Introductory Survey of Bayesian Methods
Considering Dynamic Linear Models

For the Kx Community NYC Meetup, 23 January 2020
by Devon McCormick, CFA
devonmcc@gmail.com
What is Bayesian Statistics?

Bayesian: "Evidence about the true state of the world represents a degree of belief" versus

Frequentist: "Evidence represents only measured frequencies of data"

\[ P(H | D) = \frac{P(H) P(D | H)}{P(D)} \]

versus

\[ P(H) = \frac{\text{# Successes}}{\text{# Trials}} \]

"State of mind" versus "State of the world"
"Subjective" versus "Objective"

PROBABILITY DOES NOT EXIST

The abandonment of superstitious beliefs about the existence of the Phlogiston, the Cosmic Ether, Absolute Space and Time, ... or Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs. (p. x)

- de Finetti, "Theory of Probability"
**Inverse Probability**

**Conventional Probability**

10% have disease
Sample of 20 tested...

Two should be infected.

**Inverse Probability**

Two out of 20 tested have the disease...

How many in the population have the disease?
The Fall and Rise of Bayesian Statistics

Early Timeline: the genesis and decline of Bayesianism...

Thomas Bayes  Pierre-Simon Laplace  JS *  ...  R.I.P. Bayes  R. A. Fisher
Richard Price  George Chrystal  (frequentist)
1761  1774  ...  1843  1891  1925 ...

* John Stuart Mill denounced probability as "ignorance... coined into science."

Later Timeline: the rise of Bayesianism

Émile Borel  Bruno de Finetti  Harold Jeffreys  Arthur Bailey  L.J. Savage
Frank Ramsey  A. M. Turing  I. J. Good
1920s  1926/1931  1928/1937  1939  1940  1950  1954

The Fall and Rise of Bayesianism
Growth of the Use of Bayesian Methods

Times have changed. Beginning in the early 1990s, there was an abrupt proliferation of studies using Bayesian methods in mainstream statistics.

From "Prior approval: The growth of Bayesian methods in psychology" by Mark Andrews & Thom Baguley, November 30, 2012
Motivation and Derivation of Bayes's Law

From Jaynes:

Strong syllogism (Aristotle, fourth century BCE):
Major premise: if A is true, then B is true
Minor premise: A is true
Conclusion: therefore, B is true
and its inverse:
Major premise: if A is true, then B is true
Minor premise: B is false
Conclusion: therefore, A is false

Weaker syllogism "epagoge":
Major premise: if A is true, then B is true
Minor premise: B is true
Conclusion: therefore, A becomes more plausible
and its inverse:
Major premise: if A is true, then B is true
Minor premise: A is false
Conclusion: therefore, B becomes less plausible

From Downey:

Conjunction is commutative:
\[ p(A \text{ and } B) = p(B \text{ and } A) \]

Probability of a conjunction:
\[ p(A \text{ and } B) = p(A) \cdot p(B | A) \]

Interchanging A & B:
\[ p(B \text{ and } A) = p(B) \cdot p(A | B) \]

Therefore,
\[ P(B) \cdot p(A | B) = p(A) \cdot p(B | A) \]

Dividing by \( p(B) \):
\[ \frac{p(A | B)}{p(B)} = \frac{p(A)}{p(B | A)} \]

\[ \text{Bayes's Law!} \]
Why Bayes?

The lady tasting tea... Choosing Mozart...

Flipping a coin...

How do these cases differ, based only on the evidence?

\[
P(H|D) = \frac{P(H) \cdot P(D|H)}{P(D)}
\]

Prior

Likelihood

Posterior

Normalizing constant
The Problem with Priors

Given 12 coin flips, 3 of which were tails, how likely is it that the coin used is fair?

We cannot answer this question without specifying an underlying probability model.

