Probabilistic Machine Learning:
3. Estimation

Tomasz Kajdanowicz, Piotr Bielak, Maciej Falkiewicz, Kacper Kania, Piotr Zieliński

Department of Computational Intelligence
Wroclaw University of Science and Technology

The presentation has been inspired and in some parts totally based on

1. working notes of Prof. Martin Ridout,University of Kent, UK
2. Murphy's book
3. working notes of Tamara Broderick: Variational Bayes and beyond: Bayesian inference for big data
4. working notes of David Blei: Variational Inference
5. presentation of David Blei, Rajesh Ranganat, Shakir Mohamed: Variational Inference: Foundations and Modern Methods
6. presentation of Shakir Mohamed: Variational Inference for Machine Learning

Table of Contents

Pre-reading

1. Markov Chain Monte Carlo
2. Variational Inference

Table of Contents

What is statistic?

Statistic

A **statistic** is a piece of data from a portion of a population.

Example

- a bit of information
- part of a data seti
- if you know something about 10% of people
  that's all statistic

What is a Statistic used for?

- way to understand the data that is collected
- producing some meaningful information about that data

Types of statistic

Descriptive Statistics

- describe data (e.g. sample mean or sample median)
- includes order statistics (tell something about how the data is ordered)
- charts and graphs
- anything that describes data is descriptive statistics

Estimators

- used to guess at a parameter
- something about a population
- if you know the sample mean you can use it to guess what the population mean is
- used in inferential statistics: a “best guess” about something, based on data

Test Statistics

- which are used in null hypothesis testing
- known fact about a population tested to see if it is true or nots
- some examples of test stats: t-score, and chi-square

What is an Estimator?

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Sample mean & Estimator

The **sample mean** is an estimator for the **population mean**. An **estimator** is a statistic that estimates some fact about the population.

Example

- want to know the average height of children in a school with a population of 1000 students
- take a sample of 30 children, measure them and find that the mean height is 148 cm
- this is sample mean
- from sample mean estimate that the population mean (estimand) is about 148 cm

Point vs. Interval

Estimators can be:

- a single value (like the standard deviation)
- a range of values (like a confidence interval)

When an estimator is a range of values, it's called an **interval estimate**.

When it is a single value it's called a **point estimate**.

Estimator types

Estimators can be described in several ways:

- Biased: a statistic that is either an overestimate or an underestimate.
- Unbiased: an accurate statistic that neither underestimates nor overestimates.
- Efficient: a statistic with small variances (the one with the smallest possible variance is also called the “best”). Inefficient estimators can give you good results as well, but they usually requires much larger samples.
- Invariant: statistics that are not easily changed by transformations, like simple data shifts.
- Sufficient: a statistic that estimates the population parameter as well as if you knew all of the data in all possible samples.

Table of Contents

Akaike information criterion - AIC

- (AIC) is an estimator of out-of-sample prediction error
- measures relative quality of statistical models for a given set of data
- AIC estimates the relative amount of information lost by a given model while representing the process that generated the data
- needed modification for small sample

AIC

Let k be the number of estimated parameters in the model. Let $\hat{L}$ be the maximum value of the likelihood function for the model. Then the AIC value of the model is the following

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

Bayesian information criterion - BIC

BIC

$$\mathrm{BIC} = \ln(n)k - 2\ln(\hat{L})$$

where
$\hat{L}$ is the maximized value of the likelihood function of the model M, i.e. $\hat{L} = p(x \mid \hat{\theta}, M)$
$\hat{\theta}$ are the parameter values that maximize the likelihood function
x is the observed data
n is the number of data points in x
k is the number of parameters estimated by the model

For example, in multiple linear regression:

- the estimated parameters are the intercept
- the a slope parameters
- and the constant variance of the errors

$$\boxed{\text{thus, } k = a + 2}$$

Likelihood-ratio test

- is a hypothesis test that choose the “best” model between two nested models
- nested - one model is a special case of the other:
  e.g. model one has four predictor variables (height, weight, age, ocoupation), model two has two predictor variables (age,ocupation)

Running the Test

- the null hypothesis is that the smaller model is the “best” model
- if the null hypothesis is rejected, then the larger model is a significant improvement over the smaller one

$$LRT = -2\log_e\left(\frac{L_s(\hat{\theta})}{L_g(\hat{\theta})}\right)$$

$L_s(\hat{\theta}|x)$ - likelihood of the simpler model
$L_g(\hat{\theta}|x)$ - likelihood of the greater model The test statistic approximates a chi-squared random variable with degrees of freedom equal to the difference in the number of parameters for the two models.

