Sparse Modeling for Spectrometer Based on Band Measurement

Kyoya Uemura $^1$, Tomoyuki Obuchi $^2$, and Toshiyuki Tanaka $^2$

$^1$Kyoto University
$^2$Affiliation not available

July 21, 2023
Sparse Modeling for Spectrometer Based on Band Measurement

Kyoya Uemura, Graduate Student Member, IEEE, Tomoyuki Obuchi, Member, IEEE, and Toshiyuki Tanaka, Member, IEEE

Abstract—In typical spectrometric measurement systems, a high-resolution spectrum is obtained directly via sequential observations with a narrow slit-like measurement window at the expense of sensitivity. In this paper, we propose a novel spectrometric method applicable to these typical spectrometric systems: a multiplexed low-resolution measurement with a wide measurement window, band measurement (BM), is combined with sparse-modeling-based post-processing to obtain the original high-resolution spectrum. BM is expected to improve the measurement signal-to-noise ratio because of the increase in the sample quantities reaching the detector by widening the measurement window. BM has the significant practical advantage that it can be easily implemented in spectrometric measurement systems without device alterations. To evaluate the effectiveness of our proposal both theoretically and experimentally, we formulate the sparse-modeling-based post-processing in the proposal in terms of the least absolute shrinkage and selection operator (lasso) and perform a theoretical analysis and simulation studies concerning the resulting spectrometric method named BM-lasso. In the theoretical analysis, we derive density evolution equations for belief propagation on the BM-lasso model and obtain the expected errors of estimators of BM-lasso. The numerical evaluations of the theoretical analysis result revealed that BM-lasso achieved lower mean square error than the conventional measurement method under the same parameter conditions. Furthermore, simulation studies with both artificial and actual mass spectra show that BM-lasso significantly improves the accuracy, sensitivity, and specificity compared with the conventional measurement method, demonstrating the practicality of the proposed method.

Index Terms—Sparse modeling, lasso, belief propagation, density evolution, spectroscopy, mass spectrometry, spectrometer.

I. INTRODUCTION

SPECTROMETRIC analysis is an analytical method to acquire qualitative and quantitative information regarding an analysis target from a spectrum obtained with a spectrometer, and these have been positioned as fundamental techniques in various fields of science and engineering. There are many types of spectrometric analyses; for instance, Fourier transform infrared (FT-IR), Raman, ultraviolet-visible (UV-Vis), and X-ray photoelectron spectroscopies are well known. The target of spectrometric analyses is not limited to light. For example, mass spectrometry (MS) measures the mass-to-charge ratio (m/z) of the chemical species present in a sample. The resulting spectrum, mass spectrum, has a lot of chemical information, such as molecular weights and isotopic signatures, which allows us to determine the chemical structures and quantities of chemical species. The range of applications of these analyses has been growing over the years, such as proteomics [1]–[3] and hyperspectral imaging [4]–[6], and accordingly, the analysis targets have been more complex. This trend has led to a great demand for establishing a spectrometric method to achieve high sensitivity and accuracy in a limited measurement time.

In a measurement process of a typical spectrometer, analysis targets are separated according to the magnitudes of a target property such as m/z, wavelength, and energy, and then they are selectively introduced into a detector through a measurement window such as an electromagnetic filter and a slit. In particular, a narrow slit-like measurement window is generally used to scan along a measurement range and obtain a high-resolution spectrum directly as an array of detected signals (Fig. 1 (a)). These spectrometers with a slit-like measurement window are the most typical and widely used type of spectrometric system because they are inexpensive and easy to handle. However, a narrow measurement window considerably restricts the amount of analysis targets from reaching a detector. Therefore, it often suffers from low measurement sensitivity or signal-to-noise ratio (SNR).

In order to improve the sensitivity or SNR of spectrometric measurement, several methods for various modalities of spectrometry have been suggested over the past few decades. In particular, multiplexed measurement is one of the mainstream methods to improve SNR by simultaneously observing spectral components to acquire more information at once. One example, in some types of optical spectroscopy, is to use a multi-channel detector [7]–[9]. In this approach, a measurement system separates target light by wavelength with a polychromator and simultaneously measures each spectrum divided into some spectral ranges with a multi-channel detector. Another approach is to utilize information processing techniques as post-processing to reconstruct an original spectrum from measurement signals encoded by the measurement conditions [10]. For instance, Brock et al. [11] utilized the Hadamard transform for time-of-flight (TOF) mass spectrometer and improved the
SNR by efficiently increasing the duty cycle and the ion throughput. Other studies proposed to perform several spectral scans in parallel and estimate the original spectrum by considering the sparse nature of spectra [12], [13]. As for nuclear magnetic resonance (NMR) and Raman spectroscopy, faithful reconstruction of sparse spectra was achieved by compressed sensing with non-uniform sampling in the time domain [14]–[17].

Although these approaches effectively improve SNR, they have the significant problem of requiring complicated measurement systems. Most of these multiplexed measurements need expensive hardware modifications or complex measurement systems; thus, it is hard to implement them in the most widely used type of spectrometers that employ a slit-like measurement window, such as quadrupole mass spectrometer (QMS) and optical spectrometers with a monochromator, due to the instrumental constraints. Thus, there is a strong demand for multiplexed measurement techniques applicable to these typical spectrometers.

In this paper, we propose a novel spectrometric method applicable to typical spectrometric systems employing a slit-like measurement window: a multiplexed low-resolution measurement with a wide measurement window (Fig. 1 (b)), band measurement (BM), is combined with sparse-modeling-based post-processing to obtain the original high-resolution spectrum. In the BM, it is expected to improve the measurement SNR because of the increase in the sample quantities reaching the detector by widening the measurement window. BM has the significant advantage of being easily implemented in typical spectrometric systems because a measurement window width is a controllable variable for these spectrometers; hence, device alterations or complex measurement systems are not required.

To evaluate the effectiveness of this proposal both theoretically and experimentally, we formulate the sparse-modeling-based post-processing in the proposal in terms of the least absolute shrinkage and selection operator (lasso) [18] as the most standard sparse modeling formulation, which has desirable properties for our problem setting and theoretical analysis (discussed in detail in Sec. II-C). Hereafter, the resulting spectrometric method combining BM and lasso is called BM-lasso.

In the theoretical analysis, we perform belief propagation (BP) and density evolution (DE) utilizing the (high-order) Markov property of BM and derive the asymptotic expected errors of the BM-lasso estimators. One of the main theoretical contributions of this paper is to derive the DE equations for BP on the high-order Markov chain with non-Gaussian state distribution. The numerical evaluations of the theoretical analysis result revealed that BM-lasso achieved lower mean square error than the conventional measurement method at any SNR under the same parameter conditions. Moreover, simulation studies with both artificial and actual mass spectra were performed to show the practicality of the proposed method in more realistic situations. The result shows that BM-lasso significantly improved the accuracy, sensitivity, and specificity compared with the conventional measurement method, demonstrating the practicality of the proposed method.

II. SPARSE MODELING WITH BAND MEASUREMENT

A. Linear Model of Spectroscopic Measurement

In a typical spectroscopic measurement system, the measurement process can be formulated as the following linear measurement model via appropriate discretization,

\[ y = X \beta^* + \epsilon, \] (1)

where \( y = [y_1, y_2, \ldots, y_N]^T \in \mathbb{R}^N \) is the vector representing the measurement signals, \( \beta^* = [\beta_1^*, \beta_2^*, \ldots, \beta_N^*]^T \in \mathbb{R}^N \) is the original spectrum binned in \( N \) intervals, and \( \epsilon \in \mathbb{R}^N \) is the measurement noise vector. \( X = [x^{(1)}, x^{(2)}, \ldots, x^{(N)}]^T \in \mathbb{R}^{N \times N} \) denotes the measurement matrix, and \( x^{(i)} \in \mathbb{R}^N \) denotes the vector representing the filtering property at \( i \)th observation.

