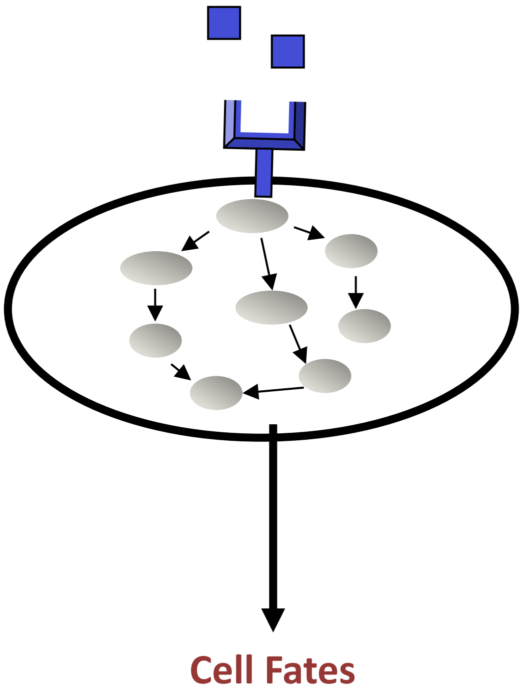
Using Partial Least Squares Regression (PLSR) to Analyze Cellular Response Data

Pamela Kreeger, Ph.D.

Department of Biomedical Engineering
University of Wisconsin-Madison
Madison, WI

Cue/Signal/Response Relationships



Cues

(e.g. growth factors, steroids, ECM)

Signals

(e.g. phosphorylation, transcription)

Due to time scale and multiple components, difficult to model with mechanistic detail

Response

(e.g., proliferation, apoptosis, migration, differentiation)

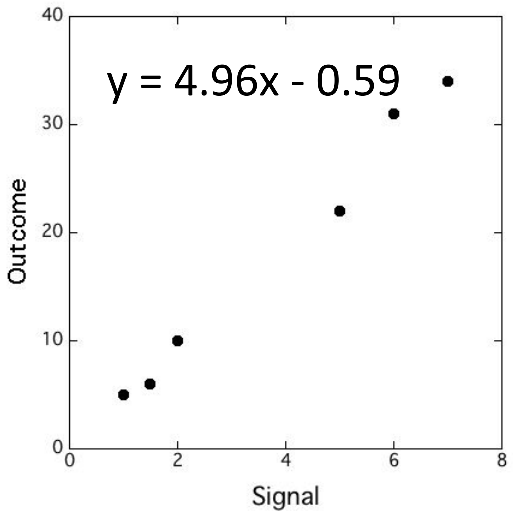
Methods for Signal/Response Modeling

$$\begin{array}{cc} \text{Signal Level} & \text{Response} \\ \begin{bmatrix}1\\2\\1.5\\5\\6\\7\end{bmatrix} & \begin{bmatrix}5\\10\\7\\24\\31\\35\end{bmatrix}\end{array}$$

We can classify and from the variation see that:

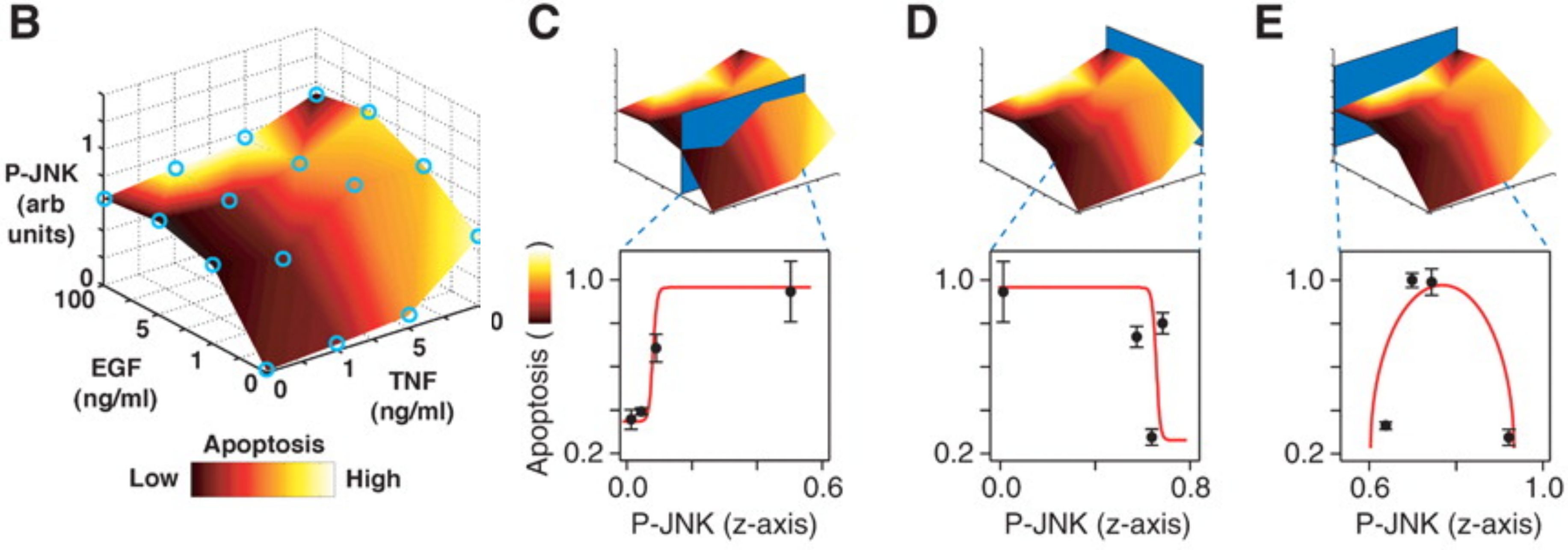
Low signal is correlated with Low response
High signal is correlated with High response

Alternatively, we may want to find a quantitative correlation between the signal and response.



Can then predict what outcome will be for new signaling data *and* what signals would be expected when a particular outcome is observed

Challenges with Univariate Relationships



The relationship between JNK activation and apoptosis appears to be highly context-dependent

→ univariate relationships are often insufficient

Janes, et al *Science* 2005

Multi-Linear Regression

In biology we often have multiple signals and multiple responses that were measured:

$$y_1 = a_1x_1 + b_1x_2 + e_1$$
$$y_2 = a_2x_1 + b_2x_2 + e_2$$

This can be written more concisely in matrix notation as:

$$\mathbf{Y = XB + E}$$

Where **Y** is a n x p matrix and **X** is a n x m matrix; minimizing **E** and solving for **B**:

$$\mathbf{B = (X^tX)^{-1}X^tY}$$

If n observations and m variables:

- m<n → no exact solution, least-squares solution possible
- m=n → one solution
- m>n → no unique solution unless we delete independent variables

  since $\mathbf{X^tX}$ cannot be inverted

  m >n is often the case in systems biology!

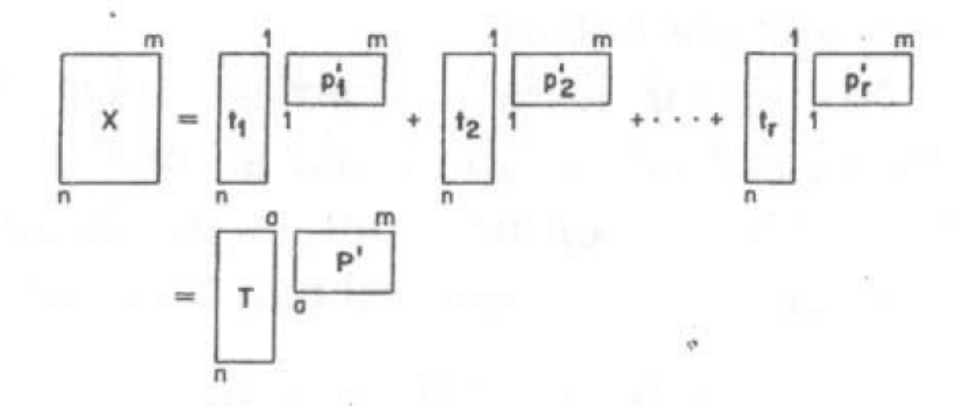
Principal Components Regression (PCR)

One solution - use the concepts from PCA to reduce dimensionality

1) Decompose **X** matrix

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{t} + \mathbf{E}$$

scores loadings residuals



As an alternative to finding the eigenvectors, the NIPALs algorithm breaks the **X** and **Y** matrices into a sum of vector products that recapitulate the eigenvectors/ eigenvalues

The components are found successively, with the first component found from **X** and the next from the residual of $\mathbf{X} - \mathbf{t}_1\mathbf{p}_1'$

