GEOPHYSICAL RESEARCH LETTERS, VOL. 26, NO. 18, PAGES 2801-2804, SEPTEMBER 15, 1999

One slope or two? Detecting statistically significant breaks of slope in geophysical data, with application to fracture scaling relationships

I.G. Main

Department of Geology & Geophysics, University of Edinburgh, Edinburgh, UK

T. Leonard, O. Papasouliotis

Department of Mathematics and Statistics, University of Edinburgh, Edinburgh, UK

C.G. Hatton

Stockton Campus, University of Durham, Stockton-on-Tees, UK

P.G. Meredith

Department of Geological Sciences, University College London, London, UK

Abstract. The scaling of displacement as a function of length is important for a variety of applications which depend on the mechanical and hydraulic properties of faults and fractures. Recently it has been suggested that the power-law exponent ν which has been found to characterise this relationship may change significantly at a characteristic length for a variety of reasons, for example when cracks begin to interact, or when faults grow to a length comparable to a characteristic size in the brittle layer. Such a break of slope requires a second straight line, requiring two extra model parameters. Here we present a new method for analysing such data, which penalises the extra parameters using a modified form of Schwarz's Information Criterion, and a Bayesian approach which represents uncertainty in the unknown parameters. We apply the method to data from the Krafla fissure zone in the north of Iceland, and find a significant break of slope, from $\nu \approx 3/2$ to $\nu \approx 2/3$, at a characteristic length of 12 m.

Introduction

This paper addresses the general question of the appropriate degree of the complexity of a statistical model in geology and geophysics. It is well known that adding extra parameters to a model will improve any curve fit to data, in the sense that the sum of squares of the residuals between the best-fitting curve and the data points will be reduced. It is therefore necessary to introduce an appropriate penalty for the extra degrees of freedom in a more complex model. Here we examine the specific case of breaks in slope for power-law scaling, using as an example data for fracture opening displacement u as a function of length l for the Krafla fissure swarm in northern Iceland. The questions to be addressed are: (1) Is there a significant break of slope in the data set?, and if so (2) where is it most likely to be?, and (3) what are the best fitting slopes for the two lines, and their associated uncertainties? The main constraints are that the data sets be comprised of pairs of observations (x_i, y_i), where the dependent variable y_i is assumed to be normally distributed, given x_i, and that the two best fit straight lines are continuous at the changepoint x^*.

Paper number 1999GL005372.
0094-8276/99/1999GL005372$05.00

Fault and fracture scaling

The scaling properties of faults and fractures is of general interest for a host of applications, including the mechanics of fault and fracture growth, fluid flow and contaminant transport in the subsurface, and for understanding the mechanics of earthquakes (Cowie et al., 1996). In general the scaling of maximum displacement u as a function of length l takes the power-law form, $u = k\, l^{\nu}$, where k and ν are constants. For example, for scale-invariant crack growth $\nu = 1$ (Scholz & Cowie, 1990).

The exponent itself may also be scale-dependent. For example Hatton et al. (1994) suggested a systematic change in the scaling of crack opening displacement as a function of length in the Krafla fissure zone in northern Iceland. In their study crack opening displacement and length were measured in a single field season by the same observers, over a bandwith in fracture length of 5 orders of magnitude. The fractures result from ongoing tensile stresses applied to a basalt deposit laid down following an eruptive period from 1975-1989, and grow through an earlier set of cooling joints with a characteristic size of 30 cm. Previously a change in exponent, from $\nu \approx 2$ to $\nu \approx 1$, at a characteristic fracture length of 3 m or so, had been inferred from fitting two independent straight lines through the data plotted on log-log axes. This break of slope may be attributed either to the effect of the characteristic scale length of the cooling joints on the crack tip process zone (Hatton et al., 1994), or to a critical crack size where the stress field generated by individual cracks begins to interact strongly with that of its neighbours, leading to co-operative behaviour in the form of scale-invariant crack growth associated with localisation of the deformation (Renshaw & Park, 1997).

Despite their importance, systematic changes in scaling are often difficult to demonstrate unequivocally, given the available data. This is mainly due to the inherent order-of-magnitude scatter in the data (u_i, l_i), not only due to measuring uncertainty, but also to an irreducible stochastic element in the physics of fracture nucleation and interaction (Cowie et al., 1996). It is also not sufficient to fit two straight lines without penalising the additional two parameters appropriately. Here we re-examine the data of Hatton et al. (1994), using a new method which takes this explicitly into account.

