Published in final edited form as:
IEEE Trans Smart Grid. 2012 September ; 3(2): 1317–1324.

A Randomized Response Model For Privacy Preserving Smart Metering

Shuang Wang,
Division of Biomedical Informatics, University of California, San Diego, San Diego, CA

Lijuan Cui,
School of Electrical and Computer Engineering, University of Oklahoma, Tulsa, OK, 74135

Jialan Que,
Division of Biomedical Informatics, University of California, San Diego, San Diego, CA

Dae-Hyun Choi,
Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX

Xiaoqian Jiang,
Division of Biomedical Informatics, University of California, San Diego, San Diego, CA

Samuel Cheng, and
School of Electrical and Computer Engineering, University of Oklahoma, Tulsa, OK, 74135

Le Xie
Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX

Shuang Wang: shw070@ucsd.edu; Lijuan Cui: lj.cui@ou.edu; Jialan Que: jiq4@pitt.edu; Dae-Hyun Choi: cdh8954@neo.tamu.edu; Xiaoqian Jiang: x1jiang@ucsd.edu; Samuel Cheng: samuel.cheng@ou.edu; Le Xie: lxie@ece.tamu.edu

Abstract

The adoption of smart meters may bring new privacy concerns to the general public. Given the fact that metering data of individual homes/factories is accumulated every 15 minutes, it is possible to infer the pattern of electricity consumption of individual users. In order to protect the privacy of users in a completely de-centralized setting (i.e., individuals do not communicate with one another), we propose a novel protocol, which allows individual meters to report the true electricity consumption reading with a pre-determinted probability. Load serving entities (LSE) can reconstruct the total electricity consumption of a region or a district through inference algorithm, but their ability of identifying individual users' energy consumption pattern is significantly reduced. Using simulated data, we verify the feasibility of the proposed method and demonstrate performance advantages over existing approaches.

Keywords

Smart metering; Data privacy; Gaussian mixture

I. Introduction

With profound changes of the electric power industry towards a smarter grid in support of sustainable energy utilization, many utility companies are in the process of replacing

Correspondence to: Shuang Wang, shw070@ucsd.edu.

conventional metering devices with smart meters [1]. Smart meters make it possible to provide near real-time price incentives to customers which could potentially reduce the need for expensive peak capacity and energy. The successful adoption of smart metering and pricing could offer many benefits, including: reduction in wholesale prices [2], enhanced reliability [3], and environmental improvement [4].

However, the massive deployment of smart meters also raises a series of concerns, for example, (1) depending on the utility, the "gap" between the operational benefits and infrastructural investment is large [4]; and (2) the fear of loss of privacy (i.e., "spy at home") may arise in ordinary customers. The first problem might be obviated by benefits in the long run but the second problem becomes increasingly more challenging [5] [6]. It would be possible for utilities to infer the type of appliances individual customers are using at every 15 minutes (i.e., when you are using your computer and when your garage door is activated). The compromise of customers' privacy would be significant if they are left unprotected.

User privacy has been an important issue in various applications involving information exchange, data sharing, and medical data dissemination [7], [8], [9]. Many previous studies have been conducted in a centralized environment, where owners have the ability to adjust the data in a global manner. Oftentimes, privacy is protected by suppression, generalization, and randomization to ensure properties such as K-anonymization [10], l-divergence [11], or t-closeness [12]. Most of these techniques intend to hide identities of individuals in a crowd of others so that no single identity can be uniquely distinguished. Recently, some researchers have also considered privacy protection in a distributed environment, e.g., making peer-to-peer communication accountable without losing privacy [13] and preserving location privacy in distributed environments [14].

Unfortunately, none of these models are suited to privacy problems of smart electricity meters, which are setup in a decentralized environment. Individual smart meters report their reading to the load serving entity (LSE) but they do not communicate with one another. It remains an open question of how to preserve the ability of the LSE to compute an approximation of the current electricity consumption (i.e., for dynamic pricing) while protecting the privacy of individual users.

Efthymiou and Kalogridis [15] suggested an approach to protecting smart metering data via anonymization but their approach requires the participation of a third party. Similarly, Quinn [16] suggested that metering data can be aggregated and encrypted so that an individual's information is anonymised to roughly the scale of a city block. Recently, Kalogridis et. al. introduced a new approach to enable privacy protection of smart meters toward undetectable appliance load signatures, which used a rechargeable battery to moderate the home's load signature in order to hide appliance usage information [17]. All these techniques, however, require a significant effort in technology development, standards, policy, and regulatory activities, which are not yet available.

In this paper, a privacy protection solution based on the existing infrastructure and technology is proposed. Most directly related to our research is a recent paper by Bohli et al [18], which added Gaussian noises to each smart meter to prevent the adversary from guessing the patterns of energy consumption correctly. There are several issues with that approach: (1) a substantial amount of smart meters are required to ensure the accurate aggregated reading and protect the privacy of individuals (i.e., 3,810,000 customers are necessary to sufficiently hide the usage of a washing machine per household); (2) it is easy to recover true readings because the Gaussian noise added to each smart meter follows the same distribution; and (3) approximately half of the smart meters report negative readings because their added noise was to cover 50% confidence interval of the typical household

power consumption, rending erroneous outputs. In other words, half of these readings are meaningless to the LSE. These limitations reduce the applicability of their method in practice.

In summary, the main contribution of this paper is twofold:

- A practical and novel privacy protection method for smart metering is proposed. In this method smart meters report samples from Gaussian Mixture Models (GMM) rather than post-randomized absolute readings to the LSE. In particular, these samples can be used to recover the distribution of energy consumption at time t and calculate the total amount of energy consumption without revealing the actual individualized energy consumption patterns.
- The performance of the proposed method is verified in two parts corresponding to: (1) privacy protection analysis; and (2) data usability preservation analysis. Both analyses evaluate two performance indices, the capability of privacy protection and the accuracy of estimating the aggregated smart meter readings using some statistical tests, respectively.

This paper is organized as follows. In Section II, the concept of the GMM is briefly reviewed. Then, the parameter estimation algorithms and the formulation of the proposed privacy protocol for smart metering are introduced. The variance estimation technique using expectation propagation (EP) is discussed with more detail in Section III. The simulation results for the proposed method are presented in Section IV. Discussion and concluding remarks are summarized in Section V and Section VI.

II. Proposed Method

In this section, we propose a novel approach to protecting the privacy of smart meter users while achieving more accurate estimate for aggregated meter readings with sometimes large errors beyond typical confidence interval of Gaussian noise. Our approach utilizes a statistical method based on GMM with mixtures of K Gaussian components to mix actual smart meter readings with faked readings from $K-1$ pre-determined distributions so that the appliance usage information of individual users is protected. In the next subsection the GMM is briefly reviewed. This is followed by the introduction of privacy protection protocol based on GMM and inference algorithms for estimating the total electricity consumption in Subsection II-B and Subsection II-C, respectively.

A. Preliminary of Gaussian mixture model

In statistics, a mixture model represents the probabilistic distribution of sub-populations within a large population. In particular, the GMM is a linear superposition of Gaussian components, which provides a richer class of density models than the single Gaussian model. The GMM has been widely used in machine learning with different applications, e.g., speech recognitions [19] and image retrieval [20]. The GMM can be written as follows:

$$p(x)=\sum_{k=0}^{K-1}\omega_k\mathcal{N}(\mu_k,\sigma_k^2), \quad (1)$$

where ω_k is the weight of each Gaussian component such that $\sum_{k=0}^{K-1}\omega_k = 1$, and μ_k and σ_k^2 are the mean and variance of each Gaussian component, respectively. In the context of statistical machine learning, the process of learning model parameters (i.e. ω_k, μ_k and σ_k) based on observations is referred to as inference process. In the past decades, many inference algorithms have been studied, among which EM citation and EP citation are two

popular inference algorithms for solving the inference problem of GMM. In the rest of this section, we will introduce the proposed privacy protocol for GMM based smart metering and two inference algorithms (EM and EP), which estimate the unknown parameters (e.g., aggregated smart meter readings).

B. Privacy protocol for smart metering

Generally, normal distribution model of smart meter readings with the independent and identically distributed (i.i.d.) assumption has been widely used in many previous smart meter studies [18], [21], [22]. Therefore, we also assume that the readings of smart meters are i.i.d., and their distributions are modeled with a normal distribution parameterized by mean and variance.

In our application, an LSE gets the readings from a set of smart meters $S = \{s_1, s_2, \cdots, s_N\}$ for $N > 1$. We denote these readings by $e_{it} \in R$ where e_{it} represents the electricity consumption measured by smart meter s_i in the period t and R is the range of measures. We use T and N to denote the total time ticks and the total number of smart meters of interest.

The readings of a period t from all smart meters $\{s_i\}$ follow a Gaussian distribution $e_{it} \sim \mathcal{N}(\mu_t, \sigma_t^2)$ with mean μ_t and variance σ_t^2. From the distribution of $\{e_{it}\}$, the LSE needs to know two critical values: (1) The marginal of $\Sigma_t = \sum_{i=1}^{N} e_{it}, \forall t \in 1, \cdots, T$, which can be also written as $\Sigma_t \sim \mathcal{N}(N\mu_t, N\sigma_t^2)$, corresponds to the total energy consumption of n smart meters at time t, which is necessary for dynamic pricing; and (2) The marginal of $\Sigma_i = \sum_{t=1}^{T} e_{it}$ for all smart meters $s_i \in S$ is needed for each billing cycle.

For the protection of users' privacy, the surrogate readings generated by GMM, $\{e'_{it}\}$, are transmitted to the LSE, rather than the actual readings $\{e_{it}\}$. We construct a mixture model with $K = 3$ mixture components which all follow the Gaussian distribution but with different parameters

$$p(e'_{it}) = \omega_0 \mathcal{N}(\mu_t, \sigma_t^2) + \omega_1 \mathcal{N}(\mu_{1t}, \sigma_{1t}^2) + \omega_2 \mathcal{N}(\mu_{2t}, \sigma_{2t}^2) \quad (2)$$

where $\mathcal{N}(\mu_t, \sigma_t^2)$ corresponds to the Main Gaussian Component (MGC) which represents the Gaussian distribution of the actual smart meter readings, while the other two components $\mathcal{N}(\mu_{1t}, \sigma_{1t}^2)$ and $\mathcal{N}(\mu_{2t}, \sigma_{2t}^2)$ introduce uncertainty into actual readings for privacy protection.

