On the Equivalence of the Weighted Least Squares and the Generalised Least Squares Estimators, with Applications to Kernel Smoothing

Luati, Alessandra and Proietti, Tommaso

30 May 2008

On the Equivalence of the Weighted Least Squares and the Generalised Least Squares Estimators, with Applications to Kernel Smoothing

Alessandra Luati
Dip. Scienze Statistiche
University of Bologna

Tommaso Proietti
S.E.F. e ME. Q.
University of Rome "Tor Vergata"

Abstract

The paper establishes the conditions under which the generalised least squares estimator of the regression parameters is equivalent to the weighted least squares estimator. The equivalence conditions have interesting applications in local polynomial regression and kernel smoothing. Specifically, they enable to derive the optimal kernel associated with a particular covariance structure of the measurement error, where optimality has to be intended in the Gauss-Markov sense. For local polynomial regression it is shown that there is a class of covariance structures, associated with non-invertible moving average processes of given orders which yield the the Epanechnikov and the Henderson kernels as the optimal kernels.

Keywords: Local polynomial regression; Epanechnikov Kernel; Non-invertible Moving average processes.

1 Introduction

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$, $p < n$. Throughout the paper we will assume that $\mathbf{X}$ is a deterministic matrix with full column rank and that the covariance matrix $\boldsymbol{\Sigma}$ is positive definite and non singular. We can relax both the assumption of normality and of deterministic regressors and replace it by the weak exogeneity assumption, $\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}, \mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \boldsymbol{\Sigma}$.

A well-known result (Aitken theorem, 1935) states that, if $\boldsymbol{\Sigma}$ is known, the best linear unbiased estimator (BLUE) of the regression parameters is the generalised least squares estimator (GLSE)

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}. \tag{2}$$

Much attention has been devoted in the literature to the search of conditions for which the ordinary least squares estimator (OLSE),

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{3}$$

is equivalent to the GLSE (2), and thus it is BLUE.

Anderson was the first who faced this problem, stating (1948, p. 48) and proving (1971, pp. 19 and 560) that equality between (2) and (3) holds if and only if there are p linear combinations of the columns of $\mathbf{X}$ that are eigenvectors of $\boldsymbol{\Sigma}$. The relevance of this result is self-evident, although Anderson's condition is not easy to verify in practice, i.e. for given matrices $\mathbf{X}$ and $\boldsymbol{\Sigma}$. Later developments in this field concerned the search of equivalent conditions for the OLSE to be BLUE. A relevant contribution in this sense was that of Zyskind (1967), who derived eight equivalent conditions, among which the commutativity relation between the covariance matrix and the orthogonal projection matrix onto the column space of $\mathbf{X}$. Commutativity is easy to verify when $\boldsymbol{\Sigma}$ is known. See also Amemiya (1985, pp. 182-183).

Further investigations concerned the search of conditions for the GLSE to be BLUE even though some hypotheses of Aitken theorem are relaxed, for example when $\mathbf{X}$ or $\boldsymbol{\Sigma}$ are not full rank (see Zyskind and Martin, 1969; Lowerre, 1974; Baksalary and Kala, 1983). Other approaches investigated equality over $\mathbf{y}$ (Krämer, 1980; Jaeger and Krämer, 1998) or for varying $\mathbf{X}$ (Watson, 1967; McElroy, 1967, Zyskind, 1969, Baksalary and Van Eijnsbergen, 1988) or in a coordinate-free setting (Kruskal, 1968; Phillips, 1992). An excellent and exhaustive review of these results is Puntanen and Styan (1989). Another strand of the literature has considered the asymptotic equivalence of $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{GLS}$; well known cases are polynomial and trigonometric deterministic regression in time series (Grenander and Rosenblatt, 1957), time series regressions with integrated regressors (Phillips and Park, 1988), ARIMA regressors (Krämer, 1986), fractionally integrated regressors (Krämer and Hassler, 1998).

This paper is concerned instead with establishing the conditions under which there exists a diagonal matrix $\mathbf{K}$ such that the GLSE is equivalent to the weighted least squares estimator (WLSE)

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}\mathbf{y}. \tag{4}$$

When these conditions are met, the diagonal elements of $\mathbf{K}$ provide the optimal kernel weights corresponding to a given covariance structure $\boldsymbol{\Sigma}$, where optimality is to be intended in the Gauss-Markov sense.

The interest in this issue arises in the context of local polynomial modelling, where we shall be able to derive a general class of kernels, isomorphic to noninvertible MA processes, that are particularly well behaved and that encompasses two very important kernels. It will turn out, in fact, that the Epanechnikov kernel is the optimal kernel in local polynomial regression with strictly noninvertible first order moving average errors. Similarly, the Henderson kernel (Henderson, 1916, see also Loader, 1999) is optimal when the error is a strictly non-invertible third order moving average process.

The plan of the paper is as follows: the main theorem, establishing the equivalence between GLSE and WLSE is stated in section 2 and proved in the appendix. Section 3 reviews local polynomial regression in a time series setting. It serves to set up the notation for the next section, which presents the main application of the theorem (section 4), dealing with the optimal kernel corresponding to a particular covariance structure. A sufficient condition for optimality is given (sections 4.1 and 4.2), and a more general result is proved for local polynomial regression with non invertible moving average errors (section 5). In section 4.2 we also provide illustration of this general result dealing with the Epanechnikov and the Henderson kernel. Section 6 addresses the inverse problem of determining the covariance structure corresponding to a given kernel. Section 7 concludes the paper.

2   Main results

Let us denote by $\mathcal{C}(\mathbf{X})$ the column space of $\mathbf{X}$, also called its range, and by $\mathcal{N}(\mathbf{X})$ its null space. If $\mathbf{W} \in \mathbb{R}^{n \times n}$ and $\text{rank}(\mathbf{W}) = n$, then $\mathbf{H}_W = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ is the (oblique) projection matrix onto $\mathcal{C}(\mathbf{X})$ along $\mathcal{N}(\mathbf{X}'\mathbf{W})$. The subspaces $\mathcal{C}(\mathbf{X})$ and $\mathcal{N}(\mathbf{X}'\mathbf{W})$ are complementary, in the sense that they have null intersection and their union is $\mathbb{R}^n$ (see Meyer, 2000).

The following theorem states a necessary and sufficient condition for equality between $\hat{\boldsymbol{\beta}}_{GLS}$ and $\hat{\boldsymbol{\beta}}_{WLS}$.

Theorem 1 *Equality between the GLS estimator (2) and the WLS estimator (4) holds if and only if $\mathbf{X} = \mathbf{V}^*\mathbf{M}$ where the p columns of $\mathbf{V}^*$ are eigenvectors of $\boldsymbol{\Sigma}\mathbf{K}$ and $\mathbf{M}$ is a non singular matrix.*

The proof is reported in appendix A. The theorem states that if there are p linear combinations of the columns of $\mathbf{X}$ that are eigenvectors of $\boldsymbol{\Sigma}\mathbf{K}$ then the GLSE with covariance matrix $\boldsymbol{\Sigma}$ is

equal to the WLSE with kernel $\mathbf{K}$. If the conditions of the theorem hold, the equality is true for all $\mathbf{y} \in \mathbb{R}^n$, i.e.

$$(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1} = (\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}$$

from which follows that

$$\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1} = \mathbf{X}(\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}.$$

The latter equality states that the projection matrix onto $\mathcal{C}(\mathbf{X})$ along $\mathcal{N}(\mathbf{X}'\boldsymbol{\Sigma}^{-1})$ is equal to the projection matrix onto $\mathcal{C}(\mathbf{X})$ along $\mathcal{N}(\mathbf{X}'\mathbf{K})$, i.e. $\mathbf{H}_{\Sigma^{-1}} = \mathbf{H}_K$. By uniqueness of the projection and complementarity of the spaces which it acts onto and along, it follows that $\mathcal{N}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}) \equiv \mathcal{N}(\mathbf{X}'\mathbf{K})$. This allows to generalise Zyskind (1967) most famous equivalent condition to Anderson theorem in the following corollary, whose proof is provided in the appendix.

Corollary 1 *A necessary and sufficient condition for equality between the GLS estimator (2) and the WLS estimator (4) is that* $\boldsymbol{\Sigma}\mathbf{K}\mathbf{H} = \mathbf{H}\boldsymbol{\Sigma}\mathbf{K}$ *where* $\mathbf{H} = \mathbf{H}_{\Sigma^{-1}} = \mathbf{H}_K$.

