Simplifying the Interpretation of ToF-SIMS Spectra and Images using Careful Application of Multivariate Analysis

Matthew S. Wagner
The Procter & Gamble Company
wagner.ms@pg.com

SIMS XV
September 13, 2005

Opportunities

Instrumentation

Relatively mature*

Sample Preparation

Growth area

Data Processing

Growth area

Data Acquisition

Relatively mature

* See talk by N. Winograd, Thursday 15-Sept, 9:40am

Goal for Data Analysis

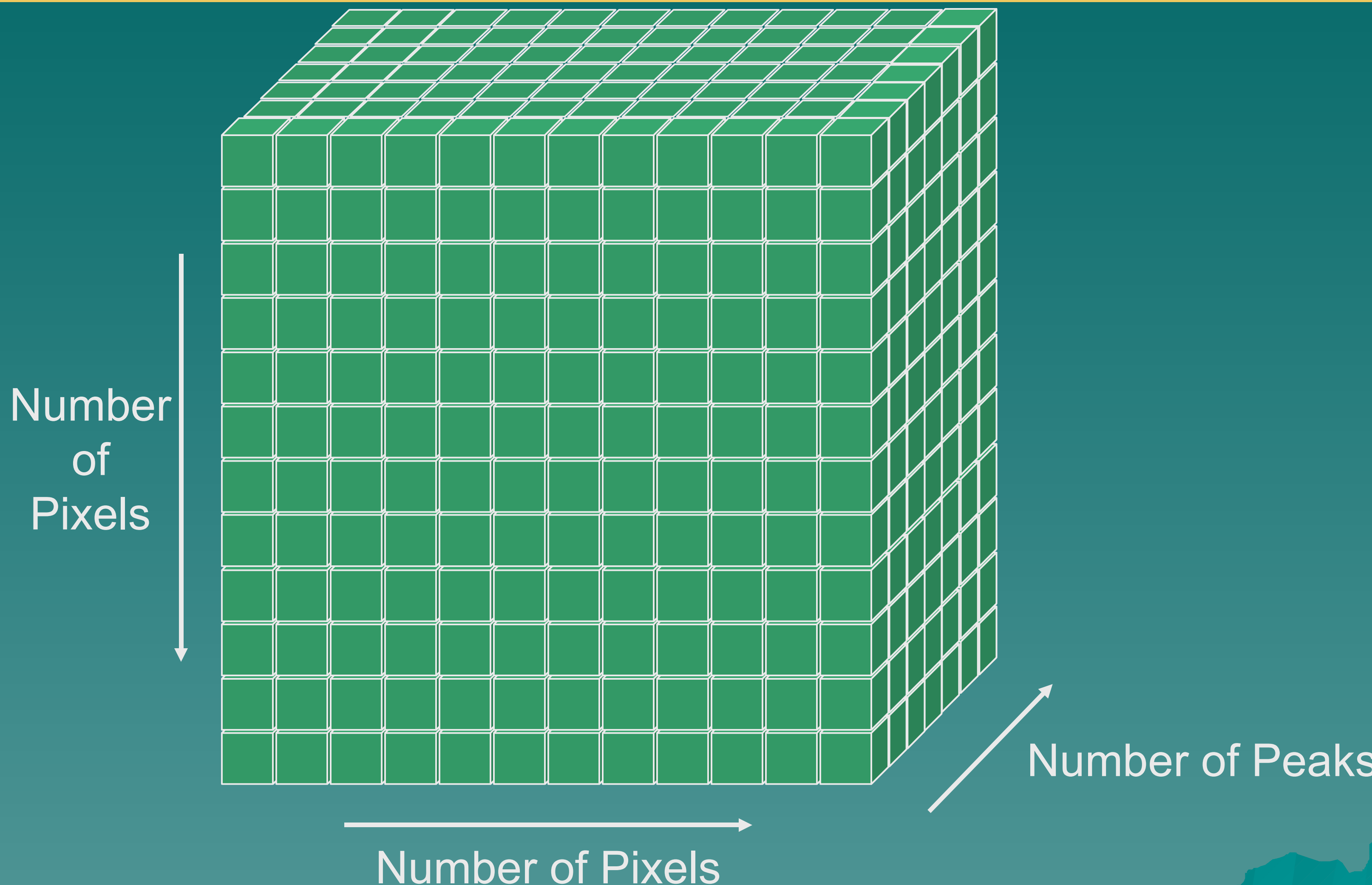
Concise and accurate chemical description of surface chemistry

Data Processing: Challenges

- Data overload
  - Large spectral and image datasets

- Use of Multivariate Analysis (MVA)
  - When is it appropriate?
  - Appropriate experimental design?
  - Appropriate pre-processing?
  - Which MVA method is best?
  - Validation of MVA methods?
  - Accurate interpretation with physically meaningful results?

Data Overload: Too many spectra!



How do we compare *multiple spectra* on the basis of *multiple peaks* in each spectrum?

Data Overload

- Generating data is (relatively) easy…

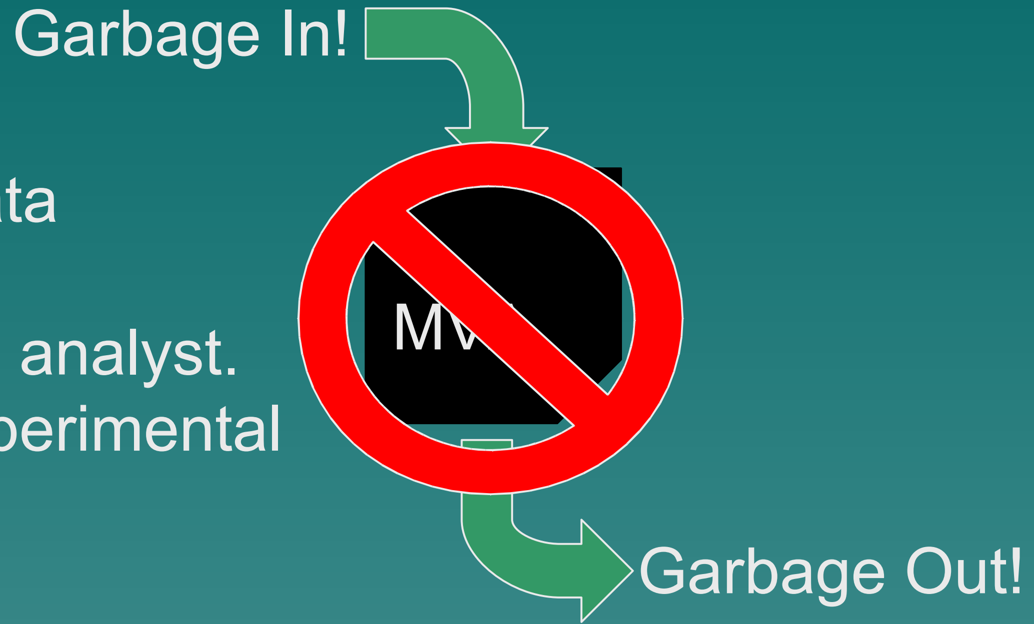
  Efficiently processing the data
  is the challenge!

- Many peaks in a spectrum…
- Peak intensities are correlated…
- Need to process spectra rapidly…
- Images present even more challenges…
  - Low signal-to-noise...
  - Large number of pixels…
  - Comparison of multiple images…

Multivariate Analysis Benefits

- Can *simplify* data analysis…
- Many examples of MVA application to SIMS data…
  - See *Surf. Sci.* **570**: 78 (2004)
- Requires good understanding of the analytical tool…

MVA: Not a Black Box!!!



MVA *is not*:

- A "black-box" tool for data analysis.
- A substitute for a skilled analyst.
- A substitute for poor experimental design.
- "Magic".

MVA *is*:

- An important and useful tool for saving the analyst time and money.
- An important and useful tool for maximizing the use of your data!

Before MVA: Data Pre-processing

- Many types of pre-treatment possible:
  - Peak selection
  - Normalization (this is a type of scaling)
  - Mean-centering, Autoscaling, Log-scaling, Mean-scaling, Poisson-scaling, etc.

- All data pre-treatments involve assumptions about the data!

- No standard method exists to determine which is best!
  - Trial-and-error approach widely used…

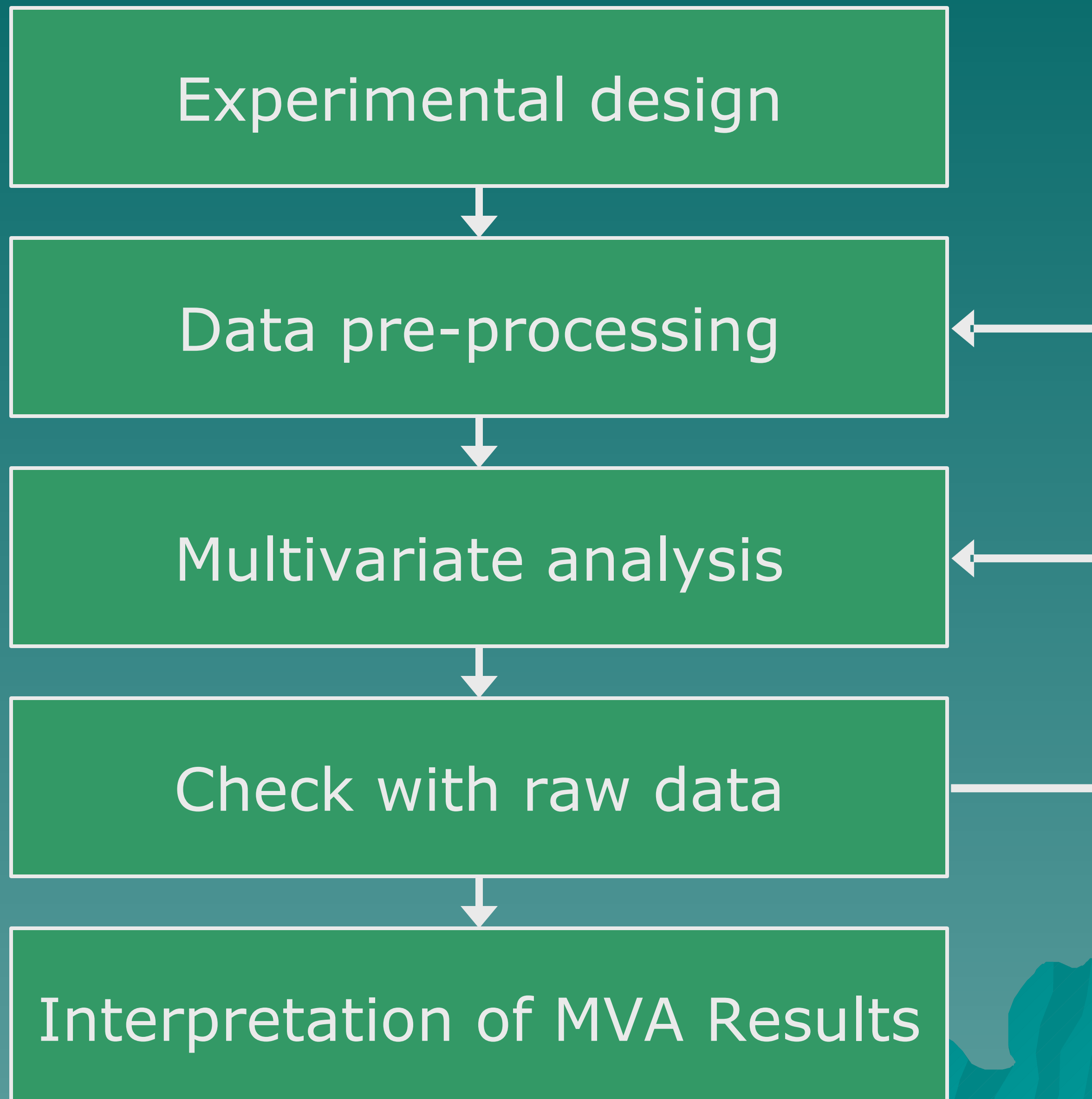
- Correct choice depends on the hypothesis being tested (and what assumptions you've made about the data)!

