Bayesian Inference on Structure Identification

He Qin

September 24, 2019

Bayesian Inference on Structure Identification

September 2019

1 Abstract

We discuss the two specific applications of Bayesian Inference[5] with deep learning[4] specifically in Structure identification problems. As deterministic modeling does not consider uncertainty during computation, it is complementary to combine it with stochastic method, for instance, Bayesian Inference when solving problem requires prior information. In the first problem, MCMC[14] is applied in randomizing the sampling of seismic waves.It is remarkably more comprehensive when the problem transfer from 1D to 2D since the randomized initialization with different MC Markov Chains do make the search of reconstruction of the seismic faster (getting convergent with more parameters) and more accurate(less inference bias misleading the inversion and less disturbed by noisy and unrelated information.) In the second problem. t-SNE[17] is conducted with modified Bayesian Information Criteria(BIC) which not only makes the computation easier and faster but also efficiently reduce the high dimensional of the data. The choice of the perplexity is also done automatically which can be updated with every iteration.

2 Introduction

In the current era, artificial intelligence[10] is developing rapidly dedicated in solving problems which is not straightforward for human or its computation is based on exponential order. The main method for machine to learn is basically either with supervision or with inference lacking supervision/semi supervised. In this paper, we combine the deep learning with Bayesian Inference on structure identification problems. Generally, the identification of structure requires human's prior knowledge about it or at least partially to construct model with some deterministic parameters. As we discuss further into seismic detection and RNA/protein sequencing, the necessity and advantage of the Bayesian inference also become clearer: Stochastic sampling helps randomize the process and against the bias and noise especially brought in by measurement. During inversion, it is also important to consider the uncertainty statically instead of global solution only based on labels at risk of no solution or overwhelming computation/time. We intentionally choose the M-H method[6] with configuration on basis of the symmetrical design of transition and ergodicity. The problem is solved with 1D model with marginal probability and applied to 2D later. And the evaluation of our model is based on 4 measures of rooted mean squared error(RMSE) with 1-norm and 2-norm roughness and depth separately after burn in.(Although the focus of the paper on seismic wave is deep learning with MCMC, to make sure the quality of the test, some basic details including preprocessing is still included in the appendix D). As for another application we discuss about is the RNA sequencing. Usually, scRNA sequenceing with the large RNA content information brings about redundancy in computation with randomized sampling. Thus, we here turn to conduct the variational inference[1]. The easiest but fastest criteria Baysian Information Criteria (BIC)[11] is modified with Kullback Leibler(KL)[13] convergence which is mainly to maximize the variational lower-bound. In some specific problem of dna/protein alignment, structural informaion can also be considered with diffusive process which here we do not introduce detailedly.

3 Methodology

In this work, the basic model is the Bayes theorem utilized in the inverse model for identification problem[9]. Generally, for a forward operator f(.):
$d = f(m) + \epsilon$,
where m is the model with N observations contained in the data vector d and ϵ is an N-dimensional vector containing the residuals, the solution can be given by the conditional posterior probability distribution:
$p(m|d) = k * \rho(m) * L(m|d)$,
where k is a normalizing constant, $\rho(m)$ is prior for N-dimensional model and L(m|d) is the likelihood function, which reflects how well a model explains the data.

As for studying the interior structure of a planet, we usually starts with the problem simplified on 1D or 2D model [18]which leads to our interest on the marginal probability distribution for a subset of the model parameters, which is,
p(m'|d) = k*ρ(m')*L(m'|d) = $\int p(m|d) dm_{s+1} dm_{s+2} ... dm_l$
where m' is a c-dimensional model vector(with s<= l). The marginal posterior pdf stands the terminal state of knowledge of m', given the variations in the l,c model parameters.With c being 1 or 2, the marginal probability distribution represents 1-D or 2-D marginal posterior pdfs, respectively standing the problem with knowledge of a single parameter model and the correlation information between any two parameters. Note that, the necessary pre-process and decompose/ simulation of P-wave, S-wave and surface wave is not introduced but included in Appendix D to make the work completed.

