Bayesian Statistics
What it is and what it can do for you

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Overview

- **The Bayesics**
- **Bayesian prediction**
- **Bayesian model inference**
- **Smoothness priors**

Bernoulli trials - frequentist

- **Data**: n trials with binary outcomes: $X_1, ..., X_n$.
  - 0 = head in coin flip. 1 = tails.
  - 0 = no rain in Tokyo. 1 = rain in Tokyo.
- **Population parameter**

$$\theta = \Pr(X = 1)$$

- θ is a fixed constant.
- **Unbiased estimator** $\hat{\theta} = s/n$. s = number of successes ($X_i = 1$). Tokyo: $\hat{\theta} = 95/365 \approx 0.247$.
- $\hat{\theta}$ varies from sample to sample. **Sampling distribution**.
- **Confidence interval**: random interval $[a, b]$ such that the true θ belongs to the interval in 95% of all possible samples of size n. Tokyo: $[0.215, 0.305]$.
- **Hypothesis test**: $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ based on the test statistic $z = (\hat{\theta} - \theta_0)/SD(\hat{\theta})$.

Bernoulli trials - Bayesian

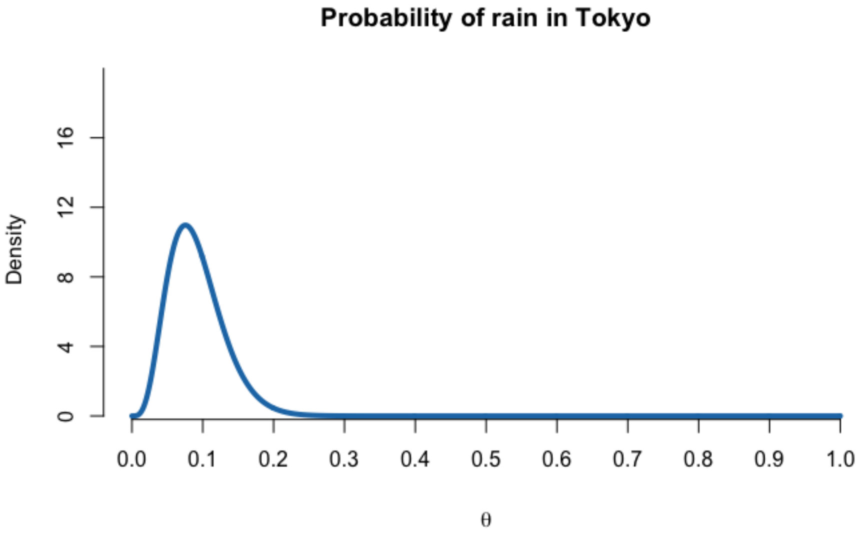
- θ may be fixed, but it is **unknown to me**. I should describe my **uncertainty** about θ in the form of a **probability distribution**.
- Probability is **subjective degree of belief.**
- **Learning from data**: given a **prior** distribution, $p(\theta)$, how do we **update** to a **posterior distribution** $p(\theta|\text{data})$?
- **Bayes theorem** for events

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

- Example:

$$p(\text{cancer}|\text{positive test}) = \frac{p(\text{positive test}|\text{cancer})p(\text{cancer})}{p(\text{positive test})}$$

