**Reduced Cost Construction of Jacobian Matrices for High-Resolution Inverse Modeling: An Application to Optimizing North American Methane Sources from TROPOMI Satellite Data**

**Authors**

Hannah Nesser1, Daniel J. Jacob1, Joannes D. Maasakkers2, Melissa P. Sulprizio1, Yuzhong Zhang1, Tia Scarpelli1

**Affiliations**

**1** Harvard University, Cambridge, Massachusetts, USA.

2 SRON Netherlands Institute for Space Research, Utrecht, the Netherlands.

**Abstract**

Global high-resolution observations of atmospheric trace gas concentrations from satellites can greatly improve our understanding of surface emissions through inverse analyses. For example, the new TROPOspheric Monitoring Instrument (TROPOMI) retrieves daily global observations of atmospheric methane concentrations at 7x7 km2 pixel resolution. Variational inverse methods can optimize surface emissions globally at this resolution but do not readily provide error characterization, including information content, for the posterior solution. In fact, the information content of the satellite data may be considerably lower than the data density would suggest because of limited retrieval success rate, instrument noise, and error correlations that propagate through the inversion. This could lead to smoothing errors in variational methods. An analytic inverse solution provides closed-form characterization of the posterior error statistics and information content but requires the construction of the Jacobian matrix relating emissions to atmospheric concentrations. Building the Jacobian matrix is computationally expensive at fine resolution because it involves perturbing each emission element, typically individual grid cells, in the atmospheric transport model. We propose a method to greatly decrease the computational cost of analytic inversions by constructing the Jacobian matrix using only the emission elements with sufficient information content from the observations. Starting from an initial estimate of the Jacobian matrix that assumes simple transport, we iteratively apply perturbations to the leading patterns of information content rather than to the individual model grid cells. The resulting matrix optimizes emissions only in areas spanned by these leading patterns. We demonstrate the method in an analytic Bayesian inversion of TROPOMI data over North America in July 2018. We confirm that the estimated Jacobian matrix produces posterior emission estimates and error covariances that are similar to an inversion conducted with the Jacobian matrix for the original model grid. Our method enables computationally efficient, high-resolution analytic inversions of high-density satellite data.

**Introduction**

Satellite retrievals of atmospheric trace gases can be used in inverse models to infer the emissions that best explain the observed concentrations. The standard inverse approach uses an adjoint model to optimize emissions given the observations. The computational cost of such variational approaches is independent of the resolution at which emissions are optimized and the approach can be used in both linear and nonlinear systems. However, the solution provides incomplete characterization of errors and information content. As the observing capacity of satellites increases, inversions will constrain surface emissions at increasing resolution. Characterizing the information content of the inverse solution is necessary to prevent over-interpretation of high-resolution inverse results. In linear systems with normally distributed errors, as in the problem of atmospheric methane, an analytic solution exists that characterizes both the error and information content of the optimized emissions. However, the computational cost of analytic inversions is limited by the resolution at which surface emissions are constrained. In this paper, we decrease the dominant computational cost in an analytic inversion, the characterization of the linear relationship between emissions and observations, given by the Jacobian matrix. We iteratively update a low-cost initial estimate of the Jacobian matrix by applying a finite difference scheme to the dominant patterns of information content in the system. The resulting low-rank approximation of the Jacobian is demonstrated in a high-resolution (0.5º x 0.625 º) inversion of one month of methane retrievals from the Tropospheric Monitoring Instrument (TROPOMI) over the North American domain.

The Tropospheric Monitoring Instrument (TROPOMI) aboard the Sentinel-5 precursor launched in October 2017 now provides daily, global retrievals of total column methane concentrations at 7 x 7 km2 nadir pixel resolution. Even at a ~3% retrieval rate, TROPOMI produces orders of magnitude more observations at higher density than the other satellite providing global retrievals of column methane concentrations, GOSAT. For example, Maasakkers et al. (2019) conducted a global inversion of ~1.2 x 106 GOSAT retrievals from 2010 to 2015. TROPOMI returned almost 6.4 x 106 successful retrievals in August 2018, five times as many retrievals as GOSAT provided over six years. TROPOMI also provides denser coverage in areas where retrievals are successful. TROPOMI is one of many existing and planned earth-observing satellites that retrieve concentrations of methane and other atmospheric trace gases at high spatial and temporal resolution. This significant increase in current and planned observing capacity supports an increase in the resolution at which inversions optimize emissions. However, satellites do not observe all grid cells equally. TROPOMI’s ~3% retrieval rate reflects its inability to observe in areas with high cloud coverage, high aerosol loading, and high albedo. As a result, the information content of the observations may not support a dramatic increase in resolution everywhere. It is therefore of interest to characterize the information content of the inverse solution along with the optimal emissions, requiring analytic solution of the inversion.

