Mathematical Statistics
Stockholm University

Krylov sequences as a tool for analysing iterated regression algorithms

Anders Björkström

Research Report 2007:11

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2007:11**,
http://www.math.su.se/matstat

Krylov sequences as a tool for analysing iterated regression algorithms

Anders Björkström*

May 2007

Abstract

We use Krylov sequences to analyze more regression methods than PLSR. Some results already proven for PLSR are shown to hold for other methods also. We prove that the well-known peculiar pattern of alternating shrinkage and inflation is not unique for PLSR. We also show that for any method in a wide class, the coefficient of determination is, for any data, at least as high as for PCR with the same number of components.

Key words: Continuum regression, shrinkage regressors, PLSR, PCR, Krylov sequence

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: bjorks@math.su.se.

1 Introduction

Several regression methods are defined in a recursive manner: A basic algorithm is executed a number of times, identifying one "factor" at each iteration, and the regression coefficients are finally determined by ordinary least-squares regression of the response variable on all the factors. Iterative methods often perform well, but a drawback is that their mathematical and statistical properties are seldom easy to analyze. However, in the important special case of PLS regression, it is often fruitful to resort to a geometrical interpretation, pointed out by Helland (1988). He demonstrated that the vector $\widehat{y}$ consisting of the values fitted by a factor PLS is the projection of the data vector y on a subspace of the span of the explanatory variables, a subspace that can be described as the span of a certain Krylov sequence. In the present paper we show that a wider class of regression methods share the property that $\widehat{y}$ can be interpreted as the projection of y on the span of a Krylov sequence. Consequently, conclusions about PLSR, drawn by utilizing the properties of Krylov sequences, will, after appropriate modification, be valid for these other regression methods also. We discuss two examples of this. One concerns shrinkage properties, the other is a comparison with principal components regression (PCR).

1.1 Notation and terminology

Throughout this paper, we assume that the response variable y is univariate. We denote by X the centered $n \times p$ matrix of data on explanatory variables, by y the centered n-vector of response data. We assume that y is in the column span of X, since the component of y orthogonal to span(X) plays no role in standard regression methods. The singular value decomposition of X is denoted $X = U\Lambda^{1/2}V^T$, where U is $n \times p$, Λ is $p \times p$ diagonal, and V is $p \times p$. Consequently, we have the spectral decompositions $XX^T = U\Lambda U^T$ and $X^TX = V\Lambda V^T$. The eigenvalues of Λ will be indexed in nonincreasing order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. We use the letter b (with different superscripts or indices) for regression coefficient vectors (p-vectors), and g for the coefficients when expressing b as linear combinations of the column vectors of V: $b = Vg$. Similarly we use w for the coefficient vector for y when written

as a linear combination of the column vectors of U: $y = Uw$. We treat the case $\mathrm{rank}(X) = p \leq n - 1$, but similar conclusions will be valid for other cases also.

We use the term *regression method* (or just *method*) for any algorithm or other instruction that uniquely defines a p-vector b given any matrix X and any data vector y with the same number of rows as X. In mathematical terminology, a method is a function from $\mathbf{R}^{n\times p} \times \mathbf{R}^n$ to $\mathbf{R}^p$. We write $b = \mathcal{B}(X, y)$.

1.2 Iterative methods

For any regression method except OLSR, the residual $r = y - Xb$ may have a nonzero component in the column span of X. It may then be of interest to "use the residual as input", *i.e.*, evaluate $b' = \mathcal{B}(X, r)$. By finding a new regression vector b_2 which is a linear combination of b and b' one often improves not only the coefficient of explanation but also the predictive capability of the method. If this refinement is repeated $a-1$ times, we may denote the resulting iterated method $\mathcal{B}_{\mathrm{it}}(X, r; a)$. The rationale for this recursive procedure is the belief that the variation in data is caused by a number a of latent variables. It is tacitly assumed that only the most influential latent variable will be captured adequately by the first factor *i.e* by $b_1 = \mathcal{B}(X, y)$, so a sequence of adjustments are necessary to correct for the influence of the second, third, and so on. For the purpose of the present paper, however, this statistical motivation is less essential than the exact procedure of the iteration. Therefore, we give a more detailed description of it in Appendix A.

