The Annals of Statistics
2018, Vol. 46, No. 2, 711–744
https://doi.org/10.1214/17-AOS1565

MULTISCALE BLIND SOURCE SEPARATION

BY MERLE BEHR[*,1], CHRIS HOLMES[†,2] AND AXEL MUNK[*,‡,3]

University of Goettingen[*], *University of Oxford*[†], *and*
Max Planck Institute for Biophysical Chemistry[‡]

We provide a new methodology for statistical recovery of single linear mixtures of piecewise constant signals (sources) with unknown mixing weights and change points in a multiscale fashion. We show exact recovery within an ε-neighborhood of the mixture when the sources take only values in a known finite alphabet. Based on this we provide the SLAM (Separates Linear Alphabet Mixtures) estimators for the mixing weights and sources. For Gaussian error, we obtain uniform confidence sets and optimal rates (up to log-factors) for all quantities. SLAM is efficiently computed as a nonconvex optimization problem by a dynamic program tailored to the finite alphabet assumption. Its performance is investigated in a simulation study. Finally, it is applied to assign copy-number aberrations from genetic sequencing data to different clones and to estimate their proportions.

1. Introduction.

As the presented methodology requires a quite broad range of techniques, we will briefly introduce them in this section for explanatory purposes. Details are given in subsequent sections and the Supplementary Material [5].

1.1. *The statistical blind source separation problem.*

We will start by introducing a particular kind of the blind source separation (BSS) problem which will be considered throughout this paper. More generally, in BSS problems (for a review, see Section 1.8) one observes a mixture of signals (sources) and aims to recover these sources from the available observations, usually corrupted by noise. The blindness refers to the fact that neither the sources nor the mixing weights are known. Of course, without any additional information on the sources the BSS problem is unsolvable as the weights and sources cannot be separated, in general. However, under the additional assumption that the sources take values in a known finite alphabet, we will show that estimation of all quantities and inference for these is indeed possible.

Received July 2016; revised January 2017.

[1] Supported by DFG RTG 2088 and CRC 803 Z.

[2] Supported by the Oxford-Man Institute, an EPSRC programme grant i-like, the Medical Research Council UK, and the Alan Turing Institute

[3] Supported by DFG CRC 803 Z and FOR 916 B1, B3, Z.

MSC2010 subject classifications. Primary 62G08, 62G15; secondary 92D10.

Key words and phrases. Multiscale inference, honest confidence sets, change point regression, finite alphabet linear mixture, exact recovery, genetic sequencing.

Motivated by several applications mainly from digital communications [e.g., the recovery of mixtures of multi-level PAM signals (see [56, 69])] and cancer genetics (see Section 1.7), we assume, from now on, that the m source functions f^i, $i = 1, \ldots, m$, consist of arrays of constant segments, that is, step functions with unknown jump sizes, numbers, and locations of change points (c.p.'s), respectively. More specifically, let for a finite (known) ordered alphabet $\mathfrak{A} = \{a_1, \ldots, a_k\} \subset \mathbb{R}$, with $a_1 < \cdots < a_k$, each source function be in the class of step functions on $[0, 1)$

$$\mathcal{S}(\mathfrak{A}) := \left\{ \sum_{j=0}^{K} \theta_j \mathbb{1}_{[\tau_j, \tau_{j+1})} : \theta_j \in \mathfrak{A}, 0 = \tau_0 < \cdots < \tau_K < \tau_{K+1} = 1, K \in \mathbb{N} \right\}. \tag{1}$$

Note that this implies that for each source function the number $K(f^i)$ of c.p.'s is assumed to be finite, possibly different, and unknown. We will assume $\theta_j \neq \theta_{j+1}$ for $j = 0, \ldots, K$ to ensure identifiability of the c.p.'s τ_j. Note that without further specification $\mathcal{S} := \mathcal{S}(\mathfrak{A})$ is an extremely flexible class of functions, including any discretized source function taking values in $\mathfrak{A}$. Moreover, we define the set of all possible (linear) mixtures with m components each in $\mathcal{S}$ as

$$\mathcal{M} := \mathcal{M}(\mathfrak{A}, m) = \left\{ \omega^\top f = \sum_{i=1}^{m} \omega_i f^i : \omega \in \Omega(m) \text{ and } f \in \mathcal{S}(\mathfrak{A})^m \right\}, \tag{2}$$

with mixing weights ω in the m-simplex

$$\Omega(m) := \left\{ \omega \in \mathbb{R}^m : 0 \leq \omega_1 \leq \cdots \leq \omega_m \text{ and } \sum_{i=1}^{m} \omega_i = 1 \right\}. \tag{3}$$

For a set $\tilde{\Omega} \subset \Omega(m)$ we define $\mathcal{M}(\mathfrak{A}, \tilde{\Omega})$ analogously. Throughout the following, we assume that m is known. Extension to unknown m is akin to a model selection type of problem and beyond the scope of this paper.

In summary, in this paper we will be concerned with the *statistical blind source separation regression* model.

The SBSSR-model. For a given finite alphabet $\mathfrak{A}$ and a given number of mixture components $m \in \mathbb{N}$ let $g = \sum_{i=1}^{m} \omega_i f^i \in \mathcal{M}$ be an arbitrary mixture of m piecewise constant source functions $f^i \in \mathcal{S}$. Suppose we observe

$$Y_j = g(x_j) + \sigma \varepsilon_j, \qquad j = 1, \ldots, n, \tag{4}$$

at sampling points $x_j := (j-1)/n$, s.t. the error $(\varepsilon_1, \ldots, \varepsilon_n)^\top \sim \mathcal{N}(0, I_n)$, $\sigma > 0$, that is, i.i.d. centered normal random variables with variance σ^2.

EXAMPLE 1.1. In Figure 1, a mixture g of $m = 3$ source functions f^1, f^2, f^3, taking values in the alphabet $\mathfrak{A} = \{0, 1, 2\}$, is displayed. The mixing weights are given by $\omega^\top = (0.11, 0.29, 0.6)$. Normal noise with standard deviation $\sigma = 0.22$ is added according to the SBSSR-model, $n = 7680$. Both, n and σ were chosen close to our data example in Section 5.

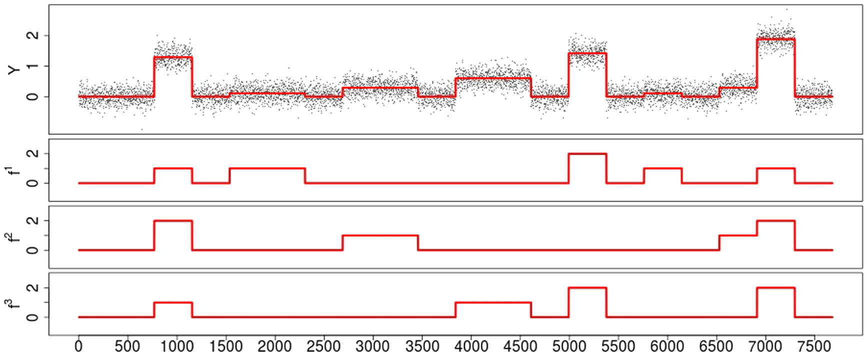

FIG. 1. *The mixture* $g = 0.11f^1 + 0.29f^2 + 0.6f^3$, *together with the observations* Y (*gray dots*), *and the sources* f^1, f^2, f^3 *from Example* 1.1 (*from top to bottom*).

In summary, the unknowns in the SBSSR-model are:

1. the weights $\omega = (\omega_1, \dots, \omega_m)^\top$ and
2. the source functions f^i, $i = 1, \dots, m$, that is, their
   (a) number of c.p.'s $K(f^i)$,
   (b) c.p. locations τ_j^i, $j = 1, \dots, K(f^i)$, and
   (c) function values $f^i(x)$ $(\in \mathfrak{A})$ at locations $x \in [0, 1)$.

In this paper, we will address estimation of all the quantities in 1. and 2. and, in addition, we will construct under further assumptions:

3. a uniform (i.e., honest) confidence region $\mathcal{C}_{1-\alpha}$ for the weights ω and
4. asymptotically uniform multivariate confidence bands for the source functions $f = (f^1, \dots, f^m)^\top$.

REMARK 1.2.

(a) For simplicity, we assume throughout the following that g in (4) is sampled equidistantly at $x_j = (j-1)/n$, $j = 1, \dots, n$ and that all functions are defined on the domain $[0, 1)$. We stress that extensions to more general domains $\subseteq \mathbb{R}$ and sampling designs are straightforward under suitable assumptions (see, e.g., [11]) but will be suppressed to ease notation.

(b) Further, for sake of brevity, we will assume that in (4) the variance σ^2 is known, otherwise one may pre-estimate it $\sqrt{n}$-consistently by standard methods; see, for example, [20, 21, 39, 50] and Section 5.

1.2. *Identifiability and exact recovery.* Before we introduce estimators for ω and f, we need to discuss identifiability of these parameters in the SBSSR-model, that is, conditions when g determines them uniquely via $g = \sum_{i=1}^m \omega_i f^i$.

Although, deterministic finite alphabet instantaneous (linear) mixtures, that is, $\sigma = 0$ in the SBSSR-model (4), received a lot of attention in the literature [22, 37, 43, 54, 59, 66, 72], a complete characterization of identifiability remained elusive and has been recently provided in [6], which will be briefly reviewed here as far as it is required for our purposes. Obviously, not every mixture $g \in \mathcal{M}$ in (2) is identifiable. Consider, for example, $\omega \in \Omega(m)$ in (3) such that $\omega_1 = \omega_2$. Then a jump in the source function f^1 has the same effect on the mixture g as a jump in f^2 and hence, f^1 and f^2 cannot be distinguished from the mixture g. Likewise, when ω_1 and ω_2 are close, that is, $\omega_2 - \omega_1 \to 0$, it becomes arbitrarily difficult to separate f^1 and f^2 from the observations Y in the SBSSR-model. For statistical estimation, it is therefore necessary that different source function values $f(x) = (f^1(x), \dots, f^m(x)) \in \mathfrak{A}^m$ are sufficiently well separated by the mixing weights ω. This is quantified by the *alphabet separation boundary* [6]

$$\mathrm{ASB}(\omega) = \mathrm{ASB}(\omega, \mathfrak{A}) := \min_{a \neq a' \in \mathfrak{A}^m} \left|\omega^\top a - \omega^\top a'\right|. \tag{5}$$

A necessary identifiability condition in the SBSSR-model is $\mathrm{ASB}(\omega) > 0$ (see [6], Section 3.A), where the size of $\mathrm{ASB}(\omega)$ can be understood as a conditioning number for the difficulty of separating the sources in the SBSSR-model, that is, the smaller $\mathrm{ASB}(\omega)$, the more difficult separation of sources. Therefore, to quantify the estimation error of any method which serves the purposes in 1.–4. we must restrict to submodels of mixing weights which sufficiently separate different alphabet values in $\mathfrak{A}^m$, that is, for given $\delta > 0$ we introduce

$$\Omega^\delta = \Omega^\delta(\mathfrak{A}, m) := \{\omega \in \Omega(m) : \mathrm{ASB}(\omega) \geq \delta\}. \tag{6}$$

Note further that $\mathrm{ASB}(\omega) > 0$ implies that any jump in the source vector f (i.e., at least one source f^i jumps) occurs as well in the mixture $g = \omega^\top f$ and that $\mathrm{ASB}(\omega)$ coincides with the minimal possible jump height of g.

Just as we have restricted the possible ω's in (6), it is necessary to further restrict the set of possible source functions $f \in \mathcal{S}(\mathfrak{A})^m$ in (1). Consider for example the case of two sources, $m = 2$, such that $f^1 = f^2$. Then $g = \omega_1 f^1 + \omega_2 f^2 = f^1$, independently of ω, and hence, ω cannot be determined from g. Therefore, a certain kind of variability of the sources f^i is necessary to ensure identifiability of the mixing weights ω. We employ from [6] the following simple sufficient identifiability condition.

DEFINITION 1.3. A vector of source functions $f = (f^1, \dots, f^m)^\top \in \mathcal{S}(\mathfrak{A})^m$ is *separable* if there exit intervals $I_1, \dots, I_m \subset [0, 1)$ such that f is constant on I_r with function values

$$f|_{I_r} \equiv [A]_r, \qquad r = 1, \dots, m, \tag{7}$$

with

$$A := a_1 E_m + (a_2 - a_1) I_m = \begin{pmatrix} a_2 & a_1 & a_1 & \dots & a_1 \\ a_1 & a_2 & a_1 & \dots & a_1 \\ \vdots & & & & \vdots \\ a_1 & a_1 & \dots & a_1 & a_2 \end{pmatrix} \in \mathfrak{A}^{m \times m}, \tag{8}$$

where E_m denotes the matrix of ones, I_m the identity matrix, and $[A]_r$ the rth row of A.

The notation "separable" is borrowed from identifiability conditions for nonnegative matrix factorization [2, 24, 57]; see Section 1.8 for details. Separability in Definition 1.3 means that for each of the m sources f^i there is a region where only this source function is "active" (taking the second smallest alphabet value a_2) and all the other sources are "silent" (taking the smallest alphabet value a_1). For example, if we have an alphabet of the form $\mathfrak{A} = \{0, 1, a_3, \dots, a_k\}$, A becomes the identity matrix and separability means that each of the mixing weights ω_i appears at least once in the mixture $g = \omega^\top f$. Note that separability in Definition 1.3 only requires that the values $[A]_r \in \mathfrak{A}^m$ are attained *somewhere* by the source functions $f^1, \dots, f^m$ and does not specify the location. For specific situations it is possible to replace the matrix A in (8) by a different invertible [as a function from $\Omega(m)$ to $\mathbb{R}^m$] matrix if this matrix induces enough variability in the sources for the weights to be identifiable from their mixture (see [6]). Here, however, we consider arbitrary alphabets and number of sources and the separability condition in Definition 1.3 ensures identifiability for arbitrary $\mathfrak{A}$ and m, in general. Note that when the source functions $f = (f^1, \dots, f^m)^\top$ attain all k^m possible function values in $\mathfrak{A}^m$ somewhere in $[0, 1)$, the case of maximal variation, then, in particular, f is separable (see [6] for further examples). We stress that the above assumption (7) on the variability of f is close to being necessary for identifiability (see [6], Theorem 3.1). Hence, without such an assumption no method can provide a unique decomposition of g into the f^i's and its weights ω_i, $i = 1, \dots, m$, even in the noiseless case. Summing up, we will, in the following, restrict to those mixtures g in the SBSSR-model, which are in

$$\mathcal{M}^\delta := \left\{ \omega^\top f = \sum_{i=1}^m \omega_i f^i : \omega \in \Omega^\delta \text{ and } f \in \mathcal{S}(\mathfrak{A})^m \text{ is separable} \right\}. \tag{9}$$

For instance, in Example 1.1 f is separable and $\omega \in \Omega^{0.02}$, that is, $g \in \mathcal{M}^{0.02}$.