See talk by B. Tyler, Thursday 15-Sept, 11:20am

MVA Toolbox

- Pattern Recognition/Factor Analysis
  - Principal Component Analysis
  - Multivariate Curve Resolution
- Classification
  - Neural Networks
  - Cluster Analysis
- Regression
  - Principal Component Regression
  - Partial Least Squares Regression
- Image Analysis

MVA Process


Experimental design
Data pre-processing
Multivariate analysis
Check with raw data
Interpretation of MVA Results

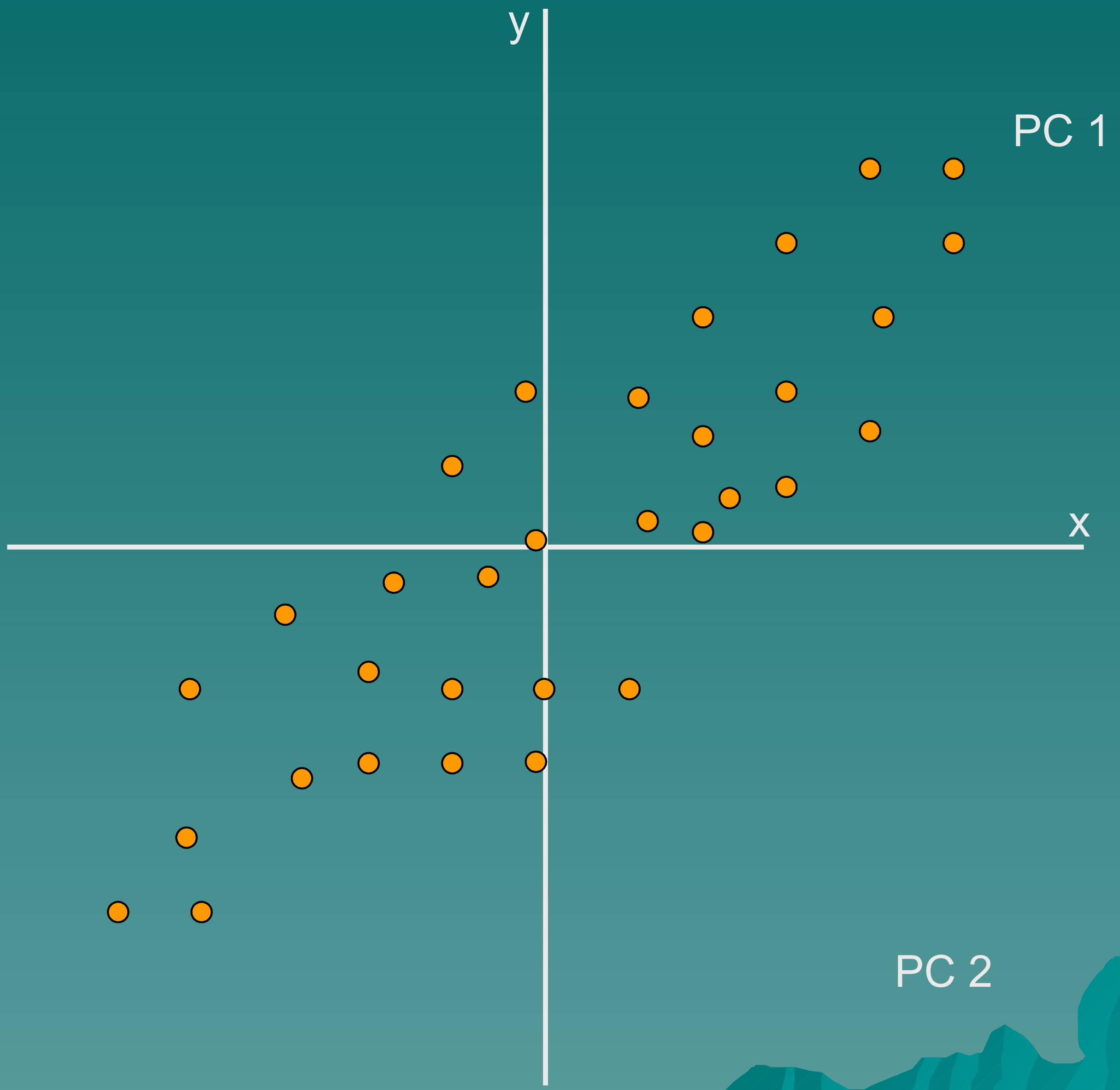
Pattern Recognition

Principal Component Analysis

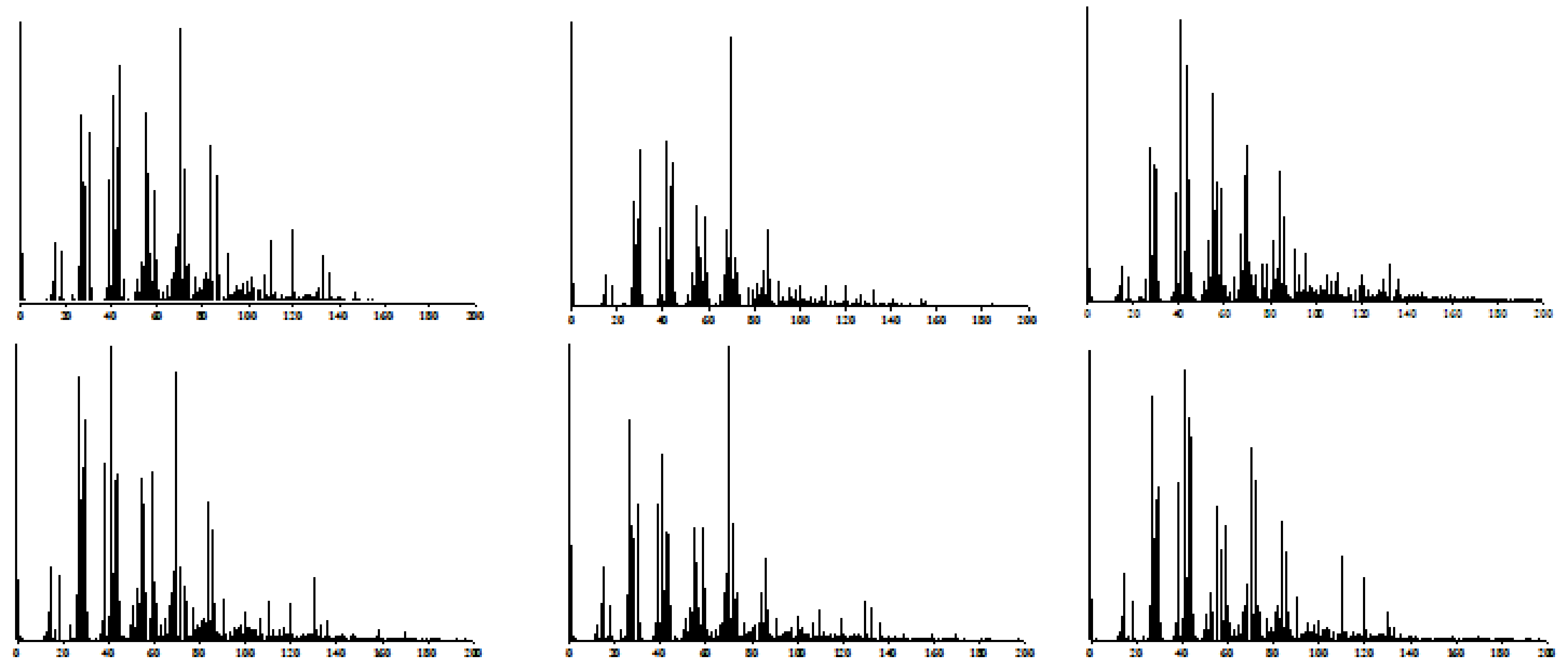
$$X = SL^T + E$$

- PCA decomposes data (X) into scores (S) and loadings (L)
- PCs capture orthogonal directions of variance
- PCA commonly used for SIMS data analysis
- For more information:
  - *Chemom. Intel. Lab. Syst.* **2**: 37 (1987)
  - J.E. Jackson *A User's Guide to Principal Components* (1991)

PCA is axis rotation



Adsorbed Proteins



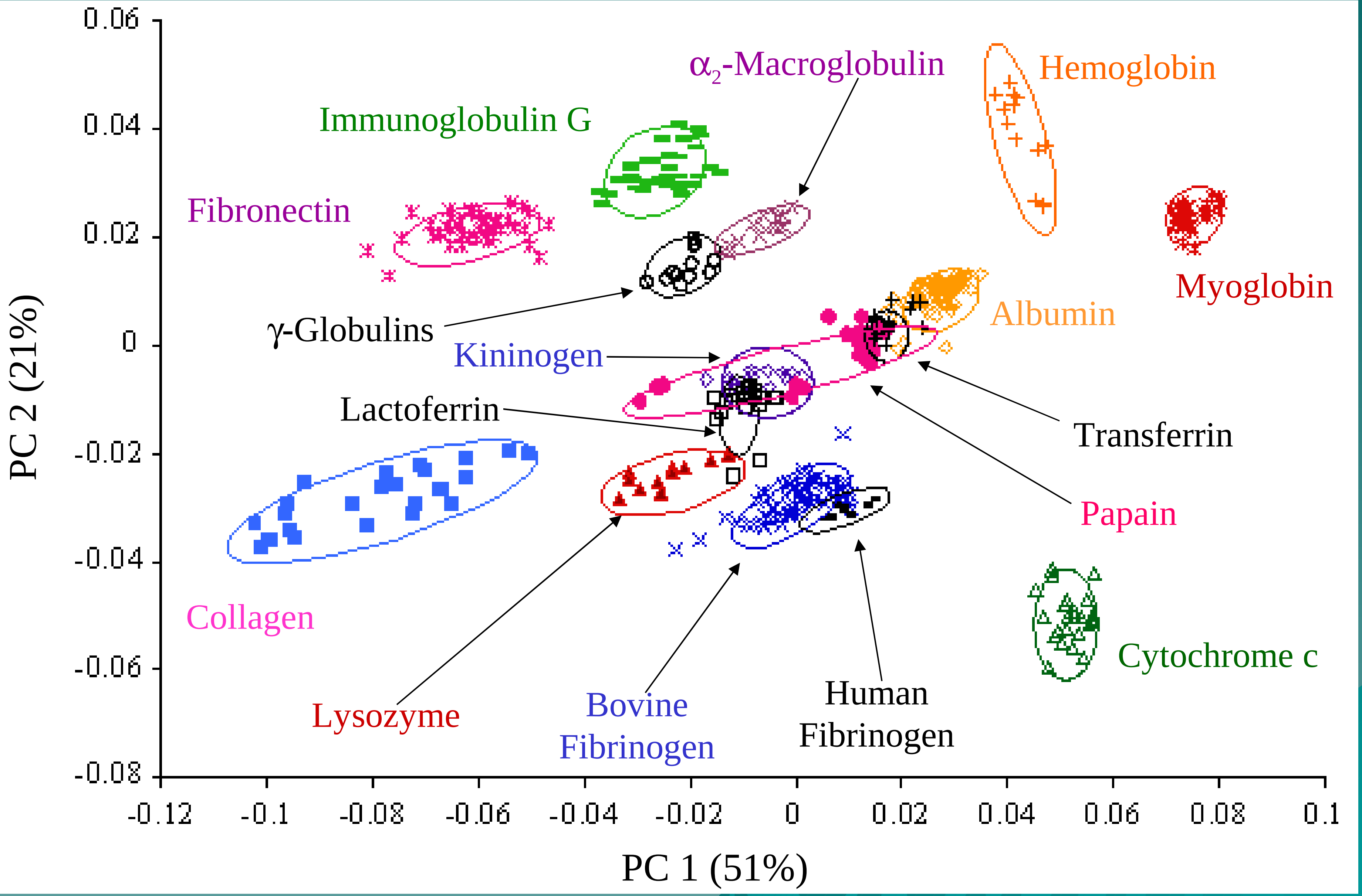
No unique, identifying peaks are present in the spectra of different adsorbed proteins.

Langmuir **17**: 4649 (2001)

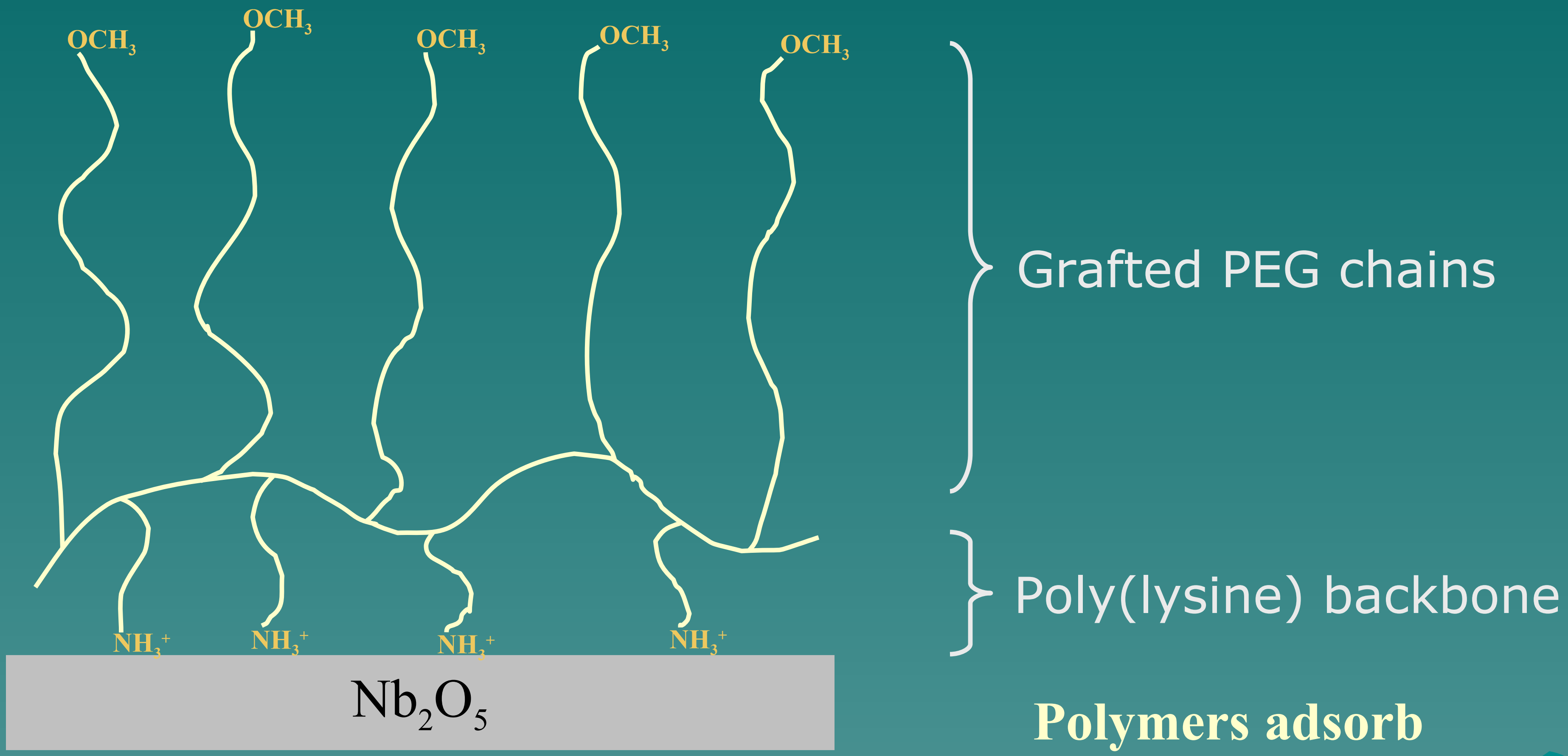
Data Pre-processing

- Amino acid-related peaks selected from positive ion spectra (37 total).
  - Inclusion of all peaks in $0 \leq m/z \leq 200$ prevented discrimination between proteins.
- ToF-SIMS spectra normalized to sum of selected peaks.
  - Assumption: Relative peak intensities are chemically important.
- Mean-centered
  - Assumption: Variance around mean is chemically important.

