J Geod (2008) 82:167–175
DOI 10.1007/s00190-007-0169-6

ORIGINAL ARTICLE

On a stronger-than-best property for best prediction

P. J. G. Teunissen

Received: 8 November 2006 / Accepted: 22 May 2007 / Published online: 18 August 2007

Abstract The minimum mean squared error (MMSE) criterion is a popular criterion for devising best predictors. In case of linear predictors, it has the advantage that no further distributional assumptions need to be made, other then about the first- and second-order moments. In the spatial and Earth sciences, it is the best linear unbiased predictor (BLUP) that is used most often. Despite the fact that in this case only the first- and second-order moments need to be known, one often still makes statements about the complete distribution, in particular when statistical testing is involved. For such cases, one can do better than the BLUP, as shown in Teunissen (J Geod. doi: 10.1007/s00190-007-0140-6, 2006), and thus devise predictors that have a smaller MMSE than the BLUP. Hence, these predictors are to be preferred over the BLUP, if one really values the MMSE-criterion. In the present contribution, we will show, however, that the BLUP has another optimality property than the MMSE-property, provided that the distribution is Gaussian. It will be shown that in the Gaussian case, the prediction error of the BLUP has the highest possible probability of all linear unbiased predictors of being bounded in the weighted squared norm sense. This is a stronger property than the often advertised MMSE-property of the BLUP.

Keywords Minimum mean squared error (MMSE) prediction · Least-squares collocation · Universal Kriging · Best linear unbiased prediction (BLUP) · Maximum probability of bounded prediction error

P. J. G. Teunissen (✉)
Delft Institute for Earth Observation and Space Systems (DEOS), Delft University of Technology, Kluyverweg 1, 2629 HS Delft, The Netherlands
e-mail: P.J.G.Teunissen@tudelft.nl

1 Introduction

We speak of prediction if a function of an observable random vector $y \in R^m$ is used to 'guess' the outcome of another random, but unobservable, vector $y_0 \in R^{m_0}$. If the function is given as G, then $G(y)$ is said to be the predictor of y_0. If $G(y)$ is a predictor of y_0, then $e_0 = y_0 - G(y)$ is its prediction error. When predicting spatially and/or temporal varying variates on the basis of observations of these variates at some locations in space and/or instances in time, one often uses the minimization of the mean squared prediction error as the criterion for optimal prediction. If $G(y)$ is the predictor, then $E||y_0 - G(y)||^2$ is its mean squared error (MSE). Note, since both y_0 and y are random, that the expectation E, or mean, is taken with respect to their joint probability density function (PDF), $f_{y_0 y}(y_0, y)$. Thus, $E||y_0 - G(y)||^2 = \int\int ||y_0 - G(y)||^2 f_{y_0 y}(y_0, y) dy_0 dy$. The predictor that succeeds in minimizing this mean squared prediction error is referred to as the best predictor. Important examples of such best prediction methods are least-squares collocation, universal Kriging, Wiener filtering, or recursive Kalman filtering (e.g., Moritz 1980; Cressie 1991; Kailath 1981). Under the correct conditions, all of these methods can be viewed as particular representations of the method of best linear unbiased prediction (BLUP). The BLUP achieves its minimum mean squared error (MMSE) within the class of linear unbiased predictors.

In Teunissen (2006), it has been shown, however, that for the same linear model on which the BLUP is based, meaningful predictors can be devised that have smaller mean squared prediction errors than the BLUP. Hence, if one really values the property of obtaining the smallest possible MSE, these predictors are to be preferred over the BLUP.

No distributional assumptions, other then about the first- and second-order moments, need to be made to establish the

MMSE-property of the BLUP. The MMSE-property implies that one can expect the squared-norm of the prediction error vector of the BLUP to be smaller on the average than the squared-norm of the prediction error vector of any other linear unbiased predictor. This is a nice property and often the best achievable in the absence of any further distributional information. The MMSE-property does, however, not reveal information about the frequency with which one can expect repeated outcomes of the prediction error to be close to zero.

To be able to compute such a probability, one would need information about the complete distribution. In the ideal case, one would then like to be in a position to select the predictor that has the highest possible probability of a bounded prediction error. In the present contribution, we will show that the BLUP is such a predictor in the Gaussian case. That is, the Gaussian BLUP has the highest possible probability of bounding the prediction error of all linear unbiased predictors. This is a stronger property than the MMSE-property of the BLUP.

This contribution is organised as follows. In Sect. 2, we introduce the linear model on which our prediction analysis is based. We also give a useful, but unconventional, representation of the class of linear unbiased predictors. This representation provides for an efficient derivation of the BLUP and its minimum error variance property in Sect. 3. In Sect. 4, we introduce an origin-centred ellipsoid of arbitrary shape and size, and show that its probability content based on the error PDF of the BLUP is the largest possible within the class of linear unbiased predictors. In Sect. 5, we show by means of examples the general applicability of the linear model on which the BLUP is based. Hence, the Gaussian predictors used in each of these applications also have the largest possible probability of bounding the prediction error.

In Sect. 6, we consider the maximum probability of bounded prediction error for other best predictors. In Sect. 7, we show that, with respect to the linear model used, estimation can be seen to be a special case of prediction. Hence, the best linear unbiased estimator (BLUE) can be seen to be a special case of the BLUP. As a consequence, the maximum probability of bounded prediction error, now becomes a maximum probability of bounded estimation error for the BLUE. This is clearly a stronger property than the minimum variance property as described by the well-known Gauss–Markov theorem.

