Data Science and Scientific Computation Core Course

Understanding and visualizing data

March 6, 2023

Contents

1 Introduction

More and more frequently we are faced with large datasets, e.g., long temporal signals or high-dimensional samples of interesting processes. These data can be either continuous (e.g., as is typical of physics experimental data) or discrete (e.g., DNA / text sequences, qualitative descriptors). Here we will be concerned mostly with continuous signals, although many approaches we will describe have been developed also for the discrete cases. Non-trivial statistical structure in the data, which we can model or exploit for prediction, almost by definition must be due to some sort of correlation structure, and our goal here will be to quantify and visualize the simplest types of correlation.

We will introduce and review, as necessary, basic and generically applicable approaches to characterizing a new dataset. We will want to know, for example, if the collected samples are statistically independent, or whether a certain temporal signal is stationary or not. Further, we will examine the first- and second-order statistical structure in the data. This is useful for several reasons, out of which it is worth highlighting the following three.

First, knowing the low-order statistical structure in the data often allows us to optimize further data analysis, either by picking a more appropriate method (e.g., some methods work best if the data is nearly Gaussian), by preprocessing the signals (e.g., using filtering or applying nonlinear transformations), or by visualizing the data in an intuitively informative way (e.g., projecting the data along axes with largest variance or clustering the data).

Second, datasets are often of sufficient size to permit direct sampling of first- and second-order statistics, allowing us to approach the data in an as assumption-free way as possible. Sometimes, data collection or experimental setups leave their imprints on low-order statistics and it is useful to understand how such systematic effects may impact the results of various analyses.

Third, there exist processes that can be fully described by their first- and second-order statistics: if we are lucky, then our data is described by such simple processes, but even if not, these processes may serve as good baseline or null models against which to compare real data.

1.1 Topics

- Histograms obtained by binning the data, empirical PDFs and CDFs, mean / std / skewness / kurtosis, quantile statistics.
- Two-point statistics (covariance), different normalization conventions (covariance, Pearson correlation), covariance eigendecomposition and principal component analysis (PCA).
- Subsampling and error bar estimation using boostrap / jackknife methods.
- Stationary and non-stationary processes, timescales and auto-correlation functions.
- Looking for higher-order statistical structure: K-means clustering, multi-dimensional scaling (MDS), t-SNE, and independent component analysis (ICA).

1.2 Data

This is a hands-on course. You will apply the principles from the lectures to the actual data. The used data sets are described below: understanding of the data acquisition can aid the data analysis.

Microelectrode array data. Two voltage traces (`trace1, trace2`) from two single electrodes in a microelectrode array that records extracellular activity of retinal ganglion cells. This is a (pre)amplified signal that contains a mixture of noise, background "rumbling" of cells that are a significant distance away from the electrode but nevertheless induce voltage fluctuations in it, and closeby cell(s) that, when they spike, cause a large, sharp echo of the action potential, or spike, on the electrode. Spikes are unitary neural events that neurons use for communication; they consist of stereotyped voltage fluctuations, usually $\sim 1-3$ ms long, across the neuron's membrane, and they can propagate along the neuron's axon over long distances. Usually, these spikes are signals of interest that need to be isolated from experimental data. The data vector consists of $2 \cdot 10^6$ voltage samples (in μV), sampled at $2 \cdot 10^4$ Hz, for a total of 100 seconds of recording. Data is by Olivier Marre, details in Ref [1]. We will use this voltage trace as a running example to demonstrate the topics discussed in the course.

Natural images data. An ensemble of 45 calibrated grayscale natural images from Ruderman et al, *Phys Rev Lett* **73** (1994), with resolution of 256×256 pixels, used for some homeworks.

Read the spike sorting paper by Olivier Marre et al., "Mapping a complete neural population in the retina," Journal of Neuroscience **32:** 14859–73 (2012) [1]. This is a technical paper; solving homeworks below does not require such detailed understanding. The main purposes of this reading assignment are to: **(i)** give you some background on the actual experiment where our data comes from and introduce you to the spike sorting problem; **(ii)** to demonstrate a difficult data analysis problem in a realistic setup where theoretical ideas are a useful starting point, but for an actual working algorithm one needs to pay attention to the details of the system and the apparatus; direct applications of out-of-the-box methods tend not to do well on this problem.

1.3 Why are we analyzing data and building models?

One way to answer this question is by thinking of what types of insight about the world we can extract from data. Broadly, we may be interested in **making predictions**, **testing hypotheses in a statistically rigorous fashion,** or looking for **principled understanding** of the natural phenomena. These goals are by no means exclusive, but focusing on one aspect rather than the other may lead us to pick a different data analysis or modeling approach. For example, predictive models can be very successful when evaluated on withheld / future data generated by the same process that created the training data, and if the training data is plentiful, such models can be very complex, with millions of parameters. This may lead to superb performance but very low interpretability; i.e., it may be difficult for humans to understand precisely *what* structure of the data the model has learned and made use of for prediction. In contrast, models chosen to support human interpretability are usually simpler, and their components (e.g., parameters or assumed processes) usually map in a relatively straightforward fashion to physical reality. Here, the ability to generalize to withheld data generated by the same process is desired but not the ultimate objective; most often, generalization of model predictions to qualitatively new kind of data is preferred.

Different disciplines also emphasize to various amounts different tradeoffs faced in model construction and inference. The principal balance studied by statistical learning theory is the balance between model complexity and amount of training data (to control for overfitting of the models to training data). The principal balance emphasized in natural sciences is, in contrast, the balance between the model complexity and the richness of model predictions, assuming sufficient data to actually perform model inference.

Similarly, different fields also differ in their fundamental conception about what constitutes the model and its specification. Models in statistics or machine learning thus span the range from "null models" used for hypothesis testing or linear regression models, where the hypothesis and the assumptions are made explicit, the models are interpretable and usually data is not limiting, to models that can capture arbitrary functional relationships, such as neural networks, which are powerful, have large numbers of parameters that are hard to interpret, and are usually poorly constrained by data. In biology, models are often graphical schemes resembling engineering block diagrams that summarize biological processes and their interactions. It is hard to use these models for quantitative predictions (except for certain cases where these graphical models can be mapped to, e.g., systems of differential equations for chemical kinetics but which still feature lots of typically unknown parameters). On the other hand, these models make explicit and experimentally testable qualitative predictions, e.g., what happens when a process is disrupted or a component removed from a system. Lastly, a typical model in natural sciences is, for instance, a "natural law", say, Newton's law of gravitation, with $F_g = Gm_1m_2/r^2$. While this is a very simple equation, the key to its development has been to identify what are the relevant quantities (masses, forces, distance) that enter the equation, and identify how these quantities fit with the consistency requirements from other physical laws. To "infer" or learn the law of gravity, it was also crucial to abstract away from the raw data (e.g., from data on the proverbial measurements of falling objects from the leaning tower of Pisa) the systematic effects of air friction, which dominates our everyday experience. Once learned, however, this equation provides tremendous generalization performance, which extends across spatial scales from particles invisible to the eye to stellar objects. Again, these are not typical, but rather extremal examples from each discipline, chosen to illustrate the variability of modeling approaches; the situation in reality is more mixed.

Lastly, one should also consider the appropriate scale at which the modeling and analysis should take place. Here one can differentiate between **mechanistically detailed models** and **phenomenological models**; in the first class, the observed results are explained in terms of certain "elementary processes" given by prior / theoretical knowledge (e.g., molecular dynamics, chemical reactions, etc), usually parametrized by many parameters. In contrast, in the phenomenological approach microscopic details are often abstracted away to get an effective model with a small number of parameters. Sometimes, the transition from the microscopic into effective model is done formally (as in the application of statistical physics or renormalization group apparatus). Typical examples (going from physical sciences towards life sciences) include: the mechanics of colliding hard spheres in a box (detailed mechanical model) can be understood as ideal gas (a course grained, "thermodynamic" model); metal atoms in a crystal lattice can be understood as giving rise to Ohm's law or Ising magnetism; all-atom protein dynamics can be understood phenomenologically as a network of stochastic conformational transitions; and a detailed, ODE-based model of neural spike generation (Hodgkin-Huxley model of nonlinear differential equations with 20 parameters) can be simplified into a leaky integrate-and-fire model of neural spiking (one equation with a few parameters). Often, phenomenological models can be more precise or predictive for the *selected* question of interest, but their parameters may be harder to map to experimentally controllable quantities than in case of the microscopic models.

An interesting biological example of the two scales of description is provided by the problem of the so-called Planar Cell Polarity in epithelial tissues. An epithelium is a 2D tissue (like the skin), in which cells have a well-defined top (apical) and bottom (basal) surface. Often, on such tissues, a secondary macroscopic direction emerges: a studied case has been in the fruit fly *Drosophila*, where each epithelial cell grows a hair, and all the hairs share a common orientation in the plane perpendicular to the apical-basal normal. The question of how all the cells break the symmetry and decide on a common direction for the hairs has deserved substantial modeling effort. Burak and Shraiman (*PLOS Comput Biol* **5:** e1000628 (2009)) present a phenomenological, two-equation model with ~ 5 parameters for the process, to be contrasted with the microscopically-detailed model by Amonlirdviman et al (*Science* **307:** 423 (2005)), which is specified by a system of reaction-diffusion equations with roughly 40 parameters. The issue with the latter approach is not only the fact that most of these parameters are not known and exploring robustness within such a large space is hard, but that the actual list of processes that give rise to the equations is not known to be correct or complete. On the positive side, predicting the effects of mutations in detailed models is trivial, since most mutations simply correspond to running the model with certain chemical species or reactions missing. In contrast, a single phenomenological model can encompass many microscopic models and distills down the essential components needed to explain the phenomenon, but may be harder to link to accessible experimental perturbations.

Taken together, these – very broad and philosophical – considerations nevertheless should motivate you to think very carefully about the following questions: *Why you are building a model in the first place? What do you hope to extract and learn from it? What kind of the model is most appropriate for the question at hand, or does even specifying the model mathematically already raise interesting new questions?* These considerations will strongly influence the best approach to take.

2 Basic inspection of the data

A segment of the roughly 100 s worth of data in `trace1` is shown in Fig 1. A quick look suggests that the trace can be understood as a superposition of slow baseline fluctuations (happening on a ~ 0.5 s timescale), large excursions in voltage that on the plot appear almost instantaneous and likely correspond to neural spikes that we wish to find, and the residual, fast timescale fluctuation. In the course of the lectures we will make these intuitions more precise.

