APPLIED (STATISTICAL) SIGNAL PROCESSING
LECTURE NOTES

Lecturer: Saleem Zaroubi

CONTENTS

1 GENERAL INFORMATION

TEACHER :
Prof. Saleem Zaroubi,
Kapteyn Astronomical Institute,
Room 141, Tel: 3634055,
Email: saleem@astro.rug.nl

TEACHING ASSISTANT :
Mr. Sarrvesh S. Sridhar,
Kapteyn Astronomical Institute,
Room 183, Tel: 3634083,
Email: sarrvesh@astro.rug.nl

COURSE DESCRIPTION: The aim of this course is to introduce the students to the basics of Statistical Signal Processing with emphasis on the application of this field to data and image analysis. Such methods play a crucial role in the analysis and interpretation of data in almost every field of science. The course will present the general mathematical and statistical framework of Statistical Signal Processing with special emphasis on examples from Astronomy and physics. The course will cover the topics of Random vectors and processes, Estimation theory, Moments analysis, Filtering and Sampling theory. Problem sets and computer assignments are substantial and integral part of the course.

LITERATURE: The course will rely on the lecture slides/notes and will not follow a specific book in detail. However, there are two books that give a general overview of the material that will be covered in the course.

1. Fundamentals of Statistical Signal Processing: Estimation Theory v. 1 (Prentice Hall Signal Processing Series), by Steven M. Kay
2. Discrete Random Signals and Statistical Signal Processing (Prentice Hall Signal Processing), by Charles W. Therrien

COURSE HOURS PER WEEK: 4 hours of lecture and 2 hours of tutorial

GRADING: The homework assignments and computer project are an integral and mandatory part of the course. The student must submit a substantial number of the homework (80%) and all computer projects in order to be allowed in the final exam. The final grade will be composed as follows: 60% Homework and Computer projects 40% Final exam.

2 INTRODUCTION

In experimental and observational sciences one is often faced with the task of drawing conclusions about nature from uncertain, incomplete and often complex data. Signals are almost always contaminated by errors which are statistical in nature; their relation to the underlying physical theory can be complicated, especially in sciences like astronomy where one can not isolate the physical processes that one wants to study, and in many cases the nature of the underlying signal itself is fundamentally of statistical nature, like in quantum theory. Therefore, we are forced to model and study such signals in probabilistic and statistical fashion.

One can bring many examples here but I'll focus on examples from physics and astronomy. Let us assume we want to measure the speed of a car that is traveling in a straight line and departs from point 1 to point 2 and we measure its position at different times. Obviously, all what we know about the car motion are the distances we measure. Its motion, if it was moving in a constant speed, is given by the black line, which we normally model as:

$$D = D_0 + vt \tag{1}$$

The example is shown in Fig. 1 where the red crosses show the measurement and the black line is the underlying true motion, in case the speed is constant. We obviously would like to estimate the speed and the initial distance (our parameters) from the data. The main causes to the random nature in such case are the measurement errors.

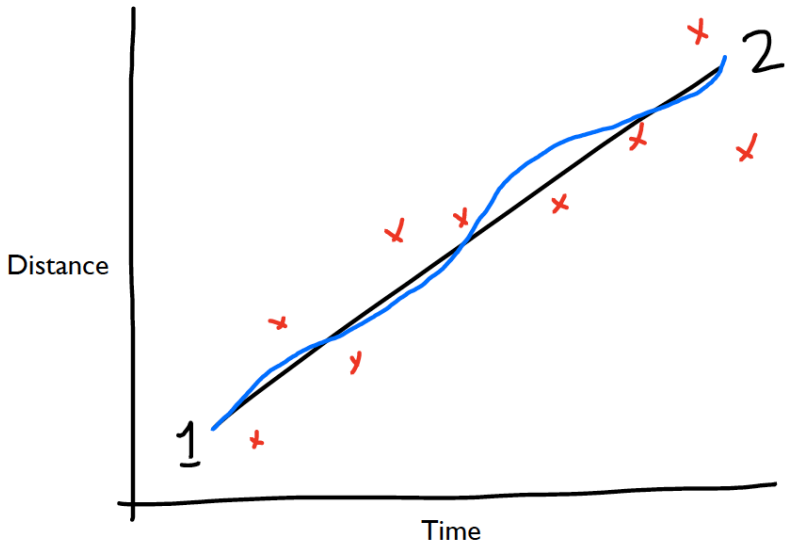

Figure 1: Example: Measurement of distance as a function of time

Another situation that reflects more what often happens in nature is that the underlying speed of the car is almost constant but not exactly, rather it has a stochastic component due to changes that have to do with many factor (the driver, the road conditions, the inclination of the road, the traffic, etc.). In such a case the statistical nature of the problem is more fundamental. In fact, in most cases in nature the underlying signal is stochastic and the best we can do is that we model it under a certain set of assumptions.

The Statistical Signal Processing Problem:

In general, every statistical signal processing problem has a number of essential elements, which we list below:

MEASUREMENT: One normally measured a quantity (or more), say x, which is measured N times . Hence the vector $\mathbf{x} = \{x[1], x[2], \ldots, x[N]\}^T$ represents the data.

MODELING: This gives the statistical model that describes the relation between the underlying quantities, normally parameters' vector $\boldsymbol{\theta}$, and the data. In probabilistic terms this can be written as $p(\mathbf{x};\boldsymbol{\theta})_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}$. Here $\boldsymbol{\Theta}$ is the space from the parameters are drawn. Notice that besides $\boldsymbol{\theta}$ we assume that we know the PDF (Probability Density Function) of the process.

Inference: gives the best value of $\boldsymbol{\theta}$ best fits the data. This generally has a number of components which often are referred to as: 1. Detection and parameters estimation (both can be viewed as "estimation"); 2. Prediction (inferring the value of a signal $\mathbf{y}$ given an observation of a related, yet different, signal, $\mathbf{x}$; 3. Learning which means that we learn about a relation between two stochastic signals $\mathbf{x}$ and $\mathbf{y}$ from the data we have.

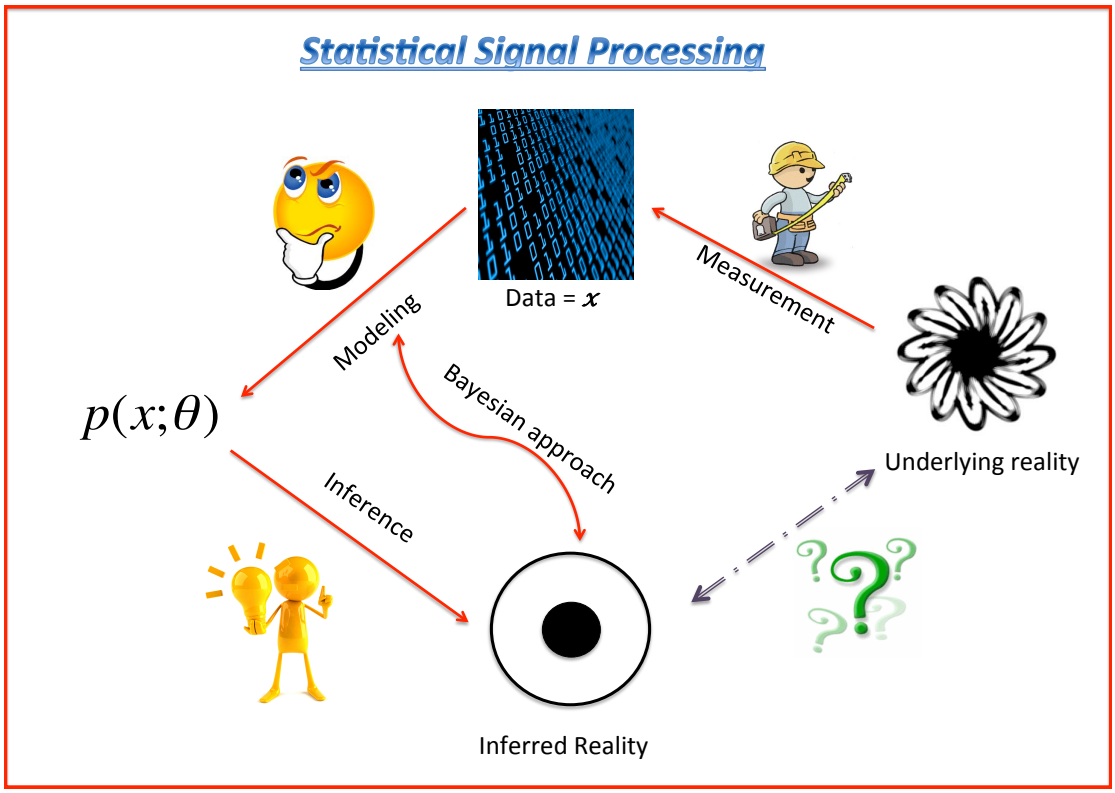

Figure 2: Example: Measurement of distance as a function of time

We wish to determine $\boldsymbol{\theta}$ for a given the data vector $\boldsymbol{x}$. This is normally given by a, so called, estimator of $\boldsymbol{\theta}$ denoted by $\hat{\boldsymbol{\theta}} = g(\boldsymbol{x})$ where g is some function. This is basically the problem of parameter estimation.

The first step in devising a good estimator, $\hat{\boldsymbol{\theta}}$, is to mathematically model the data. Since, as mentioned earlier the data is inherently random we describe it using a parameterized PDF, $p(\mathbf{x};\boldsymbol{\theta})$. Notice that this is a major decision about modeling the data as one not only decides how to choose the parameter but more importantly to which class of PDFs our model belongs.

As an example, consider the problem of measuring distance to the Galactic center. Assume that this measured through some distant indicator and one has N such measurements $x[n]$ where $n \in 1, \ldots, N$. Assume that the parameter one wants to estimate is the real distance to the galactic center, θ, then one can write the PDF of each measurement as:

$$p(x[n];\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n]-\theta)^2\right], \quad (2)$$

where here σ is the standard deviation of the measurement. Assuming that each measurement is independent the total PDF is hence given by,

$$p(\boldsymbol{x};\theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x[n]-\theta)^2\right]. \tag{3}$$

Obviously, in practice we are not given the PDF but we must choose it in a manner that is consisted with the problems constraints but also with the prior knowledge we have about the problem. For example, the distant to the Galactic center cannot be negative and should be roughly of the order of magnitude of 1 or 10 kpc. If the model you assume give you results that violate these constraints then it means either the measurement, the model or both are wrong.

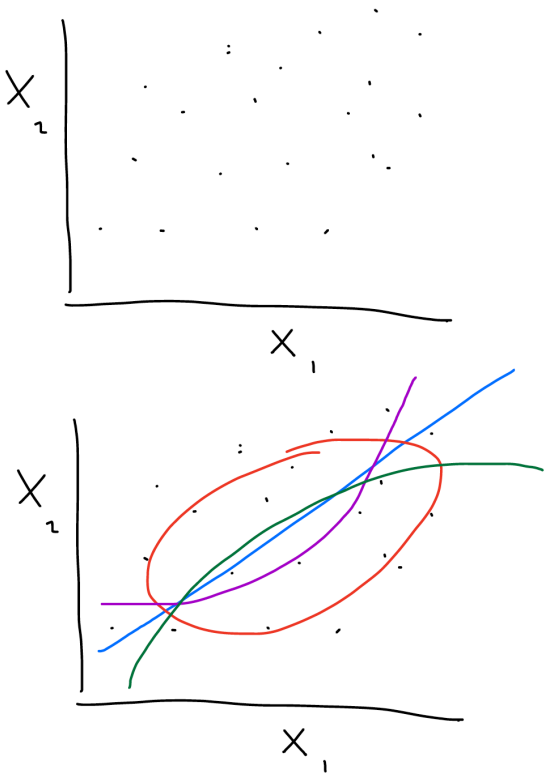

Figure 3: Example: Finding the best model that fits the data. All the models drawn can potentially be good fit for the data, yet at the same time they all can be very bad fits for the data. This all depends on the error or uncertainty that each point carries.

One sometimes also encounter a case where a number of models can be chosen and can fit the data. In Fig. 3 such an example of shown. The data that relates X_1 to X_2 shown in the upper panel can be fit by a number of models. Depending on the measurement uncertainty of each of the data points, all these models can be equally good, or conversely, equally bad. What decides this are the data and our knowledge of the physical system we are measuring. Furthermore, one would like also to be able to judge whether the inferred results one obtains from the data about the system under exploration are reliable, i.e., whether the model that has been used is a good model.

In case the parameters are of statistical nature themselves, then it make sense to write the joint PDF

$$p(\boldsymbol{x};\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{4}$$

where $p(\boldsymbol{\theta})$ is the prior PDF which summarizes our knowledge about $\boldsymbol{\theta}$ before the data has been taken, and $p(\boldsymbol{x}|\boldsymbol{\theta})$ is the conditional PDF which gives the probability of the data given certain value of the parameters.

This is the heart of the so called, Bayesian estimation, which incorporates the previous knowledge about the quantity we want to estimate, together with the current data.

Estimators and their performance:

Consider the data shown in Fig. 4 of repeated measurement of distance between two moving points. Where N stands for the number of measurements and index i indicates the measurement number number. It seems that such data can be modeled as a constant distance measured with noise, i.e.,

$$x[i] = d_0 + w[i] \qquad i = 1, 2, \ldots, N \tag{5}$$

Where $w[i]$ stands for white Gaussian noise with zero mean and variance of σ^2, such white Gaussian noise PDF is normally written as $\mathcal{N}(0, \sigma^2)$.

In order to estimate the value of d_0 we can choose a number of, so called, estimators. **An estimator is a certain mathematical operation or rule that is applied to the data in order to obtain the desired quantity.** The most natural estimator in this case is the mean of the sample data which can be written as,

$$\hat{d_0} = \frac{1}{N} \sum_{i=1}^{N} x[i], \tag{6}$$

where the sign $\hat{d_0}$ indicates that this is an estimator and the underlying value. Now for each such estimator we have to ask how close is it to the real value and whether this is the best estimator.

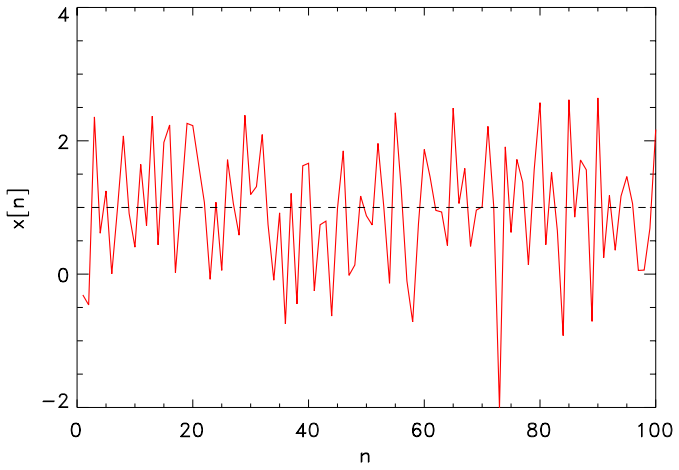

Figure 4: Measurement of a fixed distance with errors

In order to determine how close our estimator to the real value one should explore the statistical properties of the estimator. The simplest two statistics to inspect are the expectation value of this estimator and its variance (i.e. error). Assuming that the measurement is unbiased, i.e., the mean of the errors one makes in each measurement is zero, one can write the estimator as follows:

$$E(\hat{d_0}) = E\left(\frac{1}{N} \sum_{i=1}^{N} x[i]\right) = \frac{1}{N} \sum_{i=1}^{N} E(x[i]) = d_0 \tag{7}$$

so that the average of the estimator produces the correct value. This is good because it means our estimator is not biased and fully uses the data at hand.

Now one can also calculate the variance of the estimator, in this case it yields the following result,

$$\text{(8)} \quad \mathrm{Var}(\hat{d}_0) = \mathrm{Var}\left(\frac{1}{N}\sum_{n=1}^{N} x[i]\right) = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(x[i]) = \frac{\sigma^2}{N},$$

which is also good because it means that not only we get the average right but we also get the variance smaller the more we add independent measurements.

PROBLEM: *Consider the following estimator of* d_0*,*

$$\check{d}_0 = \frac{1}{N+2}\left(2x[1] + \sum_{i=2}^{N-1} x[i] + 2x[N]\right).$$

Do you think this is a better, worse or equal estimator than the one given in Eq. 6 and why?

To summarize, a *good* estimator should be unbiased and converges to the real value when more data points exist. Of course the later point makes sense, the more information one has about the parameter the more accurately one can estimate it. Obviously, this is correct only if the assumptions we assume hold all the time. During the course we will address the topic of *best* estimators in a more detailed manner.

