Mathematical Tools
for Neural and Cognitive Science

Fall semester, 2022

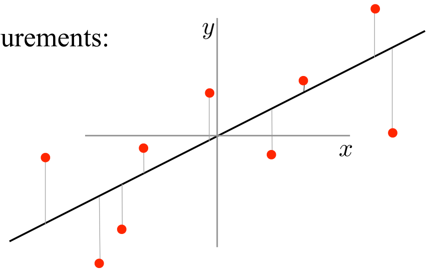
Section 2: Least Squares

Least squares regression:

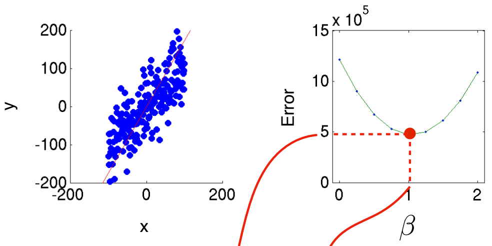
“objective” or “error" function

$$\min_{\beta} \boxed{\sum_n (y_n - \beta x_n)^2}$$

In the space of measurements:



[Gauss, 1795 - age 18!]



$$\hat{\beta} = \arg\min_{\beta} \sum_n (y_n - \beta x_n)^2$$

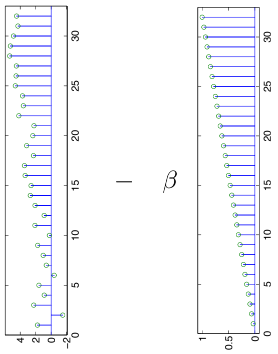
$$\min_{\beta} \sum_n (y_n - \beta x_n)^2$$

$$\min_\beta \sum_n (y_n - \beta x_n)^2$$

can solve this with calculus… *[on board]*

... or, with linear algebra!

$$\min_\beta ||\vec{y} - \beta\vec{x}||^2$$



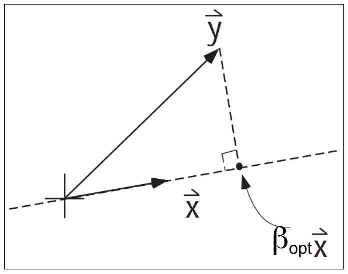
Observation $\vec{y}$

Regressor $\vec{x}$

$$\min_\beta ||\vec{y} - \beta\vec{x}||^2$$

Geometry:

Note: this is a 2-D cartoon of the N-D vectors, not the two-dimensional *(x,y)* measurement space of previous plots!



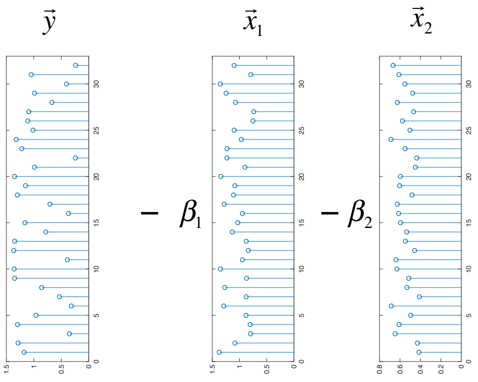
Note: partition of sum of squared data values:

$$||\vec{y}||^2 = ||\beta_{\text{opt}}\vec{x}||^2 + ||\vec{y} - \beta_{\text{opt}}\vec{x}||^2$$

Multiple regression:

$$\min_{\vec{\beta}} ||\vec{y} - \sum_k \beta_k \vec{x}_k||^2 = \min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2$$

2D example:



(observation)

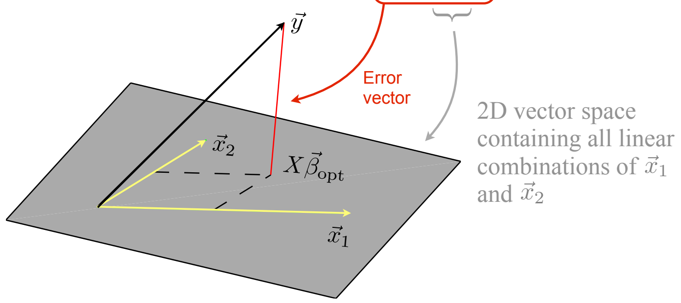
(regressor 1)

(regressor 2)

Solution via the “Orthogonality Principle”:

Construct matrix X, containing columns $\vec{x}_1$ and $\vec{x}_2$

Orthogonality: $X^T \boxed{\left(\vec{y} - X\vec{\beta}\right)} = \vec{0}$


$\vec{y}$
Error vector
$\vec{x}_2$
$X\vec{\beta}_{\text{opt}}$
$\vec{x}_1$
2D vector space containing all linear combinations of $\vec{x}_1$ and $\vec{x}_2$

Alternatively, can solve using SVD...

