Fisher Information Matrix

Jayanti Prasad
DST-SERB (Young Scientist) PDF

Inter-University Centre for Astronomy & Astrophysics (IUCAA)
Pune, India (411007)

April 29, 2015

Introduction

"The terrific utility of the Fisher Information matrix is that, if you can compute it, it enables you to estimate the parameters errors before you do the experiment. If it can be computed quickly, it also enables one to explore different experimental setups and optimize the experiment. This is why the Fisher matrix approach is so useful in survey design. Also complementary of different, independent, and uncorrelated experiments (i.e., how in combination they can lift degeneracies) can be quickly explored: the combined Fisher matrix is the sum of the individual matrices."

[L. Verde (*Statistical Methods in cosmology*)]

Why to bother about Fisher Information ?

- It is a measure of the ability to estimate a parameter and so plays an important role in parameter estimation.
- It is a measure of the state of disorder of a system or phenomenon [How it is related to entropy, Shannon information etc.,]
- The Fisher information is a measure for the amount of information that an observed random variable provides about an unknown parameter.

An Estimator

- An estimator or "point estimate" is a statistic/rule for calculating an estimate of a given quantity based on observed data.
- There are two type of estimators : point estimators and interval estimators.
- Example of estimators are "sample mean", "sample variance"

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{1}$$

- Then the bias of this estimator is defined as:

$$\text{bias}(\hat{\theta}) = \left\langle \hat{\theta} \right\rangle - \theta = \left\langle \hat{\theta} - \theta \right\rangle \tag{2}$$

  and an estimator is said to be unbiased if it has no bias.
- Mean-squared error e^2 is defined as:

$$e^2 = \left\langle [\hat{\theta} - \theta]^2 \right\rangle \tag{3}$$

Estimation

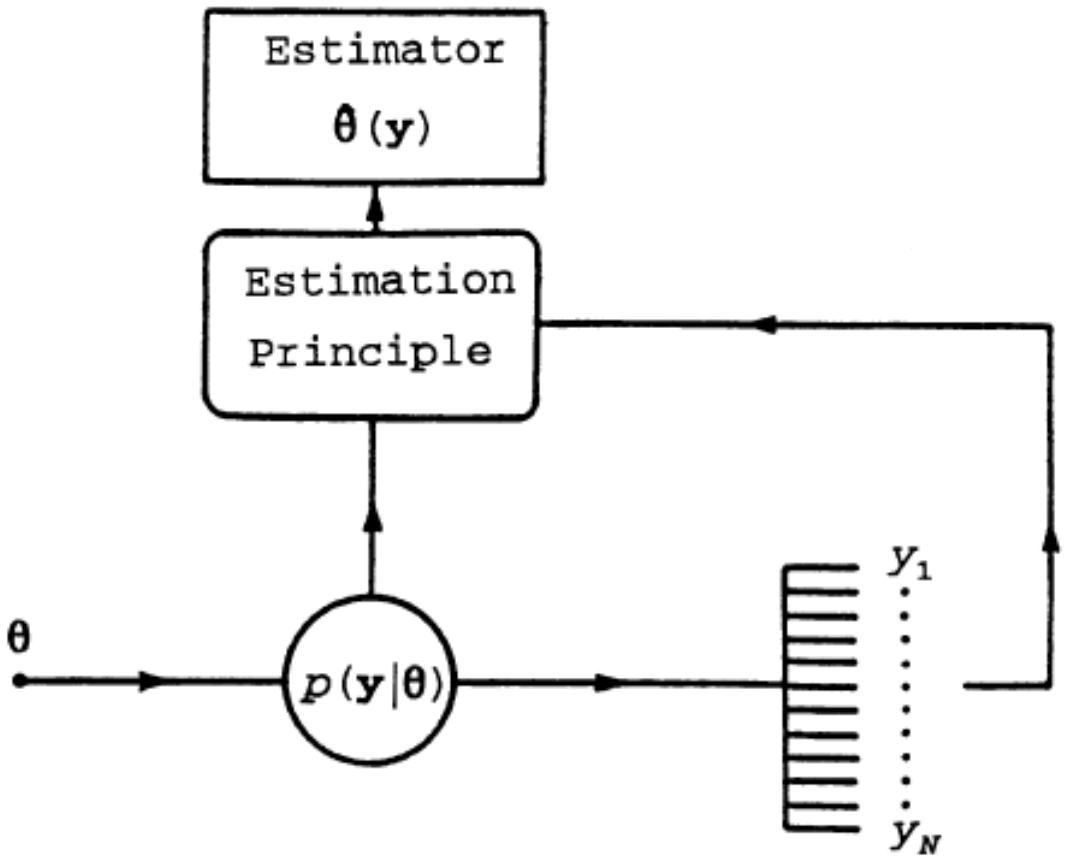

Fig. 1.1. The parameter estimation problem of classical statistics. An unknown but fixed parameter value θ causes intrinsic data $\mathbf{y}$ through random sampling of a likelihood law $p(\mathbf{y}|\theta)$. Then, the random likelihood law and the data are used to form the estimator $\hat{\theta}(\mathbf{y})$ via an estimation principle. (Reprinted from Frieden, 1991, by permission of Springer-Verlag Publishing Co.)

Frequentest vs Bayesian

Frequentest

- We are given some data y, consisting of true signal x plus additive noise n.
- We select a point estimator $\hat{\theta}(x)$ e.g., ML estimator.
- We characterize statistical error as the fluctuations of $\hat{\theta}(x)$, computed over a very long series of independent experiments where the parameters are kept fixed.

Bayesian

- No true values of parameters is assumed and we deal with their prior probability distribution $p(\theta)$.
- We compute posterior probability $p(\theta|x)$ from the prior $p(\theta)$ and likelihood $p(x|\theta)$.
- Statistical error in a single experiment is given by the spread of the posterior distribution $p(\theta|y)$,

Bayesian Analysis

Two rules of probability

- Sum Rule:

$$P(A|B) + P(\bar{A}|B) = 1 \tag{4}$$

- Product Rule:

$$P(A, B|C) = P(A|C)P(B|A, C) = P(B|C)P(A|B, C) \tag{5}$$

Bayes' Theorem

Substituting $A = \theta, B = y, C = I$ we get :

$$P(\theta|y, I) = \frac{P(y|\theta, I)P(\theta|I)}{P(y|I)} \tag{6}$$

with

$$P(y|I) = \sum_i P(y|\theta_i, I)P(\theta_i|I) \tag{7}$$

to make sure that

$$\sum P(\theta_i|y, I) = 1 \tag{8}$$

The two problems in statistical inference

- Model selection: Which of two or more competing models is most probable given our present state of knowledge?
- Parameter estimation: Assuming the truth of a model, find the probability density function for each of its parameters.

Information gain in Bayesian analysis

In Bayesian parameter estimation the amount of information gain-the information in the posterior relative to the prior is:

$$I = \int P(\theta|y) \ln \left(\frac{P(\theta|y)}{P(\theta)} \right) \tag{9}$$

Fisher Matrix : An introduction

- Close to the global maximum we can expand the log likelihood in Taylor series:

$$\ln \mathcal{L}(\theta) = \ln \mathcal{L}(\theta_0) + \frac{1}{2} \sum_{ij} (\theta_i - \theta_{0,i}) \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \mid_{\theta=\theta_0} (\theta_j - \theta_{0,j}) + ... \quad (10)$$

- By truncating this expansion to the quadratic term we assume that the likelihood surface is locally a multi-variate Gaussian.
- The Hessian matrix is defined as:

$$\mathcal{H}_{ij} = \frac{\partial^2 \ln \mathcal{L}}{\partial \Theta_i \partial \theta_j} \quad (11)$$

- The Hessian matrix encloses the information of the parameters errors and their covariance.

