sensors

MDPI

Article

Kernel Sparse Representation with Hybrid Regularization for On-Road Traffic Sensor Data Imputation

Xiaobo Chen [1,2,*], Cheng Chen [2], Yingfeng Cai [1], Hai Wang [2] and Qiaolin Ye [3]

[1] Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China; yfcai@ujs.edu.cn
[2] School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China; 13851301126@126.com (C.C.); 1000004061@ujs.edu.cn (H.W.)
[3] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China; yqlcom@njfu.edu.cn
* Correspondence: 1000003032@ujs.edu.cn; Tel.: +86-187-5296-5624

Received: 1 August 2018; Accepted: 29 August 2018; Published: 31 August 2018

Abstract: The problem of missing values (MVs) in traffic sensor data analysis is universal in current intelligent transportation systems because of various reasons, such as sensor malfunction, transmission failure, etc. Accurate imputation of MVs is the foundation of subsequent data analysis tasks since most analysis algorithms need complete data as input. In this work, a novel MVs imputation approach termed as kernel sparse representation with elastic net regularization (KSR-EN) is developed for reconstructing MVs to facilitate analysis with traffic sensor data. The idea is to represent each sample as a linear combination of other samples due to inherent spatiotemporal correlation, as well as periodicity of daily traffic flow. To discover few yet correlated samples and make full use of the valuable information, a combination of l_1-norm and l_2-norm is employed to penalize the combination coefficients. Moreover, the linear representation among samples is extended to nonlinear representation by mapping input data space into high-dimensional feature space, which further enhances the recovery performance of our proposed approach. An efficient iterative algorithm is developed for solving KSR-EN model. The proposed method is verified on both an artificially simulated dataset and a public road network traffic sensor data. The results demonstrate the effectiveness of the proposed approach in terms of MVs imputation.

Keywords: sparse representation; elastic net; kernel method; missing data; imputation

1. Introduction

With the rapid development of finance and society, traffic congestion has become an urgent worldwide problem causing the waste of travel time, environmental pollution, etc. To alleviate traffic congestion and improve road capacity, intelligent transportation system (ITS) [1] was developed by integrating various techniques, such as computer, communication, artificial intelligence, and so on. Data acquisition through different sensors is the foundation of ITS, where higher-level functions including traffic forecasting, route planning, etc., usually depend on the quality and quantity of sensor data. Unfortunately, traffic data acquired by sensors are usually incomplete, indicating the loss of many entries, due to various unpredictable factors, including sensor malfunction, transmission network anomaly, storage equipment damage, etc. For example, for a dense road network in Melbourne city, about 8% of its sensors can reach up to 56% missing data. Likewise, about 10% of daily traffic volume in Beijing is missing [2], and according to Turner et al. [3], more than 5% of data within the PeMS database are lost. The loss of traffic sensor data is unavoidable under existing technical level and objective conditions, therefore posing great challenge for effective deployment of ITS. For instance,

most of the traffic state forecasting algorithms [4], either statistical approaches such as support vector regression (SVR) [5–7] and neural networks (NNs) [8–10], or model based approaches [11–14] such as extended Kalman filter (EKF) [13] and Hamilton-Jacobi equations [12], cannot leverage incomplete data for model training without proper preprocessing.

During the last few decades, data imputation has drawn much attention of researchers because it provides a rational way to make use of incomplete data. Here, imputation or recovery refers to the procedure providing plausible estimations for those missing values (MVs) given other observed values. After imputation, original incomplete data can be converted into complete data and then used in subsequent data analysis tasks. It is feasible to obtain accurate estimation of MVs because data generated in real-world usually possess inherent structure such that close relationship between MVs and observed values can be reliably established based on a set of incomplete samples. In traffic scenarios, due to the connectivity of road segments and the periodicity of people's travel behavior, traffic flow data acquired by different sensors installed in the same road network often have strong spatiotemporal correlation. For instance, traffic flow in different days of a week shows a certain degree of similarity. The upstream traffic and the neighboring downstream traffic will drastically influence each other. As a result, there exist underlying structures hidden from the raw high-dimensional sensor data. Until now, many MVs imputation algorithms have been developed in literatures, including mean imputation (MI), K-nearest neighbor (KNN) [15], support vector machine (SVM) [16], singular value decomposition (SVD) [17], probability principal component analysis (PPCA) [2], low-rank matrix completion (LRMC) [18,19], etc.

Recently, a sample self-representation based method was put forward [20] to recover MVs from incomplete data. Different from the above algorithms, it assumes that original data are derived from multiple low-dimensional linear subspaces [21,22]. In such a case, each sample can be well reconstructed by linearly combining a few of the other samples that belong to the same subspace. Then, MVs can be recovered such that the discrepancy between the observed values and their reconstructed values, namely the reconstruction error, should be as small as possible. In addition, l_1-norm and l_2-norm are utilized as regularization on combination coefficients (weights) to avoid possible overfitting. It also reported that l_1-norm is more preferable than l_2-norm in terms of recovery performance, which is consistent with the advantages of sparse representation (SR) models [21,23], where each sample can be sparsely instead of densely reconstructed by other samples within the same class. Overall, this method has demonstrated improved imputation performance in comparison with other competing algorithms [24,25], thus confirming to some extent the rationality of SR in MVs imputation. However, SR based on either l_1-norm or l_2-norm may fail to take full advantage of all the information contained in samples. It has been revealed that l_1-norm, despite sparsity-inducing characteristic, tends to select only one variable from a group of highly correlated variables [26]. It may not work well for the case where samples are highly correlated to each other. From another point of view, performing SR in the input data space potentially assumes linear relations among data and thus may fail to analyze data with complex nonlinear structure [27,28]. It is seldom that data in the actual environment, such as traffic sensor data, completely conform to such linear relations. As a consequence, the existing SR based method may lead to suboptimal imputation performance when this prerequisite is violated.

Motivated by the above observations, we propose in this paper a novel MVs imputation approach termed as kernel sparse representation with elastic net regularization (KSR-EN), which is applied to road network traffic sensor data. Our work is interesting from the following aspects:

- We perform MVs imputation in kernel-induced high-dimensional feature space instead of the original input space [29,30]. By doing so, nonlinear relationship among data samples can be discovered and leveraged for recovery performance improvement. To the best of our knowledge, this is the first work to which KSR was applied for MVs imputation.
- We propose to apply the combination of l_1-norm and l_2-norm, namely elastic net in statistics literature [26], as regularization on the representation coefficients, with the hope that enough information can be extracted from those highly correlated samples for recovering MVs.

- An iterative algorithm is developed for solving the resulting KSR-EN model by integrating monotone fast iterative shrinkage thresholding algorithm (FISTA) [31–33], as well as projected gradient descent (PGD) approach [24].
- The proposed model is evaluated on both synthetic data and real-work traffic sensor data. The results demonstrate that KSR-EN outperforms other competing algorithms in terms of MVs imputation.

The remainder of this paper is organized as follows. In Section 2, we briefly review some popular approaches for MVs imputation. In Section 3, KSR-EN model and corresponding solution algorithm are proposed. Section 4 reports some experiments on simulated as well as real-world traffic sensor data. Finally, we give some conclusions and future works in Section 5.

2. Related Work

In order to solve the problem of MVs imputation, many algorithms have been put forward by researchers from different perspectives. The most widely applied approaches can be roughly divided into three categories which are discussed in sequel.

2.1. Probabilistic Model Based Methods

For this type of method, a statistical model responsible for producing complete data needs to be specified. One of the most common choices is the multivariate Gaussian model and its extension [34]. Such a model is used to describe the inherent relationship between variables, forming the basis for MVs recovery. The model parameters, along with those MVs, can be estimated simultaneously via alternating optimization. In probabilistic principal component analysis (PPCA) [2,35], the data is assumed to be drawn from a single low-dimensional linear subspace. Then, the likelihood of observed values is derived and maximized through the well-established expectation maximization (EM) algorithm [36]. Furthermore, Bayesian PCA [37] combines the advantages of Bayesian learning [38] and PPCA such that the dimensionality of latent space in PPCA can be automatically inferred based on data. These methods are suitable for data with dominant structure. However, PPCA may perform poorly when missing ratio in data is high [39]. Moreover, how to specify a proper statistical model for real-world data is not a trivial task. In practice, it may be infeasible to postulate a uniformed model for different types of data.

2.2. Regression Model Based Methods

The regression-based methods attempt to establish regression equation to characterize the relationship between a number of observed variables and other variables with MVs based on a set of samples. Various concrete regression techniques can be adopted for this purpose, such as linear regression, support vector machine (SVM) [40,41], and neural networks (NNs) [42]. Then, MVs are replaced with the conditional expectation of the regression results. Local least squares (LLS) regression is a typical regression based imputation method which has been successfully applied in different situations [43]. By using K-nearest neighboring search, LLS first chooses a small number of variables which are most similar to the target variable. Then, least squares criterion is used to establish the regression equation based on the observed data. By doing so, LLS succeeds in exploiting underlying local similarity structure among data and thus is more applicable when data distribute in a nonlinear way. However, LLS may fail in the case of high missing ratio due to unreliable estimation of nearest neighbors or improper regression model.

2.3. Matrix Completion Based Methods

As stated in Section 1, traffic sensor data in the same road network usually exhibit strong spatiotemporal correlation because of road structure, as well as people's travel behavior. As a consequence, the data matrix usually has low-rank property, implying the number of independent rows (and

columns) is much smaller than the size of matrix. Under such a circumstance, low-rank matrix completion (LRMC) [18,19] can be used to recover MVs through rank (or its surrogate nuclear norm) minimization on the whole matrix. Many efficient optimization algorithms have been developed to solve the problem, e.g., SVT [18,44], FPCA [45], ADMM [46], etc. However, LRMC depends on global linear correlation, which is restricted for real data, such as traffic sensor data [19]. Recently, sparse representation (SR) [23] based subspace clustering [21] has drawn much attention because it supposes that the samples are drawn from a union of multiple subspaces, instead of a single one. Due to such advantage, it can reveal complex structure of data, thus leading to better imputation performance [20]. Nevertheless, many datasets in practice are not necessarily well characterized by multiple linear models, and in such cases, existing algorithms may produce suboptimal imputation results [47].

