Modeling and clustering disease progression for correlation with genetic and demographic factors

Robert Kingan
ProSanos Corp.
April 6, 2005

SSIFT: Stratification and Synchronization Inference Technology

"To address […] common diseases, which include schizophrenia, depression, and breast cancer, it is essential to incorporate observations of the clinical progression of the disease to refine the definition of phenotype." – Michael N. Liebman, U. Penn.

Agenda

- Introduction
- Overview of SSIFT
  - Assumptions—what is SSIFT-able
  - Other constraints on data selection
  - Outline of technique
  - Identifying variables
  - Modeling disease progression
  - Parameterizing different models
  - Clustering patients by progression patterns
  - Interpreting the results
- Example: SSIFT analysis of NIDDK liver transplant data
  - About NIDDK
  - SSIFT and transplant data
  - Variable selection
  - Modeling
  - Results

What makes a dataset "SSIFT-able"?

1. One should be able to model the disease's progression using one or more variables that have an initial value, a period of change, and a final value.
2. The timing of the periods of change between the different variables should be fairly consistent among patients in a particular group.
3. The patterns of change in the variables should be relevant to the patient's outcome.

SSIFT is best suited for the analysis of chronic diseases and other conditions that extend over a great deal of time, for example cancer, diabetes, and organ transplant.

