Piecewise Linear Regression under Noise Level Variation via Convex Optimization

Hiroki Kuroda and Jun Ogata

Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology

***Abstract*—Piecewise linear regression is a fundamental challenge in science and engineering. For typical applications where noise level varies in observations, the problem becomes much more challenging. In this paper, we propose a convex optimization based piecewise linear regression method which incorporates variation of the noise level. More precisely, we newly design a convex data-fidelity function as a weighted sum of approximation errors to mitigate effect of the noise level variation. The weights are automatically adjusted to the varying noise level within the framework of convex optimization. Numerical examples show performance improvements by the proposed method.**

***Index Terms*—Piecewise linear regression, noise level variation, convex optimization, change detection**

I. Introduction

Piecewise linear regression is a challenging task which aims to decompose an observed signal into segments where linear regressions fit well. This task is ubiquitous in science and engineering, and particularly important for change detection applications, e.g., [1]–[10] where endpoints of the segments should reflect significant changes in the signal. We call the endpoints *change points*. For typical applications, e.g., change detection in speech [1], [2] and seismic waves [3]–[5], the task becomes even more challenging due to variation of noise level in the observed signal.[1]

Since direct formulations for searching globally optimal change points result in hard nonconvex optimization [11], [12], most existing methods, e.g., [13]–[15] heuristically detect change points based on local statistics to reduce computational cost. These methods are local in nature, and thus prone to be sensitive against noise, as mentioned in [16], [17].

Recently, promising alternative approaches [16]–[20] translate the piecewise linear regression into estimation of a sparse support in strategically designed representation spaces. At the cost of disregarding the noise level variation, a sparse support can be effectively estimated by convex optimization such as ℓ_1 regularized least squares. Namely, under the technical assumption of a uniform noise level, these approaches find certain optimal change points via efficient convex optimization. Thus, although the estimation of a sparse support under the noise level variation still remains an open problem, it seems promising to extend the approach to incorporate the noise level variation.

In this paper, by designing a new convex optimization formulation for the sparse support estimation under the noise level variation, we improve the robustness of the sparsity-aware approach against variation of the noise level. More precisely, we newly design a convex data-fidelity function under consideration of the noise level variation. We focus on the point that the existing approaches neglect the noise level variation due to the uniform sum of approximation errors. Motivated by this observation, we design the data-fidelity function by introducing weights of approximation errors which are adapted to the varying noise level via convex optimization. The design of the proposed data-fidelity function is inspired by the idea in, e.g., [21]–[23] where a single weight is used to estimate a *uniform* noise level. In contrast with [21]–[23], the weights of the proposed data-fidelity function are encouraged to be piecewise constant to cope with abrupt changes of the noise level.

To demonstrate the effectiveness of the proposed method for piecewise linear regression under variation of the noise level, we compare the proposed method against the ℓ_1 regularized least squares based method. The results show performance improvements by the proposed technique. In addition, in our experiments, the performance of the proposed method is not sensitive against the tuning parameter which controls the degree of noise level variation.

Notations: $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{R}_{>0}$ denote the sets of nonnegative integers, real numbers, and positive real numbers respectively. For matrices or vectors, we denote the transpose by $(\cdot)^\top$. For $\boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{X} \in \mathbb{R}^{N\times M}$, $[\boldsymbol{x}]_n$ and $[\boldsymbol{X}]_{n,m}$ respectively denote the n-th component of $\boldsymbol{x}$ and the (n,m) entry of $\boldsymbol{X}$. We define the support of $\boldsymbol{x} \in \mathbb{R}^N$ by $\mathrm{supp}(\boldsymbol{x}) := \{n \in \{1,\ldots,N\} \mid [\boldsymbol{x}]_n \neq 0\}$. The ℓ_1 norm and the ℓ_0 pseudo-norm of $\boldsymbol{x} \in \mathbb{R}^N$ are respectively denoted by $\|\boldsymbol{x}\|_1 := \sum_{n=1}^N \left|[\boldsymbol{x}]_n\right|$ and $\|\boldsymbol{x}\|_0 := |\mathrm{supp}(\boldsymbol{x})|$. We define the null space of $\boldsymbol{X} \in \mathbb{R}^{N\times M}$ by $\mathcal{N}(\boldsymbol{X}) := \{\boldsymbol{u} \in \mathbb{R}^M \mid \boldsymbol{X}\boldsymbol{u} = \boldsymbol{0}\}$.

II. Preliminaries

A. Problem Formulation

We consider that the observed signal y_n $(n = 1,\ldots,N)$ is modeled by a piecewise linear regression model with variation of the noise level:

$$y_n := \begin{cases} \boldsymbol{\xi}_n^\top \boldsymbol{\vartheta}_1^\star + \sigma_1^\star \varepsilon_n, & (n \in [1, n_1^\star]), \\ \vdots & \\ \boldsymbol{\xi}_n^\top \boldsymbol{\vartheta}_\ell^\star + \sigma_\ell^\star \varepsilon_n, & (n \in (n_{\ell-1}^\star, n_\ell^\star]), \\ \vdots & \\ \boldsymbol{\xi}_n^\top \boldsymbol{\vartheta}_{L+1}^\star + \sigma_{L+1}^\star \varepsilon_n, & (n \in (n_L^\star, N]), \end{cases} \quad (1)$$

The first author is now with Ritsumeikan University (e-mail: kuroda@media.ritsumei.ac.jp).