3. Kernel Sparse Representation with Elastic Net Regularization

3.1. Linear and Kernel SR-EN

Given data matrix $X = [x_1, x_2, \cdots, x_N] \in R^{p \times N}$, where sample $x_i = [x_i(1), x_i(2), \cdots, x_i(p)]^T \in R^p$, p and N denote the number of features and samples, respectively. In the case we study, not all of the entries in sample X_i are known. Therefore, let Ω indicate the index set of observed entries in X, that is, for all $(i,j) \in \Omega$, $x_i(j)$ is observable. The task here is to obtain an accurate estimation for those $x_i(j)$, $(i,j) \notin \Omega$. Sharing similar flavor with SR [23] and sparse subspace clustering [21,48], we hope that each sample can be well approximated as a linear combination of other samples, i.e., $x_i \approx \sum_{j=1}^{N} x_j w_i(j)$ where $w_i(j)$ is an element in coefficient matrix $W \in R^{N \times N}$, indicating the contribution of x_j in reconstructing x_i. Additionally, the constraint $w_i(i) = 0$, i.e., $\mathrm{diag}(W) = 0$ in matrix form, is added to avoid trivial solution. Besides this reconstruction criterion, we also impose a penalty to W so as to alleviate possible overfitting. In this work, elastic net regularization [26], as a hybrid of l_1-norm and l_2-norm, is adopted since it not only achieves sparse variable selection but also would benefit from highly correlated variables. By integrating the above ingredients together, we propose the following MVs imputation model termed as SR with elastic net regularization (SR-EN):

$$\begin{aligned} \min_{X,W} \frac{1}{2}\|X - XW\|^2 + C\alpha\|W\|_1 + \frac{C(1-\alpha)}{2}\|W\|^2 \\ \text{s.t. } \mathrm{diag}(W) = 0,\ x_i(j) = m_i(j),\ (i,j) \in \Omega \end{aligned} \tag{1}$$

where $\|W\|_1 = \sum_{i=1}^{N}\sum_{j=1}^{N}|w_i(i)|$, $\|W\|^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}(w_i(i))^2$, $m_i(j)$ is the observed value for $x_i(j)$, $C > 0$, and $1 \geq \alpha \geq 0$ are two parameters used to balance the role of regularization and l_1-norm, respectively. When $\alpha = 1$ or $\alpha = 0$, elastic net regularization will degenerate to pure l_1-norm or l_2-norm regularization.

The above SR-EN model is able to recover missing values when samples distribute in a union of multiple linear subspaces. However, it may produce suboptimal imputation when applied to data with nonlinear structure. Therefore, we further extend SR-EN to deal with samples distributed nonlinearly in raw input space. To achieve this goal, motivated by kernel method [5], we first map the original input space R^p to a reproducing kernel Hilbert space (RKHS) H with higher or even infinite dimensionality, by employing nonlinear mapping function $\phi : R^p \to H$. Let $\phi(x) \in H$ be the image of sample x in feature space H, and $\phi(X) = [\phi(x_1), \phi(x_2), \cdots, \phi(x_N)]$ denotes the entire sample matrix after mapping. Similar to classic kernel-based learning, we make an assumption that nonlinear distribution of a given dataset in original input space R^p can be well converted into linear distribution in feature space H with much higher dimensionality, and thus facilitating the application of SR-EN. Based on the above discussion, the kernel SR-EN (KSR-EN) we propose can be expressed as:

$$\min_{X,W} \frac{1}{2}\|\phi(X) - \phi(X)W\|^2 + C\alpha\|W\|_1 + \frac{C(1-\alpha)}{2}\|W\|^2 \quad \text{s.t. } \mathrm{diag}(W) = 0,\ x_i(j) = m_i(j),\ (i,j) \in \Omega \tag{2}$$

Figure 1 presents the effect of nonlinear mapping. Note that the above KSR-EN model reduces to its linear version (1) when the mapping function ϕ is linear, namely, $\phi(x) = x$. Therefore, in what follows, we will concentrate on KSR-EN (2) and develop an effective algorithm for solving it. We let $\lambda_1 = C\alpha$ and $\lambda_2 = C(1-\alpha)$ to simplify notations when deriving optimization algorithm.

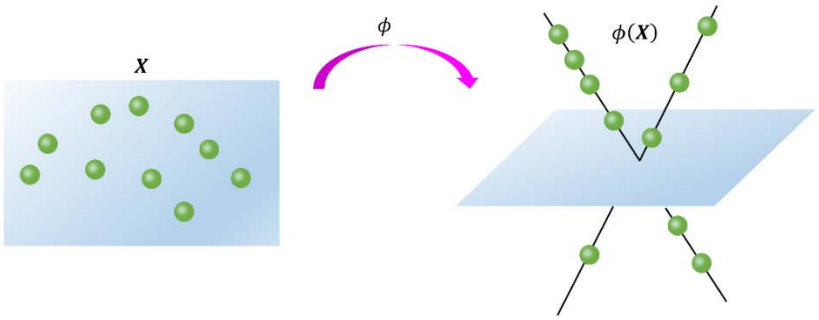

Figure 1. Nonlinear mapping in the proposed method.

3.2. Optimization Algorithm

As we can see from (2), the coupling between decision variables X and W makes it difficult to find optimal solutions for X and W simultaneously. Therefore, in this work, we choose alternating optimization scheme to find the optimal solution iteratively. Specifically, we attempt to solve (2) by alternatively optimizing over W and X while holding the other variable fixed.

Optimize W while fixing X. In such a case, (2) leads to the following optimization problem:

$$\min_{W} \frac{1}{2}\|\phi(X) - \phi(X)W\|^2 + \lambda_1\|W\|_1 + \frac{\lambda_2}{2}\|W\|^2 \tag{3}$$

In terms of our problem (3), the objective can be decomposed into the sum of two parts $f(W) + g(W)$, where $f(W) = \frac{1}{2}\|\phi(X) - \phi(X)W\|^2$ and $g(W) = \lambda_1\|W\|_1 + \frac{\lambda_2}{2}\|W\|^2$. Let us focus on the first part, $f(W)$. It is obvious that $f(W)$ is convex and differentiable and can be reformulated as $f(W) = \frac{1}{2}\mathrm{tr}(\phi(X)^T\phi(X) - \phi(X)^T\phi(X)W - W^T\phi(X)^T\phi(X) + W^T\phi(X)^T\phi(X)W)$, where $\mathrm{tr}(\cdot)$ denotes the trace operator of a matrix. Furthermore, based on the rule of matrix derivative [49], the gradient of $f(W)$ w.r.t. W is given by $\nabla f(W) = \phi(X)^T\phi(X)W - \phi(X)^T\phi(X)$. We notice immediately that both $f(W)$ and its gradient, $\nabla f(W)$, depend exclusively on the inner products between the images of all pairs of samples in feature space, without having to give explicit representation for ϕ. In other words, we can introduce kernel Gram matrix $K = \phi(X)^T\phi(X)$, with element K_{ij} computed as $K_{ij} = \phi(x_i)^T\phi(x_j)$. Then, we have:

$$f(W) = \frac{1}{2}\mathrm{tr}\left(K - KW - W^TK + W^TKW\right) \text{ and } \nabla f(W) = KW - K \tag{4}$$

According to Mercer's theorem [50], any matrix K could be a valid kernel as long as it is positive semi-definite (PSD). Some commonly used kernels include polynomial kernel, radial basis function (RBF) kernel, sigmoid kernel, etc. In this work, RBF kernel, defined as $K_{ij} = \phi(x_i)^T\phi(x_j) = e^{-\gamma\|x_i - x_j\|^2}$, is used because of its simplicity, along with good empirical performance in various kernel-based learning algorithms. In RBF kernel, γ is a free parameter controlling the smoothness degree of the kernel.

On the other hand, the second part, $g(W)$, i.e., the regularization term, is convex yet nondifferentiable, thus restricting the application of traditional gradient descent algorithm.

Nevertheless, considering the facts that $\nabla f(W)$ is Lipschitz-continues and $g(W)$ has a closed-form proximity operator, we develop a first-order algorithm to find the optimal solution of (3) under the proximal gradient descent framework. This framework, also known as fast iterative shrinkage thresholding approach (FISTA), has been widely used to solve various sparsity-related problems, such as [33].

