From Robust Chip to Smart Building: CAD Algorithms and Methodologies for Uncertainty Analysis of Building Performance

(Invited Paper)

Xiaoming Chen, Xin Li
Electrical and Computer Engineering Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
Email: {xchen3,xinli}@ece.cmu.edu

Sheldon X.-D. Tan
Department of Electrical and Computer Engineering
University of California, Riverside
CA 92521, USA
Email: stan@ece.ucr.edu

***Abstract*—Buildings consume about 40% of the total energy use in the U.S. and, hence, accurately modeling, analyzing and optimizing building energy is considered as an extremely important task today. Towards this goal, uncertainty/sensitivity analysis has been proposed to identify the critical physical and environmental parameters contributing to building energy consumption. In this paper, we propose to apply sparse regression techniques to uncertainty/sensitivity analysis of smart buildings. We consider the orthogonal matching pursuit (OMP) algorithm as a case study to demonstrate its superior efficacy over other conventional approaches. Experimental results reveal that OMP achieves up to 18.6× runtime speedups over the conventional least-squares fitting method without surrendering any accuracy.**

I. Introduction

Buildings, including residential and commercial buildings, consume more energy than the transportation and industry sectors in the U.S., accounting for about 40% of the total U.S. energy consumption [1]. The global energy demand of buildings keeps upward in recent years, mainly due to the growth in the population and the increasing demand for building services and comfort levels [2]. For this reason, energy efficiency of buildings has become a burning issue for the international community. Consequently, it is critical to reduce the energy demands of buildings by adopting a number of "smart" tools for modeling, analysis, optimization and control.

Nowadays, building simulation is widely used at different stages to predict the thermal performance and energy consumption, especially for improving the energy efficiency at the design stage [3]. Typically, the inputs of a building simulation program include the building system and components, the climate, internal gains from lighting, equipments and occupants, heating and cooling systems, schedules of occupants, equipments and lighting, etc. The simulator runs the heat balance algorithm and then predicts the energy and temperature of the building [4]. Through building simulation, we can achieve many advantages for designing energy-efficient smart buildings. For instance, strategies for improving the energy efficiency can be conveniently investigated. Building simulation has become an essential decision support tool for designing smart buildings.

When predicting the energy and temperature by building simulation, there are many uncertainties associated with a smart building [5], including physical scenarios, design and algorithm uncertainties, etc. It is important to quantify the impact of these uncertainties on the energy and thermal performance by a probabilistic approach. Such an uncertainty analysis is usually accompanied with a sensitivity analysis, which aims to identify the most important parameters contributing to the uncertainties of the building performance [6]. Thus, we can obtain the key parameters that influence the building performance most.

In the literature, there are two broad categories of approaches to perform uncertainty/sensitivity analysis: (i) local analysis and (ii) global analysis [7]. Local analysis focuses on the effects of uncertain parameters around their nominal values, whereas global analysis focuses on the influences of uncertain parameters over the entire variation space.

The global approach is often considered to be more reliable than the local approach in practice. Conventional global approaches include linear regression methods [5], [8]–[10], macroparameter-based method [11], grouping-based methods [12], [13], etc. Among them, regression methods are widely used. Since there are a large number of parameters that can influence the building performance, regression methods require a large number of simulation samples to train a high-dimensional regression model, resulting in expensive computational cost. Although grouping-based and macroparameter-based methods aim to reduce the number of samples for building performance modeling, they require priori knowledge to construct the groups or macroparameters. However, the priori knowledge may be unavailable or incorrect in practice, as will be demonstrated by our experimental example in Section IV-B.

In most practical applications, there are only a small number of parameters that can greatly influence the building perfor-

978-1-4673-8388-2/15/$31.00 ©2015 IEEE

mance. In other words, although there are a large number of parameters in total, many of them do not contribute to the building performance variation. As a result, most coefficients of the regression model are close to zero, rendering a unique sparse structure.

Motivated by this observation, we propose to adopt sparse regression methods [14]–[23] for uncertainty/sensitivity analysis of smart buildings. Sparse regression aims to solve a large number of model coefficients from a small set of simulation samples without over-fitting, by exploiting the sparsity of model coefficients. In this paper, we will take the orthogonal matching pursuit (OMP) algorithm [14], [18], [22], [23], one of the well-known sparse regression techniques, to demonstrate its superior efficacy for smart building applications. As will be demonstrated by a case study in Section IV, OMP can achieve up to 18.6× runtime speedups over the conventional least-squares fitting method for uncertainty/sensitivity analysis of buildings.

The remainder of this paper is organized as follows. In Section II, we present the background knowledge about modeling and simulation of smart buildings. We discuss the sparse regression algorithm for uncertainty/sensitivity analysis in Section III. The efficiency of the sparse regression algorithm is demonstrated by a case study in Section IV. Finally, we conclude in Section V.

II. Smart Building

In this section, we will review the background knowledge about smart buildings, covering two important topics: (i) energy consumption and simulation, and (ii) uncertainty/sensitivity analysis.