For $\mathbf{K} = \mathbf{I}$, the identity matrix, we find Zyskind condition for OLSE to be BLUE. The generalisation is not straightforward, given that Zyskind proof is based on the symmetry of both $\boldsymbol{\Sigma}$ and $\mathbf{H}_I$, the orthogonal projection matrix onto $\mathcal{C}(\mathbf{X})$, that enables to show that the two matrices have the same eigenvectors and therefore commute. When $\mathbf{K}$ is not the identity or more generally a scalar matrix, then neither $\mathbf{H}$ nor $\boldsymbol{\Sigma}\mathbf{K}$ are symmetric and in fact our proof of the corollary, revolves around the equality between $\boldsymbol{\Sigma}\mathbf{K}\mathbf{H}$ and $\mathbf{H}\boldsymbol{\Sigma}\mathbf{K}$. In any case, the corollary establishes that the matrices $\boldsymbol{\Sigma}\mathbf{K}$ and $\mathbf{H}$ commute and therefore have the same eigenvectors. Given that a complete set of eigenvectors of $\mathbf{H}$ spans $\mathbb{R}^n$, the matrix $\boldsymbol{\Sigma}\mathbf{K}$ can be reduced to a diagonal form through the same matrix that diagonalises $\mathbf{H}$. This provides a further condition to verify if equality holds between (2) and (4).

Typically, the design matrix $\mathbf{X}$ and either $\boldsymbol{\Sigma}$ or $\mathbf{K}$ are known. The first use of the above results is to obtain the diagonal matrix $\mathbf{K}$ from the pair $\mathbf{X}, \boldsymbol{\Sigma}$, as the optimal kernel that yields the best linear unbiased predictor of $\mathbf{y}$ given $\mathbf{X}$, assuming the covariance structure $\boldsymbol{\Sigma}$. For this purpose, we need to be able to determine the matrix $\mathbf{M}$ of theorem 1. This is achieved in the next section, which deals with local polynomial regression with equally spaced design points, for which the matrix $\mathbf{M}$ has a very specialised structure.

3   Local polynomial regression

The leading case of interest for the application of the above results is local polynomial regression in a time series setting. Essential references are Fan and Gjibels (1996) and Loader (1999). Let us assume that y_t is a time series, measured at discrete and equally spaced time points, that can be decomposed as $y_t = \mu_t + \varepsilon_t$, where μ_t is the signal (trend) and $\varepsilon_t \sim \mathrm{NID}(0, \sigma^2)$ is the noise. The signal is approximated locally by a polynomial of degree d, so that in the neighbourhood of time

t we can write

$$y_{t+j} = m_{t+j} + \varepsilon_{t+j}, \quad m_{t+j} = \beta_0 + \beta_1 j + \beta_2 j^2 + \cdots + \beta_d j^d, j = 0, \pm 1, \cdots, \pm h.$$

In matrix notation, the local polynomial approximation can be written as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{5}$$

where $\mathbf{y} = [y_{t-h}, \cdots, y_t, \cdots, y_{t+h}]'$, $\boldsymbol{\varepsilon} = [\varepsilon_{t-h}, \cdots, \varepsilon_t, \cdots, \varepsilon_{t+h}]'$,

$$\mathbf{X} = \begin{bmatrix} 1 & -h & h^2 & \vdots & (-h)^d \\ 1 & -(h-1) & (h-1)^2 & \vdots & [-(h-1)]^d \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 1 & 0 & 0 & \vdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 1 & h-1 & (h-1)^2 & \vdots & (h-1)^d \\ 1 & h & h^2 & \vdots & h^d \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix},$$

and $\boldsymbol{\Sigma} = \{\sigma_{ij}, i, j = -h, \ldots, h\}$.

Using this design, the value of the trend at time t is simply given by the intercept, $m_t = \beta_0$. Provided that $2h \geq d$, the $d+1$ unknown coefficients $\beta_k, k = 0, \ldots, d$, can be estimated by the method of generalised least squares, giving $\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$. In order to obtain $\hat{m}_t = \hat{\beta}_0$, we need to select the first element of the vector $\hat{\boldsymbol{\beta}}_{GLS}$. Hence, denoting by $\mathbf{e}_1$ the $d+1$ vector $\mathbf{e}_1 = [1, 0, \ldots, 0]'$,

$$\hat{m}_t = \mathbf{e}_1'\hat{\boldsymbol{\beta}}_{GLS} = \mathbf{e}_1'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} = \mathbf{w}'\mathbf{y} = \sum_{j=-h}^{h} \mathrm{w}_j y_{t-j},$$

which expresses the estimate of the trend as a linear combination of the observations with coefficients

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{e}_1. \tag{6}$$

We notice in passing that expression (6) can be equivalently derived as the solution of the constrained minimisation problem:

$$\min_{\mathbf{w}}\{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}\} \text{ subject to } \mathbf{w}'\mathbf{X} = \mathbf{e}_1',$$

where the linear constraints $\mathbf{w}'\mathbf{X} = \mathbf{e}_1'$ enforce the condition that the trend estimate reproduces a polynomial of degree d (i.e. if $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$, $\hat{m}_t = \mathbf{w}'\mathbf{y} = \beta_0$). See Hannan (1970, p. 186-187), and Wallis (1983).

Estimates of $\boldsymbol{\beta}$ can be also obtained by the method of weighted least squares, which consists of minimising with respect to the β_k's the objective function:

$$S(\hat{\beta}_0, \ldots, \hat{\beta}_d) = \sum_{j=-h}^{h} \kappa_j \left(y_{t+j} - \hat{\beta}_0 - \hat{\beta}_1 j - \hat{\beta}_2 j^2 - \cdots - \hat{\beta}_d j^d \right)^2,$$

where $\kappa_j \geq 0$ is a set of weights that define, either explicitly or implicitly, a kernel function. In general, kernels are chosen to be symmetric and non increasing functions of j, in order to weight the observations differently according to their distance from time t; in particular, larger weight may be assigned to the observations that are closer to t. As a result, the influence of each individual observation is controlled not only by the bandwidth h but also by the kernel. In matrix notation, setting $\mathbf{K} = \mathrm{diag}(\kappa_{-h}, \ldots, \kappa_{-1}, \kappa_0, \kappa_1, \ldots, \kappa_h)$, the WLS estimate of the coefficients is $\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}\mathbf{y}$ and the elements of the vector $\mathbf{w} = \mathbf{K}\mathbf{X}(\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{e}_1$ constitute the so called equivalent kernel. Note that the notation $\mathbf{w}$ is used both for the GLS coefficients (6) and for the equivalent kernel arising from WLS estimation, since we will mainly focus on the case when their elements are identical. If this should not be the case, then which one of the two meanings is to be intended will be clear from the context.

4 The optimal kernel in local polynomial regression

We address the question of the equivalence of the GLSE and the WLSE in the local polynomial regression problem described above. When the conditions of theorem 1 are satisfied, we shall refer to the diagonal elements of $\mathbf{K}$ as the optimal kernel weights. We stress that here optimality is in the Gauss-Markov sense and expresses the fact that using $\mathbf{K}$ is equivalent to using $\boldsymbol{\Sigma}$ for computing the optimal estimate of the signal and its time derivatives.

The conditions under which the equivalence holds are typically difficult to check, but in the local polynomial regression framework considered in the previous section, the particular structure of the design matrix, and consequently of the matrix $\mathbf{M}$ of theorem 1, leads to a considerable simplification.

The matrix $\mathbf{M}$ can be chosen as upper triangular with further zeros along the secondary, fourth, and so on, (upper) diagonals. This follows from the algebraic structure of $\mathbf{X}'\mathbf{K}\mathbf{X}$ and $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$. In fact, $\mathbf{X}'\mathbf{K}\mathbf{X}$ is a Hankel matrix whose elements are the values $S_r = \sum_{j=-h}^{h} j^r \kappa_j$, for $r = 0, 1, ..., 2d$, from S_0 to S_d in the first row and from S_d to S_{2d} in the last column. Note that for symmetric kernel weights satisfying $\kappa_j = \kappa_{-j}$, $S_r = 0$ for odd r and therefore $\mathbf{X}'\mathbf{K}\mathbf{X}$ has null elements along the secondary, fourth, and so on, diagonals. The matrix $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$ has not Hankel structure but has zeros along the secondary, fourth, and so forth diagonals as well, which stems from the fact that the covariance matrix of a stationary stochastic process is a symmetric Toeplitz matrix. Illustrations will be provided in section 5.

Now, $\mathbf{M}$ is such that $\boldsymbol{\Sigma}\mathbf{K}\mathbf{X}\mathbf{M}^{-1} = \mathbf{X}\mathbf{M}^{-1}\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix (see Appendix A), or, equivalently, $\boldsymbol{\Sigma}^{-1}\mathbf{X}\mathbf{M}^{-1}\mathbf{D} = \mathbf{K}\mathbf{X}\mathbf{M}^{-1}$. As a result, the linear combinations of the columns

of $\mathbf{X}$ yielding the required p eigenvectors of $\mathbf{\Sigma K}$, are provided by $\mathbf{XM}^{-1}$. This gives an operative procedure to get $\mathbf{K}$ by $\mathbf{\Sigma}$, formalised in $d+1$ conditions that directly follow by the sparse upper triangular structure of $\mathbf{M}$. In section 5 we shall provide explicit conditions in terms of the generic elements of $\mathbf{\Sigma}^{-1}$ and of $\mathbf{K}$ for $d \leq 3$, which are the most frequently encountered degrees for the fitting polynomial.