So, for the case where the experiment was designed to specify 12 coin flips, we would assume a binomial distribution:

\[ L_1(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \binom{12}{9} \theta^9 (1-\theta)^3 \]

However, if the experiment instead had specified that we would continue flipping until we saw three tails, we would use a negative binomial distribution:

\[ L_2(\theta) = \binom{r+x-1}{x} \theta^x (1-\theta)^r = \binom{11}{9} \theta^9 (1-\theta)^3 \]

The two distributions give different results when evaluated:

\[ \alpha_1 = P_{\theta=1/2}(X \geq 9) = \sum_{j=9}^{12} \binom{12}{j} \theta^j (1-\theta)^{12-j} = 0.075 \]

\[ \alpha_2 = P_{\theta=1/2}(X \geq 9) = \sum_{j=9}^{\infty} \binom{2+j}{j} \theta^j (1-\theta)^3 = 0.0325 \]
Estimating the Probability of a Rare Event

To estimate the prevalence $\Theta$ of an infectious disease in a small city, we test a sample of 20 people, giving us the number "y" of infected people in this group.

So, the parameter and sample spaces are

$\Theta = [0,1]$  \  $y=\{0,1,...,20\}$

**Sampling Model**

Before testing, let "Y" be the number of infected people we will determine. For unknown $\Theta$, a reasonable sampling model for $Y$ might be the binomial($20,\Theta$) distribution, so

$Y|\Theta \sim \text{binomial}(20,\Theta)$

Say we check a sample of 20 people for infection, and we find two cases. Based on this, what is our estimate of rate of infection for the city as a whole?
Choosing and Updating a Prior Distribution

Studies from around the country show that infection rates in comparable cities range from 0.05 to 0.20, with an average of 0.10.

We want to build a prior distribution with a substantial portion in the range (0.05, 0.20) with an expected value close to 0.10.

We don't know the actual family of distributions from which to draw, so we'll use the beta distribution since it's flexible and easily interpretable.

A beta distribution is defined by two parameters "a" and "b", for which the expectation $\Theta$ is $a/(a+b)$ and the most probable value is $(a-1)/(a-1 + b-1)$.

So, $\Theta \sim \text{beta}(2,18)$ gives an expectation of $2/20 = 0.10$ and a most probable value of $1/18 = 0.06$.

If $Y|\Theta \sim \text{binomial}(n,\Theta)$ and $\Theta \sim \text{beta}(a,b)$ and we observe a value "y" of Y, the posterior is a $\text{beta}(a+y, b+n-y)$ distribution.

Graphing this in R:

```r
a<-2; b<-18; n<-20; y<-0
curve(dbeta(x, a+y, b+n-y), lty=1, lwd=3, from=0, to=1, xlab="Percentage Infected in Population", ylab="Probability")
curve(dbeta(x, a, b), add=TRUE, lty=3, lwd=3)
legend(0.5, 8, c("p(\Theta|y)", "p(\Theta)"), lty=c(1, 3), lwd=c(3, 3))
```

(example from "A First Course in Bayesian Statistical Methods by Peter D. Hoff")
Markov Chain Monte Carlo

Probabilities of transition (edge) represented by a transition matrix.

NB. 5-point equiprobable random walk w/wrap: simple ring.

Finding the Stationary Distribution

```
sqrMat^:(i.5) ] rw5tm0
0 0.5 0 0 0.5
0.5 0 0.5 0 0
0 0.5 0 0.5 0
0 0 0.5 0 0.5
0.5 0 0 0.5 0
```

```
getDiag sqrMat^:_ ] rw5tm0
0.2 0.2 0.2 0.2 0.2
```

What is the probability of ending up at a given node?
MCMC continued...

NB. 5-point ring random walk: 0<->3
rw5tm1=: "\.&a:-.--", _1 LF, 0 : 0
0 0.33 0 0.33 0.34
0.5 0 0.5 0 0
0 0.5 0 0.5 0
0.2 0 0.4 0 0.4
0.5 0 0 0.5 0
)
sample=: 4 : '1 i.~ (?0)&<:x{y"0 2

MCdraw=: 3 : 0
'tm nd stval'=. y  NB. Transition mat, # draws, start node
states=. (stval) 0}nd$0
tm=. +/"1 tm  NB. Form for "sample"
for_ix. }.i.#states do. states=. ((states{~<:ix) sample tm) ix}states end.
(#states)%~<:/:~(i.#tm), states
NB. EG MCdraw rw5tm1; 1e5; 0