A. Energy Consumption and Simulation

The energy consumption of a building comes from various building services, including the heating, ventilation, and air conditioning (HVAC) system, lighting, equipments, refrigeration, cooking, etc. Among them, the HVAC system consumes the most energy which can be up to 50% of the total energy use [2]. Many parameters may affect the energy consumption of a building, such as the temperature/weather, properties of materials (i.e., specific heat, thermal resistance, conductivity, thermal absorptance, etc.), user behaviors, power of lighting and equipments, set points of heating and cooling, etc.

Building simulation has been extensively studied since the 1960s. There are many softwares that have been widely used for building performance simulation, including EnergyPlus [24], DOE-2 [25], ESP-r [26], TRNSYS [27], DeST [28], etc. The simulation flows and methodologies of these softwares are similar. Typically, a building simulation software takes a text file as the input which describes the building construction, material details, heating and cooling systems, user behaviors, the weather and climate, etc. A few softwares (like DeST) can also take a database as the input. Next, differential algebraic equations (DAE) are created based on the input information, built-in thermal models, and heat balance. After solving the DAEs by a numerical engine, simulation results, including the energy consumption and temperature, are reported. To accurately predict the building performance, a building is usually partitioned into many zones distributed over different spatial locations. These zones define the spatial resolution of the simulation results reported by the simulation software.

B. Uncertainty/Sensitivity Analysis

In practice, we may not exactly know many important parameters (e.g., temperature/weather, properties of materials, etc.) that affect the energy consumption of a building, especially at an early stage for planning. In this case, these facts must be modeled as a set of random variables, instead of deterministic values. Therefore, the energy consumption and temperature cannot be deterministically solved, posing the need of uncertainty/sensitivity analysis.

Uncertainty/sensitivity analysis aims to determine the performance variation of a building and the key parameters contributing to it. Let $\mathbf{x} = [x_1, x_2, \cdots, x_K]^T$ be the uncertain parameters of a building, where K is the number of these parameters. A specific building performance f is expressed as a function of all these uncertain parameters, i.e.,

$$f = f(x_1, x_2, \cdots, x_K). \tag{1}$$

To calculate the sensitivity of each parameter x_i ($i = 1, 2, \cdots, K$), and, hence, predict the uncertainty of f, we need to collect a number of samples by running building simulation. Typically, an uncertainty/sensitivity analysis flow includes the following three major steps [7].

1) Define K random variables $\mathbf{x} = [x_1, x_2, \cdots, x_K]^T$ and their probability distributions (e.g., normal distributions, uniform distributions, triangular distributions, etc.). Generate a number of (say, N) samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(N)}$.
2) Perform N independent building simulations for these N samples. Collect the energy and/or temperature values (i.e., the values of f) of the N samples from the simulation results.
3) Perform uncertainty/sensitivity analysis by processing the collected data.

As mention in Section I, there are two broad categories of approaches for uncertainty/sensitivity analysis: (i) local analysis and (ii) global analysis. Local analysis approximates the sensitivity of each parameter by a difference quotient [29]:

$$S_i = \frac{f(x_1, \cdots, x_i + \Delta x_i, \cdots, x_K) - f(x_1, \cdots, x_K)}{\Delta x_i}, \tag{2}$$

where S_i is the sensitivity of x_i, and Δx_i is a small perturbation around its nominal value x_i. The local approach is easy to implement and only requires $K + 1$ simulations. However, the local sensitivities only carry the information of $f(x_1, x_2, \cdots, x_K)$ around its nominal value. To overcome this limitation, global analysis is often used in practice.

When applying global analysis, we consider the following linear regression model to approximate the specific building performance f [5], [8]–[10], [16]:

$$f(x_1, x_2, \cdots, x_K) = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_k = \sum_{k=0}^{K} \alpha_k x_k, \quad (3)$$

where $\{\alpha_k; k = 0, 1, \cdots, K\}$ are the model coefficients, and x_0 is always 1. The linear regression model is widely used in the literature. The model coefficients $\boldsymbol{\alpha} = [\alpha_0, \cdots, \alpha_K]^T$ can be determined by solving the following over-determined linear system:

$$\mathbf{X}\boldsymbol{\alpha} = \mathbf{f}, \quad (4)$$

where $\mathbf{X} \in \mathbb{R}^{N\times(K+1)}$ is the regressor matrix containing all the sampling values of uncertain parameters, i.e., $X_{n,k} = x_k^{(n)}$, and $\mathbf{f} = [f^{(1)}, \cdots, f^{(N)}]^T$ is the response vector containing the sampling values of the building performance f. The least-squares fitting method [30] is usually used to solve the over-determined equation Eq. (4) and its solution can be expressed as the following equation in theory:

$$\boldsymbol{\alpha} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{f}. \quad (5)$$

The conventional least-squares fitting method requires that the number of samples (i.e., N) must be larger than the number of unknowns (i.e., $K + 1$). Consequently, a large number of simulation samples must be collected in order to train a high-dimensional regression model where K is large. For our application of building performance analysis, this requirement leads to a prohibitively high computational cost to repeatedly run a large number of building simulations.