4 Application on low dimension problem: wave analysis[16] with radial seismic wave[15]

For the interior structure identification based on wave analysis,the simplest model which consider the misfit simply as the likelihood during the inversion of amplitude and phase:

L(m) is proportional to exp$(-\sum \frac{|A_{obs}(\omega_i)-A_{approx}(\omega_i)|^2}{2*(\sigma_i^A)^2})$,

where the first term denotes to the observed and approximated amplitude while the second term stands for the angle between observed and approximated phase, both evaluated at the frequency ω_i considering the sources of the seismic are unknown. Here, the optimal $A_{obs} = \varsigma\chi_{l'} * M_{l,l'}$, and $A_{approx}(0\omega l) = \varsigma\chi_{l'} * \frac{k}{1+k'(0\omega l'-0\omega l'')}$, where the $M_{l,l'} = \frac{k}{[1+k'*(0\omega l'-0\omega l'')^2]}$ is used in forward model for simulation the wave.(can be seen in Appendix D)

When takes it as the nonlinear inversion without analytical formulation, the M-H algorithm is applied for configuration in each iteration. The transition probability is then proportional to the acceptance ratio α:

$\alpha = min(1, \frac{L(m_{approx}|d)*\rho(m_{approx})*p(m_{approx}->m_i)}{L(m_i|d)*\rho(m_i)*p(m_i->m_{approx})})$

, where the p is the pdf of posterior generating model perturbations at each iteration i. Note that, due to Markov process's symmetries and egotistic, with the configuration(derivation see Appendix A), the chain is guaranteed to move neither too often (ratio smaller than 1) nor too rarely (ratio larger than 1) which converges to the prior $\rho(m)$

As we cannot eliminate the exception to symmetric distribution which occurs when the model parameterization evolves as part of the inversion, we thus sample the prior pdf of a global measure of model structure S(especially, considering roughness for 2D or 3D cases). Under such circumstances, the asymmetry of the $p(S(m_i- > S(m_{approx})))$ is supposed to be taken into account, which gives rise to the ratio being:

$\alpha = min(1, \frac{\rho(S(m_{approx}))*p(S(m_{approx})->S(m_i))}{\rho(S(m_i))*p(S(m_i)->S(m_{approx}))})$

, where in the numerator, $\rho(S(m_{approx}))$ is the prior probability of the chosen model structure metric, and $p(S(m_{approx})- > S(m_i))$ is the probability of proposing a given model structure when using an underlying symmetric proposal pdf $p(S(m_{approx})- > S(m_i))$ for the individual model parameters. The approximated pdf in terms of model structure is often asymmetric and depends strongly on m_i in the denominator of the ratio, andm_{approx}. The absence of an analytical expression for the proposal pdfs in the equation, requires us to estimate it numerically at each iteration, which is the to fit the unknown pdf with the gamma distribution for different degrees of symmetry(the asymmetry occurs concurrently for both positive and negative skewness while reduces to symmetry in the limit when the shape parameter goes to infinity as the distribution approximates normal distribution.):

$f(x;\beta,\theta) = \frac{x^{\beta-1}*e^{-\frac{x}{\theta}}}{\theta^{\beta}*\Gamma(\beta)}, x > 0$

In this equation, $\beta > 0$ represents the shape parameter which determines the skewness and $\theta > 0$ is the scale parameter determines the dispersion of the pdf.

Because the domain for gamma pdf is [0, inf] while the pdfs is required to be on a subdomain [a,b] where $a > 0 and b < \inf$. We thus introduce the shift parameter μ_0 translating the pdfs with: $f(x;\beta,\theta)$ obtained by the computation of gamma function with:
skewness $= \frac{2}{\sqrt{(\beta)}}$, Variance $= E[(x-\mu)^2] = \beta * \theta^2, Mean = \mu_0 + \mu = \mu_0 + \beta * \theta$

4.1 Structural Algorithm

(1)With the model at iteration i: $m_i = m_1, m_2, ..., m_M$, compute the m_{approx} with the symmetric configuration equation and the according measure of structure $S(m_{approx})$
(2)similarly, generate P new realizations from result of step 1: $p(m_i- > m_{approx})$
(3)for each i and any p belongs to 1,2,...P,compute the corresponding structure $S_p = S(m_p)$
(4)estimate the pdfs with $f(x;\beta,\theta)$ from the obtained samples: $S_1, S_2, ..S_P$ in step 3 until iterations reaches the required number

