ABACUS: Unsupervised Multivariate Change Detection via Bayesian Source Separation

Wenyu Zhang, Daniel Gilbert, David Matteson

Department of Statistical Science, Cornell University

Cornell University

Objectives

For multivariate data with multiple change points, ABACUS (*Automatic BAyesian Changepoints Under Sparsity*) integrates sparse Bayesian blind source separation with a change detection framework to:

- Recover lower-dimensional latent signals;
- Utilize multi-level sparsity to achieve both dimensionality reduction and modeling of signal changes;
- Detect additive outliers and level shifts separately.

Introduction

Offline multiple change detection in multivariate data is studied, specifically where the data exhibit mean changes that can occur simultaneously in several channels. The direction and magnitude of change can be different across channels. The multivariate data are assumed to be generated by low-dimensional latent source signals through linear mixing according to the model $Y = MS + E$, shown in Figure 1.

Observed mean changes manifest from the latent space, and changes are detected by estimating these latent source signals, which possess 'semantic' meaning of the underlying states and are free of noise.

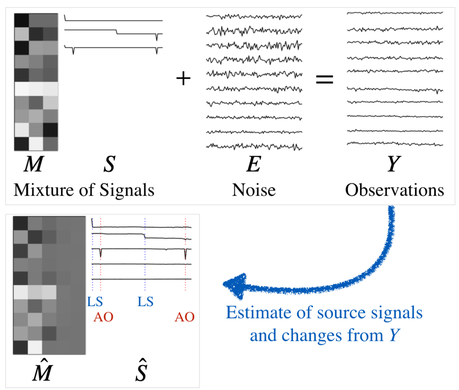
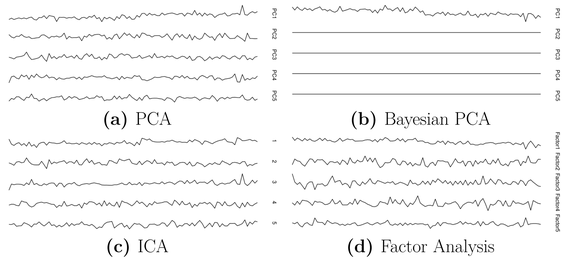


Figure 1: Given observations generated by the linear mixing of signals contaminated by noise, ABACUS estimates the source signals and detect additive outliers (AO, red) and level shifts (LS, blue). In M, darker and lighter cells represent negative and positive values respectively, and medium gray cells represent zero.

Related Works

Other matrix decomposition methods do not recover piecewise-constant latent signals, and do not correctly recover the effective dimensionality of the latent space.



(a) PCA **(b)** Bayesian PCA **(c)** ICA **(d)** Factor Analysis

Problem Formulation

Observations $Y \in \mathbb{R}^{P\times N}$ is a P-dimensional data stream of length N. Let K be a user-specified upper bound for $\text{rank}(S) = r$ such that $r \le K < P$. Then the decomposition is:

$$Y_{\cdot n} = MS_{\cdot n} + E_{\cdot n}$$
$$S_{\cdot n} = S^{(0)}_{\cdot n} + S^{(1)}_{\cdot n}$$
$$S^{(0)}_{\cdot n} = V^{(0)}_{\cdot n} \text{ and } \triangle S^{(1)}_{\cdot n} = V^{(1)}_{\cdot n}$$

where

- $M \in \mathbb{R}^{P\times K}$ is the mixing matrix
- $S \in \mathbb{R}^{K\times N}$ is the source signal matrix
- $S^{(0)}, S^{(1)} \in \mathbb{R}^{K\times N}$ are component matrices of S
- $V^{(0)}, V^{(1)} \in \mathbb{R}^{K\times N}$ are sparse change matrices
- $E_{\cdot n} \sim \text{N}(0,\ \Psi)$ and $\Psi = \text{diag}\,(\psi)$

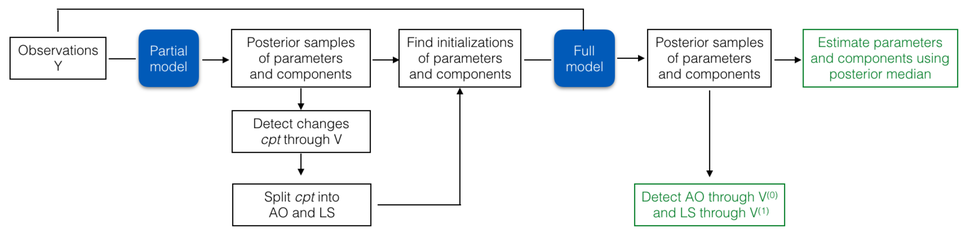


Figure 3: Implementation procedure. From observations Y, a partial model is first fit and its estimations initialize the full Bayesian model. Final estimates of source signals and change points are obtained from the median of MCMC samples.

