How challenging is modeling of a data set?

Davood Shamsi[†] and Farinaz Koushanfar[†‡]
[†]Electrical and Computer Engineering Dept., and [‡]Computer Science Dept.
Rice University, Houston, TX 77005

Abstract

We introduce a novel methodology for determining the difficulty of modeling a given data set. The method utilizes formulation of modeling as an optimization problem instance that consists of an objective function and a set of constraints. The properties of the data set that could affect the quality of optimization are categorized. In large optimization problems with multiple properties that contribute to the solution quality, it is practically impossible to analytically study the effect of each property. A number of metrics for evaluating the effectiveness of the optimization on each data set are proposed. Using the well known Plackett and Burmann fast simulation methodology, for each metric, the impact of the categorized properties of the data are determined for the specified optimization method. A new approach for combining the impacts resulting from different properties on various metrics is described. The method is illustrated on distance measurement data used for estimating the locations of wireless nodes in ad-hoc networks.

I. Introduction

Years of continuous research in statistical modeling and optimization has produced a multitude of readily available methods and tools that could be employed for building models for a given data set [1]. Often times, a new statistical modeling method is theoretically analyzed for meeting an optimality criteria under certain assumptions and/or for its runtime complexity. Many modeling practices today concern a large body of data that does not conform with typical assumptions needed to analytically declare an optimality criteria. In such scenarios, the modeling method is usually evaluated by how it performs on sets of real or simulated data. For example, some statistics of the resulting prediction error and/or a defined criterion (e.g., Bayesian information criterion (BIC)) is used for experimental evaluation of the method on the data. A relevant question to answer is if indeed modeling the pertinent data set requires introduction of a new modeling method, or the data could have been just as well addressed by other known methods. Aside from the theoretical properties, a useful new methodology is the one that can build models for difficult-to-characterize data that is hard to comprehend by other tools and methods.

Addressing the problem is important since it would introduce criteria for quantifying the difficulty of modeling a given data set. This would provide impetus for inventing newer modeling methods and tools that can address the challenging aspect of the difficult-to-characterize data. Simultaneously, formation of new tools would depend upon finding truly challenging data sets that need to be modeled, as opposed to building new models that have a limited practical usage. Formation of sets of challenging data would also build a foundation for comparison among the various modeling methods and algorithms that attempt to model the data set. The problem of finding difficult-to-characterize data is complicated by variations in properties of the underlying data sets collected by different sources. This includes difference in size, format, hidden covariates, and the form of noise present in the collected data. It is not easy to find unique metrics that could be used for comparison of different modeling methods.

To compare various modeling methods that address the same class of problems, the current practice is to use common data. The common data is typically publicly available to download and use by the researchers. Examples of public database for such data includes [2], [3]. Sensitivity of modeling error or other discrepancy metrics to the underlying noise in data has been widely studied for a number of modeling methods [4] [5]. Also, the consistency of estimators based on a number of strong assumptions on the distribution of the data has been pursued [6]. However, no generic method or tool for determining the difficulty in modeling a data set free of imposing strong assumptions – such as Normality or other closed form distributions of noise – is available. Note that the runtime complexity of a modeling method is an orthogonal concept. The complexity measures the worst-case computational time for the algorithm used for finding the model. Analyzing the worst-case runtime complexity does not help in understanding the complexity of characterizing a specific data set.

Finding an appropriate model for a data set is usually accomplished by fitting model parameters on the data such that a measure of accuracy is optimized, e.g., minimizing the mean square error. To analyze the model's performance on the data, we study the modeling optimization problem that consists of an objective function (OF) and a number of constraints. The data set is considered as the input to the optimization problem. We introduce a number of metrics that measure the complexity of the optimization problem based on the OF properties and constraints. The challenge in most optimization problems is the existence of nonlinearities that make the solution space coarse, causing bumpiness and multiple local minimums. We propose a number of measures for the smoothness of the OF and constraints space that estimate the feasibility of reaching the global minimum.

To enable studying the effectiveness of the optimization on a data set, one should characterize the properties of the pertinent data set. The properties are specific to each data set and each problem. In this paper, we focus on the problem of modeling the

location of nodes in a wireless network by using erroneous mutual distance measurements between a number of node pairs. However, we emphasize that our method is generic and can be used for determining the challenge in addressing any data set that includes forming an optimization problem. The location estimation problem is selected for four reasons. First, it is a very well addressed problem in the literature and there are several methods that are developed for this problem [7] [8] [9] [10] [11]. Second, there are a number of publicly available data sets for the measured distance data in the networks [2] [12] [13]. Third, the nonlinear relationship between noise in measurements data and location of nodes makes the modeling problem extremely challenging. Fourth, localization problem is an NP-complete problem, i.e., in the worst case, there is no algorithm that can solve it in polynomial time [14] [7].

We characterize a number of properties of the measurement data set that could affect the quality of location estimation. Studying the interaction between the identified data properties and optimization metrics requires long simulations and analyses. We use the well-known Plackett and Burmann [15] simulation methodology to rapidly study the complex interactions of properties. A new approach for combining the impacts resulting from different properties of data on various optimization metrics is described. The sensitivity of optimization with respect to the various parameter ranks is presented.

