Probability, Statistics, Fourier Analysis and Signal Processing

John T. Whelan

Lecture given at IUCAA, Pune, 2014 January 17

Contents

1 Probability and Statistical Inference

1.1 Logic and Probability

There are numerous interpretations of probability, but one which applies well to observational science is that of an extended logic. Let A be a proposition which could be either true or false, e.g., "The orbital period of Mars is between 686 and 687 days," "John Whelan is an Indian citizen as of Jan 1, 2020," or "My detector will collect 427 photons in the next two hours." We may know, given the information at hand, that A is definitely true or definitely false, or we may be uncertain about the answer, either because our knowledge of the situation is incomplete, or because it refers to the outcome of an experiment with a random element, which has not occurred yet. The probability of the proposition A (which we also call an "event") is a number between 0 and 1 which quantifies our degree of certainty, given the information at hand. We write this as $P(A|I)$, where I represents some state of knowledge, to emphasize that the probability we assign always depends on the information we have, the assumption that a model is correct, etc. If A is definitely true, in the context of I, then $P(A|I) = 1$. If it's definitely false, $P(A|I) = 0$.

If A represents the outcome of an experiment which we could somehow arrange to repeat under identical circumstances, then $P(A|I)$ will be approximately equal to the long-term frequency of the event A. I.e., if we do some large number N of repetitions of the experiment, at the beginning of which we recreate the situation described by I, the approximate number of experiments in which A will turn out to be true is $N \times P(A|I)$. In the classical or "frequentist" approach to statistics, this is the only sort of event to which we're allowed to assign a probability, but in the more general "Bayesian" framework we are free to assign probabilities to any logical proposition.

Several basic operations can be used to combine logical propositions:

- Negation. $\overline{A}$ is true if A is false, and vice-versa. In words, we can think of $\overline{A}$ as "not A". (Other notations include A' and $\neg A$.)
- Intersection. A, B is true if A and B are both true. In words, this is "A and B". (Other notations include $A \cap B$ and $A \wedge B$.) The advantage of the comma is that $P(A, B|I)$ is the probability that both A and B are true, given I.
- Union. $A + B$ is true if either A or B (or both) is true. In words, this is "A or B". (Other notations include $A \cup B$ and $A \vee B$.) Note the unfortunate aspect of this notation that $+$ is to be read as "or" rather than "and".

There are basic rules of probability corresponding to these logical operations:

- $P(A|I) + P(\overline{A}|I) = 1$
- The product rule: $P(A, B|I) = P(A|B, I)P(B|I)$
- The sum rule: if A and B are mutually exclusive, i.e., if $P(A, B|I) = 0$, then $P(A + B|I) = P(A|I) + P(B|I)$.

Note that in this approach, where all probabilities are conditional, the product rule is really what's fundamental. Classical approaches to probability instead define the conditional probability as $P(A|B) = \frac{P(A,B)}{P(B)}$, and therefore only entertain consideration of the conditional probability $P(A|B)$ if B is not only something to which they're allowed assign a probability, but for which that probability is nonzero.

Because the logical "and" and "or" operations are symmetrical, i.e., A, B is equivalent to B, A and $A + B$ is equivalent to $B + A$, we can write the product rule in two different ways:

$$P(A, B|I) = P(A|B, I)P(B|I) = P(B|A, I)P(A|I) \tag{1.1}$$

this can be rearranged into *Bayes's Theorem*, which says that

$$P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)} \tag{1.2}$$

which is incredibly useful when you naturally know $P(B|A, I)$ but would like to know $P(A|B, I)$. For instance, suppose A refers to "I have terrible-disease-of-the-year (TDY)", B refers to "I test positive for TDY", and I represents the information that I had no extra risk factors or symptoms for TDY but was routinely tested, 0.1% of people in such a group have TDY, the test has a 2% false positive rate (2% of people without TDY will test positive for it) and a 1% false negative rate (1% of people with TDY will test negative for it). This information tells us that:

- $P(A|I) = 0.001$ so $P(\overline{A}|I) = 0.999$.
- $P(\overline{B}|A, I) = 0.01$ so $P(B|A, I) = 0.99$.
- $P(B|\overline{A}, I) = 0.02$ so $P(\overline{B}|\overline{A}, I) = 0.98$.