Prior distribution



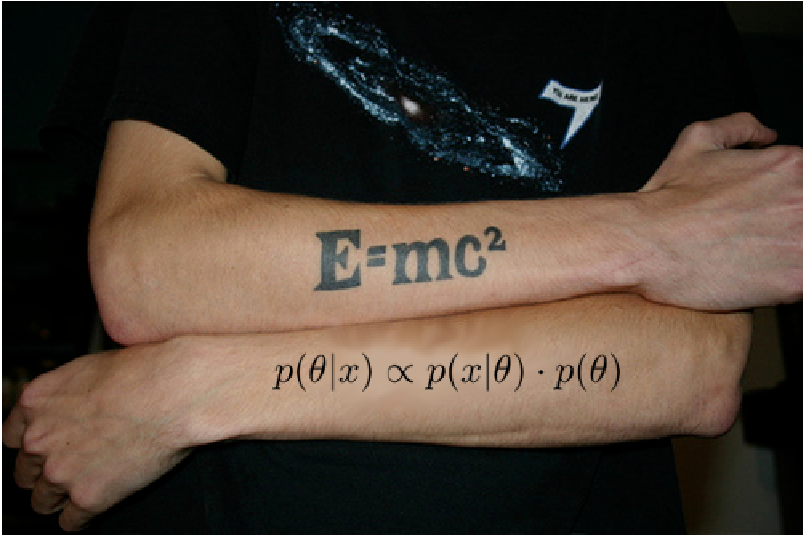
Bernoulli trials - Bayesian

- **Bayes theorem** for a continuous parameter

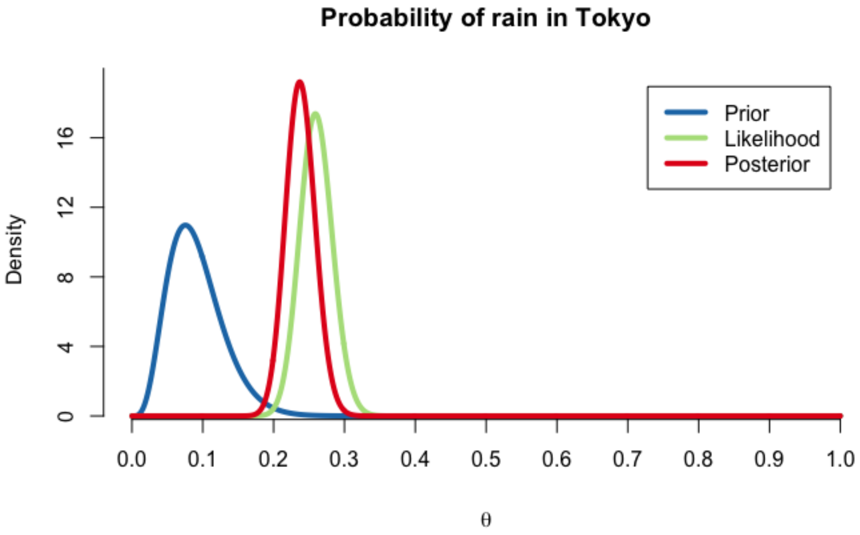
$$\underbrace{p(\theta|x_1,...,x_n)}_{\text{posterior}} = \frac{\overbrace{p(x_1,...,x_n|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(x_1,...,x_n)}_{\text{marginal likelihood}}}$$

- **Bayesian updating** in Bernoulli trials:
  - Prior: $\theta \sim \text{Beta}(\alpha, \beta)$
  - Likelihood: $\theta^s(1-\theta)^f$
  - Posterior: $\theta|x_1,..,x_n \sim \text{Beta}(\alpha+s, \beta+f)$

Great theorems make great tattoos



Bayesian analysis of rain in Tokyo



Bayesian analysis of rain in Tokyo

- The posterior is a probability distribution. We can compute probabilities by integration

$$\Pr(\theta < 0.2 | x_1, .., x_n) = 0.03$$

- In R: `pbeta(0.2, shape1 = alpha + s, shape2 = beta + n-s)`
- Bayesian **95% credible interval**

$$[0.199, 0.280]$$

- Direct probabilistic interpretation!

$$\Pr(\theta \in [0.199, 0.280] | x_1, ..., x_n) = 0.95$$

.

- In R:
  - `qbeta(0.025, shape1 = alpha + s, shape2 = beta + n-s)`
  - `qbeta(0.975, shape1 = alpha + s, shape2 = beta + n-s)`

Conjugate priors

- Previous example was nice: prior and posterior were both Beta distributions.
- Beta **prior is conjugate** to a Bernoulli model.
- Normal prior is conjugate to Normal model.
- Gamma prior is conjugate to Poisson model (count data).

Normal approximation for "large" datasets

- What if the model does not have a conjugate prior?
- Theorem: the **posterior** distribution will be a **normal distribution** in **large datasets**, for **any** prior

$$\theta|x_1, ..., x_n \overset{approx}{\sim} N(\hat{\theta}, \Sigma) \text{ for large } n$$

- $\hat{\theta}$ and Σ can be obtained by **numerical optimization**
- `optim` in R or `fminunc` in Matlab.
- Just need to code up the likelihood and the prior.

