Massive analysis of multidimensional astrophysical data by inverse regression of physical models

Sylvain DOUTÉ[1] Florence FORBES[2] Stanislaw BORKOWSKI[2] Samuel HEIDMANN[2] Luc MEYER[2]

[1]CNRS, IPAG, University of Grenoble Alpes, 38000 Grenoble, France

[2]Inria, CNRS, Grenoble INP, LJK, University of Grenoble Alpes, 38000 Grenoble, France

Résumé – Les sciences modernes de l'observation telles que la géophysique, l'astrophysique, l'imagerie médicale, etc. produisent un volume énorme de données de haute dimension. Une approche puissante pour analyser ces données et récupérer des informations d'intérêt utilise le formalisme bayésien pour inverser des modèles physiques. Dans cet article, nous montrons l'application d'une méthode basée sur une approche statistique de régression inverse - GLLiM - qui a l'avantage de produire des distributions approximant les lois à posteriori cibles. Ces distributions peuvent également être utilisées pour des prédictions plus fines à l'aide d'un échantillonnage d'importance tout en fournissant un moyen de mieux explorer le problème inverse lorsque plusieurs solutions équivalentes existent et d'effectuer une estimation du niveau d'incertitude. Dans cet article, notre objectif est de présenter une application de GLLiM à l'analyse d'une séquence d'images hyperspectrales acquises depuis l'espace pour la même scène martienne et de présenter le logiciel PlanetGLLiM.

Abstract – Modern observational sciences such as geophysics, astrophysics, medical imaging, etc. produce huge volumes of high-dimensional data. One powerful approach to analyse such data and retrieve information of interest is Bayesian formalism to inverse physical models on the data. In this paper, we show the application of a method based on a statistical inverse regression - GLLiM- that has the advantage to produce full probability distributions as approximations of the target posterior distributions. These distributions can also be used for further refined predictions using importance sampling while also providing a way to better explore the inverse problem when multiple equivalent solutions exist and to carry out uncertainty level estimation. In this paper, we present an application of GLLiM for the analysis of a sequence of hyperspectral images acquired from space for the same Martian scene and present the PlanetGLLiM software.

1 Introduction

Sensors aboard observation platforms in orbit around the Earth and other planets of the Solar System return huge volumes of data. The latter result from measurements that are high-dimensional covering space, wavelengths, time, angles, etc. in different spectral domains and different regimes (reflection, emission, active sensing). Retrieving the information of interest from such data consists of inverting a *direct or forward* model, which theoretically describes how parameters of interest $x \in X$ are translated into observations $y \in Y$. In remote sensing, the observations y are high-dimensional (of dimension D) because they represent signals in time, angles or wavelengths. Besides, many such high-dimensional observations are available and the application requires a very large number of inversions (denoted by N_{obs} in what follows). The parameters x to be predicted (of dimension L) is itself multi-dimensional with correlated dimensions. In [5] a learning technique is put forward with a Bayesian framework capable of solving the problem of inverting physical models on multi-angular data in order to estimate the value of their parameters. The method addresses: 1) the large number of observations to be analysed, 2) their high dimension, 3) the need to provide predictions for several correlated parameters, 4) the possible existence of multiple solutions and 5) the requirement to provide the latter with a confidence measure (e.g. uncertainty quantification). Here, we present a planetary application of the statistical inversion method implemented as a high-performance, documented, and open-source software **PlanetGLLiM**[i] across multiple platforms as docker images. We analyse series of hyperspectral images acquired in the visible and infrared at different angles over regions of interest at the surface of Mars. The dimension D is greater than L, with L typically smaller than ten and D up to a few hundreds, while the number of observations to be inverted N_{obs} can be of the order of a few millions.

