Data Science and Scientific Computation Core Course

Understanding and visualizing data

February 28, 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!*

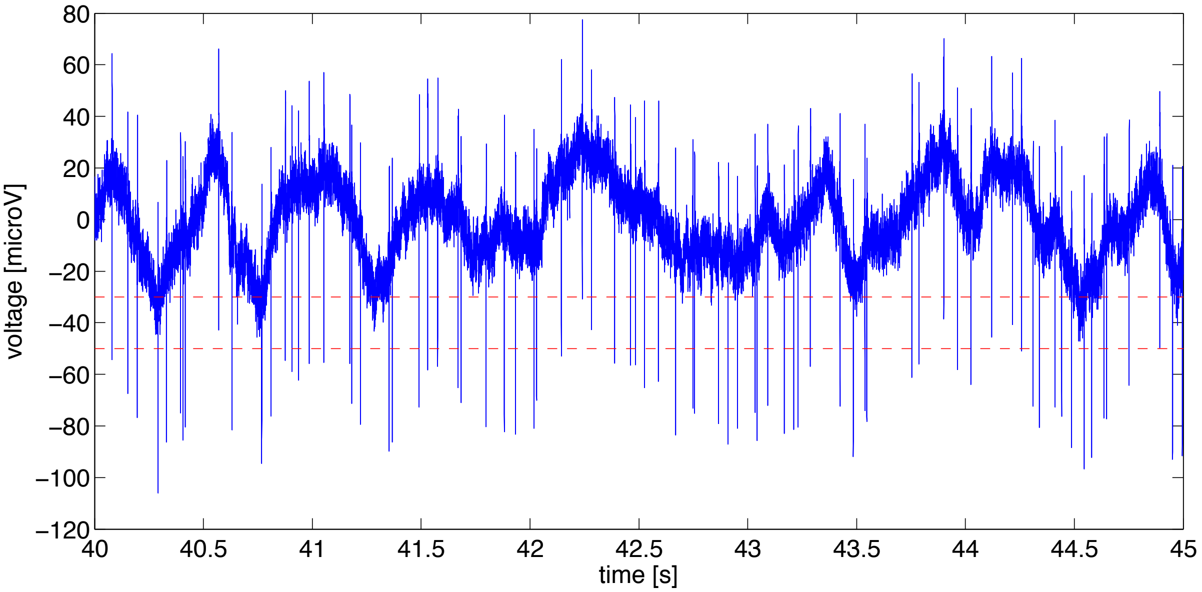

Figure 1: A 5 s example trace of the retinal recording on the multielectrode array (MEA).

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.

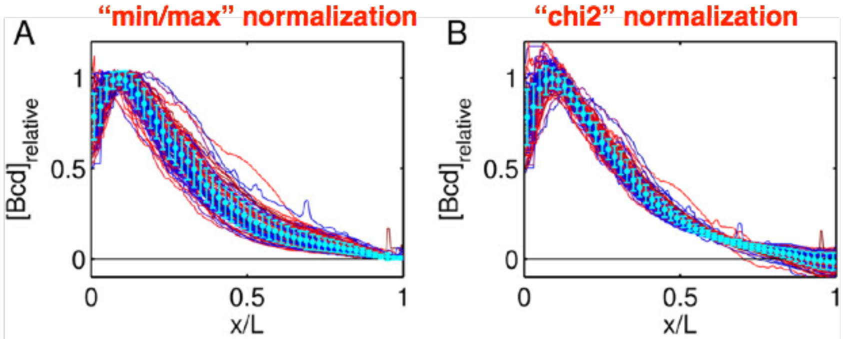

Figure 2: Comparison between normalizing each Bicoid gradient, [Bcd](x), by its minimum and maximum (to align all gradients between 0 and 1) in (A), vs aligning the gradients by adjusting an additive offset and multiplicative scale per gradient so that they best align to each other in a χ^2-sense in (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.** $P_{\mathrm{emp}}(\mathbf{x}) = \frac{1}{N}\sum_{t=1}^{N}\delta(\mathbf{x} - \mathbf{x}_t)$, 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:

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

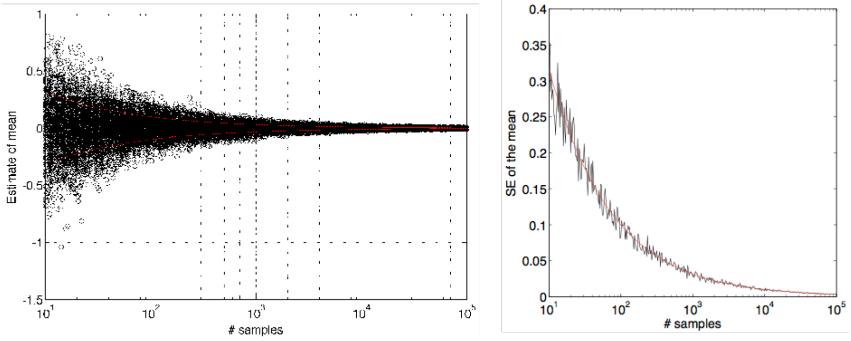

Figure 3: Individual data points are drawn IID from the standard Gaussian distribution with mean $m = 0$ and variance $\sigma^2 = 1$. **Left.** Different number of samples N are used to compute estimates of the mean according to Eq (1); each estimate is plotted as black circle. Since the estimator is 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

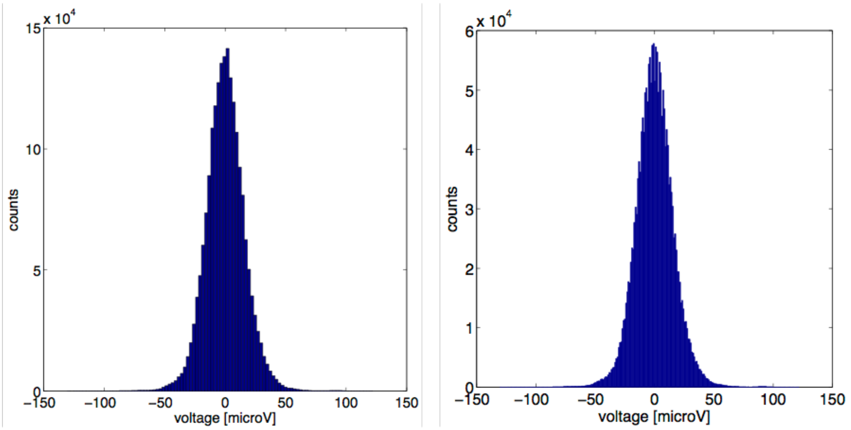

Figure 4: Raw histograms of data in `trace1`, with 100 bins **(left)** and 1000 bins **(right)**, uniformly spaced between the maximal and minimal observed values in the full trace. Interestingly, the histogram at right reveals curious "throughs" between the bins: on closer inspection it turns out that 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

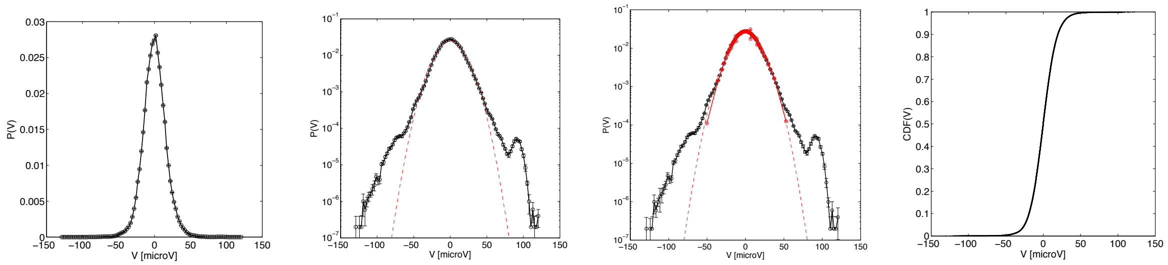

Figure 5: **Left.** Histogram of `trace1` data constructed over 100 equidistant bins and normalized into an estimate of the PDF, as in Eq (7). Note that PDFs have units, in this case $1/\mu V$, since they have to integrate to 1. **Middle left.** Logarithmic plot of the same PDF and comparison to the 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 $\int dxP(x) = 1$), 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