

RESEARCH

Open Access



DL-CHI: a dictionary learning-based contemporaneous health index for degenerative disease monitoring

Aven Samareh*  and Shuai Huang

Abstract

Effective monitoring of degenerative patient conditions is crucial for many clinical decision-making problems. Leveraging the nowadays data-rich environments in many clinical settings, in this paper, we propose a novel clinical data fusion framework that can build a contemporaneous health index (CHI) for degenerative disease monitoring to quantify the severity of deterioration process over time. Our framework specifically exploits the monotonic progression patterns of the target degenerative disease conditions such as the Alzheimer's disease (AD) and articulate these patterns with a systematic optimization formulation. Further, to address the patient heterogeneity, we integrate CHI with dictionary learning to build sets of overcomplete bases to represent the personalized models efficiently. Numerical performances on two real-world applications show the promising capability of the proposed DL-CHI model.

Keywords: Dictionary learning, Patient monitoring, Convex optimization, Personalized healthcare

1 Introduction

In this paper, we concern the problem of patient risk monitoring that is to characterize the trajectory over the course of progression. Although there is no universal definition of the concept “patient condition,” it has been a crucial concept in the communications between clinicians and frequently referenced by healthcare providers. Developing a precise contemporaneous longitudinal index (CHI) that can faithfully reflect the underlying patient condition across the course of the condition's progression holds great value for facilitating a range of clinical decision-makings. For instance, it will help early detection of patient deterioration to help reduce the number of serious incidents, i.e., it is reported that 11% of serious incidents are a function of deterioration not acted upon mainly due to the failure to recognize the sign of deterioration [1, 2]. It will also help enhance the continuity of care since a longitudinal perspective of the patient condition can be provided for clinicians and healthcare providers. Also, it may ultimately lead to development of control system engineering that can implement adaptive

interventions for better healthcare management [3–6], with a global representation of the dynamic condition in evolution.

Towards this goal, technological innovations are emerging in many healthcare applications, which have given rise to a data-rich environment where an abundance of longitudinal clinical measurements that reflect the degeneration of the health condition can be continuously collected. For example, to monitor the surgical site infection (SSI), daily wound measurements, such as the temperature, granularity, and distance of the wound, could be acquired to assess the condition of the wound, together with other non-wound-related but important clinical signals such as heart rate, morning body temperature, NG tube presence, etc. However, particular data characteristics present challenges that call for specialized data fusion models to predict patient conditions using the multivariate longitudinal data. For instance, as these multivariate longitudinal data are actually temporal realizations of an underlying disease progression in different dimensions, how to leverage our knowledge of the disease progression process to fuse the data is a challenge. Also, the fact that these data are usually sampled at irregular time points adds in another layer of complexity. And even if we

*Correspondence: asamareh@uw.edu
Industrial and Systems Engineering, University of Washington, Seattle 98105, USA

could fuse the data properly, the existence of patient heterogeneity multiplies the complexity of the problem that calls for a generic framework to personalize the model based on individual's characteristics implicitly embedded in data.

To tackle those challenges, we propose a novel framework, named as DL-CHI, that focuses on a particular category of disease conditions that follow a monotonic disease progression process. In our previous work [7], we have developed a contemporaneous health index (CHI) that fuses the irregular multivariate longitudinal time series data to quantify the severity of degenerative disease conditions to fit the monotonic degradation process of the disease condition. However, CHI is designed for average user and ignores the patient heterogeneity and therefore limits their applicability in real-world applications. For example, it is known that patients of Alzheimer's disease (AD) suffer from very diverse and heterogeneous progression processes [8–10]. A possible remedy is to build personalized model on an individual's basis. However, this demands a great amount of labeled training samples, which are very likely not feasible in many clinical settings.

Thus, this motivates us to develop the DL-CHI framework by integrating CHI with dictionary learning [8, 11]. The basic idea shared by the dictionary learning algorithms is that the input signal is approximated with a sparse linear combination of a few dictionary elements or basis [12]. DL has been used in many signal processing applications, such as signal reconstruction [13], face recognition [14], and healthcare [15, 16]. The dictionary basis provides a succinct representation that can span the space of the personalized models to capture the patient heterogeneity and reveal the hidden structures in the data (in a similar spirit as principal component analysis). It has been shown that the performance of a classification task can be improved by learning a sparsifying dictionary from the data set. [17, 18]. The reason is that the sparsifying dictionary actually plays a role in the regularization of the model learning, as the dictionary basis vectors are numerical representations of patient heterogeneity. Translating this wisdom into DL-CHI, our basic idea is to first learn individual models through the CHI formulation and then reconstruct the model parameters of the learned individual models via supervised dictionary learning. Each column of the dictionary represents a basis vector. As such, each individual model is represented as a sparse linear combination of the basis vectors.

The paper is organized as follows. In Section 2, related work in the literature will be reviewed and discussed. In Section 3, the proposed analytic framework will be presented, and the corresponding computational algorithm will be derived. In Section 4, the proposed method

will be implemented and validated using two real-world applications; one is for monitoring of brain health in AD and the other is monitoring of SSI. We will conclude the study in Section 5. Note that, in this paper, we use lowercase letters, e.g., x , to represent scalars, boldface lowercase letters, e.g., \mathbf{v} , to represent vectors, and boldface uppercase letters, e.g., \mathbf{W} , to represent matrices.

2 Related works

2.1 The CHI model

The CHI model is developed in [7] which specifically utilizes the monotonicity of disease progression to enhance the data fusion of multivariate clinical measurements taken at irregular time points. In this section, we will first briefly present the basic formulation of the CHI model and then present the DL-CHI model that integrates CHI with dictionary learning for personalized models.

The CHI model was motivated by the common characteristics of many degenerative conditions such as AD which shows monotonic progression trajectory. For example, for AD, a number of biomarkers have been developed to measure the degeneration of the neural systems, including the neuroimaging modalities such as PET and MRI scans [19, 20]. It is typical to see that, along with the disease progression, the brain volumes shown in the MRI scans continue to shrink over time. The same phenomenon could be observed on the PET scans with the persistent decrease of metabolic activities. Those monotonic patterns indicate that the disease progression, once started, tends to be worse and worse.

The task of CHI is to translate multivariate longitudinal measurements into a contemporaneous health index $h_{n,t}$ that captures patient condition changing over the course of progression. Note that different individuals could be measured with different length of time and at different time locations. As we target degenerative conditions, CHI should be monotonic, i.e., $h_{n,t_1} \geq h_{n,t_2}$ if $t_1 \geq t_2$, if we assume that higher index represents more severe condition. Since CHI is a latent construct that is not directly measurable, clinical variables associated with it can be measured over time, which provide us data to learn it.

Denote the training set by $\mathbf{x}_{n,t} = [x_{n,1,t}, \dots, x_{n,d,t}]^T \in \mathbb{R}^d$ collected from N patients. Here, each measurement $x_{n,i,t}$ is the value of the i th variable for the n th subject for a given time t , where $t \in \{1, \dots, T_n\}$ is the time index. Converting the measurements $\mathbf{x}_{n,t}$ into $h_{n,t}$ needs a mathematical model for $h_{n,t} = f(\mathbf{x}_{n,t})$. Here, for simplicity and interpretability, we start with the linear models, i.e., $h_{n,t} = \mathbf{x}_{n,t} \cdot \mathbf{w}$, where $\mathbf{w} \in \mathbb{R}^d$ is a vector of weight coefficients to combine the d variables. Denote the total number of positive and negative samples by N^+ and N^- respectively, i.e., $N^+ := |\{n|y_n = 1\}|$ and $N^- := |\{n|y_n = -1\}|$.

