analytical chemistry

pubs.acs.org/ac Article

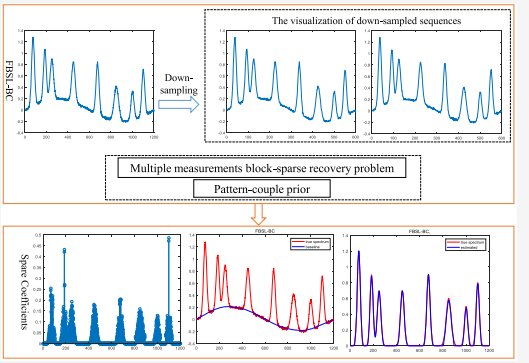
Fast Burst-Sparsity Learning-Based Baseline Correction (FBSL-BC) Algorithm for Signals of Analytical Instruments

Haoran Li, Suyi Chen, Jisheng Dai,* Xiaobo Zou,* Tao Chen, Tianhong Pan, and Melvin Holmes

Cite This: *Anal. Chem.* 2022, 94, 5113−5121 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Baseline correction is a critical step for eliminating the interference of baseline drift in spectroscopic analysis. The recently proposed sparse Bayesian learning (SBL)-based method can significantly improve the baseline correction performance. However, it has at least two disadvantages: (i) it works poorly for large-scale datasets and (ii) it completely ignores the burst-sparsity structure of the sparse representation of the pure spectrum. In this paper, we present a new fast burst-sparsity learning method for baseline correction to overcome these shortcomings. The first novelty of the proposed method is to jointly adopt a down-sampling strategy and construct a multiple measurements block-sparse recovery problem with the down-sampling sequences. The down-sampling strategy can significantly reduce the dimension of the spectrum; while jointly exploiting the block sparsity among the down-sampling sequences avoids losing the information contained in the original spectrum. The second novelty of the proposed method is introducing the pattern-coupled prior into the SBL framework to characterize the inherent burst-sparsity in the sparse representation of spectrum. As illustrated in the paper, burst-sparsity commonly occurs in peak zones with more denser nonzero coefficients. Properly utilizing such burst-sparsity can further enhance the baseline correction performance. Results on both simulated and real datasets (such as FT-IR, Raman spectrum, and chromatography) verify the substantial improvement, in terms of estimation accuracy and computational complexity.



1. INTRODUCTION

The combination of spectroscopy technology and chemometrics provides a reliable analytical technique for qualitative and quantitative analysis, especially in analytical chemistry and nondestructive testing.[1] However, baselines are usually superimposed on the pure spectrum, because of the instrumental measurement effects and disturbance during the spectral recording process. Such a baseline shift could change the essential characteristics of feature peaks (e.g., position, strength, and slope). As a consequence, the collected spectral signals are not strictly adhering to Beer−Lambert Law, which will bring a negative impact on the following qualitative and quantitative analysis.[2,3] To avoid the interference of baseline shift, some relevant baseline correction algorithms are required in spectroscopic analysis, either to enhance feature signals with weak intensities or guarantee the accuracy and stability of the calibration model.

In the past decades, numerous methodologies have been introduced to obtain better baseline correction results, such as derivative,[4] frequency analysis,[5] and curve fitting.[6] In the derivative methods, the first derivative was used to eliminate the constant offsets from the spectra and then the linear drifts can be removed by the second derivative method,[7,8] but the derivative process is usually inadequate for baseline removal and would change the shapes of the feature peak, which, in return, hampers the interpretation of the preprocessed spectral signals. Assuming the baseline and pure spectrum can be respectively captured in low and high frequency,[9] the frequency analysis aims to separate the spectra into the different frequency domains by removing the low-frequency components and reconstructing the spectra using wavelet coefficients.[10,11] However, such precise frequency separation is usually invalid in practice and will cause essential information loss.[12] Since the baseline always exhibits a smooth curve, formulating the baseline correction problem as a curve-fitting problem is another widely used strategy. Two most commonly used methods are polynomial fitting[13] and penalized least-squares.[14] The traditional polynomial fitting is to fit manual-selected spectral location points, but it is quite time-consuming and substantially relies on the user's expertise.[15] Therefore, some modified polynomial fitting methods were developed to realize the automatic selection of spectral locations points or estimate the baseline iteratively.[16,17] However, any improperly chosen order of the polynomial will result in an overfitting or

Received: December 15, 2021
Accepted: March 9, 2022
Published: March 18, 2022

https://doi.org/10.1021/acs.analchem.1c05443
Anal. Chem. 2022, 94, 5113−5121

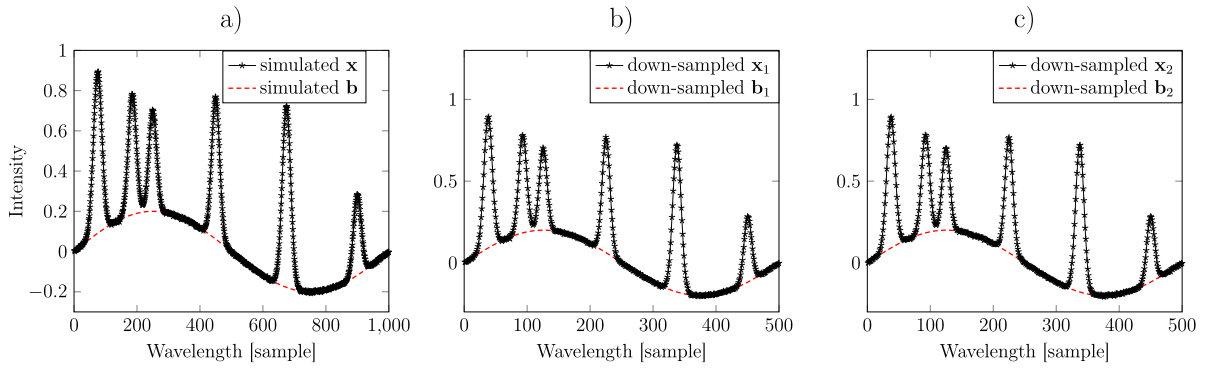


Figure 1. (a) Original simulated spectrum and baseline, (b) resampled simulated spectrum, and (c) resampled baseline.

under-fitting problem. Penalized least-squares (PLS) is the most popular strategy for baseline correction. After Eilers presented the algorithm called asymmetric least-squares (AsLS),[14] a large number of extended PLS methods were proposed in the literature, e.g., asymmetrically reweighed penalized least-squares (arPLS),[18] adaptive iteratively reweighted penalized least-squares (airPLS),[9] morphological weighted penalized least-squares (MPLS),[19] and improved asymmetric least-squares (iasLS).[20] However, these methods require tuning some user-defined parameters, which could be very time-consuming. Moreover, they also suffer from performance loss with a strong background noise.[21]

Recently, compressed sensing theory has brought a new conception for data analysis,[22,23] and sparse recovery technique has been widely applied in variable selection,[24] spectral calibration,[25] and baseline correction.[26] Ning et al. presented a baseline estimation and denoising method by using the sparse representation (BEADS), which aims to formulate the feature peaks with sparse derivatives,[27] but it is limited to the scenario that the first and last elements of the spectrum share the same value. Han et al. proposed a simultaneous spectrum fitting and baseline correction method using sparse representation (SSFBCSP), which formulates the sparse representation of spectral peak by Voigt-like or Gaussian redundant dictionaries, and solves the problem using the matching pursuits method.[28] Although the SSFBCSP method can achieve satisfactory performance for baseline correction, the adopted l_1-norm approximation could bring a performance loss and it is challenging to select the regularization terms across different datasets properly.[21,29] To tackle the drawbacks of the SSFBCSP, Li et al. proposed a sparse Bayesian learning-based baseline correction approach (SBL-BC). Thanks to the inherent learning capability of the Bayesian framework, SBL-BC does not require to manually select any regularization terms. Simulations and real dataset experiments illustrated that SBL-BC can significantly improve the baseline correction performance. Nevertheless, SBL-BC still has the following disadvantages: (i) it would poorly works for large-scale datasets, since it is computationally intractable to calculate the inverse of a large matrix iteratively; and (ii) the sparse representation of pure spectrum exhibits a burst-sparsity structure in practical implementations, which can be exploited to enhance the baseline correction performance, but SBL-BC completely ignores such structures.

