Bayes Crash Course



Gregor Kastner

Department of Statistics
University of Klagenfurt
Austria

Outline

Thinking Bayes

Doing Bayes

Numerical Bayes

Contemporary Bayes

Bayesian thinking



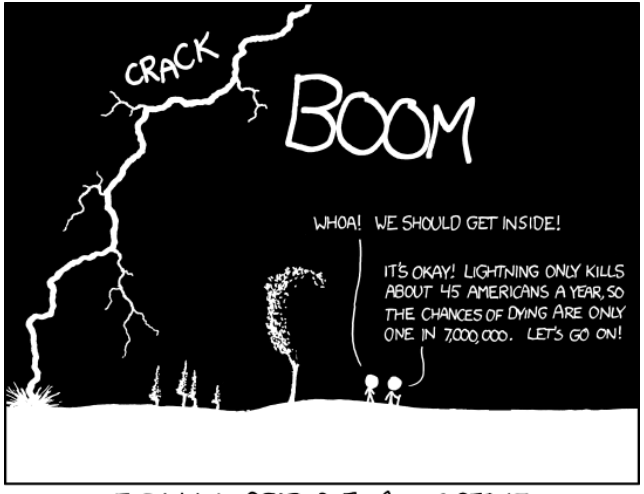
https://lebeninwien.wordpress.com/2010/03/13/studententypen-in-wien-klischee-oder-wahrheit/

Bayesian thinking

Three principles of Bayesian thinking:

- Remember your priors! (vs. base rate neglect)
- Imagine your theory is wrong. Would the world look different? (cf. classical hypothesis testing)
- Update incrementally! (vs. classical hypothesis testing)

Conditional Risk Source



Bayesian thinking is all about getting conditional probabilities: Setosa: Conditional Probability

A first formalization attempt (Bayes, 1763)



Reverend Thomas Bayes (born 1702 in London, died 1761)

$$
\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\
&= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}
\end{aligned}
$$

Seeing Theory: Bayes' Theorem

But how does this relate to statistical and econometric inference?

Outline

Thinking Bayes

Doing Bayes

Numerical Bayes

Contemporary Bayes

Likelihood

Given an iid sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ from a density $p(\boldsymbol{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the likelihood function factorizes into its marginals,

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}).$$

Considered as a function of $\boldsymbol{\theta}$, this is a fundamental entity for the analysis of the information provided about $\boldsymbol{\theta}$ by the sample $\boldsymbol{y}$, and Bayesian analysis (as well as large parts of frequentist inference) analysis relies on it to draw its inference on $\boldsymbol{\theta}$.

Example I: Coin flips

When $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)' \in \{0, 1\}^n$ is an iid sequence of n coin flips, each $y_i \sim \text{Bernoulli}(\pi)$, we get

$$p(\boldsymbol{y}|\pi) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i},$$

or equivalently

$$p(k|\pi, n) = \binom{n}{k} \pi^k (1-\pi)^{n-k},$$

where k is the number of (1)s. Seeing Theory: Likelihood

Typical frequentist approach: Maximum likelihood (ML). Find

$$\arg\max_{\pi} p(k|\pi, n) = \arg\max_{\pi} \log(p(k|\pi, n)),$$

which is given by $\hat{\pi} = k/n$.

Posterior distribution

What's new in a Bayesian analysis? Two components:

- Prior distribution $p(\boldsymbol{\theta})$ (quantifies uncertainty before we look at data)
- Likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ ("updates" the prior)

yield a joint stochastic model

$$p(\boldsymbol{y}, \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Instead of considering the likelihood alone, the Bayes' formula allows to modify it from a probability distribution on the data into a probability distribution on $\boldsymbol{\Theta}$, the posterior:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

In order to do so, a prior $p(\boldsymbol{\theta})$ has to be chosen to start the analysis.

A (very) long discussion in a nutshell...

Pro: We now have a distribution on the thing we are actually interested in!

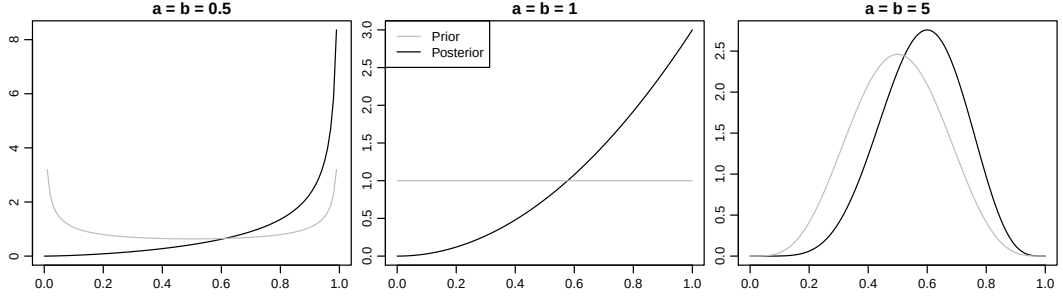
Con: We need a prior! Where is that supposed to come from?

Example I: Posterior

- Likelihood: $k \sim \text{Binomial}(n, \pi)$ Plots
- Prior: $\pi \sim \mathcal{B}(a, b)$ Plots
- Model: $p(\pi, k) = p(k|\pi)p(\pi)$
- Inference is based on the posterior $p(\pi|k)$:

$$\pi|k \sim \mathcal{B}(a + k, b + n - k)$$ Plots

Example with $n = k = 2$: Seeing Theory: Prior to Posterior



Bayesian point estimates

The posterior distribution can be considered as a density estimate of the parameter of interest. Sometimes, though, you want point estimates. The following choices are common:

- Posterior expectation:

$$\hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$$

  Minimizes the quadratic loss $\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2$.
- Posterior median, minimizes the absolute loss $|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{MED}}|$.
- Maximum a posteriori estimator (MAP):

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y})$$

  Minimizes the 0-1 loss $\mathbb{1}(\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}_{\text{MAP}})$. Often recovers the asymptotic properties of the MLE as n increases.

