Journal of Multivariate Analysis **76**, 249–266 (2001)
doi:10.1006/jmva.2000.1930, available online at http://www.idealibrary.com on IDEAL®

Extended Gauss–Markov Theorem for Nonparametric Mixed-Effects Models

Su-Yun Huang

Academia Sinica, Taipei, Taiwan, Republic of China
E-mail: syhuang@stat.sinica.edu.tw

and

Henry Horng-Shing Lu

National Chiao-Tung University, Hsinchu, Taiwan, Republic of China
E-mail: hslu@stat.nctu.edu.tw

Received April 29, 1998; published online December 21, 2000

The Gauss–Markov theorem provides a golden standard for constructing the best linear unbiased estimation for linear models. The main purpose of this article is to extend the Gauss–Markov theorem to include nonparametric mixed-effects models. The extended Gauss–Markov estimation (or prediction) is shown to be equivalent to a regularization method and its minimaxity is addressed. The resulting Gauss–Markov estimation serves as an oracle to guide the exploration for effective nonlinear estimators adaptively. Various examples are discussed. Particularly, the wavelet nonparametric regression example and its connection with a Sobolev regularization is presented.

AMS subject classifications: 62G05, 62H12.

Key words and phrases: nonparametric mixed-effects; Gauss–Markov theorem; best linear unbiased prediction (BLUP); regularization; minimaxity; normal equations; nonparametric regression; wavelet shrinkage; deconvolution.

1. INTRODUCTION

Consider a process Y observed through the model

$$Y = Af + \sigma\varepsilon. \tag{1.1}$$

The function of interest, f, is defined on an index set $\mathscr{T}$. The index set $\mathscr{T}$ can be an interval, a set of finite elements, or a set of countable elements. The error process ε is zero mean and is defined on an index set $\mathscr{J}$ with a known covariance kernel $\mathscr{R}$. A is a linear mapping from $L_2[\mathscr{T}]$ to $L_2[\mathscr{J}]$. When A is the identity mapping, the signal f is observed directly with noise.

0047-259X/01 $35.00

Otherwise, the signal is observed indirectly through A with noise. The underlying function f is modeled via nonparametric mixed-effects. That is,

$$f(t) = \sum_{k=1}^{m} \beta_k \phi_k(t) + \delta Z(t), \qquad t \text{ in } \mathcal{T}, \tag{1.2}$$

where $\phi_k(t)$ are known functions, β_k are fixed but unknown coefficients, $Z(t)$ is a zero mean process with a covariance kernel $E[Z(t)\,Z(s)] = \mathscr{W}(t, s)$ and $Z(t)$ is independent of ε. The ratio δ/σ is assumed known.[1]

The mixed-effects models described in (1.1) and (1.2) are often used in the analysis of longitudinal data or curve data. (See Laird and Ware, 1982; Ramsay, 1982; Besse and Ramsay, 1986; Ramsay and Dalzell, 1991; Anderson and Jones, 1995; among many others.) They also appear in the literature of spline smoothing and nonparametric Bayesian regression (Kimeldorf and Wahba, 1970 and 1971; Wahba, 1978 and 1990; and Barry, 1986). In this article we extend the Gauss–Markov Theorem for linear mixed-effects models to the nonparametric mixed-effects models. In contrast to using traditional linear algebra approach, we use analysis tools involving reproducing kernel Hilbert spaces. We find it quite natural to describe the space of fixed effects and the space of random effects via subspaces of a certain reproducing kernel Hilbert space (RKHS) and the penalty in the associated regularization can also be naturally represented by a semi-norm of the RKHS. The technical development of extended Gauss–Markov Theorem provides a unified and illuminating perspective for constructing linear estimators or predictors for various models. There are often parameters involved in the extended Gauss–Markov Theorem. When there is no prior knowledge about these parameters, one needs to estimate them based upon the data. Is then the extended Gauss–Markov Theorem of no practical value? The answer is no. The extended Gauss–Markov Theorem provides an oracle! Although it results in a linear estimator which often involves with unknown parameters, this linear estimator can serve as a useful guidance for exploring proper and effective nonlinear estimators. The reader is referred to Huang and Lu (2000) for an application of the extended Gauss–Markov Theorem to wavelet nonparametric regression.

The article is organized as follows. In Section 2, the Gauss–Markov estimation is extended to a more general setting to include nonparametric mixed-effects. The extended Gauss–Markov estimation is the so called best linear unbiased prediction (BLUP) because it is linear, unbiased and it

[1] Knowledge of the ratio δ/σ together with the identifiability condition for β (in Section 2) will be necessary and sufficient for the identifiability of the nonparametric mixed-effects model discussed later. One may assume that both σ and δ are known and set δ to one without loss of generality. But then the modeling assumption will be more restricted than only assuming the knowledge of the ratio.

minimizes the mean square error. In Section 3, a regularization method via functional penalized least squares (PLS) is described. The BLUP can be obtained as a functional PLS estimator, where the minimaxity of this regularization method is connected to that in Li (1982). There is also an intrinsic linkage between the functional PLS regularization and the Sobolev regularization in wavelet shrinkage presented in Section 3. The normal equations are derived in Section 4. In Section 5, a generalization to the case that the signal f is observed through an affine mapping is investigated. Illustration examples and discussion are given in Section 6.

2. THE EXTENDED GAUSS–MARKOV THEOREM

2.1. *Preliminaries.* A couple of notations and properties are introduced below.

- Let $\mathscr{H}_0$ be the Hilbert space linearly spanned by $\{\phi_k(t), k=1, ..., m\}$ and equipped with the $L_2[\mathscr{T}]$-norm. (The specific norm chosen here will not affect the result of Gauss–Markov estimation nor the corresponding regularization discussed later. It is simply a choice of convenience for writing down normal equations in Section 4.)
- Let $\mathscr{H}_{\mathscr{W}}$ be the RKHS generated by the covariance kernel $\mathscr{W}$ of the process $Z(t)$.
- Define $\mathscr{H}_1 = \mathscr{H}_0 \oplus \mathscr{H}_{\mathscr{W}}$, which is a RKHS.
- Let $\mathscr{H}_{\mathscr{R}}$ be the RKHS generated by the error covariance kernel $\mathscr{R}$. Assume that $\mathscr{R}$ has a positive discrete spectrum, i.e., an eigenfunction-eigenvalue decomposition with all (countably many) eigenvalues positive.
- Now A is a linear mapping from $L_2[\mathscr{T}]$ to $L_2[\mathscr{J}]$. When dealing with fixed effects, we confine A to be from $\mathscr{H}_0$ to $\mathscr{H}_{\mathscr{R}}$ and use the notation $A_{[\mathscr{H}_0 \to \mathscr{H}_{\mathscr{R}}]}$ for the operator itself and use notation $A^*_{[\mathscr{H}_{\mathscr{R}} \to \mathscr{H}_0]}$ for its adjoint. When dealing with random effects, we confine A to be from $\mathscr{H}_{\mathscr{W}}$ to $\mathscr{H}_{\mathscr{R}}$ and use the notation $A_{[\mathscr{H}_{\mathscr{W}} \to \mathscr{H}_{\mathscr{R}}]}$ for the operator itself and use notation $A^*_{[\mathscr{H}_{\mathscr{R}} \to \mathscr{H}_{\mathscr{W}}]}$ for its adjoint. The adjointness of a pair of operators depends on the norms employed in the involved linear spaces. The reader should be cautioned that the $\mathscr{H}_{\mathscr{R}}$-adjointness is not equal to the $L_2[\mathscr{J}]$-adjointness. For instance, one can consider the above bounded linear mapping $A_{[\mathscr{H}_1 \to \mathscr{H}_{\mathscr{R}}]}$. The $\mathscr{H}_{\mathscr{R}}$-adjoint operator, $A^*_{[\mathscr{H}_{\mathscr{R}} \to \mathscr{H}_1]}$, has the property that for f in $\mathscr{H}_1$ and h in $\mathscr{H}_{\mathscr{R}}$,

