



Least squares collocation with uncorrelated heterogeneous noise estimated by restricted maximum likelihood

Wojciech Jarmołowski

Received: 21 May 2014 / Accepted: 27 February 2015 / Published online: 3 April 2015
© The Author(s) 2015. This article is published with open access at Springerlink.com

Abstract The article describes the estimation of a priori error associated with heterogeneous, non-correlated noise within one dataset. The errors are estimated by restricted maximum likelihood (REML). The solution is composed of a cross-validation technique named leave-one-out (LOO) and REML estimation of a priori noise for different groups obtained by LOO. A numerical test is the main part of this case study and it presents two options. In the first one, the whole data is split into two subsets using LOO, by finding potentially outlying data. Then a priori errors are estimated in groups for the better and worse subset, where the latter includes the mentioned outlying data. The second option was to select data from the neighborhood of each point and estimate two a priori errors by REML, one for the selected point and one for the surrounding group of data. Both ideas have been validated with the use of LOO performed only in points of the better subset from the first kind of test. The use of homogeneous noise in the two example sets leads to LOO standard deviations equal 1.83 and 1.54 mGal, respectively. The group estimation generates only small improvement at the level of 0.1 mGal, which can also be reached after the removal of worse points. The pointwise REML solution, however, provides LOO standard deviations that are at least 20 % smaller than statistics from the homogeneous noise application.

Keywords LSC · REML · Regularization · A priori noise

1 Introduction

Least squares collocation (LSC) assumes spatial correlation of the investigated quantity. The correlation may be determined for the signal present in the data, as well as for the noise. Even if the noise is correlated, it can represent different unwanted factors, the influence of which has to be excluded from the solution. The noise, however, can be also assumed as non-correlated, which is fairly useful if its correlation is not evident or hard to assess. Uncorrelated noise is an example of quite common noise appearing in some data kinds and has been assumed also in this article. The investigations are focused on the variance of uncorrelated noise, which is heterogeneous. The restricted maximum likelihood (REML) estimation is found to be worth of implementation to investigate this heterogeneity in details. This technique has been successfully used in the estimation of various parameters contributing to the covariance matrices. This article uses REML in the analysis of the individual noise variances distributed along the diagonal in the noise covariance matrix.

The theory of LSC often starts from elementary case, where the data coincides exactly with the model (Moritz 1980). Nevertheless, it is not possible to measure a quantity with no measurement error and therefore the variables representing error variances must be taken into account in the solution. In addition to the signal covariance matrix representing variances and covariances of spatially correlated field, the noise covariance matrix of uncorrelated errors occurs as a diagonal matrix. The diagonal has to hold correct point noise variances to provide the best results of the prediction. However, there are many examples in the literature, where a regularization is applied to this diagonal and its dimension is empirically corrected (Rummel et al. 1979; Rapp and Wang 1994; Marchenko et al. 2003). The mentioned examples of LSC regularization can be seen as related or even based

W. Jarmołowski (✉)

Department of Satellite Geodesy and Navigation, Faculty of Geodesy, Geospatial and Civil Engineering, University of Warmia and Mazury in Olsztyn, ul. Oczapowskiego 2, Olsztyn 10-719, Poland
e-mail: wojciech.jarmolowski@uwm.edu.pl

on Tikhonov regularization (Moritz 1980; Koch and Kusche 2002; Eshagh and Sjöberg 2011). Rummel et al. (1979) and Kotsakis (2007) indicate D. L. Philips, as also related with this technique. The LSC regularization techniques mentioned above can have a physical justification related with some specific physical properties of the modeled spatial field. Klees et al. (2004) provide some suspected spatial factors affecting the covariance matrices of the signal and noise. Sabaka et al. (2010) analyse the regularization in time domain besides the spatial problems. This is applied in the case of time correlation for repeating observations of the same phenomena. This work starts from an assumption that although numerically the manipulation of diagonal noise covariance matrix can be comparable to the regularizations mentioned above, in fact, it is unnecessary when the errors are correctly estimated. These errors cannot be, however, based on the survey errors and the resolution has to be taken into account together with some other factors.

The issue of noise covariance level present in the data has been discussed many times, e.g. in Arabelos et al. (2007), Sadiq et al. (2010) and Filmer et al. (2013). The ratio between signal and noise is investigated for altimetry data by Hwang and Parsons (1995). They analyse noise size for different satellites and use these estimates for the weighting purposes in LSC. Arabelos and Tscherning (1998) also analyse an influence of different noise values in altimetry data on the prediction of the gravity. Among different questions analysed together with the noise covariance problems, the spatial resolution is a frequently discussed factor (Rummel et al. 1979; Eshagh and Sjöberg 2011). The example of significant influence of the resolution and a priori noise on the LSC results can be found in Lee et al. (2013). The spacing along the satellite track is used for the estimation of the standard error by Paolo and Molina (2010). Filmer et al. (2013) also point the relation between a priori error and data spacing. They also see a dependency of errors on the terrain, which can be partially associated with covariance models used, especially if height correlation is not taken into account by these models. Therefore, Bouguer gravity anomalies are applied in the numerical test here, as they are free of height dependence to the useful level. A closely related problem to a priori noise in data is filtering in different kinds of data processing, which eliminates some data, which adversely affects the result. The properties of the filter are usually selected on the basis of the information about the noise level and data resolution (Andersen and Knudsen 1998; Hwang et al. 2007) but can be also assessed with the use of some physical properties of the signal source (Strykowski 2000). The filters are frequently necessary to solve the problems of the correlated noise present in e.g. satellite gradiometry data (Schuh 2003; Reguzzoni and Tselfes 2009). The presented research focuses on the random part of the observations, considered as white noise.

This work attempts to apply a non-homogeneous a priori noise of the data in LSC, in line with the belief that large errors in the data should not be necessarily removed and treated as gross-errors. The alternative often used in data processing, especially in large datasets, is the removal of outlying observations based on the specified threshold. A handy set of techniques and possible thresholds, as well as some classification of the outliers can be found in Kern et al. (2005). The result of the removal of suspected outliers is also shown in this article, however, it is worth investigating if the errors come really from an evident mistake or they are rather observational in nature. A significant number of the data with larger observational errors can justify a need to include them in LSC to keep worse interpolation results rather than empty places. The question is then how to estimate these noise values and what is the influence of noise values on LSC results? LSC can be processed with one dataset; however, the variety of the terrestrial and still growing number of the satellite data sources, often requires combining of the data sets with different noise characteristics. The variance component estimation (VCE) is popular in assessing the existing noise variances for different data that contribute to the combined solution (Kotsakis and Sideris 1999). The influence of particular datasets is then diversified, depending on different noise variances. The examples are an application of minimum norm quadratic unbiased estimation (MINQUE) of variance component for different groups contributing to the geoid height values (Fotopoulos 2005) and maximum likelihood (ML) estimation of two groups of distance observations in the horizontal geodetic network (Groddecki 1999). Another practical example investigates the signal variance and the noise variance as the variance components (Yang et al. 2009). These variances are resolved by ML method to find their ratio. This work is focused on the estimation of a priori noise variance alone by REML. The subsets of different a priori noise variances existing within one type of the data can be extracted e.g. by the separation of some data forming groups of the supposed, different noise variance levels. This split can be done empirically by cross-validation (CV) based on the spatial correlation (Tscherning 1991). The split into the groups of different noise is of course approximate; however, significant differences in noise variance level may be worth investigating also with the use of such a simple approach. In a more detailed approach, the individual a priori noise variances can be estimated point by point with the separation of a single point only.

There is no proof that one dataset has uniform noise variance over some data region as there are at least two factors that can add to this heterogeneity. The most obvious are different measurement errors, which may sometimes be especially large, e.g. due to the extremely hard observational conditions. These errors are sometimes assessed as outliers in some data analyses, but in fact, they are poor

observations. It is also suspected that the relation between the measurement accuracy, local high-frequency signal variance and the spatial data resolution can exert some influence on the noise covariance matrix. The worst option in processing data by LSC is to keep outlying observations together with the better data and apply a homogeneous noise variance. All the regions close to the outliers are then affected in interpolation process, which is shown in numerical examples. Two schemes of REML application are proposed below to assess noise variances in groups or individually. The threshold can be set to keep outliers in a separate group of point measurements. The influence of the outliers may be eliminated by the removal of the points or equivalently by assigning them a sufficiently large value of a priori error. The estimation of prediction errors can be then used to assess the advantage of one or the other way in terms of errors distribution and the obtained resolution of interpolated spatial field. The crucial motivation in the development of efficient noise assessment techniques can be a permanently increasing number of different data types. Satellite, terrestrial, marine and airborne data compose spatio-temporal databases describing the same phenomena. Therefore, a combination of the data to obtain optimal estimates can require simultaneous, empirical noise assessment. The combination of CV and REML techniques presented in the article provides some answers in statistics related to a priori noise and prediction errors, which can be interesting especially when data are sparse, noisy or come from many sources.