PCA Reduces Dimensionality



PLL-g-PEG Monolayers



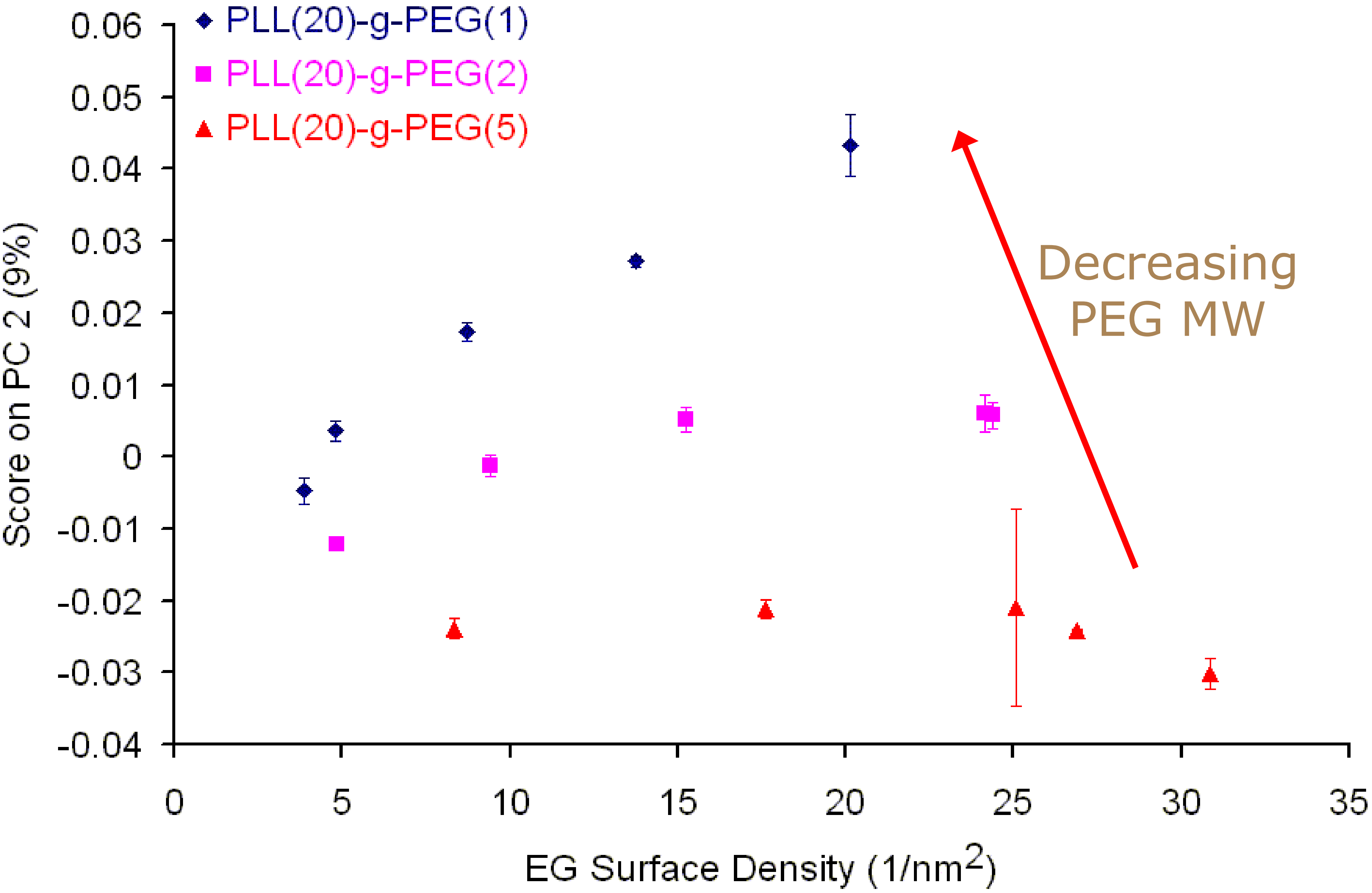
Polymers adsorb electrostatically onto negatively charged Nb_2O_5

Anal. Chem. **76**: 1483 (2004)

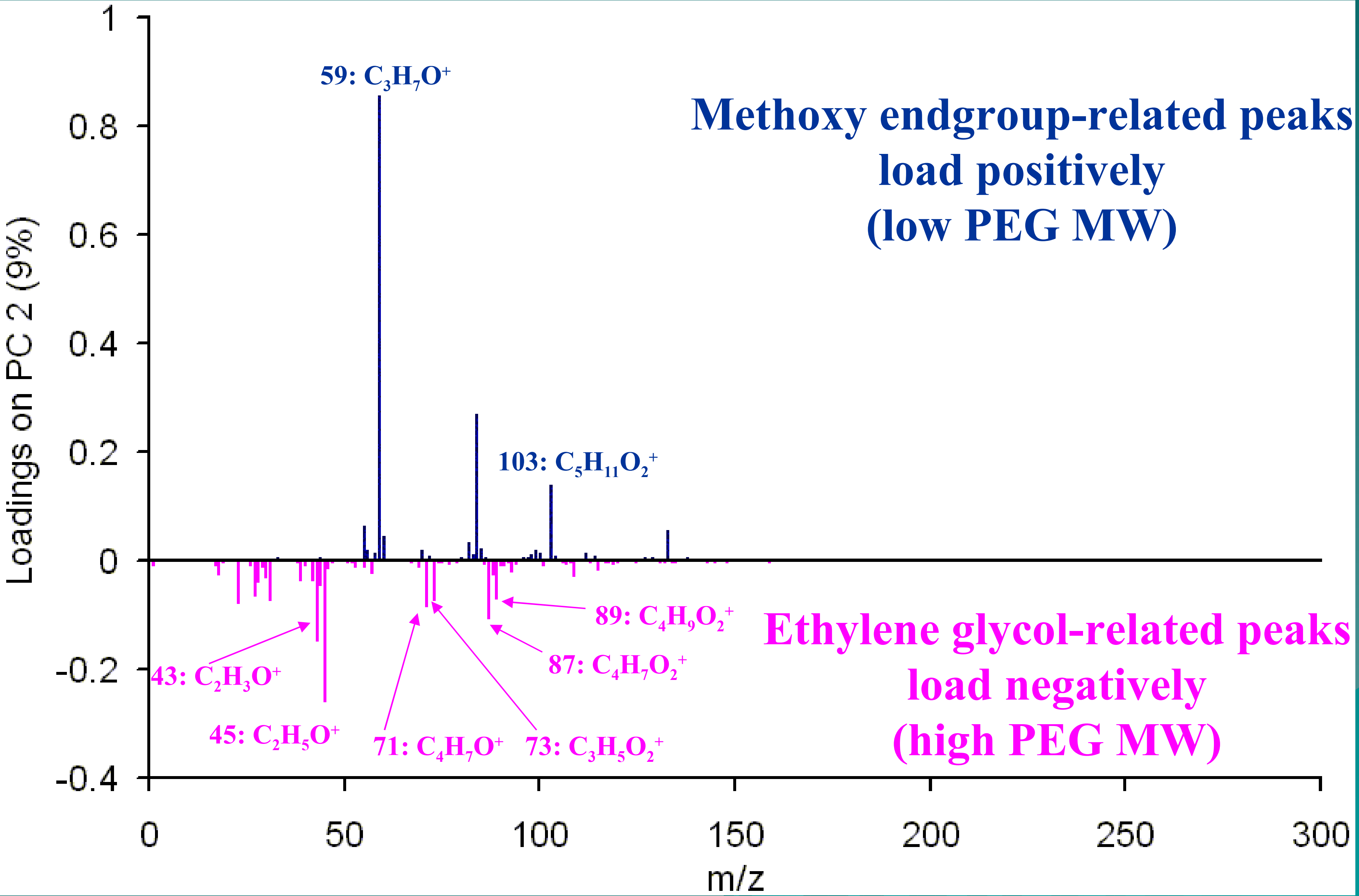
Data Pre-processing

- All peaks selected in $0 \le m/z \le 300$ range from positive ion spectra.
- ToF-SIMS spectra normalized to sum of selected peaks.
  - Assumption: Relative peak intensities are chemically important.
- Mean-centered
  - Assumption: Variance around mean is chemically important.

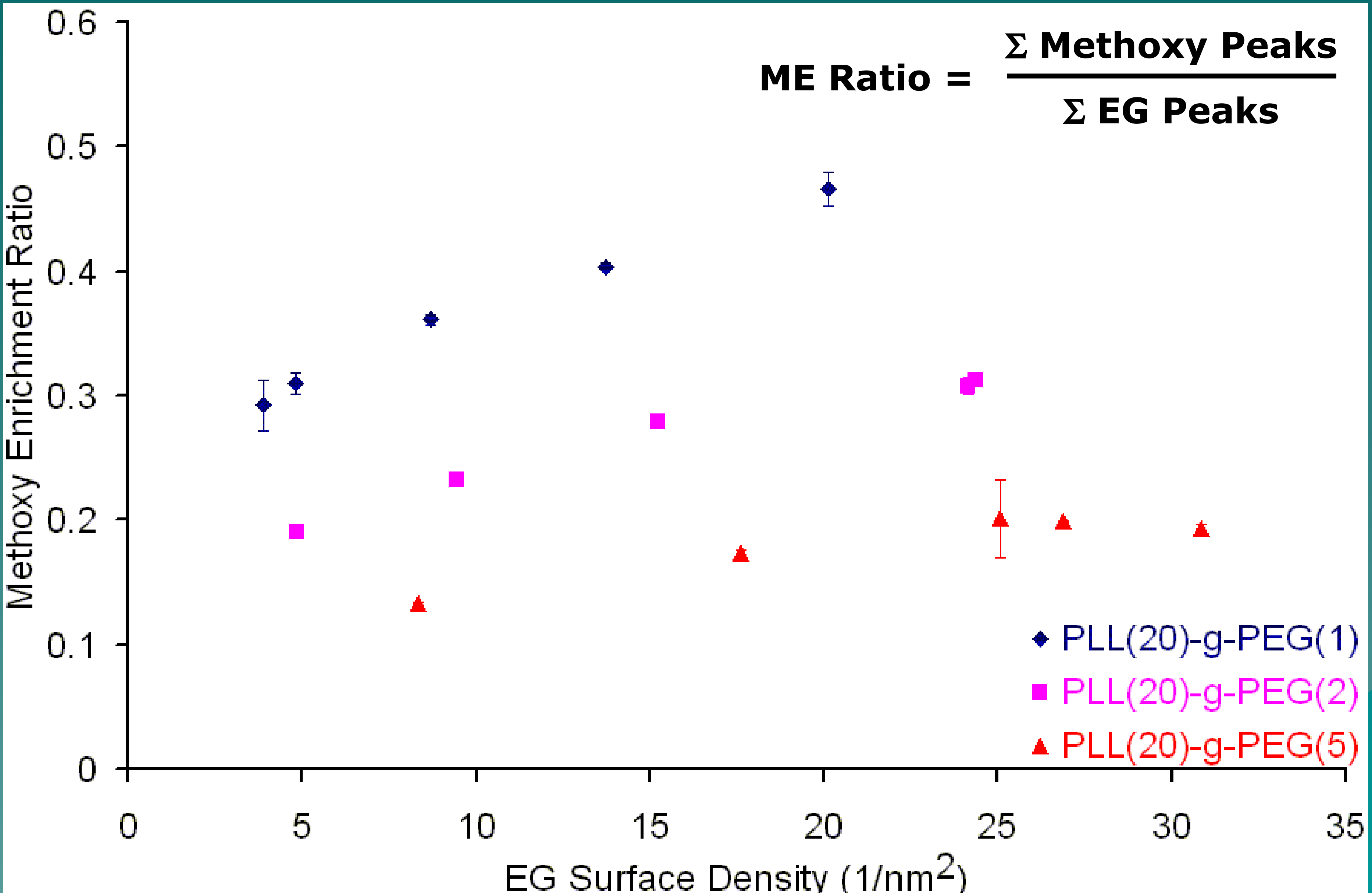
PC 2 Shows Trends w/ PEG MW



Loadings Assist Interpretation



Raw Data Confirms PCA Results



PCA Reminders

- PCA captures orthogonal directions of variance in the *pre-processed* data.
- Scores show the relationship between samples.
- Loadings show the relationship between the raw data and the PCA results.
- Check the PCA results with the raw data (especially later PCs)!