2 Linear unbiased prediction

Consider the linear model

$$\begin{bmatrix} y \\ y_0 \end{bmatrix} = \begin{bmatrix} A \\ A_0 \end{bmatrix} x + \begin{bmatrix} e \\ e_0 \end{bmatrix} \tag{1}$$

in which $x \in R^n$ is a nonrandom unknown parameter vector and $[e^T, e_0^T]^T \in R^{m+m_0}$ is a random vector, with expectation and dispersion given as

$$E\begin{bmatrix} e \\ e_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and}$$
$$D\begin{bmatrix} e \\ e_0 \end{bmatrix} = D\begin{bmatrix} y \\ y_0 \end{bmatrix} = \begin{bmatrix} Q_{yy} & Q_{yy_0} \\ Q_{y_0y} & Q_{y_0y_0} \end{bmatrix} \tag{2}$$

respectively. The matrices A and A_0 of order $m \times n$ and $m_0 \times n$, respectively, are assumed known and matrix A is assumed to be of full column rank. The dispersion matrix is also assumed known.

It is our goal to predict y_0 on the basis of y. Let $G(y) = L_0 y + l_0$ be a linear predictor of y_0. Then $G(y)$ is said to be a linear unbiased predictor of y_0 if $E(G(y)) = E(y_0)$ for all x. Hence, $L_0E(y) + l_0 = E(y_0)$ and thus $L_0Ax + l_0 = A_0x$ should hold for all x. This shows that $G(y)$ is a linear unbiased predictor of y_0, if and only if $L_0A = A_0$ and $l_0 = 0$. This result can now be used to give a representation of linear unbiased predictors that will turn out to be useful in our analysis of the best linear unbiased predictor.

Linear unbiased predictors *Let $G(y)$ be a linear unbiased predictor (LUP) of y_0. Then an $m_0 \times (m-n)$ matrix H exists such that*

$$G(y) = A_0\hat{x} + Ht \tag{3}$$

where $\hat{x} = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} y$, $t = B^T y$, and B is an $m \times (m-n)$ basis matrix of which the columns span the null space of A^T.

Proof The sought for representation follows from solving the matrix equation $L_0A = A_0$ or its transposed form $A^T L_0^T = A_0^T$. The general solution of this transposed form is given by the sum of its homogeneous solution and a particular solution. Since BH^T is the general solution of the homogeneous equation $A^T L_0^T = 0$ and $Q_{yy}^{-1}A(A^T Q_{yy}^{-1} A)^{-1} A_0^T$ is a particular solution, the general solution for L_0 follows as $L_0 = A_0(A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} + HB^T$. Substitution of this solution into $G(y) = L_0 y + l_0$ gives, with $l_0 = 0$, the result Eq. (3). □

In Eq. (3), we recognize $\hat{x}$ as the BLUE of x and $t = B^T y$ as the redundancy vector of misclosures. The vector $(\hat{x}^T, t^T)^T$ stands in a one-to-one relation with the data vector y. We have

$$\begin{bmatrix} \hat{x} \\ t \end{bmatrix} = \begin{bmatrix} (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} \\ B^T \end{bmatrix} y \quad \Leftrightarrow$$
$$y = \left[A,\ Q_{yy}B(B^T Q_{yy} B)^{-1}\right] \begin{bmatrix} \hat{x} \\ t \end{bmatrix} \tag{4}$$

Note that $E(t) = 0$ and that $\hat{x}$ and t are uncorrelated. The $(m-n)$-vector t is identically zero in the absence of redundancy (the full rank matrix A will then be a square matrix

with $m = n$). Thus in the absence of redundancy, only a single LUP exists, namely $G(y) = A_0\hat{x} = A_0A^{-1}y$. Hence, it is the presence of redundancy ($m > n$) that gives us the freedom to select a best predictor from the class of linear unbiased predictors.

3 Best linear unbiased prediction

Let $\hat{G}(y)$ be the best linear unbiased predictor (BLUP) of y_0. Then $\hat{G}(y)$ is the solution of the minimization problem

$$E||y_0 - \hat{G}(y)||^2 = \min_{G \in LUP} E||y_0 - G(y)||^2 \tag{5}$$

Note that the minimization of the MSE is restricted to the class of LUPs. This is not needed per se. In Teunissen (2006), it has been shown that one can determine meaningful best predictors in classes of predictors that are larger than the class of linear unbiased predictors. Such predictors are however nonlinear. Hence, the restriction to unbiased predictors is needed if one wants to work with a linear predictor. The following will make this clear.

With $G(y) = L_0y + l_0$, $\bar{y}_0 = E(y_0) = A_0x$ and $\bar{y} = E(y) = Ax$, we have $E||y_0 - G(y)||^2 = E||(y_0 - \bar{y}_0) - L_0(y - \bar{y}) + ((A_0 - L_0A)x - l_0)||^2$, from which it follows that

$$\begin{aligned} E||y_0 - G(y)||^2 &= E||(y_0 - \bar{y}_0) - L_0(y - \bar{y})||^2 \\ &\quad + ||(A_0 - L_0A)x - l_0||^2 \\ &= \text{trace}\left(Q_{y_0y_0} - 2L_0Q_{y_0y} + L_0Q_{yy}L_0^T\right) \\ &\quad + ||(A_0 - L_0A)x - l_0||^2 \end{aligned} \tag{6}$$

In order to minimize the MSE, we would need to minimize this objective function as function of the matrix L_0 and the vector l_0. Note, however, that the unknown parameter vector x is part of the objective function. Hence, the minimizer of the objective function would then depend on the unknown x and would therefore not result in a useable predictor. This problem does not occur if the second term on the right-hand side of Eq. (6) is absent, which is the case when one restricts the minimization to the class of LUPs.

To determine the BLUP, we make use of the representation of Eq. (3). We have

$$\begin{aligned} &E||y_0 - G(y)||^2 = E||y_0 - A_0\hat{x} - Ht||^2 \\ &= E||(y_0 - A_0\hat{x} - Q_{y_0t}Q_{tt}^{-1}t) - (H - Q_{y_0t}Q_{tt}^{-1})t||^2 \\ &= E||y_0 - A_0\hat{x} - Q_{y_0t}Q_{tt}^{-1}t||^2 + E||(H - Q_{y_0t}Q_{tt}^{-1})t||^2 \end{aligned} \tag{7}$$

since t is uncorrelated with $\hat{x}$ and uncorrelated with $y_0 - Q_{y_0t}Q_{tt}^{-1}t$. From the last equation it follows that the MSE is minimized, if matrix H is chosen as $H = Q_{y_0t}Q_{tt}^{-1}$. We are now in the position to determine the BLUP of y_0.