We denote $\boldsymbol{\Omega} = (\omega_0, \omega_1, \omega_2)$ as a vector of mixture weights, which represents the prior probabilities of the distribution components and $\sum_{k=0}^{2} \omega_k = 1$. More specifically, the LSE only knows that a smart meter is reporting a true reading with probability ω_0 and a fake reading with probability $\omega_1 + \omega_2$ which either follows the distribution $\mathcal{N}(\mu_{1t}, \sigma_{2t}^2)$ or $\mathcal{N}(\mu_{2t}, \sigma_{2t}^2)$. Here, two parameter pairs $(\mu_{1t}, \sigma_{1t}^2)$ and $(\mu_{2t}, \sigma_{2t}^2)$ are known to the LSE.

From a Bayesian perspective, the mixture of weights $\boldsymbol{\Omega}$ are prior probabilities, which often correspond to the knowledge about the population or expert opinions. In this paper, we consider the scenario where the LSE and users, through negotiation with each other, determine a unique weight vector to be programmed into individual smart meters. Large weights on MGC will lead to accurate but less private readings, and vice versa.

C. Inference algorithms for data aggregation

Given all the smart meter readings in the period t, the primary goal of the LSE is to estimate the parameters μ_t and σ_t of the Main Gaussian Component in order to aggregate the total current electricity consumption. A general technique for finding these parameters through the maximum likelihood in latent variable models is the expectation maximization (EM) algorithm [23]. EM is an efficient iterative optimization method for solving statistical parameter estimation problem in the presence of incomplete data. The EM algorithm has been widely used in data clustering and computer vision. EM works by iteratively performing two processes: expectation step (E-step) and maximization step (M-step). According to one of the most insightful explanations of EM in terms of lower-bound maximization [24], [25], E-step can be interpreted as finding an optimal local bound to the posterior distribution, then M-step maximizes the bound to refine the estimate of the unknown parameter. However, EM algorithm is sensitive to the initialization and only converges to the local maxima. Our simulation results show that EM algorithm provides accurate estimates for the mean μ_t and variance σ_t^2 of MGC, leading to the aggregation of the total electricity consumption with a high accuracy.

Another technique for statistical parameter estimation is a deterministic approximate method, which is based on analytical approximations to the posterior distribution from Bayesian inference perspective. Among deterministic approximate methods, expectation propagation (EP) [26] has a higher accuracy comparing to others (e.g. Laplace approximation (LA) and variational Bayes (VB)). Since EP algorithm has been widely used to solve different inference problems [27], [28], in this paper, we also apply EP algorithm to our parameter estimation problem. Our simulation results show that EP leads to more improvement on the estimation accuracy than EM in some cases. The details of the variance estimation using EP will be described in Section III.

III. Variance estimation with EP

In this section, we will describe the variance σ_t^2 estimation using EP algorithm. Usually, the variance estimation problem can be depicted as a learning or inference problem on graphical models, especially the factor graph. In Bayesian inference perspective, the estimation of variance corresponds to the estimation of the posterior distribution of the variance. Belief propagation (BP)-like algorithms such as the sum-product and the max-product algorithms are the popular techniques for inference problem. Thus, our variance estimation problem can be depicted on the factor graph in Fig.1, where all circle nodes denote variable nodes, and all square nodes denote factor nodes. Here, for the ease of exposition, we let $\upsilon=\sigma_t^2$, $\upsilon_1=\sigma_{1t}^2$, $\upsilon_2=\sigma_{2t}^2$, $\mu = \mu_t$, $\mu_1 = \mu_{1t}$ and $\mu_2 = \mu_{2t}$. The posterior distribution of the variance can be described as the product of all the incoming messages $p(\upsilon|\mathbf{y}) \propto m_{g\to V}(\upsilon) \prod_{i\in\mathcal{N}\setminus g(V)} m_{f_i\to V}(\upsilon)$, where $m_{g\to V}(\upsilon)$ corresponds to the *a priori* distribution $p(\upsilon)$ for υ modeled by a factor function $g(\upsilon)$, $m_{f_i\to V}(\upsilon)$ corresponds to the likelihood for observation y_i (smart meter reading (SMR)) modeled by factor function $f(y, \upsilon)$. However, it is infeasible to calculate the posterior distribution directly using BP algorithm, since the belief state of υ is a mixture of 3^N Gaussian distributions when each observation y_i follows a Gaussian mixture model with three components (see (2)). Thus, we resort to approximate the posterior distribution by $q(\upsilon)$ through EP algorithm. The key idea of EP is to sequentially compute approximate messages $\tilde{m}_{g\to V}(\upsilon)$ and $\tilde{m}_{f_i\to V}(\upsilon)$ in replace of true messages $m_{g\to V}(\upsilon)$ and $m_{f_i\to V}(\upsilon)$, then get a posterior on υ by combining these approximations together.

The formula of EP is given as follows:

1. Initialize the term approximation $\tilde{m}_{g\to V}(\upsilon)$ (i.e. the priori knowledge of the variance), and $\tilde{m}_{f_i\to V}(\upsilon)$ (i.e. the approximate likelihood of the variance based on the noise SMR).
2. Compute the posterior approximation for υ as:

$$q(\upsilon)=\frac{1}{Z}\tilde{m}_{g\to V}(\upsilon)\prod_{i\in\mathcal{N}\setminus g(V)}\tilde{m}_{f_i\to V}(\upsilon), \quad (3)$$

   where $\mathcal{N}\setminus g(V)$ denotes the set of all neighbors' indices for variable node V except the index of g, $Z=\int_\upsilon \tilde{m}_{g\to V}(\upsilon)\prod_{i\in\mathcal{N}\setminus g(V)}\tilde{m}_{f_i\to V}(\upsilon)$ is a normalization factor.
3. Until messages converge:

   For each factor node f_i (capturing the likelihood of the variance based on the noise SMR), where $i\in\mathcal{N}\setminus g(V)$

   a. Remove the approximate message $\tilde{m}_{f_i\to V}(\upsilon)$ from the posterior approximation $q(\upsilon)$ to generate

$$q^{\setminus i}(\upsilon)\propto\frac{q(\upsilon)}{\tilde{m}_{f_i\to V}(\upsilon)}, \quad (4)$$

      where $q^{\setminus i}(\upsilon)$ represent the product of all the incoming messages for variable υ except the message sent from factor node f_i.

   b. To update $q(\upsilon)$, combine $q^{\setminus i}(\upsilon)$, the current message $m_{f_i\to V}(\upsilon)$ and the normalization constant Z to get a complex posterior $\hat{p}(\upsilon)$. Then minimize the Kullback Leibler (KL) divergence $D(\hat{p}(\upsilon)\|q(\upsilon))$ by performing moment matching (**Proj**[·]). Thus,

$$q(\upsilon)=\mathbf{Proj}[\frac{1}{\mathbf{Z}}\mathbf{q}^{\setminus \mathbf{i}}(\upsilon)\mathbf{m}_{\mathbf{f_i}\to\mathbf{v}}(\upsilon)], \quad (5)$$

      where $Z=\int_\upsilon q^{\setminus i}(\upsilon)m_{f_i\to V}(\upsilon)$.

   c. Set approximate message

$$\tilde{m}_{f_i\to V}(\upsilon)=\frac{Zq(\upsilon)}{q^{\setminus i}(\upsilon)}. \quad (6)$$

A detailed derivation of the proposed EP algorithm is provided in the Appendix Section.

IV. Simulation Results

In this section, we show the simulation results of our proposed scheme in two subsections:

- Subsection IV-B: The privacy protection capability of the proposed method is evaluated.
- Subsection IV-C: The estimation accuracy of the smart meter readings aggregated by the LSE is measured.

In Subsection IV-B, we use two statistical tests, a paired F-test for individual smart meters over time and a Kolmogorov-Smirnov test (KS-test) [29] for all smart meters at a given time tick. The null hypothesis here is that the difference is significant. We would use the percentage of rejecting the null hypothesis as our index to measure the extent of privacy

protection. The reasons why we choose the aforementioned two tests as the metrics of privacy risk are as follows. From statistical inference perspective, the learning of user behavior corresponds to the estimate of the posterior distribution of user behavior based on observations. However, the inference outcome in statistics highly relies on the model observations (i.e. smart meter readings in our problem). Therefore, we can interpret the problem of minimizing privacy risk of each user as introducing significantly changes the smart meter reading distribution over temporal and spatial axes but preserving their marginals. In other words, in order to protect the privacy of users the inference outcomes about personal information (i.e., energy consumption) based on the original readings and surrogate readings should be highly different with each other. Now, we address the problem how much different the aforementioned two meter readings are. In statistics, the goal of applying these tests is to check whether the smart meter readings before and after applying our protection model are significantly different in terms of distribution and variance. That is, data collected in one situation (i.e. the original smart meter readings in our problem) is compared to data collected in a different situation (i.e. the surrogate smart meter readings after applying the proposed algorithm) with the aim of examining if the first situation produces different results from the second situation. If surrogate readings can pass the aforementioned statistical tests, we conclude that the inference outcome based on the surrogate readings will be different from that based on the original readings with a high probability in statistics. With that discussed, the proposed F-test and KS-test reference provide two valid metrics to quantify privacy risk.

In Subsection IV-C, we measure the differences (i.e., mean and standard deviation (SD)) between the aggregated actual readings and the recovered total readings using our approach. This subsection demonstrates that the introduced techniques used for data recovering (e.g., EM or EP) lead to different statistics of the recovered total readings. The results from Subsection IV-B and Subsection IV-C are based on the simulation setup for the performance analysis of the proposed method described in the next subsection.

A. Simulation Setup

We use simulated data to evaluate the performance of the proposed method in a Close-to-Real environment. We use hourly demands reported by California ISO (CAISO) [30] to simulate our data in the following steps:

1. Estimate the number of users: According to the report from US department of energy [31], a typical household in US uses 920 kW of electricity per month. By numerically integrating the hourly electricity demands within Jun 2011 from CAISO, we can obtain the total monthly electricity consumption of the whole California (CA) area. Then, the total monthly electricity consumption is divided by the average monthly usage (920 kW) to approximate the number of users in CA. By calculation, the approximate number of users in CA is about 22.44 million, which quite matches the US 2010 census data [32].
2. Estimate aggregated power consumption of every 15 minutes: Given the approximate number of users and hourly demands of the whole CA area, we can easily estimate the average hourly demands of each user. Since we are interested in the aggregated power consumption of every 15 minutes, it is required to interpolate quarter demands between hourly demands followed by numerical integration. Fig. 2 shows the average aggregated power consumption of every 15 minutes by using the proposed data simulation strategy.
3. Sample individual power consumption: As we justified in section II-B, the readings of smart meters are i.i.d. and their distributions can be modeled with a normal distribution [18], [21], [22] parameterized by mean μ and variance σ^2, where the

mean μ is equal to the corresponding aggregated power consumption estimated in step 2). Then we can sample individual power consumption from these normal distributions with different SD σ to build a Close-to-Real simulation environment.

