

A lightning tour of Bayesian Statistics and Regularisation

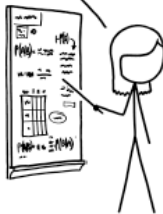
Daniel Lawson — University of Bristol

Lecture 07.1 (v4.0.0)

In anticipation

GIVEN THESE PREVALENCES,
IS IT LIKELY THAT THE TEST
RESULT IS A FALSE POSITIVE?

WELL, THIS CHAPTER IS ON
BAYES' THEOREM, SO YES.



SOMETIMES, IF YOU UNDERSTAND
BAYES' THEOREM WELL ENOUGH,
YOU DON'T NEED IT.

Signposting

- ▶ Bayesian methodology is a huge and important area
- ▶ This is to give the background for:
 - ▶ Bayes Rule
 - ▶ Signposting Bayesian tools
 - ▶ Understanding Latent Dirichlet Allocation
 - ▶ Regularisation

Questions

- ▶ What is a “prior”, and do we believe it?
- ▶ What is a “posterior” probability?
- ▶ Can we still do inference when the posterior is intractable?
- ▶ Can we justify regularization?
- ▶ Are Bayesian methods slower than frequentist methods?

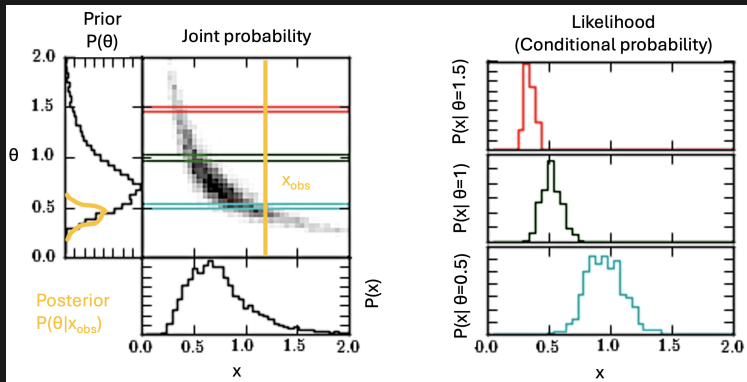
A brief aside into Bayesian Modelling

- ▶ Bayesian Models are **generative**, that is, you can simulate data from them.
- ▶ They consist of:
 - ▶ a **prior** $\Pr(\theta)$, that is conceptualised as either a model, or as beliefs,
 - ▶ and the **likelihood** $\Pr(x|\theta)$, that depends on the data.
- ▶ The task is to integrate over the prior, to find the **posterior probability** using **Bayes' theorem**:

$$\Pr(\theta|x) = \frac{\Pr(x|\theta) \Pr(\theta)}{\Pr(x)}$$

- ▶ In general $\Pr(x)$ is hard to evaluate but there are methods to avoid doing this.

Visualisation of Bayes Theorem



Example of Bayes Theorem

- ▶ One important application of Bayes' theorem is **False discovery**.
 - ▶ Imagine that we made a Bad-Guy-Detector (TM) which has a 99% chance of seeing a malicious attack if present ($\theta = 1$)...
 - ▶ But a 0.01% chance of declaring an attack when it isn't ($\theta = 0$).
 - ▶ Let p be the true frequency of malicious attacks.
 - ▶ If our BGD activates ($x = 1$), what is the probability of a true attack?
- ▶ Probability of the data: $\Pr(x = 1) = 0.99p + 0.0001(1 - p)$
- ▶ Probability of an attack: $\Pr(\theta = 1|x = 1) = 0.99p / \Pr(x = 1)$
- ▶ If $p = 0.001$ then $\Pr(\theta = 1|x = 1) \approx 0.9$
- ▶ If $p = 0.0001$ then $\Pr(\theta = 1|x = 1) \approx 0.5$
- ▶ If $p = 0.00001$ then $\Pr(\theta = 1|x = 1) \approx 0.09$
- ▶ If $p = 0.000001$ then $\Pr(\theta = 1|x = 1) \approx 0.001$

