Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Instructor: Dr. Brandon Corbett

Lecture 6 for ChE 765 | Sep 767, McMaster University

Moving to two data blocks!

- PCA
  - Summarized many variables in terms of best few latent variables
  - Good for visualizing a lot of historical data
  - Good for process monitoring
- What if we want to make predictions?

Scenario: You are a banker

- You are a banker
- Your job is to give loans to first time home buyers
- How will you decide what mortgages to approve?

Discussion of X and Y blocks

- On the board

Review: Covariance

	Cylinder temperature (K)	Cylinder pressure (kPa)	Room humidity (%)
	273	1600	42
	285	1670	48
	297	1730	45
	309	1830	49
	321	1880	41
	333	1920	46
	345	2000	48
	357	2100	48
	369	2170	45
	381	2200	49
Mean	327	1910	46.1
Variance	1188	38940	7.3

Review: Covariance

Formal definition for covariance

$$\text{Cov}\{x, y\} = \mathcal{E}\{(x - \overline{x})(y - \overline{y})\} \qquad \text{where} \qquad \mathcal{E}\{z\} = \overline{z}$$

- Covariance with itself = variance:
  $\text{Cov}\{x, x\} = \mathcal{V}(x) = \mathcal{E}\{(x - \overline{x})(x - \overline{x})\}$
- (Co)variance of centered vector = (co)variance of uncentered vector
- Covariance describes overall tendency of 2 variables

Review: Covariance

Formal definition for covariance

$$\text{Cov}\{x, y\} = \mathcal{E}\{(x - \overline{x})(y - \overline{y})\} \qquad \text{where} \qquad \mathcal{E}\{z\} = \overline{z}$$

Covariance matrix for example:

- variances are on the diagonal
- covariances on the off-diagonals (symmetric matrix!)

$$\text{Covariance} = \begin{bmatrix} & \text{Temperature} & \text{Pressure} & \text{Humidity} \\ \text{Temperature} & 1188 & 6780 & 35.4 \\ \text{Pressure} & 6780 & 38940 & 202 \\ \text{Humidity} & 35.4 & 202 & 7.3 \end{bmatrix}$$

Review: Correlation

- (Co)variance depends on units: e.g. different covariance for grams vs kilograms
- Correlation removes the scaling effect:

Formal definition for correlation

$$r(x,y) = \frac{\mathcal{E}\left\{(x-\overline{x})(y-\overline{y})\right\}}{\sqrt{\mathcal{V}\left\{x\right\}\mathcal{V}\left\{y\right\}}} = \frac{\text{Cov}\left\{x,y\right\}}{\sqrt{\mathcal{V}\left\{x\right\}\mathcal{V}\left\{y\right\}}}$$

- Divides by the units of x and y: dimensionless result
- $-1 \leq r(x,y) \leq 1$

$$\text{Correlation} = \begin{array}{c} \\ \text{Temperature} \\ \text{Pressure} \\ \text{Humidity} \end{array} \begin{array}{ccc} \text{Temperature} & \text{Pressure} & \text{Humidity} \\ 1.0 & 0.997 & 0.380 \\ 0.997 & 1.0 & 0.379 \\ 0.380 & 0.379 & 1.0 \end{array}$$

Review: Least squares

We have 2 vectors of data, **x** and **y**. Presume the relationship between them:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

ϵ term:

- unmodelled components of the linear model
- measurement error
- other random variation

Important: error is from y, not from x.

We want parameter estimates:

- $b_0 = \hat{\beta}_0$
- $b_1 = \hat{\beta}_1$
- $e = \hat{\epsilon}$
-

Review: Least squares

To make derivations easier here, we will center both $\mathbf{x}$ and $\mathbf{y}$.

Least squares model is: $\mathbf{y} = \beta_1 \mathbf{x} + \epsilon$

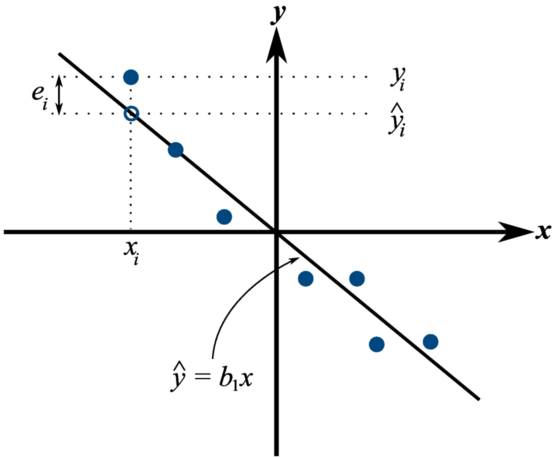
We can always recover the intercept, if we need it:

