ABACUS: Unsupervised Multivariate Change Detection via Bayesian Source Separation

Wenyu Zhang* wz258@cornell.edu
Daniel Gilbert* deg257@cornell.edu
David S. Matteson* matteson@cornell.edu

Abstract

Change detection involves segmenting sequential data such that observations in the same segment share some desired properties. Multivariate change detection continues to be a challenging problem due to the variety of ways change points can be correlated across channels and the potentially poor signal-to-noise ratio on individual channels. In this paper, we are interested in locating additive outliers (AO) and level shifts (LS) in the unsupervised setting. We propose ABACUS, *Automatic BAyesian Changepoints Under Sparsity*, a Bayesian source separation technique to recover latent signals while also detecting changes in model parameters. Multi-level sparsity achieves both dimension reduction and modeling of signal changes. We show ABACUS has competitive or superior performance in simulation studies against state-of-the-art change detection methods and established latent variable models. We also illustrate ABACUS on two real application, modeling genomic profiles and analyzing household electricity consumption.

Keywords: blind source separation; dimension reduction; latent factor model; multivariate change points; sparse signal extraction; unsupervised learning

1 Introduction

Change detection segments sequential data such that observations in each segment share the same characteristics. We can view it as a specific form of clustering where sequential data points tend to cluster together. Two common sequential orderings are time and physical location. Offline change detection segments the data retrospectively and is useful for uncovering events and systematic behaviors in data analysis tasks. It is applied in a variety of fields including energy consumption [13], genomics [22] and finance [10]. Furthermore, in the potential presence of change points, utilizing change detection prior to data modeling can help prevent building inappropriate models under the assumption of data homogeneity, and consequently supports improved prediction and statistical inference.

In this paper, we study offline multiple change detection in multivariate data, 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. Here, we refer to mean changes lasting a single time unit with an immediate return as additive outliers (AO), and mean changes with duration two or greater as level shifts (LS). We assume that the multivariate data are generated by low-dimensional latent source signals through linear mixing according to the model $Y = MS + E$, shown in Figure 1, similar to the general linear setting used in the blind source separation literature [15,21]. Notationwise, M is the mixing matrix, and Y, S and E are the observations, source signals and noise, respectively. Observed mean changes manifest from the latent space, and we detect changes by estimating these latent source signals, which possess 'semantic' meaning of the underlying states and are free of noise.

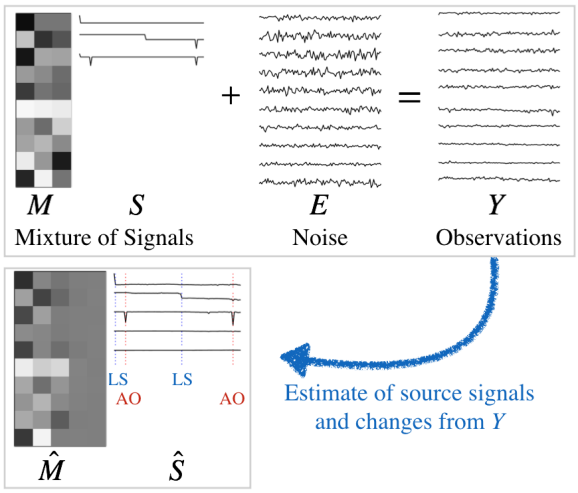


Figure 1: Given observations generated by the linear mixing of signals contaminated by noise, ABACUS estimates the source signals and detect additive outliers (AO, red) and level shifts (LS, blue). In M, darker and lighter cells represent negative and positive values respectively, and medium gray cells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

Multivariate data are readily observed in many applications in today's world, and mean changes are of particular interest since the mean is often a salient aspect of the system state. Multivariate data can be observations from multiple channels monitoring a single system, or a collection of univariate data streams from

*Cornell University

multiple related systems. Examples of the first scenario include household power consumption measured with sub-meters [13], and wine quality based on physicochemical test variables [1]. Examples of the second scenario include array comparative genomic hybridization measurements from several patients with the same medical condition [20]. In these and other examples, change points in multivariate data sometimes occur simultaneously in multiple channels because the signals may be driven by the same underlying processes. It is of interest to identify these shared change points to further analyze the relationship between channels. Running univariate change detection on each channel does not encourage identification of such shared changes.

Finding changes in multidimensional data is known to be a difficult problem. If the magnitude of change as measured by symmetric Kullback-Leibler divergence is kept constant, detectability of the change worsens when the data dimension P increases. This can hinder detection even at dimensions as low as $P = 10$ [1]. Another issue arises when the data dimensions P exceeds the sample size N. If one wishes to use hypothesis testing to test for homogeneity, naive calculations of familiar test statistics such as the Hotelling's t-squared statistic are prohibitive. Several approaches tackle multivariate data by incorporating a dimensionality reduction step [25, 26], but these either project the data onto a single dimension or require the user to select the reduced dimensionality.

Our main contribution is to successfully integrate sparse Bayesian blind source separation with a change detection framework. No previous work on latent variable modeling explicitly considered source signals with unconstrained mean changes. Bayesian variations of principle component analysis (PCA) are capable of automatic dimensionality selection [4, 28], and shrinkage priors also achieve desirable properties in trend filtering [19]. In our Bayesian latent model, we use horseshoe priors to recover the lower-dimensional source signals and to simultaneously model the change points. The two tasks complement each other since the source signals exhibit changes. We propose ABACUS, *Automatic BAyesian Changepoint Under Sparsity*, an automatic procedure that simultaneously detects additive outliers and level shifts via estimating components from the source separation problem. Figure 1 gives an example where ABACUS recovers the true latent change space of size three by estimating values in the appropriate dimensions of M and S to zero, and ABACUS also locates relevant change points. We show through simulations and real data applications that ABACUS achieves better performance in both change detection and source recovery.

