Some limiting distributions of random variables arising from high-dimensional processes

Alden Bradford

November 2021

the first component of a sphere

Let:

$$e_1 = (1, 0, \dots, 0)^T$$
$$V \sim \text{Unif}(S_{n-1})$$
$$X = e_1 \cdot V$$

In class we derived that

$$f_X(x) = \frac{1}{a}(1 - x^2)^{(n-3)/2}, \quad x \in [-1, 1].$$

This is explicit, but does not lend much intuition as $n \to \infty$.

a more intuitive statement

Let $V_n \sim \mathrm{Unif}(S_{n-1})$, $X_n = e_1 \cdot V_n$.
Then as $n \to \infty$, $\sqrt{n}X_n$ converges in distribution to $N(0,1)$, the standard normal distribution.

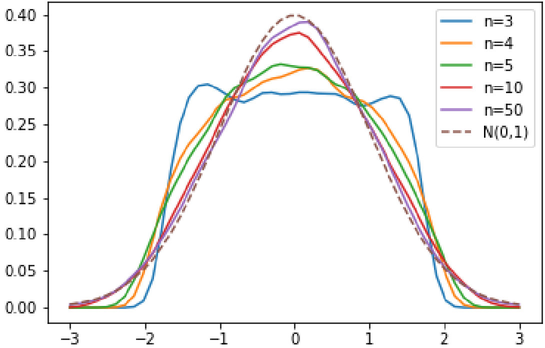


Figure: kernel density estimate based on 10,000 samples from $\sqrt{n}X_n$

Famous theorems

We can prove this quickly based on two theorems of probability theory. Write $\xrightarrow{p}$ for convergence in probability, $\xrightarrow{d}$ for convergence in distribution.

Theorem (The law of large numbers)

Given $X_1, \ldots, X_n$ i.i.d. with $|E[X_1]| < \infty$ then as $n \to \infty$,

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{p} E[X_1].$$

Theorem (Slutsky's theorem)

If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ then $X_n Y_n \xrightarrow{d} cX$.

Simple proof

Draw n samples N_i from $N(0,1)$. Then:

$$X_n \sim \frac{N_1}{\sqrt{N_1^2 + \cdots + N_n^2}}$$

$$\sqrt{n} X_n \sim N_1 \sqrt{\frac{n}{N_1^2 + \cdots + N_n^2}}$$

Notice $E[N_i^2] = 1$ so by the law of large numbers, $\frac{1}{n}(N_1^2 + \cdots + N_n^2) \xrightarrow{p} 1$. Then $\sqrt{\frac{n}{N_1^2+\cdots+N_n^2}} \xrightarrow{p} 1$. Slutsky's theorem takes us home.

The same pattern, more generally

Theorem (Diaconis-Freedman)

This holds for a wide class of high-dimensional distributions. i.e., for a wide class of high-dimensional distributions, most 1-dimensional projections of a sample from the distribution will be indistinguishable from a sample of a normal distribution.

See their article "asymptotics of graphical projection pursuit"

Bumpiness

What makes an interesting projection?

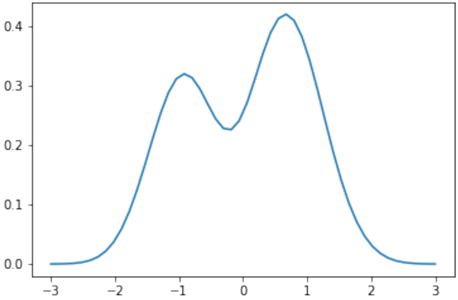


Figure: a density estimate of 50 samples from a noteworthy distribution

Forms multiple "bumps"

WithinSS

More precisely: given a real random variable X, Define the *WithinSS* as

$$W(X) = \min_{c \in \mathbb{R}} \frac{E[\mathrm{Var}(X|X < c)]}{\mathrm{Var}(X)}.$$

An interesting sample has a low W, that is, the variance is reduced considerably by thresholding at some point c and computing the variance in each half.

typical values of W

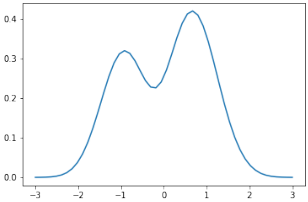
Remark

the law of total variance states

$$Var(X) = E[Var(X|Y)] + Var(E[X|Y])$$

which guarantees $0 \leq W \leq 1$.

