International Society for Bayesian Analysis, 9^{th} World Meeting,
Hamilton Island, Australia, 2008.

An Improved Bayesian Method for DNA Copy Number Estimation

P.M.V. Rancoita[1,2,3⋆], M. Hutter[4], F. Bertoni[2] and I. Kwee[1,2]

[1] IDSIA, Manno-Lugano, Switzerland
[2] Laboratory of Experimental Oncology, Oncology Institute of Southern Switzerland (IOSI),Bellinzona, Switzerland
[3] Dipartimento di Matematica, Università degli Studi di Milano, Milano, Italy
[4] RSISE/ANU/NICTA, Canberra, Australia
⋆ paola@idsia.ch

Tumors, some congenital and some hereditary disorders are related to chromosomal aberrations. One type of aberration is the change of the DNA copy number in one or more regions of the genome. Apart from the sex chromosomes, in a healthy cell the copy number is two because we inherit two copies of each chromosome from each our parents, but in a tumor the genome can present regions of deletions or gains. The aberrated copy number along the genome can be mathematically represented as a piecewise constant function. With current microarray technologies we are able to measure the copy number of DNA at several millions of positions simultaneously, however the data is very noisy.

Bayesian Piecewise Constant Regression (BPCR) is a regression method for data that are noisy observations of a piecewise constant function and has been suggested for DNA copy number estimation. Estimation of the number of segments, the boundaries, and the levels of the segments is performed in a Bayesian way. However, some estimators in the original formulation gave problems in certain situations. In particular, the boundary estimator did not take into account the dependency among the boundaries and could estimate multiple breakpoints at the same position. Here, we propose alternative estimators for these parameters that lead to a significant better performance of the original algorithm

The original BPCR estimated the number of segments and each boundary with the MAP estimator. The MAP estimator minimizes the posterior expected 0-1 error; changing the error type we can derive alternative estimators. For the number of segments, we propose alternative estimators based on the absolute error ($\widehat{K}_1$) and the squared error ($\widehat{K}_2$). For the boundaries, we propose three alternative estimators;

one of them is based on the *total 0-1 error* ($\widehat{\underline{T}}_{joint}$), which is the 0-1 error defined for the whole vector. Since the length of the vector of the estimated boundaries depends on the estimated segment number, the usage of this error leads to a problem of interpretation when the two vectors belong to different vector spaces. Hence, we changed the vector space of the boundaries by mapping them into R^{n+1} with a binary transformation. The components of the new vectors are equal to one, only at the positions corresponding to the boundaries. On these vectors we defined another type of error, called *binary error*, based on the Russel-Rao dissimilarity measure of binary vectors. The two additional estimators minimize this error with respect to different posterior probabilities ($\widehat{\underline{T}}_{BinErr}$ and $\widehat{\underline{T}}_{BinErrAk}$). On the basis of theoretical and empirical results obtained on artificial data, we found that it is best to use a combination of $\widehat{K}_2$ and $\widehat{\underline{T}}_{BinErrAk}$.

Finally, we compared our improved version of BPCR with other existing methods to estimate genomic copy number data (CBS, HMM, CGHseg and GLAD). The comparisons on artificial data showed that our improved version of BPCR generally performed best. The CBS method recovered the profiles quite well, but it was generally unable to detect segments of small width. HMM tended to divide the profiles into more and smaller segments; the number of this segments increased when the noise is higher. CGHseg sometimes had problems with the estimation of the segment levels, because it uses the arithmetic mean as estimator, which is not suitable when segments contain only few points. GLAD often did not detect obvious segments even at mild noise levels.

When the variance of the noise was much higher than the variance of the segment levels, it was difficult to choose the best performing method. In this situation the behavior of the improved BPCR (mBPCR) depended on the choice of the estimator for the variance of the noise: $\hat{\rho}$ or $\hat{\rho}_1$. In general, mBPCR with $\hat{\rho}$ (and most of the other methods) missed the smaller segments, while mBPCR with $\hat{\rho}_1$ overestimated the number of the segments leading to a copy number profile with more and smaller segments.

The comparisons on real data showed that mBPCR and CBS were among the best performing methods. However CBS still exhibited its problem in the detection of small segments. Moreover, CGHseg seemed sensitive to outliers. On these real data, the choice of the estimator for the variance of the noise minimally affected the resulting estimation made by mBPCR: the results were generally more or less the same, but with $\hat{\rho}_1$ sometimes estimating a higher number of segments.

In summary, we have proposed new parameter estimators for the BPCR algorithm to estimate piecewise constant data. In comparisons, our new method outperformed the original algorithm and existing methods.