2 Hyperspectral images of Mars

In planetary science, information on the microtexture of surface materials such as grain size, shape, roughness and internal structure can be used as tracers of geological processes. This information is accessible by remote sensing under certain conditions thanks to multi-angular optical observations. Around Mars, the CRISM instrument [6] acquires sequences of hyperspectral images in the visible and infrared from eleven different angles when the Mars Reconnaissance Orbiter flies over a site. Such observations are available for hundreds of Martian sites of interest. In the seminal work of [3], the characterisation of Martian materials by orbital spectrophotometry is conducted at one wavelength only (750 nm) thanks to the MARS-ReCO tool [1]. MARS-ReCO actually extracts a parametric model of surface reflectance (BRDF RTLS) by atmospheric correction for 544 visible and near-infrared wavelengths and for a network of several thousand points distributed over the scene. The interpretation of the BRDF (Bidirectional Reflectance

i. documentation available here

Distribution Function) extracted by "MARS-ReCO" in terms of composition and microtexture is based on the inversion of physical models of radiative transfer (Hapke and Shkuratov models) linking in a nonlinear way physical and observable parameters (functional $y = F(x)$). In this case $y = y_{obs}$ is a vector of D reflectance values for $D = 11$ geometries, x is vector of $L = 4$ or $L = 6$ photometric parameters. The great number of observations N_{obs}, organized in data cubes, results from the combination of spectral and spatial sampling of the scene and is of the order of 10^6.

3 Inversion pipeline

From an experimental vector y_{obs} of reflectance values, the objective is to estimate the mean or the most probable value(s) for each of the L parameters (the components of vector x) of the physical model F and to provide a measurement of uncertainty about the estimations. The **PlanetGLLiM** pipeline performs the estimation in four steps:

1. The generation of a database of N couples $\mathcal{D}_N = \{(x_n, y_n), n = 1 : N\}$ using the direct physical model, $y = F(x)$. More specifically, x_n values are simulated from a chosen prior distribution, e.g. uniform over the parameter range, and the direct model F is applied to produce $F(x_n)$ to which a typically Gaussian noise is added to provide y_n. F is supposed to be available as a closed-form expression (see section 4). Note that such a database can also come from real observations of (x, y) couples if available.

2. The learning phase in which the pipeline constructs from the database a direct and an inverse parametric statistical model of the functional $y = F(x)$. The learned model is a Gaussian mixture model with a structured parameterization referred to as GLLiM for Gaussian Locally Linear Mapping (see [2, 5] for details). Because it is trained on the data from step 1, it depends on the physical direct model. More specifically, GLLiM depends on a set of parameters $\theta = \{\pi_k, c_k, \Gamma_k, A_k, b_k, \Sigma_k, k = 1 : K\}$ which can be estimated with a standard Expectation Maximisation algorithm. It then provides approximations of both conditional distributions $p(x|y)$ and $p(y|x)$ as K-component Gaussian mixtures. In particular, in a Bayesian setting, the inverse model is then approximated by the following surrogate probability distribution function (PDF) expression p_{GLLiM} which is learned once, for all possible y_{obs} to be inverted,

$$p(x|y) \approx p_{GLLiM}(x|y) = \sum_{k=1}^{K} w_k(y)\mathcal{N}(x; A_k y + b_k, \Sigma_k)$$

with $w_k(y) \propto \pi_k \mathcal{N}(y; c_k, \Gamma_k)$ and where K is estimated from the database using the Bayesian information criterion (BIC).

3. This surrogate model is then used for all vectors y_{obs} (corresponding to the spectral and possibly spatial dimensions of the problem) in order to build a set of a PDFs $p_{GLLiM}(x|y_{obs})$.

4. The PDFs are then exploited each independently by different techniques to estimate the solution $\hat{x}$ corresponding to each vector y_{obs}: estimation by the mean/mode of the PDF, fusion of the components of the Gaussian mixture model into a small number of centroids (usually two or three) to identify possible multiple modes (we consider up to 2 modes in the following), importance sampling of the true target PDF $p(x|y_{obs})$ with the proposal distribution set to $p_{GLLiM}(x|y_{obs})$ around the mean or around the centroids/modes, [5].

More specifically, for each vector y_{obs}, there are six types of estimations, provided by the GLLiM approximated distribution:

- *pred_mean*: prediction by the surrogate PDF mean.
- *pred_center_1*: prediction of the first PDF mode.
- *pred_center_2*: prediction of the second PDF mode.
- *is_mean*: importance sampling estimation of the true PDF mean using the surrogate PDF as importance proposal.
- *is_center_1*: importance sampling estimation of the first true mode.
- *is_center_2*: importance sampling estimation of the second true mode.