Always make sure to do a "sanity check" of the data: look at the beginning and end of traces in case something unusual happens there (in case of spiking data, perhaps the data acquisition was already / still going on while the retina was not being stimulated), check for gaps in the data (e.g., due to experimental failures etc). This already requires you to have some understanding of how the data was collected and how you should interpret it (does a segment of zeros correspond to something that could actually have happened or is it a sign of experimental failure?). Sometimes different parts of the data are not of the same quality (here, the retina could be slowly dying already towards the end of the experiment), which needs to be taken into account. All these considerations are not restricted to this example: *It makes sense to invest some time in the beginning and understand the data collection process well before plugging the data into any kind of analysis!*

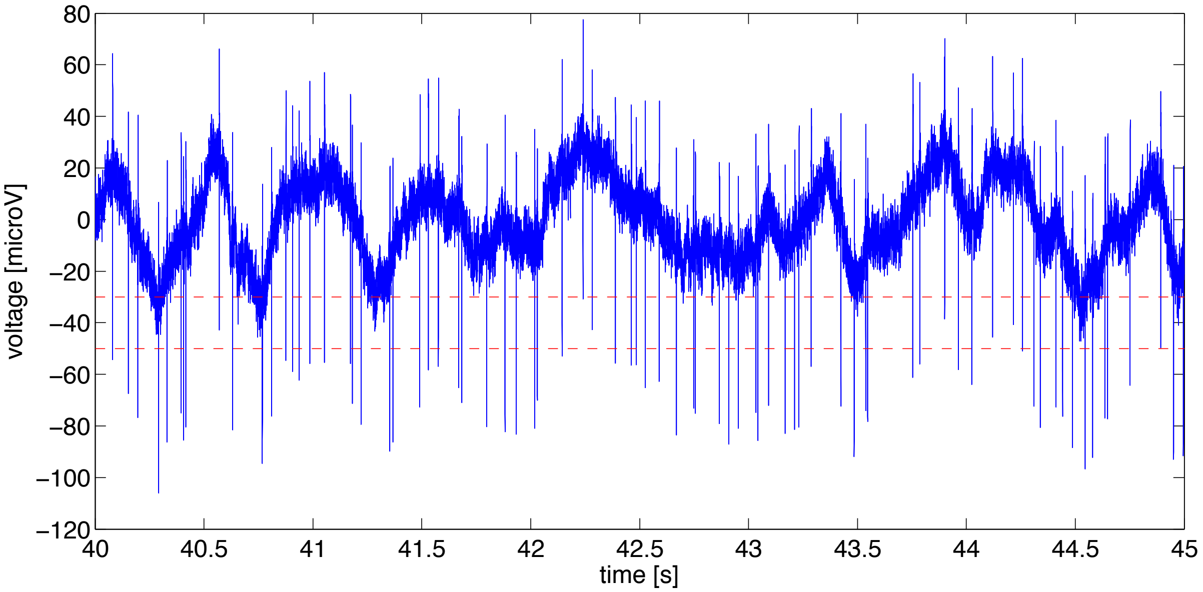

Figure 1:   A 5 s example trace of the retinal recording on the multielectrode array (MEA).

2.1 Absolute vs arbitrary units

Check what are the units in which the data is gathered: are values reported in physical units or some "arbitrary units" [a.u.]? In particular, do the absolute zero and the dynamic range have any intrinsic meaning or are they something that can be chosen arbitarily by "data centering" (i.e., subtracting the mean from the time series to center the data at 0) and "normalization" (e.g., by dividing the data by maximum or standard deviation over the time series, to get a unitless measurement)? These considerations can have very direct and practical effects. For example, if the same experiment is repeated many times (or we are comparing the signal on multiple electrodes as in our example), then — if what is being measured are really physical quantities — the signals from different electrodes or repeats should be comparable directly and plotted on the same scale. It is precisely to guarantee the same absolute zero that we ground the electrical measurements to the same reference. If the units and zero are arbitrary, i.e., a consequence of experimental measurement that can vary from repeat to repeat rather than being intrinsic to the system being measured, then the data needs to be normalized before comparing across repeats. This is very prevalent in life science, where the measurements are often quantitative (i.e., give numerical values) but not physical (numerical values don't have absolute units, but are only a proxy for some underlying physical quantity). For example, we are often interested in the protein concentration (which would have units of, e.g., nanomole per liter) but measure instead the average light intensity of a fluorescently tagged protein. The measured quantity depends not only on concentration, which is intrinsic to the biological system, but also on the microscopy settings, staining or imaging protocols, etc.

An instructive example of how important data normalization can be to scientific conclusions is provided by the measurement of Bicoid morphogen concentration gradients in developing fruit fly (*Drosophila melanogaster*) embryos. Spatial gradients of morphogens are important, since they instruct cells of a multicellular organism to differentiate into different tissues; the precision of these concentration gradients limits how precisely the cells can differentiate. In Houchmandzadeh et al, *Nature* **415:** 798 (2002) the authors measure the concentration gradients, $c_i(x)$, of Bicoid morphogen in many embryos, $i = 1, \ldots, N$, as a function of the spatial coordinate, x, that extends from anterior to posterior. Since the measurements are not of a direct physical quantity but consist of immunostaining data supposedly linearly related to the concentration, the authors first normalize the gradients before plotting them on top of each other: they subtracted the offset from every gradient so that at its minimum the gradient was zero, and divided each gradient by an amplitude so that at the maximum it was one. Then they computed the standard deviation over these gradients to quantify their embryo-to-embryo reproducibility or precision (Fig 2, left), and concluded that this precision is too small to explain the precision by which cells later decide about their differentiation fate, implying the presence of unknown other biological signals in the system. Later, Gregor et al, *Cell* **130:** 153 (2007) remeasured the same gradients and reanalyzed also the original data from Houchmandzadeh et al, by normalizing the gradients not to minimum and maximum but so that they overlapped as well as possible in the χ^2 sense. Now, the aligned gradients were much less variable in the middle, where cells decide on their fate, implying that no extra signal is needed (Fig 2, right). In retrospect, it is clear that the initial normalization to min/max squeezed out, by construction, all the variability at low and high x, and squeezed it into the middle; the χ^2 normalization by Gregor et al, in contrast, made no such biased choice. This has two implications: (i) data normalization / alignment can be very important and needs to be thought through carefully; (ii) if the measurement was of the actual physical quantity (concentration) or the absolute calibration of the staining was performed, no normalizations / alignments would be needed and all profiles could simply be compared to each other directly. This approach, however, requires much greater experimental effort; cf. Dubuis et al, *Molecular Systems Biology* **9:** 639 (2013).

Remember that similar types of considerations apply to *any* data, not just the examples from life sciences discussed here – be it genomic sequences, physics measurements, images taken from the internet as training data for image classifiers, etc. The problem of offset and scale is very generic: when referees or professors grade the students, they do so on a fixed scale (say 1 to 10), but different people grade by using different dynamic ranges (some squeeze all the grades on a scale between 8 and 10, whereas some use the whole dynamic range): how does one normalize for that? Similarly, when digitalizing signals, analog values are packed into, e.g., a discrete set of 10-bit digital values $(0, \ldots, 1023)$, losing the unit, offset and the absolute value of the dynamic range in the process, which have to be recorded separately. This is relevant, for instance, for digital images, which do not record physical quantities (light fluxes). The first step in analyzing the data is to understand such issues of data representation and biases in data collection, so that you can later asses their impact on your conclusions.

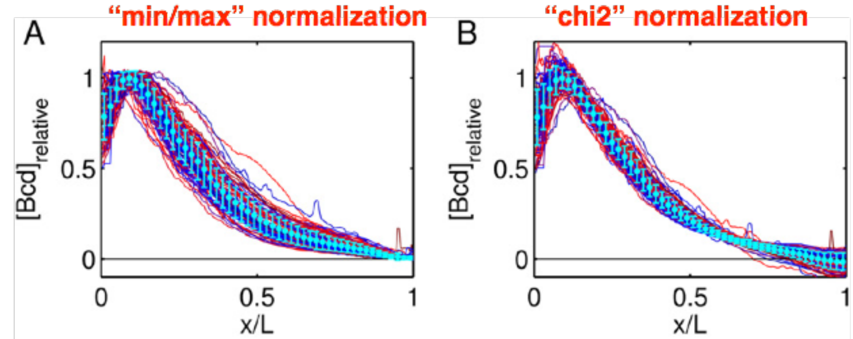

Figure 2: Comparison between normalizing each Bicoid gradient, [Bcd](x), by its minimum and maximum (to align all gradients between 0 and 1) in (A), vs aligning the gradients by adjusting an additive offset and multiplicative scale per gradient so that they best align to each other in a χ^2-sense in (B). In **(A)**, most of the variability between gradients (cyan errorbars) is squeezed from $x = 0$ and $x = 1$ towards close to the middle, at $x \sim 0.5$, whereas in **(B)** the gradients are very precise in the middle. See the corresponding text box for references and details; figure reproduced from Gregor et al, *Cell* **130:** 153 (2007).

3 Descriptive statistics

Suppose you are given N samples of some data that we will jointly denote as $\mathcal{D} = \{\mathbf{x}_t\}$, with $t = 1, \ldots, N$ indexing the samples. For now, let's think of $\mathbf{x}$ as real-valued data of dimension D, i.e., $\mathbf{x} \in \mathbb{R}^D$. In our case of the `trace1` voltage trace, the situation is simple if we view voltages as $D = 1$ dimensional scalars, and if we don't worry about the correlation between successive time points (more formally, here we will start the discussion assuming that the samples are *independent and identically distributed* (IID) – that means that the probability of observing a certain sample is independent of all the other observed samples, and that all samples are generated by the same distribution). Now, for our timeseries data, there obviously *are* temporal correlations and the IID assumption is thus patently false, but one can nevertheless look at the marginal statistics, as below (and be on a lookout about what could go wrong when making unsubstantiated IID assumptions).

We will introduce some basic notions here:

- **Empirical distribution.** $P_{\text{emp}}(\mathbf{x}) = \frac{1}{N}\sum_{t=1}^{N} \delta(\mathbf{x} - \mathbf{x}_t)$, where $\delta(\mathbf{x})$ is a Dirac-delta function, with the property that $\int_{-\infty}^{\infty} d\mathbf{x}\ \delta(\mathbf{x}) = 1$. Empirical distribution is a way to represent the given samples in $\mathcal{D}$ as a distribution composed of N infinitely sharp peaks, each corresponding to one observation. For IID data, this distribution contains equal information as the list of samples $\mathcal{D}$.

- **"True" data generating distribution** $P(\mathbf{x})$. If we repeated the same experiment or data collection, this is the distribution we would be drawing the new samples from (each time obtaining a different empirical distribution). We usually don't know P, but would like to learn about its properties from the data $\mathcal{D}$. Sometimes we make assumptions about the form of this distribution (e.g., by assuming it is a Gaussian) but treat its parameters (e.g., mean and variance) as unknowns to be determined from data; in this case we talk about *parametric statistics*. In other cases we don't want to assume the functional form for P, but nevertheless reason about it; in this case we talk about *non-parametric statistics*. For IID data, finding the true generating distribution would give us the complete statistical model of the process of interest.

- **Sample statistic** is some function that can be applied to data $\mathcal{D}$. Descriptive statistics are quantities that summarize or describe the data; estimators are functions of the data that attempt to extract or approximate the parameters of the generating distributions from data. For example, in $D = 1$

  $$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t \tag{1}$$

  is a descriptive statistic that corresponds to the *mean* of the data. In case I believed that the generating distribution is Gaussian with the mean parameter m (i.e., $\int dx\, xP(x) = m$), the formula in Eq (1), solely viewed as a function of the data, would also become the *estimator* for the mean m of the Gaussian distribution.

- **Bias and variance of the estimators.** While statistics can be any functions of the data, when they are used for estimation of the parameters of the generating distributions, there are two desired properties that good estimators should share. First, they should be *unbiased*:

  $$\langle \bar{x} - m \rangle = 0, \tag{2}$$

  i.e., on our Gaussian example, the estimator of Eq (1) averaged over many draws from the (by assumption Gaussian) distribution $P(x)$, will be equal to the true mean, m; here, the averaging over draws is denoted by brackets, $\langle \cdot \rangle$. Second, the estimator should not only be unbiased, but also efficient; intuitively, it should give the smallest possible variance around the true parameter value given some number of samples, N. In statistics, the study of good estimators is a large subject that we won't go into, and efficient estiamtors of course depend on what parameter is being estimated. But a general rule of thumb is that if samples are IID (independent, identically distributed), the variance of efficient estimators should scale as

  $$\mathrm{Var}[\bar{x}] \sim N^{-1}. \tag{3}$$

  Thus, the "error bar" on statistical estimates from N IID samples should be expected to generically decrease as $1/\sqrt{N}$. The simplest example of this scaling is shown in Fig 3. Make sure that you understand well the difference between a statistic (say, for the mean), its "error bar" (standard error of the mean; SEM), and in the Gaussian case the standard deviation of the Gaussian distribution. Standard error is a property of a particular estimator which should shrink with more samples; standard deviation of the underlying Gaussian distribution is a parameter of that distribution that is independent of the number of samples.

