

Probability distributions and likelihood functions

Christian Damgaard
Ecoscience
Aarhus University

Probability distributions

- The probability of observing a specific value
- Continuous random variables
 - Normal distribution
 - Gamma distribution – variance increase with mean
 - Beta distribution – to model proportions/probabilities
- Discrete random variables
 - Binomial distribution
 - Beta-binomial distribution - variable probability
 - Poisson distribution – count data
 - Generalized Poisson distribution – mean \neq variance





Probability distribution function

- Normal distribution

$$PDF(x; \mu, \sigma) = \frac{1}{\sqrt{2 \pi \sigma^2}} \text{Exp} \left(-\frac{(x - \mu)^2}{2 \sigma^2} \right)$$





Probability distribution function

- **Binomial distribution**

$$PDF(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- **Poisson distribution**

$$PDF(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$





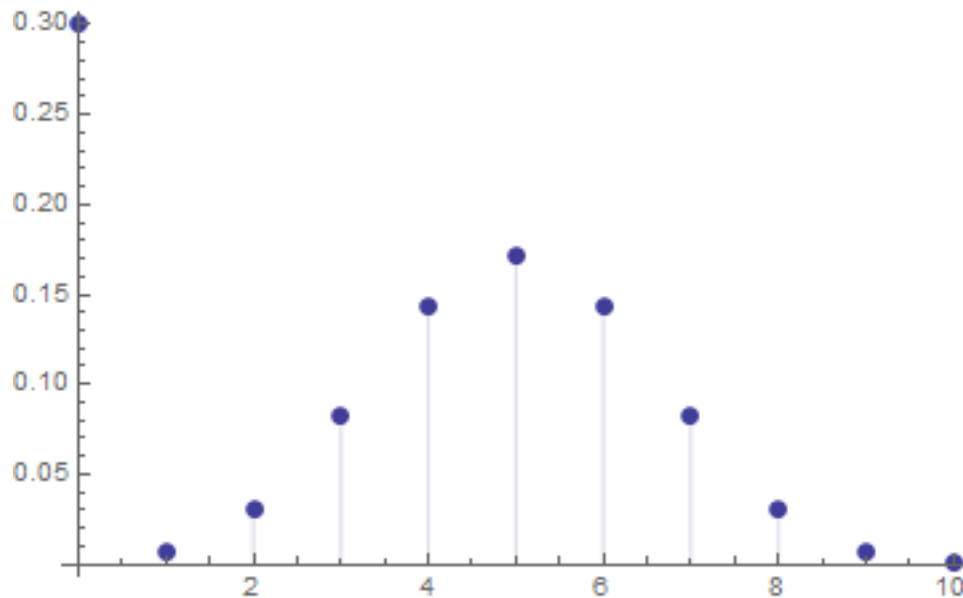
Modify probability distribution functions

- **Reparametrize**
 - make the mean a parameter
 - e.g. gamma distribution
- **Parameter mixture distributions**
 - generalize a parameter
 - e.g. beta-binomial distribution
- **Zero-inflated distributions**
 - add a zero generating process
 - e.g. zero-inflated Poisson distribution
- **Censored regression**
 - Piecewise integration of distribution function
 - e.g. right censored Poisson distribution



Zero-inflated data

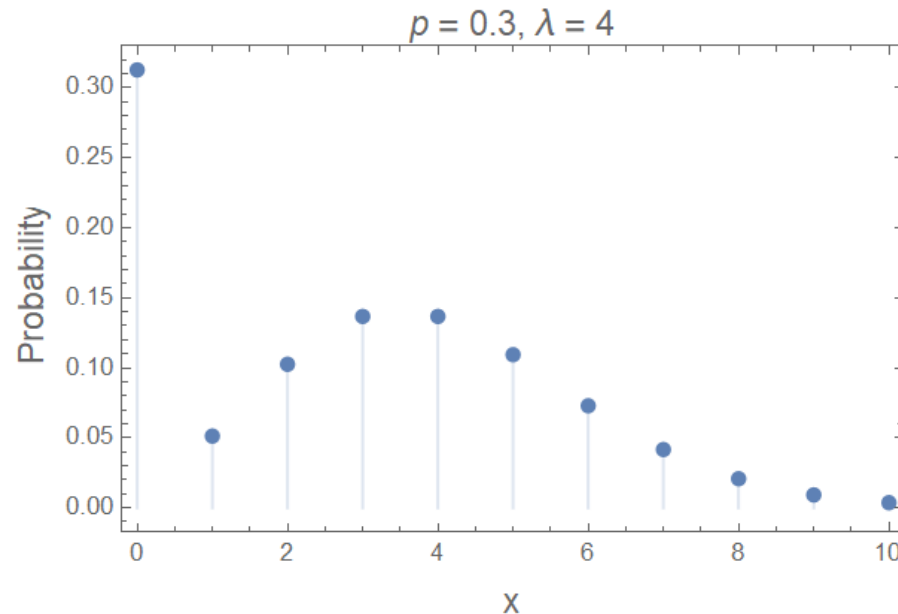
- Local abundance with colonization and extinction dynamics
- Zero-inflated binomial distribution



Zero-inflated Poisson distribution

- Zeros are generated with probability $(1 - p)$

$$PDF(x; p, \lambda) = \begin{cases} (1 - p) + p e^{-\lambda} & x = 0 \\ p \frac{e^{-\lambda} \lambda^x}{x!} & x > 0 \end{cases}$$



Censored data

- **Left censored**
 - Detection limits
- **Interval censored**
 - Population census every 24 hour
- **Right censored**
 - Maximum age

**Grain size sieve -
grain size is a continuous
variable**



Right censored regression with the Poisson distribution

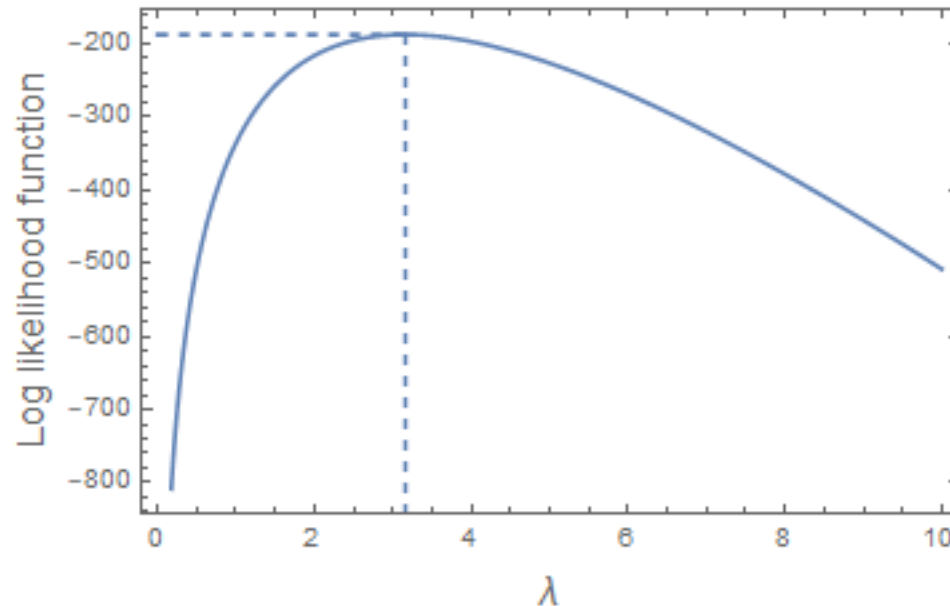
- If $x \geq 10$, the value 10 is recorded

$$PDF(x; p, \lambda) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & x < 10 \\ \prod_{x=10}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} & x \geq 10 \end{cases}$$



Likelihood functions

- Independent and identical distributed random variables (i.i.d.)
- $L(Y|\theta) = \prod_{i=1}^n PDF(y_i, \theta)$
- Maximum likelihood estimates



Likelihood ratio test

$$H_0: \theta = \theta_0$$

$$H_1: \theta$$

$$\mathcal{R} = \frac{L(\theta_0)}{L(\theta)} \Leftrightarrow \log(\mathcal{R}) = \log(L(\theta_0)) - \log(L(\theta))$$

$$\text{For } n \rightarrow \infty: -2 \log(\mathcal{R}) \sim \chi^2(\dim(\theta) - \dim(\theta_0))$$

Very general test...

The F-test is a likelihood ratio test



Loglikelihood functions in R

Normal distribution: `dnorm(y, μ , σ , log=TRUE)`

Binomial distribution: `dbinom(y, x, ρ , log=TRUE)`

Poisson distribution: `dpois(y, λ , log=TRUE)`



PIT residuals (Dunn-Smyth residuals)

X is a continuous random variable with CDF: F_X

$$X \sim f(x, \theta)$$

$$Y := F_X(X)$$

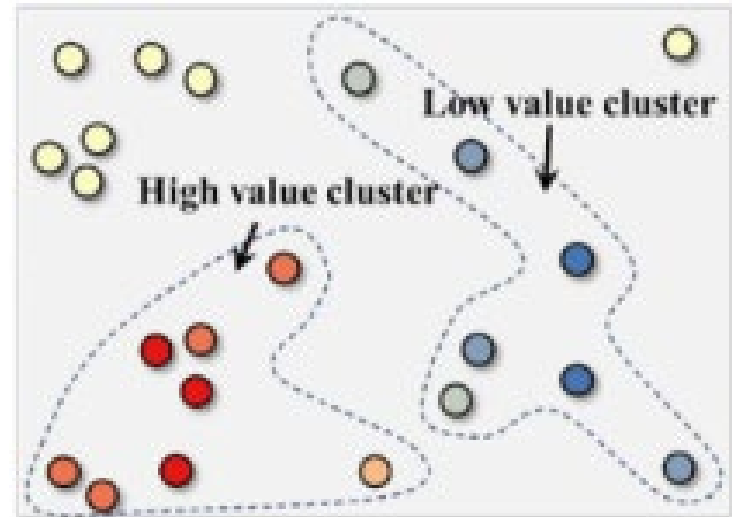
Y has a standard uniform distribution

For discrete distributions – add random noise to Y



Covariation and independence

When data covary in space or time the assumption of independence is violated

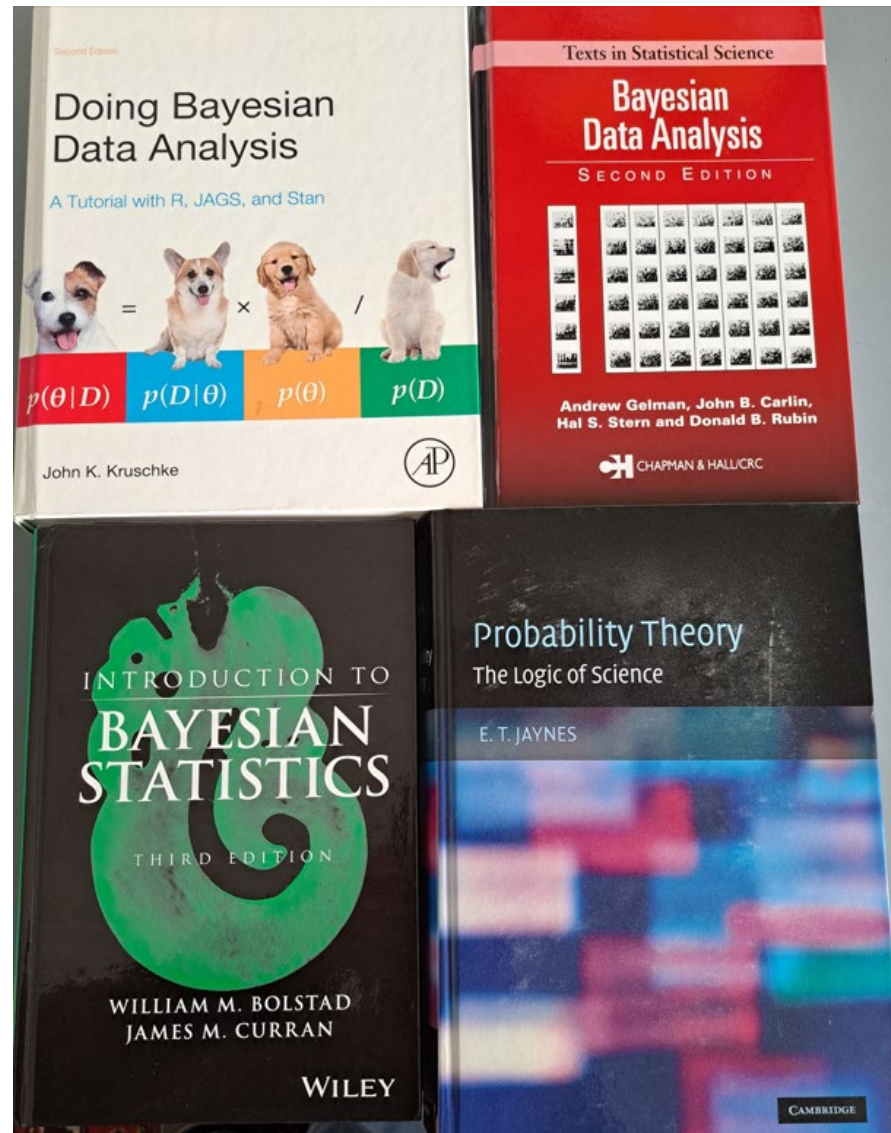


If covariation is not taken into account in the statistical model, then the statistical inference will be biased towards too low uncertainty