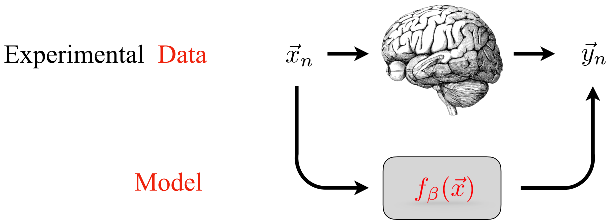
$$\begin{aligned}\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2 &= \min_{\vec{\beta}} ||\vec{y} - USV^T\vec{\beta}||^2 \\ &= \min_{\vec{\beta}} ||U^T\vec{y} - SV^T\vec{\beta}||^2 \\ &= \min_{\vec{\beta}^*} ||\vec{y}^* - S\vec{\beta}^*||^2 \\ \text{where} \quad \vec{y}^* &= U^T\vec{y}, \quad \vec{\beta}^* = V^T\vec{\beta}\end{aligned}$$

Solution: $\beta^*_{\text{opt},k} = y^*_k / s_k, \quad \text{for each } k$

or $\vec{\beta}^*_{\text{opt}} = S^{\#}\vec{y}^* \quad \Rightarrow \vec{\beta}_{\text{opt}} = VS^{\#}U^T\vec{y}$

[on board: transformations, elliptical geometry]

Fitting a parametric model (general)



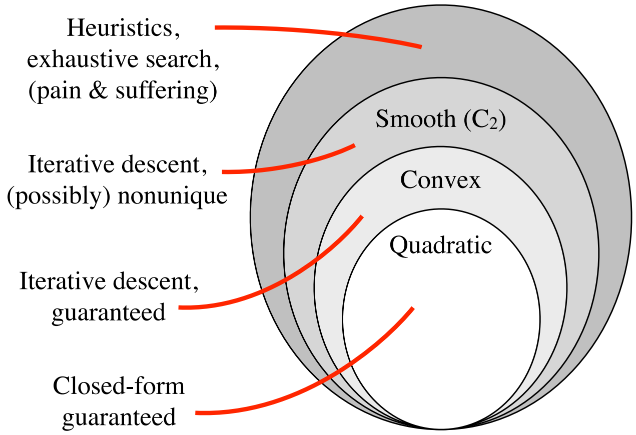
To fit model $f_\beta(\vec{x})$ to data $\{\vec{x}_n, \vec{y}_n\}$,

optimize parameters β to minimize an error function:

$$\min_\beta \sum_n E\left(\vec{y}_n, f_\beta(\vec{x}_n)\right)$$

Ingredients: data, model, error function, optimization method

Optimization



Heuristics, exhaustive search, (pain & suffering)

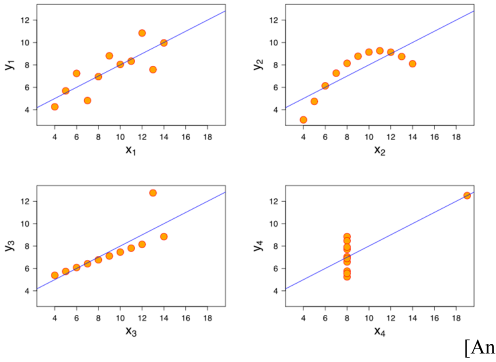
Iterative descent, (possibly) nonunique

Iterative descent, guaranteed

Closed-form guaranteed

Interpretation warning: fitting a line does not guarantee data actually lie along a line

These 4 data sets give the same regression fit, and same error:



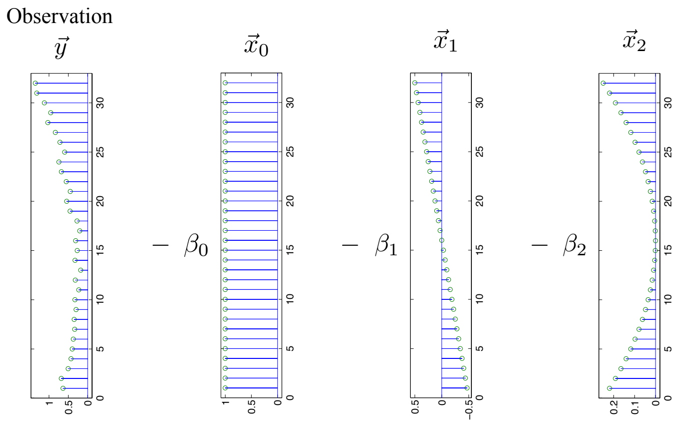
(need a better model)

(need a better error metric)

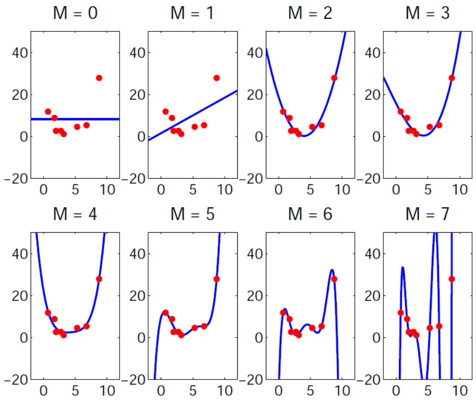
(need a better experiment)

[Anscombe, 1973]

Polynomial regression



Polynomial regression - how many terms?



(to be continued, when we get to "statistics"...)

Weighted Least Squares

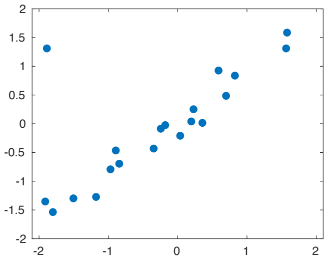
$$\min_{\beta} \sum_{n} \left[w_n (y_n - \beta x_n) \right]^2$$

$$= \min_{\beta} ||W(\vec{y} - \beta\vec{x})||^2$$

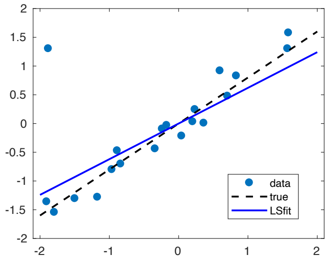
diagonal matrix

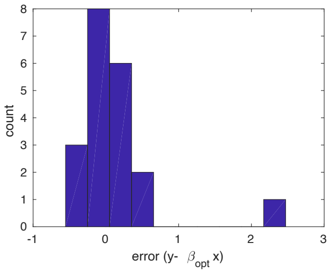
Solution via simple extensions of basic regression solution (i.e., let $\vec{y}^* = W\vec{y}$ and $\vec{x}^* = W\vec{x}$ and solve for β)

Outliers

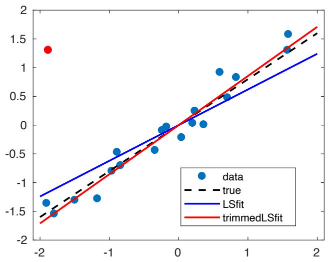


Outliers



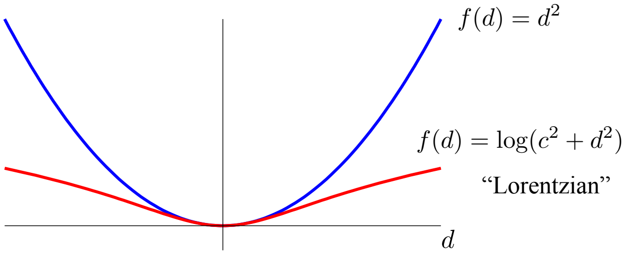


"Trimming"… discard points with large error.
Note: a special case of weighted least squares.



Trimming can be done iteratively (discard outlier, re-fit, repeat), a so-called "greedy" method. When do you stop?

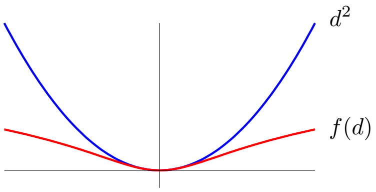
More generally, use a “robust” error metric.
For example:



Note: generally can’t obtain solution directly (i.e., requires an iterative optimization procedure).

