Probabilistic Approaches for Pattern Recognition

Cemre Zor

September 2021

Slides courtesy of Anil Rao, Andre Marquand, Richard E. Turner

Overview of PR in Neuroimaging

PR involves learning a mapping between input and output:

Overview of PR in Neuroimaging

PR involves learning a mapping between input and output:

Input

Training examples
$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$

No mathematical
Model available

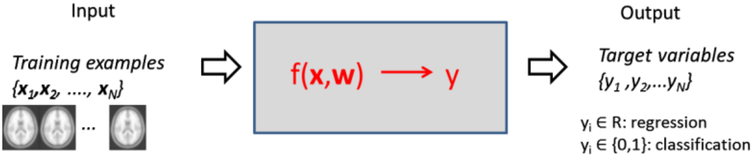
Output

Target variables
$\{y_1, y_2, \ldots y_N\}$

$y_i \in R$: regression
$y_i \in \{0,1\}$: classification

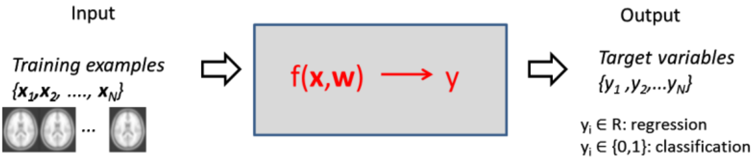
Overview of PR in Neuroimaging

PR involves learning a mapping between input and output:



Overview of PR in Neuroimaging

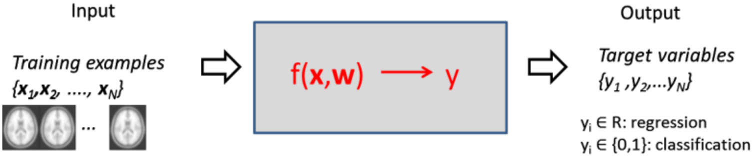
PR involves learning a mapping between input and output:



PR techniques hold two main advantages over conventional univariate analytic methods:

Overview of PR in Neuroimaging

PR involves learning a mapping between input and output:

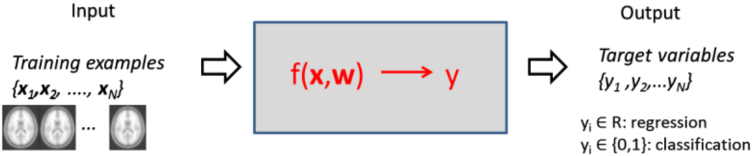


PR techniques hold two main advantages over conventional univariate analytic methods:

1. They can make **predictions** at the level of single subjects

Overview of PR in Neuroimaging

PR involves learning a mapping between input and output:



PR techniques hold two main advantages over conventional univariate analytic methods:

1. They can make **predictions** at the level of single subjects
2. They are **multivariate** (e.g. They can make use of correlations between brain regions)

Approaches to Pattern Recognition

There are many different algorithms used for PR

Algorithms

- Linear Regression
- Neural Networks
- Random Forests / Decision Trees
- Linear Discriminant Analysis
- Kernel methods (e.g. Support Vector Machines, Gaussian Processes, Relevance Vector Machines)

- Important factors in model performance
    - Accuracy
    - Speed

- Important factors in model performance
    - Accuracy
    - Speed
- Confidence intervals
    - Probabilistic Approaches

Why do we “go” probabilistic?

- Handy for
  - quantifying the “belief” in the observations
    - useful for clinical applications where it is a natural way to accurately reflect variability within clinical populations
    - e.g. the probability that a subject has a psychiatric disorder

Why do we "go" probabilistic?

- Handy for
    - quantifying the "belief" in the observations
        - useful for clinical applications where it is a natural way to accurately reflect variability within clinical populations
        - e.g. the probability that a subject has a psychiatric disorder
    - generating a continuous measure that can be correlated with behavioural variables
        - e.g. make quantitative predictions from whole-brain fMRI volumes, which may correlate with subjective pain intensity for every stimulus class

Why do we "go" probabilistic?

- Handy for
    - quantifying the "belief" in the observations
        - useful for clinical applications where it is a natural way to accurately reflect variability within clinical populations
        - e.g. the probability that a subject has a psychiatric disorder
    - generating a continuous measure that can be correlated with behavioural variables
        - e.g. make quantitative predictions from whole-brain fMRI volumes, which may correlate with subjective pain intensity for every stimulus class
    - combining classifiers

Why do we “go” probabilistic?

- Handy for
  - quantifying the “belief” in the observations
    - useful for clinical applications where it is a natural way to accurately reflect variability within clinical populations
    - e.g. the probability that a subject has a psychiatric disorder
  - generating a continuous measure that can be correlated with behavioural variables
    - e.g. make quantitative predictions from whole-brain fMRI volumes, which may correlate with subjective pain intensity for every stimulus class
  - combining classifiers
  - generating new samples
    - e.g. create brain images at different stages of AD disease, after learning their distribution over time

Probabilistic prediction for clinical applications

Coherent handling of uncertainty is especially important in medicine

Sources of uncertainty in clinical applications

- Diagnostic uncertainty (class labels may be noisy)
- Heterogeneity in disease severity and course
- Individual variability in response to treatment

In such applications predictive confidence is potentially highly informative

$p(label|\mathbf{data}) = 0.55$: ambiguous $p(label|\boldsymbol{data}) = 0.99$: confident

Probability Theory

- $p(X)$ is the *marginal* probability of X
- $p(X, Y)$ is the *joint* probability of X and Y
- $p(X|Y)$ is the *conditional* probability of X given Y

Rules

- $0 \leq p(X) \leq 1$
- p(sure thing) = 1
- probabilities must sum to one: $\sum_X p(X) = 1$
- Product rule: $p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X)$
- Sum rule: $p(X) = \sum_Y p(X, Y)$

Bayes rule is derived from the product rule

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \qquad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Probabilistic Learning

- To define a probabilistic model, we start with choosing the **likelihood** function which describes how the data labels were produced

Probabilistic Learning

- To define a probabilistic model, we start with choosing the **likelihood** function which describes how the data labels were produced

$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T\mathbf{w}$

Probabilistic Learning

- To define a probabilistic model, we start with choosing the **likelihood** function which describes how the data labels were produced

$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T\mathbf{w}$

$$p(labels \,|\, parameters) = p(\mathbf{y} \,|\, \mathbf{w}, \mathbf{x})$$

Probabilistic Learning

- To define a probabilistic model, we start with choosing the **likelihood** function which describes how the data labels were produced

$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T\mathbf{w}$

$$p(labels \mid parameters) = p(\mathbf{y} \mid \mathbf{w}, \mathbf{x})$$

Many possible choices depending on our problem eg. if we are doing regression or classification.

Probabilistic Learning

- To define a probabilistic model, we start with choosing the **likelihood** function which describes how the data labels were produced

$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T\mathbf{w}$ $\qquad p(labels \mid parameters) = p(\mathbf{y} \mid \mathbf{w}, \mathbf{x})$

Many possible choices depending on our problem eg. if we are doing regression or classification.

- We also specify our **prior** beliefs about the weight vector

$$p(parameters \mid hyperparameters) = p(\mathbf{w} \mid \theta)$$

You can think of this as similar to regularisation in non-probabilistic approaches

Probabilistic Learning

- Inference then amounts to computing the posterior distribution (Bayes rule)

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \theta) = \frac{p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w}|\theta)}{p(\mathbf{y}|\mathbf{X}, \theta)}$$

Likelihood

Prior

Posterior

Marginal Likelihood

- Gives a **distribution** for the weight vector **w** given the data. We can then find the best **w** and use it to perform predictions

Probabilistic Learning

- Inference then amounts to computing the posterior distribution (Bayes rule)

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \theta) = \frac{p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w}|\theta)}{p(\mathbf{y}|\mathbf{X}, \theta)}$$

Likelihood: $p(\mathbf{y}|\mathbf{w}, \mathbf{X})$

Prior: $p(\mathbf{w}|\theta)$

Posterior: $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \theta)$

Marginal Likelihood: $p(\mathbf{y}|\mathbf{X}, \theta)$

- Gives a **distribution** for the weight vector **w** given the data. We can then find the best **w** and use it to perform predictions

$$\arg\max_{\mathrm{w}} \; p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \theta) \propto \prod_{i=1}^{n} \; p(y \; |\mathbf{w}, \mathbf{x}) \; p(\mathbf{w}|\theta)$$

Probabilistic Learning

- Inference then amounts to computing the posterior distribution (Bayes rule)

Likelihood

Prior

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \theta) = \frac{p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w}|\theta)}{p(\mathbf{y}|\mathbf{X}, \theta)}$$

Posterior

Marginal Likelihood

- Gives a **distribution** for the weight vector **w** given the data. We can then find the best **w** and use it to perform predictions

$$\arg\max_{\mathrm{w}} \; p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \theta) \propto \prod_{i=1}^{n} \; p(y \;|\mathbf{w}, \mathbf{x}) \; p(\mathbf{w}|\theta)$$

$$\arg\max_{\mathrm{w}} \; \log p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \theta) \propto \sum_{i=1}^{n} \log p(y \;|\mathbf{w}, \mathbf{x}) + \log p(\mathbf{w}|\theta)$$

Decision Theory

In neuroimaging, we commonly divide the learning process into two phases:

1. **Inference**: computing the posterior distributions
2. **Decision**: make a prediction/decision based on the posterior

Decision Theory

In neuroimaging, we commonly divide the learning process into two phases:

1. **Inference**: computing the posterior distributions
2. **Decision**: make a prediction/decision based on the posterior

- This framework is highly flexible: e.g. we can accommodate asymmetric misclassification costs where a false negative may be costly than a false positive (medical applications)

- In contrast, many approaches combine these phases and learn a function that directly maps inputs (x) onto class labels (y).

Connection to non-probabilistic approaches

- Consider a linear model that aims to predict the output (y) using a weighted combination of the inputs (**x**)

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T \mathbf{w} + b$$

- To estimate the weights we seek to minimise the empirical risk which is penalised to restrict model flexibility

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} \sum_{i=1}^{n} L(y_i, \mathbf{x}_i, \mathbf{w}) + \lambda J(\mathbf{w})$$

Connection to non-probabilistic approaches

- Consider a linear model that aims to predict the output (y) using a weighted combination of the inputs (**x**)

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T \mathbf{w} + b$$

- To estimate the weights we seek to minimise the empirical risk which is penalised to restrict model flexibility

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} \sum_{i=1}^{n} L(y_i, \mathbf{x}_i, \mathbf{w}) + \lambda J(\mathbf{w})$$

- Probabilistic models can be viewed from a similar perspective

$$\log p(\mathbf{w} | \mathbf{y}, \mathbf{X}, \theta) \propto \sum_{i=1}^{n} \log p(y_i | \mathbf{w}, \mathbf{x}_i) + \log p(\mathbf{w} | \theta)$$

Model Selection

- The marginal likelihood (evidence) plays an important role in probabilistic modeling

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

Model Selection

- The marginal likelihood (evidence) plays an important role in probabilistic modeling

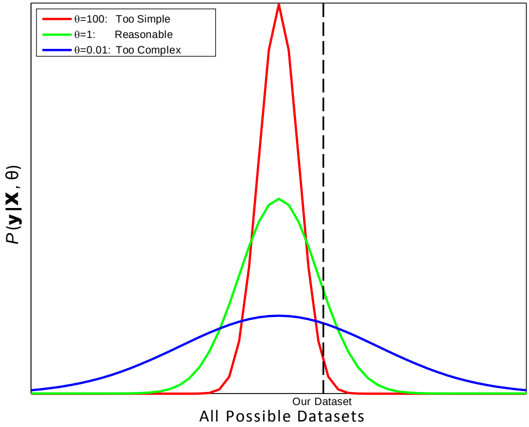
$$p(\mathbf{y}|\mathbf{X},\theta) = \int p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\theta)d\mathbf{w}$$

It embodies a tradeoff between data fit and model complexity and can be used for:

- deciding which of several competing models is most probable
- automatic optimisation of hyperparameters θ by evidence maximisation

Model Selection

- Choosing optimum value for θ



Introduction to Gaussian process models

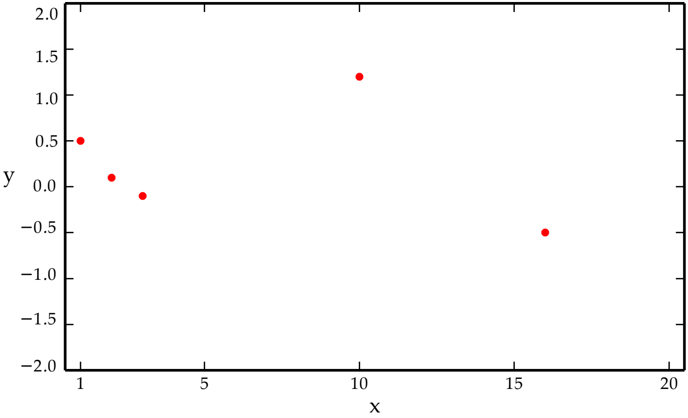
Introduction to Gaussian process models

GPs are flexible probabilistic kernel methods with many applications

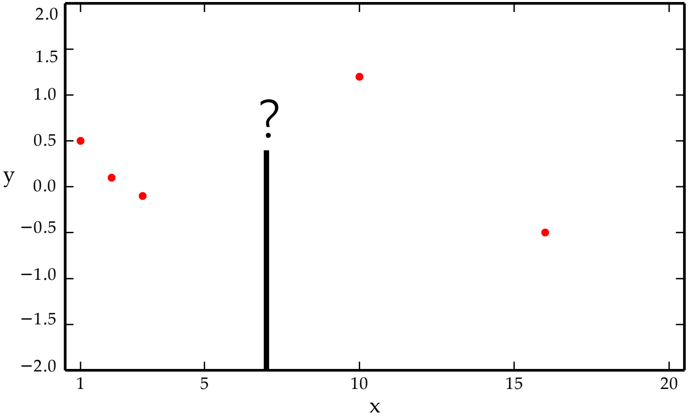
They are most easily understood as a distribution over functions, and GP inference consists of applying Bayes' rule to find the (posterior) function distribution that best approximates the training data.

