Online Parameter Selection for Gas Distribution Mapping

Victor Hernandez Bennetts[a*], Marco Trincavelli[a], Achim J. Lilienthal[a],
Victor Pomareda[b] and Erik Schaffernicht[a]

[a] Centre for Applied Autonomous Sensor Systems, Örebro University, SE-70182. Orebro, Sweden.

[b] Institute for Bioengineering of Catalonia. University of Barcelona, Barcelona, Spain.

[*] Corresponding author

Email: victor.hernandez@oru.se

Abstract - The ability to produce truthful maps of the distribution of one or more gases is beneficial for applications ranging from environmental monitoring to mines and industrial plants surveillance. Realistic environments are often too complicated for applying analytical gas plume models or performing reliable CFD simulations, making data-driven statistical gas distribution models the most attractive alternative. However, statistical models for gas distribution modelling, often rely on a set of meta-parameters that need to be learned from the data through Cross Validation (CV) techniques. CV techniques are computationally expensive and therefore need to be computed offline. As a faster alternative, we propose a parameter selection method based on Virtual Leave-One-Out Cross Validation (VLOOCV) that enables online learning of meta-parameters. In particular, we consider the Kernel DM+V, one of the most well studied algorithms for statistical gas distribution mapping, which relies on a meta-parameter, the kernel bandwidth. We validate the proposed VLOOCV method on a set of indoor and outdoor experiments where a mobile robot with a Photo Ionization Detector (PID) was collecting gas measurements. The approximation provided by the proposed VLOOCV method achieves very similar results to plain Cross Validation at a fraction of the computational cost. This is an important step in the development of on-line statistical gas distribution modelling algorithms.

Keywords - bandwidth selection, gas distribution mapping, virtual leave-one-out cross validation.

1 Introduction

Gas Distribution Modelling (GDM) can be seen as the task of deriving truthful representations of the observed gas distribution from a set of spatially and temporally distributed measurements of relevant variables, foremost gas concentration but also wind, pressure and temperature[1]. Gas distribution maps can be useful in several real world application scenarios. For example, gas leaks can be located using the implicit information contained in the gas distribution model and exposure of human operators to hazardous compounds can be minimized by avoiding areas of high concentration highlighted in the map.

Gas monitoring in industrial scenarios is commonly carried out by means of a fixed grid of sensing nodes, which are placed at a few pre-defined locations. While this approach is cost effective, it lacks the flexibility of a mobile platform. Among other advantages, gas sensitive mobile robots can provide adaptive measuring at a higher spatial resolution and they can be efficiently deployed in emergency scenarios.

The Kernel DM+V algorithm[2], is a well known approach for gas distribution modelling. Kernel DM+V not only predicts the mean gas concentration at a given query point but also provides a measurement of uncertainty in the form of a predictive variance. At the core of the algorithm is the well known Nadaraya-Watson estimator with RBF kernel[3], which is applied twice in an intertwined manner, once for the estimation of the predictive mean and once for the estimation of the predictive variance.

The predictive capabilities of a model generated with Kernel DM+V are determined by the selection of an appropriate kernel bandwidth. The kernel bandwidth is selected, from a grid search space, by minimizing a loss function using K-fold Cross Validation (CV). The computational cost of this method is high due to the need for building and evaluating multiple models on different parts of the data for CV. Therefore, the selection of the kernel has to be carried out offline.

In this work, an alternative approach towards online bandwidth selection is presented. The proposed algorithm eases the computational demands of CV by using Virtual Leave-One-Out Cross Validation (VLOOCV) for model evaluation. VLOOCV does not require dividing the data into multiple training and testing sets. Instead, a single model is trained and evaluated using the full dataset and its loss score is adjusted using a set of leverage factors. In addition, we demonstrate that the particular characteristics of the measurement data can be exploited to further reduce the computation time by obviating the calculation of said leverage factors. This opens the possibility of performing bandwidth selection, while measurements are being collected. The evaluation of the proposed method was carried out using data collected (in indoor and outdoor locations) with a mobile robot equipped with a Photo Ionization Detector (PID). The results show that model evaluation using VLOOCV can be performed at a fraction of the computational cost of CV yielding comparable models.

While the present work was evaluated in the task of GDM, this does not necessarily imply that the proposed algorithm is limited to GDM. For example, tasks such as gas discrimination[4], where the hyper parameters of a classification model have to be learned, could use VLOOCV to reduce the computational cost.