In some cases, can use iteratively re-weighted least squares (IRLS)...

Iteratively Re-weighted Least Squares (IRLS)



initialize: $w_n^{(0)} = 1$

$$\beta^{(i)} = \arg\min_{\beta} \sum_n \omega_n^{(i)} \left(y_n - \beta x_n\right)^2$$

iterate

$$\omega_n^{(i+1)} = \left| \frac{f'(y_n - \beta^{(i)} x_n)}{y_n - \beta^{(i)} x_n} \right|$$

iterate

(one of many variants)

Constrained Least Squares

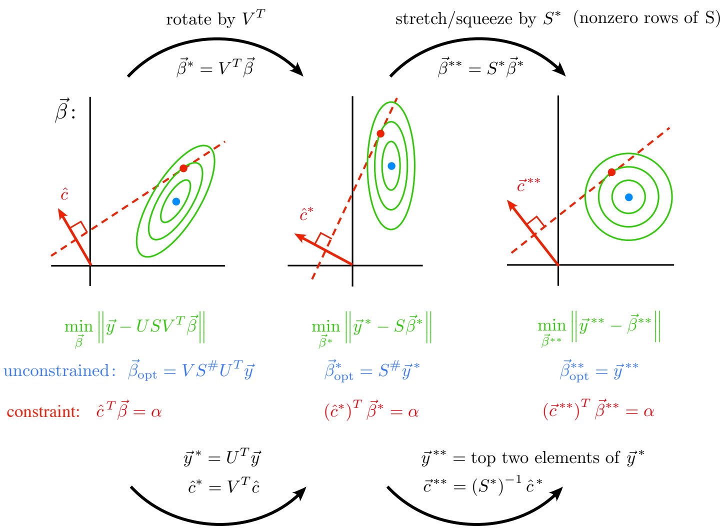
Linear constraint:

$$\arg\min_{\vec{\beta}} \left\| \vec{y} - X\vec{\beta} \right\|^2, \quad \text{where} \quad \vec{c}^T \vec{\beta} = 1$$

Quadratic constraint:

$$\arg\min_{\vec{\beta}} \left\| X\vec{\beta} \right\|^2, \quad \text{where} \quad \left\| \vec{\beta} \right\|^2 = 1$$

Can be solved exactly using linear algebra (SVD)...
[on board, with geometry]


rotate by V^T
stretch/squeeze by S^* (nonzero rows of S)
$\vec{\beta}^* = V^T\vec{\beta}$
$\vec{\beta}^{**} = S^*\vec{\beta}^*$
$\vec{\beta}$:
$\hat{c}$
$\hat{c}^*$
$\vec{c}^{**}$

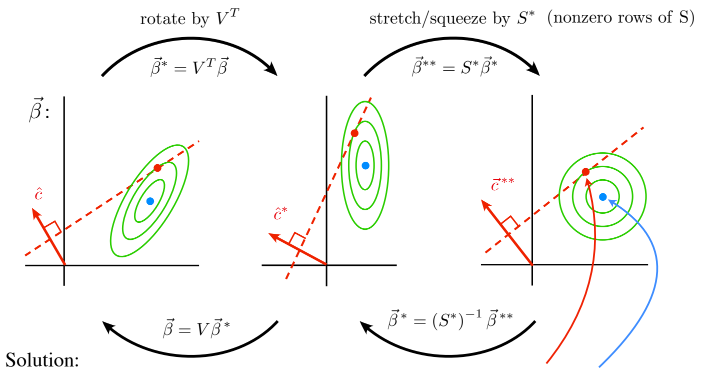
$\min_{\vec{\beta}} \left\|\vec{y} - USV^T\vec{\beta}\right\|$ $\min_{\vec{\beta}^*} \left\|\vec{y}^* - S\vec{\beta}^*\right\|$ $\min_{\vec{\beta}^{**}} \left\|\vec{y}^{**} - \vec{\beta}^{**}\right\|$

unconstrained: $\vec{\beta}_{\text{opt}} = VS^{\#}U^T\vec{y}$ $\vec{\beta}^*_{\text{opt}} = S^{\#}\vec{y}^*$ $\vec{\beta}^{**}_{\text{opt}} = \vec{y}^{**}$

constraint: $\hat{c}^T\vec{\beta} = \alpha$ $(\hat{c}^*)^T\,\vec{\beta}^* = \alpha$ $(\vec{c}^{**})^T\,\vec{\beta}^{**} = \alpha$

