# An Efficient Data Characterization and Reduction Scheme for Smart Metering Infrastructure

Sharda Tripathi and Swades De

***Abstract*—In this work, a novel characterization of smart meter data based on Gaussian mixture (GM) model is presented. It is shown that compared to the existing characterization models, the proposed GM model provides a significantly better fit for smart meter data. Further, at each smart meter, sparsity of data is exploited to devise an adaptive data reduction algorithm using compressive sampling technique such that the bandwidth requirement for smart meter data transmission is reduced with minimum loss of information. When compared to the closest competitive data compression scheme, besides being more robust to noise in transmission channel, the proposed compressive sampling based data reduction algorithm offers about $12.8\%$ and $7.4\%$ higher bandwidth saving respectively at $1$ second and $30$ seconds sampling intervals for comparable reconstruction accuracy. Proposed scheme is tested in real-time using RT-LAB.**

***Index Terms*—Smart meter, data compression, data characterization, Gaussian mixture (GM) model, compressive sampling**

## I. Introduction

Advanced metering in smart grid has emerged as a powerful paradigm to enable bi-directional information flow between utility and consumers in the electricity distribution network [1]-[2]. Unlike their analog counterparts, smart meters follow a rapid and automated data logging approach to generate loads of fine grained electricity consumption data. Though smart metering is useful for understanding and modeling of energy usage patterns, efficient communication and storage of this massive data remains a challenge. Besides, another hitch of high resolution smart meter data compared to smooth averaged load profiles is erratic load patterns. The load patterns vary considerably not only for a single user, but also across different users over a time frame. Thus, for extracting information and comprehending consumer behavior in a big data-resource constrained communication scenario, effective characterization as well as reduction of data at granular level are essential.

### A. Related Works and Motivation

In view of limitations in handling big data, strategies for smart meter data reduction have lately attained considerable research interest. These include data compression algorithms operating at appliance level as well as household level with high/ low data resolution at granular/ aggregate collection points in a distribution network. Aggregate level data compression approaches report highest compression ratios, thereby pruning data volume and enhancing energy efficiency of the transmission network. A recent such study in [3] based on singular value decomposition has proposed compression of large data sets transmitted from data collector to control center. In [4], generalized extreme value characterization has been proposed to identify load features in smart meter data which are then stored efficiently in compressed format. Other data characterization models for load profiling considered various distributions, namely, linear, Gaussian, exponential, log-normal, generalized Pareto, beta, and gamma [4], [5]. A similar study on low resolution aggregate level smart meter data is based upon dictionary learning and sparse encoding [6]. It decomposes the load profile into partial usage patterns so as to carefully preserve all required information. In [7], smart meter readings are represented as Gaussian waveforms with minimum features and burrow-wheeler transform, and entropy encoding is applied for its compression. More lossy and loss-less data compression methods for electric signal waveforms are listed in [8]. It is notable that the resolution of data considered in these studies are on the order of one sample per several minutes, whereas typical smart meter readings are on half-hourly basis and collected by the utility only once a day [9]. An algorithm working at aggregation points such as data collector and control center in Fig.1 has access to day-long / week-long data chunks at once (depending upon data collection frequency) from several smart meters. This aids in identifying daily, weekly, seasonal, or behavioral patterns in the data, and exploiting them to achieve data compression becomes relatively easy, since aggregated data at the collector from several smart meters could be huge.

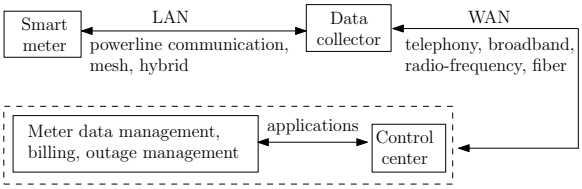


Fig. 1: Smart metering framework.

In recent studies it has been noted that increasing the resolution of smart meter data makes it more useful for near-real time applications, like energy feedback [10], demand response [11], dynamic pricing [12], load monitoring [13], and short-term load forecasting [14]. Accordingly, modern-day smart metering framework is capable of supporting capture of energy consumption data at a rate as high as 1 sample per second which generates a huge volume of data even in smart meter to data collector stage of network communica-

S. Tripathi and S. De are with the Department of Electrical Engineering and Bharti School of Telecommunication, Indian Institute of Technology Delhi, New Delhi, India (e-mail: {sharda.tripathi, swadesd}@ee.iitd.ac.in).

tion. A lossy compression method [15] produces piece-wise approximation of original data to control the smart meter data volume at the cost of accurate data reproducibility. To address this issue, performance of 4 loss-less compression algorithms (adaptive trimmed Huffman, adaptive Markov chain Huffman, tiny Lempel Ziv Markov chain, and Lempel Ziv Markov chain Huffman) was investigated in [16], [17]. A resumable compression method [18] based on differential coding also proposes compression of data sampled at 1 second interval for household level and 3 seconds interval for appliance level at the smart meter. In [19], data granularity and decimal precision of the afore-mentioned approaches was examined. It was stated that, despite their fairly well performance on the appliance level data, with coarse granularity, especially with decimal precision exceeding 2, these tend to become less effective. Besides, most differential coding based compression techniques are sensitive to small consecutive value differences in smart meter data. Their compression performance is expected to degrade with increasing sampling interval and presence of corrupted samples in data transmission/collection process. It has been observed that, although high granularity (on the order of seconds) is critical to attain substantial compressive gains, this may not be the complete picture. Consequently, compression of high resolution household level data at the smart meter remains practically challenging owing to spiky and rapidly fluctuating load patterns.

To this end, considering high resolution data at the smart meter, in this work the problem of smart meter data characterization and reduction is revisited with a perspective to achieve higher compression gains and reduce bandwidth requirement for data transmission from smart meter to the data collector.

### B. Main Contributions

In this work, a new model for characterization of smart meter data is proposed, followed by an adaptive data reduction algorithm for bandwidth saving between smart meter and the data collector. Main contributions of this work are as follows:

1) A novel Gaussian mixture based model is proposed for the characterization of high frequency smart meter data, which is used in evaluating the quality of data reduction at the smart meter. Compared to the existing models, the proposed Gaussian mixture model is shown to have a significantly better fit.
2) An adaptive data reduction scheme using compressive sampling is devised to operate at the smart meter which achieves about 40% bandwidth saving in data transmission to the nearest collection center without any appreciable loss of information.
3) Based on extensive simulations using open datasets as well as real smart meter readings, data update interval for high frequency smart meter data is empirically estimated.
4) Performance comparison of the proposed data reduction scheme with an existing competitive approach in [18] demonstrates noise robustness during data transmission. Additionally, to achieve the same order of RMSE, bandwidth saving with the proposed scheme is found to be 12.8% and 7.4% higher, respectively, for data sampled at 1 second and 30 seconds.