First and foremost, a very simple and easily checked necessary condition arises in a regression framework with an intercept, i.e. when the first column of $\mathbf{X}$ is the vector of ones, denoted $\mathbf{i} = [1, 1, \ldots, 1]'$. This will be discussed below in section 4.1.

4.1 Local constant regression and a simple necessary condition

When the degree of the fitting polynomial is equal to zero, $\mathbf{X} = \mathbf{i}$ and $\mathbf{M}$ is a scalar, so that the necessary and sufficient condition that $\mathbf{K}$ and $\mathbf{\Sigma}$ must satisfy for the WLSE to equal to the GLSE reduces to $\mathbf{\Sigma}^{-1}\mathbf{i} = \mathbf{Ki}$. Denoting by ς_{ij} the generic element of $\mathbf{\Sigma}^{-1}$, for $i, j = -h, \ldots, 0, \ldots, h$, the unnormalised kernel weights are equal to the row sums of the elements of the inverse covariance matrix, that is

$$\kappa_j = \sum_{i=-h}^{h} \varsigma_{ij}, \quad \text{for } j = -h, \ldots, h.$$

In the more general case, the first column of the matrix $\mathbf{X}$ is the vector $\mathbf{i}$, and the matrix $\mathbf{M}$ is upper triangular; hence, the first column of $\mathbf{X}$ is itself an eigenvector of $\mathbf{\Sigma K}$ corresponding to an eigenvalue, say, d_1, so that $\mathbf{\Sigma Ki} = d_1\mathbf{i}$. It therefore follows that a necessary condition for $\mathbf{K}$ to satisfy theorem 1 is that, up to the factor d_1,

$$\mathbf{Ki} \propto \mathbf{\Sigma}^{-1}\mathbf{i} \tag{7}$$

which means that the elements of $\mathbf{K}$ are (proportional to) the sum of the row elements of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$. As pointed out above, for local constant estimators belonging to the Nadaraya (1964) and Watson (1964) class, the condition is also sufficient. Hence, in the general case we suggest the following strategy:

- derive a candidate kernel from the necessary condition $\boldsymbol{\kappa} = \mathbf{\Sigma}^{-1}\mathbf{i}$;

- verify that the other conditions are met.

Obviously, for spherical errors, $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, the candidate kernel is the uniform kernel. When ε_t is the first order autoregressive process, or AR(1), $\varepsilon_t = \phi\varepsilon_{t-1} + \xi_t, \xi_t \sim \text{WN}(0, \sigma^2)$, where WN denotes a white noise process,

$$\kappa_{|h|} = 1 - \phi, \kappa_j = (1 - \phi)^2, j = 0, \pm 1, \ldots, \pm(h-1),$$

so that the kernel will be admissible, the weights will be non increasing with $|j|$, if $-1 < \phi < 0$. This example has been used in the literature to illustrate the asymptotic equivalence of OLS and
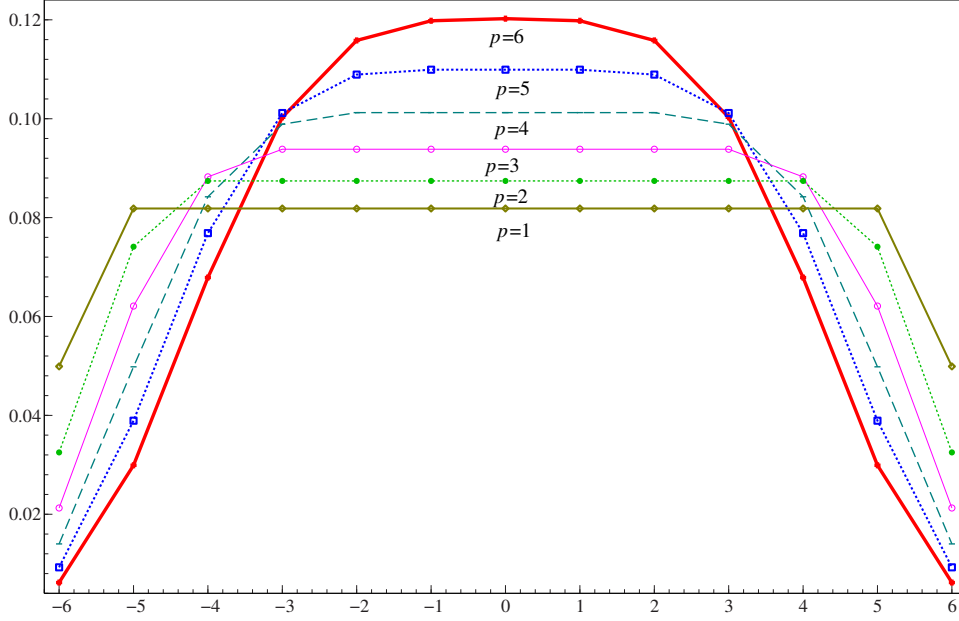
Figure 1: Plot of the kernel weights associated to the covariance matrix of the AR(p) process $(1 + 0.64B)^p \varepsilon_t = \xi_t$, for $p = 1, \ldots, 6$ and $h = 6$.

GLS for polynomial trend estimation. As a matter of fact, when h goes to infinity, the kernel tends to the uniform kernel. If $\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-1} + \xi_t, \xi_t \sim \text{WN}(0, \sigma^2)$,

$$\kappa_{|h|} = 1 - \phi_1 - \phi_2, \kappa_{|h-1|} = (1 - \phi_1)^2 - \phi_2(2 - \phi_1), \kappa_j = (1 - \phi_1 - \phi_2)^2, j = 0, \pm 1, \pm(h-2).$$

The kernel will be admissible only for some parameter combinations. In general, if $\varepsilon_t \sim \text{AR}(p)$, the central weights for $|j| \leq h - p$ will be constant. Figure 1 displays the kernel associated to the AR(p) process $(1 + 0.64B)^p \varepsilon_t = \xi_t$, where B is the backshift operator such that $B^k x_t = x_{t-k}$, for $p = 1, \ldots, 6$, and $h = 6$. The process $(1 - \phi B)^p \varepsilon_t = \xi_t$ with a positive ϕ does not yield an admissible kernel, as $\boldsymbol{\Sigma}^{-1}\mathbf{i}$ has negative elements.

4.2 Non-invertible moving average models

An important class of candidate kernels, nesting the Epanechnikov and the Henderson kernels, arises in the local polynomial regression framework, when the error ε_t is generated by the non-invertible moving average (MA) process of order q:

$$\varepsilon_t = (1 - B)^q \xi_t, \quad \xi_t \sim \text{WN}(0, \sigma^2). \tag{8}$$

From the interpretative standpoint, (8) is the roughest stationary MA(q) process, since its spectral density has q unit poles at the zero frequency and increases monotonically from 0 to the Nyquist

7

frequency. As a consequence, postulating this model amounts to impose a smoothness prior on the signal estimates.

Let us denote by $\boldsymbol{\Sigma}_q$ the covariance matrix of the process (8). This is the symmetric $2h + 1$-banded Toeplitz matrix, with the coefficients associated with B in the binomial expansion of $(1 - B)^{2q}$, displayed symmetrically about the diagonal in each row and column. For instance,

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & -1 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 \\ 0 & -1 & 2 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 2 & -1 \\ 0 & 0 & 0 & \ldots & -1 & 2 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 6 & -4 & 1 & \ldots & 0 & 0 \\ -4 & 6 & -4 & \ldots & 0 & 0 \\ 1 & -4 & 6 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 6 & -4 \\ 0 & 0 & 0 & \ldots & -4 & 6 \end{bmatrix}.$$

For determining the candidate kernel by the necessary condition $\boldsymbol{\Sigma}_q^{-1}\mathbf{i}$, we use a result due to Hoskins and Ponzo (1972), according to which the j-th row sum $\boldsymbol{\Sigma}_q^{-1}\mathbf{i}$ is

$$\sum_{i=-h}^{h} \varsigma_{q,ij} = \frac{\binom{h+j+q}{q}\binom{h-j+q}{q}}{\binom{2q}{q}}, \quad j = -h, ..., h,$$

where we have adapted Theorem 3, p. 396, of Hoskins and Ponzo (1972) to our notation and corrected a minor mistake concerning the sign. Hence,

$$\begin{aligned} \sum_{i=-h}^{h} \varsigma_{q,ij} &= \frac{1}{(2q)!} \frac{(h+j+q)!}{(h+j)!} \frac{(h-j+q)!}{(h-j)!} = \\ &= \frac{1}{(2q)!}(h+1+j)(h+2+j)\ldots(h+q+j)(h+1-j)(h+2-j)\ldots(h+q-j) \\ &= \frac{1}{(2q)!}[(h+1)^2 - j^2]\ldots[(h+q)^2 - j^2] \\ &= \kappa_{q,j}. \end{aligned}$$

In conclusion, the candidate kernel satisfying $\boldsymbol{\kappa}_q = \boldsymbol{\Sigma}_q^{-1}\mathbf{i}$ has weights

$$\kappa_{q,j} \propto [(h+1)^2 - j^2][(h+2)^2 - j^2]\ldots[(h+q)^2 - j^2], \tag{9}$$

for $j = -h, ..., h$.

