Def.- (Vector Space) A vector space over a field F is a set V together with two operations:

1) Addition: $+ : V \times V \to V$ (Takes two elements of V and returns another element of V)

2) Scalar Multiplication: $\cdot : F \times V \to V$

And must satisfy eight particular axioms (associativity of $+$, commutativity of $+$, etc.)

⊛ We refer to elements of V as vectors and elements of F as scalars.

Def.- (Linear Transformation) A linear transformation between two vector spaces V and W is a mapping $T: V \to W$ such that:

1) $T(v_1 + v_2) = T(v_1) + T(v_2)$, v_1 and $v_2 \in V$
2) $T(\alpha v) = \alpha T(v)$ for any scalar α.

We can represent linear transformations as matrices. Let

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ a_{21} & \cdots & a_{2k} \\ \vdots & \ddots & \vdots \\ a_{n1} & & a_{nk} \end{bmatrix}$$

then $A : \mathbb{R}^k \to \mathbb{R}^n$. The number of columns of a matrix encodes the number of basis vectors of the input space. In particular, every column shows where does the standard basis vectors move.

Example:

1) $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$

- $\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \overbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}^{\hat{i}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

- $\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \overbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}^{\hat{j}} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

Then, what linearity implies is that

$$A \begin{bmatrix} 5 \\ 3 \end{bmatrix} = A(5\hat{i} + 3\hat{j}) = 5 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 3 \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \cdot 1 + 3 \cdot 1 \\ 5 \cdot 1 + 3 \cdot 2 \end{bmatrix}$$

So matrix multiplication is performed by "projecting" the rows of a matrix onto every column.

2) $A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 0 \end{bmatrix}$

- $\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$
- $\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$

Then, linearity again implies

$$A \begin{bmatrix} 5 \\ 3 \end{bmatrix} = 5 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + 3 \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 5.1 + 3.1 \\ 5.1 + 3.2 \\ 5.1 + 3.0 \end{bmatrix}$$

We can conclude that in general, matrix with matrix multiplication consist on projections of the rows of some matrix A onto the columns of some other matrix B.

$$A = \begin{bmatrix} a_{1\,{1\times p}} \\ \vdots \\ a_{n\,{1\times p}} \end{bmatrix}_{n\times p} \qquad B = \begin{bmatrix} b_{1\,{p\times 1}} & \cdots & b_{k\,{p\times 1}} \end{bmatrix}_{p\times k}$$

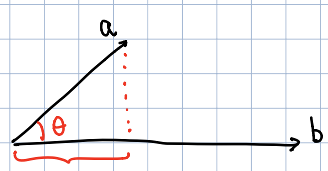
$$AB = \begin{bmatrix} a_1 \cdot b_1 & \cdots & a_1 \cdot b_k \\ \vdots & & \\ \underbrace{a_n} \cdot b_1 & \cdots & a_n \cdot b_k \end{bmatrix}_{n\times k}$$

for all these operations to be well defined we need that the row vectors of A have the same size as the colum vectors of B.

Dot Products and Projections:

The dot product between vectors a and b is computed as

$$a \cdot b = \|a\| \cos(\theta) \|b\|$$



This size is $\cos(\theta)\ \|a\|$

Notice that the dot product gives the size of the projection but scaled by the size of vector b. To get the value of this projection we normalize it by dividing the size of b.

Notice, however, that this yields a scalar. To make it point in the direction of b we can use b to create a unit vector in that direction. That is,

$$\text{proj}_b\, a = \|a\| \cos(\theta) \underbrace{\frac{b}{\|b\|}}_{\text{unit vector pointing in direction of } b}$$

$$= \underbrace{\frac{a \cdot b}{\|b\|^2}}_{} \; b$$

looks like univariate regression! We are computing the projection of a into the subspace created by the vector b.

(*) Notice that $\cos\theta = \dfrac{a \cdot b}{\|a\|\,\|b\|}$ so it doesn't depend on the scale of a and b.
this is actually a measure of correlation: $\cos(\theta) \in [-1, 1]$.