Although a wide variety of types of spectrometers have been developed and used in various domains of spectrometry, the most typical and straightforward way to measure a spectrum is to pass the sample to be analyzed through a filter with a narrow slit-like measurement window to a detector while scanning the position of the measurement window over a particular spectral range (Fig. 1 (a)). This measurement method observes each spectral component sequentially. Thus, the conventional measurement process can be formulated as

\[ y = I_N \beta^* + \epsilon, \] (2)

where \( I_N \) is the \( N \times N \) identity matrix. Then, the measurement result can be easily interpreted because the measurement signal \( y = \beta^* + \epsilon \) is simply a noisy version of the true spectrum. On the other hand, this measurement method can only detect a single spectrum component in each observation, and all the components except the detected one are discarded. Therefore, the SNR of the conventional measurement is often low, which may significantly limit the measurement sensitivity and accuracy.

Moreover, these typical spectrometric systems have another problem: it is challenging to implement recent spectrometric methods [12]–[17] in the typical spectrometers with slit-like measurement windows due to their instrumental constraints. Therefore, there is a strong demand for novel spectrometric methods that not only improve performance but also can be easily implemented in these typical spectrometers.

B. Band Measurement (BM) Model

We design a novel class of multiplexed measurements for typical spectrometers, which we call the band measurement (BM). It is a low-resolution measurement with a wide measurement window as shown in Fig. 1 (b). The BM process can be formulated as

\[ y = X_w \beta^* + \epsilon, \] (3)

where

\[ X_w = [x_w^{(1)}, x_w^{(2)}, \ldots, x_w^{(N)}]^T, \] (4)

\[ x_{ij} = \begin{cases} 1, & 0 \leq j - i < w, \\ 0, & \text{otherwise}. \end{cases} \] (5)
The measurement matrix $X_w \in \mathbb{R}^{N \times N}$ in BM is defined as a band matrix with the bandwidth $w \in \mathbb{N}$, which corresponds to the width of the slit-like measurement window. Note that any band matrix $X_w$ is of full rank because it is an upper triangular matrix without zero diagonal elements. In addition, the band matrix $X_w$ with $w = 1$ equals to $I_N$, and thus BM can also be regarded as a natural extension of the conventional measurement model (2).

A wider measurement window in BM makes the intensity of the measurement signals larger, by passing larger quantities of the sample through the filter. Therefore, the BM with $w \geq 2$ is expected to improve the SNR and sensitivity of the spectrometric measurement compared with the conventional measurement process ($w = 1$). Moreover, BM has the significant advantage of being easily implemented in typical spectrometers only by controlling the measurement conditions; hence alterations on the device or complex measurement systems are not required, unlike most of the methods proposed so far for multiplexed measurement. In fact, optical spectrometers can perform BM just by widening the exit slit, and BM for QMS has already been utilized in another context of data-independent acquisition (DIA) [19], [20].

In the case of BM with $w \geq 2$, the measurement signal consists of a mixture of $w$ consecutive spectral components. Consequently, one needs additional post-measurement signal processing to reconstruct the original spectrum $\beta^*$ from the noisy low-resolution measurement signal $y$.

C. Reconstruction of sparse spectra

For reconstructing the spectrum, we employ the property that the true spectrum $\beta^*$ has sparsity:

$$\|\beta^*\|_0 \ll N,$$

where $\|\beta^*\|_0$ denotes the number of nonzero elements in $\beta^*$. Sparsity holds exactly or approximately in some standard spectra, such as mass spectra and emission line spectra. From an information processing viewpoint, the reconstruction of the sparse spectrum $\beta^*$ from the given measurement matrix $X_w$ and the BM signal $y$ in (3) is equivalent to 1D-image deblurring and super-resolution, with $z^{(i)}_{w}$ being the blurring kernels. In this context, sparse modeling is well known to solve these problems. The most classical approach of sparse modeling is to find an estimate $\hat{\beta}$ via solving a constrained $\ell_1$-norm $\|\hat{\beta}\|_0$ optimization problem. This problem is a class of combinatorial optimization problems and is NP-hard [21]; thus, it is typically solved approximately with greedy algorithms [22]–[25]. Another approach is based on relaxations in which the non-convex and discontinuous $\ell_0$-norm is approximated by other functions such as $\ell_1$-norm ($0 < p \leq 1$) [26]. In particular, $\ell_1$-approximation is well-known as a successful approach, e.g., basis pursuit [27] and lasso [18], since there exist feasible and efficient optimization algorithms due to the convexity and the $\ell_1$-optimization problem shares the same solution as the original $\ell_0$-problem with high probability under certain conditions [28]–[30]. Moreover, novel sparse modeling techniques and their algorithms have been developed over the years, e.g., the $\ell_1$-regularization approaches based on more sophisticated statistical models [13]–[15] and Bayesian approach using a sparsity-promoting prior [31], [32].

Although there are a lot of sparse modeling techniques as discussed above, in this work, we specifically focus on the lasso, a class of $\ell_1$-regularization method, for the following three reasons. First, the lasso is suitable for our aim to theoretically evaluate the effectiveness of the concept that the BM is combined with sparse modeling: the lasso is the most standard sparse modeling method and has desirable properties for the theoretical analysis discussed in Sec. III, such as convexity and uniqueness of solutions. Second, several fast and efficient algorithms for lasso are well-known; thus, it is advantageous for practical applications with high-dimensional target spectra and limited computational resources. Finally, the lasso has a well-established criterion for model/hyperparameter selection, which is advantageous for our problem setting, as discussed below.

The lasso estimate for $\beta^*$ of the BM model (3), in its regularized form, is given by

$$\hat{\beta}_1 = \arg \min_{\beta} \mathcal{H}(\beta | y, X_w, \lambda),$$

where

$$\mathcal{H}(\beta | y, X_w, \lambda) = \frac{1}{2} \| y - X_w \beta \|_2^2 + \lambda \| \beta \|_1$$

is the objective function, which consists of the ordinary least square (OLS) term and the $\ell_1$ penalty term. The hyperparameter $\lambda \geq 0$, called the lasso regularization parameter, controls the weight of the $\ell_1$ penalty. The $\ell_1$ penalty works as a
sparsity-promoting prior and encourages the lasso estimate $\hat{\beta}_\lambda$ to be sparse while maintaining the convexity of the objective function to ease the optimization.

To obtain a reasonable estimate, an appropriate value of $\lambda$ must be selected. For example, one may want to minimize the estimation risk $\mathbb{E}[\|\hat{\beta}^* - \hat{\beta}_\lambda\|^2_2/N]$, which is the expected value of the mean square error (MSE) of the estimated spectrum $\hat{\beta}_\lambda \in \mathbb{R}^N$, with respect to $\lambda$ as

$$
\lambda_{\text{MSE}} = \arg \min_\lambda \mathbb{E} \left[ \frac{\|\hat{\beta}^* - \hat{\beta}_\lambda\|^2_2}{N} \right].
$$

However, $\lambda_{\text{MSE}}$ cannot be calculated because $\beta^*$ is unknown in general, and thus we need an alternative criterion that can be evaluated from known quantities only. The most typical way for hyperparameter selection is to perform cross-validation (CV), where the measurement signal dataset $\{y_i\}_{i=1}^N$ is split into the training dataset used to obtain an estimate and the validation dataset used to evaluate the risk. Unfortunately, however, hyperparameter selection with CV does not work well for BM model (3). This is because, in the BM model, at most only $w$ of the $N$ observed signals contain information on a particular spectral component $\hat{\beta}_i$, and the split dataset may not contain any information on the spectral component. On the other hand, for lasso, adaptive model selection criteria for efficiently selecting the optimal fit have been discussed [33]–[35] in the framework of Stein’s unbiased risk estimator (SURE) [36]. First, we considered employing the SURE of the estimation risk, but it did not work because the variance of the SURE tended to be quite large. Therefore, we employ a more stable criterion, Akaike information criterion (AIC) for lasso [34], [35], [37], which is an unbiased estimator of the prediction error.