The following simple but fundamental result will guide us later on to derive estimators for all quantities in 1. and 2. in the statistical setting (4) (see Section 1.4).

THEOREM 1.4 (Stable recovery of weights and source functions). *Let* $g = \omega^\top f$, $\tilde{g} = \tilde{\omega}^\top \tilde{f}$ *be two mixtures in* $\mathcal{M}^\delta$ *for some* $\delta > 0$ *and let* ε *be such that* $0 < \varepsilon < \delta(a_2 - a_1)/(2m(a_k - a_1))$. *If*

$$\sup_{x \in [0,1)} \left| g(x) - \tilde{g}(x) \right| < \varepsilon, \tag{10}$$

1. *then the weights satisfy the stable approximate recovery* (*SAR*) *property* $\max_{i=1,\dots,m} |\omega_i - \tilde{\omega}_i| < \varepsilon/(a_2 - a_1)$ *and*
2. *the sources satisfy the stable exact recovery* (*SER*) *property* $f = \tilde{f}$.

For a proof see Section S1.1 in the Supplementary Material [5].

1.3. *Methodology*: *First approaches*. In order to motivate our (quite involved) methodology, let us discuss briefly some attempts which may come to mind at a first glance. As a first approach to estimate ω and f from the data Y in the SBSSR-model one might pre-estimate the mixture g with some standard c.p. procedure, ignoring its underlying mixture structure, and then try to reconstruct ω and f afterwards. One problem is that the resulting step function cannot be decomposed into mixing weights $\omega \in \Omega(m)$ and source function $f \in \mathcal{S}^m(\mathfrak{A})$, in general, as the given alphabet $\mathfrak{A}$ leads to restrictions on the function values of g. But already for the initial step of reconstructing the mixture g itself, a standard c.p. estimation procedure (which does ignore the mixture structure) is unfavorable as it discards important information on the possible function values of g (induced by $\mathfrak{A}$). For example, if g has a small jump in some region, this might be easily missed (see Figure 2 for an example). Consequently, subsequent estimation of f and ω will fail as well. In contrast, a procedure which takes the mixture structure explicitly into account right from its beginning is expected to have better detection power for

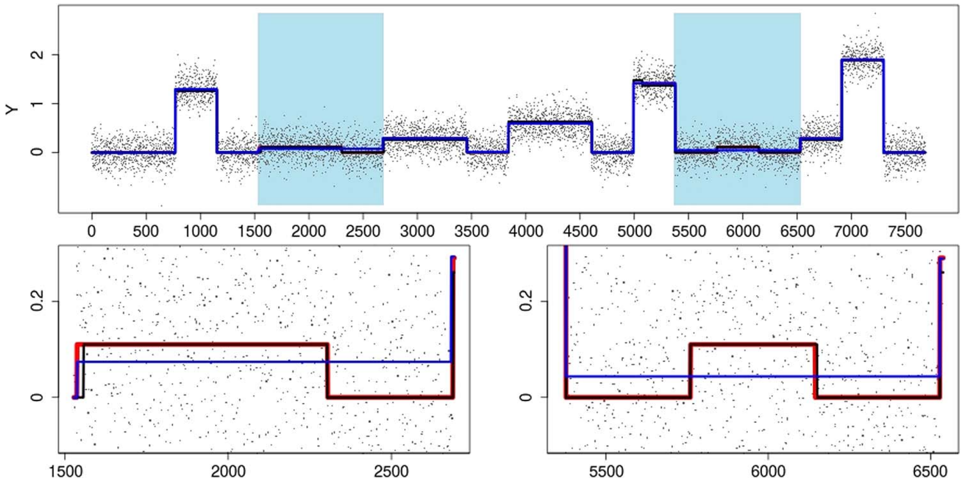

FIG. 2. *Observations Y from Example* 1.1 (*gray dots*), *together with the true underlying mixture g* (*red line*). *The blue line shows the c.p. estimate from* [32], *which does not incorporate the mixture structure. The red line shows the estimate with the proposed method* (*see Figure* 4 *for the underlying recovery of ω and the sources f*). *The blue areas display a region where g has a small jump* (*red line*), *which is not detected by the c.p. estimator* [32] (*blue line*), *but by the proposed method* (*black line*). *The bottom plots show a zoom in of the blue regions.*

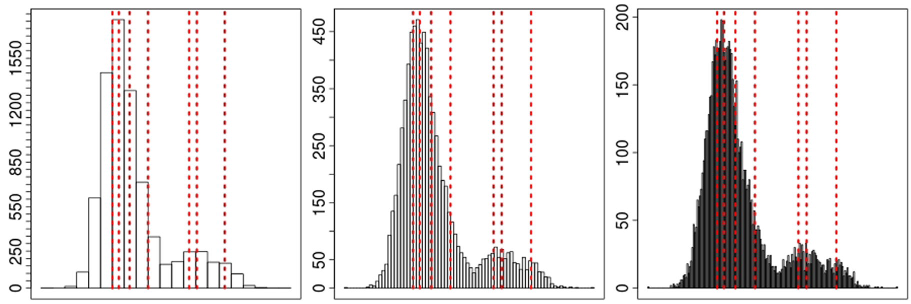

FIG. 3. *Histogram of the data from Example* 1.1 *with* 20, 100, *and* 200 *equidistant bins*, *respectively* (*from left to right*). *The vertical red lines indicate the true function values* (*modes*) *of g which have to be identified.*

a jump. As a conclusion, considering the SBSSR-model as a standard c.p. model discards important information and does not allow for demixing, in general.

A second approach which comes to mind is to first use some clustering algorithm to pre-estimate the function values of g, ignoring its serial c.p. structure, and infer the mixing weights ω from this. This pre-clustering approach has been pursued in several papers [22, 37, 72] for the particular case of a binary alphabet, that is, $k = 2$. However, as the number of possible function values of g equals k^m (recall that k is the size of the alphabet and m is the number of sources), recovery of these values in a statistical context by clustering is a difficult task in general, as it amounts to estimate the location of (at most) k^m modes correctly from the marginal distributions of the observations Y_j. In fact, this corresponds to mode hunting (see, e.g., [16, 30, 46, 53, 55, 67]) with potentially large number of modes which is known to be a hard problem. We illustrate the difficulty of this in Figure 3 employing histograms of the Y_j's in Example 1.1 with different bin widths. From this, it becomes obvious that a pre-clustering approach is not feasible for the present data.

Summing up, ignoring either of both, the c.p. and the finite alphabet mixture structure, in a first pre-estimation step discards important information which is indispensable for statistically efficient recovery. We emphasize that we are not aware of any existing method taking both aspects into account, in contrast to the method presented in this paper (SLAM), which will be briefly described now.

1.4. *Separate Linear Alphabet Mixtures* (*SLAM*).

In a first step, we will construct a confidence region $\mathcal{C}_{1-\alpha}$ for the weights ω which can be characterized by the acceptance region of a specific multiscale test with test statistic T_n, which is particularly well suited to capture both, the c.p. and the mixture structure, of g. The confidence level is determined by a threshold $q_n(\alpha)$ such that for any $g = \sum_{i=1}^m \omega_i f^i \in \mathcal{M}^\delta$

$$\{\omega \in \mathcal{C}_{1-\alpha}(Y)\} \supseteq \{T_n \leq q_n(\alpha)\}. \tag{11}$$

In a second step, we estimate f based on a multiscale constraint again. In the following section, we will introduce this procedure in more detail. We stress that the multiscale approach underlying SLAM is crucial for valid recovery of sources and mixing weights as the jumps potentially can occur at any location and any scale (i.e., interval length of neighboring sampling points).

1.4.1. *Multiscale statistic and confidence boxes underlying SLAM.* As the jump locations may occur at any place, a well established way for inferring the function values of g is to use local log-likelihood ratio test statistics in a multiscale fashion (see, e.g., [20, 29, 30, 32, 63]). Let $g|_{[x_i,x_j]} \equiv g_{ij}$ denote that g is constant on $[x_i, x_j]$ with function value g_{ij}. For the local testing problem on the interval $[x_i, x_j] \subset [0, 1)$ with some given value $g_{ij} \in \mathbb{R}$

$$H_0 : g|_{[x_i,x_j]} \equiv g_{ij} \quad \text{vs.} \quad H_1 : g|_{[x_i,x_j]} \not\equiv g_{ij} \tag{12}$$

the local log-likelihood ratio test statistic is

$$T_i^j(Y_i, \dots, Y_j, g_{ij}) = \ln\left(\frac{\sup_{\theta\in\mathbb{R}} \prod_{l=i}^j \phi_\theta(Y_l)}{\prod_{l=i}^j \phi_{g_{ij}}(Y_l)}\right) = \frac{(\sum_{l=i}^j Y_l - g_{ij})^2}{2\sigma^2(j-i+1)}, \tag{13}$$

where ϕ_θ denotes the density of the normal distribution with mean θ and variance σ^2. We then combine the local testing problems in (12) and define in our context the multiscale statistic T_n for some candidate function $\tilde{g}$ (which may depend on Y) as

$$T_n(Y, \tilde{g}) := \max_{\substack{1\le i\le j\le n \\ \tilde{g}|_{[x_i,x_j]}\equiv \tilde{g}_{ij}}} \frac{|\sum_{l=i}^j Y_l - \tilde{g}_{ij}|}{\sigma\sqrt{j-i+1}} - \text{pen}(j-i+1), \tag{14}$$

where $\text{pen}(j-i+1) := \sqrt{2(\ln(n/(j-i+1))+1)}$. The maximum in (14) is understood to be taken only over those intervals $[x_i, x_j]$ on which $\tilde{g}$ is constant with value $\tilde{g}_{ij} = \tilde{g}(x_i)$. The function values of $\tilde{g}$ determine the local testing problems (the value g_{ij} in (12)) on the single scales $[x_i, x_j]$. The calibration term $\text{pen}(\cdot)$ serves as a balancing of the different scales in a way that the maximum in (14) is equally likely attained on all scales (see [29, 32]). Other scale penalizations can be employed as well (see, e.g., [70]), but, for the ease of brevity, will not be discussed here. With the notation $\bar{Y}_i^j := \sum_{l=i}^j Y_l/(j-i+1)$, the statistic $T_n(Y, \tilde{g})$ in (14) has the following geometric interpretation:

$$T_n(Y, \tilde{g}) \le q \quad \Leftrightarrow \quad \tilde{g}_{ij} \in B(i, j) \qquad \forall 1 \le i \le j \le n \text{ with } \tilde{g}|_{[x_i,x_j]} \equiv \tilde{g}_{ij}, \tag{15}$$

for $q \in \mathbb{R}$, with intervals

$$B(i, j) := \left[\overline{Y}_i^j - \frac{q + \text{pen}(j-i+1)}{\sqrt{j-i+1/\sigma}}, \overline{Y}_i^j + \frac{q + \text{pen}(j-i+1)}{\sqrt{j-i+1/\sigma}}\right]. \tag{16}$$

In the following, we will make use of the fact that the distribution of $T_n(Y, g)$, with $g \in \mathcal{M}^\delta$ [see (9)] the true signal from the SBSSR-model, can be bounded from above with that of $T_n = T_n(Y, 0)$. It is known that $T_n \overset{\mathcal{D}}{\Rightarrow} L(\mathbb{B}) < \infty$ a.s. as $n \to \infty$, a certain functional of the Brownian motion $\mathbb{B}$ (see [28, 29]). Note that the distribution of $T_n(Y, 0)$ does not depend on the (unknown) f and ω anymore. As this distribution is not explicitly accessible and to be more accurate for small n (≤ 5000 say) the finite sample distribution of T_n can be easily obtained by Monte Carlo simulations. From this, one obtains $q_n(\alpha)$, $\alpha \in (0, 1)$, the $1 - \alpha$ quantile of T_n. We then obtain

$$\inf_{g \in \mathcal{M}^\delta} \mathbf{P}\big(T_n(Y, g) \leq q_n(\alpha)\big) \geq 1 - \alpha. \tag{17}$$

Hence, for the intervals in (16) with $q = q_n(\alpha)$ it follows that for all $g \in \mathcal{M}^\delta$

$$\mathbf{P}\big(g_{ij} \in B(i, j)\ \forall 1 \leq i \leq j \leq n \text{ with } g|_{[x_i, x_j]} \equiv g_{ij}\big) \geq 1 - \alpha. \tag{18}$$

In the following, we use the notation $B(i, j)$ for both, the intervals in (16) and the corresponding boxes $[i, j] \times B(i, j)$.

1.4.2. *Inference about the weights.* We will use now the system of boxes $\mathfrak{B} := \{B(i, j) : 1 \leq i \leq j \leq n\}$ from (16) with $q = q_n(\alpha)$ as in (17) to construct a confidence region $\mathcal{C}_{1-\alpha}$ for ω such that (11) holds, which ensures

$$\inf_{g \in \mathcal{M}^\delta} \mathbf{P}(\omega \in \mathcal{C}_{1-\alpha}) \geq 1 - \alpha. \tag{19}$$

More precisely, we will show that a certain element $B^\star \in \mathfrak{B}^m$ (denoted as the space of m-boxes) directly provides a confidence set $\mathcal{C}^\star_{1-\alpha} = A^{-1}B^\star$ for ω, with A as in (8). As $B^\star$ cannot be determined directly, we will construct a covering, $\mathfrak{B}^\star \ni B^\star$, of it such that the resulting confidence set

$$\mathcal{C}_{1-\alpha} = \bigcup_{B \in \mathfrak{B}^\star} A^{-1}B \tag{20}$$

has minimal volume (up to a log-factor) (see Section 2.4). The construction of $\mathfrak{B}^\star$ is done by applying certain reduction rules on the set $\mathfrak{B}^m$ reducing it to a smaller set $\mathfrak{B}^\star \subset \mathfrak{B}^m$ with $B^\star \in \mathfrak{B}^\star$. This is summarized in the CRW (confidence region for the weights) algorithm in Section 2.1 (and Section S2.1 in the Supplementary Material [5], respectively), which constitutes the first part of SLAM.

In Example 1.1 for $\alpha = 0.1$ this gives $\mathcal{C}_{0.9} = [0.00, 0.33] \times [0.07, 0.41] \times [0.39, 0.71]$ as a confidence box for $\omega = (\omega_1, \omega_2, \omega_3)^\top$ which covers the true value $\omega = (0.11, 0.29, 0.60)^\top$ in this case.

