Source Parameter Estimation of Atmospheric Pollution Using Regularized Least Squares

Huimin Chen  
Department of Electrical Engineering  
University of New Orleans  
New Orleans, LA, U.S.A.  
Email: hchen2@uno.edu

X. Rong Li  
Department of Electrical Engineering  
University of New Orleans  
New Orleans, LA, U.S.A.  
Email: xli@uno.edu

Abstract—This paper presents a regularized least squares method to estimate the location and release rate of atmospheric pollution. We assume that measured pollution concentration at different ground locations and meteorological conditions such as wind speed are available so that one can solve the advection-diffusion equation for a non-steady point source. However, even finding linear parameters related to the release rate is an ill-posed problem and one has to impose certain regularization technique to avoid potential overfit. We propose to use $\ell_p$-regularization $(0 \leq p \leq 1)$ and discuss its advantage over the popularly used $\ell_2$-regularization. The accuracy of source parameter estimation is examined for the cases where both the number of sources and the corresponding locations are unknown.

Keywords: Source localization, regularized least squares, model selection, air pollution model.

I. INTRODUCTION

Post-accident management is crucial for public protection from potential accidents resulting from dangerous gas leakages. The determination of source origins and release rates will be useful for the forecast of gas concentration in the atmosphere and for the management staff to prioritize off-site evacuation plans. A lot of research has been focused on detecting and localizing single or multiple plume sources with autonomous vehicles or sensor networks such as [16] for a vapor-emitting source, [3] for a nuclear source, and [11], [12] for a chemical source. In [10] the plume detection and localization problem is formulated as abrupt change detection using sparse sensor measurements. The development of a large scale testbed has been reported in [7] for plume detection, identification and tracking. Despite the abundant literature in plume detection [4], [15], [17] and localization [9], [20], [21], limited efforts have been made toward solving the source localization and parameter estimation problem jointly. The main reason is that even finding linear parameters related to the source release rate is an ill-posed problem and one has to impose certain regularization technique to avoid potential overfit. We adopt the least squares technique based on the solution to the advection-diffusion equation [13], [14] and impose $\ell_p$-regularization for $0 \leq p \leq 1$ [5], [6] to characterize the sparsity of the unknown source release rate signal. We also discuss its advantage over the popularly used $\ell_2$-regularization. The accuracy of source parameter estimation is examined for the cases where both the number of sources and the corresponding locations are unknown.

The rest of the paper is organized as follows. Section II introduces the plume localization and parameter estimation problem. Section III presents the regularized least squares solution to the parameter estimation problem. Section IV discusses the implementation of the joint model selection and parameter estimation algorithm and the choice of regularization parameter. Section V compares our approach with alternative regularization methods. Section VI provides realistic source release scenarios to assess the performance of the proposed algorithm. Concluding summary is in Section VII.

II. PROBLEM FORMULATION

We assume that a Cartesian coordinate system is used with $x$-axis oriented towards the mean wind direction, $y$-axis in the cross-wind direction and $z$-axis in the vertical direction. If the source of a pollutant is located at $(x_0, y_0, z_0)$ with release rate $q(t)$, then at time $t$, the concentration of the pollutant at some down-stream location $(x, y, 0)$ can be written as [13]

$$C(x, y, 0, t) = \int_0^t K(t, \tau)q(\tau)d\tau \quad (1)$$

where the kernel $K(t, \tau)$ is

$$K(t, \tau) = \exp[-\frac{\|x-x_0-u(t-\tau)\|^2}{4K_x(t-\tau)} - \frac{(y-y_0)^2}{4K_y(t-\tau)} - \frac{z_0^2}{4K_z(t-\tau)}]$$

$$8\pi^2(K_xK_yK_z)(t-\tau)^2 \quad (2)$$

with mean wind speed $u$ and diffusion coefficients $K_x$, $K_y$, $K_z$ along $x$, $y$ and $z$ directions, respectively.