Example I: Point estimates

Recall:

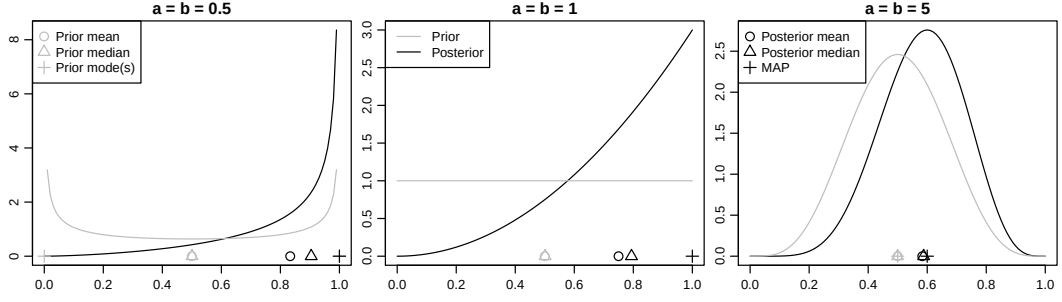
$$\pi|k \sim \mathcal{B}(a+k, b+n-k)$$ Wikipedia: Beta distribution

The posterior mean is

$$\hat{\pi} = \frac{k+a}{n+a+b} = \frac{k}{a+b+n} + \frac{a}{a+b+n}$$

$$\hat{\pi}_{\text{MED}} \approx \frac{k+a-1/3}{n+a+b-2/3}, \quad \hat{\pi}_{\text{MAP}} = \frac{k+a-1}{n+a+b-2} \quad \text{(usually)}$$

Example with $n = k = 2$:



Bayesian credible intervals

To construct "confidence intervals", Bayesians (again) rely on the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. Because $\boldsymbol{\theta}$ is modeled as a random variable, a natural definition of a credible region on $\boldsymbol{\theta}$ is to determine $C(\boldsymbol{y})$ such that

$$p(\boldsymbol{\theta} \in C(\boldsymbol{y})|\boldsymbol{y}) = 1 - \alpha,$$

where α is some predetermined level. Note that integration is done over the parameter space, not the observation space! This means that $1 - \alpha$ corresponds to the probability that a random $\boldsymbol{\theta}$ belongs to this fixed set $C(\boldsymbol{y})$, rather than to the probability that a random set contains a fixed but unknown $\boldsymbol{\theta}$.

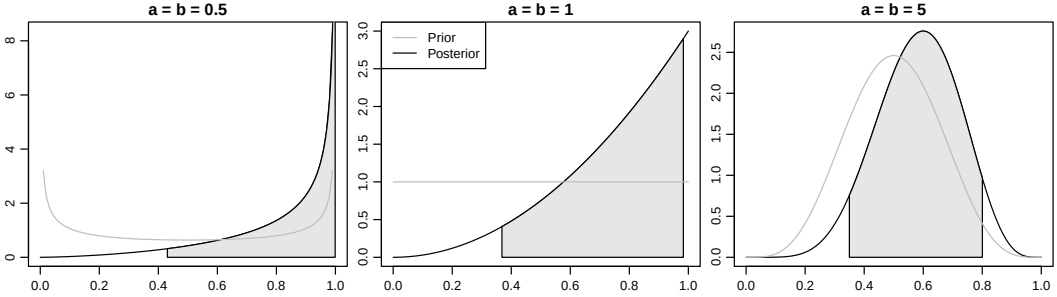
This allows us to safely say things such as:

- "The probability that $\boldsymbol{\theta}$ lies outside this region is α"
- "$C(\boldsymbol{y})$ contains $\boldsymbol{\theta}$ with probability $1 - \alpha$"

which is incorrect in the traditional (frequentist) setup.

Example I: Equal-ended 90% credible intervals

Back to our toy example with $n = k = 2$:

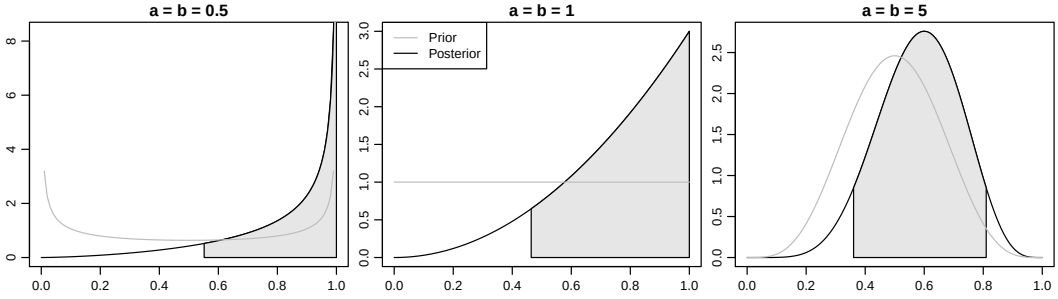


$[0.431, 0.999]$ $[0.368, 0.983]$ $[0.350, 0.800]$

Frequentist: Random interval that covers the true parameter with confidence $1 - \alpha = 0.90$ (for which we observe *one* realization!)

$$p \pm 1.6449 \frac{p(1-p)}{n} = [1, 1] \quad \text{(Ouch!)}$$

Example I: 90% highest posterior density (HPD) intervals



$[0.552, 1.000]$ $[0.464, 1.000]$ $[0.361, 0.810]$

Clopper-Pearson interval: [0.158, 1.000] (Ahhh, better!)