1. Bayesian Latent Variable Model

Sparse group priors are placed on the columns of M and rows of $V^{(d)}$ through $\lambda^{(d)}_h$ for dimensionality reduction of the latent space. Sparse group priors are also placed on the columns of $V^{(d)}$ through $\phi^{(d)}_n$ to select a subset of indices as change locations. Elementwise sparsity is placed on $V^{(d)}$ through $\gamma^{(d)}_{hn}$ to allow sparse changes for each latent variable.

For $1 \le i \le P$ and $1 \le h \le K$ and $1 \le n \le N$ and $d \in \{0, 1\}$, priors are set as:

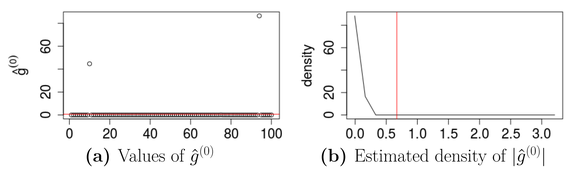
$$M_{\cdot h} | \lambda^{(0)}_h, \lambda^{(1)}_h,\ \tau^{(0)}, \tau^{(1)}, \Psi \sim \text{N}\left(0, \lambda^{(0)}_h \lambda^{(1)}_h \tau^{(0)} \tau^{(1)} \Psi\right)$$
$$V^{(d)}_{hn} | \phi^{(d)}_n, \lambda^{(d)}_h, \gamma^{(d)}_{hn},\ \tau^{(d)} \sim \text{N}\left(0, \phi^{(d)}_n \lambda^{(d)}_h \gamma^{(d)}_{hn} \tau^{(d)}\right)$$

Priors of ψ_i are Inverse Gamma and priors of shrinkage parameters are half-cauchy.

2. Change Detection

Let $f^{(d)}_n$ be the element with the largest magnitude in change matrix $V^{(d)}_{\cdot n}$. At any index n, $f^{(d)}_n$ is nonzero if and only if there is a change of type d in at least one latent variable. Finding all such indices is equivalent to finding the change locations. For robustness with empirical samples, use

$$\hat{g}^{(d)}_n = \text{median}\left(\hat{f}^{(d)}_n\right)$$



(a) Values of $\hat{g}^{(0)}$ **(b)** Estimated density of $|\hat{g}^{(0)}|$

Application: array-based comparative genomic hybridization

aCGH is a technique for studying copy number alterations in event of diseases. The dataset contains 43 samples of different individuals with bladder tumor. Each sample has 2215 probes measuring the log2 ratio between the number of transcribed DNA copies from tumorous cells and from a healthy reference.

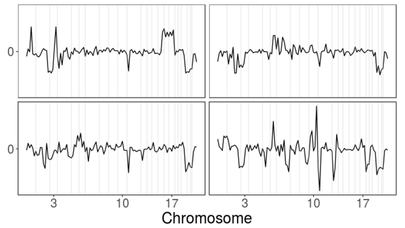


Figure 5: log2 ratio between the number of transcribed DNA copies from tumorous cells and from a healthy reference. Negative indicates deletion, positive indicates amplification.

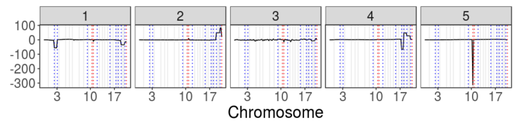


Figure 6: Latent source signals (1-5) recovered with $K = 5$

S	Chromosome arm with changes	Tumor stage
1	2q, 3q, 20p/q	pT_1
2	17p/q, 18p/q, 19p/q, 20p/q	pT_1
4	10q	pT_a, pT_1, pT_{2-4}
5	11p, 20p/q	pT_{2-4}

Table 1: Genetic aberrations corresponding to changes on latent signals

ABACUS performs consistently across different K in terms of the change points and latent source signals recovered.

References

[1] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In *AISTATS*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80. PMLR, 16–18 Apr 2009.

[2] Chuan Gao, Christopher Brown, and Barbara Engelhardt. A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects. 10 2013.