An optimal penalty for breakpoint detection in signals with variable sampling density

Toby Dylan Hocking

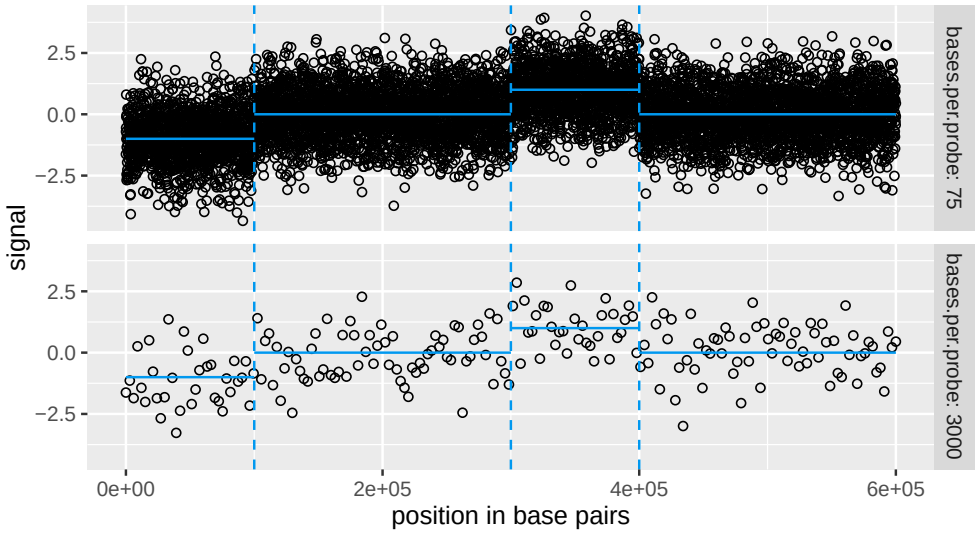
October 25, 2016

In this vignette, we use the breakpointError to derive an optimal penalty for breakpoint detection in signals of varying sampling density. First, we will present an empirical analysis of several simulated signals using the breakpointError. Then, we will discuss the relationship of our results to relevant theoretical results. This analysis was originally presented by Hocking [2012, Chapter 4].

In recent years, several authors have developed a theory of minimal penalties that can be used to accurately recover a signal from noisy observations [Arlot and Massart, 2009, Lebarbier, 2005]. These methods can be used offline to analyze some assumptions about the signal and the noise of the data. Typically, these results guarantee recovery of the correct signal with high probability. However, in this vignette we are more interested in accurate recovery of the breakpoints than the signal itself. So here we use the breakpointError to directly attack the problem of breakpoint detection rather than signal recovery.

In real array CGH data, the sampling density of probes along the genome is not uniform across samples. In fact, we see a sampling density between 40 and 4400 kilobases per probe in the neuroblastoma data set `data(neuroblastoma,package="neuroblastoma")`.

So to construct a penalty that can best adapt to this variation, we analyze the following simulation. We create a latent piecewise constant signal $\mu \in \mathbb{R}^D$ over $D = 600000$ base pairs, shown as the blue line in the figure below. We define a signal sample size $d_i \in \{200, \ldots, 8000\}$ for every noisy signal $i \in \{1, \ldots, n = 4\}$. Let $y_i \in \mathbb{R}^{d_i}$ be noisy signal i, sampled at positions $p_i \in \mathcal{X}^{d_i}$, with $p_{i1} < \cdots < p_{i,d_i}$. We sample every probe j from the $y_{ij} \sim N(\mu_{p_{ij}}, 1)$ distribution. These samples are shown as the black points in the figure below.



We would like to learn some model complexity parameter λ on the first noisy signal, and use it for accurate breakpoint detection on the second noisy signal. In other words, we are looking for a model selection criterion which is invariant to sampling density.

1 Empirical analysis of simulations

To determine an optimal penalty for breakpoint detection in simulated data, we proceed as follows. For every signal i, we use pruned dynamic programming to calculate the maximum likelihood estimator $\hat{y}_i^k \in \mathbb{R}^{d_i}$, for several model sizes $k \in \{1, \ldots, k_{\max} = 8\}$ [Rigaill, 2010]. Then, we define the model selection criteria

$$k_i^\alpha(\lambda) = \arg\min_k \lambda k d_i^\alpha + ||y_i - \hat{y}_i^k||_2^2. \tag{1}$$