Assume that $\epsilon$ in the linear model (3) is a spherically symmetric Gaussian noise, i.e., is generated according to

$$
\epsilon \sim \mathcal{N}(0_N, \sigma^2 I_N),
$$

where $0_N$ is the $N$-dimensional zero vector, and $\sigma$ is the standard deviation. Under this Gaussian noise assumption with known variance $\sigma^2$, AIC for lasso is derived as follows [34]:

$$
\text{AIC}_{\text{lasso}}(\hat{\beta}_\lambda) = \frac{\|\hat{y} - X_w \hat{\beta}^\top\|^2_2}{N\sigma^2} + 2 \frac{\|\hat{\beta}_\lambda\|_0}{N}.
$$

Note that AIC for lasso (11) can be evaluated using known quantities only. Therefore, AIC for lasso can be utilized as an appropriate criterion for selecting the value of the lasso regularization parameter $\lambda$.

From these discussions, we formulate the post-processing part of the novel spectroscopic method which we propose in this paper as what we call BM-lasso, in which an original spectrum is reconstructed according to (7). Furthermore, we employ AIC$_{\text{lasso}}$ as the adaptive model selection criterion for lasso, that is, $\lambda = \lambda_{\text{AIC}}$, where

$$
\lambda_{\text{AIC}} = \arg \min_\lambda \left[ \text{AIC}_{\text{lasso}}(\hat{\beta}_\lambda) \right] = \arg \min_\lambda \left[ \frac{\|\hat{y} - X_w \hat{\beta}^\top\|^2_2}{N\sigma^2} + 2 \frac{\|\hat{\beta}_\lambda\|_0}{N} \right].
$$

III. THEORETICAL ANALYSIS

A. Overview

In this section, we provide a theoretical framework that enables us to exactly compute the estimation risk in the large-$N$ limit, i.e., $\lim_{N \to \infty} \mathbb{E}[\|\hat{\beta}^* - \hat{\beta}_\lambda\|^2_2/N]$, given the noise and true signal distributions of certain reasonable forms. The framework is based on the technique called density evolution (DE) that tracks the average dynamical behavior of the algorithm termed belief propagation (BP). To facilitate understanding of the theoretical analysis, an overview of the DE prescription is first provided before the detailed discussion.

In Sec. III-B, as a preliminary to discussing the BM-lasso estimation in the framework of the BP algorithm, we formulate a probability distribution associated with the lasso estimation problem (7). In Sec. III-C, we show that BP is able to exactly compute the BM-lasso estimate $\hat{\beta}(y)$ depends on the measurement signal $y$ and thus has a distribution originating from the randomness of the true signal $\beta^*$ and noise $\epsilon$. DE gives the distribution of $\beta$ in a recursive manner, and then we derive the objective estimation risk $\lim_{N \to \infty} \mathbb{E}[\|\hat{\beta}^* - \hat{\beta}_\lambda\|^2_2/N]$ by considering the behaviors of the DE recursive equations in the large-$N$ limit. Finally, we give a concrete algorithm to compute the estimation risk $\lim_{N \to \infty} \mathbb{E}[\|\beta^* - \hat{\beta}_\lambda\|^2_2/N]$ in Sec. III-E.

It should be noted that, although BP is certainly applicable to compute the lasso estimate, we do not propose it for solving (7) in the current paper; BP is just introduced as an intermediate step to derive DE for the theoretical analysis. The DE result will be compared with a simulation result later, where a more efficient algorithm than BP is used in the simulation studies.

B. Formulation in terms of probability model

Following the prescription of the graphical model approach [38], we reformulate the lasso estimation (7) as a maximum-a-posteriori (MAP) estimation with respect to the following probability model:

$$
P_\gamma(\beta|X_w, y) = \frac{1}{Z} \exp \left[ -\gamma H(\beta|X_w, y) \right] = \frac{1}{Z} \left( \prod_{i=1}^N \phi_\mu(y_{\mu}|\beta, x_{\mu}^{(\mu)}) \right) \left( \prod_{i=1}^N \psi(\beta_i) \right),
$$

where

$$
\phi_\mu(y_{\mu}|\beta, x_{\mu}^{(\mu)}) = \exp \left[ -\frac{\gamma}{2} (y_{\mu} - x_{\mu}^{(\mu)}\beta)^2 \right],
$$

$$
\psi(\beta) = \exp \left[ -\gamma |\lambda|/|\beta| \right].
$$

$Z$ denotes the partition function, and $\gamma > 0$ is an “inverse temperature” parameter. The probability model (13) is
the posterior distribution of $\beta$ defined from the likelihood $\phi_{\mu}(y_\mu|\beta, x_\mu^{(\mu)})$ of $\beta$ given $y_\mu$, and the prior $\psi(\beta)$ of $\beta = (\beta_1, \ldots, \beta_N)^T$.

Although the above reformulation of the lasso estimation in terms of the MAP estimation is valid for any $\gamma$, if one takes the limit $\gamma \to \infty$, the posterior distribution $P_\gamma(\beta|X_w, y)$ concentrates on the set of the MAP solutions, i.e., $P_\gamma(\beta|X_w, y) \to \delta(\beta - \beta_{\text{MAP}}(y))$, where $\delta(\cdot)$ is the delta function. One can therefore assess the properties of the lasso estimate by studying the properties of the posterior distribution $P_\gamma(\beta|X_w, y)$ in the large-$\gamma$ asymptotics. More concretely, if the posterior distribution $P_\gamma(\beta|X_w, y)$ has a unique MAP solution, then one can expect that the posterior mean approaches the MAP distribution as asymptotics. More concretely, if the posterior $\psi(\beta)$ of (13) is regarded as a Markov chain, which can be exploited to evaluate the posterior distribution (13) in our problem can be regarded as a Markov chain of order $1$ via appropriate reformulation of the lasso estimation in terms of the MAP estimation.

We aim to evaluate the performance of BM-lasso by using the estimation risk $\mathbb{E}[\|\beta^* - \beta_\lambda\|^2/N]$. To evaluate it, we must obtain the BM-lasso estimate $\beta_\lambda$, i.e., the posterior mean of (13) in the large-$\gamma$ limit as discussed above. Although evaluation of the posterior mean of a high-dimensional probability model is in general computationally demanding since it requires high-dimensional integrations, it turns out that the posterior distribution (13) in our problem can be regarded as a Markov chain, which can be exploited to evaluate the posterior mean efficiently.

Let us describe the Markovian structure of the model (13) in more detail. For this purpose, we introduce a factor graph [38], [39], which is a bipartite graph consisting of variable nodes corresponding to $\beta_i$ and factor nodes corresponding to factors of (13), i.e., $\phi_{\mu}$ and $\psi$. In a factor graph, a variable node and a factor node are connected by an edge if and only if the factor function includes the variable as its argument. Fig. 2 represents factor graphs of the model (13) for $w = 1, 2,$ and $3$. When $w = 1$, each element of $\beta$ can be estimated separately from other elements thanks to their independence, as depicted by the disjointness of the factor graph. If the factor graph representation of the probability model of interest is connected but cycle-free (i.e., a tree), as in the case of $w = 2$, then one can apply BP on it to efficiently perform exact inference on the probability model. More precisely, when $w = 2$ the model (13) corresponds to a Markov chain of order 1, with $\beta_1, \ldots, \beta_N$ being its state variables. More generally, when $w \geq 2$ the model (13) corresponds to a Markov chain of order $(w - 1)$. Although the factor graph representation of the posterior distribution $P_\gamma(\beta|X_w, y)$ has cycles for $w \geq 3$, by following the standard prescription of reducing a higher-order Markov chain to a Markov chain of order 1 via appropriate grouping of state variables [40], one can convert it into a cycle-free one-dimensional chain, as shown in Fig. 3. The resulting “clustered” factor graph allows us to apply BP to do exact inference on it. Below we explain the detailed prescription for doing this.