Additionally, since $B = B, A + B, \overline{A}$,

$$\begin{aligned} P(B|I) &= P(B, A|I) + P(B, \overline{A}|I) \\ &= P(B|A, I)P(A|I) + P(B|\overline{A}, I)P(\overline{A}|I) \\ &= 0.99 \times 0.001 + 0.02 \times 0.999 = 0.00099 + 0.01998 \\ &= 0.02097 \end{aligned} \tag{1.3}$$

We can then use Bayes's theorem to show that

$$P(A|B,I) = \frac{0.00099}{0.02097} \approx 0.04721 \qquad (1.4)$$

I.e., if I test positive for TDY, I have about a 4.7% chance of actually having the disease. This is a lot less than $P(B|A,I)$, which is 99%!

In the context of observational science, Bayes's theorem is most commonly applied to a situation where H is a hypothesis which I'd like to evaluate and D is a particular set of data I've collected. It's usually straightforward to work out $P(D|H,I)$, the probability of observing a particular set of data values given a model, but I generally want to answer the question, what is my degree of belief in the hypothesis H after the observation. The answer, according to Bayes's Theorem, is

$$P(H|D,I) = \frac{P(D|H,I)P(H|I)}{P(D|I)} \qquad (1.5)$$

1.2 Probability Distributions

In what follows, we will often suppress the explicit mention of the background information I on which all of our probabilities are conditional. The logical propositions to which we often assign probabilities involve the values of some random or otherwise unknown quantities. So for example, $N_{\text{counts}} = 37$ or $70\,\text{km/s/Mpc} < H_0 < 75\,\text{km/s/Mpc}$. Sometimes the notation gets a bit confused between a quantity and its value, and you'll see things like X for a "random variable" and x for a value it can take on. You'd like to be able to specify the probability that $X = x$, as a function of x. In practice, this is slightly complicated by whether we think of X as taking on only discrete values, or if it can take on any value in a continuous range.

If X is discrete, we can talk about its *probability mass function* $p_X(x) = P(X = x)$. This is often just written $p(x)$ or $P(x)$. For instance, if X is the number of events in a particular interval from a stationary process in which the events are independent of one another, and the average number of events expected in the interval given the long-term event rate is μ it is described by the Poisson distribution

$$p(x) = P(X = x) = \begin{cases} \frac{\mu^x}{x!}\, e^{-\mu} & x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \qquad (1.6)$$

However, it often happens that X is continuous, so that it is vanishingly unlikely that it takes on one specific value. For instance, the height of a randomly chosen person will not be exactly 175 cm. If you measure it to more significant figures, it will turn out to be 175.25 cm or 175.24732 cm etc. So instead we want to talk about the probability for X to be in a small interval, which we call the *probability density function*

$$f(x) = \lim_{dx \to 0} \frac{P(x - \frac{dx}{2} < X < x + \frac{dx}{2})}{dx} \qquad (1.7)$$

so that

$$P(a < X < b) = \int_a^b f(x)\, dx \qquad (1.8)$$

The pdf might be called pdf(x) or even $P(x)$. A useful notation for the pdf is $\frac{dP}{dx}$, which tends to make the impact of changes of variables more obvious. In the end, it's a bit hopeless to try to stick to one letter, since you might want to talk about the joint probability distribution associated with some discrete and some continuous random variables. To give a concrete example, a common probability distribution is the Gaussian distribution with parameters μ and σ, which has pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2} \;, \qquad -\infty < x < \infty \qquad (1.9)$$

In either case, you can define an operation known as the *expectation value*

$$E\left[g(X)\right] = \begin{cases} \sum_x g(x)\,p(x) & X \text{ discrete} \\ \int_{-\infty}^{\infty} g(x)\,f(x)\,dx & X \text{ continuous} \end{cases} \tag{1.10}$$

with the mean $\mu_X = E\left[X\right]$ as a special case, and also the variance

$$E\left[(X-\mu_X)^2\right] = E\left[X^2\right] - \mu_X^2 \tag{1.11}$$

To have a sensible probability distribution, we should satisfy a normalization condition $\sum_x p(x) = 1$ or $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

1.3 Some Basic Statistical Inference

1.3.1 Bayesian Methods

Broadly speaking, the kinds of questions we'd like to answer using observational data are:[1]

- Hypothesis testing and model selection: given some observed data $\mathbf{x}$, what can we say about the relative plausibility of models $\mathcal{H}_1$ and $\mathcal{H}_2$?