Best linear unbiased predictor *Let $\hat{y}_0 = \hat{G}(y)$ be the best linear unbiased predictor (BLUP) of y_0. Then*

$$\begin{aligned} \hat{y}_0 &= A_0\hat{x} + Q_{y_0t}Q_{tt}^{-1}t \\ &= A_0\hat{x} + Q_{y_0y}Q_{yy}^{-1}(y - A\hat{x}) \end{aligned} \tag{8}$$

Proof Substitution of $H = Q_{y_0t}Q_{tt}^{-1}$ into Eq. (3) gives the first expression of Eq. (8). To determine the second expression of Eq. (8) from its first, we note that $Q_{y_0t}Q_{tt}^{-1}t = Q_{y_0y}B(B^TQ_{yy}B)^{-1}B^Ty$. With the use of the projector identity $Q_{yy}B(B^TQ_{yy}B)^{-1}B^T = I_m - A(A^TQ_{yy}^{-1}A)^{-1}A^TQ_{yy}^{-1}$, we obtain $Q_{y_0t}Q_{tt}^{-1}t = Q_{y_0y}Q_{yy}^{-1}(y - A\hat{x})$, which proves the second expression of Eq. (8). □

The first expression of Eq. (8) explictly shows the LUP structure (cf. Eq. 3). In the second expression of Eq. (8), the BLUP has been written in terms of y and $\hat{x}$.

The BLUP is the MMSE predictor in the class of LUPs. Hence, it has the smallest mean squared prediction error within this class. The BLUP is, however, also the predictor that has the smallest error variance of all LUPs. The BLUP is therefore sometimes also referred to as the minimum error variance linear unbiased predictor.

To compare the error variance of the BLUP with that of an arbitrary LUP, we first express the LUP prediction error, $e_0 = y_0 - A_0\hat{x} - Ht$, in the BLUP prediction error. This gives $e_0 = \hat{e}_0 - (H - Q_{y_0t}Q_{tt}^{-1})t$. Application of the variance propagation law, noting that $\hat{e}_0$ and t are uncorrelated, gives

$$Q_{e_0e_0} = Q_{\hat{e}_0\hat{e}_0} + (H - Q_{y_0t}Q_{tt}^{-1})Q_{tt}(H - Q_{y_0t}Q_{tt}^{-1})^T \tag{9}$$

Equation (9) shows by how much the error variance of an arbitrary LUP differs from the error variance of the BLUP. Since the second term on the right-hand side of Eq. (9) is positive semidefinite, we have $f^TQ_{\hat{e}_0\hat{e}_0}f \leq f^TQ_{e_0e_0}f$ for any $f \in R^{m_0}$. Thus, the error variances of LUPs are never smaller than the error variance of the BLUP. We summarize this minimum error variance property of the BLUP as follows.

Minimum error variance *Let $\hat{e}_0 = y_0 - \hat{y}_0$ and $e_0 = y_0 - G(y)$ be the prediction error of the BLUP and of an arbitrary LUP, respectively. Then*

$$Q_{\hat{e}_0\hat{e}_0} \leq Q_{e_0e_0} \tag{10}$$

Now let us for the moment reflect on the properties of the BLUP.

1. We know that the BLUP has a zero-mean prediction error, $E(\hat{e}_0) = 0$. This implies that we can expect the prediction error to be zero on average.
2. We know that the BLUP has the smallest possible mean squared prediction error of all LUPs, $E||\hat{e}_0||^2 \leq E||e_0||^2$. This implies that we can expect the squared-norm of the

prediction error vector of the BLUP to be smaller on the average than the squared-norm of the prediction error vector of any other LUP.

3. We know that the variance of any linear function of the prediction error vector of the BLUP is never larger than the variance of the same function of the prediction error vector of any other LUP, $f^T Q_{\hat{e}_0\hat{e}_0} f \leq f^T Q_{e_0e_0} f$.

The above properties are all nice properties indeed. They do not, however, tell us anything about the frequency with which one can expect repeated outcomes of the prediction error $\hat{e}_0$ to be close to zero. That is, they do not allow us to determine the probability that the prediction error is close to zero. In order to determine such probability, we need the complete PDF of $\hat{e}_0$. In the absence of any information other than the first two moments of the prediction error, the best one can do, if one wants to make a probabilistic statement, is to make use of the Chebyshev inequality (e.g. Stark and Wood 1986; Casella and Berger 1990; Teunissen et al. 2005). In our case, however, we would need a multivariate version of this inequality.

Multivariate Chebyshev inequality *Let e_0 be the prediction error of an arbitrary LUP. Then, for any matrix $W \geq 0$ and any $r \in R$, we have the inequality*

$$P(||e_0||_W^2 \geq r^2) \leq \frac{\text{trace}(W Q_{e_0e_0})}{r^2} \tag{11}$$

with the squared weighted norm $||.||_W^2 = (.)^T W (.)$.

