Distance Constrained Kriging Weights to Correct Large Weights to Data at the End of Strings

Olena Babak and Clayton V. Deutsch

Centre for Computational Geostatistics
Department of Civil & Environmental Engineering
University of Alberta

A characteristic feature of kriging is that end samples in strings of data receive large weights. This is theoretically valid; however, poor estimates and poor distributions of uncertainty arise when the end samples are unusually high or low. Many geological settings exhibit trends and such unusual grades at the contacts between geological domains. A number of ad-hoc corrections have been proposed, but none provide a constrained solution with a well defined measure of optimality. A new method for estimation in a finite domain is proposed. This method is referred to as DCK – Distance Constrained Kriging. The method combines the features of both Kriging and Inverse Distance estimation methods. Kriging weights to data in a string are constrained such that closer data to the location being estimated are given more weight. DCK was tested using two real data sets, one from mining and another from petroleum. In both cases, the proposed method outperforms conventional Simple and Ordinary Kriging in cross validation.

Introduction

Geostatistics has become a powerful tool in many areas of natural resources characterization. It is widely used to quantify uncertainty in energy and mineral resources such as natural gas [1], oil [2] and coal [3]. Other applications consist of generating input for flow simulation [4] and calculating the likelihood of exceeding critical threshold in contamination studies [5]. Geostatistical procedures rely on kriging for optimal estimation and to model local conditional distributions. Simulation is performed by drawing from such conditional distributions [6].

An implicit assumption of kriging is that the study area is embedded within an infinite domain. This causes kriging to give high weights to end samples in strings of data. This 'string effect' of kriging can cause serious problems. Strings of data are often observed in mining and petroleum applications where the data are collected along wells or drillholes. The artifact weighting of boundary samples can result in biased estimation, especially when the data exhibits strange trends with boundary/border effects.

A number of ad-hoc solutions have been proposed. The most common are to extend the string or wrap the string [7, 8]. These approaches attempt to fix the string effect either by changing the data configuration or the covariance function. They do not yield significant improvement in the results [7-9].

A new method for improved estimation of unsampled locations in finite domain is proposed in this paper. This method, referred to as Distance Constrained Kriging (DCK), modifies traditional Kriging approaches in that a distance weighting when estimating location of interest is considered. The weights are constrained by their distance to the location being estimated. The location in a string that are the closest to the location receives the largest weight, the second closest location receives the second largest weight and so on; the locations furthest from the location of interest receives the smallest weight.

The proposed new method has several important characteristics. Similar to Inverse Distance, the weights are ordered according to the distance. However, as in both Ordinary and Simple Kriging, the magnitude of the weights account for variogram based measures of spatial variability. The kriging weights are minimally corrected to enforce reasonable weights. For evaluation of the improvement of the proposed technique over both traditional Kriging techniques, two practical applications from mining and petroleum are considered. The advantages and characteristics of the method are discussed.

Maximum Influence of the Data in a String on the Finite Domain Estimation

This section illustrates the artificial weighting of boundary samples in string of data. The maximum influence of the data in the string $X_i, i = 1,\ldots,Ls,$ on the estimation location *u** was defined as maximum relative weight given by an estimation location to the data in the string:

$$i(u^*) = \max_{j=1,\cdots,Ls} \frac{\left|\lambda_j(u^*)\right|}{\sum_{k=1}^{Ls}\left|\lambda_k(u^*)\right|} \cdot 100\%,$$

where $\lambda(u^*) = (\lambda_1(u^*),\ldots,\lambda_{Ls}(u^*))^T$ denote the traditional (Simple or Ordinary) Kriging weights.

It is well known that the artificial weighting of boundary samples in the finite domain estimation causes more problems in Ordinary Kriging, than Simple Kriging. We believe that this is due to the implicit estimation of the mean and consequent constraint on the sum of the weights.