too narrow posterior distributions

c.f. too low p-values

Bayesian textbooks



Laplace's approximation

Christian Damgaard
Ecoscience
Aarhus University

Bernstein–von Mises theorem

The posterior distribution of parameters are approximately normal distributed

$$P(\theta | x_1, \dots, x_n) = N(\theta_0, n^{-1} I(\theta_0)^{-1}) \quad \text{for } n \rightarrow \infty$$

θ_0 : true value of θ

I : Fisher information matrix

The theorem links Bayesian and frequentist inference



Laplace's approximation

Bayes theorem: $p(\theta|x, y) = p(\theta|x) \frac{p(y|x, \theta)}{p(y|x)}$

Joint density:

$$p(y, \theta|x) = p(y|x, \theta)p(\theta|x) = p(y|x) p(\theta|y, x)$$

Laplace's approximation:

$$p(y, \theta|x) \approx p(y, \hat{\theta}|x) \exp\left(-\frac{1}{2}(\theta - \hat{\theta})S^{-1}(\theta - \hat{\theta})\right)$$

$\hat{\theta}$: mode of joint density

S^{-1} : matrix of second derivatives at $\hat{\theta}$



RTMB

- Find the mode using the Newton method and the gradients of the loglikelihood function
- RTMB
 - Template Model Builder - automatic differentiation
 - developed by Kasper Kristensen (DTU-Aqua)
 - implements Laplace approximation for random effects
 - automatic sparseness detection and code optimization





RTMB

```
1 library(RTMB)
2
3 #load data
4 data <- read.table(file="file", header=TRUE, sep = ",")
5
6 #dat and par list
7
8 dat<-list()
9 dat$District <- as.integer(as.factor(diseasedata$District))
10 ...
11
12 par<-list()
13 par$beta <- as.vector(rep(0.1, dat$nfaktor))
14 par$Zy <- numeric(max(dat$Year))
15 par$logSdY <- 0
16 ...
```



RTMB

```
17 #Joint negative log-likelihood function
18 jnll <- function(par){
19   getAll(par, dat)
20   X <- OBS(X)
21   sdY <- exp(logSdY)
22   jnll <- 0
23   jnll <- jnll -sum(dnorm(Zy, 0, sdY, log=TRUE))
24   for(i in 1:length(N)){
25     expvar_row <- Expvar[District[i], ]
26     predlin <- sum(beta * expvar_row) + Zy[Year[i]]
27     predp <- exp(predlin)/(1+exp(predlin))
28     jnll <- jnll - dbinom(X[i], N[i], predp, log = TRUE)
29   }
30 }
31 jnll
32 }
33
34 jnll(par)
35
```



RTMB

```
35  
36 obj <- MakeADFun(jnl1, par, random="Zy")  
37  
38 fit <- nlminb(obj$par, obj$fn, obj$gr)  
39  
40 sdr <- sdreport(obj)  
41 pl <- as.list(sdr,"Est")  
42 plsd <- as.list(sdr,"Std")  
43
```



A wide-angle photograph of a rolling landscape covered in dense purple heather. The terrain is hilly, with some green trees scattered across the top of the ridges. The sky is a pale, overcast blue.

Probability theory the logic of science

Christian Damgaard
Ecoscience
Aarhus University

Deductive reasoning

Aristoteles (384 BC – 322 BC) is credited with formalizing logic

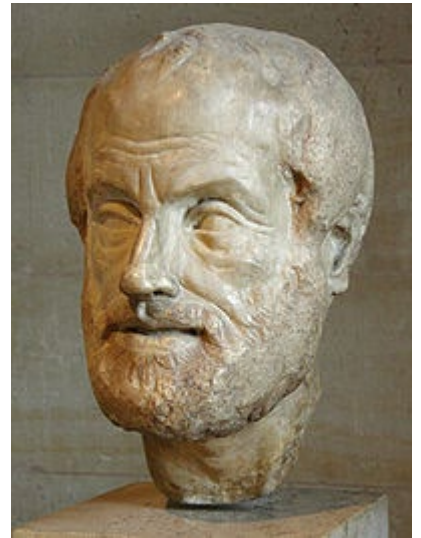
A. It rains

B. There are clouds in the sky.

If A is true, then B is true,

If B is false, then A is false

It would be nice and simple if we could use such reasoning all the time



Plausible reasoning

In science it is often not possible to be sure

A. It will start to rain before 10

B. The sky will become cloudy before 10

If B is true, then A becomes more plausible

$P(A|B)$?

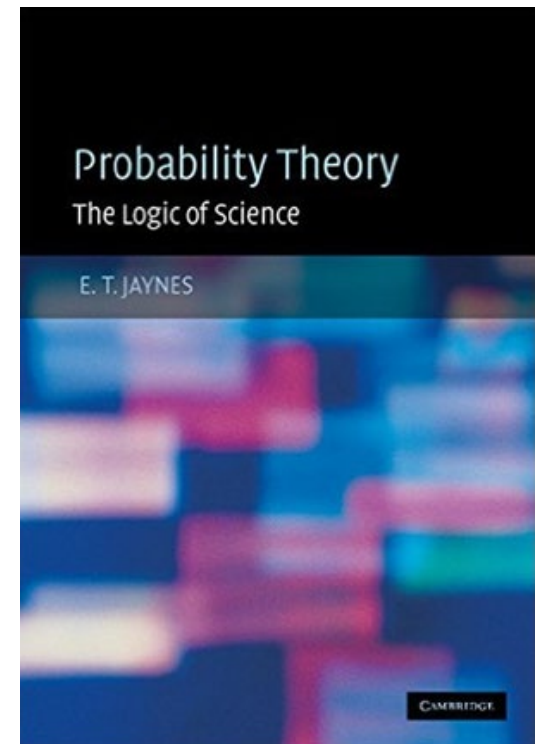


Fundament of plausible reasoning

1. Degrees of plausibility are represented by real numbers
2. Qualitative correspondence with common sense
3. Consistency

These statements are sufficient to form the basis of probability theory

(Cox 1946, Janes 2003)





Probability theory

Product rule:

$$P(AB|C) = P(A|BC)P(B|C) = P(B|AC)P(A|C)$$

Sum rule:

$$P(A \vee B|C) = P(A|C) + P(B|C) - P(AB|C)$$

$$\text{if } B = \bar{A}: \quad P(A|C) + P(\bar{A}|C) = 1$$

**A probability represent a state of knowledge
(not a physical entity)**





Bayes theorem

H : Some hypothesis

D : Data

X : Prior - information that is not entailed in the data

$$P(HD|X) = P(H|DX)P(D|X) = P(D|HX)P(H|X) \Leftrightarrow$$

$$P(H|DX) = P(H|X) \frac{P(D|HX)}{P(D|X)}$$

Posterior = prior * normalized likelihood function





Testing hypotheses

Bayes theorem: $P(H|DX) = P(H|X) \frac{P(D|HX)}{P(D|X)}$

Inverse probability: $P(H|DX)$

The posterior pdf is the probability that hypothesis H is correct given the data D and the prior information X

Testing hypotheses are done by estimating (compound) parameters

e. g. $H: \mu > 0$



Numerical estimation of posterior dist.

Bayes theorem:
$$P(H|DX) = P(H|X) \frac{P(D|HX)}{P(D|X)}$$

Generally, it is not possible to calculate $P(D|X)$, instead we use that

posterior \propto prior * likelihood function

and simulate the posterior distribution using MCMC (Markov-chain Monte-Carlo) methods



MCMC - Metropolis - Hastings algorithm

Sample a MCMC chain from the posterior distribution -
 U is the sampling chain

1. Guess a value of the parameter $U^0 = u$
2. Draw v from a candidate density $q(u)$
3. Compute $r = \frac{P(v|Y)}{P(u|Y)}$
4. Draw $z = \text{random number } [0, 1]$
5. If $r \geq z, U^t = v$
If $r < z, U^t = u$

Repeat 2-5 many times



Example: Poisson distribution

Y is an independent and identically distributed (i.i.d.) stochastic variable: $Y \sim \text{Poisson}(\lambda)$

Likelihood function: $P(Y|\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{y_i}}{y_i !}$

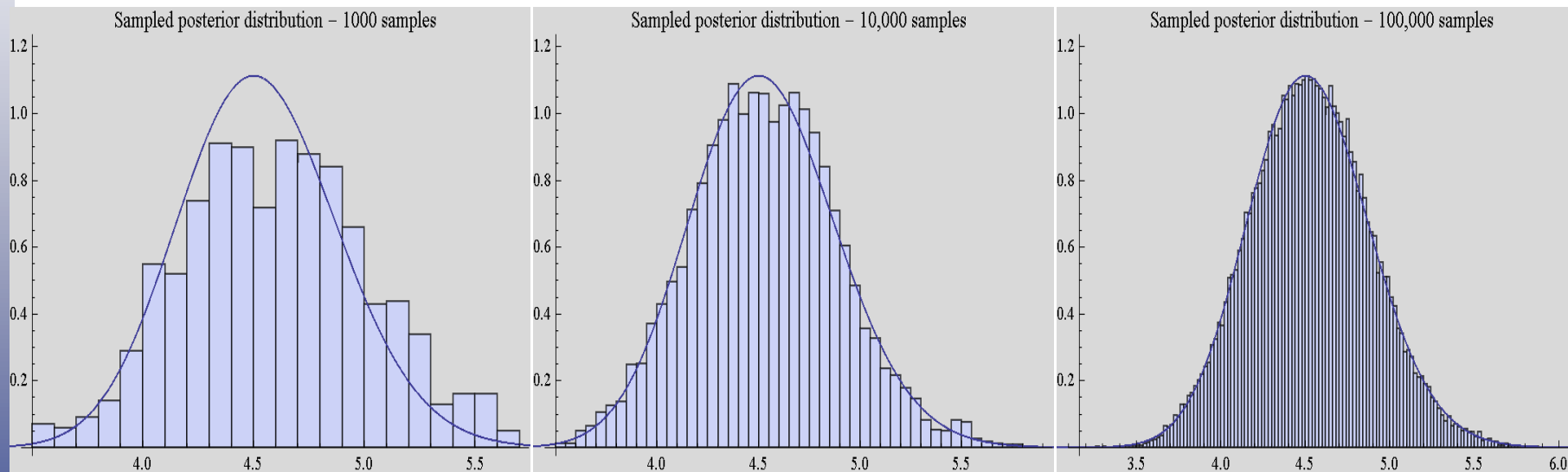
Prior distribution (gamma distribution): $P(\lambda) = \frac{e^{-\frac{\lambda}{\beta}} \beta^{-\alpha} \lambda^{\alpha-1}}{\Gamma(\alpha)}$

Posterior distribution: $P(\lambda|Y) \propto P(\lambda) P(Y|\lambda)$



Example: Poisson distribution

Numerical estimation of the posterior distribution using MCMC



Very general method

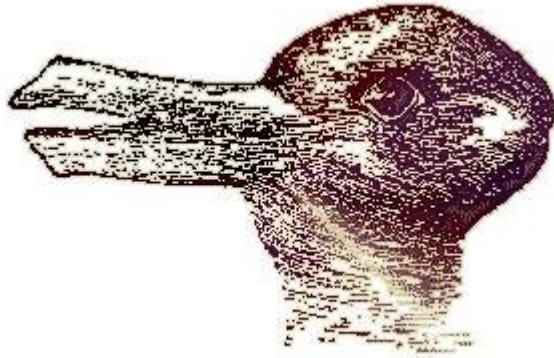


The prior...

