# General Themes

- Basic mathematical approaches to data and models are successful because of fundamental principles, not because of accidents. What are those principles?
- Use appropriate mathematical objects to represent data and models: Measurements in the lab are discrete and finite, but mathematical concepts and models typically involve continua and infinities. So the mathematical constructs should allow for a smooth transition between large but finite and infinite, and a smooth transition between sampled and continuous. Choosing the appropriate mathematical object amounts to a statement about what is important. For example, representing the transmembrane voltage of a neuron as a point process is a statement that the spikes are all-or-none, and that the details of spike shape do not matter.
- Coordinates are a necessary evil. We need them to communicate, but they lead to arbitrariness in the way that data and models are presented. We will try to develop the theory, as much as possible, without coordinates.
    - Working without coordinates, although difficult at first, will have some real advantages – for example, by showing the existence of the determinant without the use of coordinates, its main properties come "for free."
    - When coordinate are needed – for example, to work with data, we try to find natural coordinates that are based on the intrinsic features and symmetries of the problem. These will typically lead to more incisive analyses, and descriptions of the data that suggest generalization, not the specific features of the experiment, measuring devices, etc.
- We will make counter-factual assumptions (e.g., "linearity"). We make them because they are extremely useful starting points, and because there are natural ways to relax them. There seems to be a meta- principle: principled methods based on broad assumptions that are known to be only approximately correct work much better than non-principled methods based on ad hoc assumptions. Also, symmetries, even if only approximate, are often very helpful.

# What are the challenges in this material?

- Objects play multiple roles, and simple examples may not reveal what is generic
    - *numbers and transformations:* on a line, numbers act both as positions and transformations. A number $x$ can also be interpreted as the transformation $A_x$, "add $x$". Adding two numbers, $x$ and $y$, can be thought of at least three ways: (1) combining two scalars to form a third; (2) applying a transformation $A_x$ to the number line, and seeing how it transforms $y$, i.e., $A_x(y) = y + x$, and (3) combining two transformations: first apply $A_x$ and then apply $A_y$, and noting that the net result of this sequence of transformations is the transformation $A_{x+y}$. That is, for any number $z$,

$A_{x+y}(z) = A_y(A_x(z))$. This seems like hair-splitting until one considers the transformations of the surface of the sphere – here, there's clearly a distinction between points on a sphere, and rotations of the sphere. To specify a rotation, you have to specify a pair of opposite points (a diameter of the sphere), and a rotation angle; this is quite distinct from specifying two a single point on a sphere. Another fundamental difference is that transformations on a line are commutative (order doesn't matter); while rotations of a sphere are not commutative (order does matter).

- *the two sides of a dot-product:* Consider a signal and a filter. Both are functions of time: the signal is $s(t)$, the filter is $f(t)$, and the action of the filter on the signal produces a value at each instant in time. At time 0, we can write this as $s \bullet f = \int_0^\infty f(\tau)s(-\tau)d\tau$. (There's a sign convention here – positive arguments for *f* refer to the impact of *s* at previous values of time). It appears that *s* and *f* play the same role, but in fact they are different kinds of objects. One clue that this is the case is that they have different units. For example, if you change the unit of measure from volts to microvolts, you *increase* the numbers that describe *s* by a factor of $10^6$. But for *f*, we describe the gain as the size of the output for a unit input. So the gain decreases by a factor of $10^6$, as the gain goes from units of "per volt" to "per microvolt". Similar considerations apply to the distinction between an image, and a spatial activity pattern – but now the coordinates are the pixels. Or to lights, and chromatic mechanisms – the coordinates correspond to the three kinds of photoreceptors.
- Another example of this is in statistics: *the distinction between a quantity to be estimated (e.g., "the mean") and a procedure for estimating it (e.g., "take the average")*. In some situations – for example, a Gaussian prior – the best way to estimate a statistic is to apply its definition directly (the "plug-in" estimator). But more generally, this is not the case: for example, if the samples are drawn from a heavy-tailed distribution, then more accurate estimates of the mean can be obtained by discarding outliers before averaging.

- As the above examples show, one needs to be familiar with examples so that one can use them to help understand the abstractions

## Coordinate systems

The typical objects we measure are intrinsically multivariate. However, the "intuitively obvious" coordinate systems we use to describe them may not be the most natural ones. It might seem natural to describe a time series *s* in terms of its samples at a series of time steps (as above). But this has some undesirable consequences.