Fisher Matrix

- The Fisher matrix plays a fundamental role in forecasting errors from a given experimental set up and is defined as:

$$\mathcal{F}_{ij} = -\left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_i \partial\theta_j} \right\rangle = \langle \mathcal{H} \rangle \,. \tag{12}$$

  Where average is over an ensemble of observational data:

$$\mathcal{F}_{ij} = \int p(y|\theta)\mathcal{H}_{ij} dy \tag{13}$$

- Let us consider the case of one parameter : θ_i for which we can write:

$$\Delta \ln \mathcal{L} = -\frac{1}{2}\mathcal{F}_{ij}(\theta_i - < \theta_i >)^2 \tag{14}$$

  and identifying $2\Delta \ln \mathcal{L} = 1$ to $\Delta\chi^2$ corresponding to 68% CL we see that $1/\mathcal{F}_{ii}$ yields 1σ displacement for Θ_i.
- In general:

$$\sigma_{ij}^2 \geq (\mathcal{F}^{-1})_{ij}. \tag{15}$$

  This is called Crammer-Rao bound.

Fisher Matrix - Crammer-Rao bound

- For an unbiased estimator

$$\left\langle \hat{\theta}(y) - \theta \right\rangle = \int dy[\hat{\theta}(y) - \theta]p(y|\theta) = 0 \tag{16}$$

- Differentiating Eq.(16) wrt θ

$$\int dy[\hat{\theta}(y) - \theta]\frac{\partial p}{\partial \theta} - \int dy p(y|\theta) = 0 \tag{17}$$

We can use

$$\frac{\partial p}{\partial \theta} = p\frac{d\ln p}{\partial \theta} \quad \text{and} \quad \int dy p(y|\theta) = 1 \tag{18}$$

and get

$$\int dy[\hat{\theta}(y) - \theta]p\frac{d\ln p}{\partial \theta} = 1 \tag{19}$$

Using Schwartz inequality:

$$\left[\int dy \left(\frac{d\ln p}{\partial \theta}\right)^2 p(y|\theta)\right]\left[\int dy \left(\hat{\theta}(y) - \theta\right)^2 p(y|\theta)\right] \geq 1 \tag{20}$$

Fisher Matrix - Crammer-Rao bound

- We can also write :

$$Ie^2 \geq 1 \quad \text{or} \quad e^2 > \frac{1}{I} \tag{21}$$

where

$$I = \left[\int dy \left(\frac{d \ln p}{\partial \theta} \right)^2 p(y|\theta) \right] \tag{22}$$

is the Fisher information matrix and

$$e^2 = \left[\int dy \left(\hat{\theta}(y) - \theta \right)^2 p(y|\theta) \right] \tag{23}$$

is the Mean-squared error.

Fisher Matrix : two definitions

- We use :

$$\frac{\partial}{\partial\theta}\int p(y|\theta)dy = \int \frac{\partial}{\partial\theta}p(y|\theta)dy \tag{24}$$

$$\int \frac{\partial^n}{\partial\theta^n}p(y|\theta)dy = \frac{\partial^n}{\partial\theta^n}\int p(y|\theta)dy = 0 \tag{25}$$

- The mean of "score function" is zero.

$$\left\langle \frac{\partial}{\partial\theta}\ln p(y|\theta)\right\rangle = \int \frac{\partial}{\partial\theta}p(y|\theta) = 0 \tag{26}$$

- Now :

$$\begin{aligned}\mathcal{H} &= -\left\langle \frac{\partial^2 \ln p(y|\theta)}{\partial\theta^2}\right\rangle = -\int p(y|\theta)\frac{\partial}{\partial\theta}\left[\frac{1}{p(y|\theta)}\frac{\partial p(y|\theta)}{\partial\theta}\right]dy \\ &= \int \frac{1}{p(y|\theta)}\left[\frac{\partial p(y|\theta)}{\partial\theta}\right]^2 dy - \int \frac{\partial^2 p(y|\theta)}{\partial\theta^2}dy \\ &= \int p(y|\theta)\left[\frac{\partial \ln p(y|\theta)}{\partial\theta}\right]^2 dy \\ &= -\left\langle\left[\frac{\partial \ln p(y|\theta)}{\partial\theta}\right]^2\right\rangle \end{aligned} \tag{27}$$

Fisher Matrix : Computation

- Write down the likelihood for the data given the model.
- Instead of the data values (which are not known) use the theory prediction for a fiducial model.
- In the covariance matrix include expected experimental errors.
- Take derivatives with respect to the parameters.

