Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2014)
76, *Part* 3, *pp.* 495–580

Multiscale change point inference

Klaus Frick,

Interstate University of Applied Sciences of Technology, Buchs, Switzerland

Axel Munk

University of Göttingen and Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

and Hannes Sieling

University of Göttingen, Germany

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, October 16th, 2013,* Professor G. A. Young *in the Chair*]

Summary. We introduce a new estimator, the simultaneous multiscale change point estimator SMUCE, for the change point problem in exponential family regression. An unknown step function is estimated by minimizing the number of change points over the acceptance region of a multiscale test at a level α. The probability of overestimating the true number of change points K is controlled by the asymptotic null distribution of the multiscale test statistic. Further, we derive exponential bounds for the probability of underestimating K. By balancing these quantities, α will be chosen such that the probability of correctly estimating K is maximized. All results are even non-asymptotic for the normal case. On the basis of these bounds, we construct (asymptotically) honest confidence sets for the unknown step function and its change points. At the same time, we obtain exponential bounds for estimating the change point locations which for example yield the minimax rate $\mathcal{O}(n^{-1})$ up to a log-term. Finally, the simultaneous multiscale change point estimator achieves the optimal detection rate of vanishing signals as $n \to \infty$, even for an unbounded number of change points. We illustrate how dynamic programming techniques can be employed for efficient computation of estimators and confidence regions. The performance of the multiscale approach proposed is illustrated by simulations and in two cutting edge applications from genetic engineering and photoemission spectroscopy.

Keywords: Change point regression; Dynamic programming; Exponential families; Honest confidence sets; Multiscale methods

1. Introduction

Assume that we observe independent random variables $Y = (Y_1, \ldots, Y_n)$ through the exponential family regression model

$$Y_i \sim F_{\vartheta(i/n)}, \qquad i = 1, \ldots, n, \tag{1}$$

where $\{F_\theta\}_{\theta \in \Theta}$ is a one-dimensional exponential family with densities f_θ and $\vartheta : [0, 1) \to \Theta \subseteq \mathbb{R}$ a right continuous step function with an unknown number K of change points. Figs 1(a) and 1(b) depict such a step function with $K = 8$ change points and corresponding data Y for the Gaussian family $F_\theta = \mathcal{N}(\theta, \sigma^2)$ with fixed variance σ^2.

Address for correspondence: Hannes Sieling, Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstrasse 7, 37077 Göttingen, Germany.
E-mail: hsielin@math.uni-goettingen.de

 1369–7412/14/76495

that it was comprehensive, in terms of developing a method for detecting change points, the associated theory for the method and efficient approaches to implementing the method in practice. Furthermore the paper did not just consider point estimates for the location of change points, but also how to construct confidence intervals for their location, and confidence bands for the underlying function. It is this latter aspect of the paper that I would like to comment on.

One issue with constructing confidence intervals for change point problems is the difficulty in defining what such an interval should be when there is uncertainty over how many change points there are. As can be seen in Fig. 1(c), the likely location of change points can vary substantially as you change their number. The approach in the paper seems to circumvent this by working with asymptotic confidence intervals, and using the fact that the method gives a consistent estimate of the number of change points. As a result, asymptotically you know the number of change points in the data, and hence this issue vanishes. This asymptotic behaviour, however, is different from that in many real life applications where there is substantial uncertainty over the number of change points. So how meaningful are the confidence intervals in practice? How is a practitioner to interpret a figure like Fig. 1(c), particularly as the confidence level is inherently linked with the number of change points detected.

These issues seem particularly prevalent in applications with both large numbers of observations and many change points. Would pointwise confidence bands for the underlying function be more appropriate than simultaneous confidence bands in these cases? In situations where there is substantial uncertainty about the number of change points, are there any real alternatives to Bayesian approaches (e.g. Yao (1984), Barry and Hartigan (1993), Fearnhead and Liu (2007) and Fearnhead and Vasileiou (2009)) if you want to quantify the full uncertainty of any estimates?

Cheng-Der Fuh and Huei-Wen Teng (*National Central University, Jhongli City*)

We congratulate Professor Frick, Professor Munk and Dr Sieling for providing a comprehensive and challenging paper for multiple-change-point detection. There remain interesting problems that have not been addressed or have not yet been completely solved, which deserve further consideration.

(a) Besides the moving average MA(l) model mentioned in the paper for dependent data, another interesting model is the hidden Markov model. We assume that X_n is a Markov chain on a finite state space $D=\{1,\ldots,d\}$, with initial distribution $P(X_0=x_0)=\pi_\theta(x_0)$, and transition probability $P(X_n=x_n|X_{n-1}=x_{n-1})=p_\theta(x_{n-1},x_n), n=1,2,\ldots$. The observation Y_n at time n is continuous with density function $f_\theta(Y_n|x_n), n=1,2,\ldots$ (all parameterized by θ). Let K denote the number of change points, $\tau_1<\tau_2<\ldots<\tau_k$ be the change points and $\theta_1,\ldots,\theta_k$ be the associated parameters. The problem of interest is that the state transition probability and observation model conditioned on the state undergo a change from θ_i to θ_{i+1} at the change point τ_i for $i=1,\ldots,K$. In particular, the state transition from x_{τ_i-2} to x_{τ_i-1} is described by θ_{i-1}, whereas the transition from x_{τ_i-1} to x_{τ_i} is described by θ_i. Similarly, observation model $f_\theta(Y_{\tau_i-1}|x_{\tau_i-1})$ is described by θ_{i-1}, whereas $f_\theta(Y_{\tau_i}|x_{\tau_i})$ is described by θ_i. This model is suitable for a setting where the observation is causally related to the state and, hence, a change in the state transition leads to a change in the observation density. In the case of $K=1$, the joint density of $Y=\{Y_1,\ldots,Y_n\}$ is given as

