Bayesian Methods

Mark Fisher

Research
Federal Reserve Bank of Atlanta

October 27, 2017

Outline

Outline

Bayesian Method(s)

1. **Bayesian methods**: A "bag of tricks"
   - Reach in and grab one (when it's convenient)
2. **Bayesian method**: An approach to doing **inference**
   - Distinct from the "frequentist" approach
   - Instead of a *bag of tricks*, it's more like a *school of magic*
     - The tricks (Bayesian methods) emerge organically from the principles of the discipline (the Bayesian method)
   - As *Obi-wan Kenobi* said to *Luke Skywalker*
     - "You must learn the ways of the ~~force~~ **method** if you're to come with me to ~~Alderaan~~ **inference**"
3. The **method** is powerful
   - "For my ally is the ~~force~~ **method**, and a powerful ally it is." –*Yoda*
4. But the bag-of-tricks way of thinking leads to the **dark side**

   (i.e., using Bayesian methods for frequentist purposes)
   - "The dark side of the ~~force~~ **method** is a pathway to many abilities some consider to be unnatural." –*Chancellor Palpatine*

Outline

1. Bits and pieces regarding the **method in general**
    - How a Bayesian uses probability
    - Bayes' rule
        - What it is
        - Recursive updating
        - Moving targets
    - Sampling distributions versus posterior distributions
        - Conceptual issue
    - Gibbs sampler and Rao–Blackwellization
    - Regression
        - Two ways to express the model
2. **How I use the method** to learn about **mutual fund skill**
    - Linear factor model for mutual fund returns
    - Bayesian density estimation
        - Calculating the predictive distribution
    - Computing a well-informed prior
        - Starting with an open-minded prior
    - Learning about skill within fund-regimes

:(Very little discussion of the numerical methods involved

Outline

How a Bayesian uses probability

1. Probability is used to characterize **information**
   - Probability is not about long-run **frequencies** *per se*
2. Consider a hypothesis H that has a fixed, unknown truth value
   - A Bayesian can assign a probability to the truth of the hypothesis
     - Example: $\Pr[H \text{ is } \mathsf{True}] = 60\%$
3. Consider a parameter θ that has a fixed, unknown value
   - A Bayesian can assign a probability distribution to the parameter
     - Example: $p(\theta) = \mathsf{N}(\theta|2, 3)$

⋆ $\mathsf{N}(\mu, \sigma^2)$ denotes the **normal** distribution (also known as **Gaussian**)
   - with **mean** μ and **variance** σ^2

"x is normally distributed": $x \sim \mathsf{N}(\mu, \sigma^2)$

Normal PDF (Probability Density Function) for x

$$p(x) = \mathsf{N}(x|\mu, \sigma^2) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$$

Bent coin example

Probability and information

1. Coin lands concave side up 60% of the time
   - Bayesian and frequentist agree
2. One side of the coin is labeled “heads” and the other side is “tails”
   - You don’t know which
3. What is the probability the coin comes up “heads”?
   - Bayesian says 50%

Bayes' rule: Thomas Bayes (1702–1761)

An Essay towards solving a Problem in the Doctrine of Chances (1763)

1. Data and parameter(s)
   - Data: $y = (y_1, \ldots, y_n)$
   - Parameter(s): $\theta = (\theta_1, \ldots, \theta_d)$
2. **Joint** distribution factored into **conditional** and **marginal** distributions

$$\begin{aligned} p(y,\theta) &= p(y|\theta)\, p(\theta) && \text{(first way)} \\ &= p(\theta|y)\, p(y) && \text{(second way)} \end{aligned}$$

   implies

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)} \propto p(y|\theta)\, p(\theta) \qquad \textbf{(Bayes' rule)}$$

   - Note: $p(y) = \int p(y,\theta)\, d\theta = \int p(y|\theta)\, p(\theta)\, d\theta$
3. Conventional names
   - $p(\theta|y)$ — **posterior distribution**
   - $p(y|\theta)$ — **likelihood** (sample information about θ)
   - $p(\theta)$ — **prior distribution** (non-sample information about θ)
   - $p(y)$ — **marginal likelihood**

Sampling distribution and likelihood

This expression

$$p(y|\theta)$$

has **two uses**

1. **Sampling distribution** (a function of y with θ fixed)
   - for the data y when the parameter θ is known

     $$\int p(y|\theta)\,dy = 1 \qquad \text{(PDF integrates to 1)}$$

   - Used to run things "forwards"
     - Original use of probability (games of chance)
2. **Likelihood** (a function of θ with y fixed)
   - for the unknown parameter θ when the data y are observed

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

   - Used to run things "backwards" for **inverse probability**
     - Bayesian inference

Some history

For example, see McGrayne (2012) *The Theory that Would Not Die*

1. Thomas Bayes
   - Richard Price (edited and presented Bayes' paper)
2. **Pierre-Simon Laplace**
   - Independently discovered and extensively developed "Bayes' rule"
3. John Venn (and others) were unhappy with aspects of Laplace's formulation
4. Ronald A. Fisher (and others) developed alternatives
5. World War II
   - Bayesian methods used to win the war and kept secret after the war
   - Code breaking, the German tank problem, etc
6. Atom bombs and thermonuclear bombs
   - **Markov Chain Monte Carlo** (MCMC) invented to compute integrals
   - Ulam, Metropolis, Teller
7. Image reconstruction
   - Produced the Gibbs sampler
8. Fast(er) computers made recent advances possible (10^9 over 50 years)

Some comments on Bayes' rule

1. Everyone used Bayes' rule
    - Frequentists use Bayes' rule when there is a "genuine" prior distribution
        - "Genuine" means based on observed frequencies
    - Bayesians use Bayes' rule all the time
        - Observed frequencies are great and Bayesians use them when they're available
        - But Bayesians **do not restrict themselves** to only those cases where observed frequencies are available

2. Bayes' rule is about **learning**
    - Prior distribution is transformed into posterior distribution via likelihood
    - Posterior distribution gets used as the prior distribution
        - when more data becomes available

Recursive updating

1. Keep track of the observations: $y_{1:n} = (y_1, \ldots, y_n)$
2. **Likelihood** — given *conditionally* iid observations (conditional on θ)
   - **iid** means "independently and identically distributed"

$$p(y_{1:n}|\theta) = \prod_{i=1}^{n} p(y_i|\theta) \qquad \text{(conditional independence)}$$

