Baytas *et al. EURASIP Journal on Bioinformatics and Systems Biology* (2016) 2016:15
DOI 10.1186/s13637-016-0045-x

EURASIP Journal on Bioinformatics and Systems Biology

RESEARCH Open Access

Stochastic convex sparse principal component analysis

Inci M. Baytas[1], Kaixiang Lin[1], Fei Wang[2], Anil K. Jain[1] and Jiayu Zhou[1*]

Abstract

Principal component analysis (PCA) is a dimensionality reduction and data analysis tool commonly used in many areas. The main idea of PCA is to represent high-dimensional data with a few representative components that capture most of the variance present in the data. However, there is an obvious disadvantage of traditional PCA when it is applied to analyze data where interpretability is important. In applications, where the features have some physical meanings, we lose the ability to interpret the principal components extracted by conventional PCA because each principal component is a linear combination of all the original features. For this reason, sparse PCA has been proposed to improve the interpretability of traditional PCA by introducing sparsity to the loading vectors of principal components. The sparse PCA can be formulated as an ℓ_1 regularized optimization problem, which can be solved by proximal gradient methods. However, these methods do not scale well because computation of the exact gradient is generally required at each iteration. Stochastic gradient framework addresses this challenge by computing an expected gradient at each iteration. Nevertheless, stochastic approaches typically have low convergence rates due to the high variance. In this paper, we propose a convex sparse principal component analysis (Cvx-SPCA), which leverages a proximal variance reduced stochastic scheme to achieve a geometric convergence rate. We further show that the convergence analysis can be significantly simplified by using a weak condition which allows a broader class of objectives to be applied. The efficiency and effectiveness of the proposed method are demonstrated on a large-scale electronic medical record cohort.

Keywords: Sparse PCA, Convex PCA, Proximal mapping

1 Introduction

Principal component analysis (PCA) is a commonly used dimensionality reduction and data analysis tool in many areas such as computer vision [1, 2], data mining [3, 4], biomedical informatics [5, 6], and many others. The goal of PCA is to learn a linear transformation such that the learned principal components are the dimensions retaining the most of the variance in the data. Principal components are obtained by computing the eigenvalue decomposition of the covariance matrix, and it can also be computed by the singular value decomposition of the data matrix. Let $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ be the normalized covariance matrix for n training data points where each data point is in a d-dimensional feature space. The PCA of computing the top p components can be written as the following optimization problem:

$$\max_{\mathbf{Z}\in\mathbb{R}^{d\times p}} \|\mathbf{SZ}\|_F^2, \quad \text{s.t.}\mathbf{Z}^T\mathbf{Z} = \mathbf{I}, \tag{1}$$

where $\mathbf{Z}$ is an orthogonal projection matrix. In many applications, we are only interested in a few top principal components. In this case, the principal components can be computed in an iterative fashion: the leading principal component is calculated at each iteration (e.g., using power methods), and we then deflate the computed component and the next principal component now becomes the leading one [7]. Therefore, we focus on finding the leading principal component in this paper. In spite of its advantages, there is an obvious disadvantage of PCA. In the solution of Eq. (1), the principal components are linear combinations of all input variables. This means that the columns of $\mathbf{Z}$ matrix, which are called loadings of principal components, are dense. One important implication of dense loadings is that we lose the ability to interpret the

*Correspondence: jiayuz@msu.edu
[1]Computer Science and Engineering, Michigan State University, 48824 East Lansing, USA
Full list of author information is available at the end of the article

output dimensions of conventional PCA. PCA works well if we are not interested in the physical meanings of the features or if the interpretation of principal components is not crucial for the application. However, the intepretability is a significant factor when it comes to many applications such as biology, finance, and biomedical informatics. In the domain of biomedical informatics, as more and more electronic medical records (EMR) [8] of patients are available, medical researchers are interested in applying various techniques to analyze the EMR data. Each feature of the EMR data is a record/event related to a certain diagnosis. When the traditional PCA is applied to the data, those medical features are projected to a low dimensional space, in which each new feature will be the linear combination of all the original features. In this case, it is hard to comprehend the meaning of the new features.

Sparse PCA has been proposed to address this drawback. In sparse PCA, we learn sparse loading vectors which combine only few of the input variables allowing interpretation of the principal components. Sparse PCA was firstly proposed by Zou et al. in [9], where PCA was formulated as a regression problem and the sparse PCA was introduced by imposing the lasso (elastic net) constraint. Other common approaches to solve the sparse PCA problem are semi-definite programming [10, 11] and inverse power method [12]. Moreover, a more recent study [13] investigated sparse PCA with oracle property. Aforementioned approaches are generally not scalable enough to work with large-scale datasets. One way to deal with large sample sizes is using stochastic methods. We can see an example of stochastic PCA in [7]. Authors described an algorithm with computationally cheap stochastic iterations and variance reduction which was suggested in [14].

In this study, sparse PCA is posed as an ℓ_1 regularized optimization problem. Standard approaches to solve such sparse learning problems are proximal gradient methods [15–17], which require computation of the full gradient at each iteration. These methods generally work with a composite function including a smooth part and a non-smooth part. A large family of machine learning problems [18–23] can be expressed as composite functions. Traditionally, solving problems with objectives, which are not continuously differentiable, requires subgradient descent [24] which has very poor performance [25]. The recently developed proximal gradient methods can solve these composite problems with fast convergence rates [26, 27]. However, these methods are hardly scalable to large-scale problems with large sample sizes because of the computation of full gradient. Therefore, stochastic gradient-based methods are preferred in such problems. One major disadvantage of the stochastic gradient descent is the low convergence due to high variance by random sampling. Johnson and Zhang proposed a solution for this drawback in [14]. Their solution reduced the variance by using a copy of the estimated optimal point and the full gradient at this point in the gradient step. This approach exploited the strong convexity property to obtain a geometric convergence rate under expectation. Xiao and Zhang similarly presented a multi-stage scheme to progressively reduce the variance of the proximal stochastic gradient (Prox-SVRG) with a geometric convergence rate under expectation in [15]. The fundamental assumptions were Lipschitz continuity of the gradient of smooth part and the strong convexity of the objective function.

To tackle the aforementioned challenges in this paper, we introduce a novel stochastic convex sparse PCA (Cvx-SPCA) method which is extremely efficient and can handle large-scale datasets. Specifically, we propose to adopt a convex formulation of PCA [28] which provides a strongly convex function. The problem structure in this design allows us to leverage efficient scheme of Prox-SVRG [15] which leads to an exponential (geometric) convergence rate. We also investigate the convergence analysis of Prox-SVRG and present a new proof of the convergence rate which significantly reduces the conditions and assumptions required. As such, we show that the optimization scheme can be applied to a much larger class of problems to obtain the geometric convergence rate. We conducted extensive experiments on both synthetic and real datasets to illustrate the efficiency of the proposed algorithm. Because of its efficiency, we were able to apply the proposed algorithm to analyze a real EMR cohort with a large number of patients, which is hardly possible to analyze by using traditional approaches.

2 Convex sparse principal component analysis

In this section, we introduce the problem formulation and optimization scheme of the proposed approach. The problem of finding a sparse loading vector is posed as the combination of ℓ_1 sparsity inducing norm and convexity from the convex principal component analysis, which allows us to utilize an extremely efficient stochastic proximal gradient approach.