*Time series:* First, if we change the discretization (which is something that is arbitrarily determined by our instrumentation) we drastically change the representation of *s* – we change its dimensionality.

Second, we *always* record *s* through some kind of filter, again determined by our instrumentation. This filter –considered as operating in discrete time – is a linear transformation that mixes the coordinates of *s*. In continuous time, $s^{observed}(t) = \int_0^\infty f(\tau) s^{underlying}(t-\tau) d\tau$, where *f* describes the mixing process due to the filter. Discretizing time in steps of $\Delta\tau$ so that $s_j = s(j\Delta\tau)$, and similarly for *f*, this is equivalent to $s_j^{observed} = \Delta\tau \sum_{k=0}^{\infty} f_k s_{j-k}^{underlying}$. So the instrumentation filter mixes the coordinates of *s*. There's no particular reason that the ones that the instrumentation hands us is the best one to use. Wouldn't it be nice to use a coordinate system in which filters don't mix the coordinates?

Note also that we can rewrite the mixing as $s_j^{observed} = \Delta\tau \sum_{m=-\infty}^{j} f_{j-m} s_m^{underlying} = \Delta\tau \sum_{m=-\infty}^{j} F_{j,m} s_m^{underlying}$ (with $m = j-k$ and $F_{j,m} = f_{j-m}$). So the filter *f* corresponds to a matrix *F* whose elements are constant on diagonals. Now it's clearer that filters and signals are different objects: signals are vectors, while filters are a special kind of linear transformation on the vectors.

*Imaging:* All of this applies to imaging as well. The underlying image is discretized into pixels, and this process is, at the very least, affected by linear filters (e.g., spatial blurring, due to defocus, image processing algorithms to sharpen). Two dimensions of space, instead of one (unidirectional) dimension of time.

*Multichannel neural data:* While it is intuitive to use separate coordinates for each channel, other coordinates may be more useful: underlying signals mix via volume conduction at the level of local field potentials. Neural spike trains may not be cleanly separated. And even if they are, finding a lower-dimension set of coordinates that can account for all the observed channels, via linear combination, greatly simplifies things. So dimension reduction is a major theme.

# Why is the field of statistics still an active one?

It's obvious that one needs statistics: to describe experimental data in a compact way, to compare datasets, to ask whether data are consistent with a model. But why is this a hard problem (or at least, why are people still making advances)? We will be focusing on multivariate data (e.g., time series, images), but the reason that "statistics" is nontrivial emerges even when we look at univariate data. Of course multivariate data does not make things better.

## *Toy example: estimating the mean*

To set it up: we suppose that there is a collection of possible outcomes $x$ (an "ensemble" $\Omega$, or a set of possible measurements associated with their probabilities). We would like to estimate the true mean $\mu$ of the values $x$ in $\Omega$, which we write as $\mu = E(x)$.

$E(x)$ means the expected value of $x$, which we may also write as $\langle x \rangle_\Omega$, to emphasize the dependence on the ensemble $\Omega$. For discrete ensembles, $\langle x \rangle_\Omega = \sum_{x \in \Omega} xp(x)$, where $p(x)$ is the probability of drawing $x$ from $\Omega$. For continuous ensembles, $\langle x \rangle_\Omega = \int_\Omega xp(x)dx$, where $p(x)\Delta x$ is the probability of drawing a value between $x$ and $x + \Delta x$ from $\Omega$.

Clearly if we could sample all of $\Omega$, the problem would be simple: we'd just apply one of the above formulae to our exhaustive sample, to determine its average. But we only have a finite sample, not all of $\Omega$.

We now draw $N$ values from $\Omega$, say $x_1$, …, $x_N$. Our problem is to craft an "estimator" function, say, $\hat{\mu} = \hat{\mu}(x_1,...,x_N)$, to provide an estimate of $\mu$. The obvious choice is the sample mean. This is also known as the "plug-in" estimator, $\hat{\mu}_{plugin} = \frac{1}{N}\sum_{i=1}^{N} x_i$, since it is "plugging in" the measured values into the formula for the mean. But it is not the only choice.