2 Assigning noise by restricted maximum likelihood (REML)

The known LSC equation for the interpolation of gravity values in the space domain reads (Moritz 1980):

$$\tilde{\Delta \mathbf{g}}^r = \mathbf{C}_p^T \cdot (\mathbf{C}_s + \mathbf{C}_n)^{-1} \cdot \Delta \mathbf{g}^r, \quad (1)$$

where $\Delta \mathbf{g}^r$ is the residual data vector. The matrix \mathbf{C}_s is the covariance matrix of the residuals, \mathbf{C}_p is the covariance matrix between predicted residuals and data residuals and \mathbf{C}_n represents the noise covariance matrix. The objective of the numerical tests is to split the observations into groups of different noise variance. An arbitrary number of the groups can be used, however various drawbacks have to be considered, e.g. the importance of results improvement or computational efficiency. In the proposed numerical tests, which will be described in details in the next section, we have always two groups: the better subset and the worse subset. Thus, since we consider uncorrelated noise, \mathbf{C}_n becomes the block-diagonal matrix:

$$\mathbf{C}_n = \begin{bmatrix} \delta \mathbf{n}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \delta \mathbf{n}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \delta \mathbf{n}_i \end{bmatrix}, \quad \delta \mathbf{n}_i = \delta n_i^2 \cdot \mathbf{I}_m \quad (2)$$

In Eq. (1), $\delta \mathbf{n}_i$ are diagonal matrices with squared standard deviations of uncorrelated, homogeneous noise of m data points in the subsets. In this article, different noise variance values are grouped in the vector of parameters, i.e. $\boldsymbol{\theta} = \{\vartheta_i\} = \{\delta n_i\}$ when we split the data into i groups. The further investigations use single noise standard deviation δn or 2×1 vector $\boldsymbol{\theta} = \{\delta n_1, \delta n_2\}$. The better and worse subsets use the parameters δn_1 and δn_2 , respectively. Therefore, the covariance matrix $\mathbf{C}(\boldsymbol{\theta})$ will be consequently equal $\mathbf{C}_s + \mathbf{C}_n$. The noise is assumed to be non-correlated. The signal is spatially correlated and \mathbf{C}_s , as well as \mathbf{C}_p , are generated using Gauss-Markov third order (GM3) planar model (Moritz 1978).

$$\text{GM3}(\mathbf{C}_0, CL, s) = \mathbf{C}_0 \left(1 + \frac{s}{CL} + \frac{s^2}{3 \cdot CL^2} \right) \cdot \exp\left(\frac{-s}{CL}\right), \quad (3)$$

where the spherical distance denoted as s is used instead of planar one and CL is also in spherical distance units. The selection of the covariance model is partially based on the frequent use of GM3 model in the literature for geoid (Kavzoglu and Saka 2005) or gravity interpolation (Moreaux 2008). The use of this and other planar models is very popular in local gravity modeling by LSC and in the analyses of the covariance parameters (Camacho et al. 1997; Kotsakis 2007). These models can approximate the empirical covariance very well in local areas, where the removal of long-wavelength signal part can efficiently eliminate remote covariance terms. The planar models can well represent predominant, close covariance terms, which are most significant in LSC. The remaining, small values of long-wavelength correlation are systematic and are not suspected as contributing to the white noise. Since the analysis is focused on $\delta \mathbf{n}$ parameter, which forms a diagonal matrix that does not use GM3 model, the GM3 was assumed as useful in the test, after the efficient removal of long-term signal trend described later.

LSC by Eq. (1) is quite common in regular gridding of gravity anomalies. It can be used to predict regular grid in the data area, but also to predict sparse points. The latter option can be used in the CV process, which has various known forms. One of them is leave-one-out (LOO) validation, which removes one point from the dataset, used for the prediction in the position of the removed one (Eq. 4). The differences between n data values and the predictions made in the same positions are often used as a measure of the prediction precision. The difference in point p used in LOO validation reads:

$$\begin{aligned}
 & \text{LOO}_p \left((\boldsymbol{\theta}) \mid \left(\Delta \mathbf{g}_{n \times 1}^r, \Delta \mathbf{g}_{n \times 1}^r \right) \right) \\
 &= \left\{ \begin{array}{l} \left| \Delta g_p^r - \tilde{\Delta g}_p^r \right| \\ \tilde{\Delta g}_p^r = \mathbf{C}_{\mathbf{P}(n-1) \times 1}^T \cdot \left(\mathbf{C}_{\mathbf{S}(n-1) \times (n-1)} + \mathbf{C}_{n(n-1) \times (n-1)} \right)^{-1} \cdot \Delta \mathbf{g}_{(n-1) \times 1}^r \\ \wedge \\ \Delta g_p^r \notin \Delta \mathbf{g}_{(n-1) \times 1}^r \end{array} \right\}. \tag{4}
 \end{aligned}$$

LSC assumes that the mean is known and equals zero, i.e. we process residual data, which are spatially correlated. The residuals can be always produced by removing a long-wavelength signal part. This is usually done using global harmonic expansion of the physical quantity. The global harmonic expansion should represent sufficiently large degree to remove long-wavelength part. The maximum degree and order of the removed global harmonic expansion of gravity has to be equivalent to a rather smaller area size than that of the selected data. The choice of an adequate degree and order is necessary to have a Gaussian distribution of the residuals and possibly few residues of long-wavelength signal.

REML estimation of the covariance parameters assumes that pure ML can provide biased estimates due to non-zero mean in the data. Therefore, REML applies the orthogonal projection of the data based on the spatial distribution and although this test uses the residuals from the global model subtraction, a minimalistic first order trend is additionally applied in REML to follow the rule. The first order trend reads:

$$\mathbf{X} = \begin{bmatrix} 1 & \varphi_1 & \lambda_1 \\ \dots & \dots & \dots \\ 1 & \varphi_n & \lambda_n \end{bmatrix}, \tag{5}$$

The implementation of the trend can be done by the matrix \mathbf{R} , which has the relation with the so-called projection matrix \mathbf{P} and reads (Koch 2007; van Loon 2008):

$$\begin{aligned}
 \mathbf{R}(\boldsymbol{\theta}) &= \mathbf{C}(\boldsymbol{\theta})^{-1} \cdot \mathbf{P} \\
 &= \mathbf{C}(\boldsymbol{\theta})^{-1} \cdot \left\{ \mathbf{I} - \mathbf{X} \left[\mathbf{X}^T \mathbf{C}(\boldsymbol{\theta})^{-1} \mathbf{X} \right]^{-1} \mathbf{X}^T \mathbf{C}(\boldsymbol{\theta})^{-1} \right\}. \tag{6}
 \end{aligned}$$

\mathbf{R} is then used in the probability density function to separate the vector of coordinates and obtain independency of the marginal likelihood function of spatial data distribution, i.e.:

$$\begin{aligned}
 p(\Delta \mathbf{g}^r, \boldsymbol{\theta}) &= |\mathbf{C}(\boldsymbol{\theta})|^{-\frac{1}{2}} \left| \mathbf{X}^T \mathbf{C}(\boldsymbol{\theta})^{-1} \mathbf{X} \right|^{-\frac{1}{2}} \\
 &\times \exp \left[-\frac{1}{2} \Delta \mathbf{g}^{rT} \mathbf{R}(\boldsymbol{\theta}) \Delta \mathbf{g}^r \right]. \tag{7}
 \end{aligned}$$

In the case of unknown vector of the parameters $\boldsymbol{\theta}$, the process should be iterative. The use of approximate parameters may be sometimes helpful and sufficient to replace iterations. This

experiment shows values of the negative log-likelihood function (NLLF) for different combinations of at most two parameters in the vector $\boldsymbol{\theta}$. The NLLF in our case reads:

$$\begin{aligned}
 \text{NLLF}(\Delta \mathbf{g}^r, \boldsymbol{\theta}) &= \frac{1}{2} \ln |\mathbf{C}(\boldsymbol{\theta})| + \frac{1}{2} \ln \left| \mathbf{X}^T \mathbf{C}(\boldsymbol{\theta})^{-1} \mathbf{X} \right| \\
 &+ \frac{1}{2} \left[\Delta \mathbf{g}^{rT} \mathbf{R}(\boldsymbol{\theta}) \Delta \mathbf{g}^r \right]. \tag{8}
 \end{aligned}$$