3. **Bayes' rule** and its **recursive structure**

$$p(\theta|y_{1:n}) = \underbrace{\frac{\overbrace{p(y_{1:n}|\theta)}^{\text{likelihood}}\,\overbrace{p(\theta)}^{\text{prior}}}{p(y_{1:n})}}_{\text{all data in likelihood}} = \underbrace{\frac{\overbrace{p(y_n|\theta)}^{\text{likelihood}}\,\overbrace{p(\theta|y_{1:n-1})}^{\text{prior}}}{p(y_n|y_{1:n-1})}}_{\text{only new data in likelihood}}$$

where the **prior** $p(\theta|y_{1:n-1})$ is the **posterior** given by

$$p(\theta|y_{1:n-1}) = \frac{p(y_{1:n-1}|\theta)\,p(\theta)}{p(y_{1:n-1})} = \frac{p(y_{n-1}|\theta)\,p(\theta|y_{1:n-2})}{p(y_{n-1}|y_{1:n-2})}$$

and so on

Suppose the "parameter" is a "moving target"

1. **Likelihood**

$$p(y_n|\theta_n) \qquad \text{(observation equation)}$$

   - For each observation y_n there is a different "parameter" θ_n

2. **Transition probability**

$$p(\theta_n|\theta_{n-1}) \qquad \text{(law of motion)}$$

3. **Bayes' rule**

$$\overbrace{p(\theta_n|y_{1:n})}^{\text{posterior for } \theta_n} = \frac{p(y_n|\theta_n)\,\overbrace{p(\theta_n|y_{1:n-1})}^{\text{prior for } \theta_n}}{p(y_n|y_{1:n-1})} \qquad \text{(updating)}$$

   where the **prior** is given by

$$\overbrace{p(\theta_n|y_{1:n-1})}^{\text{prior for } \theta_n} = \int \underbrace{p(\theta_n|\theta_{n-1})}_{\text{law of motion}}\,\overbrace{p(\theta_{n-1}|y_{1:n-1})}^{\text{posterior for } \theta_{n-1}}\,d\theta_{n-1} \qquad \text{(prediction)}$$

4. The **Kalman filter** is a special case
   - When $p(y_n|\theta_n)$ and $p(\theta_n|\theta_{n-1})$ are Gaussian

Relation between sampling and posterior distributions

Sometimes they **appear** to be the **same** (but **they're not**)

1. $y = (y_1, \ldots, y_n)$ where $p(y|\mu, \sigma^2) = \prod_{i=1}^{n} \mathsf{N}(y_i|\mu, \sigma^2)$
   - Assume σ^2 is known
   - Define $\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$
2. **Sampling distribution**
$$\widehat{\mu} \sim \mathsf{N}(\mu, \sigma^2/n)$$
   - $\widehat{\mu}$ is a **test statistic** (a function of the data)
   - If we knew μ (the truth), then we could say where we think $\widehat{\mu}$ (i.e., the data) is likely to be
3. **Posterior distribution**
$$\mu \sim \mathsf{N}(\widehat{\mu}, \sigma^2/n)$$
   - $\widehat{\mu}$ is a **sufficient statistic** (a complete summary of the data)
   - Having seen the data (i.e., $\widehat{\mu}$), we can say where we think μ (the truth) is likely to be
4. **Mathematically**, the two density functions are **equivalent**
$$\underbrace{\mathsf{N}(\widehat{\mu}|\mu, \sigma^2/n)}_{\text{sampling}} \equiv \frac{e^{-\frac{(\mu-\widehat{\mu})^2}{2\,\sigma^2/n}}}{\sqrt{2\,\pi\,\sigma^2/n}} \equiv \underbrace{\mathsf{N}(\mu|\widehat{\mu}, \sigma^2/n)}_{\text{posterior}}$$

Relation between sampling and posterior distributions

Sometimes they **appear** to be quite **different** (as **they are**)

1. "Understanding Unit Rooters: A Helicopter Tour"
   - Sims and Uhlig (1991) *Econometrica*
2. Autoregression

$$y_t = \rho\, y_{t-1} + \varepsilon_t \qquad \varepsilon_t \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2) \qquad \text{(time series)}$$

   - Stationary autoregression $|\rho| < 1$
   - **Unit root** (random walk): $\rho = 1$
3. When ρ is near 1
   - **Sampling distribution** is highly **non-Gaussian**
     - Dickey–Fuller distribution
   - **Posterior distribution** is **Gaussian** (if the prior is flat)
     - follows from the Gaussian likelihood
4. These two distributions are quite different from each other
   - **How is one to choose??!**

What *I* want

1. *I* want to be able to make **probability statements** about
   - where parameters are likely to be
   - which models are more likely
   - where future observations are likely to be
2. The Bayesian method delivers this

3. When I started out I realized: I ***have*** to use a prior
   - The "price of admission"
   - This is a cost
4. Over time, I came to see this differently: I ***get*** to use a prior
   - This is a benefit

5. BTW, frequentists (get to) use priors **implicitly**
   - Sometimes they call it "regularization"
   - Bishop (2006) ***Pattern Recognition and Machine Learning***
     - Bayesian approach provides a "principled framework" for machine learning

Gibbs sampler

1. Let $\theta = (\theta_1, \theta_2)$
2. The **joint** posterior distribution

$$p(\theta|y) = p(\theta_1, \theta_2|y)$$

   (often) can be completely characterized by the two **full conditional** posterior distributions

$$p(\theta_1|y, \theta_2) \qquad \text{and} \qquad p(\theta_2|y, \theta_1)$$

3. Let $\{\theta^{(r)}\}_{r=1}^R = \{(\theta_1^{(r)}, \theta_2^{(r)})\}_{r=1}^R$ denote a **sample** from $p(\theta|y)$
4. Given $\theta^{(r)}$, compute $\theta^{(r+1)}$ as follows

$$\begin{aligned} \theta_1^{(r+1)} &\sim p(\theta_1|y, \theta_2^{(r)}) \\ \theta_2^{(r+1)} &\sim p(\theta_2|y, \theta_1^{(r+1)}) \end{aligned} \qquad \text{(Gibbs sampler)}$$