The data model:

A general assumption we often assume is that the relation between the measured data and the underlying quantity that one measures is linear, i.e.,

$$\text{(9)} \quad \boldsymbol{x} = \boldsymbol{R}\boldsymbol{s}(\boldsymbol{\theta}) + \boldsymbol{\epsilon},$$

where $\boldsymbol{s}$ the vector of the quantity we would like to measure, e.g., the luminosity of a star, $\boldsymbol{R}$ is a matrix that encapsulates the response function of the experimental/observational apparatus, e.g., the point spread function, and $\boldsymbol{\epsilon}$ is the noise vector. Please notice that this equation may depend on the underlying parameters, that we are primarily interested in, in a very complicated and nonlinear manner. We'll come back to this equation many times during the course.

Here is an example of how data improves with time and how previous knowledge can be incorporated in the estimates of the new data.

In the following section I will remind you of the some of the mathematical topics needed in order to proceed with the course.

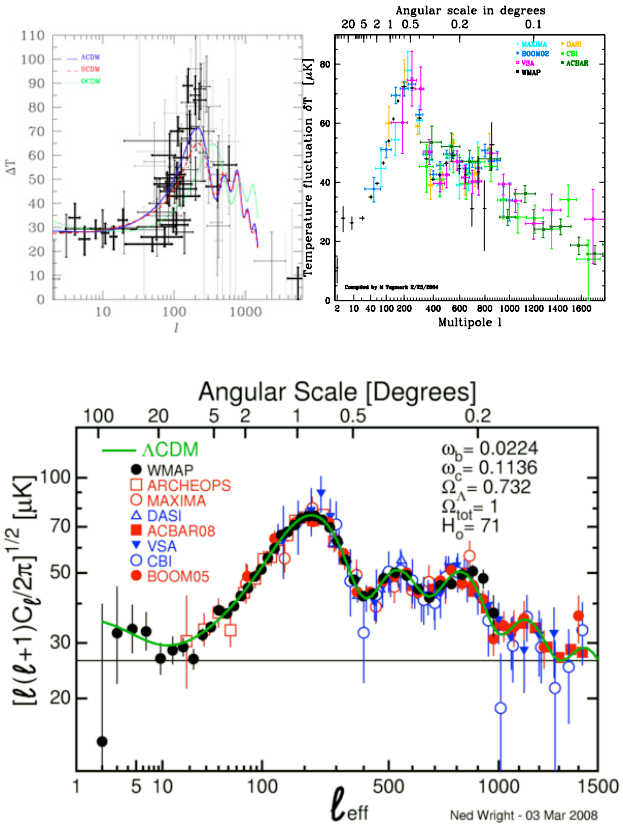

Figure 5: The Evolution of the CMB data and models with time.

3 MATHEMATICAL PRELIMINARIES

3.1 *Random vectors and Processes*

A random vector is a vector in which all the components are random variables. Let $\mathbf{v}$ be such a vector, then for a given specific value of this random vector v_0 we can define the Cumulative Distribution Function (CDF) as the probability of $\mathbf{v}$ to be smaller than v_0 component by component.

$$F(\mathrm{v}_0) = Pr(\mathbf{v} \leq \mathrm{v}_0). \tag{10}$$

The left panel of Fig. 6 show an example of such CDF of a function of two variables in x and y which is given in this case by $(1+erf(x/5))(1+erf(y/10))/4$.. This example shows the general properties of a CDF which we list here as a reminder.

1. $F(-\infty) = 0$ and $F(\infty) = 1$
2. F is a non decreasing function; i.e., in $a \leq b$ then $F(a) \leq F(b)$.

While CDF is a meaningful and reasonable function with which to understand random vectors another function is more often used, the so called Probability Density Function (PDF). The PDF is simply defined as,

$$p(x) \equiv \lim_{\Delta x \to 0} \frac{Pr(x < X < x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x}. \tag{11}$$

The PDF function is non-negative, i.e., $p(x) \geq 0$ for every x) and in normalized, i.e., $\int_{-\infty}^{\infty} p(x)dx = 1$.

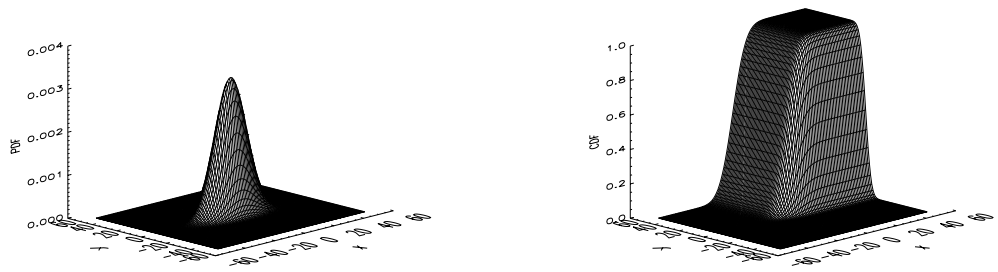

Figure 6: Examples of the joint CDF and PDF of x and y.

There are a lot of examples of widely used PDF and CDF functions, like, binomial, Poisson distribution, Gaussian (normal) distribution, χ^2, etc. The distribution that is most widely used is the Gaussian distribution (though all of these distribution are commonly used). In Fig. 6 we show a Gaussian distribution with two variables x and y (left panel) and its CDF. The PDF of one variable is given by:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-x_0)^2}{2\sigma^2}}. \quad (12)$$

Notice that there are two free parameters in this distribution mean x_0 and the standard deviation σ (show that!!)

If the PDF is a function of a number of variables, say $\mathbf{x}$ and $\mathbf{y}$ then the joint CDF and PDF give the cumulative and density probabilities of the two vectors occurring and is normally denoted $p(\mathbf{x}, \mathbf{y})$. For example the probability:

$$p(x,y) = \frac{1}{\sqrt{2\pi\sigma_1^2}\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-x_0)^2}{2\sigma_1^2}} e^{-\frac{(y-y_0)^2}{2\sigma_2^2}}, \quad (13)$$

for which we show the PDF and CDF in Fig. 6.

In this case, $p(\mathbf{x}, \mathbf{y})$ is separable, namely it can be written as $p(\mathbf{x})p(\mathbf{y})$. However, more often the joint probability is not separable. For examples:

$$p(x,y) \propto \exp\left(-\frac{x^2}{2\sigma_1^2} - \frac{xy}{\sigma_{12}^2} - \frac{y^2}{2\sigma_2^2}\right). \quad (14)$$

It is easy to see that the PDF of $\mathbf{x}$ can be derived from the joint PDF as follows

$$p(\mathbf{x}) = \int_{\infty}^{\infty} p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y}, \quad (15)$$

It is actually easy to see this in the case of a separable joint PDF but one can also see how this can be generalized. This operation of integrating the PDF over a certain random variable in order to derive a PDF for the other variables is called **marginalization** and the resultant PDF is called the **marginal** PDF.

3.2 *Conditional probabilities and Bayes' theorem*

Let A and B are two random events. The conditional probability of B provided event A has occurred is normally denoted as $Pr[B|A]$. There is a simple relation that one can derive that states that the joint probability of A and B to happen, $Pr[A, B]$ is given by the probability of B to occur provided A has occurred which has to be modified by the probability of A to occur, namely, $Pr[A, B] = Pr[B|A]Pr[A]$.

From this one can show that

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{y})}, \tag{16}$$

Notice that each p in this equation can be a completely different function. In case of events $\mathbf{x}$ and $\mathbf{y}$ are statistically independent then $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ which naturally gives the result $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})$ as excepted in such a case.

The conditional probability brings us to one of the most important rules in probability theory called Bayes' rule. This rule is derived from the fact that $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$, which gives,

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}. \tag{17}$$

The interpretation of this rule is very hotly debated between two schools of thought in statistical interpretation, the so-called *Bayesian* and *frequentist* interpretation.

Another way to represent this rule is,

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{\int_{-\infty}^{\infty} p(\mathbf{x}|\mathbf{y})p(\mathbf{y})\mathrm{d}\mathbf{y}}. \tag{18}$$

As I mentioned earlier, this is a very useful equations 17 and 18 arise frequently in problems of statistical inference where often, $\mathbf{y}$ denotes a random vector that can not be observed to measured whereas the random vector $\mathbf{x}$, which is related to $\mathbf{y}$, *can* be measured. In this context $p(\mathbf{y})$ is called the *prior* PDF (namely the PDF before the measurement of $\mathbf{x}$) and $p(\mathbf{y}|\mathbf{x})$ is called the *posterior* (the PDF after the measurement of $\mathbf{x}$). Bayes' rule is used to update the knowledge we have about $\mathbf{y}$ after the measurement of vector $\mathbf{x}$.

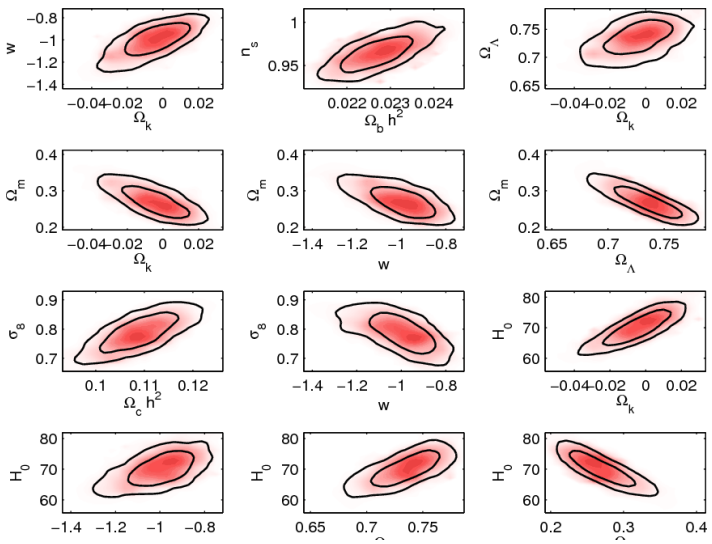

Figure 7: An example of the parameters that are evaluated from the CMB data. The parameters presented in each panel are obtained by marginalizing over the distribution over the rest of the parameters.

An example of such an application is as follows: A two dimensional random vector has the pdf

$$p(\mathbf{y}) = A_1\delta^D(\|\mathbf{y}-\mathbf{a}_1\|) + A_2\delta^D(\|\mathbf{y}-\mathbf{a}_2\|), \tag{19}$$

where $A_1 + A_2 = 1$ and $\mathbf{a}_1$ and $\mathbf{a}_2$ are constant vectors. Now the vector $\mathbf{x}$ has the conditional probability,

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{\pi} e^{-\|\mathbf{x}-\mathbf{y}\|^2} \quad (20)$$

Calculate from this $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$ and then get $p(\mathbf{x})$ by marginalizing over $\mathbf{y}$ which gives,

$$p(\mathbf{x}) = \frac{1}{\pi}\left[A_1 e^{-\|\mathbf{x}-\mathbf{a}_1\|^2} + A_2 e^{-\|\mathbf{x}-\mathbf{a}_2\|^2}\right]. \quad (21)$$

We finally reach the *posterior* PDF,

$$p(\mathbf{y}|\mathbf{x}) = A_1'(\mathbf{x})\delta^D(\|\mathbf{y}-\mathbf{a}_1\|) + A_2'(\mathbf{x})\delta^D(\|\mathbf{y}-\mathbf{a}_2\|), \quad (22)$$

where

$$A_i'(\mathbf{x}) = \frac{A_i e^{-\|\mathbf{x}-\mathbf{a}_i\|^2}}{A_1 e^{-\|\mathbf{x}-\mathbf{a}_1\|^2} + A_2 e^{-\|\mathbf{x}-\mathbf{a}_2\|^2}}. \quad (23)$$

Notice that $A_1'(\mathbf{x}) + A_2'(\mathbf{x}) = 1$.

3.3 *Expectations and Moments*

The expectation value of a random vector, $\mathbf{x}$ is defined as

$$E\{\mathbf{x}\} = \langle x \rangle = \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) \mathrm{d}\mathbf{x}. \quad (24)$$

This is actually the quantity that we measure when we average data and is often called $\boldsymbol{\mu}_{\mathbf{x}} = E\{\mathbf{x}\}$. Notice that this expectation value is not the same as the Median (which is the point at which the CDF=1/2) or the Mode (which is the maximum point of the PDF, which is also not unique). However, in some distributions these three quantities are the same, e.g., Gaussian PDF.

This concept can be made more general and one can define the *moment* n of a give distribution by,

$$E\{\mathbf{x}^n\} = \int_{-\infty}^{\infty} \mathbf{x}^n p(\mathbf{x}) \mathrm{d}\mathbf{x}. \quad (25)$$

The most used moments are the first (mean) and second order moments of the distribution. However, the 3^{rd} and 4^{th} moments are also often used (in relation to the so called skewness and kurtosis, respectively).

The variance is also a widely used statistic for a single random variable and is defined as:

$$\mathrm{Var}\{x\} = E\{x^2 - E\{x\}^2\} = \int_{-\infty}^{\infty} (x^2 - \mu_x^2) p(x) dx. \quad (26)$$

A related statistic is of course the so called Standard Deviation, σ, and is given simply by $\sigma = \sqrt{\mathrm{Var}\{x\}}$.

As an example we take the following distribution, $p(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ where $\lambda > 0$. The mean of this distribution is given by,

$$E\{x\} = \int_0^{\infty} x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}\int_0^{\infty} \xi e^{-\xi} d\xi = \frac{1}{\lambda}\left[-(\xi+1)e^{-\xi}\right]_0^{\infty} = \frac{1}{\lambda}. \quad (27)$$

The variance of x is,

(28)

$$\mathrm{Var}\{x\} = \int_0^\infty (x - \frac{1}{\lambda})^2 \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2}\left(\int_0^\infty \xi^2 e^{-\xi} d\xi - 1\right) = \frac{1}{\lambda^2}\left(\left[-(\xi(\xi+2)+2)e^{-\xi}\right]_0^\infty - 1\right) = \frac{1}{\lambda^2}.$$

One can also show that the following theorem is easy to prove: **The variance of a sum of uncorrelated random variables is equal to the sum of the variances of these random variables.**, i.e.,

$$\mathrm{Var}\{\sum_{n=1}^{N} x[n]\} = \sum_{n=1}^{N} \mathrm{Var}\{x[n]\} \quad (29)$$

Now if the random variable is a vector one can see that the definition of these momenta has to be more accurate. A widely used quantity is the so called correlation function or matrix (in the case of discrete data which is always the case) which is defined as $E\{\mathbf{x}\mathbf{x}^{*T}\}$. We'll discuss this more in Section **??**.

3.4 *Transformation with single valued inverses:*

If x is a contentious random variable with PDF $p(x)$ and $y = g(x)$ is a single valued differentiable (analytical) function then the PDF of y is given by,

$$p(y) = p(x)\left|\frac{dx}{dy}\right| = p(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|. \quad (30)$$

The term in differentiation is called the Jacobian of the transformation.

As an example consider x, a random variable drawn from a Gaussian PDF $\mathcal{N}(0,\sigma^2)$ and $y = x^2$, what is then $p(y)$? First one should calculate the Jacobian of the transformation,

$$\frac{dx}{dy} = \frac{1}{2x} = \frac{1}{\pm 2\sqrt{y}}\,, \quad (31)$$

remember y can only be positive which means that both $\pm x$ map to y, this will double the contribution to the PDF. Gence the PDF of y is

$$p_y(y) = 2.\frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}\frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{e^{-y/(2\sigma^2)}}{\sqrt{y}}. \quad (32)$$

Notice that the first 2 reflects that x is mapped twice to y. This is the so called $\chi^2(y,1)$, i.e., χ-squared distribution with one degree of freedom.

Now, one can generalize this to vector transformation, namely, $\mathbf{y} = \mathbf{g}(\mathbf{x})$, then the volume element of $\mathbf{x}$ is $V_\mathbf{x} = \Delta x_1 \Delta x_2 \Delta x_3 \ldots \Delta x_N$ which transforms to a volumes element $V_\mathbf{y} = \mathrm{J}(\mathbf{y},\mathbf{x})V_\mathbf{x}$ (see Figure 8).

The Jacobian for the transformation of random vector $\mathbf{x}$ to random vector $\mathbf{y}$ is then given by,

$$\mathrm{J}(\mathbf{y},\mathbf{x}) \equiv abs\begin{vmatrix} \frac{\partial g_1(\mathbf{x})}{dx_1} & \frac{\partial g_2(\mathbf{x})}{dx_1} & \cdots & \frac{\partial g_N(\mathbf{x})}{dx_1} \\ \frac{\partial g_1(\mathbf{x})}{dx_2} & \frac{\partial g_2(\mathbf{x})}{dx_2} & \cdots & \frac{\partial g_N(\mathbf{x})}{dx_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial g_1(\mathbf{x})}{dx_N} & \frac{\partial g_2(\mathbf{x})}{dx_N} & \cdots & \frac{\partial g_N(\mathbf{x})}{dx_N} \end{vmatrix}.$$

Hence, the distribution with the new vector $\mathbf{y}$ is,

$$p_y(\mathbf{y}) = \frac{1}{\mathrm{J}(\mathbf{y},\mathbf{x})} p_x(\mathbf{g}^{-1}(\mathbf{y})) \quad (33)$$

This is a very useful theorem and will be used extensively during the course, especially when we use linear transformations.