To overcome the aforementioned shortcomings of SBL-BC, we present a new fast burst-sparsity learning method for baseline correction. The key idea of the proposed method is to adopt a down-sampling strategy to reduce the dimension of the spectrum. In the general case, the down-sampling process will lose information contained in the original spectrum. However, we notice that the down-sampled sequences have very similar peaks in the same position (as will be shown in Figure 1). In other words, they will share the same sparse representation under some lineshapes. Consequently, we can construct a multiple measurements block-sparse (or row-sparse) recovery problem with the down-sampled sequences. Therefore, there is almost no information loss for the new proposed method. Recall that the sparse representation of spectrum actually exhibits a burst-sparsity structure, which is in the form of the nonzero coefficients occurring in clusters.[30] How to explicitly take this burst-sparsity structure into account becomes an essential step to enhance the spares recovery performance.[31] To this end, we further introduce the pattern-coupled prior into the multiple measurements SBL framework to characterize the burst sparsity.[32] Experimental results on both simulated and real datasets reveal the substantial performance improvement, in terms of estimate accuracy and computational complexity.

The rest of the paper is organized as follows. In Section 2, we first briefly review the data model and SBL-BC, and presents the construct of the multiple measurements block sparsity recovery problem with a down-sampling step and introduces the pattern-coupled prior to exploit the burst sparsity. The effectiveness of the proposed method will be illustrated through simulation and real dataset examples in Section 3. Finally, the research is concluded in Section 4.

2. EXPERIMENTAL SECTION

2.1. Data Model and Review of SBL-BC. Considering that the observed spectrum vector $\mathbf{x} \in \mathbb{R}^N$ is in the form of

$$\mathbf{x} = \mathbf{s} + \mathbf{b} + \mathbf{n} \tag{1}$$

where $\mathbf{s} \in \mathbb{R}^N$ denotes the pure spectrum vector, $\mathbf{b} \in \mathbb{R}^N$ denotes the baseline vector, and $\mathbf{n} \in \mathbb{R}^N$ denotes the noise.

According to refs 33 and 34, the pure spectrum $\mathbf{s}$ has only a few peaks, and thus we can sparsely represent it with some line shape dictionaries. For example, let $\mathbf{A}_N \in \mathbb{R}^{N\times N}$ be a Gaussian line shape dictionary, whose (i, j) entry $a_{i,j}$ is

$$a_{i,j} = \exp\left(-\frac{(i-j)^2}{2\sigma_j^2}\right) \tag{2}$$

where σ_j^2 represents an unknown variance that controls the width of the Gaussian line shape of the jth column ($\mathbf{a}_j$). For a

https://doi.org/10.1021/acs.analchem.1c05443
Anal. Chem. 2022, 94, 5113−5121

set of properly chosen σ_j^2 values, $\mathbf{s}$ can be approximately written as reported in ref 28:

$$\mathbf{s} \approx \mathbf{A}_N\mathbf{w} \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^N$ is the sparse coefficient vector with nonzero elements lying in the peak zones of $\mathbf{s}$. Substituting eq 3 into eq 1, we have

$$\mathbf{x} = \mathbf{A}_N\mathbf{w} + \mathbf{b} + \mathbf{n} \tag{4}$$

The sparse solution to eq 4 can be obtained by using the l_1-norm minimization, but the existing methods typically suffer from performance loss, because the restricted isometry property (RIP) cannot be guaranteed in practice and it is challenging to select the regularization terms across different datasets properly.[29] To overcome the shortcomings of the existing methods, our previous work provided SBL-BC for joint pure spectrum fitting and baseline correction with no regularization terms being involved.[21] In the following, we briefly review SBL-BC from the Bayesian perspective.

Assume that $\mathbf{n}$ follows an independently identically distribution (i.i.d.) zero-mean Gaussian distribution; then, we have

$$p(\mathbf{x}|\mathbf{w}, \mathbf{b}, \alpha) = \mathcal{N}(\mathbf{x}|\mathbf{A}_N\mathbf{w} + \mathbf{b}, \alpha^{-1}\mathbf{I}) \tag{5}$$

where α denotes the noise precision, $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathbf{I}$ is the identity matrix of an appropriate size. For ease of Bayesian inference, α is usually assigned as a Gamma distribution, i.e.,

$$p(\alpha) = \Gamma(\alpha|e, f) \tag{6}$$

where e and f are some small positive constants. Note that the baseline $\mathbf{b}$ varies much more slowly than $\mathbf{s}$. Left-multiplying $\mathbf{b}$ by the second-order derivative matrix $\mathbf{D}_N \in \mathbb{R}^{(N-2)\times N}$, we can obtain an approximate zero vector. To characterize this property, SBL-BC modeled $\mathbf{D}_N\mathbf{b}$ as an i.i.d. Gaussian matrix with zero mean and a precision β, i.e.,

$$p(\mathbf{b}|\beta) = \mathcal{N}(\mathbf{D}_N\mathbf{b}|\mathbf{0}, \beta^{-1}\mathbf{I}) \tag{7}$$

where $\mathbf{0}$ is a zero vector with appropriate size. Similarly, β is constructed as a Gamma distribution, i.e.,

$$p(\beta) = \Gamma(\beta|e, f) \tag{8}$$

According to the Bayesian model,[35] we can assign the sparse vector $\mathbf{w}$ as an i.i.d. Gaussian prior distribution with a precision vector $\boldsymbol{\gamma} = \{\gamma_j\}_{j=1}^N$, i.e.,

$$p(\mathbf{w}|\boldsymbol{\gamma}) = \prod_{j=1}^{N} \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathrm{diag}\{\boldsymbol{\gamma}\}^{-1}) \tag{9}$$

and $\boldsymbol{\gamma}$ is also assigned as a Gamma distribution:

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{N} \Gamma(\gamma_j|e, f) \tag{10}$$

Hence, we can formulate the joint distribution $p(\mathbf{x}, \mathbf{w}, \alpha, \boldsymbol{\gamma}, \beta, \mathbf{b})$ as

$$p(\mathbf{x}, \mathbf{w}, \alpha, \boldsymbol{\gamma}, \beta, \mathbf{b}) = p(\mathbf{x}|\mathbf{w}, \mathbf{b}, \alpha)p(\mathbf{w}|\boldsymbol{\gamma})p(\mathbf{b}|\beta)p(\boldsymbol{\gamma})p(\beta)p(\alpha) \tag{11}$$

Since it is intractable to directly infer the posterior distribution $p(\mathbf{w}, \alpha, \boldsymbol{\gamma}, \beta, \mathbf{b}|\mathbf{x})$ from eq 11, SBL-BC aims to approximately calculate the posterior distribution via VBI.[36] Although SBL-BC can tackle the shortcomings of the existing baseline correction algorithm, it still has some disadvantages: (i) it suffers from high computational complexity, especially for large-scale datasets, because it requires iteratively calculating the inverse of a large matrix whose size is $N \times N$; and (ii) it ignores the burst sparsity of $\mathbf{w}$, which actually can be exploited to enhance the baseline correction performance.

In the rest of this section, we present the new fast burst-sparsity learning method for the baseline correction. We begin by adopting the down-sampling approach to reduce the dimension of the dictionary matrix. Then, we introduce the pattern-coupled prior into the SBL framework to exploit the burst sparsity of $\mathbf{w}$. Note that we will reuse some notations, e.g., α, β, and $\boldsymbol{\gamma}$, as described previously.

2.2. Down-Sampling. For a large-scale dataset, the dimension of the observed spectrum vector would be very large. Fortunately, resampling can reduce the dimension of original data and retain the original structure. Specifically, let G be an integer factor and assume $N = GL$. Then, we can down-sample $\mathbf{x}$ into G sequences, i.e.,

$$\mathbf{x}_g = \mathbf{s}_g + \mathbf{b}_g + \mathbf{n}_g \qquad g = 1, 2, ..., G \tag{12}$$

where $\mathbf{x}_g = [x_g, x_{g+G}, ..., x_{g+(L-1)G}]^T$; the same is observed for $\mathbf{s}_g$, $\mathbf{b}_g$, and $\mathbf{n}_g$. It can be demonstrated that such down-sampling produces a very good approximation of the sequence by sampling the signal at a lower rate. For example, Figure 1a shows a noise-free simulated spectrum vector $\mathbf{x}$ of size 1000 × 1 and the corresponding baseline $\mathbf{b}$; meanwhile, Figures 1b and 1c show the down-sampling sequences $\mathbf{x}_1$ and $\mathbf{x}_2$ with an integer factor $G = 2$, as well as the down-sampled baseline $\mathbf{b}_1$ and $\mathbf{b}_2$, respectively. As shown in Figure 1, the down-sampling sequences $\mathbf{x}_1$ and $\mathbf{x}_2$ have similar peaks in the same positions, and they almost coincide with the original spectrum vector $\mathbf{x}$ up to a scale $G = 2$. So are the original baseline and down-sampled baselines.