Regression

Partial Least Squares Regression

$$Y = BX + E$$

- PLSR correlates an independent variable (X) with a dependent variable (Y) via regression coefficients (B).
- PLSR maximizes correlation between X and Y
- Cross-validation important for selecting number of factors retained
- For more information:
  - *Anal. Chim. Acta* **185**: 1 (1986)

Plasma-deposited Tetraglyme

$$H_3CO\left(CH_2\text{-}CH_2\text{-}O\right)_4CH_3$$

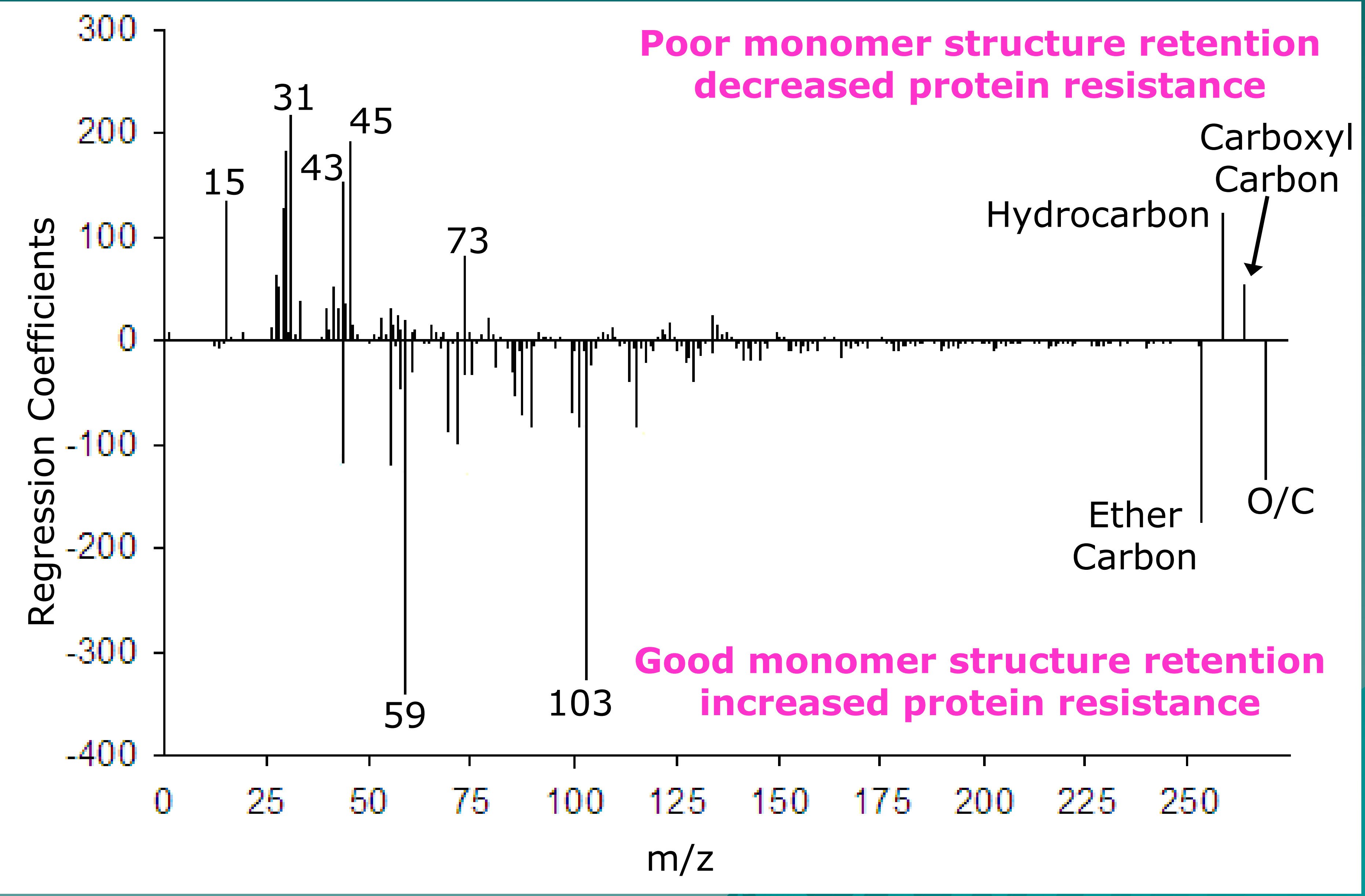
- Plasma deposition of tetraglyme monomer results in PEG-like plasma polymer.
- Reactor power determines protein resistance (higher power = more protein adsorption).
- Combination of positive ion ToF-SIMS and XPS measurements
- What differences in surface chemistry result in decreased protein resistance?

Langmuir **19**: 1692 (2003)

Data Pre-processing

- All peaks selected in 0 ≤ m/z ≤ 250 range.
- ToF-SIMS spectra normalized to most intense peak.
  - Each spectrum within the range [0 1].
- XPS data concatenated onto ToF-SIMS spectra.
  - All XPS data within the range [0 1].
- Mean-centered
  - Assumption: Variance around mean is chemically important.

RegCoeffs Explain Related Factors



PLSR Reminders

- PLSR maximizes correlation between independent and dependent variables for model dataset.
- Regression coefficients show how ToF-SIMS data relates to dependent variable.
- Cross-validation is critical for selection of appropriate number of factors, but model dataset must be appropriate for test dataset.
- Check the PLSR results (i.e. regression coefficients) with the raw data.

Image Analysis

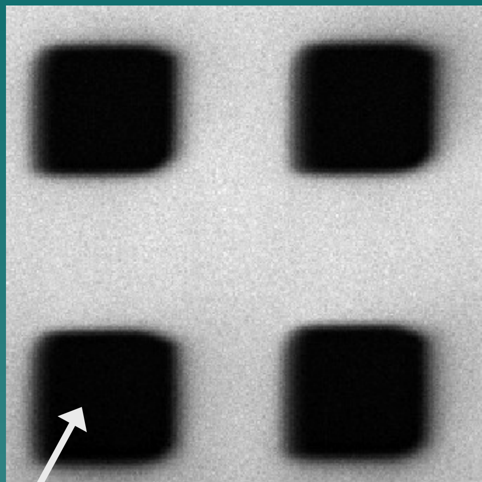
Multivariate Curve Resolution

$$X = CS^T + E$$

- MCR resolves the dataset (X) into pure component spectra (S) and concentration (C) vectors.
- Number of components and initial guess required for C or S.
- Alternating least squares with non-negativity constraints typically used.
- For more information:
  - *Chemom. Intel. Lab. Syst.* **73**: 105 (2004)

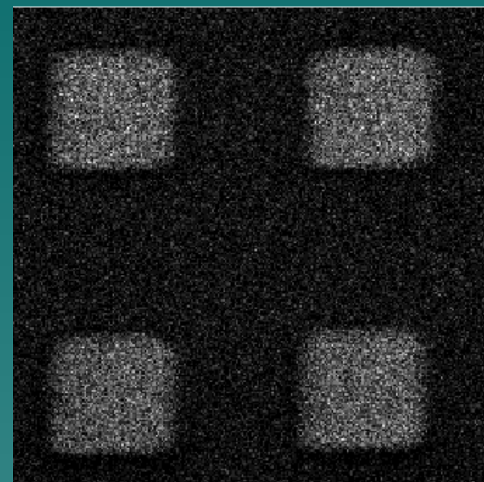
Etched Polymer (PMMA) Film

Total Ion Image



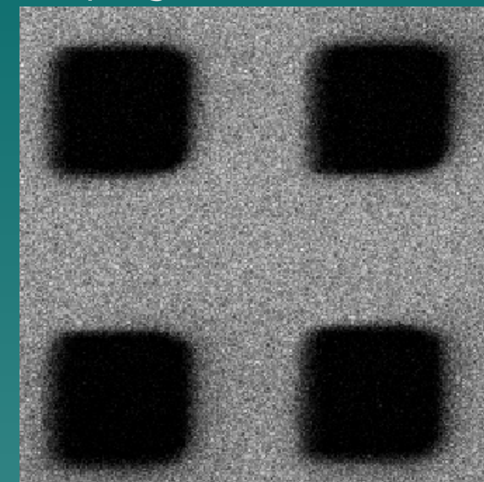
Etched region

Si^+ Image



S/N = 3.8

$C_4H_5O^+$ Image

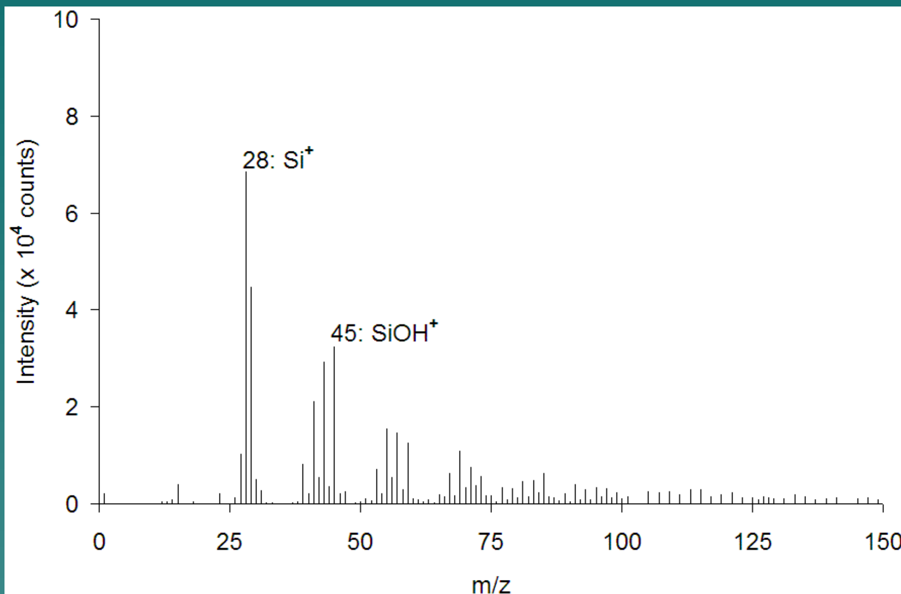


S/N = 34.0

- Image field of view: 256 μm x 256 μm,
  256 x 256 pixels
- Etched region has 24% of total pixels in image.
- Etched region has 2% of total counts in image.

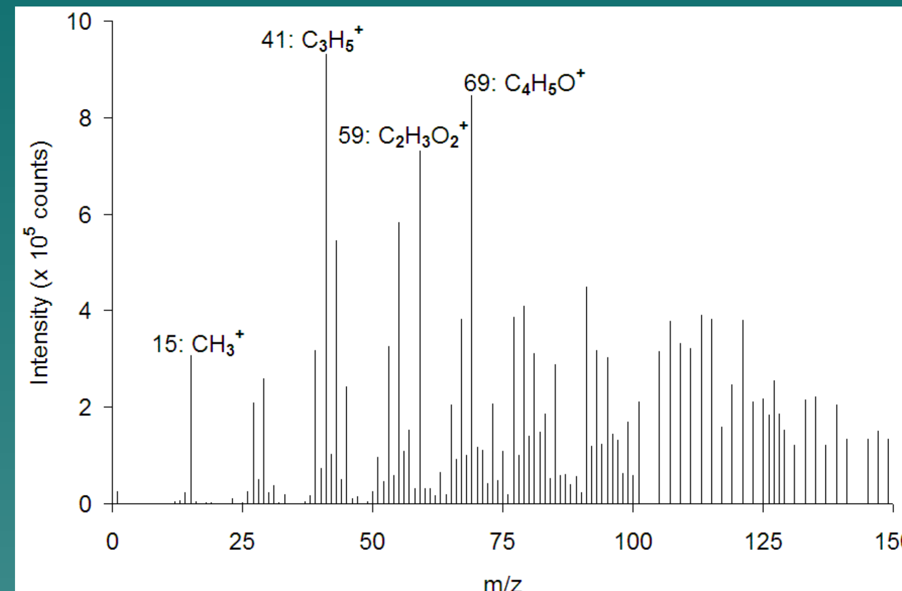
Example: Etched Polymer Film

Etched Region



Total counts: 5.6×10^5

Non-etched Region



Total counts: 2.8×10^7

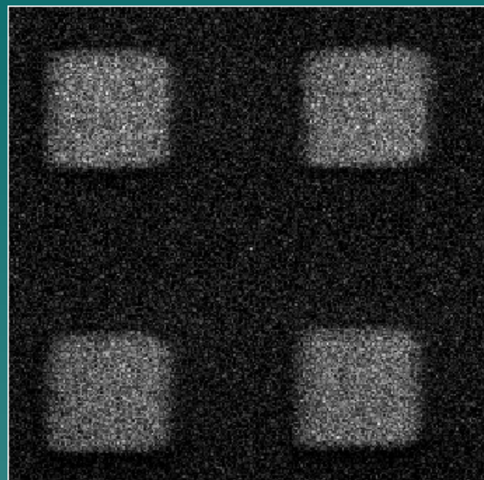
- Etched region has 24% of total pixels in image.
- Etched region has 2% of total counts in image.