Approximate posterior by simulation

- **Fast computers + simulation algorithms = Bayes popular**.
- **Markov Chain Monte Carlo** (**MCMC**) for general problems.
- **Sequential Monte Carlo** (**SMC**) for sequential (time-series) problems.
- **Integrated Nested Laplace Approximation** (**INLA**) for spatio-temporal problems.
- **Approximate Bayesian Computation** (**ABC**) when it is easy simulate data from the model, but hard to write down its probability distribution.

Bayesian prediction

- **Predicting** the observation tomorrow x_{n+1} given observations up to today: $x_1, ..., x_n$

$$\underbrace{p(x_{n+1}|x_1, ..., x_n)}_{\text{predictive distribution}} = \int \underbrace{p(x_{n+1}|\theta)}_{\text{model}} \underbrace{p(\theta|x_1, ..., x_n)}_{\text{posterior}} d\theta$$

- Obtaining the **predictive distribution by simulation**:
  1. Simulate a $\theta^\star$ from $p(\theta|x_1, ..., x_n)$
  2. Simulate tomorrow's value x_{n+1} from the model $p(x_{n+1}|\theta^\star)$
  3. Repeat Step 1 and 2 many times.
- Predictive distribution includes **three sources of uncertainty**:
  - Intrinsic model shocks/disturbances (Step 2)
  - Parameter uncertainty (Step 1)
  - Model uncertainty (explained next)

Bayesian model inference

- We usually entertain more than one model

$$
\begin{aligned}
\mathrm{M}_1 &: \; p_1(x|\theta_1) \\
\mathrm{M}_2 &: \; p_2(x|\theta_2)
\end{aligned}
$$

- Example 1:

$$
\begin{aligned}
\mathrm{M}_1 &: \; x \sim N(\theta_1, 1) \\
\mathrm{M}_2 &: \; x \sim t_1(\theta_2, 1)
\end{aligned}
$$

- Example 2:

$$
\begin{aligned}
\mathrm{M}_1 &: \; y = \alpha_1 + \beta_1 x + \epsilon \\
\mathrm{M}_2 &: \; y = \alpha_2 + \beta_2 x + \gamma_2 z + \epsilon
\end{aligned}
$$

- Example 3:

$$
\begin{aligned}
\mathrm{M}_1 &: \; x \sim \text{Bernoulli}(\theta) \\
\mathrm{M}_2 &: \; x \sim \text{Bernoulli}(0.5)
\end{aligned}
$$

Bayesian model inference

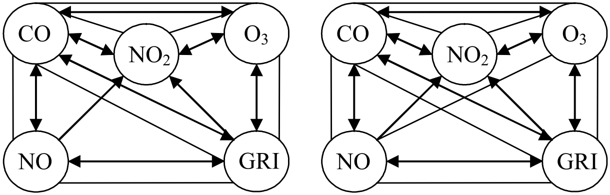
- Bayesian **posterior model distribution**

$$\underbrace{\Pr(M_k|x_1,...,x_n)}_{\text{posterior probability}} \propto \underbrace{p(x_1,...,x_n|M_k)}_{\text{marginal likelihood}} \quad \cdot \quad \underbrace{\Pr(M_k)}_{\text{prior probability}}$$

where

$$p(x_1,...,x_n|M_k) = \int p(x_1,...,x_n|\theta_k)p(\theta_k)d\theta_k.$$

- Directly generalized to any number of models.



$p(G \backslash \mathbf{X}) = 0.248$ $\quad$ $p(G \backslash \mathbf{X}) = 0.181$

- **Bayesian Model Averaging** (BMA).

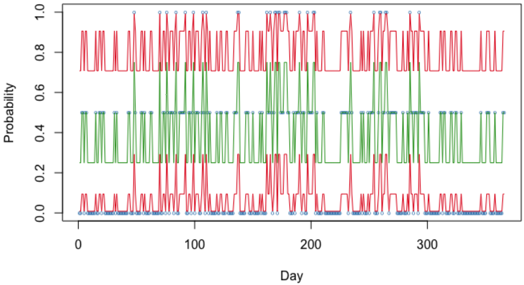
And hey! ... let's be careful out there.