As the boxes $B(i, j)$ from (16) are constructed in a symmetric way, SLAM now simply estimates ω by

$$\hat{\omega} = \frac{1}{\sum_{i=1}^m (\underline{\omega}_i + \overline{\omega}_i)} (\underline{\omega}_1 + \overline{\omega}_1, \ldots, \underline{\omega}_m + \overline{\omega}_m), \tag{21}$$

with $\mathcal{C}_{1-\alpha} =: [\underline{\omega}_1, \overline{\omega}_1] \times \cdots \times [\underline{\omega}_m, \overline{\omega}_m]$. In Example 1.1, (21) gives for $\alpha = 0.1$ $\hat{\omega} = (0.17, 0.25, 0.58)^\top$.

For $D \subset \mathbb{R}^m$ and $d \in \mathbb{R}^m$ define the maximal distance

$$\overline{\operatorname{dist}}(d, D) := \sup_{\tilde{d} \in D} \|d - \tilde{d}\|_\infty. \tag{22}$$

Further, and for all following considerations, define

$$\alpha_n = \exp(-c_1 \ln^2(n)) \quad \text{and} \quad \beta_n = \exp\biggl(-75m^2 \frac{(a_k - a_1)^2}{(a_2 - a_1)^2} c_1 \ln^2(n)\biggr), \tag{23}$$

for some constant c_1, to be specified later, see (40). Denote the minimal distance between any two jumps of $g \in \mathcal{M}^\delta$ (and hence of the f^i's, recall the discussion in Section 1.2) as λ. Then, in addition to uniform coverage in (19) for $\alpha = \alpha_n$ in (23), we will show that the confidence region $\mathcal{C}_{1-\alpha}$ from (20) covers the unknown weight vector ω with maximal distance shrinking of order $\ln(n)/\sqrt{n}$ with probability tending to one at a superpolynomial rate,

$$\mathbf{P}\biggl(\overline{\operatorname{dist}}(\omega, \mathcal{C}_{1-\alpha_n}(Y)) < \frac{c_2}{a_2 - a_1} \frac{\ln(n)}{\sqrt{n}}\biggr) \geq 1 - \exp(-c_1 \ln^2(n))$$

for all $n \geq N^\star$, for some constants $c_1 = c_1(\delta)$, $c_2 = c_2(\lambda, \delta)$ and some explicit $N^\star = N^\star(\lambda, \delta) \in \mathbb{N}$ (see Corollary 2.8).

1.4.3. *Inference about the source functions.* Once the mixing weights ω have been estimated by $\hat{\omega}$ [see (21)], SLAM estimates $f^1, \ldots, f^m$ in two steps. First, the number of c.p.'s $K(g)$ of $g = \omega^\top f \in \mathcal{M}^\delta$ will be estimated by solving the constrained optimization problem

$$\hat{K} := \min_{\tilde{g} \in \mathcal{M}(\mathfrak{A}, \hat{\omega})} K(\tilde{g}) \quad \text{s.t. } T_n(Y, \tilde{g}) \leq q_n(\beta). \tag{24}$$

Here, the multiscale constraint on the r.h.s. of (24) is the same as for $\mathcal{C}_{1-\alpha}(Y)$ in (11), but with a possibly different confidence level $1 - \beta$. Finally, we estimate $f^1, \ldots, f^m$ as the constrained maximum likelihood estimator

$$\hat{f} = (\hat{f}^1, \ldots, \hat{f}^m)^\top := \operatorname*{argmax}_{\tilde{f} \in \mathcal{H}(\beta)} \sum_{i=1}^n \ln(\phi_{\hat{\omega}^\top \tilde{f}(x_i)}(Y_i)), \tag{25}$$

with (see Section 2.2)

$$\mathcal{H}(\beta) := \{\tilde{f} \in \mathcal{S}(\mathfrak{A})^m : T_n(Y, \hat{\omega}^\top \tilde{f}) \leq q_n(\beta) \text{ and } K(\hat{\omega}^\top \tilde{f}) = \hat{K}\}. \tag{26}$$

Choosing $\alpha = \alpha_n$ and $\beta = \beta_n$ as in (23), in Section 2.4 (see Theorem 2.7) we show that with probability at least $1 - \alpha_n$, for n large enough, the SLAM estimator $\hat{f}$ in (25) estimates for all $i = 1, \ldots, m$:

1. the respective number of c.p.'s $K(f^i)$ correctly,
2. all c.p. locations with rate $\ln^2(n)/n$ simultaneously, and
3. the function values of f^i exactly (up to the uncertainty in the c.p. locations).

Obviously, the rate in 2. is optimal up to possible log-factors as the sampling rate is $1/n$. From Theorem 2.7, it follows further (see Remark 2.9) that the minimax detection rates are even achieved (again up to possible log-factors) when $\delta, \lambda \to 0$ (as $n \to \infty$).

Further, we will show that a slight modification $\tilde{\mathcal{H}}(\beta)$ of $\mathcal{H}(\beta)$ in (26) constitutes an asymptotically uniform (for given ASB δ and λ) multivariate confidence band for the source functions $(f^1, \dots, f^m)$ (see Section 2.3).

To illustrate, Figure 4 depicts SLAM's estimates of the mixture $\hat{g} = \hat{\omega}^\top \hat{f}$, with $\hat{\omega} = (0.11, 0.26, 0.63)^\top$, and the source functions $\hat{f}^1$, $\hat{f}^2$, $\hat{f}^3$ from (25) with Y as in Example 1.1, $\beta = 0.01$ (corresponding to $q_n(\beta) = 2.1$), and an automatic choice of α, the MVT-selection method explained in Section 4.6. In order to visualize $\tilde{\mathcal{H}}(\beta)$, we illustrate the provided confidence in gray scale encoding the projections of $\tilde{\mathcal{H}}(\beta)$ (recall the alphabet $\mathfrak{A} = \{0, 1, 2\}$).

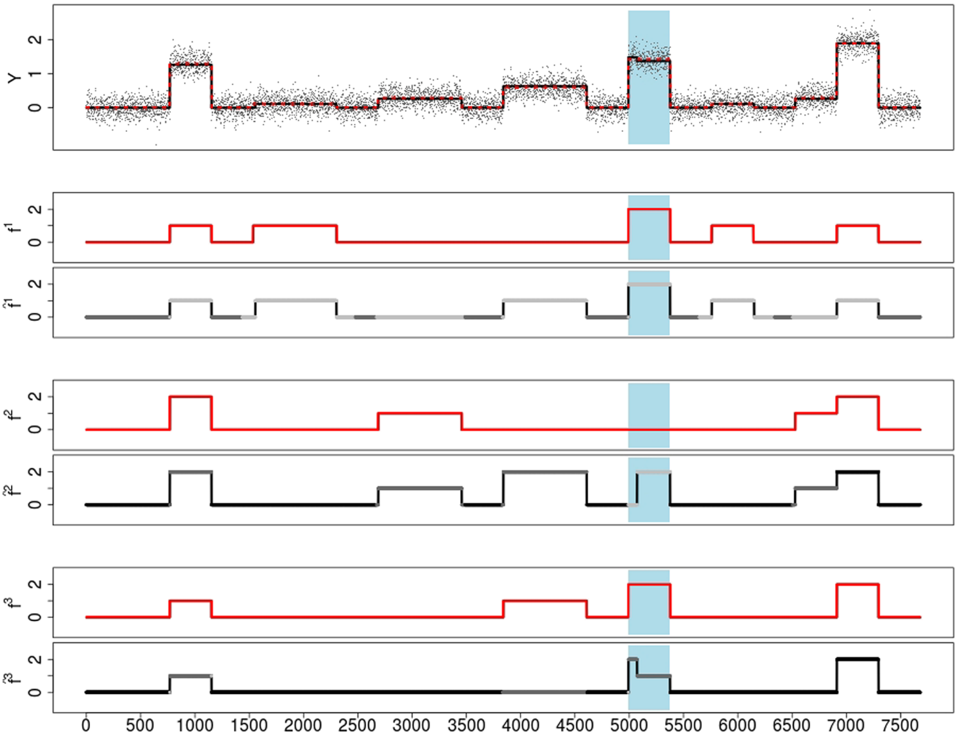

FIG. 4. *First row*: g (*red dotted line*), $\hat{g}$ (*black line*) *with* $\hat{\omega} = (0.11, 0.26, 0.63)^\top$, *and data* Y (*gray*) *from Example* 1.1. *Subsequent rows*: f^i (*red line*) *and SLAM's estimate* $\hat{f}^i$ (*gray/black line*) *for* $q_n(\alpha) = 0.2$ *and* $q_n(\beta) = 2.1$ (*see Section* 4.6). *Gray shades for segments of* $\hat{f}^i$ *indicate the confidence for the given segment*: *a maximal deviation of two* (*light gray*), *one* (*gray*), *and no deviation* (*black*) *at confidence level* $\beta = 0.01$. *The blue area displays a constant region of* g *where* $\hat{g}$ *includes a* (*wrong*) *jump and its effect on estimation of the sources.*

1.5. *Algorithms and software.* SLAM's estimate for ω (see (21) and Algorithm CRW, in Section S2.1 in the Supplementary Material [5]) can be computed with polynomial complexity between $\mathcal{O}(n^m)$ and $\mathcal{O}(n^{2m})$ (see Section 3). Using dynamic programming, the final estimate of sources can then be computed with a complexity ranging from $\mathcal{O}(n)$ and $\mathcal{O}(n^2)$ depending on the final solution (see Section 3 for details). An R-package including an implementation of SLAM is available on request.

1.6. *Simulations.* The performance of SLAM is investigated in a simulation study in Section 4. We first investigate accuracy of $\hat{\omega}$ and the confidence region $\mathcal{C}_{1-\alpha}(Y)$ as in (21) and (20). We found always higher coverage of $\mathcal{C}_{1-\alpha}(Y)$ than the nominal confidence level $1-\alpha$. In line with this, $\hat{\omega}$ appeared to be very stable under the choice of the confidence level α. Second, we investigate SLAM's estimates $\hat{f}$. A major conclusion is that if g is not well estimated in a certain region, this typically will influence the quality of the estimates of f^i in this region but not beyond (see the marked light blue region in Figure 4 where the estimator $\hat{g}$ includes a wrong jump in a constant region of g but this error does not propagate serially). This may be explained by the flexible c.p. model $\mathcal{M}^\delta$ together with the multiscale nature of SLAM, which locally "repairs" estimation errors. Finally, in Section 4.6 we comment on practical choices for α and β complementing the theoretically motivated choices in (23). To this end, we suggest a data driven selection method for α when it is considered as tuning parameter for the accuracy of the estimate $\hat{\omega}$ and $\hat{f}$ rather than a confidence level for the coverage of ω.

1.7. *Application to cancer genetics.* Blind source separation in the context of the SBSSR-model occurs in different areas, for example, in digital communications and signal transmission. The main motivation for our work comes from cancer genetics, in particular from the problem to assign copy-number aberrations (CNAs) in cell samples taken from tumors (see [48]) to its clones. CNAs refer to stretches of DNA in the genomes of cancer cells which are under copy-number variation involving deletion or duplication of stretches of DNA relative to the inherited (germline) state present in normal tissues. CNAs are known to be key drivers of tumor progression through the deletion of "tumor suppressing" genes and the duplication of genes involved in processes such as cell signaling and division. Understanding where, when and how CNAs occur during tumourgenesis, and their consequences, is a highly active and important area of cancer research (see e.g., [8]). Modern high-throughput technologies allow for routine whole genome DNA sequencing of cancer samples and major international efforts are underway to characterize the genetic make up of all cancers, for example, The Cancer Genome Atlas, http://cancergenome.nih.gov/.

A key component of complexity in cancer genetics is the "clonal" structure of many tumors, which relates to the fact that tumors usually contain distinct cell populations of genetic subtypes (clones) each with a distinct CNA profile (see, e.g., [36, 61]). High-throughput sequencing technologies act by bulk measurement of large numbers of pooled cells in a single sample, extracted by a micro-dissection biopsy or blood sample for haematological cancers.

The copy-number, that is, the number of copies of DNA stretches at a certain locus, of a single clone's genome is a step function mapping chromosomal loci to a value $i \in \{0, \dots, k\}$ corresponding to i copies of DNA at a locus, with reasonable biological knowledge of k (in our example $k = 5$; see Section 5).

From the linear properties of the measurement technologies the relative amount of DNA measured at any loci is therefore a mixture of step functions, with mixture weights given by the relative proportion of the clone's DNA in the pool. The estimation of the mixed function, that is, estimating the locations of varying overall copy numbers, has perceived considerable interest in the past (see [15, 27, 32, 41, 51, 52, 68, 73]). However, the corresponding demixing problem, that is, jointly estimating the number of clones, their proportion, and their CNAs, has only perceived more recently as an important issue and hence received very little attention in a statistical content so far and is the main motivation for this work.

In Section 5, we illustrate SLAM's ability to recover the CNA's of such clones by utilizing it on real genetic sequencing data. On hand of a special data set, with measurements not only for the mixture but also for the underlying source functions (clones) and with knowledge about the mixing weights, we are able to report on the accuracy of SLAM's estimates of the corresponding CNA profile and the mixing proportion of the clones.

1.8. *Related work.* Each BSS of finite alphabet sources (see, e.g., [10, 22, 37, 43, 45, 47, 54, 59, 72]) and the estimation of step functions, with unknown number and location of c.p.'s (see, e.g., [13, 27, 31–34, 40–42, 49, 51, 52, 62, 65, 68, 74]), are widely discussed problems. However, the combination of both, as discussed in this paper, is not. Rigorous statistical methodology and theory for finite alphabet BSS problems is entirely lacking to best of our knowledge and we are not aware of any other method which provides estimates for and confidence statements in the SBSSR-model in such a rigorous and general way. There are, however, related problems, discussed in the following.

Rewriting the SBSSR-model (4) in matrix form $Y = F\omega + \varepsilon$ with $F = (f^i(x_j))_{1 \le j \le n, 1 \le i \le m}$ shows some commonality to signal recovery in linear models. In fact, our Theorem 1.4 reveals some analogy to exact and stable recovery results in compressive sensing and related problems (see [12, 25]). We stress, however, that there are fundamental differences. There typically the systems matrix F is known and ω is a sparse vector to be recovered, having only a very few non null coefficients. Under an additional finite alphabet assumption (for known F) recovery of ω is for example, addressed in [1, 9, 18, 26]. In our setting both, F and ω are unknown.

Another related problem is nonnegative matrix factorization (NMF) (see, e.g., [2, 24, 44, 57]), where one assumes a multivariate signal $Y \in \mathbb{R}^{n \times M}$ resulting from

M different (unknown) mixing vectors, that is, $\omega \in \mathbb{R}_+^{m\times M}$, and an (unknown) nonnegative source matrix $F \in \mathbb{R}_+^{n\times m}$. There, a fundamental assumption is that $m \ll \min(n, M)$, which obviously does not hold in our case where $M = 1$. Instead we employ the additional assumption of a known finite alphabet, that is, $F \in \mathfrak{A}^{n\times m}$. Indeed, techniques and algorithms for NMF are quite different from the ones derived here, as our methodology explicitly takes advantage of the one dimensional (i.e., ordered) c.p. structure under the finite alphabet assumption.

