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 and from Richard Turner's lecture on Gaussian process.

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.

Examples: drug design

Actions: the chemical consistencies to mix, Rewards: drug effectiveness (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.

Examples: 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.

It turns out, this is equivalent to optimizing (maximizing or minimizing) a function for which:

- We do not have an explicit parametric form, e.g., how should simulated smoke look like to look realistic to people. (our function here is a mapping from smoke simulation parameters to realism as measured by humans)
- 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.

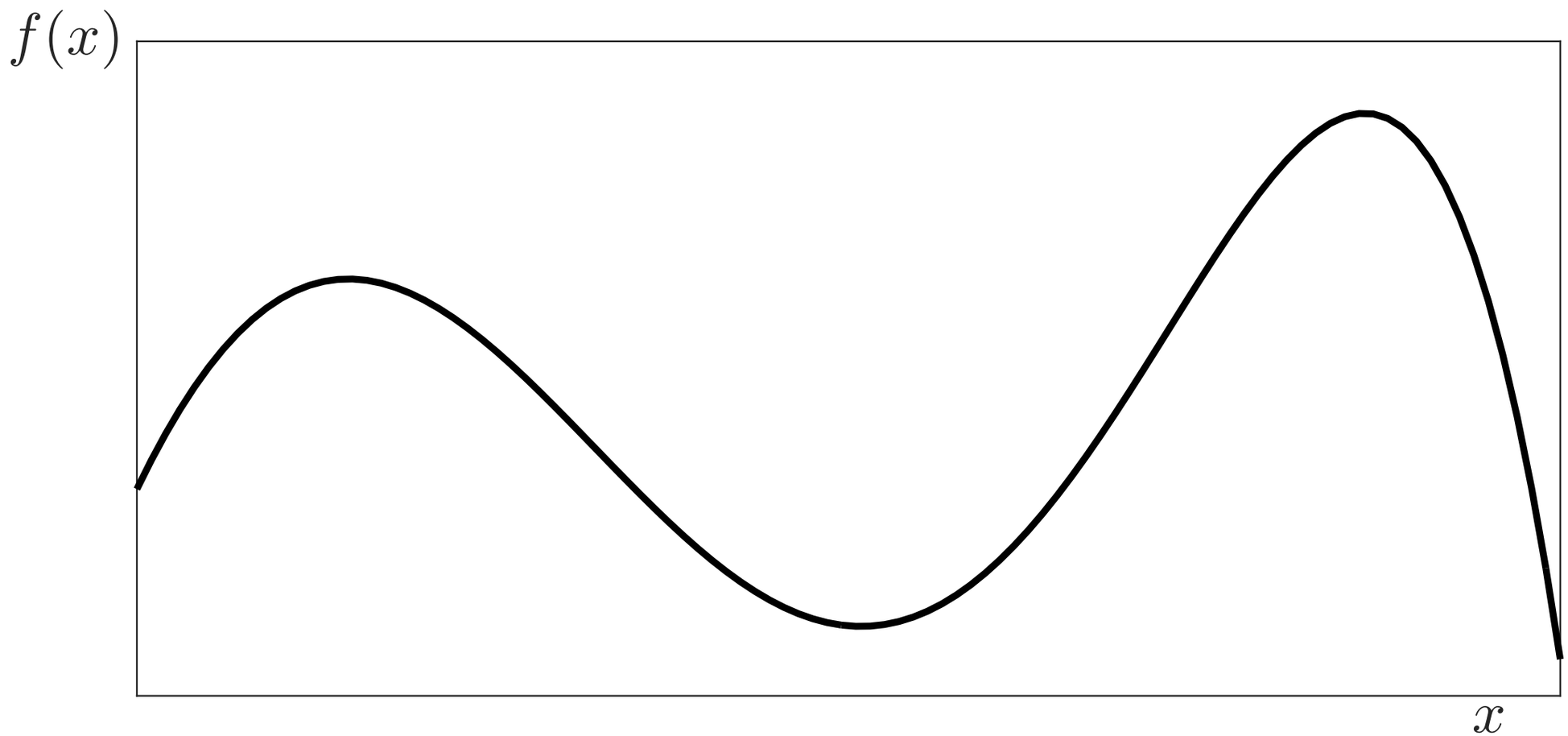
It turns out, this is equivalent to black-box (no gradients) optimization of functions.

Actions: places to evaluate the function.

Rewards: the value of the function.

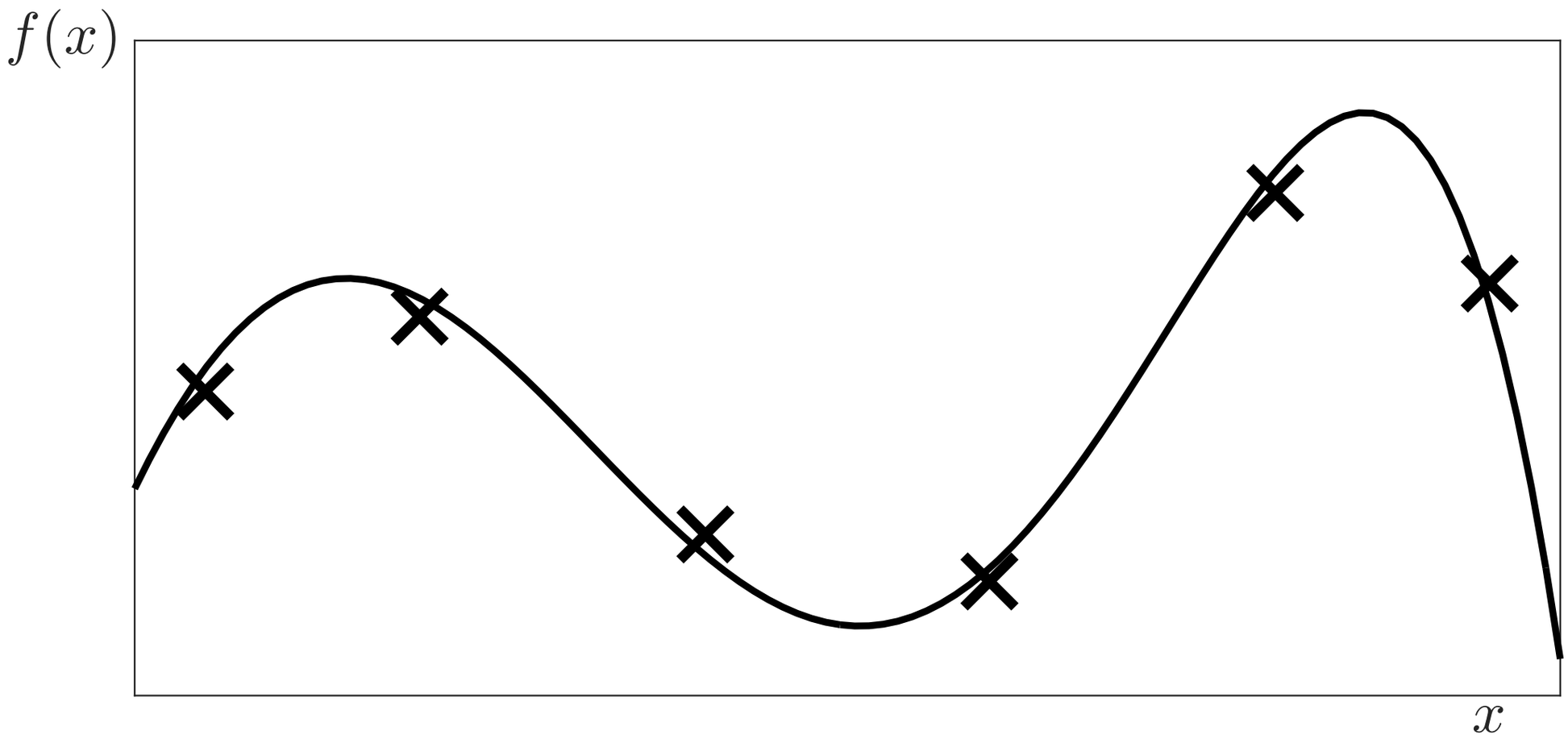
Black-box Optimisation

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



Black-box Optimisation

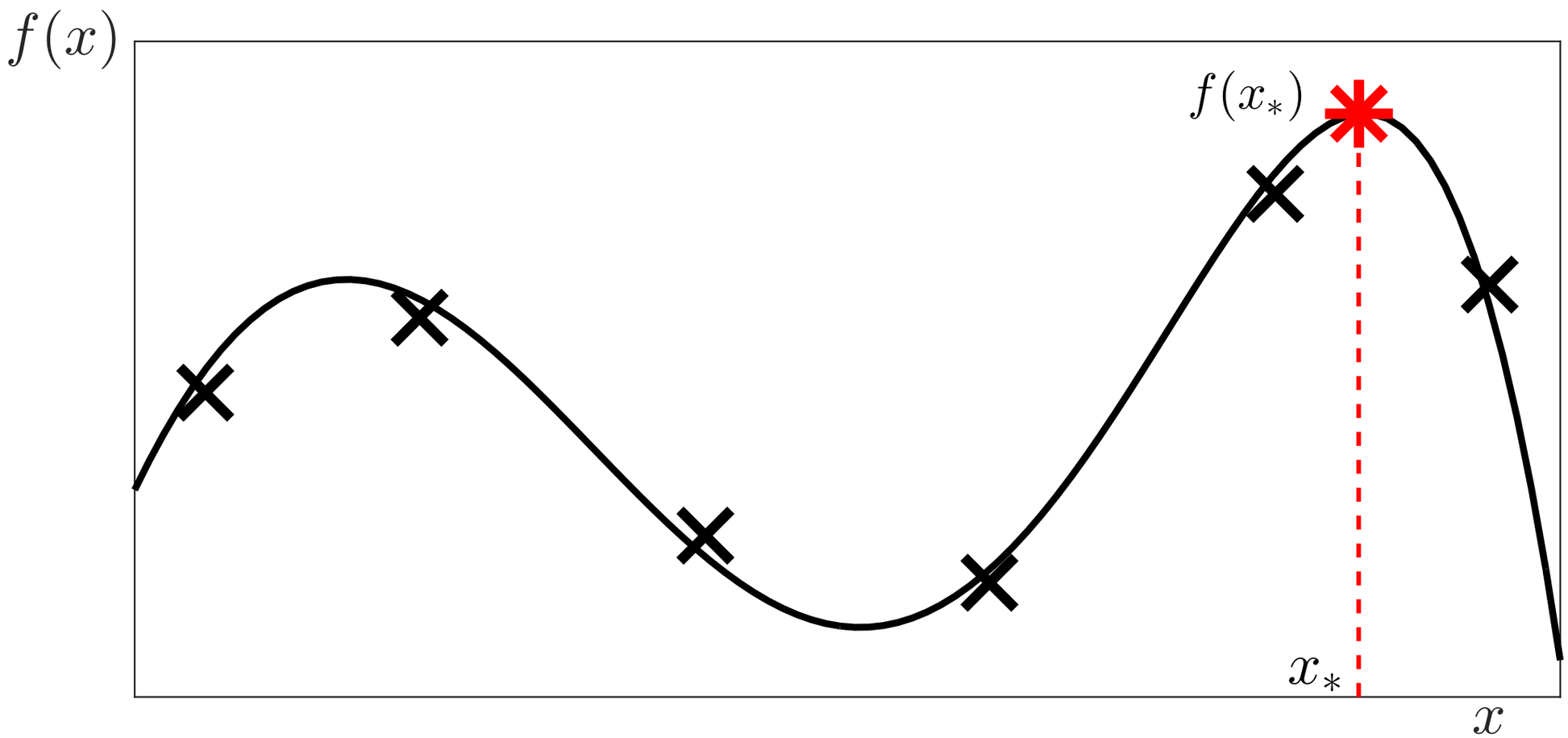
$f : \mathcal{X} \to \mathbb{R}$ is an expensive black-box function, accessible only via noisy evaluations.



Black-box Optimisation

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

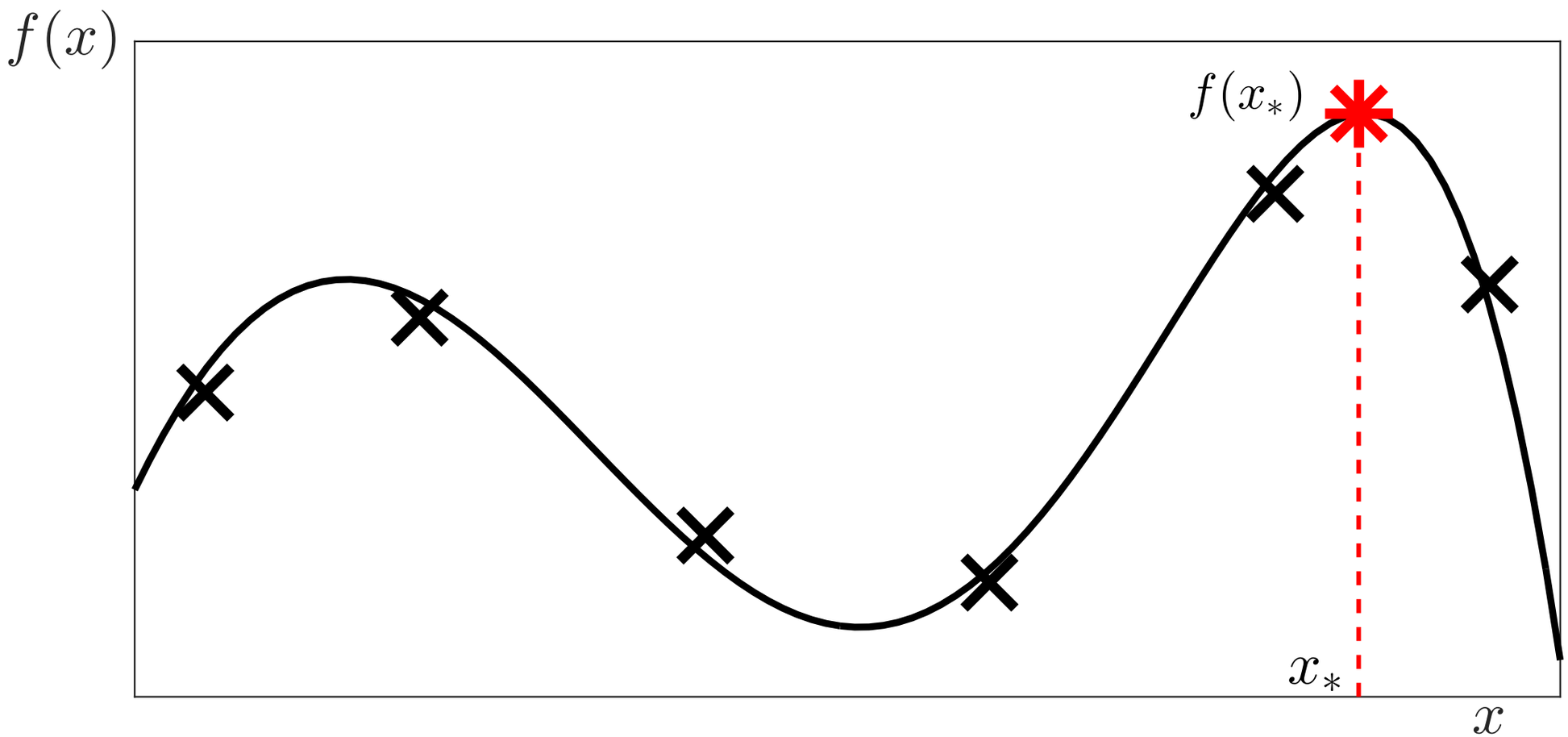
Let $x_\star = \mathrm{argmax}_x f(x)$.



Black-box Optimisation

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

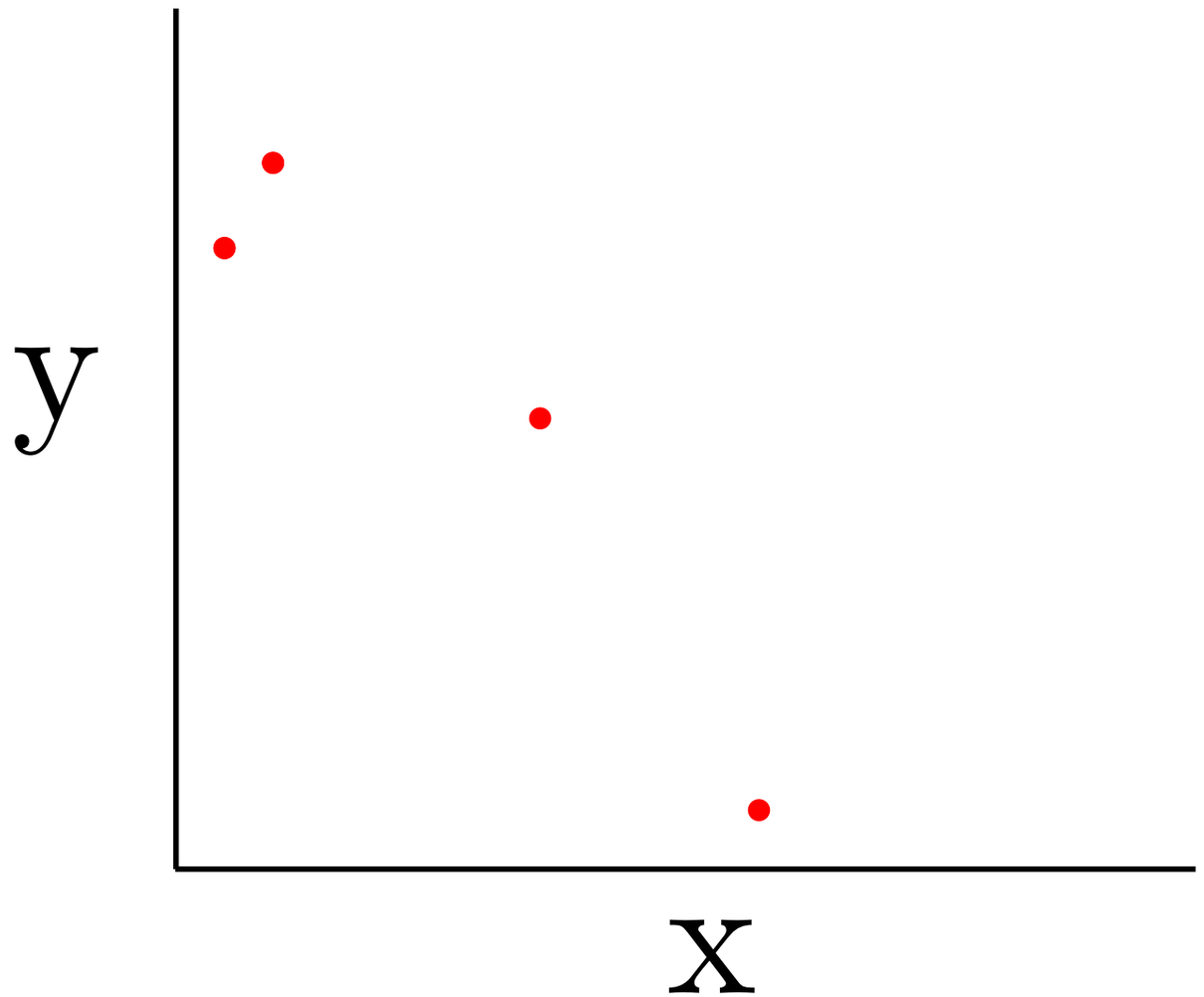
Let $x_\star = \mathrm{argmax}_x f(x)$.



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

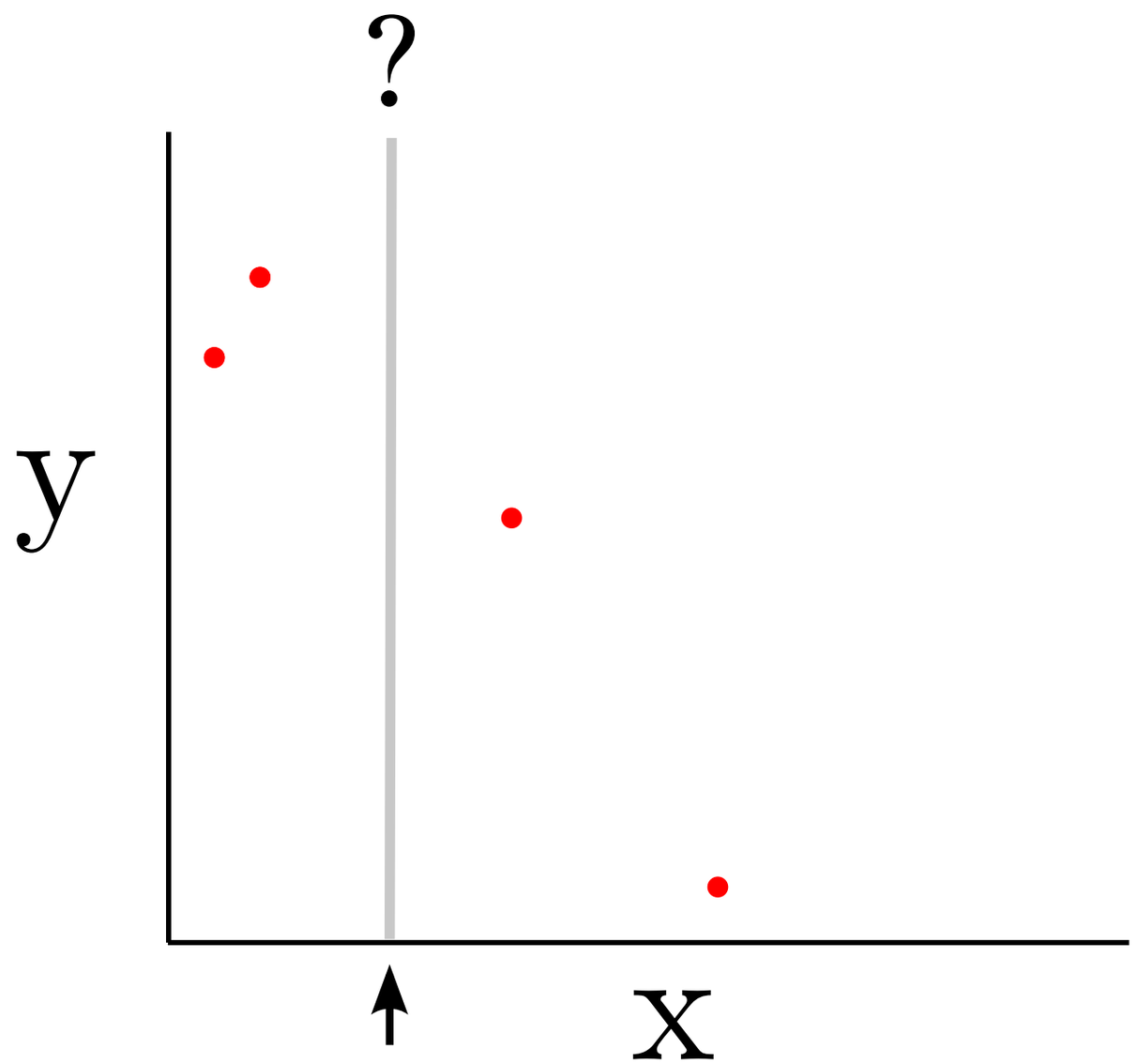
My action space: where in the x axis I should evaluate the function next.

Non-linear regression

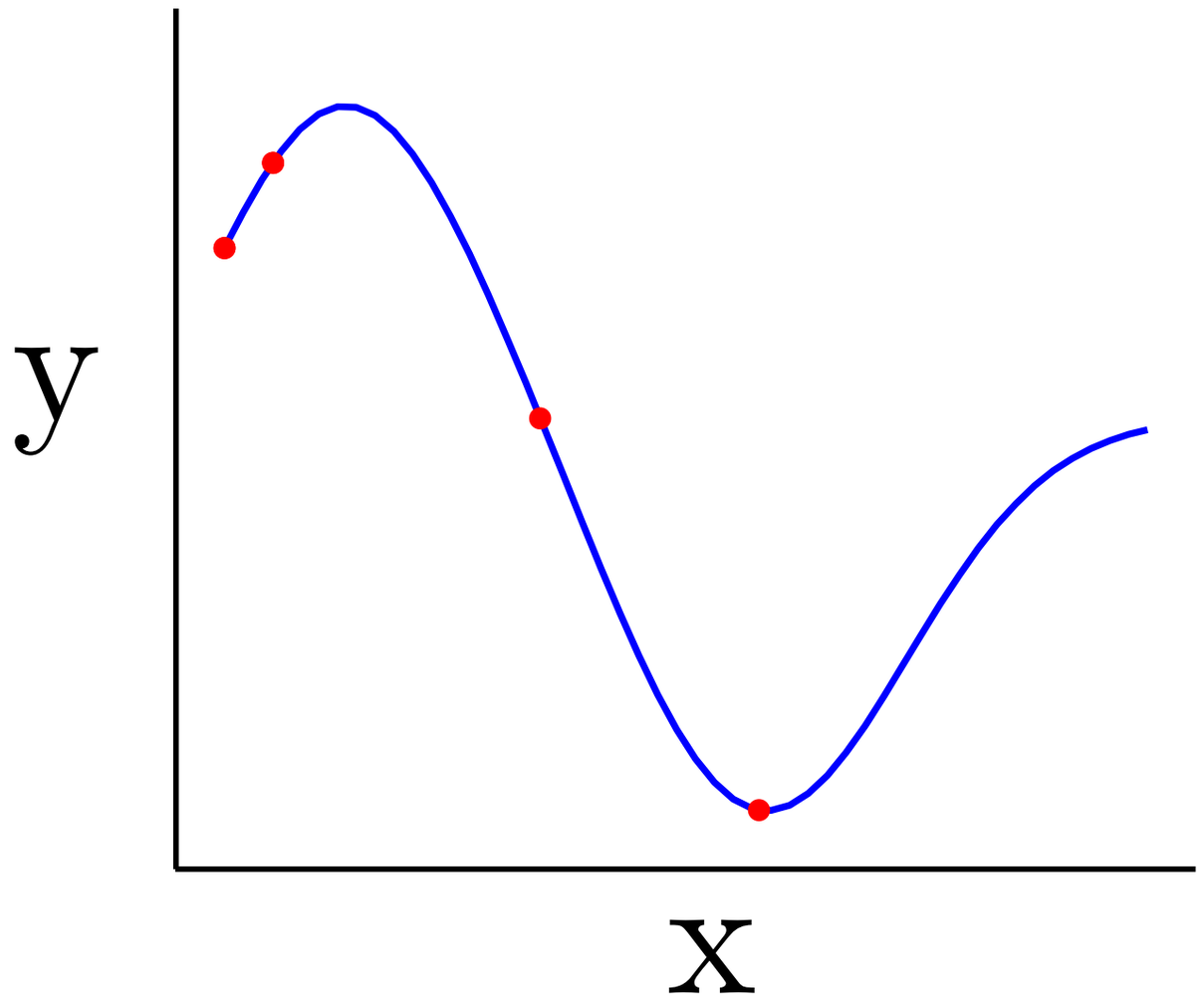


Which x location would you select next?

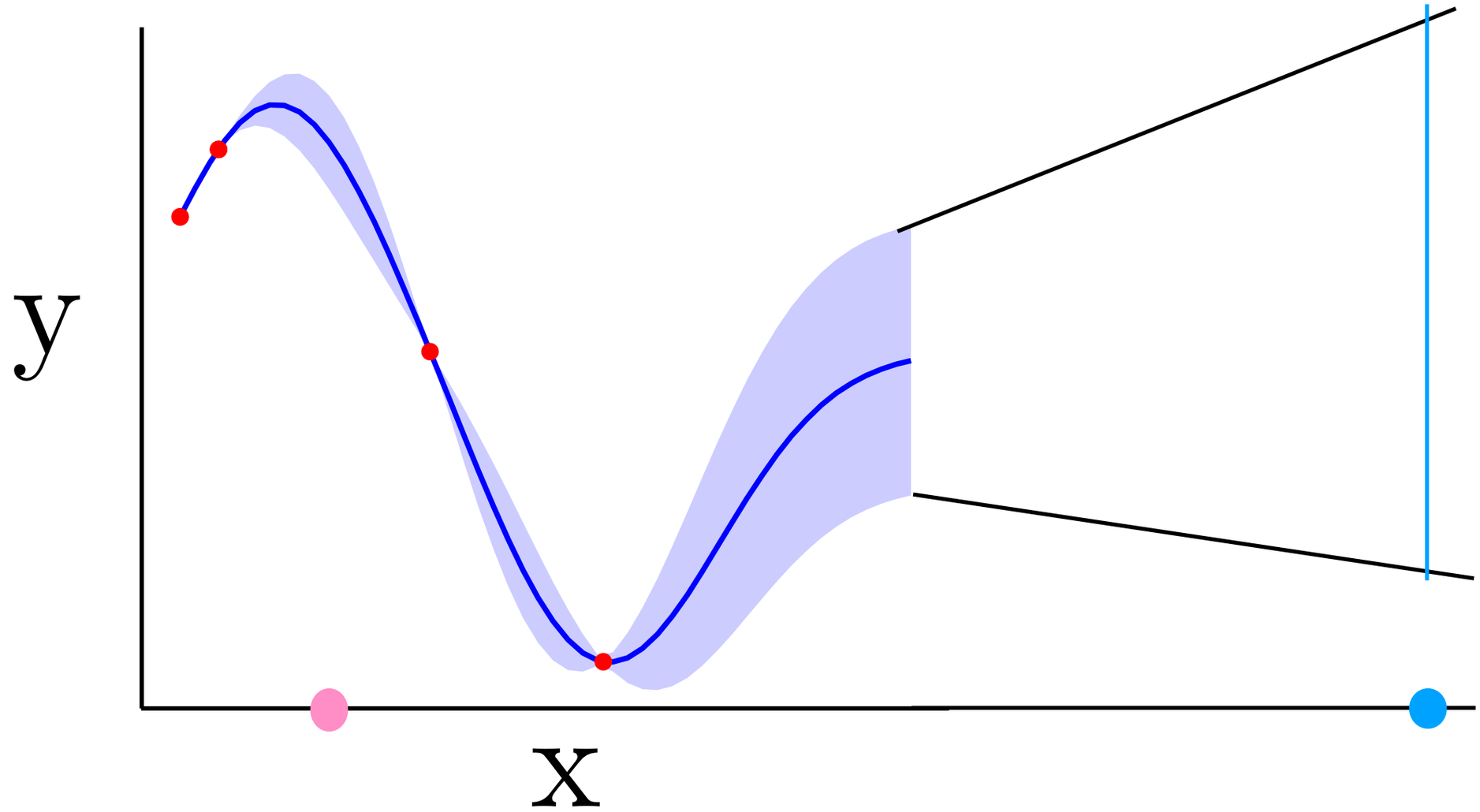
Non-linear regression



Non-linear regression



Non-linear regression with uncertainty



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

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

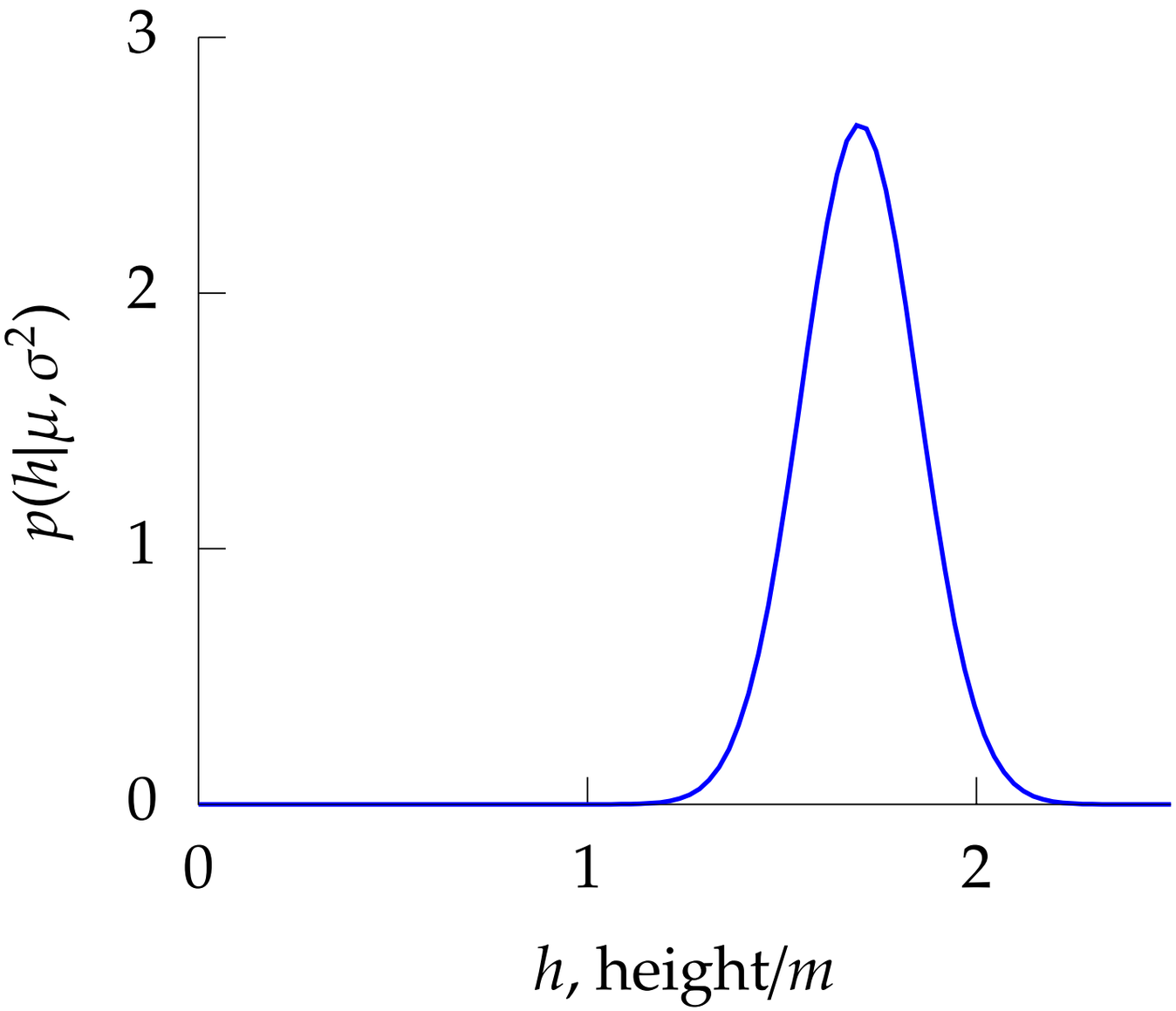
Gaussian Density

Perhaps the most common probability density

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

σ^2 is the variance of the density and μ 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.

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

Two Important Gaussian Properties

Sum of Gaussians

- Sum of Gaussian variables is also Gaussian.

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

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.

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

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.

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

Two Important Gaussian Properties

Scaling a Gaussian

- Scaling a Gaussian leads to a Gaussian.

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

And the scaled density is distributed as

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

Multivariate Consequence

- If

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

Multivariate Consequence

- If

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

- And

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

Multivariate Consequence

- If

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

- And

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

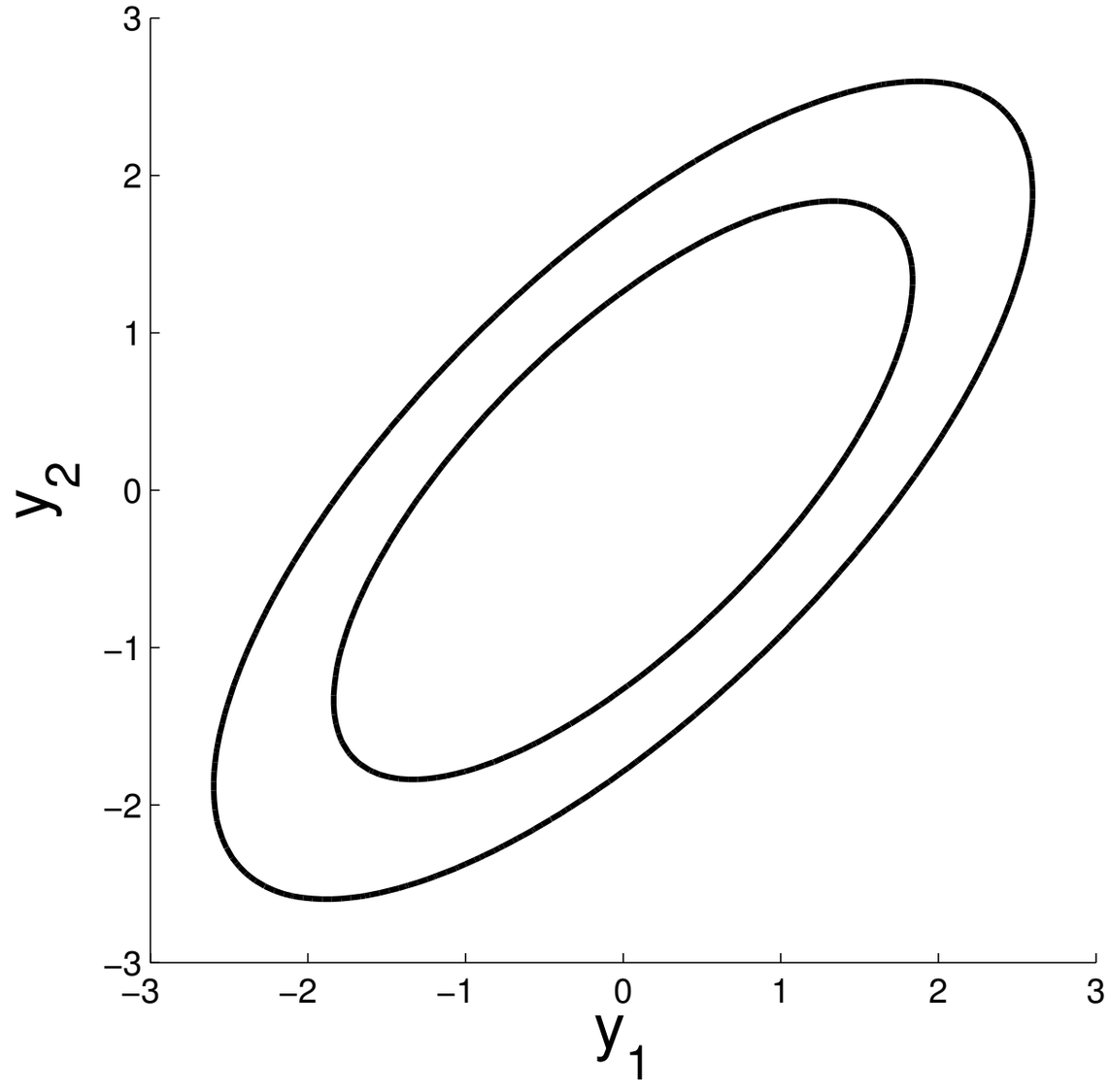
- Then

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

Gaussian Distribution

$$\Sigma_{i,j} = \mathbb{E}\left[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right]$$

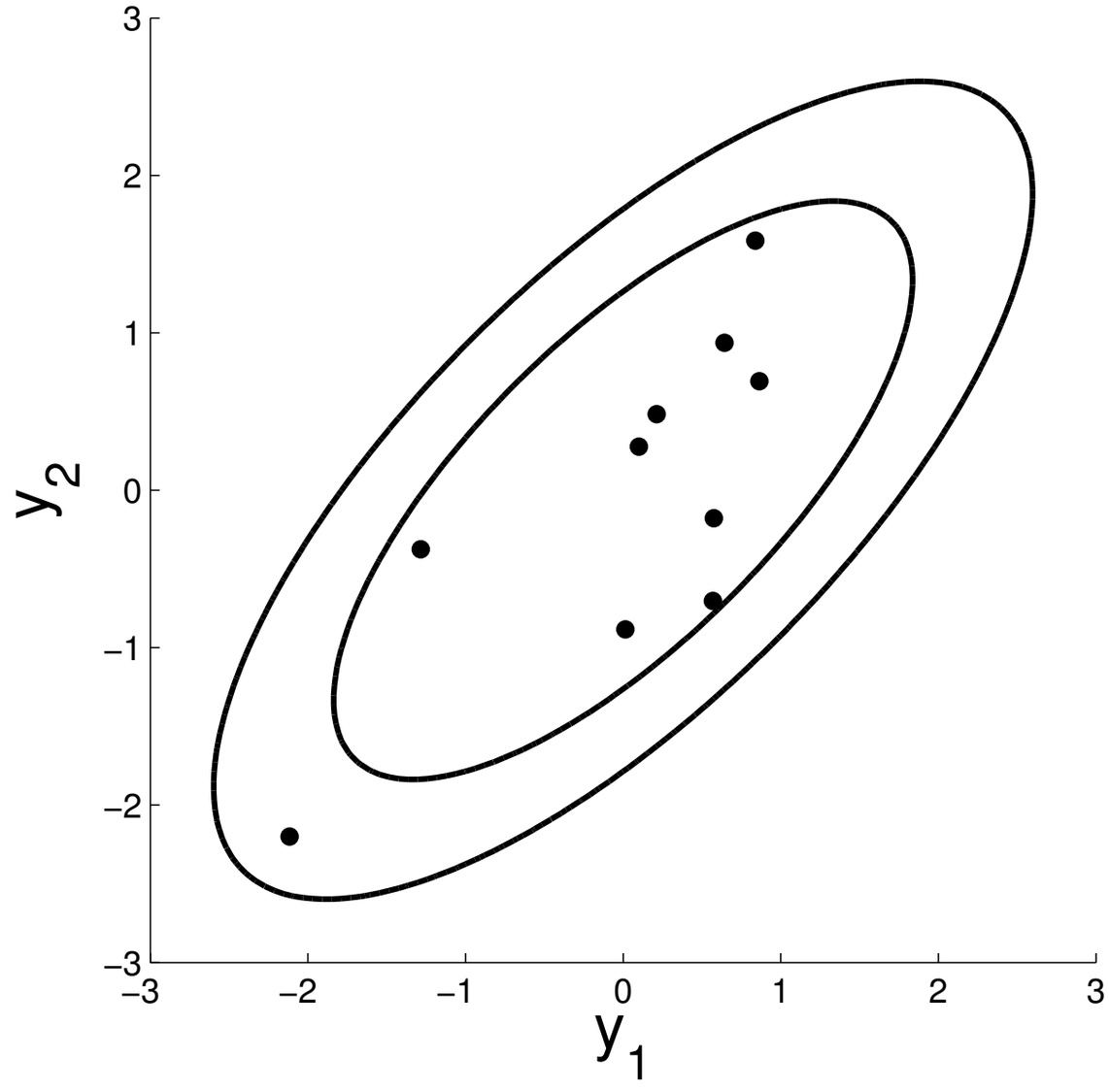
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\tfrac{1}{2}\mathbf{y}^\mathsf{T}\Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$$



Gaussian Distribution

$$\Sigma_{i,j} = \mathbb{E}\left[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right]$$

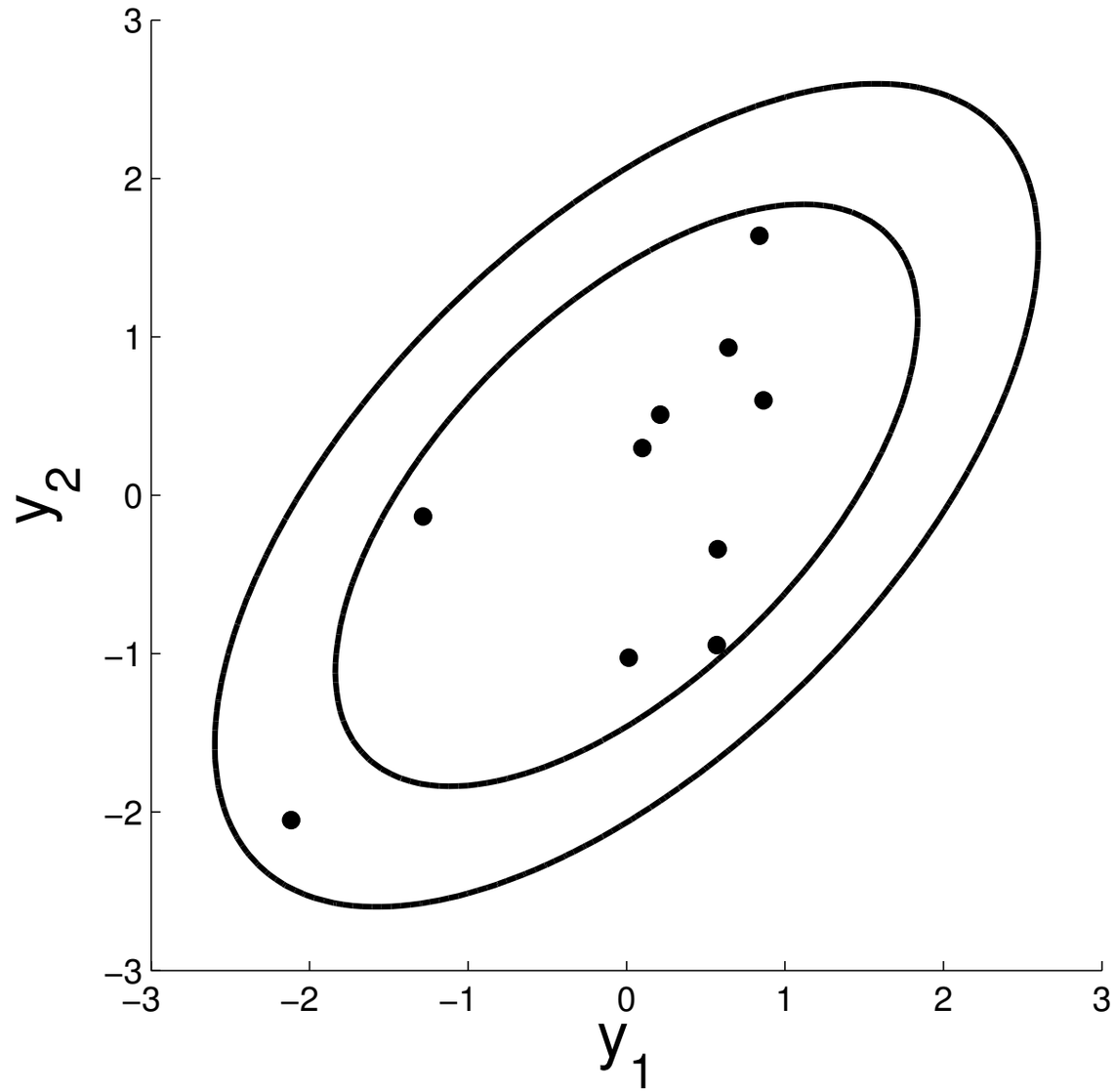
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\tfrac{1}{2}\mathbf{y}^\mathsf{T}\Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$$



Gaussian Distribution

$$\Sigma_{i,j} = \mathbb{E}\left[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right]$$

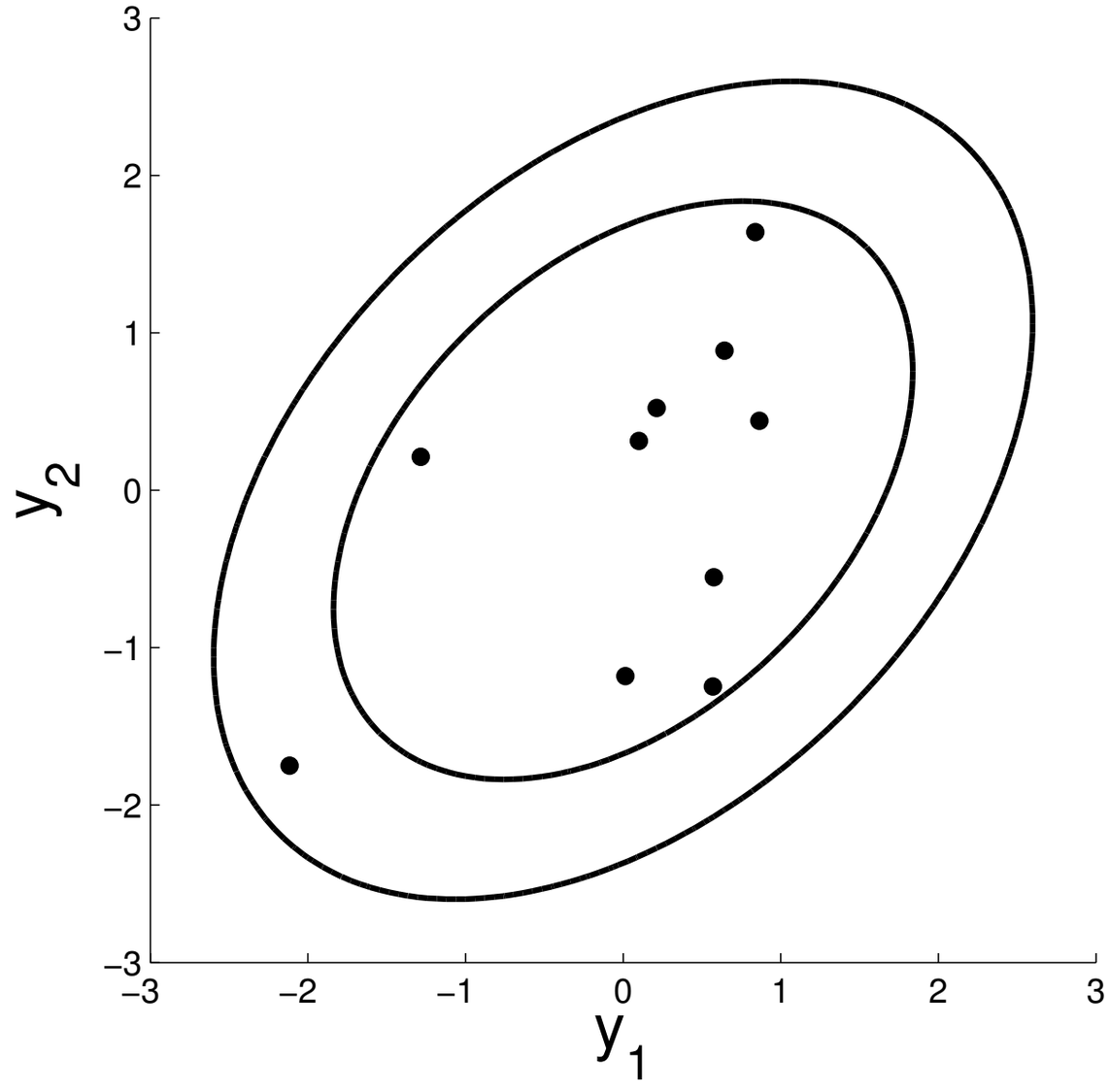
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\tfrac{1}{2}\mathbf{y}^\mathsf{T}\Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}$$



Gaussian Distribution

$$\Sigma_{i,j} = \mathbb{E}\left[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right]$$

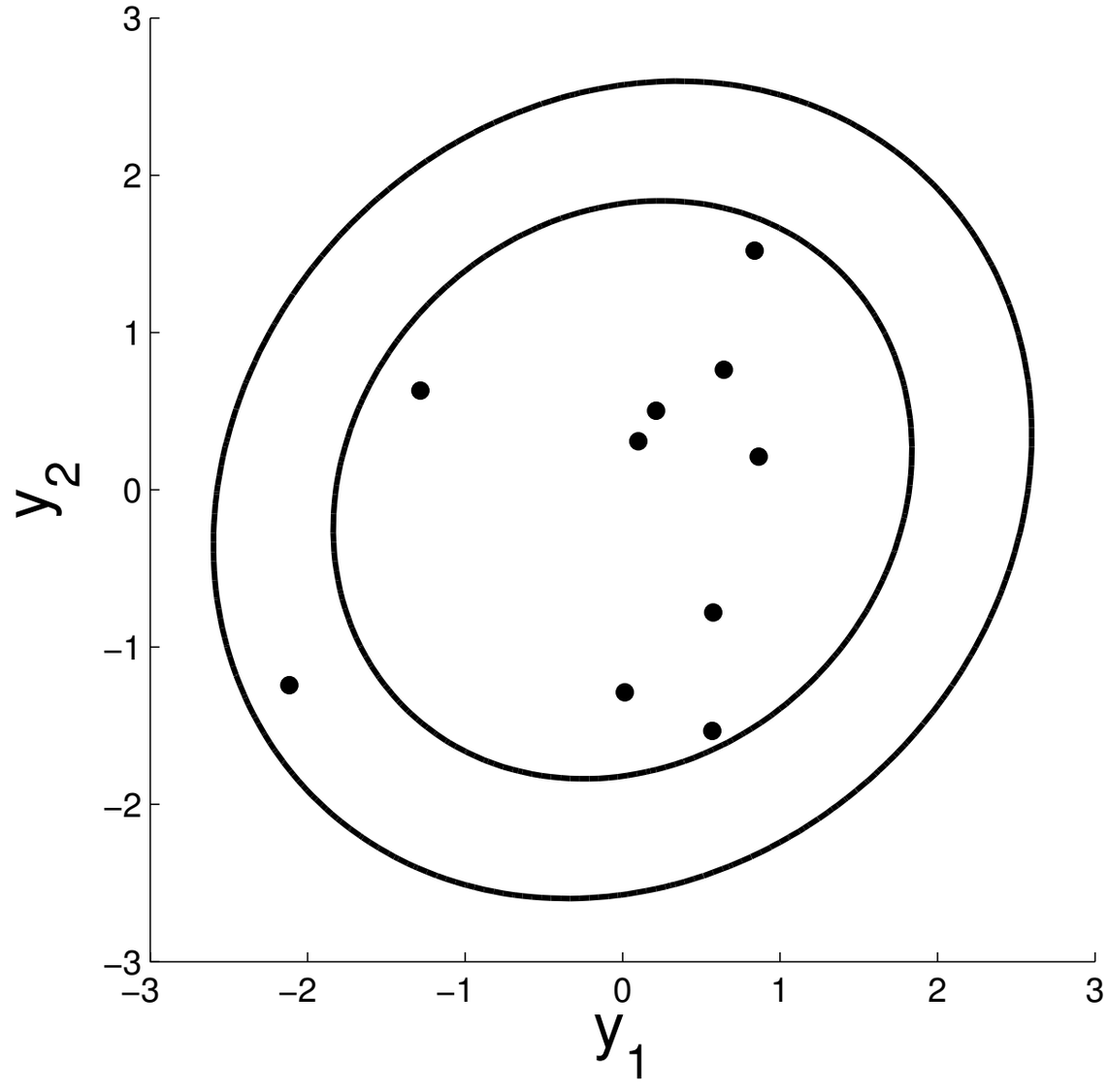
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\tfrac{1}{2}\mathbf{y}^\mathsf{T}\Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$$



Gaussian Distribution

$$\Sigma_{i,j} = \mathbb{E}\left[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right]$$

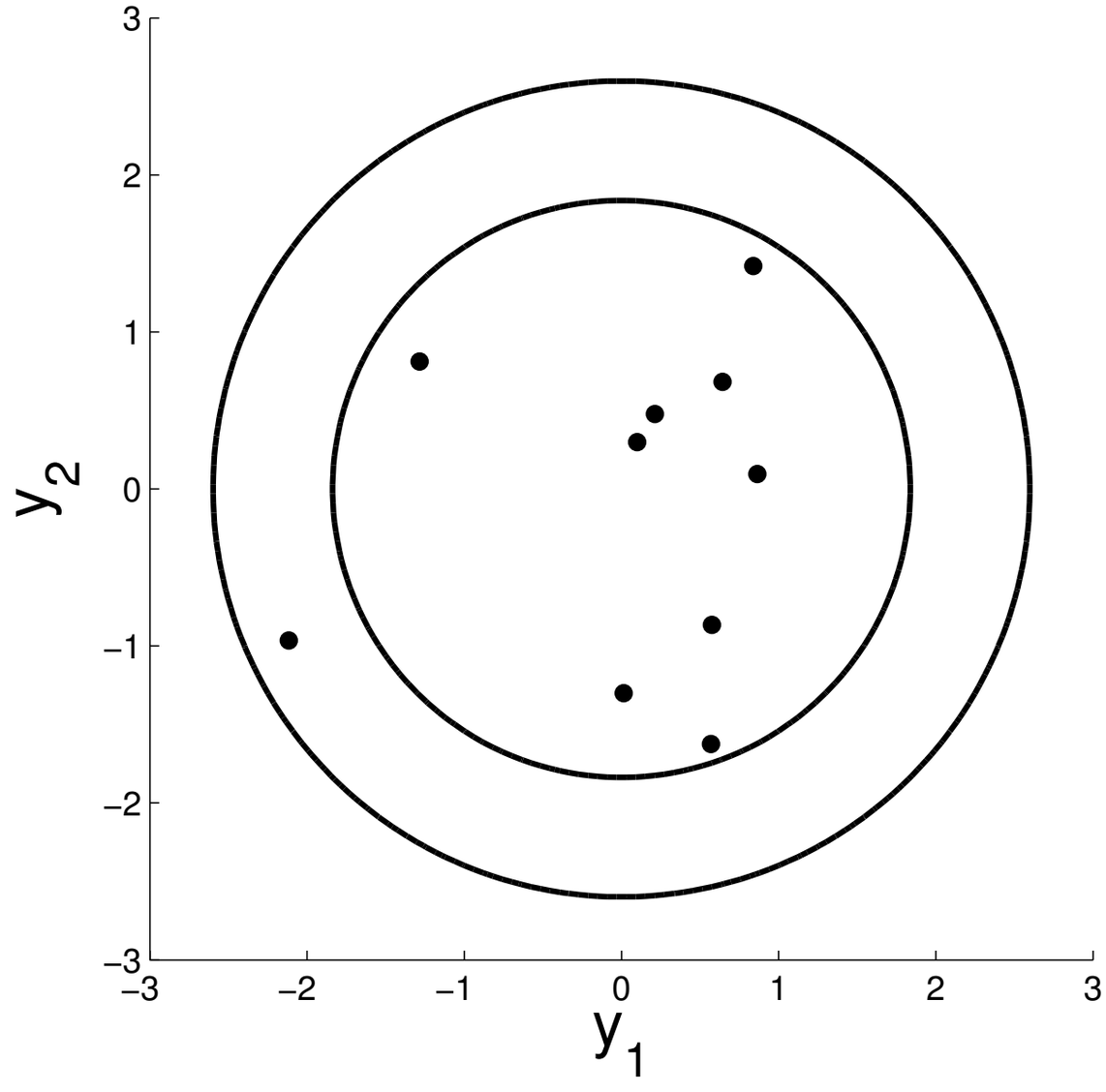
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\tfrac{1}{2}\mathbf{y}^\mathsf{T}\Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{bmatrix} 1 & .1 \\ .1 & 1 \end{bmatrix}$$



Gaussian Distribution

$$\Sigma_{i,j} = \mathbb{E}\left[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right]$$

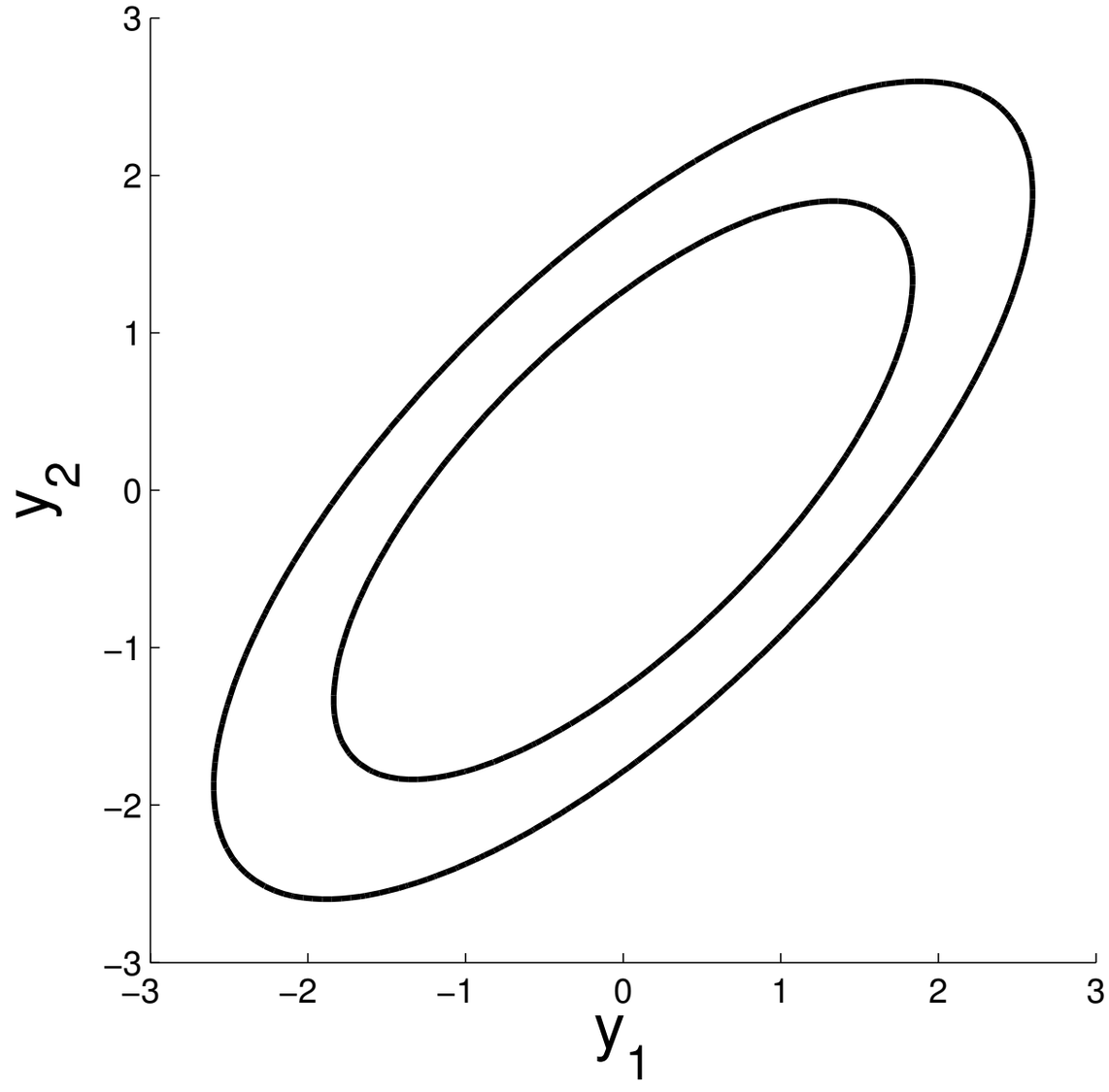
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\tfrac{1}{2}\mathbf{y}^\mathsf{T}\Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Gaussian Distribution

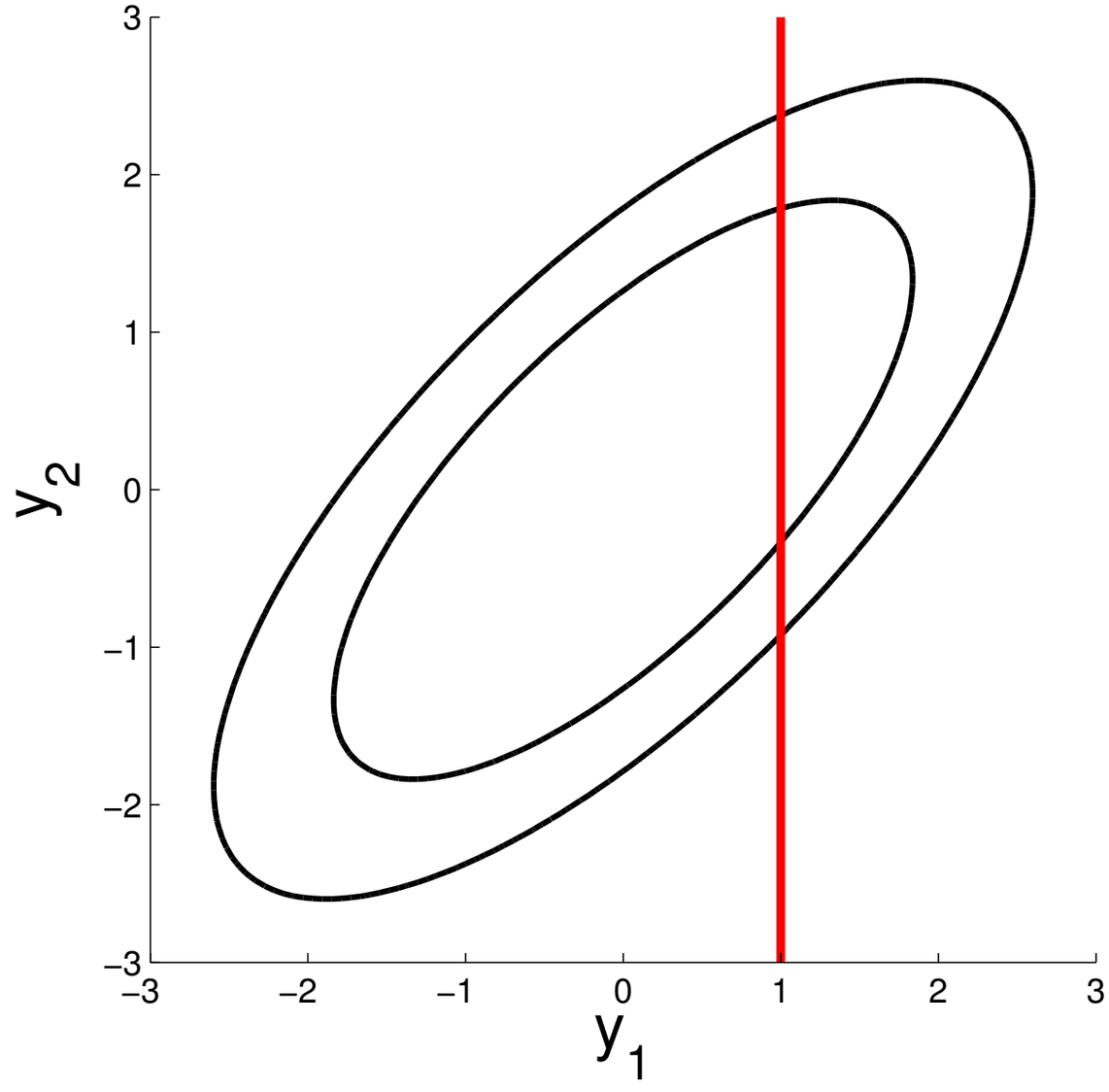
$$\Sigma_{i,j} = \mathbb{E}\left[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right]$$

$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\tfrac{1}{2}\mathbf{y}^\mathsf{T}\Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$$



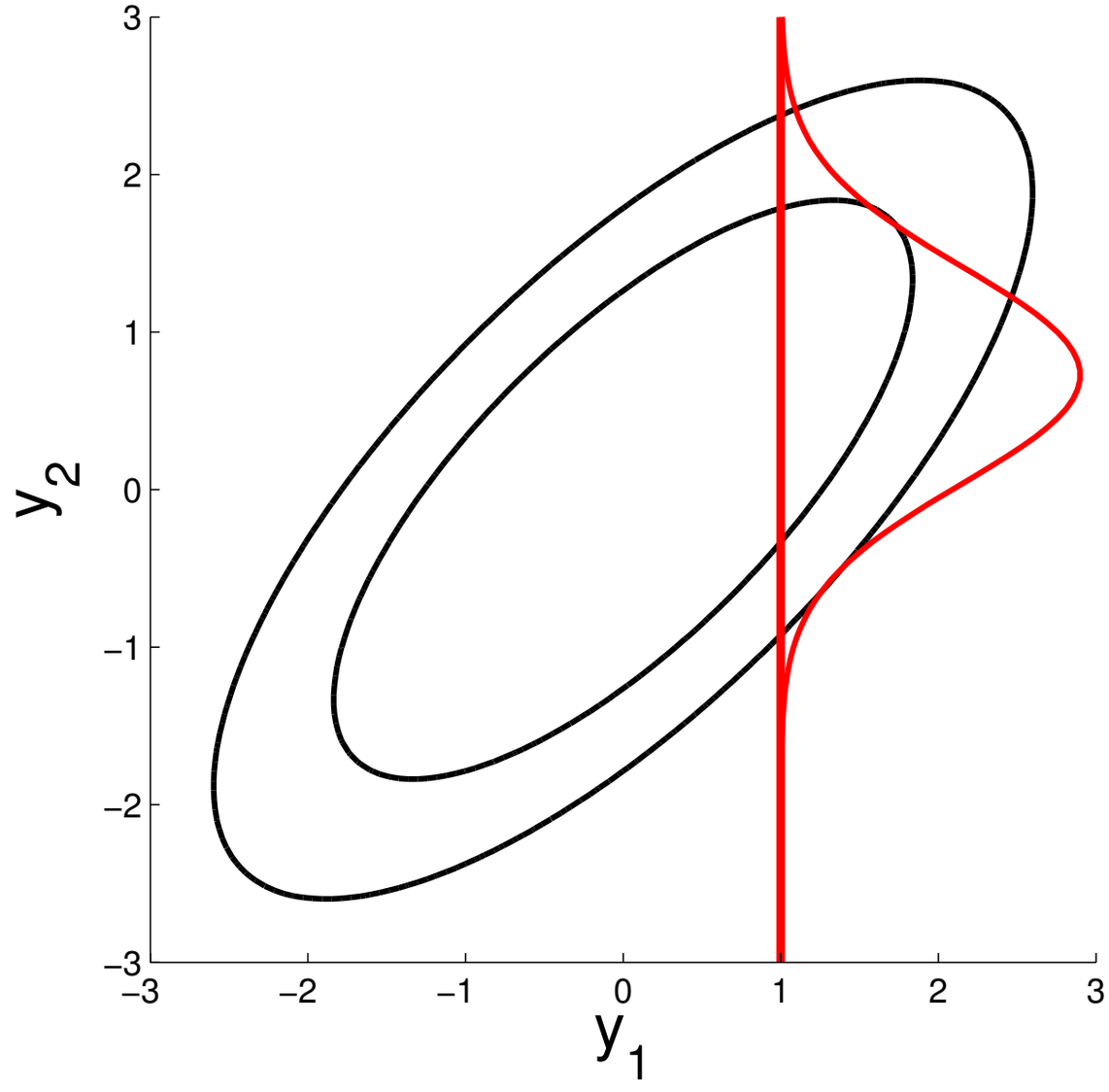
Gaussian distribution - Conditioning

$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\tfrac{1}{2}\mathbf{y}^\mathsf{T}\Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{bmatrix} 1 & .7 \\ .7 & 1 \end{bmatrix}$$



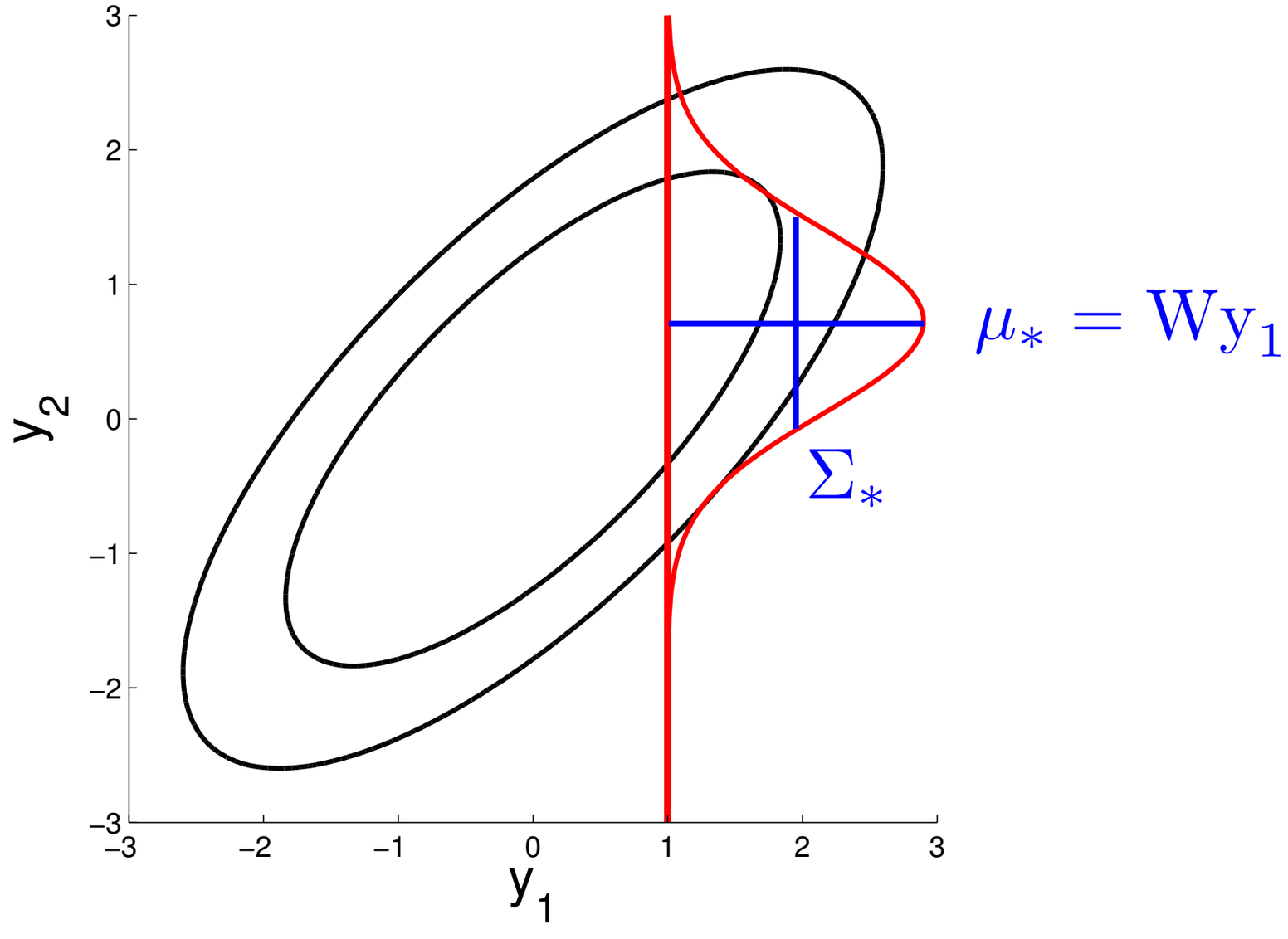
Gaussian distribution - Conditioning

$$p(\mathsf{y}_2|\mathsf{y}_1, \Sigma) \propto \exp\left(-\tfrac{1}{2}(\mathsf{y}_2 - \mu_*){\Sigma_*}^{-1}(\mathsf{y}_2 - \mu_*)\right)$$



Gaussian distribution - Conditioning

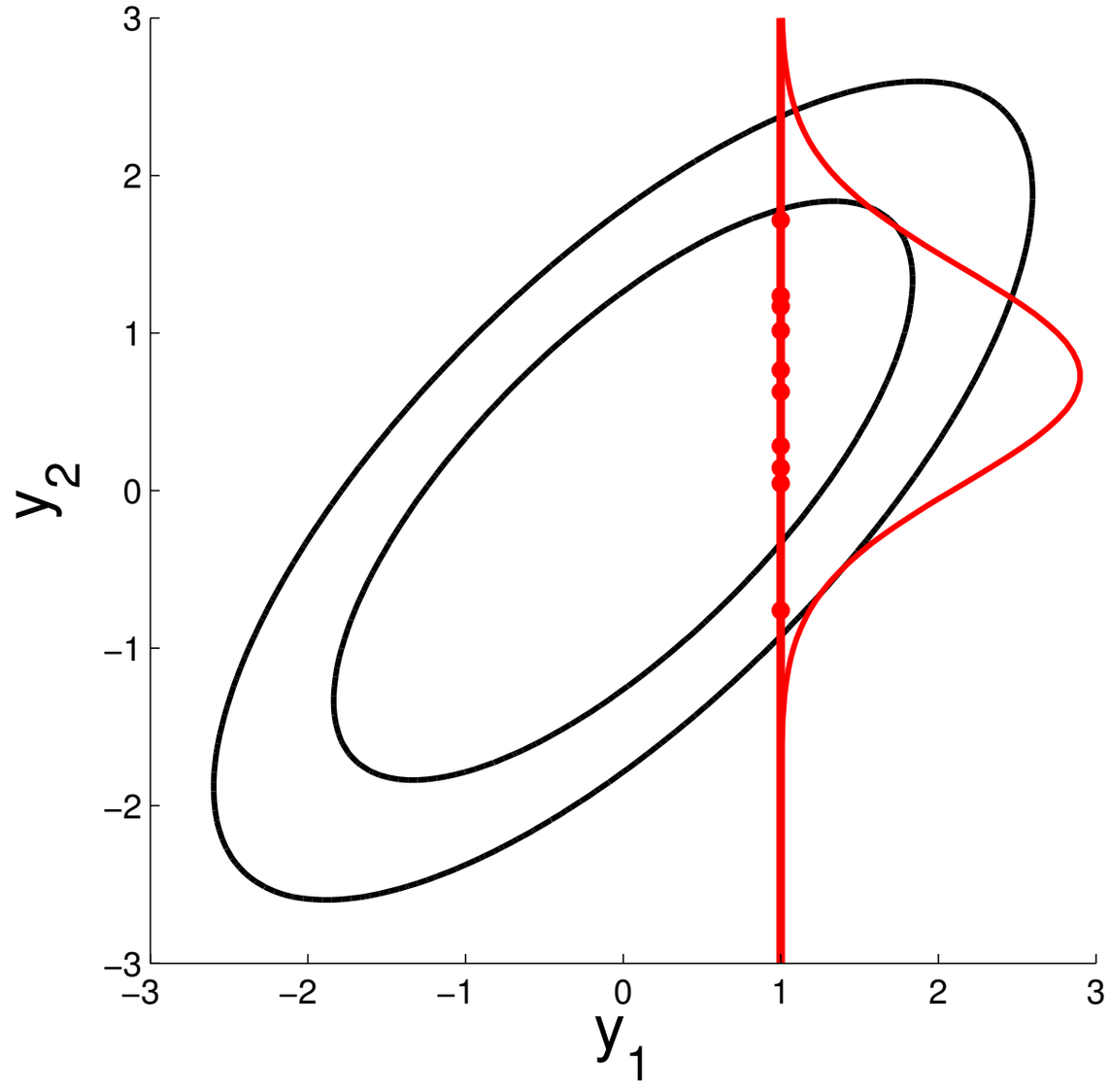
$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\tfrac{1}{2}(\mathbf{y}_2 - \mu_*){\Sigma_*}^{-1}(\mathbf{y}_2 - \mu_*)\right)$$



There are closed form solutions for μ_*, Σ_*

Gaussian distribution - Conditioning

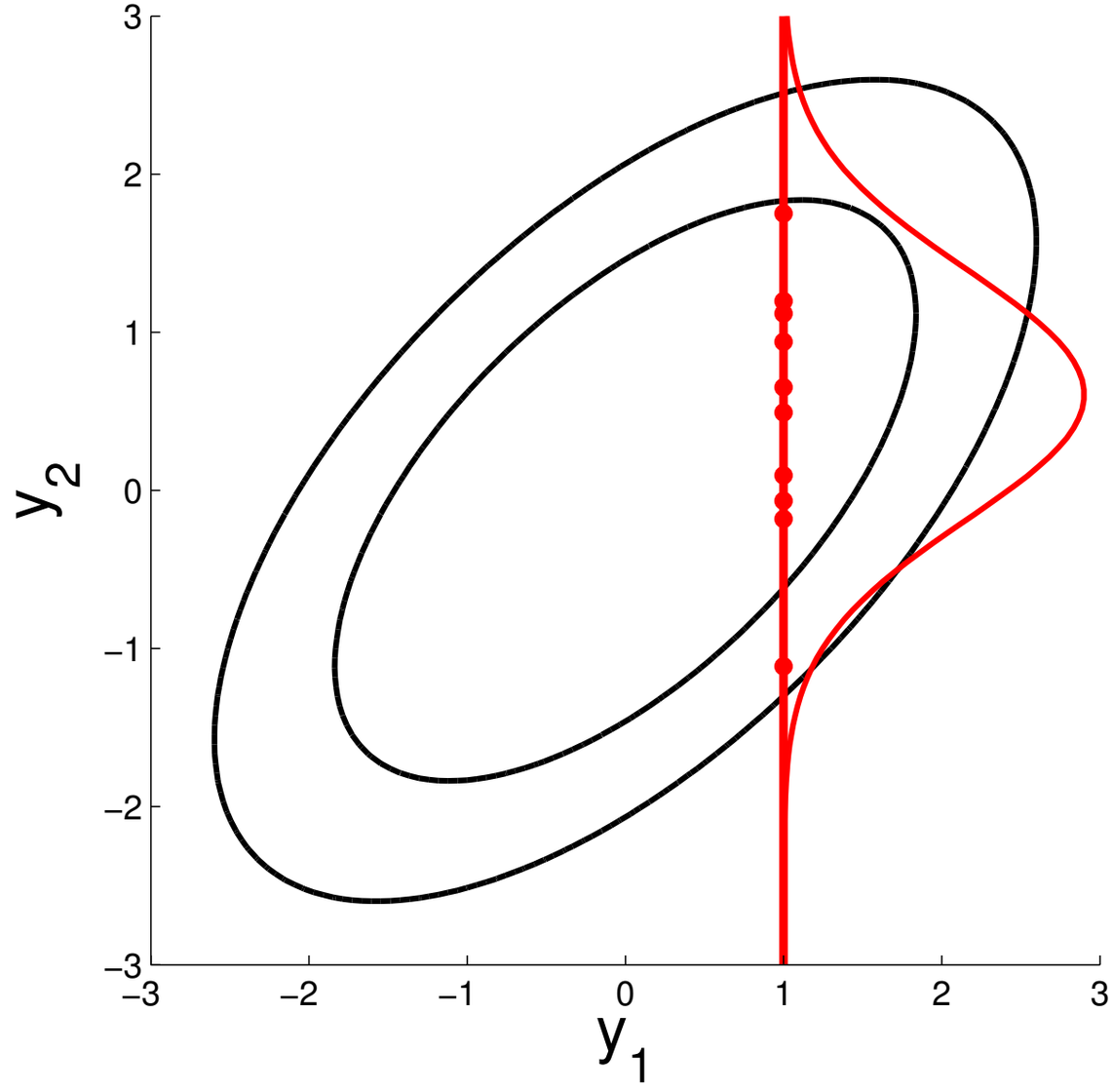
$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\tfrac{1}{2}(\mathbf{y}_2 - \mu_*){\Sigma_*}^{-1}(\mathbf{y}_2 - \mu_*)\right)$$



There are closed form solutions for μ_*, Σ_*

Gaussian distribution - Conditioning

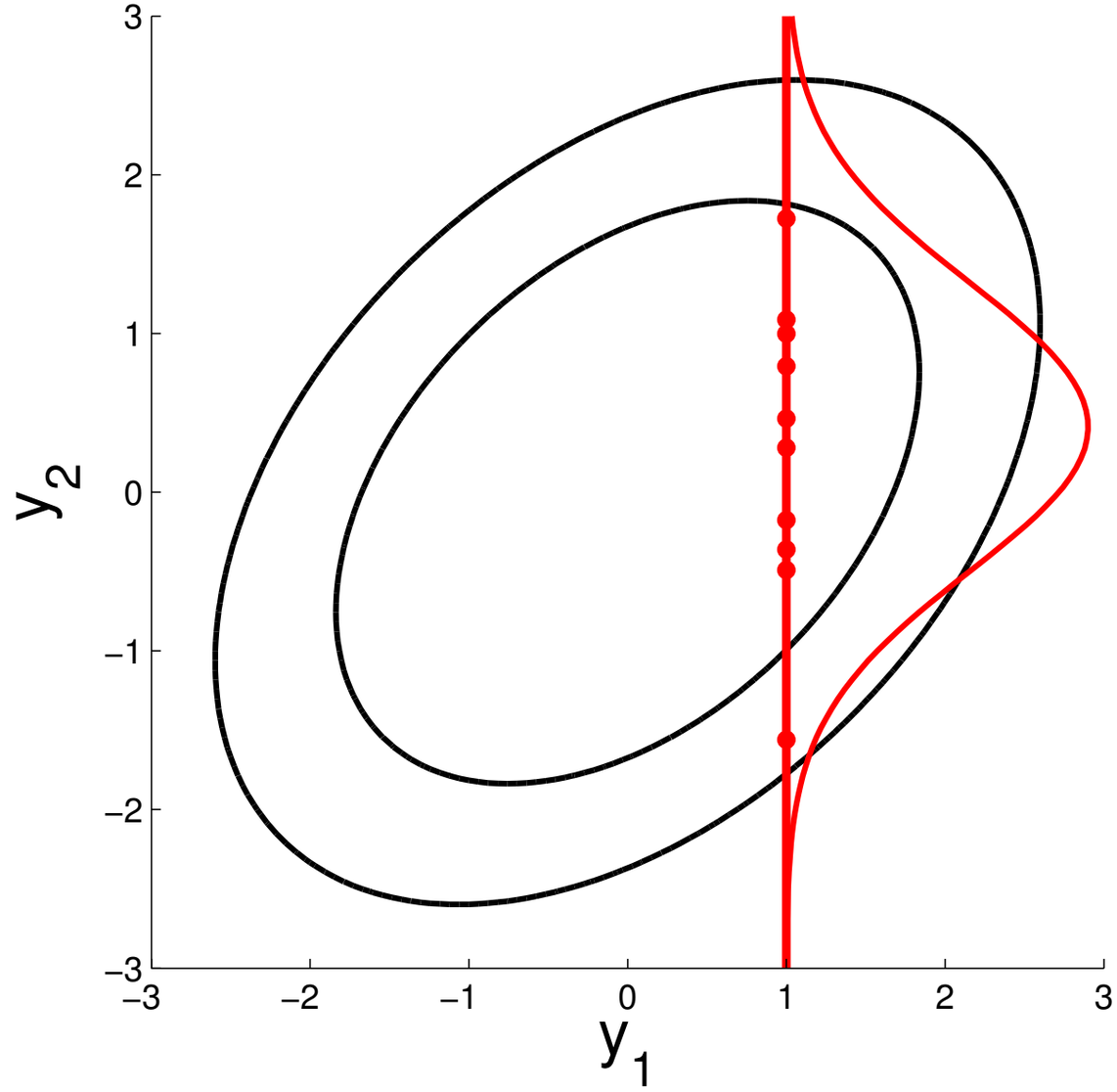
$$p(\mathsf{y}_2|\mathsf{y}_1, \Sigma) \propto \exp\left(-\tfrac{1}{2}(\mathsf{y}_2 - \mu_*){\Sigma_*}^{-1}(\mathsf{y}_2 - \mu_*)\right)$$