[1]For simplicity, the term *noise* is used to refer to modeling error as well as observation noise.

where $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N \in \mathbb{R}^K$ are the known regression vectors, $n_1^\star < \cdots < n_L^\star$ are unknown change points, $\boldsymbol{\vartheta}_1^\star, \ldots, \boldsymbol{\vartheta}_{L+1}^\star$ consist of unknown coefficients in each time-segment. In particular, this paper considers unknown scales $\sigma_1^\star, \ldots, \sigma_{L+1}^\star > 0$ associated with the noise term $\varepsilon_n \in \mathbb{R}$. Note that ε_n includes modeling error and observation noise. Although this is a sort of the heteroscedastic noise model, we focus on the particular situation where the noise variance is constant between consecutive change points which are unknown in advance.

The goal of piecewise linear regression is to estimate $n_1^\star, \ldots, n_L^\star \in \{1, \ldots, N\}$ and $\boldsymbol{\vartheta}_1^\star, \ldots, \boldsymbol{\vartheta}_{L+1}^\star \in \mathbb{R}^K$ from $(\boldsymbol{\xi}_n, y_n)_{n=1}^N$. Remark that, once $n_1^\star, \ldots, n_L^\star$ are estimated, the estimation of $\boldsymbol{\vartheta}_1^\star, \ldots, \boldsymbol{\vartheta}_{L+1}^\star$ reduces to the classical linear regression in each time-segment. Thus, the major goal of piecewise linear regression is the estimation of $n_1^\star, \ldots, n_L^\star$.

The problem shown in (1) reproduces many scenarios in the literature as its special instances. For instance, the problem (1) reproduces change detection in mean and variance of Gaussian random variables, e.g., [1] by specially setting $\boldsymbol{\xi}_n = 1$ and supposing that $\varepsilon_1, \ldots, \varepsilon_N$ are independently drawn from Gaussian distribution $\mathcal{N}(0,1)$. The problem (1) also covers segmentation of piecewise auto-regressive processes, e.g., [1]–[6] by restricting $\boldsymbol{\xi}_n$ to $\boldsymbol{\xi}_n = (y_{n-1}, \ldots, y_{n-K})^\top$. In addition, recovery of piecewise polynomial signals, e.g., [7], [8] and piecewise sinusoidal signals, e.g., [9], [10] are reproduced by setting $\boldsymbol{\xi}_n = (\varphi_1(n), \ldots, \varphi_K(n))^\top$ with $\varphi_1, \ldots, \varphi_K$ defined as polynomial functions and sinusoidal functions respectively.

B. Translating Piecewise Linear Regression into Estimation of Sparse Support

Based on [20], we briefly demonstrate how the estimation of $n_1^\star, \ldots, n_L^\star$ can be reduced to the estimation of sparse support. In the next lemma which is an application of [20, Lemma 1 & Theorem 1], we show the design of $\boldsymbol{W} \in \mathbb{R}^{(N-K)\times N}$ which makes $\boldsymbol{W}\boldsymbol{s}^\star$ sparse for *noiseless* samples

$$\boldsymbol{s}^\star := (y_1 - \tilde{\varepsilon}_1, \ldots, y_N - \tilde{\varepsilon}_N)^\top \in \mathbb{R}^N,$$

where $\tilde{\varepsilon}_n := \sigma_\ell^\star \varepsilon_n$ when $n \in (n_{\ell-1}^\star, n_\ell^\star]$ for some $\ell \in \{1, \ldots, L+1\}$.[2]

Lemma 1 (Support of $\boldsymbol{W}\boldsymbol{s}^\star$)**.** *For each $n = 1, \ldots, N-K$, define $\boldsymbol{\Phi}_n := (\boldsymbol{\xi}_n, \ldots, \boldsymbol{\xi}_{n+K})^\top \in \mathbb{R}^{(K+1)\times K}$, and let $\boldsymbol{w}_n \in \mathcal{N}(\boldsymbol{\Phi}_n^\top)\setminus\{\boldsymbol{0}\}$. Construct the matrix $\boldsymbol{W} \in \mathbb{R}^{(N-K)\times N}$ by using $\boldsymbol{w}_n$ to build its n-th row $(n = 1, \ldots, N-K)$ as*

$$([\boldsymbol{W}]_{n,1}, \ldots, [\boldsymbol{W}]_{n,N}) := (\underbrace{0, \ldots, 0}_{n-1}, \boldsymbol{w}_n^\top, \underbrace{0, \ldots, 0}_{N-n-K}). \quad (2)$$

Then, the support of $\boldsymbol{W}\boldsymbol{s}^\star$ is covered as

$$\operatorname{supp}(\boldsymbol{W}\boldsymbol{s}^\star) \subset \bigcup_{\ell=1}^{L} \{n_\ell^\star - K + 1, \ldots, n_\ell^\star\}. \quad (3)$$

The inclusion (3) implies that $\boldsymbol{W}\boldsymbol{s}^\star \in \mathbb{R}^{N-K}$ has at most KL nonzero components. Thus, if $KL \ll N$ (i.e. the number of samples is sufficiently many), $\boldsymbol{W}\boldsymbol{s}^\star$ is sparse.

[2]For convenience, we let $n_0^\star := 0$ and $n_{L+1}^\star := N$.

