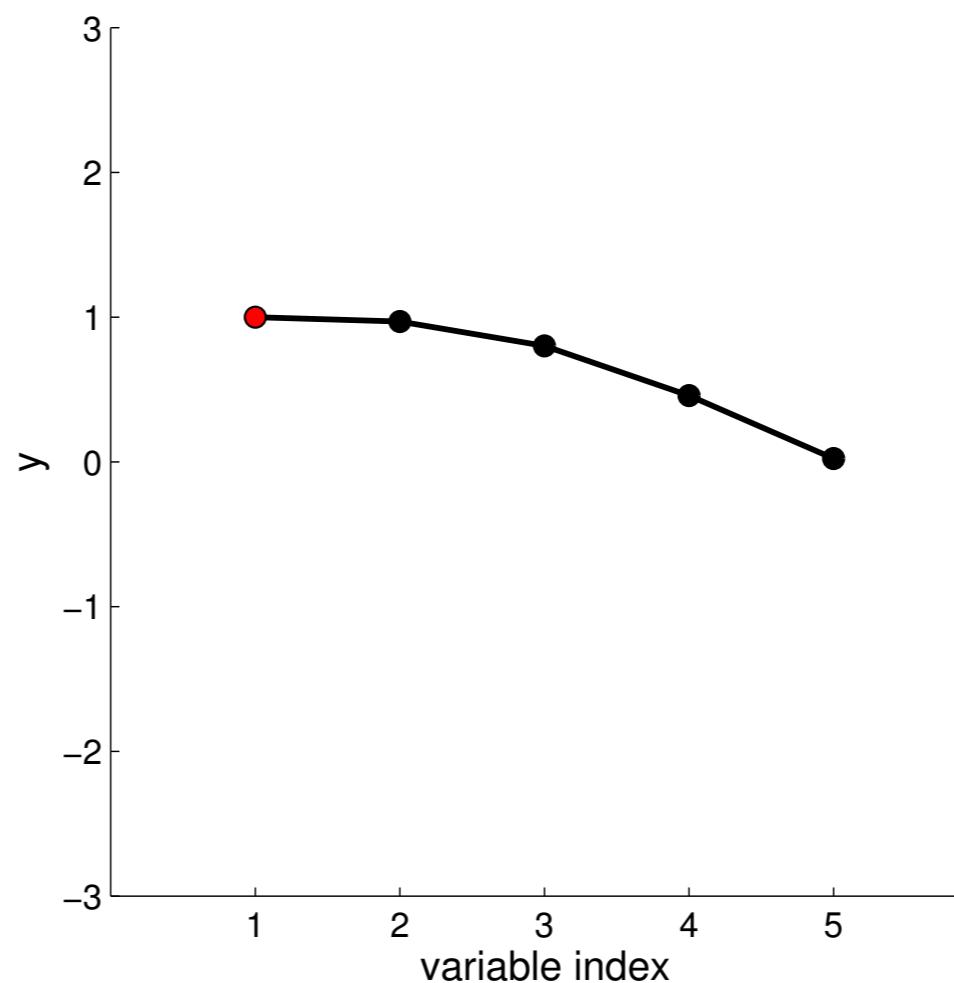
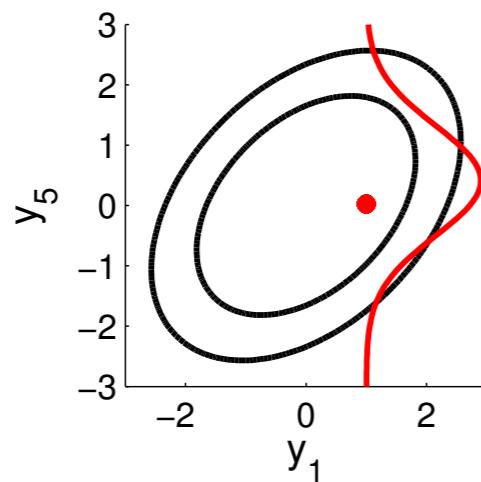
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



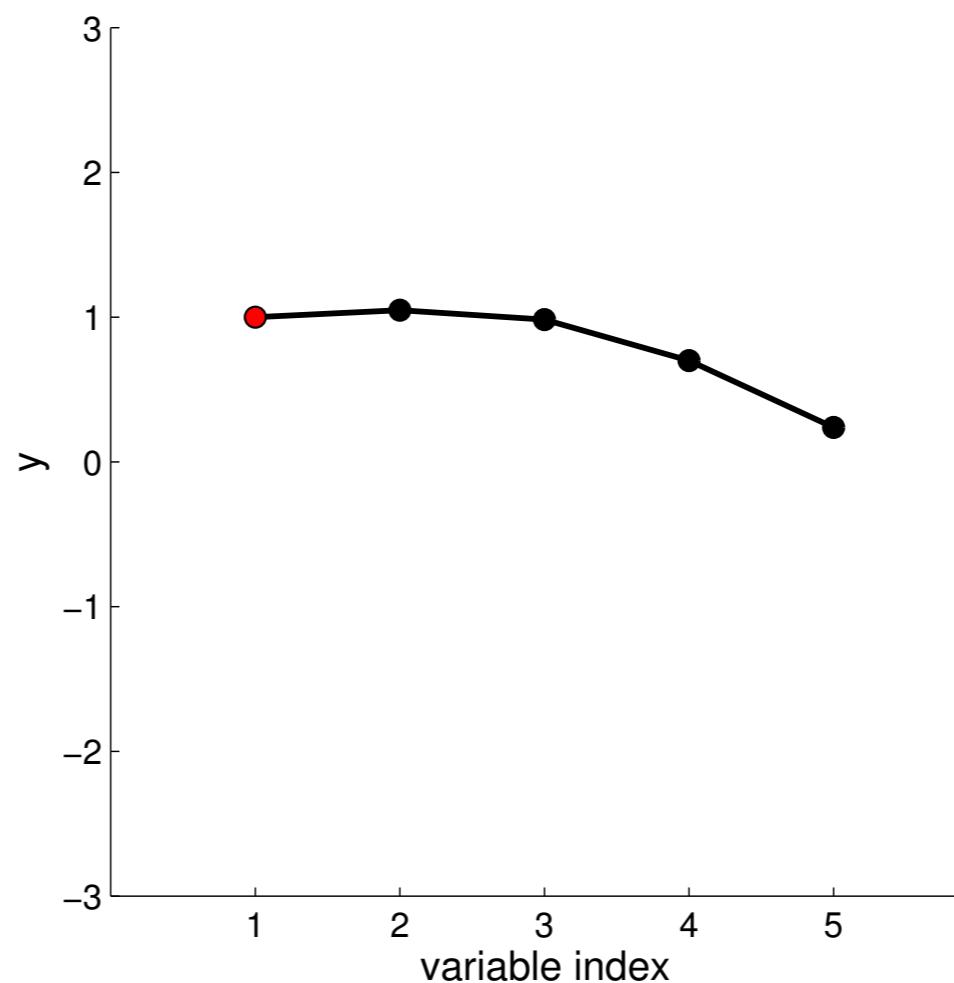
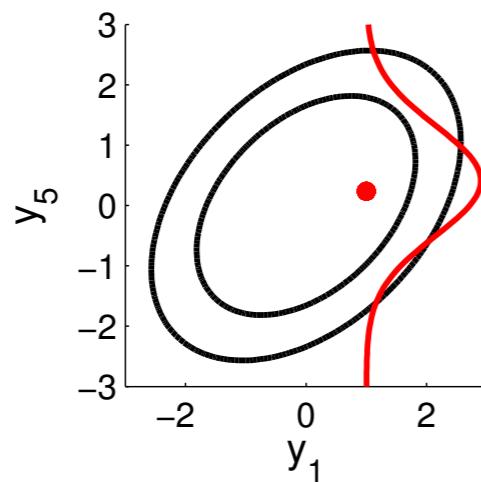
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



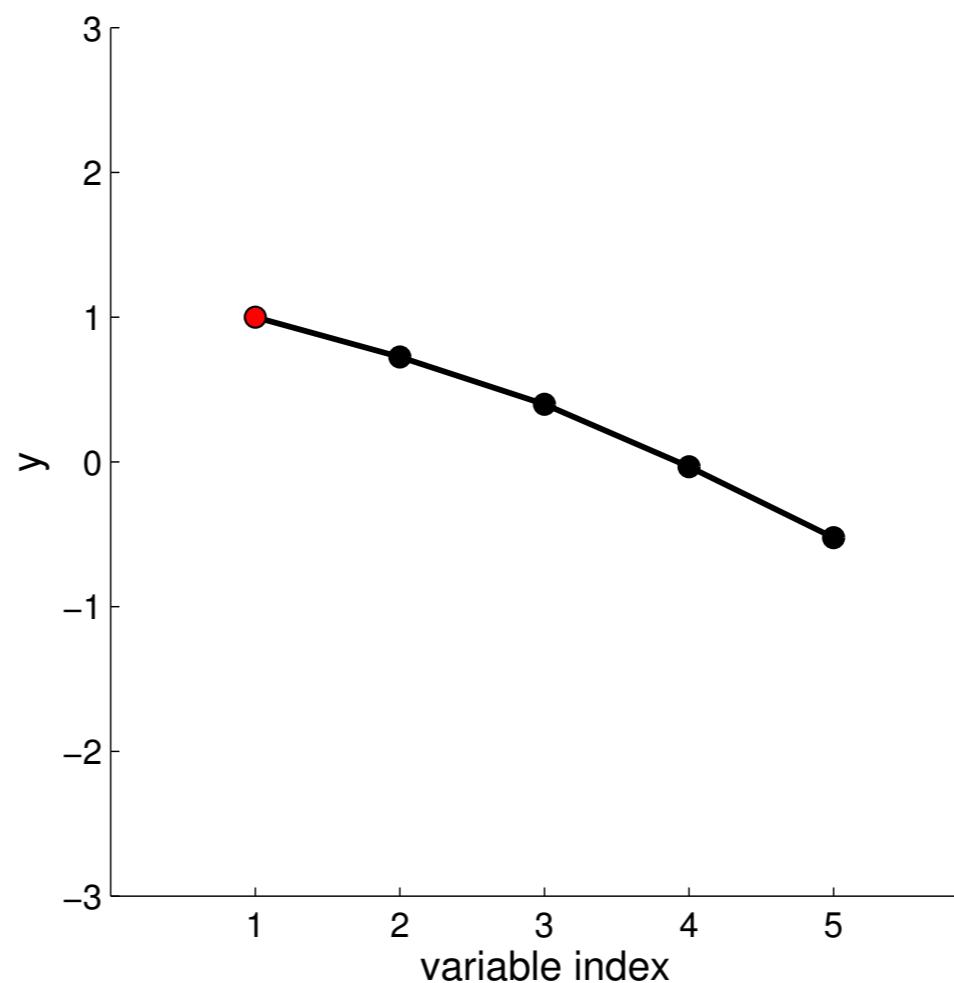
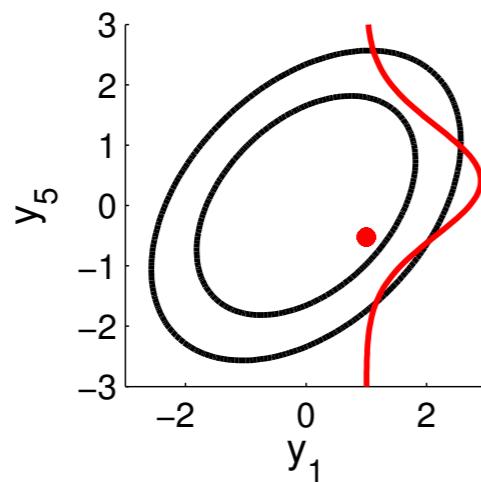
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



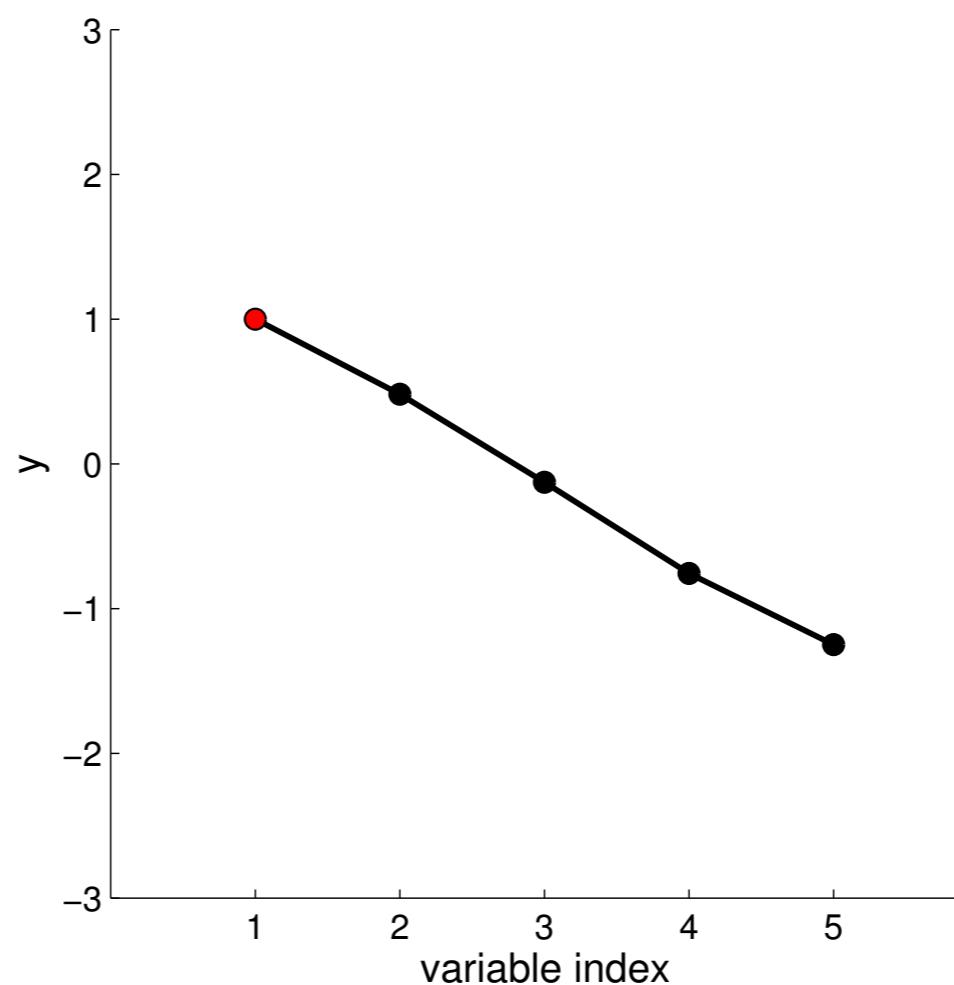
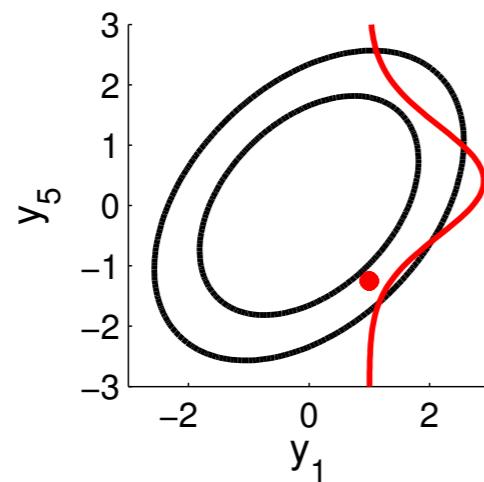
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



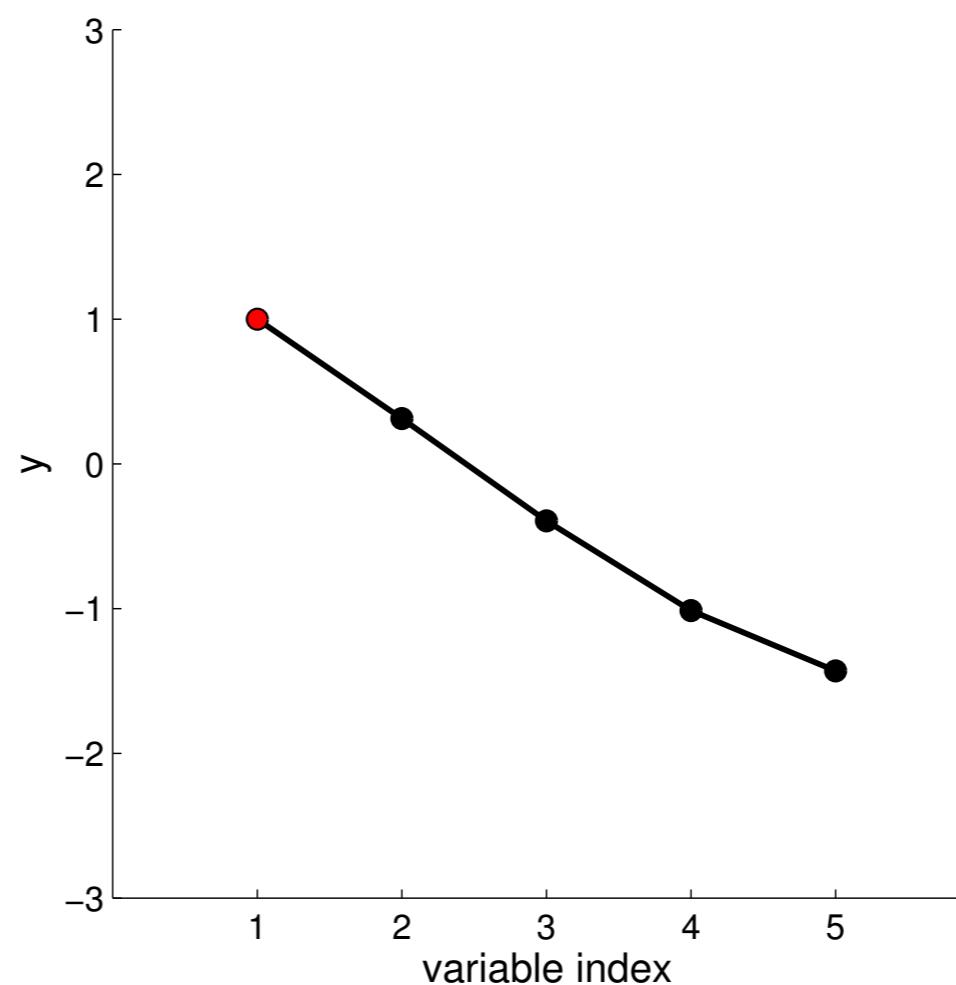
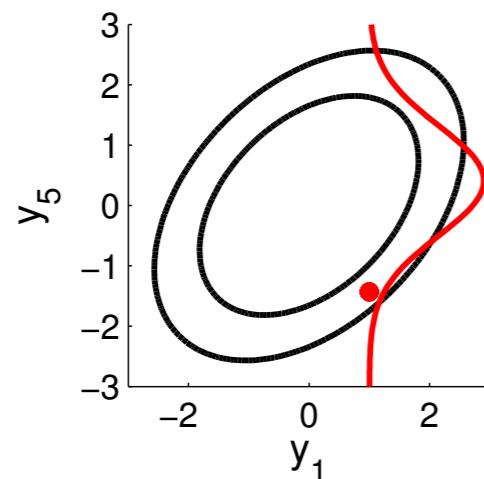
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



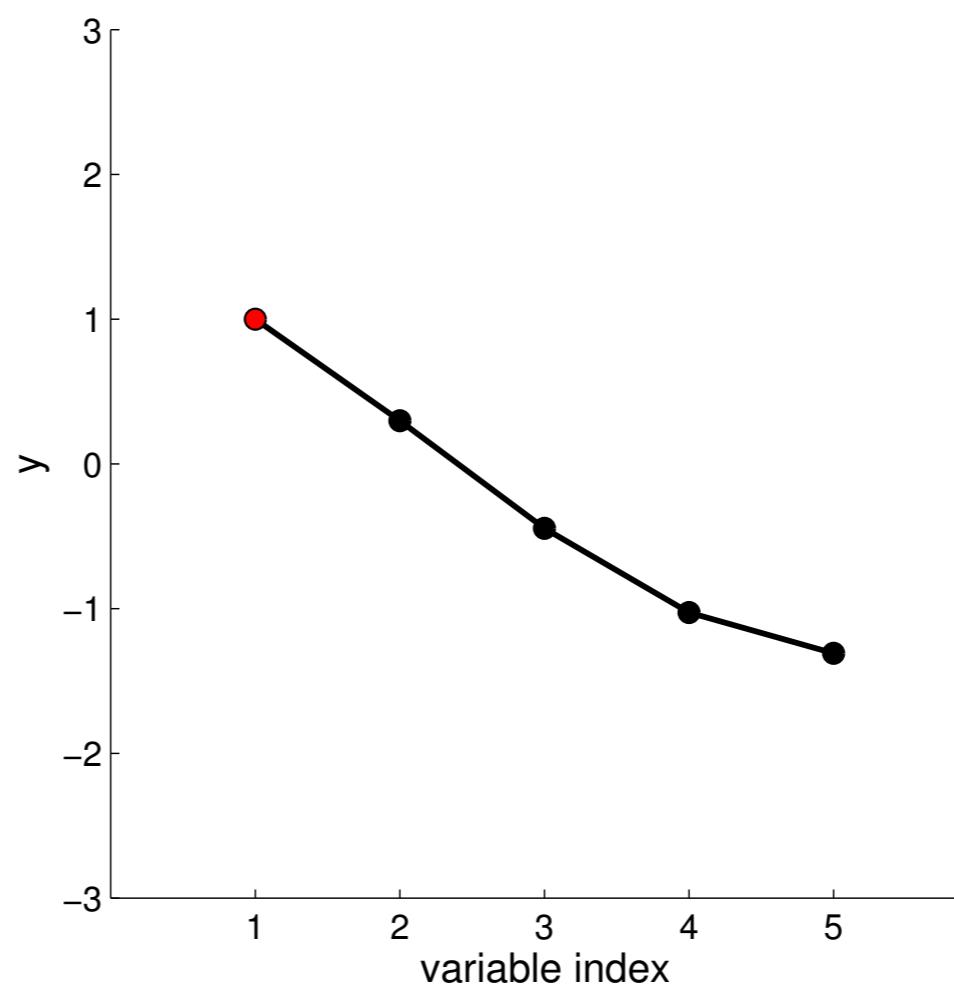
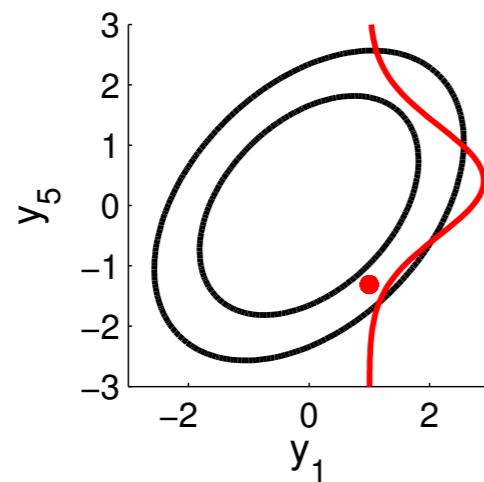
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



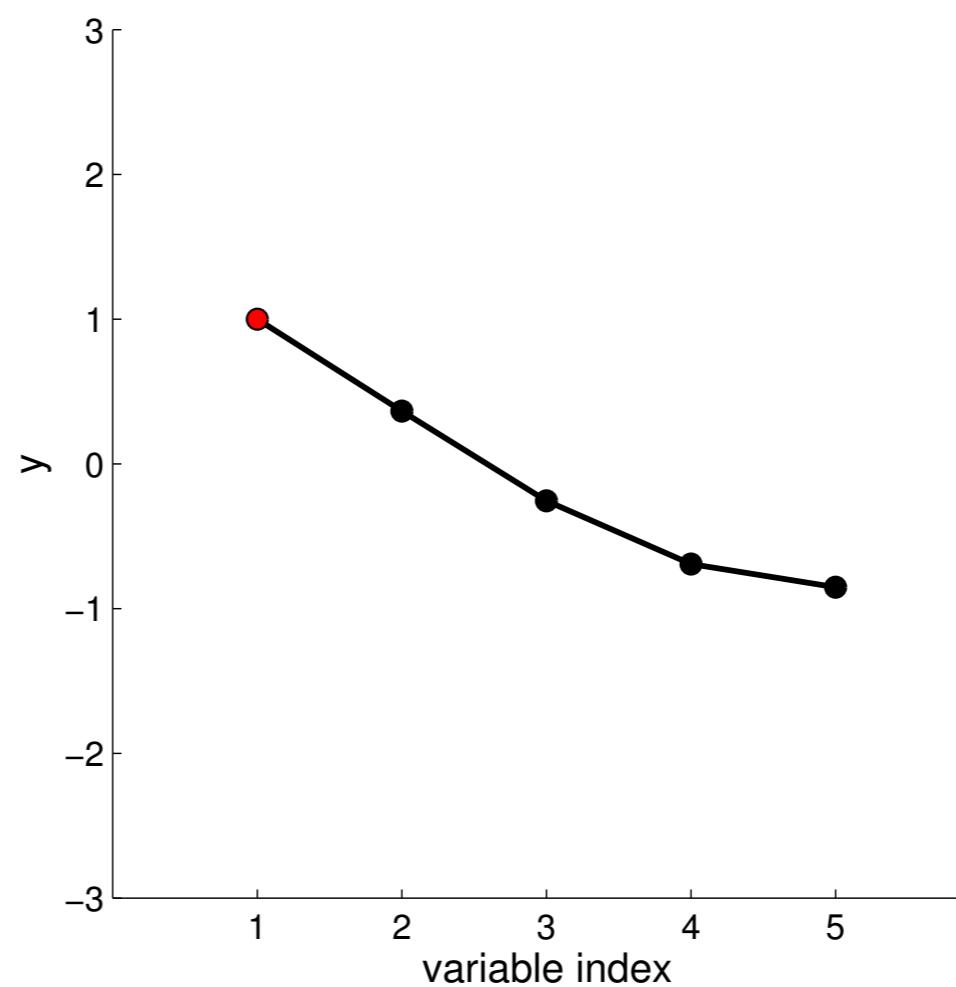
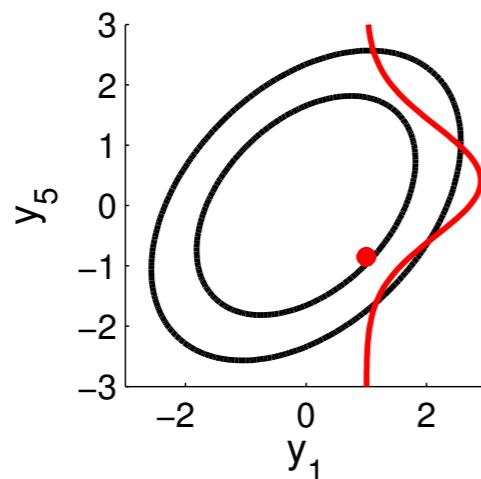
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



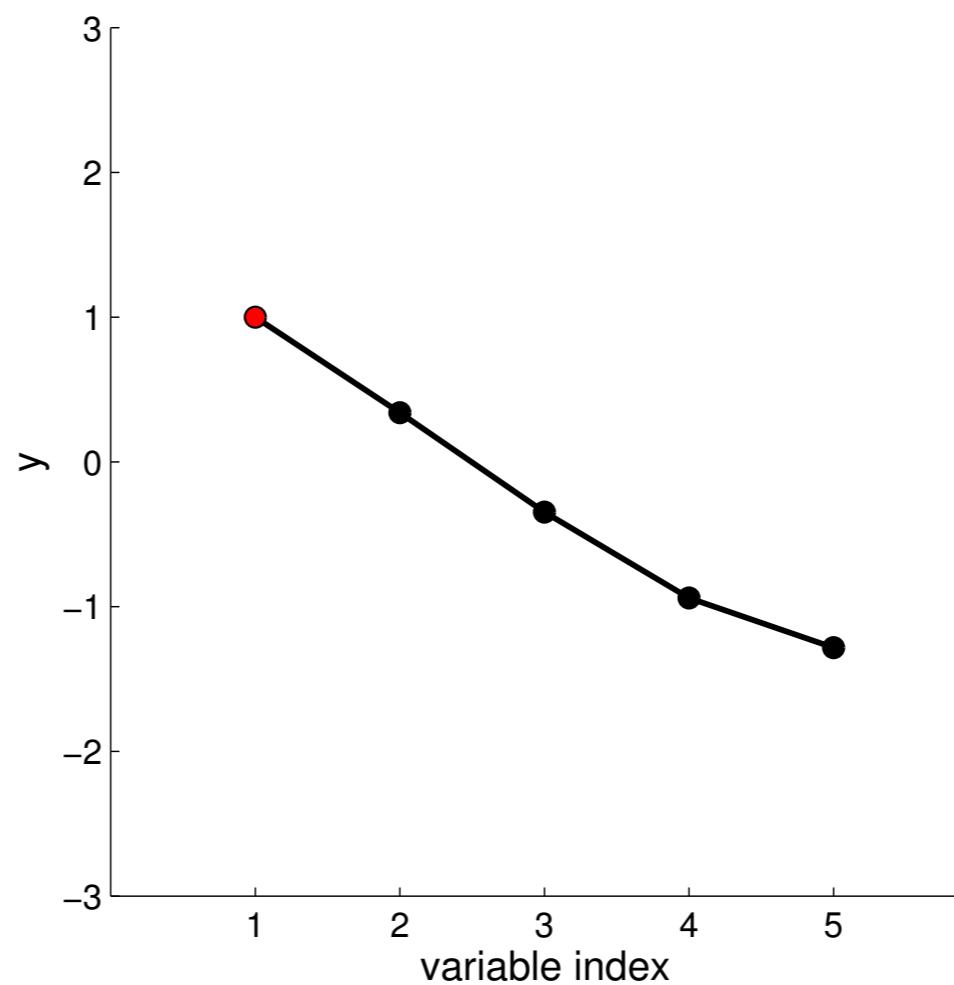
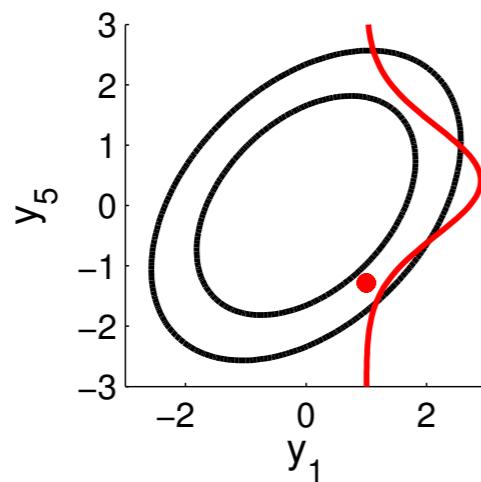
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



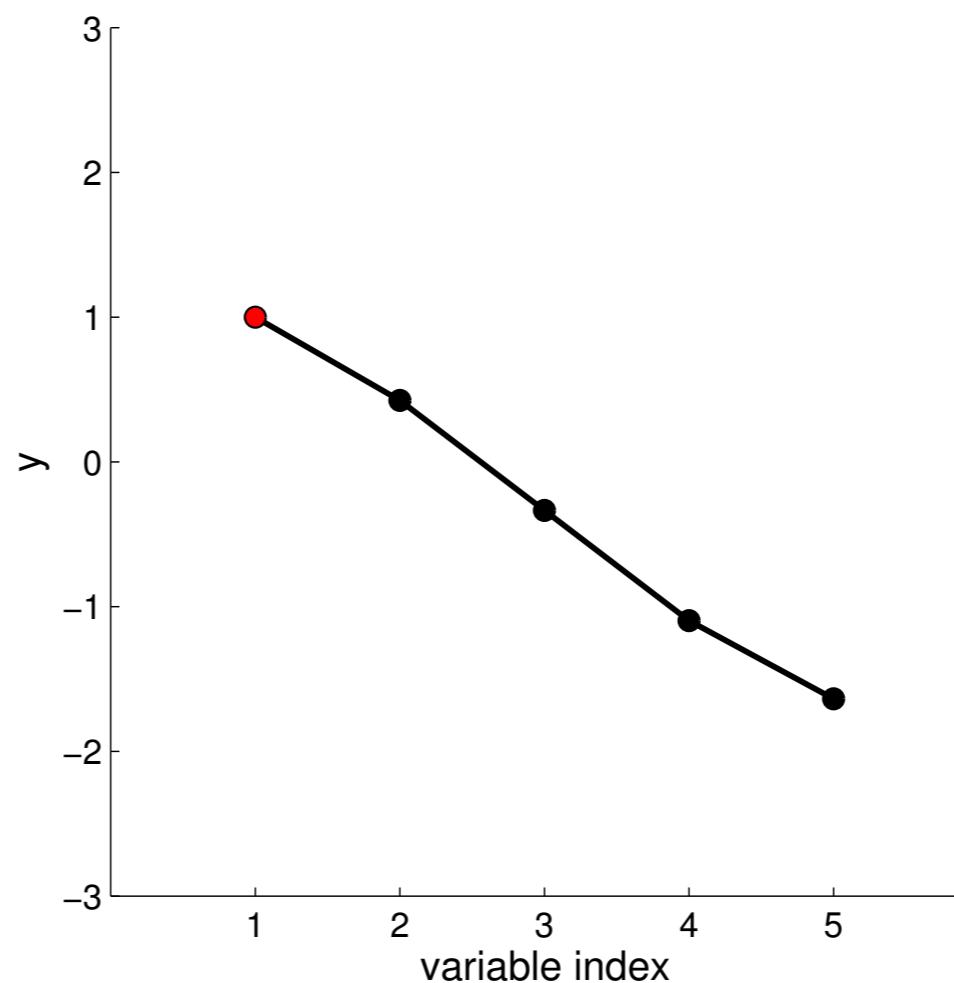
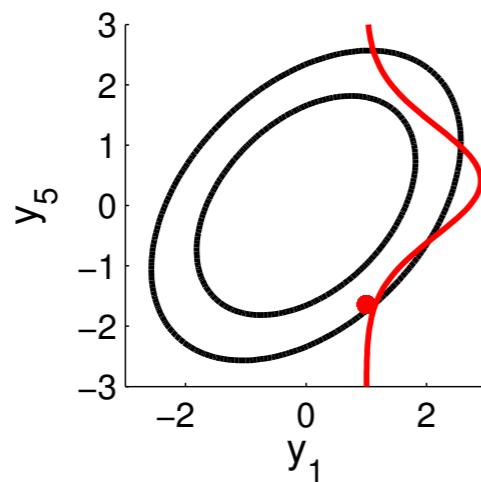
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



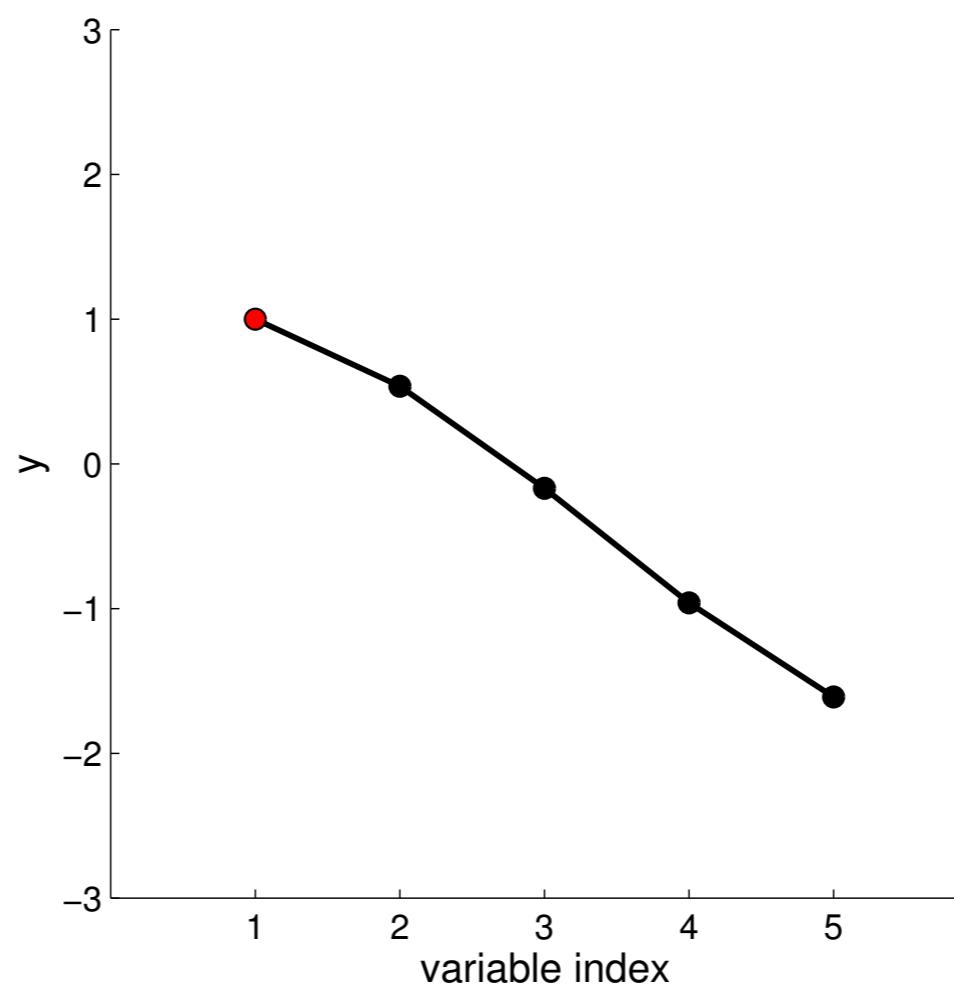
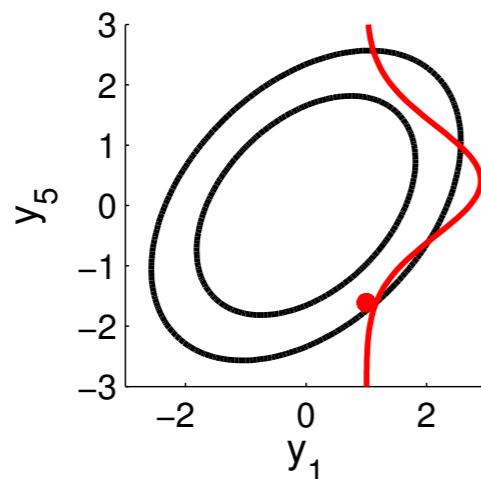
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



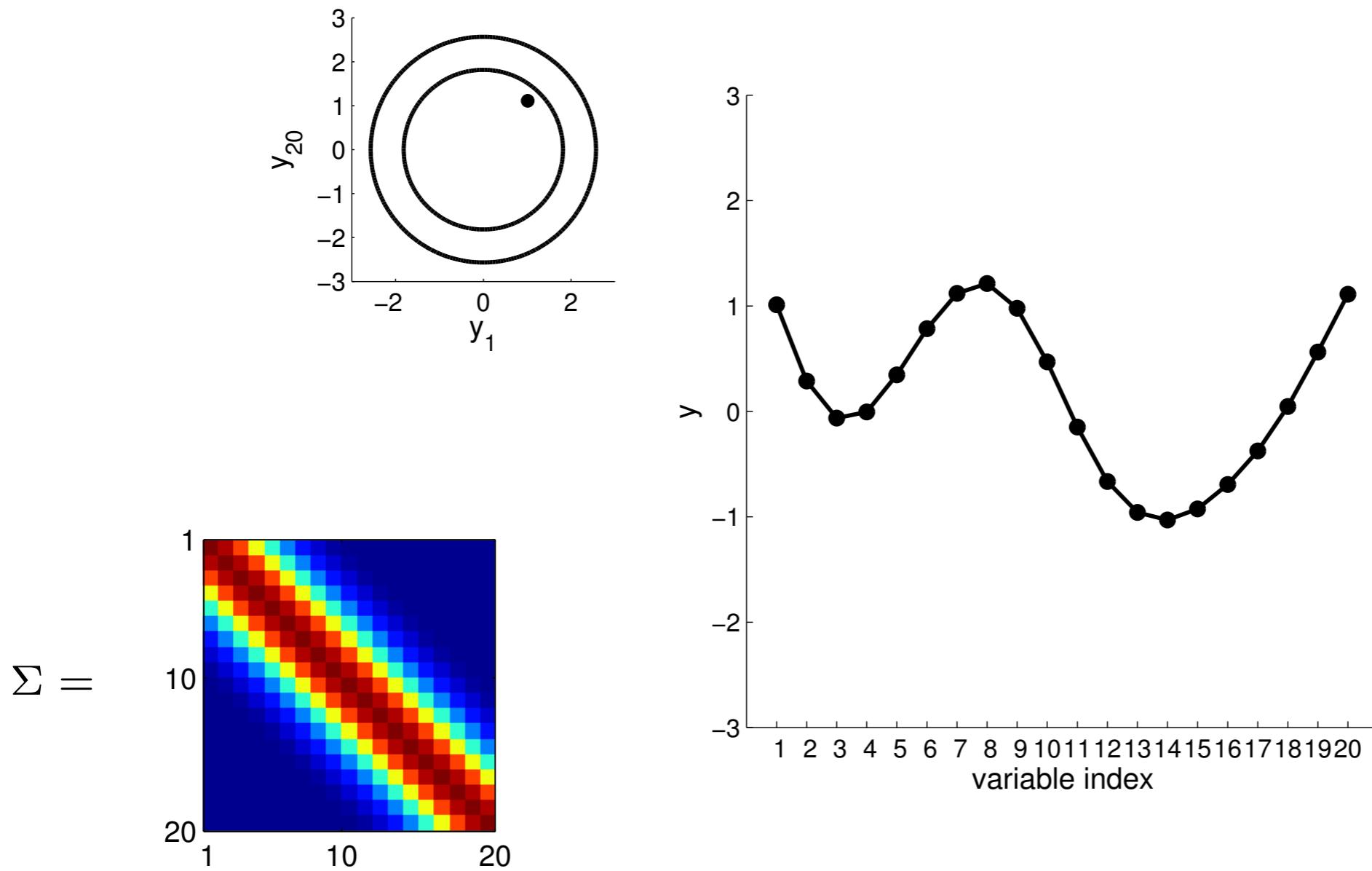
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix - conditioning



