Norwegian University of Life Sciences
Faculty of Veterinary Medicine and Biosciences
Department of Animal and Aquacultural
Sciences

Master Thesis 2014
30 credits

Use of eigenvectors in modelling longitudinal data

Matthew D. Price

Use of eigenvectors in modelling longitudinal data

Matthew David Price
Registration #983037

THESIS ANIMAL BREEDING AND GENETICS (M30-IHA)
August 2014

Norges miljø- og biovitenskapelige universitet

Education and Culture

Erasmus Mundus

Department of Animal and Aquacultural Sciences

Supervisor:
Prof. T. H. E. Meuwissen*

Examiners:
Prof. T. H. E. Meuwissen*
Dr E. Fimland†

* Department of Animal and Aquacultural Sciences, NMBU, Ås, Norway
† Retired

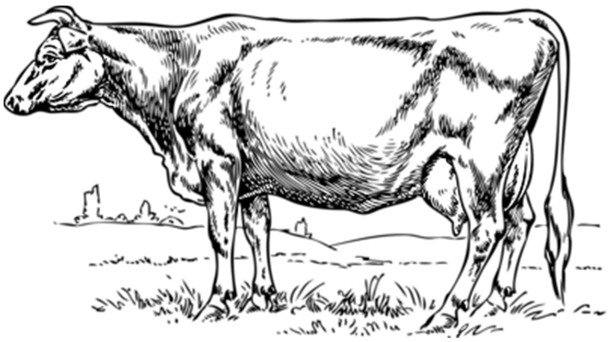
Preface

This Masters thesis was undertaken as part of the second year of study for the European Master in Animal Breeding and Genetics (EMABG) program at the Norwegian University of Life Sciences in Ås, Norway. It was the second of two theses to be produced under the program, the first of which was undertaken at Wageningen University, The Netherlands. Both theses were complemented by a series of course work, of which the relative weighting with thesis work was half-and-half.

This particular thesis investigated the use of eigenvectors in the modelling of longitudinal data, such as that of seasonal lactation yield. As part of a regression model, eigenvectors derived from the data may constitute a set of basis functions which would act as regressors in the model. The basis set of such a regression model has utilised various functions in other studies, such as polynomials, exponentials, trigonometric functions, etc., and so this study sought to investigate and compare the goodness-of-fit of an eigenvector-based model with that of models based upon other common types of basis sets. It has been a rewarding endeavour for myself, utilising multiple areas of interest such as quantitative genetics, programming, mathematics and statistics, as well as employing a measure of problem solving and method development which I find particular satisfaction with.

I wish to thank the members of both the Norwegian University of Life Sciences (Ås, Norway) and Wageningen University (Wageningen, NL) who have not only been instrumental in this project, but in this two-year Masters program as a whole. Particular thanks goes to Professor Theo Meuwissen for suggesting this thesis topic and assuming the role of supervisor for it, and to the program administrators Dieuwertje Lont and Stine Telneset for their support and guidance throughout this course of study. I wish also to thank the companies of CRV BV (Arnhem, NL) and DairyNZ (Hamilton, NZ) whose financial support via scholarship grants has made this work, and indeed my entire course of study, possible. Special thanks also goes to Dr Jim Gibbs of Lincoln University, NZ, who was so instrumental in encouraging me to join, and recommending me for, this EMABG program.

Finally, I wish to express my deepest gratitude to my parents who have encouraged and supported me in both word and prayer during this time. Moreover, I give thanks to God for his sovereign provision, guidance and love which has, amongst many things, blessed me with both the faculties within myself and the opportunities beyond myself to undertake and complete this entire course of study.



Summary

The modelling of longitudinal data by a regression model generally utilises a set of basis functions of a certain type (polynomial, exponential, etc.) to be used as regressors in the model. Along with the regression coefficients, these basis functions define the interpolating functions for fitting the data.

These continuous functions can further be used to extrapolate values for points in time beyond those from the initial dataset, although the accuracy by which they may do so depends upon the nature of the dataset and the type and order of the interpolating function (where the order is the number of basis functions used). The accuracy of the extrapolation of data, and indeed the interpolation (fitting) of data, will depend upon how well the interpolating function takes into the account the covariance of the data. If the type of function does not represent the nature of the data well, the obtaining of accuracies will be problematic, particularly for the boundary regions of the data.

