CPSC 535D - Statistical Computing & Monte Carlo Methods

Arnaud Doucet

Email: arnaud@cs.ubc.ca

16th January 2007

- Slides available on the Web before lectures:
`www.cs.ubc.ca/~arnaud/CS535D_2007.html`

- Textbook: C.P. Robert & G. Casella, *Monte Carlo Statistical Methods*, Springer, 2nd Edition.

- Additional lecture notes available on the Web.

- Textbooks which might also be of help:

  - A. Gelman, J.B. Carlin, H. Stern and D.B. Rubin, *Bayesian Data Analysis*, Chapman&Hall/CRC, 2nd edition.
  - C.P. Robert, *The Bayesian Choice*, Springer, 2nd edition.

1.1– Preliminaries

- Four assignements including programs (R or Matlab).
- Final project.
- Final weighting not determined (should be 50% or 60%).

1.2– Objectives of this course

- To provide an introduction to Bayesian statistics.

- To provide an introduction to modern computational methods used in statistics.

- To provide an introduction to complex (but realistic!) statistical models.

- At the end of this course you should be able to understand and fit complex models (and be able to assess if your analysis makes any sense!).

1.2– Objectives of this course

- Even if you have already followed the course by Kevin Murphy during the 1st term, you will still learn quite a few things.

- In this course, the emphasis is on computational methods.

- We will go through detailed case studies.

2.1– Overview of the Course Contents

- Introduction to Bayesian Statistics

$$\pi\left(\theta|\,x\right) = \frac{f\left(x|\,\theta\right)\pi\left(\theta\right)}{\int f\left(x|\,\theta\right)\pi\left(\theta\right)d\theta}.$$

- Explosion of Bayesian statistics over the past 15 years: approximately 30% of papers in top statistical reviews are about Bayesian statistics.

- Among the top 10 most cited mathematicians over the last 10 years, 5 are Bayesian statisticians!

- Over the last 5 years, 4 Copss Medals were awarded to Bayesian statisticians.

2.2– Why focusing on Bayesian Statistics?

• The Bayesian approach is very well-adapted to many application areas: bioinformatics, genetics, epidemiology, econometrics, machine learning, nuclear magnetic resonance etc.

• It allows one to incorporate in a principled way any prior information available on a given problem.

• Straightforward to handle missing data, outliers, censored data etc.

• It is a simple framework and, in my opinion, much simpler than "standard" approaches.

• It is honest and makes clear that any statistical analysis relies on a part of subjectivity.

2.3– Computational Methods

• Why have Bayesian statistics enjoyed such an increasing popularity over the last 15 years?

$\Rightarrow$ Implementation difficult and requires computational methods.

• For complex models, Bayesian methods require computing very high dimensional integrals.

• Deterministic methods are completely inefficient
$\Rightarrow$ Curse of dimensionality.

• Monte Carlo methods are the only possible way to address such problems.
$\Rightarrow$ Standard Monte Carlo methods are inefficient.

2.4– Monte Carlo Computational Methods

• Monte Carlo are stochastic algorithms with a wide range of applications in physics, chemistry, mechanics, optimization.

• Markov chain Monte Carlo (MCMC) are a very popular class of Monte Carlo algorithms

$\Rightarrow$ The Metropolis algorithm was named the top algorithm of the 20th century by mathematicians, computer scientists & physicists!

• Here you will learn MCMC algorithms and why they work (or don't work!).

2.5– Monte Carlo Computational Methods

- MCMC algorithms are iterative algorithms and hence inadequate for many problems of interest.

  - For massive datasets browsing repeatedly the data is too expensive.

  - For high-frequency volatility data or target tracking, users

    are impatient!

- Sequential Monte Carlo (SMC) also known as particle filters are a recent class of algorithms to address such probems.

3.1– Example: Capture-Recapture Experiments

• Example: You are interested in estimating the population size of the bears in BC.

• You capture some bears, mark them and release them.

• Later on, you capture more bears, mark them (some of them might be already marked) and release them... and so on.

• You build a probabilistic model and based on these data, you can come up with an estimate of the population size of the bears.

• Your model can include migration effects, birth/deaths of the individuals, etc.

3.1– Example: Capture-Recapture Experiments

- Individuals are observed once or several times.

- The repeated observations are used to infer the population size but also its dynamics.

- Numerous applications in biology and ecology (for estimating some species populations), in sociology and demography (for estimating the size of populations at risk), for fraud detection (phone, credit card etc.) or software debugging (total number of bugs).

3.1– Example: Capture-Recapture Experiments

- A Bayesian approach to such problems is natural.

- Prior on the population size.

- Probabilistic model to describe its evolution is a prior.

- We want to propagate uncertainty of our estimates from one stage to the next, missing data, etc.

- Bayesian MCMC approaches have become very popular for such problems.

3.2– Example: Regression and Classification

- y is called the response/outcome and $x = (x_1, \ldots, x_p)$ is a set of explanatory variables.

- Given n data $\{y^i, x^i\}$, we want to determine a model relating y to x.