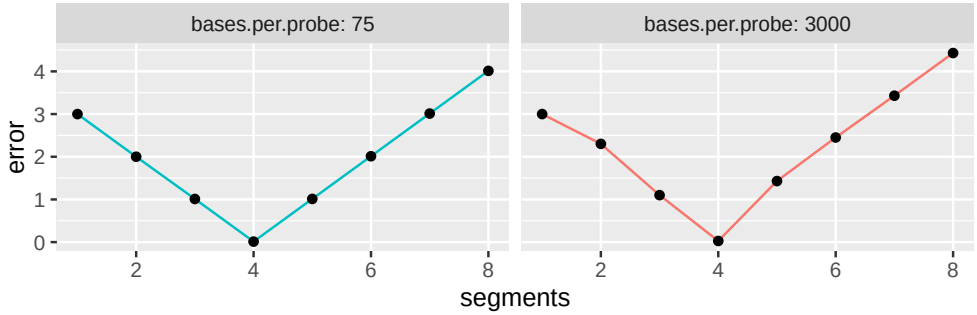
Each of these is a function $k_i^\alpha : \mathbb{R}^+ \to \{1, \ldots, k_{\max}\}$ that takes a model complexity tradeoff parameter λ and returns the optimal number of segments for signal i. The goal is to find a penalty exponent $\alpha \in \mathbb{R}$ that lets us generalize λ between different signals i.

Naïvely, one may expect that the best exponent is $\alpha = 1$, since that corresponds to an error term with the average residual:

$$k_i^1(\lambda) = \arg\min_k \lambda k + ||y_i - \hat{y}_i^k||_2^2/d_i. \tag{2}$$

However, we will show that this penalty is not optimal, by analyzing the breakpointError.

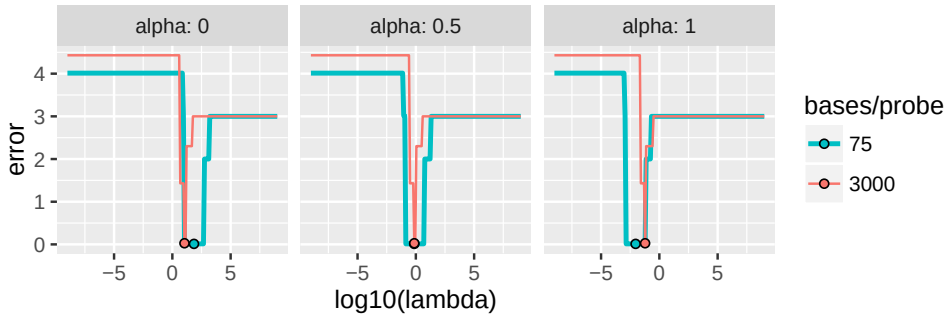
To quantify the accuracy of a segmentation for signal i, let $e_i(k)$ be the breakpointError of the model with k segments. In the figure below, we plot e_i for the 2 simulated signals i shown previously.



Now, let us define the penalized model breakpoint error $E_i^\alpha : \mathbb{R}^+ \to \mathbb{R}^+$ as

$$E_i^\alpha(\lambda) = e_i\left[k_i^\alpha(\lambda)\right]. \tag{3}$$

In the figure below, we plot these functions for the two signals i shown previously, and for several penalty exponents α.



The dots in the figure show the optimal λ found by minimizing the penalized model breakpoint detection error:

$$\hat{\lambda}_i^\alpha = \arg\min_{\lambda \in \mathbb{R}^+} E_i^\alpha(\lambda) \tag{4}$$

This figure suggests that $\alpha \approx 1/2$ defines a penalty with aligned error curves, which will result in $\hat{\lambda}_i^\alpha$ values that can be generalized between profiles.

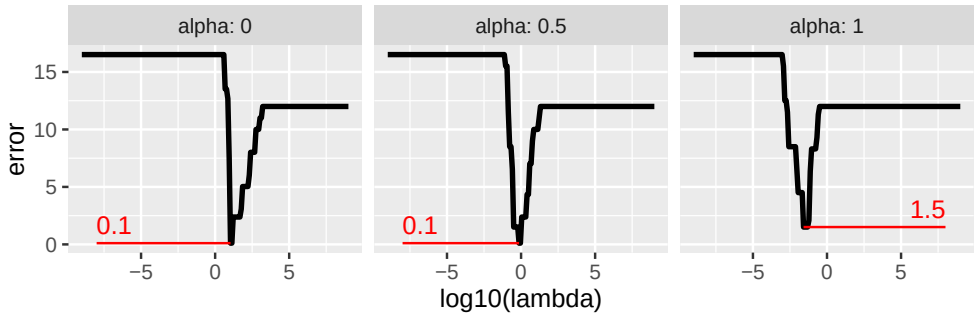
Now, we are ready to define 2 quantities that will be able to help us choose an optimal penalty exponent α. First, let us consider the training error over the entire database:

$$E^{\alpha}(\lambda) = \sum_{i=1}^{n} E_i^{\alpha}(\lambda), \tag{5}$$

and we define the minimal value of this function as

$$E^{*}(\alpha) = \min_{\lambda} E^{\alpha}(\lambda). \tag{6}$$

In the figure below, we plot these training error functions E^{α} (black) and their minimal values E^{*} (red) for several values of α.