- **Moments of the distribution.** Other useful descriptive statistics that we give as examples here are the variance, $\sigma^2 = \frac{1}{N-1}\sum_{t=1}^{N}(x_t - \bar{x})^2$, and higher order moments. If you wonder about the denominator, $(N-1)$, in formula for variance, I side with the classic textbook *Numerical Recipes in C* by Flannery et al: any standard statistical textbook will explain why unbiased variance estimate when the mean is *a priori* unknown uses $(N-1)$ instead of N, but if this distinction is important for your data, you are anyway most likely making dangerous inferences in a data-limited regime. Higher-order moments of the distribution are defined as:

  $$M_k = \frac{1}{N}\sum_{t=1}^{N}\frac{(x_t - \bar{x})^k}{\sigma^k}, \tag{4}$$

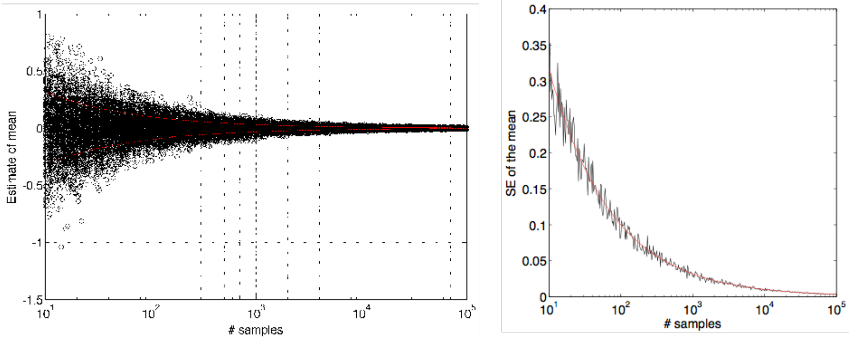

Figure 3: Individual data points are drawn IID from the standard Gaussian distribution with mean $m = 0$ and variance $\sigma^2 = 1$. **Left.** Different number of samples N are used to compute estimates of the mean according to Eq (1); each estimate is plotted as black circle. Since the estimator is unbiased, it converges to zero with increasing N, as desired. The standard deviation of the estimator is shown as the red envelope. **Right.** Standard deviation of the mean estimate, also known as the standard error of the mean (SEM), is plotted in black; superposed in red is the theoretical curve, $\mathrm{Std}[\bar{x}] = \sigma/\sqrt{N}$, a well-known formula for SEM for Gaussian distributions.

where the deviation from the mean is divided by powers of σ so that the resulting moment, M_k, is dimensionless. The well-known cases for $k = 3$ (skewness) and $k = 4$ (kurtosis) we will encounter later in the course.

4 Histograms and probability distributions

Note that for IID data all statistics (mean, moments, other functions) can be seen—and mathematically written—as functions of the empirical distribution, $P_{\mathrm{emp}}(\mathbf{x})$. A lot of information should thus be contained in a suitable representation and visualization of that distribution. To this end, we usually construct and plot raw histograms, by choosing some particular binning of the x variable (e.g., defined by the bin boundaries $x_0 \leq x_1 \leq \cdots \leq x_L$) and then plotting the number of samples, X_j, that fall into bin j. Mathematically, this corresponds to plotting

$$X_j = \sum_t (x^t \geq x_j) \wedge (x^t < x_{j+1}), \tag{5}$$

where the formula on the right-hand side simply gives a 1 if the sample falls between two bin boundaries, x_j and x_{j+1}, and 0 otherwise. This is done for `trace1` data for two different choices of data binning in Fig 4. Note that if data is intrinsically discrete, then no explicit binning is required. For continuous data, however, raw histograms with a certain choice of binning scheme give rise to several considerations:

1. Density estimation (estimating true $P(x)$ from finite number of samples for continuous x) is a hard problem that can only be solved given prior assumptions on P (e.g., on its smoothness). The simplest methods tend to convolve each data point, x^t, with a narrow Gaussian to smoothen the data and averaging across all data points, in a process known as kernel density estimation (KDE). This is beyond the scope of this course, but is a standard approach that has also been a topic in Methods of Data Analysis course.

2. Constructing raw histograms gives an obvious sense of statistical power in individual bins (since we see directly the number of samples per bin); on the other hand, the count values

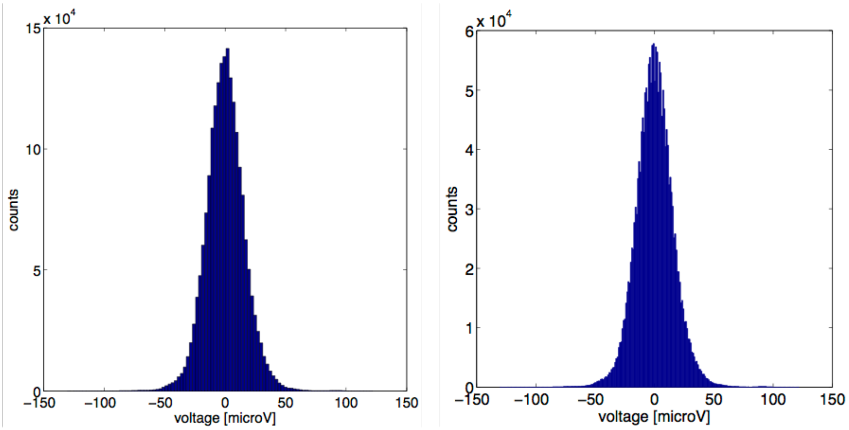

Figure 4: Raw histograms of data in `trace1`, with 100 bins **(left)** and 1000 bins **(right)**, uniformly spaced between the maximal and minimal observed values in the full trace. Interestingly, the histogram at right reveals curious "throughs" between the bins: on closer inspection it turns out that our data is not really continuous but discrete (you can check that there are only 2022 different levels in the two-million sample dataset), due to the digitalization of extracellular voltages by the recording equipment. This is a typical, rather than exceptional case, whereever physical data is captured by devices with limited resolution. For many analyses this does not matter, but for some it can: so check if your continuous-looking data is truly discrete.

are not directly comparable across the same data histogrammed using different binning widths (cf. Fig 4 left and right).

3. If bins are unequally sized, raw histograms are very hard to interpret.

4.1 Estimating PDFs

To show a distribution in a way that is independent of binning, construct an estimate of the (normalized) probability density function (PDF), $\hat{P}$, by normalizing the histogram so that:

$$\int dx\ \hat{P}(x) = 1 \tag{6}$$

If the histogram is constructed using uniform binning, i.e., $\Delta = x_1 - x_0 = x_2 - x_1 = \cdots = x_L - x_{L-1}$ with bin counts X_j in different bins, one can approximate the $\hat{P}$ as

$$\hat{P}(x_j) = \frac{1}{\Delta} \frac{X_j}{\sum_k X_k}. \tag{7}$$

Note that both histograms and estimated PDFs, Eqs (5-7), are just particular classes of descriptive statistics of the data, so that general considerations about estimation apply to them as well as to the more common statistics such as the moments.

There are a few points worth making about the choice of binning:

- You want to choose the number of bins that is high enough to capture the details of the distribution (i.e., the value of the distribution should not change much between two neighboring bins), but still small enough that each bin contains the number of counts $X_j \gg 1$, so that the empirical probability can be well estimated. There is no universal rule for doing this, and obviously there is a tradeoff between the resolution and sampling power depending on the choice of bin width, Δ.

- In the limit of $\Delta \to 0$, the estimate of the PDF, $\hat{P}$, will converge to the empirical distribution, P_{emp}, defined above. No information about data $\mathcal{D}$ is lost by such fine binning, but it provides a poor estimate to the true generating distribution, $P(x)$, since it is completely overfit to the samples.

- There are a number of small details about putting Eq (7) into practice: given that we used a discrete binning scheme to assign data points to bins, and have normalized those bins, to what value of x should the estimated value of the PDF be assigned (i.e., to the left boundary of the bin, the right one, the center, etc)? With fine enough binning this usually does not matter and the only advice is to keep in mind that the exact values depend on this consideration, but it may matter in the case of non-uniform binning (see below).

- You may choose bins that are not uniformly spaced. One popular choice is adaptive binning where bin boundaries are selected such that each bin roughly contains equal number of counts. This ensures the same statistical power in each bin, but could mean that bins differ widely in size: the bins are very densely distributed close to the peak of the distribution, and are very rare and large in the tails. In this case it does matter how bin centers are defined in the tails; for example, see Fig 5 middle right. Note that in the case of nonuniform binning of continuous variables, it is essential to plot normalized PDFs, since raw histograms no longer can be graphically interpreted. Such adaptive binning schemes are useful for higher-dimensional histograms which otherwise suffer from the curse of dimensionality. For D dimensional data uniform binning often results in large numbers of bins in the tails of the high-dimensional distribution—indeed, maybe the majority of all bins—being completely empty, which can complicate certain analyses. Instead, one can adaptively make the tail bins very large: this trades resolution in the tails for statistical power.

An example PDF estimate for our data is shown in Fig 5 at left, created from a normalized 100-bin histogram. A necessary step in inspecting the empirical distributions is to always plot them on a logarithmic scale. This quickly reveals interesting structure in the tails, as shown in Fig 5 at middle left: we can guess that the excess of low voltage excursions are exactly due to the spiking events in our electrode trace, and a naive way to set a threshold for discriminating spikes from background would be to set the threshold at the minimum separating the bulk of the distribution from the low voltage peak. This threshold would roughly correspond one of the red dashed lines in Fig 1, but you should quickly be able to convince yourself that that simple criterion will miss some of the spikes and thus generate false negatives.

4.2 Cumulative distributions

An alternative way of showing the histogram that is less dependent on the binning is by means of the cumulative density function (CDF), formally defined as:

$$C(x) = \int_{-\infty}^{x} dx' \; P(x'). \tag{8}$$

Given N samples, the estimation can be done in a nearly binning-free way, by directly using the empirical distribution, $P_{\text{emp}}(x)$, in Eq (8). For example, the following Matlab script will effectively plot a CDF estimate of `data` (make sure you understand how this works; the plot is shown in Fig 5 at right):

```
plot(sort(data,’ascend’),(1:numel(data))./numel(data),’k.’);
```

CDFs are useful in particular if the underlying true distribution is mixed, i.e., contains a continuous component as well as point probability masses; in this case, no binning scheme for PDF estimation is convenient, but the CDF is well-behaved, making large discrete steps at the locations of point masses (when CDF are estimated from finite data as above, the CDF is anyway composed of unitary steps on the y axis that have a magnitude of the inverse number of samples). Note that CDFs are also unit-free and range from 0 (at minimum of x or formally at $x = -\infty$) to 1 (at maximum of x or formally at $x = \infty$). If used to compare multiple distributions on a CDF plot, the clearest comparison is around the median, but it is hardest to compare visually the tails; for that, instead, it is often convenient to plot $1 - C(x)$ on the logarithmic scale. Kolmogorov-Smirnov test that we will introduce later to compare distributions also operates on the CDFs and similarly has highest sensitivity around the median.