Introduction to Gaussian process models

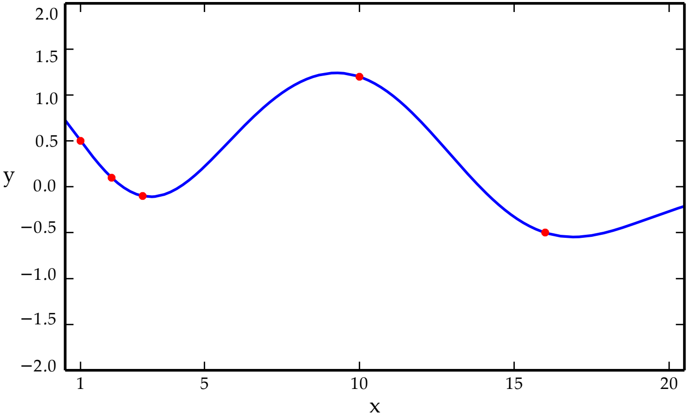
- Advantages:
  - Explicit probabilistic framework (Likelihood-Prior-Posterior)
  - Provide mechanisms for automatic parameter optimisation (optimisation of Marginal Likelihood)
  - Natural extension to binary and multi-class classification

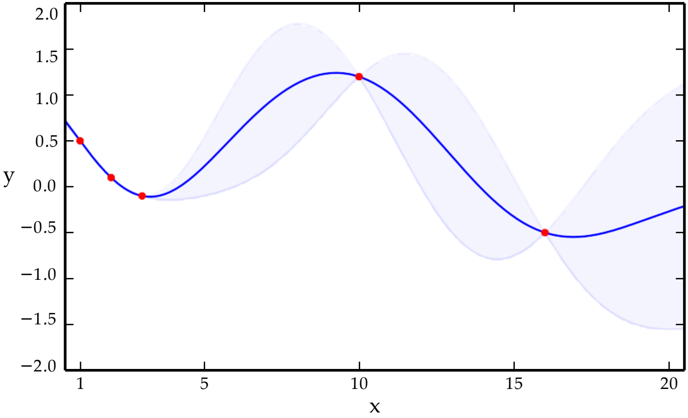


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
y
1
5
10
15
20
x

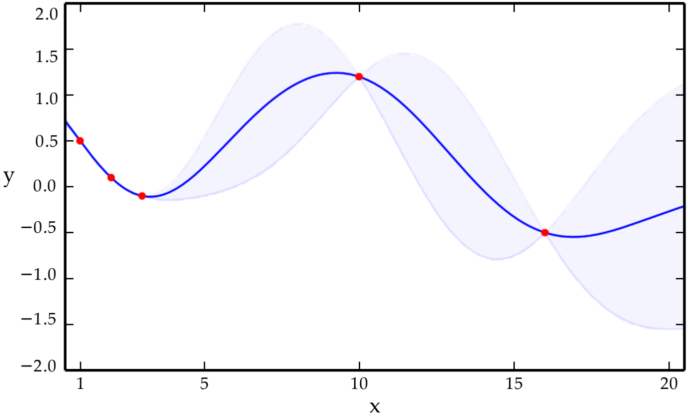


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
y
?
1
5
10
15
20
x


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
y
1
5
10
15
20
x

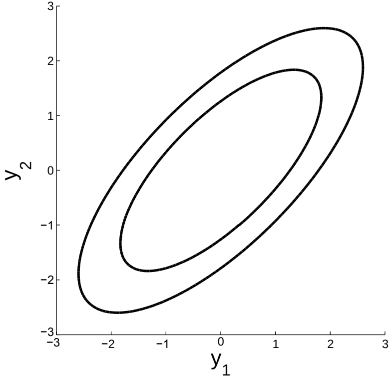

2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
y
1
5
10
15
20
x

Can we do this with a plain old Gaussian?



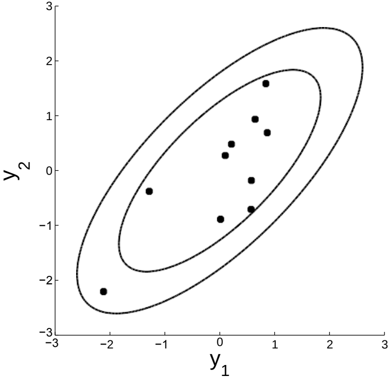
Gaussian Distribution

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



Gaussian Distribution

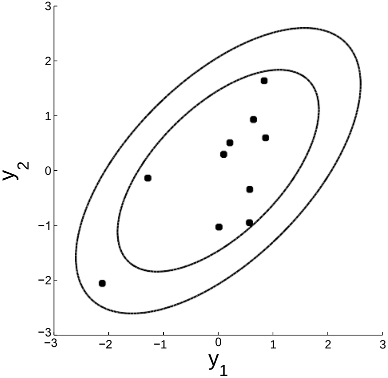
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^T \Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{matrix} 1 & .7 \\ .7 & 1 \end{matrix}$$



Gaussian Distribution

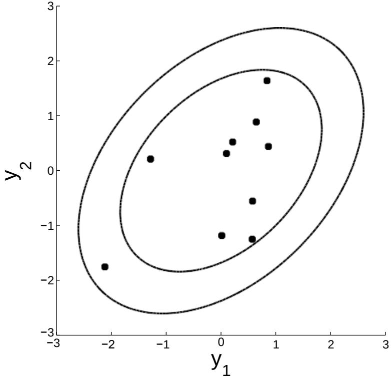
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y}\right)$$

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



Gaussian Distribution

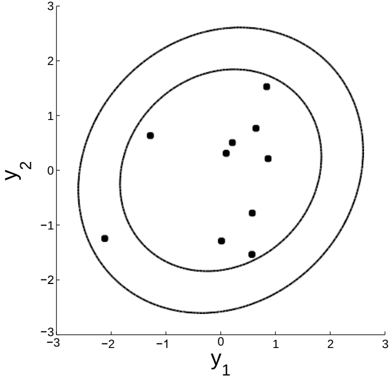
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$$



Gaussian Distribution

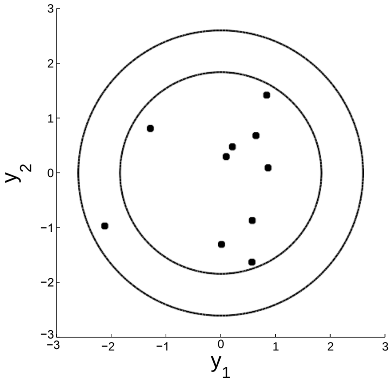
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^T \Sigma^{-1}\mathbf{y}\right)$$

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



Gaussian Distribution

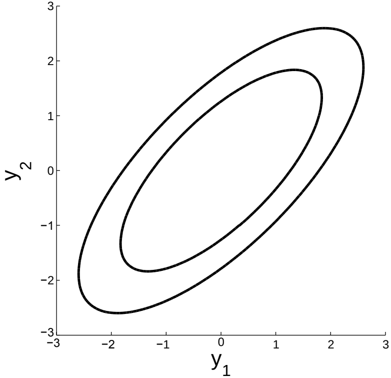
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\frac{1}{2}\mathbf{y}^T \Sigma^{-1}\mathbf{y}\right) \qquad \Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$



Gaussian Distribution

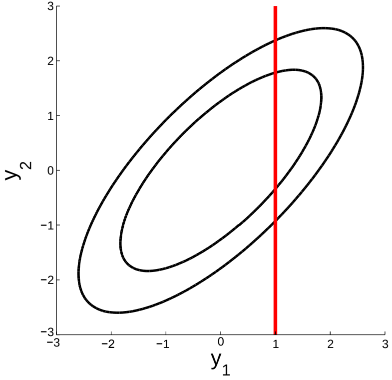
$$p(\mathbf{y}|\Sigma) \propto \exp\left(-\tfrac{1}{2}\mathbf{y}^T \Sigma^{-1}\mathbf{y}\right)$$

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



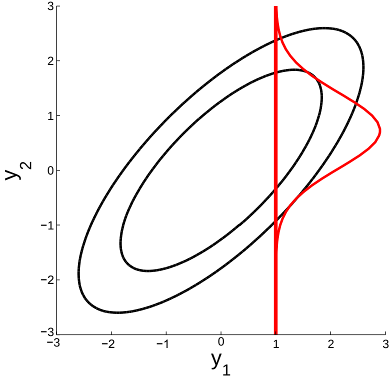
Gaussian Distribution

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



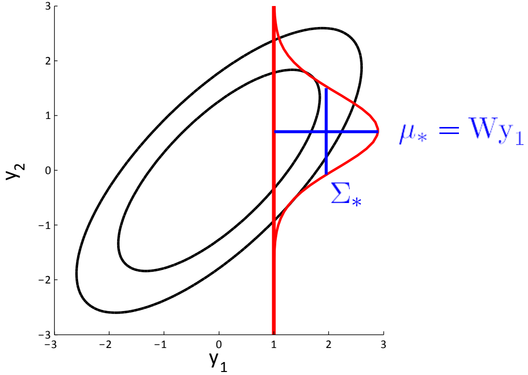
Gaussian Distribution

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



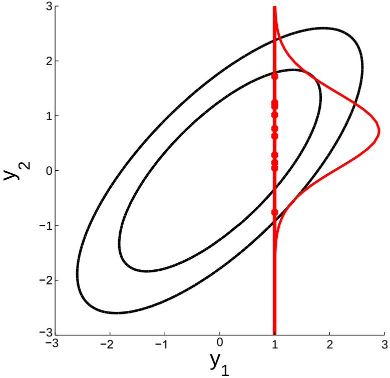
Gaussian Distribution

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



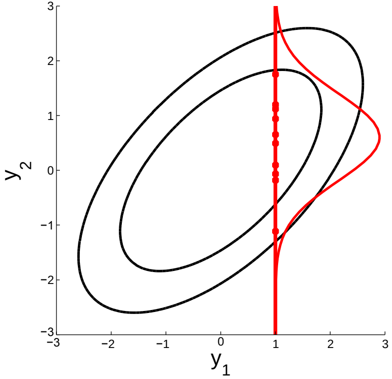
Gaussian Distribution

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



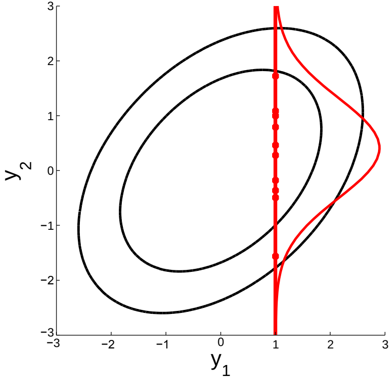
Gaussian Distribution

$$p(y_2 | y_1, \Sigma) \propto \exp\left(-\tfrac{1}{2} (y_2 - \boldsymbol{\mu}_*) \Sigma_*^{-1} (y_2 - \boldsymbol{\mu}_*) \right)$$



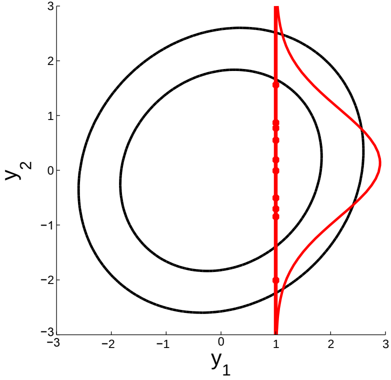
Gaussian Distribution

$$p(y_2|y_1,\Sigma) \propto \exp\left(-\tfrac{1}{2}(y_2-\boldsymbol{\mu}_*)\Sigma_*^{-1}(y_2-\boldsymbol{\mu}_*) \right)$$



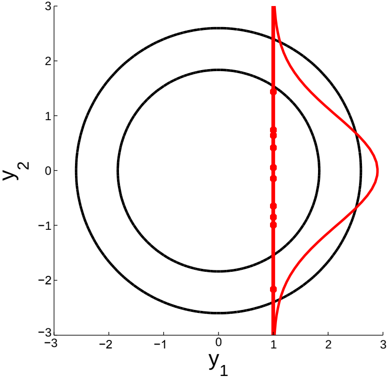
Gaussian Distribution

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

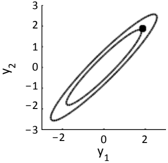


Gaussian Distribution

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

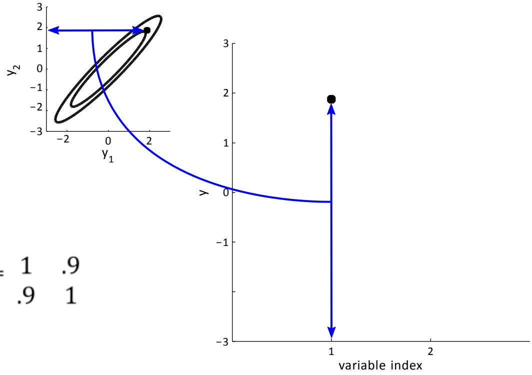


New visualisation



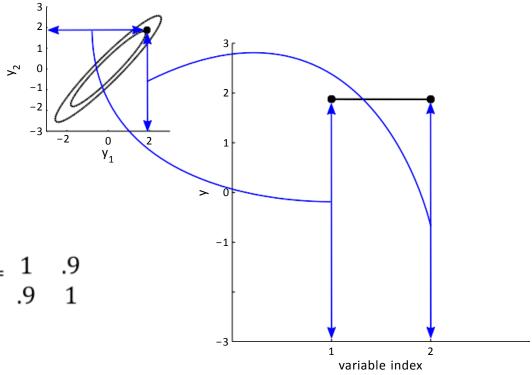
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

New visualisation



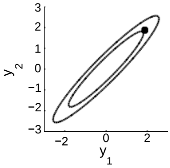
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

