Error Analysis and Propagation in Metabolomics Data Analysis

Part I

Hunter N.B. Moseley. "Error Analysis and Propagation in Metabolomics Data Analysis" *Comp Struct Biotech J*, 4, e201301006 (2013).

Presented by Hunter Moseley

Outline

- What is Error Analysis?
- Statistical Terminology and Definitions
- Biases
- Major Steps in Error Analysis
- Linear Assumptions in Error Propagation

Uncertainty

What is ~~Error~~ Analysis?

But "Error Analysis" is the accepted term across multiple fields and disciplines.

- **Error analysis is the detection, identification, and quantification of different types of uncertainty present in measurements and the propagation of this uncertainty through mathematical calculations and procedures.**
  - **This definition associates the term error more with precision and less with mistake (inaccuracy).**
- **Error (Uncertainty) Analysis has several uses:**
  i. **Quality control of experiments.**
  ii. **Selection of appropriate statistical methods for data analysis.**
  iii. **Determination of uncertainty in results.**

Why is Error Analysis So Important for Metabolomics

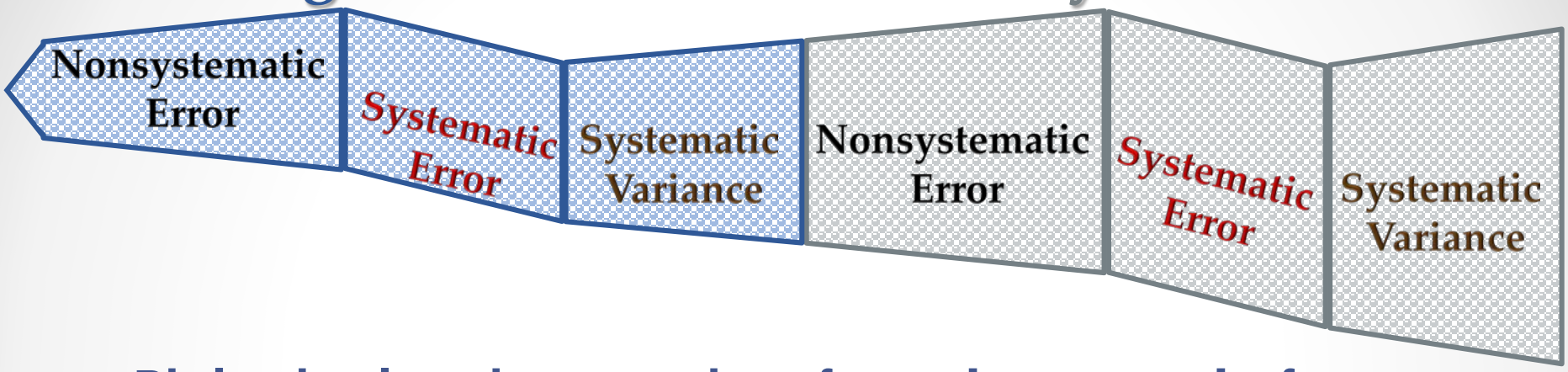
- **Error analysis plays a fundamental role in describing the amount of confidence in results.**
  - **Especially as the number and heterogeneity of measurements increases.**
- **Metabol**omics** experiments have a lot of measurements.**
- **Metabolomics has more molecular heterogeneity than other omics technologies.**
  - **Genomics - 1 type of molecular entity, DNA.**
  - **Transcriptomics - 1 type of molecular entity, RNA.**
  - **Proteomics - 1 type of molecular entity, protein.**
  - **Metabolomics - thousands of types of molecular entities.**

Basic Statistical Terminology

- **Mean:** $\bar{x} = \frac{\sum x}{N}$
  - Estimate of the expected value.
- **Variance:** $\sigma_x^2 = \frac{\sum (x - \bar{x})^2}{N - 1}$
  - Spread of repeated measured values around the mean.
- **Standard Error:** $SE_x \; or \; \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$
  - Probabilistic description of how close the mean is to the expected value.
- **Confidence Interval:**
  - Identifies a range which includes the expected value at some level of confidence (typically 95% or 99%).
- **Covariance:** $\sigma_{xy}^2 = \frac{\sum (x - \bar{x})(y - \bar{y})}{N - 1}$
  - Describes how two measured variables vary together.
- **(Pearson's) Correlation:** $r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{(N - 1)\sigma_x \sigma_y}$
  - Describes the dependence between two measured variables.

Variances

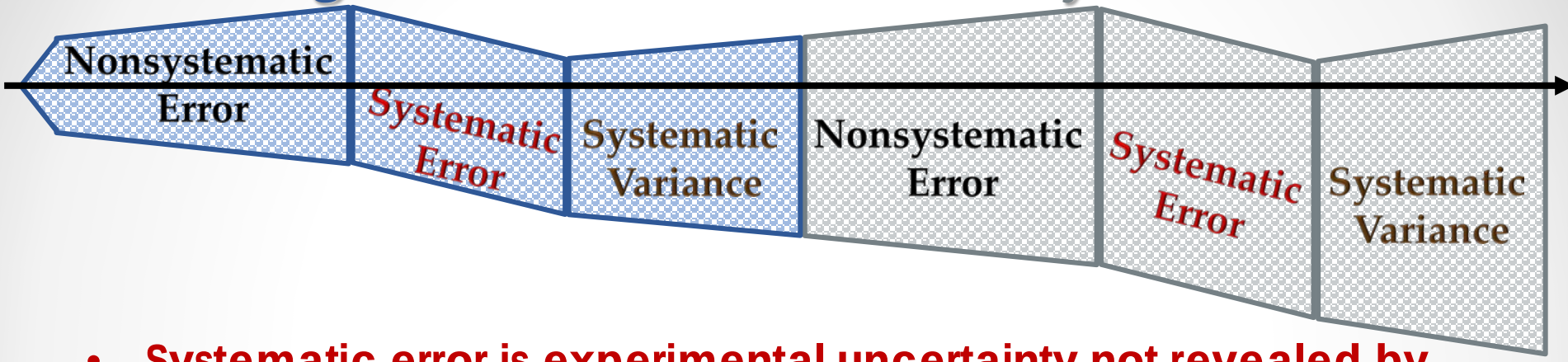
Biological Variance vs Analytical Variance



- **Biological variance arises from the spread of measured values observed from multiple biological samples.**
- **Analytical variance arises from the spread of measured values observed from multiple measurements made from the same biological sample.**

