The Seventh Asian-Pacific Symposium on Structural Reliability and Its Applications (APSSRA2020)
October 4–7 2020, Tokyo, Japan
T. Takada, I. Yoshida & T. Itoi (editors)

Model Selection for Soil Stratification by Sparse Modelling

I. Yoshida[1] and T. Shuku[2]

[1] *Department of Urban and Civil Engineering, Tokyo City University. Email: iyoshida@tcu.ac.jp*
[2] *Graduate School of Environmental and Life Science, Okayama University, Email: shuku@cc.okayama-u.ac.jp*

Abstract: Spatial distribution of soil property is often modelled by random component and trend component in Geostatistics. We developed a new method that estimates trend component based on fused lasso, which is one of the formulation for sparse modeling, and random component based on Kriging, namely Bayesian update of Gaussian random field or Gaussian Process Regression. The method requires parameters for random field such as autocorrelation distance and sparsity parameter which control the level of sparsity. This paper studies applicability of information criteria BIC to determine them. The parameters are simultaneously estimated by minimization of BIC. For eight random realizations, the trend components are estimated. They show generally good agreement with the true model.

Keywords: spatial distribution, trend component, information criteria, model selection, Bayesian update.

1. Introduction

Understanding soil variability is important in the geotechnical engineering. Ching and Phoon (2019) classified the uncertainties in geotechnical engineering into four categories, namely spatial variability, transformation uncertainty, statistical uncertainty, and measurement error. The spatial variability of soil properties is a major source of geotechnical uncertainty. Spatial variability has been extensively studied and applied to reliability analysis and optimal observation planning. Papaioannou and Straub (2017) applied Bayesian analysis to determine spatially varying soil properties. Yoshida et al. (2018) proposed a method for optimal sampling planning in terms of the number and placement of additional sampling points based on value of information, which can be computed easily by updating a Gaussian random field.

Spatial distribution of soil property is often modelled by random component and trend component in Geostatistics. The estimation of the random component is well established, whereas that of the trend component based on observation data is not. The soil stratification, namely the identification of trend component including boundary, is one of the important topics in geotechnical engineering. Nishimura et al. (2016) selected the trend and covariance functions of a property from prepared models. In this approach, however, the result strongly depends on the set of prepared models or trend functions. Ching and Phoon (2017) avoided this problem by preparing many basis functions and selected a suitable collection of basis functions for the modeling of the trend component using sparse Bayesian learning.

Sparse modeling has attracted attention in many fields, including medical science, astronomy, geophysics, and civil engineering. According to the general principle of sparsity, a phenomenon should be represented by as few variables as possible. One of the most widely adopted methodologies for sparse modeling is the least absolute shrinkage and selection operator (lasso) proposed by Tibshirani (1996). The concept, methodology, and application of lasso in many fields were discussed by Hastie et al. (2015). We might expect the trend component to be piecewise-constant over contiguous regions of the soil parameters. Lasso produces sparse solutions with many coefficients equal to 0. Fused lasso (Tibshirani et al. 2005), also known as total variation denoising or trend filtering (Kim et al. 2009, Hastie et al. 2015), additionally produces sparsity in solution differences (i.e., neighboring coefficients are similar), making some solutions identical.

The author developed a new method that estimates trend and random component in geotechnical data based on fused lasso and Kriging (e.g., Christakos 1992; Cressie 1991), namely Bayesian update of Gaussian random field or Gaussian Process Regression (Rasmussen and Williams, 2006). The proposed method simultaneously estimates the trend component, correlated random component, and uncorrelated observation noise.

2. Formulation of random and trend field

2.1 Random Component

Kriging is a probabilistic interpolation method that can also be interpreted as a probabilistic inverse problem. Here formulation for random component with assumed trend is summarized. Please refer Yoshida et al. (2018) for the detail. To numerically represent a continuous random field, it is necessary to discretize it with a finite set of random variables, represented by vector $\boldsymbol{x}$. Consider the case for which the observation vector $\boldsymbol{z}$ is sampled from a random field with given trend, autocovariance function, and observation noise.

$$\boldsymbol{x}^T = \left\{\boldsymbol{x}_1^T, \boldsymbol{x}_2^T\right\} \tag{1}$$

where $\boldsymbol{x}_1$ denotes variables at observation sites and $\boldsymbol{x}_2$ denotes variables in the region to be estimated. Separating

the prior covariance matrix corresponding to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ yields

