Advanced Computing in Industrial Mathematics

12th Annual Meeting of the Bulgarian Section of SIAM December 20–22, 2017, Sofia, Bulgaria Revised Selected Papers
The series “Studies in Computational Intelligence” (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/7092
Advanced Computing in Industrial Mathematics

12th Annual Meeting of the Bulgarian Section of SIAM December 20–22, 2017, Sofia, Bulgaria Revised Selected Papers
Preface

The 12th Annual Meeting of the Bulgarian Section of the Society for Industrial and Applied Mathematics (BGSIAM) was held in Sofia, December 20–22, 2017. The Section was formed in 2007 with the purpose to promote and support the application of mathematics to science, engineering, and technology in Bulgaria.

The goals of BGSIAM follow and creatively develop the general goals of SIAM:

- To advance the application of mathematics and computational science to engineering, industry, science, and society;
- To promote research that will lead to effective new mathematical and computational methods and techniques for science, engineering, industry, and society;
- To provide media for the exchange of information and ideas among mathematicians, engineers, and scientists.

During the BGSIAM’17 conference, a wide range of problems concerning recent achievements in the field of industrial and applied mathematics were presented and discussed. The meeting provided a forum for exchange of ideas between scientists, who develop and study mathematical methods and algorithms, and researchers, who apply them for solving real-life problems.

Among the topics of interest are high-performance computing, numerical methods and algorithms, analysis of partial differential equations and their applications, mathematical biology, control and uncertain systems, stochastic models, molecular dynamics, neural networks, genetic algorithms, metaheuristics for optimization problems, generalized nets, and Big Data.

The invited speakers were:

- Krassimir Atanassov (Bulgarian Academy of Sciences), Generalized Nets—Theory and Applications
- Peter Minev (University of Alberta, Canada), High-order Artificial Compressibility for the Navier–Stokes Equations
- Maya Neytcheva (Uppsala University, Sweden), Enhanced degree of parallelism when solving optimal control problems constrained by evolution equations
Zahari Zlatev (Aarhus University, Denmark), Application of repeated Richardson Extrapolation in the treatment of some chemical modules of environmental pollution models.

We would like to thank all the referees for the constructive remarks and criticism, which furthered considerable improvements of the quality of the papers in this book.

Sofia, Bulgaria

Krassimir Georgiev
Michail Todorov
Ivan Georgiev
Method for Indoor Localization of Mobile Devices Based on AoA and Kalman Filtering ....................................... 1
A. Alexandrov and V. Monov

Cross-Validated Sequentially Constructed Multiple Regression .... 13
Slav Angelov and Eugenia Stoimenova

How to Assess Multi-population Genetic Algorithms Performance Using Intuitionistic Fuzzy Logic ........................................ 23
Maria Angelova and Tania Pencheva

Perturbation Analysis of a Nonlinear Matrix Equation Arising in Tree-Like Stochastic Processes .............................. 37
Vera Angelova

On Two-Way Generalized Nets ............................................ 51
Krassimir Atanassov

Variational Methods for Stable Time Discretization of First-Order Differential Equations ....................................... 63
Simon Becher, Gunar Matthies and Dennis Wenzel

Numerical Solutions of Ordinary Fractional Differential Equations with Singularities .............................................. 77
Yuri Dimitrov, Ivan Dimov and Venelin Todorov

Testing Performance and Scalability of the Pure MPI Model Versus Hybrid MPI-2/OpenMP Model on the Heterogeneous Supercomputer Avitohol .............................................. 93
Vyara Koleva-Efremova

Analyses and Boolean Operation of 2D Polygons .................. 107
Georgi Evtimov and Stefka Fidanova
Ant Colony Optimization Algorithm for Workforce Planning: Influence of the Algorithm Parameters ........................................................................................................ 119
Stefka Fidanova, Olympia Roeva and Gabriel Luque

Convergence of Homotopy Perturbation Method for Solving of Two-Dimensional Fuzzy Volterra Functional Integral Equations ........................................................................................................ 129
Atanaska Georgieva and Artan Alidema

Iterative Method for Numerical Solution of Two-Dimensional Nonlinear Urysohn Fuzzy Integral Equations ........................................................................................................ 147
Atanaska Georgieva, Albena Pavlova and Svetoslav Enkov

Comparison Analysis on Two Numerical Solvers for Fractional Laplace Problems ........................................................................................................ 163
Stanislav Harizanov and Svetozar Margenov

Initial Calibration of MEMS Accelerometers, Used for Measuring Inclination and Toolface ........................................................................................................ 177
Tihomir B. Ivanov and Galina S. Lyutskanova-Zhekova

Design and Implementation of Moving Average Calculations with Hardware FPGA Device ........................................................................................................ 189
Vladimir Ivanov and Todor Stoilov

On the Exact Traveling Wave Solutions of a Hyperbolic Reaction-Diffusion Equation ........................................................................................................ 199
Ivan P. Jordanov and Nikolay K. Vitanov

Generalized Nets: A New Approach to Model a Hashtag Linguistic Network on Twitter ........................................................................................................ 211
Kristina G. Kapanova and Stefka Fidanova

On the Group Analysis of Differential Equations on the Group SL(2, R) ........................................................................................................ 223
Georgi Kostadinov and Hristo Melemov