Obviously, $\mathbf{x}_g$s have similar sparse representations. Hence, we are able to construct a multiple measurements block-sparse (or row-sparse) recovery problem as follows. Using eq 3, eq 12 can be rewritten as

$$\mathbf{x}_g = \mathbf{A}_L\mathbf{w}_g + \mathbf{b}_g + \mathbf{n}_g \qquad g = 1, 2, ..., G \tag{13}$$

Stacking all $\mathbf{x}_g$ terms into a matrix, we have

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_G] = \mathbf{A}_L\mathbf{W} + \mathbf{B} + \mathbf{N} \tag{14}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_G]$, and $\mathbf{B}$ and $\mathbf{N}$ can be expressed in a similar fashion. Since $\mathbf{w}_g$ terms share the same support, $\mathbf{W}$ becomes a row-sparse matrix. Moreover, the elements of $\mathbf{B}$ in the same row have very similar values.

2.3. Pattern-Coupled Prior. Recall that the σ_l^2 terms represent the unknown variances controlling the width of the Gaussian lineshapes in $\mathbf{A}_L$. SBL-BC tries to refine σ_l^2 s with the expectation-maximization (EM) algorithm. Although the baseline correction performance can be significantly improved by using such variance refinement, it is impossible to obtain the optimal σ_l^2 value exactly. Even worse, wide or irregular peaks in the pure spectrum need several nearby lineshapes to approximate. Hence, the sparse vectors $\mathbf{w}$ and $\mathbf{w}_g$s exhibit a burst-sparsity property, which can be exploited to further

https://doi.org/10.1021/acs.analchem.1c05443
Anal. Chem. 2022, 94, 5113−5121

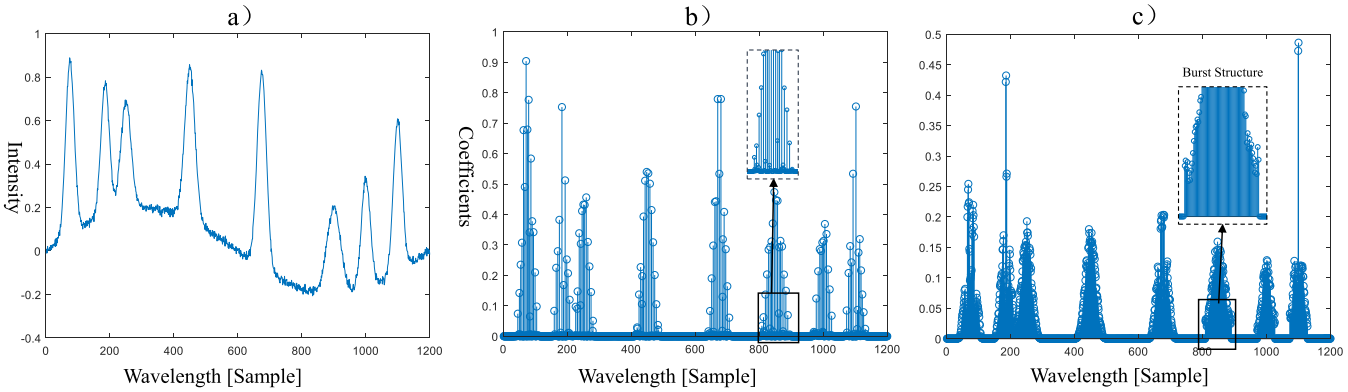


Figure 2. (a) Simulated spectrum with sinusoidal baseline, (b) simulated spectrum of the sparse vector **w** obtained by SBL-BC methods, and (c) simulated spectrum of the burst-sparsity based SBL.

enhance the baseline correction performance. For example, Figure 2 shows the coefficients of sparse vectors with SBL-BC and burst-sparsity-based SBL methods. It can be seen that denser significant elements (nonzero coefficients) appear in peak zones using the burst-sparsity-based SBL method, and the mean square error (NMS) values between estimated spectrum and simulated spectrum are 0.0232 (SBL-BC) and 0.0196 (burst-sparsity based SBL). Hence, it is nontrivial to capture the burst-sparsity structure of spectrum, which can further enhance the baseline correction performance.

To characterize the burst-sparsity of $\mathbf{w}_g$s, we introduce the pattern-coupled prior in this subsection. Assume that each element of $\mathbf{w}_g$ (denoted as $w_{l,g}$) is controlled not only by its own precision γ_l as shown in eq 9, but also by adjacent precisions γ_{l-1} and γ_{l+1}. In this case, the value of $w_{l,g}$ will be affected by its neighbors $w_{(l-1),g}$ and $w_{(l+1),g}$, to capture the burst sparsity. Note that the precision vector $\boldsymbol{\gamma} = \boldsymbol{\gamma} = [\gamma_1, \gamma_2, ..., \gamma_L]^T$ is independent of the measurement index g. For ease of the Bayesian formulation, we define an assignment vector $\mathbf{h}$ of size $L \times 1$, whose elements take values from $\{-1, 0, 1\}$ only. If $h_l = 0$, it means that $w_{l,g}$ is controlled by its own precision γ_l; otherwise, $w_{l,g}$s are controlled by the left and right adjacent precisions (γ_{l-1} and γ_{l+1}) for $h_l = -1$ and $h_l = 1$, respectively. Hence, the burst-sparsity prior can be written as

$$p(\mathbf{W}|\boldsymbol{\gamma}, \mathbf{h}) = \prod_{l,g=1}^{L,G} \{\mathcal{N}(w_{l,g}|0, \gamma_{l-1}^{-1})\}^{[h_l=-1]} \times \{\mathcal{N}(w_{l,g}|0, \gamma_l^{-1})\}^{[h_l=0]} \{\mathcal{N}(w_{l,g}|0, \gamma_{l+1}^{-1})\}^{[h_l=1]} \tag{15}$$

where $[a = b]$ returns a value of 1 if the quality $a = b$ holds true, and 0 otherwise. The burst-sparsity prior is also known as the pattern-coupled prior, as it enforces the burst-sparsity from the pattern structure.[37] For example, for a small γ_l, the corresponding $x_{l,g}$ and its adjacent $x_{(l-1),g}$ and $x_{(l+1),g}$ could have a significant value with a high probability. Since $\mathbf{h}$ is unknown in practice, we assign it a uniform categorical distribution:

$$p(\mathbf{h}) = \prod_{l,g=1}^{L,G} \left(\frac{1}{3}\right)^{[h_l=-1]} \left(\frac{1}{3}\right)^{[h_l=0]} \left(\frac{1}{3}\right)^{[h_l=1]} \tag{16}$$

Then, the joint distribution $p(\mathbf{X}, \mathbf{W}, \mathbf{h}, \boldsymbol{\gamma}, \mathbf{B}, \alpha, \beta)$ can be factorized by

$$p(\mathbf{X}, \mathbf{W}, \mathbf{h}, \boldsymbol{\gamma}, \mathbf{B}, \alpha, \beta) = p(\mathbf{X}|\mathbf{W}, \mathbf{B}, \alpha)p(\mathbf{W}|\boldsymbol{\gamma}, \mathbf{h})p(\boldsymbol{\gamma})p(\mathbf{h})p(\mathbf{B}|\beta)p(\beta)p(\alpha) \tag{17}$$

where

$$p(\mathbf{X}|\mathbf{W}, \mathbf{B}, \alpha) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{x}_g|\mathbf{A}_L\mathbf{w}_g + \mathbf{b}_g, \alpha^{-1}\mathbf{I}) \tag{18}$$

$$p(\boldsymbol{\gamma}) = \prod_{l}^{L} \Gamma(\gamma_l|e, f) \tag{19}$$

$$p(\mathbf{B}|\beta) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{D}_L\mathbf{b}_g|\mathbf{0}, \beta^{-1}\mathbf{I}) \tag{20}$$

with α and β being the precisions of noise and $\mathbf{B}$, respectively.

