Data-analysis and Inverse Methods in Astronomy

Introduction to statistical inference, including Bayesian methods

Antti Penttilä

Homepage http://wiki.helsinki.fi/display/53834

What are inverse methods?

The term *inverse method* is used quite widely and loosely, and in general this means all the situations where we have

model f that uses data x and parameters θ to produce results/to model observations y, i.e.

$$f(x; \theta) \rightarrow y$$

which is the direct problem.

Inverse problem is to find parameters θ when result y is known.

However, there are vast amount of different methods for that, depending on the problem. Next page shows my own ‘classification’ of inverse methods.

With background in statistics, I tend to feel that ‘inverse method’-people are trying to assimilate statistics as a small special case of inverse methods, whereas I feel that most of the ‘inverse methods’ are just statistical methods with a new name. In any case, do not try to ‘invent the wheel again’.

"inverse methods"

$f(x;\theta)=y$

$f(x;\theta)=y+\varepsilon$

inverse methods

$f(x;\theta)$ typically arising from theory – physics or e.g. from medical imaging. The model is often some PDE or integral equation. Randomness and/or error in result y does not play major role. e.g. Radon-transform

statistical inference/inversion/modeling

Result includes random error

frequentist inference

Maximum likelihood, linear and nonlinear models, parameter estimation, …

Bayesian inference

parameter has a priori probability distribution. Posterior distribution is searched

closed-form solution to max. likelihood estimates

computational methods

closed-form solution with conjugate prior

computational methods, MCMC the most used

Purpose of this lecture

I would like to go through the basic concepts in statistical (classical and Bayesian) inference. These include

- probability model
- likelihood and log-likelihood function
- maximum likelihood estimate
- a priori and a posteriori distributions
- Bayesian maximum probability estimate
- conjugate priori
- linear and non-linear models

All together these subjects sum to well over 20 study credits in the department of mathematics and statistics…

Frequentist inference, e.g. maximum likelihood principle

Philosophical grounds – all information from the observed data via the statistical model about the event. These two will give *the most likely* estimate of the true, unknown parameter.

In its essence there are no assumptions about Gaussian distribution etc.

…ML principle

We can start with very simple example which will introduce the concepts. Let's say that we have the heights of n people and we are interested in the distribution parameters of height in the population.

We can define our model by assuming that the population height is Gaussian distributed, and our n people are independently sampled

$$y_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2),\ i = 1, \ldots, n \quad \text{or} \quad \boldsymbol{y} \sim N_n(\mu, \sigma^2 \boldsymbol{I}_n)$$

we know $\boldsymbol{y}$ and we want to estimate the parameters μ and σ.

(see section 6.3 'Yleinen malli' in Juvela's notes)

Likelihood function

According to our model, the probability of observation y_i is given by the Gaussian probability density function $p(y_i;\mu,\sigma)$. Because the observations are independent, their joint probability is

$$p(\boldsymbol{y};\mu,\sigma) = p(y_1;\mu,\sigma)\cdots p(y_n;\mu,\sigma) =$$

$$\prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp(-\frac{1}{2\sigma^2}(y_i-\mu)^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2)$$

This is, essentially, the likelihood function L of the model and observations. One can clean up any term not involving the parameters away from L, so

$$L(\mu,\sigma;\boldsymbol{y}) \propto p(\boldsymbol{y};\mu,\sigma)$$

Log-likelihood function

The likelihood principle is simple – the parameter values that are the most likely, given the observations, are the best guess. So, we need to maximize the likelihood function $L(\mu,\sigma;\boldsymbol{y})$ for the unknown parameters using the known data $\boldsymbol{y}$.

For practical purposes, it is often easier to maximize the logarithm of the L (products become sums, exponential functions are canceled, …). This is called the log-likelihood function l. In our example

$$l(\mu,\sigma;\boldsymbol{y}) = \log(L(\mu,\sigma;\boldsymbol{y})) = -\frac{n}{2}\log(\sigma^2) - \frac{(n-1)s^2 + n(\mu-\bar{y})^2}{2\sigma^2}$$

where $\bar{y}$ is the mean of $\boldsymbol{y}$ and s^2 is the observed variance.