In our experiment, we consider a GMM with three components ($K = 3$). Moreover, to make the simulation more reasonable, the number of smart meters N in our study is set equal to 500, which is a typical size of a small community. The SD σ_t of the MGC is calculated by choosing different percentages of the mean μ_t of the corresponding MGC. In our simulation, the affection of sampling data from different σ_t's is discussed later in this paper. Furthermore, SD's σ_{it} of the other two Gaussian mixtures are always β times bigger than σ_t. The means of the other two Gaussian components are decided by setting $\mu_{1t} = \text{icdf}(\alpha, \text{icdf}(\alpha, \mu_t, \sigma_t), \beta\sigma_t)$ and $\mu_{2t} = \text{icdf}(1-\alpha, \text{icdf}(1-\alpha, \mu_t, \sigma_t), \beta\sigma_t)$, respectively, where $\text{icdf}(\alpha, \mu, \sigma)$ evaluates the inverse Gaussian cumulative distribution at the values α with parameter values given by μ and σ. Furthermore, we set $w_i = \frac{1 - w_0}{K - 1}$ for all $i \neq 0$. The above settings guarantee that the generated GMM is symmetric. In addition, the parameters α and β can provide a trade-off between privacy protection and data usability preservation. In our simulation, we choose $\alpha = 0.01$ and $\beta = 2$. Finally, all results presented below are averaged over 100 trials and the error bar in each figure indicates the corresponding standard deviation.

B. Privacy Protection Analysis

First, we study the privacy protection capability of the proposed method by selecting different σ_t, where σ_t is represented by the percentage of μ_t in our results. In this case, we choose $w_0 = 0.7$ and vary σ_t from 2% to 20% percentage of μ_t. In Fig.3, we presented the F-test and KS-test results for each user tick and KS-test for each time. The former two tests aim to verify that the variance and the distribution for each time tick before and after introducing our protection model are significantly different, while the last test focuses on verifying that the distribution changes of each user's SMR during a 24-hour period are also significant before and after implementing the proposed model. Given fixed parameters $\alpha = 0.01$ and $\beta = 2$, Fig.3 shows that the proposed method can protect nearly 90% users (KS-test for each user) on a one day time series, when $\frac{\sigma t}{\mu t}$ is larger than 8%. Moreover, the KS-test for each time in Fig.3 shows a 100% protection for all σ_t's, where 96 time ticks during a 24-hour period are included in our test. Similar to the result of KS-test for each user, the protection index of F-test decreases significantly, since σ_t is very small at $5\%\mu_t$ point. To avoid degradations of protection with respect to F-test for each user and KS-test for each time, one can increase β, so that the other two Gaussian components have large enough σ_{it} to generate a significantly mixed distribution.

Second, we study the privacy protection capability of the proposed method by changing the weights on Main Gaussian Component w_0. In this case, we fix $\sigma_t = 20\%\mu_t$ and vary w_0 from 0.7 to 0.95. In Fig. 4, we can see that almost 99% time, the proposed model can sufficiently protect each user away from privacy risk, if $w_0 < 0.8$. Furthermore, if $w_0 < 0.7$, the proposed method can efficiently protect more than 95% users on their daily electricity consumption. However, the protection indices with respect to all tests, especially the KS-tests for each time and each user, decrease very fast as the weight w_0 increase. It is due to the fact that the MGC will dominate the GMM for large w_0's.

C. Data Usability Preservation Analysis

To preserve data usability to the LSE, it requires the accurate estimate of the aggregation of all SMR at a given time tick. In this paper, we use relative accuracy (RA) and the mean square error (MSE) to quantify the estimation accuracy, where RA is defined as $1 - \frac{|\hat{\mu}_t - \mu_t|}{\mu_t}$.

First, we analyze the impact of w_0 on the estimation accuracy and compare our results with Bohil's results, where we fix $\sigma_t = 20\%\mu_t$. Fig. 5 shows that the proposed algorithm can achieve at least 99% RA by using 500 smart meters and the RA increases as w_0 increases. However, the required number of smart meters in Bohli's method is at least 12, 000 to obtain the same accuracy. Furthermore, this number in Bohil's method increases significantly as the weight w_0 increases. Thus the proposed method is more flexible to provide privacy protection at a community level with high confidence. Moreover, regarding different w_0, we also study the estimation accuracy in terms of MSE in Fig. 6. Fig. 6 shows that the proposed method provides a high accurate estimate for both μ_t and σ_t in terms of low MSE. In addition, when w_0 is small (equal or less than 0.75), EP algorithm provides a better estimation accuracy of σ_t than EM algorithm, while EM algorithm works better as w_0 is larger than 0.75.

Second, we study the estimation accuracy with respect to different σ_t, for which we set $w_0 = 0.7$. The relationship between RA and σ_t is, firstly, shown in Fig.7, where the RA decreases as σ_t increases. As a reference, the number of smart meters required in Bohli's method has also been plotted in Fig. 7. As mentioned earlier, to reach the same estimation accuracy, Bohli's method needs a huge number of smart meters (e.g. more than 3 million smart meters are needed to achieve a 99.95% accuracy). Moreover, the estimation accuracy of mean and SD with respects to changing σ_t is given in Fig.8 in terms of MSE. We can see that MSE of the estimated mean and SD increases as σ_t increases. In addition, the estimation accuracy of EP algorithm for SD estimation always outperforms EM algorithm.

Third, we investigate the accuracy of the marginal of $\Sigma_i = \sum_{t=1}^{T} e_{it}$ for all smart meters $s_i \in S$, which corresponds the total energy consumption of each user during a given period T. Since smart meter reports its reading every 15 minutes, there are total $\mathscr{T} = 96$ reports per day. Therefore, we choose $T = \mathscr{T} 30(days) = 2,880$ to study the accuracy of monthly energy consumption. Fig. 9 depicts the aggregation accuracies of the proposed privacy protection algorithm for a month period by varying the weight of MGC, where 500 smart meters are used in this experiment. In Fig. 9, the proposed algorithm achieves a high RA, (i.e. 98.78%), even though the weight ω_0 is equal to 0.7. Moreover, as the weight ω_0 increases, the RA can achieve as high as 99.5%. Although we have shown that the proposed algorithm can provide the high fidelity of monthly aggregation, it is necessary to point out, in practice, neither the user nor the LSE would like to compromise for paying the difference. Thus, a practical workaround is to incorporate a local aggregator on each smart meter, which can losslessly collect the monthly usage of each user and report the corresponding sum total at the end of each billing cycle.

In addition to estimation accuracy, we study the distributions of SMR of a given time tick before and after applying protection methods. In Fig. 10, we can see that both the proposed method and Bohli's method change the true distribution of SMR a lot. Notably, the changes introduced by the proposed method only occur at the lower and upper ends of the true distribution. It can be interpreted that our method dedicates to protect these users who really need to be protected. In the proposed method, we provide a tunable parameter α to control the protection coverage, where a larger α means a wider protection coverage of the lower and upper ends. In contrast, Bohli's method adds huge noise on all SMR, which intensely

destroys the true distribution. Fig. 10 shows that the estimated distribution of SMR generated by the proposed method quite fits the true distribution, while the estimated distribution of the data obtained from Bohli's method is totally different from the true distribution. Thus, the surrogate SMR generated by the proposed method is more acceptable, because it sufficiently preserves important information of the true distribution. Another significant advantage of the proposed method is that it can avoid negative SMR report by controlling the Gaussian mixture component at the lower end. In contrast, Bohli's method reports a large portion of negative SMR to make sure a high accuracy of the aggregated SMR. Thus, we conclude that the data generated by Bohli's method is lack of utility and practical value.

Finally, we investigate the SMR of a given user during a 24 hours period in Fig. 11. We can see that the proposed method only adds minor changes at a few time ticks to protect the individual privacy, while Bohli's method adds huge noises at almost all SMR. In Fig. 11, it can be seen clearer that more than half of the surrogate data generated by Bohli's method are negative values. Although both proposed method and Bohli's method can provide high accurate aggregated SMR and privacy protection on SMR, Bohli's method needs a huge number of smart meters and a large portion of negative SMR. The proposed algorithm sufficiently solves these drawbacks within Bohli's methods. Thus we conclude that the proposed algorithm outperforms the Bohli's method.

V. Discussion

Privacy is a less studied but important aspect of dynamic pricing associated with smart metering, particularly when readings are required to be reported in a timely manner (e.g., every 15 minutes). While previous efforts to protect privacy suggested solutions based on non-existing conditions (i.e., trusted third parties and/or rechargeable batteries), our solution is based on available resources, and it can be realized with a reasonable size of population, e.g., about 500 ~ 1000 smart meters in one load district with performance guarantees.

Without increasing computational complexity, our approach demonstrates better performance in experiments using synthetic data compared to a recent approach based on post-randomization [18]. Because the summation of Gaussian random variables is still a Gaussian random variable, our proposed framework can be generalized towards more complicated scenarios where different smart meters carry different uncertainties in reporting their readings to achieve fine-grained privacy protection.

The limitation of this study is that the authors only considered smart meter-based security protocol. For example, all kinds of energy consumption at home are reported to the LSE with the same level security. In the future, we would need more intelligent smart meter with a quality of service-based security protocol. That is, a single smart meter at home must be able to allocate different security levels for different services. Therefore, GMM model proposed in this paper may need to be updated in order to accommodate future needs of differentiated privacy levels.

VI. Conclusion

We propose a novel method to protect the privacy of customers with smart meters while assuring the capability to load serving entities to estimate the current electricity consumption. Uncertainty is introduced to the readings of smart meters by shuffling actual values with faked ones at a given prior probability. We show in our experiments that recovered total electricity consumption can approximate the aggregated actual readings with very high accuracy (~99%), whereas the ability of identifying individual usage patterns are

largely obviated. The proposed framework provides a privacy-preserving approach to utilizing smart meters at end users' levels. Future work will include testing of this proposed method with large-scale realistic smart metering data.

Acknowledgments

The authors would like to thank Professor Lucila Ohno-Machado, Division of Biomedical Informatics, University of California San Diego, San Diego, for involving the discussion and providing valuable comments.

Shuang Wang and Xiaoqian Jiang were funded in part by the National Library of Medicine (R01LM009520) and iDASH (NIH grant U54HL108460). Dae-Hyun Choi and Le Xie were supported in part by Texas Engineering Experiment Station and Power Systems Engineering Research Center.