1.3 Organization of the paper

The purpose of this paper is to derive two results concerning a certain type of regression methods. In section 2, we define the type. In section 3 we derive a conclusion about their shrinkage properties, and in section 4 we compare them to PCR.

2 Semi-linear methods

Definition 1. A method $\mathcal{B}(X, y)$ is called *semi-linear* if it can be written

$$\mathcal{B}(X, y) = \kappa\, M\, X^T y, \tag{1}$$

where the $(p \times p)$-matrix M does not depend on y. The number κ may depend on both X and y.

Comments:

- When a method is called *linear*, one means that for constant X, $\mathcal{B}$ varies linearly with y. OLSR and PCR (with a pre-determined number of factors) are linear, but most other methods are not. Several methods attempt to reduce the variance of the OLS regressor by replacing the ill-conditioned $X^T X$ by some alternative, better suited for inversion. The replacement is often dependent on y, making $\mathcal{B}$ a non-linear function. We now study a class of methods where the nonlinearity has a particularly simple form, *viz.*, in the factorization (1), only the scalar κ depends on y, while the matrix M does not.

- One-factor PLSR is a semi-linear method, with $M = I_p$, the $p \times p$ identity matrix.

- Ridge regression (RR) and its least-squares adjusted form LSRR (Björkström & Sundberg, 1999) are semi-linear methods if the ridge parameter is pre-determined. In this case, $M = (X^T X + \delta I_p)^{-1}$, where δ is the ridge parameter. An alternative, equivalent parameterization is $M = ((1-\alpha)X^T X + \alpha I_p)^{-1}$.

- Continuum regression, CR (Stone & Brooks, 1990) is not a semi-linear method. This may seem as a paradox, given the close relationship (Sundberg, 1993) between CR and RR. CR is similar to RR in that the first factor can be written $b_{CR} = \kappa\,(X^T X + \delta I_p)^{-1} X^T y$. However, in CR the ridge parameter δ is determined from a maximization criterion that involves y.

- Continuum power regression, CPR (Wise & Ricker, 1993; de Jong et al, 2001) is a semi-linear method if their method parameter μ is pre-determined. Here, $M = V \Lambda^{2(\mu-1)} V^T$.

- The role of the scalar κ is, usually, to ascertain orthogonality between residuals and fitted values (this happens when $\kappa = y^T X M X^T y / |M X^T y|^2$). If we choose κ in any other way, the resulting method is not really fit for being iterated, but it seems nevertheless convenient to let Definition 1 admit other scalar functions κ also.

- PCR with a factors is a semi-linear method, with $M = (X^T X)^{-1}$ truncated to its first a eigencomponents. However, in PCR subsequent factors are not formed by iteration as defined in Appendix A. Therefore, our conclusions (Propositions 1 and 2) are not applicable to PCR.

Helland's (1988) result says that for PLSR with a factors, the vector b of regression coefficients satisfies the equation

$$X\, b = P_a\, y \tag{2}$$

where the $n \times n$ matrix P_a denotes projection on a certain subspace of span(X), *viz.* the Krylov sequence

$$\mathcal{S}_A = \text{span}[\, XX^T y,\ (XX^T)^2 y\ \ldots (XX^T)^a y\,]. \tag{3}$$

We now formulate a generalization of this.

Proposition 1 *If $\mathcal{B}(X, y)$ is a semi-linear method, then $\mathcal{B}_{it}(X, r; a)$ yields a p-vector b_a satisfying $X b_a = P_a\, y$, where P_a denotes projection on the column span of a Krylov sequence, viz.*

$$\mathcal{S}_H(a) = \mathit{span}[\, Hy, H^2 y, \ldots, H^a y\,]$$

where $H = XMX^T$, and M is the matrix that occurs in the definition of the method $\mathcal{B}$ (equation 1).

Proof: With $M = I_p$ the proposition is a well-known result for PLSR. Our proof, also, mimics the proof for the PLSR case (Helland, 1988). The core is to establish that two subspaces $\mathcal{S}_H(a)$ and $\mathcal{S}_t(a)$ are equal, where $\mathcal{S}_t(a)$ is spanned by the a first

vectors $Xb_1, \ldots, Xb_a$ in the iteration procedure. We give the details in appendix B. By construction, b_a satisfies $Xb_a = P_a\, y$ when P_a denotes projection on $\mathcal{S}_t(a)$, so the proposition follows. ■

The subspaces $\mathcal{S}_H(a) = \mathcal{S}_t(a)$, $a = 1, 2, ...$ cannot be indefinitely increasing. From some a-value onwards, $\mathcal{S}_H(a+1)$ will be the same as $\mathcal{S}_H(a)$. This happens when $a = p$, or earlier. The corresponding b-vector will then be the OLS regressor.