On Digital Watermarking for Audio Signals ........................................................................................................ 231
Hristo Kostadinov and Nikolai L. Manev

Numerical Evaluation of Fokas’ Transform Solution of the Heat Equation on the Half-Line ........................................................................................................ 245
Mohamed Lachaab, Peter R. Turner and Athanassios S. Fokas

Molecular Dynamics Study of the Solution Behaviour of Antimicrobial Peptide Indolicidin ........................................................................................................ 257
Rositsa Marinova, Peicho Petkov, Nevena Ilieva, Elena Lilkova and Leandar Litov
A Functional Expansion and a New Set of Rapidly Convergent Series Involving Zeta Values ................................................. 267
Lubomir Markov

Simulated Annealing Method for Metal Nanoparticle Structures Optimization .......................................................... 277
Vladimir Myasnichenko, Leoned Kirilov, Rossen Mikhov, Stefka Fidanova and Nikolay Sdobnyakov

Orientation Selectivity Tuning of a Spike Timing Neural Network Model of the First Layer of the Human Visual Cortex ............. 291
Simona Nedelcheva and Petia Koprinkova-Hristova

Gaussian Model Deformation of an Abdominal Aortic Aneurysm Caused by the Interaction between the Wall Elasticity and the Average Blood Pressure .............................................. 305
N. Nikolov, S. Tabakova and St. Radev

EM Estimation of the Parameters in Latent Mallows’ Models .......... 317
Nikolay I. Nikolov and Eugenia Stoimenova

Evolution Equation for Propagation of Blood Pressure Waves in an Artery with an Aneurysm .............................................. 327
Elena V. Nikolova

The Derivatives of Polynomials Represented as a Linear Combination of Finite Differences with Step One ............................. 341
Inna Nikolova

Investigation of As-cast Light Alloys by Selected Homogenization Techniques with Microstructure Effects ............................. 355
Ludmila Parashkevova and Pedro Egizabal

Performance Analysis of Intel Xeon Phi MICs and Intel Xeon CPUs for Solving Dense Systems of Linear Algebraic Equations: Case Study of Boundary Element Method for Flow Around Airfoils ......................................................... 369
D. Slavchev and S. Margenov

Memristor CNNs with Hysteresis ........................................... 383
Angela Slavova and Ronald Tetzlaff

Change Point Analysis as a Tool to Detect Abrupt Cosmic Ray Muons Variations ......................................................... 395
Assen Tchorbadjieff and Ivo Angelov
Performance Analysis of Effective Methods for Solving Band Matrix SLAEs After Parabolic Nonlinear PDEs .................. 407
Milena Veneva and Alexander Ayriyan

Statistical Characteristics of a Flow of Substance in a Channel of Network that Contains Three Arms ......................... 421
Nikolay K. Vitanov and Roumen Borisov

A. B. Yanovski and T. I. Valchev
Method for Indoor Localization of Mobile Devices Based on AoA and Kalman Filtering

A. Alexandrov and V. Monov

Abstract The mobile devices are a widely used tools for indoor localization. By different reasons the localization of mobile devices in closed areas has not yet fully developed because of missing of a reliable positioning technology. The paper presents a hybrid method for improving the accuracy of indoor positioning approach for Bluetooth Low Energy (BLE) mobile devices based on optimized combination of Angle of Arrival (AoA) and Receive Signal Strength (RSS) technologies. We propose a hybrid optimization method for indoor positioning, realized by two stage data fusion process using Extended Kalman filtering approach and Fraser-Potter equation. The test results show that the proposed method can achieve sensitively better accuracy in a real environment compared to existing indoor localization methods and techniques.

1 Introduction

Indoor Positioning is a challenge topic in public areas, which are used by large number of people. The problem of the mobile devices localization in closed areas and buildings has become more difficult because of the big complexity and scale of the public space. Nowadays there is a lot of researches in the area of implementation of evolutionary algorithms and AI neural networks in the localization process [1, 2]. The positioning and the localization of assets in indoor spaces is useful for several reasons. Loss and theft of equipment take a large expense of the budget. When it is possible to have the position of a device in real time, a system could be developed that locates the assets through the public area. The main goal of this paper is to develop a method and algorithm for an optimized indoor positioning system for localization of mobile devices (MD) and assets in indoor public areas.

A. Alexandrov (✉)
IICT-BAS, Sofia, Bulgaria
e-mail: akalexandrov@iit.bas.bg

V. Monov
e-mail: vmonov@iit.bas.bg

© Springer Nature Switzerland AG 2019
K. Georgiev et al. (eds.), Advanced Computing in Industrial Mathematics,
Studies in Computational Intelligence 793,
https://doi.org/10.1007/978-3-319-97277-0_1
2 Related Work

A large number of solutions, such as angle of arrival (AOA) [3, 4], received signal strength (RSS) [5–7], time of arrival (ToA) [8], time difference of arrival (TDOA) [9], and etc., have been proposed to attain mobile devices localization by measuring the received radio signal between the mobile device and the localization sensor node.

