Machine Learning @ Data science 2019: Set 1

Jordi Morera Serra

February 1, 2019

Exercise 1:

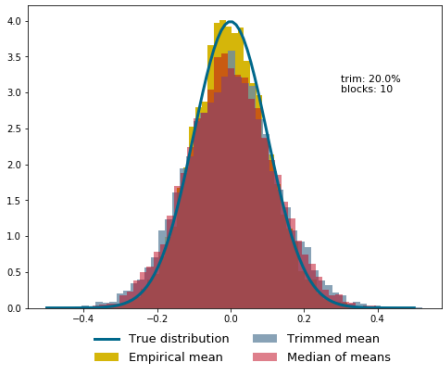
In order to compare the different estimators performance, we are going to split the analysis in two cases: thin and fat tailed distributions. The chosen ones are normal distribution and standard student-t distribution with different degrees of freedom.

$m_n$will be computed in three different ways:

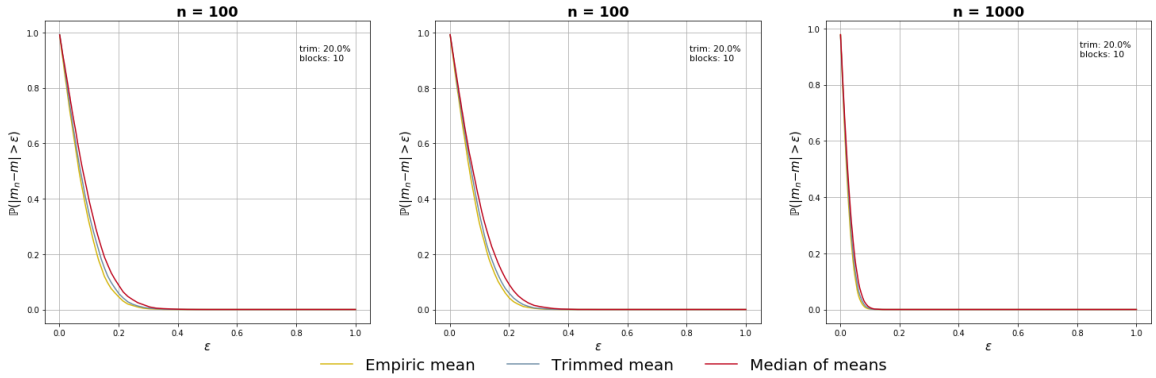
- Empirical mean
- Trimmed mean. The amount of data to be trimmed will be computed as percentage tr, instead of an absolute number, of the total amount of observations.
- Median of means with different blocks b.

Thin tailed distribution : Standard Normal

In figure 1 can be observed the mean distribution for $n = 100$ and the evolution of $P(|m_n - m| > \epsilon)$ over different ϵ for a sample of most representative number of variables $n = \{10, 100, 1000\}$.



(a) Mean distribution with n = 100



(b) Mean estimator performance analysis

Figure 1: n i.i.d Gaussian random variables mean estimators

After experimenting with different values on the parameters tr and b, one notices that the best performance is reached with values $tr = 0$ and $b = 1$ respectively. This would be equivalent, in both cases, to the empirical mean. In order to illustrate their low performance compared to the empirical mean, tr and b have been fixed to 20% and 10 respectively.

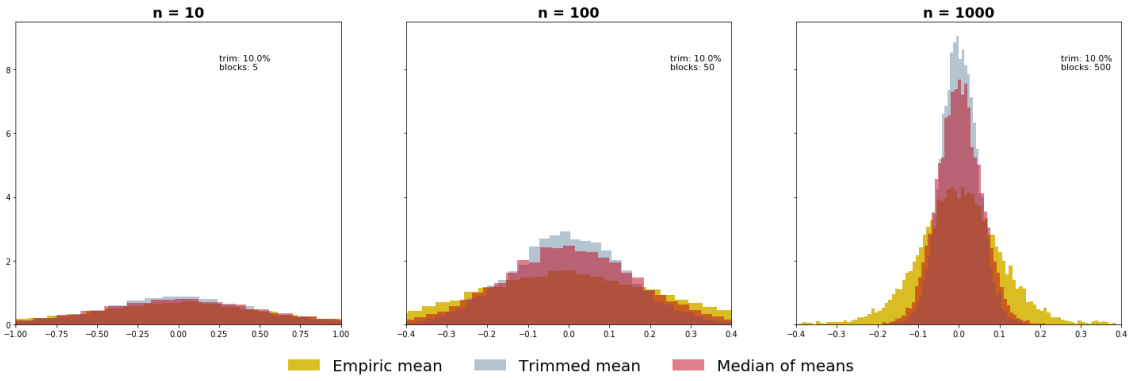
Remarks:

- The empirical mean estimator shows the best performance in all situations.
- As observed in figure 1a, both trimmed mean and median of means have wider tails than the true mean distribution. This is translated to figure 1b since for a given ϵ the probability of falling outside $\pm\epsilon$ interval is higher.
- The true distribution of the mean follows a $\mathcal{N}(0, \frac{1}{n})$. Logically, the probability that one observation is far from the true mean decreases when n increases. This is captured for all three estimators.

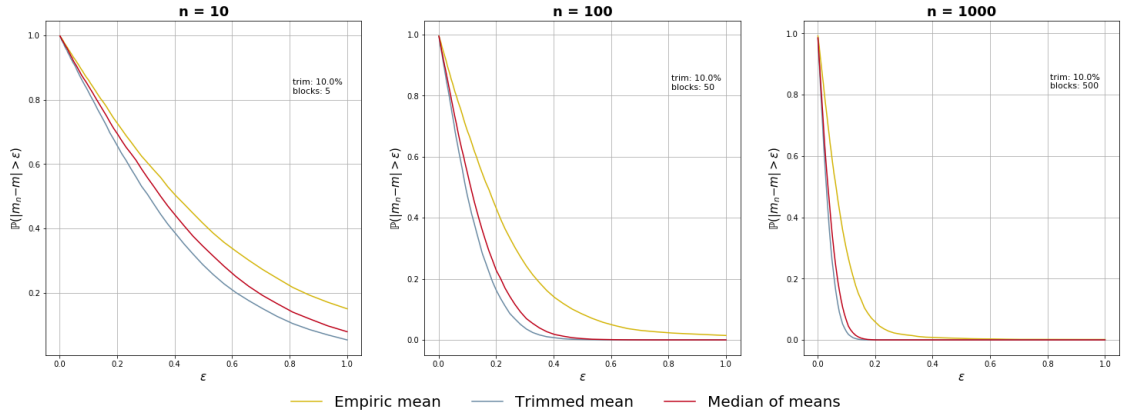
Fat tailed distribution : Student-t

As an example of a fat tailed distribution, we have chosen the student-t distribution. Note that for degrees of freedom from $1 < df \leq 2$ the variance does not exist and as dfincreases the student-t distribution tends to a normal distribution. In order to show interesting cases in the analysis of different mean estimators, we will analyze two specific cases: one, with infinite variance $df = 2$,and another one, with $df = 5$, which has fatter tails than a normal distribution and the variance exists.

df=2 Performance analysis can be found in figure 2a. The parameters for the trimmed mean and the MoM have been chosen in order to maximize its performance, i.e. low $P(|m_n - m| > \epsilon)$ for a given n.



(a) Mean distribution for different mean estimators.