Exercise

Show that for

1. Binomial probability distribution:

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{28}$$

Fisher matrix:

$$\mathcal{F}(p) = \frac{n}{p(1-p)} \tag{29}$$

2. For Poisson distribution:

$$P(x; \lambda) = \frac{\lambda^x e^{-x}}{x!} \tag{30}$$

Fisher matrix:

$$\mathcal{F}(\lambda) = \frac{1}{\lambda} \tag{31}$$

3. For Gaussian $\mathcal{N}(\mu, \sigma^2)$ Fisher matrix:

$$\mathcal{F}(\sigma) = \mathrm{dig}(1/\sigma^2, 2/\sigma^2) \tag{32}$$

Jeffery's prior

- The Jeffery's prior, is a non-informative (objective) prior distribution on parameter space and is related to determinant of the Fisher information:

$$p(\theta) \propto \sqrt{|\mathcal{F}(\theta)|} \tag{33}$$

- Jeffreys' prior has the key feature that it is invariant under reparameterization of the parameter vector θ.

$$p(\theta) \propto \sqrt{I(\theta)}$$

using the change of variables theorem and the definition of Fisher information:

$$p(\varphi) = p(\theta)\left|\frac{d\theta}{d\varphi}\right| \propto \sqrt{I(\theta)\left(\frac{d\theta}{d\varphi}\right)^2} = \sqrt{\mathrm{E}\left[\left(\frac{d\ln L}{d\theta}\right)^2\right]\left(\frac{d\theta}{d\varphi}\right)^2}$$

$$= \sqrt{\mathrm{E}\left[\left(\frac{d\ln L}{d\theta}\frac{d\theta}{d\varphi}\right)^2\right]} = \sqrt{\mathrm{E}\left[\left(\frac{d\ln L}{d\varphi}\right)^2\right]} = \sqrt{I(\varphi)}.$$