New visualisation

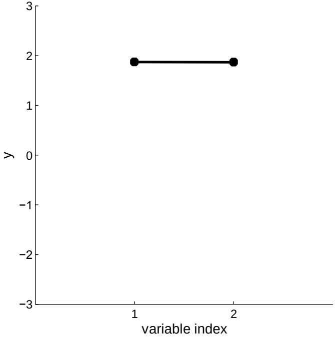


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

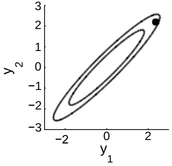
New visualisation

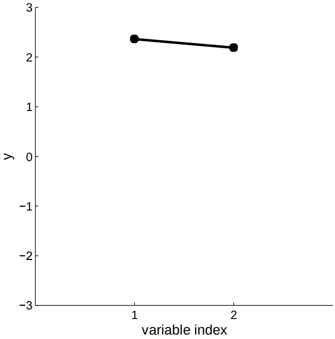


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



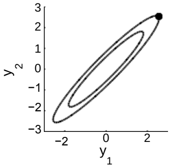
New visualisation



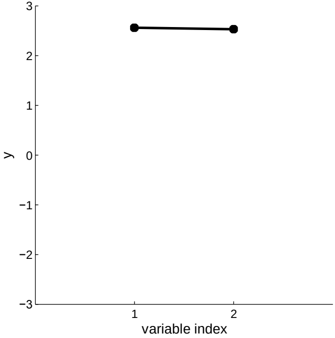


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

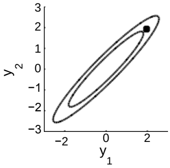
New visualisation



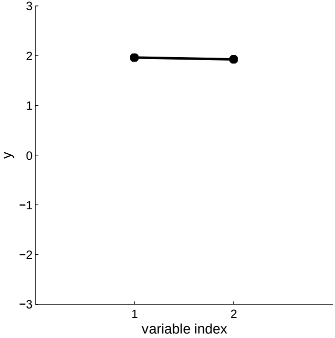
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



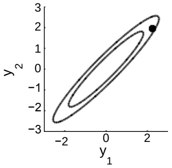
New visualisation



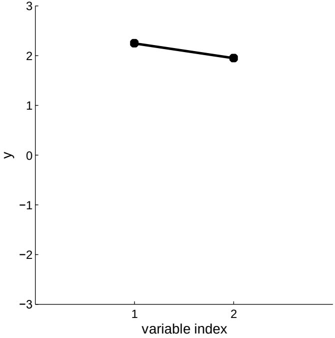
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



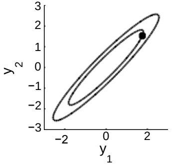
New visualisation

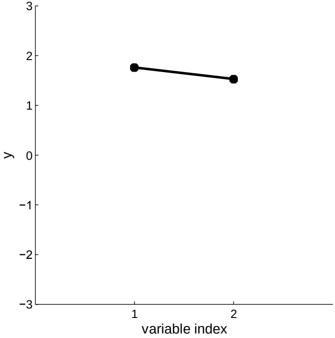


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



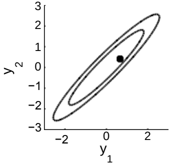
New visualisation

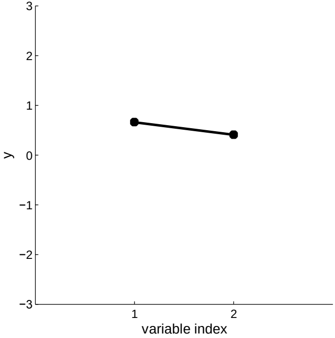




$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

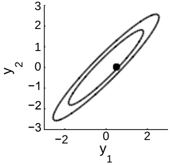
New visualisation

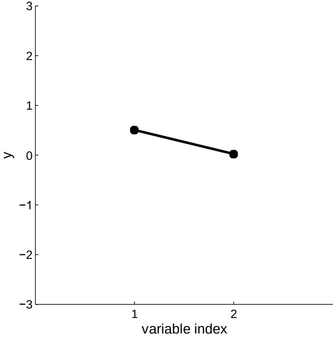




$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

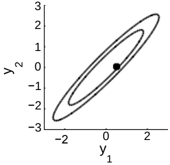
New visualisation



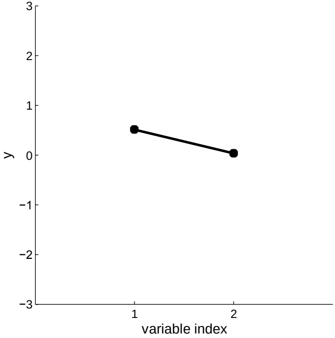


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

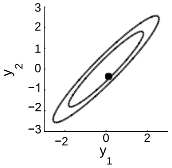
New visualisation



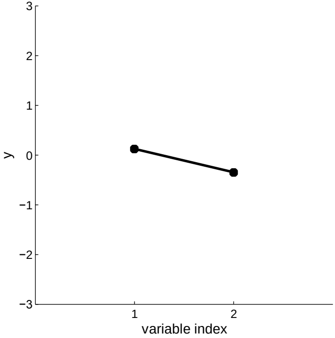
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



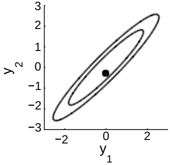
New visualisation

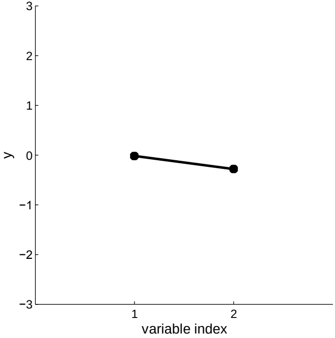


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



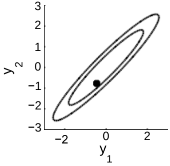
New visualisation



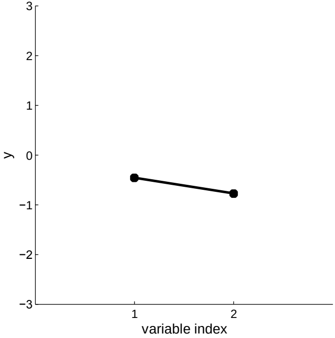


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

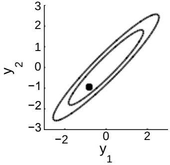
New visualisation



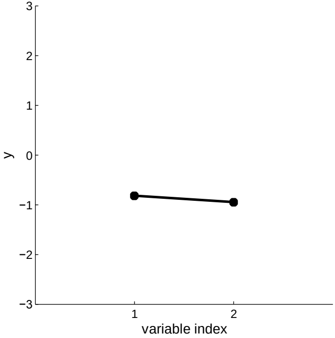
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



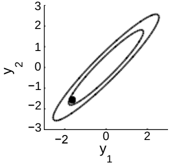
New visualisation



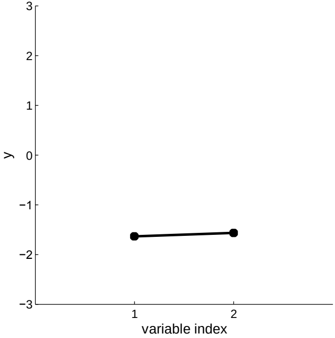
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



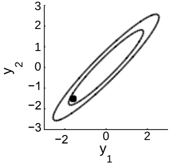
New visualisation



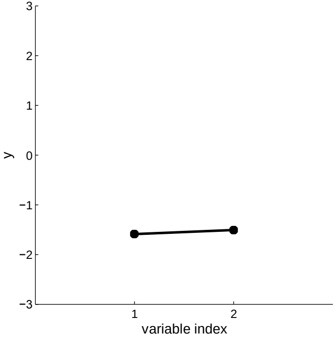
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



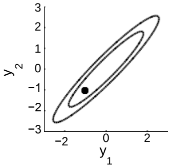
New visualisation



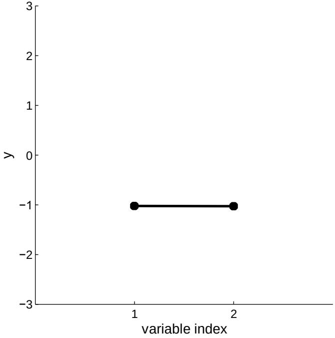
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



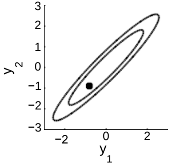
New visualisation



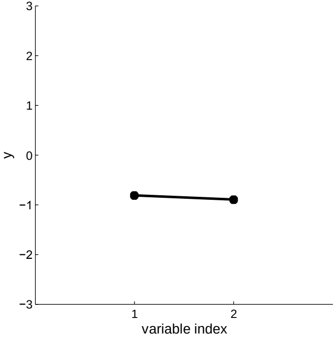
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



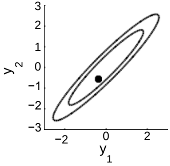
New visualisation



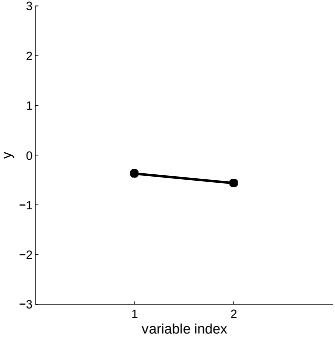
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



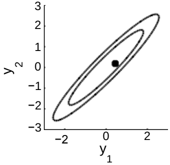
New visualisation



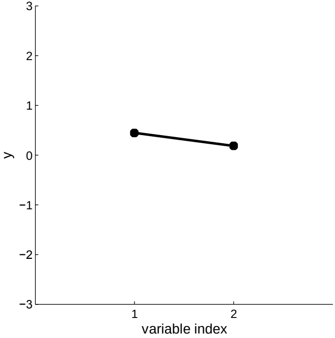
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



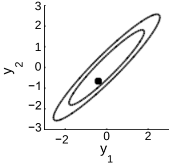
New visualisation

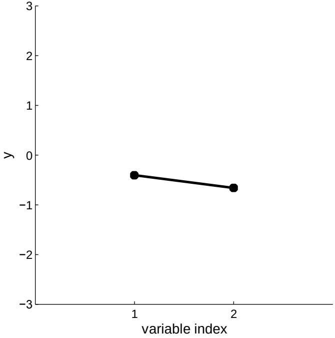


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



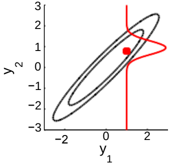
New visualisation



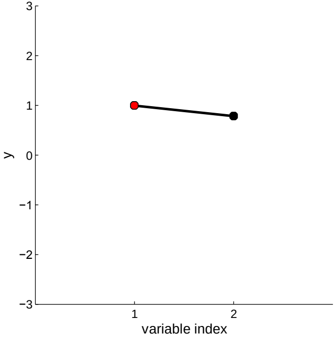


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

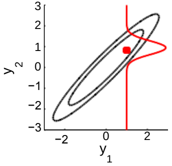
New visualisation



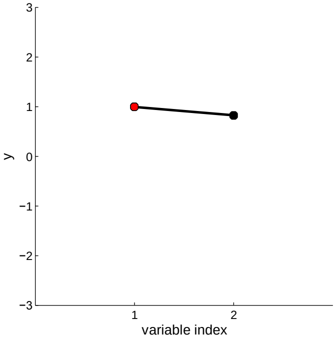
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



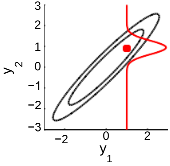
New visualisation

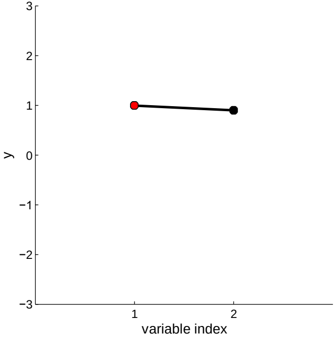


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



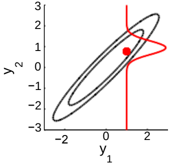
New visualisation



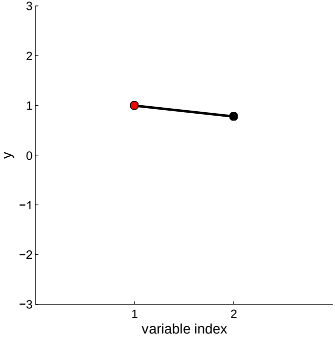


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

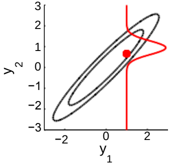
New visualisation



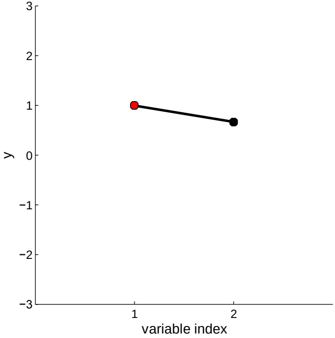
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



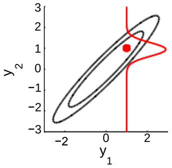
New visualisation



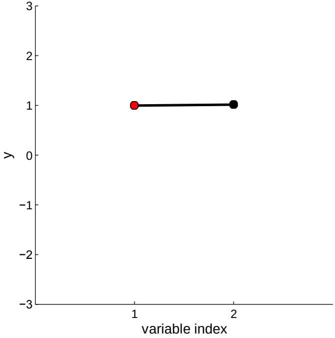
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



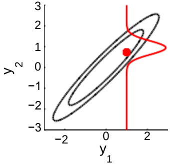
New visualisation

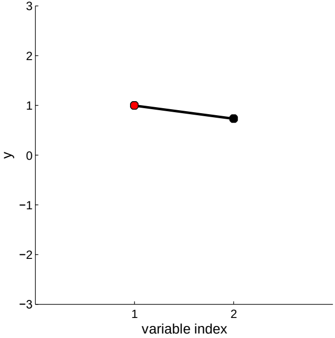


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



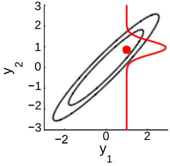
New visualisation



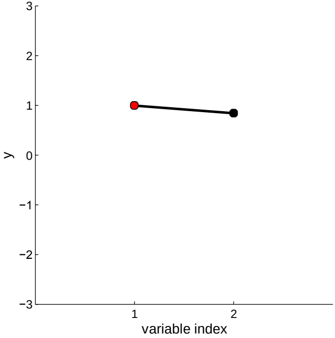


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

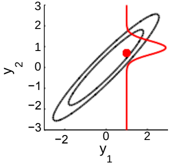
New visualisation



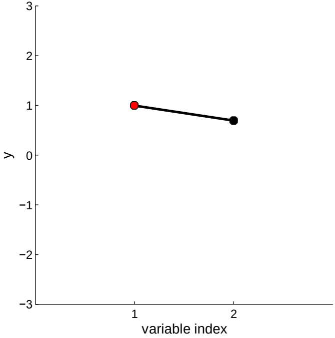
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