2.4. Bayesian Inference via VBI. The down-sampled pure spectrum and baseline can be jointly obtained if the maximum a posteriori (MAP) estimate of $p(\Omega|\mathbf{X})$ can be calculated, where $\Omega = \{\mathbf{W}, \alpha, \boldsymbol{\gamma}, \mathbf{h}, \beta, \mathbf{B}\}$ denotes the hidden variable set. Unfortunately, this MAP estimate is still intractable as in ref 21. Hence, the VBI methodology is performed to approximately factorize $p(\Omega|\mathbf{X})$ by

$$p(\Omega|\mathbf{X}) \approx q(\Omega) = q(\mathbf{W})q(\alpha)q(\boldsymbol{\gamma})q(\mathbf{h})q(\beta)q(\mathbf{B}) \tag{21}$$

There are many feasible factorizations and the optimal one should minimize the Kullback−Leibler (KL) divergence between $q(\Omega)$ and $p(\Omega|\mathbf{X})$. According to ref 36, the optimal solution is

$$\ln\, q^*(\Omega_k) \propto \langle \ln p(\mathbf{X}, \Omega)\rangle_{q_{/k}^*} \qquad k = 1, 2, ..., 6 \tag{22}$$

where Ω_k is the kth element of Ω, $q_{/k}^*$ is short for $\prod_{i\neq k} q^*(\Omega_i)$, and $\propto$ denotes equality up to a constant. Since $q^*(\Omega_k)$ is dependent on other $q^*(\Omega_i)$, $i \neq k$, the closed form for $q^*(\Omega_k)$ is difficult to find. However, the stationary solution can be obtained by following iteratively update strategy:[32]

$$\ln\, q^{(r+1)}(\Omega_1) \propto \langle \ln p(\mathbf{X}, \Omega)\rangle_{q^{(r)}(\Omega_2)q^{(r)}(\Omega_3)q^{(r)}(\Omega_4)q^{(r)}(\Omega_5)q^{(r)}(\Omega_6)} \tag{23}$$

$$\ln\, q^{(r+1)}(\Omega_2) \propto \langle \ln p(\mathbf{X}, \Omega)\rangle_{q^{(r+1)}(\Omega_1)q^{(r)}(\Omega_3)q^{(r)}(\Omega_4)q^{(r)}(\Omega_5)q^{(r)}(\Omega_6)} \tag{24}$$

https://doi.org/10.1021/acs.analchem.1c05443
Anal. Chem. 2022, 94, 5113−5121

$$\ln\, q^{(r+1)}(\Omega_3) \propto \langle \ln p(\mathbf{X}, \Omega)\rangle_{q^{(r+1)}(\Omega_1)q^{(r+1)}(\Omega_2)q^{(r)}(\Omega_4)q^{(r)}(\Omega_5)q^{(r)}(\Omega_6)} \tag{25}$$

$$\ln\, q^{(r+1)}(\Omega_4) \propto \langle \ln p(\mathbf{X}, \Omega)\rangle_{q^{(r+1)}(\Omega_1)q^{(r+1)}(\Omega_2)q^{(r+1)}(\Omega_3)q^{(r)}(\Omega_5)q^{(r)}(\Omega_6)} \tag{26}$$

$$\ln\, q^{(r+1)}(\Omega_5) \propto \langle \ln p(\mathbf{X}, \Omega)\rangle_{q^{(r+1)}(\Omega_1)q^{(r+1)}(\Omega_2)q^{(r+1)}(\Omega_3)q^{(r+1)}(\Omega_4)q^{(r)}(\Omega_6)} \tag{27}$$

$$\ln\, q^{(r+1)}(\Omega_6) \propto \langle \ln p(\mathbf{X}, \Omega)\rangle_{q^{(r+1)}(\Omega_1)q^{(r+1)}(\Omega_2)q^{(r+1)}(\Omega_3)q^{(r+1)}(\Omega_4)q^{(r+1)}(\Omega_5)} \tag{28}$$

where $(\cdot)^{(r)}$ denotes the rth iteration. Extending the derivation for SBL-BC, we can obtain the update solutions with Lemma 1.

Lemma 1. The optimal solutions to eqs 23−28 are given as follows:

$$q^{(r+1)}(\mathbf{W}) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{w}_g|\hat{\mathbf{w}}_g^{(r+1)}, \boldsymbol{\Sigma}^{(r+1)}) \tag{29}$$

$$q^{(r+1)}(\alpha) = \Gamma\left(\alpha\middle| e + \frac{1}{2}\mathrm{LG}, f_\alpha^{(r+1)}\right) \tag{30}$$

$$q^{(r+1)}(\boldsymbol{\gamma}) = \prod_{l=1}^{L} \Gamma(\gamma_l|e_l^{(r+1)}, f_l^{(r+1)}) \tag{31}$$

$$q^{(r+1)}(\mathbf{h}) = \prod_{l=1}^{L}\prod_{k\in\{-1,0,1\}} \hat{h}_{l,k}^{(r+1)}\delta(h_l - k) \tag{32}$$

$$q^{(r+1)}(\beta) = \Gamma\left(\beta\middle| e + \frac{1}{2}\mathrm{LG}, f_\beta^{(r+1)}\right) \tag{33}$$

$$q^{(r+1)}(\mathbf{B}) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{b}_g|\hat{\mathbf{b}}_g^{(r+1)}, \boldsymbol{\Sigma}_b^{(r+1)}) \tag{34}$$

where

$$\boldsymbol{\Sigma}^{(r+1)} = \left(\hat{\alpha}^{(r)}\mathbf{A}_L^{\mathrm{T}}\mathbf{A}_L + \sum_{k\in\{-1,0,1\}} \boldsymbol{\Psi}_k^{(r)}\boldsymbol{\Lambda}_k^{(r)}\right)^{-1} \tag{35}$$

$$\hat{\mathbf{w}}_g^{(r+1)} = \hat{\alpha}^{(r)}\boldsymbol{\Sigma}^{(r+1)}\mathbf{A}_L^{\mathrm{T}}(\mathbf{x}_g - \hat{\mathbf{b}}_g^{(r)}) \tag{36}$$

$$f_\alpha^{(r+1)} = f + \frac{1}{2}\sum_{g=1}^{G}\left(\left\|\mathbf{x}_g - \mathbf{A}_L\hat{\mathbf{w}}_g^{(r+1)} - \hat{\mathbf{b}}_g^{(r)}\right\|_2^2 + \mathrm{tr}\{\mathbf{A}_L\boldsymbol{\Sigma}^{(r+1)}\mathbf{A}_L^{\mathrm{T}} + \boldsymbol{\Sigma}_b^{(r)}\}\right) \tag{37}$$

$$\hat{\alpha}^{(r+1)} = \langle\alpha\rangle_{q^{(r+1)}(\alpha)} = \frac{\left(e + \frac{1}{2}\mathrm{LG}\right)}{f_\alpha^{(r+1)}} \tag{38}$$

$$e_l^{(r+1)} = e + \frac{G}{2}(\hat{h}_{l+1,-1}^{(r)} + \hat{h}_{l,0}^{(r)} + \hat{h}_{l-1,1}^{(r)}) \tag{39}$$

$$f_l^{(r+1)} = f + \frac{1}{2}\sum_{g=1}^{G}(\hat{h}_{l+1,-1}^{(r)}\varpi_{l+1,g}^{(r+1)} + \hat{h}_{l,0}^{(r)}\varpi_{l,g}^{(r+1)} + \hat{h}_{l-1,1}^{(r)}\varpi_{l-1,g}^{(r+1)}) \tag{40}$$

$$\hat{\gamma}_l^{(r+1)} = \langle\gamma_l\rangle_{q^{(r+1)}(\gamma_l)} = \frac{e_l^{(r+1)}}{f_l^{(r+1)}} \tag{41}$$