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{12}^T & \boldsymbol{M}_{22} \end{bmatrix}, \text{ where, } \boldsymbol{M}_{ij} = E[(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_j - \bar{\boldsymbol{x}}_j)^T] \tag{2}$$

We have random variables and their covariance matrix $\boldsymbol{P}$ updated by the observation vector $\boldsymbol{z}$.

$$\begin{Bmatrix} \hat{\boldsymbol{x}}_1 \\ \hat{\boldsymbol{x}}_2 \end{Bmatrix} = \begin{Bmatrix} \bar{\boldsymbol{x}}_1 \\ \bar{\boldsymbol{x}}_2 \end{Bmatrix} + \begin{bmatrix} \boldsymbol{M}_{11} \\ \boldsymbol{M}_{12}^T \end{bmatrix} \left[\boldsymbol{M}_{11} + \boldsymbol{R}\right]^{-1} \{\boldsymbol{z} - \bar{\boldsymbol{x}}_1\} \tag{3}$$

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} \\ \boldsymbol{P}_{12}^T & \boldsymbol{P}_{22} \end{bmatrix} \tag{4}$$

$$\text{where } \boldsymbol{P}_{ij} = E[\boldsymbol{M}_{ij} - \boldsymbol{M}_{i1}(\boldsymbol{M}_{11} + \boldsymbol{R})^{-1}\boldsymbol{M}_{1j}]$$

Prior covariance matrix $\boldsymbol{M}$ is often formulated based on an autocovariance function. Several types of autocovariance function have been proposed. In this paper, the following equation is used:

$$R(d) = \sigma_M^2 \exp\left\{-\left(\frac{d}{a}\right)^2\right\} \tag{5}$$

where d and a are respectively the distance between two points in $\boldsymbol{x}$ and the autocovariance distances (related to the scale of fluctuation), and σ_M^2 is the variance of random component.

2.2 Trend Component

The determination of the trend component is challenging because observation data $\boldsymbol{z}$ contain trend component $\bar{\boldsymbol{x}}_1$, random component $\boldsymbol{w}$, and observation noise $\boldsymbol{v}$ (i.e., $\boldsymbol{z} = \bar{\boldsymbol{x}} + \boldsymbol{w} + \boldsymbol{v}$). This paper proposes a method for estimating the trend component using the concept of sparse modeling. The following relation is assumed for the trend component.

$$\bar{\boldsymbol{x}}_1 = \boldsymbol{T}\bar{\boldsymbol{x}}_2 \tag{6}$$

where $\boldsymbol{T}$ is a matrix used to connect the trend component vector at observation locations and that at evaluation locations. The simplest form of $\boldsymbol{T}$ is a matrix whose entry closest to an observation point is one and all other entries in the row are zero.

Although the spatial distribution of a geotechnical property is not generally sparse, its derivatives might be sparse. The vector $\boldsymbol{x}_2$ is assumed to be a depth profile with a fixed interval of a soil parameter of interest. The matrix for calculating the first-order difference (corresponding to first-order differentiation) $\boldsymbol{D}$ is defined as:

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \tag{7}$$

If sparsity is expected in a space of first-order differentiation, the trend component $\bar{\boldsymbol{x}}_2$ can be obtained by solving the following equation.

$$\min_x \left\{ \frac{1}{2}\left\|\boldsymbol{z} - \boldsymbol{T}\bar{\boldsymbol{x}}_2\right\|_2^2 + \frac{1}{b}\left\|\boldsymbol{D}\bar{\boldsymbol{x}}_2\right\|_1 \right\} \tag{8}$$

where $\|\,.\,\|_1$, $\|\,.\,\|_2$ are an ℓ_1, ℓ_2 norm, which represent the sum of the absolute values, and squared values of the component respectively. In statistics, the problem in Eq. (8) is widely known as lasso (Tibshirani 1996, Hastie et al. 2015). The regularization parameter b controls the sparsity or complexity of the space. A large value of b allows the trend component to adapt more closely to the observation data. Conversely, a small value of b restricts the parameters, leading to a simple trend that fits the observation data less closely.

