

SVD-based Functional ANOVA For Measurement Evaluation of MALDI- TOF Mass Spectrometry of Polymers*

Z.Q. John Lu (John.lu@nist.gov)

Statistical Engineering Division, ITL

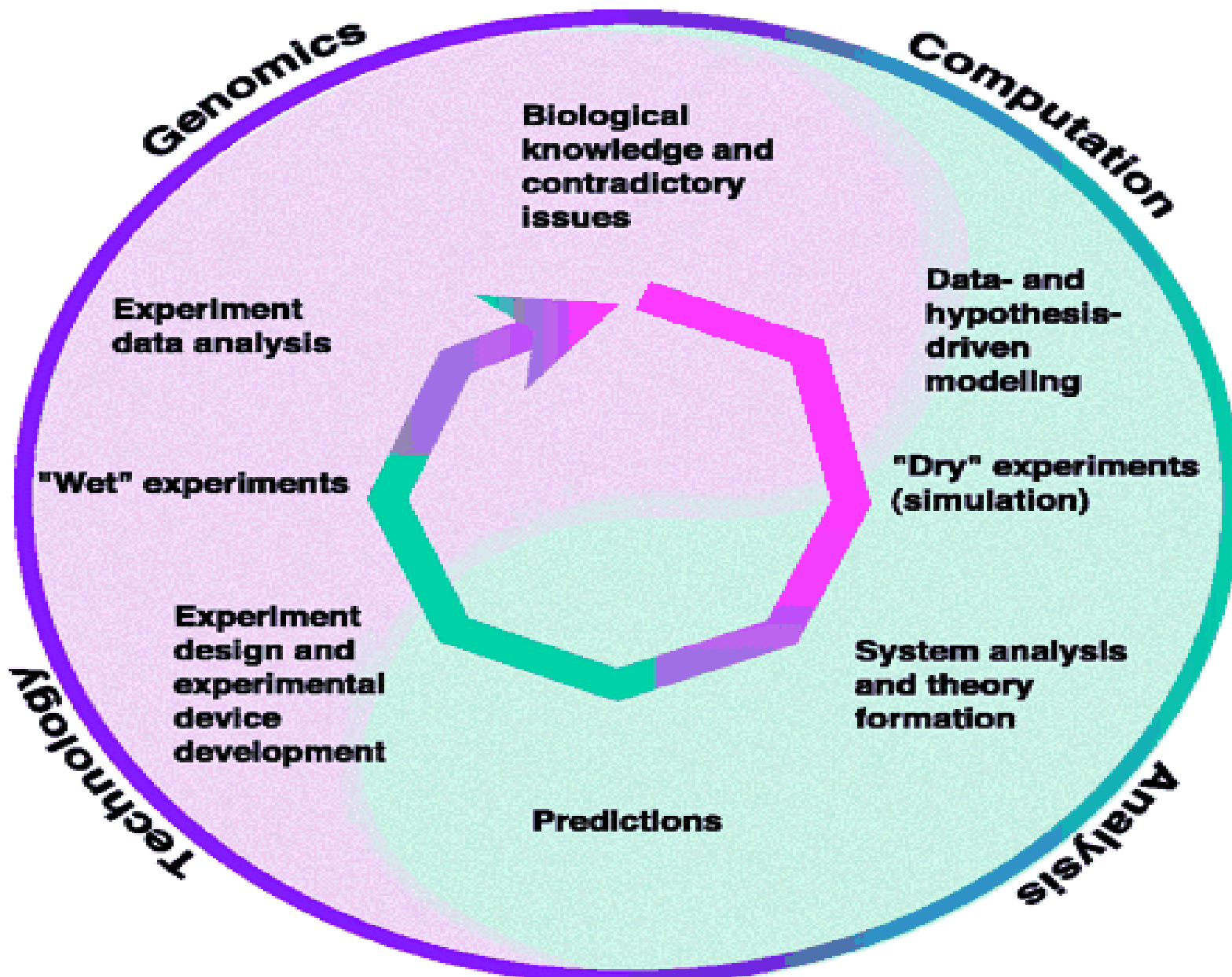
National Institute of Standards and Technology

***Acknowledgement:** Presentation based in part on collaboration with Charles Guttman (NIST), Stephanie Wetzel (NIST), Jennifer Hockett (Iowa State).

I. Background

- 1. Statistics in Biology:** beginning of mathematical statistics (R A Fisher); statistical genetics; genomics (microarrays), proteomics, system biology (Nobert Wiener).
- 2. High-throughput experiments and modern metrology:** large data (many variables, high p), and unfortunately not many samples (low n)!