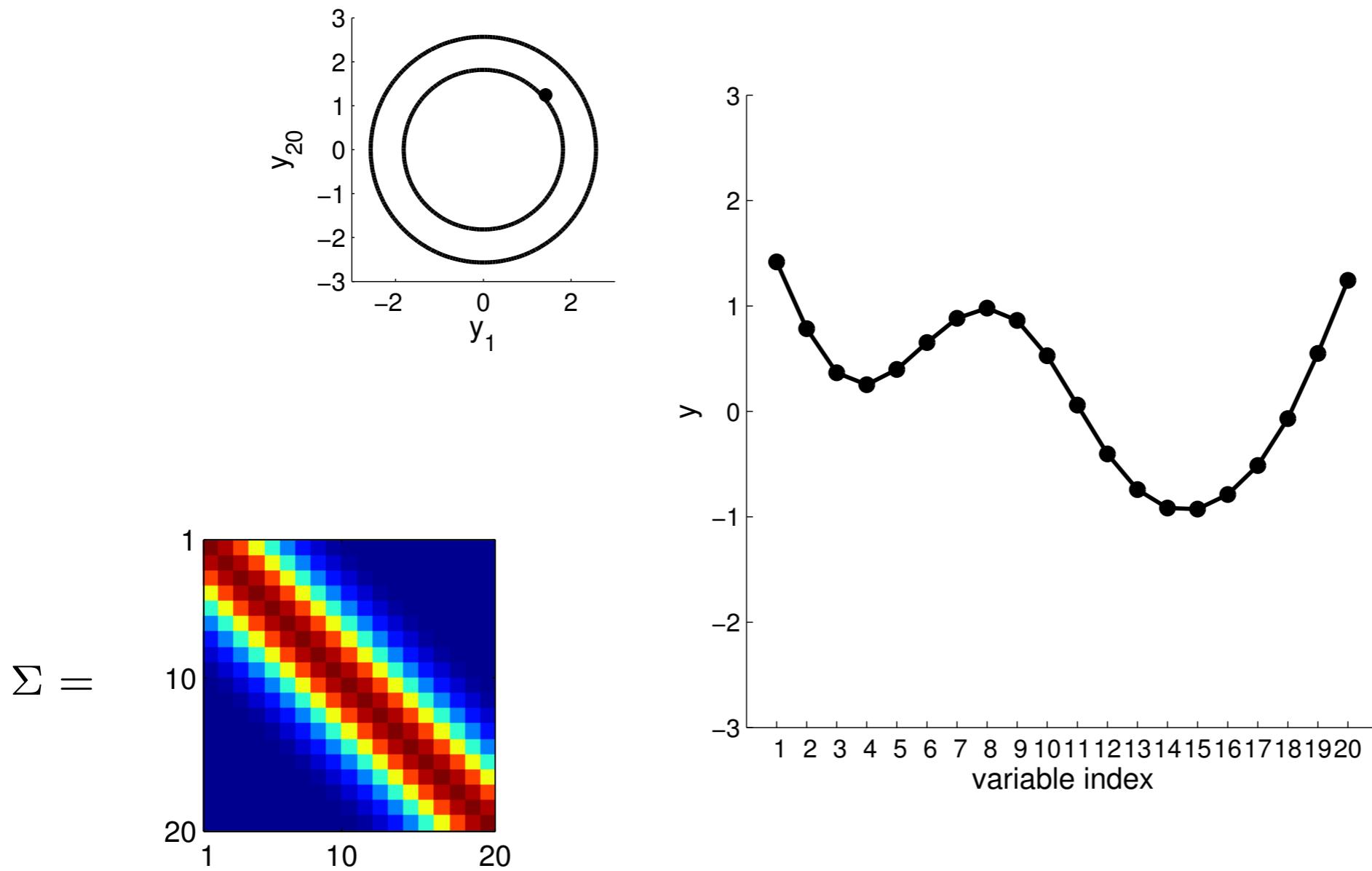
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

# Special covariance matrix

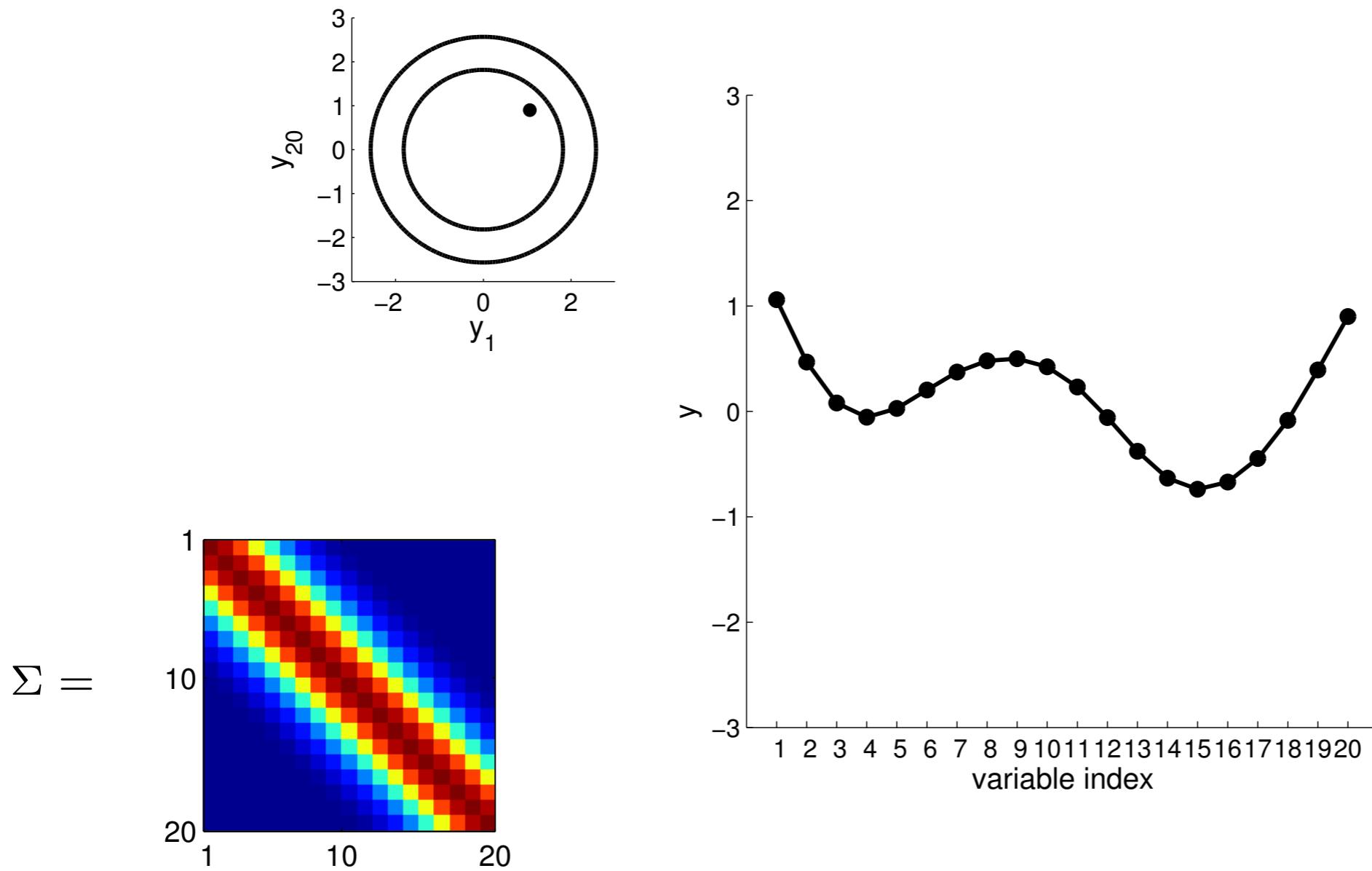


red means 1, blue means 0

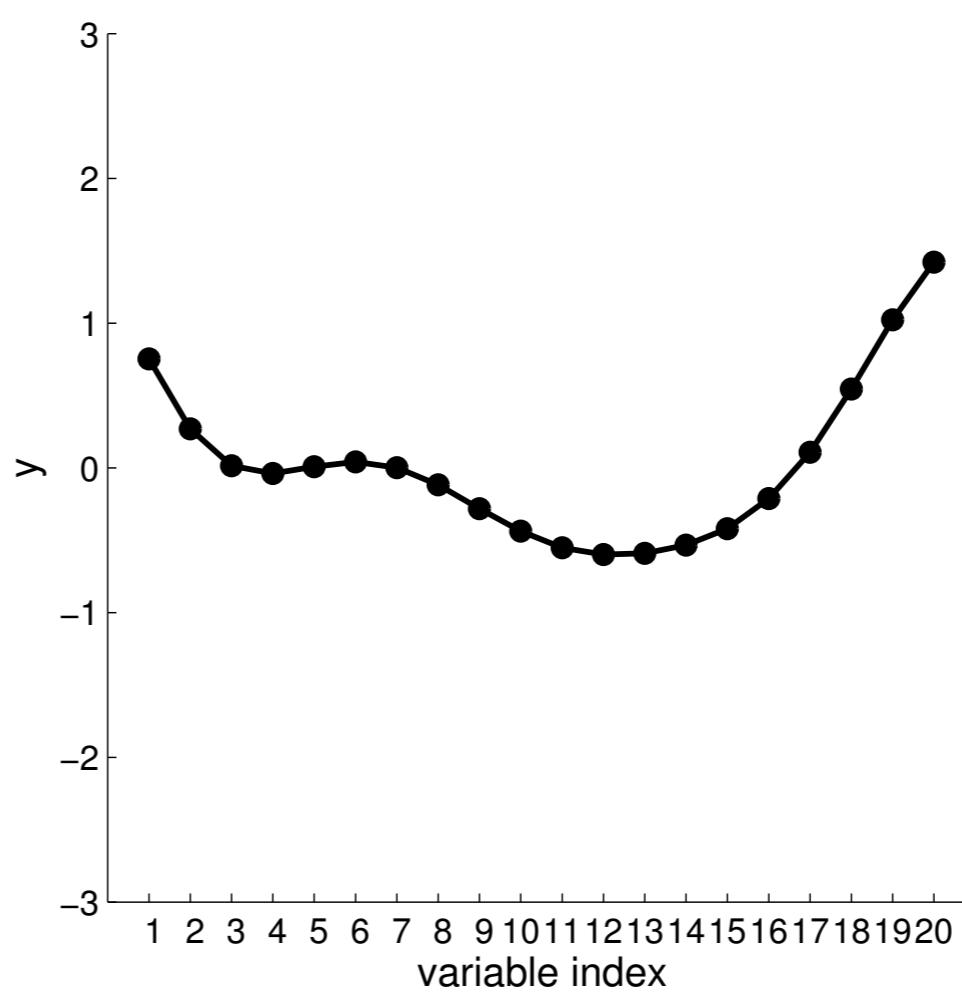
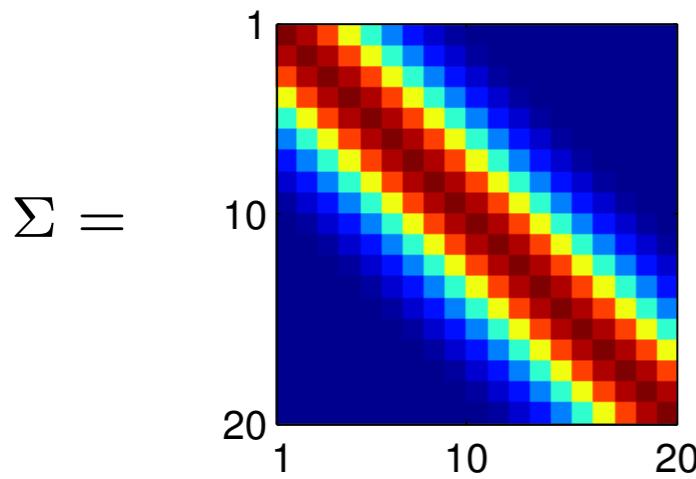
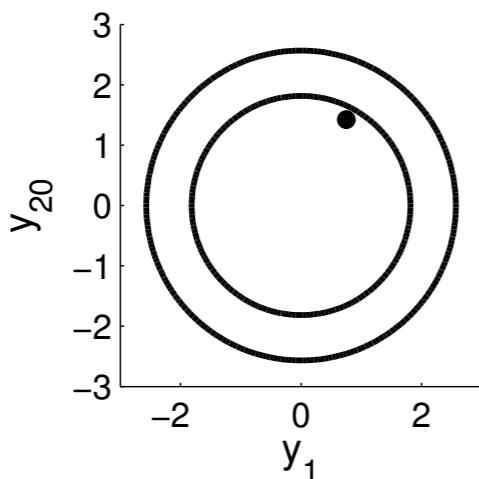
# Special covariance matrix



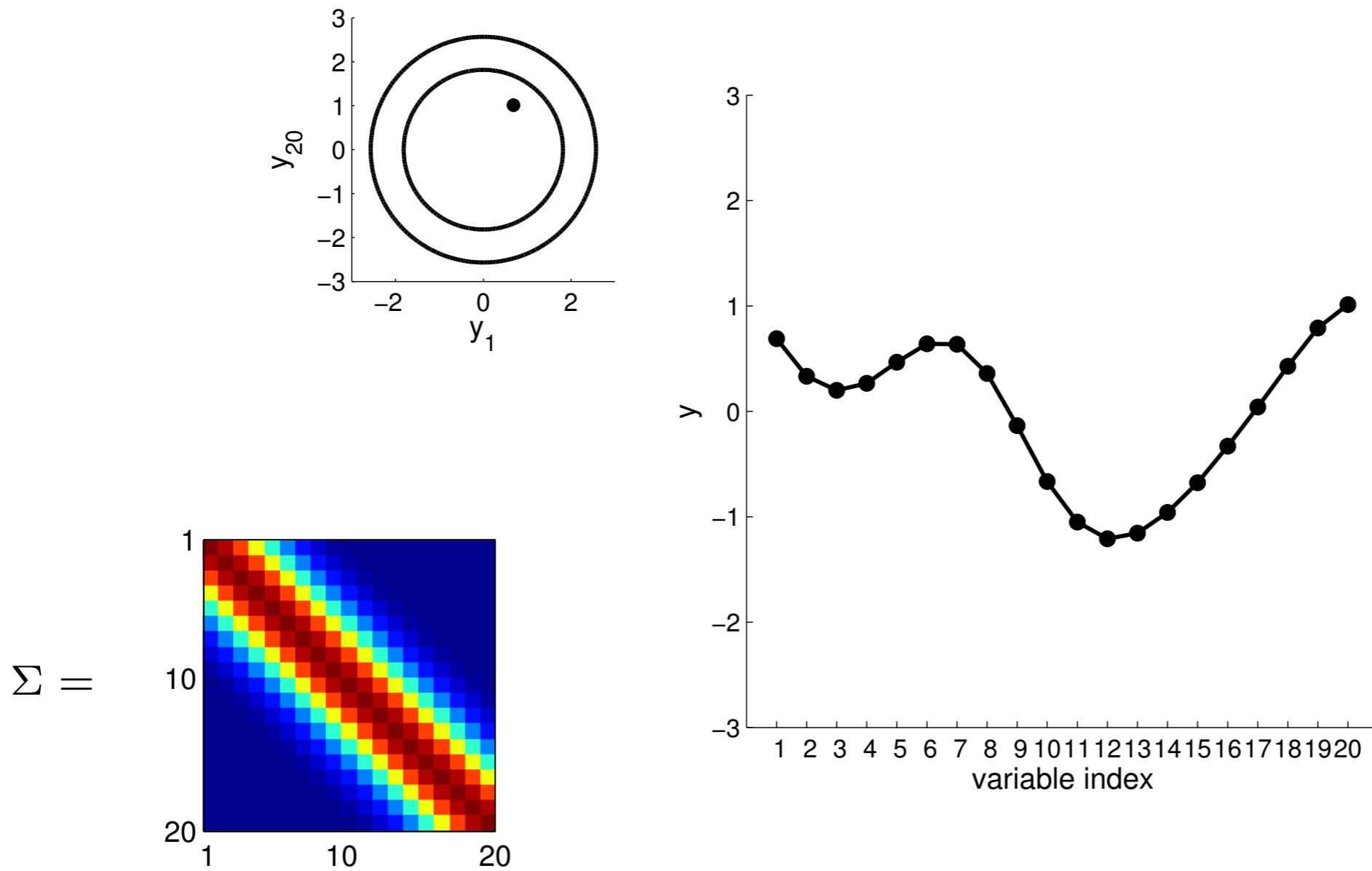
# Special covariance matrix



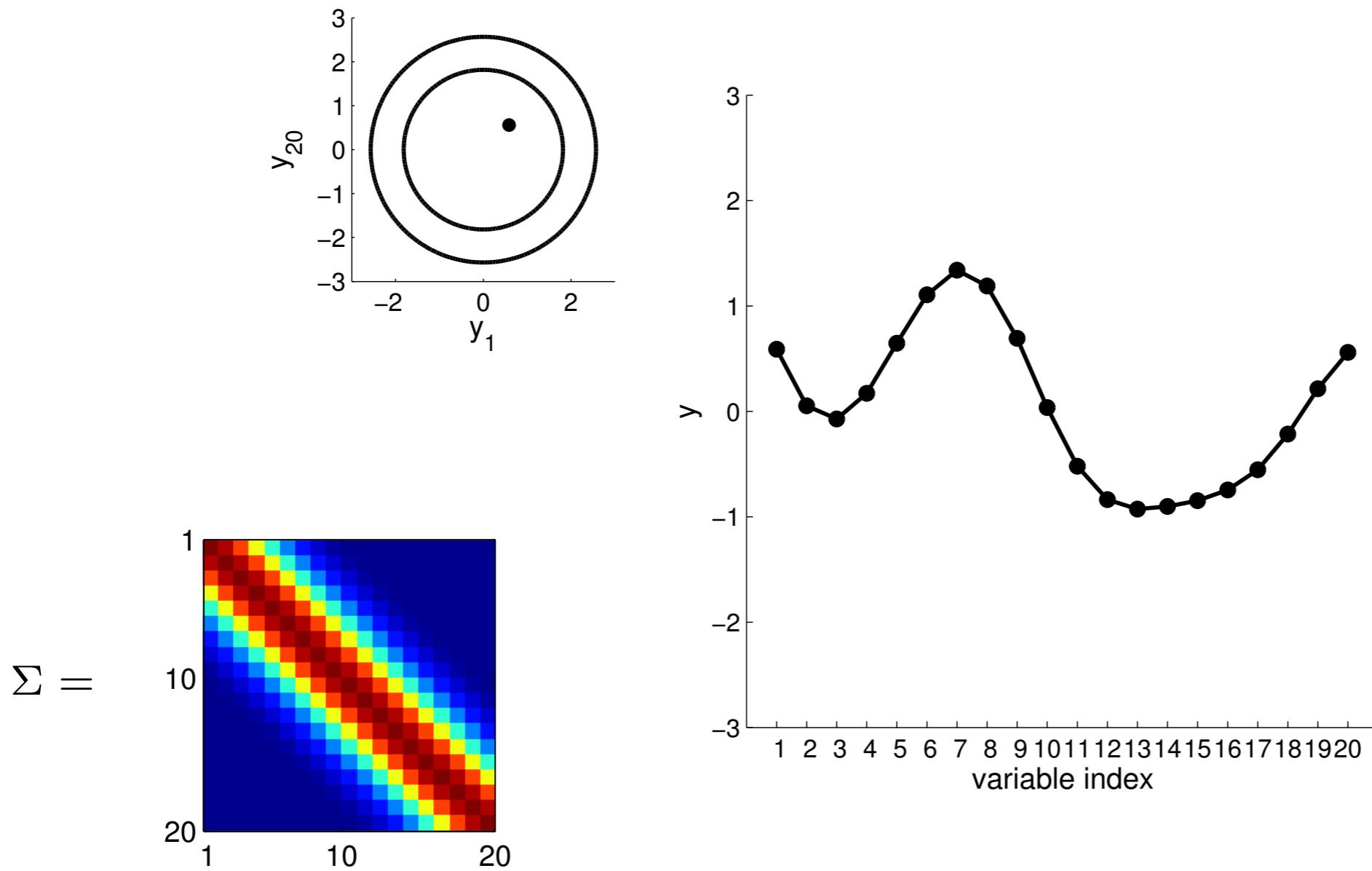
# Special covariance matrix



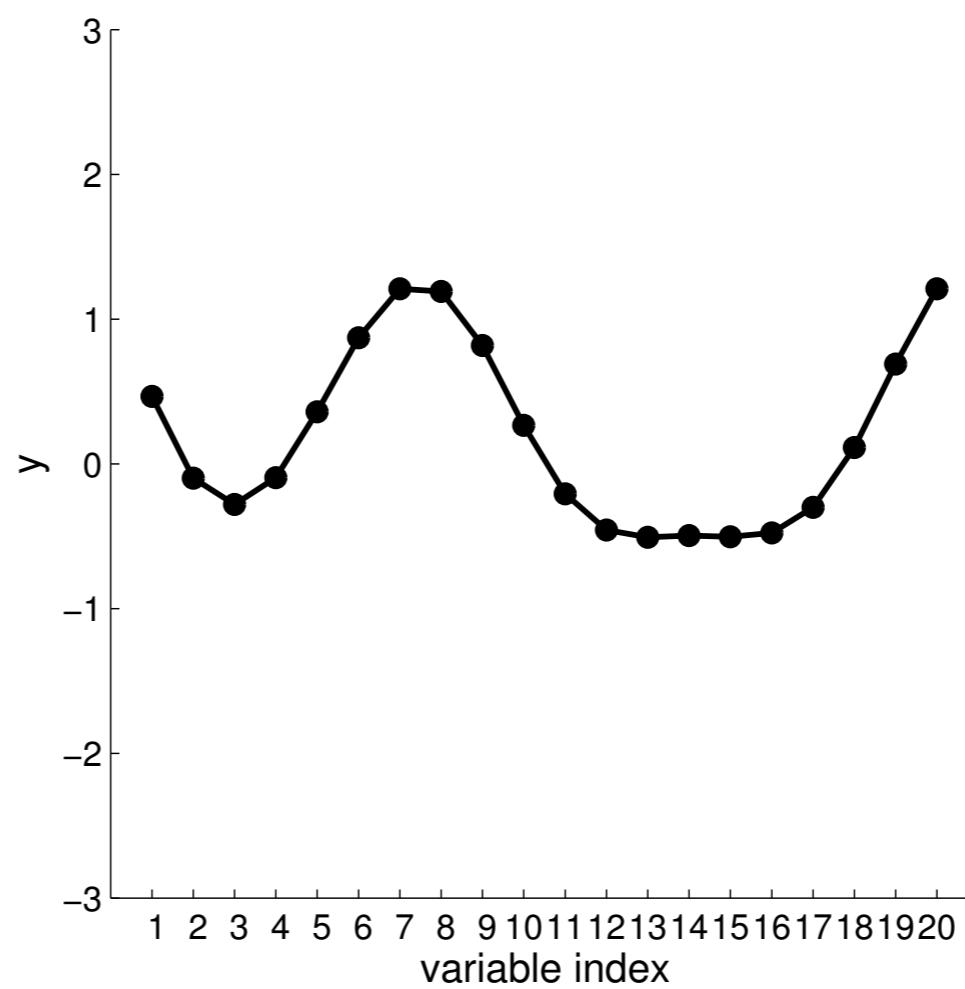
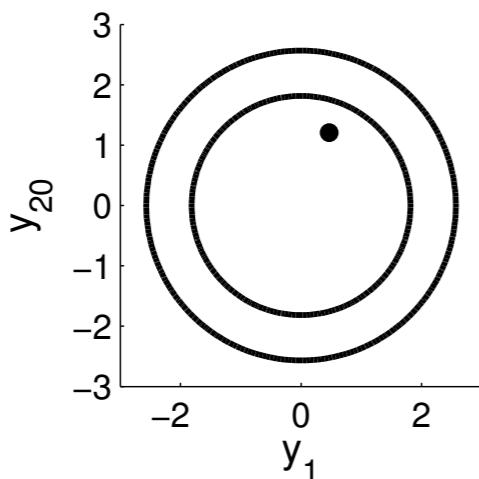
# Special covariance matrix



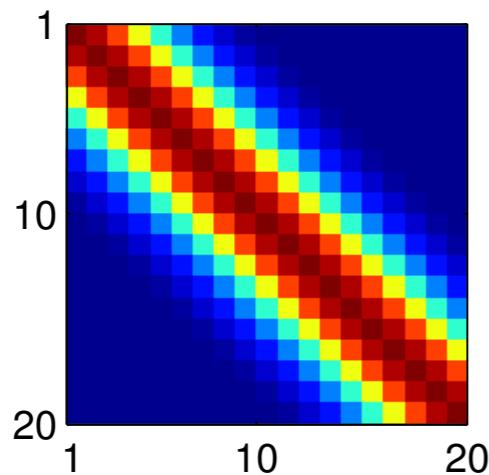
# Special covariance matrix



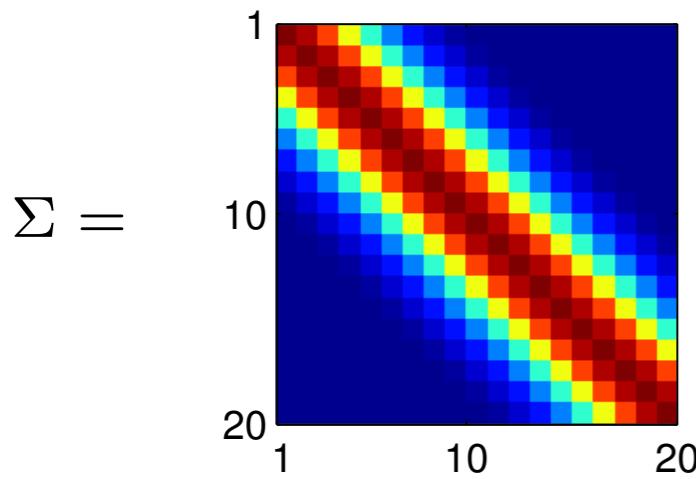
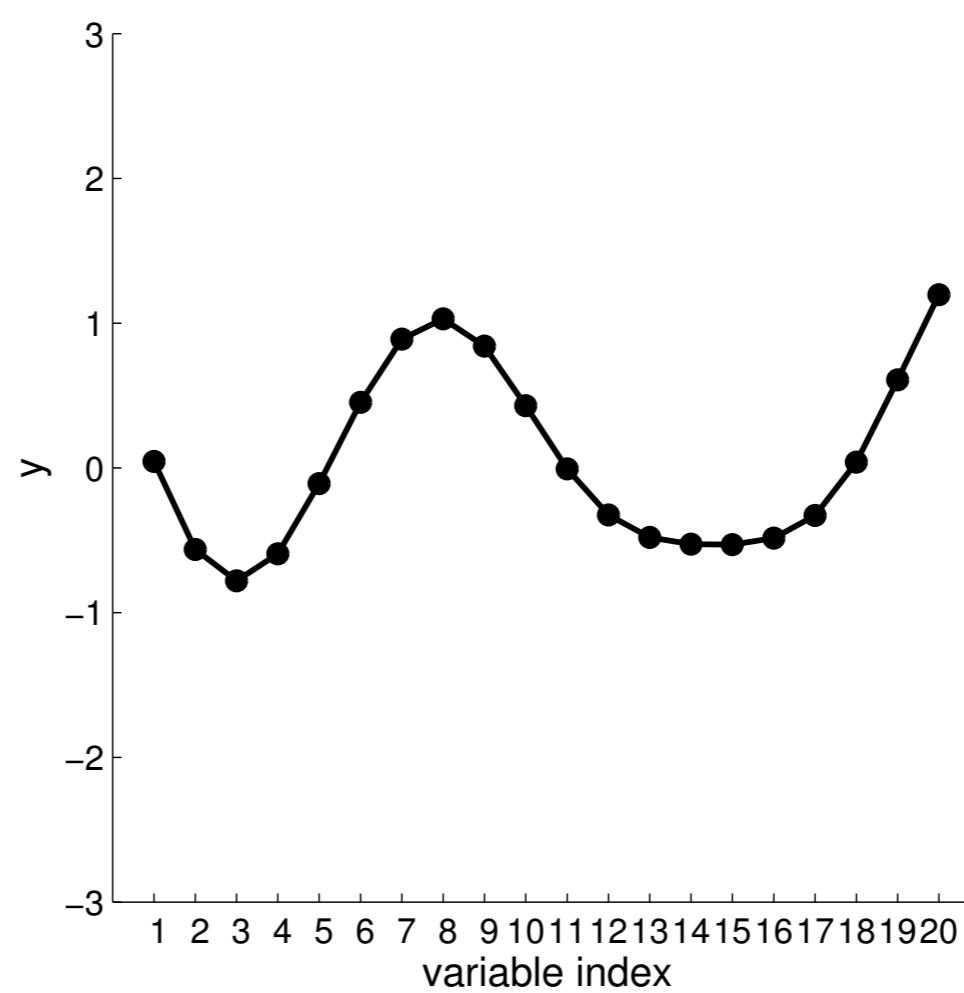
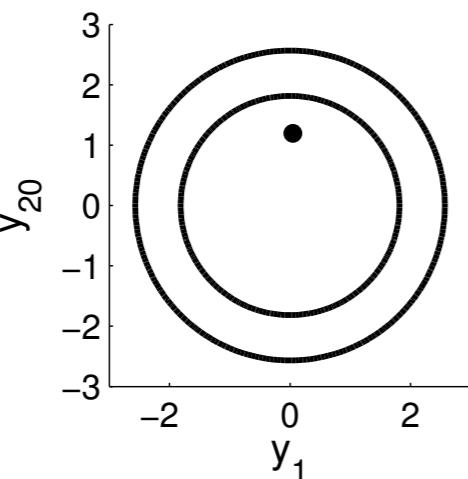
# Special covariance matrix



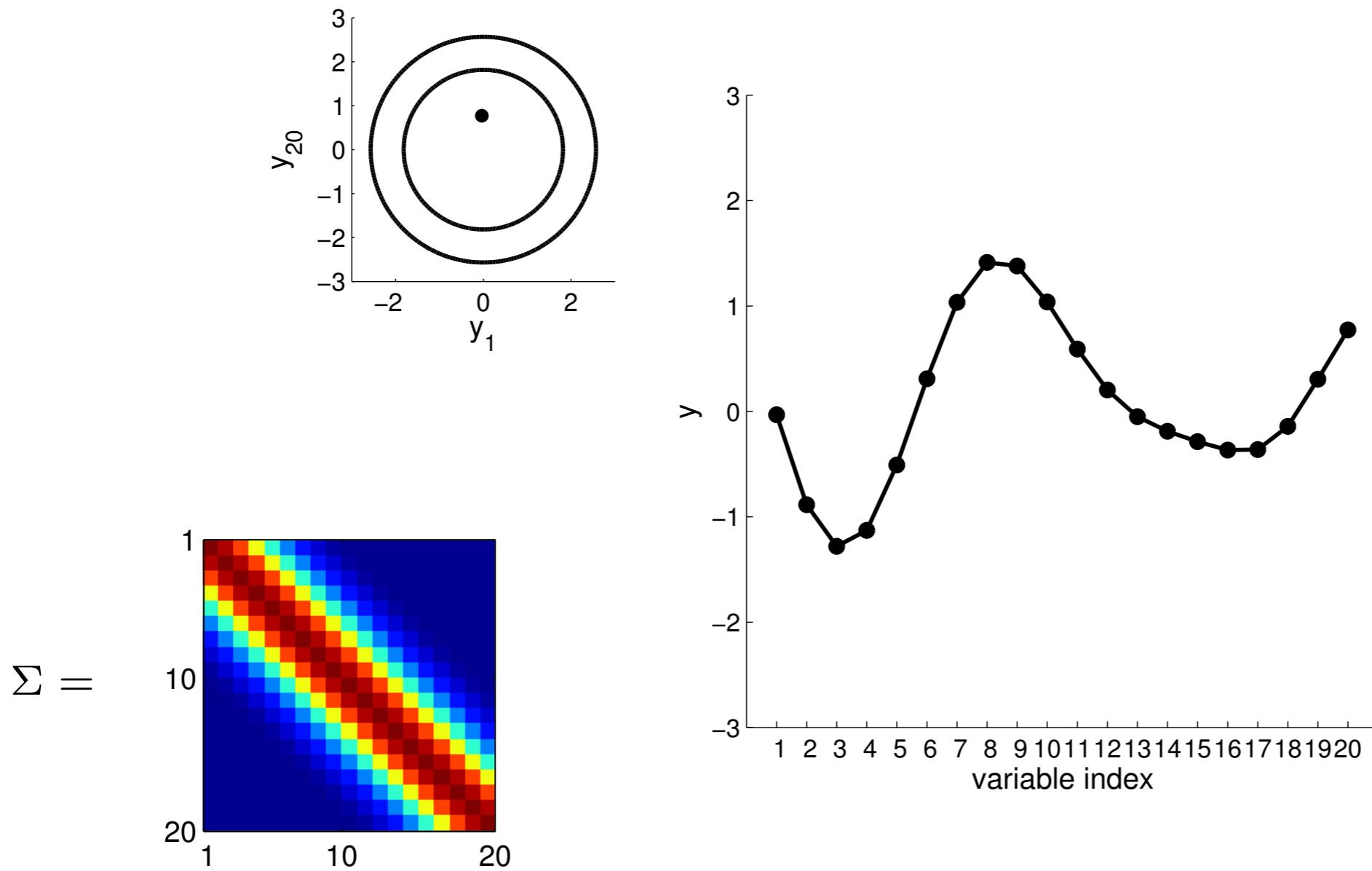
$\Sigma =$



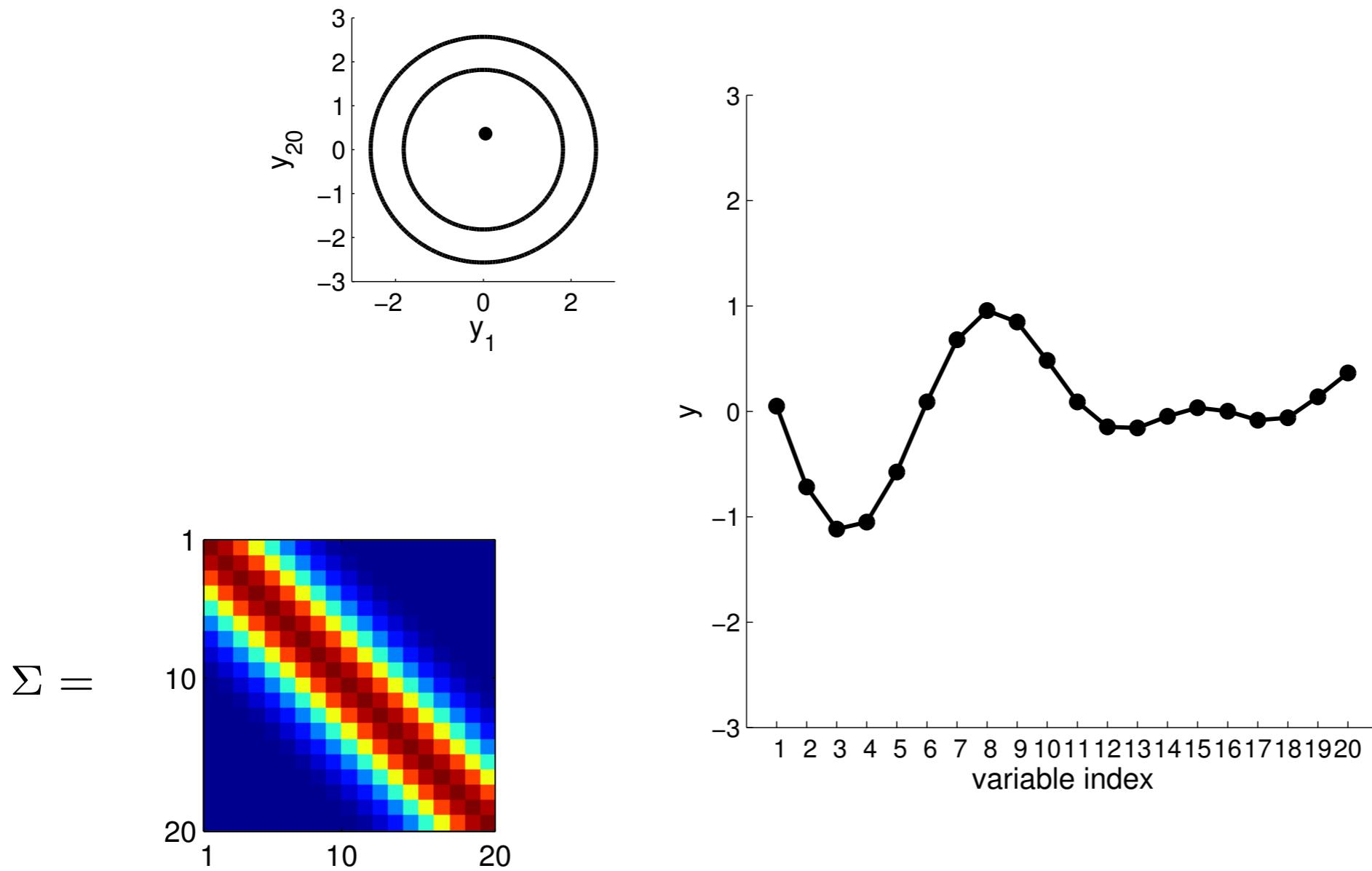
# Special covariance matrix



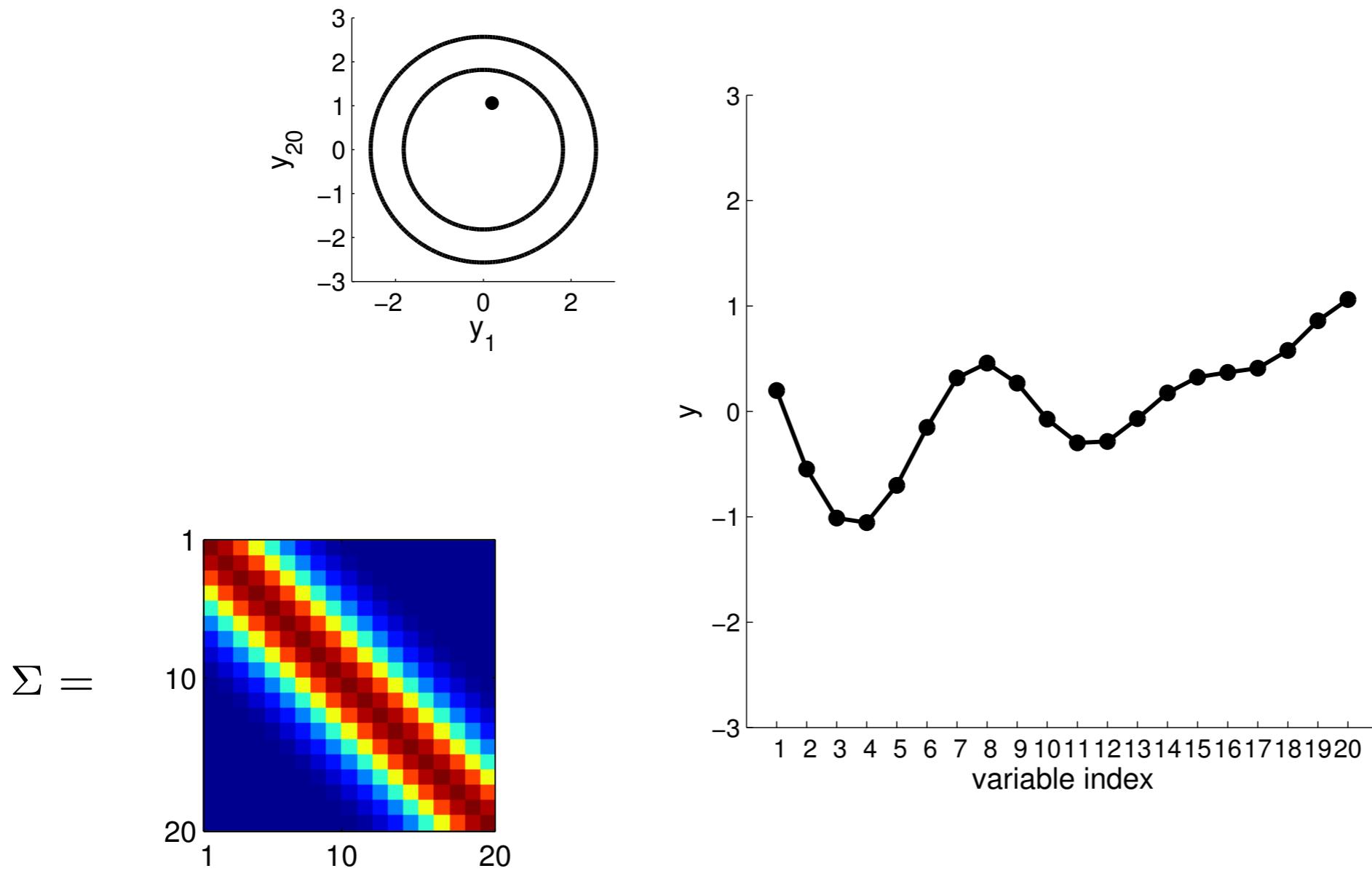
# Special covariance matrix



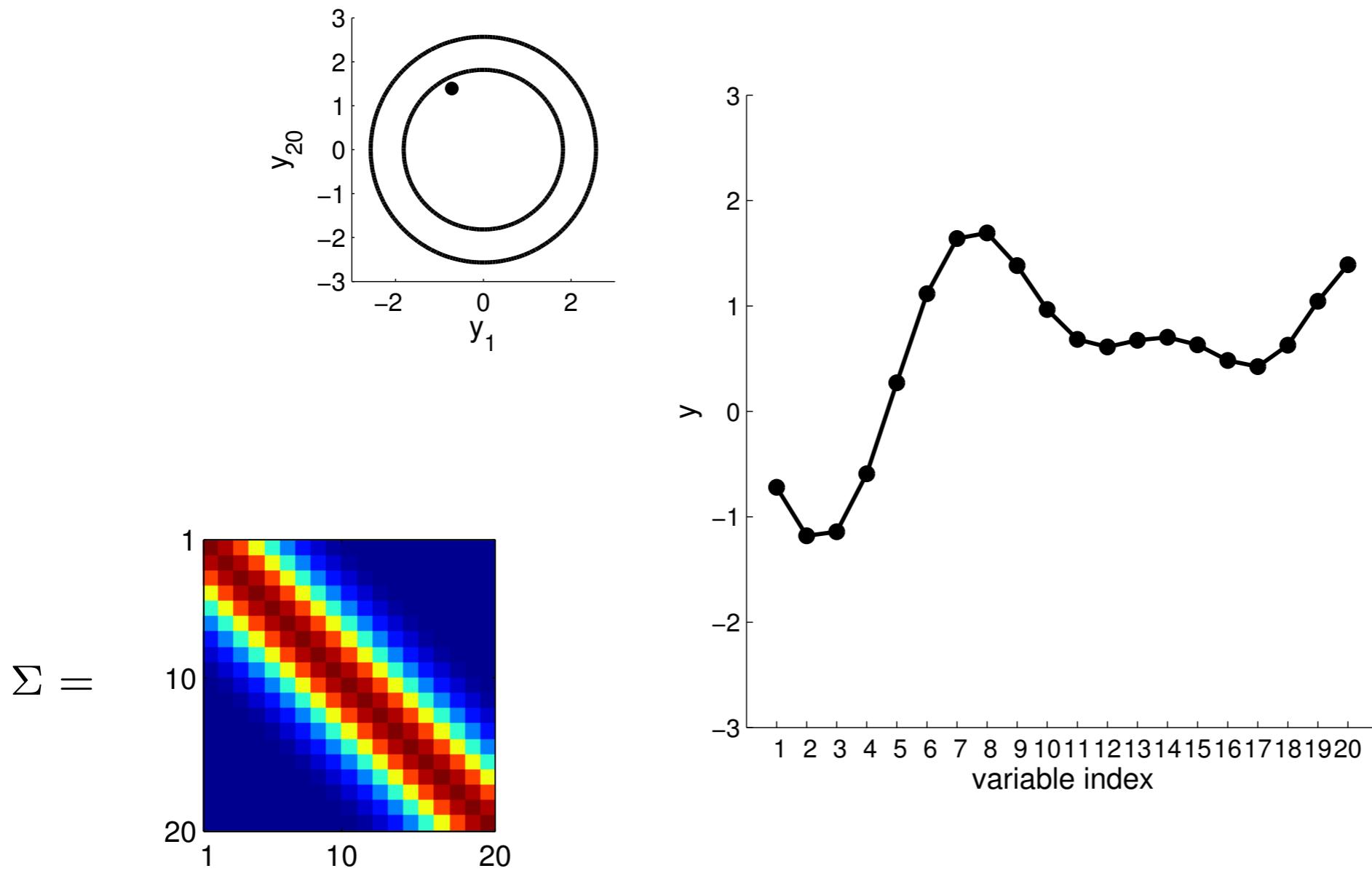
# Special covariance matrix



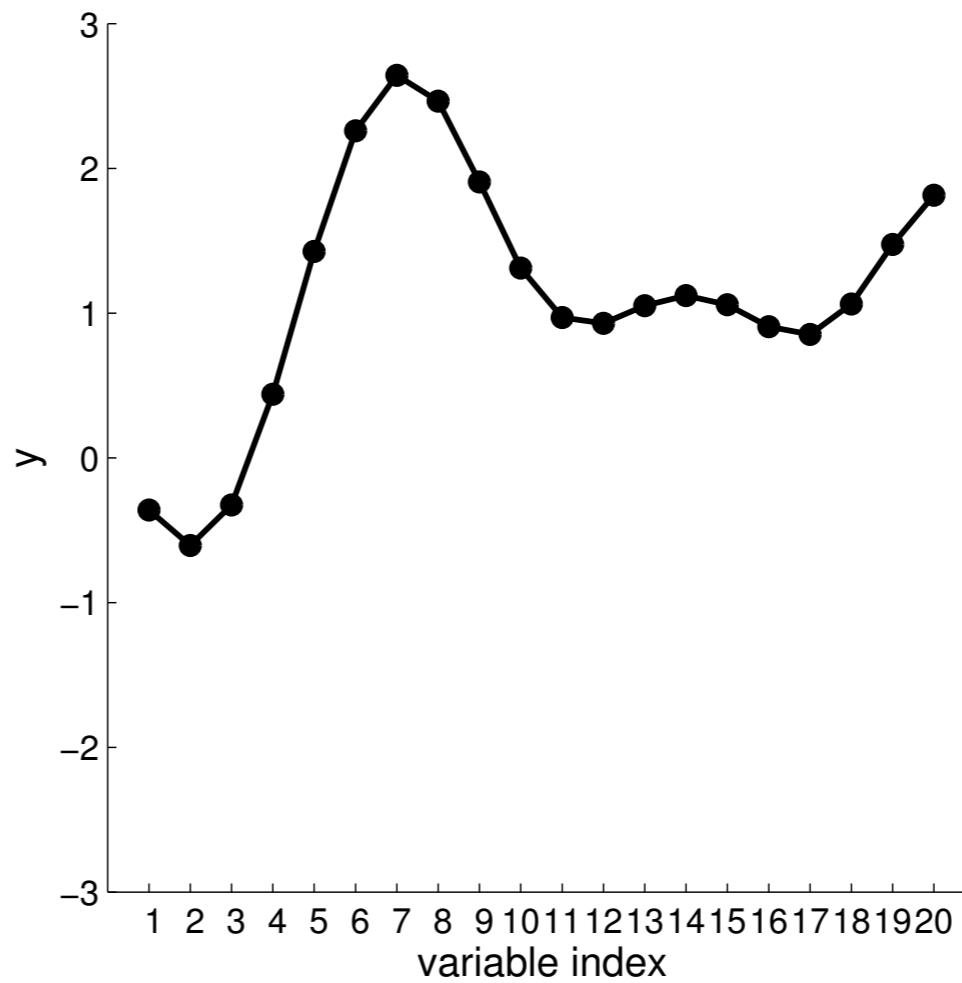
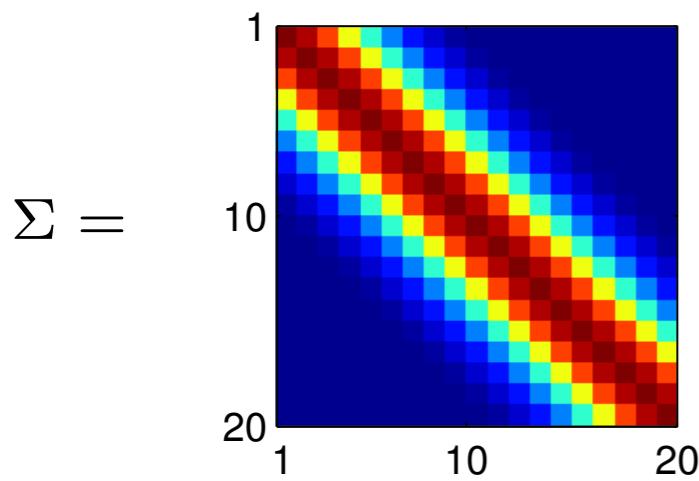
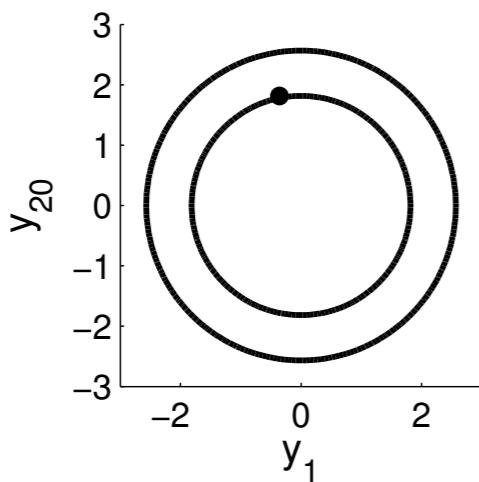
# Special covariance matrix



