Journal of Hydrology 538 (2016) 63–70

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Bayesian analysis to detect abrupt changes in extreme hydrological processes

Seongil Jo [a], Gwangsu Kim [a], Jong-June Jeon [b,*]

[a] Department of Statistics, Korea University, Seoul, Republic of Korea
[b] Department of Statistics, University of Seoul, Seoul, Republic of Korea

A R T I C L E I N F O

Article history:
Received 9 July 2015
Received in revised form 17 January 2016
Accepted 29 March 2016
Available online 6 April 2016
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Felix Frances, Associate Editor

Keywords:
Bayesian change-point analysis
Non-stationarity
Generalized extreme-value distribution

S U M M A R Y

In this study, we develop a new method for a Bayesian change point analysis. The proposed method is easy to implement and can be extended to a wide class of distributions. Using a generalized extreme-value distribution, we investigate the annual maximum of precipitations observed at stations in the South Korean Peninsula, and find significant changes in the considered sites. We evaluate the hydrological risk in predictions using the estimated return levels. In addition, we explain that the misspecification of the probability model can lead to a bias in the number of change points and using a simple example, show that this problem is difficult to avoid by technical data transformation.

1. Introduction

Non-stationarity in the hydrological processes has received much attention since (Wigley, 1985), who referred the importance of capturing non-stationarity in prediction. Using a simple example, he showed that ignoring non-stationarity in the probability model can lead to severe biases in high quantiles. His results provoked several studies on non-stationary models using extreme data in hydrology and climatology. To this effect, the Bayesian change point (BCP) analysis is one of the most popular research topics in the field.

In hydrology, Perreault et al. (2000a,b,c) studied a BCP model under a (multivariate) normal distribution. Rasmussen (2001) constructed a Bayesian regression model with abrupt changes in the intercept and slope parameters and (Seidou et al., 2007) generalized the works of Perreault et al. (2000c) and Rasmussen (2001). Kim et al. (2009) investigated the change points of the annual maximum precipitation across South Korean Peninsula. They used regional averages of annual maximum precipitations across multiple sites in South Korea, and assumed the normal distribution of the regional averages.

In climatology, Chu and Zhao (2004) proposed a BCP model to detect a single mean change in the Poisson distribution. The proposed method is based on Bayesian hypothesis testing, which compares the model evidences between a single change and no change. They analyzed frequency of annual occurrences of a tropical cyclone and showed significant evidences of existence of change points in the mean of the distribution. Zhao and Chu (2006) extended the Bayesian model to detect multiple change points in the occurrences of hurricane activities. Their model is also based on the Bayesian hypothesis testing in which hypotheses are predetermined. Tu et al. (2009) applied the Bayesian model to analyze the occurrences of typhoons, heavy rainfall, and heat waves.

The Bayesian model provides statistical evidence that supports the non-stationarity of the climatological or hydrological processes and offers decision tools for forecasting. However, to the best of our knowledge, most of studies that adopt the BCP model assume distributions in the exponential family as a probability model of observations. Moreover, they are largely based on Bayesian hypothesis testing which is computationally intensive and restricted to models with a pre-determined number of change points. To overcome these weakness (Zhao and Chu, 2010) introduced a prior distribution of the number of change points in the Bayesian model. They applied the reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995), which is widely used to control the number of mixture components.

As an alternative to the BCP model (Zhao and Chu, 2010), we propose a new model with a spike and slab prior distribution of regression parameters. In this study, a generalized extreme value

* Corresponding author.
E-mail address: jj.jeon@gmail.com (J.-J. Jeon).

http://dx.doi.org/10.1016/j.jhydrol.2016.03.065
0022-1694/

(GEV) distribution is considered as the probability model of the observations; however, the probability distribution of the model can be extended to a sufficiently large class of useful distributions. By estimating differences in the mean parameters, we detect the change points of parameters in a time domain.

The remainder of this paper is organized as follows. In Section 2, we briefly review the recent literature on change-point analyses, and explain theoretical connections between our research and the developed models. In Section 3, we propose the new Bayesian method using spike and slap prior. This section also explains the way implementation of the MCMC algorithm using OpenBUGS with an underlying GEV distribution. In Section 4, we conduct a numerical analysis of annual maximum of precipitation in the South Korean Peninsula. In Section 5, we provide concluding remarks.

2. Background of change point analysis

This section discusses the recent literature on change-point analyses and shows the relationship between existing studies and the present one. There are two major approaches to analyze extreme events using a non-stationary distribution. First is to find gradual varying changes and second is to detect abrupt changes in the stochastic quantities. The first approach originates from a generalized linear model, which is a regression model for distributions in the exponential family. In particular, when the assumption of the underlying probability model is a GEV distribution, the stationary model can be easily extended to a non-stationary model, reflecting changes in the mean trend and variance by introducing time-dependent parameters (Coles et al., 2001). The second approach is from the CUSUM test (Page, 1954), which provides statistical decisions about mean changes in production processes. The statistical evidence for the existence of a mean change is evaluated using the test statistics from the CUSUM test and a new estimation is conducted accounting for the estimated change points.