Vector Transformation Rules:

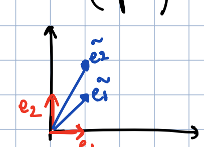
The linear mapping A is what we use to compute a new basis given by the columns of A from an old basis.

$$\begin{tikzcd} \text{OLD BASIS } \{e_1, e_2\} \arrow[d, bend right, "A"'] \\ \text{New Basis } \{\tilde{e}_1, \tilde{e}_2\} \arrow[u, bend right, "A^{-1}"'] \end{tikzcd}$$

"Covariant"

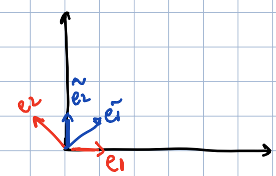
Example: • $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $e_2 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$ $A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$

$\tilde{e}_1 = A e_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $\tilde{e}_2 = A e_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$



- $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $\quad e_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ $\quad A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$

$$\tilde{e}_1 = A e_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \tilde{e}_2 = A e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$



OLD COMPONENTS in $\{e_1, e_2\}$

$A \uparrow \qquad \downarrow A^{-1}$

NEW COMPONENTS in $\{\tilde{e}_1, \tilde{e}_2\}$

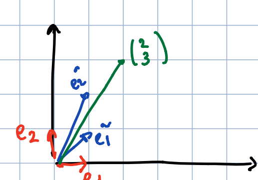
" Contra Variant "

Examples: - $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $\quad e_2 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$ $\quad A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$

$$\tilde{e}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \tilde{e}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Then, how do we express $\begin{pmatrix} 2 \\ 3 \end{pmatrix}$ in the $\{e_1, e_2\}$ basis into the new basis? We multiply the inverse of A, although it may seem unintuitive.

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$



So in the perspective of the new basis it looks like the vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

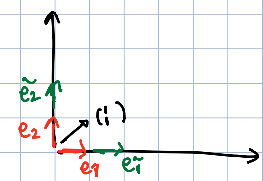
that is,

$$\begin{pmatrix} 2 \\ 3 \end{pmatrix} = 1\,\tilde{e}_1 + 1\,\tilde{e}_2 .$$

Does this make sense? Uhm, yes.

Consider two cases

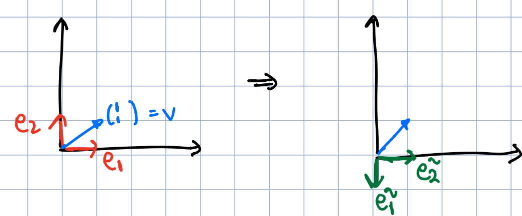
1) Pure Scale: $A = 2I = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$



The vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ in the perspective of the new basis feels half of what it feels in the old basis.

$$A^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$$ in the new basis.

2) Pure rotation: $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$



$$A^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

If the basis rotates clockwise then it feels as if v rotated counterclockwise in this new basis. That is why we say that vector components are contravariant and basis are covariant. (*) The opposite happens in a dual space!

Detour : Vector and Matrix Differentiation Rules

We start by considering a scalar function or scalar field that take vectors $x \in \mathbb{R}^n$ as input. Then, define

$$\frac{\partial f(x)_{1\times 1}}{\partial x_{n\times 1}} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

partial derivatives along a column!

As a special case we have $f(x) = a'x$. Therefore

$$\frac{\partial a'x}{\partial x} = \begin{bmatrix} \frac{\partial a'x}{\partial x_1} \\ \vdots \\ \frac{\partial a'x}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = a.$$

And recall that $a'x = x'a$ because it's a scalar, so that

$$\frac{\partial x'a}{\partial x} = \begin{bmatrix} \frac{\partial x'a}{\partial x_1} \\ \vdots \\ \frac{\partial x'a}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = a.$$

Similarly,

$$\frac{\partial f(x)_{1\times 1}}{\partial x'_{1\times n}} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \cdots & \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

$$= \left[\frac{\partial f(x)}{\partial x}\right]'$$

partial derivatives along a row!