The covariance of the data can be decomposed into a set of eigenvectors, which re-parameterises the data into the directions of greatest to lowest variance, as is used in principal component analysis (PCA). By using the eigenvectors of the directions of greatest data variance as a set of basis vectors in a regression model, they have the potential to better fit the data than interpolating functions of the same order. It was therefore the objective of this study to investigate the use of eigenvectors in a regression for modelling longitudinal data, and compare it to existing models.

Additionally, the piecewise interpolation of the eigenvectors themselves was investigated to see if a regression using the resulting eigenvector-based function set was effective in extrapolating data beyond the time-points of the initial dataset.

The study utilised a simulated dataset for which covariance and noise variance could be known and controlled. It was found that an eigenvector-based model can indeed provide better fitting of the data, as well as extrapolation of data, than some existing functions (the Legendre polynomials and a modified version of the inverse polynomials in particular). However, results were not consistent, and the noise of the dataset needs to be better taken into account in order to better ascertain what advantage an eigenvector-based model truly has over existing models.

Key words: Eigenvectors; interpolation; test day models; lactation curves

Table of Contents

1. Introduction

In this study, the accurate modelling of longitudinal data is motivated by the need for ascertaining parameter estimates for various traits which have been repeatedly measured over time for a number of individuals. These estimates may then be employed in a breeding model which takes many other traits into consideration for the purposes of determining the genetic worth of individuals, or simply in predicting those trait values for future or missing points in time. The motivating example for this study was to model milk yield in dairy cattle over time, ascertaining milk yield parameters per animal per lactation period, based upon yields from several test-days per period. Although this specific scenario has motivated the study, the methods developed herein should be more broadly applicable.

There are a number of ways in which the recorded values of the test-days (heretofore referred to as time-points) could be considered in the context of a model. The values of the time-points could be treated as those of a single trait with repeated measures over time, or as a distinct 'trait' per time-point, with each 'trait' having some measure of covariance with other 'traits'. A repeatability model was a common method in estimating the genetic parameters of milk yield traits up until 1999 (Interbull, 2000), although other models have been subsequently developed and implemented in breeding programs since then. One such model is the fixed regression model, and it is the fundamental model upon which this study is based.

It was in Germany where the fixed regression model first began to be implemented for genetic selection of milk traits on a national scale, from 1995 and for several years thereafter. Rather than use a single 305-day yield value per cow per season, the model used individual test day records, whilst also taking herd and day-of-test effects into account. Consequently, the model was able to account for the general shape of the seasonal lactation curve, per grouping of animals, as well as the curve shape per individual animal. The shape of the lactation curve is particularly important, as particular aspects of it – such as the time to peak yield, rate of increase to peak yield, and rate of decline from peak yield – are all important properties of the nature of the lactation period for which desirable selection is sought.

A general fixed regression model is described by the following equation:

$$y_{hij} = htd_h + \sum_{k=0}^{n} \phi_{jk}\beta_k + u_i + pe_i + e_{hij} \tag{1}$$

where, for test day *j* upon herd-test-day (HTD) subclass *h* for the test day record of cow *i*:

y_{hij}	=	test day record of cow *i* in HTD subclass *h* on day *j*
htd_h	=	HTD subclass *h*
ϕ_{jk}	=	*j*th element of the *k*th vector of a set of *n*+1 pre-defined basis functions evaluated across all *j* test days
β_k	=	fixed regression coefficient for aforementioned function
u_i	=	vector of animal additive genetic effects for cow *i*
pe_i	=	vector of permanent environmental effects for cow *i*
e_{hij}	=	random residual

The sum of the basis functions weighted by their corresponding regression coefficient, for each animal, produces an interpolating function of the data which should approximate the real curve underlying the data. The interpolating function will be of a predefined maximum order n, and the set of n basis functions will have varying orders of k = $\{1, \ldots, n\}$ for each basis function ϕ_k.

This study would utilise simulated data. As such, there was no need to take some of the fixed effects of the general fixed regression model into account, if such effects were not simulated for in the first place. Thus a simplified fixed regression model would be used in this study, the components of which were just the sum of the products of the regression coefficients and basis functions, and the error term.

It should be noted that the fixed regression model is not the only longitudinal model to employ a set of basis functions. Another such model is the more modern random regression model (Schaffer and Dekkers, 1994), which is quite similar to the fixed regression model, except that the regression coefficients are considered as random, not fixed, effects. For the purposes of this study the simpler fixed regression model was used, although the merits of this choice invite further consideration, as will be given in the discussion section of this study.