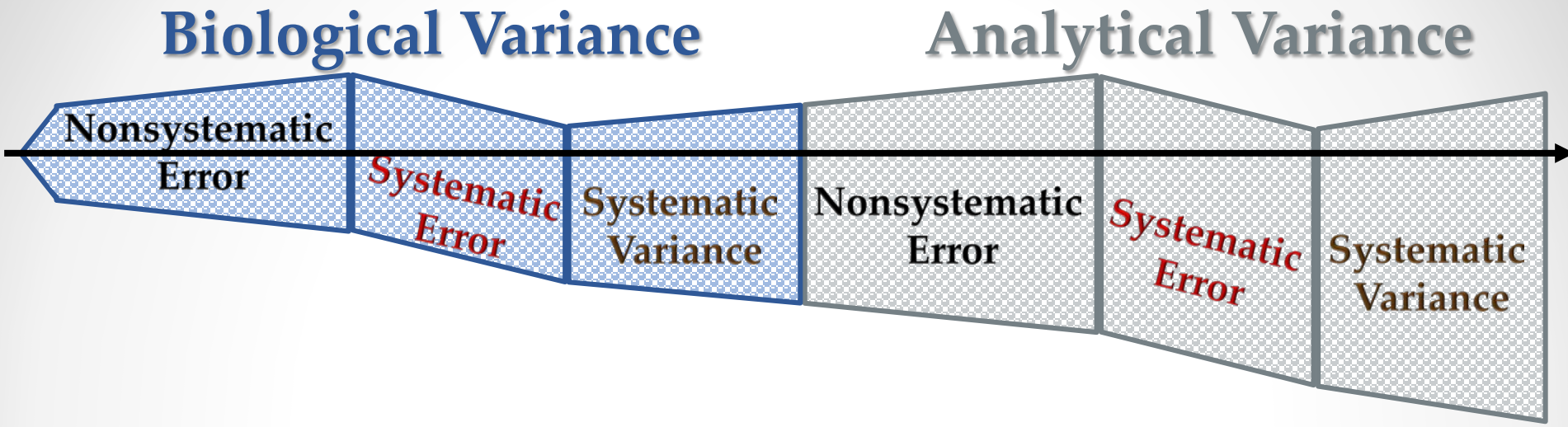
Variances and Errors

Biological Variance Analytical Variance



- **Systematic error is experimental uncertainty not revealed by repeated measurements.**
  - **Does not appreciably affect variance.**
  - **Can affect covariances and correlations between measured variables.**
  - **Typically only revealed and corrected by separate tests/experiments.**
- **Nonsystematic error (AKA error variance) is the experimental uncertainty revealed by repeated measurements.**
  - **Can be reliably estimated by statistical methods.**
- **Systematic variance represents the variance between groups of related samples in the sample set.**
  - **Specific systematic variances can be the desired signal to detect or part of the uncertainty in the measurements due to confounding factors.**
  - **In other words, one scientist's uncertainty is another scientist's usable systematic variance.**

Variances, Errors, and Biases

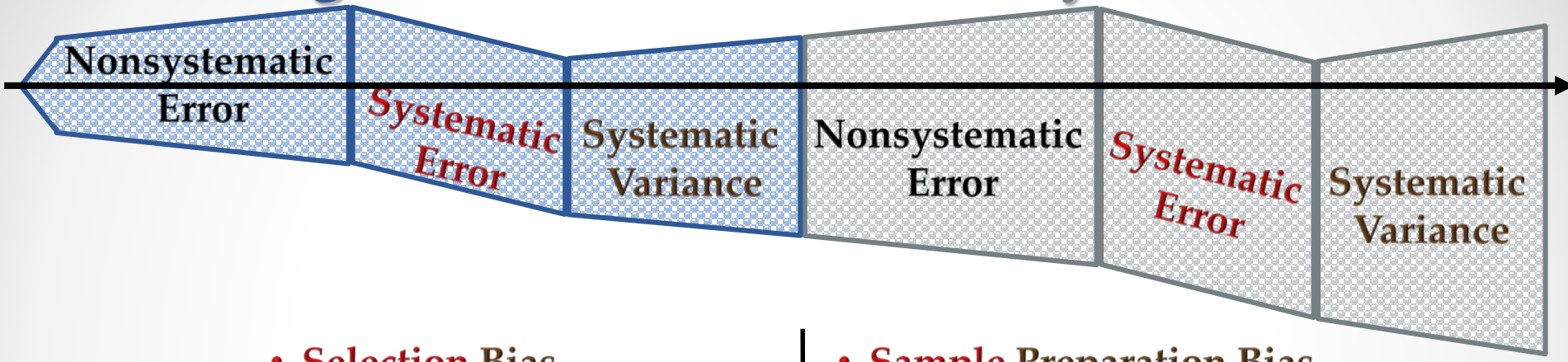


- **Bias refers to any factor that distorts the design, execution, analysis, and interpretation of a measurement.**
  - **A systematic error that distorts the measured values but does not change the variance.**
  - **A systematic variance arising from a confounding factor that is either unknown or inadequately addressed.**
  - **An inadequate or improper statistical method of analysis.**

Types of Biases

Biological Variance

Analytical Variance



- Selection Bias
  - Genetic (race/sex) Bias
  - Epigenetic Bias
    - Tissue/Cell Selection Bias
  - Temporal Selection Bias
- Biological Conditions Bias

- Sample Preparation Bias
  - Extraction Bias
  - Procedural Bias
  - Storage Bias
- Standards Bias
- Sample Complexity Bias
- Analytical Conditions Bias

Biological Biases

Analytical Biases

- Methodological Bias
  - Statistical Assumptions
  - Lack of Statistical Power
  - Multiple Testing
- Confirmation Bias

- Assignment Error
  - Metabolite Assignment Error
  - Class (Group) Assignment Error

Interpretive Biases and Errors

Why is assessment of assignment error so problematic in metabolomics?

Dealing with Biases

I. Use reasonably consistent experimental designs that exclude:
   - Partial consistencies for specific groups of samples, which may lead to a systematic variance from biological or analytical biases.
   - Trivial consistencies that may limit the generalization of results, due to systematic errors from biological or analytical biases.

II. Use effective experimental designs.
   - **Matched-pair case-control experiments limit the effects of confounding factors, especially from biological biases.**
   - Balanced case-control groups with respect to possible confounding factors (sex, age, related biological condition) prevents systematic error.
   - Equally balanced confounding factors (blocking) allow the use of more sophisticated statistical methods .
     - ANOVA instead of t-test or Welch ANOVA instead of Welch's t test.