Let A be a $m \times n$ matrix, $A = \begin{pmatrix} a_1' \\ \vdots \\ a_m' \end{pmatrix}$ where $a_j \in \mathbb{R}^n$ for $j=1,\dots,m$.

$$\frac{\partial \; Ax}{\partial x'} = \begin{pmatrix} \frac{\partial a_1'x}{\partial x'} \\ \vdots \\ \frac{\partial a_m'x}{\partial x'} \end{pmatrix} = \begin{pmatrix} a_1' \\ \vdots \\ a_m' \end{pmatrix} = A.$$

we know how to compute each one of these

$$\frac{\partial \; x'A'}{\partial x} = \begin{pmatrix} \frac{\partial x'a_1}{\partial x} & \cdots & \frac{\partial x'a_m}{\partial x} \end{pmatrix}$$
$$= \begin{pmatrix} a_1 & \cdots & a_m \end{pmatrix} = A'.$$

Multivariate chain rule: let $f(x)$ and $x(\alpha)$, $\alpha \in \mathbb{R}$

$$\underset{1\times1}{\frac{\partial f(x)}{\partial \alpha}} = \frac{\partial f(x)}{\partial x_1}\frac{\partial x_1}{\partial \alpha} + \dots + \frac{\partial f(x_n)}{\partial x_n}\frac{\partial x_n}{\alpha}$$

$$= \sum_{i=1}^{n} \underset{1\times1}{\frac{\partial f}{\partial x_i}} \underset{1\times1}{\frac{\partial x_i}{\partial \alpha}} \equiv \underbrace{\frac{\partial f}{\partial x_i}\frac{\partial x_i}{\partial \alpha}}$$

Some people use this notation where i denotes the dummy indices. It's also known as Einstein's summation.

Let $\alpha \in \mathbb{R}^r$ and $x = x(\alpha)$. Then

$$\frac{\partial x}{\partial \alpha'} = \begin{pmatrix} \frac{\partial x_1}{\partial \alpha'} \\ \vdots \\ \frac{\partial x_n}{\partial \alpha'} \end{pmatrix}_{n\times r}$$

Then,

- $$\frac{\partial f(x)_{1\times 1}}{\partial \alpha_{r\times 1}} = \begin{pmatrix} \frac{\partial f(x)}{\partial \alpha_1} \\ \vdots \\ \frac{\partial f(x)}{\partial \alpha_r} \end{pmatrix}$$

use multivariate chain rule

$$= \begin{pmatrix} \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial \alpha_1} \\ \vdots \\ \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial \alpha_r} \end{pmatrix}$$

$$= \begin{pmatrix} \left(\frac{\partial f}{\partial x}\right)' \frac{\partial x}{\partial \alpha_1} \\ \vdots \\ \left(\frac{\partial f}{\partial x}\right)' \frac{\partial x}{\partial \alpha_r} \end{pmatrix}$$

we can transpose scalars!

$$= \begin{pmatrix} \frac{\partial x'}{\partial \alpha_1} \frac{\partial f}{\partial x} \\ \vdots \\ \frac{\partial x'}{\partial \alpha_r} \frac{\partial f}{\partial x} \end{pmatrix} = \begin{pmatrix} \frac{\partial x'}{\partial \alpha_1} \\ \vdots \\ \frac{\partial x'}{\partial \alpha_r} \end{pmatrix} \frac{\partial f}{\partial x}$$

$$= \underbrace{\frac{\partial x'_{1\times n}}{\partial \alpha_{r\times 1}}}_{r\times n} \underbrace{\frac{\partial f_{1\times 1}}{\partial x_{n\times 1}}}_{n\times 1}.$$

- $$\frac{\partial f(x)_{1\times 1}}{\partial \alpha'_{1\times r}} = \begin{pmatrix} \frac{\partial f(x)}{\partial \alpha_1} & \dots & \frac{\partial f(x)}{\partial \alpha_r} \end{pmatrix}$$

use multivariate chain rule

$$= \begin{pmatrix} \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial \alpha_1} & \dots & \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial \alpha_r} \end{pmatrix}$$

$$= \begin{pmatrix} \left(\frac{\partial f}{\partial x}\right)' \frac{\partial x}{\partial \alpha_1} & \dots & \left(\frac{\partial f}{\partial x}\right)' \frac{\partial x}{\partial \alpha_r} \end{pmatrix}$$

$$= \left(\frac{\partial f}{\partial x}\right)' \begin{pmatrix} \frac{\partial x}{\partial \alpha_1} & \dots & \frac{\partial x}{\partial \alpha_r} \end{pmatrix}$$

$$= \underbrace{\frac{\partial f_{1\times 1}}{\partial x'_{1\times n}}}_{1\times n} \underbrace{\frac{\partial x_{n\times 1}}{\partial \alpha'_{1\times r}}}_{n\times r}$$