Meanwhile, Lemma 1 shows that the locations of change points $n_1^\star, \ldots, n_L^\star$ are reflected in $\operatorname{supp}(\boldsymbol{W}\boldsymbol{s}^\star)$. This suggests that the estimation of $\operatorname{supp}(\boldsymbol{W}\boldsymbol{s}^\star)$ is the key step for the estimation of $n_1^\star, \ldots, n_L^\star$. Indeed, if $\operatorname{supp}(\boldsymbol{W}\boldsymbol{s}^\star)$ is correctly estimated, $\hat{n}_1, \ldots, \hat{n}_{\hat{L}}$ obtained by the following algorithm are consistent with $n_1^\star, \ldots, n_L^\star$ under technical conditions which usually hold (see [20, Theorem 3] for detail). Note that information on L is not explicitly utilized in this paper, though it would be beneficial for the estimation of $n_1^\star, \ldots, n_L^\star$.

Algorithm 1 (Estimation of change points $n_1^\star, \ldots, n_L^\star$)**.** *Let $\operatorname{supp}(\boldsymbol{W}\hat{\boldsymbol{s}})$ be the estimate of $\operatorname{supp}(\boldsymbol{W}\boldsymbol{s}^\star)$ (For the estimation of $\operatorname{supp}(\boldsymbol{W}\hat{\boldsymbol{s}})$ under the noise level variation, see Section III below). Then, we obtain the estimates $\hat{n}_1, \ldots, \hat{n}_{\hat{L}}$ as*

$$\begin{aligned}
\hat{n}_{\hat{L}} &= \max\{n \in \{1, \ldots, N-K\} \mid [\boldsymbol{W}\hat{\boldsymbol{s}}]_n \neq 0\}, \\
\hat{n}_{\hat{L}-1} &= \max\{n \in \{1, \ldots, \hat{n}_{\hat{L}} - K\} \mid [\boldsymbol{W}\hat{\boldsymbol{s}}]_n \neq 0\}, \\
&\vdots \\
\hat{n}_1 &= \max\{n \in \{1, \ldots, \hat{n}_2 - K\} \mid [\boldsymbol{W}\hat{\boldsymbol{s}}]_n \neq 0\},
\end{aligned}$$

where $\hat{L}$, the estimate of L, is defined so that $\{n \in \{1, \ldots, \hat{n}_1 - K\} \mid [\boldsymbol{W}\hat{\boldsymbol{s}}]_n \neq 0\} = \varnothing$.

III. Proposed Method

The proposed piecewise linear regression method first estimates $\operatorname{supp}(\boldsymbol{W}\boldsymbol{s}^\star)$, and then estimates $n_1^\star, \ldots, n_L^\star$ by Algorithm 1.

The key of the proposed method is the convex optimization formulation which incorporates the noise level variation for estimation of $\operatorname{supp}(\boldsymbol{W}\boldsymbol{s}^\star)$. In Section III-A, we newly design a convex data-fidelity function $\psi(\boldsymbol{s})$ which measures the error between $y_1, \ldots, y_N$ and components of $\boldsymbol{s}$ under consideration of the noise level variation. Then, in Section III-B, we present a solver for the proposed formulation which combines the newly designed data-fidelity function and $\|\boldsymbol{W}\boldsymbol{s}\|_1$ promoting the sparsity of $\boldsymbol{W}\boldsymbol{s}$, i.e.,

$$\min_{\boldsymbol{s} \in \mathbb{R}^N} \psi(\boldsymbol{s}) + \lambda\|\boldsymbol{W}\boldsymbol{s}\|_1, \quad (4)$$

where $\lambda > 0$ controls the importance of sparsity.

A. Design of the Data-Fidelity Function

We first design a nonconvex data-fidelity function $\tilde{\psi}(\boldsymbol{s})$, and then obtain $\psi(\boldsymbol{s})$ as a certain convex relaxation of $\tilde{\psi}(\boldsymbol{s})$. The proposed data-fidelity function is inspired by the idea in, e.g., [21]–[23] for joint estimation of *uniform* noise level and regression coefficients. In our context, the data-fidelity function of [21]–[23] is defined as

$$\min_{\sigma > 0} \left(\sum_{n=1}^{N} \frac{(y_n - [\boldsymbol{s}]_n)^2}{\sigma} + \sigma \right),$$

where σ is introduced to estimate the uniform noise level. Introduction of σ in this way is beneficial in terms of convexity of the overall cost function in $(\boldsymbol{s}, \sigma) \in \mathbb{R}^N \times \mathbb{R}_{>0}$, when substituted into (4) for ψ. If $n_1^\star, \ldots, n_L^\star$ are known, this data-fidelity function can be naturally extended to cope with multiple noise levels:

$$\min_{\sigma_1,\ldots,\sigma_{L+1}>0} \sum_{\ell=1}^{L+1} \left(\sum_{n=n_{\ell-1}^{\star}+1}^{n_{\ell}^{\star}} \frac{(y_n - [\boldsymbol{s}]_n)^2}{\sigma_\ell} + \sigma_\ell \right). \tag{5}$$

