Reference Manual for Package “sparse-msrf”

J. Ray and J. Lee
Sandia National Laboratories, Livermore, CA
{jairay,jlee3}@sandia.gov

1 Introduction

This is a manual for the package “sparse-msrf”, a Matlab package for estimating fossil-fuel CO_2 (ffCO2) emissions from sparse measurements of its concentration at a limited number of locations. This software was used in the study documented in [1, 2, 3]. It is suggested that you read either [1] or [2, 3] before delving into this software package.

Estimation of fields (defined on a grid), from sparse data, cannot be performed naively i.e. by assuming that each grid-cell is an independent parameter to be estimated - there is simply not enough observations to estimate all the cells. If the field is smooth, and adjoining grid-cells are constrained to assume field values which are not too different, one can use Gaussian Random Field [4] to model them and drastically reduce their dimensionality (making it feasible to estimate the lower dimensional model from limited data). When fields are not smooth, one can model them as a (linear) combination of proxies that contain similar patterns or one can use basis sets with compact support e.g., wavelets. A wavelet encoding/model of a field is not low-dimensional, unless one can “mask” (or eliminate) a large fraction of the wavelets using a guess/easily observed proxy of the field being modeled (which then results in a random field model). Since wavelets are a multi-resolution basis, one can called a masked collection of wavelets a Multiscale Random Field model (MsRF).

Given the paucity of data, even the model may not be sufficiently low dimensional. In such a case, one takes recourse to sparse reconstruction (shrinkage regression) which identifies which wavelet weights can be estimated from observations while turning off the rest. This is most efficient when the parameters being estimated (wavelet weights) are independent (alternatively, when the bases are orthogonal); since wavelets are an orthogonal basis set, this hold true for us.

The software contains the following capabilities:

- The ability to search through a number of wavelet families to determine the sparsity of encoding a given non-stationary field
- The functionality to create a MsRF model by masking specific wavelets in a wavelet encoding of a field. The field being used to create the masking could be an emission field or an easily-observed proxy (as was in our case).

- Functionality to construct various random sampling matrices and compare their sampling efficiency versus the atmospheric transport model.
- A sparse reconstruction functionality, based on StOMP [5], to estimate an emission field in an irregular (non-rectangular) domain while ensuring non-negativity (ffCO2 emissions cannot be negative).

The software package can be easily adapted to conduct network design or Observation System Simulation Experiments for ffCO2 estimation. It can be modified to also estimate emission patterns for gases which can be approximated as non-reactive tracers e.g. methane emissions from landfills, when the number, and distribution of landfills is ambiguous [6].

Citation: The proper way to cite this package is: "J. Ray, J. Lee and S. Lefantzi, "sparse-msrf: A package for sparse modeling and estimation of fossil-fuel CO2 emission fields", `http://www.sandia.gov/~jairay/kfcs.html`.

2 Description of the package

We envision that users will use this package to (1) experiment with wavelets to model non-stationary field (not just ffCO2 emission fields) and (2) use the software to perform source or emission field inversions for passive tracers. To that end, we also provide a realistic transport model.

The contents of the package are separated into different directories.

- `data/`: This directory contains miscellaneous data, required as input for the study in our paper.
  - `vulcan.8day.1/`: Folder containing ffCO2 emissions from the Vulcan inventory for US [7], aggregated to 1 degree and 8 day resolution. This is the "ground truth" emission field and is used to select wavelet family for the MsRF.
  - `nlights-0.1x0.1/`: A folder containing radiances of lights at night over North America. This is an easily-observed proxy for ffCO2 emissions and is used to choose which wavelets to mask.
  - `edgar2005.1x1/`: Emission field from the EDGAR inventory [8] which is used to provide a constant-in-time approximation of time-variant ffCO2 emission fields to be estimated.
  - `tower.obs`: Names of towers where ffCO2 concentrations are measured.
- `doc/`: Contains this manual and bibliography.