Specifically, given the Lipschitz constant L of $\nabla f(W)$ and the current solution W_k of (3) at the k-th iteration, it is possible to construct an approximating function $q(W, W_k)$ majorizing the original $f(W)$ at W_k:

$$q(W, W_k) = f(W_k) + \langle W - W_k, \nabla f(W_k) \rangle + \frac{L}{2}\|W - W_k\|^2 + g(W) \tag{5}$$

where the definitions of $f(W_k)$ and $\nabla f(W_k)$ are given in (4). From the definition of $q(W, W_k)$, we have:

$$q(W, W_k) \geq f(W), \ \forall W \tag{6}$$

where the equality holds if and only if $W = W_k$. This fact motivates the following update:

$$W_{k+1} = \underset{W}{\operatorname{argmin}} q(W, W_k) \tag{7}$$

We can prove that W_{k+1} will lead to an improved objective value for (3) because:

$$f(W_k) = q(W_k, W_k) \geq q(W_{k+1}, W_k) \geq f(W_{k+1}) \tag{8}$$

In order to solve (7), we rewrite the objective (5) as:

$$q(W, W_k) \propto \frac{L}{2}\left\|W - \left(W_k - \frac{\nabla f(W_k)}{L}\right)\right\|^2 + g(W) \tag{9}$$

Let $V_k = W_k - \frac{\nabla f(W_k)}{L}$ and incorporating the definition of $g(W)$, W_{k+1} can be obtained by:

$$W_{k+1} = \underset{W}{\operatorname{argmin}} \frac{L}{2}\|W - V_k\|^2 + \lambda_1\|W\|_1 + \frac{\lambda_2}{2}\|W\|^2 \tag{10}$$

Notice that all of the entries in W are independent of each other, and thus can be optimized separately and parallelly. For the sake of simplicity, consider the proximity operator for elastic net regularization with univariate as follows:

$$\text{Prox}(b) = \underset{a}{\operatorname{argmin}} \frac{1}{2}(a - b)^2 + \lambda_1|a| + \frac{\lambda_2}{2}a^2 \tag{11}$$

where $a, b \in R$. Problem (11) has a closed-form optimal solution as:

$$\text{Prox}(b) = \text{sign}(b)\left(\frac{|b| - \lambda_1}{1 + \lambda_2}\right)_+ \tag{12}$$

where $(a)_+ = \max\{0, a\}$. Incorporating the above result into (10), one can obtain the solution W_{k+1} as:

$$W_{k+1} = \text{sign}(V_k)\left(\frac{L|V_k| - \lambda_1}{L + \lambda_2}\right)_+ \tag{13}$$

The whole algorithm for solving (3) is summarized in Algorithm 1.

Algorithm 1. Monotone fast iterative shrinkage thresholding algorithm for solving (3)

Input: Estimated data matrix X, trade-off parameters λ_1, λ_2, Gaussian kernel parameter γ
Initialize W_0
Output: Coefficient matrix W
Procedure
Calculate $N \times N$ kernel matrix K with $K_{ij} = e^{-\gamma\|x_i - x_j\|^2}$, $L = 1.1\|K\|_2$, $Y_1 = W_0$, $k = 1$, $t_1 = 1$
while not converge **do**
$Z_k = Y_k - \frac{KY_k - K}{L}$
$Z_k = \text{sign}(Z_k)\left(\frac{L|Z_k| - \lambda_1}{L + \lambda_2}\right)_+$
$\text{diag}(Z_k) = 0$
$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
$W_k = \min_W f(W) + g(W)$ s.t. $W = Z_k, W_{k-1}$
$Y_{k+1} = W_k + \left(\frac{t_k}{t_{k+1}}\right)(Z_k - W_k) + \left(\frac{t_k - 1}{t_{k+1}}\right)(W_k - W_{k-1})$
$k = k + 1$
end while

Optimize X while fixing W. In such a case, (2) leads to the following optimization problem with equality constraint:

$$\begin{aligned} \min_X & \text{tr}\left(K - KW - W^T K + W^T K W\right) \\ & \text{s.t. } x_i(j) = m_i(j),\ (i,j) \in \Omega \end{aligned} \tag{14}$$

where we have used the results in (4). To solve (14), we employ the projected gradient descent method with Armijo step size rule. The essential ingredient is the calculation of derivative w.r.t. decision variables.

Denoting the objective in (14), as $h(X) = \text{tr}\left(K - KW - W^T K + W^T K W\right)$, its partial derivative w.r.t. the kernel matrix K is given by:

$$\frac{\partial h(X)}{\partial K} = (I - W)(I - W)^T \tag{15}$$

Furthermore, for RBF kernel $K_{ij} = e^{-\gamma\|x_i - x_k\|^2}$, we can easily compute:

$$\frac{\partial K_{ik}}{\partial x_i} = -2\gamma(x_i - x_k)e^{-\gamma\|x_i - x_k\|^2} \tag{16}$$

To calculate the derivative of $h(X)$ w.r.t. the entries in X, we apply the chain rule to get:

$$\frac{\partial h(X)}{\partial x_i(j)} = \sum_k \frac{\partial h(X)}{\partial K_{ik}} \frac{\partial K_{ik}}{\partial x_i(j)} \tag{17}$$

where $\frac{\partial h(X)}{\partial K_{ik}}$ and $\frac{\partial K_{ik}}{\partial x_i(j)}$ can be obtained from (15) and (16), respectively. For clarity, the algorithm for solving (14) is presented in Algorithm 2.

Algorithm 2. Gradient descent algorithm for solving (14)

Input: Coefficient matrix W
Output: Estimated data matrix X
Procedure
Initialize data matrix X
while not converge **do**
Compute the derivative of $h(X)$ w.r.t. X, that is $\frac{\partial h(X)}{\partial X}$ using (17)
Let $\frac{\partial h(X)}{\partial x_i(j)} = 0$, for any $(i,j) \in \Omega$
Find step size l with Armijo rule, that is, choose $l = \max\left\{1, \frac{1}{2}, \frac{1}{4}, \cdots\right\}$ such that

$$h(X) - h\left(X - l\frac{\partial h(X)}{\partial X}\right) \geq \frac{l}{4}\left\|\frac{\partial h(X)}{\partial X}\right\|^2$$

Update $X = X - l\frac{\partial h(X)}{\partial X}$
end while

4. Experiments

4.1. Configuration

Besides the proposed KSR-EN (and its linear version SR-EN), we include some typical imputation algorithms including LLS, PPCA, and LRMC in order to comprehensively evaluate their performance. As stated in Section 2, LLS is a regression based method, PPCA relies on statistical assumption of data, and LRMC imposes low-rank property of sample matrix. Some recent studies [2,51] have manifested that PPCA and LRMC are two effective approaches for traffic sensor data. All of these methods are implemented in MATLAB 2015a on a PC with Core i7 2.4 GHz CPU and 12GB RAM. Following [20,24], the parameters in each method, such as the number of K-nearest neighbors in LLS, the dimensionality of latent space in PPCA, are tuned to give optimal performance.

To measure the accuracy of each method, we randomly produce MVs and then employ different methods to obtain corresponding estimations. Finally, the estimated values are compared with the real values and the difference between them are calculated. In this work, two widely applied metrics, i.e., root mean square error (RMSE) and relative error (RELERR), are calculated as follows:

$$RMSE = \sqrt{\frac{1}{T}\sum_{(i,j)\notin\Omega}(x_i(j) - \hat{x}_i(j))^2} \tag{18}$$

$$RELERR = \sqrt{\frac{\sum_{(i,j)\notin\Omega}(x_i(j) - \hat{x}_i(j))^2}{\sum_{(i,j)\notin\Omega}(x_i(j))^2}} \tag{19}$$

where T denotes the total number of missing entries, and $\hat{x}_i(j)$ and $x_i(j)$ denote the imputed value and the real value, respectively. Obviously, the smaller the RMSE and RELERR, the better the imputation performance. In addition, we repeat each test 10 times and report the mean imputation error and the associated standard deviation, so as to reduce the potential bias caused by randomness.

4.2. Synthetic Data

We first evaluate the proposed method on nonlinear synthetic dataset to intuitively illustrate its behavior. The simulated samples are shown as the red points in Figure 2. The specific equations for generating these samples are described as follows:

$$\text{Arc 1}: \ x_1 = sin(t), \ x_2 = cos(t) - 1, \ x_3 = t$$
$$\text{Arc 2}: \ x_1 = 1 - cos(t), \ x_2 = -sin(t), \ x_3 = t$$

where latent variable t is sampled from a one-dimensional uniform distribution in the interval $\left[-\frac{\pi}{2}, 0\right]$. As we can see, the entire set of samples is drawn from two arcs which intersect at the

origin. The intrinsic dimension for each arc is 1. Gaussian noise with zero mean and 0.05 standard deviation is added to each sample. We synthesize 100 samples from each arc and the whole sample set, organized in matrix $X \in R^{3\times 200}$ in this case, forms a nonlinear structure in three-dimensional space. To get incomplete matrix, a randomly selected entry for each sample is removed, thus resulting in an incomplete matrix with missing ratio of 33.33%. Then, different imputation approach is applied to restore the missing entries of the data matrix. Figure 3 show the results in one experiment. Notice that in Figure 3, the first column depicts the real samples (red points) and the imputed samples (blue points), the second column shows the scatter plot of real and estimated values, and the last column further shows the residual. The averaged errors obtained by different methods across 10 tests are summarized in Table 1, where the best results are highlighted in bold.

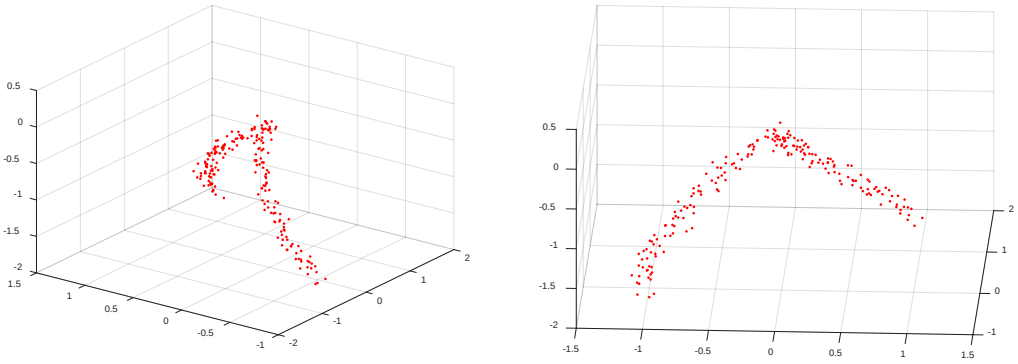

Figure 2. Illustration of simulated data from two different viewpoints.