The minima of NLLF, which are a base for the parameters choice, can be observed graphically in Fig. 3. These results are also calculated using the scoring process (Grodecki 1999; van Loon 2008). This method is often called Fisher scoring and is iterative. The vector of the parameters is iterated using the following algorithm:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \mathbf{S}^{-1}(\boldsymbol{\theta}_k) \cdot \mathbf{d}_k(\boldsymbol{\theta}_k), \quad k \in \{1, 2 \dots 10\}. \tag{9}$$

\mathbf{S} is the Fisher information matrix, which is positive definite and is produced using first derivatives of the covariance matrix $\mathbf{C}(\boldsymbol{\theta})$:

$$\begin{aligned}
 \mathbf{S} &= \begin{bmatrix} S_{1,1} & S_{1,2} & \dots & S_{1,j} \\ S_{2,1} & S_{2,2} & \dots & S_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ S_{i,1} & S_{i,2} & S_{i,3} & S_{i,j} \end{bmatrix}, \quad S_{i,j} = \text{tr}(\mathbf{R}\mathbf{C}_i\mathbf{R}\mathbf{C}_j), \quad i, j = \delta n_i. \tag{10}
 \end{aligned}$$

The vector of the scores $\mathbf{d}(\boldsymbol{\theta}_k) = \mathbf{t}(\boldsymbol{\theta}_k) - \mathbf{u}(\boldsymbol{\theta}_k)$ and

$$\mathbf{u} = \{u_i\}, \quad u_i = \Delta \mathbf{g}^{rT} \cdot \mathbf{R}\mathbf{C}_i\mathbf{R} \cdot \Delta \mathbf{g}^r. \tag{11}$$

$$\mathbf{t} = \{t_i\}, \quad t_i = \text{tr}(\mathbf{R}\mathbf{C}_i). \tag{12}$$

Derivatives of the covariance matrix are computed with respect to the covariance parameters, i.e.:

$$\mathbf{C}_i = \frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \vartheta_i}. \tag{13}$$

If we can accept C_0 and CL derived from the fitting of the covariance model and fix them here, only the noise standard deviation parameter is analysed in the numerical tests. Consequently, differentiation in Eq. (13) is fairly easy. The signal part of $\mathbf{C}(\boldsymbol{\theta})$ becomes constant and the variable part is block diagonal \mathbf{C}_n added to \mathbf{C}_s . Such a noise covariance matrix can hold only uncorrelated part of the data.

3 Initial data assessment

Gravity data come from the U.S. gravity database available at the website of University of Texas at El Paso (Hildenbrand et al. 2002; Keller et al. 2006). Two sets of gravity data are selected for the testing. The first is located in Wyoming and the second in New Mexico. The first dataset consists

of 492 point Bouguer gravity anomalies with spherical distances between the points around 0.05° , whereas the second one stores 608 points of the same data with similar resolution. The topography structures of both selected areas are similar. The Wyoming sample is an area of moderate roughness at altitude between around 1050 and 1400 m with respect to WGS84. Only southwestern corner of the selected region includes a mount slope that is 2800 m high. The topography of New Mexico sample is mainly between 870 and 1400 m with one mount slope up to 2000 m, also in the south-west. The U.S. gravity database consists of large number of surveys from different time epochs and therefore the accuracy varies, depending on the survey techniques used. The data used in this analysis is extracted geographically from the composed database and no accuracy estimate is provided in the files. The codes indicating the survey campaigns are only available. The information about the measurement errors is therefore mixed and difficult to assess with no access to the National Geodetic Survey (NGS) reports. The choice of the testing areas was approximately based on the gravity data analysis of the U.S. gravity base that can be found in [Saleh et al. \(2013\)](#). They provide a detailed investigation of gravity errors by the nearest neighbor and crossover analysis, which may be substantially useful for data processing by LSC. This work follows the empirical approach and assumes that a priori noise variance depends not only on the survey error, but also on the data resolution. The criterion was to find some larger errors in the datasets to assign different a priori errors amongst one dataset. The data are often regarded as having homogeneous noise variance by many, especially if one technique is applied in the survey over a small area. However, the analyses like in [Saleh et al. \(2013\)](#) prove heterogeneity of the noise. The specific question in data analysis is the problem of outliers. The outlying observation can be a gross-error that has its value distorted by the factor different from typical observational conditions. This factor, in gravity dataset, may be e.g. the coordinate that is false due to severe conditions in GNSS positioning ([Bakuła 2012](#)). The data with noise significantly larger than average are often regarded as gross-errors and removed. In practice, large datasets may contain a considerable number of observations with significant observational errors, which cause problems in the simultaneous processing. These problems occur in LSC, which needs weighting of the data or removal of outlying values. The question to pose is whether different noise parameters for different groups can provide more accuracy than using homogeneous noise or removing of the outliers? Are the observations with a larger noise indeed removable and what will be the accuracy in the data gaps? Finally, what is the advantage or disadvantage of using individual noise values for the points? These questions are investigated in the paper with the specific use of REML technique.

In case of large datasets, the manual search of the outlying observations can be inefficient. Different techniques are used to automatize the process, including the very popular LSC ([Tscherning 1991](#); [Vergos et al. 2005](#)). The rule is to compare the prediction with originally measured value in the same positions. A special task is to find the threshold value, which is responsible for the removal of points. The same technique is used in this paper in LOO validation, performed to split the data into groups of different noise levels. This means that the residual observation used for comparison does not apply in point LSC prediction ([Kohavi 1995](#)). The data used in LOO was previously detrended using global harmonic expansion of geopotential. The long-wavelength Bouguer gravity part was generated from the EGM2008 to degree and order 360. The distribution of the residuals is shown in [Fig. 2c, d](#). The residuals of the better subset were then used in the estimation of ECF by averaging the products of point values separated by similar distance. The analytical Gauss-Markov model ([Eq. 3](#)) is subsequently fitted into ECF ([Fig. 2a, b](#)) by the manual manipulation of C_0 and CL parameters and graphical assessment. C_0 and CL are then fixed to the values from covariance function fitting and δn is set roughly to 0.5 mGal, which is afterwards found to be not so far from the average noise here. The radius of selection of the points used in every interpolation was set to $4CL$ and typical LOO validation has been done for all data points. The subsequent step is different than removal, however the removal option is also made for comparison in a separate calculation. Rather than being removed, the outlying values are first stored in another file, which is called worse subset and obtains different a priori error than the remaining data in the better subset. The threshold was set to 6 mGal for both datasets and this choice is in some sense an example with no mathematical rule associated with it. On the other hand, smaller outliers are quite frequent in the validation, especially in Wyoming, and therefore seen as typical rather than outlying. The threshold for both datasets is intentionally equal, in the belief that this can assure a more comparable analysis. As a result, the Wyoming data have been split into 474 better points and 18 worse points, whereas New Mexico data have been split into 601 better points and 7 worse points. [Figure 1](#) shows datasets and worse data as rounded integer LOO differences by [Eq. \(4\)](#).

Some assumptions are made prior to the REML estimation of noise variances. If the noise variance is several times smaller than the signal variance, the inexactness between the true signal variance and C_0 estimated from the data variance affects a priori errors to a relatively small degree. Many authors estimate a combination or ratio between noise variance and signal variance ([Pardo-Igúzquiza 1998](#); [Camacho et al. 1997](#); [Yang et al. 2009](#)). This kind of ratio is not applied here, but fitting of the functions is assumed as quite a good approximation of C_0 and CL . This is also based on the previous studies, where CL estimated by the fitting is very similar