$\vec{y}^* = U^T\vec{y}$
$\hat{c}^* = V^T\hat{c}$

$\vec{y}^{**}$ = top two elements of $\vec{y}^*$
$\vec{c}^{**} = (S^*)^{-1}\,\hat{c}^*$


rotate by V^T
stretch/squeeze by S^* (nonzero rows of S)
$\vec{\beta}^* = V^T\vec{\beta}$
$\vec{\beta}^{**} = S^*\vec{\beta}^*$
$\vec{\beta}$:
$\hat{c}$
$\hat{c}^*$
$\vec{c}^{**}$
$\vec{\beta} = V\vec{\beta}^*$
$\vec{\beta}^* = (S^*)^{-1}\,\vec{\beta}^{**}$

Solution:

$\vec{\beta}_{\text{c,opt}} = V\,(S^*)^{-1}\,(\vec{y}^{\,**} + \gamma\vec{c}^{\,**})$

Write solution as: $\vec{\beta}^{**}_{\text{c,opt}} = \vec{y}^{\,**} + \gamma\vec{c}^{**}$

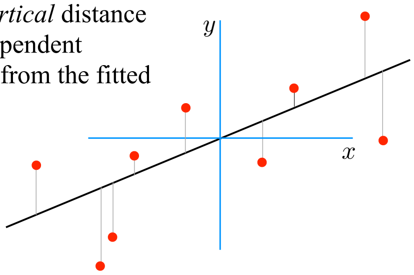
Solve for γ:

$$(\vec{c}^{\,**})^T\,\vec{\beta}^{**} = (\vec{c}^{\,**})^T\,(\vec{y}^{\,**} + \gamma\vec{c}^{\,**}) = \alpha$$

$$\Rightarrow \quad \gamma = \frac{\alpha - (\vec{c}^{\,**})^T\,\vec{y}^{\,**}}{(\vec{c}^{\,**})^T\,\vec{c}^{\,**}}$$

Standard Least Squares regression

Error is *vertical* distance (in the “dependent variable”) from the fitted line...

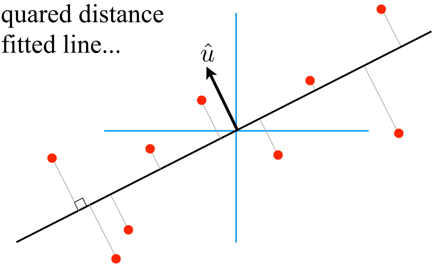


$$\arg\min_{\beta} ||\vec{y} - \beta\vec{x}||^2$$

Total Least Squares Regression

(a.k.a "orthogonal regression")

Error is squared distance from the fitted line...



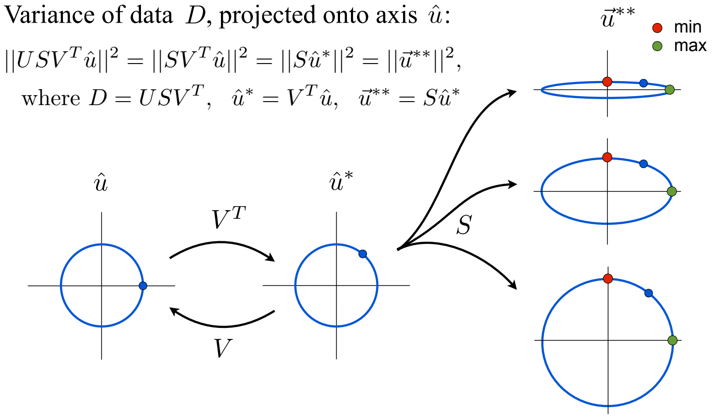
expressed as: $\min_{\hat{u}} ||D\hat{u}||^2, \quad \text{where } ||\hat{u}||^2 = 1$

Note: "data" matrix D now includes both x and y coordinates

Variance of data D, projected onto axis $\hat{u}$:

$$||USV^T\hat{u}||^2 = ||SV^T\hat{u}||^2 = ||S\hat{u}^*||^2 = ||\vec{u}^{**}||^2,$$

$$\text{where } D = USV^T, \quad \hat{u}^* = V^T\hat{u}, \quad \vec{u}^{**} = S\hat{u}^*$$

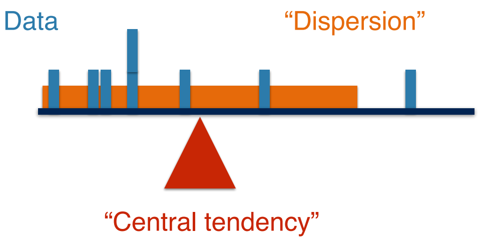


Set of $\hat{u}$'s of length 1 (i.e., unit vectors)

Set of $\hat{u}^*$'s of length 1 (i.e., unit vectors)

First two components of $\vec{u}^{**}$ (rest are zero!), for three example S's.