To the best of our knowledge, this is the first work that systematically studies the impact of the data set on the optimization problem employed for building statistical models. Most of the previous works are devoted to modeling and analysis of the worst case complexity. The results of our analyses could be directly used for constructing benchmarks for the problem. The proposed work aims at creating a unified framework based on real data that can help evaluation and comparison of desperate efforts that address the same problem.

The remainder of the paper is organized a follows. In the next section, location estimation problem and our notations are formally defined. In Section III, we devise a number of metrics that are used for OF evaluation. The simulation methodology is described in Section IV. In Section V, we illustrate how the results of different metrics can be combined. We have applied the derived method on the measurements from a real network in Section VI. We conclude in Section VII.

II. Preliminaries

In this section, we present the formal definition of the problem. We also describe the notations that are used throughout the paper.

Location estimation problem: Given a set of N nodes denoted by $V = \{v_1, v_2, \dots v_N\}$ in $\mathbb{R}^d$ ($d = 2, 3$). For a given subset of node pairs denoted by $E \subset V \times V$, mutual distance of nodes are measured, *i.e.*, for all $(v_i, v_j) \in E$, $l(v_i, v_j) = d(v_i, v_j) + \epsilon_{i,j}$ is known; $d(v_i, v_j)$ is the Euclidean distance between the nodes v_i and v_j; $\epsilon_{i,j}$ is the distance measurement error. Moreover, there is a subset with $M(> 2)$ nodes denoted by $V_B = \{v_1, \dots v_M\}$, $V_B \subset V$ such that the nodes in V_B have their exact location information (coordinates). The nodes in the set V_B are called the *beacon* nodes.

Question: find the location of all possible nodes.

In this paper, we focus on two-dimensional networks. Extension to three-dimensional networks is straightforward. Coordinates of the node v_i are denoted by (x_i, y_i).

The location estimation problem can be formulated as an optimization problem. The goal is to find the coordinates of $K = N - M$ non-beacon nodes such that the discrepancy (error) between the measured distance data and the nodes' distances estimated from the final coordinates is minimized. In other words,

$$F_L(x_{M+1}, y_{M+1}, x_{M+2}, y_{M+2}, \dots x_N, y_N) = \sum_{(v_i,v_j)\in E} L(e_{v_i,v_j}) \tag{1}$$

$$e_{v_i,v_j} = l(v_i, v_j) - \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Where $L : \mathbb{R} \to \mathbb{R}^+$ is a function that is typically a metric (measure) of error. $F_L : \mathbb{R}^{2K} \to \mathbb{R}^+$ is known as objective function (OF) of the optimization problem.

Note that the OF of the location estimation problem is not necessarily a linear or convex function. There are a number of fast and efficient tools that are developed for linear and convex programming. However, there is no oracle algorithm that can solve all optimization problems. To find the minimum of a nonlinear problem like location estimation, there are a number of heuristic methods that may be employed. The nonlinear system solvers have a tendency to get trapped in a local minimum and do not necessarily lead to the global minimum. Although there are a variety of minimization algorithms, most of them are common in one subcomponent that starts from an initial point and follow the steepest decent to reach the minimum. The algorithms differ in how they choose the starting point, how they select the direction in the search space, and how they avoid local (non-global) minima. Thus, the shape of the OF around the global minimum is an important factor in finding the solution.

Data set: The measurement data used in this problem consists of measured distances between a number of static nodes in the plane. Measurements are noisy; there are multiple measurements for each distance. The true location of the nodes is known and will be known as the ground truth. As explained in Section I, we sample the data set to obtain instances with specific properties.

Parameters: We will define a number of parameters that can be extracted from the data set. The sensitivity of the location estimation to the variations in each parameter will be studied.

We study the effect of different parameters in the location estimation problem and identify the hard instances of measurement data. Ten parameters are studied:

- P_1 – *Number of nodes* (N): the total number of nodes in the network.
- P_2 – *Number of beacons* (B): the number of beacon nodes with known locations.
- P_3 – *Mean squared error* ($\overline{\epsilon^2}$): mean squared error of distance measurements.
- P_4 – *Maximum allowed squared error* ($\text{MAX}_{\epsilon_m^2}$): the maximum squared error that can possibly exist in distance measurements.
- P_5 – *Percentage of large errors* ($PER_{\epsilon_0^2}$): percentage of squared distance measurement noises that are higher than a specific value ϵ_0^2.
- P_6 – *Mean degree* ($\overline{D}$): mean degree of the nodes in the network. Degree of a node v_i is define as number of nodes that know their mutual distance to v_i.
- P_7 – *Minimum length* (MINL): possible minimum length of the measured distances between nodes in the network.
- P_8 – *Maximum length* (MAXL): possible maximum length of the measured distances between nodes in the network.
- P_9 – *Mean length* ($\overline{l}$): mean length of the measured distances between nodes in the network.
- P_{10} – *Minimum degree* (MIND): a lower bound on the possible minimum degree of the nodes in the network.