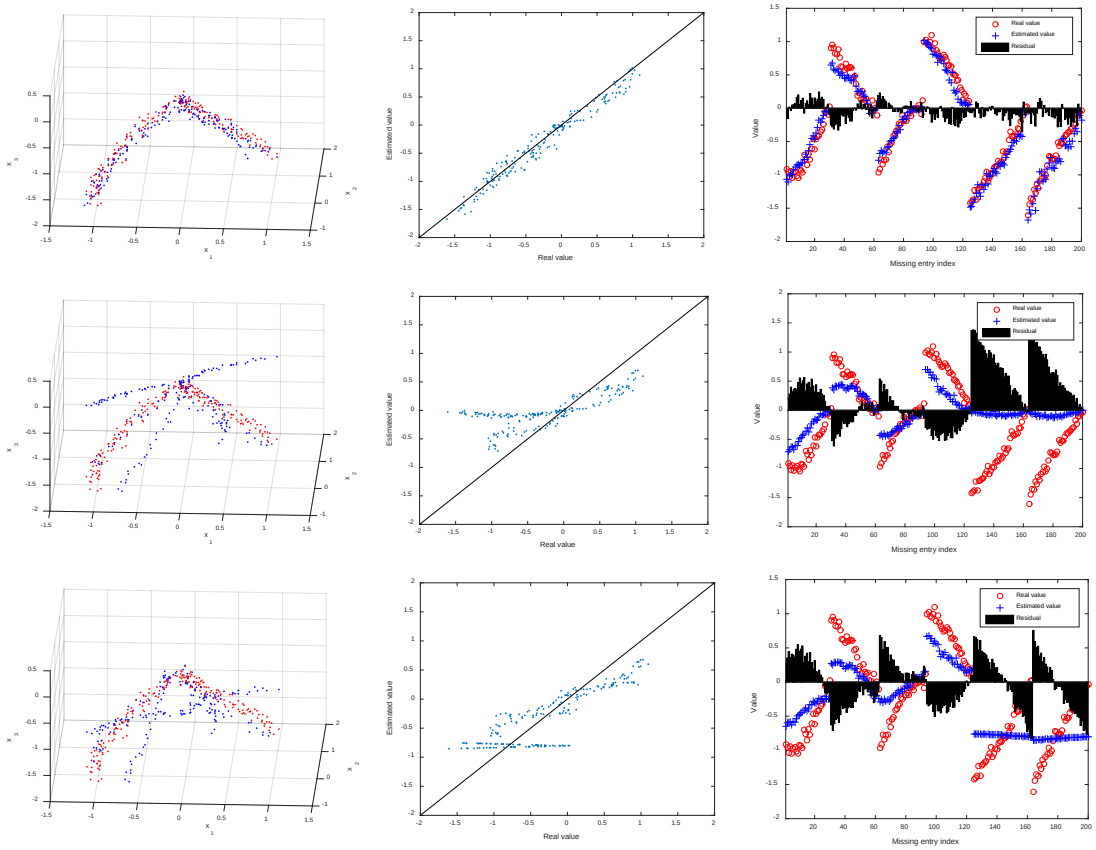

Figure 3. *Cont.*

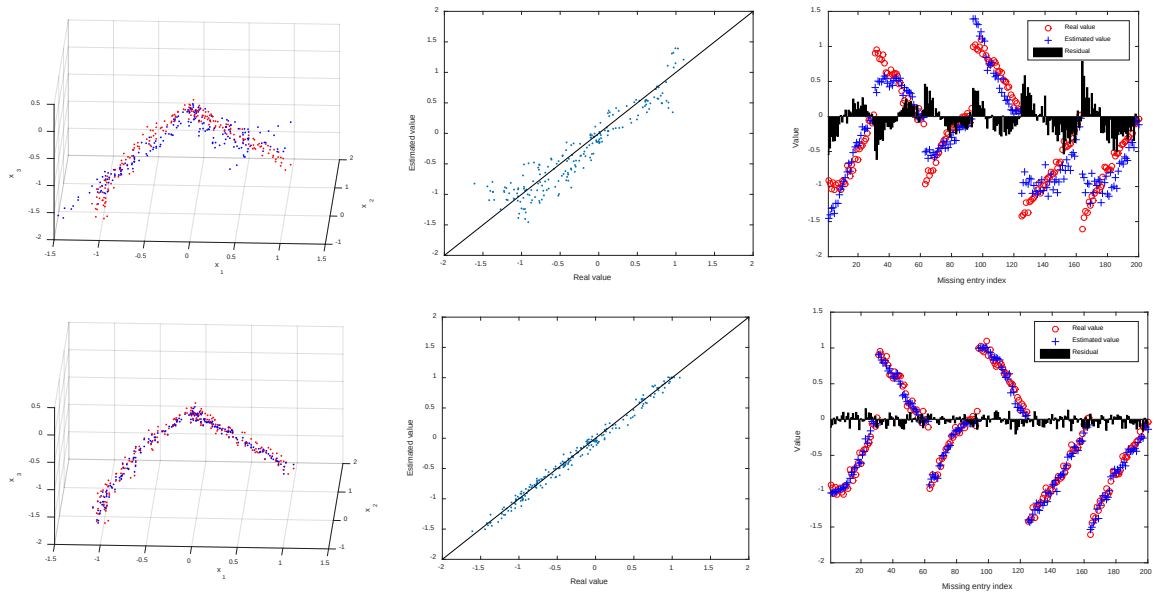

Figure 3. Imputation results obtained by local least squares (LLS), low-rank matrix completion (LRMC), probability principal component analysis (PPCA), sparse representation with elastic net regularization (SR-EN), and kernel sparse representation with elastic net regularization (KSR-EN) (from **top** to **bottom**).

Table 1. Imputation error root mean square error (RMSE) and relative error (RELERR) obtained by different approaches. The best results are highlighted in bold.

Metrics	Compared Methods			Ours	
	LLS	**LRMC**	**PPCA**	**SR-EN**	**KSR-EN**
RMSE	13.89 ± 1.30	56.05 ± 1.86	37.77 ± 1.55	25.07 ± 2.80	$\mathbf{7.00 \pm 0.37}$
RELERR	19.23 ± 1.85	77.60 ± 2.79	52.23 ± 2.24	34.68 ± 3.60	$\mathbf{9.69 \pm 0.51}$

As can be seen from these results, LRMC performs worst since it heavily depends on the global low-rank structure of data, which is violated in our case. Instead of minimizing the rank of data for recovery, PPCA makes use of maximum likelihood estimation and EM algorithm to jointly optimize model parameters and MVs. As a result, PPCA works better than LRMC, although it also imposes Gaussian distribution for data. SR-EN further improves the results of PPCA, although it is also based on linear structure of data. Different from PPCA, SR-EN is applicable to data lying on or close to a union of linear subspaces [48]. LLS achieves significantly smaller errors than LRMC, PPCA, and SR-EN. It may be because LLS is an imputation algorithm based on local rather than global linear relationship between samples. Finally, as expected, KSR-EN, as nonlinear extension of SR-EN, consistently outperforms all other imputation algorithms under both performance metrics. For example, the imputed results obtained by KSR-EN fit well with the true values according to Figure 3. For more clearly illustrating this, we present in Figure 4 the coefficient matrix W obtained by SR-EN and KSR-EN. Note that the sample index in the figure roughly reflects the proximity relation between samples. We can see that W derived from KSR-EN exhibits clear sparse and diagonal structure, indicating that in the high-dimensional feature space, each sample can be well represented as a linear combination of its neighboring samples within the same class. It also confirms that KSR-EN successfully discovers the underlying nonlinear structure of multiple subspaces. In contrast, the diagonal structure spreads in wider range in W obtained by SR-EN. For KSR-EN, we also notice from the figure of KSR-EN that some nonzero values in W appear far away from the main diagonal, seemingly violating our hypothesis. After careful analysis, we found that these off-diagonal nonzero values are mainly caused by the samples drawn from the intersection area of two arcs. Since the samples in this area are heavily mixed, it may be impossible to distinguish them completely.

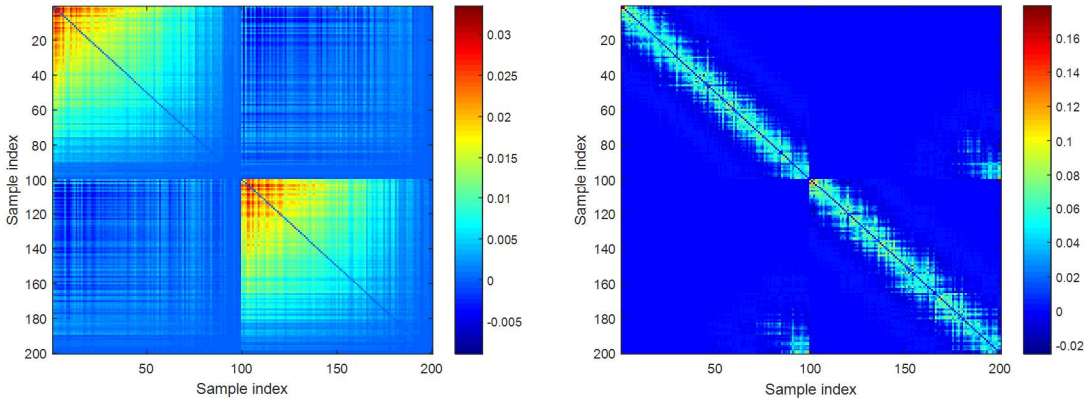

Figure 4. Illustration of coefficient matrix obtained by SR-EN (**left**) and KSR-EN (**right**).