The formulation of the CHI learning framework is shown in below:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad (1a)$$

$$+ \beta \sum_{n \in \{1, \dots, N\}} \max \left(0, 1 - y_n \left(\mathbf{x}_{n, T_n}^\top \mathbf{w} + b \right) \right) \quad (1b)$$

$$+ \alpha \sum_{\substack{n \in \{1, \dots, N\} \\ t \in \{1, \dots, T_n - 1\}}} \max \left(0, 1 - \mathbf{z}_{n, t}^\top \mathbf{w} \right) \quad (1c)$$

$$+ \frac{\lambda}{2} \left(\frac{1}{N^+} \sum_{n \in \{N^+ | y_n = 1\}} \left(\left(\mathbf{x}_{n, T_n} - \bar{\mathbf{x}}_{T_n}^+ \right)^\top \mathbf{w} \right)^2 \right) \quad (1d)$$

$$+ \frac{\lambda}{2} \left(\frac{1}{N^-} \sum_{n \in \{N^- | y_n = -1\}} \left(\left(\mathbf{x}_{n, T_n} - \bar{\mathbf{x}}_{T_n}^- \right)^\top \mathbf{w} \right)^2 \right) \quad (1e)$$

$$+ \gamma \|\mathbf{w}\|_1. \quad (1f)$$

Items in (1) can be explained as follows:

- The first term (1a) and second term (1b) are the SVM formulation that aims to utilize the label information to enhance the discriminatory power of CHI. Here, $y_n \in \{1, -1\}$ is the label of the n th sample that indicates if the n th subject is diseased or not.
- The term (1c) is invented to enforce the monotonicity of the learned health index, i.e., $h_{n, t_1} \geq h_{n, t_2}$ if $t_1 \geq t_2$. Here, $\mathbf{z}_{n, t}$ is the difference of two successive data vectors $\mathbf{z}_{n, t} := \mathbf{x}_{n, t+1} - \mathbf{x}_{n, t}$.
- Items (1d) and (1f) are invented to encourage the homogeneity of CHI within the group that has the same health status. Here, $\bar{\mathbf{x}}_{T_n}^+$ and $\bar{\mathbf{x}}_{T_n}^-$ represent the center of data vectors at time T_n for all positive and negative samples, respectively, that are:

$$\bar{\mathbf{x}}_{T_n}^+ := \frac{1}{N^+} \sum_{n \in \{n | y_n = 1\}} \mathbf{x}_{n, T_n}$$

$$\bar{\mathbf{x}}_{T_n}^- := \frac{1}{N^-} \sum_{n \in \{n | y_n = -1\}} \mathbf{x}_{n, T_n}.$$

- The last term, (1f), is the L_1 -norm penalty that is used to encourage sparsity of the features.

Note that the proposed formulation generalized many existing models, such as SVM, sparse SVM, LASSO, etc. The CHI model could be efficiently solved by using the block coordinate descent algorithm that is illustrated in [Appendix: "CHI model formulation"](#) section.

2.2 Dictionary learning

Developing models like CHI helps us to capture changes in various aspects of the disease trajectory. But as CHI assumes the same model for the whole population, it

ignores heterogeneity of degenerative diseases and therefore limits its applicability in real-world applications that have shown great patient heterogeneity [21, 22]. Recently, it has been shown that learning a dictionary can overcome the above limitations [14, 23, 24]. The basic idea of dictionary learning algorithms is to approximate training samples as a sparse linear combination of the few dictionary elements. Hence, dictionary learning algorithm can be considered as a way to represent low-dimensional structure of high-dimensional data.

DL was applied to many applications and achieved state-of-the-art performances, such as image denoising [13] and inpainting [25], clustering [26, 27], classification [28, 29] etc. It is known that the conventional DL framework was designed for a reconstruction task instead of adapting to classification. It is believed that classification performance will be further improved if we carefully learn a classification-oriented dictionary. For instance, in [12] a sparse representation-based classification (SRC) method was proposed for robust face recognition and achieved very impressive results. SRC treats the original data set as a dictionary, wherein the class-specific training sets are sub-dictionaries contributing to discrimination. Inspired by SRC, Yang et al. proposed a meta-face learning [14] to learn an adaptive dictionary for each class, and Ramirez et al. [17] added another term to derive more delicate classification-oriented dictionaries.

The use of dictionary learning for personalization of prediction models is also achieved by proposing novel transfer learning approaches. For example, in [6] personalization task was performed in two phases: learning user-specific source classifiers and learning a distribution-to-classifier mapping via implementing dictionary learning. Another approach is to perform multi-modal task-driven dictionary learning algorithm under the joint sparsity constraint to enforce collaborations among multiple homogeneous/heterogeneous sources of information. In task-driven formulation, the multi-modal dictionaries are learned simultaneously with their corresponding classifiers. The resulting multi-modal dictionaries can generate discriminative sparse codes from the data that are optimized for a given task such as binary or multi-class classification [30].

There are various dictionary learning algorithms that are effective for classification tasks [31–34]. Zhang and Li proposed discriminative K-SVD to simultaneously achieve a dictionary which has a good representation power while supporting optimal discrimination of the classes [33]. The name K-SVD refers to updating a dictionary with K vectors. A collection of training vectors corresponding to the dictionary vector in its approximation are taken by minimizing the Frobenius norm of the approximation error by solving for the dictionary vector at each iteration. This algorithm starts with an initial dictionary and initial

sparse code coefficients, and then, one dictionary vector is updated at each iteration. The corresponding sparse coefficient is changed before proceeding to update the next dictionary vector. The minimization is done through singular value decomposition (SVD). Another example is the iterative least squares dictionary learning algorithms (ILS-DLA) presented in [31, 32], where assumes known sparse code coefficients at each iteration and derives the best possible dictionary using either the orthogonal matching pursuit (OMP) or Focal Under-determined System Solver (FOCUSS). ILS-DLA method deploys a second order update which makes it nearly impractical in reasonable dimensions due to its matrix inversion step. Another example is the recursive least squares dictionary learning algorithm RLS-DLA, which is an online alternation of ILS-DLA. In the online alternation, each training signal is processed one at a time to improve the dictionary. One of the larger challenges with ILS-DLA and K-SVD is to find a good initial dictionary. The online nature of RLS-DLA prevents getting stuck in a local minimum close to the initial dictionary contrary to the K-SVD and ILS-DLA. RLS-DLA uses the forgetting factor to improve the convergence properties of the algorithm and hence makes the algorithm less dependent on the initial dictionary. However, RLS-DLA method requires to permute the order of training vectors and adapt the forgetting factor to satisfy the randomness and convergence properties of the online nature of the algorithm.

There are several properties that should be considered in the search for a successful dictionary training algorithm. *Flexibility*: The algorithm should be flexible enough to run with various sparse approximation algorithm such as pursuit algorithm which involves finding the best projections of input signal onto the span of an overcomplete dictionary \mathbf{D} . The flexibility property would enable different choices in favor of run-time constraints. Usually, methods that are flexible enough would separate the dictionary updates with sparse coding stage. *Adaptivity*: An overcomplete dictionary \mathbf{D} either can be chosen as a pre-determined set of functions or designed to iteratively getting updated to better fit the data. Choosing a pre-specified dictionary is appealing because it is simpler and may lead to a fast algorithm. However, the dictionary that leads to the best representation for each member in this set, under strict sparsity constraints, is needed. Such dictionaries have the potential to outperform the commonly used pre-specified dictionaries. *Efficiency*: A dictionary learning algorithm should lead to a numerically efficient and fast convergence. For example, ILS-DLA has a second-order update which makes it nearly impractical in reasonable dimensions.