This data-fidelity function mitigates the effect of noise level variation by adjusting the weights $1/\sigma_1, \ldots, 1/\sigma_{L+1}$, i.e., assigning larger σ_ℓ if $(y_n - [\boldsymbol{s}]_n)^2$ for $n \in (n_{\ell-1}^{\star}, n_{\ell}^{\star}]$ have larger errors. However, since $n_1^{\star}, \ldots, n_L^{\star}$ are unknown, this data-fidelity function is not available in practice. Thus, we propose to use the minimum over possible change points:

$$\tilde{\psi}(\boldsymbol{s}) := \min_{\substack{\sigma_1,\ldots,\sigma_{L+1}>0,\\ (n_1,\ldots,n_L)\in\mathcal{I}}} \sum_{\ell=1}^{L+1} \left(\sum_{n=n_{\ell-1}+1}^{n_\ell} \frac{(y_n - [\boldsymbol{s}]_n)^2}{\sigma_\ell} + \sigma_\ell \right), \tag{6}$$

where $\mathcal{I} := \{(n_1, \ldots, n_L) \in \{1, \ldots, N\}^L \mid n_1 < \cdots < n_L\}$ is the set of possible change point locations.

The data-fidelity function $\tilde{\psi}(\boldsymbol{s})$ is expected to effectively approximate (5) because of the following reason. Roughly speaking, the function in (6) is suppressed when a single weight is assigned to consecutive $(y_n - [\boldsymbol{s}]_n)^2$'s of close values. Meanwhile, since we suppose that the noise level is constant in $(n_{\ell-1}^{\star}, n_\ell^{\star}]$, $(y_n - [\boldsymbol{s}]_n)^2$ for $n \in (n_{\ell-1}^{\star}, n_\ell^{\star}]$ will have errors of the same level. Thus, the minimum in (6) is expected to be attained around $n_1^{\star}, \ldots, n_L^{\star}$.

Since the grid search of $(n_1, \ldots, n_L) \in \mathcal{I}$ involved in the evaluation of $\tilde{\psi}(\boldsymbol{s})$ is computationally intractable, we instead use $\psi(\boldsymbol{s})$ which is obtained via convex relaxation of the function to be minimized in (6). Notice that $\tilde{\psi}(\boldsymbol{s})$ can be equivalently written as

$$\tilde{\psi}(\boldsymbol{s}) = \min_{\boldsymbol{\sigma}\in\mathbb{R}_{>0}^N} \left(\sum_{n=1}^{N} \frac{(y_n - [\boldsymbol{s}]_n)^2}{[\boldsymbol{\sigma}]_n} + [\boldsymbol{\sigma}]_n \right) \quad \text{subject to } \|\boldsymbol{D\sigma}\|_0 \le L,$$

where $\boldsymbol{D}$ is the first discrete difference operator defined as

$$\boldsymbol{D} := \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(N-1)\times N}.$$

By replacing the ℓ_0 pseudo-norm with the ℓ_1 norm, we finally obtain the proposed data-fidelity function which can be evaluated by convex optimization:

$$\psi(\boldsymbol{s}) = \min_{\boldsymbol{\sigma}\in\mathbb{R}_{>0}^N} \left(\sum_{n=1}^{N} \frac{(y_n - [\boldsymbol{s}]_n)^2}{[\boldsymbol{\sigma}]_n} + [\boldsymbol{\sigma}]_n \right) \quad \text{subject to } \|\boldsymbol{D\sigma}\|_1 \le \alpha, \tag{7}$$

where $\alpha > 0$ is a tuning parameter to control the degree of noise level variation. Plugging (7) into (4) yields convex optimization in $(\boldsymbol{s}, \boldsymbol{\sigma}) \in \mathbb{R}^N \times \mathbb{R}_{>0}^N$, for which efficient solvers are available (see Section III-B). We also present the following variant by adding the constraint into the data-fidelity function:

$$\psi(\boldsymbol{s}) = \min_{\boldsymbol{\sigma}\in\mathbb{R}_{>0}^N} \left(\sum_{n=1}^{N} \frac{(y_n - [\boldsymbol{s}]_n)^2}{[\boldsymbol{\sigma}]_n} + [\boldsymbol{\sigma}]_n \right) + \kappa\|\boldsymbol{D\sigma}\|_1, \tag{8}$$

with a tuning parameter $\kappa > 0$. Note that, by letting

$$g(\boldsymbol{v}) = \iota_{B_1^\alpha}(\boldsymbol{v}) := \begin{cases} 0, & \text{if } \|\boldsymbol{v}\|_1 \le \alpha; \\ \infty, & \text{otherwise,} \end{cases}$$

for (7) and $g(\boldsymbol{v}) = \kappa\|\boldsymbol{v}\|_1$ for (8), the proposed data-fidelity function can be written in a unified form as

$$\psi(\boldsymbol{s}) = \min_{\boldsymbol{\sigma}\in\mathbb{R}_{>0}^N} \left(\sum_{n=1}^{N} \frac{(y_n - [\boldsymbol{s}]_n)^2}{[\boldsymbol{\sigma}]_n} + [\boldsymbol{\sigma}]_n \right) + g(\boldsymbol{D\sigma}). \tag{9}$$

B. Solver for Proposed Formulation

The proposed formulation (4) with (9) is a convex optimization problem in $(\boldsymbol{s}, \boldsymbol{\sigma}) \in \mathbb{R}^N \times \mathbb{R}_{>0}^N$. Moreover, mainly from [24], we can efficiently compute the proximity operators of building blocks of the optimization problem. Thus, we can apply proximal splitting techniques, e.g., [25]–[29] to solve the problem. In particular, with the first order methods [25]–[28], we can implement an iterative solver which needs only $\mathcal{O}(KN)$ arithmetic operations per iteration because $\boldsymbol{W}$ has at most $K+1$ nonzero entries in each row (see (2)).