2.1 Regular semi-linear methods

Proposition 1 is applicable to any semi-linear method. To proceed, we need an additional condition, which we define as follows:

Definition 2. A semi-linear method is called *regular* if there exists a function $\mu(x)$ such that M in Definition 1 can be written

$$M = V\ \mathrm{diag}(\mu(\lambda_1), \ldots, \mu(\lambda_p))\ V^T,$$

where V and Λ are from the spectral decomposition $X^TX = V\Lambda V^T$. ■

Definition 2 requires that M depends only on X^TX, that is, the covariances between the x-variables, and not on any other properties of the training data X. Most methods are of this kind. PLSR is a regular semi-linear method with $\mu(x) = 1$ (a constant function). In the following, we shall be particularly interested in two families of functions $\mu(x)$:

$$\mu(x) = \frac{1}{(1-\alpha)x\ + \alpha},$$

where α is a nonnegative number, and

$$\mu(x) = x^\alpha,$$

where $\alpha \geq -1$. PLSR is a member of both families, corresponding to $\alpha = 1$ in the first case and $\alpha = 0$ in the second case.

3 Shrinkage properties of regular semi-linear methods

3.1 Shrinkage functions

Variance reduction is the major motivation for seeking alternatives to OLSR. A useful descriptive measure characterizing an alternative method is its so-called shrinkage factors. They are defined as $\widetilde{f}_j = \widetilde{g}_j/\widehat{g}_j$, $j = 1, \ldots, p$, where the numbers $\widehat{g}_j$ descibe the OLSR regressor as a linear combination of the eigenvectors of $X^T X$, and the $\widetilde{g}_j$ do the same for the alternative method. In other words, $b_{OLS} = V\widehat{g}$ and $\widetilde{b} = V\widetilde{g}$. (It is well-known that $\widehat{g} = \Lambda^{-1/2}w$, cf section 1.1.) The shrinkage factors indicate whether the j:th singular component is "shrunk" ($\widetilde{f}_j < 1$) or "inflated" ($\widetilde{f}_j > 1$) by the alternative regressor, relative to OLSR. The *shrinkage function* $f(\lambda_j)$ relates the factor $\widetilde{f}_j$ to the corresponding eigenvalue λ_j. For example, for RR, the shrinkage function is $f(\lambda) = \lambda/(\lambda + \delta)$. For PLSR with a factors, an important property is that the shrinkage function can be written $f_j = \Phi_0(\lambda_j)$, where $\Phi_0(x)$ is a certain polynomial of degree a.

3.2 The shrinkage properties of PLSR

It has been shown (Butler & Denham, 2000 or Lingjærde & Christophersen, 2000) that PLSR always shrinks the smallest eigencomponent, ($f_p < 1$) while, for other components, shrinking and inflation alternates in an intricate way. If the sequence of shrinkage factors is arranged in order of increasing eigenvalues, elements less than one and greater than one will form a sequence that consists of $a+1$ "runs" (by a run we mean an unbroken sequence of numbers on the same side of 1). Thus, for example, the largest eigencomponent will be shrunk, $f_1 < 1$, if a is even, and expanded, $f_1 > 1$, if a is odd. In proving this, equations (2) and (3) play an important role. We now generalize the proofs to arbitrary regular semi-linear methods.