Geladi *Analytica Chimica Acta* 1986

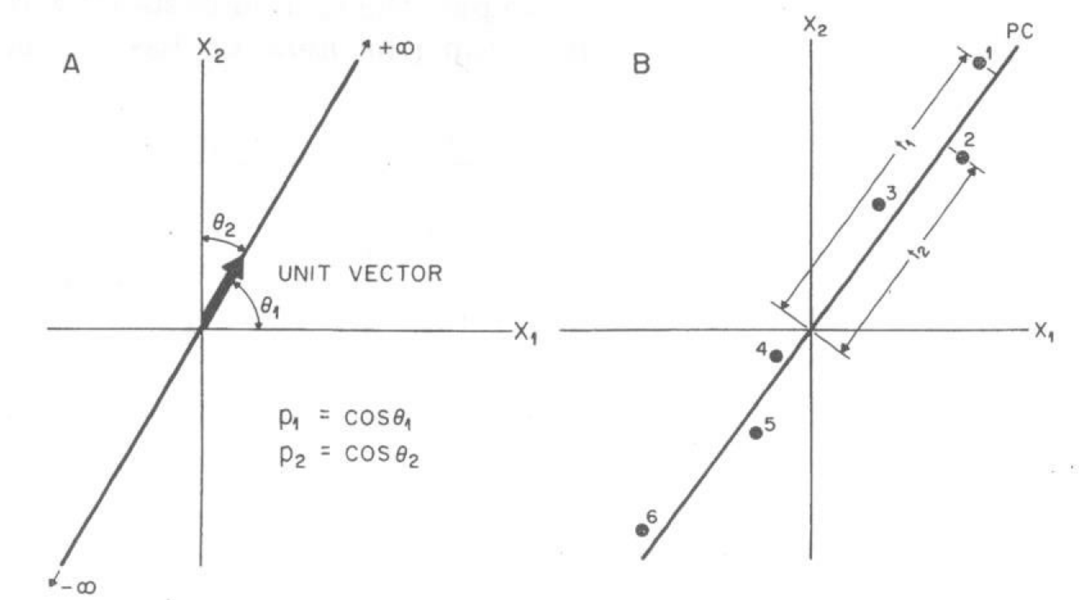
Principal Components Regression (PCR)

One solution - use the concepts from PCA to reduce dimensionality

1) Decompose **X** matrix

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E}$$

scores loadings residuals



Loadings (**p**) are the direction of the principal component in space

Scores (**t**) are the magnitude of where an observation is along the principal component

Geladi *Analytica Chimica Acta* 1986

Principal Components Regression (PCR)

One solution - use the concepts from PCA to reduce dimensionality

1) Decompose **X** matrix

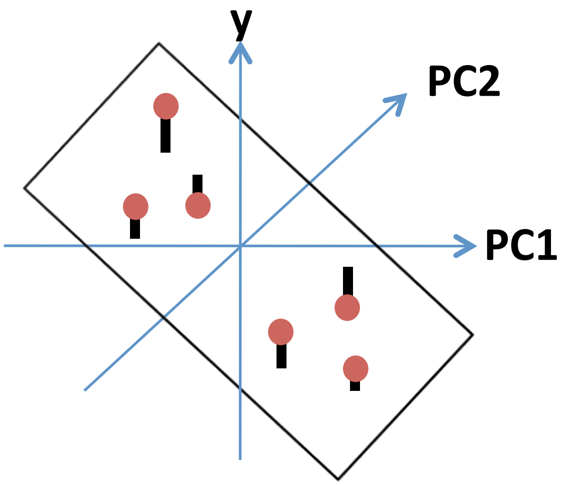
$$\mathbf{X} = \underset{\substack{\uparrow\\ \text{scores}}}{\mathbf{T}}\underset{\substack{\nwarrow\\ \text{loadings}}}{\mathbf{P}^{t}} + \underset{\substack{\nwarrow\\ \text{residuals}}}{\mathbf{E}}$$

2) Regress **Y** again the scores (Scores describe observations – by using them we link **X** and **Y** for each observation)

$$\mathbf{Y} = \mathbf{TB} + \mathbf{E}$$

Principal Components Regression (PCR)

Result – for each observation (●), there is a residual (▮) between the actual y value and the value for the y plane fit to the principal components.



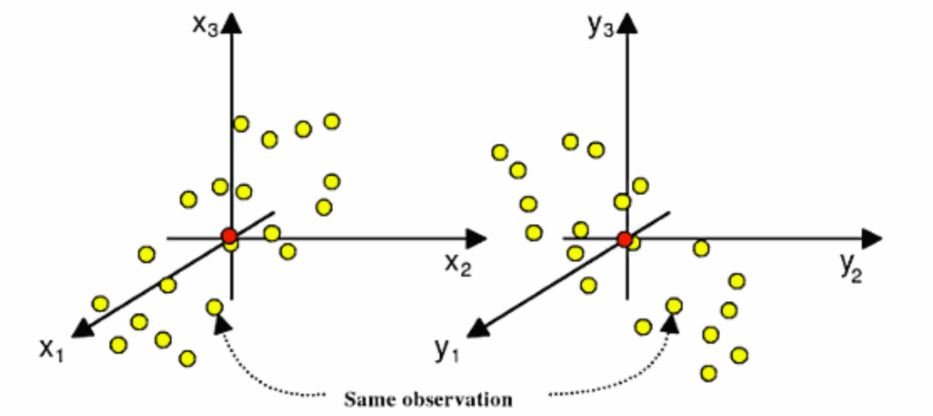
Problem – PCs for the **X** matrix do not necessarily capture **X**-variation that is important for **Y**

Example – the first components capture signaling that is related to another cell fate, while the signals that co-vary for this particular y are buried in later components

PLSR

PLSR = partial least squares regression
OR projection to latent structures

Data has values in both X and Y spaces for each observation



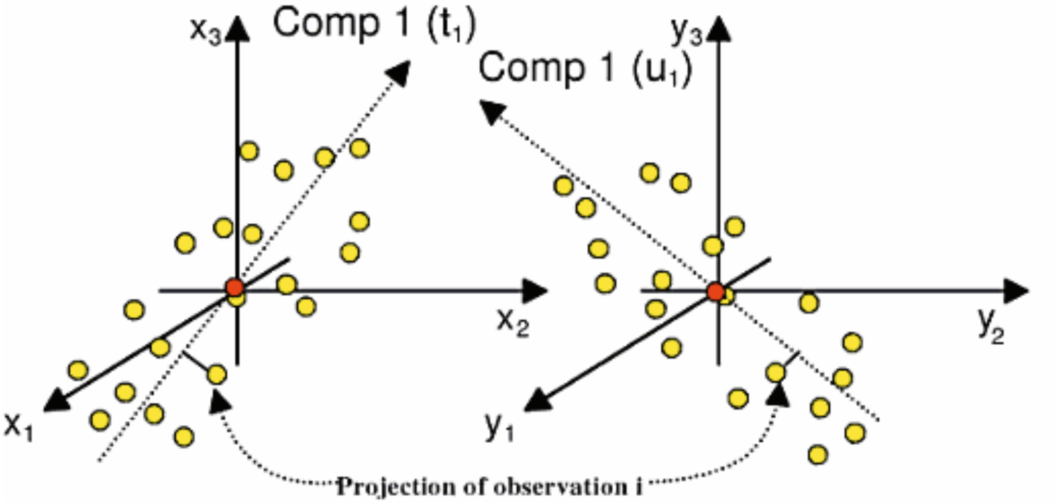
Find PCs for both matrices (while emphasizing the parts of **X** that correlate with **Y**) – will use NIPALs algorithm to construct the principal components.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E}$$

$$\mathbf{Y} = \underset{\substack{\uparrow \\ \text{scores}}}{\mathbf{U}}\,\underset{\substack{\nwarrow \\ \text{loadings}}}{\mathbf{Q}^t} + \underset{\substack{\nwarrow \\ \text{residuals}}}{\mathbf{F}}$$

Eriksson, et al. Multi- and Megavariate Data Analysis 2006

PLSR – NIPALs with Scores Exchanged



Steps for each component (h)

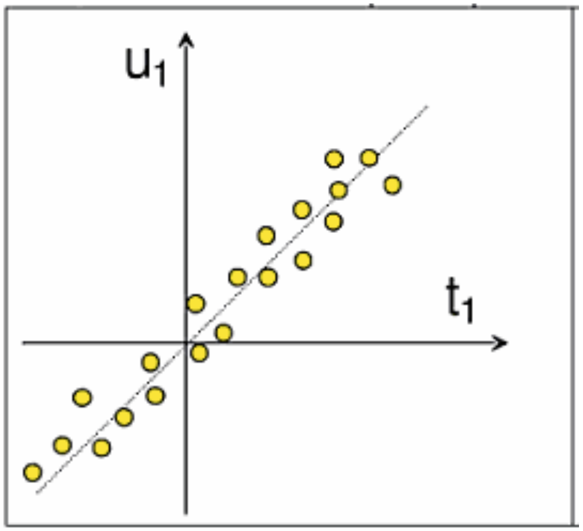
1) Find scores for **Y** ($\mathbf{u}_h$)