Let us fix $w \geq 2$ and assume $N \equiv 0 \mod (w - 1)$ to simplify the notation in the following discussion. Let $I_t = \{(t - 1)(w - 1) + 1, (t - 1)(w - 1) + 2, \ldots, tw - 1\}$. The index set $\{1, 2, \ldots, N\}$ is partitioned into the collection $\{I_t\}_{t=1,2,\ldots,N/(w-1)}$ of the index subsets, each of which is of size $(w - 1)$. We then define the following clustered variables and clustered potential functions, with the clustering according to the partition $\{I_t\}_{t=1,2,\ldots,N/(w-1)}$:

$$\beta_t = (\beta_i)_{i \in I_t} = [\beta_i(t(w-1)+1), \ldots, \beta_i(t(w-1))]^T,$$

$$y_t = (y_i)_{i \in I_t} = [y_i(t(w-1)+1), \ldots, y_i(t(w-1))]^T,$$

$$\Phi_{\mathcal{T}}(\beta_t, \beta_{t+1} | y_t) = \prod_{\mu \in I_t} \phi_{\mu}(y_i | \beta_i, x_\mu^{(\mu)}),$$

$$\Psi(\beta_t) = \prod_{i \in I_t} \psi(\beta_i),$$

where $t = 1, 2, \ldots, N/(w - 1)$, and where $\mathcal{T}$ in $\Phi_{\mathcal{T}}$ is a shorthand notation for the pair $(t, t+1)$ and expresses that we regard this clustered potential function to be located between $\beta_t$ and $\beta_{t+1}$ (note that the right-hand side of (18) depends on $\beta$ only through $\beta_t$ and $\beta_{t+1}$). Then, the original factor graph for BM-lasso (e.g., Fig. 2 (c)) can be redrawn in terms of these clustered variables and clustered potential functions as the one-dimensional chain shown in Fig. 3. In other words, the clustered factor graph can be regarded as a Markov model of order 1, with a sequence $(\beta_t)$ of states. Accordingly, the posterior distribution (13) is rewritten in terms of the clustered expression as

$$P_{\gamma}(\beta|X_w, y) = \frac{1}{Z} \left\{ \prod_{t=1}^{N/(w-1)-1} \Phi_{\mathcal{T}}(\beta_t, \beta_{t+1} | y_t) \right\} \Phi_{\mathcal{N}}(\beta | y_{N-1})$$

$$\times \prod_{\mu \in N/(w-1)} \phi_{\mu}(y_\mu | x_\mu^{(\mu)}, y_\mu),$$

where

$$\Phi_{\mathcal{N}}(\beta | y_{N-1}) = \prod_{\mu \in N/(w-1)} \phi_{\mu}(\beta_\mu | x_\mu^{(\mu)}, y_\mu).$$

We apply BP on this clustered factor graph and obtain the following set of equations:

$$\mathcal{M}_{t \to t+1}(\beta_{t+1}) = \int \Phi_{\mathcal{T}}(\beta_t, \beta_{t+1} | y_t) \mathcal{M}_{t \to t}(\beta_t) d\beta_t,$$

$$\mathcal{M}_{t \to t}(\beta_{t}) = \frac{1}{Z_t} \Psi(\beta_t) \mathcal{M}_{t \to t}(\beta_t),$$

where

$$Z_t = \int \Psi(\beta_t) \mathcal{M}_{t \to t}(\beta_t) d\beta_t,$$

and $d\beta_t = \prod_{i \in I_t} d\beta_i$. We call $\mathcal{M}_{t \to t}$ a message which expresses a local marginal distribution of $\beta_t$ when $t$th clustered potential function $\Phi_{\mathcal{T}}$ is absent. We also term $\mathcal{M}$ a bias message. These recurrence equations are solved recursively in the forward direction along the chain, starting from

$$\mathcal{M}_{1 \to 1}(\beta_{1}) = \frac{1}{Z_1} \Psi(\beta_1), \quad Z_1 = \int \Psi(\beta_1) d\beta_1.$$

Similarly, we can consider backward messages $\mathcal{M}'_{t+1 \to t}$ and backward bias messages $\mathcal{M}'_{t \to t}$, as well as recurrence
We call these quantities the rates of the forward bias messages and the forward messages, assuming that these quantities are $O(1)$ in the limit $\gamma \to \infty$. Rewriting (22) and (23) in terms of these rates in the large-$\gamma$ asymptotics yields

$$
\hat{J}_{\tau \rightarrow t+1}(\beta_{t+1}) = \max_{\beta_t'} \left\{ \frac{1}{\gamma} \log \Phi_{\tau}(\beta_t',\beta_{t+1} | \eta_t) + J_{t \rightarrow \tau}(\beta_t') \right\},
$$

$$
J_{t \rightarrow \tau}(\beta_t) = \frac{1}{\gamma} \log \Psi(\beta_t) + \hat{J}_{\tau \rightarrow t}(\beta_t)
$$

$$
- \max_{\beta_t'} \left\{ \frac{1}{\gamma} \log \Phi_{\tau}(\beta_t') + \hat{J}_{\tau \rightarrow t}(\beta_t') \right\},
$$

where the maximization with respect to $\beta_t'$ in (30) and (31) appears through the Laplace/saddle-point method in the limit $\gamma \to \infty$ applied to the respective integrative in (22) and (24). We also have similar reformulations for the backward messages and the backward bias messages, yielding recurrence equations for the rates $\hat{J}_{t+1 \rightarrow \tau}(\beta_{t+1})$ and $J_{t+1 \rightarrow \tau}(\beta_{t+1})$ of the backward bias messages and the backward messages, respectively. The marginal distribution of a clustered-variable node $P_{\beta_t}(\beta_t)$ is also characterized in the asymptotics as $P_{\beta_t}(\beta_t) \propto e^{\gamma J_t(\beta_t)}$, the rate of which is given by

$$
J_t(\beta_t) = \frac{1}{\gamma} \log \Psi(\beta_t) + \hat{J}_{\tau \rightarrow t}(\beta_t) + \hat{J}_{t \rightarrow \tau}(\beta_t)
$$

$$
- \max_{\beta_t'} \left\{ \frac{1}{\gamma} \log \Phi_{\tau}(\beta_t') + \hat{J}_{\tau \rightarrow t}(\beta_t') + \hat{J}_{t \rightarrow \tau}(\beta_t') \right\}.
$$

These yield the BP algorithm in the limit $\gamma \to \infty$. The component $\hat{\beta}_\lambda$ of the lasso estimator corresponding to the index set $I_\lambda$ is obtained as the mean of the marginal posterior $P_{\beta_t}(\beta_t)$ in the limit $\gamma \to \infty$, which can be evaluated by applying the Laplace/saddle-point method, yielding

$$
\left( \hat{\beta}_\lambda \right)_{I_\lambda} = \arg \max_{\beta_t} \left\{ J_t(\beta_t) \right\}.
$$

Thus, we can obtain the BM-lasso estimate $\hat{\beta}_\lambda$ via the rates.

One might consider applying the idea of the approximate message passing (AMP) [38], [41], [42], a widely used approximation technique to avoid complex calculations in BP, to obtain simpler formulas than the BP just formulated. Unfortunately, it does not work well in our case. The AMP is based on the Gaussian approximation of the messages $\hat{M}$ and $M$, which would allow us to convert the functional updates of the BP (22)–(24) to the updates of the parameters characterizing the Gaussians. This approximation is justified by the independence of the incoming messages into a factor node and the central limit theorem based on the large number of incoming messages. However, in our clustered graphical model (Fig. 3), the number of the incoming message is only one; thus there is no place in our BP formulation where one can apply the Gaussian approximation yielding the AMP. One might alternatively consider applying the AMP idea to the original factor graph (Fig. 2). It turns out not to work, either, because the number of messages sent to a particular node is limited (at most $2(w-1)$ messages) and the independence of the incoming messages does not hold. Thus, we evaluate the BP equations (22)–(24) directly in the theoretical analysis.
D. Density Evolution

Although the BP algorithm is certainly applicable to evaluate lasso estimates, our primary objective of formulating the BP algorithm is not to propose it as a practical algorithm, as several algorithms based on convex optimization would find a solution efficiently. Rather, we would like to utilize the BP algorithm in order to analyze performance of the lasso estimator. For this purpose, we employ what is called DE, which is a method to systematically analyze average-case behaviors of the BP algorithm. In the discussion of DE below, we focus on the large-

$N$ dimensional regime and thus the approach from the large-$N$ limit because of the following two reasons. One is that some natural performance measures such as the MSE $\|\beta^* - \beta\|_2^2/N$ obey the law of large numbers or the self-averaging property under some appropriate setting on the distributions of $\beta^*$ and $\epsilon$; hence the quantitative discussion free from stochastic variations is possible in the limit. The other is that the expected application of BM-lasso is in the high-dimensional regime and thus the approach from the large-$N$ limit is quantitatively comparable and practically useful.