To reduce the number of simulation samples required for linear regression, macroparameter-based and grouping-based methods are proposed. The idea of the macroparameter-based method [11] is to lump the uncertain parameters with the same physical meaning and similar sensitivity magnitude into macroparameters. As such, regression model is simplified with reduced dimensionality and, hence, the number of required simulation samples is reduced. The idea of the grouping-based methods [12], [13] is similar. They partition all the uncertain parameters into several groups where the parameters in the same group should share the same sign for sensitivity. Next, an iterative approach is performed where a significant group is selected or an insignificant group is eliminated at each iteration step. As a result, the number of required simulation samples can be greatly reduced, because the number of groups is often substantially less than the number of uncertain parameters.

Both the macroparameter-based and grouping-based methods require priori knowledge to construct the macroparameters or groups. Typically, if no additional prior knowledge is available, grouping the uncertain parameters with the same physical meaning is often considered as an appropriate choice. However, as will be demonstrated by our experimental example in Section IV-B, multiple parameters with the same physical meaning may show significantly different sensitivity values in terms of their signs and/or magnitudes. It, in turn, indicates the fact that appropriately extracting the correct priori knowledge for the macroparameter-based or grouping-based methods is not a trivial task in practice.

III. Uncertainty/Sensitivity Analysis by Sparse Regression

In most practical applications, there are only a small number of important parameters that can significantly affect the building performance. In other words, most components of $\boldsymbol{\alpha} = [\alpha_0, \cdots, \alpha_K]^T$ are close to zero. As a result, we can adopt sparse regression methods to solve the model coefficients from a small set of simulation samples, by exploring the sparsity of $\boldsymbol{\alpha}$. In this paper, we will consider the L_0-norm regularization method [14], [17]–[23] which is one of the well-known sparse regression methods to find the nonzeros of $\boldsymbol{\alpha}$. By applying L_0-norm regularization, the linear system in Eq. (4) can be under-determined. The L_0-norm regularization method determines the solution of Eq. (4) by solving the following optimization problem:

$$\begin{aligned} &\text{minimize} && ||\mathbf{r}||_2 = ||\mathbf{X}\boldsymbol{\alpha} - \mathbf{f}||_2, \\ &\text{subject to} && ||\boldsymbol{\alpha}||_0 \leq \lambda, \end{aligned} \quad (6)$$

where $||\cdot||_0$ and $||\cdot||_2$ are the L_0-norm and L_2-norm of a vector, respectively. L_0-norm means the number of nonzeros in a vector. The L_0-norm $||\boldsymbol{\alpha}||_0$ measures the sparsity of $\boldsymbol{\alpha}$. Therefore, by constraining the L_0-norm of $\boldsymbol{\alpha}$, the optimization in Eq. (6) attempts to find a sparse solution $\boldsymbol{\alpha}$ which minimizes the L_2-norm of the residual vector $\mathbf{r} = \mathbf{X}\boldsymbol{\alpha} - \mathbf{f}$. The optimization problem in Eq. (6) is non-deterministic polynomial-time (NP) hard [23] and, hence, is difficult to solve. In Section III-A, we will describe an efficient heuristic algorithm to solve the optimization problem in Eq. (6) using OMP [14], [18], [22], [23].

The parameter λ in Eq. (6) explores the tradeoff between the sparsity of $\boldsymbol{\alpha}$ and the accuracy of the fitted model. A large λ can result in a small residual. However, a small residual does not necessarily mean a small modeling error, because the model in Eq. (4) may be over-fitted for the given samples. For example, we consider an extreme case where Eq. (4) is under-determined and λ is sufficiently large. In this case, we can always find a solution such that the residual is exactly 0. However, such a solution is likely to be useless because it over-fits the given samples and may present large errors for other samples that are not used for model training. In practice, λ can be optimally determined by the cross-validation technique [16], as will be discussed in Section III-B.

A. Orthogonal Matching Pursuit

The OMP algorithm [14], [18], [22], [23] heuristically solves the optimization problem in Eq. (6) by identifying a small set of important parameters and using them to approximate the performance function f. For other unimportant parameters, the corresponding model coefficients are set to 0. In what follows, we will describe the details of OMP, including a basic theory of parameter selection and an iterative flow to solve the model coefficients.

1) Parameter Selection: The purpose of OMP is to identify a subset of important parameters which significantly impact the performance function f. OMP uses the inner product between the performance function f and parameter x_i to measure the importance of x_i ($i = 0, 1, \cdots, K$). When all the parameters are appropriately normalized and statistically independent, the inner product between f and x_i exactly equals the model coefficient α_i, i.e.,

$$\langle f, x_i \rangle = \sum_{k=0}^{K} \alpha_k \langle x_i, x_k \rangle = \alpha_i. \tag{7}$$

This is the reason why the inner product can be adopted as a good criterion to measure the importance of each parameter. In other words, if $\langle f, x_i \rangle$ is far away from 0, the parameter x_i is highly correlated with the performance function f, hence, it should be selected to approximate the performance function. On the contrary, if $\langle f, x_i \rangle$ is close to 0, the parameter x_i is almost uncorrelated with f, hence, the corresponding model coefficient α_i can be set to 0.

