Macroeconometrics

Lecture 5

Bayesian estimation

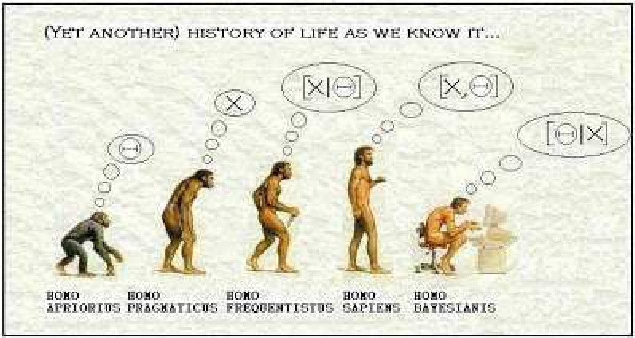
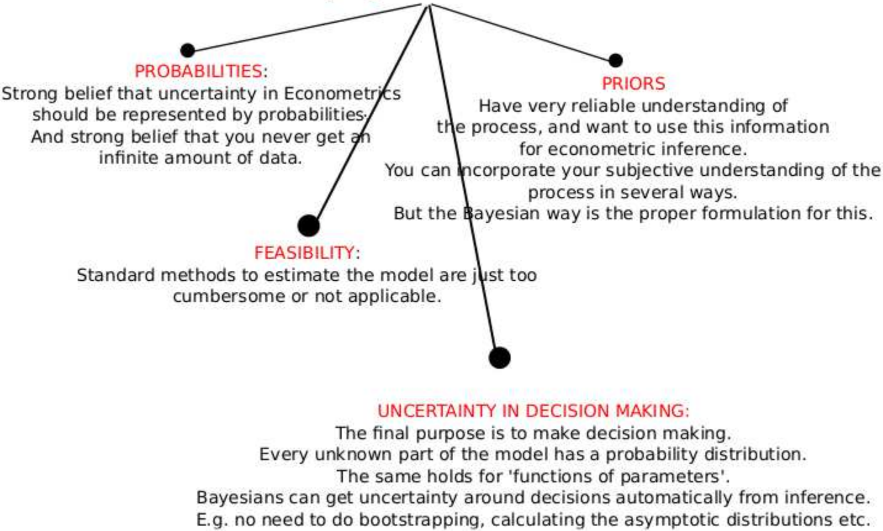


Figure: No offence!!!!

Why Bayesian Econometrics



PROBABILITIES:
Strong belief that uncertainty in Econometrics should be represented by probabilities.
And strong belief that you never get an infinite amount of data.

PRIORS
Have very reliable understanding of the process, and want to use this information for econometric inference.
You can incorporate your subjective understanding of the process in several ways.
But the Bayesian way is the proper formulation for this.

FEASIBILITY:
Standard methods to estimate the model are just too cumbersome or not applicable.

UNCERTAINTY IN DECISION MAKING:
The final purpose is to make decision making.
Every unknown part of the model has a probability distribution.
The same holds for 'functions of parameters'.
Bayesians can get uncertainty around decisions automatically from inference.
E.g. no need to do bootstrapping, calculating the asymptotic distributions etc.

Where does the name ‘Bayesian’ come from?

Reverend Thomas Bayes (1702–1761)
(Bayes Theorem 1763)

For two events A and B: $p(A, B) = p(A|B)p(B)$

Books and references

- Zellner, A. (1971), An Introduction to Bayesian Inference in Econometrics, Wiley.
- Geweke, J. (2005), Contemporary Bayesian Econometrics and Statistics, Wiley.
- Koop, G. (2003), Bayesian Econometrics, Wiley.
- Canova, F. (2007), Methods for Applied Macroeconomic Research, Princeton University Press, (Chapters 9-11)
- Fernandez-Villaverde, J.: check his webpage
- (Classical) Hamilton, J. (1994), Times Series Analysis, Princeton University Press.

Your first mountain-bike trip

Possible outcomes:

- FUN with probability p,
- HOSPITAL with probability $1 - p$.

Before the trip what do you think p is?

- $p < 0.5 \rightarrow$ do not go to the trip
- $p \geq 0.5 \rightarrow$ go to the trip.

You did go to that trip! Ended up in the HOSPITAL

Will you go again?

Not a Bayesian → You will not go again.

$$pr(y|p) = p^0(1-p)^1 = 1 - p$$

Maximize your likelihood:

$$p = 0$$

You will definitely end up in the HOSPITAL next time.

Your first mountain-bike trip

You did go to that trip! Ended up in the HOSPITAL
Will you go again?

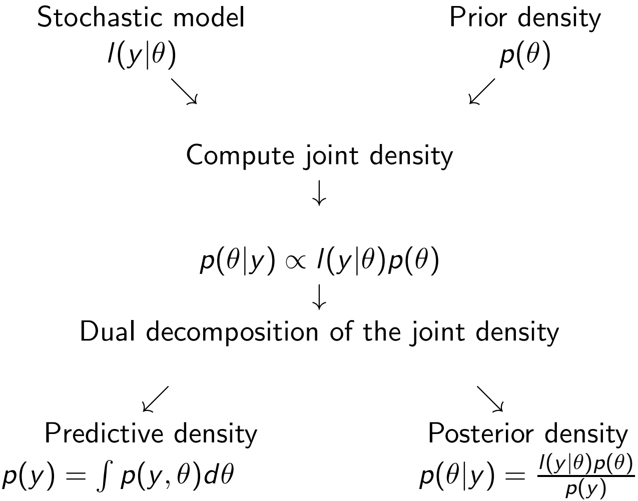
Bayesian version
Prior for p: $pr(p) = 2p$

$$pr(p|y) \propto pr(y|p)pr(p) = p^0(1-p)^1 \times 2p = 2(1-p)p$$

You still have a choice to make.

Note: $\propto$ denotes that the right hand side is a probability density kernel (pdf apart from a scaling constant).

Bayes' rule



Stochastic model

$l(y|\theta)$

Prior density

$p(\theta)$

Compute joint density

$$p(\theta|y) \propto l(y|\theta)p(\theta)$$

Dual decomposition of the joint density

Predictive density

$$p(y) = \int p(y,\theta)d\theta$$

Posterior density

$$p(\theta|y) = \frac{l(y|\theta)p(\theta)}{p(y)}$$

Bayes' rule

- $p(y)$ plays the role of normalizing constant (just like $\sqrt{2\pi}$ in the case of the normal distribution).
- Focus on the posterior kernel $p(\theta|y) \propto l(y|\theta)p(\theta)$.
- $p(y)$ is very relevant for model selection and forecasting.

Helicopter Tour Comparison

Frequentist inference	**Bayesian inference**
Parameters θ are fixed unknown constraints. There is a true value $\theta = \theta_0$	*Parameters θ are stochastic variables.* One defines a priori distribution on the parameter space.
Data y are used to estimate θ and check validity of postulated model, by comparing data with data set from model	*Data y* are used as evidence to update the state of the mind: data transform the prior into the posterior distribution by the likelihood.

Helicopter Tour Comparison

Frequentist inference	**Bayesian inference**
Objective concept of probability: a prob. is *the* fraction of occurrences when a process is repeated infinitely often.	*Subjective concept of probability*: a prob. is *a* degree of belief that an event occurs.
One can use the *maximum likelihood* estimator as an estimator $\hat{\theta}_{ML}$ of θ.	One uses *Bayes' theorem* to obtain the posterior distribution of θ.

A gentle introduction to Bayesian Econometrics(1)

- Take a linear regression model

$$y_t = \theta x_t + \epsilon_t, \qquad \epsilon_t \sim NID(0, \sigma^2)$$

  equivalently,

$$y = x\theta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I).$$

- Example:
  - x_t is (log) net income at time t,
    y_t is (log) consumption at time t
  - Model parameters: θ, σ^2

A gentle introduction to Bayesian Econometrics (2)

- What do we know from standard econometrics?
- Use OLS $\hat{\theta} = (x'x)^{-1}x'y$
- Use confidence intervals to address uncertainty in estimation.
- α% confidence interval for OLS estimate

  includes the true value θ with probability α

- Main message: *There is a true value θ!!*

  We are getting an estimate of this true value.

A gentle introduction to Bayesian Econometrics(3)

- How ‘many people’ explain Bayesian philosophy:

  There is no true/fixed value for parameters θ, σ^2*, these are random parameters.*

- The above statement is actually PARTIALLY TRUE:

  Parameters can be random or constant, but Bayesians treat any uncertainty, including parameter uncertainty, using probabilistic measures.

  (θ, σ^2) *may have a true value, but it is unknown, hence the need to define a probability distribution to account for this uncertainty.*

A gentle introduction to Bayesian Econometrics(4)

- Implication for a statistician/econometrician:

  Define a probability measure for (θ, σ^2) based on your subjective view.

  Then let the data talk: update your subjective view.

- Back to the example:

  My subjective view on θ: $\theta \in [0, 1]$.

  - I know people do not decrease their consumption when they have more money:

    $$\theta \geq 0$$

  - I know people spend responsibly (not more than the amount they get)

    $$\theta \leq 1$$

  I am quite uncertain about what the exact value is: define a uniform distribution on $\theta \in [0, 1]$.

A gentle introduction to Bayesian Econometrics(5)

- Implication for a statistician/econometrician:

  Define a probability measure for (θ, σ^2) based on your subjective view.