However, the identifiability conditions (6) and (7) from Section 1.2 are similar in nature to identifiability conditions for the NMF problem [2, 24], from where the notation "separable" originates. In order to ensure identifiability in the NMF problem, the "α-robust simplicial" condition (see, e.g., [57], Definition 2.1) on the mixing matrix $\omega \in \mathbb{R}_+^{m\times M}$ and the "separability" condition (see, e.g., [57], Definition 2.2) on the source matrix $F \in \mathbb{R}_+^{n\times m}$ are well established [2, 24, 57].

There, the "α-robust simplicial" condition assumes that the mixing vectors $\omega_{1\cdot}, \ldots, \omega_{m\cdot} \in \mathbb{R}_+^M$ constitute vertices of an m-simplex with minimal diameter (distance between any vertex and the convex hull of the remaining vertices) α. This means that different source values $F_{i\cdot} \in \mathbb{R}^m$ are mapped to different mixture values $F_{i\cdot}\omega \in \mathbb{R}_+^M$ by the mixing matrix $\omega \in \mathbb{R}_+^{m\times M}$. This condition is analog to the condition $\mathrm{ASB}(\omega) \geq \delta$ in (6), which also ensures that different source values $f(x) \in \mathfrak{A}^m$ are mapped to different mixture values $\omega^\top f(x) \in \mathbb{R}$ via the mixing weights $\omega \in \Omega(m)$, with minimal distance δ between different mixture values.

The "separability" condition in NMF is the same as in Definition 1.3 but with A replaced by the identitiy matrix (recall that in NMF the sources can take any positive value in $\mathbb{R}_+$, in contrast to the SBSSR-model where the sources can only take values in a given alphabet $\mathfrak{A}$) and the intervals $I_r \subset (0, 1]$ are replaced by measurement points $i_r \in \{1, \ldots, n\}$ (recall that the SBSSR-model considers a change-point regression setting, in contrast to NMF where observations do not necessarily come from discrete measurements of an underlying regression function). In both models (NMF and SBSSR), the separability condition ensures a certain variability of the sources in order to guarantee identitfiability of the mixing matrix and vector, respectively, from their mixture.

Besides NMF, there are many other matrix-factorization problems, which aim to decompose a multivariate signal $Y \in \mathbb{R}^{n\times M}$ in two matrices of dimension $n \times m$ and $m \times M$, respectively. A popular example is independent component analysis (ICA) (see, e.g., [3, 7, 17]), which exploits statistical independence of the m different sources. We stress that this approach becomes infeasible in our setting where $M = 1$ as the error terms then sum up to a single error term and ICA would treat this as one observation. Other matrix-factorization methods assume a certain sparsity of the mixing-matrix [64]. Similar to NMF methods, in general all these methods, however, again rely on the assumption that $M > 1$ (most of them even require $M \geq m$) as otherwise the signal is not even identifiable, in contrast to our situation again due to the finite alphabet.

Minimization of the ℓ_0 norm using dynamic programming (which has a long history in c.p. analysis; see, e.g., [4, 31, 33, 42]) for segment estimation under a multiscale constraint has been introduced in [11] (see also [19] and [32]) and here we extend this to mixtures of segment signals and in particular to a finite alphabet restriction.

The SBSSR problem becomes tractable as we assume that our signals occur with sufficiently many alphabet combinations which may be present already on small scales on the one hand, and on the other hand we also observe long enough segments (large scales) in order to estimate reliably the corresponding mixing weights on these [see the identifiability condition in (7)]. Both assumptions seem to be satisfied in our motivating application, the separation of clonal copy numbers in a tumor.

To the best of our knowledge, the way we treat the problem of clonal separation is new; see, however, [14, 23, 38, 48, 60, 71]. Methods suggested there, all rely on specific prior information about the sources f and cannot be applied to the general SBSSR-model. Moreover, most of them treat the problem from a Bayesian perspective.

2. Method and theory.

2.1. *Confidence region for the weights.* Let Y and $g = \omega^\top f \in \mathcal{M}^\delta$ be as in the SBSSR-model (4). Our starting point for the recovery of the weights ω and the sources f is the construction of proper confidence sets for ω which is also of statistical relevance by its own as the source functions are unknown which hinders direct inversion of a confidence set for g.

Consider the system of boxes $\mathfrak{B} = \{B(i, j) : 1 \leq i \leq j \leq n\}$ from (16) with $q = q_n(\alpha)$ as in (17) for some given $\alpha \in (0, 1)$, as described in Section 1.4.1.

As the underlying sources f are assumed to be separable [see Definition 1.3 and (9)] there exist intervals $[x_{i_r^\star}, x_{j_r^\star}] \subset (0, 1]$, for $r = 1, \dots, m$, such that

$$f|_{[x_{i_r^\star}, x_{j_r^\star}]} \equiv [A]_r, \tag{27}$$

with A as in (8). Assume for the moment that these intervals would be known and let $B^\star := B(i_1^\star, j_1^\star) \times \cdots \times B(i_m^\star, j_m^\star) \in \mathfrak{B}^m$ be the corresponding m-box. Then a $1 - \alpha$ confidence region for ω is given as

$$\mathcal{C}_{1-\alpha}(i_1^\star, j_1^\star, \dots, i_m^\star, j_m^\star) := A^{-1} B^\star. \tag{28}$$

To see that (28) is, indeed, a $1 - \alpha$ confidence region for ω, note that

$$\{\omega \in \mathcal{C}_{1-\alpha}(i_1^\star, j_1^\star, \dots, i_m^\star, j_m^\star)\} \supset \bigcap_{1 \leq r \leq m} \{g|_{[x_{i_r^\star}, x_{j_r^\star}]} \equiv \omega^\top [A]_r \in B(i_r^\star, j_r^\star)\}$$

and

$$\{T_n(Y, g) \leq q_n(\alpha)\} = \bigcap_{\substack{1 \leq i \leq j \leq n \\ g|_{[x_i, x_j]} \equiv g_{ij}}} \{g_{ij} \in B(i, j)\}.$$

This implies that

$$\{\omega \in \mathcal{C}_{1-\alpha}(i_1^\star, j_1^\star, \dots, i_m^\star, j_m^\star)\} \supset \{T_n(Y, g) \le q_n(\alpha)\} \tag{29}$$

and therefore it holds uniformly in $g \in \mathcal{M}^\delta$ that

$$\mathbf{P}(\omega \in \mathcal{C}_{1-\alpha}(i_1^\star, j_1^\star, \dots, i_m^\star, j_m^\star)) \ge \mathbf{P}(T_n(Y, g) \le q_n(\alpha)) \ge 1 - \alpha. \tag{30}$$

Of course, as the source functions f are unknown, intervals $[x_{i_r^\star}, x_{j_r^\star}]$ which satisfy (27) are not available immediately, and thus, one cannot construct the m-box $B^\star$ required for (28) directly.

For this reason, we will describe a strategy to obtain a subsystem of m-boxes, that is, a subset $\mathfrak{B}^\star \subset \mathfrak{B}^m$, which covers $B^\star$ conditioned on $\{T_n(Y, g) \le q_n(\alpha)\}$ almost surely. To this end, observe that for any random set $\mathcal{C}^\star(Y) \subset \mathbb{R}^m$ with

$$\mathbf{P}(\mathcal{C}^\star(Y) \supset \mathcal{C}_{1-\alpha}(i_1^\star, j_1^\star, \dots, i_m^\star, j_m^\star) | T_n(Y, g) \le q_n(\alpha)) = 1 \tag{31}$$

(29) and (30) imply $\mathbf{P}(\omega \in \mathcal{C}^\star(Y)) \ge 1 - \alpha$. We then define $\mathcal{C}_{1-\alpha}$ as in (20). To this end, $\mathfrak{B}^\star$ is constructed such that the diameter of the resulting $\mathcal{C}_{1-\alpha}$ is of order $\ln(n)/\sqrt{n}$ (see Corollary 2.8). The construction will be done explicitly by an algorithm which relies on the application of certain reduction rules to $\mathfrak{B}^m$ to be described in the following.

Let $\mathrm{proj}_r : \mathfrak{B}^m \to \mathfrak{B}$, for $r = 1, \dots, m$, denote the rth projection [i.e., $\mathrm{proj}_r(B_1 \times \cdots \times B_m) := B_r$] and define the set of boxes on which any signal fulfilling the multiscale constraint is non constant (nc) as

$$\mathfrak{B}_{\mathrm{nc}} := \{B(i, j) \in \mathfrak{B} : \exists [s, t], [u, v] \subset [i, j] \text{ with } B(s, t) \cap B(u, v) = \varnothing\}. \tag{32}$$

R1. Delete $B \in \mathfrak{B}^m$ if there exists an $r \in \{1, \dots, m\}$ such that $B(i, j) := \mathrm{proj}_r(B) \in \mathfrak{B}_{\mathrm{nc}}$ as in (32).

The reasoning behind R1 is as follows. $g|_{[x_{i_r^\star}, x_{j_r^\star}]}$ is constant for $r = 1, \dots, m$ as $f^1, \dots, f^m$ are constant on $[x_{i_r^\star}, x_{j_r^\star}]$. Consequently, all m-boxes that include a box $B(i, j) \in \mathfrak{B}$ such that g cannot be constant on $[x_i, x_j]$ [conditioned on $T_n(Y, g) \le q_n(\alpha)$] can be deleted in order to preserve coverage of $B^\star$. Let $[x_i, x_j]$ be an interval on which g is constant (say $g|_{[x_i, x_j]} \equiv c$) and assume that there exist intervals $[s, t], [u, v] \subset [i, j]$ such that $B(s, t) \cap B(u, v) = \varnothing$. Then by construction of the boxes $B(s, t)$ and $B(u, v)$, $T_n(Y, g) \le q_n(\alpha)$ implies that $c \in B(s, t)$ and $c \in B(u, v)$, which contradicts $B(s, t) \cap B(u, v) = \varnothing$. In other words, $\mathfrak{B}_{\mathrm{nc}}$ (nc $\hat{=}$ non constant) in (32) includes all boxes $B(i, j)$ such that all function $\tilde{g} \in \mathcal{M}^\delta$ which fulfill the multiscale constraint $T_n(Y, \tilde{g}) \le q_n(\alpha)$ cannot be constant on $[x_i, x_j]$. Note that, in contrast to the following two reduction rules, the reduction rule R1 does not depend on the specific matrix A in the identifiablity condition in (7).

R2. Delete $B \in \mathfrak{B}^m$, with $[\underline{b}_r, \overline{b}_r] := \operatorname{proj}_r(B)$ if at least one of the following statements holds true:

1. $\overline{b}_1 \leq a_1$ or $\underline{b}_1 \geq a_1 + \frac{a_2 - a_1}{m}$,
2. for any $2 \leq r \leq m$
$$\frac{a_2 + (m-1)a_1 - \sum_{j=1}^{r-1} \underline{b}_j}{m - r + 1} \leq \underline{b}_r \quad \text{or} \quad \underline{b}_{r-1} \geq \overline{b}_r,$$
3. $\sum_{j=1}^{m} \overline{b}_j \leq a_2 + (m-1)a_1$.

R2 1. comes from the fact that $0 < \omega_1 < 1/m$, R2 2. from $\omega_{i-1} < \omega_i < (1 - \sum_{j=1}^{i-1} \omega_j)/(m - i + 1)$, and R2 3. from $\sum_{j=1}^{m} \omega_j = 1$, together with the specific choice of the matrix A in (8). For a different choice of A in (7) the equations in R2 can be modified accordingly.

In what follows, define for $k = 1, \dots, n$,

$$\mathcal{J}_k := \{[i, j] : k \in [i, j] \text{ and } B(i, j) \notin \mathfrak{B}_{\text{nc}}\}. \tag{33}$$

R3. Delete $B \in \mathfrak{B}^m$, if there exists a $k \in \{1, \dots, n\}$ such that for all $[i, j] \in \mathcal{J}_k$

$$\Big[\max_{i \leq u \leq v \leq j} \underline{b}_{uv}, \min_{i \leq u \leq v \leq j} \overline{b}_{uv}\Big] \cap \{\tilde{\omega}^\top a : a \in \mathfrak{A}^m \text{ and } \tilde{\omega} \in A^{-1}B\} \tag{34}$$

is empty, with $[\underline{b}_{uv}, \overline{b}_{uv}] := B(u, v) \in \mathfrak{B}$.

Conditioning on $T_n(Y, g) \leq q_n(\alpha)$ implies $\omega \in A^{-1}B^\star$, and, in particular, that there exists an $\tilde{\omega} \in A^{-1}B^\star$ such that $\operatorname{Im}(g) := \{g(x_1), \dots, g(x_n)\} \subset \{\tilde{\omega}^\top a : a \in \mathfrak{A}^m\}$. Moreover, for every $k \in \{1, \dots, n\}$ there exists an interval $[x_i, x_j]$ where g is constant with $g|_{[x_i, x_j]} \equiv g(x_k) \in \operatorname{Im}(g)$. So, $T_n(Y, g) \leq q_n(\alpha)$ implies $g(x_k) \in B(u, v)$ for all $[u, v] \subset [i, j]$ and, therefore, for $B = B^\star$ (34) is not empty [conditioned on $T_n(Y, g) \leq q_n(\alpha)$].

REMARK 2.1 (Incorporating prior knowledge on minimal scales).

(a) If we restrict to a minimal scale $\lambda \in (0, 1)$ on which a jump of g may occur, that is, for τ_j, $j = 0, \dots, K + 1$, being the c.p.'s of g,
$$\lambda := \min_{j \in \{0, \dots, K\}} |\tau_{j+1} - \tau_j| > 0, \tag{35}$$
we can modify R3 with $\mathcal{J}_k$ in (33) replaced by $\mathcal{J}_k \cap \{[i, j] : j - i + 1 \geq n\lambda\}$.

(b) In many applications (see Section 5), it is very reasonable to assume apriori knowledge of a minimal interval length $\lambda^\star$ of $[x_{i_r^\star}, x_{j_r^\star}]$ in (27). This means that there exists some interval $I_r \subset [0, 1)$ of minimum size $\lambda^\star$, where $(f^1, \dots, f^m)$ take the value $[A]_r$ as in (8) for $r = 1, \dots, m$. This is summarized in the following reduction rule.

R4. Knowing that $j_r^\star - i_r^\star + 1 \geq \lambda^\star n$ for $r = 1, \dots, m$ in (27), delete $B \in \mathfrak{B}^m$ if there exists an $r \in \{1, \dots, m\}$ such that for $B(i, j) := \mathrm{proj}_r(B)$ $j - i + 1 < \lambda^\star n$.

R1–R4 is summarized in Algorithm CRW, in Section S2.1 in the Supplementary Material [5], for constructing a confidence region for ω.