$$\hat{h}_{l,k}^{(r+1)} = \frac{\exp(\nu_{l,k}^{(r+1)})}{\sum_{k\in\{-1,0,1\}}\exp(\nu_{l,k}^{(r+1)})} \tag{42}$$

$$f_\beta^{(r+1)} = f + \frac{1}{2}\sum_{g=1}^{G}(\left\|\mathbf{D}_L\hat{\mathbf{b}}_g^{(r)}\right\|^2 + \mathrm{tr}\{\mathbf{D}_L\boldsymbol{\Sigma}_b^{(r)}\mathbf{D}_L^{\mathrm{T}}\}) \tag{43}$$

$$\hat{\beta}^{(r+1)} = \langle\beta\rangle_{q^{(r+1)}(\beta)} = \frac{\left(e + \frac{1}{2}LG\right)}{f_\beta^{(r+1)}} \tag{44}$$

$$\boldsymbol{\Sigma}_b^{(r+1)} = (\hat{\alpha}^{(r+1)}\mathbf{I} + \hat{\beta}^{(r+1)}\mathbf{D}_L^{\mathrm{T}}\mathbf{D}_L)^{-1} \tag{45}$$

$$\hat{\mathbf{b}}_g^{(r+1)} = \hat{\alpha}^{(r+1)}\boldsymbol{\Sigma}_b^{(r+1)}(\mathbf{x}_g - \mathbf{A}_L\hat{\mathbf{w}}_g^{(r+1)}) \tag{46}$$

with $\boldsymbol{\Psi}_k^{(r)} = \mathrm{diag}\{\hat{h}_{1,k}^{(r)}, \hat{h}_{2,k}^{(r)}, ..., \hat{h}_{L,k}^{(r)}\}$, $\boldsymbol{\Lambda}_k^{(r)} = \mathrm{diag}\{\hat{\gamma}_{1+k}^{(r)}, \hat{\gamma}_{2+k}^{(r)}, ..., \hat{\gamma}_{L+k}^{(r)}\}$, $\varpi_{l,g}^{(r+1)} = \langle w_{l,j}^2\rangle_{q^{(r+1)}(w_j)} = [\hat{w}_{l,g}^{(r+1)}]_j^2 + [\boldsymbol{\Sigma}^{(r+1)}]_{l,l}$, $\nu_{l,k}^{(r+1)} = \frac{G}{2}(\widehat{\ln\gamma_{l+k}})^{(r+1)} - \frac{1}{2}\sum_{g=1}^{G}\hat{\gamma}_{l+k}^{(r+1)}\varpi_{l,g}^{(r+1)}$, and $(\widehat{\ln\gamma_l})^{(r+1)} = \langle\ln\gamma_l\rangle_{q^{(r+1)}(\gamma_l)} = \Psi(e_l^{(r+1)}) - \ln(f_l^{(r+1)})$.

■

The proof of Lemma 1 and the pseudocode of the proposed FBSL-BC method is expressed in the Supporting Information.

3. RESULTS AND DISCUSSION

3.1. Simulated Examples. In this section, we evaluate the performance of the proposed FBSL-BC algorithm with simulated datasets and real data sets, respectively. The compared algorithms are airPLS (with the tunable parameter λ^{air}), ALS (with the tunable parameters λ^{ALS} and p^{ALS}), SSFBCSP (with the tunable parameters λ_1 and λ_2), BEADS (with the tunable parameter d, f_c, and r) and SBL-BC. Note that all the tunable parameters will be obtained by grid searching and will be specified in each trial. The initialization of FBSL-BC is set as $\hat{\alpha}^{(0)} = 1$, $\hat{\gamma}_l^{(0)} = 1$, $\hat{h}_{l,k} = 1/3$, $\forall l, k$, $\boldsymbol{\Sigma}_b^{(0)} = (\mathbf{I} + 10^5\cdot\mathbf{D}_L^{\mathrm{T}}\mathbf{D}_L)^{-1}$, $\hat{\mathbf{b}}_g^{(0)} = \boldsymbol{\Sigma}_b^{(0)}\mathbf{x}_g$, $\forall l, k, g$. The partial least-squares regression (PLSR) will be used to set up the calibration model, and the number of latent variables (LVs) is obtained by leave-one-out cross-validation (LOOCV). The root-mean-square error prediction (RMSEP), root-mean-square error cross validation (RMSECV), and coefficient of determination (R^2) are used to evaluate the accuracy of the prediction results.

SBL-BC and SSFBCSP are selected as the benchmarks to evaluate the performance of FBSL-BC in Simulation 1. The parameters of SSFBCSP were optimized by searching the sets $\{10^2, 10^3, ..., 10^{10}\}$ and $\{10^{-1}, 10^{-2}, ..., 10^{-5}\}$, respectively. For fairness, the redundant dictionary used in SSFBCSP is replaced by the one used in FBSL-BC and SBL-BC. Consider that a simulated spectrum consists of 1500 channels and 11 Gaussian peaks, i.e.,

https://doi.org/10.1021/acs.analchem.1c05443
Anal. Chem. 2022, 94, 5113−5121

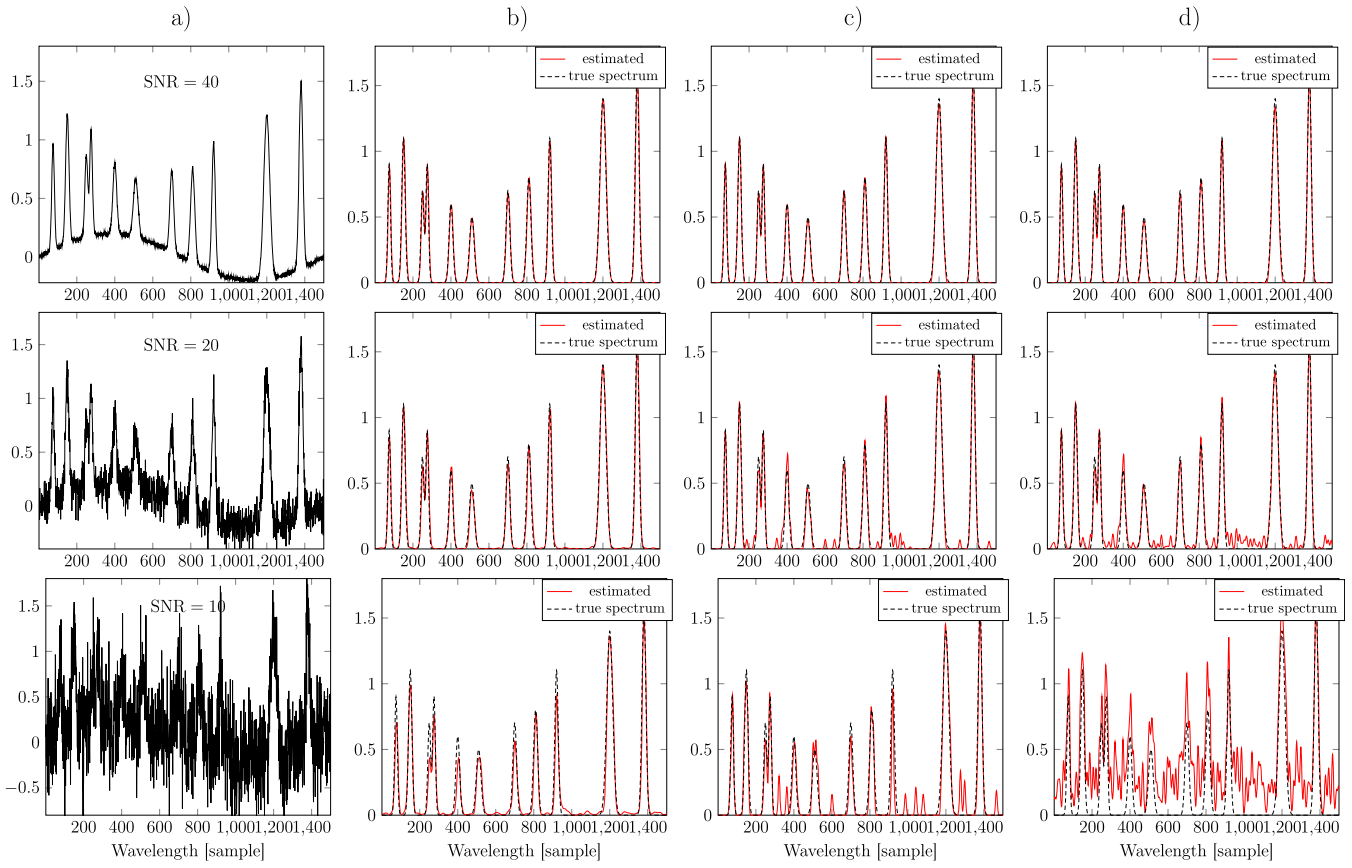


Figure 3. Comparison among FBSL-BC, SBL-BC, and SSFBCSP for simulated spectrum with a sinusoidal baseline. (a) simulated spectrum and (b) proposed FBSL-BC. (c) SBL-BC, (d) SSFBCSP with $\lambda_1 = \{10^4, 10^6, 10^5\}$, and $\lambda_2 = \{10^{-3}, 10^{-1}, 10^{-1}\}$.