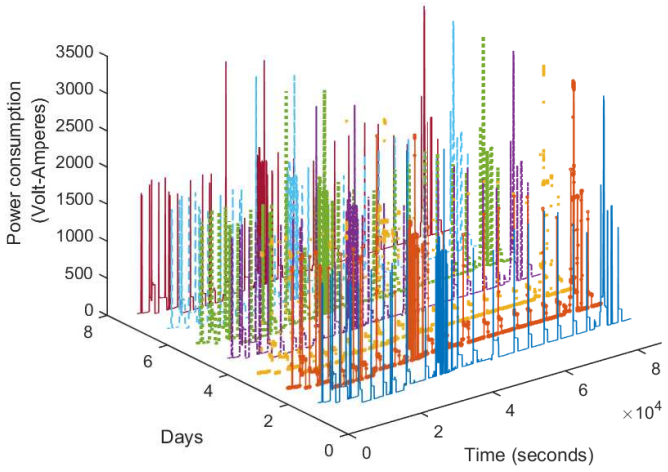


Fig. 2: Daily power consumption of house 1 for 7 days.

5) Online implementation of the proposed system level design on Simulink is tested in real-time using RT-LAB on real smart meter data.

Unlike other data compression algorithms, the proposed compressive sampling based data reduction scheme exploits the inherent rapidly fluctuating nature of high resolution smart meter data. It is observed that, although the data appears incoherent in time domain, it can actually be concisely represented in a sparsifying basis. Thus, adaptively choosing the sparsity over optimum batch size before data transmission can be utilized for substantial reduction in data volume. To the best of the authors' knowledge, exploiting temporal data stochasticity at the smart meter to reduce the volume of transmitted samples without compromising on reconstruction accuracy has not been studied so far.

### C. Paper Organization

Layout of the paper is as follows: Section II briefly describes the datasets used in this study, followed by a discussion on GM model technique for characterization of high frequency smart meter data. In Section III, adaptive compressing sampling algorithm for reduction of smart meter data is proposed. Numerical results based on large scale simulations are discussed in Section IV. Finally, the paper is concluded in Section V.

## II. Characterization of Smart Meter Data

### A. Dataset

In this work, Reference Energy Disaggregation Dataset (REDD) published by Massachusetts Institute of Technology [20] is used to perform simulations. This is an openly available dataset containing detailed power usage information from 6 houses. There are a total of 116 load profiles from all houses consisting of averaged as well as circuit-wise breakup of power consumption of each house. Specifically, the working set in this study is based on the averaged meter readings. These are labeled as "mains" in the dataset and logged in

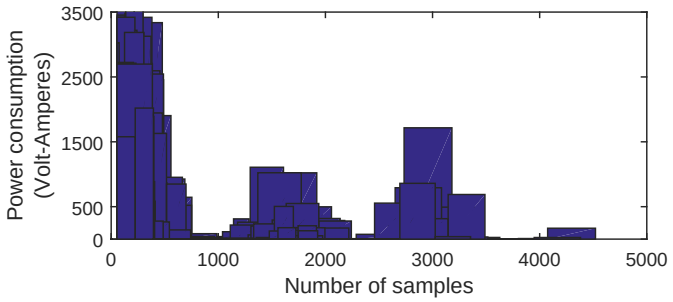


Fig. 3: Power consumption distribution across 7 days for house 1.

at a frequency of 1 sample per second. With such sampling rate, 1 day smart meter reading comprises of 86400 samples. Further, 2 mains outlet from house 1 have meter readings of around 18 days each. Likewise, for houses $2, 3, 4, 5$, and 6, the counts are $13, 16, 19, 3$, and 10 days, respectively. Thus, a total of 158 daily load profiles are considered. Another dataset used in the current study is based on available real smart meter data collected at 30 seconds sampling interval. For this dataset collection, 3 floors of a residential building have been equipped with EM6400 metering device to capture daily power consumption of the households. For statistical consistency, 6 such datasets are investigated in this work.

### B. Data characterization model

Smart meter data is vulnerable to noise in the form of very high usage, short interval spikes (less than 5 seconds) due to aberrant device behavior, or user mistakes in device operation. In order to eliminate this artifact, the spikes are replaced with extrapolated neighboring values. Fig. 2 shows the daily power consumption of a household for 7 days. Histogram of daily power consumption profile is plotted in Fig. 3. It can be observed from Fig. 3 that probability distribution function consists of multiple Gaussian components. This leads to our intuition of GM model for the characterization of smart meter data. This data characterization elaborates the structural features of energy consumption data and is in general useful for synthetic smart meter data generation. In this work, the developed GM model is used to assess the performance of the proposed adaptive data reduction scheme (presented in next section) to evaluate that the structural features are preserved during compression and transmission of smart meter data.

GM model [21] is a weighted superimposition of multiple Gaussian components. If $N$ i.i.d sample points $x = \{x_1, x_2, \cdots, x_N\}$ are observed, then a $k$-component GM model is expressed as:

$$f_k(x) = \sum_{j=1}^{k} w_j \mathcal{N}(x|\mu_j, \sigma_j), \text{with} \quad w_j \geq 0 \quad \text{and} \quad \sum_{j=1}^{k} w_j = 1, \tag{1}$$

where $\mathcal{N}(x|\mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}}$ is the $j$th Gaussian component and $w_j$ is the mixing coefficient corresponding to each Gaussian component. To specify the GM model, optimal value of parameters $k$ and $\mu_j$, $\sigma_j$ and $w_j$, $\forall j \in \{1, \cdots, k\}$ are obtained such that the likelihood of observed data given the model parameters is maximized. The log likelihood function is defined as:

$$L(x, f_k) \triangleq L_k = \sum_{i=1}^{N} \ln \left\{ \sum_{j=1}^{k} w_j \mathcal{N}(x|\mu_j, \sigma_j) \right\}$$

### C. Model Parameter Estimation

Optimal parameters $w_j$, $\mu_j$, and $\sigma_j$ are estimated using Expectation-Maximization (EM) algorithm [22]. The algorithm works iteratively such that maximum likelihood of observed data increases with each subsequent iteration, thereby converging to a saddle point. Each iteration consists of E and M steps. For the start of algorithm, parameters $w_j$, $\mu_j$, and $\sigma_j$ are initialized respectively to $1/k$, 0, and 1. During E-step, the posterior probability $p(j|x_i)$ for each GM component $j$ is evaluated corresponding to every data point $x_i$ using the initialized parameters. Thereafter, in the M-step, posterior probabilities obtained in E-step are maximized with respect to the model parameters to obtain updated values of $w_j$, $\mu_j$, and $\sigma_j$. This process is repeated unless the difference between old and new likelihood estimates falls below an acceptable error margin which is taken as $10^{-6}$ in this work.