4.2 Convergence estimation

We measure the model structure with four deterministic quantities: Depth(l1-norm) sum of the absolute differences between each model parameter and a prior reference m_{ref}:
$S_{D_1} = \sum\sum |m_{i,j} - m_{ref}|$
Depth(l2-norm) sum of the squared differences between each model parameter and a prior reference m_{ref}:
$S_{D_2} = \sum\sum (m_{i,j} - m_{ref})^2$
Roughness(l1-norm) sum of the absolute differences between neighboring parameters:
$S_{R_1} = \sum\sum |m_{i,j} - m_{l,k}|$
(i,j) unequals to (l,k)
Roughness(l2-norm) sum of the squared differences between neighboring parameters:
$S_{R_2} = \sum\sum (m_{i,j} - m_{ref})^2$
(i,j) unequals to (l,k)

Practically, we calculated the SR1 and SD1 as evaluation of the data of one riverbank in India, 5 independent MCMC chains are run with the structural prior M-H acceptance ratio (v is the velocity according to real data): $\alpha = min(1, exp(l(m_{approx}|d) - l(m_i|d)) * (\frac{p(S(m_{approx})->S(m_i))}{p(S(m_i)->S(m_{approx}))})^v)))$

For 1D case(results with log2 preprocessed only is shown. more detal in table and rawdata result can be seen in appendix C), in the left figure, the posterior density of model structure is measured with gamma distributed(bins = 50, 10 ks and 10 thetas separately) synthetic data while the right one is tested with gaussian (bins = 50, 10 mus and 10 sigmas separately) both comparing

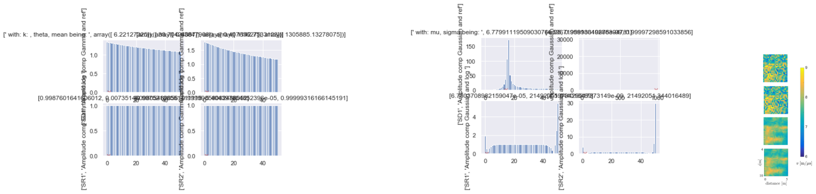

Figure 1:

with one real data(mean location marked with cross) through SD measure and comparing with one another synthetical data through SR measure. (For gamma distribution, meanSR1 =0.999912, meanSR2=0.999993, p2=0.006874 meanSD1 = 1.211205, meanSD2= 1.428754, p2 = 0.000144 while for normal distribution, meanSR1 =0.730871e-5, meanSR2=0.291341e-9, p1 = 0.033512, meanSD1 = 18.211205, meanSD2= 928.754145, p2 = 0.724534), for gamma prior, both SD and SR(norm1 and norm 2) gives small differences and according to the ttest result, the difference between two synthetica data (generated separately after resampling) and the real data both follows the same distribution. However, for the Gaussian prior, the SD gives abnormal result with real data and the ttest also reports significantly different distribtion for the SD although the SR difference is still of the same distribution. The possible reason might be the error caused in preprocess with parameter mu and sigma setting.

For 2D case, the amplitude difference of love and rayleigh wave after burn in under SR1, SR2, SD1 and SD2 measurements are shown in pixels.(The convergence is measured through RMSE(Appendix B, and all the auto-correlation function are made sure with time lag smaller than 0.2) being 0.7144, 0.4126, 0.0105, 0.0103)
The convergence for the 1D realization is of not too much difference among the 2 measurements. However, for the MCMC inversion on 2D grids 5m*4m field size, because 5 independent MCMC chains are run parallelly with 0.15 perturbation as noise to the parameters, the independent Gaussian realization is thus of significant difference(rho1 = 0.00041, rho2 = 0.00214) between different measures. The depth metrics are smoother than the roughness metrics while the 1-norm and 2-norm does not give significant difference.(rho1=0.4814, rho2 =0.3325)

5 Application on high dimension problem: Single-cell RNA Sequencing[8]

As the structure information is also quite useful in dna or protein alignment, Bayes Inference can also be applied on alignment combining diffusive structure information which is modeled with Hidden Markov[12] combining RNN[2]as sequence data.X_n. The predicted dna/proteinY_n is processed based on simialrly again forward status $\alpha_{cell} = P(Y_{s_j;\theta})$, and backforward status: $\beta_{cell} = P(Y_{t+1:t|s_{j+1};\theta})$

and the posterior is thus $\gamma = \frac{\alpha_t * \beta_t}{P(t)}$
. Since our interest here is to show the advantage of the model with high dimensional RNA sequencing data, this work is not introduced here.Instead, we are going to discuss the scRNA sequencing problem with the t-SNE with modified BIC.