MCdraw rw5tm1; 1e5; 0
0.22202 0.16524 0.18316 0.25299 0.17659

A More Interesting Graph

(From "Probabilistic Graphical Models - Principles and Techniques by Daphne Koller and Nir Friedman")

Grade

SAT
MCMC-based Methods

Metropolis-Hastings Algorithm

Given $X^{(t)} = x^{(t)}$,

1. Generate $Y_t \sim q(y|x^{(t)})$.
2. Take
   
   $$X^{(t+1)} = \begin{cases} 
   Y_t & \text{with probability } \rho(x^{(t)}, Y_t), \\
   x^{(t)} & \text{with probability } 1 - \rho(x^{(t)}, Y_t),
   \end{cases}$$

   where

   $$\rho(x, y) = \min \left\{ \frac{\pi(y)}{\pi(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}.$$ 

Metropolis-Hastings often has sub-optimal convergence or may have other convergence problems. However, it provides a good baseline solution that is simple to implement and may be combined with other methods.

Gibbs Sampling

For a multi-variate $\Theta$ with differing distributions for each $\Theta_i$, do the following steps:

0. Assign a vector of starting values, $S$, to the parameter vector:
   
   $$\Theta^{j=0} = S.$$
1. Set $j = j + 1$.
2. Sample $(\theta_1^j | \theta_2^{j-1}, \theta_3^{j-1}, \ldots, \theta_k^{j-1})$.
3. Sample $(\theta_2^j | \theta_1^j, \theta_3^{j-1}, \ldots, \theta_k^{j-1})$.
   
   $\vdots$
4. Sample $(\theta_k^j | \theta_1^j, \theta_2^j, \ldots, \theta_{k-1}^j)$.
5. Return to step 1.

Both Metropolis-Hastings and Gibbs sampling have many implementations in R as well as in other languages.

Gibbs sampling is also implemented in a number of packages such as OpenBUGS and JAGS.
Bayesian Networks

Therefore, according to this scenario, it's more likely the grass is wet because it's raining than because the sprinkler is on.

\[
p(S=1|W=1) = \frac{\sum p(C=c, S=1, R=r, W=1)}{p(W=1)}
\]

\[
= \frac{0.278/0.647}{0.43} = 0.43
\]

\[
p(R=1|W=1) = \frac{\sum p(C=c, S=s, R=1, W=1)}{p(W=1)}
\]

\[
= \frac{0.456/0.647}{0.71} = 0.71
\]

where

\[
p(W=1) = \sum p(C=c, S=s, R=r, W=1) = 0.647
\]

However, S and R are conditionally dependent because of their common child W, so:

\[
p(S=1|W=1, R=1) = 0.1945
\]

It's less likely the sprinkler is on if it's raining.

A node is independent of its ancestors given its parents.

By the chain rule of probability, the joint probability of all the nodes in the graph above is

\[
\]

By using conditional independence relationships, we can rewrite this as

\[
P(C, S, R, W) = P(C) * P(S|C) * P(R|C) * P(W|S,R)
\]
Hidden nodes of a network may be the outcome in which we are interested or they may simplify the complexity of a network, perhaps providing us a way to reason about causality.

Bayesian network tools: Microsoft’s MSBNx, Matlab’s Bayes Net Toolbox, Bayesia. Various packages in R: bnlearn (network structure learning), deal, ebdbNet.

Figure 4. Inference model for Bayesian network multimodal sensor fusion with weights.
Dynamic Linear Models

DLMs are "state space" models - they characterize a state by multidimensional set of points and embody a system equation underlying the observation equation.