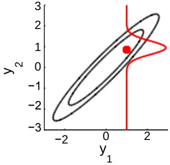
New visualisation

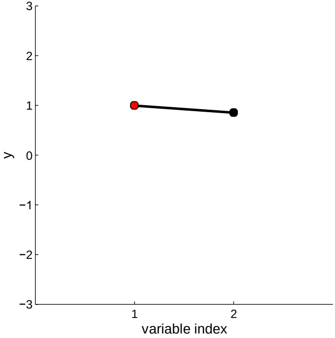


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



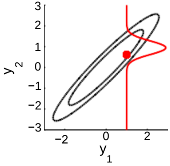
New visualisation



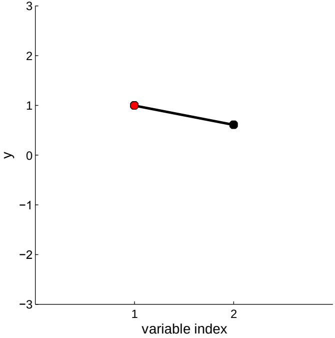


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

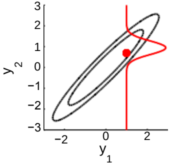
New visualisation



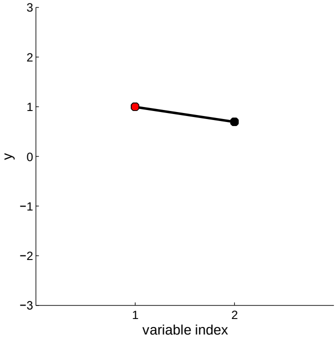
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



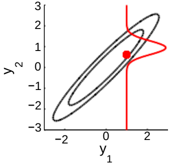
New visualisation

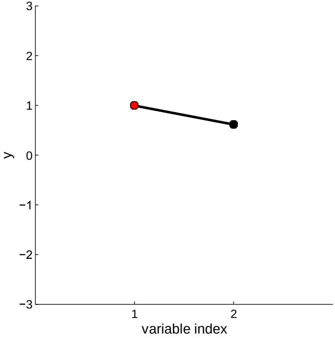


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



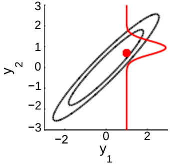
New visualisation



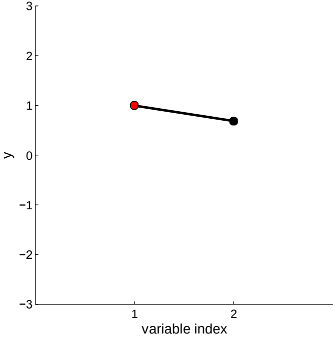


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

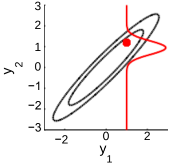
New visualisation



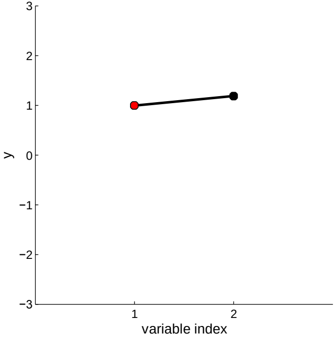
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



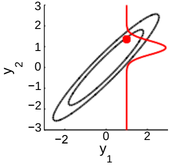
New visualisation



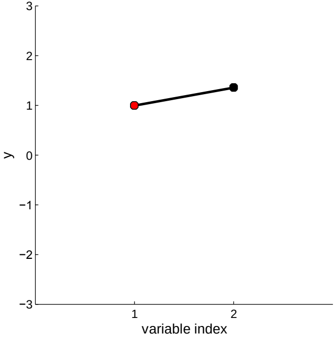
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



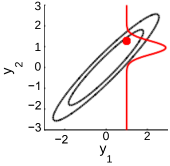
New visualisation



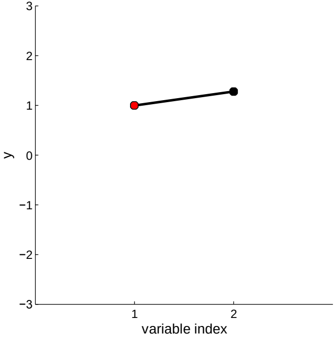
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



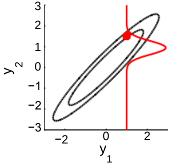
New visualisation



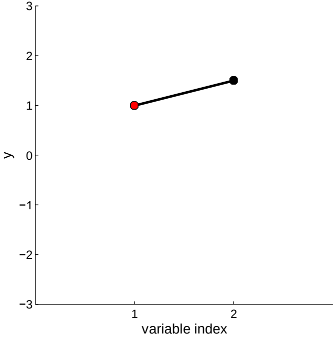
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



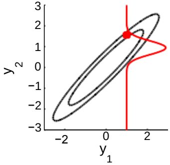
New visualisation



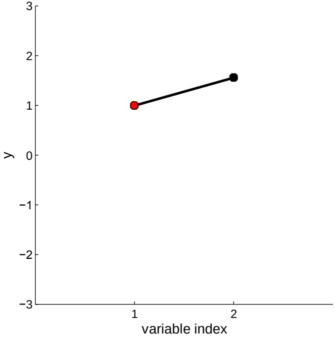
$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



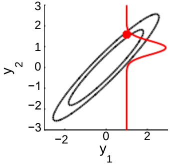
New visualisation

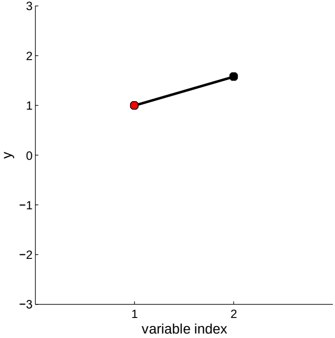


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$



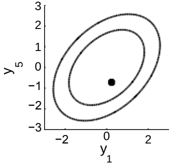
New visualisation



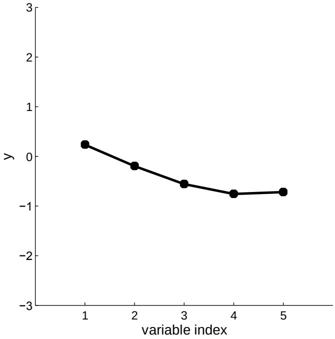


$$\Sigma = \begin{matrix} 1 & .9 \\ .9 & 1 \end{matrix}$$

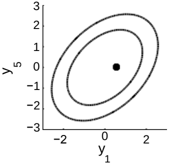
New visualisation

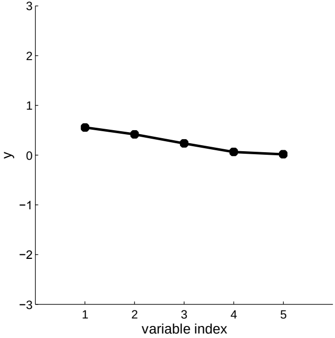


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



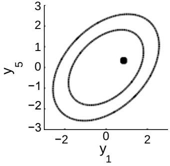
New visualisation

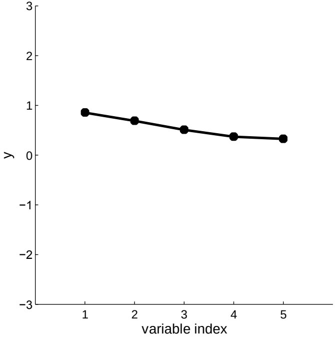




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

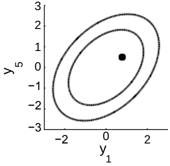
New visualisation

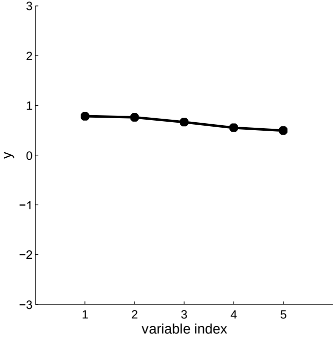




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

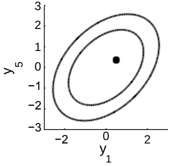
New visualisation



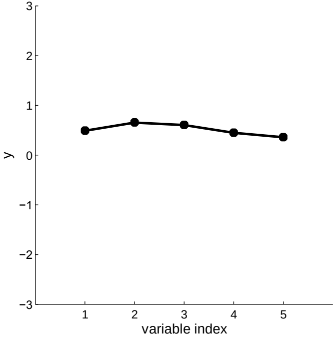


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

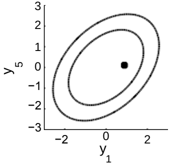
New visualisation



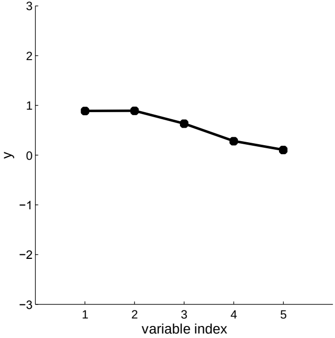
$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



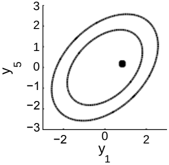
New visualisation

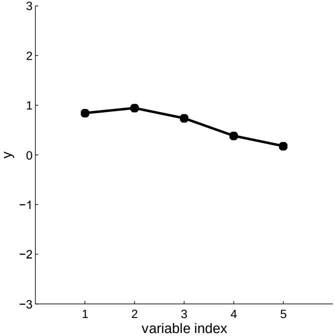


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



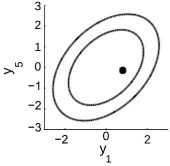
New visualisation

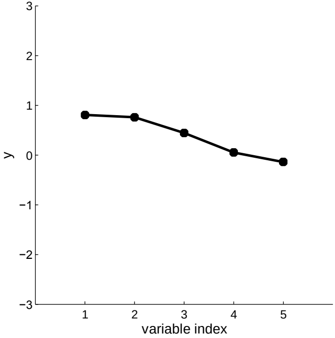




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

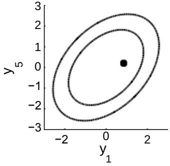
New visualisation

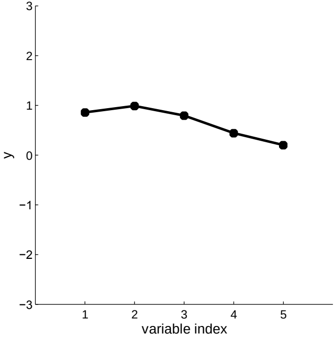




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

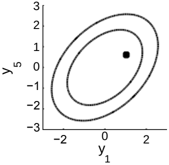
New visualisation



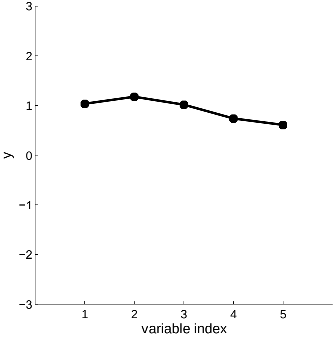


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

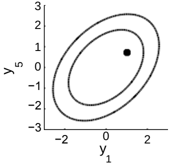
New visualisation



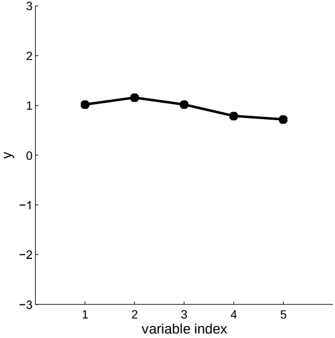
$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



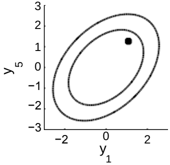
New visualisation

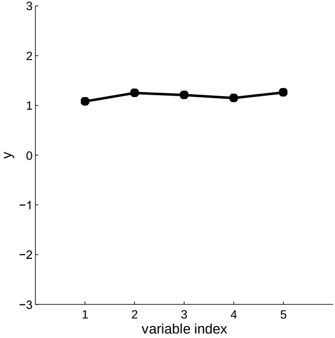


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



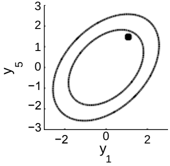
New visualisation

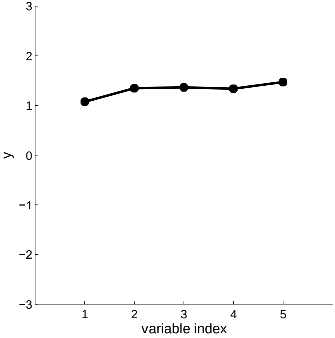




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

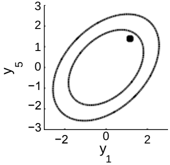
New visualisation

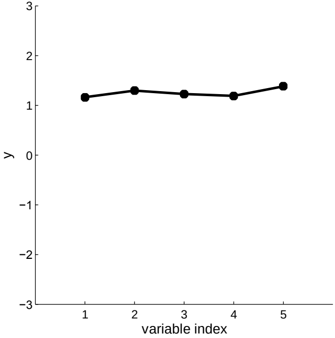




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

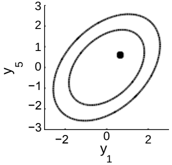
New visualisation

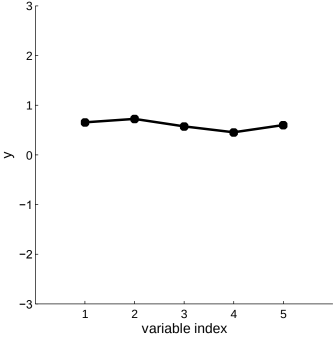




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

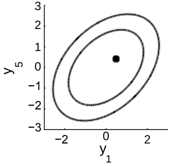
New visualisation



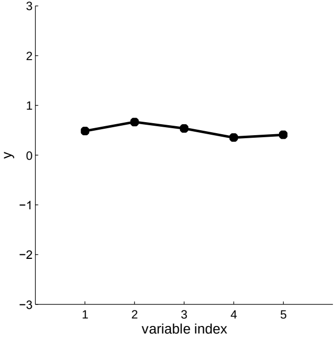


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

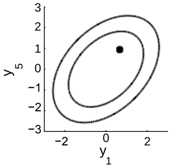
New visualisation



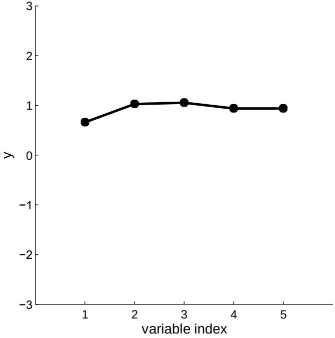
$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



