3D Data-driven Site Characterization using Geotechnical Lasso with Basis Functions

K.K. Phoon* and T. Shuku

*Architecture and Sustainable Design/Information Systems Technology and Design, Singapore University of Technology, Singapore

Graduate School of Environment and Life Science, Okayama University Japan

Abstract

Geotechnical lasso (Glasso), which is a data-driven site characterization (DDSC) method proposed by the authors, is advantageous to identification of abrupt changes in data such as soil layer boundaries because it is formulated with piecewise functions. In Glasso, target grounds are split into many cells or cuboids, and geotechnical parameters for all the cells/cuboids need to be identified. It requires massive computation time and memory space and has limitations in high degree-of-freedom (DoF) problems. This study newly proposes a Glasso based on continuous basis functions (BFs) to reduce DoF, i.e., to address the limitations of the existing Glasso. The applicability of the proposed Glasso with BFs is investigated through DDSC for virtual grounds. This study also sheds light on the advantages and disadvantages of two types of Glasso based on the comparison.

Keywords: data-driven site characterization, lasso, basis functions, benchmark examples

1. Introduction

Decision making in geotechnical engineering is always linked to a project conducted at a specific site, and it is natural for data-driven site characterization (DDSC) to receive attention in data centric geotechnics. The primary goal of DDSC is to map 3D stratigraphic profile below a project site and to assign properties to each spatial point. This approach is called maximum likelihood (ML) method with variance \( \sigma^2 \)

\[ y = f(x) + \varepsilon \]  

where \( \varepsilon \in \mathbb{R}^{N \times 1} \) is a zero mean Gaussian random variable with variance \( \sigma^2 \).

In the proposed method, we employ linear combinations of fixed nonlinear functions of the input variables as the structure \( f(x) \), of the form

\[ f(x) = f(x, w) = \sum_{k=1}^{M} w_k \varphi_k(x) \]  

where \( w_k \) is an unknown coefficient and \( \varphi_k \) is the \( k \)-th BF. There are many possible choices for BFs, such as polynomials, sigmoidal, wavelet, and Legendre. Original Glasso uses a piecewise function \( w = [w_1, w_2, \ldots, w_N]^T \in \mathbb{R}^{N \times 1} \) for \( f(x) \). In this study, shifted Legendre polynomial basis is used as the BF to compare the performance of the proposed method with that of the existing method, sparse Bayesian learning (SBL) that is proposed by Ching and Phoon [6]. The first few shifted Legendre polynomials are:

\[ \varphi_0(x) = 1 \]

\[ \varphi_1(x) = 2x - 1 \]

\[ \varphi_2(x) = 6x^2 + 6x + 1 \]

\[ \varphi_3(x) = 20x^3 - 30x^2 + 12x - 1 \]

Once a specific structure is given, the subsurface modelling becomes the problem of identifying unknown coefficient vector \( w \). A straightforward approach to estimate \( w \) and \( \sigma^2 \) is by maximizing the likelihood function defined by

\[ p(y \mid w, \sigma^2) = \left(2\pi\sigma^2\right)^{-N/2} \exp \left(-\frac{1}{2\sigma^2} \| y - \Phi w \|^2 \right) \]

where \( \Phi \) is an \( N \times M \) matrix, called the design matrix, whose elements are given by \( \Phi_{ij} = \varphi_i(x_j) \):

\[ \Phi = \begin{bmatrix}
\varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_M(x_1) \\
\varphi_0(x_2) & \varphi_1(x_2) & \cdots & \varphi_M(x_2) \\
\vdots & \vdots & \ddots & \vdots \\
\varphi_0(x_N) & \varphi_1(x_N) & \cdots & \varphi_M(x_N)
\end{bmatrix} \]

This approach is called maximum likelihood (ML) method and usually leads to an excessively complex model that over-fits the data. To achieve reasonable modelling, the Bayesian approach has been widely used in data-driven site characterization. Based on Bayes’ rule, the posterior probability density function (PDF) of \( w \) and \( \sigma^2 \), \( p(w, \sigma^2 \mid y) \), is given by:

* E-Mail: kkphoon@sutd.edu.sg
\[ p(w, \sigma^2 | y) = \frac{p(y | w, \sigma^2) p(w, \sigma^2)}{p(y)} \] (6)

where \( p(w, \sigma^2) \) is the prior PDF of \( w \) and \( \sigma^2 \), and \( p(y) \) is a normalization term. This normalization term is often left out, and the posterior PDF can be written as

\[ p(w, \sigma^2 | y) \propto p(y | w, \sigma^2) p(w, \sigma^2) \] (7)

Lasso assumes the following PDF for the prior of coefficient vector \( w \):

\[ p(w | \kappa) \propto \exp(-\kappa \|w\|_1) \] (8)

where \( \kappa \) is the diversity parameter in the Laplace PDF, and \( \| : \| \) is an L1 norm, which stands for sum of absolute values of \( w_i \). By substituting Eqs. 4 and 8 into Eq. 7, we get the following objective function \( J \) for this problem:

\[ J = \frac{1}{2} \| y - \Phi w \|^2 + \lambda \|w\|_1 \] (9)

where \( \lambda = \kappa \sigma^2 \). The parameter \( \lambda \) is called regularization parameter that controls the complexity of the polynomial model \( \Phi w \).

The \( L_1 \) regularization term plays a role to select the important polynomials in all the predetermined Legendre polynomials. This means that the best model \( f(x, w) \) for the given data \( y \) is automatically selected by minimizing Eq. (9) without solving model selection problem independently from estimating the coefficients \( w \). This study used the alternating direction method of multipliers (ADMM[7]) for this convex minimization problem.

In the proposed method, the DoF of the problem to be solved, i.e., the number of unknown parameters, does not depend on the physical scale of the problem and the spatial resolution of unknown vector unlike original Glasso.

3. Examples for Basis Function Selection

This section demonstrates the proposed method in the following synthetic examples on selecting basis functions, which was solved by Ching and Phoon [6] using SBL:

1. EX1: \( \kappa(x) = 100 + 200x \); 
2. EX2: \( \kappa(x) = 100 + 100x - 1000x^2 + 1000x^3 \); and
3. EX3: \( \kappa(x) = 100 + 100 \cos(4\pi x) \).

All the numerical setup is same as that used in [6], i.e., spatial variability \( \varepsilon(x) \) is with standard deviation \( \sigma = 20 \) and scale of fluctuation (SoF) \( \delta = 0.1 \). The shifted Legendre polynomials with order up to 20 \((M = 20)\) are adopted as the BFs. The regularization parameter \( \lambda \) was determined based on an empirical method called \( L \)-curve method [8].

Fig. 1 shows one simulated realization of data \( y \) (thin continuous line) and the estimated models \( f(x) \) (thick grey line) using Glasso with BFs. Fig. 2 shows the absolute values of coefficients \( w_i \) finally identified using ADMM. For all examples, many \( w_i \) values are zeros, and simple models are selected due to the \( L_1 \) regularization term. In EX1 and EX2, the parameters selected by Glasso are same as those by SBL [6]. We found that Glasso works similarly to SBL in model selection problems. There may be some theoretical connections and equivalences between lasso and SBL, and they are worth investigating for future study.

4. Benchmarking

This chapter demonstrates the Glasso with BFs through the benchmark examples proposed by Phoon et al. [5] which are designed for supporting unbiased and competitive evaluation of emerging DDSC methods. In the benchmark examples, four types of virtual grounds that have different stratigraphy (Fig. 3) were proposed using cone penetration test (CPT) data. The physical volume of the virtual ground for all the benchmark examples is defined by 20 m long \( \times \) 20 m wide \( \times \) 10 m deep cuboid. Synthetic CPT data, consisting of cone tip resistance \( q_t \) and...
sleeve friction $f_s$, were generated from a bivariate normal distribution. The simulated $q_t$ and $f_s$ values are assigned into the mid-point of the cells. The training and validation datasets consist of depth profiles of $q_t, f_s,$ and $I_c$ “measured” at various prescribed locations in the virtual grounds. Fig. 4 shows the plan view of the sounding locations for these training and validation datasets.