   - Looks like cheating (it's not)
   - Draws are **not iid** (they're serially dependent)
     - Equivalent number of independent draws is less than R

Rao–Blackwellization

1. You have draws $\{(\theta_1^{(r)}, \theta_2^{(r)})\}_{r=1}^R$ from posterior distribution $p(\theta_1, \theta_2|y)$
2. You want to plot the **marginal distribution** for θ_1
3. You could use a histogram of $\{\theta_1^{(r)}\}_{r=1}^R$
4. Or you could **Rao–Blackwellize**

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)\, d\theta_2 = \int p(\theta_1|y, \theta_2)\, \overbrace{p(\theta_2|y)\, d\theta_2}^{dP(\theta_2|y)}$$
$$\approx \frac{1}{R}\sum_{r=1}^{R} p(\theta_1|y, \theta_2^{(r)})$$

By taking an indirect route you get a smooth approximation

- Using the draws $\{\theta_2^{(r)}\}_{r=1}^R$

5. Note: Follows from Rao–Blackwell theorem

Regression: How to express the model

1. **Traditional way**

$$y_i = \alpha + \beta\, x_i + \varepsilon_i \qquad \text{for } i = 1, \ldots, n$$

where

$$\varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$$

2. **Alternative way** (Bayesians do it this way)

$$p(y|x, \theta) = \prod_{i=1}^{n} p(y_i|x_i, \theta)$$

where $\theta = (\alpha, \beta, \sigma^2)$ and

$$p(y_i|x_i, \theta) = \mathsf{N}(y_i|\underbrace{\alpha + \beta\, x_i}_{\mu_i}, \sigma^2)$$

Outline

Mutual fund data

1. Jones and Shanken (2005)
   - "Mutual fund performance with learning across funds"
2. U.S. Equity funds
   - Number of funds: $n = 5{,}136$
3. Monthly observations, January 1961 to June 2001
   - Returns not available for all funds on all dates
     - Some funds come into existence after beginning
     - Some funds go out of existence before end
   - Minimum 12 observations per fund, mean 77.3 months
   - Total number of observations: 396,820
4. Returns adjusted for risk-free rate, before fees and taxes
5. Four-factor model (Fama–French and Cathcart)
   - EMRF — excess market return
   - SMB — small minus big (market capitalization return)
   - HML — high minus low (book-to-market equity return)
   - MOM — momentum (past one-year)

Factor model for mutual fund returns

1. Returns net of the risk-free rate
   - There are n mutual funds: $Y_{1:n} = (Y_1, \ldots, Y_n)$
   - Fund i has $T_i - \tau_i + 1$ observations: $Y_i = (y_{i\tau_i}, \ldots, y_{iT_i})$
     - **Not a panel** since no requirement that $\tau_i = \tau_j$ or $T_i = T_j$
   - f_t is a vector of **factors** at time t
     - F_i is a matrix of factors aligned with (τ_i, T_i)
2. **Likelihood** (this is just a regression)

$$p(Y_{1:n}|F_{1:n}, \alpha, \beta, \varsigma^2) = \prod_{i=1}^{n} p(Y_i|F_i, \alpha_i, \beta_i, \varsigma_i^2)$$

   where

$$p(Y_i|F_i, \alpha_i, \beta_i, \varsigma_i^2) = \prod_{t=\tau_i}^{T_i} \mathsf{N}(y_{it}|\alpha_i + f_t^\top \beta_i, \varsigma_i^2)$$

   β_i is a vector of factor coefficients for fund i
3. $\alpha_i > 0$ represents **skill** for fund i
   - Question: Which funds display skill and which don't?

Likelihood for skill (equals posterior with a flat prior)

1. α_i is the **parameter of interest** for fund i
2. (β_i, ς_i^2) are **nuisance parameters** — for now
3. **Jeffreys prior** for the nuisance parameters: $p(\beta_i, \varsigma_i^2) \propto 1/\varsigma_i^2$
4. Then (suppressing F_i in the notation)

$$p(Y_i|\alpha_i) = \iint \frac{p(Y_i|F_i, \alpha_i, \beta_i, \varsigma_i^2)}{\varsigma_i^2} d\beta_i d\varsigma_i^2 = \mathsf{Student}(\alpha_i|\widehat{\alpha}_i, \tau_i^2, \nu_i)$$

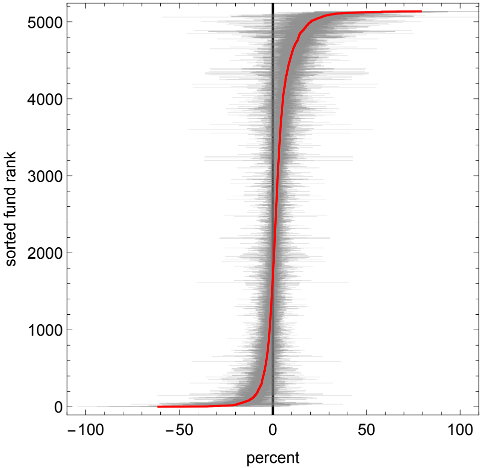
   where $(\widehat{\alpha}_i, \tau_i^2, \nu_i)$
   - depends only on the data (Y_i, F_i)
   - is a **sufficient statistic** for (Y_i, F_i)
5. In particular, $\widehat{\alpha}_i$ is the ordinary least squares (**OLS**) estimate
   - Let $\phi_i = (\alpha_i, \beta_i)$
   - Then $\widehat{\alpha}_i = \widehat{\phi}_{i1}$ where

$$\widehat{\phi}_i = (F_i^\top F_i)^{-1} F_i^\top Y_i$$

6. **Posterior with a flat prior**
   - If $p(\alpha_i) \propto 1$, then $p(\alpha_i|Y_i) = p(Y_i|\alpha_i)$

The data

in the form of the likelihoods $p(Y_i|\alpha_i) = \mathsf{Student}(\alpha_i|\widehat{\alpha}_i, \tau_i^2, \nu_i)$



1. Plot of 90% confidence intervals for α_i, sorted by $\widehat{\alpha}_i$ (shown in red)
2. If $p(\alpha_i) \propto 1$ then these are HPD regions for posterior distributions