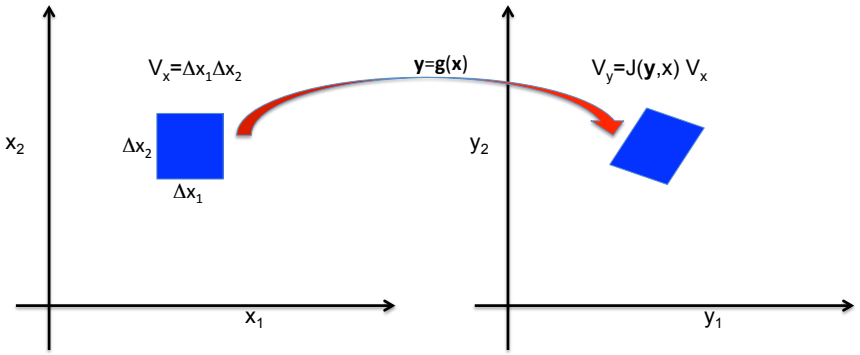

Figure 8: The interpretation of the transformation Jacobian, which relates the new volume element to the old volume element.

3.5 *Fourier Transforms (FT, DFT, FFT)*

i: Basic properties

Now we move to a topic that is not related to the previous issues we discussed but is needed at various points in the course. Frequently in physics one encounter integral transforms which are often very useful in various situations. The most useful and widely used such transform in physics, math, signal processing, etc., is the so called, Fourier Transform. This transform has the form,

$$\mathcal{F}\{f(t)\} = \tilde{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt, \tag{34}$$

and its inverse is defined as,

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{f}(\omega)e^{i\omega t}d\omega. \tag{35}$$

Notice that the inverse includes the factor $1/2\pi$. One is actually free to choose where to place this constant in the forward, inverse or both transformations. This is actually the source of a lot of confusion when one deals with FT. Here we'll stick to the way I have defined it.

Example Fourier Transform of a Gaussian field $\mathcal{N}(0,\sigma^2)$. One can show that,

$$\mathcal{F}\{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}\} = e^{-\frac{\omega^2\sigma^2}{2}}. \tag{36}$$

That is an FT of a Gaussian is also a Gaussian.

Here is a list of general properties of FT,

1. Linearity: $\mathcal{F}\{f+g\} = \mathcal{F}\{f\} + \mathcal{F}\{g\}$
2. Time shift: $\mathcal{F}\{f(t-t_0)\} = e^{-i\omega t_0}\mathcal{F}\{f(t)\}$
3. Modulation: $\mathcal{F}\{e^{i\omega_0 t}f(t)\} = \tilde{f}(\omega-\omega_0)$
4. Scaling: $\mathcal{F}\{f(at)\} = \frac{1}{|a|}\tilde{f}(\frac{\omega}{a})$
5. Conjugation: If $h(t) = f^*(t)$ then $\tilde{h}(\omega) = \tilde{f}(-\omega)$. If f is real then $\tilde{f}^*(\omega) = \tilde{f}(-\omega)$.
6. Convolution: Convolution is defined as $f(t) * g(t) \equiv \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$. In Fourier space convolution becomes multiplication, namely, $\mathcal{F}\{f(t) * g(t)\} = \tilde{f}\cdot\tilde{g}$

7. Parseval's theorem: $\int_{-\infty}^{\infty} f(t)g(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{f}^*(\omega)\tilde{g}(\omega)d\omega$. This relation gives $\int_{-\infty}^{\infty} f^2(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\tilde{f}(\omega)|^2 d\omega$

Fourier transform can be generalized to higher dimensions in which the same rules applying also to these dimensions.

In many applications in which the frequency spectrum of an analogue signal is required, the best that can be done is to sample the signal $f(t)$ a finite number of times at fixed intervals and then use a discrete Fourier transform to estimate discrete points on $\tilde{f}(\omega)$. In order to see which points we are looking discretize the transform and yet have the Fourier functions still orthogonal over the chosen points. This translates to a very specific requirement on the discretization in both spaces, real and Fourier.

Consider a set of N measurements done within time T with constant interval $T/N \equiv \Delta$, i.e., $t[j] = j\Delta$ where $j = 1, \ldots, N$. Notice that since we only sample N points in the function this limits the types of functions we can use in such a transformation so as to make representative of the normal FT. For every sampling interval Δ there is a special $\omega_c = \frac{2\pi}{2N}$ called the Nyquist frequency which gives the limit with which we know the signal (we'll discuss this at the end of the course when we discuss sampling theorem). Therefore we would like to sample the discrete Fourier space with frequencies up to Nyquist frequency $\omega[k] = \frac{2\pi k}{N\Delta}$ where $k = 1, \ldots, N$. Hence DFT at the frequency point k is defined as,

$$\tilde{f}[k] = \sum_{j=1}^{N} f[j] e^{-\frac{2\pi i j k}{N}} \tag{37}$$

DFT fulfills a number of properties like orthogonality,

$$\sum_{j=1}^{N} \left(e^{-i\omega_1 t[j]}\right)^* e^{-i\omega_2 t[j]} = \sum_{j=1}^{N} e^{-\frac{2\pi i j (k_2 - k_1)}{N}} = N\delta^K_{k_2,k_1}, \tag{38}$$

periodicity,

$$\tilde{f}[k+N] = \sum_{j=1}^{N} f[j] e^{-\frac{2\pi i j (k+N)}{N}} = \sum_{j=1}^{N} f[j] e^{-\frac{2\pi i j k}{N}} e^{-2\pi i j} = \tilde{f}[k] \; . \tag{39}$$

The inverse transform is also given by,

$$f[j] = \frac{1}{N}\sum_{k=1}^{N} \tilde{f}[k] e^{\frac{2\pi i j k}{N}} \; . \tag{40}$$

These two transformations have very similar properties to this we listed for FT.
Fast Fourier Transforms (FFT):

FFT is a very fast way to calculate DFT. This calculation scales like $N \log N$ computationally instead on N^2 one expects for DFT. The only assumption one needs to add is that N is some power of 2 and that the Fourier space is samples at points,

$$\omega(k) = \begin{cases} \frac{2\pi(k-1)}{N\Delta} & \text{if } k \leq N/2+1; \\ -\frac{2\pi(N-k+1)}{N\Delta} & \text{if } k > N/2+1. \end{cases}$$

I will not explain this here but direct the student to other resources that describe FFT (e.g., the book Numerical Recipes by Press et al.).

4 ESTIMATION THEORY

4.1 *Properties of estimators*

There are some generic properties that we desire an estimator, $\hat{\boldsymbol{\theta}}_N$ to have. The subscript N was added to indicate the number of data points that the estimator depends on. Here is a list of these properties.

1. *Unbiased:* The estimator $E\{\hat{\boldsymbol{\theta}}_N\} = g(\mathbf{x})$ is said to be unbiased if

   $$E\{\hat{\boldsymbol{\theta}}_N\} = \boldsymbol{\theta} \tag{41}$$

   otherwise, the estimator is said to be *biased* with bias $b(\hat{\boldsymbol{\theta}}) = E\{\hat{\boldsymbol{\theta}}_N\} - \boldsymbol{\theta}$.

2. *Consistent:* An estimator is said to be consistent if it converges to towards the true value of $\boldsymbol{\theta}$, i.e.,

   $$\lim_{N\to\infty} \hat{\boldsymbol{\theta}}_N = \int_{-\infty}^{\infty} g(\mathbf{x}) p(\mathbf{x}) dx = \boldsymbol{\theta}. \tag{42}$$

   In other words, it is consistent if the more data one has the closer it is to the real value.

3. *Efficient:* An estimator $\hat{\boldsymbol{\theta}}_N$, is said to be efficient with respect to another estimator, say $\check{\boldsymbol{\theta}}_N$ if

   $$\mathrm{Var}\{\hat{\boldsymbol{\theta}}_N\} < \mathrm{Var}\{\check{\boldsymbol{\theta}}_N\}$$

In what follow we shall discuss how we look for an estimator.

i: Minimum Variance unbiased estimator

In order to find a good estimator one often needs to adopt some optimality criterion. A natural choice of such a criterion is the so called *mean square error* (MSE) which is defined as,

$$mse(\hat{\boldsymbol{\theta}}) = E\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\}. \tag{43}$$

Although this criterion seems to make sense it unfortunately leads to problematic estimators that depend on the bias itself and hence requires some extra knowledge, sometimes on the very value we are trying to measure. Here is how it depends on the bias,

$$mse(\hat{\boldsymbol{\theta}}) = E\{\big[(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\}) + (E\{\hat{\boldsymbol{\theta}}\} - \boldsymbol{\theta})\big]^2\} \tag{44}$$

$$= E\{\big[(\hat{\boldsymbol{\theta}} - E\{\hat{\boldsymbol{\theta}}\}) + b(\boldsymbol{\theta})\big]^2\} \tag{45}$$

$$= \mathrm{Var}\{\hat{\boldsymbol{\theta}}\} + b^2(\boldsymbol{\theta}). \tag{46}$$

Here of course I assumed that the bias is statistically independent of the $\hat{\boldsymbol{\theta}}$. This result naturally makes sense as it implies that the error one gets due to the estimator is composed of two components, one due to the variance of the estimator and the other due to its bias.

Here we give an example of why such an estimator can be problematic. Assume that an estimator of a constant value is given by,

$$\hat{A} = a\frac{1}{N}\sum_{n=1}^{N} x[n]. \tag{47}$$

We would like to find a that results in the minimum MSE. Remember that $E\{\hat{A}\} = aA$ and $\mathrm{Var}\{\hat{A}\} = a^2\sigma^2/N$ (see equations 7 and 8). Clearly,

$$mse(\hat{A}) = \frac{a^2\sigma^2}{N} + (a-1)^2A^2\ . \tag{48}$$

Now lets find a that minimizes MSE by taking it derivative with respect to a and equate to zero. This yields an optimal value for a given by,

$$a_{opt} = \frac{A^2}{A^2 + \sigma^2/N}. \quad (49)$$

This is clearly not what we want as the optimal estimator depends on the value of the parameter that we wish to measure A.

This approach in fact is not very useful. Better to find an unbiased estimator and then minimize the variance, known also as Minimum Variance Unbiased (MVU) estimator. Finding such estimators will be the topic of the next few sections.

4.2 *The Cramer-Rao Lower bound*

Sine all the information we have is embodied in the data and the prior knowledge about the system given in the form of the PDF. It is clear that the more the PDF dependent on the parameter we like to estimate that more chance we have on determining its value more accurately. This statement can be shown for the following example (similar to the one described in Eq. 5),

$$x[0] = A + w[0], \quad (50)$$

where $w[0] \sim \mathcal{N}(0, \sigma^2)$, and we want to estimate A. The conditional PDF is then written as

$$P(x[0]|A) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x[0]-A)^2}{2\sigma^2}\right] \quad (51)$$

In general, when $p(\mathbf{x}|\boldsymbol{\theta})$ is viewed as a function of the unknown parameters (with fixed $\mathbf{x}$), it is termed the *likelihood function* – we'll discuss this function in detail soon. In the two examples I have given, one can intuitively see that the sharpness of the likelihood function determines how well one can estimate the unknown parameter A. To quantify the degree of sharpness one can measure the the negative of the second derivative of the logarithm of the likelihood function at its peak. For our example this

$$\ln p(x[0]|A) = -\ln\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x[0]-A)^2 \quad (52)$$

For our estimator, $\hat{A} = x[0]$, this gives the following relations,

$$\mathrm{Var}\{\hat{A}\} = \sigma^2 = \frac{1}{-\frac{\partial^2 \ln p(x[0]|A)}{\partial A^2}}. \quad (53)$$

The variance clearly increases as the curvature decreases.

Although in this example the second derivative does not depend on $x[0]$, this is not true in general. Hence it is more appropriate to define the curvature measure as,

$$-E\{\frac{\partial^2 \ln p(x[0]|A)}{\partial A^2}\} \quad (54)$$

Notice that the expectation is taken with regard to $p(x[0]|A)$ whereas the left hand side of equation 52 is calculated with regard to $\hat{A}$.

It turns out that this result can be generalized in the form of the so called Cramer-Rao Lower Bound (CRLB). We'll show this theorem in the general case, but first we show it and prove it for the simple case of one scalar parameter. In this case the CRLB theorem is givens follows: **CRLB theorem (the scalar case):** Assume that the PDF, $p(\mathbf{x}|\theta)$, is known, we are looking for θ that satisfies the regularity condition, i.e.,

$$E\left[\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta}\right] \quad \text{exists and is finite for all } \theta$$

where the expectation is taken with respect to $p(\mathbf{x}|\theta)$. Then the variance of any unbiased estimator $\hat{\theta}$ satisfies the two equivalent inequalities,

$$\text{Var}\{\hat{\theta}\} \geq \frac{1}{-E\left[\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2}\right]} \tag{55}$$

or

$$\text{Var}\{\hat{\theta}\} \geq \frac{1}{E\left[\left(\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta}\right)^2\right]} \tag{56}$$

where the derivative is evaluated at the true value of θ and the expectation is taken with respect to $p(\mathbf{x}|\theta)$. Furthermore, an unbiased estimator may be found that attains the bound for all θ if and only if

$$\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta} = I(\theta)(\hat{\theta} - \theta), \tag{57}$$

where $\hat{\theta}$ is the MVU estimator and variance is the minimum variance given by $1/I(\theta)$.

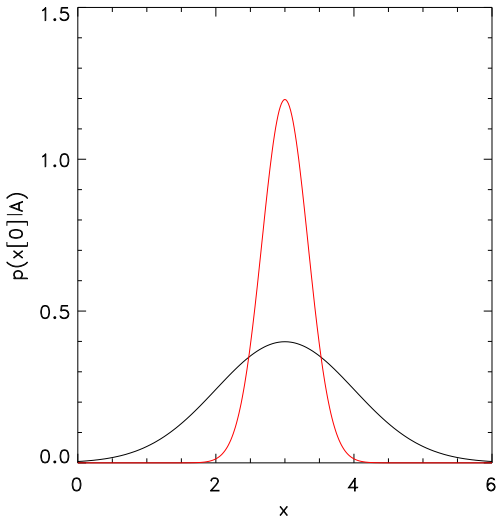

Figure 9: Logic behind CRLB: sharpness of PDF.

Proof of CRLB theorem:

We start with remembering that $\hat{\theta}$ is unbiased, i.e., $E\{\hat{\theta}\} = \theta$, i.e.,

$$E\left[(\hat{\theta} - \theta)\right] = \int_{-\infty}^{\infty} p(\mathbf{x}|\theta)\,(\hat{\theta} - \theta)\,d\mathbf{x} = 0, \tag{58}$$

Now we take the partial derivative with respect to θ, notice that we assume that such a derivative exists for all components within the integral, namely, all functions are well behaved. This yields,

$$\int_{-\infty}^{\infty} \left(\frac{\partial p(\mathbf{x}|\theta)}{\partial \theta}(\hat{\theta} - \theta) - p(x|\theta)\right) d\mathbf{x} = 0 \tag{59}$$

or

$$\int_{-\infty}^{\infty} \frac{\partial p(\mathbf{x}|\theta)}{\partial \theta}(\hat{\theta} - \theta) d\mathbf{x} = 1. \tag{60}$$

Then we make the transition to

$$\int_{-\infty}^{\infty} \frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta} p(\mathbf{x}|\theta)(\hat{\theta}-\theta)d\mathbf{x} = 1. \tag{61}$$

Finally, taking the square and then use the Cauchy-Schwarz inequality we can recast this equation in the form,

$$1 = \left(\int_{-\infty}^{\infty} \left(\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta} p(\mathbf{x}|\theta)^{1/2} \right) \left((\hat{\theta}-\theta) p(\mathbf{x}|\theta)^{1/2} \right) d\mathbf{x} \right)^2 \tag{62}$$

$$\leq \left(\int_{-\infty}^{\infty} \left(\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta} \right)^2 p(\mathbf{x}|\theta)d\mathbf{x} \right) \left(\int_{-\infty}^{\infty} (\hat{\theta}-\theta)^2 p(\mathbf{x}|\theta)d\mathbf{x} \right) \tag{63}$$

The last term in the RHS is $\mathrm{Var}(\hat{\theta})$ which immediately gives the second for of the CRLB shown above. The Cauchy-Schwart relation becomes an equality only if,

$$\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta} p(\mathbf{x}|\theta)^{1/2} = \frac{1}{c(\theta)} (\hat{\theta}-\theta) p(\mathbf{x}|\theta)^{1/2}, \tag{64}$$

where c is only a function of θ. Now it is easy to see that $c(\theta) = \frac{1}{I(\theta)}$ by taking the derivative

$$\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2} = -\frac{1}{c(\theta)} + \frac{\partial c^{-1}}{\partial \theta}(\hat{\theta}-\theta). \tag{65}$$

Since $\hat{\theta} = \theta$ (unbiased estimator) we obtain,

$$-E\Big\{ \frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2} \Big\} = -\frac{1}{c(\theta)} = I(\theta). \tag{66}$$

Which also proves the equality statement in the CRLB theorem.