To decide the optimal number of Gaussian components $k$ of the GM model, Hellinger's distance [23] is used as a measure of goodness of fit. This metric quantifies the similarity between two probability distributions and assumes values in $[0, 1]$ such that a 0 is indicative of perfect similarity. As this value increases towards 1, statistical properties of the two distributions begin to deviate and a 1 signifies complete discrepancy between them. The idea is to choose $k$ such that the value of Hellinger's distance falls below an acceptable threshold of 0.05 [24]. For discrete probability distributions $P = \{p_1, p_2, \cdots, p_n\}$ and $Q = \{q_1, q_2, \cdots, q_n\}$, Hellinger's distance between them is defined as:

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n} (\sqrt{p}_i - \sqrt{q}_i)^2}. \tag{2}$$

## III. Data reduction using Adaptive Compressive Sampling

In this section, an adaptive data reduction scheme based on compressive sampling is presented. The scheme is proposed to operate at the smart meter in order to reduce the bandwidth requirement for transmission of individual smart meter data while preserving the characteristics obtained in Section II.

Compressive sampling technique [25] exploits the sparsity in a dataset to reconstruct the compressed signal from far fewer samples than required by Nyquist sampling theorem. Let $n$ be the length of sample window over which data transmission takes place. If $x = \{x_1, x_2, \cdots, x_n\}$ be the samples in the data window, then $x$ can be expressed as:

$$x = \psi f \tag{3}$$

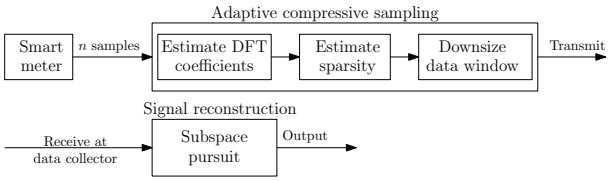


Fig. 4: Adaptive compressive sampling for smart meter data.

where $\psi$ is the sparse basis matrix of size $n \times n$ and $f$ is the column vector of coefficients corresponding to $\psi$. Only $m$ ($m \ll n$) samples out of $n$ in a data window are randomly chosen for transmission. This downsizing is performed in the interest of bandwidth saving over the communication channel. Using (3), the transmitted samples are denoted by,

$$y = \phi x = \phi \psi f \tag{4}$$

where $\phi$ is an $m \times n$ sensing matrix. Accurate signal reconstruction from $m$ samples map to the problem of solving an underdetermined system of linear equations. In this work, subspace pursuit algorithm [26] has been used to recover the signal at the data collector. Discrete Fourier transform (DFT) and random Gaussian matrix are chosen as sparse basis and sensing matrix, respectively, such that incoherence and restricted isometry property are satisfied for successful signal reconstruction [25]. Fig. 4 shows block diagram of the proposed adaptive compressive sampling scheme for smart meter data reduction. Unlike the conventional compressive sampling, where sparsity is known a priori and remains constant throughout the execution of algorithm, in the proposed scheme, sparsity $s$ is decided for each data window by estimating the number of DFT coefficients containing 99.99% energy of samples in the data window. It helps to capture the rapidly varying behavior of smart meter data and is essential in reducing the count of transmitted samples without compromising on the reconstruction accuracy in a dynamic environment. Accordingly, $m$ samples to be transmitted are randomly chosen using $m = s \log n$.

## IV. Results

In this section, first the model selection in $k$-GM technique is discussed, followed by a comparison of proposed $k$-GM model with existing data characterization models. Subsequently, empirical estimation of optimum data collection interval and comparative performance analysis with respect to a recent competitive approach in [18] are presented. Finally, real-time testing of Simulink based online implementation of the system model is discussed for assessing feasibility of the proposed adaptive compressive sampling algorithm.

### A. Model Selection

As discussed in Section II, Hellinger's distance metric is estimated for different $k$ values in $k$-GM model. Fig. 5 shows the variation of Hellinger's distance with increasing components in the GM model averaged over 7 daily load profiles from house 1. It can be observed that, beyond $k = 4$ Hellinger's distance consistently hovers below the threshold

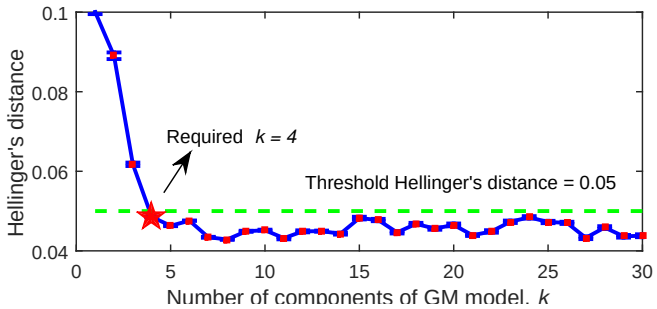


Fig. 5: Model selection using Hellinger's distance.

TABLE I: 4-GM model parameter estimates for smart meter data

| $k$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\mu_j$ (VA) | 58.053 | 131.50 | 291.20 | 1783.6 |
| $\sigma_j$ (VA) | 5.2967 | 106.2834 | $8.001 \times 10^3$ | $1.221 \times 10^5$ |
| $w_j$ | 0.098 | 0.529 | 0.34 | 0.033 |

TABLE II: Variation of Hellinger's distance across houses and days for the test load profiles.