For $X \sim \text{Unif}(\{-1, 1\})$, $W(X) = 0$. with $c = 0$.
For $X \sim N(0, 1)$, $W(X) = 1 - \frac{2}{\pi} = 0.363$, with $c = 0$.
For the empirical distribution shown below, $W(X) = 0.155$.



Anomalous W values

- Problem: the distribution shown above is a sample of 50 points from a standard Gaussian distribution. (I drew 1,000,000 such 50-point samples and chose the one with the lowest W)
- Solution: describe the distribution of W on an empirical sample of n points from a standard Gaussian, so that we know how good a given value of W is.
- Idea: as we get more samples, the ideal threshold c will surely go to 0. Then

$$W \approx \frac{E[\mathrm{Var}(X|X<0)]}{\mathrm{Var}(X)} = \frac{\mathrm{Var}(|X|)}{\mathrm{Var}(X)}.$$

My theorem

Letting Z be a standard Gaussian random variable, define

$$\begin{aligned} \mu &= E[|Z|] = \sqrt{2/\pi}, \\ \sigma^2 &= \mathrm{Var}(|Z|) = 1 - 2/\pi, \\ \kappa^2 &= E\left[\left((|Z|-\mu)^2 - \sigma^2|Z|^2\right)^2\right] = 8(\pi-3)/\pi^2. \end{aligned}$$

Theorem (due to me!)

Let $X_1, \ldots, X_n$ be independent standard Gaussian random variables. Let $\overline{X} = \frac{1}{n}\sum X_i$ and let $\overline{X}' = \frac{1}{n}\sum |X_i|$. Then the quantity

$$\frac{\sqrt{n}}{\kappa}\left[\frac{\frac{1}{n}\sum_{i=1}^n (|X_i| - \overline{X}')^2}{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2} - \sigma^2\right]$$

converges in distribution to a standard Gaussian random variable as $n \to \infty$.

How good is this estimate?

Let's try 100,000 samples for various n.

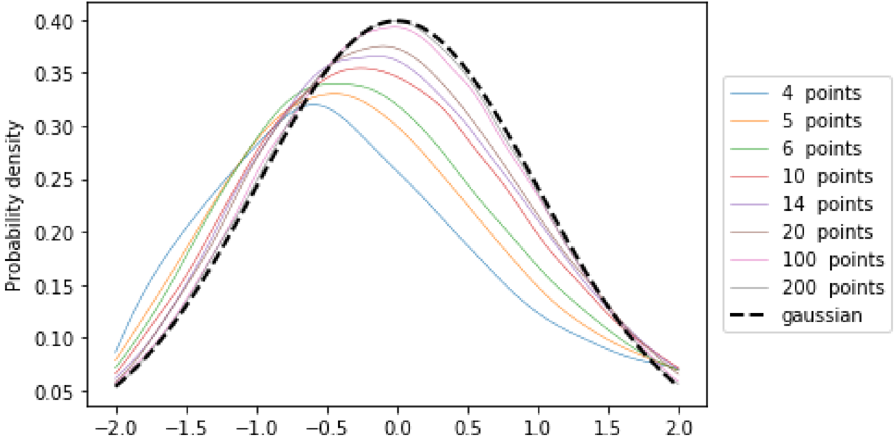


Figure: $\frac{\sqrt{n}}{\kappa}\left[\frac{\text{Var}(|Z|)}{\text{Var}(Z)} - \sigma^2\right]$

How good was the approximation $W \approx \frac{\mathrm{Var}(|Z|)}{\mathrm{Var}(Z)}$?

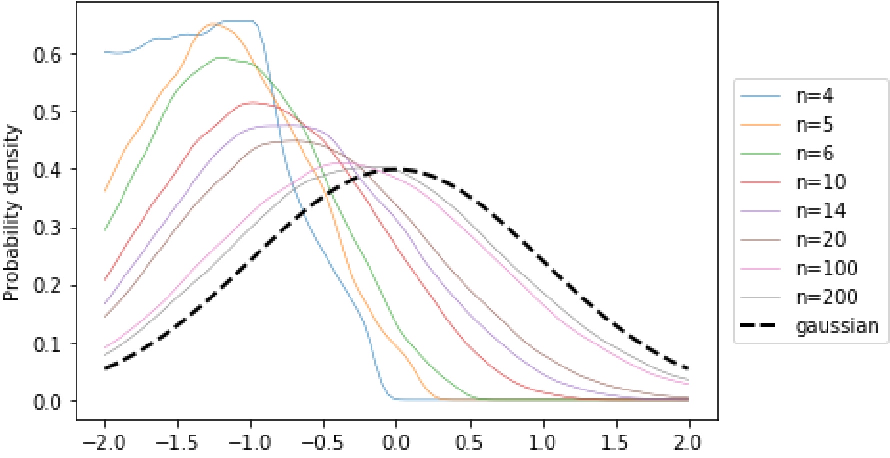


Figure: $\frac{\sqrt{n}}{\kappa}\left[W - \sigma^2\right]$

Can we get the means to line up?