Special cases of the previous rules are

$$\frac{\partial\, a'x_{n\times 1}}{\partial \alpha_{r\times 1}} = \frac{\partial x'}{\partial \alpha}\frac{\partial\, a'x}{\partial x} = \frac{\partial x'}{\partial \alpha}\, a .$$

$$\frac{\partial\, a'x_{1\times 1}}{\partial \alpha'_{1\times r}} = \frac{\partial\, a'x}{\partial x'}\frac{\partial x}{\partial \alpha'} = a'\frac{\partial x}{\partial \alpha'} .$$

This allows us to generalize to the derivative of a vector

$$\frac{\partial\, Ax_{m\times 1}}{\partial \alpha'_{1\times r}} = \begin{pmatrix} \frac{\partial (a_1'x)}{\partial \alpha'} \\ \vdots \\ \frac{\partial (a_m'x)}{\partial \alpha'} \end{pmatrix} = \begin{pmatrix} a_1' \frac{\partial x}{\partial \alpha'} \\ \vdots \\ a_m' \frac{\partial x}{\partial \alpha'} \end{pmatrix}$$

$$= \begin{pmatrix} a_1' \\ \vdots \\ a_m' \end{pmatrix} \frac{\partial x}{\partial \alpha'} = \underset{m\times n}{A} \frac{\partial x}{\partial \alpha'}_{n\times r}$$

$$\frac{\partial\, x'A'_{1\times m}}{\partial \alpha_{r\times 1}} = \begin{pmatrix} \frac{\partial x'a_1}{\partial \alpha} & \cdots & \frac{\partial x' a_m}{\partial \alpha} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial x'}{\partial \alpha} a_1 & \cdots & \frac{\partial x'}{\partial \alpha} a_m \end{pmatrix}$$

$$= \frac{\partial x'}{\partial \alpha} \begin{pmatrix} a_1 & \cdots & a_m \end{pmatrix} = \frac{\partial x'}{\partial \alpha} A'$$

Notice that

$$\frac{\partial x}{\partial x'} = \begin{pmatrix} \frac{\partial x_1}{\partial x'} \\ \vdots \\ \frac{\partial x_n}{\partial x'} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & \vdots & & \\ 0 & 0 & \cdots & 1 \end{pmatrix} = I .$$

Let $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$ and A is $m \times n$. Then

- $$\frac{\partial (z'Ax)_{1\times 1}}{\partial x_{n\times 1}} = \frac{\partial (c'x)}{\partial x} \text{, where } c = A'z$$
$$= c = A'z.$$

- $$\frac{\partial (z'Ax)_{1\times 1}}{\partial z_{m\times 1}} = \frac{\partial (z'd)}{\partial z} \text{, where } d = Ax$$
$$= d = Ax.$$

$$\frac{\partial (z'Ax)}{\partial \alpha} = \begin{pmatrix} \frac{\partial z'Ax}{\partial \alpha_1} \\ \vdots \\ \frac{\partial z'Ax}{\partial \alpha_r} \end{pmatrix} = \begin{pmatrix} \frac{\partial z'Ax}{\partial x'} \frac{\partial x}{\partial \alpha_1} + \frac{\partial z'Ax}{\partial z'} \frac{\partial z}{\partial \alpha_1} \\ \vdots \\ \frac{\partial z'Ax}{\partial x'} \frac{\partial x}{\partial \alpha_r} + \frac{\partial z'Ax}{\partial z'} \frac{\partial z}{\partial \alpha_r} \end{pmatrix}$$

use the fact that we can transpose scalars.

$$= \begin{pmatrix} \frac{\partial x'}{\partial \alpha_1} \frac{\partial z'Ax}{\partial x} + \frac{\partial z'}{\partial \alpha_1} \frac{\partial z'Ax}{\partial z} \\ \vdots \\ \frac{\partial x'}{\partial \alpha_r} \frac{\partial z'Ax}{\partial x} + \frac{\partial z'}{\partial \alpha_r} \frac{\partial z'Ax}{\partial z} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial x'}{\partial \alpha_1} \frac{\partial z'Ax}{\partial x} \\ \vdots \\ \frac{\partial x'}{\partial \alpha_r} \frac{\partial z'Ax}{\partial x} \end{pmatrix} + \begin{pmatrix} \frac{\partial z'}{\partial \alpha_1} \frac{\partial z'Ax}{\partial z} \\ \vdots \\ \frac{\partial z'}{\partial \alpha_r} \frac{\partial z'Ax}{\partial z} \end{pmatrix}$$

$$= \frac{\partial x'}{\partial \alpha} A'z + \frac{\partial z'}{\partial \alpha} Ax.$$