| Averaged over days | | Averaged over houses | |
|---|---|---|---|
| House | Hellinger's distance | Day | Hellinger's distance |
| 1 | 0.045 | Monday | 0.0348 |
| 2 | 0.045 | Tuesday | 0.0379 |
| 3 | 0.035 | Wednesday | 0.0340 |
| 4 | 0.038 | Thursday | 0.0344 |
| 5 | 0.048 | Friday | 0.0314 |
| 6 | 0.031 | Saturday | 0.0342 |
| | | Sunday | 0.0359 |

and thus gain in model fitness is considerably small. Thus, $k = 4$ is chosen as the optimal number of Gaussian components in the $k$-GM model. It may be noted that computation complexity of $k$-GM model is $\mathcal{O}(kN^2)$. Hence, a smaller value of $k$ is preferred. Other parameters $\mu_j$, $\sigma_j$, and $w_j$ as obtained from EM algorithm for each of the $k$ components are presented in Table I. As observed from Fig. 2, the individual plots may seem to appear quite similar due to the daily behavioural pattern of consumers in a house, but they may vary significantly across different houses. Consequently, parameter estimates of the 4-GM characterization model, namely, mean, variance, and weights of Gaussian mixture components will tend to differ for each house. However, it has been observed in the simulations that, despite varying user behaviour, the proposed 4-GM model for characterization of high frequency smart meter data remains valid. This is elaborated via Table II, where the Hellinger's distance between the empirical and 4-GM modelled probability density functions for different houses averaged over all days and for different days averaged over all houses using 151 test load profiles is presented. It can be observed that the test load profiles support 4-GM characterization of high frequency smart meter data with an average Hellinger's distance of 0.0404 for different houses and 0.0346 for different days, both of which are within the acceptable limit [24].

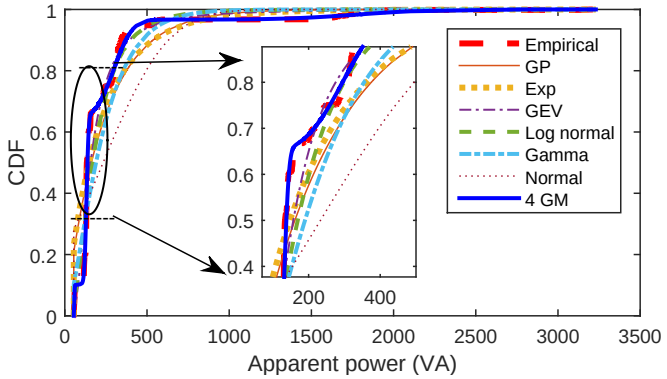


Fig. 6: CDF plot for different GM components.

TABLE III: Hellinger's distance for various models against empirical distribution

| Distribution fits | Hellinger's distance |
|---|---|
| Normal | 0.0872 |
| Exponential | 0.0866 |
| Generalized Pareto (GP) | 0.0866 |
| Gamma | 0.0832 |
| Log normal | 0.0803 |
| Generalized extreme value (GEV) | 0.0784 |
| 2 GM model | 0.0725 |
| 3 GM model | 0.0446 |
| 4 GM model | 0.0379 |
| 5 GM model | 0.0373 |
| 6 GM model | 0.0370 |

### B. Comparison with State-of-art

CDF of 4-component Gaussian mixtures is compared with the existing data characterization models against the empirical CDF in Fig. 6. It is clearly visible that, CDF of the 4-GM model closely follows the empirical CDF. The corresponding Hellinger's distances are presented in Table III. The metric values are noted to be higher and above the acceptable threshold of 0.05 in case of the existing data characterization models: normal, exponential, generalized Pareto, gamma, log normal, and generalized extreme value, thus indicating a poorer fit in comparison with the GM model with $k \geq 4$. Additionally, Hellinger's distance remains fairly constant up to 3 decimal places for $k \geq 4$, thus validating the choice of $k = 4$. This study is repeated on 5 other household datasets at different time spans and the same conclusion is found to be true. *Thus, daily power consumption data at 1 Hz sampling frequency by the smart meter can be reasonably characterized by a 4-component GM model.* This data characterization will be used to evaluate the quality of data reduction at the smart meter.

### C. Optimum Data Update Interval Estimation

Signal reconstruction error is measured in terms of root mean square error (RMSE) at the data collector. Fig. 7 presents the variation of bandwidth saving and RMSE with number of samples in the data window $n$. Here, $(n-m)/n$ is used as a measure of bandwidth saving. It can be observed that both RMSE and bandwidth saving reduce with increasing values of $n$, thus there exists a trade-off between them. Beyond an

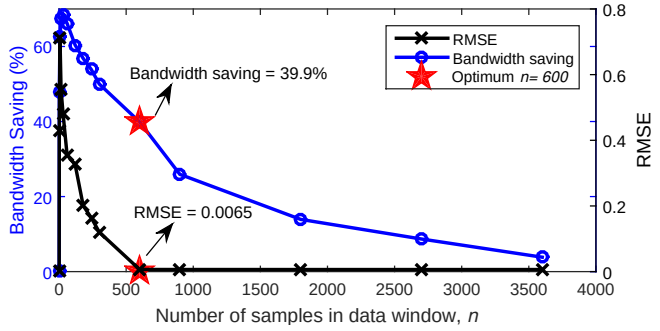


Fig. 7: Variation of bandwidth saving and RMSE with number of samples in data window.

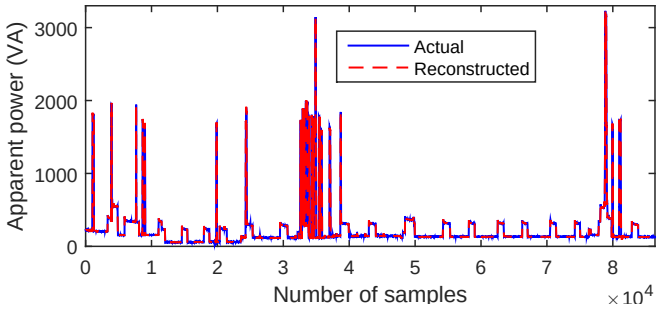


Fig. 8: Reconstructed data for 10 minutes interval versus actual data.

optimum $n$, RMSE nearly saturates. Hence, this is the required window size $n$ over which data samples must be periodically transmitted to the data collector. From the plots in Fig. 7, it can be observed that for $n \geq 600$, further reduction in RMSE is negligibly small. Consequently, the optimum $n$ is selected to be 600 samples. The corresponding bandwidth saving is $39.9\%$. Since the data considered in this study is sampled at 1 Hz, $n = 600$ samples correspond to a data collection interval of 10 minutes. *Thus, by applying adaptive compressive sampling and updating data at the collector every* 10 *minutes, about* $40\%$ *reduction in bandwidth resource requirement can be achieved in transmission of data from each smart meter.*