Proof Let $f_{e_0}(\alpha)$ be the PDF of e_0. Then $E||e_0||_W^2 = \int_{R^{m_0}} ||\alpha||_W^2 f_{e_0}(\alpha)\mathrm{d}\alpha = \int_{||\alpha||_W^2 \leq r^2} ||\alpha||_W^2 f_{e_0}(\alpha)\mathrm{d}\alpha + \int_{||\alpha||_W^2 \geq r^2} ||\alpha||_W^2 f_{e_0}(\alpha)\mathrm{d}\alpha \geq \int_{||\alpha||_W^2 \geq r^2} ||\alpha||_W^2 f_{e_0}(\alpha)\mathrm{d}\alpha \geq r^2 \int_{||\alpha||_W^2 \geq r^2} f_{e_0}(\alpha)\mathrm{d}\alpha = r^2 P(||e_0||_W^2 \geq r^2)$. Furthermore, we have $E||e_0||_W^2 = E(e_0^T W e_0) = E(\text{trace}(W e_0 e_0^T)) = \text{trace}(W Q_{e_0e_0})$, since $E(e_0) = 0$. From this and the inequality $E||e_0||_W^2 \geq r^2 P(||e_0||_W^2 \geq r^2)$, the stated result follows. □

This result states that the probability that the prediction error of any LUP resides outside the origin-centred ellipsoid $||e_0||_W^2 = r^2$ is bounded from above by $\text{trace}(W Q_{e_0e_0})/r^2$. Hence, the probability that the prediction error resides outside the ellipsoid, will become more tightly bounded when the precision of the prediction error improves. With reference to Eq. (10), this implies that the upperbound for the BLUP will be smaller than the corresponding upperbound for any other LUP. From this, one may not conclude, however, that the probability that the prediction error resides outside the ellipsoid is smaller for the BLUP than for any other LUP. Ideally, however, one would like to be in a position to be able to make such an optimality statement. In the next section, we will show when this is the case.

4 Optimality of the BLUP in the Gaussian case

So far, we only made use of the first- and second-order moments of the random vectors y and y_0. That is, no further distributional assumptions were made about these random vectors. This implies that the two optimality properties of the BLUP, the MMSE property and the minimum error variance property, both hold true irrespective of the distributions of y and y_0. This, of course, is a nice result, which is also often stressed in the literature (e.g. Arnold 1981; Bar-Shalom and Li 1993; Koch 1987; Myers and Milton 1991; Rao and Toutenburg 1995; Stapleton 1995; Stark and Wood 1986; Sengupya and Jammalamadaka 2003). However, if one really values the mentioned two properties, why not aim for predictors that have these two properties in a class larger than the class of LUPs? After all, such predictors, when they exist, will have a smaller mean squared prediction error and a smaller error variance than the BLUP. In Teunissen (2006), it has been shown that such predictors of y_0 indeed exist. They are found in the class of equivariant predictors and in the class of integer equivariant predictors.

Does the result of Teunissen (2006) make the BLUP obsolete? The answer must be yes if one prefers predictors with smaller mean squared prediction errors and smaller error variances. However, as the present section will show, the BLUP has another optimality property and one which is stronger than the above-mentioned two. This stronger optimality property holds true in case the joint distribution of y and y_0 is Gaussian.

Let $\hat{e}_0$ be the prediction error of the BLUP and let e_0 be the prediction error of an arbitrary LUP. If y and y_0 of the linear model of Eqs. (1) and (2) are Gaussian distributed, then so are the zero-mean random vectors $\hat{e}_0$ and e_0. We have $\hat{e}_0 \sim \text{N}(0, Q_{\hat{e}_0\hat{e}_0})$ and $e_0 \sim \text{N}(0, Q_{e_0e_0})$, with $Q_{\hat{e}_0\hat{e}_0} \leq Q_{e_0e_0}$. Now let us first transform the two prediction errors, $\hat{e}_0$ and e_0, as $\hat{u} = F\hat{e}_0$ and $v = Fe_0$, respectively, in which F is an arbitrary, but invertible, $m_0 \times m_0$ matrix. Then $\hat{u} \sim \text{N}(0, Q_{\hat{u}\hat{u}} = F Q_{\hat{e}_0\hat{e}_0} F^T)$ and $v \sim \text{N}(0, Q_{vv} = F Q_{e_0e_0} F^T)$, with $Q_{\hat{u}\hat{u}} \leq Q_{vv}$. If we define the random vector $u = Q_{\hat{u}\hat{u}}^{1/2} Q_{vv}^{-1/2} v$, then u has a distribution that is identical to that of $\hat{u}$ (note: the square-root matrix notation $M^{1/2}$ denotes a matrix satisfying $M = M^{1/2} M^{1/2}$; thus $M^{-1/2} M M^{-1/2} = I$). We therefore have the probabilistic equality

$$P(||\hat{u}||^2 \leq r^2) = P(||u||^2 \leq r^2) \tag{12}$$

The probability that $\hat{u}$ resides in an origin-centred hypersphere with radius r is thus equal to the probability that u resides in the same hypersphere. The squared norm of u can be expressed in the squared norm of v as

$$\begin{aligned} ||u||^2 &= ||Q_{\hat{u}\hat{u}}^{1/2} Q_{vv}^{-1/2} v||^2 \\ &= ||v||^2 + v^T Q_{vv}^{-1/2} (Q_{\hat{u}\hat{u}} - Q_{vv}) Q_{vv}^{-1/2} v \end{aligned} \tag{13}$$

Note that $||u||^2 \leq ||v||^2$, since $Q_{\hat{u}\hat{u}} \leq Q_{vv}$. Substitution of the last expression of Eq. (13) into Eq. (12) gives

$$\begin{aligned} P(||\hat{u}||^2 \leq r^2) &= P(||v||^2 + v^T Q_{vv}^{-1/2} \\ &\quad \times (Q_{\hat{u}\hat{u}} - Q_{vv}) Q_{vv}^{-1/2} v \leq r^2) \\ &\geq P(||v||^2 \leq r^2) \end{aligned} \tag{14}$$

Since $\hat{u} = F\hat{e}_0$ and $v = Fe_0$, we may write the squared norms of $\hat{u}$ and v, as weighted squared norms of $\hat{e}_0$ and e_0: $||\hat{u}||^2 = ||\hat{e}_0||_W^2$ and $||v||^2 = ||e_0||_W^2$, with $W = F^T F$. We may therefore express the probabilistic relation of Eq. (14) directly in terms of the prediction errors. We therefore have the following optimality result for the BLUP.