4.3. Traffic Sensor Data Imputation

4.3.1. Data Collection

In this experiment, we evaluate the proposed KSR-EN algorithm on a real-world traffic flow dataset which is publicly accessible at http://portal.its.pdx.edu/. The data is collected from 40 loop detectors installed at Interstate 205 (I205) interstate highways. The selected sub-area road network is shown in Figure 5. The traffic volume is aggregated every 15 min and the unit is vehicles per 15 min (veh/15 min). For each loop detector, 96 data points is recorded per day. We use traffic data of 30 weekdays in the year of 2015 since the traffic flow profile on weekends and holidays are very different from weekdays. The total number of traffic volumes is $96 \times 40 \times 30 = 115{,}200$. The whole data is organized as a 96×1200 matrix with each column denoting a sample. Figure 6 illustrates the traffic flow profiles captured by 40 sensors in the same day. As can be seen, despite of overall similarity, the variation of traffic flow at different road segments shows remarkable difference in terms of maximum traffic flow, the duration of rush hours, etc.

Figure 5. The selected sub-area road network of Portland, OR, USA.

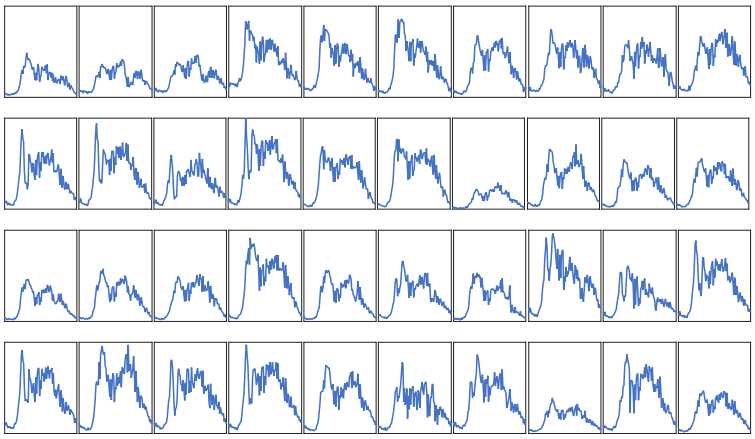

Figure 6. Illustration of traffic flow profiles from 40 sensors in the same day.

In order to comprehensively evaluate the recovery performance of different imputation algorithms under complex environments, we artificially inject MVs into original data by simulating three different missing patterns [2]. (i) Missing completely at random (MCAR), which means the presence or absence of MVs is completely independent of observed values and other parameters of interest. (ii) Missing at random (MAR), indicating that the occurrence of MVs depends on its neighboring points. (iii) Mixture of MCAR and MAR (MIXED), where half of MVs obey MCAR and the other half are from MAR. Generally speaking, MCAR is easier than MAR in the sense of recovery because MVs in the former case appear as isolated points randomly distributed, while MVs in the latter case often look like continuous missing, thus making accurate imputation more challenge. In addition, we also change the missing ratio δ, that is, the ratio of the number of missing entries to the total number of entries in data matrix, from 0.1 to 0.5 with step 0.1. Obviously, the larger δ, the harder the imputation task.

4.3.2. Imputation Error

Tables 2–4 summarize the imputation errors obtained by different methods under MCAR, MAR, and MIXED missing patterns, respectively. The best results are highlighted in bold for clarity. From these results, we can obtain some interesting observations. As a typical local structure (instead of global structure) based approach, LLS is able to produce small imputation errors when the missing ratio is low. However, the performance deteriorates rapidly with increased missing ratio. It may be because that the reliable estimation of local structure or neighboring samples turns out to be very difficult given a set of samples with many unknown entries. Different from LLS, LRMC and PPCA all depend on the global linear subspace structure of data. PPCA works slightly better than LRMC on this dataset, and both outperform LLS when δ is larger than 0.3. However, these methods may fail to deliver good performance on the dataset with complex intrinsic structure. SR-EN, our proposed linear imputation approach, works better than the above conventional approaches, since it successfully accounts for the potential multiple linear subspace structure, thus avoiding the assumption of single subspace. KSR-EN further relax the restriction on linear subspace structures of SR-EN via effective nonlinear mapping from original input space to a high-dimensional feature space. By doing so, KSR-EN is able to explore multiple linear subspace structure in feature space, which is nonlinear in input space, and thus enhance the performance of SR-EN. As a result, KSR-EN is superior to SR-EN and the other competing approaches w.r.t. imputation errors, regardless of specific missing pattern or missing ratio. It suggests that the integration of nonlinear kernel mapping, as well as SR-EN, allows significant performance gain for the recovery of MVs in traffic scenarios.

Table 2. Imputation error RMSE and RELERR obtained by different approaches under missing completely at random (MCAR) missing pattern. The best results are highlighted in bold.

δ	Metrics	Compared Methods			Ours	
		LLS	**LRMC**	**PPCA**	**SR-EN**	**KSR-EN**
0.1	RMSE	76.62 ± 0.60	80.63 ± 0.86	79.06 ± 1.36	74.12 ± 1.22	$\mathbf{67.52 \pm 0.43}$
	RELERR	11.22 ± 0.09	11.80 ± 0.08	11.58 ± 0.15	10.85 ± 0.21	$\mathbf{9.89 \pm 0.04}$
0.2	RMSE	82.26 ± 0.70	83.30 ± 0.24	82.98 ± 0.24	76.43 ± 0.23	$\mathbf{70.65 \pm 0.33}$
	RELERR	12.06 ± 0.12	12.21 ± 0.05	12.16 ± 0.05	11.20 ± 0.02	$\mathbf{10.36 \pm 0.06}$
0.3	RMSE	89.94 ± 0.70	86.58 ± 0.74	85.47 ± 0.72	82.62 ± 1.84	$\mathbf{75.05 \pm 0.96}$
	RELERR	13.18 ± 0.08	12.69 ± 0.06	12.53 ± 0.06	12.11 ± 0.23	$\mathbf{11.00 \pm 0.10}$
0.4	RMSE	99.01 ± 0.56	90.94 ± 0.50	89.69 ± 0.63	85.83 ± 1.46	$\mathbf{80.42 \pm 0.63}$
	RELERR	14.49 ± 0.08	13.31 ± 0.04	13.13 ± 0.06	12.56 ± 0.19	$\mathbf{11.77 \pm 0.07}$
0.5	RMSE	113.03 ± 0.36	95.91 ± 0.44	95.83 ± 0.69	91.85 ± 1.46	$\mathbf{87.11 \pm 0.43}$
	RELERR	16.54 ± 0.04	14.04 ± 0.04	14.02 ± 0.07	13.46 ± 0.25	$\mathbf{12.75 \pm 0.04}$

Table 3. Imputation error RMSE and RELERR obtained by different approaches under missing at random (MAR) missing pattern. The best results are highlighted in bold.

δ	Metrics	Compared Methods			Ours	
		LLS	**LRMC**	**PPCA**	**SR-EN**	**KSR-EN**
0.1	RMSE	84.31 ± 1.11	98.31 ± 0.46	90.14 ± 0.88	87.44 ± 0.75	$\mathbf{77.84 \pm 0.99}$
	RELERR	12.00 ± 0.16	13.99 ± 0.10	12.83 ± 0.03	12.45 ± 0.22	$\mathbf{11.08 \pm 0.20}$
0.2	RMSE	91.89 ± 0.56	100.86 ± 1.17	95.13 ± 1.27	90.57 ± 1.41	$\mathbf{82.41 \pm 0.97}$
	RELERR	13.12 ± 0.04	14.40 ± 0.13	13.58 ± 0.17	12.93 ± 0.16	$\mathbf{11.76 \pm 0.11}$
0.3	RMSE	100.07 ± 1.51	103.06 ± 1.32	97.45 ± 1.28	94.56 ± 1.16	$\mathbf{87.63 \pm 0.85}$
	RELERR	14.34 ± 0.19	14.77 ± 0.15	13.96 ± 0.15	13.55 ± 0.16	$\mathbf{12.56 \pm 0.12}$
0.4	RMSE	112.54 ± 1.17	106.19 ± 0.79	101.50 ± 1.21	99.54 ± 0.73	$\mathbf{92.12 \pm 1.73}$
	RELERR	16.18 ± 0.17	15.26 ± 0.12	14.59 ± 0.16	14.31 ± 0.11	$\mathbf{13.24 \pm 0.27}$
0.5	RMSE	130.85 ± 1.96	109.78 ± 0.69	105.50 ± 0.87	104.55 ± 1.00	$\mathbf{99.21 \pm 0.98}$
	RELERR	18.84 ± 0.27	15.81 ± 0.11	15.19 ± 0.10	15.06 ± 0.15	$\mathbf{14.29 \pm 0.14}$

Table 4. Imputation error RMSE and RELERR obtained by different approaches under MIXED missing pattern. The best results are highlighted in bold.

