OUTLIER REMOVAL FOR IMPROVED SOURCE ESTIMATION IN ATMOSPHERIC INVERSE PROBLEMS

Marta Martinez-Camara, Martin Vetterli

School of Computer and Communication Sciences,
Ecole Polytechnique Fédérale de Lausanne (EPFL),
marta.martinez-camara@epfl.ch
martin.vetterli@epfl.ch

Andreas Stohl

Norwegian Institute for Air Research,
(NILU),
ast@nilu.no

ABSTRACT

Estimation of the quantities of harmful substances emitted into the atmosphere is one of the main challenges in modern environmental sciences. In most of the cases, this estimation requires solving a linear inverse problem. A key difficulty in evaluating the performance of any algorithm to solve this linear inverse problem is that the ground truth is typically unknown. In this paper we show that the noise encountered in this linear inverse problem is non-Gaussian. Next, we develop an algorithm to deal with the strong outliers present in the measurements. Finally, we test our approach on three different experiments: a simple synthetic experiment, a controlled real-world experiment, and real data from the Fukushima nuclear accident.

1. INTRODUCTION

Many of today's environmental problems are related to emissions of harmful substances into the atmosphere. The Fukushima nuclear accident, volcano eruptions, and greenhouse gases emissions are some examples.

To predict the consequences of these emissions is a difficult but crucial exercise. The main tool that we have for this task are the atmospheric dispersion models.

The use of these models requires knowledge of the source term: how much material was released, where and when. This information is generally unknown. Hence, source estimation is an important problem in its own right.

Various implementations of atmospheric dispersion models, such as Lagrangian dispersion models (LDM) [1], permit us to relate the source to the measurements in a linear way:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \qquad (1)$$

where $\mathbf{y}$ is the measurement vector, $\mathbf{x}$ is the source term, $\mathbf{A}$ is the model matrix and $\mathbf{e}$ is an additive error.

It is clear from (1) that estimating the source means solving a linear inverse problem, which is essentially a signal recovery problem.

Most environmental scientists use a least squares approach with Tikhonov (ℓ_2-norm) regularization, or variants of this method, to recover an estimate $\hat{\mathbf{x}}$ of the source:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 + \lambda\|\mathbf{x}\|_2, \qquad (2)$$

where $\lambda \geq 0$ is a regularization parameter.

For example, in [2, 3], Tikhonov regularization combined with a smooth second derivative constraint is used. In [4] Tikhonov regularization with a non-negative constraint is used. A slightly different approach is the use of a sparsity constraint together with a non-negative constraint as in [5].

All these approaches minimize the energy of the disagreement between the model and the observations, while at the same time keeping the energy of the solution in check. While this is a reasonable approach, no metrics of real performance are (or can be) given in most of these studies, simply because no knowledge of the ground truth is available. This fact makes it impossible to evaluate the true performance of any of these approaches.

1.1. Contributions

In this paper we use one of the rare cases where both the source and the measurements are available to show that the Gaussian assumption implicit in (2) does not hold. The noise is characterized by a heavy-tailed, highly non-Gaussian distribution.

We show that solving the problem without using the measurements associated with the largest errors improves the source reconstruction, compared to using all the measurements. Hence, we propose a novel signal processing method to detect such measurements blindly, tailored to this particular problem setting of environmental inverse problems.

Furthermore, we show the improvement brought by the new method in different setups: first on a toy problem using synthetic data, and second on the real-world measurements from the controlled experiment in [6]. In both cases we can observe clearly an improvement over previous methods. Finally, we apply the new method to the recovery of the source term of the Fukushima nuclear accident, and discuss the findings.