We assume that there are $J$ sensors deployed at fixed locations where sensor $j$ is located at $(x_j, y_j, 0)$ and collects $N$ concentration readings $c_j = \{C(x_j, y_j, 0, t_n)\}_{n=1}^N$. Denote by $c = \{c_j\}_j$ the collection of all sensor readings. Denote by $q = \{q(\tau_i)\}_{i=1}^M$ the discretized source release sequence where $q(\tau_i)$ is the release rate at time $\tau_i$. Ideally, we have the following observation equation

$$c = A(p)q \quad (3)$$

where $p = (x_0, y_0, z_0)$ denotes the unknown source location. Note that for measurement $c_j(x_j, y_j, 0, t_n)$, the corresponding
element $a_{(j,n,k)}$ in $A(p)$ is given by
\[ a_{(j,n,k)} = K(t_n, \tau_k)\beta_{nk} \] (4)
where $\beta_{nk}$ is a quadrature weight [13], [14]. The estimation of source location $p$ and release rate $q$ can be formulated as the least squares problem given by
\[ (p^*, q^*) = \arg\min_{p,q} ||c - A(p)q||_2^2 \] (5)
Note that this formulation is valid only for a single source. To extend the estimation problem to include multiple sources, we assume that the concentration readings are the results of aggregation from multiple source releases. Assume that there are $s$ sources with unknown locations $\{p(i)\}_{i=1}^s$ and release rate sequences $\{q(i)\}_{i=1}^s$. Then we have the following ideal observation equation
\[ c = \sum_{i=1}^s A(p(i))q(i) \] (6)
The source parameter estimation problem becomes identifying the number of sources, the corresponding origins and the release sequences jointly using only the concentration readings from multiple sensors.

III. REGULARIZED LEAST SQUARES

In the single source case, the matrix $A$ in (5) can be analyzed for various source locations. By ranking the singular values of $A$, the authors of [13] found that the discrete time least squares problem (5) is in general ill-posed and suggested to use the Tikhonov regularization to ensure certain smoothness of $q$. However, for a source with an instantaneous release, the sequence $q$ may have only one spike that violates the smoothness assumption. Nevertheless, for multiple sources with instantaneous releases, we will observe the aggregated sparse signal with an unknown number of spikes. In fact, the sparsity assumption is crucial for the identification of multiple sources with instantaneous releases at different times. To encourage the sparsity of the release rate sequence estimate, we propose to use $\ell_p$-regularized least squares as the objective function, i.e.,
\[ J(p, q) = ||c - A(p)q||_2^2 + \lambda||q||_p \] (7)
where the regularization parameters $p$ controls the sparsity of the solution $q$ and $\lambda$ makes the tradeoff between the goodness-of-fit to the observations and the complexity of the model. Note that $p = 1$ is popularly used in compressed sensing [8] due to its numerical reliability. In fact, for any given $p$, minimizing (7) becomes a convex program if one chooses $p = 1$. However, our $\ell_1$-regularized least squares problem still requires non-convex optimization without knowing the source location $p$. In addition, when choosing the regularization term with $0 < p < 1$, one favors a more sparse solution than that using $\ell_1$-regularization [6]. In this case, $||\cdot||_p$ is not a norm, but $d(x, y) = ||x - y||_p^p$ is still a metric.

When the concentration readings are the aggregation of individual releases, we have to identify the number of sources and find their locations. In this case, we are facing a model selection problem, where model $s$ corresponds to $s$ sources with unknown locations $\{p(i)\}_{i=1}^s$ and release rate sequences $\{q(i)\}_{i=1}^s$. Assuming that different sources have different instantaneous release times so that there is no identifiability issue among the models, we can choose model $s$ that minimizes a modified version of (7)
\[ J(s, p(s), q(s)) = ||c - \sum_{i=1}^s A(p(i))q(i)||_2^2 + \text{pen}(p(s), q(s)) \] (8)
where
\[ \text{pen}(p(s), q(s)) = \lambda \sum_{i=1}^s ||q(i)||_p + 3s \log N \] (9)
Note that the first term in the penalty encourages the sparsity of each identified source release sequence and the second term is for model complexity of the source location parameter based on the Bayesian information criterion [18]. The second term is necessary because one does not want to treat one source with two instantaneous releases (sparsity of 2) as two different sources with instantaneous releases at different time instances (sparsity of 1). In practice, when the locations of two sources are close, they could be identified as a single source with aggregated release rate sequence. This seems to be acceptable when the locations of multiple sources are within the range of the localization accuracy obtained by minimizing (8).