The analysis of our data cubes leads to different estimations of the L physical parameters and corresponding uncertainties for thousands of spatial points and hundreds of wavelengths that are treated independently in each dimension. A joint processing along the spatial and spectral axes can be envisioned to take advantage of the continuity on the spatial and/or spectral evolution of the parameters to be inferred. For example, the solution obtained by GLLiM at a given wavelength (modes and associated covariance matrices) could be used to calculate a prior for the next wavelength according to a prognostic model operator. There may be other ways to account for such structured information but this is left for future investigations.

4 The Hapke radiative transfer model

The Hapke model is a radiative transfer model introduced by B. Hapke in [4]. This model makes it possible to explain a reflectance measurement with relatively few parameters like the absorptivity of the particles, their scattering cross sections, their phase function and the macroscopic roughness of the material. The model proposed in [4] links the physical parameters $(w, b, c, \theta_H, B_0, h)$ (see Table 1 for details) to the bidirectional reflectance by the following formula:

$$R(i, e, G) = \frac{w}{4\pi} \frac{\mu_{0eG}(\theta_H)}{\mu_{0eG}(\theta_H) + \mu_{eG}(\theta_H)} [P_G(\alpha, b, c)(1 + B_G(B_0, h) + M(\mu_{0eG}(\theta_H), \mu_{eG}(\theta_H))] \times S_G(i, e, \alpha, \theta_H) \quad (1)$$

G encodes the geometric measurement configuration as a set of 3 angles characterizing the planetary surface illumination and observation: the incident i and emerging e angles (μ_0 and μ_e their cosines), and phase α the angle between the scattering and incidence directions. The functions P_G, M, B_G, and S_G express respectively the single and multiple scattering of light within the granular medium, the opposition effect at small phase angles, and the anisotropy of the reflectance due to roughness.

5 The PlanetGLLiM software

We applied the GLLiM method to inverse the Hapke physical model on the CRISM data using the **PlanetGLLiM software**. It is an application specifically tailored to handle inversions of reflectance models given cubes of multi-spectral

Symbol	Parameter	Physical meaning	Range
ω	Single scattering albedo	Scattering at the grain level	[0 , 1]
b	Anisotropy parameter	b=0 isotropic scattering b=1 directional scattering	[0 , 1]
c	Backward or forward scattering coefficient	c > 0.5 backscattering c < 0.5 forward scattering	[0 , 1]
θ_H	Photometric roughness	Mean slope angle averaged on all scales	[0° , 90°]
B_0	Amplitude of the shadow hiding opposition effect	Opacity of the grains	[0 , 1]
h	Width of the shadow hiding opposition effect	Porosity of the granular medium	[0 , 1]

Table 1 – Description of the Hapke parameters

data and geometries typically produced by planetary remote sensing. PlanetGLLiM offers an easy to use graphical user interface and implements the Hapke and the Shkuratov models. Custom models can be added by the user in the form of a single Python class. The application is built around a computationally efficient kernel, implemented in C++, that can handle situations where the signals to be inverted present a moderately high number of dimensions and are in large numbers. The kernel, called Kernelo, can be used as a C++ library or a Python module, and as such it can be used to inverse other problems having direct models but unrelated to reflectance or remote sensing. Both PlanetGLLiM [ii] and Kernelo [iii] are open source and freely available under the CeCILL license.

6 Data analysis: an illustration

In [5] a validation of the GLLiM method is performed by inverting the 6 parameter Hapke model on controlled synthetic data with GLLiM and two MCMC schemes. This allows a quantitative comparison of the different predictions based on the prediction errors with the reference x_{obs} and on the reconstruction errors with the synthetic y_{obs} vectors. In this section, we further illustrate the capabilities of the GLLiM method by applying it to the analysis of a CRISM multi-angular sequence of hyperspectral images (sec. 2) identified by FRT0000B385 to recover surface granularity.

The target is a semi-circular depression in the Eos Chasma of Valles Marineris (Fig. 1). Of special interest is the mineral hematite (a mineral specie composed of iron(III) oxide with the formula Fe_2O_3) that is concentrated on the adjacent raised plateau and the depressed terrains that surround it. Also outcrops of high albedo terrain rich in sulfate-bearing material can be distinguished. Note that hematite on Mars is widespread in dust in nanophase form, with crystals only tens of nanometers in size. There are also fine-grained (red) and coarse-grained (gray) hematite that exhibit absorption features around 0.5 and 0.9 μm. The latter type of hematite is the byproduct of the interaction of an acidic fluid with basaltic rocks that occurred in aquifers hundreds of millions years ago below the Martian surface. This byproduct is usually found in sulfate-rich sedimentary materials like those present in Eos Chasma.