- If $y \in \mathbb{R}^k$: regression problem.

3.3– Example: Regression and Classification

- Linear regression

$$
\begin{aligned}
y &= \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon \text{ where } \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right) \\
&= f\left(x_i\right) + \varepsilon
\end{aligned}
$$

- Example: Predict the weights of children given the weights of its parents.

- Analysis of Bayes linear model and connections to standard approaches.

- Prior selection, prediction and all that.

3.3– Example: Regression and Classification

- How to handle outliers?

- How to handle noisy explanatory variables?
[Most people claim they are thinner than they really are].

- How to handle missing data?
[Data are never perfect]

- How to handle censored/categorical data?
[Heterogeneous databases]

3.3– Example: Regression and Classification

• Most data are not normal and do not depend linearly on the explanatory variables.

• Consider a dichotomous model where the outcome $y \in \{0, 1\}$ (death in a medical study, unemployement in a socioeconomic study, migration in a capture-recapture study, etc.).
$\Rightarrow$ Normality assumption just does not make any sense.

• Logit regression

$$\Pr\left(y = 1 \middle| x\right) = \frac{\exp\left(f\left(x_i\right)\right)}{1 + \exp\left(f\left(x_i\right)\right)}.$$

• Probit regression

$$\Pr\left(y = 1 \middle| x\right) = \Phi\left(f\left(x_i\right)\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{f(x_i)} \exp\left(-\frac{u^2}{2}\right) du.$$

3.3– Example: Regression and Classification

- Assume you have access to some counting data; e.g. Gamma ray counts for geology, network traffic inference using link count data, astronomy, panel count data (epilepsy, number of patents etc.)

- Poisson regression might be used

$$y|\, x \sim \mathcal{P}\left(\exp\left(f\left(x_i\right)\right)\right)$$

3.3– Example: Regression and Classification

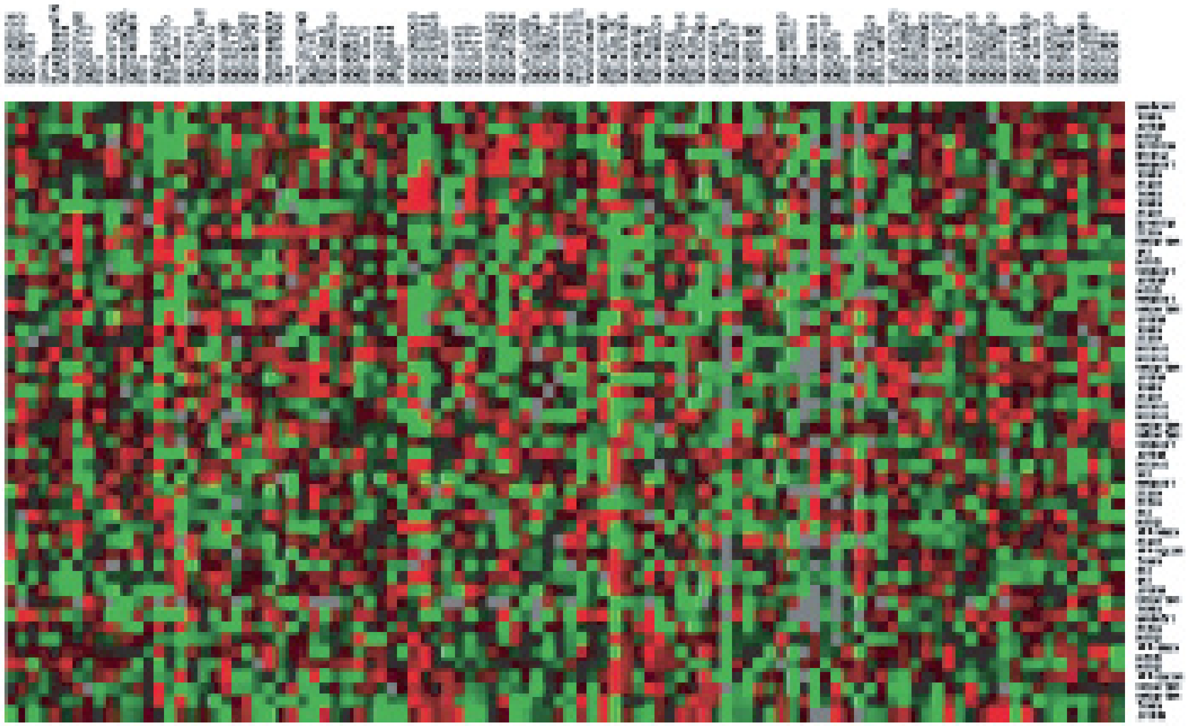
• Assume the outcome $y \in \{1, ..., k\}$ but a ranking information is available; e.g. $k = 5$ and the outcome corresponds to a grade level $E < D < C < B < A$.

• Applications include consumers rating, gene ranking, dynamic ranking of chess players or football teams.