K-SVD algorithm is flexible and works with any pursuit algorithm. In addition, it leads to the best representation for each training vector. Given the merits of DL

in overcoming heterogeneity of models, and the classification performance, here we used the idea of DL and developed the DL-CHI framework using K-SVD dictionary learning algorithm. Therefore, we reconstructed our model parameter of each individual sample to be linear combination of dictionary elements. We further compared our methodology with CHI and other dictionary models K-SVD, ILS-DLA, and RLS-DLA. Note, that DL-CHI formulation is personalized and not designed for average users unlike the above methods.

3 The proposed DL-CHI model

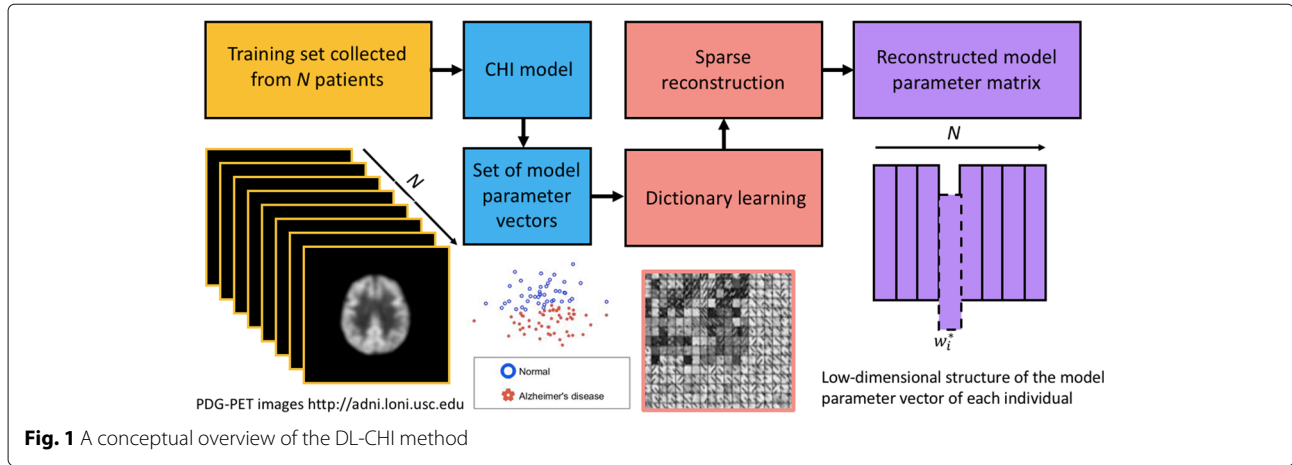
3.1 Rational and formulation

To extend CHI for personalized models, our approach is built on the dictionary learning framework [35]. As we have mentioned, the dictionary learning aims to identify a set of representative vectors that could characterize the low-dimensional structure embedded in a high-dimensional vector space [36–38]. Particularly, here, taking the model parameter vectors of all the individuals as the high-dimensional vector space, we seek a dictionary to represent these model parameter vectors. The dictionary will be learned from data, and it helps regularize the learning of the models since it requires the model parameter vectors to be (sparse) linear combination of the dictionary bases. The whole pipeline of this DL-CHI model is shown in Fig. 1.

From this point of view, the dictionary learning could be viewed as a trade-off made between two extremes. In one extreme, there is only one model for all the individuals, i.e., the “one-size-fits-all” model. On the other extreme, there is one distinct model for all the individuals and these models are all independent with each other. As a trade-off, dictionary learning exploits the dependency and difference of the individuals simultaneously.

To fulfill this idea, here, we denote the set of model parameter vectors of all the individuals as $\mathbf{W}^* = [\mathbf{w}_1^* \dots, \mathbf{w}_i^*, \dots, \mathbf{w}_N^*]$, where \mathbf{w}_i^* represents weight coefficient vector of the i^{th} patient learned from the CHI model. Using dictionary learning, we aim to find an overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{d \times k}$ that contains k independent columns referred as the basis vectors, $\{d_i\}_{i=1}^k$. A model parameter vector \mathbf{w}^* can be represented as a linear combination of these basis vectors, satisfying the approximation condition $\mathbf{w}^* \approx \mathbf{D}\mathbf{a}$, where \mathbf{a} is the coefficient vector which can be considered as the representation of \mathbf{w}^* over the dictionary \mathbf{D} .

In order for \mathbf{D} to be flexible and robust to noise, we set the dictionary to be overcomplete ($k > d$). On the other hand, given any \mathbf{w}^* with a overcomplete dictionary, we need to find the smallest set of basis vectors from the dictionary to represent it. When we set the dictionary to be overcomplete, an infinite number of solutions are available for the representation; hence, constraints on



the solution must be set. The solution with the fewest number of nonzero coefficients in \mathbf{a} to represent \mathbf{w}^* is certainly an appealing representation. This strategy is called sparse coding that is often used in dictionary learning representations. In this setting, sparse coding amounts to computing the following:

$$\begin{aligned} \{\mathbf{A}, \mathbf{D}\} &= \min_{\mathbf{D}, \mathbf{A}} \sum_i^N \|\mathbf{w}_i^* - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda_1 \|\mathbf{a}_i\|_0 \\ &= \min_{\mathbf{D}, \mathbf{A}} \|\mathbf{W}^* - \mathbf{D}\mathbf{A}\|_2^2 + \lambda_1 \|\mathbf{A}\|_0 \end{aligned} \quad (2)$$

Here, $\|\cdot\|_0$ is the ℓ^0 norm, counting the nonzero entries of a vector, and $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_N]$ is the coefficient matrix of the sparse decomposition. In order to achieve sparse representations given a set of training vectors, we adapt a dictionary that leads to the best representation for each vector in this training set, under strict sparsity constraints.

3.2 Computational algorithm

In DL-CHI, we used the K-SVD dictionary learning algorithm [39, 40] for sparse representation as an optimization problem, which can be efficiently solved via orthogonal matching pursuit (OMP) and singular value decomposition (SVD). The K-SVD approach is an iterative procedure that consists of two steps, and both steps in the algorithm are coherent with each other, working towards the minimization of the overall objective function.

First, we considered the sparse coding stage where we assumed that \mathbf{D} was fixed and considered the optimization problem in (2) as a search for sparse representation with coefficients summarized in the matrix \mathbf{A} . The sparsity term of the constraint was relaxed so that the number of nonzero entries of each column \mathbf{a}_i could be more than

1 and less than a number T_0 . In doing so, the relaxed objective function becomes:

$$\min_{\mathbf{a}_i} \|\mathbf{w}_i^* - \mathbf{D}\mathbf{a}_i\|_2^2 \quad \text{s.t.} \quad \forall i, \|\mathbf{a}_i\|_0 \leq T_0, i = 1, 2, \dots, N \quad (3)$$

In (3) \mathbf{D} was first fixed such that we could focus on learning the coefficient matrix \mathbf{A} using the orthogonal matching pursuit method, as long as it could supply a solution with a fixed and predetermined number of nonzero entries T_0 . OMP is an iterative greedy algorithm that selects the column best correlated with the residual part of the signal and represents the sub-optimal solution to the problem of sparse signal representation. The major advantage of the OMP is its simplicity and fast implementation. The problem in (3) consists of N distinct problems.