An advantage of the CDF for visualization is the ease with which we can directly read off the quantile statistics. In comparison with central statistics / moments, defined by mean and variance formulas and by Eq (4), quantile statistics ask about values on the x axis of the CDF at which the cumulative data crosses some threshold probability; mathematically, quantile Q_θ for the threshold θ is the value where:

$$\theta = \int_{-\infty}^{Q_\theta} dx' \; P(x') = C(Q_\theta). \tag{9}$$

For instance, median (a special name for the $\theta = 0.5$ quantile) is the value at which the CDF crosses 0.5; similarly, one can look at the points where the CDF crosses 0.25 and 0.75, and define the *interquartile range* (IQR) as the range of x-values that contain 50% of the total probability weight (that is, the range between $Q_{0.25}$ and $Q_{0.75}$). This is an example of robust statistic, since its value is independent of the presence of small number of extreme outliers—in contrast with, say, variance. A simple way to compare two distributions graphically, including in the tails, is to make a quantile-quantile plot: each point represents the value of a particular quantile in the first distribution (to be shown on x-axis) vs the value of the same quantile the second distribution (on y-axis). Another robust statistic for quantifying the spread of the distribution (although not a quantile statistic) is the *mean absolute deviation*,

$$AD = \frac{1}{N} \sum_{t=1}^{N} |x^t - \bar{x}|. \tag{10}$$

4.3 Gaussian distributions and z-scoring

Returning now to normalized PDFs, one important feature of plotting a normalized PDF on the log plot is that you can easily compare it with standard distributions, e.g., the normal distribution with mean and variance selected to match the empirical estimate from the data (you should know the Gaussian distribution by heart, including the normalization factor):

$$\mathcal{N}(x; \bar{x}, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\bar{x})^2}{\sigma^2}}. \tag{11}$$

Such a comparison is shown in Fig 5 in the middle left; together with the ability to estimate the error bars on our PDF estimates that we will discuss next, this brings us closer to being able to assess the significance of deviations between data and model PDFs, such as the normal distribution. (As a side note: despite the extremely simple nature of these suggested steps, I

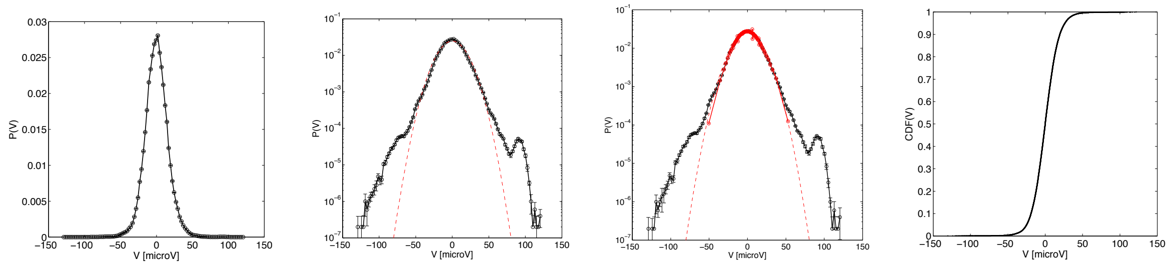

Figure 5: **Left.** Histogram of `trace1` data constructed over 100 equidistant bins and normalized into an estimate of the PDF, as in Eq (7). Note that PDFs have units, in this case $1/\mu V$, since they have to integrate to 1. **Middle left.** Logarithmic plot of the same PDF and comparison to the Gaussian distribution of matching mean and variance (dashed red line) plotted directly from Eq (11), shows the excess weight in the tails of the data. A useful reference to remember when interpreting PDF estimates is the value of the PDF that corresponds to seeing the data point once in the whole dataset (e.g., the left-most extremal values at the logarithmic plot). **Middle right.** Comparison of the distribution constructed from equally spaced bins (black) with the distribution constructed using 100 adaptive bins selected such that the number of samples per bin is equal (and the resulting error bars, see below, are also equal). This provides a good agreement near the mode of the distribution, but can lose details in the tail, where the bins span large segments of the x-axis. **Right.** Cumulative distribution of the same data can be constructed without binning and is useful for reading off quantile statistics.

regularly see students plot raw histograms also for continuous data and have trouble plotting model distributions on top of their data correctly—make sure that you can do all these steps routinely and automatically.)

At this point it is useful to also write down the generalization of the Gaussian distribution to the multivariate case,

$$\mathcal{N}(\mathbf{x};\bar{\mathbf{x}},\mathbf{C}) = (2\pi)^{-D/2}|\mathbf{C}|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^T\mathbf{C}^{-1}(\mathbf{x}-\bar{\mathbf{x}})}, \tag{12}$$

where $\mathbf{C}$ is the $D \times D$ covariance matrix and $\bar{\mathbf{x}}$ is the D-dimensional mean column vector. A few basic facts to remember about Gaussian distributions:

- They are symmetric around the mean and have a single peak, so that the mode, median, and mean coincide. They have a strong central tendency, i.e., a faster-than-exponential drop in probability with the distance from the mean.

- They are fully specified by the first (mean) and second (covariance) moments. All higher moments, M_k, can be expressed in terms of the low-order moments. All odd-order moments ($k \geq 3$, k is odd) are zero due to symmetry. For one-dimensional case:

$$M_k = \begin{cases} 0 & \text{if } k \text{ is odd} \\ \sigma^k(k-1)!! & \text{if } k \text{ is even} \end{cases} \tag{13}$$

- Integrals of the Gaussian distributions are analytically tractable; in the multi-variate case, integrating over any subset of components of $\mathbf{x}$ also results in a Gaussian (marginal) distribution.

- In 1D, interval $[\bar{x}-\sigma, \bar{x}+\sigma]$ contains $\sim 68\%$ of the total PDF weight, and $[\bar{x}-2\sigma, \bar{x}+2\sigma]$ contains $\sim 95\%$ of the weight.

- They are the most random distributions for continuous real-valued variables with a given mean and covariance; formally, they are distributions that maximize the entropy, $S[P] = -\int dx\, P(x) \log P(x)$, with given first- and second-order moments.
- Gaussian distributions are limiting distributions for sums of IID random variables that are, each, drawn from an underlying distribution with a finite mean and variance. In that case, the mean (variance) of the sum is the sum of the means (variances) of the underlying variables. This is known as the Central Limit Theorem (CLT), which also underlies many arguments about the $N^{-1/2}$ scaling of the standard error of various efficient estimators that we mentioned previously.

Figure 5 should convince you that how we visualize data is important: plotting on a linear scale reveals no surprising features, while plotting the same data on the logarithmic scale does. As illustrated by the caveat example below, the "visualization" should not be taken too far, to manipulate scientific conclusions.

> An example of data analysis where data visualization probably played a critical role for advancing a dubious scientific statement is in Bar-Even et al, *Nature Genet* **38:** 636 (2006). This paper, appearing in a reputable journal, was published there mainly due to its title claim, that "Noise in protein expression scales with natural protein abundance". More precisely, the authors implied that the variance in protein expression normalized by the squared mean expression (i.e., the coefficient of variation, CV, squared) is related to the mean protein expression linearly on a log-log plot. To show that linearity, the authors scatter-plotted the two quantities of interest on the log-log plot across many genes and many conditions in Figure 2 of their paper. These points involve a lot of scatter: so the authors first excluded the region of high and low abundance at arbitrary thresholds ("...chosen by eyeballing" from the Methods section of the paper), and then iteratively linearly fit the data, excluding outliers *and plotting those outliers in light gray in Figure 2.* This way of visualizing the points included (and arbitrarily excluded) from the fit plays well with the human visual system so as to clearly imply a linear relationship where there hardly is one.

As in the example of morphogen gradients in the fruit fly development where the gradients from different embyos had to be "normalized" and compared, we often need to compare distributions of a certain quantity, either across repeats of the same experiment or accumulated across some related phenomena. If such distributions differ in mean or variance, comparing their shape is difficult; a useful approach is to "z-score" the values, i.e., transform data x^t into:

$$z^t = \frac{x^t - \bar{x}}{\sigma_x}. \tag{14}$$

This subtracts from each data point the mean over all data and divides by the standard deviation. If the actual data $\mathcal{D}$ were drawn from a Gaussian distribution, the resulting distribution of z should be the standard zero-mean unit-variance Gaussian. If not, we can show the actual distribution of z-scored values and compare it across conditions to see if the conditions *only* differed in mean/variance, or also in the full distribution shape (i.e., higher-order moments).

Homework 1. How would you propose to generalize "z-scoring" (e.g., subtraction of the mean, normalization by the standard deviation) from the 1D case to the multivariate case, where $\mathbf{x} \in \mathbb{R}^D$? Generate a synthetic dataset with 10^4 data points drawn from bivariate Gaussian distribution with different means and standard deviations for both variables (e.g., $\bar{x}_1 = 10$, $\bar{x}_2 = -1$, and $\sigma_1 = 2$, $\sigma_2 = 1$), and for three different correlation coefficients (e.g., $\rho = 0, 0.5, 0.95$). Does your proposed transformation alter the covariance matrix?

Sometimes, z-scoring can lead to a case where distributions of values generated by similar, but not identical, processes (that differ in mean and variance) collapse onto a universal distribution. This is referred to as a "data collapse," which can indicate several important aspects about the data. One possibility is that the data acquisition method (experimental protocol) induces variation in mean/variance from repeat to repeat of the experiment and z-scoring is a way to eliminate that variation. A more interesting possibility is that the underlying process that generates the data is universal (hence has the same z-scored distribution) but is modulated by some low-dimensional factors that subsequently affect the mean and variance. An example of this is the distribution of light intensities across pixels of natural images: the distribution has a universal shape given by the laws of optics, statistics of the objects in our visual environment, and their reflectivity. This universal distribution is modulated strongly in its mean simply by the fact that we observe nature at different overall amounts of light illumination (due to variation in time-of-day etc). A less intuitive but no less intriguing example is the recent work of Brenner et al, *arxiv.org:1503.01046*, where the authors showed that the distribution of protein amount expressed from different promoters in *Escherichia coli* bacterium, accumulated from single-cell measurements across time, collapses onto a universal distribution when z-scored. The authors claim that this is not due to experimental variability, implying that the gene expression could be described by a (unknown) universal stochastic process that can be modulated by a single extra variable which jointly influences its mean and variance.

Homework 2. Plot the voltage signal $x(t)$ `trace1` from the microelectrode array and visually examine it. Spikes are very fast downward voltage excursions (sometimes reaching to $\sim -100\ \mu V$), followed by a small overshoot (zoom in to a few spikes to see how they typically look like).

To get some sense of the signal, plot a probability distribution function (properly normalized, so that $\int dx P(x) = 1$), of $x(t)$. Estimate the error bars on the PDF by splitting the data multiple times into halves and compute the SD over PDF estimates constructed from halves of the data. Is there any obvious feature for negative voltages in the histogram where you could draw a threshold to recognize the spikes easily? To identify the spikes, you can set a threshold. Scan a range of thresholds, from $-70\ \mu V$ and $-30\ \mu V$; whenever the signal crosses the threshold in a downward direction (please pay attention to this definition!), identify a putative spike, and plot the number of spikes as a function of the threshold. By examining the trace in detail, can you claim that any specific threshold is a good choice for spike detection?

The problem seems to be in slow baseline fluctuations in the recorded voltage, $x(t)$, an intuition that could be made precise using spectral (Fourier) methods. For now, estimate the slow baseline variation by smoothing the original signal over the timescale of $T = 100$ consecutive time points. The simplest way to do this is to define a new time series, $\tilde{x}(t)$, such that each value of $\tilde{x}$ corresponds to a moving average of the original time series, e.g.,

$$\tilde{x}(t) = \frac{1}{T} \sum_{t'=t-T/2+1}^{t+T/2} x(t') \tag{15}$$

Subtract this slowly varying component from the original signal. Do you see spikes more clearly now? Plot the number of detected spikes as a function of the threshold, for thresholds between $-75\ \mu V$ and $-30\ \mu V$. How dependent is the number of spikes on the threshold now?