A key concept in the modelling of the longitudinal data is that of the relationship between the values of differing time-points (or 'traits' as they shall be considered for the purposes of this study). The assumption is that the values of these traits are not independent of each other, but that there is a real relationship between them which can be described in terms of a function in time or by the covariance structure of the data. It is important that the model used takes these relationships into account, as the prediction of unknown trait values depends upon this trait interdependence, and furthermore the estimation of trait effects can be made with greater accuracy when information on related traits can be incorporated into the estimation method.

It is the choice of pre-defined basis functions within this regression model which this study is primarily concerned with. The choice of the type of basis function set will determine how well the model will fit the data overall, as well as how well different aspects of the underlying curve are modelled. In particular, the boundary areas of the data are of special interest, as model bias can become particularly pronounced in these areas due to local random variation (Macciotta *et al.*, 2005).

Three types of basis sets were compared in this study: the Legendre polynomial basis functions, a modified version of the inverse polynomial functions, and a new type of basis set, unique to this study; an eigenvector set.

The primary objective of this study was thus to determine whether the fixed regression model using an eigenvector basis set was not only feasible, but whether it had an advantage over currently used basis sets.

To this end, both the estimation of data for the finite set of given data-points themselves, as well as for the continuous longitudinal range between such data-points (as could be assessed with reference to additional generated data from the same simulated data function), would be undertaken, the results of which would be analysed via several methods to determine the merits of the various basis sets.

2. Materials and Methods

Because the nature of this study was one which involved not only the implementation of a set of methods, but one which also involved the development of methods for implementation, this process of method development itself falls within the category of materials and methods. As such, some discussion will first be given here to the process in which the particular basis sets investigated in this study were chosen.

Once the choice of basis sets has been established, the process by which they may be implemented and evaluated will then be described, including the data used, the models implemented, and the analysis done.

2.1 Choice of basis function sets

2.1.1 Polynomials

The most common type of basis function set is a polynomial one, and of these there are several subtypes. The general form of each polynomial basis function is given by:

$$\phi_k(t) = \sum_{m=0}^{k} \alpha_{m+1} t^m = \alpha_1 + \alpha_2 t + \alpha_3 t^2 + \cdots + \alpha_{k+1} t^k \qquad (2)$$

where the coefficients α_m, $m = \{0, \ldots, k\}$, of the powers of time variable t are specified for each basis function of a particular basis function set.

The simplest set of polynomial basis functions is the monomial one, where each $\phi_k(t) = t^k$. Another is the Newton set of basis polynomials, where each Newton polynomial is of the form:

$$\pi_k(t) = \begin{cases} 1 & if\ k = 1 \\ \prod_{d=1}^{k-1} (t - t_d) & if\ k > 1 \end{cases} \qquad (3)$$

Other sets of polynomial basis functions include the Lagrange set, the Legendre set, and the Chebyshev set. Although the α-coefficients of the basis polynomials will differ per set, and the β-coefficients of a subsequent regression will also differ per set, the resultant interpolating polynomials of the regression will be the same per animal. The main difference in the choice of function set has to do with the computational time, components of which are the β-coefficient calculation part, and the interpolating polynomial evaluation part. A set such as the monomial set has a straight forward evaluation of interpolating polynomials, but the computation of β-coefficients requires the full solution of a linear system of equations (which moreover may be ill-conditioned), whilst the opposite is true for other basis sets, such as Lagrange.

As the computational time is of little consequence in this study, any one of the available polynomial basis function sets would be adequate in the comparison of the polynomial-type basis set with the other two in this study. The Legendre polynomial basis set was chosen, as it is the most commonly used basis function set in data interpolation; it isn't prone to ill-conditioned linear systems; it makes no assumptions about the shape of the curve for which it is approximating; and it is relatively straight-forward to apply.

Despite the popularity of polynomial data interpolation, there is a particular drawback to it known as the Runge phenomenon (Runge, 1901). This phenomenon describes how, for higher order polynomials interpolating equispaced points, the edges of the range of the data suffer from high oscillation of the predicted values between the interpolation points. As the accurate modelling of these boundary areas is a particular concern of this study, a second type of basis function set was sought for comparison with the eigenvector basis set, in which the boundary areas were better-behaved.