When $q = 1$, $\varepsilon_t = (1 - B)\xi_t$ and $\boldsymbol{\kappa}_1$ is the Epanechnikov (1969) kernel, with elements $\kappa_{1,j} \propto [(h+1)^2 - j^2]$, or, equivalently,

$$\kappa_{1,j} \propto \frac{3}{4}\left[1 - \left(\frac{j}{h+1}\right)^2\right], \ j = -h, ..., h.$$

The Epanechnikov kernel minimises the asymptotic mean integrated square error (see Priestley and Chao, 1972, and Benedetti, 1977) and the efficiency of any kernel estimator is generally measured with respect to it (see Wand and Jones, 1995).
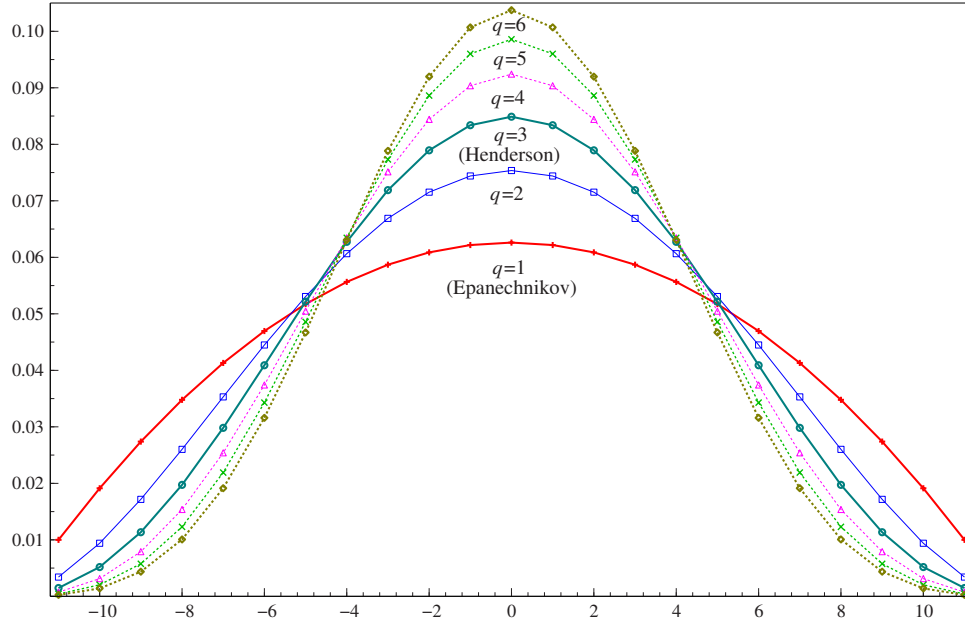
8

Figure 2: Plot of the normalised kernel weights associated to the covariance matrix of the non-invertible MA(q) process $\varepsilon_t = (1-B)^q \xi_t$, for $q = 1, \ldots, 6$ and $h = 10$.

Also another popular kernel, the Henderson kernel (Henderson, 1916) is nested in (9), arising when $q = 3$:

$$\kappa_{3,j} \propto [(h+1)^2 - j^2][(h+2)^2 - j^2][(h+3)^2 - j^2]. \tag{10}$$

The Henderson filter (see Henderson, 1916, Kenny and Durbin, 1982, Loader, 1999, Ladiray and Quenneville, 2001) arises as the weighted least squares estimator of a local cubic trend at time t using $2h + 1$ consecutive observations. The filter has a long tradition for trend-cycle estimation in economic time series. The relevance of Henderson's contribution to modern local regression is stressed in Loader (1999). Currently, the Henderson filters are employed for trend estimation in the X-12-ARIMA procedure, the official seasonal adjustment procedure in the U.S., Canada, the U.K. and many other countries. See Dagum (1980), Findley *et al.* (1998) and Ladiray and Quenneville (2001) for more details.

Henderson (1916) addressed the problem of defining a set of kernel weights that maximise the smoothness of the estimated local cubic trend, in the sense that the variance of its third differences is as small as possible. In local cubic regression, with $d = 3$, the GLSE of the trend when the covariance matrix $\boldsymbol{\Sigma}_3$ is the symmetric Toeplitz matrix with nonzero elements $\sigma_{ii} = 20$, $\sigma_{i,i+1} = \sigma_{i,i-1} = -15$, $\sigma_{i,i+2} = \sigma_{i,i-2} = 6$, $\sigma_{i,i+3} = \sigma_{i,i-3} = -1$, is equivalent to the WLSE obtained using the kernel (10).

Hannan (1970, p. 186-187), and Wallis (1983) observed this equivalence by referring to

the derivation of the Henderson filter as the solution of the constrained minimisation problem: $\min_{\mathbf{w}}\{\mathbf{w}'\mathbf{\Sigma}_3\mathbf{w}\}$ subject to $\mathbf{w}'\mathbf{X} = \mathbf{e}_1'$, where the linear constraints enforce the condition that the trend estimate reproduces a cubic polynomial. In the next section we prove a more general result that encompasses this equivalence. Notice also that in our approach, the Henderson kernel need not be associated to local cubic polynomial estimation, as it can be defined for any value of d.

5 Local polynomial regression and higher order conditions

This section discusses whether the candidate kernel satisfies the additional equivalence conditions. It will turn out, for instance, that when $\mathbf{X}$ is a polynomial of order $d \geq 1$, the kernel derived above for the AR(p) process $(1 - \phi B)^p \varepsilon_t = \xi_t$ does not satisfy the other conditions. On the other hand, these conditions are automatically satisfied by the candidate kernels arising from the strictly non-invertible MA process $\varepsilon_t = (1 - B)^q \xi_t$, as it is stated in the following proposition.

Proposition 1 *The kernel (9) is optimal for the non-invertible MA(q) process (8).*

The proof, provided in the appendix, is based on the fact that the covariance matrix of the non-invertible MA(q) process (8) is associated with the finite difference operator of order $2q$, $\Delta^{2q}, \Delta = (1 - B)$, subject to null boundary conditions. At the same time, the matrix $\mathbf{K}_q$, which is the diagonal matrix with diagonal elements given by the candidate kernel (9), has elements that lie on a polynomial of the same order, $2q$. In the local polynomial regression setting considered so far, the convolution of these operators act onto symmetric or skew-symmetric vectors, such as the columns of $\mathbf{X}$, leaving unchanged their symmetric and polynomial structure. As a result the column space of $\mathbf{\Sigma}_q \mathbf{K}_q \mathbf{X}$ coincides with that of $\mathbf{X}$.

5.1 Local linear regression

When $d = 1$, then, following the considerations in section 4, $\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}$ and $\mathbf{X}'\mathbf{K}\mathbf{X}$ are diagonal, and so is the matrix $\mathbf{M}$ satisfying $\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^{-1}\mathbf{D} = \mathbf{K}\mathbf{X}\mathbf{M}^{-1}$. It therefore follows that necessary and sufficient conditions for $\hat{\boldsymbol{\beta}}_{GLS} = \hat{\boldsymbol{\beta}}_{WLS}$ are $\mathbf{\Sigma}^{-1}\mathbf{x}_r \propto \mathbf{K}\mathbf{x}_r$ for $r = 1, 2$, where $\mathbf{x}_r$ denotes the r-th column of the $\mathbf{X}$ matrix, i.e.

$$\sum_{i=-h}^{h} \varsigma_{ij} \propto \kappa_j, \quad \sum_{i=-h}^{h} i\varsigma_{ij} \propto j\kappa_j, \quad j = -h, ..., h \tag{11}$$

Alternatively, using the matrix equations $\mathbf{\Sigma}\mathbf{K}\mathbf{X} = \mathbf{X}\mathbf{M}^{-1}\mathbf{D}\mathbf{M}$, which for this case reduce to $\mathbf{\Sigma}\mathbf{K}\mathbf{X} = \mathbf{X}\mathbf{D}, \mathbf{D} = \mathrm{diag}(d_1, d_2)$, and writing $\mathbf{\Sigma} = \{\sigma_{ij}\}$, the necessary and sufficient conditions become

$$\sum_{i=-h}^{h} \sigma_{ij}\kappa_i = d_1, \quad \sum_{i=-h}^{h} i\sigma_{ij}\kappa_i = d_2 j, \quad j = -h, ..., h. \tag{12}$$

It is straightforward to see that the candidate kernels derived for the AR(p) process do not satisfy the above conditions. On the other hand, (11) can be verified using the expression of ς_{ij} in Lemma 5, p. 397, of Hoskins and Ponzo (1972), whereas (12) can be verified either directly or using Theorem 1, p. 394, of Hoskins and Ponzo (1972).