H. Kitano 2002, Science, System Biology

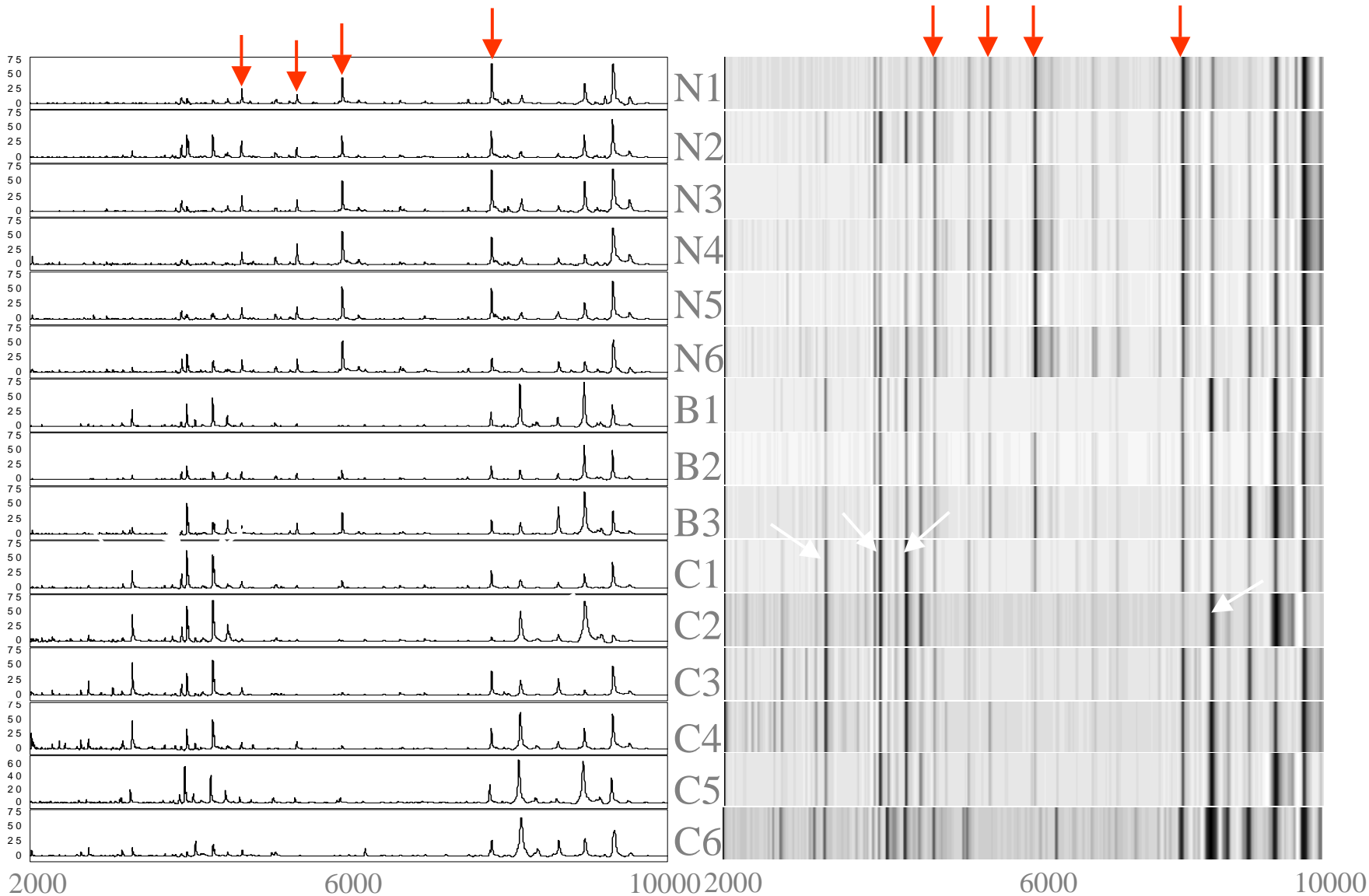


3. Biomarker hunting from serum mass spectra

- Qu, Y, B. Adam, et al (2003): Biometrics, 59, 143-151.
 - $P=48,538$, $n=248$ (167 cancer patients and 81 controls)
 - Test samples= 45
- Claim: data too large for PCA or SVD
 - Opts to use wavelet transform, K-L information criterion for variable selections, then apply Fisher's discriminant analysis, evaluating classifier: sensitivity=98%, specificity=99%.

SELDI Serum Protein Profile Analysis- Prostate Cancer

Bao-ling Adam et al 2004 Interface



II. Statistical and Metrological Issues

- High-dimensional pattern recognition and regression modeling (prediction)
 - What are the underlying assumptions for pattern recognitions such as the simple hyperplane classifier?
 - Hidden and intrinsic low-dimensional predictors and classifying variables?
 - Need robust and data and user-friendly, algorithmic like, and fast, scale well
- Repeatability and reproducibility controls (Baggerly et al 2004, Bioinformatics)
- Experimental design to validate conclusions: sample size requirements