Distance Constrained Kriging Weights

In order to correct the structure of the maximum influence of end samples in a string when estimating a finite domain, we propose to estimate it using a linear estimator that minimizes estimation variance but constrains the weights to have certain influence structure. Specifically, the weights assigned by each estimation location to the string of data are ordered with respect to distance: the closest data in the string to the location of interest is constraint to receive the largest weight; the second closest data to the location of interest is constraint to receive the second largest weight and so on. In that way, the data in the string located furthest from the estimation location is assigned the smallest weight.

Let us consider *n* adjacent data $i = 1,\ldots,n,$ at locations $u_i, i = 1,\ldots,n,$ aligned in a string. Consider now the problem of estimating the value of a variable of interest X at an unsampled location *u** using the Distance Constrained Kriging approach. The usual Kriging technique is modified in that it weights the samples in the sting of data according to the distance from the estimation location to each data in a string. More precisely, the Distance Constrained Simple Kriging (DCSK) provides a model of the unsampled value $X(u^*)$ as the following linear combination of the data in a string $X_i = X(u_i), i = 1,\ldots,n,$ and the population mean *m*

$$X_{DCSK}^* = \sum_{i=1}^{n} \lambda_{DCSK,i} X_i + \left(1 - \sum_{i=1}^{n} \lambda_{DCSK,i}\right) m, \tag{1}$$

where $\lambda_{DCSK,i},$ denotes the DCSK weight of the *i*-th sample, $i = 1,\ldots,n,$ found by minimizing the estimation variance σ_{est}^2

$$\min_{\lambda} \sigma_{est}^2 = \sigma^2 - 2\sum_{i=1}^{n} \lambda_{DCSK,i} Cov(X_i, X_{DCSK}^*) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{DCSK,i}\lambda_{DCSK,j} Cov(X_i, X_j) \tag{2}$$

subject to

$$\lambda_{DCSK,i} > \lambda_{DCSK,j}, \text{ if } d_i < d_j, \text{ for each } i, j = 1,\ldots,n, \tag{3}$$

where σ^2 denotes the population variance; $Cov(X_i, X_j)$ and $Cov(X_i, X_{DCSK}^*)$, $i, j = 1,\ldots,n,$ denote the data-to-data covariance and the data-to-estimation point covariance, respectively; and d_i denotes the distance from the estimation location to the *i*-th data point in the string, $i = 1,\ldots,n.$

The Distance Constrained Ordinary Kriging estimates the value at the location of interest *u** as

$$X_{DCOK}^{*} = X(u^{*}) = \sum_{i=1}^{n} \lambda_{DCOK,i} X_i, \tag{4}$$

where $\lambda_{DCOK,i}$, denotes the DCOK weight of the *i*-th sample, $i = 1, \ldots, n,$ found by minimizing the estimation variance σ_{est}^2

$$\min_{\lambda} \sigma_{est}^2 = \sigma^2 - 2\sum_{i=1}^{n} \lambda_{DCOK,i} Cov(X_i, X_{DCOK}^{*}) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{DCOK,i}\lambda_{DCOK,j} Cov(X_i, X_j) \tag{5}$$

subject to

$$\lambda_{DCOK,i} > \lambda_{DCOK,j}, \quad \text{if} \quad d_i < d_j, \; i, j = 1, \ldots, n, \tag{6}$$

and

$$\sum_{i=1}^{n} \lambda_{DCOK,i} = 1, \tag{7}$$

where, as before, σ^2 denotes the population variance; $Cov(X_i, X_j)$ and $Cov(X_i, X_{DCOK}^{*})$, $i, j = 1, \ldots, n,$ denote the data-to-data covariance and the data-to-estimation point covariance, respectively; and d_i denotes the distance from the estimation location to the *i*-th data point in the string, $i = 1, \ldots, n.$

Properties of estimates produced by Finite Domain Kriging

As the 'usual' Kriging techniques, Distance Constrained Kriging has the following characteristics:

1 Distance Constrained Kriging estimator is unbiased.

2 The Distance Constrained Kriging estimator is an exact interpolator, that is, a conditioning data is reproduced at its exact location.