Now we will construct a curve similar to an often-used performance measure for binary classification: the "receiver-operating characteristic" or an ROC curve[a]. In classification scenarios, one often uses a threshold to determine whether some event belongs to a particular class (is "positive" or P, when, e.g., above the threshold) or not (is "negative" or N, when, e.g., below the threshold). If we posses the "ground truth", that is, we know with certainty how each event should be assigned to P and N categories, we can compare the threshold-based classifier with this ground truth, by computing two quantities: the "true positive rate" (TP) and the "false positive rate" (FP). These quantities obviously depend on the value of the threshold, and when plotted one against the other as a function of that threshold, we obtain an ROC curve.

In our case, to see how important it is to subtract the slowly varying baseline, let's start by picking a particular threshold of $-50\ \mu V$. Then, on the baseline-subtracted trace, identify all the spikes and declare them to be correct identifications ("ground truth"). Now, go back to to the non-baseline subtracted case, and identify the spikes using different threshold values. For every spike identified on the non-baseline-subtracted trace, you can ask whether that was a true detection or not compared to your ground truth. Plot the TP and FP rates as a function of the threshold – this is a plot closely analogous to the ROC curve.

Why is this plot is closely analogous, but not exactly equal to standard ROC curve, and what is the reason for the difference? What kind of shape on TP vs FP plot corresponds to good classification performance? Related to that, check out what AOC means and how it relates to the ROC curve – this is one of the standard measures of classifier performance.

[a] Check out the Wikipedia page for ROC curve.

5 Estimating error bars

In this section we will describe how to obtain error bars on various statistics, including on the PDF estimates. Until now, we only mentioned the standard error of the mean (SEM), which is connected to the variance of the underlying Gaussian distribution (generic, since it emerges under the central limit theorem) by a known formula; and we have discussed the *scaling* of typical errors for statistics (as inverse square root of the number of *independent* samples). For quantitative estimates, however, we need **(i)** to determine not only the scaling, but the actual numerical value for the error bar given some number of samples; **(ii)** to treat non-IID samples.

A non-parametric way to estimate error bars is by means of *bootstrap resampling*. To illustrate the idea on a concrete example, let's consider the simplest estimation problem. Imagine we have collected N total samples and each sample is either white or black. In the dataset, there are M black samples, and we would like to estimate the frequency of black samples. You can see this as the simplest problem of histogram / probability distribution estimation: we are estimating probabilities of white vs black, i.e., a histogram or distribution that has support for two different values. Our considerations will generalize to examples with multiple bins.

Clearly, $\hat{p} = M/N$ is the empirical estimate for the probability of obtaining a black sample. But what is the error on that estimate? Well, if we considered a repeat experiment where we were again collecting N samples, and with probability $\hat{p}$ each would be black, we have binomial expectation for the number of black samples Z. On average, we expect $\hat{p}N$ black balls, and the variance in repeat observations in the number of black Z would be $\mathrm{Var}[Z] = N\hat{p}(1-\hat{p})$. Since on repeat experiments I would estimate probability of black as $q = Z/N = \hat{p}$, it follows that $\mathrm{Var}[q] = \mathrm{Var}[Z]/N^2$, and thus $\mathrm{Std}[q] = \frac{1}{\sqrt{N}}\sqrt{\hat{p}(1-\hat{p})}$. In case black is rare, $\hat{p} \ll 1$ (this is the typical regime relevant for constructing empirical PDFs, where we select binning such that the number of samples in any one bin is a small fraction of the total number of samples), and then the standard deviation of my estimate of probability q (which I take to be the error on the empirical estimate $\hat{p}$):

$$\mathrm{Std}[\hat{p}] = \mathrm{Std}[q] \approx \sqrt{\frac{\hat{p}}{N}}. \tag{16}$$

How would boostrapping work in this case, where the data is identically distributed and independent?

One version of bootstrap would proceed as follows:

1. Subsample the data $\mathcal{D}$, by selecting randomly only half of the data points, to get a subset $\mathcal{D}_\mu$ of size $N/2$.

2. Evaluate the statistic, f, on the subsampled data, to get f_μ.

3. Repeat the procedure 1-2 multiple times (say, 10-100 times) and approximate the error bar on f as $\mathrm{Std}[f_\mu]$ over subsamples μ.

Let's check what we would expect in our problem with black and white samples. When we draw a subset of half the samples, above, let's consider how we draw black samples assuming, again, that they are rarer than white ($\hat{p} \ll 1$). Then I am sampling each of M black samples with a probability 1/2 into my subsample; the rest of the samples will be white. On average, I expect $\frac{1}{2}M$ black balls in my subsample, with the variance of $\frac{1}{4}M$ from binomial theorem across the subsamples. From this, you can estimate the variance in the estimate of probability of black across subsamples (with factors of 2 cancelling out), to get again $\mathrm{Std}[\hat{p}] \approx \sqrt{\frac{\hat{p}}{N}}$, as above.

The key to bootstrapping is to look at the variability in the statistic over artificially constructed subsets of data and use that as a proxy for the variability over the complete dataset. Clearly, this is only a rough estimate, but we can see how well it does for estimating the errors on PDFs, as illustrated in Fig 6. In that case, as above, we can reason what the errors should be independently of the resampling process: namely, if we observe X_j counts in bin j in the raw histogram and can think of those counts approximately as multinomial or Poisson samples (the distinction is not crucial so long as no single bin contains the large majority of the sampled events — e.g., under the same assumption as our simple example above), then the error on the

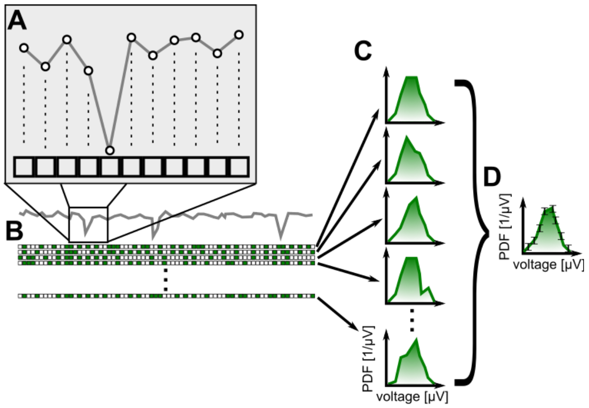

Figure 6: **Schematic of bootstrapping by selecting halves of data to determine error bars on empirical PDF estimate.** This prescription makes sense for uncorrelated data (independent samples). For time series data, such as our voltage example, the independence assumption is likely violated. **(A)** Example time series data (gray line). Zoom in: dots represent measurements, squares below illustrate uniform discrete sampling in time. **(B)** Construction of random resamplings without replacement. Each resampling (a single row) consists of a choice of half of the time points (denoted by green boxes), chosen randomly. **(C)** Given the selected time points, each resampling yields a PDF, estimated over an identical binning grid. **(D)** Standard deviation across individual resampled PDFs yields error bars on the PDF bins.

raw histogram in bin j should be $\approx \sqrt{X_j}$, from which you can derive the error in the estimated PDF. We can then compare whether the simple bootstrap estimates of the error, following the above prescription, match the Poisson expectation in Fig 7. While the match in Fig 7 looks encouraging and the deviation of our PDF estimate from Gaussian clearly looks significant, we should bear in mind the following caveats:

- The bootstrap prescription described above chooses random halves of the data, but the fact that we are choosing halves (not thirds, etc.) appears not to enter our reasoning. In fact, that is not entirely true: you can check analytically that if exactly halves of the data are chosen without replacement for the bootstrap outlined above, the error estimates from the bootstrap will agree with the poisson expectation.

- The poisson estimate is not exact if the total number of samples observed, N, is considered fixed. With a fixed total number of samples, a better model is multinomial (i.e., where the counts drawn in each bin occur with an unknown true probability p_j, and then the expected variance in that bin is $Np_j(1-p_j)$). This would, e.g., properly account for the fact if a single bin contains all the counts, its empirical error is zero given a fixed number of samples observed, N.

- The fact that the total number of observed samples is given, the error estimates on various bins are not independent, as you can convince yourself by thinking of an extreme example where the histogram only has two bins. Estimating a histogram can thus be seen as an estimation problem where multiple statistics at the same time are being estimated, and the error structure of this estimation is fully captured by the covariance matrix of errors. In practice, for estimating PDFs this is very rarely relevant, but it may be for other types of estimates.

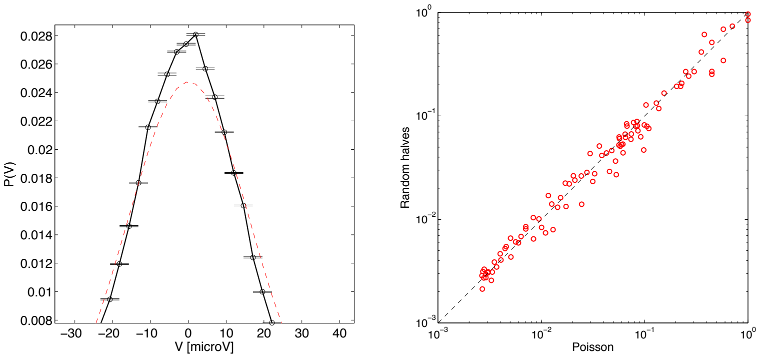

Figure 7: **Left.** Zoom-in of the peak of the histogram of `trace1` values from Fig 5 with bootstrap estimates of the error bars (black), compared to the Gaussian expectation for the peak (dashed red line). **Right.** Error estimated as standard deviation over halves of the data (y-axis) compared to the Poisson (square-root of the number of counts) based estimate on x-axis. Each point is an error bar for one of the 100 bins used to construct the PDF estimate at right.

- The major error in our application of reasoning above to the `trace1` data is the assumption of IID samples, which—as we will see—is patently false. Subsequent samples acquired at 10 kHz are by no means independent: while we nominally have 2M samples, the *effective number of independent samples* is much smaller, and our error estimates should take this into account.

Before addressing these concerns on real data, let's look at the estimation of central moments and quantiles on synthetic data, where we can guarantee that the IID sampling is correct. The goal here is to demonstrate in practice that estimates of the central moments can be very sensitive to outliers and thus many samples might be required to obtain central moments with small error bars; in contrast, quantile statistics are much more robust. In Fig 8 at left we show two very similar looking distributions: a normal (Gaussian) in red, and the log-normal distribution (where the logarithms of the random variable are Gaussian distributed) in black. Skewness, the third central moment, M_3, of Eq. (4) is zero for the Gaussian distribution and ≈ 0.3 for the lognormal distribution shown here. Kurtosis, the fourth central moment, M_4, is exactly 3 for the Gaussian distribution (which is why people often report normalized kurtosis, $K = M_3 - 3$) and 3.17 for the lognormal distribution. By these two measures, therefore, the lognormal differs only marginally from the Gaussian distribution.

Figure 8 middle shows the estimates of the skewness and normalized kurtosis as a function of the number of IID samples drawn from the lognormal distribution. Error bars are estimated as the std of these statistics over repeated independent draws. The convergence of the kurtosis and skewness to their true values is slow, with $\geq 10^3$ samples needed for the estimates to stabilize and their error bars to become significantly different from 0. This can be contrasted with the quantile estimation shown in Fig 8 at right, where the convergence of quantiles is fast and error bars comparatively small already at an order of magnitude fewer samples.