REMARK 2.2 (Noninformative m-box). If $\mathfrak{B}^\star = \varnothing$, we formally may set $\mathcal{C}_{1-\alpha} := \Omega(m)$, the trivial confidence region. As $\{\mathfrak{B}^\star = \varnothing\} \subset \{T_n(Y, g) > q_n(\alpha)\}$, the probability that this happens can be bounded from above by α. This is in general only a very rough bound, simulations show that $\mathfrak{B}^\star = \varnothing$ is hardly ever the case when α is reasonably small. For instance, in 10,000 simulations of Example 1.1 with $n = 1280$, $\sigma = 0.1$, $\alpha = 0.1$, it did not happen once. Of course, when $\alpha \nearrow 1$, $\mathfrak{B}^\star = \varnothing$ finally, as no mixture $g \in \mathcal{M}^\delta$ can fulfill the multiscale constrained $T_n(Y, g) \leq q$ for arbitrarily small q.

REMARK 2.3 (Shape of $\mathcal{C}_{1-\alpha}$). The previous construction of the confidence set $\mathcal{C}_{1-\alpha}$ does not ensure that the confidence set is of m-box form

$$[\underline{\omega}_1, \overline{\omega}_1] \times \cdots \times [\underline{\omega}_m, \overline{\omega}_m]. \tag{36}$$

In general it is a union of m-boxes. However, we can always take the smallest covering m-box of $\mathcal{C}_{1-\alpha}$, given by

$$\Big[\inf_{\tilde{\omega} \in \mathcal{C}_{1-\alpha}} \tilde{\omega}_1, \sup_{\tilde{\omega} \in \mathcal{C}_{1-\alpha}} \tilde{\omega}_1\Big] \times \cdots \times \Big[\inf_{\tilde{\omega} \in \mathcal{C}_{1-\alpha}} \tilde{\omega}_m, \sup_{\tilde{\omega} \in \mathcal{C}_{1-\alpha}} \tilde{\omega}_m\Big], \tag{37}$$

in order to get a confidence set as in (36). Note, that $\overline{\mathrm{dist}}(\omega, \mathcal{C}_{1-\alpha}) =: d$ remains the same when we replace $\mathcal{C}_{1-\alpha}$ by (37). To see this, consider $\hat{\mathcal{C}} := \omega + [-d, d]^m$, which is a covering m-box of $\mathcal{C}_{1-\alpha}$, so in particular $\hat{\mathcal{C}}$ covers (37), with $\overline{\mathrm{dist}}(\omega, \hat{\mathcal{C}}) = d$.

Summing up, we have now constructed a confidence set $\mathcal{C}_{1-\alpha}$ for the mixing vector ω in the SBSSR-model. Given $\mathcal{C}_{1-\alpha}$ SLAM estimates ω as in (21). From this, in the next section we derive estimators for the sources $f^1, \dots, f^m$.

2.2. *Estimation of source functions.* SLAM estimates $f = (f^1, \dots, f^m)$ by solving the constraint optimization problem (25), which admits a solution if and only if

$$\min_{\tilde{f} \in \mathcal{S}(\mathfrak{A})^m} T_n\big(Y, \hat{\omega}(\alpha)^\top \tilde{f}\big) \leq q_n(\beta). \tag{38}$$

(38) cannot be guaranteed in general but it can be shown that it holds asymptotically with probability one (see Theorem S1.1 in the Supplementary Material [5]), independently of the specific choice of $\hat{\omega} \in \mathcal{C}_{1-\alpha}(Y)$ in (21). For finite n our sim-

ulations show that violation of (38) is hardly ever the case. For instance, in 10,000 simulation runs of Example 1.1 with $\alpha = \beta = 0.1$ it did not happen once. Therefore, in practice, failure of (38) might rather indicate that the model assumption is not correct (e.g., due to outliers) and could be treated by pre-processing of the data. Another strategy can be to decrease β and hence the constraint in (38) as for $\beta > \beta'$ it holds that $q_n(\beta') > q_n(\beta)$.

REMARK 2.4 (Incorporating identifiability conditions in SLAM). The separability condition in (7) could be incorporated in the estimator (25), which provides a further restriction on $\mathcal{H}(\beta)$ in (26). This may yield a finite sample improvement of SLAM, however, at the expense of being less robust if such a particular identifiability condition is violated [see Section 4.5.1 for a simulation study of SLAM when the identifiability conditions in (6) and (7) are violated].

2.3. *Confidence bands for the source functions.* Obviously, uniform confidence sets for f cannot be obtained if we allow for an arbitrarily small distance between two c.p.'s of g (as for any c.p. problem, see [32]). However, if we restrict to a minimal scale λ as in (35), the SLAM estimation procedure in (25) leads to asymptotically uniform confidence bands for the source functions $f^1, \dots, f^m$. To this end, we introduce

$$\mathcal{M}^\delta_\lambda := \Big\{ g \in \mathcal{M}^\delta : \min_{j \in \{0,\dots,K(g)\}} |\tau_{j+1} - \tau_j| \geq \lambda \Big\}, \tag{39}$$

where, as in (1), $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K(g)} < \tau_{K(g)+1} = 1$ denote c.p.'s of g. Moreover, let $\tilde{T}_n$ be as in (14), but with $\text{pen}(j - i + 1)$ replaced by $\text{pen}(j - i + 1) + ((a_2 - a_1)\ln(n)/m + \sqrt{8\sigma^2 \ln(e/\lambda)/\lambda})\sqrt{(j - i + 1)/n}$ and let $\tilde{\mathcal{H}}(\beta)$ be as in (26) but with T_n replaced by $\tilde{T}_n$. Then $\tilde{\mathcal{H}}(\beta)$ constitutes an asymptotically uniform confidence band as the following theorem shows.

THEOREM 2.5. *Consider the SBSSR-model and let $\hat{\omega}$ be the SLAM estimator from (21) for $\alpha = \alpha_n$ as in (23). Then, for $\tilde{\mathcal{H}}(\beta)$ as in (26) with T_n replaced by $\tilde{T}_n$, $\tilde{\mathcal{H}}(\beta)$ provides an asymptotically uniform confidence region for the sources f,*

$$\lim_{n \longrightarrow \infty} \inf_{g \in \mathcal{M}^\delta_\lambda} \mathbf{P}\big((f^1, \dots, f^m) \in \tilde{\mathcal{H}}(\beta)\big) \geq 1 - \beta.$$

For a proof see Section S1.3 in the Supplementary Material [5].

2.4. *Consistency and rates.* In the following, we investigate further theoretical properties of SLAM. As in Theorem 2.5 our results will be stated uniformly over the space $\mathcal{M}^\delta_\lambda$ in (39), that is, for a given minimal length λ of the constant parts of the mixture g and a given minimal ASB δ as in (5). Define the constants

$$c_1 = \frac{\delta^2 (a_2 - a_1)^2}{48600 \sigma^2 m^2 (a_k - a_1)^2}, \qquad c_2 = \frac{\delta + \sqrt{2\sigma^2 \ln(e/\lambda)}}{\sqrt{\lambda}}. \tag{40}$$

Further, let $N^\star \in \mathbb{N}$ be the smallest integer, s.t.

$$\sqrt{\frac{2\ln(eN^\star/\ln^2(N^\star))}{\ln^2(N^\star)}} + \frac{\sqrt{6\ln(3e/\lambda)}}{\sqrt{N^\star\lambda}} \leq \frac{\delta}{4\sigma}, \quad \text{and} \tag{41}$$

$$\frac{\ln(N^\star)}{\sqrt{N^\star\lambda}} \leq \frac{\delta(a_2 - a_1)/(a_k - a_1)}{2m(\delta + \sqrt{2\sigma^2\ln(e/\lambda)})}. \tag{42}$$

REMARK 2.6 (Behavior of $N^\star$). Note that the left-hand side in (41) and (42) is decreasing in $N^\star$, respectively. For fixed λ and $\delta/\sigma \searrow 0$, (41) dominates the behavior of $N^\star$ as it is essentially of the form $\sigma/\delta \leq c(\lambda)\sqrt{\ln(N^\star)}$, whereas (42) is of the form $\sigma/\delta \leq c(\lambda, \mathfrak{A}, m)\sqrt{N^\star}/\ln(N^\star)$. Conversely, for fixed δ/σ and $\lambda \searrow 0$, (42) dominates the behavior of $N^\star$ as it is essentially of the form $\lambda^{-1}\ln(\lambda^{-1}) \leq c(\delta/\sigma, \mathfrak{A}, m)N^\star/\ln^2(N^\star)$ whereas (41) is of the form $\lambda^{-1}\ln(\lambda^{-1}) \leq c(\delta/\sigma)N^\star$.

THEOREM 2.7. *Consider the SBSSR-model with $g \in \mathcal{M}_\lambda^\delta$. Let $\hat{\omega}$ and $\hat{f} = (\hat{f}^1, \dots, \hat{f}^m)$ be the SLAM estimators from (21) and (25), respectively, with $\alpha = \alpha_n$ and $\beta = \beta_n$ as in (23). Further, let $\hat{\tau}^i$ and τ^i be the vectors of all c.p. locations of $\hat{f}^i$ and f^i, respectively, for $i = 1, \dots, m$. Then for all $n > N^\star$ in (41) and (42) and for all $i = 1, \dots, m$:*

1. $K(\hat{f}^i) = K(f^i)$,
2. $\max_j |\hat{\tau}_j^i - \tau_j^i| \leq 2\frac{\ln^2(n)}{n}$,
3. $\max_j |\hat{f}^i|_{[\hat{\tau}_j, \hat{\tau}_{j+1})} - f^i|_{[\tau_j, \tau_{j+1})}| = 0$, *and*
4. $|\hat{\omega}_i - \omega_i| \leq \frac{c_2}{a_2 - a_1}\frac{\ln(n)}{\sqrt{n}}$,

with probability at least $1 - \exp(-c_1\ln^2(n))$, *with* c_1 *and* c_2 *as in* (40).

From the proof of Theorem 2.7 (see Section S1.2 in the Supplementary Material [5]) it also follows that assertions 1.–4. hold for any $\hat{\omega} \in \mathcal{C}_{1-\alpha}(Y)$ and we obtain the following.

COROLLARY 2.8. *Consider the SBSSR-model with $g \in \mathcal{M}_\lambda^\delta$. Let $\mathcal{C}_{1-\alpha}(Y)$ be as in (20) and α_n as in (23). Further, let $\overline{\text{dist}}$ be is as in (22). Then for all $n > N^\star$ in (41) and (42)*

$$\overline{\text{dist}}(\omega, \mathcal{C}_{1-\alpha_n}(Y)) < \frac{c_2}{a_2 - a_1}\frac{\ln(n)}{\sqrt{n}}$$

with probability at least $1 - \exp(-c_1\ln^2(n))$, *with* c_1 *and* c_2 *as in* (40).

REMARK 2.9 (SLAM (almost) attains minimax rates).

(a) (C.p. locations.) Theorem 2.7 states that we can recover the c.p. locations of f^i in probability with rate $\ln^2(n)/n$. Obviously, the estimation rate of the c.p. locations is bounded from below by the sampling rate $1/n$. Consequently, the rate of Theorem 2.7 differs from the optimal rate only by a $\ln^2(n)$ factor.

(b) (Weights.) By the one-to-one correspondence between the weights and the function values of g the weights' detection rate $\ln(n)/\sqrt{n}$ immediately follows from the box height in (16) with $q_n(\alpha_n) \in \mathcal{O}(\ln(n))$ and coincides with the optimal rate $\mathcal{O}(1/\sqrt{n})$ up to a $\ln(n)$ term.

(c) (Dependence on λ.) The minimal scale λ in Theorem 2.7 may depend on n, that is, $\lambda = \lambda_n$. In order to ensure consistency of SLAM's estimates $\hat{\omega}$ and $(\hat{f}^1, \ldots, \hat{f}^m)$, Theorem 2.7 requires that (41) and (42) holds (for a sufficiently large $N^\star$) and that $c_2 \ln(n)/\sqrt{n} \to 0$, as $n \to \infty$. By Remark 2.6, this is fulfilled whenever $\lambda^{-1}\ln(\lambda^{-1}) \in \mathcal{O}(n/\ln^2(n))$. This means that the statements 1.–4. in Theorem 2.7 hold true asymptotically with probability one as the minimal scale λ_n of successive jumps in a sequence of mixtures g_n does not asymptotically vanish as fast as of order $\ln^3(n)/n$. We stress that no method can recover finer details of a bump signal (including the mixture g) below its detection boundary which is of the order $\ln(n)/n$, that is, SLAM achieves this minimax detection rate up to a $\ln^2(n)$ factor (see [30, 32]).

(d) (Dependence on δ.) Just as the minimal scale λ, the minimal ASB δ in Theorem 2.7 may depend on n as well, that is, $\delta = \delta_n$. Analog to 2.9), the SLAM's estimates remain consistent whenever $\delta^{-1} \in \mathcal{O}(\sqrt{\ln(n)})$, that is, the statements 1.–4. in Theorem 2.7 hold true asymptotically with probability one if the minimal ASB δ_n in a sequence of mixtures g_n does not decrease as fast as of order $1/\sqrt{\ln(n)}$. We stress that no method can recover smaller jump heights of the mixture g below its minimax detection rate, which in $1/\ln(n)$. To see this, note that statement 2. in Theorem 2.7 provides asymptotic detection power one for $2\ln(n)^2$ i.i.d. observations with mean δ_n (recall that the ASB corresponds to the minimal possible jump height of the mixture g). Hence, SLAM achieves the minimax rate up to a $\sqrt{\ln(n)}$ factor.

REMARK 2.10 (SLAM for known ω). If ω is known in the SBSSR-model, the second part of SLAM can be used separately. We may then directly solve (24) without pre-estimating ω, that is, in Section 1.4.3, we simply replace $\hat{\omega}$ by ω. Then, Theorem 2.5 is still valid for $\tilde{\mathcal{H}}(\beta)$ replaced by $\mathcal{H}(\beta)$. Further, a careful modification of the proof of Theorem 2.7 shows that the assertions in Theorem 2.7 hold for a possibly smaller $N^\star$ in (41) and (42) and for c_1 replaced by $75m^2(a_k - a_1)^2 c_1/(a_2 - a_1)^2$. We stress that the finite alphabet assumption is still required and the corresponding identifiability assumption $\mathrm{ASB}(\omega) \geq \delta$ must be valid.

3. Computational issues. SLAM is implemented in two steps. In the first step, for a given $\alpha \in (0,1)$ a confidence region for the mixing weights ω is computed as in Algorithm CRW (see Section 2.1 and S2.1). To this end, each of the n^{2m} m-boxes in $\mathfrak{B}^m = \{B(i,j) : 1 \leq i \leq j \leq n\}^m$ needs to be examined with the reduction rules R1–R4 for validity as a candidate box for the intervals $[i_1^\star, j_1^\star] \times \cdots \times [i_m^\star, j_m^\star]$, which yields the complexity $\mathcal{O}(n^{2m})$. There are, however, important pruning steps, which can lead to a considerably smaller complexity.