2. NOISE CHARACTERIZATION

2.1. System model

Modeling and predicting the behaviour of the atmosphere is a complicated task. Thousands of input parameters have to be taken into consideration, and the propagation of errors in the model is an unavoidable fact. On the other hand, measurements are collected with high quality instrumentation, but are still not perfect. These two different types of error - model and measurement errors - have very different characteristics. The former is a multiplicative error, $\mathbf{N}$, and the latter an additive one, $\mathbf{n}$. Taking both into account, the problem

can be reformulated as

$$\mathbf{y} = (\mathbf{A} + \mathbf{N})\mathbf{x} + \mathbf{n} = \mathbf{A}\mathbf{x} + \mathbf{e} \qquad (3)$$

where $\mathbf{e}$ is now the total error.

2.2. Controlled tracer experiment

Only a few controlled tracer experiment have ever been performed - the most important ones in Europe and in the US. They are exceptional opportunities to study model and measurement errors, as well as to test the various source recovery algorithms.

The ETEX experiment [6] was performed in France in 1994. A greenhouse gas was released in a controlled fashion, and detected by sensors placed in a radius of 1000 km around the source. Hence, in this case we have access to the measurements $\mathbf{y}$, the true source $\mathbf{x}$, and an estimation of the model $\mathbf{A}$. This permits us to study the errors $\mathbf{N}$, $\mathbf{n}$, and $\mathbf{e}$.

Let us model the components e_i of the vector $\mathbf{e}$ as random and iid. We can approximate their empirical probability distribution by plotting the histogram of the elements e_i.
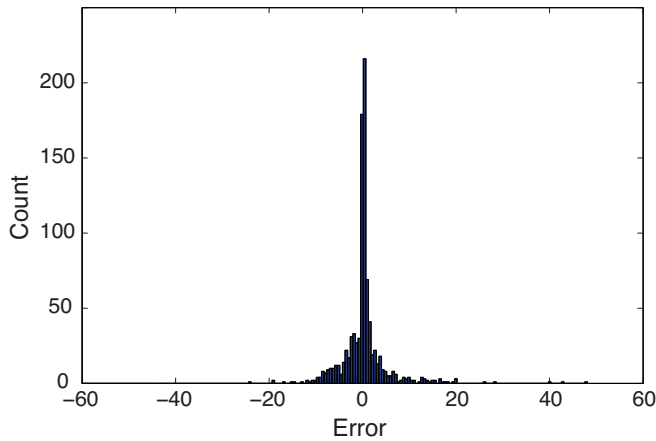

Fig. 1. Histogram of the total error $\mathbf{e}$. For clarity, the zero-error bin has been omitted here.

Fig. 1 shows graphically that the error has a heavy-tailed distribution. It can be appreciated in the histogram that the distribution clearly deviates from a Gaussian one. This is confirmed by calculating the excess kurtosis of the sample distribution. The value of $g = 123.64$ indicates that the underlying distribution is strongly super-Gaussian.

Using the ℓ_2 norm in the loss function in (2) is optimal when the residuals are Gaussian - which is not our case. Even worse, this loss function is very sensitive to outliers - just like the ones present in the heavy-tailed distribution shown in Fig. 1. Hence, the performance of (2) and its variants could be improved by additional processing, aimed at removing and/or marginalizing the outliers. In this paper we propose and demonstrate a novel scheme for this additional processing.

3. SOURCE ESTIMATION WITH COMPENSATION OF MODEL ERRORS

Imagine that we have an oracle which identifies to us the measurements corresponding to the largest errors (i.e. the outliers). If we remove these measurements, the performance of (2) in terms of the mean square error (MSE) improves significantly. In order to illustrate this, we have removed measurements associated with larger errors (sorted in magnitude) in the ETEX experiment. Fig. 2 shows how the MSE improves if we identify by hand these measurements and remove them from the estimation process.
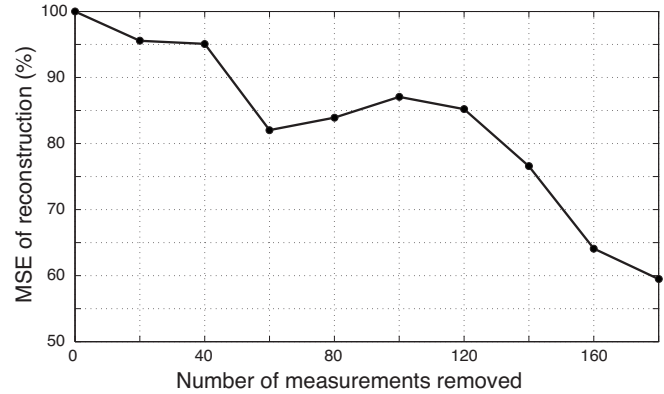

Fig. 2. The MSE decrease as we remove more outliers.