δ	Metrics	Compared Methods			Ours	
		LLS	**LRMC**	**PPCA**	**SR-EN**	**KSR-EN**
0.1	RMSE	79.48 ± 1.67	89.57 ± 1.01	84.40 ± 2.17	80.48 ± 1.09	$\mathbf{71.98 \pm 1.79}$
	RELERR	11.46 ± 0.28	12.91 ± 0.18	12.17 ± 0.32	11.60 ± 0.18	$\mathbf{10.38 \pm 0.29}$
0.2	RMSE	86.60 ± 0.47	92.73 ± 0.05	89.37 ± 0.71	84.96 ± 1.44	$\mathbf{77.36 \pm 1.38}$
	RELERR	12.52 ± 0.12	13.40 ± 0.09	12.92 ± 0.07	12.28 ± 0.26	$\mathbf{11.18 \pm 0.16}$
0.3	RMSE	93.75 ± 0.86	95.18 ± 1.10	91.37 ± 1.22	87.88 ± 0.92	$\mathbf{81.06 \pm 2.35}$
	RELERR	13.58 ± 0.11	13.78 ± 0.13	13.23 ± 0.15	12.73 ± 0.14	$\mathbf{11.74 \pm 0.31}$
0.4	RMSE	104.31 ± 0.52	99.00 ± 0.57	95.26 ± 0.62	93.85 ± 0.71	$\mathbf{88.28 \pm 1.68}$
	RELERR	15.13 ± 0.06	14.36 ± 0.05	13.82 ± 0.09	13.61 ± 0.07	$\mathbf{12.80 \pm 0.24}$
0.5	RMSE	118.92 ± 1.31	103.14 ± 0.96	100.83 ± 0.65	98.33 ± 0.78	$\mathbf{94.02 \pm 1.00}$
	RELERR	17.28 ± 0.17	14.99 ± 0.12	14.65 ± 0.08	14.23 ± 0.10	$\mathbf{13.66 \pm 0.12}$

4.3.3. Influence of Parameters

In this work, the linear combination of l_1-norm and l_2-norm, namely elastic net, is used as regularization to encourage highly correlated samples can be selected for reconstructing each target sample. As a result, the trade-off parameter α plays an indispensable role in the final model. On one hand, small α means that l_2-norm dominates the regularization and makes the solution dense. On the other hand, large α will increase the portion of l_1-norm and thus leads to solution with more sparsity. To this end, we first study the impact of parameter α. In this example, we focus on MCAR missing pattern and missing ratio $\delta = 0.2$. To investigate the influence of α, we fix $C = 2^{-5}$ and change α from 0 to 1 with step 0.2. The variation of RMSE and RELERR w.r.t. α is shown in Figure 7. Moreover, one sample is selected to show the variation of representation coefficients w.r.t. α. The results are shown in Figure 8. As we can see from Figure 7, small α and large α both degrade the imputation performance. From the viewpoint of sparsity, when α equals to zero, the resulting coefficients are dense. Large α tends to shrink many elements in the coefficients towards zero. We observe from extensive experiments that in most cases, optimal performance is achieved when α is between 0 and 1, thus confirming the effectiveness of elastic net regularization.

In what follows, we investigate the influence of trade-off parameter C. Through tuning parameter C, we can control the strength of regularization on coefficient matrix. Small C will weaken the role of regularization and leads to coefficient matrix with large magnitude values. It will make the resulting model fit well on the observed values but perform worse on other MVs. This is also known as overfitting in machine learning [52]. In contrast, large C tends to drive many coefficients to be small (or exact zero) and produces a model incapable of characterizing the inherent structure of data sufficiently. Similar to the above experiment, we fix $\alpha = 0.6$ and tune the value of C from $\{2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}\}$. The variation of RMSE and RELERR is shown in Figure 9. Accordingly, the variation of representation coefficients of one selected sample is illustrated in Figure 10. We notice that when C is not too large or small, the obtained performance is satisfactory. This observation is consistent with the above analysis we made.

Next, we investigate the influence of aggregation time on the imputation performance of each method. Larger aggregation time leads to smoother traffic flow profile, thus causing loss of detail information. Due to different aggregation time, the variation range of traffic flow in unit interval is much different. Therefore, relative error is more suitable when comparing imputation performance under different aggregation time. The aggregation time is set to be 30 min and the experimental results are shown in Figure 11. As can be seen, the proposed KSR-EN outperforms all the other approaches regardless of specific missing pattern or missing ratio, thus verifying the effectiveness of KSR-EN again.

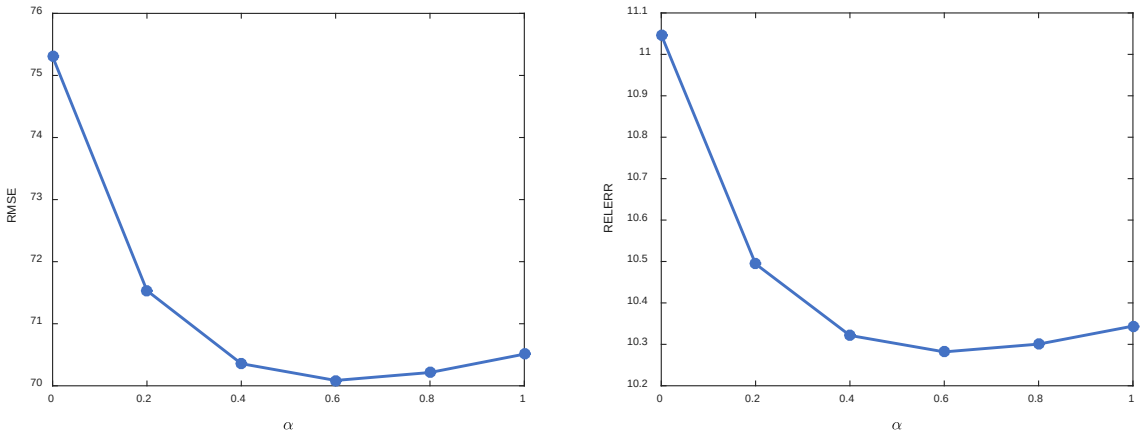

Figure 7. Performance variation of KSR-EN w.r.t. different values of α.

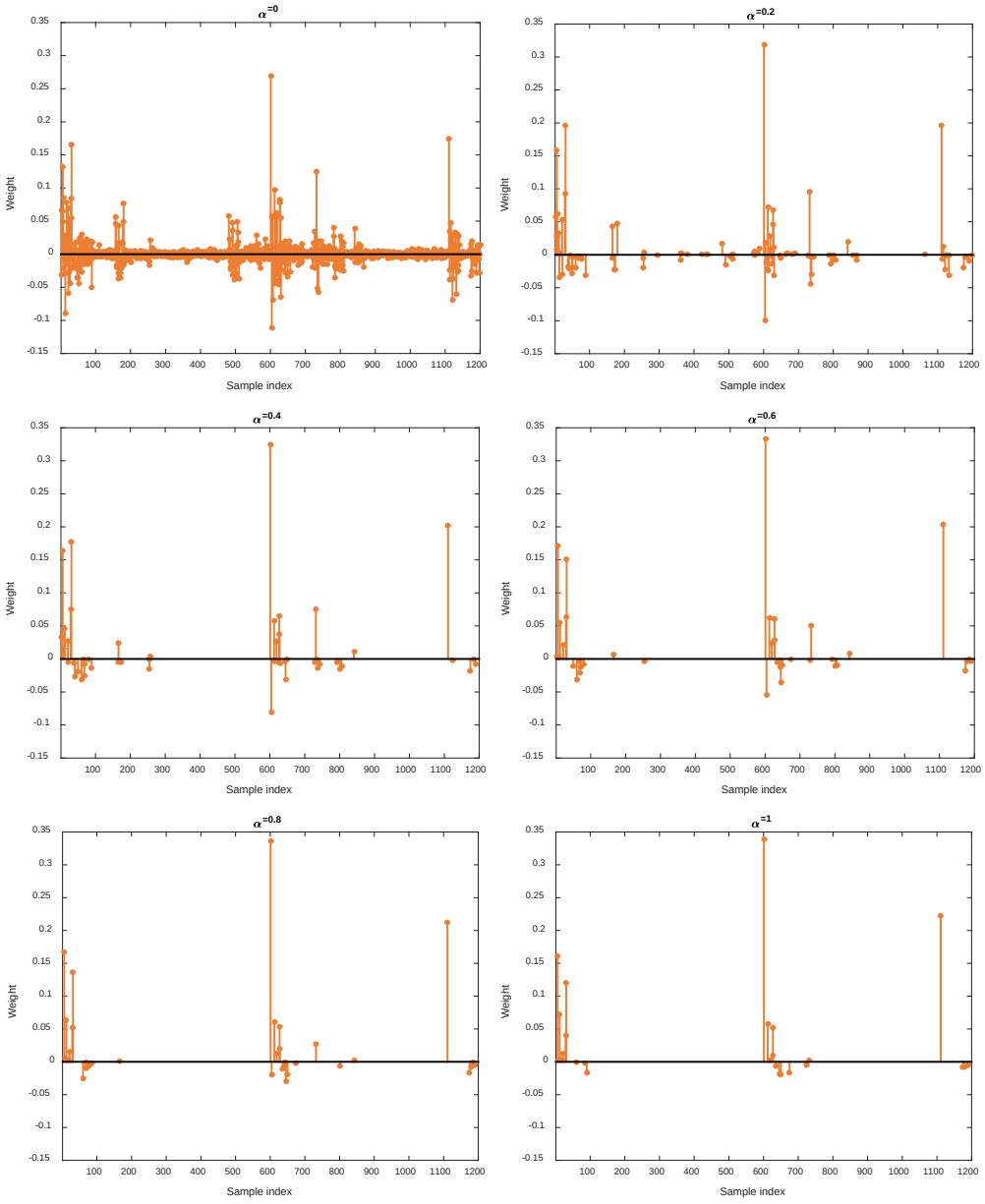

Figure 8. Variation of sparsity w.r.t. different values of α.

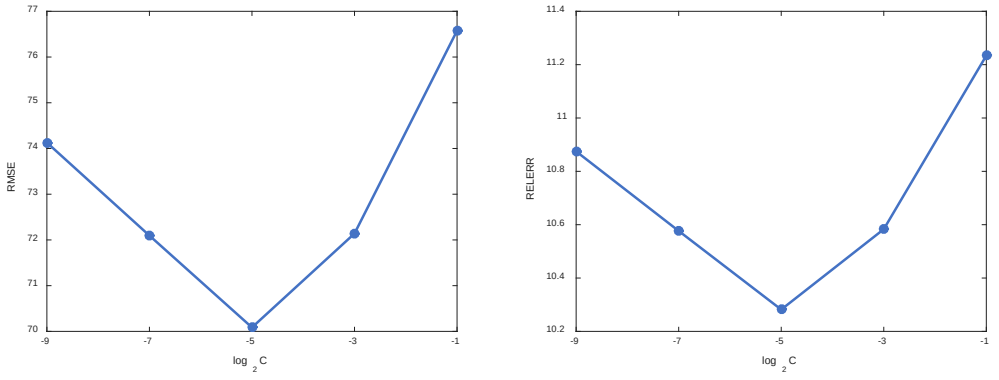

Figure 9. Performance variation of KSR-EN w.r.t. different values of C.