To be precise, we minimize the lower semicontinuous convex envelope of the cost function (4) with (9). That is,

$$\min_{(\boldsymbol{s},\boldsymbol{\sigma})\in\mathbb{R}^N\times\mathbb{R}^N} \sum_{n=1}^{N} f_n([\boldsymbol{s}]_n, [\boldsymbol{\sigma}]_n) + g(\boldsymbol{D\sigma}) + \lambda\|\boldsymbol{Ws}\|_1, \tag{10}$$

where

$$f_n(s, \sigma) := \begin{cases} \dfrac{(y_n - s)^2}{\sigma} + \sigma, & \text{if } \sigma > 0; \\ 0, & \text{if } s = y_n \text{ and } \sigma = 0; \\ \infty, & \text{otherwise.} \end{cases}$$

Note that this modification is slight because we only additionally suppose that $\frac{(y_n - [\boldsymbol{s}]_n)^2}{[\boldsymbol{\sigma}]_n} = 0$ for $[\boldsymbol{s}]_n = y_n$ and $[\boldsymbol{\sigma}]_n = 0$.

From [24, Example 2.4], the proximity operator of f_n can be computed as

$$\operatorname{prox}_{f_n}(s, \sigma) = \begin{cases} (y_n, 0), & \text{if } \sigma + \frac{(s-y_n)^2}{4} \le 1; \\ (y_n, \sigma - 1), & \text{if } s = y_n \ \&\ \sigma > 1; \\ \left(s - x\frac{s-y_n}{|s-y_n|}, \sigma - 1 + \frac{x^2}{4}\right), & \text{otherwise,} \end{cases} \tag{11}$$

where $x > 0$ is the unique positive root of

$$x^3 + 4(\sigma + 1)x - 8|s - y_n| = 0,$$

and can be explicitly given via Cardano's formula as follows. Let $p = 4(\sigma+1)$, $q = -8|s - y_n|$, and $D = -\frac{q^2}{4} - \frac{p^3}{27}$. Then,

$$x = \begin{cases} \sqrt[3]{-\frac{q}{2} + \sqrt[2]{-D}} + \sqrt[3]{-\frac{q}{2} - \sqrt[2]{-D}}, & \text{if } D < 0; \\ 2\sqrt[3]{-\frac{q}{2}} & \text{if } D = 0; \\ 2\sqrt[3]{\sqrt[2]{\frac{q^2}{4} + D}} \cos\left(\frac{\arctan(-2\sqrt[2]{D}/q)}{3}\right) & \text{if } D > 0, \end{cases}$$

where $\sqrt[2]{\cdot}$ and $\sqrt[3]{\cdot}$ denote the positive square root and the real cubic root respectively. For $\gamma > 0$, the proximity operator of $\gamma\|\cdot\|_1 \colon \mathbb{R}^d \to \mathbb{R}$ is given for each $i = 1, \ldots, d$ as

$$[\mathrm{prox}_{\gamma\|\cdot\|_1}(\boldsymbol{x})]_i = \mathrm{sign}([\boldsymbol{x}]_i) \max\left\{\left|[\boldsymbol{x}]_i\right| - \gamma, 0\right\}. \quad (12)$$

The proximity operator of $\iota_{B_1^\alpha}$, which reduces to the ℓ_1 ball projection, can be computed as

$$\mathrm{prox}_{\iota_{B_1^\alpha}}(\boldsymbol{v}) = \begin{cases} \boldsymbol{v}, & \text{if } \|\boldsymbol{v}\|_1 \le \alpha; \\ (\mathrm{sign}([\boldsymbol{v}]_n) a_n)_{n=1}^{N-1}, & \text{otherwise}, \end{cases} \quad (13)$$

where $a_n := \max\{|[\boldsymbol{v}]_n| - (\sum_{m=1}^{M} \rho_m - \alpha)/M, 0\}$ with $M := \max\{m \in \{1, \ldots, N-1\} \mid (\sum_{i=1}^{m} \rho_i - \alpha)/m < \rho_m\}$ and $\rho_1, \ldots, \rho_{N-1}$ the sorted copy of $|[\boldsymbol{v}]_1|, \ldots, |[\boldsymbol{v}]_{N-1}|$ in descending order.

By applying the linearized augmented Lagrangian method [27], [28], we show an iterative solver for (10) as Algorithm 2. The sequence $(\boldsymbol{s}_j, \boldsymbol{\sigma}_j)_{j=1}^{\infty}$ generated by Algorithm 2 converges to the solution of (10) if β in Algorithm 2 is chosen to be larger than the maximum singular value of $\begin{pmatrix} \boldsymbol{W}, & \boldsymbol{O}, & -\boldsymbol{I}, & \boldsymbol{O} \\ \boldsymbol{O}, & \boldsymbol{D}, & \boldsymbol{O}, & -\boldsymbol{I} \end{pmatrix}$ where $\boldsymbol{I}$ and $\boldsymbol{O}$ are the identity and zero matrices of appropriate sizes.