Now let us show that the two forms of CRLB are equivalent,

$$\frac{\partial^2}{\partial \theta^2} \left(\int_{-\infty}^{\infty} p(x|\theta)d\mathbf{x} \right) = \frac{\partial^2}{\partial \theta^2}(1) = 0 \tag{67}$$

Which gives,

$$0 = \frac{\partial}{\partial \theta} \left(\int_{-\infty}^{\infty} \frac{\partial \ln p(x|\theta)}{\partial \theta} p(\mathbf{x}|\theta)d\mathbf{x} \right) \tag{68}$$

$$= \int_{-\infty}^{\infty} \left(\frac{\partial^2 \ln p(x|\theta)}{\partial \theta^2} p(\mathbf{x}|\theta) + \left(\frac{\partial \ln p(x|\theta)}{\partial \theta} \right)^2 p(\mathbf{x}|\theta) \right) d\mathbf{x} \tag{69}$$

hence

$$E\left[\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2} \right] = -E\left[\Big(\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta} \Big)^2 \right] \tag{70}$$

Example: In this example we show a case in which the estimator will never reach the CRLB. Assume we wish to estimate the phase, ϕ, of a sinusoidal function which Gaussian noise, i.e.,

$$x[n] = A\cos(\omega_0 t[n] + \phi) + w[n], \quad \text{for} \quad n = 1, \ldots, N \tag{71}$$

where $t[n]$ are the times of measurements (which we assume are known very accurately), A the amplitude and ω_0 is the angular frequency. Now the PDF is of course given by,

$$p(\mathbf{x};\phi) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\big(x[n] - A\cos(\omega_0 t[n] + \phi)\big)^2\right\}. \tag{72}$$

Now we want to estimate the the CRLB. We first calculate the second derivative of the log-likelihood, i.e.,

$$\frac{\partial^2 \ln p(\mathbf{x};\phi)}{\partial\phi^2} = -\frac{A}{\sigma^2}\sum_{n=1}^{N}\big(x[n]\cos(\omega_0 t[n] + \phi) - A\cos(2\omega_0 t[n] + 2\phi)\big) \tag{73}$$

Now we take the negative expected value of this we get,

$$\begin{aligned}
-E\left\{\frac{\partial^2 \ln p(\mathbf{x};\phi)}{\partial\phi^2}\right\} &= \frac{A}{\sigma^2}\sum_{n=1}^{N}\big(E\{x[n]\}\cos(\omega_0 t[n] + \phi) - A\cos(2\omega_0 t[n] + 2\phi)\big) && (74)\\
&= \frac{A^2}{\sigma^2}\sum_{n=1}^{N}\big(\cos^2(\omega_0 t[n] + \phi) - \cos(2\omega_0 t[n] + 2\phi)\big) && (75)\\
&= \frac{A^2}{\sigma^2}\sum_{n=1}^{N}\big(\frac{1}{2} + \frac{1}{2}\cos(2\omega_0 t[n] + 2\phi) - \cos(2\omega_0 t[n] + 2\phi)\big) && (76)\\
&= \frac{A^2}{\sigma^2}\sum_{n=1}^{N}\big(\frac{1}{2} - \frac{1}{2}\cos(2\omega_0 t[n] + 2\phi)\big) = \frac{NA^2}{2\sigma^2} + \frac{A^2}{2\sigma^2}\sum_{n=1}^{N}\big(\cos(2\omega_0 t[n] + 2\phi)\big) && (77)\\
&\approx \frac{NA^2}{2\sigma^2}. && (78)
\end{aligned}$$

Therefore, $\mathrm{Var}(\hat{\phi}) \geq \frac{2\sigma^2}{NA^2}$. Notice that here the estimator gets better the larger A/σ is and the more measurements we have. In this example the condition for the bound to hold is clearly not satisfied. **Therefore, a phase estimator that is unbiased and attains the CRLB does not exist.**

CRBL for transformed parameters: Now we present the more general case where we are not interested directly in the unbiased estimator $\hat{\theta}$ but in another function of it, $\psi(\hat{\theta})$. In this case the CRLB theorem takes the form,

$$\mathrm{Var}\big\{\hat{\psi}(\theta)\big\} \geq \frac{(\partial\psi/\partial\theta)^2}{-E\big(\partial^2 \ln p(\mathbf{x}|\theta)/\partial\theta^2\big)}. \tag{79}$$

For example, in the case shown in Eq. 5 if one wants to estimate A^2 instead of A then the CRLB gives,

$$\mathrm{Var}\big\{\hat{A^2}\big\} \geq \frac{(2A)^2}{N/\sigma^2} = \frac{4A^2\sigma^2}{N}. \tag{80}$$

Notice that nonlinear transformations tend to change the properties of the estimator. To demonstrate this consider the last example where we show that the estimator of $\hat{A} = E\{x\}$ is efficient (it is an MVU estimator) as it attains the CRLB. However the estimator $\hat{A^2} = E\{x^2\}$ is not efficient since it gives,

$$E(x^2) = E^2(x) + \mathrm{Var}(x) = A^2 + \sigma^2/N \neq A^2 \tag{81}$$

One can immediately see from this that only linear transformation maintain the efficiency of an esimator. It is easy to prove this as $\psi(\theta) = a\theta + b$ then the CRLB is simply $a^2\mathrm{Var}(\theta)$ which means if $\hat{\theta}$ is efficient then $\hat{\psi(\theta)}$ still fulfills the conditions for efficiency.

CRLB for the general case of multiple parameters: In the case of multiple parameters, say M parameters,

we define $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_M]^T$. The Fisher information matrix is then defined as

$$I_{i,j} = -E\left[\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]. \tag{82}$$

Let us also define the vector transformation $\boldsymbol{\psi}(\boldsymbol{\theta})$ and the covariance matrix

$$\mathrm{Cov}(\boldsymbol{\psi}) = E\big(\,(\boldsymbol{\psi} - E(\boldsymbol{\psi}))\,(\boldsymbol{\psi} - E(\boldsymbol{\psi}))^T\,\big), \tag{83}$$

where the superscript ^{T} stands for transpose. The CRLB in this case is given by

$$\mathrm{Cov}(\boldsymbol{\psi}) \geq \frac{\partial\boldsymbol{\psi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \boldsymbol{I}^{-1} \frac{\partial\boldsymbol{\psi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^T. \tag{84}$$

Notice that $\frac{\partial\boldsymbol{\psi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ is the transformation Jacobian matrix, i.e., the $\{i,j\}$ term is given by $\frac{\partial\boldsymbol{\psi}_i(\boldsymbol{\theta})}{\partial\theta_j}$. Matrix inequality in this case has the following meaning: $\mathbf{A} \geq \mathbf{B}$ means $(\mathbf{A} - \mathbf{B})$ is a positive semidefinite matrix. A special case of this formulation is if $\boldsymbol{\psi}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and we focus only on the diagonal terms, namely

$$\mathrm{Var}(\theta_i) = \mathrm{Cov}(\boldsymbol{\theta})_{i,i} \geq \left[\boldsymbol{I}(\boldsymbol{\theta})^{-1}\right]_{i,i} \tag{85}$$

Let us consider the problem of a set of N observations of a constant quantity with white Gaussian noise,

$$x[n] = A + w[n] \qquad n = 1, 2, \ldots, N. \tag{86}$$

The conditional PDF is then written as

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x[n]-A)^2}{2\sigma^2}\right] \tag{87}$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\sum_{n=1}^{N} \frac{(x[n]-A)^2}{2\sigma^2}\right] \tag{88}$$

Now we would like to estimate both A and σ^2, hence our parameters vector is $\boldsymbol{\theta} = [A, \sigma^2]^T$. For this case the Fisher information matrix is

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} -E\left\{\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2}\right\} & -E\left\{\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial\sigma^2}\right\} \\ -E\left\{\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\sigma^2 \partial A}\right\} & -E\left\{\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\sigma^{2\,2}}\right\} \end{bmatrix} = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}$$

Which means that $\mathrm{Var}(\hat{A}) \geq \sigma^2/N$ and $\mathrm{Var}(\hat{\sigma}^2) \geq 2\sigma^4/N$, where the first diagonal term in the CRLB gives the variance of the signal and the second gives the variance of the variance, respectively.

Figure 10 shows an example of the use of CRLB where the solid contour show the bound whereas the red and green points show the actual calculation of the variance. The left panel shows function which we try to fit to the data and right panel shows the results for various parameters.

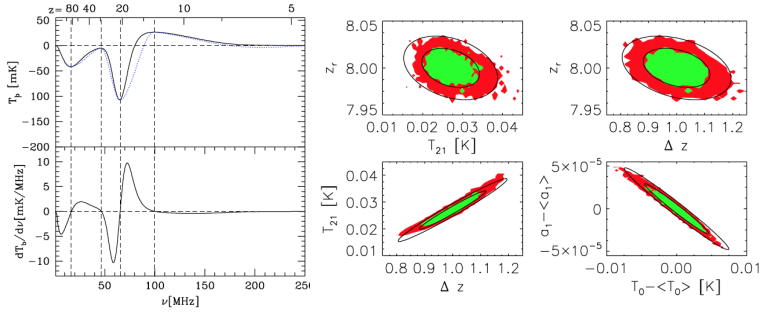

Figure 10: An example of CRLB vs. Full calculation. This figure shows an example of the use of CRLB where the solid black contours show the bound whereas the red and green points show the actual calculation of the variance. The left panel shows function which we try to fit to the data and right panel shows the results for various parameters.

4.3 *Linear Models*

So far we have focused on the limits of estimators. In the next subsections we'll explore how to actually find good estimators. But first we would like to spend some time on discussing linear models which have very simple properties.

As an example we start with the simple case of the following data model (straight line fitting),

$$x[n] = A + Bs[n] + w[n], \quad \text{for} \quad n = 1, \ldots, N \tag{89}$$

where $s[n]$ is the independent quantity with respect to which we measure the data, $w[n]$ is WGN and A and B are free parameters that we would like to set. We can recast this equation in matrix notation as,

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}, \tag{90}$$

where

$$\mathbf{x} = [x[1], x[2], \ldots, x[N]]^T \tag{91}$$

$$\mathbf{w} = [w[1], w[2], \ldots, w[N]]^T \tag{92}$$

$$\boldsymbol{\theta} = [A\ B]^T \tag{93}$$

The matrix $\mathbf{H}$ is obviously the following $N \times 2$ matrix,

$$\mathbf{H} = \begin{bmatrix} 1 & s[1] \\ 1 & s[2] \\ \vdots & \vdots \\ 1 & s[N] \end{bmatrix}. \tag{94}$$

Since we assume that the noise vector is Gaussian we can write it as $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. As we have shown in the discussion on the CRLB theorem, the equality constraint can be obtained if

$$\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta}). \tag{95}$$

for $\mathbf{g}(\mathbf{x}) = \hat{\boldsymbol{\theta}}$. Then the estimator will be an MVU estimator. Here we want to show that this is exactly what we get in the case of a linear model.

One can easily show that for our case the first derivative is,

$$\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \boldsymbol{\theta}}\left[\mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}\right] \tag{96}$$

$$= \frac{1}{\sigma^2}\left[\mathbf{H}^T\mathbf{x} - \mathbf{H}^T\mathbf{H}\boldsymbol{\theta}\right] \tag{97}$$

$$= \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\left[(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} - \boldsymbol{\theta}\right]. \tag{98}$$

Which is exactly of the form of Eq. 95. Namely, this gives $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ and $\mathbf{I}(\boldsymbol{\theta}) = \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}$. Hence the covariance matrix of the MVU estimator of $\boldsymbol{\theta}$ is

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{I}^{-1}(\boldsymbol{\theta}) = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}. \tag{99}$$

One of assumptions here is that the matrix $\mathbf{H}^T\mathbf{H}$ is invertible. This result means that also the parameter vector $\boldsymbol{\theta}$ follows a Gaussian field $\sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^T\mathbf{H})^{-1})$

This derivation can obviously be generalized for more than two parameters, i.e. if $\mathbf{H}$ is an $N \times q$ matrix where q is the number of parameters ($q < N$). The general formulation is then as follows: For a data that can be modeled as,

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}. \tag{100}$$

Then the MVU estimator of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \tag{101}$$

and the covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by,

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{I}^{-1}(\boldsymbol{\theta}) = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1}. \tag{102}$$

The Gaussian nature of the MVU estimator allows us to determine the exact statistical properties of the estimated parameters.

We show now a couple of very useful examples of this theorem related to fitting curves. Lets us first see in detail the meaning of these equations for the two parameters A and B then from equation 94 one gets

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ s[1] & s[2] & \dots & s[N] \end{bmatrix} \begin{bmatrix} 1 & s[1] \\ 1 & s[2] \\ \vdots & \vdots \\ 1 & s[N] \end{bmatrix} = \begin{bmatrix} N & \sum_{n=1}^{N} s[n] \\ \sum_{n=1}^{N} s[n] & \sum_{n=1}^{N} s[n]^2 \end{bmatrix}. \tag{103}$$

and

$$\mathbf{H}^T\mathbf{x} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ s[1] & s[2] & \dots & s[N] \end{bmatrix} \begin{bmatrix} x[1] \\ x[2] \\ \vdots \\ x[N] \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} x[n] \\ \sum_{n=1}^{N} s[n]x[n] \end{bmatrix} \tag{104}$$

Now assume that $s[n] = n\Delta s$, where Δs is constant, therefore we get,

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} N & \frac{N(N+1)}{2}\Delta s \\ \frac{N(N+1)}{2}\Delta s & \frac{N(N+1)(2N+1)}{6}\Delta s^2 \end{bmatrix}. \tag{105}$$

In case the noise is correlated with a correlation matrix $\mathbf{C}$, then the estimator vector and its covariance

are given by,

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} \tag{106}$$

and,

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}, \tag{107}$$

respectively.

The key to show this is by remembering that the correlation matrix $\mathbf{C}$ and its inverse are positive definite and hence we can write $\mathbf{C}^{-1} = \mathbf{D}^T\boldsymbol{\Sigma}^{-1}\mathbf{D}$, with $\mathbf{D}$ is an orthogonal matrix (eigen value problems). The matrix D is often called the whitening matrix simply because,

$$\mathbf{E}\{(\mathbf{Dw})(\mathbf{Dw})^T\} = \mathbf{DCD}^T = \boldsymbol{\Sigma}, \tag{108}$$

where,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_N^2. \end{bmatrix}. \tag{109}$$

In other words this transformation decorrelates the noise and makes each component a WGN again.

This brings us to the interesting case of a multi-variate Gaussian which is normally defined in the following easy,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\mathbf{C}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \tag{110}$$

where $\boldsymbol{\mu}$ is the mean vector of $\mathbf{x}$. We'll return to this later when we discuss SVD and PCA methods.

4.4 *Maximum Likelihood Estimation*

We'll now turn our attention from MVU estimators and explore the very widely use method for finding estimators, namely, the so called maximum likelihood estimation. There are a number of reasons for why this estimator is the most widely used estimator but chief among them is the simple fact that it has a very clear and often straightforward way to calculate it. In other words, it does not only have the desired theoretical properties but also has a clear practical way to calculate it.

The Maximum Likelihood Estimator (MLE) is simply given by maximizing the likelihood with respect to the parameter we would like to estimate. Usually this is written is the following way,

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\mathbf{x}|\boldsymbol{\theta}). \tag{111}$$

This can be obtained by one of the two equations

$$\frac{\partial p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{ML} = 0 \tag{112}$$

$$\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{ML} = 0. \tag{113}$$

The two equations are equivalent because the logarithm function is a monotonic function.