Luck and perspicacity

1. **Luck** is in the **likelihood**
   - The likelihood encapsulates the data (the observations)
   - How do we know that a fund manager wasn't just lucky?
2. **Perspicacity** is in the **prior**

   (discernment, keen perception)
   - The prior encapsulates what has been learned from **other sources**
   - If we knew what the distribution of alphas was, then we would be able to better evaluate the likelihood
     - More like **wisdom** than perspicacity, but wisdom doesn't start with a "p"
3. There are $n - 1$ other sources for every fund
   - Jones and Shanken pointed this out
   - They made an important contribution, but they didn't go far enough
4. We seek a **well-informed prior**
   - We must first assemble an **open-minded prior**
     - Jones and Shanken's prior wasn't open-minded
     - they required that skill have a normal distribution
   - An open-minded prior allows for a wide variety of distributions

Density estimation is the key

1. Bayesian density "estimation" amounts to
   - **Computing** a ***predictive*** **distribution**
2. We conduct the analysis in terms of predicting x_{n+1} given $x_{1:n}$
   - First we will let $\underline{x_i = \widehat{\alpha}_i}$, which we **observe**
   - Later we will let $\underline{x_i = \alpha_i}$, which we **do not observe**
3. **"Boilerplate"** (what follows is both *everything* and *nothing*)
   - **Likelihood** for parameters ψ

$$p(x_{1:n}|\psi) = \prod_{i=1}^{n} p(x_i|\psi) \qquad \text{(conditionally independent)}$$

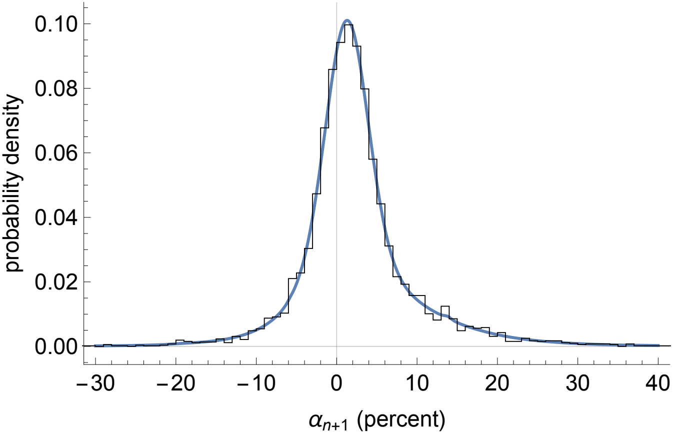
   - **Prior** for the parameters: $p(\psi)$
   - **Posterior** for parameters ψ

$$p(\psi|x_{1:n}) = \frac{p(x_{1:n}|\psi)\, p(\psi)}{p(x_{1:n})} \qquad \text{(Bayes' rule)}$$

   - **Predictive distribution** for **next observation** x_{n+1}

$$p(x_{n+1}|x_{1:n}) = \int p(x_{n+1}|\psi)\, p(\psi|x_{1:n})\, d\psi \approx \frac{1}{R} \sum_{r=1}^{R} p(x_{n+1}|\psi^{(r)}) \qquad (\star)$$

Cross-sectional distribution for $\{\widehat{\alpha}_i\}_{i=1}^n$



- Histogram of $\widehat{\alpha}_{1:n}$ and **predictive distribution** $p(\widehat{\alpha}_{n+1}|\widehat{\alpha}_{1:n})$

Specify the framework for an open-minded prior

equivalent to a Dirichlet Process Mixture (DPM) model

1. **Likelihood** is an **infinite mixture**

$$p(x_i|\psi) = \sum_{c=1}^{\infty} w_c \, f(x_i|\theta_c)$$

where $\psi = (w, \theta)$

- **mixture weights**: $w = (w_1, w_2, \ldots)$ where $w_c \geq 0$ and $\sum_{c=1}^{\infty} w_c = 1$
- **mixture component parameters**: $\theta = (\theta_1, \theta_2, \ldots)$
- ***kernel***

$$f(x_i|\theta_c) = \mathsf{N}(x_i|\mu_c, \sigma_c^2)$$

where $\theta_c = (\mu_c, \sigma_c)$ — **location** and **scale** (mean and std. dev.)

2. **Prior**

- $p(\psi) = p(w, \theta) = p(w)\, p(\theta)$

$$w \sim \mathsf{Stick}(\xi) \qquad \text{(}\textbf{stick-breaking}\text{ distribution)}$$

$$\theta_c \stackrel{\text{iid}}{\sim} H \qquad \text{(}\textbf{base}\text{ distribution)}$$

ξ is the **concentration parameter**

How to cope with an infinite number of components

1. The **number of observations** is **finite**: $n < \infty$
   - Thus: the number of "occupied" mixture components is always finite (and usually far less than n)
2. "Unoccupied" components can be consolidated (by averaging)
   - into a single component with a finite weight
3. Alternatively: **truncate** the sum (make the mixture finite)
   - But make the upper bound large enough to ensure there are always a few "unoccupied" components (2 will do, 5 is good, 10 is plenty)

$$p(x_i|\psi) = \sum_{c=1}^{M} w_c \, f(x_i|\theta_c)$$

M is the upper bound

Specify the framework for an open-minded prior
(continued)

(This material will not be on the test.)

1. Stick-breaking prior

$$w_c = v_c \prod_{\ell=1}^{c-1} (1 - v_\ell) \qquad \text{where } v_c \overset{\text{iid}}{\sim} \mathsf{Beta}(1, \xi)$$