- Parameter estimation: if our model $\mathcal{H}$ depends on some parameters $\boldsymbol{\theta}$, how do the data affect our judgments about the plausible values of $\boldsymbol{\theta}$?

In principle, parameter estimation is a special case of hypothesis testing, since we could consider different hypotheses corresponding to different parameter values, but in practice the notation is slightly different.

In any case, the easiest thing to construct is the probability distribution for the data given the model and any parameters: $P(\mathbf{x}|\mathcal{H},\boldsymbol{\theta},I)$. This is often used to compare models or parameter values, and as such is considered a function of $\mathcal{H}$ and/or $\boldsymbol{\theta}$ and called the *likelihood function*. It can be related to probabilities for $\mathcal{H}$ and/or $\boldsymbol{\theta}$ using Bayes's theorem.

Putting aside the question of parameters, a comparison between models $\mathcal{H}_1$ and $\mathcal{H}_2$ would be to compare $P(\mathcal{H}_1|\mathbf{x},I)$ to $P(\mathcal{H}_2|\mathbf{x},I)$, where Bayes's theorem tells us

$$P(\mathcal{H}|\mathbf{x},I) = \frac{P(\mathbf{x}|\mathcal{H},I)P(\mathcal{H}|I)}{P(\mathbf{x}|I)} \tag{1.12}$$

if we take the ratio of these two probabilities, we get

$$\frac{P(\mathcal{H}_1|\mathbf{x},I)}{P(\mathcal{H}_2|\mathbf{x},I)} = \frac{P(\mathbf{x}|\mathcal{H}_1,I)}{P(\mathbf{x}|\mathcal{H}_2,I)}\frac{P(\mathcal{H}_1|I)}{P(\mathcal{H}_2|I)} \tag{1.13}$$

To get the actual ratio, we'd need to know the ratio $\frac{P(\mathcal{H}_1|I)}{P(\mathcal{H}_2|I)}$ of probabilities that we'd assign in the absence of the data, but it's more unambiguous just to quote the factor by which the ratio changed, which is the *Bayes factor*

$$\frac{P(\mathbf{x}|\mathcal{H}_1,I)}{P(\mathbf{x}|\mathcal{H}_2,I)} \tag{1.14}$$

which is just the likelihood ratio.

If we want to assume a particular hypothesis and make a statement about the parameters in light of the observed data, we again use Bayes's theorem to construct

$$P(\boldsymbol{\theta}|\mathbf{x},\mathcal{H}) = \frac{P(\mathbf{x}|\boldsymbol{\theta},\mathcal{H})\,P(\boldsymbol{\theta}|\mathcal{H})}{P(\mathbf{x}|\mathcal{H})} \tag{1.15}$$

It is conventional to call $P(\boldsymbol{\theta}|\mathbf{x},\mathcal{H})$ the *posterior probability distribution* on $\boldsymbol{\theta}$ and $P(\boldsymbol{\theta}|\mathcal{H})$ the *prior probability distribution.* This is in some sense an artificial distinction, but in reflects the fact that the latter is adjusted in light of the data to give the former.

[1] We write both $\mathbf{x}$ and $\boldsymbol{\theta}$ as vectors to emphasize the fact that there will in general be multiple data points and multiple parameters.