With a learned \mathbf{A} , we searched for the best dictionary \mathbf{D} . The search process is to update only one column of the dictionary, \mathbf{d}_k , at each time corresponding to i^{th} row in \mathbf{A} , denoted as \mathbf{a}_T^i (this is not the vector \mathbf{a}_i which is the i^{th} column in \mathbf{A}). The process of updating only one column of \mathbf{D} at a time has a straightforward solution based on the singular value decomposition (SVD). The problem becomes looking only at the training vectors that uses only one column of the dictionary vector in its approximation, minimizing the approximation error \mathbf{E}_k . The matrix $\mathbf{E}_k = \mathbf{W}^* - \sum_{j \neq k}^k \mathbf{d}_j \mathbf{a}_T^j$ stands for the error for all the training samples when the k th basis is removed, and \mathbf{a}_T^k is the k th row in \mathbf{A} . The SVD finds the closest rank-1 matrix (in Frobenius norm) that approximates \mathbf{E}_k . Hence, we re-wrote the penalty term in (3) as:

$$\sum_i^N \|\mathbf{w}_i^* - \mathbf{D}\mathbf{a}_i\|_2^2 = \|\mathbf{W}^* - \mathbf{D}\mathbf{A}\|_F^2 \quad (4)$$

The notation $\|A\|_F$ stands for the Frobenius norm, defined as $\|A\|_F = \sqrt{\sum_{ij} A_{ij}^2}$. Then, the penalty term in (2) can be rewritten as:

$$\begin{aligned} \|\mathbf{W}^* - \mathbf{D}\mathbf{A}\|_F^2 &= \left\| \mathbf{W}^* - \sum_{j=1}^k \mathbf{d}_j \mathbf{a}_T^j \right\|_F^2 \\ &= \left\| \left(\mathbf{W}^* - \sum_{j \neq k} \mathbf{d}_j \mathbf{a}_T^j \right) - \mathbf{d}_k \mathbf{a}_T^k \right\|_F^2 \quad (5) \\ &= \left\| \mathbf{E}_k - \mathbf{d}_k \mathbf{a}_T^k \right\|_F^2 \end{aligned}$$

Hence, we updated the $\left\| \mathbf{E}_k - \mathbf{d}_k \mathbf{a}_T^k \right\|_F^2$, assuming fixed coefficients \mathbf{A} and error \mathbf{E}_k . The constraint is over the j th orthonormal basis \mathbf{D}_j . By decomposing the multiplication $\mathbf{D}\mathbf{A}$ into the sum of K rank 1 matrices, we can assume that the other $K - 1$ terms were fixed, and the k th remains unknown. Then, the singular value decomposition finds the closest $K - 1$ terms that approximate \mathbf{E}_k , and this will effectively minimize the error in Eq. (5).

The above solution of vector \mathbf{a}_T^k is very likely to be filled, because the sparsity constraint is not enforced. To enforce the sparsity constraint, we define ω_k as the group of indices pointing to examples \mathbf{w}_i^* that use basis \mathbf{d}_k and entries of $\mathbf{a}_T^k(i)$ that are non-zero. Thus, $\omega_k = \{i | 1 \leq i \leq N, \mathbf{a}_T^k(i) \neq 0\}$. Then, we compute $\mathbf{E}_k = \left\| \mathbf{W}^* - \sum_{j \neq k} \mathbf{d}_j \mathbf{a}_T^j \right\|_F^2$ by only choosing columns corresponding to ω_k . We then apply the SVD decomposition $\mathbf{E}_k^R = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$. The solution for \mathbf{d}_k is the first column of \mathbf{U} , and the updated coefficient vector is the first column of $\mathbf{U} \times \mathbf{\Lambda}(1, 1)$.

3.3 Summary of DL-CHI

Putting all together, an overview of the DL-CHI method can be seen from Fig. 2. A full description of the DL-CHI algorithm is also given in Algorithm 1. It can be seen in Algorithm 1 that we have to learn \mathbf{W} , \mathbf{A} , and \mathbf{D} . We split the algorithm into two phases for learning personalized CHI and dictionary learning. In the phase I, we intend to solve \mathbf{w}^* via CHI using the Algorithm 2 described in Appendix: “Algorithm” section. In this phase, we learn the model parameter vectors of all individuals, which lead to the construction of the matrix \mathbf{W}^* . In the phase II, we use the K-SVD method to learn the dictionary by first computing the best representation matrix \mathbf{A} via (3) using the matching pursuit algorithm and then searching for the best dictionary. With a learned dictionary, the representations of the individual’s models could be identified and further used as the final individual models. Specifically, from the dictionary algorithm we can find the the

low-dimensional structure of the model parameter matrix $\mathbf{W}^* \approx \mathbf{D}\mathbf{A}$, where each column of \mathbf{W}^* is a reconstructed model parameter vector of each individual to be linear combination of dictionary elements.

Algorithm 1 The DL-CHI algorithm

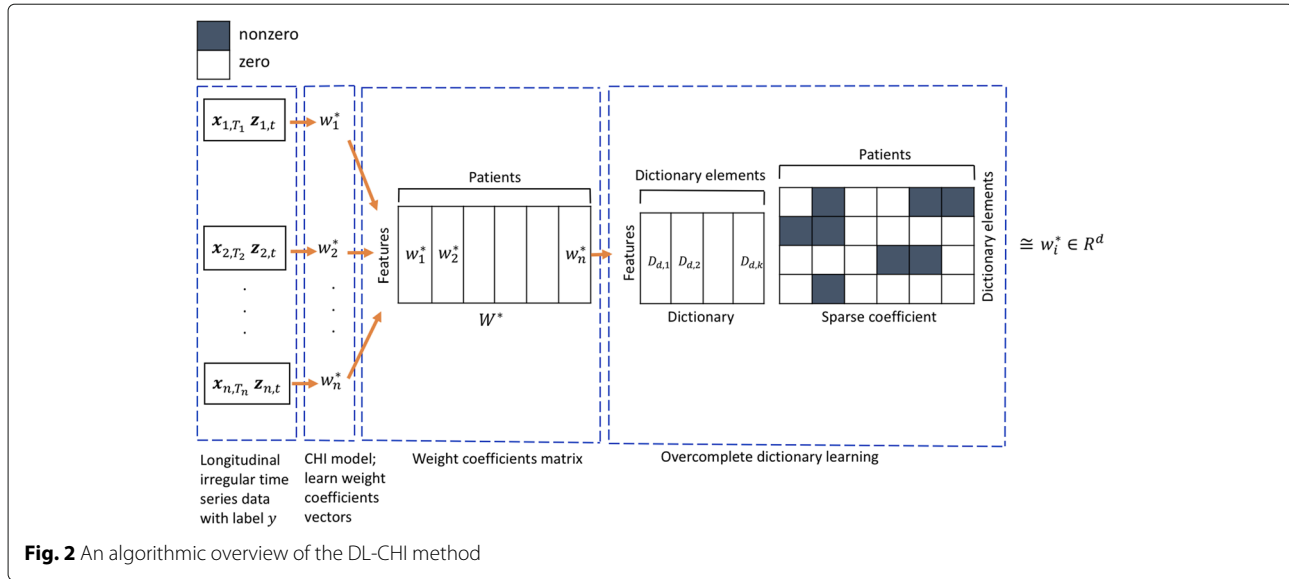
Require: $\mathbf{D}^{(0)} \in \mathbb{R}^{d \times k}$, $\mathbf{W}^* \in \mathbb{R}^{d \times n}$, and $\lambda \in \mathbb{R}$
Ensure: Find a dictionary $\mathbf{D} \in \mathbb{R}^{d \times k}$, and a corresponding coefficient matrix $\mathbf{A} \in \mathbb{R}^{k \times n}$ such that the representation error, $\mathbf{R} = \mathbf{W} - \mathbf{D}\mathbf{A}$ is minimized and \mathbf{A} fulfill sparseness criterion