Our objective is to get information on the microtexture of the surface materials of the scene targeted by FRTB385 such as grain size, shape, roughness and internal structure. For this observation MARS-ReCO extracts a parametric model of surface bidirectional reflectance for 1648 pixels distributed across the scene and for 344 wavelengths between 0.44 and 2.60 microns (i.e. $N_{obs} = 566912$). The kmeans clustering of the photometric curves at a wavelength of 755 nm, defined as the series of reflectance values calculated at this wavelength with the model for the 11 geometries of the CRISM observation, leads to 5 classes of terrains (Figure 1). The latter are meaningful since they are spatially correlated to geological structures. The Hapke model (1) is massively inverted on the full surface dataset according to the method described in Section 3. We note that for a large majority of y_{obs} the reconstruction error obtained with the posterior mean *is_mean* is lower than the error with the centroids *is_center_1* or *is_center_2*. That means that the posterior pdf is unimodal with a quite large lobe. Then, we calculate from the individual results, the mean spectrum of each Hapke parameter for the different terrain classes. The error bars are estimated on the basis of the root mean square of the pixel-wise uncertainties. We note an excellent continuity in the spectral domain even though the analysis is performed independently for each wavelength, which suggests the physical validity of the solutions. As an example, the single scattering and photometric roughness mean spectra are given in Figure 2 for the hematite rich terrains (blue class). The former parameter, which is directly related to optical properties of the constitutive materials, shows very clearly the distinctive shape and absorption band of grey hematite at 1 μm. The photometric roughness is intermediate between that of dust and that of basalt sand. The micro-texture can be qualitatively interpreted based on a comparison between the retrieved phase

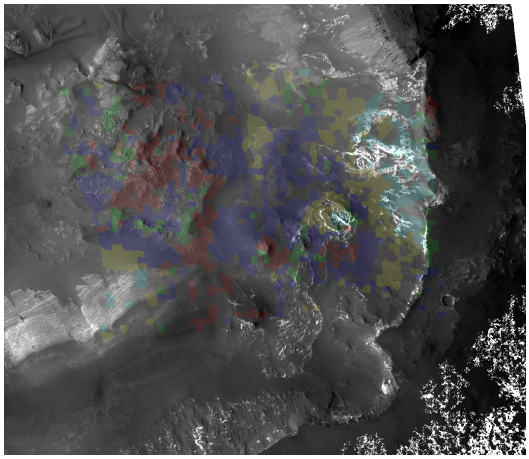

Figure 1 – Classification map superimposed on a context CTX image of a semi-circular depression in the Eos Chasma, Mars. The classification of the terrains is based on a kmeans clustering of their CRISM photometric curves at a wavelength of 755m.

ii. link PlanetGLLiM
iii. link kernelo

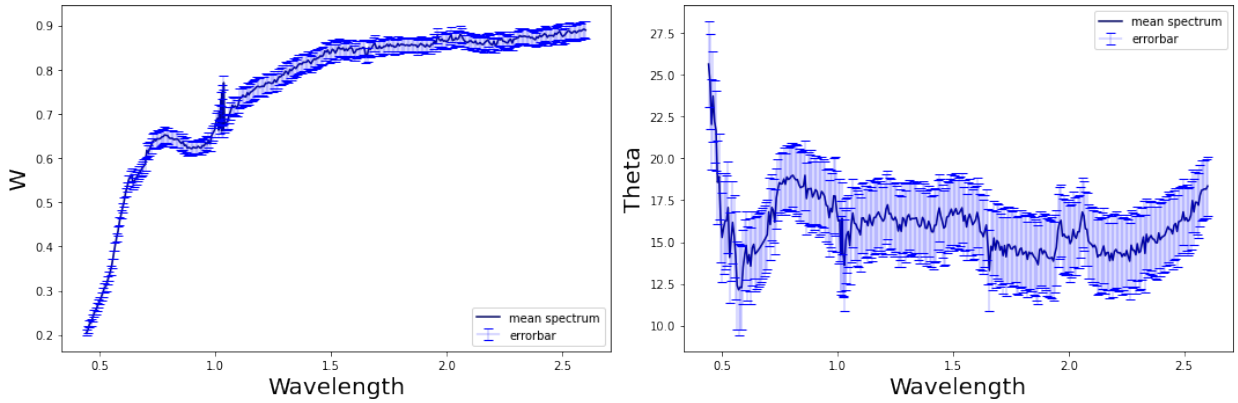

Figure 2 – Spectrum of the Hapke parameters ω and θ_H averaged over the blue terrain class. Uncertainties are indicated as error bars.

