Forecasting in the Bayesian way

Andreas E. Murr
Department of Politics & International Studies
University of Warwick
a.murr@warwick.ac.uk

Coventry, 20 March 2017

Objectives

- Introduce basic ideas of Bayesian inference.
- Highlight its advantages and disadvantages.
- Illustrate the process of Bayesian prediction and forecasting.
- Show how to estimate models and interpret their results.

Definitions

Probability

Frequentist: Long-run frequency of event.

Bayesian: Degree of belief.

Statistical inference

Draw conclusions from observed data y about unobserved parameters θ or a new observation $\tilde{y}$.

Bayesian inference

Draw conclusions in terms of probability statements.

Condition on the observed value of y: $p(\theta|y)$ or $p(\tilde{y}|y)$.

Example: A biased coin?

1. Set up a probability model (parametric or non-parametric):

$$p(y|\theta) = \text{Bin}(y|n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

2. Specify prior:

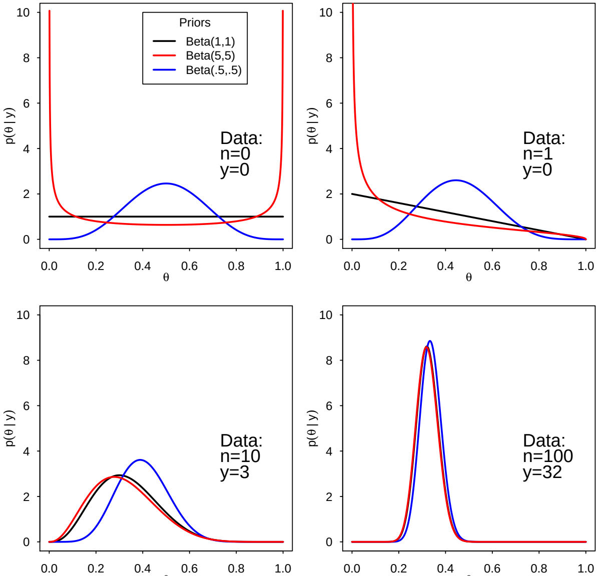
$$p(\theta) = \text{Beta}(\alpha, \beta)$$

3. Summarise posterior distribution:

$$p(\theta|y,n) = \text{Beta}(y+\alpha, n-y+\beta)$$

4. (Evaluate model adequacy.)

Example: A biased coin?



Bayes' rule

According to Bayes' rule:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1}$$

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{2}$$

'Bayesian mantra'

The posterior distribution $p(\theta|y)$ is proportional to the prior distribution $p(\theta)$ times the likelihood $p(y|\theta)$.

Requirement of Bayesian statistics:

Express prior belief about the parameter in the form of a probability distribution.

Similarities and differences to frequentist approach

Both approaches:

- Begin with a probability model (data generating process).
- Relate observed data y with a set of unknown parameters θ.
- Include fixed, known covariates x.
- Denote probability model as $p(y|\theta, x)$ or $p(y|\theta)$.

Similarities and differences to frequentist approach

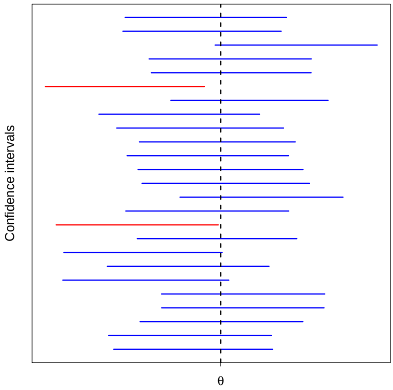
	Frequentist	**Bayesian**
Parameters (unknown)	Fixed	Random
Data (known)	Random	Fixed
Probability Model	$L(\theta\|y)$	$L(\theta\|y)p(\theta)$

Treating unknowns as random and knows as fixed has several advantages.

Confidence

Which of these is the correct interpretation of a 95% confidence interval?

- An interval that has a 95% chance of containing the true value of the parameter.
- An interval that over 95% of replications contains the true value of the parameter, *on average*.



Intuitive interpretation of findings

Frequentist approach:

- 95% confidence interval for θ is [1.5, 2.4].
- If we were to repeatedly draw from our population, 95% of our confidence intervals would contain the population parameter.
- But we do not know whether the present confidence interval contains the population parameter.