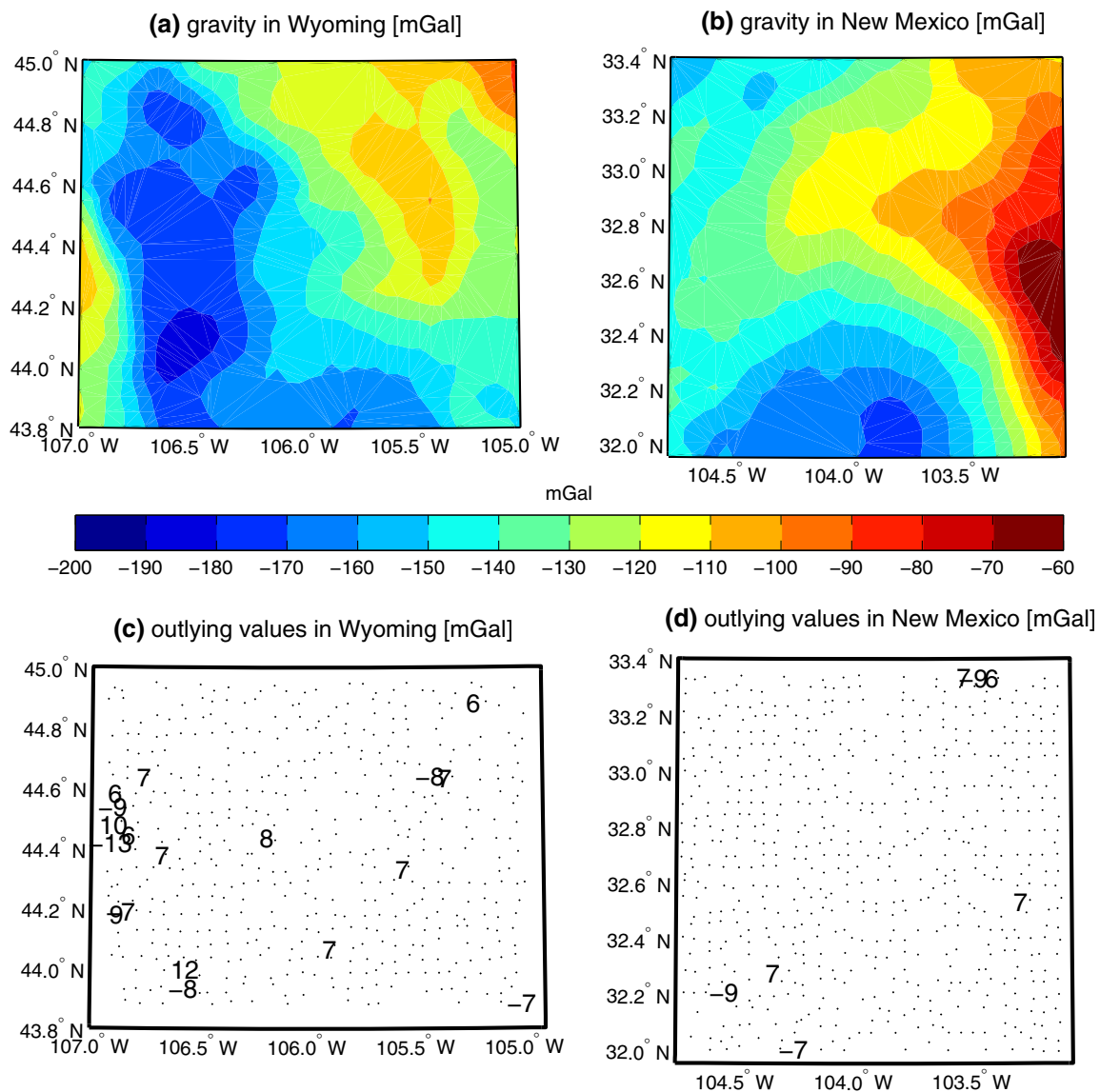


Fig. 1 Bouguer gravity anomalies and scheme of split with better subsets represented by dots and worse subsets by integer, rounded LOO differences (Eq. 4)

to that from CV results (Jarmołowski 2013). Even if this is not true at all, the examples prove that small changes of these parameters can affect the estimation of δn by REML only negligibly (Jarmołowski and Bakula 2014). Fixing of C_0 and CL enables a more clear view on δn and makes REML process more effective in the estimation of δn . This parameter is most difficult to estimate by the covariance function fitting and therefore needs an application of REML.

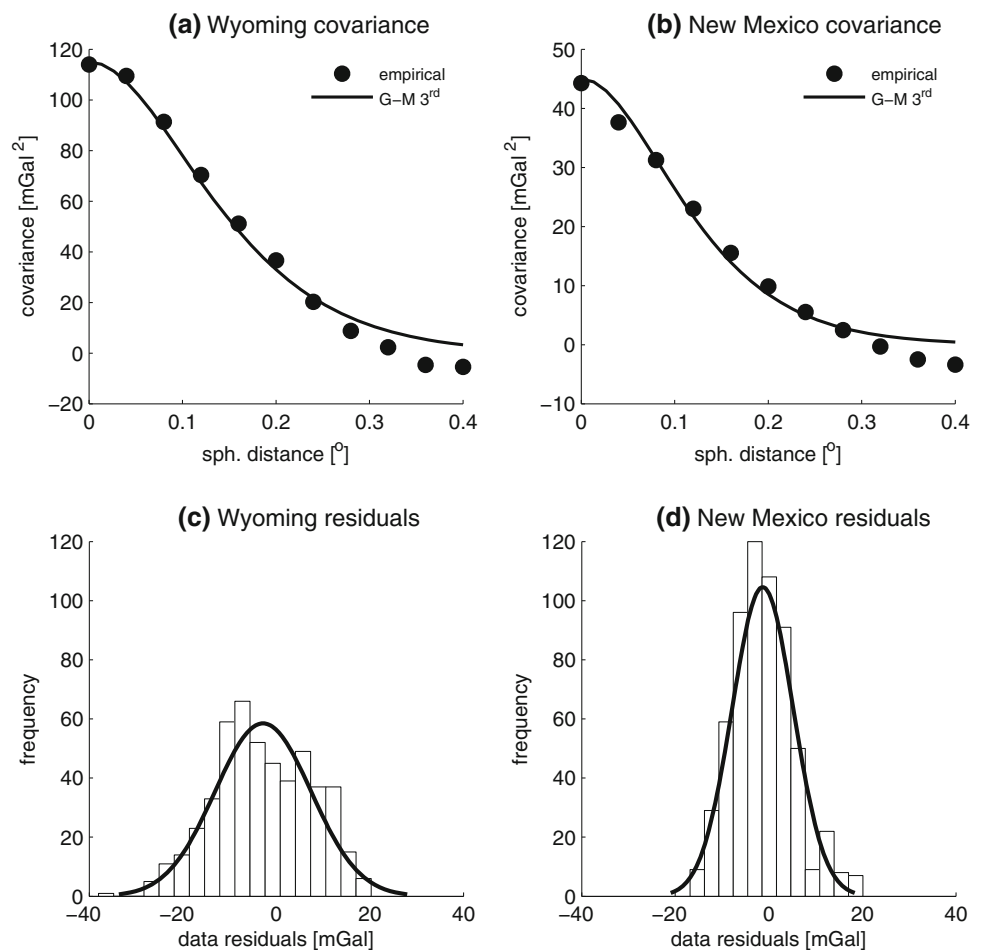
C_0 and CL estimated by the model fitting into the values of ECF are respectively, 115 mGal^2 and 0.060° for Wyoming and 45 mGal^2 and 0.050° for New Mexico (Fig. 2a, b). These values are then fixed in REML and LOO processes, but we have to remember that the gravity field can be only locally assumed as stationary and the covariance parameters are valid only locally. LOO validation uses a limited distance range for

the points used in point computation. It is four times larger than CL and this choice is based on the covariance function shapes. This distance assures the selection of the points within the area of positive covariance, with the assumption that small negative values are systematic in nature and therefore do not affect non-correlated δn . This rule is also used for the point REML estimation. δn_1 is there applied for one calculated point and δn_2 for the subset limited by $4CL$.

4 Numerical experiment

The presented numerical test uses the fitting of the analytical planar covariance model (Eq. 3) into the empirical covariance values (Fig. 2a, b). I decided to fix C_0 at the begin-

Fig. 2 Empirical covariance functions of residuals, fitted GM3 function (a, b) and histograms of residuals (c, d)



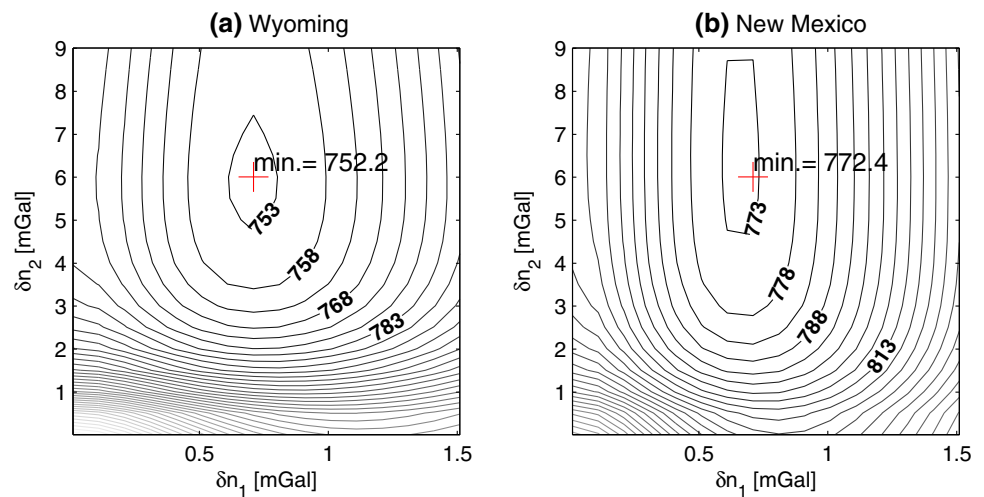
ning, since the covariance matrix based on Eqs. (1) and (3) implies a correlation between signal and noise variances (Jarmołowski and Bakuła 2014). The fitting of the analytical covariance model into empirical covariance function (ECF), estimates CL parameters that are close to those found by CV when C_0 approximates the signal variance (Jarmołowski 2013). Thus, the two mentioned parameters can be assessed as proper at some level of accuracy. This is assumed to be sufficient to estimate the noise variance with the error several times smaller than C_0 error, especially if δn is much smaller than C_0 .