  Then let the data talk: update your subjective view.

- Back to the example:

  My subjective view on θ:

  a uniform distribution on $\theta \in [0, 1]$.

$$p(\theta) \sim \text{Unif(0,1)}.$$

A gentle introduction to Bayesian Econometrics(6)

- Implication for a statistician/econometrician:

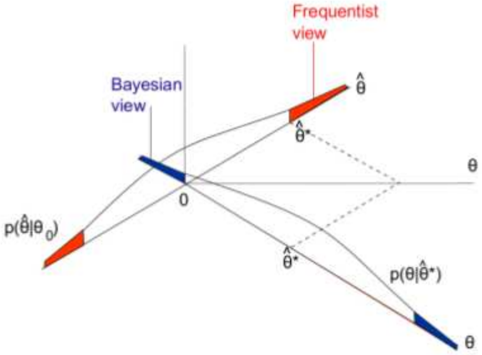
  Define a probability measure for (θ, σ^2) based on your subjective view.

  Then let the data talk: update your subjective view.

- Back to the example:
- How do the 'data talk': Famous Bayes' rule for two events A and B

$$p(A, B) = p(A|B)p(B)$$

Difference between frequentist an Bayesian



- Frequentist one provides a ‘probability density’ of the estimated parameter.
- Only Bayesian view provides a ‘probability density’ for the true parameter value.

⇒ Any frequentist econometrician who does more than ‘rejecting’ or ‘not rejecting’ a hypothesis is using the Bayesian interpretation of his results.
E.g. a coefficient being ‘strongly significant’ is already a Bayesian statement.

Bayes' Theorem (1763)

- Consider a model with:
    - data $y = (y_1, \ldots, y_N)$,
    - parameter vector $\theta = (\theta_1, \ldots, \theta_N)$, consisting of **random** elements.
- Two ways of writing the joint density of y and θ:

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

- Rewriting yields Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta)$$

where

$$\begin{aligned} p(\theta|y) &= \text{posterior density of } \theta \text{ given } y \\ p(y|\theta) &= \text{likelihood function} \\ p(\theta) &= \text{prior density of } \theta \\ p(y) &= \text{marginal density of } y. \end{aligned}$$

Principles of Bayesian econometrics (1)

- **Basic idea:** prior and posterior density are **subjective** evaluations of possible states of nature and/or outcomes of some process (or action).
- Famous quote from De Finetti:

  PROBABILITIES DO NOT EXIST

- That is, probabilities are not physical quantities that one can measure in practice, but they are **states of the mind**.
- Different priors $p(\theta)$ lead to different posterior densities $p(\theta|y)$...
- One may have a fundamental problem with this:

  'I don't like the idea of subjective outcomes!'

- But: Choice of size level (5%, 10%, ?%), model, data sample may also be subjective.

Principles of Bayesian econometrics (2)

Interpretation of Bayes' rule:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- One starts with the prior density $p(\theta)$; this contains intuitive, theoretical or other ideas on θ through.
- Then one learns from (new) data through the likelihood function $p(y|\theta)$. This yields the posterior $p(\theta|y)$.
- Briefly stated, Bayes' paradigm is a **learning principle**.
- Note that we can apply this rule sequentially.

Principles of Bayesian econometrics (3)

- Using the symbol $\propto$ ('is proportional to'), one can write Bayes' theorem

  $$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

  as

  $$p(\theta|y) \propto p(y|\theta)p(\theta) \qquad p(y|\theta)p(\theta) \text{ is the 'kernel' of } p(\theta|y)$$

  because $p(y) = \int p(y|\theta)p(\theta)d\theta$ only serves as a scaling constant, i.e. it does not depend on θ. Just like integrating constant $\sqrt{2\pi}$ in normal density.

- In words:

posterior density $\propto$	likelihood function	$\times$ prior density
beliefs after data $\Leftarrow$	influence of data	& beliefs before data

General note on conjugacy

- Conjugacy: Case in which posterior density has the same shape as the prior density. E.g. Normal prior applied to normal DGP results in normal posterior.
- **Advantage:** Conjugacy simplifies Bayesian analysis.
- **Disadvantage:** Conjugacy priors often more driven by convenience than by realism.

DGP(likelihood)	**Prior**	**Posterior**
normal	normal	normal
binomial	binomial	binomial
Poisson	Gamma	Gamma

Simple Example, classical inference

- Let the data generating process be

$$y_t = \mu + \epsilon_t \quad \epsilon_t \sim N(0,1)$$

- Look for the estimator that minimizes the Residual Sum of the Squares

$$\mu_{ols} = \arg\min \sum_{t=1}^{T} \epsilon^2 = \arg\min \sum_{t=1}^{T} (y_t - \mu)^2$$

Simple Example, classical inference

- Look for the estimator that maximizes the probability (likelihood) of having observed the sample $y_1, ..., y_T$

$$\mu_{ml} = \arg\max\ (2\pi)^{-T/2} \exp\left\{-1/2 \sum_{t=1}^{T} (y_t - \mu)^2\right\}$$

- Classical estimators

$$\mu_{ml} = \mu_{ols} = \overline{y}_T = 1/T \sum y_t$$

are consistent and unbiased.

- Since y_t is a gaussian white noise, CLT (it hold also under stationarity and ergodicity)

$$\left(\lim_{T\to\infty}\right) \sqrt{T}(\mu_{ols} - \mu) \sim N(0,1)$$

Simple Example, Bayesian inference

- Let the data generating process be (Zellner, 1971)

$$y_t = \mu + \epsilon_t \quad \epsilon_t \sim N(0,1)$$

- Assume a normal prior for μ

$$p(\mu) = (2\pi\sigma^2)^{-1/2} \exp\left\{ -1/2 \left(\frac{\mu - m}{\sigma} \right)^2 \right\}$$

- The posterior distribution is given by

$$p(\mu | y^T) = \frac{p(\mu) p(y^T | \mu)}{p(y^T)}$$

- After some algebra (▸ link), you get

$$p(\mu | y^T) = (2\pi\sigma^2)^{-1/2} \exp\left\{ -1/2 \left(\frac{\mu - \widehat{\mu}}{\sigma_\mu} \right)^2 \right\}$$

  where

$$\sigma_\mu^2 = \frac{\sigma^2}{T\sigma^2 + 1} \qquad \widehat{\mu} = \frac{T\sigma^2}{T\sigma^2 + 1}\overline{y} + \frac{1}{T\sigma^2 + 1} m$$

Simple Example, Bayesian inference

- The posterior is normal

$$p(\mu|y^T) \sim N(\widehat{\mu}, \sigma_\mu^2)$$

  where

$$\sigma_\mu^2 = \frac{\sigma^2}{T\sigma^2 + 1} \qquad \widehat{\mu} = \frac{T\sigma^2}{T\sigma^2 + 1}\mu_{ols} + \frac{1}{T\sigma^2 + 1}m$$

- See how prior information enters elegantly into the analysis
  Tight priors: when $\sigma \to 0$, $\widehat{\mu} \to m$
  Loose priors: when $\sigma \to \infty$, $\widehat{\mu} \to \mu_{ols}$
  As T gets large the sample information prevails over prior assumptions: when $T \to \infty$, $\widehat{\mu} \to \overline{y}$
- Asymptotically

$$\sqrt{T}(\mu - \widehat{\mu})|y^T \sim N(0, T\sigma_\mu^2) \to N(0,1)$$

Linear regression model under the natural conjugate prior for β given σ^2

- Consider again the linear regression model with k exogenous variables:

$$y_t = \theta x_t + \epsilon_t, \qquad \epsilon_t \sim NID(0, \sigma^2)$$

- Splitting up the prior into two parts:

$$p(\sigma^2) \propto 1/\sigma^2, \qquad p(\beta|\sigma^2) = N(b, \sigma^2 B)(conjugateprior)$$

$$\begin{aligned} p(\beta, \sigma^2) &= p(\beta|\sigma^2)p(\sigma^2) \propto N(b, \sigma^2 B) \times (1/\sigma^2) \\ &\propto (\sigma^2)^{-\dfrac{k+2}{2}} |B|^{-\dfrac{1}{2}} \exp\left(\frac{1}{2\sigma^2}(\beta - b)' B^{-1}(\beta - b)\right) \end{aligned}$$

Posterior density

$$
\begin{aligned}
p(\beta, \sigma^2 | y) &= p(\beta|\sigma^2)p(\sigma^2)p(y|\beta,\sigma^2) \\
&\propto \left(\frac{1}{\sigma^2}^{\dfrac{T}{2}} \right) \exp\left(-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2} \right) \\
&\times (\sigma^2)^{-\dfrac{k+2}{2}} |B|^{-\dfrac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - b)' B^{-1} (\beta - b) \right)
\end{aligned}
$$

Marginal posteriors

- After some tedious algebra, posterior density is also a normal density (conjugacy):

  $$\beta|y \sim N(\bar{\beta}, \bar{\sigma}^2(X'X + B^{-1})^{-1})$$

  where

  $$\bar{\beta} = (X'X + B^{-1})^{-1}(X'y + B^{-1}b)$$
  $$\bar{\sigma}^2 = \frac{1}{T}\Big((y - X\beta)'(y - X\beta) + (b - \bar{B})'B^{-1}(b - \bar{B})\Big)$$

- if prior variance B is small (large) the influence of the prior on the posterior mean and posterior variance is high (low).
- A conjugate prior with a very large variance is uninformative.
- If the number of observations T is large, the influence of the prior becomes less as $X'X = \sum_{i=1}^{T} x_i'x_i$ and $X'y = \sum_{i=1}^{T} x_i'y_i$ become large.

