Automated Groove Identification in 3D Bullet Land Scans (we'll change the title)

Kiegan Rice *
Department of Statistics, Iowa State University
and
Nathaniel Garton
Department of Statistics, Iowa State University
and
Ulrike Genschel
Department of Statistics and CSAFE, Iowa State University
and
Heike Hofmann
Department of Statistics and CSAFE, Iowa State University

December 17, 2018

Abstract

Keywords: 3 to 6 keywords, that do not appear in the title

*The authors gratefully acknowledge ...

1 Background

Thanks to (1), we can do stuff. Hamby et al. (1)

2 Data Source

3 Methodology

We first need to remove the global structure of the bullet land.

3.1 Global Structure Removal

The non-traditional data structure necessitates employing non-traditional methods to model and remove the global structure. The data are made up of two competing structures: the LEA data, of which we would like to model the global structure, and the GEA data, which we would like to consider as outlying data. Traditional statistical modeling techniques minimize the least squared vertical distance from each data point to a fit line; this results in undue influence by GEA points, which pull any fit lines towards their unusual points.

While bullets are traditionally circular, it is unwise to use a rigidly quadratic model to fit the global structure. We cannot assume that fired bullets will retain a neatly circular shape, especially at the level of detail scans are captured. The significant amount of physical pressure that acts upon bullets as they are fired through a barrel also can lead to some warping or slight deformations *(find a citation from JFS or AFTE about warping/deformation of bullets?)*. Finally, the placement of the land relative to the plane of reference when a 3D scan is being captured can vary slightly, meaning that the 2D crosscuts can be slightly tilted or rotated. This will not translate into a clean quadratic-shaped(?) crosscut.

To avoid the potential risks arising from using a quadratic linear model, we instead use a locally weighted regression (LOESS) which fits linear regression models on small pieces of the data and combines predictions to result in a non-parametric predicted fit of the data structure.

However, since LOESS is still rooted in traditional regression techniques, it is unable to adequately identify and address the separation between GEA and LEA data structures. To address this, we implement a robust version of LOESS which iteratively downweights unusual data points and re-fits a LOESS model to each land. This robust LOESS is an adapted version of the robust LOESS proposed by (2).

This model is fit as follows: (add more formulaic language here...)

1. Fit a LOESS model (span = 1) to an entire crosscut to predict y using values of x. Assign weights of 1 to each data point for this fitting procedure.

2. Obtain predicted values of y from the model fit in step 1.

3. Calculate residual values using the predicted y values.

4. Calculate bisquare weights for each residual value using the following formula:

$$max(1 - (residual/(6 * mar))^2, 0)^2$$

5. Assign weights to each data point according to its residual value. If the residual value is positive, assign the bisquare downweight. If the residual is zero or negative, leave the weight at 1.

6. Repeat steps 1-5 with updated weights at each iteration for k iterations, with 20 iterations as the default.

7. After k iterations of updating the weight vector, fit a LOESS model (span = 1) and obtained predicted and residual values.

The subsequent prediction methods for shoulder location are based on the residuals calculated from the fit to the global structure of each land. One method uses penalized two-class classification techniques to classify each data point into "LEA" or "GEA", while the second uses Bayesian changepoint analysis to predict the data points at which the shoulders begin on either side.

3.2 Two-Class Classification

3.3 Bayesian Changepoint Analysis

The idea behind the changepoint approach is that within either the left GEA, right GEA, or the LEA, the global structure is consistent and can either be described by a line with zero slope, a line with positive slope for the right GEA, or a line with negative slope for the left GEA. Finding the points where the GEAs and LEA meet is treated as a problem of model selection. That is, the best fitting statistical model, in terms of the magnitude of the likelihood, should be the one which assumes that the points at which the global structure changes align with where the GEAs and LEA meet. These points of global structural change are what we will call changepoints. Thus, our model will be defined in a piecewise fashion. In practice there are also complex additional patterns which may exist for a number of reasons, but this large scale structural assumption remains generally reasonable. The complex smaller scale patterns can be thought of as the dependence in the data after accounting for the global structure. Because of the nature of the model which we consider, it becomes necessary for computational reasons to perform a couple of additional data preprocessing steps. Specifically, we will scale the residuals from the robust LOESS procedure, and we will impute missing values. In the next section, we describe the model that we will use to identify changepoints, after which we will describe the estimation procedure which we use. Details of the additional data preprocessing steps can be found in the appendix.

3.4 Bayesian Model Formulation

Before introducing the model, we introduce some notation. First, let $\{Y(x_i) : i = 1, 2, ..., n\}$ denote the set of random variables representing the residuals from the robust LOESS procedure at the values x_i. For simplicity, also assume that $x_1 < x_2 < ... < x_n$. Also, let c_l be the value of the left changepoint and c_r be the value of the right changepoint. Here, the left changepoint is where the left GEA meets the LEA, and the right changepoint is where the right GEA meets the LEA. Also, denote the median centered x values as $x_i' = x_i - \tilde{x}$ where $\tilde{x}$ is the median x value. As mentioned in the previous paragraph, the complex

small scale patterns, such as the striae, will be modeled through a covariance structure on the data that will be allowed to differ between each GEA and between the GEAs and LEA. We will construct the covariance matrices from the exponential covariance function $K(x, x'; \sigma, \ell) = \sigma^2 e^{-\frac{|x-x'|}{\ell}} = cov(Y(x), Y(x'))$. The differences in covariance matrices for the GEAs and LEA will be reflected in the parameters σ and ℓ. The data model that we consider is then,