- In Bayesian statistics we have to specify a prior distribution for the parameters
- Is it possible to design an experiment without using prior knowledge?
- Updating prior knowledge with data
- We interpret and value scientific information differently
 - subjective assessments



The prior...



What animal?

Bird, hare or both

When you first have seen both animals, the information content of the picture will have changed

The prior...



Objective

“Chair”

Subjective

Synthetic a priori judgments: causality, concept of time, Euclidean space,...



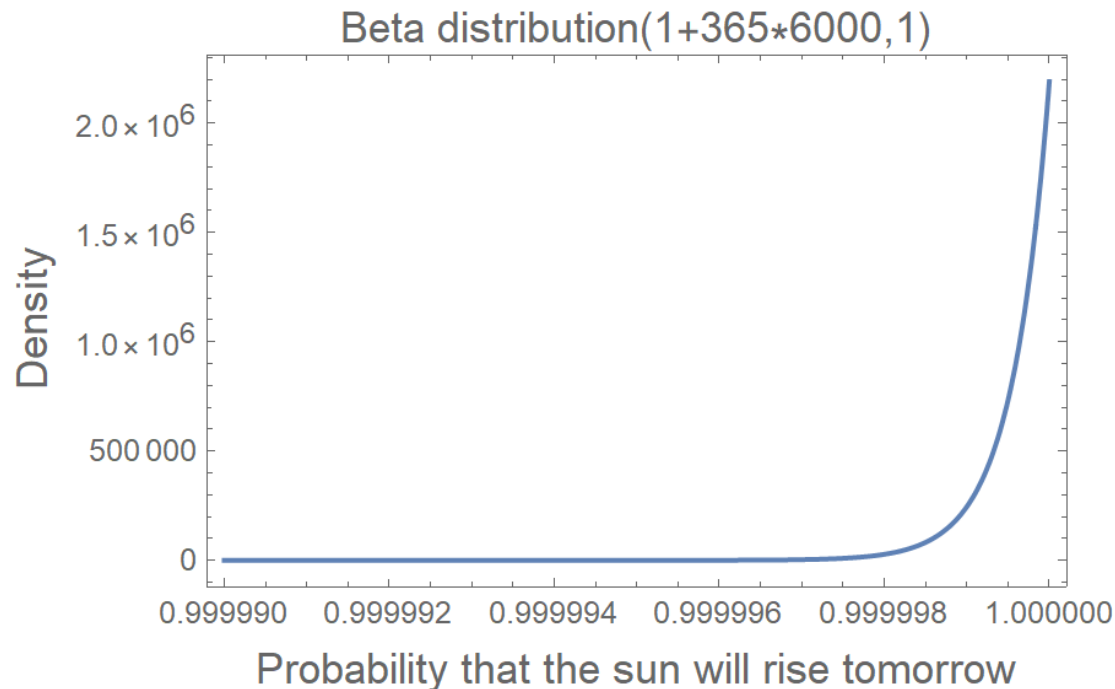
Inductive reasoning

- **An often-used method to obtain general results from observations**
 - e.g. from observation points to a regression line
- **The method of induction has been criticized as being illogical (Hume, Popper)**
- **However, if we generalize the Aristotelian logic to a quantitative logic (scientific logic), then the method of induction is firmly rooted in probability theory**
- **We may not be able to state that all swans are white, but we can calculate the probability that the next swan we encounter is white**



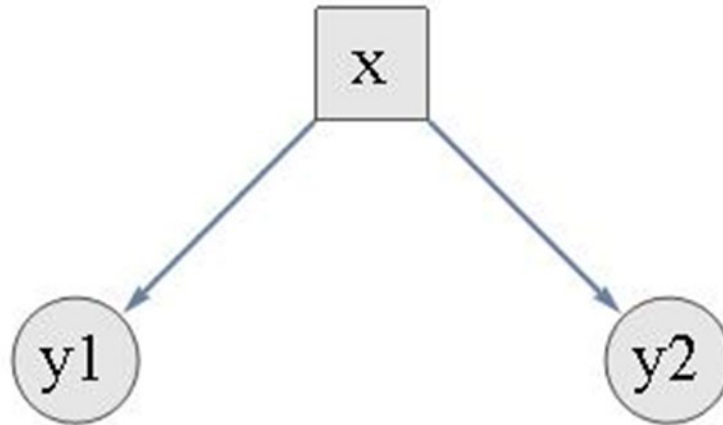
Example: the sun rise problem

- Laplace calculated the probability that the sun would rise the next day in 1814
- He assumed that the sun had risen each morning for 6000 years, and a uniform prior probability distribution

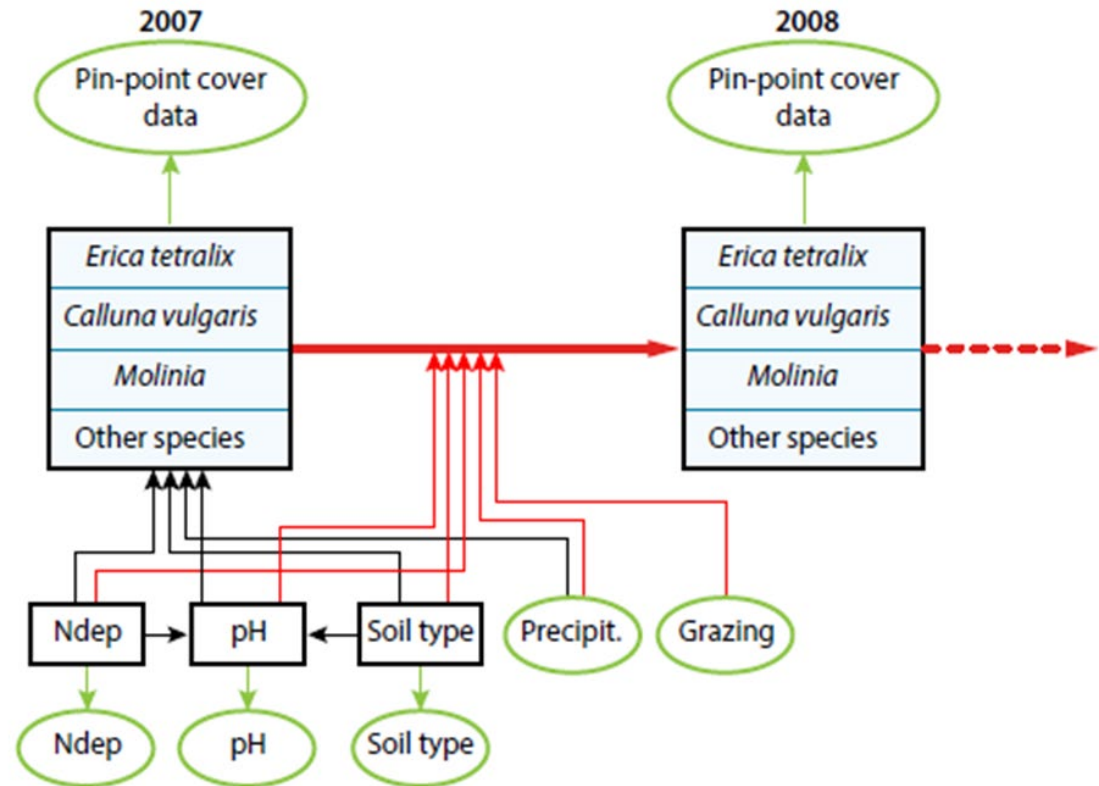
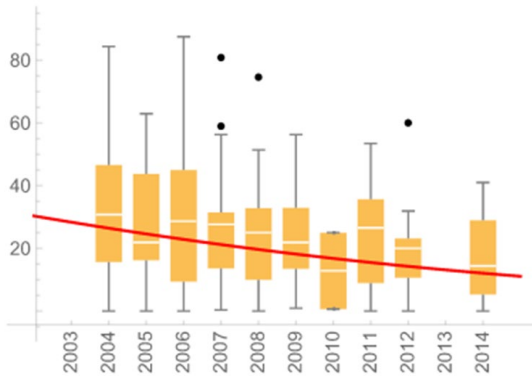


Scientific logic and causality

- Correlation is not causality
- In order to understand a phenomenon we need to know the causal pathways
 - Why instead of what
- Investigate a causal hypothesis by fitting it to data



SEM example – wet heathlands



Structural equation modelling

Christian Damgaard
Ecoscience
Aarhus University

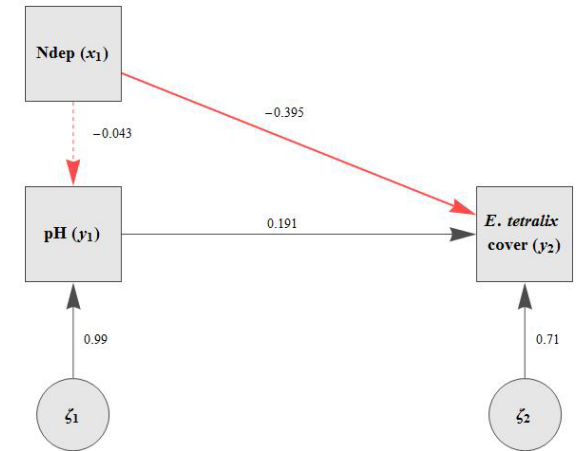
Structural Equation Modelling

Hypothesized causal mechanisms are specified in graphical models

Tests for conditional independence

Estimation of direct and indirect effects

Ecological predictions with a quantified degree of uncertainty



Correlation vs. causality

y1 and y2 are correlated

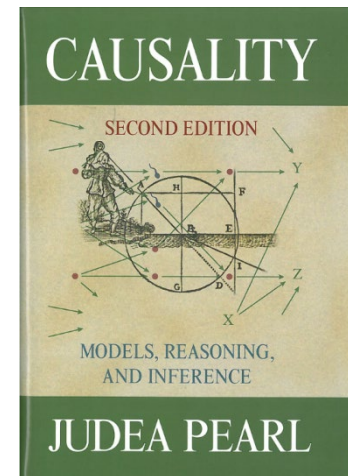
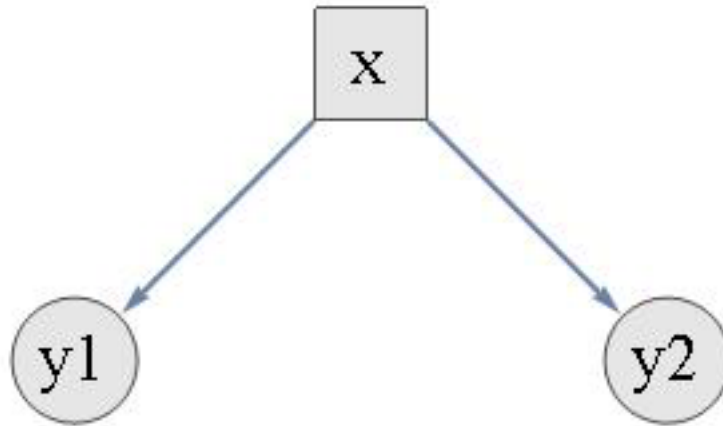