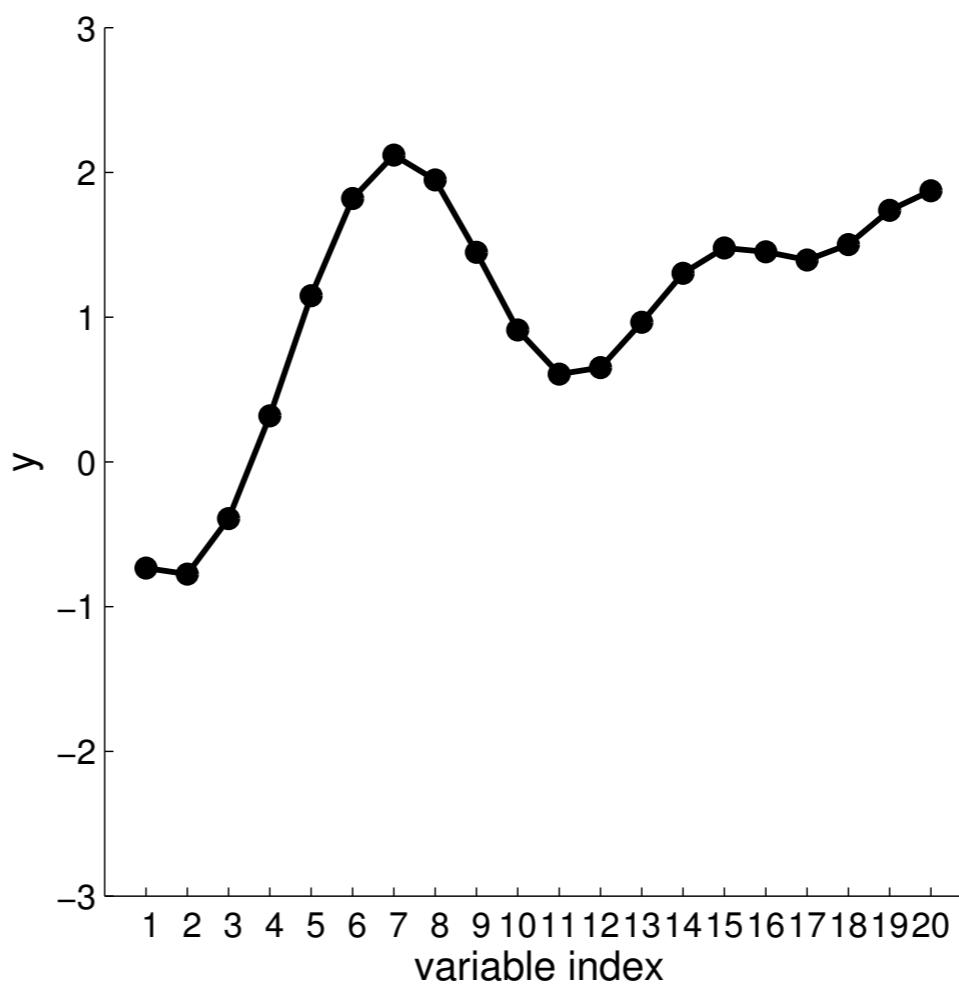
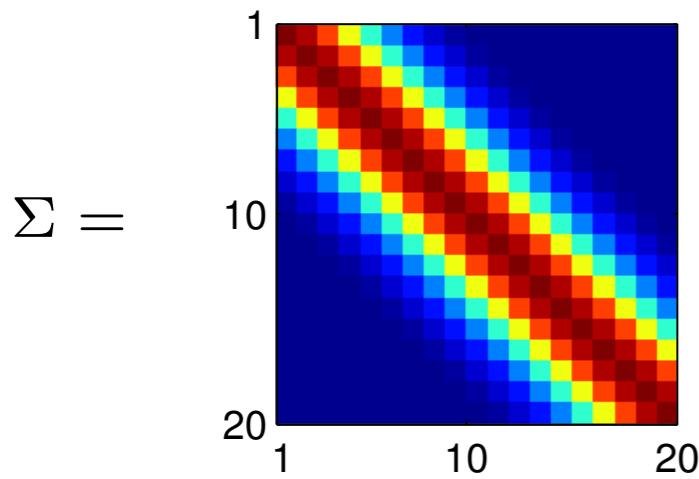
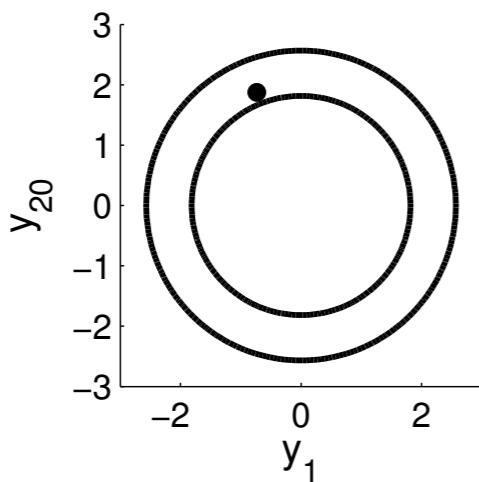
# Special covariance matrix



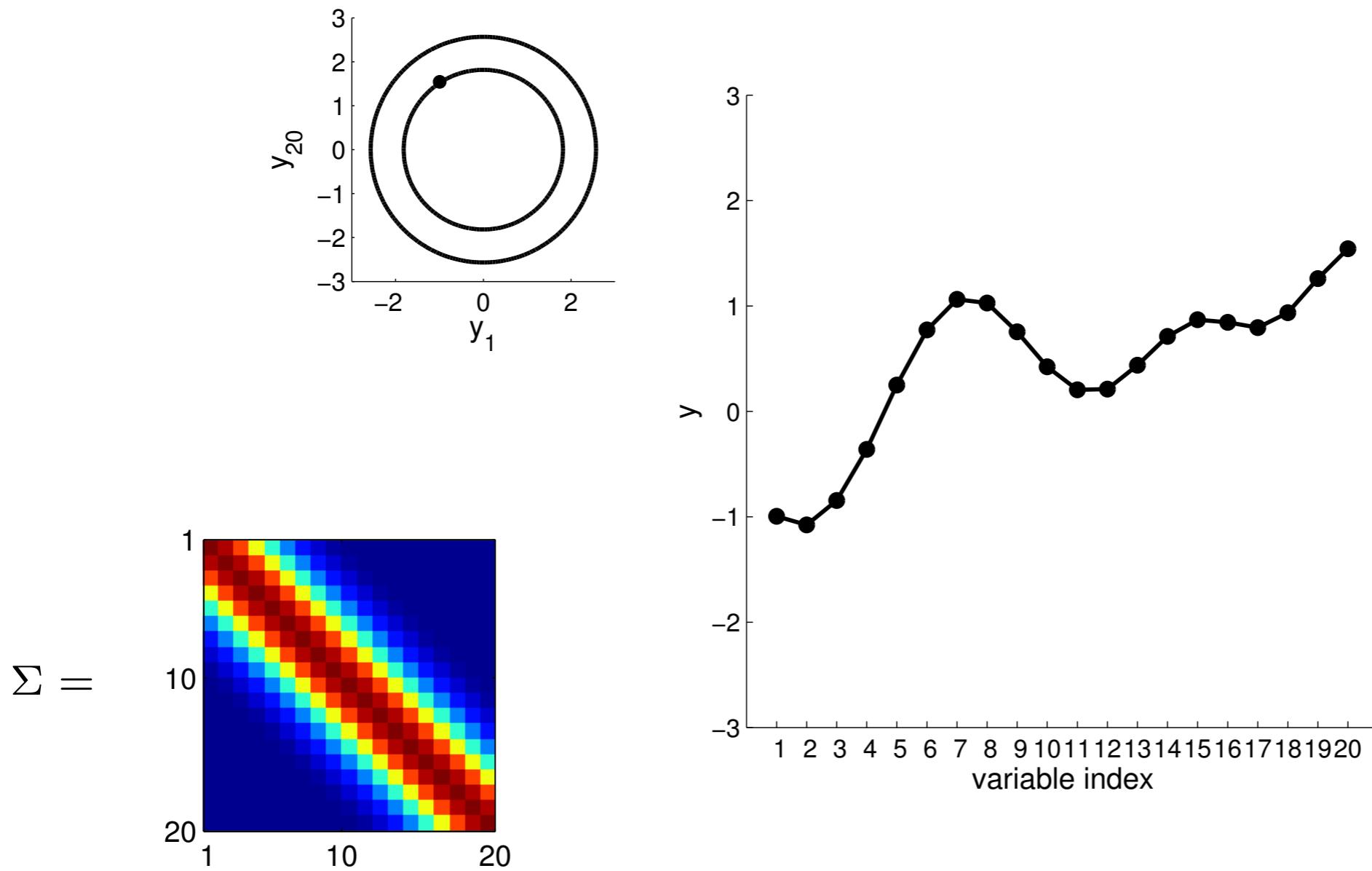
# Special covariance matrix



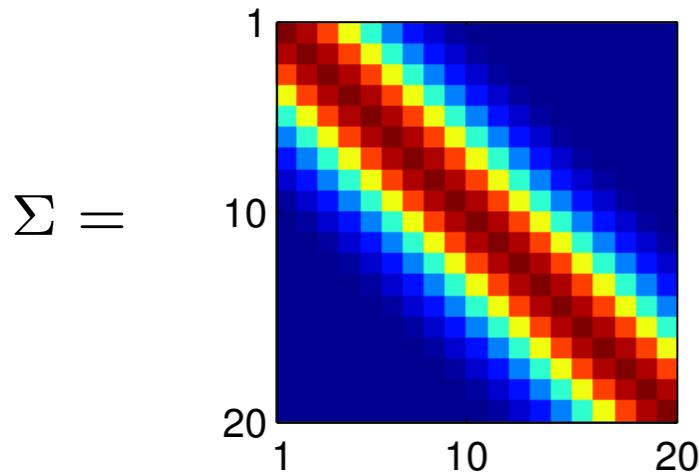
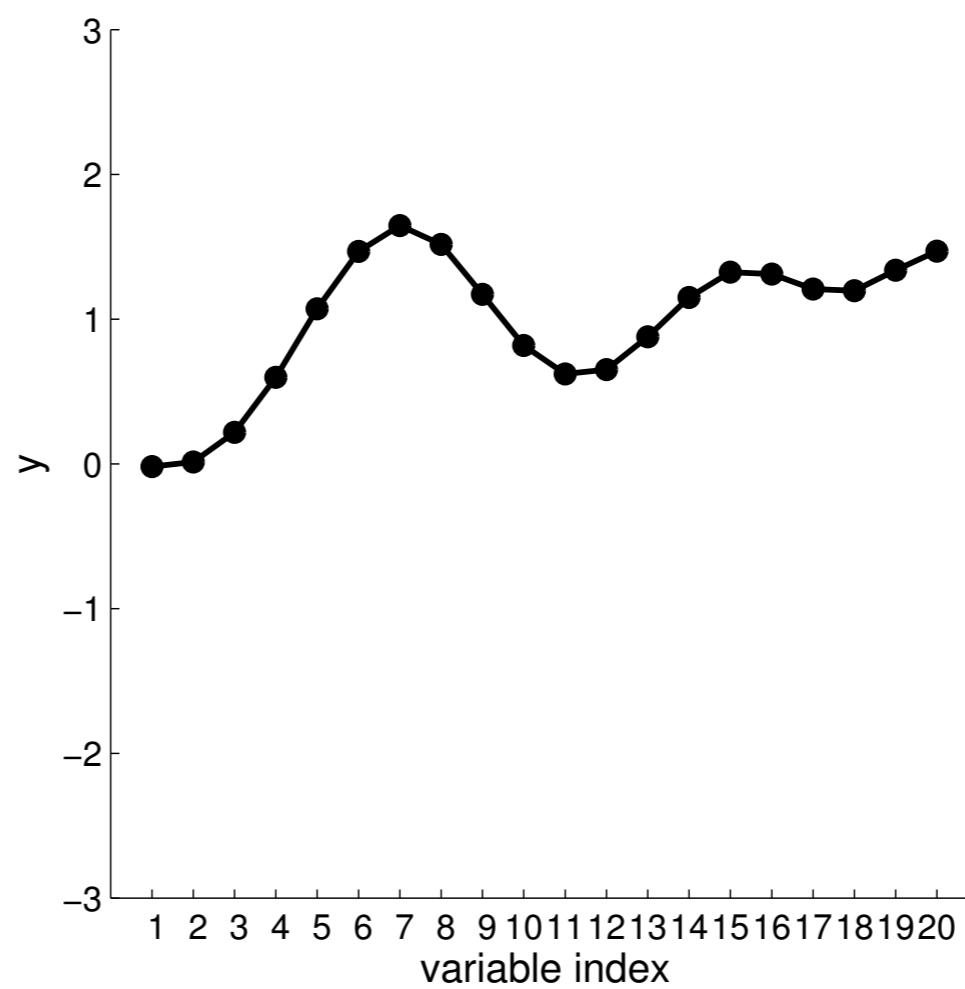
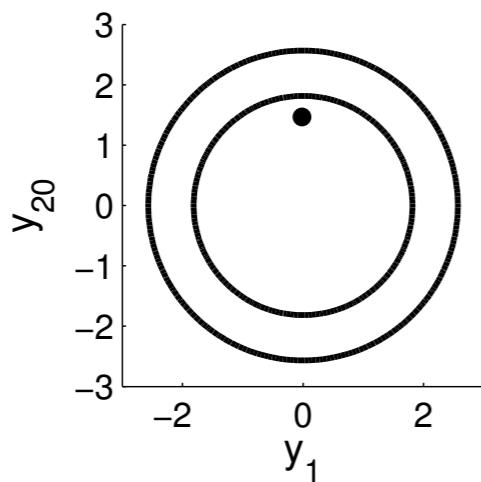
# Special covariance matrix



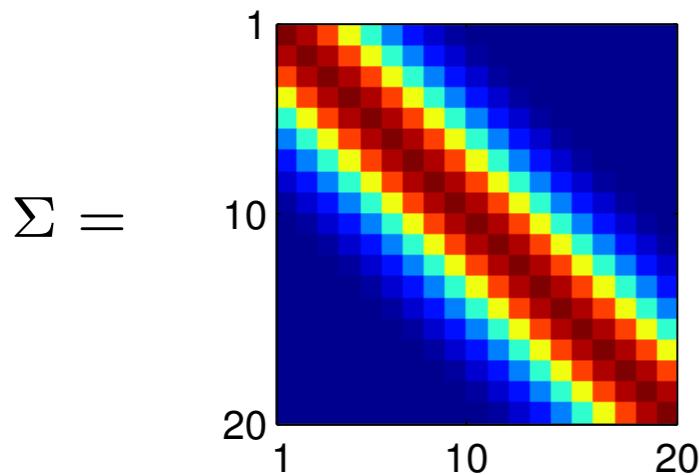
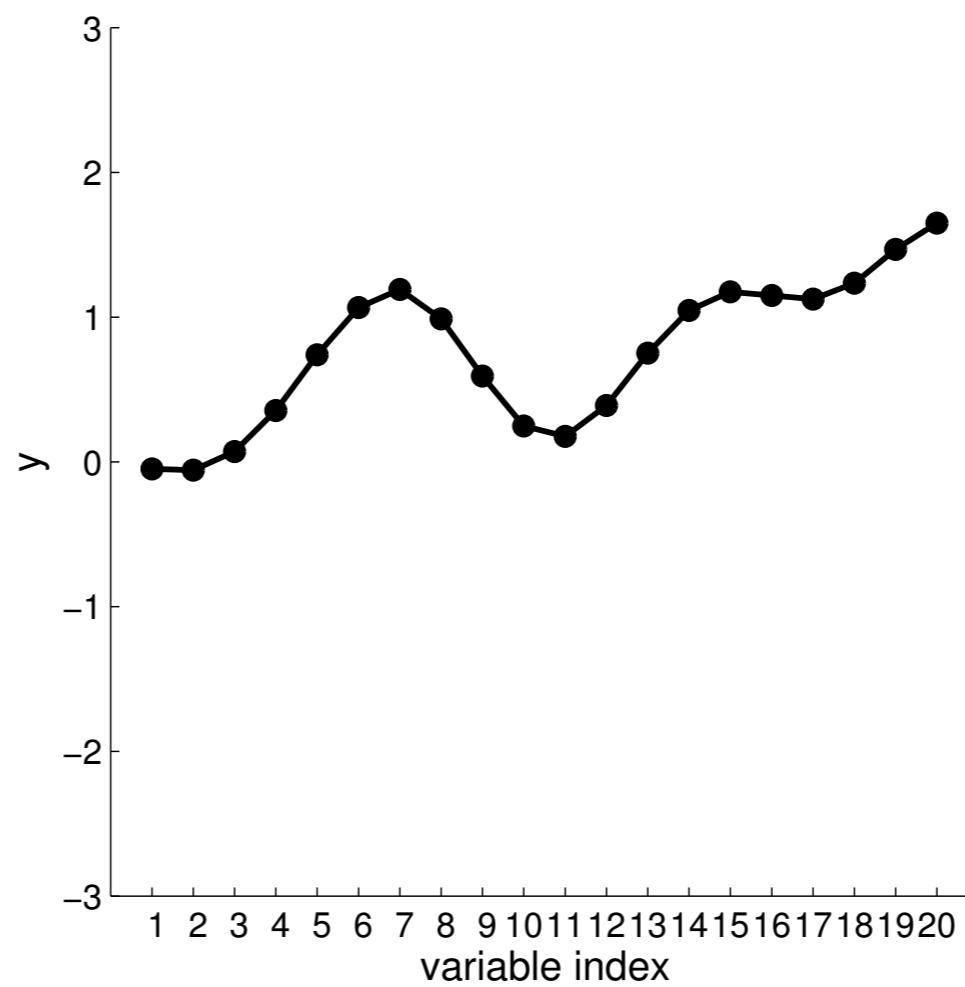
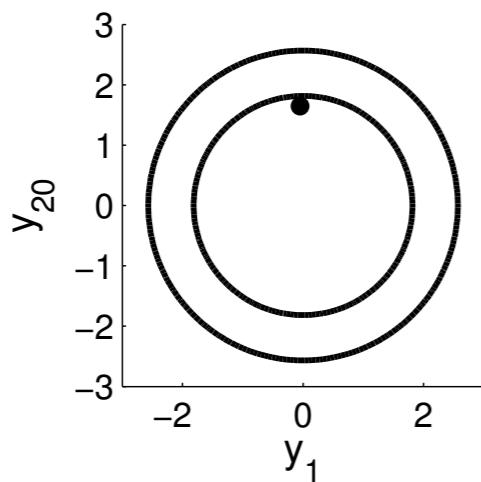
# Special covariance matrix



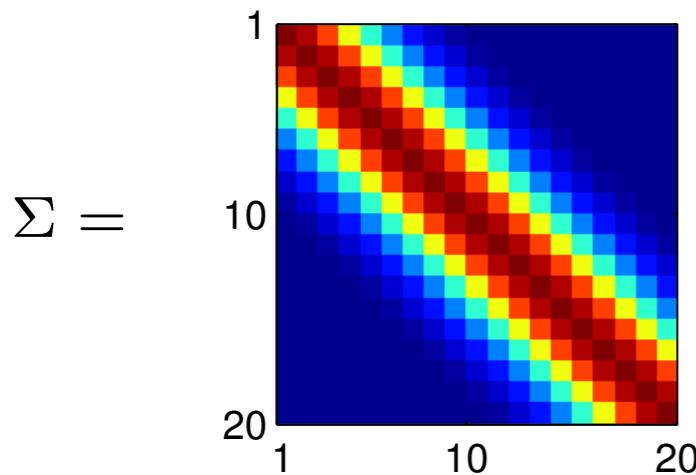
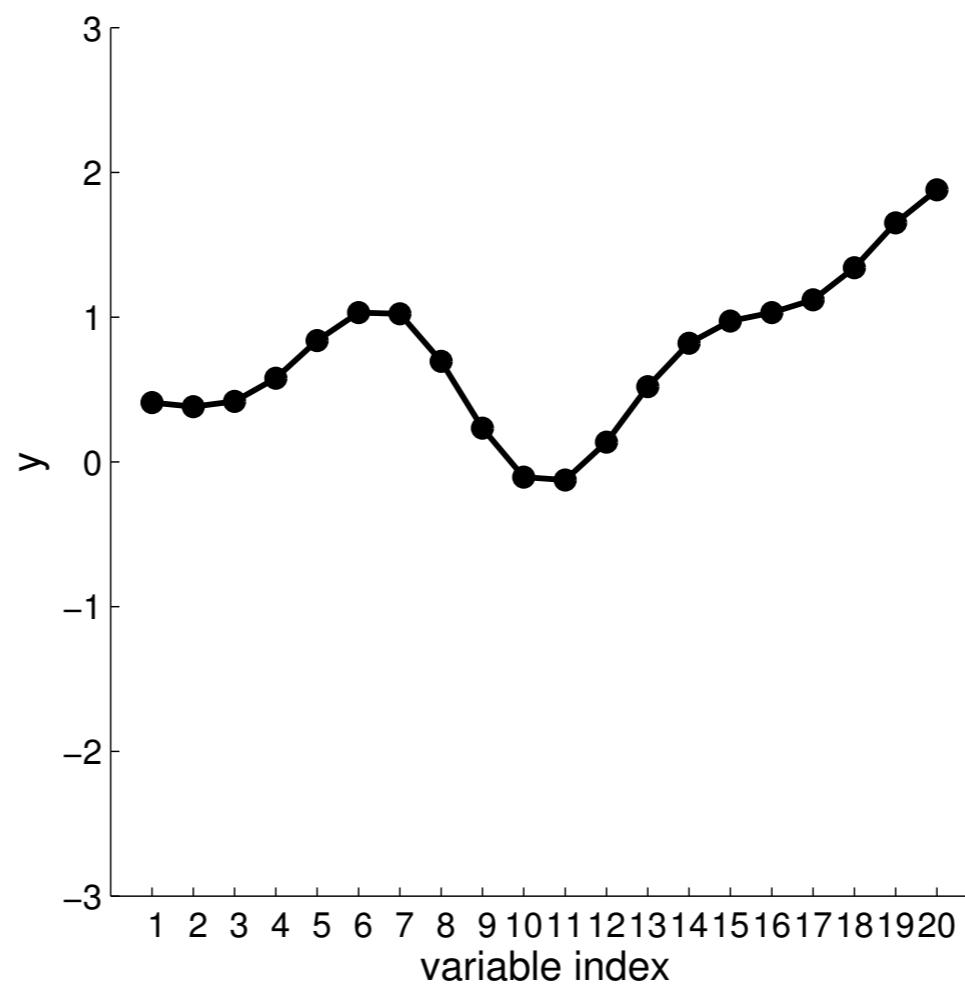
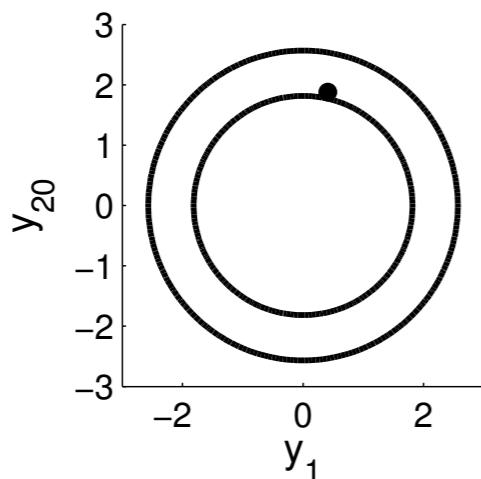
# Special covariance matrix



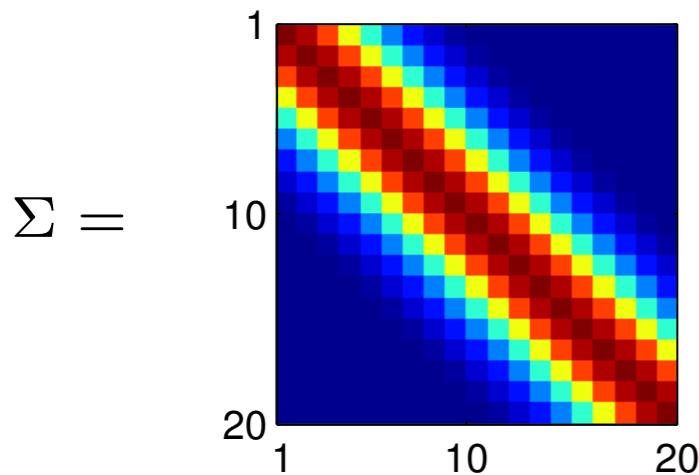
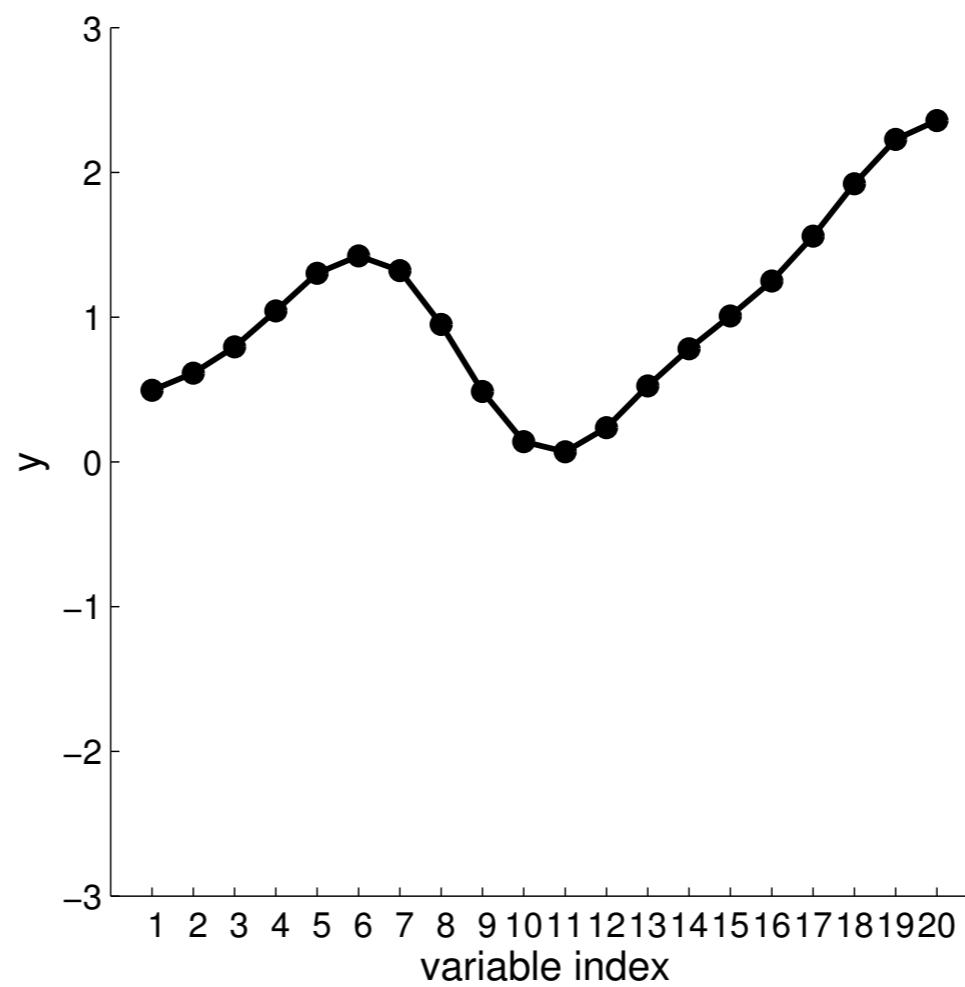
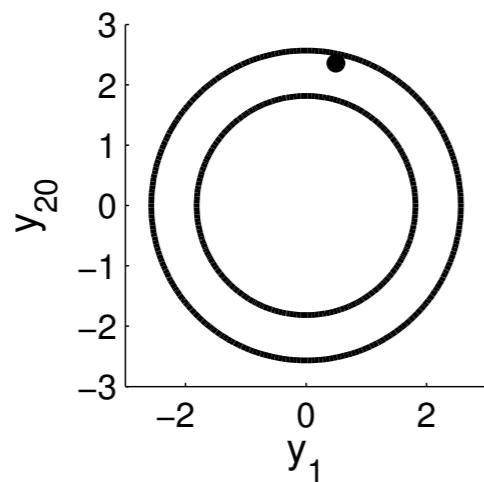
# Special covariance matrix



# Special covariance matrix

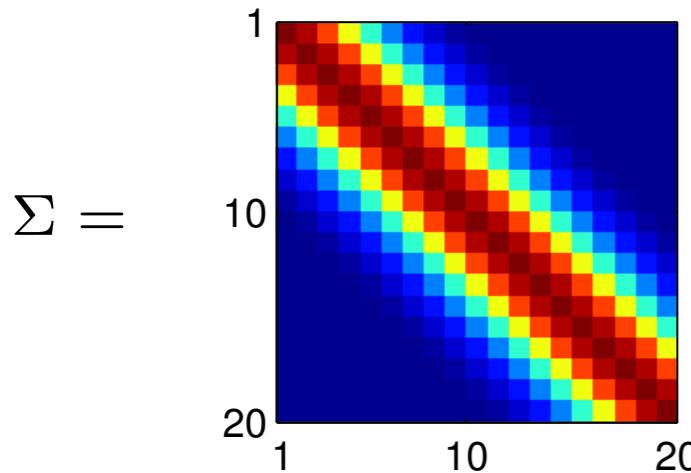
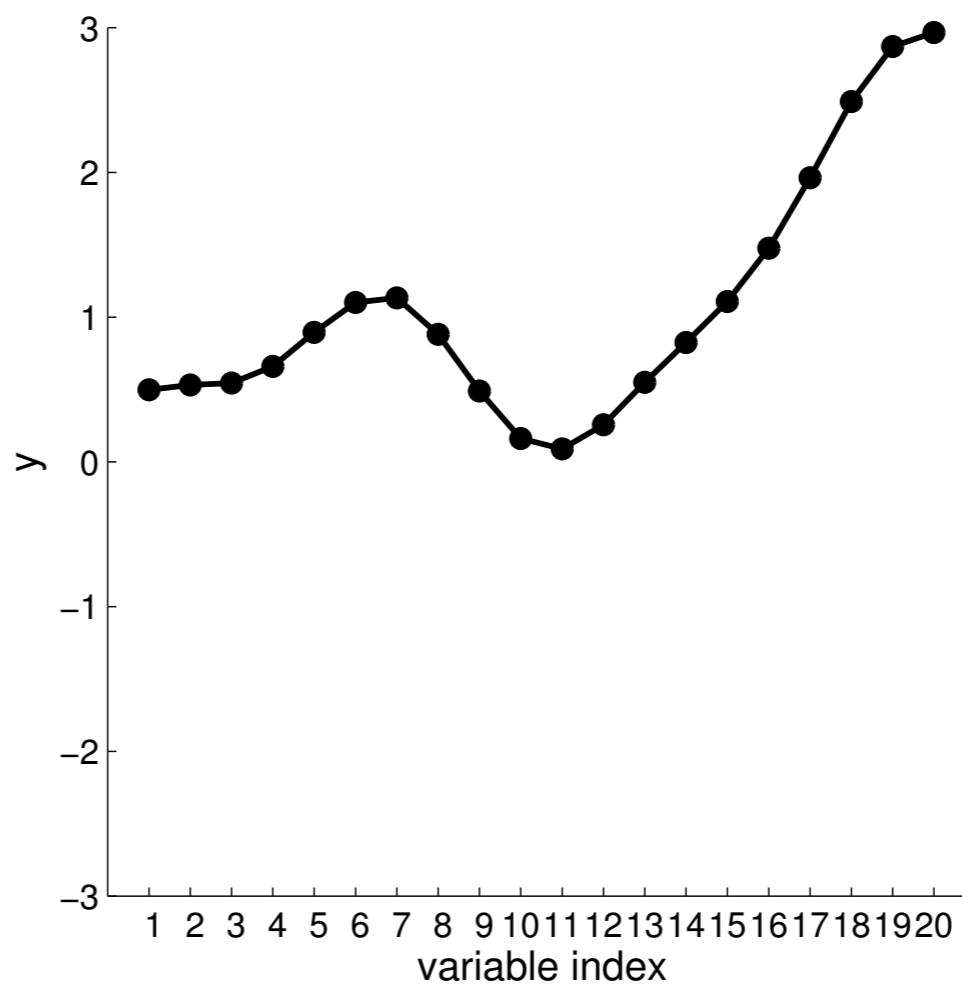
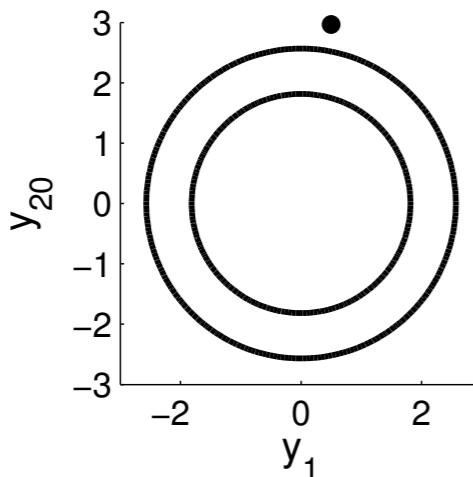


# Special covariance matrix

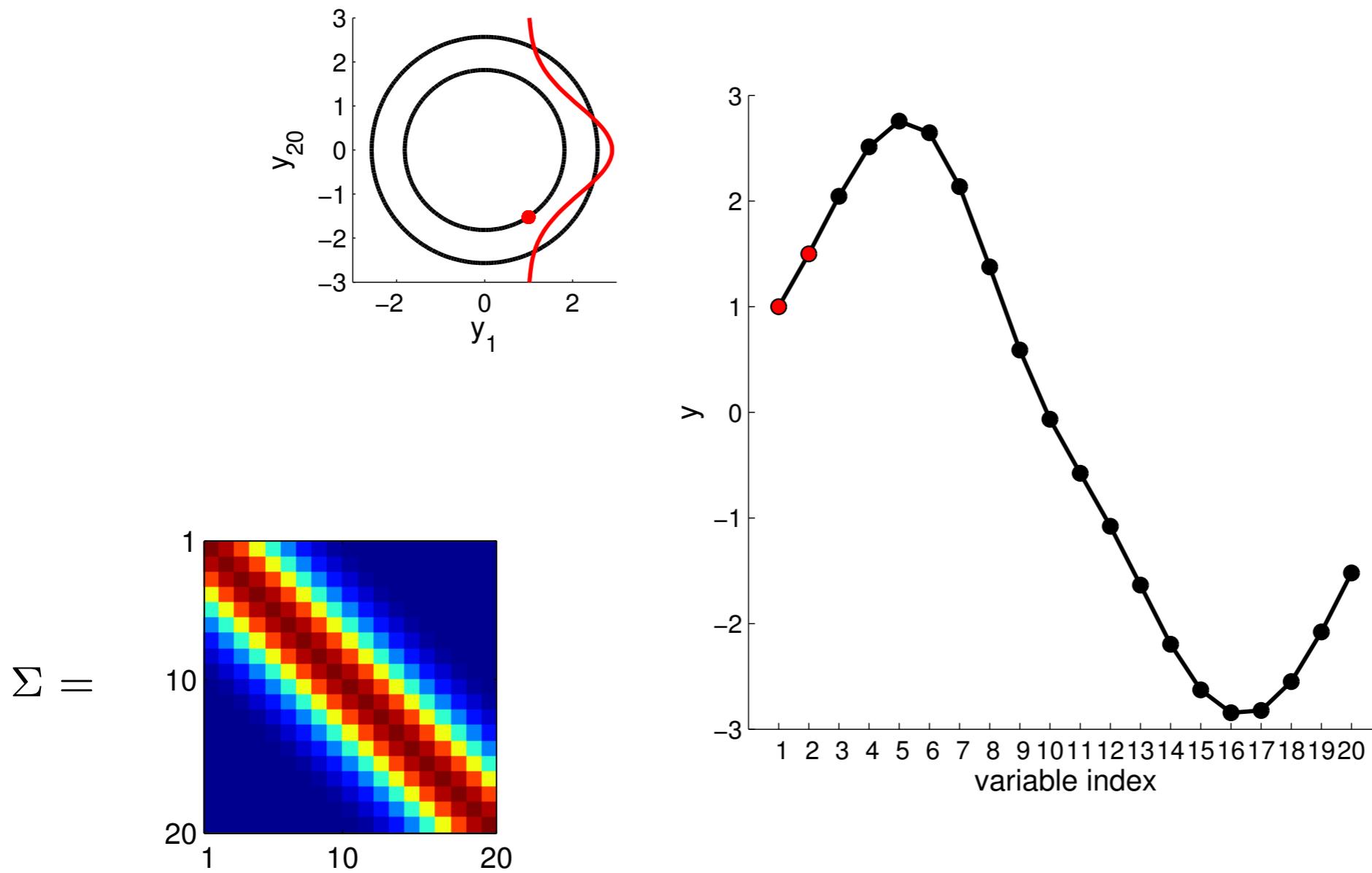


# Special covariance matrix

What do those samples look like? Just smooth functions. Our prior is that functions are smooth: in neabry points, we see nearby values