2.1 Convex sparse PCA

The goal of sparse PCA is to learn sparse loading vectors such that the principal components will be linear combinations of a few key variables instead of all the variables. We propose the following convex optimization problem:

$$\min_{\mathbf{z}\in\mathbb{R}^d} \{P(\mathbf{z}) = F(\mathbf{z}) + R(\mathbf{z})\}, \tag{2}$$

where the convex PCA loss [28] is given by:

$$F(\mathbf{z}) = \tfrac{1}{2}\mathbf{z}^T(\lambda\mathbf{I} - \mathbf{S})\mathbf{z} - \mathbf{w}^T\mathbf{z}$$

and the regularization term $R(\mathbf{z}) = \gamma\|\mathbf{z}\|_1$ is the ℓ_1 norm of the loading vector $\mathbf{z}$, $\gamma \in \mathbb{R}$ is the regularization

parameter controlling the sparsity of the loading vector, $\lambda > \lambda_1(\mathbf{S})$ is the convexity parameter, $\mathbf{w} \in \mathbb{R}^d$ is a random vector, and $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$. Here, $\lambda_1(\mathbf{S})$ represents the largest eigenvalue of the covariance matrix $\mathbf{S}$ and $\mathbf{w}$ is a vector of normally distributed random numbers. An upper bound for the regularization term γ can be derived by using standard subgradient analysis [25]: if the regularization parameter γ is larger than the maximum of absolute value of the elements of the vector $\mathbf{w}$, i.e., $\|\mathbf{w}\|_\infty$, we will end up with trivial solutions (solutions with only zeros). This thus guides us to use a parameter range of $\gamma \in [0, \|\mathbf{w}\|_\infty]$.

In the above approach, we use a convex optimization formulation of finding the first principal component inspired by the work in [28]. Even though $R(\mathbf{z})$ is not strongly convex, the overall cost function in Eq. (2) is a strongly convex function in which the strong convexity comes from $F(\mathbf{z})$. The structure of the problem defined in Eq. (2) allows us to use gradient based algorithms to obtain the global solution. Moreover, the strong convexity usually ensures nice convergence properties for stochastic gradient schemes as well. Therefore, we can also benefit from the faster convergence rate of the proximal stochastic scheme proposed in [15]. We note that the objective function of traditional PCA as shown in Eq. (1) does not define a convex problem, and thus, the analysis in this paper cannot be applied to it.

The most common methods to solve problems such as Eq. (2), where the objective function is comprised of the average of smooth component functions and a non-smooth function, are proximal gradient methods. In the next section, the method used to solve convex optimization problem given in Eq. (2) will be explained.

2.2 Optimization scheme

In this paper, we propose to use a proximal stochastic gradient method with progressive variance reduction approach [15] to solve the problem in Eq. (2). The function denoted by $F(\mathbf{z})$ can also be written as the sum of n smooth functions:

$$F(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\mathbf{z}^T\left(\lambda\mathbf{I} - \mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{z} - \mathbf{w}^T\mathbf{z}. \tag{3}$$

When n is very large, calculating the full gradient at each gradient descent iteration is an expensive operation. Hence, stochastic gradient methods are preferred to solve such problems. In stochastic approach, instead of calculating gradients for all of the data points, one data point is randomly sampled and the gradient at this point is calculated at each iteration. Therefore, the number of calculations decreases. However, the drawback of the stochastic gradient methods is the high variance introduced because of random sampling. As a result of the high variance, we suffer from poor convergence rates. As discussed previously, there are solutions to reduce the variance and increase the convergence rate. One of the studies which mitigates the high variance problem of stochastic gradient method is proximal stochastic gradient method with progressive variance reduction [15]. The study in [15] showed that the variance of the gradient can be upper bounded by using a multi-stage scheme which progressively reduces the variance. When the algorithm converges to optimal point, variance also converges to zero. Therefore, this approach can achieve better convergence rates than conventional stochastic gradient even with constant step sizes. We refer the readers to Section 3.1 in [15] for detailed proof of bounding the variance.

In this paper, we also follow the approach in [15]. The algorithm used in this study is given in Algorithm 1.

Algorithm 1 The proposed stochastic gradient descent with variance reduction algorithm for solving Cvx-SPCA

Input: $\lambda, [x_1, x_2, \ldots, x_n], \mathbf{S}, \mathbf{w}, \mathbf{z}_0, \eta, \gamma, m, T$
Output: $\mathbf{z}$

for $s = 1, 2, \ldots T$ **do**
  $\tilde{\mathbf{z}} = \tilde{\mathbf{z}}_{s-1}$
  $\tilde{\mathbf{v}} = (\lambda\mathbf{I} - \mathbf{S})\,\tilde{\mathbf{z}} - \mathbf{w}$
  $\mathbf{z}_0 = \tilde{\mathbf{z}}$
  for $k = 1, 2, \ldots, m$ **do**
    Pick $x_{ik} \in \{x_{1k}, \ldots, x_{nk}\}$ randomly
    $\mathbf{v}_k = \left(\lambda\mathbf{I} - x_{ik}x_{ik}^T\right)\left(\mathbf{z}_{k-1} - \tilde{\mathbf{z}}\right) + \tilde{\mathbf{v}}$
    $\mathbf{z}_k = \mathrm{prox}_{\eta\gamma}\left(\mathbf{z}_{k-1} - \eta\mathbf{v}_k\right)$
  end for
  $\tilde{\mathbf{z}}_s = \frac{1}{m}\sum_{k=1}^{m} \mathbf{z}_k$
end for
return $\tilde{\mathbf{z}}_T$

In the algorithm, $\mathbf{z}_0$ is the initial value for loading vector $\mathbf{z}$, η is the constant step size, γ is the regularization term to control sparsity of $\mathbf{z}$, m is the number of iterations for each epoch s, and T is the maximum number of epochs. At each epoch, full gradient at the point $\tilde{\mathbf{z}}$ is calculated periodically. The cost of calculating the full gradient is the product of a $d \times d$ matrix and a d dimensional vector. Therefore, the most time consuming part in our algorithm is the multiplications with covariance matrix, when the feature dimension is high. $\tilde{\mathbf{z}}$ is an estimate of the optimal point and it is updated at each epoch to be utilized in gradient calculations. During m stochastic gradient steps, we first sample a data point randomly and compute the gradient $\mathbf{v}_k$. If we take the expectation of the gradient calculated in Eq. (4), we can see that $\mathbf{v}_k$ is also an estimate of the full gradient as in conventional stochastic gradient methods.

This shows that $\mathbf{v}_k$ given below is in the same direction as the full gradient under expectation.

$$\begin{aligned}\mathbf{v}_k &= \nabla f_{ik}\left(\mathbf{z}_{k-1}\right) - \nabla f_{ik}\left(\tilde{\mathbf{z}}\right) + \nabla F\left(\tilde{\mathbf{z}}\right) \\ &= \left(\lambda \mathbf{I} - x_{ik}x_{ik}^T\right)\left(\mathbf{z}_{k-1} - \tilde{\mathbf{z}}\right) + \left(\lambda \mathbf{I} - \mathbf{S}\right)\tilde{\mathbf{z}} - \mathbf{w},\end{aligned} \tag{4}$$

where $\nabla F\left(\tilde{\mathbf{z}}\right)$ is the average gradient of functions $f_i\left(\mathbf{z}\right), i = 1, \ldots, n$ or the full gradient at point $\tilde{\mathbf{z}}$, $\nabla f_{ik}\left(\mathbf{z}_{k-1}\right)$ is the gradient of the function calculated by using the data point x_{ik} sampled at the kth iteration and $\tilde{\mathbf{z}}$ is the average of $\mathbf{z_k}$, $k = 1, .., m$ at the end of an epoch.

After the gradient computation, we update $\mathbf{z}_k$ by using the proximal mapping for ℓ_1 norm as follows.

$$\begin{aligned}\mathbf{z}_k &= \text{prox}_{\eta,\gamma}\left(\mathbf{z}_{k-1} - \eta \mathbf{v}_k\right) \\ &= \text{sign}\left(\mathbf{z}_{k-1} - \eta \mathbf{v}_k\right) \max\left(0, \left|\mathbf{z}_{k-1} - \eta \mathbf{v}_k\right| - \eta\gamma\right).\end{aligned}$$

In this algorithm, variance of the stochastic gradient $\mathbf{v}_k$ is reduced progressively, while both $\tilde{\mathbf{z}}$ and $\mathbf{z}_{k-1}$ are converging to the optimal point $\mathbf{z}_* = \arg\min_{\mathbf{z}} P\left(\mathbf{z}\right)$ [15]. Since the full gradient is utilized to modify stochastic gradients and function F is an average of smooth component functions, variance can be bounded. In the next section, we will give the convergence analysis of the aforementioned algorithm.