This paper is structured as follows: In Section 2, the Kernel DM+V equations are introduced. Section 3, describes the proposed bandwidth selection approach. In Section 4 and Section 5, the experimental set-up and results are presented. Section 6 summarizes the paper and lists the conclusions of this work.

2 The kernel DM+V Algorithm

As a model free approach, Kernel DM+V does not make strong assumptions about the particular functional form of the modelled gas distribution[2]. Kernel DM+V discretizes the exploration area into a grid of cells and uses the Nadaraya-Watson estimator[3] twice to model the mean distribution and the predictive variance ($\hat{c}(\mathbf{x}_k)$ and $\hat{v}(\boldsymbol{x}_k)$ respectively) at the center of each cell k in the grid. In its iterative form, for each new concentration measurement c_i , acquired at a given location $\mathbf{x_i}$, the mean distribution map is given by:

$$\omega_i^k = \omega_{i-1}^k + \mathcal{N}(|x_i - x_k|, \sigma) \tag{1}$$

$$C_i^k = \frac{C_{i-1}^k \cdot \omega_{i-1}^k + \mathcal{N}(|x_i - x_k|, \sigma) \cdot c_i}{\omega_i^k} \tag{2}$$

$$\alpha_i^k = 1 - e^{-\frac{\omega_i^k}{g_s}} \tag{3}$$

$$\hat{c}_i(x_k) = \alpha_i^k \cdot C_i^k + (1 - \alpha_i^k) \cdot c_0 \tag{4}$$

In Equations (1) and (2), $\mathcal{N}$ is the RBF kernel that weights the importance of measurement c_i, according to its distance from the cell center x_k. Thus, for each cell in the grid, a weighted average C_i^k is computed. The term σ (i.e. ”kernel bandwidth”) controls the smoothing level of $\mathcal{N}$ and thus, a proper selection of σ determines the predictive capabilities of the model.

Equation (3) assigns a confidence value (bounded between 0 and 1) to each cell, according to the number of neighbouring measurements. The term g_s, is a scaling parameter[2] that can be tied to the selection of σ using $g_s = \mathcal{N}(0, \sigma)$. The final estimation of the mean distribution map in Equation (4), is given by a weighted sum between C_i^k and c_0, which is a prior assumption regarding the gas concentration, for example, the global mean of the acquired measurements.

The variance map is computed using Equations (5) and (6). Notice that, contrary to the computation of the mean map, it is not expressed iteratively. This is due to the fact that the weighted square prediction error (Equation (5)) has to be computed for each measurement, using the latest update on $\hat{c}_i$. Hence, the computational cost of the Kernel DM+V is dominated by the calculation of $\hat{v}(x_k)$ and is linear to the number of measurements.

$$V_i^k = \frac{\sum_{j=1}^{i}[\mathcal{N}(|x_i - x_k|, \sigma) \cdot (c_j - \hat{c}_i(x_k))^2]}{\omega_i^k} \tag{5}$$

$$\hat{v}_i(x_k) = \alpha_i^k \cdot V_i^k + (1 - \alpha_i^k) \cdot v_0 \tag{6}$$

2.1 Bandwidth Selection

The performance of the regression model, derived by Kernel DM+V, depends on the selection of the kernel bandwidth σ. A common method to select σ is to perform Cross Validation (CV) over a search space $\boldsymbol{\sigma}$=[σ_1,…,σ_m] with m being the number of bandwidths to evaluate. In CV, the training set is randomly partitioned into *K* folds, where *K-1* partitions are used to train a model and the remaining fold is used for validation purposes. This process requires to train and test $K \times m$ models.

The optimal bandwidth σ_o can be found in the search space by minimizing a loss function *E* that measures the predictive capabilities of the model. For uncertain predictive models, as in the case of Kernel DM+V, *E* should not only rank the model's accuracy, but it should also aim for more balanced models in which, wrong predictions made with high certainty are penalized. The Negative Log Predictive Density (NLPD) is a loss function that favours rather conservative models[5]. This means that models that are not over confident in their prediction are preferred. Under the assumption of a Gaussian posterior $p(c_i|x_i)$, the NLPD of a set of *D* unseen measurements $\{c, x\}$ is computed by:

$$E = \frac{1}{2D}\sum_{i=1}^{D}\left(log\ \hat{v}(x_i) + \frac{(c_i - \hat{c}(x_i))^2}{\hat{v}(x_i)}\right) + \frac{1}{2}log\ (2\pi) \tag{7}$$

Thus, as mentioned above, the computation of a gas distribution model using Kernel DM+V is dominated by the computation of the variance map since, contrary to the mean distribution map, it has to be computed from scratch each time a new measurement is acquired. An update of the maps together with selection of an optimal kernel bandwidth thus requires to perform $K \times m \times N \times N_g$ operations, with N being the number of measurements and N_g being the number of cells in the map.