There are closed form solutions for μ_*, Σ_*

Gaussian distribution - Conditioning

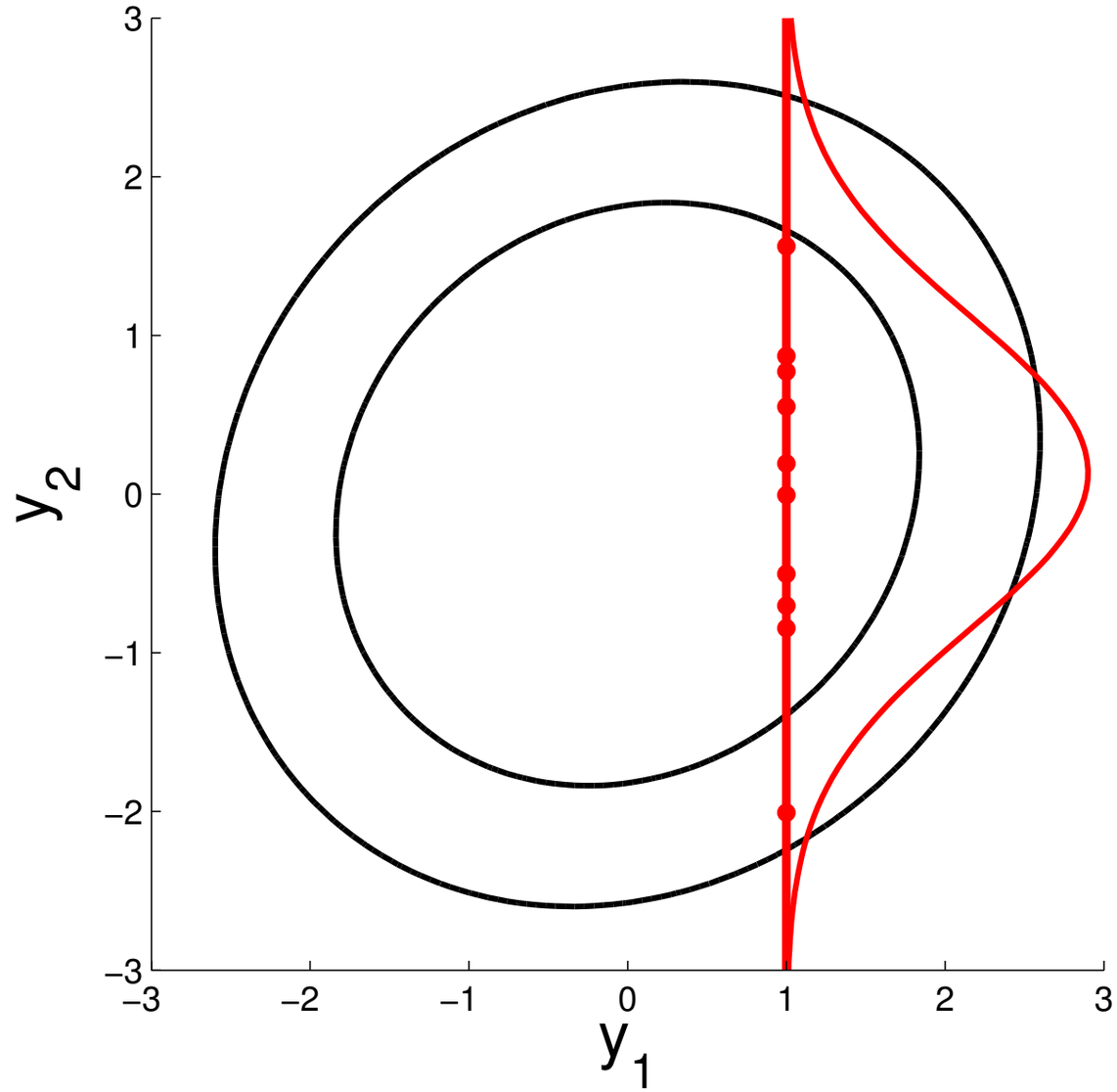
$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\tfrac{1}{2}(\mathbf{y}_2 - \mu_*){\Sigma_*}^{-1}(\mathbf{y}_2 - \mu_*)\right)$$



There are closed form solutions for μ_*, Σ_*

Gaussian distribution - Conditioning

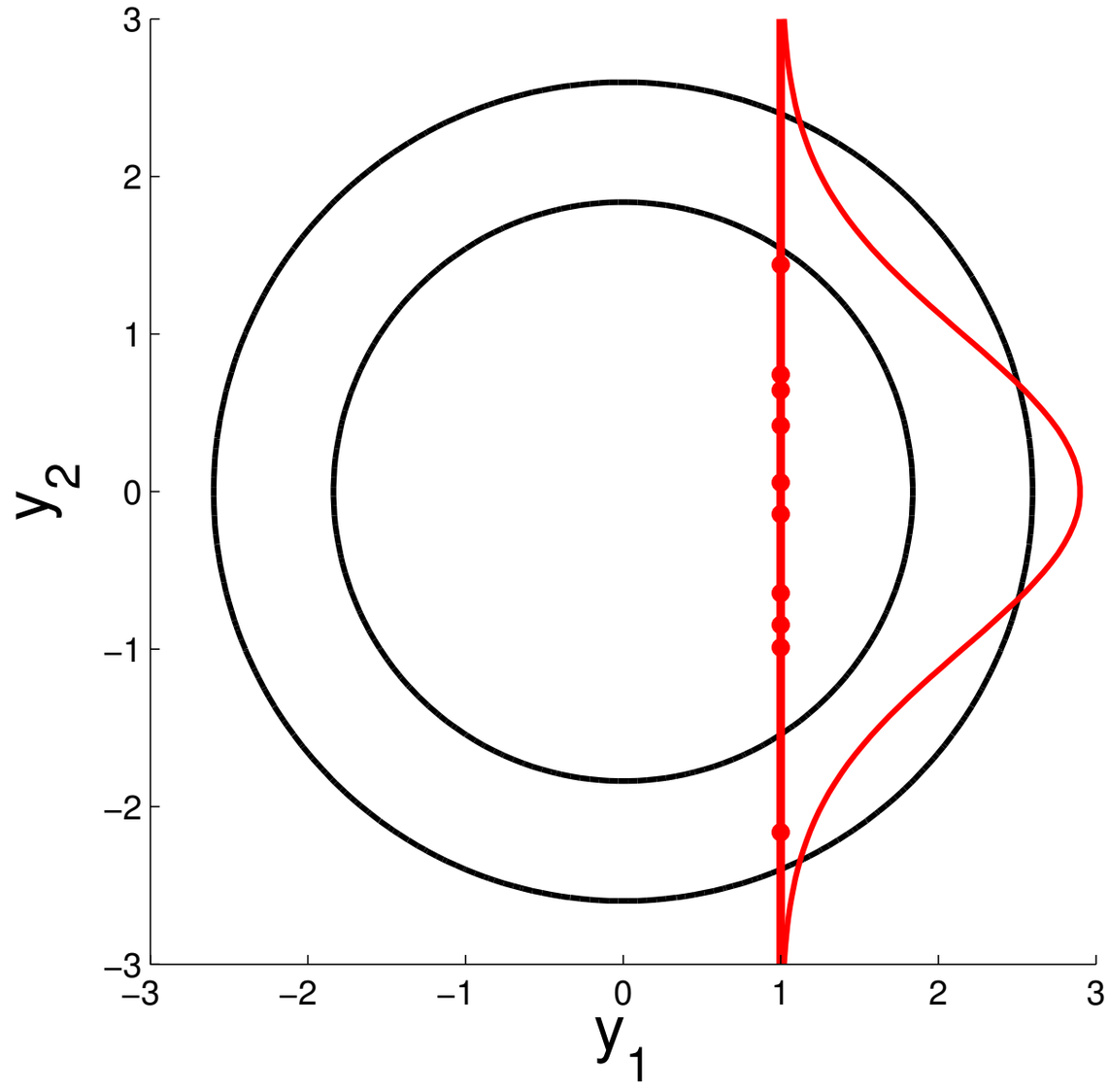
$$p(\mathsf{y}_2|\mathsf{y}_1, \Sigma) \propto \exp\left(-\tfrac{1}{2}(\mathsf{y}_2 - \mu_*){\Sigma_*}^{-1}(\mathsf{y}_2 - \mu_*)\right)$$



There are closed form solutions for μ_*, Σ_*

Gaussian distribution - Conditioning

$$p(\mathbf{y}_2|\mathbf{y}_1, \Sigma) \propto \exp\left(-\tfrac{1}{2}(\mathbf{y}_2 - \mu_*){\Sigma_*}^{-1}(\mathbf{y}_2 - \mu_*)\right)$$



There are closed form solutions for μ_*, Σ_*

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*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} \tag{4.12}$$

Then the marginals are given by

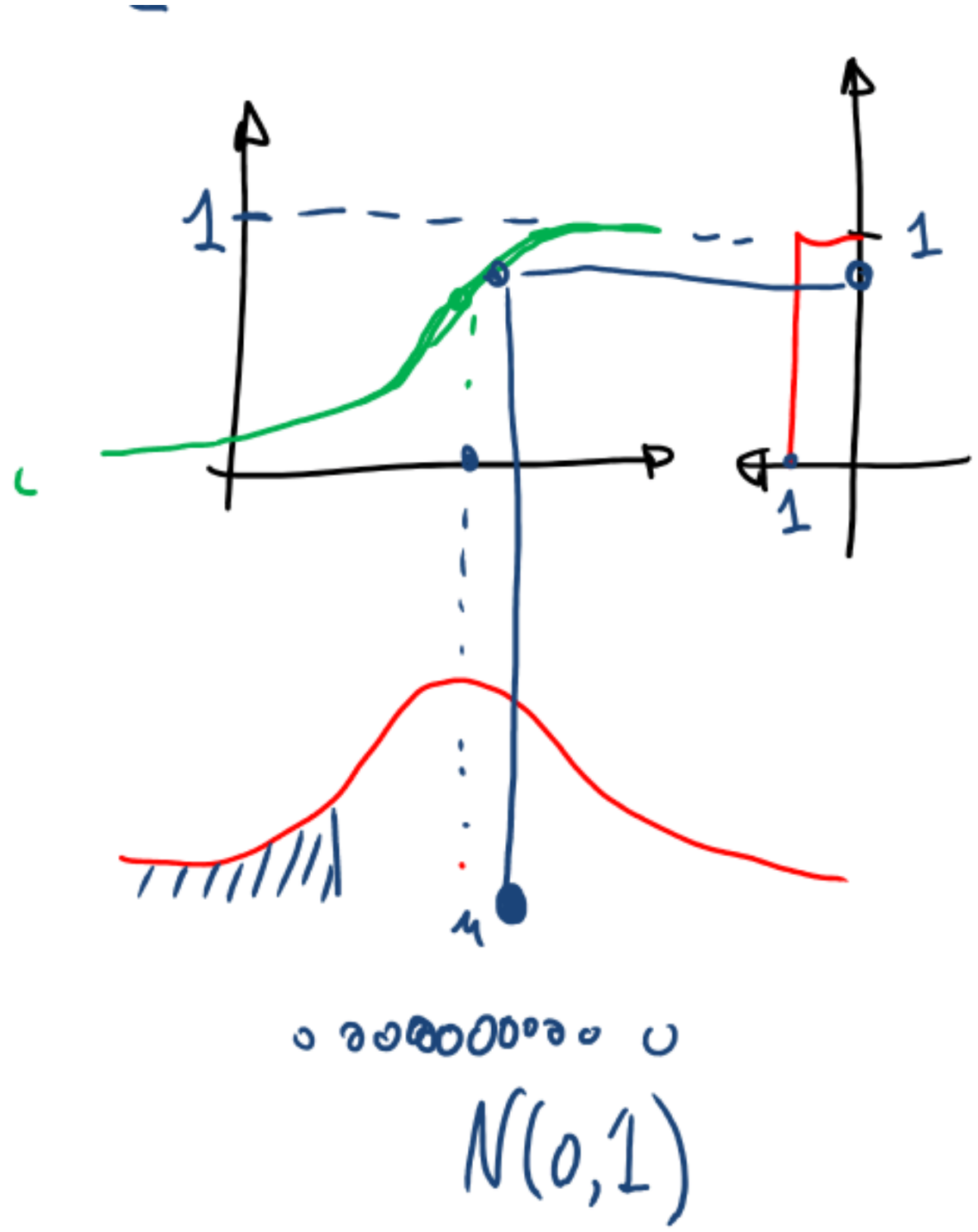
$$\begin{aligned} p(\mathbf{x}_1) &= \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ p(\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \end{aligned}$$

and the posterior conditional is given by

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

Sampling from a Gaussian density

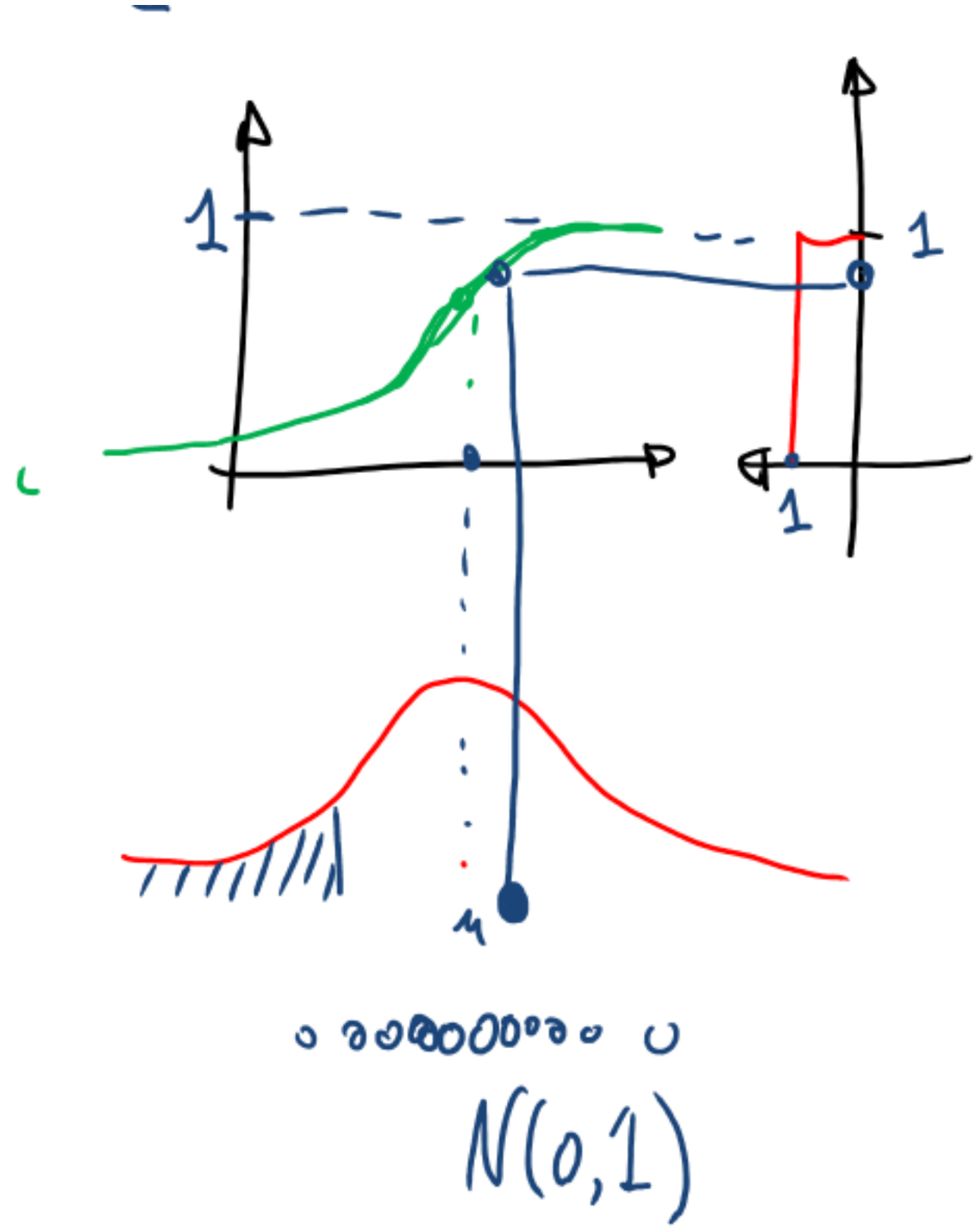
$$x_i \sim \mathcal{N}(0,1)$$



Sampling from a Gaussian density

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

$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$
$$\sim \mu + \sigma\mathcal{N}(0,1)$$



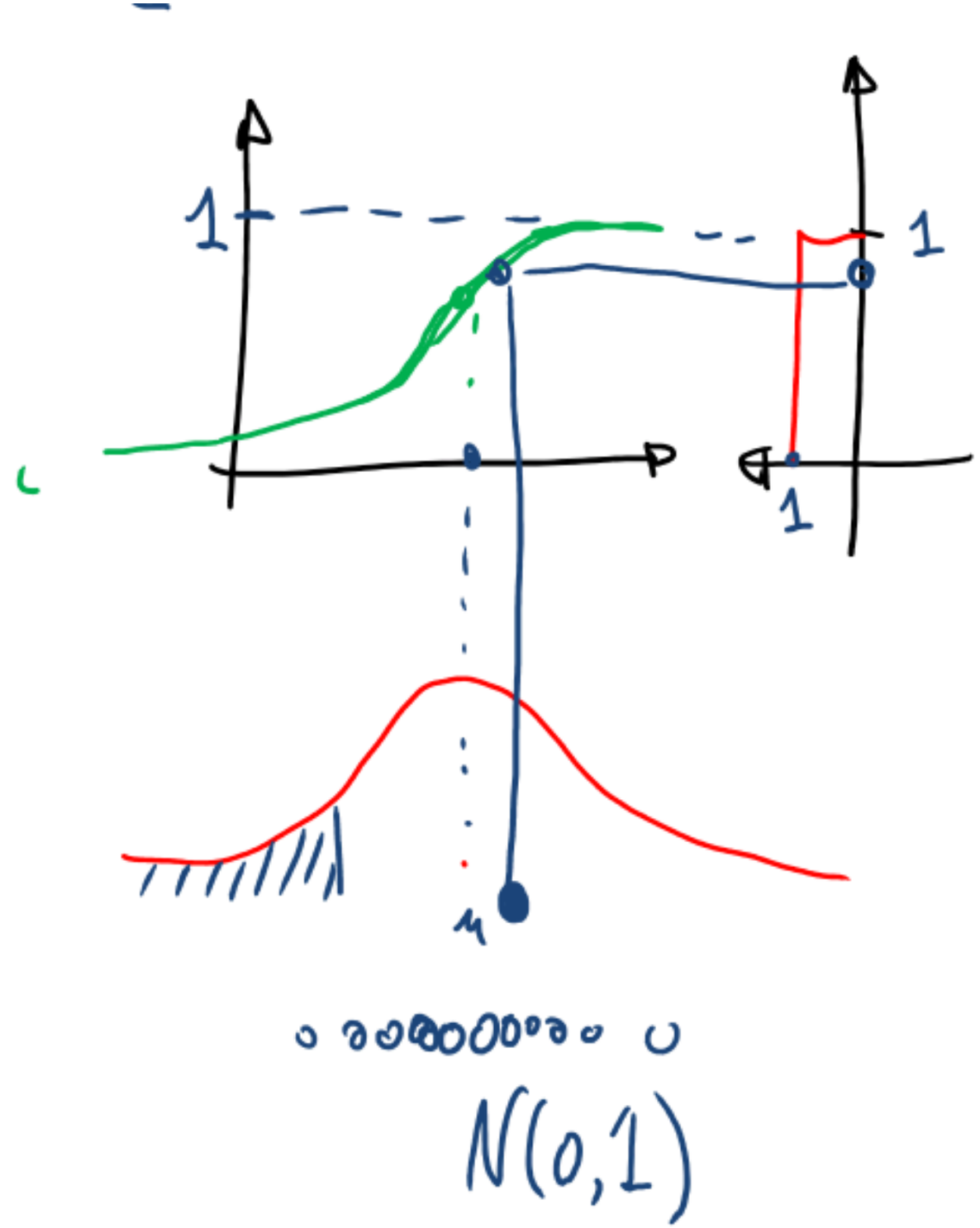
Sampling from a Gaussian density

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

$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$
$$\sim \mu + \sigma\mathcal{N}(0,1)$$

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

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



Sampling from a Gaussian density

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

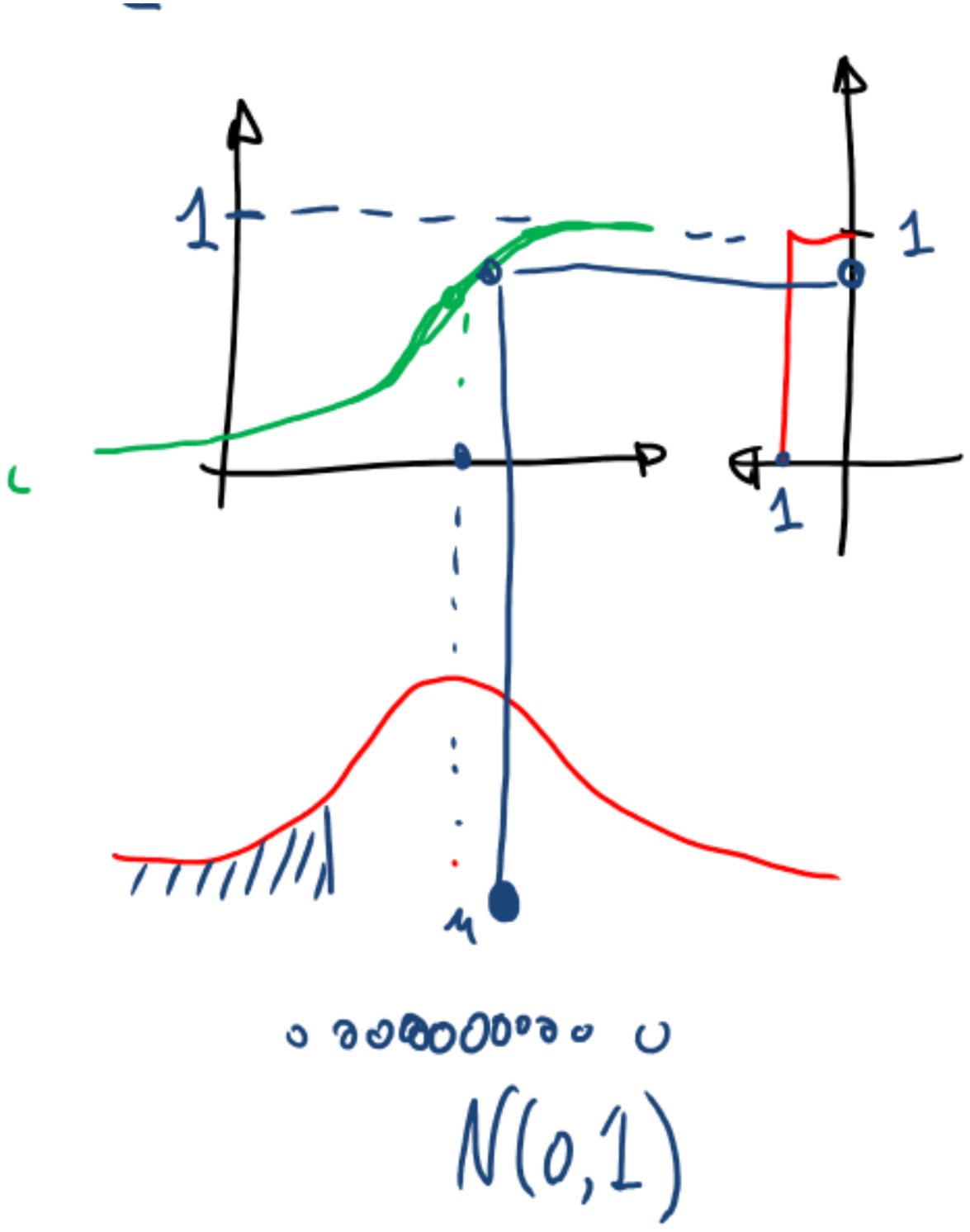
$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$
$$\sim \mu + \sigma\mathcal{N}(0,1)$$

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

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

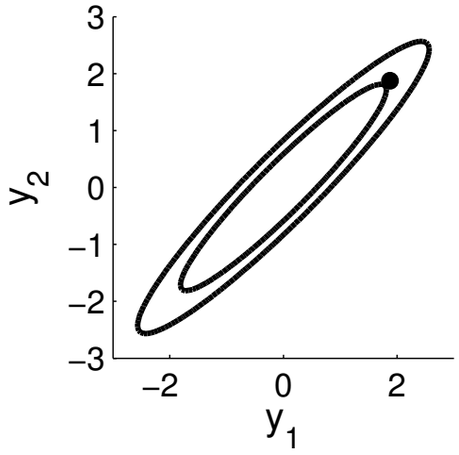
x μ Σ

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



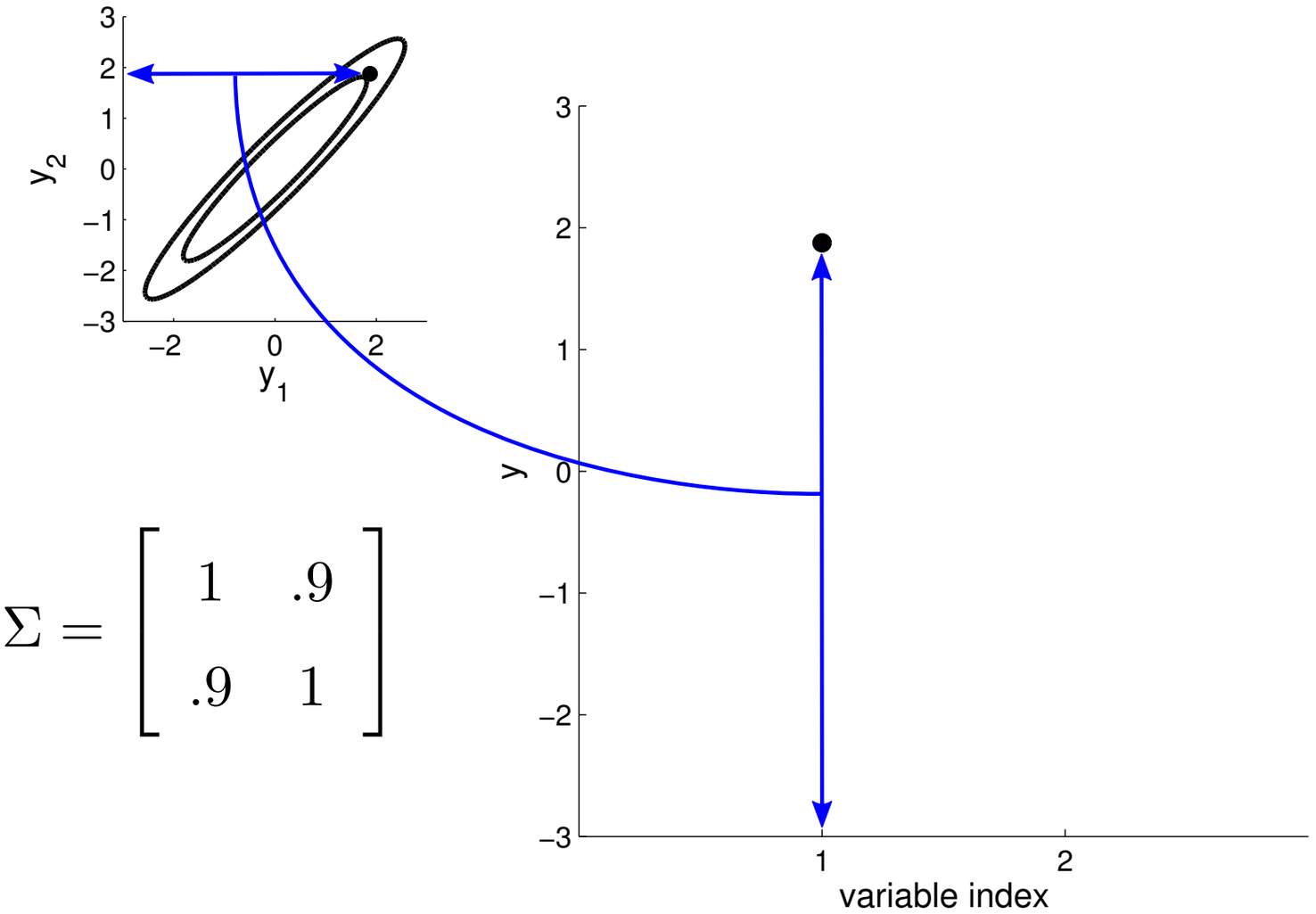
Cholesky decomposition $\Sigma = LL^T$

Towards higher dimensional gaussians - New Visualisation



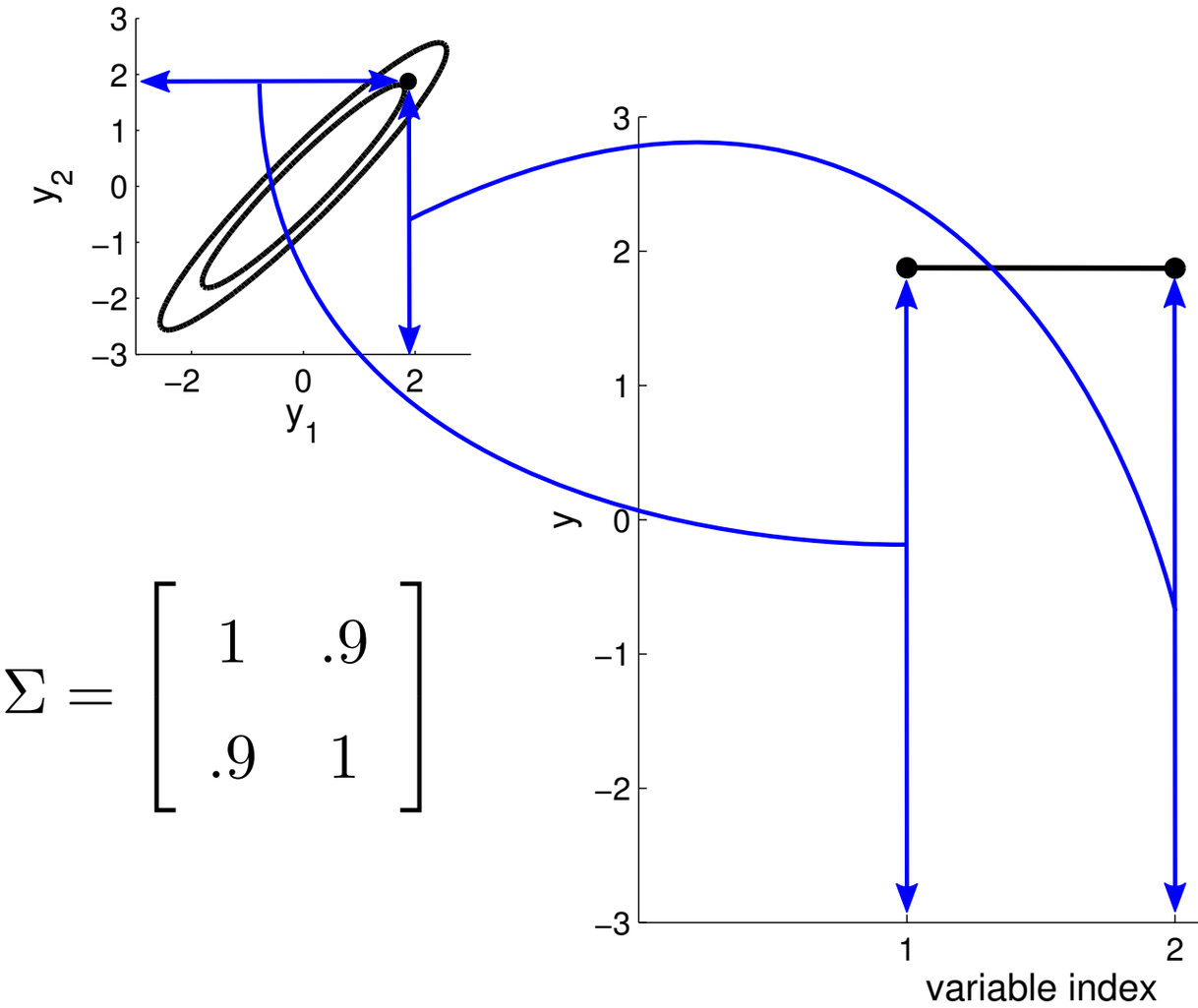
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New Visualisation



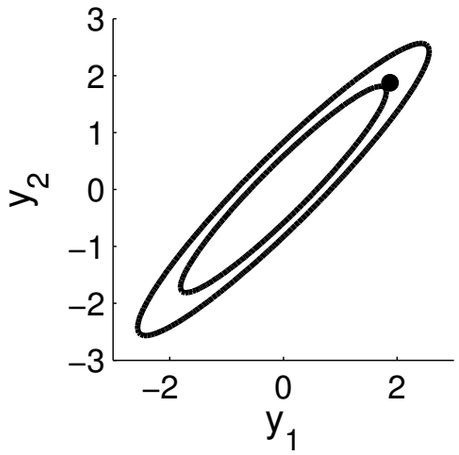
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New Visualisation

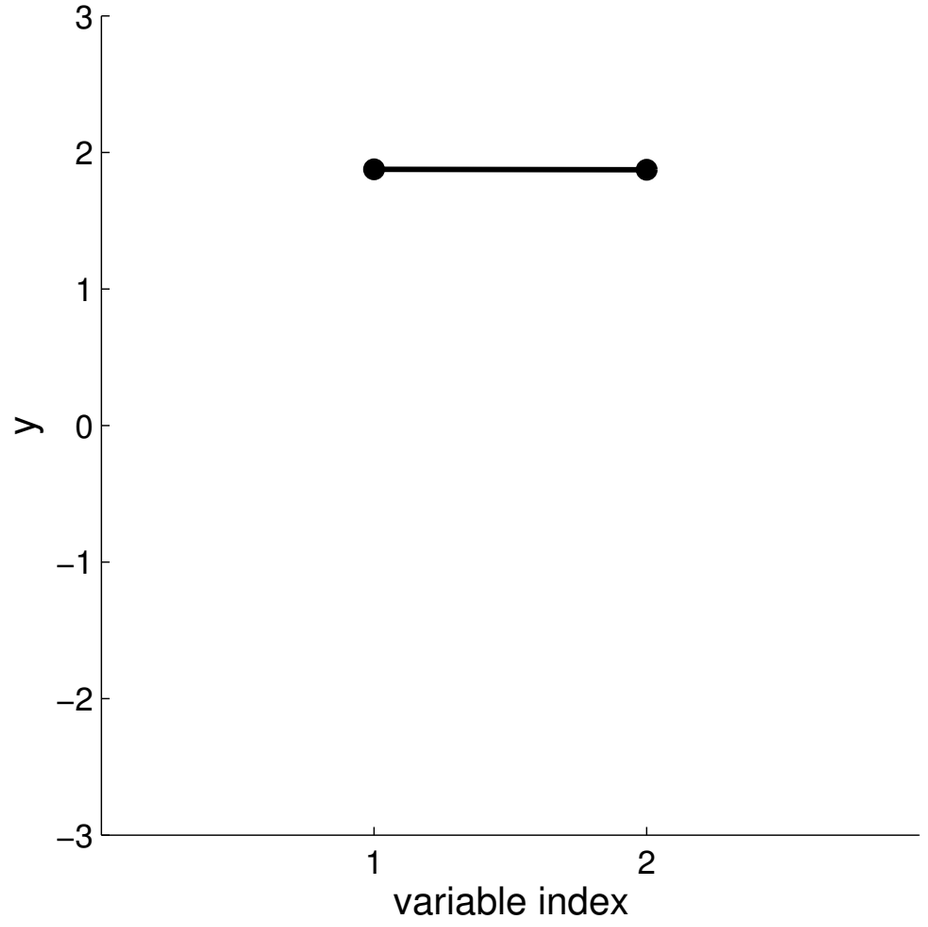


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

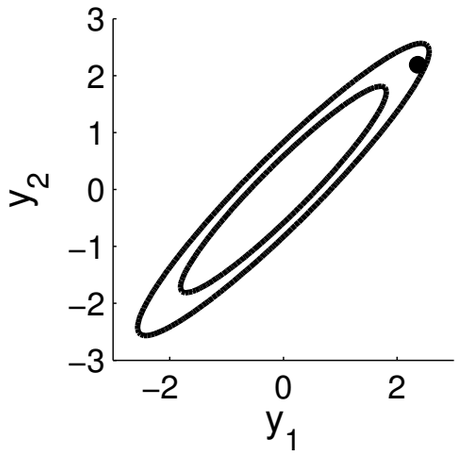
New Visualisation



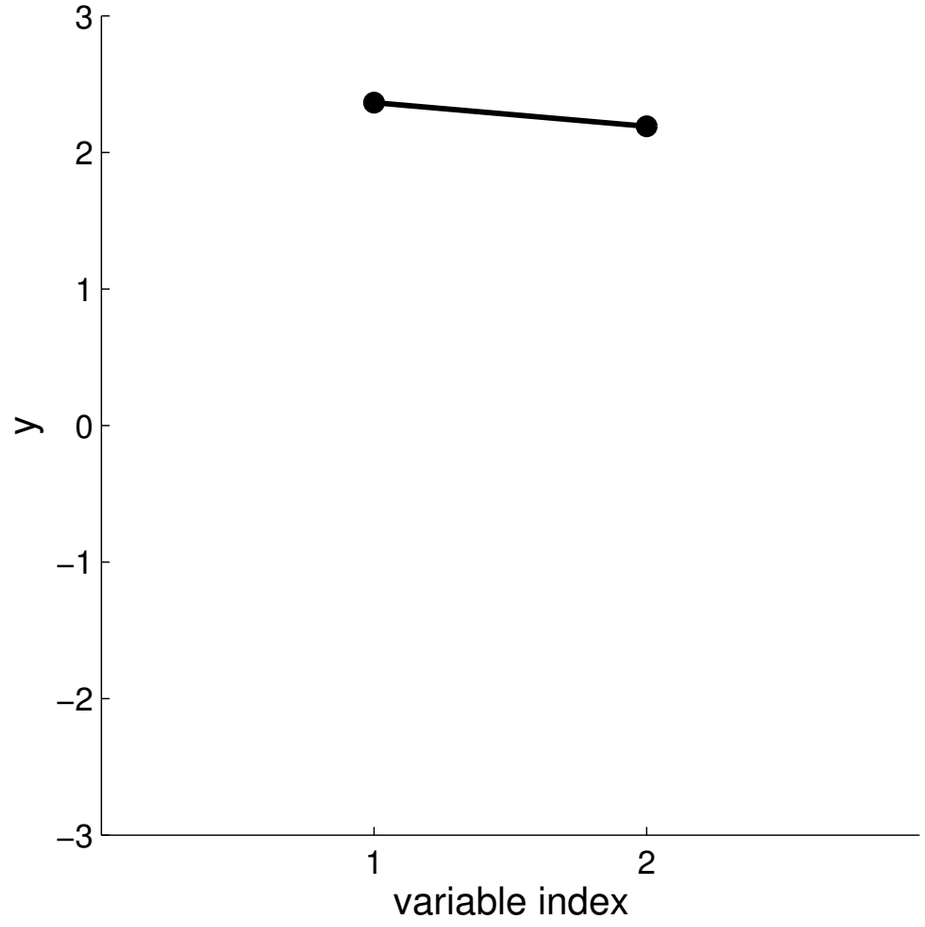
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



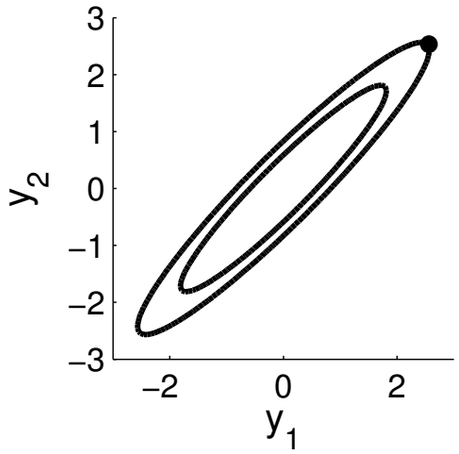
New Visualisation



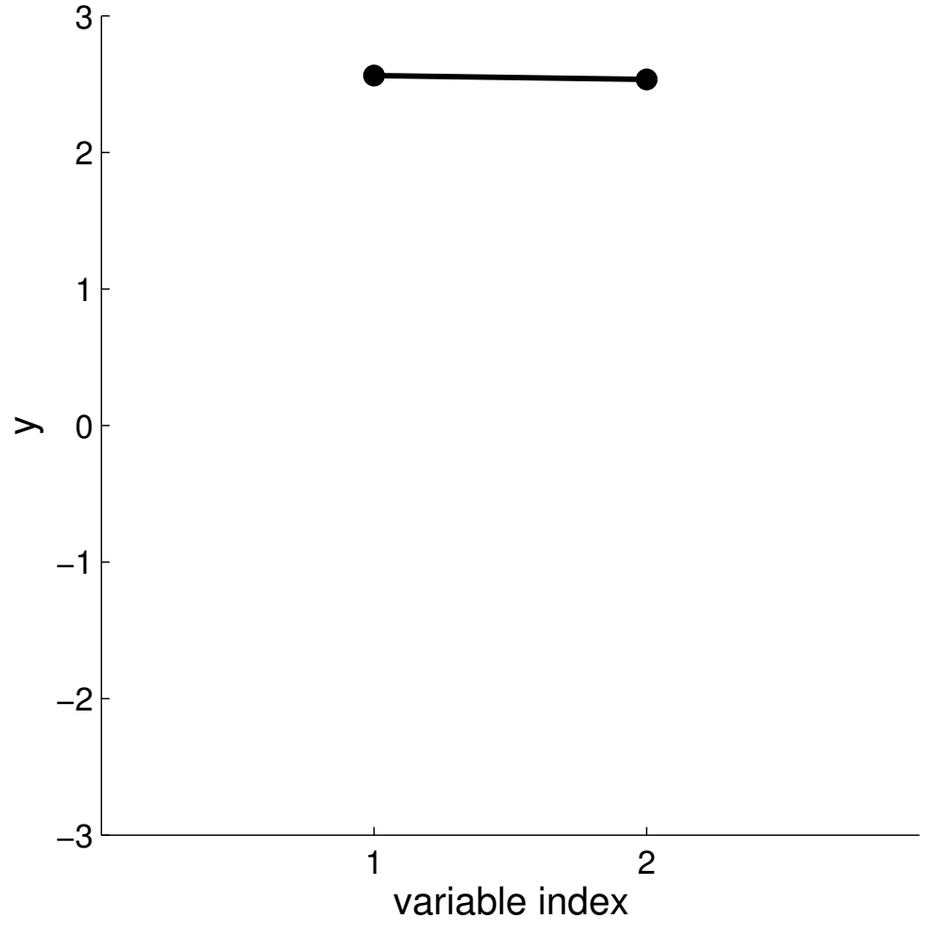
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



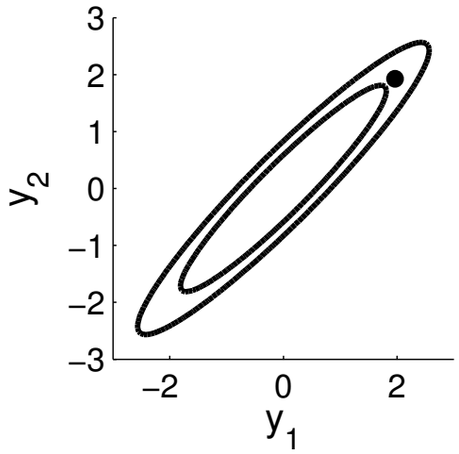
New Visualisation



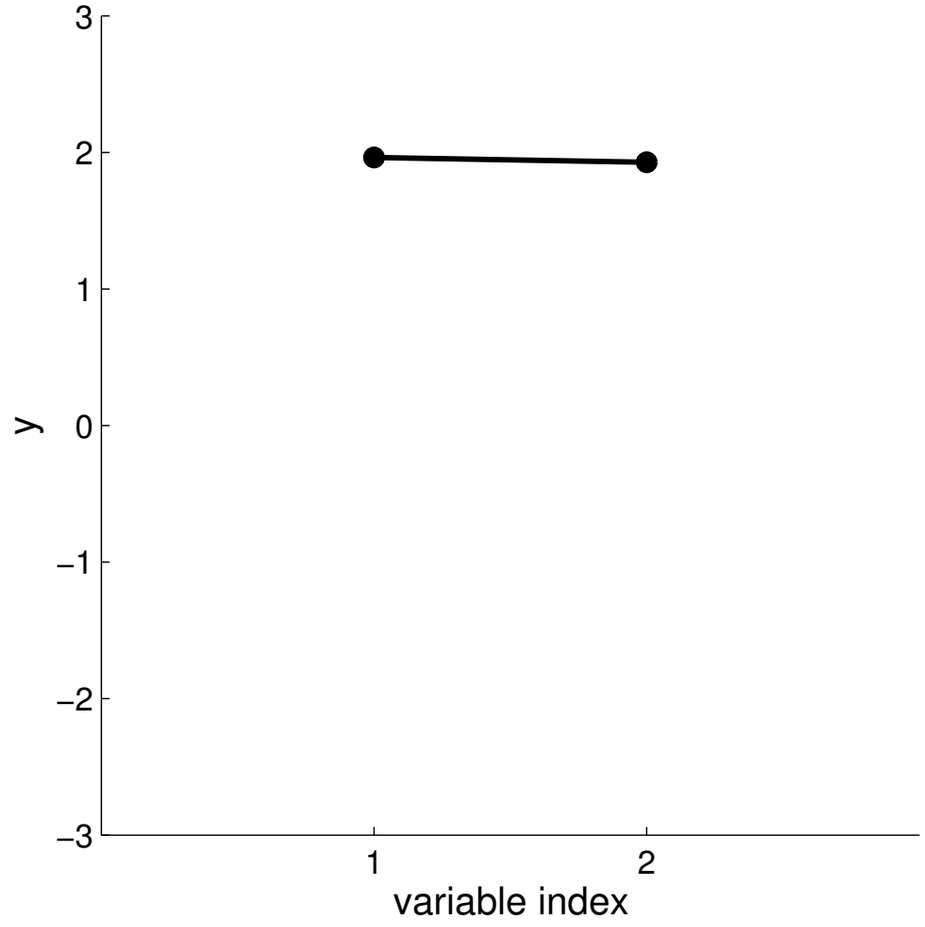
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



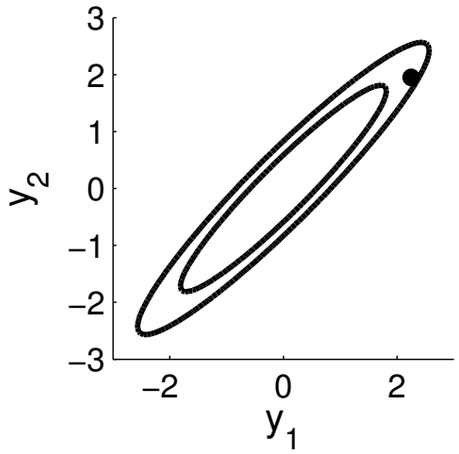
New Visualisation



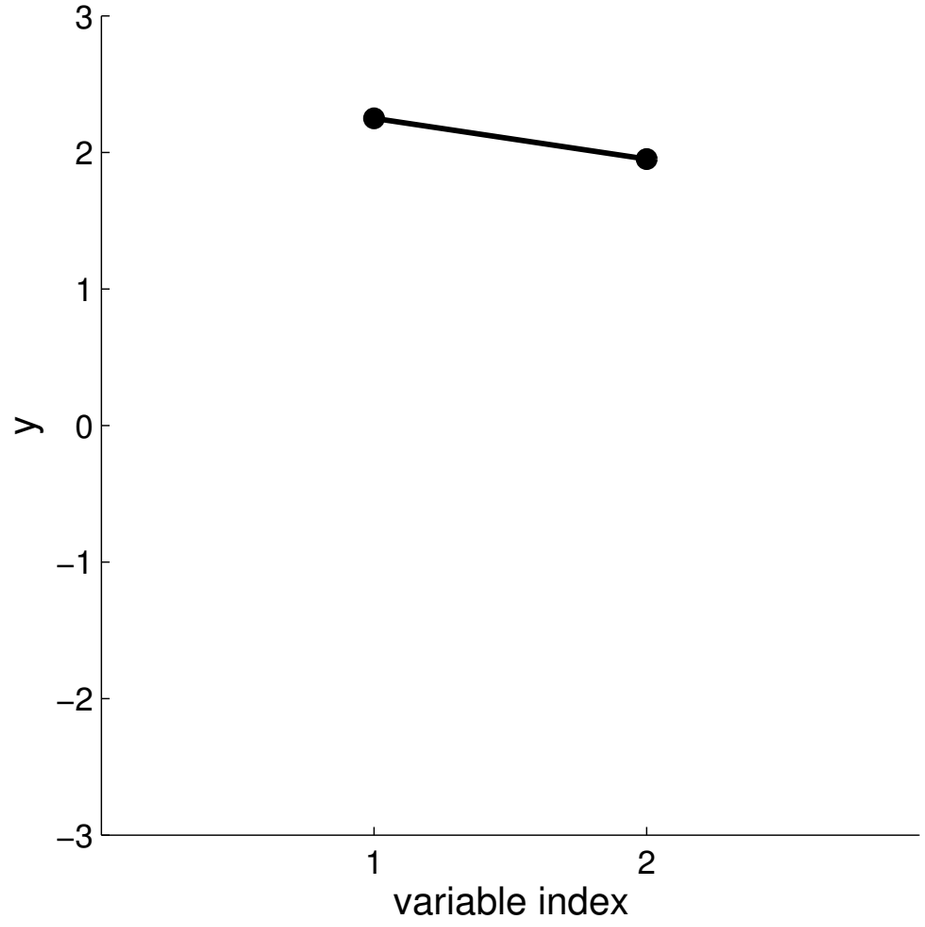
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



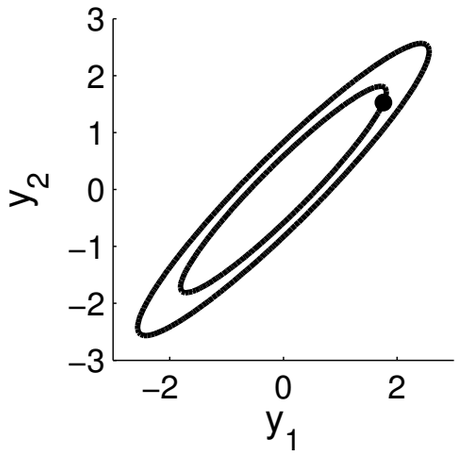
New Visualisation



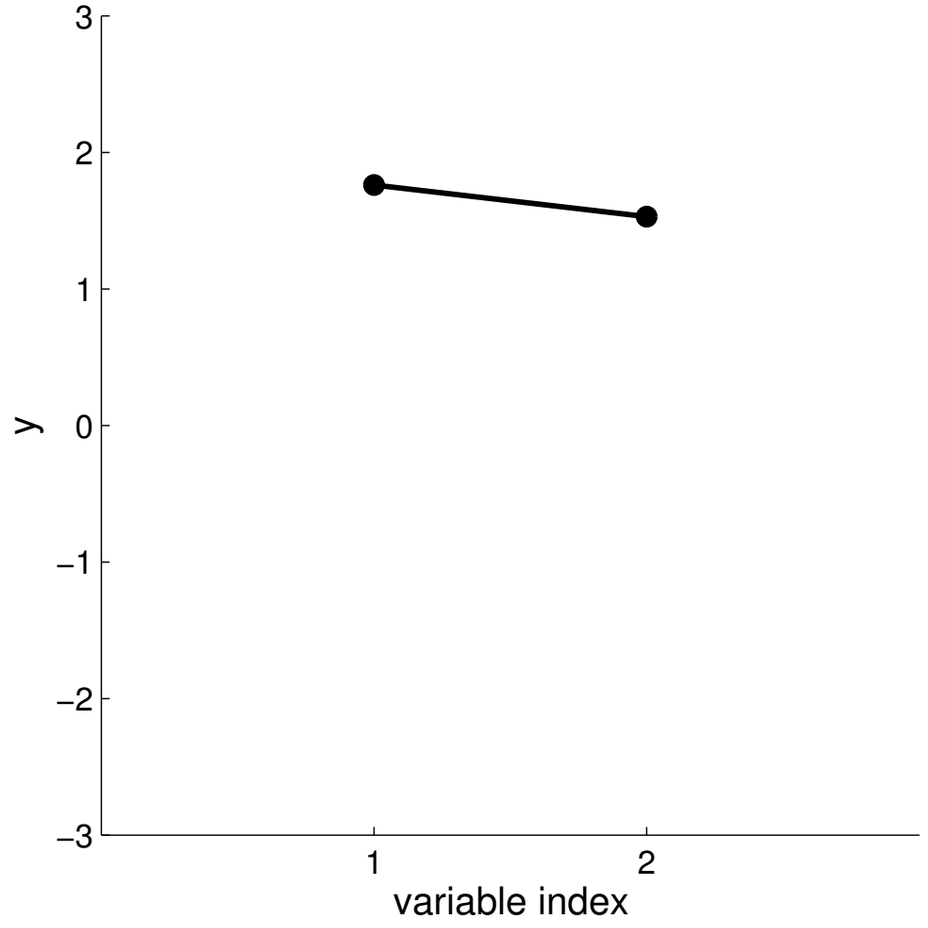
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



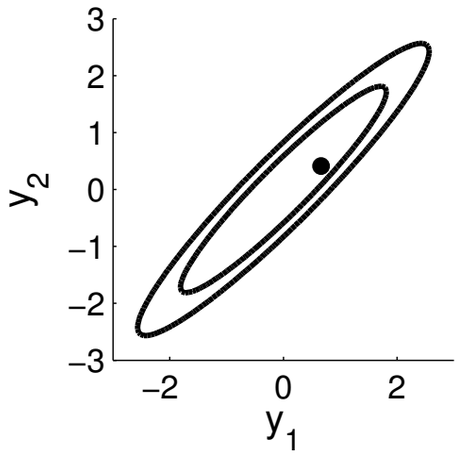
New Visualisation



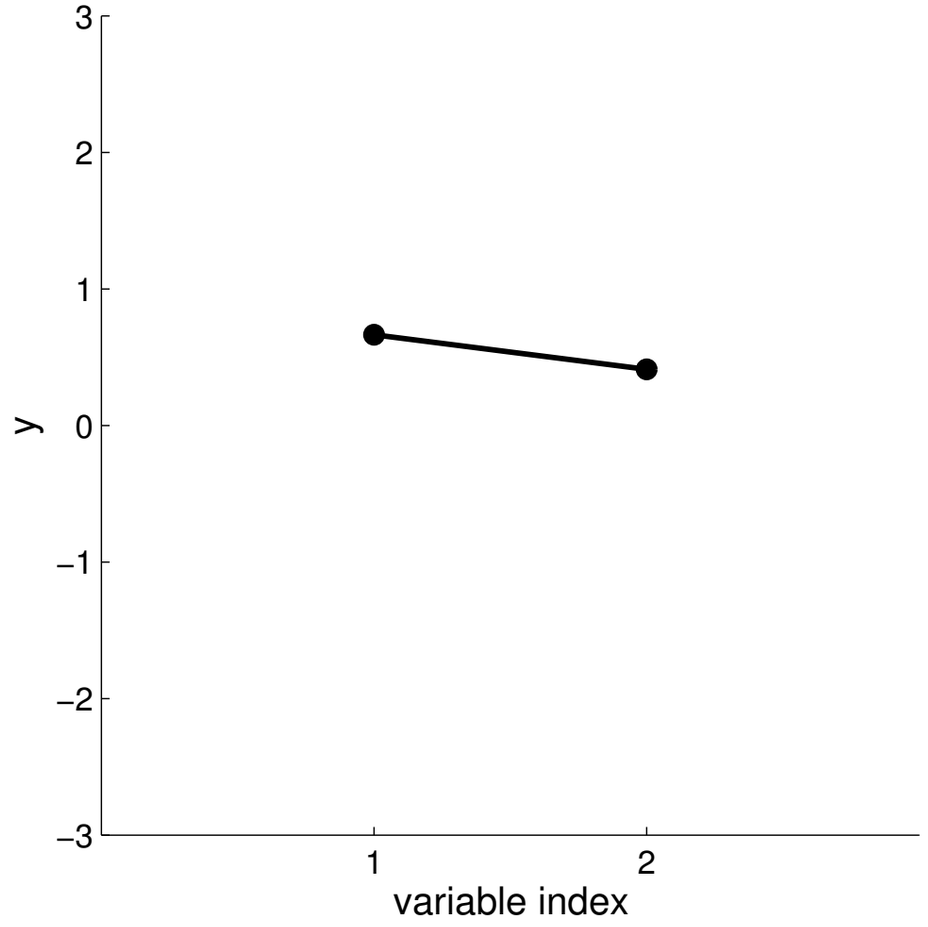
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



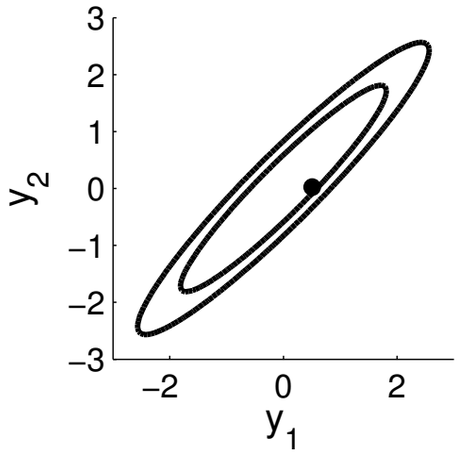
New Visualisation



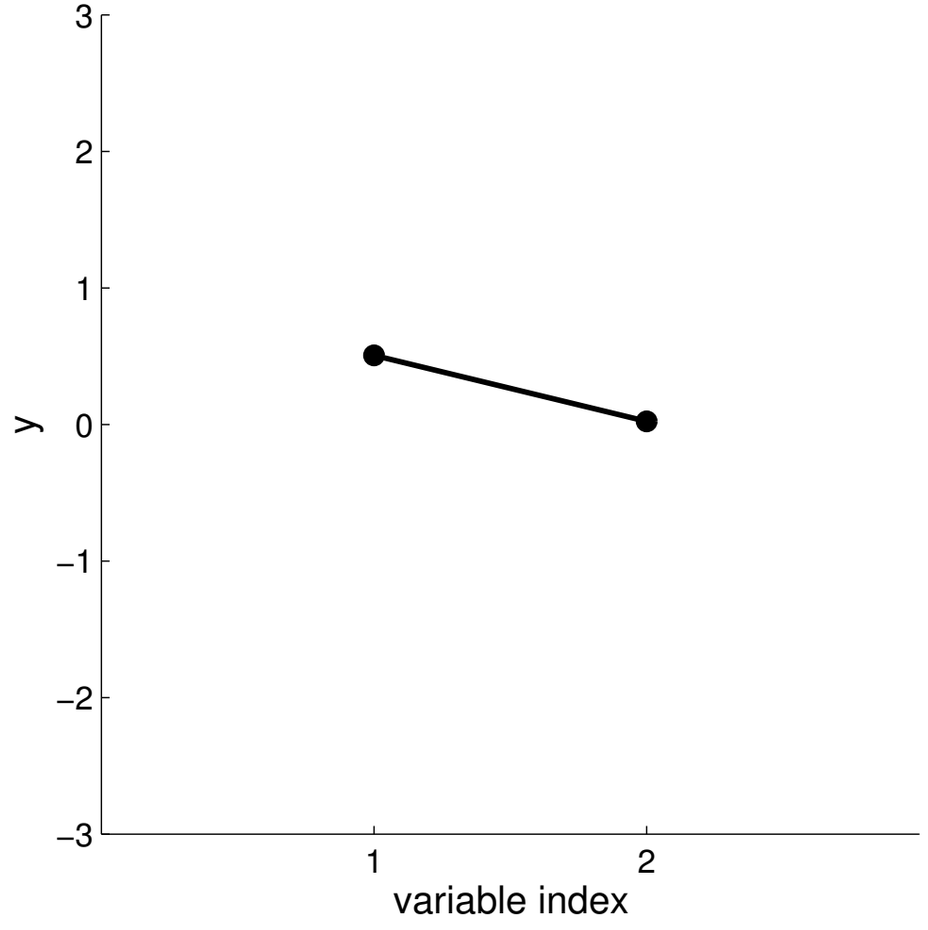
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



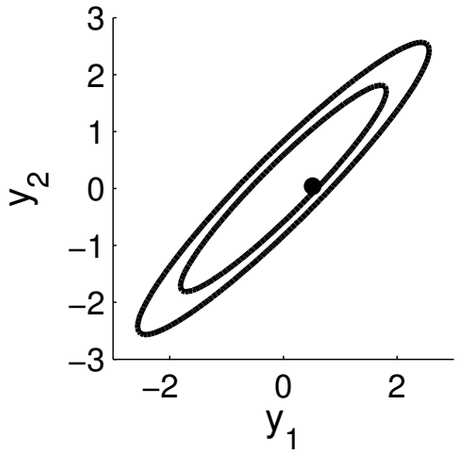
New Visualisation



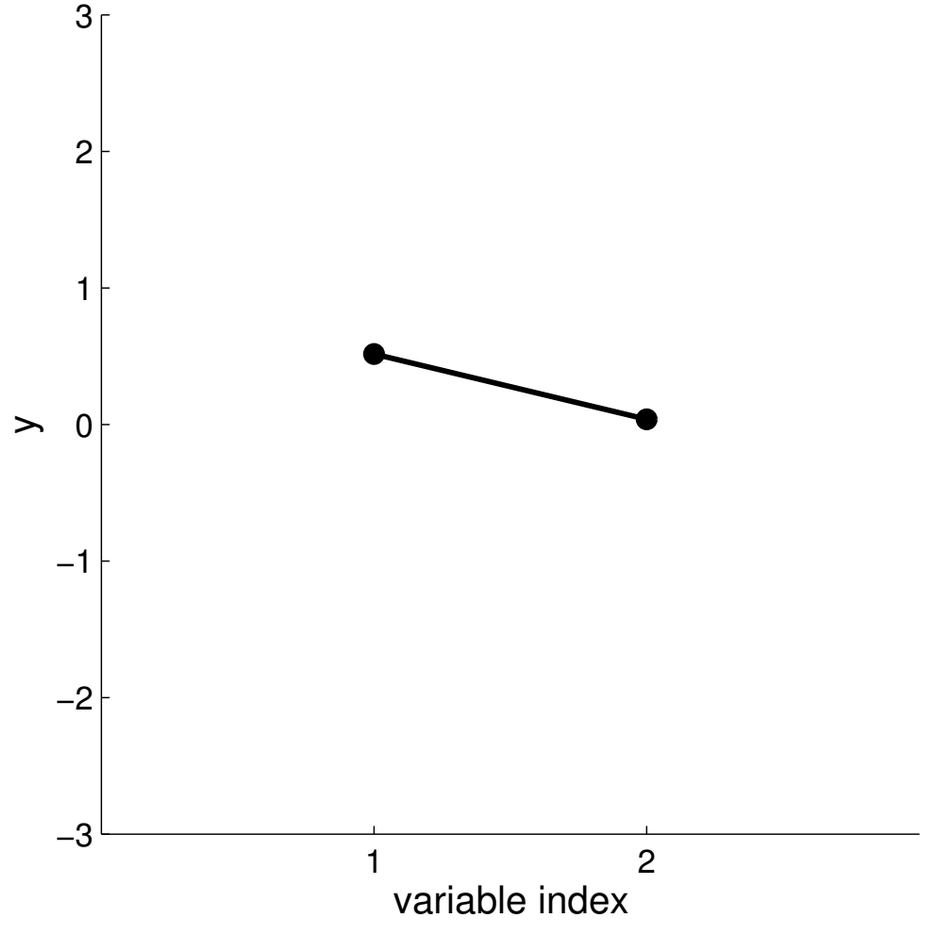
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



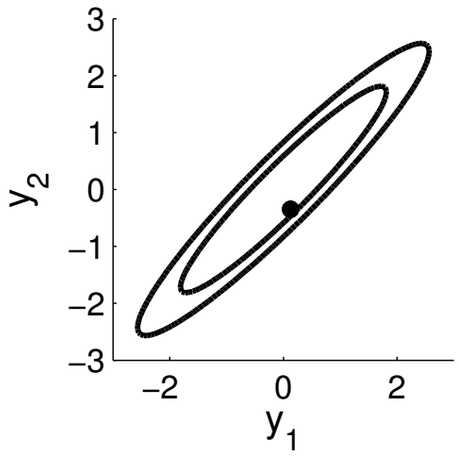
New Visualisation



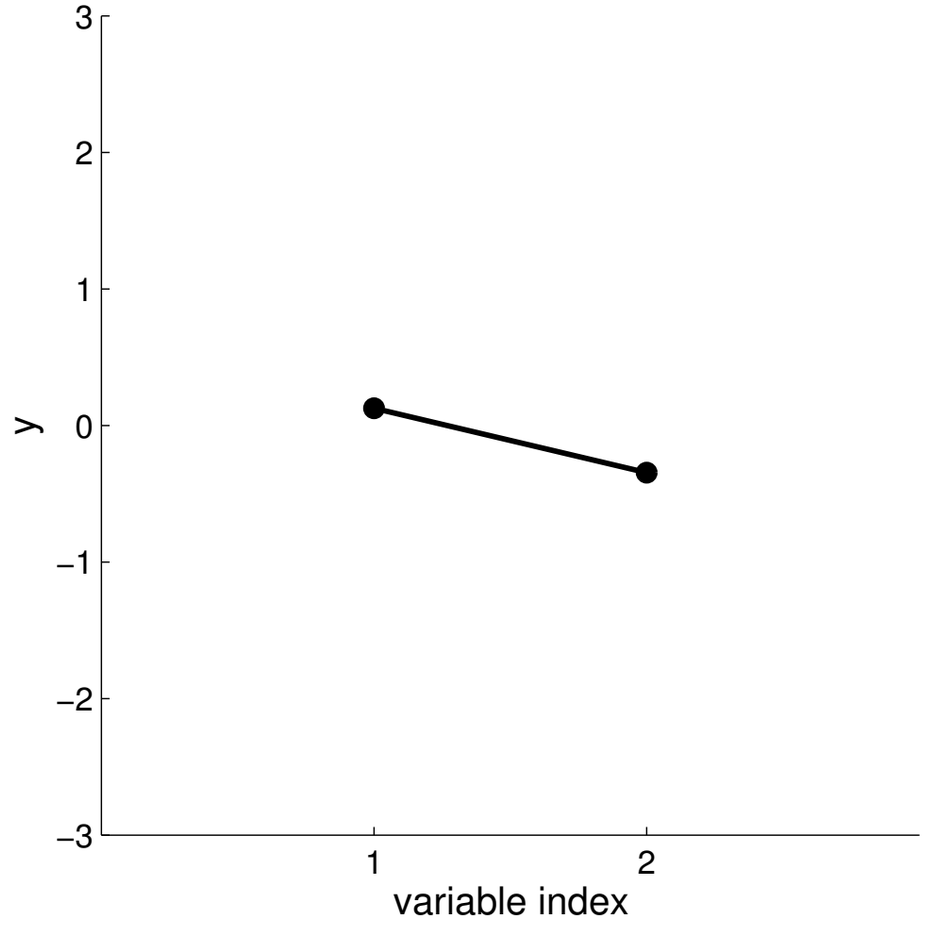
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



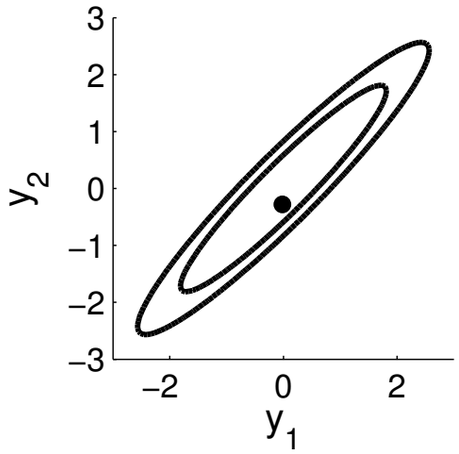
New Visualisation



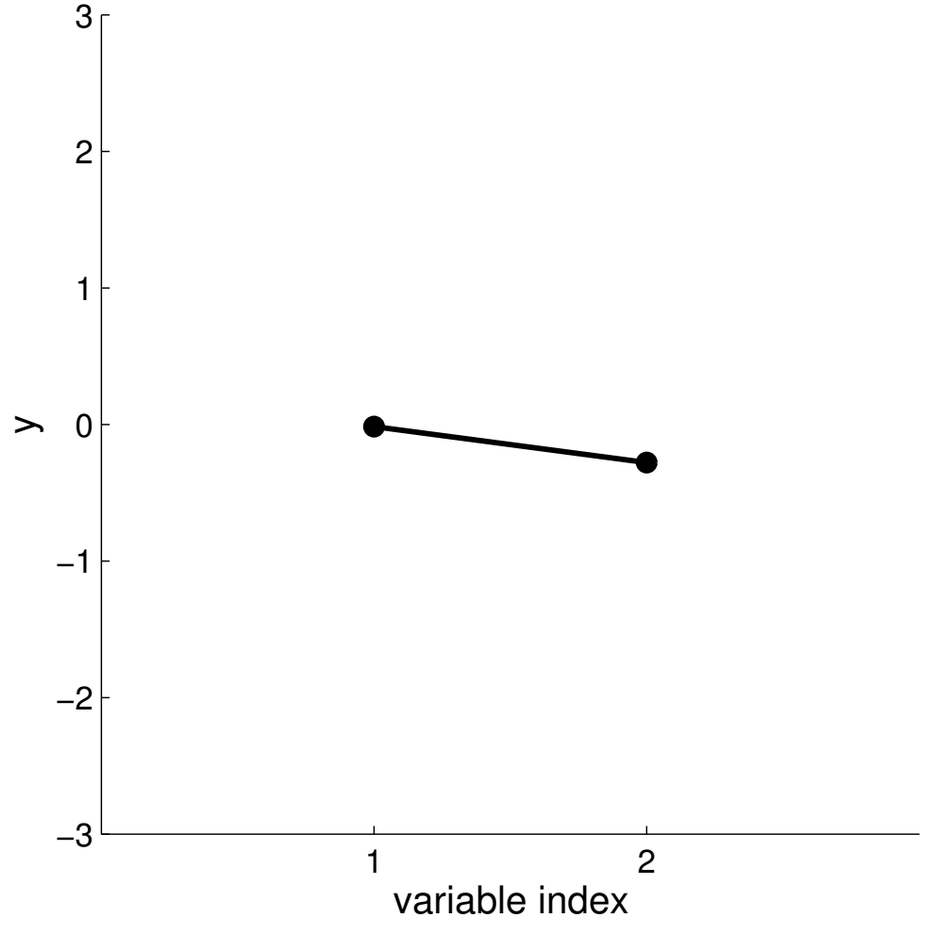
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



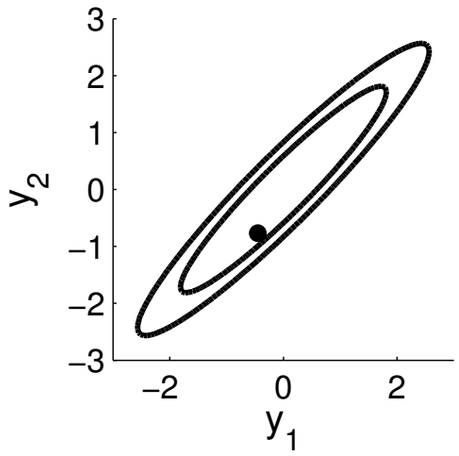
New Visualisation



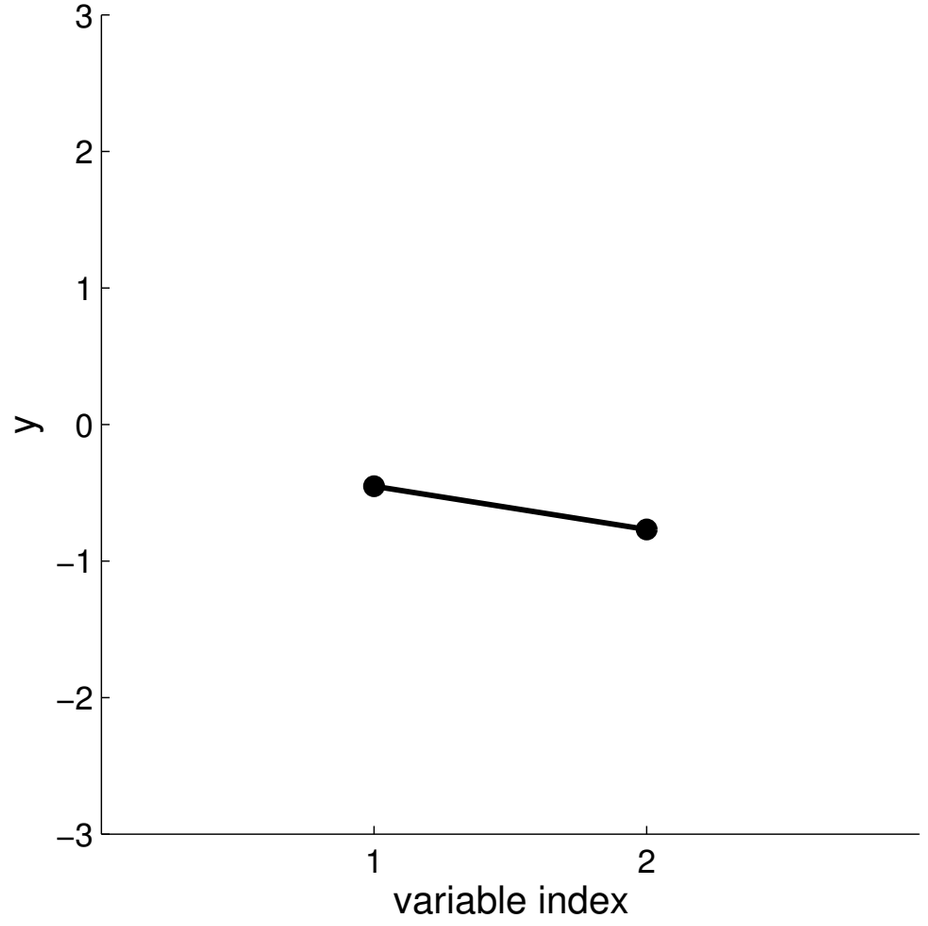
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



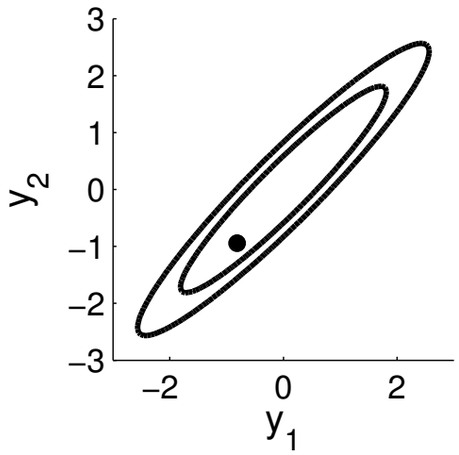
New Visualisation



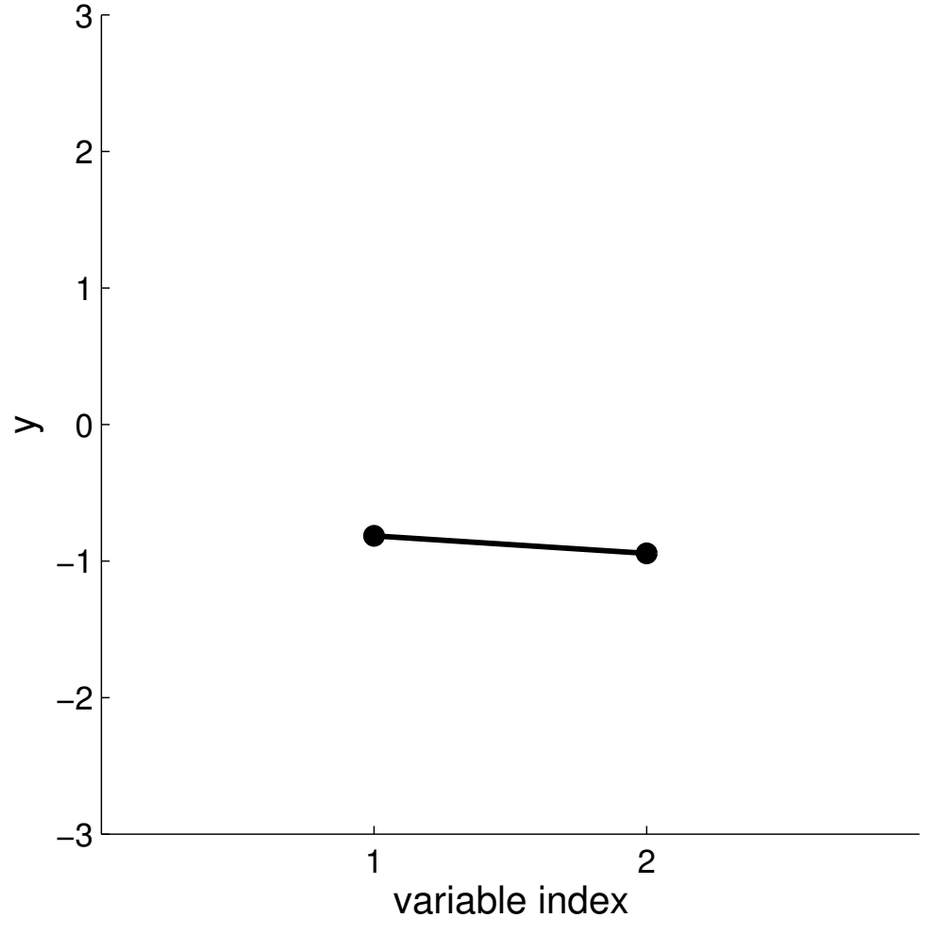
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



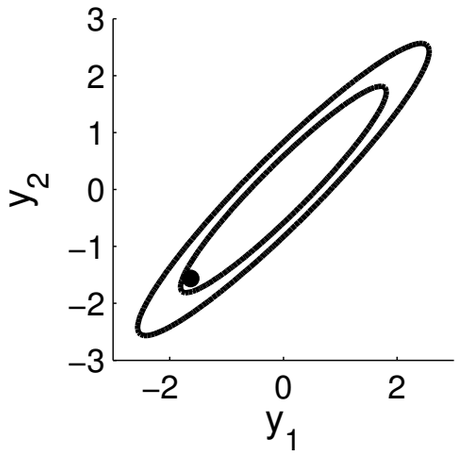
New Visualisation



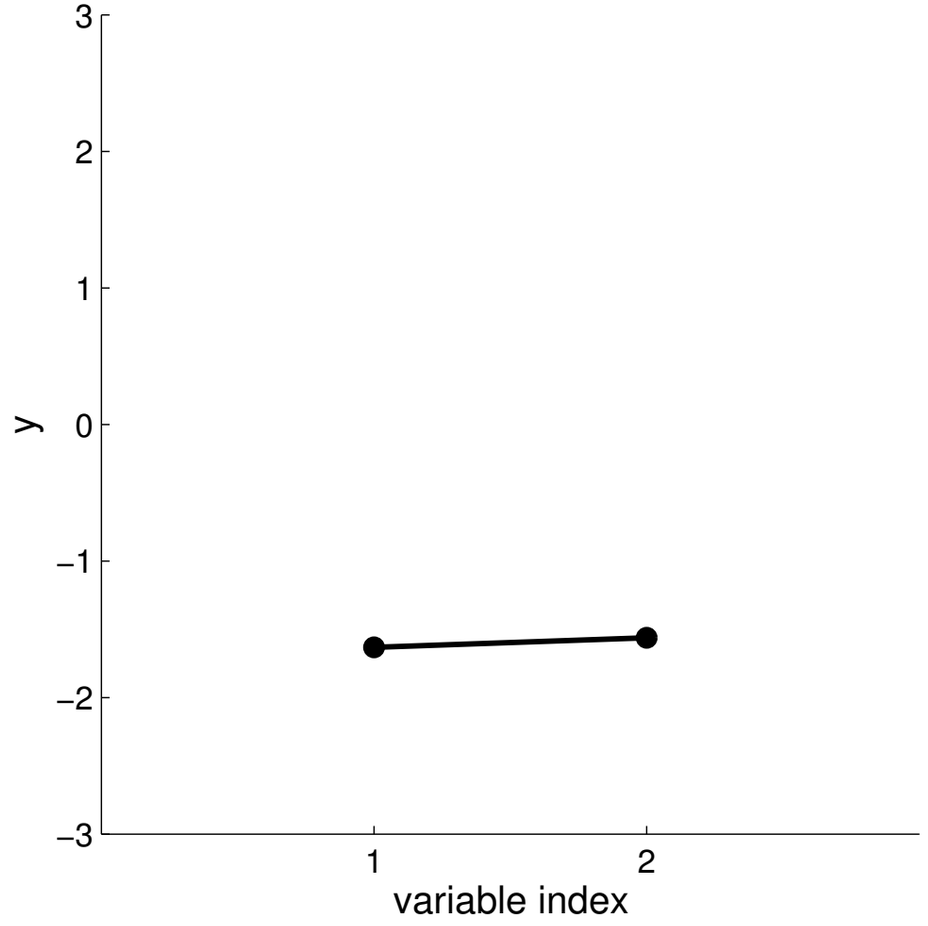
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



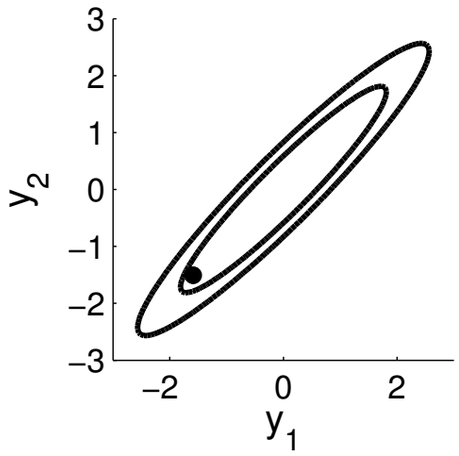
New Visualisation



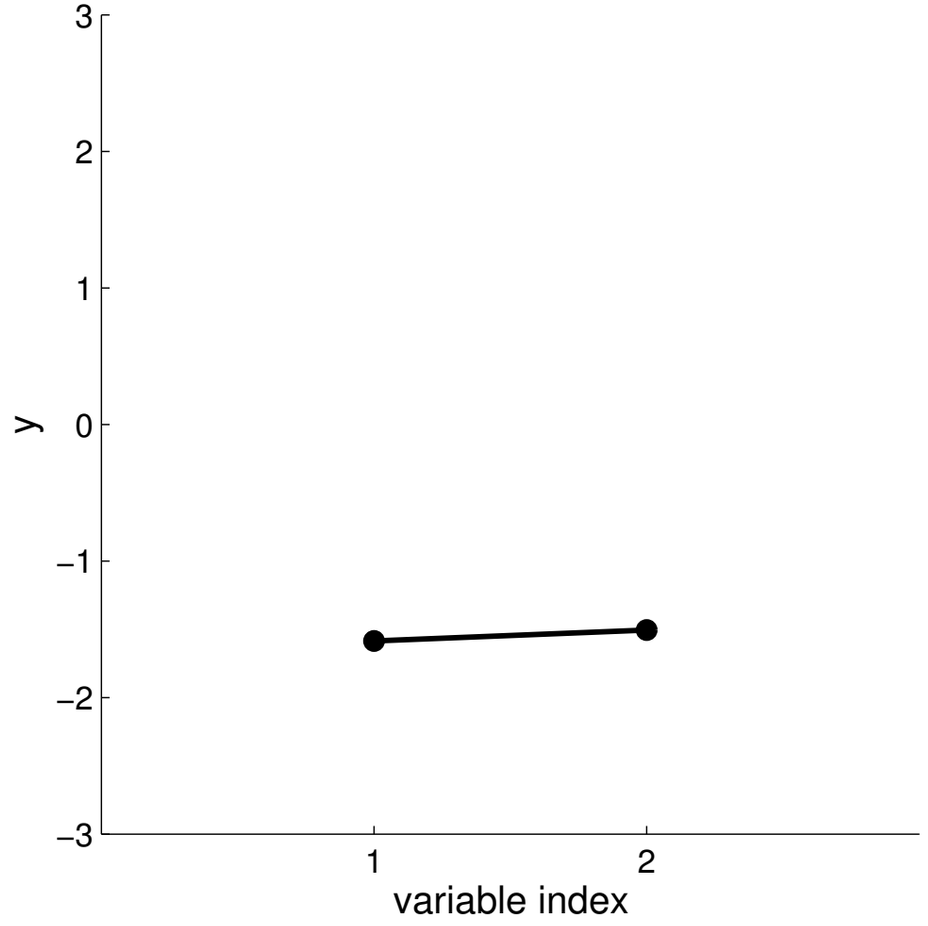
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