As a special case we have

$$\frac{\partial x'Ax}{\partial x} = \frac{\partial x' \overbrace{Ax}^{c}}{\partial x} + \frac{\partial \overbrace{x'A}^{d'}x}{\partial x}$$

$$= \underbrace{Ax}_{c} + \underbrace{A'x}_{d} = (A + A')x$$

Notice that we can also differentiate a scalar along two dimensions

$$\frac{\partial z' A x_{1\times 1}}{\partial A_{m\times n}} = \begin{pmatrix} \frac{\partial z'Ax}{\partial A_{11}} & \cdots & \frac{\partial z'Ax}{\partial A_{1n}} \\ \vdots & & \\ \frac{\partial z'Ax}{\partial A_{m1}} & \cdots & \frac{\partial z'Ax}{\partial A_{mn}} \end{pmatrix}$$

$$= \begin{pmatrix} z_1 x_1 & \cdots & z_1 x_n \\ \vdots & & \vdots \\ z_m x_1 & \cdots & z_m x_n \end{pmatrix} = z \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}$$

$$= z\, x'.$$

Application to GMM criterion function:

Let $z \in \mathbb{R}^{\ell}$, $x \in \mathbb{R}^{k}$, $b \in \mathbb{R}^{k}$, $y \in \mathbb{R}$, and W is $\ell \times \ell$ and symmetric.
Then

$$\frac{\partial}{\partial b}\Big[(\overbrace{z(y - x'b)}^{c})' \underset{\text{symmetric}}{W} (z(y - x'b)) \Big] = \frac{\partial c'}{\partial b} W' c + \frac{\partial c'}{\partial b} W c$$

because of symmetry

$$= 2 \frac{\partial (z(y-x'b))'}{\partial b} W (z(y-x'b))$$

$$= 2 \frac{\partial}{\partial b}\Big[(zy - zx'b)'\Big] W (z(y-x'b))$$

$$= -2 \frac{\partial\, b'_{1\times k} x_{k\times 1} z'_{1\times \ell}}{\partial b} W (z(y-x'b))$$

$$= -2\, x z' \, W (z(y - x'b))$$

Now, the real criterion function uses $Z_{n\times \ell}$, $X_{n\times k}$, and W is $\ell \times \ell$.
Then,

$$\frac{\partial}{\partial b}\Big[(Z'(Y - Xb))' W (Z'(Y - Xb)) \Big] = -2 \frac{\partial\, b'X'Z}{\partial b} W (Z'(Y-Xb))$$

$$= -2\, X'Z \; W (Z'Y - Z'Xb)$$

Example: Non linear GMM

Consider the criterion function

$$Q_n(\theta) = \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n} \underbrace{g(W_i,\theta)'}_{1\times \ell}\right] \underbrace{A'A}_{\substack{\ell\times\ell \\ \text{Symmetric}}} \left[\frac{1}{n}\sum_{i=1}^{n} \underbrace{g(W_i,\theta)}_{\ell\times 1}\right]$$

$$\frac{\partial Q_n(\theta)_{1\times 1}}{\partial \theta_{k\times 1}} = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(W_i,\theta)'}{\partial\theta}\right] A'A \left[\frac{1}{n}\sum_{i=1}^{n} g(W_i,\theta)\right]$$

$$\frac{\partial^2 Q_n(\theta)}{\partial\theta'_{1\times k}\,\partial\theta_{k\times 1}} = d'\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(W_i,\theta)}{\partial\theta}\right]$$