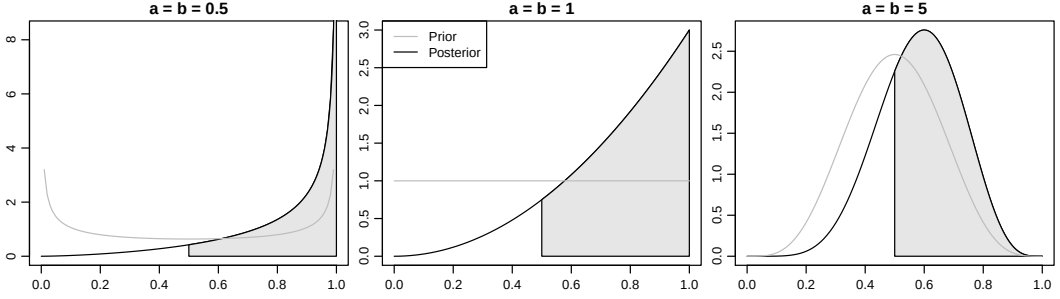
Hypothesis testing

Hypothesis: Coin shows (1) more often than a fair coin.

$$H_0: \quad \pi \leq 1/2$$
$$H_1: \quad \pi > 1/2$$

frequentist: $P(k \geq 2|\pi = 1/2) = P(k = 2|\pi = 1/2) = 0.25 \quad (p\text{-value})$

Bayesian: $P(\pi > 1/2|k) = \ldots$



0.924 0.875 0.726

Two-sided hypothesis testing

Hypothesis: Coin is unfair.

$$H_0: \quad \pi = 1/2$$
$$H_1: \quad \pi \neq 1/2$$

frequentist:
$P(k \geq 2 \text{ or } k \leq 0 | \pi = 1/2) = P(k = 2 \text{ or } k = 0 | \pi = 1/2) = 0.5$ (p-value)

Bayesian: $P(\pi = 1/2 | k) = 0$ (Ouch!)

Ways out:

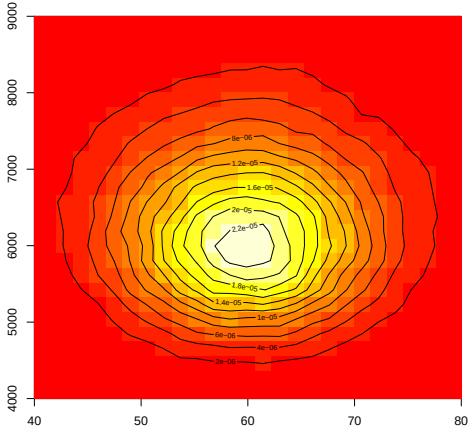
- Compute a posterior $(\alpha - 1)$-interval and check if it includes 0.5
- Use Bayes factors
- Stop thinking in terms of NHST and embrace probability

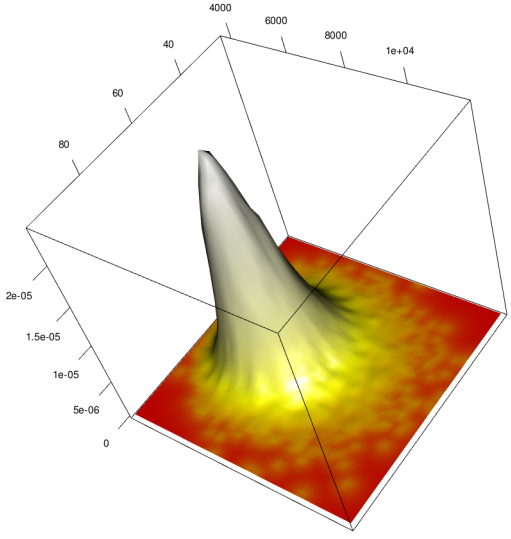
Example II: The normal model with conjugate priors

- Observation model: $\boldsymbol{y}|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$
- Prior distribution: $\mu|\sigma^2 \sim \mathcal{N}(m, M\sigma^2)$ and $\sigma^2 \sim \mathcal{IG}(a, b)$
- Posterior distribution: $p(\mu, \sigma^2|\boldsymbol{y}) = p(\mu|\sigma^2, \boldsymbol{y})p(\sigma^2|\boldsymbol{y})$ with

$$\begin{aligned}
\mu|\sigma^2, \boldsymbol{y} &\sim \mathcal{N}\left(\frac{m + Mn\bar{y}}{Mn+1}, \frac{M\sigma^2}{Mn+1}\right), \\
\sigma^2|\boldsymbol{y} &\sim \mathcal{IG}\left(a + \frac{n}{2}, b + \frac{1}{2}\left(\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2 + \frac{n}{Mn+1}(\bar{y} - m)^2\right)\right).
\end{aligned}$$

Example II: Posterior





Example III: Bayesian Regression

- Observation equation: $\boldsymbol{y}|\beta, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I})$
- Prior distributions: $\beta|\sigma^2 \sim \mathcal{N}(\boldsymbol{b_0}, \sigma^2 \boldsymbol{B_0})$, $\sigma^2 \sim \mathcal{IG}(c_0, C_0)$