### D. Performance of Proposed Adaptive Compressive Sampling Algorithm

Signal reconstruction for optimum data collection interval against the actual meter readings is plotted in Fig. 8. It can be observed that reconstructed data closely follows the actual data owing to a low RMSE of 0.0065. To further validate signal reconstruction accuracy, the reconstructed data is characterized using 4-GM model as proposed in Section II. Hellinger's distance metric between empirical and reconstructed smart meter data is found to be 0.0398. From Fig. 9 it can be observed that, CDFs of the actual empirical data versus the data modeled using 4-GM model and reconstructed data modeled using 4-GM model are very closely matched. Table IV presents the parameter estimates of 4-GM model for the reconstructed smart meter data. A comparison of 4-GM parameter estimates from Table I and Table IV reveals that the structural features of data at the smart meter before compression are restored after

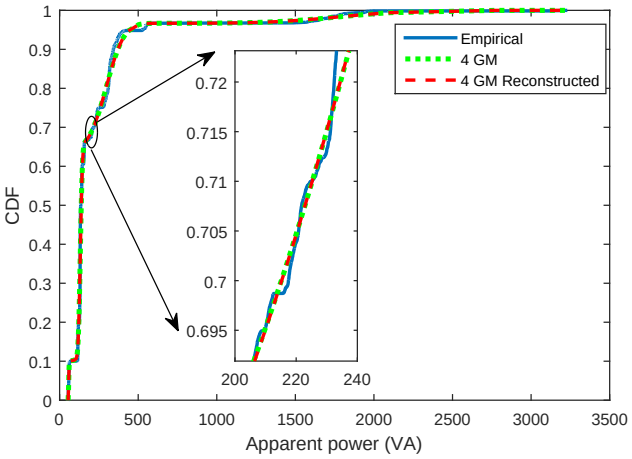


Fig. 9: Comparison of CDFs of empirical data versus 4-GM modeled data and 4-GM modeled reconstructed data over 10 mins.

TABLE IV: 4-GM model parameter estimates for reconstructed smart meter data.

| $k$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\hat{\mu}_j$ (VA) | 58 | 131.9 | 297.3 | 1782.9 |
| $\hat{\sigma}_j$ (VA) | 5.5633 | 106.4793 | $8.081 \times 10^3$ | $1.221 \times 10^5$ |
| $\hat{w}_j$ | 0.0991 | 0.5421 | 0.3257 | 0.0331 |

data reconstruction. *Thus, bandwidth saving is achieved with minimal information loss in the data compression process.*

It may be noted that although RMSE tends to be small, sometimes maximum difference between actual samples and reconstructed samples could be large, especially in the data windows having more number of spikes. Fig.10 captures the statistics of maximum and minimum reconstruction error per sample averaged over daily load profiles for each house in REDD dataset. It can be observed that maximum and minimum reconstruction error lie in the order of $10^1$ and $10^{-5}$, respectively. It has been found in the simulations that, although the absolute values of maximum reconstruction error are high, with respect to actual values it corresponds to less than $4\%$ error across all data sets. Additionally, occurrence of samples with large error (order of $10^1$) in daily load profile is below $0.5\%$ in the datasets considered in this study. Consequently, reconstruction error is very small for most of the data samples, hence an error order of $10^1$ corresponding to very few samples do not alter the structural features of dataset. It has been validated via simulations that for every household dataset Hellinger's distance is under the acceptable threshold.

### *E. Comparative Performance Analysis*

In this section, performance of the proposed adaptive compressive sampling technique is compared with the competitive resumable load data compression approach based on entropy coding [18], which also aims at bandwidth reduction between smart meter and data collector in the electricity distribution network. In [18], REDD dataset with 1 second sampling interval is used to demonstrate loss-less compression of high frequency smart meter data. A comparison of bandwidth saving between adaptive compressive sampling and resumable data compression using REDD data set is presented in Fig. 11. Note that these meter readings have a decimal precision of 2. Here bandwidth saving is a measure of relative difference between actual and compressed batch size before transmission without including any overhead. It can be observed that, although resumable data compression is loss-less, adaptive compressive sampling outperforms in bandwidth reduction with an acceptable reconstruction error for smaller data collection intervals. However, with increasing interval size, adaptive compressive sampling requires more samples for data reconstruction in order to maintain low RMSE, thus reducing the bandwidth saving. At the optimum data collection interval of 10 minutes, adaptive compressive sampling saves $23.7\%$ more bandwidth compared to resumable data compression.

It has been observed that, if an input dataset with higher precision values is used, then simultaneously attaining high compressive gains and lossless reconstruction with resumable data compression is not possible. It causes either bandwidth savings to diminish in order to maintain the precision, or a small reconstruction error is introduced due to truncation of meter readings to smaller precision. In Table V, performance of adaptive compressive sampling and resumable data compression is compared for real smart meter readings with

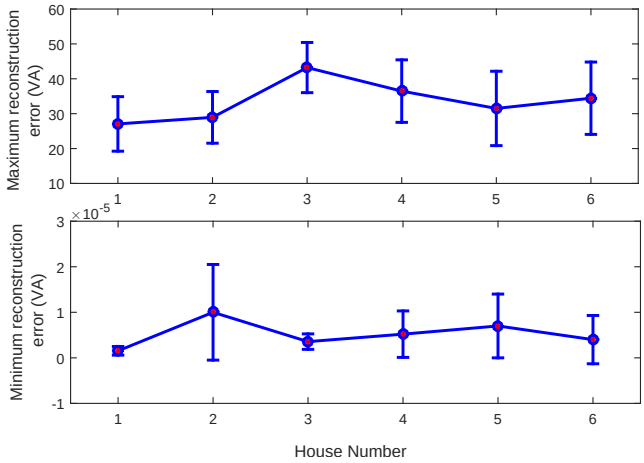


Fig. 10: Maximum and minimum reconstruction error for all houses.

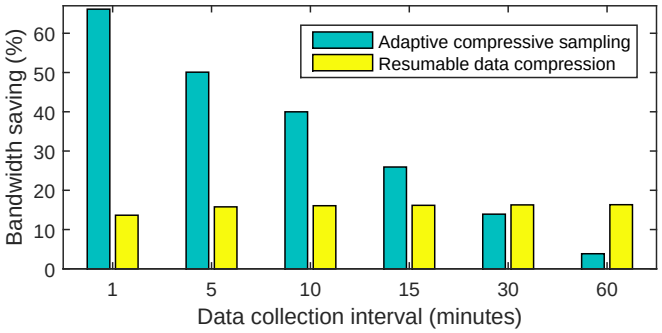


Fig. 11: Performance comparison of adaptive compressive sampling and the resumable data compression [18] at different data collection intervals, with samples collected at 1 sample/sec.