- $b_0 = \overline{\mathbf{y}} - b_1 \overline{\mathbf{x}}$

We want predictions from our model:

- For a new x-observation: x_{new}
- prediction is $= \hat{y}_{\text{new}} = b_1 x_{\text{new}}$

Review: Least squares



Review: solving the least squares model

Has to be an optimization problem: **minimizing** the sum of squared errors

- Easy to solve! Unconstrained optimization problem

$$\min f(b_1) = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - b_1 x_i)^2$$

$$\begin{aligned} \frac{\partial f(b_1)}{\partial b_1} &= -2 \textstyle\sum_i^n (x_i)(y_i - b_1 x_i) = 0 \\ b_1 &= \frac{\sum_i (x_i y_i)}{\sum_i (x_i)^2} = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{x}'\mathbf{x}} \end{aligned}$$

Remarks

1. $\sum_i e_i = 0$
2. Easily prove that $\sum_i (x_i e_i) = \mathbf{x}^T \mathbf{e} = 0$
   - The residuals are uncorrelated with the input variables, $\mathbf{x}$
   - There is no information in the residuals that is in the $\mathbf{x}$'s
3. Prove and interpret that $\sum_i (\hat{y}_i e_i) = \hat{\mathbf{y}}^T \mathbf{e} = 0$
   - The fitted values are uncorrelated with the residuals

Notation for MLR

The general linear model for observation i

$$
\begin{aligned}
y_i &= \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \epsilon_i \\
y_i &= [x_1, x_2, \ldots, x_K] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \epsilon_i \\
y_i &= \underbrace{x^T}_{(1\times K)} \underbrace{\beta}_{(K\times 1)} + \epsilon_i
\end{aligned}
$$

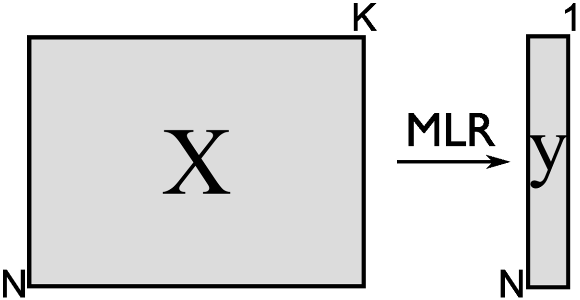
- where each x_k column (variable) and the y column have been centered

Notation for MLR

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,K} \\ x_{2,1} & x_{2,2} & \dots & x_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,K} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

- $\mathbf{y}$: $N \times 1$
- $\mathbf{X}$: $N \times k$
- $\mathbf{b}$: $K \times 1$
- $\mathbf{e}$: $N \times 1$



Estimating the model parameters via optimization

Objective function: minimize sum of squares of the errors

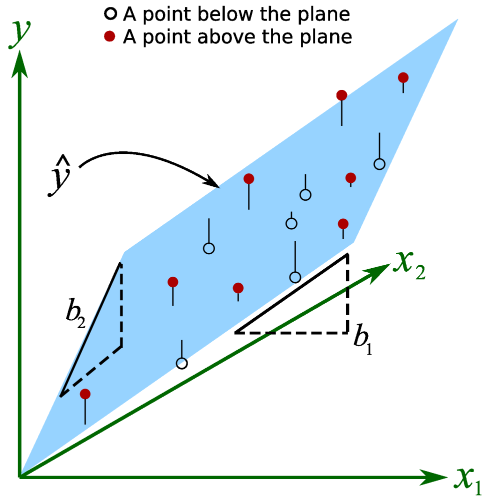
$$
\begin{aligned}
f(\mathbf{b}) &= \mathbf{e}^T\mathbf{e} \\
&= (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{b} + \mathbf{b}\mathbf{X}^T\mathbf{X}\mathbf{b}
\end{aligned}
$$

- Solving $\dfrac{f(\mathbf{b})}{\partial \mathbf{b}} = 0$ gives $\boxed{\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}}$
- $\mathcal{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1} S_E^2$
- $S_E = \sqrt{\dfrac{\mathbf{e}'\mathbf{e}}{N-K}} \approx$ standard deviation of the residuals