- 1: **while** not converge **do**
- 2: **Start iterations** $t := 1, 2, \dots$ **do**
- 3: **Update** $\mathbf{A}^{(t)}$
- 4: **for** $i = 1, 2, \dots, N$,
- 5: $\mathbf{a}_{(t)} = \min_{\mathbf{a}_i} \left\| \mathbf{w}_i^* - \mathbf{D}^{(t-1)} \mathbf{a}_i \right\|_2^2 \forall i, \|\mathbf{a}_i\|_0 \leq T_0$
- 6: **end for**
- 7: **Update** $\mathbf{D}^{(t)}$
- 8: **for** $k = 1, \dots, k$ **Update the** k_{th} **column of** $\mathbf{D}^{(t)}$:
- 9: Define $\omega_k = \{i | 1 \leq i \leq d, \mathbf{a}_T^k(i) \neq 0\}$
- 10: Compute: $\mathbf{E}_k = \mathbf{W}^* - \sum_{t \neq k} \mathbf{d}_t \mathbf{a}_T^t$
- 11: In \mathbf{E}_k , choose only columns corresponding to ω_k
- 12: Apply the SVD to the corresponding \mathbf{E}_k
- 13: $\mathbf{E}_k = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$.
- 14: The updated \mathbf{d}_k is the first column of \mathbf{U} .
- 15: The updated \mathbf{a}_T^k is first column of $\mathbf{U} \times \mathbf{\Lambda}(1, 1)$.
- 16: **end for**
- 17: Reconstructed individual model parameter: $\mathbf{W}^* \approx \mathbf{D}\mathbf{A}$

4 Numerical studies

4.1 Real-world applications

We implement the DL-CHI model on two real-world datasets that were collected in Alzheimer’s disease (AD) and surgical site infection (SSI) research. Both diseases exhibit monotonic disease progression and significant patient heterogeneity. For the Alzheimer’s disease data, we use the FDG-PET images of 162 subjects (Alzheimer’s Disease: 74, Normal aging: 88) downloaded from the ADNI (www.loni.usc.edu/ADNI). For each subject, there are at least three time points and at most seven time points. The data has been preprocessed and the Automated Anatomical Labeling has been used to segment each image into 116 anatomical volumes of interest (AVOIs). We select 90 AVOIs that are in the cerebral cortex in our study. Each AVOI becomes a variable here. The measurement data of each region, according to the mechanism of FDG-PET, is the regional average FDG binding counts, representing the degree of glucose metabolism. Extensive evidences in the literature have shown that the glucose metabolism will decline as a function of the aging,



while the pathology of neurodegenerative diseases such as AD will further accelerate the declination, providing a perfect application example for implementing and testing the proposed DL-CHI method.

The SSI data exhibit similar characteristics as the AD data. There have been many models developed to monitor individuals who are subject to developing SSI [41–43], based on daily wound measurements, such as the temperature, granularity, and distance of the wound, together with other non-wound-related but important clinical signals such as heart rate, morning body temperature, and NG tube presence, etc. Figure 3 shows the longitudinal trend of a wound-related variable collected in our data, which clearly shows the monotonic degradation process of the SSI patients. The SSI data include longitudinal wound measurements from 857 patients, among which 169 are SSI patients and 539 are normal control. The data include wound measurement variables, for example, wound edge

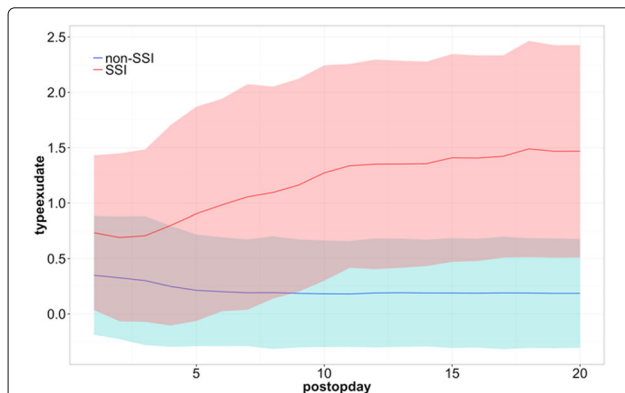


Fig. 3 Example of the longitudinal data of wound assessment that could gradually separate the SSI group with the non-SSI group as the condition progresses over time [7]

distance, temperature, include exudate amount, etc. Some other physiological variables such as heart rate are also provided in the data. Subjects were measured in time length ranging from 3 days to 21 days.

4.2 Parameter tuning and validation

For each experiment, we randomly split the data into two equal parts, one for training and one for testing. For training, we used 10-fold cross validation to tune the parameters. As CHI is a complex data fusion mechanism that synthesizes monotonicity of the disease progression, label information, and statistical homogeneity, we use a comprehensive scheme to compare DL-CHI with CHI. Specifically, we compared the two models (1) when only monotonicity is used for model training (i.e., by setting $\beta = 0$ and optimizing for α), (2) when only the label information is used for model training (i.e., by setting $\alpha = 0$ and optimizing for β), and (3) when a full model is used (i.e., by optimizing for both α and β). In addition, we performed in each of the settings by randomly down-sampling the training data, i.e., only using a proportion of the data ranging from 15 to 75%, to train both models. A model that can maintain good performances with less training data in obviously more promising in healthcare applications while data collection is relatively more costly than other real-world applications.

4.3 Results

Comparison between CHI and DL-CHI across a wide range of scenarios aforementioned are reported in Table 1. In general, it is observed that the DL-CHI model could significantly improve CHI model by accounting for the patient heterogeneity. This makes sense, since enforcing the constraint that the individual CHI model

Table 1 AUC performance for ADNI and SSI data across different ratio of training and testing datasets obtained by 10-fold cross-validation

| $\alpha = 0, \beta^*$ | | | |
|-----------------------|-----------|-------------------|-------------------|
| Data | Ratio (%) | CHI | DL-CHI |
| ADNI | 15 | 0.870 ± 0.024 | 0.887 ± 0.021 |
| | 20 | 0.883 ± 0.021 | 0.890 ± 0.016 |
| | 35 | 0.889 ± 0.014 | 0.936 ± 0.051 |
| | 50 | 0.890 ± 0.031 | 0.940 ± 0.047 |
| | 75 | 0.927 ± 0.012 | 0.959 ± 0.036 |
| SSI | 15 | 0.850 ± 0.055 | 0.867 ± 0.039 |
| | 20 | 0.861 ± 0.036 | 0.877 ± 0.020 |
| | 35 | 0.871 ± 0.012 | 0.886 ± 0.020 |
| | 50 | 0.862 ± 0.015 | 0.892 ± 0.041 |
| | 75 | 0.889 ± 0.024 | 0.914 ± 0.027 |
| $\alpha^*, \beta = 0$ | | | |
| ADNI | 15 | 0.780 ± 0.016 | 0.863 ± 0.034 |
| | 20 | 0.799 ± 0.054 | 0.873 ± 0.024 |
| | 35 | 0.804 ± 0.012 | 0.844 ± 0.034 |
| | 50 | 0.818 ± 0.019 | 0.869 ± 0.064 |
| | 75 | 0.855 ± 0.064 | 0.905 ± 0.024 |
| SSI | 15 | 0.829 ± 0.064 | 0.860 ± 0.023 |
| | 20 | 0.860 ± 0.021 | 0.879 ± 0.016 |
| | 35 | 0.870 ± 0.034 | 0.883 ± 0.034 |
| | 50 | 0.880 ± 0.042 | 0.892 ± 0.036 |
| | 75 | 0.883 ± 0.026 | 0.895 ± 0.016 |
| α^*, β^* | | | |
| ADNI | 15 | 0.865 ± 0.021 | 0.872 ± 0.025 |
| | 20 | 0.871 ± 0.023 | 0.881 ± 0.014 |
| | 35 | 0.874 ± 0.032 | 0.890 ± 0.026 |
| | 50 | 0.891 ± 0.021 | 0.910 ± 0.041 |
| | 75 | 0.901 ± 0.020 | 0.919 ± 0.036 |
| SSI | 15 | 0.741 ± 0.032 | 0.814 ± 0.041 |
| | 20 | 0.758 ± 0.034 | 0.820 ± 0.030 |
| | 35 | 0.770 ± 0.013 | 0.831 ± 0.036 |
| | 50 | 0.791 ± 0.026 | 0.887 ± 0.015 |
| | 75 | 0.806 ± 0.010 | 0.862 ± 0.036 |