To study the effect of the parameters, we construct a variety of network instances with different properties. The networks are constructed by selecting subsets of an implemented network. Having specific values for parameters, we use Integer Linear Programming (ILP) to extract each subset such that it meets specified properties. To do so, we model parameter constraints as linear equalities and inequalities. Some parameters such as the mean squared error, $\overline{\epsilon^2}$, can be easily stated by linear equalities and inequalities. But some parameters such as the mean degree of the nodes, $\overline{D}$, need a mapping to be stated in linear terms. The description of the exact procedure of modeling by linear constraints is beyond the scop of this paper [13].

III. Metrics

In this section, we introduce metrics for error and OF that are used for evaluating the importance of different parameters for location estimation. Three error metrics and four OF metrics are presented. Thus, a total of twelve combined metrics are used to evaluate the importance of parameters.

A. Error Metrics

The three error metrics studied in this paper are: L_1, L_2, and the maximum likelihood (ML). L_1 and L_2 are the common error norms in the L_p family defined as:

$$L_p(e_{v_n,v_m} \in E) = (\sum_{(v_n,v_m)\in E} |e_{v_n,v_m}|^p)^{1/p} \quad if \quad 1 \leq p < \infty.$$

To find the error metric corresponding to ML, we need to model the noise in distance measurements. To model the noise, the probability density function (PDF) of errors, f_m, for the distance measurements should be approximated. Different methods are developed to approximate PDF of noise, f_m [13]. We have used kernel fitting that is a simple and known PDF approximation method [1]. To have the maximum likelihood estimation for the nodes' locations, we find the nodes' coordinates such that they maximize

$$\prod_{(v_n,v_m)\in E} f_m(e_{v_n,v_m}) = \exp\{\sum_{(v_n,v_m)\in E} \ln(f_m(e_{v_n,v_m}))\} \tag{2}$$

or equivalently minimize

$$\sum_{(v_n,v_m)\in E} -\ln(f_m(e_{v_n,v_m})). \tag{3}$$

Note that we assume noise in distance measurements are independently identically distributed. Following the same notations as Equation 1 for Equation 3, for the ML estimation we consider the following error metric:

$$L_{ML}(e_{v_n,v_m}) = -\ln(f_m(e_{v_n,v_m})). \tag{4}$$

B. Objective Function (OF) Metrics

We describe metrics that are used for evaluating OFs. The metrics are introduced based on the properties of OF that are effective in optimization. These metrics are such that they assign larger values to the more difficult-to-optimize OFs. In defining the OF metrics, we assume that there is a fixed instance of location estimation data. Thus, for a fixed error metric, the OF would be fixed. Metrics of OF are denoted by $M : \mathcal{C} \rightarrow R^+$ where $\mathcal{C}$ is the functional space that contains all OFs.

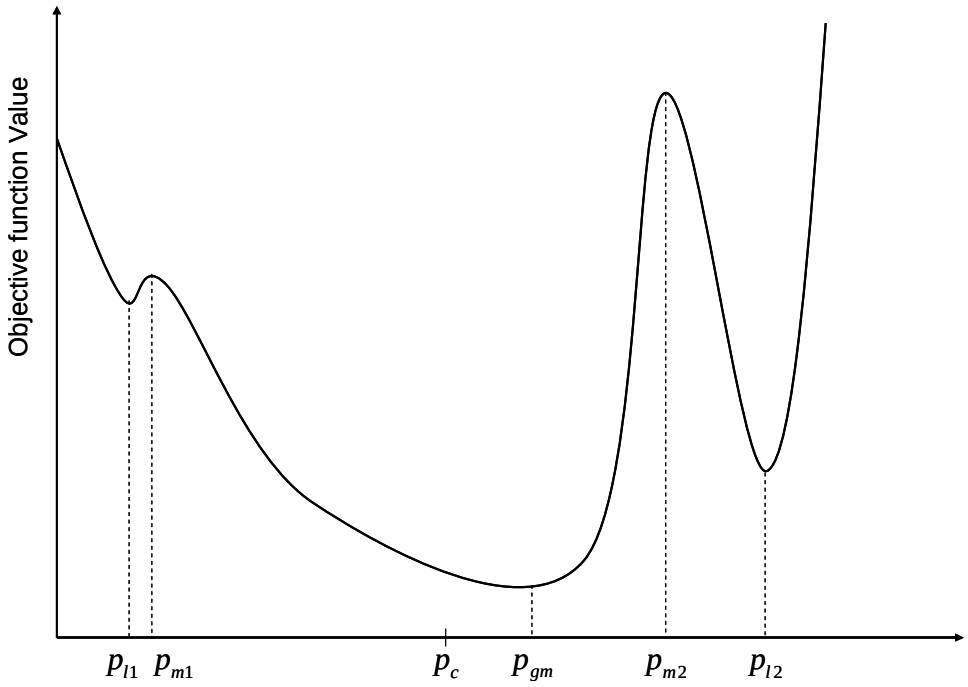

Fig. 1. Metrics and objective function (OF).

1) Drifting of Objective Function (OF) : Since there is noise in distance measurements, true location of the nodes is often not the global minimum of the OF. Location of the OF's global minimum is a measure of the goodness of the OF. Figure 1 illustrates the effect of noise on the OF. For simplicity of presentation, a one-dimensional OF is shown. In this figure, p_c is the correct nodes' location. However, the global minimum of the OF is drifted to p_{gm} because of the noise. We consider the distance between p_c and p_{gm} as an OF metric and denote it by *drifting*.