- `examples/`: This folder contains 5 examples that drive the modeling and algorithmic machinery of the package. 3 examples in (`examples/ex01/`) concern wavelet modeling of ffCO2 fields using a variety of wavelets, the sparsity of representation and generally exercise the machinery for representing a masked collection of wavelets. `examples/ex02` contains a large example that uses Haar wavelet and nightlight radiances to construct a MsRF, uses EDGAR to create a prior emission field and uses StOMP to perform the estimation. This code is quite complicated and ought to be learnt piecemeal. `examples/ex03/` contains code that compares the sampling efficiency of the transport model versus traditional random sampling matrices.

- `H-matrices/`: ffCO2 concentration at a space-time location i, $\mathbf{y}(x_i, t) = \mathbf{H}_i\mathbf{z}$ where $\mathbf{z}$ is the emission field and $\mathbf{H}$ is the linear operator that serves as the transport model over North America, with emission at 1 degree and 8-day resolution. This directory contains the $\mathbf{H}_i$ for 30+ observation towers.

- `incoherence/`: Folder with software machinery to compute random sampling matrices and their coherence with the wavelet bases constituting the MsRF.

- `msrf-selection/`: This folder contains the software machinery for representing a fields as a masked collection of wavelets of a number of families and orders. The masking field (in our case, a nightlight matrix) is an argument to the Matlab functions that perform the masking.

- `sparse-inversion/` : This folder contains the bulk of the code. It has functions that select the wavelets for masking using nightlights, use EDGAR to make the guess emission field, and perform estimation of the non-negative field (Steps I and II, as described in [3], are clearly marked). All the functions are documented, but it is expected that you might need help. Mail Jaideep Ray.

  Do not try to understand the code by looking through `msrf-selection/` and `sparse-inversion/`; rather start with the examples and work your way down.

3 Installation

The package depends on two open-source Matlab toolboxes, SparseLab [9] and WaveLab [10]. Install them first. They contain their own installation manuals.

Uncompress the `sparse-msrf.tgz` file

In each of the examples, you will have to set the path to the two toolboxes before you can run them. These specifications are at the top of the example Matlab scripts.

4 Description of examples

`examples/ex01/Fig02.m`: This example shows how different types of wavelets provide wavelet coefficients of different average values (and their variance) at different spatial scales, when encoding annually-averaged ffCO2 emissions from Vulcan. Run this once so that all useful function in the toolboxes are in your path.

Use this example to explore the encoding different wavelet families and different orders inside the same family can provide. If you are trying to model a different field e.g., methane emissions, chances are that Haar wavelets (which we found to be best) is the wrong choice. Type `help MakeONFilter` to determine the wavelet families and corresponding order that are implemented in WaveLab. All user-settable parameters are at the top of the Matlab script.

`examples/ex01/Fig23.m`: In any wavelet encoding, the bulk of the wavelets are at the finest scales; the ability to mask off a large fraction of them results in an enormous reduction of dimensionality (a *sparse* representation). This examples show how certain wavelet families do not manage to yield a sparse representation; wavelet weights at fine scale are all rather small, but not zero, and all about the same magnitude, making it impossible to construct as masking threshold.

Use this example to explore how the masking metric becomes absolutely impossible as one progresses to smoother wavelets with larger support, while the ffCO2 field has sharp discontinuities at the 1 degree resolution scale.

`examples/ex01/Fig24.m`: The more wavelets we remove, the more approximate the model and bigger the error when we try to model Vulcan emissions with them. This code allows you to see how model fidelity drops as one strives for a sparse model.

`examples/ex02/driver01.m`: This is a complicated example that, as distributed, performs an inversion over a 3-Week period (Week 31 to 33, where a Week is 8 days). One needs a buffer of 2 Weeks ahead of the period of interest. This is the place to start if you wish to delve into the workings of the modeling and inversion code. The example will take about 10 minutes to run.