2.1.2 Inverse polynomials

A promising candidate for such a second type of basis functions were the inverse polynomials, as described by Nelder, 1966. The general form of an inverse polynomial function is given as:

$$\phi_k(t) = \frac{t}{\sum_{m=0}^{k} \alpha_{m+1} t^m} = \frac{t}{\alpha_1 + \alpha_2 t + \cdots + \alpha_{k+1} t^k} \qquad (4)$$

The particular advantage of these functions is the existence of an x-asymptote which flattens the function in the boundary areas of an interval centred about function's origin. The advantage brought by this function was not without its own particular challenges though.

Unlike the standard polynomials, where the addition of two polynomials of order $\leqslant n$ in polynomial vector space $\mathbb{P}_n$ will result in another polynomial in that same $\mathbb{P}_n$ space, the same property does not hold for the inverse polynomials. That is, the sum of two inverse polynomials is not another inverse polynomial. Furthermore, as a polynomial can effectively be shifted horizontally by the addition of a lower order polynomial, a basis set of standard polynomials can thus be considered originless; they have no fixed origin. Inverse polynomials, on the other hand, each have a fixed origin, and so aspects of the nature of the curve underlying the data must be assumed beforehand in order to calibrate the basis inverse polynomials accordingly. An additional complication lies in the existence of y-asymptotes when the denominator of equation 4 equals zero, given certain conditions of the α-coefficients.

A modified version of an inverse polynomial basis set, which sought to address the aforementioned issues, was thus developed and implemented in this study.

2.1.3 Eigenvectors

Finally, the new eigenvector type of basis set would supply a set of basis vectors directly to the regression model, without the need for a function evaluation for time-points j per animal i. Because the set of eigenvectors would be determined from the eigendecomposition of the covariance matrix of the data, the eigenvectors would still take into account the relationship between data from different time-points, without the need for a set of continuous functions which would otherwise describe the relationship between time-point data via function parameters.

The eigendecomposition of a symmetrical d×d matrix yields a set of d eigenvector-eigenvalue pairs, which essentially represent a re-parameterisation of the vector space $\mathbf{F}^d$ of the matrix in such a way that the new eigenvectors (which are, incidentally, orthogonal to each other) describe the directions of greatest variance in the data (for the eigenvector with maximum eigenvalue); of second-greatest variance in the data (for the eigenvector with second-highest eigenvalue); and so forth. It is thus possible to discard the eigenvectors corresponding to directions of minimal or no data variance, and in doing so, the data has been re-parameterised to a lower dimensional space. This is essentially the approach used in principal component analysis (PCA), which seeks a lower dimensional representation of a dataset. In the case of the modelling of longitudinal milk data, for example, the data associated with d test-day yields per animal can instead be represented by a linear combination of $\leqslant d$ eigenvectors.

Although the potential loss in dimensionality is a definite computational advantage (as the effective number of traits being modelled can be reduced), it is the nature of the way in which the eigenvectors explain the variance of the data which is of real significance in their potential implementation as a basis set in a regression model. In a similar fashion to the way in which increasing the order of a basis function set should increase the predictive power of the model, Increasing the number of eigenvectors used, from those of highest eigenvalue to lowest, should also increase the predictive power of the model. Additionally, the eigenvector approach would seem to have a clear advantage over the basis function sets, as while the basis function sets would start with very simple basis functions of low order (which presumably would not initially accurately represent the data), the eigenvector basis set would begin with the vectors explaining maximal variance first, and so it would seem likely that the eigenvector basis set would fare much better than the basis function sets for low orders of n.

The eigenvector-based model would not suffer from the Runge phenomenon (Runge, 1901), simply because the eigenvectors can only be evaluated for the time-points of the dataset, and not for the intervals in-between. Although this limitation to only estimate values for the specified time-points is a considerable drawback to the model, the possibility exists to perform a piecewise interpolation of the eigenvectors themselves, and thus convert them into piecewise functions for all t within the longitudinal range of the data. In this study, both a linear piecewise interpolation and a natural cubic spline interpolation would be used upon the eigenvectors.