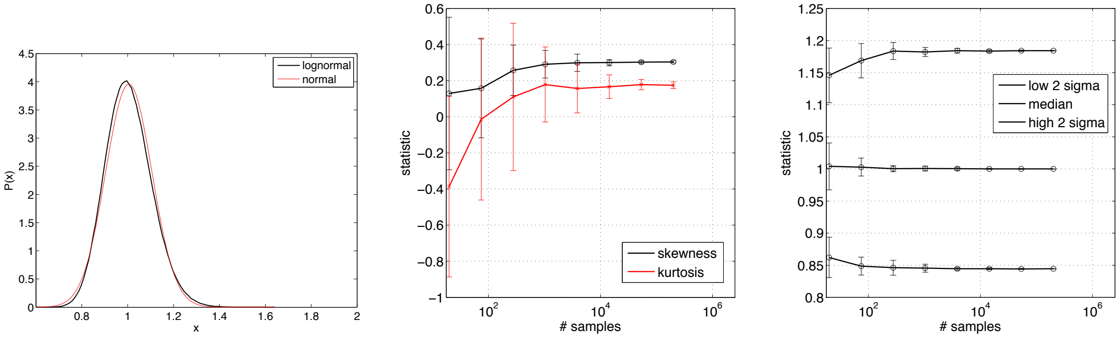

Figure 8: **Left.** Normal (red) and lognormal (black) distributions from which we will draw IID samples. $M_3 = 0$ (0.3) for the normal (lognormal, respectively) distribution; $M_4 = 3$ (3.17) for normal (lognormal, respectively) distribution. Middle: M_3 (black) and $K = M_4 - 3$ (red) estimates from the lognormal distribution. **Right.** Quantile estimates (middle line: median, top line: quantile corresponding to $+2\sigma$ Gaussian deviation, bottom line: quantile corresponding to -2σ Gaussian deviation.

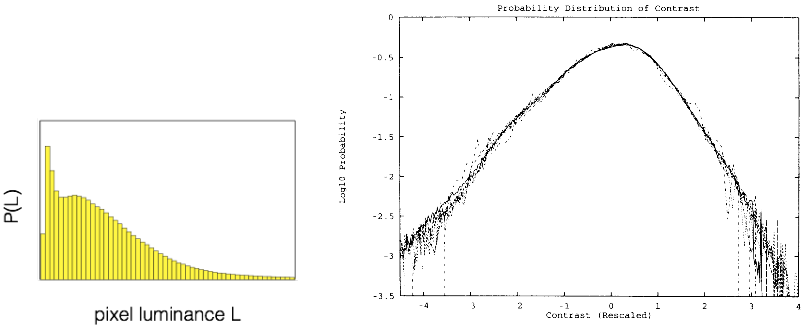

Figure 9: **Left.** The distribution of pixel luminance across a calibrated natural image dataset from Tkacik et al, *PLOS One* **6:** e20409 (2011) shows a large, right-ward skew. This is due to a combination of factors, including the time-of-day and weather variability as well as the overrepresentation of the dark patches in natural scenes due to shade and occlusion with very bright pixels. **Right.** Distribution of $\log L/L_0$, where L_0 is the mean contrast over each calibrated image; the scaling by the mean collapses the distributions sampled from multiple images onto a single distribution. Log-transform reduces the right-ward skew: the resulting distribution is much better described as central and nearly symmetric, with exponential tails. Figure at right reproduced from Ruderman et al, *Phys Rev Lett* **73** (1994).

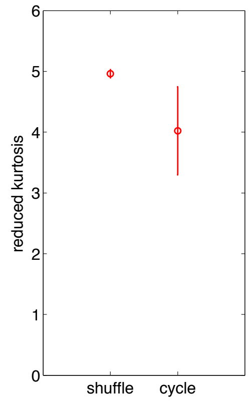

Figure 10: Mean and error bars on kurtosis, M_4, estimated by a simple bootstrap resampling procedure where subsamples of 1/20th of the data are selected randomly from the trace multiple times ("shuffle", at left) or contiguous blocks of 1/20th of the time series are taken multiple times, and the error bars are taken to be the standard deviation of the kurtosis statistic across these subsamples.

It is potentially useful to transform the raw data nonlinearly so that its distribution takes a form more amenable to analysis. Here the goal is not to reduce or reject "outliers" (i.e., large possibly wrongly measured data points) but to deal with distributions that in their raw form have special shapes that make parameter estimation difficult, e.g., long, power-law decaying tails. An intuitive example is the luminance of natural images. One can take a calibrated camera and collect natural imagery; because it is calibrated, it is possible to transform the output of (today, digital) cameras into physical units, i.e., energy fluxes at a certain wavelength. The luminance distribution over pixels of multiple images is a non-negative quantity with a very large, rightward skew, as shown in Fig 9 at left. Using a log transform and normalization per-image, as in Fig 9 at right, achieves a data collapse that we discussed in the context of gene expression before, and ensures that the resulting distribution is much better described by central moments. Sometimes, nonlinear transformations of the data can also be motivated by theoretical considerations: e.g., in chemical kinetics, the relevant quantity is often log concentration (chemical potential) as we know from the form of the relevant equations, so even if what is measured is absolute concentration, it may make sense to log-transform that before analysis. Make sure you understand analytically how distributions transform under nonlinear functions.

After this detour, let's return to estimating the error bars on statistics on real, possibly correlated data. As with your lognormal synthetic example, we can ask about the skewness and kurtosis of the `trace1` time series, to find $M_3 = 0.11$ and $M_4 = 4.97$; i.e., we see slightly higher than Gaussian skew and considerably higher kurtosis, as expected due to heavy tails. Let's now carry out a simple bootstrap resampling procedure to look at the error bars on the moments, and consider two just slightly different resampling scenarios: (i) in the "shuffle" case, select fractions (e.g. 1/20ths) by picking data points randomly; (ii) in the "cycle permute" case, take contiguous blocks of data of the same (e.g., 1/20th) size. As shown in Fig 10, this leads to error bars that are different almost by a factor of 10, a huge difference!

As you might have guessed, the difference comes because the data samples in the time trace are not IID samples: there is a strong correlation between the successive values in the time series. In the first, "shuffle," estimation, we do bootstrap resampling as if the data were IID, leading us to think that there are more independent samples than there really are. In the second, "cycle," method, we don't break any temporal correlations within the contiguous blocks that we take from `trace1`, properly maintaining temporal correlations. By comparing the two estimated error bars, we may even estimate how many independent samples there really are in the data: the ratio of both standard deviations is about 10, thus the ratio of the two variances is ~ 100. Since the variance of an estimator typically drops with N^{-1} as we discussed previously, that implies that the real number of effectively independent samples is $\sim 100\times$ less than the nominal number of samples $(2 \cdot 10^6)$, so about $2 \cdot 10^4$.

In sum, bootstrap procedure depends strongly on whether the data is correlated. For **uncorrelated, IID data**:

- Resample (with replacement, or with synthetically added noise if the noise distribution is known) from the empirical data many times N samples.
- Evaluate the error bar of the statistic of interest as a standard deviation of the statistic evaluated over resampling draws.

Alternatively, you may resample (without replacement) exactly halves of the data, as we did above; in this case, factors of 2 happen to cancel out to give you proper estimates.

For **correlated data** where the dominant source of correlation is serial (i.e., neighboring values are most strongly correlated) and of finite range, much shorter than the full dataset size:

- Split the data into contiguous blocks, for blocks of different sizes.
- Evaluate the statistic on each block and compute its Std across the blocks of a given size.
- Extrapolate the Std of the statistic, as a function of the block size, towards the whole data set size.

The last, extrapolation step is necessary to address one of the bootstrapping concerns mentioned above: the dependence on our choice of the size of the subsample. Clearly, if I compute the statistic on half of the data, I should be overestimating the error bar by roughly $\sqrt{2}$. Bootstrapping prescription above takes this scaling into account by empirical extrapolation, illustrated in Fig 11, for estimating the variance of the "residual" fast-noise part of our `trace1` signal.

Mathematical aside: Error bars for an Ornstein-Uhlenbeck process. Let's verify that the bootstrapping procedure for correlated data makes sense on data that we can synthetically generate and where we have a full control over its statistics. The simplest stochastic process that results in a Gaussian marginal distribution of values as well as exponentially decaying autocorrelation function is the so called Ornstein-Uhlenbeck process, defined by the solution of the following differential equation:

$$\dot{x} = -\gamma x + \xi(t) \tag{17}$$

where $\xi(t)$ is uncorrelated zero-mean Gaussian noise with standard deviation A, i.e., $\langle \xi(t) \rangle = 0$ and $\langle \xi^2 \rangle = A^2$. This process, defined in continuous time, can be simulated in discrete time (with proper care in how the random term is implemented); the discrete time version of this process is known in statistics as an AR(1) (auto-regressive process of order 1). The process is stationary and has a Gaussian distribution of x, with zero mean and variance $\sigma_x^2 = \frac{A^2}{2\gamma}$, as well as exponentially decaying autocorrelation function $C(t) = C_0 \exp(-\gamma t)$.

We can simulate this process numerically for three values of $\gamma = 0.1, 1, 10$, which drastically change correlation length in the timeseries, while keeping the number of samples fixed to $N_{\text{tot}} = 10^7$. We adjust A to keep the std of the distribution of x constant ($\sigma_x = 0.2$). On these three synthetic timeseries, we can now implement the bootstraping scheme to estimate the error bar on the standard deviation in x, as shown in Fig 12. We observe log-log bootstrap extrapolation slope for the error bar with inverse data set size close to -0.5, as expected for efficient estimators. Assuming independence of the timeseries, we estimate the estimator error at $E_0 \sim 4 \cdot 10^{-5}$, obviously independently of γ (and thus correlation time); this we can do by choosing random halves of the data, as discussed in the notes. Our bootstrap analysis with correlated data chunks, in contrast, suggests $E(\gamma = 1) \approx 0.00067$, $E(\gamma = 0.1) \approx 0.0023$, $E(\gamma = 10) \approx 0.00022$.

Do these values make sense? Let's first compare the $\gamma = 1$ analysis with assumed IID samples. The ratio error variances is $(E(\gamma = 1)/E_0)^2 = (6.7 \cdot 10^{-4}/4.1 \cdot 10^{-5})^2 \sim 260$. The discretization used is $\Delta t = 0.005$, which yields 200 timebins per correlation time for $\gamma = 1$. Thus, indeed, the number of independent samples is by a factor of ~ 200 smaller than the total number of samples, as observed.

How about the comparison between different error estimates for three values of γ? Observe that $(E(\gamma = 0.1)/E(\gamma = 1))^2 \sim 12$, and $(E(\gamma = 1)/E(\gamma = 10))^2 \sim 9$. Both fractions are close to 10, which is the ratio of the correlation timescales given by γ. The bootstrap error estimate therefore correctly captures the scaling of the error with the effective number of independent samples $N_{\text{eff}} \approx N_{\text{tot}} \gamma \Delta t$.

It is worth mentioning a few other uses of resampling methods:

- **Jackknife estimate of variance.** A simple way to estimate an error bar on the statistic of interest, f, is to evaluate f on the full dataset, $\mathcal{D}$, as well as on all subsamples of $\mathcal{D}$ where exactly one sample is left out, assuming IID data. The resulting error bar on the statistic is then given by

  $$\sigma_f^2 \approx \frac{N-1}{N} \sum_{t=1}^{N} \left(f_{\mathcal{D}\setminus t} - f_{\mathcal{D}} \right)^2 , \tag{18}$$

  where $f_{\mathcal{D}}$ indicates evaluation on the full data, and $f_{\mathcal{D}\setminus t}$ indicates leave-one-out evaluation. For small datasets, this is a fast approximation to the full bootstrap.