2) Use $\mathbf{u}_h$ to find the loadings for **X** ($\mathbf{p}_h$)

3) Use $\mathbf{p}_h$ to find scores for **X** ($\mathbf{t}_h$)

4) Use $\mathbf{t}_h$ to find **Y** loadings ($\mathbf{q}_h$)

5) Use $\mathbf{q}_h$ to calculate $\mathbf{u}_h$



Repeat until get convergence

The scores vectors are related by:

$$\mathbf{u}_h = b_h \mathbf{t}_h$$

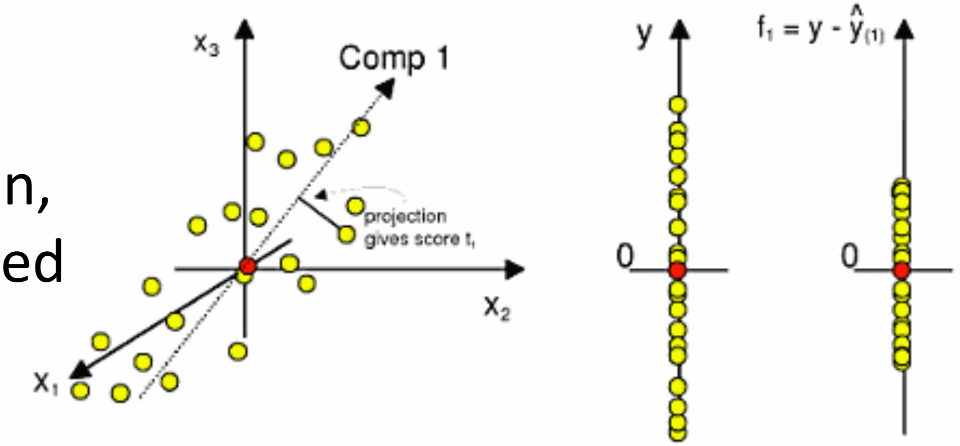
Note: Data is mean-centered for PLSR. Unit variance scaling can also be applied if the magnitudes of **X** values are not considered important

Eriksson, et al. Multi- and Megavariate Data Analysis 2006

PLSR – NIPALs with Scores Exchanged

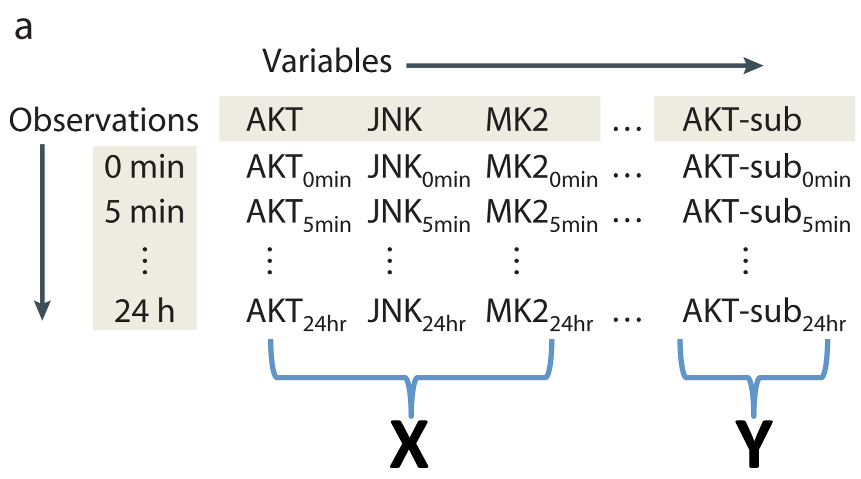
By forcing the **X** and **Y** matrices to swap scores vectors we rotate the principal components toward the independent variables that link most strongly to the dependent variables.

The first component still captures the most information, and what is in PC1 is subtracted before PC2 is calculated.



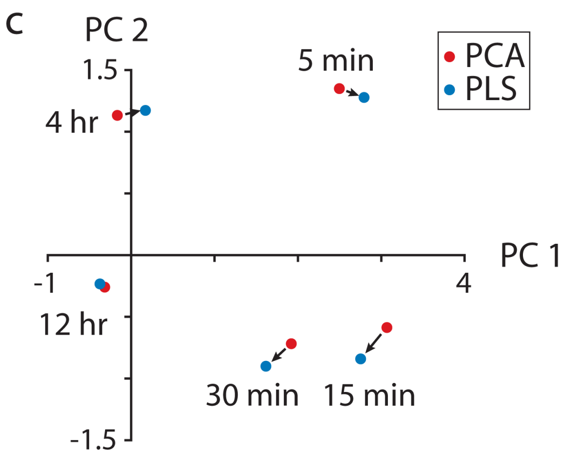
Eriksson, et al. Multi- and Megavariate Data Analysis 2006

Components in PLSR and PCA Differ

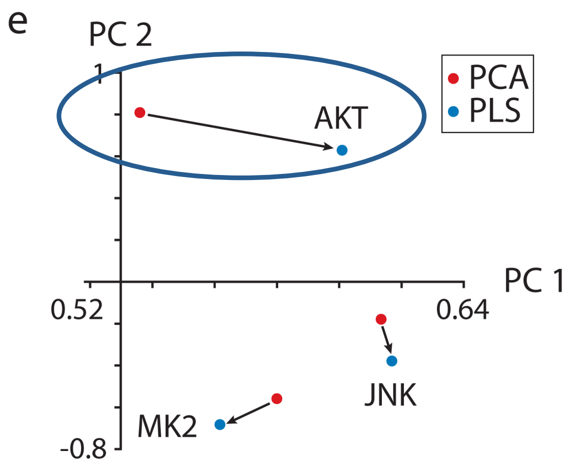


Observations	AKT	JNK	MK2	…	AKT-sub
0 min	$\mathrm{AKT}_{0\mathrm{min}}$	$\mathrm{JNK}_{0\mathrm{min}}$	$\mathrm{MK2}_{0\mathrm{min}}$	…	$\mathrm{AKT\text{-}sub}_{0\mathrm{min}}$
5 min	$\mathrm{AKT}_{5\mathrm{min}}$	$\mathrm{JNK}_{5\mathrm{min}}$	$\mathrm{MK2}_{5\mathrm{min}}$	…	$\mathrm{AKT\text{-}sub}_{5\mathrm{min}}$
⋮	⋮	⋮	⋮		⋮
24 h	$\mathrm{AKT}_{24\mathrm{hr}}$	$\mathrm{JNK}_{24\mathrm{hr}}$	$\mathrm{MK2}_{24\mathrm{hr}}$	…	$\mathrm{AKT\text{-}sub}_{24\mathrm{hr}}$

Scores plot



Loadings plot



Compare 2 models:

1) PCA on the **X** matrix

2) PLSR of the **X** and **Y** matrix

For example, AKT has a larger loading in PC1 in PLSR than in PCA

Janes *Nat Rev MCB* 2006

Determining the Number of Components

The optimal model will have enough components to accurately fit data and be predictive, but remain simple enough for interpretation. Additionally, the model is subject to over-fitting constraints.

Three metrics are used to evaluate the utility of adding a new component (a):

R^2X: sum of squares for the variation in the **X** matrix

$$R^2X = 1 - \frac{\Sigma(X_{model,a} - X_{obs})^2}{\Sigma(X_{obs}^2)}$$

R^2Y: sum of squares for the variation in the **Y** matrix

$$R^2Y = 1 - \frac{\Sigma(Y_{model,a} - Y_{obs})^2}{\Sigma(Y_{obs}^2)}$$

Q^2Y: fraction of the total variation in the **Y** matrix that can be predicted

$$Q^2Y = [1.0 - \Pi(PRESS/SS)_a]$$

PRESS = Prediction Error Sum of Squares

1) Remove an individual data element (i,k)
2) Fit model
3) Predict the element i,k that was withheld