5.2 Local quadratic regression

In the case $d = 2$, the expressions for $\mathbf{M}$ and its inverse are:

$$
\mathbf{M} = \begin{bmatrix} m_{11} & 0 & m_{13} \\ 0 & m_{22} & 0 \\ 0 & 0 & m_{33} \end{bmatrix}, \quad \mathbf{M}^{-1} = \begin{bmatrix} m^{(11)} & 0 & m^{(13)} \\ 0 & m^{(22)} & 0 \\ 0 & 0 & m^{(33)} \end{bmatrix}
$$

and, therefore, the following further condition, besides (11) and (12), is required:

$$
\sum_{i=-h}^{h} i^2 \varsigma_{ij} = \frac{1}{d_3} \left[j^2 + \frac{m^{(13)}}{m^{(33)}} \left(1 - \frac{d_3}{d_1} \right) \right] \kappa_j \quad \text{for } j = -h, ..., h. \tag{13}
$$

In the first order moving average case it is convenient to work with $\mathbf{\Sigma K X} = \mathbf{X M}^{-1}\mathbf{D M}$. The first two conditions are as before, and the third can be written as the difference equation:

$$
\kappa_{j+1} + \kappa_{j-1} = -\frac{m^{(13)}}{m^{(33)}}(d_3 - d_1) + (d_3 + d_1 - 2d_2)j^2. \tag{14}
$$

It is immediate to check that (14) holds for the Epanechnikov kernel and the higher order kernels (9).

5.3 Local cubic regression: the Henderson filters

In the case $d = 3$,

$$
\mathbf{M} = \begin{bmatrix} m_{11} & 0 & m_{13} & 0 \\ 0 & m_{22} & 0 & m_{24} \\ 0 & 0 & m_{33} & 0 \\ 0 & 0 & 0 & m_{44} \end{bmatrix}, \quad \mathbf{M}^{-1} = \begin{bmatrix} m^{(11)} & 0 & m^{(13)} & 0 \\ 0 & m^{(22)} & 0 & m^{(24)} \\ 0 & 0 & m^{(33)} & 0 \\ 0 & 0 & 0 & m^{(44)} \end{bmatrix}
$$

so that a fourth condition besides (11), (12) and (13) has to be satisfied, which involves odd powers of j,

$$
\sum_{i=-h}^{h} i^3 \varsigma_{ij} = \frac{1}{d_4} \left[j^3 + \frac{m^{(24)}}{m^{(44)}} \left(1 - \frac{d_4}{d_2} \right) j \right] \kappa_j \quad \text{for } j = -h, ..., h,
$$

where the proportionality constant is d_4^{-1}.

In terms of the difference equation $\mathbf{\Sigma K X M}^{-1} = \mathbf{X M}^{-1}\mathbf{D}$, when ε_t is a first order moving average error term, the conditions that a kernel has to satisfy are the following:

$$- \kappa_{j-1} + 2\kappa_j - \kappa_{j+1} = d_1, \tag{15}$$

$$\kappa_{j-1} - \kappa_{j+1} = j(d_2 - d_1), \tag{16}$$

$$\kappa_{j-1} + \kappa_{j+1} = -\frac{m^{(13)}}{m^{(33)}}(d_3 - d_1) + (d_3 + d_1 - 2d_2)j^2, \tag{17}$$

$$(d_1 - 3d_2 + d_3 - d_4)j^3 = j \left[\frac{m^{(24)}}{m^{(44)}}(d_4 - d_2) - (d_2 - d_1) - 3\frac{m^{(13)}}{m^{(33)}}(d_3 - d_1) \right]. \tag{18}$$

Note that for a strictly non-invertible MA process, (18) is always satisfied by the Epanechnikov kernel, given that both $d_1 - 3d_2 + d_3 - d_4$ and $\frac{m^{(24)}}{m^{(44)}}(d_4 - d_2) - (d_2 - d_1) - 3\frac{m^{(13)}}{m^{(33)}}(d_3 - d_1)$ are null quantities.

6   Kernel smoothing

In this section we consider the inverse problem of reconstructing, if there exists, a covariance structure (i.e. some stochastic process) for which a given kernel estimator is BLUE. Hence, the starting point of this section is a set of kernel weights. With respect to local polynomial regression, that has a long tradition for smoothing time series (see Macaulay, 1931), kernel estimators for the fixed design regression problem (5) are of more recent origin (Priestley and Chao, 1972). The equivalence between the two methods has been explored by Müller (1987), who pointed out how kernel estimation is a particular case of local polynomial regression where locally weighted averaging is performed instead of locally weighted regression and kernel weights are given explicitly as $\mathrm{w}_j = \kappa_j(\sum_{j=-h}^{h} \kappa_j)^{-1}$.

Writing, as before, $\boldsymbol{\kappa} = [\kappa_h, \ldots, \kappa_1, \kappa_0, \kappa_1, \ldots, \kappa_h]'$, the vector containing the elements of a given symmetric and positive kernel with associated diagonal matrix $\mathbf{K}$, up to some constant, we can express condition (7) as follows:

$$\boldsymbol{\Sigma}\boldsymbol{\kappa} = \mathbf{i}. \tag{19}$$

We assume that $\boldsymbol{\Sigma}$ represents the covariance structure of a stationary stochastic process, and therefore that it is a symmetric, positive definite and Toeplitz matrix completely characterised by its first row or column elements, collected in the vector $\boldsymbol{\sigma} = [\sigma_{11}, \sigma_{12}, \sigma_{13}, ..., \sigma_{1,2h+1}]'$. Hence, (19) can be written as

$$\mathcal{K}\boldsymbol{\sigma} = \mathbf{i} \tag{20}$$

where

$$
\mathcal{K} = \begin{bmatrix}
\kappa_h & \kappa_{h-1} & \dots & \kappa_{h-1} & \kappa_h \\
\kappa_{h-1} & \kappa_h + \kappa_{h-2} & \dots & \kappa_h & 0 \\
\vdots & \vdots & \dots & \dots & \vdots \\
\kappa_0 & 2\kappa_1 & \dots & 0 & 0 \\
\vdots & \vdots & \dots & \dots & \vdots \\
\kappa_{h-1} & \kappa_h + \kappa_{h-2} & \dots & \kappa_h & 0 \\
\kappa_h & \kappa_{h-1} & \dots & \kappa_{h-1} & \kappa_h
\end{bmatrix}.
$$

It is evident that the linear system (20) is redundant: the last h rows of the complete matrix $[\mathcal{K}|\mathbf{i}]$ can be deleted thus giving rise to a system of $h+1$ equations in $2h+1$ unknown variables, let us denote it by $\mathcal{K}^\dagger \boldsymbol{\sigma} = \mathbf{i}^\dagger$, where the symbol $\dagger$ indicates that only the first $h+1$ rows of $\mathcal{K}$ and $\mathbf{i}$ are selected. As long as the rank of $\mathcal{K}^\dagger$ is equal to that of $[\mathcal{K}^\dagger|\mathbf{i}^\dagger]$, the system admits infinite solutions depending on the values assumed by h variables, namely $\sigma_{1,h+2}, \sigma_{1,h+3}, \dots, \sigma_{1,2h+1}$. Choices for the latter values that restitute a unique solution can be obtained by setting all the free variables equal to zero or by selecting the minimum norm solution which is the orthogonal projection onto the row space of $\mathcal{K}^\dagger$. These are not always amenable choices, since they may lead to non positive definite or singular covariance matrices.

When $h = 1$, explicitly solving (20) gives a symmetric Toeplitz matrix whose first row or column elements, depending on any value of σ_{13}, are:

$$
\begin{aligned}
\sigma_{11} &= \kappa_0^{-1} - 2\kappa_1\kappa_0^{-1}\sigma_{12} \\
\sigma_{12} &= (\kappa_0 - \kappa_0\kappa_1\sigma_{13} - \kappa_1)(\kappa_0^2 - 2\kappa_1^2)^{-1} \\
\sigma_{13} &= \text{free parameter.}
\end{aligned}
\tag{21}
$$

When $h > 1$, analytic solutions become rather complicate to calculate. Anyway, exact numerical solutions may be found by solving the linear system (20) using scale reduction algorithms. For example, take the $\mathbf{QR}$ decomposition of $\mathcal{K}^\dagger$ and then back-solve $\mathbf{R}\boldsymbol{\sigma} = \mathbf{Q}'\mathbf{i}^\dagger$.

Admissible solutions exist for $h \geq 1$ when the Epanechnikov, the biweight or the tricube kernels are chosen. The latter arise for values of s equal to 2 and 3, respectively, in the following equation

$$
\kappa_j \propto \left(1 - \left|\frac{j}{h+1}\right|^s\right)^s, \quad j = -h, \dots, h.
\tag{22}
$$

and are the suggested weighting functions in the robust locally weighted regression method (loess) developed by Cleveland (1979).