Let us introduce DE. The recurrence equations in the BP algorithm are deterministic when $y$ is fixed. However, assuming the measurement model (3) with noise and a prior distribution of $\beta^*$, one can regard $y$ as being stochastic. To evaluate the estimation risk, we need to consider the stochasticity of $y$, and consequently that of the rates and $\beta_t$. DE is a method that considers the distributions of the messages (or the rates) in a message-passing algorithm like BP and tracks the evolution of the distributions under a certain stochasticity assumption.

As mentioned above, the stochasticity of $y$ under the measurement model (3) is governed by the stochasticity of $(\beta^*, \epsilon)$. For the stochasticity of $(\beta^*, \epsilon)$, we assume that the elements of $\beta^*$ are independent and identically distributed (i.i.d.), and that the elements of $\epsilon$ are also i.i.d., that is,

$$P_{\beta_t}(\beta^*) = \prod_{i=1}^{N} P_{\beta_t}(\beta^*_i),$$

$$P_{\epsilon}(\epsilon) = \prod_{i=1}^{N} P_{\epsilon}(\epsilon_i).$$

(34)

(35)

Hereafter, conditional distributions are denoted according to the standard way: for example, the distribution of $J_{t-1}$ conditioned by $\beta_t^*$ is denoted as $P_{J_{t-1} \mid \beta_t^*}$.

Using these notations, we derive the update rule of the distributions of the rates $J$ and $\tilde{J}$, which corresponds to the transition probabilities of the Markov chain. According to (3), $y_t$ is given as a function of $(\beta_t^*, \beta_{t+1}^*)$ and $\epsilon_t$, that is, $y_t = \tilde{y}_t(\beta_t^*, \beta_{t+1}^*, \epsilon_t)$. From the discussion in Sec. III-C and the i.i.d. assumption above, the directed graphical model for the rates $J$ and $\tilde{J}$ and the BM-lasso estimate is described in Fig. 4. Note that the distributions of the forward and backward rates $\tilde{J}_{t-1} \rightarrow t$ and $\tilde{J}_{t} \rightarrow t$ that are to be sent to a target clustered node $t$ become independent if conditioned on $\beta_t^*$. Moreover, it is clear that the conditional rate distributions $P_{J_{t-1} \rightarrow t}(\tilde{J} \mid \beta_t^*)$, $P_{J_{t} \rightarrow t}(J \mid \beta_t^*)$ have Markov property. Hence, we employ the conditional rate distributions $P_{J_{t-1} \rightarrow t}(\tilde{J} \mid \beta_t^*)$, $P_{J_{t} \rightarrow t}(J \mid \beta_t^*)$ as DE objects instead of the conventional DE objects $P_{J_{t-1} \rightarrow t}(\tilde{J})$, $P_{J_{t} \rightarrow t}(J)$, and below we describe the update rule of these distributions.

Suppose $P_{J_{t-1} \rightarrow t}(\tilde{J} \mid \beta_t^*)$ is given. Then $P_{J_{t} \rightarrow t}(J \mid \beta_t^*)$ is derived as the pushforward measure of $P_{J_{t-1} \rightarrow t}(\tilde{J} \mid \beta_t^*)$ via the mapping (31). Putting aside mathematical rigor, this could conceptually be written as

$$P_{J_{t} \rightarrow t}(J \mid \beta_t^*) = \int D\tilde{J} \prod_{\beta_t} \left[ \delta \left( J(\beta_t) - \{ -\lambda \|\beta_t\|_1 + \tilde{J}(\beta_t) \} \right) \right] P_{J_{t-1} \rightarrow t}(\tilde{J} \mid \beta_t^*),$$

(40)

where $\prod_{\beta_t}$ denotes the product over all configurations of $\beta_t$, and where $D\tilde{J} = \prod_{\beta_t} d\tilde{J}(\beta_t)^2$.

As shown in Fig. 4, the distribution of $\tilde{J}_{t-1} \rightarrow t+1$ conditional on $\beta_t^*, \beta_{t+1}^*, \epsilon_t$ is the pushforward measure of $P_{J_{t} \rightarrow t}(J \mid \beta_t^*)$.
via the mapping (30):
\[
P_{\tilde{J}_{t+1} | \tilde{J}_t}(\tilde{J} | \beta_t^*, \beta_{t+1}^*, \epsilon_t)
= \int D\tilde{J} \prod_{\beta_t} \left[ \delta(\tilde{J}(\beta_{t+1})) 
- \max_{\beta_t^*} \left\{ \frac{1}{\gamma} \log P_\epsilon(\beta_t', \beta_{t+1}^* | y_t) + J(\beta_t') \right\} \right] P_{\epsilon_{t+1}}(J | \beta_t).
\]

By taking the average over \(\beta_t^*\) and \(\epsilon_t\), we obtain
\[
P_{\tilde{J}_{t+1} | \tilde{J}_t}(\tilde{J} | \beta_{t+1}^*)
= \int d\beta_t^* d\epsilon_t P_{\beta_t}^1(\beta_t^*)P_{\epsilon_t}(\epsilon_t)P_{\tilde{J}_{t+1} | \tilde{J}_t}(\tilde{J} | \beta_t^*, \beta_{t+1}^*, \epsilon_t) .
\]

Thus, the update rule for \(P_{\tilde{J}_{t+1} | \tilde{J}_t}(\tilde{J} | \beta_t^*)\) is derived from (40)–(42). Similar formulas can be obtained for the distributions of the backward rates. These constitute our DE equations.

As mentioned at the beginning of this subsection, the graphical model (Fig. 4) defines a Markov chain with the states \((\tilde{J}_{t-1}^\dagger, \tilde{J}_{t-1}^\prime, t)_{t=1}^{N/(w-1)}\). It is homogeneous in \(t\) under the i.i.d. assumptions (34) and (35) on the elements of \(\beta^*\) and \(\epsilon\). We further assume ergodicity of the Markov chain. It will allow performance evaluation in the large-\(N\) asymptotics, where one may be able to evaluate the stationary distributions of the message rates and safely ignore the transient behaviors of the Markov process which would reside near the ends of the one-dimensional chain. Then, we expect that \(P_{\tilde{J}_{t+1} | \tilde{J}_t}(\tilde{J} | \beta_t^*)\) converges to a stationary distribution
\[
P_{\tilde{J}_{t+1} | \tilde{J}_t}(\tilde{J} | \beta_t^*) \rightarrow P_{\tilde{J}}(\tilde{J} | \beta^*)\quad \text{as } t \text{ increases.}
\]
Convergence to a stationary distribution is also expected for the distributions of the backward rates, and we assume this stationarity hereafter. This implies that the rates located in the bulk of the one-dimensional chain (i.e., far enough from the boundaries) obey these stationary distributions. In the large-\(N\) limit, the contributions from the bulk will dominate the overall performance, with negligible contributions from the near-boundary regions.