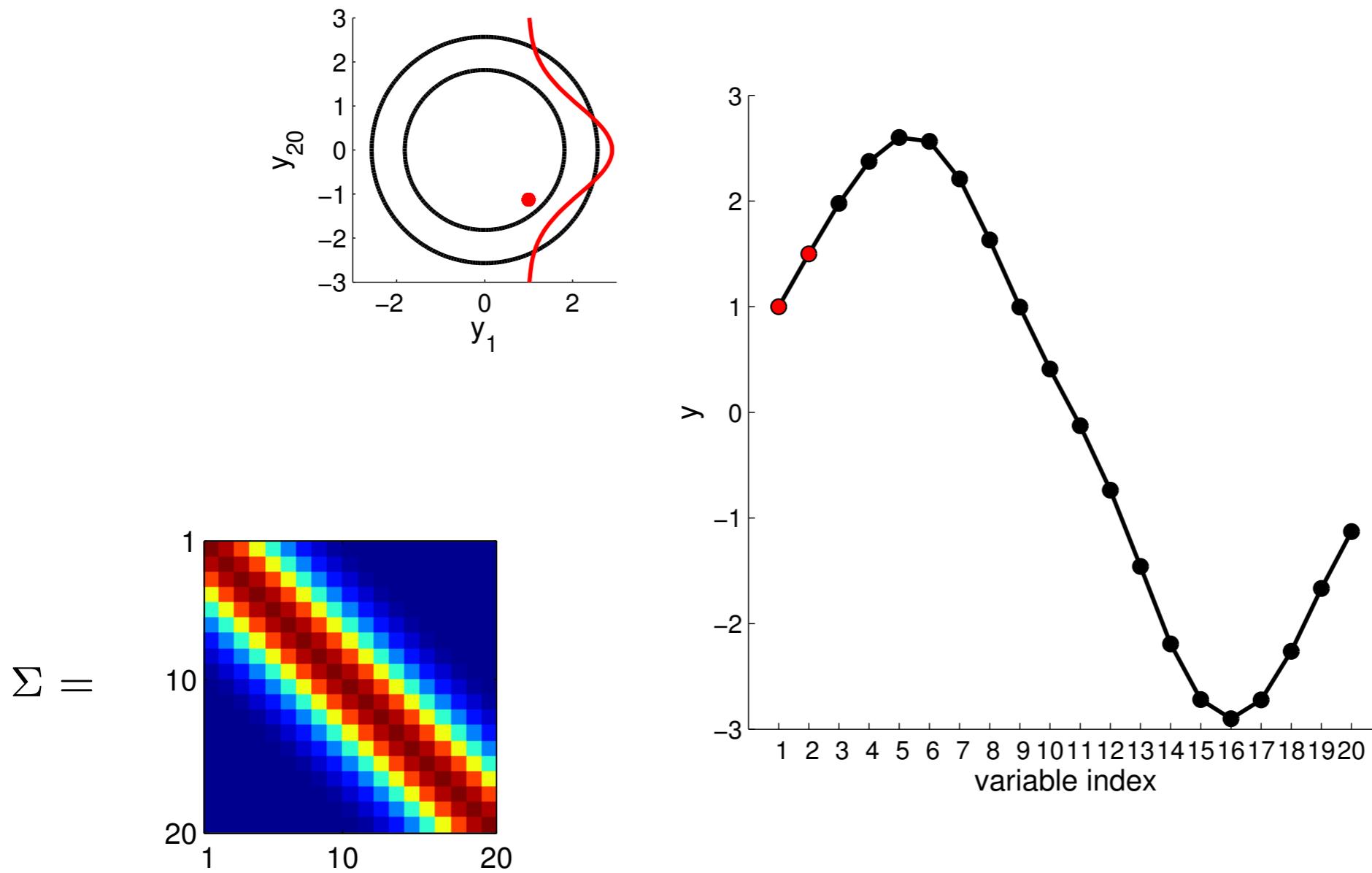


# Special covariance matrix - conditioning



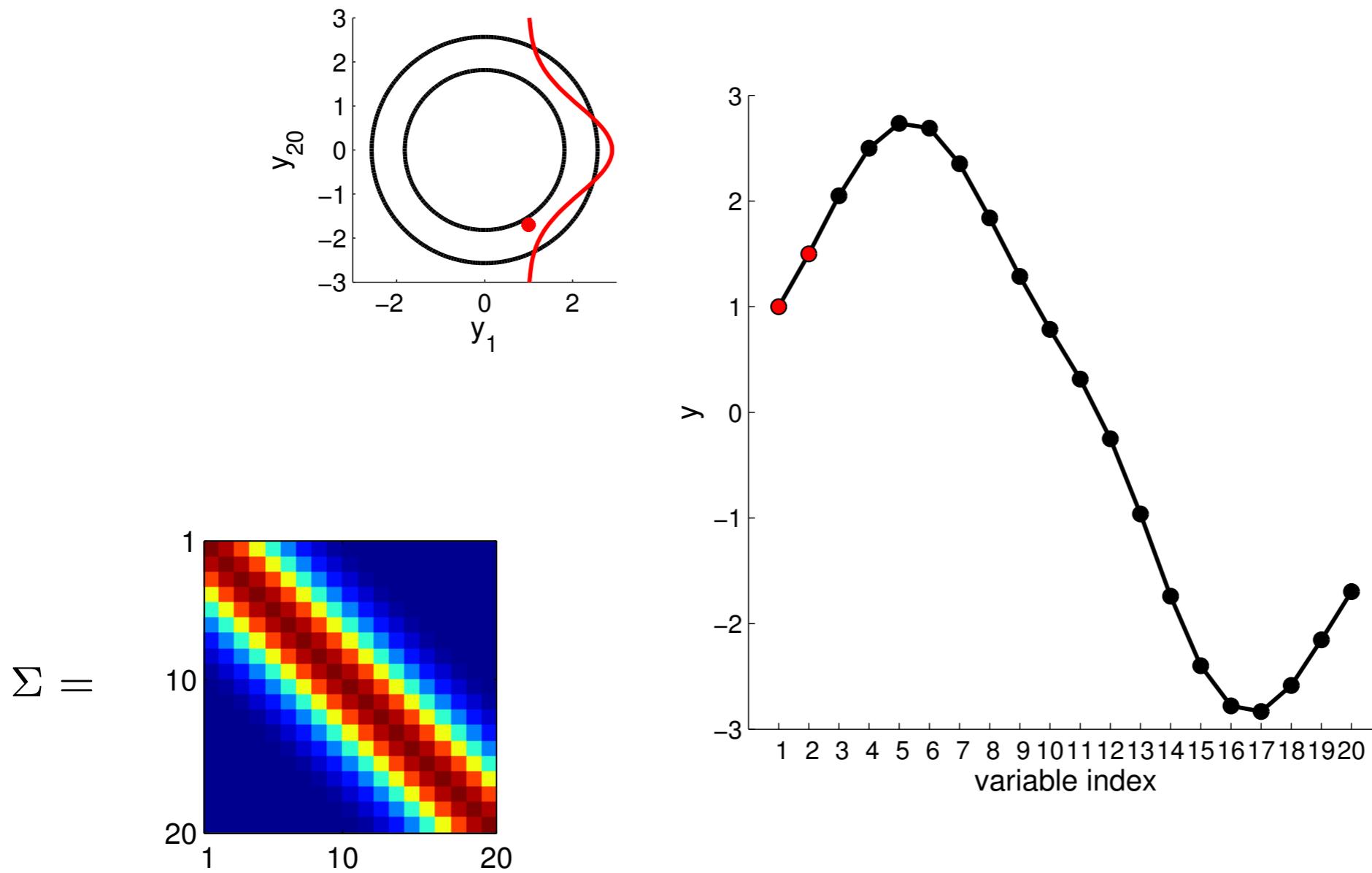
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



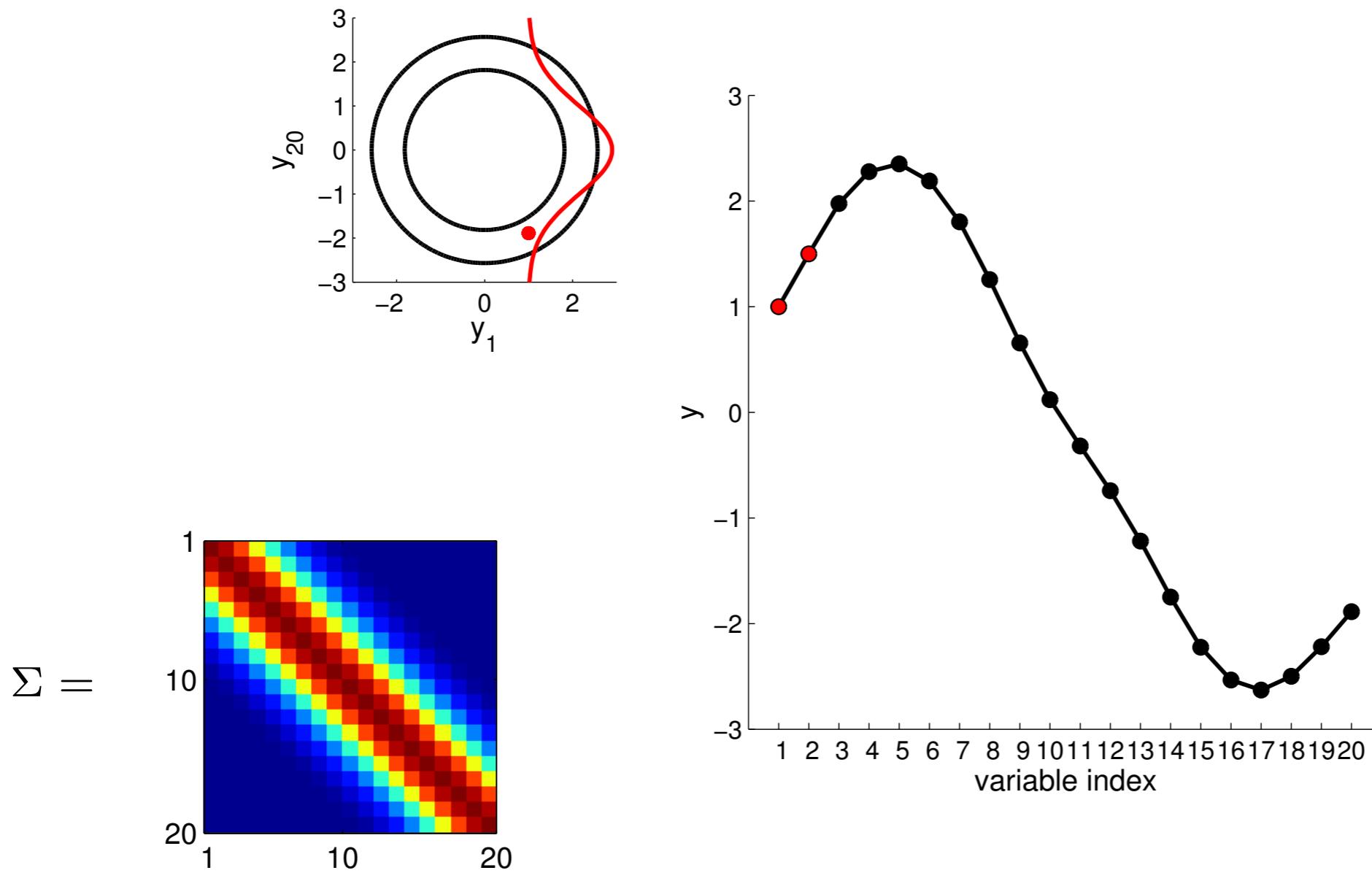
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



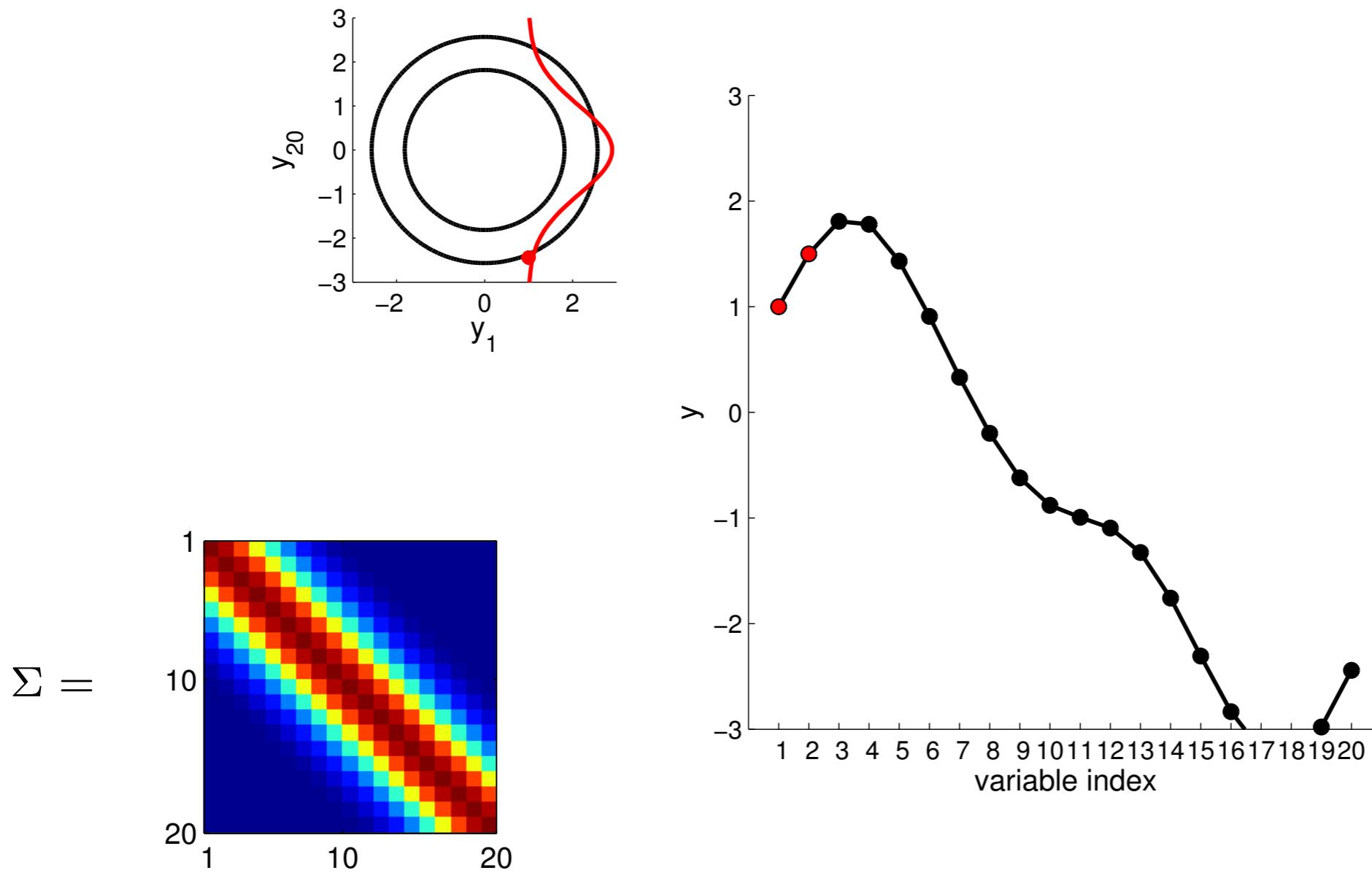
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



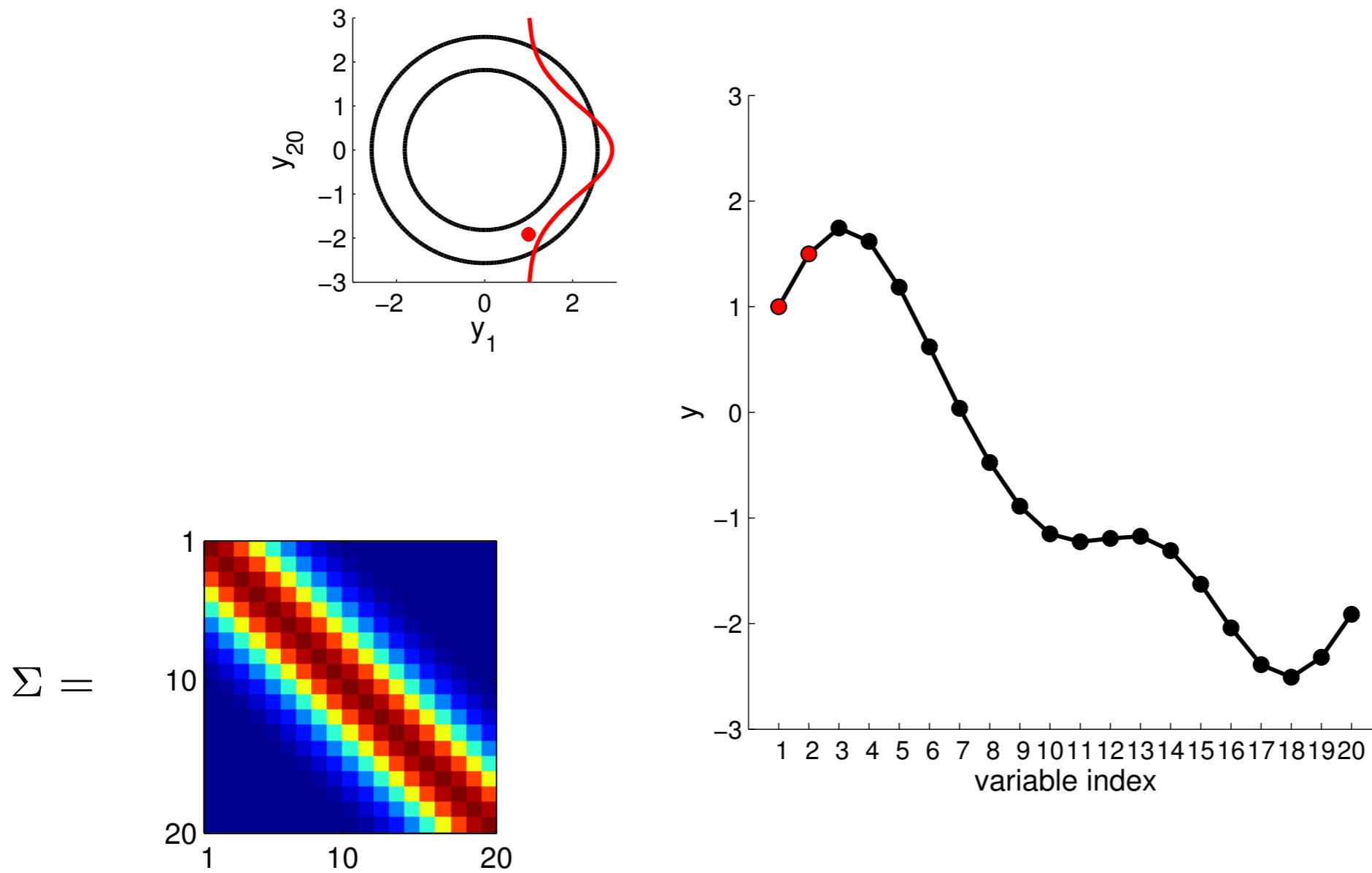
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



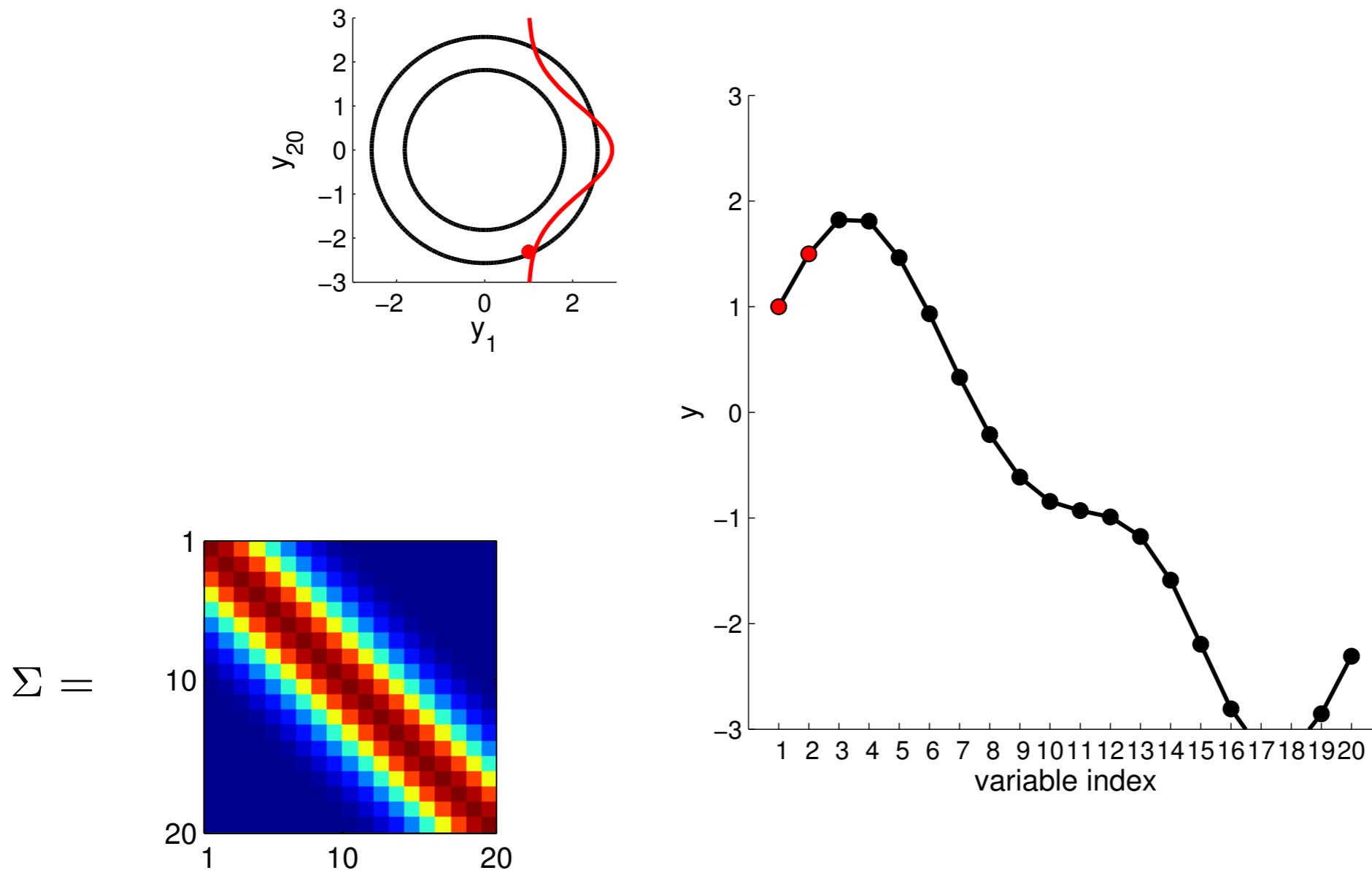
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



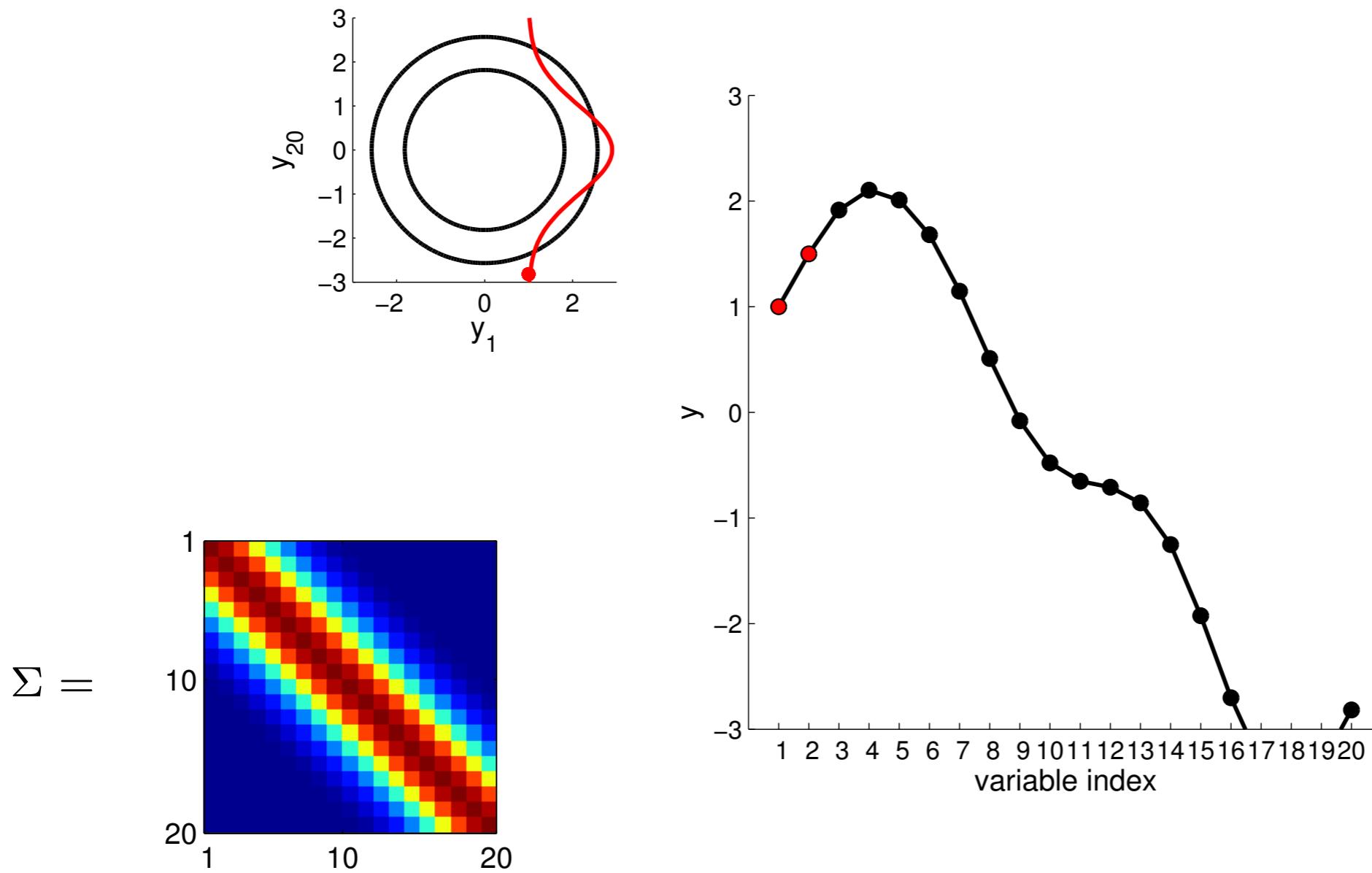
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



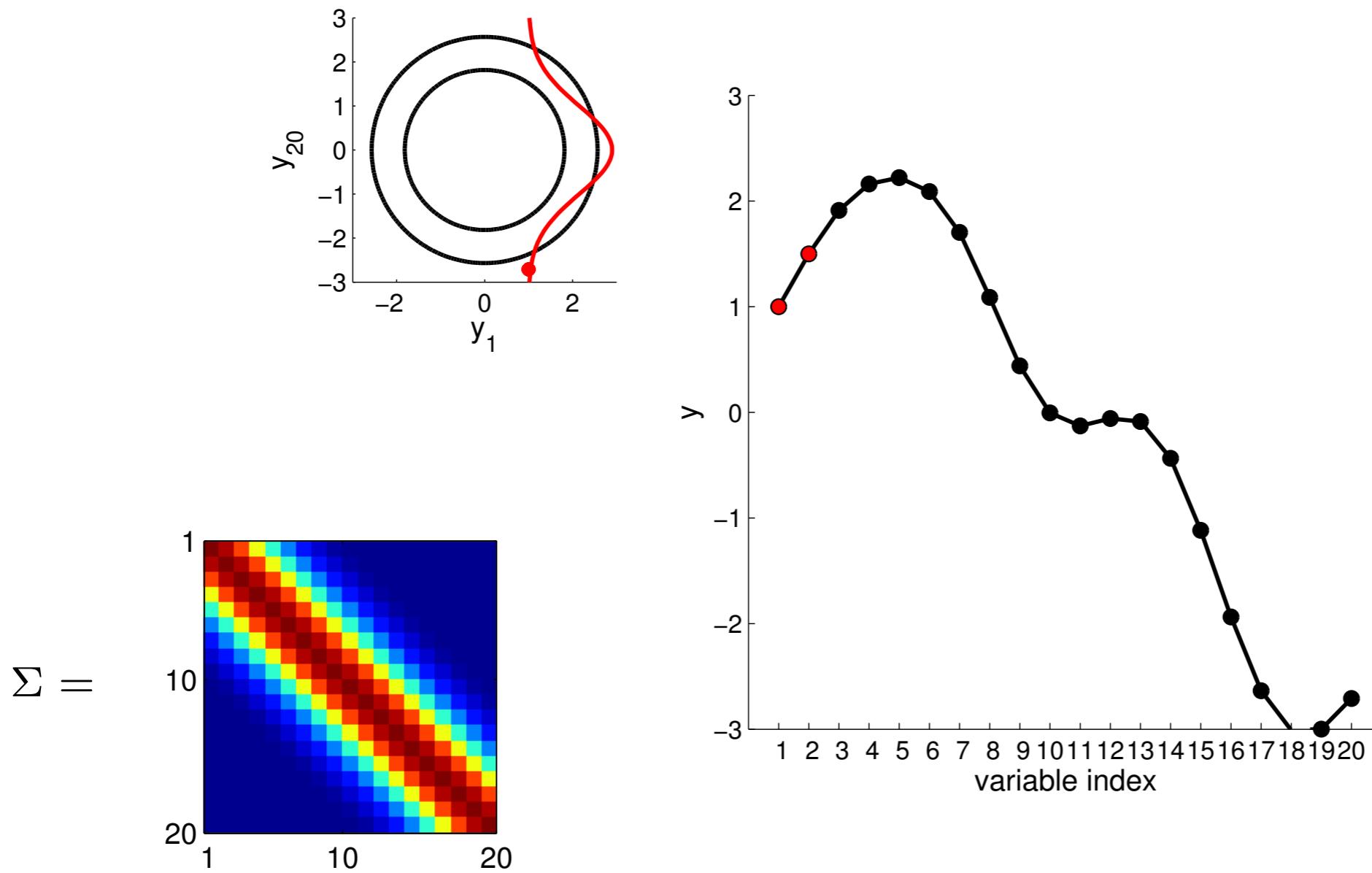
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



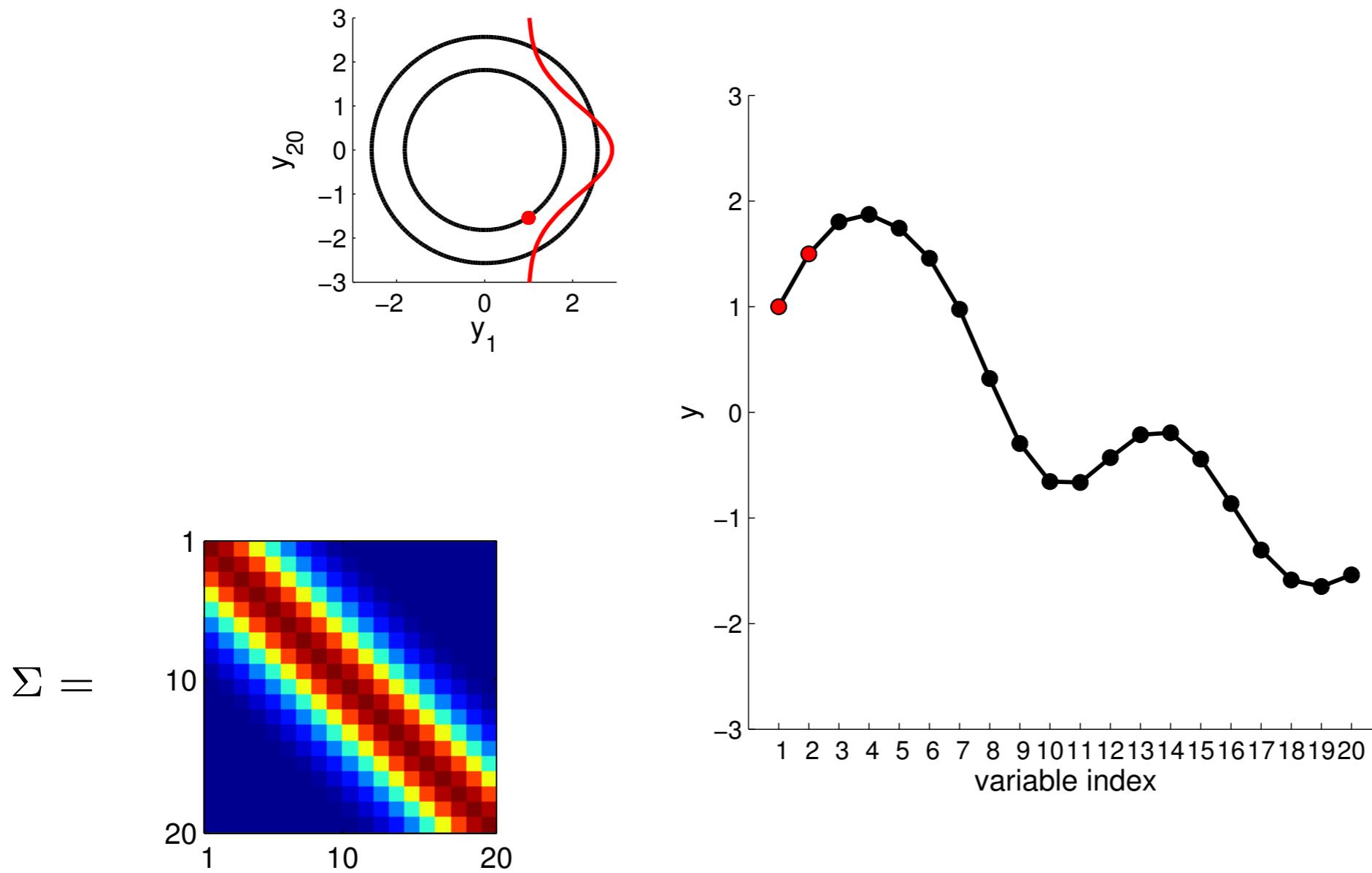
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



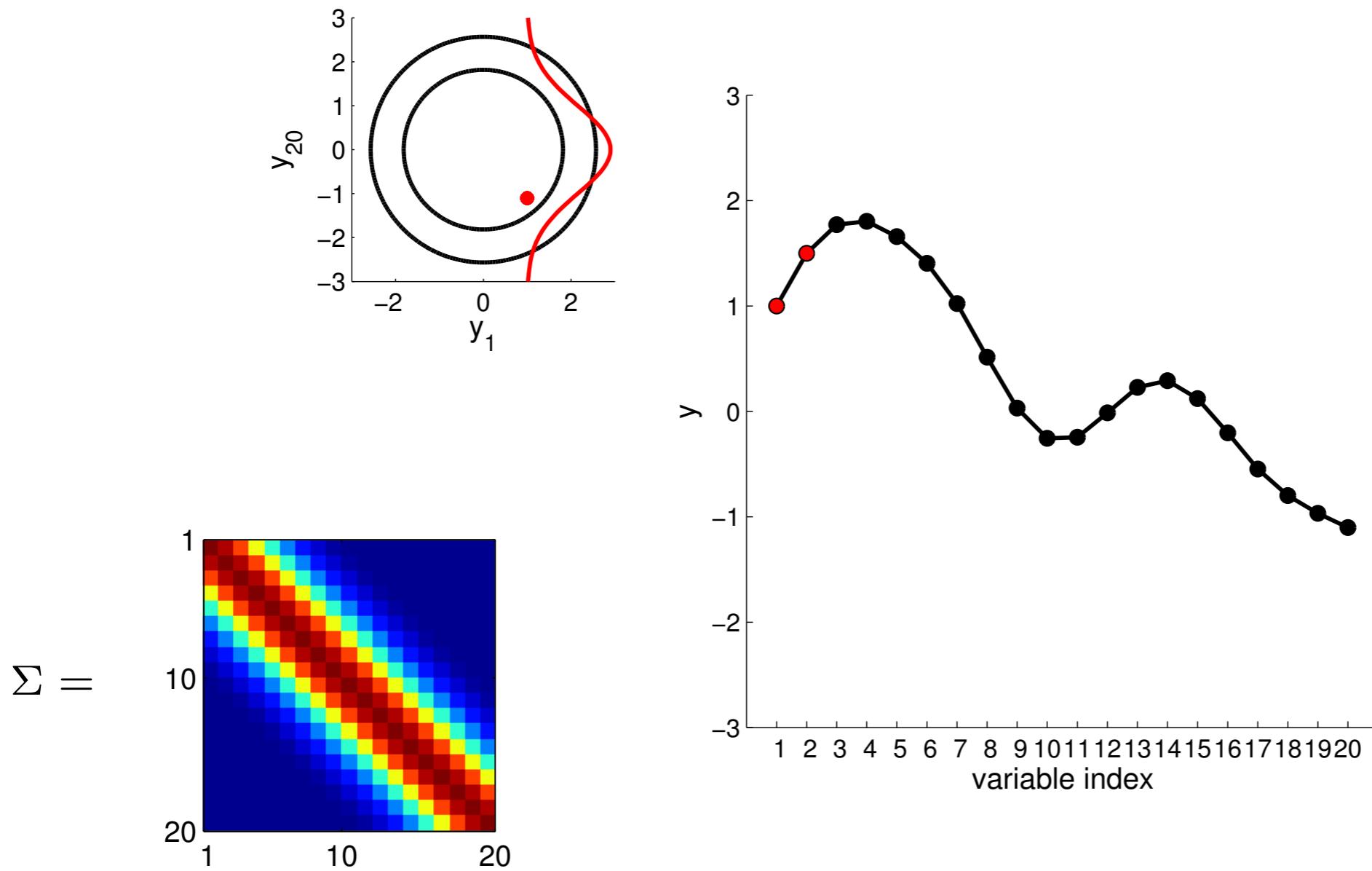
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



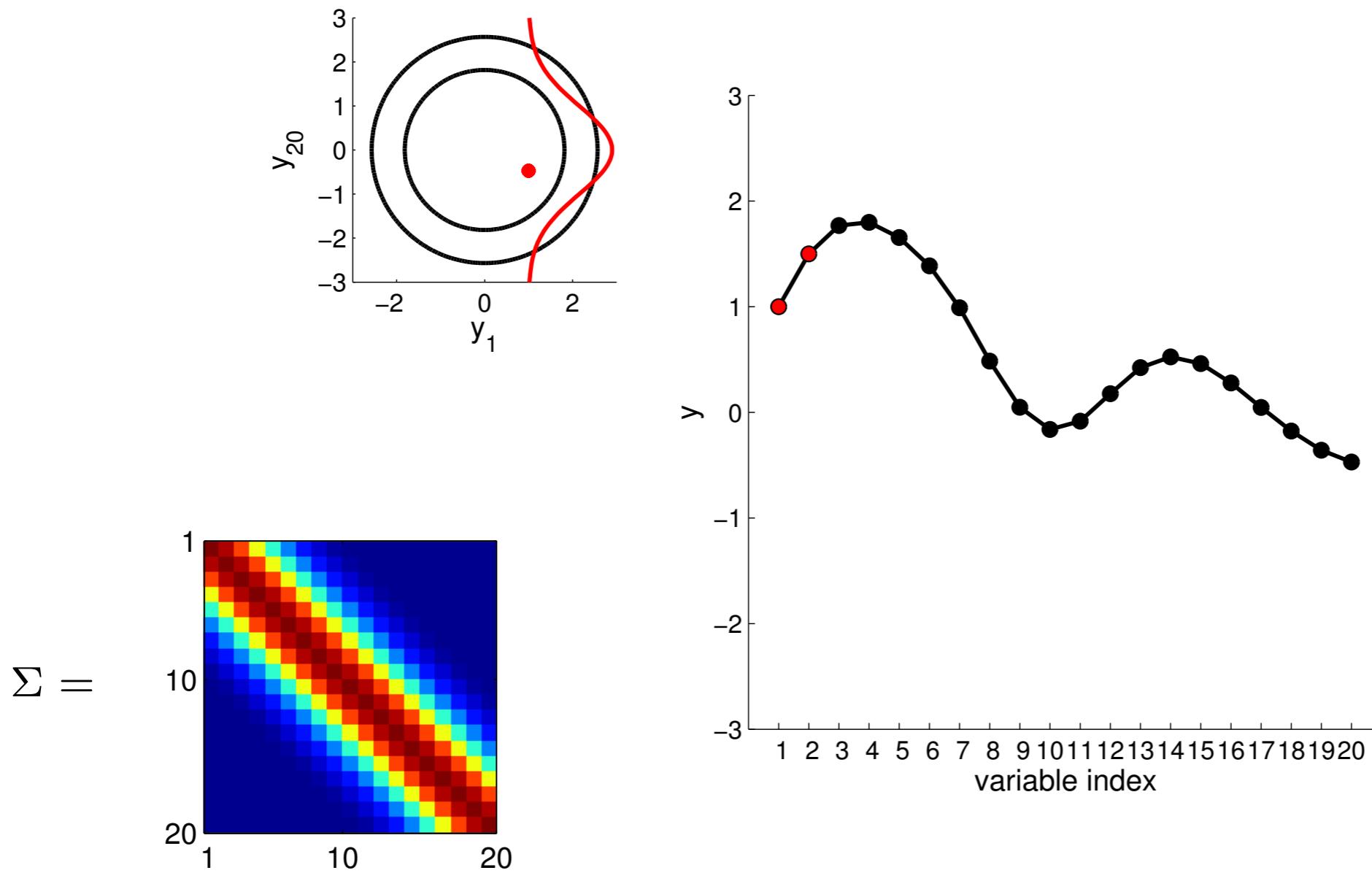
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