Together with the clear way of calculating the MLE it also has a number of theoretical properties that are generally desired in estimators. Here we will list these properties but not prove them. The first is that the MLE is asymptotically (for very large data sets) unbiased and asymptotically attains the CRLB, therefore, it is asymptotically efficient. In the case of a one-to-one transformation of the parameters the MLE for these

parameters is also the MLE of the transformation of the parameters. Of course, one of issues that one has to pay attention to is how much data one needs to get close to these properties! In what follows we will show a number of examples of the use of this estimator.

Assume one has the same problem we considered earlier again, namely, a set of N observations of constant quantity with white Gaussian noise,

$$x[n] = A + w[n] \qquad n = 1, 2, \ldots, N. \tag{114}$$

However here the we will assume that both variance and mean are A. Therefore, the conditional PDF is then written as

$$p(\mathbf{x}|A) = \frac{1}{(2\pi A)^{N/2}} \exp\left[-\sum_{n=1}^{N} \frac{(x[n]-A)^2}{2A}\right]. \tag{115}$$

We would like to calculate the estimator $\hat{A}$. The way to do that is straightforward, calculate $\hat{A}$ that satisfies,

$$\left.\frac{\partial \ln p(\mathbf{x}|A)}{\partial A}\right|_{A=\hat{A}} = 0. \tag{116}$$

This calculation gives the following estimator,

$$\hat{A} = -\frac{1}{2} + \sqrt{\frac{1}{N}\mathbf{x}\cdot\mathbf{x} + \frac{1}{4}} \tag{117}$$

Notice that this is a biased estimator because,

$$E(\hat{A}) = E\left(-\frac{1}{2} + \sqrt{\frac{1}{N}\mathbf{x}\cdot\mathbf{x} + \frac{1}{4}}\right) \neq -\frac{1}{2} + \sqrt{E\left(\frac{1}{N}\mathbf{x}\cdot\mathbf{x}\right) + \frac{1}{4}} = -\frac{1}{2} + \sqrt{A + A^2 + \frac{1}{4}} = A \tag{118}$$

where we have used the relation $E\left(\frac{1}{N}\sum_n x[n]^2\right) = \mathrm{Var}(x) + E^2(x) = A + A^2$. However for $N \to \infty$ the inequality becomes equality since, $\frac{1}{N}\sum_n x[n]^2 = \mathrm{Var}(x) + E^2(x) = A + A^2$.

We have calculated the estimator now let us calculate its variance. This is a bit more difficult but can be analytically done in the limit of very large N, which gives $\frac{1}{N}\sum_n x[n]^2 \approx A + A^2$. In this case we can use Taylor expansion of the function $g(u) = -\frac{1}{4} + \sqrt{u + \frac{1}{4}} = g(u_0) + dg/du|_{u_0}(u - u_0)$. This yields the following expression for the estimator,

$$\hat{A} \approx A + \frac{\frac{1}{2}}{A + \frac{1}{2}}\left[\frac{1}{N}\sum_n x[n]^2 - (A + A^2)\right] \tag{119}$$

Now the variance at this limit can be calculated as follows,

$$\mathrm{Var}(\hat{A}) = \frac{\frac{1}{2}^2}{N(A + \frac{1}{2})^2}\mathrm{Var}(x^2) \tag{120}$$

We can calculate $\mathrm{Var}(x^2) = E(x^4) - E(x^2)^2$. One can easily show that for $x[n] \sim \mathcal{N}(\mu, \sigma^2)$, $E(x^2) = \mu^2 + \sigma^2$ and $E(x^4) = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$. For our case $\mu = \sigma^2 = A$ we get $\mathrm{Var}(x^2) = 4A^3 + 2A^2$, hence,

$$\mathrm{Var}(\hat{A}) = \frac{\frac{1}{2}^2}{N(A + \frac{1}{2})^2}(4A^3 + 2A^2) = \frac{A^2}{N(A + \frac{1}{2})}. \tag{121}$$

Here we were lucky because we could estimate the asymptotic value of the variance and the mean analytically and hence decide whether our data sample is large enough for a given accuracy we would like to reach. However in general that is not possible and one often has to test how many data points needed to

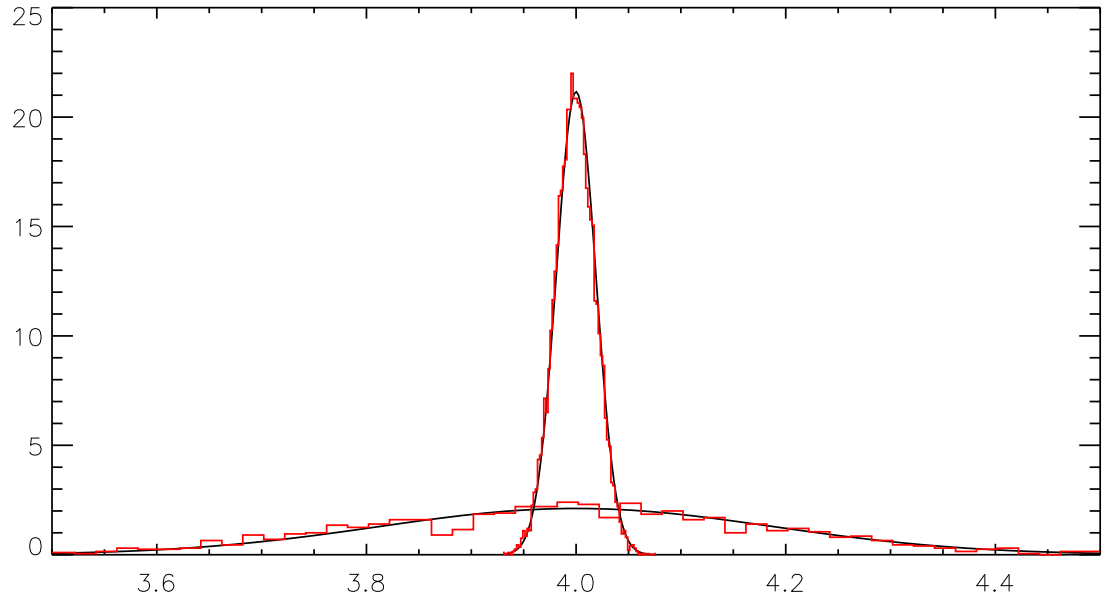

Figure 11: Monte Carlo simulations to find A and $\mathrm{Var}(A)$ as compared to their asymptotic values. The broad one is done for 100 data points and the highly peaked results is for 10000 data points.

reach a given accuracy using so called Monte Carlo techniques. Figure 11 shows a result of a Monte Carlo estimation of the error.

We also show how to numerically find the minimum and the variance numerically (see figure 12). What is done here is that we simply substitute possible values of A that cover the range in which this parameter makes the likelihood function attain its maximum. Here I simply calculated $p(\mathbf{x}|A)$ for values of A in the range of 2 to 6 with 1000 steps. In order to interpret this function statistically one has to normalize it so that $\int L(A)dA = 1$ and the mean an variance can be calculated as,

$$E(A) = \int AL(A)dA \quad \text{and} \quad \mathrm{Var}(A) = \int A^2 L(A)dA - E^2(A). \tag{122}$$

We also show a Gaussian fit to $L(A)$ which shows that it is not exactly a Gaussian.

The maximum likelihood method could be also used to compare between different hypothesis using the so called maximum likelihood ratio test. We might return back to this later in the course when we (hopefully) discuss hypothesis testing.

Before finishing with MLE there are two theorems that we should see, the first shows the Invariance property of the MLE and the second has to do with the asymptotic properties of MLE.

INVARIANCE OF MLE: The MLE of $\boldsymbol{\alpha}$ which is a function of parameter vector $\boldsymbol{\theta}$ such that $\boldsymbol{\alpha} = \mathbf{f}(\boldsymbol{\theta})$ where $\mathbf{f}$ is a well behaved function (though not necessarily invertible), then the following relation holds, $\hat{\boldsymbol{\alpha}} = \mathbf{f}(\hat{\boldsymbol{\theta}})$.

PROOF: The value of $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood, $p_{\boldsymbol{\theta}}$, also maximizes the likelihood $p_{\boldsymbol{\alpha}}$ at $\boldsymbol{\alpha} = \mathbf{f}(\boldsymbol{\theta})$ because of the relation

$$p_{\boldsymbol{\alpha}}(\mathbf{x}|\boldsymbol{\alpha}) = p_{\boldsymbol{\alpha}}(\mathbf{x}|\mathbf{f}(\boldsymbol{\theta})) = \mathrm{J}^{-1}(\boldsymbol{\alpha},\boldsymbol{\theta}) p_{\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}), \tag{123}$$

where J is the Jacobian of the transformation. Notice that the key to this is that the Jacobian is data independent.

The other property has to do with the asymptotic behavior of MLE, the first is the property of consistency and the second is of asymptotic normality.

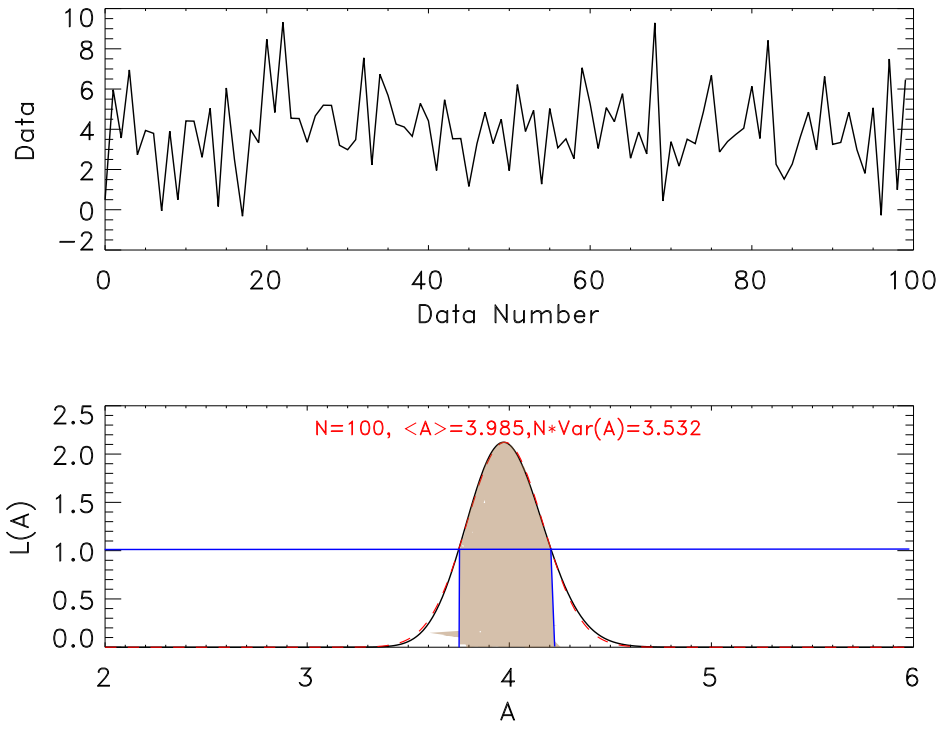

Figure 12: Numerical calculation of the MLE for 100 points data vector. The mean of the underlying PDF is 4 and the variance is also 4 ($\sigma = 2$). The estimated value of the mean A and the error is shown in the figure. The red dashed line shows a Gaussian fit to the likelihood. The area shown in the light brown region is the one that corresponds to $\approx 68.3\%$ of the probability (normally referred to as 1σ significance – the 2σ and 3σ uncertainty correspond to $\approx 95.4\%$ and $\approx 99.7\%$ probability respectively).

CONSISTENCY OF MLE: The MLE is statistically consistent. This property means the MLE converges to the true value of $\boldsymbol{\theta}_0$ for a very large number of data points.

$$\lim_{N\to\infty} \hat{\boldsymbol{\theta}}_{MLE} \to \boldsymbol{\theta}_0 \tag{124}$$

We will not prove this here but merely point out that proving this theorem require a number of conditions to hold. These conditions are: 1- Identification (there is a unique global maximum of the likelihood); 2- Compactness (the parameter space $\boldsymbol{\theta}$ is compact); 3- Continuity in $\boldsymbol{\theta}$; 4- Dominance (there is a dominant region in the space of $\boldsymbol{\theta}$.

ASYMPTOTIC NORMALITY OF THE MLE: The MLE has asymptotically Gaussian distribution with mean $\boldsymbol{\theta}_0$ and variance given by the Fisher Information matrix, i.e.,

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0\right) \sim \mathcal{N}(0, \mathbf{I}^{-1}), \tag{125}$$

namely, the MLE attains the CRLB. The proof of this theorem relies on the central limit theorem which states that under certain (generic) conditions the mean of very large independent random variables (not necessarily *iid*s) follows a Gaussian distribution.

4.5 *Least Squares method*

We now move to another very widely used estimators called the method of *least squares*. This method departs significantly from the estimators we have pursued so far in the sense that it does not look for an MVU and in many cases is hard to assign to it any statistical optimality, however, often using it just makes sense. This method is especially used in overdetermined system in which there is much more data points than unknowns, e.g., fitting a curve with a few parameters to many data points. This method was first used by F. Gauss (at the age of 18) when he used it to study planetary motions.

A main feature of this method is that no probabilistic assumptions about the data are made and only a single model is assumed. This gives it a very significant advantage is terms of its broad range of possible applications. On the negative side it is not possible to make any assertions about the optimality (bias, MVU, etc) of the estimator without further assumptions on the system. In other word out assumptions in this case about the system are very limited and hence our interpretation of the results is also limited. However, such simple assumptions allows a quick finding of an estimator!

The Least Square Estimator (LSE) basically attempts to find the minimum difference between the data and the assumed signal in the least square sense. In other word for data vector $\mathbf{x} = [x[1], x[2], \ldots, x[N]]^T$, assumed to be a function of $\mathbf{s}$, $\mathbf{f}(\mathbf{s};\boldsymbol{\theta})$, the LSE is given by the parameters $\boldsymbol{\theta}_{LSE}$ that minimizes the equation,

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N}(x[i] - f[i])^2. \tag{126}$$

namely,

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0. \tag{127}$$

Note that no probabilistic assumptions about the data and noise have been made.

As an example let us take the simple example of measuring a quantity with a constant value, A. The LSE is the one that minimizes,

$$J(A) = \sum_{i=1}^{N}(x[i] - A)^2. \tag{128}$$

Namely, the one that satisfies the equation,

$$\frac{\partial J(A)}{\partial A} = -2\sum_{i=1}^{N}(x[i] - A) = 0 \tag{129}$$

This gives the following estimator,

$$\hat{A} = \frac{1}{N}\sum_{i=1}^{N} x[i]. \tag{130}$$

Although it might seem that we have reached the MVU estimator we have discussed earlier one has to be careful. This is also true in case that the signal can be written as $x[i] = A + w[i]$ where $w[i]$ is WGN with zero mean. This however is not necessarily true in other situation.

To emphasize this point we show an example of data with WGN noise and with Rayleigh distributed noise i.e., $x[i] = A + \epsilon[i]$ (the noise $\epsilon[i] = \sqrt{w_1[i]^2 + w_2[i]^2}$ where w_1 and w_2 are *iid* from $\sim \mathcal{N}(0,1)$. Figure 13 shows these two examples where the later case is shown clearly to be biased.

Now let us consider the problem in which the model of the data is

$$x[i] = A\cos(\omega_0 s[i]) + \epsilon[i]. \tag{131}$$

where we would like to estimate A , namely $\theta = A$. In such case, $J(A) = \sum_{i=1}^{N}\left(x[i] - A\cos(\omega_0 s[i])\right)^2$ and we have to solve two equations,

$$\frac{\partial J(A)}{\partial A} = -2\sum_{i=1}^{N}\cos(\omega_0 s[i])(x[i] - A\cos\left(\omega_0 s[i]\right)) = 0 \tag{132}$$

Which gives the simple closed form estimator,

$$\hat{A} = \frac{\sum_{i=1}^{N}\cos(\omega_0 s[i])x[i]}{\sum_{i=1}^{N}\cos^2(\omega_0 s[i])}. \tag{133}$$

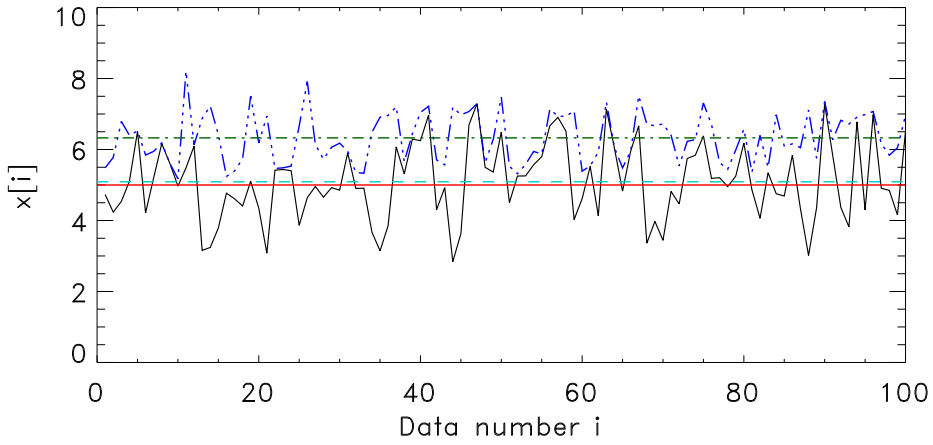

Figure 13: This figure shows the case of $x[i] = A + w[i]$ where first $w[i]$ is a WGN with zero mean and $\sigma = 1$ (solid black line) and with $w[i]$ is Rayleigh distributed, i.e., chi-squared with 2 *d.o.f.*, (blue triple dotted dashed line). The horizontal line show the real value of A (red solid line), the WGN case estimator (cyan dashed line) and the one estimated from the data with the Rayleigh distributed noise (green dotted-dashed line).