NLLF values described in Eq. (8) are dependent on $\mathbf{C}(\boldsymbol{\theta})$ and therefore also on the parameters in the vector $\boldsymbol{\theta}$. C_0 and CL are fixed, as their accuracy from the fitting (Fig. 2a, b) is assessed as sufficient to estimate reliable values of δn . The validity of this assumption is confirmed later by the results of REML and LOO estimation. In the first numerical test the vector $\boldsymbol{\theta}$ comprises of two average standard deviations of a priori noise representing two subsets of each dataset: better and worse. These standard deviations are δn_1 and δn_2 respectively. In the second numerical test the split of the data is not considered. Individual δn values are estimated in the

following way. For each point, the closest data is employed in REML estimation, using $4CL$ radius of the selection. The same parameters δn_1 and δn_2 indicate noise standard deviation of one point from the whole dataset and respective quantity for the group of remaining points, which are located closer than $4CL$ to the one selected. The shape of $\boldsymbol{\theta}$ vector is the same as in the first test, however, δn_1 is reserved for the investigated point only, i.e. $m = 1$ for the subset (Eq. 2). The remaining points obtain the parameter δn_2 and the same process, as in group estimation, is repeated n times for a limited number of the observations that is different each time. The subset of m processed points obtains then group δn_2 value of the noise standard deviation, whereas one central point is distinguished and has an individual value δn_1 . δn_1 is large if the point value is outlying. The parameter δn_2 is variable and it may have different values, some of which may also be large if poor observations are present in the subset. Finally, δn_2 is neglected in the further investigation and δn_1 values represent point noise variances.

The process of scoring (Eq. 9) in the presented form has been extensively discussed in the literature and some problems with solvability have been found (Kubik 1970; Pardo-

Fig. 3 NLLF values for variable noise standard deviations of better (δn_1) and worse (δn_2) subsets



Igúzquiza 1997; Grodecki 1999). Therefore, a decision was taken to control the estimation of the parameters by scoring from the beginning. The mentioned problems are numerically illustrated in the previous work (Jarmołowski and Bakula 2014), however, this example presents successful scoring, consistent with NLLF values. Nevertheless, the improvement of scoring convergence should be investigated in the future works. There are many examples of more advanced or robust techniques e.g. in Smyth (2002) and Kusche (2003). To control the scoring results here, NLLF has been calculated for variable δn_1 and δn_2 after fixing C_0 and CL . The empirically selected range of the search is visible in Fig. 3, where 0.1 mGal is the step for δn_1 and 0.5 mGal for δn_2 . The global minimum of NLLF indicates a smaller δn_1 for better subsets and several times larger δn_2 for the worse subsets (Fig. 3).

The noise standard deviations δn_1 and δn_2 , which can be graphically assessed from Fig. 3, are also estimated by the scoring. The parameters obtained for Wyoming are following: $\delta n_1 = 0.71$ mGal and $\delta n_2 = 5.85$ mGal, and for New Mexico: $\delta n_1 = 0.66$ mGal and $\delta n_2 = 6.24$ mGal. These results are quite consistent with Fig. 3, which confirms the correctness of the scoring. The iterations are shown in Figs. 4b and 5b. It should be noted here that scoring for some additional data samples was ineffective via this simple approach.

The main objective of the current test is to prove the usefulness of REML estimation of a priori noise for different groups of observations (Figs. 4, 5), as well as for the point approximations of a priori noise in the second test (Fig. 6). In the first test, the scoring has been additionally used with no split of the data, i.e. for the whole datasets. In this case, a priori noise standard deviation is denoted just as δn and one value is determined for each area (Figs. 4a, 5a). This value is subsequently used in the repeated LOO validation test (Figs. 4c, 5c), as well as two values found for better and worse subsets (Figs. 4d, 5d). All of the repeated LOO tests have been performed only in the positions of the better subsets

to eliminate the data with significant noise from the comparisons. This assures that the measured values are close to the true, correlated gravity field and enables more accurate validation. More precisely, the predictions are made in the positions of the better subset using the whole set. The distance limit for the data used in the point prediction is $4CL$, because positive covariance values reach approximately this distance (Fig. 2a, b). LOO values for the homogeneous noise are shown as scatter plot in Figs. 4c and 5c. Figures 4d and 5d present LOO values in the case of different a priori standard deviations estimated for split data. All scatter plots have the areas of the circles increasing exponentially, to magnify LOO validation results. The areas have been drawn using absolute values of LOO differences in the power, i.e. $3^{|\text{LOO}|}$. These values have been empirically found as most effective graphically. This idea reveals 0.1 mGal of differences in standard deviation between different solutions, which can be observed in pairs of figures, i.e. Figs. 4c, d and 5c, d.

The mean and the standard deviation of LOO differences for Wyoming data are improved for split data (Fig. 4d) in relation to respective statistics when the homogeneous a priori noise is used (Fig. 4c). The places with a significant concentration of worse points, after assigning $\delta n_2 = 5.85$ mGal, obtain better LOO values in the better points. No improvement is observed in the places that are away from the worse points. Sometimes LOO results are even worse than for the homogeneous noise in these places. This proves that larger $\delta n = 1.25$ mGal was a better choice for these places (Fig. 4c) than smaller one, i.e. $\delta n_1 = 0.71$ mGal (Fig. 4d). Such worse LOO differences can result from the larger actual noise in the mentioned places and may suggest more detailed split of the data, e.g. into three subsets.

Although New Mexico data have fewer worse points after applying the threshold equal to 6 mGal, the standard deviation of repeated LOO differences has also decreased a little. The places of the worse data have lower LOO values in the