should be represented by a dictionary plays a role in the regularization of the model learning, as the dictionary basis vectors are numerical representations of patient heterogeneity. It is shown that in all of the three scenarios, using only monotonicity ($\beta = 0$), using only the label information $\alpha = 0$, or the full-model DL-CHI model, achieve satisfying results. Another observation is that enforcing monotonicity constraint alone leads

to satisfactory performance for the DL-CHI model. As shown in Table 1, the DL-CHI method is also robust to small sample size. We investigate DL-CHI model's capability by selecting only 15% of the data as the training data, while the 10-folder cross validation was used to identify the optimal parameters in the model. The results show that our method achieves better prediction performance than the CHI model that uses the same ratio of the training data. Overall, the results show that the DL-CHI has a great potential for clinical applications to overcome the limitation of the CHI method in mitigating patient heterogeneity.

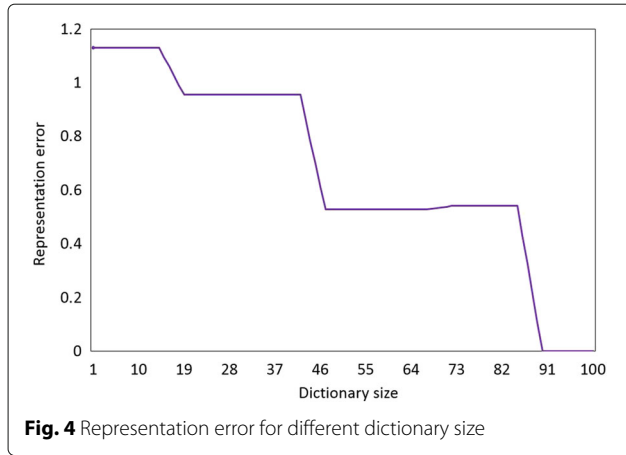
Table 2 shows the performance comparison of personalized DL-CHI method with the CHI model and three dictionary methods: K-SVD, ILS-DLA, and RLS-DLA. While for each model training, 10-fold cross validation is used on the training data and the AUC is evaluated on the testing data. Results in Table 2 show that the integration of dictionary learning with the CHI model improves the performance of the algorithm. The performance of RLS-DLA is in general considerable better than that of ILS-DLA and K-SVD. However, interestingly DL-CHI model performance demonstrates that it is superior to the RLS-DLA despite its convergence as an online algorithm and its ability for reconstruction purposes.

4.4 Representation capacity of dictionary learning

Figure 4 provides the results regarding the number of basis vectors needed for a sufficient representation of patient heterogeneity from AD. Apparently, the larger the dictionary size, the lower the representation error. On the other hand, we can also observe that the error of representation drops quickly with the increasing number of basis vectors in the dictionary. As the optimal dictionary size is not known in advance; hence, we first obtained it through an initial dictionary \mathbf{D}_0 of large size K . The initial dictionary $\mathbf{D}_0 \in \mathbb{R}^{d \times k}$ is obtained by selecting K samples randomly from input signals. The dictionary \mathbf{D}_0 helps minimizing the reconstruction error, and it is not yet optimal. For our experiment, we selected the number of basis based on the minimum error of representation given various dictionary sizes. To satisfy the overcompleteness, we choose the size

Table 2 AUC performance comparison for ADNI and SSI data for CHI, DL-CHI, K-SVD, ILS-DLA and RLS-DLA models obtained by 10-fold cross-validation

| Data | ADNI | SSI |
|---------|-------------------|-------------------|
| DL-CHI | 0.951 ± 0.025 | 0.902 ± 0.032 |
| CHI | 0.920 ± 0.021 | 0.880 ± 0.010 |
| RLS-DLA | 0.903 ± 0.030 | 0.873 ± 0.065 |
| K-SVD | 0.850 ± 0.043 | 0.803 ± 0.014 |
| ILS-DLA | 0.723 ± 0.012 | 0.653 ± 0.063 |



of \mathbf{D}_0 to be sufficiently larger than the dimension of an input signal.

5 Conclusion

In this paper, we presented a DL-CHI formulation to help build personalized contemporary health index (CHI) to monitor patient condition over time. Through applications on two real-world datasets of AD and SSI, the DL-CHI model is shown to be better than the CHI model in patient prediction and can achieve robust results with small sample sizes. In the future, we may further enhance the DL-CHI method in the following directions. First, note that, in the current DL-CHI formulation, the individual models have to be learned via the CHI formulation without information from the dictionary. Only with a learned dictionary, the representations of the individual's models are identified and further used as the final individual models. This is a possibility that a joint learning of both steps could further enhance the performance of DL-CHI by incorporating the dictionary into the CHI formulation. Second, the need of transfer learning when the supply of training data is limited is vital. One way to tackle this problem is by exploring the transfer learning through model-based transfer, where the prior knowledge from the generic recognizer enters through a modified regularization term in the CHI model. Last but not least, we can also consider an integration between data-based and model-based transfer learning. Where, by re-weighting the input source data, we can minimize the discrepancy between the source and the target distributions, and then allowing CHI to be biased toward the parameters of another model.

Appendix

CHI model formulation

For completeness of DL-CHI, here we present more details of the CHI formulation (1). The CHI formulation is convex but contains multiple non-smooth terms such as (1b), (1c), and (1f). To solve this formulation, we could

merge the smooth terms and derive the dual optimization problem, and finally train it via the block coordinate descent algorithm. Specifically, we can simplify Eq. (1) in a quadratic form by defining:

$$\begin{aligned} \|\mathbf{w}\|_Q^2 &:= \mathbf{w}^\top \mathbf{Q} \mathbf{w} = \|\mathbf{w}\|^2 \\ &+ \frac{\lambda}{2} \left(\frac{1}{N^+} \sum_{n \in \{N^+ | y_n = 1\}} \left((\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}_{T_n}^+)^T \mathbf{w} \right)^2 \right) + \\ &+ \frac{\lambda}{2} \left(\frac{1}{N^-} \sum_{n \in \{N^- | y_n = -1\}} \left((\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}_{T_n}^-)^T \mathbf{w} \right)^2 \right). \end{aligned} \quad (6)$$

where \mathbf{Q} is defined as

$$\begin{aligned} \mathbf{Q} &:= \mathbf{I} + \lambda \left(\frac{1}{N^+} \sum_{n \in \{n | y_n = 1\}} (\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}_{T_n}^+) (\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}_{T_n}^+)^T \right. \\ &\quad \left. + \frac{1}{N^-} \sum_{n \in \{n | y_n = -1\}} (\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}_{T_n}^-) (\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}_{T_n}^-)^T \right). \end{aligned}$$