The objective function for the minimization can be written using the related covariance matrices shown in a previous section:

$$J = \frac{1}{2}(\boldsymbol{z} - \boldsymbol{T}\bar{\boldsymbol{x}}_2)^{\mathrm{T}}\left[\boldsymbol{M}_{11} + \boldsymbol{R}\right]^{-1}(\boldsymbol{z} - \boldsymbol{T}\bar{\boldsymbol{x}}_2) + \frac{1}{b}\left\|\boldsymbol{D}\bar{\boldsymbol{x}}_2\right\|_1 \tag{9}$$

For the minimization of this objective function, we use the alternating direction method of multipliers (ADMM) (Hastie et al. 2015), which blends the decomposability of the dual-ascent method with the superior convergence properties of the method of multipliers. Equation (9) is rewritten as follows:

$$J = f(\boldsymbol{x}_f) + g(\boldsymbol{x}_g), \quad \text{s.t.} \quad \boldsymbol{x}_g = \boldsymbol{D}\boldsymbol{x}_f \tag{10}$$

where

$$\boldsymbol{x}_f = \bar{\boldsymbol{x}}_2, \quad f(\boldsymbol{x}_f) = \frac{1}{2}(\boldsymbol{z} - \boldsymbol{T}\boldsymbol{x}_f)^{\mathrm{T}}\left[\boldsymbol{M}_{11} + \boldsymbol{R}\right]^{-1}(\boldsymbol{z} - \boldsymbol{T}\boldsymbol{x}_f),$$

$$g(\boldsymbol{x}_g) = \frac{1}{b}\left\|\boldsymbol{x}_g\right\|_1$$

The objective function of Eq. (9) is solved using the Lagrange multiplier method, as shown in Eq. (10). The algorithm for the minimization of J is summarized as follows:

Step 1: initialize $\boldsymbol{x}_f$, $\boldsymbol{x}_g$, and $\boldsymbol{u}$ and set $k = 0$

Step 2:
$$\begin{aligned} \boldsymbol{x}_f[k+1] = &\left(\boldsymbol{T}^{\mathrm{T}}\left[\boldsymbol{M}_{11} + \boldsymbol{R}\right]^{-1}\boldsymbol{T} + \mu\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\right)^{-1} \\ &\left(\boldsymbol{T}^{\mathrm{T}}\left[\boldsymbol{M}_{11} + \boldsymbol{R}\right]^{-1}\boldsymbol{z} + \mu\boldsymbol{D}^{\mathrm{T}}(\boldsymbol{x}_g[k] - \boldsymbol{u}[k])\right) \end{aligned} \tag{11}$$

Step 3: $$\boldsymbol{x}_g[k+1] = \boldsymbol{s}_{1/\mu b}\left(\boldsymbol{D}\boldsymbol{x}_f[k+1] + \boldsymbol{u}[k]\right) \tag{12}$$

$$\text{where,} \quad [\boldsymbol{s}_c(\boldsymbol{y})]_i = \begin{cases} y_i - c & y_i > c \\ 0 & -c \le y_i \le c \\ y_i + c & y_i < -c \end{cases}$$

Step 4: $\boldsymbol{u}[k+1] = \boldsymbol{u}[k] + (\boldsymbol{D}\boldsymbol{x}_f[k+1] - \boldsymbol{x}_g[k+1])$

(13)

Step 5: if $|\boldsymbol{D}\boldsymbol{x}_f - \boldsymbol{x}_g| < \varepsilon$, then stop;
otherwise, $k = k + 1$ and go to Step 2.

where ε is the parameter for the stopping criterion, μ is the penalty parameter associated with the constraint, $\boldsymbol{u}$ are Lagrange multipliers, and $\boldsymbol{S}_c(\boldsymbol{y})$ is the soft thresholding operator (Hastie et al. 2015). The algorithm involves a ridge regression update for $\boldsymbol{x}_f$, a soft-thresholding step for $\boldsymbol{x}_g$, and a simple linear update for $\boldsymbol{u}$.

The lasso gives shrinkage estimation. The estimated coefficients are shrunken towards zero compared to the least-squares solution without a regularization term. In the case of the first-order differential lasso, namely fused lasso, the estimated trends are shrunken towards the mean of all observation data. To avoid this shrinkage problem, the following two-step algorithm is employed. In step 1 (model selection by lasso), certain coefficients are set to zero by the minimization of Eq. (9) and hence excluded from the model. In step 2, least-squares regression is applied to the selected variables. This method is basically the same as relaxed lasso (Meinshausen et al. 2007).