References

1. Faruqui A. The economics of dynamic pricing for the mass market. The Brattle Group, April. 2007
2. Vasconcelos J. Survey of regulatory and technological developments concerning smart metering in the european union electricity market. 2008
3. Collier S. Ten steps to a smarter grid. Industry Applications Magazine, IEEE. 2010; vol. 16(no. 2): 62–68.
4. Neenan B, Hemphill R. Societal benefits of smart metering investments. The electricity journal. 2008; vol. 21(no. 8):32–45.
5. McDaniel P, McLaughlin S. Security and privacy challenges in the smart grid. IEEE Security & Privacy. 2009:75–77.
6. Khurana H, Hadley M, Lu N, Frincke D. Smart-grid security issues. Security & Privacy, IEEE. 2010; vol. 8(no. 1):81–85.
7. Fienberg, S.; Martin, M. Sharing research data. National Academy Press; 1985.
8. Fienberg S. Privacy and confidentiality in an e-commerce world: Data mining, data warehousing, matching and disclosure limitation. Statistical Science. 2006; vol. 21(no. 2):143–154.
9. Donaldson, M.; Lohr, K. Health data in the information age: use, disclosure, and privacy. National Academy Press; 1994.
10. Sweeney L, et al. k-anonymity: A model for protecting privacy. International Journal of Uncertainty Fuzziness and Knowledge Based Systems. 2002; vol. 10(no. 5):557–570.
11. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD). 2007; vol. 1(no. 1): 3–es.
12. Li, N.; Li, T.; Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. IEEE 23rd International Conference on Data Engineering (ICDE); IEEE; 2007. p. 106-115.
13. Belenkiy, M.; Chase, M.; Erway, C.; Jannotti, J.; Kupcu, A.; Lysyanskaya, A.; Rachlin, E. Making p2p accountable without losing privacy. Proceedings of the 2007 ACM workshop on Privacy in electronic society; ACM; 2007. p. 31-40.
14. Zhong, G.; Hengartner, U. A distributed k-anonymity protocol for location privacy. IEEE International Conference on Pervasive Computing and Communications (PerCom); IEEE; 2009. p. 1-10.
15. Efthymiou, C.; Kalogridis, G. Smart grid privacy via anonymization of smart metering data. First IEEE International Conference on Smart Grid Communications (SmartGridComm); IEEE; 2010. p. 238-243.
16. Quinn E. Smart metering & privacy: Existing law and competing policies. A Report for the Colorado Public Utilities Commission. 2009
17. Kalogridis, G.; Efthymiou, C.; Denic, S.; Lewis, T.; Cepeda, R. Privacy for smart meters: towards undetectable appliance load signatures. First IEEE International Conference on Smart Grid Communications (SmartGridComm); IEEE; 2010. p. 232-237.
18. Bohli, J.; Sorge, C.; Ugus, O. A privacy model for smart metering. IEEE International Conference on Communications Workshops (ICC); IEEE; 2010. p. 1-5.

19. Reynolds D, Rose R. Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing. 1995; vol. 3(no. 1):72–83.
20. Permuter, H.; Francos, J.; Jermyn, I. Gaussian mixture models of texture and colour for image database retrieval. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03); IEEE; 2003. p. III–569
21. Varodayan, D.; Khisti, A. Smart meter privacy using a rechargeable battery: minimizing the rate of information leakage. IEEE International Conference on Acoustics, Speech, and Signal Processing; IEEE; 2011. p. 1932-1935.
22. Clement-Nyns K, Haesen E, Driesen J. The impact of charging plug-in hybrid electric vehicles on a residential distribution grid. IEEE Transactions on Power Systems. 2010; vol. 25(no. 1):371–380.
23. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological). 1977; vol. 39(no. 1):1–38.
24. Neal R, Hinton G. A view of the em algorithm that justifies incremental, sparse, and other variants. Learning in graphical models. 1998; vol. 89:355–368.
25. Minka, T. Expectation-maximization as lower bound maximization. 1998. *Tutorial published on the web at* http://www-white.media.mit.edu/tpminka/papers/em.html
26. Minka T. Expectation propagation for approximate Bayesian inference. Uncertainty in Artificial Intelligence. 2001; vol. 17:362–369.
27. Wang, S.; Cui, L.; Cheng, S. Noise Adaptive LDPC Decoding Using Expectation Propagation; IEEE Global Telecommunications Conference; 2011.
28. Cui L, Wang S, Cheng S. Adaptive Slepian-Wolf decoding based on expectation propagation. Accepted by IEEE Communications letter. 2011
29. Massey F. The kolmogorov-smirnov test for goodness of fit. Journal of the American Statistical Association. 1951; vol. 46(no. 253):68–78.
30. http://www.caiso.com/.
31. [accessed on July 7, 2011] http://www.oe.energy.gov/information_center/faq.htm
32. [accessed on July 7, 2011] http://quickfacts.census.gov/qfd/states/06000.html
33. Opper M. A Bayesian approach to on-line learning. On-line Learning in Neural Networks. 1999:363–378.

Appendix

A. IG Distribution Approximation using EP

Since the inverse gamma (IG) distribution is a conjugate prior for Gaussian distribution with variance as a parameter, we choose the *a priori* distribution as an IG distribution. Then we will explain the EP algorithm on the problem of variance estimation based on the IG distribution.

Each iteration of EP based IG distribution approximation proceeds as follows:

1. Initialize the prior messages (i.e. initial knowledge of the variance) for the variance variable

$$m_{g\to V}(\upsilon)=IG(\upsilon,\alpha^0,\beta^0)=\frac{\beta^{0\alpha^0}}{\Gamma(\alpha^0)}\upsilon^{-\alpha^0-1}\exp(-\frac{\beta^0}{\upsilon}) \quad (7)$$

with

$$\alpha^0=1,\beta^0=\upsilon^0(\alpha^0+1),\text{ and } z^0=\frac{\beta^{0\alpha^0}}{\Gamma(\alpha^0)}, \quad (8)$$

where υ^0 is the initial variance, β^0 and α^0 are scale and shape parameters for IG distribution, respectively.

2. Initialize the likelihood messages

$$\tilde{m}_{f_i \to V}(\upsilon) = IG(\upsilon, \alpha_i, \beta_i) = z_i \upsilon^{-\alpha_i - 1} \exp\left(-\frac{\beta_i}{\upsilon}\right) \quad (9)$$

with

$$\beta_i = 0,\ \alpha_i = -1 \text{ and } z_i = 1, \quad (10)$$

3. Initialize the posterior probability distributions of the variance

$$q(\upsilon) = Z \upsilon^{-\alpha^{\text{new}} - 1} \exp\left(-\frac{\beta^{\text{new}}}{\upsilon}\right) \quad (11)$$

with

$$\alpha^{\text{new}} = \alpha^0, \beta^{\text{new}} = \beta^0, \text{ and } Z = z^0. \quad (12)$$

4. Until all messages converge:

   For each factor node f_i, where $i \in \mathcal{N} \backslash g(V)$

   a. Remove $\tilde{m}_{f_i \to V}(\upsilon)$ from the posterior $q(\upsilon)$

$$\alpha^{\text{tmp}} = \alpha^{\text{new}} - (\alpha_i + 1); \beta^{\text{tmp}} = \beta^{\text{new}} - \beta_i. \quad (13)$$

   b. Update α^{new} and β^{new} according to moment matching. (See Section VI-B for more details about the derivation of α^{new}, β^{new} updates.)

   c. Set approximate message

$$\alpha_i = \alpha^{\text{new}} - (\alpha^{\text{tmp}} + 1); \beta_i = \beta^{\text{new}} - \beta^{\text{tmp}};$$
$$z_i = Z \frac{\beta^{\text{new}\,\alpha^{\text{new}}}}{\Gamma(\alpha^{\text{new}})} \frac{\Gamma(\alpha^{\text{tmp}})}{\beta^{\text{tmp}\,\alpha^{\text{tmp}}}} \frac{\Gamma(\alpha_i)}{\beta_i^{\alpha_i}}. \quad (14)$$

B. Moment Matching for IG distribution

By the technique of moment matching [33], $q(\upsilon)$ is obtained by matching the mean and variance of $q(\upsilon)$ to those of $\hat{p}(\upsilon)$. Then we get α^{new} and β^{new}, the parameters of $q(\upsilon)$. We simplify the notations and let $\alpha' = \alpha^{\text{tmp}}$, $\beta' = \beta^{\text{tmp}}$.

The mean and variance of IG distribution are matched by the following updates

$$m_1 = \frac{w_1 C + w_2 \frac{\beta'}{\alpha' - 1} N(y; \mu_1, \upsilon_1) + w_3 \frac{\beta'}{\alpha' - 1} N(y; \mu_2, \upsilon_2)}{w_1 C \frac{(\alpha' - \frac{1}{2})}{B} + w_2 N(y; \mu_1, \upsilon_1) + w_3 N(y; \mu_2, \upsilon_2)},$$

$$m_2=\frac{w_1C+\frac{w_2\beta'^2}{(\alpha'-1)(\alpha'-2)}N(y;\mu_1,\upsilon_1)+\frac{w_3\beta'^2}{(\alpha'-1)(\alpha'-2)}N(y;\mu_2,\upsilon_2)}{w_1C\frac{(\alpha'-\frac{1}{2})(\alpha'-\frac{3}{2})}{B^2}+w_2N(y;\mu_1,\upsilon_1)+w_3N(y;\mu_2,\upsilon_2)} \quad (15)$$

$$\begin{aligned}\alpha^{\text{new}}&=\frac{m_2}{m_2-m_1^2}+1,\\ \beta^{\text{new}}&=m_1(\alpha^{\text{new}}-1),\end{aligned} \quad (16)$$

where $C=\frac{\beta'^{\alpha'}}{\Gamma(\alpha')}\frac{1}{\sqrt{2\pi}}\frac{\Gamma(\alpha'-\frac{1}{2})}{(\beta'+\frac{(y-\mu)^2}{2})^{(\alpha'-\frac{1}{2})}}$, $B=\beta'+\frac{(y-\mu)^2}{2}$.