Theorem (BLUP's maximum probability of bounded prediction error): *Let y and y_0 have a joint Gaussian distribution with first- and second-order moments as given in Eqs. (1) and (2). Further, let $\hat{y}_0$ be the BLUP of y_0 and let $G(y)$ be any LUP of y_0. Then*

$$P(||y_0 - \hat{y}_0||_W^2 \leq r^2) \geq P(||y_0 - G(y)||_W^2 \leq r^2) \tag{15}$$

for any $W > 0$ and any $r \in R$.

This result states that in the Gaussian case, given an origin-centred ellipsoid of arbitrary size and shape, the prediction error of the BLUP has of all LUP errors, the highest probability of residing in this ellipsoid. In practical terms this implies that in case of a repeated experiment, one can expect an origin-centred ellipsoid of arbitrary shape and size, to catch more of the BLUP error-scatter than of any other LUP error-scatter. Hence, the above result is a much stronger property, than the first- and second-order moment-based BLUP properties of minimum error variance or minimum mean squared prediction error.

A closer look at the derivation on which the above theorem is based, reveals that the result of the theorem can even be strenghtened. The only place where the assumption of Gaussianity is needed in the derivation is in the proof of Eq. (12). Thus a strengthening of the theorem is realized, if we are able to show that this same probabilistic equality can also hold true for other distributions. Assume therefore that the PDF's of $\hat{u}$ and v are given as

$$f_{\hat{u}}(\alpha) = \frac{h(\alpha^T Q_{\hat{u}\hat{u}}^{-1}\alpha)}{\sqrt{|\det Q_{\hat{u}\hat{u}}|}} \quad \text{and} \quad f_v(\alpha) = \frac{h(\alpha^T Q_{vv}^{-1}\alpha)}{\sqrt{|\det Q_{vv}|}} \tag{16}$$

for some function $h : R_0^+ \mapsto R$. Then it follows from an application of the PDF transformation rule, that $u = Q_{\hat{u}\hat{u}}^{1/2} Q_{vv}^{-1/2} v$ has the same distribution as $\hat{u}$. Hence, the probabilistic equality of Eq. (12) holds true for all distributions of the type given in Eq. (16). Note that the choice $h(x) = (2\pi)^{-m_0/2} \exp(-x/2)$ leads to the Gaussian distribution. Also note that PDFs of the above type are elliptically contoured, i.e. their contour surfaces are ellipsoids, just like in case of the Gaussian distribution. Thus the result of the theorem can be strengthened by stating that it holds true for all elliptically contoured distributions.

From the above theorem we conclude that the BLUP is the preferred predictor in the Gaussian (or elliptically contoured) case, even with the knowledge that other predictors exist that outperform the BLUP as far as their error variance or mean squared prediction error is concerned.

At this point it is also worthwhile to make the following remark. The above shown maximum probability property should not be confused with the approach of maximizing the differential probability (i.e. density), which is implicit in the principle of maximum likelihood and which has been used by Gauss in his first justification of the method of least-squares (Gauss 1809; Waterhouse 1990). This, despite the fact that terms as 'maximum probability' and 'most probable' are still used in this context. In particular, the maximum probability property of the BLUP should not be confused with the property of the maximum likelihood solution of the posterior probability density function, which unfortunately is often referred to as the maximum a posteriori probability (MAP) solution. The principle of the MAP is not to maximize the probability, but rather, just like the likelihood principle, to maximize a likelihood function, which in case of random parameters is equivalent to maximizing a conditional density, i.e. the derivative of probability. We therefore agree with Scharf (1991) that it would be better to give the MAP a different name. This could be the maximum likelihood predictor (MLP) or, as proposed by Scharf (ibid), the maximum a posteriori likelihood solution, though the acronym MAL might not please some readers.

5 Some applications

The results obtained so far are based on the linear model as described by Eqs. (1) and (2). The representation of this model, although suited for the derivation and analysis of the BLUP, is probably not in a form that directly appeals to concrete applications. We will therefore show, by means of some important examples, the wide range of prediction problems that can be covered with this model. Since the Gaussian assumption is often made in these applications, the probabilistic optimality property of Eq. (15) holds true for their best predictors as well.

5.1 Prediction of a random vector with unknown mean

Many applications can be described as observing a random vector x', with unknown mean x, in the presence of additive noise e. The goal is then to predict x' on the basis of the vector of observables y. In the linear case, the model can be described as $y = Ax' + e$. We assume that the variance matrix $Q_{x'x'}$ of x' is known and that e is a zero-mean random

vector, uncorrelated with x', with known variance matrix Q_{ee}. To set the stage for predicting x', we set in Eqs. (1) and (2), $e \to A(x'-x)+e$, $y_0 \to x'$, $A_0 \to I$, and $e_0 \to x'-x$. With these settings, we have $Q_{y_0 y} \to Q_{x'x'}A^T$ and $Q_{yy} \to AQ_{x'x'}A^T + Q_{ee}$. The BLUP of x' follows then, from Eq. (8), as

$$\hat{x}' = \hat{x} + Q_{x'x'}A^T(AQ_{x'x'}A^T + Q_{ee})^{-1}(y - A\hat{x}) \tag{17}$$

This representation of $\hat{x}'$ is known as the variance-form. Using the well-known matrix inversion lemma, the corresponding information-form follows as

$$\hat{x}' = \hat{x} + (Q_{x'x'}^{-1} + A^T Q_{ee}^{-1}A)^{-1}A^T Q_{ee}^{-1}(y - A\hat{x}) \tag{18}$$

The solution $\hat{x}'$ is referred to as the batch solution. Under certain conditions on A, $Q_{x'x'}$ and Q_{ee} (e.g. Teunissen 2001), one can also formulate a recursive solution, which ultimately leads to the well-known Kalman filter. Hence, if the state vectors and the observables on which the Kalman filter is based, are Gaussian-distributed, the Kalman filtered state will not only have the often advertised MMSE property (e.g. Kailath 1981; Sorenson 1985; Bar-Shalom and Li 1993), but also the maximum probability property of Eq. (15). The Gaussian assumption is often made in Kalman filtering, in particular in case of model validation for the detection and identification of model misspecifications.