3.3 A generalization

In connection with Proposition 1, we note that the vector $Xb_a = P_a y$, being in $\mathcal{S}_H(a)$, can be written as a sum $\Sigma_{m=1}^{a} \phi_m H^m y$, for some coeficients ϕ_m, $m = 1, \ldots, a$. Since the projection is orthogonal, the coefficients ϕ_m will be the numbers that

minimize

$$|y - \Sigma_{m=1}^{a} \phi_m H^m y|^2 \ . \tag{4}$$

Since $H = XMX^T$ and $M = \mu(X^TX)$, we get $H^m = UD^mU^T$, (where U comes from $X = USV^T$), with $D = \text{diag}(d_1 \ldots, d_p)$, each $d_j = \lambda_j \mu(\lambda_j)$. Recalling $y = Uw$ we see that the quantity to be minimized is

$$|w - \Sigma_{m=1}^{a} \phi_m D^m w|^2 \ = \ \Sigma_{j=1}^{p} (w_j - \Sigma_{m=1}^{a} \phi_j {d_j}^m w_j)^2 = \Sigma_{j=1}^{p} (1 - \Phi(d_j))^2 {w_j}^2 \tag{5}$$

where $\Phi(x)$ is an intercept-free polynomial of degree a. We denote by $\Phi_0(x)$ the polynomial that minimizes (5). Its coefficients are denoted ϕ_m, that is, $\Phi_0(x) = \Sigma_{m=1}^{a} \phi_m x^m$.

Proposition 2 *The shrinkage factors for the iterated method $\mathcal{B}_{it}(X, y; a)$ are the numbers $\Phi_0(d_j)$.*

Proof: We need to find the expression for the regressor in canonical form, that is, find g_a such that $b_a = Vg_a$. It follows from (4) that $P_a y = \Phi_0(H)y$, which is equal to $U\Phi_0(D)U^TUw$, or shorter $U\Phi_0(D)w$, so from $Xb_a = P_a y$ we get $U\Lambda^{1/2}V^TVg_A = U\Phi_0(D)w$. Since $U^TU = I_p$, we get $\Lambda^{1/2}g_A = \Phi_0(D)w$, and, since diagonal matrices commute, $g_a = \Phi_0(D)\Lambda^{-1/2}w = \Phi_0(D)g_{OLS}$. We see thus that the numbers $\Phi_0(d_j)$ will be the shrinkage factors as asserted in the proposition. ■

Graphically, to find the numbers $\Phi_0(d_j)$ we want to find a polynomial curve of degree a that passes through the origin (i.e., no constant term), and that comes as close as possible to the points with coordinates $(d_j \ , \ 1)$. Figure 1 illustrates the situation.

It turns out that the following holds:

Proposition 3 *The iterated method $B_{it}(X, y; a)$ will shrink some of the singular components of X and expand others. The smallest eigencomponent will be shrunk, and there will be a total of $a + 1$ runs of shrinkages and inflations.*

Proof: See Appendix C.

Again PLSR is an important application. In PLSR we have $\mu(x) = 1$, so $d_j = \lambda_j$. We now apply proposition 3 to other regression methods.

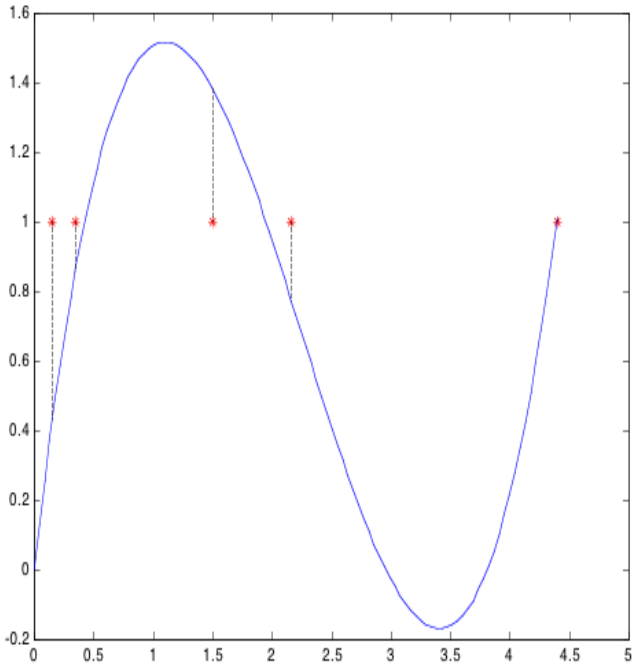

Figure 1: Illustration of an intercept-free cubic polynomial $y = \Phi(x)$ minimzing (5) given $p = 5$ eigenvalues λ_j. The quantity to be minimized is the (weighted) sum of squares of the deviation from $y = 1$ at $x = \lambda_j$, indicated by vertical dashed lines. Here, the two smallest eigencomponents are shrunk, number three is inflated and number four shrunk. The largest eigencomponent is fit almost exactly.