Maximum likelihood estimate

The maximum likelihood estimate (MLE) of the unknown model parameters is now the values that maximizes l. By looking at l

$$l(\mu, \sigma; \boldsymbol{y}) = -\frac{n}{2}\log(\sigma^2) - \frac{(n-1)s^2 + n(\mu - \bar{y})^2}{2\sigma^2}$$

we can see that the value that maximizes l for μ, called $\hat{\mu}$ is $\bar{y}$

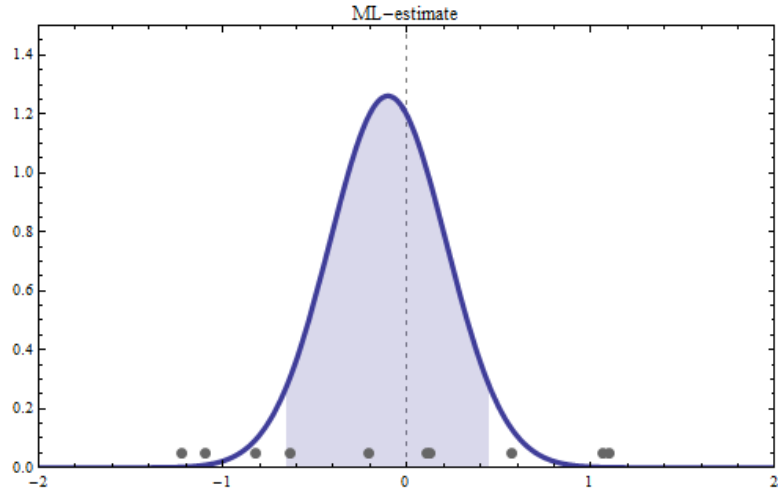
For the MLE of σ^2 we can take the 1st derivative of l in respect to σ^2 and solve its root. It turns out that

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

So, quite as expected, our answer to this problem is that we will guess that the population mean is the sample mean and the population variance is the (biased) sample variance. Obvious, but the same ML principle applies to more complicated problems, too.

Maximum likelihood inference

ML estimate is a *point estimate* for the (unknown) parameter value. Often error estimate is needed – confidence intervals.



10 observations from $N(0,1)$ and ML-estimate of the mean value with 95% confidence interval.

Where is the Gaussian assumption?

The correct distribution of the ML-estimate cannot always be (analytically) derived. However, central limit theorem states that it approaches Gaussian. Confidence intervals etc. can be found using that assumption.

Bayesian inference

Philosophical difference to 'classical' frequentist statistics – parameter θ is random variable. Furthermore, there is a way to include subjective information via *a priori* distribution.

The Bayes theorem

$$\mathrm{P}(B|A) \propto \mathrm{P}(B)\ \mathrm{P}(A|B)$$

Probability of B given A is proportional to probability of B (*a priori* information) times probability of A given B (model).

Note that (subjective) decision about a priori information is always done in Bayesian inference, at least implicitly.

…Bayesian inference

The Bayesian inference is actually not that different from the frequentist inference with the ML principle. The Bayes theorem says, that

$$\mathrm{P}(B|A) \propto \mathrm{P}(B)\ \mathrm{P}(A|B)$$

We can apply that to modeling:

A – observed data

B – parameters

probability of B – *a priori* distribution $D_{\mathrm{pr}}(\theta)$

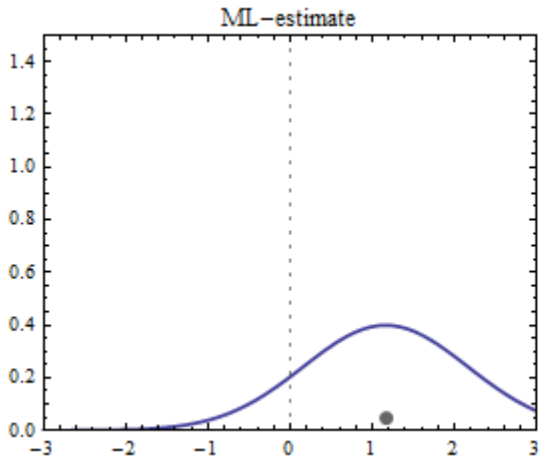
prob. of observations given the parameters – same likelihood function as before