$d' := \frac{1}{n}\sum \frac{\partial g'}{\partial\theta} A'A$

$c := A'A\left[\frac{1}{n}\sum g\right]$

$$+ \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta'}\left[\frac{\partial g(W_i,\theta)'}{\partial\theta}\right]\right] c$$

$$= \left[\underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(W_i,\theta)'}{\partial\theta}}_{k\times\ell}\right] \underbrace{A'A}_{\ell\times\ell} \left[\frac{1}{n}\sum_{i=1}^{n}\underbrace{\frac{\partial g(W_i,\theta)}{\partial\theta'}}_{\ell\times k}\right]$$

$$+ \underbrace{\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta'_{k\times k}}\left[\frac{\partial g(W_i,\theta)'}{\partial\theta_{k\times\ell}}\right]\right]}_{\frac{\partial}{\partial\theta'}\underbrace{\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(W_i,\theta)'}{\partial\theta}\right]_{k\times\ell}}_{M'_{k\times\ell}}} \underbrace{A'A\left[\frac{1}{n}\sum g(W_i,\theta)\right]_{\ell\times 1}}_{v_{\ell\times 1}}$$

We need a new operator!

(*) Property:

$M'v = [I_d \otimes v']\,\mathrm{vec}(M)$

where d is the number of columns of M.

$$= \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(W_i,\theta)'}{\partial\theta}\right] A'A \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial g(W_i,\theta)}{\partial\theta'}\right]$$

$$+ \frac{\partial}{\partial\theta'}\underbrace{\left(I_k \otimes \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} g(W_i,\theta)'\, A'A\right)}_{1\times\ell}\right)}_{k\times k\ell} \underbrace{\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}\left(\underbrace{\frac{\partial g(W_i,\theta)}{\partial\theta'}}_{\ell\times k}\right)\right]}_{k\ell\times 1}$$

only affects this argument

$$= \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial g(w_i,\theta)'}{\partial \theta} \right] A'A \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial g(w_i,\theta)}{\partial \theta'} \right]$$

$$+$$

$$\left[I_k \otimes \left(\frac{1}{n} \sum_{i=1}^{n} g(w_i,\theta)' A'A \right) \right] \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta'} \mathrm{vec}\left(\frac{\partial g(w_i,\theta)}{\partial \theta'} \right) \right]$$

and you will notice that now dimensions match, so that we get a $k \times k$ matrix.

To see why the vec operator property holds, consider:

$$M = \underbrace{\left[\underbrace{m_1}_{\ell \times 1} \; \cdots \; \underbrace{m_d}_{\ell \times 1} \right]}_{\ell \times d} \quad \text{and} \quad v \in \mathbb{R}^{\ell}$$

$$M'v = \begin{bmatrix} m_1' \\ \vdots \\ m_d' \end{bmatrix} v = \begin{bmatrix} m_1'v \\ \vdots \\ m_d'v \end{bmatrix}$$

$$= \begin{bmatrix} v' & 0 & \cdots & 0 \\ 0 & v' & \cdots & 0 \\ & \cdots & & \\ 0 & \cdots & 0 & v' \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_d \end{bmatrix}$$

$$= (I_d \otimes v') \, \mathrm{vec}(M)$$

where

$$\underbrace{A}_{m \times n} \otimes \underbrace{B}_{p \times q} = \underbrace{\begin{bmatrix} a_{11} B & \cdots & a_{1n} B \\ \vdots & \ddots & \vdots \\ a_{m1} B & \cdots & a_{mn} B \end{bmatrix}}_{mp \times nq} .$$

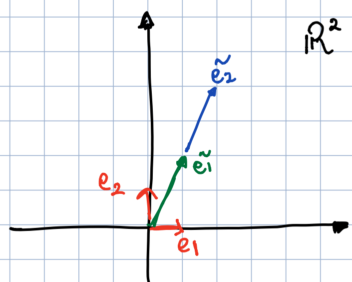
Linear System of Equations

$$\underbrace{A}_{l\times m}\ \underbrace{x}_{m\times 1} = \underbrace{b}_{l\times 1}$$

⊛ In the case of simultaneous equations $A = \Pi_2'$, $x = \gamma_1$, $b = \Pi_2'$.