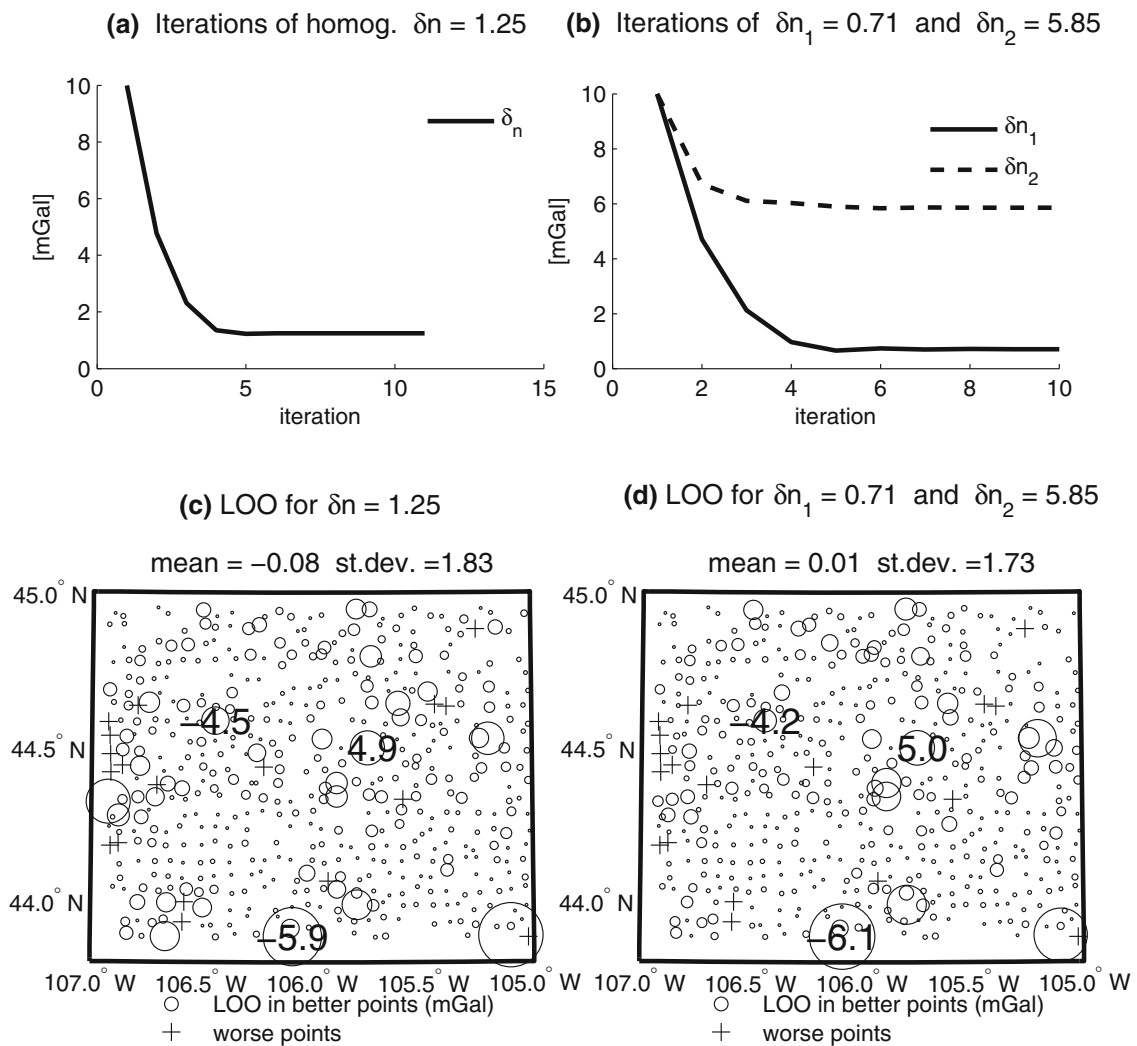


Fig. 4 Scoring iterations and repeated LOO validation of the estimated a priori standard deviations for Wyoming data. The crosses indicate worse points. The areas of circles increase exponentially. A few numbers are used to represent the scale

better points in Fig. 5d than in Fig. 5c. LOO values in the better points that are far from worse subset show no bigger differences between the two solutions. This is due to the small differences between homogeneous noise $\delta n = 0.85$ mGal and that for better subset after split, i.e. $\delta n_1 = 0.66$ mGal. To provide a more detailed answer on the usefulness of REML estimation in groups of the data, a typical remove procedure has been also applied and compared to the mentioned results. The Fisher scoring has been applied to the better subsets alone and δn has been estimated as practically the same values as in group estimation. These values have been checked in LOO validation performed after the removal of the worse data and the statistics were: mean 0.01 mGal and standard deviation 1.73 mGal in Wyoming; mean 0.00 mGal and standard deviation 1.45 mGal in New Mexico. These statistics in comparison with Figs. 4d and 5d prove no practical difference between group estimates used and outliers removal at

the same threshold of split and removal. The group estimation by REML can assign a priori errors to outlying data that are able to marginalize their influence to the same extent as outliers removal. Therefore, REML can be considered as useful tool for such purposes, especially if fast scoring technique is applied.

The split solution and outliers removal show only small improvement in relation to the homogeneous noise assumption and therefore the decision was taken to perform the second test. REML estimation has been made n times for all the points in Wyoming and New Mexico. This test is performed with the use of NLLF minimum search only, since the robustness of the scoring is not improved in this work. The ranges of both parameters start from 0.1 mGal and then follow from 0.5 mGal, with step 0.5 to 8 mGal for δn_2 and from 1 mGal, with step 1 to 12 mGal for δn_1 . Minimum NLLF indicating optimal values of both parameters is individually found

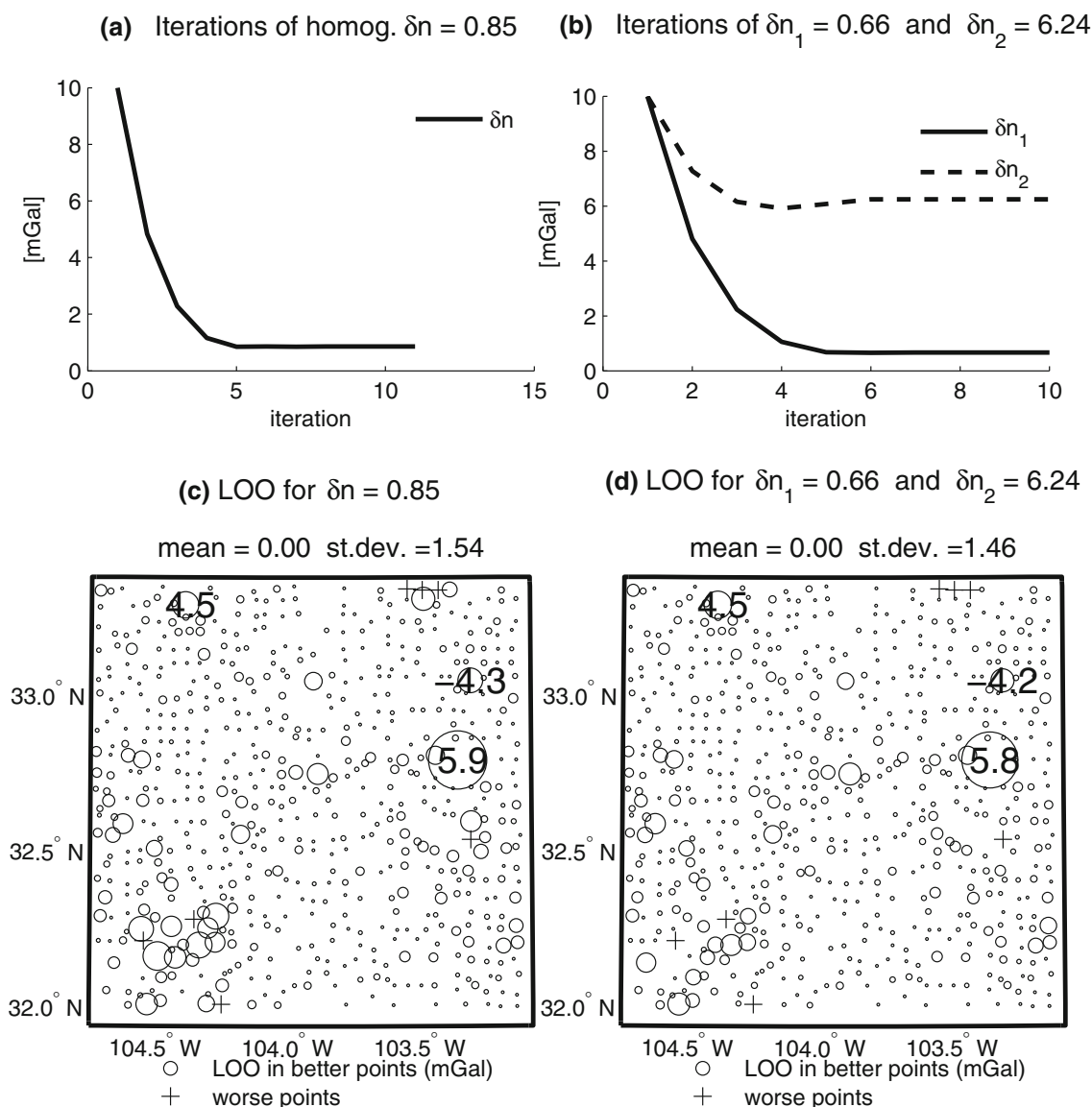


Fig. 5 Scoring iterations and repeated LOO validation of the estimated a priori standard deviations for New Mexico data. The crosses indicate worse points. The areas of circles increase exponentially. A few numbers are used to represent the scale

for every point. Every n repeat finds δn_1 for the analysed point p and δn_2 for the group of the remaining points in the selected neighborhood of p . The values of the estimated δn_2 are then omitted and not presented in Fig. 6. Point values of δn_1 are shown in Fig. 6a, b. The numbers show the estimated a priori noise exceeding or equal 2 mGal only to keep some clarity of the figure. It is immediately noticeable that many values exceed average a priori noise estimated for the better subset in the first test. The use of individual a priori noise values in LOO validation decreases standard deviations of the differences computed consequently in the positions of the better points only (Fig. 6c, d). This improvement is evident in comparison to the improvement from the group estimation and provides LOO standard deviations equal 1.44 mGal in

Wyoming and 1.22 mGal in New Mexico. These values are about 20 % smaller in relation to respective standard deviations from the use of homogeneous noise.