2.3 Model Selection with Information Criteria

Cross validation (Hastie et al. 2015) is widely used to determine regularization parameter b in lasso. Unlike the ordinary lasso formulation, three parameters, namely the variances of the random component and observation error and the autocovariance distance, have to be determined simultaneously in addition to regularization parameter b because they are related to each other. To reduce the computation time, we use the AIC or BIC (Akaike 1974, Schwartz 1978) to determine these parameters, including regularization parameter b.

The AIC and BIC are given by

$$AIC = -2\ln L + 2n_h \tag{14}$$

$$BIC = -2\ln L + \ln(m)n_h \tag{15}$$

where -ln L is the negative log-likelihood, expressed as

$$-\ln L = \frac{1}{2}(\boldsymbol{z} - \bar{\boldsymbol{x}}_1)^T(\boldsymbol{M}_{11} + \boldsymbol{R})^{-1}(\boldsymbol{z} - \bar{\boldsymbol{x}}_1) + \frac{1}{2}\ln|\boldsymbol{M}_{11} + \boldsymbol{R}| - \frac{m}{2}\ln(2\pi) \tag{16}$$

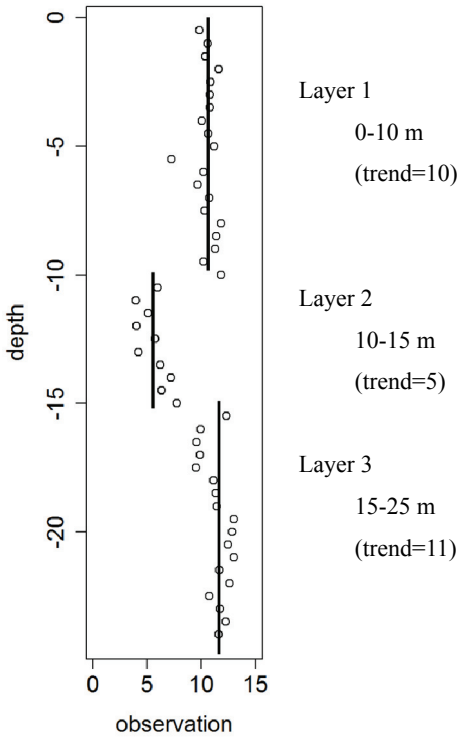


Figure 1. Assumed trend component and hypothetical observation data.

where n_h is the number of hyperparameters and m is number of observation data (size of $\boldsymbol{z}$). In this stratification formulation, hyperparameter n_h is the number of values of the trend and the location of the boundary

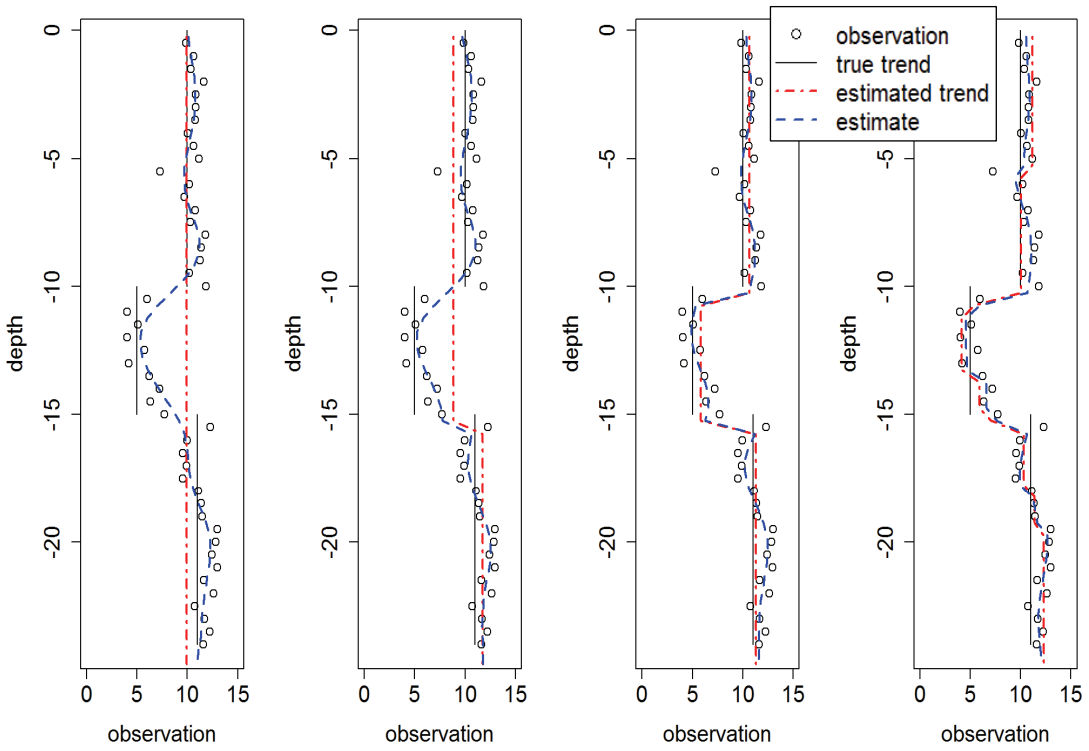
3. Numerical example of the proposed method