2.2 Data

The simulated data used in this study was generated from a function (equation 5) in which the response variable y per record i was a linear combination of two normal distribution density functions, evaluated at time t, plus an random residual.

$$y_i(t) = x_{1i}\varphi_1(t) + x_{2i}\varphi_2(t) + e_i \tag{5}$$

In this model, normal distribution density functions $\varphi_1(t)$ and $\varphi_2(t)$ were defined by normal distributions $\mathrm{N}(\mu=1, \sigma=1)$ and $\mathrm{N}(\mu=4, \sigma=2)$, respectively. The coefficients x_{1i} and x_{2i} of these normal functions were randomly sampled, per record, from the multivariate normal distribution, which was defined by $MVN\left(\mu = 0, \sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$. The random residual was sampled from the normal distribution $\mathrm{N}(\mu=0, \sigma= \sigma_s)$ per record, with a certain value of σ_s per generated dataset, from σ_s ={0, 0.2, 0.5, 1}. As such, four different datasets would be generated, each one with differing noise levels, although the primary dataset used would be the one for which the noise was sampled form the normal distribution $\mathrm{N}(\mu=0, \sigma=1)$.

The simulated dataset was comprised of N = 1,000 records y_i, i = {1, …, 1000}; evaluated at d = 6 different time points t_j , j = {1, …, 6}, of **t** = [0 1 2 3 4 5].

The resulting dataset would have high variation at points $t_2 = 1$ and $t_5 = 4$, and lower variation at the end points $t_1 = 0$ and $t_6 = 5$. Coupled with the covariation present between the two underlying density functions, this would hopefully present a realistic challenge for each model in capturing the total variation present across the entire interval.

2.3 Fixed regression model

As the simulated data could be generated devoid of any unnecessary effects, the fixed regression model used in this study could likewise be simplified to the form:

$$y_{ij} = \sum_{k=0}^{n} \phi_{jk}\beta_{ik} + e_{ij} \tag{6}$$

where for each data value y_{ij} for individual i at time-point j, ϕ_{jk} was the jth element of the kth basis vector (either an eigenvector or a basis function evaluated at time-points **t**) of a set of n vectors, β_{ik} was the regression coefficient of the kth basis vector for individual i, and e_{ij} was a random residual per individual i, per time-point j.

The vectors of regression coefficients were calculated per individual, so that the d-dimensional data-space of the d traits (observations at different time-points) per individual would be effectively re-parameterised to an n-dimensional coefficient-space. The advantage of this would not only be that each individual had a reduced number of ‘traits’ for the purposes of breeding value estimation, but that the variance of these β-coefficients might also be per calculated per time-point.

For a given basis set type of a given order n, a basis matrix Φ_n would be derived from the data, whether it was from the evaluation of each basis function for the time vector **t** (explained in detail in §2.4), or directly from the eigendecomposition of the covariance matrix of the data (explained in detail in §2.4.3). Each row of the basis matrix would correspond to each basis function/eigenvector, with each column corresponding to each time-point of the data. The dimension of the matrix would thus be n×d, where d = 6 in this study.

$$\Phi_n = \begin{bmatrix} \leftarrow & \phi_1(\mathbf{t}) & \rightarrow \\ \leftarrow & \phi_2(\mathbf{t}) & \rightarrow \\ & \vdots & \\ \leftarrow & \phi_n(\mathbf{t}) & \rightarrow \end{bmatrix} \tag{7}$$

In matrix notation, the regression model per individual i was thus:

$$y_i(\mathbf{t}) = \Phi_{\boldsymbol{n}}^{\mathbf{T}}\boldsymbol{\beta}_i + \boldsymbol{e}_i(\mathbf{t}) \tag{8}$$

Or, for the whole dataset:

$$\mathrm{Y} = \beta\Phi_n + \mathrm{E} \tag{9}$$

Where Y was the N×d = 1,000×6 matrix of the data, Φ_n was the basis matrix of order n, β was the N×n matrix of regression coefficients, and E was the N×d matrix of random residuals. The estimates for Y were calculated thusly:

$$\hat{\mathrm{Y}} = \beta\Phi_n \tag{10}$$

where
$$\beta = \mathrm{Y}\Phi_{\boldsymbol{n}}^{\mathbf{T}}(\Phi_n\Phi_{\boldsymbol{n}}^{\mathbf{T}})^{\mathbf{-1}} \tag{11}$$

2.4 Longitudinal models

There were three longitudinal models used to fit the data – the Legendre polynomial model, the modified inverse polynomial model, and the new eigenvector model. Each model had the capacity for a variable number of n model parameters. A model of n parameters was referred to as a model of order n, and would have a basis set consisting of n basis functions, each of which contained up to n parameters (with at least one containing exactly n parameters).