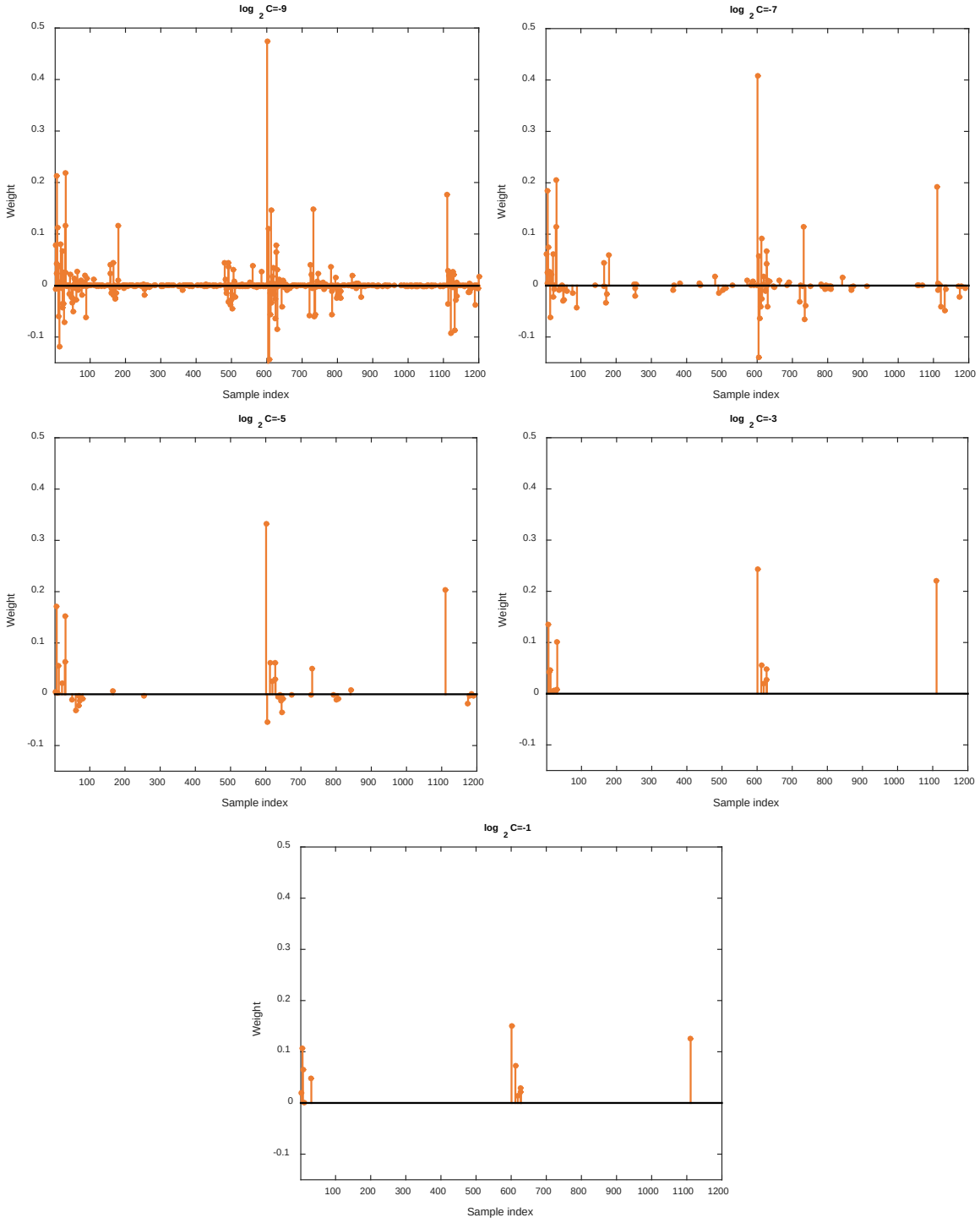

Figure 10. Variation of sparsity w.r.t. different values of *C*.

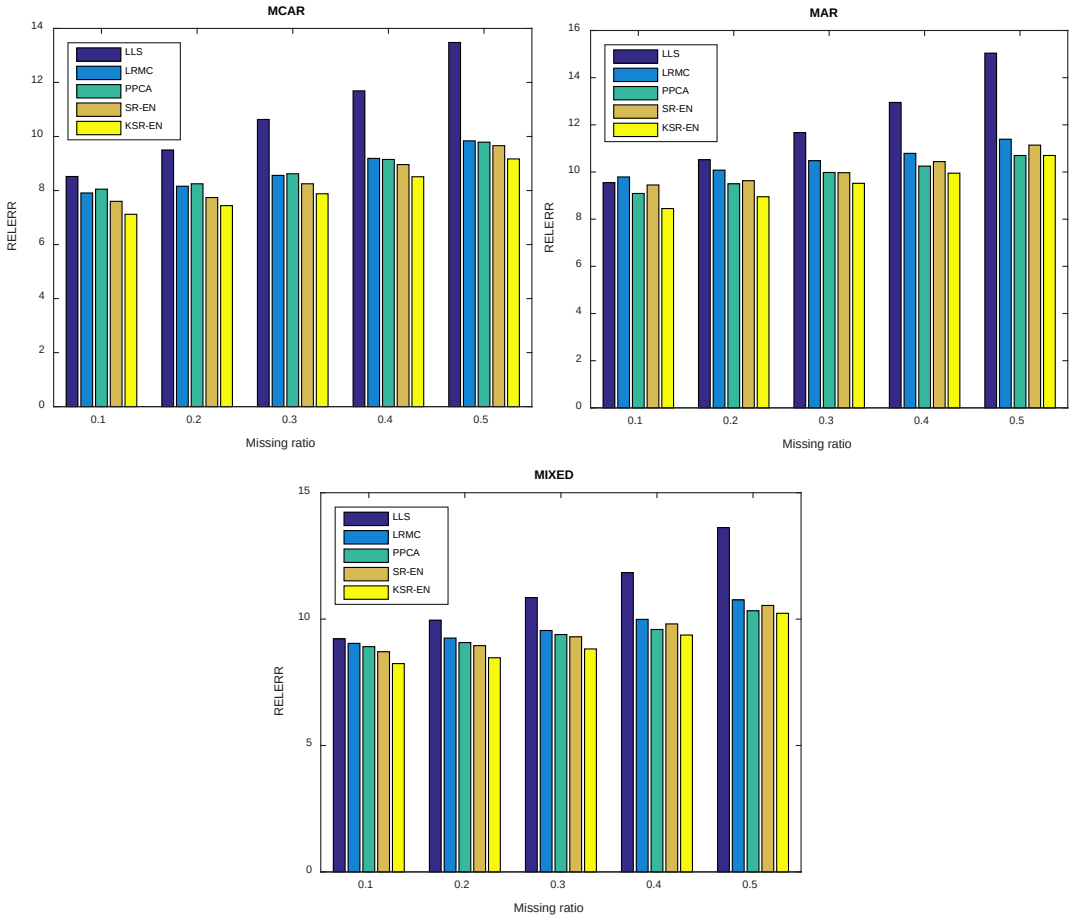

Figure 11. Relative error of each method with 30 min aggregation time.

4.3.4. Computational Time

In this section, we report the computational time of different approaches employed in this paper. We take MIXED missing pattern as it is a combination of MCAR and MAR cases. The experimental results are shown in Table 5 as follows. As we can see, the proposed KSR-EN performs slower than other traditional approaches. It is well-known that sparse representation is a time-consuming procedure especially when the number of samples is large. In addition, gradient descent generally leads to slow convergence. To deal with this problem and improve the efficiency of our approach, several strategies can be exploited. For example, we can approximately solve sparse representation problem through greedy algorithms, such as matching pursuit (MP) or orthogonal MP (OMP) [53], etc. Alternatively, quasi-Newton approach can be used to replace the gradient descent, owing to its affordable memory request but fast convergence. We will concentrate on improving the efficiency of this model in the future work.

Table 5. Computational time(s) of different approaches.

δ	Compared Methods			Ours	
	LLS	**LRMC**	**PPCA**	**SR-EN**	**KSR-EN**
0.1	2.281	3.008	2.815	75.153	178.690
0.2	3.576	4.098	3.520	76.455	192.777
0.3	5.662	5.993	5.325	78.657	223.926
0.4	8.453	8.654	8.138	80.897	249.088
0.5	12.483	12.381	11.961	81.419	274.759

5. Conclusions

The problem of missing traffic sensor data imputation is studied in this work. Conventional sparse representation based imputation may lead to the loss of information conveyed in highly correlated samples. The application to real-world data with complex nonlinear structure is also problematic due to the drawbacks of the linear model. Therefore, we propose the KSR-EN model, which integrates elastic net regularization and the kernel method in a unified framework. In such a way, MVs imputation is performed in high-dimensional feature space rather than original input data space, benefiting accurate imputation. To solve the resulting model, an iterative algorithm is further developed by optimizing the representation coefficients and MVs alternatively. Experiments on both synthetic and traffic sensor data verify that exploiting nonlinear sparse representation, along with the combination of l_1-norm and l_2-norm, can provide more accurate imputation than other competing approaches.

In current work, MVs are estimated from a statistical perspective. Despite reduced estimation errors, the dynamic property of traffic flow is ignored in this work. Many traffic state estimation models have been developed using different techniques, such as extended Kalman filter (EKF) [13], Hamilton-Jacobi equations [12], etc. In future work, we will try to incorporate the traffic state estimation models into our MVs imputation approach to exploit the inherent relationship between variables. Another aspect deserving investigation is the theoretical supporting argument of our proposed model, such as the condition that accurate imputation can be guaranteed. We will focus on the problems in future work.

