Machine Learning Problem Set 1

Sebastian Wolf

31.01.2019

Contents

1 Mean estimators

For this problem, I sample from a Gaussian, a t-distribution with 2 and 3 degrees of freedom to test the performance of the empirical mean, the trimmed mean, and the median of means.

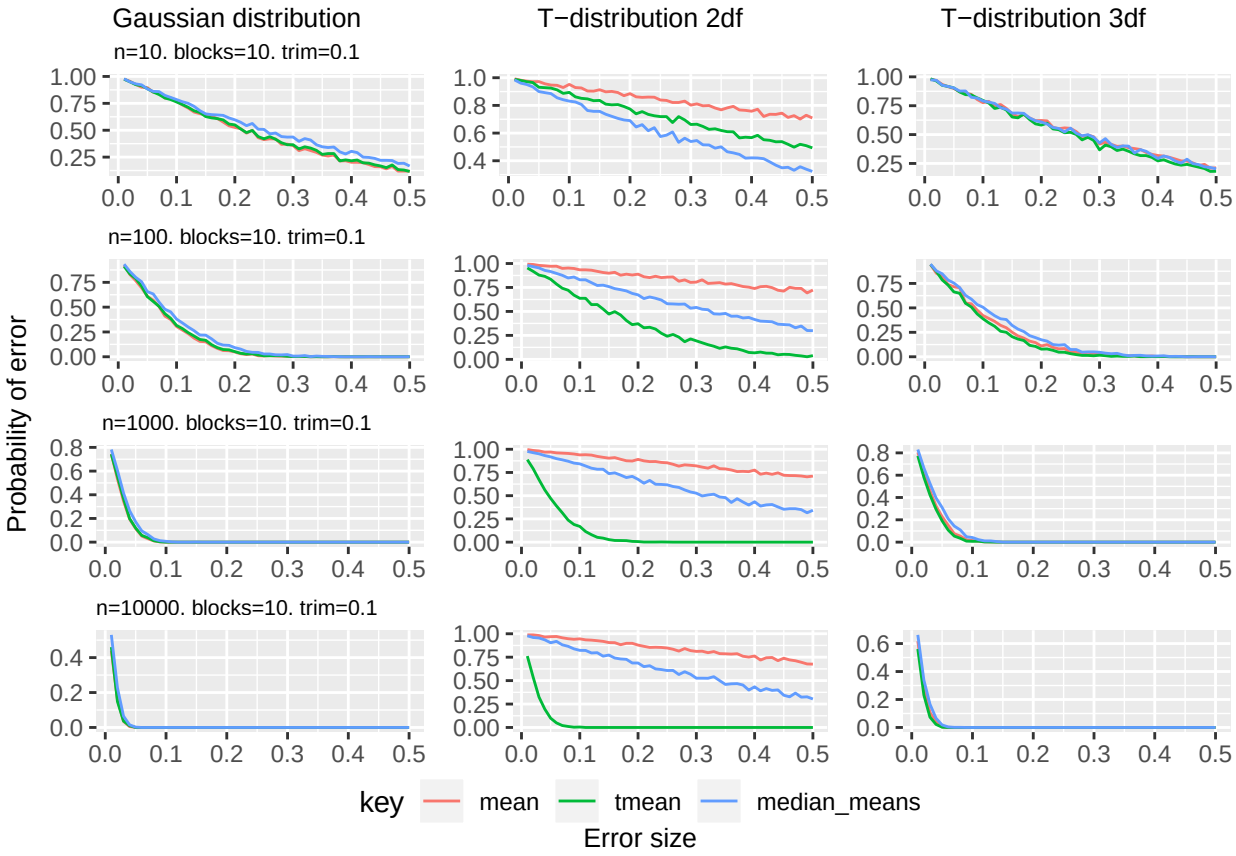
I compare the probability of error (meaning the number of times, out of all trials, that the difference between estimated mean and true mean is more than the error threshold set, divided by the total number of trails) of the three estimators for different sample sizes, block-sizes (for the median of means) and trim-sizes (for the trimmed mean) over the error range [0-0.5]. I do this in two steps:

(Figure 1) generate random vectors, one of each distribution, of lengths 10, 100, 1000 and 10000, and calculate the mean estimators keeping trim-size fixed at 10% of the sample and the number of blocks fixed at 10. (Figure 2) generate random vectors, one of each distribution, of lengths 20, and calculate the mean estimators, varying the trim-size of the trimmed mean from 0.1 to 0.4, and varying the number of blocks of the median of means estimator from 4-20

Throughout, I generate 1000 trials to compute the probabilities.

