Intro to Bayes

Upcoming

Bayesian programming
Error Propagation
Forecasting competition 1!

Things to keep in mind:
Be thinking about a forecasting project.
(more details to come)

I am going to look over your lab 4 before next class (next Thursday). Reach out if you need help, but I am gone Monday to Wednesday.

Intro to Bayes

Forecasting time:
~ 5% to build model make mean prediction.
~ 95% to fully quantify and propagate sources of uncertainty.

Why Bayes?

The era of raging debate between Bayesians and frequentists has mostly ended….

Why Bayes?

The era of raging debate between Bayesians and frequentists has mostly ended….

1. Well developed theoretical basis
2. Explicit quantification of probability (impossible with frequentist or ML)
3. Easy to partition uncertainty into different sources (difficult with frequentist or ML).
4. Easily handles missing data or uncertainty in data
5. Prior-posterior updating given new data allows for updating forecasts based on new data.

Why Bayes?

The era or raging debate between Bayesians and frequentists has mostly ended….

1. Well developed theoretical basis

2. Explicit quantification of probability (impossible with frequentist or ML)

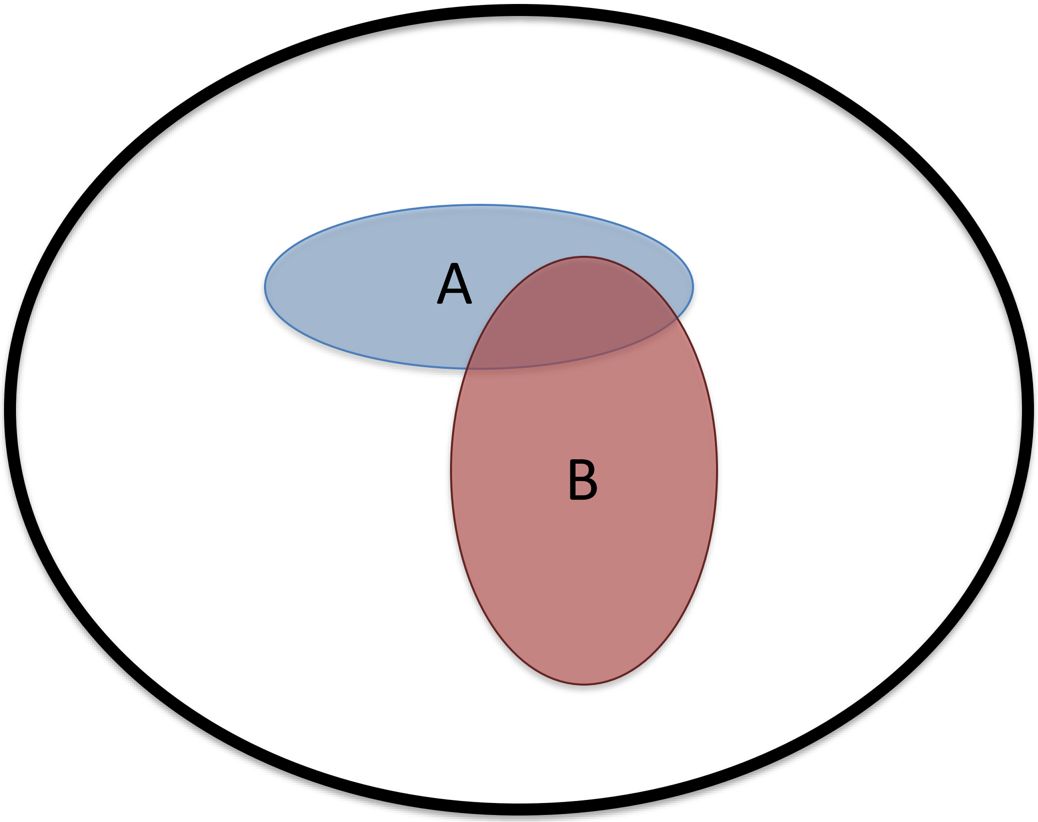
3. Easy to partition uncertainty into different sources (difficult or impossible with frequentist or ML).

4. Easily handles missing data or uncertainty in data

5. Prior-posterior updating given new data allows for updating forecasts based on new data.

This doesn't mean that frequentist or ML approaches can't make valid and useful forecasts, Bayesian approaches just have a number of distinct advantages.