3 Distance Constrained Kriging approach provides a model for the unsampled value of the variable of interest according to its spatial continuity by the covariance function. Distance Constrained Kriging takes into account the redundancy of data in the string and closeness of the data in the sting to an estimation location.

Recall also that the Distance Constrained Kriging estimate is obtained as a linear combination of data in sting that minimizes estimation variance; however, in the Distance Constrained Kriging approach the estimation variance is minimized according to the distance constraint, thus, the estimate produced is characterized by the same or higher estimation variance, that is,

$$\sigma_{est}^{DCK} \geq \sigma_{est}^{K},$$

where σ_{est}^{K} and σ_{est}^{DCK} denote the estimation variance in the traditional Kriging and in the Distance Constrained Kriging.

Note also that Finite Domain Kriging approaches are also characterized by the following property:

4 The weights in the Finite Domain Kriging approach are assigned to data in string sorted according to the distance from these data to an estimation location.

Thus, the distance from the data to the estimation location has an affect on the resulting estimate; however, unlike the Inverse Distance technique, this affect is additionally corrected by the spatial continuity of the variable.

To compare the kriging weights obtained using the 'traditional' Kriging with the Distance Constrained Kriging approaches several small studies were performed. The weights were calculated for four estimation locations, (1, 7), (1.8, 7), (2.8, 7) and (3.8,7), based on the string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. Isotropic spherical variograms with a contribution of one and ranges of correlation 2 and 20 are considered for analysis. Results for the weights are shown in Figure 3 for comparison of the Ordinary Kriging and the Distance Constrained Ordinary Kriging data weighting and in Figure 4 for comparison of the Simple Kriging and the Distance Constrained Simple Kriging weighting. Note how the Distance Constrained Kriging approaches reduce the artificially higher weighs given to the end samples of the sting. Furthermore, also note that when the estimation location is located on the shortest distance to one of the two boundary samples in the string of data, then the Distance Constrained Simple Kriging results in the same estimate as the Simple Kriging.

Finite Domain Kriging: Generalization to the Case of Multiple Singles

Let us consider *K* strings of $n_k, k=1,\ldots,K$ data each at locations $u_i^k, k=1,\ldots,K, i=1,\ldots,n_k$. Then the value of the variable of interest X at an unsampled location *u** $X(u^*)$ in the Distance Constrained Simple Kriging approach is given by the following linear combination of the data in strings $X_i^k = X(u_i^k), k=1,\ldots,K, i=1,\ldots,n,$ and the population mean *m*

$$X_{DCSK}^* = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\lambda_{DCSK,i}^k X_i^k + \left(1-\sum_{k=1}^{K}\sum_{i=1}^{n_k}\lambda_{DCSK,i}^k\right)m, \tag{8}$$

where $\lambda_{DCSK,i}^k,$ denotes the DCSK weight of the *i*-th sample, $i=1,\ldots,n,$ in the *k*-th string, $k=1,\ldots,K,$ found by minimizing the estimation variance σ_{est}^2

$$\min_{\lambda}\ \sigma_{est}^2 = \sigma^2 - 2\sum_{k=1}^{K}\sum_{i=1}^{n_k}\lambda_{DCSK,i}^k Cov(X_i^k, X_{DCSK}^*) + \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{K}\sum_{j=1}^{n_k}\lambda_{DCSK,i}^k\lambda_{DCSK,j}^l Cov(X_i^k, X_j^l), \tag{9}$$

subject to

$$\lambda_{DCSK,i}^k > \lambda_{DCSK,j}^l, \text{ if } d_i^k < d_j^k, \text{ for each } k,l=1,\ldots,K, i=1,\ldots,n_k, j=1,\ldots,n_{l,} \tag{10}$$

where σ^2 denotes the population variance; $Cov(X_i^k, X_j^l)$ and $Cov(X_i^k, X_{DCSK}^*)$, $k,l=1,\ldots,K, i=1,\ldots,n_k, j=1,\ldots,n_{l,}$ denote the data-to-data covariance and the data-to-estimation point covariance, respectively; and d_i^k denotes the distance from the estimation location to the *i*-th data point in the *k*-th string, $k=1,\ldots,K, i=1,\ldots,n_k$.

