Chapter 2

**Interaction Screening for Ultra-High Dimensional Data (Title of iForm paper)**

2.1 Introduction

Variable selection is a heavily researched area in statistical literature. Aspects that need to be considered when selecting predictors for a final model are incorporating meaningful predictors, maintaining accuracy of estimation and the computational cost of conducting the procedure. Fan and Lv (2008) proposed the property of sure independent screening as a way of identifying meaningful variables. Their property ensures that important variables survive after applying variable screening procedures with probability tending to 1.

Recently, Candes and Tao have proposed the Dantzig selector using *L*1-regularization and showed that it achieves the ideal risk up to a logarithmic factor log*.p/*. Their innovative procedure and remarkable result are challenged when the dimensionality is ultrahigh

Sure independence screening : a property that all the important variables survive after applying a

variable screening procedure with probability tending to 1.

Interaction models provide a better approximation to the response surface, improve prediction accuracy, and bring new insight on the interplay between predictors. (Interaction Screening for Ultra-High Dimensional Data, Hoa and Zhang 2014)

*Nelder 1994 on Marginality principle and why to include main effects before interactions*

“ When we fit sequences of quantitative terms such as x1, x2, x1x2, x21, x22,..., we have to ask which sequences make sense. if we fit x1 without an intercept, then the response must go through the origin, i.e. zero must be a special point on the x-scale where y is zero. Similarly, if x21 fitted without an x1 term then the turning-point must occur at the origin (not impossible, but very unlikely). For if x1 might just as well be x1 −a then (x1 −a)2 = x21 −2ax1 +a2 and the linear term re-appears. Both terms must be fitted in the order x1, then x21, and we say that x1 is f-marginal to x21. With two continuous variable x1 and x2, new effects arise: if x1x2 is fitted without x1 and x2 then the response surface must be centered on a col (saddle-point) for the process to make sense. In general there is no reason to expect such a centering to occur, so x1 and x2 must be fitted before x1x2. ...”

Both marginality principle and invariance principle suggest that the selected model should keep the hierarchical structure. (Hao and Zhang 2014; Note on …)

2.2 Model

* iForm Results
  + Adjusted R-Sqaures
  + Marginality principle (M in iForm) offers important guidance for variable selection in interaction models
    - Dynamically selects interaction and main effects to avoid determining multiple cut-offs like in two stage model selections
  + Markers selected in final model
    - Identify cis and trans-QTLs
  + Look for the “hotspots” identified in 2010 c. elegans paper

iForm Selection Algorithm

D is chose as a reasonable upperbound of to terminate the procedure. BIC can also be used for the optimal stopping position. Since the main effects and interaction effects are chosen dynamically we only need to use the selection criteria 1 time.

2.3 Results

2.4 Discussion

5 Conclusion (Hao and Zhang 2014 A Note…)

This note aims to clarify some important issues on variable selection for linear model with interactions. The key concepts and methods presented here also apply to generalized linear models, and models with higher order interaction terms and more complex hierarchical structures. When choosing between main effect models, two-way interaction models or higher interaction models, one needs to consider the bias-variance tradeoff. In general, adding the interactions terms to the model will lower the bias but increase the variance.