Change-Point Detection Methods and Applications. Preliminary program.

September 11-12$^{\text{th}}$, 2008

■ **Thursday 11th**, 9h30-12h00. Chair : *Chair 1.*

9h00-9h30	Welcome coffee	...
9h30-10h30	**M. Lavielle**	A model selection approach for the change-point problem. Application to biomedical signal processing
10h30-10h45	Break	...
10h45-11h45	**D. Siegmund**	BIC Applied to Change-point Model Selection When the Number of Change-points is Large

■ **Thursday 11th**, 13h30-17h00. Chair : *Chair 2.*

13h30-14h30	**N. Akakpo**	Estimating a discrete distribution via histogram selection
14h30-14h45	Break	...
14h45-15h45	**A. Célisse**	Segmentation in the mean of heteroscedastic data via resampling or cross-validation
15h45-16h00	Break	...
16h00-17h00	**P. Perron**	Testing Jointly for Structural Changes in the Error Variance and Coefficients of a Linear Regression Model

■ **Friday 12th**, 9h00-12h20. Chair : *Chair 3.*

9h00-10h00	**N. Zhang**	Detecting Simultaneous Change-points in Multiple Sequences
10h00-10h15	Coffee break	...
10h15-11h15	**J.-Y. Tourneret**	Joint segmentation of wind speed and direction using a hierarchical model
11h15-11h30	break	...
11h30-12h30	**E. Lebarbier**	Joint Segmentation of multivariate Gaussian processes using linear models

■ **Friday 12th**, 14h-17h30. Chair : *Chair 4.*

14h00-15h00	**Y. Guedon**	Exploring the segmentation space for multiple change-point models
15h00-15h15	break	...
15h15-16h15	**E. Terzi**	Problems and algorithms for segmenting sequential data

MARC LAVIELLE, *Univ. Paris XI Orsay*, Marc.Lavielle@math.u-psud.fr

A model selection approach for the change-point problem. Application to biomedical signal processing.

A methodology for model selection based on a penalized contrast is developed. This methodology is applied to the change-point problem, for estimating the number of change points and their location. We aim to complete previous asymptotic results by constructing algorithms that can be used in diverse practical situations.

First, we propose an adaptive choice of the penalty function for automatically estimating the dimension of the model, that is, the number of change points.

In a Bayesian framework, we define the posterior distribution of the change-point sequence as a function of the penalized contrast. A MCMC procedure is shown to be very efficient for sampling this posterior distribution. The parameters of this distribution are estimated with a stochastic version of the EM algorithm (SAEM).

An application to EEG analysis and some Monte-Carlo experiments illustrate these algorithms. The Matlab codes are available at http://www.math.u-psud.fr/~lavielle/programs/.

D. SIEGMUND, *Stanford University*, siegmund@stat.stanford.edu

BIC Applied to Change-point Model Selection When the Number of Change-points is Large.
joint work with N. Zhang.

In a previous paper (*Biometrics*, 2006, pp. 22-32) we derived a Bayes Information Criterion (BIC) for determining the number of change-points in a sequence of independent observations when the number m of change-points is assumed to remain bounded as the number of observations increases. Here we discuss the scenario where the number of change-points increases with the sample size. Compared to the case of bounded m, this scenario seems to be more appropriate for some applications, such as the analysis of DNA copy number data. Whereas in the previous case it seemed reasonably unambiguous where to terminate the asymptotic expansion of the Bayes factors used to define the BIC, in this case there is a conflict between (i) maintaining maximal independence from prior assumptions and (ii) including all terms that are infinitely large in the asymptotic limit. In addition to this conceptual issue in defining the BIC, some theoretically interesting terms, which are functionals of Brownian motion studied previously in change-point detection, arise out of a more detailed analysis.

N. Akakpo, *Univ. Paris XI Orsay*, nathalie.akakpo@math.u-psud.fr

Estimating a discrete distribution via histogram selection. *Joint work with C. Durot.*

We aim at estimating the joint distribution of a finite sequence of independent categorical variables. Given a collection of partitions and the associated histograms, we select from the data a best histogram by minimizing a penalized least-squares criterion. In fact, we only consider a reduced collection of partitions, the partitions into dyadic intervals, a choice inspired from an approximation result due to DeVore and Yu. Our estimator satisfies a non-asymptotic oracle-type inequality and some adaptivity properties in the minimax sense. Moreover, its computational complexity is only linear in the length of the sequence. As an application, we use that estimator during the preliminary stage of a hybrid procedure for detecting multiple change-points in the joint distribution of the sequence. That second procedure still satisfies adaptivity properties. We thus obtain a new algorithm for detecting change-points in DNA sequences that may be very long.