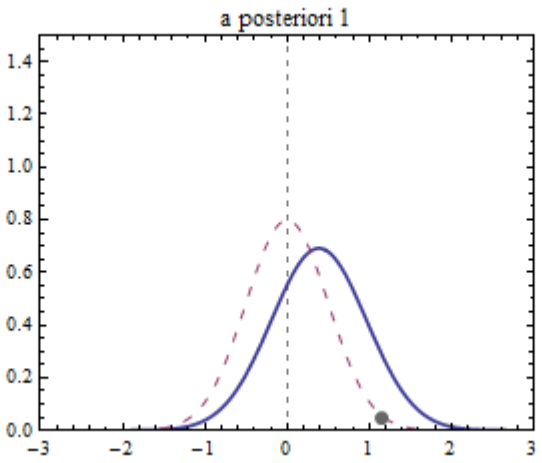
prob. of parameters given the observations – *a posteriori* distribution $D(\theta|y)$

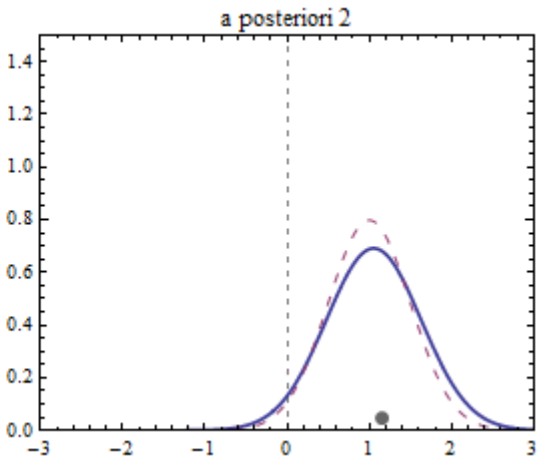
$$D(\theta|y) \propto D_{pr}(\theta)\ L(\theta; y)$$

Example of posterior distributions

Observations from normal distribution. Estimate μ in three cases – classical ML (left), a posteriori (middle) with correct a priori, and a posteriori (right) with misleading a priori

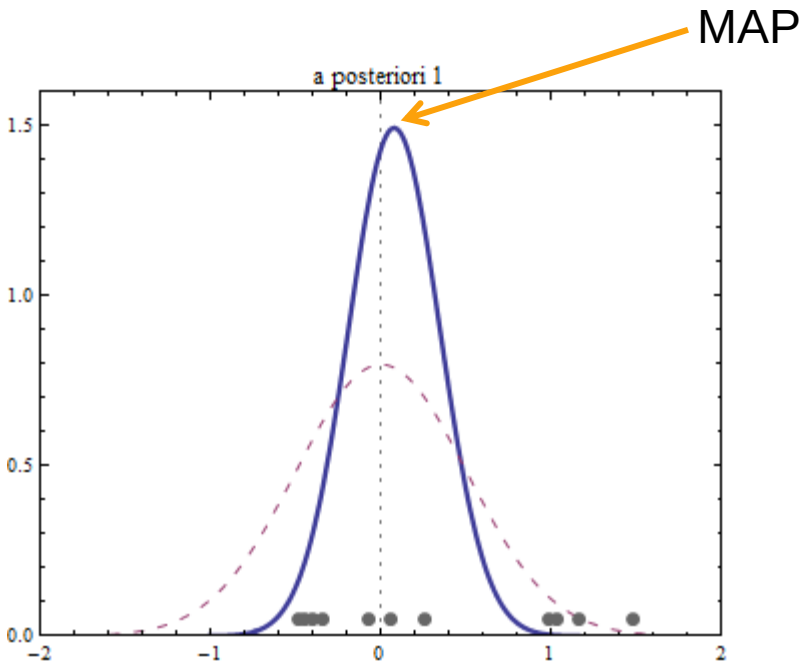






Maximum a posteriori estimate

With Bayesian inference the whole posterior distribution is the answer, However, if some (point)estimate for the parameters is needed, it is the *maximum a posterior* (MAP) estimate. So, instead of L as in freq. inference one maximizes $D_{\text{pr}}(\theta)\, L(\theta;y)$.