Monte Carlo integration

- For (more) complicated models (than the previous linear ones), it's usually impossible to find analytical solution.
- Integration is not feasible.
- Monte Carlo integration.

Monte Carlo integration: Motivation

- The importance of Monte Carlo integration in Bayesian inference
- Several simulation methods useful for Monte Carlo integration:
    - Direct sampling: inversion method.
    - Indirect sampling: rejection sampling, importance sampling.
    - Markov Chain Monte Carlo: Metropolis-Hastings algorithm, Gibbs sampling, data augmentation
- Due to the use of Monte Carlo integration, Bayesian inference is also called simulation based inference.

Monte Carlo integration: Motivation (continued)

Bayes' theorem: $p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)}$

where:

- $p(\theta|y)$: posterior density of θ given y
- $p(y|\theta)$: likelihood function
- $p(\theta)$: prior density of θ
- $p(y)$: predictive density of y.

Note: integral in denominator (predictive density)
$\Rightarrow$ integration needed to obtain exact **posterior density**.
Posterior mean of θ (expectation of θ given y):

$$E[\theta|y] = \int \theta p(\theta|y) d\theta$$

Note: again an integral has to be evaluated.

Monte Carlo integration: Motivation (continued)

- So, one needs integration in order to know:
  - (exact) posterior density
  - posterior mean, variance, etc. of θ
  - posterior odds ratio (for model comparison)
- Note: In linear models these integrals can be computed analytically.
- For more complicated models, it is usually impossible to find analytical solutions.
- **In general, we need numerical integration methods for Bayesian inference.**
- Two options:
  - deterministic integration
  - Monte Carlo (MC) integration

Monte Carlo integration: Motivation (continued)

- Suppose we want to evaluate: $E\left[g(\theta)\right] = \int g(\theta)p(\theta|y)d\theta$
- **Deterministic integration**
  1. Evaluate integrand $f(\theta) = g(\theta)p(\theta|y)$ in many fixed points $\theta^1, \ldots, \theta^n$.
  2. Use weighted sum of evaluations as approximation to integral:

$$\int f(\theta)d\theta \approx \sum_{i=1}^{n} w_i f(\theta^i)$$

  where $w_1, \ldots, w_n$ are weights.

- **Examples of deterministic integration methods**
  - trapezoid rule;
  - Simpson's rule;
  - Gaussian integration rules.

Monte Carlo integration: Motivation (continued)

- m-dimensional deterministic integration with n evaluation points in each direction:

$$\int \dots \int f(\theta_1, \dots, \theta_m) d\theta_1 \dots d\theta_m$$

$$\approx \sum_{k=1}^{n} w_{mk} \dots \left[\sum_{j=1}^{n} w_{2j} \left[\sum_{i=1}^{n} w_{1i} f(\theta_{1i}, \theta_{2j}, \dots, \theta_{mk}) \right] \right]$$

  Note: n^m function evaluations needed.

- **'Curse of dimensionality'**: The number of function evaluations n^m increases exponentially with m, the dimension of θ.
- Deterministic integration infeasible for high-dimensional integration problems!Example: 50 evaluation points for each dimension of a 6-dim. integral:

$$50^6 = 15625000000 > 15 \text{ billion}$$

  function evaluations.

Monte Carlo integration: Motivation (continued)

- Deterministic integration infeasible for high-dimensional integration problems
- For Bayesian analysis of θ with ‘high’ dimension, a different integration method is needed that does not suffer from the ‘curse of dimensionality’

Solution: Monte Carlo integration!

Monte Carlo integration: Motivation (continued)

- Basic difference becomes clear from:

$$E\left[\theta|y\right] = \int \theta p(\theta|y) d\theta$$

- Deterministic integration evaluates the right-hand integral explicitly
- Monte Carlo integration focuses on the left-hand posterior expectation:
  - draw θ's from the posterior density $p(\theta|y)$
  - use this sample of θ's to investigate characteristics of the posterior distribution.

Monte Carlo integration: Motivation (continued)

1) Collect a sample $\theta^1, \ldots, \theta^n$ from posterior distribution with density $p(\theta|y)$.
2) Estimate expectation $\underline{E\left[g(\theta)|y\right]}$ by corresponding sample mean:

$$E\left[g(\theta)|y\right] \approx \frac{1}{n}\sum_{i=1}^{n} g(\theta^i)$$

Examples:

- estimate posterior mean of θ by sample mean of θ's;
- estimate posterior $Prob\left[\theta \in D\right]$ by fraction of θ's in D, etc.

Advantage: CLT: convergence at rate $1/\sqrt{n}$, independent from m

$\Rightarrow$ No 'curse of dimensionality' in Monte Carlo integration!!

Monte Carlo integration: Sampling methods

In Monte Carlo integration methods one typically needs to draw from the posterior distribution. We group the sampling methods according to **two characteristics**:

- **Direct/indirect sampling**:
  Do we directly draw from the posterior, or do we need a correction mechanism (e.g. acceptance-rejection step, weighting of draws)?
- **Independence/dependence sampling**:
  Are the obtained draws independent from each other?

Direct Sampling

In the ideal situation, we can directly draw from the posterior distribution (without requiring an acceptance-step or weighting the draws).

Some direct sampling methods:

1. uniform sampling;
2. inversion method.

[1] Uniform sampling

Any sampling algorithm is based on collecting draws from the uniform $U(0,1)$ distribution!!

Independence Sampling vs. Dependence sampling

We also group the sampling methods according to the following question:

Are the obtained draws independent from each other?

Answer Yes	⇒ Independence sampling (Law of Large Numbers & Central Limit Theorem apply)
Answer No	⇒ Dependence sampling: Markov chain Monte Carlo (Metropolis-Hastings algorithm, Gibbs sampling. These methods rely on Markov chain theory).

Indirect Independence Sampling

- Recall: In ideal situation, we can directly draw from posterior distribution. However: Direct sampling is mostly very difficult or very slow. Solution: Indirect Sampling methods.
- Principle of indirect sampling methods:
  (1) Draw points from a certain 'candidate' distribution.
  (2) Use 'correction mechanism' to get characteristics of posterior distribution of interest (the 'target' distribution).
- 'candidate' distribution: easy to sample from, and (hopefully) a reasonable approximation to the 'target' distribution.
- Possible 'correction mechanisms':
  - acceptance-rejection step
  - weighting of draws.

Indirect Independence Sampling (continued)

Two indirect independence sampling methods:

I. **Rejection sampling** (alias 'the acceptance-rejection method'): draw points from 'candidate' distribution, and accept with a certain probability.

II. **Importance sampling**: draw points from 'candidate' distribution, and give all draws certain weights.

Rejection sampling (acceptance-rejection method)

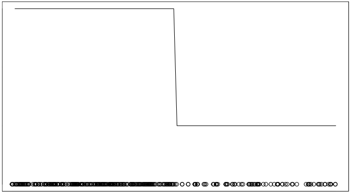
- An attempt to sample one x from target distribution with density $p(x)$:
  - I) Generate y from candidate distribution with density $q(y)$.
  - II) Accept y with probability $\frac{p(y)}{cq(y)}$

    (by generating u from $Y(0,1)$ and accepting if $u \leq \frac{p(y)}{cq(y)}$)
- Here: c is a constant such that $p(x) \leq cq(x), \forall x$ so we must have $0 \leq \frac{p(y)}{cq(y)} \leq 1$ (necessary condition for a probability) Note: the number of draws from the candidate necessary to obtain a certain number N of draws from the target density $p(x)$ is itself random.
- The higher the probability of an acceptance, the less ‘attempts’ are required to obtain N draws from the target density
- The faster the Monte Carlo integration algorithm works.

Rejection sampling example: good and bad candidates

Target density $p(x)$: $p(x) = \begin{cases} 3/2 & \text{if } 0 < x \le 1/2 \\ 1/2 & \text{if } 1/2 < x < 1 \\ 0 & \text{else} \end{cases}$

good candidate: $q(x) = \text{unif}(0,1)$



acceptance rate $= 0.65$

- Higher acceptance rate is important (indication of success)
- Still, very high acceptance rate is possible although the candidate is bad!
- In more complicated models this becomes a real issue.

Rejection sampling (continued)

The probability of an acceptance is:

$$\begin{aligned}
Pr\left[U \leq \frac{p(Y)}{cq(Y)}\right] &= \int Pr\left[U \leq \frac{p(y)}{cq(y)}\right] q(y)dy \\
&= \int \frac{p(y)}{cq(y)} q(y)dy \\
&= \frac{1}{c}\int p(y)dy = \frac{1}{c}
\end{aligned}$$

Rejection sampling (continued)

- The larger the probability of acceptance, the faster we get a sample of x's. Choose c as small as possible (such that restriction $p(x) \leq cq(x) \forall x$):

$$\text{optimal } c = \max_x \frac{p(x)}{q(x)}.$$

- Optimal c is small (and probability of acceptance is high) if variation in the ratio $p(x)/q(x)$ is small.
- Candidate distribution should be a 'good approximation' to target distribution.