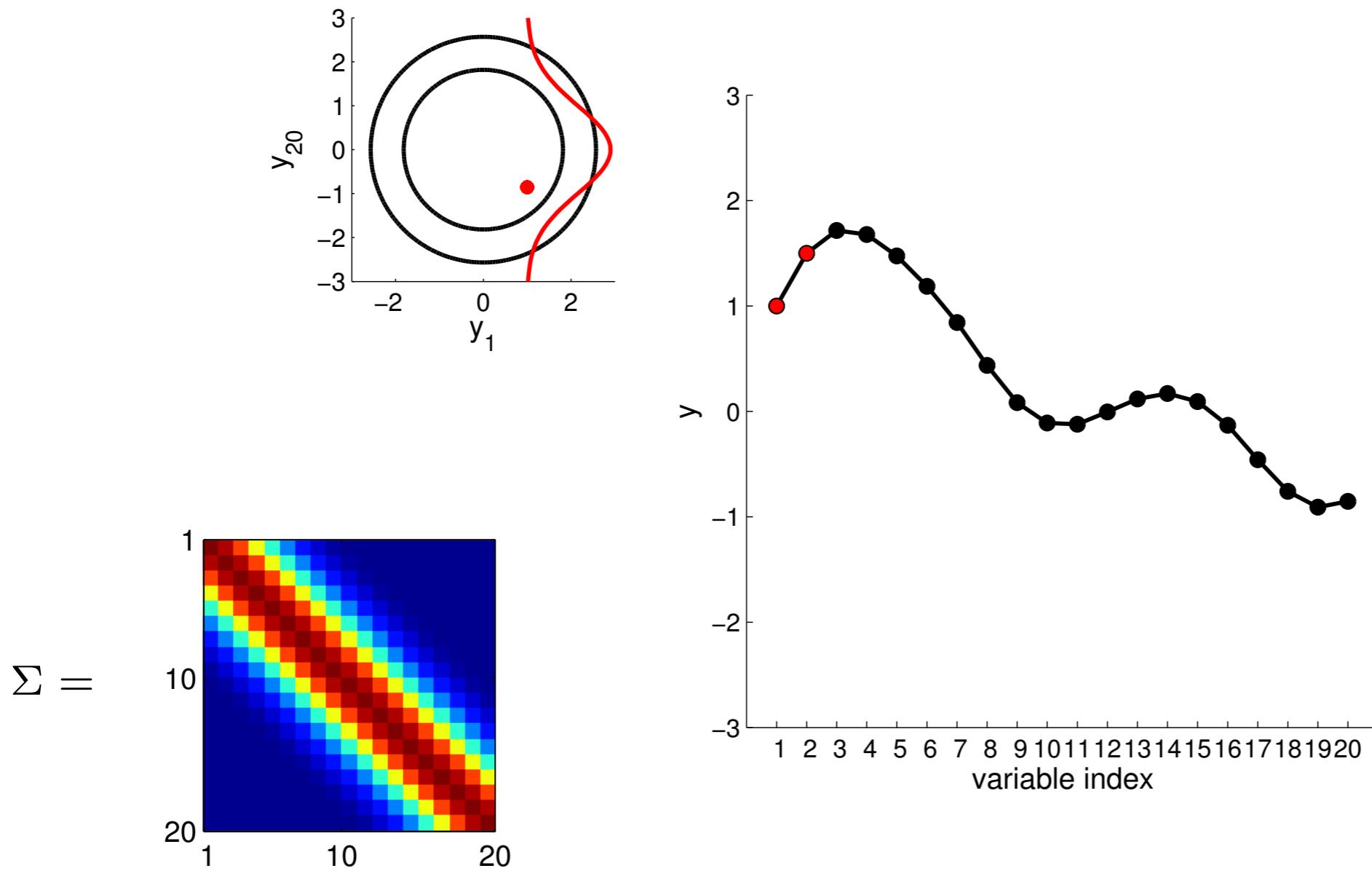
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



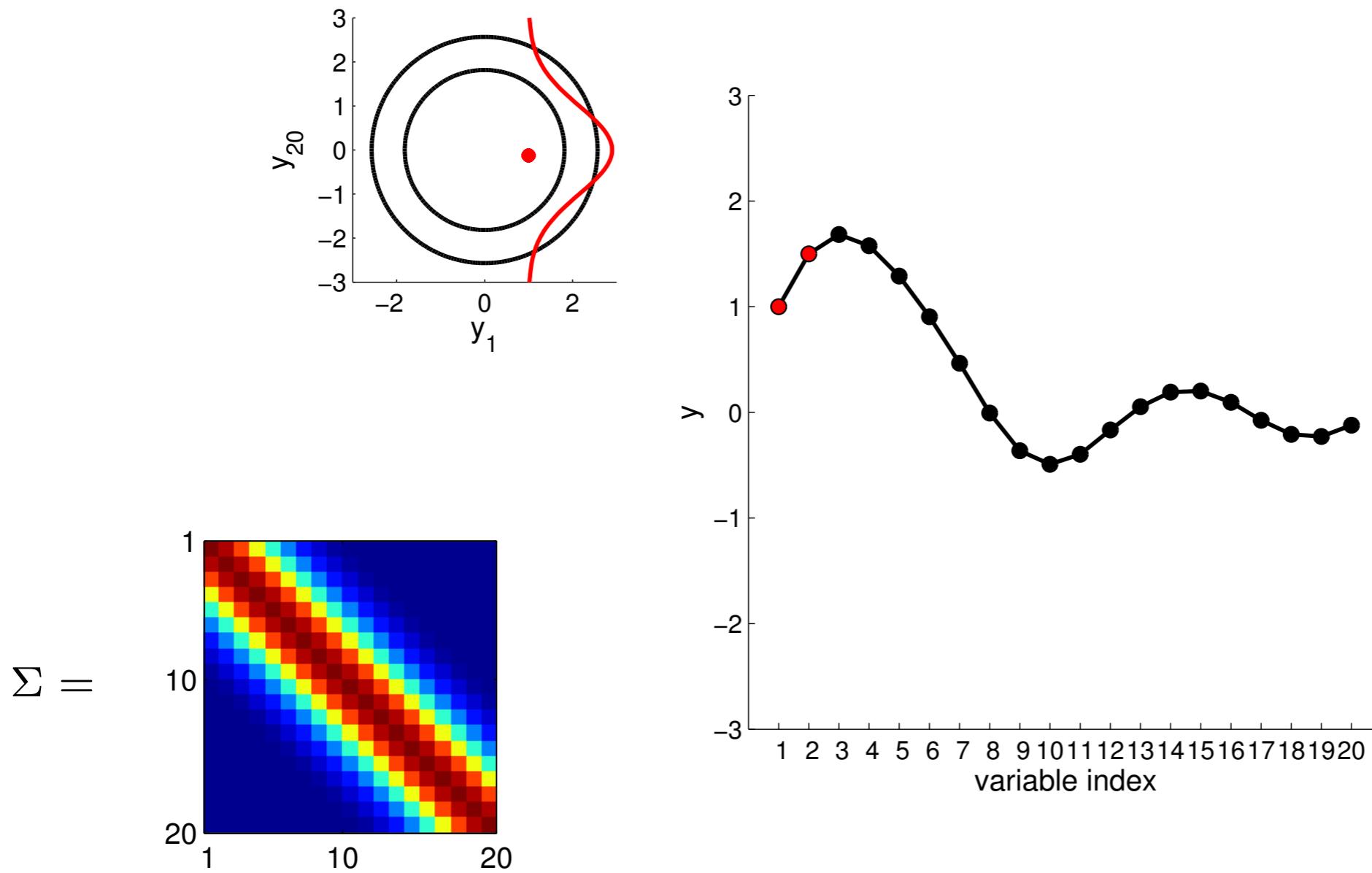
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



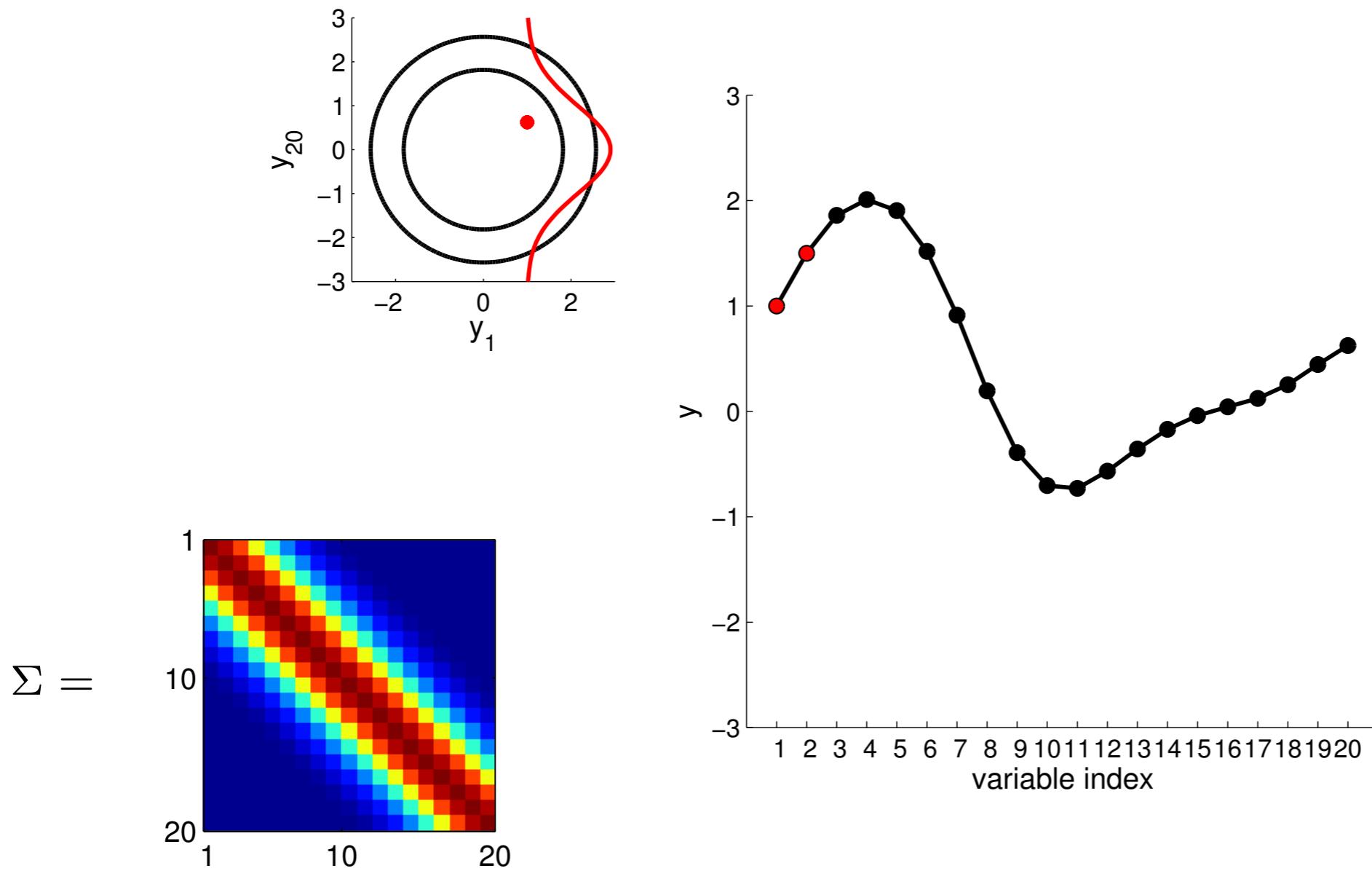
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



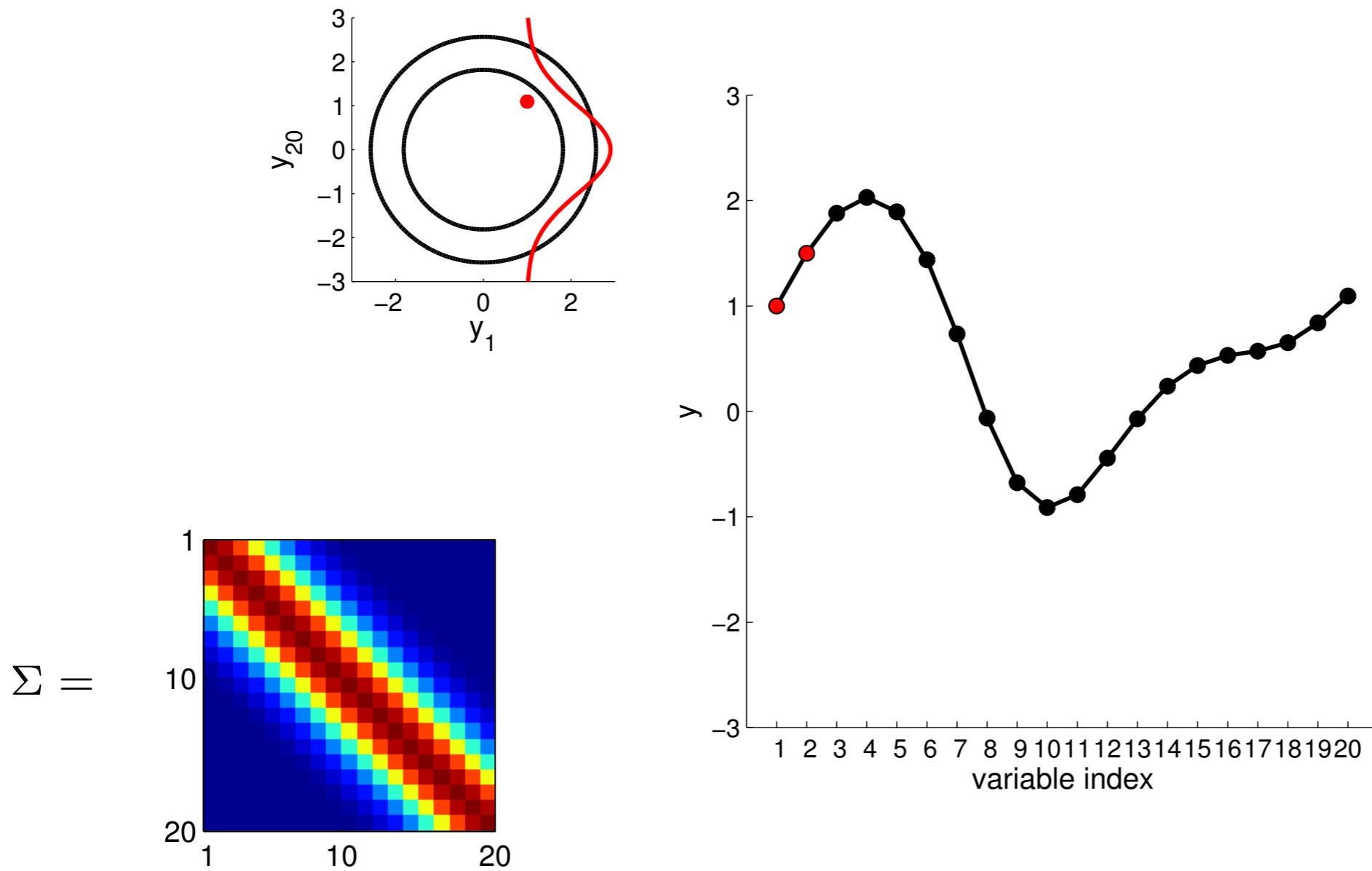
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



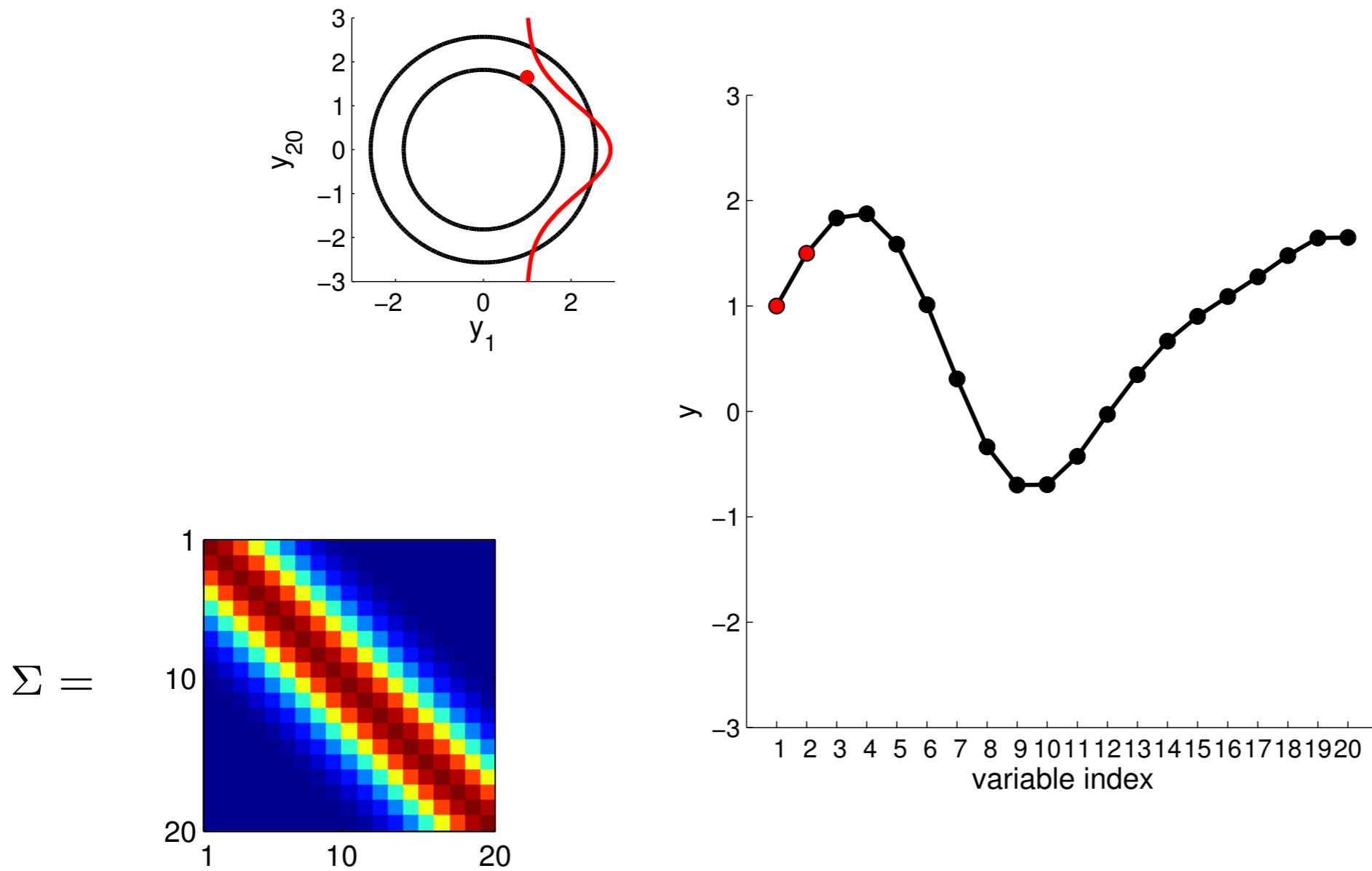
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



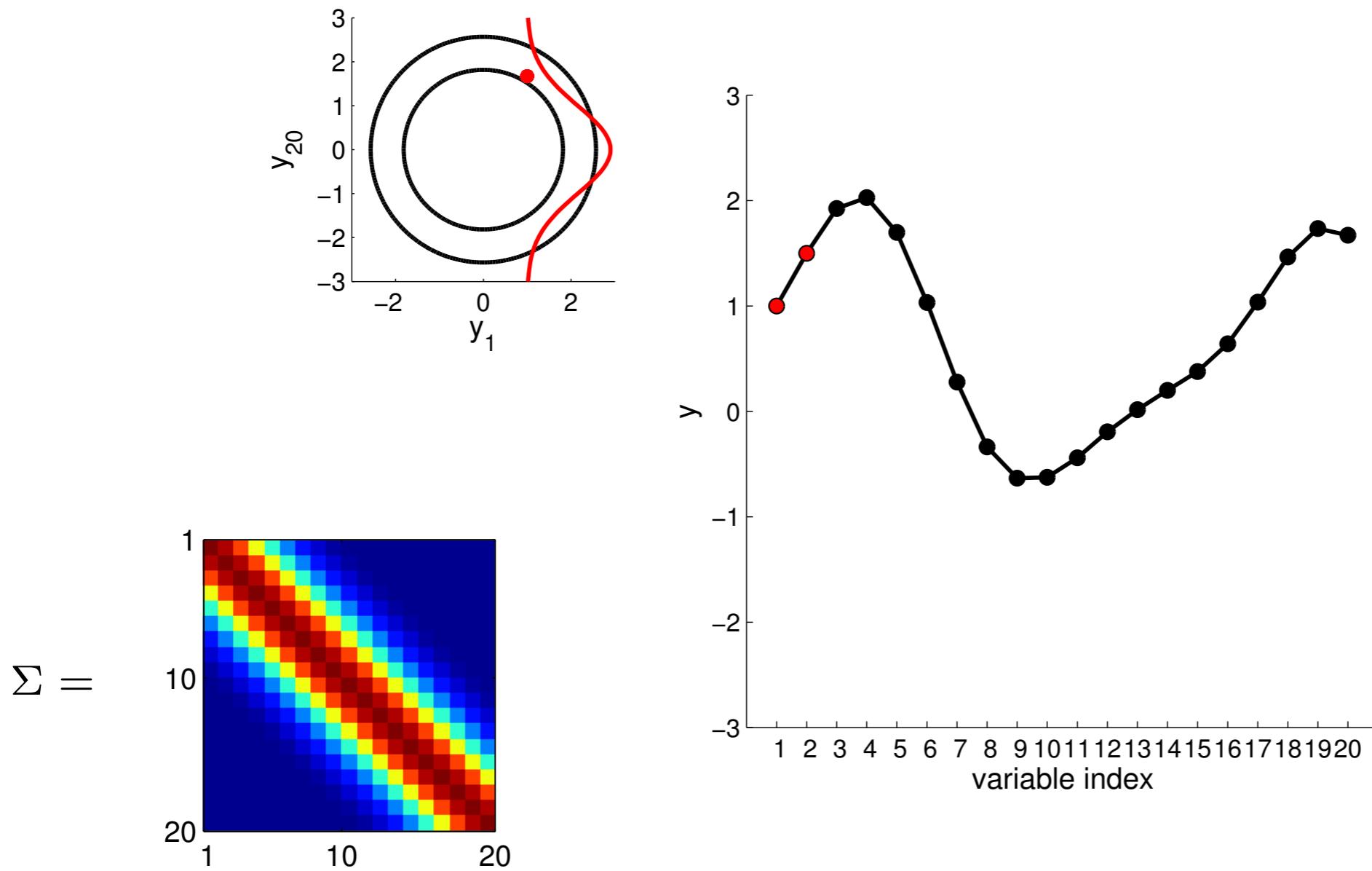
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



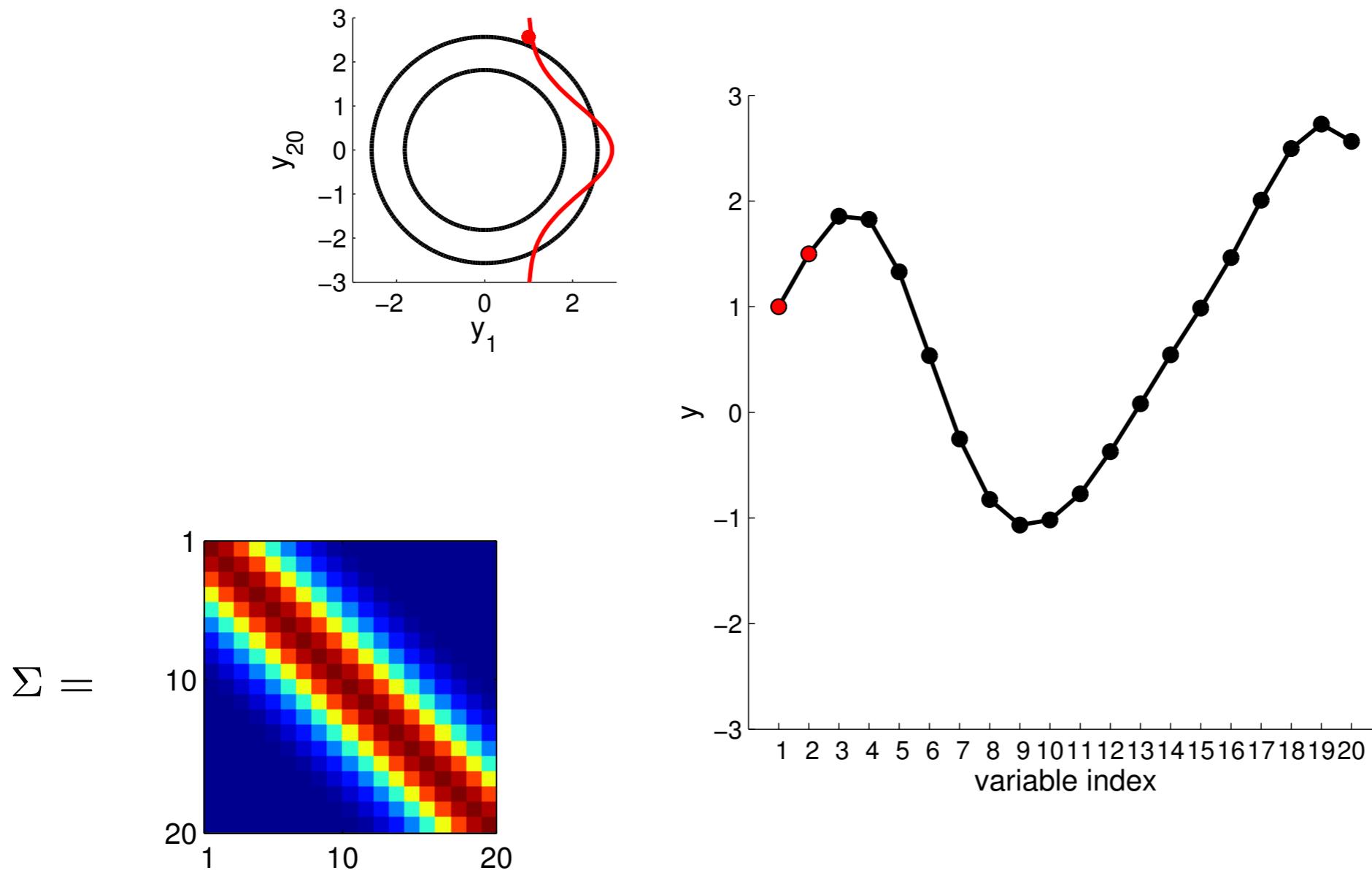
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



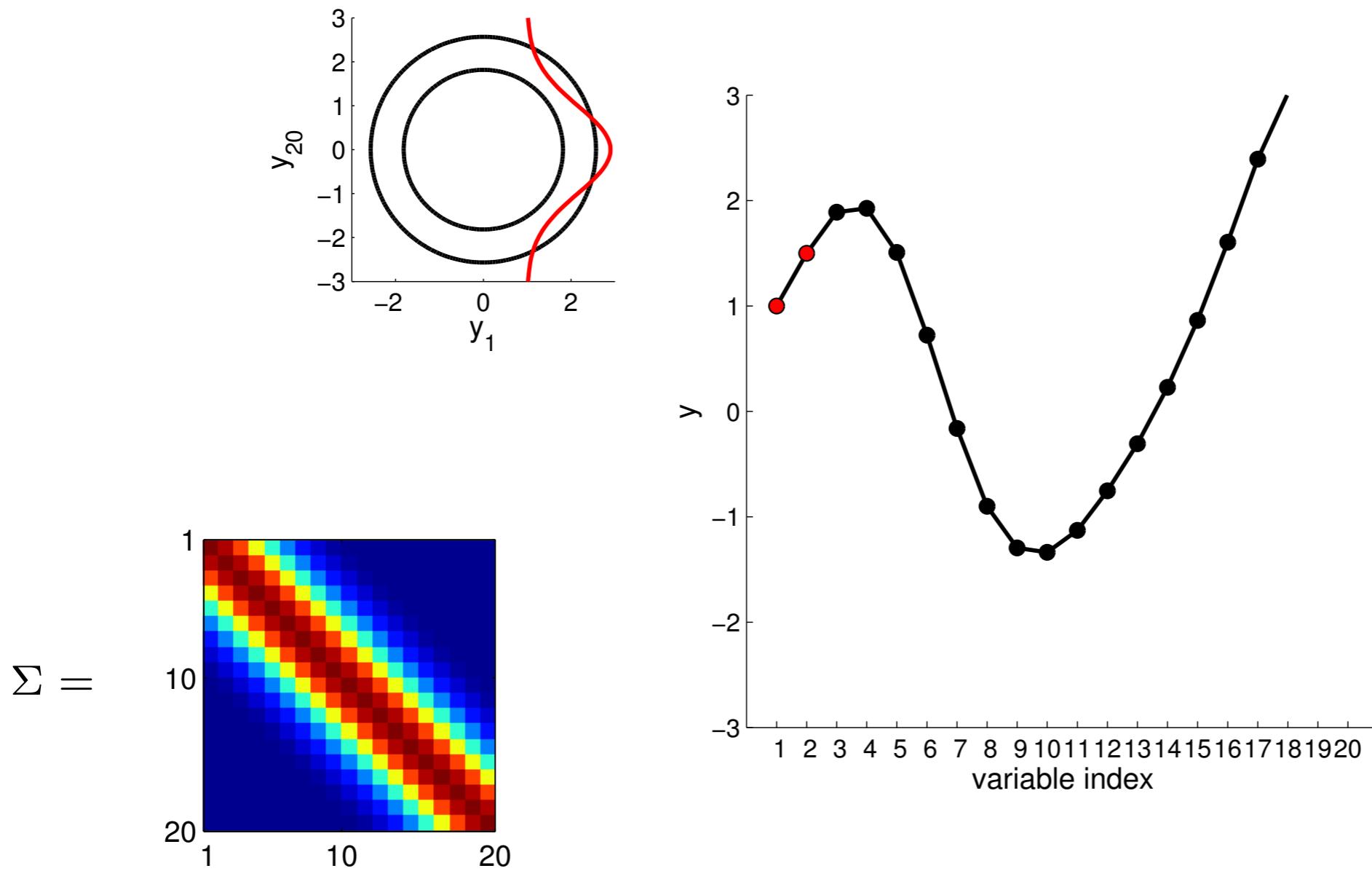
Conditioning on  $y_1$  and  $y_2$

# Special covariance matrix - conditioning



Conditioning on  $y_1$  and  $y_2$

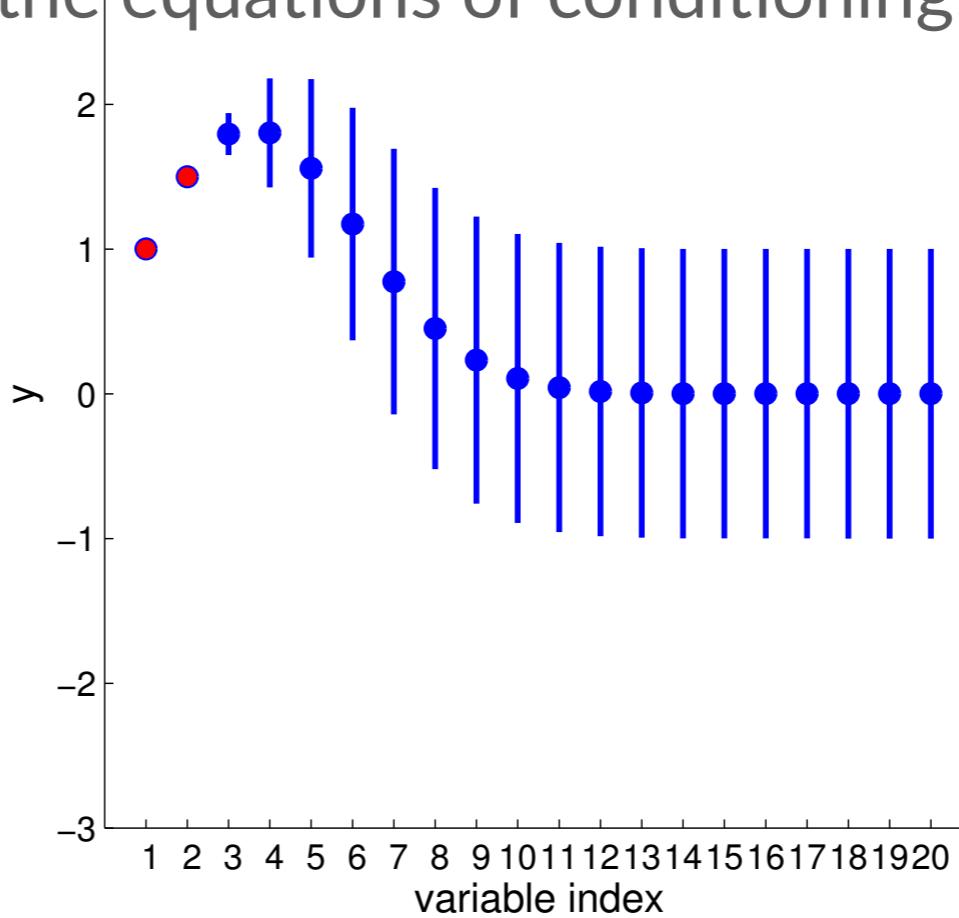
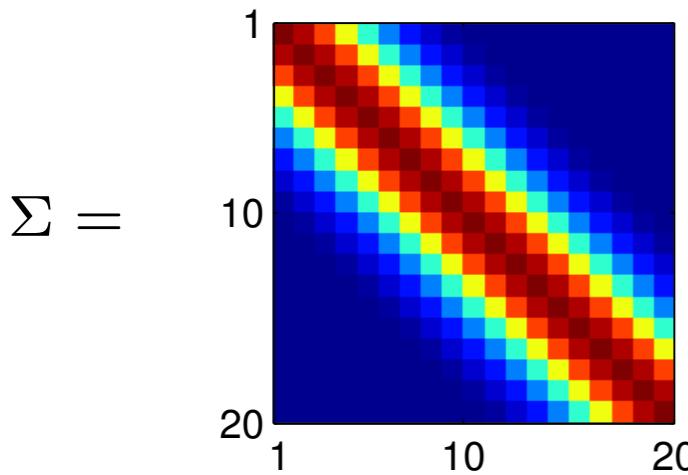
# Special covariance matrix - conditioning



Conditioning on  $y_1$  and  $y_2$

# Regression Using Gaussians

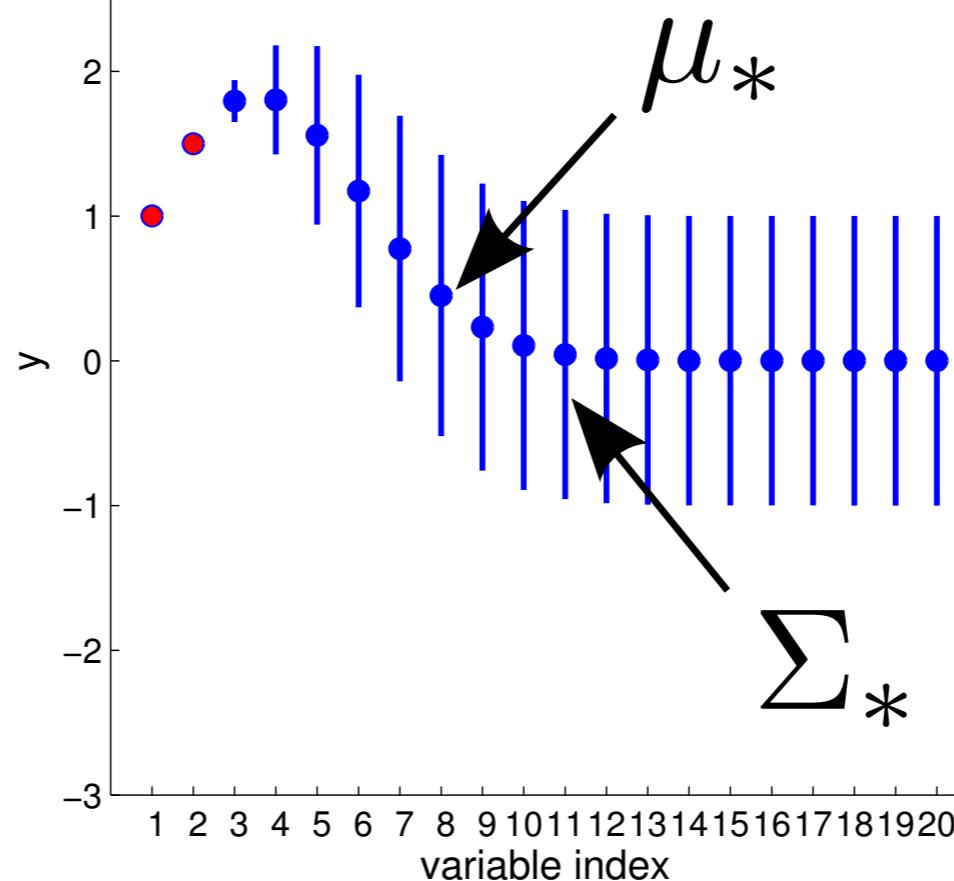
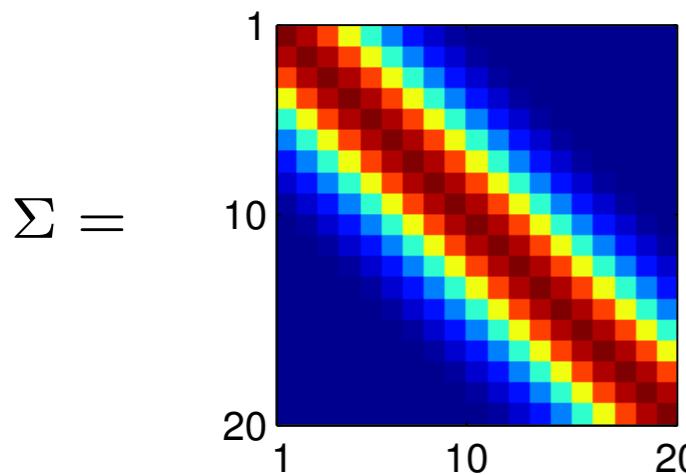
- If we average over the samples we can get mean and variance for each of the variables, conditioning on the observed red values! Exactly what we were looking for: [Regression with error bars](#).
- Actually we do not need to average! We will compute means and variances analytically, using the <sup>3</sup> equations of conditioning!



Conditioning on  $y_1$  and  $y_2$

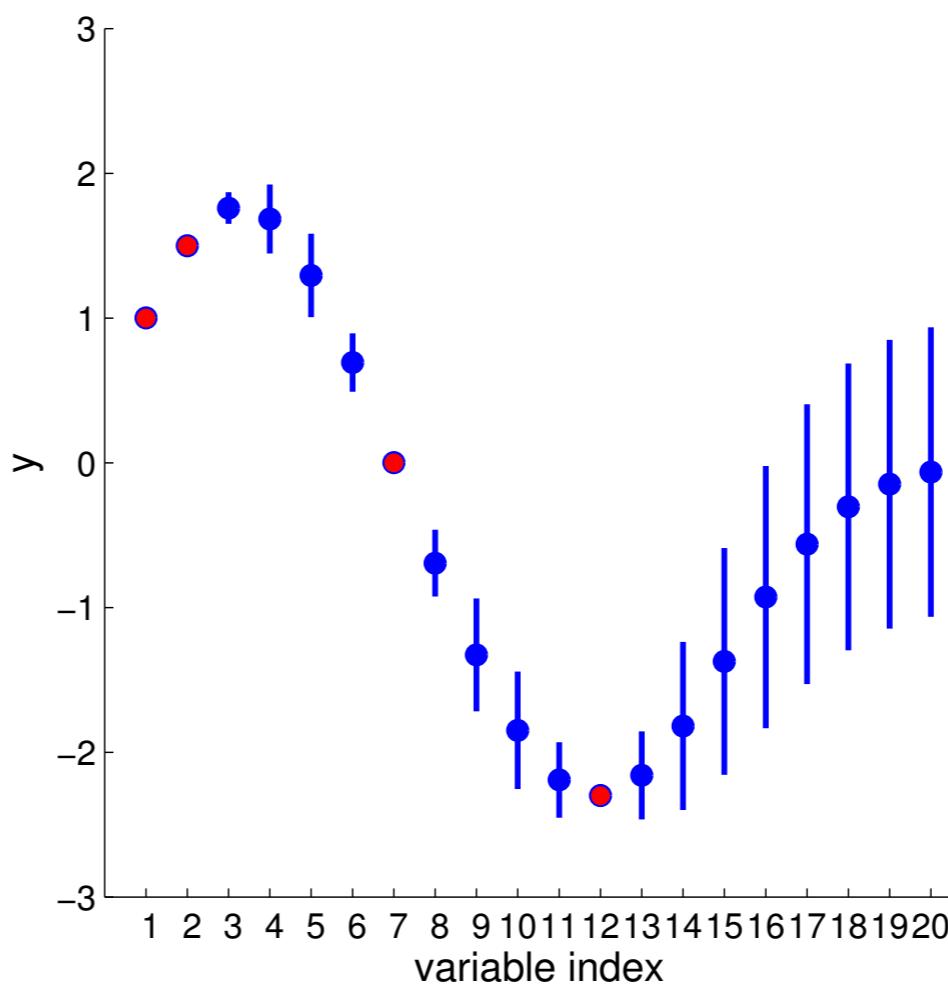
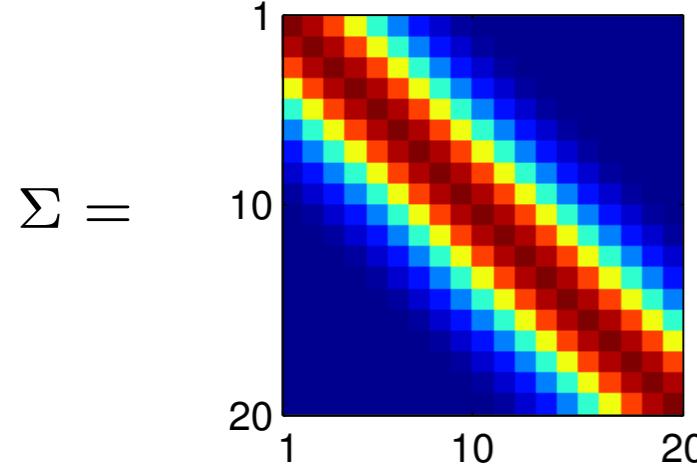
# Regression Using Gaussians

- If we average over the samples we can get mean and variance for each of the variables, conditioning on the observed red values! Exactly what we were looking for: [Regression with error bars](#).
- Actually we do not need to average! We will compute means and variances analytically, using the <sup>3</sup> equations of conditioning!