Remark 1 (Combination with other sparse representation)**.** We focus on the combination of the proposed data-fidelity function and the sparse representation in [20], but it is straightforward to combine with other sparse representation [16]–[19]. Note that, for piecewise linear regression under the uniform noise level, the lower dimensional representation space designed in [20] yields superior performance than [16]–[19]. Thus, it is expected that the combination of the proposed data-fidelity function and the sparse representation in [20] yields better performance for the situation where the noise level varies as well.

IV. Numerical Experiments

To show the effectiveness of the proposed method for piecewise linear regression under the noise level variation, we conduct numerical experiments on the estimation of piecewise sinusoidal signals and piecewise linear signals. More precisely, in (1), we set $\boldsymbol{\xi}_n = (\varphi_1(n), \ldots, \varphi_K(n))^\top$ $(n = 1, \ldots, N)$ with $\varphi_1, \ldots, \varphi_K$ specified below.

a) (Piecewise sinusoidal signal) We define φ_k as

$$\varphi_{2k'+1}(t) = \cos(\tfrac{2\pi k'}{K} t)\ (k' = 0, \ldots, (K-1)/2),$$
$$\varphi_{2k'}(t) = \sin(\tfrac{2\pi k'}{K} t)\ (k' = 1, \ldots, (K-1)/2),$$

with $K = 5$. We generate components of $\boldsymbol{\vartheta}_\ell^\star$ by a uniform distribution $[-0.5, 0.5]$.

b) (Piecewise linear signal) We define $\varphi_1(t) = 1$ and $\varphi_2(t) = t$. We set $\boldsymbol{\vartheta}_\ell^\star$ so that $\sum_{k=1}^{K} [\boldsymbol{\vartheta}_\ell^\star]_k \varphi_k(t)$ interpolates $(n_{\ell-1}^\star, h_{1,\ell})$ and $(n_\ell^\star, h_{2,\ell})$ for $h_{1,\ell}$ and $h_{2,\ell}$ generated from a uniform distribution $[-1, 1]$.

We set $N = 250$, $(n_1^\star, \ldots, n_L^\star) = (40, 70, 120, 190)$, $(\sigma_1^\star, \ldots, \sigma_{L+1}^\star) = (1, 100, 10, 5, 50)$, and generate $(\varepsilon_n)_{n=1}^N$ as the white Gaussian noise.

Algorithm 2: Solver for proposed formulation (10)

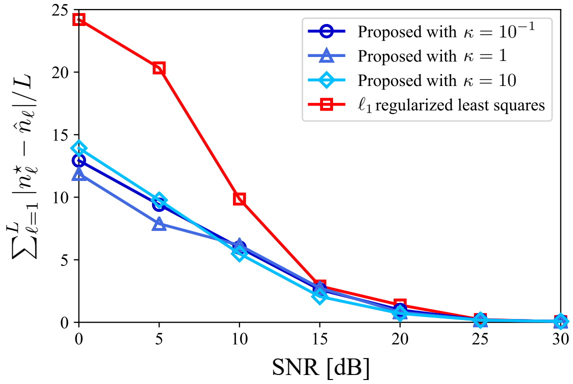
Input: $\beta > 0, \mu \in (0,1), \boldsymbol{s}_0, \boldsymbol{\sigma}_0, \boldsymbol{u}_0, \boldsymbol{v}_0, \boldsymbol{\eta}_0, \boldsymbol{\zeta}_0$
while *a stopping criterion is not satisfied* **do**
- $\bar{\boldsymbol{s}}_{j+1} = \boldsymbol{s}_j - \frac{1}{\beta^2} \boldsymbol{W}^\top (\boldsymbol{W}\boldsymbol{s}_j - \boldsymbol{u}_j) + \frac{1}{\beta} \boldsymbol{W}^\top \boldsymbol{\eta}_j$
- $\bar{\boldsymbol{\sigma}}_{j+1} = \boldsymbol{\sigma}_j - \frac{1}{\beta^2} \boldsymbol{D}^\top (\boldsymbol{D}\boldsymbol{\sigma}_j - \boldsymbol{v}_j) + \frac{1}{\beta} \boldsymbol{D}^\top \boldsymbol{\zeta}_j$
- $\bar{\boldsymbol{u}}_{j+1} = (1 - \frac{1}{\beta^2}) \boldsymbol{u}_j + \frac{1}{\beta^2} \boldsymbol{W}\boldsymbol{s}_j - \frac{1}{\beta} \boldsymbol{\eta}_j$
- $\bar{\boldsymbol{v}}_{j+1} = (1 - \frac{1}{\beta^2}) \boldsymbol{v}_j + \frac{1}{\beta^2} \boldsymbol{D}\boldsymbol{\sigma}_j - \frac{1}{\beta} \boldsymbol{\zeta}_j$
- **for** $n = 1, \ldots, N$ **do**
  - $([\tilde{\boldsymbol{s}}_{j+1}]_n, [\tilde{\boldsymbol{\sigma}}_{j+1}]_n) = \mathrm{prox}_{f_n}([\bar{\boldsymbol{s}}_{j+1}]_n, [\bar{\boldsymbol{\sigma}}_{j+1}]_n)$
- $\tilde{\boldsymbol{u}}_{j+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}(\bar{\boldsymbol{u}}_{j+1})$
- $\tilde{\boldsymbol{v}}_{j+1} = \mathrm{prox}_g(\bar{\boldsymbol{v}}_{j+1})$
- /* *See* (11), (12) *and* (13) *for computation of* prox */
- $\tilde{\boldsymbol{\eta}}_{j+1} = \boldsymbol{\eta}_j - \frac{1}{\beta}(\boldsymbol{W}\tilde{\boldsymbol{s}}_{j+1} - \tilde{\boldsymbol{u}}_{j+1})$
- $\tilde{\boldsymbol{\zeta}}_{j+1} = \boldsymbol{\zeta}_j - \frac{1}{\beta}(\boldsymbol{D}\tilde{\boldsymbol{\sigma}}_{j+1} - \tilde{\boldsymbol{v}}_{j+1})$
- $(\boldsymbol{s}_{j+1}, \ldots, \boldsymbol{\zeta}_{j+1}) = (1-\mu)(\boldsymbol{s}_j, \ldots, \boldsymbol{\zeta}_j) + \mu(\tilde{\boldsymbol{s}}_{j+1}, \ldots, \tilde{\boldsymbol{\zeta}}_{j+1})$
- $j \leftarrow j + 1$