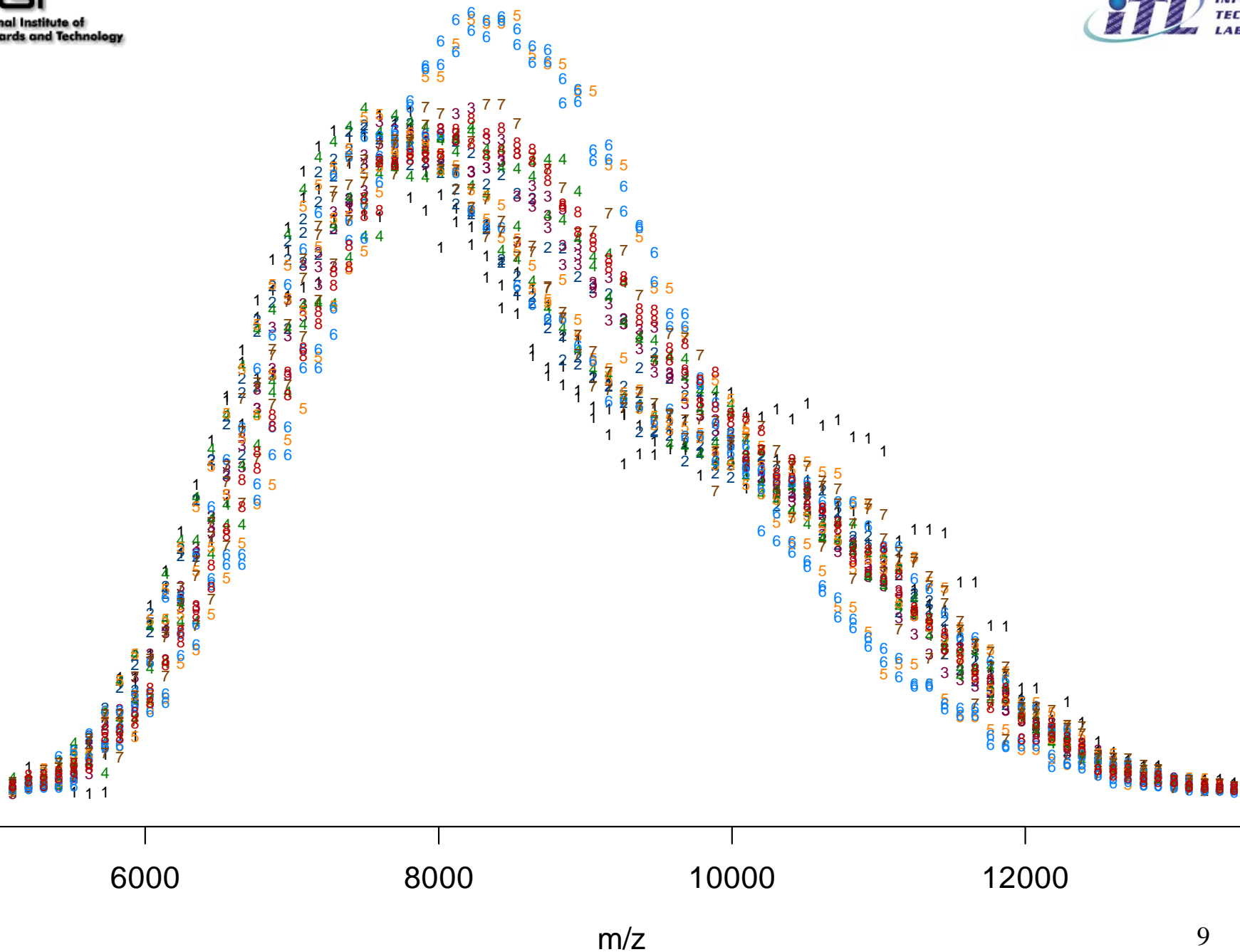
Biostatistics to the rescue?