$$\langle A^*_{[\mathscr{H}_{\mathscr{R}} \to \mathscr{H}_1]} h, f \rangle_{\mathscr{H}_1} = \langle h, Af \rangle_{\mathscr{H}_{\mathscr{R}}},$$

while the $L_2[\mathscr{J}]$-adjoint operator, $A^*_{[L_2(\mathscr{J})\to\mathscr{H}_1]}$, means that the adjoint operator is derived as if the linear space $\mathscr{H}_{\mathscr{R}}$ is equipped with the $L_2[\mathscr{J}]$-norm. The adjoint operator $A^*_{[L_2(\mathscr{J})\to\mathscr{H}_1]}$ has the property that for f in $\mathscr{H}_1$ and h in $\mathscr{H}_{\mathscr{R}}$,

$$\langle A^*_{[L_2(\mathscr{J})\to\mathscr{H}_1]}h, f\rangle_{\mathscr{H}_1} = \langle h, Af\rangle_{L_2[\mathscr{J}]}.$$

- Let $L_{[\mathscr{H}]}$ denote an arbitrary bounded linear functional on an arbitrary RKHS $\mathscr{H}$ of functions on $\mathscr{T}$. The symbol $L_{[t,\mathscr{H}]}$ may be also used to denote its dependence on a particular t in $\mathscr{T}$.
- Let $\ell_{[t,\mathscr{H}]}$ denote the evaluation functional on $\mathscr{H}$ defined by $\ell_{[t,\mathscr{H}]}(h) = h(t)$ for h in $\mathscr{H}$.
- Let $\mathscr{M}$ be the reproducing kernel of a RKHS $\mathscr{H}$. The notation $(L_{1,[\mathscr{H}]}\otimes L_{2,[\mathscr{H}]})\mathscr{M}$ is used to denote the bi-linear functional

$$L^{(s)}_{1,[\mathscr{H}]}L^{(t)}_{2,[\mathscr{H}]}\mathscr{M}(s,t),$$

where $L^{(\cdot)}_{i,[\mathscr{H}]}$ means $L_{i,[\mathscr{H}]}$ is applied to what follows as a function of $(\cdot)$.

The subscripts may be suppressed from various symbols to keep notation simple when there is no ambiguity.

2.2. *Extended Gauss–Markov Theorem.* Adopting the conventional terminology, the estimators of random effects are *predictors* and the estimators of fixed effects are *estimators*. If there is no ambiguity, estimators or predictors are used without distinction.

DEFINITION. A predictor $\hat{f}(t)$ is the best linear unbiased prediction (BLUP) for $f(t)$ if and only if

- $\hat{f}(t)$ is linear in Y in the sense that $\hat{f}(t)$ can be represented as $L_{[t,\mathscr{H}_{\mathscr{R}}]}Y$,
- $\hat{f}(t)$ is unbiased in the sense that, $E\hat{f}(t) = Ef(t)$ for all t in $\mathscr{T}$, and
- $\hat{f}(t)$ has the minimum mean square error, that is, $E(\hat{f}(t)-f(t))^2 \leqslant E(\tilde{f}(t)-f(t))^2$ for all t in $\mathscr{T}$, among all linear unbiased estimators $\tilde{f}(t)$.

The covariance kernel for random effects is supposed to satisfy the conditions:

$$\int_{\mathscr{T}} \mathscr{W}(t,t)\,dt < \infty \qquad \text{and} \qquad \int_{\mathscr{T}}\int_{\mathscr{T}} \mathscr{W}^2(t,s)\,ds\,dt < \infty. \tag{2.1}$$

If $\mathscr{T}$ is a finite set or a set of countable elements, the Lebesgue measure dt should be replaced by the counting measure. The condition $\int_{\mathscr{T}} \mathscr{W}(t,t)\,dt < \infty$ ensures that the sample path of $Z(t)$ is in $L_2[\mathscr{T}]$ *a.s.* Meanwhile, the

condition $\int_{\mathscr{T}}\int_{\mathscr{T}} \mathscr{W}^2(t,s)\,ds\,dt < \infty$ ensures that the kernel $\mathscr{W}$ has an eigenfunction-eigenvalue decomposition by the Hilbert–Schmidt theorem (Reed and Simon, 1972; Wahba, 1990). Thus, there exists a complete orthonormal basis of $L_2[\mathscr{T}]$, $\{\psi_1(t), \psi_2(t), ...\}$, and eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant 0$ such that $\mathscr{W}$ has the decomposition

$$\mathscr{W}(s,t) = \sum_{\nu} \lambda_\nu \psi_\nu(s)\,\psi_\nu(t).$$

Thus, the process $Z(t)$ has a representation, the so called Karhunen–Loève representation,

$$Z(t) \sim \sum_{\nu} \gamma_\nu \psi_\nu(t), \tag{2.2}$$

where "$\sim$" means "equal in distribution", $\gamma_\nu = \int_{\mathscr{T}} Z(t)\,\psi_\nu(t)\,dt$, $\nu = 1, 2, ...$ are a sequence of uncorrelated random variables with zero means and variances $\lambda_1, \lambda_2, ...$

When there are only finitely many non-zero eigenvalues, the prior model (1.2) reduces to a linear parametric mixed-effects model:

$$f(t) = \sum_{k=1}^{m} \beta_k \phi_k(t) + \sum_{\nu=1}^{J} \delta\gamma_\nu \psi_\nu(t), \qquad t \text{ in } \mathscr{T}. \tag{2.3}$$

The Gauss–Markov Theorem for parametric models (i.e., the BLUE for fixed-effects models or the BLUP for mixed- or random-effects models) has been discussed, for instance, in Harville (1976) and Robinson (1991). As for the nonparametric model, there are discussions in the literature of spline smoothing (Kimeldorf and Wahba, 1971; Wahba, 1978; Wang, 1998). This article complements as well as generalizes the Gauss–Markov estimation in Kimeldorf and Wahba (1971, Section 7), Wahba (1978) and Harville (1976). It also aims to provide an interesting link to the work by Li (1982). Following is a key lemma for establishing the Gauss–Markov Theorem later.

LEMMA 2.1. *Let $\mathscr{H}_{\mathscr{M}}$ be a RKHS of functions defined on index set $\mathscr{J}$ with kernel $\mathscr{M}$. Assume the kernel $\mathscr{M}$ has a positive discrete spectrum: $\mathscr{M}(\zeta,\zeta') = \sum_j \lambda_j^{(\mathscr{M})} h_j(\zeta)\,h_j(\zeta')$, where $\lambda_1^{(\mathscr{M})} \geqslant \lambda_2^{(\mathscr{M})} \geqslant \cdots > 0$ and $\{h_j\colon j \geqslant 0\}$ forms a complete orthonormal basis for $L_2[\mathscr{J}]$. Let $\mathscr{H}$ be a RKHS of functions defined on index set $\mathscr{T}$ and $A_{[\mathscr{H}\to\mathscr{H}_{\mathscr{M}}]}$ be a one-to-one linear mapping from $\mathscr{H}$ to $\mathscr{H}_{\mathscr{M}}$. Then the unique solution to the following minimization problem*

$$\min_{L_{[\mathscr{H}_{\mathscr{M}}]}} (L_{[\mathscr{H}_{\mathscr{M}}]} \otimes L_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M}, \qquad \text{subject to} \quad L_{[\mathscr{H}_{\mathscr{M}}]} A_{[\mathscr{H}\to\mathscr{H}_{\mathscr{M}}]} = \ell_{[t,\mathscr{H}]} \tag{2.4}$$

is given by

$$L^{opt}_{[\mathscr{H}_{\mathscr{M}}]} = \ell_{[t,\mathscr{H}]}(A^*_{[\mathscr{H}_{\mathscr{M}}\to\mathscr{H}]}A_{[\mathscr{H}\to\mathscr{H}_{\mathscr{M}}]})^{-1}A^*_{[\mathscr{H}_{\mathscr{M}}\to\mathscr{H}]}.$$