Bayesian approach:

- 95% credible interval for θ is [1.5, 2.4].
- After observing the data, there is a 95% chance that the parameter falls between 1.5 and 2.4.

Advantages and disadvantages

Advantages

- Intuitive interpretation of findings.
- Easy computation of quantities of interest.
- Incorporation of prior information.
- Fitting of realistic (complex) models.
- Handling of missing values.
- Inference with small samples.
- ...

Disadvantages

- Elicit and defend subjective information.
- Show that results do not depend on which prior is used.
- Computing the posterior distribution can be challenging.

Monte Carlo method

- Analytically summarising posterior distributions is often impossible or too cumbersome.
- Use Monte Carlo methods.
- (General method used also by frequentist approach.)

Monte Carlo principle

Anything we want to know about a random variable θ can be learned by sampling many times from $p(\theta)$, the density of θ.

Example: Compute posterior expected value

Analytical:

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

Computational:

- Produce random sequence of T draws $\theta^{(1)}$, $\theta^{(2)}$, ..., $\theta^{(T)}$ from $p(\theta|y)$.
- $E(\theta|y) \approx \frac{1}{T} \sum_{t=1}^{T} \theta^{t}$.

Markov chain Monte Carlo

- Bayesian inference relies typically on Markov chain Monte Carlo.
- (MCMC can also be used by the frequentist approach, but this is not widespread yet.)
- The sequence of draws $\theta^{(1)}$, $\theta^{(2)}$, ..., $\theta^{(T)}$ are dependent.
- Each draw $\theta^{(t+1)}$ depends only on the previous draw $\theta^{(t)}$ (Markov chain).
- Construct algorithms so that the Markov chain converges to the target distribution.
- The two most common algorithms are
  - the Gibbs sampling algorithm and
  - the Metropolis-Hastings algorithm.
- Gibbs is a special case of Metropolis-Hastings.

Computation

These and other algorithms allow us to sample from

- multidimensional distributions (e.g., Gibbs), and
- any distribution irrespective its shape (e.g., Metropolis).

Many extensions, such as:

- Metropolis-Hastings.
- Metropolis-within-Gibbs.

Always run diagnostic tests, such as:

- Are traceplots stationary?
- Do chains with different starting values converge?

Software

- In R, use JAGS, rjags, coda, and superdiag.
- Example: Linear regression model with semi-conjugate priors

```
model {
  # likelihood
  for (i in 1:N){
    y[i] ~ dnorm(mu[i], tau)
    mu[i] ~ alpha + beta*x[i]
  }
  # prior
  alpha ~ dnorm(0, 0.001)
  beta ~ dnorm(0, 0.001)
  tau ~ dgamma(0.001, 0.001)
}
```

- The number of models you can estimate is pretty much unlimited.

Prediction

Why?

- To impute missing or censored data.
- To predict replicate datasets in order to check adequacy of model.
- To know what happens in the future.

Bayesian prediction

Bayesians want the appropriate posterior predictive distribution for $\tilde{y}$ to account for all sources of uncertainty.

Sources of uncertainty:

- Uncertainty about $E(\tilde{Y})$,
- sampling variability of $\tilde{Y}$ around its expectation,
- uncertainty about the size of this variability, and
- the correlations between these components.

Estimation

- *Frequentist*: Easy to get point prediction, harder to get predictive distribution.
- *Bayesian*: Trivial to get predictive distribution using MCMC.

Software

- No need to explicitly include the quantities to be predicted in the model description.
- Expand the data set by including missing data indicated as `NA`.
- For instance, instead of

  ```
  data=list("x"=c(8,1,3),"y"=c(2,4,7))
  ```

  write

  ```
  data=list("x"=c(8,1,3,2),"y"=c(2,4,7,NA))
  ```

  This computes the predictive distribution of $\tilde{y}|x = 2$.

Time Series ('Bayesian forecasting')

Time series

Data arising in sequence over time.

Observations are likely to be dependent.