The Distance Constrained Ordinary Kriging (DCOK) estimates the value at the location of interest *u** as

$$X_{DCOK}^* = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\lambda_{DCOK,i}^k X_i^k, \tag{11}$$

where $\lambda_{DCOK,i}^k,$ denotes the DCOK weight of the *i*-th sample, $i=1,\ldots,n,$ in the *k*-th string, $k=1,\ldots,K,$ found by minimizing the estimation variance σ_{est}^2

$$\min_{\lambda}\ \sigma_{est}^2 = \sigma^2 - 2\sum_{k=1}^{K}\sum_{i=1}^{n_k}\lambda_{DCOK,i}^k Cov(X_i^k, X_{DCOK}^*) + \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{K}\sum_{j=1}^{n_k}\lambda_{DCOK,i}^k\lambda_{DCOK,j}^l Cov(X_i^k, X_j^l), \tag{12}$$

subject to

$$\lambda^k_{DCOK,i} > \lambda^l_{DCOK,j}, \text{ if } d^k_i < d^k_j, \text{ for each } k,l = 1,\ldots,K,\ i = 1,\ldots,n_k,\ j = 1,\ldots,n_{l,} \quad (13)$$

$$\text{and} \quad \sum_{k=1}^{K}\sum_{i=1}^{n_k} \lambda^k_{DCOK,i} = 1, \quad (14)$$

where, as before, σ^2 denotes the population variance; $Cov(X^k_i, X^l_j)$ and $Cov(X^k_i, X_{DCOK}*)$, $k,l = 1,\ldots,K,\ i = 1,\ldots,n_k,\ j = 1,\ldots,n_{l,}$ denote the data-to-data covariance and the data-to-estimation point covariance, respectively; and d^k_i denotes the distance from the estimation location to the *i*-th data point in the *k*-th string, $k = 1,\ldots,K,\ i = 1,\ldots,n_k$.

Note that in the case of multiple string the four properties for the Finite Domain Kriging estimator outlined in the section 2.1 also hold.

Practical Applications

To assess the overperformance of the Finite Domain Kriging approaches over the traditional Simple and Ordinary Kriging, a case study of the two real data sets was conducted. One data set was chosen from a petroleum reservoir (data set 1) and one from a mineral deposit (data set 2). Both data sets contain the information from several vertical wells or drillholes. The results are in a paper available from the authors; however, the performance of the Finite Domain Kriging approaches and their traditional Kriging counterparts in cross validation was assessed based on the following summary measures:

1. Correlation between true values of the variable of interest and estimates produced by the modeling technique. An estimator that results in high correlation between the truth and estimate is desirable.
2. Mean of the error distribution. Errors are calculated as difference between true values of the variable of interest and estimates produced by the modeling technique. Errors should be centered on zero.
3. Standard deviation of the error distribution and the sum of absolute errors. The spread of errors around zero should be as narrow as possible.

We also considered a summary measure for comparing Distance Constrained Kriging approaches with their traditional counterparts. These measure the improvement of the Distance Constrained Simple Kriging over Simple Kriging and an improvement of the Distance Constrained Ordinary Kriging over Ordinary Kriging. These measures are defined as follows

$$\begin{aligned}\text{Impovement of DCSK over SK} &= \frac{(\text{Sum of abs vaslues of residuals of SK-Sum of abs vaslues of residuals of DCSK})}{\text{Sum of abs vaslues of residuals of SK}} \cdot 100\% \\ \text{Impovement of DCOK over OK} &= \frac{(\text{Sum of abs vaslues of residuals of OK-Sum of abs vaslues of residuals of DCOK})}{\text{Sum of abs vaslues of residuals of OK}} \cdot 100\%\end{aligned} \quad (15)$$