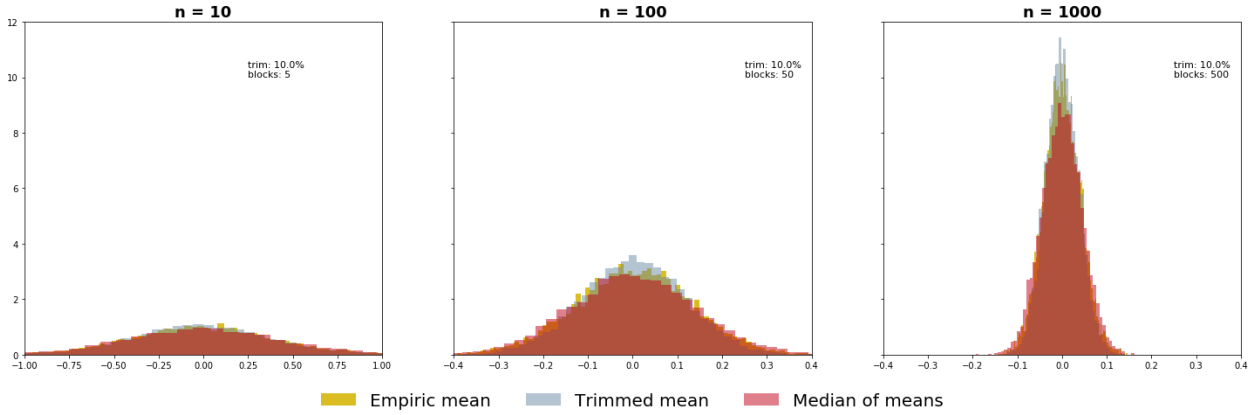
(b) Mean estimator performance analysis

Figure 2: n i.i.d student random variables mean estimators

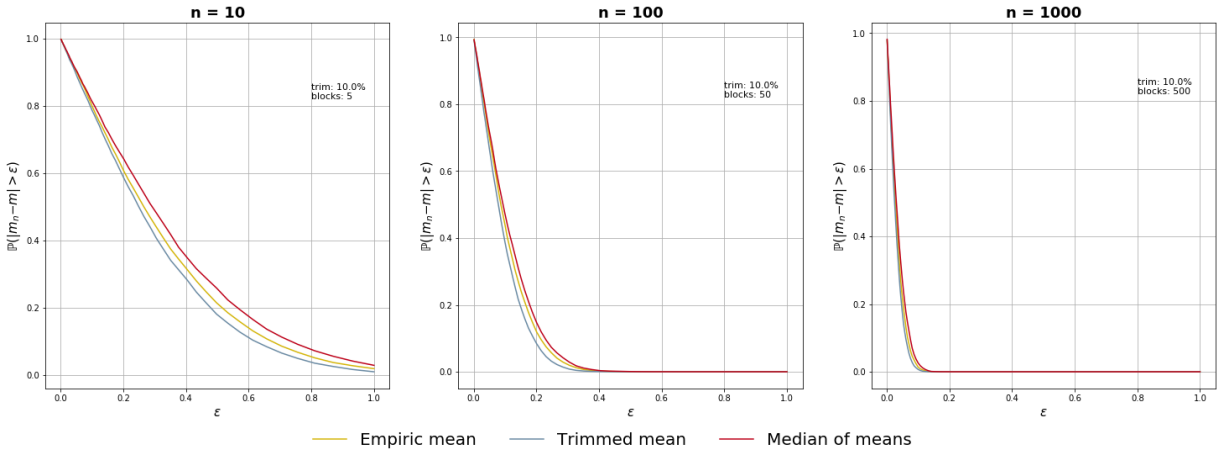
Remark

- In figure 2a one can see that MoM and trimmed mean estimator are more robust when applied to fat tailed distributions. It is to say, it concentrates around the expected mean with lower variance. This translates, as we can see in figure 2b, to a lower probability of falling outside $\pm\epsilon$, for a given n and ϵ.
- As expected, for all m_n, its variance decreases when n increases.
- When computing MoM, if the number of $b = n$, the estimator is directly the median. It turns out that the median is also a robust estimator for fat tailed distribution, and, in fact, it performs as well as the trimmed mean estimator.
- Increasing b when n increases, improves MoM performance. Similarly, trimmed mean also improves its performance trimming out more data for higher n.

df=5 Performance analysis can be found in figure 3. Again, the parameters for the trimmed mean and the MoM have been chosen in order to maximize its performance, i.e. low $P(|m_n - m| > \epsilon)$ for a given n.



(a) Mean distribution for different mean estimators.



(b) Mean estimator performance analysis

Figure 3: n i.i.d student random variables mean estimators

Remarks:

- Here we find and intermediate situation between the two previous cases. Although trimmed mean is still the best performer, as df increases, the performance of our estimators tends to the performance of the normal distribution case.