Rejection sampling (continued)

Disadvantages of rejection sampling:

We accept y with probability $\frac{p(y)}{cq(y)}$ where $c = \max_x \frac{p(x)}{q(x)}$:

(1) maximization required: this may take a lot of time;

(2) $\max_x \frac{p(x)}{q(x)}$ may not even exist

$\Rightarrow$ in that case rejection sampling is impossible with this ‘candidate’ distribution.

(3) $c = \max_x \frac{p(x)}{q(x)}$ may be very small

$\Rightarrow$ many draws may be required from the ‘candidate’ to generate a draw from the target.

Importance Sampling

Importance sampling: Difference with rejection sampling:

- Rejection sampling: draws get either full weight (acceptance) or no weight at all (rejection).
- Importance sampling: draws get weights that can take any possible non-negative value, reflecting the relative importance of draws.

Advantages of Importance Sampling over Rejection Sampling:

- We do not need to find $x = \max_x \frac{p(x)}{q(x)}$.
- We do not throw away draws (information), but give them certain weights.

  $\Rightarrow$ In general, IS yields better estimates than Rejection Sampling.

Importance Sampling - Basic idea

- We want to evaluate the expectation $E[g(X)]$, with $g(\cdot)$ a function - e.g. $g(x) = x$ for the mean, where X is a random variable with (target) density *kernel* p.
- Importance Sampling is based on:

$$E[X] = \int g(x) \frac{p(x)}{\int p(x)dx} dx = \frac{\int g(x)p(x)dx}{\int p(x)dx} = \frac{\int g(x)[w(x)q(x)]dx}{\int [w(x)q(x)]dx}$$
$$= \frac{\int [g(x)w(x)]q(x)dx}{\int w(x)q(x)dx} = \frac{E[w(Y)g(Y)]}{E[w(Y)]}$$

where:

- Y is a random variable with (candidate) density q;
- $w(x) \equiv p(x)/q(x)$ is the weight function.

Note: If exact density p known, then we also have:
$E[X] = E[w(Y)g(Y)]$.

Importance Sampling (continued)

Simulation is possible form the *candidate density* q then the Importance Sampling:

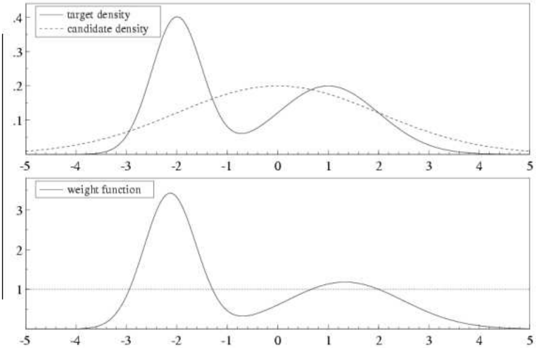
$$E[X] = \frac{\int [g(x)w(x)]q(x)dx}{\int w(x)q(x)dx} \approx \frac{\frac{1}{n}\sum_{i=1}^{n} w(y_i)g(y_i)}{\frac{1}{n}\sum_{i=1}^{n} w(y_i)} = \frac{\sum_{i=1}^{n} w(y_i)g(y_i)}{\sum_{i=1}^{n} w(y_i)}$$

where

- Y is a random variable with (candidate) density q,
- $y_1, \ldots, y_n$: realizations from candidate,
- $w(x) \equiv p(x)/q(x)$ is the weight function,
- $w(y_1), \ldots, w(y_n)$: corresponding weights.

Importance Sampling (continued)

Illustration of importance sampling (IS):



Note: Points for which candidate $<$ ($>$) target: sampled too rarely (often) $\Rightarrow$

relatively large (small) IS weights correct this 'under-sampling' ('over-sampling').

Importance Sampling (continued)

- Disadvantage of IS: IS cannot yield a sample of draws from the (target) distribution of X, but can only give an estimate of an expectation $E[g(X)]$ for a function g.
- Advantage of IS: Almost each property of interest can be written as an expectation $E[g(X)]$
- For example, recall: $Pr[\theta \in D] = E[I\{\theta \in D\}]$, where D is some region, $I = 0/1$ indicating whether $\theta \in D$.
- Using a suitable indicator function for the function g one can use IS to construct histograms of marginal densities.

Importance Sampling (continued)

- The performance of IS is greatly affected by the choice of the candidate:
  - $q(y)$ inappropriate
  - $w(y) = p(y)/q(y)$ varies much
  - only a few points y_i with extremely large weights determine the IS estimate of $E[g(X)]$:

$$g_{IS} = \frac{\sum_{i=1}^{n} w(y_i) g(y_i)}{\sum_{i=1}^{n} w(y_i)}$$

- Example: especially if $p >> q$ in the tails of the distribution

  $\Rightarrow$ few points with large IS weights at extreme locations

  $\Rightarrow$ disastrous effect on g_{IS}.

- **Conclusion:** if target might have fat tails, use candidate with fat tails (like Student's t with few degrees of freedom).

Dependent sampling: MCMC methods - Basic idea

- Obtain a sequence of draws from the desired target density $p(x)$ by *cleverly* constructing a Markov chain. Markov chain is a sequence of random variables for which the Markov property holds:

$$p(x_{t+1}|X_t, X_{t-1}, X_{t-2}, ...) = p(x_{t+1}|X_t)$$

- The key problem is finding a transition probability function $p(x, y)$ such that the Markov chain converges to the desired target probability function $p(x)$, which will be the limiting "invariant" probability of the Markov chain:

$$\sum_{x \in S} p(x)p(x, y) = p(y)$$

- Metropolis et al. (1953) and Hastings (1970) found a solution to this problem by constructing a **time-reversible** Markov chain.

MCMC methods

- Time-reversibility: if a MC has the same transition probabilities as its reversal.
- Once the MC has reached the target distribution, it will never leave it, because the probabilities will stay constant.
- The target distribution is a limiting distribution of the MC.
- Under the conditions of the irreducibility (all states are accessible from each other) and aperiodicity (the number of transitions necessary to return to a state is not necessary a multiple of some integer), the limiting distribution is unique and equal to the target distribution.

MH algorithm

- Simulating X, random variable with (target) density *kernel* $p(\cdot)$. Simulation is possible form the *candidate density* q.

Initialization: Choose feasible X_0 in S. Do for $t = 1, 2, \ldots, n$

  - Obtain y from candidate transition probability function $q(X_{t-1}, y)$,
  - Accept y (i.e. $X_t = y$) with probability α :

$$\alpha(X_{t-1}, y) = \min\left\{\frac{p(y)q(y, X_{t-1})}{p(X_{t-1})q(X_{t-1}, y)}, 1\right\}$$

  - If y is rejected, set $X_t = X_{t-1}$.

Note:

- The acceptance/rejection step is done by drawing $u \sim U(0, 1)$
- So far this is the first method we get correlated draws

MH algorithm example

Simple example: Suppose we want a Metropolis-Hastings Markov chain $X_t|t = 1, 2, \ldots$ of which the distribution converges to the target distribution:

$$Pr[X_t = 0] = 1/3 = \pi(0) \quad Pr[X_t = 1] = 2/3 = \pi(1)$$

where we choose the candidate transition probability function $q(x, y)$ as:

$$\begin{array}{ll} q(0,0) \equiv Pr[y = 0|X_{t-1} = 0] = 1/2 & q(1,0) \equiv Pr[y = 0|X_{t-1} = 1] = 1/2 \\ q(0,1) \equiv Pr[y = 1|X_{t-1} = 0] = 1/2 & q(1,1) \equiv Pr[y = 1|X_{t-1} = 1] = 1/2 \end{array}$$

Interpretation:

$$\text{flip a coin:} \quad \begin{array}{lcl} \text{head} & \Rightarrow & y = 0 \\ \text{tail} & \Rightarrow & y = 1 \end{array}$$

MH method

Usually and most basically, $q(x, y)$ is specified in one of the following ways:

- In an **independence chain** the candidate state y is drawn independently from the current state x:

$$q(y, x) = q(y)$$

- In a **random walk chain** the candidate transition step is chosen instead of the candidate state y

$$q(y, x) = q(y - x)$$

It is often chosen as $y_{t+1} = y_t + \epsilon_{t+1} \quad \epsilon_{t+1} \sim N(0, \sigma_\epsilon^2)$

Random Walk Metropolis-Hastings Algorithm(1)

Lets make the MH Algorithm operational

- Let give a functional form for $q(\theta^*|\theta^{(s-1)}, V)$, i.e. a normal so that $\theta^* \sim N(\theta^{(s-1)}, V)$. Then, $q(\theta^*|\theta^{(s-1)}, V) = q(\theta^{(s-1)}|\theta^*, V)$
- The binomial RV simplifies to
  - $\theta^{(s)} = \theta^*$ with probability

$$\alpha(\theta^*|\theta^{(s-1)}, V) = \min\left\{\frac{p(\theta^*|y^T)}{p(\theta^{(s-1)}|y^T)}, 1\right\}$$

  - $\theta^{(s)} = \theta^{(s-1)}$ otherwise

Random Walk Metropolis-Hastings Algorithm(2)

- Interpretation:
  - If we are moving uphill, i.e. $p(\theta^*|y^T) > p(\theta^{(s-1)}|y^T)$ we always keep the draw. If we are moving downhill we keep the draw with probability $\widetilde{\alpha} = \frac{p(\theta^*|y^T)}{p(\theta^{(s-1)}|y^T)}$
  - You explore the entire parameter space and not only the high probability regions
  - Two good reasons: (1) you do not want to get stuck in a local maximum, (2) you want to characterize also the tails of your distribution.
- How do you draw from a binomial distribution X with probability $\widetilde{\alpha}$? you draw r from a uniform [0,1] and if $r < \widetilde{\alpha}$ you accept, otherwise reject.