Other hypotheses tests

- Wald Test - based on the (standardized) distance between θ_0 and $\hat{\theta}$
- Likelihood Ratio Test - based on the distance from $L(\theta_0)$ to $L(\hat{\theta})$
- Rao Score Test - based on the gradient of the loglikelihood at θ_0.

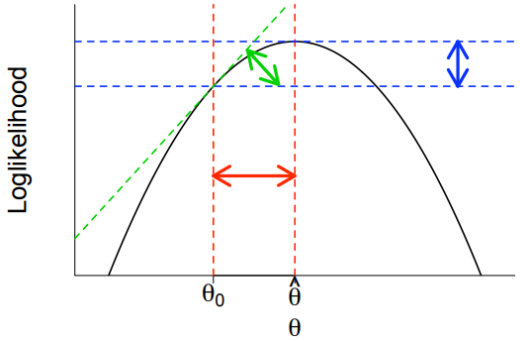


Table of Contents

Least squares

Objective

adjusts the parameters of a model function to best fit a data set

Setup

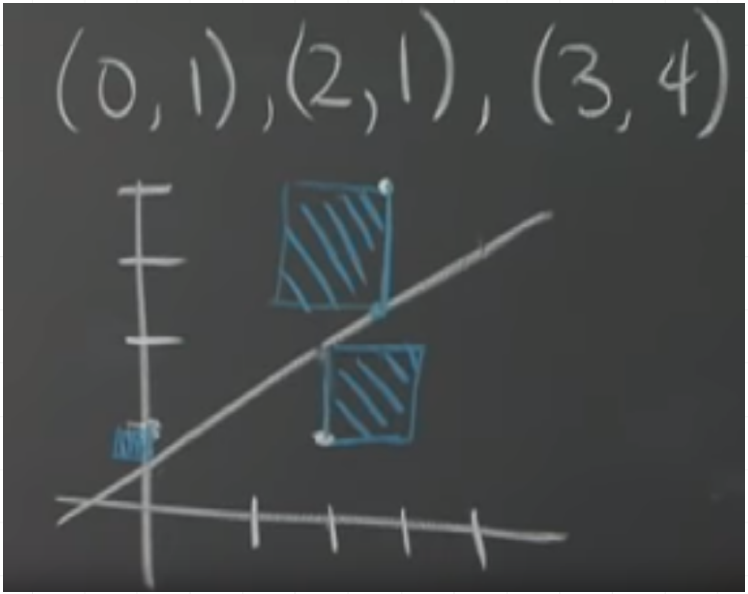
- data: n data pairs (x_i, y_i), $i = 1, \ldots, n$, where x_i is an independent variable and y_i is a dependent variable
- model:$f(x, \beta)$ - m adjustable parameters are held in the vector β
- the fit of a model to a data point is measured by its residual

$$r_i = y_i - f(x_i, \beta)$$

Least squares method

$$\arg\min_{\beta} \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (x_i - f(x_i, \beta))^2$$

Example



Solving the least squares problem

The minimum of the sum of squares is found by **setting the gradient to zero**.
There are m gradient equations:

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0, \; j = 1, \ldots, m,$$

Since $r_i = y_i - f(x_i, \boldsymbol{\beta})$, the gradient equations become

$$-2 \sum_i r_i \frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_j} = 0, \; j = 1, \ldots, m$$

The gradient equations apply to all least squares problems. Each particular problem requires particular expressions for the model and its partial derivatives.