2. Prior for the concentration parameter

$$p(\xi) = \frac{1}{(1+\xi)^2}$$

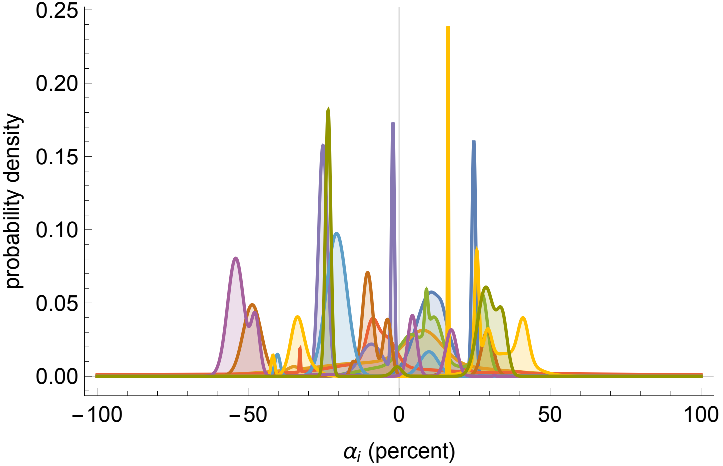
3. Base distribution: $p(\theta_c) = p(\mu_c)\, p(\sigma_c)$

$$p(\mu_c) = \mathsf{N}(\mu_c | 0, s^2) \qquad \text{(Normal)}$$

$$p(\sigma_c) = \frac{(3/A^2)\, \sigma_c}{(1 + (3/A^2)\, \sigma_c^2)^{2/3}} \qquad \text{(Singh–Maddala)}$$

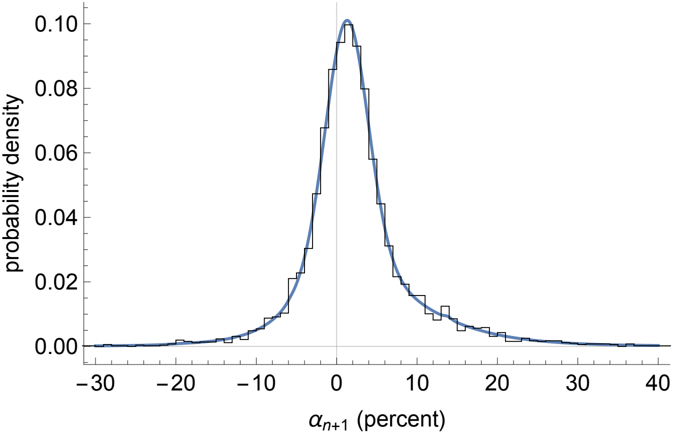
10 draws from the open-minded prior

each draw is a probability distribution



- $p(x_i|\psi)$ plotted for each of ten draws of ψ from $p(\psi)$

Reminder: Cross-sectional distribution for $\{\widehat{\alpha}_i\}_{i=1}^n$



- Histogram of $\widehat{\alpha}_{1:n}$ and predictive distribution $p(\widehat{\alpha}_{n+1}|\widehat{\alpha}_{1:n})$

Latent variable density estimation

because we don't observe the alphas

1. **What we *would* do** if we observed the alphas

$$p(\alpha_{n+1}|\alpha_{1:n}) = \int p(\alpha_{n+1}|\psi)\, p(\psi|\alpha_{1:n})\, d\psi \qquad \text{(predictive)}$$

2. **What we *know*** about the alphas given what we actually observe

$$p(\alpha_{1:n}|Y_{1:n}) \qquad \text{(posterior)}$$

3. Combine **what we would do** with **what we know**

$$p(\alpha_{n+1}|Y_{1:n}) = \int \underbrace{p(\alpha_{n+1}|\alpha_{1:n})}_{\text{predictive}}\, \underbrace{p(\alpha_{1:n}|Y_{1:n})}_{\text{posterior}}\, d\alpha_{1:n} \qquad (\star)$$

- $(\star)$ is **latent variable density estimation**
- it's a **weighted average** of the predictions (what we would do)
- the **weights** come from the posterior distribution (what we know)

A more computationally friendly expression

for latent variable density estimation

1. R draws from $p(\psi|Y_{1:n})$

$$\{\psi^{(r)}\}_{r=1}^{R} = \{(w^{(r)}, \theta^{(r)})\}_{r=1}^{R}$$

   where
   - $w^{(r)} = (w_1^{(r)}, w_2^{(r)}, \ldots, w_M^{(r)})$
   - $\theta^{(r)} = (\theta_1^{(r)}, \theta_2^{(r)}, \ldots, \theta_M^{(r)})$
     Notes
     - M is the upper bound
     - $\theta_c = (\mu_c, \sigma_c^2)$

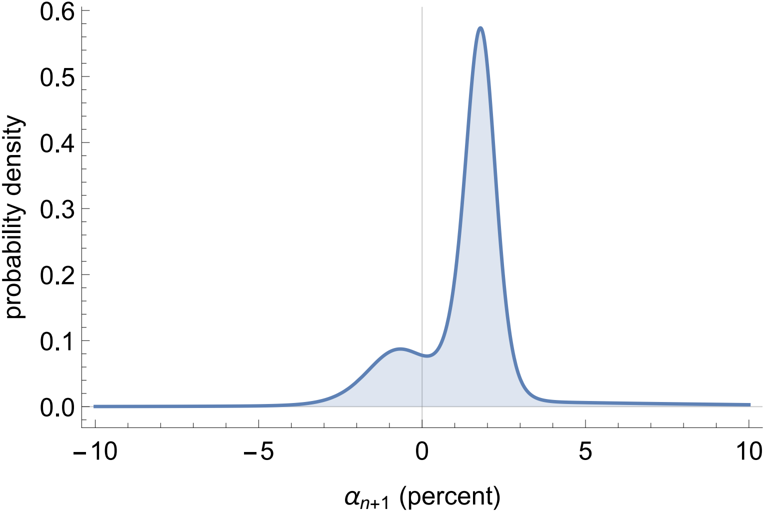
2. A **computationally friendly** expression uses the **posterior for** $\boldsymbol{\psi}$

$$\begin{aligned} p(\alpha_{n+1}|Y_{1:n}) &= \int p(\alpha_{n+1}|\psi)\, p(\psi|Y_{1:n})\, d\psi \\ &\approx \tfrac{1}{R} \textstyle\sum_{r=1}^{R} p(\alpha_{n+1}|\psi^{(r)}) \\ &\approx \tfrac{1}{R} \textstyle\sum_{r=1}^{R} \sum_{c=1}^{M} w_c^{(r)}\, \mathsf{N}(\alpha_{n+1}|\mu_c^{(r)}, \sigma_c^{2\,(r)}) \end{aligned} \qquad (\star)$$