Descriptive statistics



Descriptive statistics: **Central tendency**

- We often summarize data with averages. Why?
- Average minimizes the squared error (as in regression!):

$$\bar{x} = \arg\min_c \frac{1}{N}\sum_{n=1}^{N}\left(x_n - c\right)^2 = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\vec{1}^T\vec{x}$$

- In general: minimize L_p norm: $\arg\min_c \left[\frac{1}{N}\sum_{n=1}^{N}|x_n - c|^p\right]^{1/p}$
  - $p = 1$: median, m_x
  - $p \to 0$: mode (location of maximum)
  - $p \to \infty$: midpoint of range
- Issues: outliers, asymmetry, bimodality

Descriptive statistics: **Dispersion**

- Sample variance (squared standard deviation):

$$s_x^2 = \min_c \frac{1}{N}\sum_{n=1}^{N}(x_n - c)^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

$$= \frac{1}{N}\sum_{n=1}^{N} x_n^2 - \bar{x}^2 = \frac{1}{N}\|\vec{x}\|^2 - \bar{x}^2$$

Mean absolute deviation (MAD) about the median:

$$d_x = \frac{1}{N}\sum_{n=1}^{N}\left|x_n - m_x\right|$$

- Quantiles (eg: “90% of data lie in range [1.5 8.2]”)

Descriptive statistics: Multi-D

Data points: $\left\{\vec{d}_n\right\}$ $\quad n \in [1 \ldots N],$ $\quad$ in 2-D: $\vec{d}_n = \begin{bmatrix} x_n \\ y_n \end{bmatrix}$

As in 1D: define central tendency as vector that minimizes the sum of squared distances to all data points:

$$\bar{d} \equiv \arg\min_{\vec{c}} \sum_n \|\vec{d}_n - \vec{c}\|^2$$

$$= \arg\min_{c_x, c_y} \sum_n (x_n - c_x)^2 + (y_n - c_y)^2 \qquad \text{(in 2-D)}$$

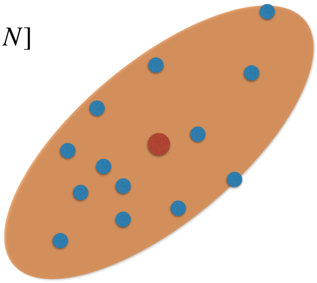
$$= \frac{1}{N}\sum_{n=1}^{N}\vec{d}_n = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}$$

Descriptive statistics: Multi-D

Data points: $\left\{ \vec{d}_n \right\} \quad n \in [1 \ldots N]$

Sample mean (average):

$$\bar{d} = \frac{1}{N} \sum_{n=1}^{N} \vec{d}_n = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}$$



Sample covariance:

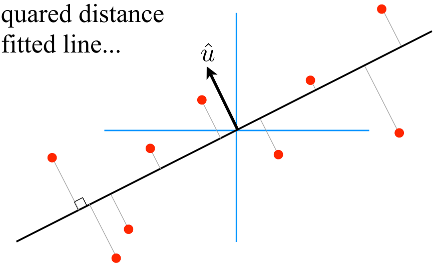
$$C_d = \frac{1}{N} \sum_{n=1}^{N} (\vec{d}_n - \bar{d})(\vec{d}_n - \bar{d})^T = \frac{1}{N} \sum_{n=1}^{N} \vec{d}_n \vec{d}_n^T - \bar{d}\bar{d}^T$$

$$= \frac{1}{N} \begin{bmatrix} ||\vec{x}||^2 & \vec{x}^T \vec{y} \\ \vec{y}^T \vec{x} & ||\vec{y}||^2 \end{bmatrix} - \begin{bmatrix} \bar{x}^2 & \bar{x}\bar{y} \\ \bar{y}\bar{x} & \bar{y}^2 \end{bmatrix} = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix}$$