2.1 Typical AoA Based Method

In the AoA based techniques [3, 4, 8], the localization sensors using directional antennas have the capability of localization the Radio Frequency (RF) signal angle of arrival. For this purpose, some techniques like RF signal angle diversity are used in order to determine the directionality of the receivers antenna system. AoA method can fix the 2-D coordinate of mobile device (MD) with two angle values measured from two localization sensors (LS) to the MD. As it is shown in Fig. 1, LS1 and LS2 represent two LSs which coordinates are already known and MD is supposed to be the mobile device. \( \theta_1 \) is the measured angle between LS1 and MD and \( \theta_2 \) is the angle between LS2 and MD.

In this case we have

\[
\begin{bmatrix}
\tan \theta_1 \\
\tan \theta_2
\end{bmatrix} = \begin{bmatrix}
(y - y_1)/(x - x_1) \\
(y - y_2)/(x - x_2)
\end{bmatrix}
\]

(1)

By solving (1), we can obtain the coordinate of MD as follows,

\[
\begin{bmatrix}
x \\
y
\end{bmatrix} = \begin{bmatrix}
y_2 - y_1 + x_1 \tan \theta_1 - x_2 \tan \theta_2 \\
y_2/\tan \theta_2 - y_1 \tan \theta_1 - x_2\tan \theta_1 - x_2
\end{bmatrix}
\]

(2)

The measurement angle error has some deviations which is caused by the measuring equipment and environmental noise. Therefore, more than one possible coordinates of MD should be calculated by using any two LS, and these calculated position points will vary around the real coordinate of MD. Varieties of methods have been
proposed to restrain the influence caused by environment noise. According to the distribution character of these varying points, the simplest method is taking the mean of these point coordinates.

2.2 RSS Method

The RSS based techniques, use a distance measurement based on the attenuation introduced by the propagation of the signal from the MD to the LS. In [5, 7, 10] is proposed an empirical mathematical model to calculate the distance according to signal propagation:

\[ p(R) = p(R_0) - 10n \log \frac{R}{R_0} - \begin{cases} nW \times WAF & (nW < C) \\ C \times WAF & (nW \geq C) \end{cases} \]

(3)

In the presented above formula (3), \( R \) denotes the distance between the MD and the LS, \( R_0 \) is an already known distance acting as reference, \( p(R) \) and \( p(R_0) \) represent the signal strength received at \( R \) and \( R_0 \) respectively, \( nW \) is the number of obstructions between the MD and the LS, \( WAF \) is the average attenuation coefficient of the wall, \( C \) is the maximum number of attenuation barriers between the MD and the LS, and \( n \) is the routing attenuation coefficient which could be empirically determined. Based on the represented RSS technology, a few methods have been proposed to estimate the position of the MD. For example, the fingerprint based solution [7] for target positioning is a typical application of RSS technology. In general, we can divide the fingerprint methodology into two phases: offline detection and sampling and online matching. For the sampling phase realization, a database is created offline to store the RF signal parameters including the geographical positions and the corresponding signal levels. In the online matching phase, the corresponding RF signals collected for the MD are compared against the already stored in the database records. By this way, it will be able to calculate the MD coordinates, as long as any RF signal record in the database is matched.

2.3 ToA Method

ToA localization algorithm as shown in Fig. 2 is based on ToA circumference equation, through the different combination of intersecting lines between corresponding circles, and generates different positioning equations [1].

The geometry model represents the signal transmit from MD to LSs. The measured distance is \( d_i \):

\[ d_i = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad i = 1, 2, 3 \]

(4)
Fig. 2 ToA localization principle

\((x_1, y_1)\) is the LSs coordinates, \((x_t, y_t)\) is coordinate of the MD. Then the MD must be located on the circumference of a circle with radius \(d_i\) centered on the LS. When there are three LS coordinates, ToA measurement equation can be expressed as follow:

\[
(x_2 - x_1)x_t + (y_2 - y_1)y_t = \frac{1}{2}[(x_2^2 + y_2^2) - (x_1^2 + y_1^2) + (d_1^2 - d_2^2)] \tag{5}
\]

\[
(x_3 - x_2)x_t + (y_3 - y_2)y_t = \frac{1}{2}[(x_3^2 + y_3^2) - (x_2^2 + y_2^2) + (d_2^2 - d_3^2)] \tag{6}
\]

By these equations, we can calculate the coordinates of the MD:

\[
x_t = \frac{(y_2 - y_1)D_3 - (y_3 - y_2)D_1}{[(x_3 - x_2)(y_2 - y_1) - (x_2 - x_1)(y_3 - y_2)]} \tag{7}
\]

\[
y_t = \frac{(x_2 - x_1)D_3 - (x_3 - x_2)D_1}{[(y_3 - y_2)(x_2 - x_1) - (y_2 - y_1)(x_3 - x_2)]} \tag{8}
\]

where:

\[
D_1 = \frac{1}{2}[(x_2^2 + y_2^2) - (x_1^2 + y_1^2) + (d_1^2 - d_2^2)] \tag{9}
\]

\[
D_3 = \frac{1}{2}[(x_3^2 + y_3^2) - (x_2^2 + y_2^2) + (d_2^2 - d_3^2)] \tag{10}
\]
2.3.1 Time Difference-of-Arrival (TDoA)

This technology [9] uses two different kind RF signals - one transmitted from LS and another from referent base station. It is common practice to use equal frequency RF signals with different polarization. The time offset between the received two radio signals is used to calculate the MD’s position. The calculation is based on the following equation:

\[ \frac{R_1}{c_1} - \frac{R_2}{c_2} = t_1 - t_2 \]  \hspace{1cm} (11)

where in vacuum \( c_1 = c_2 = c \). The velocity of the RF signals can vary depend of the medium.