3.1 Hypothetical One-dimensional Random Field

The trend component of a one-dimensional random field is assumed, as shown in Fig. 1. There are three layers, whose trend component values are 10, 5, and 11 respectively. Autocovariance distance a and variation of random component in Eq. (5) are 2 and 1, respectively. The covariance matrix of observation error $\boldsymbol{R}$ is $\sigma_R^2\boldsymbol{I}$, where $\boldsymbol{I}$ is a identify matrix. Variation of observation error σ_R^2 is 1. The hypothetical observation, which is sum of trend component, random realization of random component, and observation error, is also shown in the figure. It is assumed that observation data are obtained at 0.5-m intervals.

3.2 Estimation of Trend Component with True Parameters

Fig.2 shows the trend component estimated by minimizing Eq. (9) with regularization parameter b = 0.1, 0.2, 0.3, and 0.8. The above mentioned two-step algorithm with ADMM was used for the minimization. The random component was also estimated using Eq. (3). They were calculated using the true values of variances of random component and observation error, and autocovariance distance. When b was 0.1, a uniform



(1) b=0.1 (2) b=0.2 (3) b=0.3 (4) b=0.8

Figure 2. Estimated trend and random component by two step algorithm with respect to sparsity parameter b.

trend component was obtained. When b was 0.2, two layers were obtained, and the boundary at 16 m was detected. When b was 0.3, all boundaries were correctly detected. When b was 0.8, there were many layers with small differences at the boundaries.

Fig.3 shows the relationship between AIC, BIC, -ln(L), and regularization parameter b. The true values of three parameters, namely the variances of the random component and observation error and autocovariance, are used to examine the information criteria. -ln(L) is the negative log-likelihood, which expresses the fitting between the trend component and observation data. -ln(L) decreases with increasing b. When b is large, the trend adapts to the random components and observation noise, which is called overlearning. In Fig. 3, the graphs of AIC and BIC are downward convex, and thus the information criteria have a minimum region, which is around b = 0.2-0.3. The estimated trend with b = 0.3 agrees with true trend component as shown in Fig. 2.

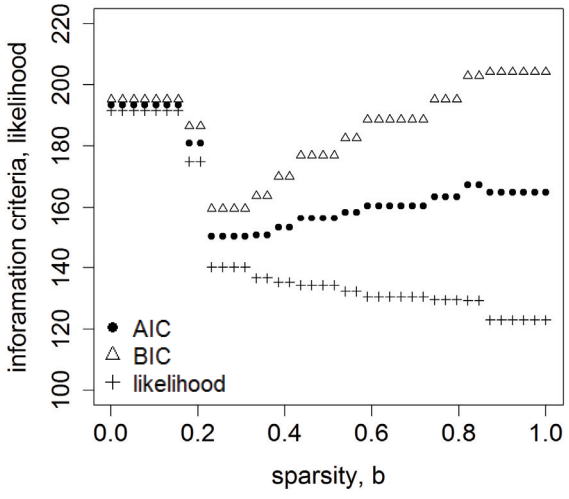


Figure 3. Information criteria AIC, BIC and sparsity parameter b.