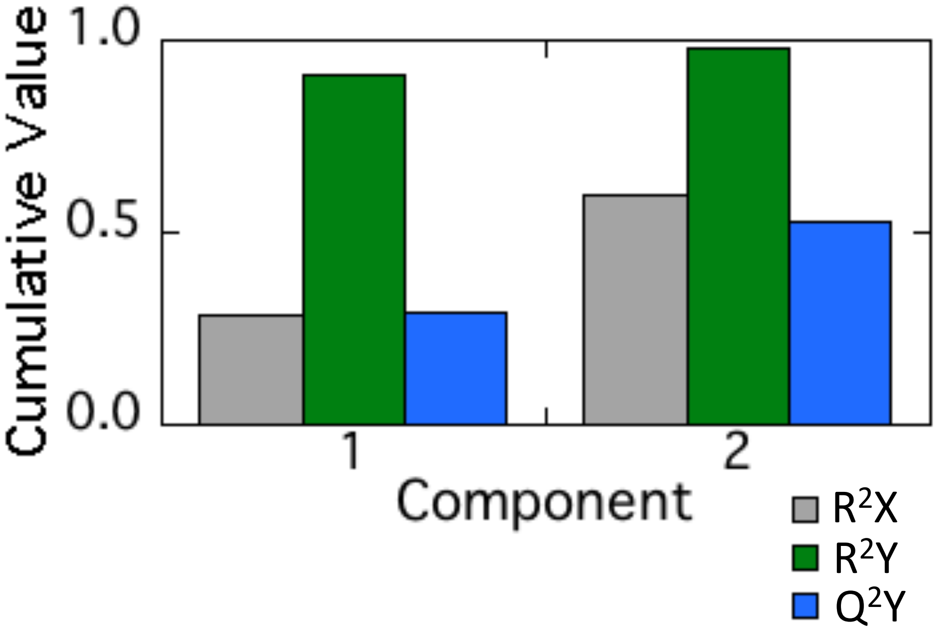
   $(observed_{i,k} - predicted_{i,k})^2$
4) Repeat until each element has been withheld once and only once

Determining the Number of Components

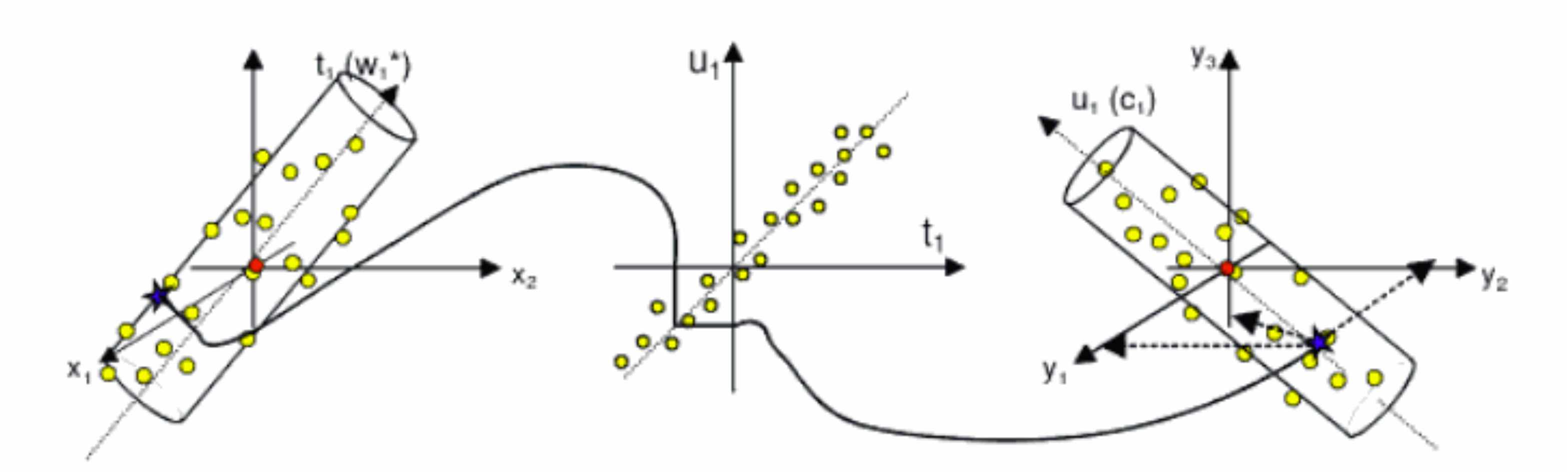
Each component contributes to these metrics – we evaluate those contributions and the cumulative value to determine if adding a new component is beneficial (Q^2Y is prioritized in this evaluation).

With each new component, evaluate the change to the cumulative Q^2Y

- Q^2Y increases significantly (>0.05), keep the component and evaluate the effect of adding another component
- Q^2Y goes down or has minimal change, stop model at the previous component



Utilizing PLSR for Predictions



Once the PLSR function has been defined, it can be used to predict the Y values for a new set of X values.

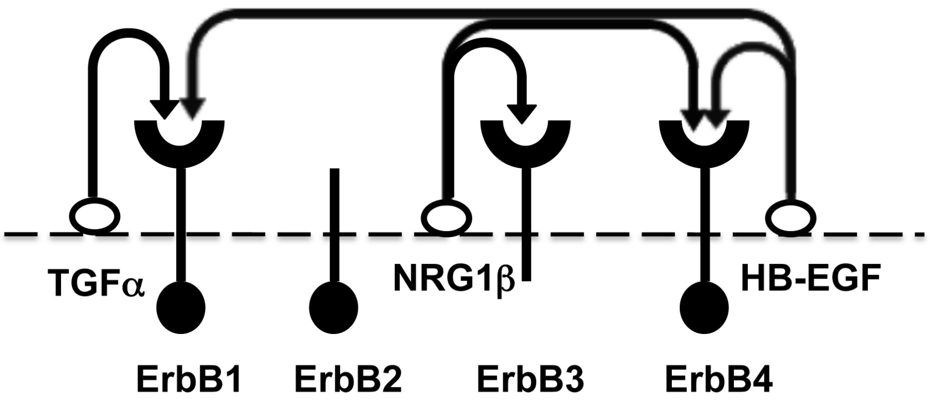
Can evaluate prediction accuracy:

DModY = s_i/s_o where s_i is the distance of the predictions and s_o is a normalization term accounting for the residual standard deviation in the model (smaller DModY indicates better prediction)

Eriksson, et al. Multi- and Megavariate Data Analysis 2006

PLSR to Study ErbB in Ovarian Cancer

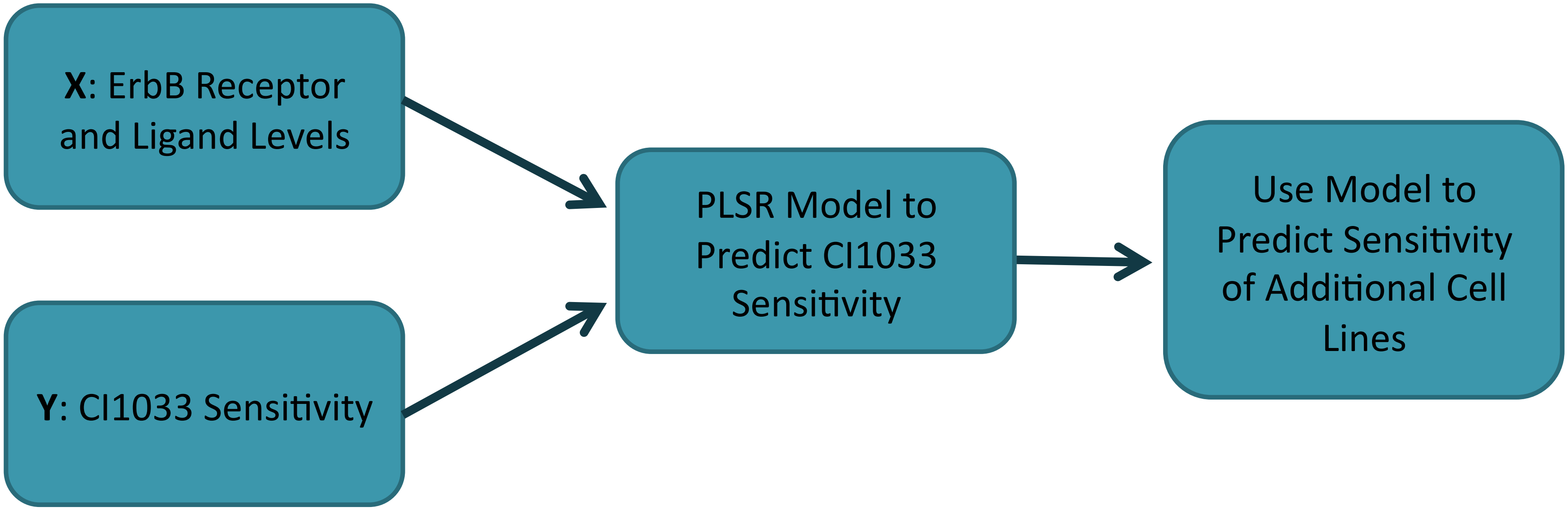
- Advanced tumors express multiple receptors/ligands
- Clinical trials with ErbB inhibitors have had little success
- Trials have not targeted inhibitors to particular sub-groups → how to identify these groups?



