Journal of Applied Geophysics 109 (2014) 266–280

Contents lists available at ScienceDirect

Journal of Applied Geophysics

journal homepage: www.elsevier.com/locate/jappgeo

Principal component analysis for filtering and leveling of geophysical data

Alexander Y. Davydenko [a], Alexander V. Grayver [b,*]

[a] Department of Geology and Geophysics, Irkutsk State University, Karl Marx Str., 664003 Irkutsk, Russia [1]
[b] Institute of Geophysics, ETH Zurich, Sonneggstrasse 5, 8092 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 8 April 2014
Accepted 16 August 2014
Available online 23 August 2014

Keywords:
Multivariate statistics
Principal component analysis
Moving window
Leveling
Trend removal

ABSTRACT

In this study, we investigate the use of multivariate statistical methods for geophysical data filtering. For this purpose, a measured scalar field is vectorized using a moving window technique and mean vector and covariance matrix are calculated by employing memory-efficient numerical algorithms. These multivariate statistics are then used to conduct principal component analysis (PCA). Namely, covariance matrix is decomposed into a set of eigenvalues and eigenvectors. By selecting a subset of eigenvectors, a PCA-based filter is realized. We demonstrate how properties of the filter are determined by the chosen subset of the eigenvectors, which in turn depend on spectrospatial properties of the field. In particular, we presented approaches to construct low-pass and spatial directional PCA-based filters. As an application, we aim at suppressing leveling errors commonly occurring in airborne data sets. The devised PCA filter was analyzed using a real aeromagnetic data set and synthetic leveling errors. The scenarios of statistically dependent and independent leveling errors were studied. Finally, we successfully applied it to real aero-electromagnetic data leveling.

1. Introduction

Multivariate statistical methods provide a large variety of techniques aimed at both analyzing and filtering data consisting of multiple variables (Härdle and Simar, 2007). These methods find numerous applications in geophysics (Reynolds, 2011). For instance, full-tensor gradiometry data can be viewed as a set of statistical variables and analyzed using multivariate statistical methods. These methods, however, cannot be directly applied to scalar fields such as those obtained from aero-magnetic or aero-electromagnetic surveys. Nevertheless, by using a moving window approach (also known as sliding window or rolling window), data can be represented as a set of statistical variables and thus become suitable for multivariate statistical methods. One of such methods is principal component analysis (PCA), which has found many applications in geophysics (Chung and Nigam, 1999; Egbert and Booker, 1986; Guo et al., 2009; Minsley et al., 2012; Smirnov and Egbert, 2012). Here, we apply the combination of the moving window approach and PCA to analyze structure of the field and suppress leveling errors in airborne geophysical data.

Leveling errors typically occur during collection of geophysical data along profile lines over large areas and result in a random, possibly spatially correlated, displacement of measurements from neighboring profile lines. This creates a characteristic washboard or strip pattern perpendicular to the survey profile direction. The effect has long been recognized when working with aero-magnetic (Yarger et al., 1978) or aero-electromagnetic (Valleau, 2000) data sets. Leveling errors can be caused by diurnal Earth's magnetic field drift (Minty, 1991), temperature variations in measuring devices (Siemon, 2009), flight direction changes or calibration errors (Huang, 2008) and eventually hinder the interpretation and can cause severe artifacts when inverting data.

Several methods exist to remove or suppress this type of errors. Some approaches are based on FFT filtering (Ferraccioli et al., 1998; Minty, 1991; Nelson, 1994; Siemon, 2009), whereas other techniques make use of different fitting methods such as least squares, polynomials or B-splines (Beiki et al., 2010; Mauring et al., 2002; Yarger et al., 1978). Many of the methods depend on so called tie-lines — a few lines measured perpendicular to the main survey profiles. Differences at the cross points between the original profiles and the tie-lines are then used to quantify leveling errors. If no tie-lines are available, some authors suggest choosing a reference line that is presumably free of leveling errors (Huang, 2008) or applying robust median filters using a moving window (Mauring and Kihle, 2005). Alternatively, one could build a directional filter using discrete wavelet transform (Fedi and Quarta, 1998; Paoletti et al., 2007). Few researchers have addressed the problem of suppressing leveling errors by means of multivariate statistical methods.