• In such case, we can have for example

$$\Pr(y|x) = \Phi(\alpha_y - f(x_i)) - \Phi(\alpha_{y-1} - f(x_i)).$$

3.4– Example: Variable Selection



Row = examples, Columns = features (genes).

Large p, small n.

3.4– Example: Variable Selection

• If p is large, we might have little information to obtain precise estimator; large number of genes p and small number of samples n ($p >> n$).

• In other words, we will increase the explanatory scope of the regression model but not necessarily its explanatory “power”.

• Moreover some of the explanatory variables might be useless: e.g. the output is the temperature and an explanatory variable is your weight.

• It is important to be able to decide which explanatory variables should be kept in a model that balances good explanatory power with good estimation performances.

3.4– Example: Variable Selection

- This is a decision problem: all potential models have to be considered in parallel against a criterion that ranks them.

- There are 2^p potential models.

- Stepwise greedy methods or LASSO can be used

$\Rightarrow$ Sensitive to initialization and/or Regularization parameter.

- Bayesian variable selection/model averaging but requires Monte Carlo if $2^p >> 1$.

3.5– Example: Nonlinear Regression

- Wavelet regression

$$f(x) = \sum_{i=1}^{k} \beta_i \Psi_i(x) + \varepsilon$$

- Radial basis regression

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i K\left(\left\|x - x^i\right\|\right) + \varepsilon$$

- P-splines: unknown numbler of knots, location of the knots etc.

- Neural nets.

3.5– Example: Nonlinear Regression

- Harmonic regression

$$y_n = \sum_{i=1}^{k} \beta_i \cos\left(2\pi f_i n + \phi_i\right) + b_n$$

Signal represented as a sum of sinusoidal components (Fourier basis).

- Widely used in NMR, music signal processing, radar etc.

- In this case, $\left\{k, \beta_{1:k}, f_{1:k}, \sigma^2\right\}$ are unknown.

3.6– The need for computational methods

- For the simplest Bayesian linear model, inference can be performed in closed-form.

- For all the other problems, advanced Monte Carlo methods will be required.

- It can be argue that models should be developed independently of the algorithms fitting them.

- Pragmatically, this is not really true. What's the point of having a super model which is impossible to fit?

- Monte Carlo methods allow users to define much more flexible models but they have limitations too.

3.7– Example: Mixture Models and Hidden Markov Models

- Finite Mixture Models

$$f\left(x\right)=\sum_{i=1}^{k}p_{i}f_{i}\left(x\right)$$

- Finite mixture models appear everywhere and are used for

  modelling multimodal distributions.

  allows to model populations as mixture of several subpopulations.

  data clustering.

3.7– Example: Mixture Models and Hidden Markov Models

- Example: Finite Mixture of Gaussians

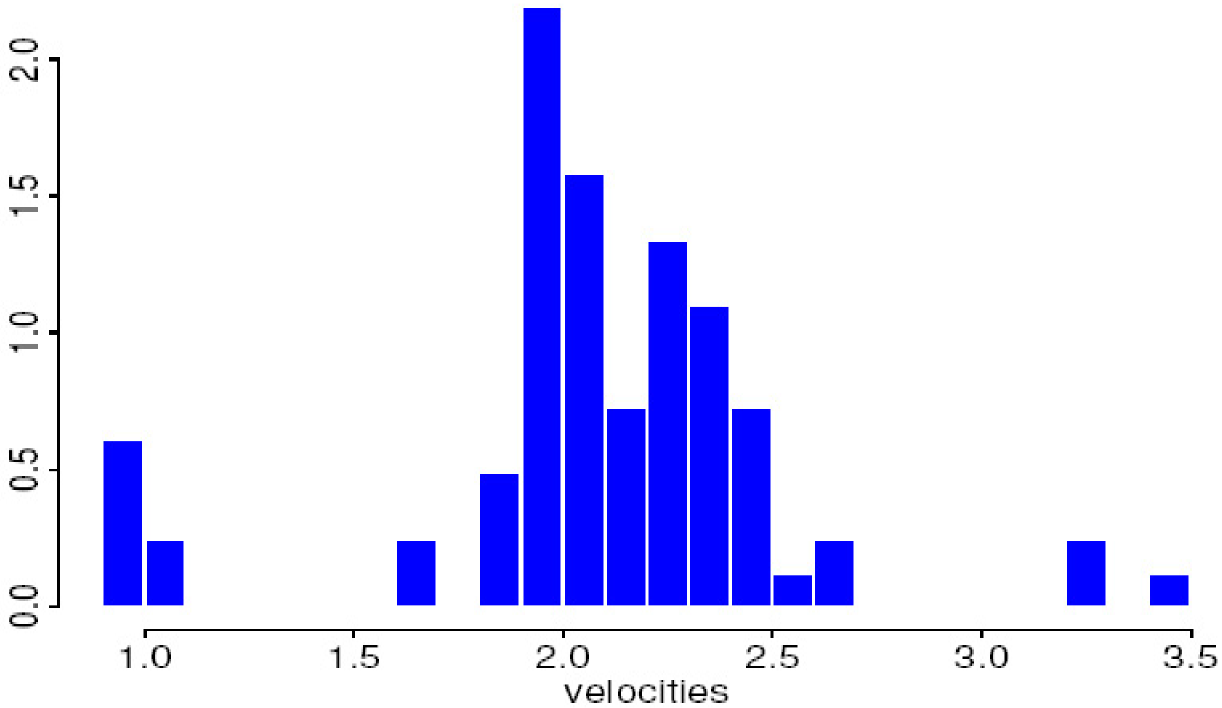
$$f\left(x|\,\theta\right)=\sum_{i=1}^{k}p_i\mathcal{N}\left(x;\mu_i,\sigma_i^2\right)$$