According to Bayes formula, the posterior density is given through

$$
\begin{aligned}
p(\beta, \sigma^2|\boldsymbol{y}) &\propto \underbrace{p(\boldsymbol{y}|\beta, \sigma^2)}_{\text{likelihood}} \underbrace{p(\beta, \sigma^2)}_{\text{prior}} = p(\boldsymbol{y}|\beta, \sigma^2) p(\beta|\sigma^2) p(\sigma^2) \\
&\propto \left(\frac{1}{\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)\right\} \times \\
&\quad \left(\frac{1}{\sigma}\right)^p \exp\left\{-\frac{1}{2\sigma^2}(\beta - \boldsymbol{b_0})' \boldsymbol{B_0}^{-1}(\beta - \boldsymbol{b_0})\right\} \times \\
&\quad \left(\frac{1}{\sigma^2}\right)^{c_0+1} \exp\left\{-\frac{C_0}{\sigma^2}\right\}
\end{aligned}
$$

Example III: Bayesian Regression

The posterior $\beta, \sigma^2 | \boldsymbol{y}$ follows a "normal-inverse-gamma" distribution, for which it can be shown that

$$\hat{\beta}_{\text{Bayes}} := \mathbb{E}[\beta | \boldsymbol{y}] = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B_0}^{-1})^{-1}(\boldsymbol{B_0}^{-1}\boldsymbol{b_0} + \boldsymbol{X}'\boldsymbol{y})$$

Setting $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{B_0}^{-1})^{-1}\boldsymbol{X}'\boldsymbol{X}$, we can interpret $\hat{\beta}_{\text{Bayes}}$ as the weighted mean of prior expectation $\boldsymbol{b_0}$ and OLS estimate $\hat{\beta}_{\text{OLS}}$:

$$\hat{\beta}_{\text{Bayes}} = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{b_0} + \boldsymbol{A}\hat{\beta}_{\text{OLS}}$$

Note that when the diagonal elements of $\boldsymbol{B_0}$ are large, $\boldsymbol{A}$ approaches $\boldsymbol{I}$ and thus $\hat{\beta}_{\text{Bayes}}$ approaches $\hat{\beta}_{\text{OLS}}$. Vice versa, when $\boldsymbol{B_0}$ has small diagonal elements, $\boldsymbol{A}$ approaches $\boldsymbol{0}$, thus $\hat{\beta}_{\text{Bayes}}$ approaches $\boldsymbol{b_0}$.

Equivalent to Ridge regression, the minimizer of $||\boldsymbol{y} - \boldsymbol{X}\beta||_2 + \lambda||\beta||_2$ when $\boldsymbol{b_0} = \boldsymbol{0}$ and $\boldsymbol{B_0}^{-1} = \lambda\boldsymbol{I}$.

A Bayesian predictive framework

Within a Bayesian framework, a natural way of assessing the predictive performance of a given model is through its *predictive density* (sometimes also referred to as *posterior predictive distribution*). It is given through

$$p(\boldsymbol{y}_{t+1}|\boldsymbol{y}^o_{[1:t]}) = \int_{\boldsymbol{\kappa}} p(\boldsymbol{y}_{t+1}|\boldsymbol{y}^o_{[1:t]}, \boldsymbol{\kappa}) \times p(\boldsymbol{\kappa}|\boldsymbol{y}^o_{[1:t]})\, d\boldsymbol{\kappa}, \tag{1}$$

where $\boldsymbol{\kappa}$ denotes the vector of all unobservables, i.e., parameters and possible latent variables.

Univariate predictive marginals for 20 exchange rates | Bivariate predictive marginals for 20 exchange rates

Bayesian take-aways

- Bayesian analysis is a proper probabilistic inferential assessment conditional on the data $\boldsymbol{y}$.
- It puts an end to the prosecutor's fallacy[1] (conducted on a regular basis when interpreting p-values).
- No more need for tongue-twisting: "Given that the null hypothesis is true, the probability of observing this or something more extreme is ..."
- Instead, posterior probabilities allow you to say things like "The probability that β is larger than 0 is ...".
- Priors must be specified!
- Closed-form solutions such as Example I are rare.

[1] https://en.wikipedia.org/wiki/Sally_Clark

Outline

Thinking Bayes

Doing Bayes

Numerical Bayes

Contemporary Bayes

Markov chain Monte Carlo (MCMC)

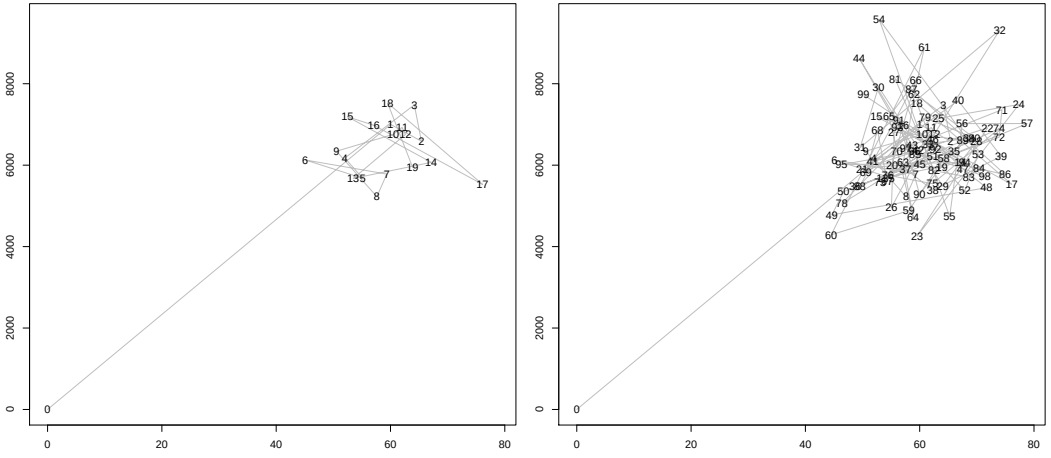
Suppose you can decompose the parameter vector into components $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$ for which the full conditional distributions are available.