III. Directly test how well a set of measured values for a given measured variable fits an expected/assumed analytical nonsystematic error distribution.
   - The Shapiro-Wilk and the Anderson-Darling tests are two of the best tests for normality (normal distribution).

Dealing with Biases

IV. Validate results with temporally-separated datasets to detect the presence of biases.
- However, passing this analytical cross-validation does not guarantee a bias-free approach.

V. Use blinded metabolomics experiments to reduce bias.
- The double-blind randomized control trial is considered the gold standard.
  - Reduces researcher-introduced performance bias.
  - Does have known masking biases due to the psychological effects of the trial itself.
- Even the blinding of analytical and/or statistical researchers can reduce performance biases.

VI. Use analytical controls to prevent or correct for analytical biases.
- **Use periodic controls or time-stamped near-random controls to track analytical conditions.**
  - Use Latin square or 2D near-random patterns on plates.
- Use blind controls to detect and correct performance and other analytical biases.
- Use a series of controls composed of complex mixtures of representative or chemically similar metabolites to determine systematic error arising from sample extraction methods and mixture interaction effects.

VII. Fully document the experiment and results.
- Document:
  - The biological and analytical experimental procedures.
  - The statistical procedures used in the analysis of the dataset.
  - A detailed list of all known or potential biases and assumptions, along with results of any analysis and testing of these bias and assumptions.
  - Adequate measures of uncertainty and confidence or at least a good explanation for why uncertainty and confidence measures are not provided.
- Enables thorough peer-review and facilitates future meta-analyses.
- Minimum reporting standards for (plant specific) metabolomics experiments exist.
  - No metabolomics standards for reporting known and potential sources of bias.
  - Can borrow from well-documented clinical standards like STARD and CONSORT.

Major Steps of Standard Error Analysis

1. **Error estimation and probability distribution testing.**
   - Involves:
     - Testing of common distributions like normal, Poisson, binomial, and Lorenztian.
       - 8 to 10 replicates are considered the minimum needed with the Shapiro-Wilks test (normality test).
       - 20 to 30 replicates are typically desired for significant power.
     - Calculation, estimation, modeling, and comparison of nonsystematic error, variance, and covariance arising from biological and analytical sources.
       - 13 replicates (12 + 1) are considered the minimum for calculating variances with ~half-width confidence intervals and at least the 90% confidence level when approximately normally distributed.
       - 30 replicates are required to calculate variances with ~half-width confidence intervals at the 99% confidence level.
   - One central question: "Will analytical nonsystematic error, variances, and covariances prevent the detection and interpretation of biological systematic variance in a given metabolomics dataset?"
     - Determine any analytical nonsystematic error, variance and covariance that could interfere with biological interpretation.
   - These issues really need to be part of the experimental design.
     - How many replicates are needed at each stage of the experimental protocol in order to have the necessary dataset for thorough error analysis.
     - Address issues of statistical power for the expected statistical methods.
       - Probability that a statistical test will properly reject the null hypothesis and not make a false negative decision (Type II error).
       - Minimum expected statistical power ≥ 0.8 at $\alpha=0.05$ (significance level) with "reasonable" statistical assumptions.
     - Test "reasonable" assumptions by increasing analytical replicates for a subset of the samples.
     - Address failed assumptions and lack of statistical power:
       i. Increase analytical replicates to deal with analytical nonsystematic error.
       ii. Correct for factors that (may) cause analytical systematic variance.
       iii. Switch to statistical methods that can handle the failed assumption(s).
         - Nonparametric Wilcoxon-Mann-Whitney test is preferred to a t-test when the data is significantly non-normal.
         - Neither test works well if the data is highly skewed.
       iv. Incorporate estimates of analytical variance and covariance into more sophisticated statistical methods.

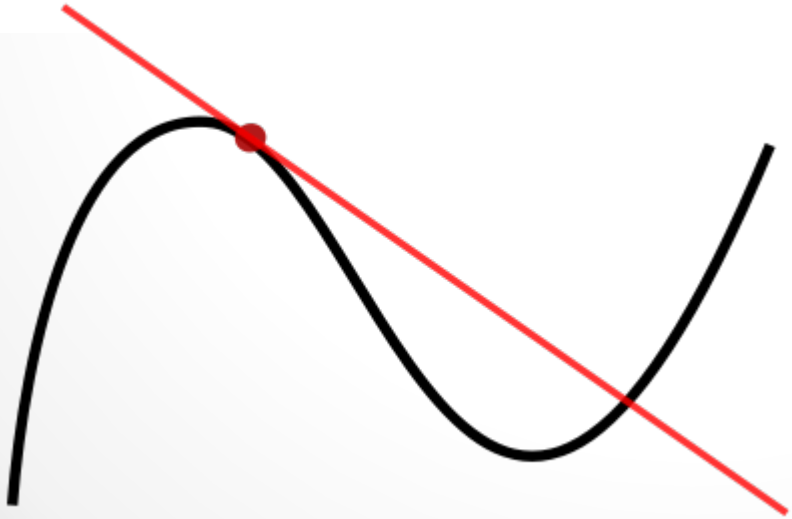
Major Steps of Standard Error Analysis

2. **Error (uncertainty) propagation analysis.**
   A. **Mathematical (analytical) derivation and approximation.**
      - **With few exceptions, almost all analyses of error propagation via mathematical derivation and approximation are performed from a linear perspective.**
      - **This linear assumption is used, whether the functions and algorithms being analyzed are linear or nonlinear.**

   B. **Numerical analysis.**
      - **Often more accurate than mathematical approximation, especially for nonlinear functions.**
      - **Very computationally expensive in many instances.**
      - **Typically requires writing programs to perform the analysis where some form of the Monte Carlo method is usually employed.**
        - **The Monte Carlo method is simply sampling a given function or algorithm via the use of random input values.**

Linear Assumption in Error Propagation

$$y = f(x_1, \ldots, x_n)$$



http://en.wikipedia.org/wiki/File:Tangent_to_a_curve.svg

The first order terms represent a tangent to the function at a specific point $f(\bar{x}_1, \ldots, \bar{x}_n)$.