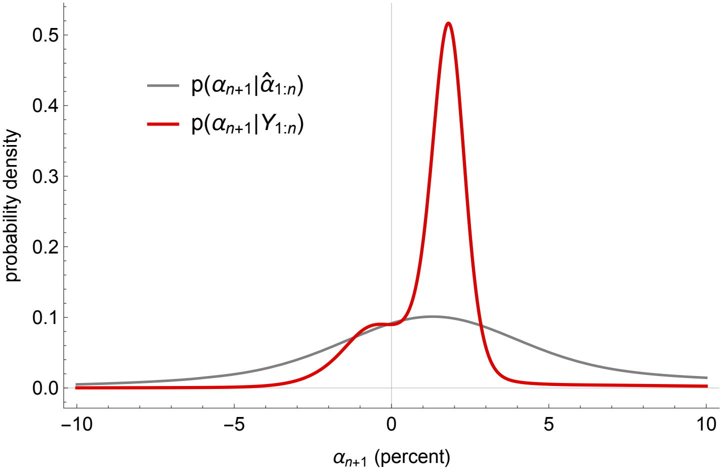
If $R = 1000$ and $M = 20$, then this is a mixture of 20,000 Gaussians

The well-informed prior

computed via latent variable density estimation



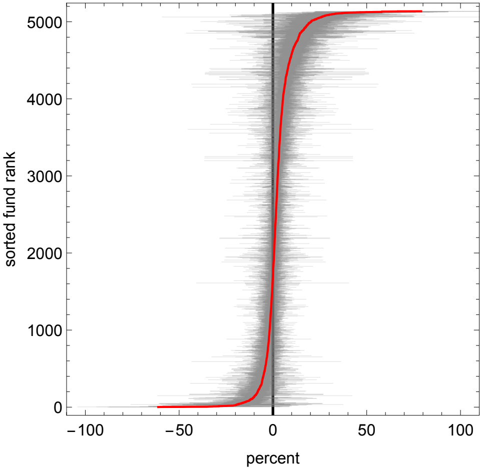
Comparison



- Comparison of well-informed prior $p(\alpha_{n+1}|Y_{1:n})$ with the smoothed histogram $p(\alpha_{n+1}|\widehat{\alpha}_{1:n})$

Remember this? "Before"

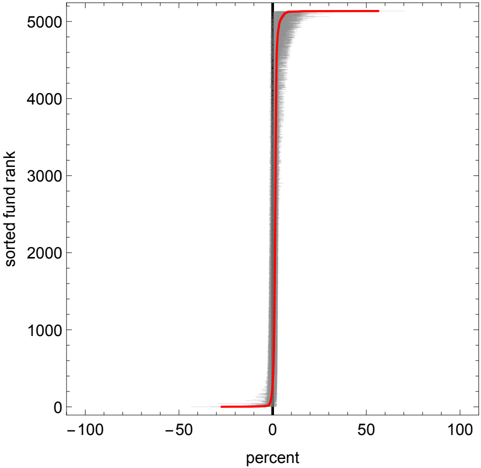
in the form of the likelihoods $p(Y_i|\alpha_i) = \mathsf{Student}(\alpha_i|\widehat{\alpha}_i, \tau_i^2, \nu_i)$



- Plot of 90% confidence intervals for α_i, sorted by $\widehat{\alpha}_i$ (shown in red)

Posterior intervals: “After”

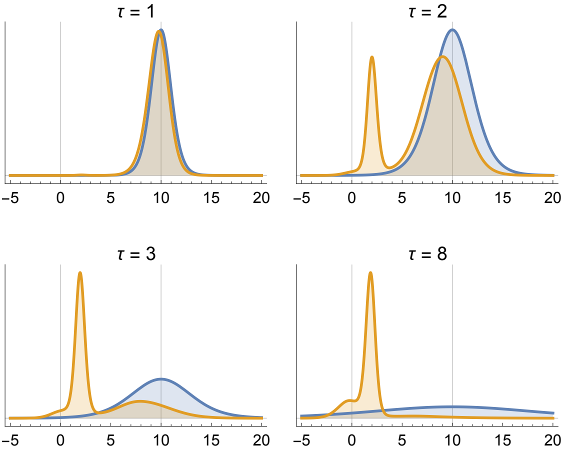
They display shrinkage relative to the likelihoods



- Plot of 90% posterior probability intervals for α_i, sorted by $E[\alpha_i|Y_{1:n}]$ (shown in red)

Vary the likelihood and see how the posterior changes

increase the width of the likelihood (as measured by τ)



- Likelihoods $p(Y_{n+1}|\alpha_{n+1}) = \mathsf{Student}(\alpha_{n+1}|10, \tau_{n+1}^2, 10)$ and associated posterior distributions

Skill related not to fund, but to fund-regime

1. Thus far, we have assumed **skill** was associated with a **fund**
   - But it's probably more like **skill** is associated with a fund **manager**
     - and managers move from one fund to another
   - But even while a manager is at a given fund, he or she may change the investment **strategy** and the **skill** associated with that manager/strategy might change
2. The upshot
   - Without a lot of additional information, we can't be sure which observations from a given fund constitute a **fund-regime**
3. Let's ask the return data (that we already have) to try to sort this out
   - Let the data tell us when the coefficients change
   - We can use a **change-point** model for this

Change-point model

1. Let s_{it} denote the **fund-regime number** for fund i at time t
   - Start numbering at 1 ($s_{i\tau_i} = 1$)
   - Each time there's a change of regime increase s_{it} by 1
2. Let q_i denote the **probability of a regime change** for fund i

$$p(s_{i,t+1} = m' | s_{it} = m, q_i) = \begin{cases} q_i & m' = m + 1 \\ 1 - q_i & m' = m \end{cases}$$

3. **Likelihood** within a fund-regime

$$p(y_{it} | s_{it} = m) = \mathsf{N}(y_{it} | \alpha_{im} + \beta_{im}\, f_t, \varsigma_{im}^2)$$

   - All the parameters ($\alpha_{im}, \beta_{im}, \varsigma_{im}^2$) are regime-dependent
4. Infinite-order mixtures for
   - α_{im}
   - each component of β_{im}
   - ς_{im}^2
   - q_i