To find the drifting distance, we start from the true locations as the initial point. Next, the steepest descent direction of the OF is followed until a local minimum is reached. The Euclidean distance between the true locations and this local minimum quantifies the drifting metric (denoted by M_1) for the pertinent OF.

2) Nearest Local Minimum : Having a number of local minimums around the global minimum in an OF may cause the optimization algorithm to get trapped in one of the non-global local minimums. It is challenging to minimize such an OF since the global minimum is hard to reach. Figure 1 illustrates the phenomena. The OF has multiple local minimums at points p_{m1}, p_{m1} and so on. The steepest decent method leads to the global minimum if and only if we start from a point between p_{m1} and p_{m2}. Hence, having a small distance between p_{m1} and p_{m2} would complicate the selection of the initial starting point.

We introduce a method to measure the distance of the true location from the local minimums around the global minimum. Because of curse of dimensionality, it is not possible to find all the local minimums around the global minimum. We randomly sample the OF in multiple directions. The nearest local minimum is computed for each randomly selected direction. We statistically find the distance to the nearest local minimum by using multiple samples.

Assume $F : \mathbb{R}^{2K} \rightarrow \mathbb{R}^{+}$ is the OF. A random direction in $\mathbb{R}^{2K}$ is a vector in this space. Let us denote it by $v \in \mathbb{R}^{2K}$. First, we define a new function $h : \mathbb{R}^{+} \rightarrow \mathbb{R}^{+}$ such that $h(t) = F(p_c + tv)$ where p_c is a vector containing the true locations of nodes. Second, we find the local minimum of h with the smallest positive t and denote it by t_1. We repeat this procedure for T times and find all t_i's. T is the number of samples. Finally, since it is expected that the defined metric has a larger value for more difficult-to-optimize OF, we define the nearest local minimum metric to be

$$M_2(F) = \left(\frac{1}{T} \sum_{i=1}^{T} t_i \right)^{-1}. \tag{5}$$

3) Measuring the Slope of OF Around the Solution : The slope of OF, *i.e.*, the norm of OF's gradient, around the global minimum is a very important parameter in the convergence rate of the optimization algorithm. OFs with a small slope around the true location converge to the global minimum very slowly.

Thus, measuring the slope of the OF around the global minimum can be used to quantify the goodness of OF. Again, we measure slope of the OF in multiple random directions around the true locations, and statistically compute this metric. OFs with sharp slopes around the global minimum are easier to optimize. This can be seen in Figure 2 where the right side of the global minimum, p_{gm}, has a sharp slope. If the initial point of steepest descent algorithm is between p_{gm} and p_{m2}, it

converges to the global minimum very fast. However, on the left side of global minimum, p_{gm}, there is a gradual slope. Thus, the steepest descent algorithm would converge very slowly on the left side. We define the true locations' slope metric as

$$M_3(F) = \left(\frac{1}{T} \sum_{i=1}^{T} \text{slope in i-th random direction} \right)^{-1}. \quad (6)$$

Note that the slope of the i-th random direction, v_i, is measured at $p_{gm} + \sigma v_i$ where σ is a small number and is a user's defined criterion.

4) Depth of the Non-Global Local Minima: Optimization problems that have an OF with deep local minimums around the global minimum are difficult to solve. A number of heuristic optimization methods take advantage of the shallow local minimums to avoid non-global local minimums, e.g., simulated annealing [16]. In figure 2, avoiding the local minimum at p_{l1} is much easier than local minimum at p_{l2}.

We define the third metric for quantifying the goodness of an OF on the data, as the depth of the non-global local minimums. We randomly select T local minimums around the true locations. Assuming that m_i is the OF value at the randomly selected local minimums, define

$$M_4(F) = \left(\frac{1}{T} \sum_{i=1}^{T} m_i \right)^{-1}. \quad (7)$$

IV. Simulation methodology

To find the effect of each parameter, we study all combinations of parameters. Assume each parameter has just two values. If we have k parameters then we have to study 2^k combinations that is computationally intractable. Instead, we use Plackett and Burman (PB) [15] fast simulation methodology that is a very well known method for reducing the number of simulations. Number of simulation in PB is proportional to the number of parameters.

In PB design, two values are assigned to each parameter: a normal value and an extreme value. The normal value is the typical value of the parameter while the extreme value is the value that is outside the typical range of the parameter. The extreme value often makes the problem either harder or easier to solve. A number of experiments with normal and extreme values of parameters are conducted.

Experiments are arranged based on a given matrix denoted by the *design matrix*. Design matrix has k columns (k is the number of parameters) and s rows where s is the number of experiments the should be set up as follows. The elements of the design matrix are either 0 or 1. We set up an experiment for each row. Values of the parameters depend on the elements on the row: 0 indicates that the normal value of the parameter is used and 1 indicates that the extreme value of the parameter is used in the experiment corresponding to the row.