I expect the following from the results: the central limit theorem, which provides an exponential bound for the empirical mean for the probability of a tail event only holds asymptotically, so for small sample sizes we can bound the probability of a tail event for the empirical mean only by Chebychev's inequality, which is linear in n. The median of means has an exponential bound for the probability of a tail event that holds at any sample size, so we would expect that it bounds tail events better for small n, and for heavy tailed distributions which generate many tail events. We haven't looked at the bounds for the trimmed mean, but since this estimator cuts outliers it can be expected to perform well in heavy tailed distributions, too.

1.1 Performance of mean estimators for various sample sizes - Figure 1



The results show that for the Gaussian distribution the empirical mean performs best at all sample sizes, whereas for the t-distribution of 2 and 3 degrees of freedom the trimmed mean performs best. The median of means performs well only at small sample sizes (when we use 10 blocks the median of mean is equivalent to the median in a sample of size 10).

In figure 2, we focus on small samples and compare the empirical mean to the median of means with different block sizes and the trimmed mean with different trim sizes.

1.2 Performance of mean estimators for various trim sizes and numbers of blocks parameters - Figure 2

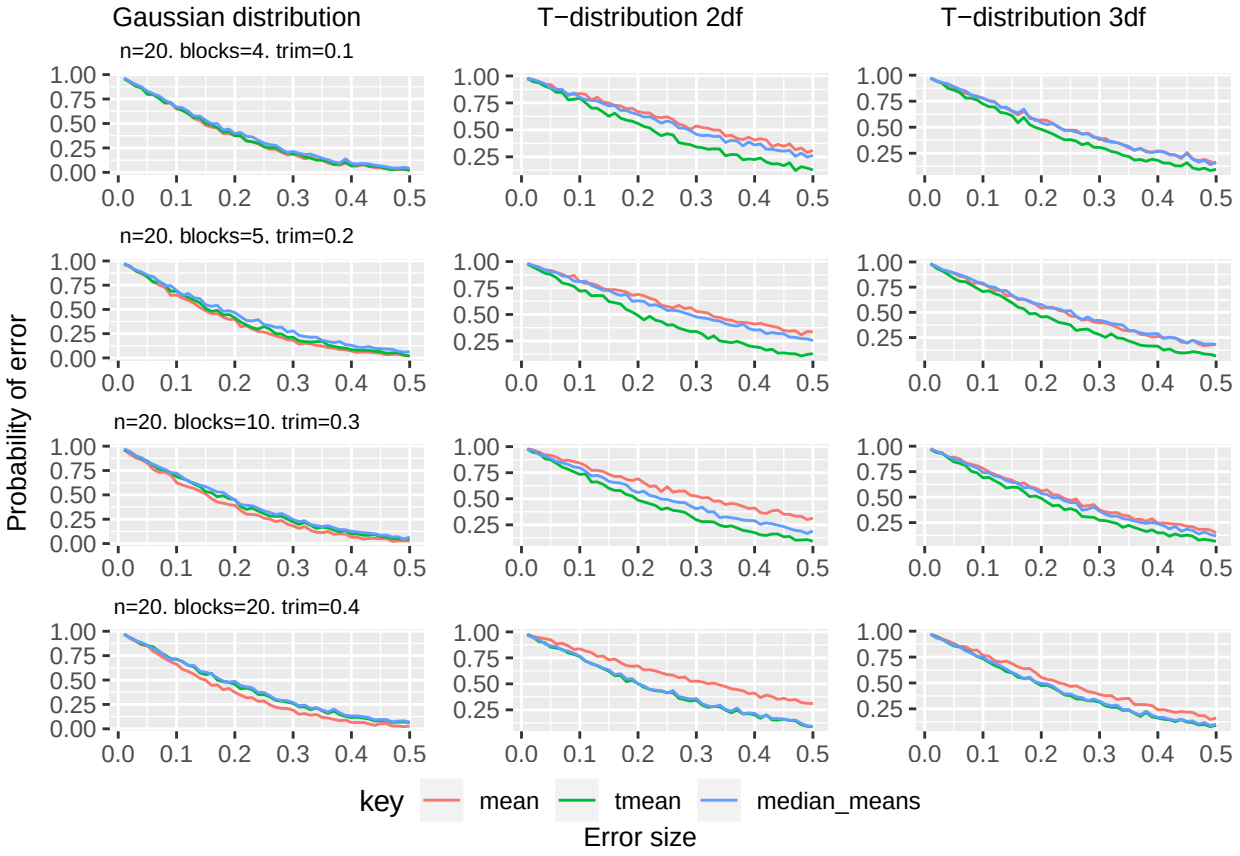


Figure 2 shows that under a gaussian distribution the trimmed mean performs better when few observations are trimmed (as then it is similar to the empirical mean), and the median of means performs better when few blocks are used (again, then it is similar to the empirical mean). Under a t-distribution with 2 degrees of freedom, the median of means performs slightly better with more blocks, and the trimmed mean with higher trim values. These tendencies are less pronounced for the t-distribution with 3 degrees of freedom - since here the tails are less heavy.