To compare the performance of DDSC methods, root-mean-square error (RMSE) of $q_t$ and a simple identification rate (IR) based on soil behaviour type (SBT) are used:

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - x_i)^2} \quad (10)$$

$$\text{IR} = \frac{1}{m} \sum_{i=1}^{m} I_i \cdot I_i = \begin{cases} 1 & \text{SBT}(y_i) = \text{SBT}(x_i) \\ 0 & \text{SBT}(y_i) \neq \text{SBT}(x_i) \end{cases} \quad (11)$$

where $y_i$ and $x_i$ is measured and estimated $q_t$ values at the $i^{th}$ depth, SBT($y_i$) and SBT($x_i$) are the SBTs corresponding to $y_i$ and $x_i$, respectively.

We used the benchmark example named “S-VG2 (Fig. 3)” with training set 2 (T2) and compared the performance (RMSE and IR) of the proposed method with that of original Glasso for this benchmark dataset.

The $q_t$ profiles estimated by the Glasso with BFs and original Glasso are shown in Fig. 5. In the figure, continuous (black) lines indicate validation data, dashed (blue) lines indicate the estimations by Glasso, and dotted (red) lines indicate the estimations by Glasso with BFs. Both methods well captures the profiles of data. While original Glasso can capture the abrupt changes in the data, Glasso with BFs cannot capture them and gives smoother profiles because it is based on continuous basis functions. In the comparison of No.381 validation data, the estimation of Glasso with BFs around the bottom of the ground is highly fluctuated. This can be interpreted as so-called “overfitting” in the context of machine learning. Since estimating profile there is “extrapolation” rather than “interpolation”, such “overfitting” occurs.

Figs 6 and 7 show the comparison of RMSEs of $q_t$ and IR between original Glasso and Glasso with BFs respectively. Although both methods give similar RMSEs
and IRs, original Glasso shows better performance (smaller RMSEs and higher IR) at all the locations.

The performance of both methods are summarized in Table 1. The “Runtime” is also included in the table because it is important for practice. The runtime means the time required for training and validation in this paper. Original Glasso required about 5 hours to reconstruct the 3D image of the virtual ground. The proposed method, however, needed about only 7 minutes and does not have the limitations about computation time and memory original Glasso has.

<table>
<thead>
<tr>
<th></th>
<th>Glasso</th>
<th>Glasso with BFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best RMSE of $q_t$</td>
<td>1.087</td>
<td>1.255</td>
</tr>
<tr>
<td>Median RMSE of $q_t$</td>
<td>1.458</td>
<td>2.346</td>
</tr>
<tr>
<td>Worst RMSE of $q_t$</td>
<td>2.904</td>
<td>4.704</td>
</tr>
<tr>
<td>Best IR</td>
<td>0.82</td>
<td>0.69</td>
</tr>
<tr>
<td>Median IR</td>
<td>0.74</td>
<td>0.65</td>
</tr>
<tr>
<td>Worst IR</td>
<td>0.56</td>
<td>0.42</td>
</tr>
<tr>
<td>PC Specs</td>
<td>Intel Core i7-8700 CPU 3.2 GHz with 32GB RAM</td>
<td></td>
</tr>
<tr>
<td>Runtime [sec]</td>
<td>19.347.6</td>
<td>399.0</td>
</tr>
</tbody>
</table>

5. Summary

This study proposed a new Glasso with basis functions to deal with the limitations of original Glasso, high computational complexity and high memory consumption. The formulation of the new method is presented, and its advantage was demonstrated in the examples for basis function selection. In addition, the performance of the proposed method was evaluated through the existing benchmark example (S-VG2) and was compared to that of original Glasso in terms of RMSE, IR, and runtime.

Although the proposed method, Glasso with BFs, reasonably estimated the $q_t$ profiles and stratifications of validation data, the performance is less than the original Glasso. However, the proposed method is much faster than the original Glasso and does not require large memory.

Acknowledgment

This research was partly supported by JSPS KAKENHI Grant Number JP18K03880 and JP20H00442.

References