Prasasya, et al. *Biotech and Bioeng* 2012

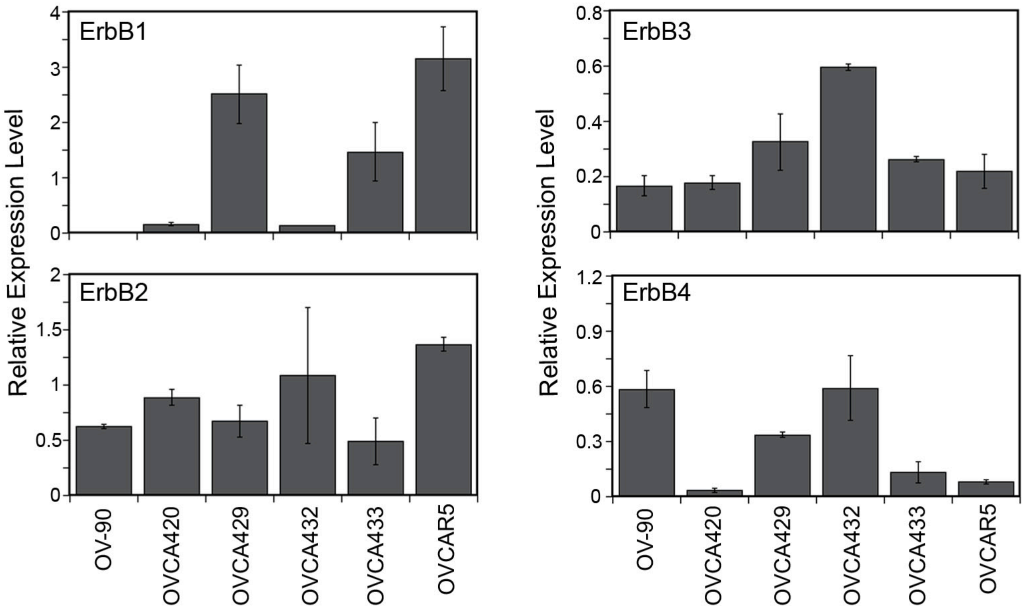
Hypothesis:

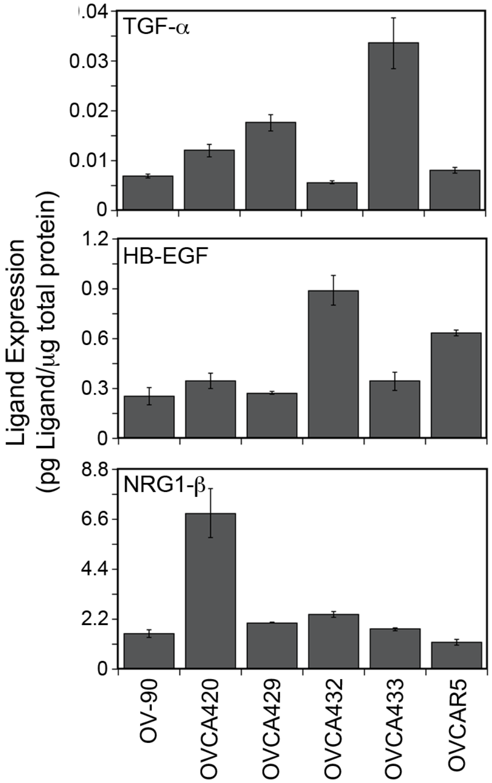
Sensitivity to ErbB Inhibitors is a Function of ErbB Network Composition


X: ErbB Receptor and Ligand Levels
Y: CI1033 Sensitivity
PLSR Model to Predict CI1033 Sensitivity
Use Model to Predict Sensitivity of Additional Cell Lines

Prasasya, et al. *Biotech and Bioeng* 2012

OvCa Express ErbB Ligands and Receptors



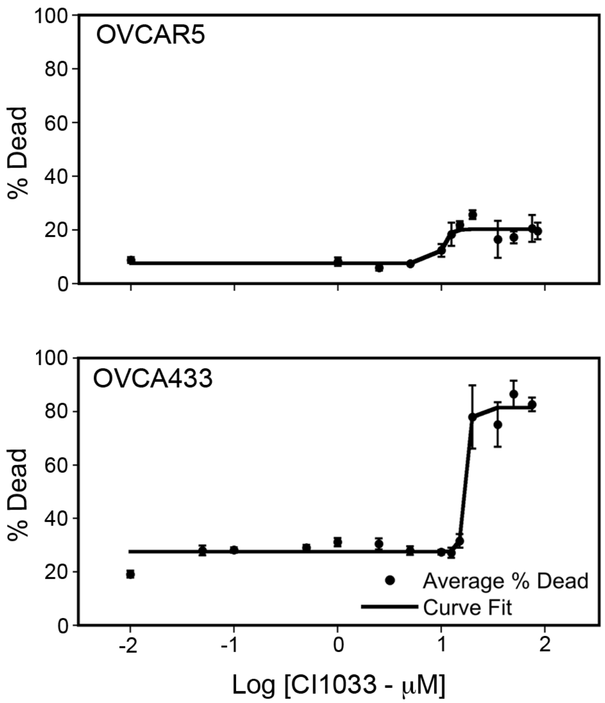


Six ovarian cancer cell lines were examined for ErbB receptor (Western blots) and ligand (ELISA) levels

- The levels of each individual protein varied widely across the panel
- The receptor/ligand combinations also varied across the panel

Prasasya, et al. *Biotech and Bioeng* 2012

OvCa Have Different Sensitivity to CI-1033



Cell Line	Sensitivity (%)
OV-90	53.2
OVCA420	91.1
OVCA429	46.9
OVCA432	48.5
OVCA433	53.9
OVCAR5	12.6

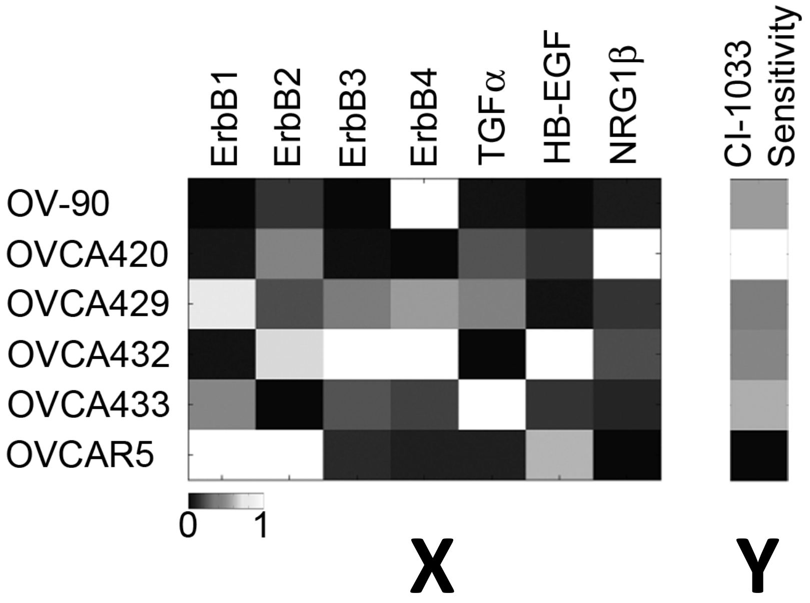
These cell lines were treated with increasing doses of CI-1033 and the level of cell death determined by CytoTox Glo (Promega)

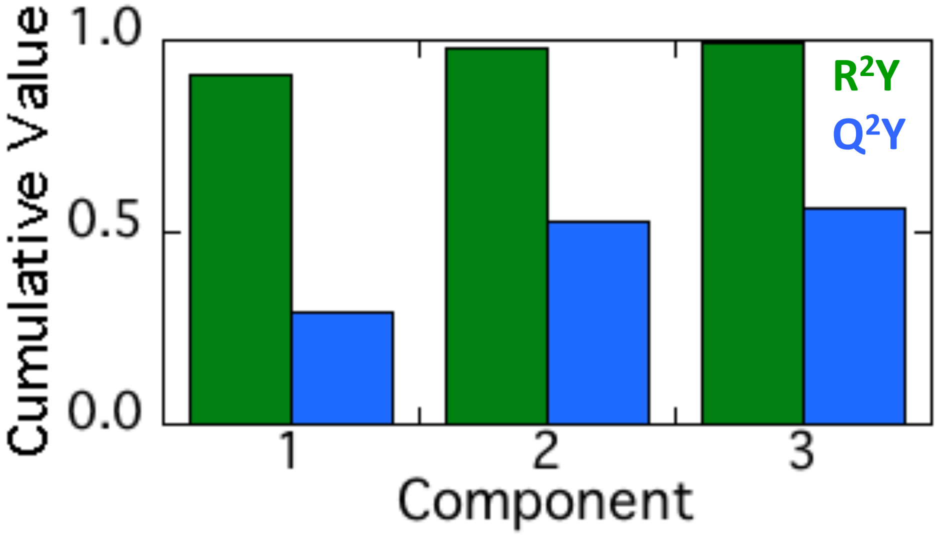
- EC50 showed small variation (3-20µM)
- The maximum increase in cytotoxicity varied greatly across the panel

