Principal component analysis (blended lecture)

DS-GA 1013 / MATH-GA 2824 Mathematical Tools for Data Science

https://cims.nyu.edu/~cfgranda/pages/MTDS_spring20/index.html

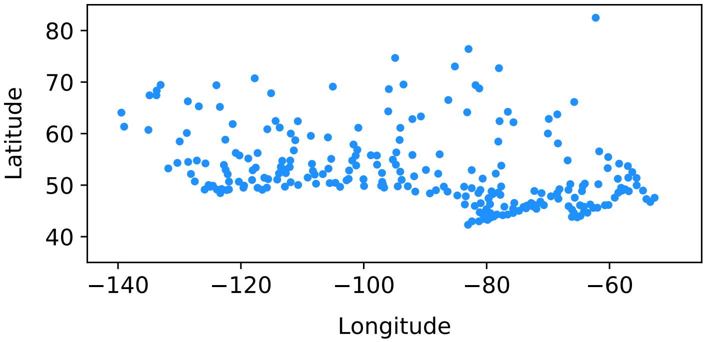
Carlos Fernandez-Granda

The spectral theorem

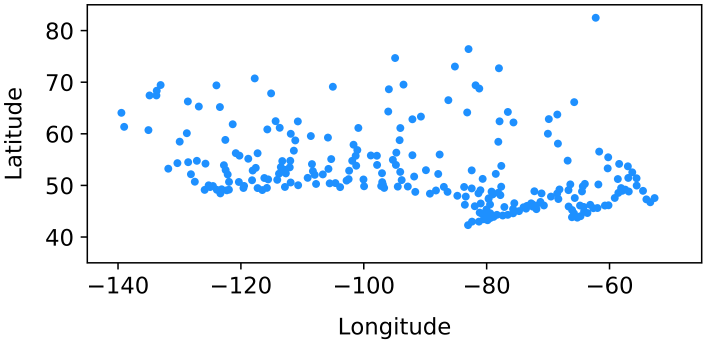
PCA of temperature data

Gaussian vectors in high dimensions

Cities in Canada



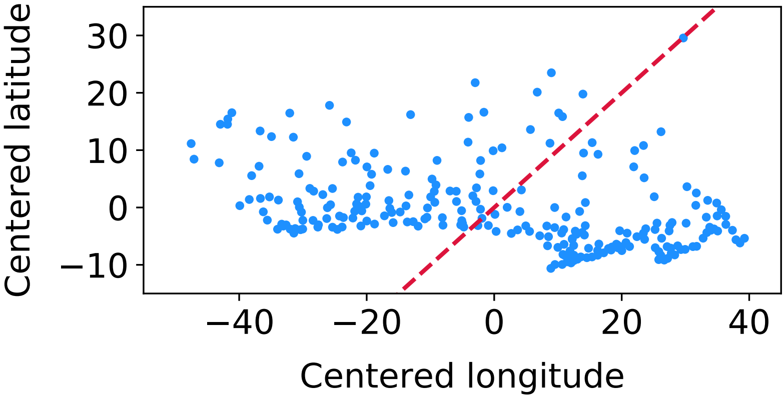
Cities in Canada



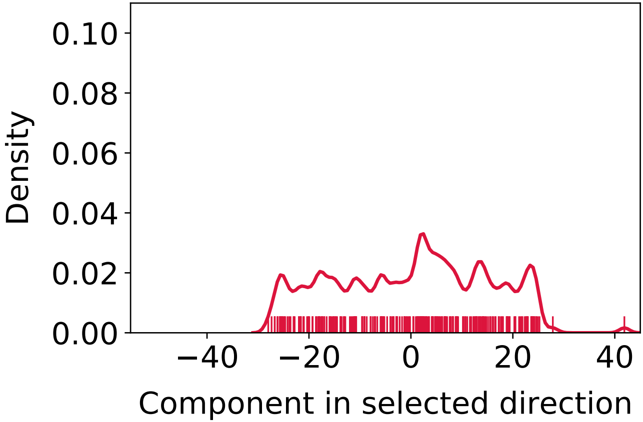
Sample covariance matrix:

$$\Sigma_X = \begin{bmatrix} 524.9 & -59.8 \\ -59.8 & 53.7 \end{bmatrix}$$

Projection onto a fixed direction



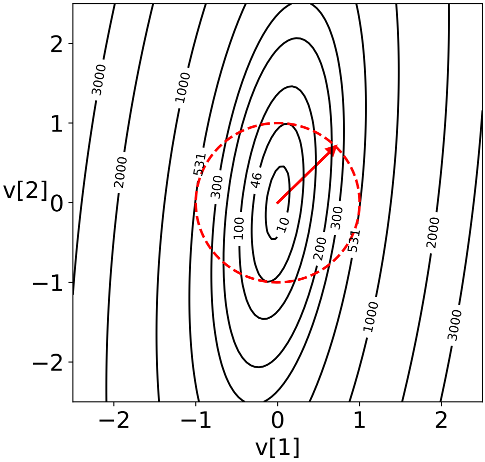
Projection onto a fixed direction



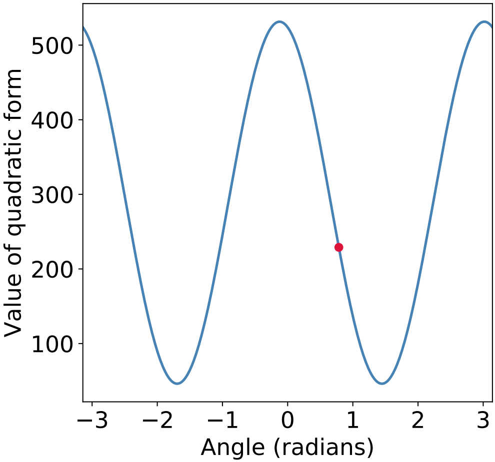
Projection onto a fixed direction

$$\sigma_{X_v}^2 = v^T \Sigma_X v$$

$f(v) := v^T \Sigma_X v$ for $||v||_2 = 1$



$f(v) := v^T \Sigma_X v$ for $||v||_2 = 1$



Spectral theorem

If $A \in \mathbb{R}^{d \times d}$ is symmetric, then it has an eigendecomposition

$$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}^T,$$

Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are real

Eigenvectors u_1, u_2, $\ldots$, u_n are real and orthogonal

Spectral theorem

$$\lambda_1 = \max_{||x||_2=1} x^T A x$$

$$u_1 = \arg\max_{||x||_2=1} x^T A x$$

$$\lambda_k = \max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x, \quad 2 \leq k \leq d$$

$$u_k = \arg\max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x, \quad 2 \leq k \leq d$$

Goal

Gain some mathematical intuition about spectral theorem

Quadratic form

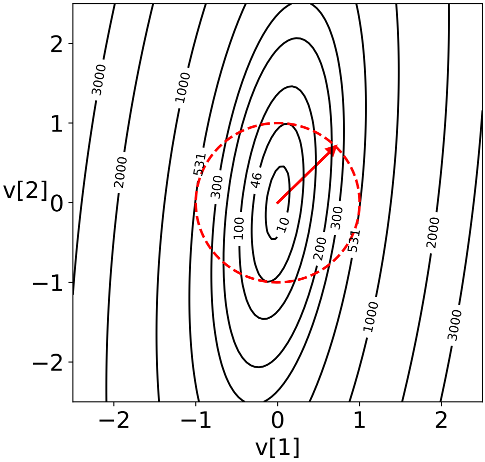
Function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(x) := x^T A x$$

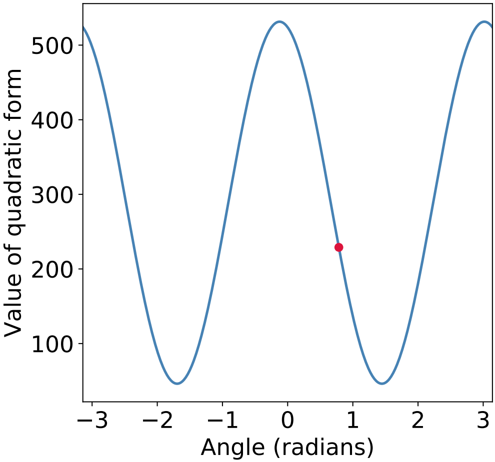
where A is a $d \times d$ symmetric matrix