Now, if instead of A we want to estimate ω_0 one can easily see that there is no closed form solution for $\hat{\omega_0}$. The main difference between this case and the former case (i.e., with $\theta = A$) is that the former minimization yields a linear dependence of the estimator on the data whereas the later (i.e., with $\theta = \omega_0$) the dependence is nonlinear and must be solved with numerical methods. It is therefore clear that the Linear Least Squares (if they can be constructed) are much simpler. In what follows we discuss their solutions.

Assume that $\mathbf{x} = \mathbf{s} + \boldsymbol{\epsilon}$ where $\mathbf{s} = \mathbf{H}\boldsymbol{\theta}$ and $\mathbf{H}$ is a known $N \times p$ matrix with $N > p$. This is naturally the linear model we saw earlier when we discussed the MVU estimators. In such case, the LSE is found by minimizing,

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}). \tag{134}$$

Differentiating this equation with respect to $\boldsymbol{\theta}$ yields,

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}. \tag{135}$$

Setting the gradient to zero gives the LSE,

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} \tag{136}$$

Notice that this equation is identical to equation 101, however, that does not mean it is an MVU estimator or that it attains the CRLB, for that to hold the noise has to be Gaussian which is a condition not made when we obtain the LSE.

One also can find the value of minimum error is,

$$J_{min}(\boldsymbol{\theta}) = \mathbf{x}^T\left(\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\right)\mathbf{x}. \tag{137}$$

Notice that here we used the fact that the matrix $\mathbf{M} = \left(\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\right)$ is idempotent matrix, i.e., it satisfies $\mathbf{M}^2 = \mathbf{M}$. This can also be expressed as,

$$J_{min}(\boldsymbol{\theta}) = \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \tag{138}$$

$$= \mathbf{x}^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \tag{139}$$

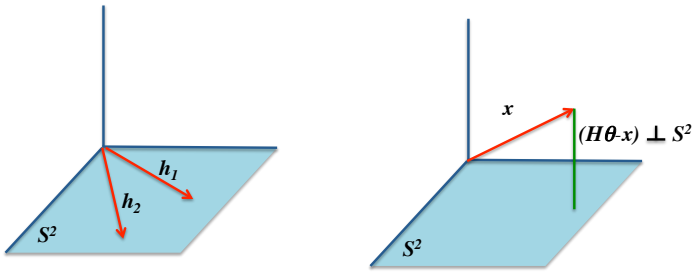

Figure 14: This figure show how the LSE find a solution that is perpendicular to the subspace (plane in this case) defined by the vectors $\mathbf{h_i}$.

The geometric interpretation of this equation is simple, it finds the solution id such a way that the vector of the error $\boldsymbol{\epsilon} = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})$ is perpendicular to the subspace spanned by all the vectors $\mathbf{h}_i$ such that $\mathbf{H} = \begin{bmatrix}\mathbf{h}_1 \mathbf{h}_2 \ldots \mathbf{h}_p\end{bmatrix}$. This is easy to see since the vector,

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{h}_1 = 0 \tag{140}$$

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{h}_2 = 0 \tag{141}$$

$$\vdots = \vdots \tag{142}$$

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{h}_p = 0. \tag{143}$$

Which translates to the equation,

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{H} = 0. \tag{144}$$

which in turns gives the LSE solution. Therefore, the LSE solution gives as error vector that is perpendicular to the space defined by the vector elements of $\mathbf{H}$ (see Figure 14.)

Now we can generalize the LSE to the so-called weighted least square estimator. In such a case,

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{W} (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}). \tag{145}$$

Where $\mathbf{W}$ is a positive definite, hence symmetric, weighting matrix with rank $N \times N$. In this case the general form of the weighted LSE is,

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T \mathbf{W} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{W} \mathbf{x} \tag{146}$$

With the minimum error value being

$$J_{min}(\boldsymbol{\theta}) = \mathbf{x}^T \left(\mathbf{W} - \mathbf{W}\mathbf{H}(\mathbf{H}^T \mathbf{W} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W}\right) \mathbf{x}. \tag{147}$$

As an example for weighted LSE assume that $x[i] = A + w[i]$ but here $w[i] \sim \mathcal{N}(0, \sigma_i^2$. Namely, the covariance matrix of the errors is given by $\mathbf{C} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$. A popular choice of weighting the data is with inverse the covariance matrix[1], i.e., $\mathbf{W} = \mathbf{C}^{-1} = diag(\sigma_1^{-2}, \sigma_2^{-2}, \ldots, \sigma_N^{-2})$. Here also $\mathbf{H} = \mathbf{I}$. This gives the following LSE,

$$J(A) = \sum_{i=1}^{N} \frac{(x[i] - A)^2}{\sigma^2[i]}. \tag{148}$$

[1]This is actually a result of the so called Best Linear Unbiased Estimator (know as BLUE) which is best MVU one can achieve in case the errors are uncorrelated and are drawn from a Gaussian distribution with zero mean and variance that is different for each data point. We haven't discussed this estimator for lack of time but it is an estimator with theoretical importance although very little practical use in general.

From Equation 146 with the one obtains the following LSE,

$$\hat{A} = \frac{\sum_{i=1}^{N} \frac{x[i]}{\sigma^2[i]}}{\sum_{i=1}^{N} \frac{1}{\sigma^2[i]}}. \quad (149)$$

This weighting scheme is easy to interpret as it gives more weight to data points with smaller error (variance).

Finding the LSE is often complicated even in the linear case. The problem is normally reduced to quadratic form, assuming that we are close to the minimum and then a solution of this quadratic formula is found. We will show such an example for solving a general linear least squares problem when we discuss the Singular Value Decomposition (SVD) method later in the course. For the nonlinear case, obtaining a solution is even more difficult and a number of methods have been developed including the family of the so called "Conjugate Gradient Methods", "Variable Metric Methods", "Simulated Annealing", etc. Specifically in the werkcollege you will learn couple of Variable Metric methods, the Broyden type methods specifically the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method and another very heavily used method called the Levenberg-Marquardt algorithm.

Lineary constrained LSE: Consider now the case in which the parameters we have are somehow related and not all independent. This is generally an interesting problem to solve. The constraints can be generally of nonlinear form, however, here we make further simplification and assume that the relation between them is linear and the is given by the relation,

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b}. \quad (150)$$

In order to find an LSE estimator that satisfy this constraint we can include it through a Lagrange multipliers vector $\boldsymbol{\lambda}$. Namely, the minimization problem becomes,

$$J_c(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) + \boldsymbol{\lambda}^T(\mathbf{A}\boldsymbol{\theta} - \mathbf{b}). \quad (151)$$

Which gives the solution

$$\hat{\boldsymbol{\theta}}_c = \hat{\boldsymbol{\theta}} - (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\boldsymbol{\lambda}/2 \quad (152)$$

and with simple manipulation we obtain,

$$\boldsymbol{\lambda}/2 = \left[\mathbf{A}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{A}^T\right]^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}} - \mathbf{b}) \quad (153)$$

4.6 *Bayesian Estimation (and philosophy)*

As we discussed earlier in the course, the heart of Bayesian estimation is the assumption that the parameters of interest, $\boldsymbol{\theta}$, constitute random vectors whose specific realization for the problem at hand must be statistically estimated. In principle this approach goes much more that this statement, since instead of finding the maximum likelihood estimator as we have done so far, i.e., maximize the probability of the data given the parameter vector, $\boldsymbol{\theta}$, we maximize the probability of the parameter vector given the data.

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}. \quad (154)$$

Notice, that the MLE approach and the Bayseian estimation approach coincide for a flat *prior*. This approach in general is more desirable when it can be implemented since $p(\boldsymbol{\theta}|\mathbf{x})$ gives in effect the probability of our theory given the data. In other words, it gives how well our model stands in face of accumulating data. Furthermore, the Bayesian approach allows us to include our *prior* knowledge or preconception on the parameter vector in the determination of the probability of the model given the data and even updating this *prior* with the accumulation of information. This is of great advantage as it allows, normally, a gradual improvement of our model of reality with the gradual accumulation of data. Obviously, sometimes

the improvement of our model of reality is very abrupt and can even contradict our previous model and preconception of reality. The Bayesian approach gives the proper vehicle with which to update our models even in such cases.

As in example for the incorporation of prior knowledge in our estimation we take the case of a DC measurement again with $x_i = A + w_i$ where $i = 1, \ldots, N$ and $w_i \sim \mathcal{N}(0, \sigma^2)$. We also assume that we have a non-flat *prior* on the value of A such that $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$. From Bayes Rule we obtain,

$$p(A|\mathbf{x}) = \frac{p(\mathbf{x}|A)p(A)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|A)p(A)}{\int p(\mathbf{x}|A)p(A)dA} \tag{155}$$

$$= \frac{e^{-\frac{\sum_i (x_i - A)^2}{2\sigma^2}} e^{-\frac{(A-\mu_A)^2}{2\sigma_A^2}}}{\int e^{-\frac{\sum_i (x_i - A)^2}{2\sigma^2}} e^{-\frac{(A-\mu_A)^2}{2\sigma_A^2}} dA} \tag{156}$$

$$= \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2), \tag{157}$$

where

$$\tilde{\sigma}^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} \tag{158}$$

and

$$\tilde{\mu} = \tilde{\sigma}^2 \left(\frac{N\langle x \rangle}{\sigma^2} + \frac{\mu_A}{\sigma_A^2} \right), \tag{159}$$

where $\langle x \rangle = \sum_i x_i / N$. The variance of $p(A|\mathbf{x})$ is clearly given by $\tilde{\sigma}$ and the mean by $\tilde{\mu}$. The variance, $\tilde{\sigma}$, is dominated by the smaller variance of the two variances the one given by the data and the one given by the *prior*. One can also see that the new mean is the weighted some of the two means such that $\tilde{\mu} = f\langle x \rangle + (1 - f)\mu_A$ where

$$f = \frac{\frac{N}{\sigma^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}}. \tag{160}$$

Obviously, $f \approx 1$, i.e., dominated by the data, in case the $\frac{\sigma^2}{N} \ll \sigma_A^2$ and $f \approx 0$, i.e., dominated by the *prior*, in case $\frac{\sigma^2}{N} \gg \sigma_A^2$. Figure 15 shows the behavior of $p(A|\mathbf{x})$ in the case in which the *prior* dominates (left panel) and the data dominate (right panel); whereas the case in which both are equally important is shown in the middle panel.

This example shows a general property of the *posterior*, $p(\boldsymbol{\theta}|\mathbf{x})$. It is a compromise between the *prior* and the likelihood. When the data quality is very good then it dominates over the *prior* and vice versa. The other interesting quantity that one should pay attention to is the *evidence*. This basically measures the suitability of the *prior* to the data. The larger the evidence the more the model is suitable.

Maximum A Posteriori (MAP) Estimation

A maximum a posteriori (MAP) estimate is the mode of the posterior distribution. It is similar to the method of maximum likelihood method except that the function we seek to maximize incorporates in it the *prior*. Formally, the MAP estimator is defined as,

$$\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\boldsymbol{\theta}|\mathbf{x}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}); \tag{161}$$

or equivalently,

$$\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\ln p(\mathbf{x}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \right]. \tag{162}$$

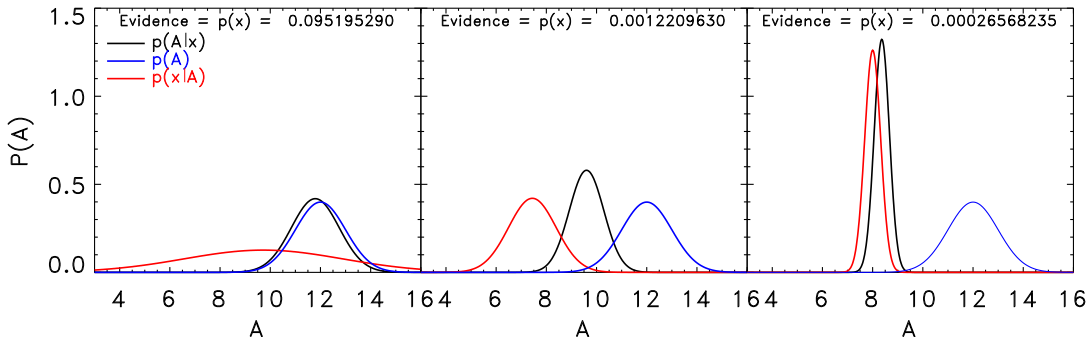

Figure 15: This figure shows three cases of the Gaussian data (red solid line) and Gaussian prior (blue solid line). The Bayesian updated probability is show with the solid black line. The three panels shows the cases in which the *prior* dominates (left panel), both *prior* and data are equally important (middle panel) and the case in which the data is very good and dominates over the *prior*. The evidence, $p(\mathbf{x})$ is also shown in each panel. Clearly, the evidence is the largest in the first case when the model is dominant. In general the evidence measures the quality of the model with respect to the data.

One of the properties of the posterior is that it is dominated by the $p(\mathbf{x}|\boldsymbol{\theta})$ at the limit of infinite amount of data. Therefore, it is clear that asymptotically the MAP estimator is the same as the MLE.

$$\lim_{N\to\infty} \hat{\boldsymbol{\theta}}_{MAP} = \lim_{N\to\infty} \hat{\boldsymbol{\theta}}_{MLE} = \boldsymbol{\theta}_0. \tag{163}$$

As an example consider N *iid* data points that are drawn from an exponential distribution,

$$p(x_i|\theta) = \begin{cases} \theta e^{-\theta x_i} & \text{if } x_i \geq 0; \\ 0 & \text{if } x_i < 0. \end{cases}$$

Since all the data points are *iid*s,

$$p(\mathbf{x}|\theta) = \prod_{i=1}^{N} p(x_i|\theta). \tag{164}$$

Let's also assume that the *prior* is also given by an exponential distribution,

$$p(\theta) = \begin{cases} \lambda e^{-\lambda\theta} & \text{if } \theta \geq 0; \\ 0 & \text{if } \theta < 0. \end{cases}$$

We now want to obtain the MAP estimator by maximizing,

$$\ln p(\theta|\mathbf{x}) = \ln p(\mathbf{x}|\theta) + \ln p(\theta) = N\ln\theta - N\theta\langle x\rangle + \ln\lambda - \lambda\theta \tag{165}$$

then

$$\frac{d}{d\theta} p(\theta|\mathbf{x})|_{\theta=\hat{\theta}_{MAP}} = \frac{N}{\hat{\theta}_{MAP}} - N\langle x\rangle - \lambda = 0 \tag{166}$$

The MAP estimator is then

$$\hat{\theta}_{MAP} = \frac{N}{N\langle x\rangle + \lambda} = \frac{1}{\langle x\rangle + \lambda/N}. \tag{167}$$

Again here at the limit of large amount of data the data dominates whereas when the data quality is poor the estimator is dominated by the *prior*. See Figure 16.

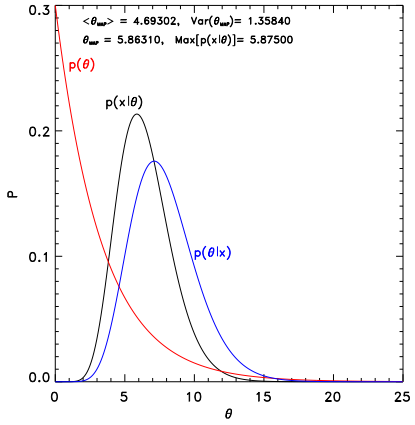

Figure 16: This figure shows the exponential example for a data set with $N = 10$ and drawn from a $\theta_0 = 5$. exponential distribution. For the *prior* a $\lambda = 0.3$ exponential distribution is used. The figure shows $p(\mathbf{x}|\theta)$ (blue solid line), the *prior*, $p(\theta)$ (red sold line) and the *posterior*, $p(\theta|\mathbf{x})$, (black solid line). The figure shows the MAP estimator both from the Equation 167 and actual position of the maximum in the figure. It also shows the mean and variance of the MAP from 1000 Monte Carlo realization of 10 data points that follow the same distribution. Notice that $p(\mathbf{x}|\theta)$ is getting close to Gaussian due to the Central Limit theorem.