The proposed method first estimates $\boldsymbol{s}^\star$ by Algorithm 2 solving (10) with $g(\cdot) = \kappa\|\cdot\|_1$, and then estimates $n_1^\star, \ldots, n_L^\star$ by Algorithm 1. For construction of $\boldsymbol{W}$ shown in Lemma 1, since there exists an arbitrariness of scalar multiplication in $\boldsymbol{w}_n \in \mathcal{N}(\boldsymbol{\Phi}_n^\top) \backslash \{\boldsymbol{0}\}$, as a typical example, we here use $\boldsymbol{w}_n$ of unit Euclidean norm $\|\boldsymbol{w}_n\| = 1$ $(n = 1, \ldots, N-K)$. In Algorithm 2, β is set to be larger than the maximum singular value of $\begin{pmatrix} \boldsymbol{W}, & \boldsymbol{O}, & -\boldsymbol{I}, & \boldsymbol{O} \\ \boldsymbol{O}, & \boldsymbol{D}, & \boldsymbol{O}, & -\boldsymbol{I} \end{pmatrix}$, $\mu = 1 - 10^{-4}$, and iteration is terminated when the difference between successive iterates is below the threshold 10^{-5}. Meanwhile, to simply measure the error between $\hat{n}_1, \ldots \hat{n}_{\hat{L}}$ and $n_1^\star, \ldots, n_L^\star$, it is beneficial to have $\hat{L} = L$. Thus, we here modify Algorithm 1 by neglecting components of $\boldsymbol{W}\hat{\boldsymbol{s}}$ whose values are small enough. More precisely, we choose $(\hat{n}_1, \ldots \hat{n}_L) \in \{1, \ldots, N-K\}^L$ by

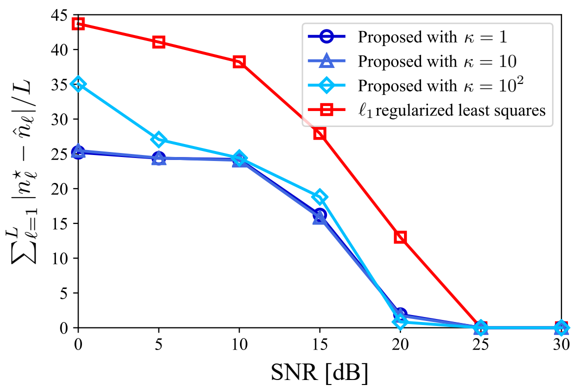
$$\min_{\hat{n}_1 < \cdots < \hat{n}_L} \sum_{n \notin \bigcup_{\ell=1}^{L} \{\hat{n}_\ell - K + 1, \ldots, \hat{n}_\ell\}} \left|[\boldsymbol{W}\hat{\boldsymbol{s}}]_n\right|.$$

Note that, since $\hat{\boldsymbol{s}}$ is given at this step, this minimization can be exactly done in a greedy way. Then, we measure the error by $\sum_{\ell=1}^{L} |n_\ell^\star - \hat{n}_\ell|/L$.

In Fig. 1 where the results are averaged over 100 realizations of noises, we show the estimation accuracy of the proposed formulation (4) with (8) for several choices of κ and the ℓ_1 regularized least squares formulation which corresponds to setting $\psi(\boldsymbol{s}) = \sum_{n=1}^{N} (y_n - [\boldsymbol{s}]_n)^2$ in (4). Note that λ, the weight of $\|\boldsymbol{W}\boldsymbol{s}\|_1$, is chosen so that the performance becomes best, independently for the proposed formulation with each choice of κ and the ℓ_1 regularized least squares



(a) Results for piecewise sinusoidal signals.



(b) Results for piecewise linear signals.

Fig. 1: Estimation accuracy of change points by the proposed formulation (4) with (8) and the ℓ_1 regularized least squares under noise level variation. The results are averaged over 100 realizations of noises for each SNR.

formulation. The results show that the proposed data-fidelity function improves the performance under the variation of the noise level, in particular for lower SNRs. In addition, it can also be seen that the performance of the proposed method is not sensitive against the choice of κ.