Statistical method

The general formulation of the linear statistical model is given in Draper & Smith (1998). Here we consider a special case of the form

$$y_i=\gamma(x_i) + \varepsilon_i, \qquad \text{for } i=1,...,n, \ : \tag{1}$$

$$\gamma(x_i) = a + b_0 [x_i I(x_i<x^*) + x^* I(x_i \geq x^*)] + b_1 (x_i - x^*) I(x_i \geq x^*),$$

where ε_i are the independently normally-distributed error terms, each having zero mean and unknown variance σ^2, n is the sample size, and I is the indicator function (I=1 when the inequality in the brackets holds, and I=0 otherwise). The five model parameters are the intercept a, the slopes b_0 and b_1, the changepoint x^*, and the variance σ^2. Here $x_i=ln(l_i)$ and $y_i=ln(u_i)$, so the slopes b_0 and b_1 are equivalent to the power-law exponents ν_0 and ν_1. We compare the changepoint model with a simple linear regression model, with intercept a_R and slope b_R, and no changepoint.

The likelihood function of the unknown parameters, a, b_0 and b_1 and σ^2, given $\mathbf{y} = [y_1, y_2, ..., y_n]$, when x^* is fixed, is denoted by

$$\ell(a,b_0,b_1,\sigma^2|\mathbf{y},x^*) \propto (\sigma^2)^{\frac{-n}{2}} \exp\left\{-\sum_{i=1}^{n} \frac{[y_i-\gamma(x_i)]^2}{2\sigma^2}\right\},$$

$$-\infty < a,b_0,b_1 < \infty \ ; \ 0 < \sigma^2 < \infty . \tag{2}$$

Taking the natural log of (2), and maximising with respect to the unknown parameters, gives the maximised log likelihood

$$L(\mathbf{y},x^*) = -\frac{n}{2}\ln(S_R^2) : \ S_R^2 = \sum_{i=1}^{n}[y_i - \hat{\gamma}(x_i)]^2 \quad . \tag{3}$$

S_R^2 is the residual sum of squares, and the hat symbol denotes the corresponding maximum likelihood estimate, given fixed x^*, of $\gamma(x_i)$. Confidence intervals for the two slopes, conditionally on x^*, can be obtained using standard multiple regression theory as

$$\hat{b}_0 \pm t_{n-3}(\alpha/2)u_0 \ ; \ \hat{b}_1 \pm t_{n-3}(\alpha/2)u_1 ,$$

where u_0 and u_1 are the estimated standard errors of b_0 and b_1 (Draper & Smith, 1998, p142), and $t_{n-3}(a/2)$ is the upper $a/2$ point of the t_{n-3} distribution.

A selection of the changepoint x^*, if it exists, is made by maximising a modified version of Schwarz's Information Criterion proposed by Leonard & Hsu (1999, p8). For the changepoint model (1), this information criterion, denoted BIC_{max}, maximises

$$BIC(x^*) = L(\mathbf{y},x^*) - \frac{1}{2}p\ln\left(\frac{n}{2\pi}\right) \tag{4}$$

with respect to x^*. Here p=5 is the number of unknown parameters in the model. (The modification to Schwarz's criterion is the factor 2π). For the straight line model with unknown variance p=3, and the appropriate formula is

$$BIC_R = L(\mathbf{y}) - \frac{1}{2}p\ln\left(\frac{n}{2\pi}\right) \tag{5}$$

where $L(\mathbf{y})$ is defined as in equation (3), based on the residual sum of squares for the linear model. By comparing (4) and (5), we can quantitatively infer situations where a double-slope assumption is better statistically than a single-slope assumption, while taking into account an extra penalty for the increase in the complexity of the model. Specifically, the data do not justify a changepoint when $BIC_R > BIC_{max}$.

The *BIC* criterion provides an alternative to Akaike's (1978) Information Criterion *AIC*, which uses an empirical factor 2 instead of $\ln[n/(2\pi)]$ in (4) & (5). For cases where n>46, computer simulations have shown that *BIC* is superior (e.g. Koehler & Murphree, 1988). Here n=80.