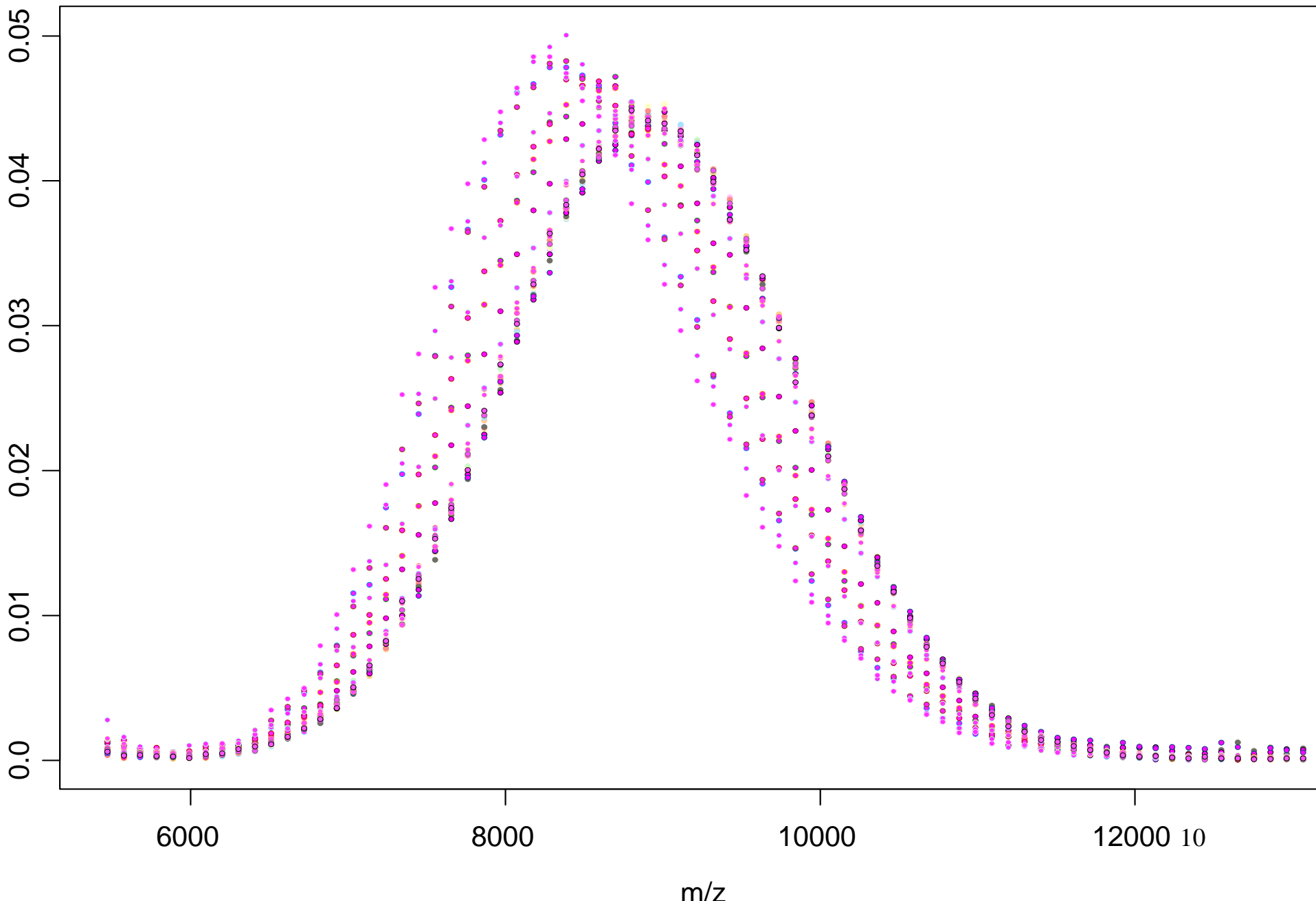


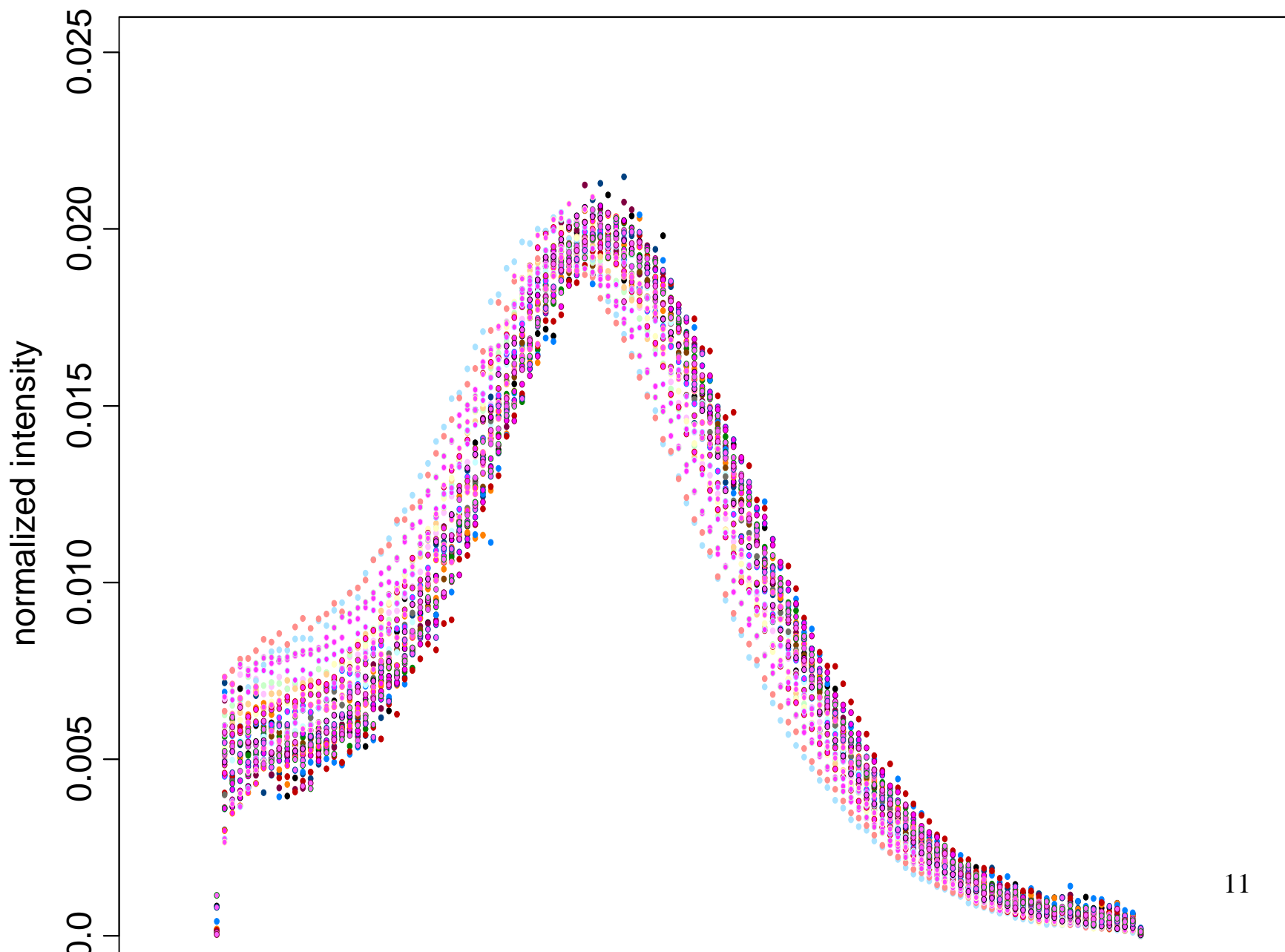
- Boguski and McIntosh 2003: “The analysis of proteomics data is currently informal and relies heavily on expert opinion”.
- **Recommend:** better **study design** to avoid confounding factors
- Study of **biological variability and reproducibility**