2 Related Works

Authors of [6] formulated multivariate change detection as a group fused Lasso, and showed empirically that detection probability approaches one with increasing P when noise is small. Variants of binary segmentation produce approximately optimal segmentations by iteratively detecting single change points [20, 22]. Dynamic programming with a suitable multivariate goodness-of-fit metric can recursively the data [27]. The above methods directly segment the observations and some assume independence across channels [11, 25]. We recover the latent change space with prior belief that only the latent signals are independent given model parameters.

Some works use a two-step procedure with compression onto a low dimension $K \ll P$ followed by change detection. Projection onto one dimension enables univariate change detection [11]. For $K > 1$, [23] applies univariate change detection on each latent signal after Independent Component Analysis (ICA). Fixed or time-varying random projection is paired with hypothesis testing [26]. Using compressive measurements, where the projection matrix is a random projection or drawn from a Gaussian ensemble, [2] derives the number of observations required for a target detection delay. For the above methods, the user has to specify the compression ratio through K. Our proposed method ABACUS is more robust to the specification of K due to automatic dimensionality selection by our sparsity assumptions. In contrast to the latent variable model that we employ, these methods also ignore estimating the mixing matrix.

Bayesian approaches in change detection typically rely on using indicator variables to denote the presence of change points. The BCP method [3, 10] assumes that observations in each segment are independent and identically distributed as Gaussian, and updates posterior segment means conditional on the segmentation at each iteration of an MCMC scheme. A uniform prior $\mathrm{U}(0, q)$ is put on the change point probabilities, and the user tunes the chances of discovering shorter or longer segments through q. In [13], given the segmentation informed by the indicator variables, a Wilcoxon rank sum test is performed at each index of the data and the resulting p-values are modeled as a Beta-Uniform mixture. The data likelihood is written as a composite marginal likelihood of the p-values. The formulation makes no assumption on the distributional form of the data.

ABACUS similarly utilizes the sparsity of changes by applying horseshoe priors, modeling the presence and absence of changes, but also the change directions and magnitudes. We utilize the horseshoe prior as it is known for robustness and superior shrinkage properties [7]. Empirically, differences in neighboring non-change location means are effectively shrunk to zero.

3 Problem Formulation

We observe $Y \in \mathbb{R}^{P\times N}$, a P-dimensional data stream of length N. Each column take the form $Y_{\cdot n} = MS_{\cdot n}+E_{\cdot n}$, where $M \in \mathbb{R}^{P\times r}$ is the mixing matrix, $S_{\cdot n}$ is the r-dimensional source signal, and $E_{\cdot n}$ is the P-dimensional noise vector, at index n. This is the general formulation of the cocktail party problem with P microphones and r conversations observed for N time points. Here, Y is not necessarily a time series, but data which are indexed sequentially. S is assumed to have full row rank.

We assume that the source signals are piecewise constant. Each segment can be of any length, and adjacent segments have different means. Latent variables are driven by the same underlying system state, and hence may share change locations, but change directions and magnitudes are not necessarily the same. We assume that the linearly-mixed signals are corrupted by independent Gaussian noise, but noise variances are not necessarily the same across channels. In the cocktail party analogy, this means that each microphone is subject to a different amount of noise due to the environment and microphone quality. The Gaussian assumption is standard in parametric change detection models [3, 22, 26].

We aim to decompose Y into its components without further information. Although the decomposition solution is not unique, [12] reports that sparsity formulations in their Bayesian latent variable model helped to stabilize fitting. We similarly apply multiple levels of sparsity in our model, as described in the next section.

4 Proposed Method: ABACUS

We introduce our Bayesian data model and estimation method, as well as our change detection approach which makes use of MCMC posterior samples.

4.1 A Bayesian Latent Variable Model We decompose source signals further into components consisting of either additive outliers (AO) or level shifts (LS). Additive outliers are abrupt mean changes lasting for only one index, while level shifts persist for two or more indices. This decomposition allows us to naturally distinguish between the two types of changes, such that they can be studied separately, e.g., a user may remove additive outliers and retain level shifts for analysis. Let K be a user-specified upper bound for $\text{rank}(S) = r$ such that $r \le K < P$. Then our modified formulation is

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

where M is the $P \times K$ mixing matrix, S is the $K \times N$ source signal matrix, E is the $P \times N$ error matrix, $S^{(0)}$ and $S^{(1)}$ are the $K \times N$ component matrices of S consisting AO and LS respectively, $V^{(0)}$ and $V^{(1)}$ are $K \times N$ 'sparse' matrices, and $\triangle$ is the differencing operator. The diagonal covariance matrix of $E_{\cdot n}$ is denoted by $\Psi = \text{diag}\,(\psi)$, so $E_{\cdot n} \sim \text{N}(0,\ \Psi)$.

We place sparse group priors on the columns of M and rows of $V^{(0)}$ and $V^{(1)}$ for dimensionality reduction of the latent space. Furthermore, we place sparse group priors on the columns of $V^{(0)}$ and $V^{(1)}$ to select a subset of indices as change locations. We also use elementwise sparsity on $V^{(0)}$ and $V^{(1)}$ to allow sparse changes for each latent variable.

We choose to use horseshoe priors because the horseshoe-shaped shrinkage profile discovers null values without diminishing strong signals. [7]. We extend the global-local shrinkage hierarchy to impose sparsity in the model at the element and group level.