3 Convergence analysis

In this section, we present the convergence analysis of the proposed algorithm. The objective function used in this paper is suitable to follow the convergence analysis in [15]. Therefore, our analysis is mostly adapted from [15]. However, we use much weaker conditions which allow a broader family of objective functions to fit in this scheme and to enjoy the geometric convergence. We retain the following assumption used throughout in [15]:

Assumption 1 *The function $R\left(\mathbf{z}\right)$ is lower semi-continuous and convex, and its effective domain, $dom(R) := \left\{\mathbf{z} \in \mathbb{R}^d | R\left(\mathbf{z}\right) < +\infty\right\}$ is closed. Each $f_i\left(\mathbf{z}\right)$, for $i = 1, \ldots, n$, is differentiable on an open set that contains $dom(R)$, and their gradients are Lipschitz continuous. That is, there exist $L_i > 0$ such that for all $\mathbf{z}, \mathbf{y} \in dom(R)$,*

$$\left\|\nabla f_i\left(\mathbf{z}\right) - \nabla f_i\left(\mathbf{y}\right)\right\| \leq L_i \left\|\mathbf{z} - \mathbf{y}\right\|,$$

which also implies that the gradient of the average function $F\left(\mathbf{z}\right)$ is also Lipschitz continuous, i.e., there is an $L > 0$ such that for all $\mathbf{z}, \mathbf{y} \in dom(R)$,

$$\left\|\nabla F\left(\mathbf{z}\right) - \nabla F\left(\mathbf{y}\right)\right\| \leq L \left\|\mathbf{z} - \mathbf{y}\right\|,$$

where $L \leq (1/n)\sum_{i=1}^{n} L_i$.

In [15], convergence analysis was done for general F and R functions and both of them were assumed to be strongly convex. On the other hand, we only assume that functions $F\left(\mathbf{z}\right)$ and $R\left(\mathbf{z}\right)$ are convex, but not necessarily strongly convex. Thus, we are relaxing this strong assumption at this point. Strong convexity provides good properties and is relevant for faster convergence rates. However, objective functions are not always strongly convex in many cases. Therefore, a simplified version of the analysis will be preferable, when the objective functions do not have necessarily strong convexity property.

Although our overall objective function is strongly convex, $R\left(\mathbf{z}\right)$ is not strongly convex as it was mentioned in the previous section. Therefore, we drop the strong convexity assumption at two steps in the original analysis of [15] and obtain the convergence rate given in the following theorem.

Theorem 1 *Under the assumption that Assumption 1 holds and $0 < \eta < 1/\left(4L_Q\right)$, where $L_Q = max_i L_i$, the convergence rate is obtained as follows:*

$$\begin{aligned}&\rho = \frac{1}{\ell\left(1 - 4L_Q\eta\right)m\eta} + \frac{4L_Q\eta\left(m+1\right)}{\left(1 - 4L_Q\eta\right)m} < 1, \\ &\mathbb{E}\left\{P\left(\tilde{\mathbf{z}}_s\right)\right\} - P\left(\mathbf{z}_*\right) \leq \rho^s\left[P\left(\tilde{\mathbf{z}}_0\right) - P\left(\mathbf{z}_*\right)\right],\end{aligned} \tag{5}$$

where $\mathbf{z}_* = \arg\min_{\mathbf{z}} P\left(\mathbf{z}\right)$.

Proof The proof of Theorem 1 starts with investigating the distance between $\mathbf{z_k}$ and $\mathbf{z}_*$; $\|\mathbf{z_k} - \mathbf{z}_*\|^2$. According to the stochastic gradient mapping definition in [15], $\mathbf{z_k}$ can be written as $\mathbf{z_{k-1}} - \eta\mathbf{g_k}$.

$$\begin{aligned}\left\|\mathbf{z_k} - \mathbf{z}_*\right\|^2 &= \left\|\mathbf{z_{k-1}} - \eta\mathbf{g_k} - \mathbf{z}_*\right\|^2 \\ &= \left\|\mathbf{z_{k-1}} - \mathbf{z}_*\right\|^2 - 2\eta\mathbf{g_k}^T\left(\mathbf{z_{k-1}} - \mathbf{z}_*\right) \\ &\quad + \eta^2\left\|\mathbf{g_k}\right\|^2.\end{aligned} \tag{6}$$

The term $\left(-\mathbf{g_k}^T\left(\mathbf{z_{k-1}} - \mathbf{z}_*\right) + \frac{\eta}{2}\left\|\mathbf{g_k}\right\|^2\right)$ can be bounded by using the definition of the proximal update as shown below.

$$\begin{aligned}\mathbf{z_k} &= \text{prox}_{\eta R}\left(\mathbf{z_{k-1}} - \eta\mathbf{v_k}\right) \\ &= \arg\min_{y}\left\{\frac{1}{2}\left\|\mathbf{y} - \left(\mathbf{z_{k-1}} - \eta\mathbf{v_k}\right)\right\|^2 + \eta R\left(\mathbf{y}\right)\right\}\end{aligned}$$

According to the optimality condition,

$$\mathbf{z_k} - \left(\mathbf{z_{k-1}} - \eta\mathbf{v_k}\right) + \eta\xi = 0,$$

where $\xi \in \partial R\left(\mathbf{z_k}\right)$ is the subgradient of $R\left(\mathbf{z}\right)$ at $\mathbf{z_k}$. If we combine the stochastic gradient mapping definition with the optimality condition, we obtain the following expression.

$$\mathbf{z_k} - \left(\mathbf{z_k} + \eta\mathbf{g_k} - \eta\mathbf{v_k}\right) + \eta\xi = 0 \Rightarrow \xi = \mathbf{g_k} - \mathbf{v_k}$$

By using the convexity of $F(\mathbf{z})$ and $R(\mathbf{z})$, we can write the following inequality.

$$\begin{aligned} P(\mathbf{y}) &= F(\mathbf{y}) + R(\mathbf{y}) \\ &\geq F(\mathbf{z_{k-1}}) + \nabla F(\mathbf{z_{k-1}})^T (\mathbf{y} - \mathbf{z_{k-1}}) \\ &\quad + R(\mathbf{z_k}) + \xi^T (\mathbf{y} - \mathbf{z_k}) \end{aligned} \tag{7}$$

Convergence analysis of [15] utilized strong convexity of F and R in 7. However, we will show that strong convexity is not required at this point. Since $F(\mathbf{z})$ is assumed to be Lipschitz continuous with Lipschitz constant L, $F(\mathbf{z_{k-1}})$ can also be bounded by using Theorem 2.1.5 in [29].

$$\begin{aligned} F(\mathbf{z_{k-1}}) \geq F(\mathbf{z_k}) &- \nabla F(\mathbf{z_{k-1}})^T (\mathbf{z_k} - \mathbf{z_{k-1}}) \\ &- \frac{L}{2} \|\mathbf{z_k} - \mathbf{z_{k-1}}\|^2 \end{aligned} \tag{8}$$

If we combine Eqs. (7) and (8), we obtain the following inequality.

$$\begin{aligned} P(\mathbf{y}) &\geq F(\mathbf{z_k}) - \nabla F(\mathbf{z_{k-1}})^T (\mathbf{z_k} - \mathbf{z_{k-1}}) \\ &\quad - \frac{L}{2} \|\mathbf{z_k} - \mathbf{z_{k-1}}\|^2 + \nabla F(\mathbf{z_{k-1}})^T (\mathbf{y} - \mathbf{z_{k-1}}) \\ &\quad + R(\mathbf{z_k}) + \xi^T (\mathbf{y} - \mathbf{z_k}) \\ &\geq P(\mathbf{z_k}) - \nabla F(\mathbf{z_{k-1}})^T (\mathbf{z_k} - \mathbf{z_{k-1}}) \\ &\quad - \frac{L}{2} \|\mathbf{z_k} - \mathbf{z_{k-1}}\|^2 + \nabla F(\mathbf{z_{k-1}})^T (\mathbf{y} - \mathbf{z_{k-1}}) \\ &\quad + \xi^T (\mathbf{y} - \mathbf{z_k}) \end{aligned}$$