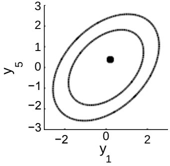
New visualisation

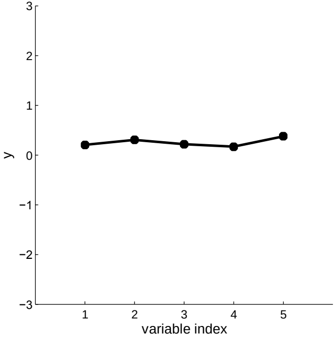


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



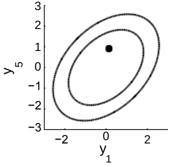
New visualisation

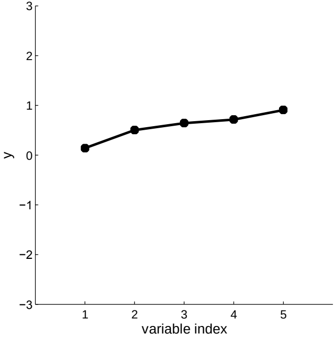




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

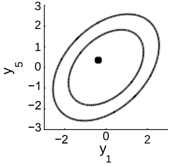
New visualisation



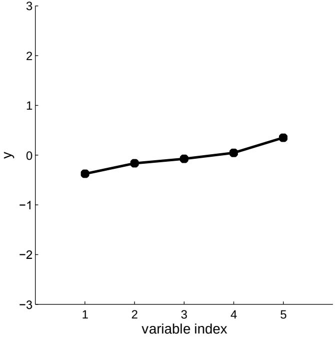


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

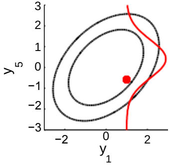
New visualisation



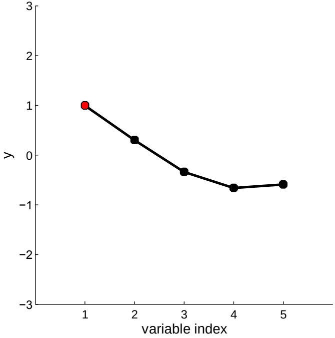
$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



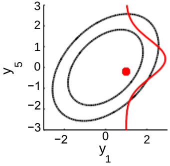
New visualisation

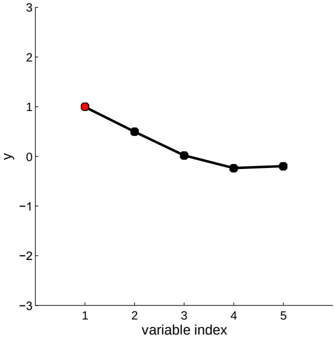


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



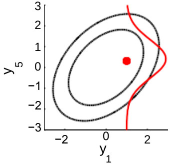
New visualisation

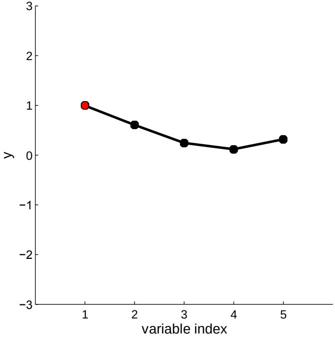




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

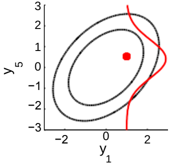
New visualisation

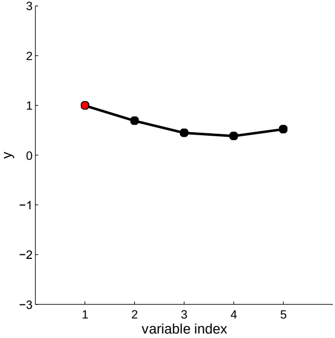




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

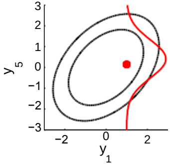
New visualisation

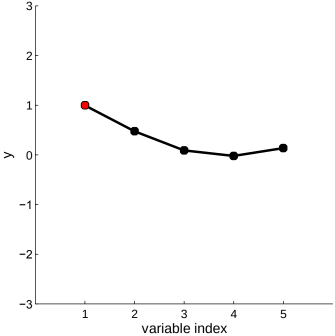




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

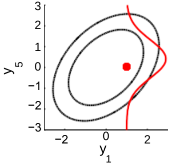
New visualisation



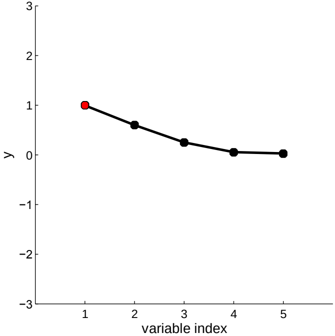


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

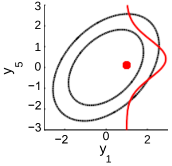
New visualisation

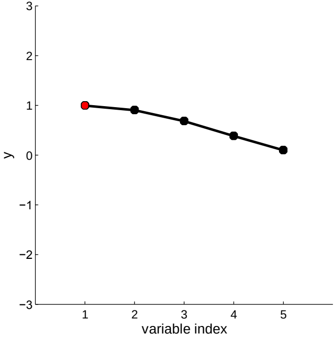


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



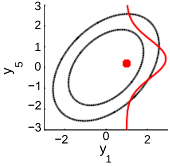
New visualisation

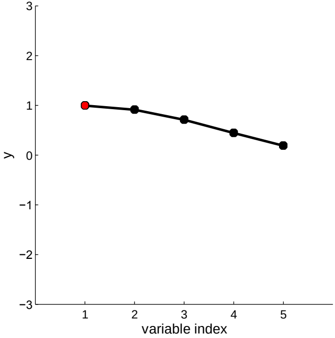




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

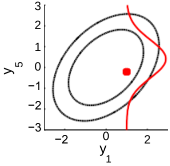
New visualisation

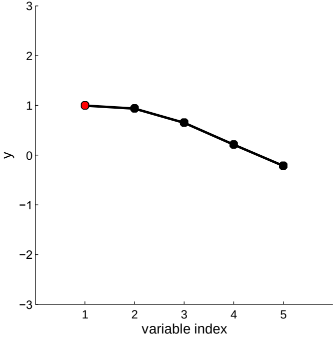




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

New visualisation

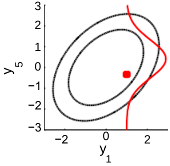


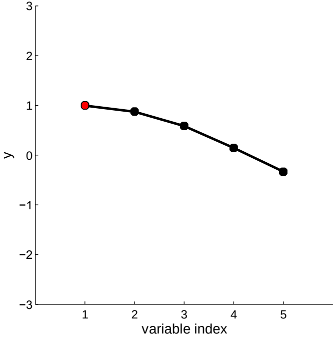


$\Sigma =$

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

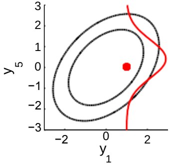
New visualisation



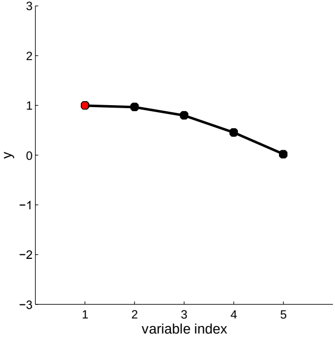


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

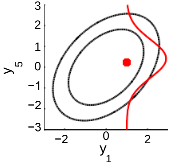
New visualisation

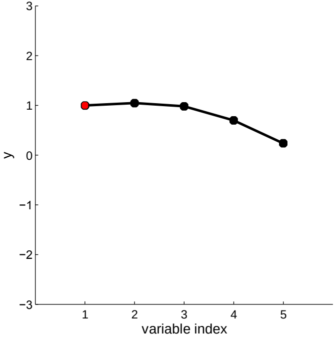


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



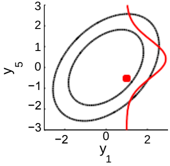
New visualisation

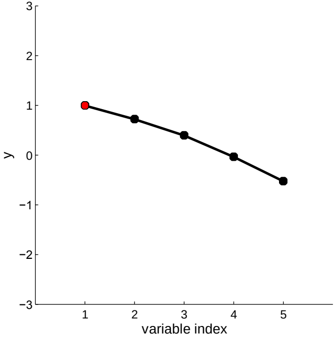




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

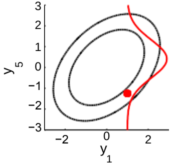
New visualisation

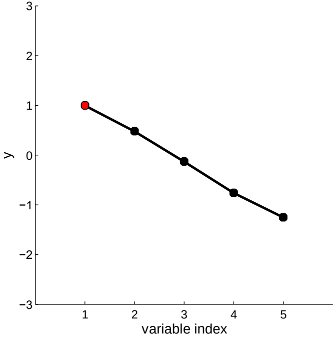




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

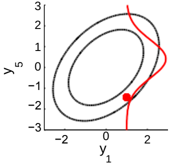
New visualisation

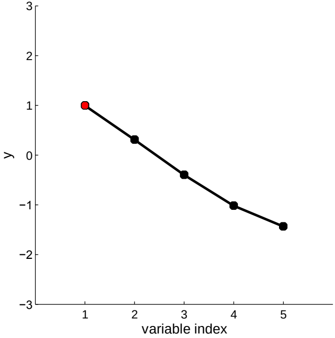




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

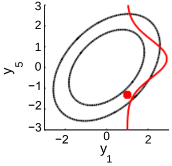
New visualisation



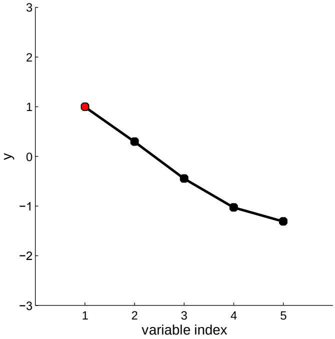


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

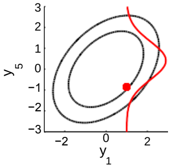
New visualisation

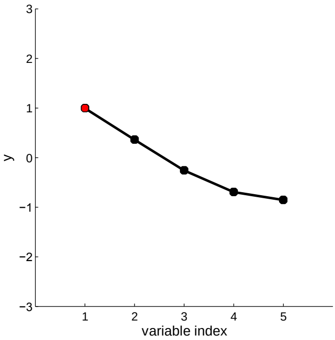


$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$



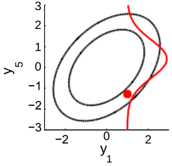
New visualisation

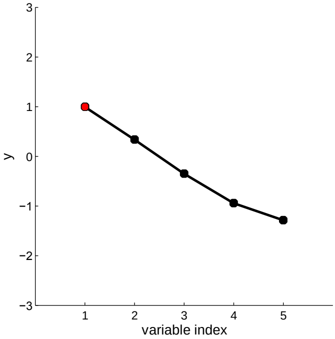




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

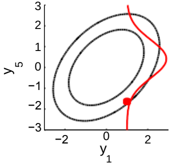
New visualisation

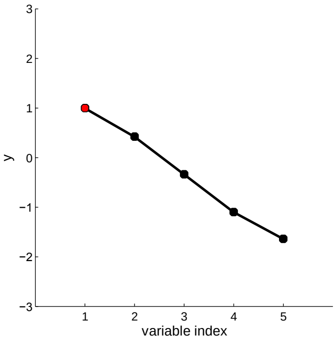




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

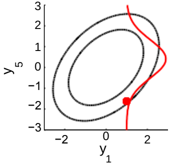
New visualisation

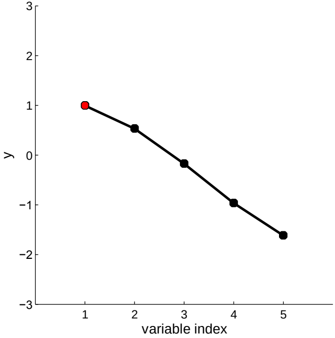




$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

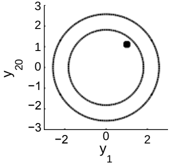
New visualisation

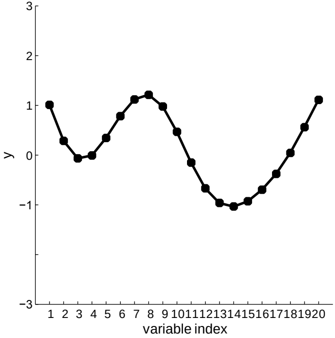




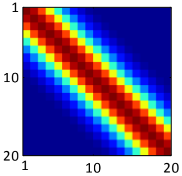
$$\Sigma = \begin{matrix} 1 & .9 & .8 & .6 & .4 \\ .9 & 1 & .9 & .8 & .6 \\ .8 & .9 & 1 & .9 & .8 \\ .6 & .8 & .9 & 1 & .9 \\ .4 & .6 & .8 & .9 & 1 \end{matrix}$$