From the above discussion, any clustered variable located in the bulk follows the same distribution irrespective of its position in the one-dimensional chain. More concretely, let us consider an arbitrary clustered variable \(\beta_{loc} \in \mathbb{R}^{w-1}\) located in the bulk. Following the prescription for BP, one sees that the marginal distribution of \(\beta_{loc}\) conditional on the corresponding true signal \(\beta^*_{loc} \in \mathbb{R}^{w-1}\) is given, in terms of the stationary rate distributions, by
\[
P_{\beta_{loc} | \beta^*_{loc}}(\beta_{loc} | \beta^*_{loc})
\rightarrow \infty, N \rightarrow \infty, \quad \int D\tilde{J} \prod_{\tilde{J}} P_{\tilde{J}}(\tilde{J} | \beta^*_{loc}) P_{\tilde{J}}(\tilde{J} | \beta^*_{loc})
\times \delta(\beta_{loc} - \beta_{loc}(\tilde{J}, \beta^*_{loc})),
\]
where
\[
\beta_{loc}(\tilde{J}, \beta^*_{loc}) = \arg \max_{\beta_{loc}} \left\{ -\lambda \| \beta_{loc} \|_1 + \tilde{J} \beta_{loc} \right\}
+ \tilde{J} \beta^*_{loc},
\]
which corresponds to (32). From this distribution of the estimator, the estimation risk \(\mathbb{E}[(\beta^* - \beta)^2] / N\) in the limit \(N \rightarrow \infty\) can be derived as follows:
\[
\mathbb{E} \left[ \frac{\| \beta^* - \beta \|^2}{N} \right]_{N \rightarrow \infty} = \mathbb{E} \left[ \frac{\| \beta_{loc} - \beta_{loc} \|^2}{w-1} \right]
= \int d\beta_{loc}^* d\beta_{loc} P_{\beta}^1(\beta_{loc}^*)P_{\beta_{loc}^*}(\beta_{loc} | \beta_{loc}) \| \beta_{loc}^* - \beta_{loc} \|^2 / w-1.
\]
Hence, for computing the estimation risk, it is sufficient to obtain the stationary distribution \(P_{\tilde{J}}(\tilde{J} | \beta^*)\) from the DE equations.

### E. Method for Solving DE Equations

Unfortunately, it is highly nontrivial to compute the stationary distribution from the update rule (40)–(42) both analytically and numerically. The origin of the difficulty is that the distribution \(P_{\tilde{J}_{t+1} | \tilde{J}_t}(\tilde{J} | \beta_t^*)\) is a functional of \(\tilde{J}_{t+1} | \tilde{J}_t(\beta_t)\), and \(\tilde{J}(\beta_t)\) are represented as order-\((w-1)\) tensors by appropriately restricting and discretizing the domain of \(\beta_t\). Further, we borrow the idea of particle filters [43] to represent probability distributions in terms of collections of points, that is, we represent the distributions \(P_{\tilde{J}}\) and \(P_{\tilde{J}^*}\) in terms of populations of the tensors of \(\tilde{J}\) and \(\tilde{J}^*\), and update them via a Monte Carlo algorithm which is straightforwardly implemented by using the Markovian properties of the conditional rates. With the approximation and the algorithm, we can numerically evaluate \(P_{\tilde{J}}\) in a systematic manner.

Although this strategy is feasible in principle for any \(w\), some computational issue prevents us from taking the above approach with large \(w\). Suppose we discretize the domain of \(\beta_t\) into \(L\) bins. In that case, the dimensionality of the discretized cluster variable \(\beta_t\) becomes \(L^{w-1}\), and hence the DE equations require summations and optimizations over a space of huge dimension when \(w\) is large. The discretization is also necessary for treating \(\beta^*\), which further increases the computational cost. These restrict us from applying the above approach with large \(w\). Due to this computational issue, in the following we specifically show the algorithm with \(w = 2\), whereas it should be noted that, putting aside the computational issue, it would be straightforward to extend the algorithm to the cases with \(w > 2\) by considering the order-\((w-1)\) tensors instead of the \(L\)-dimensional vectors.

When \(w = 2\), the clustered variable is scalar, \(\beta_t = \beta_t^1\). Let us write the discretized vector representations of \(\beta\) and \(\beta^*\) as \(\tilde{\beta} = [z_1, \ldots, z_L]^\top \in \mathbb{R}^{L^2}\) and \(\tilde{\beta}^* = [z_1^*, \ldots, z_L^*]^\top\), respectively; \(L_z\) and \(L_z^*\) are the dimensions of those vectors, which may be different from each other in general. Correspondingly, \(\tilde{J}(\beta_t)\) and \(\tilde{J}(\beta_t^*)\) also become \(L_z\)-dimensional vectors, and are denoted as \(\tilde{J}_t\) and \(\tilde{J}_t^*\), respectively. The distribution \(P_{\tilde{J}_{t+1} | \tilde{J}_t}(\tilde{J} | \beta_t)\) is represented by a population
Algorithm 1 MC algorithm for DE with $w = 2$

Input: $\bar{x}^*, \bar{z}, \lambda, N_{\text{pop}}$ and $T$

Output: $\{(\tilde{\beta}^{(i)}_T | z^*_a)^{N_{\text{pop}}}_{i=1} | a \in \{1, \ldots, L_z\}\}$

1: Construct the initial population $\{(\tilde{\beta}^{(i)}_0 | z^*_a)^{N_{\text{pop}}}_{i=1} | b \in \{1, \ldots, L_z\}\}$ by $\tilde{\beta}^{(i)}_0 = -\lambda \times \text{abs}(\bar{z})$, where $\text{abs}(\cdot)$ is the function which returns the absolute value of the given argument componentwise.
2: for $t = 1$ to $T - 1$
3: for $a \in \{1, \ldots, L_z\}$
4: for $b \in \{1, \ldots, L_z^*\}$
5: for $j = 1$ to $N_{\text{pop}}$
6: Draw $\tilde{J}^{(j)}_{t+1}$ uniformly from $\{\tilde{J}^{(i)}_t | z^*_a \}^{N_{\text{pop}}}_{i=1}$.
7: Generate $\epsilon_t$ from $P_\epsilon$.
8: $F \leftarrow F(z^*_b, \epsilon_t)$
9: $\tilde{J}^{(j)}_{t+1} \leftarrow \max \left\{ F + \text{repeat}(\tilde{J}^{(j)}_t, L_z) \right\}$
   where $\text{repeat}(\tilde{J}^{(j)}_t, L_z)$ denotes a $L_z \times L_z$ matrix whose each row vector is identical to $\tilde{J}^{(j)}_t$.
10: end for
11: Collect $J^{(j)}_{t+1}$ as $\{(\tilde{J}^{(j)}_{t+1} | z^*_a, z^*_b)^{N_{\text{pop}}}_{i=1} | b \in \{1, \ldots, L_z^*\}\}$
12: end for
13: for $k = 1$ to $N_{\text{pop}}$
14: Draw $\tilde{J}^{(k)}_t$ from $\{\{(\tilde{J}^{(i)}_{t+1} | z^*_a, z^*_b)^{N_{\text{pop}}}_{i=1}\} | b \in \{1, \ldots, L_z^*\}\}$ according to $P_{\beta^*}$.
15: end for
16: Collect $J^{(k)}_{t+1}$ as $\{(\tilde{J}^{(i)}_{t+1} | z^*_a, z^*_b)^{N_{\text{pop}}}_{i=1}\} | b \in \{1, \ldots, L_z^*\}$
17: end for
18: for $l = 1$ to $N_{\text{pop}}$
19: $\tilde{J}^{(l)}_{t+1} \leftarrow -\lambda |\bar{z}| + \tilde{J}^{(l)}_{t+1} - \max \left\{-\lambda |\bar{z}| + \tilde{J}^{(l)}_{t+1}\right\}$
20: end for
21: Collect $J^{(l)}_{t+1}$ as $\{(\tilde{J}^{(i)}_{t+1} | z^*_a)^{N_{\text{pop}}}_{i=1}\} | a \in \{1, \ldots, L_z\}$.
22: Obtain $\{(\tilde{J}^{(i)}_{t+1} | z^*_a)^{N_{\text{pop}}}_{i=1} | a \in \{1, \ldots, L_z\}\}$.
23: end for
24: return $\{(\tilde{J}^{(i)}_T | z^*_a)^{N_{\text{pop}}}_{i=1} | a \in \{1, \ldots, L_z\}\}$.