- Be especially careful with Bayesian model comparison when
  - The compared models are
    - very different in structure
    - severly misspecified
    - very complicated (black boxes).
  - The priors for the parameters in the models are
    - not carefully elicited
    - only weakly informative
    - not matched across models.
  - The data
    - has outliers (in all models)
    - has a multivariate response.

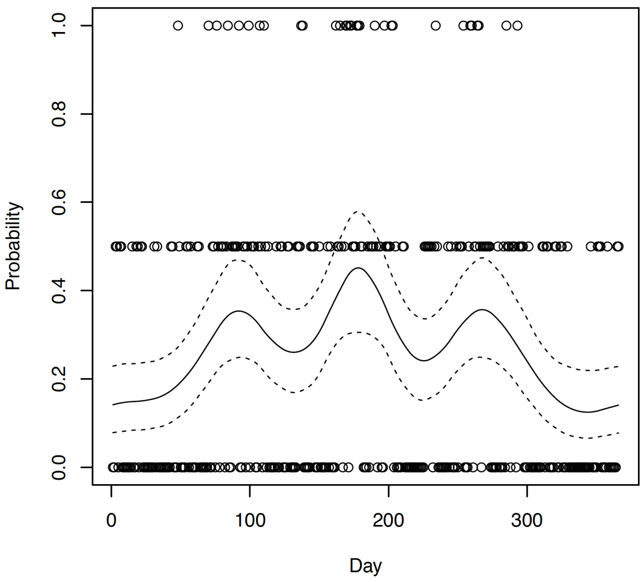
Smoothness priors

- Example: rain in Tokyo. Rain probability is likely not the same on every day.
- More general model:
$$x_i|\theta_i \sim \mathrm{Bern}(\theta_i)$$
- Very flexible: every day has its own probability θ_i.
- Smoothness prior
$$\begin{aligned} x_i|\theta_i &\sim \mathrm{Bern}(\theta_i) \\ \mathrm{Logit}(\theta_i) &= f(\mathrm{day}) \\ f &\sim \mathrm{GaussianProcess} \end{aligned}$$
- Natural extension to spatial problems. Smooth latent fields.

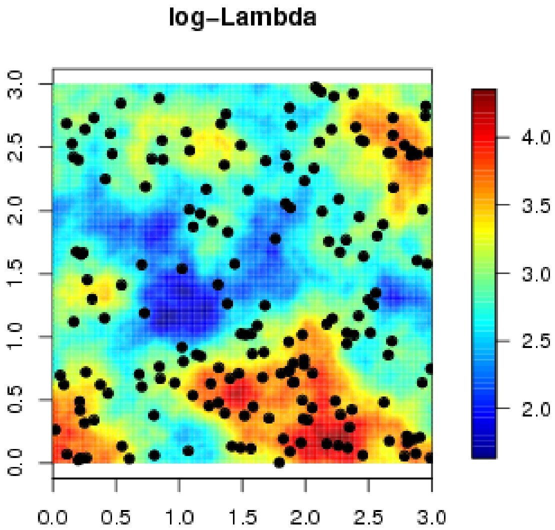
Tokyo rain - 2 years of data - no smooth



Tokyo rain - 2 years of data - smooth



Log Gaussian Cox process for spatial count data



Log Gaussian Cox process for spatial count data

- Log Gaussian Cox Process over a spatial domain $\mathbf{s} \in \mathcal{S}$.
- Spatial intensity $\lambda(s)$ surface.
- Counts in a subregion $\tilde{\mathcal{S}} \subset \mathcal{S}$ is

$$N_y(\tilde{\mathcal{S}}) \sim \mathsf{Poisson}\left(\int_{\tilde{\mathcal{S}}} \lambda(\mathbf{s}) d\mathbf{s}\right).$$

- Log intensity model

$$\log \lambda(\mathbf{s}) = \alpha + \boldsymbol{x}(\mathbf{s})\boldsymbol{\beta} + \xi(\mathbf{s})$$

  where
  - α is an intercept
  - $\boldsymbol{x}(\mathbf{s})$ are spatial covariates
  - $\boldsymbol{\beta}$ are regression coefficients
  - $\xi(\mathbf{s})$ is a Gaussian process (GP)