Interpretation of the model coefficients

The coefficients have meaning

$$y = b_1 x_1 + b_2 x_2$$

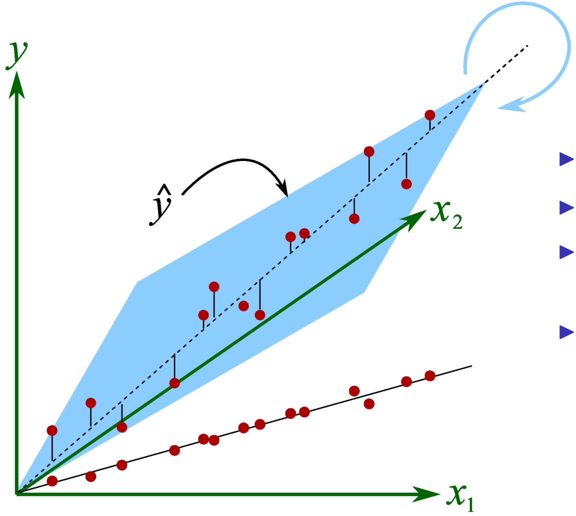


Least squares: What can go wrong?

1. Missing values
   - $\hat{y}_{\text{new}} = b_1 x_{1,\text{new}} + b_2 x_{2,\text{new}} + \ldots + b_K x_{K,\text{new}}$
   - There is nothing we can do if any $x_{k,\text{new}}$ terms go missing

Least squares: What can go wrong?

2. Highly correlated variables in **X**



- $\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$
- $\mathcal{V}(\mathbf{b}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1} S_E^2$
- Inflated confidence intervals for **b**
- Cannot interpret coefficients reliably

Leads to unstable regression coefficients. *Example on your own*.

Least squares: What can go wrong?

3. Noisy **x**-variables

   - LS model is: $\mathbf{y} = \beta_1 \mathbf{x} + \epsilon$
   - Note that model assumes error in **y**.
   - We say, "LS has a model for error" in the **y**'s.
   - Or alternatively, "model for error in the **y**-space". This means:
     - We can always compare our y error to S_E
     - see if error is large; then try to find out why
   - LS assumes that **x** is exact (no model for **x**-space error)

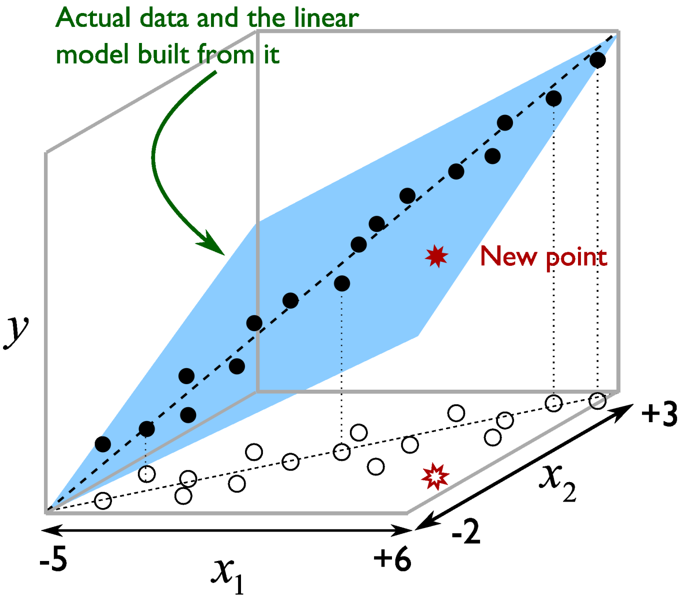
Least squares: What can go wrong?

4. Non-sensical input (related to previous point)

    - Extreme noise in **x**'s, or garbage input
    - Will go undetected, and you will always get a prediction:
    - $\hat{y}_{\text{new}} = b_1 x_{1,\text{new}} + b_2 x_{2,\text{new}} + \ldots + b_K x_{K,\text{new}}$
    - There is no **x**-space error model to catch these problems

Least squares: What can go wrong?