Consider $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$. Since it has only one independent column so it has $\text{rank}(A) = 1$. Notice that $A : \underbrace{\mathbb{R}^2}_{\text{domain}} \to \underbrace{\mathbb{R}^2}_{\text{codomain}}$ (where it can come out)

Domain space



Finding the null space:

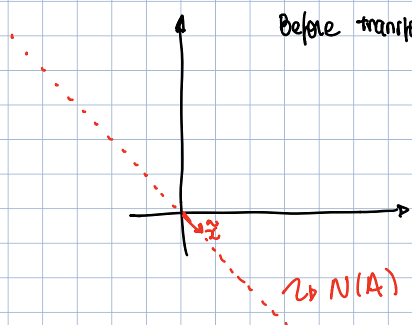
$$A\tilde{x} = 0$$

$$\tilde{x}_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \tilde{x}_2 \begin{bmatrix} 2 \\ 4 \end{bmatrix} = 0$$

$$\tilde{x} = \begin{bmatrix} 1 \\ -1/2 \end{bmatrix}$$

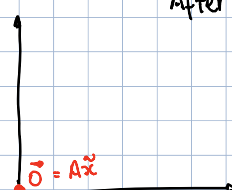
Notice that any other scaling of $\tilde{x}$ satisfies the homogeneous equation. In other words $N(A)$ is a subspace.

Before transformation



$\longrightarrow$

After transformation



The transformation squishes the $\mathbb{R}^2$ space into a range given by the rank of A.

Theorem.- (Rank Nullity) Let $A : V \to W$ be a linear transformation. Then

$$\text{Rank}(A) + \dim(N(A)) = \dim V$$

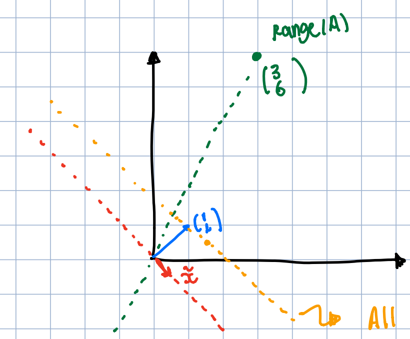
So in this case $\text{rank}(A) = 2 - \underbrace{1}_{}$.

$N(A)$ is a line, so it has dimension 1.

Therefore, the rank is a measure of how much space gets squished!

For example, the zero matrix maps every vector to 0, so it's something that squishes more. In that case rank = 0.

$$A\begin{bmatrix}1\\1\end{bmatrix} = 1\begin{bmatrix}1\\2\end{bmatrix} + 1\begin{bmatrix}2\\4\end{bmatrix} = \begin{bmatrix}3\\6\end{bmatrix}$$



All of these vectors are $\binom{1}{1} + \underbrace{\tilde{x}}_{\text{for } \tilde{x} \in N(A)}$ and all of them are mapped to $\binom{3}{6}$.

The problem of finding $Ax = \binom{3}{6}$ is that it is not a bijective mapping, i.e. from $\binom{3}{6}$ we don't have information about which vector in the orange line was the one that got transformed. That is essentially why inversion of A in the standard way it's not gonna work.

Def.- (Moore-Penrose inverse) The MP inverse of a matrix A (denoted by A^+) is a generalized inverse that satisfies

Reflexive Generalized Inv { Generalized Inv →
(1) $A A^+ A = A$
(2) $A^+ A A^+ = A^+$
(3) $A^+ A$ is symmetric
(4) $A A^+$ is symmetric

(*) The benefit of this is that it's unique!

Lemma.- A general solution to the homogenous system of linear equations $Ax = 0$ is

$$x = (I_m - A^+ A) q \quad \text{where } q \text{ is an arbitrary vector.}$$

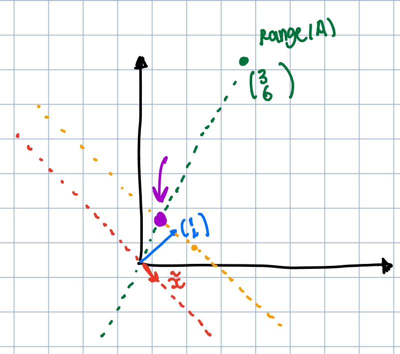
proof:

$$A(I_m - A^+A) q = (A - \underbrace{AA^+A}_{=A \text{ by } (1)}) q = 0.$$

In our example we got that the red line is the null space of A. Any generalized inverse of A should return a vector in the orange line.

$$A^+ = \begin{bmatrix} 0.04 & 0.08 \\ 0.08 & 0.16 \end{bmatrix}$$

And $A^+ b = \begin{bmatrix} 0.6 \\ 1.2 \end{bmatrix}$ which returns the following vector



And we know that any other $\tilde{x}$ in $N(A)$ that we sum still satisfies the equation.