Homework 1

1. Please understand the mechanics that implements partial derivatives calculation for:
   - Linear least squares - the model is a linear combination of the parameters (linear regression)
   - Non-linear least squares
2. Consider and check literature fo the regularization of LS:
   - Ridge regression: $\|\beta\|^2$, the L_2-norm of the parameter vector, is not greater than a given value
   - Lasso regression: $\|\beta\|$, the L_1-norm of the parameter vector, is not greater than a given value

Table of Contents

Maximum likelihood estimation

The goal of maximum likelihood estimation is to find the values of the model parameters that maximize the likelihood function over the parameter space $\theta \in \Theta$ that is

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \widehat{L}_n(\theta, \mathbf{x})$$

In practice, it is often convenient to work with the logarithm of the likelihood function, called the **log-likelihood**:

$$\log L_n(\theta, x)$$

- the logarithm is a monotonic function
- the maximum of $\log L_n$ occurs at the same value as does the maximum of L_n
- for independent and identically distributed random variables x likelihood is the product of univariate density functions -> product transforms to sum of logs
- summarizing is faster and no loss of precision in multiplication

Explicit MLE

If L is diffretiable in $\theta = [\theta_1, \theta_2, \ldots, \theta_k]^T$ the necessary conditions for the occurrence of a maximum (or a minimum) are:

$$\frac{\partial L}{\partial \theta_1} = 0, \frac{\partial L}{\partial \theta_2} = 0, \ldots, \frac{\partial L}{\partial \theta_k} = 0$$

- for some models, these equations can be explicitly solved for $\hat{\theta}$
- but in general no closed-form solution to the maximization problem is known or available
- MLE can only be found via numerical optimization
- $\hat{\theta}$ is indeed a (local) maximum if Hessian matrix is negative semi-definite at $\hat{\theta}$ (ocal concavity)
- exponential family probability distributions are logarithmically concave

Iterative procedures

Methods starting from initial guess of θ seels to obtain convergent sequence $\left\{\hat{\theta}_r\right\}$.

1. Hill climbing

$$\widehat{\theta}_{r+1} = \widehat{\theta}_r + \eta_r \mathbf{d}_r(\widehat{\theta})$$

   where $\mathbf{d}_r(\widehat{\theta})$ indicates the direction of the rth "step" and the scalar η_r captures the "step length"

2. Gradient descent
   $\eta_r \in \mathbb{R}^+$ that is small enough for convergence and $\mathbf{d}_r(\widehat{\theta}) = \nabla \ell(\widehat{\theta}_r; \mathbf{x})$
3. Some other:
   - Newton–Raphson method
   - Natural Gradient (Fisher information matrix) for linear models

Table of Contents

MAP

- the prior density, $p(\theta)$, tells us the likely values that θ may take before looking at the sample
- combining this with what the sample data tells is the likelihood density, $p(x|\theta)$, using Bayes' rule
- posterior density of θ tells us the likely θ values after looking at the sample

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$$

- evaluating the integrals may be quite difficult, except in cases where the posterior has a nice form
- when the full integration is not feasible use the **maximum a posteriori (MAP) estimate**

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|X)$$

Practical hints

- if we have no prior reason to favor some values of θ, then the prior density is flat and the posterior will have the same form as the likelihood,
- numerically MAP is an optimization over prior $\times$ likelihood

Minimum mean square error

MMSE is a Bayesian estimator:

- θ is a random variable with pdf $p(\theta)$
- θ and x have joint pdf $p(\theta, x)$

Definition

The MMSE estimator of θ is the function $\hat{\theta} = g(x)$ that minimizes the MSE=$E[(\theta - \hat{\theta})^2]$

$$\hat{\theta} = E[\theta|x] = \int \theta p(\theta|x) d\theta$$

Markov Chain Monte Carlo

- MCMC techniques are often applied to solve integration and optimisation problems in large dimensional space
- in Bayesian inference and learning:
  - Normalisation - to obtain the posterior $p(x|y)$ given the prior $p(x)$ and likelihood $p(y|x)$, the normalising factor in Bayes' theorem needs to be computed

$$p(x|y) = \frac{p(y|x)p(x)}{\int_{x'} p(y|x')p(x')dx'}$$

  - Marginalisation - from joint posterior we want marginal posterior

$$p(x|y) = \int_z p(x, z|y)dz$$

  - Expectation - obtain summary statistics of the form

$$E_{p(x|y)}(f(x)) = \int_x f(x)p(x|y)dx$$

The Monte Carlo principle

Monte Carlo

The idea of Monte Carlo simulation is to draw an i.i.d. set of samples $\{x^{(i)}\}_{i=1}^{N}$ from a target density $p(x)$ defined on a high-dimensional space X.
N samples can be used to approximate the target density with the empirical point-mass function

$$p_N(x) = \frac{1}{N}\sum_{i=1}^{N}\delta_{x^i}(x)$$

where $\delta_{x^i}(x)$ denotes the delta-Dirac mass located at $x(i)$.