### *Silly choices of estimators*

There are some clearly silly choices:
(a) a fixed, *a priori* guess, independent of the data
(b) throw out the even-numbered measurements, and take the sample mean of the rest,
(c) take the sample mean, and add a fixed number (say, 7),
(d) take the sample mean, and add a number that depends on $N$, say, $1/N$,
(e) choose one value from the data,
(f) take the sample mean, add a random number of mean zero and variance 1.

Why do we know that these choices are silly? Conversely, what properties do we want an estimator to have?

## *Desirable properties of an estimator*

A good estimator should be unbiased (i.e., it should not systematically over- or under-estimate), and should converge to the correct value as one has more and more data. That is, as $N$ grows, we would like the expected value of the estimator. $E(\hat{\mu})$ to converge to the correct answer, $\mu$. This has two components: "bias" and "consistency."

An unbiased estimator gives the correct value, on average, when applied to a finite dataset. That is, the bias is the difference between the expected value of the estimator for a sample size $N$, and the true value, $E(\hat{\mu}) - \mu$, where the expectation is taken over all datasets of size $N$.

A consistent estimator is one that eventually converges to the correct answer, almost all of the time. A basic way to formalize this is $\lim_{N\to\infty} E\left(\left(E(\hat{\mu}) - \mu\right)^2\right) = 0$ : i.e., the mean-squared scatter of the estimator around the correct value eventually goes to 0 as more data are accumulated. Occasionally, one is interested in alternative definitions of consistency – for example, bounding the largest possible error, rather than bounding the expected mean-squared error. Or bounding the largest possible error that can occur with probability > 0.

Both the lack-of-bias and the consistency conditions can be phrased in stricter forms, in which the speed of approach to 0 of the above limit is specified, typically $K / N$ for some constant $K$. More *efficient* estimators converge more rapidly, i.e., have a smaller *K*. One can't typically expect to do better than a bias that decreases as $K / N$, for some *K*.

How can one evaluate (theoretically) the performance of a statistic? One does a thought experiment in which one draws multiple sets of $N$ values from $\Omega$, calculates the estimator of interest, and compares it with the true value. At a second level, one could postulate a family of ensembles, say $\Omega(\beta)$, one for each value of an unknown parameter $\beta$ (or, more typically, a set of parameters $\beta_k$), and then see how sensitive the above analysis is to knowing $\beta$. That is, how strongly does the merit of the statistic depend on knowing, precisely, the form of the distribution? Sometimes one can do this analytically.

One can imagine a situation in which an estimator performs wonderfully for a particular family of ensembles, but performs terribly for ensembles that are not in this family – i.e., an estimator that is highly sensitive to model error.

## *Back to the toy example*

For Gaussian ensembles, the plug-in estimator for the mean is unbiased, and consistent, and it is the most efficient estimator ( $K$ is as small as possible.) But the Gaussian ensemble is the *only* ensemble for which this is true. For ensembles that are not Gaussian, there is always a more efficient estimator of the mean, i.e., one for which the convergence to the true value is faster or more certain. The plug-in estimator remains useful because (a) it is unbiased, (b) it is consistent, (c) it is simple, (d) its properties are simple to calculate, (e) often the improvements conferred by other estimators are fragile, i.e., they are highly sensitive to the assumed shape of the distribution $\Omega$.

Here is an important, practical example of a situation in which the plug-in estimator for the mean can be improved on:

Say the ensemble $\Omega$ is known to have the following structure: most of the values come from a Gaussian distribution (mean and variance unknown), but a small fraction, say $\varepsilon$, of the values are corrupted by a large measurement error. That is, $\varepsilon$ of the values consist of samples from the underlying Gaussian, to which a large quantity $M$ is either added or subtracted – or even, that the data are replaced by a large quantity $\pm M$. (Think of this as modeling a typographical error, or a rare but very large artifact.) Now, construct an estimator in two steps. First, throw out some fraction $f > \varepsilon$ of the extreme values. Then, take the mean of the remaining values. This is known as a "trimmed mean" estimator.

Since the first step gets rid of much of the variability, the resulting estimator converges faster than the plug-in estimator. If, say, we choose $f = 2\varepsilon$, then nearly all of the time, all outliers will have been eliminated. The trimmed mean is an example of a "robust" estimator (one that is relatively insensitive to outliers). The median is the limiting case of the trimmed mean (discard all but one of the measurements as extreme).