Observation equation: \[ Y_t = F_t \theta_t + \nu_t, \quad \nu_t \sim N[0, V_t], \]

System equation: \[ \theta_t = \theta_{t-1} + \omega_t, \quad \omega_t \sim N[0, W_t], \]

Initial information: \( (\theta_0 | D_0) \sim N[m_0, C_0] \)

A dynamic linear model is characterized by a set of quadruples \( \{F, G, V, W\}_t \)

For each time \( t \) where

- \( F_t \) is a known \( (n \times r) \) matrix; (regression)
- \( G_t \) is a known \( (n \times n) \) matrix; (state transition)
- \( V_t \) is a known \( (r \times r) \) variance matrix; (observation variance)
- \( W_t \) is a known \( (n \times n) \) variance matrix. (unobserved system variance)
DLMs Encompass Many Other Types of Models

West and Harrison, in "Bayesian Forecasting and Dynamic Linear Models", devote a few pages to demonstrating that a number of popular forecasting techniques are, in fact, reducible to some of the more simply-parameterized DLMs.

Among these are:

(a) Holt's point predictor.
\[ M_t = \alpha Y_t + (1 - \alpha) M_{t-1}, \quad (t > 1). \]

(b) Exponentially weighted moving averages (EWMA).
Given the sequence \( Y_t, \ldots, Y_1 \), the EWMA with parameter \( \delta, \ (0 < \delta < 1) \), is defined as
\[ M_t = \frac{(1 - \delta)^{t-1}}{(1 - \delta^t)} \sum_{j=0}^{t-1} \delta^j Y_{t-j}. \]

(c) Brown's exponentially weighted regression (EWR).
The EWR estimator (Brown, 1962) for a locally constant mean of \( Y_t, \ldots, Y_1 \), is defined as the value \( M_t = \mu \) that minimises the discounted sum of squares
\[ V_t(\mu) = \sum_{j=0}^{t-1} \delta^j (Y_{t-j} - \mu)^2 \]
for some given discount factor \( \delta \), \( (0 < \delta < 1) \).

(d) An alternative ARIMA(0,1,1) model representation.
The predictors of Box and Jenkins (1976) ...
...their widely applied ARIMA(0,1,1)...
is a] closed, constant, first-order polynomial DLM. (DLM \( \{1,1,V,W\} \))
**Dynamic Evolution**

First-Order Polynomial DLM, with Constant Variance $V$

\[
Y_t = \mu_t + \nu_t, \quad \nu_t \sim N(0, V),
\]

\[
\mu_t = \mu_{t-1} + \omega_t, \quad \omega_t \sim T_{n_t-1}[0, W_t].
\]

Information:

\[
(\mu_{t-1} | D_{t-1}) \sim T_{n_{t-1}}[m_{t-1}, C_{t-1}],
\]

\[
(\phi | D_{t-1}) \sim G \left[ \frac{n_t - 1}{2}, \frac{n_{t-1} S_{t-1}}{2} \right].
\]

Forecast:

\[
(Y_{t} | D_{t-1}) \sim T_{n_{t-1}}[f_t, Q_t],
\]

\[
R_t = C_{t-1} + W_t,
\]

\[
Q_t = R_t + S_{t-1}.
\]

Updating Recurrence Relationships

\[
(\mu_t | D_t) \sim T_{n_t}[m_t, C_t],
\]

with

\[
m_t = m_{t-1} + A_t e_t,
\]

\[
C_t = A_t S_t.
\]

\[
(\phi | D_t) \sim G \left[ \frac{n_t}{2}, \frac{n_t S_t}{2} \right],
\]

with

\[
n_t = n_{t-1} + 1,
\]

\[
S_t = S_{t-1} + S_{t-1} \left( \frac{e_t^2}{Q_t} - 1 \right),
\]

where \( e_t = Y_t - f_t, \) and \( A_t = R_t/Q_t. \)

**k-Step Forecast Distributions**

Bayesian Forecasting and Dynamic Models by West and Harrison

\[
(Y_{t+k} | D_t) \sim T_{n_t}[m_t, Q_t(k)], \quad (X_t(k) | D_t) \sim T_{n_t}[km_t, L_t(k)],
\]

with \( Q_t(k) = C_t + \sum_{j=1}^{k} W_{t+j} + S_t, \)

and

\[
L_t(k) = k^2 C_t + \sum_{j=1}^{k} j^2 W_{t+k+j} + k S_t,
\]

where for \( j > 0 \) and scale-free variances \( W^*_{t+j}, \)

\[
W_{t+j} = S_t W^*_{t+j}.
\]

Consider the DLM specified by

\[
\theta_{t-1} | y_{1:t-1} \sim N(m_{t-1}, C_{t-1}).
\]

Then the following statements hold.