where $\theta=\left\{\mu_i,\sigma_i^2,p_i\right\}_{i=1,\ldots,k}$ is estimated from some data $\left(x_1,...,x_n\right)$.

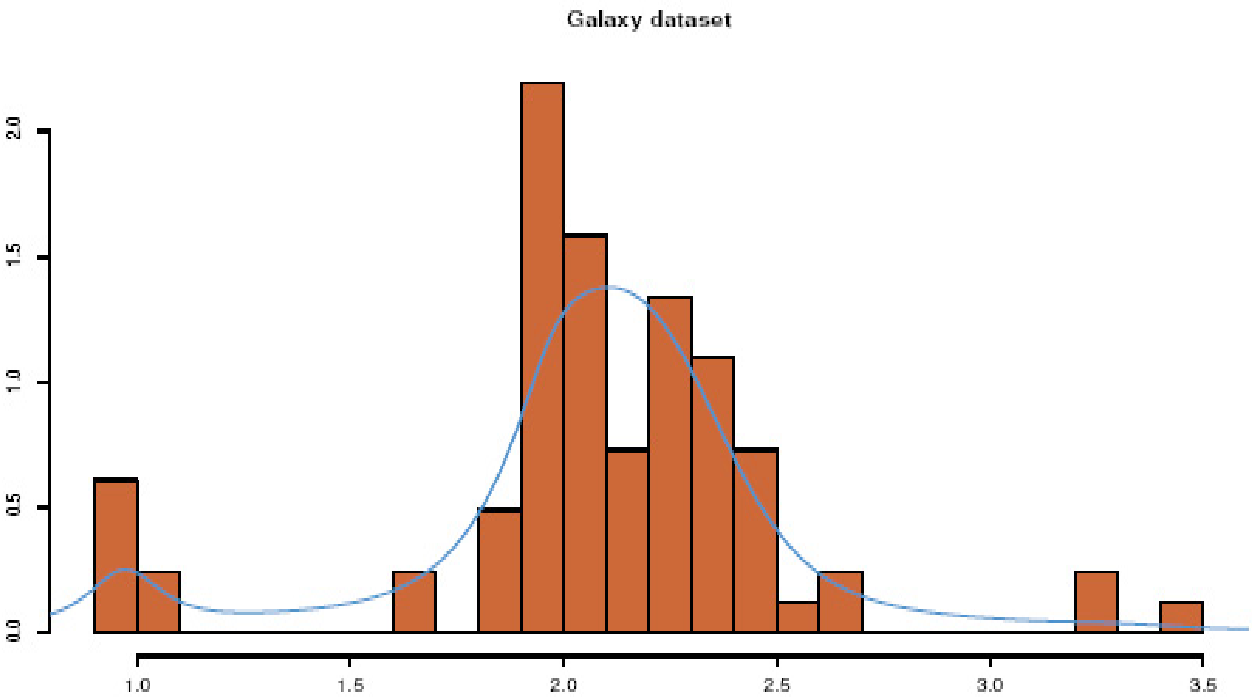
- A standard approach consists of finding a local maximum of the log-likelihood

$$\sum_{i=1}^{n}\log f\left(x_i|\,\theta\right).$$

- Problem: The likelihood is unbounded and k might be unknown.



Velocity (km/sc) of galaxies in the Corona Borealis Region



Predictive distribution for the galaxy dataset.

3.7– Example: Mixture Models and Hidden Markov Models

- Finite mixture models cannot model dependent data.

- Hidden Markov Models (HMM) are very useful in such cases.

- HMM are used for example to model DNA sequences.

3.7– Example: Mixture Models and Hidden Markov Models

- A hidden/unbserved Markov process is defined, call it $\{x_t\}$.