Probability



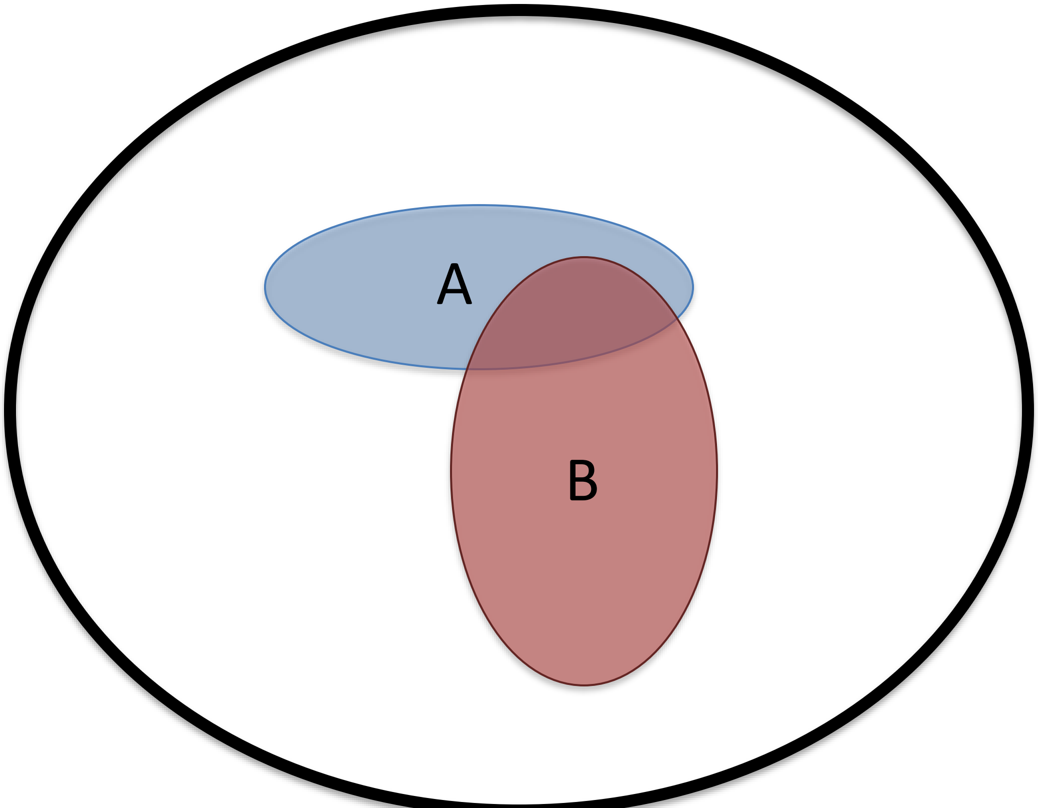
Marginal Probability

$$P(A) = Area\ of\ A$$

$$P(B) = Area\ of\ B$$

Adapted from Hobbs and Hooten 2015

Probability



Marginal Probability

$$P(A) = Area\ of\ A$$

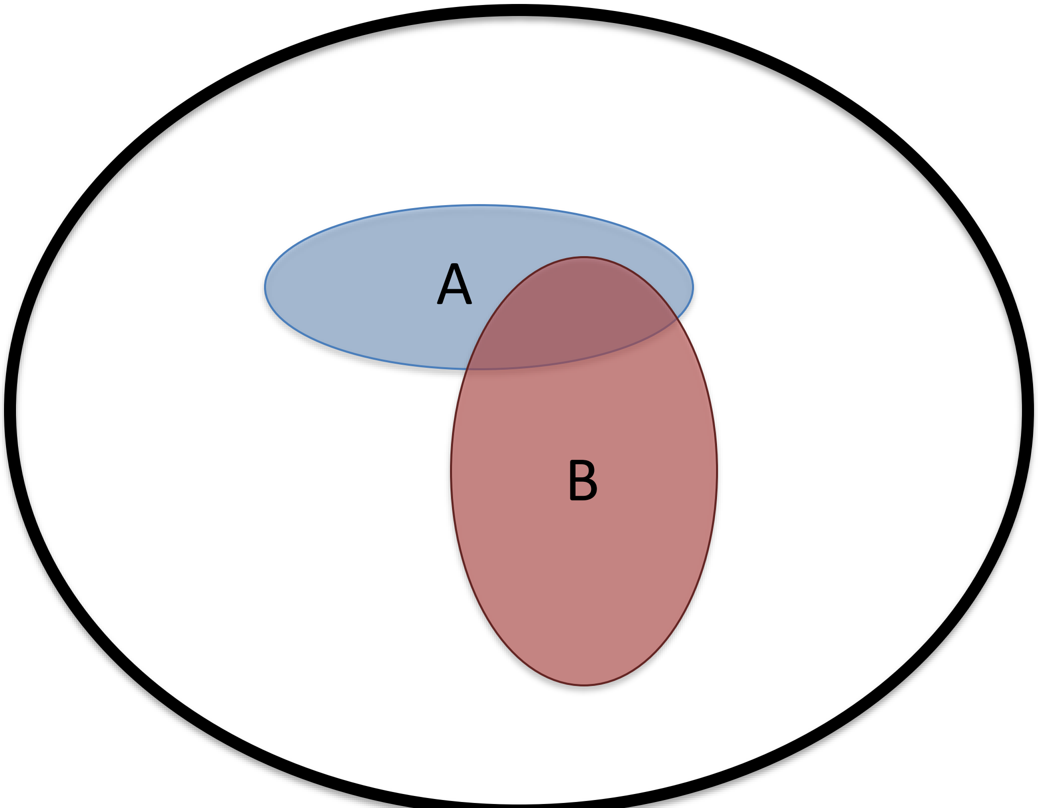
$$P(B) = Area\ of\ B$$

Joint Probability

$$P(A, B) = Shared\ Area\ of\ A\ \&\ B$$

Adapted from Hobbs and Hooten 2015

Probability



Marginal Probability

$$P(A) = Area\ of\ A$$

$$P(B) = Area\ of\ B$$

Joint Probability

$$P(A,B) = Shared\ Area\ of\ A\ \&\ B$$

Conditional Probability

$$P(B|A) = \frac{Shared\ Area\ of\ A\ \&\ B}{Area\ of\ A}$$

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

Adapted from Hobbs and Hooten 2015

Probability

$$P(B|A) = \frac{P(A,B)}{P(A)} \xrightarrow{\hspace{3em}} P(A,B) = P(B|A)\Pr(A)$$

Algebraic rearrangement

$$P(A|B) = \frac{P(A,B)}{P(B)} \xrightarrow{\hspace{3em}} P(A,B) = P(A|B)\Pr(B)$$

Probability

$$P(B|A) = \frac{P(A,B)}{P(A)} \longrightarrow P(A,B) = P(B|A)\mathrm{Pr}(A)$$

Algebraic rearrangement