With that, Eq. (6) is simplified to Eq. (7) as follows:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|_Q^2 + \gamma \|\mathbf{w}\|_1 \\ & + \alpha \sum_{\substack{n \in \{1, \dots, N\} \\ t \in \{1, \dots, T_n - 1\}}} \max(0, 1 - \mathbf{z}_{n,t}^\top \mathbf{w}) \\ & + \beta \sum_{n \in \{1, \dots, N\}} \max(0, 1 - y_n (\mathbf{x}_{n,T_n}^\top \mathbf{w} + b)). \end{aligned} \quad (7)$$

By introducing two relaxation variables ξ and ϵ , Eq. (7) is equivalent to Eq. (8) as follows:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|_Q^2 + \alpha \mathbf{1}^\top \xi + \beta \mathbf{1}^\top \epsilon + \gamma \|\mathbf{w}\|_1 \\ \text{s.t.} \quad & \mathbf{1} - \mathbf{Z}^\top \mathbf{w} - \xi \leq \mathbf{0} \\ & \mathbf{1} - \hat{\mathbf{X}}^\top \mathbf{w} - b \mathbf{y} - \epsilon \leq \mathbf{0} \end{aligned} \quad (8)$$

where

$$\begin{aligned} \xi &= (\xi_{1,1}, \dots, \xi_{1,T_1-1}, \dots, \xi_{N,1}, \dots, \xi_{N,T_N-1})^\top, \\ \mathbf{Z} &= (\mathbf{z}_{1,1}, \dots, \mathbf{z}_{1,T_1-1}, \dots, \mathbf{z}_{N,1}, \dots, \mathbf{z}_{N,T_N-1}), \\ \epsilon &= (\epsilon_1, \dots, \epsilon_N)^\top, \\ \mathbf{y} &= (\mathbf{y}_1, \dots, \mathbf{y}_N)^\top, \\ \hat{\mathbf{X}} &= (\mathbf{y}_1 \mathbf{X}_{1,T_1}, \dots, \mathbf{y}_N \mathbf{X}_{N,T_N}). \end{aligned}$$

We then can derive the dual formulation of (8) by substituting the ℓ_1 -norm penalty in (8) by its conjugate norm $\|\mathbf{w}\|_1 = \max_{\|\mathbf{s}\|_\infty \leq 1} \langle \mathbf{s}, \mathbf{w} \rangle = \max_{\|\mathbf{s}\|_\infty \leq 1} -\langle \mathbf{s}, \mathbf{w} \rangle$, and then introducing two new dual variables \mathbf{u} and \mathbf{v} which leads to the following formulation:

$$\min_{\substack{\mathbf{w}, \mathbf{b} \\ \epsilon \geq 0 \\ \xi \geq 0}} \max_{\substack{\mathbf{u} \geq 0 \\ \mathbf{v} \geq 0 \\ \|\mathbf{s}\|_\infty \leq \gamma}} \frac{1}{2} \|\mathbf{w}\|_Q^2 + \alpha \mathbf{1}^\top \xi + \beta \mathbf{1}^\top \epsilon - \langle \mathbf{w}, \mathbf{s} \rangle \\ + \langle \mathbf{u}, \mathbf{1} - \mathbf{Z}^\top \mathbf{w} - \xi \rangle + \langle \mathbf{v}, \mathbf{1} - \hat{\mathbf{X}}^\top \mathbf{w} - \mathbf{b}\mathbf{y} - \epsilon \rangle.$$

This can be rewritten as the following constrained smooth convex optimization problem, which can be solved efficiently:

$$\begin{aligned} \min_{\mathbf{s}, \mathbf{u}, \mathbf{v}} \quad & F(\mathbf{s}, \mathbf{u}, \mathbf{v}) := \frac{1}{2} \|\mathbf{s} + \mathbf{Z}\mathbf{u} + \hat{\mathbf{X}}\mathbf{v}\|_{Q^{-1}}^2 - \langle \mathbf{1}, \mathbf{u} \rangle - \langle \mathbf{1}, \mathbf{v} \rangle \\ \text{s.t.} \quad & \mathbf{0} \leq \mathbf{u} \leq \alpha \mathbf{1} \\ & \mathbf{0} \leq \mathbf{v} \leq \beta \mathbf{1} \\ & \langle \mathbf{v}, \mathbf{y} \rangle = 0 \\ & \|\mathbf{s}\|_\infty \leq \gamma. \end{aligned} \quad (9)$$

Then the solution \mathbf{w}^* to Eq. (9) can be obtained by: $\mathbf{w}^* = Q^{-1}(\mathbf{s}^* + \mathbf{Z}\mathbf{u}^* + \hat{\mathbf{X}}\mathbf{v}^*)$.

Algorithm

The block coordinate descent algorithm [44] to solve the dual problem in Eq. (9) is an iterative procedure as follows:

Algorithm 2 Block Coordinate Descent for Solving CHI

Require: Problem parameters $\{Z, Q, \hat{X}, \mathbf{y}, \alpha, \beta, \gamma\}$ and Optimization parameters $\eta_s, \eta_u, \eta_v, \rho \in (0, 1)$ (step sizes η_s, η_u and η_v can be adaptively decided using linear search alternatively)

Ensure: \mathbf{w}^*, b^*

- 1: Initialize $k = 0$
- 2: **While** not converge do
- 3:
- 4: $\mathbf{s}_{k+1} = \max(-\gamma \mathbf{1}, \min(\gamma \mathbf{1}, \mathbf{s}_k - \eta_s \nabla_s F(\mathbf{s}_k, \mathbf{u}_k, \mathbf{v}_k)))$
- 5: $\mathbf{u}_{k+1} = \max(\mathbf{0}, \min(\alpha \mathbf{1}, \mathbf{u}_k - \eta_u \nabla_u F(\mathbf{s}_{k+1}, \mathbf{u}_k, \mathbf{v}_k)))$
- 6: $\mathbf{v}_{k+1} = \text{Proj}_{\substack{\mathbf{0} \leq \mathbf{v} \leq \beta \mathbf{1} \\ \langle \mathbf{v}, \mathbf{y} \rangle = 0}}(\mathbf{v}_k - \eta_v \nabla_v F(\mathbf{s}_{k+1}, \mathbf{u}_{k+1}, \mathbf{v}_k))$
- 7: $k \leftarrow k + 1$
- 8: Recover the primal variables

$$\mathbf{w}^* = Q^{-1}(\mathbf{s} + \mathbf{Z}\mathbf{u} + \hat{\mathbf{X}}\mathbf{v})$$

$$b^* = \sum_{\{i | \mathbf{v}_i \in (0, \beta)\}} \mathbf{y}_i - \mathbf{w}^{*T} \mathbf{x}_i$$

Acknowledgements

The authors acknowledge funding support from the National Science Foundation under Grants CMMI-1536398 and CCF-1715027. Authors are also grateful for ADNI, Heather Evans, and Bill Lober for data to demonstrate our method.

Authors' contributions

SH and AS conceived the project. AS and SH completed the algorithm development, data analysis, and interpretation. Both authors contributed to the manuscript writing and approved the final manuscript.