First, note that it suffices to consider m-boxes which are maximal elements with respect to the partial order of inclusion, that is, for $B^1 = [\underline{b}_1^1, \overline{b}_1^1] \times \cdots \times [\underline{b}_m^1, \overline{b}_m^1]$, $B^2 = [\underline{b}_1^2, \overline{b}_1^2] \times \cdots \times [\underline{b}_m^2, \overline{b}_m^2] \in \mathfrak{B}^m$

$$B^1 \preccurlyeq B^2 \quad \Leftrightarrow \quad [\underline{b}_i^1, \overline{b}_i^1] \subseteq [\underline{b}_i^2, \overline{b}_i^2] \qquad \text{for all } i = 1, \dots, m,$$

where an element a of a partially ordered set P is maximal if there is no element b in P such that $b > a$. To see this, assume that an m-box B is not deleted by the reduction rule R3 in the second last line of Algorithm CRW, then an m-box $B' \in \mathfrak{B}^m$ with $B' \prec B$ does not influence the confidence region $\mathcal{C}_{1-\alpha}$ (see the last line of Algorithm CRW), as $A^{-1}B' \subset A^{-1}B$. Conversely, if an m-box B is deleted by the reduction rule R3 in the second last line of Algorithm CRW, then an m-box $B' \in \mathfrak{B}^m$ with $B' \prec B$ will be deleted by R3 as well, such that B' does not need to be considered either.

Second, note that the parameter ω, which is inferred in Algorithm CRW, is global and, hence, one can restrict to observations on a subinterval $[x_i, x_j] \subset [0,1)$ as long as $g|_{[x_i, x_j]}$ fulfills the identifiability conditions of $\mathcal{M}^\delta$.

The explicit complexity of Algorithm CRW depends on the finial solution $\hat{f}$ itself. Depending on the final $\hat{f}$, the above mentioned pruning steps yield a complexity between $\mathcal{O}(n^m)$ and $\mathcal{O}(n^{2m})$. $\hat{\omega}$ is then computed as in (21).

In the second step, for a given $\beta \in (0,1)$ and given $\hat{\omega}$ SLAM solves the constrained optimization problem (25), which can be done using dynamic programming. Frick et al. [32] provide a pruned dynamic programming algorithm to efficiently solve a one-dimensional version of (25) without the finite alphabet restriction in (72). As this restriction is crucial for SLAM we outline the details of the necessary modifications in Section S2.2 in the Supplementary Material [5]. These modifications, however, do not change to complexity of the algorithm. Frick et al. [32] show that the overall complexity of the dynamic program depends on the final solution $\hat{g}$ and is between $\mathcal{O}(n)$ and $\mathcal{O}(n^2)$.

We stress finally that significant speed up (which is, however, not the subject of this paper) can be achieved by restricting the system of intervals in T_n and $\mathfrak{B}$, respectively, to a smaller subsystem, for example, intervals of dyadic length, which for example reveals the complexity of the second step as $\mathcal{O}(n \ln(n))$.

4. Simulations. In the following, we investigate empirically the influence of all parameters and the underlying signal on the performance of the SLAM estimator. As performance measures, we use the mean absolute error, MAE, for $\hat{\omega}$ and the mean absolute integrated error, MIAE, for $\hat{f}$. Further, we report the centered mean, $\operatorname{Mean}(\hat{K}) - K$, the centered median, $\operatorname{Med}(\hat{K}) - K$, of the number of c.p.'s of $\hat{f}$, the frequency of correctly estimated number of c.p.'s for the single source functions f^i, $\operatorname{Mean}(\hat{K} = K)_i$, and for the whole source function vector f, $\operatorname{Mean}(\hat{K} = K)$. To investigate the accuracy of the c.p. locations of the single estimated source functions $\hat{f}^1, \dots, \hat{f}^m$, we report the mean of $\max_i \min_j |\tau_i - \hat{\tau}_j|$ and $\max_j \min_i |\tau_i - \hat{\tau}_j|$, where τ and $\hat{\tau}$ denotes the vector of c.p. locations of the true signal and the estimate, respectively. Furthermore, we report common segmentation evaluation measures for the single estimated source functions $\hat{f}^1, \dots, \hat{f}^m$, namely the entropy-based V-measure, V_1, with balancing parameter 1 of [58] and the false positive sensitive location error, FPSLE, and the false negative sensitive location error, FNSLE, of [35]. The V-measure, taking values in [0, 1], measures whether given clusters include the correct data points of the corresponding class. Larger values indicate higher accuracy, 1 corresponding to a perfect segmentation. The FPSLE and the FNSLE capture the average distance between true and estimated segmentation boundaries, with FPSLE being larger if a spurious split is included, while FNSLE getting larger if a true boundary is not detected (see [35] for details). To investigate the performance of the confidence region $\mathcal{C}_{1-\alpha}$ for ω, we use $\overline{\operatorname{dist}}(\omega, \mathcal{C}_{1-\alpha})$ from (22), the mean coverage $\operatorname{Mean}(\omega \in \mathcal{C}_{1-\alpha})$, and the diameters $\overline{\omega}_i - \underline{\omega}_i$, where $\mathcal{C}_{1-\alpha} = [\overline{\omega}_1, \underline{\omega}_1] \times \cdots \times [\overline{\omega}_m, \underline{\omega}_m]$. Further, we report the mean coverage of the confidence band $\tilde{\mathcal{H}}(\beta)$, that is, $\operatorname{Mean}(f \in \tilde{\mathcal{H}}(\beta))$. In order to reduce computation time, we only considered intervals of dyadic length as explained in Section 3, possibly at expense of detection power. Simulation runs were always 10,000.

4.1. *Number of source functions m.* In order to illustrate the influence of the number of source functions m on the performance of SLAM, we vary $m = 2, \dots, 5$ while keeping the other parameters in the SBSSR-model fixed.

We investigate a binary alphabet $\mathfrak{A} = \{0, 1\}$ and set $f^i = \mathbb{1}_{[(i-1)/5, i/5)}$ for $i = 1, \dots, 5$, simple bump functions. For each $m \in \{2, 3, 4, 5\}$, we choose ω such that $\operatorname{ASB}(\omega) = 0.02$ in (5) (see Table S3.1 in the Supplementary Material [5]). For $\sigma = \delta = 0.02$, $n = 1000$, and $\alpha = \beta = 0.1$, we compute $\hat{\omega}$, $\mathcal{C}_{0.9}$, $\hat{f}^1, \dots, \hat{f}^m$, and $\tilde{\mathcal{H}}(0.1)$ for each $m \in \{2, 3, 4, 5\}$, incorporating prior knowledge $\lambda \geq 0.025$ [see (35) and Remark 2.1] (with truth $\lambda = 0.05$). The results are displayed in Table S3.2. A major finding is that as the number of possible mixture values equals k^m, the complexity of the SBSSR-model grows exponentially in m such that demixing becomes substantially more difficult with increasing m.

4.2. *Number of alphabet values k.* To illustrate the influence of the number of alphabet values k, we consider three different alphabets $\mathfrak{A}_k = \{0, \dots, k\}$ for $k = 2, 3, 4$. For $m = 2$, we set

$$(43) \qquad f^1 = \sum_{i=0}^{15} (i \mod k) \mathbb{1}_{[i,i+1)/16}, \, f^2 = \sum_{i=0}^{[15/k]} (i \mod k) \mathbb{1}_{k[i,i+1)/16},$$

step functions taking successively every alphabet value in $\mathfrak{A}^2$ (see Figure S3.1 in the Supplementary Material [5]). Further, we set $\omega = (0.02, 0.98)$ such that $\mathrm{ASB}(\omega) = 0.02$ for $k = 2, 3, 4$. For $\sigma = 0.05$, $n = 1056$, and $\alpha = \beta = 0.1$ we compute $\hat{\omega}$, $\mathcal{C}_{0.9}$, $\hat{f}^1, \dots, \hat{f}^m$, and $\tilde{\mathcal{H}}(0.1)$ for each $k = 2, 3, 4$, incorporating prior knowledge $\lambda \geq 1/32$ [see (35) and Remark 2.1] (with truth $\lambda = 1/16$). The results are displayed in Table S3.3 in the Supplementary Material [5]. From this, we find that an increasing k does not influence SLAM's performance for $\hat{\omega}$ and $\mathcal{C}_{1-\alpha}$ too much. However, the model complexity k^m increases polynomially (for $m = 2$ as in Table S3.3 quadratically) in k, reflected in a decrease of SLAM's performance for the estimate of the source functions $\hat{f}$.

4.3. *Confidence levels α and β.* We illustrate the influence of the confidence levels α and β on SLAM's performance with f and ω as in Example 1.1, that is, $m = 3$, $\mathfrak{A} = \{0, 1, 2\}$, $\omega = (0.11, 0.29, 0.6)$, and f as displayed in Figure 1. For $\sigma = 0.02, 0.05, 0.1$ and $n = 1280$, we compute $\hat{\omega}$, $\mathcal{C}_{1-\alpha}$, $\hat{f}^1, \dots, \hat{f}^m$, and $\tilde{\mathcal{H}}(\beta)$ for each $(\alpha, \beta) \in \{0.01, 0.05, 0.1\}^2$, incorporating prior knowledge $\lambda \geq 0.025$ [see (35) and Remark 2.1] (with truth $\lambda = 0.05$). Results are displayed in Table S3.4 and Table S3.5 in the Supplementary Material [5]. These illustrate that SLAM's estimate $\hat{\omega}$ for the mixing weights is very stable under the choice of α. The diameters $\overline{\mathrm{dist}}(\omega, \mathcal{C}_{1-\alpha})$ and $\overline{\omega}_i - \underline{\omega}_i$, respectively decrease slightly with increasing α, as expected. Further, we found that the coverage $\mathrm{Mean}(\omega \in \mathcal{C}_{1-\alpha})$ is always bigger than the nominal coverage $1 - \alpha$ indicating the conservative nature of the first inequality in (30). With increasing β the multiscale constraint in (24) becomes stricter leading to an increase of $\hat{K}$. However, as Table S3.5 illustrates, this effect is remarkably small, resulting also in a high stability of $\hat{f}$ with respect to α and β. In contrast to the uniform coverage of the confidence region $\mathcal{C}_{1-\alpha}$ for ω for finite n [recall (19)], this holds only asymptotically for the confidence band $\tilde{\mathcal{H}}(\beta)$ (see Theorem 2.5). This is reflected in Table S3.5, where with increasing σ the coverage $\mathrm{Mean}(f \in \tilde{\mathcal{H}}(\beta))$ can be smaller than the nominal $1 - \beta$. Nevertheless, the coverage of the single source functions remains reasonably high even for large σ (see Table S3.5). In summary, we draw from Table S3.4 and S3.5 a high stability of SLAM in the tuning parameters α and β, for both, the estimation error and the confidence statements, respectively.

4.4. *Prior information on the minimal scale λ.* In the previous simulations, we always included prior information on the minimal scale λ [see (35) and Remark 2.1]. In the following, we demonstrate the influence of this prior information on SLAM's performance in Example 1.1, that is, $m = 3$, $\mathfrak{A} = \{0, 1, 2\}$, $\omega = (0.11, 0.29, 0.6)$, and f as displayed in Figure 1. For $\sigma = 0.02$, $n = 1280$, and $\alpha = \beta = 0.1$ we compute $\hat{\omega}$, $\mathcal{C}_{0.9}$, $\hat{f}^1, \dots, \hat{f}^m$, and $\tilde{\mathcal{H}}(0.1)$ under prior knowledge $\lambda \geq 0.05, 0.04, 0.025, 0.015, 0.005$ (with truth $\lambda = 0.05$). The results in Table S3.9 in the Supplementary Material [5] show a certain stability for a wide range of prior information on λ. Only when the prior assumptions on λ is of order 0.1λ (or smaller) SLAM's performance gets significantly worse.

4.5. *Robustness of SLAM.* Finally, we want to analyze SLAM's robustness against violations of model assumptions.

4.5.1. *Robustness against nonidentifiability.* Throughout this work, we assumed $g \in \mathcal{M}^\delta$, that is, $\omega \in \Omega^\delta(m)$ as in (6) and $f \in \mathcal{S}(\mathfrak{A})^m$ separable as in (7), in order to ensure identifiability. In the following, we briefly investigate SLAM's behavior if these conditions are close to be, or even violated.

Alphabet separation boundary δ. We start with the identifiability condition $\omega \in \Omega^\delta(m)$, that is, $\mathrm{ASB}(\omega) \geq \delta > 0$ as in (5). We reconsider Example 1.1, that is, $m = 3$, $\mathfrak{A} = \{0, 1, 2\}$, and f as displayed in Figure 1, but with ω chosen randomly, uniformly distributed on $\Omega(3)$. For $\sigma = 0.05$, $n = 1280$, and $\alpha = \beta = 0.1$ we compute $\hat{\omega}$, $\mathcal{C}_{1-\alpha}$, $\hat{f}^1$, $\hat{f}^2$, $\hat{f}^3$, and $\tilde{\mathcal{H}}(\beta)$, incorporating prior knowledge $\lambda \geq 0.025$ [see (35) and Remark 2.1] (with truth $\lambda = 0.05$). Consequently, for each run we get a different ω and $\mathrm{ASB}(\omega)$, respectively.

We found that SLAM's performance of $\hat{\omega}$ and $\mathcal{C}_{1-\alpha}$, respectively, is not much influenced by $\mathrm{ASB}(\omega)$ [see Table S3.7, where the average mean squared error of $\hat{\omega}$ and $\overline{\mathrm{dist}}(\omega, \mathcal{C}_{1-\alpha})$ remain stable when $\mathrm{ASB}(\omega)$ becomes small]. The situation changes of course, when it comes to estimation of f itself. $\mathrm{ASB}(\omega) = 0$ in (5) implies nonidentifiability of f, that is, it is not possible to recover f uniquely. Therefore, it is expected that small $\mathrm{ASB}(\omega)$ will lead to a bad performance of any estimator of f. This is also reflected in Theorem 2.7 where δ, with $\mathrm{ASB}(\omega) \geq \delta$, appears as a "conditioning number" of the SBSSR-problem. The results in Table S3.8 in the Supplementary Material [5] confirm the strong influence of $\mathrm{ASB}(\omega)$ on the performance of SLAM's estimate for f. However, as SLAM does not only give an estimate of f but also a confidence band $\tilde{\mathcal{H}}(\beta)$ this (unavoidable) uncertainty is also reflected in its coverage. To illustrate this define a local version of $\mathrm{ASB}(\omega)$ as $\mathrm{ASB}_x(\omega) := \min_{a \neq f(x) \in \mathfrak{A}^m} |\omega^\top a - \omega^\top f(x)|$. Intuitively, $\mathrm{ASB}_x(\omega)$ determines the difficulty to discriminate between the source functions at a certain location $x \in [0, 1)$. Now, define the local size of $\tilde{\mathcal{H}}(\beta)$ as $|\tilde{\mathcal{H}}_x(\beta)| := \#\{a \in \mathfrak{A}^m : \exists f \in \tilde{\mathcal{H}}(\beta) \text{ s.t. } f(x) = a\}$. Table S3.8 in the Supplementary Material [5] shows that the uncertainty in $|\tilde{\mathcal{H}}_x(\beta)|$ increases in nonidentifiable regions, that is, when $\mathrm{ASB}_x(\omega)$ is small.