<u>Lemma</u>.- $Ax = b$ has solution iff $\operatorname{rank}(A) = \operatorname{rank}([A \;\; b])$

<u>intuition of the proof:</u> Notice that $A = \left[\underset{\ell\times 1}{a_1} \;\; \underset{\ell\times 1}{a_2} \; \cdots \; \underset{\ell\times 1}{a_k} \right]_{\ell\times k}$, $x = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}_{k\times 1}$

Then $Ax = x_1 a_1 + x_2 a_2 + \dots + x_k a_k$.

$\operatorname{Rank}(A) = \operatorname{Rank}([A \;\; b])$ means that b is linearly dependent i.e. can be written as a linear combination of the colums of A which is literally the same as saying that $Ax = b$ has a solution.

<u>Lemma</u>.- $Ax = b$ has a solution iff $AA^+b = b$.

<u>Proof:</u> $(\Rightarrow)$ Suppose x is a solution of $Ax = b$. Then by MP inverse property

$$\underbrace{AA^+A}_{AA^+A \,=\, A} x = b$$

$$AA^+\underbrace{Ax}_{b} = b \;\Rightarrow\; AA^+b = b.$$

$(\Leftarrow)$ Suppose $AA^+b = b$. Set $\tilde{x} = A^+b$. Then $A\tilde{x} = AA^+b = b$ so that it is a solution.

<u>Lemma</u>.- If $Ax = b$ has a solution, then it takes the following form

$$x = A^+b + (I_m - A^+A)\,q$$ where q is an arbitrary vector.

Then notice that we can always get a solution in the least squares problem even with under identification. Now we will see two special cases of this general solution.

Special Case 1 : Exact Identification

$$\underset{l\times k}{A}\ \underset{k\times 1}{x} = \underset{k\times 1}{b}$$

- A has full column rank K
- $l = k$

$\Rightarrow$ A is invertible so $x = A^{-1} b$.

Special case 2 : Over identification

$$\underset{l\times k}{A}\ \underset{k\times 1}{x} = \underset{l\times 1}{b}$$

- A has full column rank K
- $l > k$

($Z'X\ \beta = Z'Y$ maybe now it looks more familiar)

Consider some positive definite and symmetric matrix W.

$$\underset{l\times l}{W^{1/2}}\ \underset{l\times k}{A}\ \underset{k\times 1}{x} = \underset{l\times l}{W^{1/2}}\ \underset{l\times 1}{b}$$

$$A'\ W^{1/2}\ W^{1/2}\ A\ x = A'\ W^{1/2}\ W^{1/2}\ b$$

$$\underbrace{\underset{k\times l}{A'}\ \underset{l\times l}{W}\ \underset{l\times k}{A}}_{\text{now this has full rank!}}\ x = A'\ W\ b$$

$$x = \underbrace{(A'WA)^{-1}\ A'W}_{A^*}\ b$$

We'll see what conditions we need so that the generalized inverse A^* is a Moore Penrose inverse.

(1) $A\ A^*\ A = A\ \underbrace{(A'WA)^{-1} A'WA}_{I} = A$ ✓

(2) $A^*\ A\ A^* = (A'WA)^{-1}\ \underbrace{A'WA\ (A'WA)^{-1}}_{I} A'W = A^*$ ✓

(3) To be symmetric $(A'WA)^{-1}\ A'WA = A'WA\ (A'WA)^{-1} = I$ ✓

$\Rightarrow A^+ A = A' A^+$

(4) To be symmetric $A(A'WA)^{-1} A'W = W'A\ (A'WA)^{-1} A'$ ✓

which is only satisfied if $W = I$.

Therefore, the MP inverse of A is

$$A^{+} = (A'A)^{-1}A' .$$

Useful things to know :

- If $X_{n \times k}$ has full column rank then so does $X'X$.
  To see this consider a $z \in N(X)$, that is, $Xz = 0$. Then notice that $X'Xz = 0$ as well. This implies $N(X) \subseteq N(X'X)$.

  Now take a $w \in N(X'X)$, that is, $X'Xw = 0$. Then we can see that it must also be the case that $w'X'Xw = 0$ so it follows that $Xw = 0$. This implies $N(X) \subseteq N(X'X)$

  Therefore $N(X) = N(X'X) \iff \text{rank}(X) = \text{rank}(X'X)$, by the rank nullity theorem.