3.3.1 Application to ridge-type regressors

Proposition 3 is applicable to LSRR. In this case, (using one of several possible parametrizations) one has $M = ((1-\alpha)X^TX + \alpha I_p)^{-1}$. This corresponds to $\mu(x) = 1/((1-\alpha)x + \alpha)$ which gives

$$d_j = \frac{\lambda_j}{(1-\alpha)\lambda_j + \alpha}.$$

We see that the special case $\alpha = 1$ gives $d_j = \lambda_j$, which is PLSR. The other special case, $\alpha = 0$, gives $d_j = 1$ for all j. At this extreme, all the points to be approximated collapse into the point (1,1). It is possible to catch this one point exactly, with a straight line (a first-degree polynomial) through the origin. This illustrates that with OLSR one factor is enough to catch all there is to capture.

For a third special case we may let α grow larger than 1. As α approaches $\lambda_1/(\lambda_1 - 1)$ we see that the denominator in the expression for d_1 tends to zero, implying that the point $(d_1, 1)$ is very far to the right. Then, if one approximates all the points with a straight line, the great leverage effect of the remote point will force a solution where $\phi(d_1) \approx 1$ while $\phi(d_j) \approx 0$ for $j \geq 2$. This is equivalent to first-factor PCR.

3.3.2 Application to continuum power regression

Continuum power regression, CPR, is a variation of PLSR where the matrix $X = U\Lambda^{1/2}V^T$ is replaced by X^μ (defined as $U\Lambda^{\mu/2}V^T$) and PLSR is performed using this modified matrix instead of X. The metaparameter μ is zero or positive. The regressor then obtained, $\widetilde{b} = \mathcal{B}_{PLS}(X^\mu, y)$ is back-transformed to yield the final result $b_{CPR} = V\Lambda^{(\mu-1)/2}V^T\widetilde{b}$. The definition of $\mathcal{B}_{PLS}$ and some linear algebra yield $b_{CPR} \propto V\Lambda^{\mu-1}V^TX^Ty$, that is, CPR is a semi-linear method and the expression for the horizontal coordinates d_j in equation (5) now is

$$d_j = {\lambda_j}^\mu.$$

We see that $\mu = 0$ or 1 yields $d_j = 1$ or λ_j, so that OLSR and PLSR are special cases of CPR as they are of LSRR. As $\mu \to \infty$, the point $(d_1, 1)$ will be far to the right of all other points $(d_j, 1)$, so that a leverage effect will lead to first-factor PCR, similarly to LSRR.

4 Regular semi-linear methods fit closer than PCR

When proving that "PLS fits closer than PCR", de Jong (1993) did the following, in summary: Note that the shrinkage factors corresponding to PCR with a factors are

$$f_j = \begin{cases} 1 & \text{for } j = 1, \ldots, a \\ 0 & \text{for } j = a+1, \ldots, p \end{cases}$$

There exists an intercept-free a-degree polynomial, denote it $\Phi_*(x)$, such that $\Phi_*(\lambda_j) = 1$ for $j = 1, \ldots, a$. It can be shown that Φ_* satisfies the inequalities $0 \leq \Phi_*(x) \leq 1$ for all $0 \leq x \leq \lambda_{a+1}$. Therefore, if we were to construct a candidate regressor with

the numbers $\Phi_*(\lambda_j)$ as shrinkage factors for $j = 1, \ldots, p$, this candidate would obtain better fit than PCR with a factors. But PLSR with a factors corresponds to using the polynomial that has best fit among all intercept-free a-degree polynomials, so PLSR cannot give worse fit than our candidate regressor. Therefore PLS fits closer than PCR.

Essentially the same argument is valid for any regular semi-linear method. We have the following propostition:

Proposition 4 *If the function $x\mu(x)$ is increasing, the regular semi-linear method in Definition 2 fits any data set better than PCR with the same number of factors.*

Proof: The proof is completely analogous to that by de Jong (1993). The only adjustment necessary is that the horizontal coordinates for the points to be approximated (the numbers d_j in (5)) are not λ_j but $\lambda_j\,\mu(\lambda_j)$. It is therefore necessary that $d_1 \geq d_2 \geq \ldots \geq d_p$. It is to this end that we need the requirement that the function $x\mu(x)$ must be increasing. Some further details are given in Appendix D. ■

5 Discussion and conclusions

It is known that PLSR has peculiar shrinkage properties. We have now demonstrated that at least two other methods, CPR and LSRR, share these troublesome properties. In one way, this is not surprising. Both methods include PLSR as a special case, and the shrinkage factors vary continuously with the metaparameter, so the same shrinkage pattern should reasonably prevail in a neighborhood of PLSR, too. As we have now seen, the pattern is present for any value of the metaparameter.