Forecasting

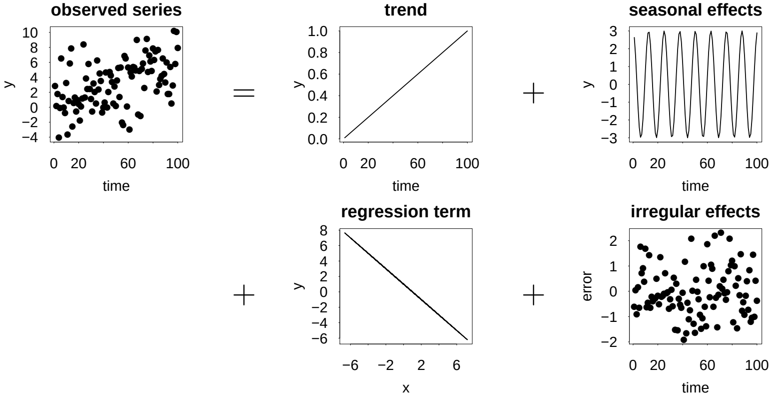
Extrapolating series into the short-, medium, or long-term future.

Use dependency through time: e.g., $\tilde{y}_{t+1} = \hat{\alpha} + \hat{\beta} y_t$.

Use know future values of input: e.g., $\tilde{y}_{t+1} = \hat{\alpha} + \hat{\beta} x_{t+1}$.

Typical model

observed series = trend + seasonal effects + regression term + irregular effects.



Dynamic Linear Models

Regression coefficients and variance of irregular effects may vary over time.
Consider the usual linear regression model

$$y_t = \mathbf{X}_t \beta + \epsilon_t \text{ ('observation model')}$$

but with changing coefficient vector β_t such that

$$\beta_t = \mathbf{M}_t \beta_{t-1} + \omega_t \text{ ('state model')}$$

where $\mathbf{M}_t$ is a transition matrix.
ϵ_t and ω_t can have time-dependent variances $\mathbf{V}_t$ and $\mathbf{W}_t$.

Some common simplifications:

- Assume $\mathbf{V}_t$ and $\mathbf{W}_t$ are constant over time ($\mathbf{V}$ and $\mathbf{W}$).
- Assume state parameters vary independently of each other, so matrix $\mathbf{W}_t$ reduces to a vector W_t.
- Assume that $\mathbf{M}_t$ is known and fixed in time (e.g., $\mathbf{M}_t = \mathbf{I}$ the identity matrix so $\beta_t = \beta_{t-1} + \omega_t$).

Software

Dynamic linear regression model where $y_t = x_t\beta + \epsilon_t$ and $\beta_t = \beta_{t-1} + \omega_t$ with constant V and W in JAGS:

```
model{
  # observation model
  for (t in 1:T){
    y[t] ~ dnorm(mu[t], V.inv)
    mu[t] <- x[t]*beta[t]
  }
  # state model
  for (t in 2:T){
    beta[t] ~ dnorm(beta[t-1], W.inv)
  }
  # settings for t=1
  beta[1] ~ dnorm(10,0.01)
  # priors
  ...
}
```

Conclusions

Bayesian methods allow you

- to answer questions like “What is the probability that ...”,
- to easily make predictions based on your model, and
- to fit “models with many parameters and complicated multilayered probability specifications”.

The software for estimating Bayesian methods is free and relatively easy to use.

References

- Bayesian Analysis for the Social Sciences (Simon Jackman, Wiley)
- Bayesian Data Analysis (Andrew Gelman et al., Chapman & Hall/CRC)
- Bayesian Methods (Jeff Gill, Chapman & Hall/CRC)
- Data Analysis Using Regression and Multilevel/Hierarchical Models (Andrew Gelman and Jennifer Hill, CUP)
- The Theory That Would Not Die (Sharon McGrayne, Yale)

Motivation

What is the probability that ...

- ... Andrew Jackson was the eighth president of the United States?
  Frequentist statistics cannot make probability statements about single-events.
- ... austerity measures improve the economy ($t = 60, n = 50$)?
  Frequentist statistics struggles to make inferences if the sample is the population.
- ... defeat in war leads to revolution in Latin America ($n = 76$)?
  Frequentist statistics is too conservative if the sample size is small.

Gibbs sampling

Use a sequence of draws from conditional distributions to characterise the joint target distribution.