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

$$\begin{aligned} M_{\cdot h}|\lambda^{(0)}_h, \lambda^{(1)}_h,\ \tau^{(0)}, \tau^{(1)}, \Psi &\sim \text{N}\left(0, \lambda^{(0)}_h \lambda^{(1)}_h \tau^{(0)} \tau^{(1)} \Psi\right) \\ V^{(d)}_{hn}|\phi^{(d)}_n, \lambda^{(d)}_h, \gamma^{(d)}_{hn},\ \tau^{(d)} &\sim \text{N}\left(0, \phi^{(d)}_n \lambda^{(d)}_h \gamma^{(d)}_{hn} \tau^{(d)}\right) \\ \psi_i &\sim \Gamma^{-1}\,(1,\ 1) \\ \tau^{(d)}|\xi^{(d)} &\sim \Gamma^{-1}\left(\frac{1}{2},\ \frac{1}{\xi^{(d)}}\right) \\ \lambda^{(d)}_h|\eta^{(d)}_h &\sim \Gamma^{-1}\left(\frac{1}{2},\ \frac{1}{\eta^{(d)}_h}\right) \\ \phi^{(d)}_n|\omega^{(d)}_n &\sim \Gamma^{-1}\left(\frac{1}{2},\ \frac{1}{\omega^{(d)}_t}\right) \\ \gamma^{(d)}_{hn}|\zeta^{(d)}_n &\sim \Gamma^{-1}\left(\frac{1}{2},\ \frac{1}{\zeta^{(d)}_{hn}}\right) \\ \xi^{(d)}, \eta^{(d)}_h, \omega^{(d)}_n, \zeta^{(d)}_{hn} &\sim \Gamma^{-1}\left(\frac{1}{2},\ 1\right) \end{aligned}$$

where $N()$ denotes the Gaussian distribution and $\Gamma^{-1}()$ denotes the Inverse Gamma distribution. Marginally, the shrinkage parameters $\tau^{(d)}, \lambda^{(d)}_h,\ \phi^{(d)}_n$ and $\gamma^{(d)}_{hn}$ are half-Cauchy, as in the horseshoe setup. Given the shrinkage parameters, we impose the prior belief that the source signals are independent, but the posterior is not necessarily so.

Let $D^{(1)}$ be the matrix representation of $\triangle$ such that $S^{(1)}\left[D^{(1)}\right]^T = V^{(1)}$, and let $D^{(0)} = I$ such that $S^{(0)} = V^{(0)}$. Now, we define the expression $F = SS^T + \text{diag}\left(\tau^{(0)}\tau^{(1)}\lambda^{(0)}\lambda^{(1)}\right)^{-1}$, which appears below.

For $1 \le i \le P$, $1 \le n \le N$, and $d \in \{0, 1\}$, we derive the full conditionals for the posterior distribution of the

main model components below. First,

$$M_{i\cdot}|\cdot \sim \mathrm{N}\left(F^{-1}SY_{i\cdot},\ \psi_i F^{-1}\right)$$

$$\psi_i|\cdot \sim \Gamma^{-1}\left(1+\frac{N}{2},\ 1+\frac{1}{2}(Y_{i\cdot}-M_{i\cdot}S)^T(Y_{i\cdot}-M_{i\cdot}S)\right)$$

and for $V_{\cdot n}^{(d)}$, the full conditional distribution is

$$\mathrm{N}\left(\left[B^{(n)}\right]^{-1}M^T\Psi^{-1}C^{(n)}\left[D^{(d)}\right]_{\cdot n}^{-1},\ \left[B^{(n)}\right]^{-1}\right)$$

where

$$B^{(n)} = M^T\Psi^{-1}M\left(\left[D^{(d)}\right]_{n\cdot}^{-T}\left[D^{(d)}\right]_{\cdot n}^{-1}\right) + \mathrm{diag}\left(\phi_n^{(d)}\lambda^{(d)}\gamma_{\cdot n}^{(d)}\tau^{(d)}\right)^{-1}$$

$$C^{(n)} = Y - MS + MV_{\cdot n}^{(d)}\left[D^{(d)}\right]_{n\cdot}^{-T}.$$

We use Gibbs sampling to approximate the posterior. The procedure is easily parallelized. Furthermore, the number of model components and parameters depend on K and correctly setting a small K can significantly reduce computational time.

In our modified $Y = MS+E$ model, multiple levels of sparsity regulate the transformations each solution pair M and S can take to reach a different solution pair, but we cannot identify the sign and scaling of M and S. To recover the components and parameters empirically, we use the median of the posterior samples to provide robustness against possible movements of the sampling path between different solutions.

4.2 Change Detection In our data model, $V^{(0)}$ and $V^{(1)}$ contain the changes for each latent variable at each index. The matrices are sparse since only entries which correspond to changes are nonzero. Let $f_n^{(d)}$ be the element with the largest magnitude in $V_{\cdot n}^{(d)}$. At any index n, $f_n^{(d)}$ 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. We use the median defined

$$\widehat{g}_n^{(d)} = \mathrm{median}\left(\widehat{f}_n^{(d)}\right)$$

for robustness with empirical samples.

Since we impose horseshoe priors on $V^{(d)}$, the entries are shrunk to approximately zero but not exactly zero. To identify the approximately zero values in the estimated $\widehat{g}^{(d)}$, we apply kernel density estimation on $|\widehat{g}^{(d)}|$ with a rectangular kernel and set the cutoff to be at the first minimum in the density function such that the minimum value is below threshold δ. The threshold ensures that the approximately zero and non-zero values are sufficiently different. We set $\delta = 10^{-10}$ for all our experiments.

5 Implementation

We fit the full Bayesian latent variable model in Section 4.1 by first fitting a partial model. The partial model differs only in that it does not include $S^{(0)}$ or $V^{(0)}$ and their associated parameters, and hence we drop the superscripts when referring to its components and parameters. Change points *cpt* detected by the partial model are a mix of additive outliers (AO) and level shifts (LS), with the former being detected as two consecutive mean changes of opposite signs in $\widehat{g}$. We distinguish between the two types of changes according to this observation with Algorithm 1, and produce additive outliers *cpt*0 and level shifts *cpt*1. We decompose the estimated components and parameters from the partial model according to *cpt*0 and *cpt*1, and pass them to the full model as initialization. For example, $V^{(0)}$ is initialized with values from $\widehat{V}$ at *cpt*0, and $V^{(1)}$ is initialized with values from $\widehat{V}$ at *cpt*1.