In practice, we do not know the analytical form of f and the inner product in Eq. (7) should be numerically calculated from a set of samples. In this paper, we adopt the Latin hypercube sampling method [31] to generate N random samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(N)}$, based on the probability distributions of the parameters $\mathbf{x} = [x_1, x_2, \cdots, x_K]^T$. Then the inner product is approximated by

$$\langle f, x_i \rangle = \frac{1}{N} \sum_{n=1}^{N} f^{(n)} x_i^{(n)} = \frac{1}{N} \mathbf{X}_i^T \mathbf{f}, \tag{8}$$

where $\mathbf{X}_i \in \mathbb{R}^{N \times 1}$ represents the ith column of the matrix $\mathbf{X}$.

According to Eq. (7) and Eq. (8), we notice that Eq. (8) is a statistical estimator for the model coefficient α_i. Such an estimation is calculated from a set of randomly generated samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(N)}$ and the corresponding responses $\mathbf{f}$, which may carry large fluctuations. For this reason, the estimation in Eq. (8) is only used to measure the correlation between each parameter x_i and the performance function f to select the important parameters. Once these important parameters are chosen, the corresponding model coefficients are solved by least-squares fitting. In addition, to further improve the accuracy of parameter selection, OMP applies an iterative approach where the most important parameter is selected at each iteration step. In what follows, we will describe the iterative algorithm in detail.

2) Iterative Algorithm: OMP applies an iterative algorithm to solve the optimization problem in Eq. (6). At each iteration step, it finds a parameter which is most correlated with the performance function by evaluating the approximate inner product defined in Eq. (8). The iteration continues until λ important parameters have been found. Finally, the performance function f is approximated by these selected parameters.

At the beginning of the iteration process, the first important parameter x_{s1} is selected such that x_{s1} is maximally correlated with the performance function f. Namely, $|\langle f, x_{s1} \rangle|$ takes the largest value over the set $\{|\langle f, x_k \rangle|; k = 0, 1, \cdots, K\}$, where the inner product is estimated by Eq. (8). Once x_{s1} is chosen, OMP approximates f by using x_{s1}, i.e.,

$$f \approx \alpha_{s1} x_{s1}, \tag{9}$$

where the model coefficient α_{s1} is determined by solving the following least-squares fitting problem:

$$\underset{\alpha_{s1}}{\text{minimize}} \ ||\alpha_{s1} \mathbf{X}_{s1} - \mathbf{f}||_2. \tag{10}$$

Next, OMP removes the component $\alpha_{s1} x_{s1}$ from the performance function f and calculates the residual, i.e.,

$$\mathbf{r} = \mathbf{f} - \alpha_{s1} \mathbf{X}_{s1}. \tag{11}$$

To select the second important parameter, OMP calculates the inner products between the residual $\mathbf{r}$ and all the unselected parameters, and then selects the optimal parameter corresponding to the largest inner product magnitude. Let x_{s2} be the second important parameter selected by OMP. Now, OMP approximates f by both x_{s1} and x_{s2}, i.e.,

$$f \approx \alpha_{s1} x_{s1} + \alpha_{s2} x_{s2}, \tag{12}$$

where the model coefficients α_{s1} and α_{s2} are solved from the following optimization problem:

$$\underset{\alpha_{s1}, \alpha_{s2}}{\text{minimize}} \ ||\alpha_{s1} \mathbf{X}_{s1} + \alpha_{s2} \mathbf{X}_{s2} - \mathbf{f}||_2. \tag{13}$$

Note that in the second iteration step, α_{s1} is re-calculated instead of directly using its previous value calculated in the first iteration step. The value of α_{s1} calculated by Eq. (10) may be different from that calculated by Eq. (13), because the sampled data $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_K$ may not be orthogonal even though the parameters $x_0, x_1, \cdots, x_K$ are statistically independent. Therefore, all the model coefficients corresponding

Algorithm 1 The OMP algorithm.

Input: The linear system $\mathbf{X}\boldsymbol{\alpha} = \mathbf{f}$ and the L_0-norm constraint λ.

Output: The model coefficients $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \cdots, \alpha_K]^T$.

1: Initialize the index set $\mathcal{S} = \emptyset$, and the residual $\mathbf{r} = \mathbf{f}$.
2: **for** $p = 1$ to λ **do**
3: Determine the index s such that $|\mathbf{X}_s^T \mathbf{r}|$ takes the largest value over the set $\{|\mathbf{X}_i^T \mathbf{r}|; i \notin \mathcal{S}\}$. Update the index set $\mathcal{S} = \mathcal{S} \cup \{s\}$.
4: Solve the following least-squares fitting problem:

$$\underset{\alpha_i, i \in \mathcal{S}}{\text{minimize}} \ ||\sum_{i \in \mathcal{S}} \alpha_i \mathbf{X}_i - \mathbf{f}||_2, \tag{14}$$

and then approximate f by the selected parameters, i.e.,

$$f \approx \sum_{i \in \mathcal{S}} \alpha_i x_i. \tag{15}$$

5: Update the residual $\mathbf{r} = \mathbf{f} - \sum_{i \in \mathcal{S}} \alpha_i \mathbf{X}_i$.
6: **end for**
7: Set the model coefficients corresponding to the unselected parameters to 0, i.e., $\alpha_i = 0, i \notin \mathcal{S}$.