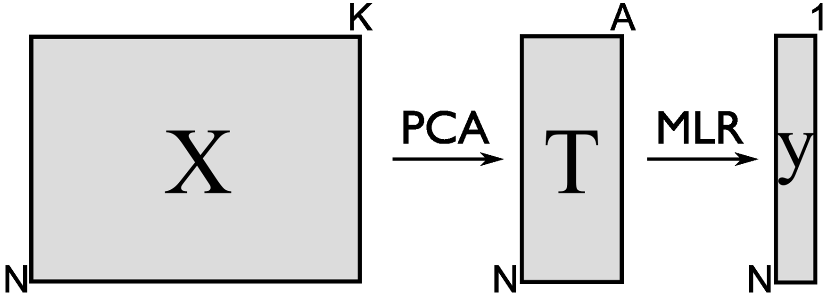
Misleading strategy that's often-used by people:



Other problems with linear regression

- MLR requires $N > K$. Problem with spectral data, and other data sets.
- If you have multiple $\mathbf{Y}$ variables: one MLR model per column in $\mathbf{Y}$

Principal component regression (PCR)



Two step model:

1. $\mathbf{T} = \mathbf{XP} + \mathbf{E}$ ordinary PCA
2. $\widehat{\mathbf{y}} = \mathbf{Tb}$ and can be solved as $\mathbf{b} = \left(\mathbf{T}'\mathbf{T}\right)^{-1}\mathbf{T}'\mathbf{y}$
   Regress the **y** onto the scores **T** to get regression coefficients **b**

Principal component regression (PCR)

Advantages:

- $\mathbf{T}$ is orthogonal: $(\mathbf{T}'\mathbf{T})^{-1}$ easily calculated
- so less need for variable selection to get a full rank $\mathbf{X}$
- PCA step handles missing values
- $\mathbf{T}$ has much less error than $\mathbf{X}$
- **Best part**: a free consistency check from T^2 and SPE
- PCA step uses fewer variables ($A < K$), we will likely meet the $N > K$ requirement in the regression step

Important point

If PCA step uses $A = K$, then predictions from PCR are same as MLR

Principal component regression (PCR)

Using a PCR model on new data

1. Center and scale the raw data as usual for PCA: $\mathbf{x}'_{\text{new}}$
2. Calculate the new scores: $\mathbf{t}'_{\text{new}} = \mathbf{x}'_{\text{new}}\mathbf{P}$
3. Consistency check: are SPE_{new} and T^2_{new} below the limits?
4. Use the MLR prediction: $\hat{y}_{\text{new}} = \mathbf{t}'_{\text{new}}\mathbf{b}$

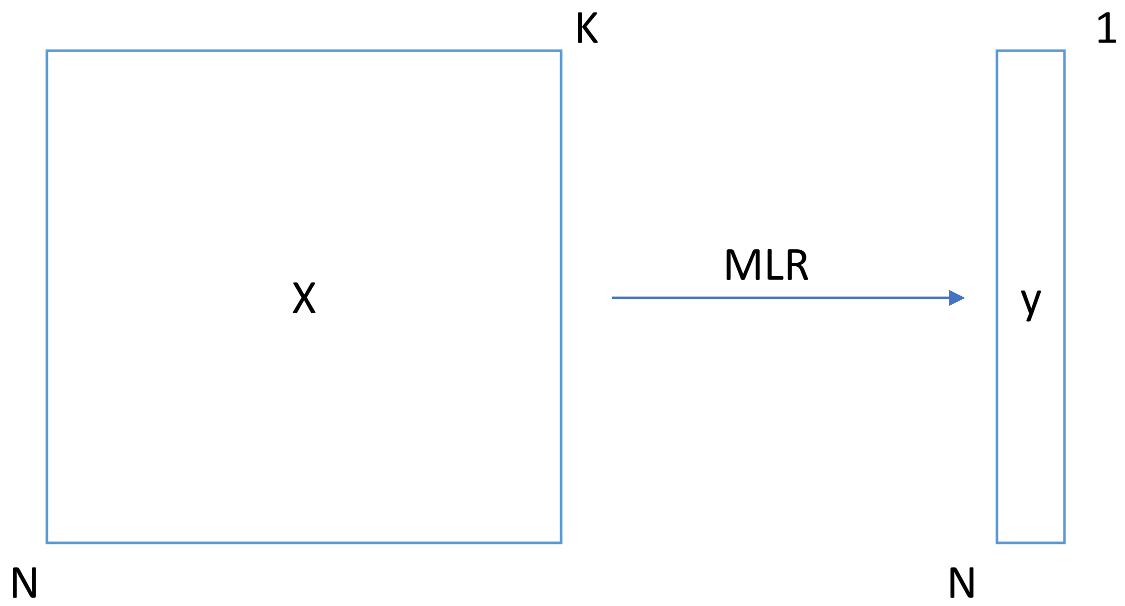
PCR: disadvantages

1. PCA components calculated without knowledge of $\mathbf{y}$
   - not necessarily predictive of $\mathbf{y}$
   - because steps 1 and steps 2 are performed sequentially
2. As a result, we often need to add additional, noisy components in PCA step
   - Add components beyond usual cross-validation

Where are we headed?