So far we have followed a maximum likelihood approach similar to that of Quandt (1958) but where our information criterion first clarifies whether or not a changepoint exists. However, in order to more fully represent the uncertainty in the data regarding x^*, we can now proceed to follow a fully Bayesian approach by constructing a prior distribution on the unknown parameters. In the prior assessment, we assume that x^*, a, b_0, b_1, and $\ln(\sigma^2)$ are independent, where x^* is uniformly distributed on some bounded interval $[x_{min}, x_{max}]$, and the other parameters are uniformly distributed on the whole real line. At this stage, any additional prior information could also be included in a straightforward way.

Under the above assumptions, the posterior density of the changepoint x^* is

$$\pi(x^*|\mathbf{y}) \propto \left|\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right|^{-1/2} \left(S_R^2\right)^{-(n-3)/2} \ : \ x_{min} \leq x^* \leq x_{max} \ , \tag{6}$$

where $\mathbf{z}_i = (1, x_i, (x_i - x^*)I[x_i \geq x^*])^T$.

To avoid practical problems with the tails, we make the pragmatic choices that x_{min} is the third smallest distinct value of x_i, and x_{max} is the third largest. Conditional on x^*, the posterior densities of

$$(b_0 - \hat{b}_0)/u_0 \text{ and } (b_1 - \hat{b}_1)/u_1$$

are respectively Student's t on n-3 degrees of freedom. Consequently, the unconditional posterior densities of b_0 and b_1 can be computed by appropriate numerical integrations with respect to the posterior density of x^*. One novelty of our techniques, with respect to the Statistics literature (e.g. Choy & Broemeling, 1980), relates to the use of the modified version of *BIC* (equation 4), rather than more complex prior assumptions.

Results

The results of applying *BIC* for equation (1) to the Kelduhverfi data set of Hatton et al. (1994: fig 3a) are shown in Figure 1. For completeness we have also added *BIC* for a linear (dashed line) and a quadratic model (solid line), both of

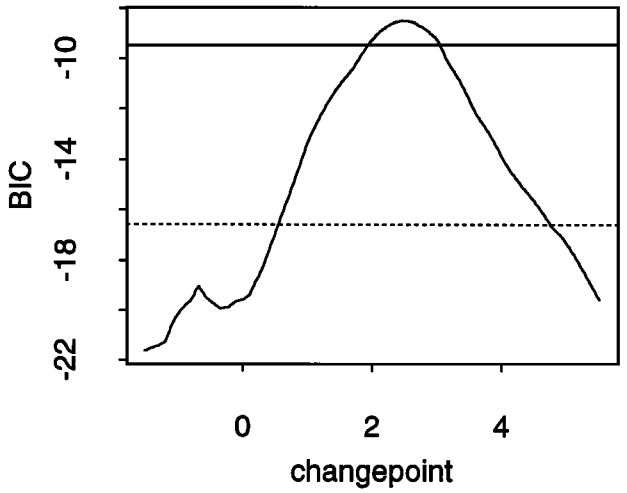

Figure 1. Plot of *BIC*, as defined in equation (4), as a function of the postulated changepoint x^* for the double-slope distribution, compared to a linear (horizontal dashed line) and a quadratic fit (horizontal solid line), for the Kelduhverfi area of the Krafla fissure swarm, (data from fig. 3a of Hatton et al., 1994).

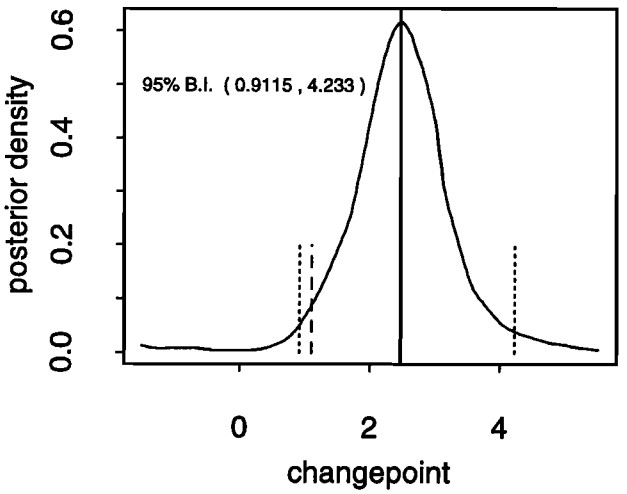

Figure 3. Posterior density function for the changepoint x^*. The posterior mean is indicated by the solid vertical line, and the previous estimate of x^* from Hatton et al., (1994) is indicated by the dashed vertical line. The 95% Bayesian interval for the changepoint is shown by the outer vertical dotted lines.

which are independent of x^*, and plot as horizontal lines. The quadratic fit is better than the linear fit, but the peak value BIC_{max} outperforms both significantly.