5.2 The trend-signal-noise model of collocation

The so-called trend-signal-noise model of collocation is another special case of the model in Eqs. (1) and (2). It has found wide-spread application in the spatial and Earth sciences (e.g. Moritz 1973, 1980; Rummel 1976; Dermanis 1980; Sanso 1986; Journel and Huijbregts 1991; Cressie 1991; Wackernagel 1995). In this model, the observable vector y is written as a sum of three terms, $y = Ax + s + n$, with Ax a deterministic trend, with an unknown parameter vector x, s a zero-mean random signal vector, and n a zero-mean random noise vector. Often one can extend the trend-signal-noise model so as to hold true for an unobservable vector $y_0 = A_0 x + s_0 + n_0$, in which s_0 and n_0 are uncorrelated zero-mean random vectors, and n_0 is also uncorrelated with n. For instance, y_0 could be a functional of the same type as y, but evaluated at a different location in space or at a different instant in time. To set the stage for predicting y_0, s_0 and n_0, we set in Eqs. (1) and (2), $e \to s+n$, $y_0 \to (y_0^T, s_0^T, n_0^T)^T$, $A_0 \to (A_0^T, 0, 0)^T$, and $e_0 \to \big((s_0+n_0)^T, s_0^T, n_0^T\big)^T$. With these settings, we obtain from Eq. (8), the BLUP of $(y_0^T, s_0^T, n_0^T)^T$ as

$$\begin{aligned}\hat{y}_0 &= A_0\hat{x} + Q_{s_0 s}(Q_{ss}+Q_{nn})^{-1}(y - A\hat{x})\\ \hat{s}_0 &= Q_{s_0 s}(Q_{ss}+Q_{nn})^{-1}(y - A\hat{x})\\ \hat{n}_0 &= 0\end{aligned}$$

with $\hat{x} = \big(A^T(Q_{ss}+Q_{nn})^{-1}A\big)^{-1}A^T(Q_{ss}+Q_{nn})^{-1}y$. These are the well-known results of least-squares collocation (e.g. Moritz 1980), or universal Kriging (e.g. Wackernagel 1995).

5.3 Predicting error components

We give two examples in which the prediction of error components is of interest. Let e in $y = Ax + e$ be given as $e = E\epsilon$, with matrix E known and where ϵ is a zero-mean random vector with variance matrix $Q_{\epsilon\epsilon}$. As an application of this formulation, the entries of ϵ can be thought of as being the individual error components that contribute to the overall error vector e. This model is known as the so-called mixed model and it often forms the basis for variance component estimation (e.g., Rao and Kleffe 1988).

To set the stage for predicting ϵ, we set in Eqs. (1) and (2), $e \to E\epsilon$, $y_0 \to \epsilon$, $A_0 \to 0$, and $e_0 \to \epsilon$. With these settings, we obtain from an application of Eq. (8), the BLUP of ϵ as

$$\hat{\epsilon} = Q_{\epsilon\epsilon}E^T(EQ_{\epsilon\epsilon}E^T)^{-1}(y - A\hat{x}) \tag{19}$$

with $\hat{x} = (A^T(EQ_{\epsilon\epsilon}E^T)^{-1}A)^{-1}A^T(EQ_{\epsilon\epsilon}E^T)^{-1}y$. Note that for the special case $E = I$, the BLUP of e is obtained as $y - A\hat{x}$.

As another application, one may consider the case where derived observables, instead of the original observables, are used to formulate the linear model. In many applications not the original data vector y is used to set up the observation equations, but rather linear functions of y. In case of Global Navigation Satellite Systems, for instance, the double-difference carrier-phase observations are often used, rather than the undifferenced carrier-phase observations (e.g. Teunissen and Kleusberg 1998; Misra and Enge 2006). In the case of levelling, the observed height difference of a levelling line is often used, rather than the individual readings (e.g., Kahmen and Faig 1987).

The use of derived observables is often done with the purpose of reducing the number of unknowns by eliminating the so-called nuisance parameters. As a result, the linear model takes the form $D^T y = Ax + D^T e$, with e the error component of the original data vector and where $D^T y$ is the vector of derived observables. Although one works in this set up with the derived vector of observables $D^T y$, one often still has the need to recover the error component of the original data vector y. If we use the settings $y \to D^T y$, $e \to D^T e$, $y_0 \to e$, $A_0 \to 0$, and $e_0 \to e$, in the linear model defined in Eqs. (1) and (2), we have $Q_{y_0 y} \to Q_{ee}D$ and $Q_{yy} \to D^T Q_{ee} D$. Hence, with these settings, we obtain from Eq. (8), the BLUP of e as

$$\hat{e} = Q_{ee}D(D^T Q_{ee} D)^{-1}(D^T y - A\hat{x}) \tag{20}$$

with $\hat{x} = (A^T(D^T Q_{ee} D)^{-1} A)^{-1} A^T (D^T Q_{ee} D)^{-1} D^T y$. Note that the linear model with derived observables reduces to the linear model of condition equations when $A = 0$.