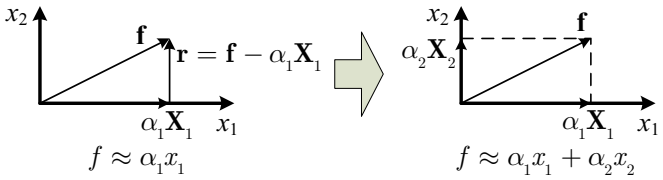


Fig. 1: A simple bivariate example is used to illustrate the OMP algorithm.

to the selected parameters are re-calculated at each iteration step in order to minimize the residual.

The aforementioned iteration process will continue until a desired number (i.e., λ) of parameters have been selected. Algorithm 1 summaries the overall flow of OMP.

To intuitively understand the OMP algorithm, we consider a simple bivariate example in Fig. 1. Assume that $\mathbf{X}_1$ has a stronger correlation with the performance $\mathbf{f}$ than $\mathbf{X}_2$. Hence, x_1 is first selected to approximate f, i.e., $f \approx \alpha_1 x_1$. The residual is $\mathbf{r} = \mathbf{f} - \alpha_1\mathbf{X}_1$, which is orthogonal to $\mathbf{X_1}$, i.e., $\mathbf{X}_1^T\mathbf{r} = 0$. Next, during the second iteration step, since $\mathbf{X}_2$ has a stronger correlation with the residual $\mathbf{r}$ than $\mathbf{X}_1$, x_2 is selected and both x_1 and x_2 are used to approximate f , i.e., $f \approx \alpha_1 x_1 + \alpha_2 x_2$, where the coefficients α_1 and α_2 are both solved by least-squares fitting. The residual $\mathbf{r} = \mathbf{f} - \alpha_1\mathbf{X}_1 - \alpha_2\mathbf{X}_2$ now becomes zero.

B. Cross-Validation

The OMP algorithm (i.e., Algorithm 1) requires the input parameter λ to constraint the L_0-norm of $\boldsymbol{\alpha}$, i.e., the number of nonzeros in $\boldsymbol{\alpha}$. In practice, λ is not known in advance. The value of λ must be carefully determined by considering the following two important facts. First, if λ is too small, only a small number of parameters are selected to approximate the performance function, thereby resulting in large modeling error. On the other hand, if λ is too large, it will lead to over-fitting which also results in an inaccurate regression model. Consequently, in order to create a regression model with the minimum modeling error, we should accurately estimate the modeling error for different λ values and then select the optimal λ such that the modeling error is minimized.

In practice, the modeling error must be accurately estimated from a limited number of given samples. The samples used for model training cannot be used for error estimation; otherwise over-fitting cannot be detected. In other words, the samples used for model training and error estimation should be different and independent. Cross-validation [16] is an efficient method for model validation that has been widely used in the statistics community. In this paper, we use a 5-fold cross-validation, as shown in Fig. 2. The given samples are partitioned into 5 groups. The modeling error is estimated by 5 independent runs. In the ith ($1 \leq i \leq 5$) run, the ith group is used for error estimation and the other 4 groups are used to train the model coefficients. Therefore, we will get 5 modeling errors from 5 runs. The final modeling error is simply the mean value of the 5 individual errors.

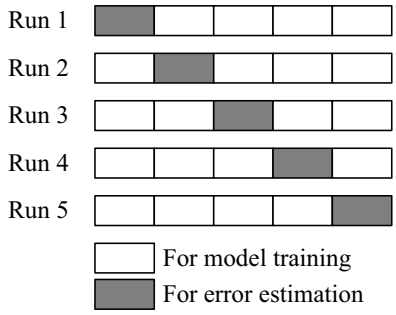


Fig. 2: A five-fold cross-validation is used for error estimation.

For our application of uncertainty/sensitivity analysis for building performance, we first perform cross-validation to determine the optimal value of λ based on a number of given samples. The OMP algorithm is used to train the model coefficients for different λ values during each cross-validation run. The optimal λ is then selected such that the modeling error is minimized. After the optimal λ is known, we run the OMP algorithm again with all available samples to generate the final regression model.

Cross-validation is time-consuming, because we need to run the OMP algorithm for multiple times to determine the optimal λ. However, in our application, the computational cost is dominated by the building performance simulation which is used to generate the required samples. Consequently, the computational cost of cross-validation is negligible, as will be demonstrated by the case study in Section IV-C.

IV. Case Study

A. Simulation Setup

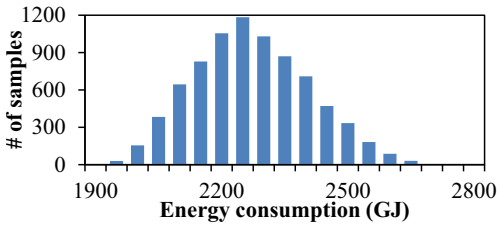
In this section, we will use a building example to demonstrate the efficiency of the aforementioned OMP algorithm for uncertainty/sensitivity analysis of building performance. A building with 10 storeys is created. Each storey has 9 zones so there are 90 zones in total. When running building simulation, we assume that each zone has its own temperature and there is no temperature variation within a single zone.

In this example, 1106 parameters are used to define the uncertainties associated with temperature/weather, properties of materials, etc. The probability distributions of these uncertainties are defined in Table I. Here, we consider both global and local uncertainties for the properties of all materials. For a given physical parameter, all the 10 storeys share the same global variation, and each storey has its own local variation.