When we apply the value of x^* which maximises *BIC* in Figure 1, we obtain the best fit solution to equation (1) as shown by the piecewise linear curve in Figure 2. The maximum likelihood estimate of the associated break of slope occurs at a value of x^*= 2.48, corresponding to l^*=12.0 m, compared to l^*=3.0 m found by Hatton et al. (1994). The posterior density function defined by equation (6), shown on Figure 3, has a peaked distribution, with an appropriate 95% Bayesian interval for x^* in the range (0.912, 4.23) or l^* in the range (2.49, 68.9 m). This range only just includes the previous estimate of Hatton et al. (1994), also shown on the diagram for reference. The complete solution, with 95% limits (based on the unconditional posterior probability distribution with x^* unspecified) is a=−4.91±0.29, b_0=1.49±0.29, b_1=0.644±0.407. Thus the two slopes b_0, b_1 are distinguishable in statistical terms. The resolving power of the method for this data set is illustrated in Figure 4, which shows the conditional posterior densities of the slopes, given (*a*) our optimal x^*, (dashed lines), or (*b*) unconditional upon x^* (solid lines).

Our analysis is based upon the assumption that the error terms ε_i are normally distributed with constant variance. We checked this assumption by an analysis of residuals as

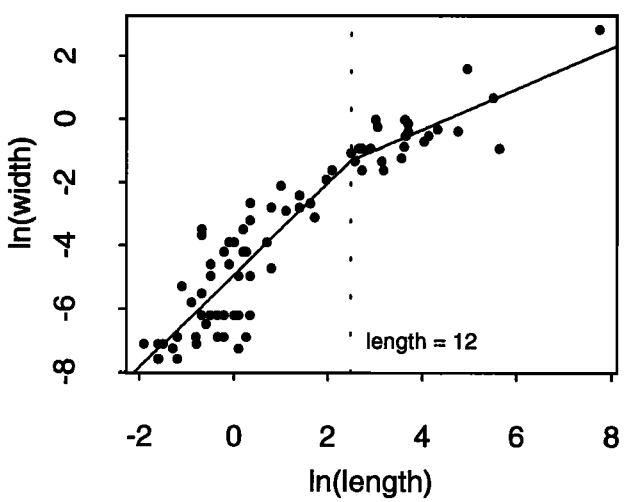

Figure 2. Fracture width (opening displacement) u versus length l for the same data as in Figure 1. The data are shown on log-log scales using natural logarithms. The best fitting lines using the Bayesian method described in the main text are shown. The location of the best fitting break of slope is indicated by the vertical dashed line.

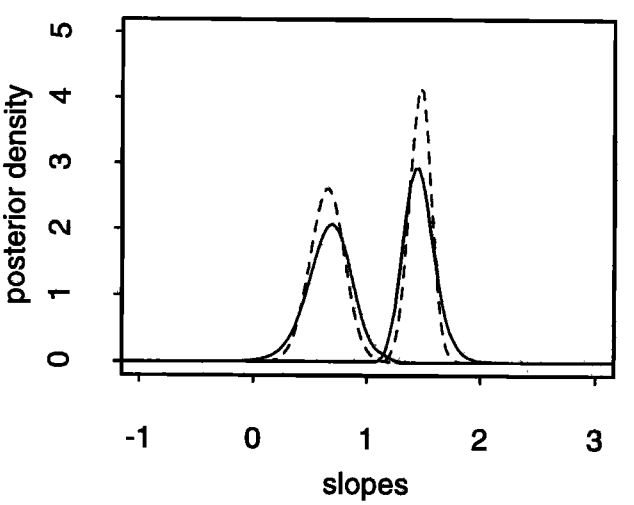

Figure 4. Posterior density function for the two best fitting slopes b_0 (right-hand peak) and b_1 (left-hand peak). Dashed lines show the conditional posterior density corresponding to the changepoint with the highest posterior probability, and the solid lines show the unconditional posterior density when integrated over our designated range of changepoints.

described in Draper & Smith (1998, pp 59-70), and found that the normal assumption is quite reasonable for this data set, although more complex models with changing variances may be required in other cases.