IV. MODEL SELECTION AND IMPLEMENTATION ISSUES

Finding the optimal solution to (8) requires solving a high dimensional nonlinear optimization problem for any fixed regularization parameters $p$ and $\lambda$. In practice, the number of sources is usually small and a strong source can have the dominant effect on the sensor readings. Thus it would be meaningful to identify and localize one source at a time by treating the impact from the remaining possible sources as additive noise. In this case, assuming the source location is given, one can obtain the sparse solution of the release rate sequence by solving the following optimization problem:
\[ \min_q ||c - Aq||_2^2 + \lambda||q||_p \] (10)
When $p = 1$, the problem becomes a convex program and is highly related to LASSO [19]. Once we obtain the release rate of the source, we can refine the estimate of source location by solving the regular nonlinear least squares problem given by
\[ \min_p ||c - A(p)q||_2^2 \] (11)
Note that for the newly estimated source location, the sparsity (non-zero locations) of the solution to (10) may change. We can iteratively update the release rate and source location estimate until the residue is comparable to the noise level of sensor readings.

We can extend the above procedure to deal with unknown number of sources. We apply a greedy heuristic algorithm that iteratively refines the estimate of signal sparsity and the
noise level to determine the appropriate regularization parameter. The algorithm simultaneously determines the number of sources, the corresponding locations and release rate sequences by the following steps.

1) Set \( s = 1 \).
2) **Initialization:** Set \( k = 0 \), \( q(s)_k = 0 \) with an initial guess of source location \( p(s)_k \).
3) **Refining the estimate:** Use Newton-Ralphson update
\[
q(s)_{k+1} = q(s)_k + A(p(s)_k)^T [c - A(p(s)_k)q(s)_k]
\]
(12) to refine the estimated source release sequence.
4) **Choosing regularization parameter:** Compute the median of the residue \( |c - A(p(s)_k)q(s)_{k+1}| \) and choose \( \lambda \) proportional to the estimated noise level.
5) **Denoising by soft thresholding:** Compute the sparse approximation of \( q(s)_{k+1} \) by \( q(s)_{k+1}^* = T(q(s)_{k+1}) \) where
\[
T(x) = \text{sign}(x) \max\{|x| - \lambda, 0\}
\]
(13)
6) **Source localization:** Solve the nonlinear least squares problem
\[
p(s)_{k+1} = \arg\min_{p(s)_k} ||c - A(p(s)_k)q(s)_{k+1}||_2^2
\]
(14)
7) **Model selection:** Set \( k = k + 1 \) and iterate until \( q(s) \) converges to a sparse solution \( q(s)^* \) or a predetermined maximum number of iterations \( k_{\text{max}} \) has reached. Subtract out the identified source from sensor readings.
\[
c \leftarrow c - A(p(s)^*)q(s)^*, \; s = s + 1
\]
(15)
Repeat steps 2–6 until
\[
J(s-1, p(s-1)^*, q(s-1)^*) < J(s, p(s)^*, q(s)^*)
\]
(16)
8) Declare the number of sources \( s-1 \), the corresponding locations \( p(s-1)^* \) and release rate sequences \( q(s-1)^* \).

For any given \( \lambda \) and \( p(s)^* \), the above iterative procedure converges to the optimal solution of (10) for \( p = 1 \) [2]. We used the median estimator of the residue to obtain the noise level which is robust against outliers. It is less sensitive to possible model mismatch than using the mean of the residue when we initially assume that there is a single (strong) source which results in the concentration readings while treating other (weak) sources as noise. Note that the dimension of \( p(s) \) only depends on the model order, i.e., the number of sources, which is usually much lower than the dimension of release sequences \( q(s) \). Thus solving the nonlinear least squares problem (14) is less computationally demanding than solving (8) directly.