As the RNA sequence data are usually of high dimension, instead of MCMC, the variational method is usually utilised. Variational inference can be regarded as the optimization problem whcih thus is of lower computation consumption. The criteria is usually based on the minimisation of the Kullback-Leiber(KL) divergence , the log difference between observed and approximated posterior distribution. The method we use is modified on t-Distributed Stochastic Neighbor Embedding(t-SNE) which converts pairwise distances in high dimensional space with data points x_i, to corresponding embedding points y_i pairwise join distributions in low dimensional spaces, which respectively follows:
$\frac{q_{i,j}=(1+|y_i-y_j|^2)^{-1}}{\sum(1+|y_s-y_t|^2)^{-1}}$
while high dimensional one is defined in symmetrical conditions: $p_{i,j} = (p_{i|j} + p_{j|i})/2n)$, where
$p_{i|j} = \frac{exp(-|x_i - x_j|^2/2\sigma_{j2})}{\sum exp(-|x_s - x_t|^2/2\sigma j^2))}$ and the KL to be optimised is thus:
$KL(P||Q) = \sum p_{i,j} * log\frac{p_{i,j}}{q_{i,j}}$ Note that the σ_j is optimized through bisectional search automatically with the pre-specified perplexity $Perplex(p_j) = 2^{H(p_j)}$, where $H(P_j) = -\sigma_j p_{i|j} * log_2^{i|j}$ ao that $Perplex(p_j) = Perplex$, where Perplex is the hyperparameter of the t-SNE central to the final cluster.

Large Perplex usually leads to the embedding suboptimal in detecting the pattern of the data(In the limit, when the Perplex goes to the number of data points, the corresponding embedding form a Gaussian or uniform like distribution and fails to be useful for structure detection at all) and thus, we design a new criteria:
$S(Perplex) = KL(P||Q) + log(n) * \frac{Perplex}{n}$
with inspiration of the Bayesian Information Criteria(BIC):
$BIC = -2 * log(L) + log(n) * k$
, where the first term stands for the goodness-of-fit of the maximum-likelihood-estimation and the second controls the complexity of the model with penalty k scaled by log(n). Intuitively, when Perplex increases, differences among points will become less and less significant with regard to the length of the kernel in distribution P, and P will tend to uniform.The forward form of KL has large cost for under-estimating probability but not for over-estimating. That is, if $p_{i,j}$is large and $q_{i,j}$ is small, KL divergence is large while in the opposite direction, KL is not affected. Increasing Perplex leads to larger σ_j and more uniform $p_{i,j}$ so it is easier is for the student-t distribution in low dimensional space to assign mass for all probability points sufficiently. This is the so called crowding problem: When projecting from high to low dimensional space, there is not enough room in lower dimensional space.Generally, increasing Perplex relaxes the problem and reduces the amount of structure to be modelled with less error according to KL while pays a cost in the second term.

With the practical test, we apply it on the MC GC cell expression classification with t-SNE

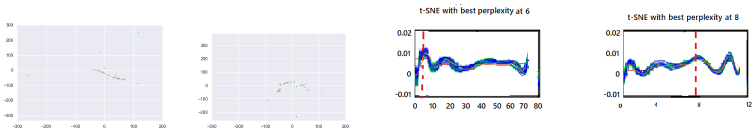

Figure 2:

The top figure left is the trajectory of the gcs and mc classified under 10 groups while the top right figure gives the result classified with 5 groups and the perplex are given as 8 and 6 separately as can be seen in the bottom figure.(The optimization is based on KL minimization with fast-gradient search)

The modified BIC v.s.perplexity shows the best classification is with 10 classifications with perplexity being 8. Both of the figures are shown on the first 2 dimensions of the embedding and it can be observed that since the reduction of the perplex is sharper, the classifications are also more easily separable non-linearly (closer within one classification while more alienated with inter classifications.) In fact, the standard deviation are all not too large along the whole tested perplex domain showing the good-to-fitness of the t-SNE which is as well of certain stability. As it being said, t-SNE performs generally well on high dimension classification with the modified BIC.