Linear Assumption in Error Propagation

$$\bar{y} = \frac{1}{N}\sum_{a}^{N} f(x_{1,a}, \dots, x_{n,a}) \approx \frac{1}{N}\sum_{a}^{N}\left(f(\bar{x}_1, \dots, \bar{x}_n) + \sum \frac{\partial f(\bar{x}_1, \dots, \bar{x}_n)}{\partial x_i}(x_{i,a} - \bar{x}_i)\right)$$

$$\approx f(\bar{x}_1, \dots, \bar{x}_n) + \sum \frac{\partial f(\bar{x}_1, \dots, \bar{x}_n)}{\partial x_i}\frac{1}{N}\sum_{a}^{N}(x_{i,a} - \bar{x}_i) \approx f(\bar{x}_1, \dots, \bar{x}_n) + \sum \frac{\partial f(\bar{x}_1, \dots, \bar{x}_n)}{\partial x_i}(0)$$

$$\approx f(\bar{x}_1, \dots, \bar{x}_n)$$

$$\sigma_y^2 = \frac{1}{N-1}\sum_{a}^{N}(y_a - \bar{y})^2 = \frac{1}{N-1}\sum_{a}^{N}\left(f(x_{1,a}, \dots, x_{n,a}) - f(\bar{x}_1, \dots, \bar{x}_n)\right)^2$$

$$\approx \frac{1}{N-1}\sum_{a}^{N}\left(f(\bar{x}_1, \dots, \bar{x}_n) + \sum \frac{\partial f(\bar{x}_1, \dots, \bar{x}_n)}{\partial x_i}(x_{i,a} - \bar{x}_i) - f(\bar{x}_1, \dots, \bar{x}_n)\right)^2$$

Linear Assumption in Error Propagation

$$\approx \frac{1}{N-1}\sum_{a}^{N}\left(\sum \frac{\partial f(\bar{x}_1,\ldots,\bar{x}_n)}{\partial x_i}\left(x_{i,a}-\bar{x}_i\right)\right)^2$$

$$\approx \frac{1}{N-1}\sum_{a}^{N}\left(\begin{array}{c}\sum\left(\frac{\partial f(\bar{x}_1,\ldots,\bar{x}_n)}{\partial x_i}\left(x_{i,a}-\bar{x}_i\right)\right)^2 \\ +\sum\sum_{j\neq i}\frac{\partial f(\bar{x}_1,\ldots,\bar{x}_n)}{\partial x_i}\frac{\partial f(\bar{x}_1,\ldots,\bar{x}_n)}{\partial x_j}\left(x_{i,a}-\bar{x}_i\right)\left(x_{j,a}-\bar{x}_j\right)\end{array}\right)$$

$$\approx \underbrace{\sum\left(\frac{\partial f(\bar{x}_1,\ldots,\bar{x}_n)}{\partial x_i}\right)^2\sigma_{x_i}^2}_{variance\ sum} + \underbrace{\sum\sum_{j\neq i}\frac{\partial f(\bar{x}_1,\ldots,\bar{x}_n)}{\partial x_i}\frac{\partial f(\bar{x}_1,\ldots,\bar{x}_n)}{\partial x_j}r_{x_ix_j}\sigma_{x_i}^2\sigma_{x_j}^2}_{covariance\ sum}$$

Gaussian Error Propagation (GEP)

Linear Assumption in Error Propagation

$$\sigma_y^2 \approx \underbrace{\sum \left(\frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i} \right)^2 \sigma_{x_i}^2}_{variance\ sum} + \underbrace{\sum_{j \neq i} \sum \frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i} \frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_j} r_{x_i x_j} \sigma_{x_i}^2 \sigma_{x_j}^2}_{covariance\ sum}$$

$$\sigma_y^2 \approx \mathbf{j}(\bar{\mathbf{x}})^{\mathrm{T}} \mathbf{C_x}\, \mathbf{j}(\bar{\mathbf{x}})$$

vector of 1st order partial derivatives at $\bar{\mathbf{x}}$

covariance matrix for **x**

$$\mathbf{C_y} \approx \mathbf{J}_F(\bar{\mathbf{x}})\, \mathbf{C_x}\, \mathbf{J}_F(\bar{\mathbf{x}})^{\mathrm{T}}$$

covariance matrix for **y**

Jacobian matrix at $\bar{\mathbf{x}}$

Error Analysis and Propagation in Metabolomics Data Analysis

Part II

Numerical Error Propagation Analysis

- **The Monte Carlo Method**
  - A large collection of methods with a wide variety of applications involving the sampling of a given function or algorithm via the use of random input values.
- **A simple Monte Carlo Method for error propagation analysis:**

$$\mathbf{y}_i = f(\mathbf{x}_i) \text{ where } \mathbf{x}_i \in X \text{ and } X_j \sim D_j$$

$\mathbf{x}_i$ - pseudo-random input vectors of values.
X – the set of pseudo-random input vectors of values used in the sampling.
$X_j \sim D_j$ – the probability distribution D_j for input variable X_j in the vector.

  - The sampling of *f* produces a set of vectors $\mathbf{y}_i \in Y$, that can be directly analyzed in an analogous manner as experimental data:
    - Probability distribution testing.
    - For common probability distributions, the calculation of:
      - Expected values.
      - Variances
      - Standard errors.
      - Correlations.

Why is error propagation analysis via a Monte Carlo method so popular?

Methods to Generate Pseudo-Random Values

- **Built-in R functions:**
  - rnorm – generates normally distributed random numbers.
  - rlnorm – generates log normally distributed random numbers.
  - rbinom – generates binomially distributed random numbers.
  - rpois – generates Poisson distributed random numbers.
- **Several straight-forward algorithms available:**
  - Typically use uniformly distributed pseudo-random numbers U[0,1].
  - Different algorithms for the common probability distributions.
    - By definition, the inverse of a cumulative distribution function can be used to calculated pseudo-random values from U[0,1] distributed values.
  - Example – Box Muller method.
    - Popular, because it is easy to implement.
    - Uses a pair of U[0,1] values to generate a pair of normally distributed values.
- **Even complex or unknown distributions can be estimated.**
  - Use a two-sample Kolmogorov–Smirnov test.
  - Sets of pseudo-random values are generated based on bootstrap-derived statistical parameters and tested against an experimentally derived set of measured values using the two-sample K-S test.
- **Correlation can be introduced by several methods.**

Knuth DE (2006) *The art of computer programming*: Addison-Wesley.

Massey Jr FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association* 46: 68-78.

Vale CD, Maurelli VA (1983) Simulating multivariate nonnormal distributions. *Psychometrika* 48: 465-471.

Headrick TC, Sawilowsky SS (1999) Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika* 64: 25-35.