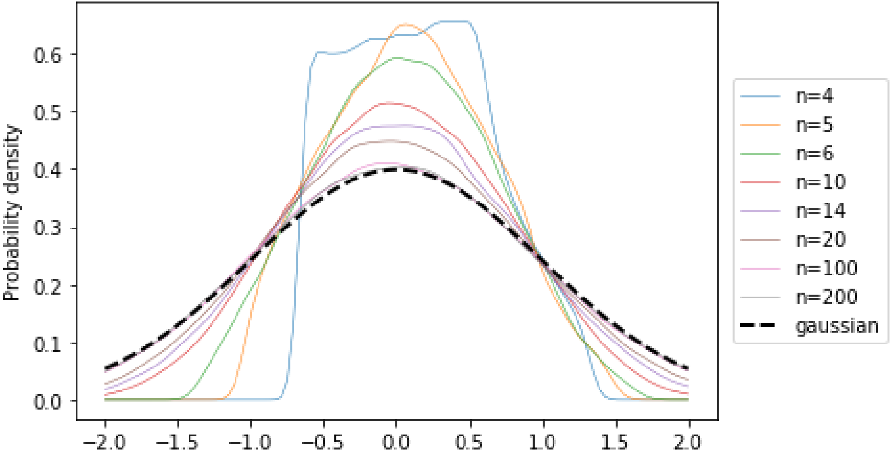


Figure: $\frac{\sqrt{n}}{\kappa}\left[W - \sigma^2 + \frac{1}{n}\right]$

Can we line up the spread?

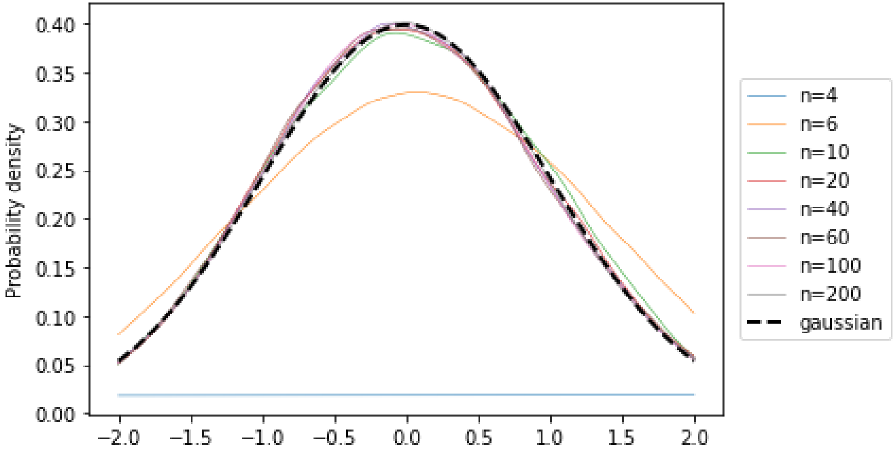


Figure: $\left(W - \sigma^2 + \frac{1}{n}\right) / \sqrt{\frac{\kappa^2}{n} - \frac{0.4}{n^{1.9}}}$

Conclusion

- For n samples drawn from a standard Gaussian, W will be approximately distributed as a Gaussian with center $\sigma^2 - \frac{1}{n}$ and standard deviation $\frac{\kappa^2}{n} - \frac{0.4}{n^{1.9}}$.
- We can use this to estimate the likelihood of a given W observation based on the hypothesis that the samples were drawn from a Gaussian.
- For example, the anomalous distribution we looked at before had a W score of W(X)$=$ 0.155.
- Using this approximation with $n = 50$, we find a 1 in 60,000 chance of getting a W that low or lower. Not likely for a single sample, but very likely if you are taking 1,000,000 attempts and taking the lowest W.