Here, we again use stochastic gradient mapping; $\mathbf{z_k} - \mathbf{z_{k-1}} = -\eta \mathbf{g_k}$ to obtain the following inequality.

$$\begin{aligned} P(\mathbf{y}) &\geq P(\mathbf{z_k}) + \nabla F(\mathbf{z_{k-1}})^T (\mathbf{y} - \mathbf{z_k}) \\ &\quad + \xi^T (\mathbf{y} - \mathbf{z_k}) - \frac{L}{2} \eta^2 \|\mathbf{g_k}\|^2 \end{aligned}$$

If we substitute ξ with $\mathbf{g_k} - \mathbf{v_k}$, then add and subtract $\mathbf{z_{k-1}}$ from the term $(\mathbf{y} - \mathbf{z_k})$:

$$\begin{aligned} P(\mathbf{y}) &\geq P(\mathbf{z_k}) + (\mathbf{v_k} - \nabla F(\mathbf{z_{k-1}}))^T (\mathbf{z_k} - \mathbf{y}) \\ &\quad + \mathbf{g_k}^T (\mathbf{y} + \mathbf{z_{k-1}} - \mathbf{z_{k-1}} - \mathbf{z_k}) - \frac{L}{2} \eta^2 \|\mathbf{g_k}\|^2 \end{aligned}$$

$$\begin{aligned} P(\mathbf{y}) &\geq P(\mathbf{z_k}) + \mathbf{g_k}^T (\mathbf{y} - \mathbf{z_{k-1}}) + \left(\eta - \frac{L}{2}\eta^2\right) \|\mathbf{g_k}\|^2 \\ &\quad + (\mathbf{v_k} - \nabla F(\mathbf{z_{k-1}}))^T (\mathbf{z_k} - \mathbf{y}) \end{aligned}$$

Under the assumption of $0 < \eta < 1/4L_Q < 1/L$, $\left(\eta - \frac{L}{2}\eta^2\right) = \frac{\eta}{2}(2 - L\eta)$ can be taken as $\eta/2$. Because $(2 - L\eta)$ is between $(1, 2)$ according to the assumption, therefore, eliminating $(2 - L\eta)$ does not change the inequality. Now we will use the result derived above for the term $\left(-\mathbf{g_k}^T (\mathbf{z_{k-1}} - \mathbf{z_*}) + \frac{\eta}{2} \|\mathbf{g_k}\|^2\right)$ in Eq. (6).

$$\begin{aligned} \|\mathbf{z_k} - \mathbf{z_*}\|^2 \leq \|\mathbf{z_{k-1}} - \mathbf{z_*}\|^2 &+ 2\eta (P(\mathbf{z_*}) - P(\mathbf{z_k})) \\ &- 2\eta \Delta^T (\mathbf{z_k} - \mathbf{z_*}), \end{aligned} \tag{9}$$

where $\Delta = \mathbf{v_k} - \nabla F(\mathbf{z_{k-1}})$ and $\mathbf{z_*}$ corresponds to $\mathbf{y}$. The term $-2\eta \Delta^T (\mathbf{z_k} - \mathbf{z_*})$ can further be bounded by using the proximal full gradient update $\bar{\mathbf{z}_k} = \text{prox}_{\eta R}(\mathbf{z_{k-1}} - \eta \nabla F(\mathbf{z_{k-1}}))$, If Cauchy-Schwarz inequality and the non-expansiveness of the proximal mapping $(\|\text{prox}_{\eta R}(x) - \text{prox}_{\eta R}(y)\| \leq \|x - y\|)$ are utilized, the following expression can be derived.

$$\begin{aligned} -2\eta \Delta^T (\mathbf{z_k} - \mathbf{z_*}) &= -2\eta \Delta^T (\mathbf{z_k} - \mathbf{z_*} + \bar{\mathbf{z}_k} - \bar{\mathbf{z}_k}) \\ &\leq 2\eta \|\Delta\| \|\mathbf{z_k} - \bar{\mathbf{z}_k}\| \\ &\quad - 2\eta \Delta^T (\bar{\mathbf{z}_k} - \mathbf{z_*}) \end{aligned}$$

If we insert the definitions of $\mathbf{z_k} = (\mathbf{z_{k-1}} - \eta \mathbf{v_k})$ and $\bar{\mathbf{z}_k} = (\mathbf{z_{k-1}} - \eta \nabla F(\mathbf{z_{k-1}}))$, we will have:

$$-2\eta \Delta^T (\mathbf{z_k} - \mathbf{z_*}) \leq 2\eta^2 \|\Delta\|^2 - 2\eta \Delta^T (\bar{\mathbf{z}_k} - \mathbf{z_*}).$$

If we combine the result shown above with Eq. (9):

$$\begin{aligned} \|\mathbf{z_k} - \mathbf{z_*}\|^2 \leq \|\mathbf{z_{k-1}} - \mathbf{z_*}\|^2 &- 2\eta (P(\mathbf{z_k}) - P(\mathbf{z_*})) \\ &+ 2\eta^2 \|\Delta\|^2 - 2\eta \Delta^T (\bar{\mathbf{z}_k} - \mathbf{z_*}). \end{aligned}$$

Now, expectations of both sides are taken with respect to $\mathbf{z_k}$.

$$\begin{aligned} \mathbb{E}\{\|\mathbf{z_k} - \mathbf{z_*}\|\} \leq \|\mathbf{z_{k-1}} - \mathbf{z_*}\|^2 &+ 2\eta^2 \mathbb{E}\{\|\Delta\|^2\} \\ &- 2\eta (\mathbb{E}\{P(\mathbf{z_k})\} - P(\mathbf{z_*})) \\ &- 2\eta \mathbb{E}\left\{\Delta^T (\bar{\mathbf{z}_k} - \mathbf{z_*})\right\} \end{aligned}$$

Since $\bar{\mathbf{z}_k}$ and $\mathbf{z_*}$ are independent from the variable $\mathbf{z_k}$; $\mathbb{E}\{\Delta^T (\bar{\mathbf{z}_k} - \mathbf{z_*})\} = \mathbb{E}\{\Delta^T\} (\bar{\mathbf{z}_k} - \mathbf{z_*}) = 0$. Because $\mathbb{E}\{\Delta^T\} = \mathbb{E}\{\mathbf{v_k} - \nabla F(\mathbf{z_{k-1}})\} = \mathbb{E}\{\mathbf{v_k}\} - \nabla F(\mathbf{v_{k-1}}) = 0$. The variance of the gradient $\mathbb{E}\{\|\Delta\|^2\}$ is upper bounded in Prox-SVRG algorithm and we will use the result of Corollary 3 in [15] which is $\mathbb{E}\{\|\Delta\|^2\} \leq 4L_Q [P(\mathbf{z_{k-1}}) - P(\mathbf{z_*}) + P(\tilde{\mathbf{z}}) - P(\mathbf{z_*})]$, where $L_Q = \max_i L_i$, $\tilde{\mathbf{z}}_s = \frac{1}{m} \sum_{k=1}^{m} \mathbf{z_k}$, and $\tilde{\mathbf{z}} = \tilde{\mathbf{z}}_{s-1} = \mathbf{z_0}$ for a fixed epoch. After incorporating the bound of the variance of the gradient into the analysis, the following expression is obtained.

$$\begin{aligned} \mathbb{E}\{\|\mathbf{z_k} - \mathbf{z_*}\|^2\} \leq \|\mathbf{z_{k-1}} - \mathbf{z_*}\|^2 &\\ - 2\eta (\mathbb{E}\{P(\mathbf{z_k})\} - P(\mathbf{z_*})) &\\ + 8\eta^2 L_Q [P(\mathbf{z_{k-1}}) - P(\mathbf{z_*})] &\\ + 8\eta^2 L_Q [P(\tilde{\mathbf{z}}) - P(\mathbf{z_*})] & \end{aligned}$$