3.3 Estimation of Trend Component without Information of True Parameters

The true values of three parameters, ~~namely~~ the variances of random component and observation error and autocovariance distance, were used in the previous section. When actual observation data are used, these three parameters in addition to regularization parameter b, should be determined simultaneously only from the observation data. This identification can be formulated as an optimization problem with respect to these four

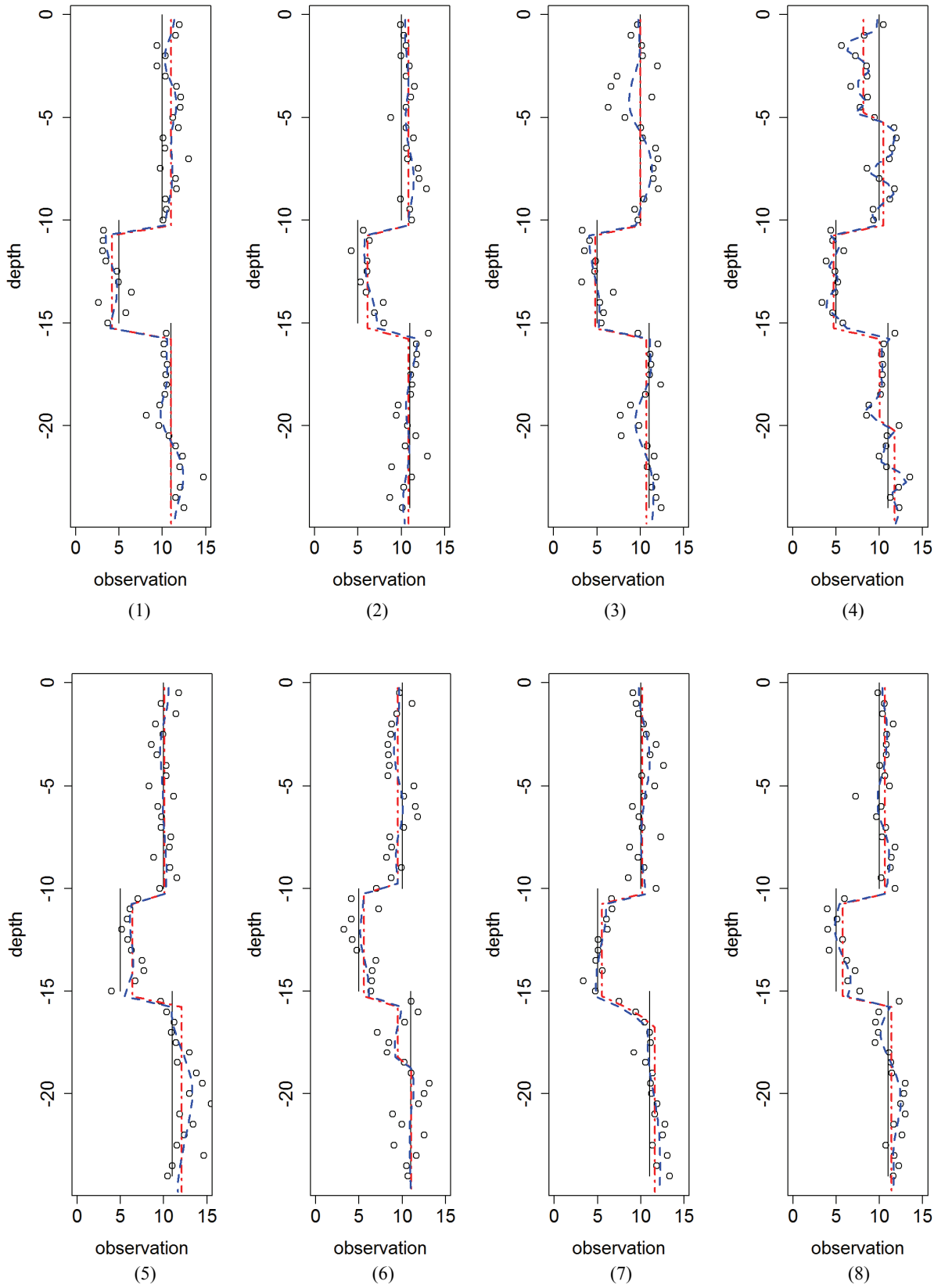


Figure 4. Estimated trend components for eight random realizations by optimized parameters.

explanatory variables, for which the objective function is the AIC or BIC. For the minimization of the objective function, a gradient-based method cannot be used because flat regions and discontinuities (the gradient is zero or infinity) with respect to parameter *b* appear in Fig. 3. Any global optimization method can be applied to the optimization problem with respect to these four variables. In this study, particle swarm optimization (PSO) is applied for optimization because of its ease of implementation and the small number of parameters required for tuning. PSO, a global optimization method proposed by Kennedy et al. (1995), is a population-based optimization technique originally inspired by the social behavior of flocking birds and schooling fish. In PSO, the potential solutions, called particles, move around in the problem space by following the current optimal particles, namely personal best and global best particles.