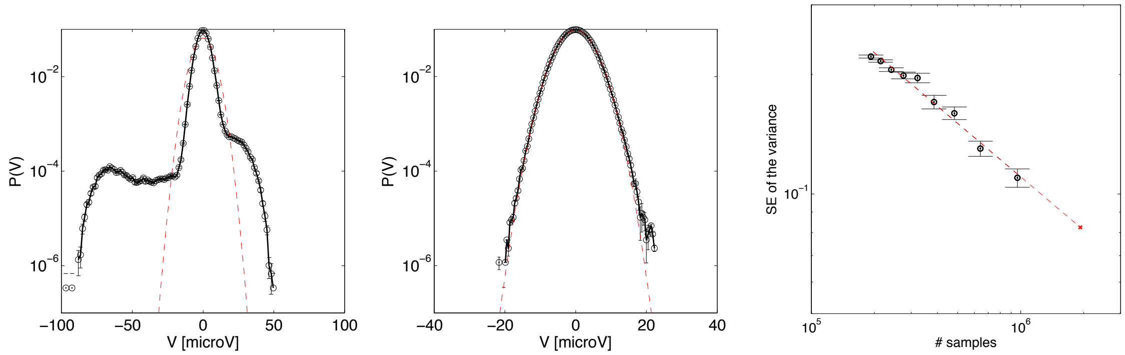

Figure 11: **Left.** Histogram of `trace1` voltage values with slow baseline fluctuations subtracted (as in the homework assignment). Note the much clearer peak at negative voltages corresponding to spiking events. **Middle.** Histogram of baseline-subtracted `trace1` signal where have further been removed, by clipping out small windows (of ~ 60 timebins in length) around the spike, matches the Gaussian distribution much more closely, with variance of $\sigma^2 \approx 16.6\ \mu\mathrm{V}^2$. **Right.** Bootstrap resampling procedure for estimating the error bar on the variance. Different data points are standard deviations of the variance estimate computed over contiguous blocks of data of the length indicated on the x-axis. Red dashed line is the linear extrapolation on the log-log plot (remember, the expectation is that the error bar on the variance decreases $\sim N^{-1/2}$, which is roughly the case here as well). We are interested in the extrapolated value (red cross) at the x-axis which corresponds to the block size equal to the total data set size, i.e., $2 \cdot 10^6$. This yields an error bar estimate of $\sigma^2 = 16.6 \pm 0.09\ \mu\mathrm{V}^2$, substantially different from the error bar estimate assuming IID, which would be $\sim 0.01\ \mu\mathrm{V}^2$.

- **Resampling for bias correction.** Some estimators are strongly biased, and the bias dominates the estimator variance (error bar) for realistic data set sizes. One such example is the naive estimator for the entropy—see the text box below.

- **Resampling for hypothesis testing.** We can often use resampling to generate null distributions against which to test certain chosen statistics. In this case, resampling procedure is crafted by hand to implement the null hypothesis. An example in Fig 13 tests whether two sets of samples have different skewness. We construct them such that they do in the case at hand, by drawing 1000 samples from normal distribution (distribution A, with true skewness zero, but due to finite sampling, any estimator will give some small nonzero value), and 1000 samples from a weakly lognormal distribution (distribution B). Let our statistic of interest be the difference of skewness ("delta skewness") between 1000 draws from each of the distributions. To create the null distribution for this "delta skewness" statistic, one can generate many resampled pairs of datasets, where a random half of the samples are taken from A and the other half from B and vice versa; by construction, such a pair of samples is statistically equivalent. We can evaluate the "delta skewness" statistic over many such pairs of samples and create the null distribution, against which one can evaluate the p-value for the statistic evaluated over the initial draw (where we compare skewness of all samples drawn from A vs all samples drawn from B).

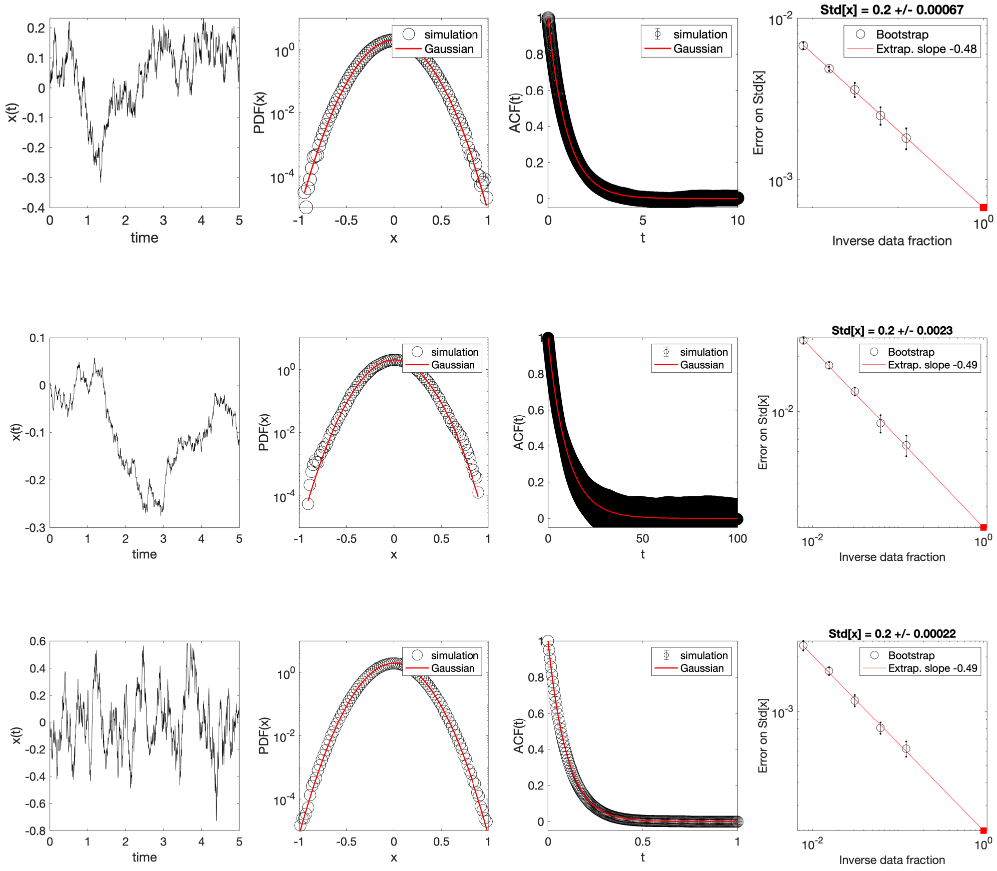

Figure 12: **Estimating the variance of Ornstein-Uhlenbeck (OU) process using bootstrap.** Each row corresponds to an instance of OU process with different correlation time (top: $\gamma = 1$; middle: $\gamma = 0.1$; bottom: $\gamma = 10$). For each process, the first column shows a sample realization of $x(t)$ for $t \in [0, 5]$. The second row shows the marginal distribution of x from the simulation (black), overlaid by the theoretical Gaussian expectation with analytically computed variance (red; here $\sigma_x = 0.2$). The third column shows the autocorrelation function estimated from the simulation (black), overlaid by the theoretical exponentially decaying function with analytically computed timescale (red). The fourth column shows the bootstrap estimate of the error bar on $sigma_x$ (the empirical mean and bootstrap-estimated error bar shown in the title of the plot). Contiguous chunks of timeseries for different fractions of the data (inverse fraction on horizontal axes) are chosen from the timeseries, and the error on std is computed across these chunks at each data fraction, and extrapolated linearly on log-log plot to full data set (fraction = 1, red square). Slope of the extrapolating line is in the legend.

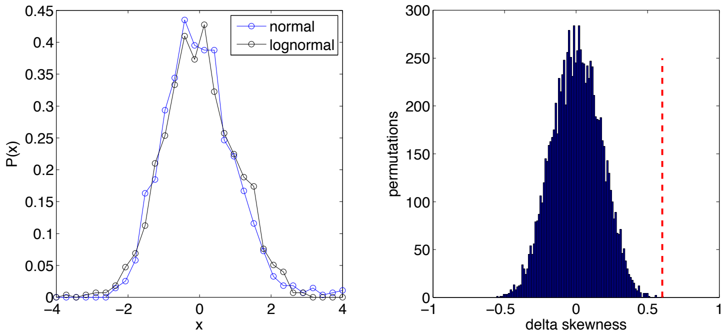

Figure 13: **Left.** A normal (blue) and weakly lognormal (black) PDF estimate, constructed from $N = 1000$ samples IID drawn from each (see main text). Both distributions have zero mean, unit variance, but different skewness. **Right.** The null distribution for the test statistic (blue), constructed as explained in the text, by computing 10^4 times the difference in skewness on resampled data. This is compared with the true difference in skewness (dashed red line), which lies far outside the null distribution, with a significance $p < 10^{-4}$.

A statistic that is known to be strongly biased is the so-called "naive" or maximum likelihood estimator for the entropy of the (discrete) distribution p, i.e., $S[p] = -\sum_i p_i \log_2 p_i$. Here, the index i runs through all discrete bins of the distribution, and p_i is an empirical estimate of the probability, i.e., $p_i = N_i/N$, where N_i is the raw counts of occurrence of state i, and $N = \sum_i N_i$ is the total number of counts. Figure 14 shows how bad the problem really is for undersampled distributions: here, the true distribution is a uniform distribution over $m = 500$ bins, which should have the entropy of $S = \log_2 500 \approx 8.97$ bits. With only a $N = 100$ samples, the naive estimator gives an entropy of ≈ 6.3 bits with a small error bar—so the dominant error of the estimator is the bias (the estimation is wrong on average), not the variance. Only in the regime where $N \gg m$ is the bias of the naive estimator small. One way to "debias" the estimate is to empirically assess how the estimate systematically changes with the sample size, N, and then extrapolate away this dependence in the infinite sample limit, as shown in Fig 14 at right. This method is most powerful when we have a theoretical expectation for the scaling of the bias with N, as we do for the entropy, and the extrapolation is used only to obtain the scaling parameters.

Homework 3. Try three different bootstrapping methods for estimating the error bar on a statistic for a synthetic data that is IID by construction. For your data set, draw a hundred samples from a normal distribution with zero mean and unit variance. On this dataset, estimate the standard deviation (ground truth value of which is unity), and its error bar. In the first method for error bar estimation, do bootstrapping by selecting random halves of the data (e.g., 1000 random splits). In the second method, do bootstrapping by resampling (e.g., 1000 times) with replacement from the full dataset. In the third method, apply the Jackknife estimate. Compare the obtained error bar estimates.

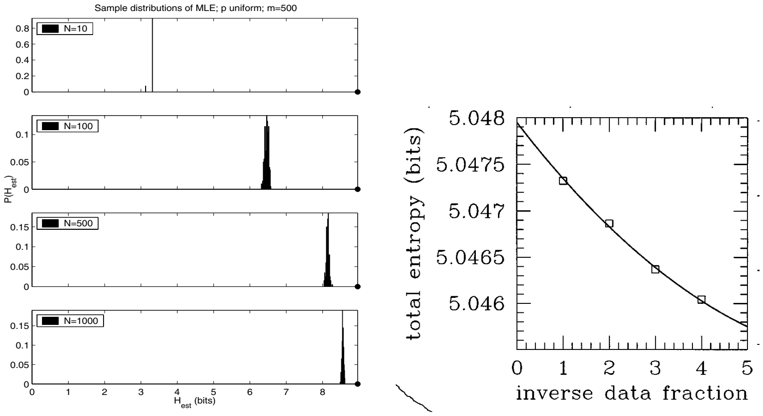

Figure 14: **Left.** Distributions of the "naive" estimator for entropy (see text), as a function of the sample size (N, in the legends), for an uniform distribution over $m = 500$ bins; the correct value of the entropy is almost 9 bits (black dot in the rightmost corner of the plot). The bias of the estimator, here the difference between the mean of the distribution and the black dot, is much larger than the error bar of the estimator (width of the distribution). Reproduced from Paninski et al, *N*eural Comput **15** (2003). **Right.** Since the bias of the naive estimator for entropy is known to scale as $1/N$ to leading order, i.e., $S(N) = S_{\text{true}} + S_1/N + S_2/N^2 + \dots$, we can undo it by subsampling: entropy is evaluated over many subsamples of the data of different size (here, "inverse data fraction" on x-axis, with, e.g., 5 representing 20% of total samples, etc.). These estimates at different fractions of the data are used to fit the extrapolation curve for $S(N)$, here quadratic, and find the true estimate, S_{true} (the value of the smooth curve at "inverse data fraction" 0, which corresponds to the $N \to \infty$ limit. Reproduced from Strong et al, *Phys Rev Lett* **80** (1998), where they estimated the entropy of neural spike trains from actual data.