1. Define target distribution: $p(\theta_1, \theta_2, ..., \theta_K | y)$.
2. Set starting values: $\theta_2^{(0)}, ..., \theta_K^{(0)}$.
3. Repeat for $t = 1, ..., T$ iterations:

$$
\begin{array}{lll}
\text{Draw } \theta_1^{(t)} & \text{from} & p(\theta_1 | \theta_2^{(t-1)}, ..., \theta_K^{(t-1)}, y) \\
\text{Draw } \theta_2^{(t)} & \text{from} & p(\theta_2 | \theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_K^{(t-1)}, y) \\
\vdots & \vdots & \vdots \\
\text{Draw } \theta_K^{(t)} & \text{from} & p(\theta_K | \theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{K-1}^{(t)}, y)
\end{array}
$$

Example: Bivariate normal distribution

Simulate from bivariate normal distribution with zero mean and unit variance for the marginals:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

Suppose we do not know how to directly sample from this joint distribution.
However, we know that

$$x|y \sim \mathrm{N}(\rho y, 1-\rho^2) \tag{3}$$

$$y|x \sim \mathrm{N}(\rho x, 1-\rho^2) \tag{4}$$

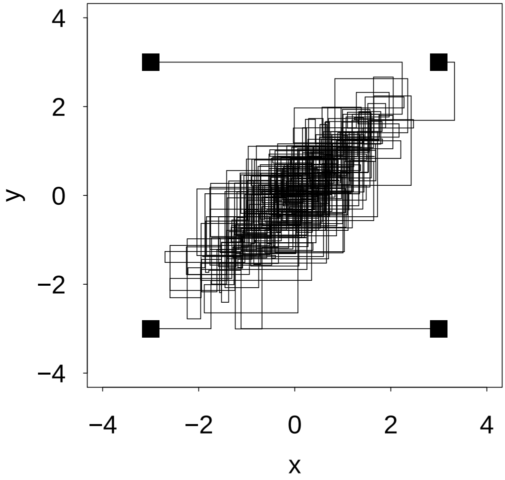
Which is the Gibbs sampler, so we can indirectly sample from the joint distribution.

Example: Bivariate normal distribution

In R we could do this as follows:

```
Gibbs <- function(n, rho, x0=0, y0=0){
  draws <- matrix(ncol=2, nrow=n)
  x <- x0
  y <- y0
  draws[1,] <- c(x, y)
  for (i in 2:n){
    x <- rnorm(1, rho * y, sqrt(1 - rho^2))
    y <- rnorm(1, rho * x, sqrt(1 - rho^2))
    draws[i,] <- c(x, y)
  }
  draws
}
```

Example: Bivariate normal distribution



Metropolis

Use a sequence of draws from a distribution from which we know how to sample to characterise a distribution from which we do not know how to sample.

1. Define target distribution: $p(\theta|y)$.
2. Set starting value: $\theta^{(0)}$.
3. Repeat for $t = 1, \ldots, T$ iterations:
    1. Sample a proposal θ^* from jumping distribution.
    2. Calculate the ratio
    $$r = \frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}. \tag{5}$$
    3. Set
    $$\theta^{(t)} = \begin{cases} \theta^* \text{ with probability } \min(r, 1) \\ \theta^{(t-1)} \text{ otherwise.} \end{cases}$$

Example: Indirectly sample from $N(0,1)$

Simulating from a normal with zero mean and unit variance using a uniform proposal distribution.

- Start the chain at $x = 0$.
- At each iteration propose an innovation $y \sim \text{Unif}(-\alpha, \alpha)$, leading to a candidate $x + y$.
- Calculate the acceptance probability $\min\left(\frac{N(x+y)}{N(x)}, 1\right)$.
- Accept candidate if acceptance probability $>$ than a random draw from $\text{Unif}(0, 1)$, reject otherwise.

Example: Indirectly sample from $N(0, 1)$

In R we could do this as follows:

```
Metropolis <- function(n, alpha, x0=0){
  draws <- rep(NA, n)
  x <- x0
  draws[1] <- x
  for (i in 2:n) {
    inno <- runif(1, -alpha, alpha)
    cand <- x + inno
    apro <- min(1, dnorm(cand)/dnorm(x))
    u <- runif(1)
    if (u < apro)
      x <- cand
      draws[i] <- x
  }
  draws
}
```

Example: Indirectly sample from $N(0, 1)$