Generalization of quadratic functions to multiple dimensions

$f(v) := v^T \Sigma_X v$

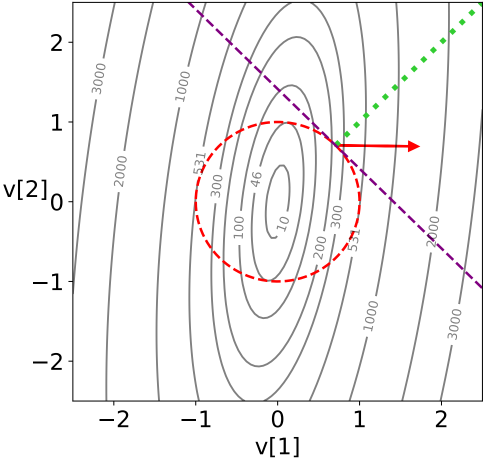


$f(v) := v^T \Sigma_X v$ for $||v||_2 = 1$



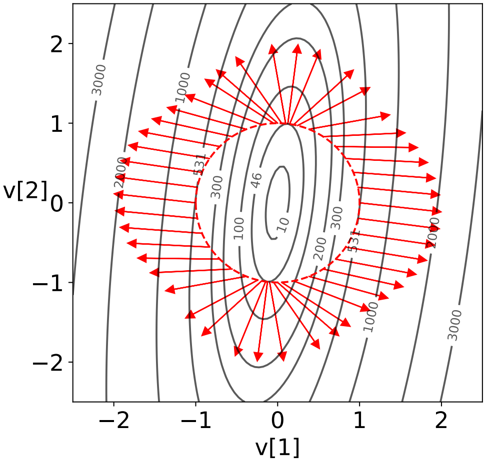
Can this point be a maximum?

Red arrow = gradient of quadratic form
Green line = gradient of $g(x) := x^T x$



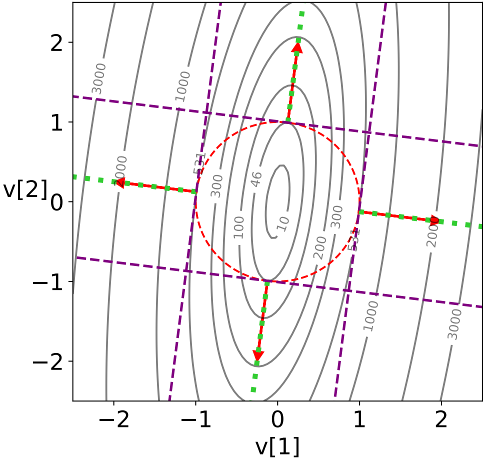
Where is the maximum?

Red arrow = gradient of quadratic form



Maximum satisfies $Ax = \lambda x$!

Red arrow = gradient of quadratic form
Green line = gradient of $g(x) := x^T x$



Some unresolved issues

Are we sure there is a maximum?

What about other local maxima?

What about minimum?

The spectral theorem

PCA of temperature data

Gaussian vectors in high dimensions

Dataset

- Hourly temperatures measured at US weather stations in 2015
- Number of features (stations): 134
- Number of examples: $24 \times 365 = 8760$

Sample covariance matrix

The sample covariance matrix of $X := \{x_1, x_2, \ldots, x_n\}$ is

$$\Sigma_X := \frac{1}{n} \sum_{i=1}^{n} c(x_i) c(x_i)^T$$

$$= \begin{bmatrix} \sigma^2_{X[1]} & \sigma_{X[1],X[2]} & \cdots & \sigma_{X[1],X[d]} \\ \sigma_{X[1],X[2]} & \sigma^2_{X[2]} & \cdots & \sigma_{X[2],X[d]} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X[1],X[d]} & \sigma_{X[2],X[d]} & \cdots & \sigma^2_{X[d]} \end{bmatrix}$$

$c(x_i) := x_i - \mu_X$
$X[j] := \{x_1[j], \ldots, x_n[j]\}$
$\sigma^2_{X[i]}$ is the sample variance of $X[i]$
$\sigma_{X[i],X[j]}$ is the sample covariance of $X[i]$ and $X[j]$

Sample covariance matrix

	Tucson, AZ	Hilo, HI	Durham, NC	Ithaca, NY
Tucson, AZ				
Hilo, HI				
Durham, NC				
Ithaca, NY				

What do you expect?