On the other hand, not all the kernels are optimal for some stochastic process. An example is the Gaussian kernel, whose weights are

$$
\kappa_j \propto \exp\left\{-\frac{1}{2}\left(\frac{j}{b}\right)^2\right\}, \quad j = -h, \dots, h,
$$

where the $b > 0$ is the smoothing parameter determining the bandwidth. The Gaussian kernel arises as the probability density function of the infinite sum of independent rectangular random

variables, and is largely applied for density estimation. Despite its popularity, for $h = 1$ there does not exist any value of σ_{13} such that the resulting Σ is positive definite and our numerical analysis seems to reveal that no admissible covariance structures may be derived for larger bandwidths. In other words, our empirical evidence induces to conclude that there does not exist any stochastic process for which the Gaussian kernel is BLUE. The same occurs with the triweight kernel

$$\kappa_j \propto \frac{35}{32} \left[1 - \left(\frac{j}{h+1} \right)^2 \right]^3 , \ j = -h, ..., h$$

and with the triangle kernel arising when $s = 0$ in (22). Note that when h is large, the weights of the polynomial kernel (10), giving the Henderson filters, become approximately proportional those of the triweight kernel (see Loader, 1999, Ex. 1.6, and Müller, 1984). When h is not too large, the approximation is not sensible and boundary conditions make the difference between the two estimators, even with respect to their Gauss-Markov optimality.

7 Conclusions

The paper has proven a general result establishing the conditions under which generalised least squares estimation is equivalent to weighted least squares estimation. The result has relevant implications for kernel smoothing in local polynomial framework. In particular it allowed to derive a class of polynomial kernels that are isomorphic to covariance structures associated to non invertible moving average processes for the errors, that encompass well known kernels such as Epanechnikov and the Henderson kernel.

A   Proofs of the main results

In this section, we provide the proofs of Theorem 1, Corollary 1 and Proposition 1. The proof of Theorem 1 requires a result concerning the simultaneous diagonalisation of two symmetric positive definite matrices (Lemma 1), which is a particular case of a well known result (see Magnus and Neudecker, 2007, Theorem 23, p. 23).

Lemma 1 *Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric and positive definite matrices of the same order. Then, a non singular matrix $\mathbf{C}$ and a positive definite diagonal matrix $\mathbf{D}$ exist, such that $\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{I}$ and $\mathbf{C}'\mathbf{B}\mathbf{C} = \mathbf{D}$ where $\mathbf{I}$ is the identity matrix.*

Proof Since $\mathbf{A}$ is symmetric and positive definite, it can be factorised as $\mathbf{A} = (\mathbf{E}^{-1})'\mathbf{E}^{-1}$ (e.g. by a Cholesky decomposition) so that $\mathbf{E}'\mathbf{A}\mathbf{E} = \mathbf{I}$. Let denote by $\mathbf{Q}$ the orthogonal matrix that diagonalises $\mathbf{E}'\mathbf{B}\mathbf{E}$, i.e. $\mathbf{Q}'\mathbf{E}'\mathbf{B}\mathbf{E}\mathbf{Q} = \mathbf{D}$. Setting $\mathbf{C} = \mathbf{E}\mathbf{Q}$ one gets $\mathbf{C}'\mathbf{B}\mathbf{C} = \mathbf{D}$ and $\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{I}$. Note that the elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{E}'\mathbf{B}\mathbf{E}$ corresponding to the eigenvectors $\mathbf{Q}$ as well as the eigenvalues of $\mathbf{A}^{-1}\mathbf{B}$ corresponding to the eigenvectors $\mathbf{C}$, given that $\mathbf{C}^{-1}\mathbf{A}^{-1}\mathbf{B}\mathbf{C} = \mathbf{D}$, as follows by $\mathbf{A}^{-1} = \mathbf{E}\mathbf{E}'$ ∎

The proof of Theorem 1 is divided into two parts. We first prove that $\mathbf{X} = \mathbf{V}^*\mathbf{M}$, where $\mathbf{V}^*$ is a matrix whose columns contain p eigenvectors of $\mathbf{\Sigma}\mathbf{K}$, and $\mathbf{M}$ is nonsingular, is a sufficient condition for $\boldsymbol{\beta}_{GLS} = \boldsymbol{\beta}_{WLS}$ and then that the equivalence $\boldsymbol{\beta}_{GLS} = \boldsymbol{\beta}_{WLS}$ implies that we can express $\mathbf{X} = \mathbf{V}^*\mathbf{M}$ (necessity).

Proof of Theorem 1. *(Sufficiency)* Let us assume that $\mathbf{X} = \mathbf{V}^*\mathbf{M}$ where $\mathbf{V}^*$ contains, as columns, p eigenvectors of $\mathbf{\Sigma}\mathbf{K}$ and $\mathbf{M}$ is a non singular matrix. The condition on $\mathbf{V}^*$ can be formalised as follows,

$$(\mathbf{\Sigma}\mathbf{K})\mathbf{V}^* = \mathbf{V}^*\mathbf{\Lambda}^*$$

where $\mathbf{\Lambda}^*$ is diagonal and its elements are the eigenvalues of $\mathbf{\Sigma}\mathbf{K}$ corresponding to the eigenvectors that are columns of $\mathbf{V}^*$. Equivalently,

$$\mathbf{V}^{*'}\mathbf{\Sigma}^{-1} = \mathbf{\Lambda}^{*-1}\mathbf{V}^{*'}\mathbf{K}$$

from which follows that

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{GLS} &= (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{y} = \\
&= (\mathbf{M}'\mathbf{V}^{*'}\mathbf{\Sigma}^{-1}\mathbf{V}^*\mathbf{M})^{-1}\mathbf{M}'\mathbf{V}^{*'}\mathbf{\Sigma}^{-1}\mathbf{y} \\
&= (\mathbf{M}'\mathbf{\Lambda}^{*-1}\mathbf{V}^{*'}\mathbf{K}\mathbf{V}^*\mathbf{M})^{-1}\mathbf{M}'\mathbf{\Lambda}^{*-1}\mathbf{V}^{*'}\mathbf{K}\mathbf{y} \\
&= \left((\mathbf{M}'\mathbf{\Lambda}^{*-1}\mathbf{M}'^{-1})(\mathbf{M}'\mathbf{V}^{*'})\mathbf{K}(\mathbf{V}^*\mathbf{M})\right)^{-1}(\mathbf{M}'\mathbf{\Lambda}^{*-1}\mathbf{M}'^{-1})(\mathbf{M}'\mathbf{V}^{*'})\mathbf{K}\mathbf{y} \\
&= \left((\mathbf{M}'\mathbf{V}^{*'})\mathbf{K}(\mathbf{V}^*\mathbf{M})\right)^{-1}(\mathbf{M}'\mathbf{V}^{*'})\mathbf{K}\mathbf{y} \\
&= (\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}\mathbf{y} \\
&= \hat{\boldsymbol{\beta}}_{WLS}.
\end{aligned}
$$

(*Necessity*) The equality between the WLSE (4) and the GLSE (2) implies

$$\mathbf{KX}(\mathbf{X}'\mathbf{KX})^{-1} = \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1},$$

or, equivalently,

$$\boldsymbol{\Sigma}\mathbf{KX} = \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{KX}). \tag{23}$$

Since $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$ and $\mathbf{X}'\mathbf{KX}$ are positive definite and symmetric, by Lemma 1, there exists a non singular matrix $\mathbf{C}$ such that $\mathbf{C}'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})\mathbf{C} = \mathbf{I}$, and $\mathbf{C}'(\mathbf{X}'\mathbf{KX})\mathbf{C} = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix containing the eigenvalues of $(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{KX})$ corresponding to the (eigenvectors) columns of $\mathbf{C}$. Hence, replacing $(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{KX}) = \mathbf{CDC}^{-1}$ into (23), gives $\boldsymbol{\Sigma}\mathbf{KX} = \mathbf{XCDC}^{-1}$, or

$$(\boldsymbol{\Sigma}\mathbf{K})\mathbf{XC} = (\mathbf{XC})\mathbf{D}.$$

The latter equality tells that the p columns of $\mathbf{XC}$ are eigenvectors of $\boldsymbol{\Sigma}\mathbf{K}$ with corresponding eigenvalues given by $\mathbf{D}$. Setting $\mathbf{XC} = \mathbf{V}^*$ and $\mathbf{M} = \mathbf{C}^{-1}$ proves the theorem ∎

Proof of Corollary 1. If Theorem 1 holds, then $(\boldsymbol{\Sigma}\mathbf{K})\mathbf{XC} = (\mathbf{XC})\mathbf{D}$. Pre-multiplying both members of the latter equation by $\mathbf{H}$ and reminding that $\mathbf{HX} = \mathbf{X}$ the result is

$$(\mathbf{H}\boldsymbol{\Sigma}\mathbf{K})\mathbf{XC} = \mathbf{HXCD} = (\mathbf{XC})\mathbf{D}.$$

On the other hand, let us consider the matrix $\boldsymbol{\Sigma}\mathbf{KH}$. It is evident that

$$(\boldsymbol{\Sigma}\mathbf{KH})\mathbf{XC} = \boldsymbol{\Sigma}\mathbf{KXC} = (\mathbf{XC})\mathbf{D}.$$