When $p(x)$ has standard form, e.g. Gaussian, it is straightforward to sample from it using easily available routines.
Otherwise we need to introduce more sophisticated techniques based on rejection sampling, importance sampling and MCMC.

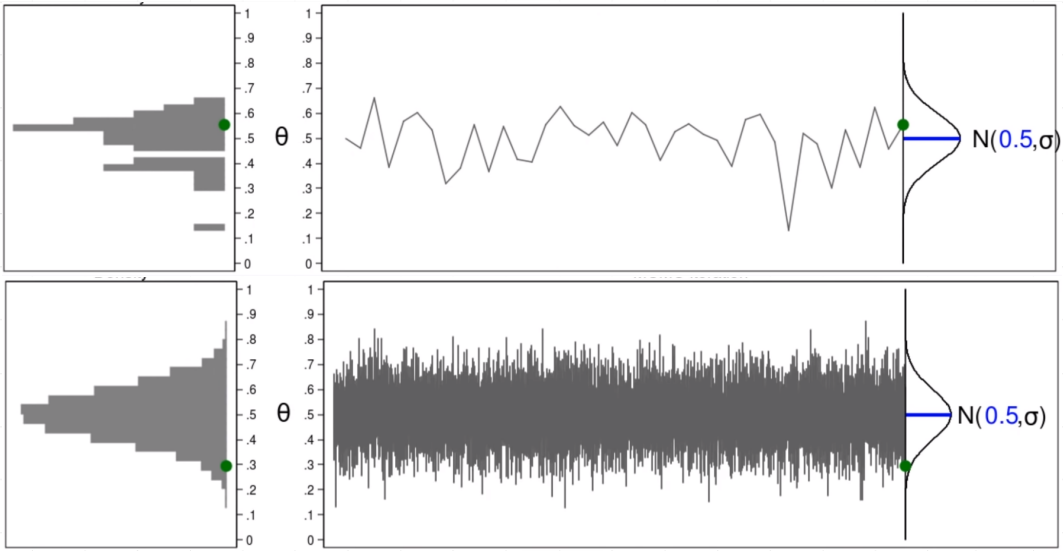
3 components

Three important concepts must be considered:

1. Monte Carlo
2. Markov Chain
3. some MCMC algorithm: e.g. Metropolis-Hastings

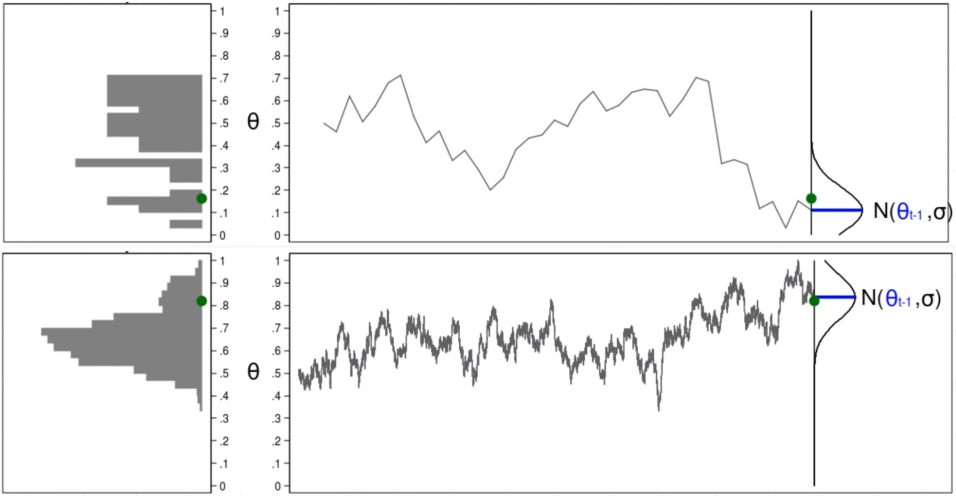
Monte Carlo

- Monte Carlo - is a random sampling used to run a simulation



Markov Chain

- MC - is a sequence numbers where each number depends on previous number in the sequence