Algorithm 1: Separating AO and LS changes

Data: Estimated $\widehat{g}$, ordered change points *cpt*
Result: Additive outliers *cpt*0, level shifts *cpt*1

```
cpt0 = cpt1 = {};
i = 1;
while not at end of cpt do
    condition 1: cpt[i+1] − cpt[i] = 1;
    condition 2: ĝ corresponding to cpt[i] and
      cpt[i+1] are of opposite signs;
    if condition 1 and 2 are True then
        add cpt[i] to cpt0;
        i = i + 2;
    else
        add cpt[i] to cpt1;
        i = i + 1;
    end
end
```

The partial model is smaller and hence can quickly estimate components and parameters for initialization. Empirically, good initialization helps the Gibbs sampler converge to solutions providing better distinction between the two types of changes in the full model. The entire procedure is shown in Figure 2. The final two boxes in green indicate the final outputs for change detection and source recovery.

6 Simulation Study

We conduct several experiments according to the model $Y = MS + E$ described in Section 3. We fix the latent space dimensionality $r = 3$, and vary N and P. Some methods require a user-specified K as an estimate for r, and we test their robustness to K. Entries of M are drawn independently from $\mathrm{Unif}(-1, 1)$, and each

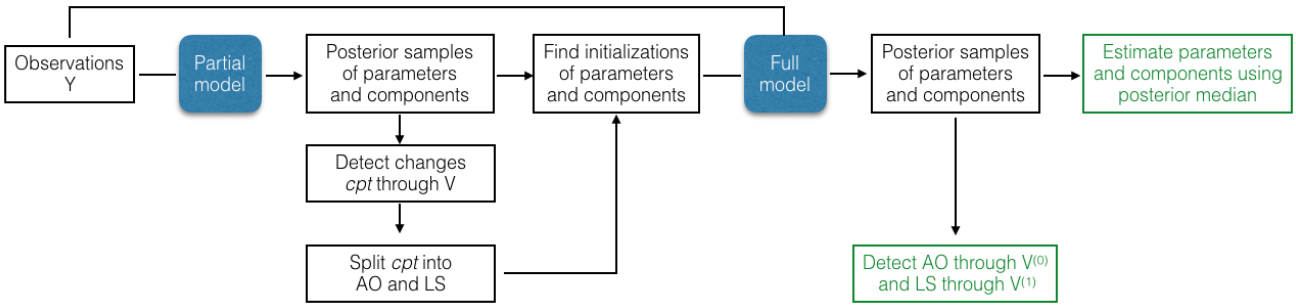


Figure 2: Implementation procedure. From observations Y, a partial model is first fit and its estimations initialize the full Bayesian model. Final estimates of source signals and change points are obtained from the median of MCMC samples.

noise variance as $\psi_i \sim \text{Unif}(0.1, 5)$. Given the number of additive outliers and level shifts, change locations are sampled uniformly at random from $\{2, 4, 6, \ldots, N-1\}$. This ensures that level shifts are at least of length two and that we do not unintentionally construct level shifts through consecutive additive outliers. To construct sparse changes, at each change location, the number of latent signals experiencing change is selected uniformly at random. Change magnitudes are drawn from $\text{Unif}(1, 5)$ with a random sign.

We compare ABACUS against state-of-the-art change detection techniques and popular latent variable models which are marked by $\times$ and $\circ$, respectively, in plots in this section. We use default parameters in software packages unless otherwise specified. To find additive outliers, the minimum segment length is set to one where possible. Detected changes are categorized into additive outliers and level shifts using Algorithm 1 without Condition 2, except for TSO mentioned below which automatically outputs different types of changes. For all MCMC procedures, number of iterations is 3000 and burn-in is 500. Each simulation is run 100 times.

Amongst competing multivariate change detection methods, GFLseg [6] finds candidate mean changes by group fused Lasso followed by selection via dynamic programming. E-divisive [20] uses binary segmentation to iteratively locate each change point through measuring between-segment distance by the energy statistic. We specify its moment index parameter $\alpha = 2$ to find level shifts, and $min.size = 2$ the smallest segment length allowed, which implies E-divisive is unable to find additive outliers. BCP [10] is a Bayesian method which models the presence of mean change at each location through an indicator variable and infers the posterior probability of change. BCP outputs a set of change points corresponding to each posterior sample, hence for evaluation we compute the average metric across all these sets. We also combine BPCA [4] and BCP to obtain a two-step Bayesian approach to first compress and then detect.

Inspect [25] transforms observations into a univariate series through cumulative sum transformation before applying wild binary segmentation. We also test three univariate methods by first applying PCA to the observations. PELT [18] is a popular parametric approach that uses dynamic programming to efficiently find the segmentation that minimizes the negative log-likelihood plus a penalty. We refer to the nonparametric version as np-PELT, which uses the empirical distribution instead [14]. A third method, TSO, jointly estimates ARIMA model parameters and change effects due to additive outliers and level shifts [8].

To fit the latent variable model, we tested against well-established methods including Independent Component Analysis (ICA), Factor Analysis (FA) and Bayesian Principal Component Analysis (BPCA). Note that ICA and FA do not impose sparsity assumptions, whereas BPCA imposes sparsity on the columns of M. For ICA, we use the FastICA implementation which measures non-Gaussianity using negentropy [16]. For FA, we use the factanal function in R [24] which automatically checks for identifiability given K and does not fit a model if K is too large to fit a unique model.