Proof. Let $L_{[\mathscr{H}_{\mathscr{M}}]}$ be an arbitrary bounded linear functional on $\mathscr{H}_{\mathscr{M}}$ satisfying the constraint $L_{[\mathscr{H}_{\mathscr{M}}]}A_{[\mathscr{H}\to\mathscr{H}_{\mathscr{M}}]} = \ell_{[t,\mathscr{H}]}$ and put $\Delta_{[\mathscr{H}_{\mathscr{M}}]} = L_{[\mathscr{H}_{\mathscr{M}}]} - L^{opt}_{[\mathscr{H}_{\mathscr{M}}]}$. Then $\Delta_{[\mathscr{H}_{\mathscr{M}}]}A_{[\mathscr{H}\to\mathscr{H}_{\mathscr{M}}]} = 0$, a zero linear functional on $\mathscr{H}$. It is true that

$$\begin{aligned}(L_{[\mathscr{H}_{\mathscr{M}}]}\otimes L_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M} &= (L^{opt}_{[\mathscr{H}_{\mathscr{M}}]}\otimes L^{opt}_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M} + (\Delta_{[\mathscr{H}_{\mathscr{M}}]}\otimes\Delta_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M}\\ &\quad + 2(L^{opt}_{[\mathscr{H}_{\mathscr{M}}]}\otimes\Delta_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M}\\ &= (L^{opt}_{[\mathscr{H}_{\mathscr{M}}]}\otimes L^{opt}_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M} + (\Delta_{[\mathscr{H}_{\mathscr{M}}]}\otimes\Delta_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M},\end{aligned}$$

if one can show that $(L^{opt}_{[\mathscr{H}_{\mathscr{M}}]}\otimes\Delta_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M} = 0$. Then the minimum is achieved by taking $\Delta_{[\mathscr{H}_{\mathscr{M}}]} = 0$. Therefore, it is left to show that $(L^{opt}_{[\mathscr{H}_{\mathscr{M}}]}\otimes\Delta_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M} = 0$.

Let $d_j = \Delta_{[\mathscr{H}_{\mathscr{M}}]}h_j$. Then

$$\begin{aligned}(L^{opt}_{[\mathscr{H}_{\mathscr{M}}]}\otimes\Delta_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M} &= (\ell_{[t,\mathscr{H}]}(A^*A)^{-1}A^*\otimes\Delta_{[\mathscr{H}_{\mathscr{M}}]})\,\mathscr{M}\\ &= \ell_{[t,\mathscr{H}]}(A^*A)^{-1}\left(\sum d_j\lambda_j^{(\mathscr{M})}A^*h_j\right).\end{aligned}$$

The term $\sum d_j\lambda_j^{(\mathscr{M})}A^*_{[\mathscr{H}_{\mathscr{M}}\to\mathscr{H}]}h_j$ is an element in $\mathscr{H}$ with the following property. For every $\phi\in\mathscr{H}$,

$$\begin{aligned}\left\langle\sum d_j\lambda_j^{(\mathscr{M})}A^*h_j,\phi\right\rangle_{\mathscr{H}} &= \left\langle\sum d_j\lambda_j^{(\mathscr{M})}h_j, A\phi\right\rangle_{\mathscr{H}_{\mathscr{M}}}\\ &= \left\langle\sum d_jh_j, A\phi\right\rangle_{L_2[\mathscr{I}]} = \Delta_{[\mathscr{H}_{\mathscr{M}}]}A_{[\mathscr{H}\to\mathscr{H}_{\mathscr{M}}]}\phi = 0.\end{aligned}$$

Thus, $\sum d_j\lambda_j^{(\mathscr{M})}A^*_{[\mathscr{H}_{\mathscr{M}}\to\mathscr{H}]}h_j = 0$ and then $(L^{opt}_{[\mathscr{H}_{\mathscr{M}}]}\otimes\Delta_{[\mathscr{H}_{\mathscr{M}}]})\mathscr{M} = 0$. Q.E.D.

Based on the eigenfunction-eigenvalue decomposition of $\mathscr{M}$, one can define M to be the bounded linear operator on $\mathscr{H}_{\mathscr{M}}$ by $Mh_j = \lambda_j^{(\mathscr{M})}h_j$. That is, M is induced by the kernel $\mathscr{M}$. Since that $A^*_{[\mathscr{H}_{\mathscr{M}}\to\mathscr{H}]} = A^*_{[L_2(\mathscr{I})\to\mathscr{H}]}M^{-1}$, Lemma 2.1 can also be expressed as below.

LEMMA 2.1*. *The solution for the minimization problem* (2.4) *can also be represented as*

$$L^{opt}_{[\mathscr{H}_{\mathscr{M}}]} = \ell_{[t,\mathscr{H}]}(A^*_{[L_2(\mathscr{I})\to\mathscr{H}]}M^{-1}A_{[\mathscr{H}\to\mathscr{H}_{\mathscr{M}}]})^{-1}A^*_{[L_2(\mathscr{I})\to\mathscr{H}]}M^{-1}.$$

A special case of Lemma 2.1 (or Lemma 2.1*), which appears constantly in linear regression problems, is given below. It can be viewed as a matrix analogue of Lemma 2.1 (or Lemma 2.1*). Let L be an n-vector, x be an m-vector, X be an $n \times m$ full rank matrix with $m < n$ and M be an $n \times n$ positive definite matrix. The unique solution to the following minimization problem

$$\min_{L \in R^n} L^T M L \qquad \text{subject to} \quad L^T X = x^T$$

is given by

$$L_{opt}^T = x^T (X^T M^{-1} X)^{-1} X^T M^{-1}.$$

From the above key lemma, the main theorem on Gauss–Markov estimation can be proved as shown below. First, there are identifiability and boundedness (continuity) conditions needed.

Identifiability condition. $A\phi_k$, $k = 1, ..., m$, are linearly independent. (This condition ensures that β_k are identifiable.)

Boundedness (continuity) condition I. The linear mapping A is bounded on $\mathscr{H}_1$.

THEOREM 2.2 (Extended Gauss–Markov Theorem). *The observation data come from the model of* (1.1) *and* (1.2) *with* $\mathscr{R}$ *having a positive discrete spectrum and* $\mathscr{W}$ *satisfying* (2.1). *Assume that the identifiability condition and the boundedness condition I are met. Then the BLUP for f is given by* $\hat{f}_{BLUP} = \hat{f}_{FE} + \hat{f}_{RE}$ *with*

$$\hat{f}_{FE} = (A_0^* M^{-1} A_0)^{-1} A_0^* M^{-1} Y,$$
$$\hat{f}_{RE} = \alpha^{-1} W A_{\mathscr{W}}^* M^{-1} (Y - A_0 \hat{f}_{FE}),$$

where

$$A_0 = A_{[\mathscr{H}_0 \to L_2(\mathscr{I})]},$$
$$A_0^* = A^*_{[L_2(\mathscr{I}) \to \mathscr{H}_0]},$$
$$A_{\mathscr{W}} = A_{[\mathscr{H}_{\mathscr{W}} \to L_2(\mathscr{I})]},$$
$$A_{\mathscr{W}}^* = A^*_{[L_2(\mathscr{I}) \to \mathscr{H}_{\mathscr{W}}]},$$

$\alpha^{-1} = \delta^2/\sigma^2$, M *is the linear operator induced by the kernel* $\mathscr{M} = \mathscr{R} + \alpha^{-1} \mathscr{W}_A$ *with* $\mathscr{W}_A = A^{(s)} A^{(t)} \mathscr{W}(s, t)$, *and* W *is the linear operator induced by the kernel* $\mathscr{W}$.