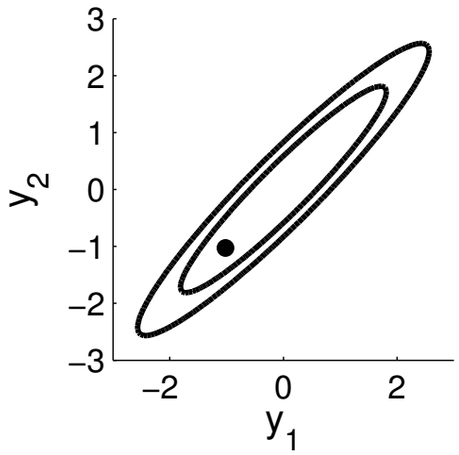
New Visualisation



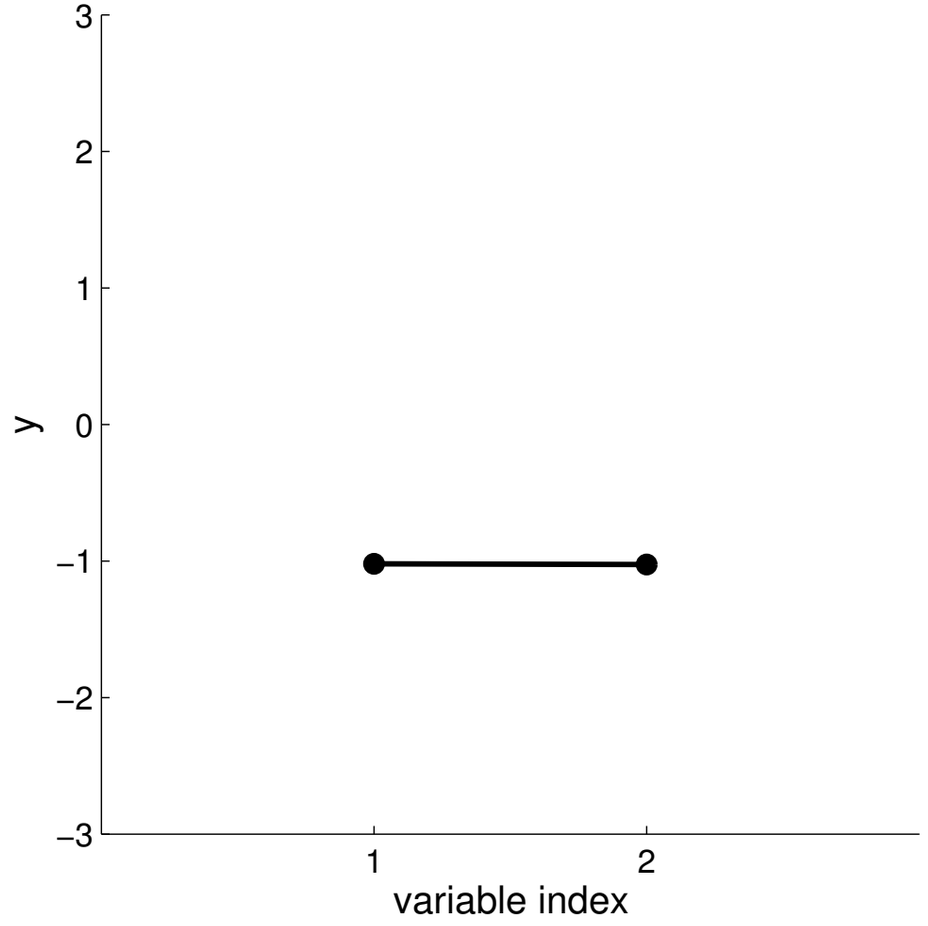
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



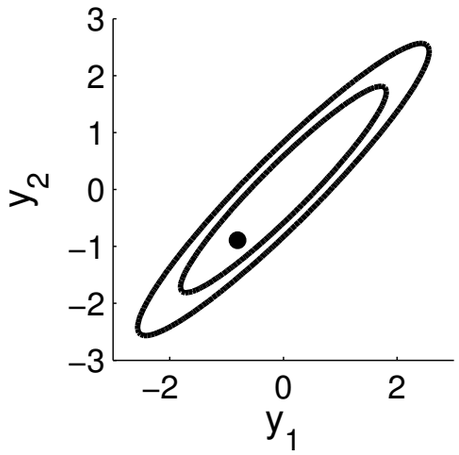
New Visualisation



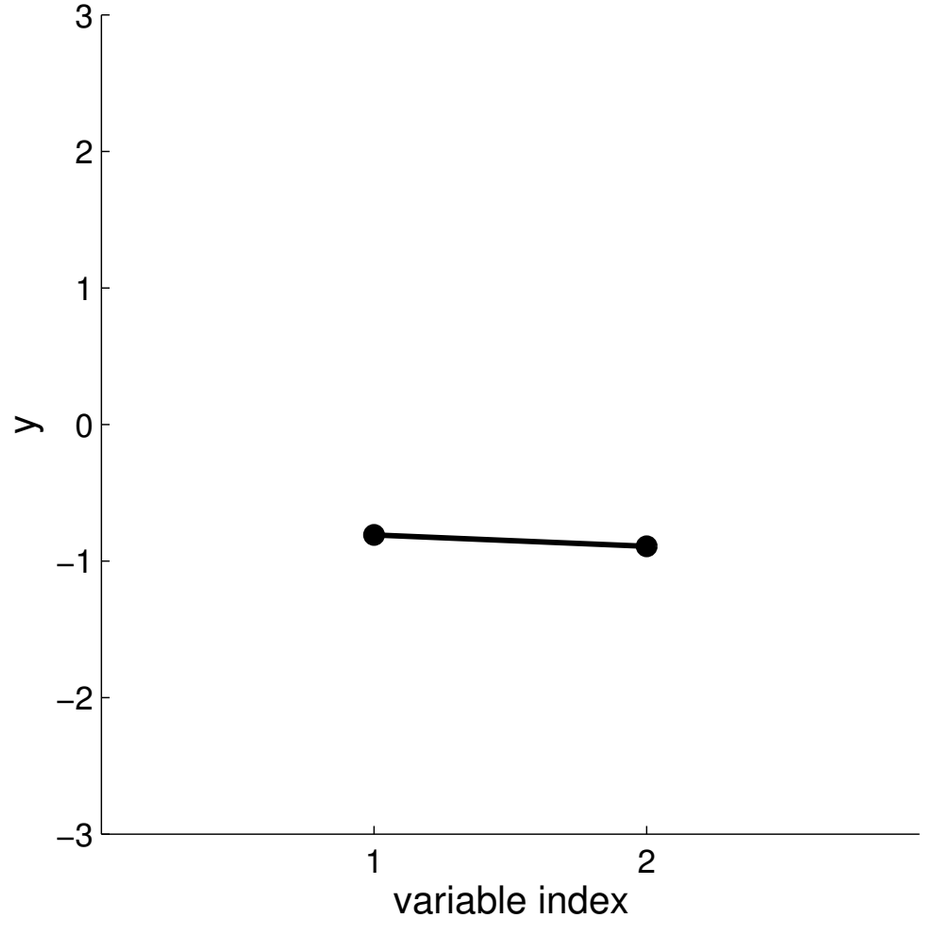
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



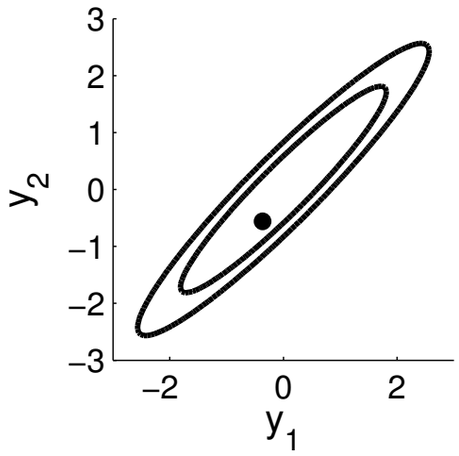
New Visualisation



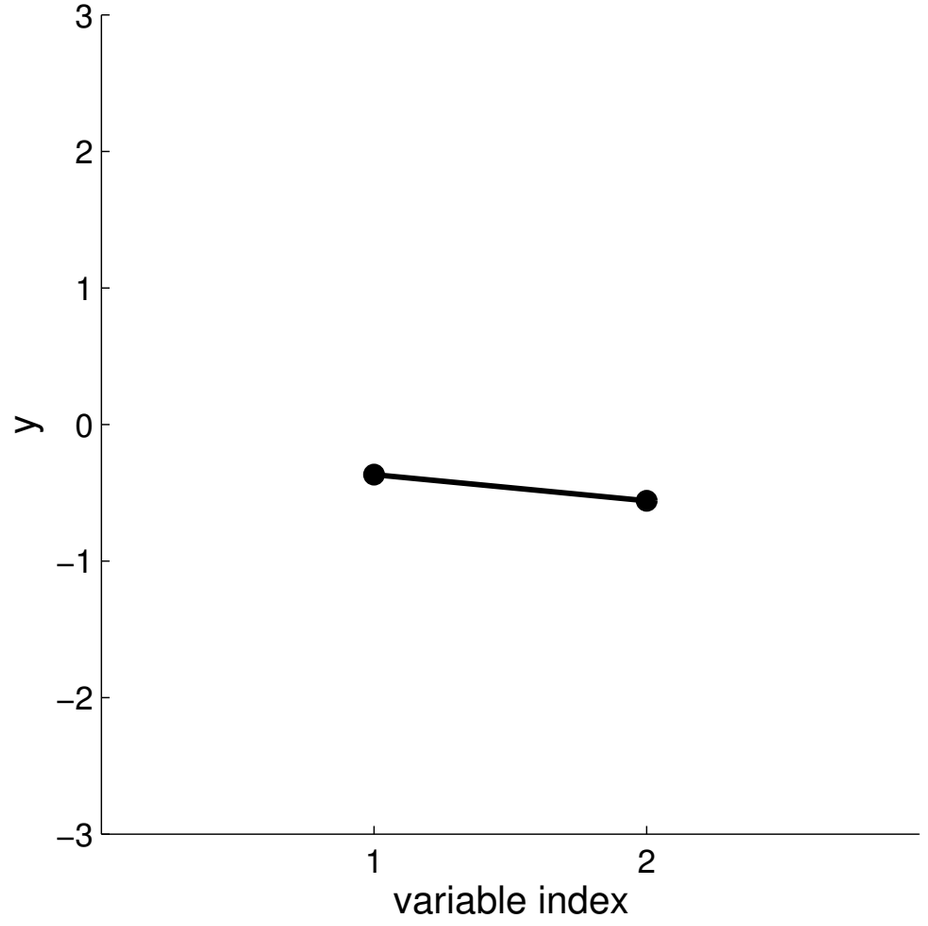
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



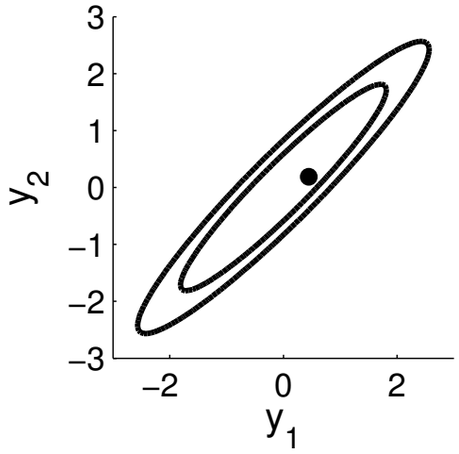
New Visualisation



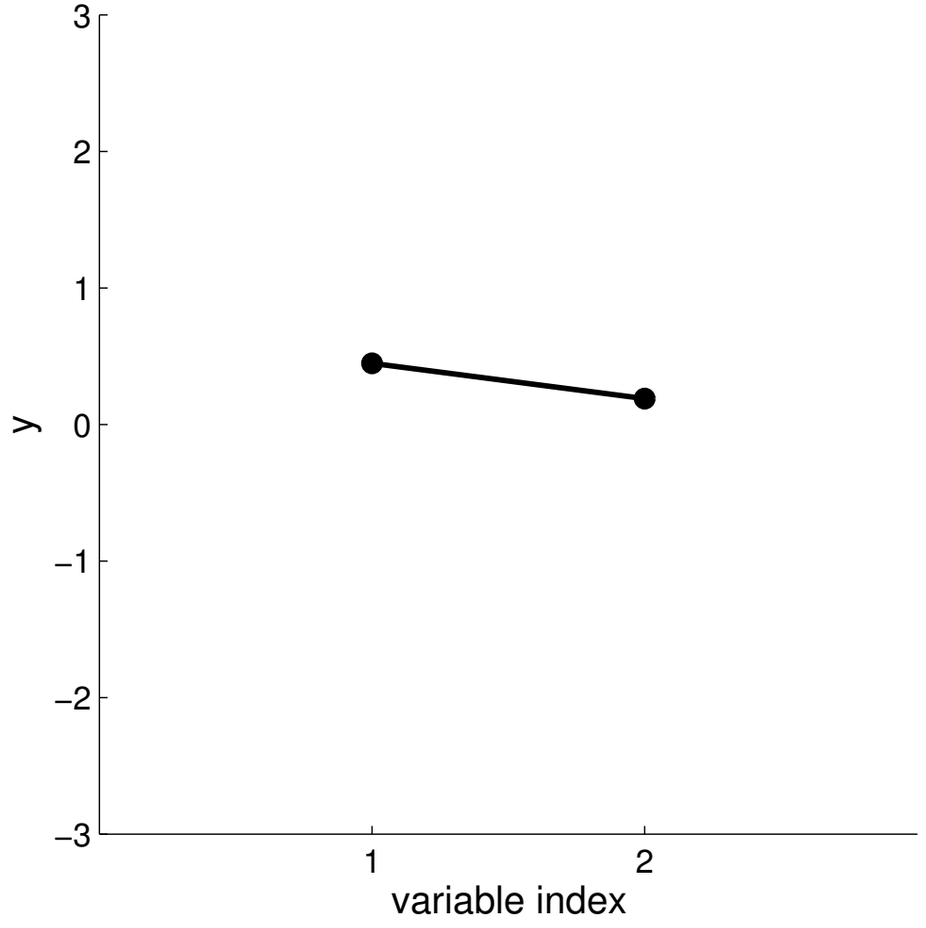
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



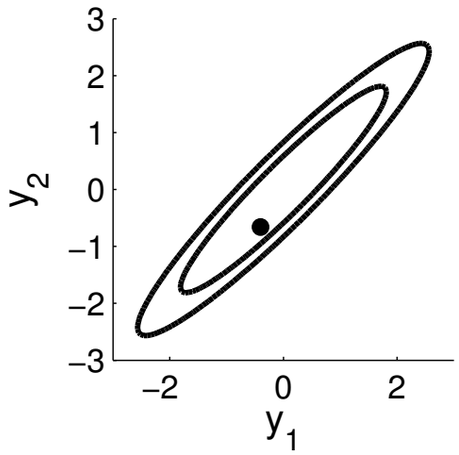
New Visualisation



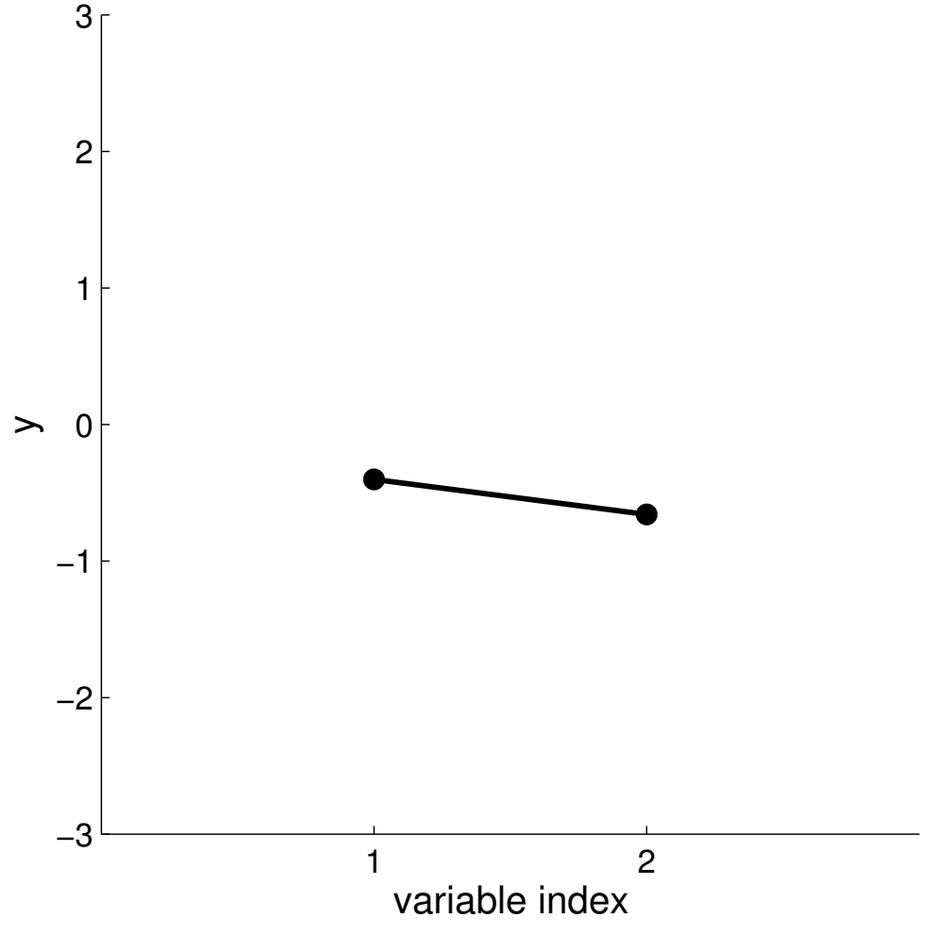
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



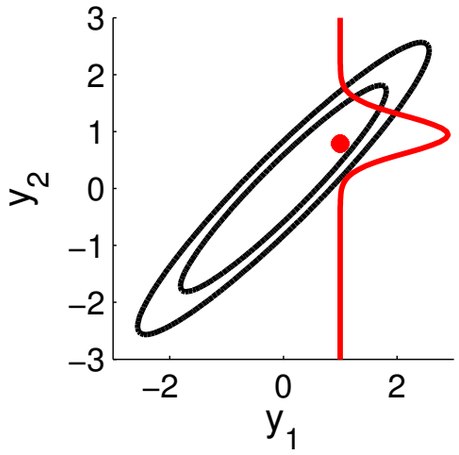
New Visualisation



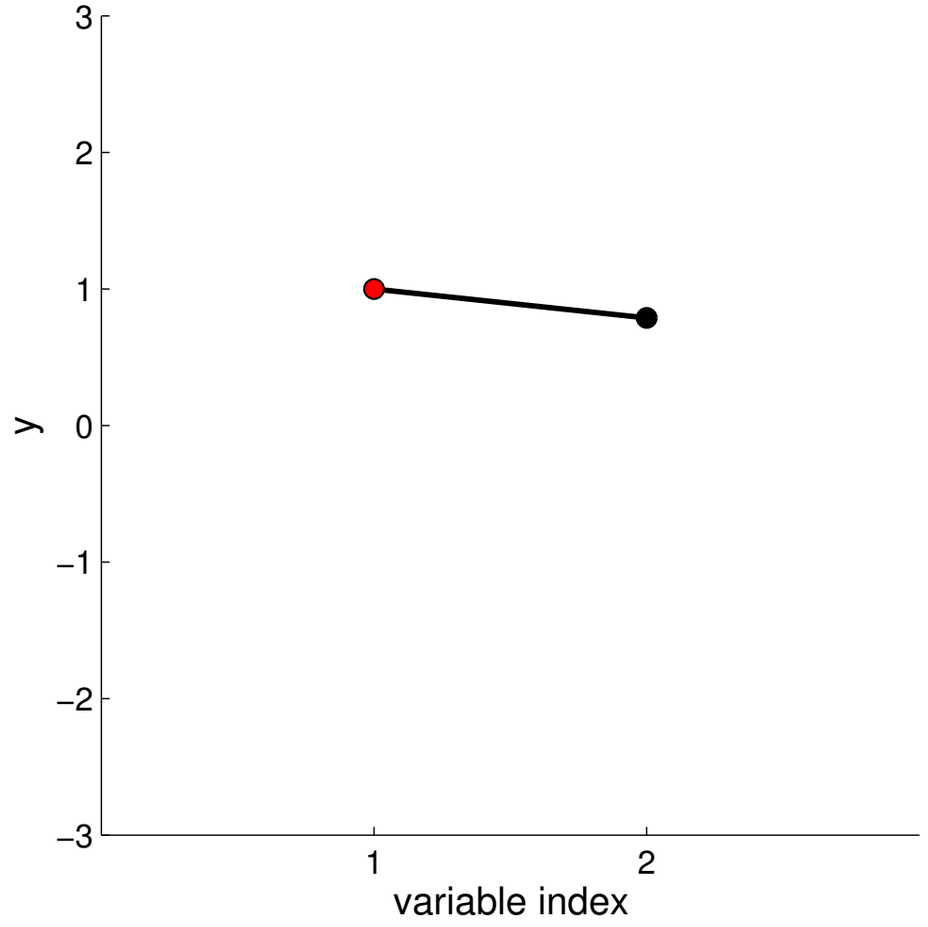
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



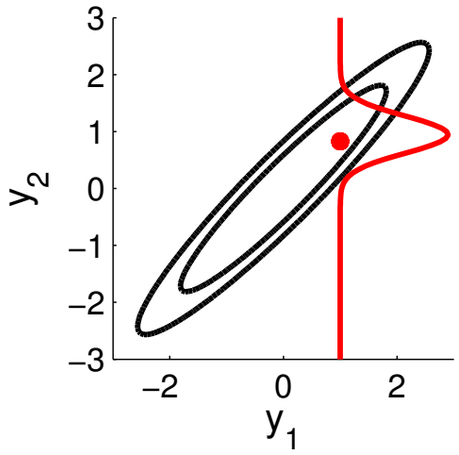
New Visualisation



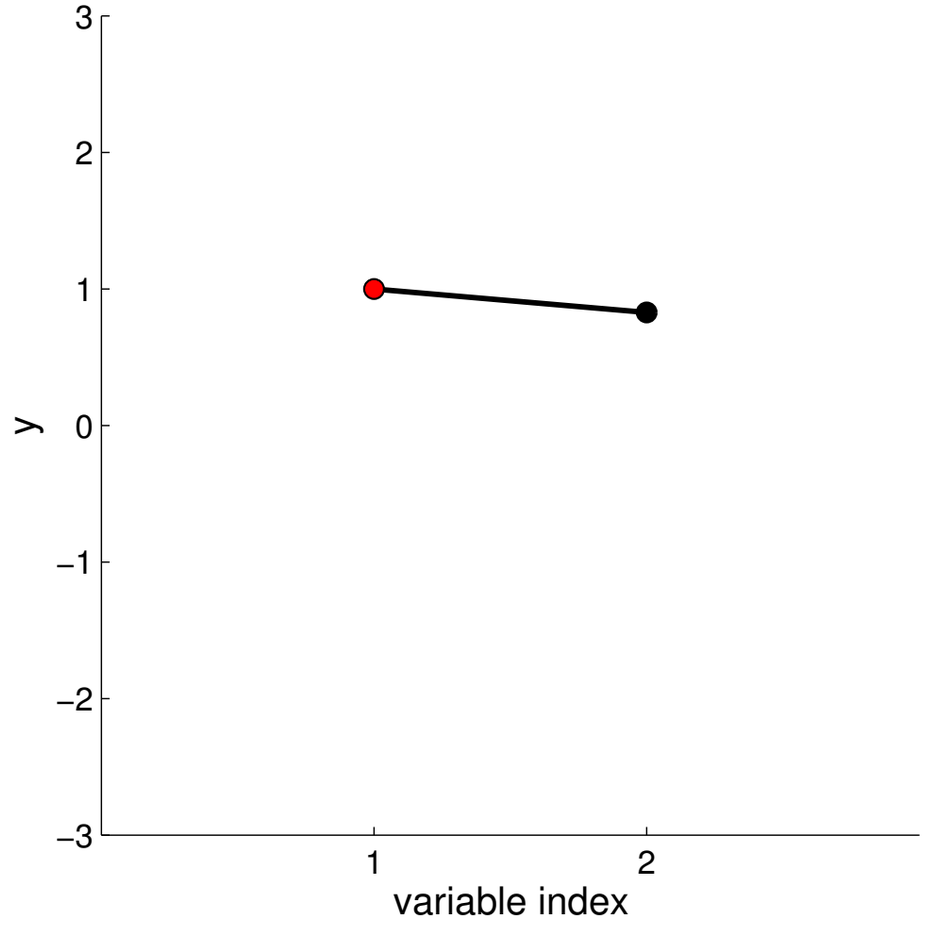
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



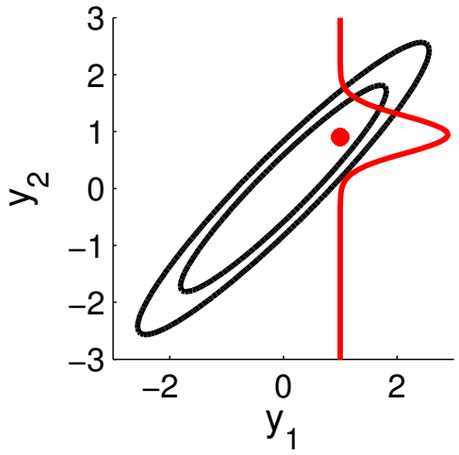
New Visualisation



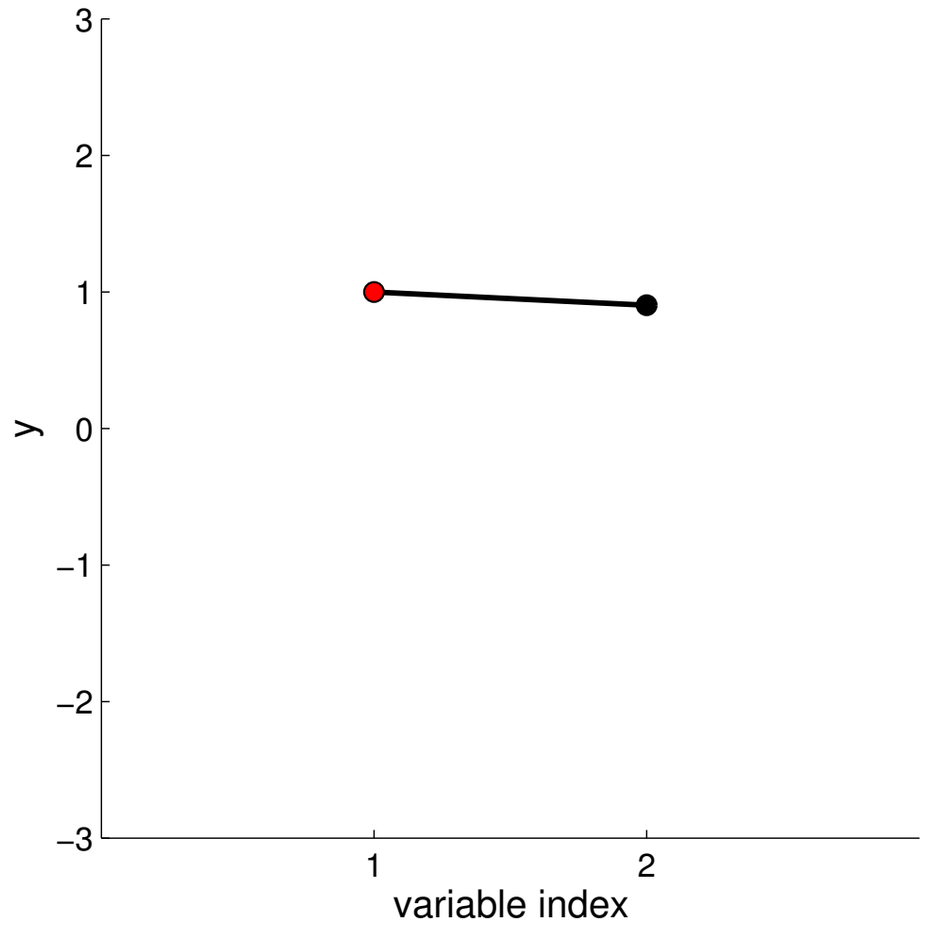
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



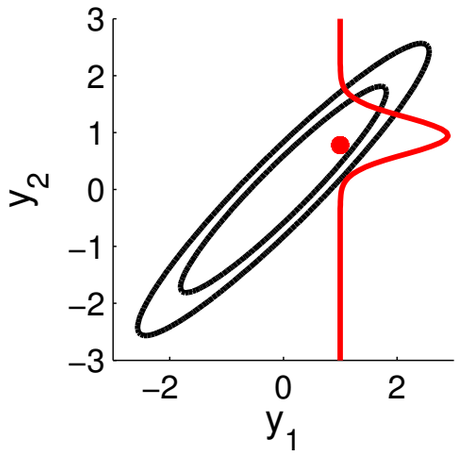
New Visualisation



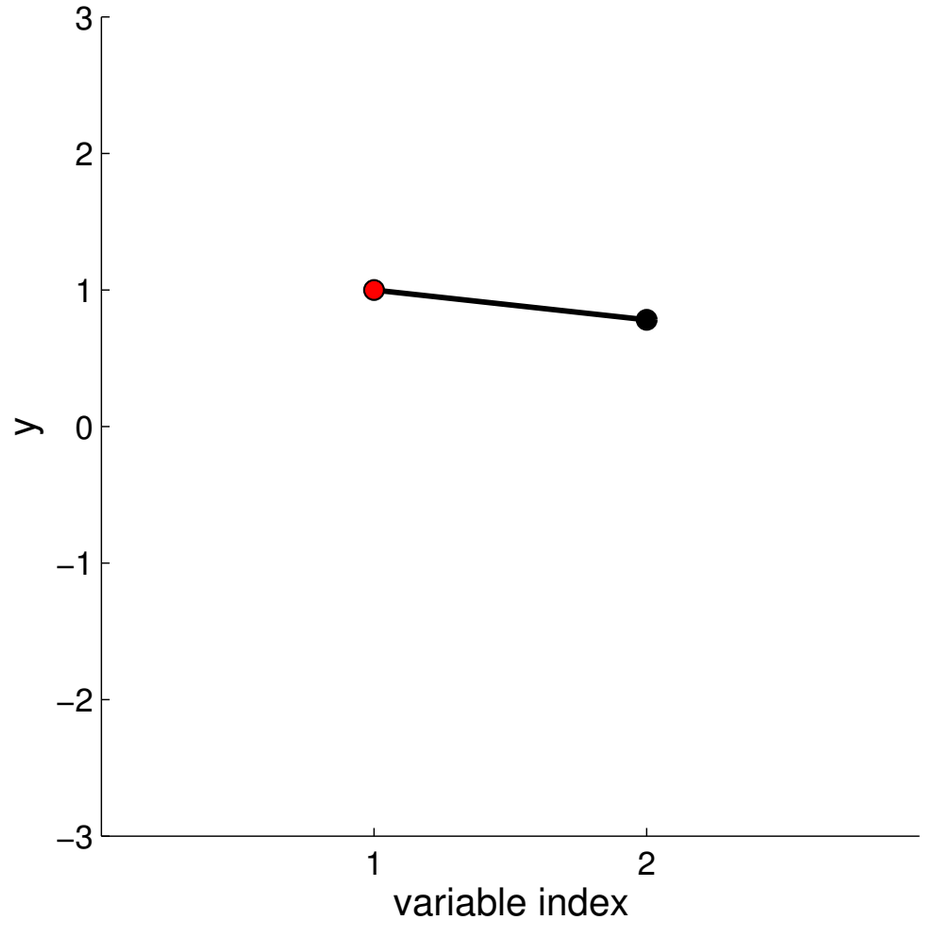
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



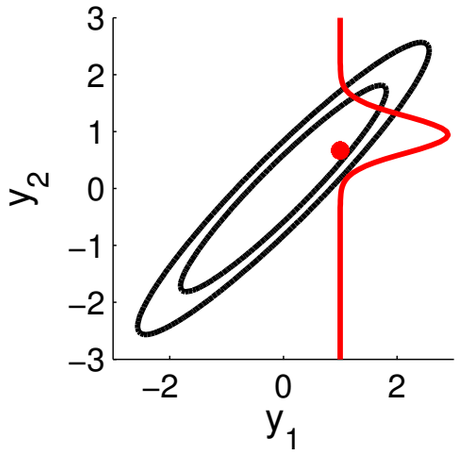
New Visualisation



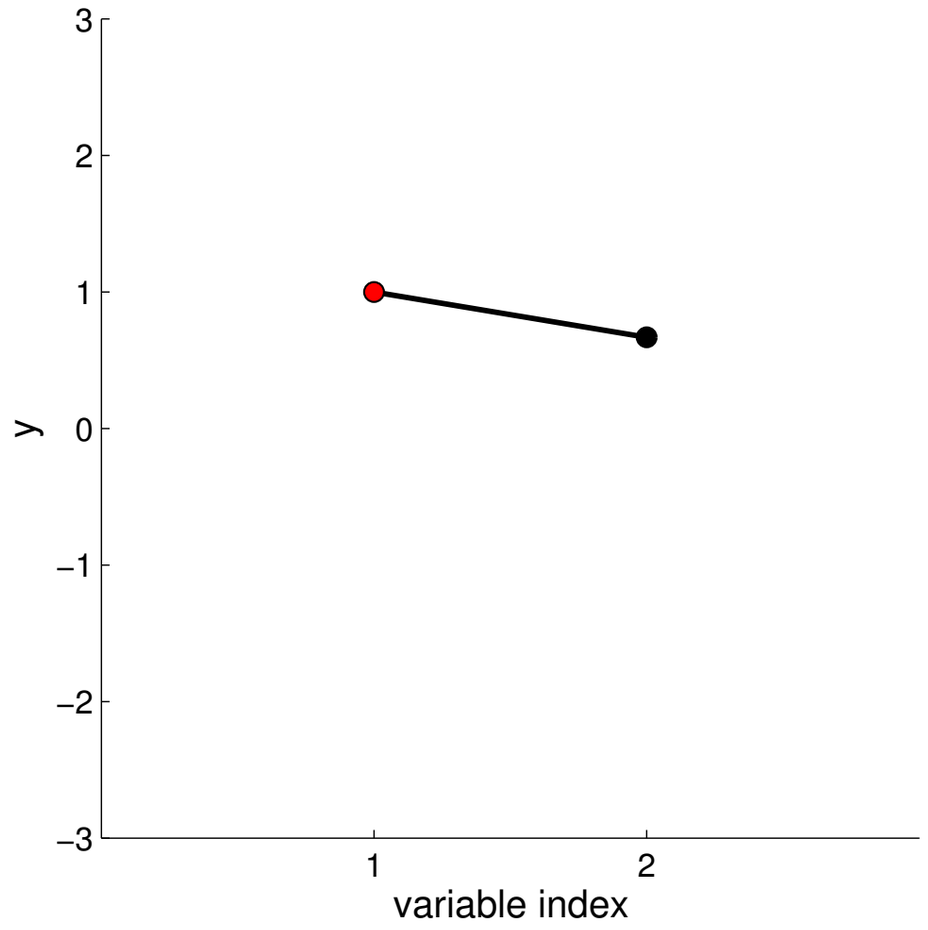
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



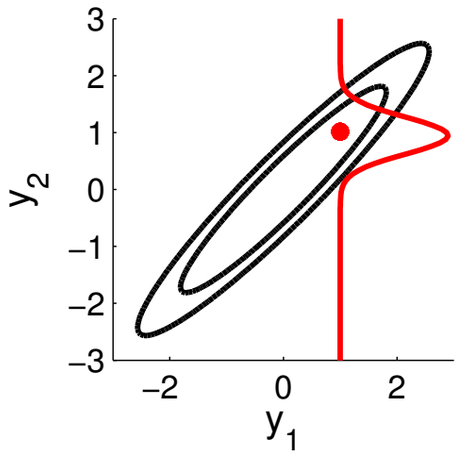
New Visualisation



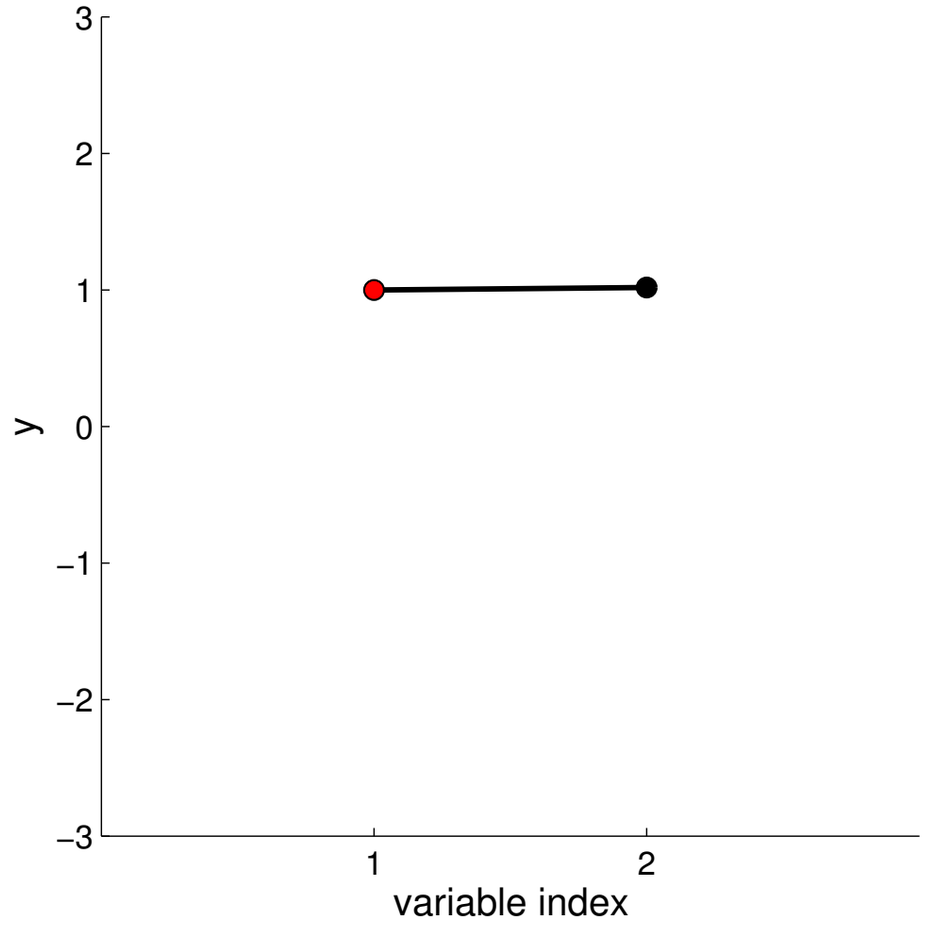
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



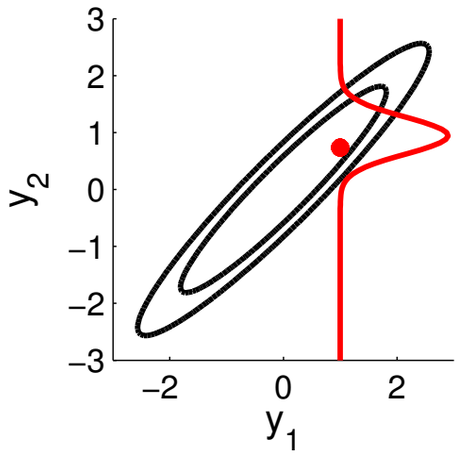
New Visualisation



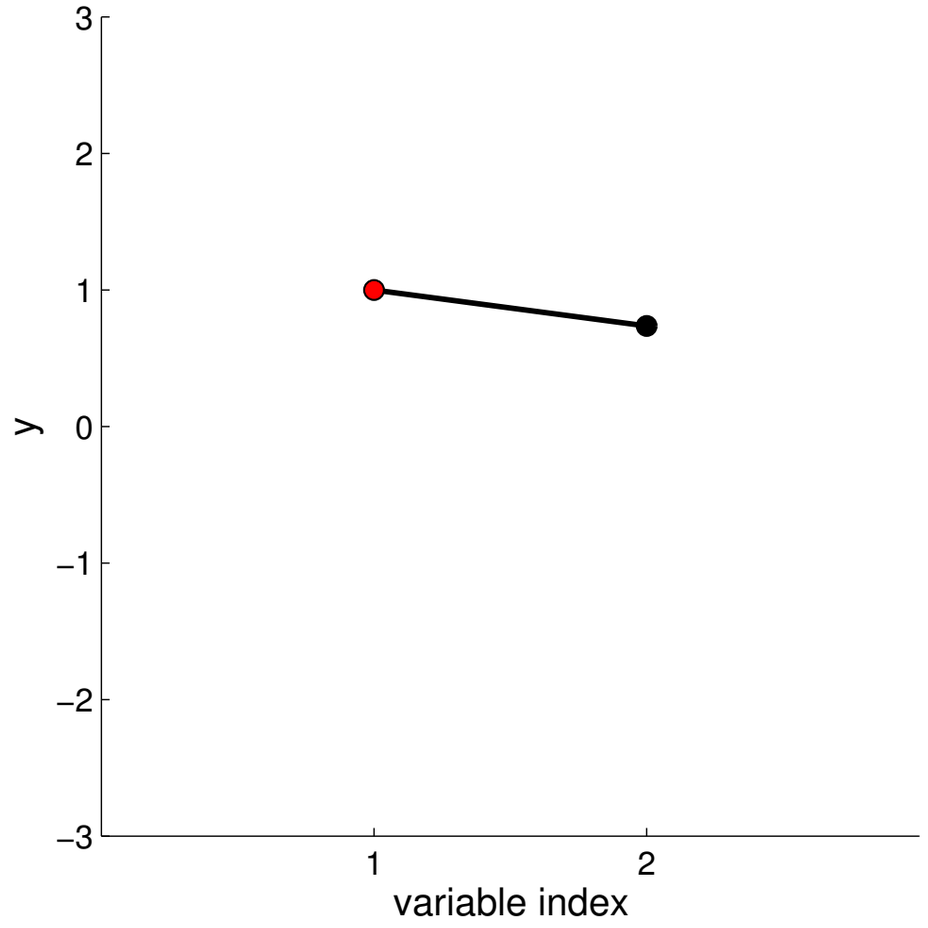
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



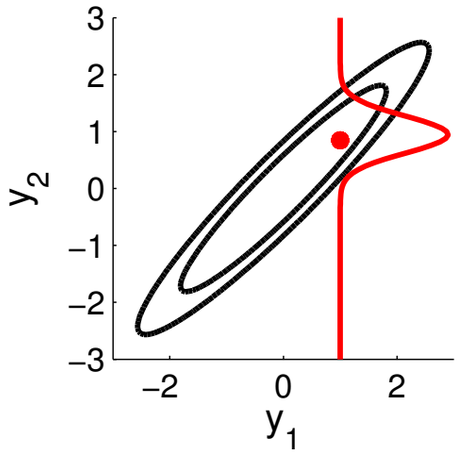
New Visualisation



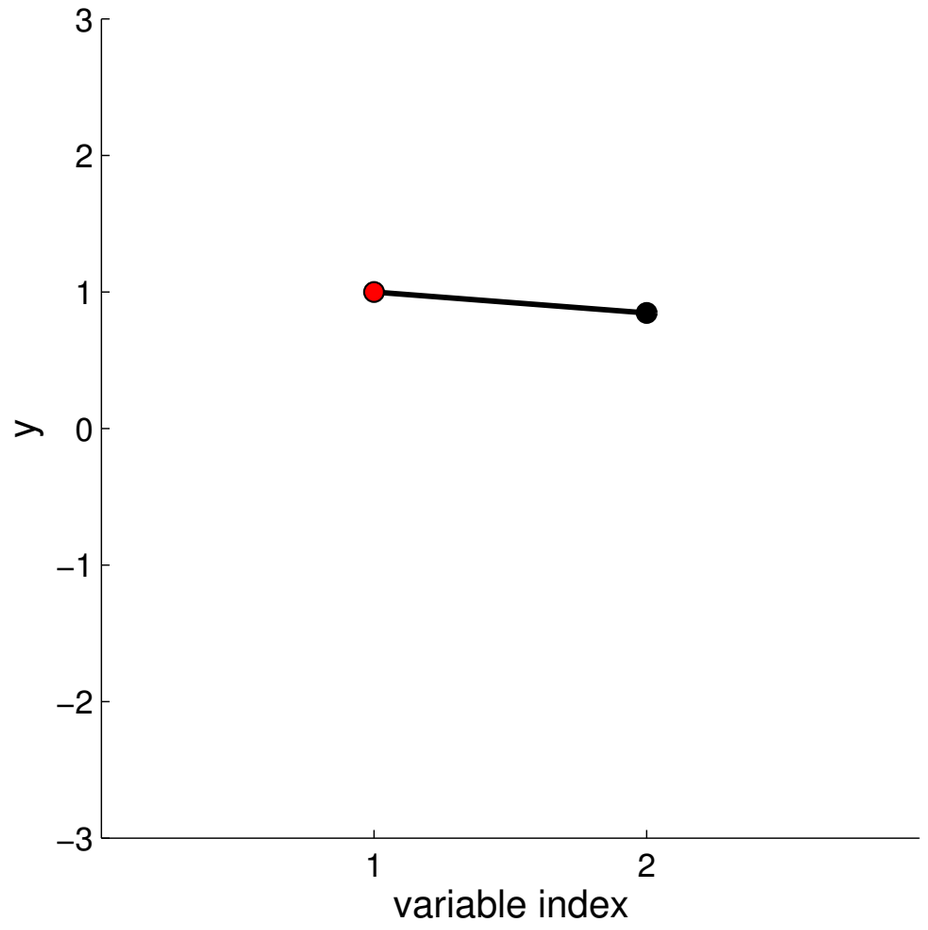
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



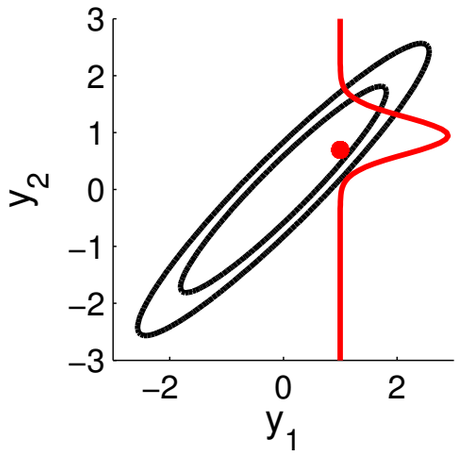
New Visualisation



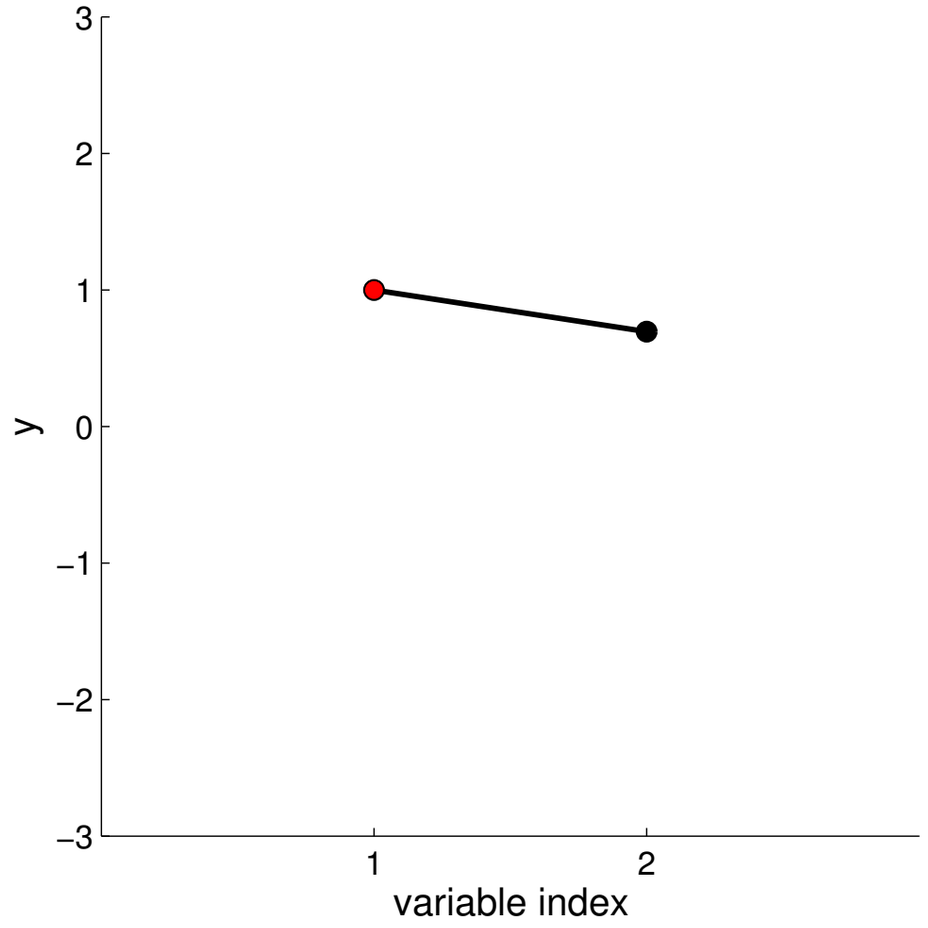
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



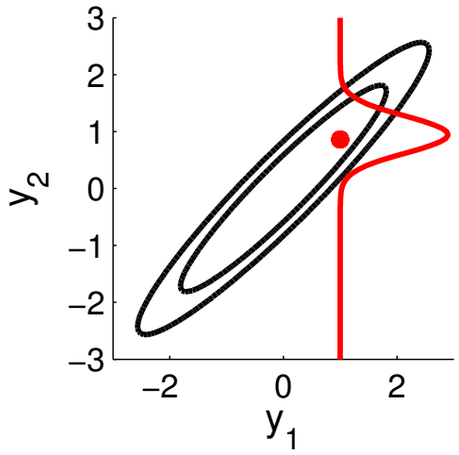
New Visualisation



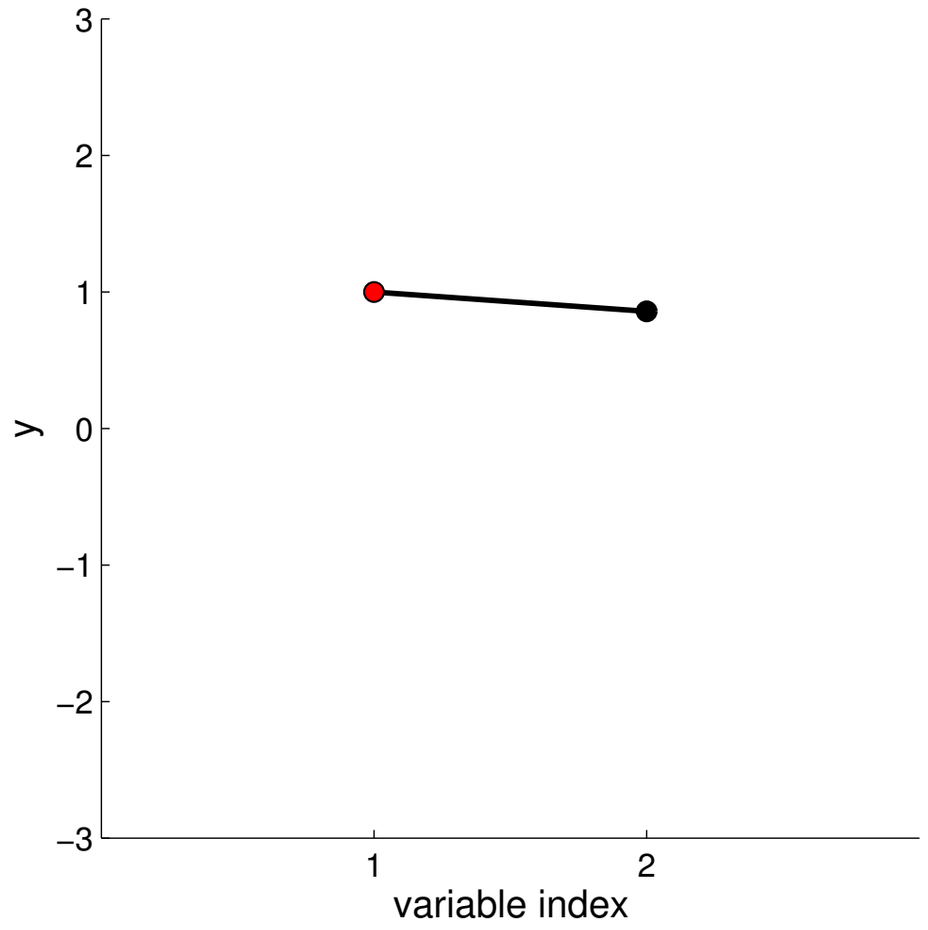
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



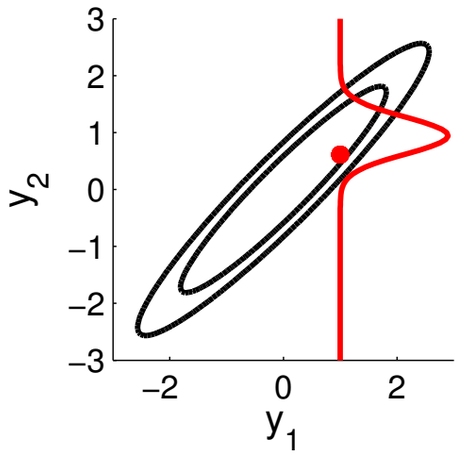
New Visualisation



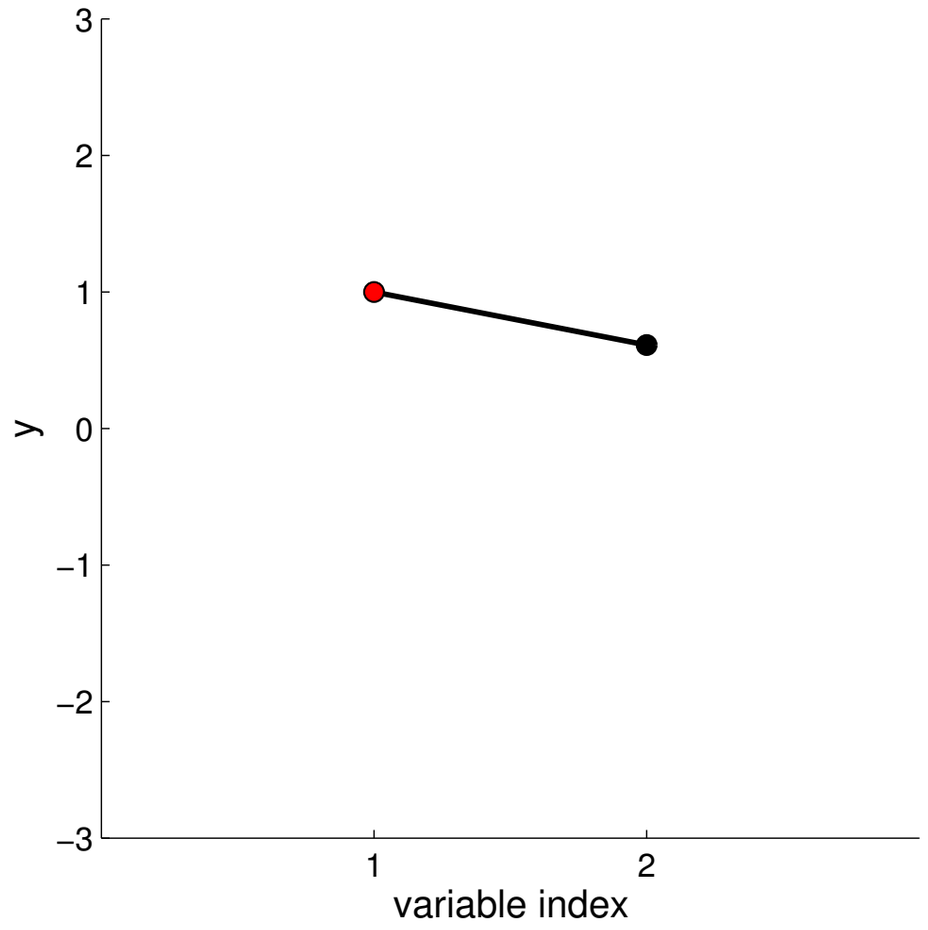
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



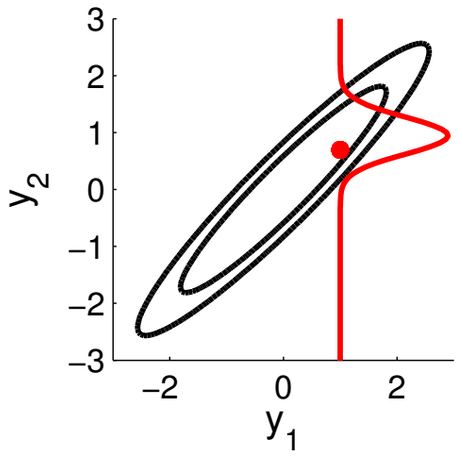
New Visualisation



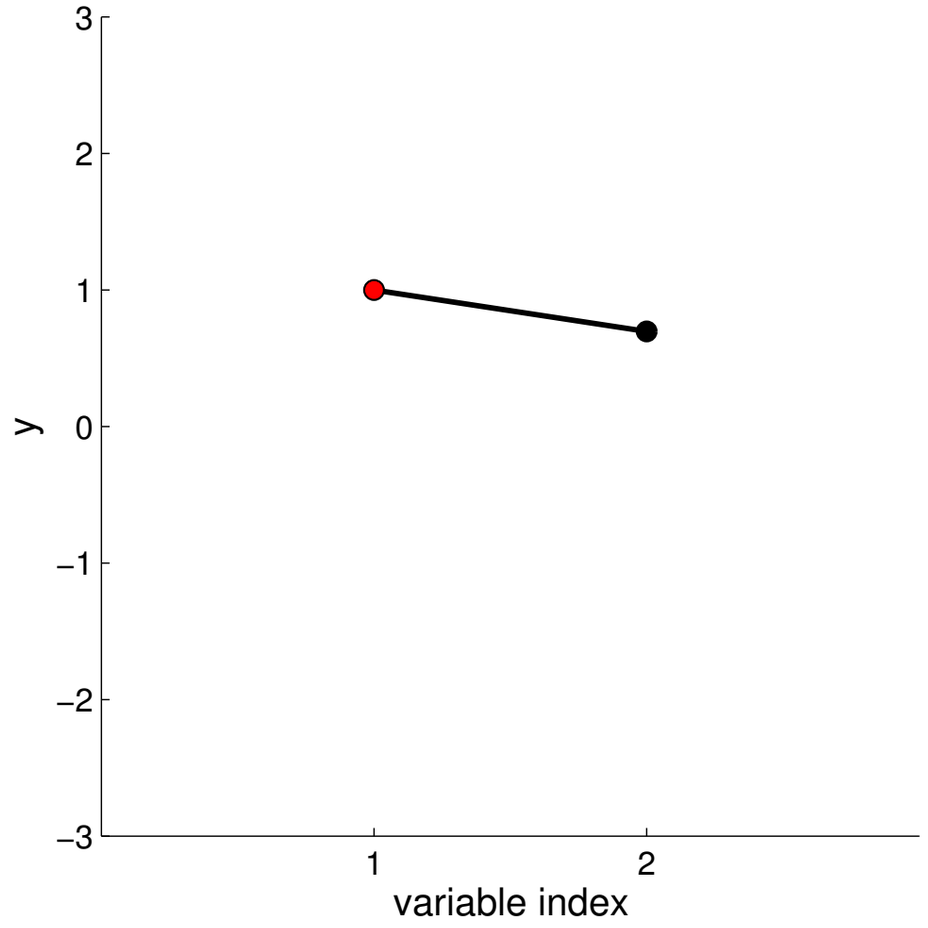
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



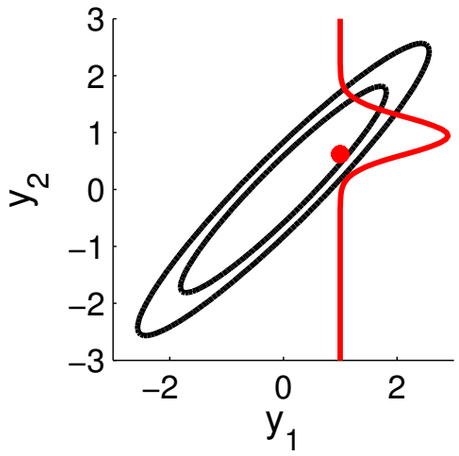
New Visualisation



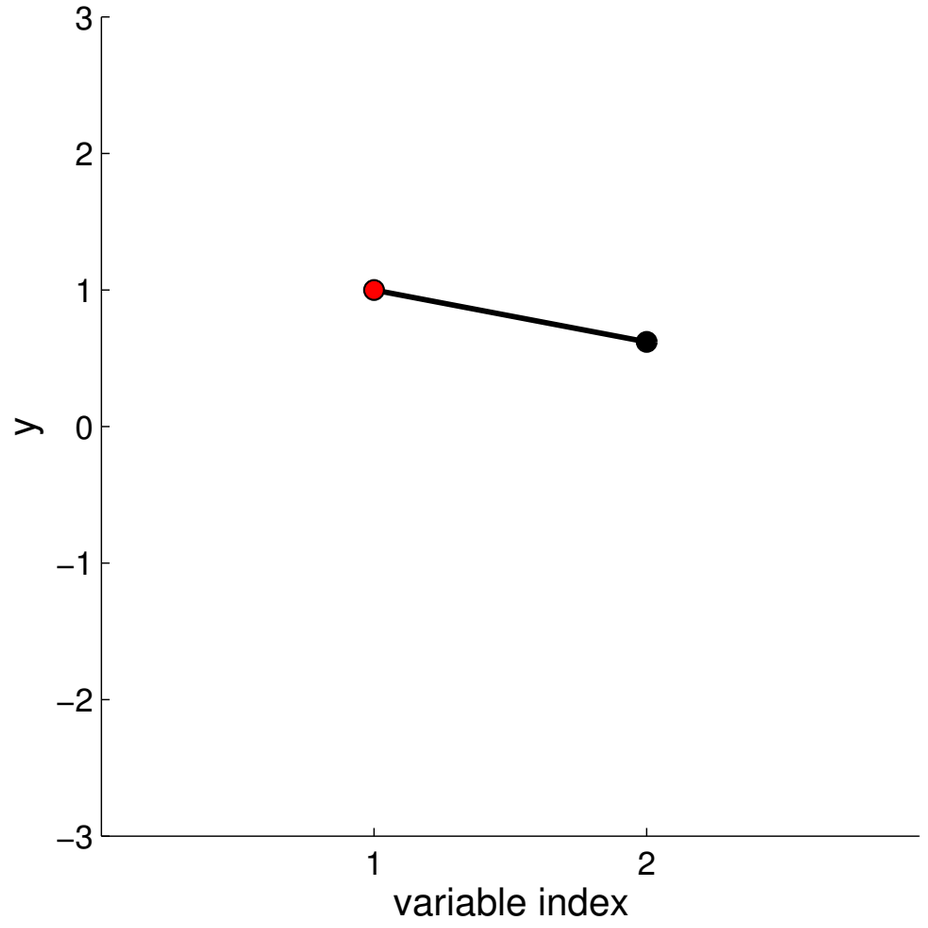
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



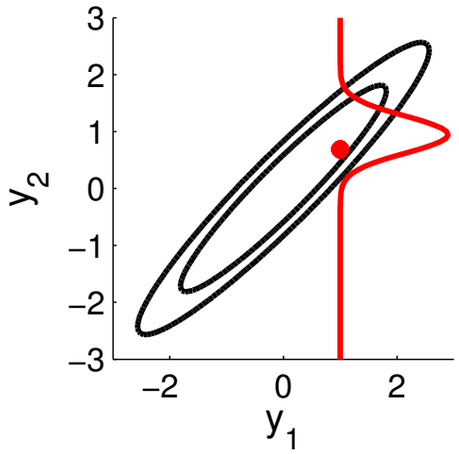
New Visualisation



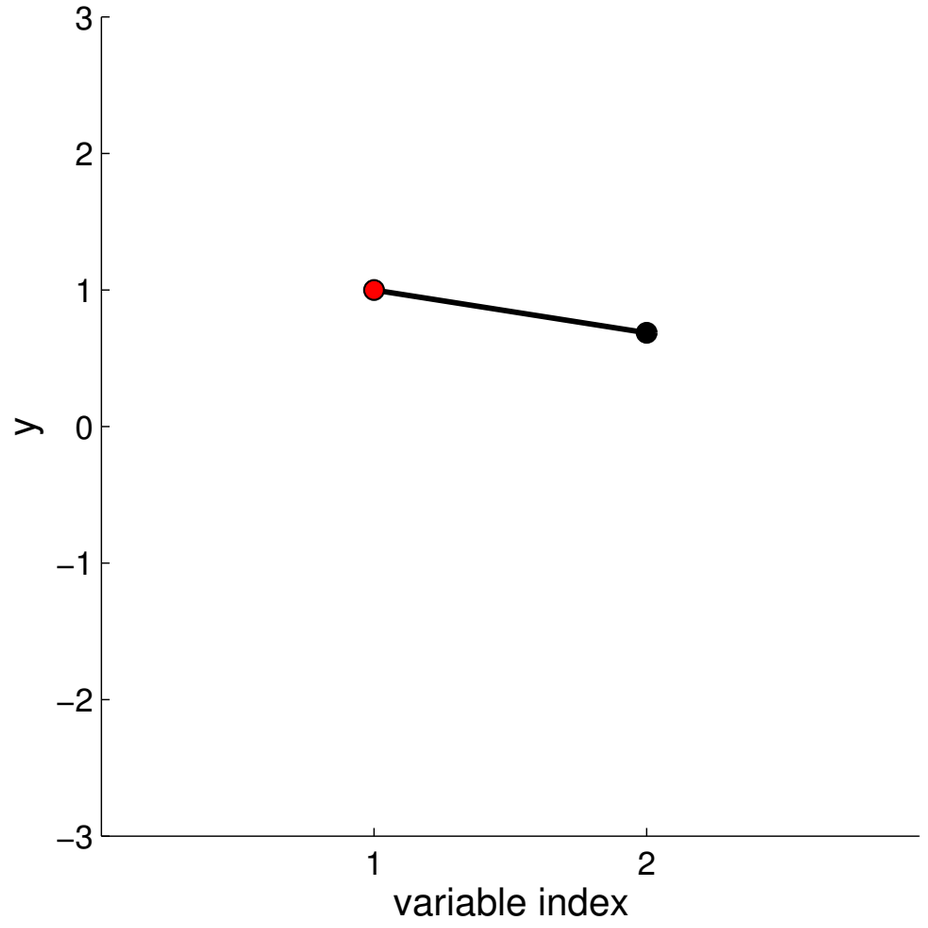
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



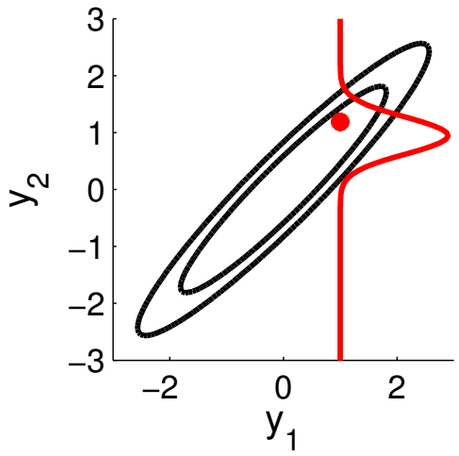
New Visualisation



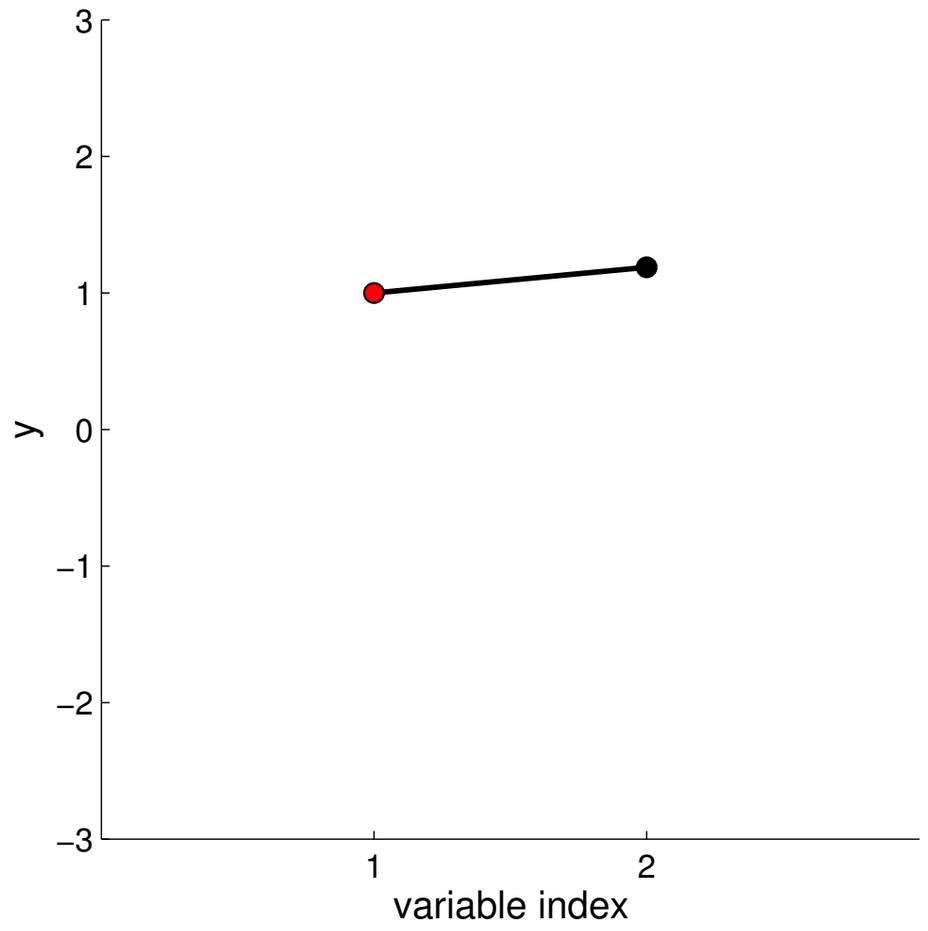
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



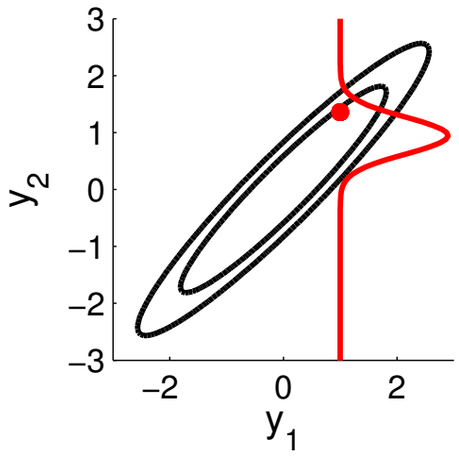
New Visualisation



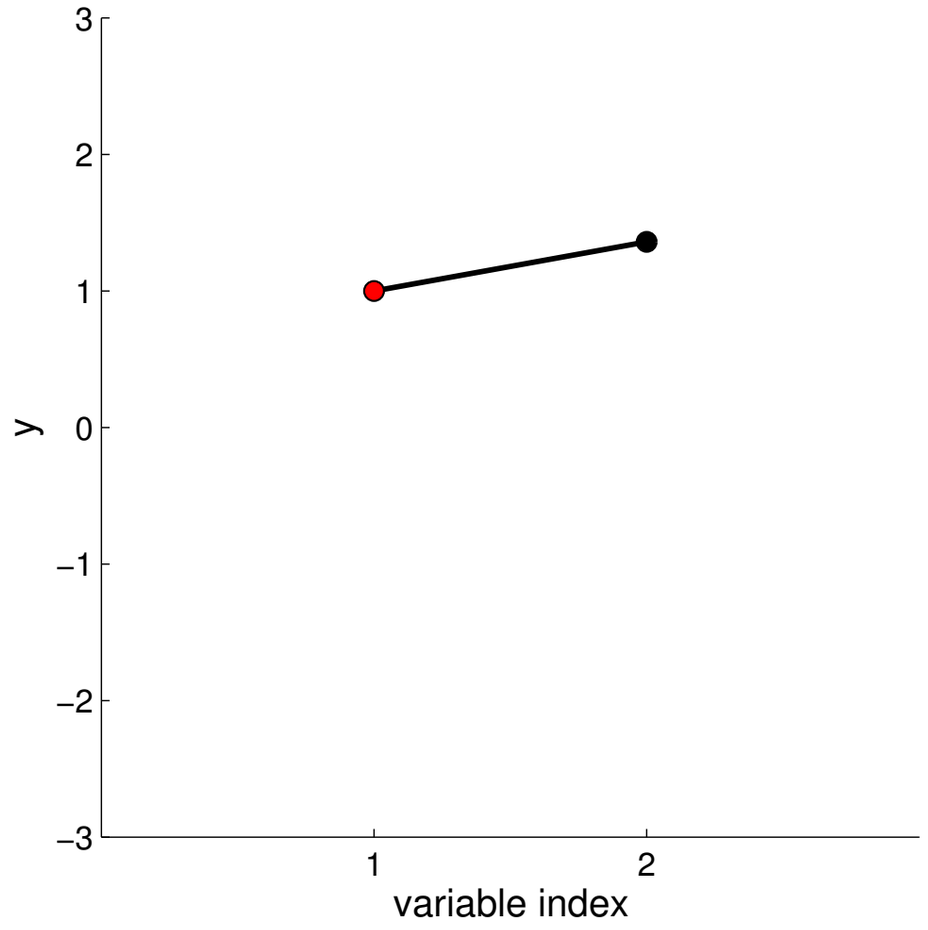
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



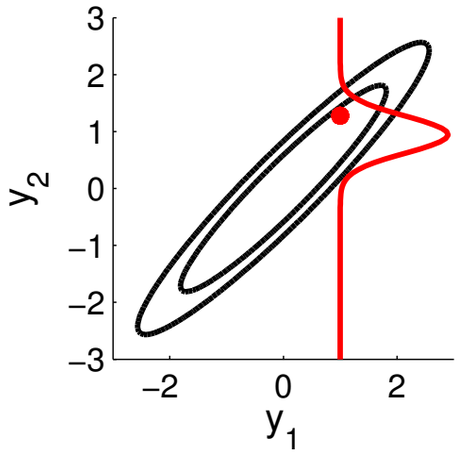
New Visualisation



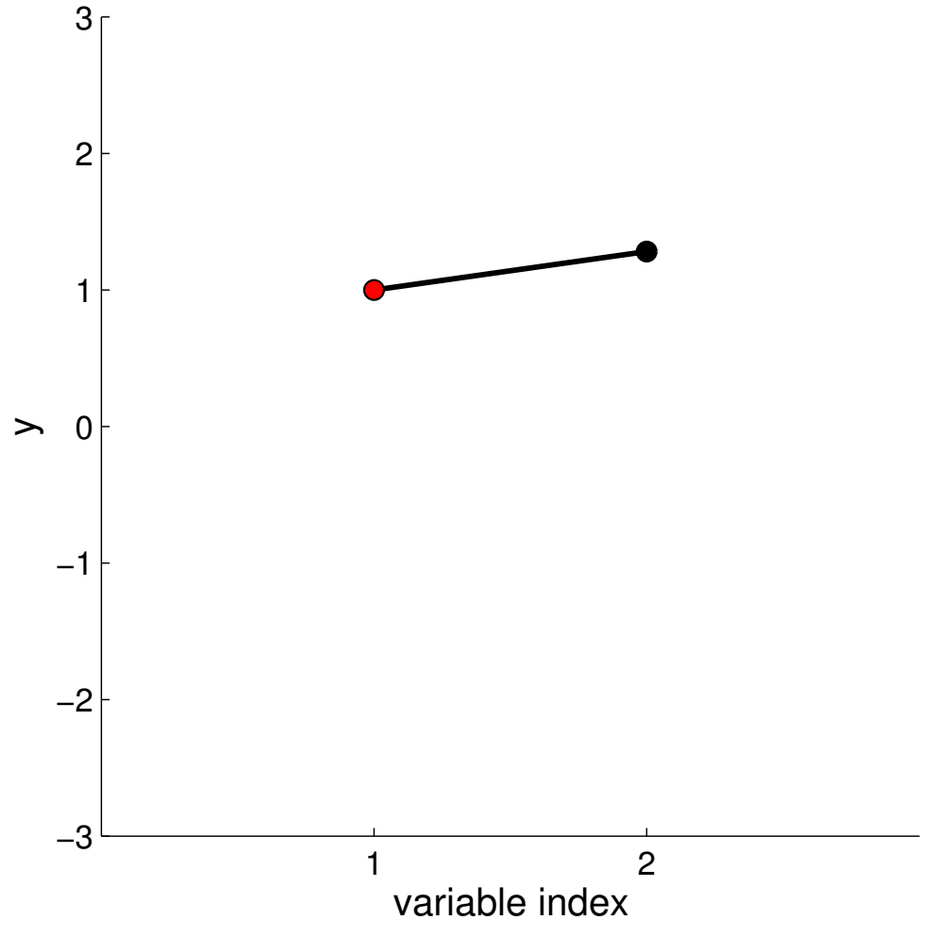
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



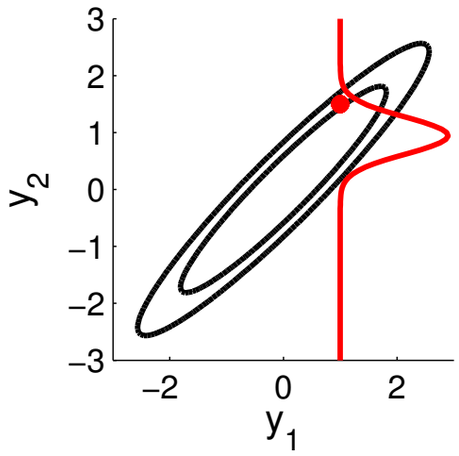
New Visualisation



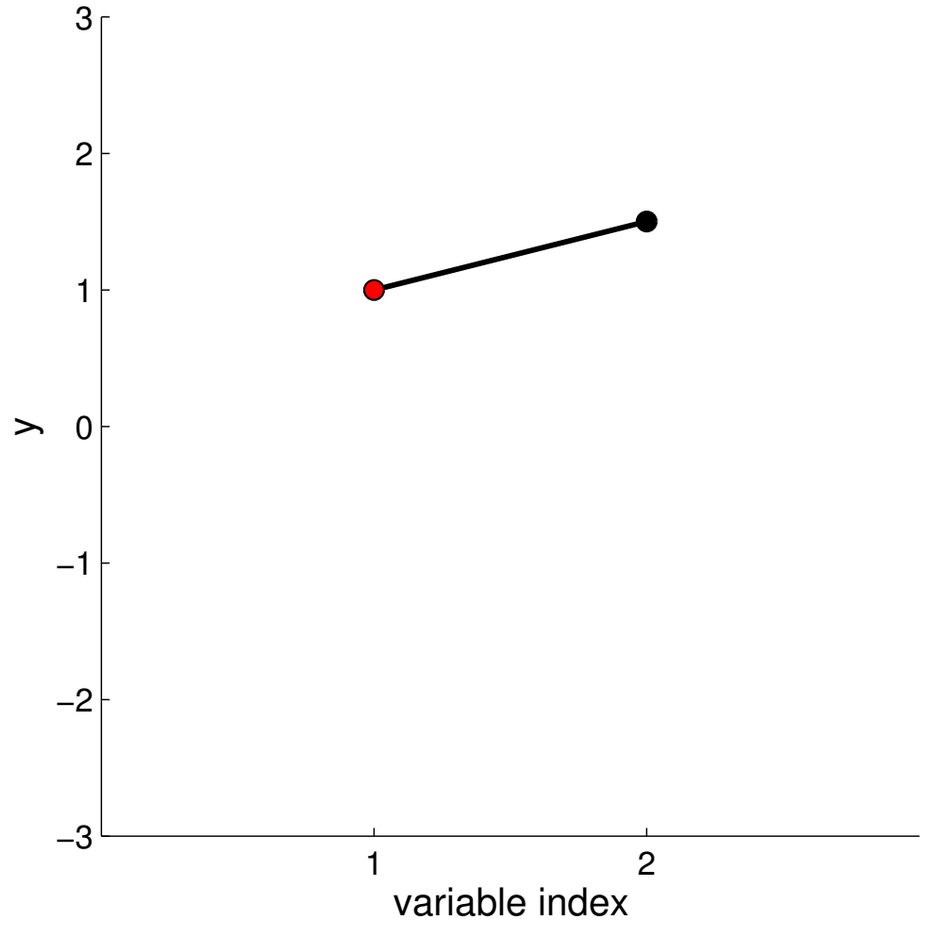
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



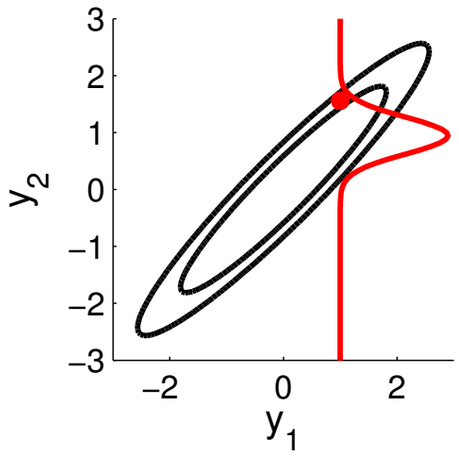
New Visualisation



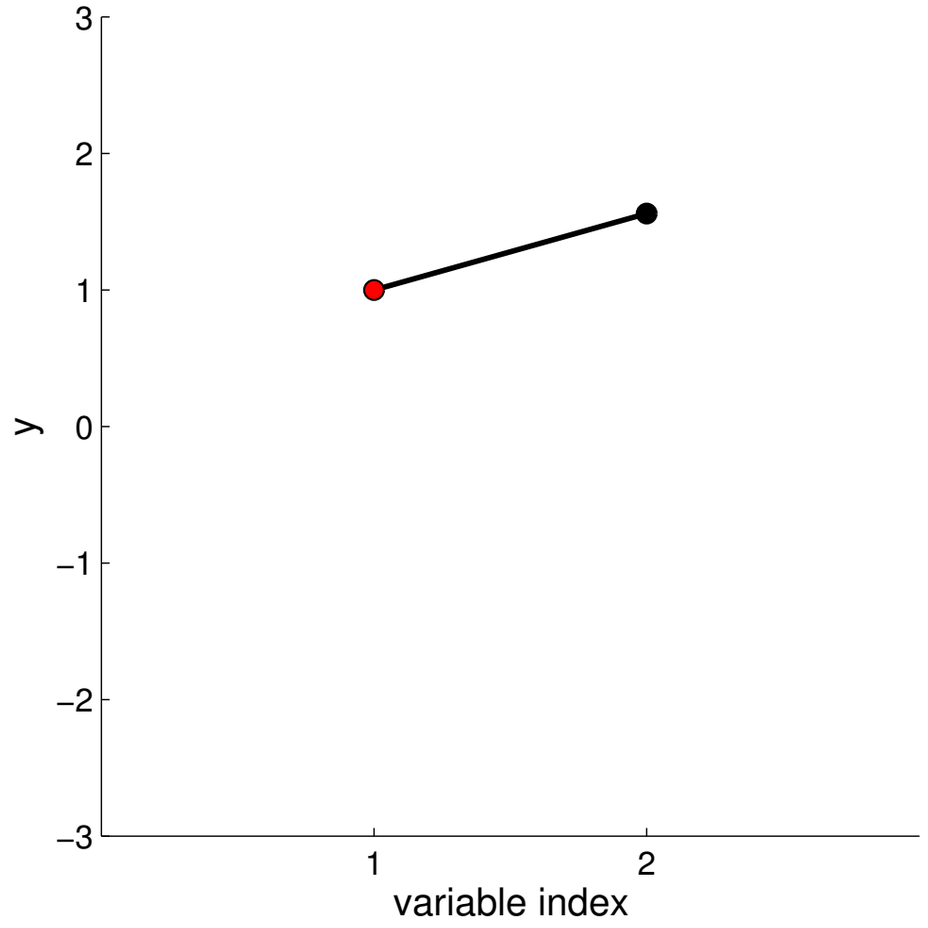
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



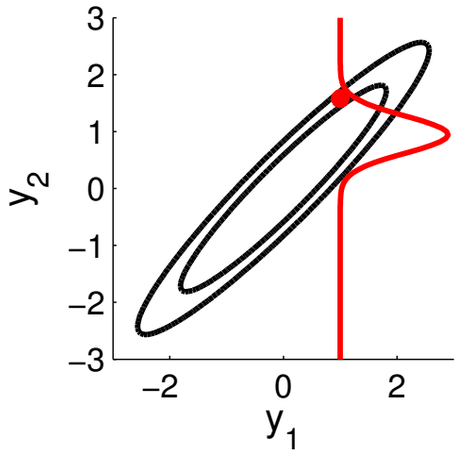
New Visualisation



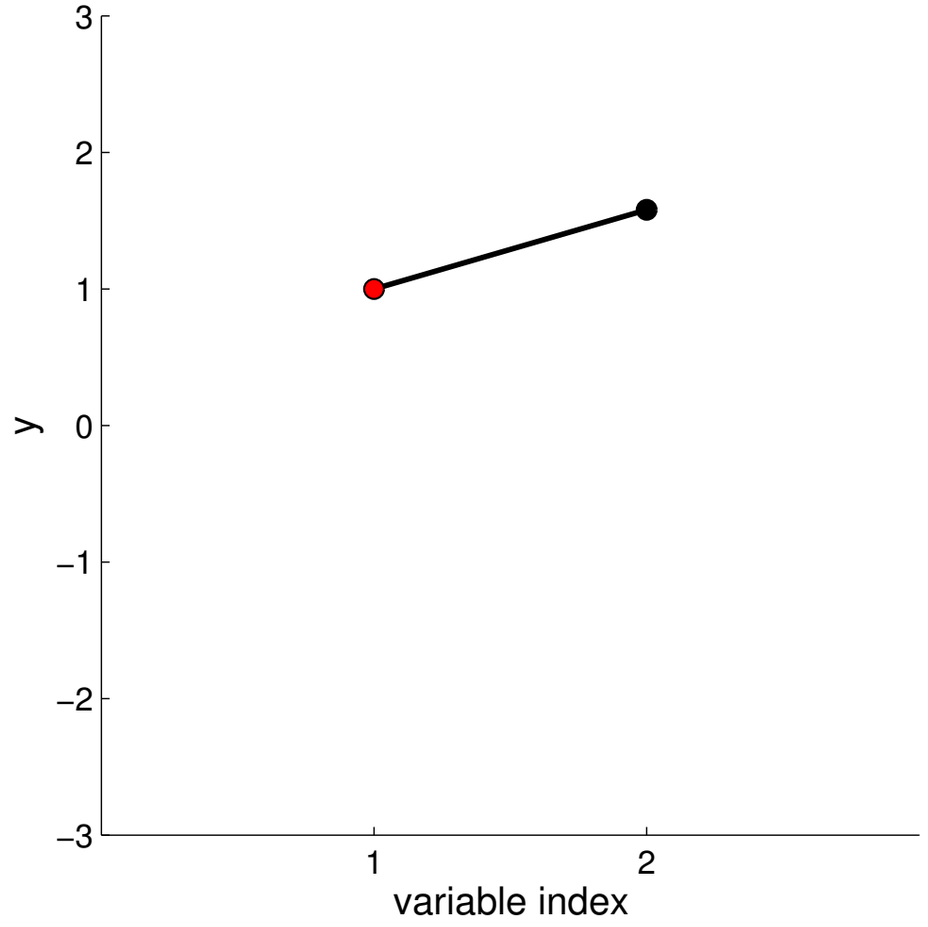
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