There are many aspects that we can discuss in Bayesian inference unfortunately however, we have to skip many of them. However, we will return to linear Bayesian estimation when we discuss Wiener filtering in the last week of the course.

5 SINGULAR VALUE DECOMPOSITION AND PRINCIPAL COMPONENT ANALYSIS

5.1 *SVD*

Every matrix, $\mathbf{M}$ of rank $m \times n$ can be decomposed in the so called Singular-Value Decomposition,

$$\mathbf{M}_{m\times n} = \mathbf{U}_{m\times m}\mathbf{W}_{m\times n}\mathbf{V}^T_{n\times n}, \tag{168}$$

Where $\mathbf{U}$ is a unitary $m \times m$ matrix consist of so called left-singular vectors which are the eigenvectors of the matrix $\mathbf{MM^T}$. The matrix $\mathbf{V}$ is $n \times n$ unitary matrix that consists of so called right-singular vectors which are the eigenvectors of the matrix $\mathbf{M^TM}$. The matrix $\mathbf{W}$ is a diagonal matrix whose diagonal terms are call the Singular Values.

It is easy to show the relation between the SVD and the eigenvalues and vectors of $\mathbf{M^TM}$ and $\mathbf{MM^T}$:

$$\mathbf{M^TM} = \mathbf{VW}^T\mathbf{UU}^T\mathbf{WV}^T = \mathbf{VW}^T\mathbf{WV}^T. \tag{169}$$

Therefore V gives the orthogonal matrix that diagonalize $\mathbf{M^TM}$ and gives the eigenvalues of square the singular values. The same goes for $\mathbf{MM^T} = \mathbf{UWW}^T\mathbf{U}^T$.

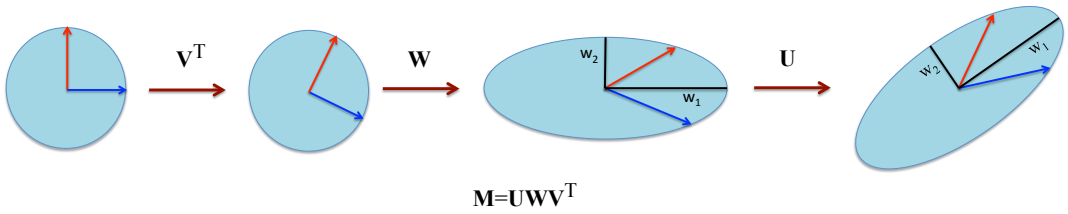

Figure 17: The Figure shows how the SVD works. Where the matrix $\mathbf{V}^T$ rotates the orthonormal vectors shown in red and blue that define a circle with unity radius), then the matrix $\mathbf{W}$ stretches the circle along the new x and y directions by the singular values w_1 and w_2, respectively. Finally, the matrix $\mathbf{U}$ rotates the two rotated and stretched vectors (i.e. the new ellipse) into final shape that gives the matrix $\mathbf{M}$ which has the new rotated and stretched vectors (shown still in red and blue) as its columns. This also applied to higher rank matrices.

One of the main uses of SVD is finding the ”inverse” of a matrix. This is a generalization of the inverse concept to the the case non-square matrices or singular matrices.There are a number of generalizations of the concept of inverse-matrix. Here we define the so called Moore-Penrose pseudoinverse normally referred to as pseudoinverse. For matrix $\mathbf{A}$ the pseudoinverse matrix $\mathbf{A}^+$ is defined as,

$$\mathbf{A}^+ = \mathbf{VW}^+\mathbf{U}^T = \mathbf{V}\,diag(w_1^{-1},\ldots,w_r^{-1},0,\ldots,0)\,\mathbf{U}^T, \tag{170}$$

where r is the rank of the matrix $\mathbf{A}$. The pseudoinverse satisfies the following four equalities,

1. $\mathbf{AA}^+\mathbf{A} = \mathbf{A}$
2. $\mathbf{A}^+\mathbf{AA}^+ = \mathbf{A}^+$
3. $(\mathbf{AA}^+)^T = \mathbf{AA}^+$
4. $(\mathbf{A}^+\mathbf{A})^T = \mathbf{A}^+\mathbf{A}$

This pseudoinverse has a lot of properties that one can discuss. Here we will mention one that is relevant to our problem. If the columns of $\mathbf{A}$ are linearly independent (so that number of data points is larger than number of parameters) then $\mathbf{AA}^T$ is invertible. In this case the formula for the pseudoinverse is,

$$\mathbf{A}^+ = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T \tag{171}$$

The application that is relevant to us here is the Linear Least Square problem which is defined as the solution, $\boldsymbol{\theta}$, of the equation

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} \tag{172}$$

minimize the Euclidean norm

(173) $\|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|$.

It is straightforward to show that this solution is given by the pseudoinverse of $\mathbf{H}$, namely,

(174) $\hat{\boldsymbol{\theta}} = \mathbf{H}^{+}\mathbf{x} = (\mathbf{H}^{T}\mathbf{H})^{-1}\mathbf{H}^{T}\mathbf{x}$

which is the same as the solution found earlier in Equation 101. See LSE examples given in the exercise session and the application to the so called Total Least Square problem.

Another use of SVD is in finding the main functions that contribute to a data set. Assume that one has a set of measurements that are produced from certain functions. and one is given the correlation function of the data. In the example here I use data produced from a random combination of random combination of the vectors shown is Figure 18 which are various powers of $\sqrt{\exp(-x^2)}$ and x and their combinations. The correlation matrix of the data is given and is shown in the left panel of Figure 19.

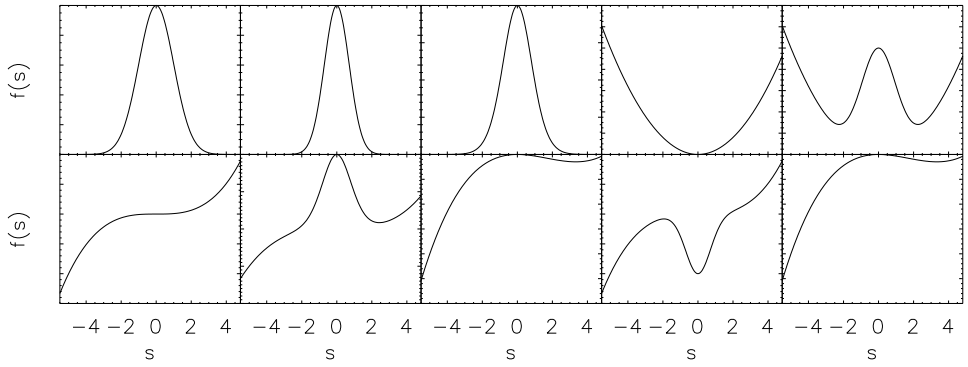

Figure 18: Data that is composed by a random combination of the vectors shown is this figure which are the following functions $\sqrt{\exp(-x^2)}$, $\exp(-x^2)$, $x^2$and x^3 and their combinations. In other words, we have only 4 independent functions from which the data is constructed.

The left panel of Figure 19 shows the correlation function of the data which as we said is given. We then apply Singular value decomposition of this matrix and obtain that singular valued which we rank order and plot in the right panel of Figure 19 which also reveals that there are only 4 significant singular values in the data.. We also show the Left Singular vectors that compose the matrix $\mathbf{U}$ which clearly show that there are only 4 independent Left Singular vectors, revealing the true amount of information in the data the other ones are clearly dominated by noise.

Another use of SVD is to stabilize the inversion. Once can clearly see from the right panel of Figure 19 that if we want to invert the singular values then the result will be dominated by the lowest values which clearly dominated by the noise and this will make the inverse explode, namely, make it unstable. A usual trick that is used in this type of analysis is to ignore the value of the inverse of these very low singular values and set their inverse to zero. This stabilizes the inversion in a natural way.

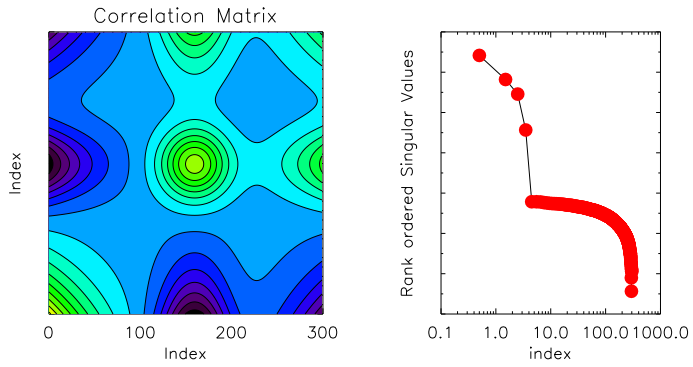

Figure 19: Rank ordered singular values. The signature of the noise is clear in the plot.

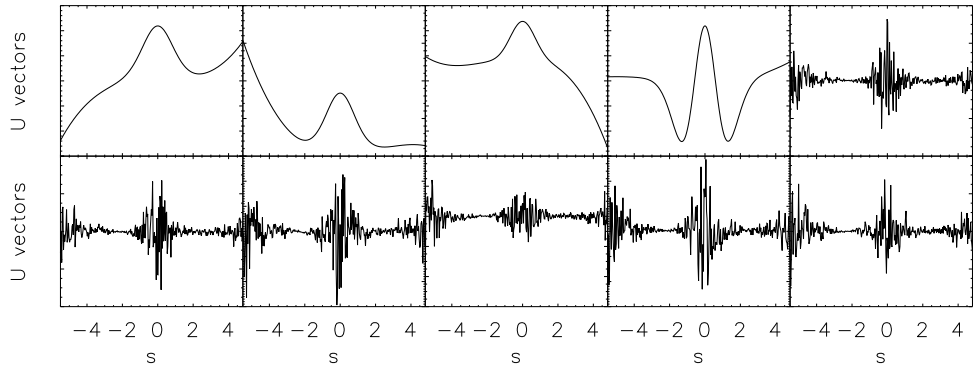

Figure 20: The Left Singular Vectors of the data showing the dominant vectors of the data. There are clearly only 4 independent functions that the SVD decomposition finds. The other vectors are clearly noise dominated.

5.2 *PCA*

Principal Component Analysis is

5.3 *Power Spectrum analysis*

6 PREDICTION AND FILTERING

Here we show a couple of examples for the for estimators that use some certain knowledge of the signal in order weather to filter, predict and reconstruct the underlying signal. The two cases we'll show here are the so called Wiener Filter (WF), known also as Optimal Filter, and the Matched Filter.

6.1 *Wiener (Optimal) Filtering*

The Wiener filter was first proposed by Norbert Wiener in 1949. The basic model for the relation between the underlying signal and the data, used to derive Wiener filter, is the standard one, namely,

$$\mathbf{x} = \mathbf{R}\mathbf{s} + \boldsymbol{\epsilon}, \tag{175}$$

where $\mathbf{x}$ is a vector of rank N, $\mathbf{s}$ is the underlying signal which is given by M diminutional vector, $\mathbf{R}$ is the so called response function/matrix (or Point Spread Function, or Selection Function, etc.) and $\boldsymbol{\epsilon}$ is the noise vector. The matrix $\mathbf{R}$ represents the **LINEAR** relation that connects the signal and the data and it has a rank $M \times N$. Wiener filter could be derived in a number of ways and it assume assumes knowledge of the first two moments of the field, $\mathbf{s}$. we wish to recover: namely its mean, $\left\langle \mathbf{s} \right\rangle$ (taken in what follows to be 0 for simplicity), and its covariance matrix,

$$\mathbf{S} = \left\langle \mathbf{s}\,\mathbf{s}^T \right\rangle \equiv \left\{ \left\langle s_i\, s_j^* \right\rangle \right\}. \tag{176}$$

Remember, $\left\langle ... \right\rangle$ denotes an ensemble average. Notice that no assumption has been made regarding the actual functional form of the probability distribution function (PDF) which governs the random nature of the field besides its first two moments. We define an optimal estimator of the underlying field, $\mathbf{s}^{\mathrm{MV}}$ (hereafter MV estimator), as the linear combination of the data, $\mathbf{x}$, which minimizes the variance of the discrepancy between the estimator and all possible realizations of the underlying field. This is obviously the LSE of the field. Thus one writes

$$\mathbf{s}^{\mathrm{MV}} = \mathbf{F}\,\mathbf{x}, \tag{177}$$

where the $\mathbf{F}$ is an $M \times N$ matrix chosen to minimize the variance of the residual $\mathbf{r}$ defined by

$$\left\langle \mathbf{r}\,\mathbf{r}^T \right\rangle = \left\langle (\mathbf{s} - \mathbf{s}_{MV})\,(\mathbf{s}^T - \mathbf{s}^{\mathrm{MV}\,T}) \right\rangle. \tag{178}$$

Carrying out the minimization of this equation with respect to $\mathbf{F}$ one finds the so-called WF,

$$\mathbf{F} = \left\langle \mathbf{s}\,\mathbf{x}^T \right\rangle \left\langle \mathbf{x}\,\mathbf{x}^T \right\rangle^{-1}. \tag{179}$$

The MV estimator of the underlying field is thus given by

$$\mathbf{s}^{\mathrm{MV}} = \left\langle \mathbf{s}\,\mathbf{x}^T \right\rangle \left\langle \mathbf{x}\,\mathbf{x}^T \right\rangle^{-1} \mathbf{x} \tag{180}$$

The variance of the residual of the α-th degree of freedom can be shown to be

$$\left\langle |r_\alpha|^2 \right\rangle = \left\langle |s_\alpha|^2 \right\rangle - \left\langle s_\alpha \mathbf{x}^T \right\rangle \left\langle \mathbf{x}\,\mathbf{x}^T \right\rangle^{-1} \left\langle \mathbf{x}\, s_\alpha \right\rangle \tag{181}$$

The noise term $\boldsymbol{\epsilon}$ is assumed to be statistically independent of the underlying field ($\left\langle \boldsymbol{\epsilon}\,\mathbf{s}^T \right\rangle = 0$) and therefore the correlation matrices appearing in equation 180 follow directly from equation 175 :

$$\left\langle \mathbf{s}\,\mathbf{x}^T \right\rangle = \left\langle \mathbf{s}\,\mathbf{s}^T \right\rangle \mathbf{R}^T \equiv \mathbf{S}\,\mathbf{R}^T \tag{182}$$

and

$$\left\langle \mathbf{x}\,\mathbf{x}^T \right\rangle \equiv \mathbf{D} = \mathbf{R}\,\mathbf{S}\,\mathbf{R}^T + \left\langle \boldsymbol{\epsilon}\,\boldsymbol{\epsilon}^T \right\rangle. \quad (183)$$

For the case in which $\boldsymbol{\epsilon}$ is expressed in terms of $\boldsymbol{\sigma}$ one gets,

$$\mathbf{N}_{\boldsymbol{\epsilon}} \equiv \left\langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \right\rangle = \mathbf{R} \left\langle \boldsymbol{\sigma}\,\boldsymbol{\sigma}^T \right\rangle \mathbf{R}^T \equiv \mathbf{R}\,\mathbf{N}_{\boldsymbol{\sigma}}\,\mathbf{R}^T, \quad (184)$$

$\mathbf{N}_{\boldsymbol{\epsilon}}$ and $\mathbf{N}_{\boldsymbol{\sigma}}$ are the correlation matrices of the noise $\boldsymbol{\epsilon}$ and $\boldsymbol{\sigma}$ respectively ($\mathbf{N}_{\boldsymbol{\epsilon}}$ and $\mathbf{N}_{\boldsymbol{\sigma}}$ are not necessarily diagonal). With these definitions, the expression for WF given in equation 180 becomes

$$\mathbf{F} = \mathbf{S}\mathbf{R}^T(\mathbf{R}\,\mathbf{S}\,\mathbf{R}^T + \mathbf{N}_{\boldsymbol{\epsilon}})^{-1} \quad (185)$$

or

$$\mathbf{F} = \mathbf{S}(\mathbf{S} + \mathbf{N}_{\boldsymbol{\sigma}})^{-1}\mathbf{R}^{-1} \quad (186)$$

Although, equations 185 and 186 are mathematically equivalent, equation 185 is often more practical computationally since it requires only a single matrix inversion.[2], However, if $\mathbf{S}$ and $\mathbf{N}_{\boldsymbol{\sigma}}$ are both diagonal, then equation 186 becomes easier to deal with numerically. Furthermore, equation 186 shows explicitly the two fundamental operations of the WF:[3] *inversion* of the response function operating on the data ($\mathbf{R}^{-1}$) and *suppression* the noise roughly by the ratio of $\frac{prior}{prior+noise}$ (if $\mathbf{S}$ and $\mathbf{N}$ are diagonal). Note that this ratio is less than unity, and therefore the method can not be used iteratively as successive applications of the WF would drive the recovered field to zero. A third operation that is done by the this filter is the prediction of the values of field $\mathbf{s}$ in locations in which there is no data which is mathematically done by the non-square nature of the matrix $\mathbf{R}$ which normally has $M \geq N$, this aspect we'll call prediction.