function properties of the terrain classes and that of reference materials measured in the laboratory. Figure 3 shows clearly that the hematite bearing materials are strongly backward scattering at visible wavelengths where internal absorption is high and becomes significantly forward scattering in the shortwave infrared from which a moderately rough and clear grains can be inferred. This type of texture was observed by the Mars Exploration Rover Opportunity for similar outcrops. Indeed the microscopic image presented in the inset of Figure 3 shows loose hematite spherules 4-6 mm in diameter on an outcrop of sediments at Eagle Crater.

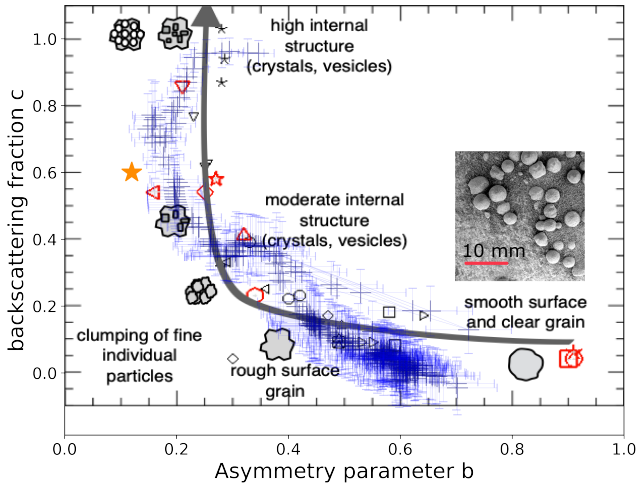

Figure 3 – Qualitative interpretation of the micro-texture of the hematite bearing terrains based on a comparison of the spectral behavior of the phase function with laboratory measurements in the Hapke (b, c) parameter space.

7 Conclusion

We have presented and illustrated the application of a statistical inversion method leveraging direct reflectance model for the analysis of high-dimensional remote sensing observations of planet Mars. The approach shows interesting capabilities both in terms of computational efficiency and inference of physically realistic, possibly multiple, solutions. The method allows, in particular, to recover not only mineral composition of Martian soil but also an estimation of granularity. We showcase the use of PlanetGLLiM, an open-source software tailored to handle inversions of reflectance models. Our results on the CRISM data contribute to the understanding of the geology and the formation of surface rocks on Mars.

References

[1] X. Ceamanos, S. Douté, J. Fernando, F. Schmidt, P. Pinet, and A. Lyapustin. Surface reflectance of Mars observed by CRISM/MRO: 1. Multi-angle Approach for Retrieval of Surface Reflectance from CRISM observations (MARS-ReCO). *Journal of Geophysical Research (Planets)*, 118:514–533.

[2] A. Deleforge, F. Forbes, S. Ba, and R. Horaud. Hyper-Spectral Image Analysis with Partially-Latent Regression and Spatial Markov Dependencies. *IEEE Journal of Selected Topics in Signal Processing*, 9(6), 2015.

[3] J. Fernando, F. Schmidt, X. Ceamanos, P. Pinet, S. Douté, and Y. Daydou. Surface reflectance of Mars observed by CRISM/MRO: 2. Estimation of surface photometric properties in Gusev Crater and Meridiani Planum. *Journal of Geophysical Research (Planets)*, 118:534–559.

[4] Bruce Hapke. *Theory of reflectance and emittance spectroscopy*. Cambridge university press, 2012.

[5] B. Kugler, F. Forbes, and S. Douté. Fast Bayesian inversion for high dimensional inverse problems. *Statistics and Computing*, 32(2):31.

[6] S. Murchie and 49 co-authors. Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) on Mars Reconnaissance Orbiter (MRO). *Journal of Geophysical Research (Planets)*, 112(E11):5–+.