How to choose prior distribution

- Select computationally easy way – use conjugate priors

- Select non-informative/vague priors

- (Best way in my opinion) Use subjective decision. Try to include the knowledge about the possible behavior of the parameters and about the likely values into a priori distribution. Do not choose too restrictive (small variance) prior. Check the sensitivity of poster to prior.

Conjugate priors

Conjugate prior $D_{\text{pr}}(\theta)$ for certain probability model (likelihood function) is such a distribution that also the posterior distribution $D(\theta|y)$ is of the same family as the prior. You will have only simple formulas to how the posterior parameters are changed from prior parameters. Example:

> Your model says that the observations should come from (discrete) Geometric distribution with parameter p. If you choose Beta-distribution as a priori for p: $D_{\text{pr}}(p) = B(\alpha,\beta)$ (hyperparameters α and β), your a posteriori distribution for p is also Beta with parameters

$$D(p|\boldsymbol{y}) = B(\alpha + n, \beta + \sum_{i=1}^{n} y_i)$$

Non-informative priors

- Use if you don't want to quantify prior information but still want to use Bayesian analysis
- One option is 'even' distributions
  - for location parameter, e.g. μ in Gaussian distribution an even dist. for large range, or some nearly even dist. like normal with huge variance
  - For scaling parameter like σ^2 in Gaussian the concept of 'even distribution' is more difficult, but e.g. $D_{\text{pr}}(\sigma^2)=1/\sigma^2$ is used (see).
  - Use of 'even' priors can lead to equivalent inference with the classical frequentist way, but not always
- Another option is the so-called Jeffreys-prior. The idea is that the a priori must have the same information for all the possible re-parametrizations of the model. This is based on the Fischer-information of the model (see).

Subjective priors

- You can use any prior distribution you want if only you can also justify it to the public.
- Sometimes it is convenient to limit the range of possible parameter values by using a priori that has zero probability in the unwanted areas of the parameter space.
- A priori can be used to guide otherwise difficult estimation problem to smaller, more likely subspace.
- Results from earlier studies can be used as a priori information
- Sensitivity analysis is recommended. Simulate and see how much your choice of a priori will drive the result. If too much, increase the uncertainty (i.e. variance) in your prior.

How to construct the a posteriori

If you are not using conjugate priors it might be impossible to have a closed-form solution to a posteriori distribution. In these cases the MCMC algorithm is very popular and convenient. Details will be presented later (Muinonen and Haario), but in a nutshell, MCMC is an algorithm that will construct a chain of numbers. This chain will automatically, at some point, converge to sample from the a posteriori distribution.

Now, you don't 'know' the a posteriori, but you can sample it infinitely, and base you estimates on these samples. E.g. mean, standard deviation, mode, median and quantiles of the a posteriori can be computed. With distribution estimation techniques, e.g. kernel-estimation (see) you can numerically construct continuous posterior distribution from samples.