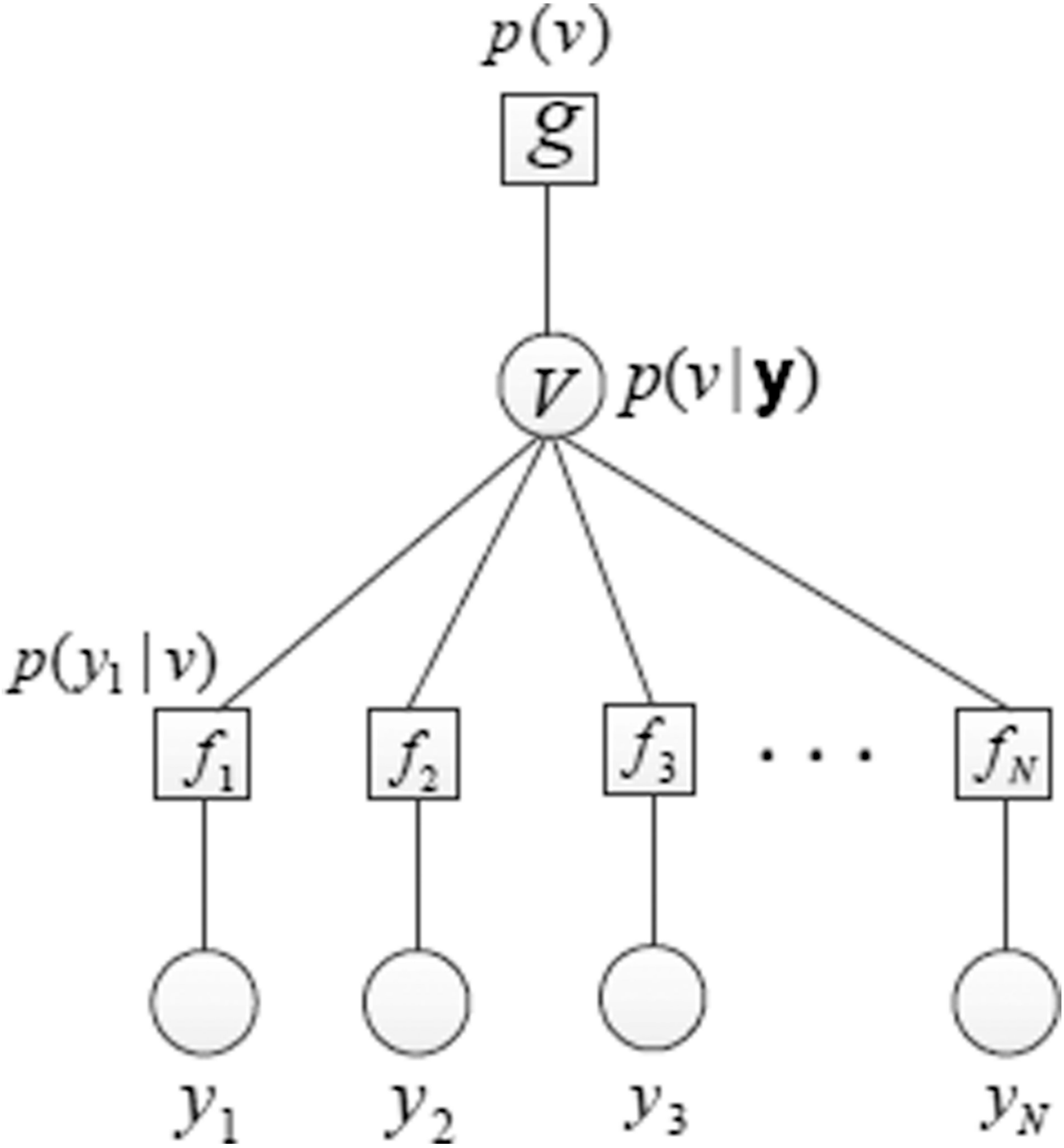

Fig. 1.
Factor graph of variance estimation with known mean.

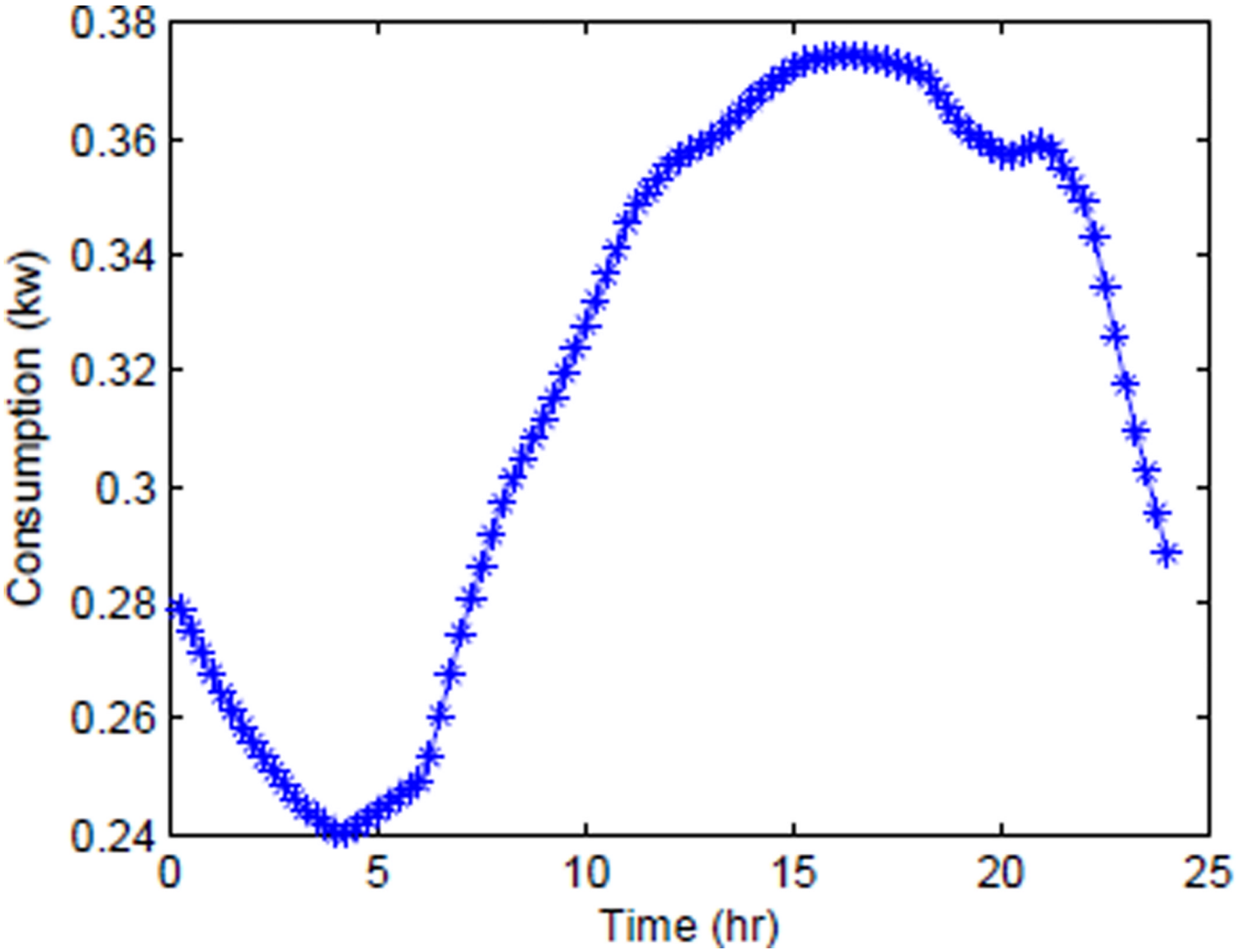

Fig. 2.
Average aggregated power consumption of every 15 minutes for each individual user