Exercise 2:

Mean

$$\mathbb{E}\left\|X\right\|^2 = \mathbb{E}(\sum_{i=1}^{d} X_i) \underset{i.i.d}{=} d\mathbb{E}X_1^2 \underset{\mathbb{E}X_1=0}{=} \frac{dVar(X_1)}{3} = \frac{d}{3}$$

Variance

$$Var(\left\|X\right\|^2) = Var(\sum_{i=1}^{d} X_i) \underset{i.i.d}{=} dVar(X_1^2)$$

The n-th moment of the uniform distribution is: $\mathbb{E}(x^n) = \frac{1}{1+n}\sum_{k=0}^{n} a^k b^{n-k}$ where $a = -1$ and $b = 1$.

$$dVar(X_1^2) = d\mathbb{E}X_1^4 - d(\mathbb{E}X_1^2)^2 = \frac{4d}{45}$$

Concentration inequalities

Markov inequality:
Since the random variable $\left\|X\right\|^2$ is non-negative,

$$\mathbb{P}(\left\|X\right\|^2 \geq t) \leq \frac{\mathbb{E}(\left\|X\right\|^2)}{t} = \frac{d}{3t}$$

Chebyshev's inequality
Since $\left\|X\right\|^2 \in \mathbb{R}$

$$\mathbb{P}(|\left\|X\right\|^2 - \mathbb{E}\left\|X\right\|^2| \geq t) \leq \frac{Var(\left\|X\right\|^2)}{t^2} = \frac{4d}{45t^2}$$

Chernoff bounds

$$\mathbb{P}(\left\|X\right\|^2 - \mathbb{E}\left\|X\right\|^2 \geq t) \leq \frac{\mathbb{E}e^{\lambda(\left\|X\right\|^2 - \mathbb{E}\left\|X\right\|^2)}}{e^{\lambda t}}$$

$$\mathbb{P}(\sum_{i=1}^{d} X_i^2 - \frac{d}{3} \geq t) = \mathbb{P}(\sum_{i=1}^{d}(X_i^2 - \frac{1}{3}) \geq t) \leq \frac{\mathbb{E}e^{\lambda(\sum_{i=1}^{d}(X_i^2 - \frac{1}{3}))}}{e^{\lambda t}} \underset{i.i.d}{=} \frac{\left[\mathbb{E}e^{\lambda(X_i^2 - \frac{1}{3})}\right]^n}{e^{\lambda t}}$$

if we now apply Hoeffding lemma:

$$\frac{\left[\mathbb{E}e^{\lambda(X_i^2 - \frac{1}{3})}\right]^n}{e^{\lambda t}} \leq e^{\frac{\lambda^2 n}{2} - \lambda t} \tag{1}$$

As $e^{f(x)}$ is an increasing monotonous function, optimizing (3) for λ is the equivalent to optimize $\frac{\lambda^2 n}{2} - \lambda t$.

$$\frac{d}{d\lambda}\left(\frac{\lambda^2 n}{8} - \lambda t\right) = 0 \ \rightarrow \ \lambda = \frac{4t}{n}$$

Plugging this result to (1), we obtain:

$$\mathbb{P}(\left\|X\right\|^2 - \mathbb{E}\left\|X\right\|^2 \geq t) \leq e^{-\frac{2t^2}{n}}$$

Order of magnitude of cosine

First let's have a look to the cosine of the angle between two vectors, X and X', uniformly distributed in the d cube.

$$cos(\alpha) = \frac{X^T X'}{\|X\|\|X'\|}$$

If $\|X\|\|X'\|$ behaves well, it should be of the order of the $\mathbb{E}\|X\|^2 = d/3$ and with a standard deviation of $\sqrt{Var(\|X\|^2)} = \sqrt{\frac{4d}{45}}$, i.e., $O(d)$. Therefore,

$$\frac{X^T X'}{\|X\|\|X'\|} \simeq \frac{X^T X'}{d} \ \rightarrow \ \mathbb{P}\big(|\frac{X^T X'}{d}| > \epsilon\big)$$

As we are interested in the order of magnitude, and $\big(\frac{X^T X'}{d}\big)$ is symmetric, the order of magnitude of

$$\mathbb{P}\big(|\frac{X^T X'}{d}| > \epsilon\big) \approx 2\mathbb{P}\big(|\frac{X^T X'}{d}| > \epsilon\big) \approx \mathbb{P}\big(\frac{X^T X'}{d} > \epsilon\big)$$

.

$$\mathbb{P}\big(\frac{X^T X'}{d} > \epsilon\big) = \mathbb{P}\big(X^T X' > d\epsilon\big) = \mathbb{P}\big(\sum_{i=0}^{d} X_i X_i^{'} > d\epsilon\big) \underset{Chernoff}{\leq} \frac{\mathbb{E}\big(e^{\lambda \sum_{i=1}^{d} X_i X_i^{'}}\big)}{e^{\lambda \epsilon d}} \underset{i.i.d}{=} \frac{\Big[\mathbb{E}\big(e^{\lambda X_1 X_1^{'}}\big)\Big]^d}{e^{\lambda \epsilon d}}$$