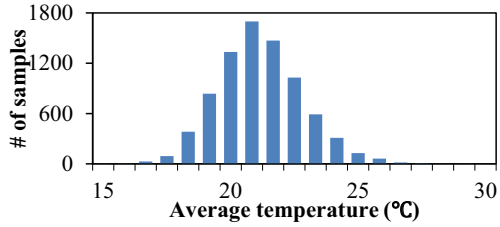
We adopt the Latin hypercube sampling method [31] to generate 8000 independent random samples where 6000 samples are used for model training and the other 2000 samples are used for error estimation. The building is simulated using EnergyPlus [24] on a desktop with a 4-core Intel i7 CPU and 16GB memory. Parallel simulation with eight processes is enabled to reduce the simulation cost. In what follows, we will compare the OMP algorithm with the conventional least-

TABLE I: Uncertain parameters of the building.

Parameter	# of parameters	Distribution
Ground temperature	12	Gaussian
Heating set point	1	Uniform
Cooling set point	1	Uniform
Lighting power per area	1	Uniform
Equipment power per area	90	Gaussian
Equipment activity factor	90	Uniform
Furniture surface area	90	Uniform
Thickness of materials	111	Gaussian
Conductivity of materials	111	Gaussian
Density of materials	111	Gaussian
Specific heat of materials	111	Gaussian
Thermal absorptance of materials	122	Gaussian
Solar absorptance of materials	122	Gaussian
Visible absorptance of materials	122	Gaussian
Thermal resistance of materials	11	Gaussian



(a) Histogram of the energy consumption.



(b) Histogram of the average temperature.

Fig. 3: Histograms of the energy consumption and the average temperature of the 66th zone in April are shown to illustrate their uncertainties.

squares (LS) fitting method in terms of both the modeling accuracy and computational cost.

B. Energy and Temperature Variability

In this example, we consider two performance metrics of interest: (i) the annual energy consumption of the building, and (ii) the average temperature of the 66th zone in April. Fig. 3 shows the histograms of the energy consumption and the average temperature. The histograms are generated from all 8000 samples.

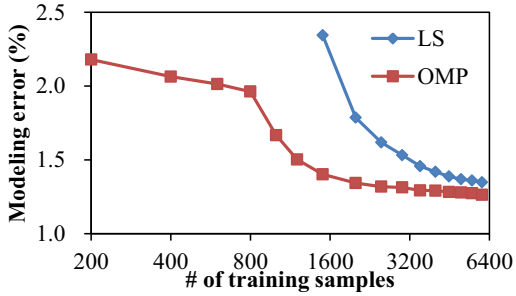
Note that substantial variations can be observed for both the energy consumption and the average temperature. The maximum energy consumption is about 50% larger than the minimum energy consumption. For the average temperature, the variation is even more significant. The maximum average temperature is nearly twice of the minimum average temperature. Consequently, it is critical to know the most important parameters that can significantly affect these performance metrics. As such, building designers are able to optimize the energy consumption and the average temperature to create energy-efficient smart buildings.

To fully understand the limitations of the macroparameter-based method [11] and the grouping-based methods [12], [13], we consider the example of average temperature here. We find that multiple parameters with the same physical meaning can show significantly different sensitivity magnitudes for the average temperature. When creating the temperature model for a specific zone (say, the zth zone), the parameters of the zth zone and its adjacent zones will have significant contributions, but those of other zones will have negligible effects. It, in turn, leads to significantly different sensitivity magnitudes for multiple parameters with the same physical meaning.

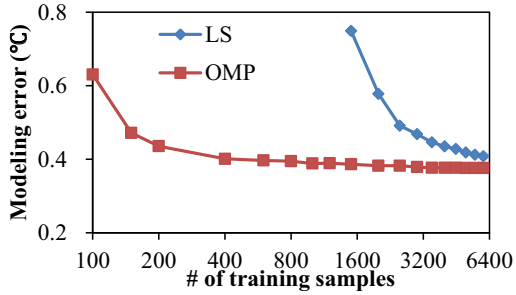
For example, when modeling the temperature of the 66th zone and considering the parameter of the equipment power of each zone, the sensitivity magnitudes of the 63rd to 67th zones (i.e., the adjacent zones) are all larger than 0.09. For other zones, the sensitivity magnitudes are substantially less than 0.09. A few other parameters, e.g., the conductivity and the thermal absorption of the floor, also show a similar pattern. These examples demonstrate an important fact that appropriately constructing groups or macroparameters requires extensive design knowledge and, hence, is not a trivial task in general.

C. Comparison on LS and OMP

Fig. 4 shows how the modeling error varies with the number of training samples. For both LS and OMP, the modeling error decreases as the number of samples increases. However,



(a) Modeling error of the energy consumption.

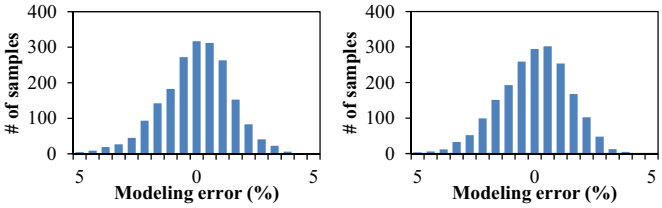


(b) Modeling error of the average temperature.