Use this example to explore how to do a year-long inversion (see line 55). It'll take 4 hours and require 20 GB RAM; memory requirements increase with the duration over which you are estimating the emissions. Also check how the inversion worsens when we do a bad job of enforcing zero emissions outside US (reduce `info.nCompSamples` on line 49)

The example uses StOMP, as implemented in `SparseLab2.1-Core/Solvers/SolveStOMP.m` (in the SparseLab toolbox that you install). The StOMP solver has some hardcoded parameters. Change

```
{[x_I, flag] = lsqr(@lsqrMat,y,OptTol,20);}
```

on line 122 to

```
{[x_I, flag] = lsqr(@lsqrMat,y,OptTol,500);}
```

and recheck.

`examples/ex03/Fig21.m` contains an example where the incoherence of the **H** matrix, with respect to the MsRF, is compared to the incoherence achieved with standard random matrices (which are generated as part of the code). The example is distributed with incoherence estimation over a 2 Week period (to save on memory).

Explore how the incoherence changes throughout the year by testing for bigger durations (not just 2 Weeks) and at different points in the year (not just Week 21-23).

5 Disclaimer of warranty

We make no warranties, explicit or implicit, that the programs contained in this collection are free of error, or are consistent with any particular standard of accuracy, or that they will meet your requirements for any particular application. They should not be relied on for any purpose where incorrect results could result in loss of property or personal injury. If you do use programs for any such purpose, it is at your own risk. The authors disclaim all liability of any kind, either direct or consequential, resulting from your use of these programs.

Acknowledgments

This work was funded by the Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

[1] J. Ray, J. Lee, S. Lefantzi, V. Yadav, A. M. Michalak, B.van Bloemen-Waanders, and S. A. McKenna. A multiresolution spatial parametrization for the estimation of fossil-fuel carbon dioxide emissions via atmospheric inversions. SAND Report SAND2013-2919, Sandia National Laboratories, Livermore, CA 94551-0969, April 2013. Unclassified and unlimited release.

[2] J. Ray, V. Yadav, A. M. Michalak, B. van Bloemen Waander, and S. A. McKenna. A multiresolution spatial parameterization for the estimation of fossil-fuel carbon dioxide emissions via atmospheric inversions. *Geoscientific Model Development Discussions*, 7:1277–1315, 2014. Accepted for publication in Geoscientific Model Development.

[3] J. Ray, J. Lee, V. yadav, S. Lefantzi, A. M. Michalak, and B. van BloemenWaanders. A sparse reconstruction method for the estimation of multiresolution emission fields via atmospheric inversion. *Geoscientific Model Development Discussions*, XX:YY–ZZ, 2014.

[4] David Higdon. A primer on space-time modeling from a Bayesian perspective. In *Statistical methods for spatio-temporal systems*, Monographs on Statistics and Applied Probability. Chapman&Hall / CRC Press, 2006.

[5] D. L. Donoho, Y. Tsaig, I. Drori, and J-L Starck. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. *IEEE Transaction on Information Theory*, (2):1094–1121, 2012.

[6] B. Hirst, P. Jonathan, F. Gonzales del Cueto, D. Randell, and O. Kosut. Locating and quantifying gas emission sources using remotely obtained concentration data. *Atmospheric Environment*, pages 141–158, 2013.

[7] Project Vulcan webpage. `http://vulcan.project.asu.edu/plots.php`.

[8] European Commission, Joint Research Centre (JRC)/Netherlands Environmental Assessment Agency (PBL). Emission Database for Global Atmospheric Research (EDGAR), release version 4.1. `http://edgar.jrc.ec.europa.eu/index.php`. EDGAR/EUROPA Web-site.

[9] Sparselab Webpage. `http://sparselab.stanford.edu/`.

[10] Wavelab Webpage. `http://statweb.stanford.edu/~wavelab/`.