Etymology of Bayes: Conjugacy and tractability

- ▶ Bayesian Inference means **integrate over** parameters:
- ▶ In a **conjugate** model the integral is analytic: i.e. we compute $\Pr(x)$ and therefore $\Pr(\theta|x)$.
 - ▶ Example: we had many Bad-Guy-Detectors $i = 1 \dots n$ and calculate $\Pr(\theta_i|x_i)$.
- ▶ Monte-Carlo methods allow **sampling of posterior parameters** $\Pr(\theta|x)$ conditional on the data without ever evaluating $\Pr(x)$
 - ▶ Example: our Bad-Guy-Detectors correlated in the form $\Pr(x_i|\{\theta\}_{i=1}^n)$ and $p(\{\theta\}_{i=1}^n)$.
- ▶ Some models are **doubly intractable**: we cannot compute $\Pr(x|\theta)$ and they cannot be sampled.
 - ▶ For example, Markov Random Fields, where the data are dependent $\Pr(x_i|\{\theta\}_{i=1}^n, x_{i-1}, x_{i+1})$.
 - ▶ Solved by e.g. **Approximate Bayesian Computation**

Conjugate models

- ▶ Conjugate models have a **known distribution** for the Prior for which observing data only updates the **parameters**.
 - ▶ We update the posterior **online** by visiting each datapoint only once.
- ▶ This means we can interpret the prior in terms of **pseudo observations**:
 - ▶ either data we have seen already,
 - ▶ or data we pretend to have seen in order to specify a prior distribution.
- ▶ The set of possible conjugate models is limited, though they can often be used as a part of a larger model.

Conjugate model example

- ▶ Example: The Beta-Bernoulli model for binary outcomes.
 - ▶ In the Bernoulli model $p(x|p)$ we flip a (biased) coin x which is heads ($x = 1$) with some unknown probability p .
 - ▶ If we parameterise the prior $p(p) = \text{Beta}(\alpha, \beta)$, with $\hat{p} = \alpha / (\alpha + \beta)$,
 - ▶ then after n observations $p(p|\{x\}) = \text{Beta}(\alpha', \beta') = \text{Beta}(\alpha + \sum_{i=1}^n x_i, \beta + (n - \sum_{i=1}^n x_i))$,
 - ▶ i.e. α was our prior number of successes (heads) and β our prior number of failures (tails).
- ▶ All discrete distributions with conjugate priors have this interpretation!
- ▶ Continuous distributions also contain a concept of *the number of observations used to form the prior estimate*.
- ▶ There is a super useful list of conjugate priors and interpretations on the [Conjugate Prior Wikipedia page](#)!

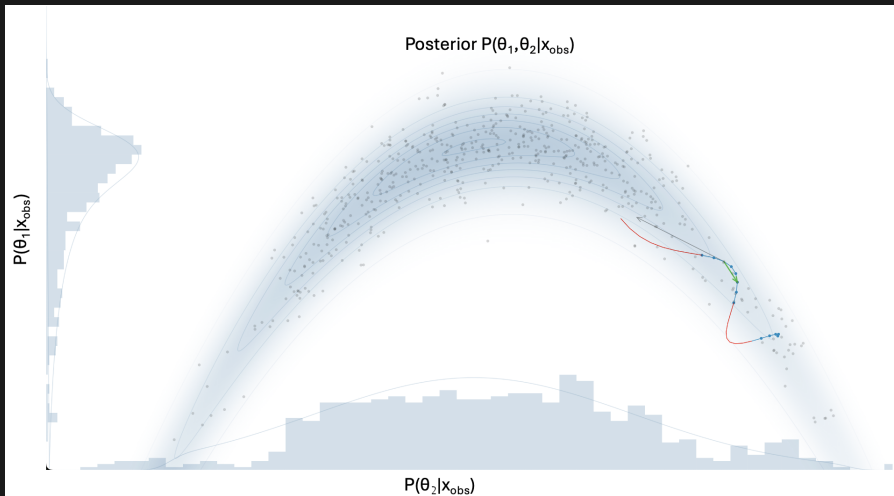
Markov Chain Monte Carlo (MCMC)

- ▶ MCMC¹ allows sampling from a posterior when we can evaluate the likelihood and the prior at any parameter value, but not integrate it.
- ▶ It **searches parameter space** by comparing the posterior at the current point to the posterior at a proposed point.
- ▶ (Somewhat surprisingly) the set of samples resembles a random sample from the posterior.
- ▶ Used for prediction, parameter estimation, etc, by averaging over the samples.
- ▶ Relatively **costly** - the number of likelihood evaluations required to obtain convergence is hard to predict.
- ▶ Careful choice of proposals is then needed.