However, in a typical real-world problem, we do not have such an oracle. The question becomes: how could one locate the outliers blindly? Robust statistics techniques such as Huber or bi-squared loss functions [7] are used often to avoid the effect of outliers. Unfortunately, in our particular problem their performance is not robust enough and does not improve the estimation accuracy. Also other heuristic methods, such as RANSAC [8] are very popular, but their combinatorial nature and the large size of our measurement datasets makes them unusable in our problem.

A key observation we would like to make concerns the model of the system, i.e. the matrix $\mathbf{A}$. It is generated by environmental scientists using meteorological data. Obviously, creating the model using two different meteorological datasets will result in two slightly different matrices. The key idea of this paper is that we can use the variation between (two or more of) these matrices as an indication of the uncertainty in the modelling process, and thus the likelihood that a certain measurement is an outlier. Intuitively, if a particular entry in the model matrix $\mathbf{A}$ varies significantly between different models of the same process, it is likely to be estimated poorly. It will thus likely cause a large entry in its element in $\mathbf{N}$, which will increase the entry in the corresponding element in $\mathbf{e}$.

Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be two model matrices generated with different meteorological datasets. We can compute the matrix of absolute differences as

$$\mathbf{D} = |\mathbf{M}_1 - \mathbf{M}_2|. \qquad (4)$$

The accumulated differences in each row $\mathbf{d}$, reveal the likelihood of an outlier occurring in each measurement

$$\mathbf{d} = \mathbf{D}\mathbf{1}, \qquad (5)$$

where $\mathbf{1}$ is a column vector whose entries are equal to one.

Our approach is to extend the original approach shown in (2) by applying weights on the residuals, together with a non-negative constraint on the solution $\hat{\mathbf{x}}$. We give more weight to the most reliable samples, and vice versa. Based on (5) we propose to use different weighting schemes. We first consider exponential weights, such that

$$\mathbf{w}_e = e^{-\mathbf{d}}, \qquad (6)$$

where the exponentiation is an element-wise operation. An alternative weight function is to define thresholded weights as

$$\mathbf{w}_\beta = \begin{cases} \frac{1}{\mathbf{d}_i+1} & \text{if } \mathbf{d}_i \geq \beta \\ 1 & \text{if } \mathbf{d}_i < \beta \end{cases} \qquad (7)$$

where β is a predefined threshold.

Our problem can then be formulated as an outlier-weighted least squares with a non-negative Tikhonov regularization:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{W}(\mathbf{Ax} - \mathbf{y})\|_2 + \lambda\|\mathbf{x}\|_2 \ \text{ s.t. } \mathbf{x} \succeq 0, \qquad (8)$$

where $\mathbf{W}$ is a diagonal weighting matrix whose main diagonal is either $\mathbf{w}_e$ or $\mathbf{w}_\beta$, and $\succeq$ is the component-wise inequality.

4. RESULTS

To illustrate our outlier mitigation algorithm, we employ it in three different experiments: a simple synthetic example; a controlled example using real-world measurements and models, and a real-world experiment without knowledge of the ground truth.