While several methods have been proposed in the past to avoid CV by introducing penalizations for complex models (i.e. Akaike, BIC)[6], these methods are not suitable for Kernel DM+V, since they base the selection only on the mean of the estimation, not considering the variance (i.e. the uncertainty in the prediction).

3 VLOOCV Parameter Selection

Monari and co-authors proposed in 2002 the Virtual Leave One Out Cross Validation (VLOOCV) method[7]. VLOOCV relies on the assumption that the withdrawal of a single example from the training set will yield a model that is not substantially different from the model that is obtained by training on the full dataset[8]. VLOOCV computes a leverage factor h_j for each of the training data points, which measures the influence of the training example j in the computation of the model. If a given data point has a large influence on the model computation, h_j will be close to 1. On the other hand, when h_j is close to 0, the data point has little effect on the model regardless of its presence or absence in the training set. VLOOCV approximates the loss function as follows:

$$E_j^{(-j)} \cong \frac{E_j}{1 - h_j} \tag{8}$$

where $E_j^{(-j)}$ is the loss on data sample j when it is left out of the training set and E_j is the loss when data sample j is included in the training error. In the specific case of an uncertain regression model, E_j can be given by the NLPD computed for the data sample j.

VLOOCV can be used to reduce the computations to select σ_o. Instead of generating K models, a single model is trained using the whole dataset for each possible σ in the search space $\boldsymbol{\sigma}$, and the leverage factors are computed from:

$$\mathbf{H}_{\sigma i} = \mathbf{Z}_{\sigma i}(\mathbf{Z}_{\sigma i}^T\mathbf{Z}_{\sigma i})^{-1}\mathbf{Z}_{\sigma i}\big|_{\sigma i \in \boldsymbol{\sigma}} \tag{9}$$

where $\mathbf{Z}$ is a $n \times m$ matrix composed by the numerical gradients of the NLDP values w.r.t. the m elements in the search space $\boldsymbol{\sigma}$. Thus, the leverage factor $h_{\sigma i}^j$ for a training point j in the model computed using σ_i is the jth element in the diagonal of matrix $\mathbf{H}_{\sigma i}$. In this way, the NLPD computation for each data sample is given as follows:

$$E_{\sigma i}^j = \frac{1}{2(1 - h_{\sigma i}^j)}\left(log\ \hat{v}(x_j) + \frac{c_j - \hat{c}(x_j)}{\hat{v}(x_j)}\right) + \frac{1}{2}log\ (2\pi) \tag{10}$$

In this way, the number of operations required to update the distribution maps, and select an optimal kernel bandwidth using VLOOCV is $m \times N \times N_g$. Thus, the computational complexity of VLOOCV is still linear in the number of measurements, as in the case of CV, but at a much smaller factor.

4 Experimental Set-up

The proposed algorithm was validated using data collected in two different locations: a robot arena and an outdoor courtyard. In both scenarios, a robotic platform equipped with a Photo Ionization Detector (PID) was used. In all the experiments, ethanol was used as a gaseous source. Ethanol is invisible in air and, in small quantities, it is harmless for humans. With a boiling point of 78.4 °C, ethanol evaporates quite quickly at room temperature and, since it is heavier than air, it propagates at ground level.

4.1 Robotic Platform

Two different robotic platforms were used to collect data. In the outdoor experiments, an all terrain ATVR-JR platform was used (Figure 1(c)). The space restrictions imposed by the robot arena require a smaller, more manoeuvrable platform, like the Pioneer P3-DX (Figure 1(a)) from MobileRobots. On both robots, a laser scanner (SICK LMS 200) and *out of the box* robotics software packages[1] were used for localization purposes.

4.2 Gas Sensing

Gas concentration measurements with both platforms were carried out using a ppb RAE 3000 photo Ionization Detector (PID) from RAE Systems. According to the manufacturer, it is linearly proportional to the concentration of the chemical compound being analyzed. The sampling frequency of the PID is 4 Hz.