$$P(A|B) = \frac{P(A,B)}{P(B)} \longrightarrow P(A,B) = P(A|B)\mathrm{Pr}(B)$$

Intro to Bayesian Inference

The goal is to ”learn” about parameters given observed data

Parameters Data

$$[\theta|y] = \frac{[\theta, y]}{[y]} \qquad [\theta, y] = [y|\theta]\,[\theta]$$

$$[\theta|y] = \frac{[y|\theta]\,[\theta]}{[y]}$$

Notation: $P(A)$ is the same as $[A]$

Intro to Bayesian Inference

The goal is to ”learn” about parameters given observed data

Likelihood Prior

Posterior

$$[\theta|y] = \frac{[y|\theta]\,[\theta]}{[y]}$$

Normalizing Constant

Intro to Bayesian Inference

The goal is to ”learn” about parameters given observed data

$$\underbrace{[\theta|y]}_{\text{Posterior}} \propto \underbrace{[y|\theta]}_{\text{Likelihood}} \underbrace{[\theta]}_{\text{Prior}}$$

Intro to Bayesian Inference

The goal is to ”learn” about parameters given observed data

Posterior Likelihood Prior

$$[\theta|y] \propto [y|\theta]\,[\theta]$$

Updated probability of the parameter value given the data

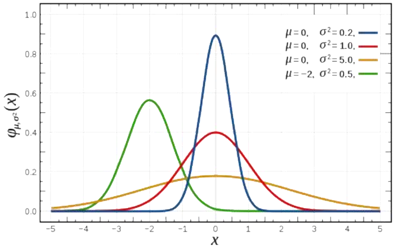
Probability of the data given parameter value