A. Célisse, *AgroParisTech*, celisse@agroparistech.fr

Segmentation in the mean of heteroscedastic data via resampling or cross-validation. *Joint work with S. Arlot.*

We tackle the problem of detecting change-points in the mean of a signal with additive noise. There is neither any distribution assumption about the signal, nor any prior knowledge on the noise level, which can be different from one point to another. The number of changes and their positions are unknown, and we want to estimate them so that the resulting estimator minimizes its quadratic risk. To this aim, we consider several model selection criteria such as cross-validation and resampling penalties that we compare to penalized criteria through simulation experiments. In several situations given the number of change-points, it appears that the breakpoint positions should not be chosen through the usual empirical risk minimization. We obtain quite better estimators by taking into account the variations of the noise level at this step of the model selection procedure. Moreover, we show that some classical penalized least-squares criteria, which were proved to be valid in the homoscedastic case by Birgé and Massart, can fail dramatically when the noise level is indeed depending on the position. On the contrary, cross-validation and resampling methods appear to be quite robust to heteroscedasticity, while having similar performances in the homoscedastic case.

P. Perron, *Boston University*, perron@bu.edu

Testing Jointly for Structural Changes in the Error Variance and Coefficients of a Linear Regression Model. *Joint work with J. Zhou.*

We provide a comprehensive treatment of the problem of testing jointly for structural changes in both the regression coefficients and the variance of the errors in a single equation system involving stationary regressors. Our framework is quite general in that we allow for general mixing-type regressors and the assumptions on the errors are quite mild. Their distribution can be non-Normal and conditional heteroskedasticity is permitted. Extensions to the case with serially correlated errors are also treated. We provide the required tools to address the following testing problems, among others : a) testing for given numbers of changes in regression coefficients and variance of the errors ; b) testing for some unknown number of changes within some pre-specified maximum ; c) testing for changes in variance (regression coefficients) allowing for a given number of changes in the regression coefficients (variance) ; d) sequential procedures to estimate the number of changes present. These testing problems are important for practical applications as witnessed by recent interests in macroeconomics and finance where documenting structural changes in the variability of shocks to simple autoregressions or Vector Autoregressive Models has been a concern. Applications to such macroeconomic time series reinforces the prevalence of changes in both their mean and variance and the fact that for most series an important reduction in variance occurred in the 80s. In many cases, however, the so-called “great moderation" can instead be viewed as a “great reversion".

N. Zhang, *Stanford University*, nzhang@stanford.edu

Detecting Simultaneous Change-points in Multiple Sequences. *Joint work H. Ji, J. Li and D. Siegmund.*

We examine the statistical problem of simultaneous detection in multiple sequences of shared change-points that may occur in only a fraction of the sequences. Motivation arises from the biological application of detecting recurrent intervals of copy number variation in multiple aligned samples of DNA. We consider the following general statistical model. For each sequence $i = 1, \ldots, N$ and position $t = 1, \ldots, T$, the random variables y_{it} are mutually independent and normally distributed with mean values μ_{it} and variances σ_i^2. The null hypothesis states that for every sample i, $\mu_{it} = \mu_i$ for all t, whereas under the alternative there exists $J \subseteq \{1, \ldots, N\}$ and $1 \leq \tau_1 < \tau_2 \leq T$, such that for each $i \in J$, $\mu_{it} = \mu_{i0} + \delta_i I_{\{\tau_1 < t \leq \tau_2\}}$ where $\delta_i \neq 0$. We propose several statistics for this testing scenario, and derive approximations for their significance level and power. Finally, we discuss computational schemes for applying these tests iteratively to detect multiple change-points and describe examples of applications to the detection of DNA copy number variation.

J.-Y. Tourneret, *ENSEEIHT*, jean-yves.tourneret@enseeiht.fr

Joint segmentation of wind speed and direction using a hierarchical model. *Joint work with N. Dobigeon.*

The problem of detecting changes in wind speed and direction is considered. Bayesian priors, with various degrees of certainty, are used to represent relationships between the two time series. Segmentation is then conducted using a hierarchical Bayesian model that accounts for correlations between the wind speed and direction. A Gibbs sampling strategy overcomes the computational complexity of the hierarchical model and is used to estimate the unknown parameters and hyperparameters. Extensions to other statistical models are also discussed. These models allow us to study other joint segmentation problems including segmentation of wave amplitude and direction. The performance of the proposed algorithms is illustrated with results obtained with synthetic and real data.