2 Random vectors in d-dimensional cubes

To derive the moments of the distribution of $\|\mathbf{X}^2\|$ I note that $\|\mathbf{X}\|^2 = \sum x_i^2$ and that each x_i is independent. I can therefore derive moments for x_i^2 and multiply by d.

2.1 Moments

2.1.1 Mean

$$E[x_i^2] = \int_{-1}^{1} x_i^2 \frac{1}{1-(-1)} dx = \frac{1}{2}\int_{-1}^{1} x_i^2 dx = \frac{1}{2}[\frac{1}{3} + \frac{1}{3}] = \frac{1}{3}$$

$$E[\|\mathbf{X}\|^2] = \sum E[x_i^2] = \frac{d}{3}$$

2.1.2 Variance

$$E[x_i^4] - E[x_i^2]^2 = \int_{-1}^{1} x_i^4 \frac{1}{1-(-1)} dx - \frac{1}{9} = \frac{1}{2}\int_{-1}^{1} x_i^4 dx - \frac{1}{9} = \frac{1}{2}[\frac{1}{5} + \frac{1}{5}] - \frac{1}{9} = \frac{4}{45}$$

$$Var[\|\mathbf{X}\|^2] = \sum Var[x_i^2] = \frac{4d}{45}$$

2.2 Concentration inequalities

2.2.1 Markov

$$P(\|\mathbf{X}\|^2 > \epsilon) \leq \frac{E(\|\mathbf{X}\|^2)}{\epsilon} = \frac{d}{3\epsilon}$$

2.2.2 Chebychev

$$P(|\|\mathbf{X}\|^2 - \frac{d}{3}| > \epsilon) \leq \frac{4d}{45\epsilon^2}$$

2.2.3 Chernoff bound

$$P(\|\mathbf{X}\|^2 - \frac{d}{3} \geq \epsilon) = P(e^{\lambda(\|\mathbf{X}\|^2 - \frac{d}{3})} \geq e^{\lambda\epsilon})$$

$$P(e^{\lambda(\|\mathbf{X}\|^2 - \frac{d}{3})} \geq e^{\lambda\epsilon}) \leq \frac{Ee^{\lambda(\|\mathbf{X}\|^2 - \frac{d}{3})}}{e^{\lambda\epsilon}} = \frac{Ee^{\lambda(\sum x_i^2 - \frac{d}{3})}}{e^{\lambda\epsilon}} = \frac{E\prod e^{\lambda(x_i^2 - \frac{1}{3})}}{e^{\lambda\epsilon}} \overset{iid}{=} \frac{\prod Ee^{\lambda(x_i^2 - \frac{1}{3})}}{e^{\lambda\epsilon}} = \frac{(Ee^{\lambda(x_i^2 - \frac{1}{3})})^d}{e^{\lambda\epsilon}}$$