In (11), \( c_1 \) represents the velocity of the first RF signal, \( c_2 \) is the velocity of second RF signal, \( t_1 \) and \( t_2 \) are the times for these two signals traveling from MD to LS and from the MD to the referent station, and \( R_1 \) is the distance between the MD and the LS and \( R_2 \) is the distance between the MD and the referent station. A large number of works have explored TDoA-based methods. For instance [9, 11–13].

2.4 Kalman Filter Observation and Transition Models

The equations of the Kalman filter approach [14, 15] can be specified in two main categories: time update equations and measurement update equations. Generally, the target of the time update equations is to predict the current state and error estimations and at the same time to obtain a priori estimates for the next time period step. The target of the measurement update equations is the feedback, i.e. the realization of a new measurement based on a priori estimate and to obtain an improved a posteriori estimate. We can describe the time update equations as predictor equations, at the same time the measurement update equations can be specified as corrector equations (see Fig. 3).

![Fig. 3 Kalman filtering approach](image)
The prediction phase starts with initial estimation of $\hat{x}_{k-1}$ and covariance vector $P_{k-1}$ and proceed with

$$\hat{x}_k = A\hat{x}_{k-1} + Bu_k + Cw_k$$

(12)

$$z_k = H_k x_k + D_k \nu_k$$

(13)

where $\hat{x}_k$ is the estimated value, $A$ is the transition state matrix of the process, $B$ is the input matrix, $C$ is the noise state transition matrix, $u_k$ is the known input value, $w_k$ is the noise, $z_k$ is the observation vector, $\nu_k$ is a variable describing observation noise, and $H_k$ is the matrix of the observed value $z_k$ and $D_k$ is a matrix describing the contribution of noise to the observation. The measurement correction adjusts the projected estimate by an actual measurement at that time. In our case, we will be focused mainly on the measurement update algorithm. The current paper doesn’t focus on details in the mathematical side of the Kalman filter measured updated equations. Details can be found in [13, 14, 16–18]. Based on the above cited papers we accept that the final extended Kalman filter measurement update equations are formulated as follows:

$$G_k = \frac{P_k H_k}{H_k P_k H_k^T + R_k}$$

(14)

$$\hat{x}_k = \hat{x}_k + G_k (z_k - H_k \hat{x}_k)$$

(15)

$$P_k = (1 - G_k H_k) P_k$$

(16)

In formula (14) $G_k$ is the so called Kalman gain, $P_k$ is an error covariance vector, $H_k$ is a matrix of the observed value vector $z_k$ and $R_k$ is the covariance matrix. The initial task during the measurement update is to compute the Kalman gain $G_k$. The next phase is to actually measure the process to receive and then to calculate a posteriori state estimate by adding the correction based on the measurement and estimation as in (15). The last phase of the process is to obtain an a posteriori error covariance estimate via (16). After each measurement update pair, the process is repeated with the previous a posteriori estimates. In the present equation of the Kalman filter, each of the measurement error covariance matrices $R_k$ can be measured before the execution of the Kalman filter data fusion process. In the case of the measurement error covariance $R_k$, especially this makes sense because there is a need to measure the process while operating the filter. We should be able to take some off-line sample measurements to determine the variance of the measurement error. In both cases, we have a good basis for choosing the parameters. Very often superior filter performance can be obtained by modification of the filter parameter $R_k$. The modification is usually performed off-line, frequently with the help of another
Kalman filter. Finally, we can note that in cases where $R_k$ is constant, both the estimation error covariance $P_k$ and the Kalman gain $G_k$ will start to stabilize very fast and then remain constant. If this is the case, the $R_k$ parameter can be precomputed by running the filter off-line.

Extended Kalman localization of the flow diagram as shown in Fig. 4, mainly includes the state variables and the error variance matrix initialization. The state equation and measurement equation are linearized, computing the Kalman gain, update the status variables, predictions and the rest four steps.

3 Hybrid Indoor Localization Method Based on RSS and AoA

In the proposed localization method, we will be focused mainly on Angle of Arrival and RSS techniques as one of the reliable and low cost methods for indoor target localization. The proposed technique is based on ranging, whereby angle approximations are obtained [8] including some syntheses of interaction information [19, 20]. In this context, geometric approaches will be used to calculate the position of the mobile device of the target as an intersection of position lines obtained from the position-related parameters. Since the RF signal measurements in real systems have a deviation, especially in indoor areas, some optimization and error reducing techniques are used to improve the accuracy of the indoor positioning. The main factors who reduce the mobile device positioning accuracy are:

- the distance between the localization sensor and the mobile device of the target.
- the level of the RF noise and the signals reflected from indoor walls and metal equipment surfaces.