Inverse models describe the dependence of emissions on atmospheric concentrations, inverting a forward model that simulates atmospheric concentrations given input emissions fields. In the absence of observational or model error, the inverse solution is simple: for *m* observations **y** and a forward model **F**, the associated emissions **x** for *n* grid boxes are **x** = **F**-1(**y**). Significant errors in both the observations and model require that the inverse solution be a statistical optimization of the emissions given the observations. Errors also produce a linearly dependent system of equations: *m* observations can constrain emissions from *n* < *m* grid boxes. Even when *m* >> *n*, as is the case for most inversions of satellite retrievals, the observations are often unable to constrain *n* unique pieces of information. Most inversions of satellite retrievals are therefore under-constrained. A prior **xA** is introduced to regularize the solution. In a Bayesian inversion, errors in the model and observations and in the prior are assumed to be normally distributed and are summarized by the observational and prior error covariance matrices **SO** and **SA**, respectively. Bayes’ theorem allows the explicit formation of a cost function **J**(**x**) that, when minimized over all **x**, maximizes the probability of the emissions given the observations:

In a linear system, the forward model can be written as **y** = **Kx** + **c**, where **K** is the Jacobian matrix that describes the sensitivity of observations to emissions, i.e. **K** = d**y**/d**x**. In this case, the cost function can be minimized in one of two ways. An adjoint model can be used to iteratively update an initial estimate for the emissions (i.e. the prior) until convergence. Or, the emissions that minimize the cost function can be found by analytic solution of . The analytic solution yields the posterior emissions as well as the associated error and the information content of posterior solution, given by the averaging kernel **A**.

Variational approaches to minimizing the cost function can be applied to linear and non-linear systems alike, and the computational cost is not limited by the dimension of the emissions vector **x**, i.e. by the number of grid boxes. However, variational approaches have a number of disadvantages. First, in high-dimensional systems, the cost function **J** is often shallow, and the variational approach may converge before the true minimum is reached. Second, variational approaches do not characterize the error or information content of the posterior solution. While ensemble approaches can approximate error, these estimates are only as good as the number of ensemble members. Third, each inverse solution requires an independent application of the adjoint model. The additional computational cost of sensitivity tests therefore mitigates the initial computational benefit. Finally, variational approaches require the continued development of adjoint models, which often lag behind state-of-the-science forward models. However, efficient automatic differentiation may reduce the lag.

Analytic solutions to inversions find the true minimum of the cost function, fully characterize the error and information content of the posterior solution, support numerous sensitivity tests at only small additional computational cost, and do not require use of the adjoint model. However, analytic solutions exist only in linear or approximately linear systems. And, the computational cost is limited by the resolution of the inversion. The

However, the computational cost of an analytic inversion increases with the resolution at which emissions are constrained. In the global inversion of GOSAT observations, Maasakkers et al. (2019) optimized emissions on a 4º x 5º grid, constraining ~1,000 terrestrial grid cells. Increasing resolution to 0.5º x 0.625º would require constraining ~64,0000 terrestrial grid cells. The increase in cost is attributable to the cost of characterizing the linear relationship between observations and emissions, due both to (1) the number of model runs and (2) the increase in model resolution.