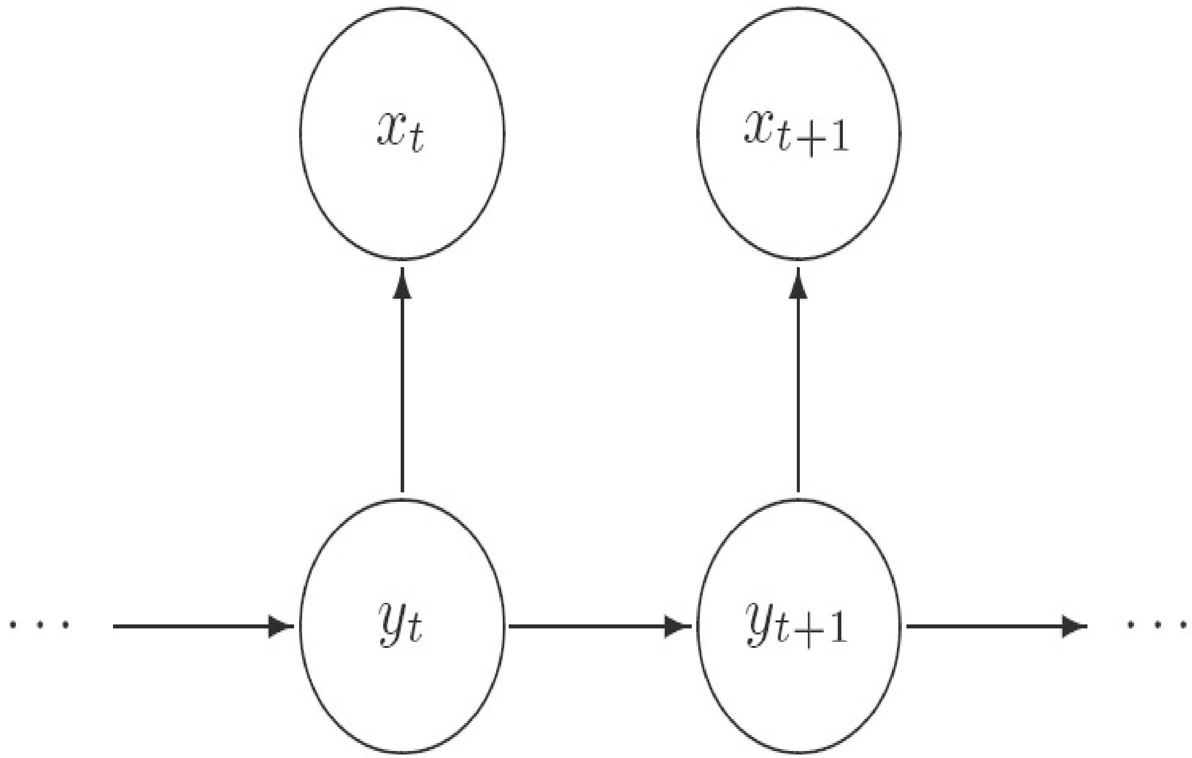
It can take values in a finite space or a continuous space.

- We only have access to an observation process $\{y_t\}$.

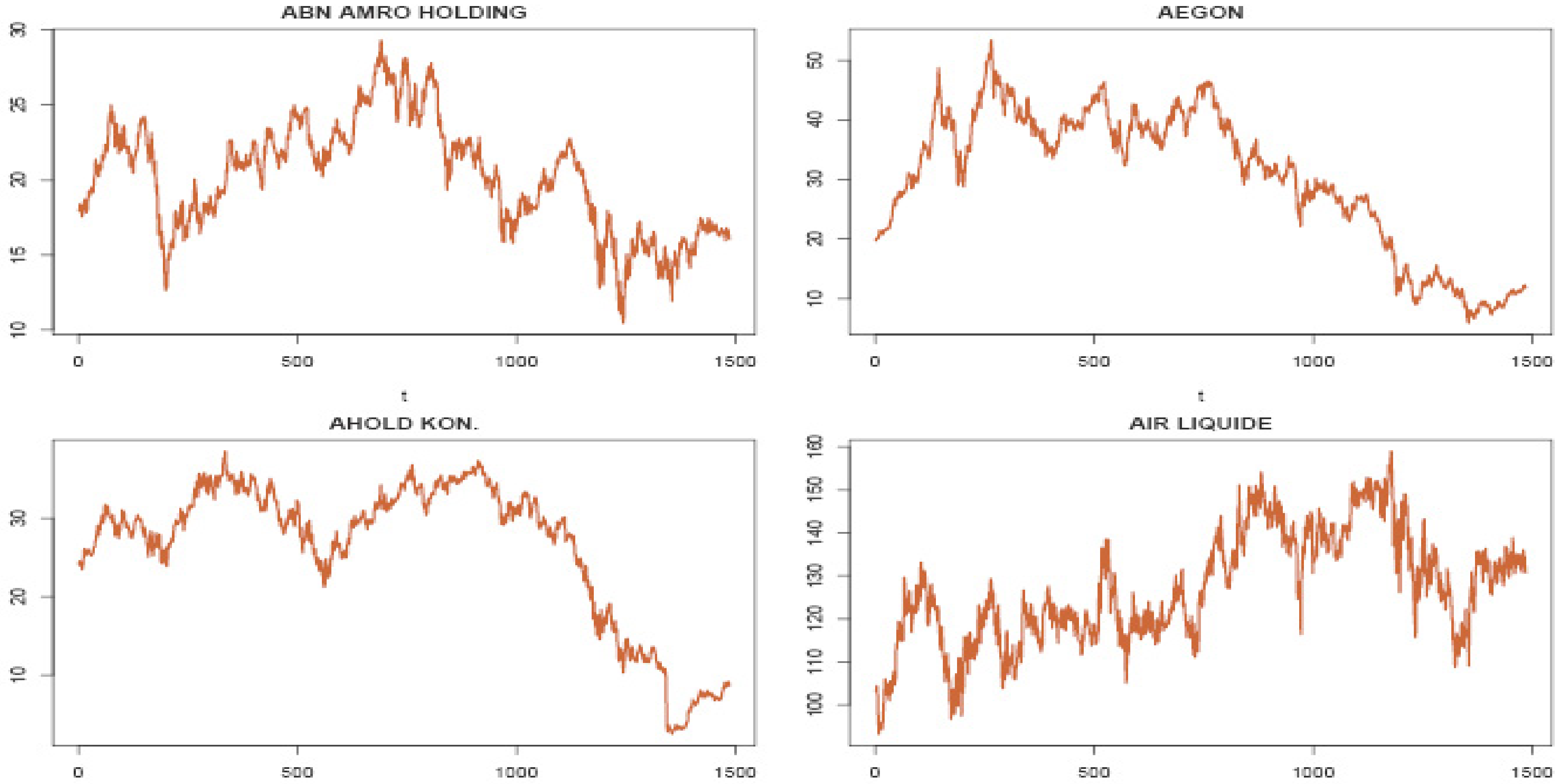
The observations are conditionally independent given $\{x_t\}$.