Violation of separability condition. Next, we consider the separability condition in (7). We consider a modification of Example 1.1, that is, $m = 3$, $\mathfrak{A} = \{0, 1, 2\}$, where we modified the source function f^1 in such a way, that it violates the separability condition in (7) for $r = 1$ (see Figure S3.2 in the Supplementary Material [5]). For $\sigma = 0.05$, $n = 1280$, and $\alpha = \beta = 0.1$, we compute $\hat{\omega}$ and $\hat{f}^1$, $\hat{f}^2$, $\hat{f}^3$ incorporating prior knowledge $\lambda \geq 0.025$ [see (35) and Remark 2.1] (with truth $\lambda = 0.05$). The results are shown in Table S3.6 in the Supplementary Material [5]. The violation of the separability condition in (7) leads to nonidentifiabilty of ω, which is naturally reflected in a worse performance of SLAM's estimate of ω. As the condition is violated for $r = 1$ this has a particular impact on $\hat{\omega}_1$. The performance for $\hat{\omega}_2$ and $\hat{\omega}_3$ remains relatively stable. The same holds true for $\hat{f}$ itself, where the estimation error of $\hat{\omega}_1$ propagates to a certain degree to the estimation of $\hat{f}^1$. The performance of $\hat{f}^2$ and $\hat{f}^3$, however, is not much influenced.

4.5.2. *Violation of normality assumption.* In the SBSSR-model, we assume that the error distribution is normal, that is, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \sim \mathcal{N}(0, I_n)$. In the following we study SLAM's performance for t-(heavy tails) and χ^2-(skewed) distributed errors. Again, we reconsider Example 1.1, that is, $m = 3$, $\mathfrak{A} = \{0, 1, 2\}$, and f as displayed in Figure 1. We add to g now t-distributed and χ^2-distributed errors, respectively, with 3 degrees of freedom, re-scaled to a standard deviation of $\sigma = 0.05$. For $n = 1280$ and $\alpha = 0.1$, we compute $\hat{\omega}$ and $\hat{f}^1$, $\hat{f}^2$, $\hat{f}^3$, incorporating prior knowledge $\lambda \geq 0.025$ [see (35) and Remark 2.1] (with truth $\lambda = 0.05$). We simulated the statistic T_n for t- and χ^2-distributed errors, respectively, and choose $q(\beta)$ to be the corresponding 90% quantile. For t-distributed errors, this gave $q(\beta) = 13.03$ and for χ^2-distributed errors $q(\beta) = 3.73$. The results (see Table S3.6 in the Supplementary Material [5]) indicate a certain robustness to misspecification of the error distribution, provided the quantiles for T_n are adjusted accordingly.

4.6. *Selection of $q_n(\alpha)$ and $q_n(\beta)$.* On the one hand, for given α and β SLAM yields confidence statements for the weights ω and the source functions f at level $1 - \alpha$ and $1 - \beta$, respectively. This suggests the choice of these parameters as confidence levels. On the other hand, when we target to estimate ω and f $q_n(\alpha)$ and $q_n(\beta)$ can be seen as tuning parameters for the estimates $\hat{\omega}$ and $\hat{f}$. Although, we found in Section 4.3 that SLAM's estimates are quite stable for a range of α's and β's, a fine tuning of these parameters improves estimation accuracy, of course. In the following, we suggest a possible strategy for this. First, we discuss $q_n(\alpha)$ for tuning the estimate $\hat{\omega}_q := \hat{\omega}(Y, q)$. Recall that for estimating ω, $q_n(\beta)$ is not required.

Minimal valid threshold (*MVT*). Theorem 2.7 yields $\ln(n)/\sqrt{n}$-consistency of $\hat{\omega}$ when $q_n(\alpha) = q_n(\alpha_n)$ with α_n as in (23), independently of the specific choice of $\hat{\omega} \in \mathcal{C}_{1-\alpha_n}$. Further, for α' [and $q_n(\alpha')$, respectively] with $\alpha' \geq \alpha_n$ [and $q_n(\alpha') \leq q_n(\alpha_n)$, respectively] $\mathcal{C}_{1-\alpha'} \subseteq \mathcal{C}_{1-\alpha_n}$ whenever $\mathfrak{B}^\star = \mathfrak{B}^\star_{q_n(\alpha')} \neq \varnothing$ in

(20). Thus, choosing the threshold q, for any discrete set $Q = \{q_1, q_2, \dots, q_N = q_n(\alpha_n)\}$, as $q^\star := \min(q \in Q : \mathfrak{B}_q^\star \neq \varnothing)$ guarantees the convergence rates of Theorem 2.7 for the corresponding estimate $\hat{\omega}(Y, q^\star)$. In practice, we found $Q = \{-1.0, -0.9, \dots, 1.9, 2.0\}$ to be a sufficiently rich candidate set.

Sample splitting (*SST*). Alternatively, we can choose q such that a given performance measure $h(q) := \mathrm{E}[L(\hat{\omega}_q - \omega)]$ for estimating ω, for example, the MSE with $L = \|\cdot\|_2^2$, is minimized. As ω is unknown, we have to estimate $h(q)$, for which we suggest a simple sample splitting procedure. Details are given in Section S4, in the Supplementary Material [5]. Simulations indicate that, especially for high noise level, the MVT-selection method outperforms the SST-selection method in terms of standard performance measures like MSE and MAE. However, in contrast to the SST-selection method, the MVT-selection method cannot be tailored for a specific performance measure h.

It remains to select $q_n(\beta)$ (and β, respectively), which is required additionally for $\hat{f}$, recall (25) and (26). Theorem 2.7 suggests to choose $q_n(\beta) = q_n(\beta_n)$ with β_n as in (23), that is, $q_n(\beta) \to \infty$ with rate $\mathcal{O}(\log(n))$. For finite n, there exist several methods for selection of $q_n(\beta)$ in c.p. regression (see, e.g., [73]), which might be used here as well. However, due to the high stability of $\hat{f}$ in q (see Section 4.3 and Figure S4.2 in the Supplementary Material [5]), we simply suggest to choose $\beta = 0.1$, which we have used here for our data analysis. This choice controls the probability of overestimating the number of jumps in g, $\mathbf{P}(K(\hat{g}) > K(g)) \leq 0.1$ asymptotically. In general, it depends on the application. A large $q_n(\beta)$ (hence small β) has been selected in the subsequent application to remove spurious changes in the signal which appear biologically not as of much relevance.

5. Genetic sequencing data.

Recall from Section 1.7 that a tumor often consists of a few distinct subpopulations, so called clones, of DNA with distinct copy-number profiles arising from duplication and deletion of genetic material groups. The copy number profiles of the underlying clones in a sample measurement correspond to the functions $f^1, \dots, f^m$, the weights $\omega_1, \dots, \omega_m$ correspond to their proportion in the tumor, and the measurements correspond to the mixture g with some additive noise.

The most common method for tumor DNA profiling is via whole genome sequencing, which roughly involves the following steps:

1. Tumor cells are isolated, and the pooled DNA is extracted, amplified and fragmented through shearing into single-strand pieces.
2. Sequencing of the single pieces takes place using short "reads" (at time of writing of around 10^2 base-pairs long).
3. Reads are aligned and mapped to a reference genome (or the patient germline genome if available) with the help of a computer.

Although, the observed total reads are discrete (each observation corresponds to an integer number of reads at a certain locus), for a sufficiently high sequencing coverage, as it is the case in our example with around 55 average stretches of DNA mapped to a locus, it is well established to approximate this binomial by a normal variate (see [48] and references there).

In the following, SLAM is applied to the cell line LS411, which comes from colorectal cancer and a paired lymphoblastoid cell line. Sequencing was done through a collaboration of Complete Genomics with the Wellcome Trust Center for Human Genetics at the University of Oxford. This data has the special feature of being generated under a designed experiment using radiation of the cell line (*"in vitro"*), designed to produce CNAs that mimic real world copy-number events. In this case therefore, the mixing weights and sequencing data for the individual clones are known, allowing for validation of SLAM's results, something that is not feasible for patient cancer samples.

The data comes from a mixture of three different types of DNA, relating to a normal (germline) DNA and two different clones. Tumor samples, even from micro-dissection, often contain high proportion of normal cells, which for our purposes are a nuisance, this is known as "stromal contamination" of germline genomes in the cancer literature. The true mixing weights in our sample are $\omega^\top = (\omega_{\text{Normal}}, \omega_{\text{Clone1}}, \omega_{\text{Clone2}}) = (0.2, 0.35, 0.45)$.

SLAM will be, in the following, applied only to the mixture data without knowledge of ω and the sequenced individual clones and germline. The latter (which serve as ground truth) will then be used only for validation of SLAM's reconstruction. We restricted attention to regions of chromosome 4, 5, 6, 18 and 20, as detailed below. Figure S3.3 in the Supplementary Material [5] shows the raw data. Sequencing produces some spatial artefacts in the data, and waviness related to the sequencing chemistry and local GC-content, corresponding to the relative frequency of the DNA bases {C, G} relative to {A, T}. This violates the modeling assumptions. To alleviate this we preprocess the data with a smoothing filter using local polynomial kernel regression on normal data, baseline correction, and binning. We used the local polynomial kernel estimator from the R package `KernSmooth`, with bandwidth chosen by visual inspection. We selected the chromosomal regions above as those showing reasonable denoising, and take the average of every 10th data point to make the computation manageable resulting in $n = 7480$ data points spanning the genome. The resulting data is displayed in Figure S3.4 in the Supplementary Material [5], where we can see that the data is much cleaned in comparison with Figure S3.3 although clearly some artefacts and local drift of the signal remain.

With $\sigma = 0.21$ pre-estimated as in [20], SLAM yields the confidence region for $\alpha = 0.1$ $C_{0.9} = [0.00, 0.31] \times [0.28, 0.50] \times [0.33, 0.72]$. With $q_n(\alpha) = -0.15$ selected with the MVT-method from Section 4.6 we obtain $\hat{\omega} = (0.12, 0.35, 0.53)$. Figure S3.5 in the Supplementary Material [5] shows SLAM's estimates for $q_n(\beta) = 2.2$ (which corresponds to $\beta = 0.01$). The top row shows the estimate for

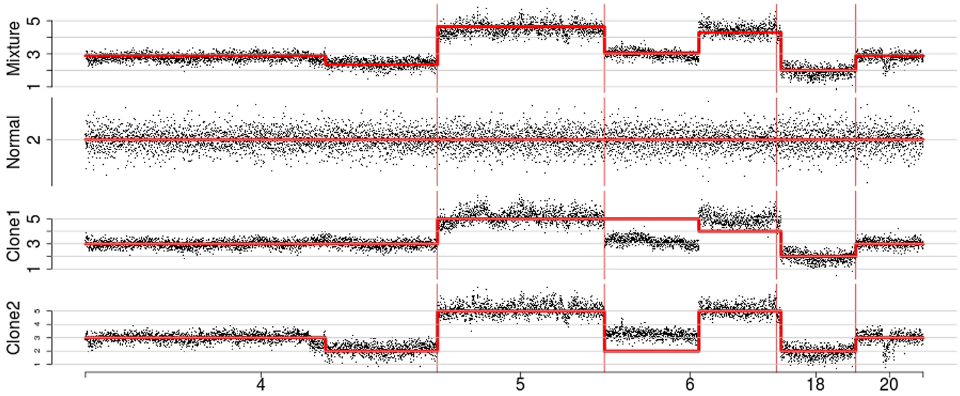

FIG. 5. *SLAM's estimates* (*red lines*) *for* $q_n(\alpha) = -0.15$ (*selected with MVT-method from Section* 4.6) *and* $q_n(\beta) = 20$. *Top row*: *total copy-number estimates across the genome. Rows* 2–4: *estimates of the CN profiles of the germline and clones.*

total copy number $\sum_j \hat{\omega}_j \hat{f}^j$ and rows 2–4 show $\hat{f}^1$, $\hat{f}^2$, and $\hat{f}^3$. We stress that the data for the single clones are only used for validation purposes and do not enter the estimation process. Inspection of Figure S3.5 shows that artefacts and local drifts of the signal result in an overestimation of the number of jumps. However, the overall appearance of the estimated CNA profile remains quite accurate. This over-fitting effect caused by these artifacts can be avoided by increasing SLAM's tuning parameter $q_n(\beta)$ at the (unavoidable) cost of loosing detection power on small scales [see Figure 5, which shows SLAM's estimate for $q_n(\beta) = 20$]. In summary, Figure 5 (and S3.5) show that SLAM can yield highly accurate estimation of the total CNA profile in this example, as well as reasonable CNA profiles and their mixing proportions for the clones, something which has not been obtainable prior to now. The analysis takes around 1 minute to run on a desktop computer with an intel core i7 processor. In future work, we aim to speed up the algorithm and explore association between the CNA patient profiles and clinical outcome data such as time-to-relapse and response to therapy.

6. Conclusion and discussion.

In this paper, we have established a new approach for separating linear mixtures of step functions with a known finite alphabet for additive Gaussian noise. This is of major interest for cancer genetics, but appears in other applications as well, for instance, in digital communications. We are not aware of any other method that deals with this problem in such a rigorous and general way. However, there are still some further generalizations and extensions to be studied.

Although we obtained a certain robustness of SLAM to misspecification of the error distribution in our simulation study, it is natural to ask how the results of this work can be extended to other types of error distributions than the normal distribution. [28, 29, 32] give several results about the multiscale statistic T_n, its limit distribution, and its geometric interpretation—which leads to the definition of the boxes $\mathfrak{B}$ [see (16)] for general one-dimensional exponential families. Combining this with the results of this work should yield extensions for such distributions.

In contrast to the noiseless case, $\varepsilon \equiv 0$ in (4), where the weights can be reconstructed in $\mathcal{O}(k^m)$ (independent of n) steps [6, 22], SLAM's estimation for the weights requires between $\mathcal{O}(n^m)$ and $\mathcal{O}(n^{2m})$ steps. Without further parallelization, this restricts the applicability of the algorithm to small number of mixtures m. Significant speed up can also be achieved when a smaller system of intervals in T_n is used (at the possible expense of finite sample detection power), for example, all intervals of dyadic length, in which case the worst case complexity reduces to $\mathcal{O}((n\ln(n))^m)$.

