Report on the PhD dissertation by Luboš PRCHAL entitled : "Selected aspects of functional estimation and testing".

I will start this report by stressing the excellent scientific level of Mr Prchal's work. The manuscript is composed of 6 Chapters and may be split into two parts. The writing is nice and clear. Chapters 1 to 3 address estimation and testing problems in the framework of functional data analysis. Chapters 4 to 6 are devoted to ROC curves. Although both topics belong to the nonparametric domain, they make it necessary to overcome different methods and independent concepts. Besides, an attentive reading of the manuscript reveals the impressive amount of statistical techniques that Mr Prchal displays. This work gave birth to six articles, two of which being accepted and the others under review.

The short Chapter I gives a comprehensive introduction to the regression model for functional inputs : $Y = \int_T \beta_0 (t) X (t) dt + \varepsilon$ where Y and ε are real random variables and X a random function. The reader will find here the basic material to understand the underlying linear inverse problem that should be solved when infering in this model as well as the spline viewpoint.

The setting of Chapter II generalizes to functional outputs the linear model above. The equation is then :

$$Y_i (t) = \int_T \beta (t, s) X_i (s) dt + \varepsilon (t) .$$

The author addresses in parallel the case of autoregression when $Y_i = X_{i+1}$ and proposes a B-spline estimator of the unknwon function $\beta (\cdot, \cdot)$ by classically minimizing a penalized mean-square criterion. Much care is given to computational aspects (matrix approximation of operators, data-driven choice of the penalization parameter...). The convergence of the estimate is mathematically proved and checked through simulations in both regression and autoregression models. A thorough real data study follows which confirms that functional methods are competitive alternatives to classical time-series.

Chapter III concentrates on testing issues. Testing is rather rarely under concern when functional statistics is at stake and I am happy to read an innovative work on this topic which, to me, deserves more attention from the community. Mr Prchal and his co-authors focus on the conditional expectation of two functional random variables : $\mathbb{E} (Y (t) | X = x) = m (t, x)$ and propose to test :

$$H_0 : \forall x, \quad m (t, x) = \mu (t) \quad \text{vs} \quad A_0 : \overline{H}_0,$$

which could be rephrased : is there a significant effect of X on the conditional mean of Y ? Two test statistics are investigated : the first stemming from a classical Fischer's F, the other computed from a sum of square on smoothed residuals. The asymptotic distributions of these statistics under the null are unknown and seem to be hard to achieve. A permutation procedure applied to simulated datasets yields both levels for the tests and powers against single alternatives. A lack-of-fit test may be derived by estimating $m (\cdot, \cdot)$

and applying the no-effect test above to residues. A complete case study is carried out (vertical atmospheric radiation profiles - the data are noisy discretized curves) in minute detail. The function m is assumed to belong to a parametric class profiled by specialists of meteorology and the tests above are duly applied.

Chapter IV introduces nonparametric ROC curves. I appreciate, like in Chapter I, the clarity and skills of Mr Prchal in explaining the grounds and foundations of his research. I recall here that ROC curves appear in supervised classification and are useful tools to monitor the performance of a discrimination method with two issues. Let Y be the classifier, $\mathcal{G}_0$ and $\mathcal{G}_1$ the two classes and set $F_0(t) = \mathbb{P}(Y \leq t|\mathcal{G}_0)$ and $F_1(t) = \mathbb{P}(Y \leq t|\mathcal{G}_1)$. The definition is given at p.91 of the dissertation :

$$ROC(p) = 1 - F_0\left(F_1^{-1}(1-p)\right), \quad p \in [0,1].$$

The main features of theoretical and empirical ROC curves, the invariance and optimality notions are outlined. Some statistical aspects such as kernel estimation, equivalence tests are also pointed out for further purpose.

As explained at the beginning of Chapter V, empirical and kernel ROC curves may significantly differ in situations where F_0 or F_1 are the cdf's of skewed random variables. The main goal of the section is to propose a modification of the kernel estimation to counterbalance these side-effects. The idea consists in applying a well-chosen transformation to the data in order to symmetrize them. The authors have the Box-Cox transformation in mind and study one of its variants. But they go beyond as well and propose a more general approach : an increasing function expressed in a B-spline basis and depending on a unknown parameter $\lambda \in \mathbb{R}^k$. The parameter λ is estimated by minimizing a MISE criterion. Several crucial issues are treated with much care : variance stabilization of the transformed data, simplified version of the minimization program, choice of the distance bewteen ROC curves, followed by a simulation and a real-data (linguistic) study. This chapter is a brilliant illustration of Mr Prchal's versatility.

The last Chapter, Chapter VI namely, gets back to testing problems and especially on testing the equivalence of ROC curves. Given two ROC curves that represent the performance of two classifiers Y and Z, the question under concern is here : are these curves significantly different ? The authors first show that the problem comes down to comparing two increasing transformations τ_0 and τ_1 expressed only in terms of the cdf's of Y and Z. These transformations are estimated on the basis of the sample and give birth to the test statistic :

$$T_n = n \int \left(\widehat{\tau}_0(s) - \widehat{\tau}_1(s)\right)^2 ds.$$

The asymptotic distribution is an infinite quadratic form of independent gaussian random variables ; an algorithm is provided to simulate this distribution and hence get levels. Then, clustering n ROC curves is possible by applying a hierarchical classification to the $n \times n$ proximity matrix (the symmetric matrix that contains the $n(n-1)/2$ p-values obtained by testing all pairs of curves). In addition, they perfect their approach by adding a preliminary step - a "forward" variable selection algorithm applied to a set of classifiers-which finally produces a superclassifier. All along the chapter, this elegant process is enlighted by frequent applications to the linguistic dataset already used in the preceding part.

As a conclusion I would like to underline again my excellent opinion on Luboš Prchal's research work. My global assessment is very positive. The manuscript provides an impressive amount of innovative research work in various fields. Therefore I believe that Luboš Prchal deserves in every respect to get his PhD from the University Toulouse 3.

Montpellier, March, 30, 2008

André Mas,
Université Montpellier 2