Data Pre-processing

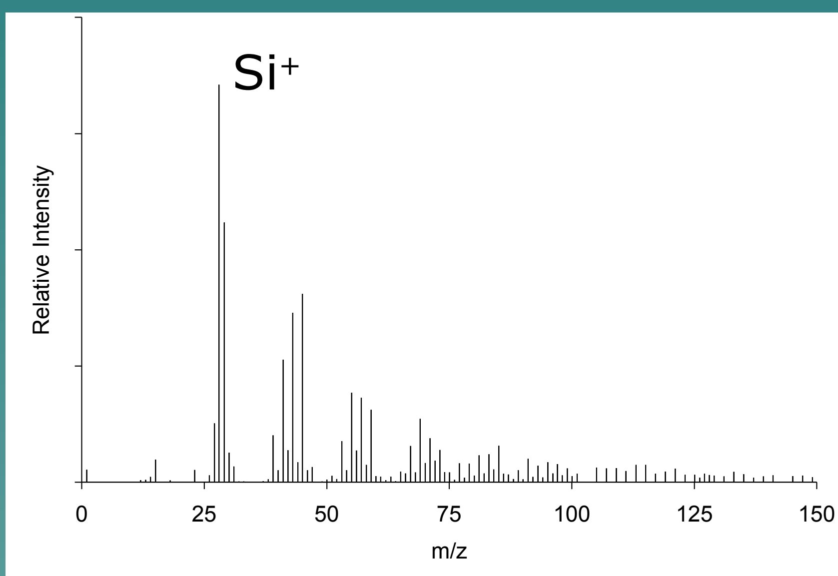
- All peaks selected in 0 ≤ m/z ≤ 150 range from positive ion image.
- ToF-SIMS dataset was scaled to minimize Poisson noise.
  - Assumption: Noise in data governed by Poisson statistics.
  - See *Surf. Interface Anal.* **36**: 203 (2004)
- MCR calculated using a ones matrix for initial spectra guess (two components fit).
- MCR results back-scaled into original spectral space.

MCR: Poisson-scaling

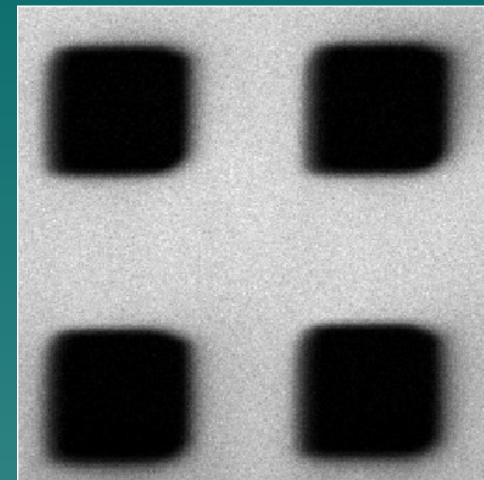
Etched Region



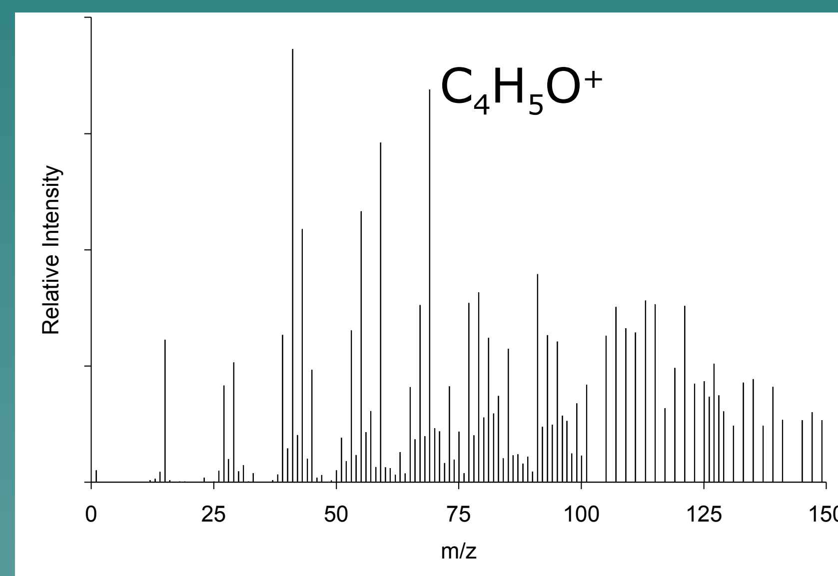
S/N = 3.9



Non-etched Region



S/N = 50.2



MCR Reminders

- MCR calculates “pure” concentration and spectrum vectors, subject to non-negativity and other constraints.
- MCR is reasonably robust to initial guess for C or S, but…
- MCR only fits the number of components you choose (choose well).
- Check the MCR results with the raw data.

Final Thoughts

Remember MVA Design!

Experimental design

↓

Data pre-processing

↓

Multivariate analysis

↓

Check with raw data

↓

Interpretation of MVA Results

Acknowledgements

Funding

- National ESCA and Surface Analysis Center for Biomedical Problems
  (NIH Grant EB002027)
- National Institute of Standards and Technology

Collaborators

- D. Graham, D. Castner, University of Washington
- S. Pasche, M. Textor, ETH-Zurich
- M. Shen, T. Horbett, B. Ratner, University of Washington

Literature Cited

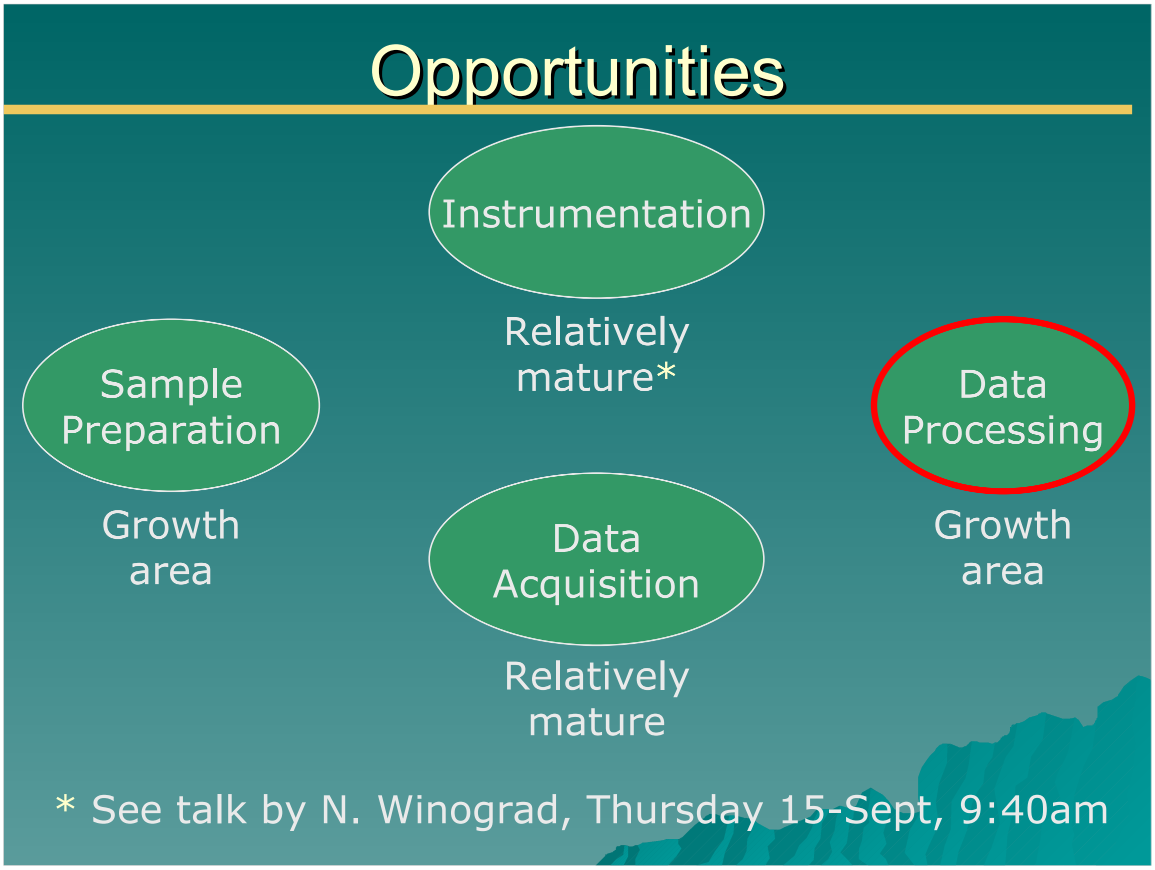
- *Surface Science* **570**: 78-97 (2004)
- *Chemometrics and Intelligent Laboratory Systems* **2**: 37 (1987)
- J.E. Jackson, *A User's Guide to Principal Components*, John Wiley & Sons: New York (1991)
- *Analytica Chimica Acta* **185**: 1 (1986)
- *Chemometrics and Intelligent Laboratory Systems* **73**: 105 (2004)
- *Surface and Interface Analysis* **36**: 203 (2004)
- *Langmuir* **17**: 4649 (2001)
- *Analytical Chemistry* **76**: 1483 (2004)
- *Langmuir* **19**: 1692 (2003)

Simplifying the Interpretation of ToF-SIMS Spectra and Images using Careful Application of Multivariate Analysis

Matthew S. Wagner
The Procter & Gamble Company
wagner.ms@pg.com

SIMS XV
September 13, 2005

Please note that multivariate analysis *simplifies* the interpretation; it does not interpret the data for you. Also note the “careful application” – it is critical that MVA users understand the capabilities and limitations of MVA for interpreting SIMS data.



Sample preparation = sectioning, cryopreservation, sugar-coating

Instrumentation = ion sources, MS, electronics, etc.

Data acquisition = automated data collection, large area imaging

Goal for Data Analysis

Concise and accurate chemical description of surface chemistry

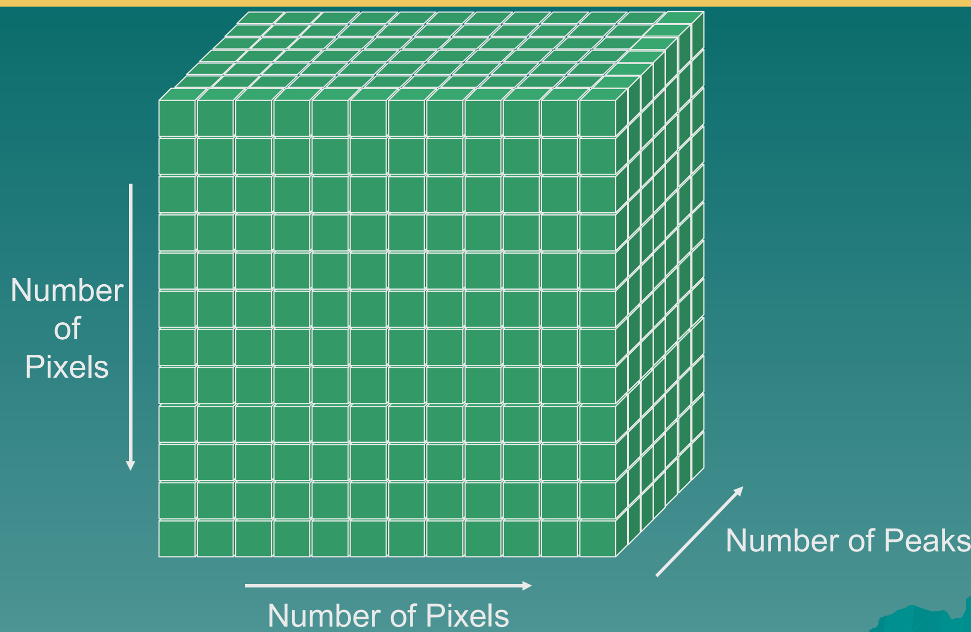
We’d like to condense the SIMS spectra into something more compact and easier to interpret. “Accurate” includes statistically relevant.