6 Best prediction

So far, we have based our analysis on the linear model as defined by Eqs. (1) and (2). The reason for choosing this linear model as starting point, is due to our belief that this model covers by far the most relevant geodetic applications. This model is characterized by the fact that the unknown means of y and y_0 are linked by means of a known relationship. Let us now, however, consider the case where no assumptions are made about the functional relationship between y and y_0. Furthermore, let us assume that the MMSE-criterion is applied for an unspecified class of predictors Ω. We now define the best predictor $\hat{G}(y)$ of y_0, as the predictor that satisfies $E||y_0 - \hat{G}(y)||^2_W = \min_{G\in\Omega} E||y_0 - G(y)||^2_W$, for any $W \geq 0$. Thus for $W = ff^T$, we have, with $\hat{e}_0 = y_0 - \hat{G}(y)$ and $e_0 = y_0 - G(y)$, that $E(f^T \hat{e}_0 \hat{e}_0^T f) \leq E(f^T e_0 e_0^T f)$ for any $f \in R^{m_0}$. Hence, if $\hat{G}(y)$ is unbiased (i.e. $E(\hat{e}_0) = 0$), then $Q_{\hat{e}_0\hat{e}_0} \leq Q_{e_0e_0}$, for any unbiased predictor $G \in \Omega$. This shows, with reference to Eq. (15), that the best predictor of class Ω, when unbiased and Gaussian distributed, will also have the maximum probability property.

Let us now consider the most relaxed class of predictors. It can be shown, if no restrictions are put on the class of predictors Ω, that the best predictor is given by the conditional mean, $\hat{G}(y) = E(y_0|y)$ (e.g. Bar-Shalom and Li 1993; Teunissen 2006). This predictor is unbiased and, in the Gaussian case, it is given, with $\bar{y}_0 = E(y_0)$ and $\bar{y} = E(y)$, as

$$\hat{G}(y) = \bar{y}_0 + Q_{y_0y} Q_{yy}^{-1} (y - \bar{y}) \tag{21}$$

The conclusion is that the probability that the error of this best Gaussian predictor resides inside an origin-centred ellipsoid is largest of all unbiased Gaussian predictors. Note that the Gaussian best predictor of Eq. (21) has the same structure as the BLUP. However, it requires, in contrast to the BLUP, that the two means, $\bar{y}_0$ and $\bar{y}$, are known.

7 Estimation as special case of prediction

Recall that we speak of prediction if a function of an observable random vector y is used to 'guess' the outcome of another random, but unobservable, vector y_0. We speak of estimation, however, if a function of y is used to 'guess' the value of a deterministic, but unknown, parameter vector x, or a function thereof. We will now show that, with respect to the linear model of Eqs. (1) and (2), estimation can be seen to be a special case of prediction.

Let us assume that e_0 in Eq. (1) is identically zero. The joint PDF of y_0 and y is then given as $f_{y_0y}(y_0, y) = \delta(y_0 - A_0x) f_y(y)$, in which $f_y(y)$ is the PDF of y and $\delta(\tau)$ is the Dirac impulse function (with the properties: $\int \delta(\tau)\mathrm{d}\tau = 1$ and $\int g(\tau)\delta(\tau - v)\mathrm{d}\tau = g(v)$). The MSE of a predictor $G(y)$ of y_0 becomes then $E||y_0 - G(y)||^2 = \int\int ||y_0 - G(y)||^2 f_{y_0y}(y_0, y)\mathrm{d}y_0\mathrm{d}y = \int ||A_0x - G(y)||^2 f_y(y)\mathrm{d}y = E||A_0x - G(y)||^2$, which is the MSE of $G(y)$ as estimator of A_0x. Hence, if e_0 is identically zero, minimizing the mean squared *prediction* error is the same as minimizing the mean squared *estimation* error.

The consequence of the above equivalence is that the BLUP-result given in Eq. (8) can be seen as a generalization of the Gauss–Markov theorem of best linear unbiased estimation (BLUE). Indeed, if e_0 is identically zero, then $Q_{y_0t} = 0$, $Q_{y_0y} = 0$ and Eq. (8) reduces to $\hat{y}_0 = A_0\hat{x}$, which is the expression for the BLUE of $E(y_0) = A_0x$. The BLUE-property of $\hat{y}_0 = A_0\hat{x}$ is a consequence of the minimum error variance property of the BLUP. The minimum error variance of $\hat{e}_0 = y_0 - \hat{y}_0$ becomes, since e_0 is identically zero and therefore $y_0 = A_0x$ is now nonrandom, a minimum variance of $\hat{y}_0$.

The fact that the BLUE is a special case of the BLUP, implies that we have a similar maximum probability property for the BLUE as the one given in Eq. (15) for the BLUP.

Corollary (BLUE's maximum probability of bounded error): *Let y have a Gaussian distribution with first- and second-order moments as given in Eqs.* (1) *and* (2)*. Further, let $\hat{y}_0 = A_0\hat{x}$ be the BLUE of $y_0 = A_0x$ and let $G(y)$ be any linear unbiased estimator (LUE) of y_0. Then*

$$P(||y_0 - \hat{y}_0||^2_W \leq r^2) \geq P(||y_0 - G(y)||^2_W \leq r^2) \tag{22}$$

for any $W > 0$ and any $r \in R$.

This result states that in the Gaussian case, given an y_0-centred ellipsoid of arbitrary size and shape, the BLUE of y_0 has of all LUEs of y_0, the highest probability of residing in this ellipsoid. This is, of course, again a much stronger property than the minimum variance property of the BLUE. Note, as in case of the BLUP, that the above result can also be shown to hold true for elliptically contoured distributions.

8 Summary and conclusions

The MMSE criterion is a popular criterion for devising best predictors. Since one can minimize the MSE over different classes of functions, there are different predictors that one can call 'best'. In the theory of the linear model, it is the BLUP which is most often used, although sometimes under different names. In the spatial and Earth science disciplines, for instance, the BLUP is also known as least-squares collocation or universal Kriging. The BLUP is the predictor that minimizes the MSE within the class of linear unbiased predictors (LUP).