6 Conclusion and Discussion

As deep learning[7] is a deterministic method which cannot consider the uncertainty during computation, mainly more models depend largely on regulation and dropouts with trim to avoid overfitting problem. However, for practical problem, especially those requires efficiency and accuracy at the same time, the combination of stochastic modeling[3], such as Bayes inference models is unassailable important. For the two problems discussed in this paper, we first imply Bayes combining MCMC to randomize the sampling of the data for getting better estimation without bias and robustly against noise. As for EEG data(seismic here),it is especially easy to get white noise disturbing the identification of 'real' information. Thus, our model based on MCMC sampling aims at working against noise both brought by measurement as well as inverse computation. To improve in the future, the neuro-network with mixed density might be considered as it can also learn non-linearly through the hidden units in the dark box which can contain more factors which is also important to the inference of interior structure and do not requires the knowing of the sources and mechanisms. It might be easier for generalizing onto more complex model especially with 3D wave equations. For the sequencing problem, the t-SNE is discussed here with modification on the BIC criteria. Comparing to the MCMC, the computation with variation is faster. The advantage of reducing high dimension can also be observed according to the small uncertainty. With the Bayesian inference, the best perplexity is chosen automatically. Improvement can be made on searching process which for now is on basis of gradient while it will be faster and better in avoiding local optimization with second-order-gradient-search. However, considering different requirement or criteria, future exploration might explore more on adding some prior as well.

7 AppendixA: The design of M-H MCMC acceptance ratio

The M-H algorithm sample a unique stationary odf: m, because it fulfills the following two conditions:

(1)Detailed balance(or reversibility) is a sufficient condition for a random walk to asymptotically reach a stationary pdf, requiring each transition$(m_i- > m_{approp})$ to be reversible and it can be stated as: given a transition matrix $T(m_{approp}|m_i)$ a stationary distribution (m) is $T(m_{approp}|m_i)(m_i) = T(m_i|m_{approp})(m_{approp})$

(2)Ergodicity of the Markov process requires that every state must be aperiodic(e.g., the system does not return to the same status at fixed intervals), positive recurrent(e.g., the expected number of steps for returning to the same state is finite) and irreducible(e.g., each status is accessible in a finite nunmers). The condition guarantees the uniqueness of the staionary pdf (m).

As mentioned, the central aspect of MCMC theory is to define transition kernels, such that the sequence of samples drawn will converge to the target pdf (m).If the chain proposes a move from m_i to m_{approp} , such that

$\rho(m_i) * q(m_i- > m_{approp}) > \rho(m_{approp}) * q(m_{approp}- > m_i)$, then this implies that the chain moves too often from m_i to m_{approp} and too rarely in the other direction. To counteract this tendency, the M-H algorithm reduces the number of moves from m_i to m_{approp} to achieve detailed balance. This is done by introducing a probability $(0 < \alpha(m_i, m_{approp}) < 1$ that the proposed move is executed. The resulting transition is defined as:

$T(m_{approp}|m_i) = q(m_i- > m_{approp})alpha(m_i, m_{approp})$ The probability $alpha(m_i, m_{approp})$is calculated to ensure that $T((m_{approp}|m_i)$satisfies the detailed balance criterion, as

$\rho(m_i)T((m_{approp}|m_i) = \rho(m_{approp})T((m_i|m_{approp})$ and thus, we have

$\rho(m_i)q(m_i- > m_{approp})alpha(m_i, m_{approp}) = \rho(m_{approp})q(m_{approp}- > m_i)alpha(m_i, m_{approp})$

According to the inequality of $\rho(m_i) * q(m_i- > m_{approp}) > \rho(m_{approp}) * q(m_{approp}- > m_i)$, the move from $m_{approp}$to m_i is not made often enough. Setting $alpha(m_i, m_{approp}) = 1$, the final acceptance ratio without data can be achieved:

$\alpha = min(1, \frac{\rho(m_{approx}) * p(m_{approx} - > m_i)}{\rho(m_i) * p(m_i - > m_{approx})})$

8 AppendixB: Some quantities for evaluation the 2D MCMC acceptance

$RMSE_M = \frac{1}{M} * ||\frac{m_{act}-m}{m_{act}}||_2$

	RMSE$_M$	R of SR1(mean)	R of SR2(mean)	R of SD1(mean)	R of SD2(mean)
Love	0.412412	0.7144	0.4126	0.0105	0.0103
Rayleigh	0.227456	0.6799	0.5522	0.0455	0.0872
ttest	0.714239	0.412202	0.213118	0.659727	0.378294

9 AppendixC: The seimic data from IRIS(descriptive review and pre-process)

Seimic wave on NS, WE and z axis are used. Prerocess as baseline correction,different transformation onto frequency domain and normaliation are conducted before forward and inverse process.