6.1 Evaluation Criteria We evaluate the detection of additive outliers and level shifts separately since some competing methods [20] detect one but not the other. We report precision and recall, and treat an estimate as accurate if it is within w of a true change location. We set $w = 1$ for the small sample experiment in Section 6.2, and $w = 3$ for the larger sample experiments in Section 6.3 and 6.4.

We evaluate the quality of model recovery through components M and S, and noise variance parameter ψ. Given true mixing matrix M and estimate $\widehat{M}$, we center and scale each row of the matrices and measure their dissimilarity using the squared trace metric in [12],

$$\epsilon_M = \frac{1}{P^2} Tr\left(MM^T - \widehat{M}\widehat{M}^T\right).$$

The metric ϵ_M is invariant to orthogonal rotation and allows cases where either MM^T or $\widehat{M}\widehat{M}^T$ is singular. Next, given true source signals S and estimate $\widehat{S}$, we measure their dissimilarity using a variation of averaged squared Euclidean distance

$$\epsilon_S = \frac{1}{r}\sum_{i=1}^{r}(1 - |\rho_i|)$$

where ρ_i is the Pearson correlation coefficient between $S_{i\cdot}$ and some $\widehat{S}_{j\cdot}$, and each pair is found greedily by descending magnitude of correlation. This measure is invariant to sign and label switching. Finally, given true noise variance ψ and estimate $\widehat{\psi}$, the difference is measured by their scaled squared norm

$$\epsilon_E = \frac{1}{P}\|\psi - \widehat{\psi}\|_2^2.$$

6.2 Simulation 1: Variations in P We test the case of small sample size $N = 100$ and varying $P \in \{10, 30, 60, 90, 110\}$. Each sample has two additive outliers and two level shifts, and K is set to 5.

Competing methods have high precision but low recall on additive outliers. An additive outlier is difficult to detect in low signal-to-noise ratio settings especially since the change only lasts for one time unit. ABACUS has comparatively high recall at the cost of a slight reduction in precision. The trade-off can be adjusted, for instance by changing the cutoff threshold for classifying change points. As P increases in Figure 3, ABACUS can locate most of the additive outliers, and is one of the best-performing methods for level shifts. Both precision and recall on level shifts decrease as P increases for BCP, possibly because parameters such as the prior on change probabilities need to be adjusted. BPCA + BCP has more consistent performance, indicating the advantage of detecting changes on latent signals. In terms of model recovery, our method also gives the lowest errors for M, S and ψ, see Figure 4.

6.3 Simulation 2: Variations in N We fix $P = 10$ and vary $N \in \{600, 800, 1000, 1200, 1400, 1600\}$. Each sample has $\frac{N}{100}$ additive outliers and $\frac{N}{100}$ level shifts, and K is set to 5. Performance of all methods is consistent across N, as shown in Figures 5 and 6. BCP shows deteriorating performance in detecting level shifts just as it did in Section 6.2, again possibly because model parameters need to be adjusted according to the sample size. Overall, ABACUS offers the best balance of precision and recall on additive outliers while all other competing change detection methods tend to miss them. ABACUS has the highest recall for level shifts, and almost always has the lowest errors for model recovery.

6.4 Simulation 3: Variations in K We fix $P = 10$ and $N = 1000$. Each sample has ten additive outliers and ten level shifts. We vary the user-specified estimate of the latent space dimensionality K between 2 and 9. The true dimensionality r is 3. The horizontal lines in Figures 7 and 8 correspond to results of methods which do not have the parameter K. According to Figure 7, the change detection results of ABACUS are consistent across K. From Figure 8, ABACUS has much more consistent error ϵ_S in S compared to competing latent variable models, whose ϵ_S increases sharply at $K \geq r$.

7 Application to Real Data

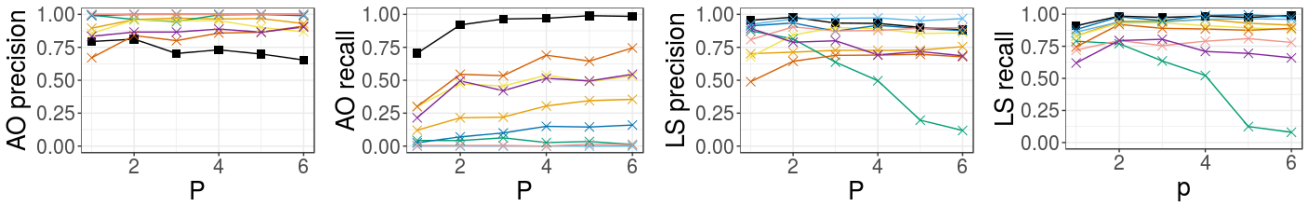
In both data applications below we set $K = 5$ and also study the robustness of ABACUS to different K values.

7.1 aCGH Data Array-based comparative genomic hybridization (aCGH) is a technique for studying copy number alterations in event of diseases. We obtain the dataset from the R package ecp [17], which has already removed sequences with more than 7% missing values, and leaves 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 [13]. A negative ratio indicates deletion, a positive ratio indicates amplification, and zero indicates an unaltered segment. We expect shared change locations for individuals with the same medical condition.

To reduce computations and ease visualization, we thin the samples by taking every 20^{th} value. We arrive at a dataset with $P = 43$ and $N = 111$. ABACUS takes approximately one minute to run on a standard desktop computer, and finds three additive outliers and seven level shifts. An additive outlier here indicates a shorter segment of genetic aberration compared to a level shift.