Gibbs sampler: Choose starting points $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_p^{(0)}$ and iterate the following steps for $s = 1, \ldots, S$:

- Sample $\boldsymbol{\theta}_1^{(s)} \sim p(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2^{(s-1)}, \ldots, \boldsymbol{\theta}_p^{(s-1)}, \boldsymbol{y})$
- Sample $\boldsymbol{\theta}_2^{(s)} \sim p(\boldsymbol{\theta}_2 | \boldsymbol{\theta}_1^{(s)}, \boldsymbol{\theta}_3^{(s-1)}, \ldots, \boldsymbol{\theta}_p^{(s-1)}, \boldsymbol{y})$
- ...
- Sample $\boldsymbol{\theta}_p^{(s)} \sim p(\boldsymbol{\theta}_p | \boldsymbol{\theta}_1^{(s)}, \ldots, \boldsymbol{\theta}_{p-1}^{(s)}, \boldsymbol{y})$

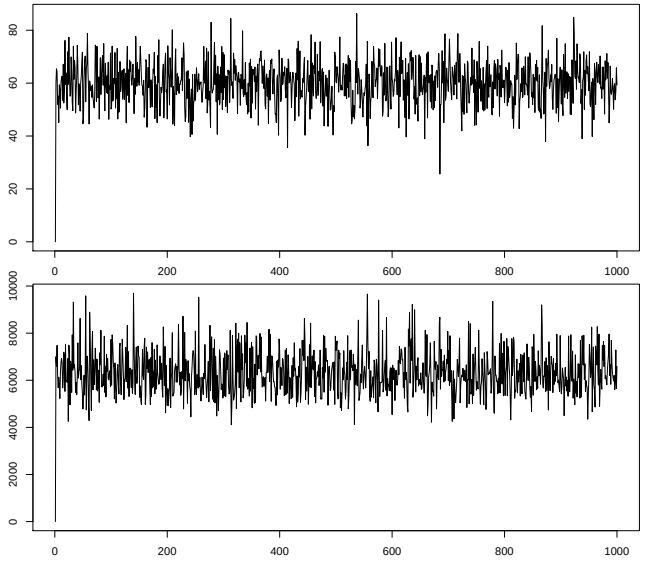
This generates a dependent sequence of draws $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)}$ whose sampling distribution approaches the posterior distribution as $s \to \infty$ (under some regularity conditions).

Example II: Draws from the Gibbs sampler I



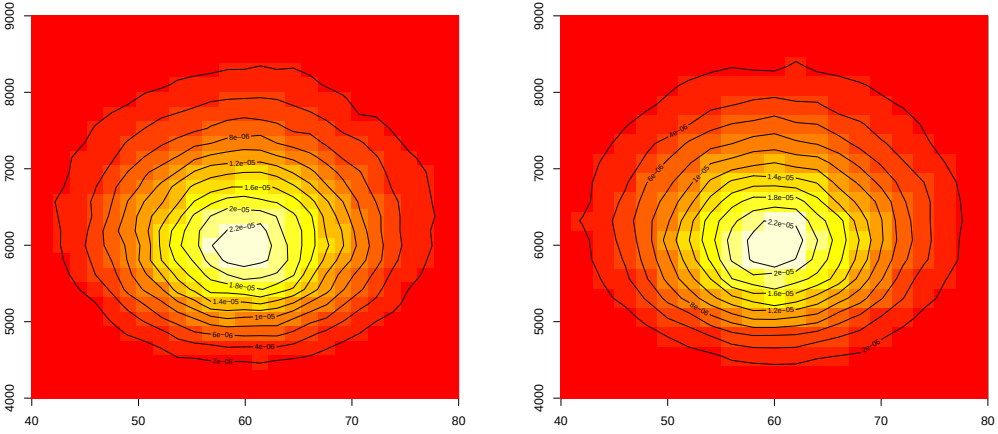
Starting value and 19 (left) resp. 99 (right) draws from the posterior via the Gibbs sampler.

Example II: Draws from the Gibbs sampler II



Trace plots obtained from the full conditional Gibbs sampler.

Example II: Draws from the Gibbs sampler III



Kernel density estimates: Direct sampler (left), Gibbs sampler (right).

Example III: Another solution

Recall: In the regression example, the posterior $\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}$ follows a "normal-inverse-gamma" distribution.

What to do if you don't speak normal-inverse-gamma-ish, or you want to find a more flexible way of learning about the posterior distribution?

$$\Downarrow$$

Use the Gibbs-sampler to "surf the posterior" by alternately simulating values from the (full) conditional parameter densities $\boldsymbol{\beta} | \boldsymbol{y}, \sigma^2$ and $\sigma^2 | \boldsymbol{y}, \boldsymbol{\beta}$.

Example III: Hands on!