Notice that $X_1 X_1^{'}$ is bounded between $[-1, 1]$ and $\mathbb{E}X_1 X_1^{'} = \mathbb{E}X_1 \mathbb{E}X_1^{'} = 0$ and we can apply Hoeffding.

$$\frac{\Big[\mathbb{E}\big(e^{\lambda X_1 X_1^{'}}\big)\Big]^d}{e^{\lambda \epsilon d}} \underset{Hoeffding}{\leq} \frac{e^{\frac{\lambda^2 d}{2}}}{e^{\lambda \epsilon d}} = e^{\frac{\lambda^2 d}{2} - \lambda \epsilon d} \tag{2}$$

As $e^{f(x)}$is an increasing monotonous function, optimizing (2) for λ is the equivalent to optimize $\frac{\lambda^2 d}{2} - \lambda \epsilon d$.

$$\frac{d}{d\lambda}\left(\frac{\lambda^2 d}{2} - \lambda \epsilon d\right) = 0 \ \rightarrow \ \lambda = \epsilon \tag{3}$$

Plugging (3) to (2) we obtain

$$\mathbb{P}\big(|cos(\alpha)| > \epsilon\big) \leq e^{-\frac{\epsilon^2 d}{2}}$$

Which translates to

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

Exercise 3:

We want to proof the following:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i < m - t\right) \leq e^{-nt/(2a^2)}$$

Let's start from the below expression and multiply both sides of the inequality by -1 and multiply both sides by n.

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i < m - t\right) = \mathbb{P}\left(-\sum X_i \geq n(t-m)\right) \tag{4}$$

We apply Chernoff bounds to (4)

$$\mathbb{P}\left(-\sum_{i=1}^{n} X_i \geq n(t-m)\right) \leq \frac{\mathbb{E}e^{-\lambda\sum_{i=1}^{n} X_i}}{e^{\lambda n(t-m)}} \underset{i.i.d}{=} \frac{\left[\mathbb{E}e^{-\lambda X_1}\right]^n}{e^{\lambda n(t-m)}} \underset{hint}{\leq} \frac{\left[\mathbb{E}[1 - \lambda X_1 + \frac{\lambda^2 n^2}{2}]\right]^n}{e^{\lambda n(t-m)}} =$$

$$\underset{\substack{\mathbb{E}X_1 = m \\ \mathbb{E}X_1^2 = a^2}}{=} \frac{\left(1 - \lambda m + \frac{\lambda^2 a^2}{2}\right)^n}{e^{\lambda n(t-m)}} \underset{\substack{1+x \simeq e \\ x\, small}}{\simeq} \frac{\left[e^{-\lambda mn + \frac{\lambda^2 a^2}{2}}\right]}{e^{\lambda n(t-m)}} = e^{\frac{n\lambda^2 a^2}{2} - \lambda nt} \tag{5}$$

As $e^{f(x)}$is an increasing monotonous function, optimizing (5) for λ is the equivalent to optimize $\frac{n\lambda^2 a^2}{2} - \lambda nt$.

$$\frac{d}{d\lambda}\left(\frac{n\lambda^2 a^2}{2} - \lambda nt\right) = 0 \;\rightarrow\; \lambda = \frac{t}{a^2}$$

Plugging this result to (5) proves our initial statement.

Exercise 4:

In figure 4 we can observe the projected standard basis vectors for an n dimensional space into a 2 dimensional space. Below there is an histogram of the pair-wise distances of the projected vectors (points).