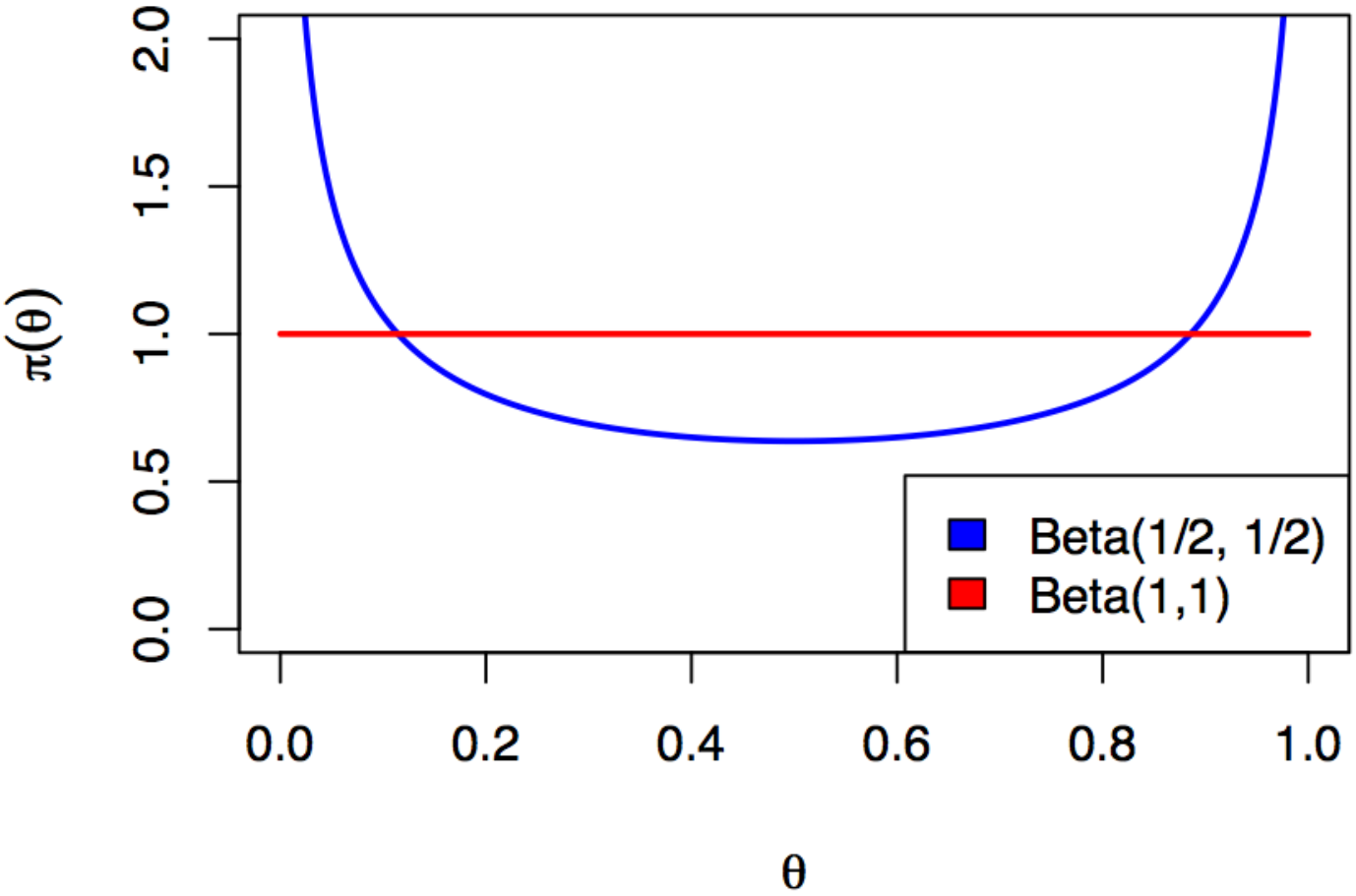

Figure 1: Jeffreys prior and flat prior densities

Fisher Information and Shannon Information

- Shannon entropy is defined as:

$$S = -\sum_{i=1}^{N} p_i \ln pi, \tag{34}$$

or in continuous form :

$$S = -\int dy p(y|\theta) \ln p(y|\theta) \tag{35}$$

- Both Fisher information I and Shannon information S are measures of probability distribution $p(y|\theta)$.
- Fisher information I is a local measure, however, S is a global once.

Fisher Information matrix as Geometric object

- A set of distributions $p_\theta = p(\theta)$ parametrized by θ_i with $i \in 1, ..., n$ is a "manifold" and Fisher information matrix can be identified with the Riemanian metric on this manifold:

$$g_{ij} = \int dx p(x|\theta) \left[\frac{\partial \ln p(x|\theta)}{\partial \theta_i}\right] \left[\frac{\partial \ln p(x|\theta)}{\partial \theta_j}\right] \tag{36}$$

- The Kullback-Leibler distance or relative entropy between two distributions $p(x)$ and $q(x)$ is defined as:

$$D(p||q) = \int dx p(x) \ln \frac{p(x)}{q(x)} \tag{37}$$

- We can related the Kullback-Leibler distance and Fisher information metric in the following way [1]:

$$D(p(x|\theta)||p(x|\theta + \delta\theta)) = \sum_{i,j} g_{ij} \delta\theta^i \delta\theta^j \tag{38}$$

[1] Jeffreys, H. (1946). *An invariant form for the prior probability in estimation problems*. Proc. Royal Soc. of London, Series A, 186, 453-461.