Properties of Y

$$\mathbf{y}_i = f(\mathbf{x}_i) \text{ where } \mathbf{x}_i \in X \text{ and } X_j \sim D_j$$

- **If *f* is linear and *X* variables are (reasonably) independent and identically distributed ($X_i \sim X_j$) with finite variance:**
  - **$y_i \in Y$ often reflects the distribution of *X* ($Y_i \sim X_j$)**
  - **Certain Y_i may even approximate a normal distribution when Y_i depend on many X_j variables (i.e. Central Limit Theorem).**
- **If *f* is nonlinear:**
  - **Drastically non-normal distributions are common for *Y* and quite distinct from *X*, even if *X* is normally distributed.**
  - **Nonlinearity is very common for metabolic models with exchange and bidirectional fluxes.**
    - **Sometimes metabolomic models can be solved (reasonably approximated) by a linearization.**

What is often the salvation of statistical analysis?

Dealing with Nonnormal Y_i

- **A median with a confidence interval is preferable to a mean with a standard error.**
  - Simply order the sampling for each Y_i and takes the interval:

    $(y_{(n+1)(1-c)}, y_{(n+1)c})$

    where c is the level of confidence as a fraction (i.e. 0.95).
  - Requires sample sizes of 1000 or 10000, depending on the desired level of confidence in these confidence intervals.
- **A Spearman's rank correlation coefficient can be used to calculate correlation in a nonparametric way.**

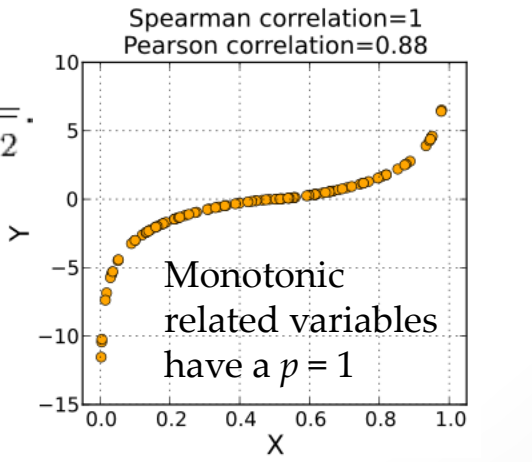
$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

where xi and yi are ranks.

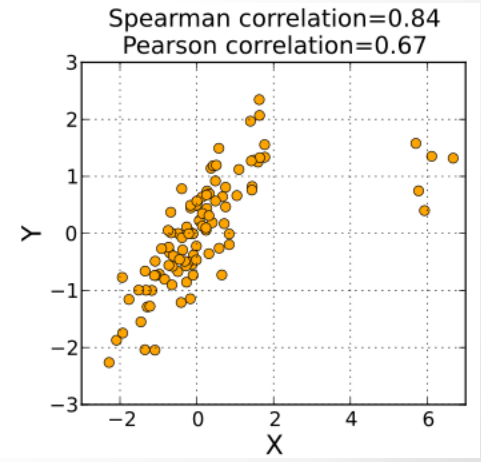
Simplified (no duplicate values)

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

where di = (xi – yi).



http://en.wikipedia.org/wiki/File:Spearman_fig1.svg



http://en.wikipedia.org/wiki/File:Spearman_fig3.svg

Buckland ST (1984) Monte Carlo confidence intervals. *Biometrics* 40: 811-817.

Spearman C (1904) The proof and measurement of association between two things. *The American Journal of Psychology* 15: 72-101.

Inverse Problems in Metabolomics

- **Often a model of relevant chemical reactions for a "known" cellular metabolic network is more easily constructed and used to calculate specific metabolite fluxes and pools (mass-related characteristics) that can be compared to experimental values, especially in a time series.**

$$\mathbf{x}_i = g(\mathbf{y}_i) \text{ where } g \approx f^{-1}$$

- **Sometimes "model" refers to:**
  - **Framework of equations (*g*).**
  - **g and fixed input parameters ($y_{i,j} = c_j$)**
  - **g and optimized parameters (y_{opt}).**

Major Metabolic Modeling Methodologies

- **Metabolic Flux Analysis (MFA)**
  - Determines a set of cellular metabolic fluxes from experimental data.
  - Uses a system of differential equations derived from a balanced stoichiometric matrix at steady-state conditions.
- **Flux Balance Analysis (FBA)**
  - Determines sets of steady-state metabolic fluxes that optimize a stated cellular objective like maximizing biomass production.
  - Uses a linearized representation of fluxes derived from a stoichiometric matrix assuming steady-state conditions.
- **Flux Ratio Analysis**
  - Determines flux ratios (relative flux) of converging pathways from experimental data.
  - Uses a system of differential equations.
- **Metabolic Control Analysis**
  - Determines control coefficients for specific components like enzymes in a metabolic network.
  - Based on how changes in enzyme concentrations affect flux through parts of the metabolic network.

Optimization of Inverse Problems

- An objective function compares the results from the model $x_i = g(y_i)$ with experimental data x_{exp} through some norm function.
  - AKA target function or energy function depending on context.
- O_s – simple objective function using an ℓ_2-norm.

$$O_s(\mathbf{y}_i) = \left\| g(\mathbf{y}_i) - \mathbf{x}_{exp} \right\|^2 = \left\| \mathbf{x}_i - \mathbf{x}_{exp} \right\|^2 = \sum_j \left(x_{i,j} - x_{exp,j} \right)^2$$

- The objective function is minimized while model parameters are optimized to y_{opt} using an optimization method of choice.
  - Often some type of Monte Carlo method by definition (cf. simulated annealing).

Problems with Inverse Problems

- **Almost all metabolomics inverse problems are ill-posed and ill-conditioned due to:**
  - **Model complexity.**
  - **Model non-linearity.**
  - **Limitations in the number and variety of measurements.**
    - **Can prior knowledge overcome limitations in the data without introducing undue bias?**
- **These issues may:**
  - **Preclude a unique solution y_{opt} to a given set of experimental measurements x_{exp} (i.e. ill-posed).**
  - **Allow the existence of multiple solutions $y_{opt,I}$ (i.e. ill-posed).**
  - **May cause discontinuities.**
  - **May cause high conditioning (i.e. large variation) in model parameters with respect to small changes in experimental measurements.**
    - **Leads to overfitting of model parameters (y_{opt}).**
    - **Amplifies uncertainty in model parameters.**

Regularization – use of additional information to prevent overfitting of an ill-conditioned problem or allow a unique solution to an ill-posed problem.