Assume that we have selected an error metric, L_i, and an objective function metric, M_j. The OF itself denoted by F_{L_i} would be fixed. For each row of the design matrix, h, we setup an experiment based on the elements of that row and measure the goodness of the objective function $M_j(F_{L_i})$ and save it in another array element denoted by $r_{i,j,h}$. The corresponding values are summed up for computing the importance factor (IF) of each parameter. For each parameter P_t, we define

$$\text{IF}_{t,i,j} = |\sum_{h=1}^{s} \alpha_{h,t} r_{i,j,h}| \quad (8)$$

where s is the number of experiments (number of rows in the design matrix), and $\alpha_{h,t}$ is 1 if the extreme value of the parameter P_t is used in the h-th experiment; otherwise, $\alpha_{h,t}$ is -1. The absolute value of IF is used to evaluate the effect of each parameter. The largest value indicates the most important parameter. For i-th error metric and j-th OF metric, $\text{IF}_{t,i,j} > \text{IF}_{u,i,j}$ means that the parameter P_t is more important than P_u. Thus, for each error metric, L_i, and for each objective function metric, M_j, we can rank parameters based on their effect on the estimated location. This ranking is denoted by $R_{i,j}$.

More precise result can be obtained by using the foldover design matrix [17]. In the foldover design matrix, all rows of the single design are repeated after its last row but 0s and 1s are exchanged in the repeated rows.

V. combining different ranks

In this section, we explain how to combine the rankings of the parameters under study to obtain a global order for them. Using the ranking method in the previous section, we would have different rankings for various error metrics and OF metrics. Since there are three error metrics and four OF metrics, there would be twelve different importance ranking lists of parameters; each parameter may have a different rank in each ranking list.

Each rank is obtained based on a specific property of the optimization problem. As it is explained in Section III, for each error and OF metric, the parameters are ranked based on the importance factor obtained from PB-design. IFs with large discrepancies lead to a stronger ranking compared to IFs with small discrepancies.

Parameter	N_S	B_S	$\overline{\epsilon_S^2}$	MAX$_{\epsilon_m^2}$	$PER_{\epsilon_0^2}$	$\overline{D_S}$	MINL$_S$	MAXL$_S$	$\overline{l_S}$	MIND$_S$
Normal Value	55	12	10 (m^2)	200 (m^2)	50	10	5 (m)	40 (m)	20 (m)	4
Extreme Value	80	3	50 (m^2)	500 (m^2)	20	6	10 (m)	60 (m)	30 (m)	3

TABLE I
NORMAL AND EXTREME VALUES FOR THE PARAMETERS.

For each ranking, $R_{i,j}$, and for each pair of parameters, P_s, P_t, we find the probability that P_s is more important than P_t. Based on the probabilities, we construct the global ranking.

Consider a specific error metric, L_i, and a specific objective function metric, M_j. Assume that the importance factor of the parameter P_t, IF$_{t,i,j}$, is normally distributed $\mathcal{N}(\lambda_{t,i,j}, \sigma^2)$. The observed value of IF$_{t,i,j}$ in a specific experiment is denoted by $if_{t,i,j}$. We normalize the importance factors to have a maximum value W. The mean of IFs are assumed to be uniformly distributed in $[0, W]$.

For each two parameters, P_s and P_t, given the BP-design experiment importance values $if_{s,i,j}$, and $if_{t,i,j}$, we find the probability that $\lambda_{s,i,j} \geq \lambda_{t,i,j}$, $Pr(\lambda_{s,i,j} \geq \lambda_{t,i,j})$. The conditional probability can be written in the Bayesian format as

$$\beta_{s,t,i,j} = Pr(\lambda_{s,i,j} \geq \lambda_{t,i,j} | IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j}) = \frac{Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j} | \lambda_{s,i,j} \geq \lambda_{t,i,j}) Pr(\lambda_{s,i,j} \geq \lambda_{t,i,j})}{Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j})}. \quad (9)$$

Since there is no prior information about the distributions of $\lambda_{s,i,j}$ and $\lambda_{t,i,j}$, we assume that $Pr(\lambda_{s,i,j} \geq \lambda_{t,i,j}) = \frac{1}{2}$. Furthermore,

$$Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = v_{t,i,j} | \lambda_{s,i,j} \geq \lambda_{t,i,j}) = \int_{x=0}^{W} \int_{y=x}^{W} Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j} | \lambda_{s,i,j} = y, \lambda_{t,i,j} = x) \frac{dy}{W} \frac{dx}{W} = \frac{1}{W^2} \int_{x=0}^{W} \int_{y=x}^{W} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y - if_{s,i,j})^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x - if_{t,i,j})^2}{2\sigma^2}} dy\ dx. \quad (10)$$

Similarly, one can find

$$\begin{aligned} Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j}) &= Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j} | \lambda_{s,i,j} \geq \lambda_{t,i,j}) Pr(\lambda_{s,i,j} \geq \lambda_{t,i,j}) \\ &+ Pr(IF_{s,i,j} = if_{s,i,j}, IF_{t,i,j} = if_{t,i,j} | \lambda_{s,i,j} < \lambda_{t,i,j}) Pr(\lambda_{s,i,j} < \lambda_{t,i,j}). \end{aligned}$$

Now, for each parameter, P_t, we define the global importance factor, if_t,

$$if_t = \sum_{i=1}^{N_{em}} \sum_{j=1}^{N_{om}} \sum_{s=1, s \neq t}^{N_p} \beta_{s,t,i,j}. \quad (11)$$

Parameters with a larger if_t have a higher probability of being important compared to the other parameters. We sort the parameters based on their corresponding if_t values.