Simple Example, MCMC-MH

- Let the data generating process be student t distributed

$$y_t = \mu + \epsilon_t \quad \epsilon_t \sim t(0,1,\nu)$$

$$p(\epsilon_t|\nu) = (\nu\pi)^{-\frac{1}{2}} \frac{\Gamma((1+\nu)/2)}{\Gamma(\nu/2)} \left[1 + \frac{1}{\nu}\epsilon_t^2\right]^{-\frac{\nu+1}{2}}$$

- Student-t looks like normals with ticker tails, i.e. they assign larger probabilities to extreme events. Stock prices returns with occasionally high or low returns. It is a symmetric distribution.
- When $\nu \to \infty$, the student-t converges to the gaussian distribution. $\nu > 2$, else the variance is infinity.
- Student-t can be expressed as the product of a gamma distribution times a normal. You are introducing a layer of heteroscedasticity.

Simulated data, $T = 250$

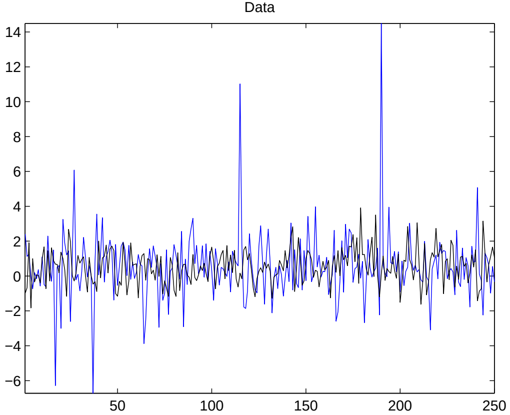


Figure: Simulated data with $\nu = 3$ and $\mu = 0.5$. In blue the student-t errors in black normal errors.

Simple Example, MCMC-MH

- Given the observed data you wish to compute analytically the posterior.
- Assume a normal prior for μ

$$p(\mu|m,\sigma) = (2\pi\sigma^2)^{-1/2} \exp\left\{-1/2\,(1/\sigma(\mu-m))^2\right\}$$

- Assume a χ-squared prior for ν

$$p(\nu|q) = \frac{2^{-q/2}}{\Gamma(q/2)} \nu^{q/2-1} e^{-\nu/2}$$

- The joint posterior distribution of ν and μ is given by

$$p(\mu,\nu|y^T) \propto (\nu\pi)^{-\frac{1}{2}} \frac{\Gamma((1+\nu)/2)}{\Gamma(\nu/2)} \left[1+\frac{1}{\nu}(y_t-\mu)^2\right]^{-\frac{\nu+1}{2}} \times p(\mu|m,\sigma) \times p(\nu|q)$$

and looks pretty ugly ...

Simple Example, MCMC-MH

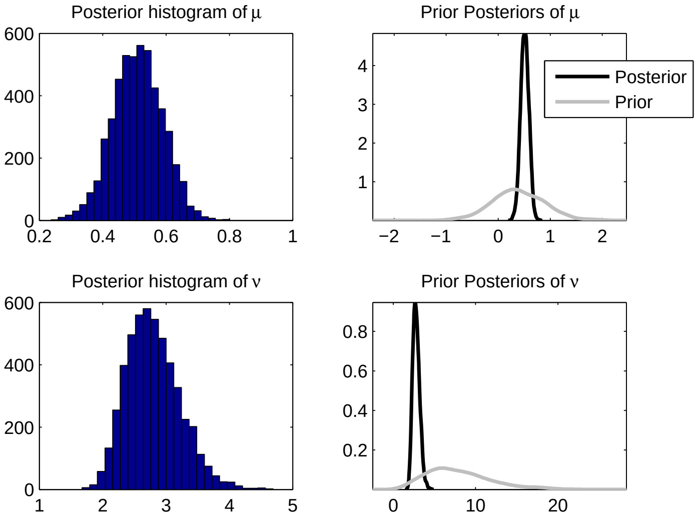
Assume assume priors values of $m = 0.4$ $\sigma = 1$ and $q = 8$. Given the observed data, we wish to characterize the posterior distribution of μ and ν. MCMC !

1. Start from $\mu^{(0)} = 1/T \sum y_t$ and $\nu_0 = q$ and $V = 0.2\ I$. Let $\theta = (\mu, \nu)$
2. Generate a candidate draw $\theta^* \sim N(\theta^{(s-1)}, V)$
3. Compute

$$p(\mu^*, \nu^*|y^T) = \left(\prod_{t=1}^{T} (\nu^* \pi)^{-\frac{1}{2}} \frac{\Gamma((1+\nu^*)/2)}{\Gamma(\nu^*/2)} \left[1 + \frac{1}{\nu^*}(y_t - \mu^*)^2 \right]^{-\frac{\nu^*+1}{2}} \right)$$
$$\times (2\pi\sigma^2)^{-1/2} \exp\left\{ -1/2\,(1/\sigma(\mu^* - m))^2 \right\} \times \frac{2^{-q/2}}{\Gamma(q/2)} (\nu^*)^{q/2-1} e^{-\nu^*/2}$$

4. Draw r form a uniform 0-1 and keep the draw if $r < p(\mu^*|y^T)/p(\mu^{(s-1)}|y^T)$
5. repeat 2-4 many times.
6. Two practical suggestions:
   - Tune the size of the jump V to target an acceptance rate of 30-40%. The smaller (larger) the step the more (less) likely you are accepting the draws.
   - MC induce correlated draws. Typically you discard a burn-in part and consider a (random) subset of accepted draws.

Posterior Simulators, Marginal distribution

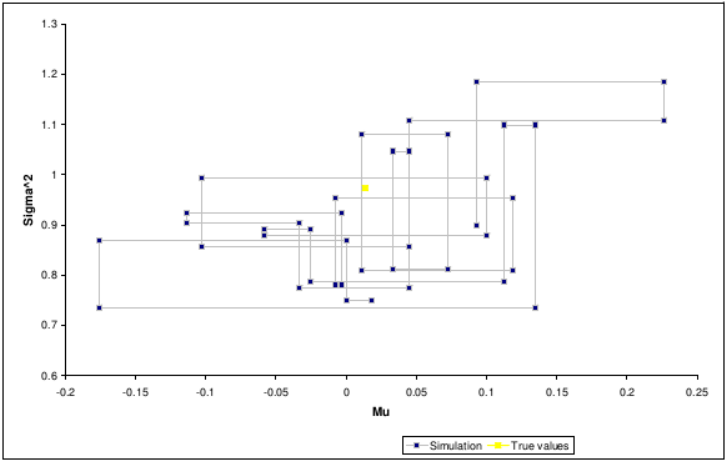


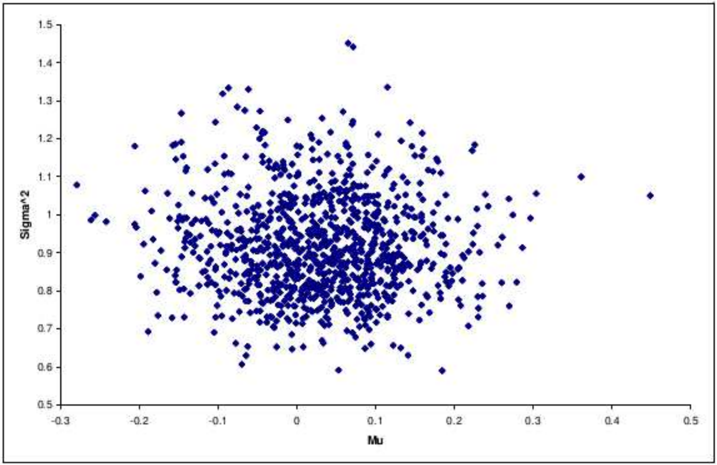
Gibbs sampling - Basic idea

- In order to simulate multivariate random variable X Gibbs sampling is 'divide and conquer':
  - Decompose X in k blocks $(X_1, X_2, \ldots, X_k)$.
  - Construct Markov chain by iteratively drawing each component of X conditional on the values of all other components.
- **Gibbs sampling procedure** works as follows:
  - Initialization: Choose feasible $x^0 = (x^0, ..., x^k)$ $[\theta^0 = (\widehat{\beta}, \widehat{\sigma^2}]$
  - Do for j=1,...,n (number of MCMC simulations)
    - Do for i=1,...,k

      Draw x_i^j from $p(x_i | x_{-i}^{j-1})$ where x_{-i}^{j-1} is the set of "most recent" values of all other components.

      [Draw $p(\beta_j | \sigma_{j-1}^2, y)$ and then $p(\sigma_j^2 | \beta_j, y)$].