Probability of the parameter value

Probability Distributions

Continuous

Normal: Any value from -Inf to Inf, Mean is independent of variance



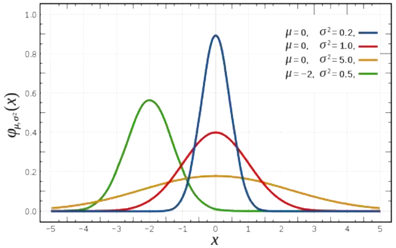
Daily carbon flux

Discrete

Probability Distributions

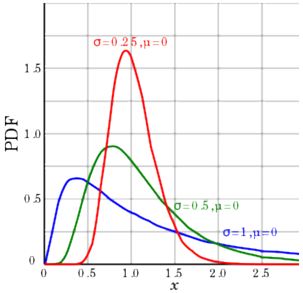
Continuous

Normal: Any value from -Inf to Inf, Mean is independent of variance



Daily carbon flux

Log Normal: Any value from >0 to Inf, Variance scales with mean



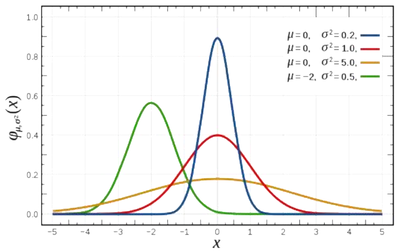
Population Density

Discrete

Probability Distributions

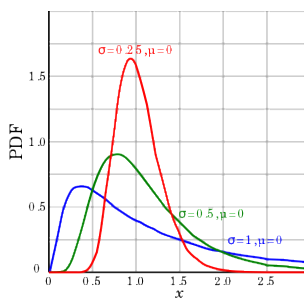
Continuous

Normal: Any value from -Inf to Inf, Mean is independent of variance



Daily carbon flux

Log Normal: Any value from >0 to Inf, Variance scales with mean



Population Density

Beta: Any value from >0 to <1, Variance scales with mean

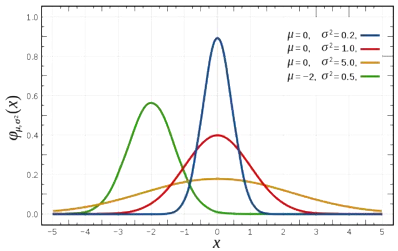
Survival Prob.

Discrete

Probability Distributions

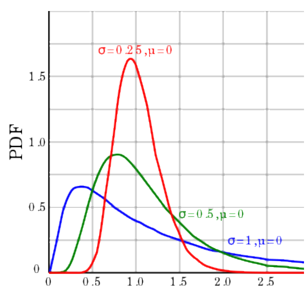
Continuous

Normal: Any value from -Inf to Inf, Mean is independent of variance



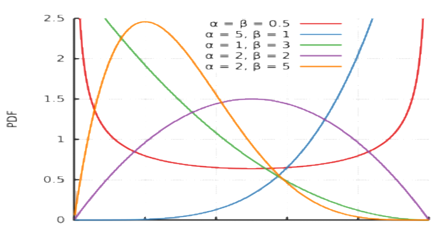
Daily carbon flux

Log Normal: Any value from >0 to Inf, Variance scales with mean



Population Density

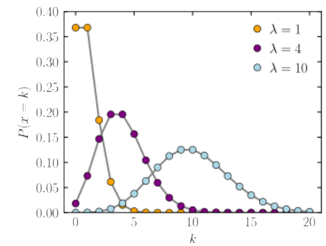
Beta: Any value from >0 to <1, Variance scales with mean



Survival Prob.

Discrete

Poisson: Any value from 0 to Inf, Mean=Variance

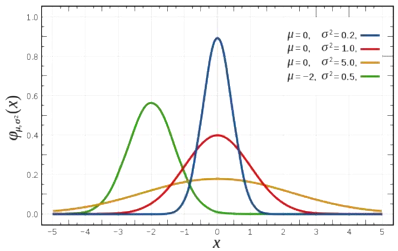


Population Counts

Probability Distributions

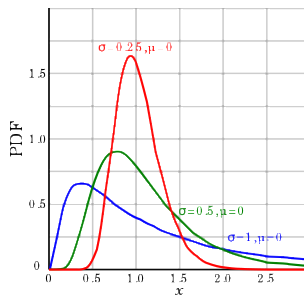
Continuous

Normal: Any value from -Inf to Inf, Mean is independent of variance



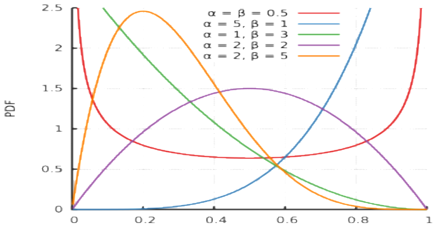
Daily carbon flux

Log Normal: Any value from >0 to Inf, Variance scales with mean



Population Density

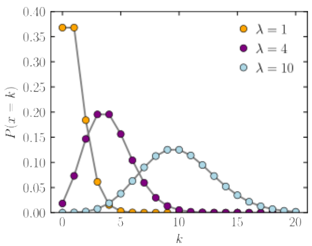
Beta: Any value from >0 to <1, Variance scales with mean



Survival Prob.