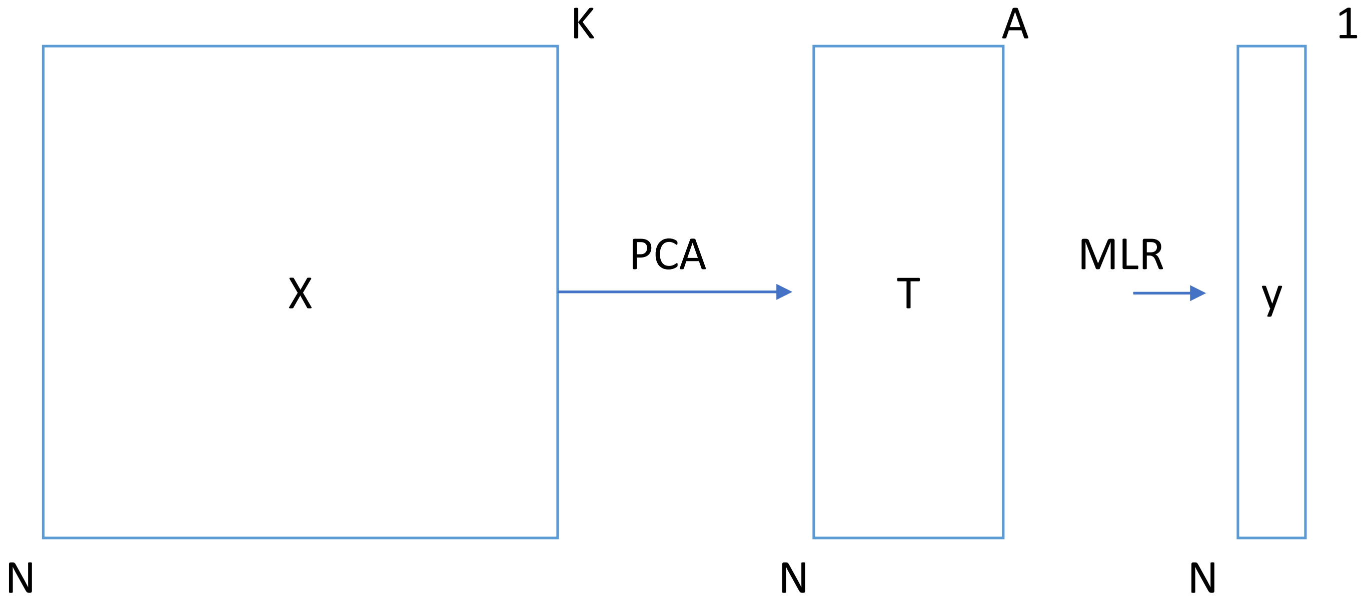
- Multiple linear regression (MLR)
- Principal component regression (PCR)
- Projection to latent structures (PLS)
    - Also called partial least squares (PLS)