Note, also, that for a hypothesis $\mathcal{H}$ concerning a model involving parameters $\boldsymbol{\theta}$ to be complete, it must also specify a prior probability distribution $P(\boldsymbol{\theta}|\mathcal{H})$ for the parameters themselves. Finally, note that if we have the numerator of (1.15) as a function of $\boldsymbol{\theta}$, we can automatically calculate the denominator by a process called marginalization, which is basically a version of the sum rule:

$$P(\mathbf{x}|\mathcal{H}) = \int P(\mathbf{x},\boldsymbol{\theta}|\mathcal{H})\,d\boldsymbol{\theta} = \int P(\mathbf{x}|\boldsymbol{\theta},\mathcal{H})\,P(\boldsymbol{\theta}|\mathcal{H})\,d\boldsymbol{\theta} \qquad (1.16)$$

1.3.2 Frequentist and Pseudo-frequentist Methods

In the frequentist formalism, you can't define things like $P(\mathcal{H}|\mathbf{x})$; instead you have to construct probabilistic statements about the random observed data $\mathbf{X}$, and then evaluate them in light of the actual observation $\mathbf{X} = \mathbf{x}$. It's usually necessary to distill the data into a single number known as a *statistic* $y(\mathbf{X})$ (or perhaps a few statistics). For example, if you want to choose between two hypotheses $\mathcal{H}_1$ and $\mathcal{H}_0$, you construct some statistic $y(\mathbf{X})$, and if it's above some threshold value y_c, you prefer $\mathcal{H}_1$, while if it's below, you prefer $\mathcal{H}_0$. Due to the randomness in the experiment, this test of the validity of $\mathcal{H}_1$ will not be perfect, i.e., there will be some chance you picked the wrong hypothesis. This is expressed by

$$\text{false alarm probability} = P(y(\mathbf{X}) > y_c|\mathcal{H}_0) \qquad (1.17\text{a})$$

$$\text{false dismissal probability} = P(y(\mathbf{X}) < y_c|\mathcal{H}_1) \qquad (1.17\text{b})$$

If you increase the threshold, you will decrease the false alarm probability, but increase the false dismissal probability (decrease the efficiency of the test). Of course, many statistics are possible, but you'd prefer to have one which minimizes the false dismissal probability for each value of the false dismissal probability. There is a result called the NeymanPearson lemma[2] which shows that this "most powerful test" is acheived by using as your detection statistic the likelihood ratio:

$$y(\mathbf{x}) = \frac{P(\mathbf{x}|\mathcal{H}_1)}{P(\mathbf{x}|\mathcal{H}_0)} \qquad (1.18)$$

It may happen, though, that you have other reasons for wanting to use a sub-optimal detection statistic $y(\mathbf{x})$. Perhaps the likelihood ratio is too expensive or difficult to compute, or perhaps $\mathcal{H}_1$ is a *composite hypothesis* with some unknown parameters $\boldsymbol{\theta}$, and so all you have access to is $P(\mathbf{x}|\mathcal{H}_1,\boldsymbol{\theta})$. (Recall that in the frequentist approach you can't even define something like $P(\boldsymbol{\theta}|\mathcal{H}_1)$, so you can't marginalize over $\boldsymbol{\theta}$ to get $P(\mathbf{x}|\mathcal{H}_1)$.) Then you can go ahead and apply the sub-optimal frequentist test.

Suppose, though, that after calculating your statistic $y(\mathbf{x})$, you decide you want to interpret things in a Bayesian way after all. (This can happen in parameter estimation especially. There may be times when *no* physically possible set of parameters produces a high enough detection statistic, and you don't really want to do something like set a negative upper limit on an event rate or energy density.) You can go ahead and do Bayesian inference using the information available, which is now just $y(\mathbf{x})$, and construct something like $P(\mathcal{H}|y(\mathbf{x}),I)$ or $P(\boldsymbol{\theta}|y(\mathbf{x}),\mathcal{H},I)$. I like to think of it as Bayesian analysis of an experiment, where the experimental data are the output of a frequentist experiment. Of course if $y(\mathbf{x})$ was chosen well, it may be that $P(\mathcal{H}|y(\mathbf{x}),I) = P(\mathcal{H}|\mathbf{x},I)$ and you haven't lost any information by calclulating $y(\mathbf{x})$ and discarding the rest of your data $\mathbf{x}$.