4.3 Robot Arena

The robot arena (Figure 1(b)) comprises an area of $5 \times 5 \times 2\ m$ and, while no artificial airflow was induced, a weak circulating airflow field (0.01-0.03 m/s) was formed in the room by natural convection. Ethanol was released at a constant rate of 0.2 l/min from a set of tubes located on the floor. A robotic platform was commanded to follow a spiral trajectory inside the arena, stopping at waypoints for 30 s.

4.4 Outdoor Courtyard

A set of experiments were conducted in an outdoor courtyard at the Örebro University main campus (Figure 1(d)). The robot was programmed to follow a random trajectory inside a $9 \times 5\ m$ area, stopping at pre-defined way-points for 30 s. An open plastic container filled with ethanol was placed on the floor with a *bubbler* that was used to facilitate evaporation. Fans were placed at the borders of the exploration area to facilitate the dispersion of ethanol over the inspection area.

5 Results

Figures 2(a) and 2(b) show NLPD plots versus kernel size for both indoor and outdoor locations. The plots are calculated using 5 fold CV (red curve), VLOOCV (green curve) and, in order to evaluate the effect of the leverage factors h, a plot of the VLOOCV without leverage factor is also included (blue line). It can be seen that, for both locations, there exists a good agreement between the minima obtained with CV, VLOOCV, and VLOOCV without leverage corrections (0.13 m, 0.12 m and 0.12 m respectively for the indoor experiments and 0.13 m, 0.10 m and 0.10 m for the outdoor experiments).

In Figure 2(c), the computation time of the three algorithms for a different number of measurement points is shown. The computational complexity of the update of the CV and VLOOCV algorithms is dominated by the update of the variance map and the calculation of the NLPD. The complexity of CV scales in addition with the number of folds while VLOOCV requires computing leverage factors which as well scales linearly with the number of measurement points. The VLOOCV algorithm is computationally less expensive, while preserving

[1] Adaptive Monte-Carlo localization packages in ROS (Robot Operating System). http://www.ros.org

the shape of the objective function. It is worth noting that correcting the negative likelihood with leverage factors does hardly change the VLOOCV result (the green and blue curves in Figures 2(a) and 2(b) are almost coincident). This suggests that the computation of the leverage factors is not needed for gas distribution mapping data obtained with mobile robots. An explanation is that the models obtained with the full dataset and with a fraction of the dataset (used in CV) are very similar as shown in Figure 2(d). The score displayed in the plot is the overlapping coefficient[9], which is the intersection of the normal distributions (1 identical distributions, 0 distributions that do not intersect) predicted by the model computed with the full dataset and with fractions (folds) of the dataset. The solid line represents the median of the overlapping coefficients while the dashed line represents the first quartile of the data. This means that at least 75% of the data lies above the dashed line. It can be observed as well in Figure 2(d) that, when the number of folds increases (towards leave one out CV), the similarity between the model obtained with the full dataset and with a fraction of the dataset (computed on K-1 folds) increases as well. This confirms the initial assumption that the withdrawal of a single example from the training does not significantly change the computation of the model. This result is expected, since the robot is constantly collecting data at 4 Hz and moving at a speed of 0.1 m/s making the training dataset highly redundant.

6 Summary and Conclusions

In this work, a method to perform online parameter selection for GDM is presented. Instead of the more traditional CV (e.g. K-folds, leave one out), we propose to use Virtual Leave One Out Cross Validation (VLOOCV), to decrease the computation time needed to perform a grid search over a set of possible mapping parameters (i.e. kernel bandwidth). VLOOCV assumes that the withdrawal of a single example from the training set will not alter substantially the model that is computed by training on the full dataset and therefore, model evaluation can be performed by introducing a leverage factor that adjusts the loss function (i.e. the NLPD).

The proposed algorithm was evaluated with data from indoor and outdoor scenarios and the results consistently indicate that, compared to K-fold CV, similar parameter selection can be achieved using VLOOCV. The key advantage of the proposed approach is that the computational time can be substantially decreased. This is due to the fact that VLOOCV needs to evaluate only one model for each of the *m* possible models that can be generated from the kernel bandwidth search space. In comparison, K-fold cross validation needs to evaluate *K* models for each kernel bandwidth in the search space.

In addition, it was observed that the computational demand can be further reduced by obviating the leverage factors, since they do not significantly alter the shape of the loss curve (i.e. NLPD vs. kernel size). This can be attributed to the particular structure of the gas sensing data which is redundant due to the relatively slow process of gas dispersion and high sampling frequency. Thus, the leverage factor at each sampling point is, for most of the cases, close to 0.