We present a method to suppress leveling errors that is based on the multivariate statistical analysis. The method does not require tie-lines and is readily valid for survey areas with arbitrary contours, gaps inside,

* Corresponding author. Tel.: +41 44 633 3154.
E-mail addresses: davydenkoay@gmail.com (A.Y. Davydenko), agrayver@erdw.ethz.ch (A.V. Grayver).

[1] Formerly Irkutsk State Technical University, Lermontov Str., Irkutsk, Russia.

http://dx.doi.org/10.1016/j.jappgeo.2014.08.006
0926-9851/

and quasi-regular observation layouts. We start with the definition of the multivariate statistics in a moving window, derive a PCA-based filter, and discuss the problem of the dependent noise in data and consequences from the statistical point of view. We then shortly describe differential polynomial fitting (DPF) approach (Beiki et al., 2010) used to compare efficiency of the PCA filter. Finally, we test the devised methodology on the real magnetic field with synthetic leveling errors, compare its efficiency with DPF and apply it to the real aero-electromagnetic data from Siberia.

2. Theory and method

2.1. Data in moving window

Before applying multivariate statistical methods, measured fields need to be vectorized and represented as a set of random variables. To this end, we introduce the moving window approach. Denote $\mathbf{d} \in \mathbb{R}^{N_d}$ a real vector consisting of N_d data measurements. Every measurement d^k is assigned a pair of coordinates $(x, y)_k$. Data are measured and arranged in N_p quasi-parallel profile lines. For clarity, we will assume that the lines are directed eastward. If this condition is not fulfilled, one can rotate profiles to point eastward. Fig. 1 shows a schematic example of a geophysical survey layout.

We define a rectangular window of size $n \times m$ points centered at a measurement point d^k, $k = 1 \ldots N_d$ with n and m being odd numbers. Fig. 1 illustrates an example of the rectangular window of size 5×3 centered at the k-th measurement point. A subset of points situated in the moving window at any position is called a window sampling. By cascading all the window rows, we can write down the measurement points in a window sampling as a vector of size $N = n \times m$. In other words, for a rectangular window of size $n \times m$ points centered at a measurement point k, we construct a window sampling vector $\mathbf{w}^k \in \mathbb{R}^N$ as

$$\begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m,1} & d_{m,2} & \cdots & d_{m,n} \end{pmatrix} \rightarrow \mathbf{w}^k = \begin{pmatrix} d_{1,1} \\ d_{1,2} \\ \cdots \\ d_{1,n} \\ d_{2,1} \\ \cdots \\ d_{m,1} \\ d_{m,2} \\ \cdots \\ d_{m,n} \end{pmatrix}. \tag{1}$$

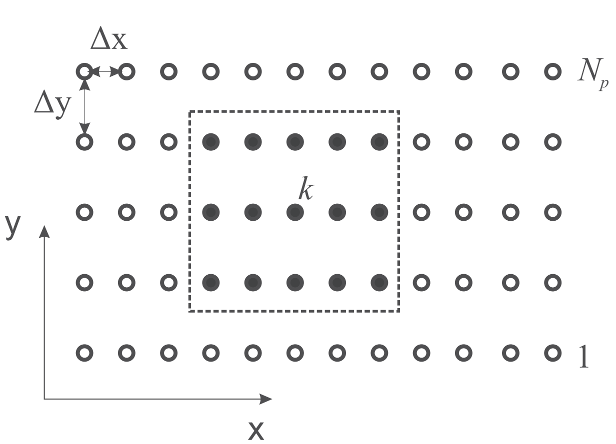

Fig. 1. Sketch of a survey layout. Black points indicate measurement positions grouped into N_p profile lines. The dashed line depicts a rectangular window of size 5×3 centered at the measurement point k.