Eight random realizations are generated and used as the hypothetical observation data for the estimation of the trend component. For each random realization, the four parameters are estimated by minimizing the BIC with PSO, and the trend and random components are estimated by minimizing Eq. (9) with the two-step algorithm with ADMM. The estimated trend and random components are shown in Fig. 4. The estimated numbers of layer for these eight cases are 3, 3, 3, 5, 3, 4, 3 and 3. In six out of eight cases, three layers of trend component were correctly estimated. The best case seems to be (3), and its estimated parameters, namely autocovariance distance, standard deviation of the random component, and standard deviation of observation error are 1.73, 0.93 and 1.19 while their true values are 2, 1, and 1. The worst case is (4). Five layers are identified and its estimated parameters are 0.53, 1.0 and 0.36.

5. Conclusion

This study proposed a method for estimating the trend and random components by assuming sparsity in the spatial differences of soil parameter without assuming any basis functions and using kriging, respectively. Fused lasso (also called trend filtering) usually considers the random component to be uncorrelated and tries to remove it before estimating the trend. The proposed method can simultaneously estimate the trend component, correlated random component, and uncorrelated observation noise.

In this paper, only eight random realizations are used for the verification of the proposed method. The estimated stratification depends on the generated random realizations. More cases should be studied to obtain quantitative conclusion. Validation of the proposed method should be also performed by using actual observation data.

References

Akaike, H. 1974. A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19: 716 – 723.

Ching, J. and Phoon, KK. 2017. Characterizing Uncertain Site-Specific Trend Function by Sparse Bayesian Learning, *Journal of Engineering Mechanics*, ASCE,143(7).

Ching, J., and Phoon, K-K. 2019. Constructing Site-Specific Multivariate Probability Distribution Model Using Bayesian Machine Learning, *Journal of Engineering Mechanics*, ASCE, 145(1): 04018126. DOI: 10.1061/(ASCE)EM.1943-7889.0001537.

Christakos, G. (1992). *Random Field Models in Earth Sciences*, Academic Press.

Cressie, N. 1991. *Statistics for Spatial Data*, John Wiley & Sons.

Hastie, T., Tibshirani, R. and Wainwright, M. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman & Hall/CRC

Kennedy, J. and Eberhart, R. 1995. "Particle swarm optimization." *Proc. of IEEE Int. Conf. on Neural Networks*, Perth, Australia, 4, 1942-1948.

Kim, S., Koh, K., Boyd, S. and Gorinevsky, D. 2009. ℓ_1 trend filtering, *SIAM Review, problems and techniques section* 51(2), 339–36.

Meinshausen, N. 2007. Relaxed Lasso, *Computational Statistics & Data Analysis*, 52: 1, 15, 374-393.

Nishimura, S., Shibata, T. and Shuku, T. 2016. Diagnosis of earth-fill dams by synthesized approach of sounding and surface wave method, *Georisk*, 10: 4, pp. 312-319.

Papaioannou, I., Straub, D., 2017. Learning soil parameters and updating geotechnical reliability estimates under spatial variability. *Georisk* 11: 1, 116-128.

Rasmussen, C. E. and Williams, C. K. I. 2006. *Gaussian Processes for Machine Learning*. London, MIT Press.

Schwarz, G. 1978. Estimating the Dimension of a Model, *The Annals of Statistics*, 6, 461-464.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B, 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. 2005. Sparsity and smoothness via the fused Lasso, *Journal of the Royal Statistical Society*, Series B 67, 91–108.

Yoshida, I., Tasaki, Y., Otake, Y., and Wu, S. 2018. Optimal Sampling Placement in a Gaussian Random Field Based on Value of Information, ASCE-ASME *Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 4:3.