Data Processing: Challenges

- Data overload
  - Large spectral and image datasets

- Use of Multivariate Analysis (MVA)
  - When is it appropriate?
  - Appropriate experimental design?
  - Appropriate pre-processing?
  - Which MVA method is best?
  - Validation of MVA methods?
  - Accurate interpretation with physically meaningful results?

Data Overload: Too many spectra!



How do we compare *multiple spectra* on the basis of *multiple peaks* in each spectrum?

Data Overload

- Generating data is (relatively) easy…

  Efficiently processing the data
  is the challenge!

- Many peaks in a spectrum…
- Peak intensities are correlated…
- Need to process spectra rapidly…
- Images present even more challenges…
  - Low signal-to-noise...
  - Large number of pixels…
  - Comparison of multiple images…

Multivariate Analysis Benefits

- Can *simplify* data analysis…
- Many examples of MVA application to SIMS data…
  - See *Surf. Sci.* **570**: 78 (2004)
- Requires good understanding of the analytical tool…

MVA: Not a Black Box!!!

MVA *is not*:

- A "black-box" tool for data analysis.
- A substitute for a skilled analyst.
- A substitute for poor experimental design.
- "Magic".

MVA *is*:

- An important and useful tool for saving the analyst time and money.
- An important and useful tool for maximizing the use of your data!

Before MVA: Data Pre-processing

- Many types of pre-treatment possible:
  - Peak selection
  - Normalization (this is a type of scaling)
  - Mean-centering, Autoscaling, Log-scaling, Mean-scaling, Poisson-scaling, etc.

- All data pre-treatments involve assumptions about the data!

- No standard method exists to determine which is best!
  - Trial-and-error approach widely used…

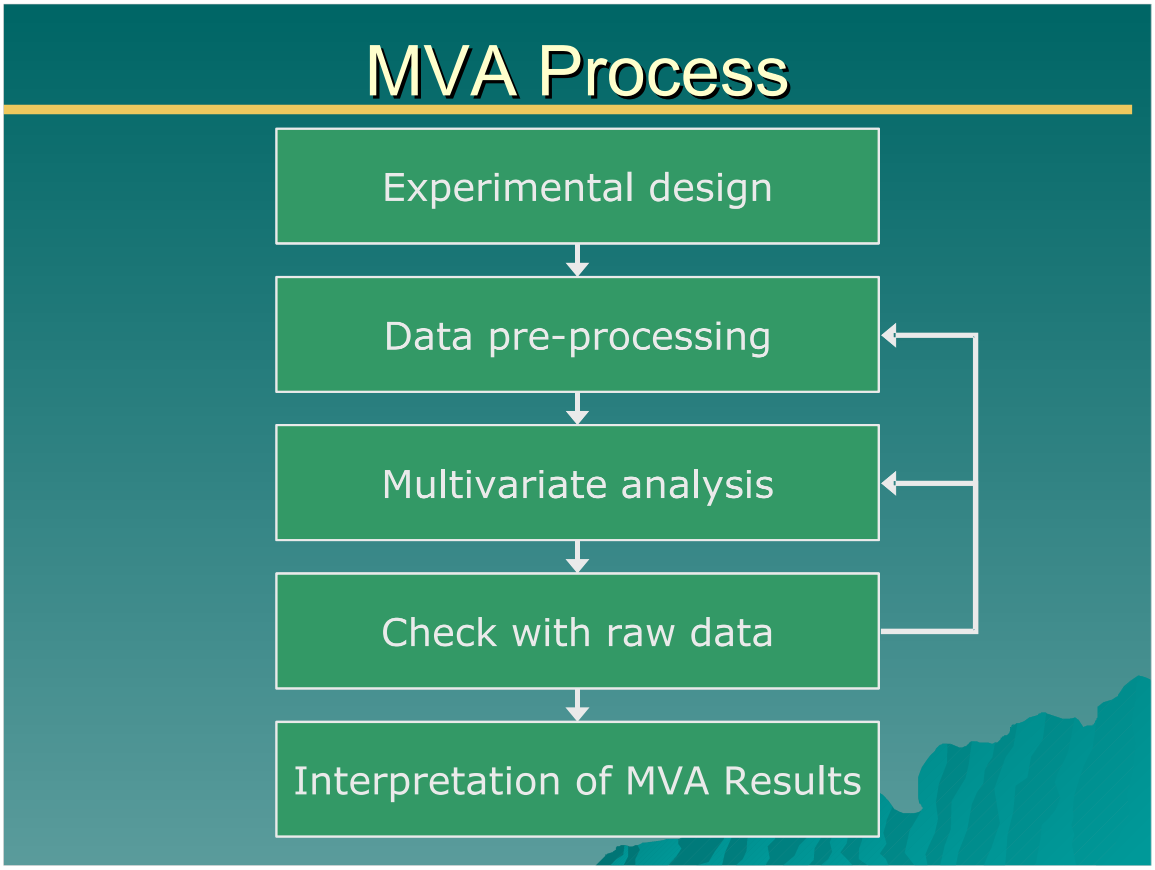
- Correct choice depends on the hypothesis being tested (and what assumptions you've made about the data)!

  See talk by B. Tyler, Thursday 15-Sept, 11:20am

MVA Toolbox

- Pattern Recognition/Factor Analysis
  - Principal Component Analysis
  - Multivariate Curve Resolution
- Classification
  - Neural Networks
  - Cluster Analysis
- Regression
  - Principal Component Regression
  - Partial Least Squares Regression
- Image Analysis

Pick the right tool for the job.



Once you’re sure the answer makes sense mathematically, you can then interpret the results physically.

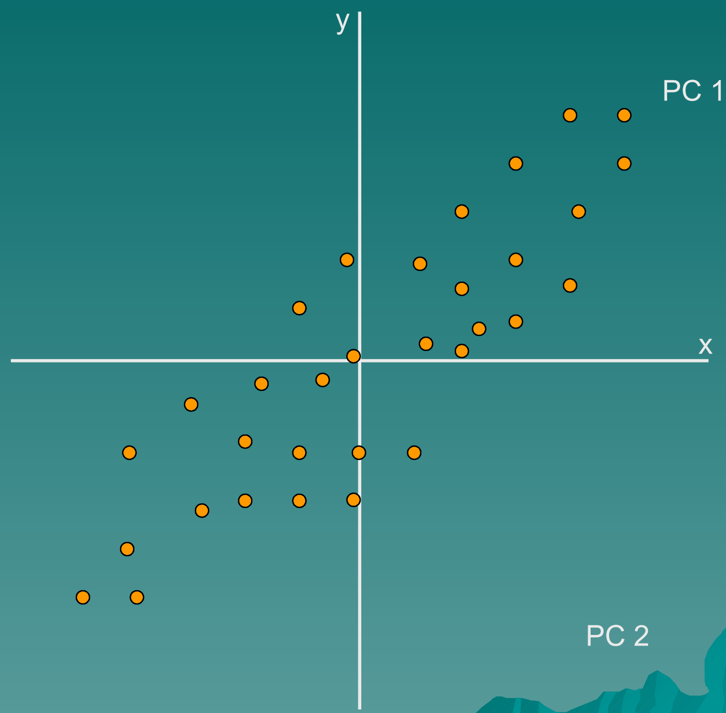
Pattern Recognition

Principal Component Analysis

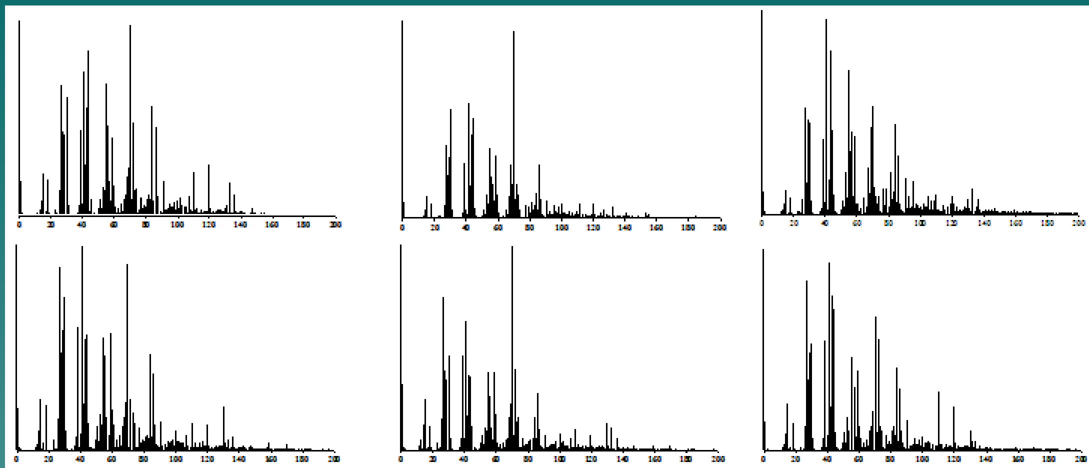
$$X = SL^T + E$$

- PCA decomposes data (X) into scores (S) and loadings (L)
- PCs capture orthogonal directions of variance
- PCA commonly used for SIMS data analysis
- For more information:
  - *Chemom. Intel. Lab. Syst.* **2**: 37 (1987)
  - J.E. Jackson *A User's Guide to Principal Components* (1991)

PCA is axis rotation



Adsorbed Proteins



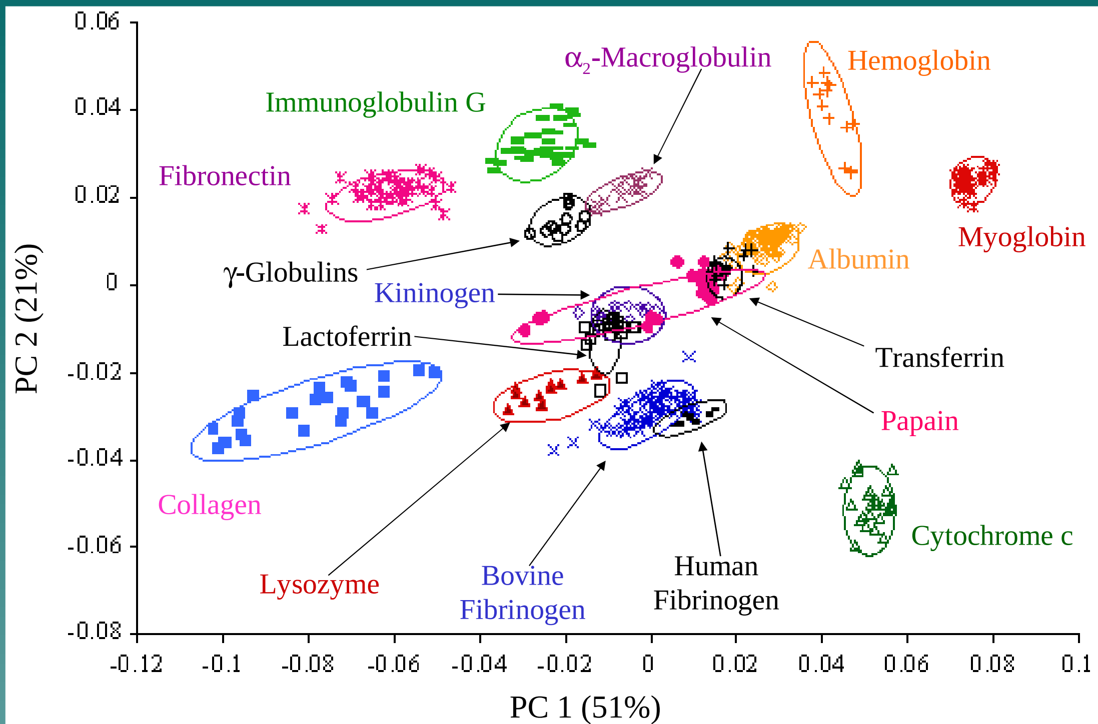
No unique, identifying peaks are present in the spectra of different adsorbed proteins.

Langmuir **17**: 4649 (2001)

Data Pre-processing

- Amino acid-related peaks selected from positive ion spectra (37 total).
  - Inclusion of all peaks in $0 \leq m/z \leq 200$ prevented discrimination between proteins.
- ToF-SIMS spectra normalized to sum of selected peaks.
  - Assumption: Relative peak intensities are chemically important.
- Mean-centered
  - Assumption: Variance around mean is chemically important.