Proof. Let $L_{[t,\mathscr{H}_{\mathscr{R}}]}Y$ be an unbiased linear estimator for $f(t)$ with $L_{[t,\mathscr{H}_{\mathscr{R}}]}$ satisfying $L_{[t,\mathscr{H}_{\mathscr{R}}]}A_{[\mathscr{H}_0\to\mathscr{H}_{\mathscr{R}}]}=\ell_{[t,\mathscr{H}_0]}$. There are two possible cases.

Case I. $A_{\mathscr{W}}$ is one to one and onto. Then the mean square error of $L_{[t,\mathscr{H}_{\mathscr{R}}]}Y$ becomes

$$\begin{aligned}E(L_{[t,\mathscr{H}_{\mathscr{R}}]}Y-f(t))^2 &= E(L(Af+\sigma\varepsilon)-f(t))^2\\ &= E(L(\delta A_{\mathscr{W}}Z+\sigma\varepsilon)-\delta Z(t))^2\\ &= E(L(\delta A_{\mathscr{W}}Z+\sigma\varepsilon)-\delta\ell_{[t,\mathscr{H}_1]}(A_{\mathscr{W}}^*A_{\mathscr{W}})^{-1}A_{\mathscr{W}}^*A_{\mathscr{W}}Z)^2\\ &= \sigma^2(L\otimes L)\mathscr{M}-2\delta^2(L\otimes\ell_{[t,\mathscr{H}_1]}(A_{\mathscr{W}}^*A_{\mathscr{W}})^{-1}A_{\mathscr{W}}^*)\mathscr{W}_A\\ &\quad+\delta^2\mathscr{W}(t,t),\end{aligned}$$

where $\mathscr{W}_A$ is the covariance kernel of AZ. Since that $\mathscr{R}$ has a positive discrete spectrum and $\mathscr{W}_A$ has a non-negative discrete spectrum, $\mathscr{M}$ has a positive discrete spectrum. Hence, it is straightforward to see that

$$\mathscr{W}_A(\zeta,\zeta')=W_A^{(\zeta')}M^{-1(\zeta')}\mathscr{M}(\zeta,\zeta'),$$

where W_A is the linear operator induced by the kernel $\mathscr{W}_A$. One then has

$$\begin{aligned}&(L\otimes L)\mathscr{M}-2\alpha^{-1}(L\otimes\ell_{[t,\mathscr{H}_1]}(A_{\mathscr{W}}^*A_{\mathscr{W}})^{-1}A_{\mathscr{W}}^*)\mathscr{W}_A\\ &\quad=(L\otimes L)\mathscr{M}-2\alpha^{-1}(L\otimes\ell_{[t,\mathscr{H}_1]}(A_{\mathscr{W}}^*A_{\mathscr{W}})^{-1}A_{\mathscr{W}}^*W_AM^{-1})\mathscr{M}\\ &\quad=(L_*\otimes L_*)\mathscr{M}-\alpha^{-2}((\ell_{[t,\mathscr{H}_1]}(A_{\mathscr{W}}^*A_{\mathscr{W}})^{-1}A_{\mathscr{W}}^*W_AM^{-1})\\ &\qquad\otimes(\ell_{[t,\mathscr{H}_1]}(A_{\mathscr{W}}^*A_{\mathscr{W}})^{-1}A_{\mathscr{W}}^*W_AM^{-1}))\mathscr{M},\end{aligned}$$

where

$$\begin{aligned}L_* &= L-\alpha^{-1}\ell_{[t,\mathscr{H}_1]}(A_{\mathscr{W}}^*A_{\mathscr{W}})^{-1}A_{\mathscr{W}}^*W_AM^{-1}\\ &= L-\alpha^{-1}\ell_{[t,\mathscr{H}_1]}WA_{\mathscr{W}}^*M^{-1}.\end{aligned}$$

It is noted that L_*A_0, as a linear functional on $\mathscr{H}_0$, satisfies the constraint

$$L_*A_0\phi=(\ell_{[t,\mathscr{H}_0]}-\alpha^{-1}\ell_{[t,\mathscr{H}_1]}WA_{\mathscr{W}}^*M^{-1}A_0)\phi$$

for $\phi\in\mathscr{H}_0$. By Lemma 2.1*, the mean square error $E(L_{[t,\mathscr{H}_{\mathscr{R}}]}Y-f(t))^2$ is minimized by taking

$$L_*=(\ell_{[t,\mathscr{H}_0]}-\alpha^{-1}\ell_{[t,\mathscr{H}_1]}WA_{\mathscr{W}}^*M^{-1}A_0)(A_0^*M^{-1}A_0)^{-1}A_0^*M^{-1}.$$

Then, the optimal L is given by

$$L_{[t,\,\mathscr{H}_{\mathscr{R}}]} = \ell_{[t,\,\mathscr{H}_0]}(A_0^* M^{-1} A_0)^{-1} A_0^* M^{-1} + \alpha^{-1} \ell_{[t,\,\mathscr{H}_1]} W A_{\mathscr{W}}^* M^{-1} (I - A_0 (A_0^* M^{-1} A_0)^{-1} A_0^* M^{-1}),$$

where I is the identity mapping on $\mathscr{H}_{\mathscr{R}}$. Therefore,

$$L_{[t,\,\mathscr{H}_{\mathscr{R}}]} Y = \hat{f}_{FE}(t) + \hat{f}_{RE}(t).$$

Case II. $A_{\mathscr{W}}$ is not one to one and onto. Let $\mathscr{H}_{\mathscr{W}}^{\text{null}}$ be the null space of $A_{\mathscr{W}}$ and let $(\mathscr{H}_{\mathscr{W}}^{\text{null}})^{\perp}$ be its orthogonal complement. Then Z has a unique decomposition $Z = Z_1 + Z_2$ with Z_1 in the null space and Z_2 in the complement. Now $A_{\mathscr{W}}$ is one to one and onto in $(\mathscr{H}_{\mathscr{W}}^{\text{null}})^{\perp}$. The best linear unbiased prediction of Z_2 proceeds as Case I above. Meanwhile, it is not hard to see that the BLUP of Z_1 is simply zero. Therefore, Theorem 2.2 also holds for Case II. Q.E.D.

The extension of Gauss–Markov type estimation to include nonparametric models can be traced back to the work of Kimeldorf and Wahba (1971). They considered a special model of (1.1) and (1.2) with random errors $\varepsilon \sim N(0, R)$, random coefficients $\beta \sim N(0, I)$, and a random component $Z(t)$ which is a zero mean Gaussian process with a known covariance kernel, where $Z(t)$, ε and β are stochastically independent. There is a subtle difference between their approach and ours in dealing with the parameters β. In Kimeldorf and Wahba (1971), the definition of unbiased estimation for nonparametric models is the same as ours, i.e., it is conditioned on a fixed but arbitrary β. In addition to the unbiasedness, the criterion for Gauss–Markov type estimation is to minimize the mean square error for estimation of f. In their work, the parameters β are assigned a standard normal distribution and the MSE is averaged over the distribution of β. The Gauss–Markov type estimation so derived is guaranteed to be unbiased for every fixed β. However, the minimum MSE is not guaranteed for every fixed β, but only guaranteed for averaging over β according to its distribution. Our approach guarantees the unbiasedness and minimum MSE for every fixed but otherwise arbitrary β.