How about linear/nonlinear models

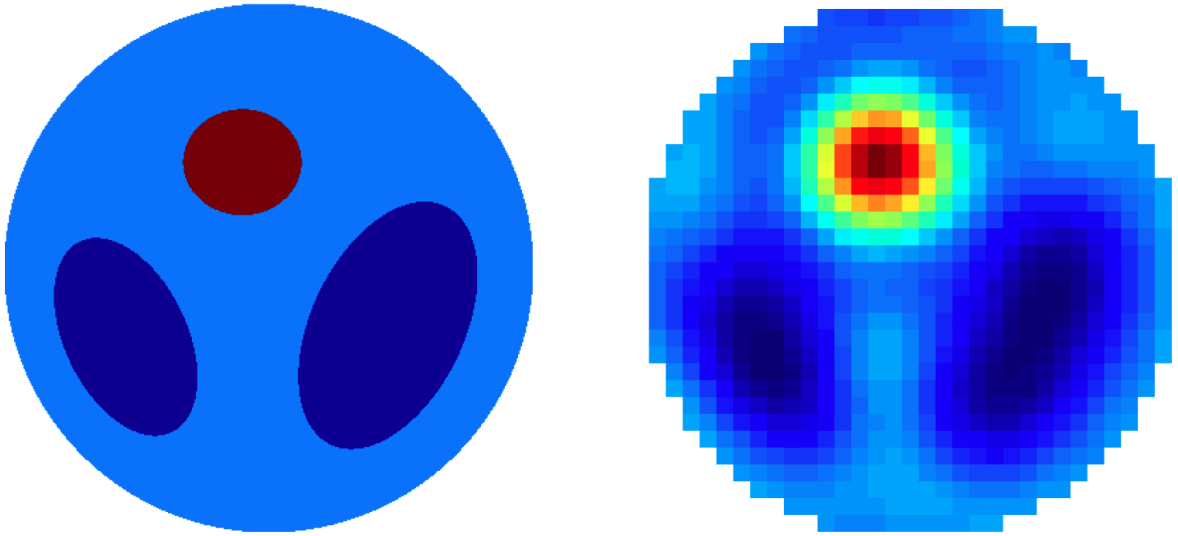
In linear/nonlinear models we have the model part, $f(x;\theta)$, which is not a probability model. The probability model comes from the random error component ε that we assume between model prediction y and the observed value $\hat{y} = y+\varepsilon$.

Usually, the error term is assumed to be Gaussian and independent. The squared errors are thus χ^2-distributed with 1 degree of freedom, and the sum of squared errors χ^2-distributed with n degrees of freedom.

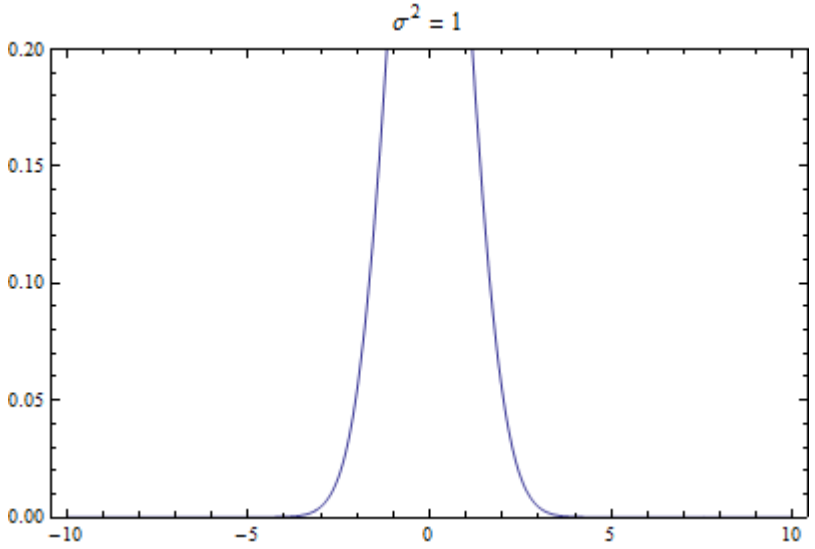
Using the ML estimation to this model produces the classical results for linear models:

$$\hat{\beta} = (X'X)^{-1}X'y \qquad \text{and} \qquad \hat{\beta} \sim N(\hat{\beta}, \frac{\sigma^2}{\sqrt{n}}C^{-1})$$