I exponentiate and apply Markov inequality. I then make use of independence of the x_i to move the expectation operator into the product. x_i^2 is bounded between [0,1] so I can apply Hoeffding's Lemma to get:

$$P(\|\mathbf{X}\|^2 - \frac{d}{3} \geq \epsilon) \leq e^{\frac{d\lambda^2}{2} - \lambda\epsilon}$$

This is a monotonous increasing function so I can optimise the exponent to get: $\lambda = \frac{\epsilon}{d}$.

$$P(\|\mathbf{X}\|^2 - \frac{d}{3} \geq \epsilon) < e^{-\frac{\epsilon^2}{2d}}$$

2.3 Cosine

I want to find a bound for the probability that the cosine is different from 0, meaning the probability that the two vectors are not orthogonal.

$$P(|cos(\alpha)| \geq \epsilon) = P(|\frac{X^T X'}{\|\mathbf{X}\|\|\mathbf{X}'\|}| \geq \epsilon)$$

I already established that $E[\|\mathbf{X}\|^2] = \frac{d}{3}$ and that $Var[\|\mathbf{X}\|^2] = \frac{4d}{45}$ and so the denominator of the above fraction is of order $\frac{d}{3} + \sqrt{\frac{4d}{45}} = O(d)$. I use this to simplify the above probability and apply a Chernoff bound:

$$P(|cos(\alpha)| \geq \epsilon) \leq 2P(X^T X') > d\epsilon) \leq 2\frac{Ee^{\lambda X^T X'}}{e^{\lambda d\epsilon}} = 2\frac{E\prod e^{\lambda x_i x_i'}}{e^{\lambda d\epsilon}} \overset{iid}{=} 2\frac{\prod Ee^{\lambda x_i x_i'}}{e^{\lambda d\epsilon}} = 2\frac{(Ee^{\lambda x_i x_i'})^d}{e^{\lambda d\epsilon}}$$

with the second last equality arising from independence of the x_i. I can now apply Hoeffding's lemma (given that $Ex_i x_i' = 0$ and that x_ix_i' is bounded from [-1,1]) to get:

$$P(\frac{X^T X'}{\|\mathbf{X}\|\|\mathbf{X}'\|} > \epsilon) \leq 2e^{\frac{\lambda^2 d}{2} - \lambda\epsilon}$$

Which is optimised with $\lambda = \epsilon$.

$$P(\frac{X^T X'}{\|\mathbf{X}\|\|\mathbf{X}'\|} > \epsilon) \leq 2e^{-\frac{\epsilon^2 d}{2}}$$

This means that the vectors will be very close to orthogonal in high dimensions:

$$|cos(\alpha)| < \sqrt{\frac{2log(\delta/2)}{d}} \quad w.p. \quad 1-\delta$$

3 Chernoff bound

To find the Chernoff bound of a non-negative random variable with given mean and variance, I start with exponentiating, applying Markov inequality, and using the independence property of the sample.

$$P(\frac{1}{n}\sum_{i=1}^{n} x_i < m - t) = P(-\sum_{i=1} x_i > n(t-m)) \leq \frac{Ee^{-\lambda(\sum x_i)}}{e^{\lambda n(t-m)}} \overset{iid}{=} \frac{E(e^{-\lambda(x_i)})^n}{e^{\lambda n(t-m)}}$$

I can then use the hint:

$$\leq \frac{E(1 - \lambda x_i + \frac{\lambda^2 x_i^2}{2})^n}{e^{\lambda n(t-m)}} = \frac{(1 - \lambda m + \frac{\lambda^2 a^2}{2})^n}{e^{\lambda n(t-m)}}$$

And using the fact that $1 + x \leq e^x$ I get:

$$P(\frac{1}{n}\sum_{i=1}^{n} x_i < m - t) \leq e^{(-\lambda m + \frac{\lambda^2 a^2}{2})^n - \lambda n(t-m)} = e^{n(\frac{\lambda^2 a^2}{2} - \lambda t)}$$