When \( 0 < p < 1 \), (10) becomes non-convex program and any iterative procedure may be trapped at a local minimum. To encourage more sparsity of the release rate sequence with smaller \( p \) and solve (10) directly, we apply iterative reweighted least squares (IRLS) update and replace the soft thresholding step by
\[
q(s)_{k+1}^{(n)} = W(n)^T A(p(s)_k)^T [A(p(s)_k)W(n)^T A(p(s)_k)^T]^{-1} c
\]
(17)
where the weighting matrix \( W(n) \) is diagonal with entries
\[
w_i(n) = \left\{ \left[ q(\tau_i (n-1)) \right]^2 + \epsilon \right\}^{p/2-1}, \; i = 1, ..., M.
\]
(18)
The damping coefficient \( \epsilon \) is chosen to be relatively large initially and decreases to a very small number when the above iteration is close to converge. Note that the IRLS algorithm converges in less than 100 iterations most of the time in our simulation study. Even though there is no theoretical guarantee that the resulting solution is globally optimal, we suspect that it does approach to near global minimum since the solution quality improves when using smaller \( p \).

V. COMPARISON WITH OTHER REGULARIZATION TECHNIQUES

Tikhonov’s regularization has been proposed in [13], [14] which essentially uses the objective function
\[
J(s, p(s), q(s)) = \left|c - \sum_{i=1}^{s} A(p(i))q(i)\right|^2 + \lambda \sum_{i=1}^{s} ||Lq(i)||_2^2
\]
(19)
where \( L \) controls the smoothness of \( q(i) \) with the approximate form
\[
||Lq(i)||_2^2 \approx \int_0^{\tau_n} \left( \frac{dNq(i)}{dt} \right)^2 d\tau
\]
(20)
The popular choice for obtaining a smooth solution is \( N = 2 \). Unfortunately, the above regularization technique only works for continuous releases from well separated sources. We rely on the sparsity of \( q(s) \) to identify the model order \( s \), which is suitable for localizing multiple sources of instantaneous release type.

Another sparsity enforced estimator was proposed in [5] which essentially minimizes the following objective function
\[
J = \left| A(p(s))^T [c - A(p(s))q(s)] \right|_{\infty} + \lambda ||q(s)||_{l_1}
\]
(21)
For known source locations, the estimated release rate guarantees to recover all possible sparse signals with a large probability [5]. However, the above objective function is a non-smooth function of \( p(s) \), which is difficult to optimize when both source locations and release rate sequences are unknown. In practice, we fix the source locations \( p(s) \) and solve (21) via linear programming. Then we fix the release rate sequence \( q(s) \) and update \( p(s) \) in its gradient descent direction. The iteration continues until \( p(s) \) reaches a stationary solution and the sparsity of \( q(s) \) does not change.

VI. MODELING APPLICATIONS

We present the simulation study of source localization and release rate estimation using multiple sensors. We are interested in both model selection and source parameter estimation accuracy.
A. Scenario Generation

Consider a single source located at \((-40, 35, 12)\) with instantaneous release of \(q(10) = 2 \cdot 10^5\). We assume that the wind speed \(u = 1.8\) along \(x\)-axis and \(K_x = K_y = 12, K_z = 0.2113\). Five sensors, located at \((0,0), (15,15), (30,30), (45,45), (60,60)\), respectively, collect concentration readings synchronously with 100 samples per sensor. All sensors are on the ground with zero elevation. We add Gaussian noise to the sensor readings with standard deviation \(4 \cdot 10^{-3}\). Each sensor will have a plume detection when the concentration reading exceeds 0.01. Fig. 1 shows one realization of the concentration readings from the five sensors. We can see that sensor 1 has early detection while sensors 3–5 have relatively large peaks in the concentration readings.

We also considered the case of two sources where one source located at \((-40, 35, 12)\) has the instantaneous release of \(q(10) = 2 \cdot 10^5\) and the other located at \((-30, 15, 15)\) has the instantaneous release of \(q(50) = 10^5\). Fig. 2 shows one realization of the concentration readings from the five sensors. Compared with Fig. 1, we can barely see the effect of the second source release due to the detection delay and source aggregation.