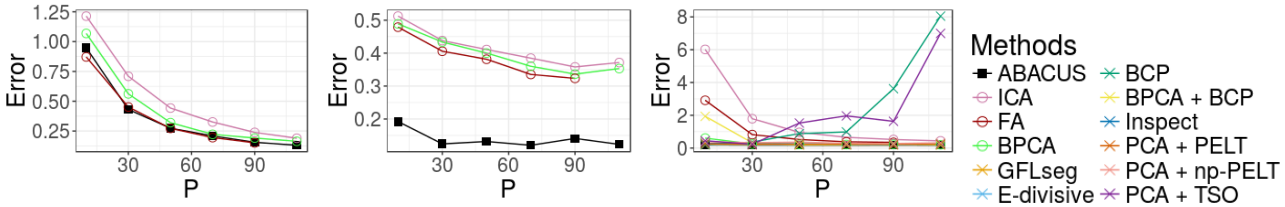
Above 99% of the variance of our estimated latent signals can be explained by four principal components, while those from ICA and FA require all five. In Figure 9, the third signal recovered exhibits no evident changes. We map the four other signals to unique sets of genetic aberrations in different tumor stages in Table 1. For instance, patients with concurrent genetic aberrations on chromosome arms 2q, 3q and 20p/q tend to be in stage pT_1, hence the changes detected can be indicative of diseases for new patients. The mapping is established based on a bladder tumor research article [5] which lists the frequent genomic alterations by chromosome arm in stages pT_a, pT_1 and pT_{2-4}. Stages are determined pathologically depending on tumor size and location.

ABACUS performs consistently across different K. Figure 10 shows that for $K \in \{10, 15, 20, 25, 30\}$, the change points and latent source signals recovered are



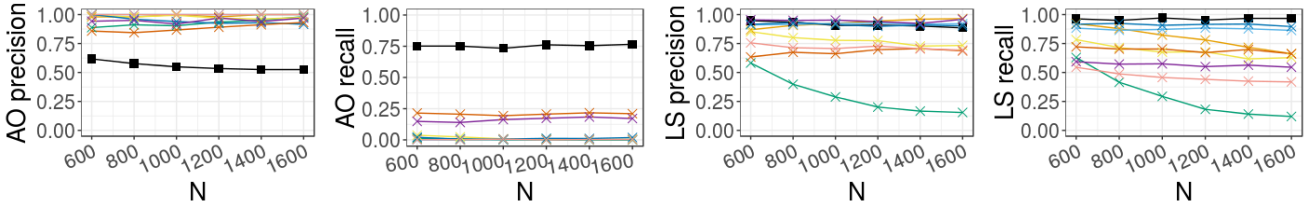
(a) Additive outliers **(b)** Level shifts

Figure 3: Average errors in change detection as data dimensionality P is varied; $N = 100$ and $K = 5$ are fixed.



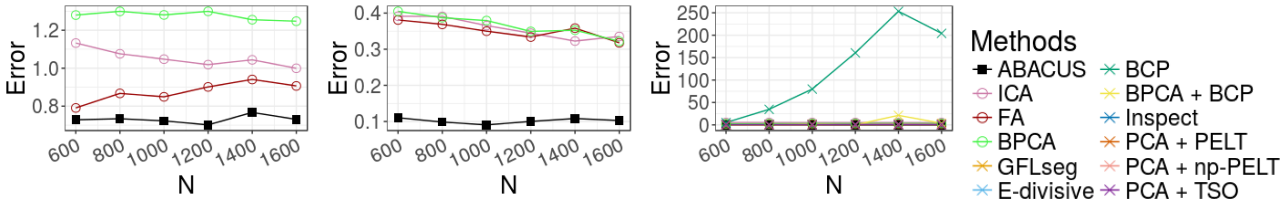
(a) Error ϵ_M for M **(b)** Error ϵ_S for S **(c)** Error ϵ_E for ψ

Figure 4: Average errors in model recovery as data dimensionality P is varied; $N = 100$ and $K = 5$ are fixed. FA does not support computations for $P = 110$ due to non-identifiability.



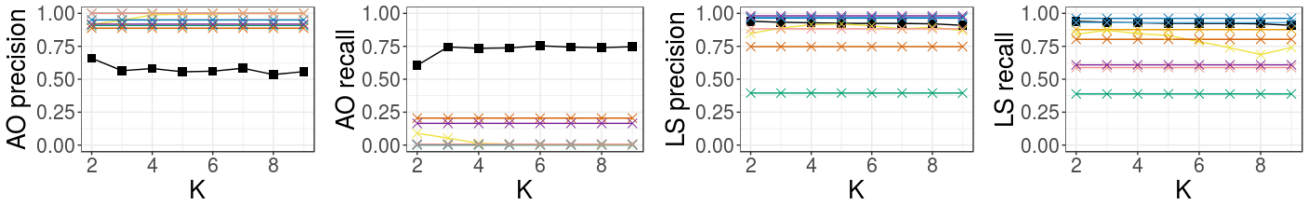
(a) Additive outliers **(b)** Level shifts

Figure 5: Average errors in change detection as sample size N is varied; $P = 10$ and $K = 5$ are fixed.



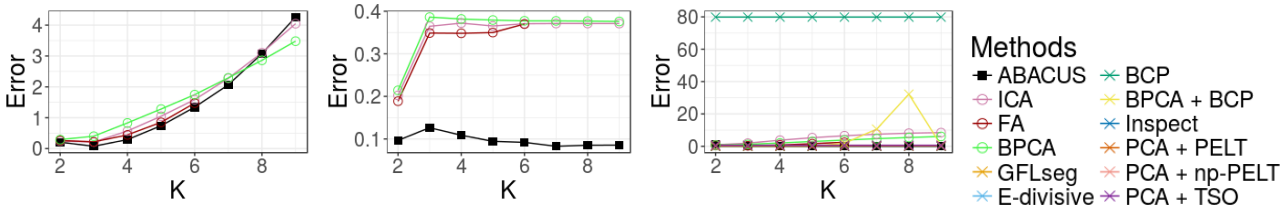
(a) Error ϵ_M for M **(b)** Error ϵ_S for S **(c)** Error ϵ_E for ψ

Figure 6: Average errors in model recovery as sample size N is varied; $P = 10$ and $K = 5$ are fixed.



