## TOPIC 1

# Least square: Cholesky, QR and Householder

## 5 Further observations [optional]

---

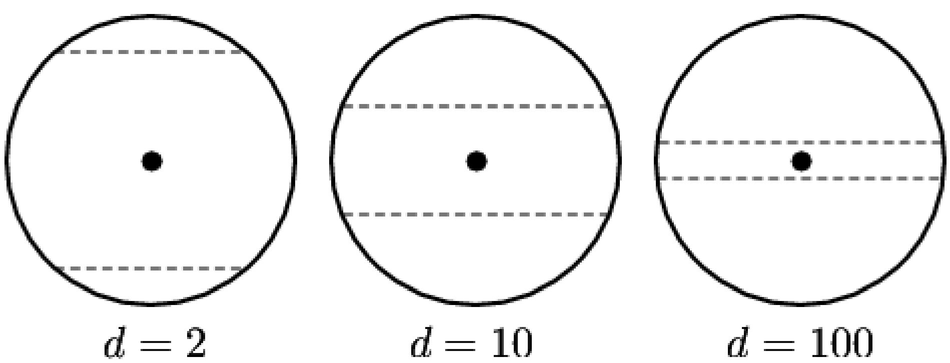
*Course:* Math 535 (http://www.math.wisc.edu/~roch/mmids/) - Mathematical Methods in Data Science (MMiDS)
*Author:* Sebastien Roch (http://www.math.wisc.edu/~roch/), Department of Mathematics, University of Wisconsin-Madison
*Updated:* Sep 13, 2020


---

### 5.1 Orthogonality in high dimension

In high dimension, orthogonality -- or more accurately near-orthogonality -- is more common than one might expect. We show this in this section.

Let $\mathbf{X}$ be a standard Normal $d$-vector. Its joint PDF depends only on the its norm $\|\mathbf{X}\|$. So $\mathbf{Y} = \frac{\mathbf{X}}{\|\mathbf{X}\|}$ is uniformly distributed over the $(d-1)$-sphere $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$, that is, the surface of the unit $d$-ball centered aroungd the origin. We write $\mathbf{Y} \sim \mathrm{U}[\mathcal{S}]$. The following theorem shows that if we take two independent samples $\mathbf{Y}_1, \mathbf{Y}_2 \sim \mathrm{U}[\mathcal{S}]$ they are likely to be nearly orthogonal when $d$ is large, that is, $|\langle \mathbf{Y}_1, \mathbf{Y}_2 \rangle|$ is likely to be small. By symmetry, there is no loss of generality in taking one of the two vectors to be the north pole $\mathbf{e}_d = (0, \ldots, 0, 1)$. A different way to state the theorem is that most of the mass of the $(d-1)$-sphere is in a small band around the equator.




(Source (https://marckhoury.github.io/blog/counterintuitive-properties-of-high-dimensional-space))

**Theorem (Orthogonality in High Dimension):** Let $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$ and $\mathbf{Y} \sim \mathrm{U}[\mathcal{S}]$. Then for any $\varepsilon > 0$, as $d \to +\infty$,

$$\mathbb{P}[|\langle \mathbf{Y}, \mathbf{e}_d \rangle| \geq \varepsilon] \to 0.$$

*Proof idea:* We write $\mathbf{Y}$ in terms of a standard Normal. Its squared norm is a sum of independent random variables. After bringing it to the numerator, we can apply Chebyshev.

*Proof:* Recall that $\mathbf{Y}$ is $\frac{\mathbf{X}}{\|\mathbf{X}\|}$ where $\mathbf{X}$ is a standard Normal $d$-vector. The probability we want to bound can be re-written as

$$\begin{aligned}
\mathbb{P}[|\langle \mathbf{Y}, \mathbf{e}_d \rangle| \geq \varepsilon] &= \mathbb{P}\left[\left|\left\langle \frac{\mathbf{X}}{\|\mathbf{X}\|}, \mathbf{e}_d \right\rangle\right|^2 \geq \varepsilon^2\right] \\
&= \mathbb{P}\left[\left|\frac{\langle \mathbf{X}, \mathbf{e}_d \rangle}{\|\mathbf{X}\|}\right|^2 \geq \varepsilon^2\right] \\
&= \mathbb{P}\left[\frac{X_d^2}{\sum_{j=1}^d X_j^2} \geq \varepsilon^2\right] \\
&= \mathbb{P}\left[X_d^2 \geq \varepsilon^2 \sum_{j=1}^d X_j^2\right] \\
&= \mathbb{P}\left[\sum_{j=1}^{d-1} (-\varepsilon^2 X_j^2) + (1 - \varepsilon^2) X_d^2 \geq 0\right].
\end{aligned}$$

We are now looking at a sum of independent (but not identically distributed) random variables

$$Z = \sum_{j=1}^{d-1} (-\varepsilon^2 X_j^2) + (1 - \varepsilon^2) X_d^2$$

and we can appeal to our usual Chebyshev machinery. The expectation is, by linearity,

$$\mathbb{E}[Z] = -\sum_{j=1}^{d-1} \varepsilon^2 \mathbb{E}[X_j^2] + (1 - \varepsilon^2)\mathbb{E}[X_d^2] = \{-(d-1)\,\varepsilon^2 + (1 - \varepsilon^2)\}$$

where we used that $X_1, \ldots, X_d$ are standard Normal variables and that, in particular, their mean is $0$ and their variance is $1$ so that $\mathbb{E}[X_1^2] = 1$.

The variance is, by independence of the $X_j$'s,

$$\begin{aligned}\mathrm{Var}[Z] &= \sum_{j=1}^{d-1} \varepsilon^4 \mathrm{Var}[X_j^2] + (1-\varepsilon^2)^2 \mathrm{Var}[X_d^2] \\ &= \{(d-1)\,\varepsilon^4 + (1-\varepsilon^2)^2\} \mathrm{Var}[X_1^2].\end{aligned}$$

So by Chebyshev

$$\begin{aligned}\mathbb{P}\,[Z \geq 0] &\leq \mathbb{P}\,[|Z - \mathbb{E}[Z]| \geq |\mathbb{E}[Z]|] \\ &\leq \frac{\mathrm{Var}[Z]}{\mathbb{E}[Z]^2} \\ &= \frac{\{(d-1)\,\varepsilon^4 + (1-\varepsilon^2)^2\} \mathrm{Var}[X_1^2]}{\{-(d-1)\,\varepsilon^2 + (1-\varepsilon^2)\}^2} \\ &\to 0\end{aligned}$$

as $d \to +\infty$. To get the limit we observed that, for large $d$, the deminator scales like $d^2$ while the numerator scales only like $d$.$\square$