In this example, $\Omega$ has heavy tails (platykurtotic). You can think of other examples in which the heavy tail extends in only one direction ($\Omega$ is skewed), or has known shape, or even, in which $\Omega$ has light tails (leptokurtotic). These lead to other estimators that all do better than the plug-in estimator, each in its own domain. The trimmed mean estimator is worse than the plug-in estimator for a Gaussian, but not much worse. So it is often a good choice. And one often does it even without thinking ("let's throw out that experiment, it's an outlier.")

Note that the above example is *extremely simple* in that we are only considering univariate quantities. Additionally, we are attempting to estimate a quantity – the mean – for which the plug-in estimator is unbiased and consistent. Neither of these are typically the case – for most statistics, even for Gaussian ensembles, the plug-in estimator is biased (for example, the sample variance). The plug-in estimator for the variance is $\frac{1}{N}\sum\left(x_i - \hat{\mu}_{plugin}\right)^2$, but this is biased; an unbiased estimator is $\frac{1}{N-1}\sum\left(x_i - \hat{\mu}_{plugin}\right)^2$.

In general, there is usually a tradeoff between bias and consistency (i.e., you can optimize the estimator for one, or for the other, but not for both.)

*In sum, one reason that statistics is not trivial is that for estimators, one does not have "one size fits all" – their merits depend on what one assumes about the ensemble, what is actually true about the ensemble, and how sensitive they are to errors about ones' assumptions.*

Another factor is that computational burden is, in fact, relevant. Prior to computers, usable statistics nearly always were those that one could determine confidence limits, etc. analytically (Gaussian, Poisson). Now more computationally-intensive approaches are practical – e.g., resampling approaches such as the bootstrap and the jackknife. But computational practicality is always a consideration, and resampling approaches are not foolproof.

As a rule of thumb, for estimators that are simple (such as a plug-in estimator), it is easier to determine their sensitivity to the choice of the ensemble, and this sensitivity is usually less severe than for a more complex estimator.

## *The Bayesian framework*

This is a brief introduction to a general approach to create principled estimators, with the algebra worked out in more detail than in class.

The idea is that we postulate that there are a family $\Omega(\beta)$ of plausible distributions, each with an assumed *a priori* probability $p(\beta)$, and we want to use our observations $X$ to modify these probabilities to some $p(\beta \mid X)$. We can then use the value $\beta$ that maximizes $p(\beta \mid X)$ to determine the most likely distribution $\Omega(\beta)$, and from this distribution $\Omega(\beta_{ML})$, determine the statistic of interest, say the mean $\mu = \mu[\Omega(\beta_{ML})]$. This is the "maximum likelihood estimator." Or, we could determine a weighted estimate, $\int \mu[\Omega(\beta)]p(\beta \mid X)d\beta$, which weights the true mean of each ensemble $\Omega(\beta)$ by the *a posteriori* probability of that ensemble. (We can even determine confidence limits in this fashion.)

The basic tool is Bayes' rule:

$$p(\beta \mid X) = \frac{p(X \mid \beta)p(\beta)}{p(X)}.$$

The most important term is $p(X \mid \beta)$, the probability that the ensemble $\Omega(\beta)$ would yield the observations $X$. We can compute this if we have a clear hypothesis about $\Omega(\beta)$. The denominator is the probability of the observations $X$. We don't know this, but it (by definition) is independent of $\beta$. The final term is the *a priori* probability $p(\beta)$. Usually, we set this to a constant $k$ independent of $\beta$ (an "uninformative prior"), indicating that we have no prior reason to favor one value of $\beta$ over another – but we could also restrict it to some range, or have it not flat. Note that there's a formal problem if we allow $\beta$ to range over $(-\infty, \infty)$, as we can't normalize it -- $\int_{-\infty}^{\infty} k d\beta$ is undefined.