$$s_i = \sum_{l=1}^{11} p_l \exp\left(-\frac{(i-\bar{\mu}_l)^2}{2\bar{\sigma}_l^2}\right) \qquad i = 1, 2, ..., 1500 \tag{47}$$

where $\mathbf{s}_i$ denotes the ith channel of $\mathbf{s}$, $\{p_l\}_{l=1}^{11}$ are set to {0.9, 1.1, 0.7, 0.9, 0.6, 0.5, 0.7, 0.8, 1.1, 1.4, 1.6}, the center points of peak zones $\{\bar{\mu}_l\}_{l=1}^{11}$ are set to {75, 150, 250, 275, 400, 510, 700, 810, 920, 1200, 1380}, and the width of peak zones $\{\bar{\sigma}_l\}_{l=1}^{11}$ are set to {50, 80, 60, 50, 100, 150, 100, 120, 80, 250, 100}. The baselines are constructed with sinusoidal and exponential functions:

$$b_i = \frac{1}{5}\sin\left(\frac{2\pi i}{1500}\right) \tag{48}$$

and

$$b_i = \frac{1}{5}\exp\left(-\frac{2i}{1500}\right) \tag{49}$$

where b_i represents the ith channel of $\mathbf{b}$.

Figure 3 shows the baseline correction results with sinusoidal baseline using the proposed method and SBL-BC and SSFBCSP (the baseline correction results with exponential baseline are provided as Supporting Information in Figure S-1). From the figures, we can observe that (i) when signal-noise ratio (SNR) is 40 dB, all the methods achieve similar baseline correction performance; (ii) when SNR = 20 dB, SBC-BC and SSFBCSP show a partial performance degradation, but the proposed method still yields a accurate and smooth pure spectrum estimation result; (iii) for a low level SNR (10 dB), SBL-BC and SSFBCSP fail to estimate the correct results, and a few extra undefined peaks were overestimated in the background zone. It is mainly because some fluctuations caused by the noise were estimated as spectral peaks; and (iv) although some imperceptible fluctuations appear in background areas for the proposed method, its performance is still much better than SBL-BC and SSFBCSP.

In Simulation 2, the Monte Carlo trails are performed to reveal the noise influence on the baseline correction. The normalized mean square error (NMSE) is selected as the evaluation metric and is calculated as

$$\mathrm{NMSE} = \frac{1}{M}\sum_{m=1}^{M}\frac{\|\mathbf{s}_m^e - \mathbf{s}_m\|_2^2}{\|\mathbf{s}_m\|_2^2} \tag{50}$$

where M is the repeat times of Monte Carlo trails and $\mathbf{s}_m^e$ denotes the pure spectrum estimation of $\mathbf{s}_m$ at the mth trail. FBSL-BC is compared with airPLS, ALS, SSFBCSP, and SBL-BC. We fix the parameters $\rho^{\mathrm{SBL}} = 10^5$, $\lambda^{\mathrm{ALS}} = 10^5$, $p^{\mathrm{ALS}} = 0.005$, $\lambda^{\mathrm{air}} = 10^4$, $\lambda_1 = 10^5$ and $\lambda_2 = 10^{-2}$. During the parameter selection process, we find that ALS with parameter $\lambda^{\mathrm{ALS}} = 10$ and $p^{\mathrm{ALS}} = 0.5$ can smooth the original simulated spectrum, which can be seemed as a denoise effect. Therefore, a denoise process is added before the baseline correction. Such an improved method is referred as two-stage ALS in the following section. It can be used as a benchmark to investigate the denoising performance because FBSL-BC has the inherent capability in simultaneous baseline correction and denoising. Figure 4 shows the NMSE of pure spectrum estimating versus SNR (NMSE of pure spectrum estimating versus SNR with

https://doi.org/10.1021/acs.analchem.1c05443
Anal. Chem. 2022, 94, 5113−5121

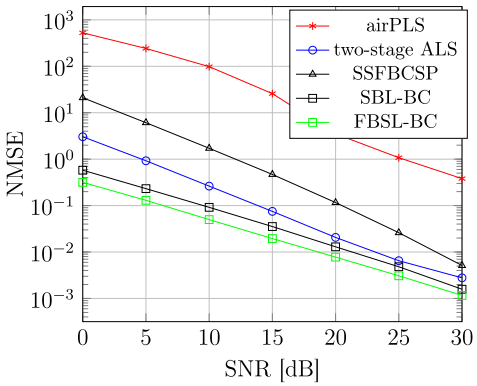


Figure 4. NMSE results under different SNR for simulated data with sinusoidal baseline, where $\lambda_1 = 10^5$, $\lambda_2 = 10^{-2}$, $\lambda^{ALS} = 10^6$, $p^{ALS} = 0.005$, and $\lambda^{air} = 10^4$.

exponential function based baseline are provided as Supporting Information (Figure S-2)). It can be seen that (i) the value of NMSE decreases with the increasing of SNR; (ii) airPLS yields the worst NMSE performance in all conditions; and (iii) FBSL-BC always outperforms others.

3.2. Applications to Real Datasets. To carefully investigate the baseline correction performance in real datasets and whether the proposed method can improve the accuracy of quantitative analysis, the first experiment is based on the corn NIR dataset that contains 80 samples and 4 analytes. The wavelength range is 1100−2498 and 2 nm intervals resulting in 700 variables (http://eigenvector.com). SBL-BC ($\rho = 10^5$), airPLS ($\lambda^{air} = 10^8$, order = 3), SSFBCSP ($\lambda_1 = 2 \times 10^5$, and $\lambda_2 = 10^{-2}$), and arPLS are used as comparison algorithms. The dataset is randomly split into two sub-datasets: the training set, which contains 50 samples, and the testing set, which contains 30 samples. The baseline correction results are provided as Supporting Information (Figure S-3). The averaged R^2 values are shown in Table 1, and the results of averaged RMSEP and

Table 1. Comparison of Averaged R^2 Results among airPLS ($\lambda^{air} = 10^8$), arPLS ($\lambda^{ar} = 10^5$), SSFBCSP ($\lambda_1 = 10^4$ and $\lambda_2 = 10^{-2}$), SBL-BC ($\rho = 10^5$), and FBSL-BC

	Averaged R^2 Results			
	moisture	oil	protein	starch
airPLS	0.9644	0.8343	0.9440	0.8808
arPLS	0.9575	0.8468	0.9333	0.8848
SSFBCSP	0.9619	0.8393	0.9484	0.8873
SBL-BC	0.9626	0.8620	0.9535	**0.8924**
FBSL-SBL	**0.9647**	**0.8671**	**0.9575**	0.8791

EMSECV are provided in the Supporting Information. It can be seen that (i) the PLS model with proposed FBSL-BC method has achieved the best R^2 (0.9647, 0.8671, and 0.9575) for the content prediction of moisture, oil, and protein, while SBL-BC obtains the best performance when for the content prediction of oil; and (ii) the execution time of FBSL-BC, SBL-BC, and SSFBCSP are 1.9210, 2.8890, and 4.1658 s, respectively. These results illustrate that the down-sampling strategy can significantly alleviate computational load for the Bayesian inference. Moreover, we would point out that the performance of comparison methods are very sensitive to the user-defined parameters, which should be identified through trial and error. In contrast, the parameter of FBSL-BC is automatically determined by the size of datasets.