1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
Sigma^2
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
Mu
Simulation
True values


Sigma^2
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
Mu

Example: bivariate normal distribution

- Given X_1 and X_2 with bivariate normal distribution: $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho = 0.98$ $\quad(\star)$.
- The conditional distributions corresponding to $(\star)$:

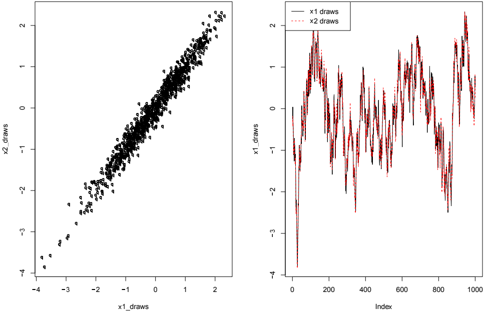
$$\begin{aligned} X_1|X_2 = x_2 &\sim N(\rho x_2, 1-\rho^2) \\ X_2|X_1 = x_1 &\sim N(\rho x_1, 1-\rho^2) \end{aligned}$$

  with $\rho = 0.98$.
- We can obtain draws from the distribution in $(\star)$ by Gibbs sampling:
  - Initialization: Choose e.g. $(X_1, X_2) = (0, 0)$.
  - Do for $t = 1, 2, \ldots, n$:
    - Draw $X_2^t \sim N(\rho X_1^t, 1-\rho^2)$
    - Draw $X_1^t \sim N(\rho X_2^t, 1-\rho^2)$

Gibbs sampling example (continued)

Gibbs draws:

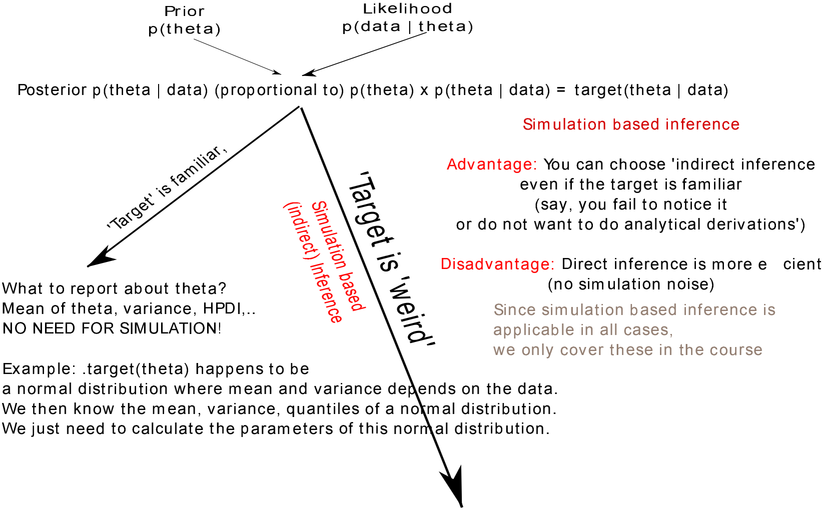
$$X_1^t \sim N(0.98X_2^{t-1}, \sigma = 1 - \rho^2) \;\; X_2^t \sim N(0.98X_1^{t-1}, \sigma = 1 - \rho^2)$$



Gibbs draws from the bivariate normal distribution
number of simulations: 1000
estimated mean:
(x_1) -0.066, (x_2) -0.068

Note:

- For $n = 1000$ draws, estimated means are still far from true values (0) (for smaller n the draws do not cover the whole domain)
- Notice high correlation between parameter draws!



Prior
p(theta)

Likelihood
p(data | theta)

Posterior p(theta | data) (proportional to) p(theta) x p(theta | data) = target(theta | data)

'Target' is familiar,

What to report about theta?
Mean of theta, variance, HPDI,..
NO NEED FOR SIMULATION!

Example: .target(theta) happens to be
a normal distribution where mean and variance depends on the data.
We then know the mean, variance, quantiles of a normal distribution.
We just need to calculate the parameters of this normal distribution.

Simulation based
(indirect) Inference

'Target is 'weird'

Simulation based inference

Advantage: You can choose 'indirect inference
even if the target is familiar
(say, you fail to notice it
or do not want to do analytical derivations')

Disadvantage: Direct inference is more efficient
(no simulation noise)

Since simulation based inference is
applicable in all cases,
we only cover these in the course

1. Use SAMPLING METHODS to get 'draws from' the posterior
(e.g. the target of the simulation method is the posterior)

2. Given these draws of theta, report findings:
Estimate mean, variance, HPDI, quantiles of theta from the draws.
(MONTE CARLO INTEGRATION)

Note: Importance sampling does points (1) and (2) at once for you
Just need to define a candidate!

Summary of sampling algorithms

	Accept/Reject	MH	IS	Gibbs
Applicable for scalar random variabe?	yes	yes	yes	no
Indirect sampling	yes	yes	yes	yes
Correlated draws	yes	yes	no	yes
Requires joint target/posterior?	yes	yes	yes	no
Can be combined with Gibbs sampler	yes	yes	no	–

Why use sampling methods? AR(1) example

Consider the AR(1) model for US GDP growth (g_t):

$$g_t = \mu + \rho g_{t-1} + \epsilon_t, \;\; \epsilon_t \sim NID(0, \sigma^2)$$

- Identify what are the model parameters: $\theta = \{\mu, \rho, \sigma^2\}$
- Write down the likelihood ($p(g_{1:T}|\theta)$) using the Bayes Rule:

$$\begin{aligned} L(g_{1:T}|\mu,\rho,\sigma^2) &= pr(g_T|g_{1:T-1},\mu,\rho,\sigma^2) \times pr(g_{1:T-1}|\mu,\rho,\sigma^2) \\ &\vdots \\ &= \prod_{t=1}^{T} pr(g_t|g_{1:t-1},\mu,\rho,\sigma^2) \times pr(g_0|\mu,\rho,\sigma^2) \\ &= \prod_{t=1}^{T} pr(g_t|g_{1:t-1},\mu,\rho,\sigma^2) \end{aligned}$$

where the last line follows since we assume that the initial observation g_0 is given (probability 1).

Why use sampling methods? AR(1) example

The model states that

$$g_t|g_{1:t-1},\mu,\rho,\sigma^2 = \mu + \rho g_{t-1} + \epsilon_t \sim NID(\mu + \rho g_{t-1}, \sigma^2).$$

i.e.

$$pr(g_t|g_{1:t-1},\mu,\rho,\sigma^2) = \phi(g_t;\mu + \rho g_{t-1},\sigma^2).$$

$\phi(g_t; a, b)$ is the density of the normal distribution with mean a and variance b at point g_t.

$$p(\text{data}|\theta) = L(g_{1:T}|\mu,\rho,\sigma^2) \quad = \quad \prod_{t=1}^{T} \phi(g_t;\mu + \rho g_{t-1},\sigma^2)$$

and $\phi(g_t; , a, b)$ is the density of the normal distribution with mean a and variance b at point g_t.

Why use sampling methods? AR(1) example

Recipe for a Bayesian: Likelihood and priors

- 'Coming up with priors'

$$p(\mu, \rho) \propto \begin{cases} 1 & \text{if } \rho \in (-1,1)\text{(I assume stationarity)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$p(\sigma^2) \propto \begin{cases} 1/\sigma^2 & \text{if } \sigma^2 > 0 \\ 0 & \textit{otherwise} \end{cases} \quad (2)$$

this is a 'flat, uninformative' prior for the variance.

Why use sampling methods? AR(1) example

- Assume independent priors for μ, ρ and σ^2. The prior for all model parameters is then:

$$p(\theta) = p(\mu, \rho, \sigma^2) \propto \begin{cases} 1/\sigma^2 & \text{if } \rho \in (-1, 1) \text{ and } \sigma^2 > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Note: The above is NOT a probability density (they do not integrate to any finite number)
- Such a prior is called an improper prior.
- Even if the priors are improper, as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them.

Why use sampling methods? AR(1) example

- Recipe for a Bayesian: Likelihood and priors
- Calculating the posterior density $\propto$ posterior density kernel $\equiv$ target density (for the sampling algorithm)
- Recall Bayes' rule:

$$
\begin{aligned}
p(\theta|\text{data}) &\propto p(\theta) \times p(\text{data}|\theta) \\
&= 1/\sigma^2 I[\rho \in (-1,1) \text{ and } \sigma^2 > 0] \\
&\times \prod_{t=1}^{T} \phi(g_t; \mu + \rho g_{t-1}, \sigma^2)
\end{aligned}
$$

and $I[\cdot]$ is the indicator function which takes the value of 1 if its argument holds, and the value of 0 otherwise.

Why use sampling methods? AR(1) example

- The posterior kernel has a very weird shape (even in this simple model)

$$
\begin{aligned}
p(\theta|\text{data}) &= p(\mu, \rho, \sigma^2|g_{1:T}) \\
&\propto 1/\sigma^2 I[\rho \in (-1,1) \text{ and } \sigma^2 > 0] \\
&\times \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(g_t - \mu - \rho g_{t-1})^2}{\sigma^2}\right)
\end{aligned}
$$

  and $I[\cdot]$ is the indicator function which takes the value of 1 if its argument holds, and the value of 0 otherwise.