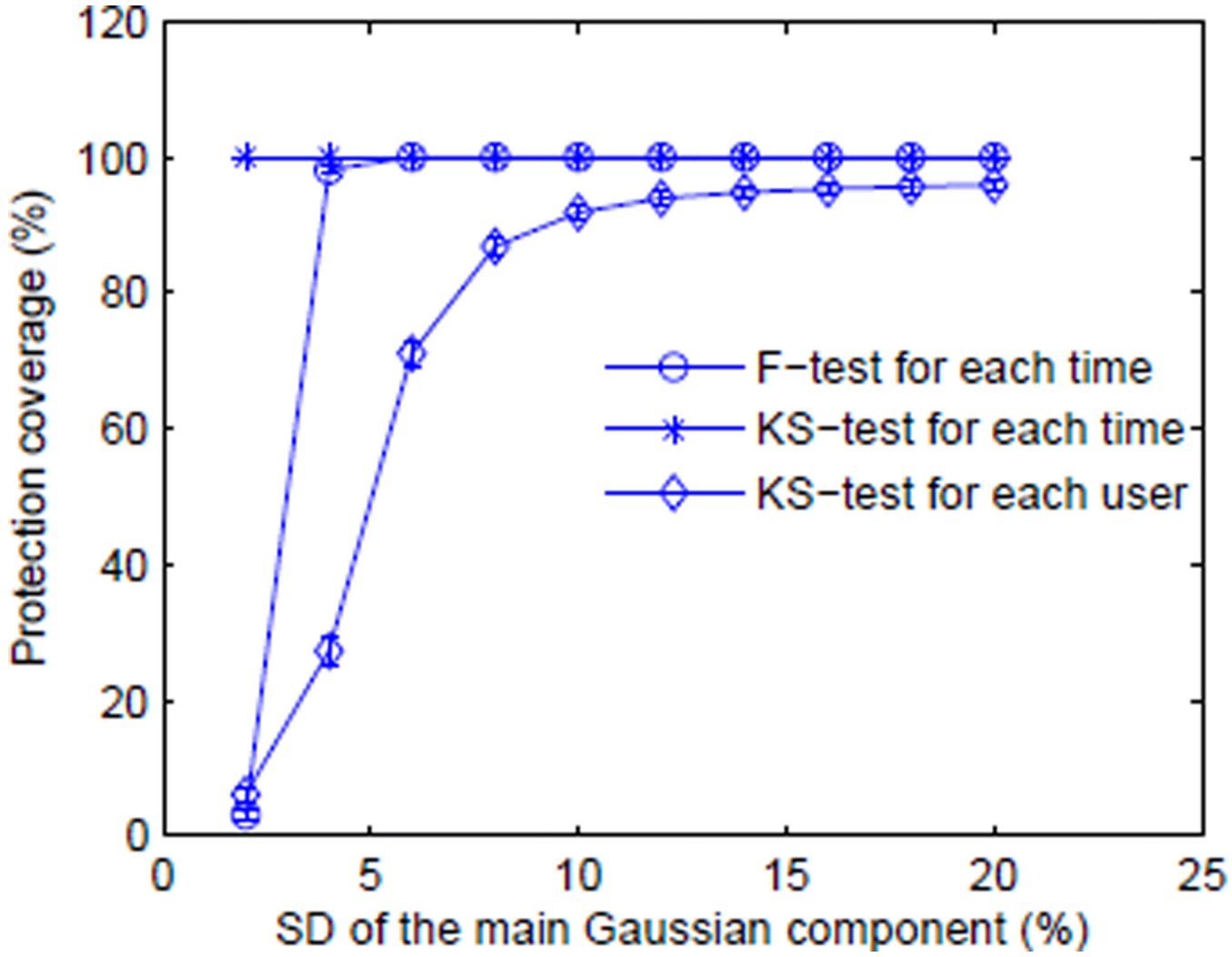

Fig. 3.
The privacy protection performance of different SD's

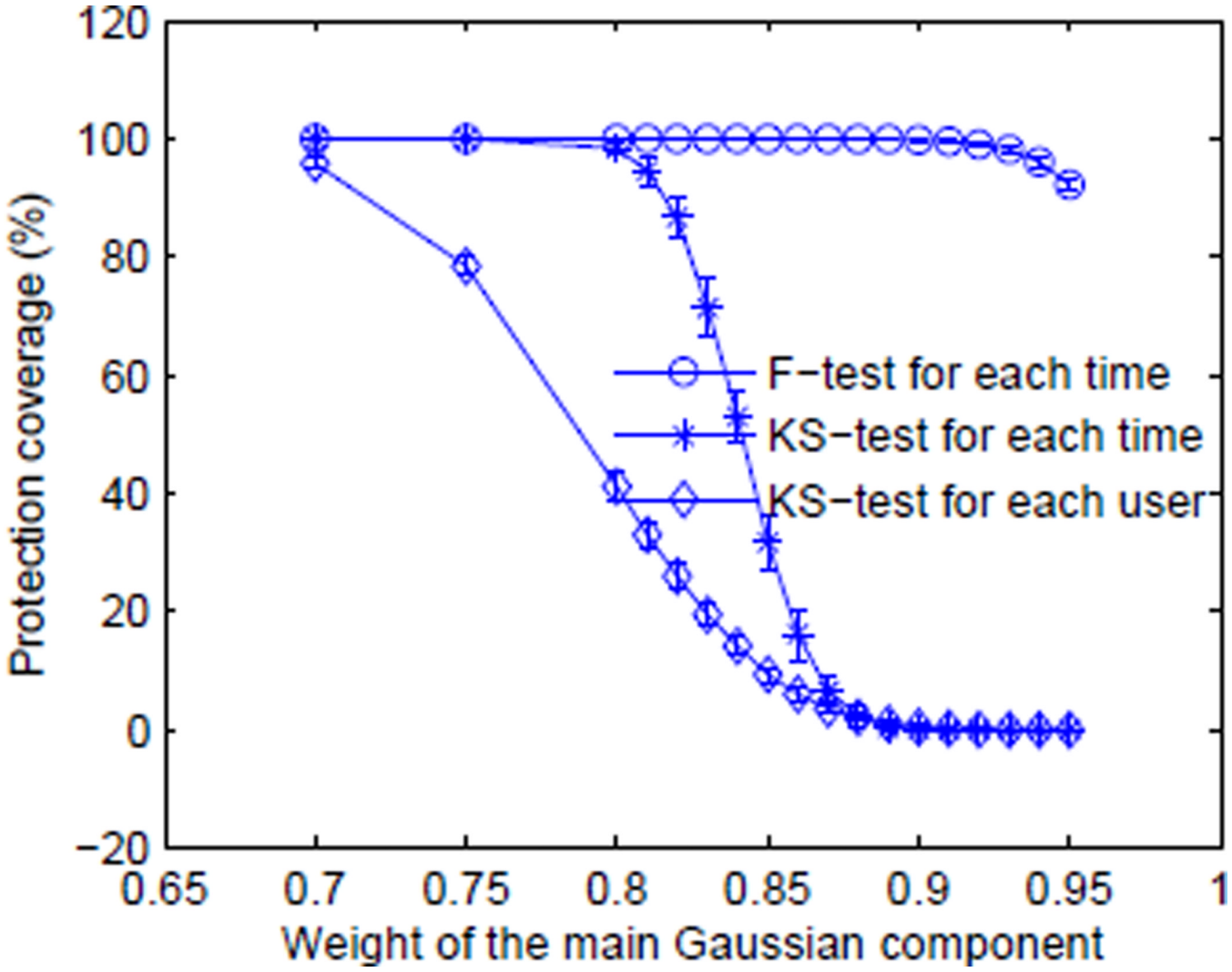

Fig. 4.
The privacy protection performance of different weights

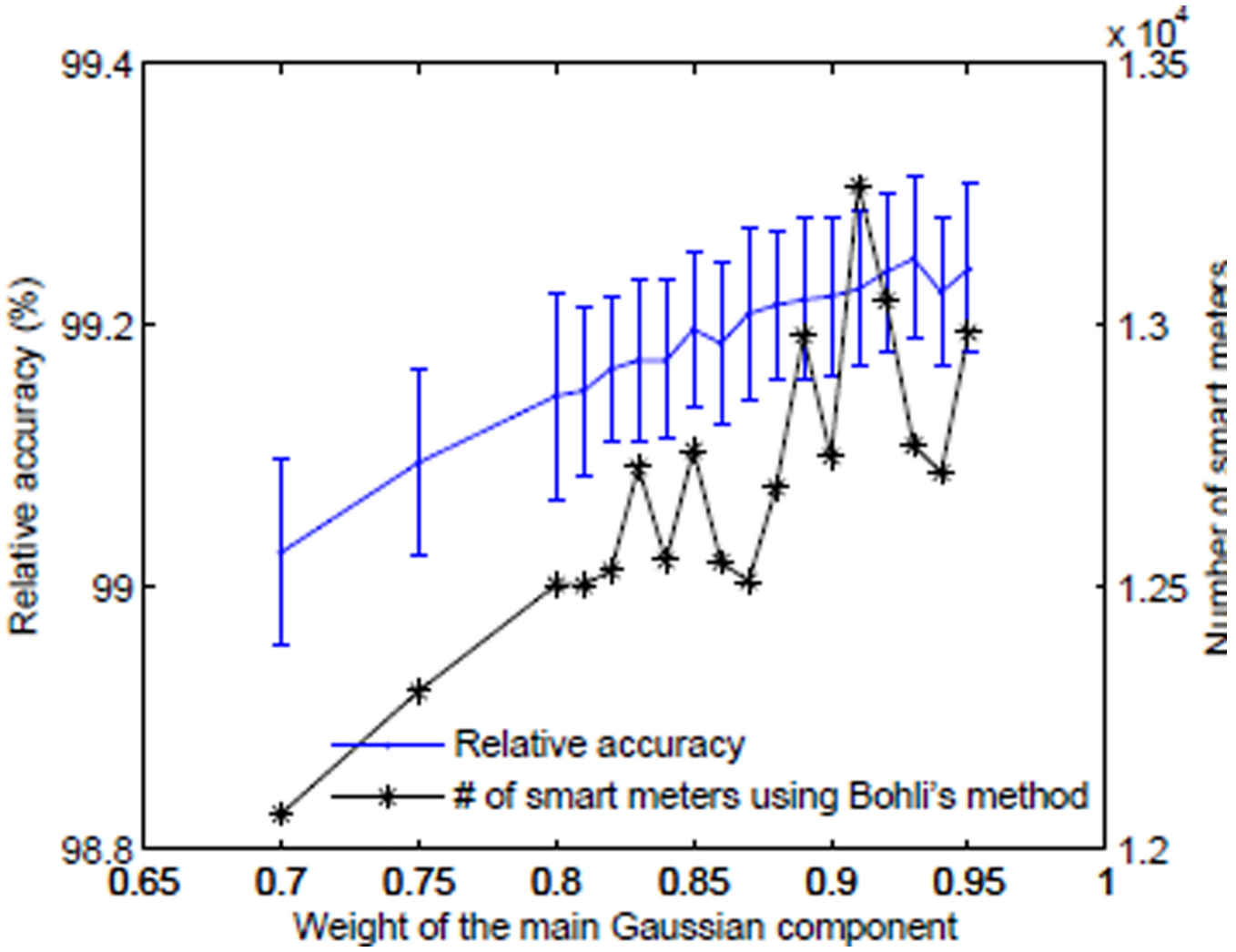

Fig. 5.
Weight of the main Gaussian component vs confidence interval.

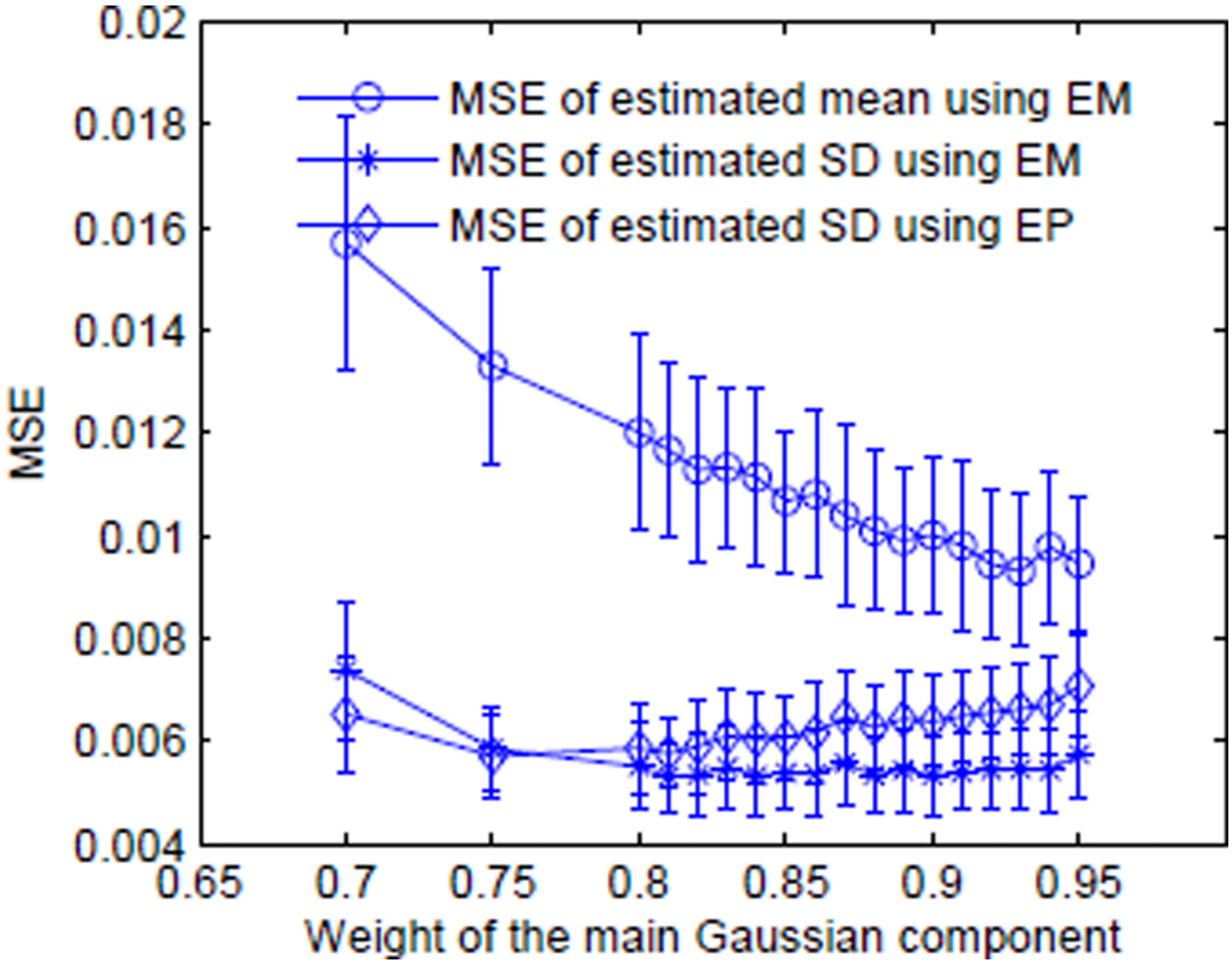

Fig. 6.
Weight of the main Gaussian component vs MSE.