Now, if we apply the inequality above repeatedly for $k = 1, \ldots, m$ and the expectation with respect to previous

random variables $\mathbf{z_1}, \ldots, \mathbf{z_m}$ are taken, then we can obtain the following inequality.

$$\begin{aligned}
&\mathbb{E}\left\{\|\mathbf{z_m} - \mathbf{z}_*\|^2\right\} + 2\eta\left[\mathbb{E}\left\{P\left(\mathbf{z_m}\right)\right\} - P\left(\mathbf{z}_*\right)\right] \\
&\quad + 2\eta\left(1 - 4\eta L_Q\right)\sum_{k=1}^{m-1}\left[\mathbb{E}\left\{P\left(\mathbf{z_k}\right)\right\} - P\left(\mathbf{z}_*\right)\right] \\
&\leq \|\mathbf{z_0} - \mathbf{z}_*\|^2 \\
&\quad + 8\eta^2 L_Q\left[P\left(\mathbf{z_0}\right) - P\left(\mathbf{z}_*\right) + m\left(P\left(\tilde{\mathbf{z}}\right) - P\left(\mathbf{z}_*\right)\right)\right]
\end{aligned}$$

Since $2\eta\left(1 - 4\eta L_Q\right) < 2\eta$, $\mathbf{z_0} = \tilde{\mathbf{z}}$ and P is convex, therefore, $P\left(\tilde{\mathbf{z}}_s\right) \leq \frac{1}{m}\sum_{k=1}^{m} P\left(\mathbf{z_k}\right)$, and we can write the following inequality.

$$\begin{aligned}
&2\eta\left(1 - 4\eta L_Q\right) m\left[\mathbb{E}\left\{P\left(\tilde{\mathbf{z}}_s\right)\right\} - P\left(\mathbf{z}_*\right)\right] \\
&\leq \|\tilde{\mathbf{z}}_{s-1} - \mathbf{z}_*\|^2 \\
&\quad + 8\eta^2 L_Q\left(m+1\right)\left(P\left(\tilde{\mathbf{z}}_{s-1}\right) - P\left(\mathbf{z}_*\right)\right)
\end{aligned}$$

By using Lemma 1 which is a weaker condition then using the strong convexity and by applying the above inequality recursively, we derive the convergence rate as follows:

$$\begin{aligned}
&\left[\mathbb{E}\left\{P\left(\tilde{\mathbf{z}}_s\right) - P\left(\mathbf{z}_*\right)\right\}\right] \\
&\leq \left(\frac{\left(\frac{2}{\ell} + 8\eta^2 L_Q\left(m+1\right)\right)}{2\eta\left(1 - 4\eta L_Q\right) m}\right)^s \left[P\left(\tilde{\mathbf{z}}_0\right) - P\left(\mathbf{z}_*\right)\right].
\end{aligned}$$

□

Lemma 1 *Consider the problem of minimizing the sum of two convex functions:*

$$\min_{\mathbf{z}\in\mathbb{R}^d}\left\{P\left(\mathbf{z}\right) = F\left(\mathbf{z}\right) + R\left(\mathbf{z}\right)\right\}.$$

A standard method for solving the above problem is the proximal gradient method. Given an initial point $\mathbf{z_0}$*, using the proximal mapping, which is shown below, iteratively generates a sequence that will converge to the optimal solution.*

$$prox_R\left(\mathbf{y}\right) = \arg\min_{\mathbf{z}\in\mathbb{R}^d}\left\{\frac{1}{2}\left\|\mathbf{z} - \mathbf{y}\right\|^2 + R(\mathbf{z})\right\}$$

Since $R\left(\mathbf{x}\right)$ *is a convex function, the optimal solution of above problem is also an optimal solution of the following problem using a tuning parameter* μ *[30] [Theorem 1].*

$$\min\frac{1}{2}\left\|\mathbf{z} - \mathbf{y}\right\|_2^2 \; s.t.\; R\left(\mathbf{z}\right) \leq \mu$$

By utilizing the optimal strong convexity condition which is a weaker condition than strong convexity [31] for a convex function R, we have the following inequality for all $\mathbf{z} \in \Omega$*:*

$$P\left(\mathbf{z}\right) - P\left(prox_E\left(\mathbf{z}\right)\right) \geq \frac{\ell}{2}\left\|\mathbf{z} - prox_E\left(\mathbf{z}\right)\right\|^2$$

where the $prox_E$ *is the Euclidean projection on to set E and* ℓ *is a positive parameter.*

We have thus removed the strong convexity condition so that we are able to apply the algorithm in [15] to more generic convex objectives.

4 Results

In this section, we present the results of two types of experiments. First, the proposed algorithm was tested on synthetic datasets to investigate the convergence of the variance reduced proximal stochastic gradient compared to traditional proximal stochastic gradient descent. In addition, running times of the proposed stochastic Cvx-SPCA and other sparse PCA methods were compared to emphasize the advantage of using a stochastic approach, when there are large number of samples. In our experiments, step size η was chosen by the following heuristic according to $0 < \eta < 1/\left(4L_Q\right)$ and L_Q was taken as the largest eigenvalue of the covariance matrix. Iteration number m was chosen as $\Theta\left(L_Q/\left(\lambda - \lambda_1\left(\mathbf{S}\right)\right)\right)$ which is suggested in [15]. Secondly, we presented our experiments on an electronic medical records data.

4.1 Synthetic dataset

In this section, we present some results of the proposed stochastic Cvx-SPCA algorithm on synthetic datasets. Synthetic datasets used in this section were all randomly generated by normally distributed random numbers with $\mathcal{N}\left(0,1\right)$. For this purpose, synthetic data with varying sample sizes were prepared by random sampling. First of all, we would like to compare the convergence of proximal stochastic gradient with variance reduction and traditional proximal stochastic gradient for our algorithm. In Fig. 1, objective versus number of epochs are plotted for using traditional proximal stochastic gradient (prox-SGD) and proximal stochastic variance reduced gradient (Prox-SVRG) methods.

In Fig. 1, convergence is observed when the maximum number of epochs is fixed to 50. We also would like to investigate how many epochs are necessary for both algorithms to converge. Therefore, we made another experiment to see how fast Cvx-SPCA with Prox-SVRG converges to a similar sparsity as Cvx-SPCA with prox-SGD. We generated another synthetic dataset with 100,000 instances and 10,000 dimensions. The result of the experiment is shown in Fig. 2. Cvx-SPCA with traditional SGD took 3646.94 s and Cvx-SPCA with SVRG took 644.60 s to converge to similar sparsity patterns.

Secondly, running times of other sparse PCA methods and the proposed method were compared in Table 1. In experiments, feature dimension was chosen as 1000. Algorithms ran until they reached similar sparsity patterns. The proposed Cvx-SPCA algorithm is more scalable, since only one gradient is computed at a time and there are no eigenvalue decomposition or SVD steps during iterations. For instance, [9] requires singular value decomposition at