In Experiment 2, we study the proposed method performance for large-scale datasets, where a private Huangshan Maofeng (a type of Chinese tea) dataset was selected, which includes 12 446 variables and 160 samples. The dataset was recorded by IRTracer-100 (Shimadzu, Japan) Fourier transform spectrometer, ranging from 10000 to 4000 cm^{-1} and the analyte is the content of sugar. We split the dataset into the training set (110 samples) and testing set (50 samples) randomly. The baseline correction results are provided in the Supporting Information (Figure S-4). The averaged RMSECV, RMSEP, R^2 and execution time are listed in Table 2. It is seen

Table 2. Averaged Evaluation Index of the Huangshan Maofeng Dataset with SSFBCSP, airPLS, SBL-BC, and FBSL-BC

	Evaluation Index			
	RMSECV	RMSEP	R^2	time [s]
airPLS	**0.8781**	1.6739	0.8892	2.2683
SSFBCSP	1.2085	1.7862	0.8746	1050.68
SBL-BC	1.14484	1.7643	0.8771	168.53
FBSL-BC	0.9520	**1.6212**	**0.8962**	97.2180

that all the methods reasonably estimate the baseline, but the PLS model with proposed FBSL-BC has achieved the best prediction performance with RMSEP and R^2 are 1.6212 and 0.8962, respectively. In addition, the proposed method is the fastest method (97.21 s) among sparse representation methods (168.53 s for SBL-BC, 1050.68 s for SSFBCSP), which significantly reduces the computational complexity.

The chromatogram dataset with an added Gaussian noise[27] was selected to verify the performance of the proposed method in Experiment 3. We compare FBSL-BC with SBL-BC BEADS, SBL-BC, and airPLS. The parameter selection of the BEADS model is the same as that observed for ref 27, i.e., $d = 1$, $f_c = 0.006$, and $r = 6$. Figure 5 shows the original chromatogram spectrum, estimated baseline, and estimated pure chromatogram for the different methods. To illustrate the differences in detail, the range between 900 and 1800 was enlarged and shown in the lower panel of Figure 5. It was observed that (i) FBSL-BC and BEADS have similar baseline correction results, and both of them can delineate the baseline well, while the proposed method yielded much sparser and smoother pure spectrum estimation results; and (ii) SBL-BC also obtained some smoother and sparser results, but the estimated baseline presented an overfitting manner.

Finally, three mineral Raman spectral datasets were adopted to thoroughly verify the generalization capability for different analytical instruments: GmeliniteNa, Cassiterite, and Marialite (http://www.handbookofmineralogy.org). The three datasets are added by zero-mean Gaussian noises with SNR = 50, 40, and 30 dB, respectively, aiming to investigate the impact of noise. The baseline correction results are shown in the Supporting Information (Figures S5−S7). It was observed that (i) airPLS can obtain an accurate baseline correction in the low noise level, but it suffered from a significant performance degradation with the increase of noise level; (ii) three sparse representation-based methods estimated the baseline adequately under all conditions, which suggests the superiority of the sparse representation methods; and (iii) although SBL-BC and SSFBCSP can obtain a satisfactory baseline correction

https://doi.org/10.1021/acs.analchem.1c05443
Anal. Chem. 2022, 94, 5113−5121

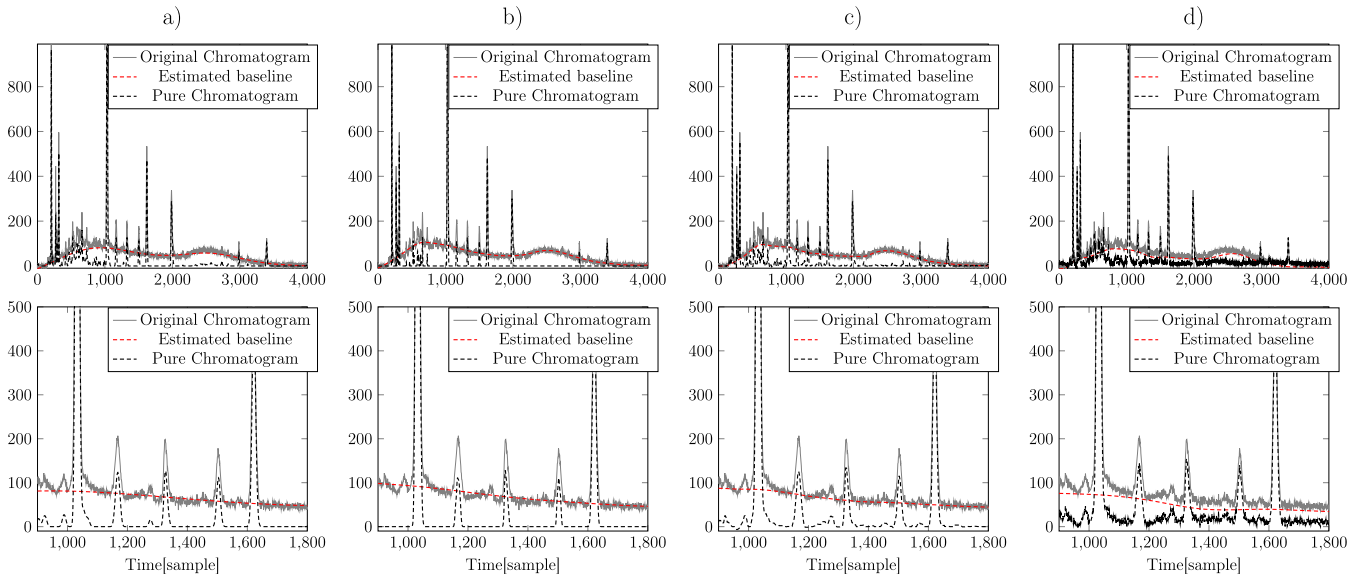


Figure 5. Estimated baseline comparison for the chromatogram dataset; the lower panel shows the enlarged part in the range of 900−1800: (a) FBSL-BC, (b) SBL-BC, (c) BEADS, AND (d) airPLS.

Table 3. Computational Time for All Experiments under Different Baseline Correction Methods

		Time [s]			
dataset	size	FBSL-BC	SBL-BC	SSFBCSP	airPLS
corn	700 × 80	1.9210	2.8890	4.1658	0.1131
cassiterite	1099 × 1	0.4501	0.6192	0.7018	0.0069
gmeliniteNa	1173 × 1	0.5177	0.6725	0.8059	0.0071
marialite	2398 × 1	1.2588	2.3737	3.1086	0.0154
chromatogram	4000 × 1	1.8447	3.6041	10.3351	0.0110
tea	12446 × 168	97.2180	168.5321	1050.68	1.2638

result, the proposed method yields a much smoother and sparser pure spectrum fitting, which means that FBSL-BC has an excellent noise reduction capability.

3.3. Discussion. Recall that FBSL-BC aims to overcome the two shortcomings of existing methods: (1) the high computation complexity for large-scale datasets; and (2) the completely ignored burst-sparsity structure. The computational time for all the experiments is shown in Table 3, where FBSL-BC is superior to others, with regard to speed, especially for dealing with high-dimension datasets. For the most common datasets with hundreds or thousands of variables, the running time of FBSL-BC is 1−3 s, which is acceptable for most applications. The airPLS is the fastest method, but it is sensitive to the parameter selection and it is time-consuming to select the appropriate parameters. Clearly, the proposed down-sampling strategy can greatly reduce the computational load.

FBSL-BC worked well for both the simulated and real datasets, and it is able to capture the burst-sparsity structure by pattern-coupled prior. As shown in Section 3.2 and Figure 2, the denser nonzero sparse coefficients (significant elements) that appear in the peak zone fully reflect the effect of the burst sparsity. The sparse coefficients for three mineral Raman data sets yielded by FBSL-BC and SBL-BC are provided as Supporting Information (Figure S8). Here, we selected and expanded the wavelength ranges containing denser feature peaks, to clearly show the difference of sparse coefficients. Similar to Figure 2, the values of coefficients in peak zones obtained by FBSL-BC are nonzero elements, which means that the burst-sparsity structure has been well captured by the pattern-coupled prior. The simulation results and real examples reaffirmed that exploiting the burst sparsity can enhance the baseline correction performance, especially for high noise levels.