Fig. 4: The modeling error is shown as a function of the number of training samples.

TABLE II: Comparison on modeling error and cost.

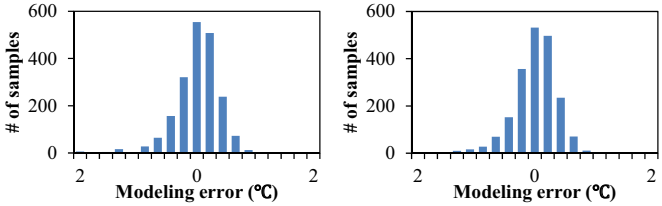
	Energy consumption		Average temperature	
	LS	OMP	LS	OMP
# of samples	3100	1200	5500	300
Modeling error	1.51 %	1.50 %	0.41 °C	0.41 °C
Simulation time	24.1 h	9.3 h	42.8 h	2.3 h
Fitting time	7.3 s	61.0 s	11.0 s	0.2 s
Total cost	24.1 h	9.3 h	42.8 h	2.3 h



(a) Modeling error of LS with 3100 training samples

(b) Modeling error of OMP with 120 training samples

Fig. 5: Histograms of modeling error are estimated from the testing samples for energy consumption.



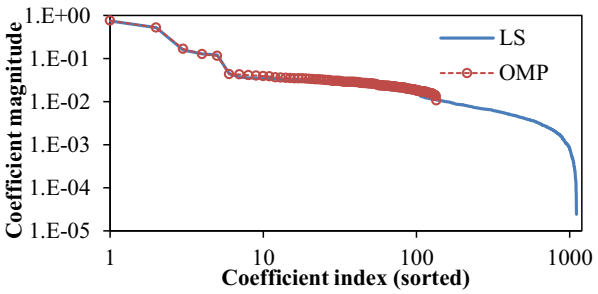
(a) Modeling error of LS with 5500 training samples

(b) Modeling error of OMP with 300 training samples

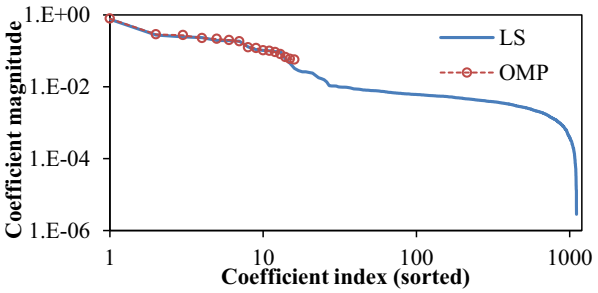
Fig. 6: Histograms of modeling error are estimated from the testing samples for average temperature.

given the same number of samples, OMP is able to achieve substantially higher accuracy than LS, especially if only few samples are available. Table II further compares the modeling error and cost between LS and OMP. The overall modeling cost includes two parts: (i) simulation time and (ii) fitting time. As shown in Table II, the overall modeling cost is dominated by the simulation time. For modeling the energy consumption and the average temperature, OMP achieves 2.6× and 18.6× runtime speedups over LS respectively, without surrendering any modeling accuracy. Fig. 5 and Fig. 6 further plot the histograms of the modeling error for both LS and OMP. These two figures further demonstrate that OMP requires less training samples than LS in order to achieve the same modeling accuracy.

Fig. 7 shows the magnitude of the model coefficients estimated by LS and OMP for the energy consumption and the average temperature, respectively. Studying Fig. 7, we notice that even though there are 1106 uncertain parameters in total, only 135 parameters contribute to the variation of energy consumption and 16 parameters contribute to the variation of average temperature. These important parameters are automatically identified by OMP to accurately model the performance metrics of interest. The underlying sparse structure of these model coefficients is essentially the necessary condition that makes the proposed OMP technique applicable to this example.



(a) Model coefficients for energy consumption.



(b) Model coefficients for average temperature.

Fig. 7: The model coefficients estimated by LS and OMP are compared to demonstrate that OMP accurately capture the dominant coefficients with large magnitude.

V. Conclusion

Nowadays, uncertainty/sensitivity analysis of building performance has been widely used when designing energy-efficient smart buildings. Conventional techniques often require to repeatedly run a large number of building simulations to extract the uncertainty/sensitivity information, thereby resulting in expensive computational cost. In this paper, we take the OMP algorithm, one of the well-known sparse regression techniques, to demonstrate its superior efficacy for uncertainty/sensitivity analysis. Experimental results by a case study show that OMP requires substantially less simulation samples than the conventional LS method in order to achieve the same modeling accuracy, and, hence, the overall modeling cost is significantly reduced (up to 18.6×).

References

[1] Energy efficiency trends in residential and commercial buildings. [Online]. Available: http://apps1.eere.energy.gov/buildings/publications/pdfs/corporate/building_trends_2010.pdf

[2] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy and Buildings*, vol. 40, no. 3, pp. 394–398, 2008.

[3] G. Augenbroe, "Trends in building simulation," *Building and Environment*, vol. 37, no. 8-9, pp. 891–902, 2002.