In Wahba (1978), $\beta \sim N(0, \xi I)$ and ξ is let to approach infinity. For fixed σ^2, δ^2, and ξ, let $E_\xi(f(t) \mid Y)$ denote the posterior mean of $f(t)$ given Y. Under normality assumptions, the posterior mean $E_\xi(f(t) \mid Y)$ is the BLUP in the sense that both the unbiasedness and the mean square error are averaged over the distribution of β instead of conditioning on a fixed value of β. By letting $\xi \to \infty$, the resulting estimate $\lim_{\xi \to \infty} E_\xi(f(t) \mid Y)$ is BLUP at *design* points (Speed 1991). This "BLUP at design points" phenomenon is re-illustrated in Wang (1998). Theorem 2.2 fills in the blanks for non-design points. Furthermore, the BLUP for the bounded linear functional of f can be derived similarly as in the next corollary.

COROLLARY 2.3. *Assume the conditions of Theorem* 2.2. *Let* $L_{[\mathscr{H}_1]}$ *be a bounded linear functional on* $\mathscr{H}_1$. *Then the BLUP for* $L_{[\mathscr{H}_1]} f$ *is* $L_{[\mathscr{H}_1]}(\hat{f}_{BLUP})$.

3. THE BLUP AND THE REGULARIZATION

The following regularization method is investigated,

$$\min_{f \in \mathscr{H}_1} -2 \langle Y, Af \rangle_{\mathscr{H}_\mathscr{R}} + \langle Af, Af \rangle_{\mathscr{H}_\mathscr{R}} + \alpha \| P_{[\mathscr{H}_1 \to \mathscr{H}_\mathscr{W}]} f \|^2_{\mathscr{H}_\mathscr{W}}, \qquad (3.1)$$

where $P_{[\mathscr{H}_1 \to \mathscr{H}_\mathscr{W}]} f$ is the projection of f in $\mathscr{H}_1$ onto $\mathscr{H}_\mathscr{W}$. Although often that the sample path of Y does not belong to $\mathscr{H}_\mathscr{R}$, $\langle Y, h \rangle_{\mathscr{H}_\mathscr{R}}$ is a well-defined random variable for every h in $\mathscr{H}_\mathscr{R}$ because the error process ε with covariance kernel $\mathscr{R}$ has the Karhunen–Loève representation. Conventionally, the above method in (3.1) can be regarded as a functional penalized least square (PLS) method:

$$\min_{f \in \mathscr{H}_1} \| Y - Af \|^2_{\mathscr{H}_\mathscr{R}} + \alpha \| Pf \|^2_{\mathscr{H}_\mathscr{W}}. \qquad (3.2)$$

The above regularization can be connected with that discussed in Li (1982). Let $A_{2,[\mathscr{H}_1 \to \mathscr{H}_2]}$ be a bounded linear mapping from $\mathscr{H}_1$ onto $\mathscr{H}_2$, a RKHS, with the null space $\mathscr{H}_0$. Our regularization method (3.1) or (3.2) penalizes directly on the $\mathscr{H}_\mathscr{W}$-norm of f, while Li's penalizes on the $\mathscr{H}_2$-norm of the transformed f by A_2:

$$\min_{f \in \mathscr{H}_1} -2 \langle Y, Af \rangle_{\mathscr{H}_\mathscr{R}} + \langle Af, Af \rangle_{\mathscr{H}_\mathscr{R}} + \alpha \| A_2 f \|^2_{\mathscr{H}_2}. \qquad (3.3)$$

It is noted that the two spaces, $\mathscr{H}_\mathscr{W}$ and $\mathscr{H}_2$, are topologically isomorphic via A_2. Similar to Li (1982), the following boundedness condition is also necessary.

Boundedness Condition II. Let $L_{A,Y}$ be a linear functional on $\mathscr{H}_1$ defined by $L_{A,Y}(g) = \langle Y, Ag \rangle_{\mathscr{H}_\mathscr{R}}$. Assume that $L_{A,Y}$ is bounded for the realization of Y.

THEOREM 3.1. *It is assumed that the identifiability condition as well as the two boundedness conditions* I *and* II *hold. Then the unique solution to problem* (3.1) *is the BLUP,* $\hat{f}_{BLUP}$, *in Theorem* 2.2 *with* $\alpha = \sigma^2/\delta^2$.

Proof. Let the minimizer of (3.1) be of the form

$$f_\alpha = f_0 + WA^*_\mathscr{W} M^{-1} h + \rho,$$

where f_0 is in $\mathscr{H}_0$, h is in $\mathscr{H}_{\mathscr{R}}$ and ρ is an element in $\mathscr{H}_{\mathscr{W}}$ which is orthogonal to $WA^*_{\mathscr{W}}M^{-1}(\mathscr{H}_{\mathscr{R}})$. Any element in $\mathscr{H}_1$ has such a representation. It is true that

$$\rho \perp WA^*_{\mathscr{W}}M^{-1}(\mathscr{H}_{\mathscr{R}}) \qquad \text{if and only if} \quad \rho \perp A^*_{\mathscr{W}}M^{-1}(\mathscr{H}_{\mathscr{R}}).$$

Then for every h in $\mathscr{H}_{\mathscr{R}}$ one has

$$\langle A\rho, M^{-1}h\rangle_{\mathscr{H}_{\mathscr{R}}} = \langle \rho, A^*_{\mathscr{W}}M^{-1}h\rangle_{\mathscr{H}_{\mathscr{W}}} = 0.$$

Thus $A\rho = 0$ and (3.2) becomes

$$\begin{aligned}&\|Y - Af\|^2_{\mathscr{H}_{\mathscr{R}}} + \alpha\,\|Pf\|^2_{\mathscr{H}_{\mathscr{W}}}\\ &\quad = \|Y - A(f_0 + WA^*_{\mathscr{W}}M^{-1}h)\|^2_{\mathscr{H}_{\mathscr{R}}} + \alpha\,\|WA^*_{\mathscr{W}}M^{-1}h + \rho\|^2_{\mathscr{H}_{\mathscr{W}}}.\end{aligned}$$

It is aimed to find f_0, h and ρ to minimize the above expression. It is evident that ρ must be zero. For an arbitrary fixed f_0 in $\mathscr{H}_0$, let $\tilde{Y} = Y - Af_0$. The minimization problem now becomes

$$\min_{h\in\mathscr{H}_{\mathscr{R}}} \|\tilde{Y} - AWA^*_{\mathscr{W}}M^{-1}h\|^2_{\mathscr{H}_{\mathscr{R}}} + \alpha\,\|WA^*_{\mathscr{W}}M^{-1}h\|^2_{\mathscr{H}_{\mathscr{W}}}. \tag{3.4}$$