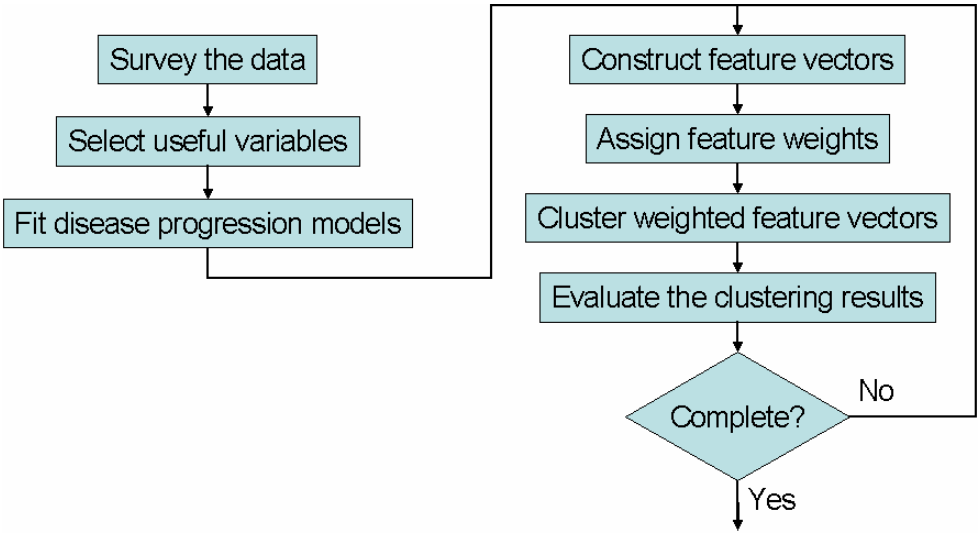

Figure 1: SSIFT workflow

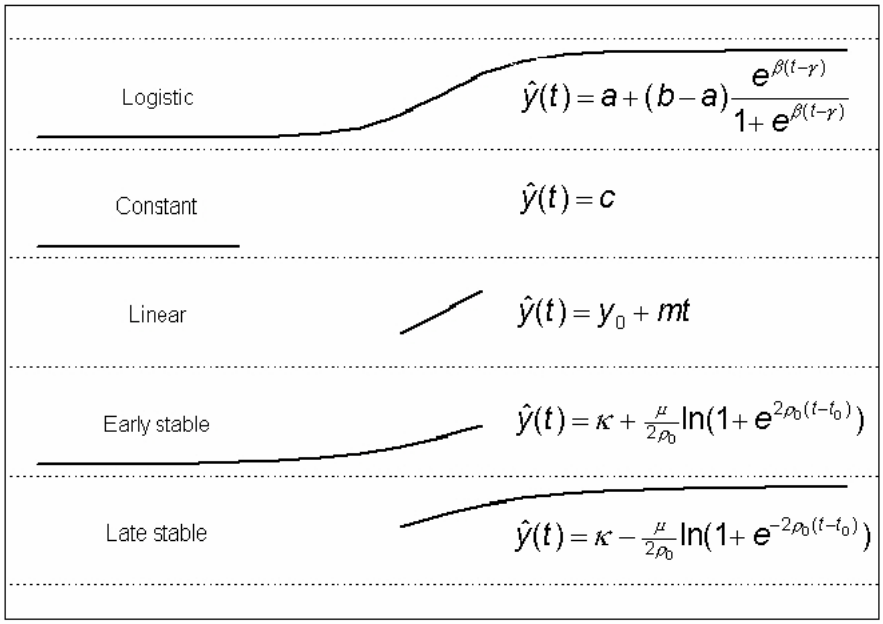

Figure 2: SSIFT curve types

Logistic	$(a,b,m,\Delta) = (a,b,\frac{\beta(b-a)}{4},\gamma - \frac{(a+b)/2 - y^*)}{m})$
Constant	$(a,b,m,\Delta) = (c,c,NULL,NULL)$
Linear	$(a,b,m,\Delta) = (\hat{y}(t_1),\hat{y}(t_n),m,\frac{y^* - y_0}{m})$
Early stable	$(a,b,m,\Delta) = (\kappa,\hat{y}(t_n),\mu,t_0 + \frac{y^* - \kappa}{\mu})$
Late stable	$(a,b,m,\Delta) = (\hat{y}(t_1),\kappa,\mu,t_0 + \frac{y^* - \kappa}{\mu})$

Table 1: Conversion from curve parameters to (a,b,m,Δ)

Modified Mahalanobis distance

$$d(p,q) = \tfrac{1}{2}(v_p - v_q)Q(g(\Sigma_p) + g(\Sigma_q))Q^T(v_p - v_q)^T$$

p,q	patients
v_p, v_q	vector of SSIFT parameters for patients p and q
Q	matrix selecting coordinates based on curve types and scaling result by number of parameters present and parameter weights
Σ_p, Σ_q	variance-covariance matrices for SSIFT parameters for patients p and q
$g()$	generalized inverse function

NIDDK liver transplant dataset:	Inclusion criteria:
• Seven year prospective study of 1563 candidates for liver transplants • Three different medical centers • 88 individual datasets	• First transplant only • At least one year transplant survival • At least three values for all variables used within SSIFT model period (weeks 2—6, month 4 and year 1)

Variable	Fraction of missing values	Variable	Fraction of missing values
α-Fetoprotein	0.80	Hemoglobin	0.03
Albumin	0.37	CSA HPLC level	0.52
Alkaline phosphatase (AP)	0.03	Potassium	0.02
Bicarbonate	0.33	CSA monoclonal level	1.00
Blood urea nitrogen (BUN)	0.04	Sodium	0.03
Calcium	0.24	Platelet count	0.11
Creatinine clearance	0.96	Prothrombin time	0.23
Cholesterol	0.51	Part. thromboplastin CT	0.40
Chlorine	0.13	Part. thromboplastin PT	0.40
Corrected PT control	0.00	CSA RIA level	1.00
Creatinine	0.02	SGOT (AST)	0.04
Direct bilirubin	0.36	SGPT (ALT)	0.20
FK506 level	0.90	Total bilirubin	0.04
Glomerular filtration rate	0.95	CSA TDX level	0.74
Gamma GTP	0.34	Total protein	0.43
Glucose	0.24	White blood cells (WBC)	0.03
Hematocrit (HCT)	0.03	Weight in KG	0.18

Table 2: Candidate variables for NIDDK liver transplant SSIFT analysis

Variable	Log?	$\Delta\hat{S}u$	Weights				$\Delta\hat{S}w$
			a	b	m	Δ	
AST	Yes	0.19	0	1	1	0	0.32
AP	Yes	0.18	0	1	1	0	0.28
Hemoglobin	No	0.13	0	1	0	1	0.24
Total bilirubin	Yes	0.15	0	1	0	1	0.21
Potassium	No	0.20	1	1	1	1	0.20
Hematocrit	No	0.19	1	1	1	1	0.19
WBC	Yes	0.17	1	1	1	1	0.17
BUN	Yes	0.14	1	1	1	1	0.14
Creatinine	Yes	0.12	1	1	1	1	0.12
Sodium	No	0.11	1	1	1	1	0.11