New Visualisation

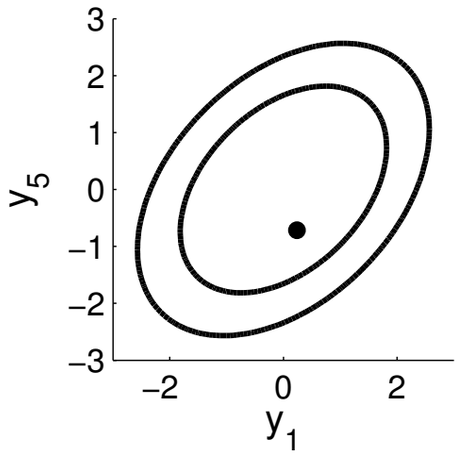


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

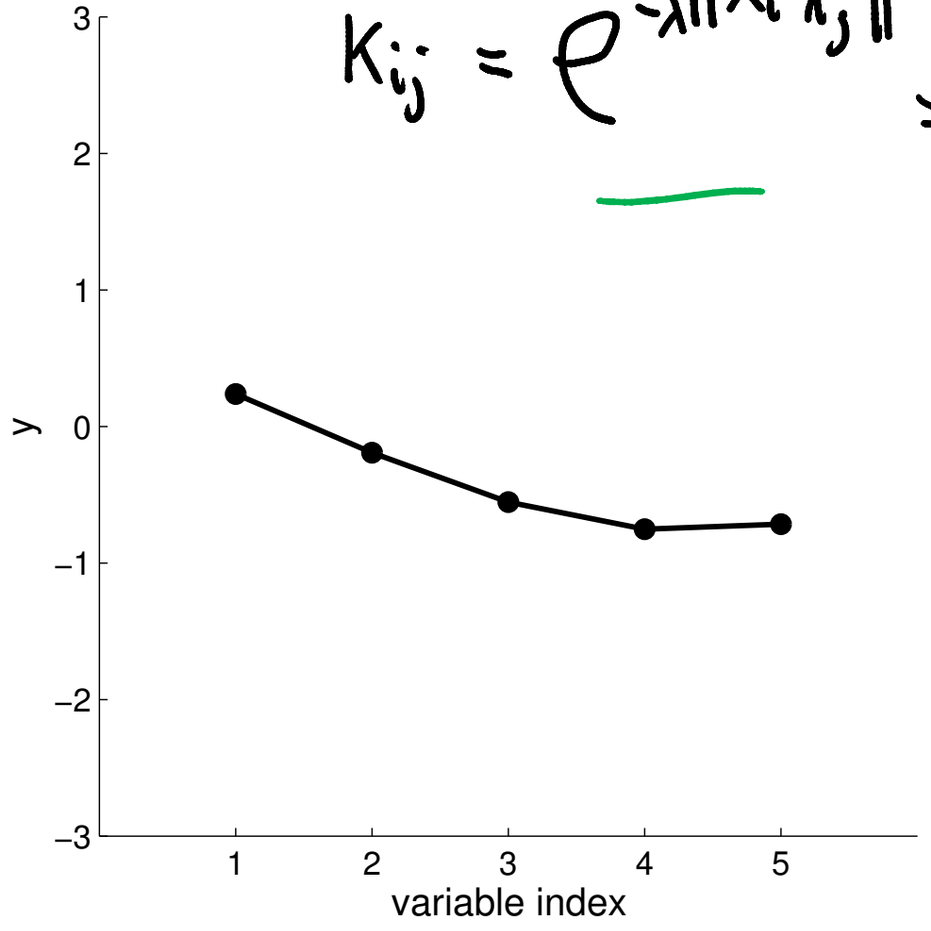


Special covariance matrix

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



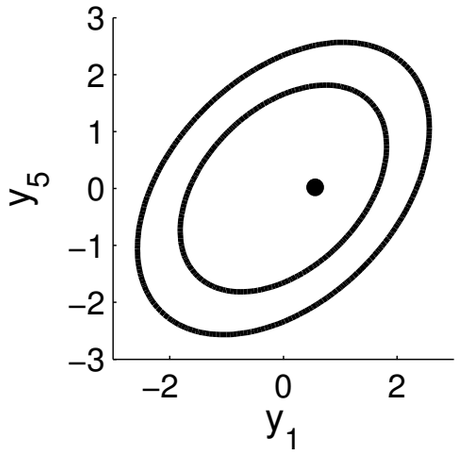
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

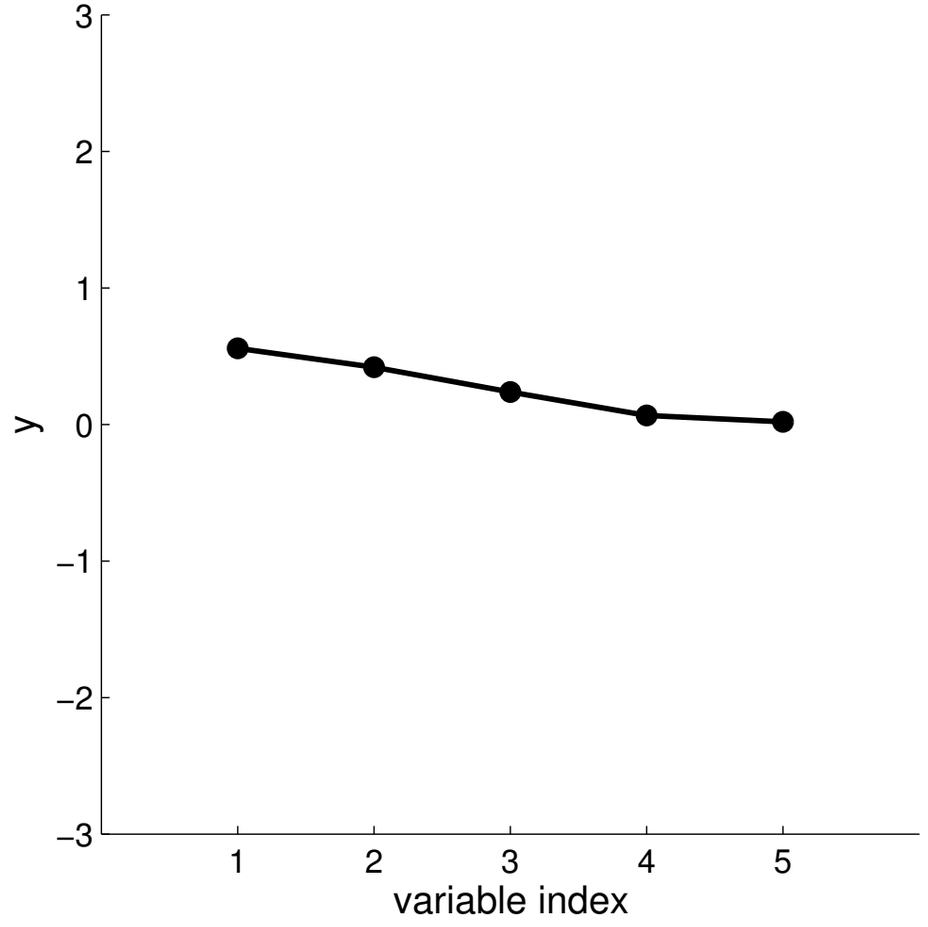


$$k_{ij} = e^{-\lambda \|x_i - x_j\|^2} \simeq \begin{cases} 0 & \|x_i - x_j\| \to \infty \\ 1 & x_i = x_j \end{cases}$$

Special covariance matrix

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

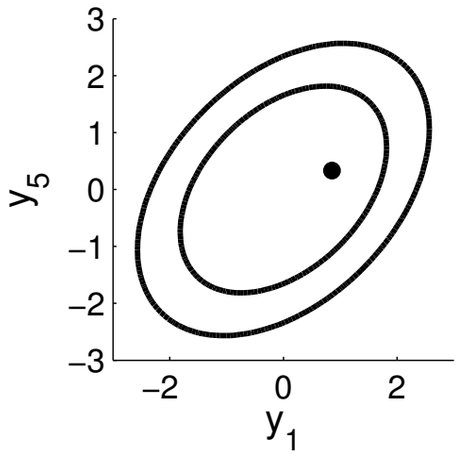


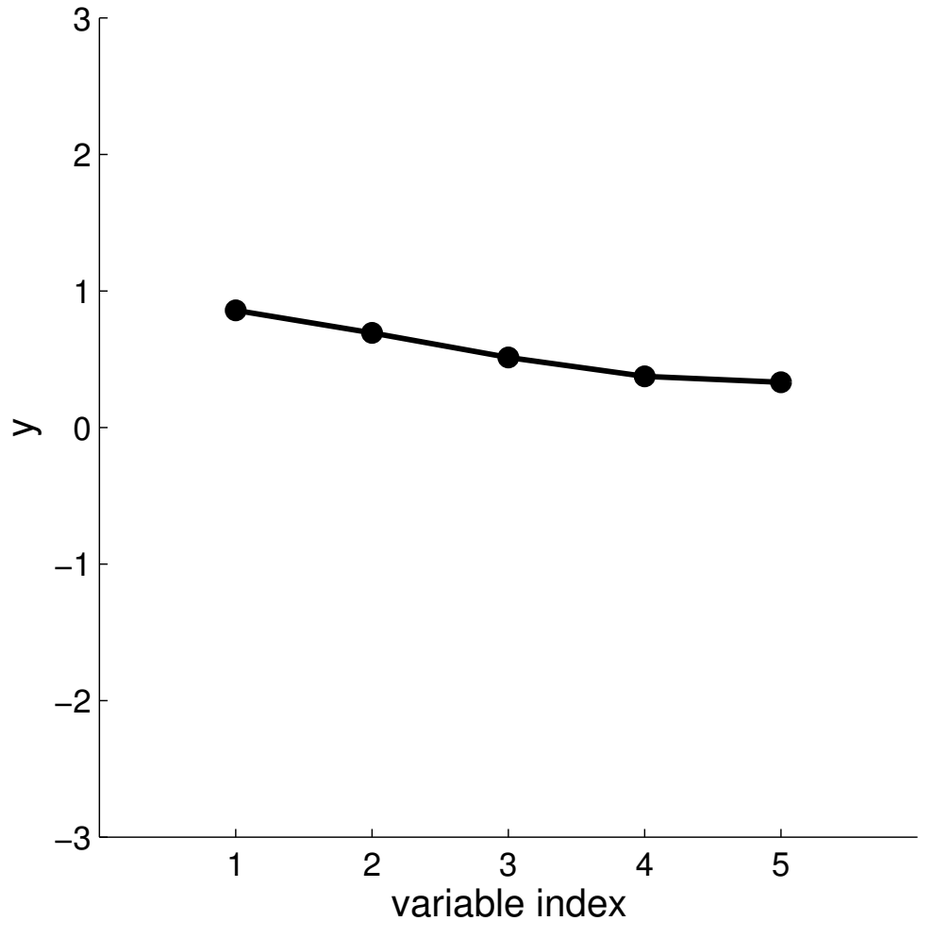


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

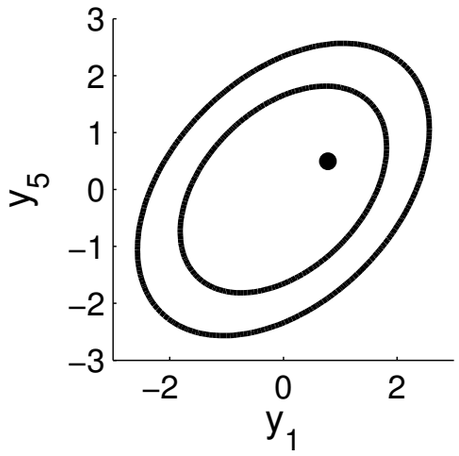


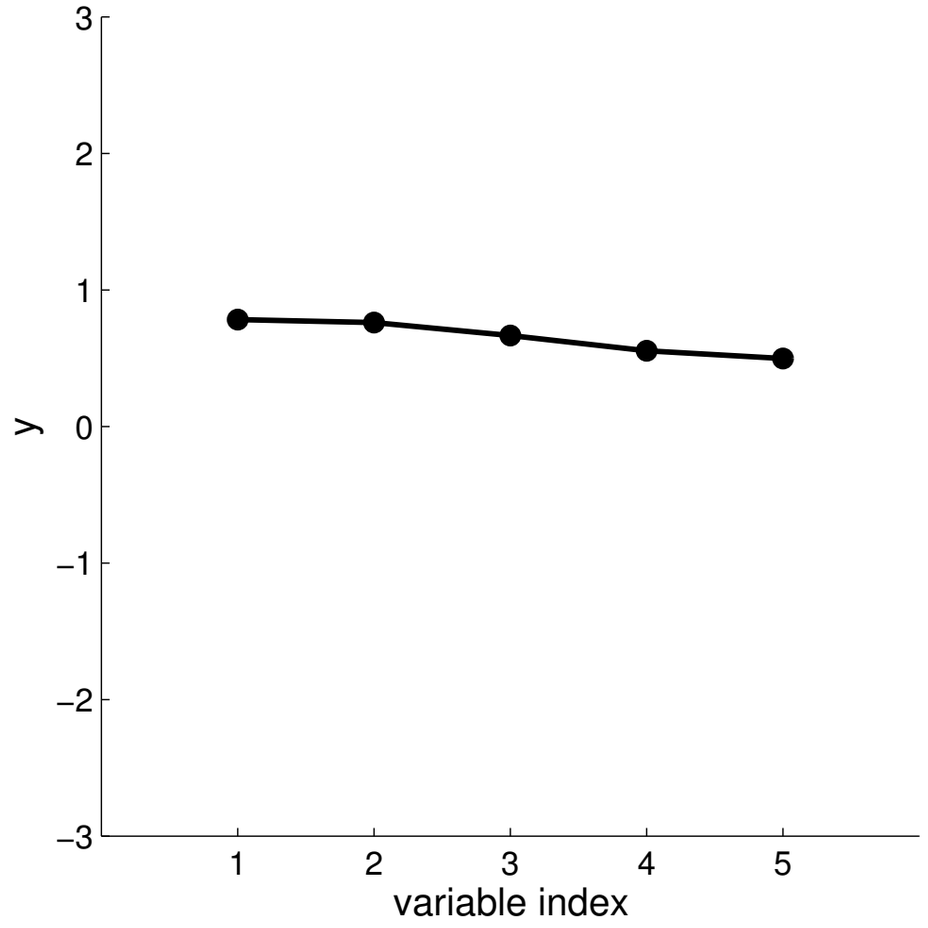


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

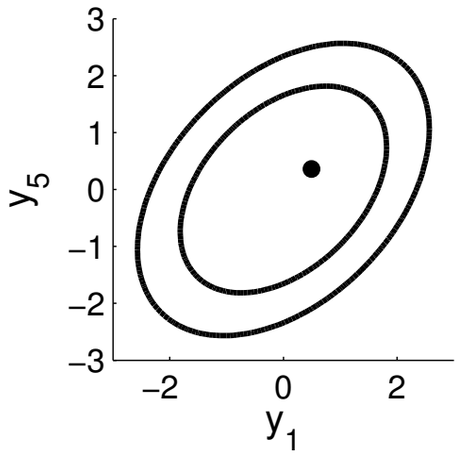


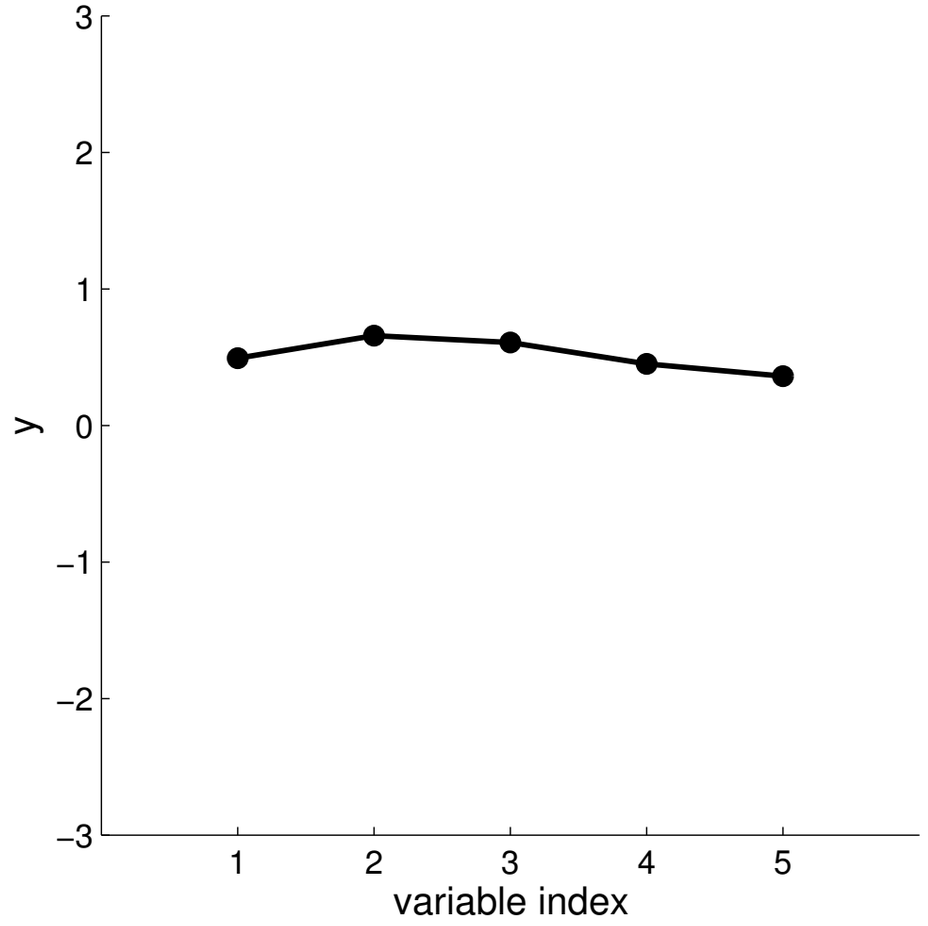


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

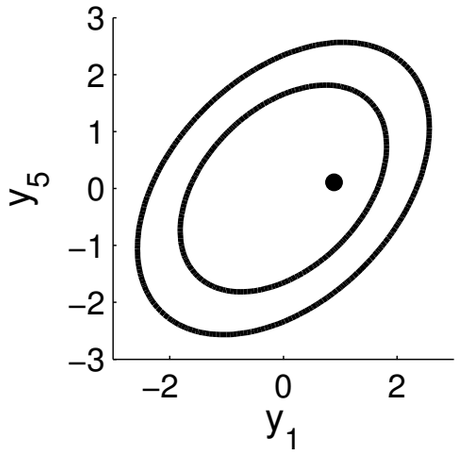


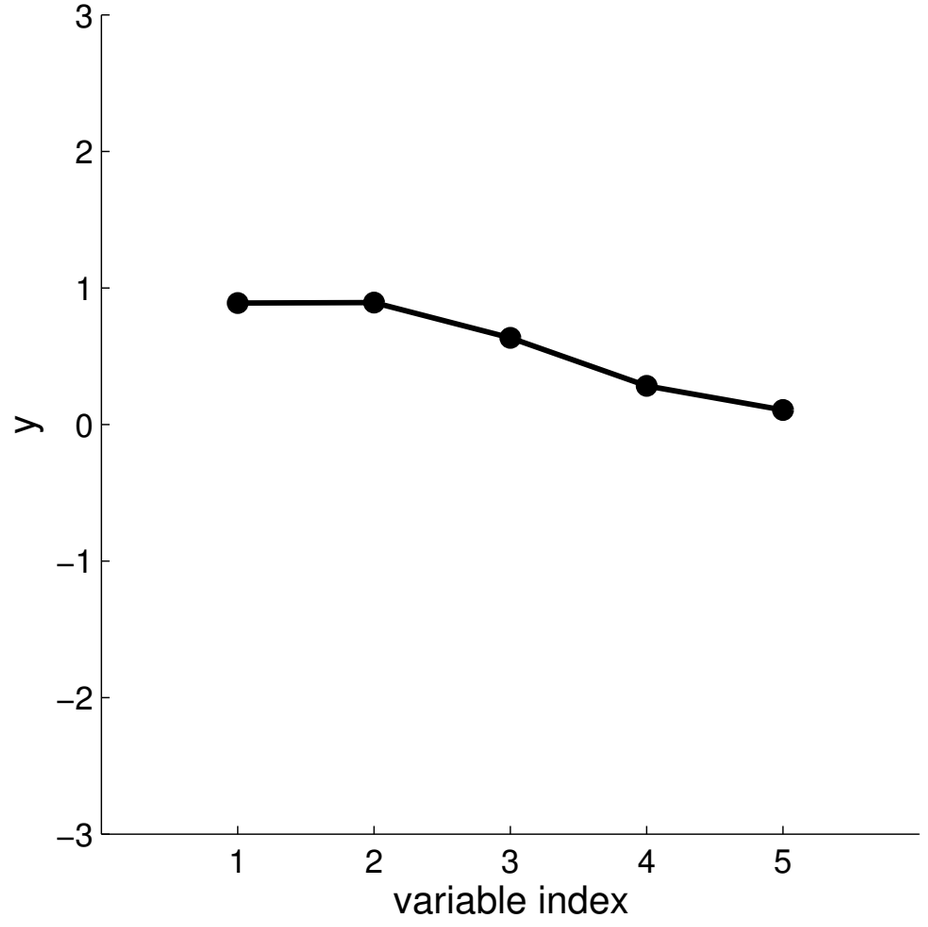


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

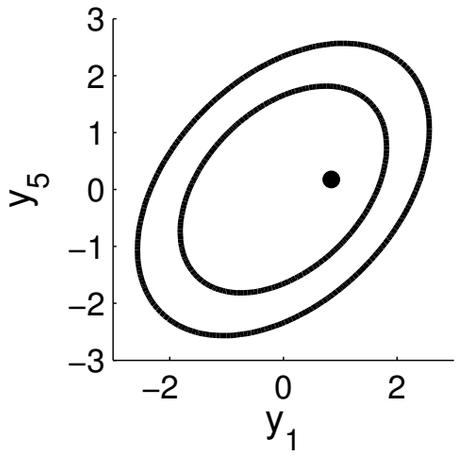


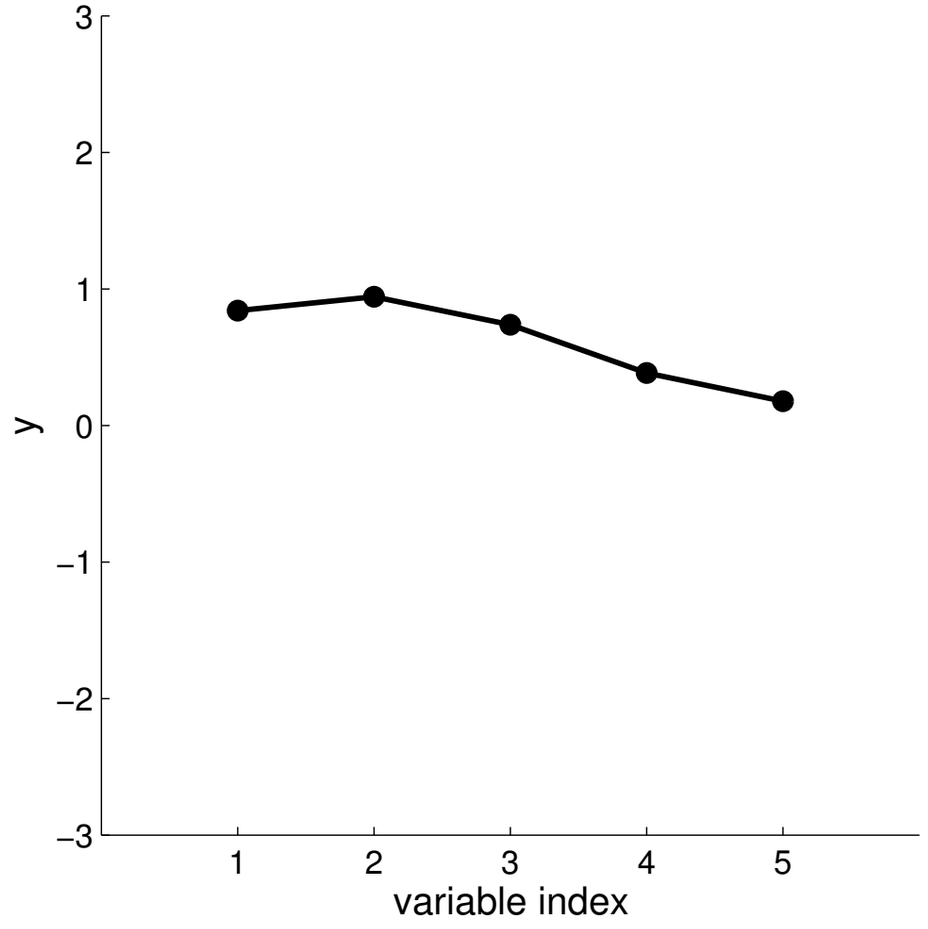


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

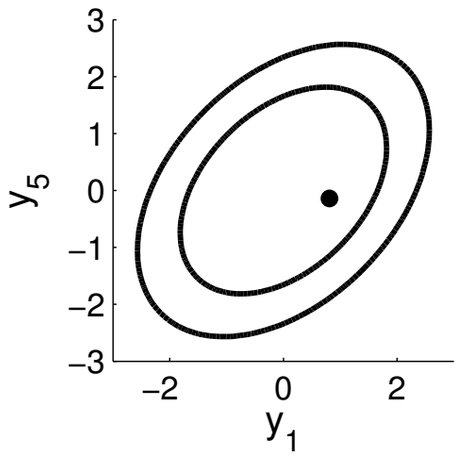


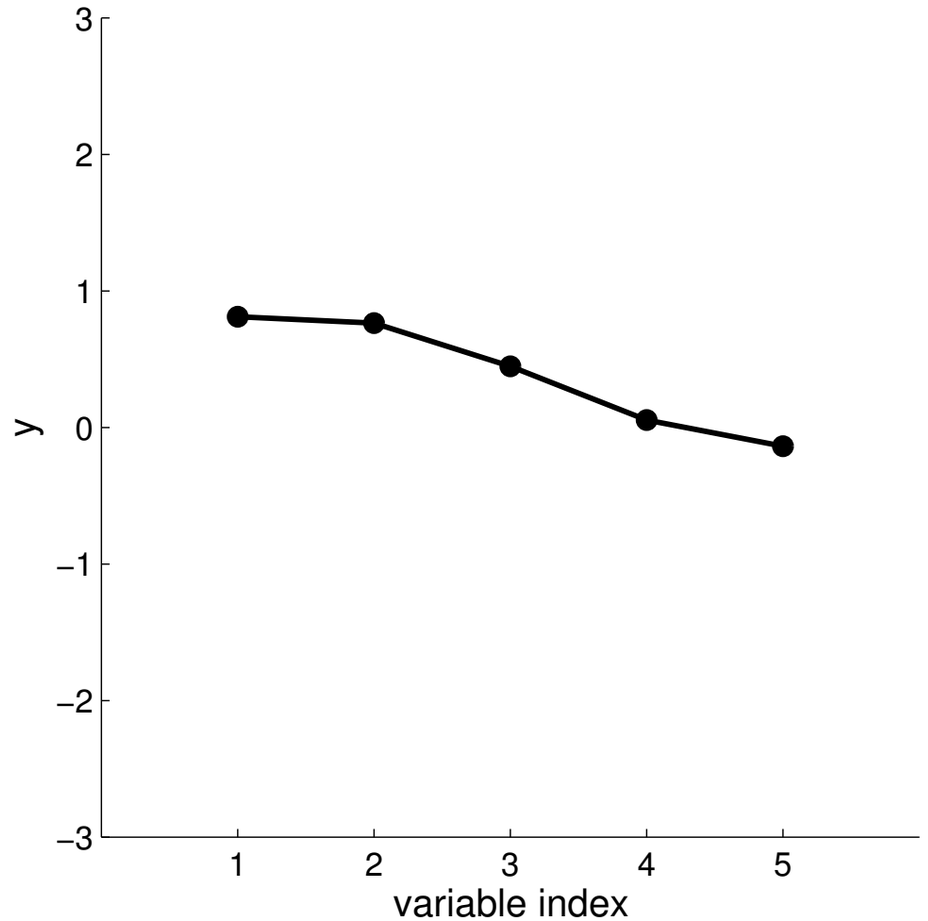


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

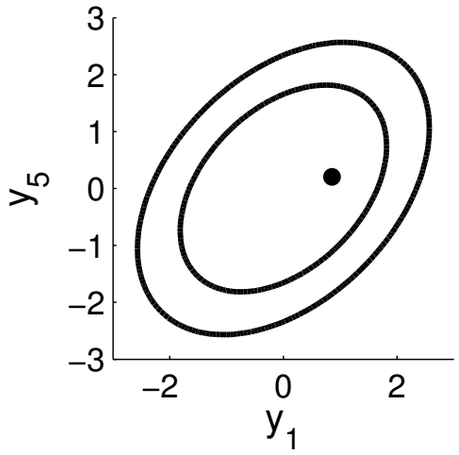


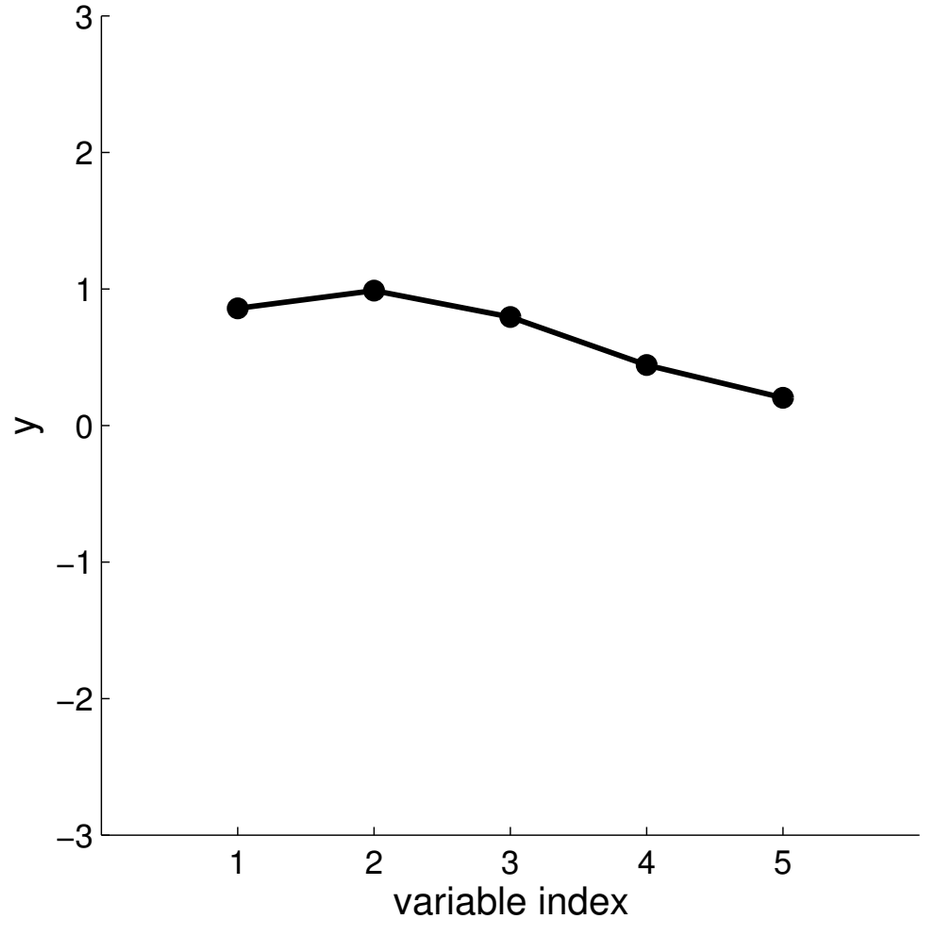


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

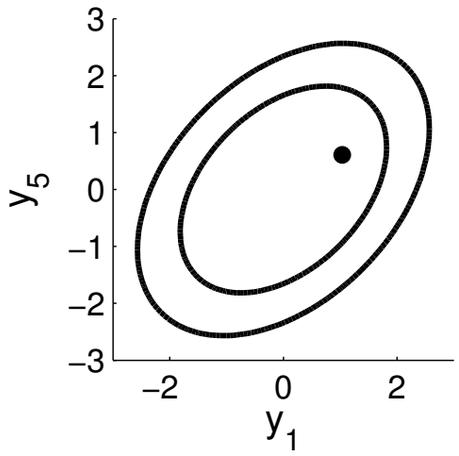


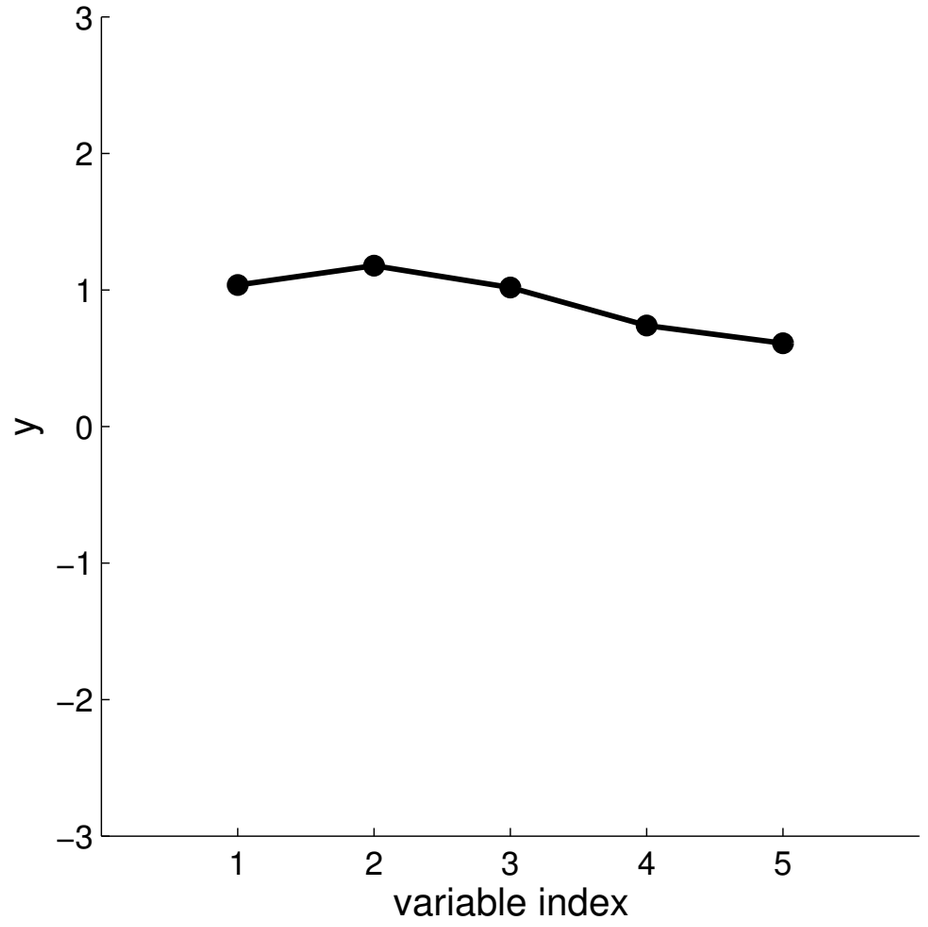


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

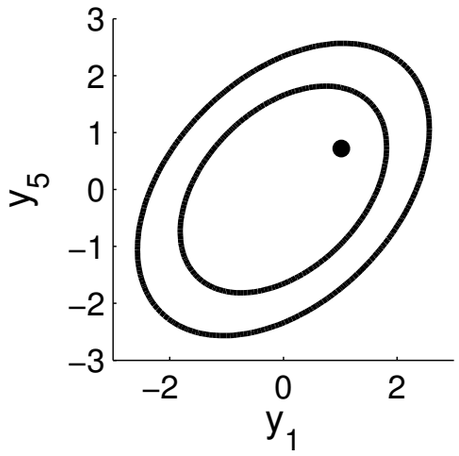


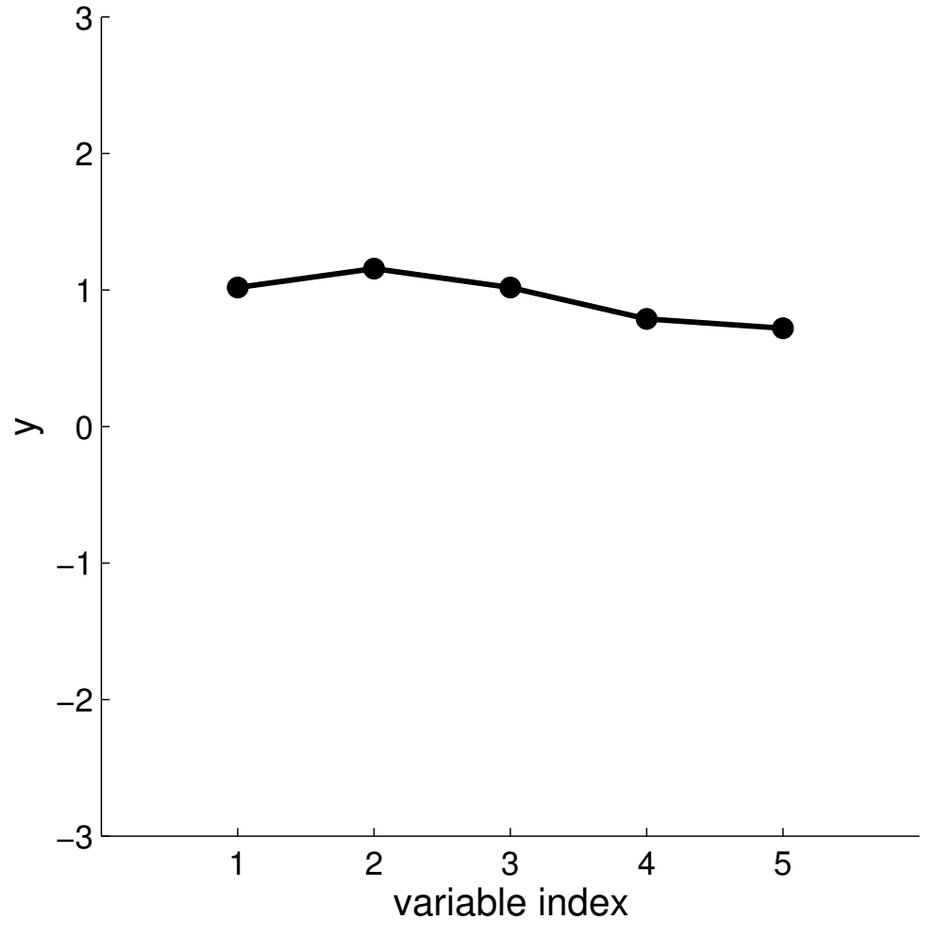


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

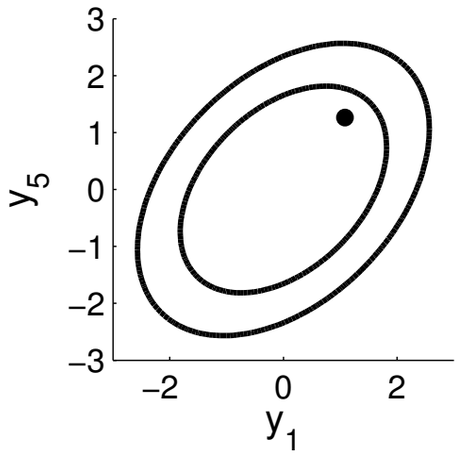


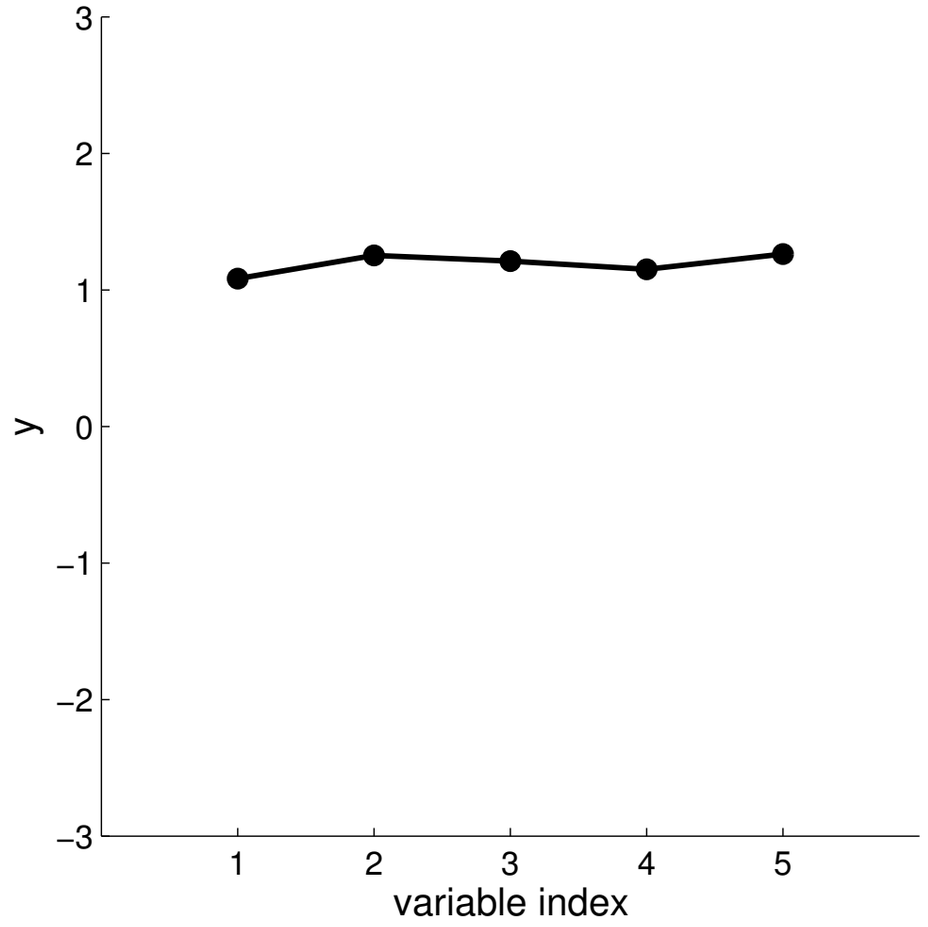


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

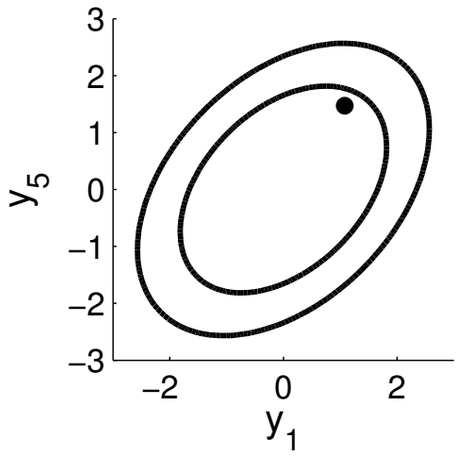


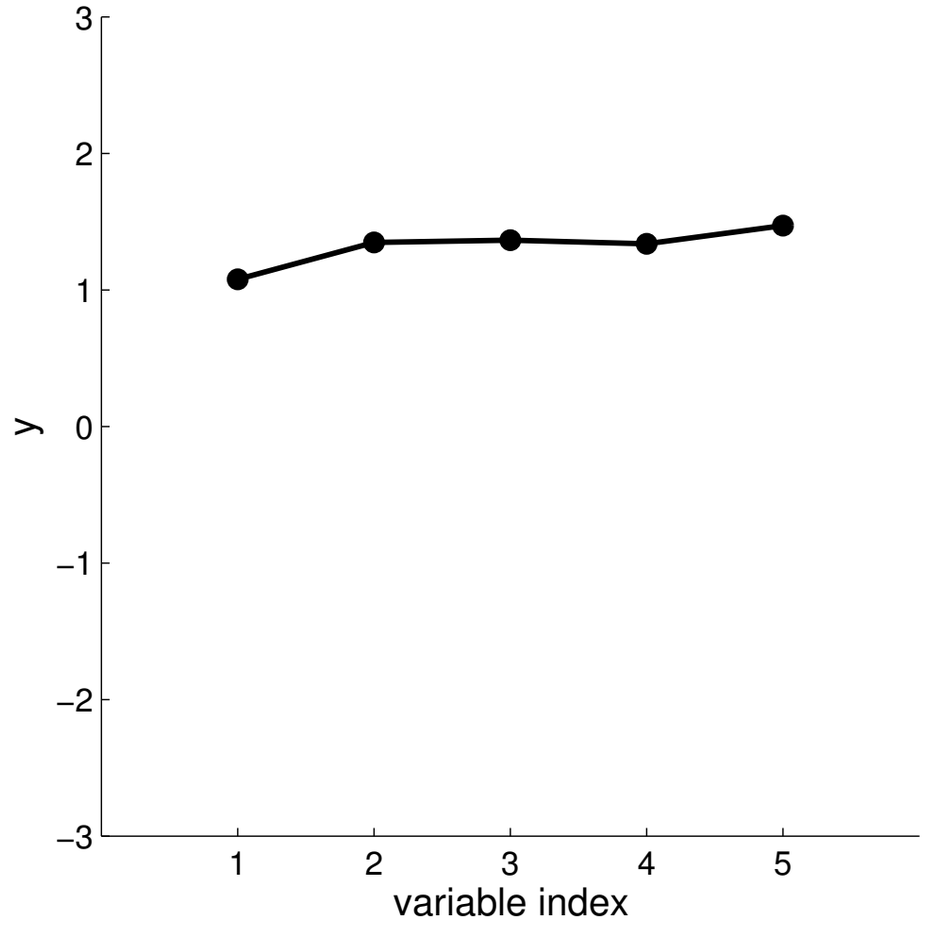


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

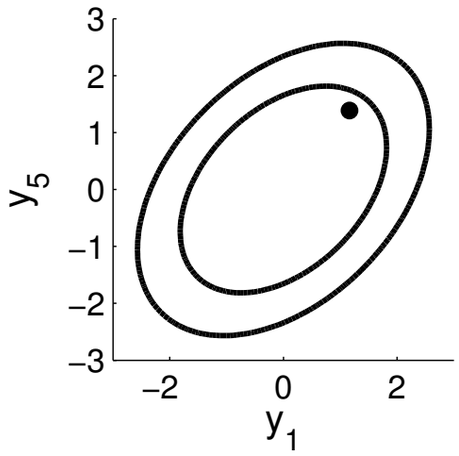


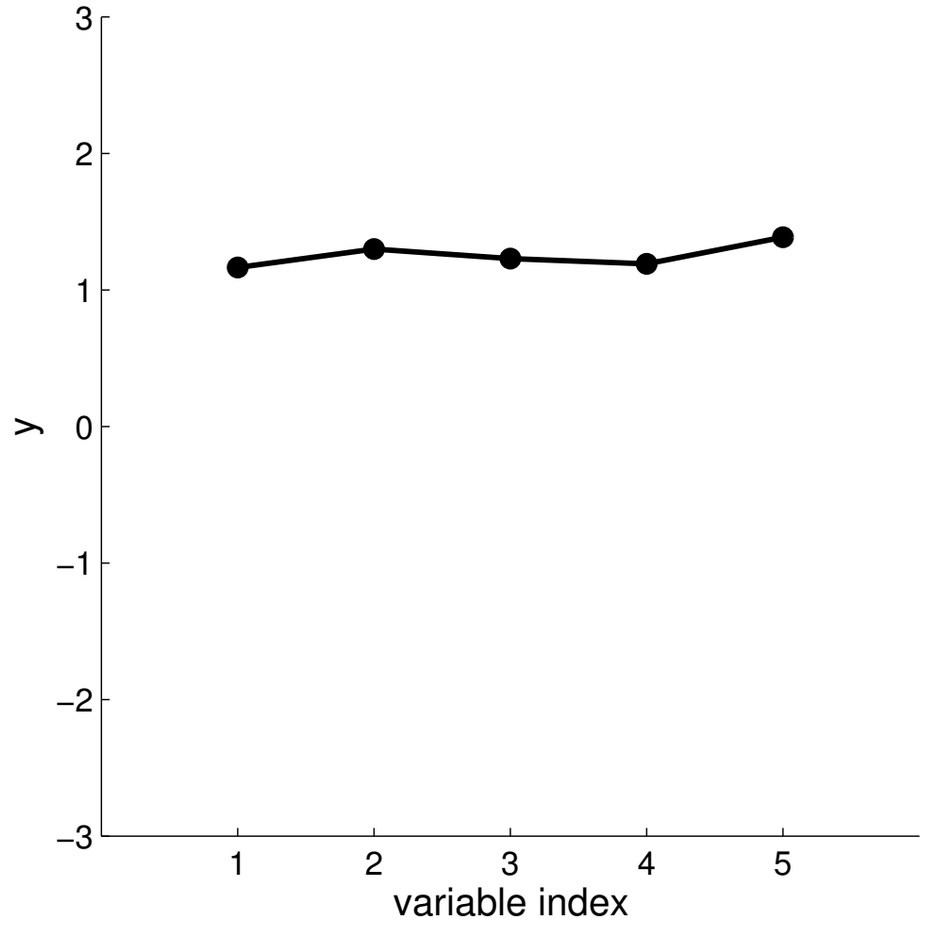


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

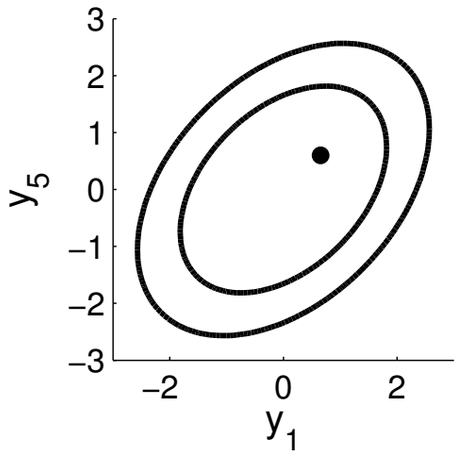


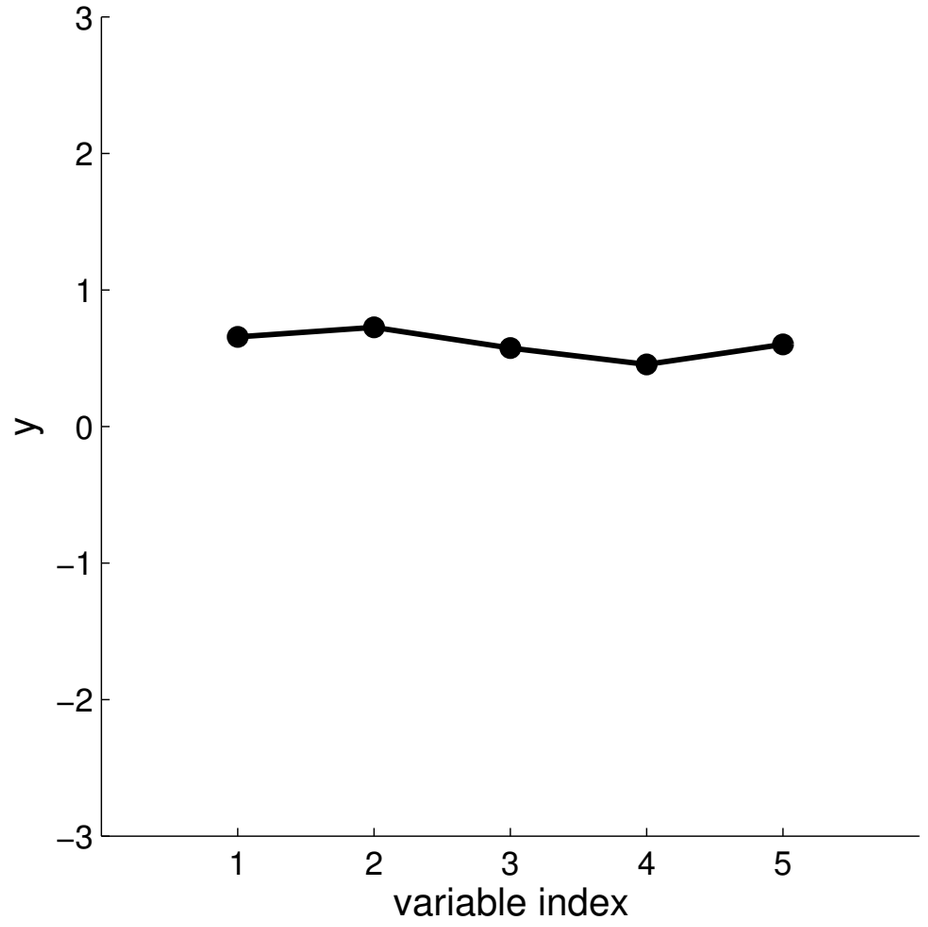


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

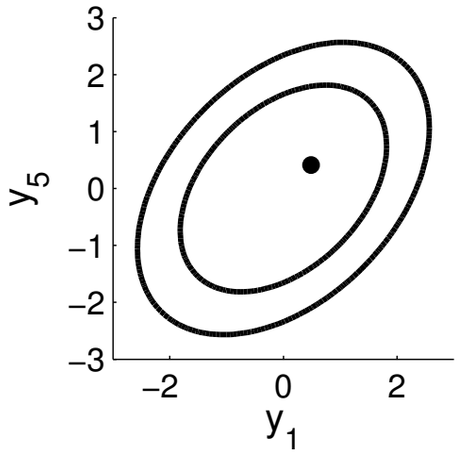


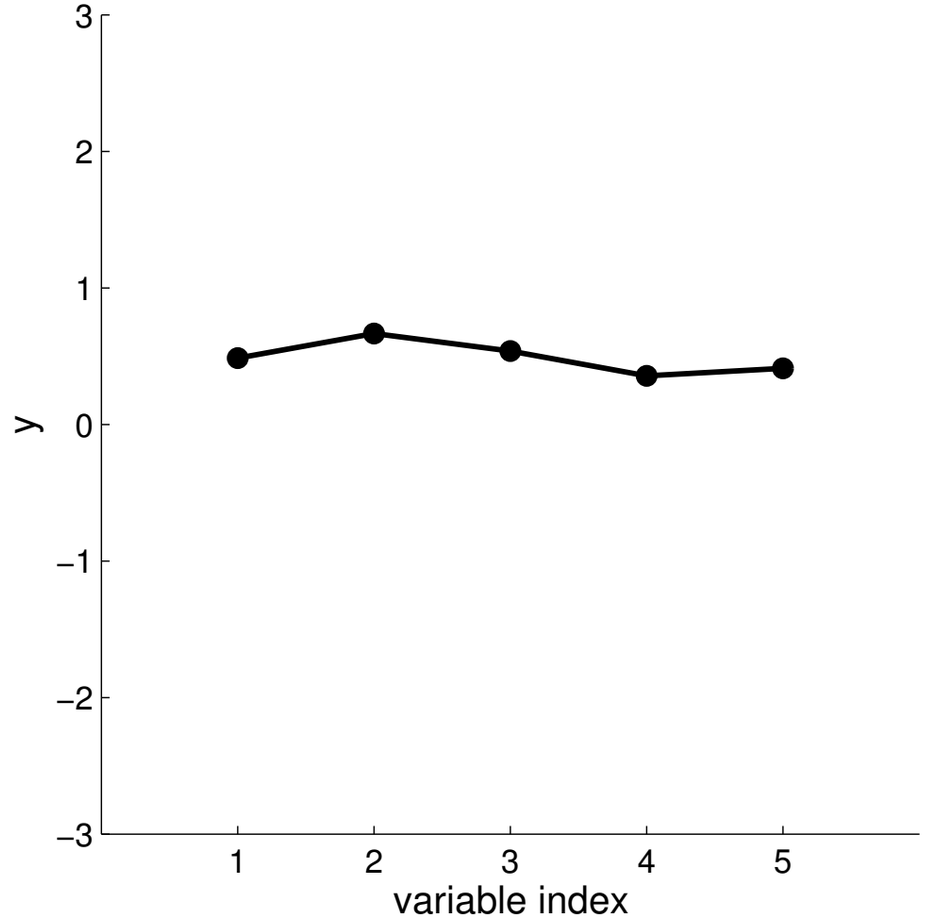


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

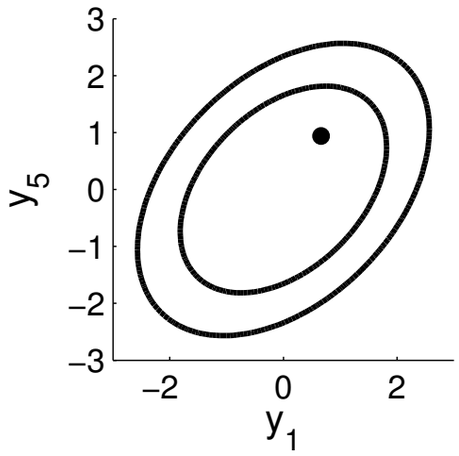


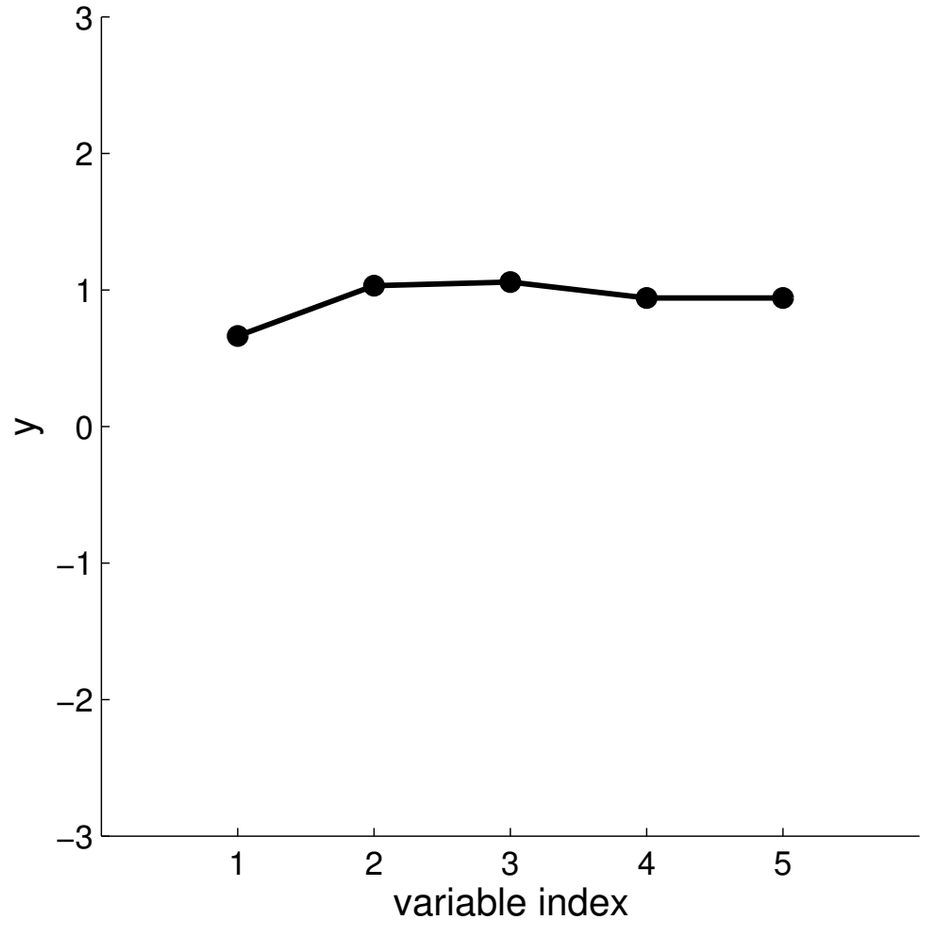


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

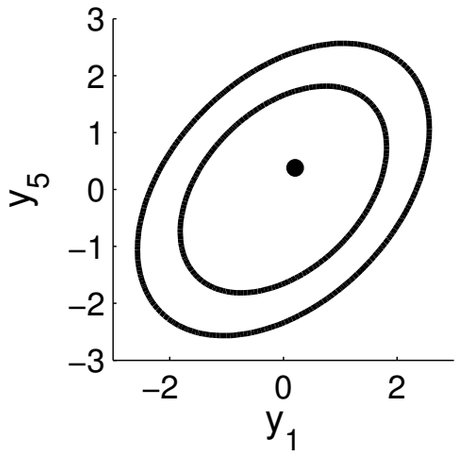


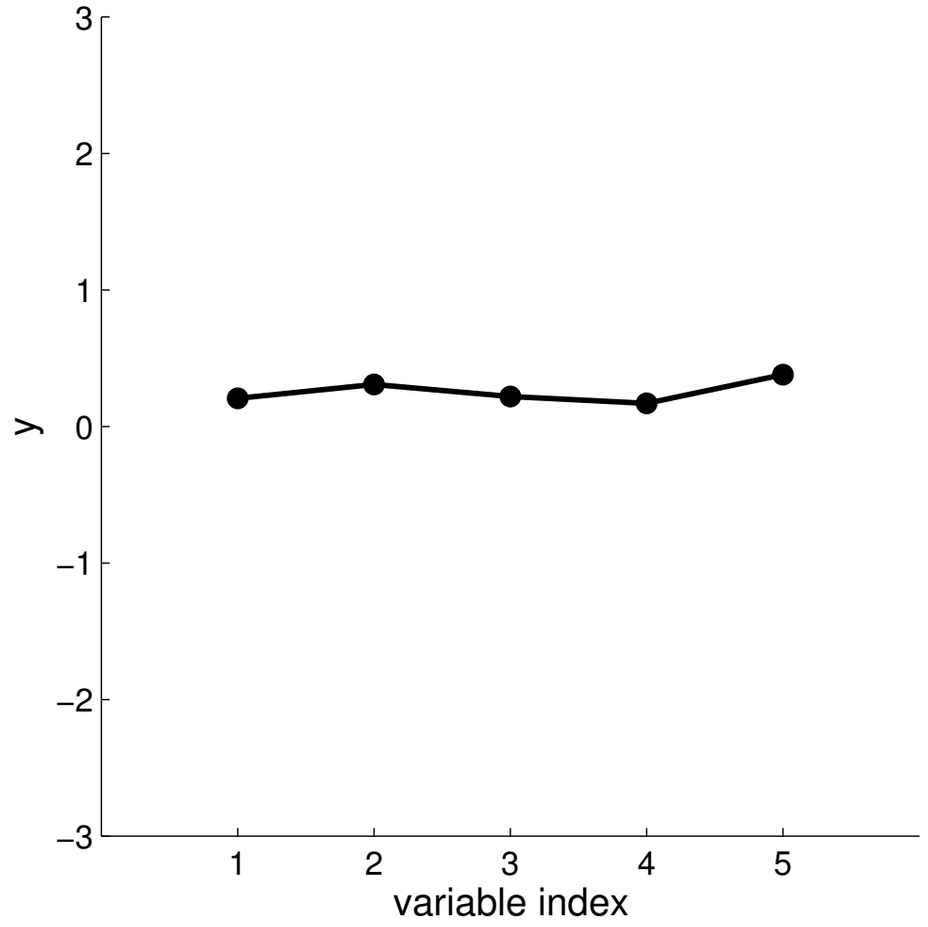


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

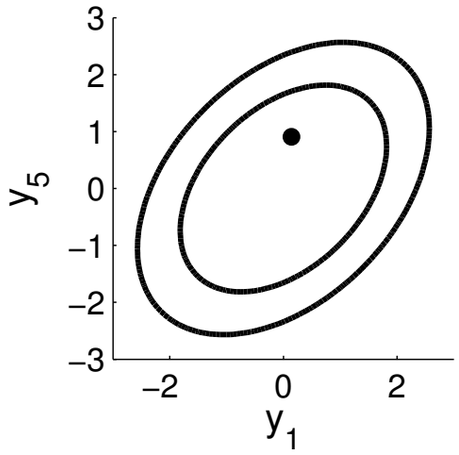


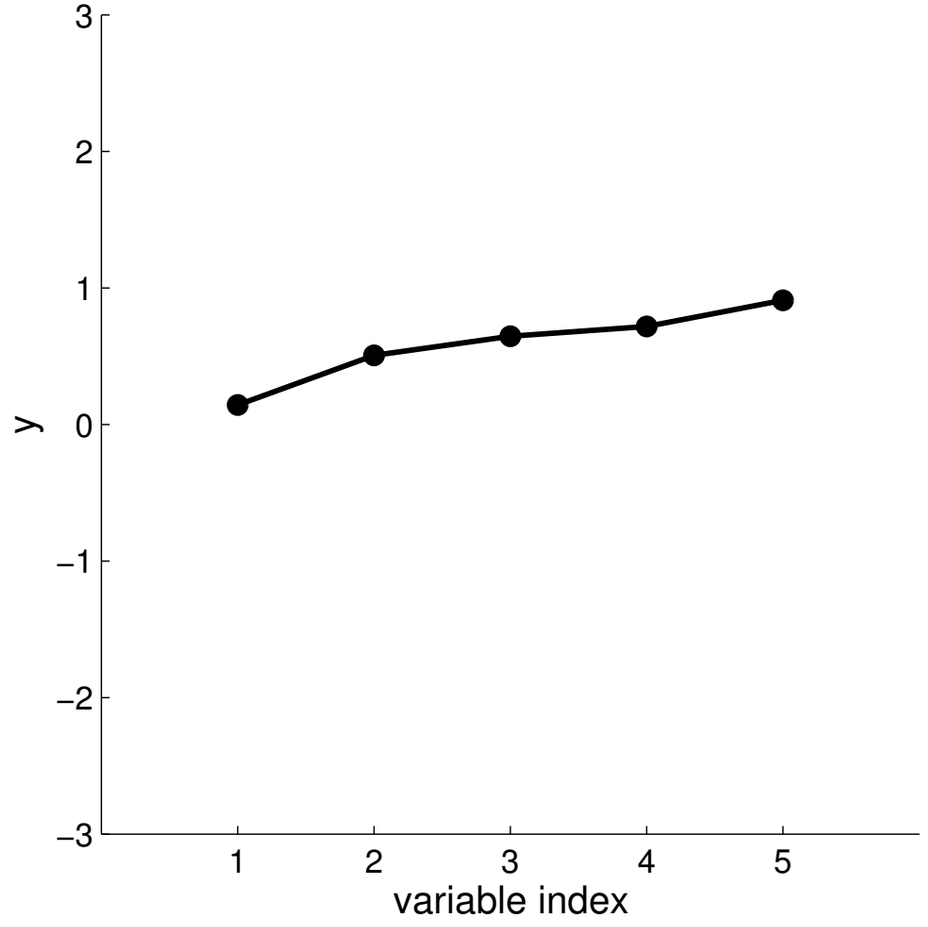


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

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

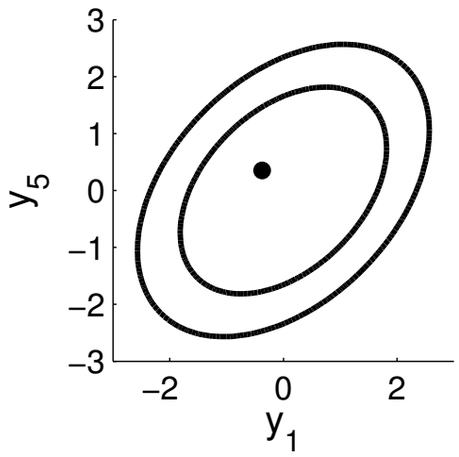


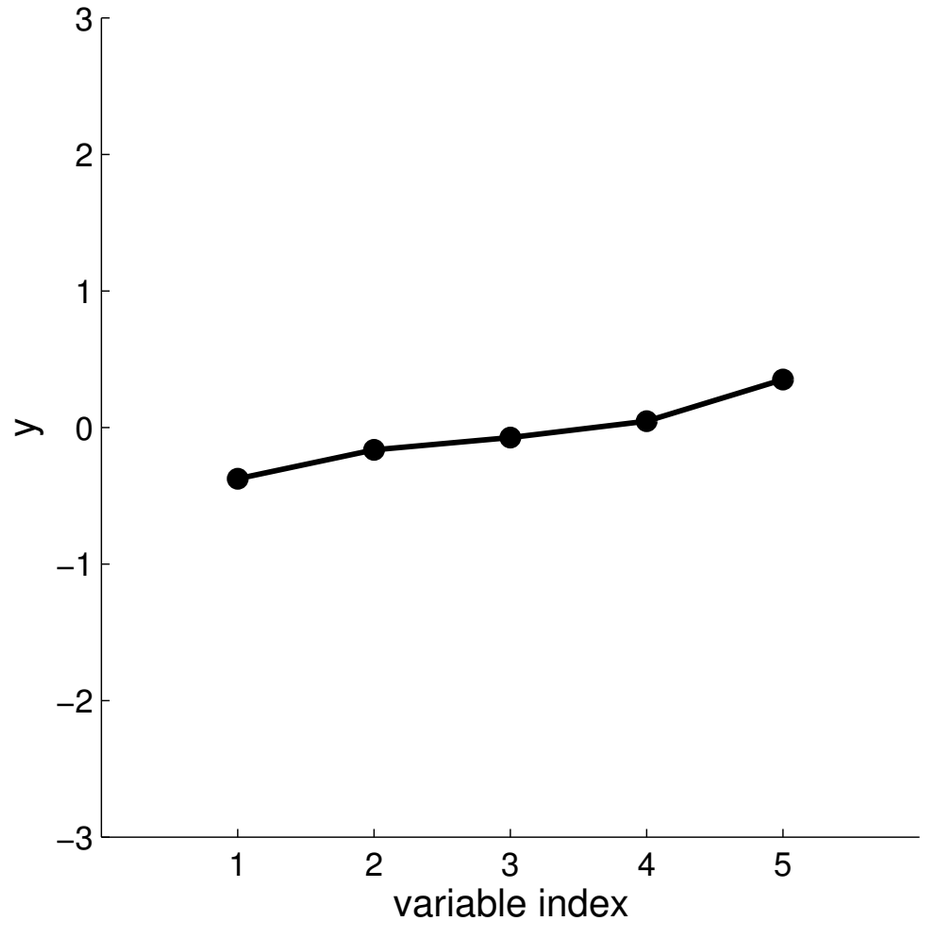


$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix

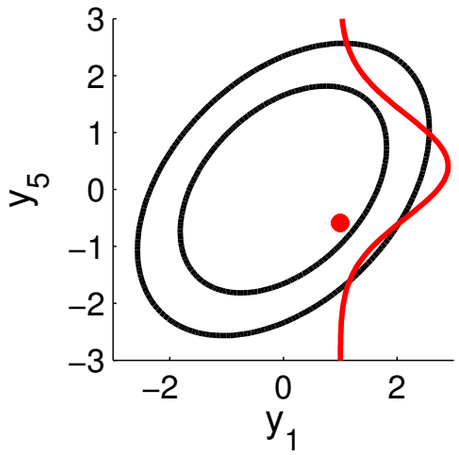
- Correlations fall off the further the indices of the variables!

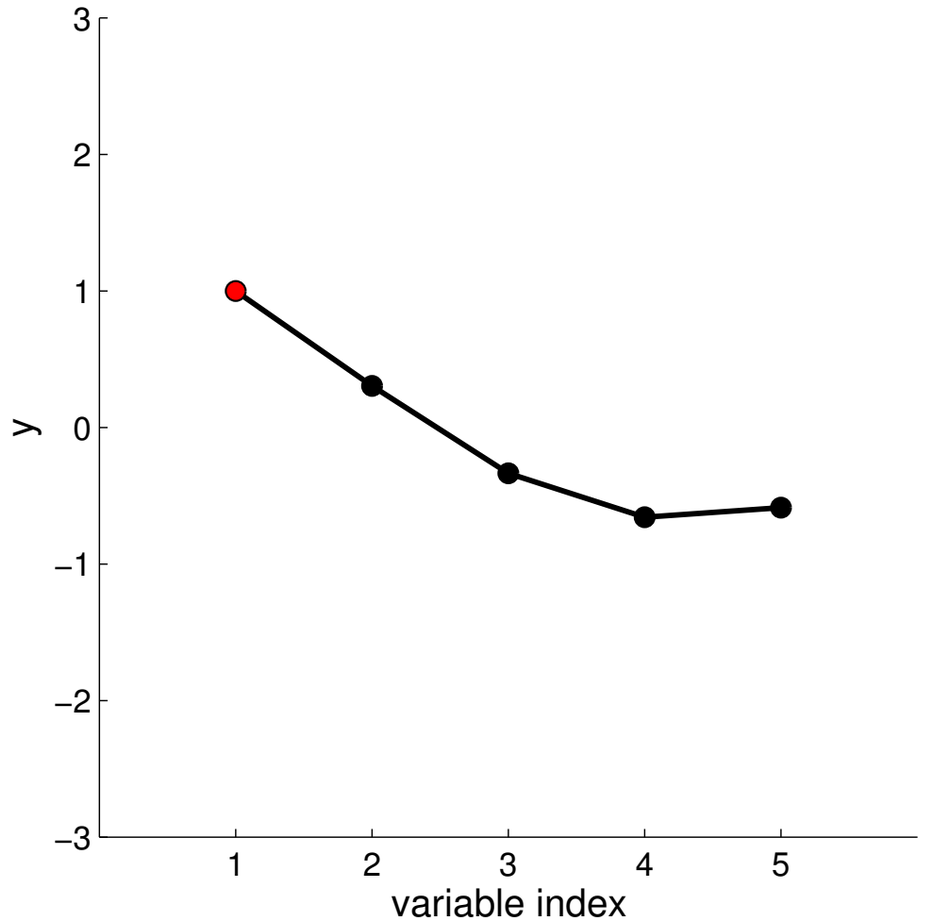




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

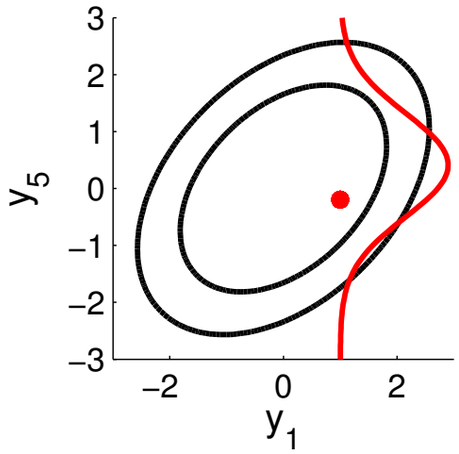
Special covariance matrix - conditioning

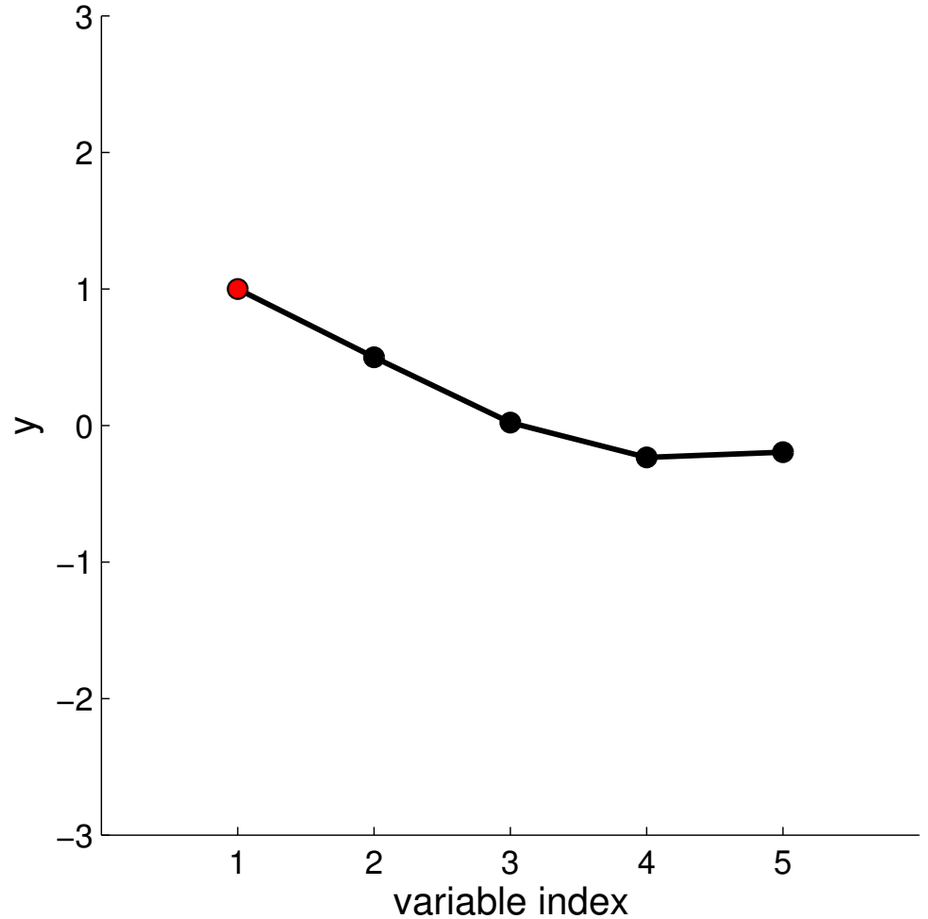




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

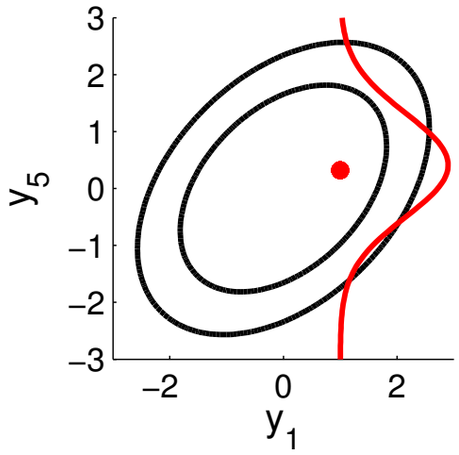
Special covariance matrix - conditioning

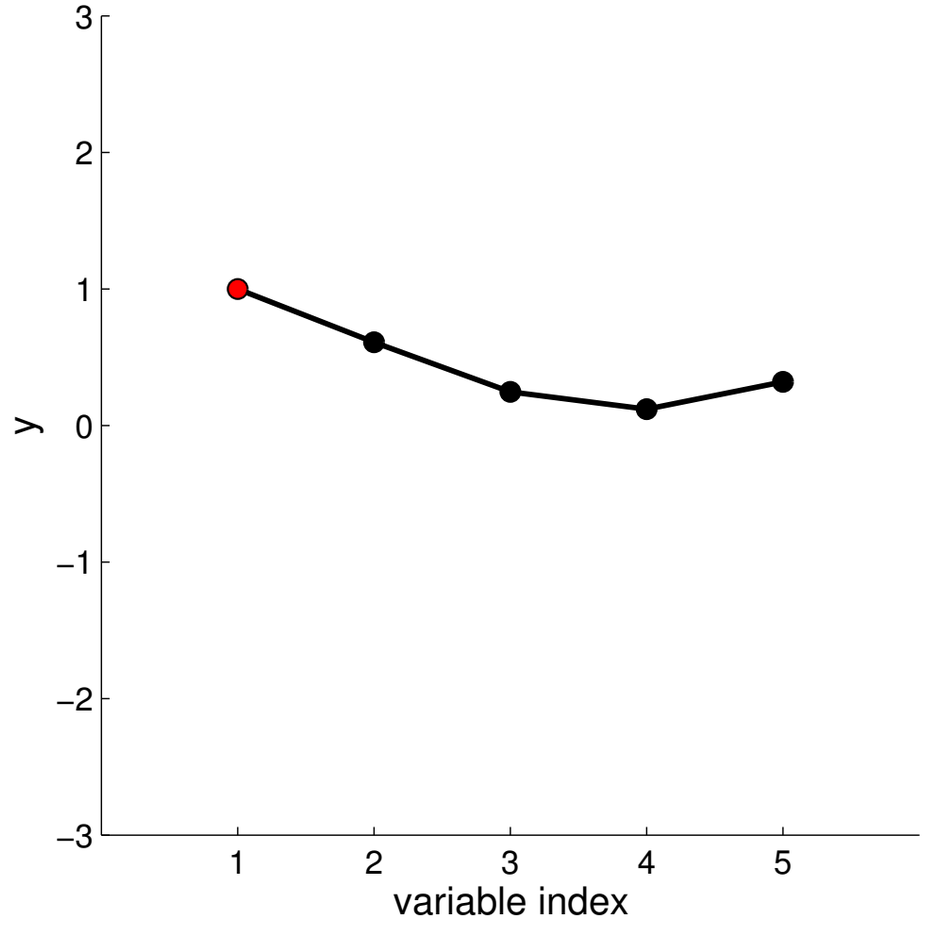




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

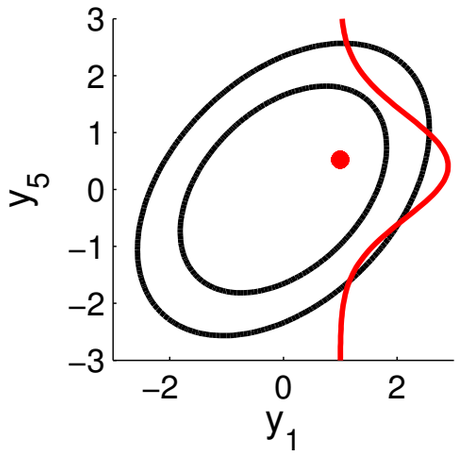
Special covariance matrix - conditioning

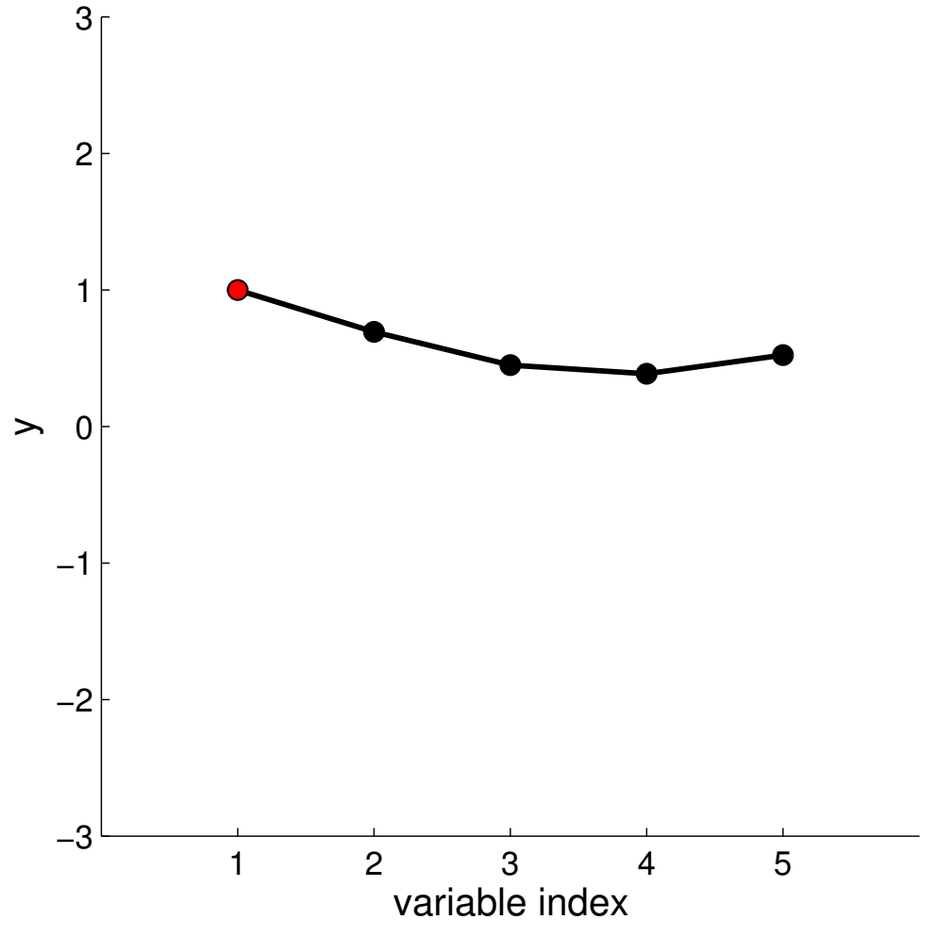




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

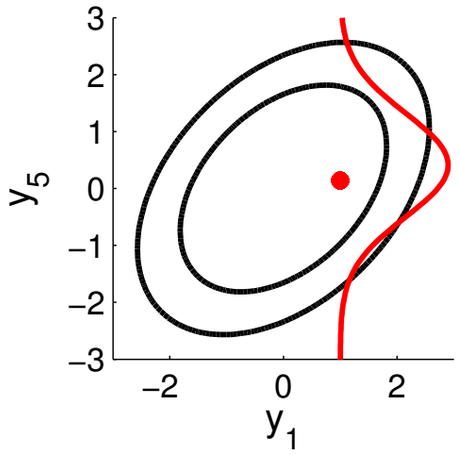
Special covariance matrix - conditioning