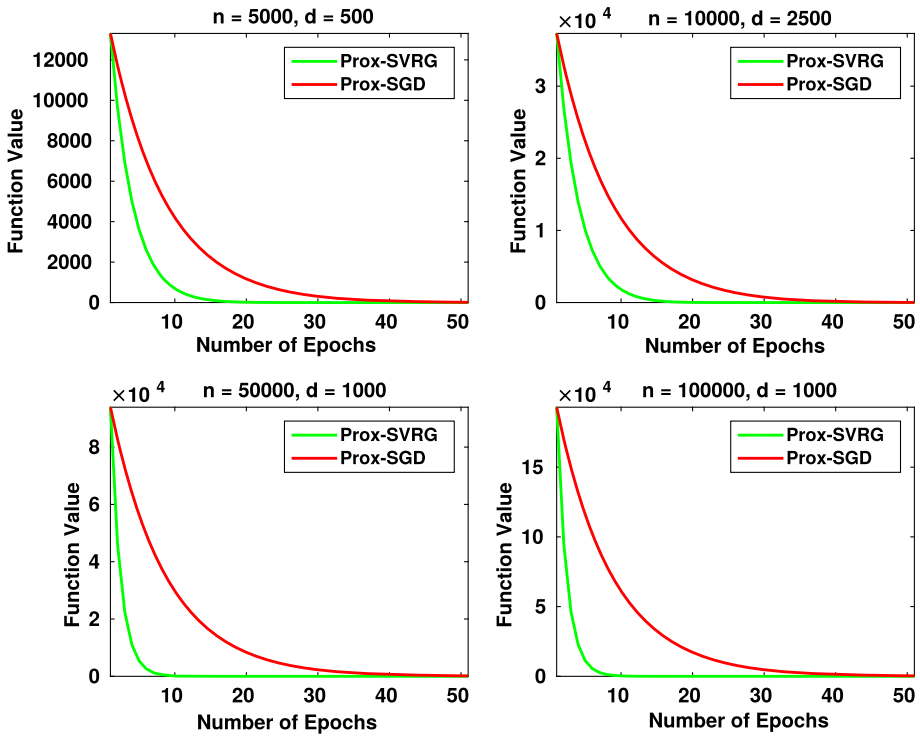

Fig. 1 Convergence for synthetic data. Convergence of the proposed stochastic Cvx-SPCA with (Prox-SVRG) and without variance reduction (prox-SGD). Proximal stochastic gradient with variance reduction has a faster convergence rate, since the variance caused by random sampling is bounded in Prox-SVRG

each iteration, which is a bottleneck in terms of running time, [12] is an inverse power method based approach, and [11] uses semi-definite programming. Therefore, scalability with respect to sample size and dimension is an issue for the aforementioned methods.

We also investigate the regularization path for the proposed algorithm. Regularization path illustrates how the solution changes for different values of regularization parameters γ which specify the level of sparsity. In order to have a suitable level of sparsity, γ should be tuned. One common way of finding an appropriate γ is the regularization path. We first generated a random sample with ten features and applied the proposed Cvx-SPCA algorithm to obtain the principal component. Then, the covariance matrix was reconstructed by using the first principal component corresponding to the largest eigenvalue with a little random noise. Loading values of principal components were computed with varying regularization parameters γ by using the reconstructed covariance matrix. We started with small γ values, and the loading vector learned from the previous step is used as the initialization for each new Cvx-SPCA step. The result is given in Fig. 3.

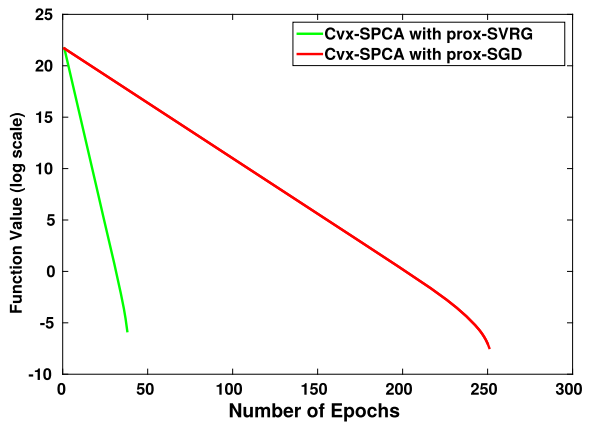

Fig. 2 Convergence of sparse pattern in the log scale. Cvx-SPCA with Prox-SGD takes 275 iterations, whereas Cvx-SPCA with Prox-SVRG takes 45 iterations to converge a similar sparsity pattern

Table 1 Running times (in seconds) of different SPCA algorithms

Sample size	Cvx-SPCA	[9]	[12]	[11]
$n = 50$ k	20.9	207.1	48.7	3002
$n = 100$ k	26.2	466.9	78.3	3237.4
$n = 500$ k	35.6	2737.06	2661.7	5276.93
$n = 1$ m	35.8	3408.59	3568	5274.26

Since proposed Cvx-SPCA does not depend on eigenvalue decomposition or semi-definite programming, it is more scalable in terms of the sample size. It also requires less iterations to reach a desired sparsity

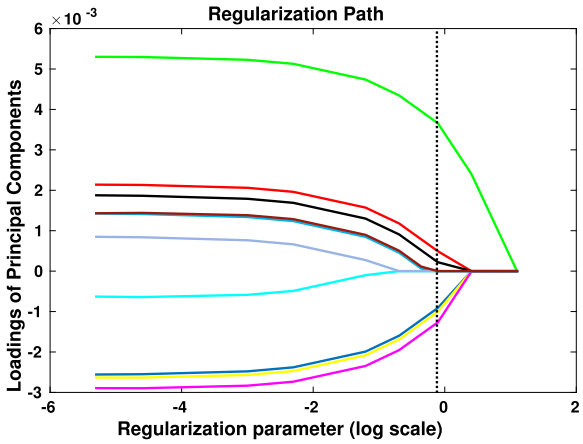

Fig. 3 Regularization path for Cvx-SPCA. We checked whether the known principal component can be recovered through the path to be able confirm that this is a valid regularization path. When regularization term was around −0.11 (*dashed line*) in logarithmic scale, we could exactly recover the non-zero loading values of the known principal component which was used to generate the data

4.2 Large-scale healthcare dataset

We applied our Cvx-SPCA algorithm to analyze disease patterns in a general patient population. The dataset we used is a real world electronic medical record (EMR) warehouse including the records of 223,076 patients over 4 years. We used the diagnosis information (in terms of ICD9 codes [32]) in our investigation, which resulted in 11,982 features in total. In this dataset, we do not have demographic information of patients explicitly. However, we investigated patient groups with different gender and age by looking at the descriptions of the ICD9 codes. We draw histograms of the number of patients with respect to the number of diagnoses each patient has in different demographic groups and in the general population as in Figs. 4 and 5, from which we can observe that the majority of the patients just have very few records. In our experiments, we eliminated the patients who have less than five records, and this resulted in 177,856 patients. As it was mentioned earlier, some of the diseases are specifically related to gender and age that let us have an idea about the demographic information of the dataset. For instance, complications of pregnancy, female genital disorders, and abortion are some of the diagnoses which are explicitly about women. Similarly, maternal complications affecting newborn and diseases such as chickenpox and measles are related to children. There are also ICD9 codes which have terms indicating the age. For instance, some of the diagnoses have the term "senile" which points out patients at least above 60 years old. Thus, we sampled female, male, old, and child patients by taking the definitions of the ICD9 codes into account. The age range of child patients can be given as from babyhood to adolescence and age of old patients can be thought as above 60 years old. In Table 2, number of patients and number

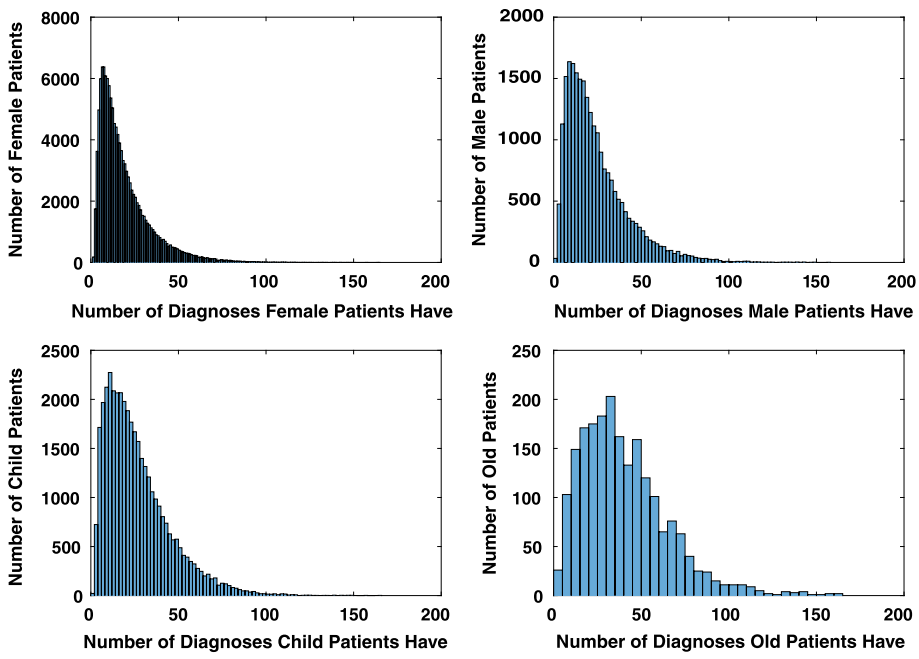

Fig. 4 Patient distribution of demographic groups. We used only diagnoses/diseases which have explicit information about demographic of the patient while sub-sampling the patients. We can observe that each group of patient has a similar trend. Most of the patients have 1–50 diagnoses entered into the record