(a) Additive outliers **(b)** Level shifts

Figure 7: Average errors in change detection as estimated latent space dimensionality K is varied; fixed $N = 1000$ and $P = 10$.



(a) Error ϵ_M for M **(b)** Error ϵ_S for S **(c)** Error ϵ_E for ψ

Figure 8: Average errors in model recovery as latent space dimensionality parameter K is varied; $N = 1000$ and $P = 10$ are fixed. FA does not support computations for $K \geq 7$ due to non-identifiability.

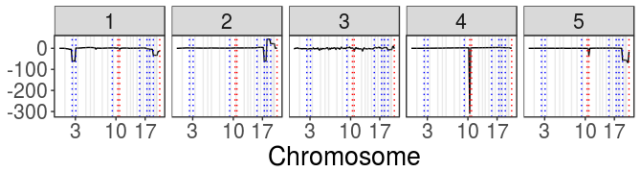


Figure 9: aCGH: Latent source signals (1-5) recovered (black), and additive outliers (red) and level shifts (blue) detected. Gray lines indicate the boundaries between chromosome pairs.

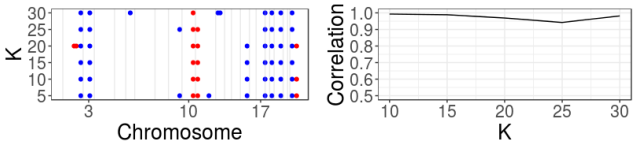
S	Chromosome arm with changes	Tumor stage
1	2q, 3q, 20p/q	pT_1
2	17p/q, 18p/q, 19p/q, 20p/q	pT_1
4	10q	pT_a, pT_1, pT_{2-4}
5	11p, 20p/q	pT_{2-4}

Table 1: aCGH: Genetic aberrations corresponding to changes detected on latent source signals. To read the table, 20p is the short arm of chromosome 20, and 20q is the long arm. Tumor stages range from a, 1 to 4 in order of severity.

very similar to those found with $K = 5$.

7.2 Electric Power Consumption Data This dataset contains per-minute measurements of electric power consumption in one household and is available on the UCI Machine Learning Repository [9]. The data has seven dimensions: global active power (GAP), global reactive power (GRP), voltage (V), global intensity (GI), and three sub-meterings for the kitchen (S1), laundry room (S2) and heating system (S3). We expect shared change points since the channels are related arithmetically, and some electrical appliances tend to be used simultaneously. For instance, $\frac{1000}{60}\text{GAP} - \text{S1} - \text{S2} - \text{S3}$ is the power consumed by appliances outside of the sub-metered zones. We analyze a full day's worth of data, that is, the observation matrix has $P = 7$ and $N = 1440$. ABACUS takes approximately fifteen minutes to run on a standard desktop computer.

The data does not follow our model assumptions exactly since the amount of fluctuations or noise is more significant in the first half of the day and there are minor trend changes in the second, but ABACUS is robust and with post-processing it finds one additive outlier and sixteen level shifts. We post-process by dynamic programming to prune the initially estimated



(a) Additive outliers (red) and level shifts (blue) **(b)** Average correlation to latent signals at $K = 5$

Figure 10: aCGH: Changes and latent source signals recovered by ABACUS are similar regardless of the specification of K.

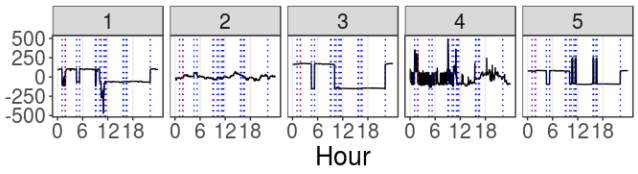
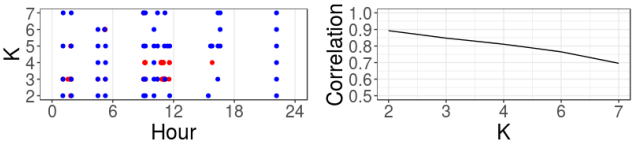


Figure 11: Power: Latent source signals (1-5) recovered (black), and additive outliers (red) and level shifts (blue) detected.



(a) Additive outliers (red) and level shifts (blue) **(b)** Average correlation to latent signals at $K = 5$

Figure 12: Power: Changes and latent source signals recovered by ABACUS are similar regardless of the specification of K.

level shifts. This is similar to GFLseg [6], except that we apply the procedure on the latent source signals which are less contaminated by noise.

The change points are indicative of the household's pattern of electricity usage, which concentrates in the first half of the day as illustrated in Figure 11. The fourth latent signal reflects the usage fluctuations and trends which differ across the two halves of the day as measured by GAP and GI. ABACUS performs consistently across different specifications of K. Figure 12 shows that for $K \in \{2, 3, 4, 6, 7\}$, the estimated change points and latent source signals recovered are similar to those found at $K = 5$.

Since the sub-meterings S1, S2 and S3 demonstrate distinct level shifts when the respective appliances are utilized, we extract ground truths for level shifts by finding positions where these signals deviate from their base levels. Compared to other change detection methods in Figure 13 and 14, ABACUS has the best overall performance with precision = 1 and recall = 0.889.