The applicability of the presented algorithm is not limited to the task of GDM. The possibility of performing parameter selection using VLOOCV could be further explored in related applications such as gas discrimination and gas quantification. In these applications, as in the case of GDM, the computation of the leverage factors would be subject to the characteristics of the collected datasets.

7 Acknowledgements

This work has partly been supported by Robotdalen[2] in the Gasbot project (project number 8140).

[2] www.robotdalen.se

8 Bibliography

1. S. Asadi, M. Reggente, C. Stachniss, C. Plagemann and A. J. Lilienthal, in *Intelligent Systems for Machine Olfaction: Tools and Methodologies*, Edited by E. L. Hines and M. S. Leeson, IGI Global, **(2011)**, pp. 153-179.

2. A. J. Lilienthal, M. Reggente, M. Trincavelli, J. L. Blanco and J. Gonzalez, in *IROS*, St. Louis, USA, **(2009)**, pp. 570-576.

3. E. Naradaya, in *Theory of Probability and Its Applications,* **(1964)**, vol. 9, no. 1, pp. 141-142,.

4. M. Trincavelli, in *Künstliche Intelligenz,* **(2011)** vol. 25, no. 4, pp. 351-354.

5. J. Quinonero Candela, C. E. Rasmussen, F. Sinz and B. Schoelkopf, in *Machine Learning Challenges*, Springer, **(2006)**, pp. 1-27.

6. M. Köhler, A. Schindler and S. Sperlich, in *A Review and Comparison of Bandwidth Selection Methods for Kernel Regression*, tech report, Courant Research Centre PEG**, (2011)**.

7. G. Monari and G. Dreyfus, in *Neural Computation*, **(2002)**, vol. 14, no. 6, pp. 1481-1506.

8. G. Monari and G. Dreyfus, in *Neurocomputing*, **(2000)**, vol. 35, pp. 195-201.

9. H. F. Inman and E. L. Bradley, in *Communications in Statistics - Theory and Methods*, **(1989)**, vol. 18, no. 10, pp. 3851-3874.

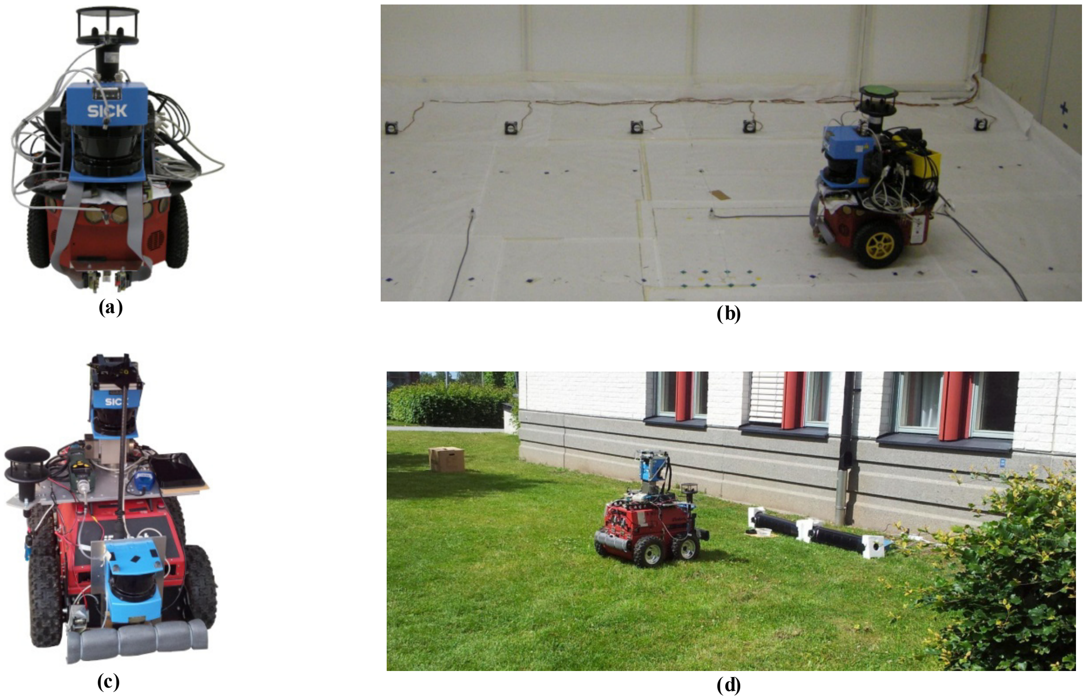

Figure 1: Experimental scenarios and robotic platforms. (a) P3-DX platform. (b) Robot arena. (c) ATVR-JR platform. (d) Outdoor courtyard.