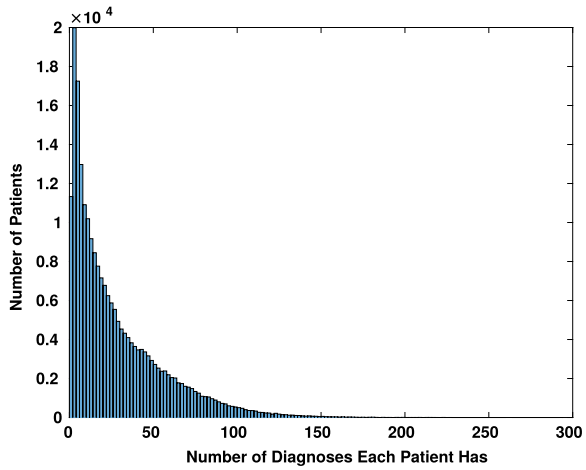

Fig. 5 Patient distribution. We observe that the majority of the patients just have very few records

of features related to female, male, people above 60 years old and children groups are given. We should note that there may be female, male, old, and child patients who we did not include into these demographic groups. For example, there should be female/male patients with diagnoses which are not gender- or age-specific. It is not always possible to guess the gender or age from diagnosis such as hypertension or infectious diseases which can be encountered in both genders. Therefore, we are reporting the demographic groups whose ICD9 codes have clear terms indicating the demographic information.

As can be seen from Table 2, the number of female-specific diseases and the number of female patients are more than other groups in the EMR dataset we used in this paper. Number of old patients is given less than other groups in the table. However, it may not mean that there are less number of old people in the whole patient population. We could not exactly extract age information of every diagnoses/diseases. For instance, hypertension or Alzheimer's were diseases commonly encountered among the people above a certain age in the past. However, these problems can be occurred in younger ages recently. For this reason, we used only diagnoses/diseases which have explicit information about demographic of the patient, while sub-sampling the patients. Distributions of different patient groups in Table 2 are given in Fig. 4.

Table 2 We sample patients who have female, male, child, and old people related features. These samples may overlap with each other. For instance, a patient may have dementia and a prostate problem together. We did not include other problems such as hypertension or kidney problems which can be encountered in every age and both genders into these groups of patients

Demographic	Number of features	Number of patients
Female	1268	130,035
Male	106	24,184
Old	66	2060
Child	596	38,434

In our experiments, we further aggregated all diagnoses belong to the same ICD9 group together, so that each patient is represented by a 918 dimensional feature vector. The value on its ith dimension represents the frequency of the ith diagnosis code appearing in the EMR of the corresponding patient. Since every patient will have a limited number of diseases, patient vectors are very sparse.

We would like to emphasize that existing sparse PCA algorithms cannot be used to analyze a dataset at this scale. We carried out both quantitative and qualitative evaluations on this dataset. We studied the convergence of the algorithm with varying number of patients, and we observe that the proposed Cvx-SPCA can still achieve a good convergence even when the sample size is very large, as shown in Fig. 6.

Next, we conducted an experiment to show how the proposed algorithm helps us to analyze the EMR data. We applied our algorithm to the whole data set and got the output features which correspond to the non-zero loading values of the leading principal component. These output features are inferred as key medical features. One of the results is summarized in Table 3. Diseases shown in this table are the features which have non-zero loadings whose absolute values are greater than a heuristic threshold. In our experiments, we observed that the most frequently encountered output features were infectious diseases, problems related to pregnancy and labor, injuries, and cancer types. This result tells us that the proposed algorithm can provide insight about the diagnoses encountered in the patient population.

We further examined the data set and divided the features into groups in terms of gender and age. We sampled the patients who have gender- and age-related problems separately and applied our algorithm to those samples

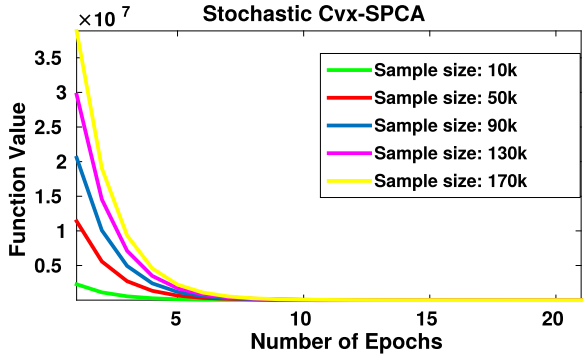

Fig. 6 Convergence for 20 epochs of Cvx-SPCA for different number of patients

Table 3 EMR data features which contributes the output dimensions after Cvx-SPCA algorithm was applied to the whole patient population. Most frequently observed problems are infections, injuries, pregnancy, and delivery related problems and cancer types

ICD9 code	Description
7	Balantidiasis/infectious
72	Mumps orchitis/infectious
115	Infection by histoplasma capsulatum
266	Ariboflavinosis/metabolic disorder
507	Pneumonitis/bacterial
695	Toxic erythema/dermatological
697	Lichen planus/dermatological
761	Incompetent cervix affecting fetus or newborn
795	Abnormal glandular papanicolaou smear of cervix
924	Contusion of thigh/injury

to analyze the output dimensions. Examples from each group are shown in Tables 4, 5, 6, and 7. We can see plausible results for the output features of each group in the tables. For example, diagnoses such as female genital disorders, perinatal problems, and anemia, which are more common among women, appeared in Table 4 where the algorithm was applied to the subset of patients who have female-related problems. Similarly, we can see from Table 5 that a subset of male patients generates prostate cancer along with other diagnoses which can be frequently seen in the general patient population as well. Cancer is a commonly encountered problem in nearly every age. We can come across cancer in the results of children and old patients as well. Another observation is that tuberculosis and bacterial infections are quite common among children.

5 Discussion

Throughout the paper, advantage of using a convex optimization approach for sparse PCA is emphasized. In this section, we would like to discuss about our conjuring of the convergence of non-convex stochastic sparse PCA by using the same framework. One surprising finding we have is if we use this non-convex PCA to construct a non-convex sparse PCA (by adding ℓ_1-norm), we still benefit from a much faster convergence rate using the stochastic scheme studied in this paper. A similar result is also presented in [7], where the authors propose a stochastic PCA approach with an exponential convergence rate by using variance reduced stochastic gradient presented in [14]. These results lead us to ask the following question: *Can we generalize the convergence analysis of proximal variance reduced stochastic gradient method further for non-convex settings?* We will investigate this problem in the future work.