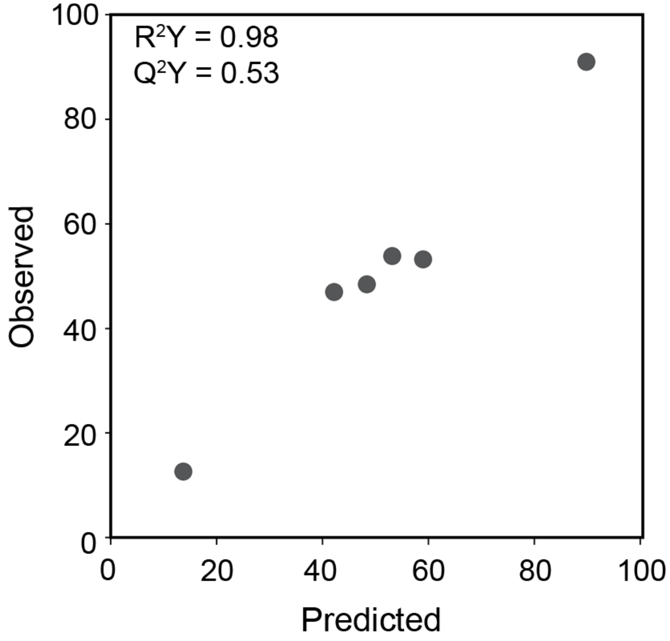
  Sensitivity = Maximum % Dead – Baseline % Dead

