Developing a Bayesian method for locating breakpoints in time series data.

Kathryn Haglich [1], Jeff Liebner [2], Sarah Neitzel [3], and Amy Pitts [4]

[1] Department of Mathematics, Lafayette College, Easton, Pennsylvania, USA

[2] Department of Mathematics, Faculty of Mathematics, Lafayette College, Easton, Pennsylvania, USA

[3] School of Biodiversity Conservation, Unity College, Unity, Maine, USA

[4] Department of Mathematics, Marist College, Poughkeepsie NY, USA

Address for correspondence: Arnošt Komárek, Department of Mathematics, Faculty of Mathematics, Lafayette College, Easton, Pennsylvania, USA.

E-mail: `liebnerj@lafayette.edu`.

Phone: (+420) 221 913 282.

Fax: (+420) 222 323 316.

Abstract: An abstract of up to 200 words should precede the text together with 5 or 6 keywords in alphabetical order to describe the content of the paper. Authors should take great care in preparing the abstract and not simply lift it from the main text. The abstract should describe the background and contribution of the manuscript and give a clear verbal description of the results and examples, and avoid citations

whenever possible. Any acknowledgements will be printed at the end of the text.

Key words: keyword a; keyword b; keyword c; keyword d; keyword e

1 Introduction

2 Math

τ is the location of breakpoints

K is the number of breakpoints

β is the regression coefficients

σ is the standard deviations

2.1 Derivations of Ratio

To start we need to find the ratio

$$\begin{aligned} ratio &= \frac{g(\tau_n K_n | x_1, \ldots, x_t)}{g(\tau_o K_o | x_1, \ldots, x_t)} \times \frac{q(\tau_o K_o | \tau_n K_n)}{q(\tau_n K_n | \tau_o K_o)} \\ &= \frac{\left[\frac{f(x_1,\ldots,x_t|\tau_n K_n)\pi(\tau_n K_n)}{\int f(x_1,\ldots,x_t|\tau_n K_n)\pi(\tau_n K_n) d\tau_n K_n}\right] q(\tau_o K_o | \tau_n K_n)}{\left[\frac{f(x_1,\ldots,x_t|\tau_o K_o)\pi(\tau_o K_o)}{\int f(x_1,\ldots,x_t|\tau_o K_o)\pi(\tau_o K_o) d\tau_o K_o}\right] q(\tau_n K_n | \tau_o K_o)} \end{aligned}$$

Then we have,

$$ratio = \frac{\left[\frac{\int f(x_1,\ldots,x_t|\tau_n K_n)\left(\pi(\tau|K)\pi(K)\pi(\beta)\pi(\sigma)\right) d\tau dK}{\int f(x_1,\ldots,x_t|\tau_n K_n)\left(\pi(\tau|K)\pi(K)\pi(\beta)\pi(\sigma)\right) d\tau dK d\beta d\sigma}\right] q(\tau_o K_o | \tau_n K_n)}{\left[\frac{\int f(x_1,\ldots,x_t|\tau_o K_o)\left(\pi(\tau|K)\pi(K)\pi(\beta)\pi(\sigma)\right) d\tau dK}{\int f(x_1,\ldots,x_t|\tau_o K_o)\left(\pi(\tau|K)\pi(K)\pi(\beta)\pi(\sigma)\right) d\tau dK d\beta d\sigma}\right] q(\tau_n K_n | \tau_o K_o)}$$

Basing priors of the paper written by Kass, DiMatteo, and Genovese (2001) we have that

$\pi(\theta) = \pi(\tau|K)\pi(K)\pi(\beta)\pi(\sigma)$ for both the θ_n and the θ_o.