To solve the minimization problem (3.4), one considers $h = \alpha^{-1}(\tilde{Y} + h_{\Delta})$, where h_{Δ} is an arbitrary element in $\mathscr{H}_{\mathscr{R}}$. Then

$$\begin{aligned}&\|\tilde{Y} - AWA^*_{\mathscr{W}}M^{-1}h\|^2_{\mathscr{H}_{\mathscr{R}}} + \alpha\,\|WA^*_{\mathscr{W}}M^{-1}h\|^2_{\mathscr{H}_{\mathscr{W}}}\\ &\quad = \|(I - \alpha^{-1}AWA^*_{\mathscr{W}}M^{-1})\,\tilde{Y} - \alpha^{-1}AWA^*_{\mathscr{W}}M^{-1}h_{\Delta}\|^2_{\mathscr{H}_{\mathscr{R}}}\\ &\qquad + \alpha^{-1}\,\|WA^*_{\mathscr{W}}M^{-1}\tilde{Y} + WA^*_{\mathscr{W}}M^{-1}h_{\Delta}\|^2_{\mathscr{H}_{\mathscr{W}}}.\end{aligned}$$

After expanding the squared norms $\|\cdot\|^2_{\mathscr{H}_{\mathscr{R}}}$ and $\|\cdot\|^2_{\mathscr{H}_{\mathscr{W}}}$, the sum of cross terms in the above expression is zero since

$$\begin{aligned}&-2\,\langle (I - \alpha^{-1}AWA^*_{\mathscr{W}}M^{-1})\,\tilde{Y}, \alpha^{-1}AWA^*_{\mathscr{W}}M^{-1}h_{\Delta}\rangle_{\mathscr{H}_{\mathscr{R}}}\\ &\qquad + 2\alpha^{-1}\,\langle WA^*_{\mathscr{W}}M^{-1}\tilde{Y}, WA^*_{\mathscr{W}}M^{-1}h_{\Delta}\rangle_{\mathscr{H}_{\mathscr{W}}}\\ &\quad = -2\,\langle M^{-1}\tilde{Y}, \alpha^{-1}AWA^*_{\mathscr{W}}M^{-1}h_{\Delta}\rangle_{L_2[\mathscr{I}]}\\ &\qquad + 2\alpha^{-1}\,\langle A^*_{\mathscr{W}}M^{-1}\tilde{Y}, WA^*_{\mathscr{W}}M^{-1}h_{\Delta}\rangle_{L_2[\mathscr{I}]} = 0.\end{aligned}$$

Thus the minimum for (3.4) is achieved by taking $h_{\Delta} = 0$. That is, the solution for (3.4) is $\hat{h} = \alpha^{-1}\tilde{Y}$.

The estimate $\hat{h} = \alpha^{-1}\tilde{Y}$ is plugged into (3.4). Then one opts for f_0 that minimizes

$$\|(I - \alpha^{-1} A W A^*_{\mathscr{W}} M^{-1})(Y - Af_0)\|^2_{\mathscr{H}_{\mathscr{R}}} + \alpha^{-1} \|W A^*_{\mathscr{W}} M^{-1}(Y - Af_0)\|^2_{\mathscr{H}_{\mathscr{W}}}. \tag{3.5}$$

To solve the minimization problem (3.5), one considers $f_0 = \sum_{k=1}^{m} \beta_k \phi_k$. The derivatives of (3.5) are taken with respect to β_k and set to be zero. Then the optimal solutions for β_k are obtained by solving the resulting equations. The calculations are straightforward and the details are omitted. We will only state the result. The solution for the minimization problem (3.5) is given by $f_0 = \hat{f}_{FE}$. The estimate $f_0 = \hat{f}_{FE}$ is then plugged into the optimal solution $\hat{h}$. The solution to the minimization problem (3.2) turns out to be

$$f_\alpha = \hat{f}_{FE} + \alpha^{-1} W A^*_{\mathscr{W}} M^{-1}(Y - A\hat{f}_{FE}),$$

which is exactly the BLUP in Theorem 2.2. Q.E.D.

Remark 1 (Minimaxity). By Theorem 2.2 of Li (1982), $L_{[\mathscr{H}_1]}(\hat{f}_{BLUP})$ is the minimax linear estimator of $L_{[\mathscr{H}_1]} f$ for f in the class:

$$\{f : f \in \mathscr{H}_1 \text{ and } \|Pf\|_{\mathscr{H}_{\mathscr{W}}} \leqslant \delta\}.$$

The interesting connection lies between the ratio σ/δ in the model, the tuning parameter α in the regularization method, and the upper bound δ in restraining functional norm $\|Pf\|_{\mathscr{H}_{\mathscr{W}}}$ for linear minimaxity.

Remark 2 (Sobolev regularization for wavelet shrinkage). Consider the nonparametric regression problem, i.e., A is the identity mapping in (1.1). For every function f in $L_2[0, 1]$, it can be represented by wavelet series. We treat the scaling coefficients as fixed effects and wavelet coefficients as random effects. Adopt the usual notations for scaling functions and wavelets. Let $\phi(t)$ be a scaling function and $\psi(x)$ be the associated wavelet function. The dilations and shifts of ϕ and ψ are given by $\phi_{j,k}(x) = \sqrt{2^j}\,\phi(2^j x - k)$ and $\psi_{j,k}(x) = \sqrt{2^j}\,\psi(2^j x - k)$ respectively. For a fixed resolution level j, f has the wavelet expansion as

$$f(t) = f_{FE}(t) + f_{RE}(t) = \sum_k \beta_k \phi_{j,k}(t) + \delta \sum_{\ell \geqslant j} \sum_k \gamma_{\ell,k} \psi_{\ell,k}(t),$$

where $\beta_k = \langle f, \phi_{j,k} \rangle$ and $\delta\gamma_{\ell,k} = \langle f, \psi_{\ell,k} \rangle$. For scaling functions and wavelets on the interval, both $\sum_k$ above are finite sums with the number of elements depending on the resolution levels. The scaling coefficients β_k are fixed effects and the wavelet coefficients $\gamma_{\ell,k}$ are random effects with zero

mean and variance $\lambda_\ell \equiv E\gamma_{\ell,k}^2$ for each resolution level ℓ. Then $f(t)$ meets the prior model (1.2) with

$$\mathscr{W}(s,t) = \sum_{\ell \geqslant j} \sum_{k} \lambda_\ell \psi_{\ell,k}(t)\, \psi_{\ell,k}(s).$$

The corresponding regularization method in (3.1) or (3.2) has the following penalty

$$\alpha \, \|Pf\|_{\mathscr{H}_{\mathscr{W}}}^2 \equiv \alpha \sum_{\ell \geqslant j} \sum_{k} |\langle f, \psi_{\ell,k} \rangle|^2 / \lambda_\ell, \tag{3.6}$$

where α is the same as previously defined. With these particular choices of ϕ, ψ, penalty in (3.6) and $\lambda_\ell = O(2^{-2\ell s})$, the procedure by (3.1) is a Sobolev regularization (Huang and Lu, 2000).