V. Conclusion

In this paper, we presented a convex optimization based piecewise linear regression method which can cope with variation of the noise level shown in (1). To this end, we design a convex data-fidelity function in (8) which mitigates effect of the noise level variation by adjusting the weights of approximation errors. Numerical examples demonstrate the effectiveness of the proposed method for situations where the noise level varies in the observed signal.

References

[1] F. Gustafsson, *Adaptive filtering and change detection*. Wiley, 2000.
[2] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 36, no. 1, pp. 29–40, Jan 1988.
[3] T. Takanami and G. Kitagawa, "A new efficient procedure for the estimation of onset times of seismic waves," *J. Physics of the Earth*, vol. 36, no. 6, pp. 267–290, 1988.
[4] T. D. Popescu, "Detection and diagnosis of model parameter and noise variance changes with application in seismic signal processing," *Mech. Syst. and Signal Process.*, vol. 25, no. 5, pp. 1598–1616, 2011.
[5] H. Kuroda, J. Ogata, H. Sakanashi, and T. Uchide, "Seismic phase picking based on sparsity-aware change detection," in *Proc. Seismological Society of Japan Fall Meeting*, Sept. 2019, (in Japanese).
[6] Z. Han, H. Chen, T. Yan, and G. Jiang, "Time series segmentation to discover behavior switching in complex physical systems," in *Proc. IEEE ICDM*, Nov. 2015, pp. 161–170.
[7] R. Sevlian and R. Rajagopal, "Detection and statistics of wind power ramps," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 3610–3620, Nov. 2013.
[8] G. C. Amaral, J. D. Garcia, L. E. Y. Herrera, G. P. Temporão, P. J. Urban, and J. P. Weid, "Automatic fault detection in WDM-PON with tunable photon counting OTDR," *J. Lightwave Technology*, vol. 33, no. 24, pp. 5025–5031, Dec. 2015.
[9] L. Aydin and A. Polydoros, "Hop-timing estimation for FH signals using a coarsely channelized receiver," *IEEE Trans. Commun.*, vol. 44, no. 4, pp. 516–526, April 1996.
[10] D. J. Torrieri, "Mobile frequency-hopping CDMA systems," *IEEE Trans. Commun.*, vol. 48, no. 8, pp. 1318–1327, Aug. 2000.
[11] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald, "Bayesian curve fitting using MCMC with applications to signal segmentation," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 747–758, March 2002.
[12] N. Dobigeon, J.-Y. Tourneret, and M. Davy, "Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1251–1263, April 2007.
[13] A. Willsky and H. Jones, "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems," *IEEE Trans. Automat. Ctrl.*, vol. 21, no. 1, pp. 108–112, Feb. 1976.
[14] U. Appel and A. V. Brandt, "Adaptive sequential segmentation of piecewise stationary time series," *Information Sciences*, vol. 29, no. 1, pp. 27–56, 1983.
[15] G. Almpanidis, M. Kotti, and C. Kotropoulos, "Robust detection of phone boundaries using model selection criteria with few observations," *IEEE Trans. Audio, Speech & Language Process.*, vol. 17, no. 2, pp. 287–298, Feb 2009.
[16] H. Ohlsson, L. Ljung, and S. Boyd, "Segmentation of ARX-models using sum-of-norms regularization," *Automatica*, vol. 46, no. 6, pp. 1107–1111, 2010.
[17] D. Angelosante and G. B. Giannakis, "Group lassoing change-points in piecewise-constant AR processes," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, March 2012.
[18] N. Ozay, M. Sznaier, C. M. Lagoa, and O. I. Camps, "A sparsification approach to set membership identification of switched affine systems," *IEEE Trans. Automat. Ctrl.*, vol. 57, no. 3, pp. 634–648, March 2012.
[19] R. Giryes, M. Elad, and A. M. Bruckstein, "Sparsity based methods for overparameterized variational problems," *SIAM J. Imag. Sci.*, vol. 8, no. 3, pp. 2133–2159, 2015.
[20] H. Kuroda, M. Yamagishi, and I. Yamada, "Exploiting sparsity in tight-dimensional spaces for piecewise continuous signal recovery," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6363–6376, Dec. 2018.
[21] P. J. Huber, *Robust Statistics*. Wiley, 1981.
[22] A. Antoniadis, "Comments on: ℓ_1-penalization for mixture regression models," *TEST*, vol. 19, no. 2, pp. 257–258, Aug 2010.
[23] T. Sun and C.-H. Zhang, "Scaled sparse linear regression," *Biometrika*, vol. 99, no. 4, pp. 879–898, Sept. 2012.
[24] P. L. Combettes and C. L. Müller, "Perspective maximum likelihood-type estimation via proximal decomposition," *Electron. J. Statist.*, vol. 14, no. 1, pp. 207–238, 2020.
[25] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators," *Adv. Comput. Math.*, vol. 38, no. 3, pp. 667–681, 2013.
[26] L. Condat, "A primal-dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms," *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 460–479, 2013.
[27] J. Yang and X. Yuan, "Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization," *Math. Comput.*, vol. 82, no. 281, pp. 301–329, 2013.
[28] M. Yamagishi and I. Yamada, "Nonexpansiveness of a linearized augmented Lagrangian operator for hierarchical convex optimization," *Inverse Problems*, vol. 33, no. 4, p. 044003, March 2017.
[29] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Programm.*, vol. 55, no. 1, pp. 293–318, April 1992.