Tikhonov Regularization

weighting factor

expected model parameters

$$O_T(\mathbf{y}_i) = \left\| g(\mathbf{y}_i) - \mathbf{x}_{exp} \right\|^2 + \alpha R(\mathbf{y}_i, \mathbf{y}_E)$$

$$= \left\| \mathbf{x}_i - \mathbf{x}_{exp} \right\|^2 + \alpha \left\| \mathbf{y}_i - \mathbf{y}_E \right\|_p^2$$

$$= \sum_j \left(x_{i,j} - x_{exp,j} \right)^2 + \alpha \left(\sum_k \left| y_{i,k} - y_{E,k} \right|^p \right)^{\frac{2}{p}}$$

p-norm

- **Issues using Tikhnov regularization:**
  - α – large enough to prevent overfitting, but small enough to prevent bias.
  - **p** – must properly weight between $||x_i||$ and $||y_i||$.
  - y_E – a "reasonable" expectation, "close enough" to y_{exact}.
- **If α and p are picked properly, a confidence region around y_E that includes y_{exact} can be estimated with respect to $\|y_i - y_E\|_p^2$ based on a Fisher distribution.**

Engl HW, Hanke M, Neubauer A (1996) *Regularization of inverse problems*: Springer.

Error-bounded Generalized Least Squares Approach

$$O_g(\mathbf{y}_i) = \left(g(\mathbf{y}_i) - \mathbf{x}_{exp}\right)^{\mathrm{T}} \mathbf{C_x}^{-1} \left(g(\mathbf{y}_i) - \mathbf{x}_{exp}\right) \leq \delta_{\mathbf{x}}$$

$$CR_{1-\beta}(\mathbf{y}_{opt}) \approx \left\{\mathbf{y}_i \middle| O_g(\mathbf{y}_i) \leq \delta_{\mathbf{x}} \approx O_g(\mathbf{y}_{opt}) + \chi^2_{n-m}(1-\beta)\right\}$$

$$CI_{y_j,1-\beta}(\mathbf{y}_{opt}) \approx \left\{y_{j0} \middle| \min O_g(\mathbf{y}_i)|_{y_j=y_{j0}} \leq \delta_{\mathbf{x}} \approx O_g(\mathbf{y}_{opt}) + \chi^2_1(1-\beta)\right\}$$

Where:

- $\mathbf{C_x}$ – analytical covariance matrix for x_{exp}.
- δ_x – error threshold.
- $\chi^2_{n-m}(1-\beta)$ – χ^2 statistic with *n-m* degrees of freedom and a p-value of 1-β.
  - *n* - # of measured experimental variables.
  - *m* - # of model parameters.
  - β - desired level of confidence.
- y_{opt} - determined by the lowest $O_g(y_i)$.

- **This approach works if:**
  - All measured variables are (approximately) normally distributed.
  - Analytical covariance matrix C_x is known or well-estimated.
- **Caveats:**
  - The residuals normalized by $C_x^{-½}$ should be tested for normality.
  - The optimization needs a large number of repetitions.
    - May be improved by the use of Jacobian and Hessian matrices .

Engl HW, Flamm C, Kügler P, Lu J, Müller S, et al. (2009) Inverse problems in systems biology. *Inverse Problems* 25: 123014.
Antoniewicz MR, Kelleher JK, Stephanopoulos G (2006) Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. *Metabolic Engineering* 8: 324-337.

The Grand Assumption: *g* is "reasonably" accurate

- **Potential for a very large (gargantuan) interpretive bias.**
- **The faith in certain metabolic models is quite troubling, given:**
  - **The lack of verified details.**
  - **Errors in metabolic databases used in the construction of models.**
    - **Especially construction of models based on eukaryotic metabolic networks.**
  - **Often more parameters than measured variables.**

Improving Model Verification

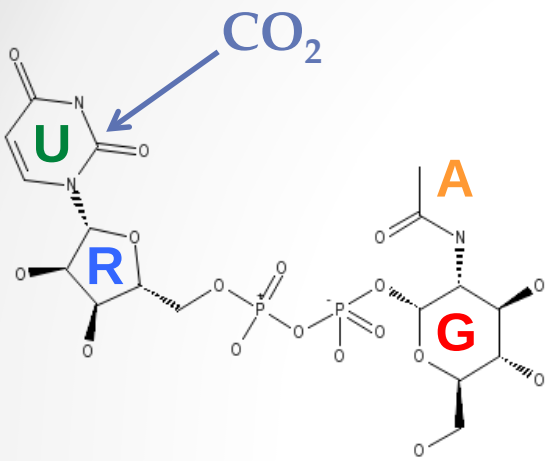
I. **Pare down a metabolic model to what is relevant to the observables.**
   a) Gross model paring.
      - Limits the model to relevant pathways and modules of a metabolic network .
   b) Specific variable pairing by independence.
      - Limits the model parameters to the smallest set of independent or "free" model parameters from which other intermediate model parameters are derived.
   c) Specific variable paring by sensitivity.
      - Removes and/or simplifies parts of a model that include insensitive model parameters with respect to measured experimental variables.

II. **Design experiments where there are enough observables to perform model selection (n >> m).**
   - Stable Isotope Resolved Metabolomics (SIRM).
      - Use of multiple stable isotopes.
      - Use of multiple source metabolites.
      - Measurements collected in a time series.

III. **Use more rigorous model verification and selection methods.**
   - Reject models where $O_g(y_{opt}) > \chi^2_{n-m}(1-\beta)$.
      - Analytical error can be adequately determined/estimated.
      - Measured variables are approximately normally distributed.
   - Select between plausible models using standard methods.
      - Akaike Information Criterion (AIC).

Fan TW-M, Lane AN, Higashi RM (2004) The Promise of Metabolomics in Cancer Molecular Therapeutics. *Current Opinion in Molecular Therapeutics* 6: 584-592.
Fan TWM, Lorkiewicz P, Sellers K, Moseley HNB, Higashi RM, et al. (2012) Stable isotope-resolved metabolomics and applications for drug development. *Pharmacology & Therapeutics* 133: 366.
Moseley HNB, Lane A, Belshoff A, Higashi R, Fan T (2011) A novel deconvolution method for modeling UDP-N-acetyl-D-glucosamine biosynthetic pathways based on 13C mass isotopologue profiles under non-steady-state conditions. *BMC Biology* 9: 37.