$$P(Y)=\begin{cases}\sum_{x_0,x_1,\ldots,x_n}\prod_{i=1}^{n} p_{\theta_1}(x_{i-1},x_i)\,f_{\theta_1}(Y_i|x_i) & \text{if } n\leqslant\tau_1-1,\\ \sum_{x_0,x_1,\ldots,x_n}\pi_{\theta_1}(x_0)\,A_{\tau_1}(Y), & \text{if } n\geqslant\tau_1,\end{cases}$$

where

$$A_{\tau_1}(y)=\prod_{i=1}^{\tau_1-1} p_{\theta_1}(x_{i-1},x_i)\,f_{\theta_1}(Y_i|x_i)\prod_{i=\tau}^{n} p_{\theta_2}(x_{i-1},x_i)\,f_{\theta_2}(Y_i|x_i).$$

The joint density for general K can be defined in a similar way.

Another setting of change point detection is that the change point is defined as the change of states for the underlying Markov chain X_n. These models are called regime switching models when the Markov chain is recurrent or are called change point models when the Markov chain is not recurrent.

(b) In finance applications, the feature is to incorporate both the stock and the options prices to identify the regime of volatility. Specifically, the option price is modelled via a variance stabilizing transform

$\log(y_{ij}) = \log\{C_{ij}(\Theta)\} + \varepsilon_{ij}$, where $C_{ij}(\Theta)$ is the model price of the option with i indicating the type of the option and j indicating the strike price. Here, the error ε_{ij} comes from market friction or model discrepancy. It is a challenging task to form change point detection using implied volatility calibrated from option prices, the volatility of the option price calculated from the error ε_{ij} and the historical volatility of the stock price.

It is interesting to see whether the simultaneous multiscale change point estimator proposed in the paper can be applied to these two problems. Alternatively, a Bayesian approach focuses on the posterior distribution of the parameters, in which Markov chain Monte Carlo methods can be implemented to sample the parameter of interest having the desired distribution.

Dario Gasbarra and Elja Arjas (*University of Helsinki*)
The paper follows consistently the frequentist statistical paradigm. Thus, for example, the number K of change points is viewed as a parameter, and consideration of different values for K is interpreted as a problem of model selection. A natural alternative, particularly in 'dynamic' applications where the data consist of noisy measurements of a signal over time, is to think of the signal as a realization of a latent stochastic process. Then the mode of estimation changes from an optimization problem to a problem formulated in terms of probabilities. Adopting a hierarchical Bayesian approach to the problem allows us similarly to quantify probabilistically also parameter uncertainty. As an illustration, we computed the posterior probabilities $P(K=k|\text{data})$ for the 'no-trend' signal (see Fig. 4(a)), using data corrupted by Gaussian noise as in the paper. For this, we applied a slightly modified version of the Gibbs sampler algorithm for marked point processes of Arjas and Gasbarra (1994).

Quite a vague prior distribution was assumed, with non-informative scale invariant priors $\pi(\sigma^2) \propto \sigma^{-2}$ and $\pi(\lambda) \propto \lambda^{-1}$respectively for the variance of the Gaussian error terms and for the Poisson intensity governing the number and the positions of the change points. For the signal levels, we adopted a Gaussian $\mathcal{N}(0.5, 10)$ prior for the initial level, and a conditionally independent and identically distributed Gaussian prior $\mathcal{N}(0, \eta^2)$ for the successive jumps, assuming a scale invariant non-informative prior also for η^2.

With these model and prior specifications the posterior probability $P(K=k|\text{ data})$ for the number of change points had its maximum value 0.41 at $k=6$, which was the true number used in the simulation. By constraining the jump sizes of the signal above the threshold 0.15, we obtained also the constrained conditional probability $P(K=6|\text{data, constraint}) = 0.78$.

The 0.95 credible interval for σ was (0.19, 0.21), the correct value being $\sigma=0.2$. Additional results, and the computer code that was used, are available from the first author.

It would be interesting to see a systematic comparison of the performance of the various methods for solving change point problems based on the two complementary statistical paradigms.

M. Hušková (*Charles University in Prague*)
I congratulate the authors for excellent work, which is important for applications. They consider a relatively simple model with multiple changes (independent observations; one-dimensional exponential family for single observations) but the paper is quite complex—it covers motivations, theoretical results, deep discussion of computational aspects, simulations, application to real data sets, discussions on single steps and discussions on possible extensions.

The disadvantage is that the paper is too long (but I have no suggestion how to shorten it) and it is difficult to extract the algorithms if one wants to apply the procedures without reading the whole paper in detail—kinds of algorithms or pseudocode would be useful.

I would like to bring your attention to Antoch and Jarušková (2013), where a different model is considered, and different procedures. It contains both theoretical results and algorithms with useful pseudocode.

Jiashun Jin (*Carnegie Mellon University, Pittsburgh*) **and Zheng Tracy Ke** (*Princeton University*)
We congratulate the authors on a very interesting paper. The paper sheds lights on a problem of great interest, and the theory and methods developed are potentially useful in many applications.

In Ke *et al.* (2013) we have investigated the problem from the variable selection perspective. Consider a linear model

$$Y = X\beta + z, \qquad X = X_{n,n}, \quad z \sim N(0, I_n), \tag{47}$$