Prasasya, et al. *Biotech and Bioeng* 2012

PLSR Relates ErbB Levels to Sensitivity

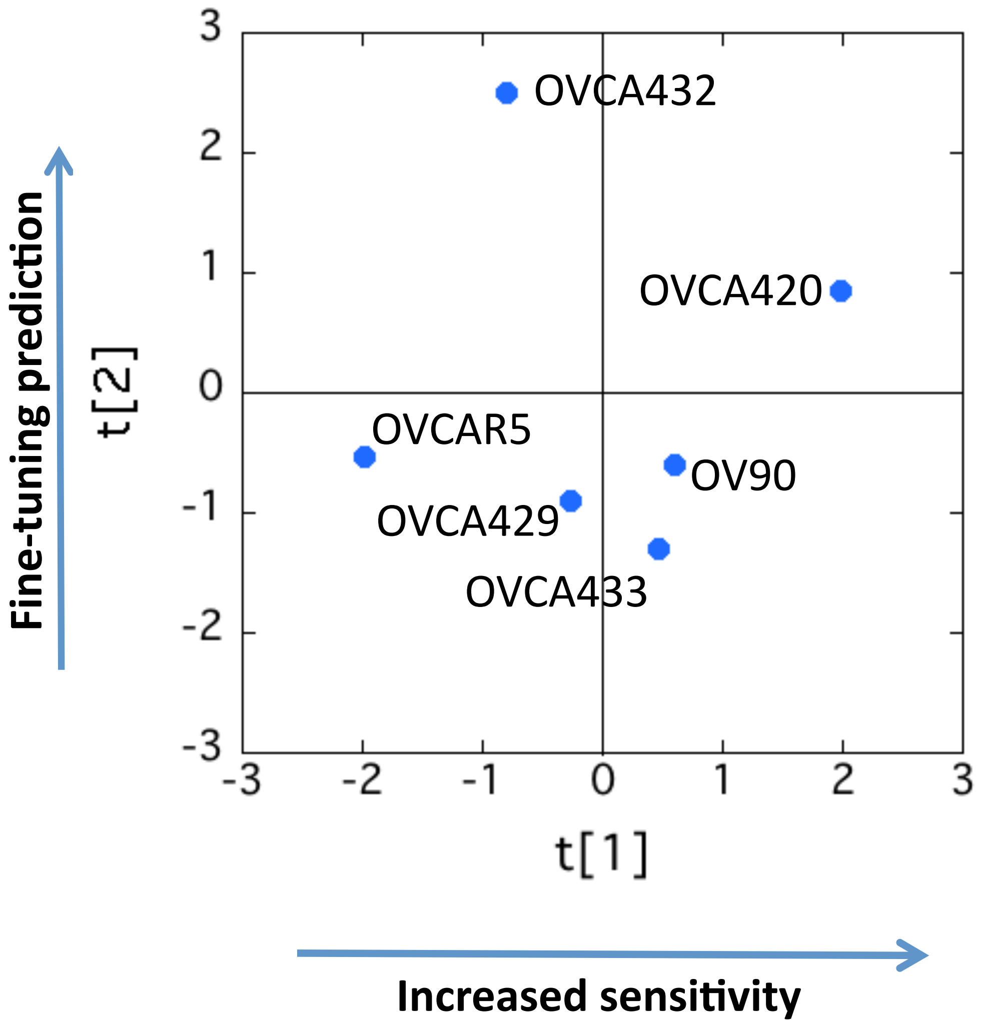






Prasasya, et al. *Biotech and Bioeng* 2012

Interpreting PLSR - Scores

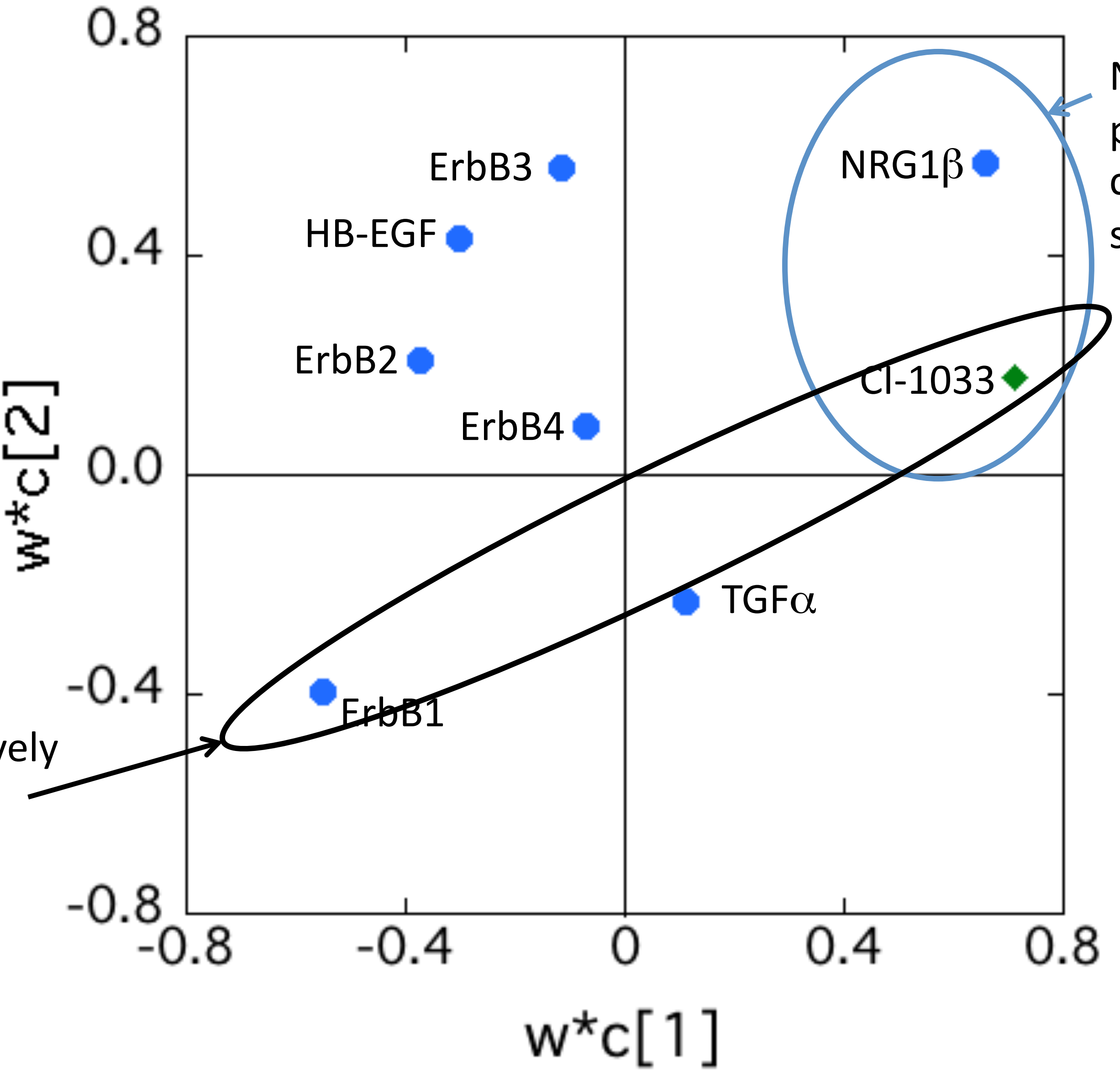


Fine-tuning prediction

Increased sensitivity

Prasasya, et al. *Biotech and Bioeng* 2012

Interpreting PLSR - Loadings

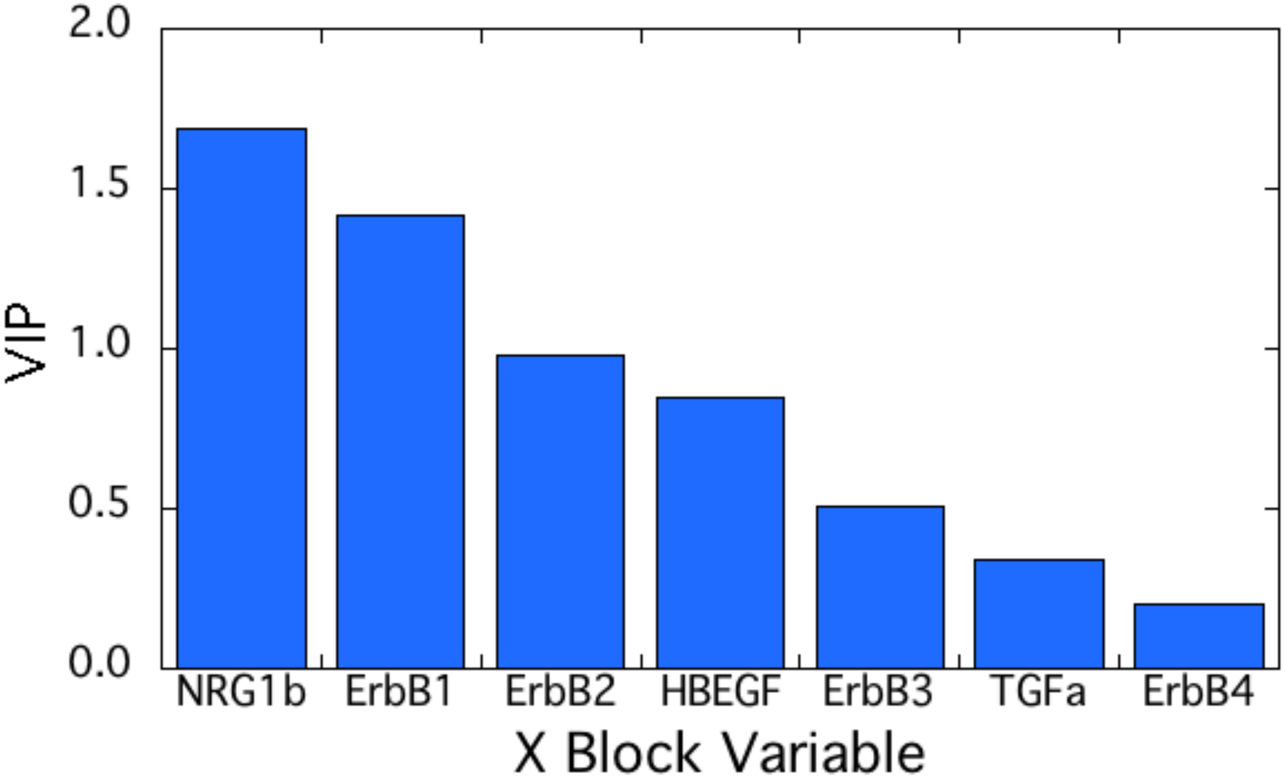


NRG1β is positively correlated with sensitivity

ErbB1 is negatively correlated with sensitivity

Prasasya, et al. *Biotech and Bioeng* 2012

Interpreting PLSR - VIP

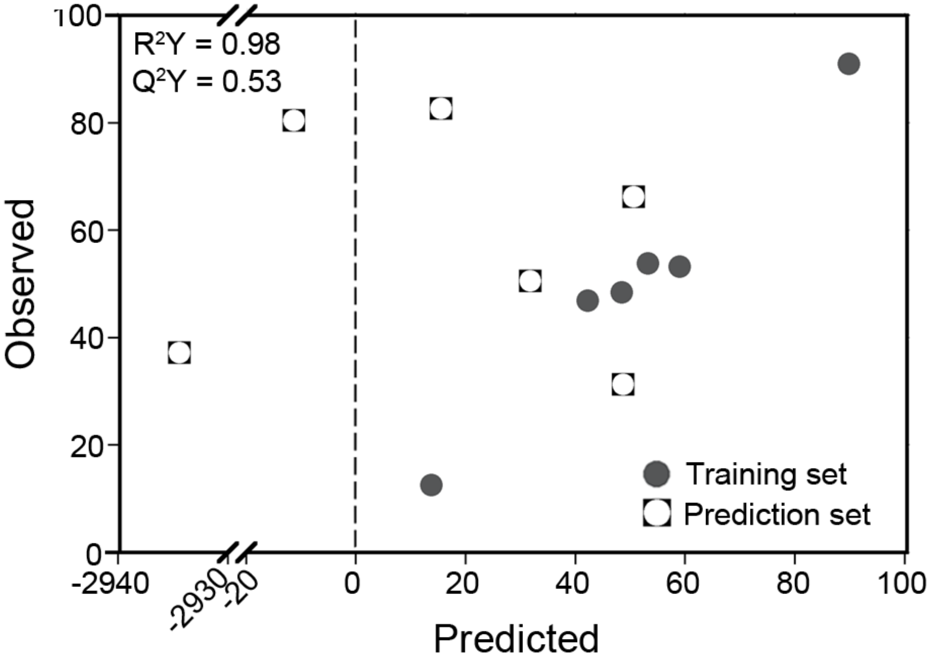


VIP = Variable Importance of Projection

- Evaluated for each **X** variable across the entire model, not for individual components
- Incorporates the weights for each variable and the variation for each respective component across the model
- Values > 1.0 indicate important variables for explaining **Y**

Prasasya, et al. *Biotech and Bioeng* 2012

PLSR Predicts Sensitivity with Mixed Results



Cell Line	Accuracy
OVCAR3	Good
TOV112D	Poor
A2780	Poor
Ca-OV3	Good
SKOV3	Poor
TOV21G	Good

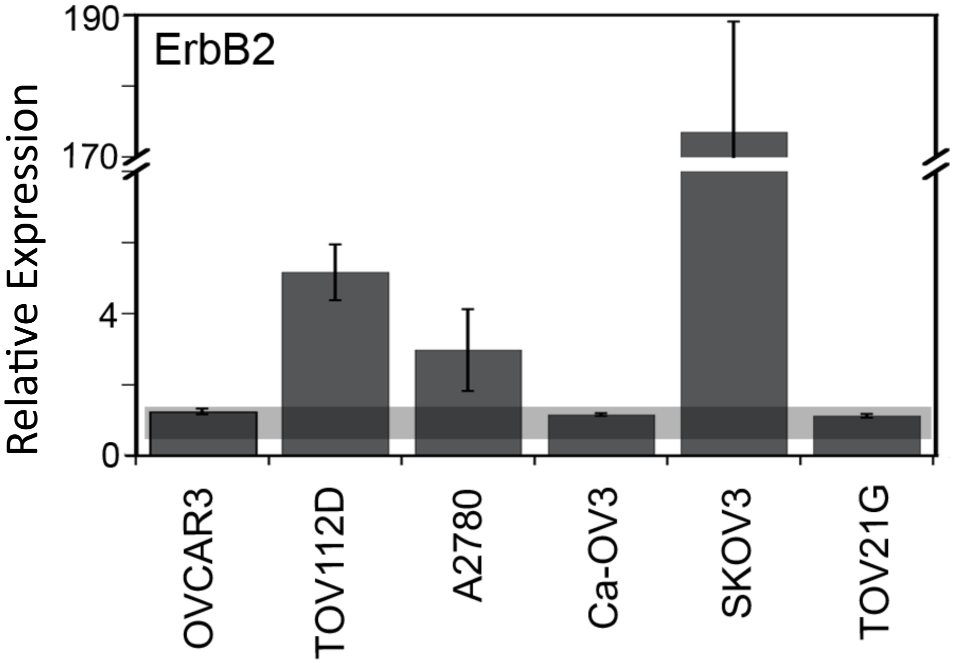
Receptor and ligand levels were determined for 6 additional cell lines and sensitivity predicted from the PLSR model.

For 3 of the 6 cells the prediction is accurate, while for 3 there are large errors.

To improve the model, need to determine source of this error.

Prasasya, et al. *Biotech and Bioeng* 2012

PLSR Predicts Sensitivity with Mixed Results



Cell Line	Accuracy
OVCAR3	Good
TOV112D	Poor
A2780	Poor
Ca-OV3	Good
SKOV3	Poor
TOV21G	Good

Examining the **X** matrix, we clearly see a connection between cells with high ErbB2 levels and failure to accurately predict.