In both cases, we conducted numerical experiments with finite $N$. In the simulations, we obtained BM-lasso estimate (7) by performing fast iterative shrinkage-thresholding algorithm (FISTA) [44] which is well-known to be efficient and reliable for obtaining the lasso estimate. We evaluated the superiority of the proposed method to the conventional measurement method by comparing the performance of $w \geq 2$ with that of $w = 1$ because a BM-lasso estimate with $w = 1$ can be regarded as a result of the conventional measurement method as discussed in Sec. II-B.

B. On Fully Artificial Model

Here we examined the performance of BM-lasso on a fully artificial model. In this model, the elements of the true signal $\beta^*$ were assumed to independently follow the distribution

$$P_{\beta^*}(\beta^*_a) = (1 - \rho) \delta(\beta^*_a) + \rho \delta(\beta^*_a - K),$$

where $K > 0$ is the magnitude of non-zero elements and where $\rho \in [0, 1]$ controls the ratio of nonzero elements in
A notable behavior is that excellent, even though the DE-based analysis is based on the solid lines) estimate (7) was obtained with the true signal and the noise $N$ conducted numerical experiments: the experiments were conducted with $w = 1$. The black dotted line shows the estimation risk when $\hat{\beta} = 0_N$.

$\beta^*$ such that $\mathbb{E}\|\beta^*\|_0 = N\rho$ holds. This model is simple yet nontrivial enough to examine the performance of the sparse estimation. Thanks to the simplicity, the discretization of the values of $\beta^*$ explained in Sec. III-E is not necessary when conducting the DE-based analysis. The measurement noise vector $\epsilon$ in this model was assumed to consist of elements that independently follow the zero-mean Gaussian distribution with variance $\sigma^2$:

$$P_\epsilon(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\epsilon_i^2}{2\sigma^2}}. \quad (48)$$

1) DE-based analysis with $w = 1, 2$: In Fig. 5, we show the plots of the MSES, which were computed with the DE-based analysis discussed in Sec. III, against the noise variance $\sigma^2$ for $w = 1$ and 2. The parameters here are $K = 5, \rho = 0.1, \lambda = 1$. A notable behavior is that the estimation risk with $w = 2$ was lower than that with $w = 1$ for all the values of $\sigma^2$ investigated, providing a firm theoretical basis supporting the superiority of BM-lasso against the conventional measurement method. For reference, the lines of the OLS estimator with $w = 1$, i.e., $\hat{\beta} = y$, and the zero estimator $\hat{\beta} = 0$ are also provided. Although the OLS result seems to be better than BM-lasso at small values of $\sigma^2$, this is because we fixed the value of the regularization parameter $\lambda$; if we optimized the value of the regularization parameter $\lambda$, the BM-lasso result became better than that of OLS.

To check the validity of our DE-based analysis, we also conducted numerical experiments: the experiments were conducted on the system with $N = 400$, and the BM-lasso estimate (7) was obtained with the true signal and the noise generated according to (47) and (48) respectively; 200 different realizations were generated to compute the mean of MSE and its standard error. The results are shown as markers in Fig. 5. The agreement between the DE-based analysis (the solid lines) and the numerical experiments (the markers) is excellent, even though the DE-based analysis is based on the analysis of the stationary distribution of the Markov chain (Fig. 4), thereby validating our theoretical treatment as well as our numerical implementation.

2) Numerical Simulations for Several $w$ and $\lambda\text{AIC}$: Next, we examined the performance of BM-lasso for a wide range of values of $w$. As we mentioned that the DE-based analysis is only computationally feasible for small $w$, here we relied on numerical experiments. The numerical experiments were conducted with $N = 400$, which should be sufficiently large to observe large-system behaviors of the BM-lasso, as demonstrated by the excellent agreement for the case $w = 2$ between the experimental results with $N = 400$ and the DE-based analysis. Besides, we checked the behavior of AIC for lasso (11). AIC for lasso is not an unbiased estimator of the MSE $\|\beta^* - \hat{\beta}_\lambda\|^2 / N$ and hence it is desirable to check whether the choice $\lambda = \lambda\text{AIC}$ gives a sufficiently small MSE compared with other values of $\lambda$. The experimental results shown here are the averaged results with the standard error obtained from 200 trials as above.

Figure 6 shows the $w$- and $\lambda$-dependence of AIC and MSE. The inverted triangles in the figure show the location of $\lambda\text{AIC}$. For each bandwidth $w$, $\lambda\text{AIC}$ was located in the region such that the MSE was small enough to be regarded as almost minimum. Thus, it was ensured that AIC for lasso worked as an appropriate hyperparameter selection criterion for the present problem.

To further evaluate the AIC-selected model, we calculated the MSE and the F1 score with $\lambda = \lambda\text{AIC}(w)$ for each bandwidth $w$ as shown in Table I. Here, the F1 score is defined as the harmonic mean of precision and recall in binary classification classifying presence/absence of non-zero elements in $\beta^*$ according to whether the corresponding elements in $\hat{\beta}_\lambda$ with $\lambda = \lambda\text{AIC}$ are non-zero or zero. We call this classification task the peak classification. Compared with the conventional measurement ($w = 1$), the proposed method with $w \geq 2$ significantly improved the MSE and the F1 score. These results indicated that the proposed method with the AIC-selected model achieved much higher estimation accuracy and performed a more appropriate peak classification.

To examine the appropriate range of bandwidth $w$, we plot the MSE values at $\lambda = \lambda\text{AIC}$ against $w$ in Fig. 7. The result shows that the MSE rapidly decreased as $w$ grew from $w = 1$, and it seemed to take a minimum around $w = 4$, and finally, it tended to grow slowly as $w$ increased. This implies that slight modifications of the bandwidth of the conventional measurement may significantly improve the estimation. Moreover, if the bandwidth is too large, the estimation becomes worse because of the ill-conditionness of the problem to be solved.

C. On Actual Mass Spectra with Artificial Noise

Here we applied the proposed method to the estimation of actual mass spectra. The data used in the experiments is one of the mass spectrometry datasets registered in Massbank, a public database for mass spectrometry. Specifically, as a true mass spectra dataset we used the dataset for a contest named Critical Assessment of Small Molecule Identification held in 2012 (CASM2012), which contains 26 mass spectra. Each of these spectra was converted from the peak-list format to
form a 3000-dimensional vector, covering the range of m/z from 50 to 650, and normalized so that the maximum signal intensity becomes 100. The measurement process of the BM was simulated according to (3) and (48) with $\sigma = 1$.

Figure 8 shows the $\lambda$-dependence on AIC and MSE. As in the fully artificial case, $\lambda_{AIC}$ was located in the region where the MSE was small enough. Thus, it was ensured that AIC for lasso is applicable to actual mass analysis as a hyperparameter selection criterion.

Table II shows the statistics which characterize the AIC-selected model for each bandwidth. From the comparison of the MSE, it was revealed that the proposed method ($w \geq 2$) was superior to the conventional one ($w = 1$) as expected. From the F1 score, BM-lasso with $w = 25$–30 appeared to achieve the best performance. Here, this optimal bandwidth was much larger than that of the fully artificial model in Sec. IV-B-2) ($w = 4$). We consider that one of the significant factors to make this difference in the optimal bandwidth is the sparsity of the original spectrum (e.g., the actual spectra used in this section are sparser ($\rho \approx 0.004$) than the artificial
shows that the proposed method (w like measurement window, in which the BM is combined with applicable to the typical spectrometric systems employing a slit-like measurement window, in which the BM is combined with also observed in Table III. The improvement in the sensitivity is times compared with the conventional of peak detection. In particular, the BM-lasso model with for w = 30 was found to improve the sensitivity by about 7

Fig. 7: Effect of bandwidth w on the MSE of BM-lasso with the artificial spectra (λ = λ_{AIC}).

spectra (ρ = 0.1)), that is, the wider BM is better when the original spectrum is sparser. This consideration was supported by the result of an additional experiment showing that the optimal bandwidth for artificial spectra with ρ = 0.01 was w = 20 (not shown).