$$P(y1|y2) = P(y2|y1) \frac{P(y1)}{P(y2)}$$



Correlation vs. causality

If both y_1 and y_2 are regulated by x then the *residual* variation of y_1 and y_2 may be independent
= conditional independence



$$P(y_2|x, y_1) = P(y_2|x) \quad \text{if } (y_2 \perp y_1|x)$$



Graphical models - SEM

Graphical model of hypothesized causal relationships

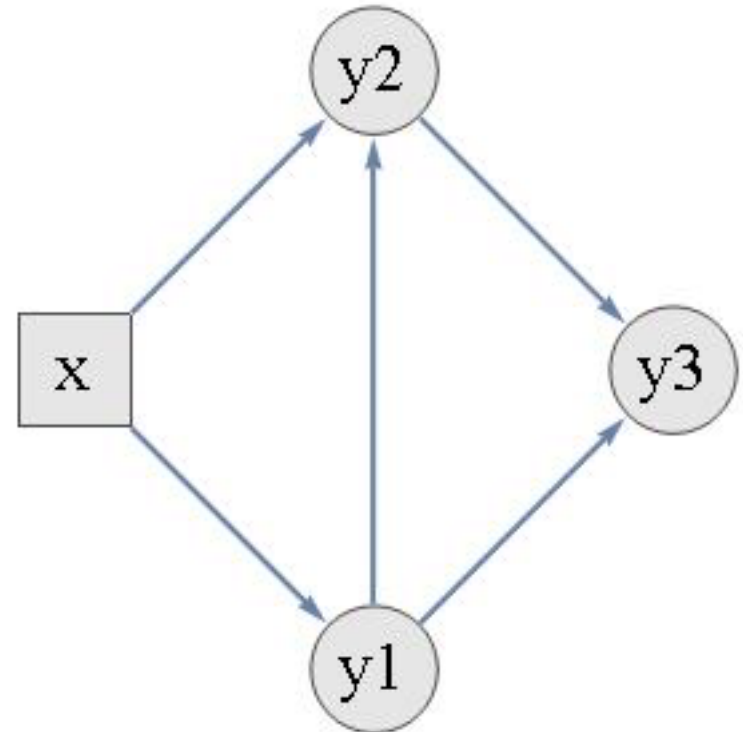
X is an exogenous variable (outside the model)

Y are endogenous variables

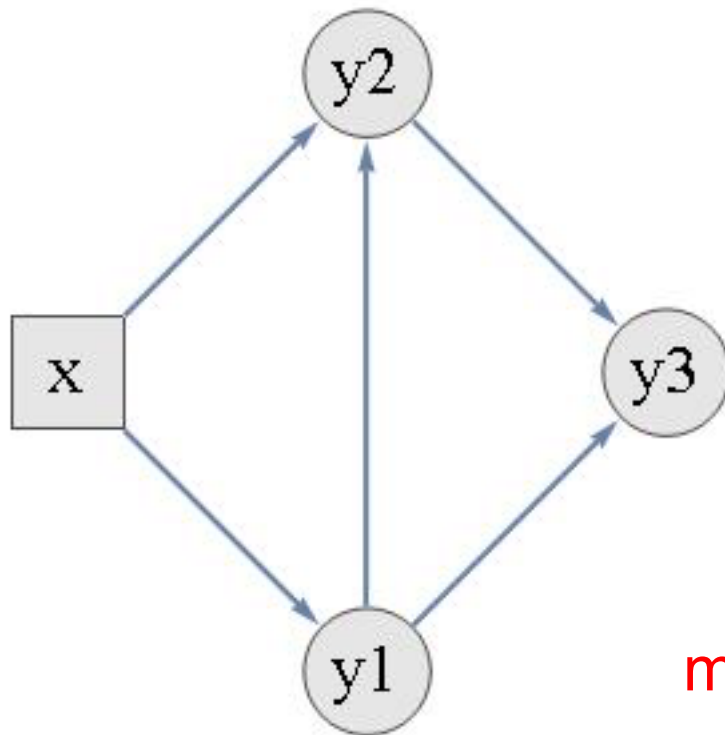
X has a direct effect on y1

X has a direct effect on y2, but also an indirect effect mediated by y1

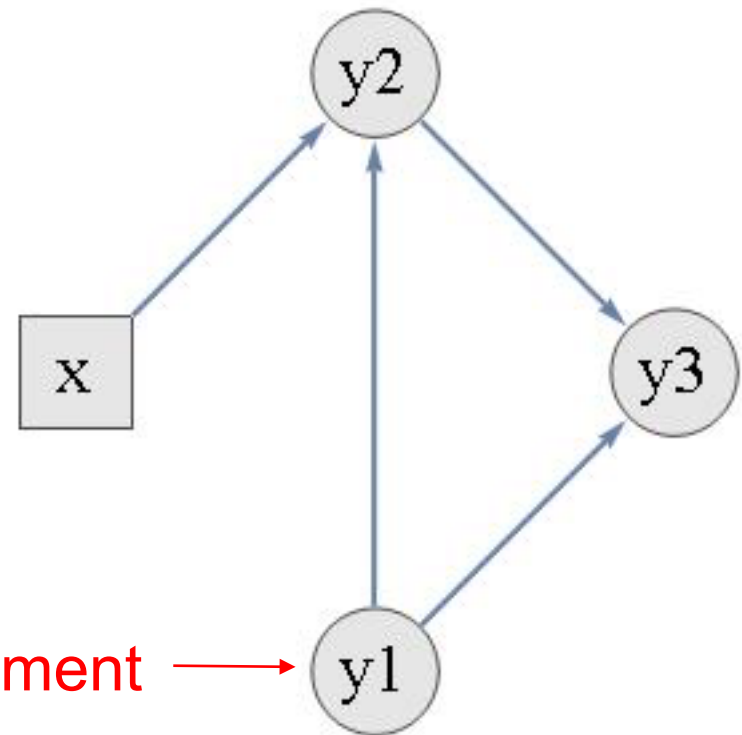
X has only indirect effects on y3



Manipulating a factor (y1) in a controlled experiment



$$P(y1, y2, y3|x)$$



$$P(y2, y3|x, do\{y1\})$$



The causal relationships are best resolved from longitudinal data

The events of today control what happens tomorrow

Conditional independence

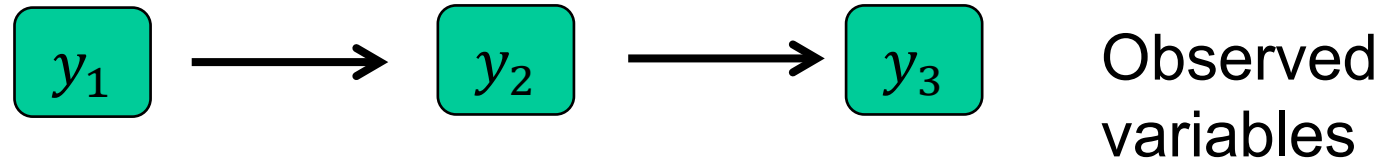
Markov chain (no history or time lag)

$$P(x_{t=1}, x_{t=2}, x_{t=3}, x_{t=4}) =$$

$$P(x_{t=1})P(x_{t=2}|x_{t=1})P(x_{t=3}|x_{t=2})P(x_{t=4}|x_{t=3})$$



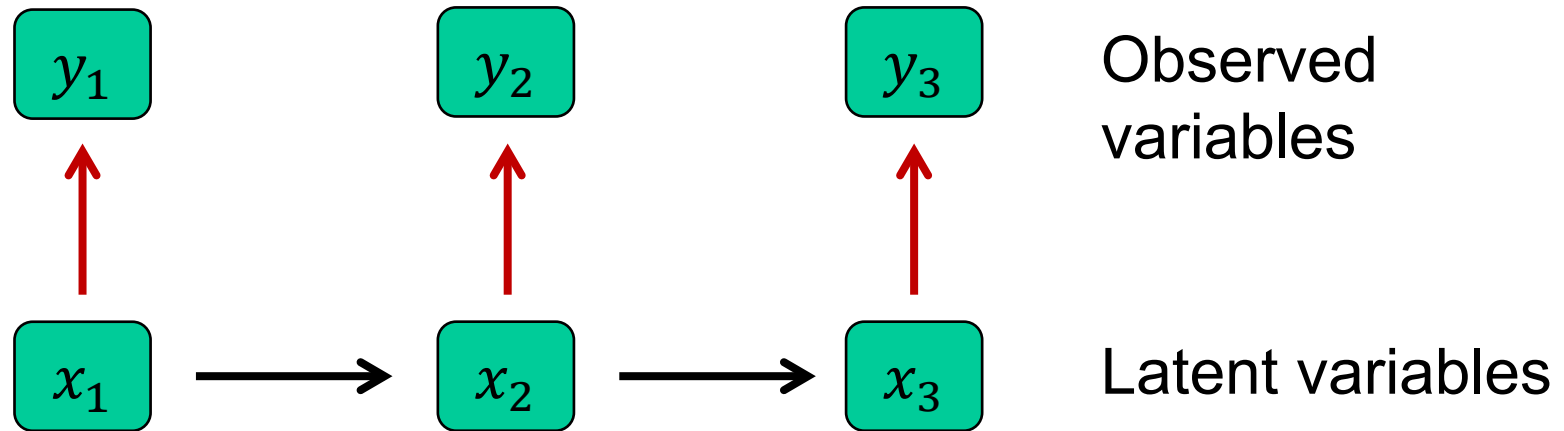
Process occurring in time



Regression analysis : $y_i = f(y_{i-1}) + \varepsilon$



SEM – hierarchical model



Process: $x_i \sim N(f(x_{i-1}, \theta), \sigma)$

Measurement: $y_i \sim \mathcal{D}(x_i, \tau)$

Separation of process error and **sampling error**



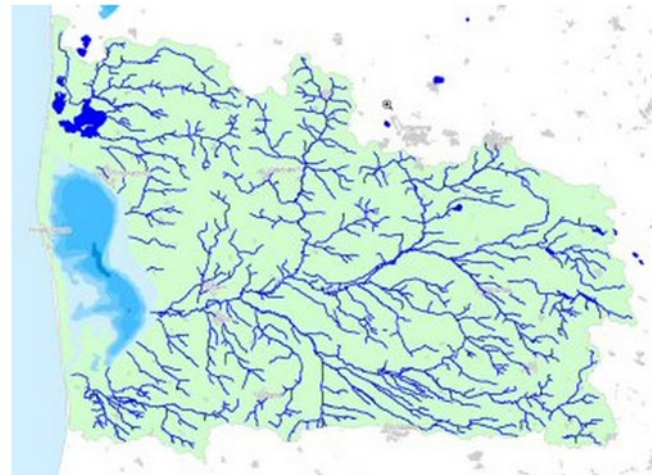
Integration of ecological processes (SEM)

In order to study causality and the effect of management and restoration at the ecosystem level we need to integrate different ecological processes in one modelling framework

Models should be explicit in space and time

Example: catchment

Topography, precipitation, soil,
food chains, competition,
biodiversity



Prediction and uncertainties

Christian Damgaard
Ecoscience
Aarhus University

Prediction of y'

$$P(y'|y) = \int P(y'|\theta) P(\theta|y) d\theta$$

$P(\cdot)$ distribution of...