1. Paragraph 4: Solving inversions (analytic)
   1. Analytic approaches: in the case that the forward model is linear (i.e. a linear relationship between emissions and the modeled observations), analytic solution to the cost function exists: posterior mean, posterior error, and information content of the solution.
      1. Disadvantages:
         1. Computational cost is limited by the cost of constructing the Jacobian, the matrix that represents the linear relationship between emissions and observations. The Jacobian is constructed via a finite differencing scheme: the dependence of observations on emissions is found by perturbing each of the grid cells for which methane emissions are constrained.
         2. As resolution increases (1) number of model runs needed increases, (2) computational cost of each model run increases.
2. Paragraph 5: Reducing computational cost of high resolution inversions
   1. Dimension reduction via grid cell aggregation (Turner, Bocquet)
   2. Reduced rank approach (Bousserez and Henze)
      1. This approach relied on the adjoint of the model and random matrix methods to construct the Jacobian
      2. For a rank n matrix, they ran the forward model n times and the adjoint n times
         1. The adjoint being computationally expensive (check IGC9 notes here--approximately 3x cost of one forward model runs), this is roughly equivalent to 4\*n model runs for a rank n Jacobian.
      3. Other approaches?
3. Paragraph 6: Proposal
   1. Here, we attempt to develop a method for constructing a Jacobian at reduced computational cost that avoids the use of the adjoint, thereby (a) reducing model development costs and adjoint lag problems and (b) reducing the total computational cost of constructing the Jacobian.
   2. We use the reduced-rank approach developed by Bousserez and Henze to propose a reduced rank Jacobian construction scheme: Rather than perturbing individual grid cells, we perturb instead the dominant patterns of information content. Because information content is a function of the forward model, this approach is iterative. We propose an iteration scheme and convergence criteria.
   3. We then demonstrate the method in the context of an inversion of TROPOMI observations over North America for July 2018 (?)
4. Methods
   1. A Bayesian inversion solves for the most likely set of emissions given a set of observations, accounting for often significant errors in the initial estimate and observations. If the forward model is assumed linear, there is an analytic solution for posterior mean, posterior mean, and information content.
   2. The solution of an analytic Bayesian inversion can be framed so that either the number of observations or number of constrained grid cells limits the computational cost; in systems with large numbers of observations, such as those given by satellites, the computational cost is limited by the number of grid cells.
   3. However, not all grid cells contain equal information: inversions are able to constrain emissions in some grid cells better than others because of prior information or observations. Notably, the information content, given by the averaging kernel, represents the sensitivity of the posterior to the truth and describes the ability of the inversion to constrain emissions in a cell. The sum of the diagonal elements gives the degrees of signals for signal (DOFS), the number of pieces of information the inversion can independently constrain. Most atmospheric inverse systems have DOFS << grid cells n (e.g. Bram's global inversion, Bram's NA inversion, Yuzhong's Permian inversion (global, regional, local examples)).
   4. Previous approaches to reducing the computational cost of these inverse solutions took advantage of the discrepancy between DOFS and the number of state vector elements and focused on reducing the dimension of the state vector. Bocquet, Turner. Disadvantages summary.
   5. Figure 1 shows the aggregation scheme developed by Bousserez and Henze. This scheme maximizes the information content of the reduced system. Specifically, they solve the inversion in the most informed subspace of the original state space. To do so, they use the most predominant patterns of information content, given by an eigendecomposition the averaging kernel, to project the original state space to a reduced dimension space. The inversion can then be solved in the reduced dimension space. However, the elements of the state vector no longer correspond to grid cells but to eigenvectors--and are therefore not easily interpreted. Bousserez and Henze define and optimal projection operator that restores the original dimension and minimizes information loss. The reduction and prolongation operators may be applied in sequence to form a projection pi--this corresponds to a projection from the original state space to a reduced rank space in which the inversion is solved. This reduced rank solution is used to find an approximation of the full rank solution.
   6. The reduced rank approach reduces the computational cost of inverting a large n x n matrix. To reduce the computational cost of constructing the Jacobian, B&H use random matrix methods to construct a rank p Jacobian. However, this approach requires p forward model runs and p adjoint runs, amounting to 4p forward model equivalents in terms of computational cost. It also relies on the adjoint, requiring continued maintenance of the adjoint.
   7. To reduce the computational cost of constructing the Jacobian, we propose perturbing the patterns of information content rather than individual grid cells.
   8. Because the information content is a function of the forward model, it is necessary to use an initial estimate of the Jacobian to find an initial set of eigenvectors, and then update the Jacobian iteratively. Specifically, we propose the following scheme:
      1. Initialize the Jacobian
      2. Perturb p eigenvectors and retrieve the model response
      3. Estimate the updated full rank Jacobian
      4. Return to (ii) and iterate until convergence [update scheme, convergence criteria]
   9. We will discuss each in turn.
      1. Jacobian initialization
         1. Mass balance approach. This produces a sparse matrix, which is suboptimal for eigendecomposition. We schmear the emissions (initially: equally around. Ultimately: using the wind vectors from GC)
      2. Perturb p eigenvectors and retrieve the model response
         1. [insert perturbation definition]
         2. This will produce a Jacobian for a dimension p space; it is reprojected using the optimal prolongation operator defined by B&H.
         3. [insert reprojection]
      3. Estimate the updated full rank Jacobian
         1. The resulting matrix is a reduced rank Jacobian. Two options:
            1. Solve for the inversion in this reduced rank space, producing the full rank approximation described in B&H. However, this requires that the initial eigenvectors to be sufficiently accurate representations of the dominant patterns of information content. This may not be the case.
            2. Estimate the full rank Jacobian. Because the projection is dimension nxn and rank p, it is not invertible. Instead, we estimate the full rank Jacobian by recognizing that something is concave (i.e. a minimum exists) and differentiating and solving for 0.
            3. This yields unstable results where the information content is low; we require that information content is sufficiently high in order to use this approach to solve for values. Otherwise, we set Jacobian values equal to 0. Functionally, this means that we optimize the inversion in areas where information content is sufficiently high and default to the prior in areas where there is insufficient information content.
      4. Iterate until convergence
         1. Iteration scheme
         2. Convergence criteria
            1. Converges to true solution as p --> n
            2. Convergence criteria
            3. Quantifying error?
5. Results
   1. Inverse system:
      1. Observations: TROPOMI observations, July 2018
      2. Observational error: residual error method applied to TROPOMI?
      3. Forward model: North American domain, GC 12.1.0, 1x1.25
      4. Prior: same as Maasakkers 2016
      5. Prior error: same
   2. Construct a true Jacobian at 1x1.25 degrees
   3. Estimate the Jacobian using mass balance approach (specifics)
   4. Iteration scheme
   5. Results
      1. Jacobian comparison
      2. Posterior mean comparison
      3. Posterior error comparison
      4. Error quantification?
6. Conclusions