To gain further insights into the peak classification capability of the BM-lasso model, we evaluated the ROC curve for each bandwidth. As shown in Fig. 9, the model with w = 30 showed the best performance for peak classification, in accordance with the results of MSE shown above. Moreover, the points corresponding to = λ_{AIC} were located upper left of the respective ROC curves. This suggests that the values of the hyperparameter λ selected via AIC were also appropriate in terms of peak classification. Besides, comparing these points, it is observed that the proposed method (w ≥ 2) improved the true positive rate while the false positive rate was maintained. This result implies that the improvement of the F1 score in Table II resulted from that of the sensitivity. Figure 10 also shows that the proposed method (w = 30) tended to be able to detect weak peaks which were comparable to the noise signal in contrast to the conventional measurement method (w = 1).

One can argue that, with the lasso formulation (7), the quantity λ/w can be regarded as an indicator of peak detection limit (see Appendix). Once accepting this, we found that λ_{AIC}/w of the proposed method (w ≥ 2) were lower than that for w = 1, which means an improvement in the sensitivity of peak detection. In particular, the BM-lasso model with w = 30 was found to improve the sensitivity by about 7 times compared with the conventional measurement method as shown in Table II. The improvement in the sensitivity is also observed in Table III.

V. CONCLUSION

In this paper, we proposed a novel spectrometric method applicable to the typical spectrometric systems employing a slit-like measurement window, in which the BM is combined with sparse-modeling-based post-processing to obtain the original high-resolution spectrum. To evaluate the effectiveness of this proposal both theoretically and experimentally, we formulated the sparse-modeling-based post-processing in the proposal in terms of the lasso and performed a theoretical analysis and simulation studies.

In the theoretical part, we gave the BM-lasso model a probability-model-based interpretation and derived the DE formulas for BP on the high-order Markov chain with non-Gaussian state distribution. We then obtained the asymptotic estimation risk of BM-lasso and performed analytical evaluations of the proposed method. Comparison of the DE-based evaluation of MSEs for w = 1 and 2 showed the superiority of w = 2 for all over the range of the signal-to-noise ratios. Numerical experiments at N = 400 were also conducted and the result showed an excellent agreement with the DE-based analysis, validating our theoretical treatment as well as our numerical implementation.

The results of numerical simulations on artificial spectra provided a more detailed quantitative perspective. AIC for lasso was proved to be an appropriate hyperparameter selection criterion, and it was shown that there exists an appropriate range of w in terms of MSE and F1 score. The overall performance of BM-lasso with appropriate w was clearly superior to that of the conventional w = 1 case. Additional numerical experiments on actual mass spectra further supported this conclusion. These results give a significant insight that the measurement window should be widened appropriately to improve the sensitivity and accuracy of the measurement in a limited measurement time.

As a future direction, it is interesting to implement BM-lasso in actual spectrometers. When doing so, some problems will emerge in the modeling level, such as the choice of the objective function which should reflect the real signal generation process. Overcoming such problems by taking the characteristics of real spectrometric data into account is the next step of the presented framework, and such a trial for QMS data is in progress. Another interesting direction is to combine BM with the existing DIA techniques, which will develop new applications in proteomics.

ACKNOWLEDGMENTS

The authors would like to thank Yasushi Ishihama and Kosuke Ogata for their discussion and comments about proteomics and mass spectroscopy. This work was partially supported by JST, CREST Grant number JPMJCR1862, Japan (KU, TO, TT) and JSPS KAKENHI Nos. 22H05117, 22K12179, 19H01812, 18K11463, 17H00764 (TO) and 21J23154 (KU).
TABLE III: Effect of the signal strength on the true positive rate ($\lambda = \lambda_{\text{AIC}}$). The numbers in parentheses represent the ratio of the number of detected peaks to the total number of peaks.

<table>
<thead>
<tr>
<th>$K$</th>
<th>$w = 1$</th>
<th>$w = 5$</th>
<th>$w = 10$</th>
<th>$w = 20$</th>
<th>$w = 30$</th>
<th>$w = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0-0.5</td>
<td>0.0% (0/41)</td>
<td>7.3% (3/41)</td>
<td>7.3% (3/41)</td>
<td>17% (7/41)</td>
<td>27% (11/41)</td>
<td>17% (7/41)</td>
</tr>
<tr>
<td>0.5-1.0</td>
<td>5.3% (1/19)</td>
<td>5.3% (1/19)</td>
<td>47% (9/19)</td>
<td>42% (8/19)</td>
<td>53% (10/19)</td>
<td>42% (8/19)</td>
</tr>
<tr>
<td>1.0-1.5</td>
<td>18% (4/22)</td>
<td>45% (10/22)</td>
<td>59% (13/22)</td>
<td>72% (16/22)</td>
<td>77% (17/22)</td>
<td>68% (15/22)</td>
</tr>
<tr>
<td>1.5-2.0</td>
<td>20% (3/15)</td>
<td>73% (11/15)</td>
<td>93% (14/15)</td>
<td>93% (14/15)</td>
<td>93% (14/15)</td>
<td>93% (14/15)</td>
</tr>
<tr>
<td>2.0-3.0</td>
<td>55% (11/20)</td>
<td>85% (17/20)</td>
<td>90% (18/20)</td>
<td>100% (20/20)</td>
<td>95% (19/20)</td>
<td>100% (20/20)</td>
</tr>
<tr>
<td>3.0-5.0</td>
<td>96% (22/23)</td>
<td>100% (23/23)</td>
<td>100% (23/23)</td>
<td>100% (23/23)</td>
<td>100% (23/23)</td>
<td>96% (22/23)</td>
</tr>
<tr>
<td>5.0-100</td>
<td>100% (168/168)</td>
<td>100% (168/168)</td>
<td>100% (168/168)</td>
<td>100% (168/168)</td>
<td>100% (168/168)</td>
<td>100% (168/168)</td>
</tr>
</tbody>
</table>

APPENDIX

A. Soft Threshold of BM-lasso

To evaluate a signal detection limit of BM-lasso, we consider a soft threshold of a BM-lasso model below.

Since the objective function $H(\beta | y, X_w, \lambda)$ of BM-lasso, given in (7), is convex downward in $\beta$, the lasso solution $\hat{\beta} = \arg \min_\beta H(\beta | y, X_w, \lambda)$ satisfies the stationarity condition $\frac{\partial H(\beta)}{\partial \beta_j} |_{\beta = \hat{\beta}} = 0$, which reads:

$$\hat{\beta}_j = \frac{1}{d_j} \left\{ \sum_{i=1}^{N} x_{ij} (y_i - \sum_{k\neq j} \hat{\beta}_k x_{ik}) - \lambda d_j \right\},$$

(49)
From the definition of a band matrix (4), if $N = w$, a soft threshold.

Here, we introduce the soft-thresholding function $S(z)$ and where $d_j$ denotes the subgradient of $|\beta_j|$ at $\beta_j = \hat{\beta}_j$, as $d_j = \begin{cases} -1 & (\beta_j < 0) \\ \in [-1, 1] & (\beta_j = 0) \\ 1 & (\beta_j > 0) \end{cases}$.

Thus, according to (53), we obtain the macroscopic BM-lasso estimate

$$\hat{\beta}_j = S \left( \frac{\sum_{i=1}^{N} x_{ij} (y_i - \sum_{k \neq j} \hat{\beta}_k x_{ik})}{a_j}, \frac{\lambda}{a_j} \right),$$

with the soft threshold $\lambda/w$, implying that $\lambda/w$ can be regarded as an indicator of a signal detection limit.

**References**


See the attached PDF for the full content.