[4] D. Zhu, T. Hong, D. Yan, and C. Wang, "Comparison of building energy modeling programs: building loads," Lawrence Berkeley National Laboratory, Tech. Rep., 2012.

[5] C. J. Hopfe and J. L. Hensen, "Uncertainty analysis in building performance simulation for design support," *Energy and Buildings*, vol. 43, no. 10, pp. 2798–2805, 2011.
[6] A. Saltelli, S. Tarantola, and F. Campolongo, "Sensitivity analysis as an ingredient of modeling," *Statistical Science*, vol. 15, no. 4, pp. 377–395, 2000.
[7] W. Tian, "A review of sensitivity analysis methods in building energy analysis," *Renewable and Sustainable Energy Reviews*, vol. 20, no. 0, pp. 411–419, 2013.
[8] A. S. Silva and E. Ghisi, "Uncertainty analysis of user behaviour and physical parameters in residential building performance simulation," *Energy and Buildings*, vol. 76, no. 0, pp. 381–391, 2014.
[9] D. G. Sanchez, B. Lacarrire, M. Musy, and B. Bourges, "Application of sensitivity analysis in building energy simulations: Combining first- and second-order elementary effects methods," *Energy and Buildings*, vol. 68, Part C, no. 0, pp. 741–750, 2014.
[10] J. S. Hygh, J. F. DeCarolis, D. B. Hill, and S. R. Ranjithan, "Multivariate regression as an energy assessment tool in early building design," *Building and Environment*, vol. 57, no. 0, pp. 165–175, 2012.
[11] G. C. Rodríguez, A. C. Andrés, F. D. M. noz, J. M. C. López, and Y. Zhang, "Uncertainties and sensitivity analysis in building energy simulation using macroparameters," *Energy and Buildings*, vol. 67, no. 0, pp. 79–87, 2013.
[12] N. Rahni, N. Ramdani, Y. Candau, and P. Dalicieux, "Application of group screening to dynamic building energy simulation models," *Journal of Statistical Computation and Simulation*, vol. 57, no. 1-4, pp. 285–304, 1997.
[13] G. Watson, "A study of the group screening method," *Technometrics*, vol. 3, no. 3, pp. 371–388, 1961.
[14] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, 1993*, Nov 1993, pp. 40–44.
[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
[16] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2005.
[17] X. Li and H. Liu, "Statistical regression for efficient high-dimensional modeling of analog and mixed-signal performance variations," in *Design Automation Conference (DAC) 2008*, June 2008, pp. 38–43.
[18] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4655–4666, Dec 2007.
[19] X. Li, "Finding deterministic solution from underdetermined equation: large-scale performance modeling by least angle regression," in *Design Automation Conference (DAC) 2009*, July 2009, pp. 364–369.
[20] W. Zhang, T.-H. Chen, M.-Y. Ting, and X. Li, "Toward efficient large-scale performance modeling of integrated circuits via multi-mode/multi-corner sparse regression," in *Design Automation Conference (DAC) 2010*, June 2010, pp. 897–902.
[21] X. Li, W. Zhang, and F. Wang, "Large-scale statistical performance modeling of analog and mixed-signal circuits," in *Custom Integrated Circuits Conference (CICC) 2012*, Sept 2012, pp. 1–8.
[22] X. Li, "Finding deterministic solution from underdetermined equation: large-scale performance variability modeling of analog/RF circuits," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 29, no. 11, pp. 1661–1668, Nov 2010.
[23] I. Rish and G. Grabarnik, *Sparse modeling: theory, algorithms, and applications*, 1st ed. Boca Raton, FL, USA: CRC Press, Inc., 2014.
[24] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann, W. Buhl, Y. Huang, C. O. Pedersen, R. K. Strand, R. J. Liesen, D. E. Fisher, M. J. Witte, and J. Glazer, "EnergyPlus: creating a new-generation building energy simulation program," *Energy and Buildings*, vol. 33, no. 4, pp. 319–331, 2001.
[25] DOE-2. [Online]. Available: http://www.doe2.com/
[26] ESP-r. [Online]. Available: http://www.esru.strath.ac.uk/Programs/ESP-r.htm
[27] TRNSYS. [Online]. Available: http://www.trnsys.com/
[28] D. Yan, J. Xia, W. Tang, F. Song, X. Zhang, and Y. Jiang, "DeST-An integrated building simulation toolkit Part I: Fundamentals," *Building Simulation*, vol. 1, no. 2, pp. 95–110, 2008.
[29] C. Spitz, L. Mora, E. Wurtz, and A. Jay, "Practical application of uncertainty analysis and sensitivity analysis on an experimental house," *Energy and Buildings*, vol. 55, no. 0, pp. 459–470, 2012.
[30] A. Charnes, E. Frome, and P.-L. Yu, "The equivalence of generalized least squares and maximum likelihood estimates in the exponential family," *Journal of the American Statistical Association*, vol. 71, no. 353, pp. 169–171, 1976.
[31] M. D. McKay, "Latin hypercube sampling as a tool in uncertainty analysis of computer models," in *Proceedings of the 24th Conference on Winter Simulation*, 1992, pp. 557–564.