- Examples: HMM are used for example to model DNA sequences,

speech processing, econometrics, etc..

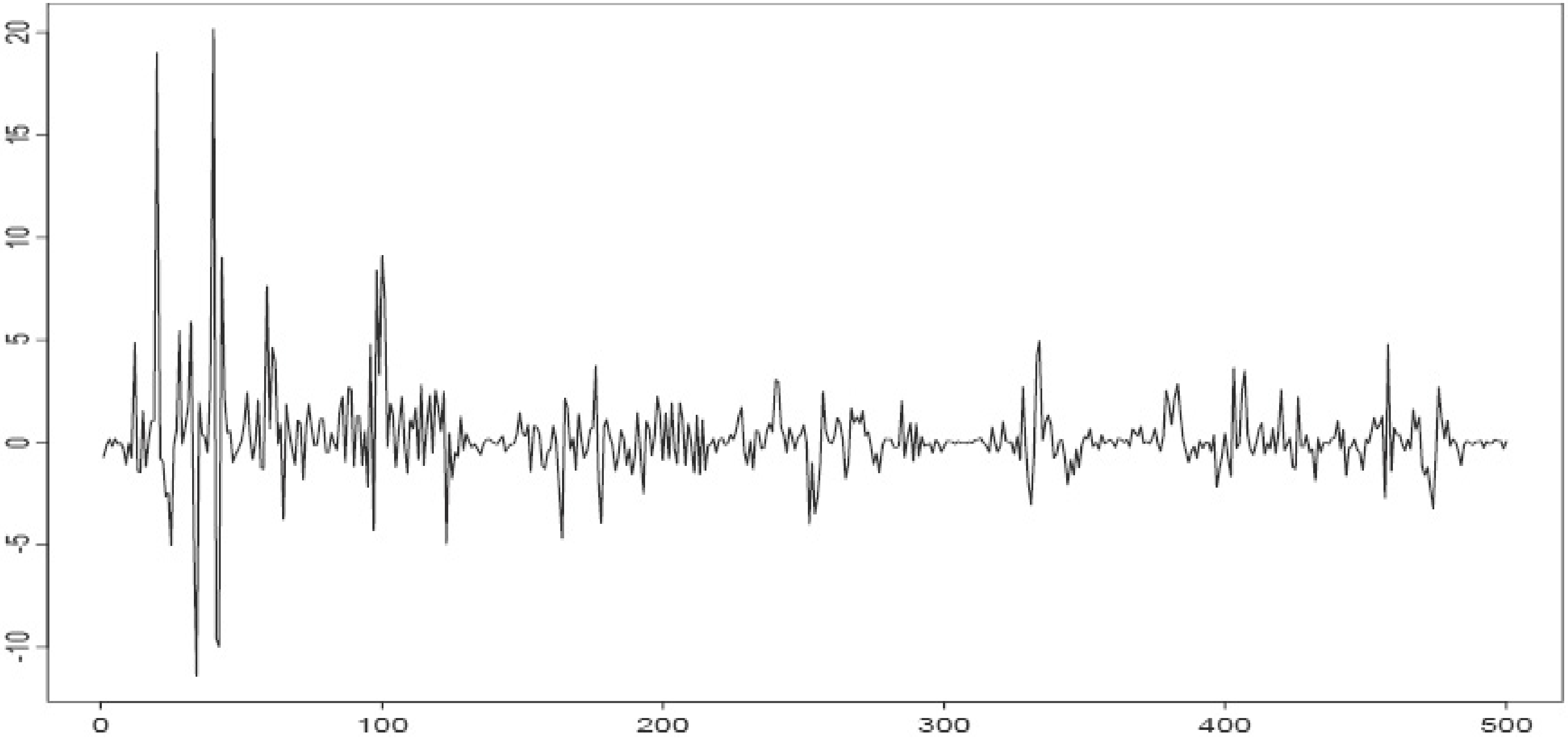


Graphical model representation of HMM



Four stock data

3.8– Example: Financial Time Series



Log return of a stock

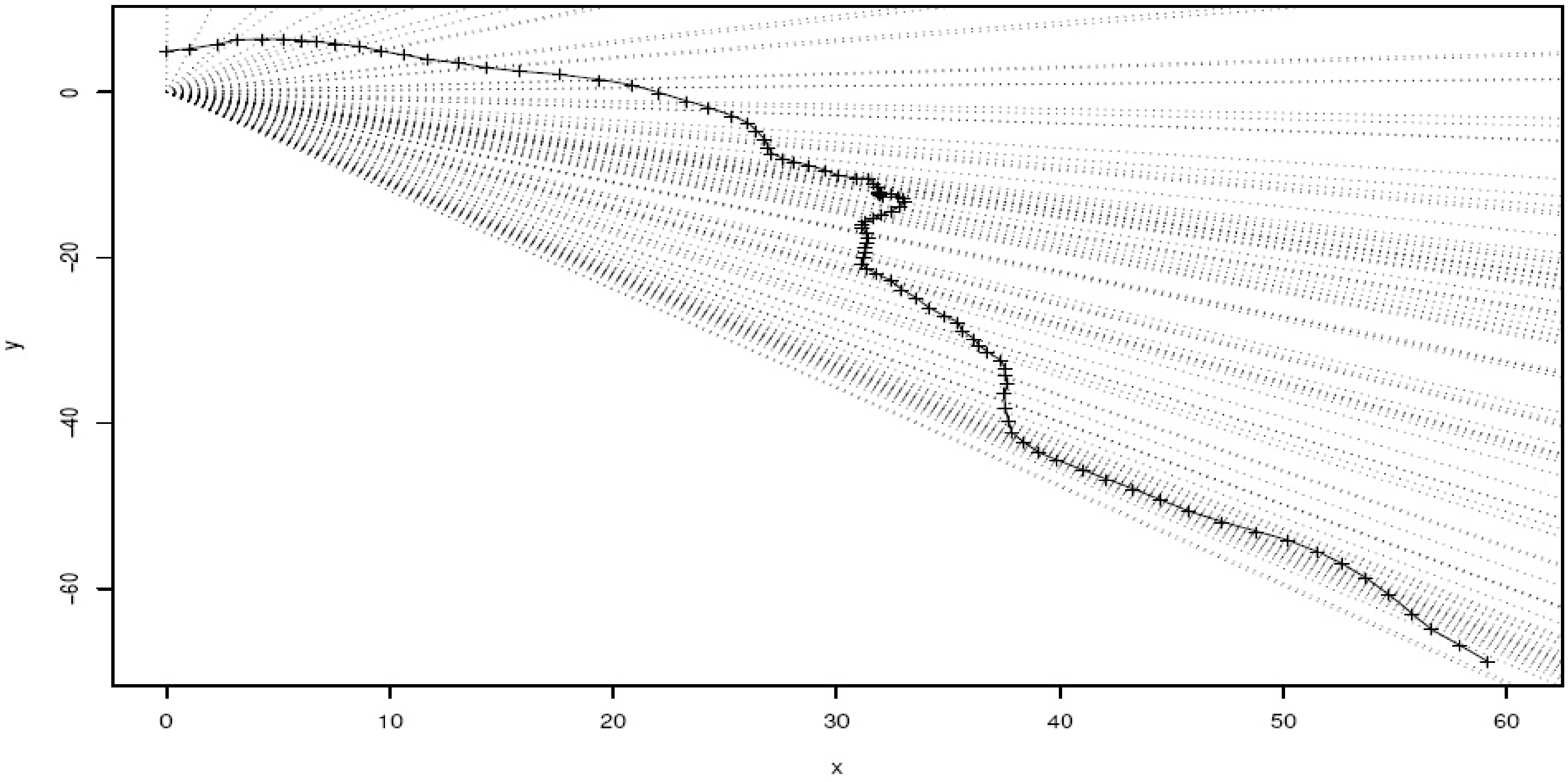
- **Example**: Consider the log-return sequence of a stock then a popular model in financial econometrics is the stochastic volatility model

$$x_t = \alpha x_{t-1} + \sigma v_t \text{ where } v_t \sim \mathcal{N}(0,1)$$

$$y_t = \beta \exp(x_t/2)\, w_t \text{ where } w_t \sim \mathcal{N}(0,1)$$

where the process (y_t) is observed but $(x_t, \alpha, \sigma, \beta)$ are unknown.