- It is hard to see if the right hand side is a known MULTIVARIATE density function for the three parameters μ, ρ, σ^2!
- Solution: Instead of simplifying this function, and trying to find if this is a known density function, we will simulate parameters $\theta = \{\mu, \rho, \sigma^2\}$ and use Monte Carlo integration to find their mean, variance, quantiles etc.

Which sampling method? Acceptance-Rejection for model parameters

- We need to simulate (μ, ρ, σ^2) together from the target density (posterior kernel):

$$\begin{aligned} p(\theta|\text{data}) &= p(\mu, \rho, \sigma^2 | g_{1:T}) \\ &\propto 1/\sigma^2 I[\rho \in (-1,1) \text{ and } \sigma^2 > 0] \\ &\times \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(g_t - \mu - \rho g_{t-1})^2}{\sigma^2}\right) \end{aligned}$$

  and $I[\cdot]$ is the indicator function which takes the value of 1 if its argument holds, and the value of 0 otherwise.
    - Recall: Need a 'candidate distribution' and a constant 'c' for this target to be above the candidate.
    - Recall: Need to find an 'ok' candidate such that I do not end up with '0' acceptance rate.
- Finding the constant 'c' is really difficult for multiple parameters

Which sampling method? Importance Sampling

- Recall: We need to define a candidate distribution that covers a wide range of parameters.
- It is hard to 'guess' where this distribution should be.
- Simple choice: Independent candidate distributions for each parameter:

$$q(\theta) = q(\mu) \times q(\rho) \times q(\sigma^2) = \underbrace{T(\mu; 3)}_{q(\mu)} \underbrace{U(\rho; -1, 1)}_{q(\rho)} \underbrace{\chi^2(\sigma^2; 3)}_{q(\rho;\sigma^2)}$$

There are also methods to find a 'good multivariate candidate'

Which sampling method? Importance Sampling

- The target density is given by:

$$
\begin{aligned}
p(\theta|\text{data}) &= p(\mu, \rho, \sigma^2 | g_{1:T}) \\
&\propto 1/\sigma^2 I[\rho \in (-1,1) \text{ and } \sigma^2 > 0] \\
&\times \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(g_t - \mu - \rho g_{t-1})^2}{\sigma^2}\right)
\end{aligned}
$$

- The candidate density is give by

$$q(\theta) = q(\mu) \times q(\rho) \times q(\sigma^2) = \underbrace{T(\mu;3)}_{q(\mu)}\underbrace{U(\rho;-1,1)}_{q(\rho)}\underbrace{\chi^2(\sigma^2;3)}_{q(\rho;\sigma^2)}$$

- The IS estimation of the parameters is given by:

$$E[\theta] = \frac{E[w(\theta_q)g(\theta_q)]}{E[w(\theta_q)]}$$

where:

- θ_q is a random variable simulated from the (candidate) density $q(\theta)$;
- $w(\theta_q) \equiv p(\theta_q|\text{data})/q(\theta_q)$ is the weight function.

Which sampling method? Importance Sampling

Pseudo-code

- Simulate M draws $\theta_q^{(1)}, \ldots, \theta_q^{(M)}$ from the 'candidate' $q(\theta)$ (these are independent draws $\Rightarrow$ Independence sampling method)
- Calculate $q(\theta_q^{(1)}), \ldots, q(\theta_q^{(M)})$, density of the 'candidate' for M draws.
- Calculate the target density (posterior kernel) $p(\theta_q^{(1)}|g_{1:T}), \ldots, p(\theta_q^{(M)}|g_{1:T})$ for M draws.
- Calculate weights $w(\theta_q^{(m)}) \equiv p(\theta_q^{(m)}|\text{data})/q(\theta_q^{(m)})$ for each draw $m = 1, \ldots, M$.

  Numerical stability: Better to calculate log-weights (similar to log-likelihood maximization):

$$\ln(w(\theta_q^{(m)})) = \ln(p(\theta_q^{(m)}|\text{data})) - \ln(\theta_q^{(m)})$$

Which sampling method? Importance Sampling

Pseudo-code (continued)

- Calculate e.g. mean values of θ ($g(\theta) = \theta$) using Monte Carlo integration

$$E[\theta] = \frac{E[w(\theta_q)\theta_q]}{E[w(\theta_q)]} \approx \frac{\frac{1}{M}\sum_{m=1}^{M} w(\theta_q^{(m)})\theta_q^{(m)}}{\frac{1}{M}\sum_{m=1}^{M} w(\theta_q^{(m)})}$$

- Calculate e.g. variance values of θ ($g(\theta) = \theta^2$) using Monte Carlo integration

$$E[\theta^2] = \frac{E[w(\theta_q)\theta_q^2]}{E[w(\theta_q)]} \approx \frac{\frac{1}{M}\sum_{m=1}^{M} w(\theta_q^{(m)})\left(\theta_q^{(m)}\right)^2}{\frac{1}{M}\sum_{m=1}^{M} w(\theta_q^{(m)})}$$

$$var(\theta) = E[\theta^2] - (E[\theta])^2$$

Which sampling method? Importance Sampling

- Effect of M (Central Limit Theorem):
  The higher M, the more precise parameter estimates
- Effect of a 'bad candidate':
  Importance weights w will be too close to 0
  (Effective number of observations is very small)
- Independence sampling (IS)
  We did not have to run a loop for m draws,
  all calculations can be made as 'vector operations'.
- Independence sampling (IS)
  No need to 'burn-in' or 'trim' draws since they are independent
  and we did not 'initialize' the algorithm.

Note: We did not use CLT for the data, goodness of the simulation is independent of the sample size T.

Which sampling method? Importance Sampling

Effect of observation sample size
(common for all sampling methods you can use):

- If the observation sample is not informative (e.g. small T), the likelihood (hence the posterior) will not be representative of the data generating process.
- This problem is not linked to the 'goodness' of Bayesian inference.
  There is simply not much information in the data
  $\Rightarrow$ Theoretically this will just lead to high variances of θ, just in the case of classical estimation.
- The problem of a small will be there whatever simulation method you use.
- Notice the estimation of mean, variance etc of θ
  We did not use any asymptotic T results for this purpose
  $\Rightarrow$ No need to derive small sample properties in case of a small sample.
  $\Rightarrow$ Also when testing parameter restrictions, no need to derive the asymptotic properties or small sample properties of tests.

The advantage of going through these simulation based inference is that we now do not need to derive test properties.

Bayesian estimation of State Space Models

- Take a simple state space model

$$
\begin{aligned}
\mathbf{y}_t &= \mathbf{Z}_t\boldsymbol{\alpha}_t + \mathbf{G}_t\boldsymbol{\epsilon}_t, \qquad t = 1, 2, \ldots, n, \\
\boldsymbol{\alpha}_t &= \mathbf{T}_t\boldsymbol{\alpha}_{t-1} + \mathbf{H}_t\boldsymbol{\eta}_t,
\end{aligned}
$$

  where $\boldsymbol{\epsilon}_t \sim \mathrm{NID}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\eta}_t \sim \mathrm{NID}(\mathbf{0}, \mathbf{I})$, and $\mathrm{E}(\boldsymbol{\epsilon}_t\boldsymbol{\eta}_t') = \mathbf{0}$.

- The initial conditions are specified as $\boldsymbol{\alpha}_1|\boldsymbol{\delta} \sim \mathrm{N}(\tilde{\alpha}^*_{1|0} + \mathbf{W}_1\boldsymbol{\delta}, \mathbf{P}^*_{1|0})$.
- The model has a *Gaussian* structure.

Gibbs sampler

- Let $\boldsymbol{\Xi}$ denote the stack of the hyperparameters.
- A typical Gibbs sampling iteration is:
  - draw $\boldsymbol{\alpha}^{(i)} \sim p(\boldsymbol{\alpha}|, \boldsymbol{\Xi}^{(i-1)}, \mathbf{y})$ (simulation smoother)
  - draw $\boldsymbol{\Xi}^{(i)} \sim p(\boldsymbol{\Xi}|\boldsymbol{\alpha}^{(i-1)}, \mathbf{y})$

  see Kim and Nelson (1999), Carter and Kohn (1994, 1996) for the discrete filter.

The simulation smoother

- The simulation smoother is an algorithm which draws samples from the conditional distribution of the states and the disturbances given the observations and the hyperparameters. We focus on the simulation smoother proposed by Durbin and Koopman (2002).
- Let us define $\mathbf{x}_t$ denote a random vector (e.g. a selection of states or disturbances) and let $\tilde{\mathbf{x}} = \mathrm{E}(\mathbf{x}|\mathbf{y})$, where $\mathbf{x}$ is the stack of the vectors $\mathbf{x}_t$; $\tilde{\mathbf{x}}$ is computed by the Kalman filter and smoother.
- We can write $\mathbf{x} = \tilde{\mathbf{x}} + \mathbf{e}$, where $\mathbf{e} = \mathbf{x} - \tilde{\mathbf{x}}$ is the smoothing error, with conditional distribution $\mathbf{e}|\mathbf{y} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$, such that the covariance matrix $\mathbf{V}$ does not depend on the observations, and thus does not vary across the simulations (the diagonal blocks are computed by the smoothing algorithm).

