FDA Problems that I Like to talk about

Jim Ramsay, McGill University

Overview

- These problems seem to resurface over and over again in interesting applications.
- Most of them haven't received much attention up to now, but progress on some of them is now urgent.
- Naturally they are "hard" problems for reasons such as:
    - they are relatively unfamiliar to statisticians, although rather less so in other branches of the mathemtical, biological and physical sciences
    - they are at this point rather poorly understood
    - progress may require mathematical expertise not normally possessed by statisticians
    - computational challenges can be formidable
    - they may just be really deep
- For these reasons they seem to me to represent exciting opportunities for fearless young researchers

Outline

Time-like and space-like continua

- A time-like continuum, indexed by t, is ordered, and an event located at point t_1 can only be *informative* with respect to an event located at point t_2 if $t_1 \leq t_2$.
- A space-like continuum, indexed by s, locates events s that be informative for all points in a neighborhood of s.
- Let x be an output function with values $x(s, t)$, and u be an input or covariate function with values $u(s, t)$.

Differential equations for time and space

- Let L be a differential operator with respect to t, and M be a differential operator with respect to s.
- Differential equation $Lx = u$ can be used to *forecast* future variation over t.
- Differential equation $Mx = u$ can be used to approximate *neighboring* variation over s.
- *Ordinary differential equations* of the form $Lx = u$ define smooth variation over time.
- *Equilibrium differential equations* of the form $Mx = u$ define smooth variation over space.
- *Partial differential equations* of the form $Lx + Mx = u$ define smooth variation over both space and time.

Why differential equations are important

- Differential equations link the impact of an input function to the current *state* of the system $x(t, s)$ and
- its rate of change defined by some order of derivative.
- They are natural where the capacity of a system to respond to change is limited by one or more *conservation* principles, such as
  - energy (physical or biological)
  - mass
  - heat
  - momentum
  - money supply
- They almost always involve parameters that must be estimated from data.
- But differential equation solutions can seldom be represented explicitly.

Outline

Functional summaries of variation

- Most statistical technology estimates specific functionals of distributions, such as
  - expectation
  - variance
  - a quantile
- Can we estimate the entire distribution of potentially observable data?
- This is a functional parameter problem, usually expressed as nonparametric *density* estimation.

Quantile functions as representation of variation

- But the *quantile function* $Q(u)$, the functional inverse of the distribution function $F(x)$, has important properties that are especially useful, such as
  - Q transforms linearly under linear transformation of data
  - $Q(u)$ immediately specifies the event x associated with risk u
  - a random uniform deviate immediately provides a random x value
- Quantile functions have been championed by John Tukey and Emmanuel Parzen, among others.
- Giles Hooker and I are getting interesting results using monotone function estimation.
- We can estimate quantile surfaces $Q(u; s, t)$ varying smoothly over s and t.
- W. Gilchrist (2000) *Statistical Modeling with Quantile Functions*, CRC Press is a great reference with a strong applied flavour.

Copulas as representations of dependence

- A copula is a multivariate distribution with uniform marginal distributions.
- Copulas have a largely unrecognize potential for being a functional representation of dependency that is free the effects of the marginal distributions.
- Quantile functions are the link between conventional cumulative distribution functions and copulas.
- There has been a great deal of recent progress on copula estimation and their use in inference.

Outline

Challenges

- Current registration methods remove phase variation from data, and amplitude variation is then estimated as a second stage.
- Can we fold estimation of phase and amplitude variation together into the same analysis?
- Alois Kneip and I have developed a joint phase/amplitude procedure for principal components analysis.
- What about doing this for functional linear models, for example?

Outline

Challenges

- Functional data analyses can involve huge numbers of parameters.
- These can often be divided into these classes:
  - *nuisance* (such as coefficients of basis expansions)
  - *structural* (such as parameters defining differential operators)
  - *complexity* controlling parameters (such as bandwidths for smoothing)
- The usual approach to nuisance parameters is to integrate over them with respect to some prior measure.
- But this seldom is possible without expensive Monte Carlo integration

Parameter cascading

- Nuisance parameters are defined as functions $c(\theta)$ of structural parameters.
- Structural parameters are defines as functions $\theta(\lambda)$ of complexity parameters.
- These functional relations are often defined intrinsically by fitting criteria with regularization to enforce smooth functional relations.
- We find that this approach
  - gives parameter estimates as good as those coming from marginalization
  - with much less programming and computational overhead
  - and greater reliability

Outline

1. Modeling functional data with dynamic systems
2. Functional representation of variation, (quantile functions)
3. Representing phase and amplitude variation through registration
4. New strategies for estimating high-dimensional models
5. **Functional methods for covariance estimation**
6. Moving beyond asymptotic performance assessment to perturbation methods
7. Variational representations and computational methods
8. Continued progress on software

Challenges

- In linear and nonlinear mixed effects models, we want to estimate how random effects vary over records or clusters.
- The number N of clusters is often much less than the within-cluster sample sizes n_i.
- We need definitions of inter-cluster variance that are not connected to the number of parameters defining within-cluster random effects.
- We are experimenting with a functional version of the Choleski decomposition, with the triangular factor being defined by piece-wise linear continuous functions defined over hexagonal regions (Lagrange finite elements).

Outline

Challenges

- The tradition in statistics is to assess the performance of statistical methods for asymptotic sample sizes.
- In functional estimation situations, the more data we have, the more complex our models are almost surely going to be.
- Asymptotic performance can be a poor indicator of actual performance (as illustrated by the rather troubled history of smoothing methods.)
- Mapping data perturbations into estimate perturbations is an alternative, often referred to as *fiducial* or *sensitivity* analysis.
- Might not this approach be more informative or at least provide alternative performance indicators?

Outline

Challenges

- The use of basis function expansions converts functional models into multivariate models.
- Sure this works, but there are often perfectly reasonable *variational* approaches expressed in terms of functional analysis that are both more elegant and may bring important advantages.
- Much has already been achieved in representing functional data analysis in strictly functional terms, but a comprehensive treatment is badly needed at this time.

Outline

Challenges for R

- The impact of available, reliable and well documented software on the development of statistics has been enormous.
- Statisticians love R, but R lacks essential infrastructure for functional problems including:
  - quality and variety of methods for numerical approximation of solutions to ordinary and partial differential equations.
  - and the same for optimization
  - R graphics is NOT object oriented
  - R syntax has many design flaws that make debugging code harder than it should be

- But these problems can easily be fixed if
  - software development is given the respect that it deserves as a research activity
  - we get involved ourselves in the development of the language
  - we train our graduate students in software development
  - we are willing to put our research funds behind software development

Other software challenges

- Easily portable code in low-level object languages such as C++ and Java is a big priority.
- A greater variety of data are needed for training and research in functional modeling methods
- Aspects of the numerical analysis of algorithms needs more attention. For example, it is better to compute indefinite integrals by recasting problems as the solution of ordinary differential equations.