VI. EVALUATION RESULTS

We have applied the developed method to real distance measurement data for location estimation problem. Parameters that were described in Section II are ranked using our methodology. We illustrate how the various ranking lists differ. Then, we combine the rankings to obtain a global ranking.

The distance measurements data from the CENS lab [18] is used to evaluate the effect of each parameter. This database is based on the real distance measurements for SH4 nodes [13]. 91 nodes are located in fixed locations. Distance measurement is done multiple times and in different days. The measurements are based on the time of flight (ToF) [19] of the signals. In this method, the time of flight of an acoustic signal is used to determine the distance between two nodes. It was previously shown that the noise in the measurements is strongly non-static [20]. Therefore, parametric methods based on optimizing the results according to a fixed noise distribution do not yield good location estimations.

We have used Integer Linear Programming (ILP) to sample the database for drawing instances with specific properties. In each experiment, the PB-design matrix implies a specific value for each parameter. Extreme and normal values for parameters are shown in Table I. The values are determined based on the real measurements' error. In all experiments, ϵ_0^2 is equal to $20(m^2)$.

The following abbreviations are used in this section.

Parameter	DOF			NLM			SMAS			DNGLM		
	L_1	L_2	ML	L_1	L_2	ML	L_1	L_2	ML	L_1	L_2	ML
N_S	4	4	2	6	5	6	2	2	2	3	2	1
B_S	2	1	1	4	2	3	4	9	4	1	4	3
$\overline{\epsilon_S^2}$	1	2	3	2	3	4	3	3	3	5	9	6
$\text{MAX}_{\epsilon_m^2}$	6	8	7	9	10	10	6	4	5	7	7	8
$PER_{\epsilon_0^2}$	7	9	10	10	8	5	7	8	7	9	10	9
$\overline{D_S}$	3	3	4	1	1	1	1	1	1	2	1	2
MINL_S	8	6	8	7	6	9	8	5	10	4	3	10
MAXL_S	10	10	9	5	9	8	10	10	8	10	5	4
$\overline{l_S}$	9	5	5	8	7	7	9	7	9	6	6	5
MIND_S	5	7	6	3	4	2	5	6	6	8	8	7

TABLE II
IMPORTANCE OF DIFFERENT PARAMETERS FOR DIFFERENT OBJECTIVE FUNCTIONS AND METRICS.

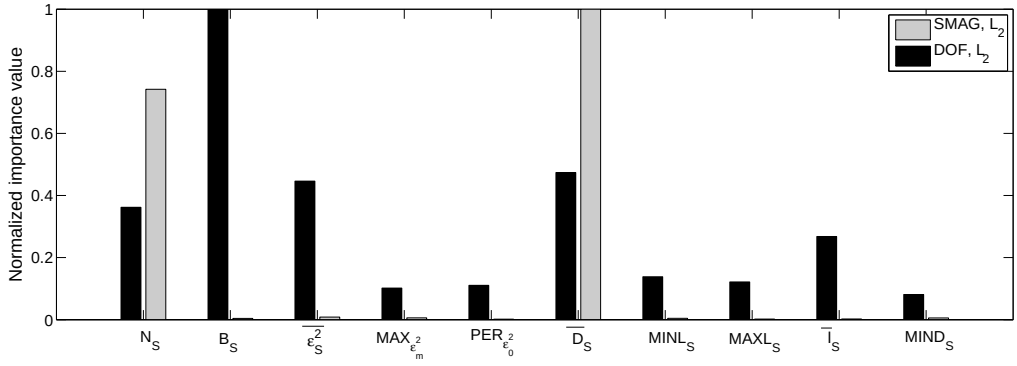

Fig. 2. Importance of different parameters for different objective functions and metrics.

- ML : Maximum Likelihood
- DOF : Drifting of the Objective Function (M_1)
- NLM : Nearest Local Minimum (M_2)
- SMAS : Slope Measurement Around the Solution (M_3)
- DNGLM: Depth of Non-Global Local Minimum (M_4)

Table II shows the result of PB-based evaluations. Each parameter is ranked based on the specific error metric and the specific OF metric. It can be seen that a specific parameter has different rankings under various error metrics and OF metrics. For example, the total number of nodes, N_S, is ranked 1, 2, 3, 4, 5, and 6 in different cases. Thus, a specific parameter does not have the same importance under various metrics. It can be seen that the number of nodes, N_S, and the number of beacons, B_S, are the two most important parameters in most evaluations; $PER_{\epsilon_0^2}$ and MAXL_S have overall low rankings.

The comparative ranks of parameter pairs tend to vary as well. Figure 2 shows the normalized importance factor (IF) for two cases: DOF and SMAS with L_2 error metric. For DOF, the number of beacons B_S is strongly more important than the mean squared error $\overline{\epsilon_S^2}$. The mean degree of nodes, $\overline{D_S}$, is weakly more important than the mean squared error $\overline{\epsilon_S^2}$. The same behavior can be seen in SMAS. From our visual inspections, the number of nodes N_S and the mean degree of nodes $\overline{D_S}$ are the most important while others almost have the same importance factor (IF). The ranks of the mean squared error $\overline{\epsilon_S^2}$ and maximum edge length MAXL_S are 3 and 10 respectively. However, their importance factors are very close.