TABLE V: Performance comparison of adaptive compressive sampling and resumable data compression at 30 second sampling interval for different datasets.

| Dataset | Adaptive compressive sampling | | Resumable data compression |
|---|---|---|---|
| | RMSE | Bandwidth saving | Bandwidth saving |
| #1 | 0.0277 | 22.63% | -3.35% |
| #2 | 0.0574 | 5.75% | -5.35% |
| #3 | 0.0598 | 27.79% | 0.8% |
| #4 | 0.0683 | 16.58% | -4.17% |
| #5 | 0.0611 | 16.88% | -9.8% |
| #6 | 0.0437 | 27.58% | 4.92% |

TABLE VI: Performance comparison of adaptive compressive sampling and resumable data compression with varying decimal precision of input dataset at 1 second sampling interval.

| Decimal precision | Adaptive compressive sampling | | Resumable data compression | |
|---|---|---|---|---|
| | RMSE | Bandwidth saving | RMSE | Bandwidth saving |
| 1 | 0.0065 | 39.9% | $1.2\times10^{-3}$ | 27.13% |
| 2 | 0.0063 | 39.9% | 0 | 16.11% |

TABLE VII: Performance comparison of adaptive compressive sampling and resumable data compression with varying decimal precision of input dataset at 30 seconds sampling interval.

| Decimal precision | Adaptive compressive sampling | | Resumable data compression | |
|---|---|---|---|---|
| | RMSE | Bandwidth saving | RMSE | Bandwidth saving |
| 0 | 0.0326 | 26.73% | $2.8\times10^{-2}$ | 19.3% |
| 1 | 0.0326 | 26.73% | $6.5\times10^{-3}$ | 12.9% |
| 2 | 0.0330 | 26.73% | $4.8\times10^{-4}$ | 2.4% |
| 3 | 0.0328 | 26.73% | $5.9\times10^{-5}$ | -6.3% |
| 4 | 0.0324 | 26.73% | 0 | -9.3% |

decimal precision of 4, and collected at the sampling interval of 30 seconds. It may be noted that resumable data compression is based on very small consecutive value difference in high frequency smart meter data. Increasing sampling interval violates this condition, eventually causing compressed batch size to exceed the actual batch size, which is signified by negative bandwidth saving with resumable data compression in Table V. Consequently, at 30 seconds sampling interval, the proposed adaptive compressive sampling technique saves around $22.4\%$ more bandwidth at the cost of increased RMSE as compared to resumable data compression. Further, on reducing sampling frequency correlation between consecutive samples also reduces, thereby deteriorating compression performance. Accordingly, as compared to 1 second, bandwidth savings for data sampled at 30 seconds interval have a mean reduction of $20.37\%$ and $33.26\%$, respectively, with adaptive compressive sampling and resumable data compression. It is worthwhile to mention here that, the primary objective is to faithfully reconstruct the smart meter data at the collection center, and it can be achieved within certain error tolerance that is characterized by a tolerable Hellinger's distance between

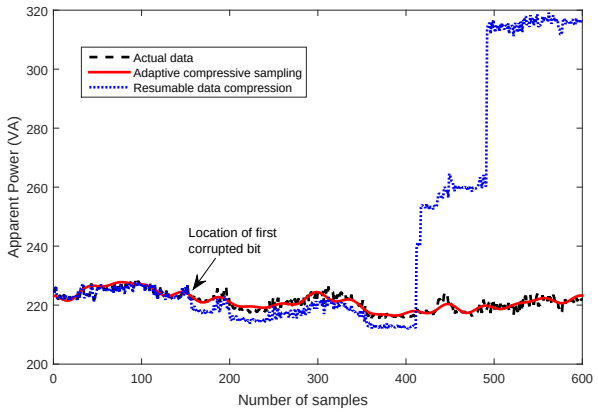


Fig. 12: Data reconstruction with 1% corrupted samples in adaptive compressive sampling and resumable data compression.

reconstructed data distribution and the original distribution [24]. Hence, loss-less reconstruction, as targeted in [18], is not absolutely necessary. *From this viewpoint, performance of the proposed adaptive compressive sampling algorithm can be considered overall superior than resumable data compression because of its much higher bandwidth saving while limiting the error performance to an acceptable value.*

In Table VI and Table VII, variation in performance of adaptive compressive sampling and resumable data compression with increasing precision of input meter readings is presented, respectively, for 1 second and 30 seconds sampling interval. From both the tables, it can be observed that increasing precision of input values does not affect the performance of adaptive compressive sampling, as the bandwidth requirement and reconstruction accuracy of adaptive compressive sampling depends on the sparsity of dataset contained in 10 minutes window size, which remains unaltered with increasing precision of input meter readings. On the contrary, bandwidth saving with resumable data compression reduces with increasing precision due to the compression of additional digits per measurement value. Also, since the original meter readings are precise up to 2 and 4 decimal places, respectively, for 1 second and 30 seconds sampling interval, small reconstruction error is observed at lower precision values. From Tables VI and VII it can be observed that, to achieve same reconstruction accuracy, improvement in bandwidth saving with adaptive compressive sampling over resumable data compression is about $12.8\%$ and $7.4\%$, respectively, for sampling interval of 1 second and 30 seconds. *Thus, unlike resumable data compression, performance of the proposed adaptive compressive sampling algorithm is independent of the precision of input dataset. Additionally, compared to loss-less resumable data compression, adaptive compressive sampling not only has higher bandwidth saving with acceptable RMSE, it also outperforms for same order of reconstruction accuracy in both the algorithms.*

Finally, Fig. 12 compares the reconstruction performance of adaptive compressive sampling and resumable data compression with $1\%$ corrupted samples in the transmission window. Respective RMSEs are found to be $0.046$ and $0.7155$. It can be observed that, in adaptive compressive sampling data is still recoverable although at the cost of reconstruction accuracy,