B. Model Selection and Parameter Estimation Accuracy

We want to compare our \(\ell_p\)-regularization method with Tikhonov’s method [13], [14] (denoted by \(p = 2\)) and Dantzig selector [5] (denoted by \(p = \infty\)) for both one-source and two-source cases. Note that Tikhonov’s method is not appropriate for estimating instantaneous release rate, which is non-smooth. However, it is meaningful to study how the incorrect assumption in regularization may affect model selection accuracy. We estimated the probability of identifying the correct number of sources based on 100 realizations of each case. For those instances where the number of sources is correctly identified, we also computed the root mean square (RMS) error of the location estimate for each source. In the case of \(s = 2\), the RMS error of the second source is in parentheses. The results are listed in Table I. We can see that in the single source case, our \(\ell_p\)-regularization method can identify the correct number of sources almost perfectly. In the two-source case, Tikhonov’s method failed to identify the second source most of the time and Dantzig selector can only identify the correct number of sources with 64 out of 100 cases. Surprisingly, the proposed \(\ell_p\)-regularization method is able to find the correct model order with higher than 80% probability. As we reduce \(p\), there is a slight increase in the probability of obtaining the correct number of sources due to the strong enforcement of sparsity. Among all cases where the first source is correctly identified, the root mean square error of the estimated release rate is \(4.6 \cdot 10^2\) with \(p = 1\). Note that the root mean square error of estimated location of the first source increases when we have a second source aggregated to it. Note also that the algorithm assuming the correct model order can only achieve the root mean square error of estimated location of the second source around 18 using \(\ell_p\)-regularized method with \(p = 1\). These observations suggest that the \(\ell_p\)-regularized least squares method is effective in joint model selection and parameter estimation for instantaneous source release.

C. Model Mismatch to Continuous Release Source

Consider a single source located at \((-40, 35, 12)\) with continuous release rate
\[
q(t) = \left[e^{-0.1(t-10)}u(t-10) + 2e^{-0.5(t-50)}u(t-50)\right] \cdot 10^5
\]
One realization of the concentration readings from the five sensors is shown in Fig. 3. Note that the concentration readings from sensors 2–5 have not reached their peaks by the end of the samples. This will in general make the source parameter estimation more difficult. In 100 realizations, the $\ell_p$-regularized least squares method with $p = 1$ identified one source in 92 times and two sources in 8 times with their estimated locations close to each other. The incorrect identification of model order is due to the abrupt release at the two time instances $t = 10$ and $t = 50$ with exponential decay of the release rate. The root mean square error of the two time instances identification of model order is due to the abrupt release at their estimated locations close to each other. The incorrect one source in 92 times and two sources in 8 times with $\ell_p$-parameter estimation more difficult. In 100 realizations, the end of the samples. This will in general make the source readings from sensors 2–5 have not reached their peaks by another important research theme and demands future work.

![Fig. 3. Sensor readings for one source with continuous release.](image-url)

**VII. DISCUSSION AND CONCLUSIONS**

We have presented an $\ell_p$-regularized least squares method to estimate the location and release rate of atmospheric pollution. For $0 \leq p \leq 1$, the method enforces sparsity of the release sequence of each identified source. The proposed method can identify multiple sources of instantaneous release type and can also localize sources of continuous release. The accuracy of source parameter estimation has been examined for the cases where the number of sources and the corresponding locations are unknown. In general, the least squares method does not provide any measure of the estimation error. However, one can examine the residue and make additional assumptions such as additive Gaussian noise in order to quantify the covariance of the localization error. Through simulation study, we found that the proposed method is effective in localizing instantaneous release sources and has certain degree of tolerance to model mismatch. It is worth noting that the sensor locations, sampling rate and measurement accuracy can affect the source localization performance significantly. Finding the best sensor placement and sensing strategy in a given surveillance area is another important research theme and demands future work.

**ACKNOWLEDGMENT**

This work was supported in part by Air Force Research Lab # FA8650-07-M-1161, Navy through Planning Systems Contract # N68335-05-C-0382, and Project 863 through grant 2006AA01Z126.

**REFERENCES**