Review: Multiple linear regression



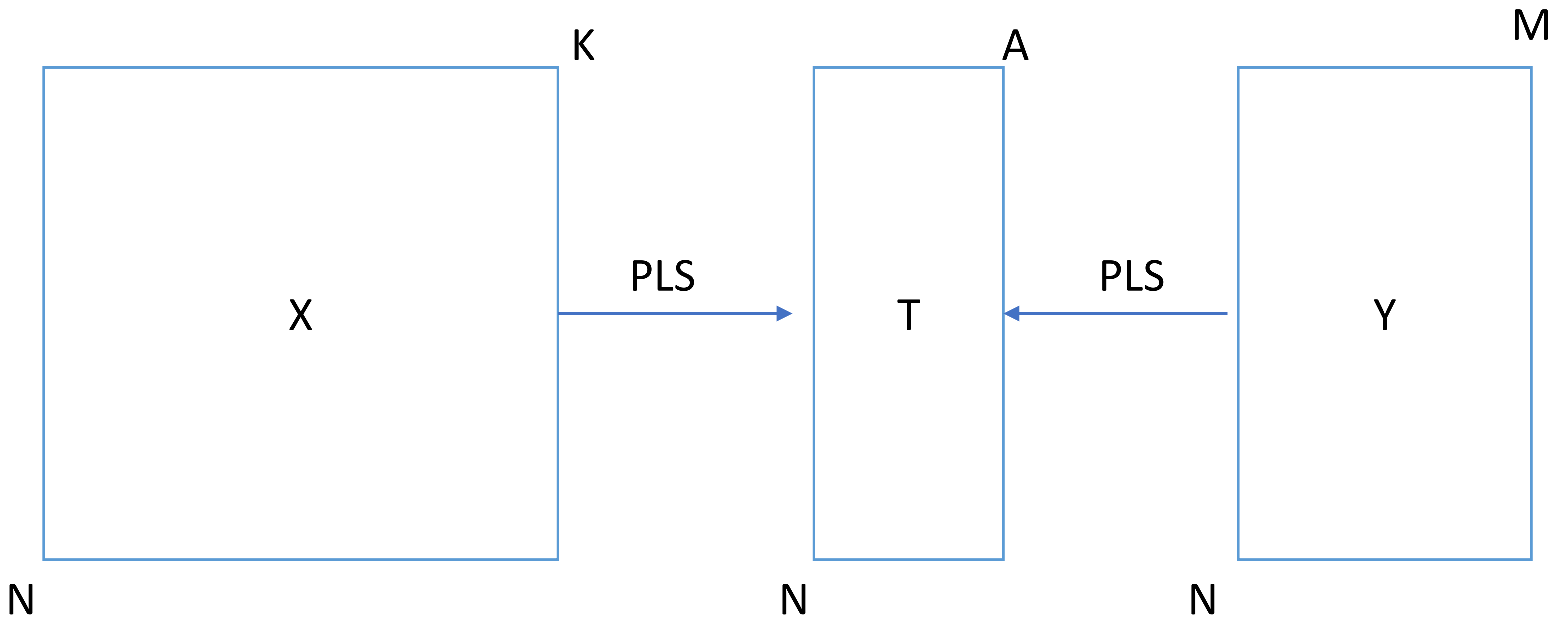
$$y = Xb$$
$$b = (X^T X)^{-1} X^T y$$

Review: principal component regression (PCR)



$$T = XP$$
$$\hat{y} = Tb$$

Projection to Latent Structures (PLS)



- 2 blocks of data
- Often used to predict Y given X
- Also used for monitoring, optimization, product development

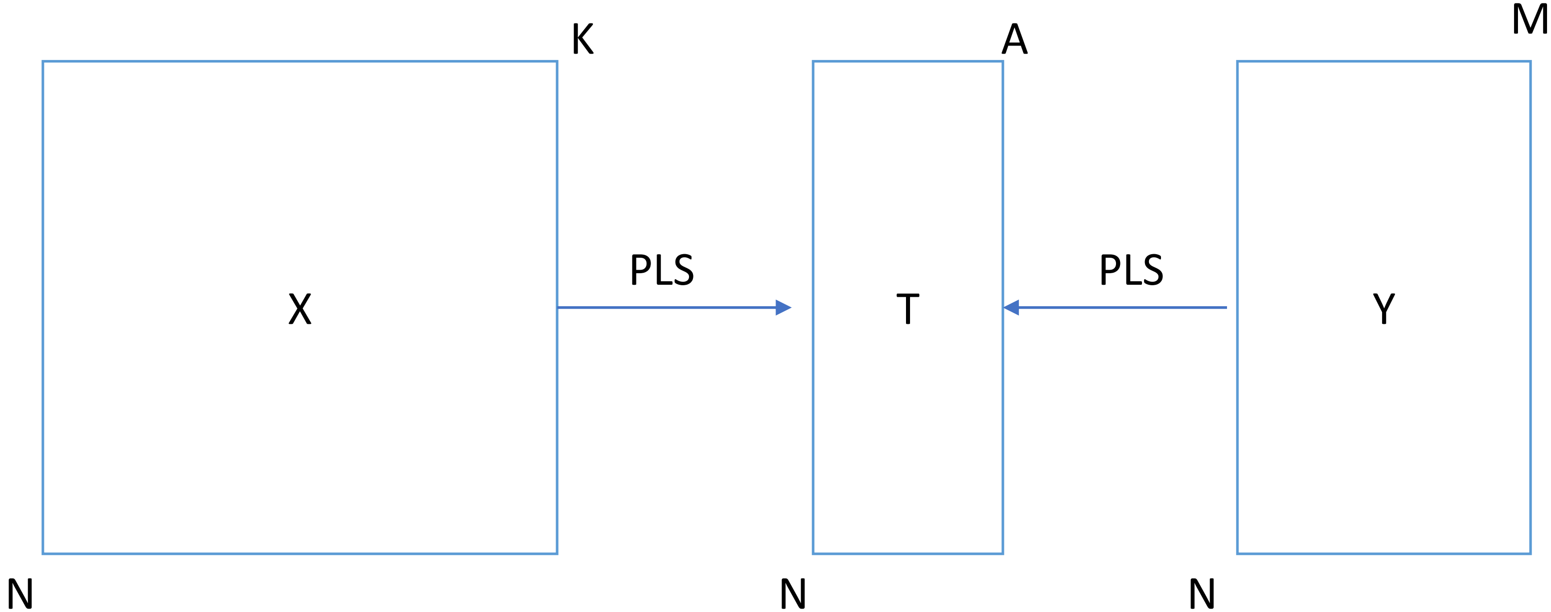
Projection to Latent Structures (PLS)