The regularizing method of the regression model has methodologically bridged two different approaches. The new method serves as a tool to detect one or multiple change points using a fused lasso (Tibshirani et al., 2005). The fused lasso method shrinks the differences between the adjacent estimated parameters in a time domain, such that the estimators are regularized along the direction of parameter averages. Using the fused penalty function, identical estimated coefficients for a neighborhood are obtained. Thus, the fused lasso provides estimators with consecutive homogeneous groups of mean parameters, which contain information about the change points. The primary advantage in using a fused lasso is its flexibility or easy extension to other probability models that detect change points. By replacing the loss function used in the fused lasso, various models that capture change points can be developed.

In Bayesian statistics, the regularizing method affects the development of new prior distributions. For a linear regression, Park and Casella (2008) and Hans (2009) explained lasso Tibshirani (1996), one of the most popular regularization methods, as a maximum a posteriori (MAP) estimate with independent Laplace priors for regression coefficients from a Bayesian perspective. Using a new class of prior distributions, Kyung et al. (2010) explain the group lasso, fused lasso, and elastic net, which are generalizations of the lasso within the framework of hierarchical Bayesian models. The family of prior distributions is motivated by the regularization method and is called the shrinkage priors, which include the Laplace (double exponential) distribution, Student's t-distribution, generalized double Pareto distribution, and horseshoe-type priors distribution (Bae and Mallick, 2004; Johnstone and Silverman, 2004; Carvalho et al., 2010; Armagan et al., 2013). The shrinkage priors provide the Bayesian models with a useful framework to achieve variable selections in the regression model.

Corresponding to the variable selection problem in the Bayesian regression model, Armagan et al. (2013) and Castillo et al. (2015) proposed a more elaborated model that uses a mixture of continuous shrinkage priors and discrete distributions. Generally, with conventional prior distributions in a regression model, it is difficult to explain variable selections using the obtained posterior distribution. However, using the mixture distribution, the theoretical foundation of the variable selection in the Bayesian regression model is constructed and the selection appears to perform well. The proposed model is motivated by the extension of the fused lasso (Tibshirani et al., 2005) to a Bayesian model using this mixture prior distribution. We define change point by time with non-zero differences in the adjacent parameters and assume spike and slab prior distributions (Ishwaran and Rao, 2005) of the differences in parameters. The spike and slab prior distribution is a type of selection prior distribution, which is a mixture of distributions with a point mass at zero and a Laplace distribution. The proposed model provides an estimation result for the abrupt change in stochastic quantities, which does not follow the normal distribution if the appropriate likelihood function is defined. In addition, the implementation is rather simple, since the OpenBUGS (http://www.openbugs.net/w/FrontPage), which is the most popular program used in a Bayesian analysis, supports the proposed model's MCMC method.

3. Bayesian change point model

3.1. Likelihood

A Generalized extreme value (GEV) distribution is defined by a limiting distribution of block-wise maxima of identical and independent random quantities. A GEV distribution has three parameters corresponding to the location (μ), scale (σ), and shape (ξ) of its probability density function. The cumulative distribution function is given by

$$F(y|\mu,\sigma,\xi) = \begin{cases} \exp[-\{1+\xi(y-\mu)/\sigma\}^{-1/\xi}], & \texttt{If} \quad \xi \neq 0 \\ \exp[-\exp\{-(y-\mu)/\sigma\}], & \texttt{If} \quad \xi = 0 \end{cases}$$

where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter, and ξ is the real valued shape parameter. The sign of ξ determines the support of the distribution. If $\xi > 0$, the support of the distribution is bounded below $\mu - \sigma/\xi$; if $\xi < 0$, the support is bounded above by $\mu - \sigma/\xi$. The kth moment of the distribution exists only for $\xi < 1/k$, and the variance and skewness are defined for $\xi < 1/3$. As long as variance and skewness exist, they do not depend on μ. On the other hand, the α-quantile and the mean of the distribution are affine functions of μ. Thus, a GEV distribution with a changing mean/median can be constructed using the location parameter depending on time. Let $y_t \sim \mathrm{GEV}(\mu_t, \sigma, \xi)$ for $t = 1, \ldots, n$, then the α-quantile of y_t is given by

$$F^{-1}(\alpha|\mu_t,\sigma,\xi) = \begin{cases} \mu_t + \sigma\frac{(-\log(\alpha))^{-\xi}-1}{\xi}, & \texttt{If} \quad \xi \neq 0 \\ \mu_t + \sigma\log(-\log(\alpha)), & \texttt{If} \quad \xi = 0 \end{cases}$$

and the mean of y_t is given by

$$\mathrm{E}[y_t|\mu_t,\sigma,\xi] = \begin{cases} \mu_t + \sigma\frac{\Gamma(1-\xi)-1}{\xi}, & \texttt{If} \quad \xi < 1 \\ \mu_t + \sigma\gamma, & \texttt{If} \quad \xi = 0 \\ \infty, & \texttt{If} \quad \xi \geqslant 1 \end{cases}$$

The conditional likelihood of $\mathbf{y} = (y_1, \ldots, y_n)$ given $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, σ, and ξ is written as