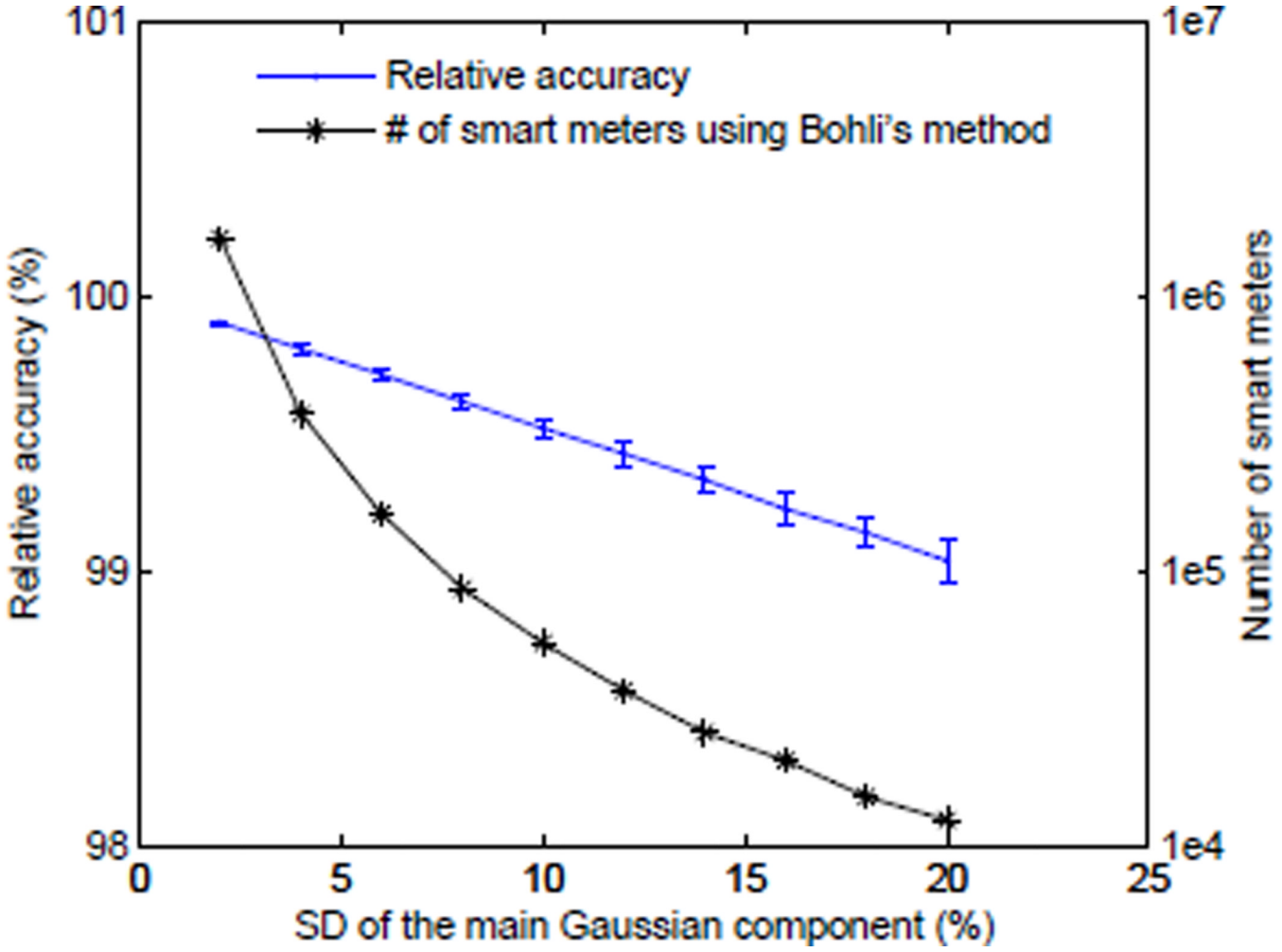

Fig. 7.
SD of the main Gaussian component (%) vs Confidence interval.

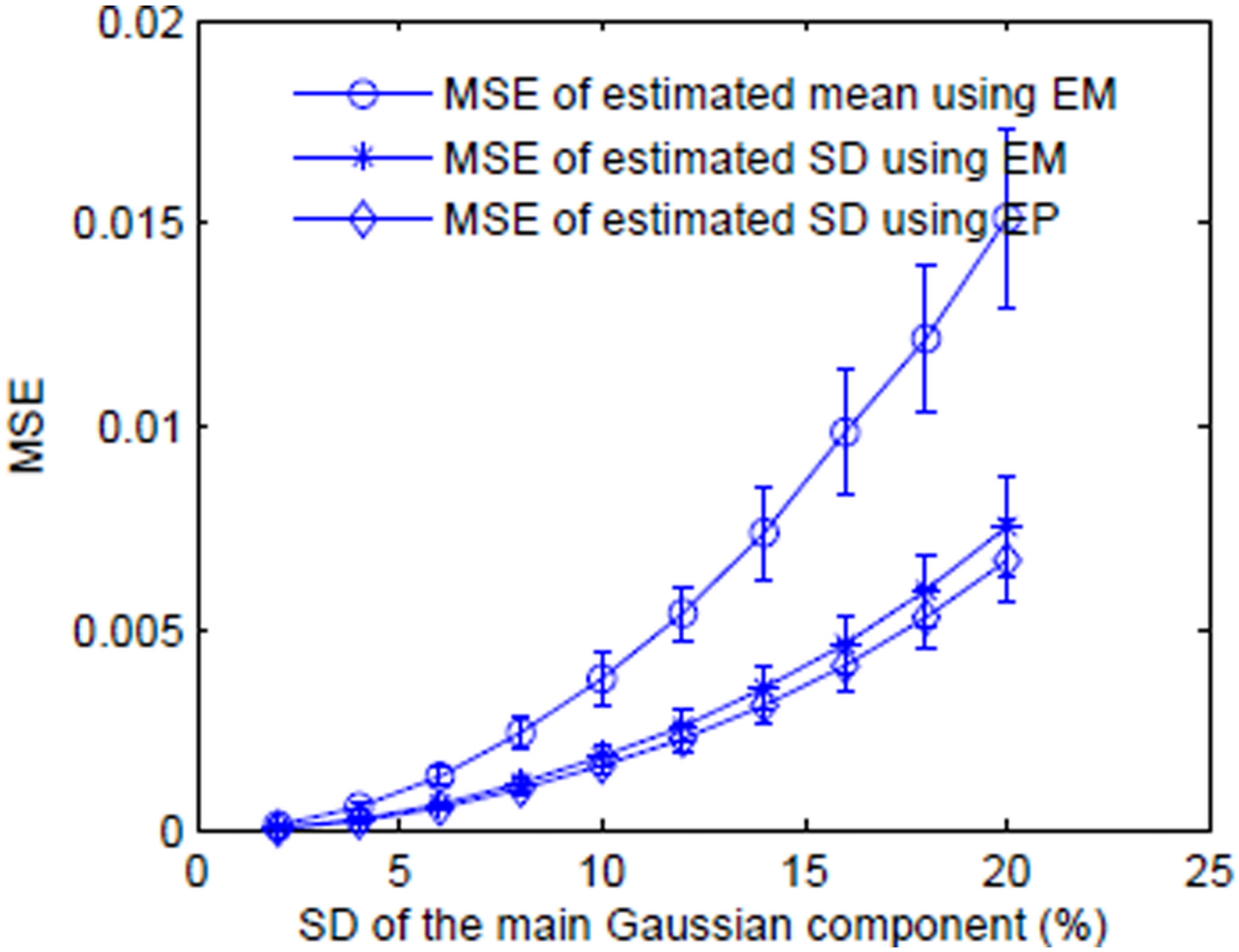

Fig. 8.
SD of the main Gaussian component (%) vs MSE.