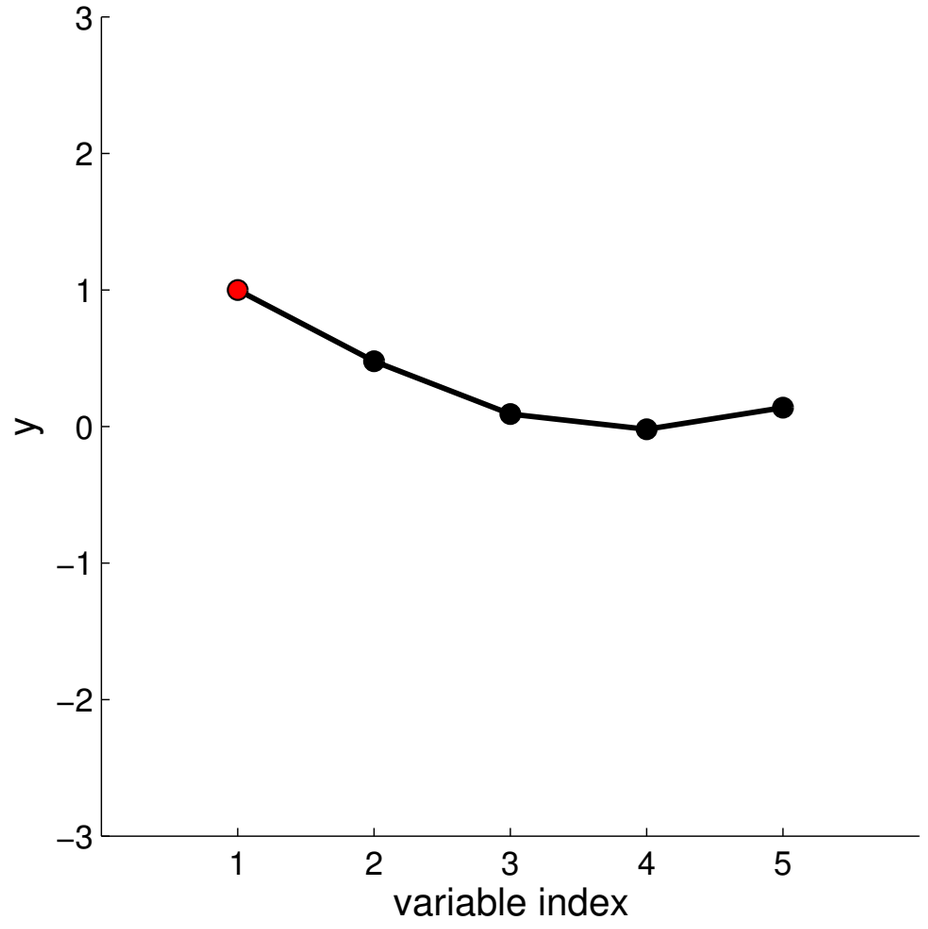




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

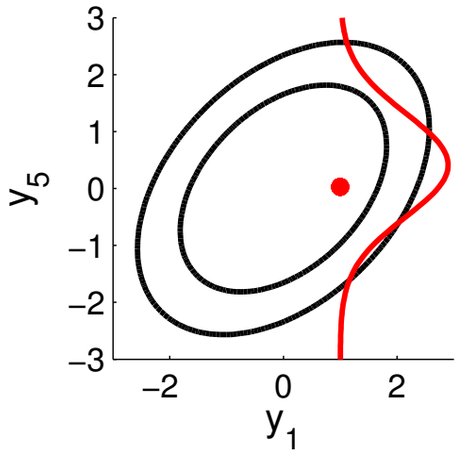
Special covariance matrix - conditioning

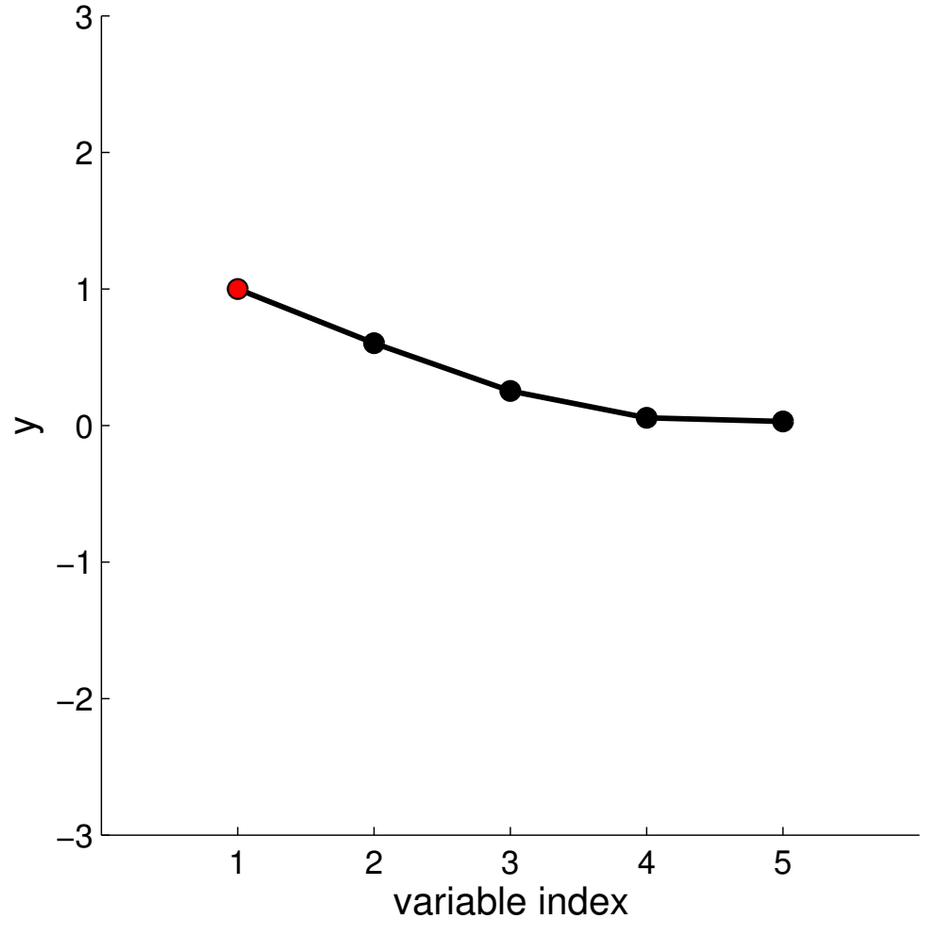




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

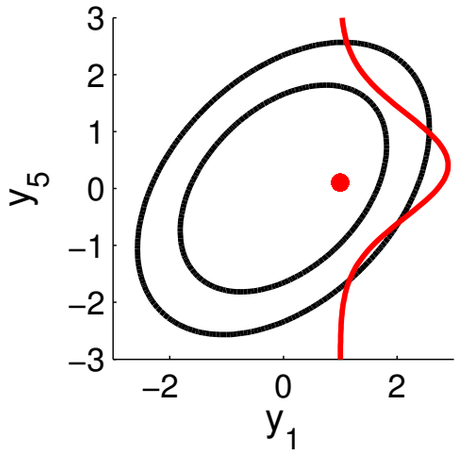
Special covariance matrix - conditioning

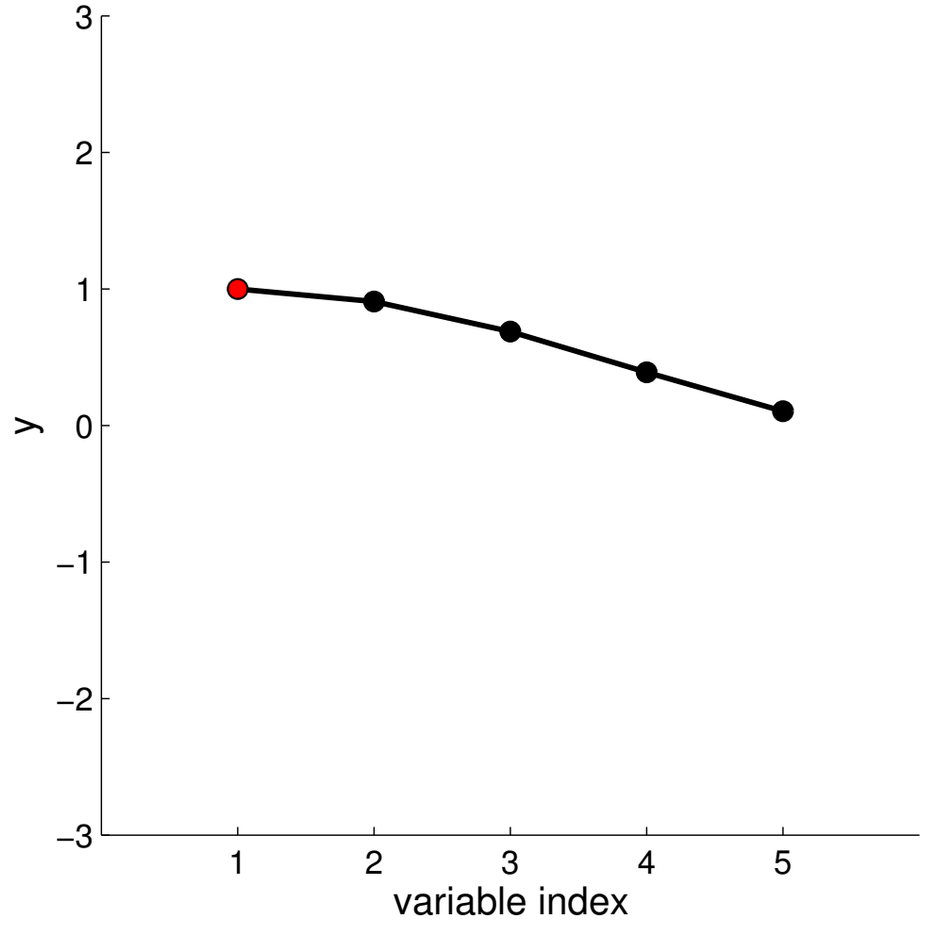




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

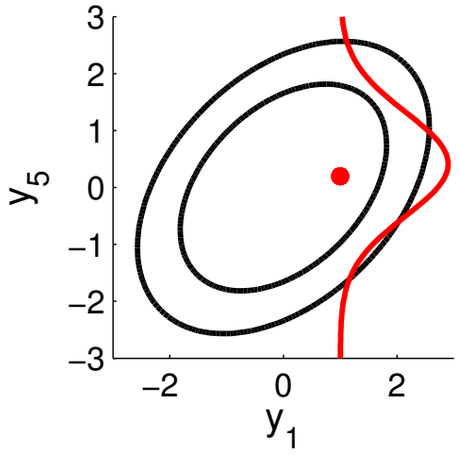
Special covariance matrix - conditioning

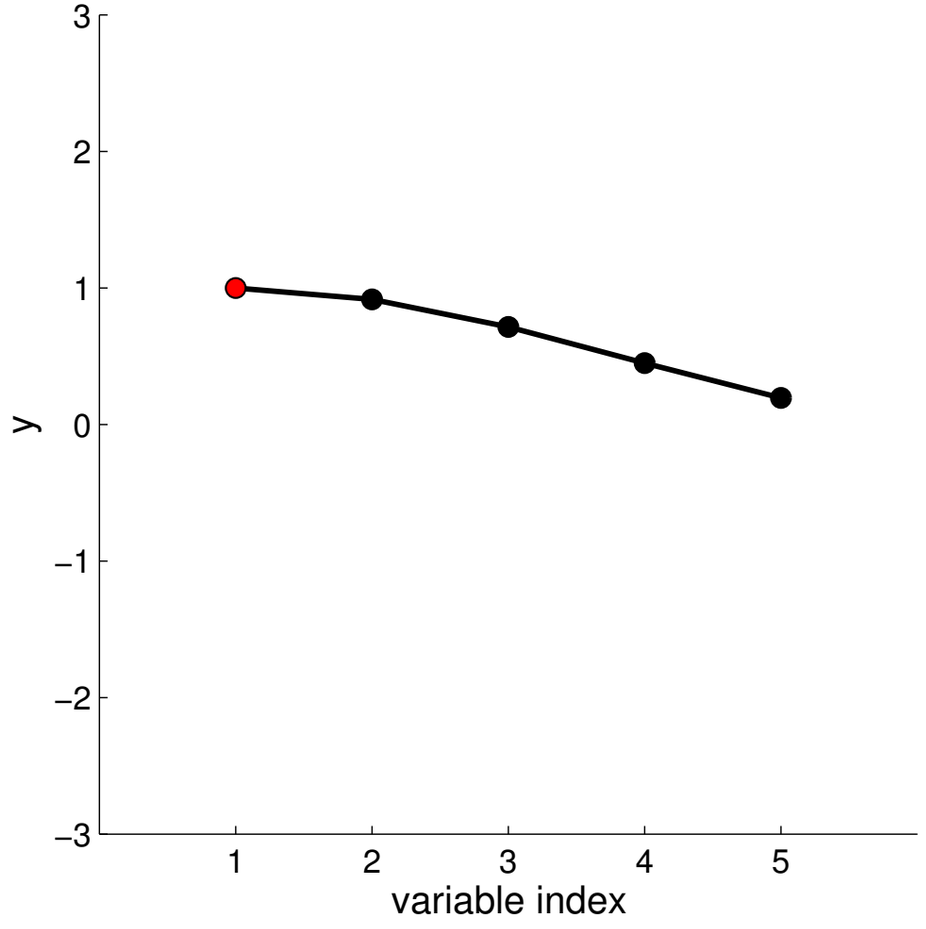




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

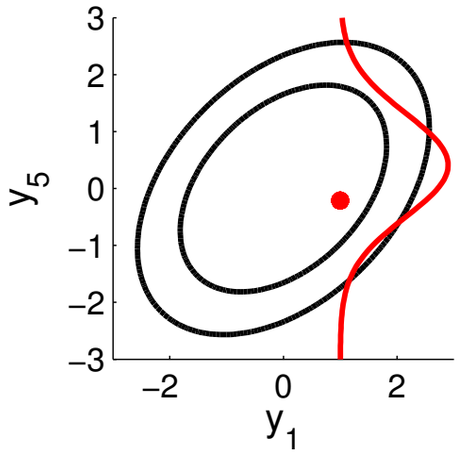
Special covariance matrix - conditioning

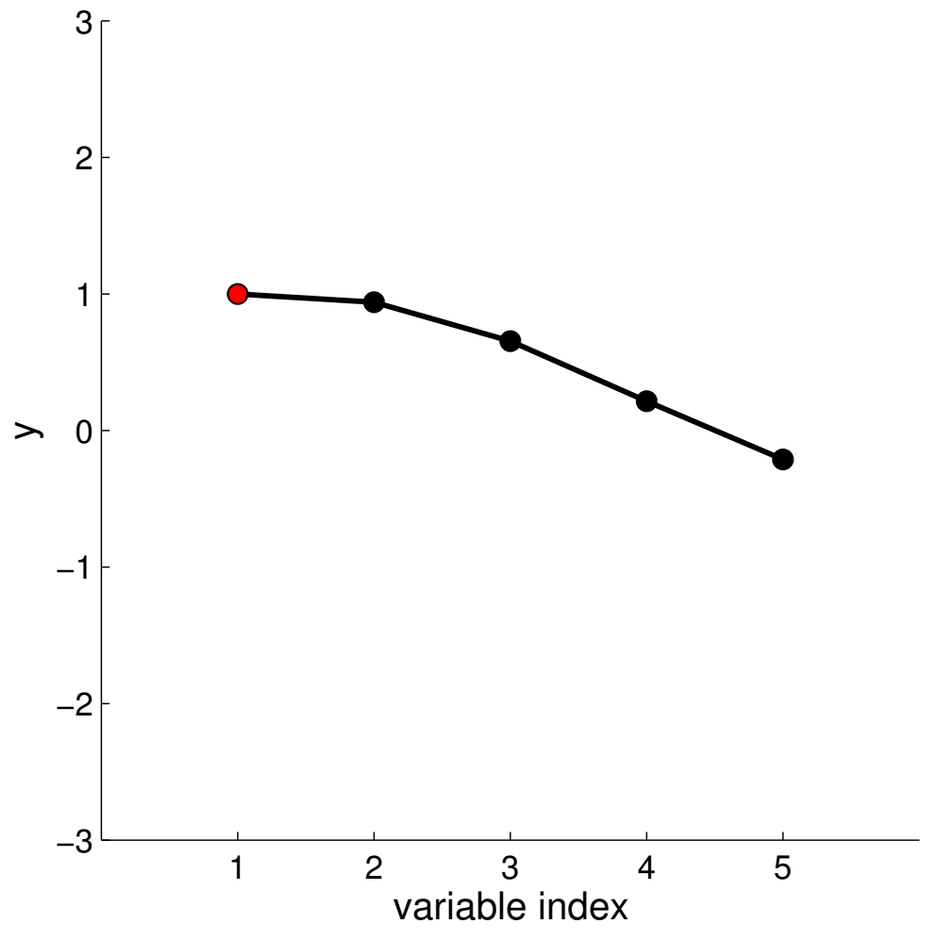




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

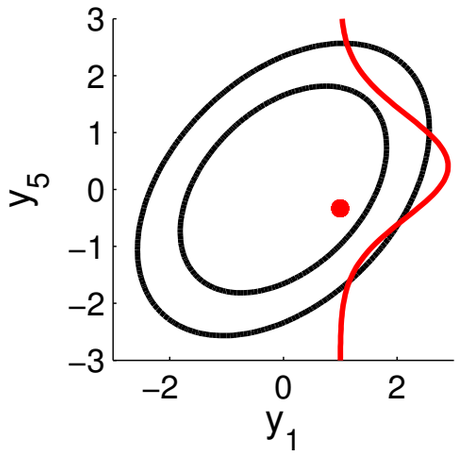
Special covariance matrix - conditioning

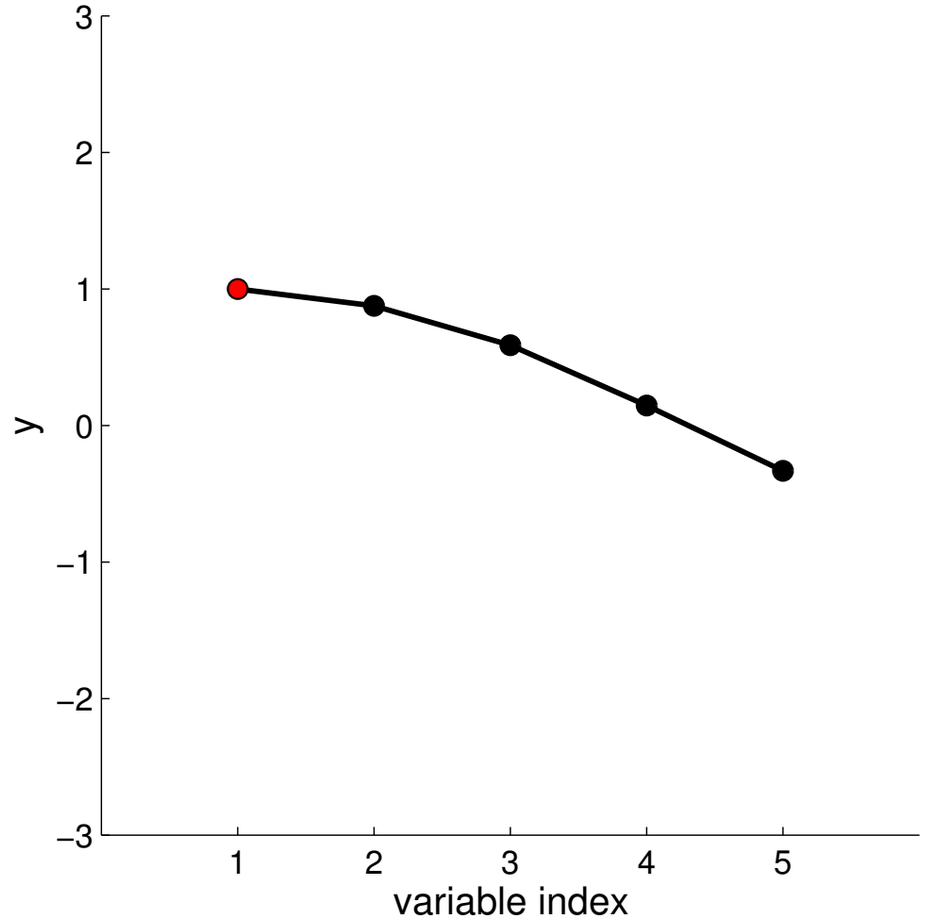




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

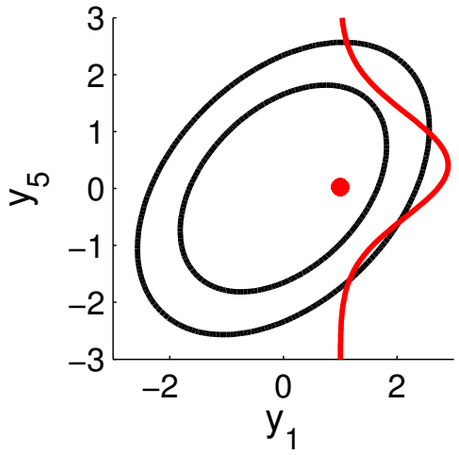
Special covariance matrix - conditioning

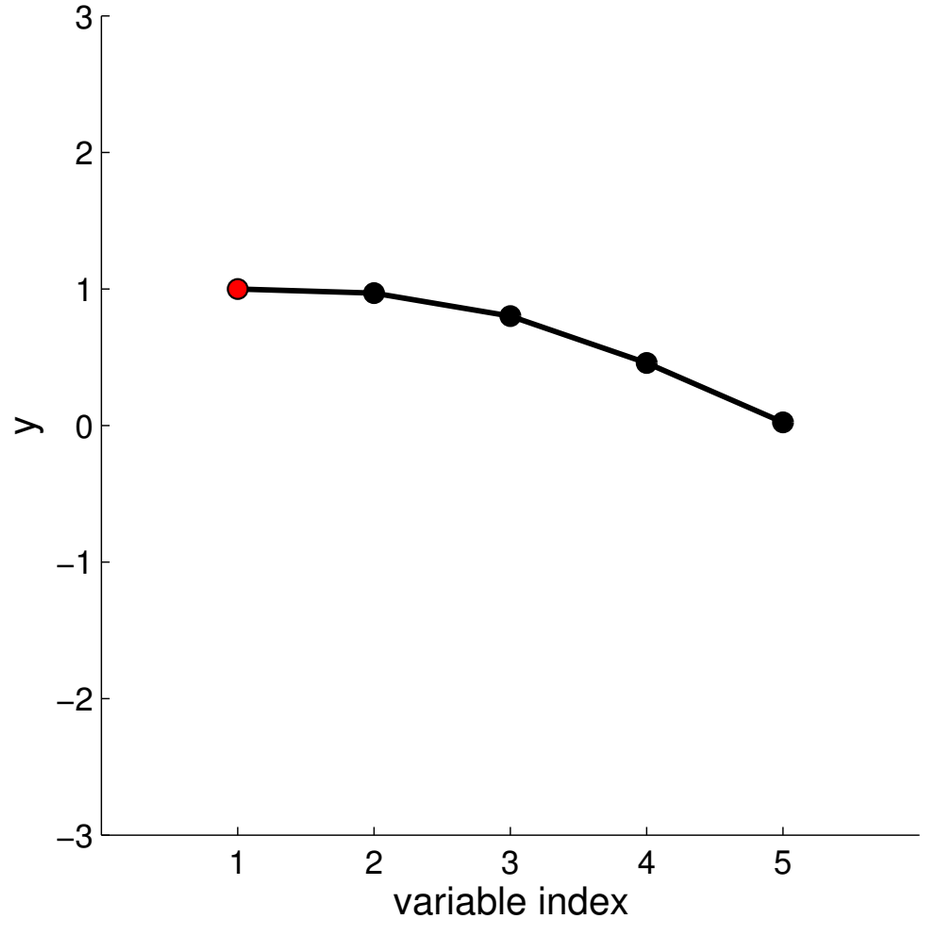




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

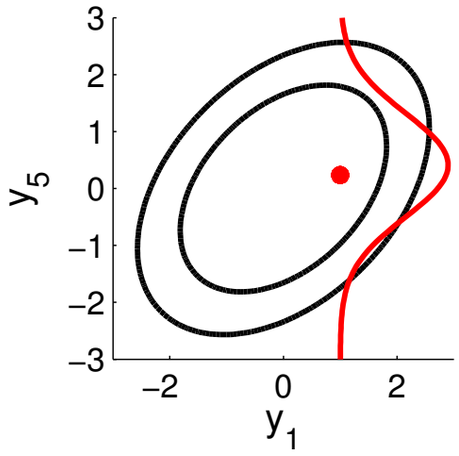
Special covariance matrix - conditioning

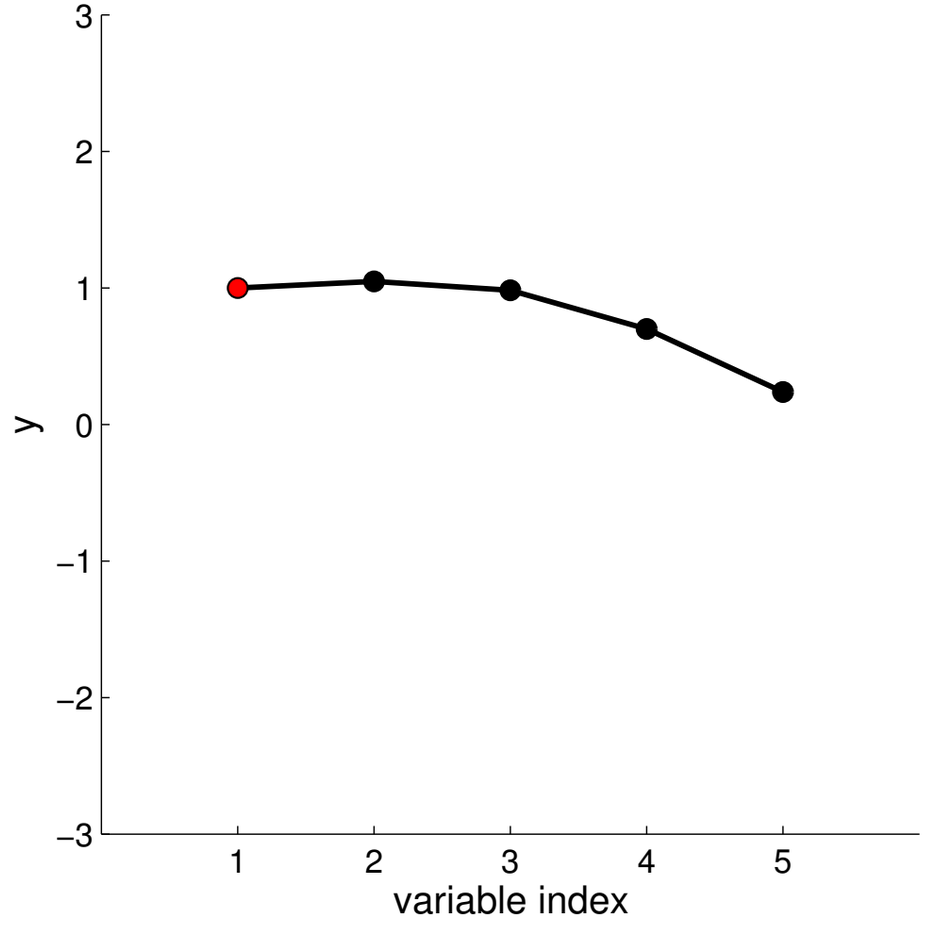




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

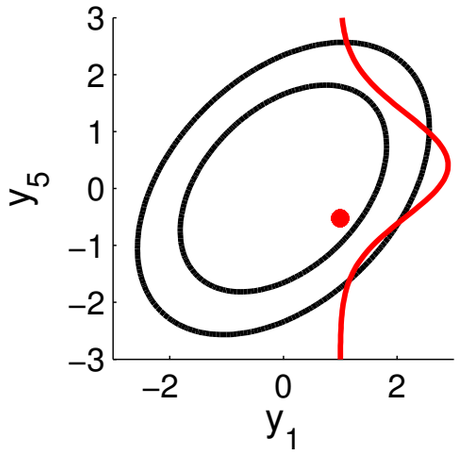
Special covariance matrix - conditioning

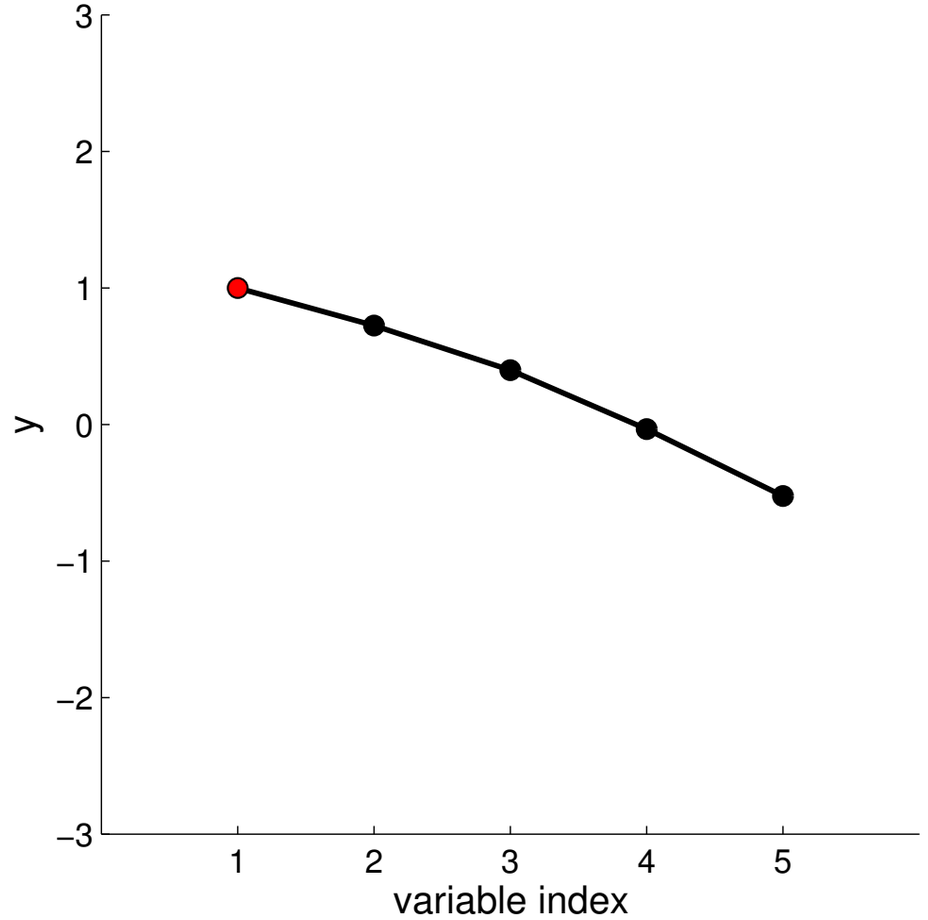




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

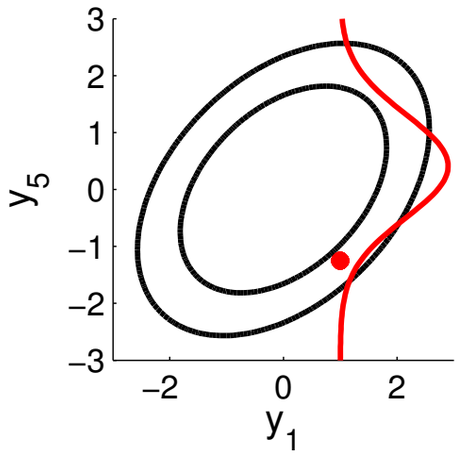
Special covariance matrix - conditioning

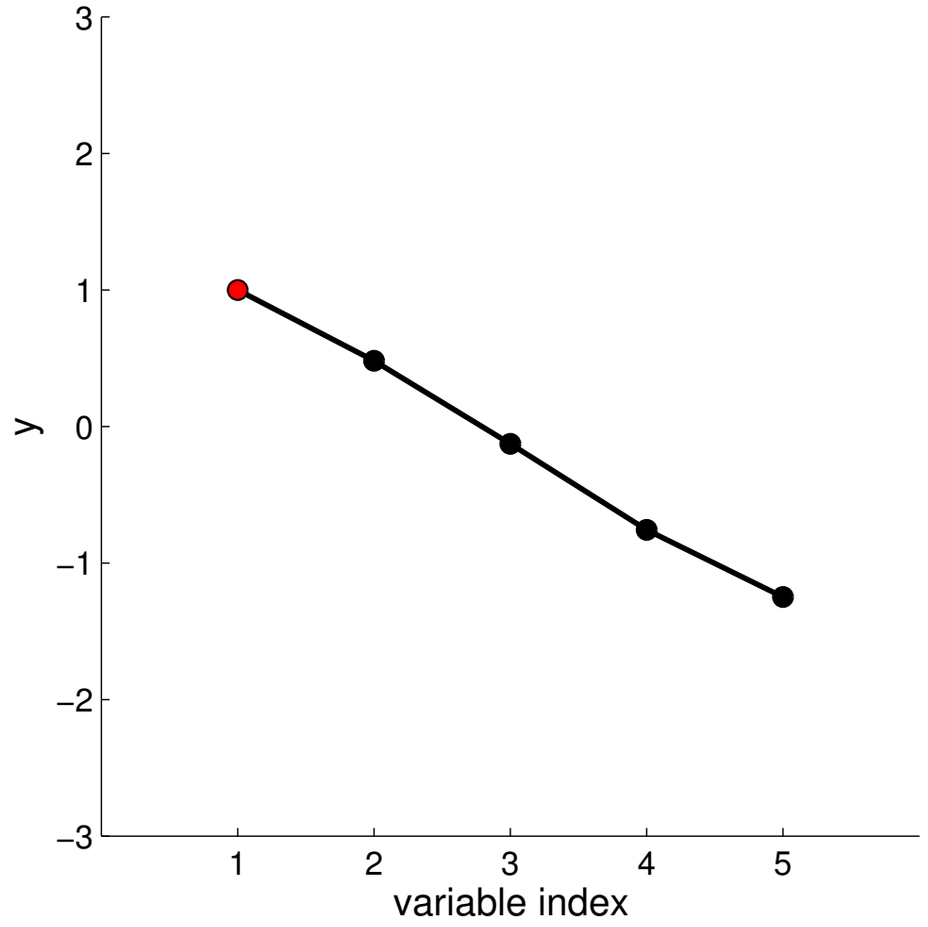




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

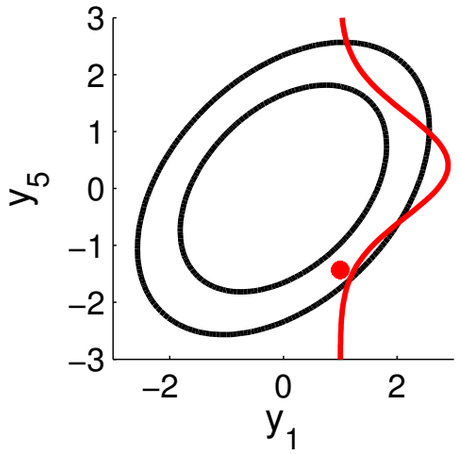
Special covariance matrix - conditioning

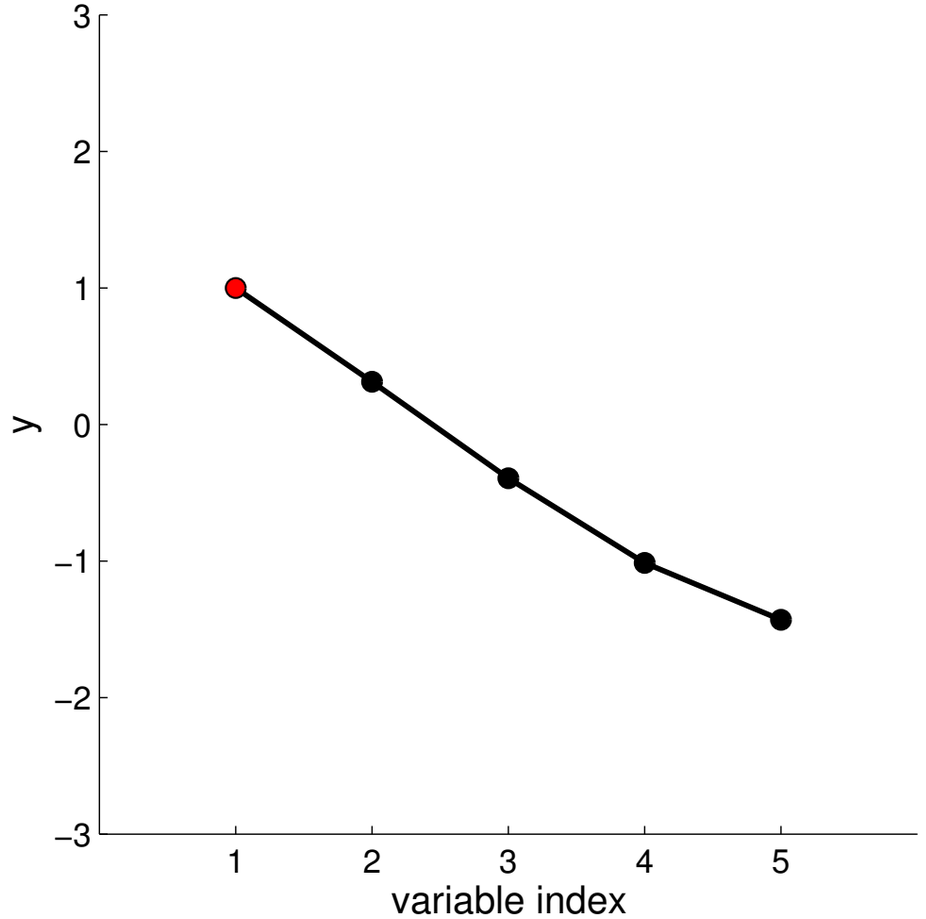




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

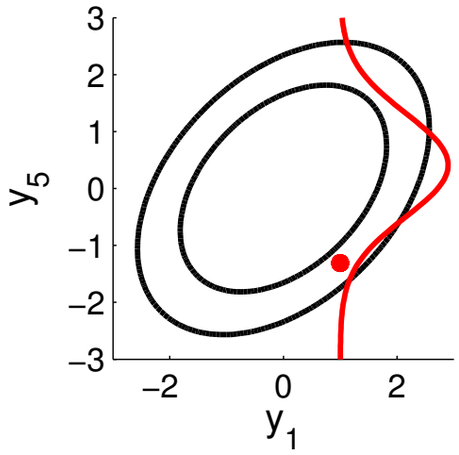
Special covariance matrix - conditioning

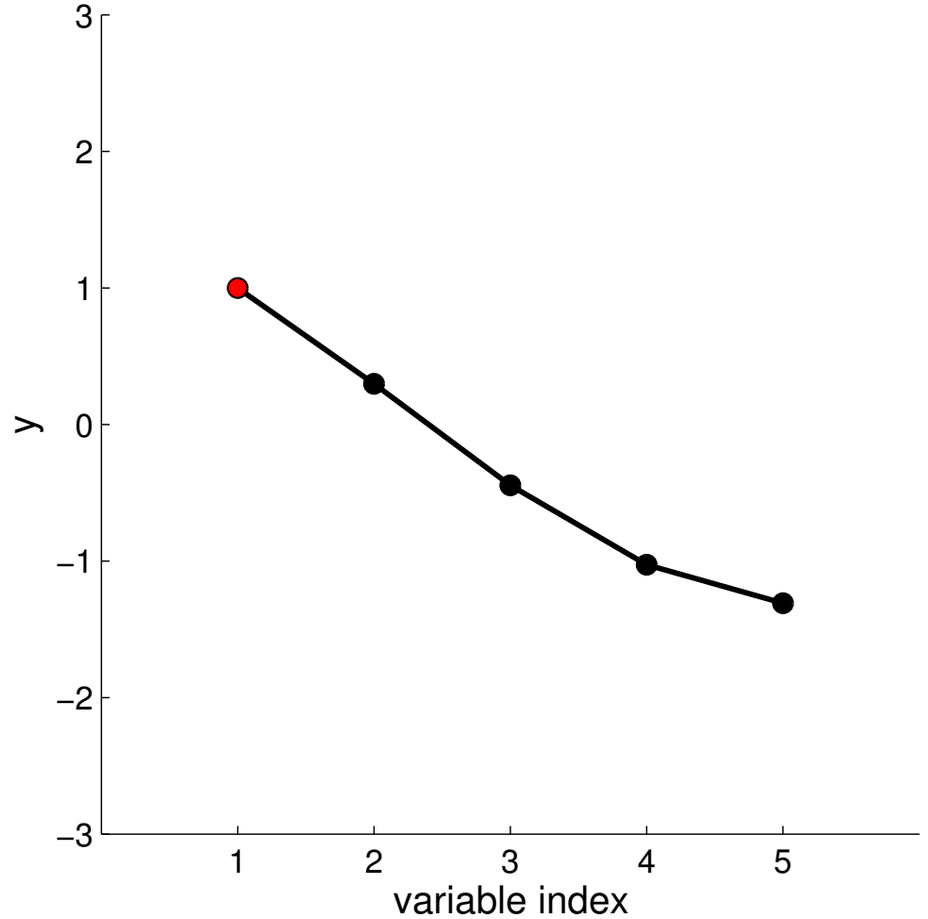




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

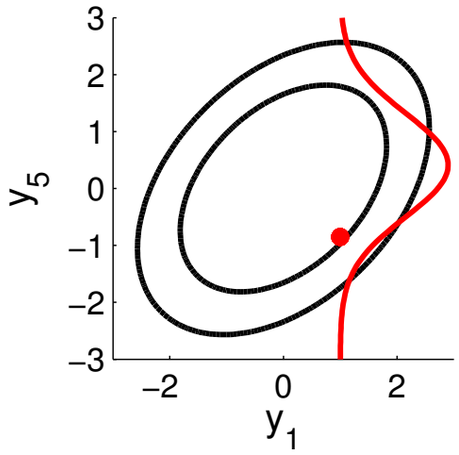
Special covariance matrix - conditioning

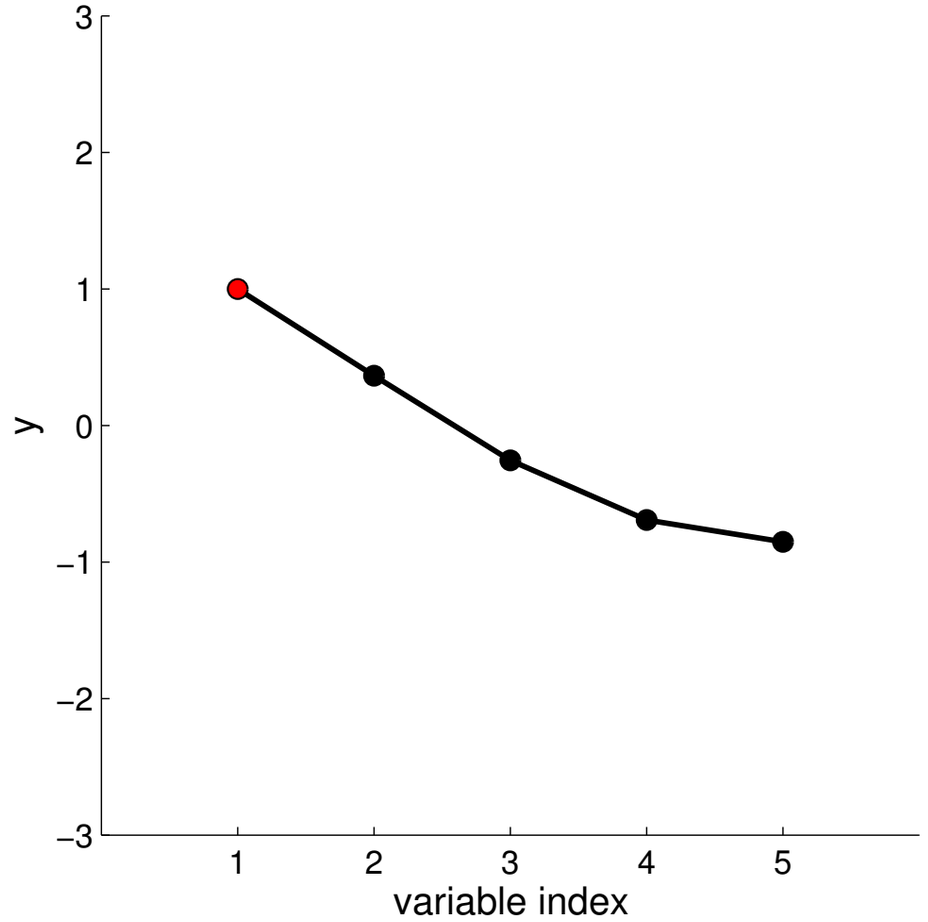




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

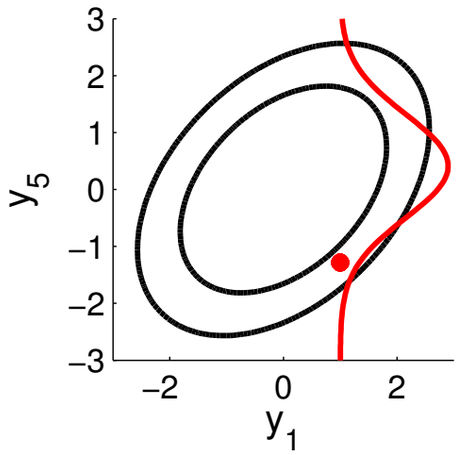
Special covariance matrix - conditioning

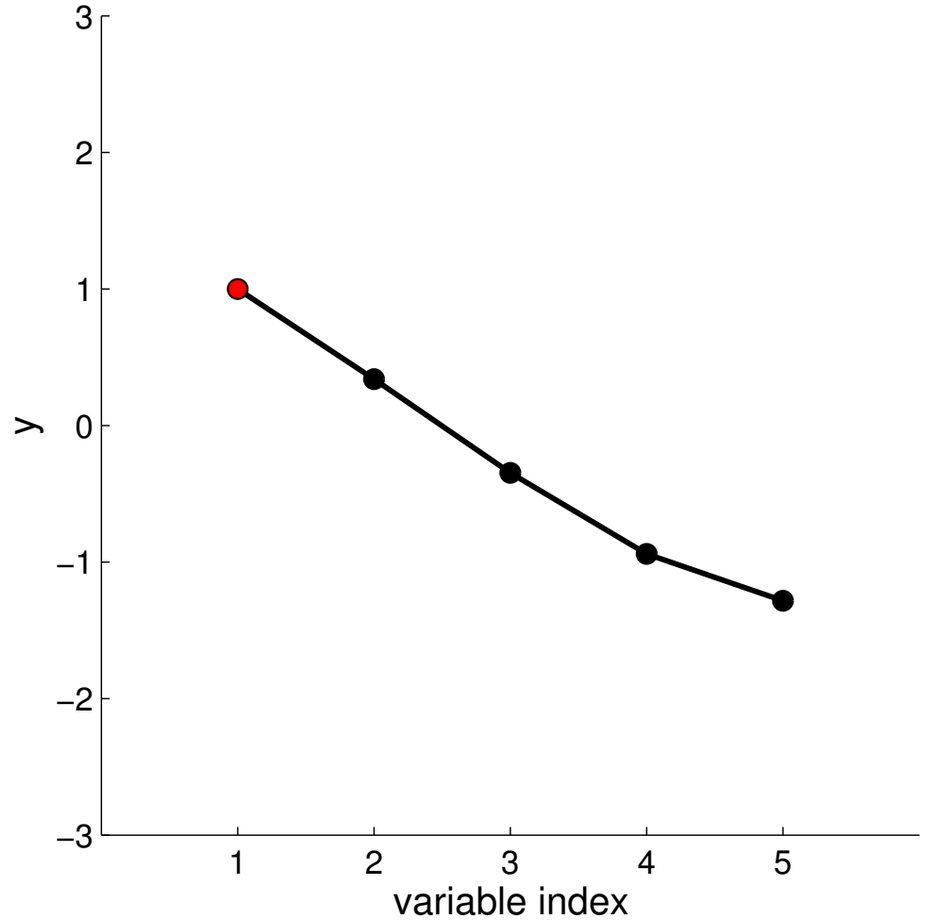




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

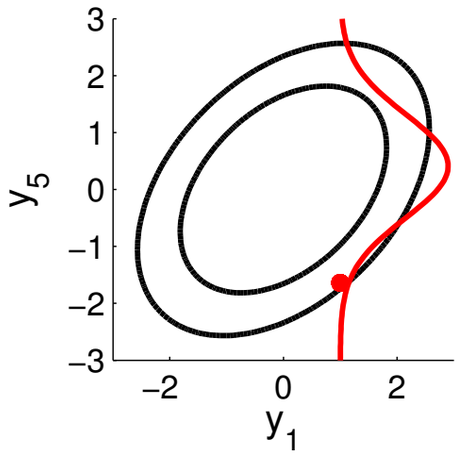
Special covariance matrix - conditioning

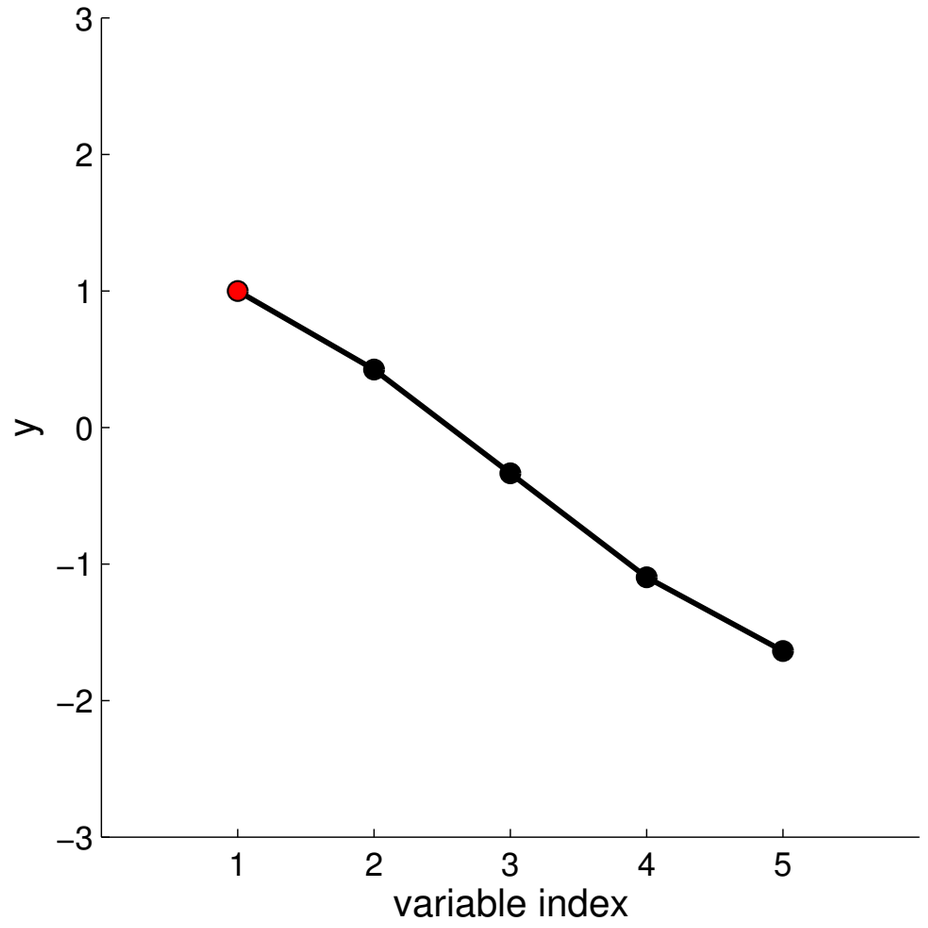




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

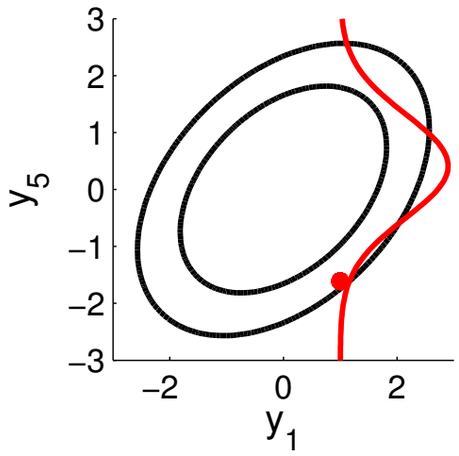
Special covariance matrix - conditioning

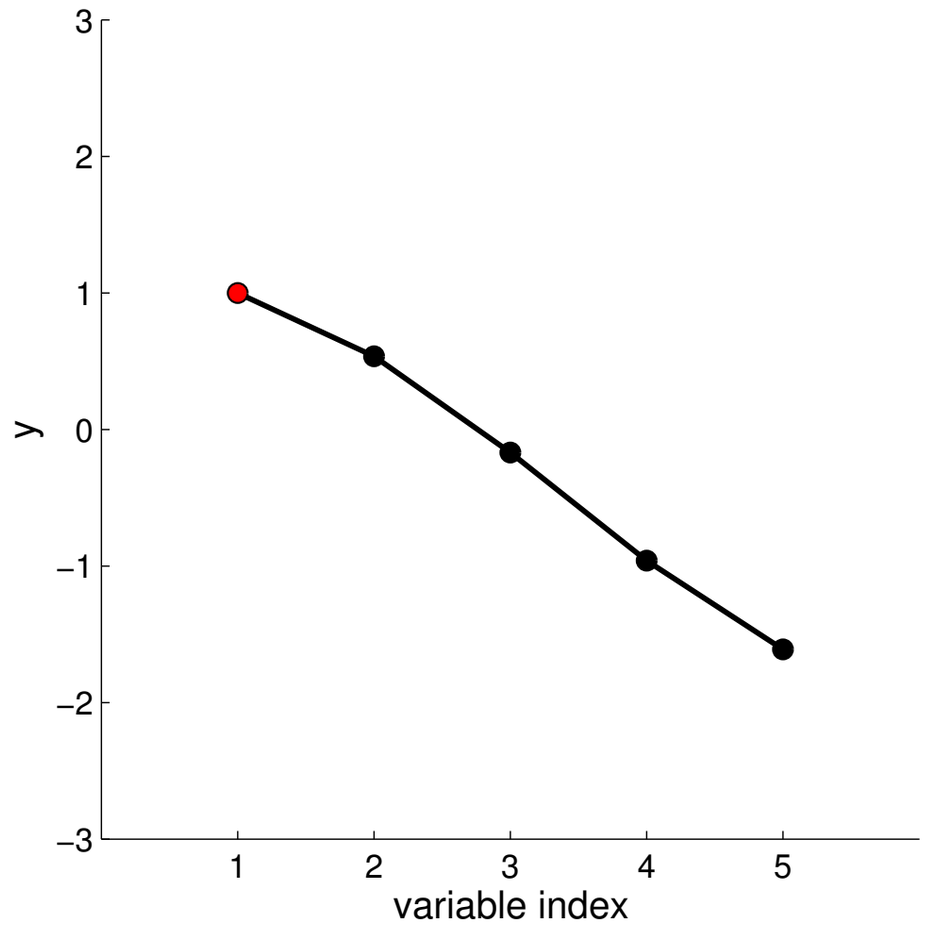




$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

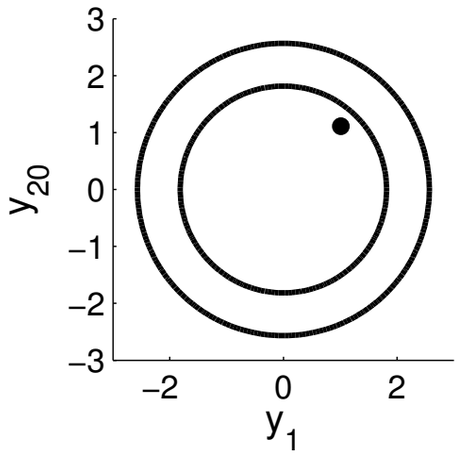
Special covariance matrix - conditioning

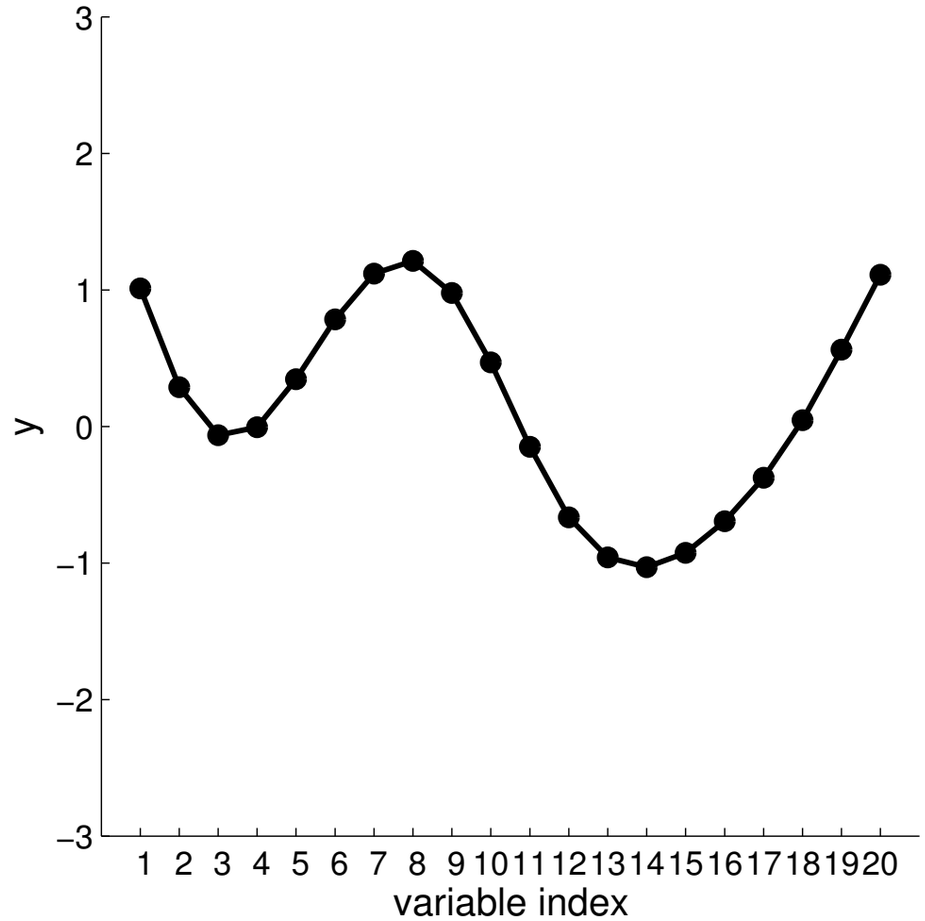




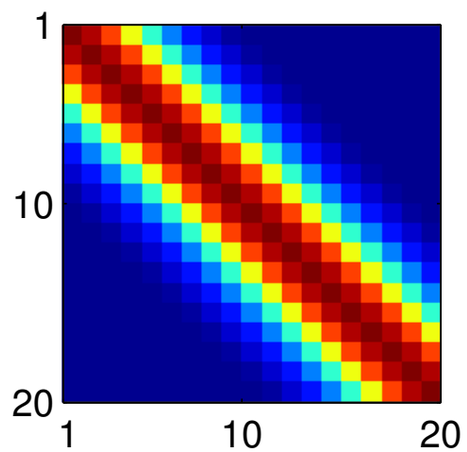
$$\Sigma = \begin{bmatrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{bmatrix}$$