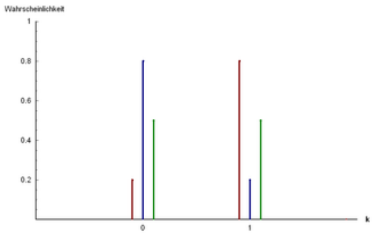
Discrete

Poisson: Any value from 0 to Inf, Mean=Variance



Population Counts

Bernoulli: 0 or 1, Single parameter Prob. That event will occur

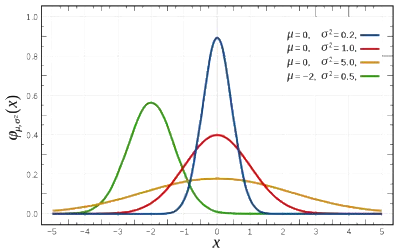


Individual Survival

Probability Distributions

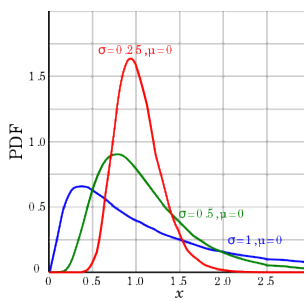
Continuous

Normal: Any value from -Inf to Inf, Mean is independent of variance



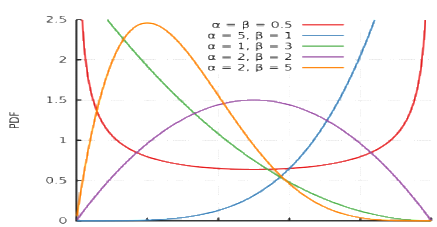
Daily carbon flux

Log Normal: Any value from >0 to Inf, Variance scales with mean



Population Density

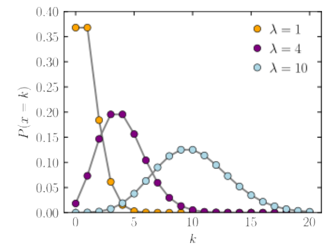
Beta: Any value from >0 to <1, Variance scales with mean



Survival Prob.

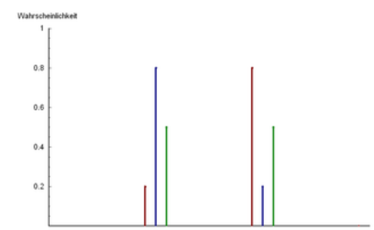
Discrete

Poisson: Any value from 0 to Inf, Mean=Variance



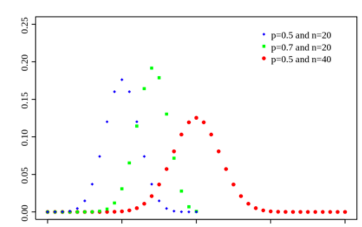
Population Counts

Bernoulli: 0 or 1, Single parameter Prob. That event will occur



Individual Survival

Binomial: 0 to Inf., Outcome of multiple Bernoulli events

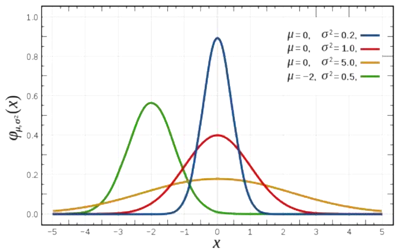


Deaths in a population

Probability Distributions

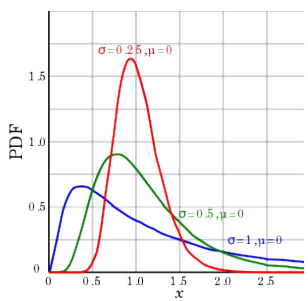
Continuous

Normal: Any value from -Inf to Inf, Mean is independent of variance



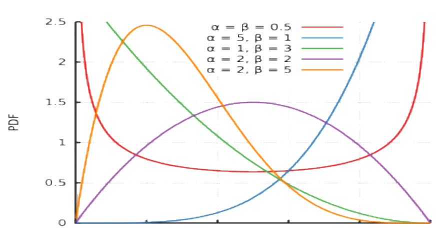
Daily carbon flux

Log Normal: Any value from >0 to Inf, Variance scales with mean



Population Density

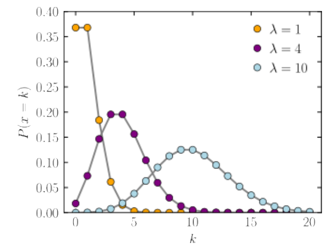
Beta: Any value from >0 to <1, Variance scales with mean



Survival Prob.