# Regression Using Gaussians

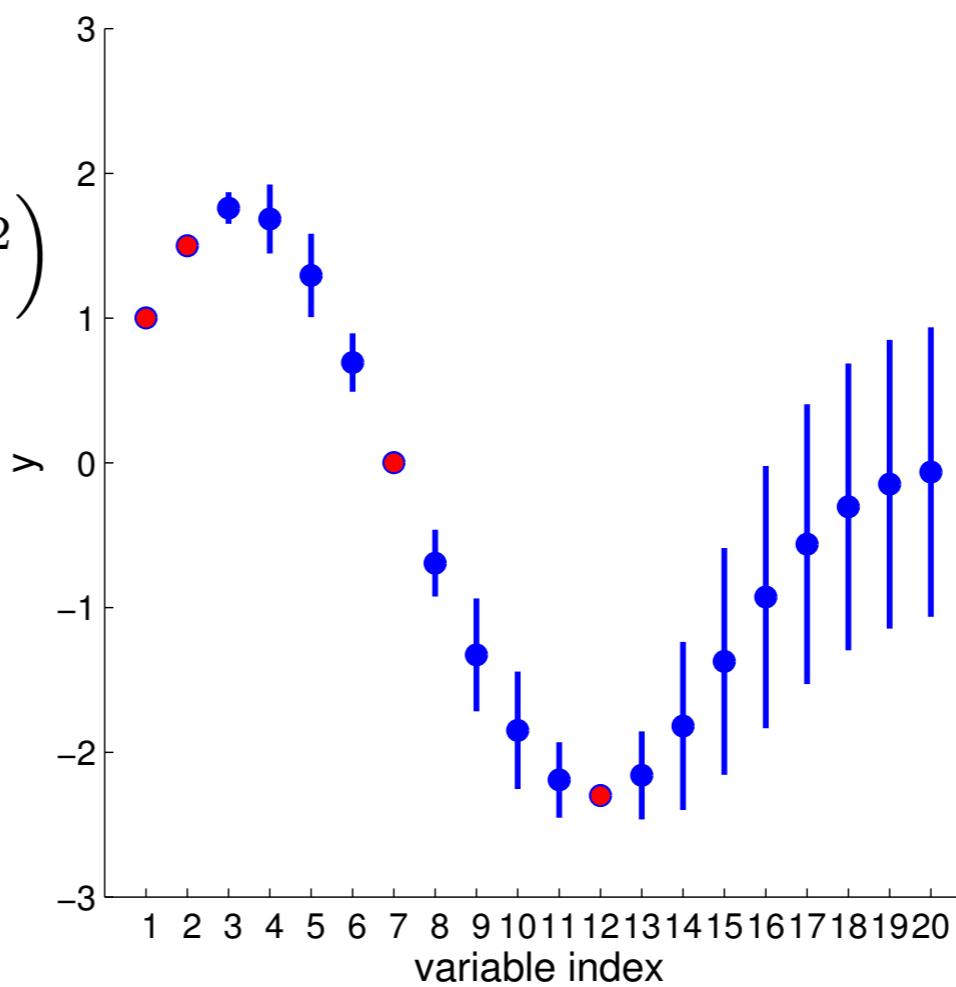
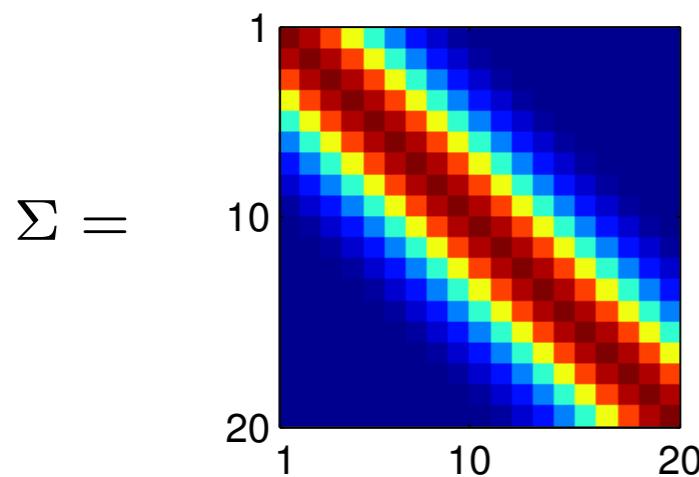
We can also condition on non-contiguous indices



# Regression Using Gaussians

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

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



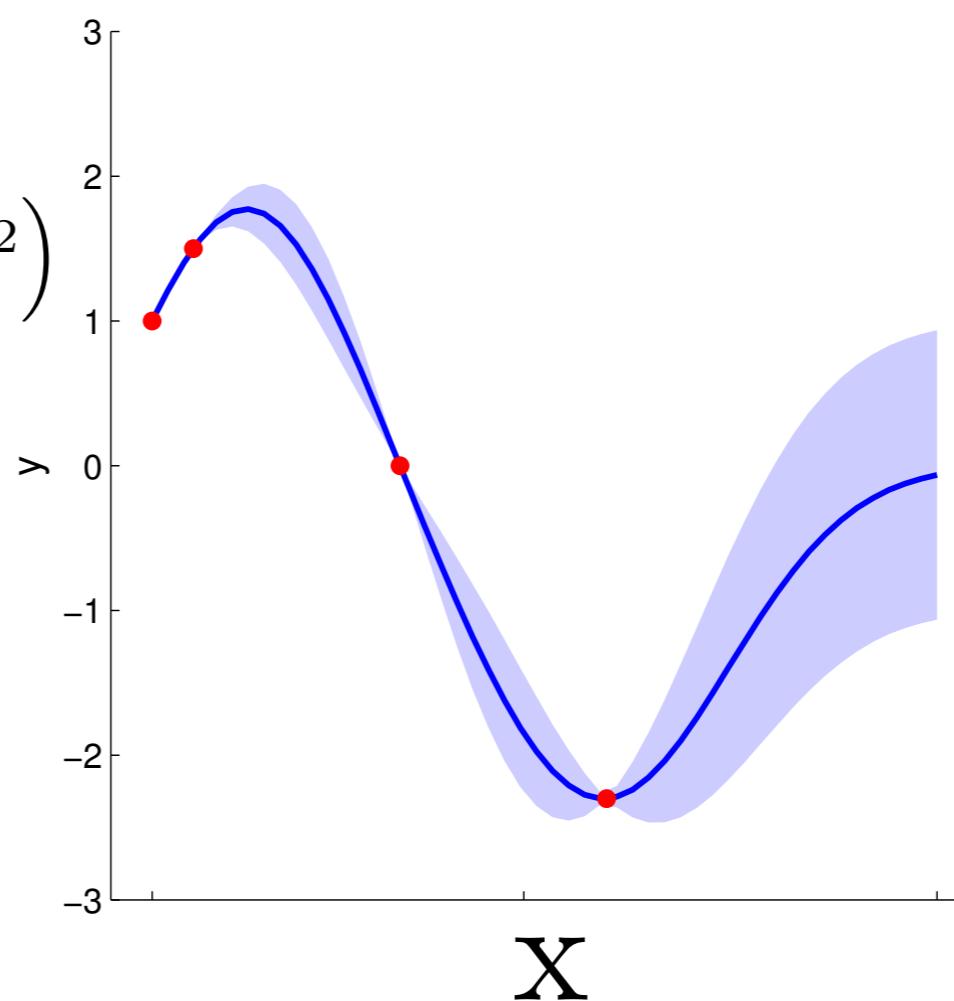
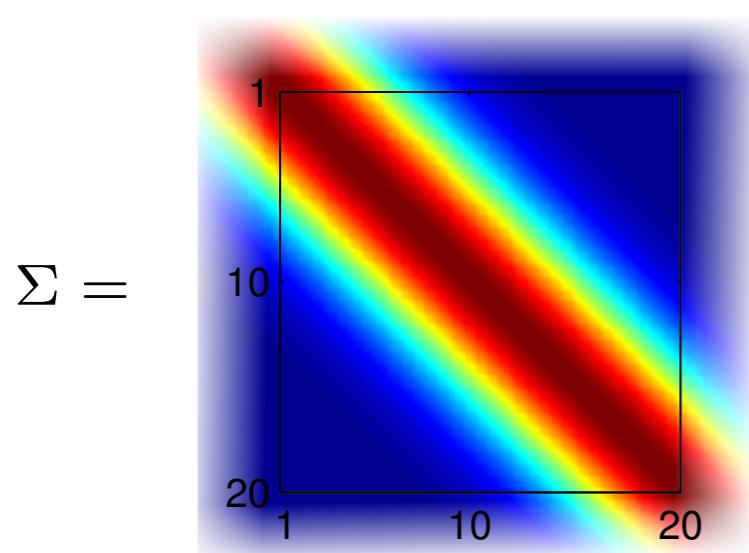
Q: Do  $x_1, x_2$  need to be integers?

# From multivariate Gaussian distributions to Gaussian Processes

GP: a multivariate Gaussian over an uncountably infinite number of variables with infinite mean vector and infinite times infinite covariance matrix

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

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

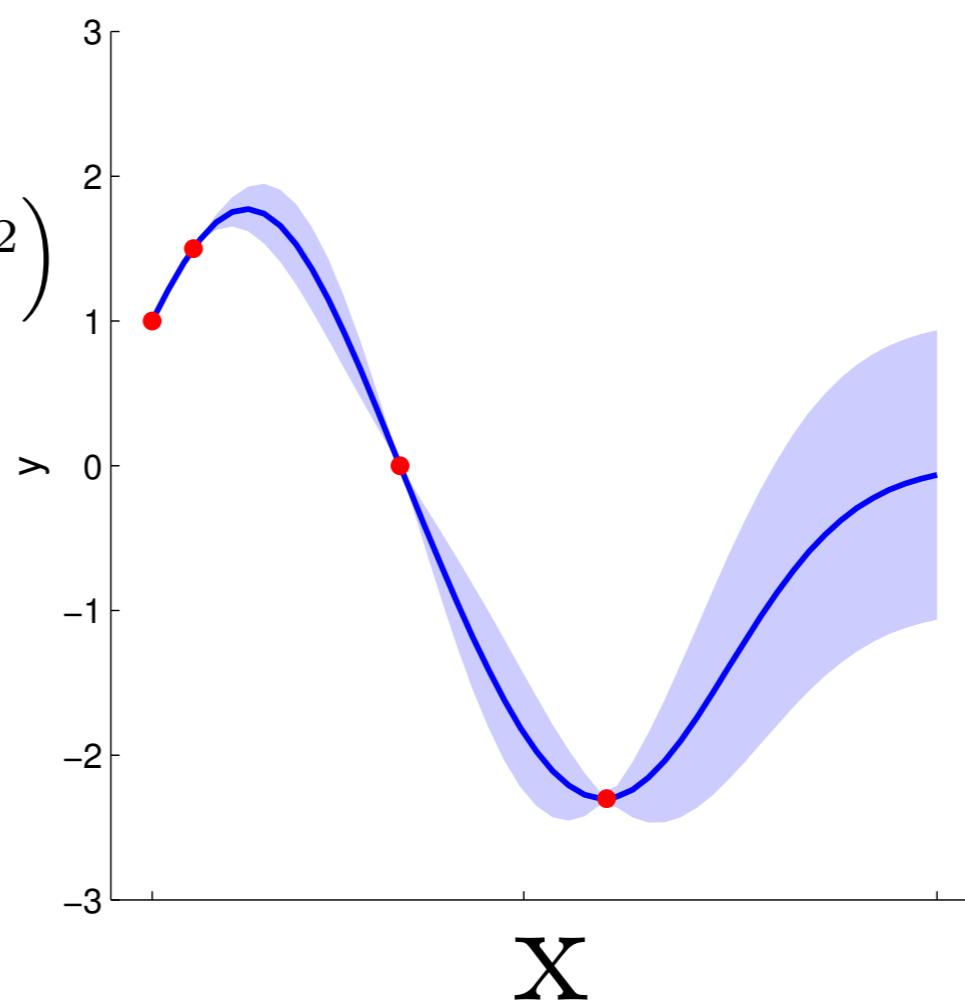
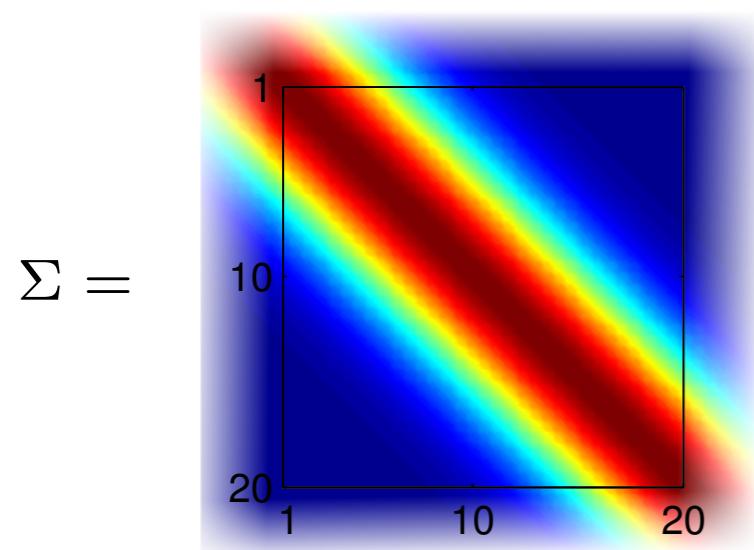


# Regression: probabilistic inference in function space

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

$$\Sigma(\mathbf{x}_1, \mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2) + I\sigma_y^2$$

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



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

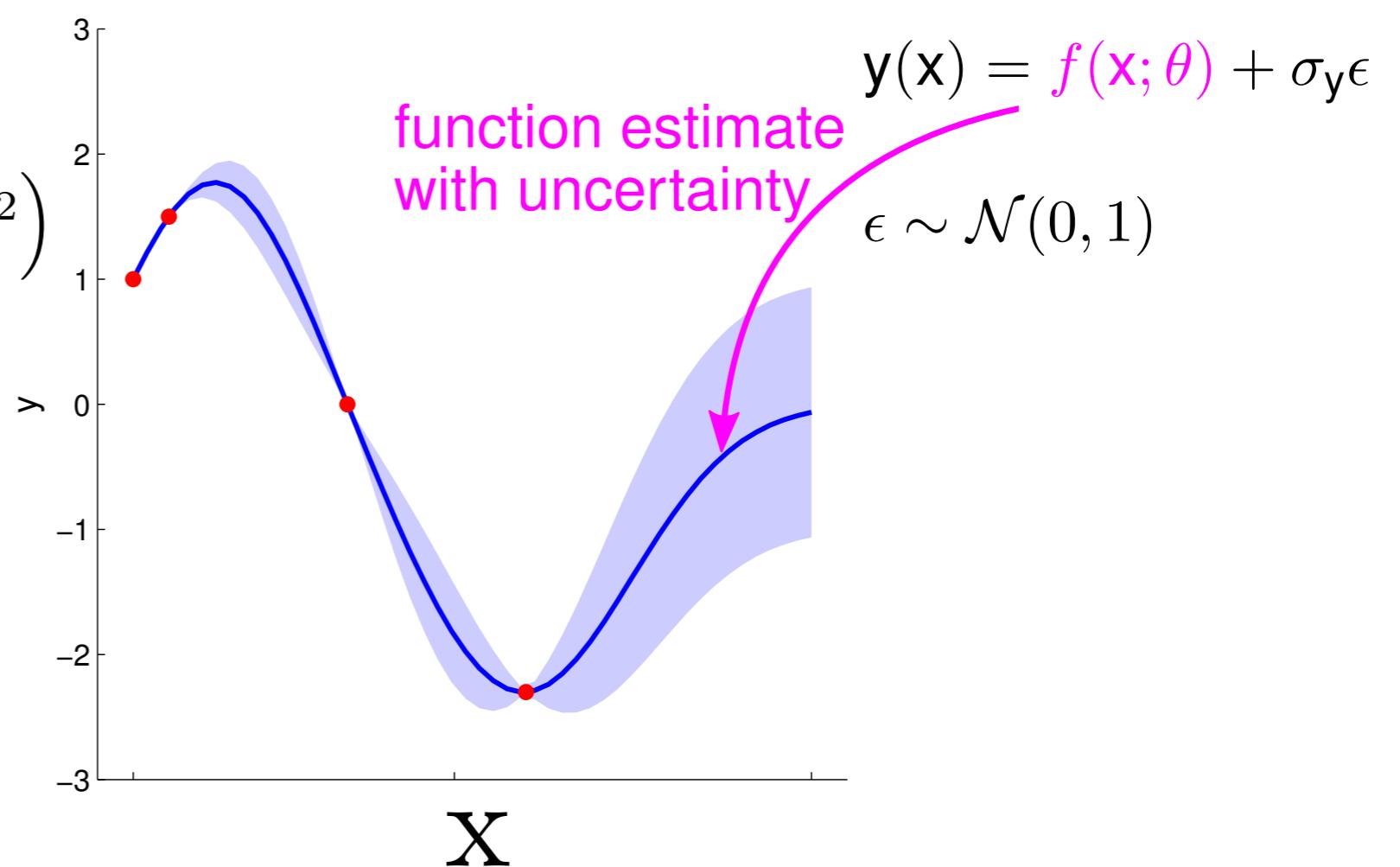
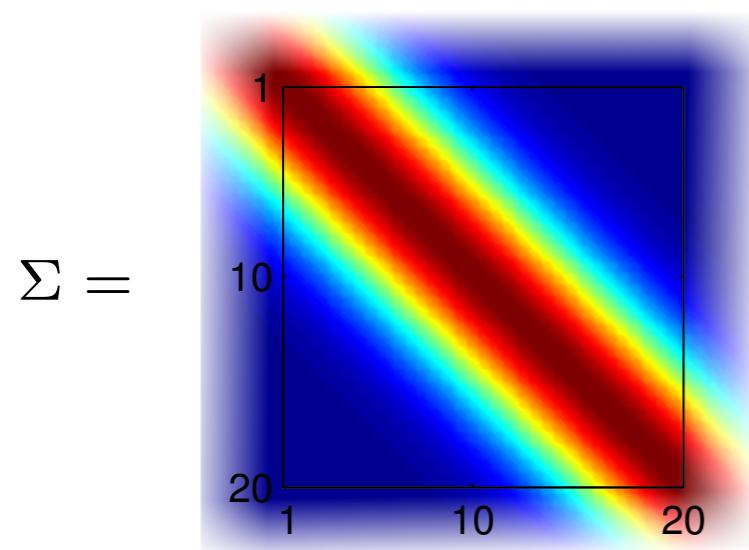
$$\epsilon \sim \mathcal{N}(0, 1)$$

# Regression: probabilistic inference in function space

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

$$\Sigma(\mathbf{x}_1, \mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2) + I\sigma_y^2$$

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

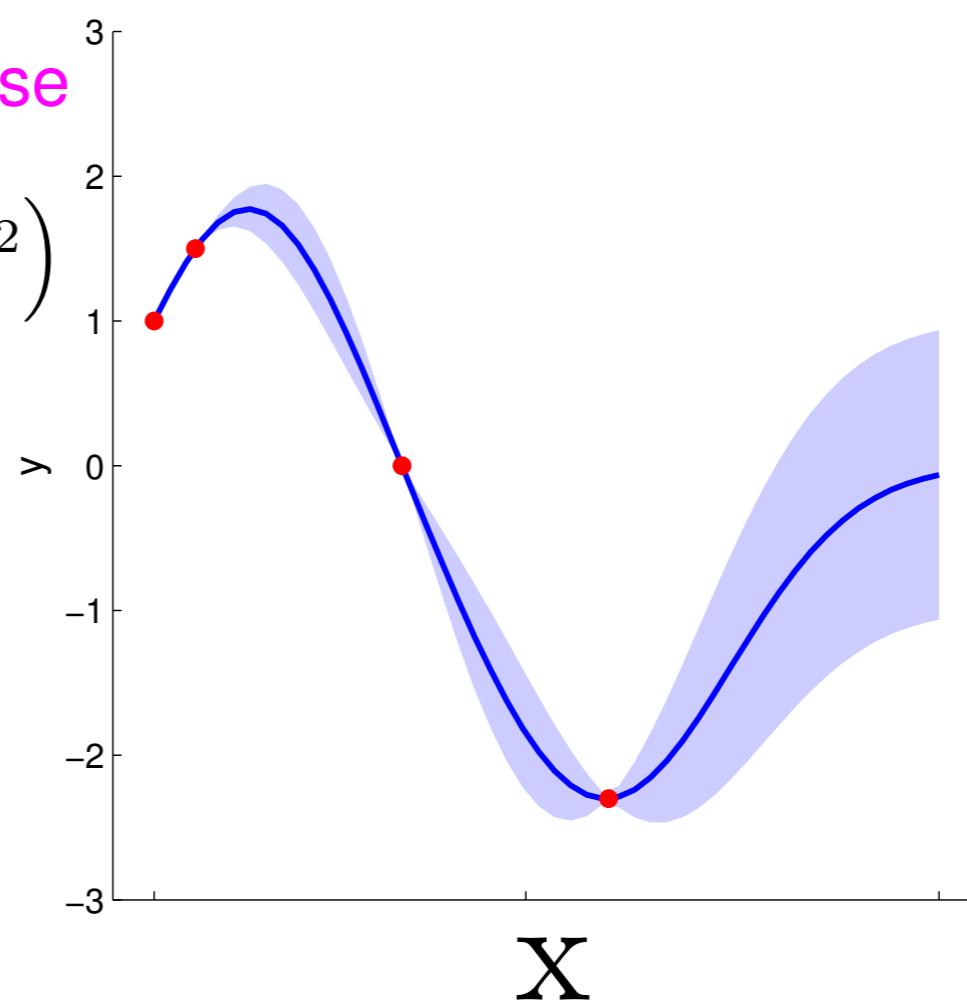
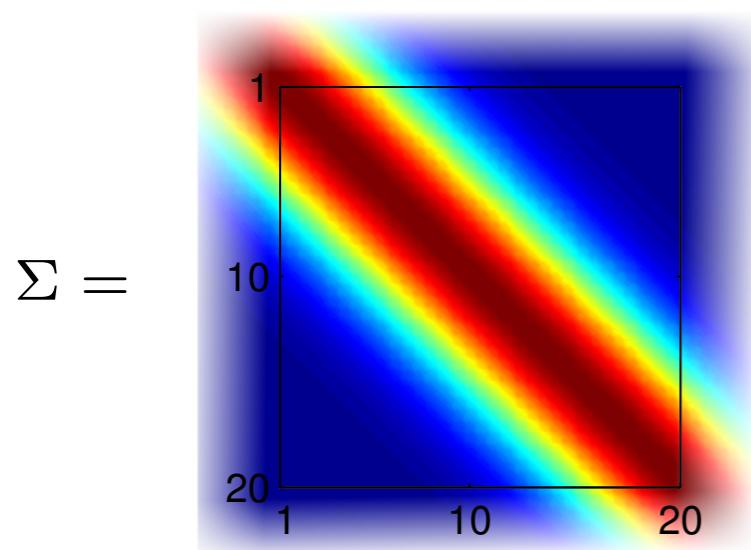


# Regression: probabilistic inference in function space

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

$$\Sigma(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{K}(\mathbf{x}_1, \mathbf{x}_2) + \textcolor{magenta}{I\sigma_y^2} \leftarrow \text{noise}$$

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



$$y(\mathbf{x}) = f(\mathbf{x}; \theta) + \textcolor{magenta}{\sigma_y \epsilon}$$

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

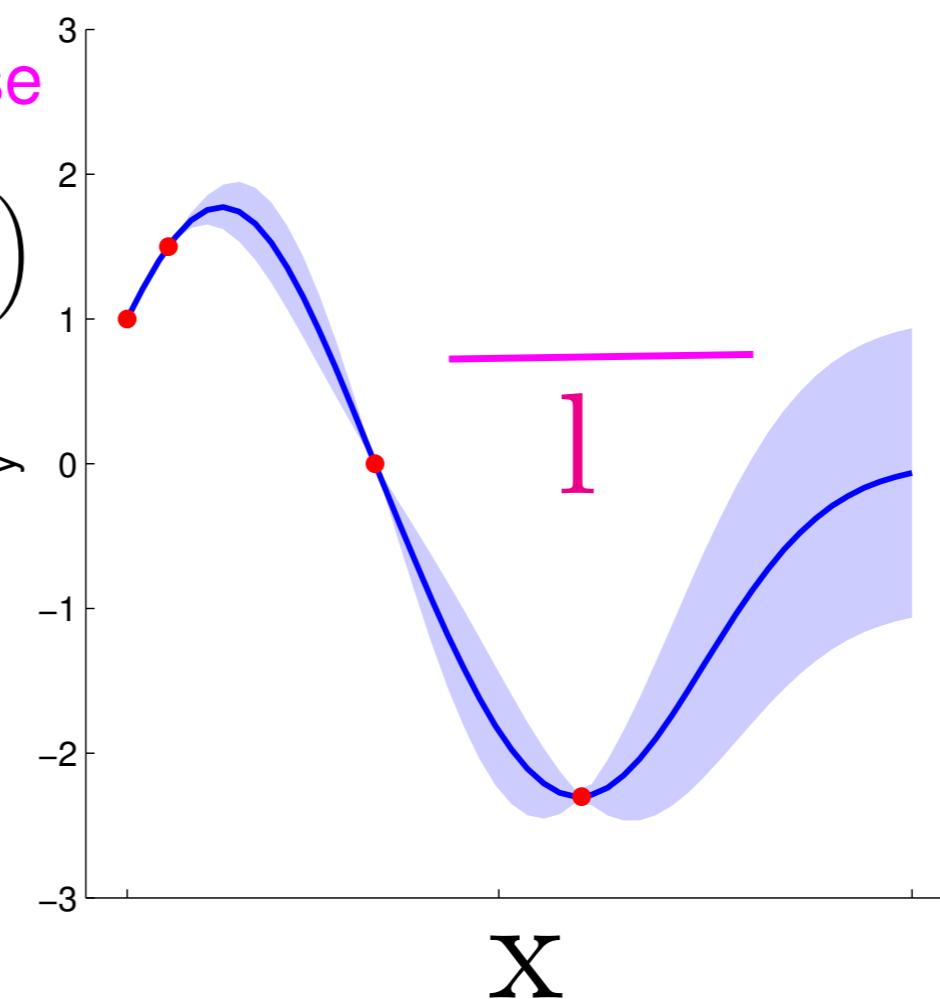
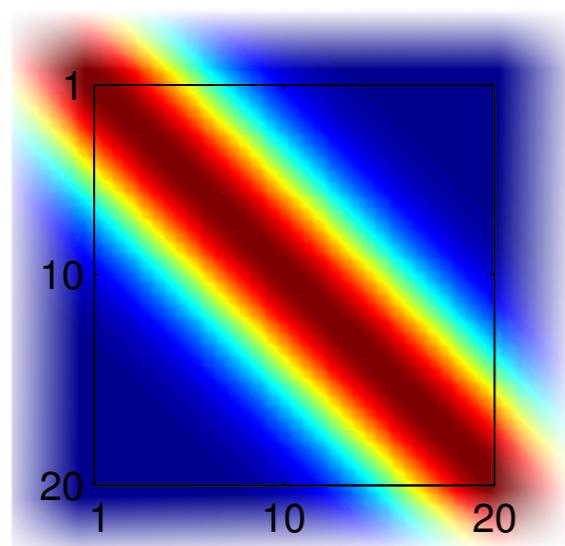
# Regression: probabilistic inference in function space

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

$$\Sigma(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{K}(\mathbf{x}_1, \mathbf{x}_2) + I\sigma_y^2 \quad \text{← noise}$$

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

$$\Sigma =$$



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

How fast the correlations fall off..

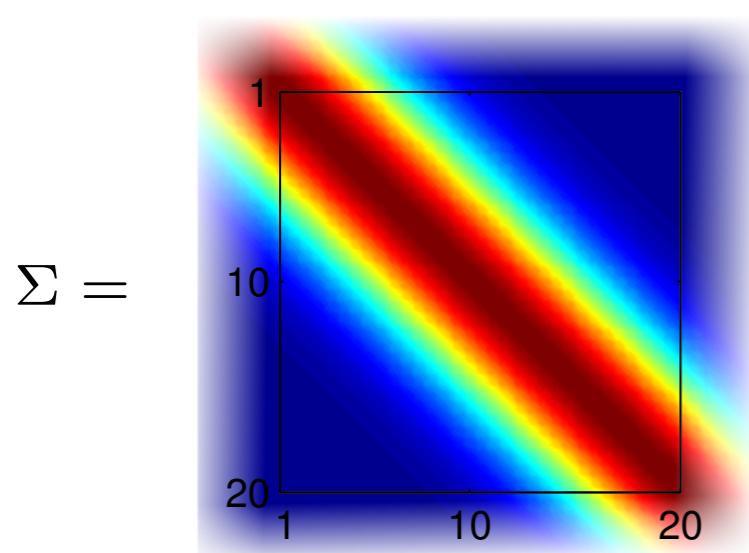
# Regression: probabilistic inference in function space

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

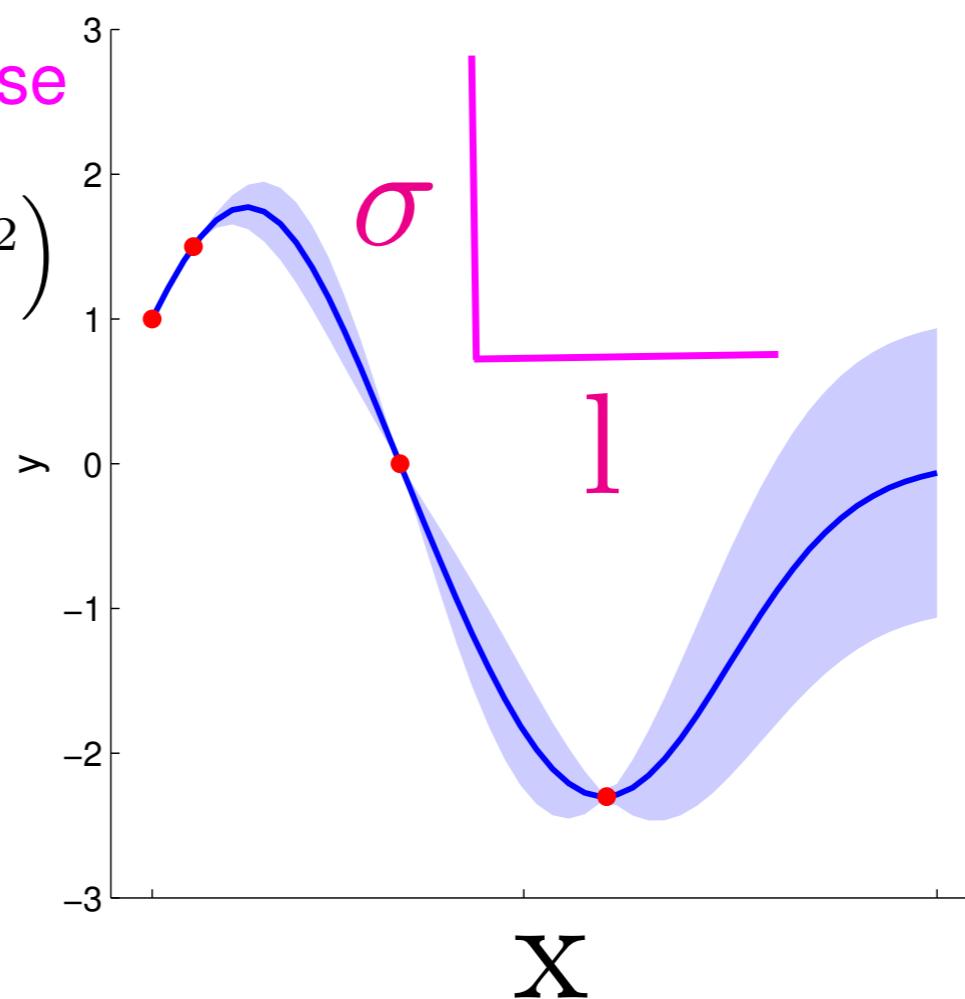
$$\Sigma(\mathbf{x}_1, \mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2) + I\sigma_y^2 \quad \text{noise}$$

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

vertical-scale    horizontal-scale



$$\Sigma =$$



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

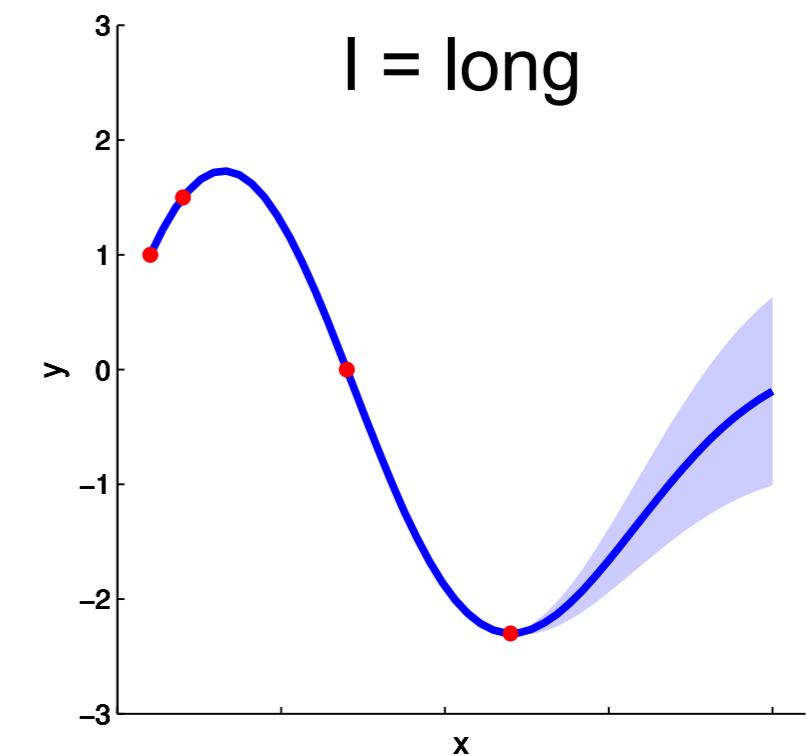
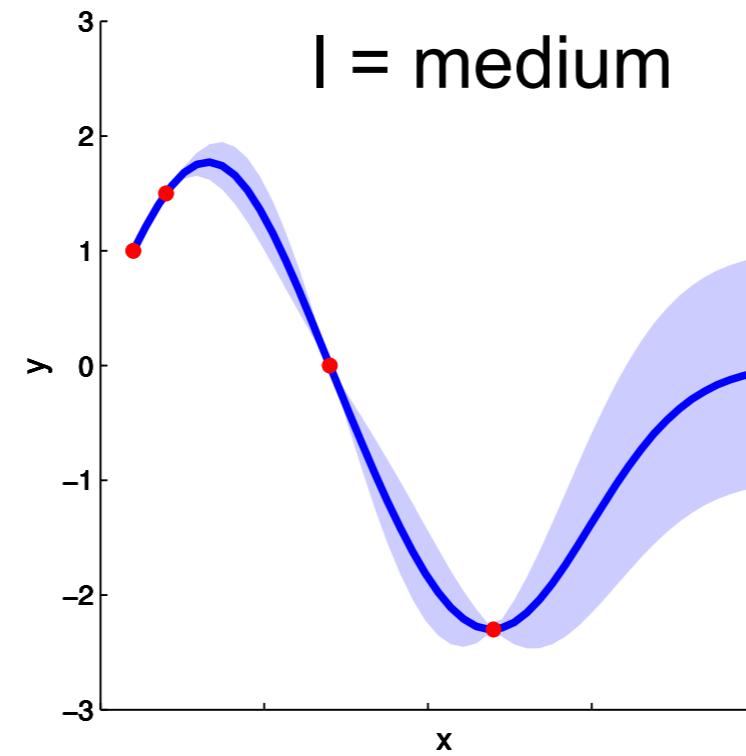
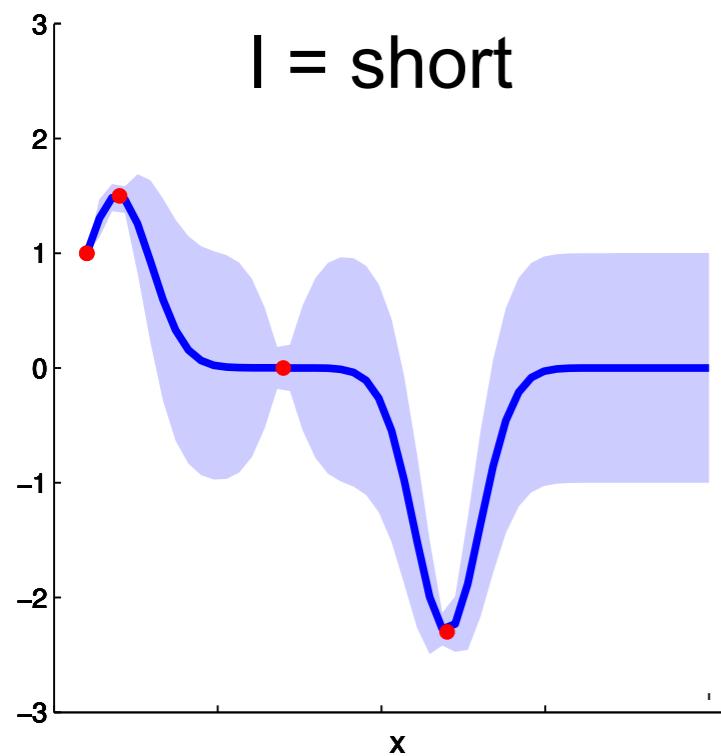
How fast the correlations fall off..

# What effect do the hyper-parameters have?

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

Hyper-parameters have a strong effect

- $l$  controls the horizontal scaling
- $\sigma^2$  controls the vertical scale of the data



# Gaussian Processes: Definition

Gaussian process = generalization of multivariate Gaussian distribution to infinitely many variables

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

# Gaussian Processes: Definition

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

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

A Gaussian distribution is fully specified by a mean vector,  $\mu$ , and a covariance matrix  $\Sigma$ :

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

# Gaussian Processes: Definition

Gaussian process = generalization of multivariate Gaussian distribution to infinitely many variables

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

A Gaussian distribution is fully specified by a mean vector,  $\mu$ , and a covariance matrix  $\Sigma$ :

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

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

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

# Mathematical justification

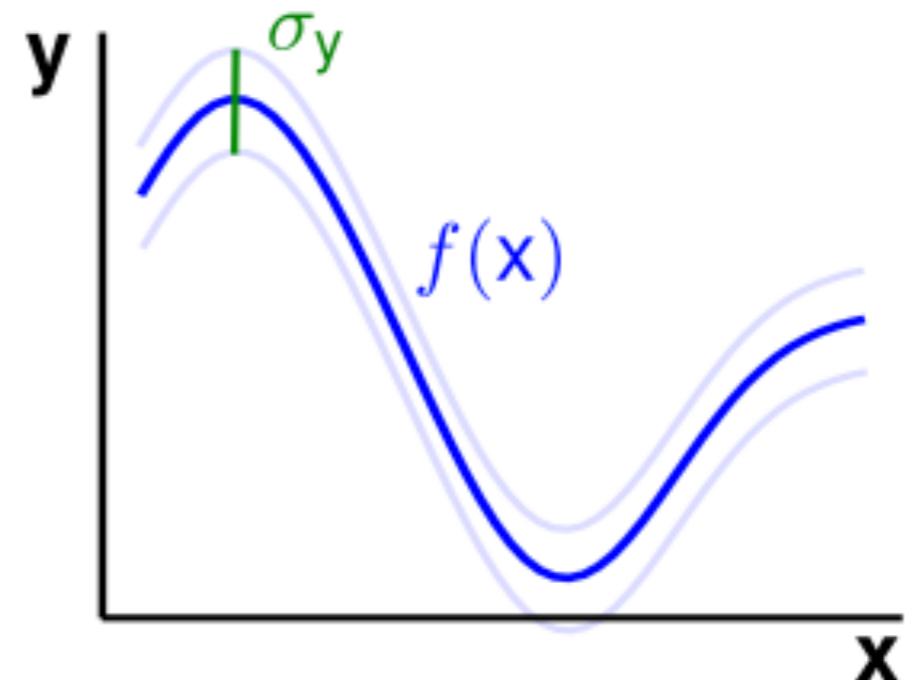
Q: What is a formal justification for how we are using GPs for regression?

# Mathematical justification

Q: What is a formal justification for how we are using GPs for regression?

Generative model (like non-linear regression)

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



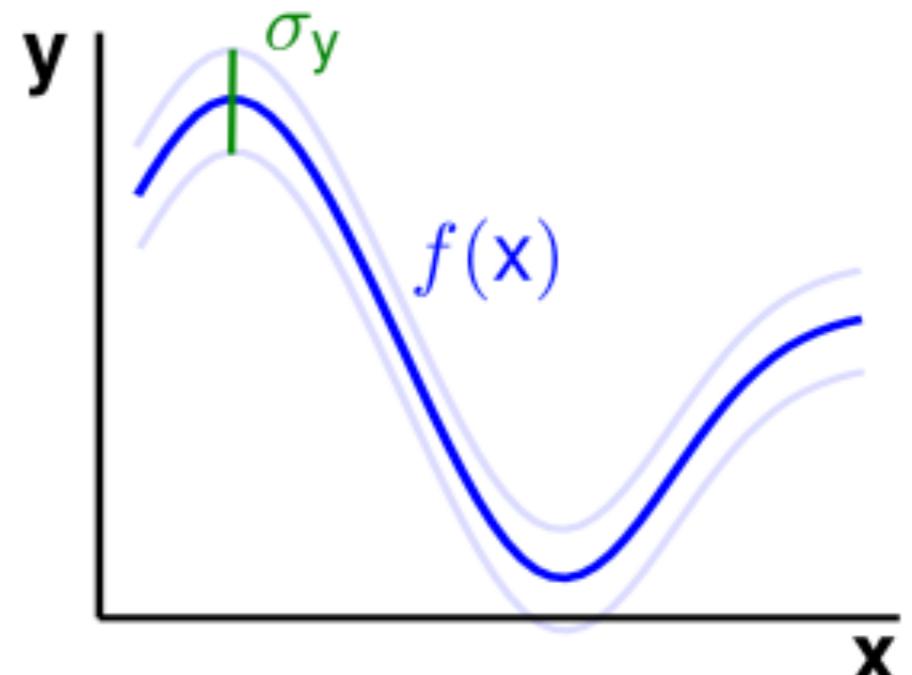
# Mathematical justification

Q: What is a formal justification for how we are using GPs for regression?

Generative model (like non-linear regression)

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

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



# Mathematical justification

Q: What is a formal justification for how we are using GPs for regression?

Generative model (like non-linear regression)

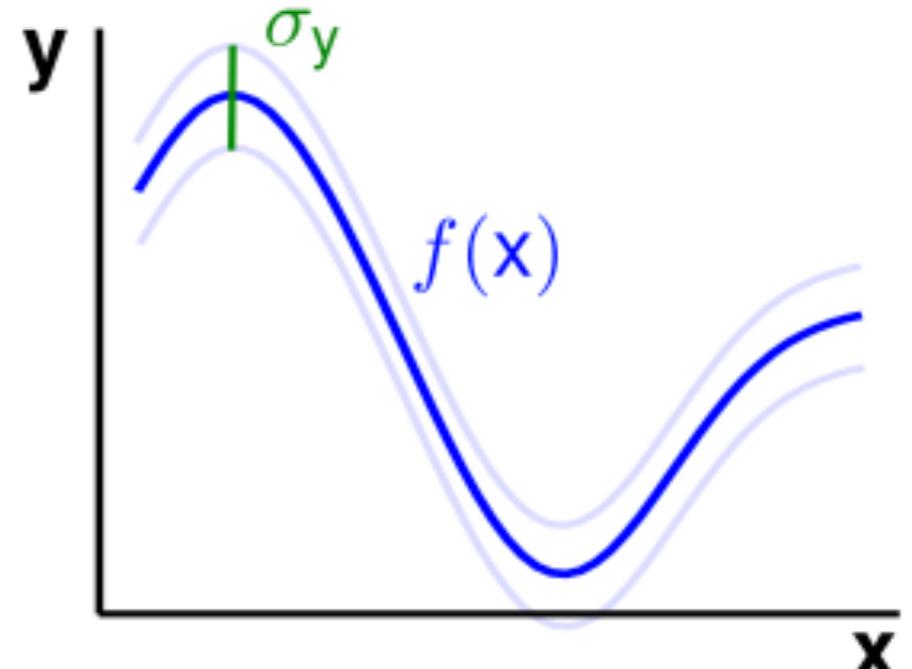
$$y(x) = f(x) + \epsilon \sigma_y$$

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

place GP prior over the non-linear function

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

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



(smoothly wiggling functions expected)

# Mathematical justification

Q: What is a formal justification for how we are using GPs for regression?