$$(Y(x_1), Y(x_2), ..., Y(x_{k_1})) \sim N(\beta_{01}\mathbb{1} + \beta_{11}x'_{1:k_1}, \Sigma_1(\sigma_1, \ell_1)) \quad (1)$$

$$(Y(x_{k_1+1}), Y(x_{k_1+2}), ..., Y(x_{k_2})) \sim N(0, \Sigma_2(\sigma_2, \ell_2)) \quad (2)$$

$$(Y(x_{k_2+1}), Y(x_{k_2+2}), ..., Y(x_n)) \sim N(\beta_{02}\mathbb{1} + \beta_{12}x'_{k_2+1:n}, \Sigma_3(\sigma_3, \ell_3)), \quad (3)$$

where $x_{k_1} < c_l \leq x_{k_1+1}$ and $x_{k_2} < c_r \leq x_{k_2+1}$ Here, $x_{1:k}$ denotes the column vector $(x_1, x_2, ..., x_k)^\top$, and $\mathbb{1}$ denotes the vector of ones. Indpendence is assumed between each of these three distributions for simplicity.

Thus the parameters that need to be estimated include the four mean parameters in the GEAs, the six covariance parameters (two for each of the three areas), and the two changepoint parameters, c_l and c_r.

The above model encapsulates the essence of the approach. However, there are a few difficulties. The first difficulty is that there are not always two GEAs in a particular land. There may be one GEA, or the land may only consist of the LEA. Thus, the above model is actually conditional on there being two GEAs in the data. We also define models for when there is one GEA on the left, one GEA on the right, or no GEAs. The models are defined in an essentially identical way. Conditional on there being only one GEA, the left GEA model is defined as,

$$(Y(x_1), Y(x_2), ..., Y(x_k)) \sim N(\beta_0\mathbb{1} + \beta_1 x'_{1:k}, \Sigma_1(\sigma_1, \ell_1)) \quad (4)$$

$$(Y(x_{k+1}), Y(x_{k+2}), ..., Y(x_n)) \sim N(0, \Sigma_2(\sigma_2, \ell_2)), \quad (5)$$

and the right GEA model is defined as,

$$(Y(x_1), Y(x_2), ..., Y(x_k)) \sim N(0, \Sigma_1(\sigma_1, \ell_1)) \tag{6}$$

$$(Y(x_{k+1}), Y(x_{k+2}), ..., Y(x_n)) \sim N(\beta_0 \not\Vdash + \beta_1 x'_{k+1:n} \Sigma_2(\sigma_2, \ell_2)). \tag{7}$$

Finally, conditional on there being no GEAs in the data, the model is simply

$$(Y(x_1), Y(x_2), ..., Y(x_n)) \sim N(0, \Sigma(\sigma, \ell)). \tag{8}$$

Thus, estimating the changepoint locations also involves selecting the most appropriate model. In order to avoid confusion, we have slightly abused notation and, for example, $\Sigma_1(\sigma_1, \ell_1)$ as it is estimated in the two changepoint model is *not* the same as $\Sigma_1(\sigma_1, \ell_1)$ from either of the one changepoint models, and $\Sigma_1(\sigma_1, \ell_1)$ is also *not* the same between the two one changepoint models. As another example, β_0 is *not* the same between each of the one changepoint models. So, to be clear, duplication of notation in *different* models is not meant to imply that those parameters are shared between models.

Ultimately, these above four models are each individually fitted, and each model above is given a prior. From there, we do model selection in the formal Bayesian way, selecting the most probable model. Simultaneously with selecting the most probable model, we also use the maximum a posteriori estimator for the changepoint locations.

In order to complete a Bayesian model specification, we need priors on each of the parameters in each model as well as each model. We will assume independence between each parameter a priori. For each length scale ℓ, we will assume $\ell \sim \text{Gamma}(3, 5)$. For each standard deviation, we will assume $\sigma \sim \text{Half-Normal}^+(0, 1)$, where $\text{Half-Normal}^+(\cdot, \cdot)$ is notation for the normal distribution restricted to the positive real numbers. For intercept parameters, $\beta_{01}, \beta_{02}, \beta_0 \sim N(0, 10)$. For the slope parameters, the preceding trend deviates slightly. For any slope that corresponds to the *left* GEA, β_1 or β_{01}, we will assume that the slope can not be positive. That is, $\beta_1, \beta_{01} \sim \text{Half-Normal}^-(0, 10)$, where $\text{Half-Normal}^-(\cdot, \cdot)$ is notation for the normal distribution restricted to the negative real numbers. Contrastingly, for any slope that corresponds to the *right* GEA, β_1 or β_{02}, we will assume that the slope can not be negative. That is, $\beta_1, \beta_{01} \sim \text{Half-Normal}^+(0, 10)$. For the changepoint locations, we assume a uniform prior $\pi(c_l, c_r) \propto I(a < c_l < c_r - \gamma < b - \gamma)$. Here, a and b

are some values close to the edges of the data. How close those values are to the edges is a parameter that is set manually. Further, we include another hyperparameter, γ, which can be set so that the changepoints are not allowed to be too close to each other. This is also a parameter that is set manually. Lastly, we assume a uniform prior over all four models.

4 Results

5 Conclusions

6 References

References

1. Hamby JE, Brundage DJ, Thorpe JW. The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries. AFTE Journal 2009;41(2):99–110.

2. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association 1979;74(368):829–836.