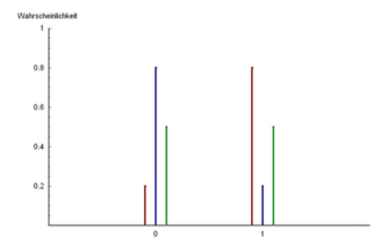
Discrete

Poisson: Any value from 0 to Inf, Mean=Variance



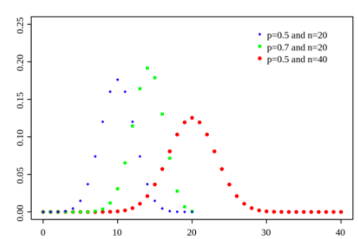
Population Counts

Bernoulli: 0 or 1, Single parameter Prob. That event will occur



Individual Survival

Binomial: 0 to Inf., Outcome of multiple Bernoulli events



Deaths in a population

Picking a likelihood

Likelihood should match the data and data generating process.

Picking a likelihood

Likelihood should match the data and data generating process.

What matters is the conditional distribution not the raw data distribution.

Picking a likelihood

Likelihood should match the data and data generating process.

What matters is the conditional distribution not the raw data distribution.

Questions to ask:

Are the data continuous or discrete?

What are the range of possible values that data can take?

How does the variance change as a function of the mean?

Picking a likelihood

Likelihood should match the data and data generating process.

$Normal(\mu, \sigma^2)$

$Poisson\ (\lambda)$

Picking a likelihood

Likelihood should match the data and data generating process.

$$Normal(\mu, \sigma^2)$$

- Continuous
- Values from –Inf to Inf
- Variance constant and independent of mean
- No Skew

$$Poisson\ (\lambda)$$

- Discrete
- Values from 0 to Inf
- Variance=mean
- Skewed at low values

Picking priors

Priors represent existing belief or knowledge about a parameter.

Picking priors

Priors represent existing belief or knowledge about a parameter.

What range can a parameter take mathematically?
What values are biologically realistic given our current understanding?
Is there an appropriate conjugate prior?

Picking priors

Priors represent existing belief or knowledge about a parameter.

$$Normal(\mu, \sigma^2)$$

$$Poisson\ (\lambda)$$

Continuous, Can take on any value.

Continuous, must be positive

Continuous, must be positive

Picking priors

Priors represent existing belief or knowledge about a parameter.

$$Normal(\mu, \sigma^2)$$

Continuous, Can take on any value.

$$Normal(\mu_p, \sigma_p^2)$$

Continuous, must be positive

Gamma(a_p, b_p)

$$Poisson\ (\lambda)$$

Continuous, must be positive

Gamma(a_p, b_p)

Picking priors

Priors represent existing belief or knowledge about a parameter.

When do priors need to be informative? When can they be uninformative?

Some parameters can almost reliably estimated by data. E.g.: Regression parameters

Some parameters will rarely ever be indefinable without prior information. E.g.: Observation Error

Picking priors

Priors represent existing belief or knowledge about a parameter.

When do priors need to be informative?
When can they be uninformative?

In practice, selecting priors can be part of the model development process.

Start with vague priors. If a parameter is not identifiable, additional data may be needed.

Why are priors important?

First, they allow for the incorporation of expert knowledge.

Second, they often provide essential information to put risk or forecast in context.

Why are priors important?

First, they allow for the incorporation of expert knowledge.

Second, they often provide essential information to put risk or forecast in context.

Disease test with 95% sensitivity is given to 100,000 people, 7,000 come back positive. (Data)

Why are priors important?

First, they allow for the incorporation of expert knowledge.

Second, they often provide essential information to put risk or forecast in context.

Disease test with 95% sensitivity is given to 100,000 people, 7,000 come back positive. (Data)

The disease has a 1% prevalence (prior probability)

Why are priors important?

First, they allow for the incorporation of expert knowledge.

Second, they often provide essential information to put risk or forecast in context.

Disease test with 95% sensitivity is given to 100,000 people, 7,000 come back positive. (Data)

The disease has a 1% prevalence (prior probability)

P(Disease | Positive)= (P(Positive | Disease)*P(Disease))/P(Positive)

Why are priors important?