E. Lebarbier, *AgroParistech*, lebarbie@agroparistech.fr

Joint Segmentation of multivariate Gaussian processes using linear models. *Joint work with E. Budinska, F. Picard, S. Robin and B. Thiam.*

In this presentation, we consider the problem of segmenting jointly multiple series, for which the purpose is to find breaks which are characteristic of individual series as well as breaks that occur in multiple series. We use a mixed linear model to account for both covariates and correlations between signals, a "time" effect being used to catch changes that are common across series. We propose an estimation algorithm based on EM which involves dynamic programming for the segmentation step. We also propose to solve a computational issue that has been raised in the case of segmentation using linear models, and show the computational efficiency of this procedure.

Segmentation methods have been successfully applied to the mapping of chromosomal abnormalities when using CGH microarrays, with an emphasis on cancer genome analysis. Current methods can deal with one CGH profile only, and do not integrate multiple arrays, whereas the CGH microarray technology becomes widely used to characterize chromosomal defaults at the cohort level. In this work, we propose to apply our methodology to the joint characterization ok multiple CGH profiles.

Y. Guédon, *CIRAD*, yann.guedon@cirad.fr

Exploring the segmentation space for multiple change-point models.

With regards to the retrospective or off-line multiple change-point detection problem, much effort has been devoted in recent years to the selection of the optimal number of change points. Here, we explore another research direction which focuses on exploring the space of possible segmentations for successive numbers of change points in an aim of model assessment and model comparison. The knowledge of solely the most probable segmentation of a sequence (for a fixed number of change points) tells us nothing about the remainder of the segmentation space. Questions of interest are :

- Is the most probable segmentation most probable by a long way or are there other segmentations with near-optimal probability ?
- Are these near-optimal segmentations very similar to the most probable segmentation or do they differ greatly ?

Methods for exploring the space of possible segmentations may be divided into two categories : (i) enumeration of possible segmentations, (ii) change-point or segment profiles i.e. possible segmentations summarized in a $J \times T$ array where J is the number of segments and T the length of the sequence. Various dynamic programming and smoothing-type algorithms belonging to these two categories are presented. The dynamic programming algorithms rely on additive contrast functions (e.g. sum of squared deviations from the mean in the Gaussian case) while the smoothing-type algorithms rely on additive log-likelihood functions. Hence, the smoothing-type algorithms apply to a more restricted class of multiple change-point models than the dynamic programming algorithms. Models of this class are characterized by a separability property, i.e. there is no global parameter that depends on within-segment parameters. Due to the deterministic succession of segments, most of the proposed algorithms have transdimensional properties, that is, the output of an algorithm for K segments, with $K = 2, \ldots, J-1$, can be computed as an almost free byproduct of the application of this algorithm for J segments.

The proposed methods are illustrated by examples corresponding to different multiple change-point models. We show using these examples that the proposed methods may help to compare alternative multiple change-point models (e.g. Gaussian model with piecewise constant variances or global variance), predict supplementary change points, highlight overestimation of the number of change points and summarize the uncertainty concerning the location of change points.

E. Terzi, *IBM Almaden Research Center*, eterzi@us.ibm.com

Problems and algorithms for segmenting sequential data.

The analysis of sequential data is required in many diverse areas such as telecommunications, stock-market analysis, and bioinformatics. A basic problem related to the analysis of sequential data is the sequence-segmentation problem. A sequence segmentation is a partition of the sequence into a number of non-overlapping segments that cover all data points, such that each segment is as homogeneous as possible. This problem can be solved optimally using a standard dynamic-programming algorithm.

In the first part of the talk I will present a new approximation algorithm for the sequence-segmentation problem. This algorithm has smaller running time that the optimal dynamic-programming algorithm, while it has bounded approximation ratio.

In the second part of the talk I will give some brief overview of some altternative segmentation models, namely clustered segmentations, segmentations with rearrangements. I will show how segmentation segmentation models can benefit from dimensionality reduction techniques.

Finally, I will discuss the problem of aggregating results of segmentation algorithms on the same set of data points. In this case we are interested in producing a segmentation that agrees as much as possible with the input segmentations. I will demonstrate some interesting algorithms for this problem and present their practical utility.