- $\beta|\boldsymbol{y}, \sigma^2 \sim \mathcal{N}(\boldsymbol{b_n}, \boldsymbol{B_n})$ with

$$\begin{aligned} \boldsymbol{B_n} &= \left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{B_0}^{-1}\right)^{-1} \\ \boldsymbol{b_n} &= \boldsymbol{B_n}\left(\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{y} + \frac{1}{\sigma^2}\boldsymbol{B_0}^{-1}\boldsymbol{b_0}\right) \end{aligned}$$

- $\sigma^2|\boldsymbol{y}, \beta \sim \mathcal{IG}(c_n, C_n)$ with

$$\begin{aligned} c_n &= c_0 + \frac{n}{2} + \frac{p}{2} \\ C_n &= C_0 + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta) + \frac{1}{2}(\beta - \boldsymbol{b_0})'\boldsymbol{B_0}^{-1}(\beta - \boldsymbol{b_0}) \end{aligned}$$

Outline

Thinking Bayes

Doing Bayes

Numerical Bayes

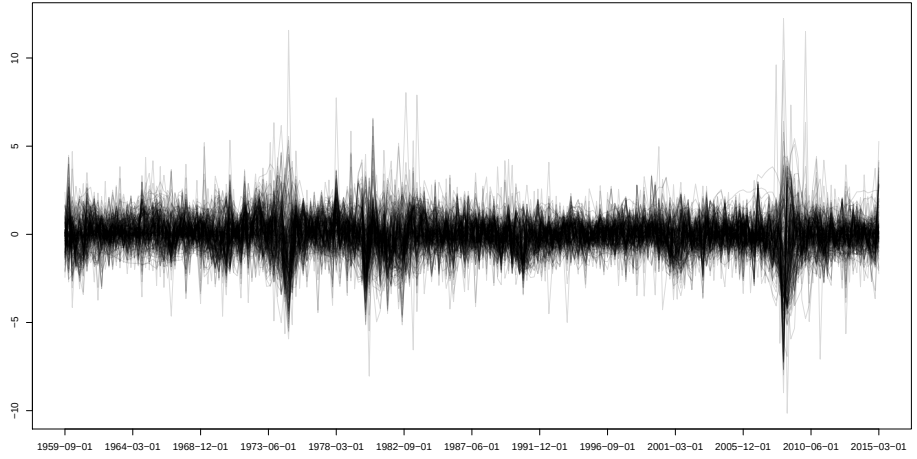
Contemporary Bayes

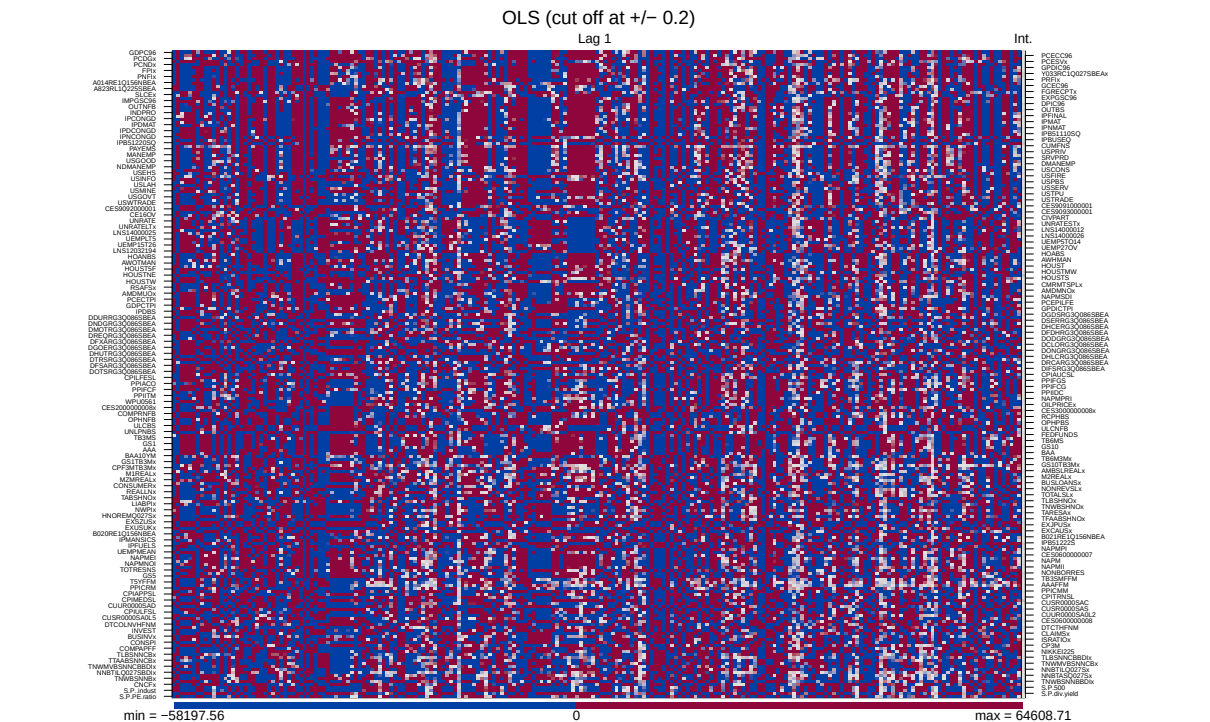
High dimensional application I

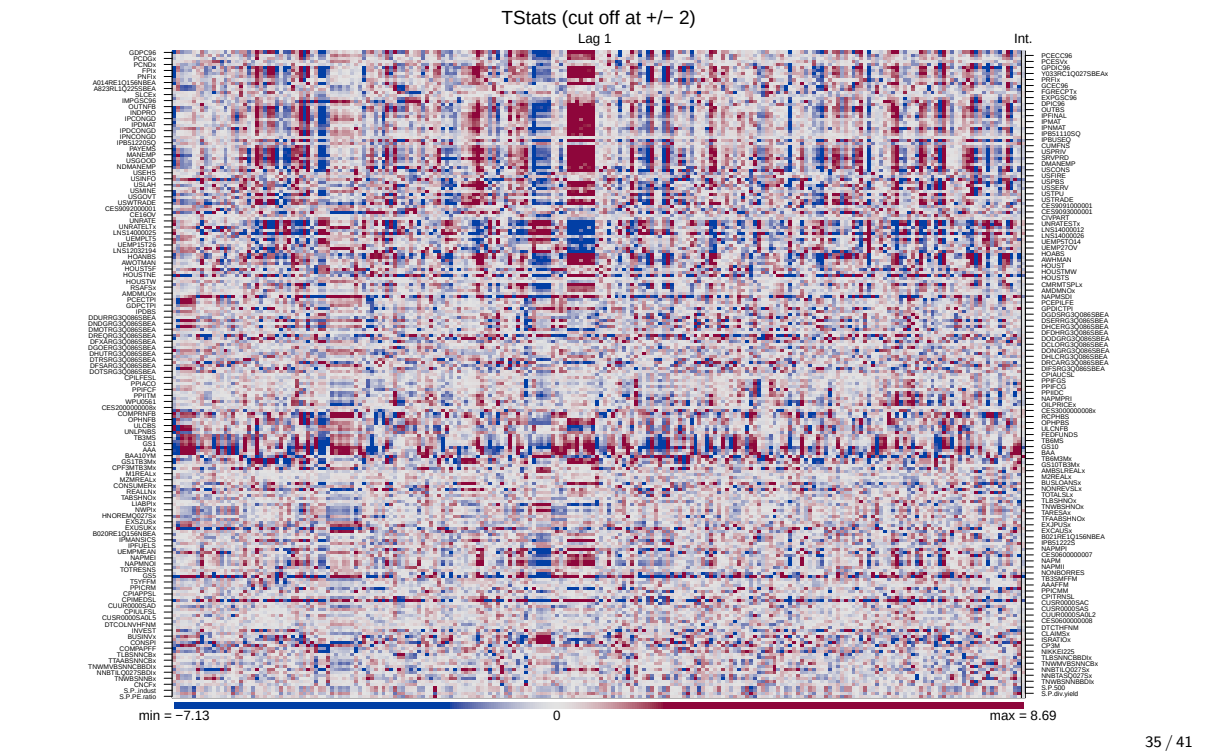
Kastner and Huber (2020) use sparse heteroskedastic VARs,

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim \mathcal{N}_m(\boldsymbol{0}, \Omega_t),$$

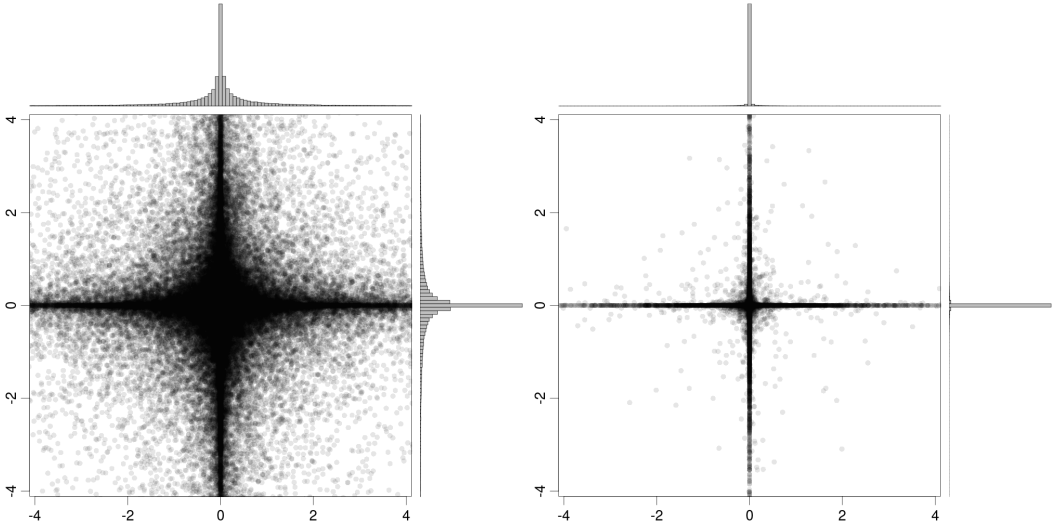
to model the 215-dimensional quarterly dataset from McCracken and Ng (2016).