The discrepancy in the rank and comparative ranks confirm our postulation that averaging the parameter ranks is not the best way for combining them. Thus, we use the combining method that was introduced in Section V. The probability comparisons for the values in Figure 2 are shown in Tables III and IV. The tables compare the importance of parameters. For example, for the DOF-L_2, Figure 2 states that B_S is strongly more important than $PER_{\epsilon_0^2}$. Table III shows that the probability that the mean of B_S is larger than the mean of $PER_{\epsilon_0^2}$ is 0.984. Similarly, $\text{MAX}_{\epsilon_m^2}$ and $PER_{\epsilon_0^2}$ have approximately the same importance. The probability that the mean of $\text{MAX}_{\epsilon_m^2}$ is larger than the mean of $PER_{\epsilon_0^2}$ is 0.49. This probability value is close to 0.5, meaning that there is not enough information to compare the values.

Table IV compares the importance factors of SMAS for the L_2 error metric. Table IV confirms the result. The rows corresponding to N_S, and $\overline{D_S}$ have values close to 1 confirming the high importance of the two parameters. When comparing other parameters, the probability that one parameter is greater than the other is about 0.5. It confirms our previous postulation that simple rankings are not sufficient for concluding the global parameter ordering and the importance factors are significant as well.

Parameter	N_S	B_S	$\overline{\epsilon_S^2}$	$\text{MAX}_{\epsilon_m^2}$	$PER_{\epsilon_0^2}$	$\overline{D_S}$	MINL_S	MAXL_S	$\overline{l_S}$	MIND_S
N_S	0	0.071	0.417	0.725	0.716	0.403	0.708	0.691	0.598	0.748
B_S	0.929	0	0.899	0.993	0.984	0.884	0.977	0.981	0.939	0.984
$\overline{\epsilon_S^2}$	0.583	0.101	0	0.787	0.786	0.476	0.756	0.754	0.660	0.798
$\text{MAX}_{\epsilon_m^2}$	0.275	0.007	0.213	0	0.490	0.193	0.464	0.477	0.354	0.515
$PER_{\epsilon_0^2}$	0.284	0.016	0.214	0.510	0	0.202	0.469	0.499	0.357	0.528
$\overline{D_S}$	0.597	0.116	0.524	0.807	0.798	0	0.785	0.795	0.678	0.821
MINL_S	0.292	0.023	0.244	0.536	0.531	0.215	0	0.519	0.392	0.545
MAXL_S	0.309	0.019	0.246	0.523	0.501	0.205	0.481	0	0.371	0.537
$\overline{l_S}$	0.402	0.061	0.340	0.646	0.643	0.322	0.608	0.629	0	0.671
MIND_S	0.252	0.016	0.202	0.485	0.472	0.179	0.455	0.463	0.329	0

TABLE III
DRIFTING OF OBJECTIVE FUNCTION AND L_2 METRIC: $Pr(\lambda_{i,j,s} \geq \lambda_{i,j,t} | V_{i,j,s} = v_{i,j,s}, V_{i,j,t} = v_{i,j,t})$ WHERE THE FIRST COLUMN IS P_s AND THE FIRST ROW IS P_t.

Parameter	N_S	B_S	$\overline{\epsilon_S^2}$	$\text{MAX}_{\epsilon_m^2}$	$PER_{\epsilon_0^2}$	$\overline{D_S}$	MINL_S	MAXL_S	$\overline{l_S}$	MIND_S
N_S	0	0.947	0.947	0.936	0.944	0.285	0.939	0.931	0.937	0.958
B_S	0.053	0	0.493	0.506	0.504	0.017	0.501	0.496	0.500	0.504
$\overline{\epsilon_S^2}$	0.053	0.507	0	0.509	0.505	0.018	0.504	0.516	0.507	0.511
$\text{MAX}_{\epsilon_m^2}$	0.064	0.494	0.491	0	0.505	0.017	0.504	0.506	0.499	0.499
$PER_{\epsilon_0^2}$	0.056	0.496	0.495	0.495	0	0.017	0.492	0.496	0.502	0.500
$\overline{D_S}$	0.715	0.983	0.982	0.983	0.983	0	0.975	0.984	0.974	0.980
MINL_S	0.061	0.499	0.496	0.496	0.508	0.025	0	0.506	0.501	0.488
MAXL_S	0.069	0.504	0.484	0.494	0.504	0.016	0.494	0	0.502	0.494
$\overline{l_S}$	0.063	0.500	0.493	0.501	0.498	0.026	0.499	0.498	0	0.505
MIND_S	0.042	0.496	0.489	0.501	0.500	0.020	0.512	0.506	0.495	0

TABLE IV
SMAS AND L_2 METRIC: $Pr(\lambda_{i,j,s} \geq \lambda_{i,j,t} | V_{i,j,s} = v_{i,j,s}, V_{i,j,t} = v_{i,j,t})$ WHERE THE FIRST COLUMN IS P_s AND THE FIRST ROW IS P_t.