The second influencing factor can be sensitively reduced by using Bayesian filtering like Kalman filter and related statistical methods as Fraser-Potter smoothing. For the AoA implementation we use custom designed grid of directed antennas oriented at different azimuth and evaluation positions. Instead of the use of two different
localization sensors situated on different places to measure the angle of arrival as described in Sect. 2.1, we use only one localization sensor. The received signal angle information can be calculated because the antenna array geometry is known. The proposed AoA principle doesn’t need of time synchronization between localization sensors. If there is a need for 2D localization only a single localization sensor is enough to find the target coordinates. The case of 3D localization needs a minimum of two sensor localization nodes to calculate the mobile device x, y and z coordinates.

In the current paper, we are focused only on the 2D indoor localization problem as shown on Fig. 5.

Implementation of 2D position localization using RSS and AoA

In Fig. 6, the 2D tracking plane is shown. When a localization sensor detects a mobile device, the system can return the \( x, y \) coordinates of the target. Usually, the height \( (h = 1.2 \text{ m}) \) component is fixed for all the measurements.

The presented approach relies on precise range determination, using combination of directional grid antennas, RSS and Angle of Arrival techniques for the device position calculation. The measured data pass a process of optimization, based on the extended Kalman filter approach and the Fraser-Potter equation. In this case, each reference sensor node (with known position) sends a ranging request to the BLE mobile devices in the area. Then the mobile device replies the sensor request, which is received by the grid of directional antennas. Based on the antenna directional diagram and an embedded in the receiver RSSI (Received Signal Strength Indication) capability could be calculated the azimuth, the evaluation and the strength of the received RF signal. Then the standard deviation of the measured data is calculated and it starts the Kalman filter based local optimization process for the last 6 measurements.

In the current case we used an experimental test system. The localization sensor node with integrated data fusion algorithm realizes a number (in our case 6 which depends
from the processor memory capacity) consistent measurements with 100 ms period. The implemented algorithm of the node fuses the results using Kalman filtering method. Based on a decision criterion (Fig. 7), the node sends the fused data of the localization measurement to the control center.
As shown on Fig. 7 the proposed indoor localization process can be divided into two main phases. In Phase 1 the RF signal measurement and target position calculation are executed. The measuring process starts simultaneously through two independent channels - AoA and RSS measurement channels. During the first phase, some RF signal requests are sent to the mobile device and the RF response is received by one or more localization sensors (LS) with fixed coordinates. In Phase-2 the process of measured data optimization starts by two level Kalman filter optimization. At the first level, a KF based optimization process is executed for each measurement channel AoA and RS, respectively. The local optimization process is combined with smoothing procedure based on Fraser-Potter equations

\[ x_n = \sigma_n^2 (\sigma_1^{-2} \hat{x}_1 + \sigma_2^{-2} \hat{x}_2 + \cdots + \sigma_6^{-2} \hat{x}_6) \]  

(17)

where \( \sigma_n^2 = \sigma_1^{-2} + \sigma_2^{-2} + \cdots + \sigma_6^{-2} \) is the variance of the combined estimate and \( x_n \) represents the combined measurement. Finally, the calculated and optimized data from each measurement channel are optimized at the second level by Extended Kalman filter approach.

The proposed two stage data fusion method was tested by experimental localization indoor system using rotating antenna, Qualcomm BLE tags and regular mobile devices (Fig. 8).

The results from the test of experimental BLE based indoor localization (Fig. 9) show that the performance of the overall system is increased sensitively because of the reduced communication traffic to the control center and the reliability of the measured data from the localization sensor node compared to regular Kalman filtering. At the same time the decentralized Kalman filter algorithm effectively reduces and measurement noises.

Fig. 8  Experimental AoA/RSS localization method system
4 Conclusions and Future Work

In this paper, a new hybrid method for improving indoor localization was presented. The method utilizes two independent channel position calculation processes based on RSS and AoA, respectively. Two level optimization is implemented. At the first level Kalman Filter and Fraser-Potter equation statistical approach is realized. The second level employs Extended Kalman filter which works as data fusion device between the channels. It additionally reduces the noise and improves the reliability of the measured data. The performance of the proposed method was tested in real environment by an experimental indoor localization system. The presented method allows the information from a consistent number of measurements to be combined and integrated in real time. Our approach gives the flexibility and reliability of the process of the mobile devices indoor localization. The experimental results show that the implementation of the proposed hybrid method improves substantially the accuracy in typical office building applications. The future research directions include the noise characteristics analysis of the received signal strength and choosing the most appropriate filtering initial parameters and correction coefficients.

References


Abstract  In this paper will be briefly presented a technique that we call Cross-validated sequentially constructed multiple regression. It is applied to a multiple regression model with at least 2 variables. This technique combines sequentially some of the model variables into components using a leave-one-out cross-validation procedure while taking into account the correlations between the model variables. The uncombined model variables along with the obtained components are then used to estimate a regression model. The newly obtained model is with lower multicollinearity, and it tends to give better out-of-sample error while recalculated with additional observations. The proposed method is tested on a real accounting data concerning the Bulgarian gas utilities.