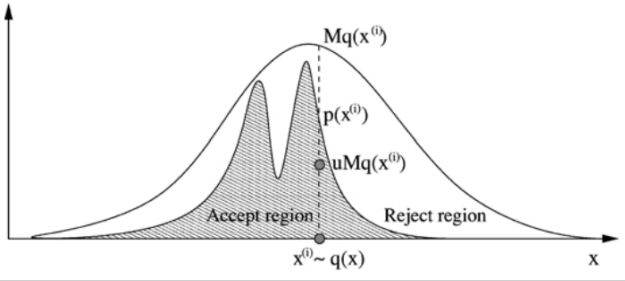
Rejection sampling

- We can sample from a distribution $p(x)$, which is known up to a proportionality constant, by sampling from another easy-to-sample proposal distribution $q(x)$ that satisfies $p(x) \leq Mq(x)$, $M < \inf$, using the accept/reject procedure

Set $i = 1$

Repeat until $i = N$

1. Sample $x^{(i)} \sim q(x)$ and $u \sim \mathcal{U}_{(0,1)}$.
2. If $u < \frac{p(x^{(i)})}{Mq(x^{(i)})}$ then accept $x^{(i)}$ and increment the counter i by 1. Otherwise, reject.



Metropolis-Hastings algorithm

- used to decide which proposed value accept or reject

1. Initialise $x^{(0)}$.
2. For $i = 0$ to $N - 1$
   - Sample $u \sim \mathcal{U}_{[0,1]}$.
   - Sample $x^\star \sim q(x^\star | x^{(i)})$.
   - If $u < \mathcal{A}(x^{(i)}, x^\star) = \min\left\{1, \frac{p(x^\star)q(x^{(i)}|x^\star)}{p(x^{(i)})q(x^\star|x^{(i)})}\right\}$

     $$x^{(i+1)} = x^\star$$

     else

     $$x^{(i+1)} = x^{(i)}$$

The fraction in min means that if a $x*$ is more likely than the current x, then we accept $x*$.

Metropolis-Hastings algorithm

- simple, but it requires careful design of the proposal distribution $q(x * |x)$
- dependence on starting values (solution: discard some starting values drawn - burn-in)
- autoceorrelation due to Markov Chain
- works well in high dimensional spaces as opposed to Gibbs sampling and rejection sampling

Visit this web for intuitive explanation.

What is a Variational Method?

Definition

Variational methods are a general family of methods for approximating complicated densities by a simpler class of densities. They turn inference into optimization.

Variational Inference (VI) - Setup

Variational Inference:

- Suppose we have some data x, and some latent variables z (e.g. paramaters of Mixture of Gaussians)
- We're interested in doing posterior inference over z
- This would consist of calculating:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(z,x)}{p(x)} = \frac{p(z,x)}{\int_{z'} p(z',x)}$$

- the numerator is easy to compute for given z, x
- the denominator is, in general, intractable

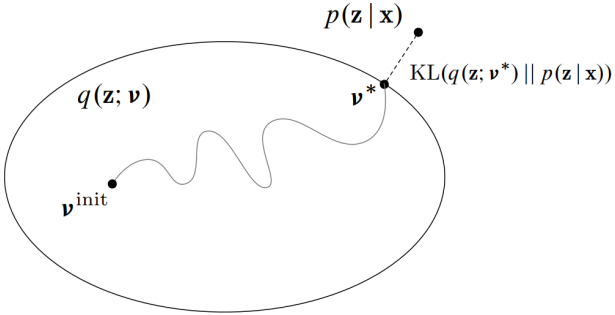
The Variational Distribution

- Rather than calculate the posterior $p(z|x)$ exactly, we can approximate it with some distribution q
- The approximating distribution q is called the variational distribution
- We'll choose q to have some nice form, so that we can feasibly calculate $q(z|x)$

KL Divergence

p and q

Since q is intended to approximate p, we want them to be as similar as possible.