4.1. Toy problem

In this section the measurements are generated using a model matrix $\mathbf{A} \in \mathbb{R}^{260 \times 10}$ with random uniform entries and a piece-wise constant source vector $\mathbf{x}$ (see, Fig. 3) as

$$\mathbf{y} = \mathbf{Ax}. \qquad (9)$$

We carry out several experiments; in each of them we build $\mathbf{M}_1$ and $\mathbf{M}_2$ by changing some of the entries of $\mathbf{A}$

$$\mathbf{M}_1 = \mathbf{A} + \mathbf{N}_1$$
$$\mathbf{M}_2 = \mathbf{A} + \mathbf{N}_2, \qquad (10)$$

where $\mathbf{N}_1$ and $\mathbf{N}_2$ are random sparse matrices. The goal is to recover $\mathbf{x}$ using $\mathbf{y}$ and the matrices with errors $\mathbf{M}_1$ and $\mathbf{M}_2$.

These matrices are well conditioned. Hence, the Tikhonov regularization term is not necessary. We recover $\mathbf{x}$ using traditional least squares, as well as weighted least squares with the weights defined in (6). We run (8) for different ratios of error entries in $\mathbf{M}_1$ and $\mathbf{M}_2$. For each ratio, we perform 1000 simulations.

Fig. 3 shows the MSE for each ratio. Clearly, the weights improve the overall performance and the best performance is achieved using the exponential weights.

4.2. Controlled, real-data experiment

In the second example we used the ETEX experiment data [6] to characterise the noise. We will use it to test the performance of our method on real dataset (measurements and system model).

In this experiment 340 kg of a greenhouse gas were released at a constant rate over a period of 12 hours. To build $\mathbf{x}$, time is discretized into 1-hour steps; in total we give $\mathbf{x}$ a window of 120 steps (5 days).

System model matrices are now generated using FLEXPART [1]. To generate both $\mathbf{M}_1$ and $\mathbf{M}_2$, we use as input to FLEXPART two different atmospheric analyses: one generated by the ERA 40 database [9], and the other generated by the ERA INTERIM database [10], both of them developed by the European Centre for Medium-Range Weather Forecasts.

Using $\mathbf{M}_1$ and $\mathbf{M}_2$, the matrix of absolute differences, $\mathbf{D}$, and the accumulated sum for each row, $\mathbf{d}$, are generated. They are shown
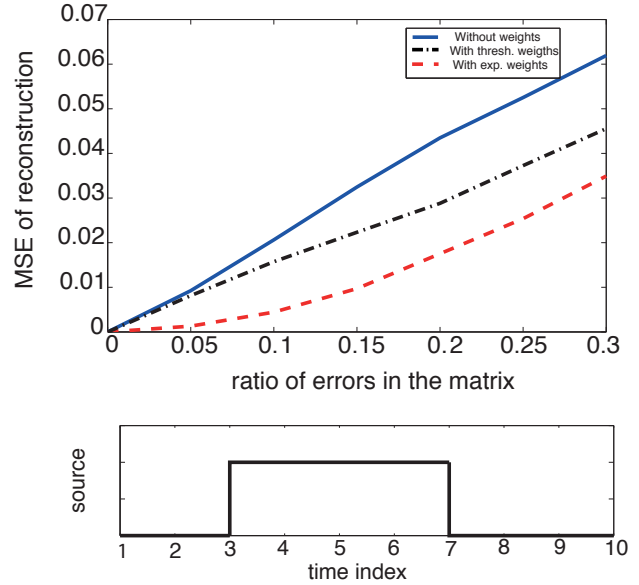

Fig. 3. In the upper part of the figure we see the MSE in the toy problem when the solution is recovered with and without weights. In the lower part we can see the source that we use in the toy problem.
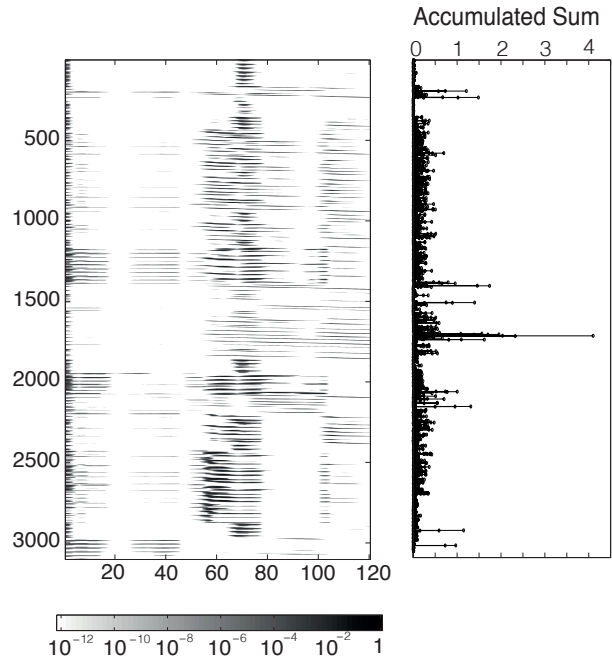