1 Introduction

Let us observe the multiple regression as a tool for prediction. A well-known fact is that under the Gaus-Markov conditions the least squares estimator is the best linear unbiased estimator for multiple regression, see for example [12, p. 41]. Let us assume that the G-M conditions hold, the input data are suitable for linear model, and that the coefficients are estimated by the least squares estimator. Another important thing is that we are assuming that we have independent model variables. However, usually in practice, we have correlations between the variables, and as a result, the model has multicollinearity problem. Multicollinearity causes some of the coefficients in the model to be estimated with high variance [9] or even to be biased because of the presence of suboptimal solutions. This leads to unstable model, and the final result is
poor predictions. Moreover, if there is a presence of multicollinearity, and the model is recalculated by a data set with outliers than the adverse effects will be increased. If we examine the outliers as a separate problem [6], we should say that they can be influential to the model or not. Usually, even if they are not too influential they have a higher impact over the model than the other observations, and their individual or group effect causes the model to be biased which again may lead to poor prediction results for observations out of the learning set. Additionally, poor prediction result may be caused by overfitting the model - there is a small number of observations per model variable, and as a result, the estimated model is misleading [2]. Under our initial assumptions, we can summarize that to obtain a linear model with better prediction performance than a multiple regression we need to take into consideration all of the mentioned factors.

To simplify the understanding of the aim of this paper, we will define one task. Let us assume that over a chosen learning set we have estimated a regression model with the desired model statistics and prediction results checked over some test sets. We want to assure that this model will remain with as nice diagnostics and prediction performance as possible after adding more observations to its estimation. The method proposed in this paper is an instrument for such a handling.

2 Cross-Validated Sequentially Constructed Multiple Regression

We start this section with some notions, then we present the core of the method which incorporates two model variables.

We are observing a multiple regression model:

\[ Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_k X_k + \xi, \]

where \( Y \) is the predicted variable, \( \alpha_i \) is the coefficient in front of predictor \( X_i \) for \( i = 1, \ldots, k \), \( \alpha_0 \) is the model intercept term, and \( \xi \) is the error term.

A standard approach to measure the model’s quality of fit is to split the primary data set into a test set and a learning set. We estimate the model over the learning set and then test it over the test set using a function over the model residuals. Such a function is the Root mean squared error (RMSE), Eq. 2, for more information see [3].

\[ \text{RMSE}(A) = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon^2(i)}{n}}, \]

where \( A \) is the test set, \( \varepsilon(i) \) is the error of the i-th observation, and \( n \) is the number of observations in set \( A \).

When the number of the observations in the primary data set is small it is not very convenient to split the set into two parts, as a result we have to estimate the model’s
quality of fit by the \( \text{RMSE} \) measure over the primary set (the full data set). However, the RMSE estimated over the learning set leads to a misleading result because it is in general biased downwards, see [7, p. 292]. To obtain a more realistic estimate of how good is the model in that case we can use the Root mean squared error after cross-validation (RMSECV) measure, see Eq. 3.

\[
\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}},
\]

where \( n \) is the number of the observations in the primary data set, \( e_i \) is the error for the \( i \)-th observation which is obtained from the regression model estimated over the primary data set without the \( i \)-th observation. RMSECV consists of out of sample errors which makes it more realistic measure for the model fit. RMSECV measure is based on a Leave-one-out cross-validation (LOOCV). This type of cross-validation is the one with the smallest variance of the error compared to the other types of cross-validation, to see more about the topic check [8, 10].

The technique that we are proposing merges some of the model variables into components reducing in that manner the total number of variables. The way we merge two chosen variables into one component is the core of the method. Each component is forged in a procedure that involves two model variables (or other already obtained components). Let us assume for simplicity that we want to combine variables \( X_1 \) and \( X_2 \) into one component \( Z \). Moreover, the estimate \( a_1 \) of the coefficient in front of \( X_1 \) is bigger in absolute value than the estimate \( a_2 \) of the coefficient in front of \( X_2 \). Then \( Z(k) = X_1 + k \star X_2 \), where we are searching for the optimal \( k \in (k_0 - \mu, k_0 + \mu) \), \( k_0 = \frac{a_2}{a_1} \), \( \mu \) is a small number usually less than 1, while minimizing the RMSECV function for the multiple regression with the rest of the model variables (except the mentioned two) and the new component \( Z(k) \), see Eq. 4. For \( k = k_0 \) we will attain the initial regression model (after substituting \( Z(k_0) \) with \( X_1 \) and \( X_2 \), \( k_0 \) is in the interval \([-1,1]\) by the definition of \( Z(k) \), \( \mu \) is usually small because we do not expect estimates much more different than the regression model’s ones.

\[
\min_k \text{RMSECV}(k) = \min_k \sqrt{\frac{\sum_{i=1}^{n} e_i^2(k)}{n}}, \quad k \in (k_0 - \mu, k_0 + \mu), \quad \mu > 0,
\]

where \( e_i(k) = Y(i) - f_i(k) \) is the error from the \( i \)-th observation from the model \( f_i(k) \). \( f_i(k) \) is the multiple regression model with input variables \( Z(k), X_3, X_4, ..., X_n \) estimated with the full data set except the \( i \)-th observation. The RMSECV(k) function is continuous and positive. Moreover, it is bounded from above, the simplest way to see this, without going into details, is to note that RMSECV(k) is constructed from errors derived from regression models, and these errors are expected to be as small as possible. Thus, searching for the global minimum of the RMSECV(k) function is a valid operation.
2.1 The Proposed Method’s Framework for More Than Two Variables

The proposed method has the following algorithm:
We have a multiple regression model with \( n \) variables.