Generative model (like non-linear regression)

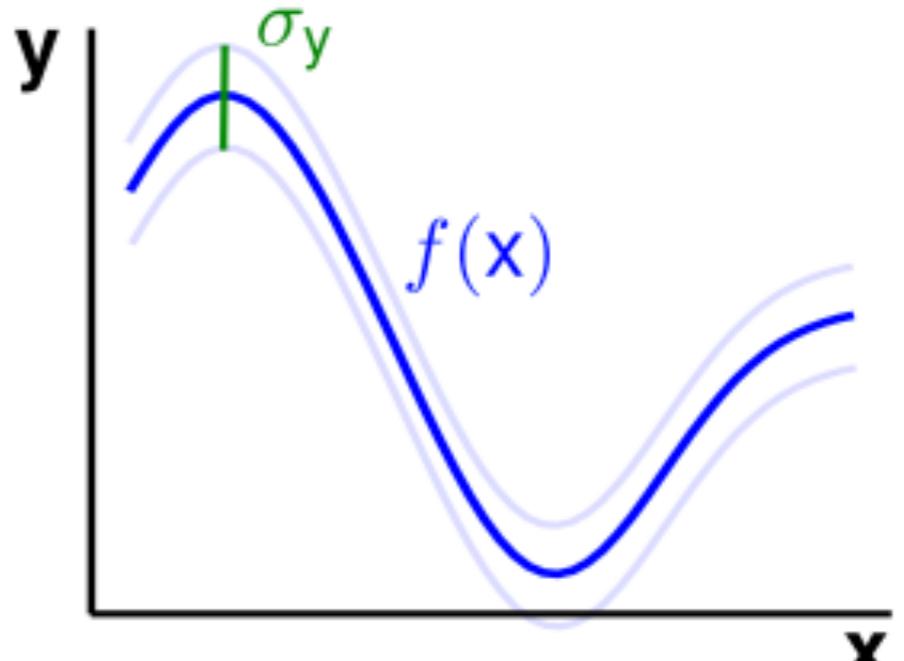
$$y(x) = f(x) + \epsilon \sigma_y$$

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

place GP prior over the non-linear function

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

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



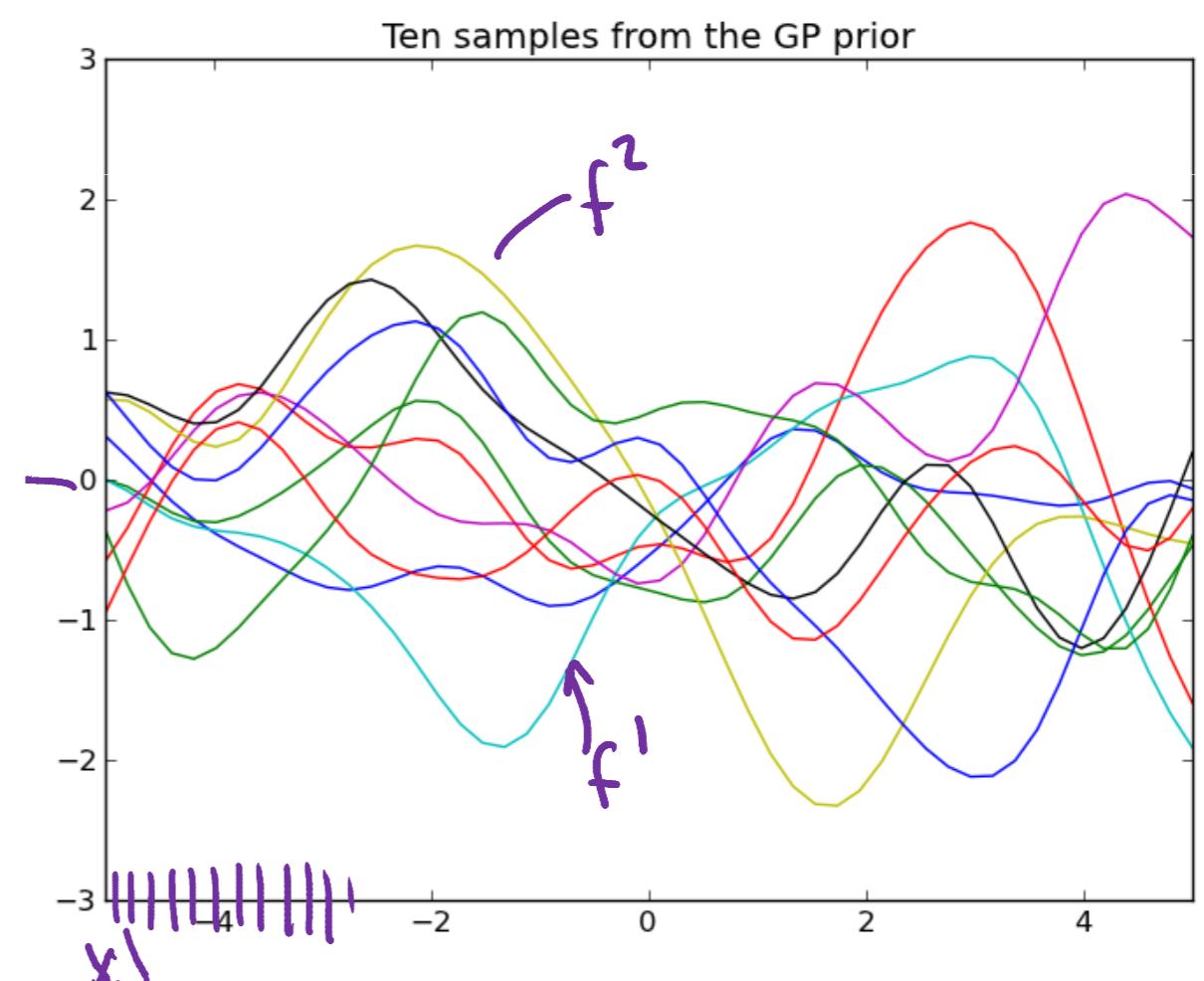
(smoothly wiggling functions expected)

since the sum of two Gaussians is a Gaussian, the model induces a GP over  $y(x)$ :

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

# Sampling from the GP prior

1. Create  $x_{1:N}$  (the N points where we will evaluate our function)
2. Compute mean  $\mu = 0_N$  and covariance matrix  $K$
3. Compute the Chelsky decomposition  $K = LL^T$
4.  $f^i \sim \mathcal{N}(\mu, K)$   
 $\sim L\mathcal{N}(0, I)$



# Sampling from the GP prior

```
from __future__ import division  
import numpy as np  
import matplotlib.pyplot as pl
```

```
def kernel(a, b):  
    """ GP squared exponential kernel """  
    sqdist = np.sum(a**2, 1).reshape(-1, 1) + np.sum(b**2, 1) - 2 * np.dot(a, b.T)  
    return np.exp(-.5 * sqdist)
```

$n = 50$  ↴ # number of test points.

$X_{\text{test}} = \text{linspace}(-5, 5, n).$  reshape(-1, 1) # Test points.

$K_ = \text{kernel}(X_{\text{test}}, X_{\text{test}})$  ↴ # Kernel at test points.

# draw samples from the prior at our test points.

$L = \text{np.linalg.cholesky}(K_ + 1e-6 * \text{np.eye}(n))$

$f_{\text{prior}} = \text{np.dot}(L, \text{np.random.normal}(\text{size}=(n, 10)))$  ↴  $\mathcal{N}(0, I)$

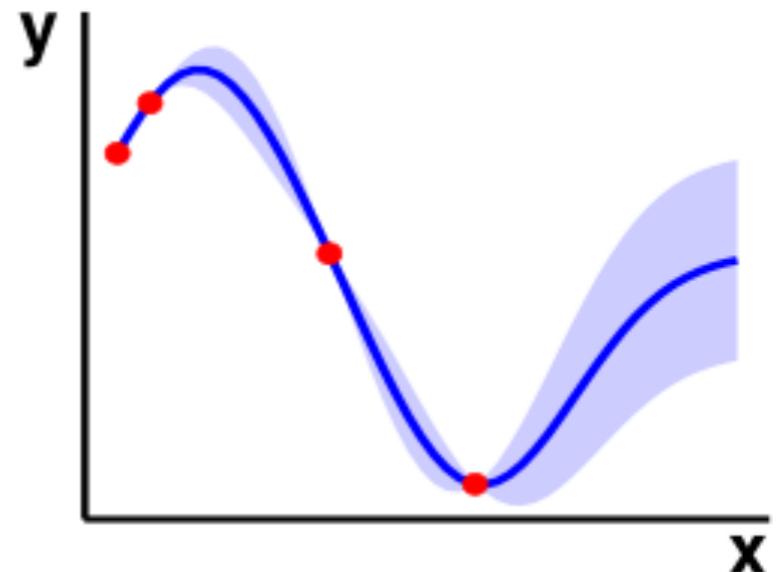
pl.plot(Xtest, f\_prior) ↴

# Gaussian Processes for regression

# Gaussian Processes for regression

**Reminder:** the conditioning equations for multivariate Gaussian distributions

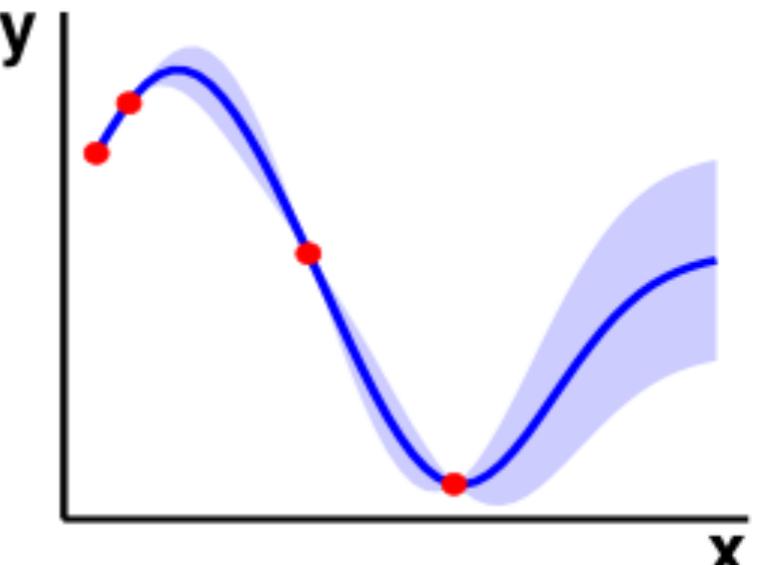
$$p(y_1 | y_2) = \frac{p(y_1, y_2)}{p(y_2)}$$



# Gaussian Processes for regression

**Reminder:** the conditioning equations for multivariate Gaussian distributions

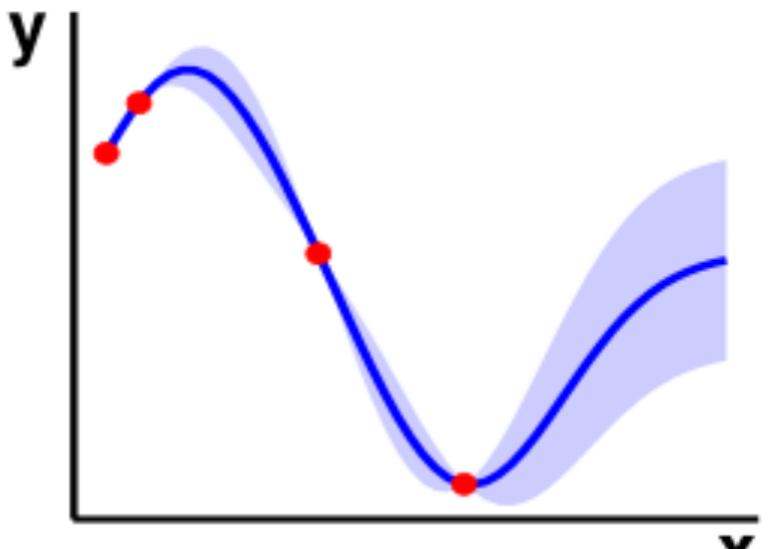
$$p(y_1, y_2) = \mathcal{N} \left( \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \right)$$
$$p(y_1 | y_2) = \frac{p(y_1, y_2)}{p(y_2)}$$



# Gaussian Processes for regression

**Reminder:** the conditioning equations for multivariate Gaussian distributions

$$p(y_1, y_2) = \mathcal{N} \left( \begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \right)$$
$$p(y_1 | y_2) = \frac{p(y_1, y_2)}{p(y_2)} \quad p(y_2) = \mathcal{N}(b, C)$$

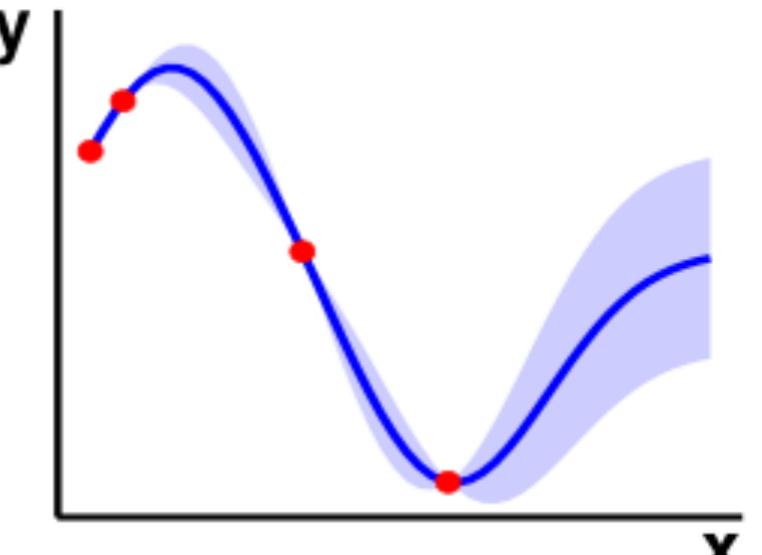


# Gaussian Processes for regression

**Reminder:** the conditioning equations for multivariate Gaussian distributions

$$p(y_1, y_2) = \mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right)$$
$$p(y_1 | y_2) = \frac{p(y_1, y_2)}{p(y_2)} \quad p(y_2) = \mathcal{N}(b, C)$$

$$p(y_1 | y_2) = \mathcal{N}\left(a + BC^{-1}(y_2 - b), A - BC^{-1}B^T\right)$$



# Gaussian Processes for regression

**Reminder:** the conditioning equations for multivariate Gaussian distributions

$$p(y_1, y_2) = \mathcal{N}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right)$$

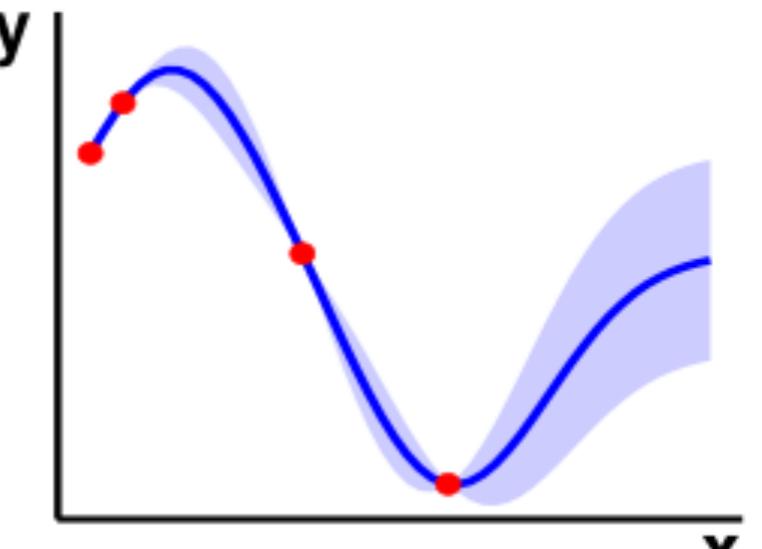
$$p(y_1 | y_2) = \frac{p(y_1, y_2)}{p(y_2)} \quad p(y_2) = \mathcal{N}(b, C)$$

$$p(y_1 | y_2) = \mathcal{N}\left(a + BC^{-1}(y_2 - b), A - BC^{-1}B^T\right)$$

**predictive mean:**  $\mu_{y_1|y_2} = a + BC^{-1}(y_2 - b)$

**predictive covariance:**  $\Sigma_{y_1|y_2} = A - BC^{-1}B^T$

Predictive uncertainty = prior uncertainty - reduction in uncertainty



# Gaussian Processes for regression: noiseless

What are the means and variances of the values  $f^*$  at points  $X^*$ , given observed values  $f$  at points  $X$ .

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \boldsymbol{\mu}(X) \\ \boldsymbol{\mu}(X_*) \end{pmatrix}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix} \right)$$

Conditioning:

$$p(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{f}) = \mathcal{N}(\mathbf{f}_* | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$$

$$\boldsymbol{\mu}_* = \boldsymbol{\mu}(\mathbf{X}_*) + \mathbf{K}_*^T \mathbf{K}^{-1} (\mathbf{f} - \boldsymbol{\mu}(\mathbf{X}))$$

$$\boldsymbol{\Sigma}_* = \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{K}_*$$

# Gaussian Processes for regression: noisy

What are the means and variances of the values  $f_*$  at points  $X_*$ , given (noisy) observed values  $y$  at points  $X$ .

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \boldsymbol{\mu}(X) \\ \boldsymbol{\mu}(X_*) \end{pmatrix}, \begin{pmatrix} \mathbf{K}_y & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix} \right) \quad \mathbf{K}_y = \mathbf{K} + \sigma_y^2 \mathbf{I}_N$$

Conditioning:

$$\begin{aligned} p(\mathbf{f}_* | X_*, X, \mathbf{y}) &= \mathcal{N}(\mathbf{f}_* | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\ \boldsymbol{\mu}_* &= \boldsymbol{\mu}(X_*) + \mathbf{K}_*^T \mathbf{K}_y^{-1} (\mathbf{y} - \boldsymbol{\mu}(X)) \\ \boldsymbol{\Sigma}_* &= \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{K}_* \end{aligned}$$

# Gaussian Processes for regression: noisy

What is the mean and variance of the value  $f_*$  at **single point**  $x_*$ , given (noisy) observed values  $y$  at points  $X$ .

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \mu(X) \\ \mu(x_*) \end{pmatrix}, \begin{pmatrix} \mathbf{K}_y & \mathbf{k}_* \\ \mathbf{k}_*^T & \mathbf{k}_{**} \end{pmatrix} \right) \quad \mathbf{K}_y = \mathbf{K} + \sigma_y^2 \mathbf{I}_N$$

Conditioning:

$$\mathbf{k}_{**} = K(\mathbf{x}, \mathbf{x}) + \sigma_y^2 \text{ scalar!}$$

$$p(\mathbf{f}_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{f}_* | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$$

$$\boldsymbol{\mu}_* = \boldsymbol{\mu}(\mathbf{x}_*) + \mathbf{k}_*^T \mathbf{K}_y^{-1} (\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}))$$

$$\sigma_*^2 = \mathbf{k}_{**} - \mathbf{k}_*^T \mathbf{K}_y^{-1} \mathbf{k}_*$$

# Computational Cost

prediction task

- train on N points
- test on M points

prediction equations

$$\mu_M = \mathbf{K}_{MN} \mathbf{K}_{NN}^{-1} \mathbf{y}_N$$

$$\Sigma_{MM} = \mathbf{K}_{MM} - \mathbf{K}_{MN} \mathbf{K}_{NN}^{-1} \mathbf{K}_{NM}$$

Full cost  $\mathcal{O}((N+M)^3)$

Without special structure, computation is limited to  $\mathcal{O}(1000)$  variables

Computational cost is a major limitation of GPs

# Numerical Computations Considerations

For GP  $\mu(x) = 0$

$$p(\mathbf{f}_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{f}_* | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$$

$$\boldsymbol{\mu}_* = \mathbf{k}_*^T \mathbf{K}_y^{-1} \mathbf{y}$$

$$\sigma_*^2 = \mathbf{k}_{**} - \mathbf{k}_*^T \mathbf{K}_y^{-1} \mathbf{k}_*$$

$$\mathbf{k}_{**} = K(\mathbf{x}, \mathbf{x}) + \sigma_y^2$$

$$\mathbf{K}_y = \mathbf{L} \mathbf{L}^T$$

$$\boldsymbol{\alpha} = \mathbf{K}_y^{-1} \mathbf{y} = \mathbf{L}^{-T} \mathbf{L}^{-1} \mathbf{y}$$

**Algorithm:** GP Regression

1.  $\mathbf{L} = \text{cholesky}(\mathbf{K}_y)$

2.  $\boldsymbol{\alpha} = \mathbf{L}^T(\mathbf{L} \backslash \mathbf{y})$

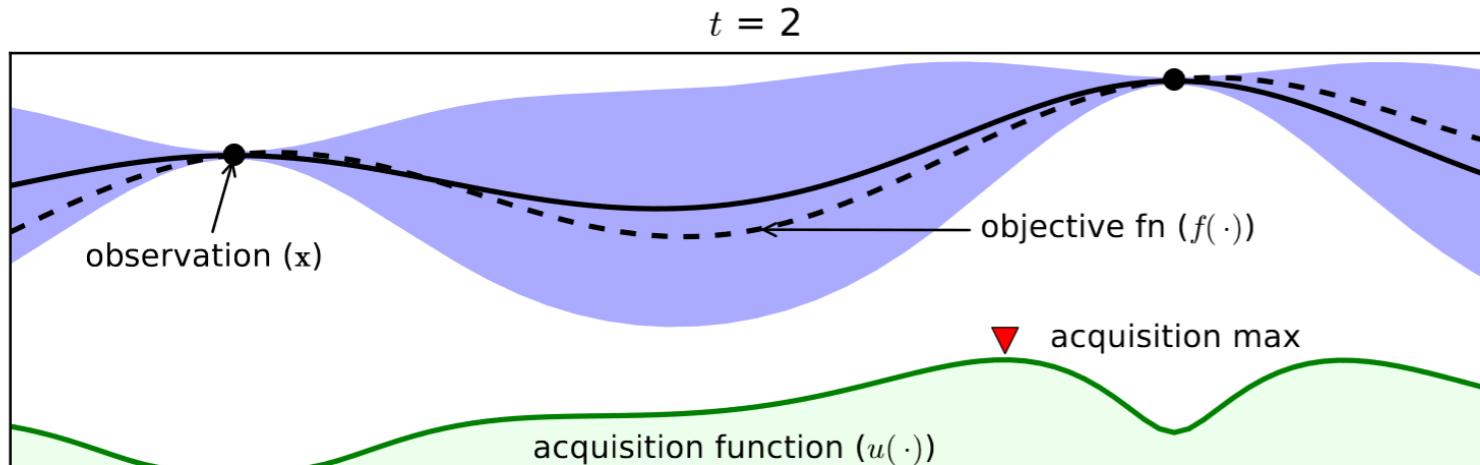
3.  $\mu_* = \mathbb{E}[f_*] = \mathbf{k}_*^T \boldsymbol{\alpha}$

4.  $\mathbf{v} = \mathbf{L} \backslash \mathbf{k}_*$

5.  $\sigma_*^2 = \text{var}[f_*] = \mathbf{k}_{**} - \mathbf{v}^T \mathbf{v}$

# Bayesian Optimization with Gaussian processes

# Bayesian Optimization with Gaussian processes



for  $t = 1, 2, \dots$  do

1. Find  $\mathbf{x}_t$  by combining attributes of the posterior distribution in a utility function  $u$  and maximising:  
$$\mathbf{x}_t = \arg \max_{\mathbf{x}} u(\mathbf{x} | \mathcal{D}_{1:t-1})$$

2. Sample the objective function:

$$y_t = f(x_t) + \varepsilon_t$$

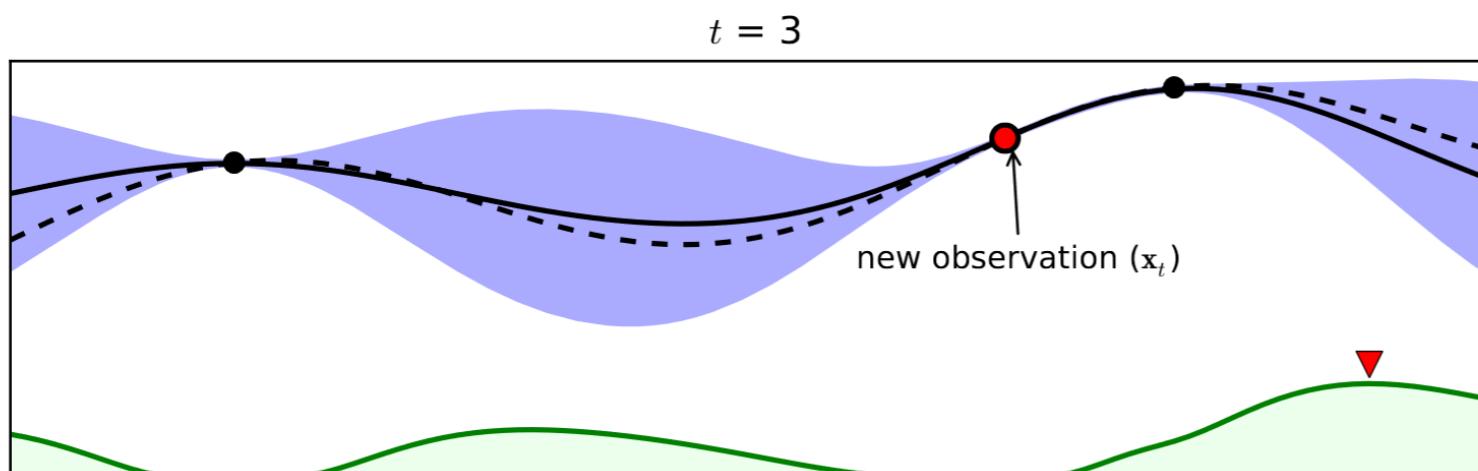
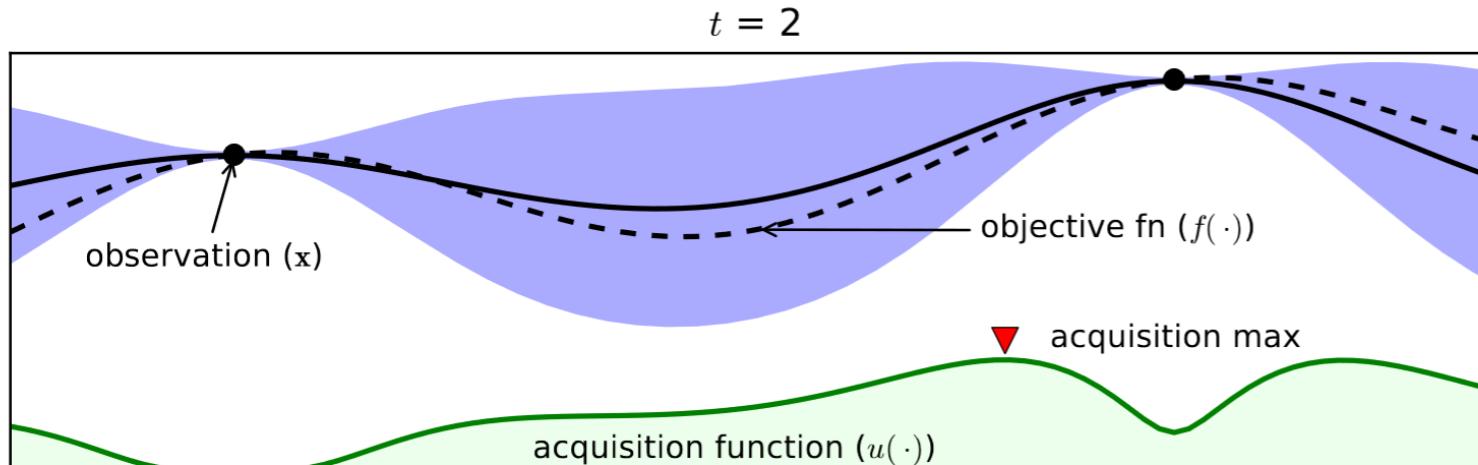
3. Augment the data :

$$\mathcal{D}_{1:t} = \left\{ \mathcal{D}_{1:t-1}, (\mathbf{x}_t, y_t) \right\}$$

and update the GP.

end for

# Bayesian Optimization with Gaussian processes



for  $t = 1, 2, \dots$  do

1. Find  $\mathbf{x}_t$  by combining attributes of the posterior distribution in a utility function  $u$  and maximising:  
$$\mathbf{x}_t = \arg \max_{\mathbf{x}} u(\mathbf{x} | \mathcal{D}_{1:t-1})$$

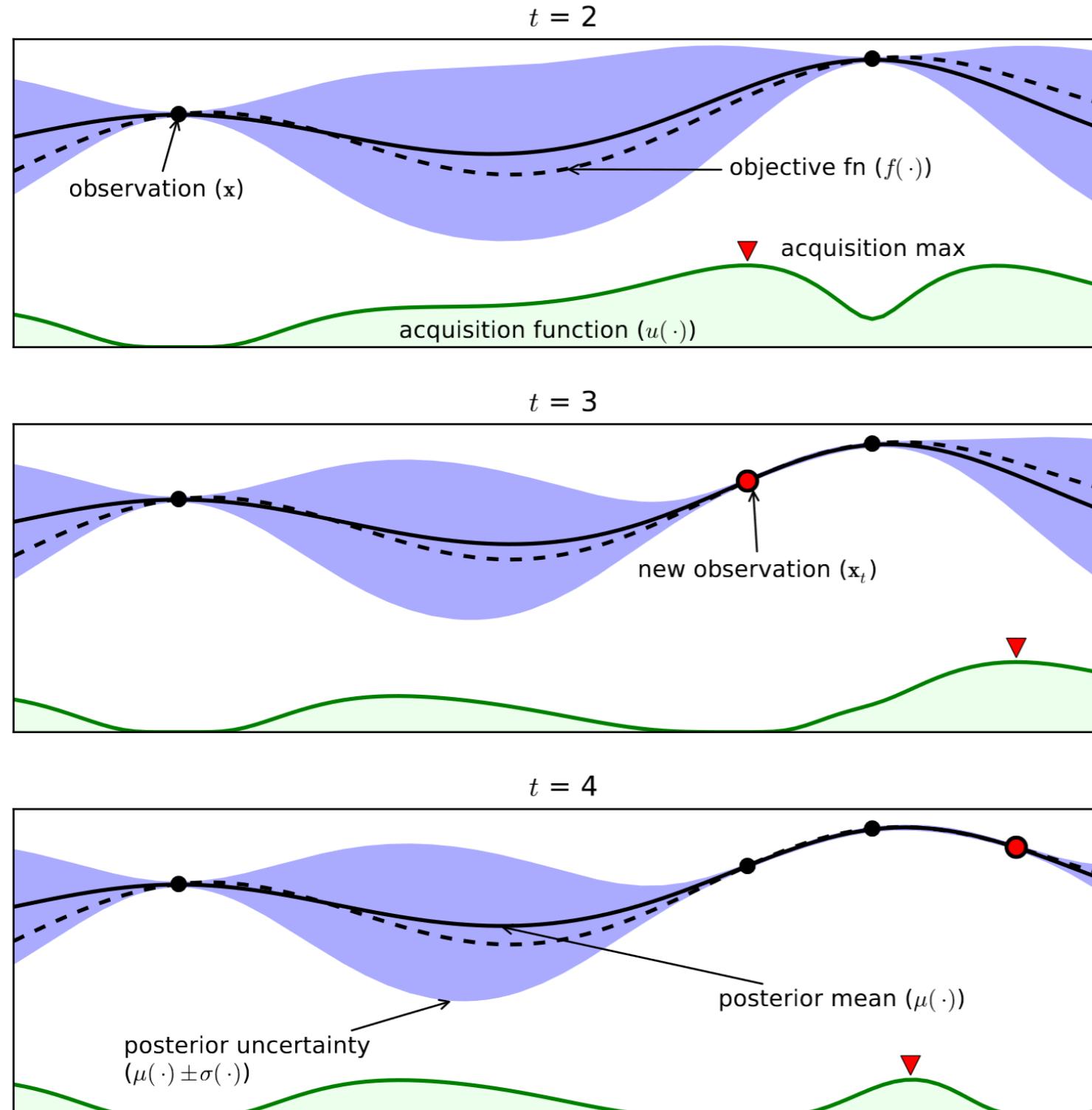
2. Sample the objective function:

$$y_t = f(x_t) + \varepsilon_t$$

3. Augment the data :  
$$\mathcal{D}_{1:t} = \left\{ \mathcal{D}_{1:t-1}, (\mathbf{x}_t, y_t) \right\}$$
 and update the GP.

end for

# Bayesian Optimization with Gaussian processes



for  $t = 1, 2, \dots$  do

1. Find  $\mathbf{x}_t$  by combining attributes of the posterior distribution in a utility function  $u$  and maximising:  
$$\mathbf{x}_t = \arg \max_{\mathbf{x}} u(\mathbf{x} | \mathcal{D}_{1:t-1})$$

2. Sample the objective function:

$$y_t = f(x_t) + \varepsilon_t$$

3. Augment the data :

$$\mathcal{D}_{1:t} = \left\{ \mathcal{D}_{1:t-1}, (\mathbf{x}_t, y_t) \right\}$$

and update the GP.

end for

# Exploration-Exploitation Tradeoff

How should we pick the next point  $x$  to evaluate?

GP prediction in the special case of **one test point**  $x_{t+1}$ :

$$P(y_{t+1} | \mathcal{D}_{1:t}, \mathbf{x}_{t+1}) = \mathcal{N}\left(\mu_t(\mathbf{x}_{t+1}), \sigma_t^2(\mathbf{x}_{t+1}) + \sigma_{\text{noise}}^2\right)$$

$$\mu_t(\mathbf{x}_{t+1}) = \mathbf{k}^T \left[ \mathbf{K} + \sigma_{\text{noise}}^2 \mathbf{I} \right]^{-1} \mathbf{y}_{1:t}$$

$$\sigma_t^2(\mathbf{x}_{t+1}) = k(\mathbf{x}_{t+1}, \mathbf{x}_{t+1}) - \mathbf{k}^T \left[ \mathbf{K} + \sigma_{\text{noise}}^2 \mathbf{I} \right]^{-1} \mathbf{k}$$

We should choose the next point  $x$  where the mean is high (exploitation) and the variance is high (exploration).

We can balance exploration and exploitation with an acquisition function  $u$ :

$$\mu(x) + \kappa \sigma(x)$$

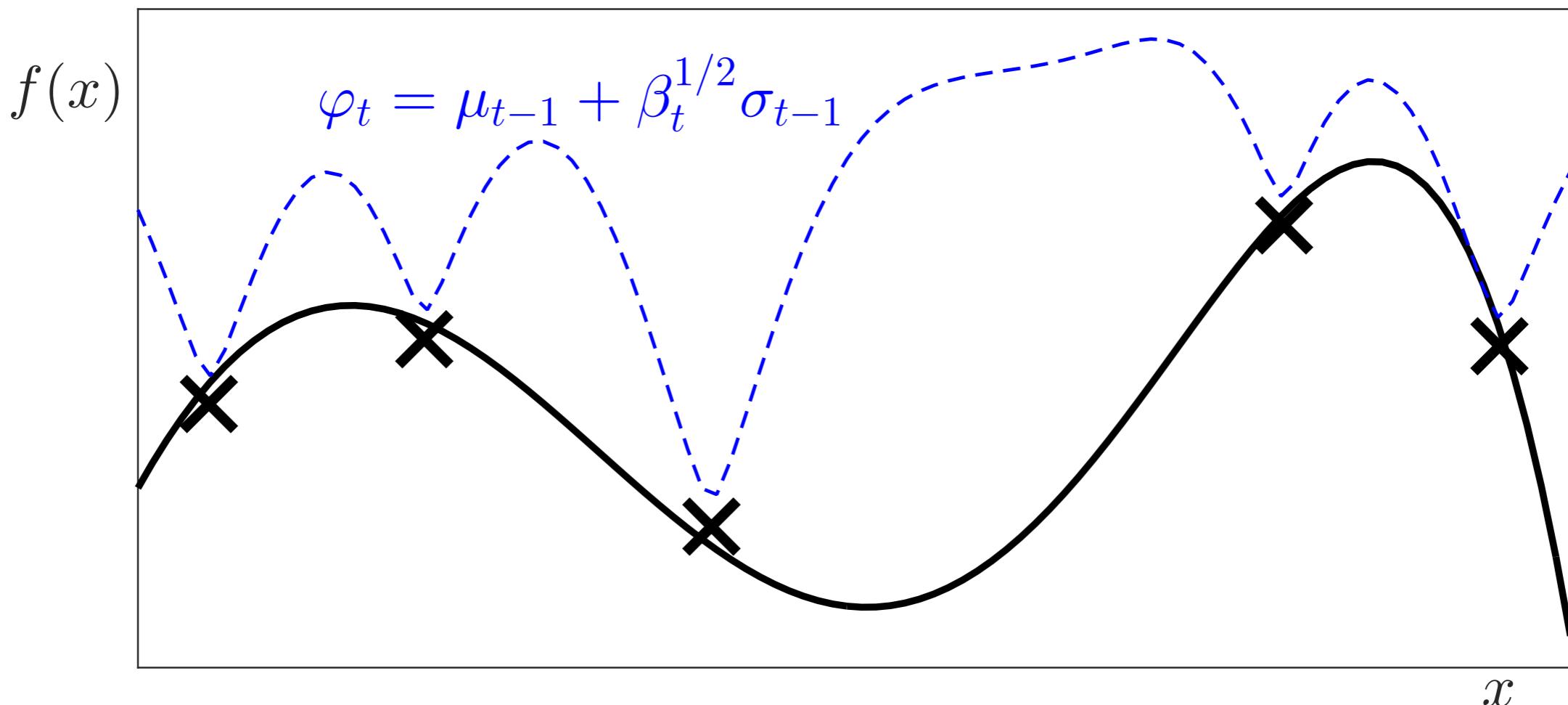
But then we need to **maximize our acquisition function to pick the next point**. For this we use some vanilla black box optimizer.