^{13}C Tracing in UDP-Hexose Biosynthesis



UDP-GlcNAc

$^{13}C_6$-Glucose

Glycolysis

$^{13}C_3$-Pyruvate

pyruvate dehydrogenase complex

$^{13}C_2$-Acetyl-CoA

CAC

$^{13}C_?$-Oxaloacetate

aspartate aminotransferase

$^{13}C_?$-Aspartate + Carbamoyl phosphate + PRPP

carbamoyl phosphate synthetase II

CO_2

$^{13}C_6$-Glucose

PPP ($^{13}C_5$-Ribose)

Pyrimidine Biosynthesis

$^{13}C_6$-Glucose

hexokinase

$^{13}C_6$-Glucose-6-P

phosphohexose isomerase

$^{13}C_6$-Fructose-6-P

glucosamine -fructose 6-P aminotransferase

$^{13}C_6$-Glucosamine-6-P

$^{13}C_2$-Acetyl-CoA

glucosamine 6-P N-acetyltransferase

CoA

$^{13}C_8$-N-acetylglucosamine-6-P

P acetyl glucosamine mutase

$^{13}C_8$-N-acetylglucosamine-1-P

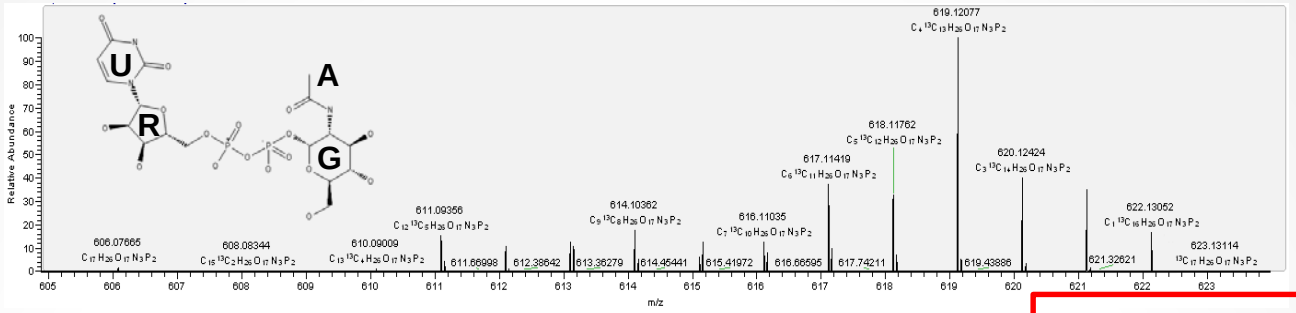
UDP-N-acetylglucosamine pyrophosphorylase

$^{13}C_?$-UDP-N-acetylglucosamine (UDP-GlcNAc)

UDP-glucosamine epimerase

$^{13}C_?$-UDP-N-acetylgalactosamine (UDP-GalNAc)

UDP-Glc|GalNAc FT-ICR-MS Data



^{12}C monoisotopic peak →

$^{13}C_2{}^{12}C_{15}{}^{1}H_{25}{}^{16}O_{17}{}^{31}P_2$

g0r0a2u0 + g0r0a0u2

Each isotopologue represents a set of mass-equivalent positional isotopomers.

$$I_{Norm,i} = \frac{I_i}{\sum I_x}$$

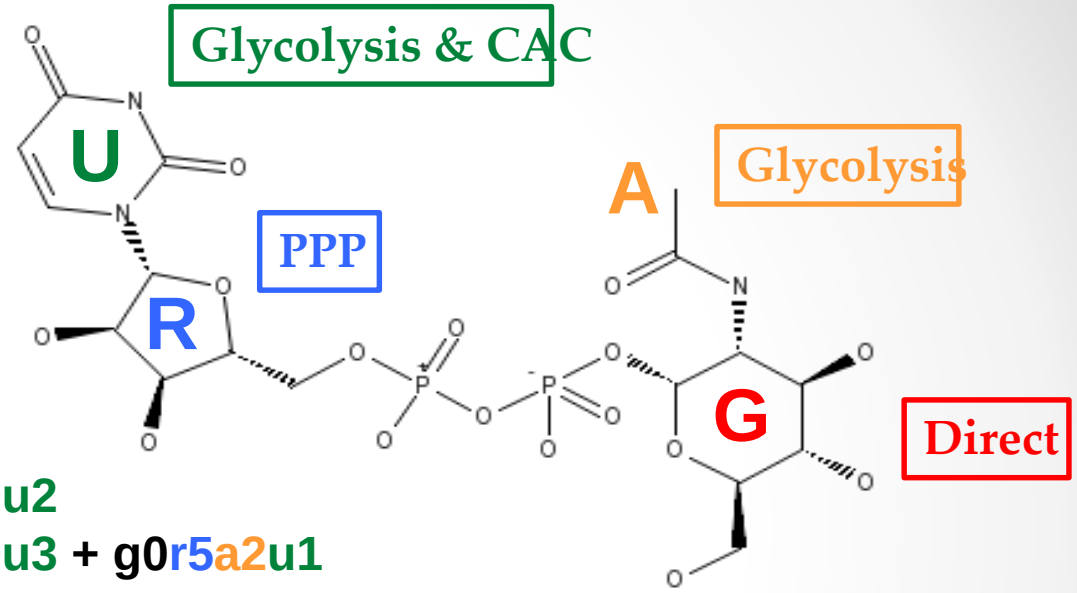
Normalization is an excellent internal reference.

m/z	i	Norm. Intensity	Corrected Intensity
606.0751	0	0.0050033	0.0060476
607.0779	1	0.00094257	0
608.0817	2	0.00099737	0.0010778
609.0844	3	0.00065213	0.00057741
610.0885	4	0.0037484	0.0042294
611.0919	5	0.041111	0.046380
612.0953	6	0.029762	0.027433
613.0990	7	0.036908	0.037537
614.1020	8	0.046745	0.047317
615.1054	9	0.017722	0.014439
616.1087	10	0.033593	0.034823
617.1125	11	0.11357	0.11867
618.1160	12	0.099003	0.096692
619.1191	13	0.29721	0.30518
620.1226	14	0.12134	0.11181
621.1260	15	0.10877	0.10728
622.1289	16	0.043753	0.041821
623.1295	17	0.00056993	0.000092779 ≈ 0

Moiety Model of Isotopologue Intensities

6_G1R1A1U3

I_0 = g0r0a0u0
I_1 = g0r0a0u1
I_2 = g0r0a0u2 + g0r0a2u0
I_3 = g0r0a0u3 + g0r0a2u1
I_4 = g0r0a2u2
I_5 = g0r5a0u0 + g0r0a2u3
I_6 = g6r0a0u0 + g0r5a0u1
I_7 = g6r0a0u1 + g0r5a2u0 + g0r5a0u2
I_8 = g6r0a2u0 + g6r0a0u2 + g0r5a0u3 + g0r5a2u1
I_9 = g6r0a0u3 + g6r0a2u1 + g0r5a2u2
I_{10} = g6r0a2u2 + g0r5a2u3
I_{11} = g6r5a0u0 + g6r0a2u3
I_{12} = g6r5a0u1
I_{13} = g6r5a0u2 + g6r5a2u0
I_{14} = g6r5a0u3 + g6r5a2u1
I_{15} = g6r5a2u2
I_{16} = g6r5a2u3
I_{17} = NA contribution only.