1. We choose the model variable which has the worst estimate based on its absolute t-value. Then we find the most correlated with it model variable, and we combine them into one component;
2. The regression model is estimated with the new component instead of the two chosen variables from Step 1 and all of the rest model variables;
3. Step 1 is repeated with the \( n-1 \) model variables;
4. Step 2 is repeated with the obtained component from Step 3 and so on.

The procedure ends when we have achieved absolute t-values over a chosen threshold for all of the derived model variables or when the model variables are reduced to a predefined percentage from their initial number. We suggest that the procedure should stop when all of the model variables are with absolute t-values over 4.5, or the number of the derived model variables is around 50% of the initial number of variables.

The obtained components are then used instead of the model variables from which they are made from.

2.2 The Idea Behind the Approach

First, let us note that we are merging the variable with the highest estimation error for its coefficient with a variable that is most highly correlated with it. Thus, we are reducing the level of multicollinearity for the model, and as a consequence, we hope to achieve better-estimated coefficients for the model with the new component.

Second, while merging variables into components we reduce the total number of model variables, as a result we have smaller amount of model coefficients to compute while the model is updated with additional observations. Smaller number of variables along with bigger number of observations is a preposition for a better estimated model.

Third, each component is produced while minimizing RMSECV. RMSECV can be regarded as the RMSE over the learning set while performing LOOCV. Its goal is to see how well a regression model is capable of predicting out of sample observations. While minimizing RMSECV(k), we are searching for a component \( Z(k) \) which allows the model that uses it to be as robust to changes in the learning set as possible concerning the out of sample prediction performance. To make this statement more clear we will present one valuable property of the LOOCV errors which shows the connection between them and their related errors from a multiple regression.
Theorem 1  The out-of-sample residuals $\varepsilon_i$, $i = 1, \ldots, n$ used in the RMSECV measure for Model 1 can be computed by the following formula:

$$\varepsilon_i = \frac{\hat{\xi}_i}{1 - h_{i,i}}, \quad i = 1, \ldots, n,$$

(5)

where $\hat{\xi}_i$ is the error of the $i$-th observation from Model 1, $n$ is the number of observations, and $h_{i,i}$ is the $i$-th element from the diagonal of the hat matrix $H = X(X'X)^{-1}X'$, where $X$ is the design matrix of Model 1, $X'$ is its transposed matrix.

Proof  This proof will follow a similar construction as the Theorem which is for the support vector regression case [4]. It is essential to show that for $i \in \{1, \ldots, n\}$:

$$\hat{\alpha} - \hat{\alpha}_i = \frac{\xi(i)}{1 - h_{i,i}}(X'X)^{-1}x_i,'$$

(6)

where $\hat{\alpha}$ is the vector of the estimates of the coefficients of Model 1, i.e., the regression with the full data set; $\hat{\alpha}_i$ is the vector of the estimates of the coefficients of the regression model without the $i$-th observation; $X_i$ is the design matrix of the model without the $i$-th observation, i.e., it is $X$ without the $i$-th row. To prove (6) we need the following:

$$[(X'X)^{-1} + \left(\frac{1}{1 - x_i'(X'X)^{-1}x_i'}\right)(X'X)^{-1}x_i'x_i^{-1}(X'X)^{-1}][(X'X) - x_i'x_i'] = I,$$

(7)

where $I$ is the unit matrix, $x_i$ is the vector with the values for the $i$-th observation. Equation 7 can be easily verified if we note that $x_i'(X'X)^{-1}x_i' \neq 1$ is a number. From (7) by right multiplication with $(X'X) - x_i'x_i'^{-1}$ we have:

$$(X'X - x_i'x_i'^{-1})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i'x_i^{-1}(X'X)^{-1}}{1 - x_i'(X'X)^{-1}x_i}. $$

(8)

Let us note that (8) can be directly obtained by the Bartlett’s matrix inversion formula [1]. To continue, we use one trick from the matrix algebra, $X_i'X = \sum_{i=1}^n x_i'x_i$. From it we have:

$$X_i'X_i = X'X - x_i'x_i' \implies (X_i'X_i)^{-1} = (X'X - x_i'x_i')^{-1}. $$

(9)

From (8) and (9), and a right multiplication with $X_i'Y_i$, we have:

$$(X_i'X_i)^{-1}X_i'Y_i = \left((X'X)^{-1} + \frac{(X'X)^{-1}x_i'x_i^{-1}(X'X)^{-1}}{1 - x_i'(X'X)^{-1}x_i}\right)X_i'Y_i.$$

(10)
We know that the coefficients of a regression model can be obtained with the formula
\[ \hat{\alpha} = (X'X)^{-1}X'Y, \]
see [12, p. 30]. Additionally, we have \( X'_i Y_i = X'Y - x_i y_i \). Then (10) becomes:
\[ \hat{\alpha}_i = \left( (X'X)^{-1} + \frac{(X'X)^{-1}x_i x_i' (X'X)^{-1}}{1 - x_i' (X'X)^{-1} x_i} \right) (X'Y - x_i y_i). \] (11)