Change-point model details

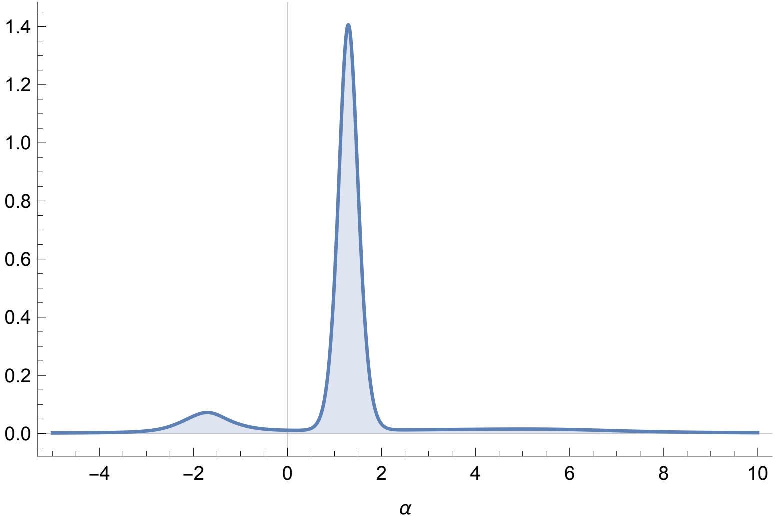
1. Number of regimes is about 12,000
   - Number of funds is about 5,000
   - About 2.4 regimes per fund on average
2. All funds have about the same probability of regime change
   - $q_i \approx .015$

How were the draws (from the posterior) made?

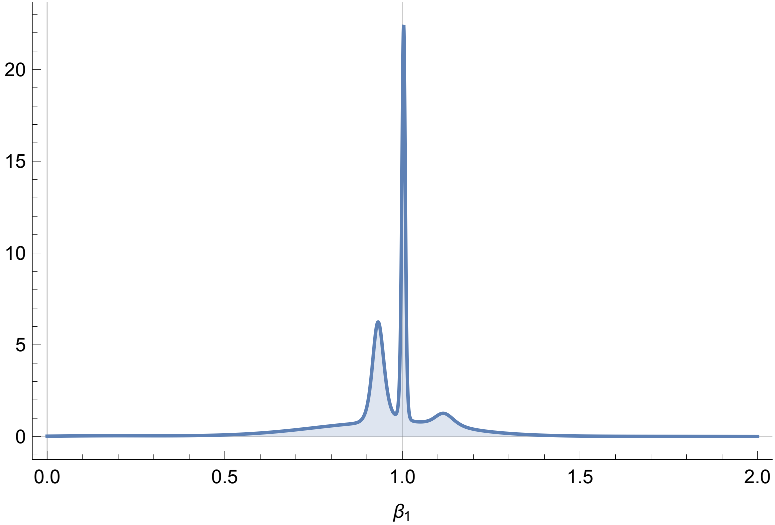
1. Gibbs sampler
   - Mixture models rely on latent classifications
     - $z_i = c$ means fund i is classified with component c
     - etc, etc, etc

   Insert here: **Too Much Information**
2. All calculations in ***Mathematica***
   - Took 330 hours (14 days) using 12 cores
   - Part of the calculation was parallelized
   - Each "draw" took about 10 seconds
   - Made 120,000 draws
     - Discard first 60,000
     - Keep every 6th of the next 60,000 (for 10,000)
     - Use every 10th of those (for 1,000)
3. Why ***Mathematica***?
   - **It's what I know** (and I know it pretty well)
   - I started with Version 2 in early 1990s
   - No packages for sampler: Code is all written by me (hand crafted)
   - Extensive use of `Compile` (to speed things up)

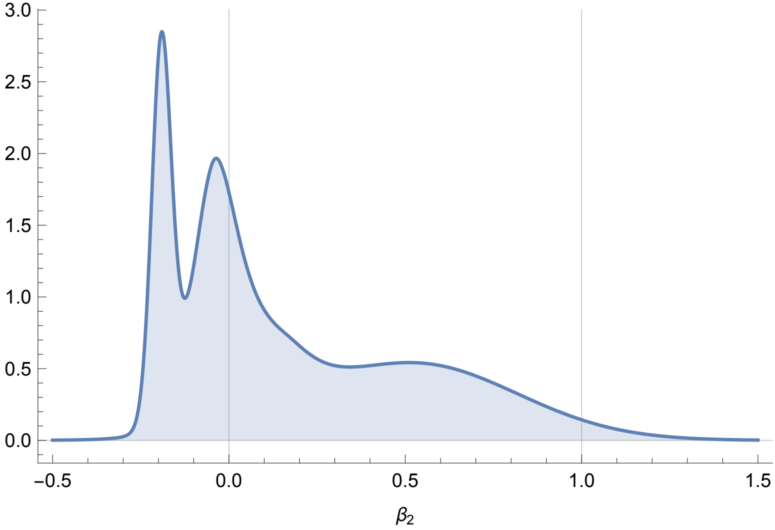
Predictive distribution for alpha



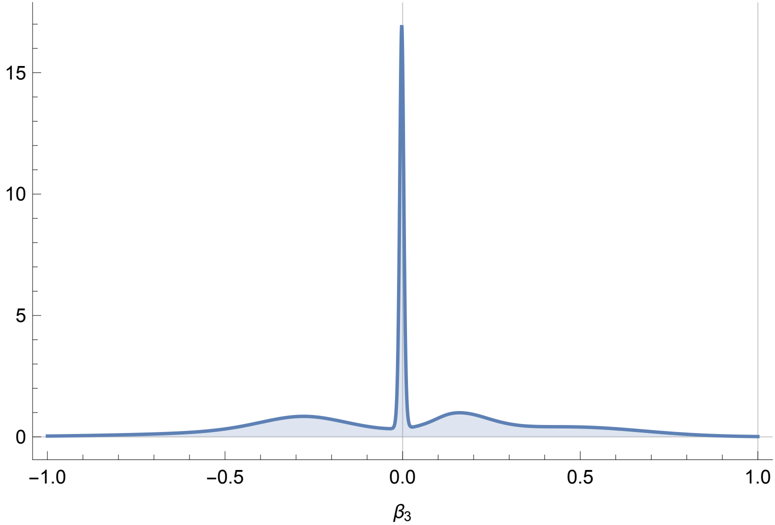
Predictive distribution for beta 1 (market)