Kullback–Leibler

We can measure this in many ways - the most common is KL divergence

$$KL(q(z|x)||p(z|x)) = \int_z q(z|x) \log \frac{q(z|x)}{p(z|x)} = E_q \log \frac{q(z|x)}{p(z|x)}$$

We can't calculate this expression, since we don't have $p(z|x)$

The ELBO

- We can re-write KL divergence as follows:

$$
\begin{aligned}
& KL(q(z|x)||p(z|x)) \\
& \qquad = E_q \log \frac{q(z|x)}{p(z|x)} \\
& \qquad = E_q \log q(z|x) - E_q \log p(z|x) \\
& \qquad = E_q \log q(z|x) - E_q[\log p(z,x) - \log p(x)] \\
& \qquad = E_q[\log q(z|x) - \log p(z,x)] + \log p(x) \\
& \qquad = -E_q[\log p(z,x) - \log q(z|x)] + \log p(x)
\end{aligned}
$$

- The second term here is a constant, $\log p(x)$ (does not depend on q)
- The first term is called the "ELBO" - Evidence Lower BOund
- So maximizing the ELBO is equivalent to minimizing the KL divergence between the posteriors!

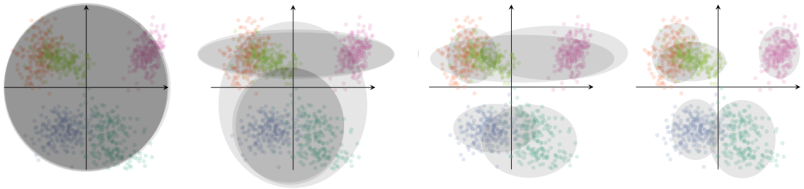
ELBO notes

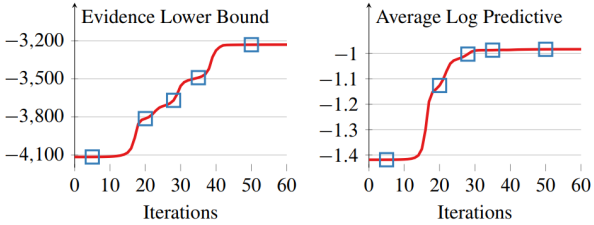
- The ELBO is important in generative modelling, a field of machine learning which is very popular at the moment
- It is how variational autoencoders (VAEs) work - the ELBO is a lower bound on $\log p(x)$
- It is often written with two terms: the ”reconstruction loss” and a KL divergence with a prior over the latent variables

$$ELBO = E_q[\log p(x|z)] - KL(q(z|x)||p(z))$$

- q is an approximate posterior distribution
- Remember we estimate the ELBO by sampling from q
- scaling up VI to massive data only with stochastic optimization

Example: mixture of gaussians





[images by Alp Kucukelbir]

Why Variational Inference?

Disadvantages:

- An **approximate posterior** q only - not always guaranteed to find exact posterior in the limit
- **Difficulty in optimisation** — can get stuck in local minima
- Typically **under-estimates the variance** of the posterior and can bias maximum likelihood parameter estimates
- **Limited theory** and guarantees for variational methods

Advantages:

- Applicable to almost **all probabilistic models**: non-linear, non-conjugate, high-dimensional, directed and undirected
- Transforms problem of **integration into one of optimisation**
- Easy **convergence assessment**
- **Compact representation** of the posterior distribution
- Can be used on **modern computing architectures** (CPUs and GPUs)

Approximate Bayesian Inference - summary

- Instead of an optimization approach to approximate posterior with general function f

$$\arg\min_{q \in Q} f(q(\cdot), p(\cdot, x))$$

  use KL divergence and maximize ELBO
- q is comes from selection of exponential distributions
- there is one more trick: Mean-field variational Bayes

$$Q_{MFVB} = \left\{ q : q(\theta) = \prod_{j=1}^{J} q_j(\theta_j) \right\}$$

Approximate Bayesian inference