Total Least Squares Regression

(a.k.a "orthogonal regression")

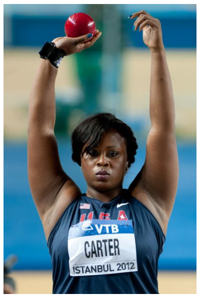
Error is squared distance from the fitted line...



expressed as: $\min_{\hat{u}} ||D\hat{u}||^2, \quad \text{where } ||\hat{u}||^2 = 1$

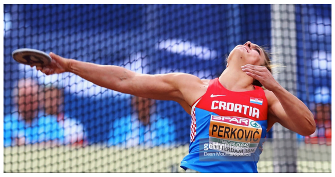
Note: "data" matrix *D* now includes both *x* and *y* coordinates

Olympic gold medalists (Rio, 2016)



Michelle Carter (USA)

Thomas Röhler (Germany)



Sandra Perković (Croatia)

3D geometry:
Javelin, Discus, Shotput…

Principal Component Analysis (PCA)

The shape of a data cloud can be summarized with an ellipse (ellipsoid), centered around the mean, using a simple procedure:

(1) Subtract mean of all data points, to re-center around origin

(2) Assemble centered data vectors in rows of a matrix, D

(3) Compute the SVD:

$$D = USV^T$$

or just use the smaller matrix

$$C = D^T D = VS^T SV^T$$
$$= V\Lambda V^T$$

(4) Columns of V are the *principal components* (axes) of the ellipsoid, diagonal elements s_k or $\sqrt{\lambda_k}$ are the corresponding principle radii, and their product is the volume.

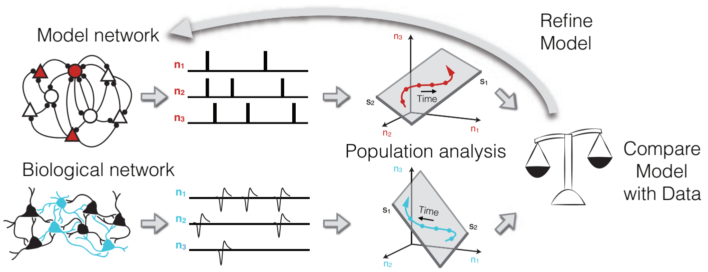
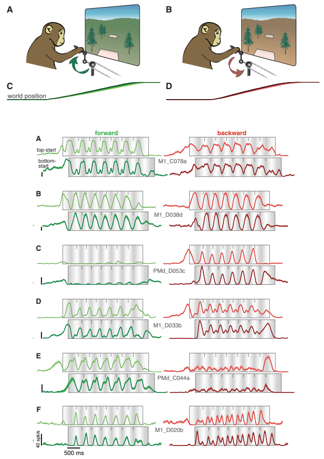
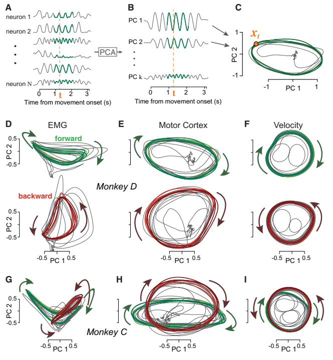


Fig 1. Relating biological and model networks using population analyses: Because a model network typically does not attempt to replicate the precise anatomical connectivity of a biological network, there is not a one-to-one correspondence of each biological neuron with a model neuron. Dimensionality reduction can be used to obtain a concise summary of the population activity from each network. This provides common ground for incisive comparisons between biological and model networks. Discrepancies in the population activity structure between biological and model networks can then help to refine model networks.

[Williamson, Doiron, Smith, Yu 2019]

Example: PCA for dimensionality reduction and visualization





[Russo et. al., 2018]

Eigenvectors/eigenvalues

- An *eigenvector* of a matrix is a vector that is rescaled by the matrix (i.e., the direction is unchanged)
- The corresponding scale factor is called the *eigenvalue*

- For matrix $C = D^T D = V \Lambda V^T$ the columns of V (denoted $\hat{v}_k$) are eigenvectors, with corresponding eigenvalues λ_k :

$$\begin{aligned} C\hat{v}_k &= V \Lambda V^T \hat{v}_k \\ &= V \Lambda \hat{e}_k \\ &= \lambda_k \hat{v}_k \end{aligned}$$