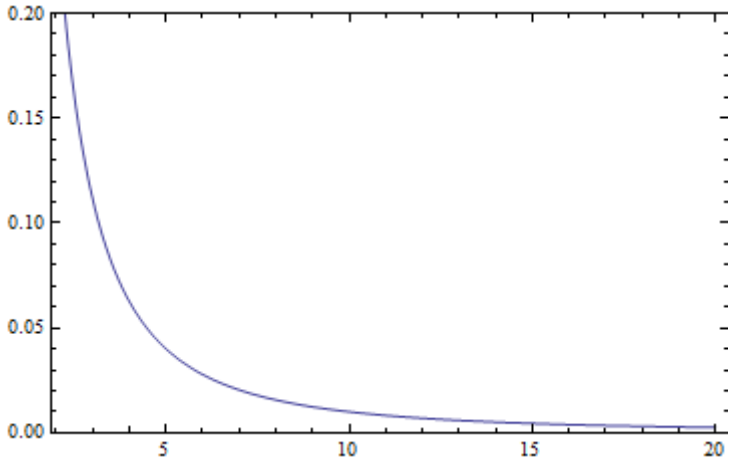
Nonlinear models have the same asymptotic behavior, but the lecture on these has to be left to another time… (see Juvela's notes, sec 4 and 6)



Reconstruction of simulated cross-section of human chest from electrical impedance tomography data using a novel reconstruction method (*courtesy of CoE in Inverse problems research, Univ. of Helsinki*)

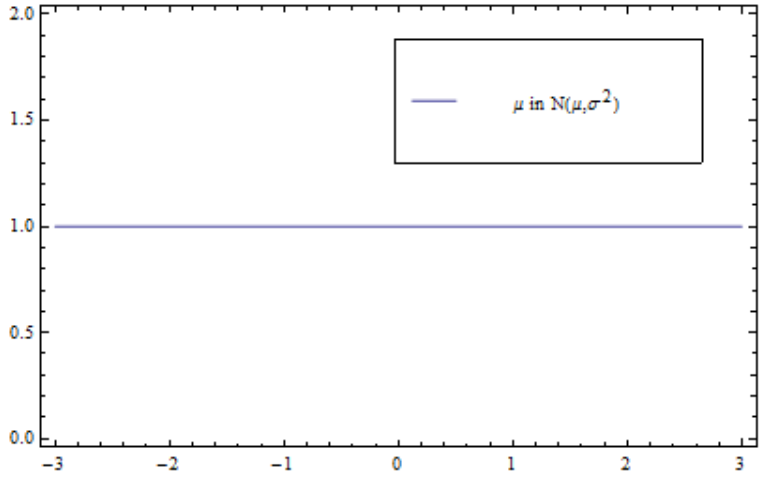


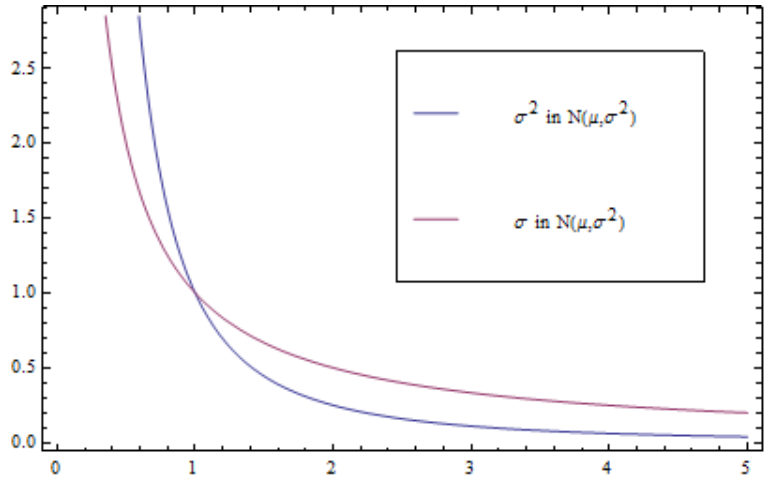
’even’ distribution for location parameter