III. MALDI-TOF Mass Spectrometry of Synthetic Polymers

- 1. The molecular mass (weight) distribution (MMD) of synthetic polymer is studied using MALDI-TOF MS. The MWDs and the moments derived from MALDI are compared with the values determined by traditional methods including size exclusion chromatography, light scattering, and osmometry.**
- 2. Evaluate effects from experimental design of various instrument settings and sample preparation.**







Classical Polymer Approaches

M_n = number-average molecular weight

First moment

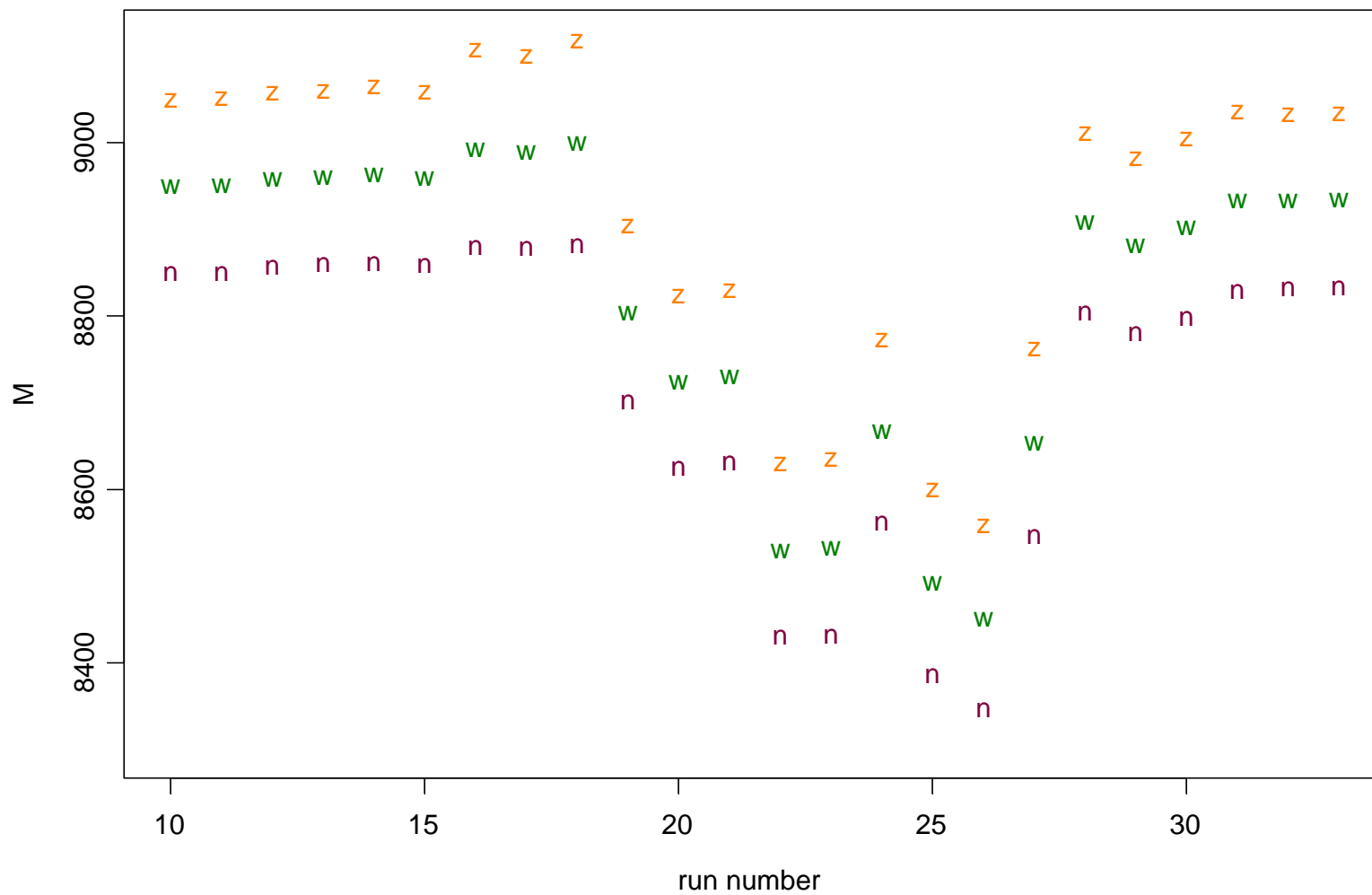
– M_w = weight-average molecular weight

2nd moment/1st moment

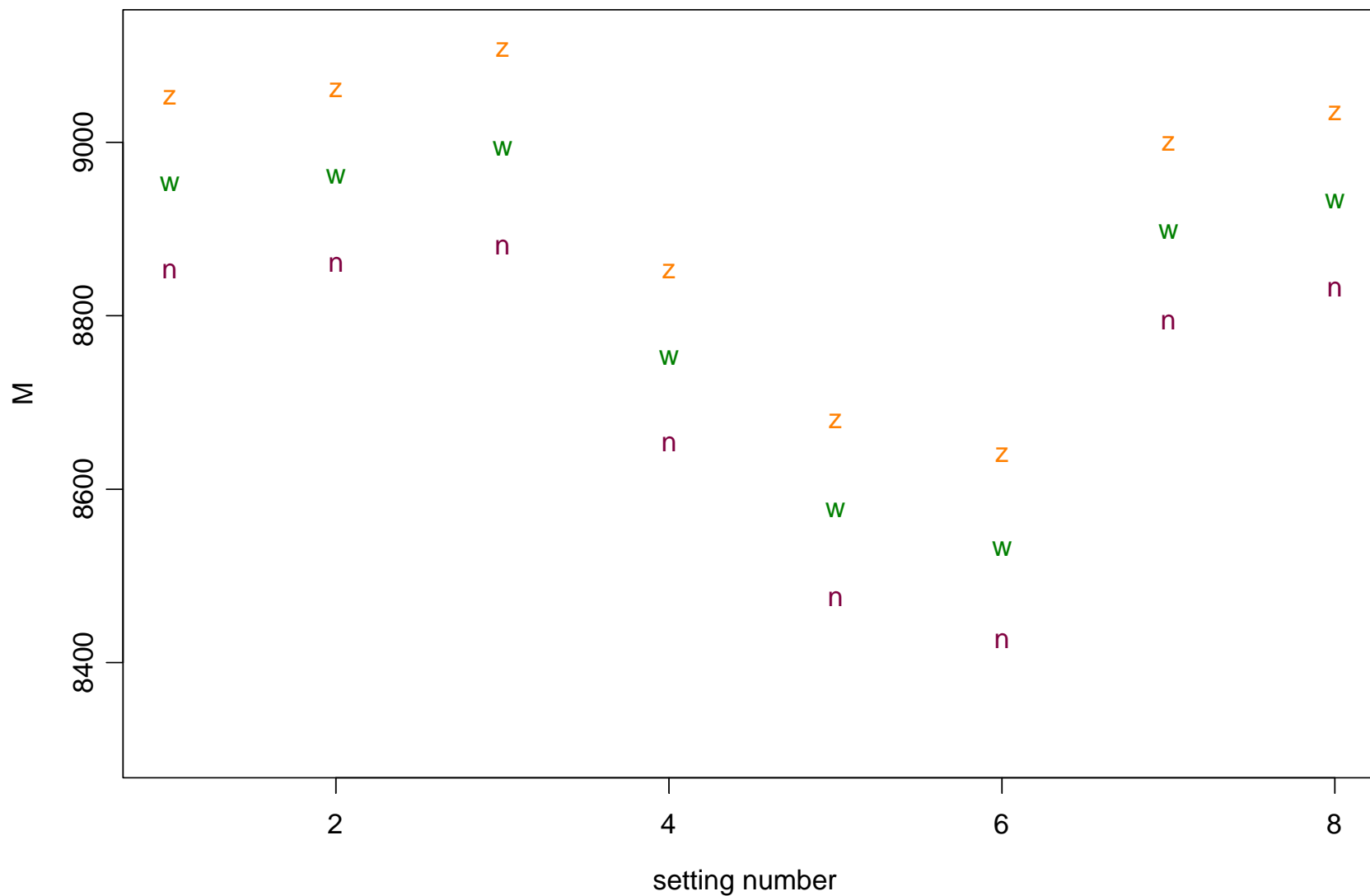
– M_z = z-average molecular weight

Third moment/2nd moment

molecular moments



average molecular moments

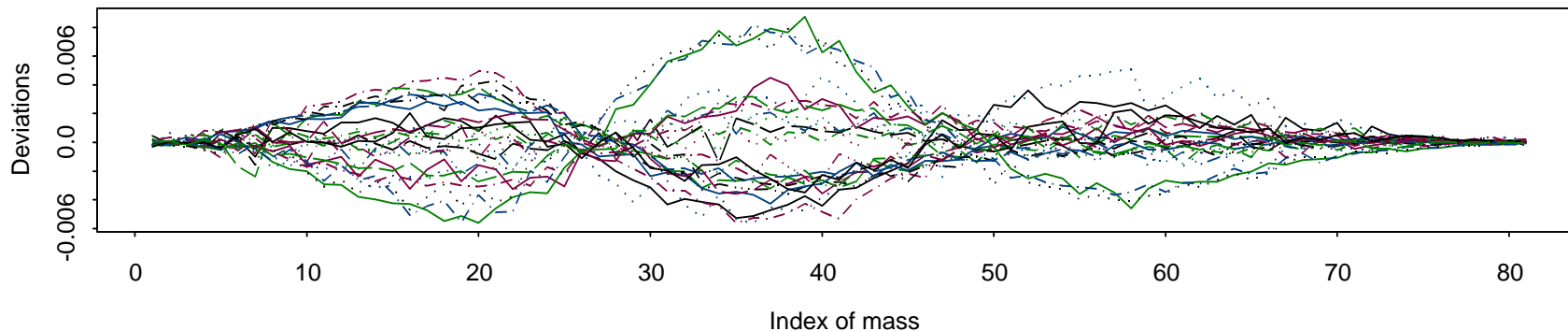


Proposed Analysis Methods

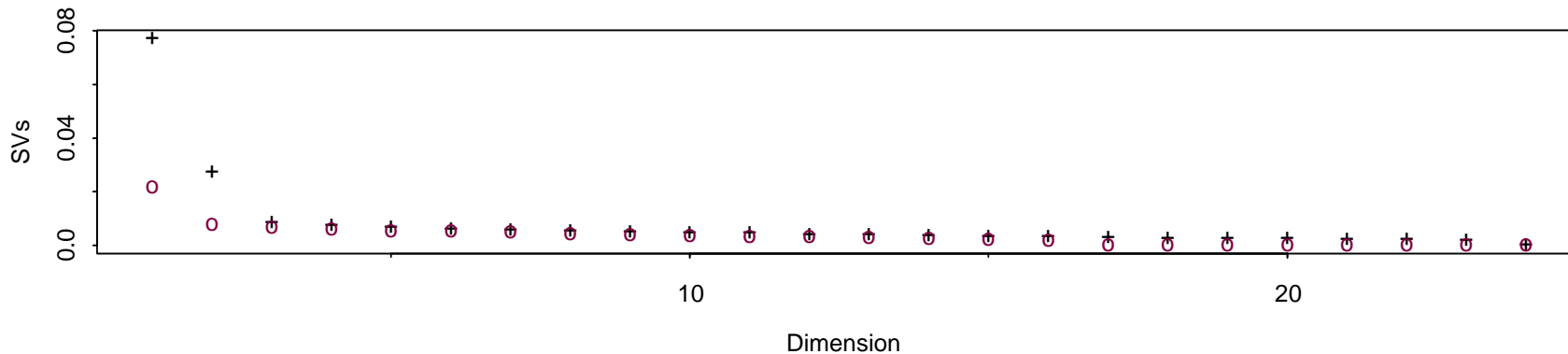