[2] J. Neyman and E. S. Pearson, *Philosophical Transactions of the Royal Society A* **231**, 694 (1933)

1.4 Exercise: estimation with known Gaussian errors

Suppose we are making a series of n measurements $\{x_i\}$ of some unknown quantity θ, each of which has a known Gaussian error of standard deviation σ_i associated with it. This means that the pdf for $\mathbf{x}$ is

$$P(\mathbf{x}|\theta,\mathcal{H},I) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x_i-\theta}{\sigma_i}\right]^2\right) \tag{1.19}$$

where the hypothesis $\mathcal{H}$ includes the values of the σ_i. Show that

$$\chi^2(\mathbf{x},\theta) = \sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma_i}\right)^2 = \left(\frac{\theta-\theta_0(\mathbf{x})}{\sigma_\theta}\right)^2 + \chi_0^2(\mathbf{x}) \tag{1.20}$$

where

$$\sigma_\theta^{-2} = \sum_{i=1}^{n} \sigma_i^{-2} \tag{1.21a}$$

$$\theta_0(\mathbf{x}) = \sigma_\theta^2 \sum_{i=1}^{n} \sigma_i^{-2} x_i \tag{1.21b}$$

$$\chi_0^2(\mathbf{x}) = \sum_{i=1}^{n} \sigma_i^{-2} x_i^2 - \sigma_\theta^{-2}\theta_0(\mathbf{x})^2 \tag{1.21c}$$

This means that

$$P(\mathbf{x}|\theta,\mathcal{H},I) = \frac{1}{\prod_{i=1}^{n}\sigma_i\sqrt{2\pi}} \exp\left(-\frac{[\theta-\theta_0(\mathbf{x})]^2}{2\sigma_\theta^2} + \frac{\chi_0^2(\mathbf{x})}{2}\right) \tag{1.22}$$

Show that we can construct the posterior $P(\theta|\mathbf{x},\mathcal{H},I)$ corresponding to a specified $P(\theta|\mathcal{H},I)$ using only the statistic $\theta_0(\mathbf{x})$, and that we can construct the "evidence" $P(\mathcal{H}|\mathbf{x},I)$ corresponding to a specified prior probability $P(\mathcal{H}|I)$ (suitable for the Bayes factor construction) using only the statistic $\chi_0^2(\mathbf{x})$.

1.5 Further Reading

- Jaynes, E. T., *Probability Theory: The Logic of Science* (Cambridge, 2003)
- Sivia, D. S., *Data Analysis: A Bayesian Tutorial*, 2nd edition (Oxford, 2006)
- Gregory, P., *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge, 2005)

2 Fourier Analysis

2.1 Continuous Fourier Transforms

You're probably familiar with the continuous Fourier transform

$$\widetilde{x}(f) = \int_{-\infty}^{\infty} dt\, x(t)\, e^{-i2\pi f(t-t_0)} \tag{2.1}$$

and its inverse

$$x(t) = \int_{-\infty}^{\infty} df\, \widetilde{x}(f)\, e^{i2\pi f(t-t_0)} \tag{2.2}$$

Notes:

- Lots of conventions, but note using f instead of ω gets rid of annoying 2π normalizations.
- If $x(t)$ is really a function of time, the origin/epoch t_0 is arbitrary and has no physical meaning. If it's a function of time *difference*, then $t_0 = 0$ makes sense.

The identity

$$\int_{-\infty}^{\infty} df\, e^{i2\pi f(t-t')} = \delta(t-t') \tag{2.3}$$

is useful for proving properties of continuous Fourier transforms.

2.2 Discrete Fourier Transforms

Real data is neither continuous nor infinite in duration. Consider discretely-sampled time series data of duration $T = N\delta t$:

$$x_j = x(t_j) = x(t + j\delta t) \qquad j = 0, 1, \ldots, N-1 \tag{2.4}$$

Its discrete Fourier transform is

$$\widehat{x}_k = \sum_{j=0}^{N-1} x_j \, e^{-i2\pi f_k (t_j - t_0)} = \sum_{j=0}^{N-1} x_j \, e^{-i2\pi jk/N} \tag{2.5}$$

where $f_k = k\delta f$, and

$$\delta f \, \delta t = \frac{\delta t}{T} = \frac{1}{N} \; . \tag{2.6}$$

We can define $\widehat{x}_k$ for any integer k, but there are only N independent values, thanks to the identifications

$$\widehat{x}_{N+k} = \widehat{x}_k \qquad \text{always} \tag{2.7a}$$

$$\widehat{x}_{-k} = \widehat{x}_k^* \qquad \text{if } \{x_j\} \text{ real} \tag{2.7b}$$