A sample $\mathbf{x}^*$ from $\mathbf{x}|\mathbf{y}$ is constructed as follows:

- Draw $(\mathbf{x}^+, \mathbf{y}^+) \sim p(\mathbf{x}, \mathbf{y})$.
  As $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$, this is achieved by first drawing $\mathbf{x}^+ \sim p(\mathbf{x})$ from an unconditional Gaussian distribution, and constructing the pseudo observations $\mathbf{y}^+$ recursively from

$$\begin{aligned} \boldsymbol{\alpha}_t^+ &= \mathbf{T}_t \boldsymbol{\alpha}_{t-1}^+ + \mathbf{H}_t \boldsymbol{\eta}_t^+, \\ \mathbf{y}_t^+ &= \mathbf{Z}_t \boldsymbol{\alpha}_t^+ + \mathbf{G}_t \boldsymbol{\epsilon}_t^+, t = 1, 2, \ldots, n, \end{aligned}$$

  where the initial draw is $\boldsymbol{\alpha}_0^+ \sim \mathrm{N}(\mathbf{0}, \mathbf{H}_0 \mathbf{H}_0')$, so that $\mathbf{y}^+ \sim p(\mathbf{y}|\mathbf{x})$.
- The Kalman filter and smoother computed on the simulated observations $\mathbf{y}_t^+$ will produce $\tilde{\mathbf{x}}^+$, and $\mathbf{x}^+ - \tilde{\mathbf{x}}^+$ will be the required draw from $\mathbf{e}|\mathbf{y}$.
- Hence , $\tilde{\mathbf{x}} + \mathbf{x}^+ - \tilde{\mathbf{x}}^+$ is a sample from $\mathbf{x}|\mathbf{y} \sim \mathrm{N}(\tilde{\mathbf{x}}, \mathbf{V})$.

Proof of the simulation smoother

- Let

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_y & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy}, & \boldsymbol{\Sigma}_x \end{pmatrix}\right]$$

  with $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}'_{yx}$, then $\mathbf{x}|\mathbf{y} \sim \mathrm{N}(\tilde{\mathbf{x}}, \mathbf{V})$,

$$\tilde{\mathbf{x}} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)$$

$$\mathbf{V} = \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yx}$$

- Notice that $\mathrm{E}(\mathbf{x}|\mathbf{y})$ is linear in $\mathbf{y}$ and that $\mathbf{V}$ is invariant to $\mathbf{y}$.

We aim at drawing a random sample from $p(\mathbf{x}|\mathbf{y})$.

Let $(\mathbf{x}^+, \mathbf{y}^+) \sim p(\mathbf{x}, \mathbf{y})$ denote a draw from the joint distribution and

$$\tilde{\mathbf{x}}^+ = \mathrm{E}(\mathbf{x}^+|\mathbf{y}^+) = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-1}(\mathbf{y}^+ - \boldsymbol{\mu}_y).$$

(computed by the KFS).

Since $\mathbf{V}$ does not depend on $\mathbf{y}$, $\mathrm{Var}(\mathbf{x}^+|\mathbf{y}^+) = \mathbf{V}$, as well.

Also, unconditionally (LIE),

$$\mathrm{E}(\mathbf{x}^+ - \tilde{\mathbf{x}}^+) = \mathbf{0}, \mathrm{Var}(\mathbf{x}^+ - \tilde{\mathbf{x}}^+) = \mathbf{V},$$

and thus

$$\mathrm{E}[\tilde{\mathbf{x}} + (\mathbf{x}^+ - \tilde{\mathbf{x}}^+)|\mathbf{y}] = \tilde{\mathbf{x}}$$

$$\mathrm{Var}[\tilde{\mathbf{x}} + (\mathbf{x}^+ - \tilde{\mathbf{x}}^+)|\mathbf{y}] = \mathrm{E}[(\mathbf{x}^+ - \tilde{\mathbf{x}}^+)(\mathbf{x}^+ - \tilde{\mathbf{x}}^+)'|\mathbf{y}] = \mathbf{V}.$$

DSGE estimation

The DSGE model after linearization, see can be written as a linear state space model:

- Measurement:

$$y_t = \Psi(\theta) + \Psi_1(\theta)t + \Psi_2(\theta)\boldsymbol{\alpha} + u_t$$

- State transition:

$$\alpha_t = \Phi(\theta)\alpha_{t-1} + \Phi_\epsilon(\theta)\epsilon_t$$

- Joint density for the observation and latent states is:

$$p(\mathrm{Y}_{1:T}, \boldsymbol{\alpha}_{1:T}|\theta) = \prod_{t=1}^{T} p(y_t, \alpha_t|\mathrm{Y}_{1:t-1}, S_{1:t-1}, \theta)$$

$$\prod_{t=1}^{T} p(y_t|\alpha_t, \theta)p(\alpha_t|\alpha_{t-1}, \theta)$$

Estimation

- Let $\boldsymbol{\Xi}$ denote the stack of the hyperparameters.
- A typical Gibbs sampling iteration is:
    - draw $\boldsymbol{\alpha}^{(i)} \sim p(\boldsymbol{\alpha}|\boldsymbol{\Xi}^{(i-1)}, \mathsf{Y}_t)$ (simulation smoother)
    - draw $\boldsymbol{\Xi}^{(i)} \sim p(\boldsymbol{\Xi}|\boldsymbol{\alpha}^{(i-1)}, \mathsf{Y}_t)$

Algebra

- The Likelihood of the data

$$p(y^T|\mu) = (2\pi)^{-T/2} \exp\left\{-1/2 \sum (y_t - \mu)^2\right\}$$

Algebra

- The Likelihood of the data

$$p(y^T|\mu) = (2\pi)^{-T/2} \exp\left\{-1/2 \sum (y_t - \mu)^2\right\}$$

- The prior

$$p(\mu) = (2\pi\sigma^2)^{-1/2} \exp\left\{-1/2 \left(\frac{\mu - m}{\sigma}\right)^2\right\}$$

Algebra

- The Likelihood of the data

$$p(y^T|\mu) = (2\pi)^{-T/2} \exp\left\{-1/2 \sum (y_t - \mu)^2\right\}$$

- The prior

$$p(\mu) = (2\pi\sigma^2)^{-1/2} \exp\left\{-1/2 \left(\frac{\mu - m}{\sigma}\right)^2\right\}$$

- The posterior distribution is given by

$$p(\mu|y^T) = \frac{p(\mu)p(y^T|\mu)}{p(y^T)}$$

Algebra

- The Likelihood of the data

$$p(y^T|\mu) = (2\pi)^{-T/2} \exp\left\{-1/2 \sum (y_t - \mu)^2\right\}$$

- The prior

$$p(\mu) = (2\pi\sigma^2)^{-1/2} \exp\left\{-1/2 \left(\frac{\mu - m}{\sigma}\right)^2\right\}$$

- The posterior distribution is given by

$$p(\mu|y^T) = \frac{p(\mu)p(y^T|\mu)}{p(y^T)}$$

- Lets analyze first the numerator

The Numerator

$$
\begin{aligned}
p(y^T|\mu)p(\mu) &= (2\pi)^{-(T+1)/2} \exp\left\{-1/2\left[\sum(y_t-\mu)^2+\left(\frac{\mu-m}{\sigma}\right)^2\right]\right\} \\
&= (2\pi)^{-(T+1)/2} \exp\left\{-1/2\left[\sum y_t^2 + T\mu^2 - 2\mu\sum y_t + \mu^2/\sigma^2 + m^2/\sigma^2 - 2\mu m/\sigma^2\right]\right\} \\
&= (2\pi)^{-(T+1)/2} \exp\{-1/2[\mu^2(T+1/\sigma^2) - 2\mu(T\overline{y} + m/\sigma^2) + T\overline{y}^2 \\
&+ m^2/\sigma^2 + \sum(y_t-\overline{y})^2]\} \\
&= (2\pi)^{-(T+1)/2} \exp\left\{-1/2\left[1/\sigma_\mu^2(\mu-\widehat{\mu})^2 + Q\right]\right\} \\
&= (2\pi)^{-1/2} \exp\left\{-1/2\frac{(\mu-\widehat{\mu})^2}{\sigma_\mu^2}\right\}(2\pi)^{-T/2}\exp\left\{-1/2Q\right\}
\end{aligned}
$$

where

$$
\begin{aligned}
\sigma_\mu^2 &= 1/(T+1/\sigma^2) \\
\widehat{\mu} &= \sigma_\mu^2(T\overline{y} + m/\sigma^2) \\
Q &= T\overline{y}^2 + \mu^2/\sigma^2 + \sum(y_t-\overline{y})^2 - \sigma_\mu^2(T\overline{y}+m/\sigma^2)^2
\end{aligned}
$$

The denominator

$$
\begin{aligned}
p(y^T) &= \int p(y^T|\mu)p(\mu)d\mu \\
&= \int (2\pi)^{-1/2} \exp\left\{-1/2\frac{(\mu-\widehat{\mu})^2}{\sigma_\mu^2}\right\}(2\pi)^{-T/2}\exp\left\{-1/2Q\right\}d\mu \\
&= (2\pi)^{-T/2}\exp\left\{-1/2Q\right\}\int (2\pi)^{-1/2}\exp\left\{-1/2\frac{(\mu-\widehat{\mu})^2}{\sigma_\mu^2}\right\}d\mu
\end{aligned}
$$

Hence

$$p(\mu|y^T) \sim N(\widehat{\mu}, \sigma_\mu^2)$$

▸ return