(i) The one-step-ahead predictive distribution of \( \theta_t \) given \( y_{1:t-1} \) is Gaussian, with parameters

\[
a_t = E(\theta_t | y_{1:t-1}) = G_t m_{t-1},
\]

\[
R_t = \text{Var}(\theta_t | y_{1:t-1}) = G_t C_{t-1} G_t' + W_t.
\]

(ii) The one-step-ahead predictive distribution of \( Y_t \) given \( y_{1:t-1} \) is Gaussian, with parameters

\[
f_t = E(Y_t | y_{1:t-1}) = F_t a_t,
\]

\[
Q_t = \text{Var}(Y_t | y_{1:t-1}) = F_t R_t F_t' + V_t.
\]

(iii) The filtering distribution of \( \theta_t \) given \( y_{1:t} \) is Gaussian, with parameters

\[
m_t = E(\theta_t | y_{1:t}) = a_t + R_t F_t' Q_t^{-1} e_t,
\]

\[
C_t = \text{Var}(\theta_t | y_{1:t}) = R_t - R_t F_t' Q_t^{-1} F_t R_t,
\]

where \( e_t = Y_t - f_t \) is the forecast error.

**Dynamic Linear Models in R**

by Petris, et al.
Signal to Noise Ratio

"A first-order polynomial DLM \( \{1, 1, V_t, W_t\} \), constant or otherwise, depends heavily on choosing appropriate values for the variances, \( V_t \) and \( W_t \)." (West and Harrison, p.49)

The signal-to-noise ratio corresponds to the ratio of the two variances \( V_t \) (observation) and \( W_t \) (system). This ratio balances the model's response between the two systems of equations, as illustrated in this example graphed using R's "dlm" package.

```r
#* NileDLMeg.R: Nile levels example from "Dynamic Linear Models
# in R" by Petris et al.
NilePoly <- dlmModPoly(order = 1, dV = 15100, dW = 1468)
unlist(NilePoly); NileFilt <- dlmFilter(Nile, NilePoly)
str(NileFilt, 1); n <- length(Nile)
attach(NileFilt)
dlmSvd2var(U[[n + 1]], D.C[n + 1, ])

plot(Nile, type="o", col = c("blue"), lwd=2, xlab = "", ylab = "Annual
Average Discharge (m^3/s) of Nile River at Aswan")
mod1 <- dlmModPoly(order = 1, dV = 15100, dW = 755)
NileFilt1 <- dlmFilter(Nile, mod1)
lines(dropFirst(NileFilt1$m), lty = "longdash", lwd=2, col=c("darkgreen"))
mod2 <- dlmModPoly(order = 1, dV = 15100, dW = 7550)
NileFilt2 <- dlmFilter(Nile, mod2)
lines(dropFirst(NileFilt2$m), lty = "dotdash", col=c("red"), lwd=2)
leg <- c("data", paste("filtered, W/V = ", format(c(W(mod1) / V(mod1)),
W(mod2) / V(mod2))))
legend("bottomright", legend = leg,
col=c("blue", "darkgreen", "red"),
lt = c("solid", "longdash", "dotdash"),
lwd= c(2,2,2),
pch = c(1, NA, NA), bty = "n")
```

[Graphs and plots illustrating the signal-to-noise ratio and Nile levels data]
Introductory Survey of Bayesian Statistical Methods with Consideration of Dynamic Linear Models

Kx Community NYC Meetup
23 January 2020

First-Order Polynomial DLM, with Constant Variance $V$

Observation: $Y_t = \mu_t + \nu_t$, $\nu_t \sim N(0, V)$,
System: $\mu_t = \mu_{t-1} + \omega_t$, $\omega_t \sim T_{n_t=1}[0, W_t]$.
Information: $(\mu_{t-1} | D_{t-1}) \sim T_{n_t=1}[m_{t-1}, C_{t-1}]$,
$(\phi | D_{t-1}) \sim G\left[\frac{n_t-1}{2}, \frac{n_t-1}{2}\right]$.