Advantages over PCR

- A single model is more efficient
  - Often fewer components than PCR
  - Easier to interpret
- PLS handles multiple Y-variables
- Assumes there is error in X and in Y

PLS overview

- Extracts each component sequentially (like PCA)
- Uses cross-validation to check the number of components
- Scores calculated from X and Y simultaneously
- Makes engineering sense: system driven by underlying latent variables

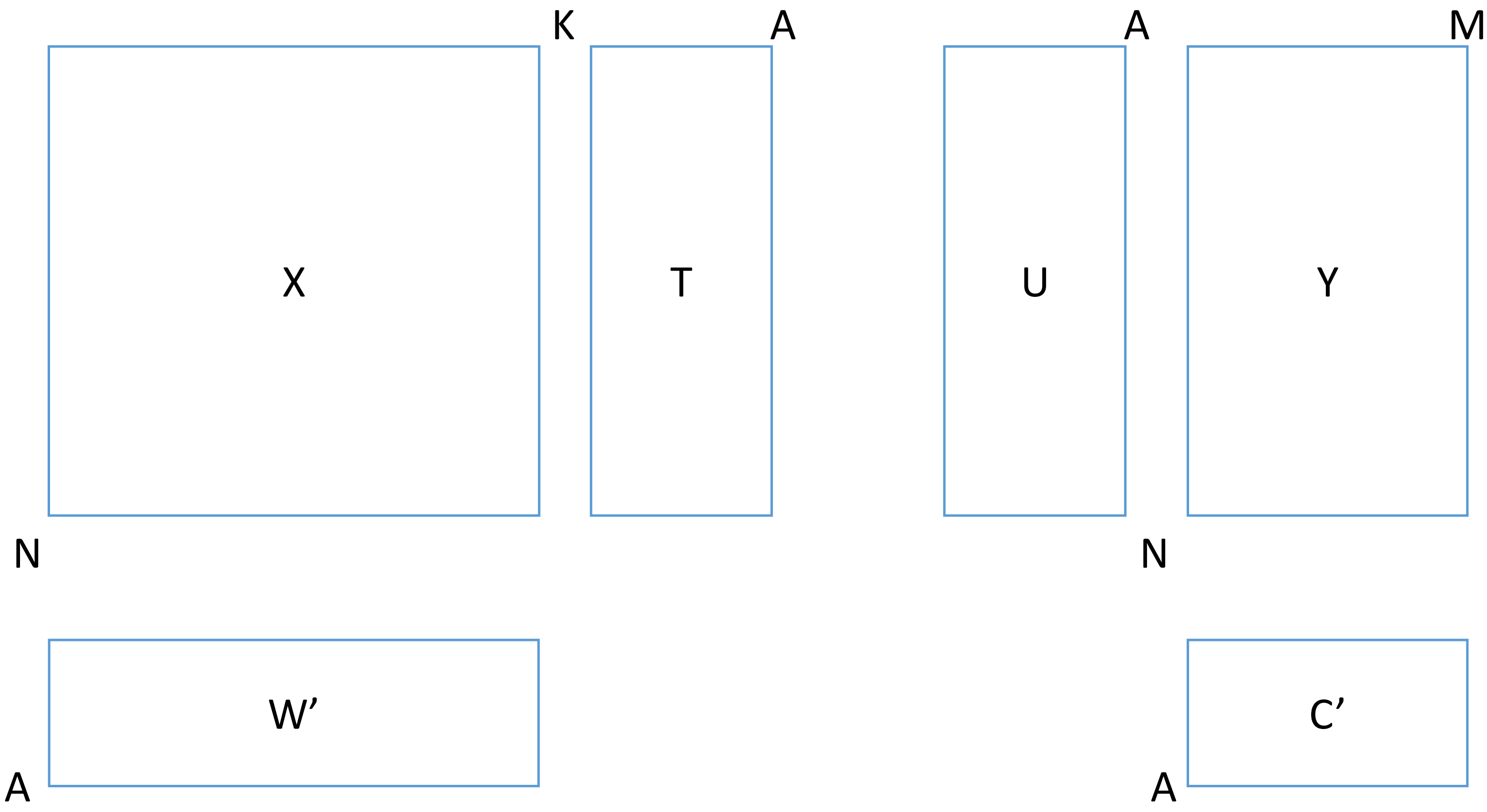


PLS Objective

- Objective of PCA: best explanation of the X-space

- What we want from PLS:
    1. Best explanation of X-space
    2. Best explanation of Y-space