Up to now we have proved that if theorem 1 holds, then $\mathbf{H}\boldsymbol{\Sigma}\mathbf{K}$ and $\boldsymbol{\Sigma}\mathbf{KH}$ share p eigenvectors (the same of $\boldsymbol{\Sigma}\mathbf{K}$ that are linear combinations of the columns of $\mathbf{X}$) associated to equal eigenvalues. If we show that $\mathbf{H}\boldsymbol{\Sigma}\mathbf{K}$ and $\boldsymbol{\Sigma}\mathbf{KH}$ also share other $2h - p$ independent eigenvectors associated to equal eigenvalues we have proved that the two matrices are equal. To do that, remind by section 2 that $\mathbf{H}$ is the (oblique) projection matrix onto $\mathcal{C}(\mathbf{X})$ along $\mathcal{N}(\mathbf{X}'\boldsymbol{\Sigma}^{-1})$ (see Meyer, 2000, pag. 634), or equivalently along $\mathcal{N}(\mathbf{X}'\mathbf{K})$, since the projector is unique. Therefore $\mathbf{H}$ is diagonalisable and has p eigenvectors in $\mathcal{C}(\mathbf{X})$ associated to eigenvalues equal to one and $n - p$ eigenvectors in $\mathcal{N}(\mathbf{X}'\boldsymbol{\Sigma}^{-1})$ or $\mathcal{N}(\mathbf{X}'\mathbf{K})$ associated to null eigenvalues. As such, the latter eigenvectors are all those $\mathbf{z} \in \mathbb{R}^n$ such that $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{z} = \mathbf{X}'\mathbf{Kz} = \underline{0}$. The same $\mathbf{z}$ are eigenvectors of $\mathbf{H}\boldsymbol{\Sigma}\mathbf{K}$ and $\boldsymbol{\Sigma}\mathbf{KH}$ associated to zero eigenvalues as well. In fact, $\forall \mathbf{z} \in \mathcal{N}(\mathbf{X}'\mathbf{K}), \boldsymbol{\Sigma}\mathbf{KHz} = \underline{0}$ and $\mathbf{H}\boldsymbol{\Sigma}\mathbf{Kz} = \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Kz} = \underline{0}$.

On the other hand, if $\mathbf{H}_{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\Sigma}\mathbf{K} = \boldsymbol{\Sigma}\mathbf{KH}_K$, then $\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K} = \boldsymbol{\Sigma}\mathbf{KX}(\mathbf{X}'\mathbf{KX})^{-1}\mathbf{X}'\mathbf{K}$ and pre-multiplying by $\boldsymbol{\Sigma}^{-1}$ and post-multiplying by $\mathbf{K}^{-1}$ one obtains $\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}' = \mathbf{KX}(\mathbf{X}'\mathbf{KX})^{-1}\mathbf{X}'$ that is $\mathbf{H}'_{\boldsymbol{\Sigma}^{-1}} = \mathbf{H}'_K$, implying equality of the WLSE obtained with kernel $\mathbf{K}$ and the GLSE with covariance matrix $\boldsymbol{\Sigma}$∎

Proof of Proposition 1. Let us define the matrix $\mathbf{X}^* \in \mathbb{R}^{(2(h+q)+1)\times(d+1)}$,, $0 < q < 2h$, as the local polynomial regression design matrix with $d+1$ columns and bandwidth $h^* = h + q$ (see section 3). The element in the $(h^* + j + 1) - th$ row and r-th columns, is j^{r-1}, $j = -h^*, \ldots, h^*$, $r = 1, \ldots, d+1$.

Let also $\boldsymbol{\Sigma}_q^* \in \mathbb{R}^{(2h+1)\times(2h^*+1)}$, denote the matrix formed from the central $2h + 1$ rows of the $(2h^*+1)$ dimensional covariance matrix of the noninvertible MA(q) process $\epsilon_t = (1-B)^q \xi_t$, $\xi_t \sim$ WN$(0,1)$, $t = -h^*, \ldots, h^*$, where, for instance,

$$
\boldsymbol{\Sigma}_1^* = \begin{bmatrix}
-1 & 2 & -1 & 0 & \ldots & 0 & 0 \\
0 & -1 & 2 & -1 & \ddots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & \ddots & -1 & 2 & -1 & 0 \\
0 & \cdots & \cdots & 0 & -1 & 2 & -1
\end{bmatrix}.
$$

Notice that $\boldsymbol{\Sigma}_q$ can be obtained from $\boldsymbol{\Sigma}_q^*$ by deleting the first and last q columns.

The matrix $\boldsymbol{\Sigma}_q^*$ is associated with the difference operator $(1-B)^{2q}$ subject to null boundary conditions. Specifically, $\boldsymbol{\Sigma}_q^*$ acts onto any polynomial vector of degree d by lowering its order to $d - 2q$ and by annihilating its first and last q components. Hence, for $d < 2q$, $\boldsymbol{\Sigma}_q^* \mathbf{X}_q^* = \mathbf{0}$, where $\mathbf{0}$ is the null matrix in $\mathbb{R}^{(2h+1)\times(d+1)}$, or, equivalently, $\mathcal{C}(\mathbf{X}^*) \subset \mathcal{N}(\boldsymbol{\Sigma}_q^*)$.

As the elements of each of the rows of the matrix are the coefficients of B in the expansion of $(1-B)^{2q}$, we can define a vector $\boldsymbol{\kappa}_q^*$, whose elements lie on a polynomial of degree $d = 2q$, subject to suitable boundary conditions, such that $\boldsymbol{\Sigma}_q^* \boldsymbol{\kappa}_q^* \propto \mathbf{i}$. In particular, the vector $\boldsymbol{\kappa}_q^*$ has to satisfy the following properties:

(p1) the elements of $\boldsymbol{\kappa}_q^*$ are non negative and describe a polynomial of order $2q$ in j, denoted $v_q(j)$, for $j = -(h+q), -h-(q-1), ..., h+q$;

(p2) the polynomial is null for $j = h+1, h+2, ..., h+q$ and $j = -(h+1), -(h+2), ..., -(h+q)$.

The property (p2) gives exactly $2q$ roots of $v_q(j)$. The latter can be therefore factorised as follows:

$$
\begin{aligned}
v_q(j) &= [(h+1) - j][(h+2) - j] \cdots [(h+q) - j][(h+1) + j][(h+2) + j]...[(h+q) + j] \\
&= [(h+1)^2 - j^2][(h+2^2 - j^2)] \cdots [(h+q)^2 - j^2].
\end{aligned}
$$

When combined, (p1) and (p2) give the symmetric kernel $v_q(j)$, so that $\boldsymbol{\kappa}_q^* = (\mathbf{0}_q', \boldsymbol{\kappa}_q', \mathbf{0}_q')'$ is the vector of kernel weights $\boldsymbol{\kappa}_q$ extended by inserting q zeros before and after.

Let us now define the matrix $\mathbf{K}_q^*$ which has the vector $\boldsymbol{\kappa}_q^*$ on the main diagonal and zero elements elsewhere. Hence,

$$
\mathbf{K}_q^* = \begin{bmatrix}
\mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{K}_q & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0}
\end{bmatrix}.
$$

The matrix $\boldsymbol{\Sigma}_q^* \mathbf{K}_q^*$ has row elements that are the coefficients of the convolutions of the polynomial $\upsilon_q(j)$, with the difference operator $(1 - B)^{2q}$. When applied to $\mathbf{X}^*$, the operator $\boldsymbol{\Sigma}_q^* \mathbf{K}_q^*$ leaves unchanged (up to a linear transformation) the degree and the structure of the columns of $\mathbf{X}^*$ and annihilates the first and last q elements. In other words, since the columns of $\mathbf{X}^*$ are symmetric or skew-symmetric vectors defining a polynomial basis, premultiplication by matrix $\boldsymbol{\Sigma}_q^* \mathbf{K}_q^*$ which is the product of a matrix which annihilates a polynomial of degree $2q$ and which raises the degree of a polynomial term by $2q$, yields a compensating effect, so that $\boldsymbol{\Sigma}_q^* \mathbf{K}_q^* \mathbf{X}^* \subseteq \mathcal{C}(\mathbf{X})$ or, more generally,

$$\mathbf{T}_q(\mathcal{C}(\mathbf{X}^*)) \subseteq \mathcal{C}(\mathbf{X}), \tag{24}$$

where $\mathbf{T}_q$ is the linear operator associated with $\boldsymbol{\Sigma}_q^* \mathbf{K}_q^*$, i.e. $\mathbf{T}_q(\mathbf{x}) = \boldsymbol{\Sigma}_q^* \mathbf{K}_q^* \mathbf{x}$, and $\mathbf{T}_q(\mathcal{C}(\mathbf{X}^*)) = \{\mathbf{T}_q(\mathbf{x}), \mathbf{x} \in \mathcal{C}(\mathbf{X}^*)\}$.