Competing interests

Both authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 7 August 2017 Accepted: 13 February 2018

Published online: 05 March 2018

References

1. R Thomson, D Luettel, F Healey, S Scobie, Safer care for the acutely ill patient: learning from serious incidents. Natl. Patient Saf. Agency (2007)
2. RP Gaynes, DH Culver, TC Horan, JR Edwards, C Richards, JS Tolson, National Nosocomial Infections Surveillance System, Surgical site infection (SSI) rates in the United States, 1992–1998: the national nosocomial infections surveillance system basic SSI risk index. Clin. Infect. Dis. **33**(Supplement_2), S69–S77 (2001)
3. B Spring, M Gotsis, A Paiva, D Spruijt-Metz, Healthy apps: mobile devices for continuous monitoring and intervention. IEEE Pulse. **4**(6), 34–40 (2013)
4. DE Rivera, Optimized behavioral interventions: What does system identification and control engineering have to offer? IFAC Proc. Vol. **45**(16), 882–893 (2012)
5. S Deshpande, DE Rivera, JW Younger, NN Nandola, A control systems engineering approach for adaptive behavioral interventions: illustration with a fibromyalgia intervention. Transl. Behav. Med. **4**(3), 275–289 (2014)
6. G Zen, L Porzi, E Sangineto, E Ricci, N Sebe, Learning personalized models for facial expression analysis and gesture recognition. IEEE Trans. Multimedia. **18**(4), 775–788 (2016)
7. Y Huang, Q Meng, H Evans, W Lober, Y Cheng, X Qian, J Liu, S Huang, CHI: A contemporaneous health index for degenerative disease monitoring using longitudinal measurements. J. Biomed. Inform. **73**, 115–124 (2017)
8. JL Cummings, Cognitive and behavioral heterogeneity in Alzheimer's disease: seeking the neurobiological basis. Neurobiol. Aging. **21**(6), 845–861 (2000)
9. MF Folstein, Heterogeneity in Alzheimer's disease. Neurobiol. Aging. **10**(5), 434–435 (1989)
10. E Friedland, JV Koss, RP Haxby, CL Grady, J Luxenberg, J Schapiro, MB Kaye, Annals Intern. Med. **109**(4), 298–311 (1988)
11. BA Olshausen, DJ Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. **381**(6583), 607 (1996)
12. J Wright, A Yang, AY Ganesh, SS Sastry, Y Ma, Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 210–227 (2009)
13. M Elad, M Aharon, Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans. Image Process. **15**(12), 3736–3745 (2006)
14. M Yang, L Zhang, J Yang, D Zhang, in *Image Processing (ICIP), 2010 17th IEEE International Conference On*. Metaface learning for sparse representation based face recognition (IEEE, 2010), pp. 1601–1604
15. Q Xu, H Yu, X Mou, L Zhang, J Hsieh, G Wang, Low-dose X-ray CT reconstruction via dictionary learning. IEEE Trans. Med. Imaging. **31**(9), 1682–1697 (2012)
16. Y Chen, X Yin, L Shi, H Shu, L Luo, C Coatrieux, J-L Tournoulin, Phys. Med. Biol. **58**(16), 5803 (2013)
17. I Ramirez, P Sprechmann, G Sapiro, in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference On*. Classification and clustering via dictionary learning with structured incoherence and shared features (IEEE, 2010), pp. 3501–3508
18. R Raina, A Battle, H Lee, B Packer, AY Ng, in *Proceedings of the 24th International Conference on Machine Learning*. Self-taught learning: transfer learning from unlabeled data (ACM, 2007), pp. 759–766
19. SG Mueller, MW Weiner, LJ Thal, RC Petersen, C Jack, W Jagust, JQ Trojanowski, L Toga, W ABeckett, The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. **15**(4), 869–877 (2005)
20. JR Petrella, RE Coleman, PM Doraiswamy, Neuroimaging and early diagnosis of Alzheimer disease: a look to the future. Radiology. **226**(2), 315–336 (2003)
21. J Zhou, J Liu, J Narayan, VA Ye, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

- Modeling disease progression via fused sparse group lasso (ACM, 2012), pp. 1095–1103
22. J Zhou, J Liu, J Narayan, Ye VA, ADN Initiative, et al., Modeling disease progression via multi-task learning. *NeuroImage*. **78**, 233–248 (2013)
 23. J Mairal, M Elad, G Sapiro, Sparse representation for color image restoration. *IEEE Trans. Image Process.* **17**(1), 53–69 (2008)
 24. Z Jiang, Z Lin, LS Davis, Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(11), 2651–2664 (2013)
 25. M Elad, Y Figueiredo, MA Ma, On the role of sparse and redundant representations in image processing. *Proc. IEEE*. **98**(6), 972–982 (2010)
 26. B Cheng, J Yang, S Yan, Y Fu, TS Huang, Learning with l1-graph for image analysis. *IEEE Trans. Image Process.* **19**(4), 858–866 (2010)
 27. J Wright, Y Ma, J Mairal, G Sapiro, S Huang, TS Yan, Sparse representation for computer vision and pattern recognition. *Proc. IEEE*. **98**(6), 1031–1044 (2010)
 28. JA Bagnell, DM Bradley, in *Advances in Neural Information Processing Systems*. Differentiable sparse coding (Curran Associates, Inc., 2009), pp. 113–120
 29. J Mairal, J Ponce, G Sapiro, A Zisserman, FR Bach, in *Advances in Neural Information Processing Systems*. Supervised dictionary learning (Curran Associates, Inc., 2009), pp. 1033–1040. <http://papers.nips.cc/paper/3448-supervised-dictionary-learning.pdf>
 30. S Bahrampour, A Nasrabadi, NM Ray, WK Jenkins, Multimodal task-driven dictionary learning for image classification. *IEEE Trans. Image Process.* **25**(1), 24–38 (2016)
 31. K Engan, SO Aase, JH Husoy, in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference On*. Method of optimal directions for frame design, vol. 5 (IEEE, 1999), pp. 2443–2446
 32. K Engan, SO Aase, JH Husoy, Multi-frame compression: theory and design. *Signal Process.* **80**(10), 2121–2140 (2000)
 33. Q Zhang, B Li, in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference On*. Discriminative k-svd for dictionary learning in face recognition (IEEE, 2010), pp. 2691–2698
 34. K Engan, K Skretting, JH Husoy, Family of iterative ls-based dictionary learning algorithms, ils-dla, for sparse signal representation. *Digit. Signal Process.* **17**(1), 32–49 (2007)
 35. J Mairal, G Sapiro, M Elad, Learning multiscale sparse representations for image and video restoration. *Multiscale Model. Simul.* **7**(1), 214–241 (2008)
 36. K Kreutz-Delgado, JF Murray, BD Rao, K Engan, T-W Lee, TJ Sejnowski, Dictionary learning algorithms for sparse representation. *Neural Comput.* **15**(2), 349–396 (2003)
 37. M Donoho, DL Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proc. Natl. Acad. Sci.* **100**(5), 2197–2202 (2003)
 38. Z Mallat, SG Zhang, Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **41**(12), 3397–3415 (1993)
 39. Z Jiang, Z Lin, LS Davis, in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*. Learning a discriminative dictionary for sparse coding via label consistent k-svd (IEEE, 2011), pp. 1697–1704
 40. M Aharon, M Elad, A Bruckstein, *rmk*-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **54**(11), 4311–4322 (2006)
 41. A Dipiro, RG Martindale, JT Bakst, PF Vacani, P Watson, MT Miller, Infection in surgical patients: effects on mortality, hospitalization, and postdischarge care. *Am. J. Health-Syst. Pharmacy*. **55**(8), 777–781 (1998)
 42. E Lawson, BL Hall, CY Ko, Risk factors for superficial vs deep/organ-space surgical site infections: implications for quality improvement initiatives. *JAMA Surg.* **148**(9), 849–858 (2013)
 43. L Saunders, M Perennec-Olivier, P Jarno, F L'Héritau, A-G Venier, L Simon, M Giard, J-M Thiolet, J-F Viel, et al, Improving prediction of surgical site infection risk with multilevel modeling. *PLoS ONE*. **9**(5), e95295 (2014)
 44. P Tseng, S Yun, A coordinate gradient descent method for nonsmooth separable minimization. *Math. Prog.* **117**(1-2), 387–423 (2009)

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com