Finally, our method applies strictly to measured data pairs, x_i,y_i, and not to frequency data, despite the importance of this type of data for other forms of characteristic size effects involving a break of slope in rock and earthquake physics (e.g. Sornette et al., 1996). Accordingly, a separate (entropy-based) method is being developed to tackle this problem explicitly.

Conclusion

A new general method for estimating the existence and location of a break in slope in data pairs is proposed, based on a modified version of Schwarz's Information Criterion, *BIC*. When applied to the specific case of displacement-length data for the Kelduhverfi data from the Krafla fissure swarm in north Iceland, the method first confirms the existence of a significant break of slope, despite the penalty for the extra two parameters required. The location of the changepoint is between 2.49 and 68.9 m, with a median at a length of around 12 m. At this point, the slope changes from b_0=1.49±0.29 below this length to b_1=0.644±0.407 for greater lengths, where the stated ranges represent 95% Bayesian intervals. The best estimate for the characteristic length l^* is more consistent with the physical model of Renshaw & Park (1997), involving crack-crack interactions, rather than the trapped process zone model of Hatton et al (1994), although neither explanation can be ruled out within the uncertainties specified by the available data.

Acknowledgements. OP was partly supported by NERC Connect grant GR3/C0022, with matching funding from BP Exploration Ltd. We thank Didier Sornette and an anonymous reviewer for constructive and insightful comments on an earlier draft.

References

Akaike, H., A Bayesian analysis of the minimum AIC procedure, *Annals of the Institute of Statistical Mathematics, 30:A*, 9-14, 1978.

Choy, J.H.C. & L.D. Broemeling, Some Bayesian inferences for a changing linear model, *Technometrics, 22*, 71-78, 1980.

Cowie, P.A., R. Knipe & I.G. Main, Scaling laws for fault and fracture populations: Introduction to the special issue, *J. Struct. Geol., 18*, v-xi, 1996.

Draper, N.R. & H. Smith, *Applied Regression Analysis*, 3rd ed., Wiley, New York, 1998.

Hatton, C.G., I.G. Main & P.G. Meredith. Non-universal scaling of fracture length and opening displacement, *Nature, 367*, 160-162, 1994.

Koehler, A.B. & E.S. Murphree, A comparison of Akaike and Schwarz criteria for selecting model order, *Appl. Stats., 37*, 187-195, 1988.

Leonard, T., & Hsu, J.S.J., *Bayesian Methods*, Cambridge University Press, New York, 1999.

Renshaw, C.E. & J.C. Park, Effect of mechanical interactions on the scaling of fracture length and aperture, *Nature* 386, 482-484, 1997.

Scholz, C.H. & P.A. Cowie, Determination of total strain from faulting using slip measurements, *Nature, 346*, 837-838, 1990.

Sornette, D., L. Knopoff, Y.Y. Kagan & C. Vanneste, Rank-ordering statistics of extreme events: application to the distribution of large earthquakes, *J. Geophys. Res., 101*, 13883-13893, 1996.

Quandt, R.E., The estimation of the parameters of a linear regression system obeying two separate regimes, *J. Amer. Stat. Assoc., 53*, 873-880, 1958.

I.G. Main, Department of Geology & Geophysics, West Mains Road, Edinburgh EH9 3JW, UK. (e-mail: imain@glg.ed.ac.uk).

T. Leonard & O. Papasouliotis, Department of Mathematics & Statistics, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, UK.

C.G. Hatton, Stockton Campus, University of Durham, University Boulevard, Stockton-on-Tees, TS17 6BH, UK.

P.G. Meredith, Department of Geological Sciences, University College London, Gower Street, London WC1E 6BT, UK.

(Received April 15, 1999; revised: July 20, 1999;
accepted: July 27, 1999)