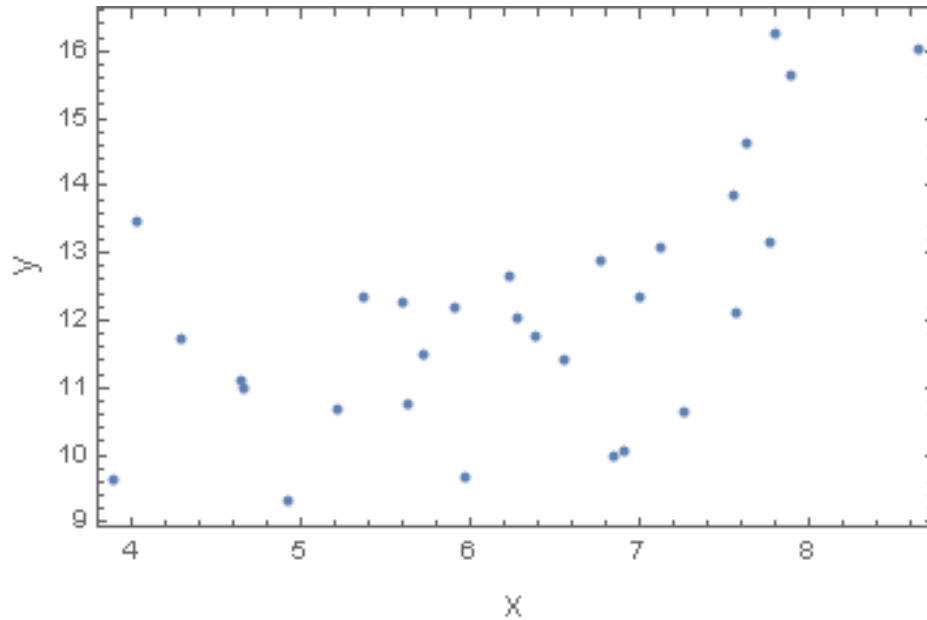
y' predicted data point

y data and knowledge

θ model and parameters



Prediction from sampled data



Dependent variable (y) ~ independent variable (x)

Abundance of plant species ~ abiotic gradient

... ~ ...

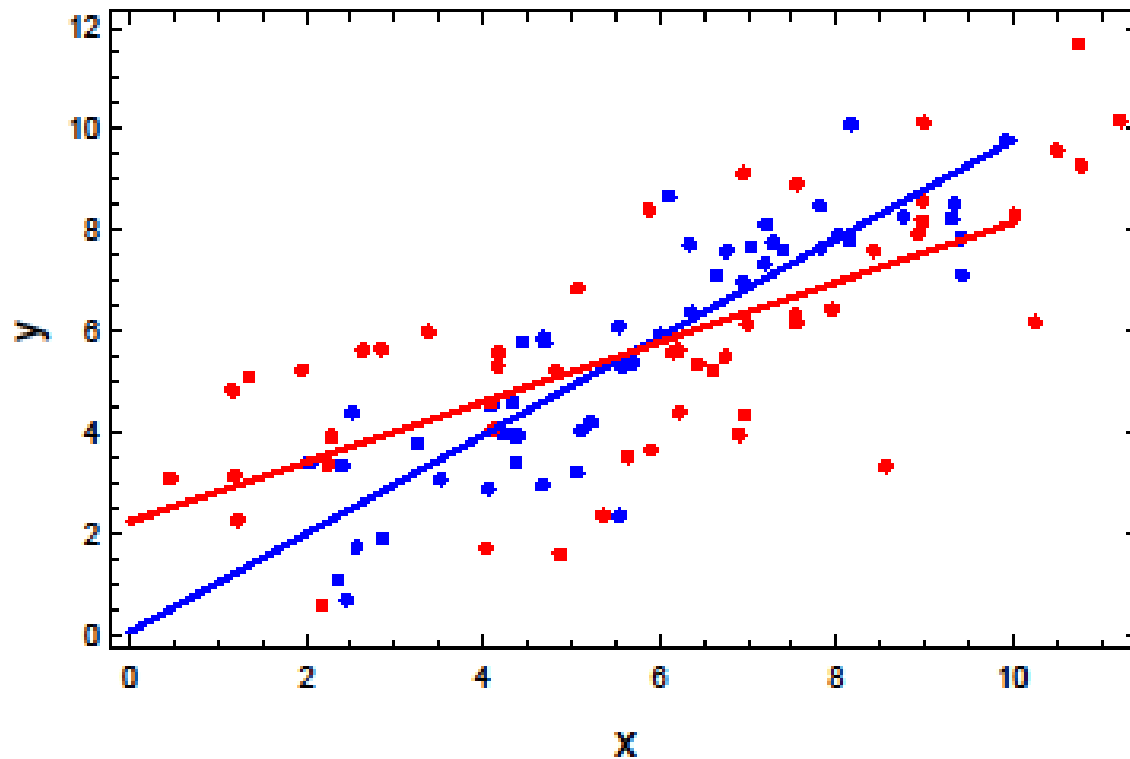


Regression dilution

Same data generation model:

Blue points: no measurement- and sampling error in X

Red points: random measurement- and sampling error in X

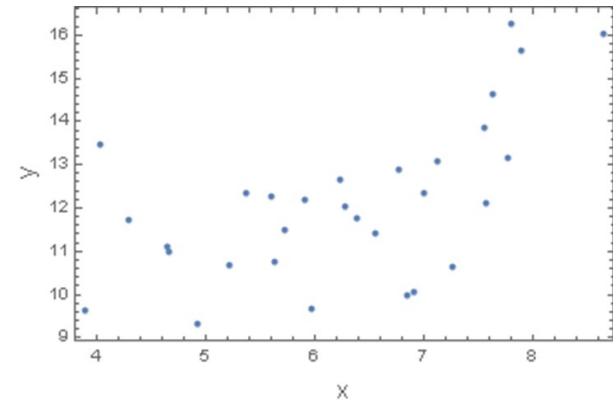


Prediction in a linear model

Linear model is fitted to data:

$$y = f(\alpha, x) + \epsilon, \quad \epsilon \sim N(0, \sigma)$$

$$E(y' | x') = f(\hat{\alpha}, x')$$



Uncertainty is due to:

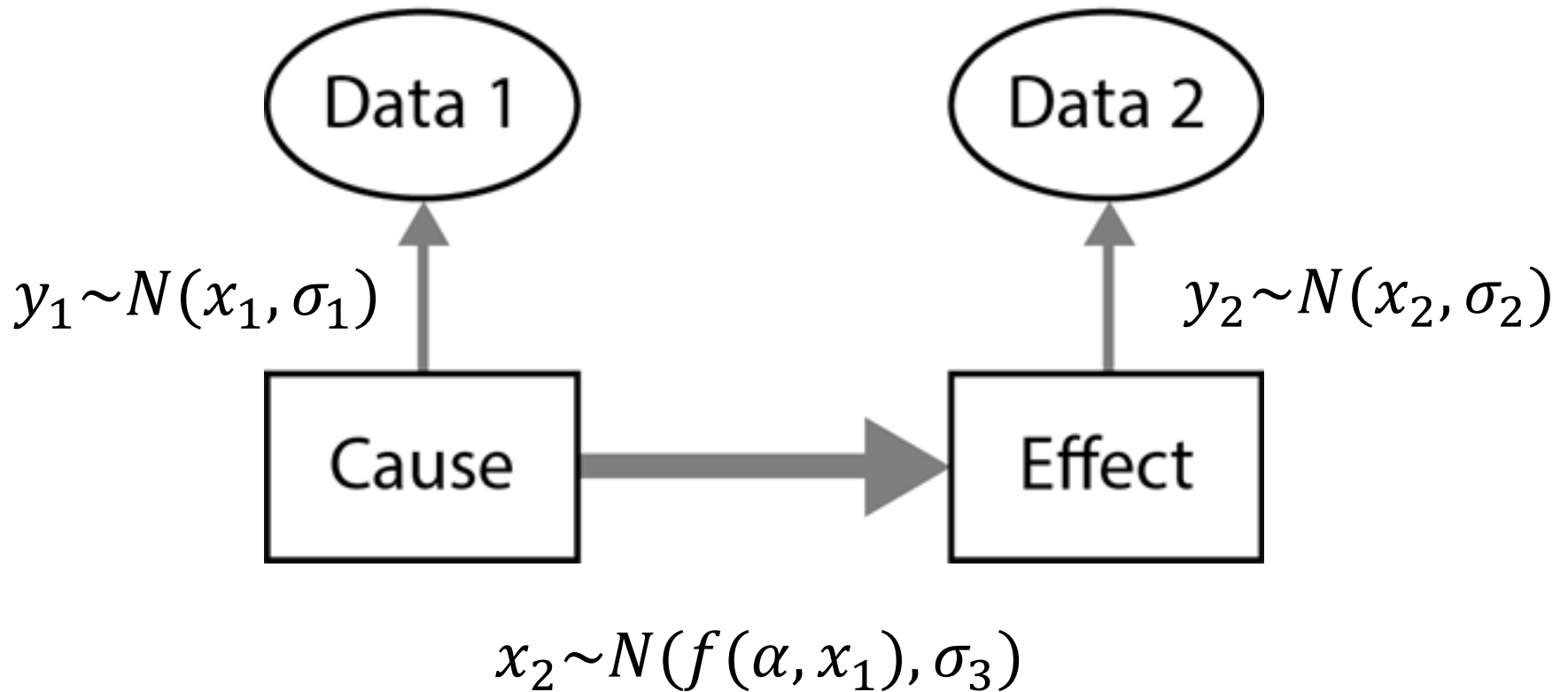
Residual variation σ

Error in the determination of $\hat{\alpha}$

x and x' are assumed known without error



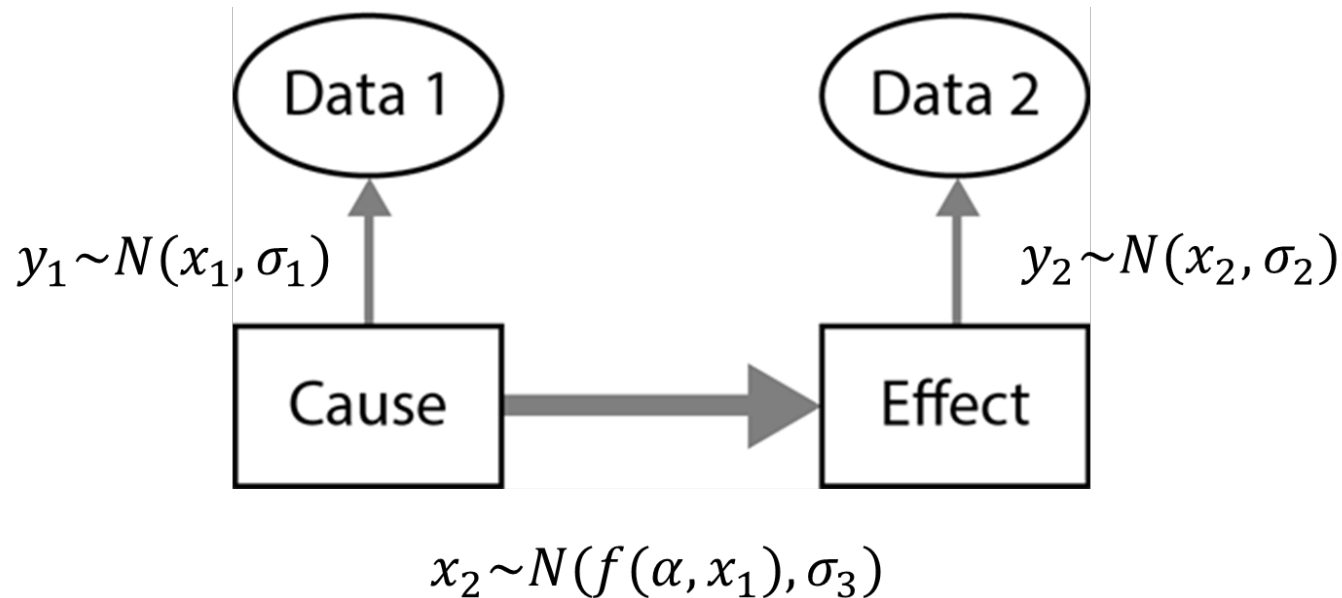
General hierarchical model



Partitioning of the variance into measurement errors and process error



Prediction of effect

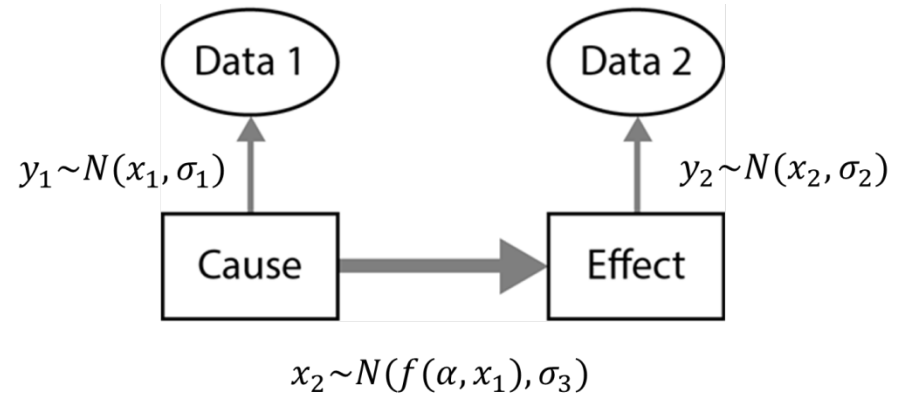
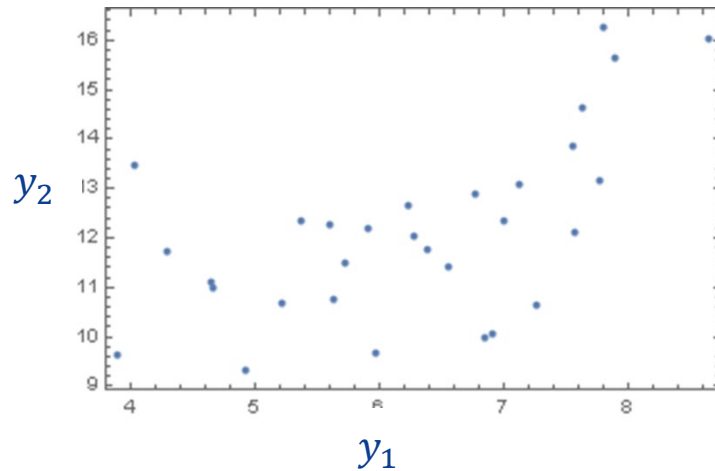


Prediction of effect depends on σ_3 , maybe σ_1 , but not σ_2

The framework may be extended to include many variables and processes (structural equation modelling)



Bayesian example



Joint posterior distribution of the parameters $\theta = (\alpha, \sigma_1, \sigma_2, \sigma_3)$ are calculated using MCMC, and predictions are made using

$$P(y'|y) = \int P(y'|\theta) P(\theta|y) d\theta$$

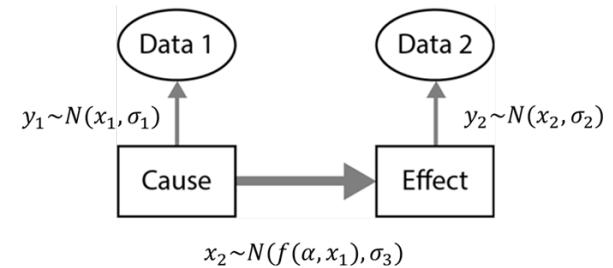
Bayesian example

Simulation with $n = 30$

Subsample = 1 : no stable convergence

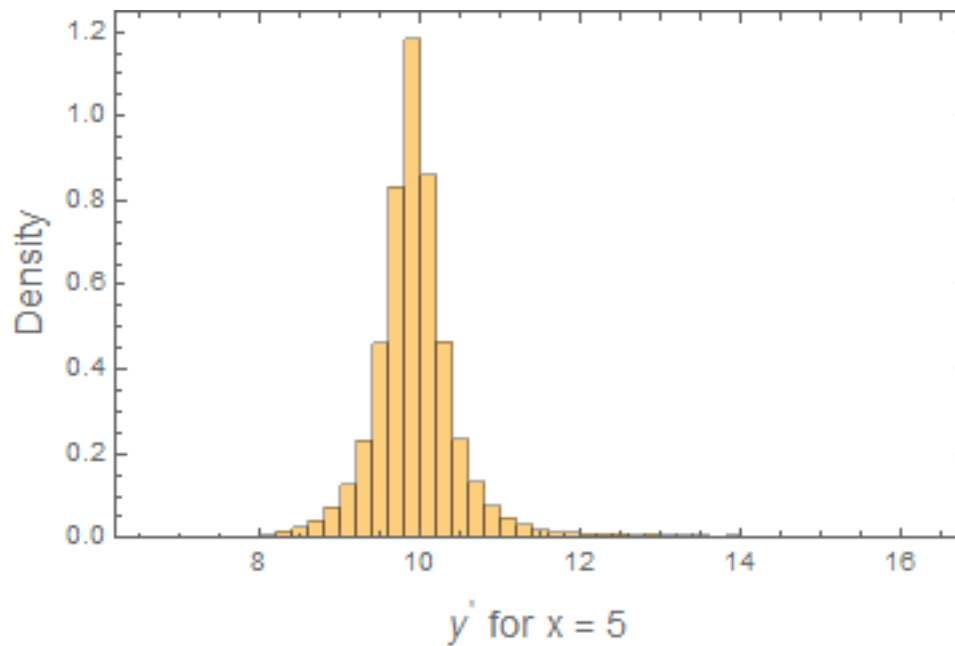
Subsample = 3 : acceptable fit

Parameter	True	2.50%	50%	97.50%
α	2	1.91	1.98	2.21
σ_1	1	0.75	0.89	1.68
σ_2	1	0.82	0.99	1.58
σ_3	0.1	0.054	0.331	1.04

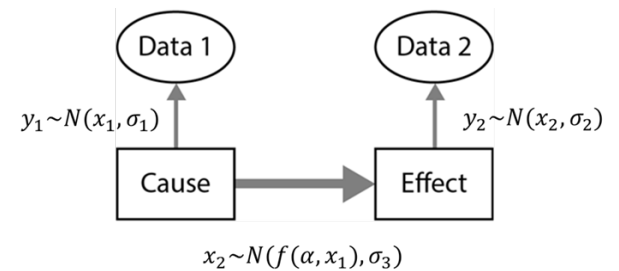


Bayesian example

Prediction for $x = 5$

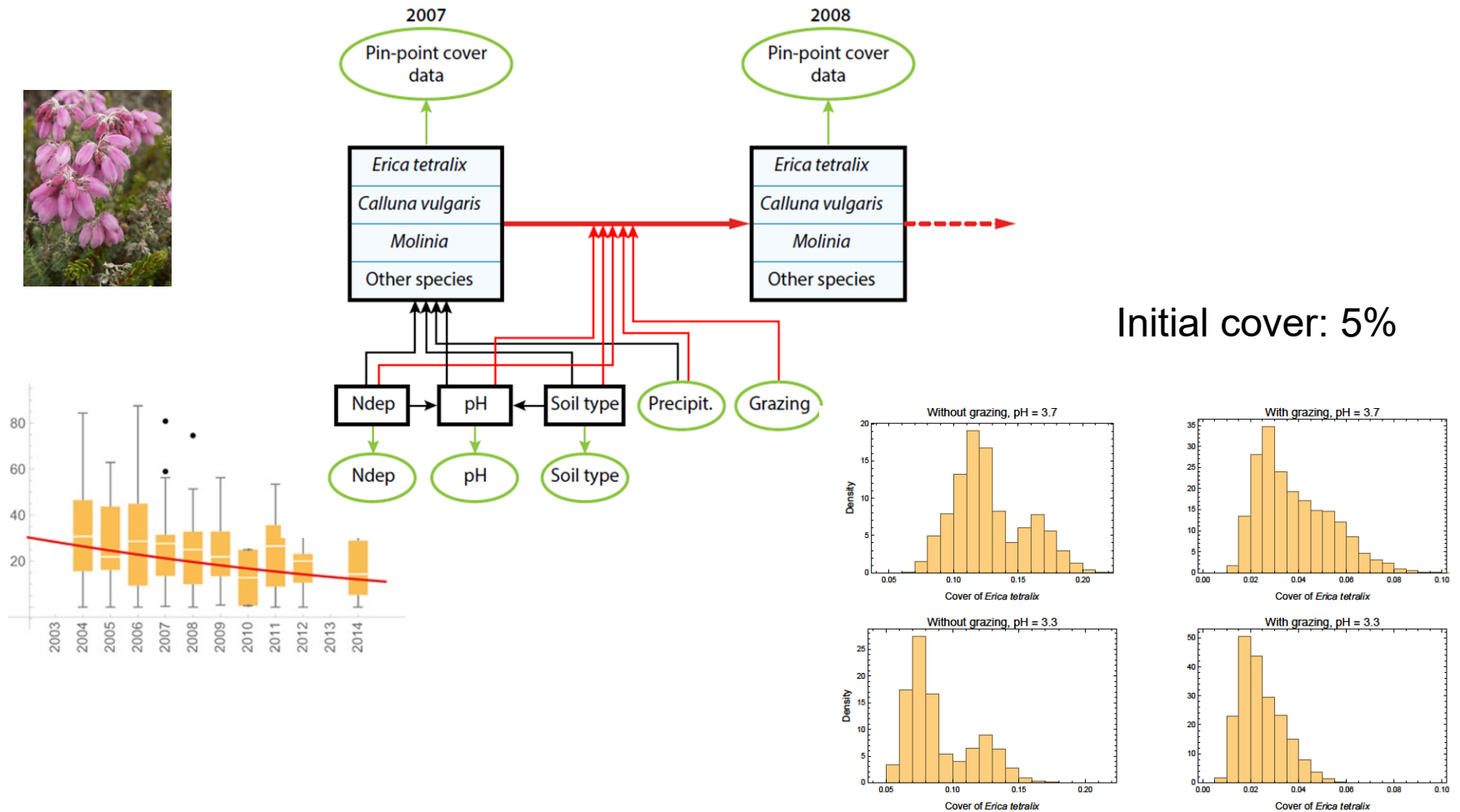


{8.87, 9.91, 11.56}



$$P(y') = P(\alpha) x + N(0, P(\sigma_3))$$

Ecological forecasts with uncertainties



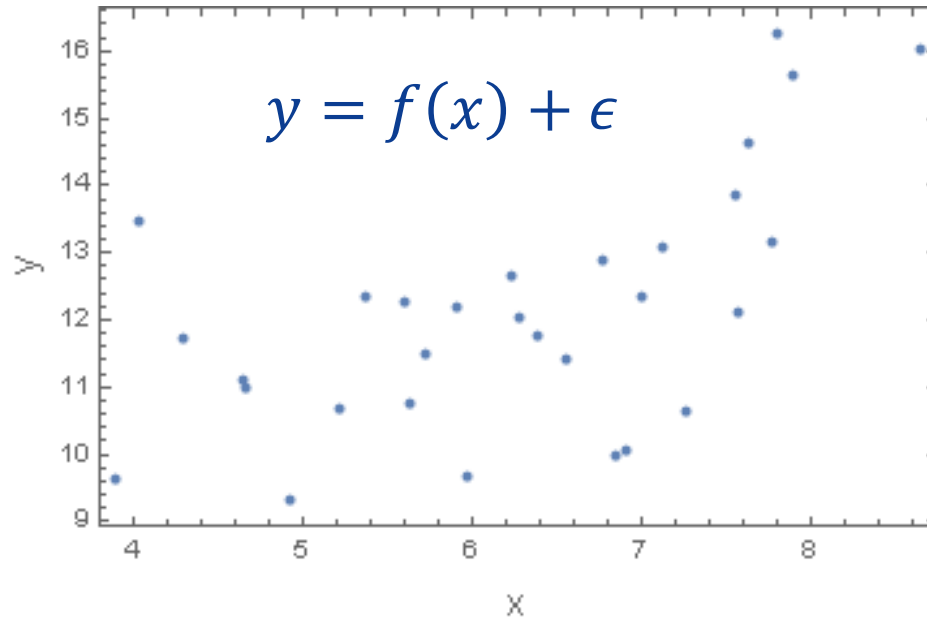
Initial cover: 5%



Uncertainty in Ecology

Christian Damgaard
Ecoscience
Aarhus University

Regression



x are assumed to be fixed!!!