This means, for a real time series $\{x_j\}$, the N real numbers in the Fourier domain are (assuming N even)

- 1 real value x_0
- $\frac{N}{2} - 2$ complex values $\{x_k | k = 1, \ldots \frac{N}{2} - 1\}$
- 1 real value $x_{-N/2} = x_{N/2}$

The identity

$$\sum_{k=0}^{N-1} e^{i2\pi(j-\ell)k/N} = N \, \delta_{j,\ell \bmod N} \tag{2.8}$$

shows us the inverse transform

$$x_j = \frac{1}{N} \sum_{k=0}^{N-1} \widehat{x}_k \, e^{i2\pi jk/N} = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \widehat{x}_k \, e^{i2\pi jk/N} \tag{2.9}$$

If we consider (2.5) to be an approximation of the integral in (2.1), we'd identify

$$\delta t \, \widehat{x}_k \sim \widetilde{x}(f_k) \tag{2.10}$$

If we plug (2.2) into (2.5) we can get the actual formula

$$\delta t \, \widehat{x}_k = \int_{-\infty}^{\infty} df \, \delta_{N,\delta t}(f_k - f) \widetilde{h}(f) \tag{2.11}$$

with a kernel

$$\delta_{N,\delta t}(x) = \delta t \sum_{j=0}^{N-1} e^{-i2\pi j \delta t \, x} \tag{2.12}$$

this is not quite a Dirac delta function for two reasons:

1. It is periodic with period $\frac{1}{\delta t}$, so it's peaked at $x = 0$, $x = \frac{1}{\delta t}$, $x = -\frac{1}{\delta t}$, etc.
2. It has an oscillatory "ringing" behavior around its peaks.

The second point is related to an issue known as spectral leakage which we won't go into; the first is known as aliasing, and it means that actually $\delta t \, \widehat{x}_k$ is a sum of not only $\widetilde{h}(f_k)$ but also $\widetilde{h}(f_k + \frac{1}{\delta t})$, $\widetilde{h}(f_k - \frac{1}{\delta t}) = \widetilde{h}^*(\frac{1}{\delta t} - f_k)$, etc. This means that to avoid confusion of different frequency components, the original time series $h(t)$ should have undergone some analog processing so that $\widetilde{h}(f)$ is negligible unless

$$-\frac{1}{2\,\delta t} < f < \frac{1}{2\,\delta t} \tag{2.13}$$

which defines the Nyquist frequency $f_{\text{Ny}} = \frac{1}{2\,\delta t}$ which is half the sampling rate $\frac{1}{\delta t}$.

3 Random Data

We'll often be interested in cases where the data $\{x_i\}$ are random with some mean and variance defined by the expectation values

$$E[x_j] = \mu_j \tag{3.1}$$

$$E[(x_j-\mu_j)(x_\ell-\mu_\ell)] = \sigma^2_{j\ell} \tag{3.2}$$

If the data are Gaussian, these are enough to define a probability density[3]

$$P(\mathbf{x}) = (\det 2\pi\boldsymbol{\sigma}^2)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\sigma}^{-2}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{3.3}$$

where $\mathbf{x}$ and $\boldsymbol{\mu}$ are column vectors made up out of $\{x_j\}$ and $\{\mu_j\}$, respectively, $\boldsymbol{\sigma}^2$ is a matrix made of $\{\sigma_{j\ell}\}$ and $\boldsymbol{\sigma}^{-2}$ is its inverse. For simplicity we'll assume the data have zero mean. We'll also start in the continuous picture; the random process associated with $x(t)$ is stationary if

$$E\left[x(t), x(t')\right] = K_x(t-t') \tag{3.4}$$

which defines the autocorrelation function $K_x(t-t')$ (in general it would have to be written $K_x(t,t')$). The Fourier transform of the autocorrelation function is the two-sided power spectral density

$$S_x^{\text{2-sided}}(f) = \int_{-\infty}^{\infty} d\tau\, K_x(\tau)\, e^{-i2\pi f\tau} \tag{3.5}$$