Possible solutions:

1) Expand training set data to include a cell line that overexpresses ErbB2
2) Remove ErbB2 from the model

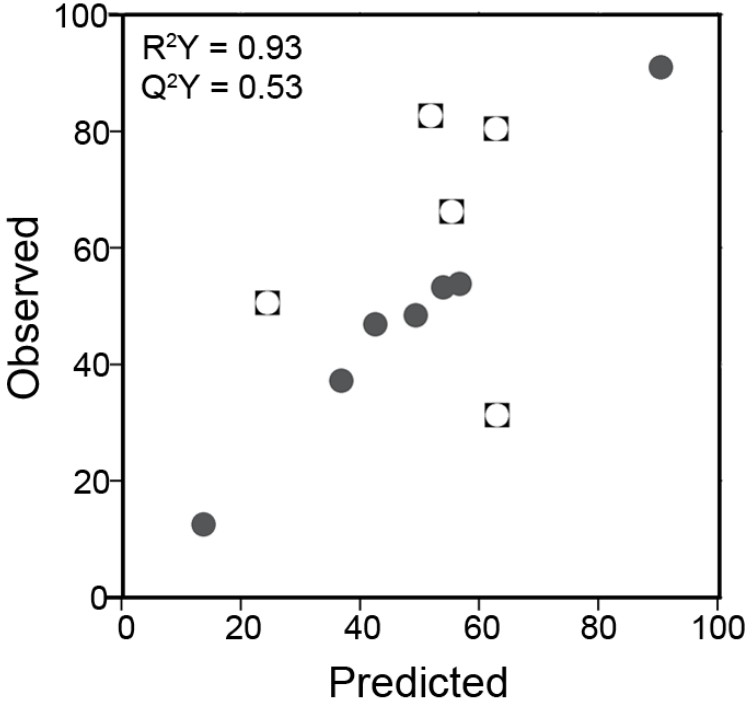
Prasasya, et al. *Biotech and Bioeng* 2012

Expanded Training Set Improves Prediction

Including cells that moderately overexpress ErbB2 is not sufficient to improve SKOV3 prediction.

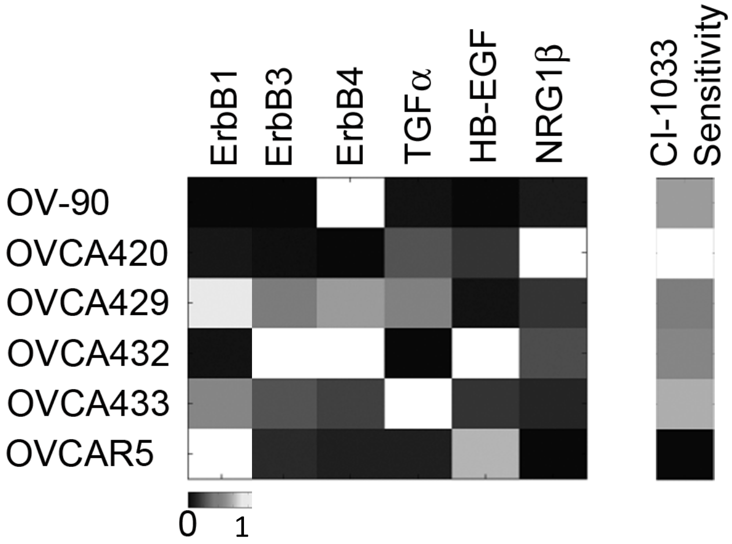
By including SKOV3 in the training set, can predict remaining cell lines with improved accuracy.

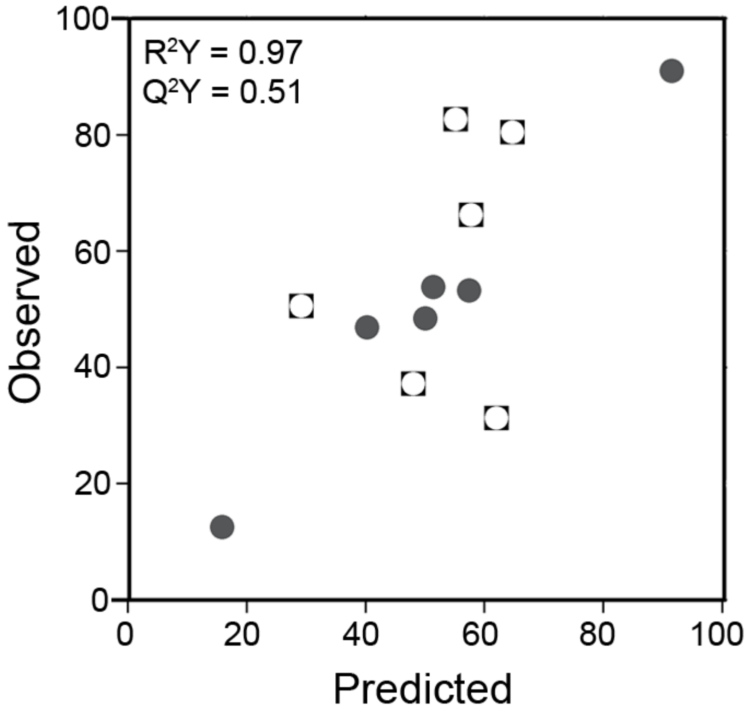
Training data must capture full range of **X** and **Y** variation!



Prasasya, et al. *Biotech and Bioeng* 2012

Smaller Model Improves Prediction





ErbB2 is rarely overexpressed in ovarian cancer.

Rebuild model without ErbB2 in **X** matrix and get accurate prediction of all 6 cell lines.

<u>Leaving data out can improve prediction accuracy</u>.

Prasasya, et al. *Biotech and Bioeng* 2012

Optimal Models Need ErbB1 & ErbB Ligands

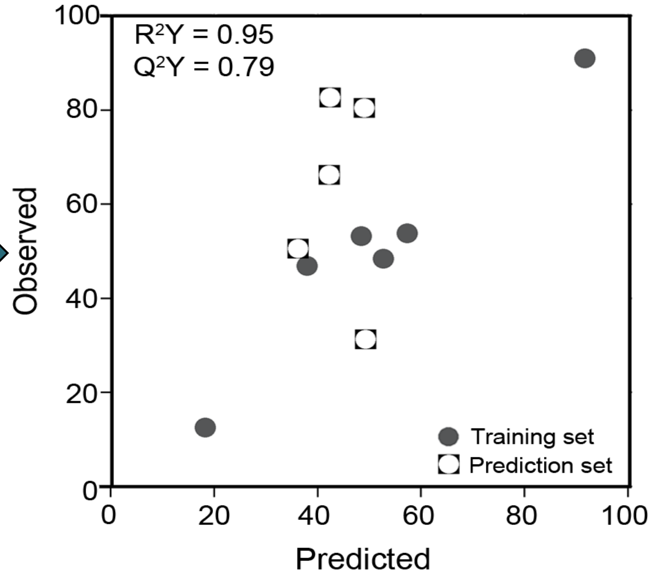
Tried 127 model variants (all possible combinations of **X** matrix)

ErbB	ErbB2	ErbB3	ErbB4	TGF-α	HB-EGF	NRG1-β	R^2Y	Q^2Y	DmodY
▨	□	▨	□	□	▨	▨	0.91	0.72	1.92
▨	□	▨	□	▨	□	▨	0.91	0.62	2.05
▨	□	□	▨	□	▨	▨	0.94	0.52	2.23
▨	□	□	□	□	▨	▨	0.94	0.82	2.39
▨	□	□	□	▨	□	▨	0.95	0.79	2.95
▨	□	□	▨	▨	▨	▨	0.96	0.55	3.09
▨	□	□	□	▨	▨	▨	0.97	0.82	3.57
▨	□	▨	▨	▨	▨	▨	0.97	0.51	3.64
▨	□	▨	□	▨	▨	▨	0.98	0.76	4.62

▨ = included in X matrix

□ = excluded from X matrix

DmodY = distance of prediction data set to model



Best models do not include ErbB2.

Best models include at least 2 ligands suggesting autocrine loops are linked to sensitivity.

Prasasya, et al. *Biotech and Bioeng* 2012

PLSR Variants

- DPLS – Discriminant PLS
  - The response matrix consists of classifications such as control = 0, treated = 1
- OPLS/O2PLS – Orthogonal PLS
  - OPLS – the **X** matrix is broken down into parts that predict **Y** and parts that are unrelated to **Y**
  - O2PLS – both matrices are broken down into related and unrelated parts

Summary

PLSR vs. PCA

PCA – has an **X** matrix; maximize the variance

PLSR – has an **X** and **Y** matrix; maximize the covariance

Interpreting PLSR

R^2X, R^2Y, Q^2Y (maximum value of 1)

Using Q^2Y to determine number of components

Scores/loadings

DModY (lower = better prediction)

VIP (>1 indicates important)