A further important issue is an extension for unknown number of source functions m. Clearly, this is a model selection problem, which might be approached with standard methods like the BIC or AIC criterion in conjunction with SLAM, a topic for further research.

One may also ask the question, whether the SBSSR-model can be treated for infinite alphabets $\mathfrak{A}$. The condition $\operatorname{ASB}(\omega) > 0$ in (6) remains necessary in order to guarantee identifiability, that is, different mixture values must be well separated. This condition, however, becomes significantly more restrictive when the size of the alphabet increases. Even for the most simple (infinite) alphabet $\mathfrak{A} = \mathbb{N}$ there exists no $m \geq 2$, $\omega \in \Omega(m)$ which fulfills $\operatorname{ASB}(\omega) > 0$, that is, no method can be valid in this situation. To see this, fix some $\omega \in \Omega(m)$ and w.l.o.g. assume that $\omega_1 \in \mathbb{Q}$, that is, $\omega_1 = n/d$ with $n, d \in \mathbb{N}$ and $d > n$. Then, $\tilde{d} := (d-n)d \in \mathbb{N}$, $n \cdot d \in \mathbb{N}$, and $\operatorname{ASB}(\omega) \leq |(\tilde{d}\omega_1 + 0 \cdot (1-\omega_1)) - (0 \cdot \omega_1 + nd(1-\omega_1))| = 0$. Hence, finiteness of the alphabet $\mathfrak{A}$ is fundamental for identifiability in the SBSSR-model.

Another issue is the extension to unknown (but finite) alphabets. If only certain parameters of the alphabet are unknown, for example, an unknown scaling constant, the alphabet is of the form $\mathfrak{A} = \{La_1, \ldots, La_k\}$ with a_i's known but L unknown, we speculate that generalizations should be possible and will rely on corresponding identifiability conditions, which are unknown so far. An arbitrary unknown alphabet, however, clearly leads to an unidentifiable model. This raises challenging issues, which we plan to address in the future.

Acknowledgements. Helpful comments of Hannes Sieling, Christopher Yau, and Philippe Rigollet are gratefully acknowledged. We are also grateful to two referees and one Editor for their constructive comments which led to an improved version of this paper.

SUPPLEMENTARY MATERIAL

Supplement to Multiscale Blind Source Separation (DOI: 10.1214/17-AOS1565SUPP; .pdf). Proofs of Theorem 1.4, Theorem 2.5, and Theorem 2.7 (Section S1); additional details on algorithms (Section S2); additional figures and tables from Section 4 and 5 (Section S3); details on the SST-method (Section S4).

REFERENCES

[1] AÏ SSA-EL-BEY, A., PASTOR, D., SBAÏ, S. M. A. and FADLALLAH, Y. (2015). Sparsity-based recovery of finite alphabet solutions to underdetermined linear systems. *IEEE Trans. Inform. Theory* **61** 2008–2018. MR3332994

[2] ARORA, S., GE, R., KANNAN, R. and MOITRA, A. (2012). Computing a nonnegative matrix factorization—provably. In *STOC'*12*—Proceedings of the* 2012 *ACM Symposium on Theory of Computing* 145–161. ACM, New York. MR2961503

[3] ARORA, S., GE, R., MOITRA, A. and SACHDEVA, S. (2015). Provable ICA with unknown Gaussian noise, and implications for Gaussian mixtures and autoencoders. *Algorithmica* **72** 215–236. MR3332931

[4] BAI, J. and PERRON, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66** 47–78. MR1616121

[5] BEHR, M., HOLMES, C. and MUNK, A. (2018). Supplement to "Multiscale blind source separation." DOI:10.1214/17-AOS1565SUPP.

[6] BEHR, M. and MUNK, A. (2015). Identifiability for blind source separation of multiple finite alphabet linear mixtures. *IEEE Trans. Inform. Theory* **63** 5506–5517.

[7] BELKIN, M., RADEMACHER, L. and VOSS, J. (2013). Blind signal separation in the presence of Gaussian noise. *J. Mach. Learn. Res. Proc.* **30** 270–287.

[8] BEROUKHIM, R., MERMEL, C. H., PORTER, D., WEI, G., RAYCHAUDHURI, S., DONOVAN, J., BARRETINA, J., BOEHM, J. S., DOBSON, J., URASHIMA, M. et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* **463** 899–905.

[9] BIOGLIO, V., COLUCCIA, G. and MAGLI, E. (2014). Sparse image recovery using compressed sensing over finite alphabets. *IEEE Int. Conf. Image Process.* (*ICIP*) 1287–1291.

[10] BOFILL, P. and ZIBULEVSKY, M. (2001). Underdetermined blind source separation using sparse representations. *Signal Process.* **81** 2353–2362.

[11] BOYSEN, L., KEMPE, A., LIEBSCHER, V., MUNK, A. and WITTICH, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348

[12] CANDES, E. J. and TAO, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory* **52** 5406–5425. MR2300700

[13] CARLSTEIN, E., MÜLLER, H.-G. and SIEGMUND, D., eds. (1994). *Change-Point Problems. Lecture Notes—Monograph Series* **23**. IMS, Hayward, CA. MR1477909

[14] CARTER, S. L., CIBULSKIS, K., HELMAN, E., MCKENNA, A., SHEN, H., ZACK, T., LAIRD, P. W., ONOFRIO, R. C., WINCKLER, W., WEIR, B. A. et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30** 413–421.

[15] CHEN, H., XING, H. and ZHANG, N. R. (2011). Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays. *PLoS Comput. Biol.* **7** e1001060. MR2776334

[16] CHENG, M.-Y. and HALL, P. (1999). Mode testing in difficult cases. *Ann. Statist.* **27** 1294–1315. MR1740110

[17] COMON, P. (1994). Independent component analysis, a new concept? *Signal Process.* **36** 287–314.

[18] DAS, A. K. and VISHWANATH, S. (2013). On finite alphabet compressive sensing. *IEEE Int. Conf. Acoust., Speech Signal Process.* (*ICASSP*) 5890–5894.

[19] DAVIES, L., HÖHENRIEDER, C. and KRÄMER, W. (2012). Recursive computation of piecewise constant volatilities. *Comput. Statist. Data Anal.* **56** 3623–3631. MR2943916

[20] DAVIES, P. L. and KOVAC, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.* **29** 1–65. MR1833958

[21] DETTE, H., MUNK, A. and WAGNER, T. (1998). Estimating the variance in nonparametric regression—what is a reasonable choice? *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 751–764. MR1649480

[22] DIAMANTARAS, K. I. (2006). A clustering approach for the blind separation of multiple finite alphabet sequences from a single linear mixture. *Signal Process*. **86** 877–891.

[23] DING, L., WENDL, M. C., MCMICHAEL, J. F. and RAPHAEL, B. J. (2014). Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet*. **15** 556–570.

[24] DONOHO, D. and STODDEN, V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? *Adv. Neural Inf. Process. Syst*. **16**.

[25] DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189

[26] DRAPER, S. C. and MALEKPOUR, S. (2009). Compressed sensing over finite fields. *Proceedings of the* 2009 *IEEE international conference on Symposium on Information Theory* **1** 669–673.

[27] DU, C., KAO, C.-L. M. and KOU, S. C. (2016). Stepwise signal extraction via marginal likelihood. *J. Amer. Statist. Assoc*. **111** 314–330. MR3494662

[28] DÜMBGEN, L., PITERBARG, V. I. and ZHOLUD, D. (2006). On the limit distribution of multiscale test statistics for nonparametric curve estimation. *Math. Methods Statist*. **15** 20–25. MR2225428

[29] DÜMBGEN, L. and SPOKOINY, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist*. **29** 124–152. MR1833961

[30] DÜMBGEN, L. and WALTHER, G. (2008). Multiscale inference about a density. *Ann. Statist*. **36** 1758–1785. MR2435455

[31] FEARNHEAD, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Stat. Comput*. **16** 203–213. MR2227396

[32] FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **76** 495–580. MR3210728

[33] FRIEDRICH, F., KEMPE, A., LIEBSCHER, V. and WINKLER, G. (2008). Complexity penalized M-estimation: Fast computation. *J. Comput. Graph. Statist*. **17** 201–224. MR2424802

[34] FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist*. **42** 2243–2281. MR3269979

[35] FUTSCHIK, A., HOTZ, T., MUNK, A. and SIELING, H. (2014). Multiscale DNA partitioning: Statistical evidence for segments. *Bioinformatics* **30** 2255–2262.

[36] GREAVES, M. and MALEY, C. C. (2012). Clonal evolution in cancer. *Nature* **481** 306–313.

[37] GU, F., ZHANG, H., LI, N. and LU, W. (2010). Blind separation of multiple sequences from a single linear mixture using finite alphabet. *IEEE Int. Conf. Wirel. Commun. Signal Process*. (*WCSP*) 1–5.

[38] HA, G., ROTH, A., KHATTRA, J., HO, J., YAP, D., PRENTICE, L. M., MELNYK, N., MCPHERSON, A., BASHASHATI, A., LAKS, E. et al. (2014). TITAN: Inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res*. **24** 1881–1893.

[39] HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528. MR1087842

[40] HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc*. **105** 1480–1493. MR2796565

[41] JENG, X. J., CAI, T. T. and LI, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *J. Amer. Statist. Assoc*. **105** 1156–1166. MR2752611

[42] KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc*. **107** 1590–1598. MR3036418

[43] KOFIDIS, N., MARGARIS, A., DIAMANTARAS, K. and ROUMELIOTIS, M. (2008). Blind system identification: Instantaneous mixtures of n sources. *Int. J. Comput. Math*. **85** 1333–1340. MR2451492

[44] LEE, D. and SEUNG, S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401** 788–791.

[45] LEE, T. W., LEWICKI, M. S., GIROLAMI, M. and SEJNOWSKI, T. J. (1999). Blind source separation of more sources than mixtures using overcomplete representations. *Signal Process. Lett.* **6** 87–90.

[46] LI, J., RAY, S. and LINDSAY, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.* **8** 1687–1723. MR2332445

[47] LI, Y., AMARI, S. I., CICHOCKI, A., HO, D. W. and XIE, S. (2006). Underdetermined blind source separation based on sparse representation. *IEEE Trans. Signal Process.* **54** 423–437.

[48] LIU, B., MORRISON, C. D., JOHNSON, C. S., TRUMP, D. L., QIN, M., CONROY, J. C., WANG, J. and LIU, S. (2013). Computational methods for detecting copy number variations in cancer genome using next generation sequencing: Principles and challenges. *Oncotarget* **4** 1868.

[49] MATTESON, D. S. and JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.* **109** 334–345. MR3180567

[50] MÜLLER, H.-G. and STADTMÜLLER, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15** 610–625. MR0888429

[51] NIU, Y. S. and ZHANG, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.* **6** 1306–1326. MR3012531

[52] OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostat.* **5** 557–572.

[53] OOI, H. (2002). Density visualization and mode hunting using trees. *J. Comput. Graph. Statist.* **11** 328–347. MR1938139

[54] PAJUNEN, P. (1997). Blind separation of binary sources with less sensors than sources. *IEEE Int. Conf. Neural Netw.* **3** 1994–1997.

[55] POLONIK, W. (1998). The silhouette, concentration functions and ML-density estimation under order restrictions. *Ann. Statist.* **26** 1857–1877. MR1673281

[56] PROAKIS, J. G. (1995). *Digital Communications*. McGraw-Hill, New York.

[57] RECHT, B., RE, C., TROPP, J. and BITTORF, V. (2012). Factoring nonnegative matrices with linear programs. *Adv. Neural Inf. Process. Syst.* **25** 1214–1222.

[58] ROSENBERG, A. and HIRSCHBERG, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *EMNLP-CoNLL* **7** 410–420.

[59] ROSTAMI, M., BABAIE-ZADEH, M., SAMADI, S. and JUTTEN, C. (2011). Blind source separation of discrete finite alphabet sources using a single mixture. *IEEE Stat. Signal Process. Workshop* (*SSP*) 709–712.

[60] ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-CÔTÉ, A. and SHAH, S. P. (2014). PyClone: Statistical inference of clonal population structure in cancer. *Nat. Methods* **11** 396–398.

[61] SHAH, S. P., ROTH, A., GOYA, R., OLOUMI, A., HA, G., ZHAO, Y., TURASHVILI, G., DING, J., TSE, K., HAFFARI, G. et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486** 395–399.

[62] SIEGMUND, D. (2013). Change-points: From sequential detection to biology and back. *Sequential Anal.* **32** 2–14. MR3023983

[63] SIEGMUND, D. and YAKIR, B. (2000). Tail probabilities for the null distribution of scanning statistics. *Bernoulli* **6** 191–213. MR1748719

[64] SPIELMAN, D. A., WANG, H. and WRIGHT, J. (2012). Exact recovery of sparsely-used dictionaries. *J. Mach. Learn. Res. Proc.* **23** 37.1–37.18.

[65] SPOKOINY, V. (2009). Multiscale local change point detection with applications to value-at-risk. *Ann. Statist.* **37** 1405–1436. MR2509078

[66] TALWAR, S., VIBERG, M. and PAULRAJ, A. (1996). Blind separation of synchronous co-channel digital signals using an antenna array—Part I. algorithms. *IEEE Trans. Signal Process.* **44** 1184–1197.

[67] TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 411–423. MR1841503

[68] TIBSHIRANI, R. and WANG, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostat.* **9** 18–29.

[69] VERDÚ, S. (1998). *Multiuser Detection*. Cambridge University Press, Cambridge.

[70] WALTHER, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.* **38** 1010–1033. MR2604703

[71] YAU, C., PAPASPILIOPOULOS, O., ROBERTS, G. O. and HOLMES, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 37–57. MR2797735

[72] YUANQING, L., CICHOCKI, A. and ZHANG, L. (2003). Blind separation and extraction of binary sources. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **86** 580–589.

[73] ZHANG, N. R. and SIEGMUND, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32. MR2345571

[74] ZHANG, N. R. and SIEGMUND, D. O. (2012). Model selection for high-dimensional, multi-sequence change-point problems. *Statist. Sinica* **22** 1507–1538. MR3027097

M. BEHR
A. MUNK
INSTITUTE FOR MATHEMATICAL STOCHASTICS
UNIVERSITY OF GOETTINGEN
GOLDSCHMIDTSTR. 7
37077 GÖTTINGEN
GERMANY
E-MAIL: behr@math.uni-goettingen.de
munk@math.uni-goettingen.de

C. HOLMES
DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
24-29 ST GILES'
OXFORD. OX1 3LB
UNITED KINGDOM
E-MAIL: cholmes@stats.ox.ac.uk