The variance of the residual given in equation 181 can be calculated easily using equation 186 . This calculation gives,

$$\left\langle \mathbf{r}\,\mathbf{r}^T \right\rangle = \mathbf{S}\,(\mathbf{S} + \mathbf{N}_{\boldsymbol{\sigma}})^{-1}\mathbf{N}_{\boldsymbol{\sigma}}. \quad (187)$$

In the rest of the paper we will consider the case where the uncertainties are expressed explicitly in the observational domain and the uncertainty matrix is assumed to be $\mathbf{N} = \mathbf{N}_{\boldsymbol{\epsilon}}$.

2c. Conditional Probability

We now consider the case where the *prior*model is extended to have a full knowledge of the random nature of the underlying $\mathbf{s}$ field, which is mathematically represented by the PDF of the field, $P(\mathbf{s})$. Knowledge of the measurement, sampling and selections effects implies that the joint PDF, $P(\mathbf{s},\mathbf{x})$, can be explicitly written. The conditional mean value of the field given the data can serve as an estimator of $\mathbf{s}$,

$$\mathbf{s}_{\rm mean} = \int \mathbf{s} P(\mathbf{s}|\mathbf{x})\,{\rm d}\mathbf{s}. \quad (188)$$

The standard model of cosmology assumes that the primordial perturbation field is Gaussian, and therefore on large scales where the fluctuations are still small the present epoch perturbations field will be very close to Gaussian. The statistical properties of the GRF depend only on its two-point covariance matrix; in particular the PDF of the underlying field is a multivariate Gaussian distribution,

$$P(\mathbf{s}) = \frac{1}{\left[(2\pi)^N {\rm det}(\mathbf{S})\right]^{1/2}} \exp\Big(-\frac{1}{2}\mathbf{s}^T\,\mathbf{S}^{-1}\mathbf{s}\Big), \quad (189)$$

determined by the covariance matrix $\mathbf{S}$.

[2] In general, the matrices are not square; in these cases inversion refers to the pseudo-inverse, e.g. as defined in terms of Singular Value Decomposition discussion.

[3] Some authors refer to the ratio, (prior/prior+noise), as the WF. However, it is not always possible to separate it from $\mathbf{R}^{-1}$; consequently our notation WF contains both the operations noise suppression and inversion of the response function.

Now, if the noise is an independent GRF, then the joint PDF for the signal and data is,

$$P(\mathbf{s},\mathbf{x}) = P(\mathbf{s},\boldsymbol{\epsilon}) = P(\mathbf{s})\,P(\boldsymbol{\epsilon}) \propto \exp -\frac{1}{2}\Big(\mathbf{s}^T\,\mathbf{S}^{-1}\mathbf{s} + \boldsymbol{\epsilon}^T\mathbf{N}^{-1}\boldsymbol{\epsilon}\Big), \tag{190}$$

while the conditional PDF for the signal given the data is the shifted Gaussian,

$$P(\mathbf{s}|\mathbf{x}) = \frac{P(\mathbf{s},\mathbf{x})}{P(\mathbf{x})} \propto P(\mathbf{s})\,P(\boldsymbol{\epsilon}) \propto \exp\left[-\frac{1}{2}\Big(\mathbf{s}^T\,\mathbf{S}^{-1}\mathbf{s} + (\mathbf{x}-\mathbf{R}\mathbf{s})^T\mathbf{N}^{-1}(\mathbf{x}-\mathbf{R}\mathbf{s})\Big)\right]. \tag{191}$$

Note also that the second term in the exponent, in equation 194 , is $-\frac{1}{2}$ the classical χ^2 distribution. Following RP and Bertschinger (1987) we rewrite equation (194) by completing the square for $\mathbf{s}$:

$$P(\mathbf{s}|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\big(\mathbf{s}-\mathbf{S}\mathbf{R}^T(\mathbf{R}\mathbf{S}\mathbf{R}^T+\mathbf{N})^{-1}\mathbf{x}\big)^T\big(\mathbf{S}^{-1}+\mathbf{R}^T\mathbf{N}^{-1}\mathbf{R}\big)\big(\mathbf{s}-\mathbf{S}\mathbf{R}^T(\mathbf{R}\mathbf{S}\mathbf{R}^T+\mathbf{N})^{-1}\mathbf{x}\big)\right]. \tag{192}$$

The integral of equation 12 is trivially calculated now to yield $\mathbf{s}_{\rm mean} = \mathbf{s}^{\rm MV}$, the residual from the mean coincides with $\mathbf{r}$, which is Gaussian distributed with a zero mean and whose covariance matrix is $\big(\mathbf{S}^{-1} + \mathbf{R}^T\mathbf{N}^{-1}\mathbf{R}\big)^{-1}$. The important result is that for GRFs the WF minimal variance reconstruction coincides with the conditional mean field.

Another estimator can be formulated from the point of view of Bayesian statistics. The main objective of this approach is to calculate the *posterior* probability of the model given the data, which is written according to Bayes' theorem as $P(\mathrm{model}|\mathrm{data}) \propto P(\mathrm{data}|\mathrm{model})P(\mathrm{model})$. The estimator of the underlying field (i.e., model, in Bayes' language) is taken to be the one that maximizes $P(\mathrm{model}|\mathrm{data})$, which is the most probable field. The Bayesian *posterior* PDF is given by:

$$P(\mathbf{s}|\mathbf{x}) \propto P(\mathbf{s})P(\mathbf{x}|\mathbf{s}), \tag{193}$$

now, in the general case which is given by equation 175, where the *prior* given in Eq. 191 assumed to be a Gaussian, equation give:

$$P(\mathbf{s}|\mathbf{x}) \propto \exp -\frac{1}{2}\Big(\mathbf{s}^T\,\mathbf{S}^{-1}\mathbf{s} + (\mathbf{x}-\mathbf{R}\mathbf{s})^T\mathbf{N}^{-1}(\mathbf{x}-\mathbf{R}\mathbf{s})\Big) \tag{194}$$

the Bayesian estimator, as it corresponds to the most probable configuration of the underlying field given the data and Gaussian *prior*, coincides with the $\mathbf{s}^{\rm MV}$.

In summary, maximizing equations 194 with respect to the field $\mathbf{s}$, yields yet another estimator, namely the maximum *a posteriori* estimate (MAP) of the field; it is easily shown that the MAP estimator coincides with the WF and conditional mean field i.e., $\mathbf{s}^{\rm MV} = \mathbf{s}_{\rm mean} = \mathbf{s}^{MAP}$.

WF Time Series Example: Deconvolution of Noisy Data With Gaps

Here we show a time series example of the performance of the WF estimator. Let $\mathbf{s}$ be a random Gaussian time series, with a known correlation function, which we would like to measure in the time range $[0-200]$ (the signal and time units are arbitrary). The correlation function of the signal is given by

$$\xi(\Delta t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P(\omega)e^{i\omega\Delta t}d\omega, \tag{195}$$

were ω is the angular frequency and $P(\omega)$ is power spectrum which is given. The power spectrum of the signal is,

$$P(\omega) = \frac{A_0\omega}{(\omega/\omega_0)^3+1}\ , \tag{196}$$

where $\omega_0 = 0.1$ and $A_0 = 100$. The measurement is procduced by smoothing the signal with a Gaussian

function, G, where,

$$(197)\quad G(t) = \frac{1}{\sqrt{2\pi\sigma^2_{smooth}}} \exp(-\frac{1}{2}\frac{t^2}{\sigma^2_{smooth}})$$

with $\sigma_{smooth} = 10$. The smoothed field is then uniformly sampled at about 100 positions, except for the time range of $[90-120]$ where there is a gap in the data. A measurement white noise, ϵ, with standard deviation three times larger than the signal standard deviation is added. Mathematically the data is connected to the underlying signal with $\mathbf{d} = \mathbf{R}\mathbf{s}(t) + \epsilon$ where the matrix $\mathbf{R}$ represents a convolution with the function $G(t)$. The green solid line in Fig. 21 shows the convolved underlying signal and the connected diamonds represent the measured data. The WF estimators for this case is,

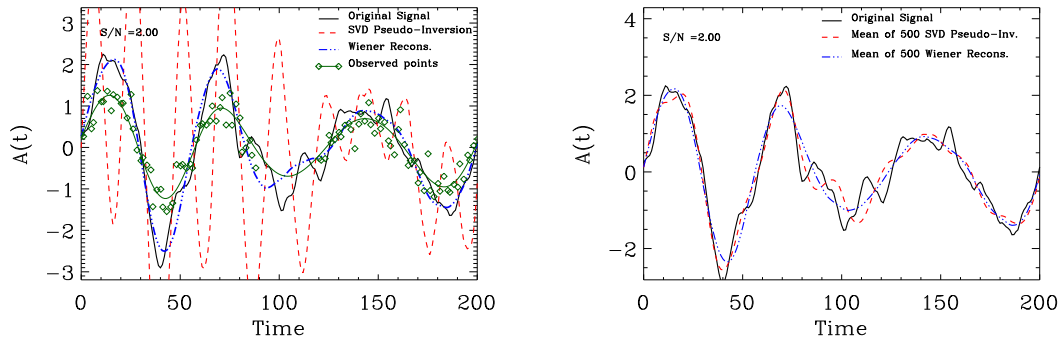

Figure 21: **Left Panel:** A one dimensional reconstruction example. The heavy-solid line shows the underlying signal $A(t)$ as a function of time; both time and amplitude has arbitrary units. The underlying signal is convolved is drawn from a correlated Gaussian Random field and is shown with solid black line. The signal is uniformly sampled with the exception of a gap in the time range of 90-120. A random noise was then added to produce the 'data' points shown with the green diamond-shaped connected points; the signal-to-noise ratio in this example is 3. The dashed red line shows an SVD Pseudo-inverse reconstructed signal, while the dashed-dotted line shows the Wiener reconstructed signal. **Right Panel:** The solid line shows the same underlying signal as the in the left panel. To this signal we add 500 noise realizations to produce Monte-Calors of the 'observed' data. The dashed line shows the mean of the 500 SVD pseudo-inversion of these realization. The dotted line shows the mean of the Wiener reconstruction of the each of each of the 500 data realizations

$$(198)\quad \mathbf{s}^{WF} = \langle \mathbf{s}\mathbf{s}^T\mathbf{R}^T \rangle \langle \mathbf{R}\mathbf{s}\mathbf{s}^+\mathbf{R}^T + \epsilon^2\mathbf{I} \rangle^{-1}\mathbf{x}.$$

We would like to deconvolve the signal from the Gaussian smoothing and recover the black solid line. A direct pseudo-inverse of $\mathbf{R}$ is unstable clearly shown in red dashed line in Fig. 21. The dotted-dashed blue line shows the WF reconstructed signal as obtained from equation 198. The figure also shows on the red dashed line a pseudo-inversion of the matrix $\mathbf{R}$ for which we used SVD method (some regularization is used here just so that the inversion does not blow up completely). The WF reconstruction is much more stable and smoother and has smaller variance than the underlying signal.

To demonstrate the differences between the two reconstructions, the right panel of Fig. 21 shows an average of 500 reconstructions of data realizations with the same underlying signal but different noise Monte-Carlos, the unbiased nature of the direct SVD inversion and the biased nature of the WF reconstructions are evident. In each SVD reconstruction we have applied the aforementioned SVD regularization, the amount of bias introduced by this procedure is clear around the extrema of the underlying signal, while the bias for the full WF application is not.

6.2 *Matched Filter*

The Matched filter has difference assumption in which we assume to know the shape of the underlying signal and we would like to find this signal's location in noisy data. Therefore, the assumption is that the

measured data is given by

(199) $\mathbf{x} = \mathbf{s} + \boldsymbol{\epsilon}$.

Our purpose here is to find the deterministic signal $\mathbf{s}$. The Matched filter assumes that we know the form of the signal, but not necessarily its position in the data. We also assume that we know the noise correlation matrix $N_{i,j} = \left\langle \epsilon_i \epsilon_j \right\rangle$ where i and j run over all the data points. Now the Matched filter is the matrix $\mathbf{F}_{MF}$ which describes a convolution with the measured data $\mathbf{x}$ so that the estimated signal is

(200) $\hat{\mathbf{s}} = \mathbf{F}_{MF}\mathbf{x} = \mathbf{F}_{MF}\mathbf{s} + \mathbf{F}_{MF}\boldsymbol{\epsilon} = \mathbf{s}' + \boldsymbol{\epsilon}'$.

In the last equation $\mathbf{s}'$ and ϵ' are the signal and noise component in the estimated signal, respectively. The Matched filter is constructed such that the contribution of the noise component is minimized relative to the contribution of the signal.

To phrase this in mathematical term we first define the vector $\mathbf{f}_i$ which corresponds to the row i in the matrix $\mathbf{F}_{MF}$, namely,

(201) $\mathbf{F}_{MF} \equiv \{\mathbf{f}_1^T, \mathbf{f}_2^T, \ldots, \mathbf{f}_i^T, \ldots, \mathbf{f}_N^T\}^T$.

Then we define the signal-to-noise ratio at point i as

(202) $SNR_i = \dfrac{\mathbf{f}_i^T \mathbf{s}\mathbf{s}^T \mathbf{f}_i}{\left\langle \mathbf{f}_i^T \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \mathbf{f}_i \right\rangle} = \dfrac{\mathbf{f}_i^T \mathbf{s}\mathbf{s}^T \mathbf{f}_i}{\mathbf{f}_i^T \mathbf{N} \mathbf{f}_i}$.

In order to maximize this terms there are a number of ways to proceed, we choose to use the one that uses the Cauchy-Schwarz inequality. To achieve this we rewrite Eq. 202 in the following way,

(203)
$$SNR_i = \frac{\left|\mathbf{f}_i^T \mathbf{s}\right|^2}{\mathbf{f}_i^T \mathbf{N} \mathbf{f}_i} = \frac{\left|\mathbf{f}_i^T \mathbf{N}^{\frac{1}{2}} \mathbf{N}^{-\frac{1}{2}} \mathbf{s}\right|^2}{\mathbf{f}_i^T \mathbf{N}^{\frac{1}{2}} \mathbf{N}^{\frac{1}{2}} \mathbf{f}_i} = \frac{\left|\left(\mathbf{N}^{\frac{1}{2}} \mathbf{f}_i\right)^T \left(\mathbf{N}^{-\frac{1}{2}} \mathbf{s}\right)\right|^2}{\mathbf{f}_i^T \mathbf{N}^{\frac{1}{2}} \mathbf{N}^{\frac{1}{2}} \mathbf{f}_i} \leq \frac{\left[\left(\mathbf{N}^{\frac{1}{2}} \mathbf{f}_i\right)^T \left(\mathbf{N}^{\frac{1}{2}} \mathbf{f}_i\right)\right]\left[\left(\mathbf{N}^{-\frac{1}{2}} \mathbf{s}\right)^T \left(\mathbf{N}^{-\frac{1}{2}} \mathbf{s}\right)\right]}{\mathbf{f}_i^T \mathbf{N}^{\frac{1}{2}} \mathbf{N}^{\frac{1}{2}} \mathbf{f}_i}.$$

The signal-to-noise ratio at point i is maximized if one attains the equality limit of the Cauchy-Schwarz inequality, which is achieved when,

(204) $\mathbf{N}^{\frac{1}{2}} \mathbf{f}_i = \alpha \mathbf{N}^{-\frac{1}{2}} \mathbf{s}$,

where α is a scalar. This final result yields the Matched filter,

(205) $\mathbf{f}_i = \alpha \mathbf{N}^{-1} \mathbf{s}$.

Clearly, α in Eq. 205 is free, so it is normally fixed by requiring the term $\mathbf{f}_i^T \mathbf{N} \mathbf{f}_i$ to be the identity matrix. Which yields the usual form of the Matched filter,

(206) $\mathbf{f}_i = \dfrac{1}{\sqrt{\mathbf{s}^T \mathbf{N}^{-1} \mathbf{s}}} \mathbf{N}^{-1} \mathbf{s}$.