In Teunissen (2006) it was shown that for the linear model other predictors exist that outperform the BLUP in the MMSE sense. Examples are the best equivariant predictor (BEP) and the best integer equivariant predictor (BIEP). The BIEP, for instance, minimizes the MSE within the class of integer equivariant predictors. Since the class of LUPs is a subset of the class of integer equivariant predictors, the MMSE of the BIEP is never larger than that of the BLUP. An advantage of the BLUP over the BIEP is that it only requires information about the first and second order moments of the distribution, whereas the complete distribution is needed for the BIEP. In the theory of the linear model, however, the distributional assumptions are often not restricted to the first and second order moments. More often than not, the complete distribution is assumed known (except for the scale, which is of no consequence for computing the prediction), for instance, to be able to apply hypothesis testing for model validation purposes. Hence, if one really values the MMSE-property in this case, one should know that the BLUP is not the 'best' predictor.

Does this mean that the BLUP can always be outperformed by other 'best' predictors? If one restricts attention to the MMSE-property, the answer must be yes. However, as we have shown in the present contribution, the BLUP has also another optimality property which is different from its MMSE-property. It was shown, for the linear model with Gaussian (or elliptically contoured) distribution, that, given an origin-centred ellipsoid of arbitrary size and shape, the prediction error of the BLUP has of all LUP errors, the highest probability of residing in this ellipsoid. Thus in this case, the BLUP has of all LUPs the highest probability of having its prediction error bounded in an ellipsoidal sense. This is a much stronger property than the MMSE-property, since it is expressed directly in terms of the probability, rather then only in the first and second order moments of the predictor.

The maximum probability property of the BLUP should not be confused with the property of the maximum likelihood solution of the posterior probability density function, which unfortunately is often referred to as the maximum a posteriori probability (MAP) solution. The principle of the MAP is not to maximize the probability, but rather, just like the likelihood principle, to maximize a likelihood function, which in case of random parameters is equivalent to maximizing a conditional density, i.e. the derivative of probability.

It was shown that also the BLUE has a maximum probability property of bounded error. Based on the recognition that the BLUE can be seen as a special case of the BLUP, we obtained as a corollary that the Gaussian (or elliptically contoured) estimation error of the BLUE has, of all LUEs, the highest probability of residing in an arbitrary, but origin-centred ellipsoid. From the maximum probability property of the BLUP and the BLUE, we therefore draw the conclusion that in case of a linear model with Gaussian (or elliptically contoured) distribution, the BLUP and BLUE are the preferred predictor and estimator, respectively, rather than their competitors having a smaller MSE. This also implies that in case of dealings with such models, the motivation for using the BLUP or the BLUE should not entirely be based anymore on their MMSE-property, as is done in the current literature, but also on their maximum probability property of bounded error.

Acknowledgments The author thanks Prof. Fernando Sanso for pointing out that the proof of the theorem admits the generalization as given by Eq. (16) and he thanks an anonymous reviewer for the reference to existing, albeit different, usage of the term maximal probability.

References

Arnold SF (1981) The theory of linear models and multivariate analysis. Wiley, New York

Bar-Shalom Y, Li X-R (1993) Estimation and tracking. Artech House, Boston, London

Casella G, Berger RL (1990) Statistical inference. Brooks/Cole, Belmont

Cressie N (1991) Statistics for spatial data. Wiley, New York

Dermanis A (1980) Adjustment of geodetic observations in the presence of signals. In: Proceedings of the international school of advanced geodesy, vol 38. Bollettino di Geodesia e Scienze Affini, pp 419–445

Gauss CF (1809) Theoria motus corporum celestium. Hamburg, 1809. Translated into English by C.H. Davis in 1857; reprinted by Dover, New York 1963

Journel AG, Huijbregts ChJ (1991) Mining geostatistics. Academic, New York

Kahmen H, Faig W (1987) Surveying. de Gruyter

Kailath T (1981) Lectures on Wiener and Kalman filtering, 2nd edn. Springer, NY

Koch KR (1987) Parameter estimation and hypothesis testing in linear models. Springer, Heidelberg

Misra P, Enge P (2006) Global positioning system: signals, measurements, and performance, 2nd edn. Ganga-Jamuna Press

Moritz H (1973) Least-squares collocation. Deutsche Geodaetische Kommission, Reihe A, No. 59, Muenchen

Moritz H (1980) Advanced physical geodesy. Herbert Wichmann Verlag Karlsruhe

Myers RH, Milton JS (1991) A first course in the theory of linear statistical models. PWS-Kent, Boston

Rao CR, Kleffe J (1988) Estimation of variance components and applications. vol. 3. North Holland, Series in Statistics and Probability

Rao CR, Toutenburg H (1995) Linear models: least-squares and alternatives. Springer, Heidelberg

Rummel R (1976) A model comparison in least-squares collocation. Bull Geod 50:181–192

Sanso F (1986) Statistical methods in physical geodesy. In: Suenkel H (ed.) Mathematical and numerical techniques in physical geodesy, Lecture Notes in Earth Sciences, vol. 7. Springer, Heidelberg 7:49–156

Scharf LL (1991) Statistical signal processing. Addison-Wesley, New York

Sengupya D, Jammalamadaka SR (2003) Linear models: an integrated approach. World Scientific, Singapore

Sorenson HW (1985) Kalman Filtering: theory and application. IEEE Press, NY

Stapleton JH (1995) Linear statistical models. Wiley, New York
Stark H, Wood JW (1986) Probability, random processes, and estimation theory for engineers. Prentice-Hall, New Jersey
Teunissen PJG (2001) Dynamic data processing: recursive least-squares. Delft University Press, Delft
Teunissen PJG (2006) Best prediction in linear models with mixed integer/real unknowns: theory and application. J Geod (in press). DOI: 10.1007/s00190-007-0140-6
Teunissen PJG, Kleusberg A (1998) GPS for geodesy. 2nd edn. Springer, Heidelberg
Teunissen PJG, Simons D, Tiberius CCJM (2005) Probability and observation theory. Lecture Notes Delft University of Technology, Delft, The Netherlands
Wackernagel H (1995) Multivariate geostatistics. Springer, Heidelberg
Waterhouse CW (1990) Gauss's first argument for least squares. Arch History Exact Sci 41(1):41–52