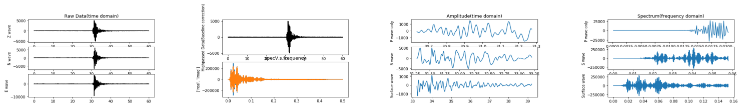

Figure 3:

10 AppendixD: forward problem of seimic wave

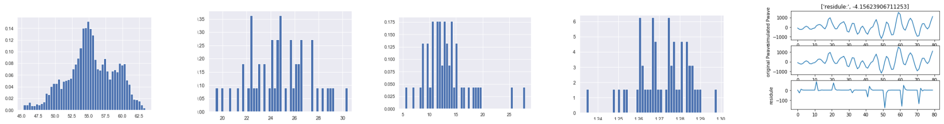

Figure 4:

References

[1] Martin Adamčík. *The Information Geometry of Bregman Divergences and Some Applications in Multi-Expert Reasoning.* Entropy, 2014.

[2] N.; Pascanu R Bengio, Y.; Boulanger-Lewandowski. *Advances in optimizing recurrent networks.* IEEE International Conference on Acoustics, Speech and Signal, 2013.

[3] N.; Pascanu R Bengio, Y.; Boulanger-Lewandowski. *Lectures on stochastic programming: Modeling and theory.* 2013.

[4] Yann; Hinton Geoffrey Bengio, Yoshua; LeCun. *Deep Learning*. Nature, 2015.

[5] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, 1973.

[6] Pierre Del Moral. *Sequential Monte Carlo samplers*. Wiley Online Library, 2006.

[7] D Deng, L.; Yu. *Deep Learning: Methods and Applications*. Foundations and Trends in Signal Processing., 2014.

[8] Dana Pe'er Elham Azizi † Sandhya Prabhakaran , Ambrose Carr. *Bayesian Inference for Single-cell Clustering and Imputing*. GENOMICS AND COMPUTATIONAL BIOLOGY, 2016.

[9] Anthony Rodgers b Gareth Brown a, Kevin Ridley b and Geoffrey de Villiers b. *Bayesian signal processing techniques for the detection of highly localised gravity anomalies using quantum interferometry technology*. SPIE, 2016.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Estimating the dimension of a model*. Annals of Statistics, 2016.

[12] H.Q. *Granule and mossy cell gene expression study: single cell sequence and alignment monitored by electroencephalography scans with Markov Chain and Bayes Hierarchical Model*. WiML2019, 2019.

[13] S Kullback. *Information Theory and Statistics*. John Wiley Sons, 1978.

[14] Faming; Wong Wing Hung Liu, Jun S.; Liang. *The Multiple-Try Method and Local Optimization in Metropolis Sampling*. Journal of the American Statistical Association, 2000.

[15] Sampsa Pursiainen. *A Mathematical Approach to Reconstructing the Interior of a Small Solar System Body via Full-Wave Computed Radar Tomography*. IEEE, 2016.

[16] Andrew P. Valentine Ralph W. L. de Wit and Jeannot Trampert. *Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks*. Geophysical Journal International, 2013.

[17] G.E van der Maaten, L.J.P.; Hinton. *Visualizing Data Using t-SNE*. Journal of Machine Learning Research, 2008.

[18] Felix Bissig 1 ·Amir Khan 1 ·Martin van Driel 1 ·Simon C. Stähler 1 Domenico Giardini 1 ·Mark Panning 2 ·Mélanie Drilleau 3 ·Philippe Lognonné 3 · Tamara V. Gudkova 4 ·Vladimir N. Zharkov 4 ·Ana-Catalina Plesa 5 · William B. Banerdt. *On the Detectability and Use of Normal Modes for Determining Interior Structure of Mars*. Spinger, 2018.