Special covariance matrix



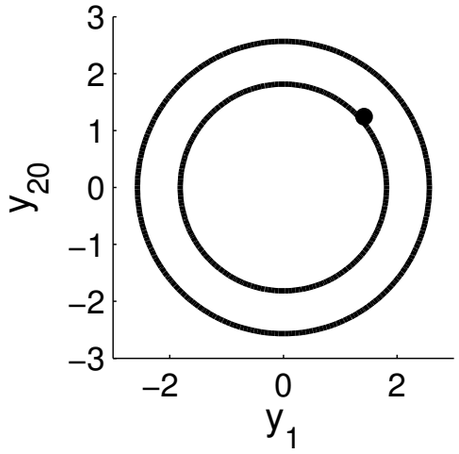


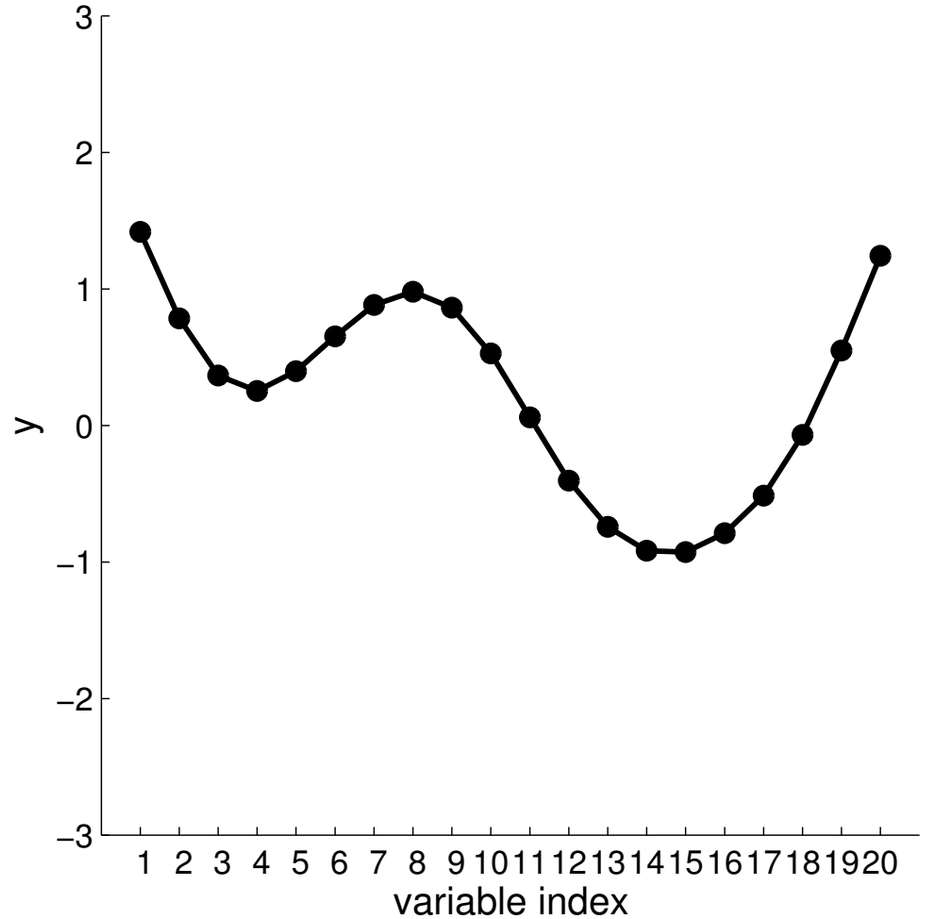
$\Sigma =$



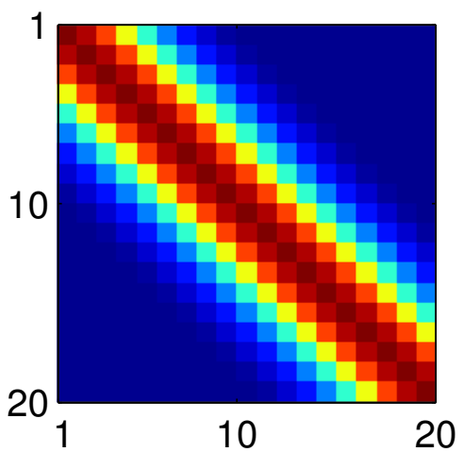
red means 1, blue means 0

Special covariance matrix

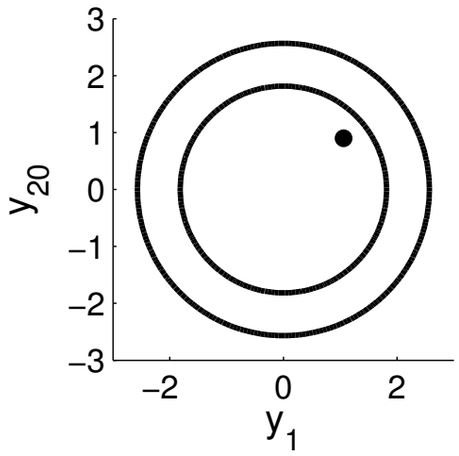


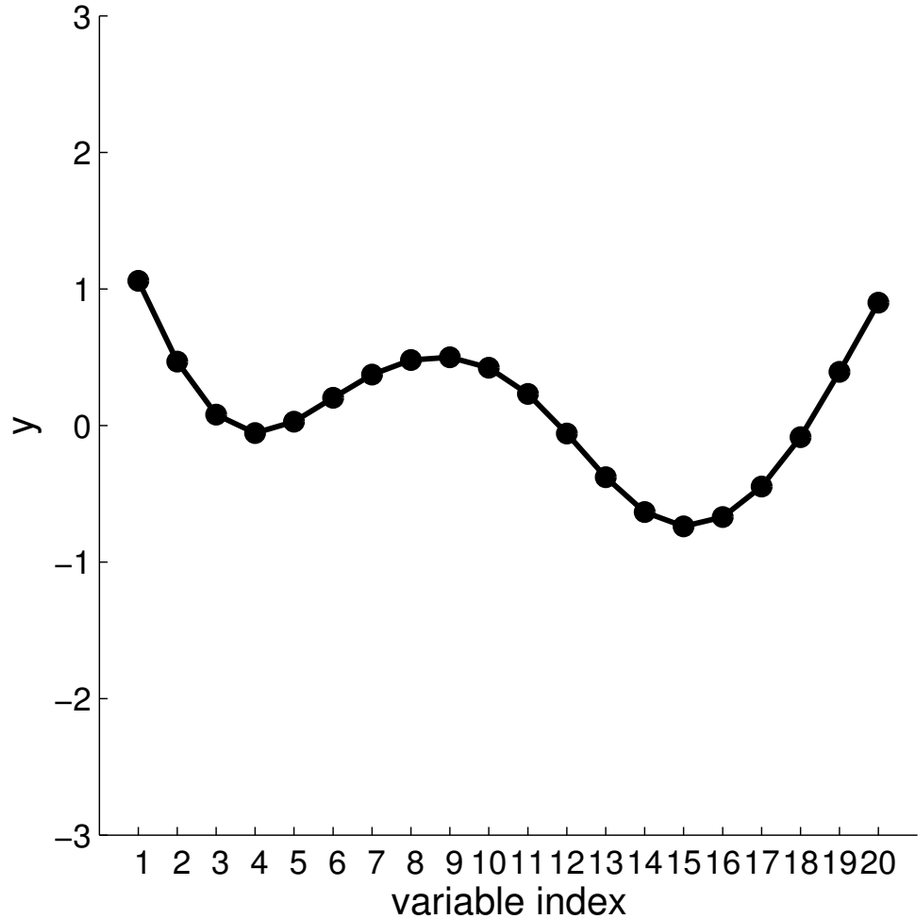


$\Sigma =$

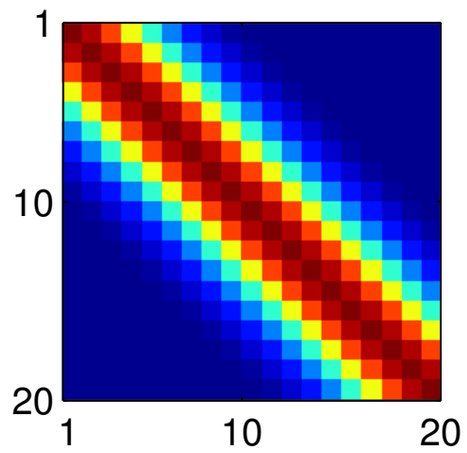


Special covariance matrix

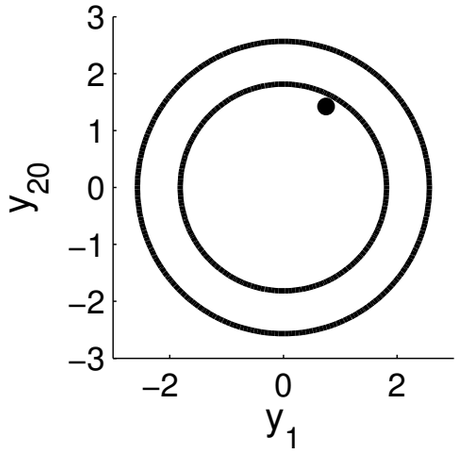


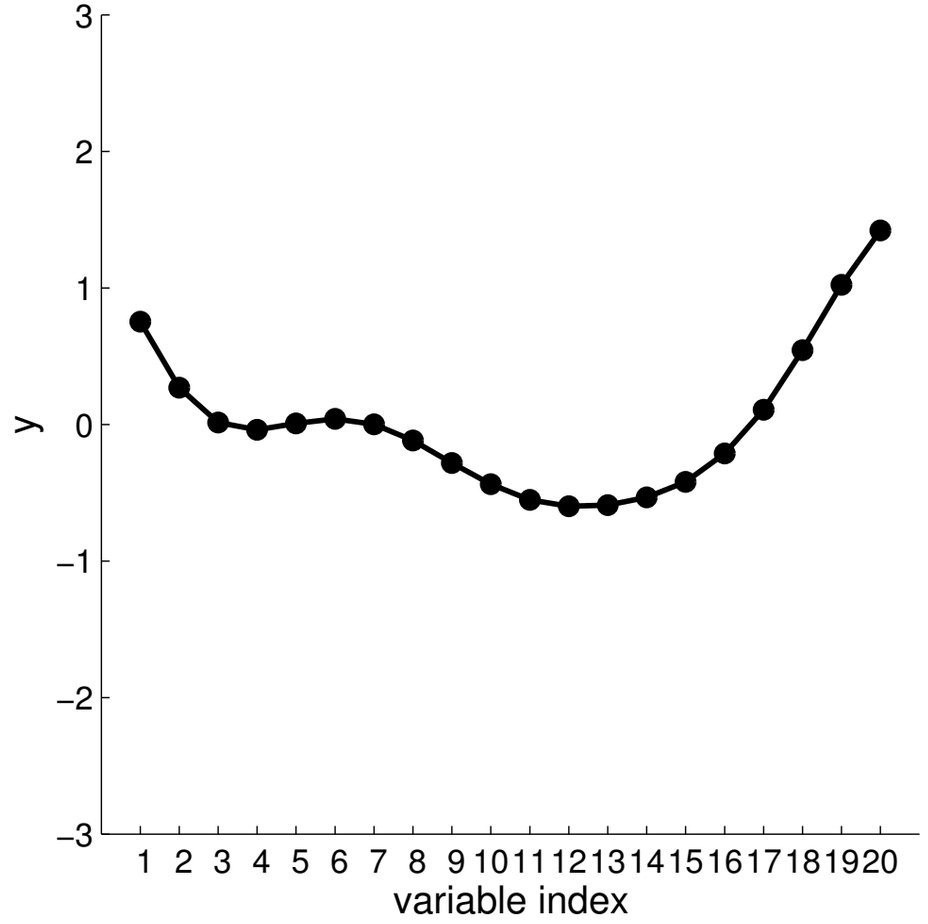


$\Sigma =$

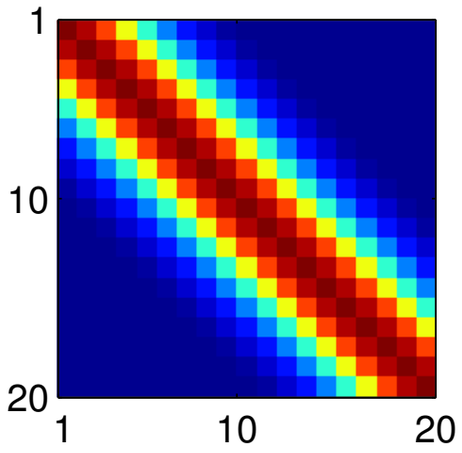


Special covariance matrix

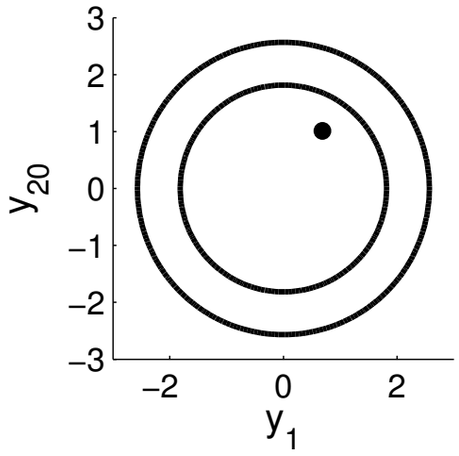


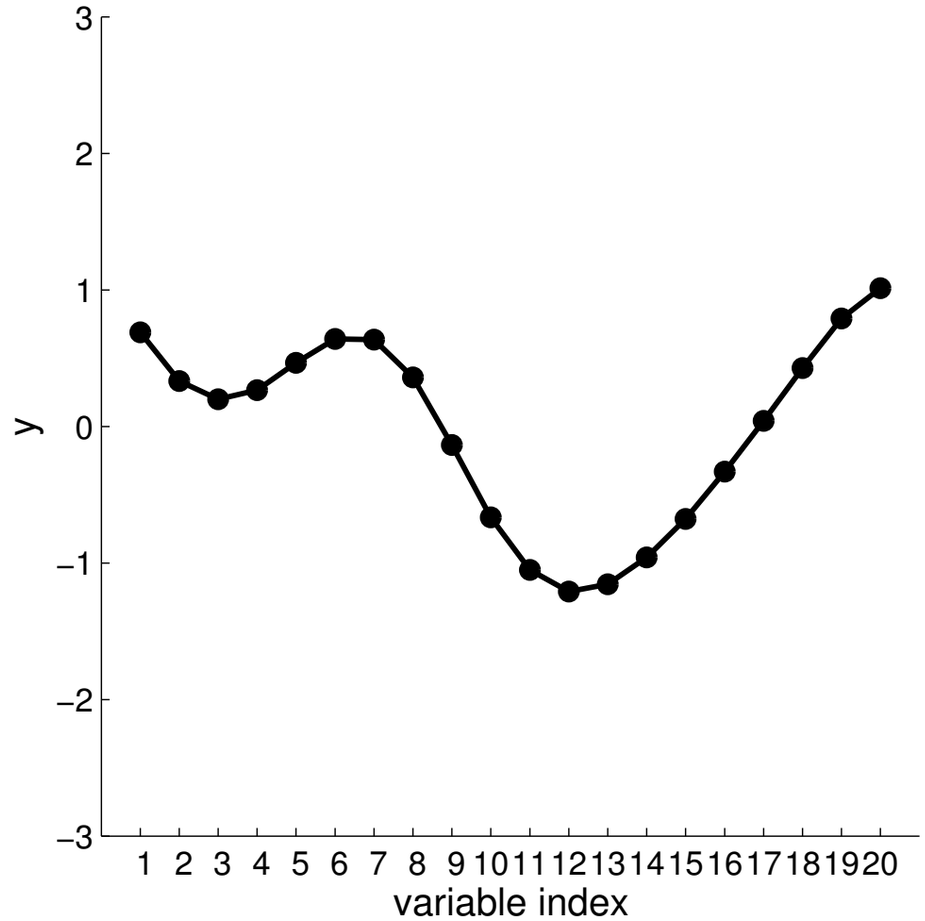


$\Sigma =$

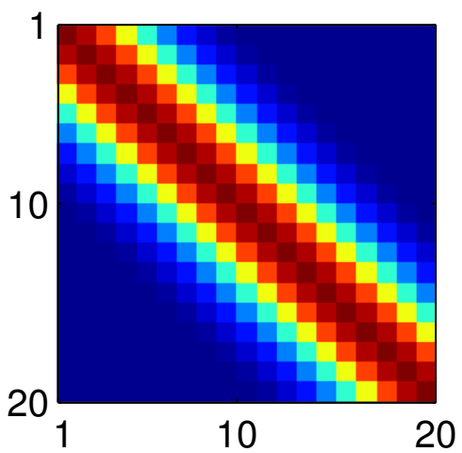


Special covariance matrix

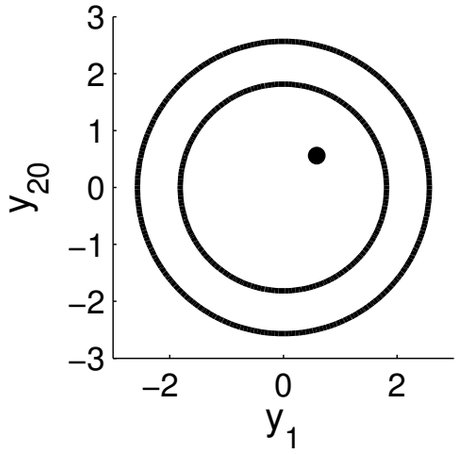


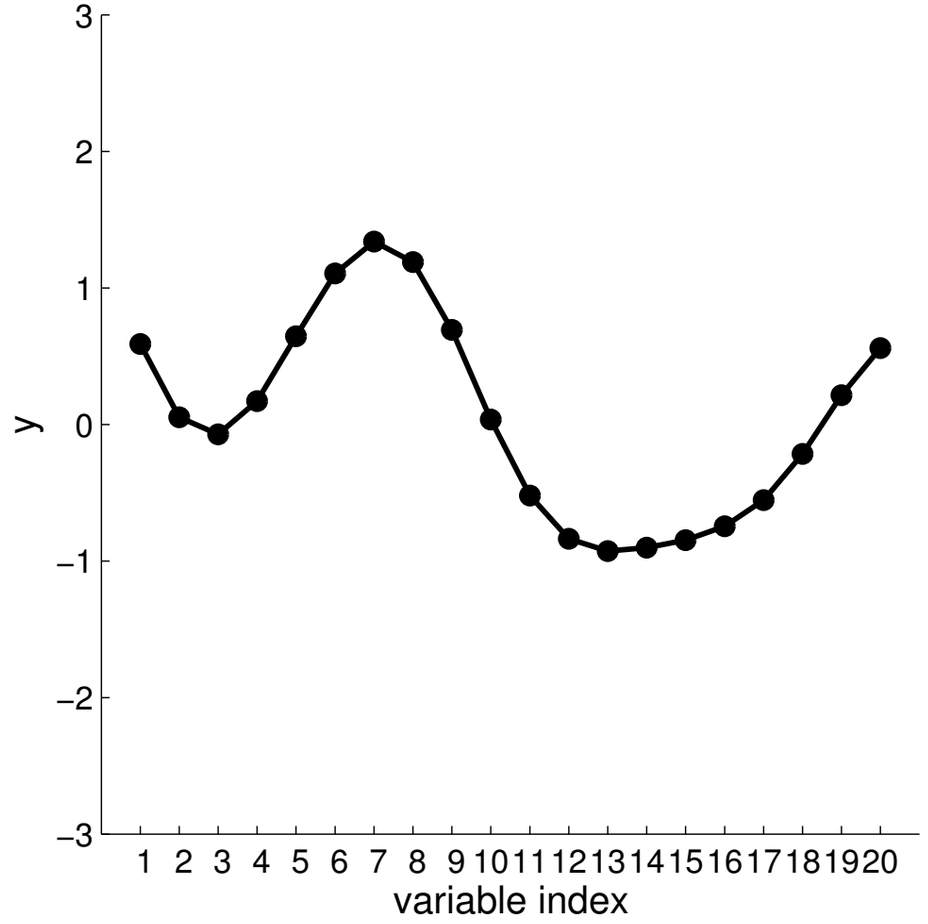


$\Sigma =$

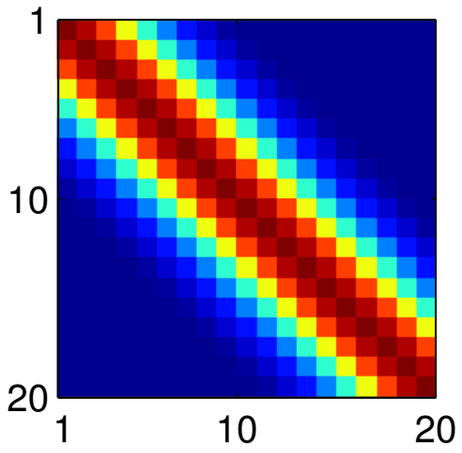


Special covariance matrix

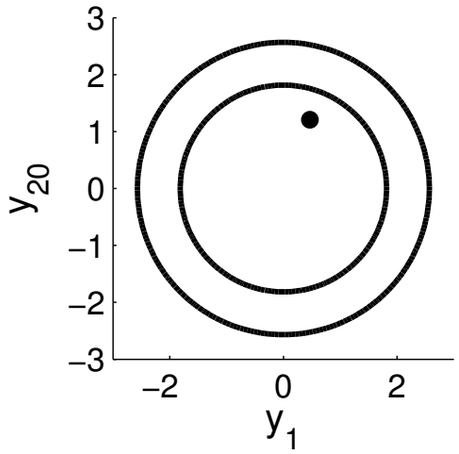


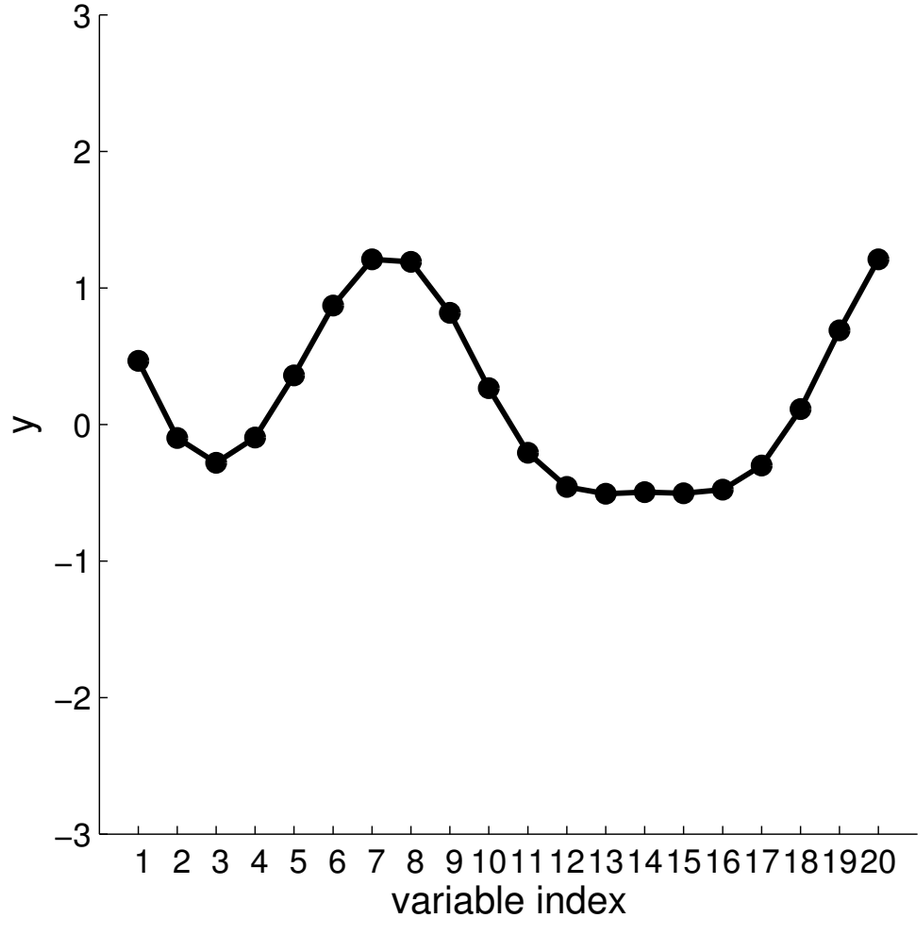


$$\Sigma =$$

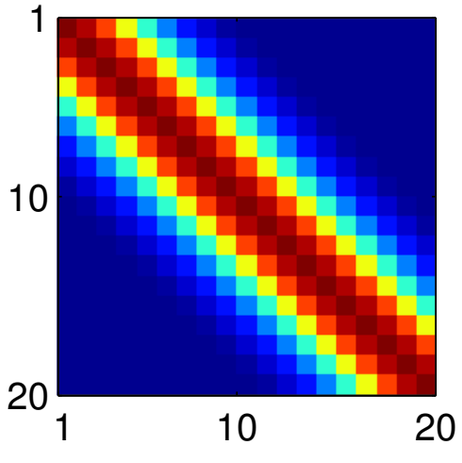


Special covariance matrix

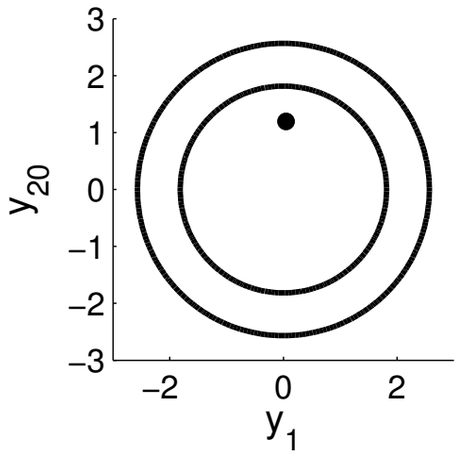


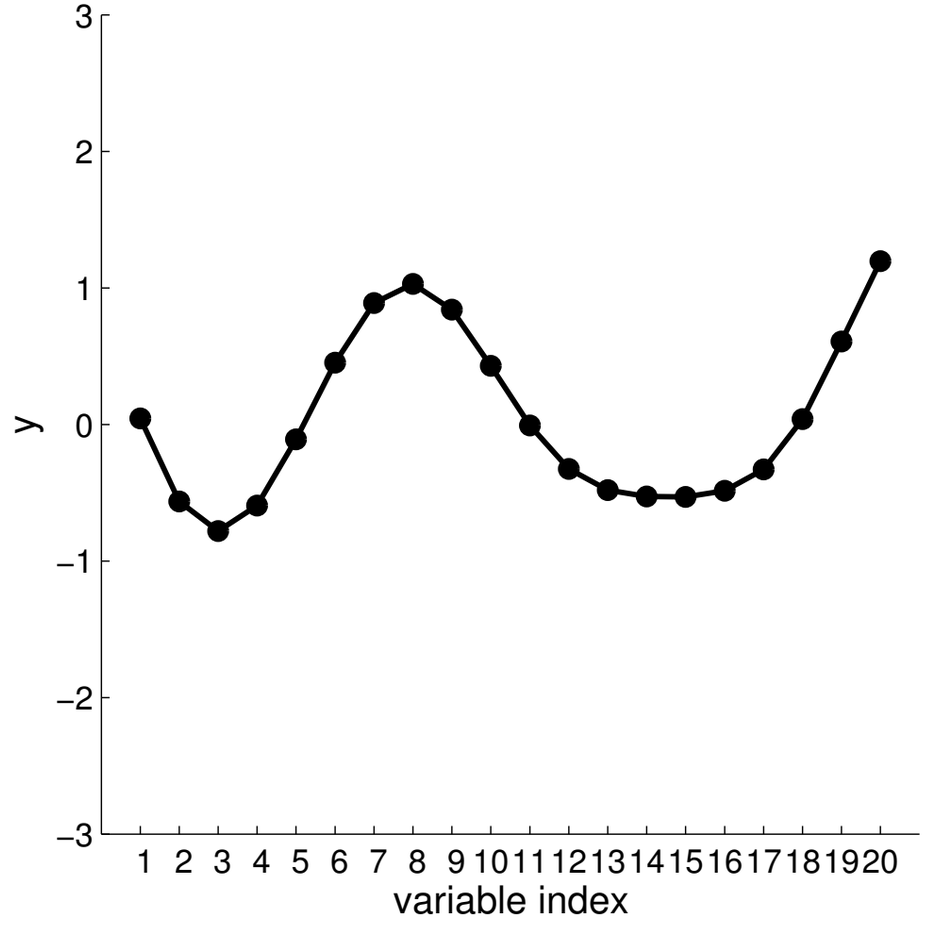


$\Sigma =$

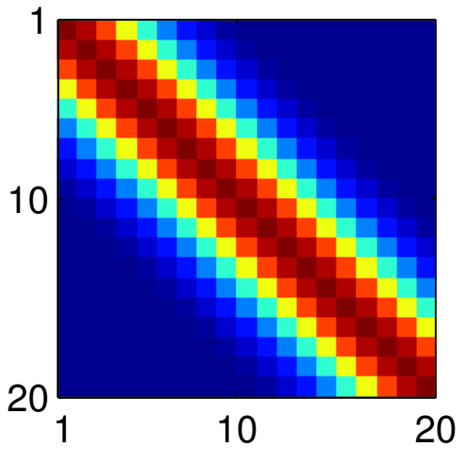


Special covariance matrix

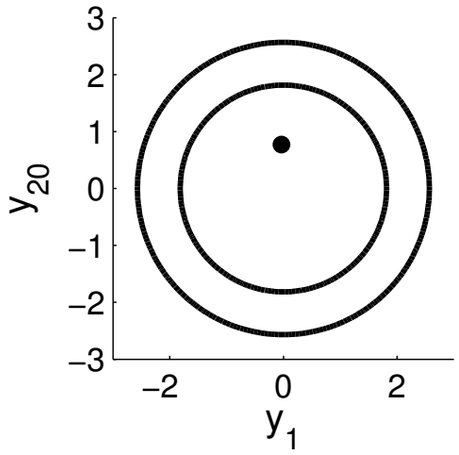


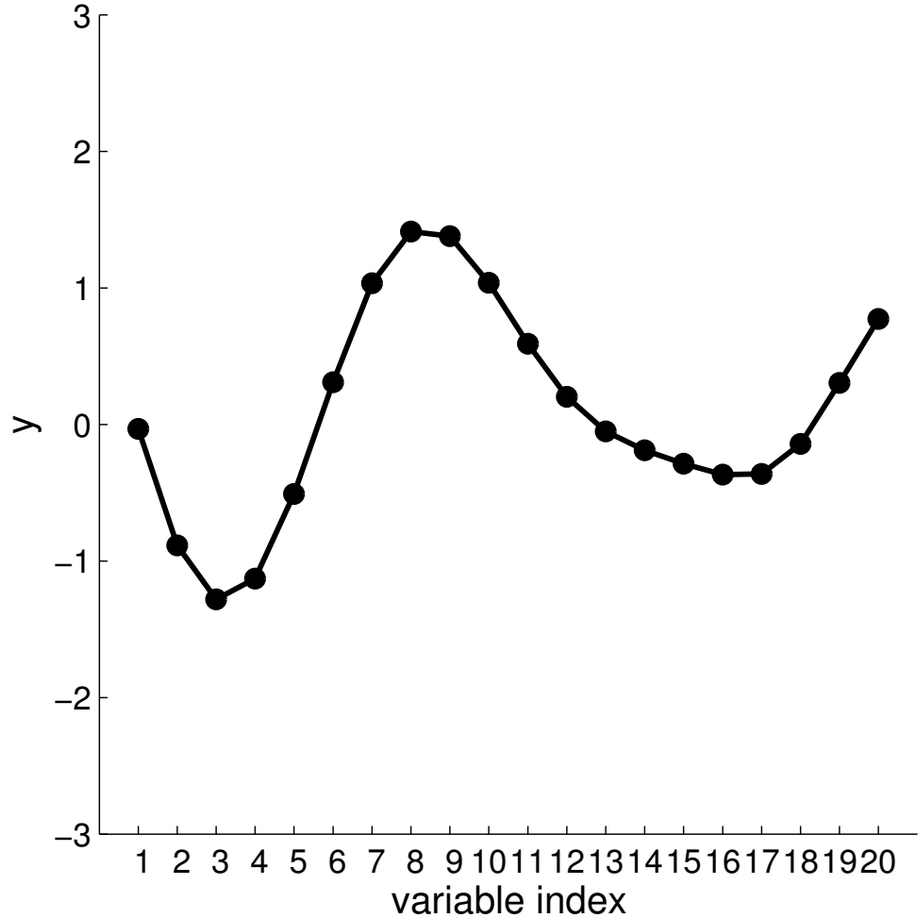


$$\Sigma =$$

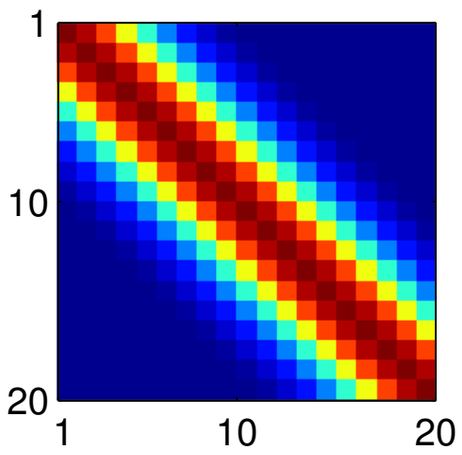


Special covariance matrix

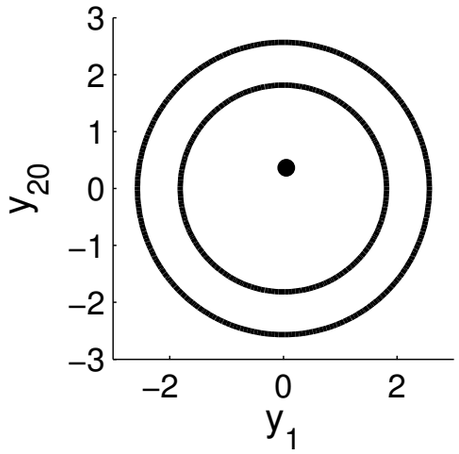


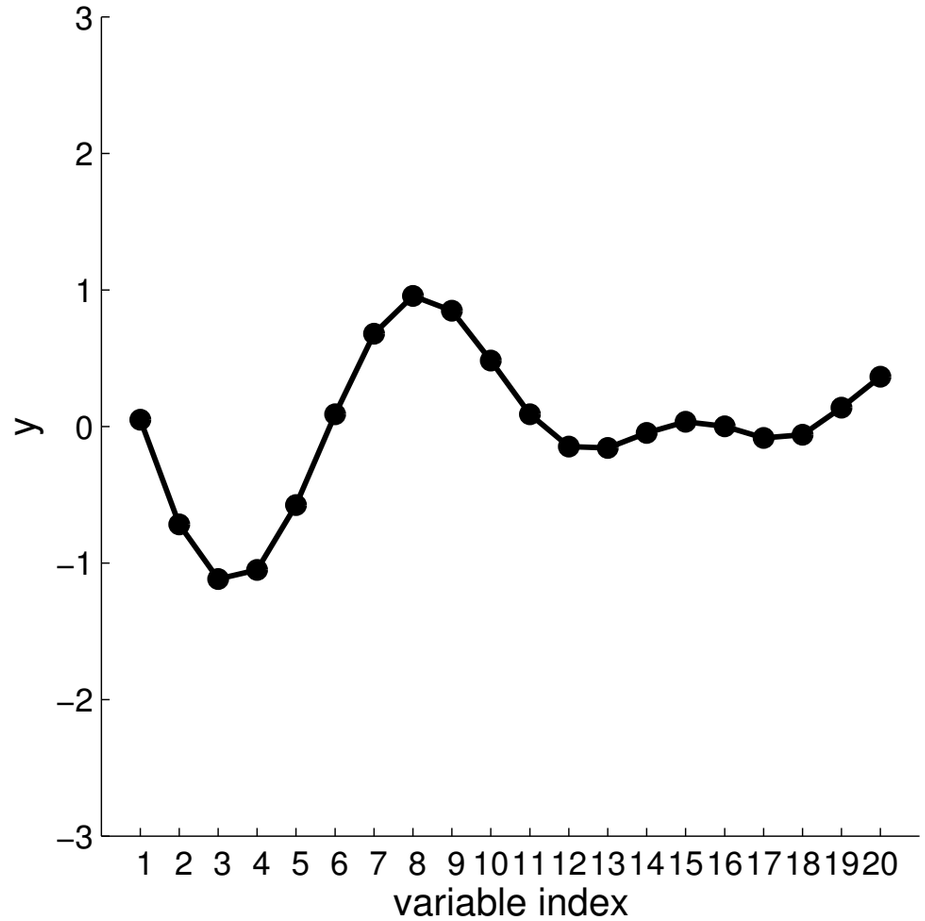


$\Sigma =$

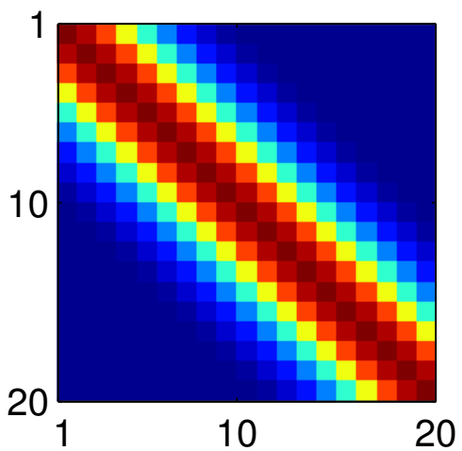


Special covariance matrix

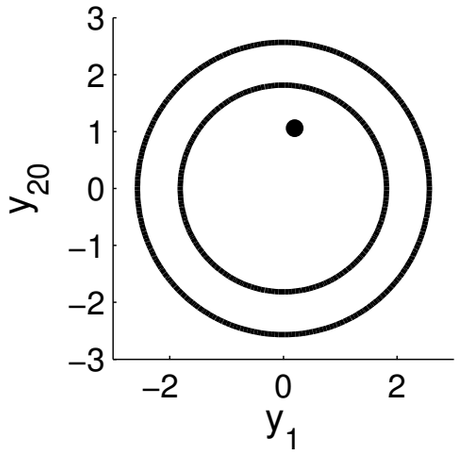


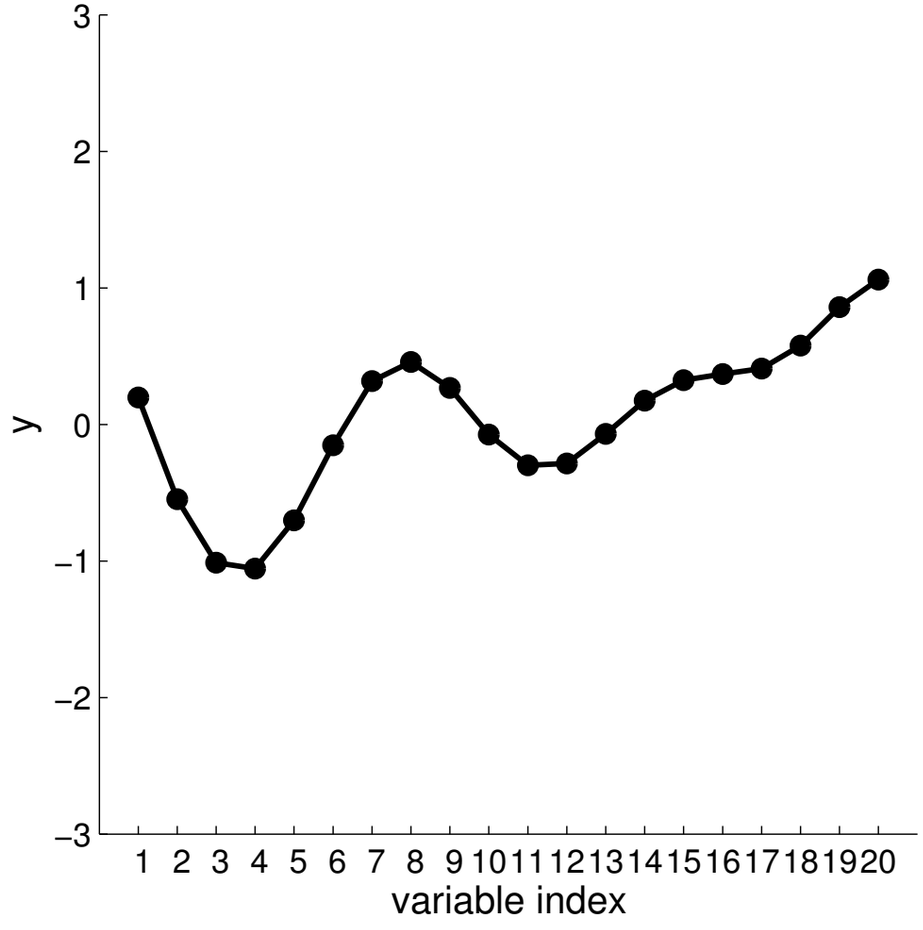


$\Sigma =$

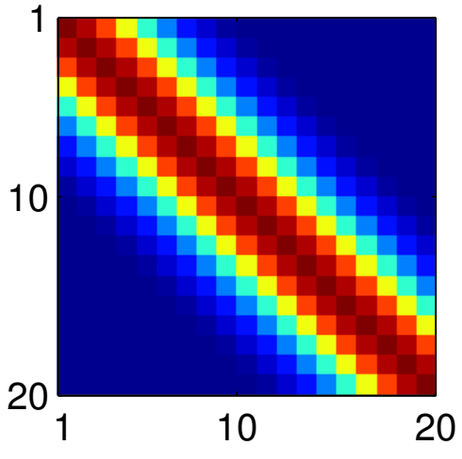


Special covariance matrix

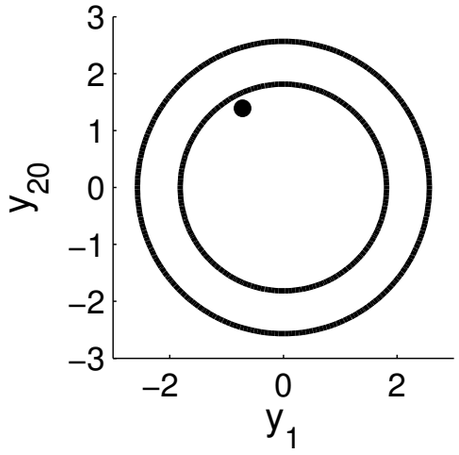


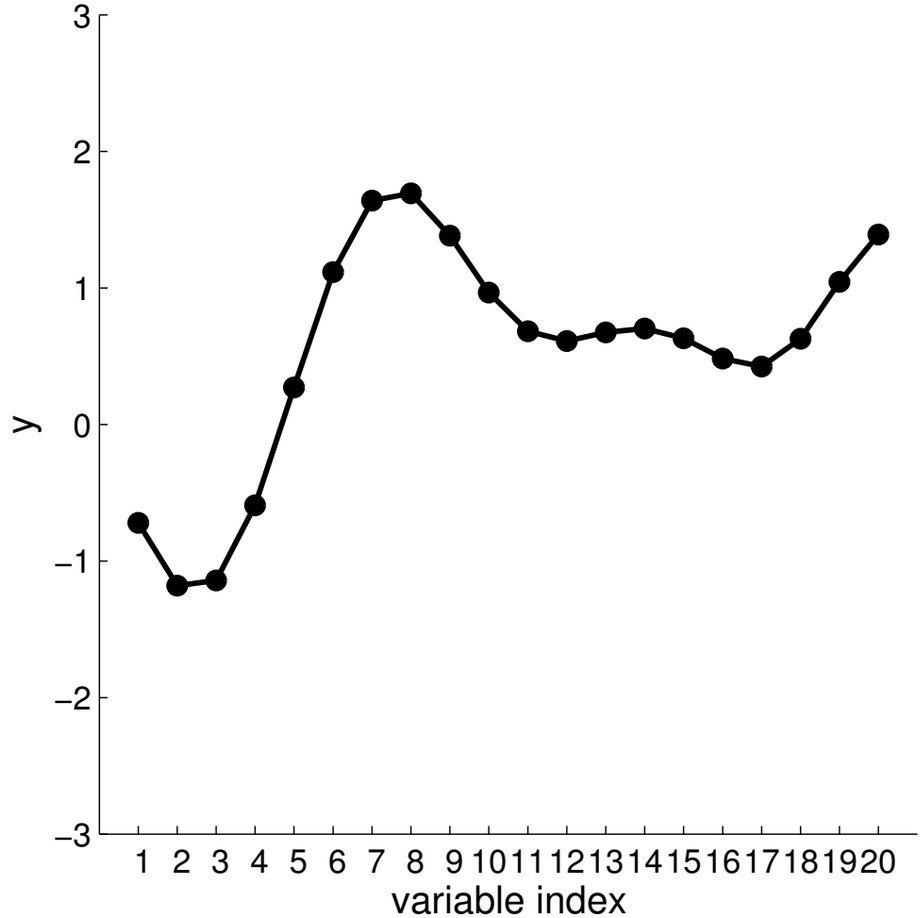


$\Sigma =$

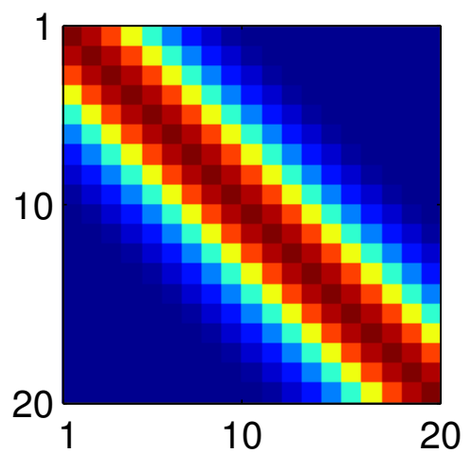


Special covariance matrix

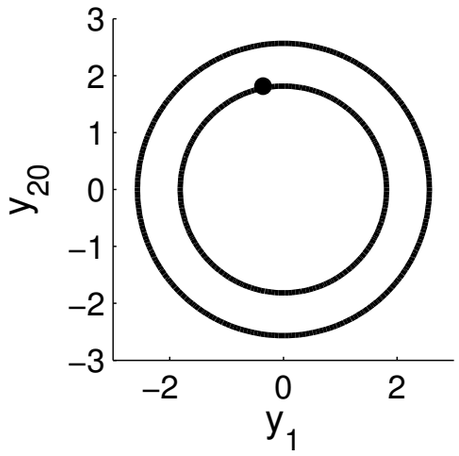


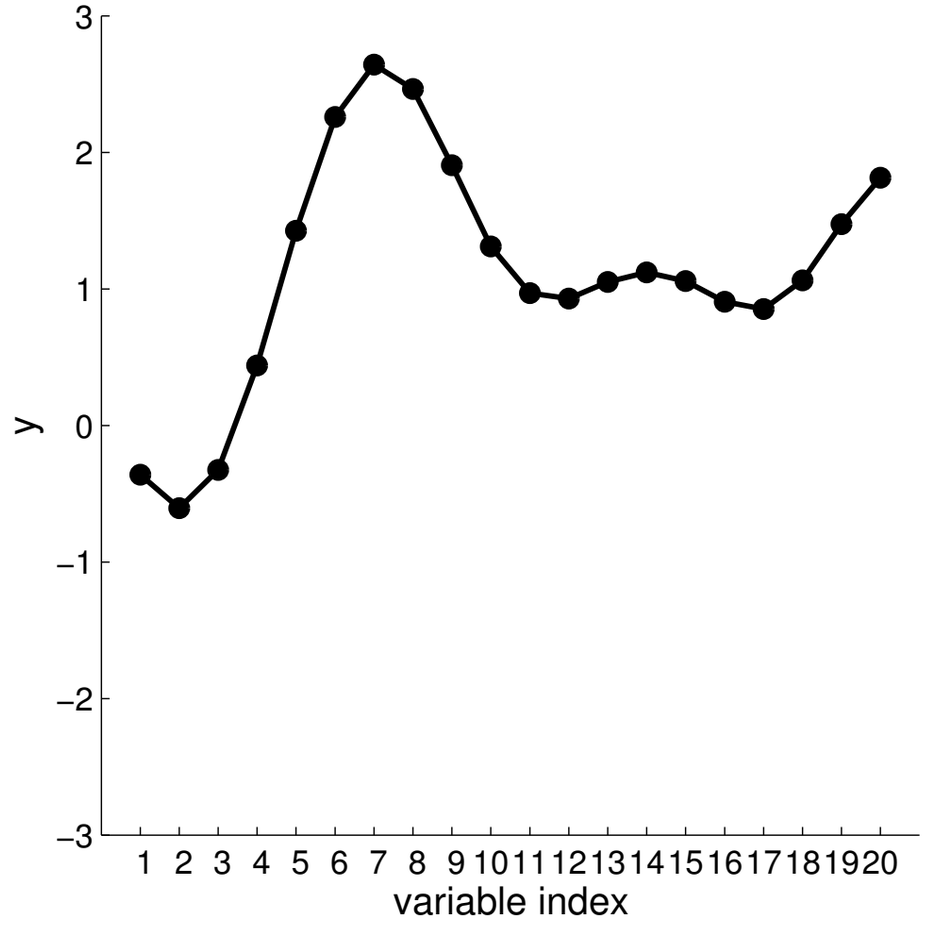


$\Sigma =$

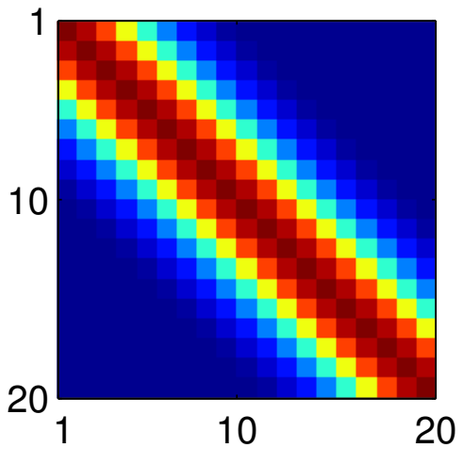


Special covariance matrix

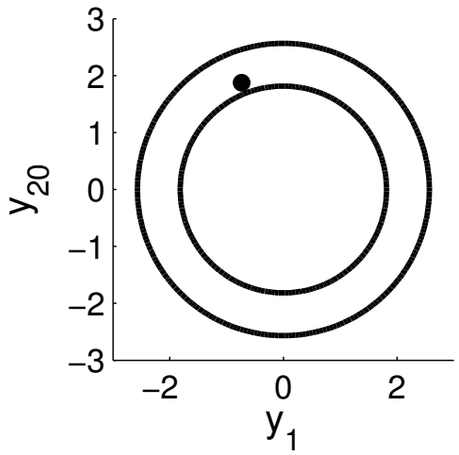


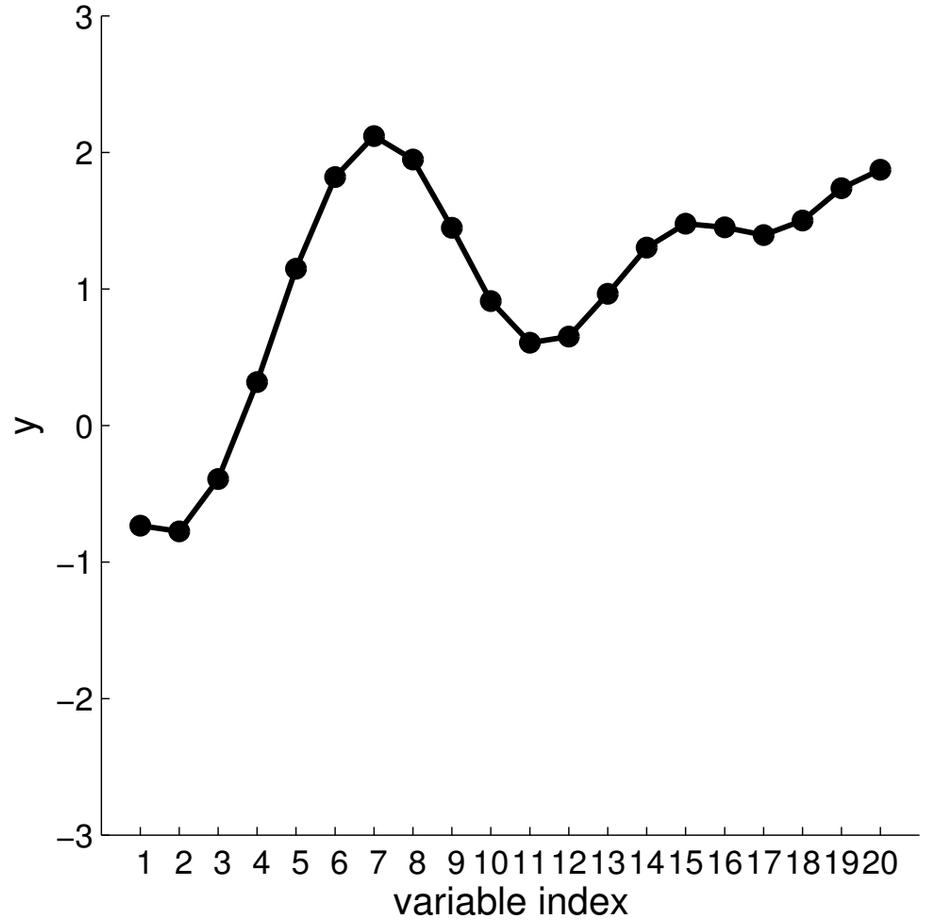


$\Sigma =$

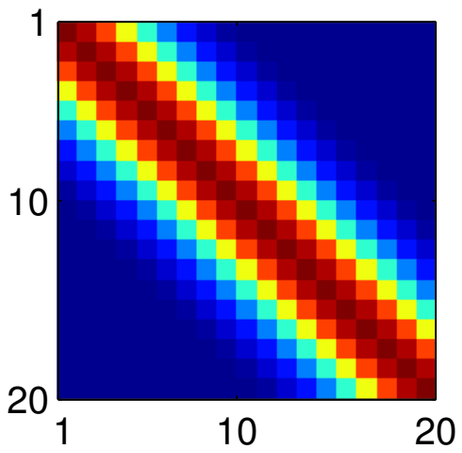


Special covariance matrix

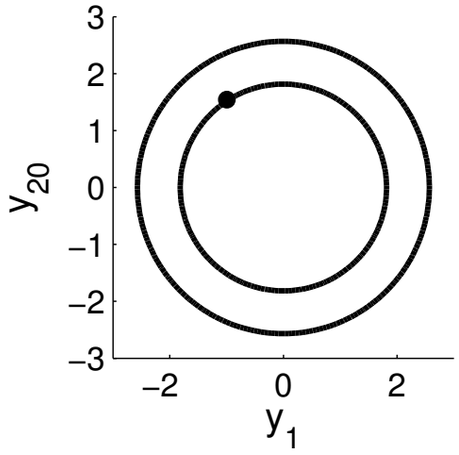


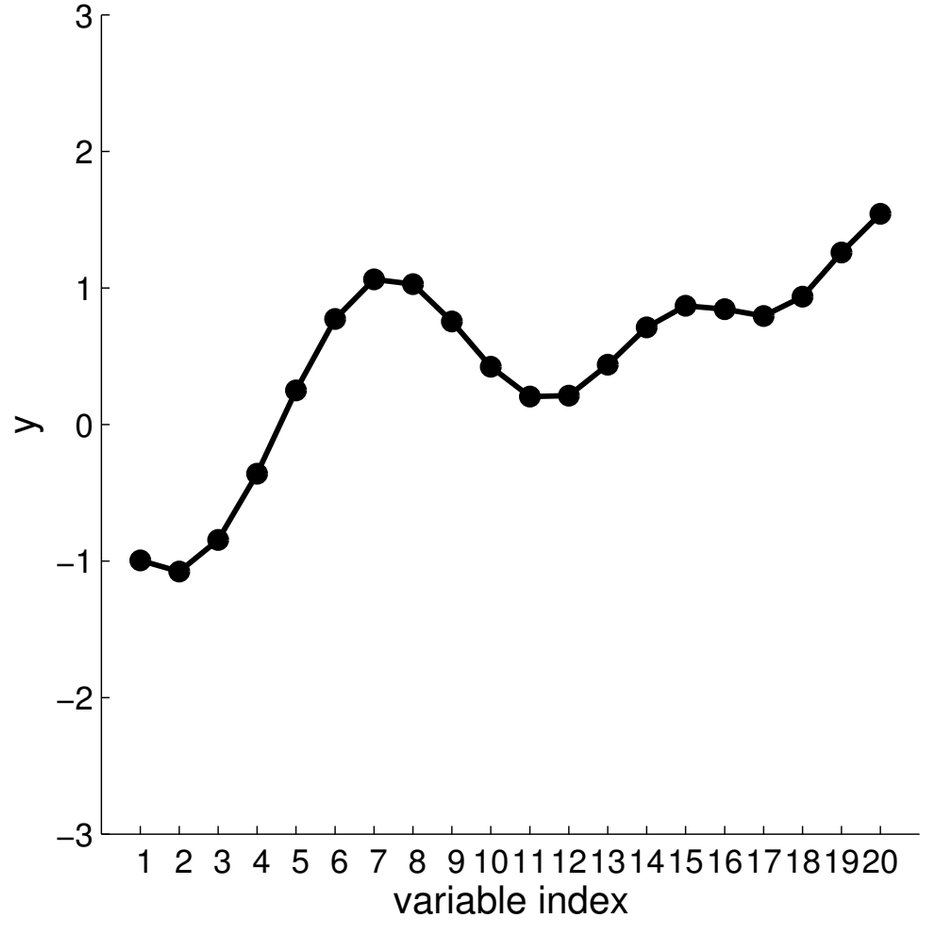


$\Sigma =$

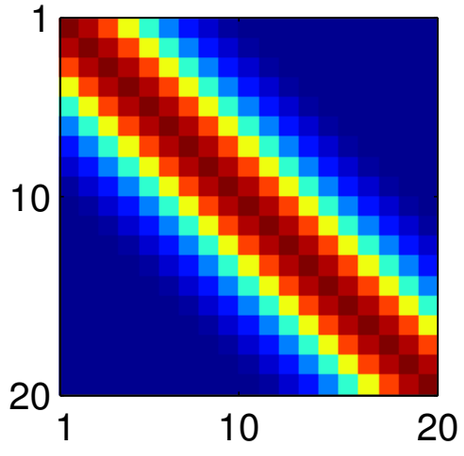


Special covariance matrix

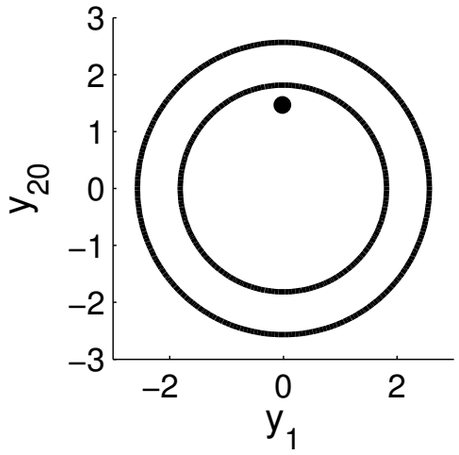


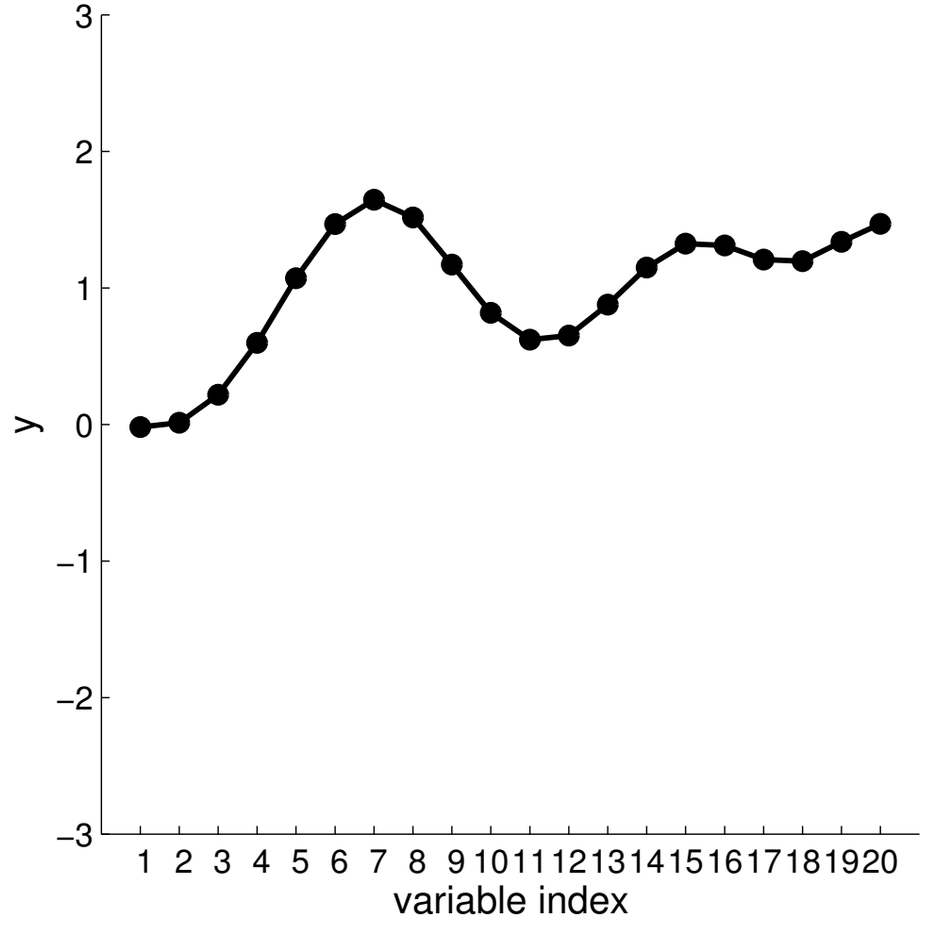


$\Sigma =$

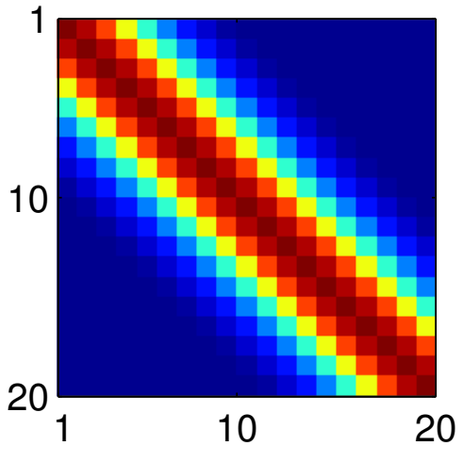


Special covariance matrix

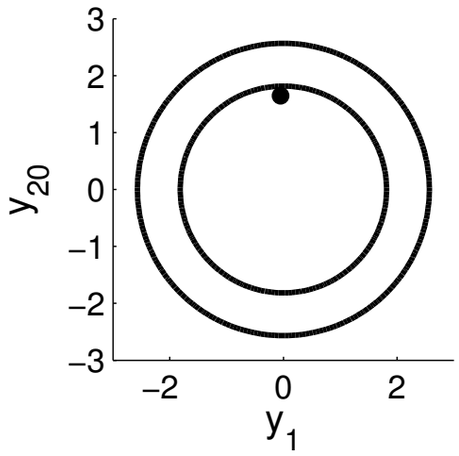


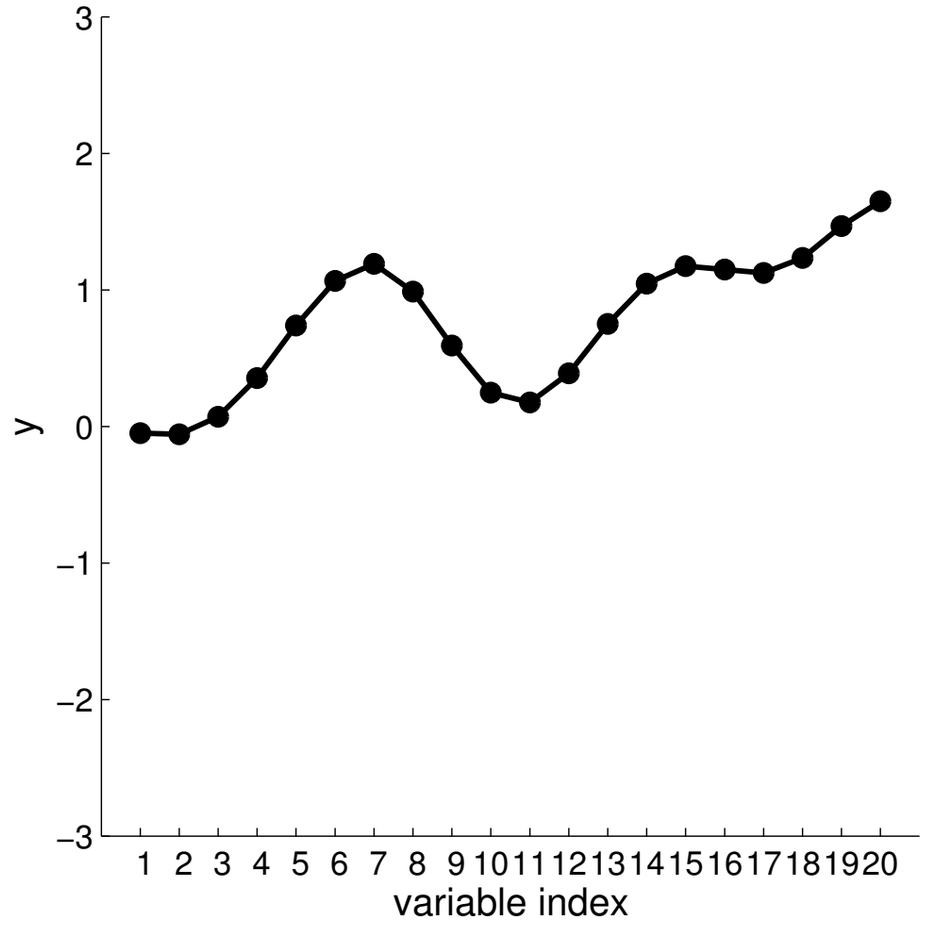


$\Sigma =$

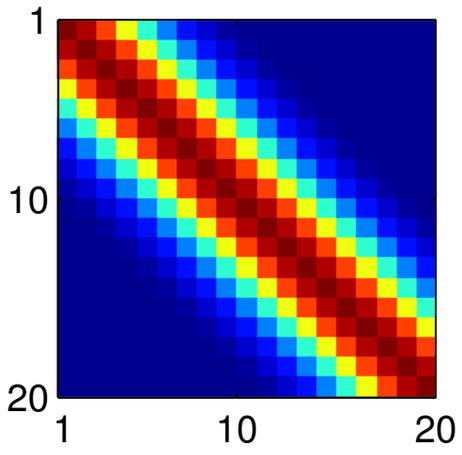


Special covariance matrix

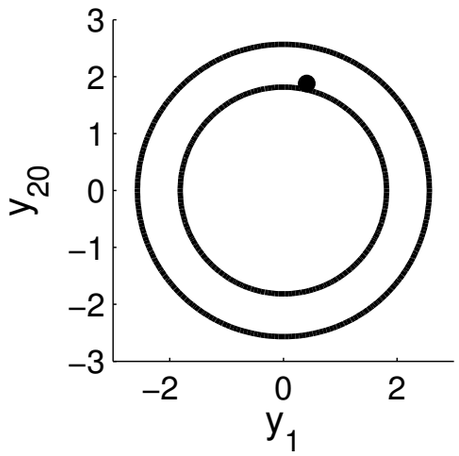


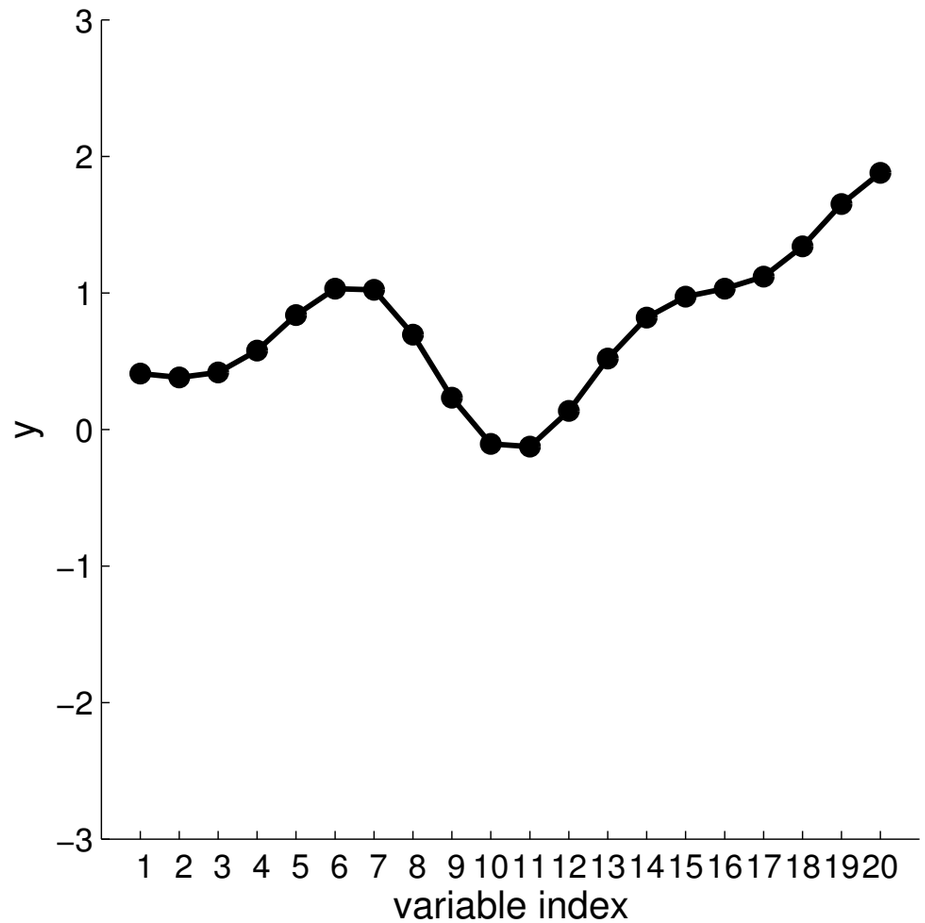


$\Sigma =$

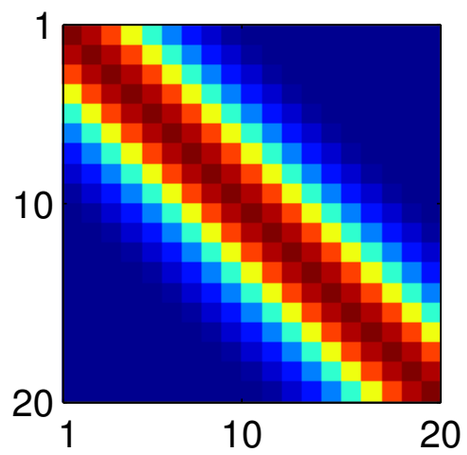


Special covariance matrix

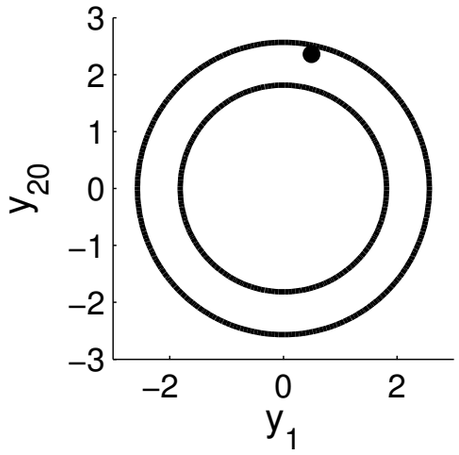


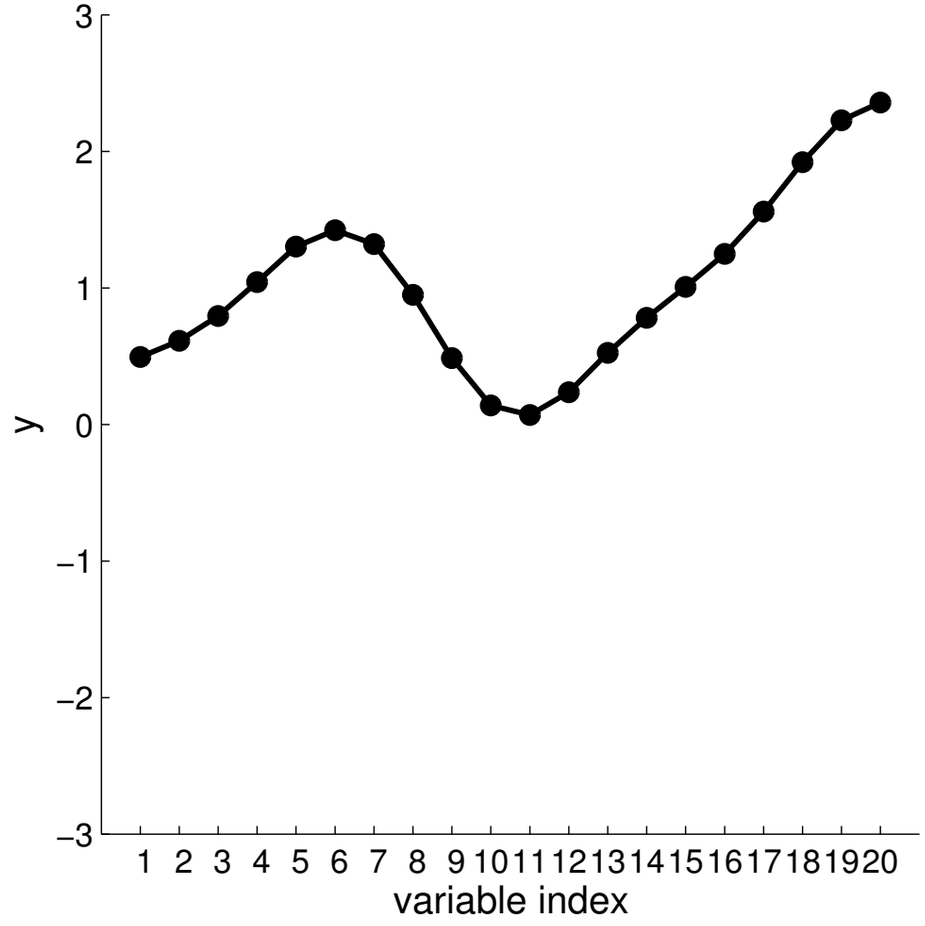


$\Sigma =$

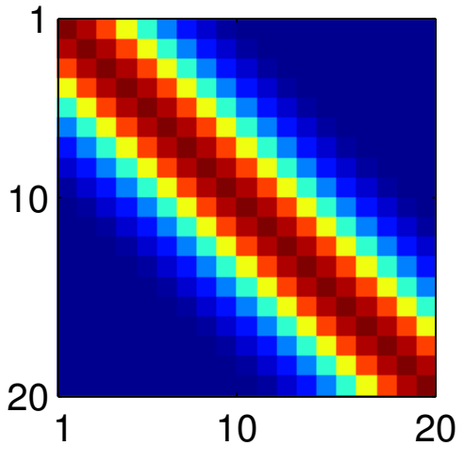


Special covariance matrix



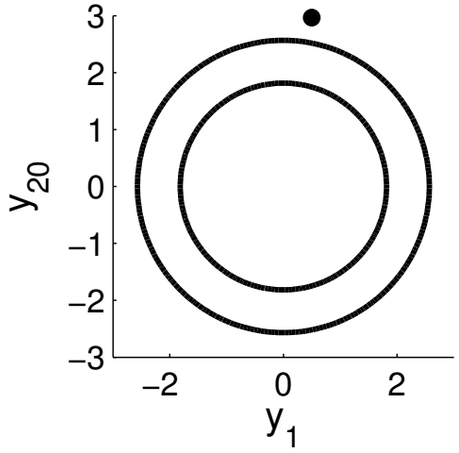


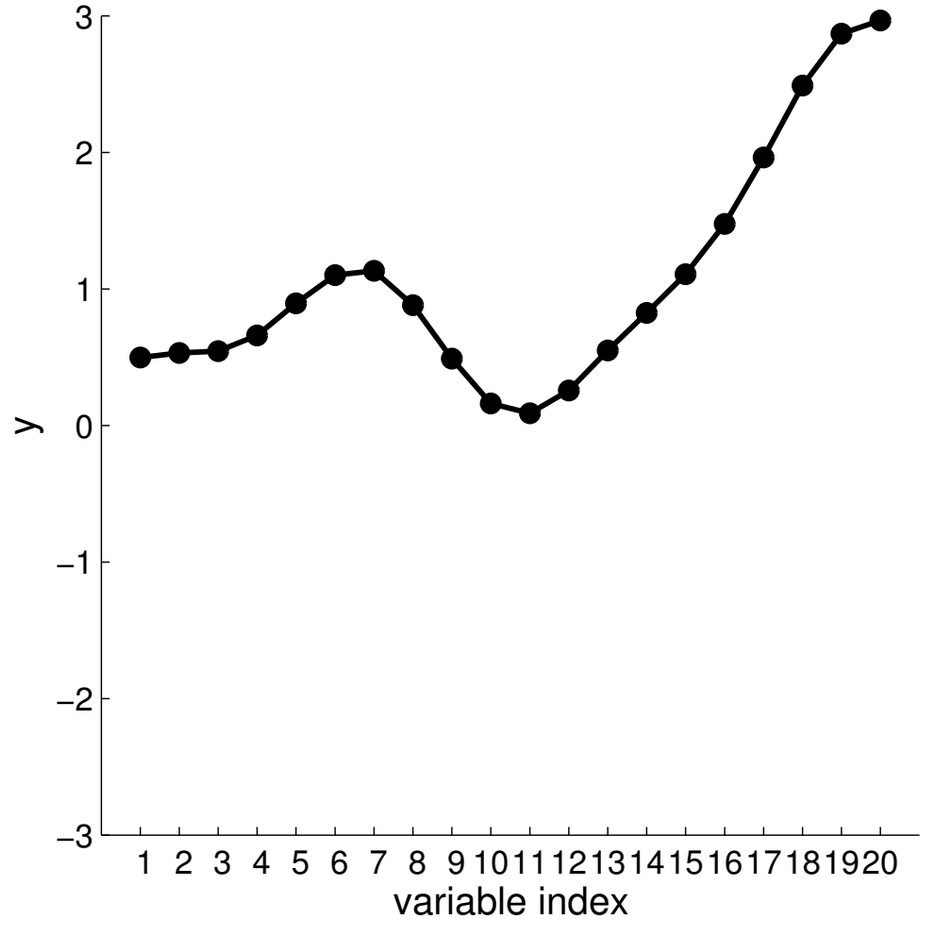
$\Sigma =$



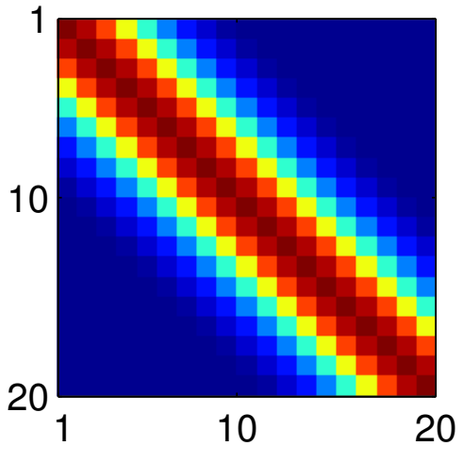
Special covariance matrix

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

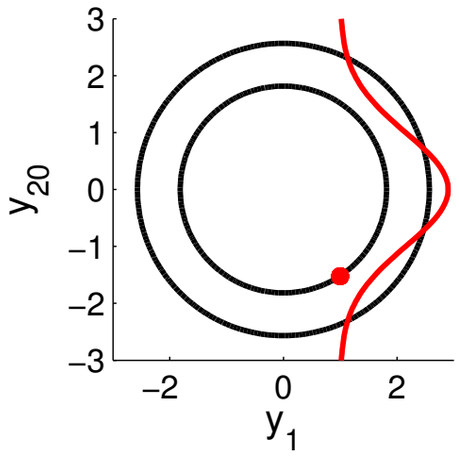


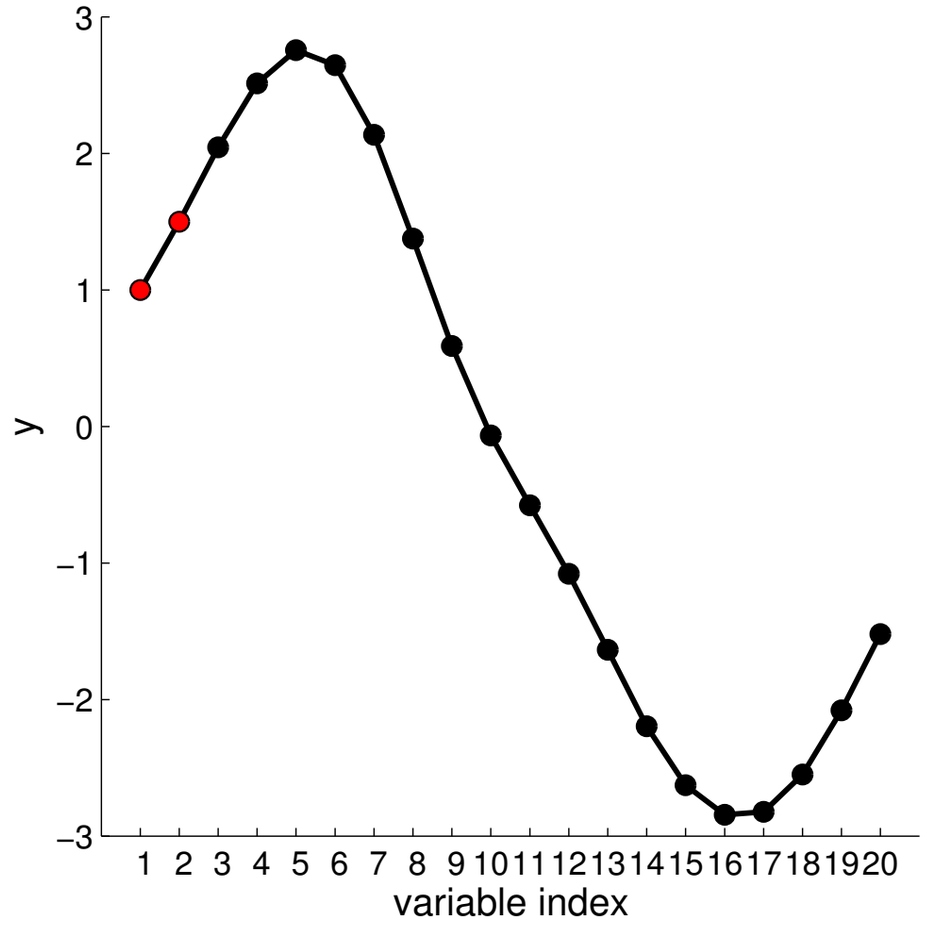


$\Sigma =$

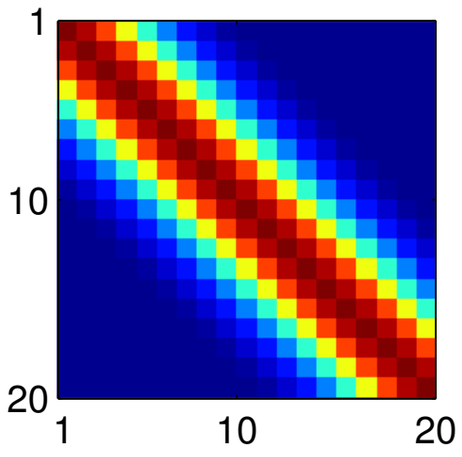


Special covariance matrix - conditioning



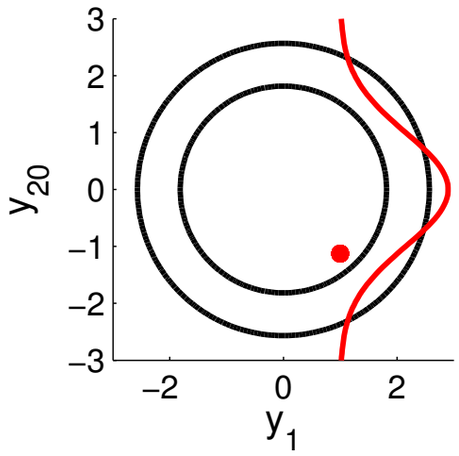


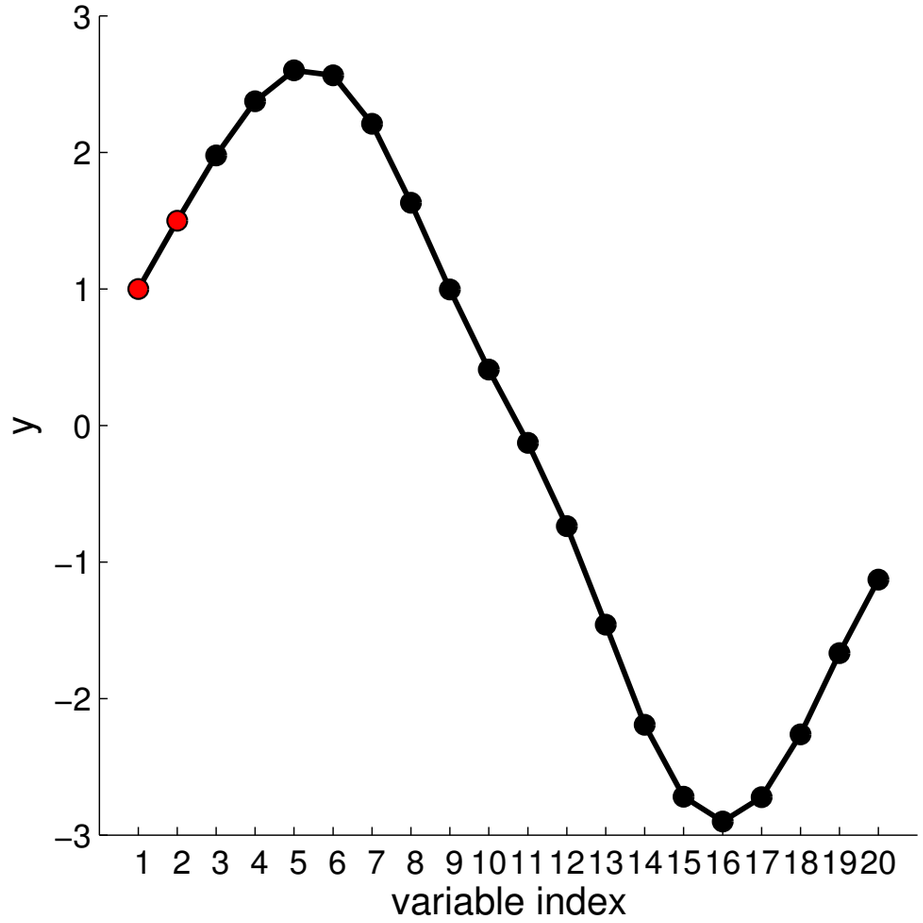
$\Sigma =$



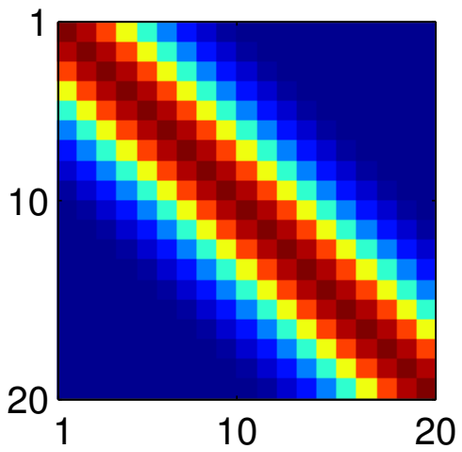
Conditioning on y1 and y2

Special covariance matrix - conditioning



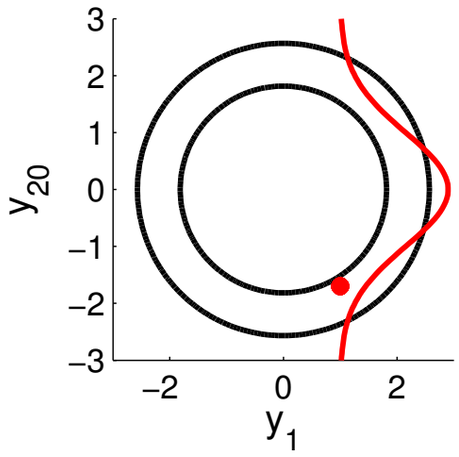


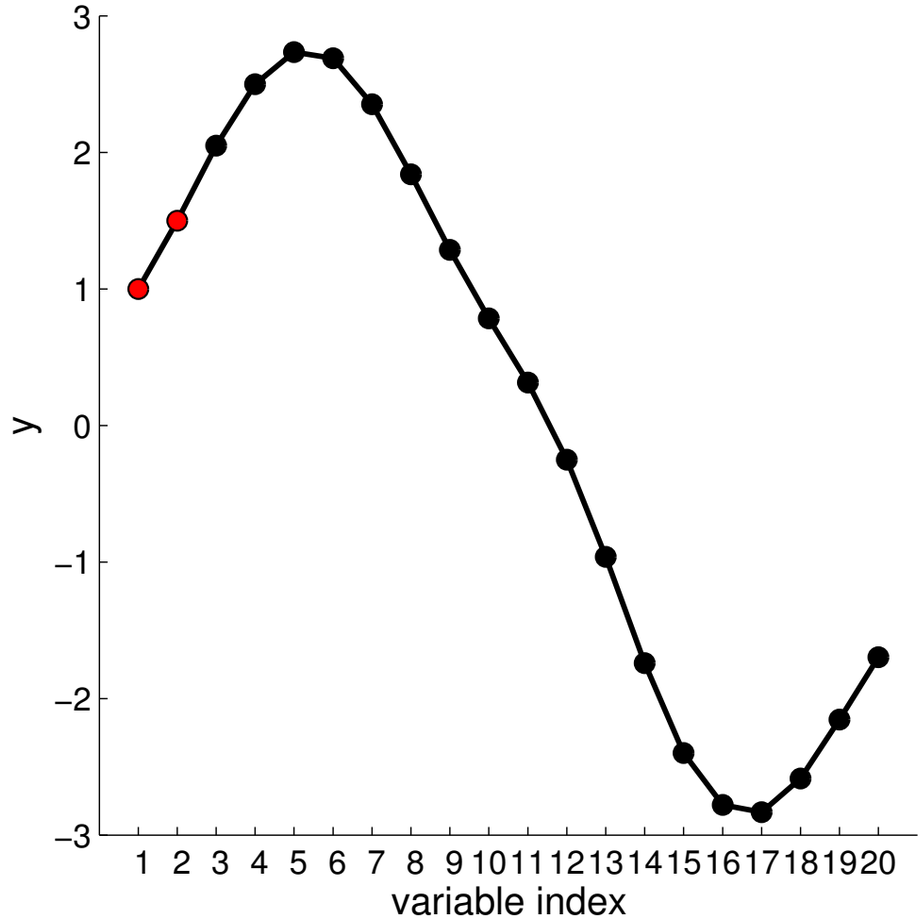
$\Sigma =$



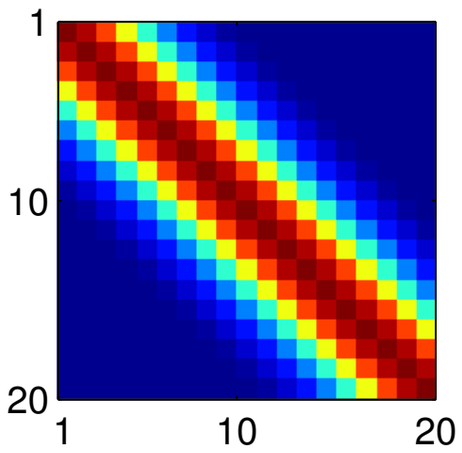
Conditioning on y1 and y2

Special covariance matrix - conditioning



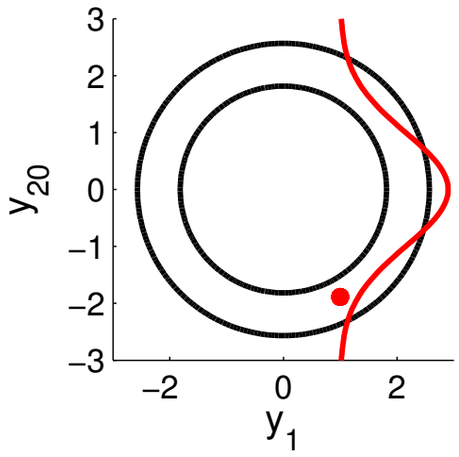


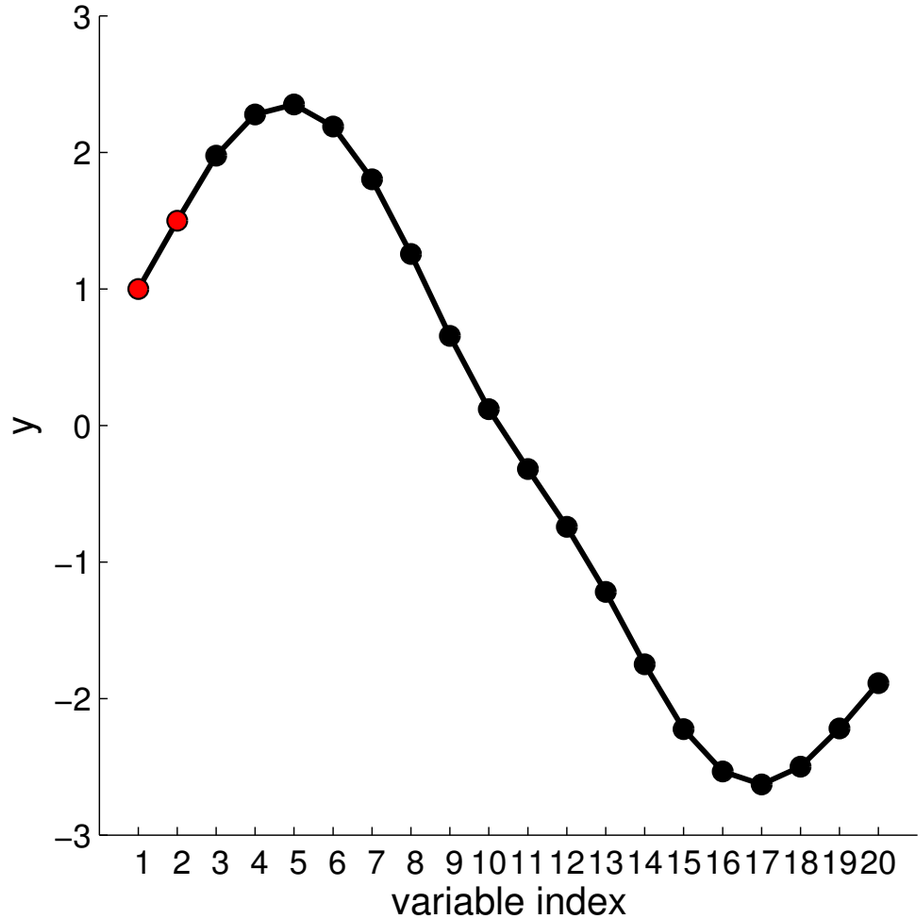
$\Sigma =$



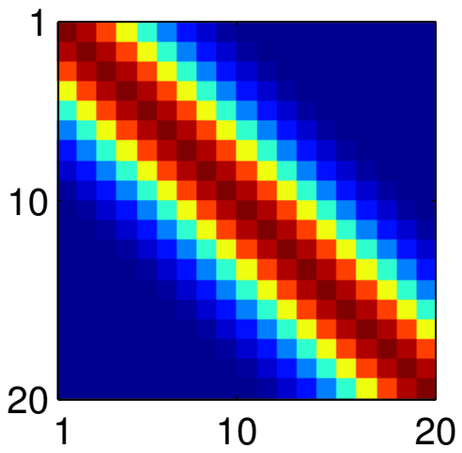
Conditioning on y1 and y2

Special covariance matrix - conditioning



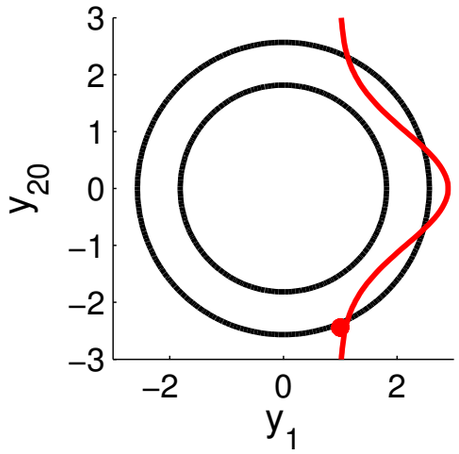


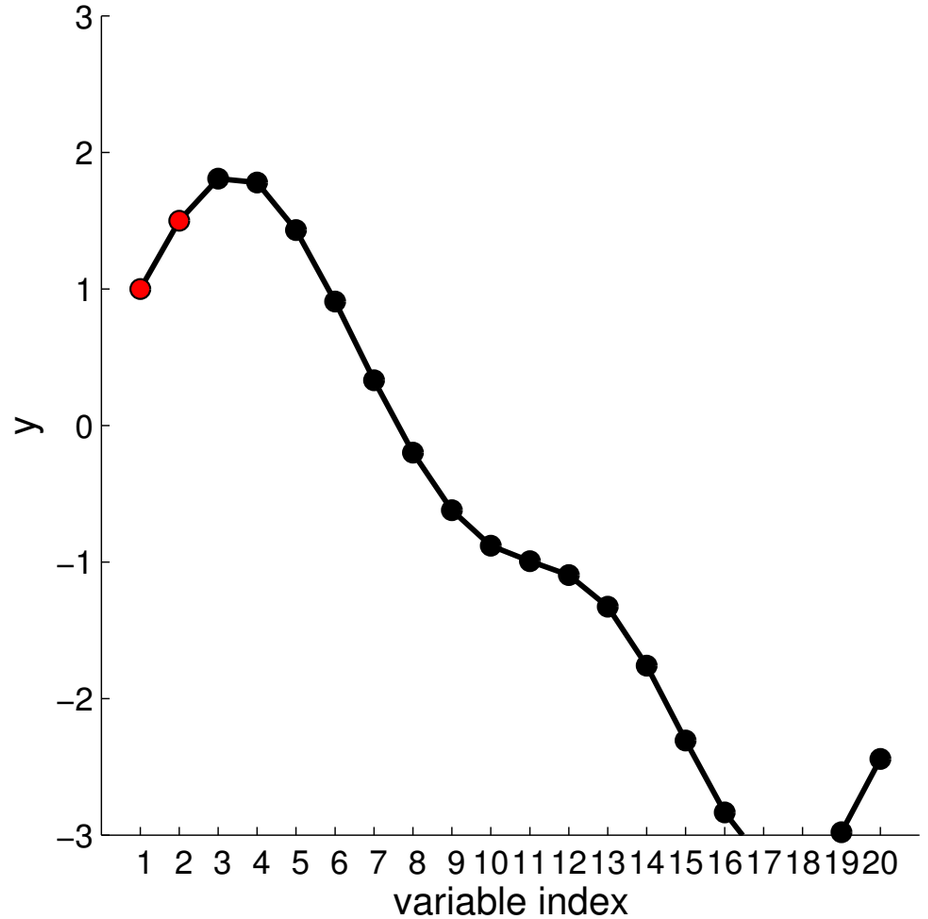
$\Sigma =$



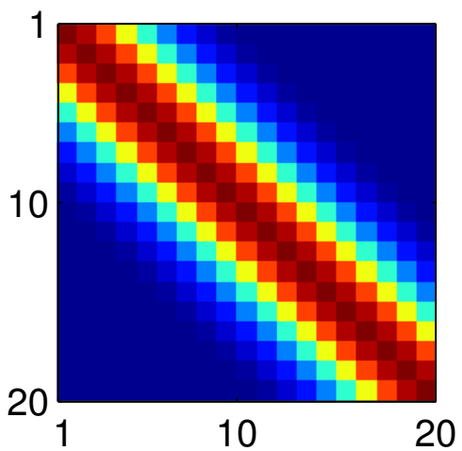
Conditioning on y1 and y2

Special covariance matrix - conditioning



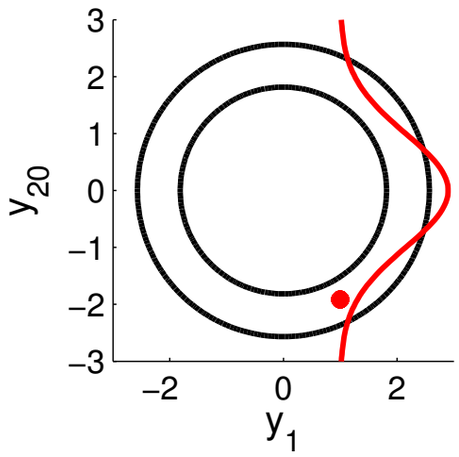


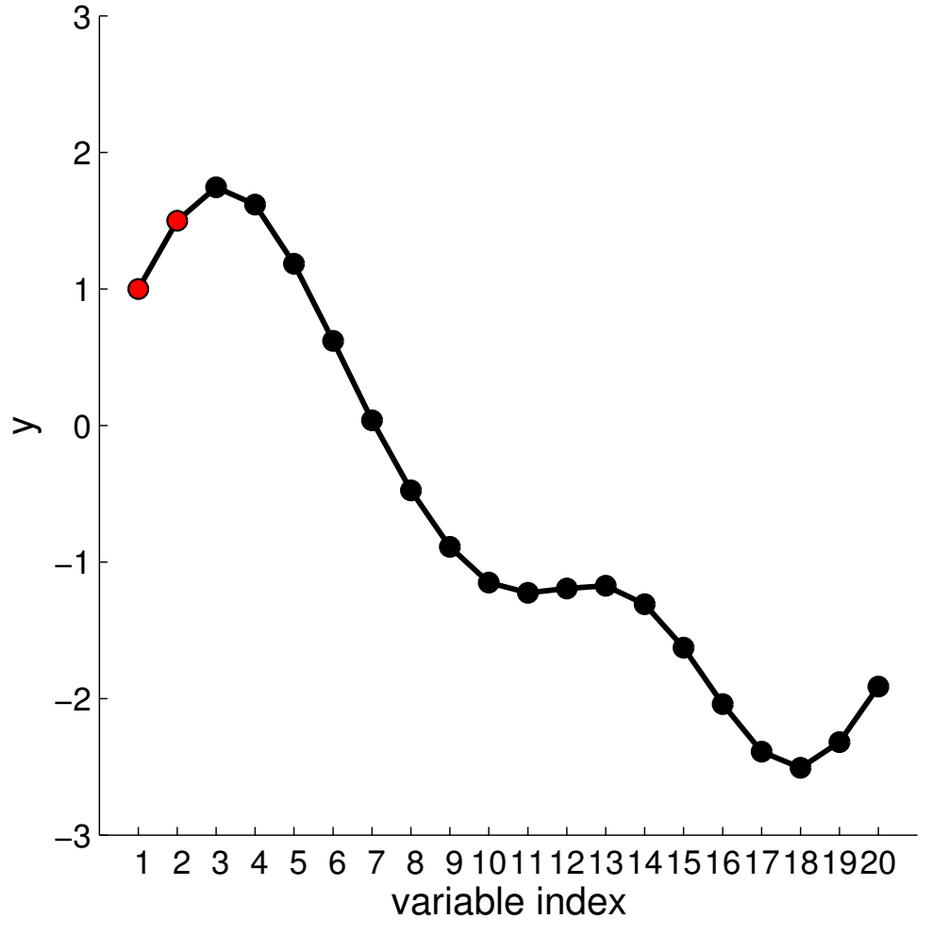
$\Sigma =$



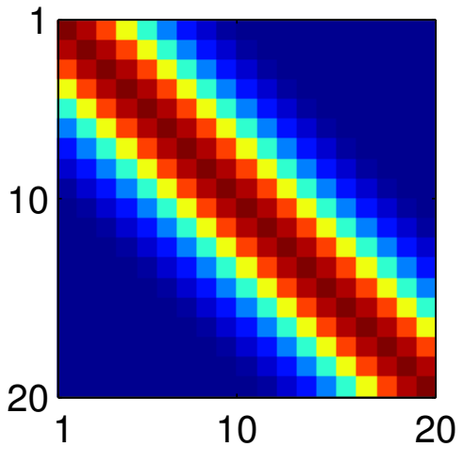
Conditioning on y1 and y2

Special covariance matrix - conditioning



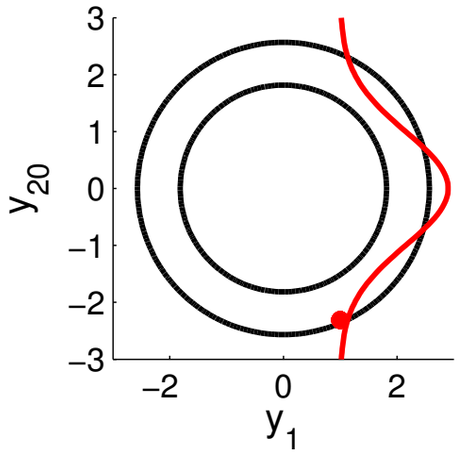


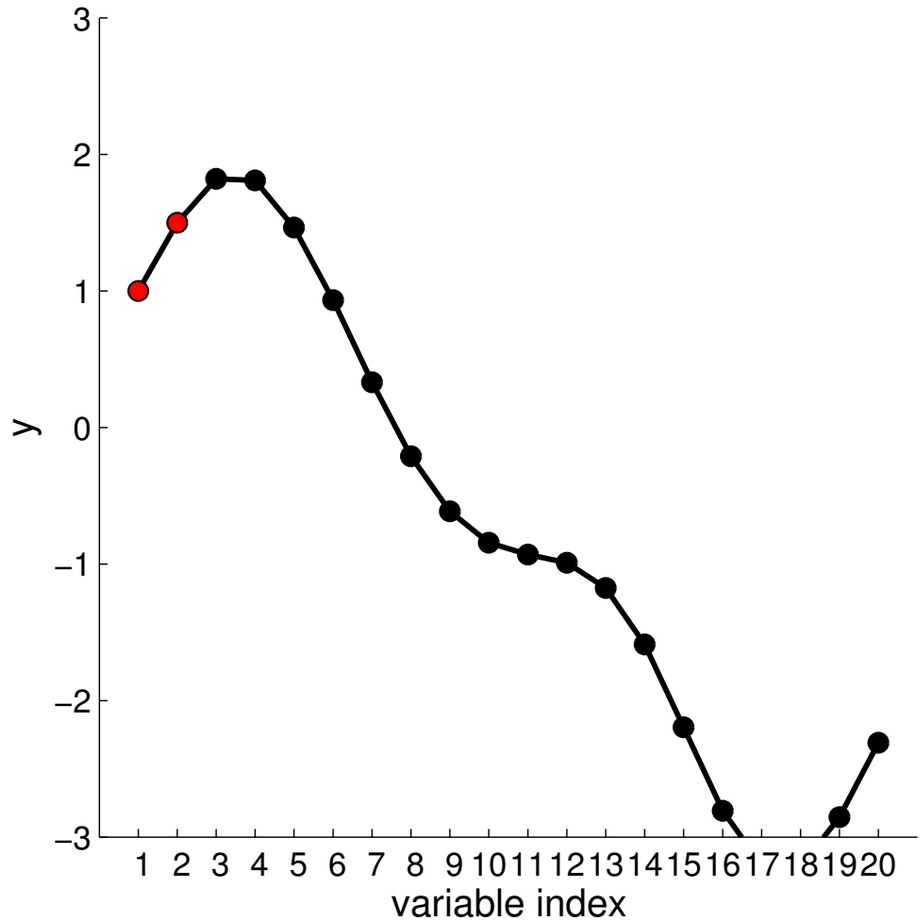
$\Sigma =$



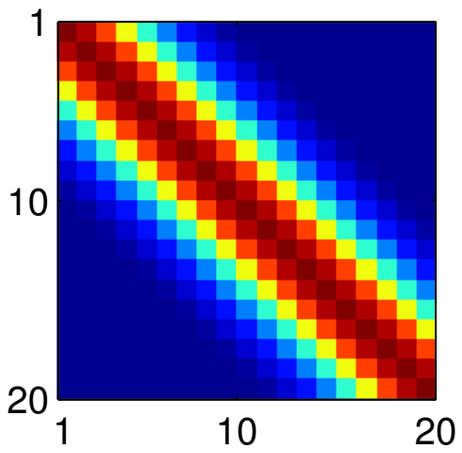
Conditioning on y1 and y2

Special covariance matrix - conditioning



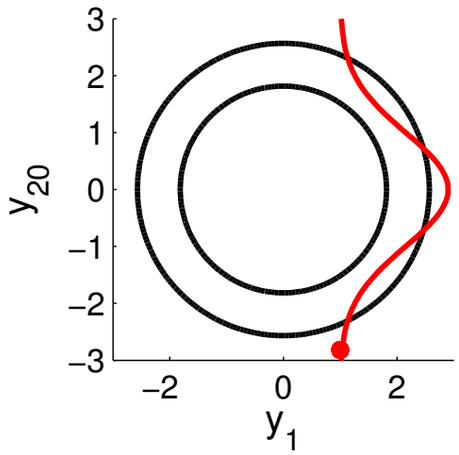


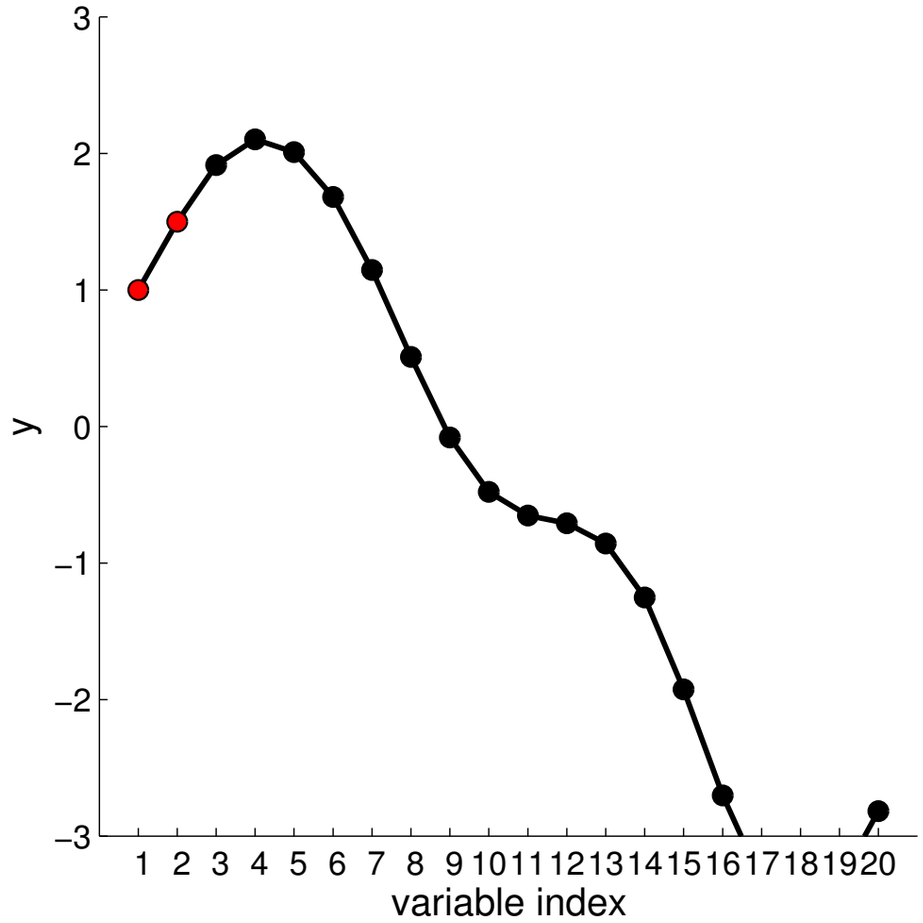
$\Sigma =$



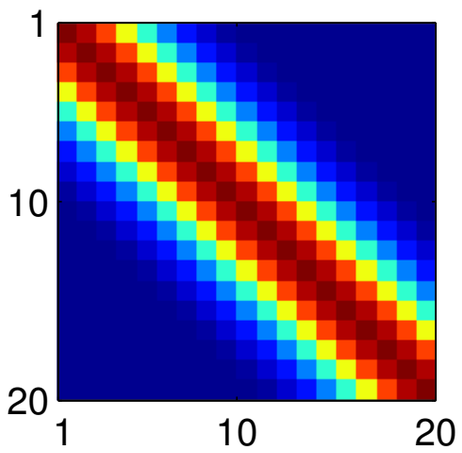
Conditioning on y1 and y2

Special covariance matrix - conditioning



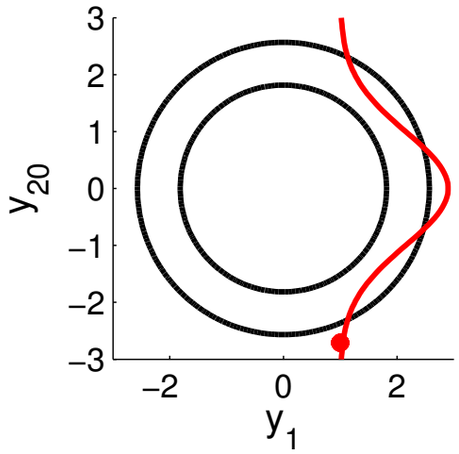


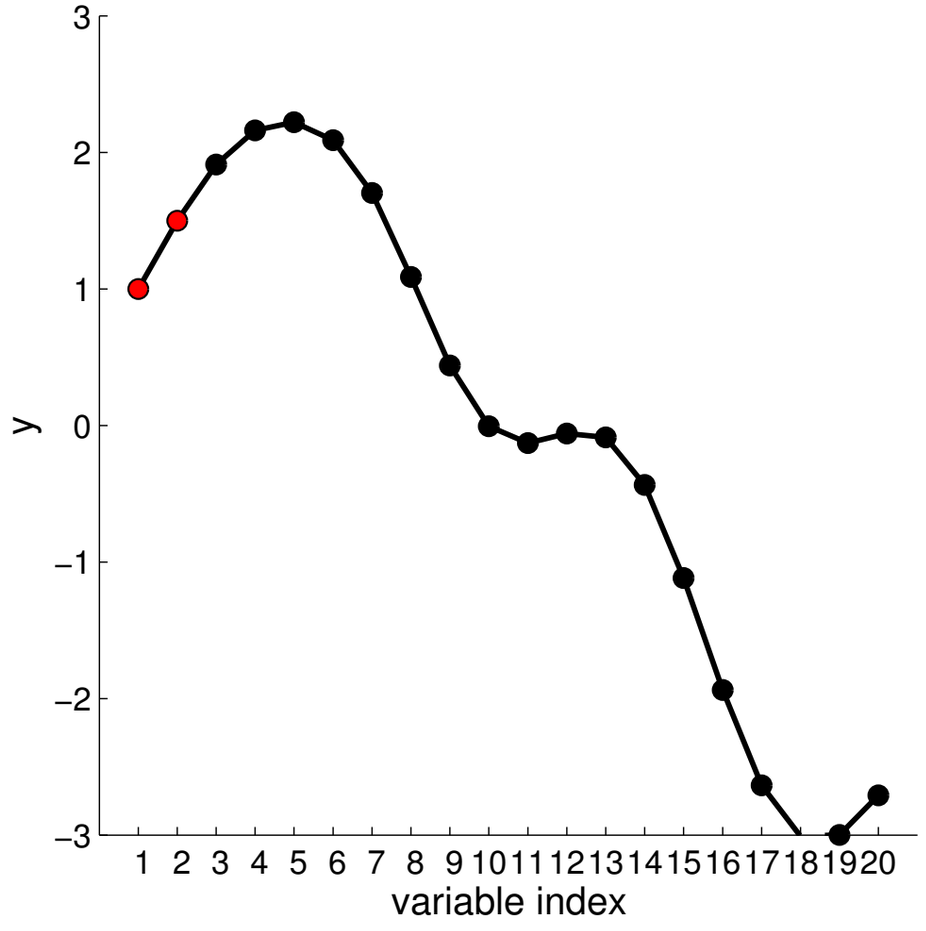
$\Sigma =$



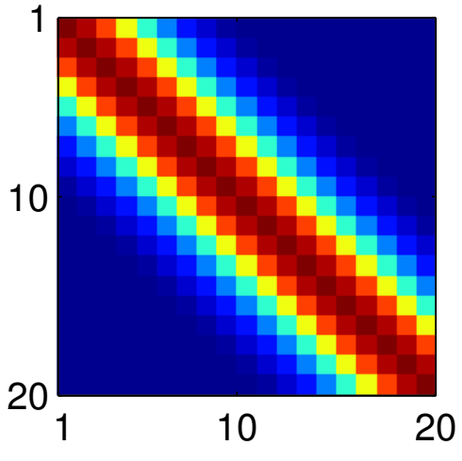
Conditioning on y1 and y2

Special covariance matrix - conditioning



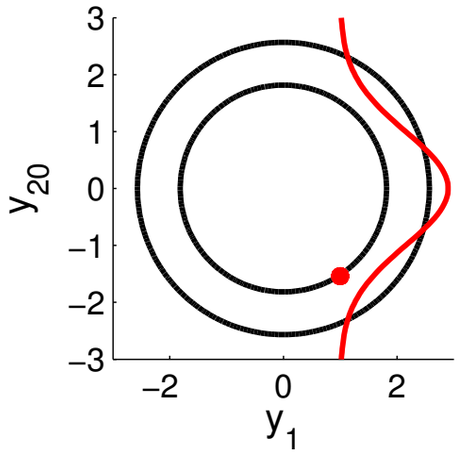


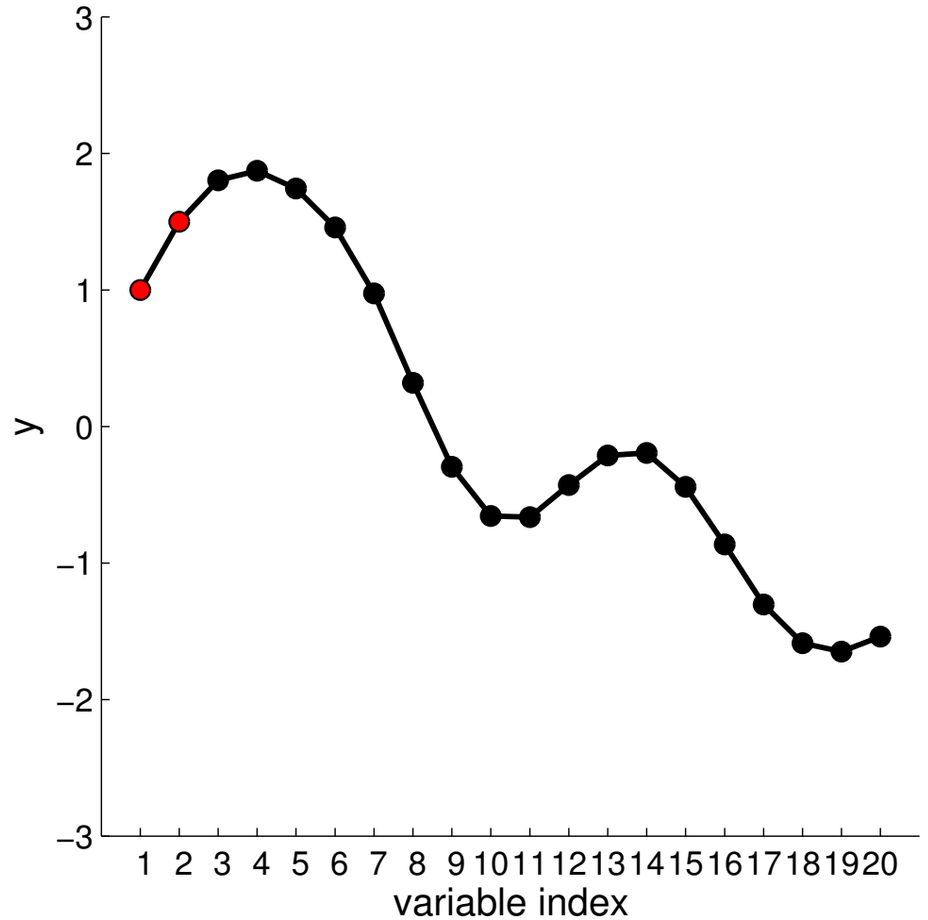
$\Sigma =$



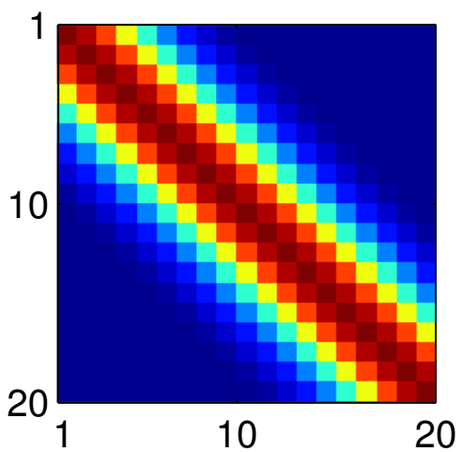
Conditioning on y1 and y2

Special covariance matrix - conditioning



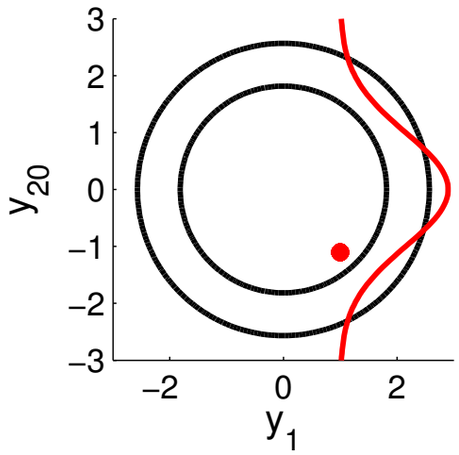


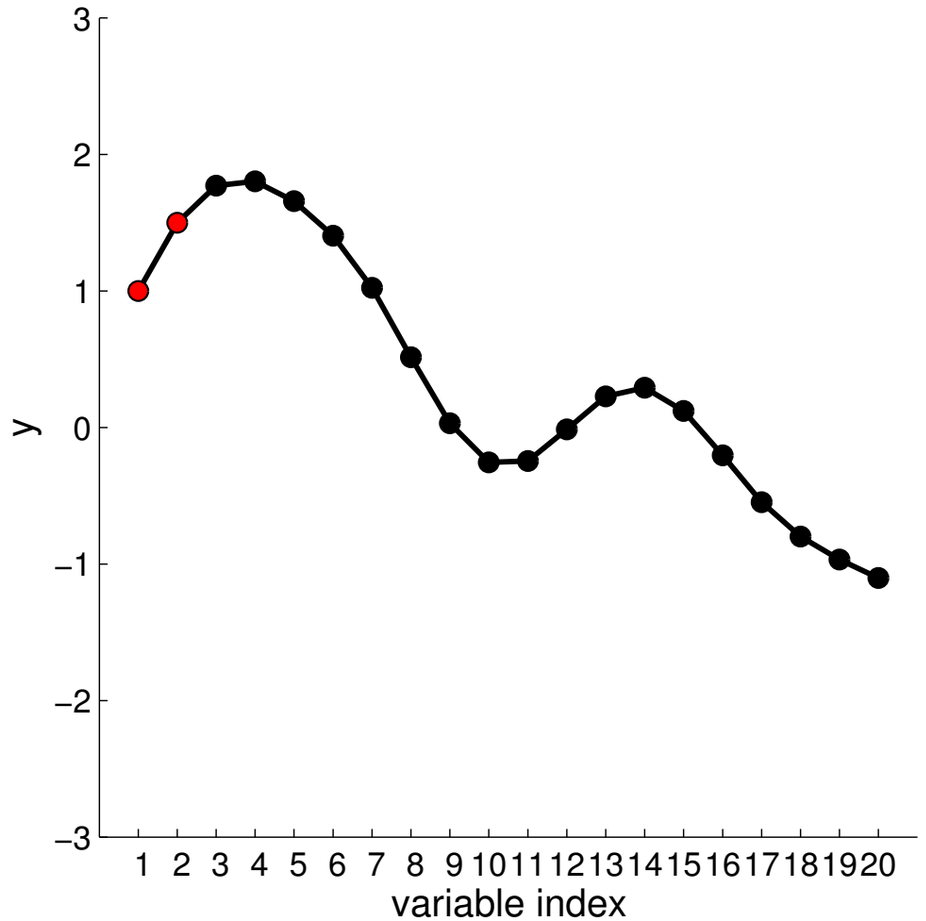
$\Sigma =$



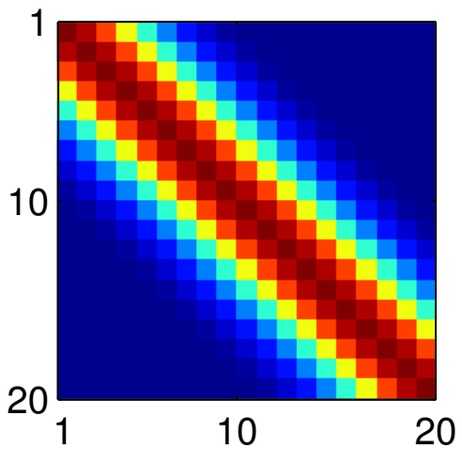
Conditioning on y1 and y2

Special covariance matrix - conditioning



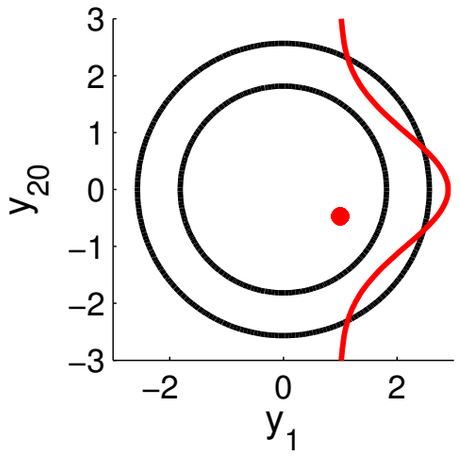


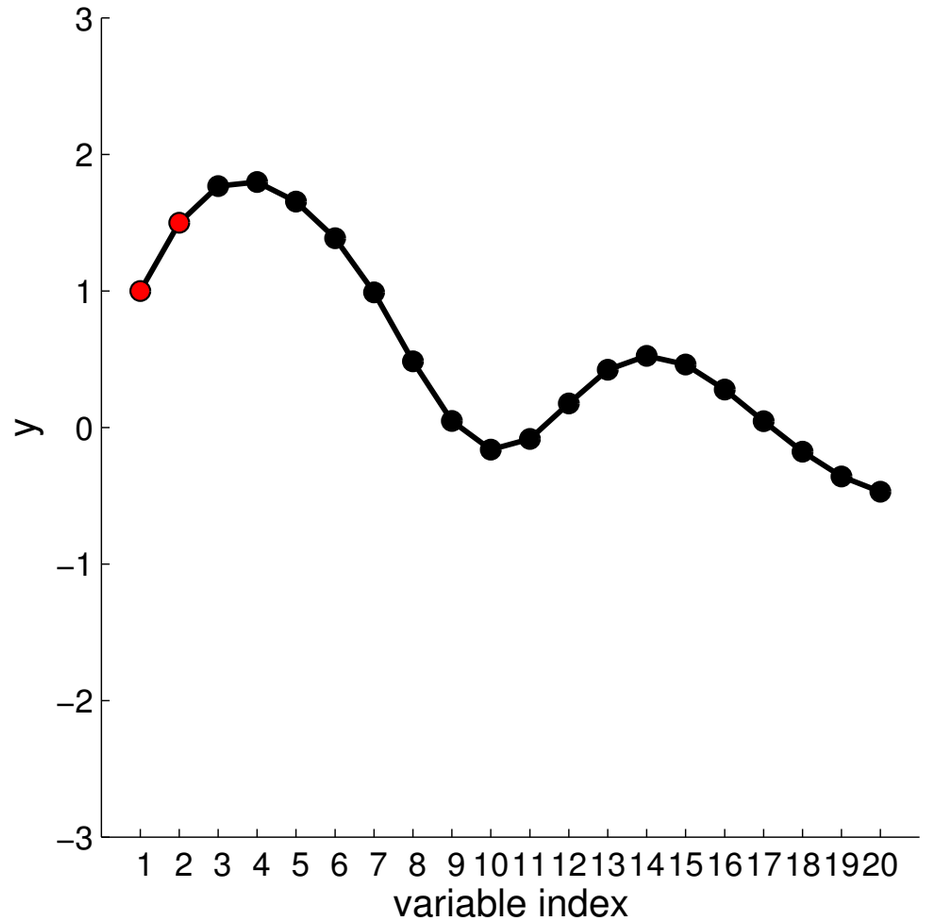
$\Sigma =$



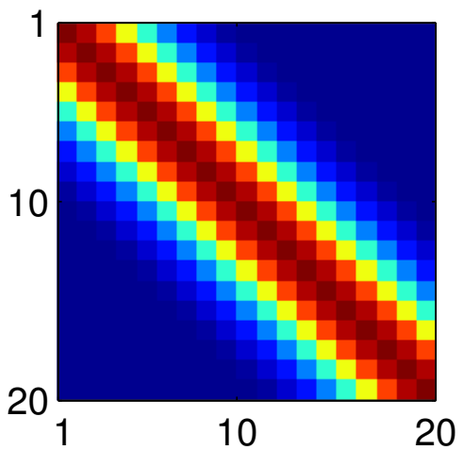
Conditioning on y1 and y2

Special covariance matrix - conditioning



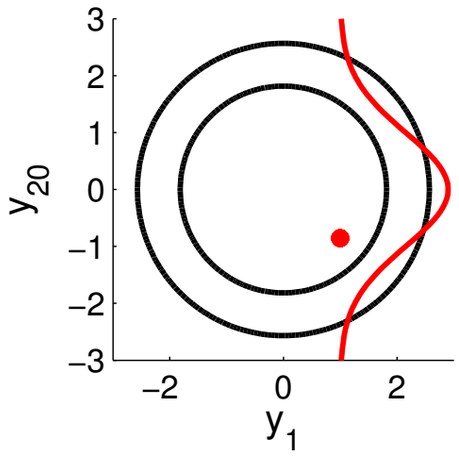


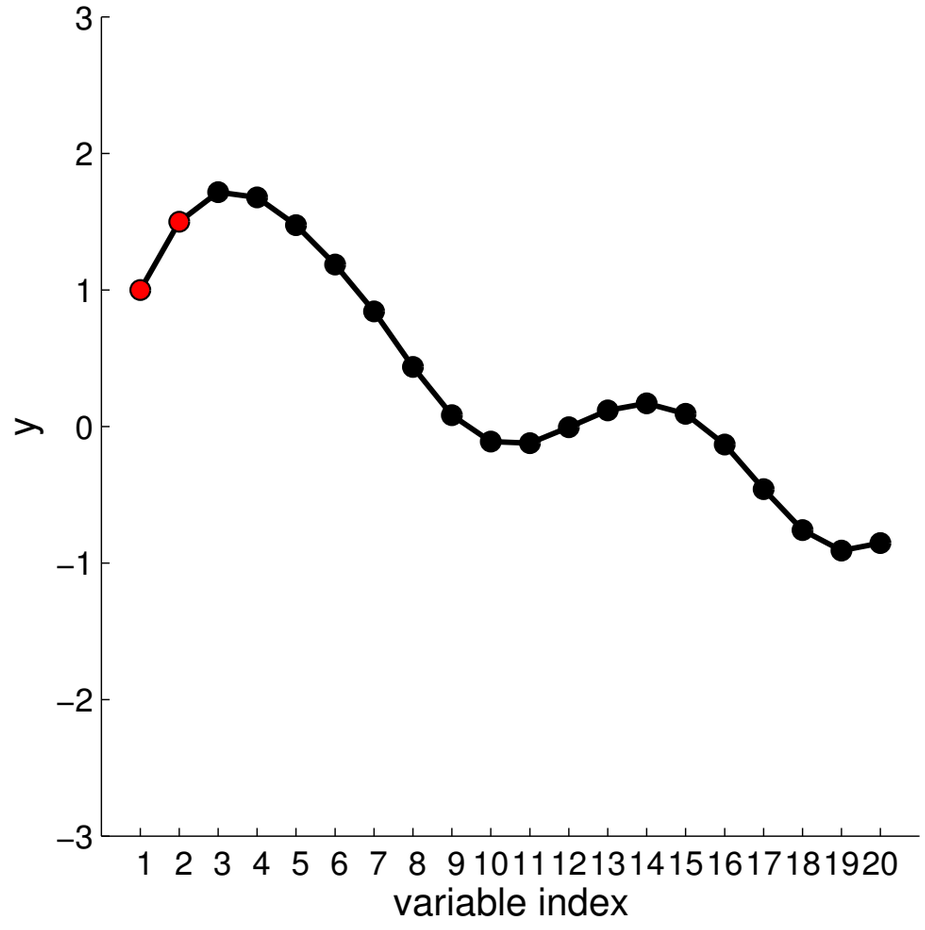
$\Sigma =$



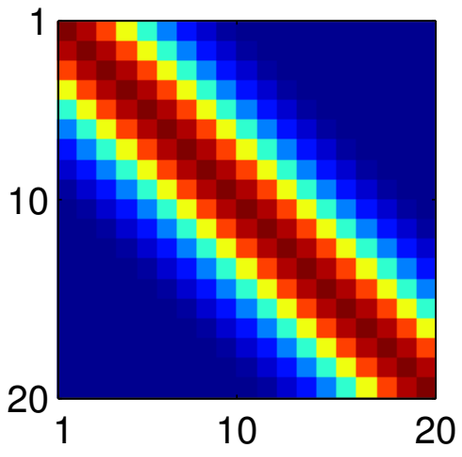
Conditioning on y1 and y2

Special covariance matrix - conditioning



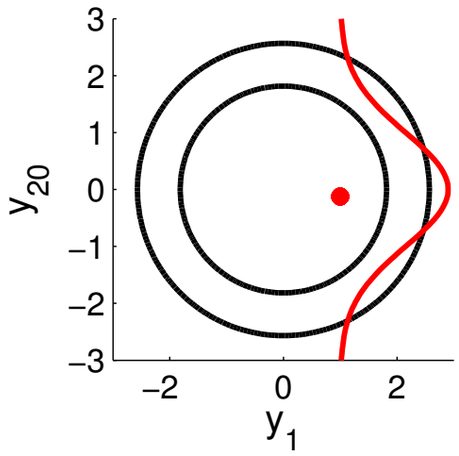


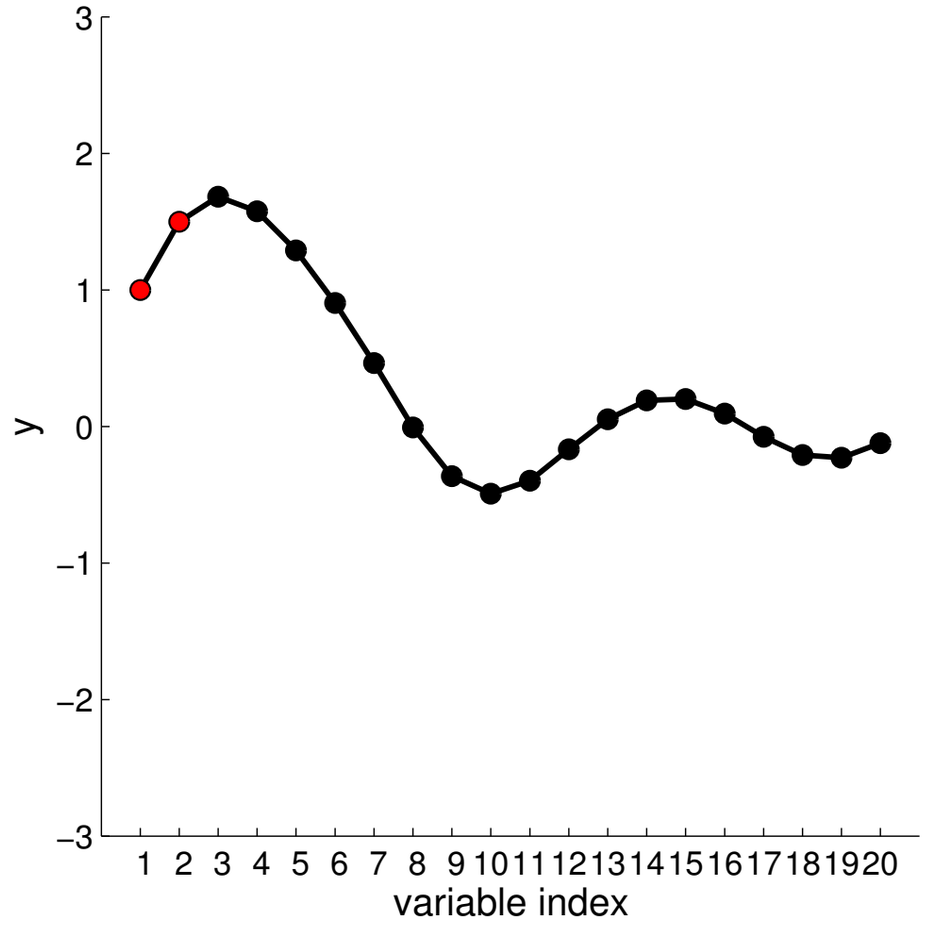
$\Sigma =$



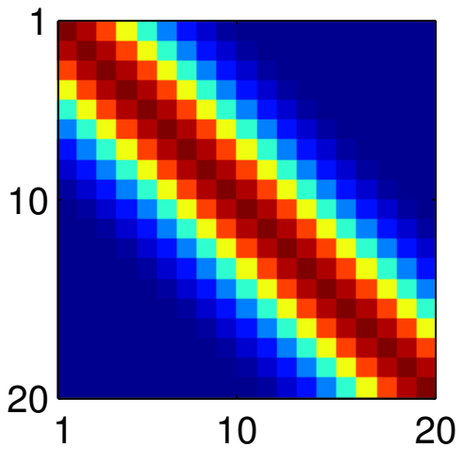
Conditioning on y1 and y2

Special covariance matrix - conditioning



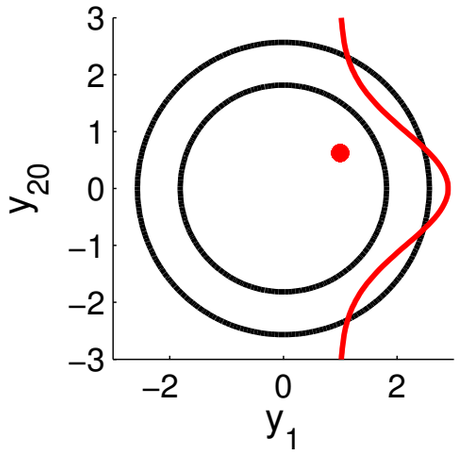


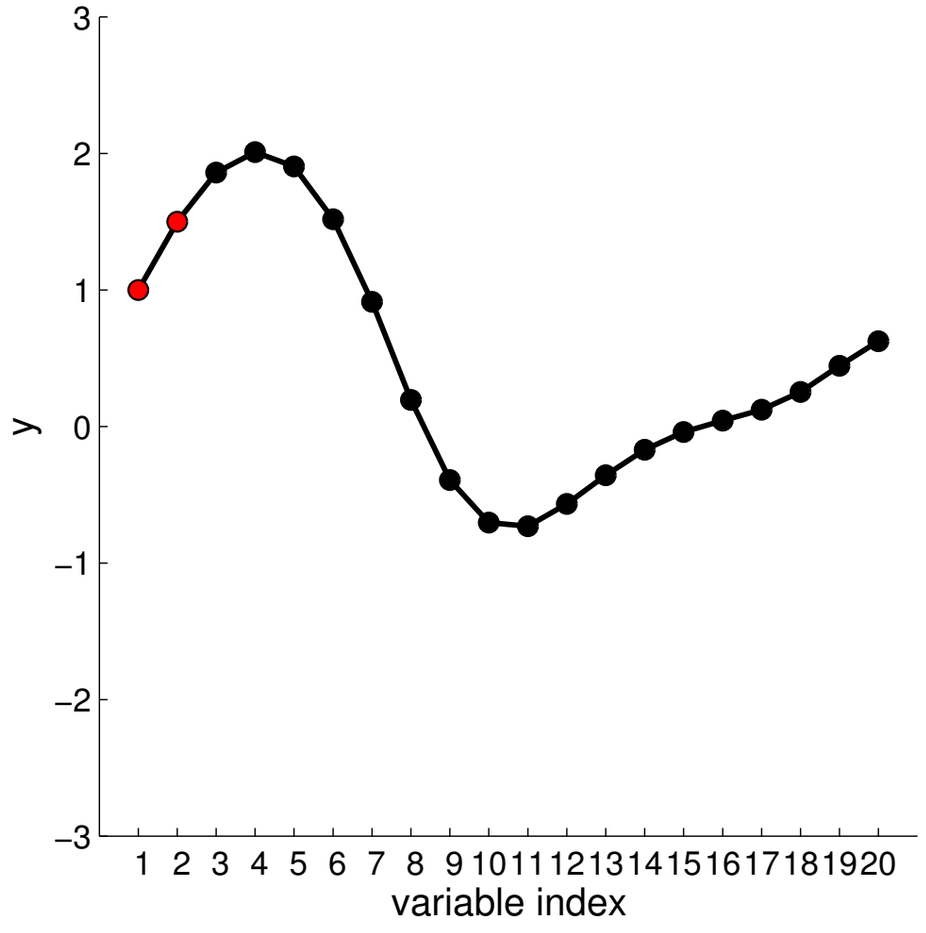
$\Sigma =$



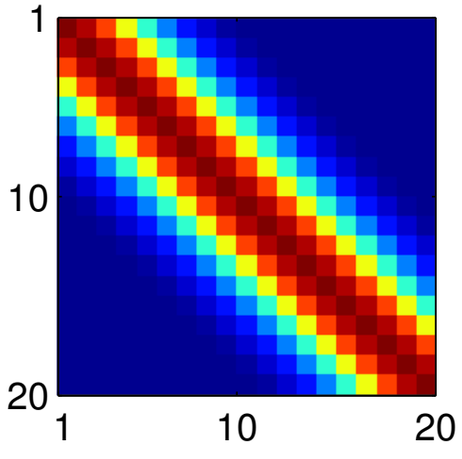
Conditioning on y1 and y2

Special covariance matrix - conditioning



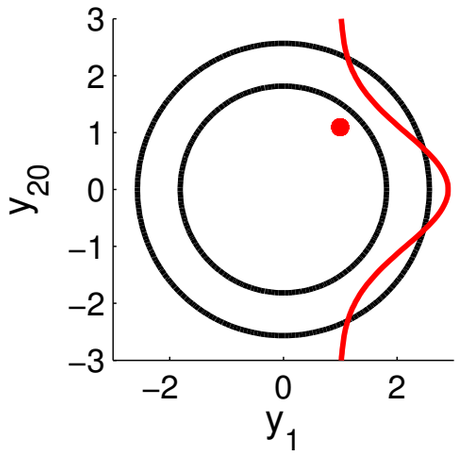


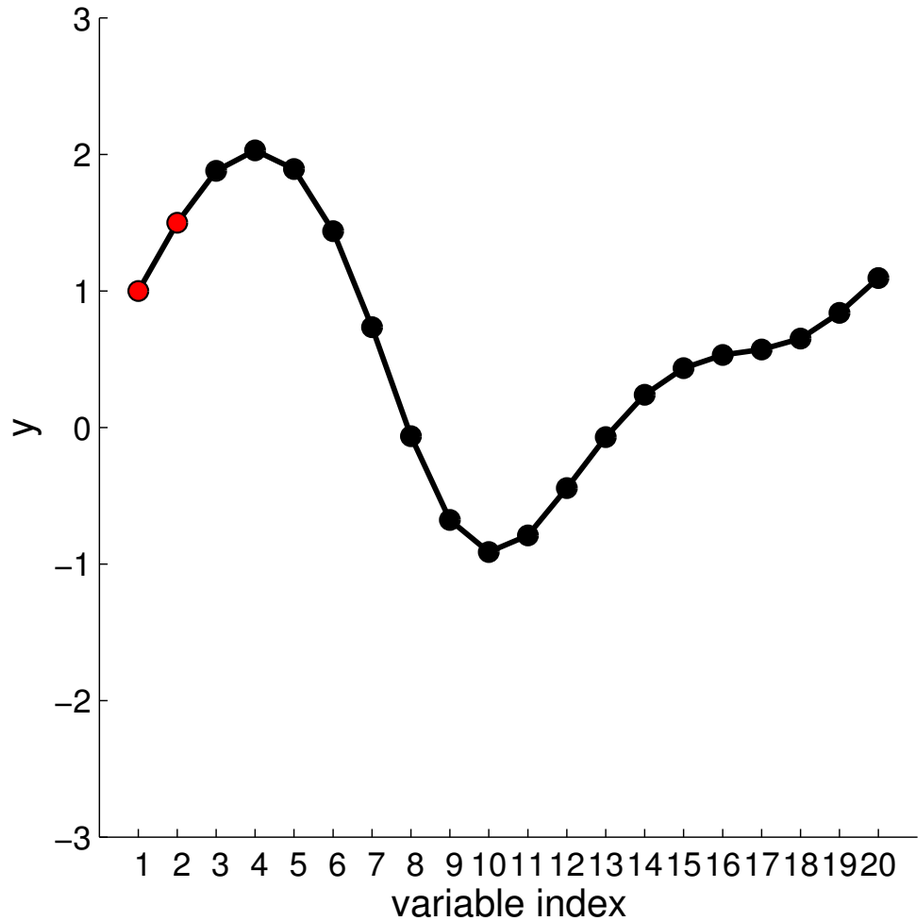
$\Sigma =$



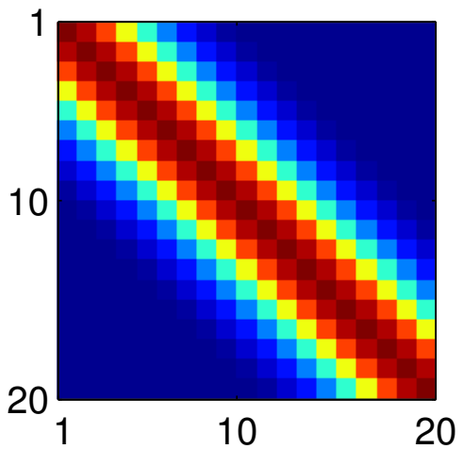
Conditioning on y1 and y2

Special covariance matrix - conditioning



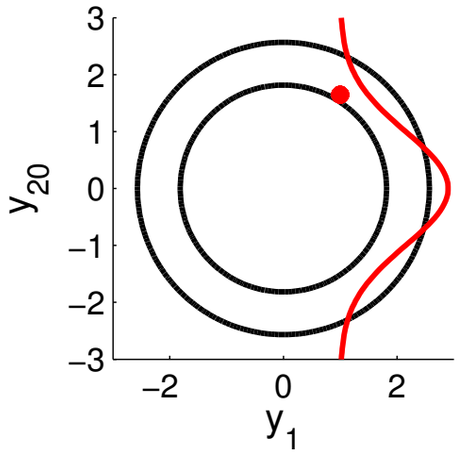


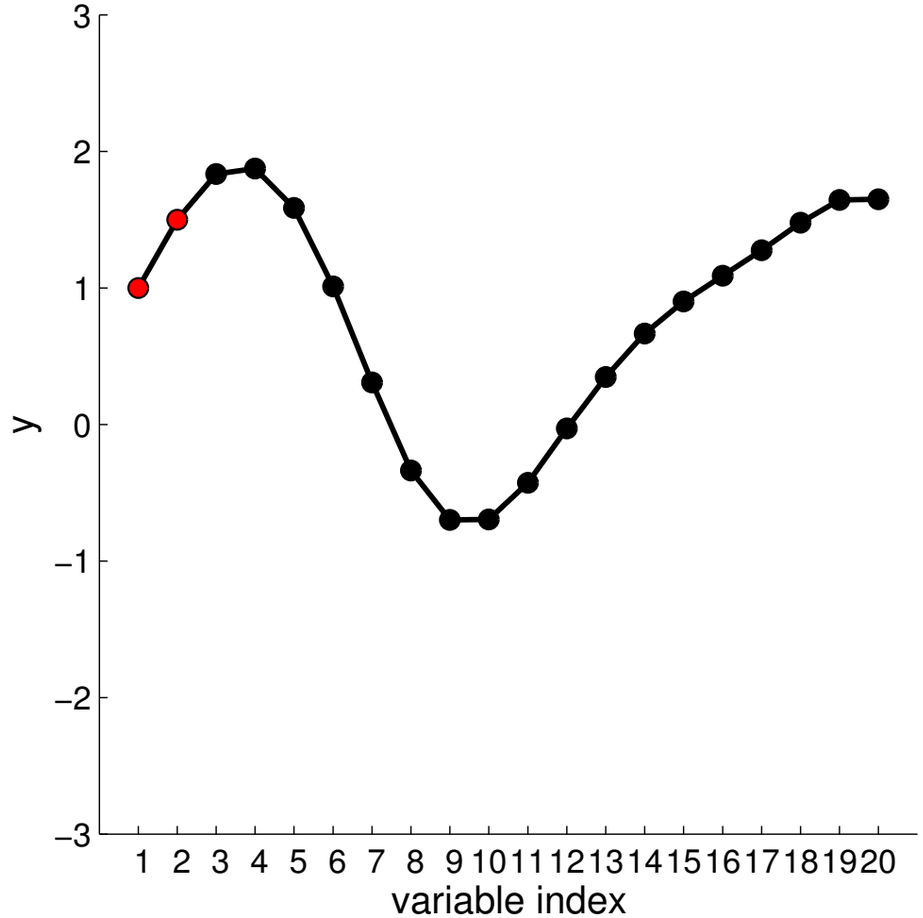
$\Sigma =$



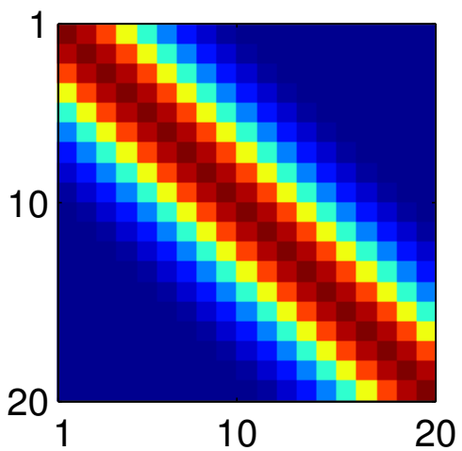
Conditioning on y1 and y2

Special covariance matrix - conditioning