The global ranking based on the introduced combining method in Section V is shown in Table V. The table indicates that the mean degree of nodes $\overline{D_S}$ is the most important parameter. This result is consistent with Table II where the mean degree of nodes $\overline{D_S}$ is the most important parameter in the seven scenarios.

The global ranking results could be used to improve the goodness of location estimations in ad-hoc networks. To deploy a network or on an already deployed network, one could exploit the results by considering the analyzed effect of each parameter on the estimated location's accuracy. Based on the constraints of the problem, the best parameters for improving the estimated locations could be determined. For example, when there are limitations for the mean degree of the graph, one can increase the number of nodes in the network to increase the accuracy of the estimated location. Note that, changing one parameter typically only improves the accuracy up to a certain point; further changing the parameter would not yield an improvement in the estimation accuracy.

VII. CONCLUSION

We introduce a systematic methodology for determining the challenge of modeling a pertinent data set. The complex modeling problem is studied as an instance of a nonlinear optimization problem that consists of an objective function (OF) and a set of constraints. The data set is the optimization input and the estimated model is the output. We characterize the input by a set of its characteristic parameters. We define four new metrics that can be used to evaluate the goodness of an input for being optimized by a specific OF. The introduced metrics are: (1) drifting of the OF, (2) distance to the nearest local minimum, (3) the slope of the OF around the solution, and (4) the depth of the non-global local minimums. We employ Plackett and Burmann simulation methodology to systematically evaluate the impact of various input parameters under each metric. Finally, we present a method for combining the effect of parameters under different metrics to determine the global impact of each parameter. We utilize the new methodology for estimating the locations of the nodes in an ad-hoc network where the distance measurement data is available. Three common forms of error metrics are considered: L_1, L_2 and L_∞. Our evaluations show that the mean degree on the nodes and the number of nodes in the network are the two most important parameters for estimating the locations.

VIII. ACKNOWLEDGEMENT

This work is supported by the National Science Foundation (NSF) CAREER Award under grant number 0644289.

Parameter	N_S	B_S	$\overline{\epsilon_S^2}$	$\text{MAX}_{\epsilon_m^2}$	$PER_{\epsilon_0^2}$	$\overline{D_S}$	MINL_S	MAXL_S	$\overline{l_S}$	MIND_S
Rank	2	3	4	8	10	1	6	9	7	5

TABLE V
GLOBAL RANKS.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.
[2] "http://cricket.csail.mit.edu, seen in sep. 2007."
[3] "http://lsda.jsc.nasa.gov/scripts/datasets/dataset_detail_result.cfm?dataset_catalog=jmapap003_245, seen in sep. 2007."
[4] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. England: Wiley, April 2004.
[5] C. Bullard and A. Sebald, "Monte carlo sensitivity analysis of input-output models," *The Review of Economics and Statistics*, vol. 70, no. 4, pp. 708–712, 1988.
[6] E. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*. Springer-Verlag, 2006.
[7] T. Eren, D. Goldenberg, W. Whiteley, Y. R. Yang, A. S. Morse, B. Anderson, and P. Belhumeur, "Rigidity, computation, and randomization in network localization," *Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, pp. 2673– 2684, 2004.
[8] D. Goldenberg, A. Krishnamurthy, W. Maness, Y. Yang, A. Young, A. Morse, and A. Savvides, "Network localization in partially localizable networks," in *Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, 2006, pp. 313–326.
[9] N. Patwari, J. Ash, S. Kyperountas, A. H. III, R. Moses, and N. Correal, "Locating the nodes: cooperative localization in wireless sensor networks," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 54–69, 2005.
[10] P. Biswas and Y. Ye, "Semidefinite programming for ad hoc wireless sensor network localization," in *Information Processing in Sensor Networks (IPSN)*, 2004, pp. 2673– 2684.
[11] N. Priyantha, A. Chakraborty, and H. Balakrishnan, "The cricket location-support system," in *ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM)*, 2000, pp. 32 – 43.
[12] "http://www-osl.cs.uiuc.edu/research?action=topic&topic=sensor+networks, seen in sep. 2007."
[13] J. Feng, L. Girod, and M. Potkonjak, "Location discovery using data-driven statistical error modeling," in *Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, 2006, pp. 1–14.
[14] J. Saxe, "Embeddability of weighted graphs in k-space is strongly np-hard," in *Allerton Conf. in Communications, Control, and Computing*, 1979, pp. 480–489.
[15] R. L. Plackett and J. P. Burman, "The design of optimum multifactorial experiments," *Biometrika*, vol. 33, no. 4, pp. 305–325, 1946.
[16] S. Kirkpatrick, C. G. Jr., and M. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.
[17] D. Montgomery, *Design and Analysis of Experiments*, fifth edition ed. Wiley, 2001.
[18] "http://research.cens.ucla.edu/."
[19] S. Lanzisera, D. Lin, and K. Pister, "Rf time of flight ranging for wireless sensor network localization," in *Workshop on Intelligent Solutions in Embedded Systems (WISES)*, 2006, pp. 1–12.
[20] J. Feng, L. Girod, and M. Potkonjak, "Consistency-based on-line localization in sensor networks," in *International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2006, pp. 529–545.