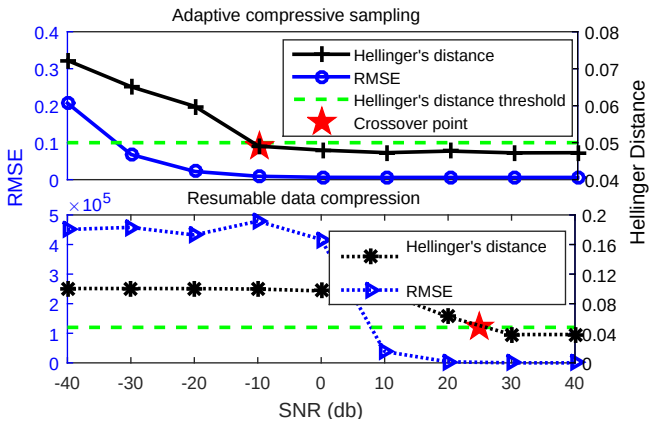


Fig. 13: Variation of RMSE and Hellinger's distance with SNR in adaptive compressive sampling and resumable data compression.

but the error rate with resumable data compression is much higher and there are chances of entire message being lost, especially when the first corrupted sample occurs earlier in the transmission sequence. The effect of channel noise on the reconstruction performance of adaptive compressive sampling and resumable data compression is quantified in Fig. 13 by considering an additive white Gaussian noise channel for data transmission with signal to noise ratio (SNR) varying from $-40$ dB to $+40$ dB. It can be observed that during reconstruction using adaptive compressive sampling, Hellinger's distance between probability density functions of actual and reconstructed data attains acceptable value at minimum SNR of $-10$ dB. This SNR value is marked as crossover to the tolerable noise region. The corresponding RMSE is $0.00628$. In case of resumable data compression, crossover is observed at $+25$ dB SNR. Various error correction techniques can be applied to both the algorithms, however they incur extra cost and complexity to the system. *Thus, it is remarked that noise robustness of the proposed adaptive compressive sampling algorithm is significantly higher, as it can reconstruct the data with reasonable accuracy at SNRs as low as $-10$ dB, whereas, with resumable data compression at least $+25$ dB SNR is required to achieve same level of reconstruction accuracy.*

### *F. Online Implementation*

In this section, system level design and real-time testing of proposed adaptive compressive sampling technique in a smart metering framework is discussed. The schematic of system model designed in Simulink is shown in Fig. 14. It comprises of 4 modules, namely, data sampling and storage, data compression, data transmission and data reconstruction. Data sampling, storage and compression takes place at individual smart meters. While data is sampled at high frequency, a buffer stores this data for optimum data collection interval (10 minutes) before adaptive compressive sampling technique can be applied on this batch. Subsequently, data reconstruction module operates at the aggregation point. This procedure is iterated for successive batches. Since transmission aspects of compressed smart meter data is not in the scope of our current study, perfect channel conditions have been assumed

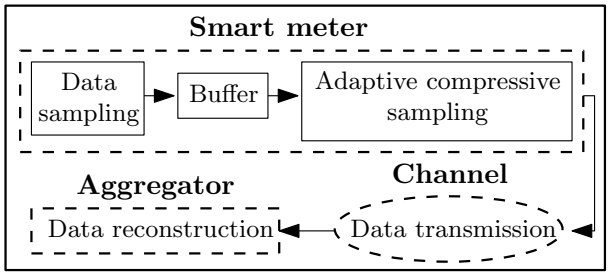


Fig. 14: Simulink implementation schematic of proposed adaptive compressive sampling algorithm.

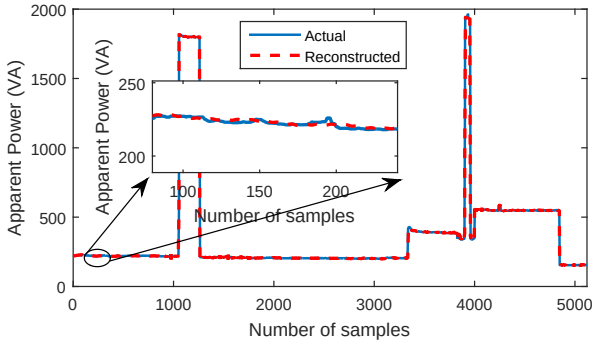


Fig. 15: Performance of Simulink based system design in RT-LAB for adaptive compressive sampling algorithm.

for error free transmission of compressed data in the implementation setup. The performance of the proposed system design on Simulink has been tested in real-time using RT-LAB on real smart meter data. RT-LAB facilitates separation of system model into subsystems, as the metering device and the aggregator, which are then executed on parallel target processors running QNX real-time operating system. For real-time execution of the proposed adaptive compressive sampling algorithm, it is required that the compression, transmission, and reconstruction operations on one batch of smart meter data is completed within the optimum data update interval. Thus, sampling time of the system is fixed at 10 minutes. It may be noted here that, owing to batch processing, sampling time of the proposed system is significantly higher compared to the actual execution time of compressive sampling based algorithms using existing VLSI technologies, which is on the order of micro seconds [27], [28]. In Fig. 15, performance of the proposed system model tested using RT-LAB over 5 batches is shown. Zero missed ticks were observed during runtime, which verify the real-time simulation. It has been observed that the performance results on RT-LAB match very well with the Matlab simulations.

## V. Concluding Remarks

To summarize, this study has sought to strike a balance between network resource requirement and quality of service in data intensive smart grid applications by exploiting the sparsity of data at the smart meter. It has been demonstrated by characterizing the smart meter data using Gaussian mixtures that the proposed data reduction algorithm at the smart meter

can significantly reduce the data volume, thereby saving $40\%$ bandwidth requirement for communication between the smart meter and the data collector, without appreciably affecting the data characteristics. Further, from comparative analysis it is observed that, with respect to resumable data compression technique, the proposed adaptive compressive sampling technique is more robust to noise in transmission channel, and for comparable reconstruction accuracy, its bandwidth saving is $12.8\%$ and $7.4\%$ higher, respectively, for data granularity of 1 second and 30 seconds. Thus, it is expected that, when the proposed smart meter data reduction algorithm works in unison with the data reduction algorithms at the collector, overall bandwidth requirement for communication of data from smart meter to control center will be considerably small.

## REFERENCES


[1] D. Alahakoon and X. Yu, "Smart electricity meter data intelligence for future energy systems: A survey," *IEEE Trans. Ind. Informat.*, vol. 12, no. 1, pp. 425–436, Feb. 2016.

[2] V. C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, and G. P. Hancke, "A survey on smart grid potential applications and communication requirements," *IEEE Trans. Ind. Informat.*, vol. 9, no. 1, pp. 28–42, Feb. 2013.

[3] J. C. S. de Souza, T. M. L. Assis, and B. C. Pal, "Data compression in smart distribution systems via singular value decomposition," *IEEE Trans. Smart Grid*, vol. 8, no. 1, pp. 275–284, Jan. 2017.