Author Contributions: X.C. conceived and designed the research; X.C. and C.C. performed the research; X.C., C.C., Y.C., H.W., and Q.Y. wrote and revised the paper.

Funding: This work was partially supported by the National Natural Science Foundation of China (Grant No. 61773184, 51875255, 6187444, U1564201, U1664258, U1762264, 61601203), the National Key Research and Development Program of China (2018YFB0105003), Six talent peaks project of Jiangsu Province (Grant No. 2017-JXQC-007), Key Research and Development Program of Jiangsu Province (BE2016149), Natural Science Foundation of Jiangsu Province (BK2017153), Key Project for the Development of Strategic Emerging Industries of Jiangsu Province (2016-1094, 2015-1084), and the Talent Foundation of Jiangsu University, China (No. 14JDG066).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, J.; Wang, F.-Y.; Wang, K.; Lin, W.-H.; Xu, X.; Chen, C. Data-driven intelligent transportation systems: A survey. *IEEE Trans. Intell. Transp. Syst.* **2011**, *12*, 1624–1639. [CrossRef]
2. Qu, L.; Li, L.; Zhang, Y.; Hu, J. PPCA-based missing data imputation for traffic flow volume: A systematical approach. *IEEE Trans. Intell. Transp. Syst.* **2009**, *10*, 512–522.
3. Turner, S.; Albert, L.; Gajewski, B.; Eisele, W. Archived intelligent transportation system data quality: Preliminary analyses of San Antonio TransGuide data. *Transp. Res. Rec. J. Transp. Res. Board* **2000**, *1719*, 77–84. [CrossRef]
4. Chen, X.; Wei, Z.; Liu, X.; Cai, Y.; Li, Z.; Zhao, F. Spatiotemporal variable and parameter selection using sparse hybrid genetic algorithm for traffic flow forecasting. *Int. J. Distrib. Sens. Netw.* **2017**, *13*. [CrossRef]
5. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
6. Vapnik, V.N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, NY, USA, 2000; 314p.
7. Chen, X.; Yang, J.; Chen, L. An improved robust and sparse twin support vector regression via linear programming. *Soft Comput.* **2014**, *18*, 2335–2348. [CrossRef]
8. Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F.-Y. Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 865–873. [CrossRef]
9. Ma, X.; Dai, Z.; He, Z.; Ma, J.; Wang, Y.; Wang, Y. Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *Sensors* **2017**, *17*, 818. [CrossRef] [PubMed]
10. Yu, H.; Wu, Z.; Wang, S.; Wang, Y.; Ma, X. Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks. *Sensors* **2017**, *17*, 1501. [CrossRef] [PubMed]
11. Work, D.; Blandin, S.; Tossavainen, O.-P.; Piccoli, B.; Bayen, A. A distributed highway velocity model for traffic state reconstruction. *Appl. Res. Math. eXpress* **2010**, *1*, 1–35.

12. Canepa, E.S.; Claudel, C.G. Networked traffic state estimation involving mixed fixed-mobile sensor data using Hamilton-Jacobi equations. *Transp. Res. Part B Methodol.* **2017**, *104*, 686–709. [CrossRef]
13. Wang, Y.; Papageorgiou, M. Real-time freeway traffic state estimation based on extended Kalman filter: A general approach. *Transp. Res. Part B Methodol.* **2005**, *39*, 141–167. [CrossRef]
14. Muñoz, L.; Sun, X.; Horowitz, R.; Alvarez, L. Traffic density estimation with the cell transmission model. In Proceedings of the 2003 American Control Conference, Denver, CO, USA, 4–6 June 2003; pp. 3750–3755.
15. Batista, G.E.; Monard, M.C. An analysis of four missing data treatment methods for supervised learning. *Appl. Artif. Intell.* **2003**, *17*, 519–533. [CrossRef]
16. Zhang, Y.; Liu, Y. Data imputation using least squares support vector machines in urban arterial streets. *IEEE Signal Process. Lett.* **2009**, *16*, 414–417. [CrossRef]
17. Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520–525. [CrossRef] [PubMed]
18. Candès, E.J.; Recht, B. Exact matrix completion via convex optimization. *Found. Comput. Math.* **2009**, *9*, 717–772. [CrossRef]
19. Chen, X.; Wei, Z.; Li, Z.; Liang, J.; Cai, Y.; Zhang, B. Ensemble Correlation-Based Low-rank Matrix Completion with Applications to Traffic Data Imputation. *Knowl.-Based Syst.* **2017**, *132*, 249–262. [CrossRef]
20. Fan, J.; Chow, T. Matrix completion by least-square, low-rank, and sparse self-representations. *Pattern Recognit.* **2017**, *71*, 290–305. [CrossRef]
21. Elhamifar, E.; Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2765–2781. [CrossRef] [PubMed]
22. Vidal, R.; Favaro, P. Low rank subspace clustering (LRSC). *Pattern Recognit. Lett.* **2014**, *43*, 47–61. [CrossRef]
23. Wright, J.; Yang, A.Y.; Ganesh, A.; Sastry, S.S.; Ma, Y. Robust face recognition via sparse representation. *Pattern Anal. Mach. Intell. IEEE Trans.* **2009**, *31*, 210–227. [CrossRef] [PubMed]
24. Chen, X.; Cai, Y.; Liu, Q.; Chen, L. Nonconvex l_p-Norm Regularized Sparse Self-Representation for Traffic Sensor Data Recovery. *IEEE Access* **2018**, *6*, 24279–24290. [CrossRef]
25. Chen, X.; Cai, Y.; Ye, Q.; Chen, L.; Li, Z. Graph regularized local self-representation for missing value imputation with applications to on-road traffic sensor data. *Neurocomputing* **2018**, *303*, 47–59. [CrossRef]
26. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]
27. Bian, X.; Li, F.; Ning, X. Kernelized Sparse Self-Representation for Clustering and Recommendation. In Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, FL, USA, 5–7 May 2016; pp. 10–17.
28. Gao, S.; Tsang, I.W.-H.; Chia, L.-T. Kernel sparse representation for image classification and face recognition. In *Computer Vision—ECCV 2010*; Springer: New York, NY, USA, 2010; pp. 1–14.
29. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, UK, 2004.
30. Thiagarajan, J.J.; Ramamurthy, K.N.; Spanias, A. Multiple kernel sparse representations for supervised and unsupervised learning. *IEEE Trans. Image Process.* **2014**, *23*, 2905–2915. [CrossRef] [PubMed]
31. Combettes, P.L.; Wajs, V.R. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.* **2005**, *4*, 1168–1200. [CrossRef]
32. Bach, F.; Jenatton, R.; Mairal, J.; Obozinski, G. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.* **2012**, *4*, 1–106. [CrossRef]
33. Beck, A.; Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2009**, *2*, 183–202. [CrossRef]
34. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B* **1999**, *61*, 611–622. [CrossRef]
35. Jolliffe, I. *Principal Component Analysis*; Wiley Online Library: New York, NY, USA, 2005.
36. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38.
37. Shi, F.; Zhang, D.; Chen, J.; Karimi, H.R. Missing value estimation for microarray data by Bayesian principal component analysis and iterative local least squares. *Math. Prob. Eng.* **2013**, *2013*, 162938. [CrossRef]
38. Zhang, Y.; Zhou, G.; Jin, J.; Zhao, Q.; Wang, X.; Cichocki, A. Sparse Bayesian classification of EEG for brain-computer interface. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *27*, 2256–2267. [CrossRef] [PubMed]

39. Tan, H.; Feng, G.; Feng, J.; Wang, W.; Zhang, Y.-J.; Li, F. A tensor-based method for missing traffic data completion. *Transp. Res. Part C Emerg. Technol.* **2013**, *28*, 15–27. [CrossRef]
40. Honghai, F.; Guoshun, C.; Cheng, Y.; Bingru, Y.; Yumei, C. A SVM regression based approach to filling in missing values. In Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Melbourne, Australia, 14–16 September 2005; Springer: New York, NY, USA, 2005; pp. 581–587.
41. Chen, X.; Yang, J.; Ye, Q.; Liang, J. Recursive projection twin support vector machine via within-class variance minimization. *Pattern Recognit.* **2011**, *44*, 2643–2655. [CrossRef]
42. Silva-Ramírez, E.-L.; Pino-Mejías, R.; López-Coello, M.; Cubiles-de-la-Vega, M.-D. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Netw.* **2011**, *24*, 121–129. [CrossRef] [PubMed]
43. Kim, H.; Golub, G.H.; Park, H. Missing value estimation for DNA microarray gene expression data: Local least squares imputation. *Bioinformatics* **2005**, *21*, 187–198. [CrossRef] [PubMed]
44. Cai, J.-F.; Candès, E.J.; Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **2010**, *20*, 1956–1982. [CrossRef]
45. Ma, S.; Goldfarb, D.; Chen, L. Fixed point and Bregman iterative methods for matrix rank minimization. *Math. Program.* **2011**, *128*, 321–353. [CrossRef]
46. Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **2011**, *3*, 1–122. [CrossRef]
47. Patel, V.M.; Vidal, R. Kernel sparse subspace clustering. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 2849–2853.
48. Wen, X.; Qiao, L.; Ma, S.; Liu, W.; Cheng, H. Sparse subspace clustering for incomplete images. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 19–27.
49. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge University Press: Cambridge, UK, 2012.
50. Schölkopf, B.; Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2002; 626p.
51. Asif, M.T.; Mitrovic, N.; Dauwels, J.; Jaillet, P. Matrix and Tensor Based Methods for Missing Data Estimation in Large Traffic Networks. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 1816–1825. [CrossRef]
52. Bishop, C. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
53. Pati, Y.C.; Rezaiifar, R.; Krishnaprasad, P.S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In Proceedings of the 1993 Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 1–3 November 1993; pp. 40–44.