Fig. 4. Matrix of differences (left) and accumulated row sum (right) in the ETEX experiment.

in Fig. 4. We observe that the rows with larger accumulated sum of differences can be clearly identified.

Using $\mathbf{d}$ we build the weights as per (6) and (7). We manually adjust β to 3.5 times the average of $\mathbf{d}$. This time around, the condi-
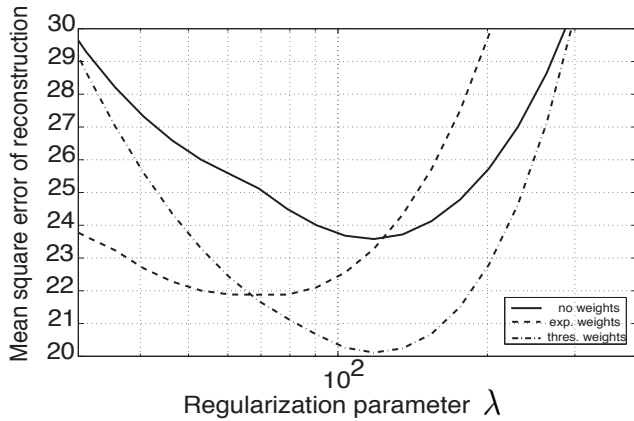
Fig. 5. Comparison of the performance of three source estimation algorithms: typical Tikhonov, and weighted Tikhonov with exponential and thresholded weights.

tion numbers of both $\mathbf{M}_1$ and $\mathbf{M}_2$ are on the order of 10^{17}. Hence, we need to include the Tikhonov regularization term to recover $\mathbf{x}$.

Fig. 5 shows the comparison of results using weighted and unweighted Tikhonov. As can be observed, also in this experiment, the proposed algorithm improves the recovery. In this case, the thresholded weights perform better than the exponential ones. However, the use of the thresholded weights requires the adjustment of the parameter β, which might be difficult to set in practice.

4.3. Fukushima

Finally, we apply our method to a real accident: the Xenon-133 emissions during the Fukushima Daiichi nuclear accident which took place in Japan in March 2011. The dataset that we use is the same as the one used in [5]. In this experiment, one system matrix has been generated using GFS data [11] and the other one using ECMRW data [12].

We use exponential weights, because as discussed earlier, it is not clear how to set β blindly. We compare the new reconstruction with the one in [5] in Fig. 6.

The new reconstruction indicates that more material was emitted on the first height level then previously thought. Also, it estimates fewer but larger emissions and in general, it delays all the emissions.