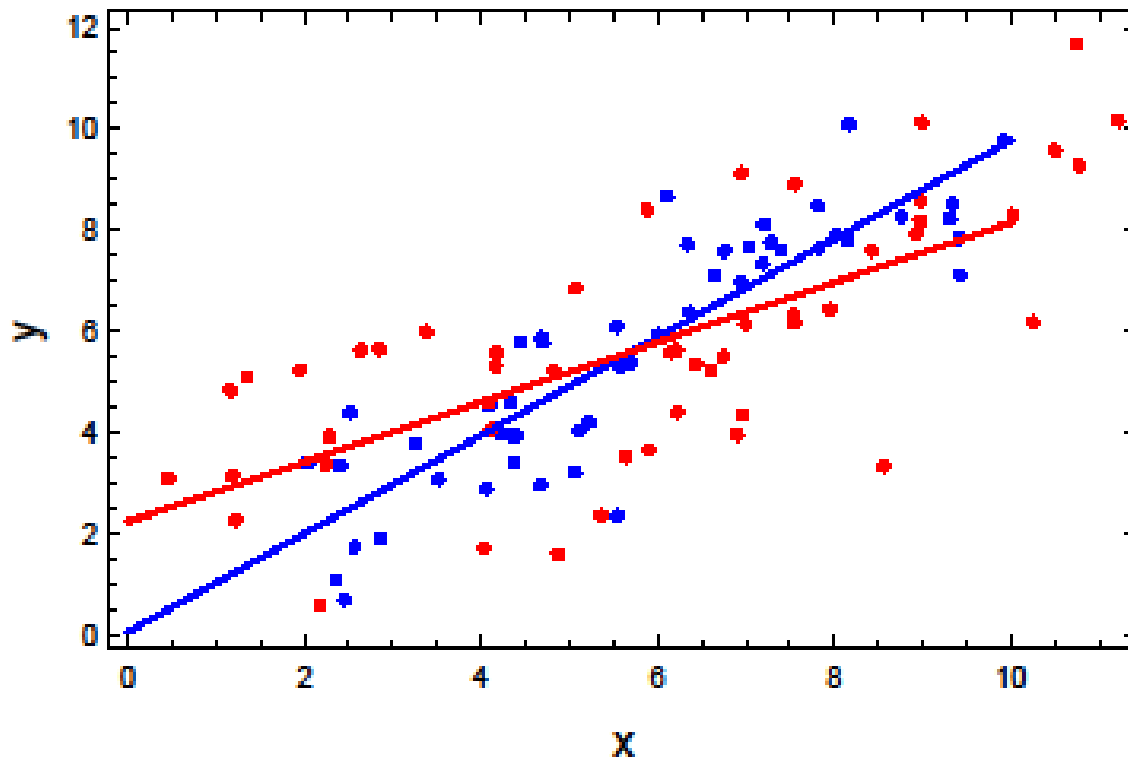
Often not correct to assume that there is no sampling- and measurement errors in x



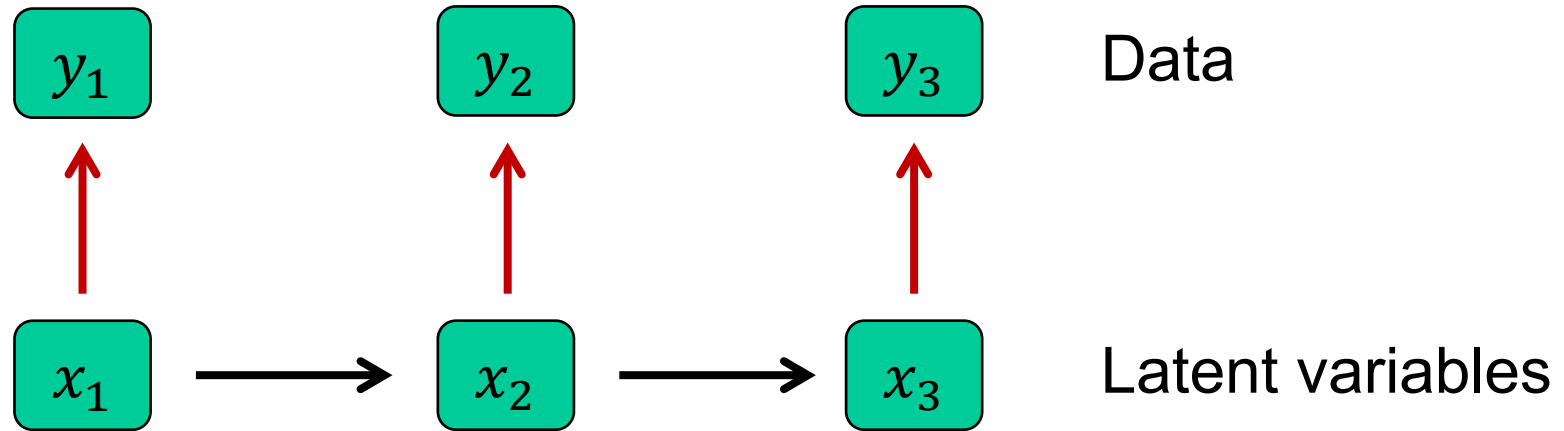
Regression dilution

Blue points: fixed X

Red points: random error in X



Hierarchical model of process



Process: $x_i \sim N(f(x_{i-1}, \theta), \sigma)$

Measurement: $y_i \sim \mathcal{D}(x_i, \tau)$

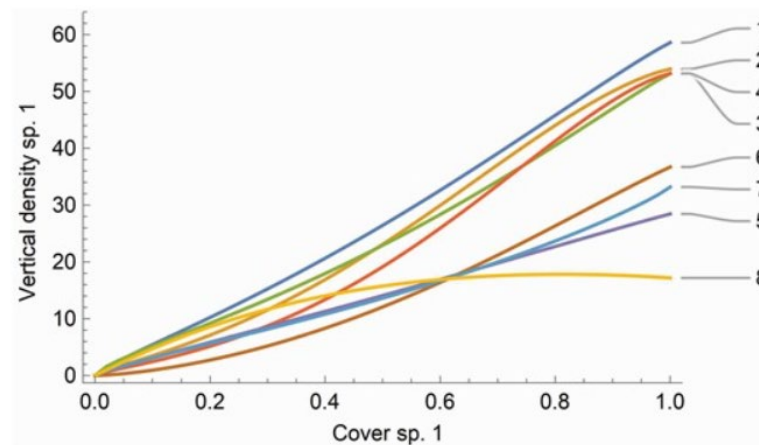
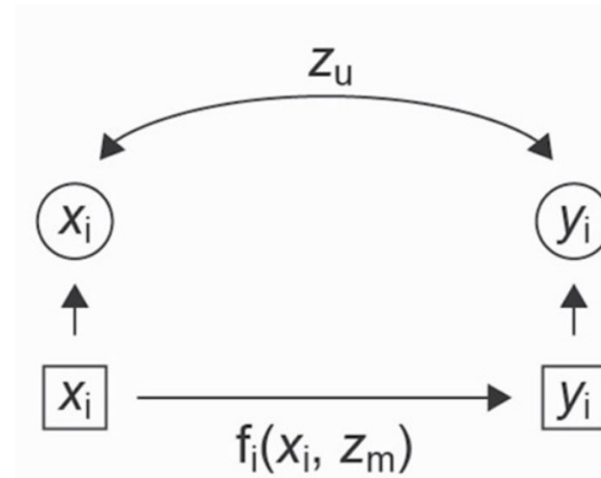
Separation of process error and **sampling- and measurement error**



Example: plant growth model

Ignoring sampling- and measurement error
 qualitative effects
 prediction bias

Model number	Frequency-dependence	Unmeasured variables	Measurement errors	WAIC	Effective number of parameters
1	-	-	-	935.89	13.86
2	+	-	-	925.99	13.25
3	-	+	-	851.05	11.23
4	+	+	-	849.88	14.00
5	-	-	+	705.20	83.39
6	+	-	+	594.80	92.01
7	-	+	+	692.06	83.84
8	+	+	+	700.73	90.44