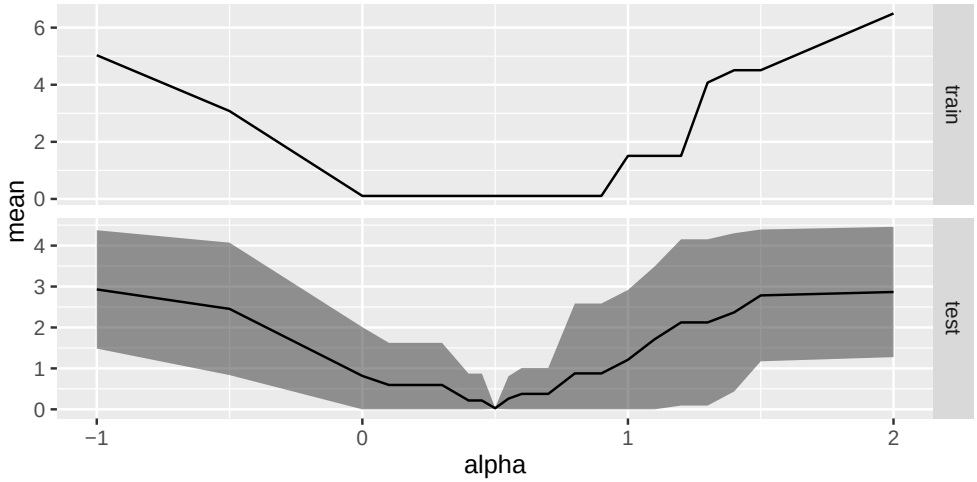


It is clear that the minimum training error is found for some penalty exponent α near $1/2$, and we would like to find the precise α that results in the lowest possible minimum $E^{*}(\alpha)$.

We also consider the test error over all pairs of signals when training on one and testing on another:

$$\text{TestErr}(\alpha) = \sum_{i \neq j} E_i^{\alpha}(\hat{\lambda}_j^{\alpha}). \tag{7}$$

In the figure below, we plot E^{*} and TestErr for a grid of α values.



It is clear that an optimal penalty is given by $\alpha = 1/2$. This corresponds to the following model selection criterion which is invariant to sampling density:

$$k_i^{1/2}(\lambda) = \arg\min_{k} \lambda k \sqrt{d_i} + ||y_i - \hat{y}_i^k||_2^2 \tag{8}$$

2 Discussion of related theoretical results

As explained by Arlot and Celisse [2010], a model selection procedure can be either efficient or consistent. An efficient procedure for model estimation accurately recovers the latent signal, whereas a consistent procedure for model identification accurately recovers the breakpoints. Since we consider the breakpoint detection error, we are attempting to construct a consistent penalty, not an efficient penalty.

In general terms, the fact that we find a nonzero exponent α for our d_i^α penalty term agrees with other results. In particular, Arlot [2008] proposed an optimal procedure to select model complexity parameters in cross-validation by normalizing by the sample size d_i.

The $\sqrt{d_i}$ term that we find here using simulations is in agreement with Fischer [2011], who use finite sample model selection theory to find a $\sqrt{d_i}$ term in a penalty optimal for clustering.

When theoretically deriving an efficient penalty for change-point model estimation in the non-asymptotic setting, Lebarbier [2005] obtained a $\log d_i$ term. This contrasts our result, which examines the identification problem using the breakpoint error and obtains a $\sqrt{d_i}$ term. But in fact this is in agreement with classical results that AIC underpenalizes with respect to the BIC, as shown in the table below.

Estimation Model	Penalty Term	Identification Model	Penalty Term
AIC	2	BIC	$\log d_i$
Lebarbier	$\log d_i$	This work	$\sqrt{d_i}$

Comparing our results with Lebarbier, in the context of classical results involving AIC and BIC. The BIC is designed for model identification and penalizes more than the AIC. Likewise, our penalty examines model identification using the breakpoint detection error, and penalizes more than the efficient penalty proposed by Lebarbier.

References

S. Arlot. V-fold cross-validation improved: V-fold penalization. *Arxiv preprint arXiv:0802.0566*, 2008.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4: 40–79, 2010.

S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279, June 2009. ISSN 1532-4435. `http://dl.acm.org/citation.cfm?id=1577069.1577079`.

A. Fischer. On the number of groups in clustering. *Statistics and Probability Letters*, 81:1771–1781, 2011.

T. D. Hocking. *Learning algorithms and statistical software, with applications to bioinformatics*. PhD thesis, Ecole Normale Superieure de Cachan, France, 2012.

E. Lebarbier. Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85:717–736, 2005.

G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. arXiv:1004.0887, 2010.