5. CONCLUSIONS

In this paper we treated linear inverse problems for the estimation of the source term using atmospheric dispersion models. We showed that the error in these models is highly non-Gaussian. To cope with this fact, we developed an algorithm to detect the outliers blindly and marginalize them using appropriate weights. We demonstrated the effectiveness of our approach on a simple synthetic example, as well as on real measurements from a controlled experiment. In both cases, our approach achieves a considerable improvement over existing techniques.
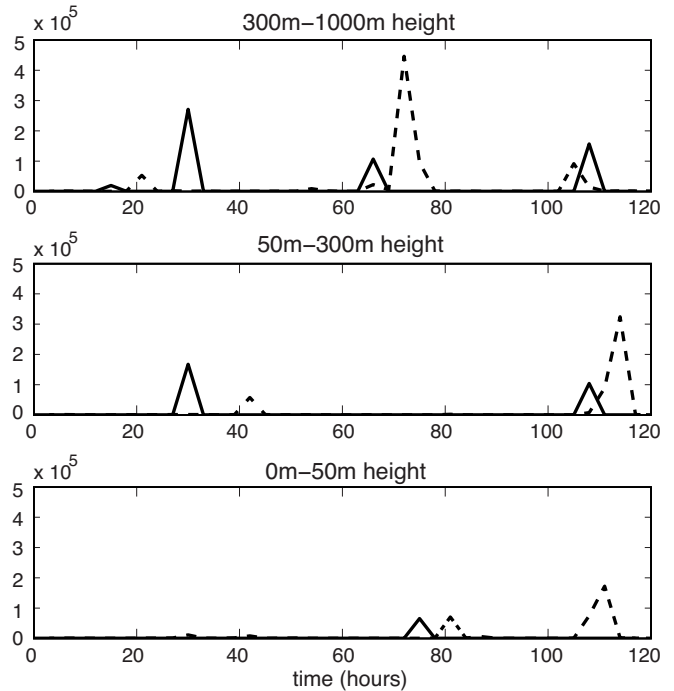
6. ACKNOWLEGDMENTS

Fig. 6. The dashed line represents the source recovered with the presented algorithm using exponential weights. The solid line represents the sources obtained in [5]. The different heights correspond to the different altitude of the emissions during the accident. The units of the emissions are GBq/s

7. REFERENCES

[1] A. Stohl, C. Foster, A. Frank, P. Seibert, and G. Wotawa, "Technical note: The lagrangian particle dispersion model flexpart version 6.2," *Atmos. Chem. Phys*, vol. 5, pp. 2461–2472, 2005.

[2] P. Seibert, A. Frank, and H. Kromp-Kolb, "Inverse modelling of atmospheric trace substances on the regional scale with lagrangian models," in *Proceedings from the EUROTRAC-2 Symphosium*, P. Midgley and M. Reuther, Eds., 2002.

[3] A. Stohl, P. Seibert, G. Wotawa, D. Arnold, J. Burkhart, S. Eckhardt, C. Tapia, A. Vargas, and T. J. Yasunari, "Xenon-133 and caesium-137 releases into the atmosphere from the Fukushima Dai-ichi nuclear power plant: determination of the source term , atmospheric dispersion, and depositon," *Atmos. Chem. Phys*, vol. 12, pp. 2313–2343, 2012.

[4] V. Winiarek, M. Bocquet, and O. S. A. Mathieu, "Estimations of errors in the inverse modeling of accidental release of atmospheric pollutant: Application to the reconstruction of the cesium-137 and iodine-131 source terms from the Fukshima Dai-ichi power plant," *Journal of Geophysical Research*, vol. 117, pp. D05122, 2012.

[5] M. Martinez-Camara, I. Dokmanic, J. Ranieri, R. Scheibler, M. Vetterli, and A. Stohl, "The Fukushima Inverse Problem," in *38th International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[6] K. Nodop, Ed., *ETEX Symposium on long-range atmospheric transport model verification and emergency response.*, Vienna (Austria), May 1997.

[7] P. J. Huber, *Robust Statistics*, John Willey, New York, 1981.

[8] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. ACM*, vol. 24, pp. 381–395, 1981.

[9] E. C. for Medium-Range Weather Forecasts, "Era 40," http://www.ecmwf.int/products/data/archive/descriptions/e4/.

[10] E. C. for Medium-Range Weather Forecasts, "Era interim," http://www.ecmwf.int/research/era/do/get/era-interim.

[11] N. Oceanic and A. Administration, "Global forecast system," http://www.emc.ncep.noaa.gov/index.php?branch=GFS.

[12] "European centre for medium-range weather forecasts," http://www.ecmwf.int.