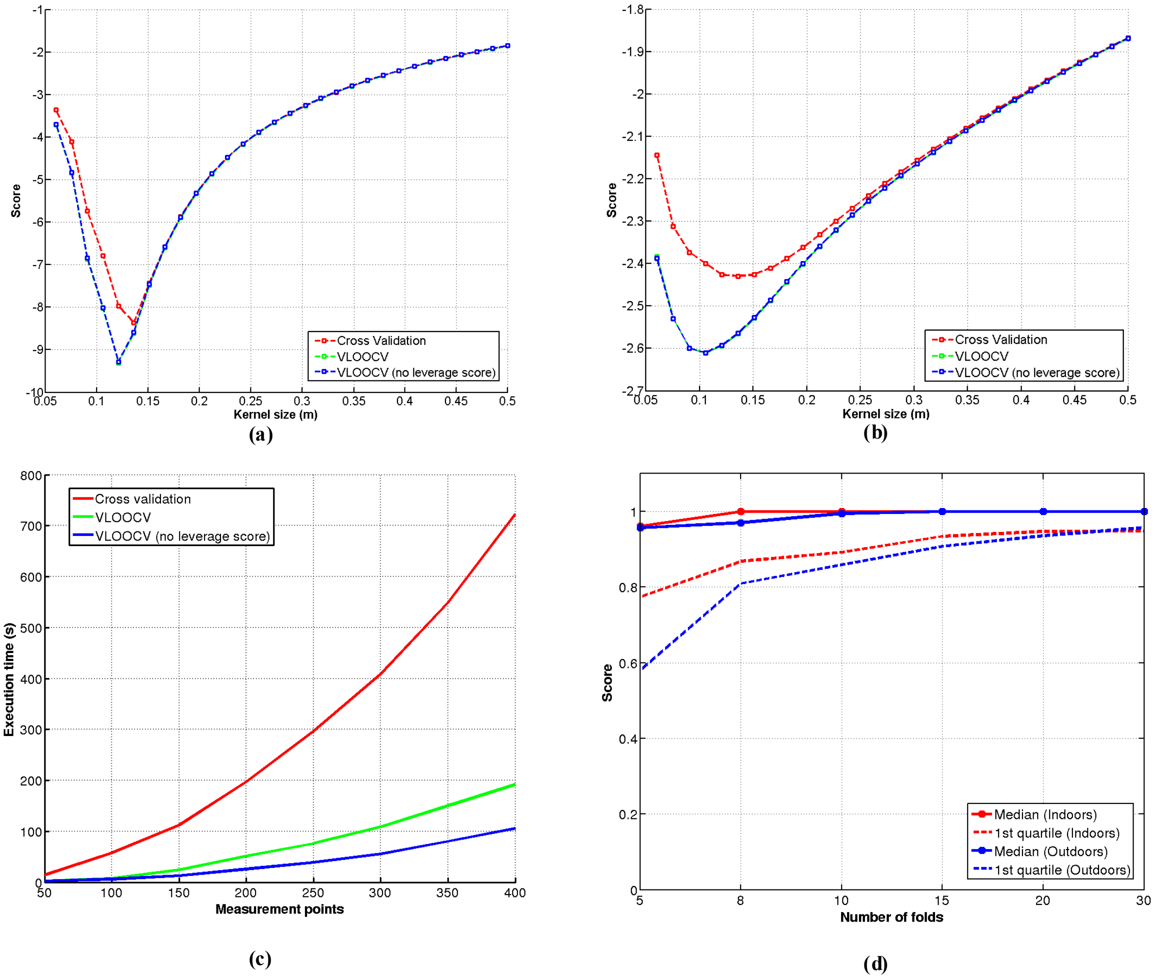

Figure 2: Experimental results in indoor and outdoor locations. (a) NLPD curves versus kernel size, indoor experiments. (b) NLPD curves versus kernel size, outdoor experiments. (c) Computation time. (d) Overlapping coefficient of the model calculated on the full dataset and the models calculated on fractions (folds) of the dataset. When the number of folds increases (towards leave one out CV), the similarity between the model obtained with the full dataset and with a fraction of the dataset (computed on K-1 folds) increases as well.