I can optimise this bound with $\lambda = \frac{t}{a^2}$, which gives the bound I am looking to prove:

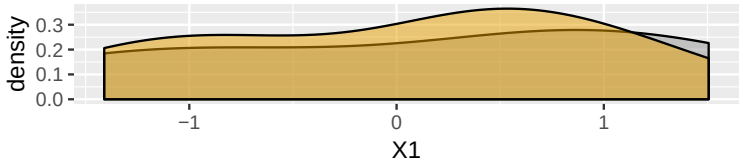
$$P(\frac{1}{n}\sum_{i=1}^{n} x_i < m - t) = e^{-\frac{nt^2}{2a^2}}$$

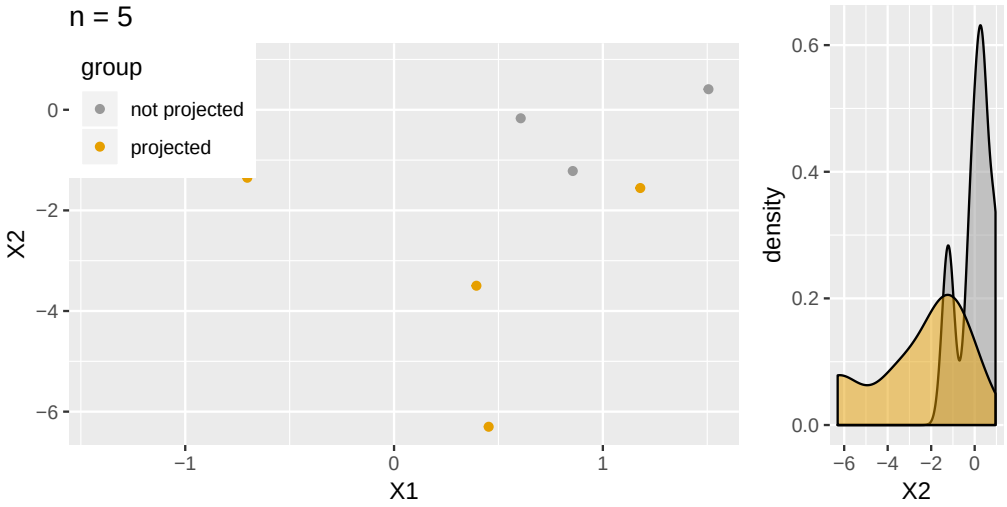
4 Random projections

I generate random projections for vectors of size 5, 50 and 1000. I do not see a difference between the projected points and the random vectors. According to the Johnson-Lindenstrauss lemma, random projections allow for drastic dimensionality reductions that preserve the pairwise distances between the projected points if we allow for some slack in the distance. In this example, I would expect all points to have approximately equidistance between each other if the Lemma applied. The Lemma holds whenever $d \geq \frac{8log(n)}{\epsilon^2}$. In this example $d = 2$ is small, so the Lemma holds only for small n and large ϵ. If we took $\epsilon = 1$ we would get $1.28 \geq n$ - not a case of interest. If we allowed for a lot of slack, say $\epsilon = 3$ then $9.48 \geq n$. However, then the initial distances would be unrecognizable.

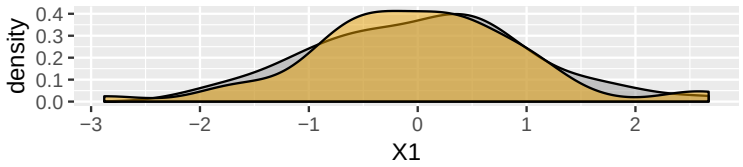
I show the three plots that overlay the projected points with the random vectors for sample sizes 5, 50 and 1000 below.