Homework 4. Use your detection of spikes on background-subtracted `trace1` data from previous homework to define a new vector z, of the same length as `trace1`, which only contains zeros or ones: ones should indicate downward threshold voltage crossings in `trace1`, with threshold you selected for spike detection, and other time bins should be zero; in other words, z should have a one whenever a spike occurred in the original voltage trace; $\sum_t z_t$ should be the total number of detected spikes in `trace1`. Note: z should only indicate threshold *crossings*, i.e., cases where voltage was above the threshold in the previous timebin and is below the threshold in the next timebin. Let's now do the following: (i) take many contiguous chunks of z of length $500, 750, 1000, 2500, 5000, \dots, 100000$ time bins, and on each chunk estimate the "spike rate", as the number of spikes in that chunk divided by the length of that chunk; (ii) estimate the error in the spike rate by computing the SD over chunks of the same length, as in the classic bootstrap; (iii) plot this error as a function of the chunk length, on the semilogx plot. Does the error bar decrease with the chunk length as you would expect? (iv) To get a better intuition, compare this plot of spike rate error bar vs chunk length for the vector z whose elements you first randomly permuted to break all temporal correlations. Check that in this permuted control, the spike rate error bar decrease as (chunk length)$^{-1/2}$. If you see a deviation from that behavior of the spike rate error bar for the original z, what could be the source of such deviations?

6 Comparing distributions

6.1 Kolmogorov-Smirnov test

Now that we know how to create histograms and estimate PDFs with error bars from data, we can look for the ways of comparing whether two distributions are the same, or, alternatively, of testing whether a data-derived distribution matches a theoretical model. For continuous data a standard test for this is the Kolmogorov-Smirnov (KS) test, which has three clear benefits: **(i)** it is non-parametric, i.e., it makes no modeling assumptions about the data; **(ii)** it doesn't require binning the data first to create histograms; and **(iii)** it is reparametrization-invariant (which means that test results are independent if carried out on the original data, x, or on some nonlinear monotonic transformation of the data, e.g., $f(x)$). Two potential drawbacks are that the test is only applicable to 1D distributions, and that its sensitivity is greatest for distribution discrepancies near the median, rather than in the tails (although generalizations of the test exist that correct for that).

Given two sets of samples of size N, the test first creates the two corresponding cumulative distribution functions, $C_1(x)$ and $C_2(x)$, and then defines as a test statistic the distance D, where $D = \max_x |C_1(x) - C_2(x)|$. The key property of this choice is that the probability of D being equal to or larger than some observed value given that the distributions don't differ (the null hypothesis) is a universal function only of D and the $\sqrt{N}$, the number of samples, which is known analytically (see *Numerical Recipes in C* if you want to know the expression) and is built in into most numerical packages.

An interesting application of the KS test is to ask whether, for instance, the background Gaussian-like noise distribution, shown in Fig 11 middle, could be *stationary*, i.e., statistically indistinguishable between the first and second part of the experiment. Indeed, up to now we

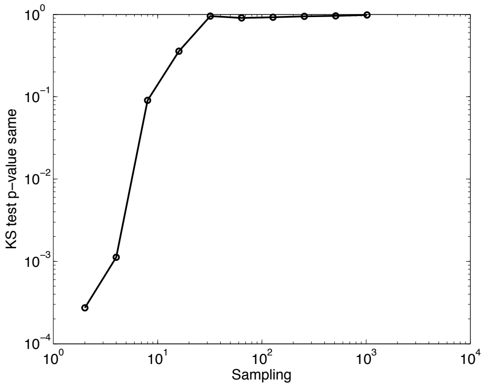

Figure 15: Raw `trace1` data has been background subtracted and spikes have been removed as in Fig 11, middle. The time series is then split into two halves, each consisting of roughly 1M samples. From this, we subsample every $k-th$ datapoint (denoted on the x-axis), and use the KS test to ask whether we can reject the null hypothesis that the (subsampled) data in the first and second halves of the trace comes from the same distribution; the resulting p-value is shown on the y-axis.

have been assuming the stationarity of our time series data, that is, the stability of all the statistics at different times, without explicitly checking for it. Figure 15 suggests that the first and the second half of the trace are nominally different (the null hypothesis that they are the same could be rejected with low p-value), but only if we don't subsample the data. Yet we know that there must be strong correlations in the data which is thus not IID, and have roughly estimated before that the effective number of independent samples could be $\sim$ 100 times lower than the nominal value. Indeed, if we subsample correspondingly (only take every 32th point or even more sparsely), the gaussian distributions in the first and second half of the trace are not significantly different. Note that we could use the KS test in a straightforward fashion not only to compare the first and second parts of the trace, but also to compare each to the Gaussian model rigorously, but same caveats about non-IID samples apply.

6.2 χ^2 test

Another standard way of comparing the distributions is useful when we are working with raw histograms, that is, with binned data. Let counts X_j constitute integer counts in a raw histogram that is to be compared to an expectation, Y_j, derived from a model (Ys need not be integers). Since we expect (from the model) the error in each bin to be $\sqrt{Y_j}$, we can form a χ^2 statistic for the deviation between the observed counts and the model:

$$\chi^2 = \sum_j \frac{(X_j - Y_j)^2}{Y_j}. \tag{19}$$

The significance of this deviation can then be looked up from the standard χ^2 distribution to assess whether the null hypothesis that the data X and the model distribution Y are the same can be rejected. Nominally, χ^2-distribution gives the probability that the sum of squares of independent gaussian variables is greater than some threshold. In Eq. (19) individual terms are clearly not Gaussian (since X are discrete), but if the counts are sufficiently large and we are summing over sufficient number of bins, the use of χ^2 is appropriate. The number of degrees-of-freedom (DoF) for the comparison is typically equal to the number of the histogram bins minus

1, if the model distribution for Y is forced by normalization to have the same total number of events, $N = \sum_j X_j$, observed in the data.

The same test can be also used to compare two data-derived histograms, as we did before when we looked at the first and second half of the time series. The only change is that now both X and Y are integers, and the denominator of Eq (19) changes into $X_j + Y_j$, since the uncertainty is in both histograms.

Lastly, χ^2 is often used as an objective function to fit theoretical distribution models to data. As an example, we fit a mixture of three Gaussians to the distribution of voltage values for `trace1` in Fig 16. This amounts to fitting 8 parameters in total (3 means and variances, and 2 prefactors) that determine the model, Y, variable in the χ^2 given by Eq (19), which is minimized using a generic nonlinear fitting routine, e.g., Levenberg-Marquardt; there is no guarantee of finding the global χ^2 minimum. There is no theoretical reason for why our data distribution should be composed of three gaussians (and, indeed, the deviations of the model from the data are quite significant), but a mixture of Gaussians is often a flexible and versatile way to phenomenologically model complex, multi-peaked distributions. It is also interesting to contrast fitting such models to data using χ^2 minimization on binned histograms vs. doing a maximum likelihood fit of the mixture of Gaussians (which is a probabilistic model) directly to unbinned data $\mathcal{D}$.

One should understand well the nature of the statistical test being done when computing, e.g., χ values in Eq. (19). If χ values across bins are squared and summed together, then *one* test is performed to assess whether a sample as a whole comes from a given distribution. A very different situation arises when deviations in individual bins are tested for significance: here, we have multiple hypothesis testing (across all bins), and our significance thresholds should be adjusted according to how many hypotheses we are testing (e.g., using Bonferroni correction). A typical case is in high energy experiments (HEP) where the Standard Model theory predicts how many events of a particular type should be detected in the particle colliders as a function of some kinematic variable, and this is compared bin-by-bin to actual data. A good example was the excitement in 2016 (which turned out to be a statistical fluctuation) of a localized deviation from Standard Model, which could be a putative particle, in *arxiv.org:1506.00962*. A localized deviation was highly significant, at 3σ; but this was correctly decreased for the “look-elsewhere effect” (physicist’s slang for multiple hypothesis testing), and fell considerably short of the 5σ standard for proclaiming new discovery (to be compared to the standard life science thresholds of $p = 0.05$ which lead to many significance reports highlighted recently). This problem is very acute when the number of effective hypotheses is immense. In fMRI scans, for instance, the brain is divided into local volume elements or “voxels” where the BOLD—blood oxygenation signal, a proxy for local energy utilization by neural processing—is measured. The number of voxels in new scanners is $> 10^5$, and with standard significance thresholds not properly corrected for this number one is likely to observe “significant” activity even when there can be none, as demonstrated by the famous fMRI scan of a dead atlantic salmon (Bennett et al, `http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf`).

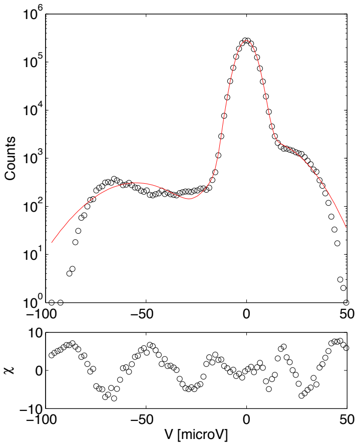

Figure 16: The raw histogram of background-subtracted voltage in `trace1` and the best-fit mixture of three Gaussians (red line), using χ^2 minimization. The plot below shows the per-bin contributions to the total χ^2. While χ^2 can be used as a good metric for fitting models to data, its precise value cannot be used directly for statistical significance testing since the data is not IID.

Homework 5. Estimate the distribution of log luminance levels in the Ruderman natural image dataset. For each image, transform it by taking the log of the pixel luminance values, and subtracting the mean value of log luminance across each image. Show the estimate of the error bars on this distribution of mean-subtracted log-luminances, y, by thinking about what "independent samples" are in this dataset (you do not need to do a full bootstrap analysis). Show that the left- and right- tails of this distribution are approximately exponential, i.e., $P_+(y) \propto \exp(-\mu_+|y|)$ and $P_-(y) \propto \exp(-\mu_-|y|)$, where $P_\pm$ stand for the positive (or, respectively, the negative) tails of the log-luminance distribution. Do this by finding the two best-fitting constants, μ_+ and μ_- and plotting the tails on top of the normalized distribution for y. A good choice to define the positive and negative tails is to use the threshold of $|y| > 1$. How you do the fitting (linear regression in log-luminance space, χ^2 fitting of raw histograms) is up to you, but explain clearly what you did.

7 Measuring correlations

Here we learn to quantify different types of (pairwise) correlations. By pairwise correlations we mean statistical relationships, in particular, deviation from statistical independence, that are evident at the level of pairs of components of data vectors, $\mathbf{x}$, or pairs of values in a time series. Pairwise independence would imply that, for a given pair of components x_μ and x_ν:

$$P(x_\mu, x_\nu) = P_\mu(x_\mu)P_\nu(x_\nu), \tag{20}$$

i.e., that the joint distribution factorizes into marginal distributions. In discretely-sampled time series, you can think of a true generating distribution over all possible sequences that could have been observed, $P(x_1, x_2, \ldots, x_N)$; samples from this distribution can be thought of as repeats of an experiment that generates the time series many times. Pairwise independence in this context