Adding Support for Theory in Open Science Big Data

John A. Miller, Hao Peng and Michael E. Cotterell
Department of Computer Science
University of Georgia
Athens, GA, USA
Outline

- Basic Definitions
- Predictive Analytics
- Using Theory in Model Development
- Representing Theory
Basic Definitions

- **Modeling Technique** – e.g., Regression, SVM
  - ScalaTion currently supports over 40 modeling techniques for prediction and classification
  - ‘caret’ R package currently supports over 200 modeling techniques

- **Model**
  - apply a modeling technique to a *dataset*
  - select *variables*
  - estimate *parameters*

- **Theory**
  - more comprehensive, more explanatory than a model
  - latest initiative for ScalaTion – include theory
Predictive Analytics

- **Regression** – the starting point
  - \( y = f(b)(x) + e \)
  - where \( y \) – response, \( b \) – parameter vector, \( x \) – predictor vector, \( e \) – error/residual
  - **Least Squares**: given \( m \) instances of data, minimize Euclidean norm of \( e = [y_i - f(b)(x_i)]_{i=1,m} \) or generally a loss function
  - If \( f \) is a linear function, \( y = b \) dot \( x \) + \( e \), so
  - \( e = y - Xb \) where \( y = [y]_{i=1,m} \) and \( X = [x]_{i=1,m} \)
  - Solve for parameter vector \( b \) using the **Normal Equations**
  - \( X^tXb = X^ty \) via **Factorization** (Cholesky, QR, SVD)
Theory in Model Development

- **Linear Regression**
  - Selecting **Higher Order** Terms
  - e.g., response $y$ varies with the square of $x_2$
  - $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_2^2$

- **Nonlinear Regression**
  - Selecting **Functional Forms**
  - e.g., **Michaelis-Menten** model for enzyme kinetics
  - $y = b_1 x_1 / (b_2 + x_1)$ or
  - $v = V_{\text{max}} [S] / (K_M + [S])$ in biochemical notation
  - one of several equations for a biochemical pathway
Using Theory in Model Development

- **Regularization**
  - $y = f(b)(x) + e$
  - $\min_b \text{loss}_f(b, x) + \lambda \text{reg}(b, x)$
  - where $\lambda$ is the **regularization parameter**
  - If $\lambda = 0$, then it is case discussed previously

- If $\text{reg}(b, x) = \|b\|_2^2$ it’s **Ridge** regression
- $\|b\|_1$ it’s **Lasso** regression
- constraints derived from theory
Using Theory in Model Development

- **Functional Data Analysis**
  - Given data sampled at n time points \{ (t_j, y_j) \}
  - \( y_j \) is an imprecise measurement at time \( t_j \)
  - Measuring a **continuous** (and **differentiable**) process
  - \( x(t) \) where \( y_j = x(t_j) + e_j \)
  - \( x(t) \) can be represented as a function
  - Flexible approach: linear combination of **basis functions**
  - \( x(t) = c \cdot p(t) \)
  - where \( c = \) vector of coefficients and \( p = \) vector of basis functions
Using Theory in Model Development

- **Basis Functions**
  - **B-Splines**, etc.
  - Useful to represent a function using order 4 (cubic) B-Splines
Smoothing

- Although B-Splines can fit all the data perfectly,
- the goal is to capture the underlying process,
- not to fit the noise.
- This can be done by
  - reducing number of knots (fewer parameters)
  - adding a penalty for lack of smoothness

\[ f(c; \lambda) = ||y - Pc||^2 + \lambda \int [D^2x(t)]^2 dt \]

where \( P = [p(t_j)]_{j=1,n} \) column-wise

\( \min_c f(c; \lambda) \)
Using Theory in Model Development

- **Functional Regression**
  - How do *predictor variables* (e.g., z) affect the *response variable* y
  - Many possible cases: both functional, neither functional, etc.

- Neither functional
  - $y_i = a + bz_i + e_i$

- z functional (e.g., *scalar-on-function* regression)
  - $y_i = a + \int b(t) z_i(t) \, dt + e_i$
Using Theory in Model Development

- **Principal Differential Analysis**
  - Given \( x(t) \) sampled at \( n \) time points \( \{(t_j, y_j)\} \)
  - that is governed by a differential equation
    \[
    Dx(t) - g(b)(x(t)) = 0
    \]
  - where \( g \) is the derivative function
  - Determine the parameters by optimizing
    \[
    \min_{b,c} \| y - Pc \|_2^2 + \lambda \int (Dx(t) - g(b)(x(t)))^2 dt
    \]
  - Instead of pushing the function towards smoothness, push it towards **compliance with the differential equation/theory**
Representing Theory

- **Represent and Solve**
  - Given observed, noisy data \{t_j, x_j, y_j,\}
  - where x and y represent coordinates, and
  - the process is governed by the laws on pendulum motion (differential equation)
    - \( D^2\theta + \frac{(g/l)}{\sin \theta} = 0 \)
    - where \( x = l \sin \theta \) and \( y = l(1 - \cos \theta) \),
  - can easily represent the equations in ScalaTion
  - and use Principal Differential Analysis to
  - estimate the parameters g and l
Representing Theory

- **Algorithms in LaTeX**
  - e.g., Cholesky Factorization

```latex
for i ← 0 until n; j ← 0 to i do
  \text{diff} = a_{ij} - l_i \cdot l_j
  l_{ij} = \text{if } i == j \text{ then sqrt (diff) else diff / } l_{jj}
endfor
```

- suitable for inclusion in papers
- can be almost directly translated into
- Unicode supporting ScalaTion
Representing Theory

- Representing First Order ODEs
  - RL Filter
    - $V = IR + L \frac{DI}{Dt}$
    - $V = V_R + V_L$
  - Newton’s Second Law of Motion
    - $Dx = v$
    - $Dv = F/m$
Representing Theory

Chemical Reaction Network
(www.cs.uga.edu/~thiab/paper25.pdf)

// d[H2]/dt = - kf1 [H2] [O] + kb1 [H] [OH] - kf3 [H2] [OH] + kb3 [H2O] [H]
def dh2_dt (t: Double) = -kf._1*c(0)*c(2) + kb._1*c(3)*c(4) - kf._3*c(0)*c(4) + kb._3*c(5)*c(3)

// d[O2]/dt = - kf2 [H] [O2] + kb2 [O] [OH]
def do2_dt (t: Double) = -kf._2*c(3)*c(1) + kb._2*c(2)*c(4)

Two of eight ODEs in a common pathway.

- Given *rate constants* kf and kb, use integrator to solve for concentrations c over time
- Given time series data, estimate the rate constants using techniques discussed
Representing Theory

- Use ontology, FOL/HOL, to help select
  - integrator (e.g., RungeKutta, DormandPrince),
  - modeling technique (e.g., Nonlinear Regression, Principal Differential Analysis) and
  - optimization method used in parameter estimation

- Automated Modeling Tool: Pret
  - Describe system behavior using a system of nonlinear differential equations
  - given a dataset
  - various reasoning techniques (e.g., qualitative reasoning, nonlinear parameter estimation reasoning) are used to select models
Questions?