- Standard Functional ANOVA
 - Analysis on the points in each spectrum at each m/z value
 - Does not take into account the continuity of the data
- Singular Value Decomposition-based
 - Singular values decomposition of fitted run spectra and setting mean spectra
 - Does take into account the continuity of the data

SVD Analysis Results: I

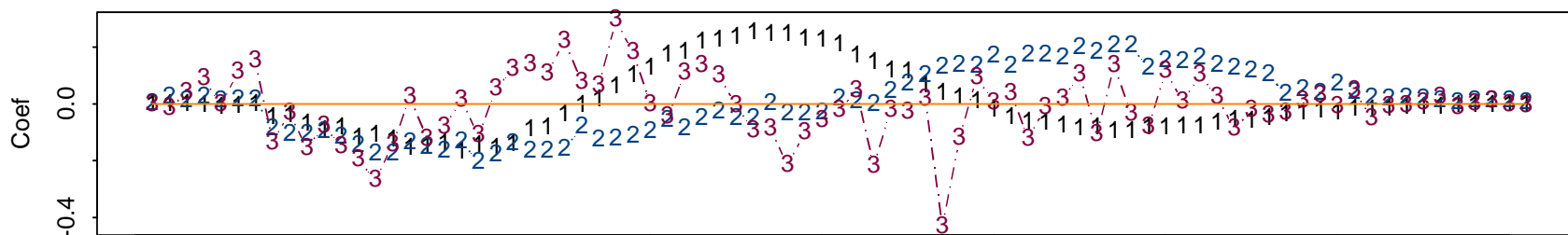
Mean-swepted mass spec



SVD spectrum of raw data (+) and of noise (o)