New visualisation

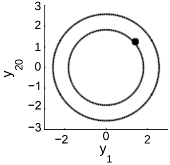




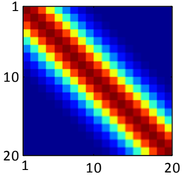
$\Sigma =$

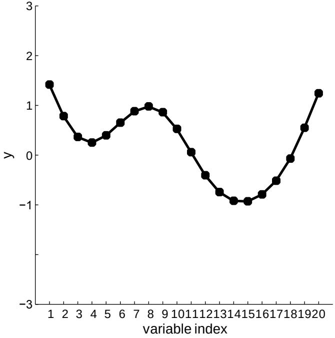


New visualisation

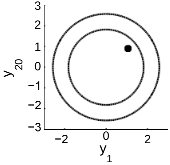


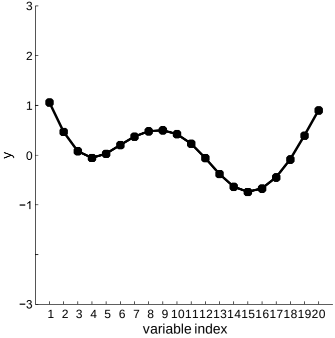
$\Sigma =$



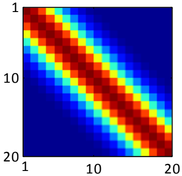


New visualisation

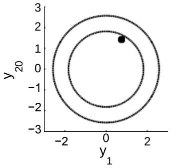




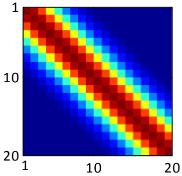
$\Sigma =$

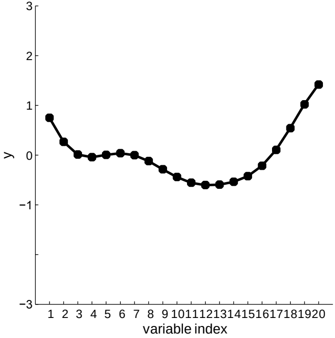


New visualisation

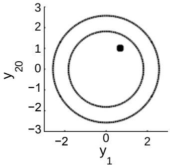


$\Sigma =$

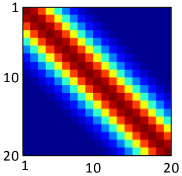


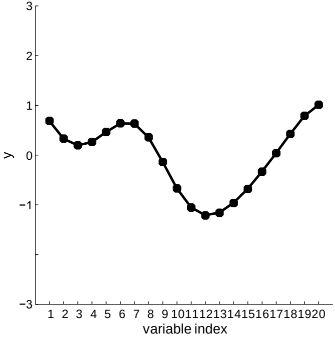


New visualisation

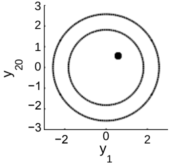


$\Sigma =$

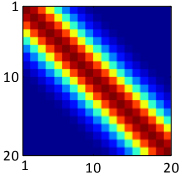


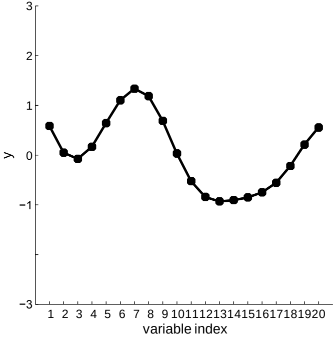


New visualisation

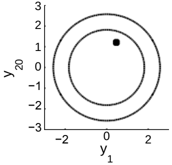


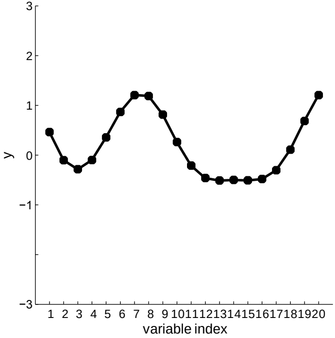
$\Sigma =$



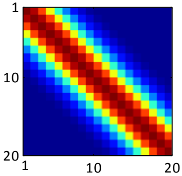


New visualisation

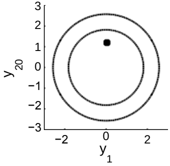




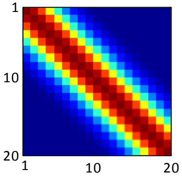
$\Sigma =$

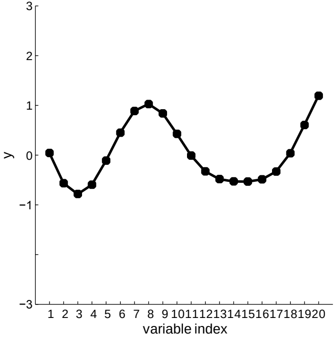


New visualisation

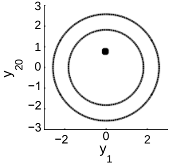


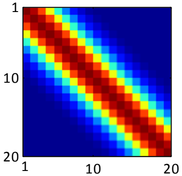
$\Sigma =$

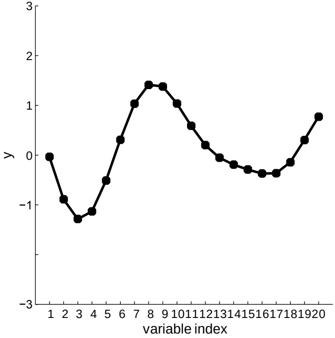




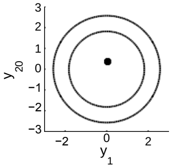
New visualisation

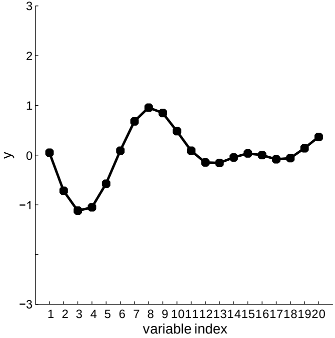


$\Sigma =$ 

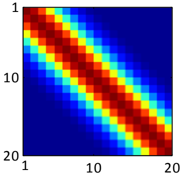


New visualisation

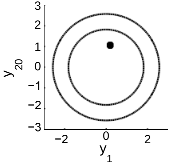


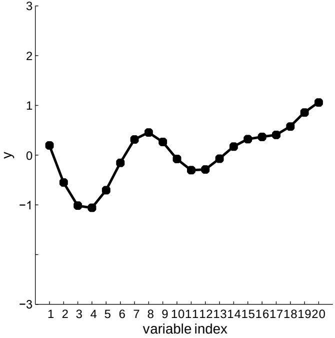


$\Sigma =$

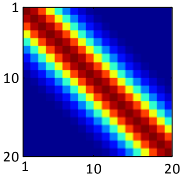


New visualisation

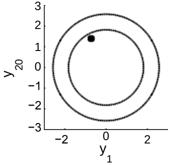


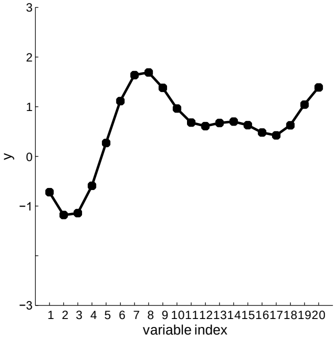


$\Sigma =$

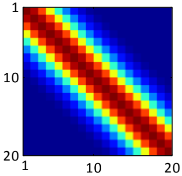


New visualisation

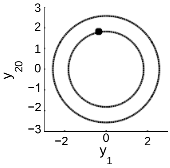




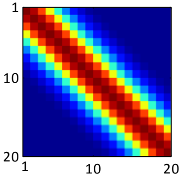
$\Sigma =$

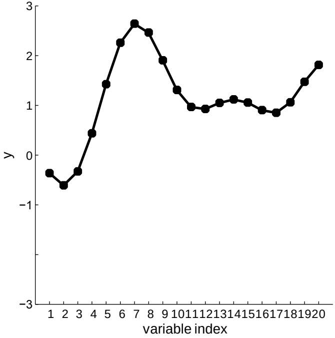


New visualisation

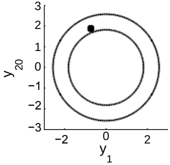


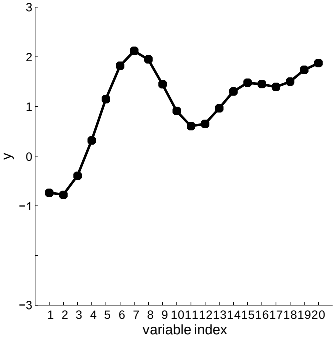
$\Sigma =$



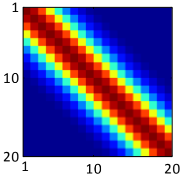


New visualisation

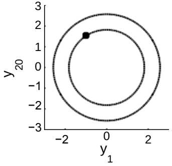


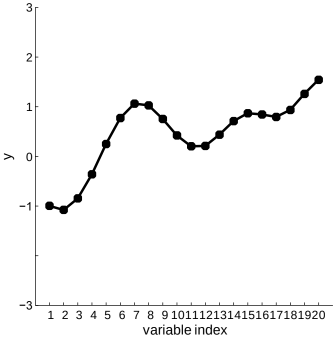


$\Sigma =$

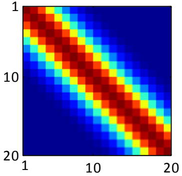


New visualisation

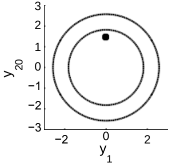




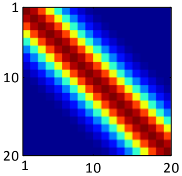
$\Sigma =$

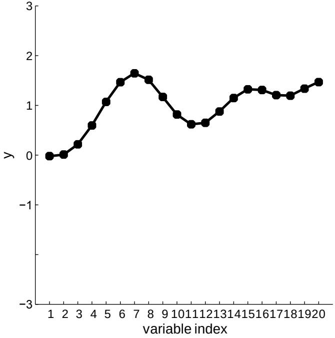


New visualisation

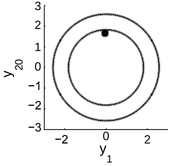


$\Sigma =$

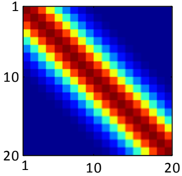


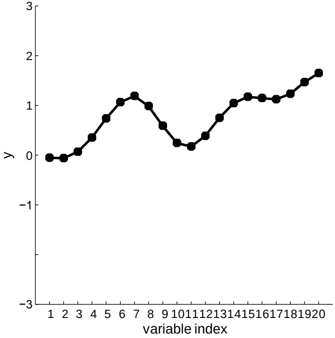


New visualisation

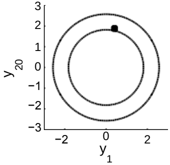


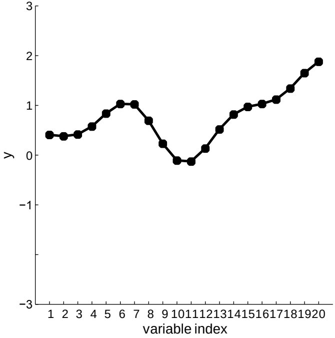
$\Sigma =$



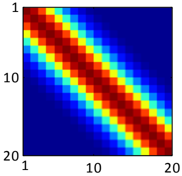


New visualisation



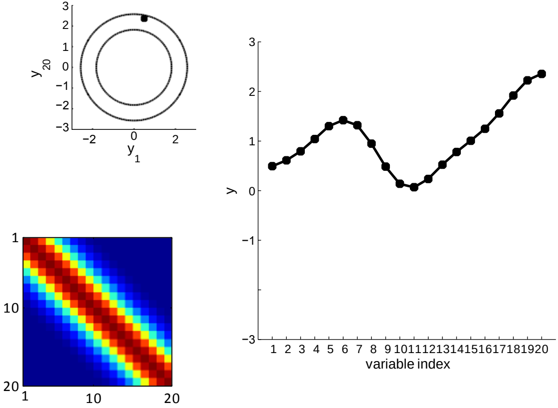


$\Sigma =$



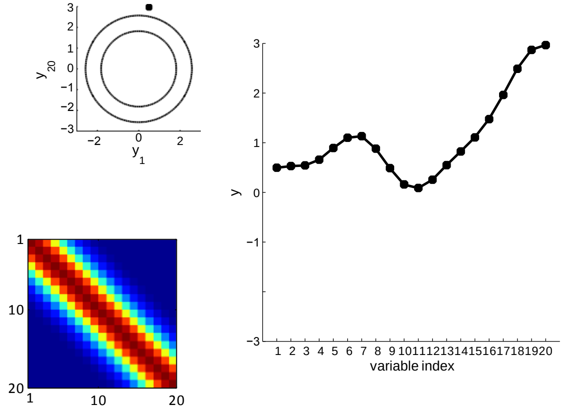
New visualisation

$\Sigma =$

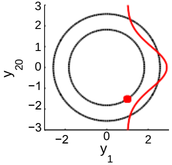


New visualisation

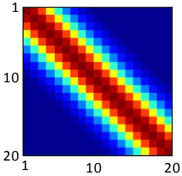
$\Sigma =$

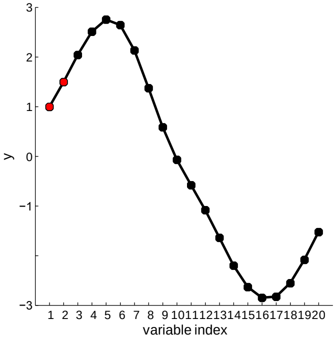


New visualisation

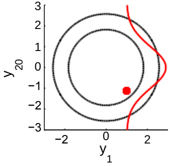


$\Sigma =$

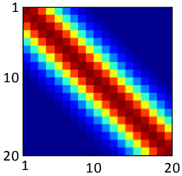


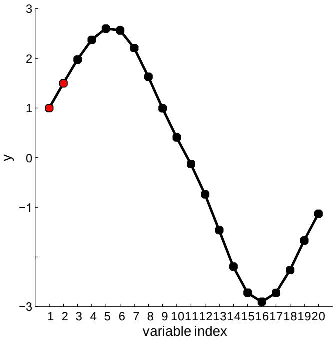


New visualisation

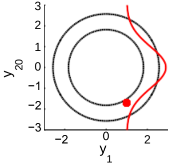


$\Sigma =$

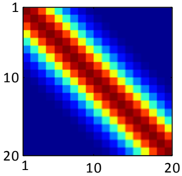


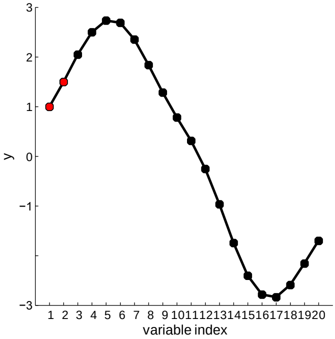


New visualisation

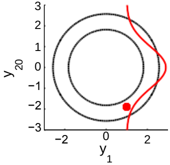


$\Sigma =$

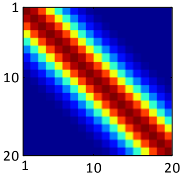


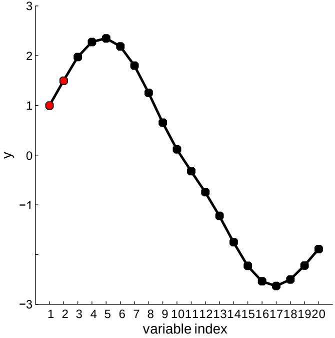


New visualisation

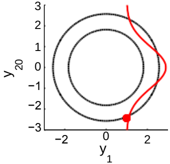


$\Sigma =$

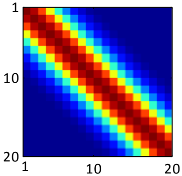


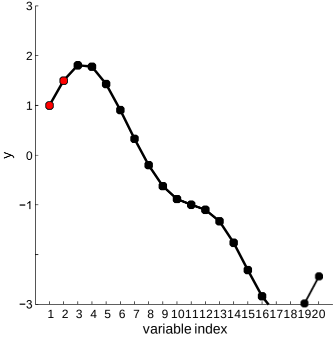


New visualisation