8 Conclusion

In this paper, we propose ABACUS, an automatic change detection procedure which makes use of Bayesian latent variable modeling. Due to the separation of additive outlier and level shift effects in the model,

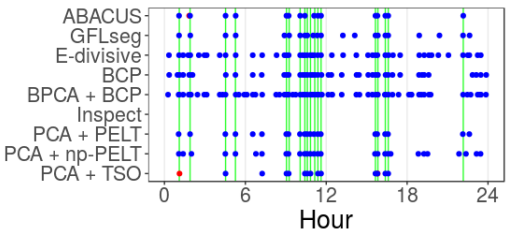
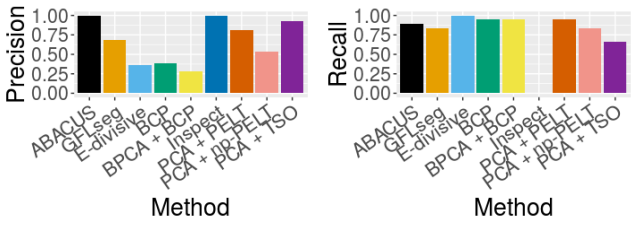


Figure 13: Power: Additive outliers (red) and level shifts (blue) estimated vs ground truth level shifts (green).



(a) Precision **(b)** Recall

Figure 14: Power: Performance in estimating level shifts.

ABACUS naturally identifies these two types of changes separately, unlike many competing approaches.

In simulations, ABACUS shows competitive or superior performance in both change detection and model recovery. In two real data applications, ABACUS found relevant change points and source signals. It is robust to over-specification of K, an important property since the true value is rarely known to the user in practice.

9 Acknowledgement

Support from Xerox PARC Faculty Award, National Science Foundation (DMS-1455172), Cornell Atkinson Center (AVF-2017), USAID, and Cornell Institute of Biotechnology & NYSTAR is gratefully acknowledged.

References

[1] C. Alippi, G. Boracchi, D. Carrera, and M. Roveri, *Change detection in multivariate datastreams: Likelihood and detectability loss*, in IJCAI, 2016.

[2] G. K. Atia, *Change detection with compressive measurements*, Signal Processing Letters, 22 (2015), pp. 182–186.

[3] D. Barry and J. A. Hartigan, *A bayesian analysis for change point problems*, JASA, 88 (1993), pp. 309–319.

[4] C. M. Bishop, *Bayesian pca*, in NIPS, Cambridge, MA, USA, 1999, MIT Press, pp. 382–388.

[5] E. Blaveri, J. L. Brewer, R. Roydasgupta, J. Fridlyand, S. DeVries, T. Koppie, S. Pejavar, K. Mehta, P. Carroll, J. P. Simko, and F. M. Waldman, *Bladder cancer stage and outcome by array-based comparative genomic hybridization*, Clinical Cancer Research, 11 (2005), pp. 7012–7022.

[6] K. Bleakley and J.-P. Vert, *The group fused lasso for multiple change-point detection*, (2011).

[7] C. M. Carvalho, N. G. Polson, and J. G. Scott, *Handling sparsity via the horseshoe*, in AISTATS, 2009.

[8] J. L. de Lacalle, *tsoutliers: Detection of Outliers in Time Series*, 2017. R package version 0.6-6.

[9] D. Dheeru and E. Karra Taniskidou, *UCI machine learning repository*, 2017.

[10] C. Erdman and J. Emerson, *bcp: An r package for performing a bayesian analysis of change point problems*, JSS, 23 (2007), pp. 1–13.

[11] P. Fryzlewicz, *Wild binary segmentation for multiple change-point detection*, The Annals of Statistics, 42 (2014), pp. 2243–2281.

[12] C. Gao, C. Brown, and B. Engelhardt, *A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects*, (2013).

[13] F. Harlé, F. Chatelain, C. Gouy-Pailler, and S. Achard, *Bayesian model for multiple change-points detection in multivariate time series*, IEEE Transactions on Signal Processing, 64 (2016), pp. 4351–4362.

[14] K. Haynes, R. Killick, P. Fearnhead, and I. Eckley, *changepoint.np: Methods for Nonparametric Changepoint Detection*, 2016. R package version 0.0.2.

[15] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*, in Wiley Interscience, 2001.

[16] A. Hyvrinen and E. Oja, *Independent component analysis: algorithms and applications*, Neural Networks, 13 (2000), pp. 411 – 430.

[17] N. A. James and D. S. Matteson, *ecp: An R package for nonparametric multiple change point analysis of multivariate data*, JSS, 62 (2014), pp. 1–25.

[18] R. Killick, P. Fearnhead, and I. Eckley, *Optimal detection of changepoints with a linear computational cost*, JASA, 107 (2012), pp. 1590–1598.

[19] D. R. Kowal, D. S. Matteson, and D. Ruppert, *Dynamic shrinkage processes*, Arxiv preprint arXiv:1707.00763, (2018).

[20] D. S. Matteson and N. A. James, *A nonparametric approach for multiple change point analysis of multivariate data*, JASA, 109 (2014), pp. 334–345.

[21] D. S. Matteson and R. S. Tsay, *Independent component analysis via distance covariance*, JASA, 112 (2017), pp. 623–637.

[22] A. B. Olshen and E. Venkatraman, *Circular binary segmentation for the analysis of array-based dna copy number data*, Biostatistics, 5 (2004), pp. 557 – 572.

[23] T. D. Popescu, *Blind separation of vibration signals and source change detection application to machine monitoring*, Applied Mathematical Modelling, 34 (2010), pp. 3408 – 3421.

[24] R Foundation for Statistical Computing, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2018.

[25] W. Tengyao and S. R. J., *High dimensional change point estimation via sparse projection*, JRSS: Series B, 80 (2018), pp. 57–83.

[26] Y. Xie, M. Wang, and A. Thompson, *Sketching for sequential change-point detection*, in GlobalSIP, Dec 2015, pp. 78–82.

[27] W. Zhang, N. A. James, and D. S. Matteson, *Pruning and nonparametric multiple change point detection*, in 2017 IEEE ICDMW, Nov 2017, pp. 288–295.

[28] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and L. Zhang, *Robust principal component analysis with complex noise*, in ICML, 2014.