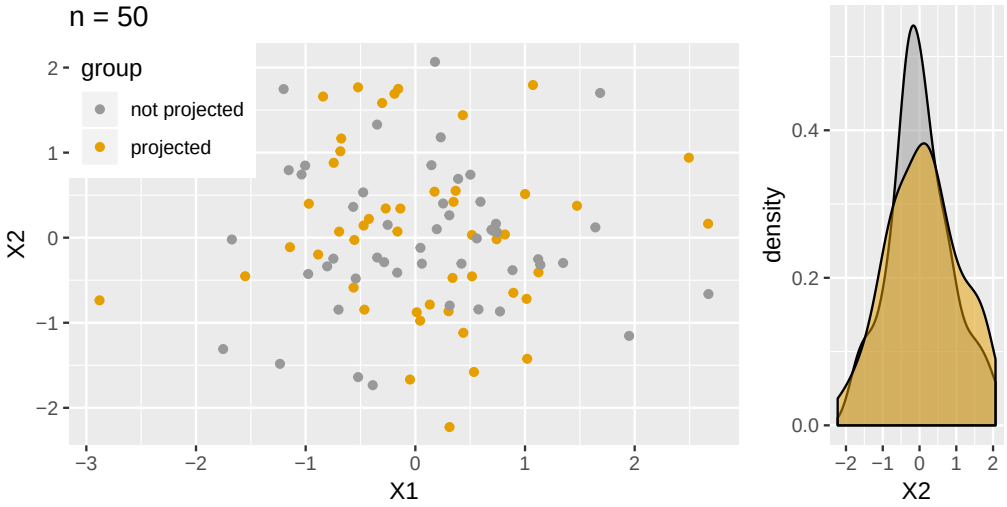
```
## [1] "Figure 3"
```



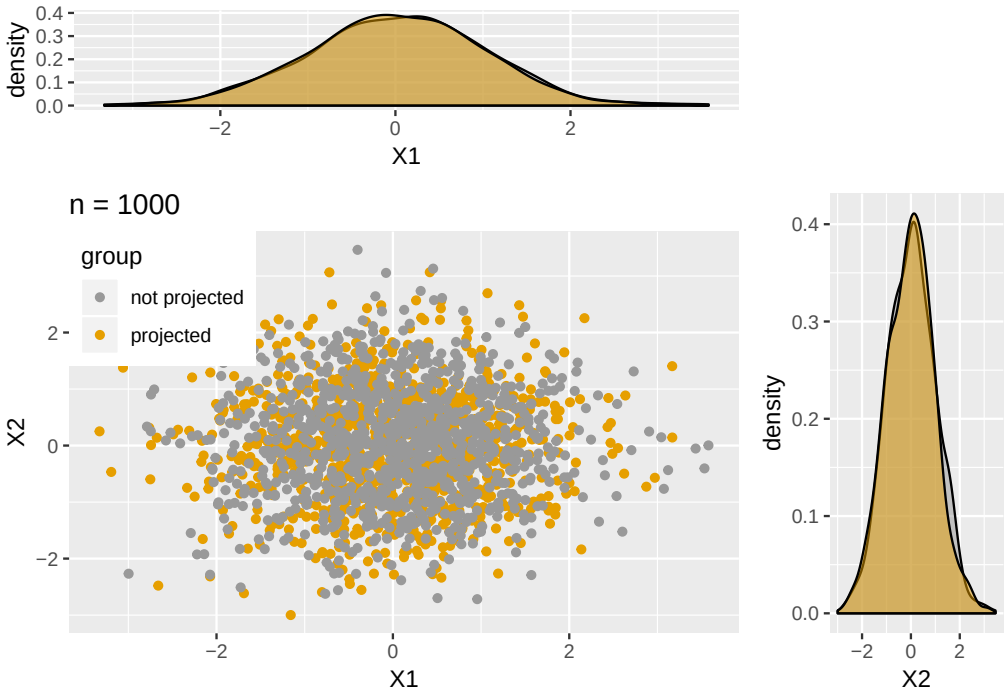



```
## [1] "Figure 4"
```






```
## [1] "Figure 5"
```


density
0.4
0.3
0.2
0.1
0.0
−2
0
2
X1
n = 1000
group
not projected
projected
X2
2
0
−2
−2
0
2
X1
density
0.4
0.3
0.2
0.1
0.0
−2
0
2
X2