But since we only care about the relative sizes of $p(\beta \mid X)$, we can ignore this – and be content with computing the relative probabilities (likelihoods) of different values of $\beta$:

$$L(\beta \mid X) = p(X \mid \beta).$$

Since $L(\beta \mid X)$ and $p(\beta \mid X)$ are proportional, the value of $\beta$ that maximizes $L$ yields the maximum-likelihood estimator. In principle, the full Bayesian estimator $\int \mu[\Omega(\beta)]p(\beta \mid X)d\beta$ can be also be computed from the likelihood, provided we can normalize $L$: $\int \mu[\Omega(\beta)]p(\beta \mid X)d\beta = \frac{\int \mu[\Omega(\beta)]L(\beta \mid X)d\beta}{\int L(\beta \mid X)d\beta}$.

Now let's implement this for a flat prior, and a set of ensembles $\Omega(\beta)$ that are Gaussians with mean $\beta$ and variance $\sigma^2$. The probability that a value drawn from $\Omega(\beta)$ is between *x* and $x+\Delta x$ given by $p(x\,|\,\beta)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\beta)^2/2\sigma^2}\Delta x$. So the probability that a sequence *X* with of *n* values, with first value between $x_1$ and $x_1+\Delta x_1$, second value between $x_2$ and $x_2+\Delta x_2$, etc., is drawn from $\Omega(\beta)$ is given by

$$p(X\,|\,\beta)=p(x_1\,|\,\beta)p(x_2\,|\,\beta)\bullet\ldots\bullet p(x_n\,|\,\beta)$$

$$=\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-(x_1-\beta)^2/2\sigma^2}\Delta x\right)\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-(x_1-\beta)^2/2\sigma^2}\Delta x\right)\bullet\ldots\bullet\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-(x_n-\beta)^2/2\sigma^2}\Delta x\right).$$

This is

$$p(X\,|\,\beta)=\left(\frac{\Delta x}{\sqrt{2\pi}\sigma}\right)^n\left(e^{-(x_1-\beta)^2/2\sigma^2-(x_2-\beta)^2/2\sigma^2-\ldots-(x_n-\beta)^2/2\sigma^2}\right)$$

$$=\left(\frac{\Delta x}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\beta)^2}$$

Finding the value of $\beta$ that maximizes this is equivalent to finding the value of $\beta$ that minimizes the sum in the exponent, $S=\sum_{k=1}^{n}(x_k-\beta)^2$. This is a quadratic function of $\beta$,

$S=n\beta^2-2\beta\sum_{k=1}^{n}x_k+\sum_{k=1}^{n}x_k{}^2$. We find this minimum by setting $\frac{dS}{d\beta}=0$:

$\frac{dS}{d\beta}=2n\beta-2\sum_{k=1}^{n}x_k$. This is 0 when $\beta=\left(\sum_{k=1}^{n}x_k\right)/n$, i.e., the plug-in estimator.

*So: for an assumed Gaussian ensemble and a flat prior, the Bayesian maximum likelihood estimator for the mean is the plug-in estimator.*

Two comments. Note that the fact that the distribution had an exponential form was very helpful –the distribution $p(X\,|\,\beta)$ is generically a product, with one term for each sample (since the samples are independent), and, since the distribution had an exponential form, the exponent in this product was a sum of terms.

Second, note that the value of $\beta$ that maximized $p(X\,|\,\beta)$ did not depend on the assumed variance. This is, unfortunately, not generic – typically, the estimated value of one statistic depends on other model parameters – so-called "nuisance parameters" – a form of model-dependence.

### *A simple situation in which it's not obvious how to construct a good estimator*

Say that you know that the ensemble $\Omega$ contains only a finite number of different values, and you want to estimate the number of different values. What can you do with a finite sample? Obviously you know that the number of different values in $\Omega$ is no less than the number of different values in your sample. But it might be larger, and there's no reasonable way of estimating how much larger it can be unless you know something about the distribution of probabilities in $\Omega$ -- especially the distribution of the very low probabilities. With a model of this, you can develop a more sophisticated estimator, but only if you can believe the model.

This example indicates the basic difficulty that confronts estimating "information" from experimental data. If we tried to use a Bayesian approach, we'd get an estimator, but the estimator would be model-dependent.

### *Optional homework*

(answers at the end of this document)

Q1. Consider the above "silly" estimators for the mean. Which are unbiased? Which are consistent?