Table 4 Output EMR data features which contributes the output dimensions after applying the proposed algorithm to the subset of patients who have female-related problems. We could observe female-specific problems and other common diseases such as heart problems and anemia

ICD9 code	Description
281	Pernicious anemia
392	Valvular and rheumatic heart disease
614	Female genital disorders
778	Serious perinatal problem affecting newborn
905	Major head injury

Table 5 Output EMR data features which contributes the output dimensions after applying the proposed algorithm to the subset of patients who have male related problems. We could observe a prostate problem which is directly related male patients. In addition, we can also see other common problems such as injuries

ICD9 code	Description
185	Malignant neoplasm of prostate
298	Depressive type psychosis
719	Effusion of joint
800	Closed fracture of vault of skull
811	Closed fracture of scapula
860	Traumatic pneumothorax

6 Conclusions

In this paper, a convex stochastic sparse PCA method is proposed. Since the problem of finding the leading eigenvector is formed as convex optimization, a well-defined convergence rate can be applied to the proposed algorithm. A proximal stochastic gradient method with variance reduction is preferred to avoid low convergence rates of traditional stochastic methods. Although strong convexity is usually required in literature, we simplify the convergence analysis of the existing Prox-SVRG algorithm

Table 6 Output EMR data features which contributes the output dimensions after applying the proposed algorithm to the subset of patients who have old age-related problems. Cancer is a commonly encountered problem in nearly every ages. In addition to this, we could observe disorders of nervous system and visual problems in the results

ICD9 code	Description
153	Malignant neoplasm of colon
173	Other malignant neoplasm of skin
337	Disorders of the autonomic nervous system
368	Visual disturbance

Table 7 Output EMR data features which contributes the output dimensions after applying the proposed algorithm to the subset of patients who have child related problems. According to our observation, tuberculosis and bacterial infections are quite common among children. Unfortunately, leukemia is also a cancer type that is seen even in small kids

ICD9 code	Description
8	Intestinal infection due to other organisms
11	Pulmonary tuberculosis
78	Other diseases due to viruses and Chlamydiae
10	Primary tuberculous infection
204	Lymphoid leukemia

by using weaker conditions. According to the experiments on several synthetic data, the proposed algorithm is shown to be more scalable due to stochastic approach. In addition, an application of sparse PCA is presented to show how sparse PCA can help to interpret electronic medical records. In future work, we would like to investigate whether sparse PCA can be used to cluster patients with respect to their medical records. For instance, we propose to apply the proposed algorithm to analyze medical records and derive clinically meaningful and structural phenotypes, which can further be helpful for patient risk stratification and clustering.

Acknowledgements
This work is supported in part by the Office of Naval Research (ONR) under grant number N00014-14-1-0631 and National Science Foundation under grant numbers IIS-1565596 and IIS-1615597.

Authors' contributions
IMB and JZ developed the algorithm. KL contributed to dropping the strong convexity section. FW provided the EMR data and contributed to the interpretation of the experimental results. IMB wrote the paper, and JZ and AKJ edited the paper. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Computer Science and Engineering, Michigan State University, 48824 East Lansing, USA. [2]School of Software Engineering, Beijing University of Technology, Beijing, China.

Received: 14 April 2016 Accepted: 4 August 2016
Published online: 09 September 2016

References

1. FD la Torre, MJ Black, in *ICCV Eighth IEEE International Conference on Computer Vision, vol. 1*. Robust principal component analysis for computer vision (IEEE, Vancouver, 2001)
2. MW Manal Abdullah, S Bo-saeed, Optimizing face recognition using pca. Int. J. Artif. Intell. Appl. (IJAIA). **3**(2), 23–31 (2012)
3. C Gokulnath, MK Priyan, E Vishnu Balan, KP Rama Prabha, in *International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*. Preservation of privacy in data mining by using pca based perturbation technique (IEEE, Chennai, 2015), pp. 202–206
4. W-Y Wang, C-X Qu, in *Second International Symposium on Information Science and Engineering*. Application and research of data mining based on improved pca method (IEEE, Shanghai, 2009), pp. 140–143
5. PP Alberto Landi, G Pioggia, in *Intelligent Systems Design and Applications, 9th International Conference on*. Backpropagation-based non linear pca for biomedical applications (IEEE, Pisa, 2009), pp. 635–640
6. D Omucheni, K Kaduki, W Bulimo, H Angeyo, Application of principal component analysis to multispectral-multimodal optical image analysis for malaria diagnostics. Malar. J. **13**(1), 485 (2014). Springer Nature
7. O Shamir, in *32nd International Conference on Machine Learning, vol. 37*. A stochastic pca and svd algorithm with an exponential convergence rate (Journal of Machine Learning Research (JMLR), Lille Grand Palais, 2015)
8. What Is an Electronic Medical Record (EMR)?. https://www.healthit.gov/providers-professionals/electronic-medical-records-emr
9. TH Hui Zou, R Tibshirani, Sparse principal component analysis. J. Comput. Graph. Stat. **15**(2), 265–286 (2006)
10. A d'Aspremont, L El Ghaoui, M Jordan, G Lanckriet, A direct formulation for sparse pca using semidefinite programming. SIAM Rev. **49**(3), 434–448 (2007)
11. AY Nikhil Naikal, SS Sastry, Informative feature selection for object recognition via sparse pca. Int. Conf. Comput. Vision, IEEE, 818–825 (2011)
12. M Hein, T Buhler, An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca. Adv. Neural Inf. Process. Syst. **23**, 847–855 (2010)
13. Z Gu, Q Wang, H Liu, Sparse pca with oracle property. Adv. Neural Inf. Process. Syst. (NIPS). **27**, 1529–1537 (2014)
14. R Johnson, T Zhang, Accelerating stochastic gradient descent using predictive variance reduction. Adv. Neural Inf. Process. Syst. **26**, 315–323 (2013)
15. L Xiao, T Zhang, A proximal stochastic gradient method with progressive variance reduction. SIAM J. OPTIM. **24**(4), 2057–2075 (2014)
16. A Nitanda, Stochastic proximal gradient descent with acceleration techniques. Neural Inf. Process. Syst. **27**, 1574–1582 (2014)
17. S Shalev-Shwartz, T Zhang, in *31 st International Conference on Machine Learning, JMLR, vol. 32*. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization (Journal of Machine Learning Research (JMLR), Beijing, 2014)
18. J Liu, J Chen, J Ye, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Large-scale sparse logistic regression (ACM, Paris, 2009), pp. 547–556
19. R Tibshirani, Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol, 267–288 (1996)
20. S Ji, J Ye, in *Proceedings of the 26th Annual International Conference on Machine Learning*. An accelerated gradient method for trace norm minimization (ACM, Montreal, 2009), pp. 457–464
21. J Zhou, L Yuan, J Liu, J Ye, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. A multi-task learning formulation for predicting disease progression (ACM, San Francisco, 2011), pp. 814–822
22. L Jacob, G Obozinski, J-P Vert, in *Proceedings of the 26th Annual International Conference on Machine Learning*. Group lasso with overlap and graph lasso (ACM, Montreal, 2009), pp. 433–440
23. J Zhou, J Chen, J Ye, *Malsar: Multi-task learning via structural regularization*. (Arizona State University, 2011), http://www.public.asu.edu/~jye02/Software/MALSAR
24. NZ Shor, *Minimization Methods for Non-differentiable Functions, vol. 3*. (Springer, Berlin Heidelberg, 2012)
25. S Boyd, L Xiao, A Mutapcic, Subgradient methods. lecture notes of EE392o, Stanford University, Autumn Quarter. **2004**, 2004–2005 (2003)
26. A Beck, M Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
27. SJ Wright, RD Nowak, MA Figueiredo, Sparse reconstruction by separable approximation. Signal Process. IEEE Trans. **57**(7), 2479–2493 (2009)
28. D Garber, E Hazan, Fast and simple pca via convex optimization (2015). arXiv:1509.05647v4 [math.OC], https://arxiv.org/abs/1509.05647
29. Y Nesterov, *Introductory Lectures On Convex Optimization: A Basic Course*, vol. 87. (Springer US, New York, 2004)
30. M Kloft, U Brefeld, P Laskov, K-R Müller, A Zien, S Sonnenburg, Efficient and accurate lp-norm multiple kernel learning. Adv. Neural Inf. Process. Syst, 997–1005 (2009)
31. J Liu, SJ Wright, Asynchronous stochastic coordinate descent: parallelism and convergence properties. SIAM J. Optim. **25**(1), 351–376 (2015)
32. International Classification of Diseases (ICD). http://www.who.int/classifications/icd/en/