# GP-Upper Confidence Bound (UCB)

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



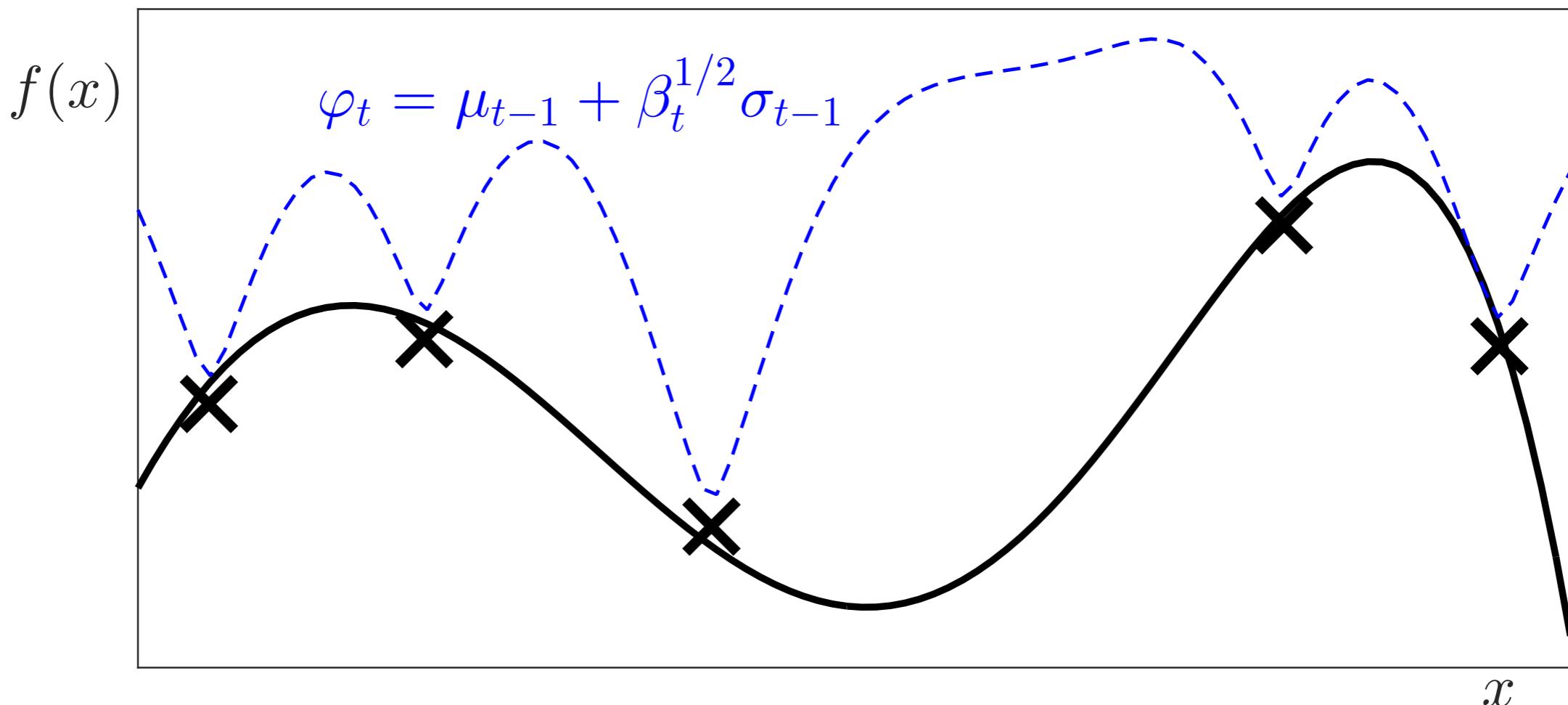
- 1) Compute posterior  $\mathcal{GP}$

# GP-Upper Confidence Bound (UCB)

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



1) Compute posterior  $\mathcal{GP}$

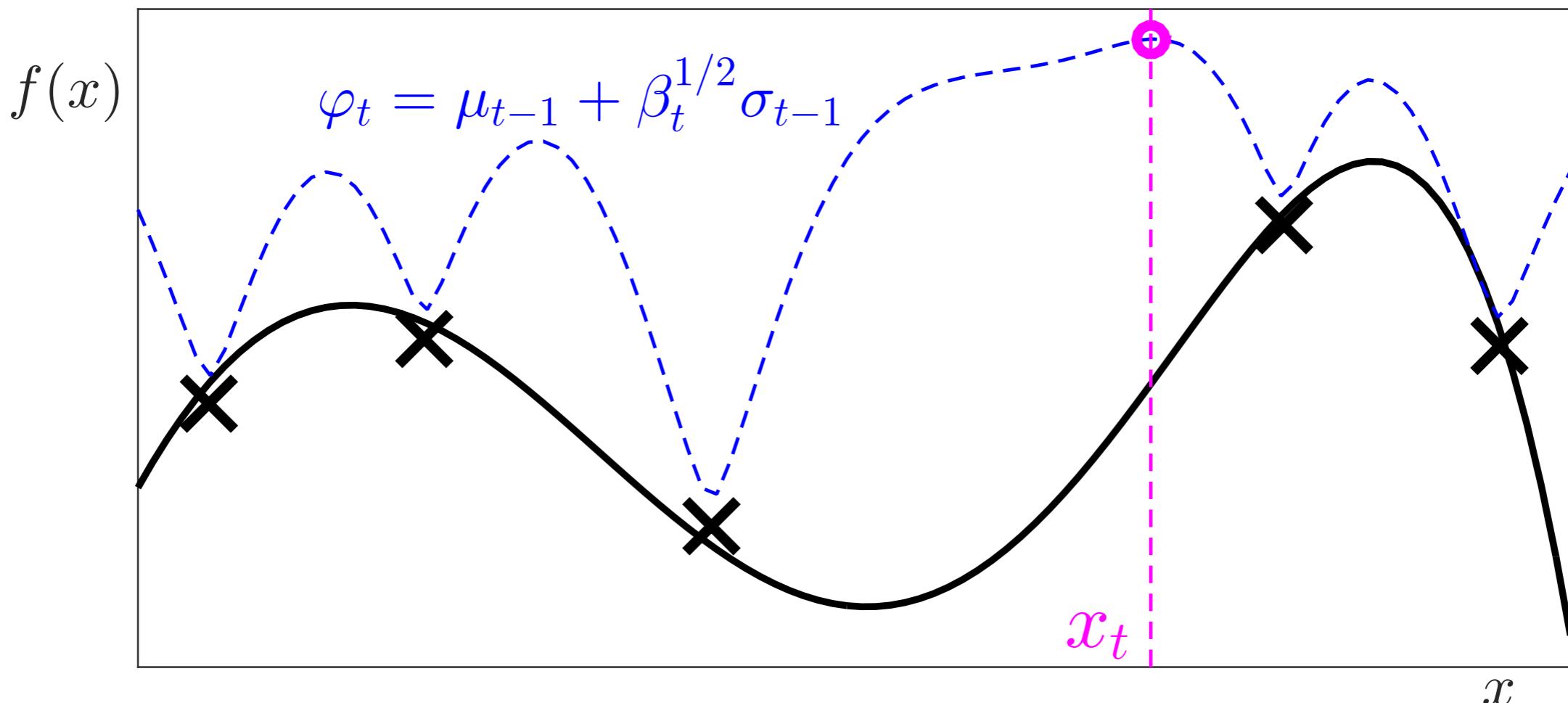
2) Construct UCB  $\varphi_t$

# GP-Upper Confidence Bound (UCB)

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



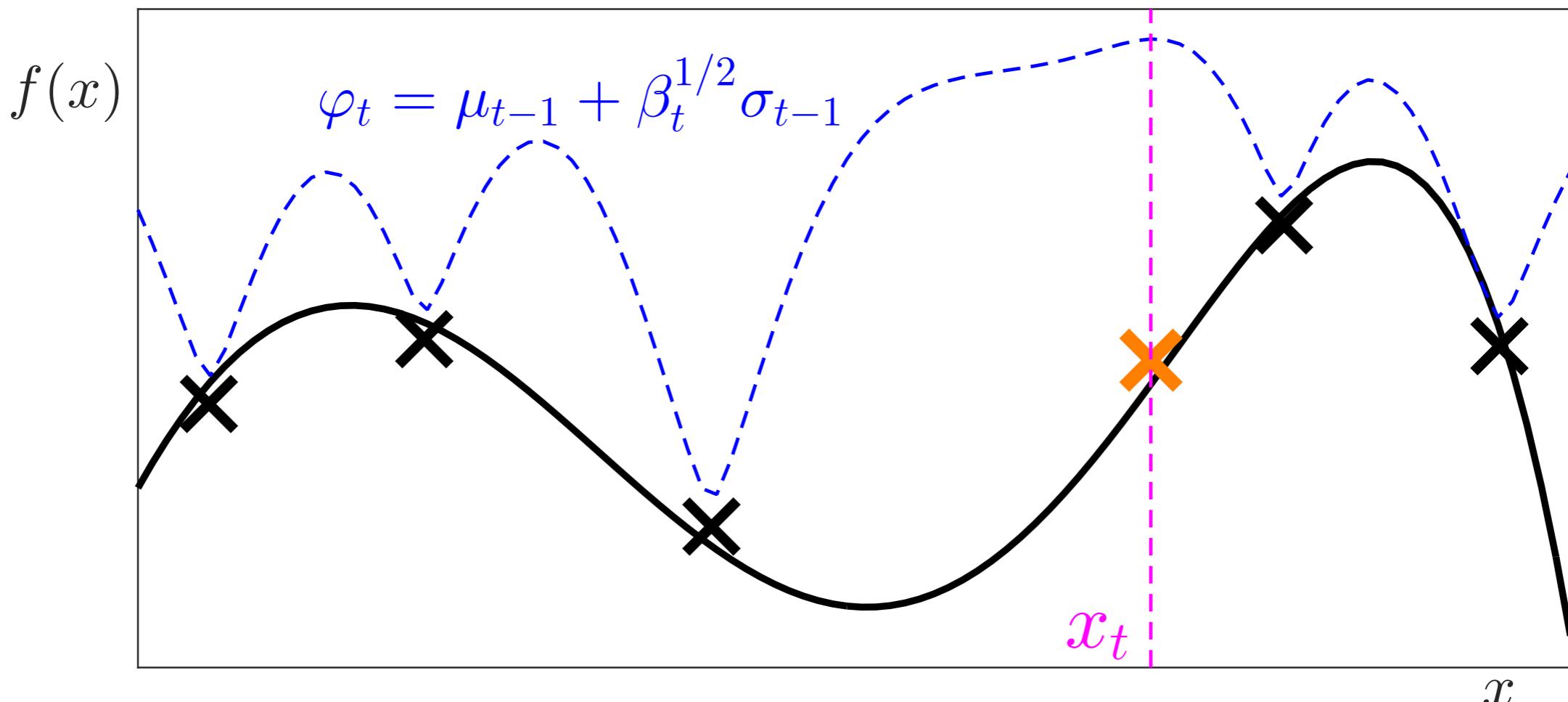
- 1) Compute posterior  $\mathcal{GP}$
- 2) Construct UCB  $\varphi_t$
- 3) Choose  $x_t = \operatorname{argmax}_{x \in D \subset (0, r)^d} \varphi_t$

# GP-Upper Confidence Bound (UCB)

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



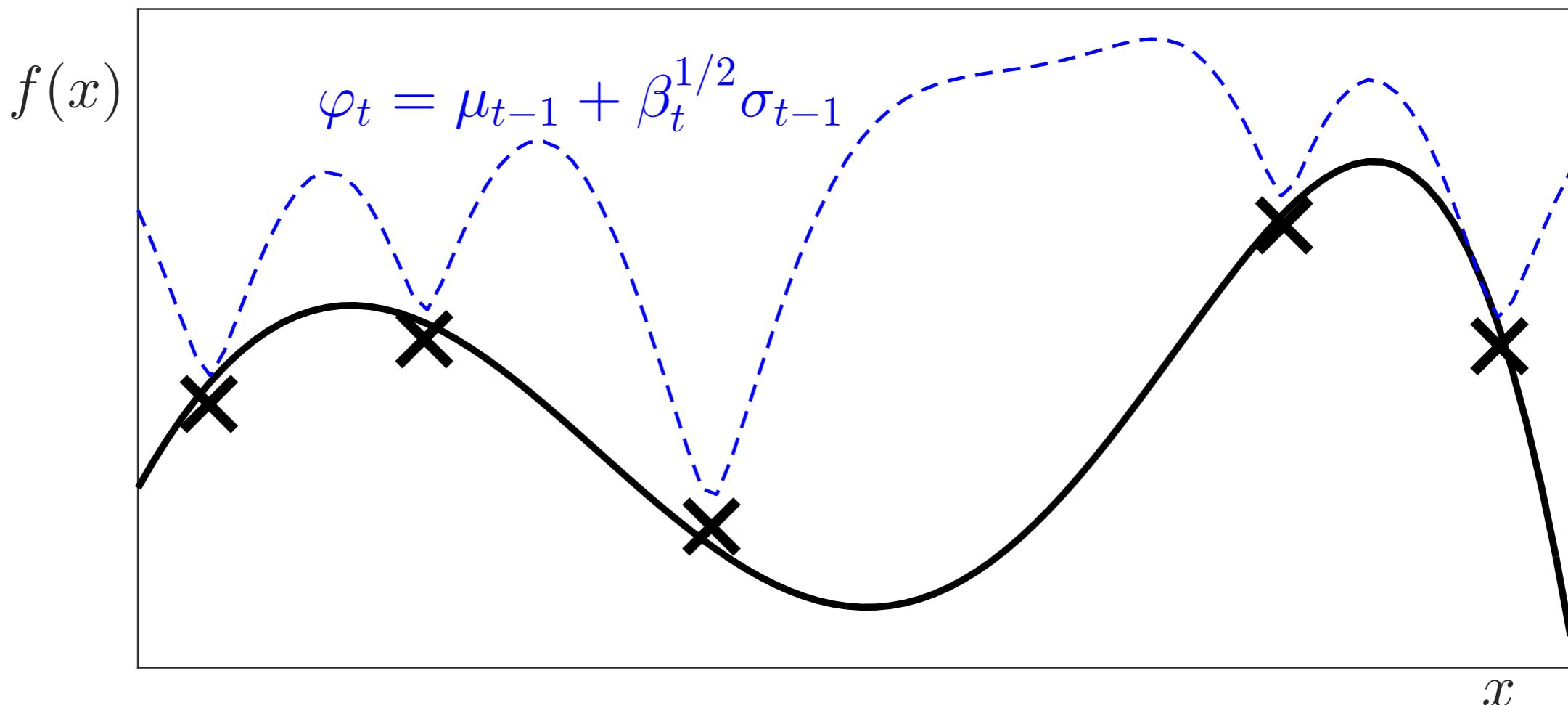
- 1) Compute posterior  $\mathcal{GP}$
- 2) Construct UCB  $\varphi_t$
- 3) Choose  $x_t = \operatorname{argmax}_{x \in D \subset (0, r)^d} \varphi_t$
- 4) Evaluate  $f$  at  $x_t$

# GP-Upper Confidence Bound (UCB)

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$ .

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



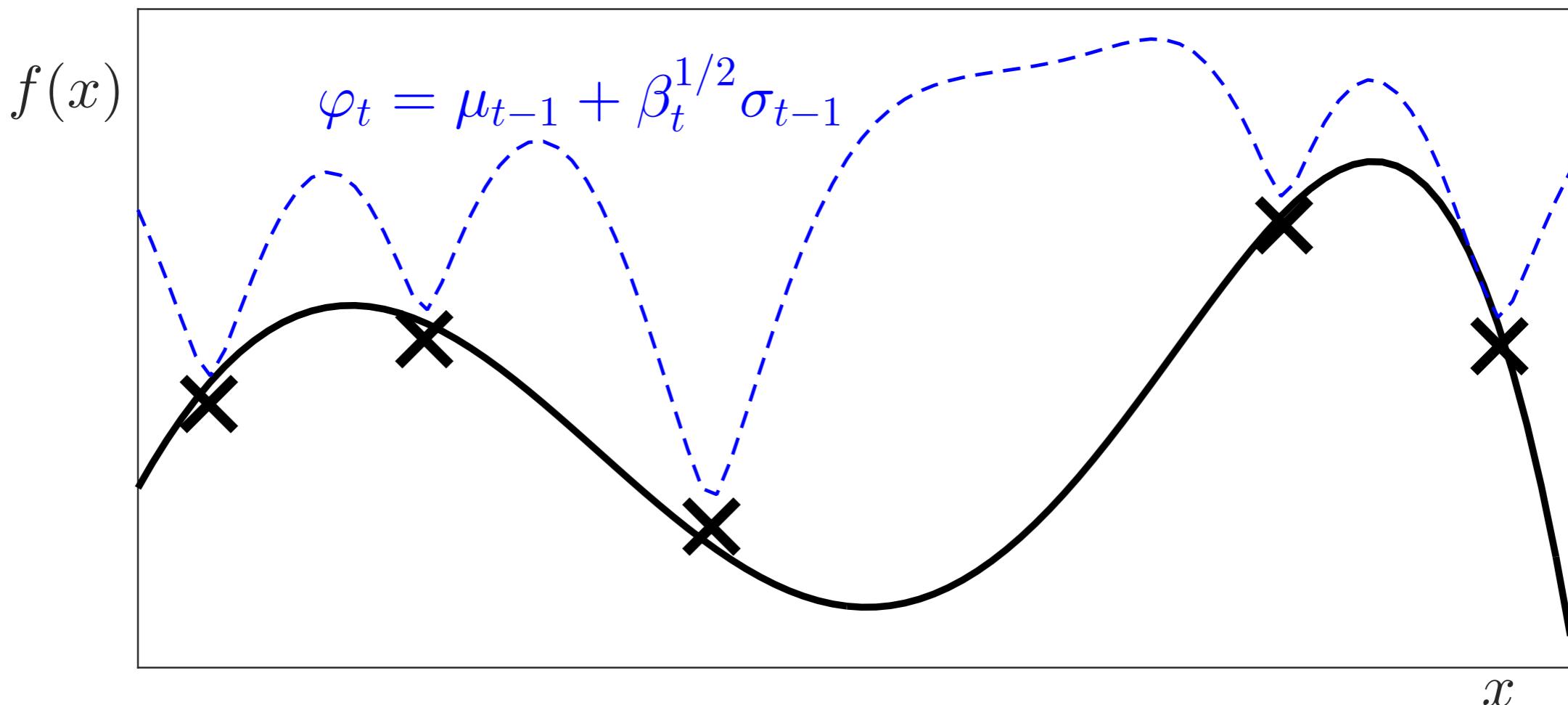
- 1) Compute posterior  $\mathcal{GP}$

# GP-Upper Confidence Bound (UCB)

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$ .

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



1) Compute posterior  $\mathcal{GP}$

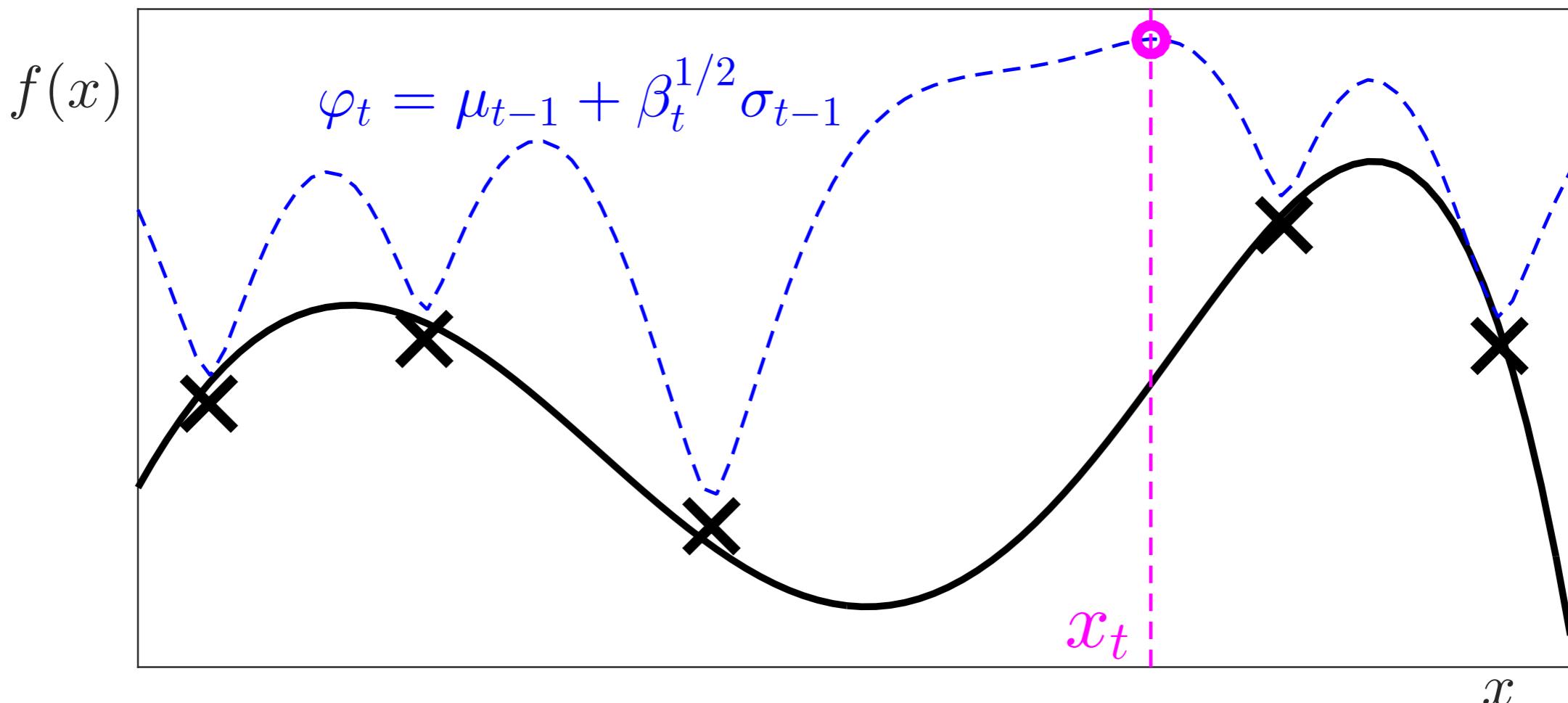
2) Construct UCB  $\varphi_t$

# GP-Upper Confidence Bound (UCB)

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$ .

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



1) Compute posterior  $\mathcal{GP}$

2) Construct UCB  $\varphi_t$

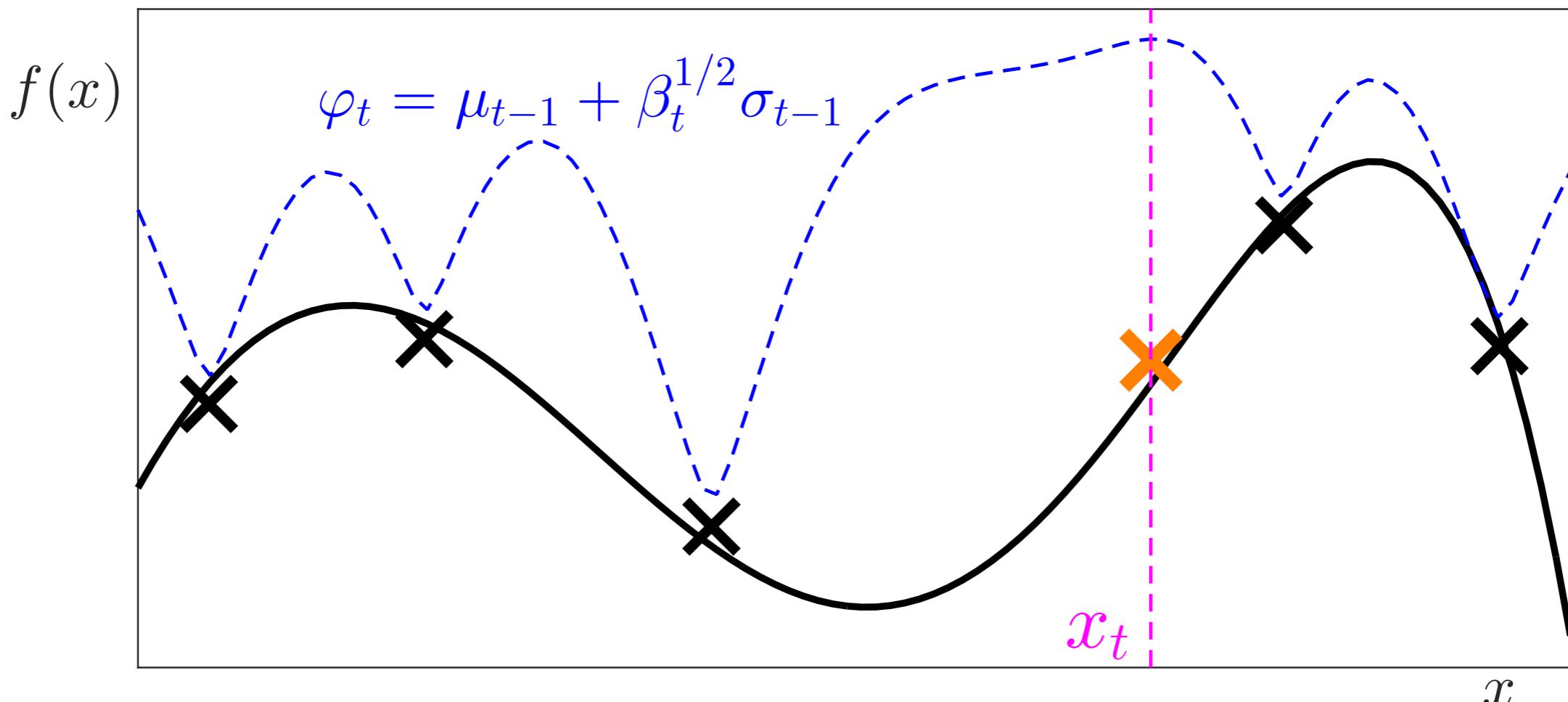
3) Choose  $x_t = \operatorname{argmax}_{x \in D \subset (0, r)^d} \varphi_t$

# GP-Upper Confidence Bound (UCB)

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$ .

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



- 1) Compute posterior  $\mathcal{GP}$
- 2) Construct UCB  $\varphi_t$
- 3) Choose  $x_t = \operatorname{argmax}_{x \in D \subset (0, r)^d} \varphi_t$
- 4) Evaluate  $f$  at  $x_t$

# GP-Upper Confidence Bound (UCB)

---

**Algorithm 1** GP-UCB

---

**Require:**  $k$

- 1:  $\mu \leftarrow 0_d$
  - 2: **for**  $t \leftarrow 1$  to  $T$  **do**
  - 3:      $\beta_t = 2 \log(t^{\frac{d}{2}+2}\pi^2/3\delta)$
  - 4:     Choose  $a_t \leftarrow \arg \max_i \mu_{t-1} + \sqrt{\beta_t} \sigma_{t-1}$
  - 5:     Observe  $y_t = f(\mathbf{x}_t) + \epsilon_t$
  - 6:      $\mu_t = k_{t-1}(\mathbf{x})^T (K_{t-1} + \sigma^2 I_d)^{-1} y_t$
  - 7:      $k_t = k(\mathbf{x}, \mathbf{x}') - k_{t-1}(\mathbf{x})^T (K_{t-1} + \sigma^2 I_d)^{-1} k_{t-1}(\mathbf{x}')$
  - 8:      $\sigma_t^2 = k_t(\mathbf{x}, \mathbf{x})$
  - 9: **end for**
- 

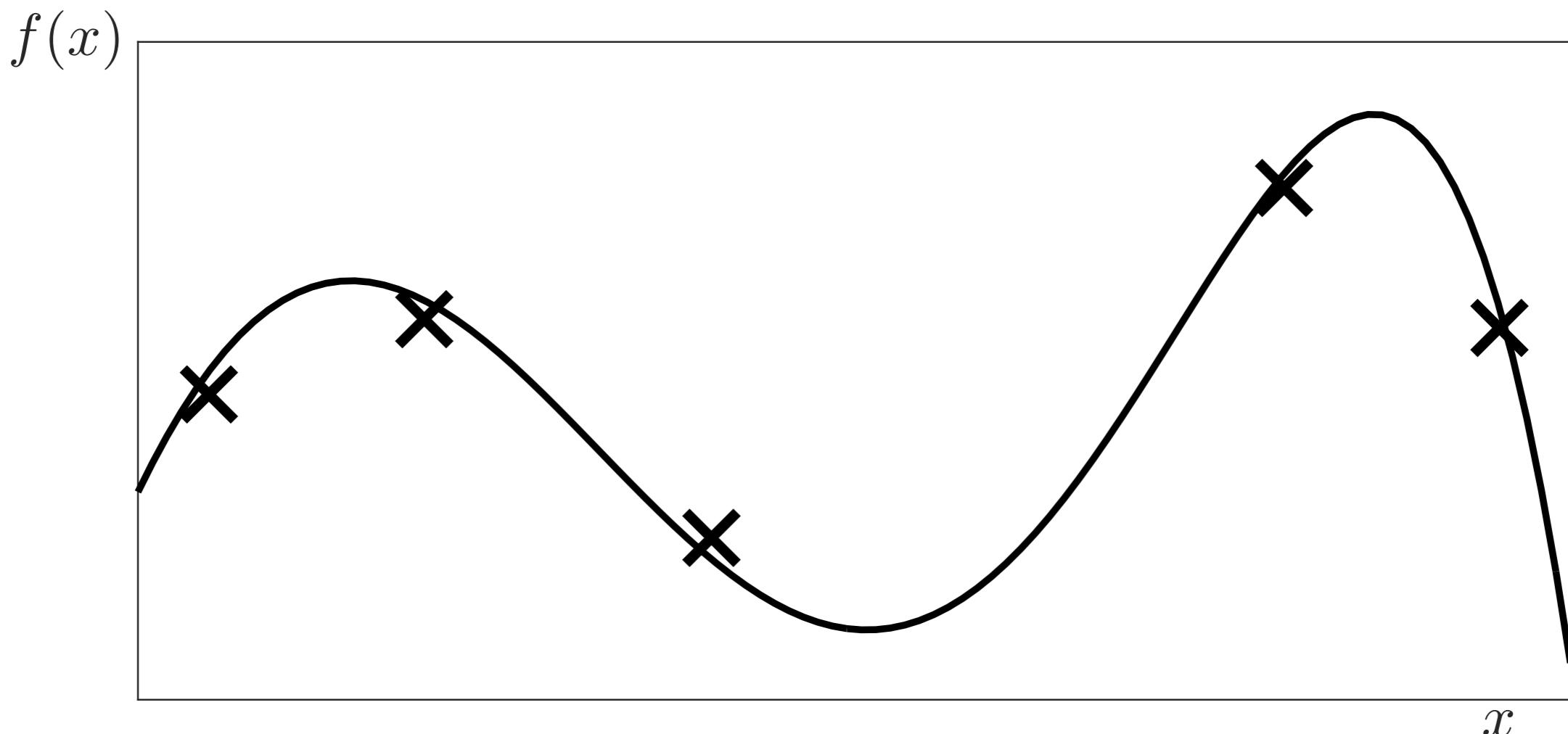
$d$ : the dimensionality of the objective,

$\delta$ : the probability that  $f(x)$  is bounded above by  $\mu_t + \beta_t \sigma_t$  and below by  $\mu_t - \beta_t \sigma_t$

# GP-Thompson Sampling

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$

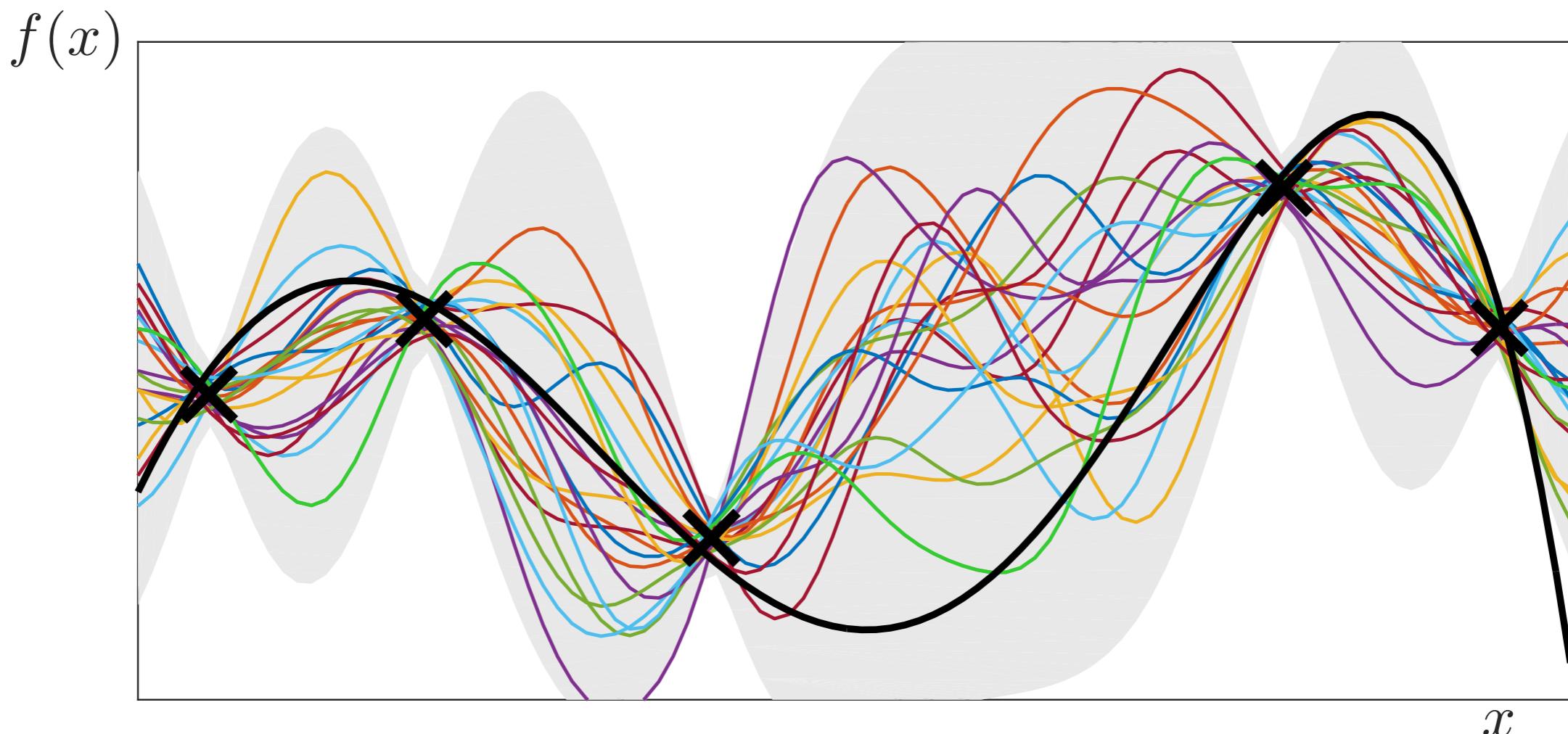
(Thompson, 1933)



# GP-Thompson Sampling

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$

(Thompson, 1933)

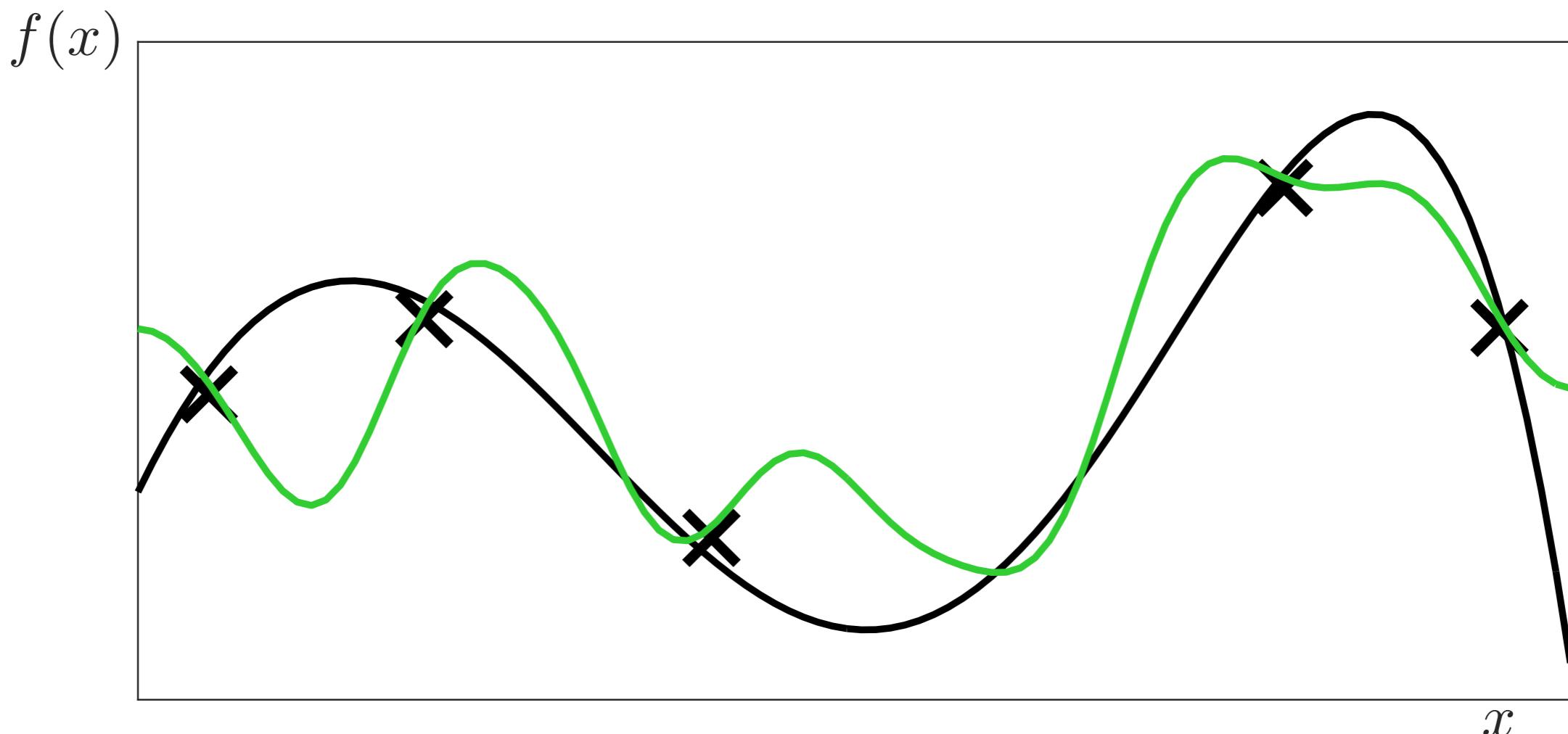


- 1) Construct posterior  $\mathcal{GP}$

# GP-Thompson Sampling

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$

(Thompson, 1933)



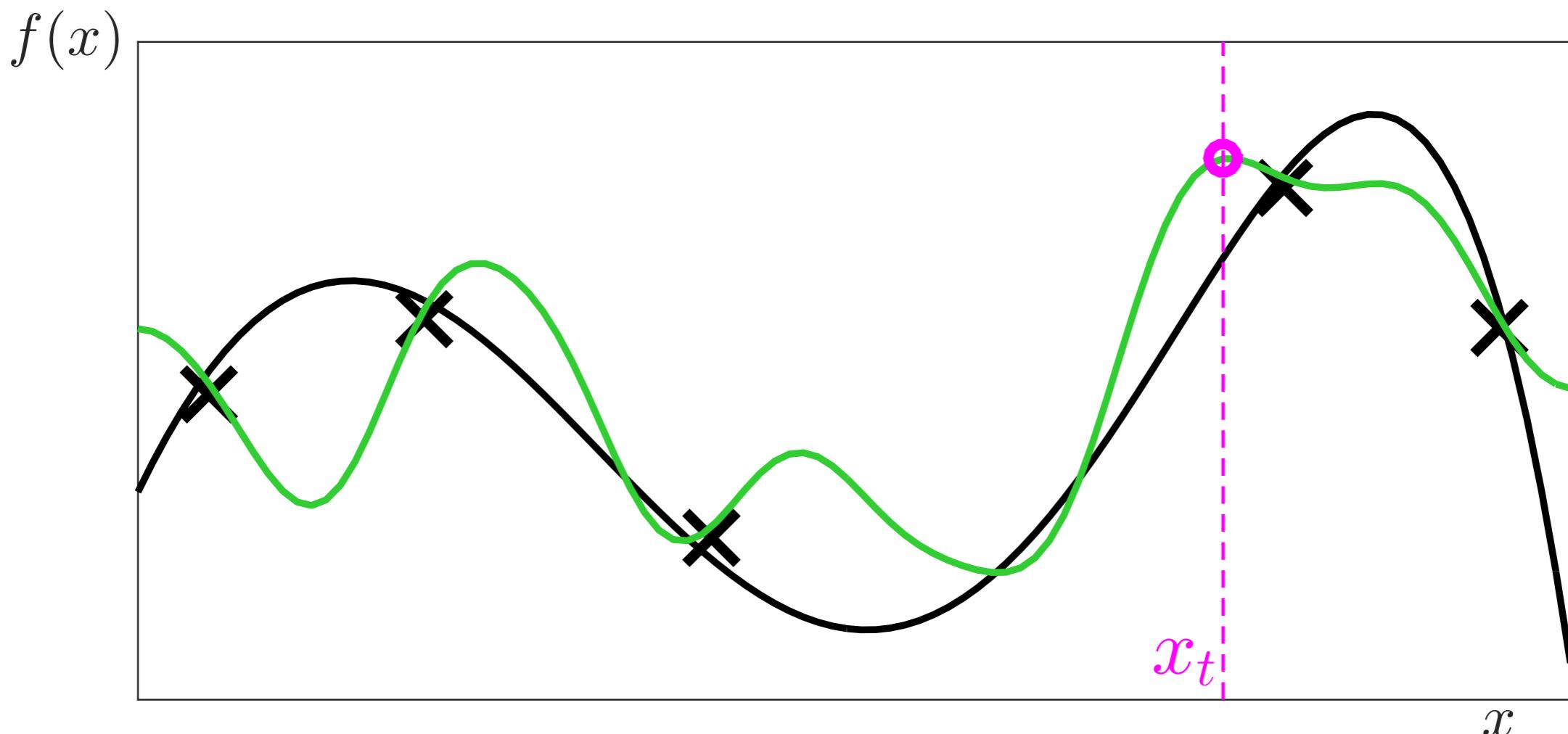
1) Construct posterior  $\mathcal{GP}$

2) Draw sample  $g$  from posterior

# GP-Thompson Sampling

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$

(Thompson, 1933)

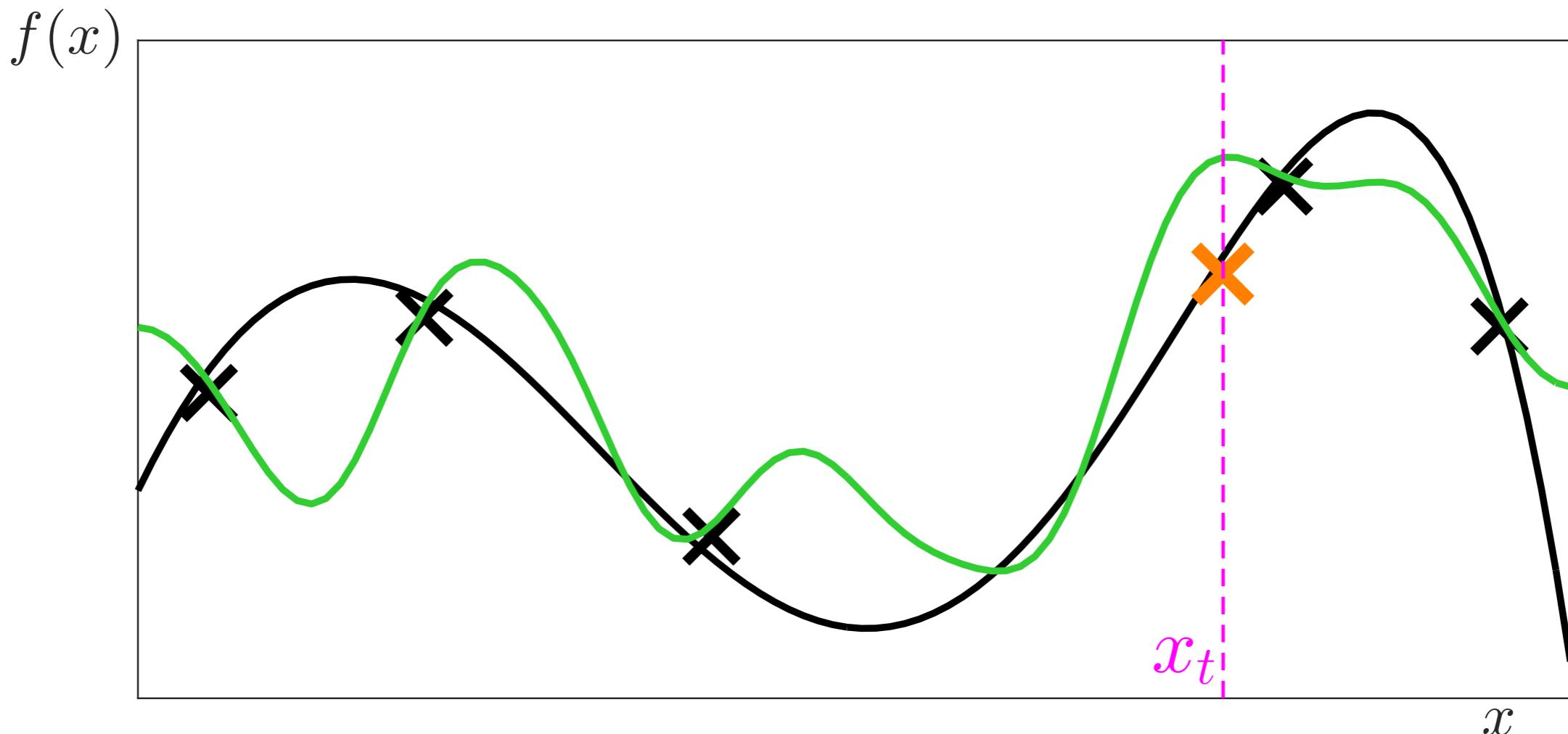


- 1) Construct posterior  $\mathcal{GP}$
- 2) Draw sample  $g$  from posterior
- 3) Choose  $x_t = \operatorname{argmax}_x g(x)$

# GP-Thompson Sampling

Model  $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$

(Thompson, 1933)



- 1) Construct posterior  $\mathcal{GP}$
- 2) Draw sample  $g$  from posterior
- 3) Choose  $x_t = \operatorname{argmax}_x g(x)$
- 4) Evaluate  $f$  at  $x_t$

# Representing Uncertainty

In later lectures we will explore representing uncertainty in regression and classification using neural networks.

We will look into:

- Bayesian neural networks, where we are estimating a distribution for each weight parameter as opposed to a point estimate,
- neural network ensembles: sets of neural networks trained on different subsets of the data and with different initializations, the entropy of their predictions quantify the uncertainty of their estimates.