Every window sampling vector $\mathbf{w}^k$ is a point in the N-dimensional space. Each dimension of such space is formed by the respective element of the moving window. Fig. 2a shows a synthetic signal $\mathbf{d}$ consisting of 13 data values. Choosing a moving window of size 3×1 yields eleven window samplings, excluding positions of the window at the first and last measurements where one data point inside the window is missing. Fig. 2b shows the window sampling vectors plotted as points in the three-dimensional space. The points colored in blue and green are the values which form window sampling vectors $\mathbf{w} \in \mathbb{R}^3$ at the 3rd and 11th positions. This simplified example serves to show that the moving window approach is in fact a mapping of the original scalar signal into some higher dimensional vector space. Such space captures the behavior of the signal in some vicinity of each measurement and facilitates application of various multivariate statistical methods.

Details on extending this approach to quasi-regular survey layouts are given in Appendix A.

2.2. Multivariate statistics in moving window

Adopting the moving window approach introduced in the previous section enables us to consider a scalar geophysical field $\mathbf{d}$ as a vector field, i.e. at any position there exists vector $\mathbf{w} \in \mathbb{R}^N$ as defined by Eq. (1). Vector $\mathbf{w}$ is a multivariate random variable that is characterized by its probability density function $f(\mathbf{w})$ with respective expected value $\boldsymbol{\mu}_w$ and covariance matrix $\mathbf{C}_w$. These quantities can be written following Tabachnick et al. (2001) using

$$\mu_w = E(\mathbf{w}) = \int_{\infty}^{\infty} \mathbf{w} f(\mathbf{w}) d\mathbf{w} \tag{2a}$$

$$\mathbf{C}_w = E\left(\mathbf{w}\mathbf{w}^T\right) - \boldsymbol{\mu}_w \boldsymbol{\mu}_w^T. \tag{2b}$$

In practice, we want to get discrete estimations of the multivariate statistics presented in Eqs. (2a) and (2b). Therefore, we first define a moving window with N points and construct sampling vectors for all N_d points in the data set $\mathbf{d}$. These vectors can be combined into a matrix $\mathbf{F}$ of size $N_d \times N$ where the kth row is given by $(\mathbf{w}^k)^T$. Following the approach adopted in this work, every column of this matrix is a random variable, whereas the whole matrix is a statistical sampling with the uniquely defined mean vector and covariance matrix. Calculating mean values for all columns and arranging them in a vector yield a sample mean vector that approximates $\boldsymbol{\mu}_w$ from Eq. (2a). More precisely, this can be written as

$$\hat{\boldsymbol{\mu}}_w = \frac{1}{N_d} 1^T \mathbf{F}, \tag{3}$$

where $\mathbf{1}^T$ is a row vector with all elements equal one such that $\mathbf{1}^T\mathbf{F}$ is a vector which contains sums of the columns of $\mathbf{F}$. Accordingly, the covariance matrix is then given by (cf. Eq. (2b))

$$\hat{\mathbf{C}}_w = \frac{1}{N_d} \mathbf{F}^T \mathbf{F} - \hat{\boldsymbol{\mu}}_w \hat{\boldsymbol{\mu}}_w^T. \tag{4}$$

As was mentioned in the previous section, a set of window sampling vectors is a point cloud in the N-dimensional space characterized by its discrete joint probability distribution. The sample mean vector in Eq. (3) gives a point in the space that minimizes the sum of Euclidean distances from all points to it

$$\hat{\boldsymbol{\mu}}_w = \underset{\mathbf{v} \in \mathbb{R}^N}{\arg min} \; ||\mathbf{v} - \mathbf{w}^k||_2, \tag{5}$$

whereas sample covariance matrix provides the estimation of linear dependence between all pairs of random variables (or, equivalently,