Sample covariance matrix

	Tucson, AZ	Hilo, HI	Durham, NC	Ithaca, NY
Tucson, AZ	78.6	14.7	54.8	65.0
Hilo, HI	14.7	8.4	9.5	11.8
Durham, NC	54.8	9.5	89.4	97.4
Ithaca, NY	65.0	11.8	97.4	137.3

Sample covariance matrix

	Tucson, AZ	Hilo, HI	Durham, NC	Ithaca, NY
Tucson, AZ	78.6	14.7	54.8	65.0
Hilo, HI	14.7	8.4	9.5	11.8
Durham, NC	54.8	9.5	89.4	97.4
Ithaca, NY	65.0	11.8	97.4	137.3

How can we normalize to evaluate whether correlation is high or not?

Correlation coefficient

Pearson correlation coefficient of $\tilde{x}$ and $\tilde{y}$

$$\rho_{\tilde{x},\tilde{y}} := \frac{\mathrm{Cov}\left(\tilde{x},\tilde{y}\right)}{\sigma_{\tilde{x}}\sigma_{\tilde{y}}}.$$

Covariance between $\tilde{x}/\sigma_{\tilde{x}}$ and $\tilde{y}/\sigma_{\tilde{y}}$

By the Cauchy-Schwarz inequality

$$\left|\mathrm{Cov}\left(\tilde{x},\tilde{y}\right)\right| \leq \sigma_{\tilde{x}}\sigma_{\tilde{y}} \qquad \text{and} \qquad -1 \leq \rho_{\tilde{x},\tilde{y}} \leq 1$$

Same holds for sample statistics

Sample correlation matrix

	Tucson, AZ	Hilo, HI	Durham, NC	Ithaca, NY
Tucson, AZ	1	0.57	0.65	0.63
Hilo, HI	0.57	1	0.35	0.35
Durham, NC	0.65	0.35	1	0.88
Ithaca, NY	0.63	0.35	0.88	1

Principal components

For dataset X containing $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$

1. Compute sample covariance matrix Σ_X
2. Eigendecomposition of Σ_X yields principal directions u_1, ..., u_d
3. Center the data and compute principal components

$$pc_i[j] := u_j^T c(x_i), \quad 1 \leq i \leq n, \; 1 \leq j \leq d,$$

where $c(x_i) := x_i - \mu_X$

Principal directions

1. Top 5 coefficients:
   Jamestown ND (0.12), Goodridge MN (0.12), Northgate ND (0.12)
   Bottom 5 coefficients:
   Mauna Loa HI (0.01), Hilo HI (0.01), Bodega CA (0.12)

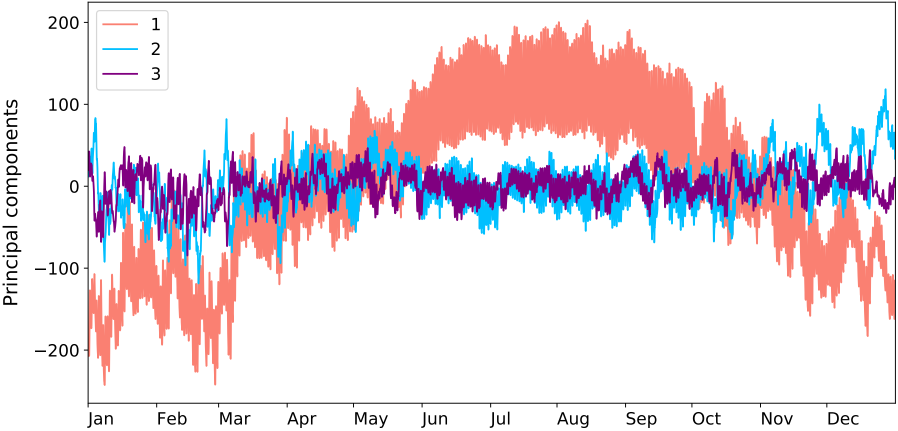
Principal directions

1. Top 5 coefficients:
   Jamestown ND (0.12), Goodridge MN (0.12), Northgate ND (0.12)
   Bottom 5 coefficients:
   Mauna Loa HI (0.01), Hilo HI (0.01), Bodega CA (0.12)
2. Top 5 coefficients:
   Elkins WV (0.13), Ithaca NY (0.12), Crossville TN (0.12)
   Bottom 5 coefficients:
   Merced CA (-0.14), Riley OR (-0.13), Redding CA (-0.13)