¹e.g. Gamerman and Hedibert. Markov chain Monte Carlo: stochastic simulation for Bayesian inference.

MCMC Visualisation

Bayesian Sampling Visualisation



Tools for Bayesian Modelling using MCMC

- ▶ MCMC is very popular because it is straightforward to implement many models using it.
- ▶ Some important tools for Bayesian Inference allow models to be specified, and automatically do the inference for you using MCMC:
 - ▶ OpenBUGS (<http://openbugs.net/w/FrontPage>)
 - ▶ JAGS (<http://mcmc-jags.sourceforge.net/>)
 - ▶ STAN (<http://mc-stan.org/>)
- ▶ STAN is the current darling because it uses a clever method to sample, called the “no U-turn sampler” (NUTS) which searches parameter space with **Hamiltonian Monte Carlo**, a method that gives the search “momentum”.
 - ▶ Visualised! **Bayesian Sampling Visualisation**

Approximate Bayesian Computation (ABC)

- ▶ ABC² allows inference when the Likelihood cannot be evaluated.
- ▶ It works by:
 - ▶ **Simulating** data from a model,
 - ▶ Creating summary statistics,
 - ▶ **Comparing** the simulated data **to the real data**,
 - ▶ Accepting parameters that generate **sufficiently close** data.
- ▶ Simulation Based Inference (SBI) exploits neural networks to map data to parameters.
 - ▶ Can be computationally costly **unless the simulation is fast**.

²Beaumont, Zhang, and Balding. "Approximate Bayesian computation in population genetics." *Genetics* 162.4 (2002): 2025-2035.

Reflection

- ▶ Are Bayesian approaches inherently slow?
- ▶ When might MAP estimation and full Bayesian inference produce different predictions?
- ▶ How have we encountered regularisation previously?
 - ▶ How does it relate to **non-parametric** models?
 - ▶ How does it relate to Random Forests, decision trees and other flexible predictors?
- ▶ When would we regularise vs cross-validate?
- ▶ Keep looking for regularisation as we move through the course, especially in flexible machine learning systems such as neural networks.

References

- ▶ There is a super useful list of conjugate priors and interpretations on the [Conjugate Prior Wikipedia page](#)!
- ▶ Methodology:
 - ▶ Gamerman and Hedibert. Markov chain Monte Carlo: stochastic simulation for Bayesian inference.
 - ▶ Doucet, Godsill, and Andrieu. "On sequential Monte Carlo sampling methods for Bayesian filtering." Statistics and computing 10.3 (2000): 197-208.
 - ▶ Andrieu, Doucet, and Holenstein [Particle Markov chain Monte Carlo methods](#)
- ▶ ABC:
 - ▶ Murray, Ghahramani, and MacKay. "MCMC for doubly-intractable distributions." arXiv:1206.6848 (2012).
 - ▶ Beaumont, Zhang, and Balding. "Approximate Bayesian computation in population genetics." Genetics 162.4 (2002): 2025-2035.
- ▶ Variational Inference:
 - ▶ Blei and Jordan. "[Variational inference for Dirichlet process mixtures](#)", Bayesian analysis 1.1 (2006): 121-143.
 - ▶ [A Beginner's Guide to Variational Methods](#), by Eric Jang.

Appendix: Sequential Monte Carlo (SMC) for filtering

- ▶ Filters are a class of model that take a sample of parameters and move them (through some observed space such as time) to track a changing distribution, for example, estimates of where an object is over time.
- ▶ Hidden Markov Models (HMMs) do this analytically for discrete parameter spaces, where the observation is a random variable depending on the true state of a system.
- ▶ The Kalman Filter is famous as it can be solved analytically by tracking a Normal distribution estimate of the location.
- ▶ Sequential Monte Carlo is a tool for implementing a wide range of Bayesian models.
- ▶ It was pioneered³ and been integrated into MCMC⁴ in Bristol.

³Doucet, Godsill, and Andrieu. "On sequential Monte Carlo sampling methods for Bayesian filtering." *Statistics and computing* 10.3 (2000): 197-208.

⁴Andrieu, Doucet, and Holenstein [Particle Markov chain Monte Carlo methods](#)