    3. Maximize relationship between X and Y spaces

PLS: Notation



Geometric interpretation

On board

Review of PCA objective:

- For PCA: best explanation of X-space:

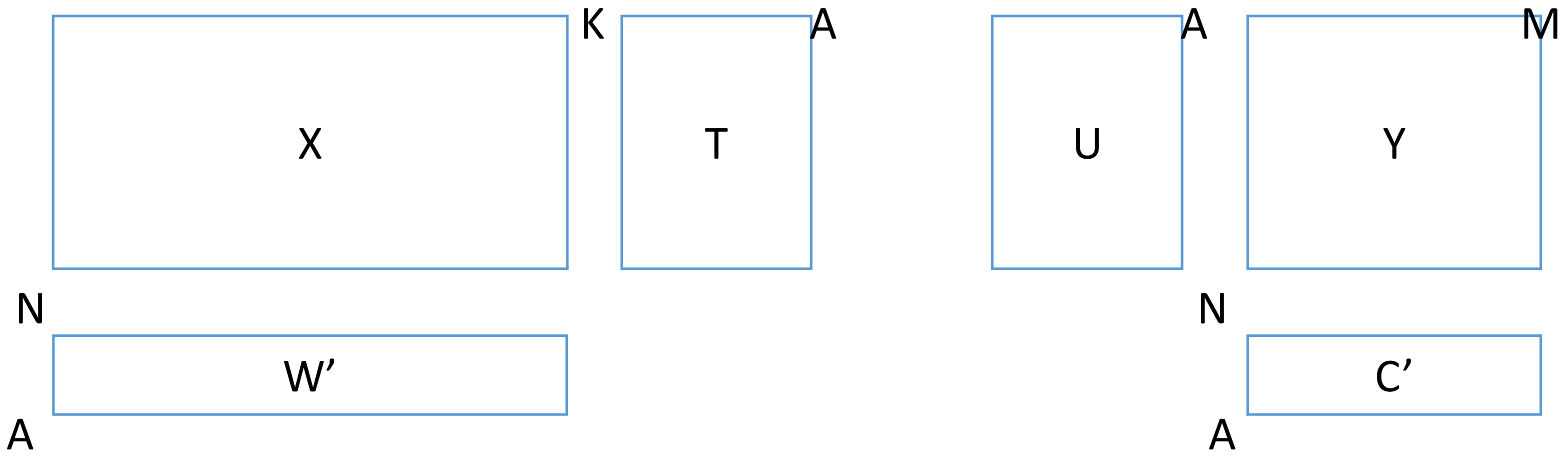
Max: $t_a^T t_a$ subject to $p_a^T p_a = 1.0$

- no other loading direction, p_a, gives greater variance of t_a

PCA Objective function:

Maximize $t_a^T t_a$

Simple PLS (SIMPLS)



- PLS scores explain X:
  - $t_a = X_a w_a$
  - $t_a^T t_a$ subject to $w_a^T w_a = 1.0$
- PLS scores also explain Y:
  - $u_a = Y_a c_a$
  - Max: $u_a^T u_a$ subject to c_a^T
- Maximize covariance (discussion on board)