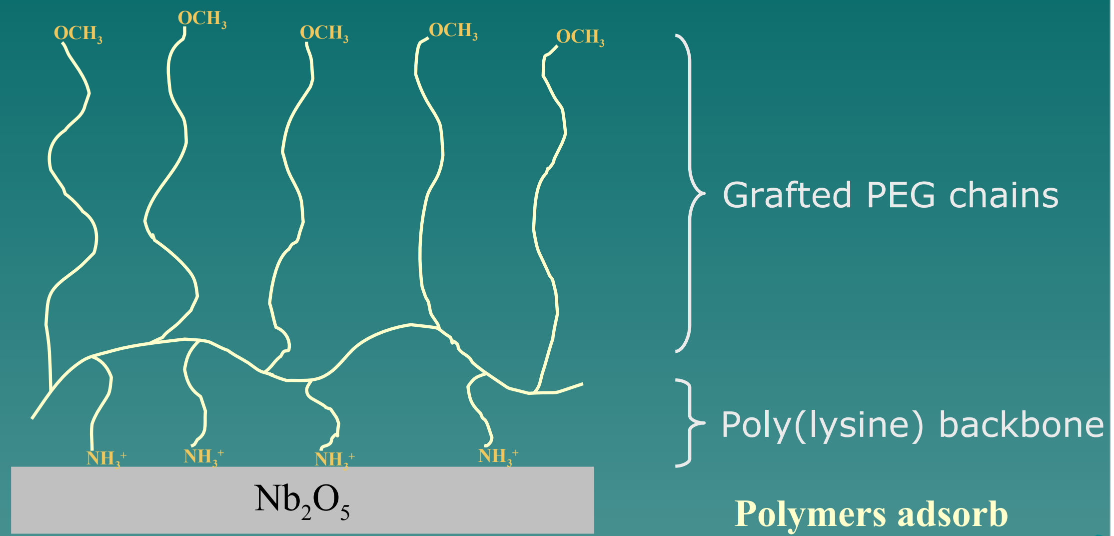
PCA Reduces Dimensionality



PCA provides:

2) Quick comparison of multiple spectra on the basis of multiple peaks in each spectrum.
3) This shows that the spectra of the different proteins are different from each other.
4) This also shows the relative reproducibility of the spectra of the different proteins.
5) Loadings give insight into amino acid composition of the proteins (data not shown).

PLL-g-PEG Monolayers

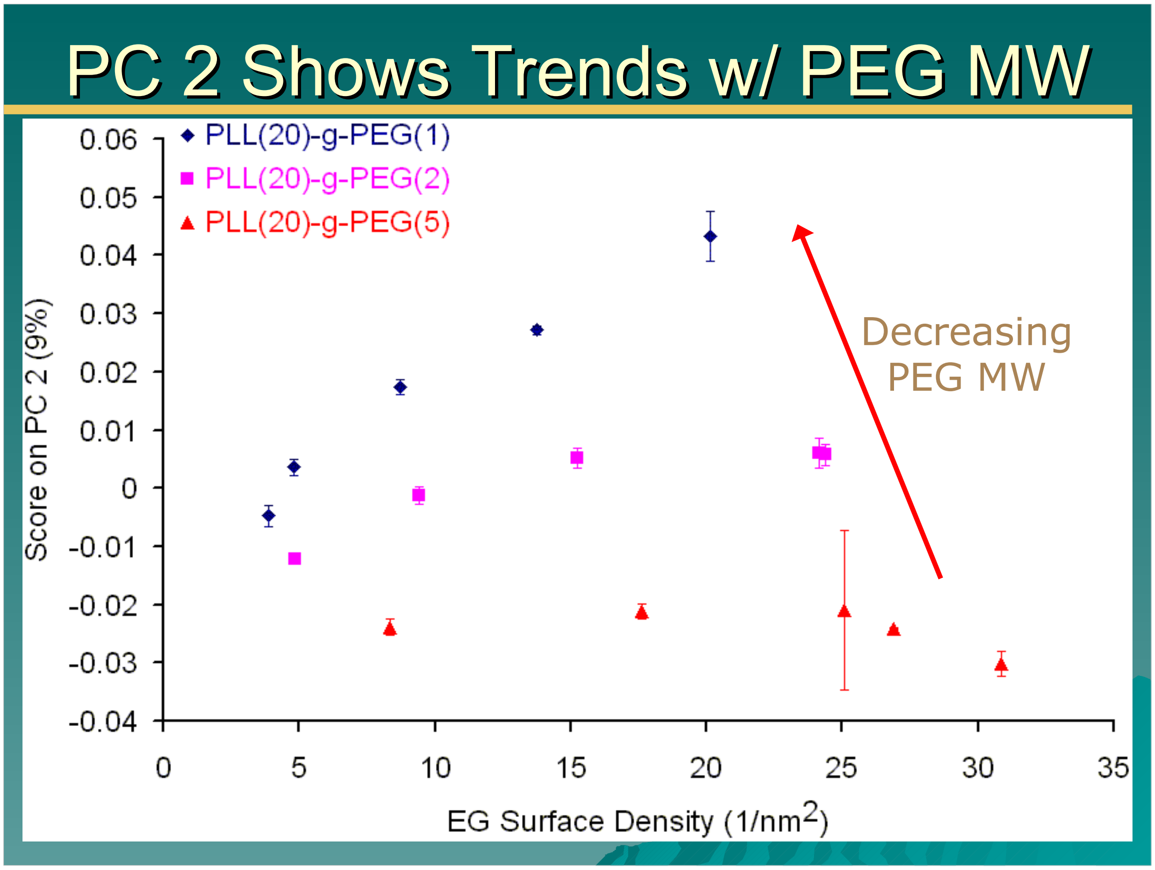


Polymers adsorb electrostatically onto negatively charged Nb_2O_5

Anal. Chem. **76**: 1483 (2004)

Data Pre-processing

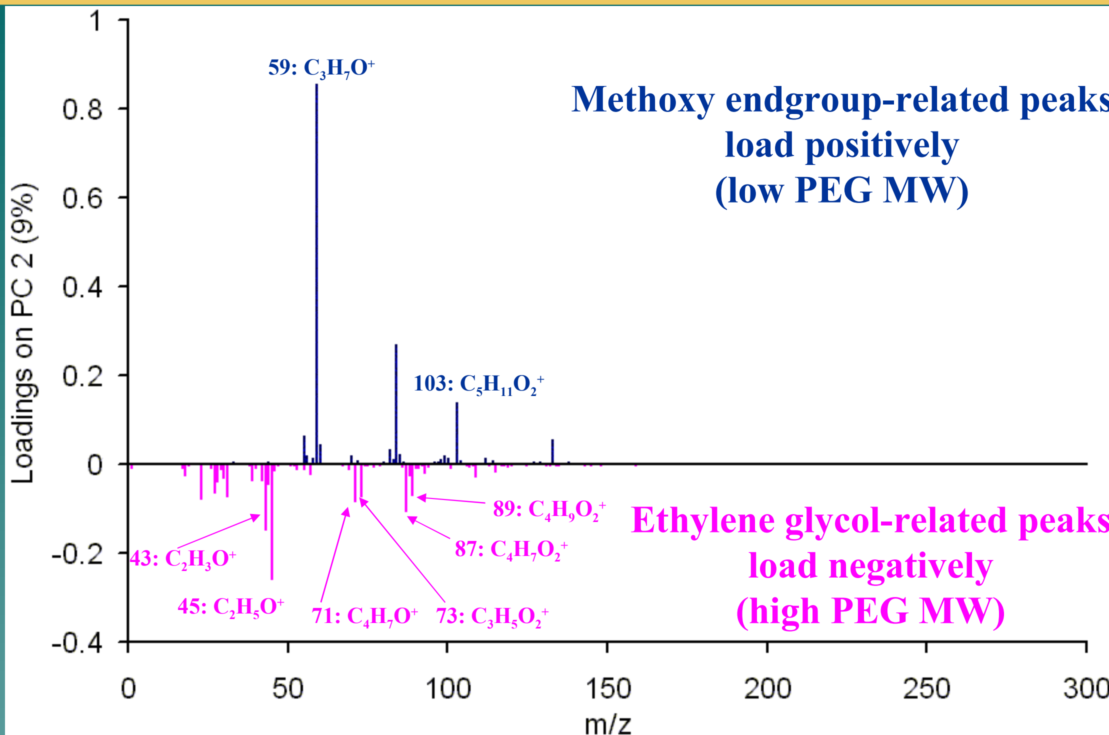
- All peaks selected in $0 \leq m/z \leq 300$ range from positive ion spectra.
- ToF-SIMS spectra normalized to sum of selected peaks.
  - Assumption: Relative peak intensities are chemically important.
- Mean-centered
  - Assumption: Variance around mean is chemically important.



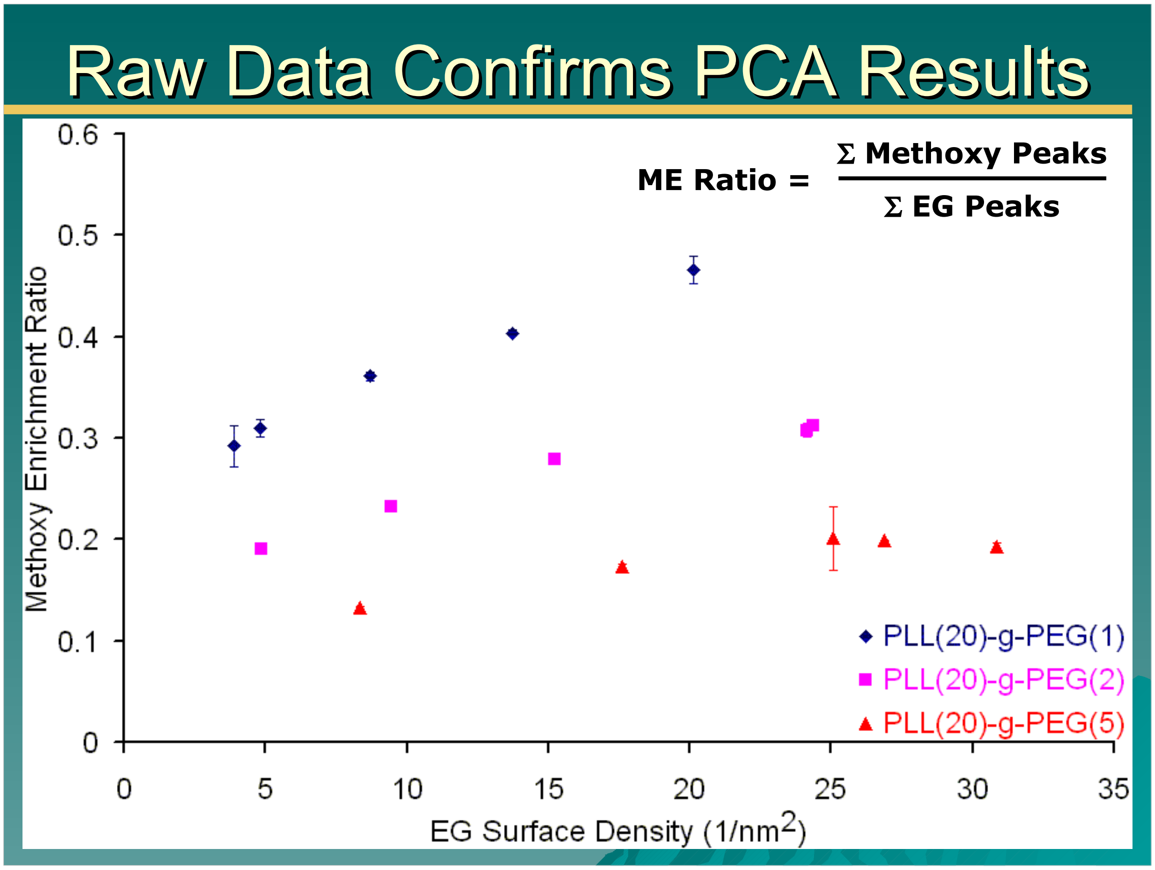
PC 1 showed trend with PEG graft density (not shown).

These scores are on PC 2, and show differences in methoxy headgroup surface concentration.

Loadings Assist Interpretation



Peaks from poly(L-lysine) (C5H10N+, 84) and Cs+ also load positively (i.e. correlated with low MW PEGs) due to thinner PEG layer.



Note that this plot uses the raw data, not the data after PC 1 has been subtracted. Correlation of later PCs with the raw data may require the variance captured in PC 1 to be subtracted. This can be considered “filtering” high directions of variance out to look at more subtle features.

PCA Reminders

- PCA captures orthogonal directions of variance in the *pre-processed* data.
- Scores show the relationship between samples.
- Loadings show the relationship between the raw data and the PCA results.
- Check the PCA results with the raw data (especially later PCs)!