4. THE NORMAL EQUATIONS

In this section, normal equations are derived from the regularization (3.1). The model (1.1), the identifiability condition and both boundedness conditions are assumed. The ϕ_ks are also assumed to be orthonormal because one can always orthonormalize them if they are not the case. The solution for (3.1) is of the form

$$f_\alpha(t) = \sum_{k=1}^{m} \beta_k \phi_k(t) + \sum_{v} \gamma_v \psi_v(t).$$

It is clear that

$$\begin{aligned} &-2 \langle Y, Af \rangle_{\mathscr{H}_{\mathscr{R}}} + \langle Af, Af \rangle_{\mathscr{H}_{\mathscr{R}}} + \alpha \, \|Pf\|_{\mathscr{H}_{\mathscr{W}}}^2 \\ &\quad = -2 \langle Y, Af \rangle_{\mathscr{H}_{\mathscr{R}}} + \langle A^*Af, f \rangle_{\mathscr{H}_1} + \alpha \langle P^*Pf, f \rangle_{\mathscr{H}_1}, \end{aligned}$$

where $A^* = A^*_{[\mathscr{H}_{\mathscr{R}} \to \mathscr{H}_1]}$ and $P^* = P^*_{[\mathscr{H}_{\mathscr{W}} \to \mathscr{H}_1]}$. By the boundedness assumption on $L_{A,Y}$ and the Riesz representation, there exists a unique $\tilde{h}_Y$ in $\mathscr{H}_1$ such that

$$L_{A,Y}(g) = \langle Y, Ag \rangle_{\mathscr{H}_{\mathscr{R}}} = \langle \tilde{h}_Y, g \rangle_{\mathscr{H}_1}$$

for all g in $\mathscr{H}_1$. Since $\tilde{h}_Y$ can be written as $\tilde{h}_Y = \sum_{k=1}^m a_k \phi_k + \sum_v b_v \psi_v$, the coefficients a_k and b_v can be calculated via

$$a_k = \langle \tilde{h}_Y, \phi_k \rangle_{\mathscr{H}_1} = \langle Y, A\phi_k \rangle_{\mathscr{H}_{\mathscr{R}}}, \qquad k = 1, ..., m$$

and

$$b_v = \lambda_v \langle \tilde{h}_Y, \psi_v \rangle_{\mathscr{H}_1} = \lambda_v \langle Y, A\psi_v \rangle_{\mathscr{H}_{\mathscr{R}}}, \qquad v = 1, 2, ...$$

By the arguments of Theorem 2.1 in Li (1982), $f_\alpha(t) = (\alpha P^*P + A^*A)^{-1}\,\tilde{h}_Y$. Then

$$f_\alpha(t) = (\alpha P^*P + A^*A)^{-1} \left(\sum_{k=1}^{m} \langle A\phi_k,\, Y\rangle_{\mathscr{H}_{\mathscr{R}}}\, \phi_k(t) + \sum_{v} \langle A\psi_v,\, Y\rangle_{\mathscr{H}_{\mathscr{R}}}\, \lambda_v \psi_v(t) \right).$$

Therefore, we have the following normal equations.

THEOREM 4.1. *Assume that ϕ_ks are orthonormal. The normal equations are given by*

$$\langle \phi_k, (A^*A + \alpha^{-1}P^*P)\, f\rangle_{\mathscr{H}_1} = \langle Y, A\phi_k\rangle_{\mathscr{H}_{\mathscr{R}}}, \qquad k = 1, ..., m, \tag{4.1}$$

$$\langle \psi_v, (A^*A + \alpha^{-1}P^*P)\, f\rangle_{\mathscr{H}_1} = \langle Y, A\psi_v\rangle_{\mathscr{H}_{\mathscr{R}}}, \qquad v = 1, 2, ... \tag{4.2}$$

5. GENERALIZATION TO AFFINE MAPPINGS

In this section, it is assumed that the signal f is observed through an affine mapping $\tilde{A}$ with random noise:

$$Y = \tilde{A}f + \sigma\varepsilon, \tag{5.1}$$

$$\tilde{A}f = \beta_0 h_0 + Af, \tag{5.2}$$

where A is a linear mapping, h_0 is a known function in $\mathscr{H}_{\mathscr{R}}$ which is orthogonal to $A(\mathscr{H}_0)$, and β_0 is a fixed but unknown parameter. The above orthogonality requirement for h_0 to $A(\mathscr{H}_0)$ can be relaxed to a condition that h_0 is linearly independent of the space $A(\mathscr{H}_0)$. The parallel results of Theorems 2.2 and 3.1 are given in Theorems 5.1 and 5.2 respectively. The proofs are similar and omitted.

THEOREM 5.1 (Gauss–Markov). *Suppose that the model* (1.1) *in Theorem* 2.2 *is replaced by* (5.1) *and* (5.2). *Under the same conditions in Theorem* 2.2, *the BLUP for f in* (5.1) *is given by* $\hat{f}_{BLUP} = \hat{f}_{FE} + \hat{f}_{RE}$ *with*

$$\hat{f}_{FE} = (A_0^* M^{-1} A_0)^{-1}\, A_0^* M^{-1} (Y - \hat{\beta}_0 h_0)$$

$$\hat{f}_{RE} = \alpha^{-1} W A_{\mathscr{W}}^* M^{-1} (Y - \hat{\beta}_0 h - A_0 \hat{f}_{FE}),$$

where

$$\hat{\beta}_0 = \langle h_0, M^{-1}Y\rangle_{L_2(\mathscr{I})} / \langle h_0, M^{-1}h_0\rangle_{L_2(\mathscr{I})}$$

and all other notations are defined as in Theorem 2.2.

Notice that, if h_0 is only independent of $A(\mathscr{H}_0)$ instead of orthogonal to $A(\mathscr{H}_0)$, the above estimate of $\hat{\beta}_0$ should be modified accordingly.

THEOREM 5.2. *Assume the identifiability condition as well as the two boundedness conditions I and II hold. Then the unique solution to problem*

$$\min_{f \in \mathscr{H}_1} -2 \langle Y, \tilde{A} f \rangle_{\mathscr{H}_{\mathscr{I}}} + \langle \tilde{A} f, \tilde{A} f \rangle_{\mathscr{H}_{\mathscr{I}}} + \alpha \| P_{[\mathscr{H}_1 \to \mathscr{H}_W]} f \|^2_{\mathscr{H}_W} \tag{5.3}$$

is the BLUP, $\hat{f}_{BLUP}$, *in Theorem* 5.1 *with* $\alpha = \sigma^2/\delta^2$.

6. EXAMPLES AND DISCUSSION

A few selected examples are briefly discussed below.

EXAMPLE 1 (Linear mixed-effects regression models). These are special cases for the problem considered in this article. Consider the linear mixed-effect model $Y = X\beta + Z\gamma + \sigma\varepsilon$, where $Var(\varepsilon) = R$ and $Var(\gamma) = \delta^2 W$. Then $\mathscr{T} = \{1, 2, ..., m+r\}$, $\mathscr{J} = \{1, 2, ..., n\}$, $A = (X, Z)$, $A^* = (X, Z)^T R^{-1}$, $P = \text{diag}(0_{m \times m}, I_{r \times r})$, $P^* = \text{diag}(0_{m \times m}, W^{-1})$ and $f = (\beta^T, \gamma^T)^T$. Plugging into Theorem 4.1, one can easily get the normal equations

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + \alpha^{-1} W^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} X^T R^{-1} Y \\ Z^T R^{-1} Y \end{pmatrix},$$

where $\alpha = \sigma^2/\delta^2$. The resulting estimates for β and γ are BLUP.

EXAMPLE 2 (Nonparametric regression, continued from Remark 2).

$$Y_i = f(t_i) + \sigma\varepsilon_i, \qquad i = 1, ..., n.$$

Suppose that we have iid observations Y_i at uniform design points. Write f in terms of the wavelet expansion

$$f(t) = \sum_k \beta_k \phi_{j,k}(t) + \delta \sum_{\ell \geqslant j} \sum_k \gamma_{\ell,k} \psi_{\ell,k}(t).$$

The reader is referred to Huang and Lu (2000) for the BLUP, the associated Sobolev regularization and an asymptotic equivalence of the BLUP. The asymptotic equivalence of the BLUP can be shown to be

$$\hat{f}_{\text{BLUP}}(t) \simeq \sum_k \hat{\beta}_k \phi_{j,k}(t) + \sum_{\ell \geqslant j} \sum_k \left(1 - \frac{\sigma^2/n}{\delta^2 \lambda_\ell + \sigma^2/n}\right) \hat{\gamma}_{\ell,k} \psi_{\ell,k}(t),$$

where $\hat{\beta}_k$ and $\hat{\gamma}_{\ell,k}$ are empirical wavelet coefficients. The quantity $\delta^2\lambda_\ell + \sigma^2/n$ can be estimated by $\hat{\gamma}^2_{\ell,k}$. Thus the BLUP practically provides a useful guidance for the nonlinear estimator proposed below,

$$\hat{f}^{\text{GCV}}_{\text{BLUPWAVE}}(t) = \sum_k \hat{\beta}_k \phi_{j,k}(t) + \sum_{\ell \geqslant j} \sum_k \left(1 - \frac{c}{\hat{\gamma}^2_{\ell,k}}\right)_+ \hat{\gamma}_{\ell,k} \psi_{\ell,k}(t),$$

where c is chosen by the generalized cross validation.