We know from the theory that the hat matrix of a regression model is \( H = X(X'X)^{-1}X' \), and its diagonal contains the leverage of each of the observations - \( h_{i,i} = x_i' (X'X)^{-1} x_i \) is the i-th diagonal element which corresponds to the leverage of the i-th observation, see [13]. Expanding the brackets in (11) we have:
\[ \hat{\alpha}_i = (X'X)^{-1}X'Y - (X'X)^{-1}x_i y_i + \left( \frac{(X'X)^{-1}x_i x_i' (X'X)^{-1}}{1 - h_{i,i}} \right) x_i y_i = \hat{\alpha} + (X'X)^{-1}x_i \left( -y_i + \frac{x_i' \hat{\alpha}}{1 - h_{i,i}} - \frac{h_{i,i} y_i}{1 - h_{i,i}} \right) = \hat{\alpha} + (X'X)^{-1}x_i \left( \frac{x_i' \hat{\alpha} - y_i}{1 - h_{i,i}} \right) \] (12)

The residual for the i-th observation is \( \hat{\varepsilon}_i = y_i - x_i' \hat{\alpha} \), then:
\[ \hat{\alpha}_i = \hat{\alpha} - \frac{\hat{\varepsilon}_i}{1 - h_{i,i}} (X'X)^{-1} x_i. \] (13)

We know that \( \varepsilon_i = y_i - x_i' \hat{\alpha}_i \) and \( \hat{\varepsilon}_i = y_i - x_i' \hat{\alpha} \). We multiply (13) by \( x_i' \):
\[ x_i' \hat{\alpha}_i = x_i' \hat{\alpha} - \frac{\hat{\varepsilon}_i}{1 - h_{i,i}} x_i' (X'X)^{-1} x_i \Leftrightarrow y_i - \varepsilon_i = y_i - \hat{\varepsilon}_i = \frac{\hat{\varepsilon}_i h_{i,i}}{1 - h_{i,i}} \Leftrightarrow \]
\[ \Leftrightarrow \varepsilon_i = \frac{\hat{\varepsilon}_i}{1 - h_{i,i}}, \; i = 1, \ldots, n. \] (14)

**Corollary 1** We can compute the RMSECV by running only one regression making the computation almost as fast as the computation of a single regression model.

From the theorem we now see that the out-of-sample residuals \( \varepsilon_i, \; i = 1, \ldots, n \) depend from the model errors \( \hat{\varepsilon}_i, \; i = 1, \ldots, n \) and from their corresponding leverage values \( h_{i,i}, \; i = 1, \ldots, n \). The leverage \( h_{i,i} \in [0, 1] \) of observation \( i \) shows how much the values of the input variables for the i-th observation will influence the model to obtain an estimate of the i-th observation close to the real output value \( y_i \) of the i-th observation, for more information see [5, 14]. For the proposed method, while making a component \( Z \) we are searching for the optimum component \( Z(k) \) that will form a design matrix which will balance between its leverage values and the obtained...
model errors in order to lower the out-of-sample errors. With other words, the *Cross-validated sequentially constructed multiple regression* procedure is trying to make a model that is more robust to the input data with respect to smaller out-of-sample errors by replacing some of the model variables with components.

### 3 An Example

The proposed step by step procedure for obtaining components is tested on a real example concerning accounting and macroeconomic information from the firms in the Bulgarian gas distribution sector in the period 2007–14. The goal is the predicting of the *Return on assets* (ROA) financial ratio for the next observed period $t$ using the input data from the current period $t - 1$. The full data set consists of 116 observations (three of these observations are omitted to improve the model). The model has seven variables one of which is a macroeconomic one and the others are financial ratios. The description of the input will be skipped because we want to highlight the proposed method, not the example.

\[
ROA(t) = \alpha_0 + \alpha_1 * X_1(t-1) + \alpha_2 * X_2(t-1) + \alpha_3 * X_3(t-1) + \alpha_4 * X_4(t-1) \\
+ \alpha_5 * X_5(t-1) + \alpha_6 * X_6(t-1) + \alpha_7 * X_7(t-1). \tag{15}
\]

The new method will be demonstrated against a multiple regression by several steps. First, the proposed method is applied to a training set $A_{2010}$ which contains the observations from years 2007–2010 (45 observations) to derive the needed components. The derived components by the method are three (43% of the primary variables):

\[
C_1 = X_1, \\
C_2 = X_2 + 0.024 * (-0.93 * (X_3 + 0.26 * X_7) + X_5), \\
C_3 = -0.66 * X_4 + X_6. \tag{16}
\]

Only these components will be used further in the example. Let us note that the components are derived from a subset over which the multiple regression has some insignificant coefficients. The aim is to demonstrate that components that are estimated over not so perfect data set can still have robust prediction results while further used. In the perfect case, it would be preferable to estimate the components from a subset over which a multiple regression model has significant model coefficients and desirable out-of-sample performance.

Second, we start to update the models that uses the three components $C_1$, $C_2$, $C_3$ and the multiple regression that uses the primary model variables $X_1$, ..., $X_7$ by adding the observations from year 2011 to the learning set $A_{2010}$. We compared them using the RMSE measure on the left out-of-sample observations, in this case they are from years 2012, 2013, 2014 (49 observations). Third, we add year 2012 to the learning