The scoring process has the main advantage in terms of computational time in REML noise estimation. However, it was not implemented in the pointwise processing in the frame of this work, because it is more complicated and needs some more numerical trials. The estimation with the use of varying parameters is not convenient and more time-consuming than scoring. However, the same estimation using LOO validation will take even more time, because it needs varying parameters in LSC computation of every point and checking the total statistics of LOO differences. The REML group estimation for Wyoming using typical “i5” processor has taken 103 s,

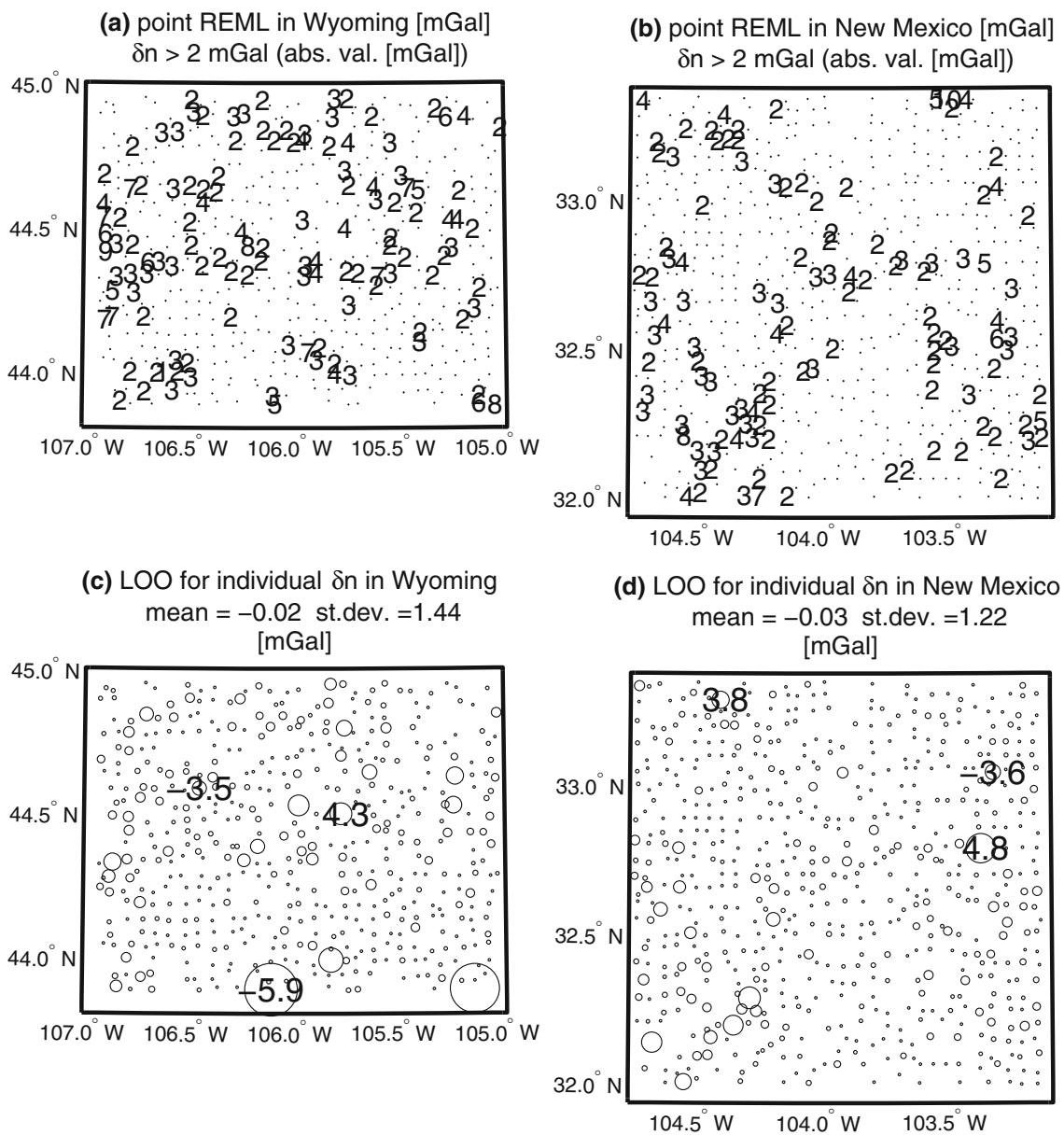


Fig. 6 Pointwise REML estimates of a priori error greater than 2 mGal in **a** Wyoming, and **b** New Mexico and LOO validation of LSC with point error values in **c** Wyoming and **d** New Mexico. The areas of *circles* and *numbers* are applied in the same way as in Figs. 4, 5

whereas New Mexico estimation has been done in 175 s. The same estimations by the scoring have been processed in 4 and 7 s, respectively. Therefore, the investigation of the scoring in point REML estimation is especially worth implementing. The pointwise estimation with the use of NLLF, which was most successful in results have been performed in 315 s for Wyoming and 242 s for New Mexico. Therefore, it can be expected that future application of the scoring in the pointwise estimation will give proportionally better time improvement, as it is found for group estimation. The time consumption and also REML effectiveness is dependent on the size of covariance matrices. It should be pointed that more

extensive test areas with a larger number of the data need to handle the data in a way that limits the size of the covariance matrix. To summarize, the time efficiency of the processes are strongly related to the size of the covariance matrices and the number of repetitions of single estimates.

All four approaches to δn in LSC i.e.: use of homogeneous noise (Figs. 4c, 5c), group estimation of noise by REML (Figs. 4d, 5d), the removal of outliers with the application of the same threshold as for groups and point estimation by REML (Fig. 6c, d) are validated in repeated LOO process on better points, together with a posteriori error estimates by the well-known formula (Moritz 1980, p. 105, Eq. 14–42).

Table 1 LSC error estimates from four approaches to δn in Wyoming (mGal)

	Min.	Max.	Mean
Homogeneous δn	0.97	5.08	1.61
Better and worse	0.71	4.85	1.37
Worse removed	0.74	4.88	1.40
Individual δn	0.48	5.13	1.36

Table 2 LSC error estimates from four approaches to δn in New Mexico (mGal)

	Min.	Max.	Mean
Homogeneous δn	0.78	3.46	1.32
Better and worse	0.67	3.39	1.24
Worse removed	0.74	3.43	1.29
Individual δn	0.40	3.22	1.21

Since LOO validations are pointwise and no grid creation has been done, the general statistics have also been calculated in sparse points, which have quite good density and coverage of the selected region. The error estimates for all four options of δn are given in Tables 1 and 2.

The statistics show the largest errors when homogeneous δn is applied with no outliers investigation. The split into two groups shows a slight advantage in relation to outliers removal, but this is not confirmed in LOO standard deviations described before in this section. Therefore, the small differences between 1.40 and 1.37 mGal in Wyoming and between 1.29 and 1.24 mGal in New Mexico can indicate only very local improvement in the resolution of the points with very limited influence on regional statistics. The advantage of point REML estimation of δn is mostly visible in the minima of the above statistics. Point δn values allow to reach the best accuracy estimates that are not available by means of any other approach.

5 Conclusions

The initial LOO validations made for the selected data samples found many observations with the errors several times larger than the average error value. Their removal can slightly limit estimated accuracy of LSC solution, which is confirmed in slightly worse error statistics. However, no significant difference in terms of repeated LOO standard deviation can be observed between processes after the removal of outliers and group error assignment. This means that group noise estimation by REML is practically equivalent to the removal of outliers in this case. Therefore, REML appears to be a helpful tool in the empirical search of the threshold for outliers and an estimator of sufficiently large noise for outliers, which

marginalizes their influence on the result. It should be pointed that REML estimates of errors are generally consistent with the results in Saleh et al. (2013).

An alternative and better option is to use individual noise variances in C_n matrix. Therefore, REML algorithm is additionally applied pointwise and individual δn are estimated for every point by the search of $NLLF$ minimum. Repeated LOO showed that the estimation of δn by REML in points can reduce the standard deviation of the differences a few times more than REML with data split or outliers removal. The pointwise REML estimation provides noticeable 20 % improvement in terms of LOO standard deviation; thus, the optimization of scoring is worth considering to obtain fast algorithm for this purpose. The detailed analysis of a priori noise by an empirical application of REML introduces some new methodology in the noise investigation, which can be presumably applied to the cases of the correlated noise in further research.

Acknowledgments I am grateful to the editors and reviewers for many valuable comments and suggestions. Gravity data originate from Gravity Database of the U.S. and were downloaded from the website of the University of Texas at El Paso (<http://research.utep.edu>).