3.9– Example: Target tracking



Bearings-only-tracking data

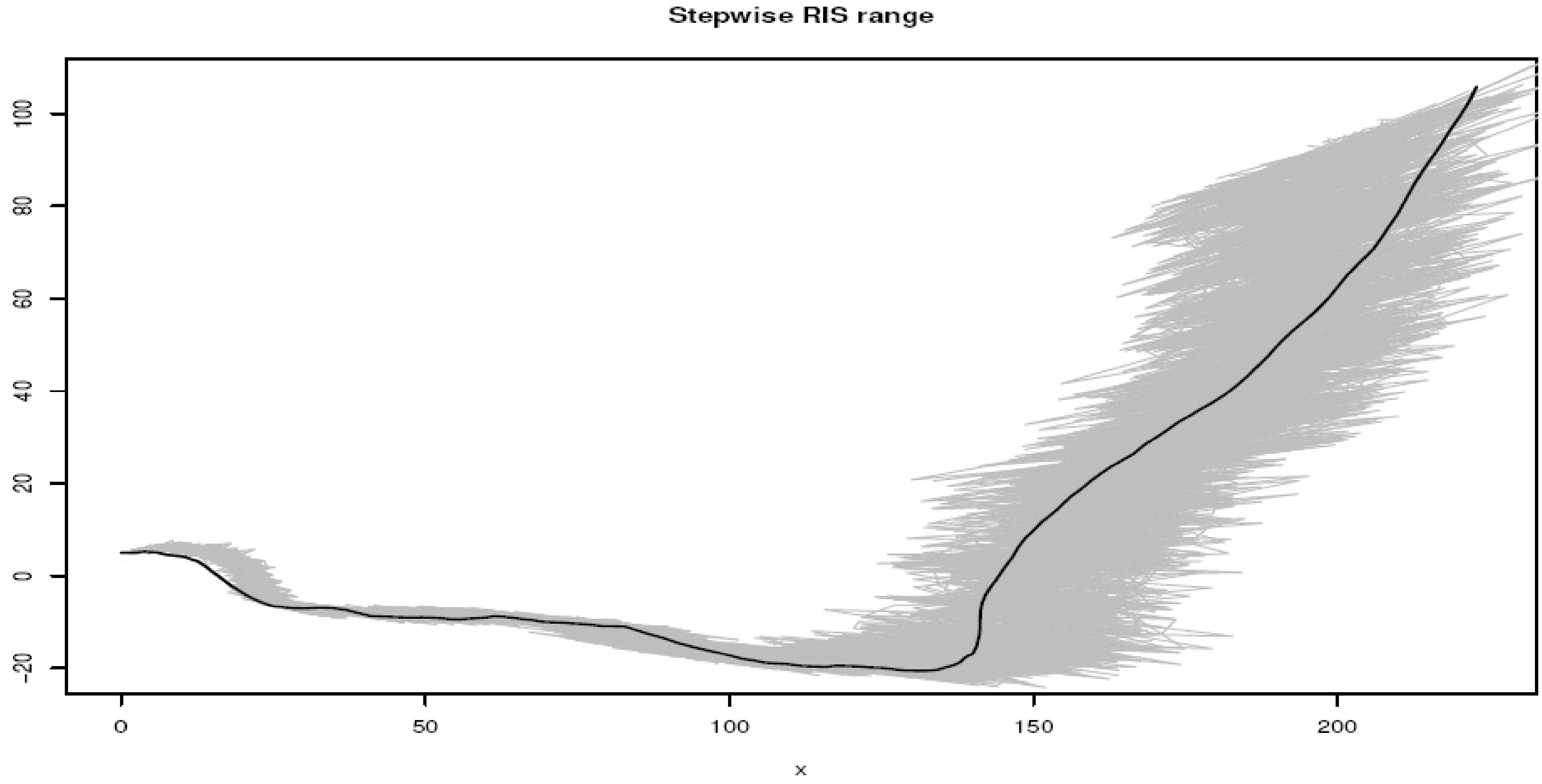
- **Example**: Consider the coordinates of a target observed through a radar.

$$
\begin{pmatrix} x_t^1 \\ \dot{x}_t^1 \\ x_t^2 \\ \dot{x}_t^2 \end{pmatrix} = \Delta \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{t-1}^1 \\ \dot{x}_{t-1}^1 \\ x_{t-1}^2 \\ \dot{x}_{t-1}^2 \end{pmatrix} + noise
$$

$$
y_t = \tan^{-1}\left(\frac{x_t^1}{x_t^2}\right) + w_t
$$

where the process (x_t) is observed but (θ_t) is unknown.

3.9– Example: Target tracking



SMC for state estimation using bearings-only-tracking data

4.1– Possible other topics

- Survival analysis.
- Curve clustering
- Source separation.
- Nonparametric Bayesian estimation.
- Suggestions are welcome.

4.2– For next week

- Please read handouts 1 & 2 by Vidakovic: available on the web.

(alternatively chapter 1 of The Bayesian Choice by C.P. Robert)

- You can also read Handouts 1 & 2 by Kevin Murphy.