Fall 2015 Description & Syllabus for
BST 531: **Nonparametric Inference**

August 31, 2015

Instructor:	Derick R Peterson, PhD
Lectures:	11:00–12:30 MW (08/31/15 – 12/09/15), SRB 1.404
Final Due:	12/16/2015 (Final Project/Homework Due)
E-mail:	`Derick_Peterson@URMC.Rochester.edu`
Phone:	275-6686
Office Hours:	By appointment (SRB 4.220)
Credit Hours:	4.0 (CRN 84959)

Prerequisites (concurrent enrollment acceptable):
Probability, Statistical Inference, Large Sample Theory, and Linear Models (BST 401, 411, 412, and 426).

Required Texts (each of which offers unique treatment, perspective, material, and utility):

1. Hollander M, Wolfe DA, and Chicken E (2014). Nonparametric Statistical Methods. Third Edition.
2. Simonoff JS (1998). Smoothing Methods in Statistics (Corrected Edition).

Additional Smoothing References (I will draw lecture material from several of these):

1. Fan J and Gijbels I (1996). Local Polynomial Modelling and Its Applications.
2. Green PJ and Silverman BW (1993). Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.
3. Hastie T and Tibshirani R (1990). Generalized Additive Models.
4. Hettmansperger TP and McKean JW (2010). Robust Nonparametric Statistical Methods. Second Edition.
5. Loader C (1999). Local Regression and Likelihood.
6. Silverman BW (1986). Density Estimation for Statistics and Data Analysis.
7. Wand MP and Jones MC (1994). Kernel Smoothing.

Workload and Grading
Homework assignments, collected approximately every 3 weeks, will consist of a mix of theoretical proofs and derivations, applied data analyses, basic software development, and small simulation studies.
There will be no exams, but there will be a take-home final project/homework due by 12/16/2015.

Topics (See page 2 for more details):

1. Nonparametric Inference (distribution free hypothesis testing, mostly from an applied perspective, from Hollander, Wolfe, and Chicken)
2. Density Estimation (Simonoff, as well as more details from Loader, Silverman, and Wand & Jones)
3. Nonparametric Regression (Simonoff, as well as more details in Fan & Gijbels, Green & Silverman, Hastie & Tibshirani, Loader, and Wand & Jones)

BST 531 Topics (Nonparametric Inference, Density Estimation, and Nonparametric Regression):

1. *Nonparametric Inference* (Hollander, Wolfe, and Chicken):
   Nonparametric estimation and inference for one-sample location and paired data, two-sample location and/or dispersion, one- and two-way layouts with and without order restrictions, tests of independence, and regression. Exact and large-sample results for some popular procedures, including the sign test and the sample median, the Mann-Whitney-Wilcoxon test and the Hodges-Lehmann location measure, and some generalizations to more complex data structures.

2. *Density Estimation* (Simonoff and others listed above):
   Given an i.i.d. sample $X_1, X_2, \ldots, X_n$ from an unknown distribution F with corrsesponding density f, we can consistently estimate f, as well as its derivatives, under very mild assumptions. Indeed, the familiar histogram, with an appropriately selected binwidth, is one such estimator. Whereas it is sometimes difficult to see the difference between two cumulative distribution functions F_1 and F_2, plots of their corresponding densities f_1 and f_2 are often quite visually distinct. In data analysis situations, estimates of the densities of variables are useful in checking parametric assumptions (such as normality), finding appropriate (symmetrizing) transformations, and simply characterizing and understanding distributions of interest. In addition, estimating the unknown density of a proposed estimator in a simulation study provides great insight into the properties of the estimator –much more insight, for example, than is provided by an estimate of its variance alone. And even if the estimator can be shown to be asymptotically normal, such simulations provide greater understanding of its finite-sample properties. Also, density estimates can often be easily integrated, yielding smooth estimated distribution functions, which are sometimes preferable to empirical distribution (step) functions; one important application is the so-called "smoothed bootstrap." And as with a histogram (binwidth), selection of the smoothing parameter almost always plays a critical role; thus this will be given particular attention, including cross-validation.

   Specific topics include: equal and unequal binwidth histograms and frequency polygons; kernel density estimation, choice of kernel, exact and asymptotic bias, variance, and mean squared error (MSE); bandwidth selection and cross-validation; local likelihood estimation; roughness penalty and spline-based methods; and multivariate density estimation.

3. *Nonparametric Regression Smoothers* (Simonoff and others listed above):
   It is often of interest to estimate the unknown functional relationship m between two variables X and Y, where $Y = m(X) + \epsilon$, and $E(\epsilon|X) = 0$. This can be accomplished using techniques very similar to those used in density estimation. The necessary assumptions, including the existence of some derivatives of m (or, in the case of density estimation, derivatives of f), and practical difficulties, such as choosing the smoothing parameter, are also similar. Thus this is a somewhat natural generalization of the ideas used in density estimation. This is, of course, not too surprising since $m(x) = E(Y|X = x) = \int y f_{Y|X}(y|x)dy$, a smooth functional of a conditional density. And, as in the case of density estimation, generalizations to higher dimensions (i.e. a k-variate function m of k predictors) are possible, though the task requires far more data for similar precision (the so-called "curse of dimensionality").

   Specific topics include: local polynomial regression, derivative estimation, choice of polynomial order, cross-validation, bandwidth selection, and locally varying bandwidths; exact and asymptotic bias, variance, and mean squared error (MSE); roughness penalty and spline-based methods; smoothing correlated data; multivariate nonparametric regression and generalized additive models (GAM).