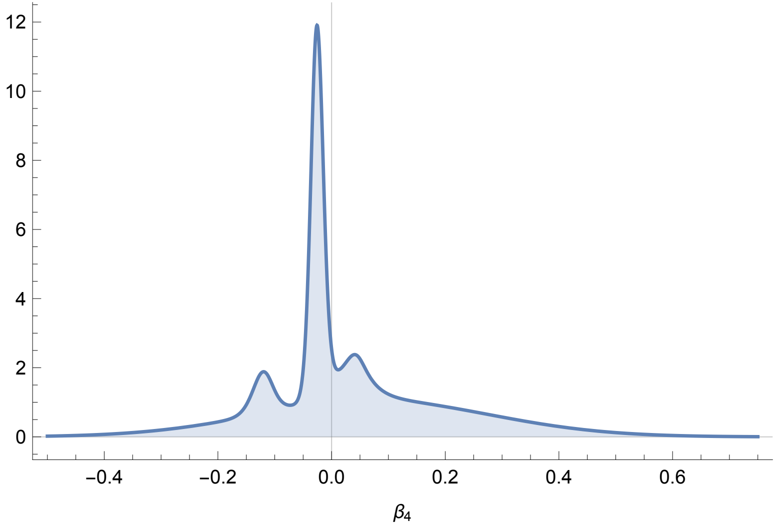
Predictive distribution for beta 2 (SMB)



Predictive distribution for beta 3 (HML)



Predictive distribution for beta 4 (MOM)



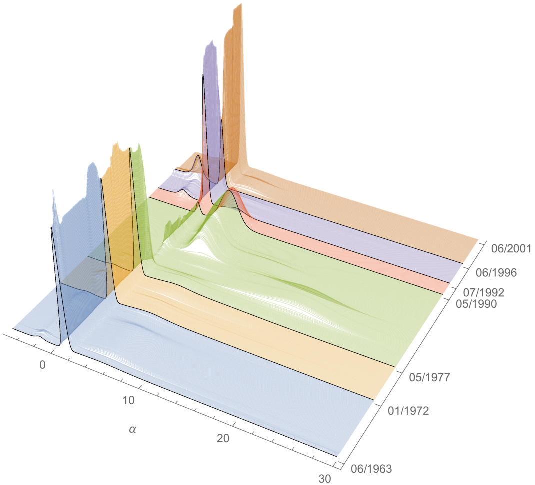
Specific funds

1. The **change-point model**
   - Each fund's **alpha can change** from one month to the next
     - same for each fund's betas (and sigma too)
2. Consequently, there is a **separate posterior distribution**
   - for **each fund-month**
   - for **each parameter**
     (about 2.4 million posterior distributions)
3. But alphas do **not have to change** from one month to the next
   - So the distributions **can be the same** from one month to the next
4. Posterior distribution for α_{it} **given all data**

$$p(\alpha_{it}|Y_{1:n}) = \int p(\alpha_{it}|s_{it})\, p(s_{it}|Y_{1:n})\, ds_{it} \approx \frac{1}{R}\sum_{r=1}^{R} p(\alpha_{it}|s_{it}^{(r)})$$

   - $p(\alpha_{it}|s_{it})$ is the posterior distribution given the **fund-regime number** s_{it}
     - it tells **which observations** in Y_i to use for the **likelihood**
     - it tells **which mixture component** $\mathsf{N}(\alpha_{it}|\mu_c, \sigma_c^2)$ to use for the **prior**

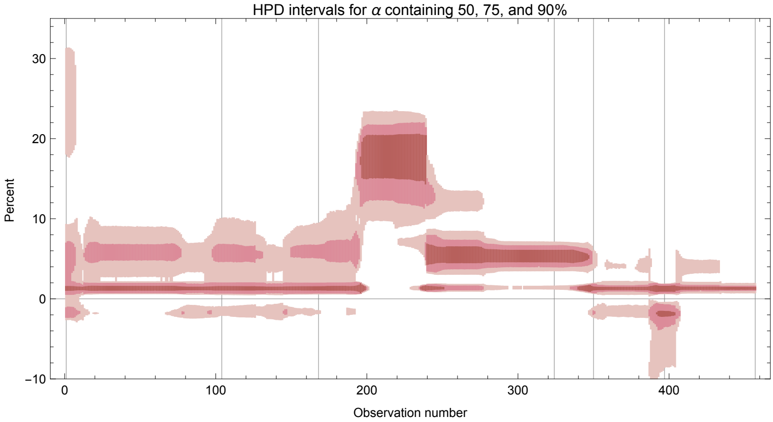
Posterior distributions for α_{it} for Magellan Fund



- Colors indicate managers: Peter Lynch is green

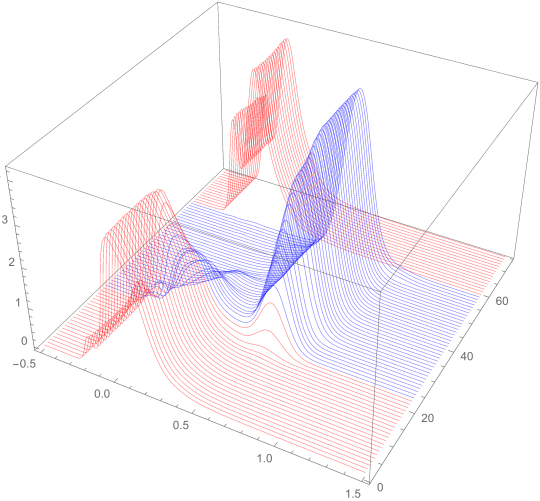
Posterior distributions for α_{it} for Magellan Fund

a different view of the same thing



- Highest Posterior Density (HPD) regions

Predictive distribution for beta 2 for fund 50



The need for machine learning

1. Applied to the **results** of our estimation

Outline

Summary: The Bayesian Method

1. Allows one to
   - combine sample and non-sample information
   - learn
     - across entities, units, regimes
2. Allows one to
   - do optimal signal extraction
   - generate hypotheses for further investigation
     - based on the extracted signals
3. Forces one to
   - confront a realistic assessment of uncertainty
   - think seriously about what one already knows
     - before seeing the new data

4. Final thought: **Decision theory is Bayesian**
   - **Optimal decision** minimizes the expected loss
     - Loss function (depends on the decision and on unknowns)
     - Expectation is computed with respect to the posterior distribution for the unknowns given the observations