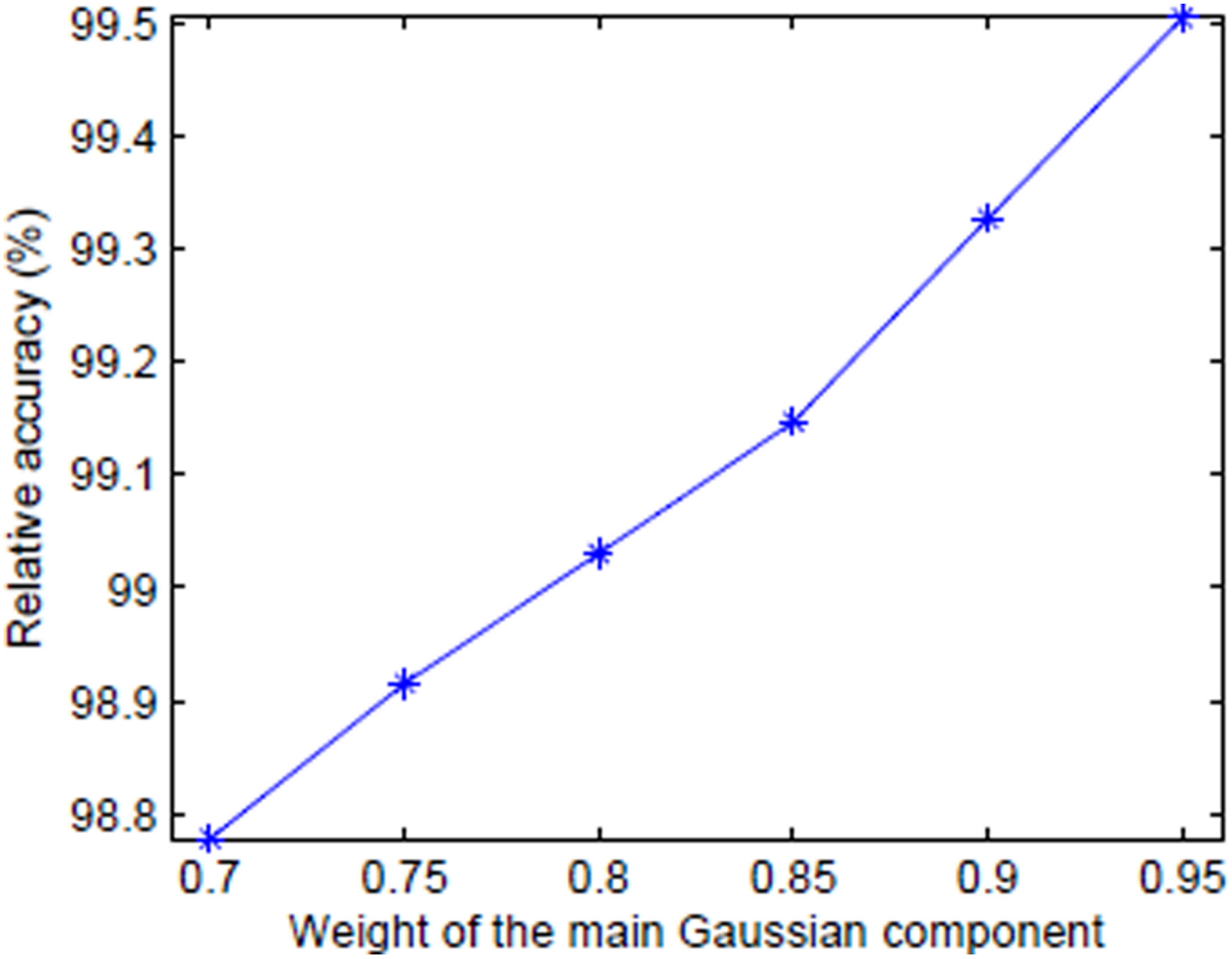

Fig. 9.
Weight of the main Gaussian component vs relative accuracy of monthly energy consumption

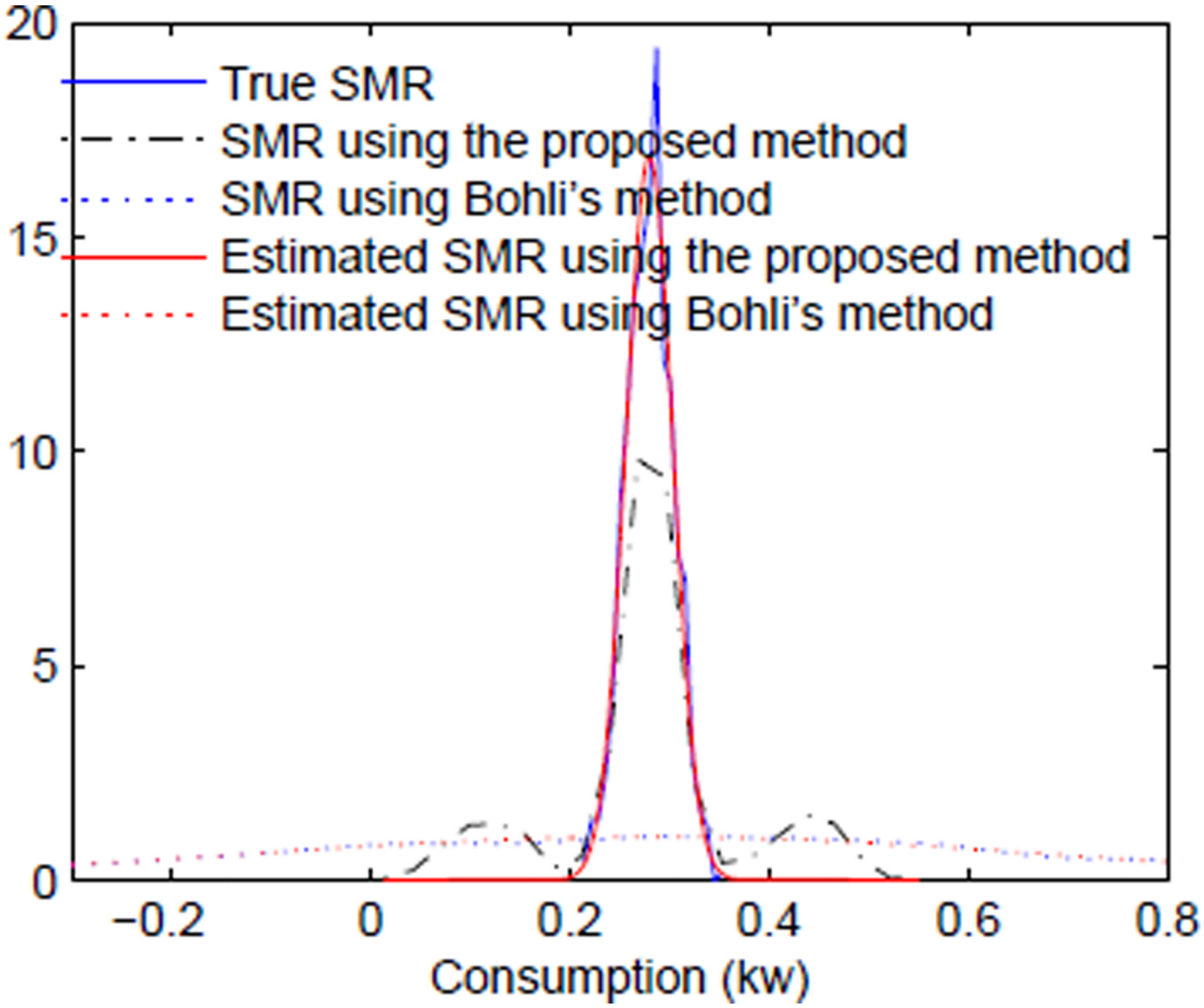

Fig. 10.
Distribution of SMR of a given time tick from a) true report (blue solid line); b) proposed method protected report (black dash-dot line); c) Bohli's method protected report (blue dots line); d) estimated distribution from protected report (red solid line); e) estimated distribution from Bohli's method protected report (red dots line).

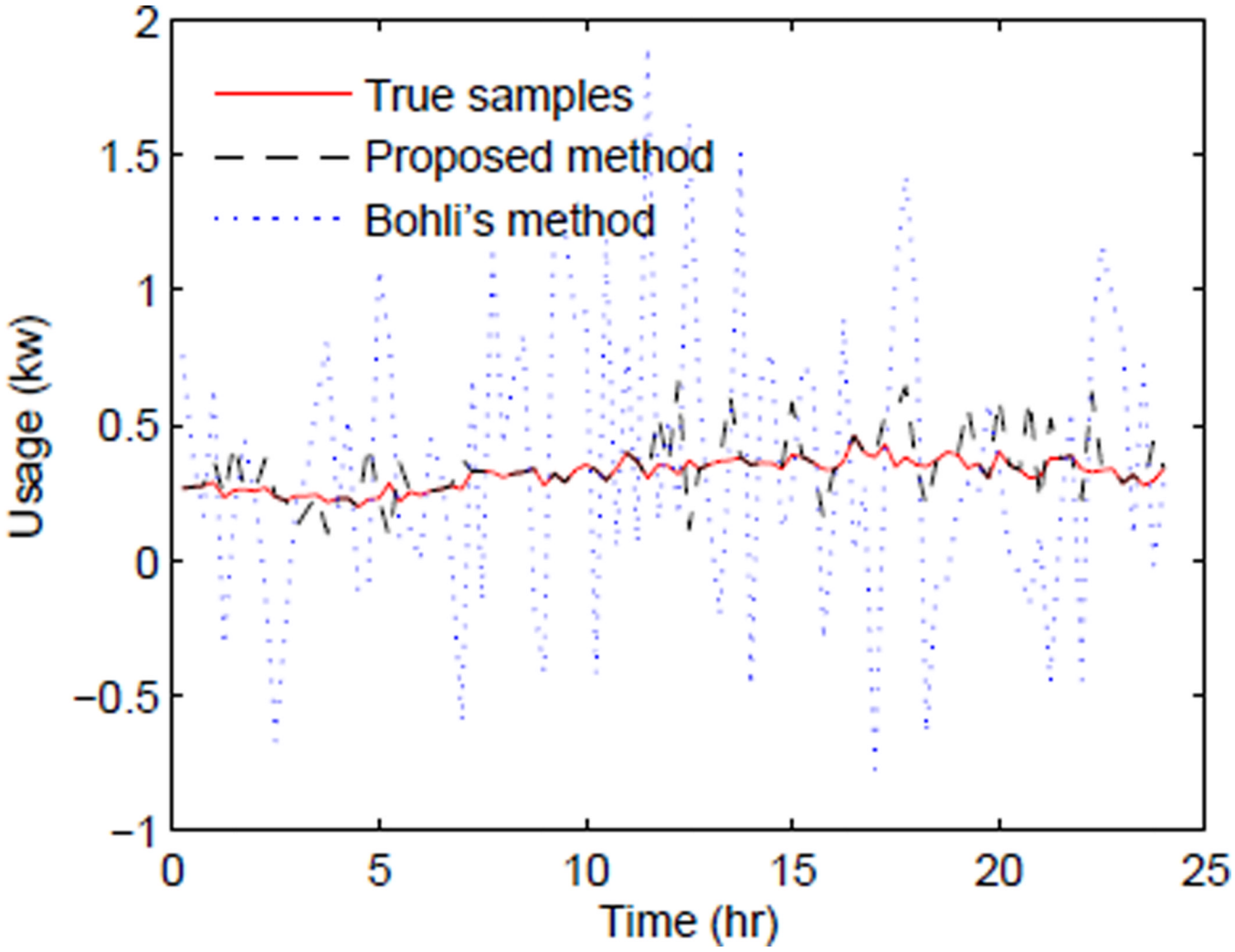

Fig. 11.
SMR of a given user from a) true report (red solid line); b) proposed protected report (black dash line); c) Bohli's method protected report (blue dots line).