$\pi(\beta)$ is an unit information prior, multivariate normal

$\pi(\sigma)$ is an inverse gamma

$\pi(\tau|K)$ might be uniform

$\pi(K)$ is a Poisson

2.1.1 BIC

The Metropolis Hastings algorithm is used to determine the acceptance of the proposed breakpoint set. The general ratio is the product of the Bayes factor, determined by the ratio of the posteriors, g, and the ratio of the Markov Chain Monte Carlo (MCMC) proposal densities, q, whose values depend on the current MCMC step.

$$ratio = \frac{g(\tau_n K_n|x_1,\ldots,x_t)}{g(\tau_o K_o|x_1,\ldots,x_t)} \times \frac{q(\tau_o K_o|\tau_n K_n)}{q(\tau_n K_n|\tau_o K_o)}$$

To be able to adequately analysis these ratios we need to put the ratio on a logarithmic scale.

$$\begin{aligned} log(ratio) = &\Big[log\big[g(\tau_n K_n|x_1,\ldots,x_t)\big] - log\big[g(\tau_o K_o|x_1,\ldots,x_t)\big]\Big] \\ &+ \Big[log\big[q(\tau_o K_o|\tau_n K_n)\big] - log\big[q(\tau_n K_n|\tau_o K_o)\big]\Big] \end{aligned}$$

As proved by Kass & Wasserman (1995), the log of the Bayes Factor can be approximated with BIC with an error on the order of $O(n^{-1/2})$ when the data size is greater than 25 and the prior follows a normal distribution. Therefore,

$$log\big[g(\tau_n K_n|x_1,\ldots,x_t)\big] - log\big[g(\tau_o K_o|x_1,\ldots,x_t)\big] \approx \frac{-\Delta BIC}{2}$$

which means that

$$log(ratio) \approx \frac{-\Delta BIC}{2} + \Big[log\big[q(\tau_o K_o|\tau_n K_n)\big] - log\big[q(\tau_n K_n|\tau_o K_o)\big]\Big]$$

In the case of addition,

$$q(\tau_o K_o|\tau_n K_n) = c \cdot d \cdot Poisson(K_{old}, \lambda), \quad q(\tau_n K_n|\tau_o K_o) = c \cdot b \cdot Poisson(K_{old}, \lambda).$$

When the chosen MCMC step is subtraction,

$$q(\tau_o K_o|\tau_n K_n) = c \cdot b \cdot Poisson(K_{new}, \lambda), \quad q(\tau_n K_n|\tau_o K_o) = c \cdot d \cdot Poisson(K_{old}, \lambda).$$

Given the equations above we have that c is the combined probability of doing an addition and subtraction step. b is the balancing birth coefficient and d a balancing death coefficient. They are in place to set the ratio of birth step to death steps. Specifically,

$$b = \frac{A_{start}}{A_{start} + K_{start} + 1} \times \frac{1}{A}$$

$$d = \frac{K_{start}}{A_{start} + K_{start} + 1} \times \frac{1}{K}$$

The first fraction is based off of starting conditions and the second fraction changes through each step. We have that A_{start} is the starting number of available spaces. An available space is any data point that is not itself a breakpoint, and endpoint, or 2 points away from an existing breakpoint. K_{start} is the starting number of breakpoints that is proposed before the function is called.

2.2 Derivations of β and σ draws

The posterior for the β coefficient is laid out in detail in the *Forecasting time series* (Pesaran Paper, 2006). Given that b_0 is the mean of the βs, B_0 is the variance co-variance matrix of the βs for the prior. Also v_0 and d_0 are the parameters of the inverse gamma prior of the inverse gamma squared (one being the shape the other rate). While S_t is the current state of the break locations, and y_t is the actual data values.

$$\beta|\sigma^2, b_0, B_0, V_0, d_0, S_t, y_t \sim N(\overline{\beta_j}, \overline{V_j})$$

Where

$$\overline{V}_j = (\sigma^{-2}x^T x + B_0^{-1})^{-1}, \quad \overline{\beta}_j = \overline{V}_j(\sigma^{-2}x^T y_t + B_0^{-1}b_0)$$

The *Forecasting time series* (Pesaran Paper, 2006) also lays out the σ posterior such that

$$\sigma_j^{-2} \sim \Gamma(v_0, d_0) \longrightarrow \sigma_j^{-2}|\beta, b_0, B_0, v_0, d_0, S_t, y_t \sim \Gamma(\overline{v}_0, \overline{d}_0)$$

Where

$$\overline{v}_0 = v_0 + \frac{n_j}{2}, \quad \overline{d}_0 = d_0 + \frac{1}{2}(y_t - x\beta)^T(y_t - x\beta)$$

3 Methods

4 Results

5 Discussion

6 Appendix

Acknowledgements

We want to thank...