[4] X. Tong, C. Kang, and Q. Xia, "Smart metering load data compression based on load feature identification," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2414–2422, Sept. 2016.

[5] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Järventausta, "Enhanced load profiling for residential network customers," *IEEE Trans. Power Del.*, vol. 29, no. 1, pp. 88–96, Feb. 2014.

[6] Y. Wang, Q. Chen, C. Kang, Q. Xia, and M. Luo, "Sparse and redundant representation-based smart meter data compression and pattern extraction," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2142–2151, May 2017.

[7] A. Abuadbba, I. Khalil, and X. Yu, "Gaussian approximation based lossless compression of smart meter readings," *IEEE Trans. Smart Grid*, vol. PP, no. 99, pp. 1–1, 2017.

[8] M. P. Tcheou, L. Lovisolo, M. V. Ribeiro, E. A. B. da Silva, M. A. M. Rodrigues, J. M. T. Romano, and P. S. R. Diniz, "The compression of electric signal waveforms for smart grids: State of the art and future trends," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 291–302, Jan. 2014.

[9] Q. Chen, D. Kaleshi, S. Armour, and Z. Fan, "Reconsidering the smart metering data collection frequency for distribution state estimation," in *IEEE SmartGridComm*, Venice, Italy, Nov. 2014, pp. 517–522.

[10] Y. Liu, C. Yuen, R. Yu, Y. Zhang, and S. Xie, "Queuing-based energy consumption management for heterogeneous residential demands in smart grid," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1650–1659, May 2016.

[11] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *IEEE Trans. Ind. Informat.*, vol. 7, no. 3, pp. 381–388, Aug. 2011.

[12] Q. Tang, K. Yang, D. Zhou, Y. Luo, and F. Yu, "A real-time dynamic pricing algorithm for smart grid with unstable energy providers and malicious users," *IEEE Internet Things J.*, vol. 3, no. 4, pp. 554–562, Aug. 2016.

[13] K. Basu, V. Debusschere, S. Bacha, U. Maulik, and S. Bondyopadhyay, "Nonintrusive load monitoring: A temporal multilabel classification approach," *IEEE Trans. Ind. Informat.*, vol. 11, no. 1, pp. 262–270, Feb. 2015.

[14] Y. H. Hsiao, "Household electricity demand forecast based on context information and user daily schedule analysis from meter data," *IEEE Trans. Ind. Informat.*, vol. 11, no. 1, pp. 33–43, Feb. 2015.

[15] F. Eichinger, P. Efros, S. Karnouskos, and K. BÃ¶hm, "A time-series compression technique and its application to the smart grid," *The VLDB Journal*, vol. 24, no. 2, pp. 193–218, 2015.

[16] M. Ringwelski, C. Renner, A. Reinhardt, A. Weigel, and V. Turau, "The hitchhiker's guide to choosing the compression algorithm for your smart meter data," in *IEEE International Energy Conference and Exhibition (ENERGYCON)*, Sept. 2012, pp. 935–940.

[17] M. Zeinali and J. S. Thompson, "Impact of compression and aggregation in wireless networks on smart meter data," in *IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2016, pp. 1–5.

[18] A. Unterweger and D. Engel, "Resumable load data compression in smart grids," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 919–929, Mar. 2015.

[19] A. Unterweger, D. Engel, and M. Ringwelski, "The effect of data granularity on load data compression," in *4th DACH Conference on Energy Informatics - Volume 9424*. New York, NY, USA: Springer-Verlag, New York, Jan. 2016, pp. 69–80.

[20] J. Z. Kolter and M. J. Johnson, "REDD: A Public Data Set for Energy Disaggregation Research," in *Proc. SustKDD Workshop Data Min. Appl. Sustain.*, San Diego, California, USA, Aug. 2011, pp. 1–6.

[21] D. A. Reynolds, "Gaussian mixture models." in *Encyclopedia of Biometrics*. Springer US, 2009, pp. 659–663.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[23] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002.

[24] L. Pardo, *Statistical Inference Based on Divergence Measures*. CRC Press, 2005.

[25] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[26] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.

[27] A. Septimus and R. Steinberg, "Compressive sampling hardware reconstruction," in *Proc. IEEE International Symposium on Circuits and Systems*, May 2010, pp. 3316–3319.

[28] J. L. V. M. Stanislaus and T. Mohsenin, "Low-complexity fpga implementation of compressive sensing reconstruction," in *Proc. IEEE Intl. Conf. Computing, Networking, and Commun. (ICNC)*, Jan. 2013, pp. 671–675.


**Sharda Tripathi** received her B. Tech. degree in Electronics and Communication Engineering from Rajiv Gandhi Technical University, Bhopal, India, in 2007 and the M.Tech. degree in Digital Communication Engineering from the Department of Electronics and Telecommunication Engineering, Maulana Azad National Institute of Technology, Bhopal, India in 2011. She is currently pursuing the Ph.D. degree in Department of Electrical Engineering, Indian Institute of Technology Delhi, India. Her current research interests include application of machine learning in smart grid communication networks.

**Swades De** (S'02-M'04-SM'14) received his B.Tech. degree in Radiophysics and Electronics from the University of Calcutta, Kolkata, India, in 1993, the M.Tech. degree in Optoelectronics and Optical communication from IIT Delhi, New Delhi, India, in 1998, and the Ph.D. degree in Electrical Engineering from the State University of New York at Buffalo, Buffalo, NY, USA, in 2004.

He is currently a Professor with the Department of Electrical Engineering, IIT Delhi. Before moving to IIT Delhi in 2007, he was a Tenure-Track Assistant Professor with the Department of ECE, New Jersey Institute of Technology, Newark, NJ, USA, from 2004–2007. He worked as an ERCIM Post-doctoral Researcher at ISTI-CNR, Pisa, Italy (2004), and has nearly five years of industry experience in India on telecom hardware and software development, from 1993–1997, 1999. His research interests include communication networks, with emphasis on performance modeling and analysis. Current directions include energy harvesting sensor networks, broadband wireless access and routing, cognitive/white-space access networks, and smart grid networks.

Dr. De currently serves as a Senior Editor of IEEE COMMUNICATIONS LETTERS, and an Associate Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE WIRELESS COMMUNICATIONS LETTERS, Springer Photonic Network Communications, and the IETE Technical Review Journal. He is a Senior Member of the IEEE and IEEE Communications and Computer Societies.