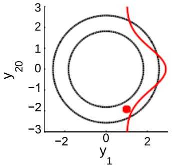


$\Sigma =$

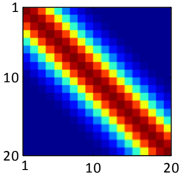


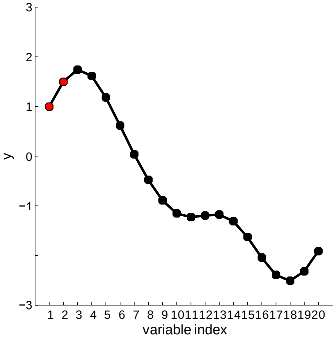


New visualisation

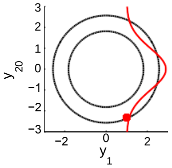


$\Sigma =$

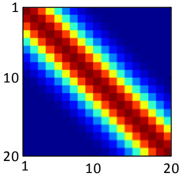


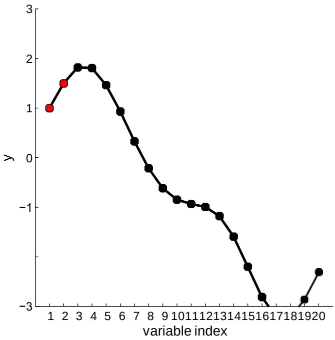


New visualisation

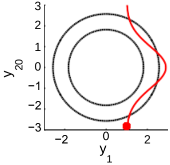


$\Sigma =$

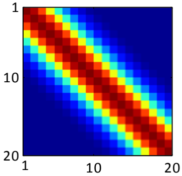


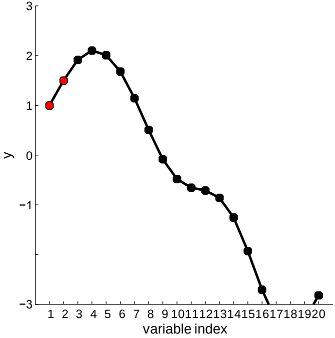


New visualisation

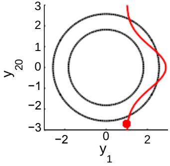


$\Sigma =$

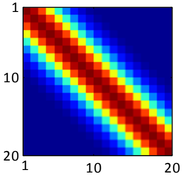


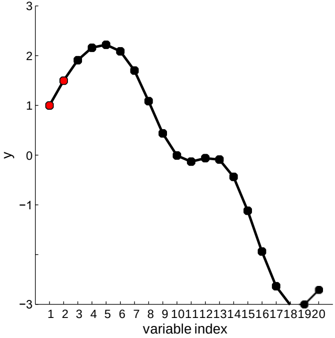


New visualisation

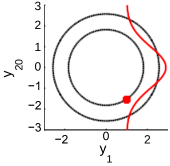


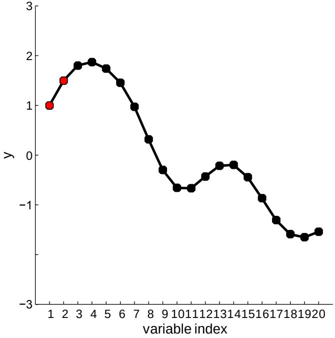
$\Sigma =$



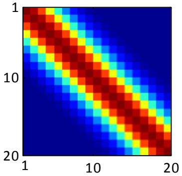


New visualisation

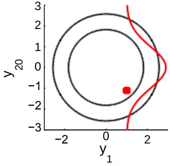




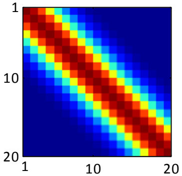
$\Sigma =$

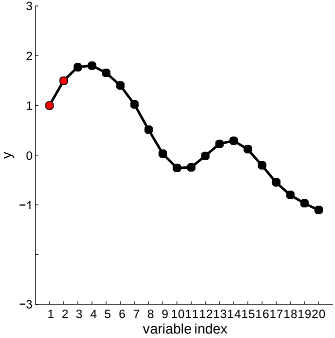


New visualisation

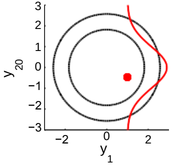


$\Sigma =$

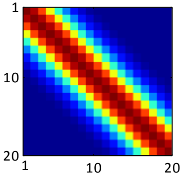


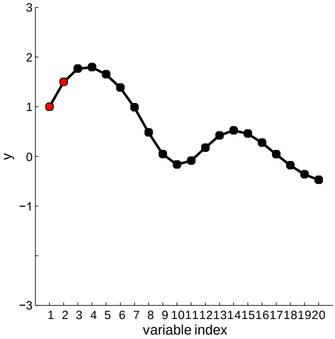


New visualisation

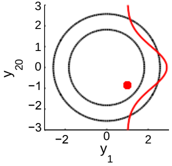


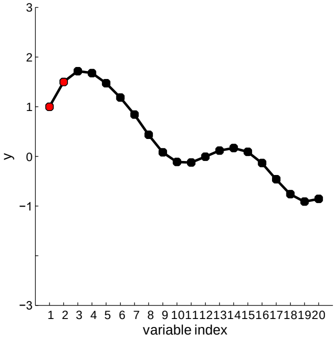
$\Sigma =$



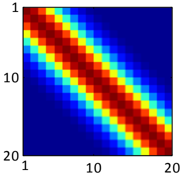


New visualisation

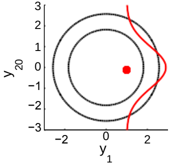


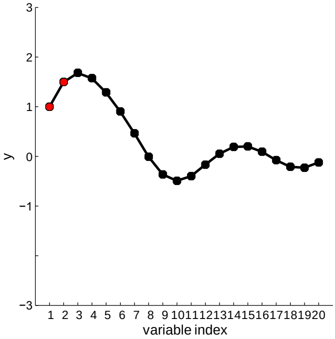


$\Sigma =$

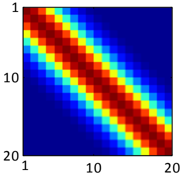


New visualisation

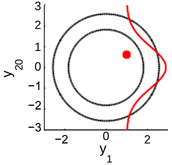




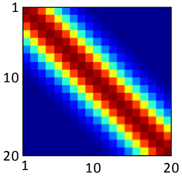
$\Sigma =$

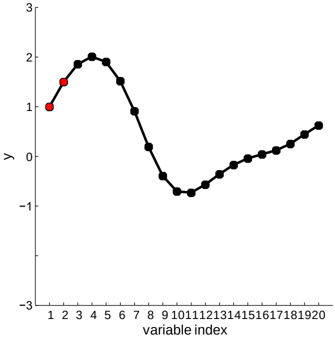


New visualisation

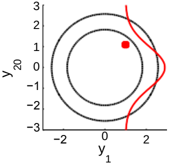


$\Sigma =$

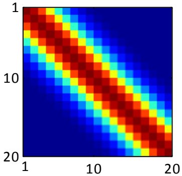


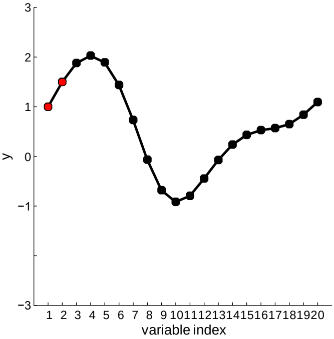


New visualisation

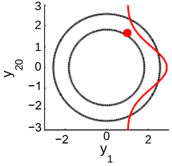


$\Sigma =$

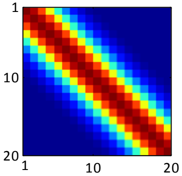


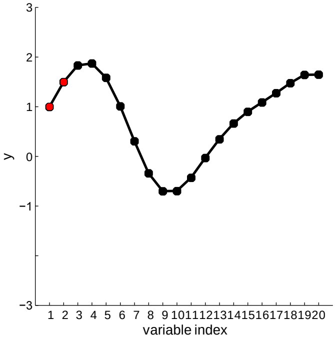


New visualisation

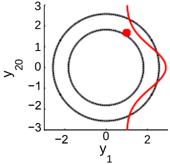


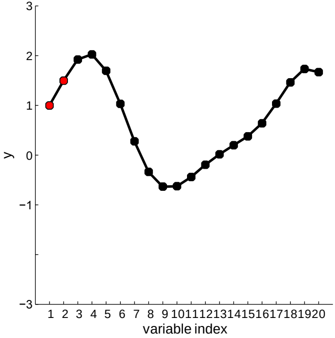
$\Sigma =$



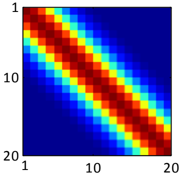


New visualisation

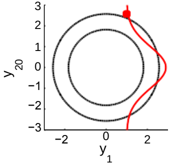




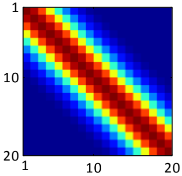
$\Sigma =$

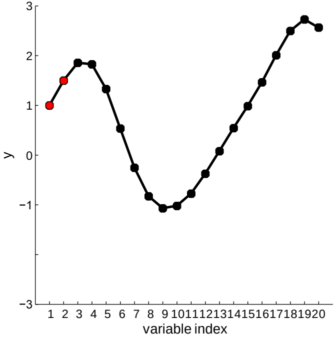


New visualisation

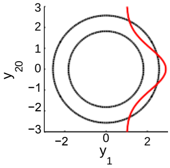


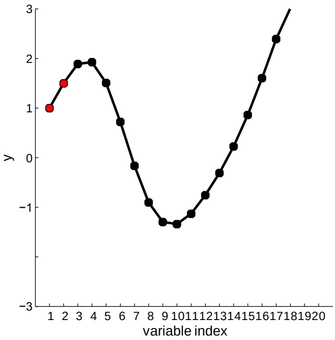
$\Sigma =$



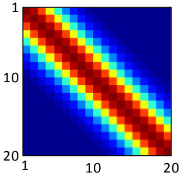


New visualisation



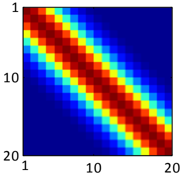


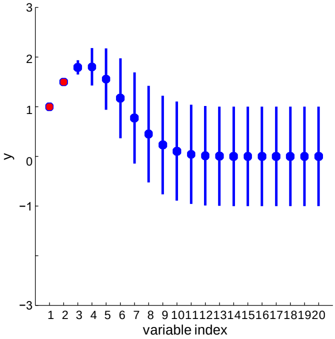
$\Sigma =$



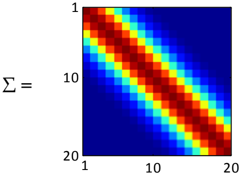
New visualisation

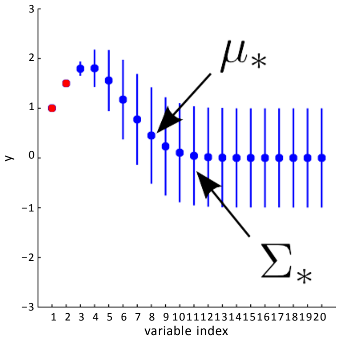
$\Sigma =$





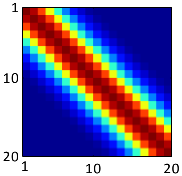
New visualisation

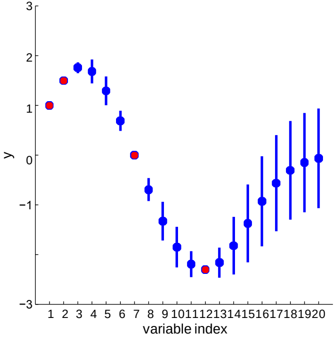




New visualisation

$\Sigma =$



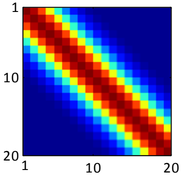


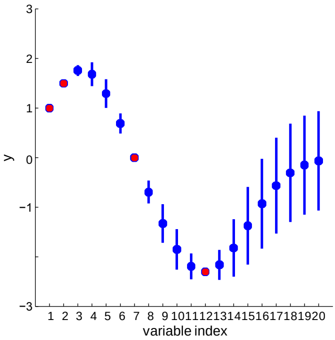
Regression using Gaussians

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\boldsymbol{\sigma}_y^2$$

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

$\Sigma =$



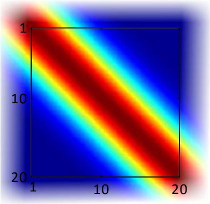


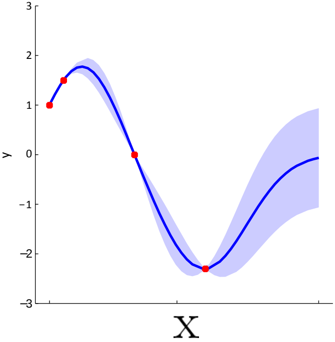
Regression: probabilistic inference in function space

$$\Sigma(x_1, x_2) = K(x_1, x_2) + \mathbf{I}\boldsymbol{\sigma}_y^2$$

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

$\Sigma =$





Mathematical Foundations: Definition

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

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

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

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

A Gaussian process is fully specified by a mean function $m(\mathbf{x})$ and covariance function $K(\mathbf{x}, \mathbf{x}^0)$:

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}^0)), \quad \text{indices} \quad \mathbf{x}$$

- Covariance function defines our prior belief about the function shape

Mathematical Foundations: Marginalisation

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

Mathematical Foundations: Marginalisation

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

We are saved by the marginalisation property:

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

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

Mathematical Foundations: Marginalisation

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

We are saved by the marginalisation property:

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

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

Mathematical Foundations: Marginalisation

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

We are saved by the marginalisation property:

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

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

Mathematical Foundations: Marginalisation

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

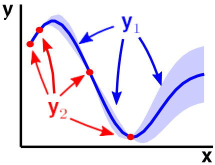
We are saved by the marginalisation property:

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

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

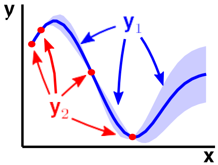
$\implies$ Only need to represent finite dimensional projections of GPs on computer.

Mathematical Foundations: Prediction



Mathematical Foundations: Prediction

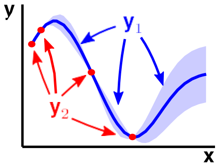
$$p(\mathbf{y}_1, \mathbf{y}_2) = \mathcal{N}\left(\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} \mathrm{A} & \mathrm{B} \\ \mathrm{B}^{\mathrm{T}} & \mathrm{C} \end{bmatrix} \right)$$



Mathematical Foundations: Prediction

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

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

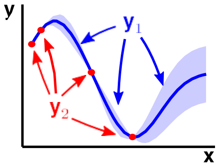


Mathematical Foundations: Prediction

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

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

$$\Longrightarrow \quad p(\mathbf{y}_1 | \mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1 ; \mathbf{a} + \mathrm{BC}^{-1}(\mathbf{y}_2 - \mathbf{b}), \mathrm{A} - \mathrm{BC}^{-1}\mathrm{B}^\mathrm{T})$$

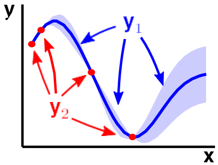


Mathematical Foundations: Prediction

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

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

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



predictive mean

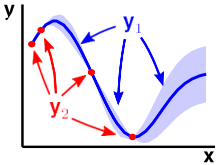
$$\mu_{\mathbf{y}_1|\mathbf{y}_2} = \mathbf{a} + \mathrm{B}\mathrm{C}^{-1}(\mathbf{y}_2 - \mathbf{b})$$

Mathematical Foundations: Prediction

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

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

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



predictive mean

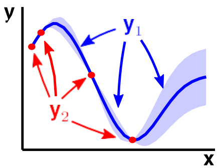
$$\mu_{\mathbf{y}_1|\mathbf{y}_2} = \mathbf{a} + \mathsf{B}\mathsf{C}^{-1}(\mathbf{y}_2 - \mathbf{b})$$

$$= \mathsf{B}\mathsf{C}^{-1}\mathbf{y}_2$$

Mathematical Foundations: Prediction

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

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



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

predictive mean

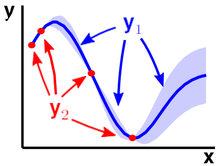
$$\begin{aligned} \mu_{\mathbf{y}_1|\mathbf{y}_2} &= \mathbf{a} + \mathsf{B}\mathsf{C}^{-1}(\mathbf{y}_2 - \mathbf{b}) \\ &= \mathsf{B}\mathsf{C}^{-1}\mathbf{y}_2 \\ &= \mathsf{W}\mathbf{y}_2 \end{aligned}$$

Mathematical Foundations: Prediction

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

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

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



predictive mean

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

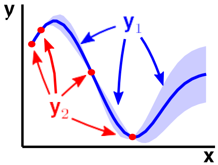
$$= \mathsf{B}\mathsf{C}^{-1}\mathbf{y}_2$$

$$= \mathsf{W}\mathbf{y}_2$$

linear in the data

Mathematical Foundations: Prediction

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



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

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

predictive mean

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

linear in the data

predictive covariance

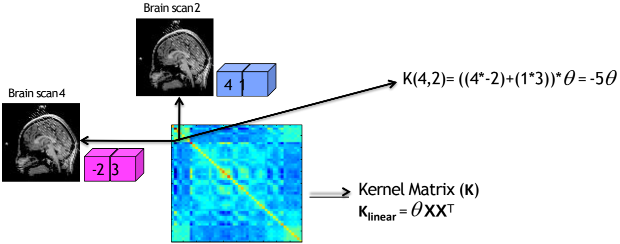
$$\Sigma_{\mathbf{y}_1|\mathbf{y}_2} = \mathsf{A} - \mathsf{B}\mathsf{C}^{-1}\mathsf{B}^\top$$

predictive uncertainty = prior uncertainty − reduction in uncertainty

predictions more confident than prior

Covariance = Kernel

- **K** can be thought of in a similar way to the kernels in eg. SVM, ie. entry *i* , *j* is the similarity of two images



- In GPs, the value of the similarity for two images defines the prior knowledge of how similar the function values are

- As for other algorithms eg. Kernel Ridge Regression we tend to use a linear kernel in neuroimaging to avoid overfitting

Link to Bayesian Linear Regression

Bayesian linear regression:

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T\mathbf{w} + b$$

where $\mathbf{w} \sim N\,(\mathbf{0}, \sigma_\omega^2 I)$, $b \sim N\,(\mathbf{0}, \sigma_b^2)$

Link to Bayesian Linear Regression

Bayesian linear regression:

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T\mathbf{w} + b$$

where $\mathbf{w} \sim N\,(\mathbf{0}, \sigma_\omega^2 I)$, $b \sim N\,(\mathbf{0}, \sigma_b^2)$

- $f(\mathbf{x}_i, \mathbf{w})$, for all *i* is Gaussian
- The joint distribution of any set of function values is Gaussian

Link to Bayesian Linear Regression

Bayesian linear regression:

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T\mathbf{w} + b$$

where $\mathbf{w} \sim N(\mathbf{0}, \sigma_\omega^2 I)$, $b \sim N(\mathbf{0}, \sigma_b^2)$

- $f(\mathbf{x}_i, \mathbf{w})$, for all i is Gaussian
- The joint distribution of any set of function values is Gaussian

→ We are specifying a Gaussian process prior on a function

Link to Bayesian Linear Regression

$cov(f_1, f_2)$ = E [f_1, f_2] = E [f_1]-E[f_2]

= E [($\mathbf{x_1}^T\mathbf{w}+b$) ($\mathbf{x_2}^T\mathbf{w}+b$)$^\mathrm{T}$]

= $\sigma_\omega^2\ \mathbf{x_1}\mathbf{x_2}^T + \sigma_b^2$

= K($\mathbf{x_1}, \mathbf{x_2}$)

Link to Bayesian Linear Regression

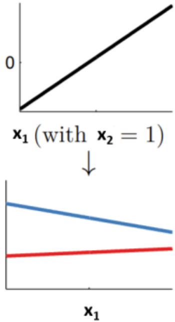
$cov(f_1, f_2) = \mathrm{E}[f_1, f_2] = \mathrm{E}[f_1]\text{-}\mathrm{E}[f_2]$

$= \mathrm{E}[(\mathbf{x_1}^T\mathbf{w}+b)(\mathbf{x_2}^T\mathbf{w}+b)^\mathrm{T}]$

$= \sigma_\omega^2\ \mathbf{x_1}\mathbf{x_2}^T + \sigma_b^2$

$= \mathrm{K}(\mathbf{x_1}, \mathbf{x_2})$

Linear kernel!



Predictive mean of Bayesian Linear Regression

= (*using the same hyp. params*)

Kernel Ridge Regression

Predictive mean of Bayesian Linear Regression

= (*using the same hyp. params*)

Kernel Ridge Regression

↓

Predictive mean of Gaussian Processes with Linear Kernels

=

Kernel Ridge Regression

- *can also be shown numerically*

Difference?

- Hyperparameter optimisation
  - KRR: Cross-validation
  - Maximum marginal likelihood estimation
    - Gradient decent
    - Grid search
    - Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) algorithm
      - GPy toolbox (for PRoNTo)

Some other covariance functions

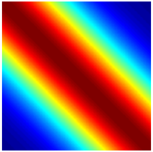
RBF (Radial Basis Function)

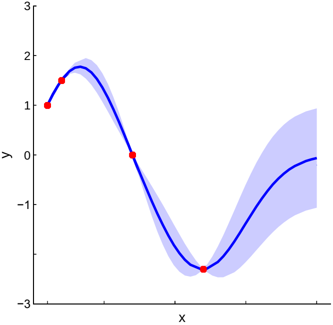
$$K(x_1, x_2) = \boldsymbol{\sigma}^2 \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

a.k.a.

Exponentiated Quadratic

Squared exponential

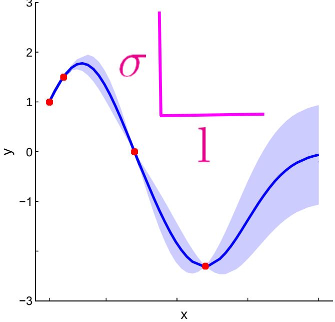
$\Sigma =$ 



RBF (Radial Basis Function)

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

vertical-scale horizontal-scale

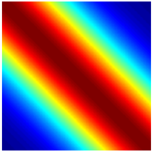


a.k.a.

Exponentiated Quadratic

Squared exponential

$\Sigma =$

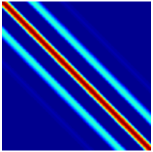


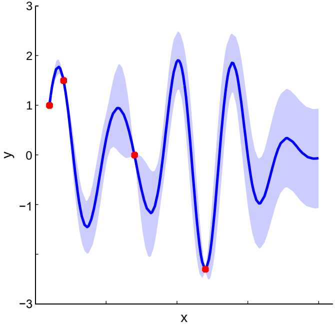
Periodic

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

sinusoid × squared exponential

$\Sigma =$





Relevance Vector Machines

- The relevance vector machine is a type of sparse Bayesian model for regression and classification (Tipping, 2001)
- For regression, the RVM uses the same Gaussian likelihood as the GP and applies a prior over the weights of the form:

$$p(\mathbf{w}|a) = \prod_i N(w|0_i, q_i)^{-1}$$

- The a_i are scaling parameters which determine the ”relevance” of each sample or voxel (MacKay, 2003). These are given flat Gamma priors.

Relevance Vector Machines

- The relevance vector machine is a type of sparse Bayesian model for regression and classification (Tipping, 2001)
- For regression, the RVM uses the same Gaussian likelihood as the GP and applies a prior over the weights of the form:

$$p(\mathbf{w}|a) = \prod_i N(w|0_i, q_i)^{-1}$$

- The a_i are scaling parameters which determine the "relevance" of each sample or voxel (MacKay, 2003). These are given flat Gamma priors.
- The RVM forces the posterior probability for the weights to concentrate on only a few of the samples/voxels. Samples/voxels with a low weight are pruned from the model (→ **Sparsity**)
- The RVM is not solvable in closed form and requires numerical approximation(s) to the posterior distribution

Conclusions

- Probabilistic approaches to pattern classification are complementary to alternative methods

Conclusions

- Probabilistic approaches to pattern classification are complementary to alternative methods
- They share many features with conventional approaches (e.g. penalised linear models)

Conclusions

- Probabilistic approaches to pattern classification are complementary to alternative methods
- They share many features with conventional approaches (e.g. penalised linear models)
- They aim to be honest about uncertainty at all stages of analysis (**coherence**)

Conclusions

- Probabilistic approaches to pattern classification are complementary to alternative methods
- They share many features with conventional approaches (e.g. penalised linear models)
- They aim to be honest about uncertainty at all stages of analysis (**coherence**)
- This provides a number of advantages, especially for clinical applications, e.g.:
    - Provide a natural way to include existing information (priors)
    - To compensate for variable class frequencies
    - To represent variabilities in illness severity

Conclusions

- Probabilistic approaches to pattern classification are complementary to alternative methods
- They share many features with conventional approaches (e.g. penalised linear models)
- They aim to be honest about uncertainty at all stages of analysis (**coherence**)
- This provides a number of advantages, especially for clinical applications, e.g.:
  - Provide a natural way to include existing information (priors)
  - To compensate for variable class frequencies
  - To represent variabilities in illness severity
- However they also have disadvantages
  - Estimating probability distributions requires more computation than just estimating a decision function.
  - Some methods may not scale as well to large datasets ($O(n^3)$)

Thanks for your attention!