4. CONCLUSION

In this paper, we extend the SBL-BC method with a fast burst-sparsity scheme that addresses the performance problems associated with other techniques. The proposed scheme is mainly based on two points. First, a down-sampling process is performed to reduce the dimension of the Gaussian line shape dictionary matrix and construct a multiple measurements block-sparse recovery problem to avoid information loss. Second, to further enhance the baseline correction performance, the burst-sparsity property in the sparse representation is carefully characterized by the pattern-coupled prior. The effectiveness of FBSL-BC was demonstrated through some simulation and real dataset examples. The simulation and Monte Carlo trials indicate that the fitted pure spectrum yield by the FBSL-BC is very similar to the simulated one, and superior performance under the high noise level condition, compared with other sparse representation methods, was also demonstrated, which means that capturing the burst-sparsity structure actually improves the baseline correction performance. In real dataset experiments 1 and 2, the PLSR model with proposed FBSL-BC has achieved superior predictive performance to other methods in most cases, which demonstrate that the FBSL-BC method can improves the prediction accuracy of the calibration model. Moreover, the recorded execution time

https://doi.org/10.1021/acs.analchem.1c05443
Anal. Chem. 2022, 94, 5113−5121

in Table 3 demonstrates that the down-sampling strategy can significantly reduce computational complexity, especially for large-scale datasets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.1c05443.

Update of solutions with Lemma 1; pseudocode of FBSL-BC; the figure of baseline correction results for different datasets; result of RMSECV and RMSEP for corn dataset (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Jisheng Dai − *School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China*; Email: jsdai@ujs.edu.cn

Xiaobo Zou − *School of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, China*; orcid.org/0000-0001-5719-7025; Email: xiaobo@ujs.edu.cn

Authors

Haoran Li − *School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China*; orcid.org/0000-0001-8399-8643

Suyi Chen − *School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China*

Tao Chen − *Department of Chemical and Process Engineering, University of Surrey, Guildford GU2 7XH, United Kingdom*

Tianhong Pan − *School of Electrical Engineering and Automation, Anhui University, Hefei 230061, China*

Melvin Holmes − *School of Food Science and Nutrition, University of Leeds, Leeds LS2 9JT, United Kingdom*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.1c05443

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Authors H.L., J.D., and S.C. thank the National Natural Science Foundation of China (under Project Grant No. 62071206). Author X.Z. thanks the earmarked fund for China Agriculture Research System.

■ REFERENCES

(1) Pasquini, C. *Anal. Chim. Acta* **2018**, *1026*, 8−36.
(2) Luo, S.-h.; Wang, X.; Chen, G.-y.; Xie, Y.; Zhang, W.-h.; Zhou, Z.-f.; Zhang, Z.-m.; Ren, B.; Liu, G.-k.; Tian, Z.-q. *Anal. Chem.* **2021**, *93*, 8408.
(3) Rinnan, Å.; Van Den Berg, F.; Engelsen, S. B. *TrAC Trends Anal. Chem.* **2009**, *28*, 1201−1222.
(4) Brown, C. D.; Vega-Montoto, L.; Wentzell, P. D. *Appl. Spectrosc.* **2000**, *54*, 1055−1068.
(5) Shao, X.; Cai, W.; Pan, Z. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 249−256.
(6) Peng, J.; Peng, S.; Xie, Q.; Wei, J. *Anal. Chim. Acta* **2011**, *690*, 162−168.
(7) Leger, M. N.; Ryder, A. G. *Appl. Spectrosc.* **2006**, *60*, 182−193.
(8) Selesnick, I. W.; Graber, H. L.; Pfeil, D. S.; Barbour, R. L. *IEEE Trans. Signal Process.* **2014**, *62*, 1109−1124.
(9) Zhang, Z.-M.; Chen, S.; Liang, Y.-Z. *Analyst* **2010**, *135*, 1138−1146.
(10) Mosier-Boss, P.; Lieberman, S.; Newbery, R. *Appl. Spectrosc.* **1995**, *49*, 630−638.
(11) Bertinetto, C. G.; Vuorinen, T. *Appl. Spectrosc.* **2014**, *68*, 155−164.
(12) Zhang, Z.-M.; Chen, S.; Liang, Y.-Z.; Liu, Z.-X.; Zhang, Q.-M.; Ding, L.-X.; Ye, F.; Zhou, H. *J. Raman Spectrosc.* **2010**, *41*, 659−669.
(13) Lieber, C. A.; Mahadevan-Jansen, A. *Appl. Spectrosc.* **2003**, *57*, 1363−1367.
(14) Eilers, P. H. *Anal. Chem.* **2004**, *76*, 404−411.
(15) Hu, H.; Bai, J.; Xia, G.; Zhang, W.; Ma, Y. *Photonic Sensors* **2018**, *8*, 332−340.
(16) Zhao, J.; Lui, H.; McLean, D. I.; Zeng, H. *Appl. Spectrosc.* **2007**, *61*, 1225−1232.
(17) Gan, F.; Ruan, G.; Mo, J. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 59−65.
(18) Baek, S.-J.; Park, A.; Ahn, Y.-J.; Choo, J. *Analyst* **2015**, *140*, 250−257.
(19) Li, Z.; Zhan, D.-J.; Wang, J.-J.; Huang, J.; Xu, Q.-S.; Zhang, Z.-M.; Zheng, Y.-B.; Liang, Y.-Z.; Wang, H. *Analyst* **2013**, *138*, 4483−4492.
(20) He, S.; Zhang, W.; Liu, L.; Huang, Y.; He, J.; Xie, W.; Wu, P.; Du, C. *Analytical Methods* **2014**, *6*, 4402−4407.
(21) Li, H.; Dai, J.; Pan, T.; Chang, C.; So, H. C. *Chemom. Intell. Lab. Syst.* **2020**, *204*, 104088.
(22) Candès, E. J.; Romberg, J.; Tao, T. *IEEE Trans. Inform. Theory* **2006**, *52*, 489−509.
(23) Donoho, D. L. *IEEE Trans. Inform. Theory* **2006**, *52*, 1289−1306.
(24) Li, H.; Dai, J.; Xiao, J.; Zou, X.; Chen, T.; Holmose, M. *Chemom. Intell. Lab. Syst.* **2022**, *221*, 104487.
(25) Lei, T.; Sun, D.-W. *Anal. Chim. Acta* **2020**, *1110*, 169−180.
(26) Yu, S.; You, X.; Mou, Y.; Jiang, X.; Ou, W.; Zhou, L. A New Approach for Spectra Baseline Correction Using Sparse Representation. In *Proceedings of the IASTED International Conference*, 2013; pp 215−221.
(27) Ning, X.; Selesnick, I. W.; Duval, L. *Chemom. Intell. Lab. Syst.* **2014**, *139*, 156−167.
(28) Han, Q.; Xie, Q.; Peng, S.; Guo, B. *Analyst* **2017**, *142*, 2460−2468.
(29) Candes, E. J.; Tao, T. *IEEE Trans. Inform. Theory* **2005**, *51*, 4203−4215.
(30) Eldar, Y. C.; Mishali, M. *IEEE Trans. Inform. Theory* **2009**, *55*, 5302−5316.
(31) Eldar, Y. C.; Kuppinger, P.; Bolcskei, H. *IEEE Trans. Signal Process.* **2010**, *58*, 3042−3054.
(32) Dai, J.; Liu, A.; So, H. C. *IEEE Trans. Signal Process.* **2019**, *67*, 1075−1087.
(33) Brown, A. J. *IEEE Trans. Geosci. Remote Sensing* **2006**, *44*, 1601−1608.
(34) Fraser, R. D.; Suzuki, E. *Anal. Chem.* **1969**, *41*, 37−39.
(35) Tipping, M. E. *J. Machine Learning Res.* **2001**, *1*, 211−244.
(36) Tzikas, D. G.; Likas, A. C.; Galatsanos, N. P. *IEEE Signal Process. Mag.* **2008**, *25*, 131−146.
(37) Fang, J.; Shen, Y.; Li, H.; Wang, P. *IEEE Trans. Signal Process.* **2015**, *63*, 360−372.

https://doi.org/10.1021/acs.analchem.1c05443
Anal. Chem. 2022, 94, 5113−5121