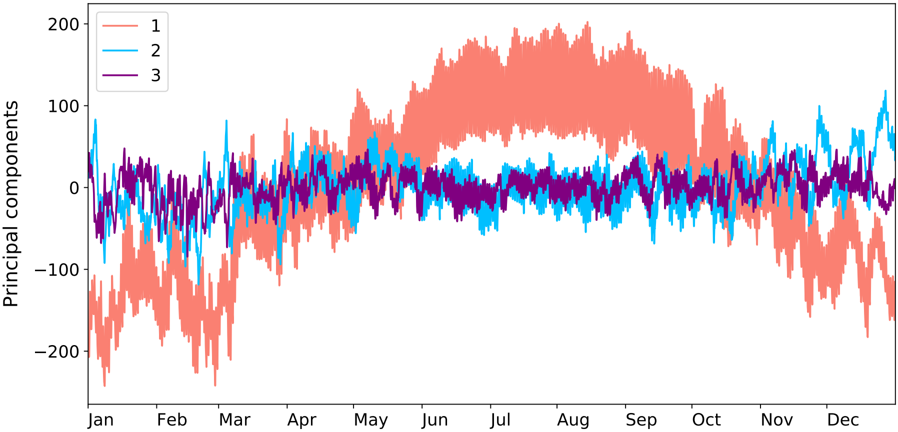
Principal directions

1. Top 5 coefficients:
   Jamestown ND (0.12), Goodridge MN (0.12), Northgate ND (0.12)
   Bottom 5 coefficients:
   Mauna Loa HI (0.01), Hilo HI (0.01), Bodega CA (0.12)
2. Top 5 coefficients:
   Elkins WV (0.13), Ithaca NY (0.12), Crossville TN (0.12)
   Bottom 5 coefficients:
   Merced CA (-0.14), Riley OR (-0.13), Redding CA (-0.13)
3. Top 5 coefficients:
   Goodridge MN (0.2), Jamestown ND (0.2), Aberdeen SD (0.18)
   Bottom 5 coefficients:
   Mauna Loa HI (0.01), Hilo HI (0.01), Bodega CA (0.12)

Principal directions



Principal directions



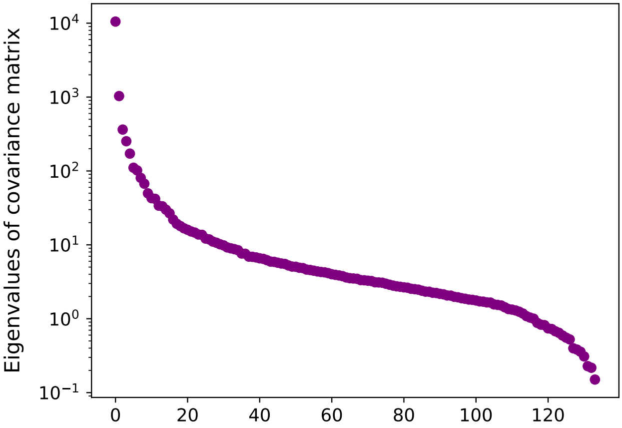
Limitations on interpretability?

Variance of principal components

How do we compute it?

Variance of principal components

How do we compute it?



The spectral theorem

PCA of temperature data

Gaussian vectors in high dimensions

Gaussian random vector

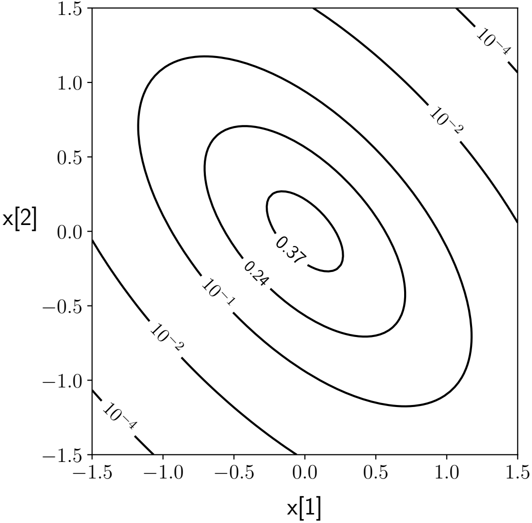
A Gaussian random vector $\tilde{x}$ is a random vector with joint pdf