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Andersen OB, Knudsen P (1998) Global marine gravity field from the ERS-1 and Geosat geodetic mission altimetry. *J Geophys Res* 103(C4):8129–8137
- Arabelos D, Forsberg R, Tscherning CC (2007) On the a priori estimation of collocation error covariance functions: a feasibility study. *Geophys J Int* 170:527–533
- Arabelos D, Tscherning CC (1998) The use of least squares collocation method in global gravity field modeling. *Phys Chem Earth* 23(1):1–12
- Bakuła M (2012) An approach to reliable rapid static GNSS surveying. *Surv Rev* 44(327):265–271
- Camacho AG, Montesinos FG, Vieira R (1997) A three-dimensional gravity inversion applied to São Miguel Island (Azores). *J Geophys Res* 102(B4):7717–7730
- Eshagh M, Sjöberg LE (2011) Determination of gravity anomaly at sea level from inversion of satellite gravity gradiometric data. *J Geodyn* 51:366–377
- Filmer MS, Hirt C, Featherstone W (2013) Error sources and data limitations for the prediction of surface gravity: a case study using benchmarks. *Stud Geophys Geod* 57(1):47–66
- Fotopoulos G (2005) Calibration of geoid error models via a combined adjustment of ellipsoidal, orthometric and gravimetric geoid height data. *J Geod* 79:111–123
- Grodecki J (1999) Generalized maximum-likelihood estimation of variance components with inverted gamma prior. *J Geod* 73:367–374
- Hildenbrand TG, Briesacher A, Flanagan G, Hinze WJ, Hittelman AM, Keller GR, Kucks RP, Plouff D, Roest W, Seeley J, Smith DA, Webring M (2002) Rationale and operational plan to upgrade the

- U.S. Gravity Database: U.S. Geological Survey Open-File Report 02 463
- Hwang C, Parsons B (1995) Gravity anomalies derived from Seasat, Geosat, ERS-1 and TOPEX/POSEIDON altimetry and ship gravity: a case-study over the Reykjanes Ridge. *Geophys J Int* 122:511–568
- Hwang C, Hsiao Y-S, Shih H-C, Yang M, Chen K-H, Forsberg R, Olesen AV (2007) Geodetic and geophysical results from a Taiwan airborne gravity survey: data reduction and accuracy assessment. *J Geophys Res* 112:B04407
- Jarmołowski W (2013) A priori noise and regularization in least squares collocation of gravity anomalies. *Geod Cartogr* 62(2):199–216
- Jarmołowski W, Bakula M (2014) Precise estimation of covariance parameters in least-squares collocation by restricted maximum likelihood. *Stud Geophys Geod* 58:171–189
- Kavzoglu T, Saka MH (2005) Modeling local GPS/levelling geoid undulations using artificial neural networks. *J Geod* 78:520–527. doi:10.1007/s00190-004-0420-3
- Keller GR, Hildenbrand TG, Kucks R, Webring M, Briesacher A, Rujawitz K, Hittleman AM, Roman DR, Winester D, Aldouri R, Seeley J, Rasillo J, Torres R, Hinze WJ, Gates A, Kreinovich V, Salayandia L (2006) A community effort to construct a gravity database for the United States and an associated web portal. In: Sinha AK (ed) *Geoinformatics: data to knowledge*, vol. 397. Geological Society of America, Special Paper, pp 21–34
- Kern M, Preimesberger T, Allesch M, Pail R, Bouman J, Koop R (2005) Outlier detection algorithms and their performance in GOCE gravity field processing. *J Geod* 78(9):509–519
- Koch KR, Kusche J (2002) Regularization of geopotential determination from satellite data by variance components. *J Geod* 76:259–268
- Koch KR (2007) *Introduction to Bayesian statistics*, 2nd edn. Springer, Berlin
- Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Melish CS (ed) IJCAI-95: Proceedings of the fourteenth international joint conference on artificial intelligence*, vol 2. Morgan Kaufmann, pp 1137–1143. <http://ijcai.org/Past%20Proceedings/IJCAI-95-VOL2/PDF/016.pdf>
- Kotsakis C, Sideris MG (1999) On the adjustment of combined GPS/levelling/geoid networks. *J Geod* 73:412–421
- Kotsakis C (2007) A covariance-adaptive approach for regularized inversion in linear models. *Geophys J Int* 171:509–522
- Klees R, Ditmar P, Kusche J (2004) Numerical techniques for large least-squares problems with applications to GOCE. In: Sansò F (ed) *International Association of Geodesy symposia, “V Hotine-Marussi Symposium on Mathematical Geodesy”*, vol 127. Springer, Berlin, pp 12–21
- Kubik K (1970) The estimation of the weights of measured quantities within the method of least squares. *Bull Geod* 95:21–40
- Kusche J (2003) A Monte-Carlo technique for weight estimation in satellite geodesy. *J Geod* 76:641–652
- Lee S, Choi C, Kim J (2013) Evaluating the suitability of the EGM2008 geopotential model for the Korean peninsula using parallel computing on a diskless cluster. *Comput Geosci* 52:132–145
- Marchenko A, Tartachynska Z, Yakimovich A, Zablotskyj F (2003) Gravity anomalies and geoid heights derived from ERS-1, ERS-2, and Topex/Poseidon altimetry in the Antarctic peninsula area. In: *Proceedings of the 5th International Antarctic Geodesy Symposium AGS’03, Lviv, Ukraine, SCAR Report No. 23*. <http://www.scar.org/publications/reports/23/>
- Moreaux G (2008) Compactly supported radial covariance functions. *J Geod* 82(431):443. doi:10.1007/s00190-007-0195-4
- Moritz H (1978) Least-squares collocation. *Rev Geophys* 16(3):421–430
- Moritz H (1980) *Advanced physical geodesy*. Herbert Wichmann Verlag, Karlsruhe
- Pardo-Igúzquiza E (1997) MLREML: a computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. *Comput Geosci* 23:153–162
- Pardo-Igúzquiza E (1998) Maximum likelihood estimation of spatial covariance parameters. *Math Geol* 30:95–108
- Paolo F, Molina E (2010) Integrated marine gravity field in the Brazilian coast from altimeter-derived sea surface gradient and shipborne gravity. *J Geodyn* 50(5):347–354
- Rapp RH, Wang YM (1994) Dynamic topography estimates using Geosat data and a gravimetric geoid in the Gulf Stream region. *Geophys J Int* 117:511–528
- Reguzzoni M, Tselis N (2009) Optimal multi-step collocation: application to the space-wise approach for GOCE data analysis. *J Geod* 83(1):13–29
- Rummel R, Schwarz KP, Gerstl M (1979) Least squares collocation and regularization. *Bull Geod* 53:343–361
- Sabaka TJ, Rowlands DD, Luthcke SB, Boy J-P (2010) Improving global mass flux solutions from Gravity Recovery and Climate Experiment (GRACE) through forward modeling and continuous time correlation. *J Geophys Res* 115:B11403. doi:10.1029/2010JB007533
- Sadiq M, Tscherning CC, Ahmad Z (2010) Regional gravity field model in Pakistan area from the combination of CHAMP, GRACE and ground data using least squares collocation: a case study. *Adv Space Res* 46:1466–1476
- Saleh J, Li X, Wang Y, Roman DR, Smith DA (2013) Error analysis of the NGS’ surface gravity database. *J Geod* 87:203–221
- Schuh WD (2003) The processing of band-limited measurements: filtering techniques in the least squares context and in the presence of data gaps. *Space Sci Rev* 108(1–2):67–78
- Smyth GK (2002) An efficient algorithm for REML in heteroscedastic regression. *J Comput Graph Stat* 11:836–847
- Strykowski G (2000) Silkeborg gravity high revisited: horizontal extension of the source and its uniqueness. *Phys Chem Earth A* 25(4):375–380
- Tscherning CC (1991) The use of optimal estimation for gross-error detection in databases of spatially correlated data. *Bulletin d’Information (Bur Gravim int)* 68:79–89
- van Loon J (2008) *Functional and stochastic modelling of satellite gravity data*. Ph.D. thesis, Delft University of Technology
- Vergos GS, Tziavos IN, Andritsanos VD (2005) Gravity database generation and geoid model estimation using heterogeneous data. In: Jekeli C, Bastos L, Fernandes J (eds) *Proceedings of International Association of Geodesy Symposia “Gravity Geoid and Space Missions”*, vol 129, pp 155–160
- Yang Y, Zeng A, Zhang J (2009) Adaptive collocation with application in height system transformation. *J Geod* 83:403–410