Q2. Construct a class of distributions for which the following estimator of the mean is unbiased, and also more efficient than the plug-in estimator: choose the highest and lowest values of the observations $x_i$, and take their mean. That is,

$\hat{\mu} = \frac{1}{2}\left(\min\{x_i\} + \max\{x_i\}\right)$. This is a kind of opposite strategy to the trimmed mean. Justify your claim (analytically or via simulation).

## A few principles

We want to be able to describe our data in a way that has meaning to others.

Some units are more natural than others (time: seconds rather than bins, space: cm (or deg) rather than pixels.

Images are obviously multivariate, but so are time series.

Often the origin is arbitrary (e.g., time).

With multidimensional datasets, it is even more of an issue: what direction should the axes point? Is there a grounded notion of "orthogonal" axes? Often at first glance there might seem to be: values at each pixel, or samples at each point in time. But when one looks at the biology or physics, one recognizes that each sample (say, at a specific pixel

or time point) reflects underlying "causes" at nearby locations or times as well (blur, filtering). So there is no strong reason that the obvious coordinates (the samples) are the natural coordinates. Blurring and filtering amount to linear transformations on the data; these are inevitable, so we might as well recognize this at the outset.

This leads to notions of:

- natural coordinates (e.g., Fourier analysis)
- data-driven coordinates (e.g., principal components analysis)
- coordinate-free descriptions (e.g., information theory)

And this (especially the notion of natural coordinates) leads us to focus on symmetries of the system. Translation in time is the paradigm.

Symmetries are often only approximate. But it is usually better to use a principled approach that is approximate, than an unprincipled one.

## Plans and options

Symmetry in the abstract (group theory)
Multivariate measurements in the abstract (vector spaces and their symmetries)
Implications of symmetry of the independent variable (how groups act on vector spaces)
Natural coordinates
Fourier analysis
    linear systems, filters
    noise and variability
Intrinsic symmetries of a vector space, and data-driven coordinates
    Principal components analysis
    Independent components analysis
    Various forms of targeted principal components analysis, e.g., "demixed"
Topological data analysis
Entropy, information, and data analysis
Graph-theoretic approaches
Point processes
Nonlinear dynamics, i.e., qualitative theory of ordinary differential equations

### *Answers to optional homework*

Q1. Silly estimators for the mean: which are unbiased, which are consistent?

(a) a fixed, *a priori* guess, independent of the data: biased, inconsistent
(b) throw out even-numbered measurements, and take the sample mean of the rest: unbiased, consistent (but inefficient)
(c) sample mean, plus a fixed number: biased, inconsistent
(d) sample mean plus $1/N$ : biased, consistent (but inefficient)

(e) choose one value from the data: unbiased, inconsistent (no improvement as the amount of data increases)
(f) sample mean plus a random number of mean zero and variance 1: unbiased, consistent, but inefficient (improves as amount of data increases, but flucutations about the true value persist even as the amount of data increases

Q2. Construct a class of distributions for which the following estimator of the mean is unbiased, and also more efficient than the plug-in estimator: choose the highest and lowest values of the observations $x_i$, and take their mean. That is,

$\hat{\mu} = \frac{1}{2}\left(\min\{x_i\} + \max\{x_i\}\right)$. This is a kind of opposite strategy to the trimmed mean. Justify your claim (analytically or via simulation).

Answer: Consider the class of "binary" distributions, i.e., those that contain only two values, $a_0$ and $a_1$ (with $a_0 < a_1$), each of which is drawn with a probability of 0.5. So the true mean is $\mu = \frac{a_0 + a_1}{2}$. Most of the time, after a large number of draws *N*, the sample minimum will be $a_0$ and the sample maximum will be $a_1$, so the estimator $\hat{\mu} = \frac{1}{2}\left(\min\{x_i\} + \max\{x_i\}\right)$ will be exact. To see how fast this estimator converges to the true mean, we note that the sample minimum and maximum will be accurate after *N* except for the fraction of trials in which the same value is drawn on each sample. This happens $2/2^N = 1/2^{N-1}$ of the time, since it requires that all of the $N-1$ draws beyond the first draw are matched to the first draw. So this estimator converges exponentially. Note that the mean-squared error of standard "plug-in" estimator decreases only like $1/N$.

Note also that if we applied this estimator to a distribution with tails, such as a Gaussian, it would be inconsistent – and in fact it would get worse and worse as we collected more data.