The improvement measures the scaled difference between the absolute residuals produced in traditional Kriging estimation and in respective Distance Constrained Kriging estimation. Note also that it follows from the above definition of the improvement measures that they can take on both positive and negative values. The positive values, of course, correspond to the fact that Distance Constrained Kriging performs better than the traditional Kriging counterpart. On the other hand, negative values imply that Distance Constrained Kriging performs worse with respect to estimation than tradition Kriging approach.

Based on cross validation results, we conclude that with respect to all cross validation measures, that is, correlation between truth and estimate, spread of errors and sum of absolute errors both Distance Constrained Kriging approaches outperform traditional Kriging approaches, while maintain virtually zero mean error distribution. Note also that improvements of Finite Domain Kriging approaches over their traditional counterparts are about 1.7% and 2% for mining and petroleum data sets, respectively.

With respect to jackknife we considered the same improvement measures as in (15). Jackknife was performed separately for each well with at least two observations. Then, based on the well improvement statistics, number of positive and negative improvements are assessed as well as an average improvement obtained over the study domain.

With respect to jackknife for petroleum data set, we observed that average improvement of Finite Domain Ordinary Kriging (FDOK) over OK is close to 5%. The average improvement of Finite Domain Simple Kriging (FDSK) over SK for this data set is more than 3%. For data set 2 the improvement was not as significant, it was about 1.5% for FDOK over OK and only about 0.6% for FDSK over SK.

Conclusions

A new approach for kriging in a finite domain using strings of data is proposed. This approach, referred to as Distance Constrained Kriging (DCK), combines characteristics of both Inverse Distance and Kriging. Similar to Inverse Distance it produces estimates that weight conditioning data based on the distance from the estimation location to the data in string, but as in traditional Kriging the estimates are optimal in a squared error sense accounting for the variogram model. Distance Constrained Kriging is an exact interpolator.

Distance Constrained Kriging is a linear unbiased estimator obtained by minimizing the estimation variance subject to distance constraints. With respect to the constraint on the sum of weights given to conditioning data, two flavors of Distance Constrained Kriging were considered: Simple and Ordinary Kriging.

The proposed approach to estimation of a finite domain using strings of data was tested using two real data sets, one data set from mining reservoir and one data set from petroleum deposit. Distance Constrained Kriging significantly reduce edge effect in estimation and perform slightly (but significantly) better for both cross validation and the jackknife.

References

Najjar, N.F., Jerome, T., and Alshammery, M., 2005. 2005 International Petroleum Technology Conference Proceedings: 1411

Ribeiro, M.T.; Hassan, S.R.; Gomes, J.S. and Bahamaish, J.N., 2004. 11th ADIPEC: Abu Dhabi International Petroleum Exhibition and Conference - Conference Proceedings, 2004: 973-982.

Kumral, M., and Dowd, P.A., 2005. A simulated annealing approach to mine production scheduling. Journal of the Operational Research Society, 56(8): 922-930.

Deutsch, C.V., 2002. Geostatistical Reservoir Modeling.

Kyriakidis P.C., and Journel, A.G., 2001. Stochastic modeling of atmospheric pollution: a spatial time-series framework. Part II: application to monitoring monthly sulfate deposition over Europe. Atmospheric Environment, 35(13): 2339-2348.

Deutsch C.V., and Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide.

Deutsch, C.V., 1994. Kriging with strings of data. Mathematical Geology 26(5): 623-638.

Deutsch, C.V., 1993. Kriging in a finite domain. Mathematical Geology 25(1): 41-52.

Saito H, McKenna SA, Zimmerman DA, et al. 2005. Geostatistical interpolation of object counts collected from multiple strip transects: Ordinary kriging versus finite domain kriging. Stochastic Environmental research and risk assessment, 19(1): 71-85.