However, both CPR and LSRR offer an additional degree of flexibility (which PLSR does not) in that the metaparameter can be varied between the steps in the iteration process. As the above shows, we *must* use this flexibility, if we are to avoid the unwanted shrinkage/inflation effects. Thus, in the model selection process, a modeller has to set values for, in effect a continuous parameters, α_m, $m = 1, \ldots, a$. where α_m = the metaparameter at the m:th iteration. Obviously, this violates a principle of parsimony, and the risk for overfitting is clear. In addition,

the metaparameter is often defined to be the argument that maximizes a function, and the only way to maximize this function is to evaluate it for a large number of values. The computational burden will be considerable.

REFERENCES

Björkström, A. & Sundberg, R. (1999) A generalized view on Continuum Regression. *Scand. J. Statist.* **26**, 17-30.

Butler, N. A. & Denham, M.C. (2000) The peculiar shrinkage properties of partial least squares regression.*J. Roy. Statist. Soc.* Ser. B **62** 595-593.

de Jong, S., Wise, B.M. and Ricker, L.N. (2001) Canonical partial least squares and continuum power regression. *J. Chemometrics*, **15**, 85-100.

de Jong, S. (1993) PLS fits closer than PCR *J. Chemometrics*, **7**, 551-557.

Helland, I. S. (1988) On the structure of Partial Least Squares Regression. *Comm. Stat Series B*, **17**, 581-607.

Lingjærde, O. & Christophersen, N. (2000) Shrinkage structure of partial least squares. *J. Roy. Statist. Soc.* , **27**, 459-473.

Stone, M. & Brooks, R. J. (1990) Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. (With discussion) *J. R. Statist. Soc.* B, **52**, 237-269; Corrigendum (1992) **54**, 906-907.

Sundberg, R. (1993) Continuum regression and ridge regression. *J. R. Statist. Soc.* B, **55**, 653-659.

Wise, B.M. & Ricker, L.N. (1993) Identification of finite impulse response models with continuum regression. *J. Chemometrics*, **7**, 1-14.

Appendices

A Iteration of methods

Let $\mathcal{B}(X, y)$ be any method. Let $A \geq 2$ be an integer. By $\mathcal{B}_{\text{it}}(X, y; A)$ we mean the following method:

1. $b_1 = \mathcal{B}(X, y)$
2. $r_1 = y - Xb_1$.
3. $t_1 = Xb_1$
4. $a = 2$
5. ${b_a}' = \mathcal{B}(X, r_{a-1})$
6. $t_a = X{b_a}'$.
7. Obtain b_a from $Xb_a = P_a y$, where P_a denotes projection on span $[t_1 \dots t_a]$. Apply a minimum-length condition on b_a if the solution is not unique.
8. If $a = A$, return $\mathcal{B}_{\text{it}}(X, y; A) = b_a$ and stop.
9. $r_a = y - Xb_a$
10. Set $a = a + 1$ and return to 5.

B Proof of proposition 1

It follows from step 7 in the iteration (Appendix A) that all we need to prove is the following: For any integer $a \geq 1$,

$$\mathcal{S}_t(a) = \mathcal{S}_H(a) \tag{6}$$

that is, the two subspaces $\mathcal{S}_t(a) = \text{span}[t_1 \dots t_a]$ and $\mathcal{S}_H(a)$ coincide. It is clear that (6) is true for $a = 1$, since $t_1 = Xb_1 = X(\kappa M X^T y) \propto Hy$. We proceed by induction and assume that (6) is true for a certain a. Then, because of Steps 5 and 6 in the iteration, $t_{a+1} = X\mathcal{B}(X, r_a)$, and because of the definition of a semi-linear