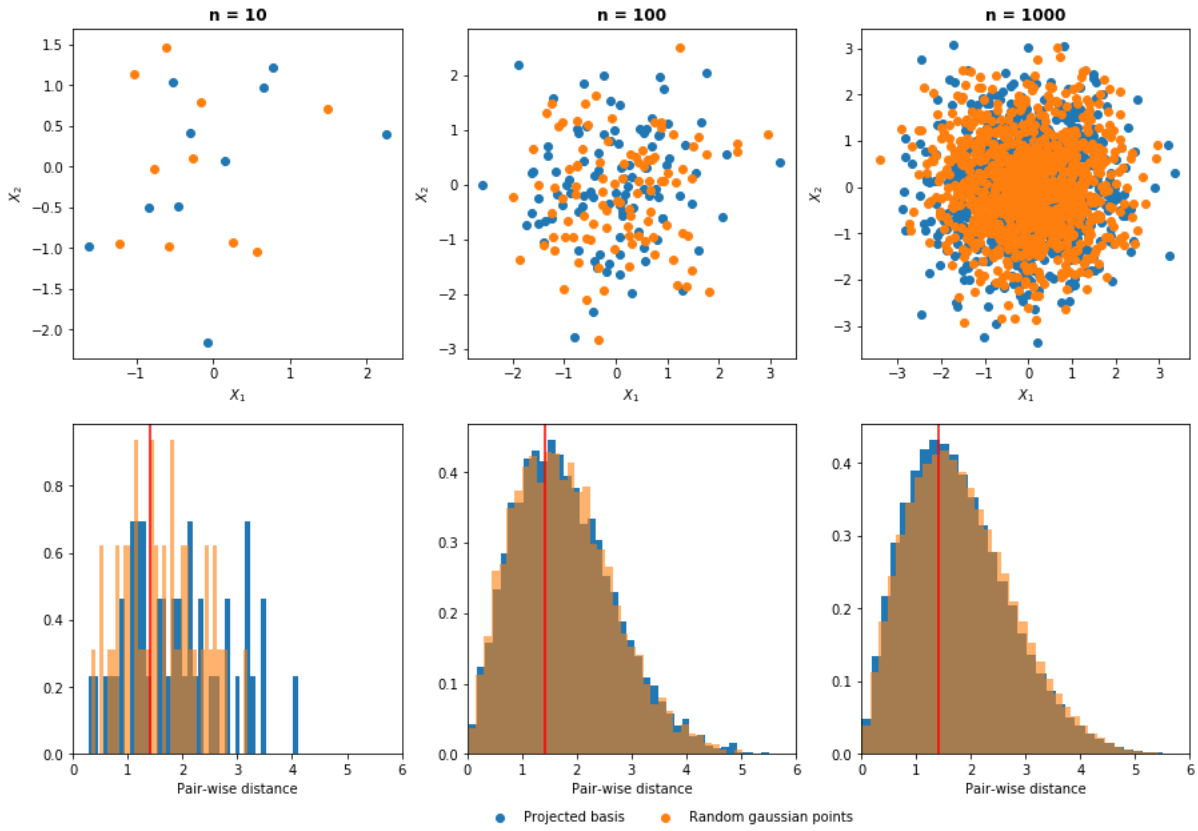


Figure 4: Random projection of n standard basis vectors

Remarks

- Mathematically, if we consider the n standard basis vectors in $\mathbb{R}^n$and we put them into a matrix appropriately ordered, we obtain $\mathbb{I}_{nxn}x\mathbb{M}_{nx2}$. Where $\mathbb{M}_{ij} \sim \mathcal{N}(0,1)$. Hence, what we see in the plots is, logically, a bi-variate random normal with the same $\mu = 0$ and $\sigma^2 = 1$. In orange we can see a bi-variate standard normal for comparison.

- Being a_i,$a_j \in n$ standard basis vectors ("points"), we can compute $\|a_i - a_j\| = \sqrt{2}\,\forall\, i,j$. As proved in class, $\mathbb{E}\|\mathbb{M}(a_i - a_j)\| = \|a_i - a_j\|$ which is exactly what we see in the below distributions, plotted with a red line. This distribution is the one you get when you compute the euclidean distance between two points whose components follow a normal distribution.

- Let's now turn our attention to Johnson–Lindenstrauss lemma which states that the distances before and after the projection to a lower dimensional space gets roughly preserved if $d \simeq O\big(\frac{log(n)}{\epsilon^2}\big)$. Where d is the reduced space, n is the number of points and ϵ is the tolerance between the distance before and after the projection. In our example n is fixed and equal to the original dimension D. Since we assume ϵ small, the condition is never fulfilled no matter the value of n we choose. Also, notice that the more we increase n, the more difficult is to maintain the distances after the projection.