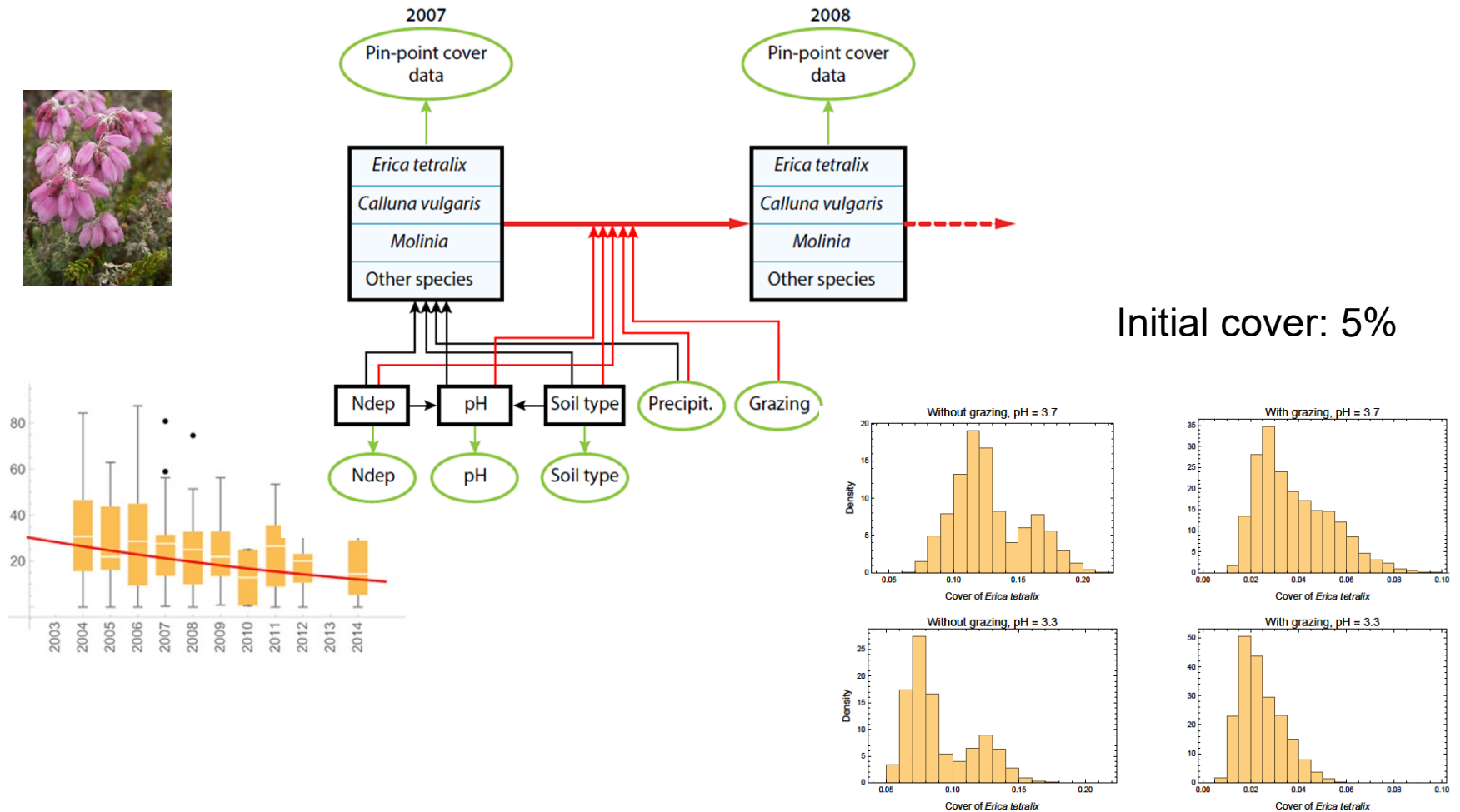


Quantitative ecological counseling

- **Ecology is complicated**
 - many interacting variables
 - different spatial scales and time lags
 - contingencies (large random events)
 - we will never have sufficient data
- **Use the data we do have access to**
- **Construct empirical models where the different sources of uncertainties are modelled**
 - test general hypothesis
- **Ecological forecasts**
 - quantitative counseling
 - local adaptive management plans

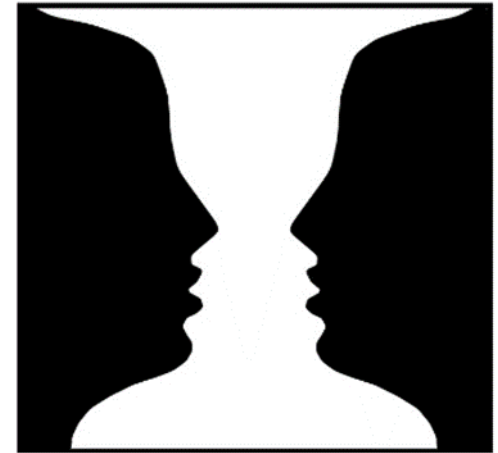


Ecological forecasts with uncertainties



Improve counseling

- **Problematic counseling**
 - general and qualitative
 - expert opinions
 - depend on researcher
 - insufficient treatment of uncertainty
- **Instead, make it an active strategy to**
 - make empirical ecosystem models where available process knowledge and data are used
 - quantify the different sources of uncertainties
 - local predictions
 - do not simplify and oversell - be honest about our uncertainties



Ecological predictions

Christian Damgaard
Ecoscience
Aarhus University

Questions

- How do you measure quality of predictions?
- How do you compare predictions on different state spaces?
- How do you know what is worth observing?
- How does quality of predictions vary as function of your data?




The effective sample size of non-random sampling

Christian Damgaard
Ecoscience
Aarhus University

Opinion

We need to talk about nonprobability samples

Robin J. Boyd ,^{1,*,@} Gary D. Powney,^{1,@} and Oliver L. Pescott^{1,@}

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 685–726
<https://doi.org/10.1214/18-AOAS1161SF>
© Institute of Mathematical Statistics, 2018

**STATISTICAL PARADISES AND PARADOXES IN BIG DATA (I):
LAW OF LARGE POPULATIONS, BIG DATA PARADOX,
AND THE 2016 US PRESIDENTIAL ELECTION¹**

BY XIAO-LI MENG
Harvard University



Non-random sampling

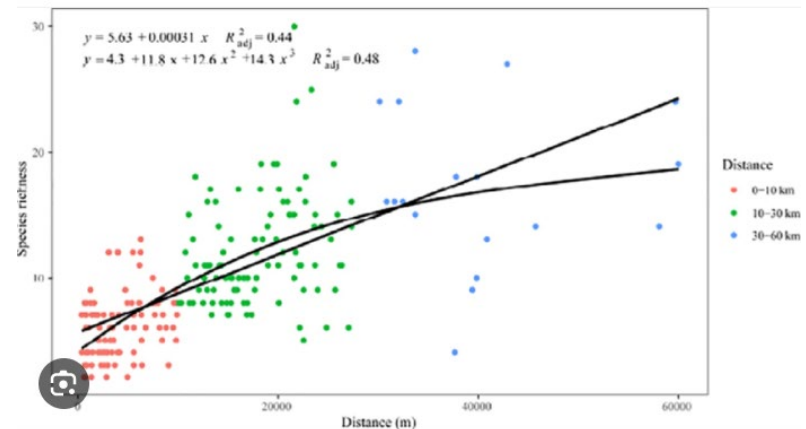
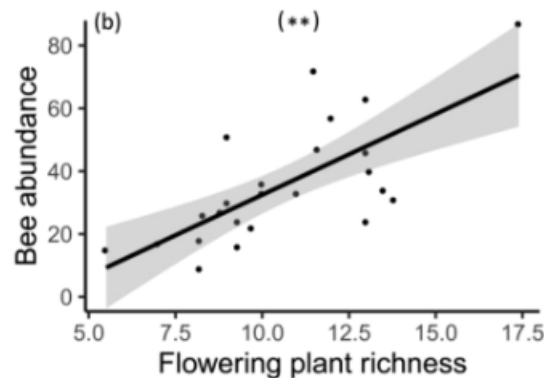
- **Citizen science**
- **Sampling at hotspots**
 - NATURA 2000 sites
- **Monitoring programs closely linked to policy**
 - metal concentration in water

- **Examples in Boyd et al (2023)**
 - problematic conclusions



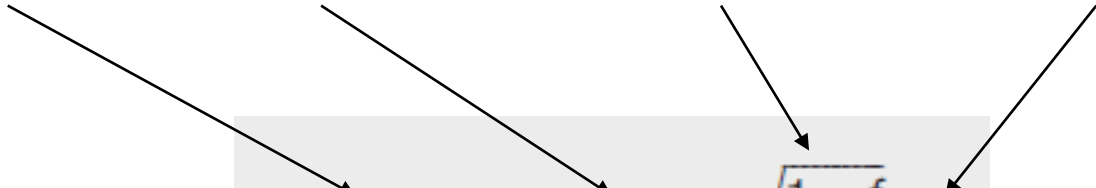
The probability of being sampled

- In a random sample the correlation between being in a sample and the variable of interest is ~ 0
- However, in a non-random sample there is a risk that the probability of being sampled is correlated with the variable of interest



Two interesting equations

bias = data quality × data quantity × problem difficulty


$$(\bar{Y}_n) - (\bar{Y}_N) = \rho(R, Y) \sqrt{\frac{1-f}{f}} \sigma_Y$$

$\rho(R, Y)$: correlation between being in a sample and Y
 f : sampling rate (n/N)

The effective sample size:

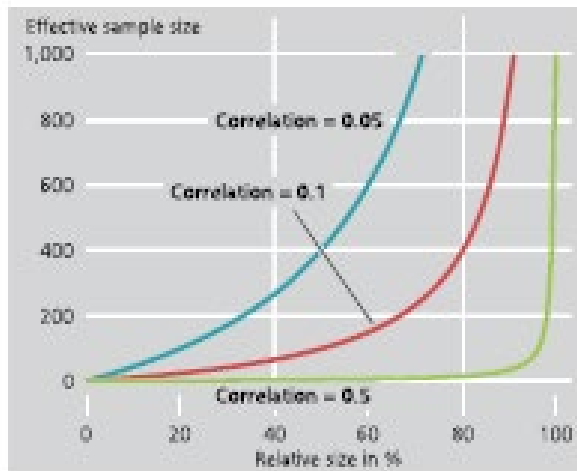
$$n_{\text{eff}} = \frac{f}{(1-f)} \frac{1}{E[\rho(R, Y)^2]} = \frac{n}{1-f} \frac{1}{E[\rho(R, Y)^2] N}$$



Being precisely wrong

- **Not just wrong, but precisely wrong**
 - Too small confidence interval centered on biased estimate
- **The effective sample size may be surprisingly low**

The effective sample size of a "Big Data" in terms of SRS size



Example with *Calluna vulgaris*

Calluna vulgaris (L.) Hull



True distribution



Sampled cells



Sampled cells with occupancy = 1

$N = 229,772$ 1 km
grid squares
 $n = 19,419$ 1 km grid
squares
 $\rho(\mathbf{R}, \mathbf{Y}) = -0.058$
 $n_{\text{eff}} = 28$
 $\bar{Y}_N = 0.299$
 $\bar{Y}_n = 0.213$

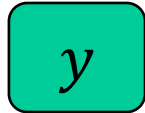
Way forward

- **Use random samples when possible**
- **Communication of possible bias in a nonrandom sample**
 - avoid being precisely wrong
 - use the effective sampling size
- **Possible to adjust for bias**
 - model the data generating process
 - “model-based inference”



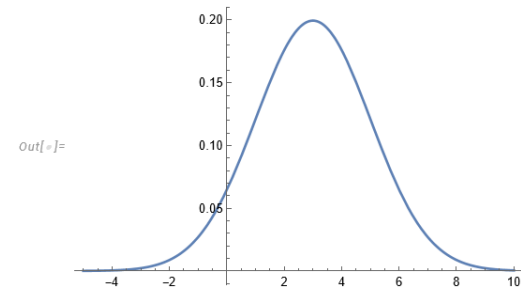
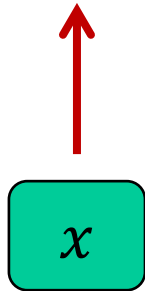
Estimering vha. af en målemodel

Data



{2.3, 3.4, MV, 3.4, UD, 2.4,...}

Estimeret fordeling af
af middelværdi



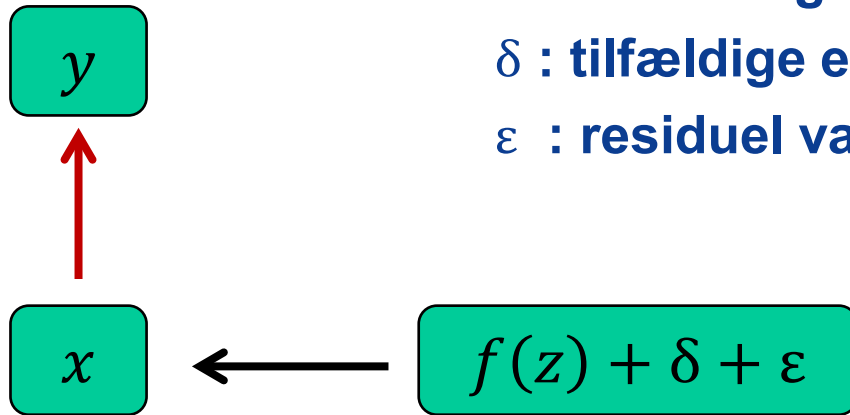
Ved estimering anvendes tilgængelig relevant information

detektionsgrænser

målt kovarians mellem stoffer

årsvariation

Forklaring og prognoser vha. af en strukturmodel



Z : sæson og forklarende variable

δ : tilfældige effekter

ϵ : residuel variation