Now, direct multiplication shows that,

$$\boldsymbol{\Sigma}_q^* \mathbf{K}_q^* \mathbf{X}^* = \boldsymbol{\Sigma}_q \mathbf{K}_q \mathbf{X}, \tag{25}$$

and combining (24) with (25) gives

$$\mathbf{T}(\mathcal{C}(\mathbf{X})) \subseteq \mathcal{C}(\mathbf{X}) \tag{26}$$

where $\mathbf{T}(\mathcal{C}(\mathbf{X})) = \{\boldsymbol{\Sigma}_q \mathbf{K}_q \mathbf{x}, \mathbf{x} \in \mathcal{C}(\mathbf{X})\}$, i.e. $\mathcal{C}(\mathbf{X})$ is an invariant subspace of $\mathbb{R}^{2h+1}$ under $\mathbf{T}$ (see Meyer, 2000, pag. 259). Since $\mathbf{T}(\mathcal{C}(\mathbf{X}))$ and $\mathcal{C}(\mathbf{X})$ have the same dimension, equal to $d + 1$, equality holds in (26). It follows that there exist $d + 1$ linearly independent vectors $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_{d+1}$ such that $\boldsymbol{\Sigma}_q \mathbf{K}_q \mathbf{X} \mathbf{c}_i = \mathbf{X} \mathbf{c}_i d_i$ for some coefficients d_i, $i = 1, 2, ..., d + 1$. In matrix notation, setting $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_{d+1}]$ and $\mathbf{D} = \text{diag}\{d_1, d_2, ..., d_{d+1}\}$, the above relation becomes

$$\boldsymbol{\Sigma}_q \mathbf{K}_q \mathbf{X} \mathbf{C} = \mathbf{X} \mathbf{C} \mathbf{D}.$$

Hence, we have proved that there exist $d + 1$ linear combinations of the columns of $\mathbf{X}$ that are eigenvectors of $\boldsymbol{\Sigma}_q \mathbf{K}_q$, i.e. the kernel (9) is optimal in the Gauss-Markov sense for the non-invertible MA(q) process (8).

Equivalently one could have noted that equation (26) implies that $\mathcal{C}(\mathbf{X}) \subseteq \mathcal{N}(\boldsymbol{\Sigma}\mathbf{K} - d\mathbf{I})$, where d is any eigenvalue of $\boldsymbol{\Sigma}\mathbf{K}$ (see Meyer, 2000, pag. 265). Since the dimension of $\mathcal{C}(\mathbf{X})$ is equal to $d + 1$, the above relation establishes that there exist $d + 1$ linear combinations of the columns of $\mathbf{X}$ that are eigenvectors of $\boldsymbol{\Sigma}_q \mathbf{K}_q$ ∎

References

Aitken, A.C. (1935), On the Least Squares and Linear Combinations of Observations, *Proceedings of the Royal Society of Edinburgh*, A, 55, 42-48.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge, U.S.

Anderson, T.W. (1948), On the Theory of Testing Serial Correlation, *Skandinavisk Aktuarietidskrift*, 31, 88-116.

Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, Wiley.

Baksalary, J.K., Kala, R. (1983), On Equalities Between BLUEs, WLSEs and SLSEs, *The Canadian Journal of Statistics*, 11, 119-123.

Baksalary, J.K., Van Eijnsbergeren, A.C (1988), A Comparison of Two Criteria For Ordinary-Least-Squares Estimators To Be Best Linear Unbiased Estimators, *The American Statistician* , 42, 205-208.

Benedetti, J.K. (1977), On the Nonparametric Estimation of Regression Functions, *Journal of the Royal Statistical Society, ser. B*, 39, 248-253.

Cleveland, W.S. (1979), Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, 64, 368, 829-836.

Epanechnikov V.A. (1969), Nonparametric Estimation of a Multivariate Probability Density, *Theory of Probability and Applications*, 14, 153-158.

Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen B. (1998). New Capabilities and Methods of the X12-ARIMA Seasonal Adjustment Program, *Journal of Business and Economic Statistics*, 16, 2.

Grenander, U., Rosenblatt, M. (1957), *Statistical analysis of stationary time series*, John Wiley and Sons, New York.

Hannan, E.J. (1970), *Multiple Time Series*, John Wiley and Sons, New York.

Henderson, R. (1916), Note on Graduation by Adjusted Average, *Transaction of the Actuarial Society of America*, 17, 43-48.

Hoskins, W.D., Ponzo P.J. (1972), Some Properties of a Class of Band Matrices, *Mathematics of Computation*, 26, 118, 393-400.

Kenny P.B., and Durbin J. (1982), Local Trend Estimation and Seasonal Adjustment of Economic and Social Time Series, *Journal of the Royal Statistical Society A*, 145, I, 1-41.

Krämer, W. (1980), A Note on the Equality of Ordinary Least Squares and Gauss-Markov Estimates in the General Linear Model, *Sankhyā*, A, 42, 130-131.

Krämer, W. (1986), Least-Squares Regression when the Independent Variable Follows an ARIMA Process, *Journal of the American Statistical Association*, 81, 150-154.

Krämer, W., Hassler, U. (1998), Limiting Efficiency of OLS vs. GLS When Regressors Are Fractionally Integrated, *Economics Letters*, 60, 3, 285-290.

Kruskal, W. (1968), When Are Gauss-Markov and Least Squares Estimators Identical? A Coordinate-Free Approach, *The Annals of Mathematical Statistics*, 39, 70-75.

Jaeger A., Krämer, W. (1998), A Final Twist on the Equality of OLS and GLS, *Statistical Papers*, 39, 321-324.

Ladiray, D. and Quenneville, B. (2001). *Seasonal Adjustment with the X-11 Method* (Lecture Notes in Statistics), Springer-Verlag, New York.

Loader, C. (1999), *Local regression and likelihood*, Springer-Verlag, New York.

Lowerre, J. (1974), Some Relationships Between BLUEs, WLSEs and SLSEs, *Journal of the American Statistical Association*, 69, 223-225.

Macaulay, F.R. (1931), *The Smoothing of Time Series*, New York: National Bureau of Economic Research.

Magnus J.R. and Neudecker H. (2007), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Third edition, John Wiley & Sons.

McElroy F.W. (1967), A Necessary and Sufficient Condition That Ordinary Least-Squares Estimators Be Best Linear Unbiased, *Journal of the American Statistical Association*, 62, 1302-1304.

Meyer C.D. (2000), *Matrix Analysis and Applied Linear Algebra*, SIAM.

Müller, H.G. (1984), Smooth Optimum Kernel Estimators of Densities, Regression Curves and Modes, *The Annals of Statistics*, 12, 2, 766-774.

Müller, H.G. (1987), Weighted Local Regression and Kernel Methods for Nonparametric Curve Fitting, *Journal of the American Statistical Association*, 82, 231-238.

Nadaraya, E.A. (1964), On Estimating Regression, *Theory of Probability and its Applications*, 9, 141-142.

Phillips P. C.B. (1992), Geometry of the Equivalence of OLS and GLS in the Linear Model, *Econometric Theory*, 8, 1, 158-159.

Phillips P.C.B., Park J.Y. (1992), Asymptotic Equivalence of OLS and GLS in Regressions with Integrated Regressors, *Journal of the American Statistical Association*, 83, 111-115.

Priestley, M.B., Chao M.T. (1972), Nonparametric Function Fitting, *Journal of the Royal Statistical Society, ser. B*, 34, 384-392.

Puntanten S., Styan, G.P.H. (1989), The Equality of the Ordinary Least Squares Estimator and the Best Linear Unbiased Estimator, *The American Statistician*, 43, 3, 153-161.

Wallis, K. (1983). Models for X-11 and X-11 Forecast Procedures for Preliminary and Revised Seasonal Adjustments. In *Applied Time Series Analysis of Economic Data* (A. Zellner, ed.), pp. 3-11. Washington DC: Bureau of the Census, 1983.

Wand M.P. and Jones M.C. (1995), *Kernel Smoothing*, Monographs on Statistics and Applied Probability, 60, Chapman&Hall.

Watson, G.S. (1967), Linear Least Squares Regression, *The Annals of Mathematical Statistics*, 38, 1679-1699.

Watson, G.S. (1964), Smooth Regression Analysis, *Sankhyā*, A, 26, 359-372.

Zyskind, G. (1967), On Canonical Forms, Non-Negative Covariance Matrices and Best and Simple Least Squares Linear Estimators in Linear Models, *The Annals of Mathematical Statistics*, 38, 1092-1119.

Zyskind, G. (1969), Parametric Argumentations and Error Structures Under Which Certain Simple Least Squares And Analysis of Variance Procedures Are Also Best, *Journal of the American Statistical Association*, 64, 1353-1368.

Zyskind, G., Martin, F.B. (1969), On Best Linear Estimation and a General Gauss-Markoff Theorem in Linear Models with Arbitrary Non-negative Structure, *SIAM Journal of Applied Mathematics*, 17, 1190-1202.