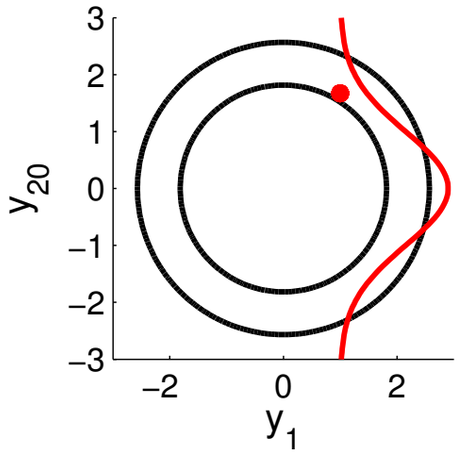


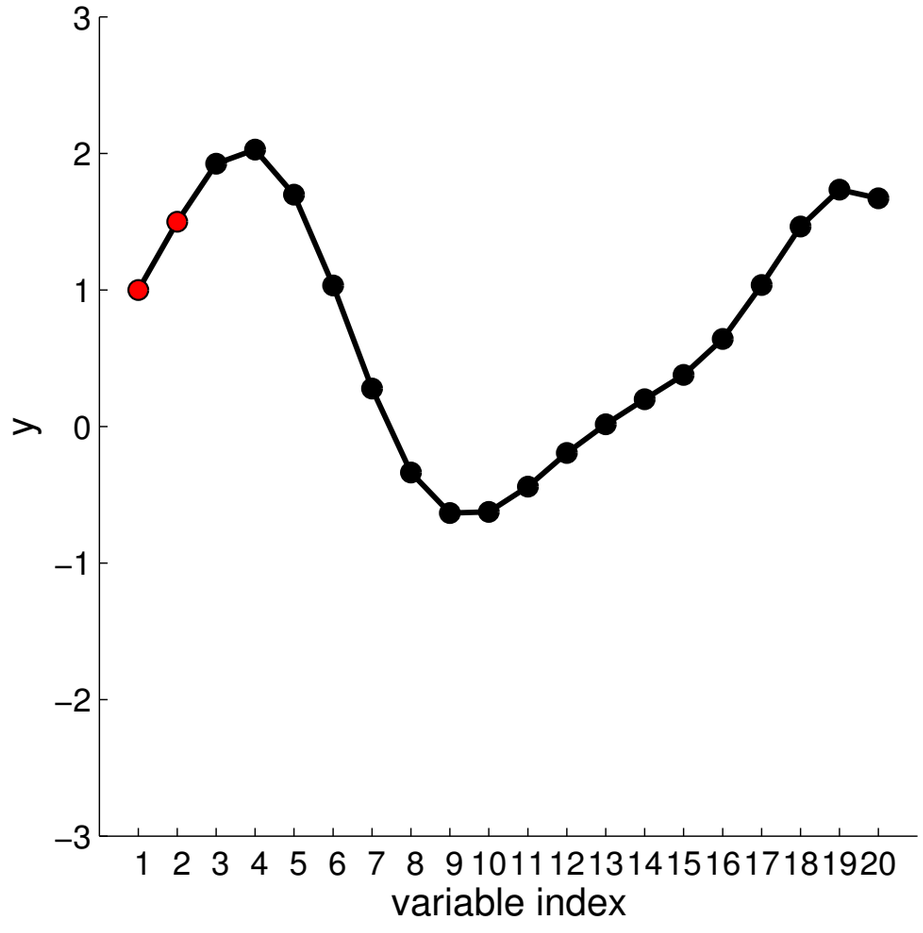
$\Sigma =$



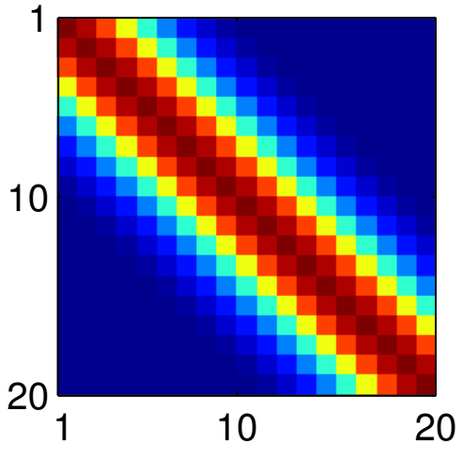
Conditioning on y1 and y2

Special covariance matrix - conditioning



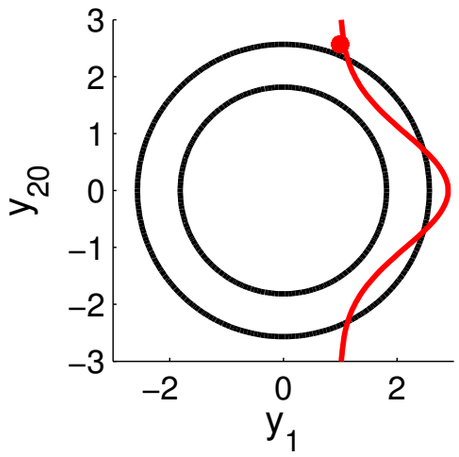


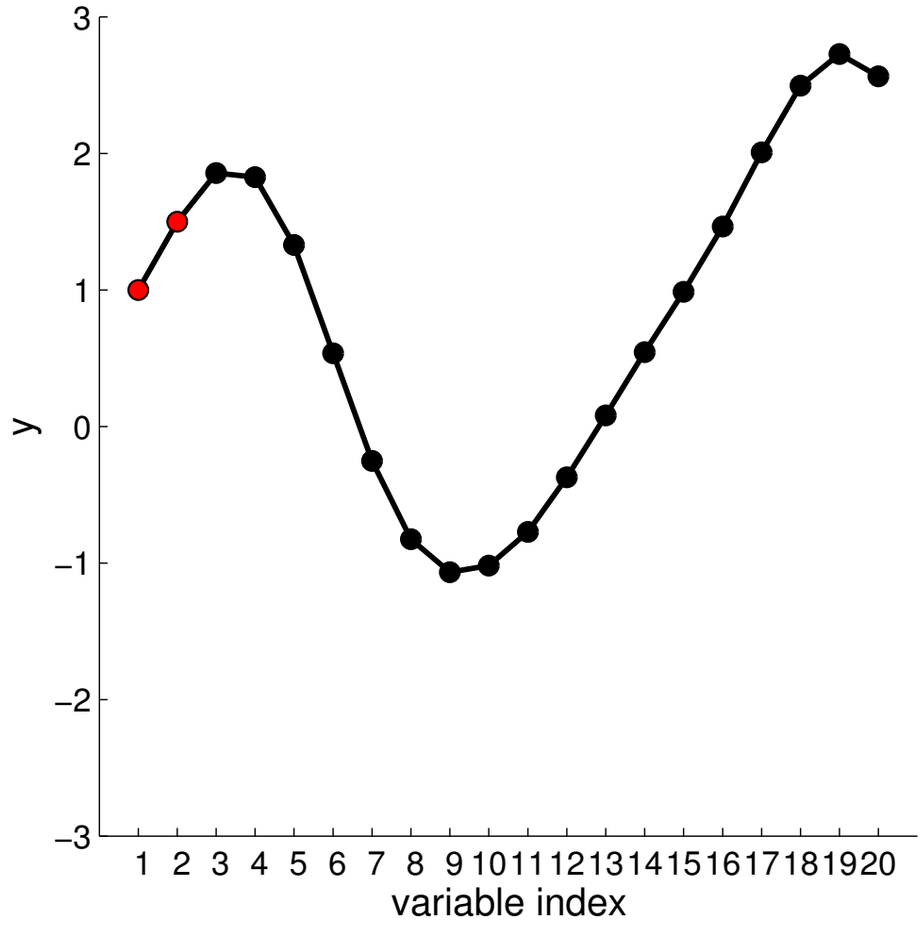
$\Sigma =$



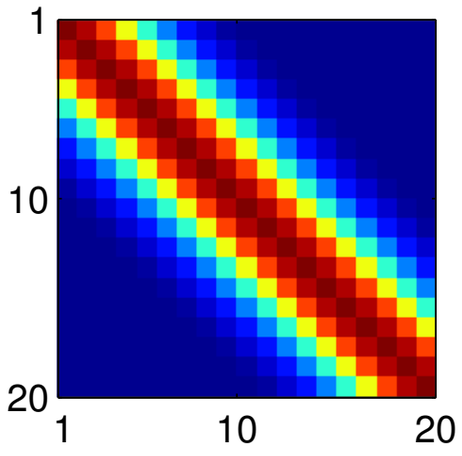
Conditioning on y1 and y2

Special covariance matrix - conditioning



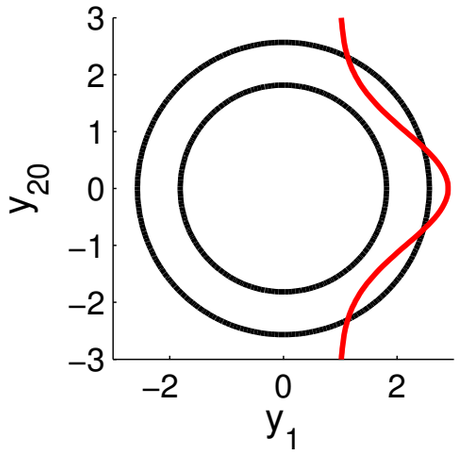


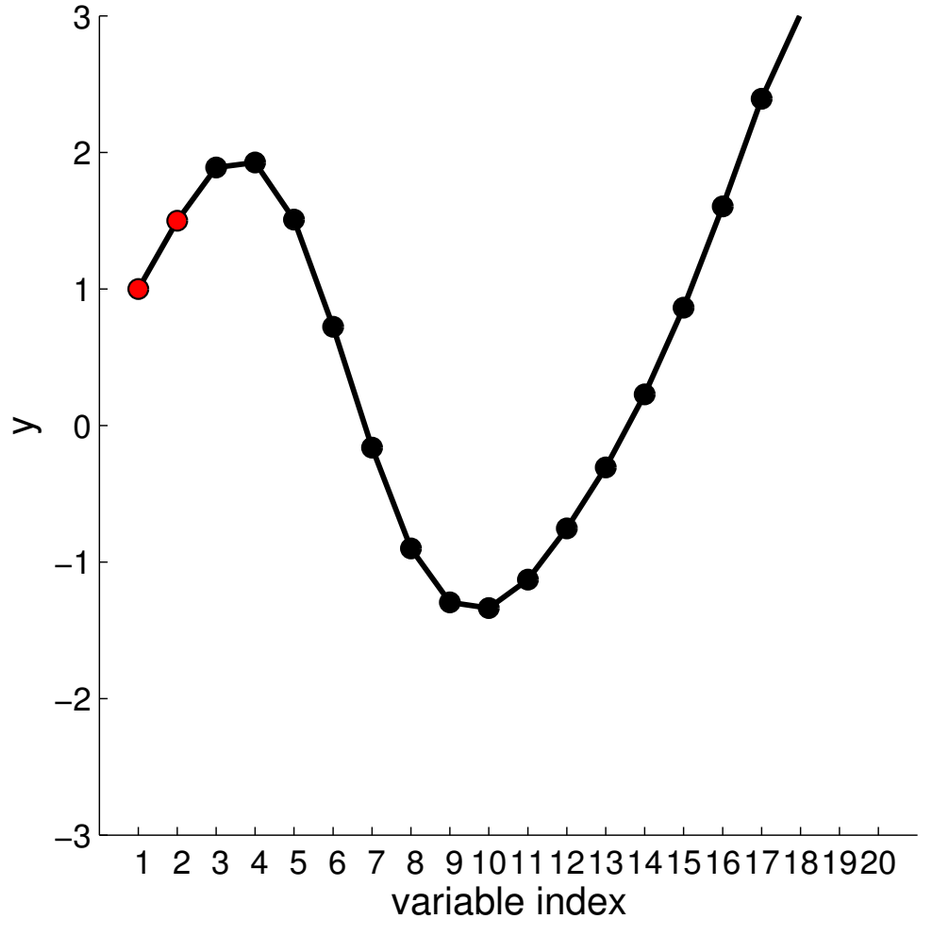
$\Sigma =$



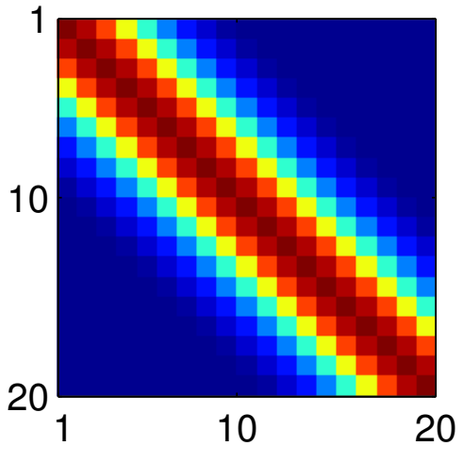
Conditioning on y1 and y2

Special covariance matrix - conditioning





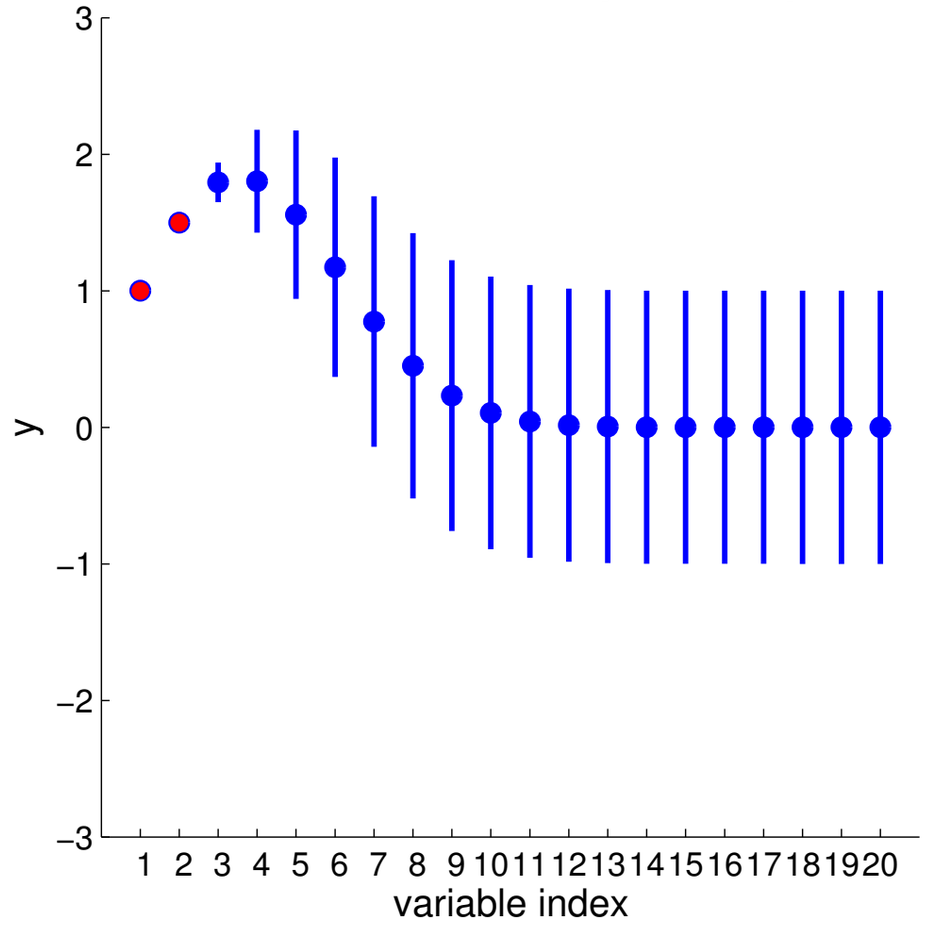
$\Sigma =$



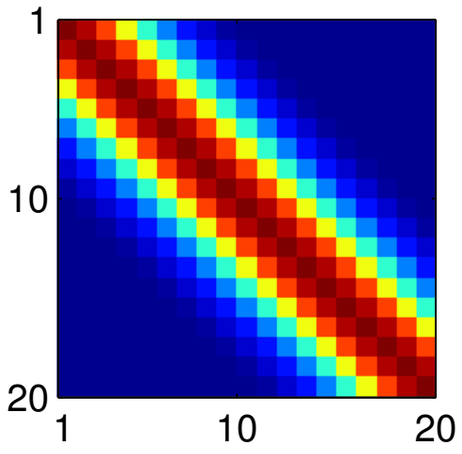
Conditioning on y1 and y2

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.
- How do we compute means and variances? Analytically, using the equations of conditioning!



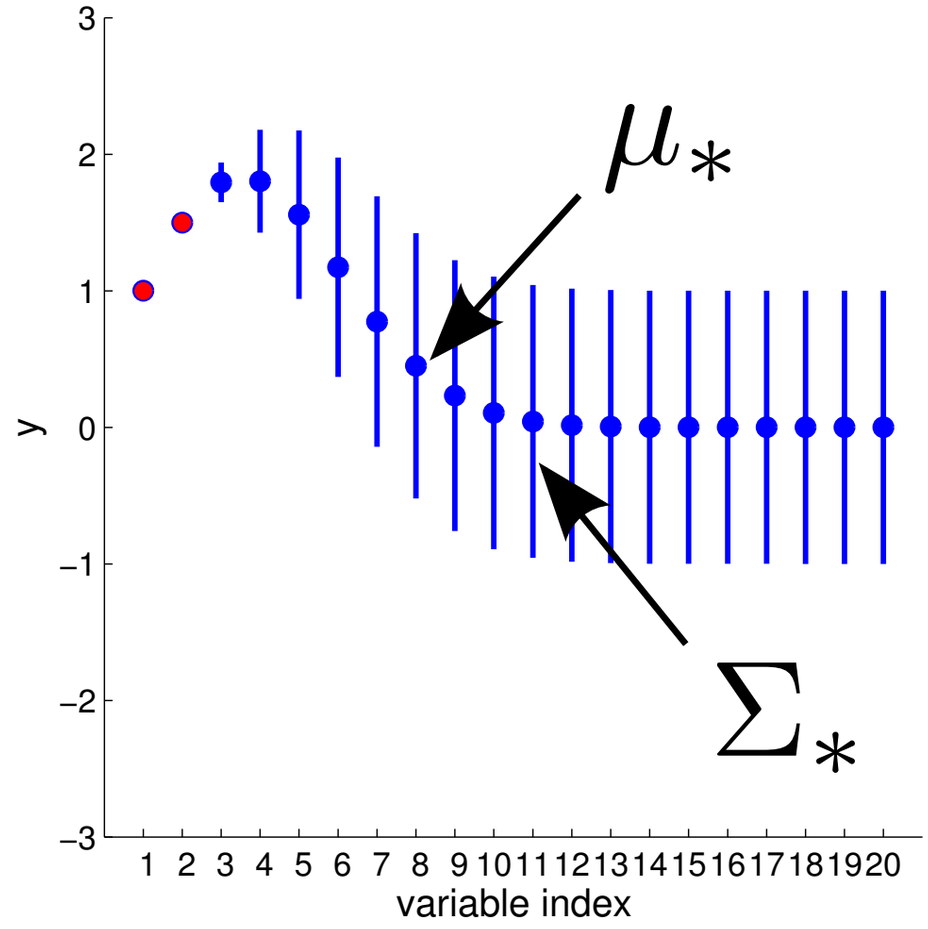
$\Sigma =$



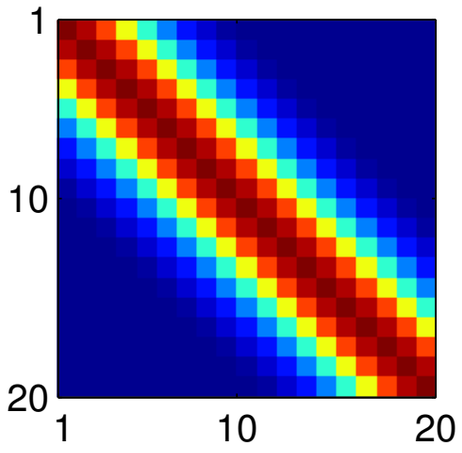
Conditioning on y1 and y2

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.
- How do we compute means and variances? Analytically, using the equations of conditioning!

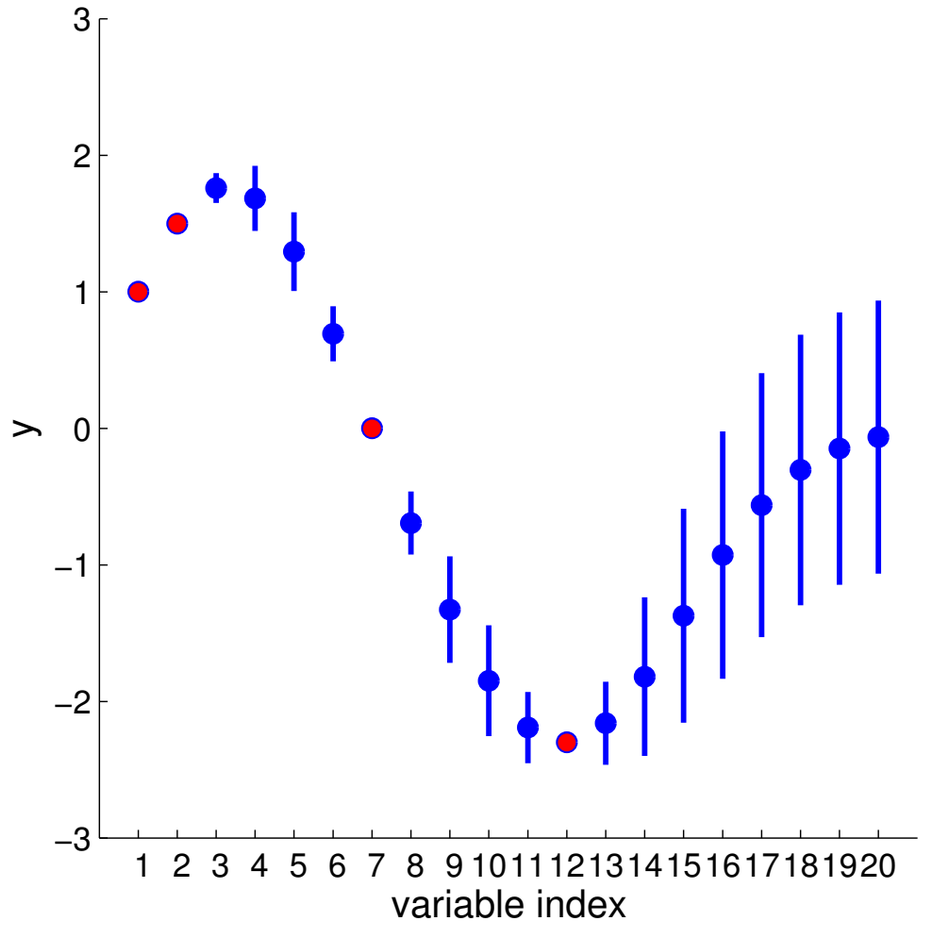


$\Sigma =$

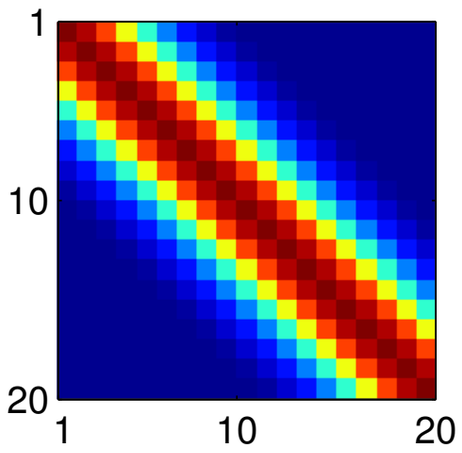


Regression Using Gaussians

We can also condition on non-contiguous indices



$\Sigma =$

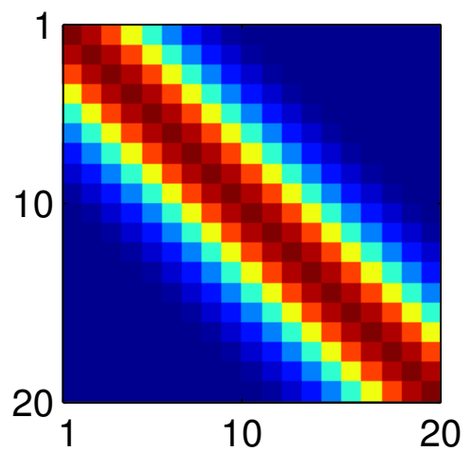


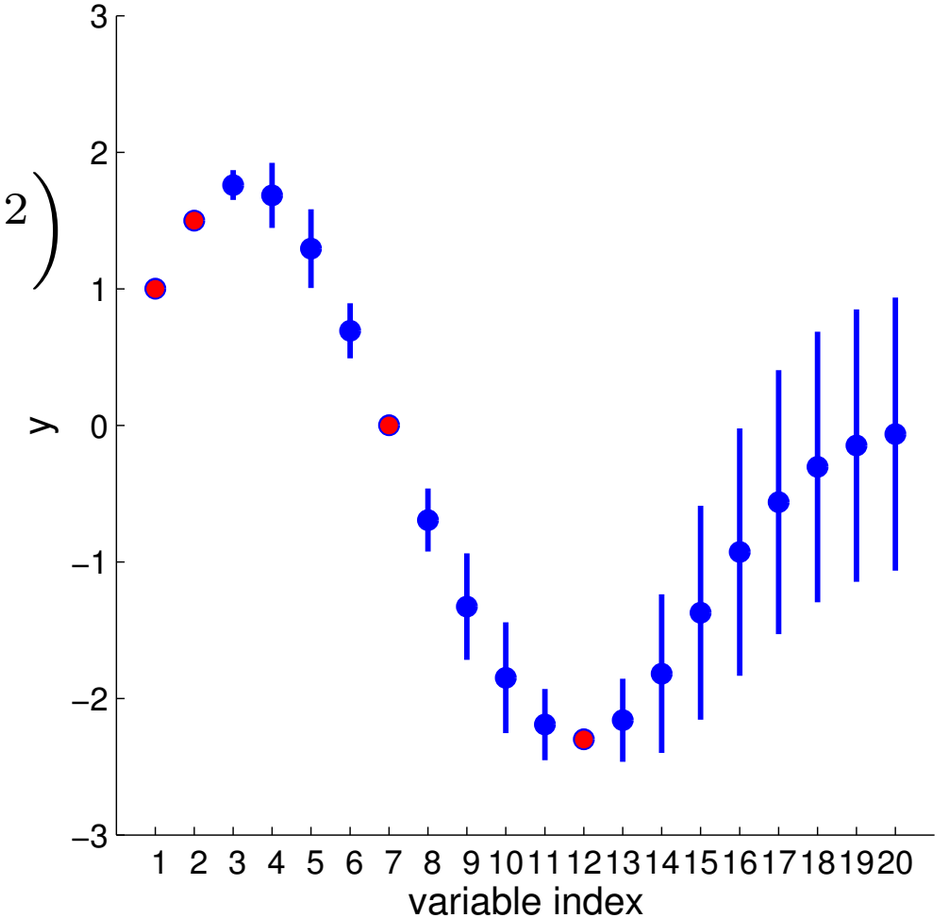
Regression Using Gaussians

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

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

$\Sigma =$





Q: Do x_1, x_2 need to be integers?

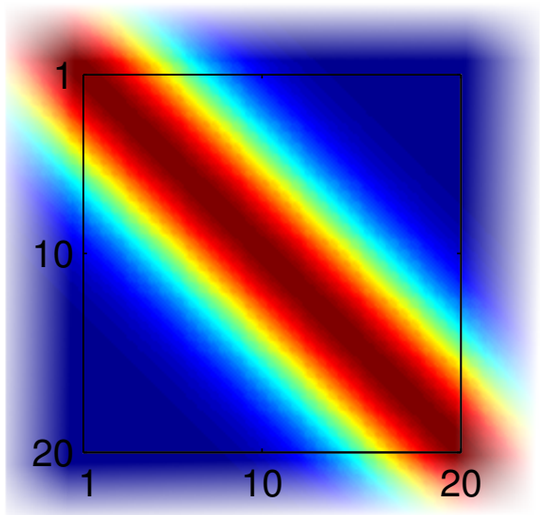
From multivariate Gaussian distributions to Gaussian Processes

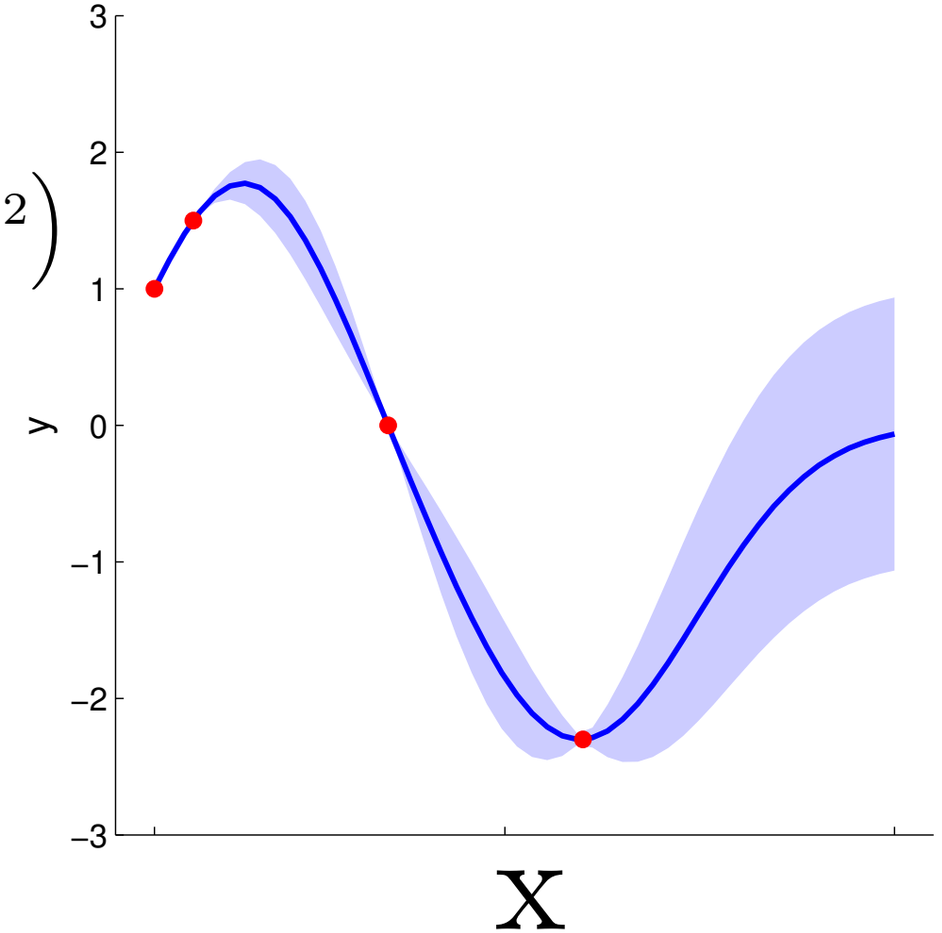
GP: a multivariate Gaussian over an uncountably infinite number of variables with infinite mean vector and infinite times infinite covariance matrix

$$\Sigma(\mathsf{x}_1, \mathsf{x}_2) = \mathsf{K}(\mathsf{x}_1, \mathsf{x}_2) + \mathsf{I}\sigma_{\mathsf{y}}^2$$

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

$\Sigma =$





Mathematical Foundations: Definition

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

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

A Gaussian distribution is fully specified by a mean vector, $\boldsymbol{\mu}$, and covariance matrix Σ:

$$\mathbf{f} = (\mathsf{f}_1, \ldots, \mathsf{f}_n) \sim \mathcal{N}(\mu, \Sigma), \quad \text{indices} \quad i = 1, \ldots, n$$

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

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

Mathematical Foundations: Regression

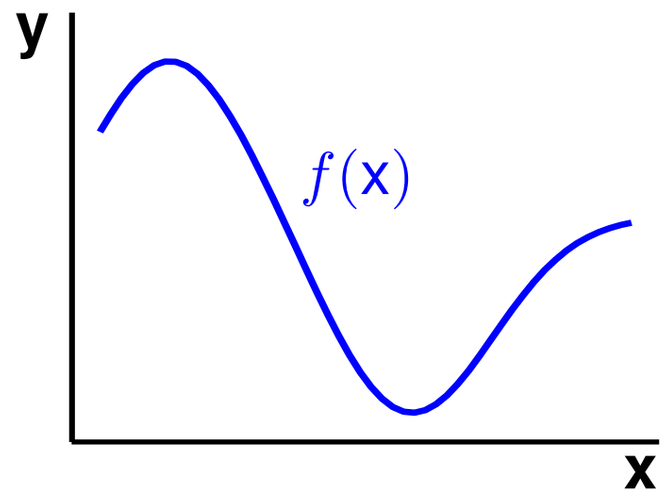
Q3. What's the formal justification for how we were using GPs for regression?

Mathematical Foundations: Regression

Q3. What's the formal justification for how we were using GPs for regression?

Generative model (like non-linear regression)

$$\mathsf{y}(\mathsf{x}) = f(\mathsf{x}) + \epsilon\sigma_{\mathsf{y}}$$



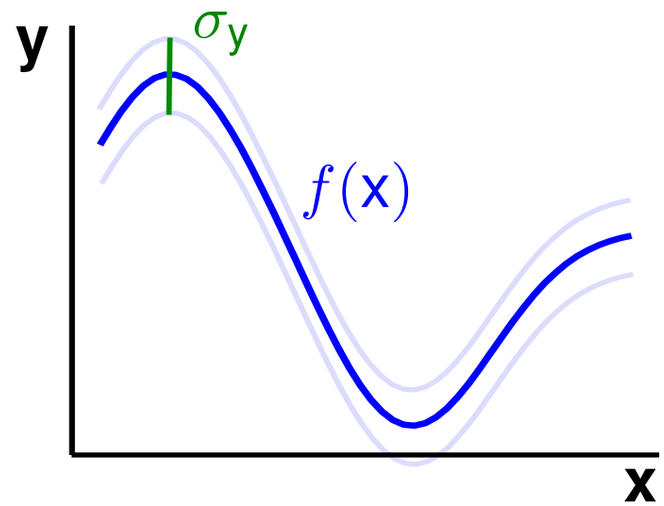
Mathematical Foundations: Regression

Q3. What's the formal justification for how we were using GPs for regression?

Generative model (like non-linear regression)

$$\mathsf{y}(\mathsf{x}) = f(\mathsf{x}) + \epsilon\sigma_{\mathsf{y}}$$

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



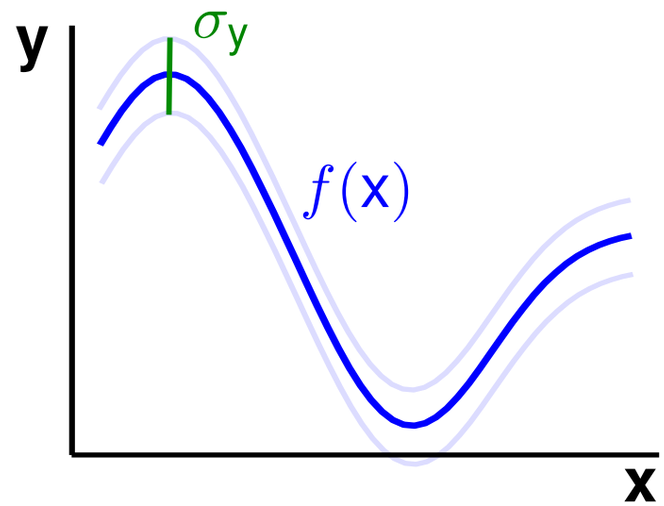
Mathematical Foundations: Regression

Q3. What's the formal justification for how we were using GPs for regression?

Generative model (like non-linear regression)

$$\mathsf{y}(\mathsf{x}) = f(\mathsf{x}) + \epsilon\sigma_{\mathsf{y}}$$

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



place GP prior over the non-linear function

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

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

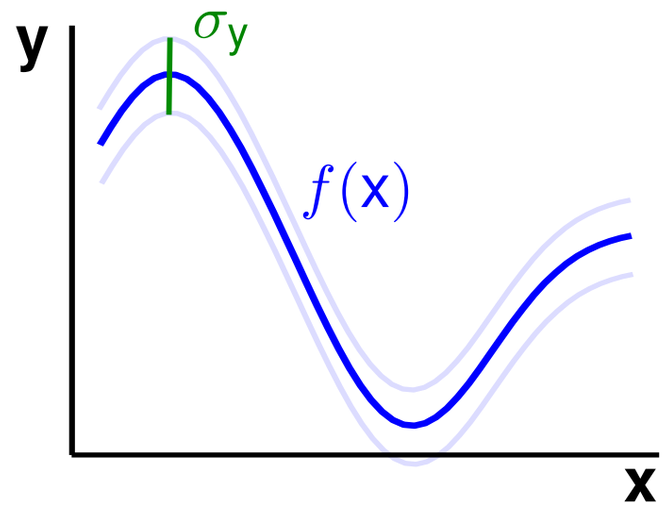
Mathematical Foundations: Regression

Q3. What's the formal justification for how we were using GPs for regression?

Generative model (like non-linear regression)

$$\mathsf{y}(\mathsf{x}) = f(\mathsf{x}) + \epsilon\sigma_{\mathsf{y}}$$

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



place GP prior over the non-linear function

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

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

since the sum of two Gaussians is a Gaussian, the model induces a GP over $\mathsf{y}(\mathsf{x})$

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

GP: A distribution over functions

A GP is a Gaussian distribution over functions:

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$$

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}\left[f(\mathbf{x})\right] \\ \kappa(\mathbf{x}, \mathbf{x}') &= \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))^T\right] \end{aligned}$$

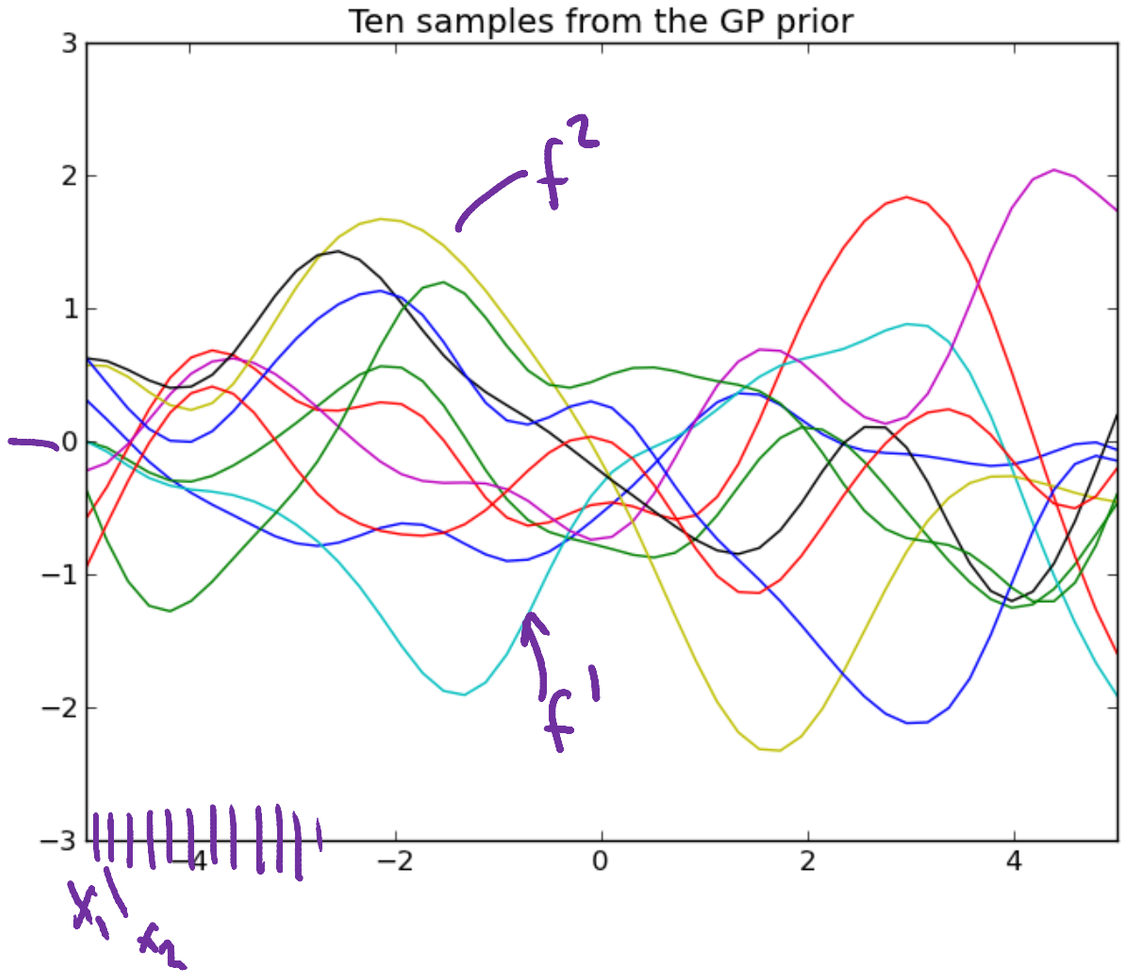
$$k(x, x') = \exp(-\tfrac{1}{2}(x - x')^2)$$

Create $x_{1:N}$

Create $\mu = O_N, \quad K$

$$K = LL^T$$

$$\begin{aligned} f^i &\sim N(O_N, K) \\ &\sim N(o, I)L \end{aligned}$$



Sampling from $P(f)$

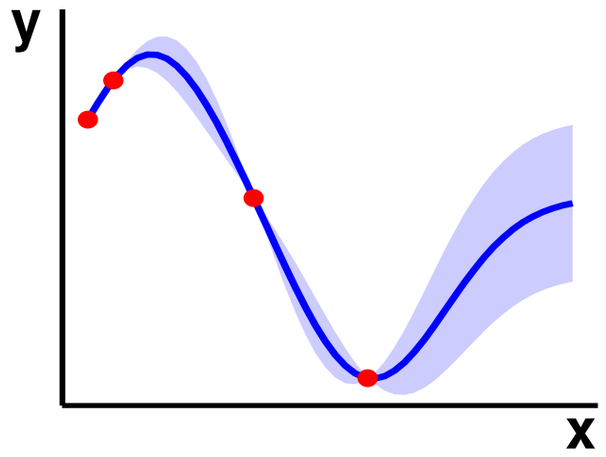
```
from __future__ import division
import numpy as np
import matplotlib.pyplot as pl

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.
Xtest = np.linspace(-5, 5, n).reshape(-1,1)  # Test points.
K_ = kernel(Xtest, Xtest)               # Kernel at test points.

# draw samples from the prior at our test points.
L = np.linalg.cholesky(K_ + 1e-6*np.eye(n))
f_prior = np.dot(L, np.random.normal(size=(n,10)))

pl.plot(Xtest, f_prior)
```

$K = LL^T$

$f = LN(0, I)$

Mathematical Foundations: Prediction

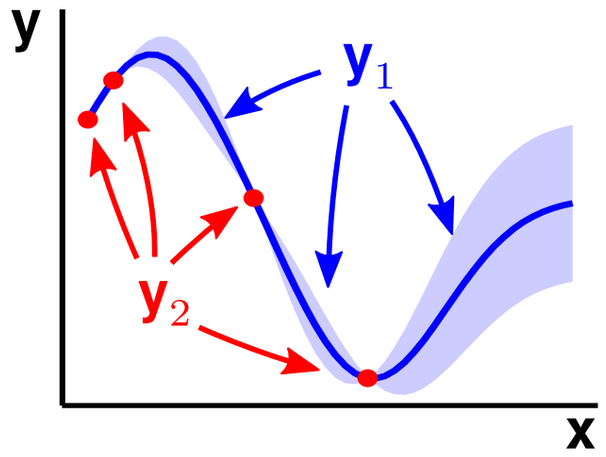
Q4. How do we make predictions?



Mathematical Foundations: Prediction

Q4. How do we make predictions?

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

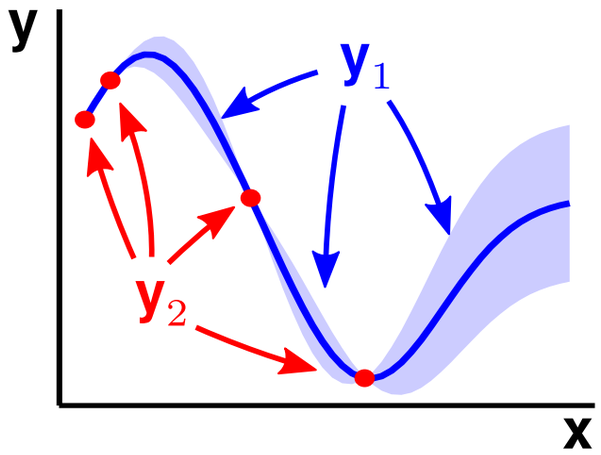


Mathematical Foundations: Prediction

Q4. How do we make predictions?

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

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

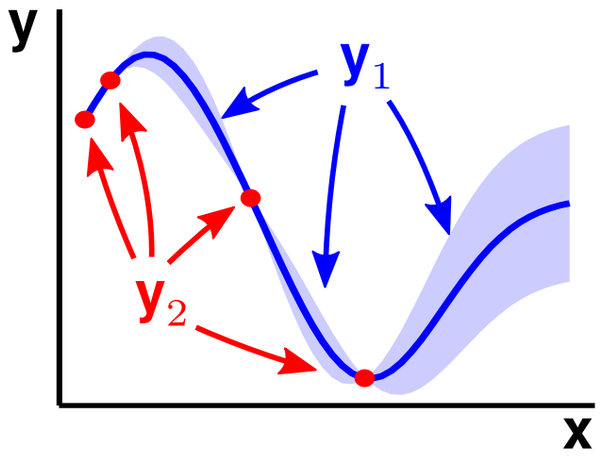


Mathematical Foundations: Prediction

Q4. How do we make predictions?

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

$$p(\mathbf{y}_1|\mathbf{y}_2) = \frac{p(\mathbf{y}_1, \mathbf{y}_2)}{p(\mathbf{y}_2)} \qquad p(\mathbf{y}_2) = \mathcal{N}(\mathbf{b}, \mathsf{C})$$



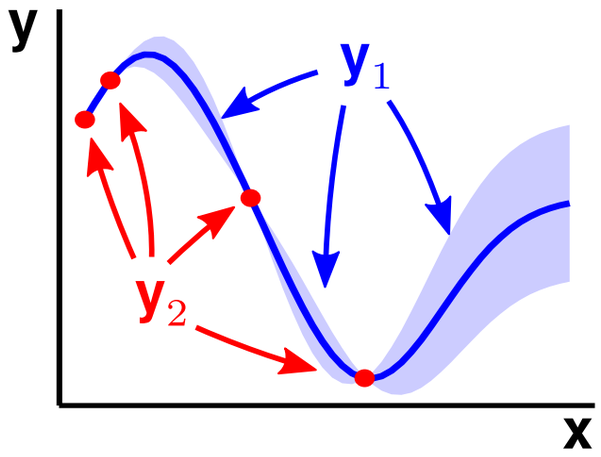
Mathematical Foundations: Prediction

Q4. How do we make predictions?

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

$$p(\mathbf{y}_1|\mathbf{y}_2) = \frac{p(\mathbf{y}_1, \mathbf{y}_2)}{p(\mathbf{y}_2)} \qquad p(\mathbf{y}_2) = \mathcal{N}(\mathbf{b}, \mathsf{C})$$

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

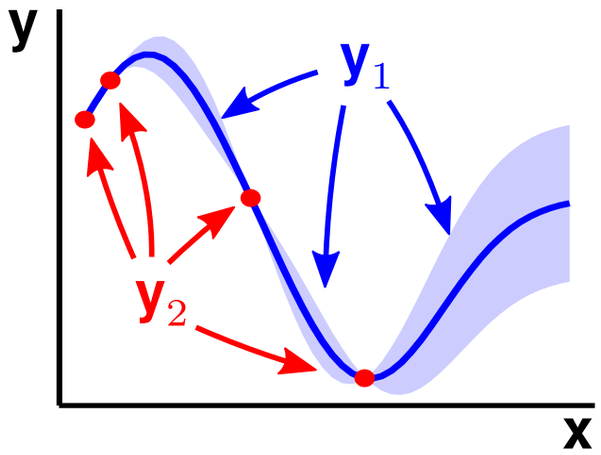


Mathematical Foundations: Prediction

Q4. How do we make predictions?