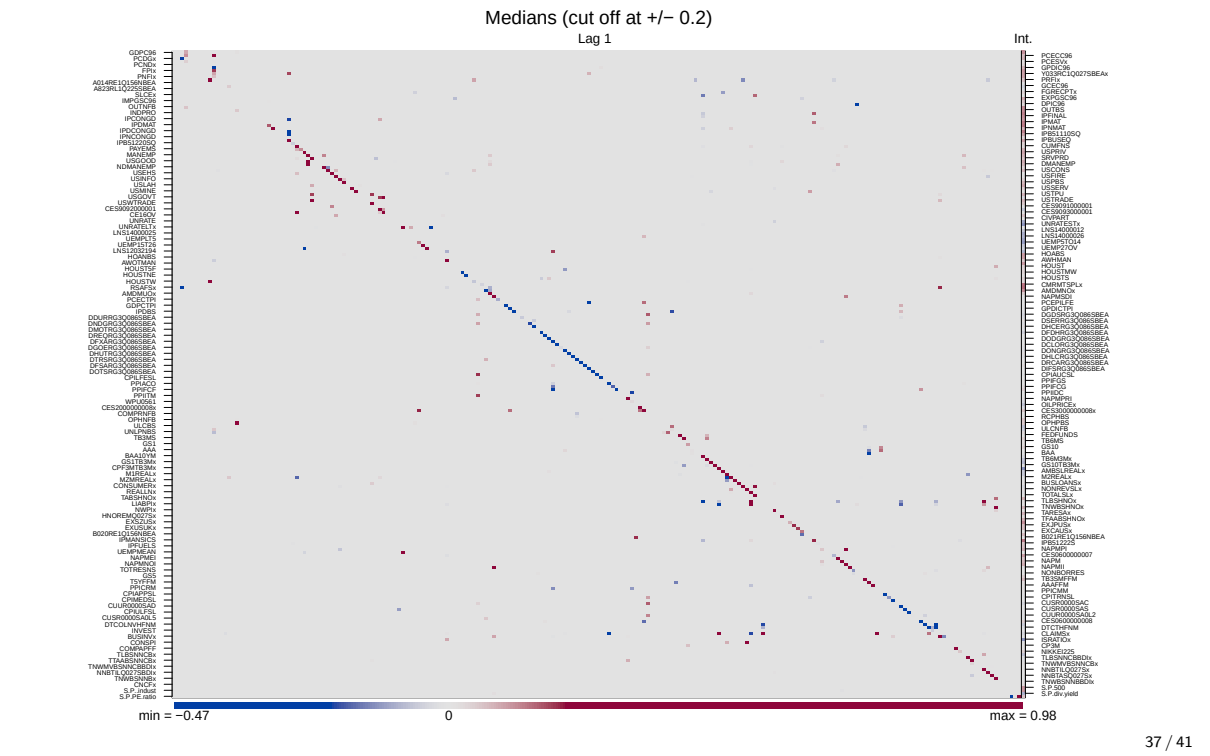



OLS (cut off at +/− 0.2)
Lag 1
Int.
min = −58197.56
0
max = 64608.71


TStats (cut off at +/− 2)
Lag 1
Int.
min = −7.13
0
max = 8.69

Shrinkage priors to the rescue (talk to Luis for more on this)




Medians (cut off at +/− 0.2)
Lag 1
Int.
min = −0.47
0
max = 0.98

High dimensional application II

Kastner (2019) uses $m + r$ conditionally independent SV volatility models,

$$h_{it} = \mu_i + \phi_i(h_{i,t-1} - \mu_i) + \sigma_i \eta_{it}, \qquad \eta_{it} \sim \mathcal{N}(0, 1)\,,$$

to define time-varying diagonal matrices

$$\boldsymbol{U}_t = \text{Diag}(e^{h_{1,t}}, \ldots, e^{h_{m,t}}), \quad \boldsymbol{V}_t = \text{Diag}(e^{h_{m+1,t}}, \ldots, e^{h_{m+r,t}}).$$

Model m-variate time series $\boldsymbol{y}_t$ through the factor model:

$$\begin{aligned} \boldsymbol{f}_t &\sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{V}_t), \\ \boldsymbol{y}_t &= \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{U}_t). \end{aligned}$$

This yields a sparse representation of a time-varying covariance matrix $\boldsymbol{\Sigma}_t$:

$$\boldsymbol{y}_t \sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{\Sigma}_t), \quad \boldsymbol{\Sigma}_t = \boldsymbol{\Lambda} \boldsymbol{V}_t \boldsymbol{\Lambda}' + \boldsymbol{U}_t.$$

Correlation matrix of 300 S&P 500 members

Summary

- Bayesian modeling is a viable alternative to the frequentist ansatz
- Priors allow to incorporate previous knowledge
- Priors allow for regularization
- Inference is usually done with computers (simulation-based methods)
- Thanks to MCMC, complex latent variable models can be estimated relatively "easy"
- Unified and coherent decision theoretic framework for econometric inference, uncertainty quantification, prediction, testing, ...

Literature

Introduction:

- Hoff, Peter D.: *A First Course in Bayesian Statistical Methods*. Springer, 2010.
- Marin, Jean-Michel & Christian P. Robert: *Bayesian Essentials with R*. Springer, 2nd edition, 2014.

The "Classics":

- Lancaster, Tony: *An Introduction to Modern Bayesian Econometrics*. Blackwell, 2004.
- Koop, Gary: *Bayesian Econometrics*. Wiley, 2006.

Additional Source:

- Blake, Andrew & Mumtaz Haroon: *Applied Bayesian econometrics for central bankers (updated)*. 2017. URL: `https://www.bankofengland.co.uk/ccbs/applied-bayesian-econometrics-for-central-bankers-updated-2017`

Bayes, T. (1763). *An Essay towards solving a problem in the doctrine of chances*. Reprinted in *Biometrika 45*, 1958, 296–315.

Kastner, G. (2019). "Sparse Bayesian time-varying covariance estimation in many dimensions". In: *Journal of Econometrics* 210, pp. 98–115.

Kastner, G. and F. Huber (2020). "Sparse Bayesian vector autoregressions in huge dimensions". In: *Journal of Forecasting*. DOI: 10.1002/for.2680.

McCracken, M. W. and S. Ng (2016). "FRED-MD: A Monthly Database for Macroeconomic Research". In: *Journal of Business & Economic Statistics* 34.4, pp. 574–589. DOI: 10.1080/07350015.2015.1086655.