First, they allow for the incorporation of expert knowledge.

Second, they often provide essential information to put risk or forecast in context.

Disease test with 95% sensitivity is given to 100,000 people, 7,000 come back positive. (Data)

The disease has a 1% prevalence (prior probability)

P(Disease | Positive)= (P(Positive | Disease)*P(Disease))/P(Positive)

(.95*.01)/0.07=.135

Only 13.5% of people that test positive have the disease!

Posterior Calculations

If prior is conjugate, we can calculate the posterior analytically.

Conjugate: Posterior distribution is the same as prior with parameters updated based on data

Posterior Calculations

If prior is conjugate, we can calculate the posterior analytically.

Conjugate: Posterior distribution is the same as prior with parameters updated based on data

Likelihood $\mathbf{y} \sim Poisson\,(\lambda)$

Prior $\lambda \sim Gamma\,(a_{prior}, b_{prior})$

Posterior $\lambda|\mathbf{y} \sim Gamma\,(a_{prior} + \sum_{i=1}^{n} y_i\,, b_{prior} + n)$

The mathematical calculations underlying this can be found online in a number of sources.

Complex problems

The power of Bayesian inference comes in modeling complex problems

Examples of complex problems: Multiple datasets, unobserved (latent) states, autocorrelation, etc.

Process Parameter

True State

Observed State (Data)

$$[z, \sigma_1, \theta, \sigma_2 | y]$$

Observation Error

Process Error

Complex problems

The power of Bayesian inference comes in modeling complex problems

Examples of complex problems: Multiple datasets, unobserved (latent) states, autocorrelation, etc.

Process Model

$$[z,\sigma_1,\theta,\sigma_2|y] \propto [y|z,\sigma_1][z|\theta,\sigma_2][\sigma_1][\theta][\sigma_2]$$

Observation Model

Parameter Models

Complex problems

The power of Bayesian inference comes in modeling complex problems

Examples of complex problems: Multiple datasets, unobserved (latent) states, autocorrelation, etc.

Process Model

$$[z, \sigma_1, \theta, \sigma_2 | y] \propto [y | z, \sigma_1][z | \theta, \sigma_2][\sigma_1][\theta][\sigma_2]$$

Observation Model

Parameter Models

This is an example of a hierarchical model

DAG (Directed Acyclic Graphs)

Thinking about and breaking down complex, hierarchical models

Process Model

$$[z, \sigma_1, \theta, \sigma_2 | y] \propto [y | z, \sigma_1][z | \theta, \sigma_2][\sigma_1][\theta][\sigma_2]$$

Observation Model

Parameter Models

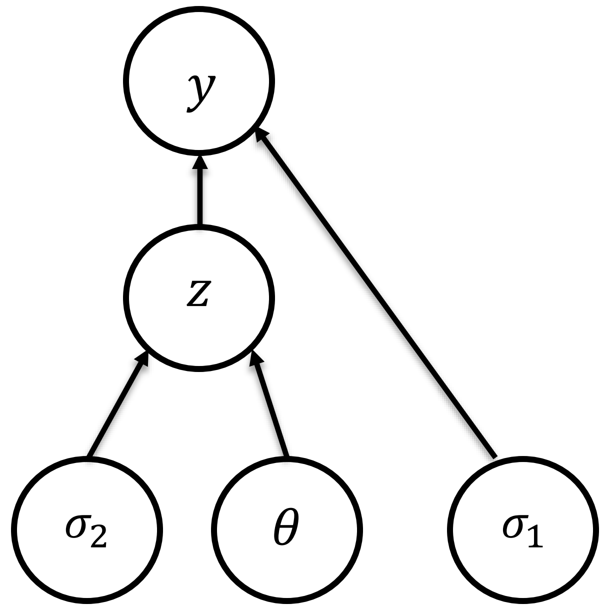
DAG (Directed Acyclic Graphs)

Thinking about and breaking down complex, hierarchical models

Observation Model

Process Model

Parameter Model



Process Model

$$[z, \sigma_1, \theta, \sigma_2 | y] \propto [y|z, \sigma_1][z|\theta, \sigma_2][\sigma_1][\theta][\sigma_2]$$