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

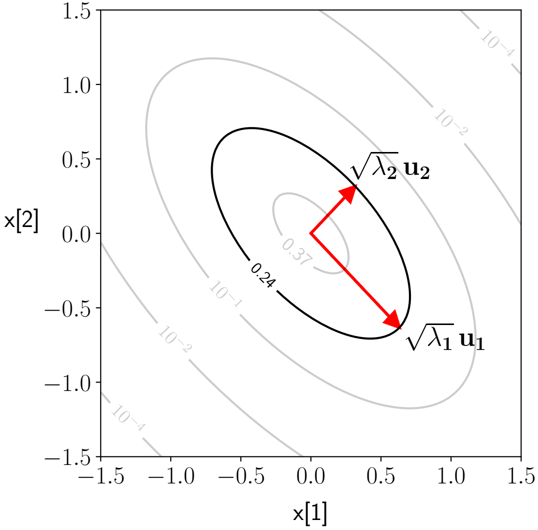
where $\mu \in \mathbb{R}^d$ is the mean and $\Sigma \in \mathbb{R}^{d \times d}$ the covariance matrix

$\Sigma \in \mathbb{R}^{d \times d}$ is positive definite (positive eigenvalues)

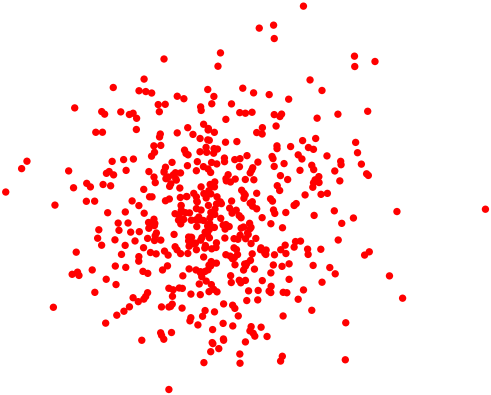
Contour surfaces



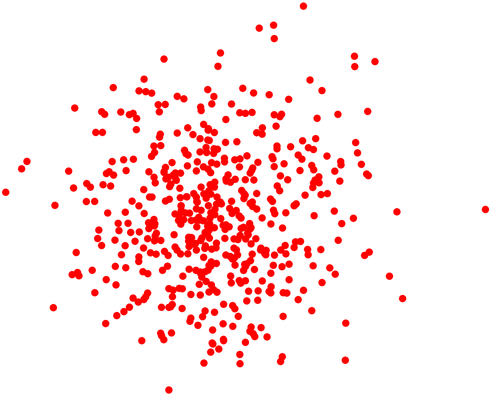
Contour surfaces



Gaussian (iid samples) in 2D



Gaussian (iid samples) in 2D



What do you expect in higher dimensions?

What statistic could be useful?

ℓ_2 norm of d-dimensional iid standard Gaussian vector $\tilde{w}$

$$\mathrm{Var}\left(||\tilde{w}||_2^2\right) = \mathrm{E}\left[\left(||\tilde{w}||_2^2\right)^2\right] - \mathrm{E}^2\left(||\tilde{w}||_2^2\right)$$

$$\mathrm{E}\left(||\tilde{w}||_2^2\right) =$$

$$\mathrm{E}\left[\left(||\tilde{w}||_2^2\right)^2\right] =$$

How does the std scale with respect to the mean?

ℓ_2 norm of d-dimensional iid standard Gaussian vector $\tilde{w}$

$$\mathrm{Var}\left(||\tilde{w}||_2^2\right) = \mathrm{E}\left[\left(||\tilde{w}||_2^2\right)^2\right] - \mathrm{E}^2\left(||\tilde{w}||_2^2\right) = 2d$$

$$\mathrm{E}\left(||\tilde{w}||_2^2\right) = \mathrm{E}\left(\sum_{i=1}^{d} \tilde{w}[i]^2\right) = d$$

$$\mathrm{E}\left[\left(||\tilde{w}||_2^2\right)^2\right] = \mathrm{E}\left[\left(\sum_{i=1}^{d} \tilde{w}[i]^2\right)^2\right]$$

$$= \sum_{i=1}^{d} \mathrm{E}\left(\tilde{w}[i]^4\right) + 2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d} \mathrm{E}\left(\tilde{w}[i]^2\right)\mathrm{E}\left(\tilde{w}[j]^2\right) = d(d+2)$$

How does the std scale with respect to the mean? $1/\sqrt{d}$

ℓ_2 norm of d-dimensional iid standard Gaussian vector