Regression

Partial Least Squares Regression

$$Y = BX + E$$

- PLSR correlates an independent variable (X) with a dependent variable (Y) via regression coefficients (B).
- PLSR maximizes correlation between X and Y
- Cross-validation important for selecting number of factors retained
- For more information:
  - *Anal. Chim. Acta* **185**: 1 (1986)

Plasma-deposited Tetraglyme

$$H_3CO\text{-}\left(CH_2\text{-}CH_2\text{-}O\right)_4\text{-}CH_3$$

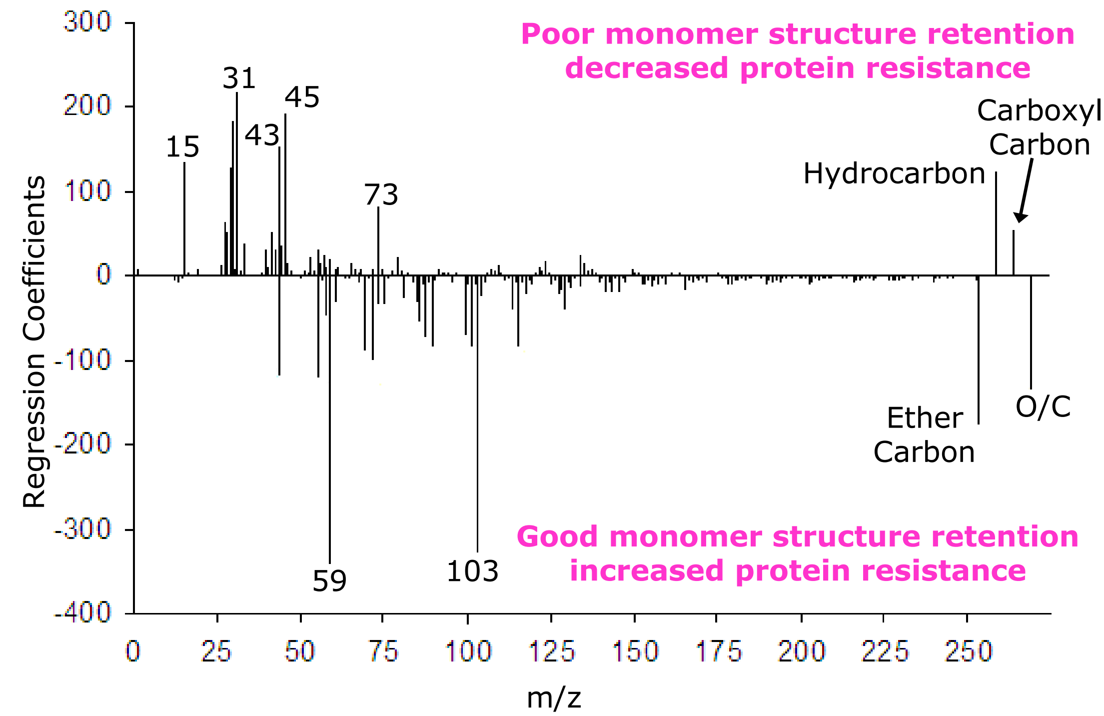
- Plasma deposition of tetraglyme monomer results in PEG-like plasma polymer.
- Reactor power determines protein resistance (higher power = more protein adsorption).
- Combination of positive ion ToF-SIMS and XPS measurements
- What differences in surface chemistry result in decreased protein resistance?

Langmuir **19**: 1692 (2003)

Data Pre-processing

- All peaks selected in $0 \le m/z \le 250$ range.
- ToF-SIMS spectra normalized to most intense peak.
  - Each spectrum within the range [0 1].
- XPS data concatenated onto ToF-SIMS spectra.
  - All XPS data within the range [0 1].
- Mean-centered
  - Assumption: Variance around mean is chemically important.

RegCoeffs Explain Related Factors



PLSR Reminders

- PLSR maximizes correlation between independent and dependent variables for model dataset.
- Regression coefficients show how ToF-SIMS data relates to dependent variable.
- Cross-validation is critical for selection of appropriate number of factors, but model dataset must be appropriate for test dataset.
- Check the PLSR results (i.e. regression coefficients) with the raw data.

Image Analysis

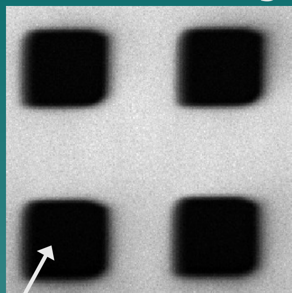
Multivariate Curve Resolution

$$X = CS^T + E$$

- MCR resolves the dataset (X) into pure component spectra (S) and concentration (C) vectors.
- Number of components and initial guess required for C or S.
- Alternating least squares with non-negativity constraints typically used.
- For more information:
  - *Chemom. Intel. Lab. Syst.* **73**: 105 (2004)

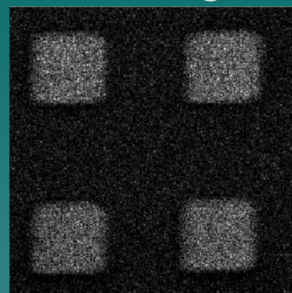
Etched Polymer (PMMA) Film

Total Ion Image



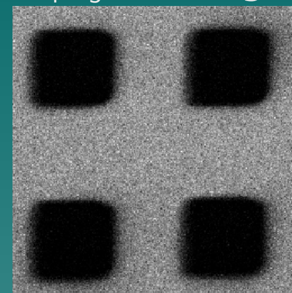
Etched region

Si^+ Image



S/N = 3.8

$C_4H_5O^+$ Image

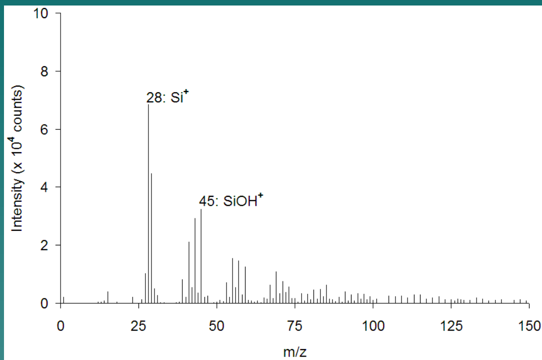


S/N = 34.0

- Image field of view: 256 μm x 256 μm,
  256 x 256 pixels
- Etched region has 24% of total pixels in image.
- Etched region has 2% of total counts in image.

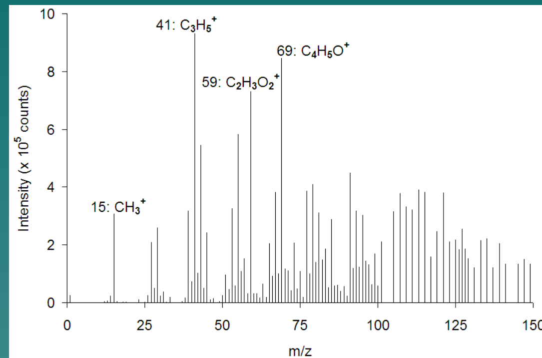
Example: Etched Polymer Film

Etched Region



Total counts: 5.6×10^5

Non-etched Region



Total counts: 2.8×10^7

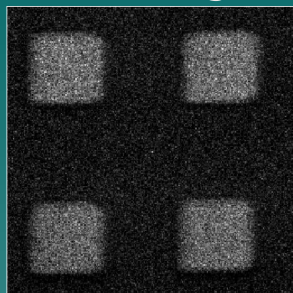
- Etched region has 24% of total pixels in image.
- Etched region has 2% of total counts in image.

Data Pre-processing

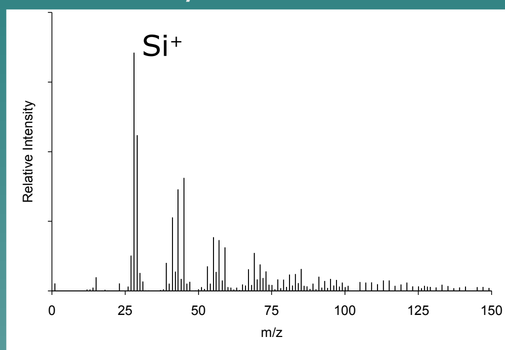
- All peaks selected in $0 \leq m/z \leq 150$ range from positive ion image.
- ToF-SIMS dataset was scaled to minimize Poisson noise.
  - Assumption: Noise in data governed by Poisson statistics.
  - See *Surf. Interface Anal.* **36**: 203 (2004)
- MCR calculated using a ones matrix for initial spectra guess (two components fit).
- MCR results back-scaled into original spectral space.

MCR: Poisson-scaling

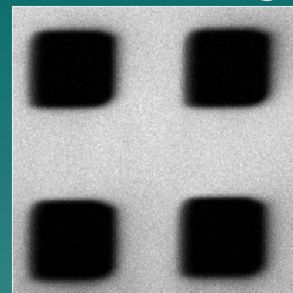
Etched Region



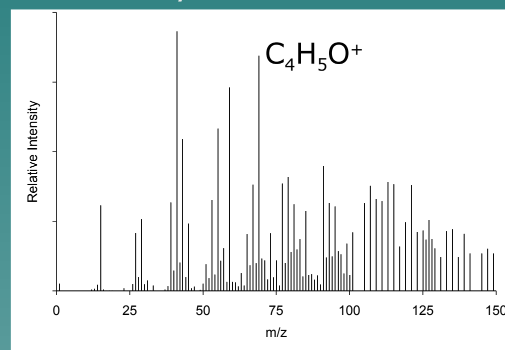
S/N = 3.9



Non-etched Region



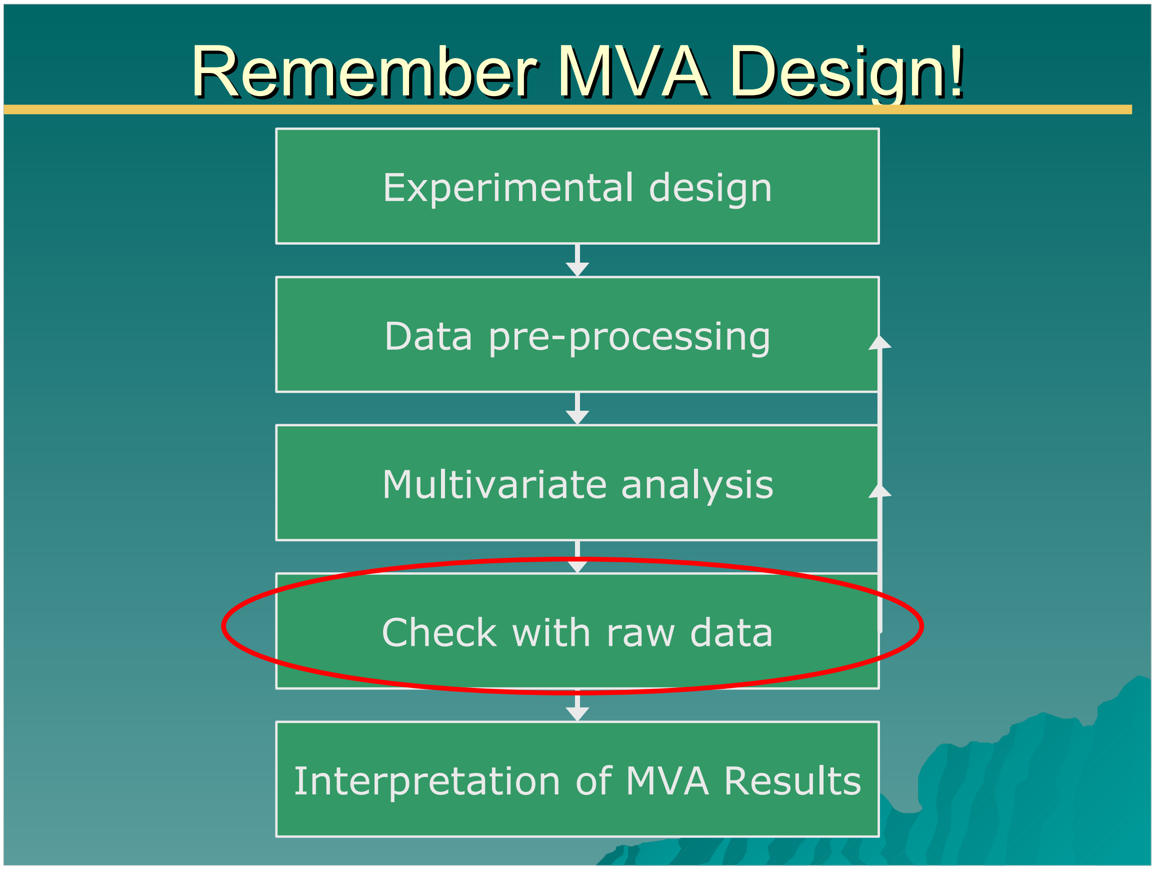
S/N = 50.2



MCR Reminders

- MCR calculates "pure" concentration and spectrum vectors, subject to non-negativity and other constraints.
- MCR is reasonably robust to initial guess for C or S, but…
- MCR only fits the number of components you choose (choose well).
- Check the MCR results with the raw data.

Final Thoughts



Once you’re sure the answer makes sense mathematically, you can then interpret the results physically.

Acknowledgements

Funding

- National ESCA and Surface Analysis Center for Biomedical Problems

  (NIH Grant EB002027)
- National Institute of Standards and Technology

Collaborators

- D. Graham, D. Castner, University of Washington
- S. Pasche, M. Textor, ETH-Zurich
- M. Shen, T. Horbett, B. Ratner, University of Washington

Literature Cited

- *Surface Science* **570**: 78-97 (2004)
- *Chemometrics and Intelligent Laboratory Systems* **2**: 37 (1987)
- J.E. Jackson, *A User's Guide to Principal Components*, John Wiley & Sons: New York (1991)
- *Analytica Chimica Acta* **185**: 1 (1986)
- *Chemometrics and Intelligent Laboratory Systems* **73**: 105 (2004)
- *Surface and Interface Analysis* **36**: 203 (2004)
- *Langmuir* **17**: 4649 (2001)
- *Analytical Chemistry* **76**: 1483 (2004)
- *Langmuir* **19**: 1692 (2003)