Observation Model

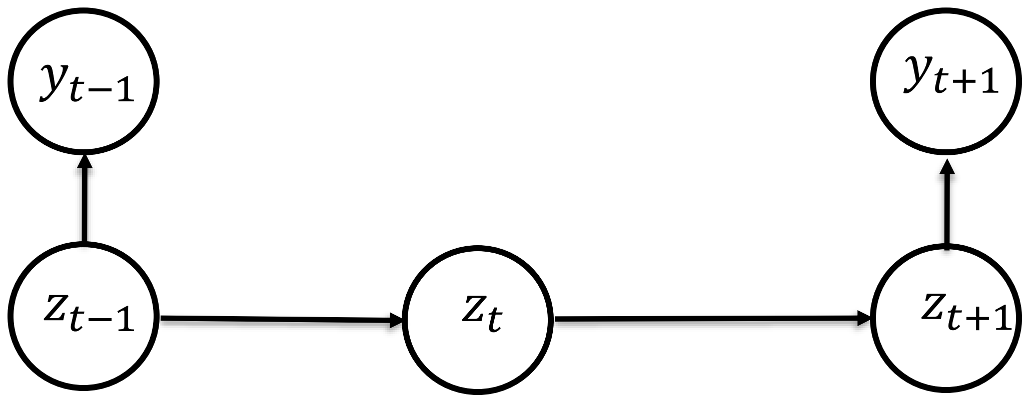
Parameter Models

Example 2: Missing data time series

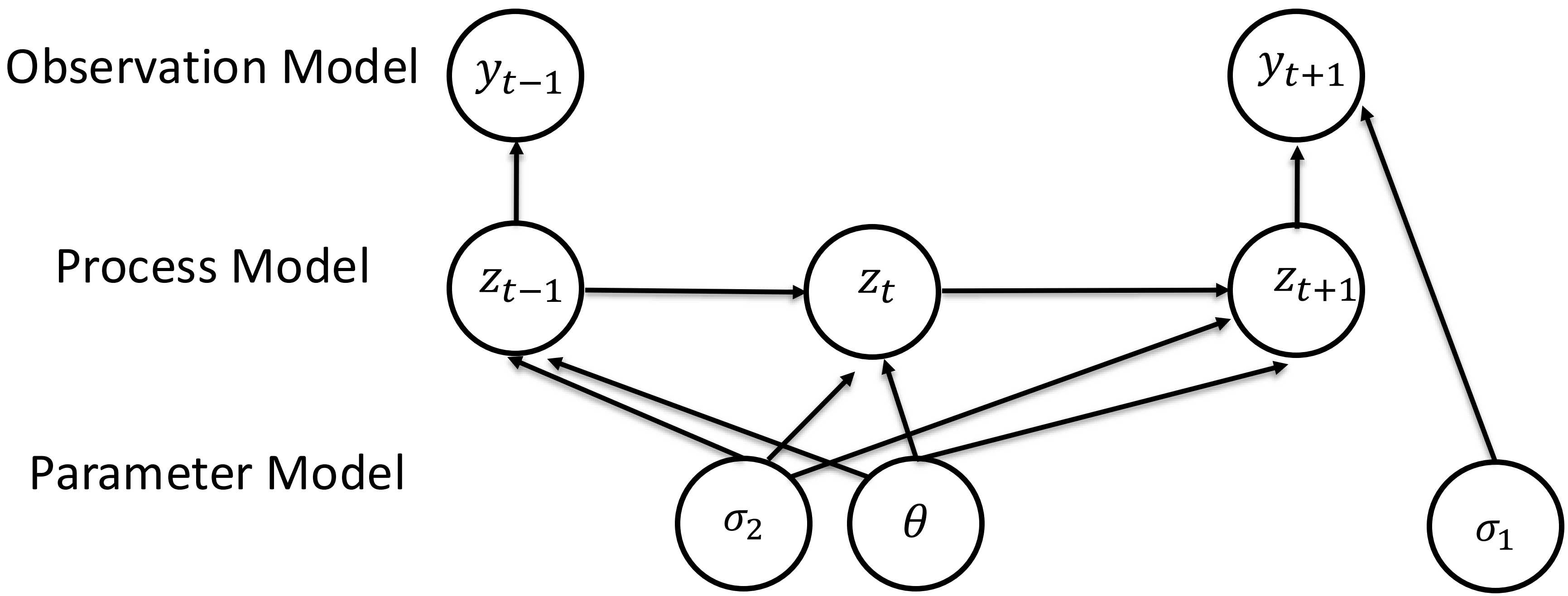
Observation Model

Process Model

Parameter Model



Example 2: Missing data time series



Example 3: Making a forecast!