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

$$p(\mathbf{y}_1|\mathbf{y}_2) = \frac{p(\mathbf{y}_1, \mathbf{y}_2)}{p(\mathbf{y}_2)} \qquad p(\mathbf{y}_2) = \mathcal{N}(\mathbf{b}, \mathsf{C})$$



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

predictive mean

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

predictive covariance

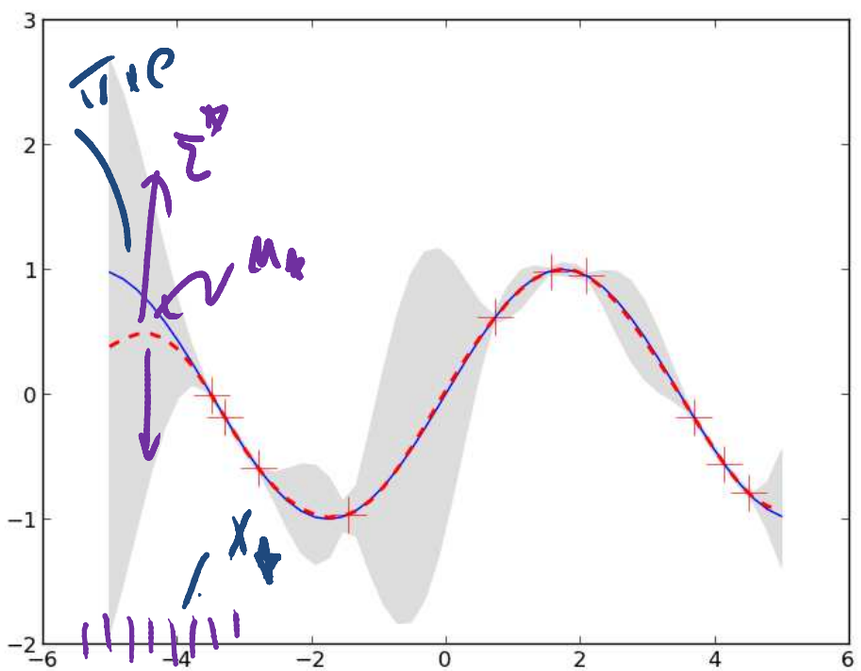
$$\Sigma_{\mathbf{y}_1|\mathbf{y}_2} = \mathsf{A} - \mathsf{B}\mathsf{C}^{-1}\mathsf{B}^\intercal$$

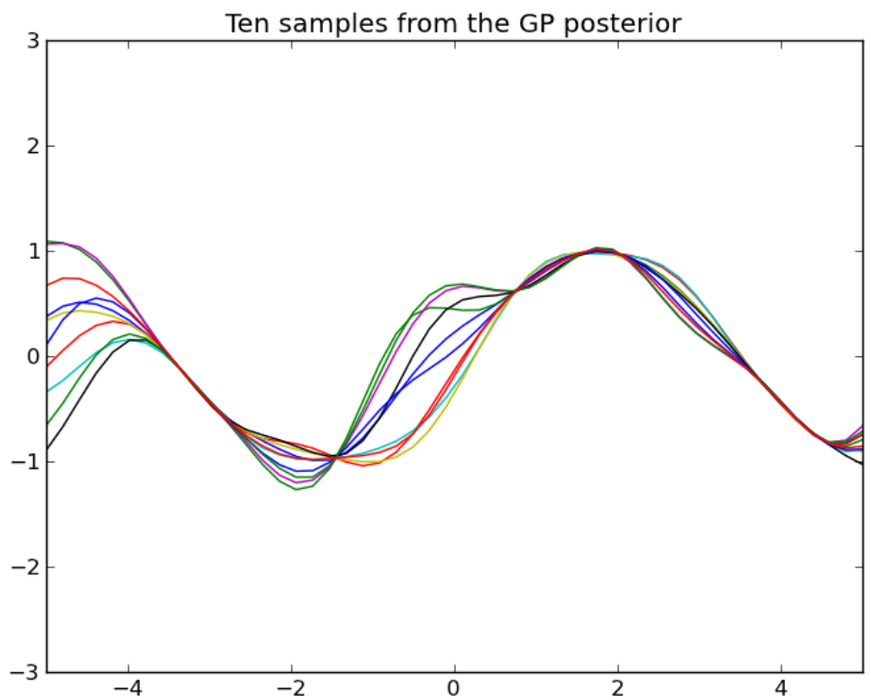
Predictive uncertainty = prior uncertainty - reduction in uncertainty

Noiseless GP Regression

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

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





Noisy GP Regression

noisy

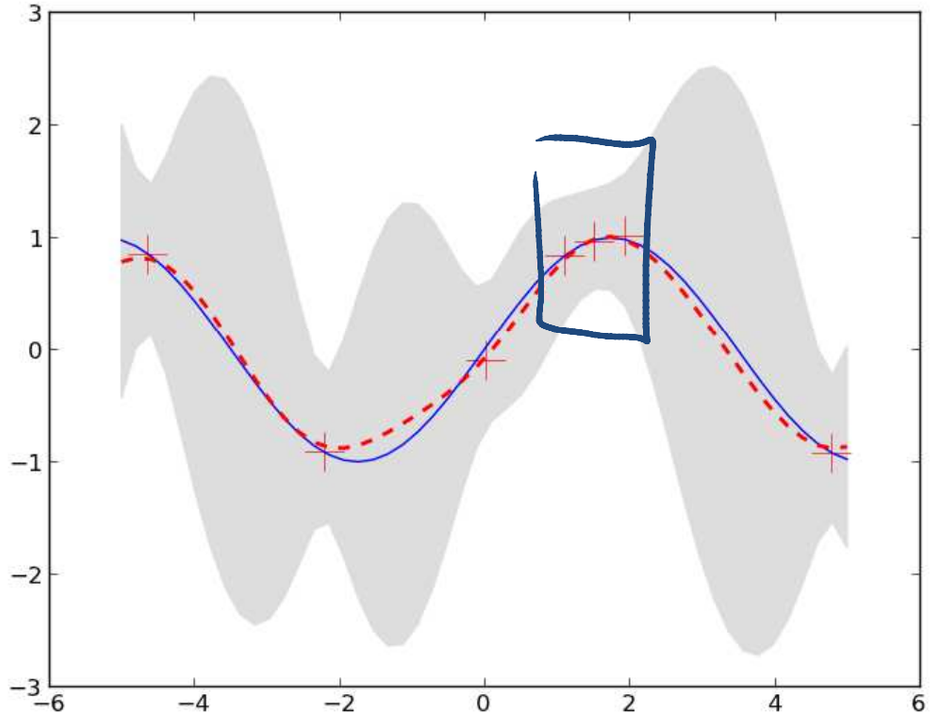
$$y = f(\mathbf{x}) + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma_y^2)$$

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X}) p(\mathbf{f}|\mathbf{X}) d\mathbf{f}$$

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$$

$$p(\mathbf{y}|\mathbf{f}) = \prod_i \mathcal{N}(y_i|f_i, \sigma_y^2)$$

$$\text{cov}[\mathbf{y}|\mathbf{X}] = \mathbf{K} + \sigma_y^2 \mathbf{I}_N \triangleq \mathbf{K}_y$$



$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_y & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right)$$

thm

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

Numerical Computations Considerations

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

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

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

Algorithm 15.1: GP regression

1 $\mathbf{L} = \text{cholesky}(\mathbf{K} + \sigma_y^2 \mathbf{I})$;

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

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

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

5 $\text{var}[f_*] = \kappa(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{v}^T \mathbf{v}$;

6 $\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^T \boldsymbol{\alpha} - \sum_i \log L_{ii} - \frac{N}{2}\log(2\pi)$

Bayesian Optimization

- Model f of the function I am trying to maximize (GPs for that)
- Aqcuisition function that takes as input the GP posteriror and suggests where to sample next
- We will see two acquisition functions:
  - UCB
  - Tompson sampling

Exploration-Exploitation Tradeoff

Recall the expressions for GP prediction:

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

Prediction Data Test

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

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

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

We could balance this tradeoff with an acquisition function as follows:

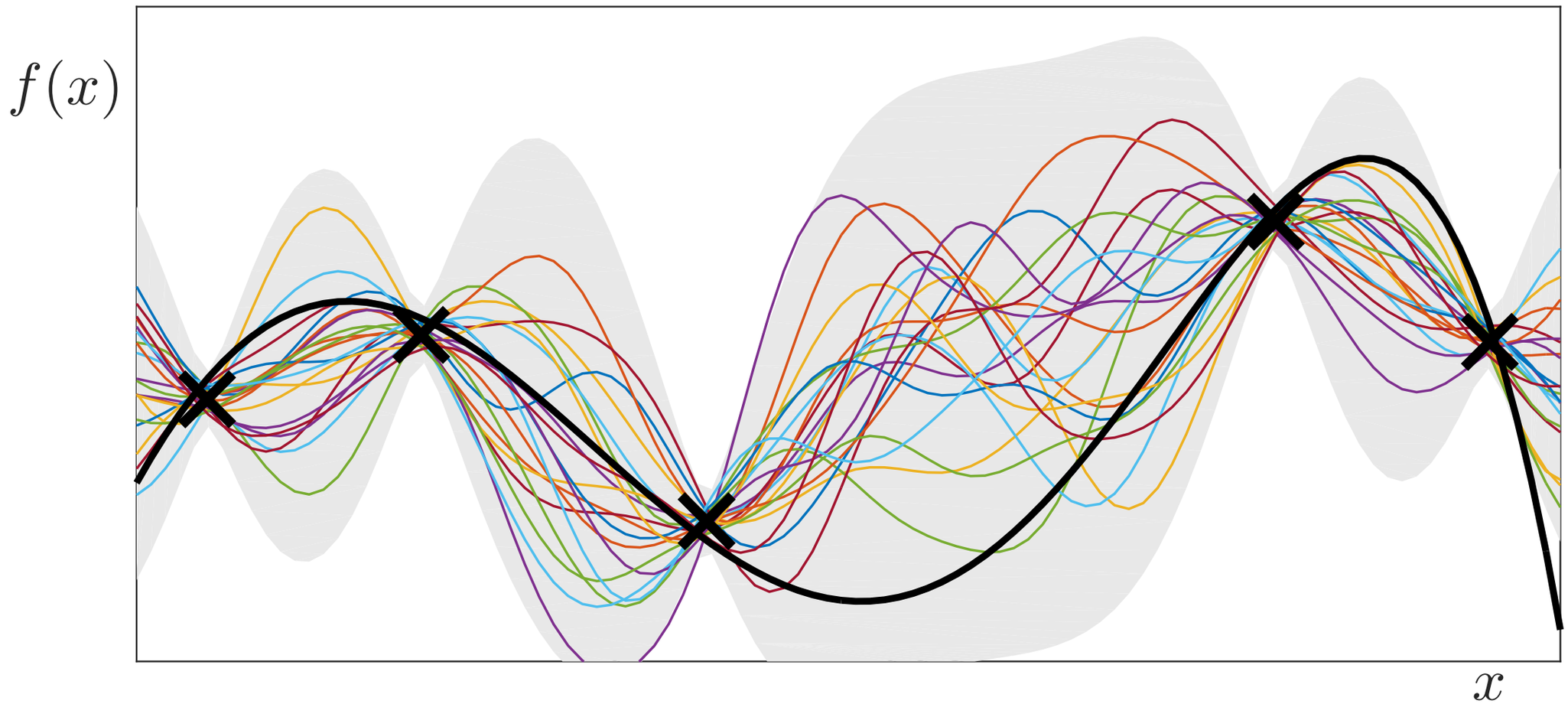
$$\mu(\mathbf{x}) + \kappa\sigma(\mathbf{x})$$

Algorithm 1: UCB in GP Bandits

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

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



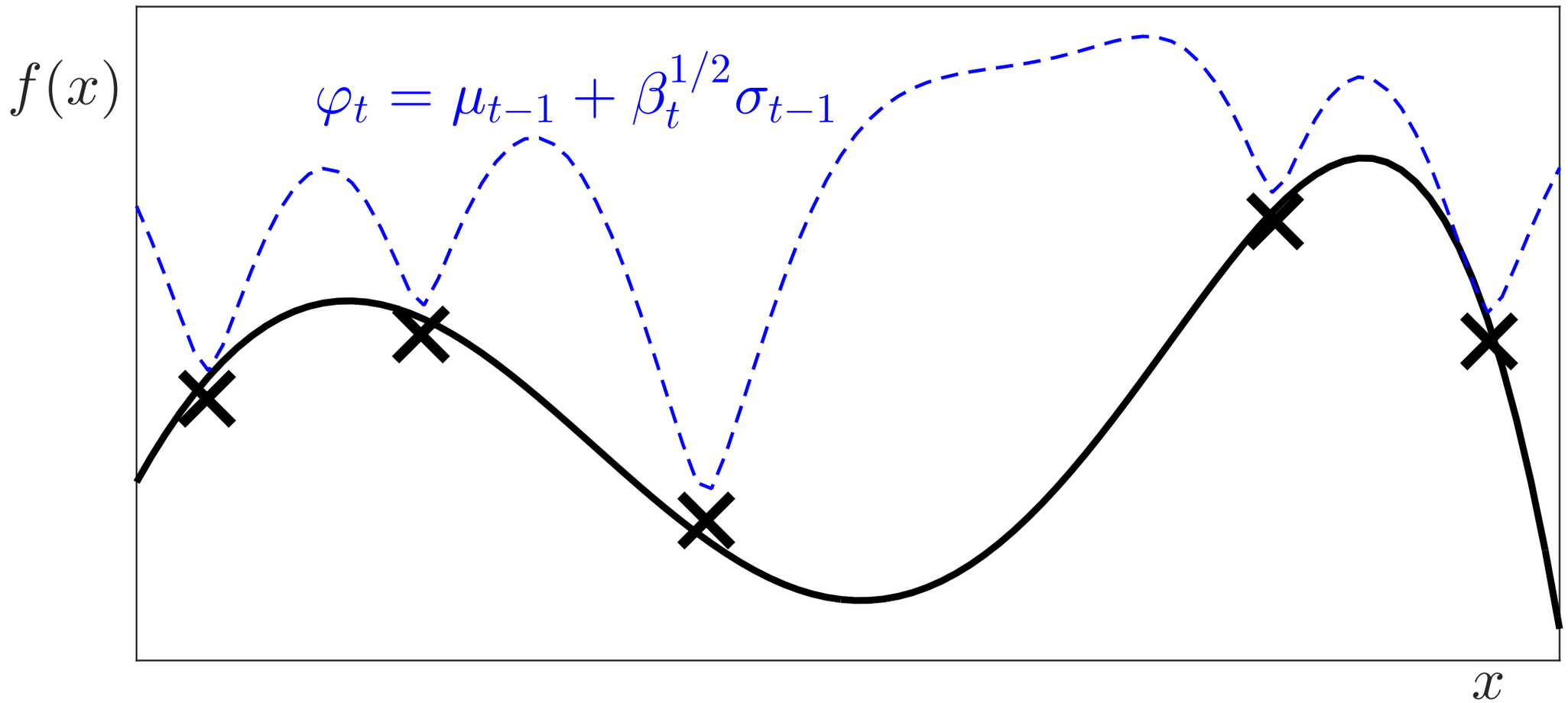
1) Compute posterior $\mathcal{GP}$.

Algorithm 1: UCB in GP Bandits

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

Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



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

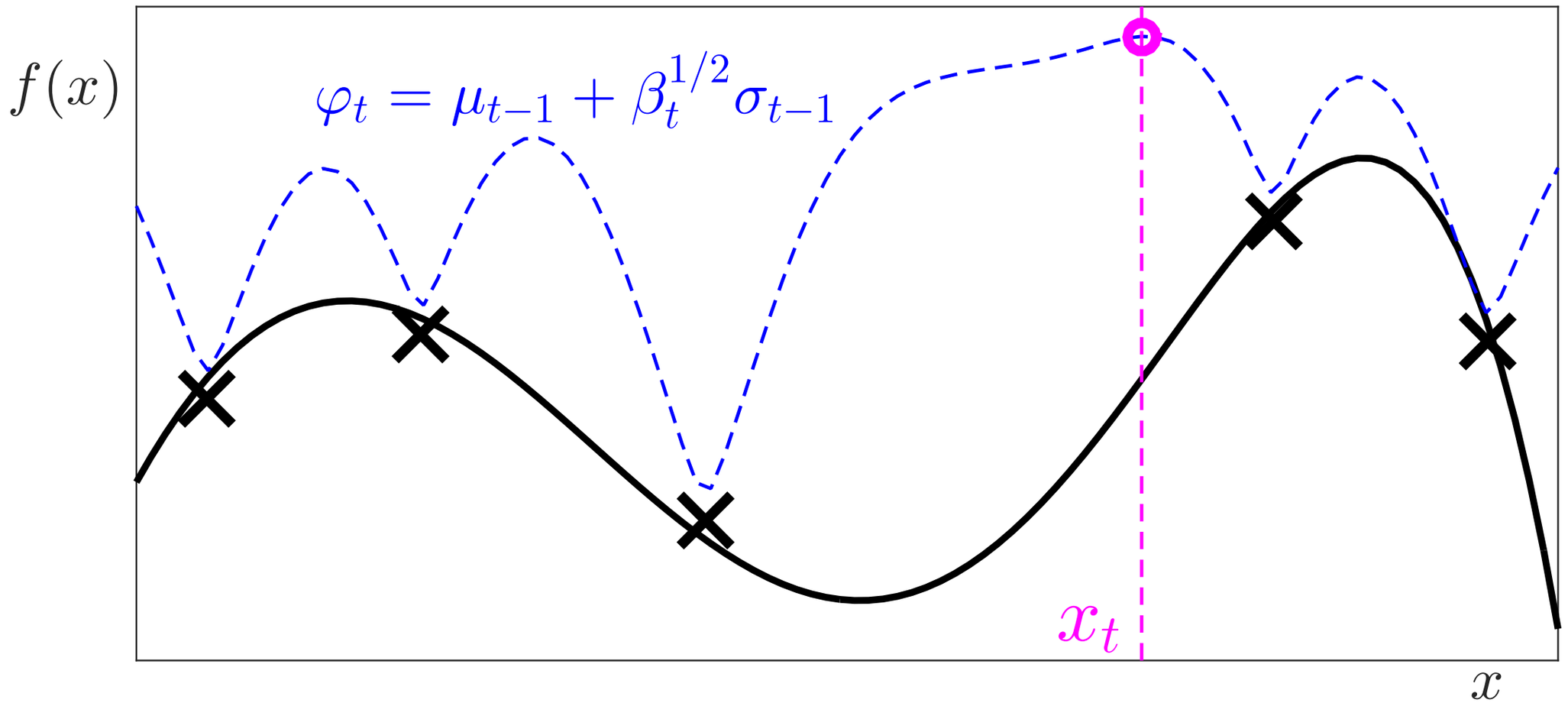
2) Construct UCB φ_t.

Algorithm 1: UCB in GP Bandits

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 φ_t.

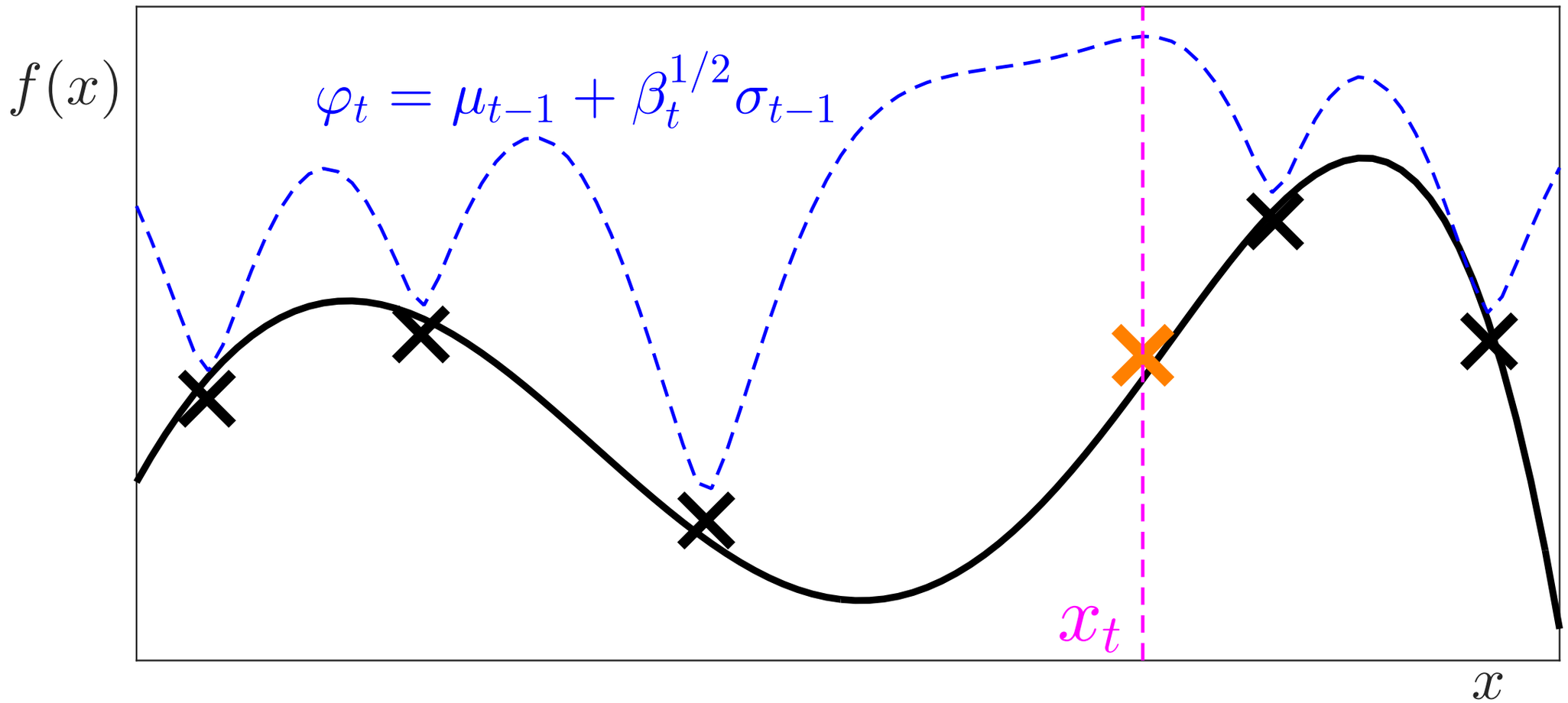
3) Choose $x_t = \mathrm{argmax}_x\, \varphi_t(x)$.

Algorithm 1: UCB in GP Bandits

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

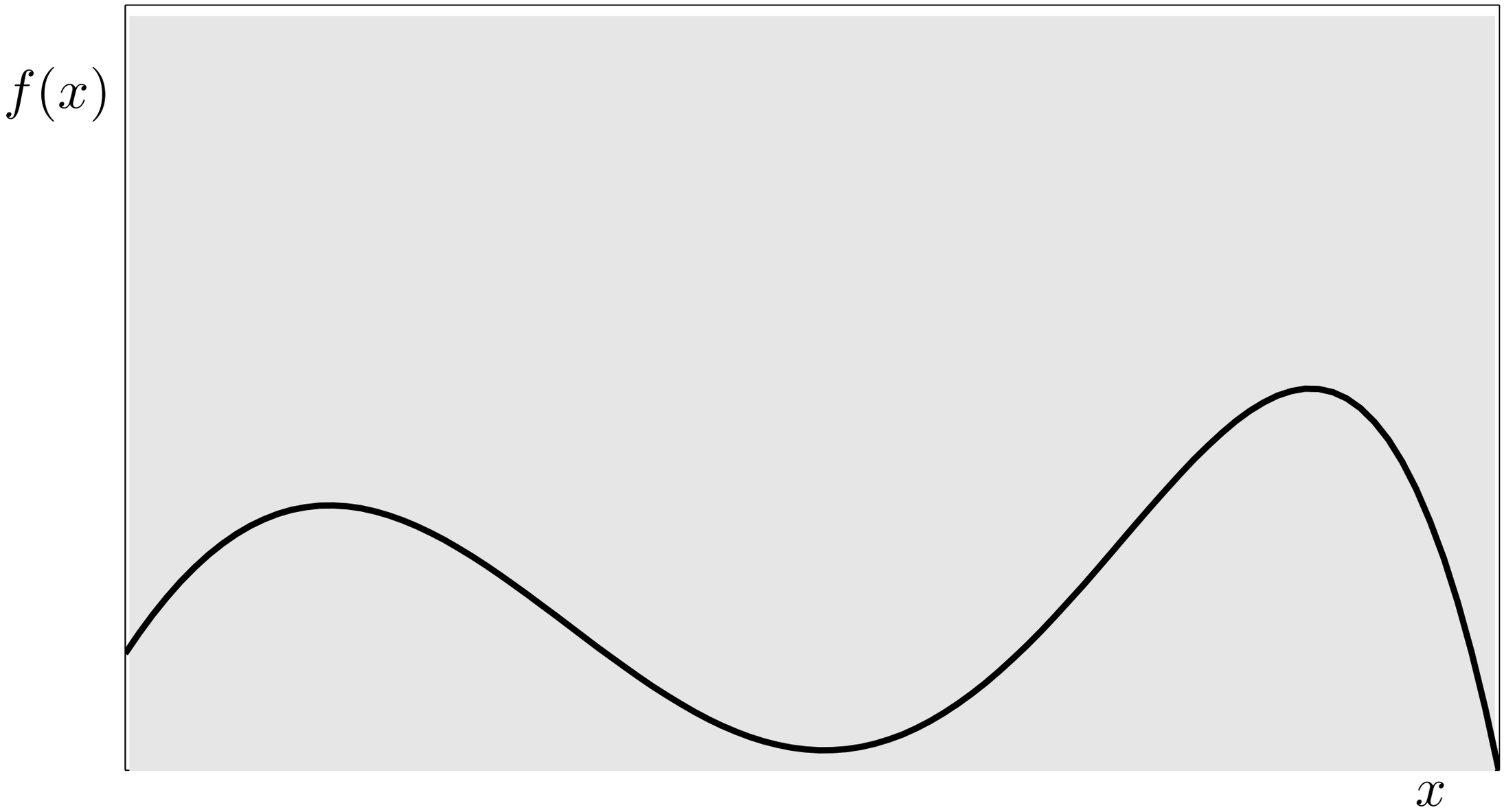
Gaussian Process Upper Confidence Bound (GP-UCB)

(Srinivas et al. 2010)



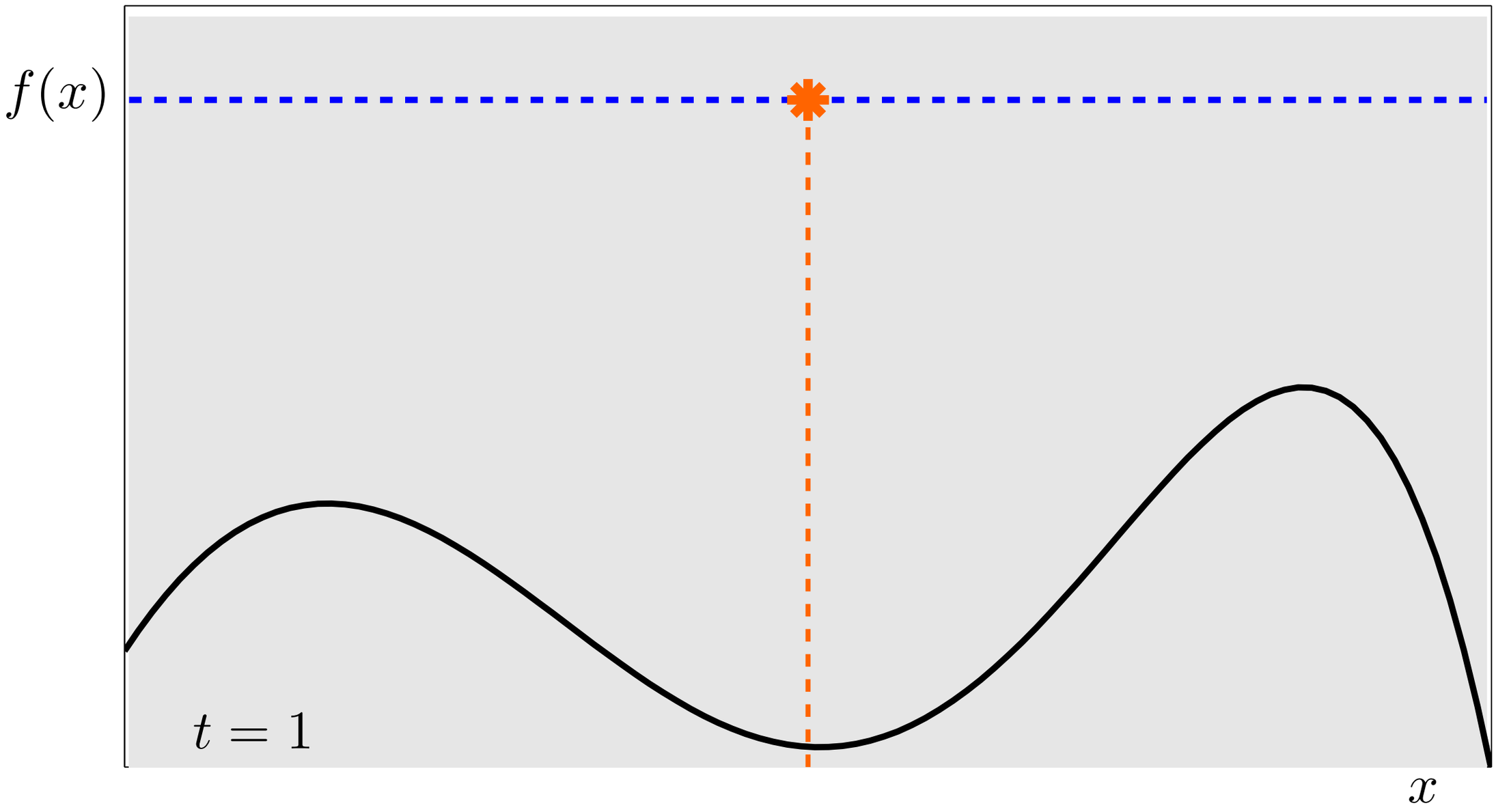
1) Compute posterior $\mathcal{GP}$.
2) Construct UCB φ_t.
3) Choose $x_t = \mathrm{argmax}_x\, \varphi_t(x)$.
4) Evaluate f at x_t.

GP-UCB



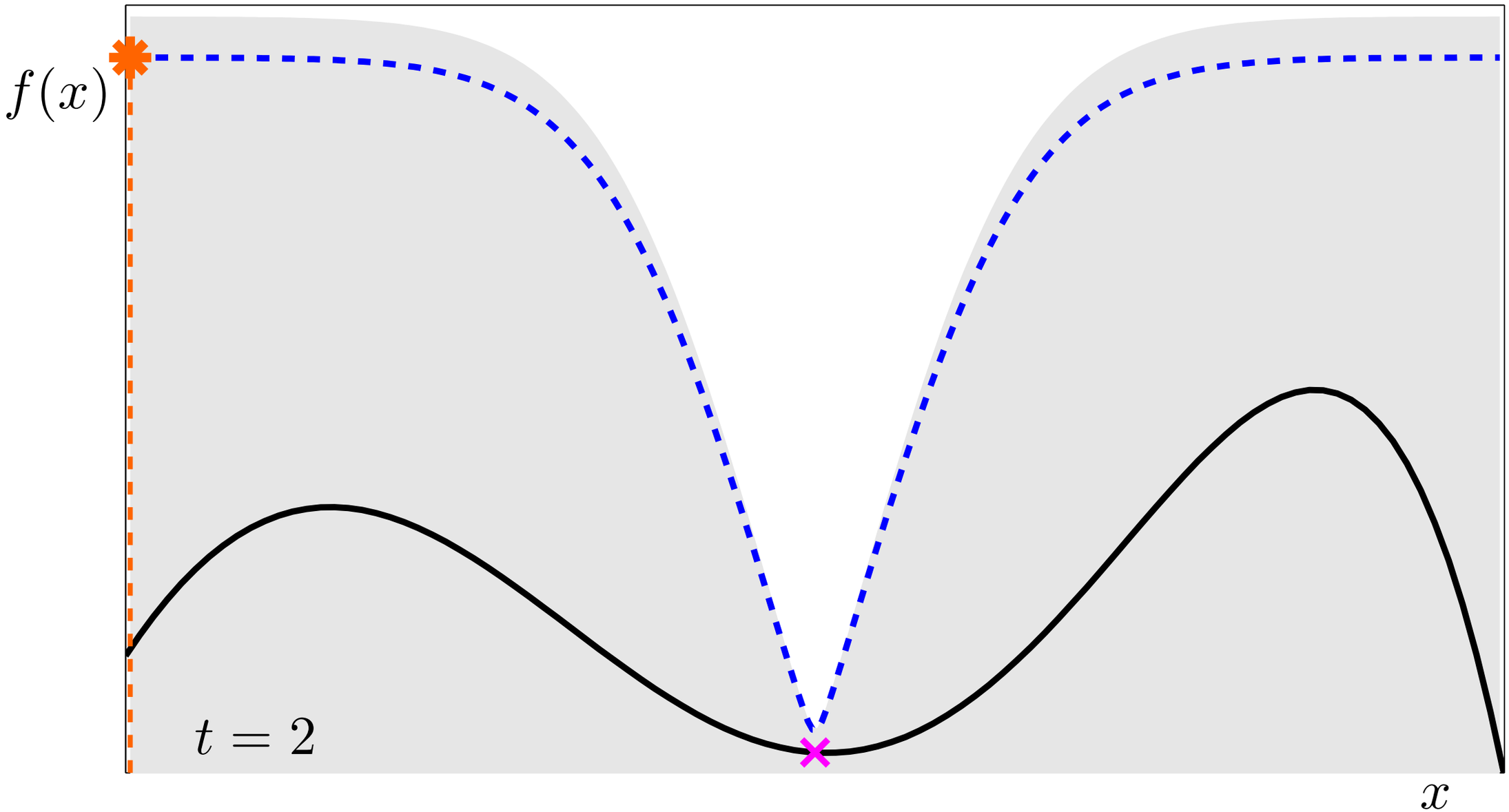
Srinivas et al. 2010

GP-UCB



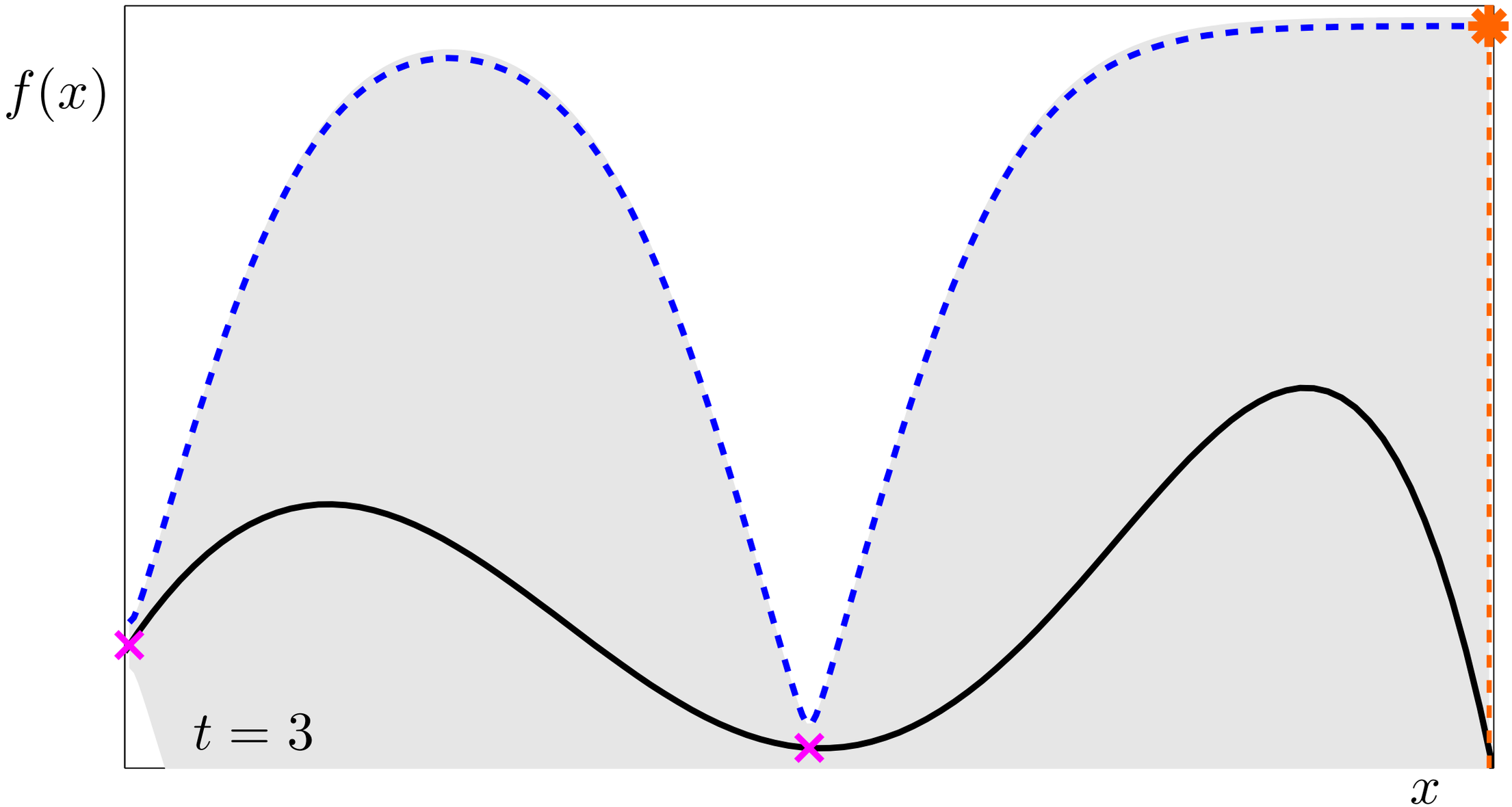
Srinivas et al. 2010

GP-UCB



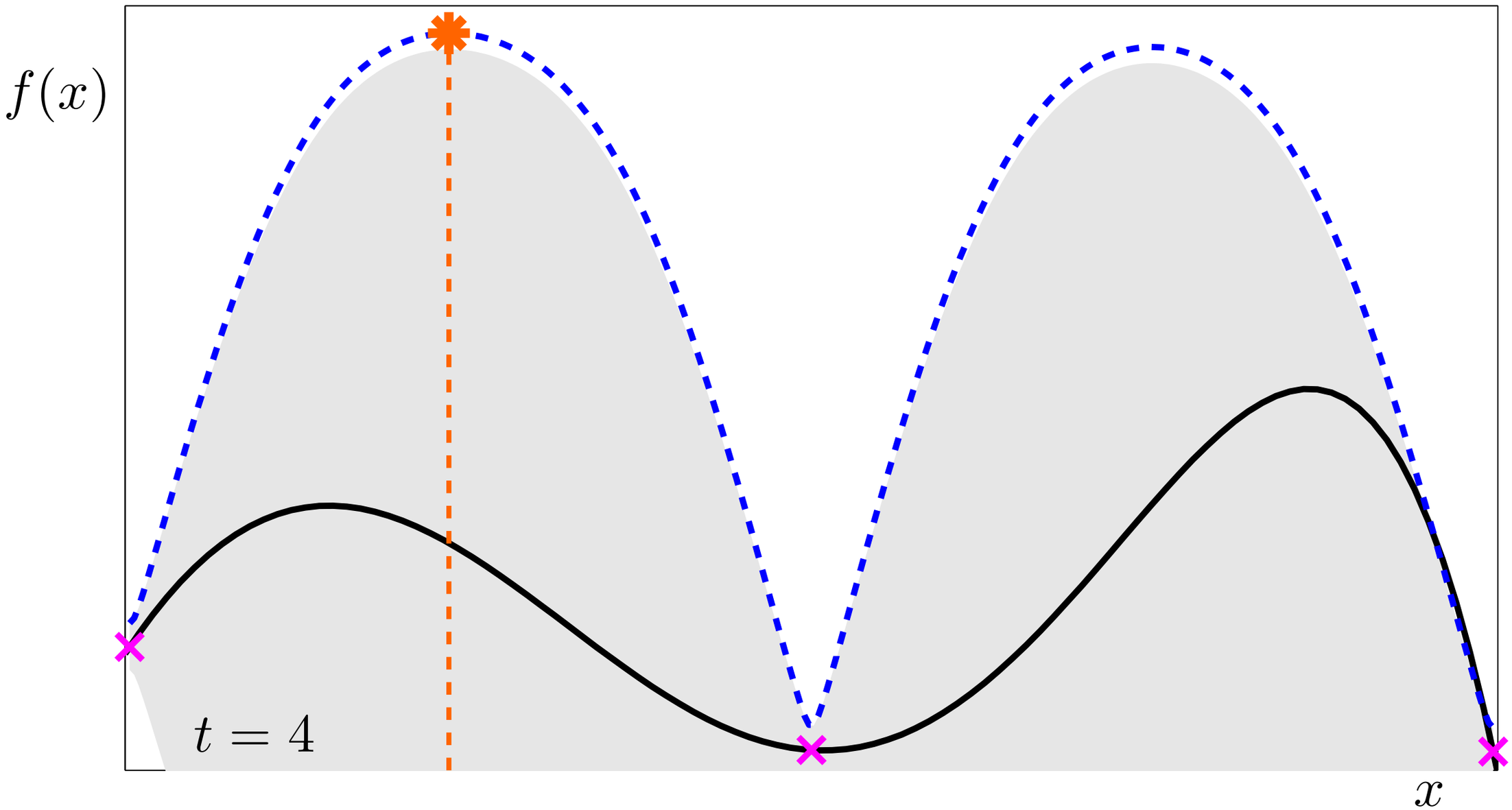
Srinivas et al. 2010

GP-UCB



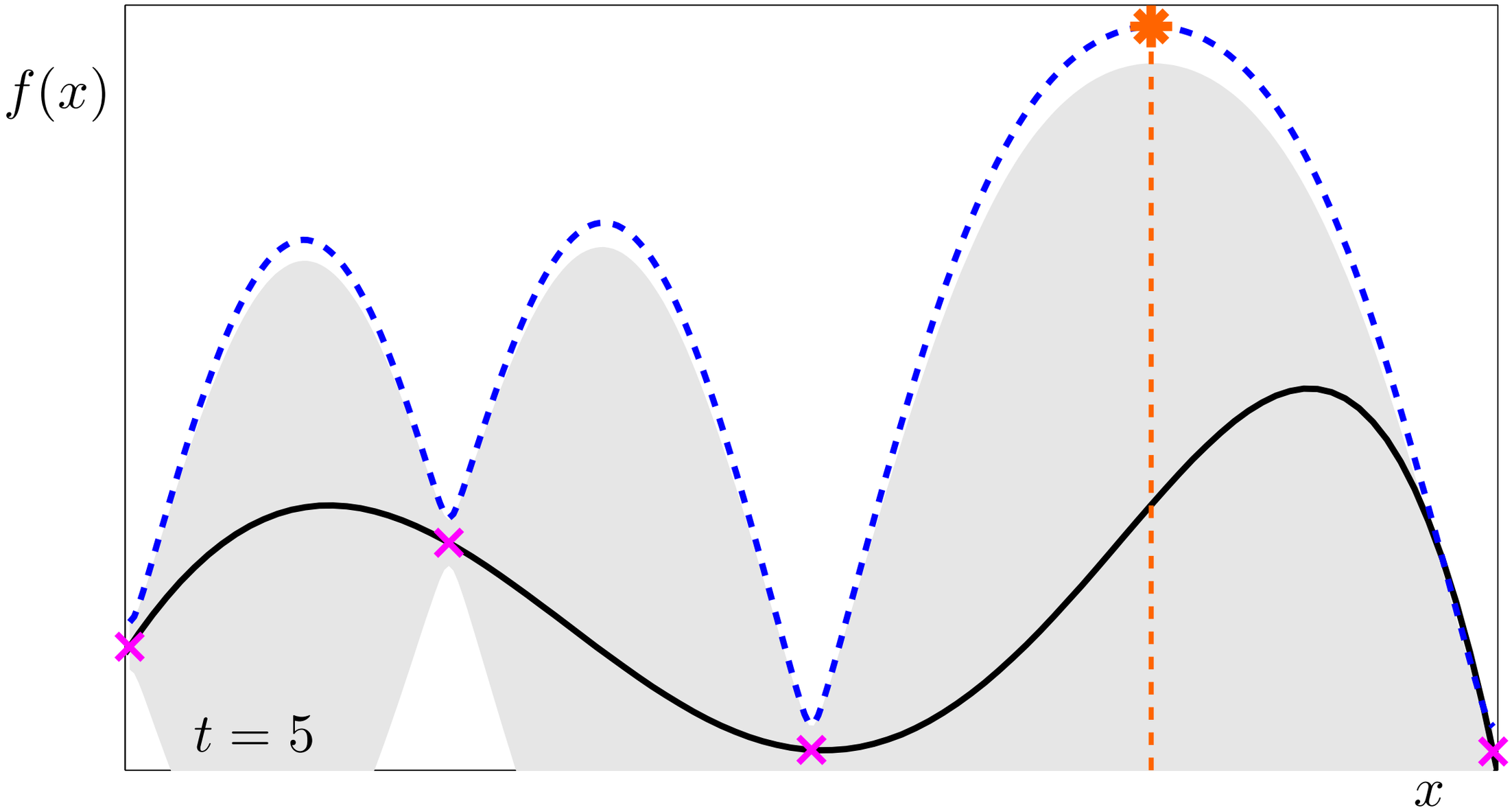
Srinivas et al. 2010

GP-UCB



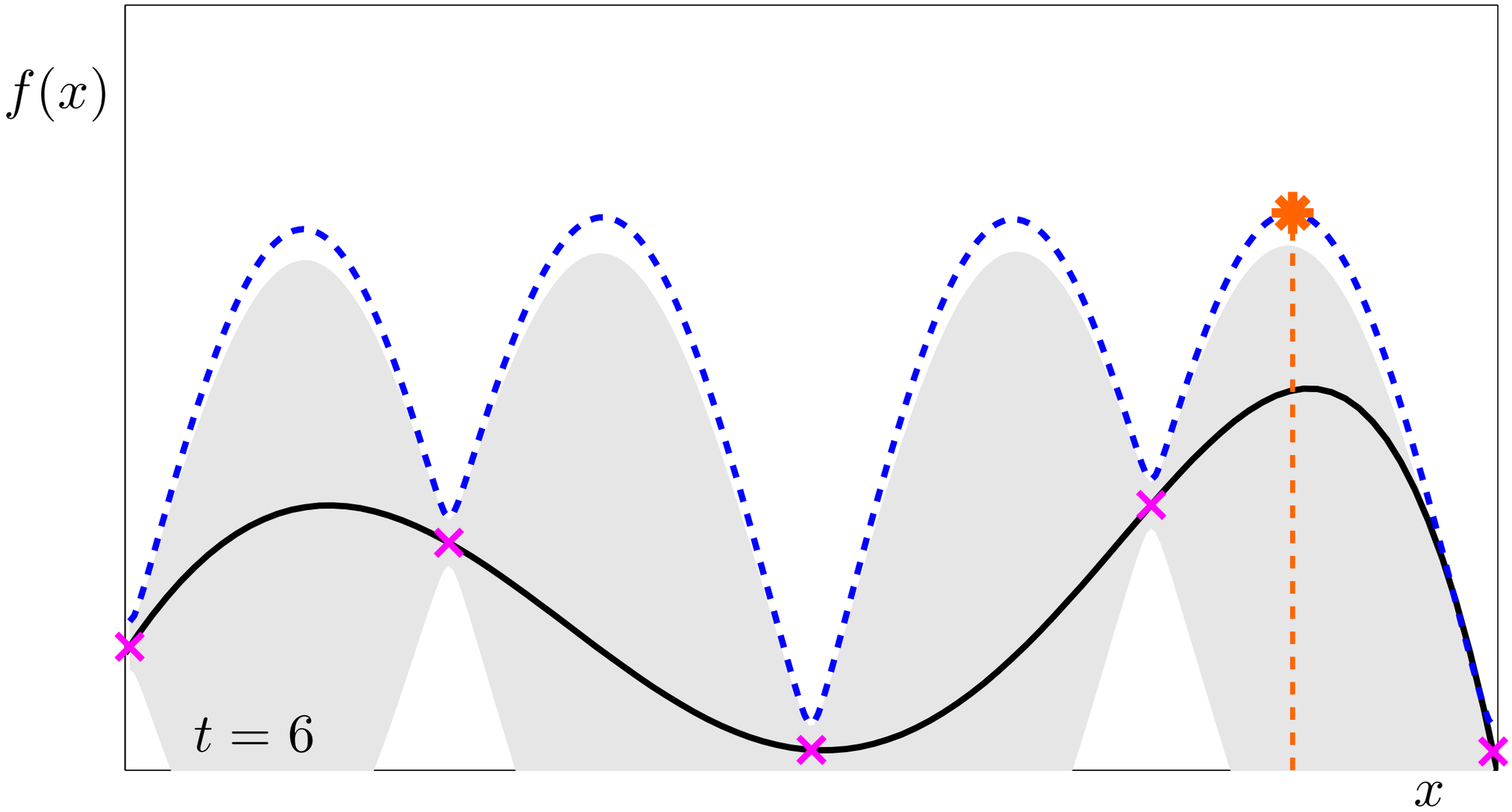
Srinivas et al. 2010

GP-UCB



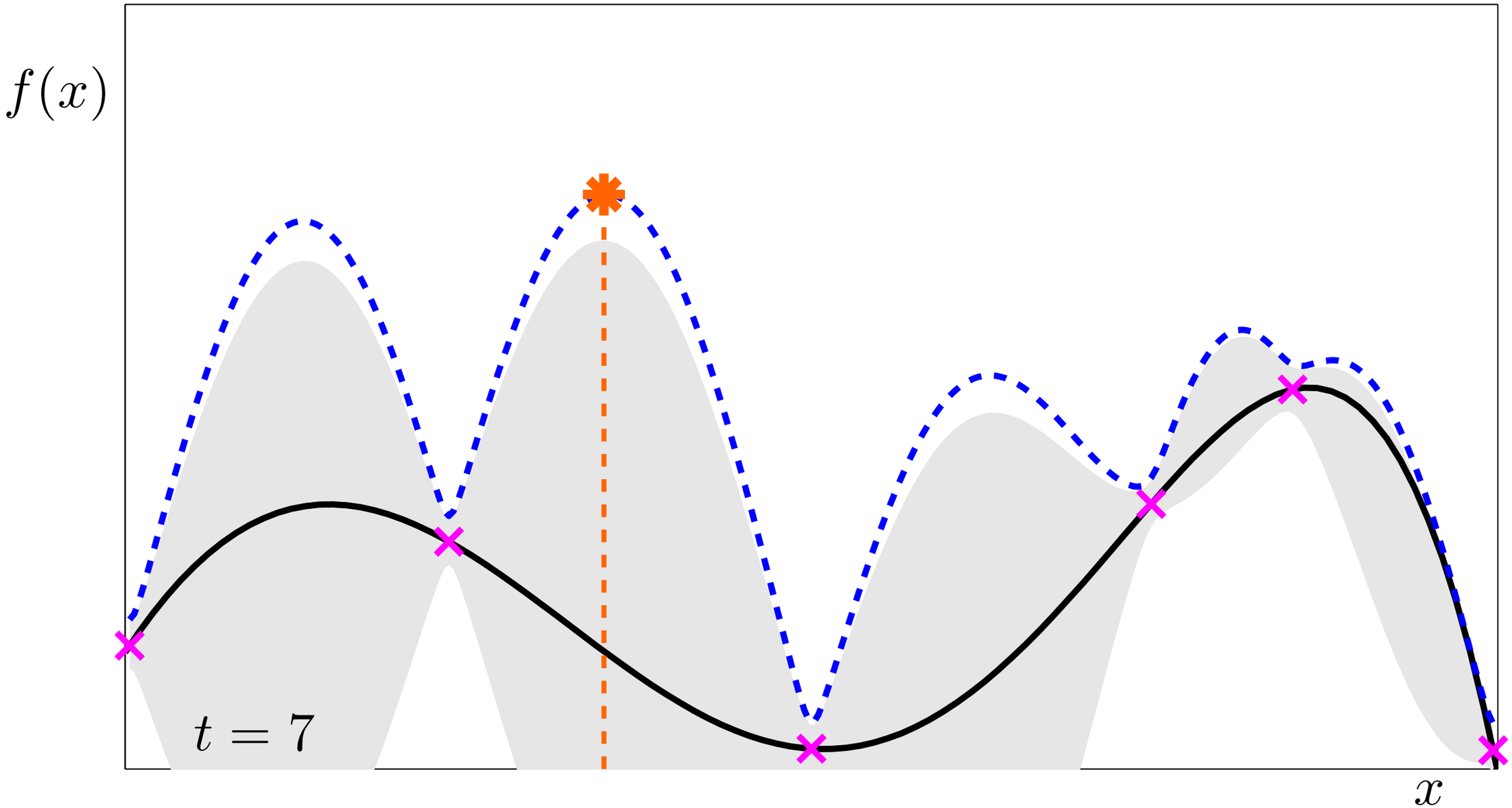
Srinivas et al. 2010

GP-UCB



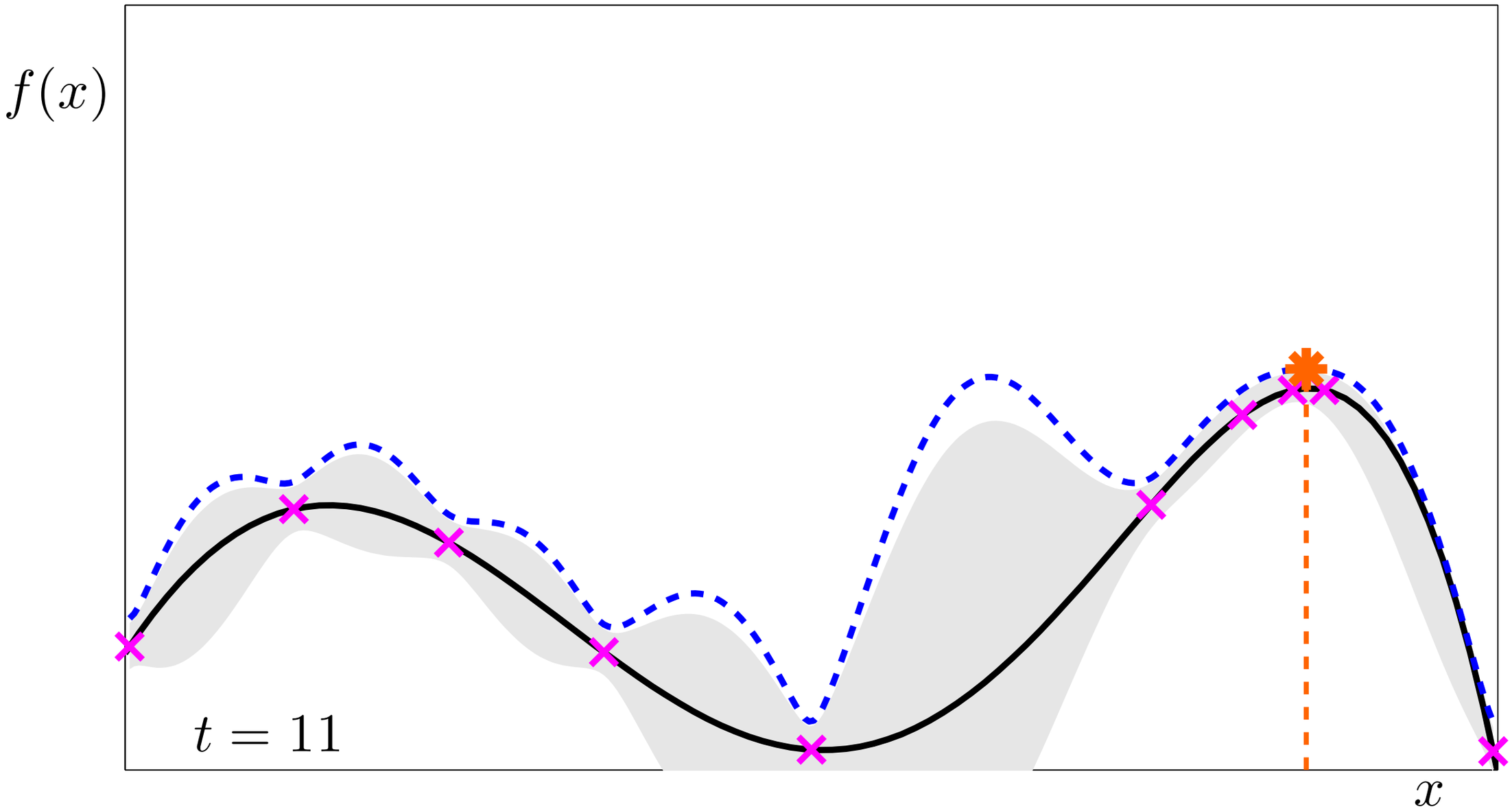
Srinivas et al. 2010

GP-UCB



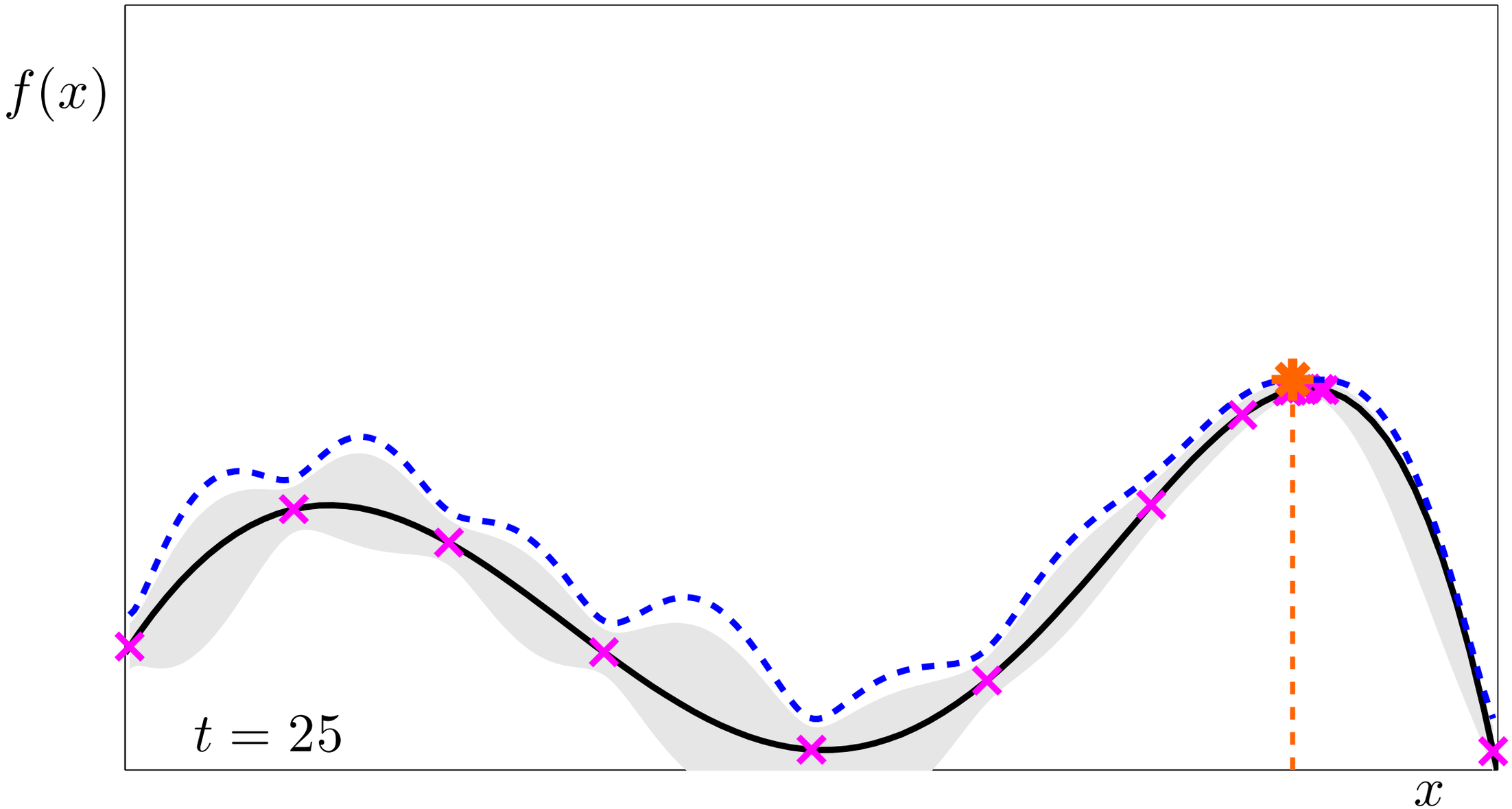
Srinivas et al. 2010

GP-UCB



Srinivas et al. 2010

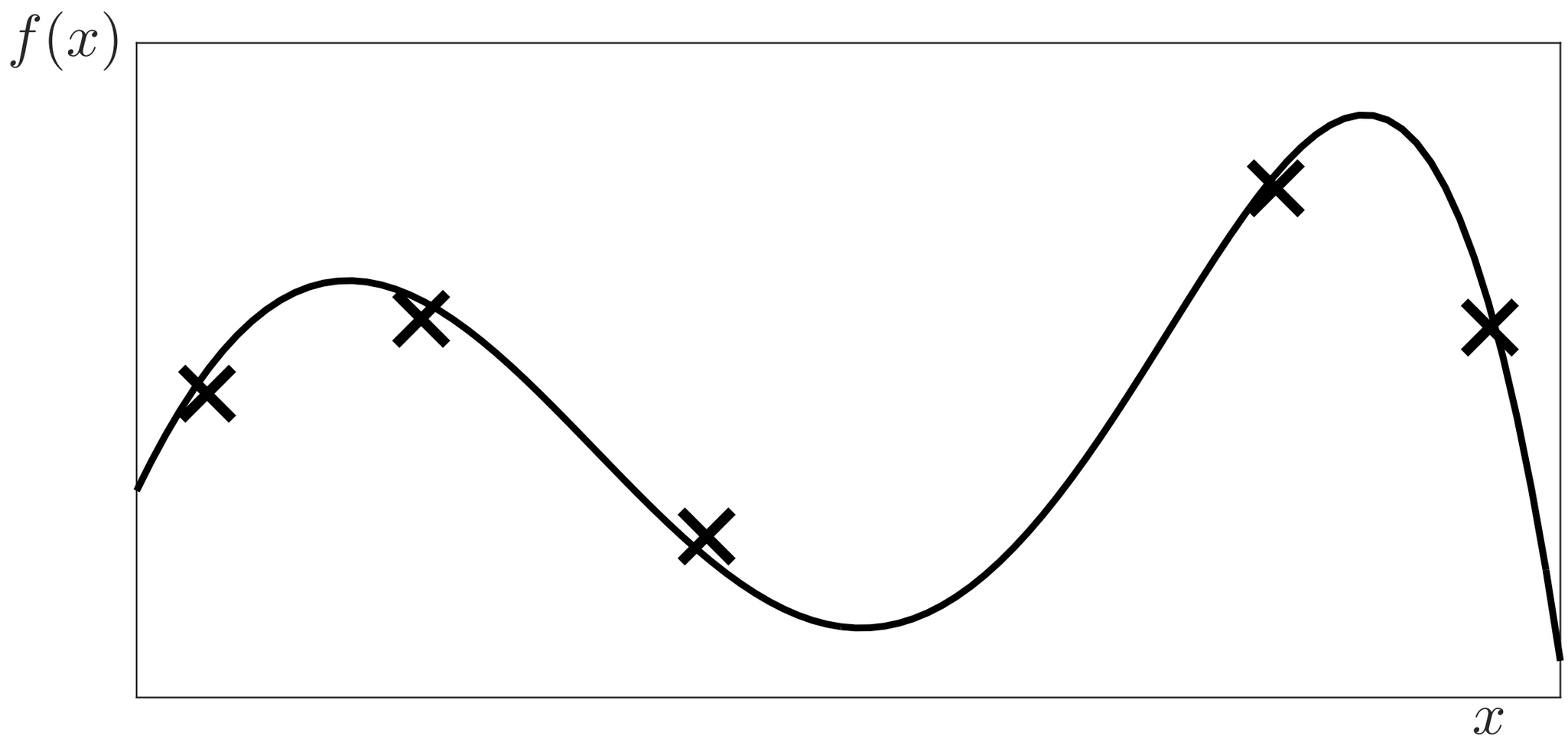
GP-UCB



Srinivas et al. 2010

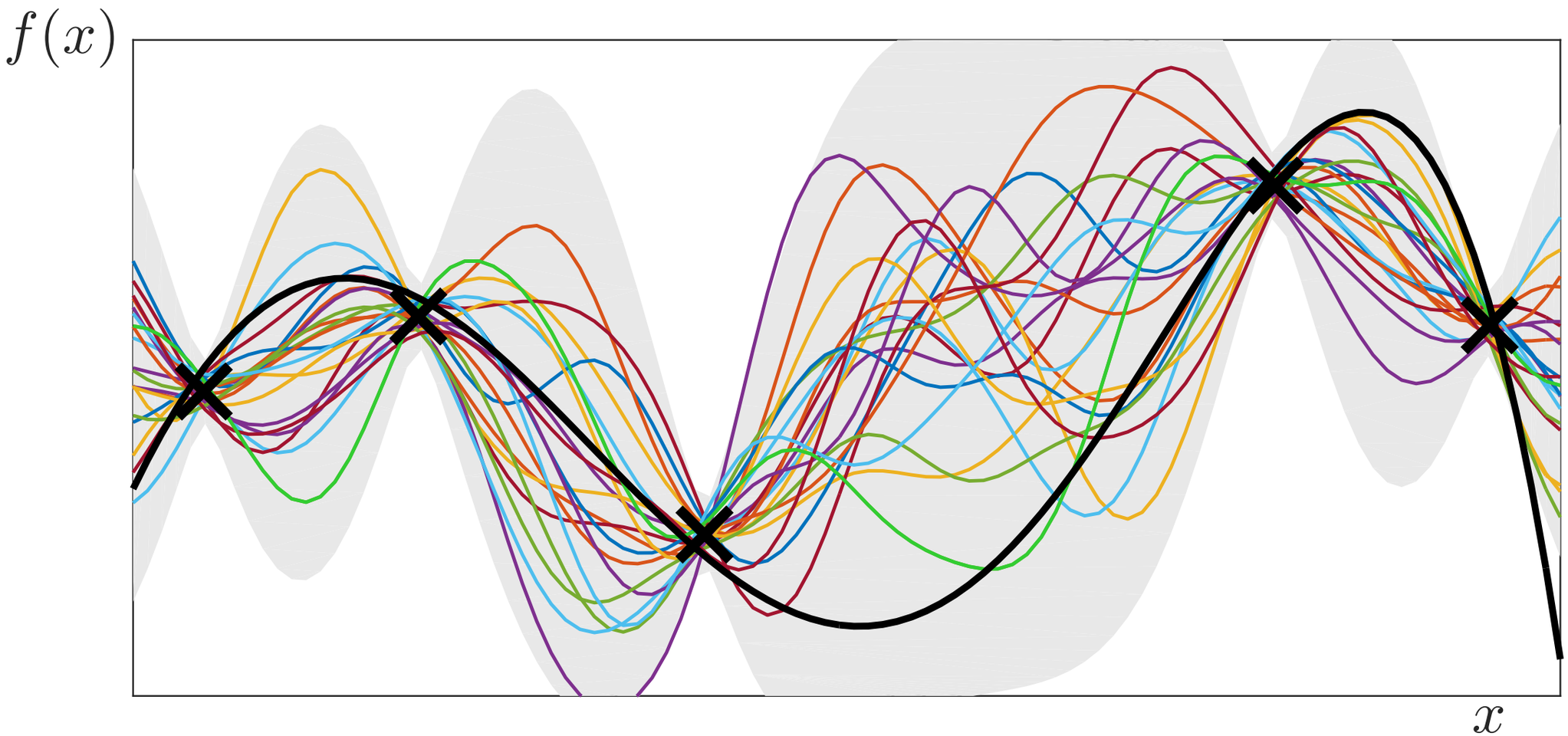
Algorithm 2: Thompson Sampling in GP Bandits

Model $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$. (Thompson, 1933)



Algorithm 2: Thompson Sampling in GP Bandits

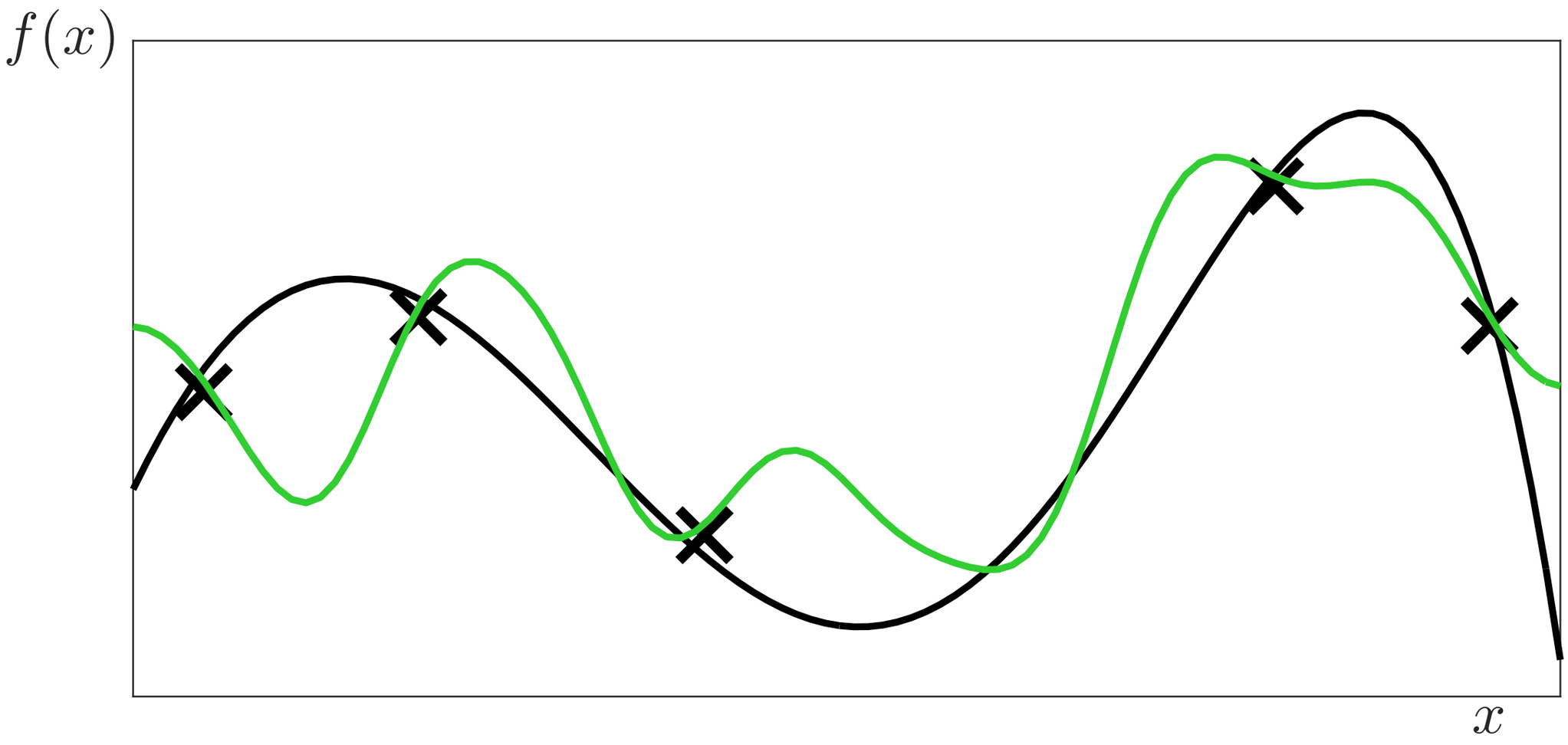
Model $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$. (Thompson, 1933)



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

Algorithm 2: Thompson Sampling in GP Bandits

Model $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$. (Thompson, 1933)

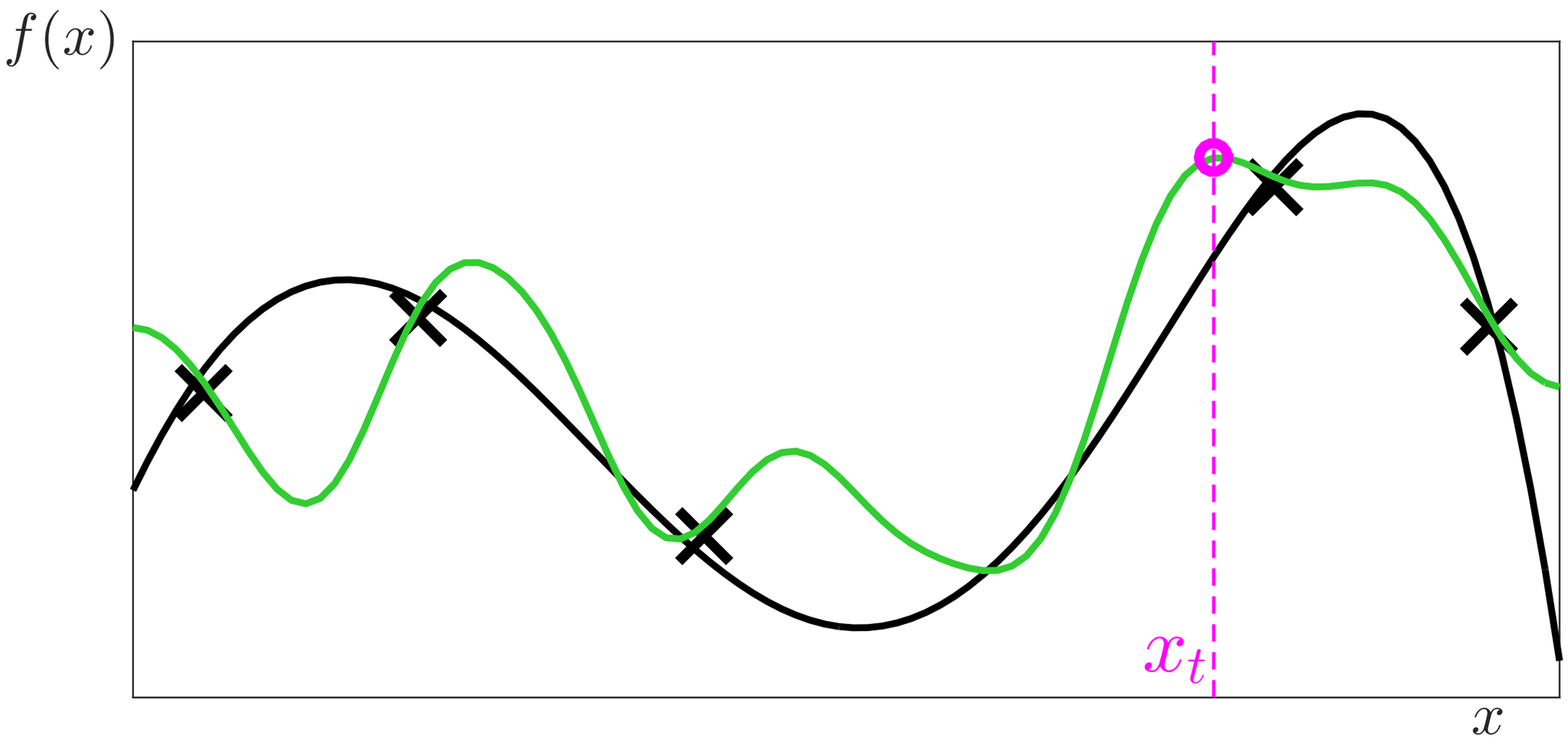


1) Construct posterior $\mathcal{GP}$.

2) Draw sample g from posterior.

Algorithm 2: Thompson Sampling in GP Bandits

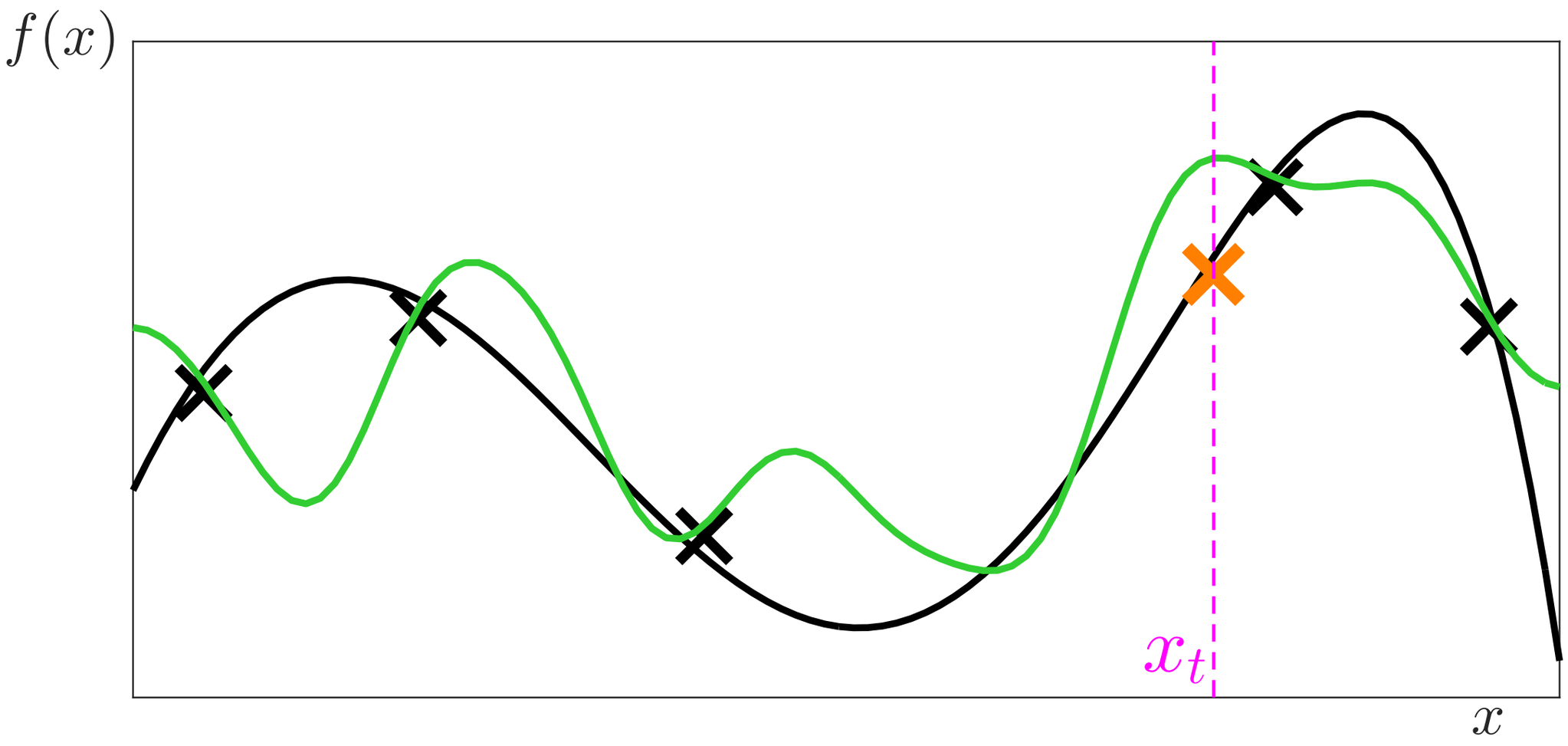
Model $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$. (Thompson, 1933)



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

Algorithm 2: Thompson Sampling in GP Bandits

Model $f \sim \mathcal{GP}(\mathbf{0}, \kappa)$. (Thompson, 1933)



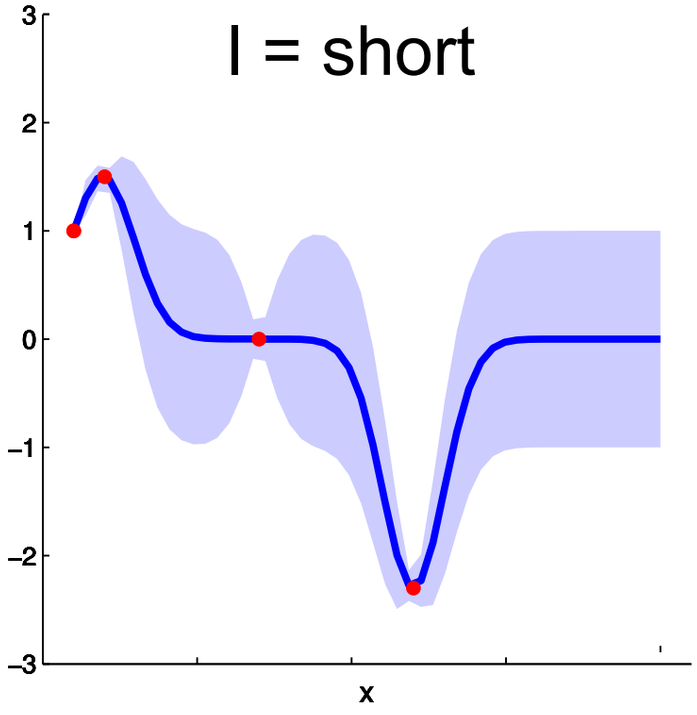
1) Construct posterior $\mathcal{GP}$.
2) Draw sample g from posterior.
3) Choose $x_t = \mathrm{argmax}_x\, g(x)$.
4) Evaluate f at x_t.

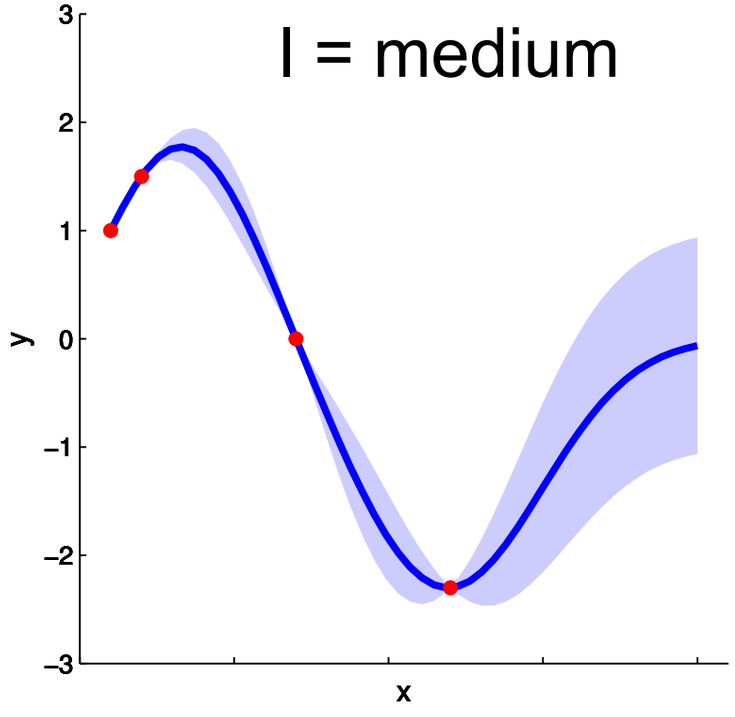
What effect do the hyper-parameters have?

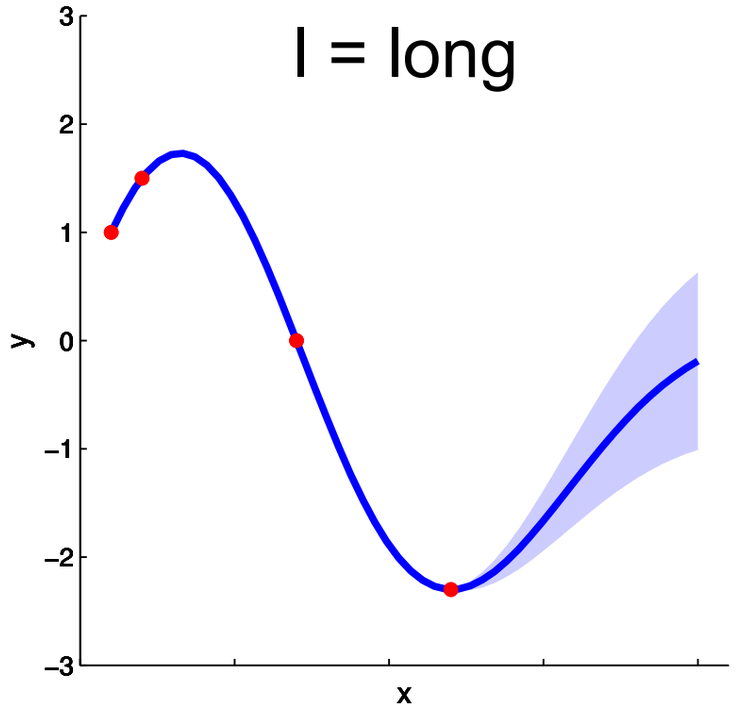
$$\mathsf{K}(\mathsf{x}_1, \mathsf{x}_2) = \sigma^2 \exp\left(-\frac{1}{2\mathsf{l}^2}(\mathsf{x}_1 - \mathsf{x}_2)^2\right)$$

- Hyper-parameters have a strong effect
  - l controls the horizontal length-scale
  - σ^2 controls the vertical scale of the data

- $\Longrightarrow$ we need automatic ways of learning the hyper-parameters from data







Computational Cost

- prediction task
  - train on N points
  - test on M points

- prediction equations

$$\mu_M = \mathsf{K}_{MN}\mathsf{K}_{NN}^{-1}\mathsf{y}_N$$
$$\Sigma_{MM} = \mathsf{K}_{MM} - \mathsf{K}_{MN}\mathsf{K}_{NN}^{-1}\mathsf{K}_{NM}$$

- Full cost $\mathcal{O}((N+M)^3)$ just variances $\mathcal{O}(N^2M)$

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

$\Rightarrow$ Computational cost is a major limitation of GPs

Representing Uncertainty

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

We will look into neural network ensembles: set of neural networks trained on different subsets of the data and with different initializations, and we will look into the entropy of their predictions to quantify the uncertainty of their estimates.