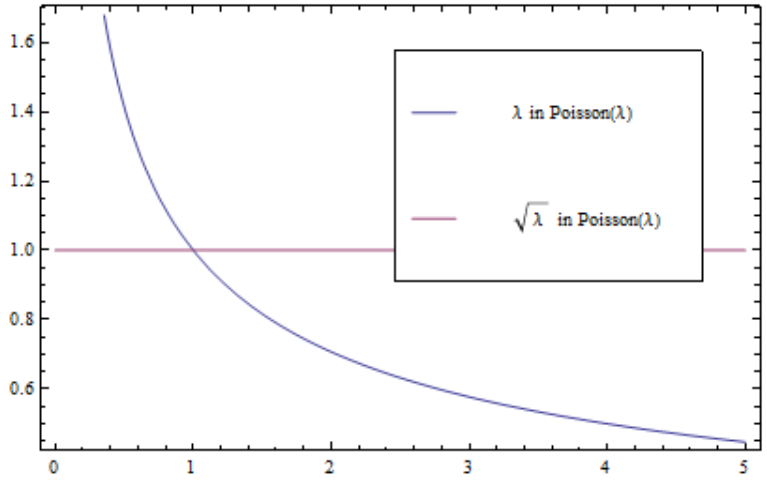


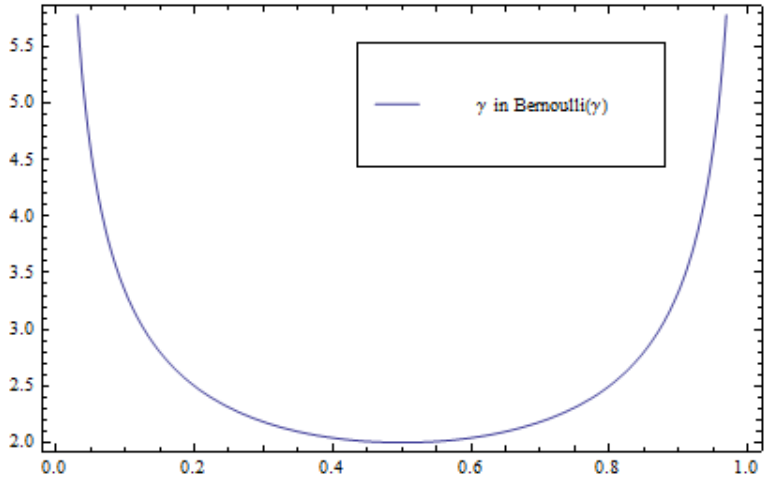
’even’ distribution for scale parameter

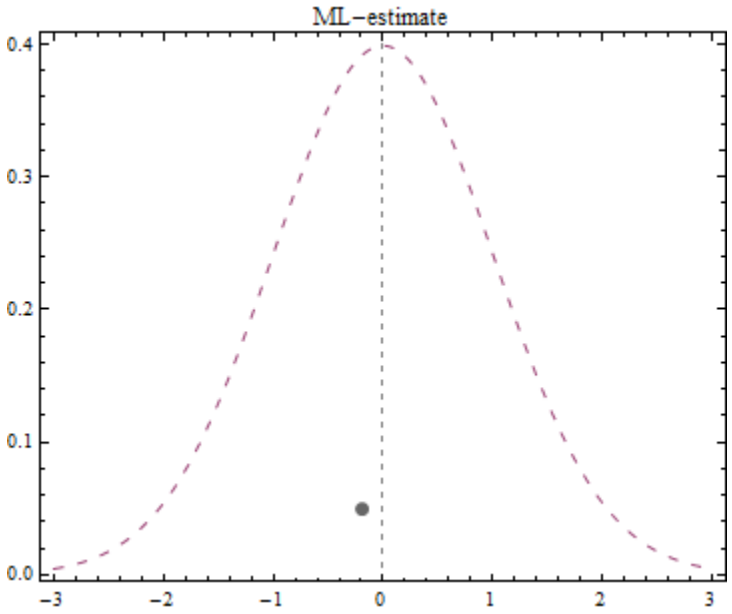
Jeffreys priors for different cases










ML−estimate
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3