Forecast: $(\mu_t | D_{t-1}) \sim T_{n_t=1}[m_t, C_t]$,
$(Y_t | D_{t-1}) \sim T_{n_t=1}[f_t, Q_t]$,
with $f_t = m_{t-1}$,
$R_t = C_{t-1} + W_t$,$Q_t = R_t + S_{t-1}$.

Updating Recurrence Relationships
$(\mu_t | D_t) \sim T_{n_t}[m_t, C_t]$,
with $m_t = m_{t-1} + A_t e_t$,$C_t = A_t S_t$,$(\phi | D_t) \sim G\left[\frac{n_t}{2}, \frac{n_t}{2}\right]$,
with $n_t = n_{t-1} + 1$,$S_t = S_{t-1} + S_{t-1} \left(\frac{e_t^2}{Q_t} - 1\right)$,
where $e_t = Y_t - f_t$, and $A_t = R_t/Q_t$.

$k$-Step Forecast Distributions
$(Y_{t+k} | D_t) \sim T_{n_t}[m_t, Q_t(k)]$,
$(X_t(k) | D_t) \sim T_{n_t}[m_t, L_t(k)]$,
with $Q_t(k) = C_t + \sum_{j=1}^{k} W_{t+j} + S_t$,
and $L_t(k) = k^2 C_t + \sum_{j=1}^{k} j^2 W_{t+k+1-j} + k S_t$,
where for $j > 0$ and scale free variances $W_{t+j}^*$,$W_{t+j} = S_t W_{t+j}^*$.

DLM Summary

Multivariate DLM: $n = 3$ : 0
tit= 'ft','Qt','mt';'Ct';'At';'et';'St' NB. Result columns
ni=. $ff [ delta= . %:delta [ 'ff gg Vt Wt y mc delta'= . 7{.y,1
dmo= a:$~#tit [ nt=. <:np=. Sp=. 1
'mp Cp='.(<Wt)$&.mc NB. Prior mean and variance: coerce shapes if scalar
at=. mp
while. ni>nt do.

Yt= y{.,nt [ Ft= . |:ff{.,nt
at= gg/+ . *mp
Wt= (delta*Cp incorporate gg*delta)-Cp incorporate gg NB. Discounting
Rt= Wt+Cp incorporate gg NB. Prior at time t: theta variance
ft= (|:Ft)+ . * at
Qt= Sp+Rt incorporate |:Ft
et= Yt-ft
nt= >:np NB. 1-step forecast +means
At= Rt/+ . *Ft+ . *%.Qt
mt= at+At/+ . * et
Ct= Rt-(:At)+ . *At+ . *Qt NB. and variance for forecast Y
St= #Sp+:(:%.Qt) incorporate et)*Sp%nt NB. Scaling factor (p. 110 (d))

Ct= Ct*St%Sp

end.
dmo= tit,dmo

)}
Further Study

Books

General
[Hoff 2009] A First Course in Bayesian Methods; Peter D. Hoff.


Bayesian Networks
[Koller 2009] Probabilistic Graphical Models - Principles and Techniques; Daphne Koller et al.

Dynamic Linear Models
[Petris 2009] Dynamic Linear Models with R; Giovanni Petris et al.

[West 1997] Bayesian Forecasting and Dynamic Linear Models; Mike West, Jeff Harrison.

Links
A Brief Introduction to Graphical Models and Bayesian Networks, by Kevin Murphy, 1998; http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html
Bayesian Thinking blog - https://learnbayes.wordpress.com

Software

The BUGS Project - www.mrc-bsu.cam.ac.uk/software/bugs/
BUGS (Bayesian inference Using Gibbs Sampling) for the Bayesian analysis of statistical models.
JAGS - mcmc-jags.sourceforge.net/ : JAGS (Just Another Gibbs Sampler) for analysis of Bayesian hierarchical models using MCMC.
R: See "https://cran.r-project.org/web/views/Bayesian.html" : list of R packages for Bayesian inference.
Matlab: Various packages - see mathworks.com