Hi Eric,

My question relates to finding the density of a function of two random variables.

The problem setting is as follows. We are given n observations of p variables. For convenient notation, we assume that each variable is centered at zero. We array the the ith variable's observations in an n-dimensional vector x_i. We want to find the pair of variables $(X_i, X_{i'})$, $i < i'$, that are most highly correlated in the dataset.

(This problem, to find the two most highly correlated variables in a dataset, has some applications in machine learning. Example #1: The treelet algorithm (Lee, Nadler, and Wasserman 2008) repeatedly merges the two most highly correlated variables in a dataset using a Jacobi rotation; the result is an analog of the wavelet transform, but for unordered data. Example #2: A robot (or other learning machine) may have a large number of tasks it wants to perform and a large number of possible variables it can observe. However, observing each variable has some cost, so it is desirable not to observe superfluous variables. One approach to determining which variables to select is to choose those with highest correlation with the predictor variables.)

Consider the following algorithm:

- Initially,
  - Set $\rho^* = -\infty$.
  - Set b arbitrarily $\in \{1, \ldots, p\}$; b is the "bridge" variable.
  - For $i = 1, \ldots, p$:
    - Set $\rho_{ib} = \frac{x_i^T x_b}{\sqrt{x_i^T x_i}\sqrt{x_b^T x_b}}$, the sample correlation between X_b and X_i. (Each x_i denotes a n-dimensional vector We assume the data is centered at zero.)
- For $i = 1, \ldots, p$,
  - For $i' = i + 1, \ldots, p$,
    - Set $r_{ii'} = \rho_{ib}\rho_{i'b} + \sqrt{(1 - \rho_{ib}^2)(1 - \rho_{i'b}^2)}$, an upper bound on the sample correlation between the variables X_i and $X_{i'}$.
    - If $r_{ii'} > \rho^*$,
      - Set $\rho_{ii'}$ to the sample correlation between X_i and $X_{i'}$, computed as above. (*)
      - If $\rho_{ii'} > \rho^*$, set $\rho^* = \rho_{ii'}$.

The $(X_i, X_{i'})$ corresponding to the last ρ^* are the most highly correlated variables.

A key question is how many times we will need to compute the sample correlation in step (*), on average, i.e., how many times $r_{ii'} > \rho$.

I have been assuming that the ρ are drawn from a uniform distribution on $[-1, 1]$. (I think it would be better to assume that they are drawn from a Gaussian distribution, since each ρ is a sum of random variables, but I wanted to start with something as simple as possible.)

I have found a lower bound on the expected number of times $r_{ii'} > \rho^*$, namely, the expected number of times that $\rho_{ii'} > \rho^*$. This is

$$\frac{H_t - 1/2}{t} \sim \frac{\log(t)}{t},$$

where H_t is the tth harmonic number, and $t = p(p-1)/2$.

I have been trying to carry out a similar analysis for the expected number of times that $r_{ii'} > \rho^*$, but here I have run into a difficulty that I have been stuck on for a long time now.

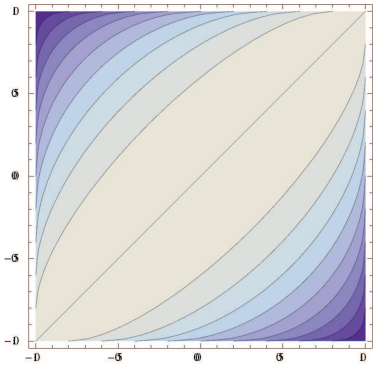
I cannot figure out how to compute the density of $r_{ii'}$. Please note that

$$r_{ii'} := r(\rho_{ib}, \rho_{i'b}) := \rho_{ib}\rho_{i'b} + \sqrt{(1-\rho_{ib}^2)(1-\rho_{i'b}^2)},$$

is a function of two uniform$[-1, 1]$ random variables. The difficulty is that r is not a "nice" function. If r were one-to-one or invertible, etc., then it would not be difficult to compute its density; but it is not.

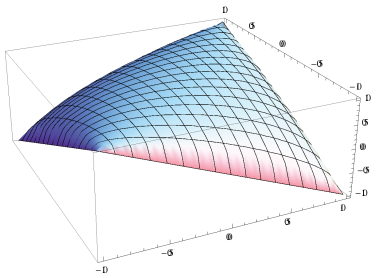
I think that the approach I need to take is as follows: I need to find a tractable parametric form for the level set $\{r_{ii'} \leq R\}$. Then I find the area of the level set and since the distribution is uniform, I can easily obtain the cdf and pdf.

The trick is to find a nice parametric form for the level set. This also I think I should be able to do, but for some reason I have been failing. Here are some plots that suggest that it should not be too hard. On the left, below, is a contour plot of $r(\rho_{ib}, \rho_{i'b})$ for $-1 < \rho_{ib}, \rho_{i'b} < 1$, and on the right is a surface plot of the same function.



(a)

and



(b)

It seems it should definitely be possible to find the area inside and outside those contours. I was wondering if you have any advice or know a place I could look to find more information to become unstuck?

Thank you,
Nate