Table 3: Selected variables, parameter weights and $\Delta\hat{S}$ scores

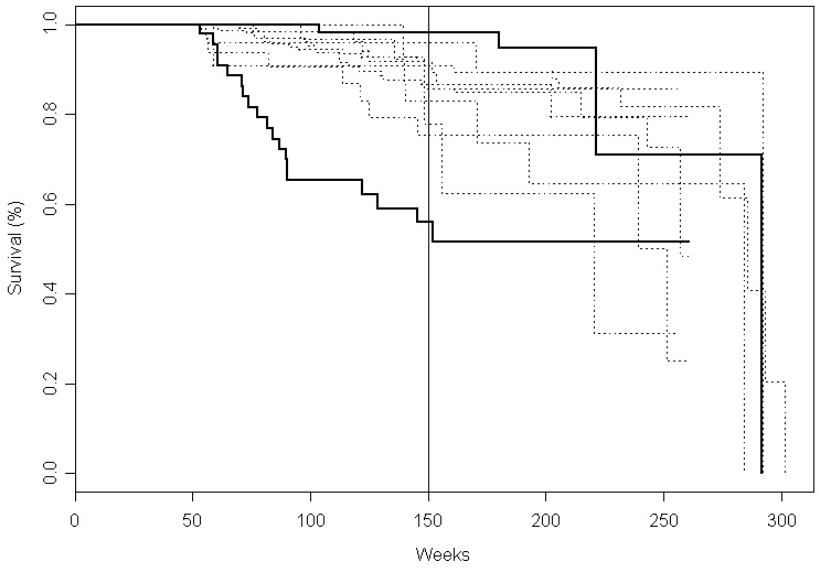

Figure 3: Kaplan-Meier estimated survival curves for liver transplant duration, stratified by clusters of SSIFT parameters for AST, AP and hematocrit, $\Delta\hat{S}=0.42$.

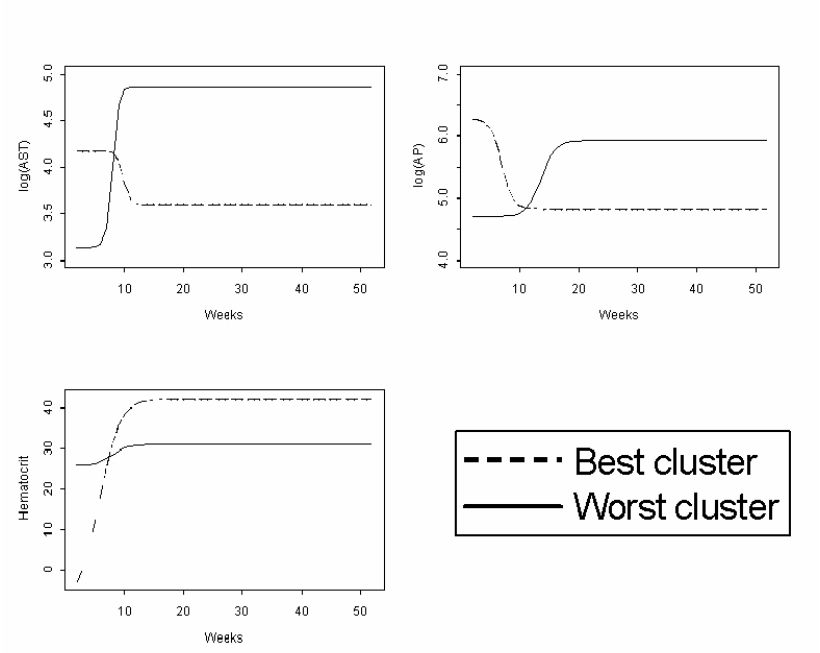

Figure 4: Curves generated from cluster mean values of AST, AP and hematocrit, best and worst clusters

References

Cluster Analysis:
L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data*. Wiley, New York, 1990

Mahalanobis Distance:
Jobson, J. (1991) *Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design*. Springer-Verlag.

NIDDK:
United States National Institute of Diabetes and Digestive and Kidney Diseases, http://www.niddk.nih.gov/

Hampel method for outlier identification:
L. Davies and U. Gather (1993). The identification of multiple outliers. *J. Amer. Statist. Assoc.* **88** (782—801).

Kaplan-Meier curves:
Insightful Corporation (2001). *S-PLUS 6 Guide to Statistics, Volume 2.* Seattle, WA: Insightful Corporation.