- Glucose: g0 + g6 = 1 ~ 1 parameter
- Ribose: r0 + r5 = 1 ~ 1 parameter
- Acetyl: a0 + a2 = 1 ~ 1 parameter
- Uracil: u0 + u1 + u2 + u3 = 1 ~ 3 parameters

6 parameters

Solving these parameter values will estimate the contribution of these metabolic pathways to ^{13}C incorporation in UDP-GlcNAc biosynthesis.

Moseley et al., *Proceedings of BIOINFORMATICS 2011, 108-115 (2011)*

Extensive Comparison of Models

AIC = -157.43 6_G0R2A1U3_g3r2r3_g6r5
AIC = -109.64 6_G1R1A1U3_a1
AIC = -136.29 6_G1R1A1U3_g5
AIC = -154.32 6_G1R1A1U3
AIC = -137.17 6_G1R1A1U3_r4
AIC = -133.12 6_G1R1A1U3_u4
AIC = -159.00 7_G0R2A2U3_g3r2r3_g6r5
AIC = -72.52 7_G0R3A1U3_g3r2r3_g6r5_g5r4
AIC = -158.25 7_G1R2A1U3_g3r2r3
AIC = -153.65 7_G1R2A1U3_r1
AIC = -159.24 7_G1R2A1U3_r2
AIC = -147.55 7_G1R2A1U3_r3
AIC = -163.39 7_G1R2A1U3_r4
AIC = -153.95 7_G2R1A1U3_g1
AIC = -153.64 7_G2R1A1U3_g2
AIC = -158.87 7_G2R1A1U3_g3
AIC = -151.21 7_G2R1A1U3_g4
AIC = -160.84 7_G2R1A1U3_g5
AIC = -154.17 8_G1R1A2U3C1

AIC = -156.58 8_G1R2A2U3_g3r2r3_g6r5_g5
AIC = -158.22 8_G1R2A2U3_g3r2r3
AIC = -154.14 8_G1R2A2U3_r1
AIC = -159.10 8_G1R2A2U3_r2
AIC = -157.39 8_G1R2A2U3_r2r3
AIC = -148.47 8_G1R2A2U3_r3
AIC = -161.97 8_G1R2A2U3_r4
AIC = -153.91 8_G2R1A2U3_g1
AIC = -155.89 9_G2R2A2U3_r2r3_g1
AIC = -154.77 9_G2R2A2U3_r2r3_g2
AIC = -156.24 9_G2R2A2U3_r2r3_g3
AIC = -152.79 9_G2R2A2U3_r2r3_g4
AIC = -156.13 9_G2R2A2U3_r2r3_g5
AIC = -155.50 9_G2R2A2U3_r2r3_g6r5_g3_g5

A Problem with OverFitting!

Akaike Information Criterion

of model parameters

Log of model likelihood

$$AIC = 2k - 2\ln(L) \approx 2k + n\left[\ln\frac{\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2}{n}\right]$$

Using MultipleTime Points to Handle Overfitting

AIC = -321.81 6_G0R2A1U3_g3r2r3_g6r5
AIC = -355.87 6_G1R1A1U3_a1
AIC = -326.98 6_G1R1A1U3_g5
AIC = -428.98 6_G1R1A1U3
AIC = -332.69 6_G1R1A1U3_r4
AIC = -308.16 6_G1R1A1U3_u4
AIC = -291.31 7_G0R2A2U3_g3r2r3_g6r5
AIC = -287.32 7_G0R3A1U3_g3r2r3_g6r5_g5r4
AIC = -290.16 7_G0R3A1U3_g3r2r3_g6r5_r4
AIC = -306.58 7_G1R1A1U3C1
AIC = -293.12 7_G1R1A1U4
AIC = -299.86 7_G1R1A2U3
AIC = -294.52 7_G1R2A1U3_g3r2r3
AIC = -308.59 7_G1R2A1U3_r1
AIC = -288.94 7_G1R2A1U3_r2
AIC = -277.44 7_G1R2A1U3_r3
AIC = -244.47 7_G1R2A1U3_r4
AIC = -318.01 7_G2R1A1U3_g1
AIC = -317.89 7_G2R1A1U3_g2
AIC = -286.93 7_G2R1A1U3_g3
AIC = -277.12 7_G2R1A1U3_g4
AIC = -252.21 7_G2R1A1U3_g5
AIC = -288.84 8_G1R1A2U3C1
AIC = -296.01 8_G1R2A2U3_g3r2r3_g6r5_g5
AIC = -288.88 8_G1R2A2U3_g3r2r3
AIC = -290.93 8_G1R2A2U3_r1
AIC = -296.67 8_G1R2A2U3_r2
AIC = -296.18 8_G1R2A2U3_r2r3
AIC = -251.87 8_G1R2A2U3_r3
AIC = -239.25 8_G1R2A2U3_r4
AIC = -303.97 8_G2R1A2U3_g1
AIC = -293.45 8_G2R1A2U3_g2
AIC = -288.32 8_G2R1A2U3_g3
AIC = -260.59 8_G2R1A2U3_g4
AIC = -236.42 8_G2R1A2U3_g5
AIC = -293.74 9_G2R2A2U3_r2r3_g1
AIC = -279.33 9_G2R2A2U3_r2r3_g2
AIC = -291.46 9_G2R2A2U3_r2r3_g3
AIC = -241.63 9_G2R2A2U3_r2r3_g4
AIC = -227.58 9_G2R2A2U3_r2r3_g5
AIC = -276.84 9_G2R2A2U3_r2r3_g6r5_g3_g5

Moseley et al., *Proceedings of BIOINFORMATICS 2011, 108-115 (2011)*

Conclusions

- **Determining the propagation of uncertainty in metabolomics data analysis is very hard.**
- **Most in the field are doing it wrong, because:**
  - **They do not understand the math.**
  - **They do not understand the analytical techniques.**
  - **They do not understand the biological problem.**
  - **They do not have the necessary datasets to determine the analytical variance.**
- **There are two ways to handle the problem:**
  - **Collect the necessary datasets to derive analytical uncertainty.**
  - **Take advantage of known correlations to estimate analytical uncertainty.**