EXAMPLE 3 (Deconvolution).

$$Y_i = \int_0^1 K(t_i - t)\, f(t)\, dt + \sigma\varepsilon_i, \qquad i = 1, ..., n.$$

Suppose that we have iid observations Y_i at uniform design points and that the convolution kernel K is known. Write f in terms of the wavelet expansion

$$f = \sum_{k \in Z} \beta_k \phi_{j,k} + \sum_{\ell \geqslant j} \sum_{k \in Z} \gamma_{\ell,k} \psi_{\ell,k} \qquad \text{for a fixed } j,$$

where $\beta_k = \langle f, \phi_{j,k} \rangle$ and $\gamma_{\ell,k} = \langle f, \psi_{\ell,k} \rangle$. Assume that the convolution kernel K satisfies the condition:

$$K * u = 0 \qquad \text{if and only if } u \text{ is a zero function.}$$

Let

$$K^*(\omega) = \int K(x)\, e^{-i\omega x}\, dx,$$

$$\phi^*_{j,k}(\omega) = \int \phi_{j,k}(x)\, e^{-i\omega x}\, dx,$$

$$\psi^*_{\ell,k}(\omega) = \int \psi_{\ell,k}(x)\, e^{-i\omega x}\, dx,$$

$$\tilde{\phi}_{j,k}(x) = \frac{1}{2\pi} \int \phi^*_{j,k}(\omega)(K^*)^{-1}(\omega)\, e^{i\omega x}\, d\omega,$$

$$\tilde{\psi}_{\ell,k}(x) = \frac{1}{2\pi} \int \psi^*_{\ell,k}(\omega)(K^*)^{-1}(\omega)\, e^{i\omega x}\, d\omega.$$

Then the two sets, $\{K * \phi_{j,k},\ K * \psi_{\ell,k},\ \ell \geqslant j,\ k \in Z\}$ and $\{\tilde{\phi}_{j,k},\ \tilde{\psi}_{\ell,k},\ \ell \geqslant j,\ k \in Z\}$, are both a complete basis for $L_2(I)$. Moreover, they are bi-orthogonal dual bases. Then the empirical coefficients are given by

$$\hat{\beta}^{\mathrm{e}}_k = \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}_{j,k}(t_i)\ Y_i,$$

$$\hat{\gamma}^{\mathrm{e}}_{\ell,k} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\psi}_{\ell,k}(t_i)\ Y_i.$$

The BLUP can be derived as

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}_{\mathrm{BLUP}} = \left\{ I - \lambda (X^T X)^{-1} \begin{pmatrix} 0 & 0 \\ 0 & \Lambda^{-1} \end{pmatrix} \right\} \begin{pmatrix} \hat{\beta}^{\mathrm{e}} \\ \hat{\gamma}^{\mathrm{e}} \end{pmatrix}, \tag{6.1}$$

where X is the design matrix (including fixed and mixed effects) based on $\{K * \phi_{j,k}(t_i),\ K * \psi_{\ell,k}(t_i) : k \in Z,\ \ell \geqslant j,\ i = 1, ..., n\}$ and Λ is a diagonal matrix given by $\Lambda = \mathrm{diag}(\lambda_j, \lambda_{j+1}, ...)^T$. Parameter selections for λ and λ_ℓ are important. Again, (6.1) serves as a guidance of possible nonlinear estimators for f with plug-in parameter estimates. However, it is not the main interest of this article and we will omit further discussion.

The extended Gauss–Markov theorem for nonparametric mixed-effects models provides theoretical enlightenments for a large spectrum of problems. The perspectives of BLUP, PLS, regularization, and minimaxity have interesting connections as demonstrated in this article. These results combining with the approximation tools, like splines or wavelets, can shed light on the reconstruction of signals from noisy data. In particular, the connection between the estimated coefficients and unknown parameters can be used to explore effective nonlinear estimators (or predictors) adaptively with the oracle provided by the extended Gauss–Markov theorem.

ACKNOWLEDGMENTS

The authors thank Professor Ker-Chau Li for helpful discussion and the referee for comments.

REFERENCES

1. S. J. Anderson and R. H. Jones, Smoothing splines for longitudinal data, *Statist. Medicine* **14** (1995), 1235–1248.
2. D. Barry, Nonparametric Bayesian regression, *Ann. Statist.* **86** (1986), 934–953.
3. P. Besse and J. O. Ramsay, Principal components analysis of sampled functions, *Psychometrika* **51** (1986), 285–311.

4. D. Harville, Extension of the Gauss–Markov theorem to include the estimation of random effects, *Ann. Statist.* **4** (1976), 384–395.
5. G. Kimeldorf and G. Wahba, A correspondence between Bayesian estimation in stochastic processes and smoothing by splines, *Ann. Math. Statist.* **41** (1970), 495–562.
6. G. Kimeldorf and G. Wahba, Some results on Tchebycheffian spline functions, *J. Math. Anal. Appl.* **33** (1971), 82–95.
7. A. Kneip and Th. Gasser, Statistical tools to analyze data representing a sample of curves, *Ann. Statist.* **20** (1992), 1266–1305.
8. N. M. Laird and J. H. Ware, Random-effects models for longitudinal data, *Biometrics* **38** (1982), 963–974.
9. K. C. Li, Minimaxity of the method of regularization on stochastic processes, *Ann. Statist.* **10** (1982), 937–942.
10. S. Y. Huang and H. H.-S. Lu, Bayesian wavelet shrinkage for nonparametric mixed-effects models, *Statist. Sinica* **10** (2000), 1021–1040.
11. E. Parzen, An approach to time series analysis, *Ann. Math. Statist.* **32** (1961), 951–989.
12. J. O. Ramsay and C. J. Dalzell, Some tools for functional data analysis, *J. Roy. Statist. Soc. Ser. B* **53** (1991), 539–572.
13. J. O. Ramsay, When the data are functions, *Psychometrika* **47** (1982), 379–396.
14. J. O. Ramsay and B. Silverman, "Functional Data Analysis," Springer-Verlag, Berlin/New York, 1997.
15. M. Reed and B. Simon, "Methods of Modern Mathematical Physics: I. Functional Analysis," Academic Press, San Diego, 1972.
16. G. K. Robinson, That BLUP is a good thing: the estimation of random effects, *Statist. Sci.* **6** (1991), 15–51.
17. T. Speed, Comment on "That BLUP is a Good Thing: The Estimation of Random Effects," by G. K. Robinson, *Statist. Sci.* **6** (1991), 15–51.
18. G. Wahba, Improper priors, spline smoothing and the problem of guarding against model errors in regression, *J. Roy. Statist Soc. Ser. B* **40** (1978), 364–372.
19. G. Wahba, "Spline Models for Observational Data," Society for Industrial and Applied Mathematics, 1990.
20. Y. Wang, Mixed effects smoothing spline analysis of variance, *J. Roy. Statist. Soc. Ser. B* **80** (1998), 159–174.