method (equation 1), $t_{a+1} \propto XMX^T r_a$, so that because of Step 9 in the iteration, $t_{a+1} \propto Hr_a = H(y - Xb_a) = Hy - HXb_a$.
Now $Xb_a \in \mathcal{S}_t(a)$ by Step 7, so $Xb_a \in \mathcal{S}_H(a)$ by our induction hypothesis. Therefore Xb_a can be written as a sum $\Sigma_{j=1}^{a} \gamma_j H^j y$ for some coefficients γ_j, and

$$t_{a+1} = Hy - \Sigma_{j=1}^{a} \gamma_j H^{j+1} y = z - \gamma_a H^{a+1} y,$$

where $z \in \mathcal{S}_H(a)$. We can thus write

$$\mathcal{S}_t(a+1) = \text{span } [A \,|\, t_{a+1}] = \text{span } [A \,|\, z - \gamma_a H^{a+1} y]$$

and

$$\mathcal{S}_H(a+1) = \text{span } [A \,|\, H^{a+1} y],$$

where A denotes any matrix whose column vectors span $\mathcal{S}_H(a)$ (or $\mathcal{S}_t(a)$). It follows that $\mathcal{S}_H(a+1) = \mathcal{S}_t(a+1)$, and the induction is complete. ■

C Proof of proposition 3

Denote with r the number of runs occurring for $\Phi_0(x)$, the best-fitting intercept-free polynomial of degree a. Assume $r \leq a$. The polynomial $\Phi_0(x)$ must then have $r-1$ "ones" (*i.e* an x such that $\Phi_0(x) = 1$) in the interval from the smallest to the largest of the numbers d_j, $j = 1, \ldots, p$. Denote these x-values δ_j, $j = 1, \ldots, r-1$. Consider the polynomial

$$\pi(d) = d(d - \delta_1)(d - \delta_2) \ldots (d - \delta_{r-1}).$$

One realizes that one of the two intercept-free a-degree polynomials $\phi(d) - \pi(d)$ and $\phi(d) + \pi(d)$ will be closer to 1 at *all* the points d_j, $j = 1, \ldots, p$ than $\phi(d)$ is. This is a contradiction, since $\phi(d)$ was defined as the best-fitting polynomial. We can thus rule out the possibility $r \leq a$ *i.e*, we have $r \geq a + 1$. On the other hand, it must also hold that $r \leq a + 1$: Since an intercept-free polynomial of degree a can have at most a "ones", it is impossible to obtain more than $a + 1$ runs. Since both $r \leq a + 1$ and $r \geq a + 1$ hold, we conclude $r = a + 1$. ■

D Comparison to PCR

For a regular semi-linear method with a factors, the error $|y - \hat{y}|^2$ is

$$\epsilon^2 \;=\; \Sigma_{j=1}^{p}\,(1 - \Phi_0(d_j))^2\,{w_j}^2. \tag{7}$$

where Φ_0 is the relevant polynomial for the method in question, as defined in equation (5).

Now let Φ_{**} denote the intercept-free a-degree polynomial that has

$$\Phi_{**}(x) = 1 \;\text{ for } x \;= d_1, \ldots, d_a. \tag{8}$$

This polynomial will satisfy $0 \;\leq\; \Phi_{**}(x) \;\leq\; 1$ for all x in the interval $[0\,,\,d_a]$ (Being a polynomial of degree a, Φ_{**} can have at most $a-1$ extreme points, and all these have to be in the interval $[d_a\,,\,d_1]$ in order for (8) to be possible. It follows that $0 = \Phi(0) \;\leq \Phi_{**}(x) \;\leq \Phi(d_a) \;= 1$ when $x \in [0\,,\,d_a]$). Therefore,

$$\Sigma_{j=1}^{p}\,(1 - \Phi_{**}(d_j))^2\,{w_j}^2 \;=\; \Sigma_{j=a+1}^{p}\,(1 - \Phi_{**}(d_j))^2\,{w_j}^2 \;\leq \Sigma_{j=a+1}^{p}\;{w_j}^2 \;=\; {\epsilon_{\text{PCR}}}^2.$$

However, the polynomial Φ_0 in (7) is by definition that which minimzes the sum $\Sigma_{j=1}^{p}\,(1 - \Phi(d_j))^2\,{w_j}^2$, so it follows that $\epsilon^2 \leq {\epsilon_{\text{PCR}}}^2$.