Fisher Information & Gravitational waves

- Fisher information is often employed as a proxy for the amount of physical information that can be gained from detection campaigns.
- Fisher matrix is defined as:

$$\mathcal{F}_{ij}(h) = (h_i, h_j) \tag{39}$$

where $h_i(t)$ is the derivative of the gravitational waveform $h(t)$ of interest with respect to the i^{th} source parameter ${}_o$i, and (,) is a signal product weighted by the expected power spectral density of detector noise.

- $\mathcal{F}_{ij}^{-1}(h)$ represents the covariance matrix of parameter errors in the parameter-estimation problem for the true signal h_0.
- Likelihood for GW signals in Gaussian noise case can be written :

$$p(s|\theta) \propto \exp[-(s - h(\theta), s - h(\theta))/2] \tag{40}$$

The case of gravitational wave signal

- The data/signal is given by:

$$s = n + h_0 \tag{41}$$

- The likelihood function:

$$p(s|\theta) \propto \exp[- < s - h(\theta)|s - h(\theta) > /2] \tag{42}$$

- Expand template about θ_0 i.e., linearized-signal approximation (LSA):

$$h(\theta) = h_0 + \Delta\theta^i h_i + ... \tag{43}$$

- The posterior :

$$p(\theta|s) \propto p(\theta) \exp\left[-\frac{1}{2} < n|n > + \Delta\theta^k < n|h_k > -\frac{1}{2}\Delta\theta^i \Delta\theta^j < h_i|h_j >\right] \tag{44}$$

Fisher matrix and GW community

- The use of Fisher matrix is very common in GW community.
- GW community is unhappy with Fisher matrix.
- Errors on estimated parameters in Bayesian analysis (MCMC) are far smaller than what are expected from Fisher analysis.
- *Fisher matrix can be a poor predictor of the amount of information obtained from typical observations, especially for waveforms with several parameters and relatively low expected signalto- noise ratios (SNR), or for waveforms depending weakly on one or more parameters, when their priors are not taken into proper consideration.* [Vallisneri M., PRD (2008), **77**, 042001]

- *a direct comparison between the FIM error estimates and the Bayesian probability density functions produced by the parameter estimation code LALINFERENCE_MCMC...we find that the FIM can greatly overestimate the uncertainty in parameter estimation achievable by the MCMC. This was found to be a systematic effect for systems composed of binary black holes, with the disagreement increasing with total mass.* [Rodriguez et. al, PRD (2013) **88** 084013.]
- *Fisher matrix approach sometimes predicts errors of* $\geq$ *100% in the estimation of parameters such as the luminosity distance and sky position. ..we conduct a Bayesian inference analysis for 120 sources situated at redshifts of between* $0.1 \leq z \leq 13.2$, *and compare the results with those from a Fisher matrix analysis. The Fisher matrix results suggest that for this particular selection of sources, eLISA would be unable to localize sources at redshifts of* $z \leq 6$. *In contrast, Bayesian inference provides finite error estimations for all sources in the study, and shows that we can establish minimum closest distances for all sources.* [Porte et. al. 2015, arXix:1502.05735v1]

Thank You !

References

1. Frieden B. R., 1998, *Physics from Fisher Information*, Cambridge.
2. Gregory P., 2005, *Bayesian Logical data analysis for physical sciences*, Cambridge.
3. Jeffreys H., 2003, *Theory of probability*, Oxford.
4. Jaranovski & Krolak, 2009, *Analysis of Gravitational Wave data*, Cambridge.
5. Dodelson C., 2003, *Modern Cosmology*, Academic Press.
6. Verde L., *Statistical methods in cosmology*, Lect.Notes Phys.800:147-177,2010.
7. Tegmark M.,*How to make maps from CMB data without losing information*, Astrophys.J.480:L87-L90,1997.
8. Tegmark M., *How to measure CMB power spectra without losing information*, Phys.Rev.D55:5895-5907,1997.
9. Vallisneri M., PRD (2008), **77**, 042001.
10. Rodriguez et. al, PRD (2013) **88** 084013.
11. Porte et. al. 2015, arXix:1502.05735v1.