We can use (2.3) to show that, formally,

$$E\left[\widetilde{x}(f')^*\,\widetilde{x}(f)\right] = \delta(f-f')\, S_x^{\text{2-sided}}(f) \tag{3.6}$$

Since $S_x^{\text{2-sided}}(f) = S_x^{\text{2-sided}}(-f)$, for real $x(t)$, define one-sided PSD

$$S_x(f) = \begin{cases} S_x^{\text{2-sided}}(0) & f=0 \\ S_x^{\text{2-sided}}(-f) + S_x^{\text{2-sided}}(f) & f>0 \end{cases} \tag{3.7}$$

Unfortunately (?) this is what most GW observers mean by PSD, so formulas have an extra factor of two ($S_x(f) = 2S_x^{\text{2-sided}}(f)$), e.g.,

$$E\left[\widetilde{x}(f)^*\,\widetilde{x}(f)\right] = \delta(f-f')\,\frac{S_x(f)}{2} \tag{3.8}$$

We can translate this into a discrete Fourier transform; just as $\widehat{x}_k \sim \widetilde{x}(f_k)$, we can show

$$E\left[|\widehat{x}_k|^2\right] \sim \frac{N}{2\delta t} S_x(f_k) \tag{3.9}$$

with the usual caveats about leakage and aliasing. Now consider the case of zero-mean Gaussian data: let $\widehat{x}_k = \xi_k + i\eta_k$ and treat ξ_0, $\{\xi_k, \eta_k | k = 1 \ldots \frac{N}{2}-1\}$, $\xi_{N/2}$ as independent and Gaussian with

$$E\left[\xi_k^2\right] = E\left[\eta_k^2\right] = \sigma_k^2 = \frac{N}{4\delta t} S_x(f_k) \tag{3.10}$$

so probability density is

$$\begin{aligned} P(\{\xi_k,\eta_k|k=1\ldots\frac{N}{2}-1\}) &= \prod_{k=1}^{N/2-1} \frac{1}{2\pi\sigma_k^2} \exp\left(-\frac{\xi_k^2}{2\sigma_k^2} - \frac{\eta_k^2}{2\sigma_k^2}\right) \\ &\propto \exp(\Lambda) \end{aligned} \tag{3.11}$$

with log-likelihood

$$\Lambda \sim -\sum_{k=1}^{N/2-1} \frac{2\delta t}{N} \frac{|\widehat{x}_k|^2}{S_x(f_k)} \sim -\sum_{k=1}^{N/2-1} 2\delta f \frac{|\widetilde{x}(f_k)|^2}{S_x(f_k)} \sim -2\int_0^\infty df\, \frac{|\widetilde{x}(f)|^2}{S_x(f)} \tag{3.12}$$

[3] We'll call this $P(\mathbf{x})$ rather than $f(\mathbf{x})$ to avoid confusion with the frequency.

This means

$$P(x) \propto e^{-\frac{1}{2}\langle x|x\rangle} \tag{3.13}$$

where the inner product is

$$\langle y|z\rangle = 4\,\mathrm{Re}\int_0^\infty df\,\frac{\widetilde{y}^*(f)\,\widetilde{z}(f)}{S_x(f)} \tag{3.14}$$

The unfamiliar factor of 4 is one factor of 2 because the integral is only over positive frequencies and one because of the use of the one-sided power spectral density.

If the data vary slowly over the observation time, it may be useful to divide it into pieces of length T and Fourier transform each of them

$$\widetilde{x}_I(f) = \int_{t_{I0}}^{t_{I0}+T} dt\,x(t)\,e^{-i2\pi f(t-t_{I0})} \tag{3.15}$$

In principle, the statistical properties of different segments will be related because of the autocorrelation function $K(t-t')$. But if the correlation length–the time over which $K(\tau)$ is non-negiligible–is small compared to T, we can neglect this, and the log likelihood function would become $P(x) \propto e^{\Lambda(x)}$ with

$$\Lambda = -2\,\mathrm{Re}\sum_I\int_0^{f_{\mathrm{Ny}}} df\,\frac{|\widetilde{x}_I(f)|^2}{S_I(f)} \tag{3.16}$$