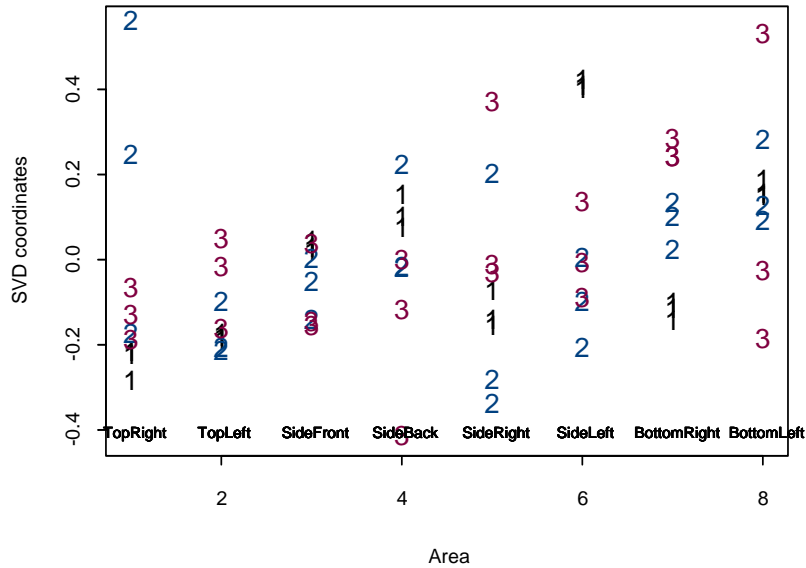


Columns of V matrix: PCA

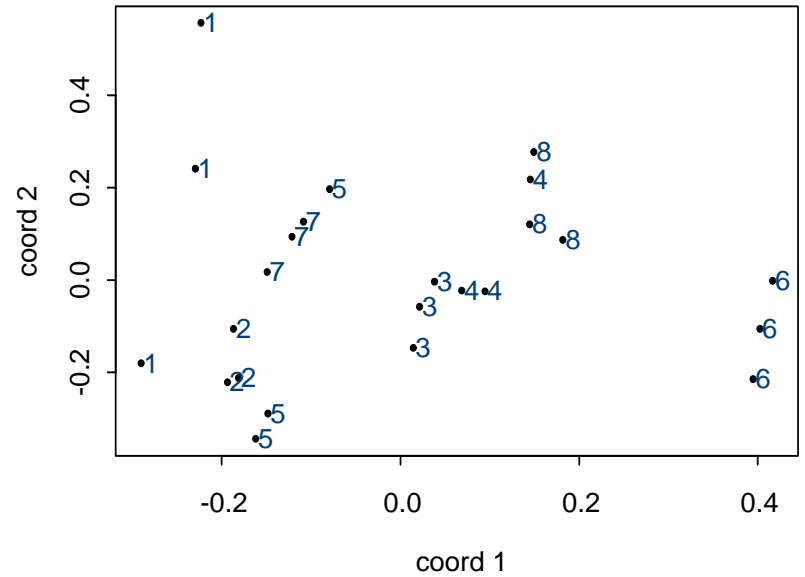


SVD Analysis Results: II

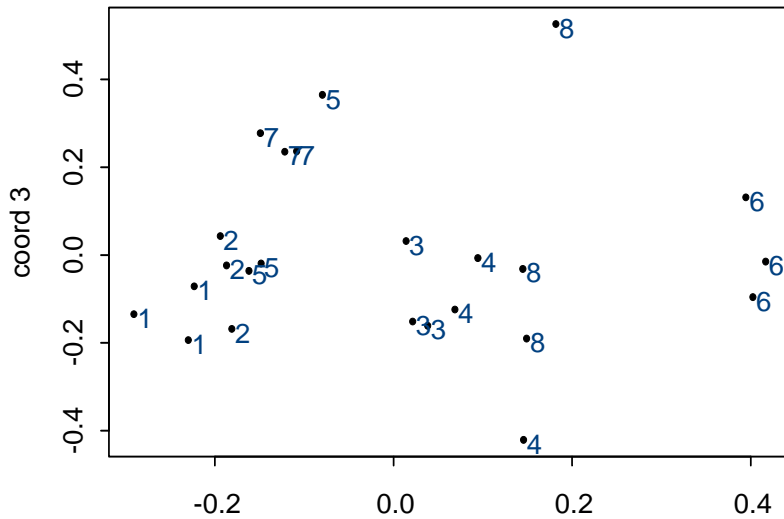
Time plot of the 3-d representational points



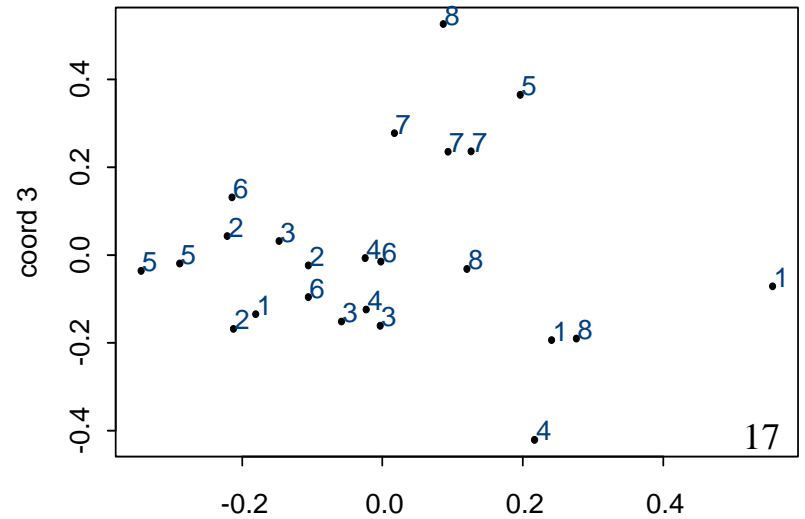
Scatter plot of coordinates 1 and 2



Scatter plot of coordinates 1 and 3



Scatter plot of coordinates 2 and 3



- Determine the significance of singular values
- Develop MANOVA to SVD-derived U vectors for reproducibility test
- Hidden structure and hidden dimension: how do they determine the sample size requirement
- Curse of dimensionality: over-hyped phenomenon (?)

Lu (1999, J. Multivariate Analysis)

- SVD-based high-dimensional regression such as multivariate locally weighted regression: popular tool for multivariate calibration in chemometrics

Related Papers:

- Z.Q. Lu (1999). Nonparametric Regression with Singular Design. *Journal of Multivariate Analysis*, 70, 177-201.
(What is curse of dimensionality?)
- Z.Q. Lu, J. Huckett, S.J. Wetzel, C.M.Guttman (2004). *Functional ANOVA For Mass Spectral Data Analysis in Synthetic Polymers*. Paper in preparation. To be submitted to *Journal of American Statistical Association*.
(The paper on statistics for synthetic polymers)
- Z.Q. Lu (2004). *Singular Value Decomposition Methods in Statistics and High-dimensional Data Analysis*. Paper in preparation.
(The SVD theory paper in statistics)