For a polynomial or inverse polynomial model of order n, each basis polynomial $\phi_k(t)$ $\{k$ = 1, …, $n\}$ had for each element $m\{m$= 1, …, $k\}$ the product of a polynomial coefficient s_{km} and a term in t with degree m-1+δ, where δ was a constant specific to the type of model (equation 12). By evaluating each basis polynomial for each time point t_j $\{j$ = 1, …, $d\}$, a set of corresponding basis vectors was created, which in turn would constitute a basis matrix Φn (equation 13).

$$\phi_k(t) = \sum_{m=1}^{k} s_{km}t^{m-1+\delta} = s_{k,1}t^{\delta} + s_{k,2}t^{1+\delta} + \cdots + s_{k,k}t^{k-1+\delta} \tag{12}$$

$$\Phi_n = \begin{bmatrix} \phi_1(\mathbf{t}) \\ \phi_2(\mathbf{t}) \\ \vdots \\ \phi_n(\mathbf{t}) \end{bmatrix} = \begin{bmatrix} s_{1,1} & 0 & \cdots & 0 \\ s_{2,1} & s_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & s_{n,n} \end{bmatrix} \begin{bmatrix} t_1^{\delta} & t_2^{\delta} & \cdots & t_d^{\delta} \\ t_1^{1+\delta} & t_2^{1+\delta} & \cdots & t_d^{1+\delta} \\ \vdots & \vdots & \ddots & \vdots \\ t_1^{n-1+\delta} & t_2^{n-1+\delta} & \cdots & t_d^{n-1+\delta} \end{bmatrix} \tag{13}$$

Under any particular model, each record of the data could be approximated as a linear combination of basis vectors. A linear regression model would then be used to determine the coefficients of each such combination of vectors.

2.4.1 Legendre polynomial model

The Legendre polynomials are a set of polynomials which satisfy Bonnet's recursion formula (equation 14), with the first two Legendre polynomials given as $\phi_1(x)$ = 1, $\phi_2(x)$ = x. An example of the Legendre polynomials up to order 7 are given in table 2.1.

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x) \tag{14}$$

Table 2.1: The first 7 Legendre polynomials

k	$\phi_k(x)$
1	1
2	x
3	$\frac{1}{2}(3x^2 - 1)$
4	$\frac{1}{2}(5x^3 - 3x)$
5	$\frac{1}{8}(35x^4 - 30x^2 + 3)$
6	$\frac{1}{8}(63x^5 - 70x^3 + 15x)$
7	$\frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5)$

The delta variable in the aforementioned general formula (equation 12) is equal to zero for this model type. It should be noted that, strictly speaking, the 'order' of a polynomial is the power of the term of highest degree, although in this study the 'order' of the basis function has been defined differently.

These polynomials are orthogonal (with respect to the L^2 inner product) on the interval $-1 \leqslant x \leqslant 1$, and it is within this interval that they formed a basis set for fitting the data. Because the data in this study did not span this particular [-1, 1] interval, the set of time points $\{t_1, \ldots, t_d\}$ was scaled to fit this interval, via its transformation from t-space into x-space thusly:

$$x(t) = \frac{2(t - t_d)}{t_d - t_1} + 1 \tag{15}$$

The estimation of each record, under the Legendre polynomial model of order n, is thus provided by the equation:

$$\hat{y}_i(t) = \sum_{k=1}^{n} \beta_{ik}\, \phi_k\big(x(t)\big) \tag{16}$$

Where, for record i, β_{ik} are the regression coefficients of each Legendre polynomial $\phi_k(x)$ evaluated for the scaled x-variable.

2.4.2 Modified inverse polynomial model

The standard inverse linear polynomial of order n in one factor x is given by the equation:

$$y(x) = \frac{x}{1 + x + x^2 + \cdots + x^n} \tag{17}$$

As is done in basis sets of standard polynomials, coefficients may be added to the terms of the inverse polynomials, specific for each basis inverse polynomial, or they could just be left without coefficients, as is the case with the monomial polynomials, as indeed was also the case in the original paper in which their application was developed (Nelder, 1966). Our basis inverse polynomials might then have a form such as this:

$$\phi_k^{-1}(x) = \frac{s_{k,1}}{x} + s_{k,2} + s_{k,3}x + \cdots + s_{k,k}x^{k-2} \tag{18}$$

However, such a form proves to be inadequate for the data in this study, and moreover, for the data we might generally try to work with, due to a few issues.

The first problem is the issue of vertical asymptotes for values of t<0. These asymptotes occur when the underlying polynomial (that is, the denominator in equation 17) has real roots. These asymptotes may further occur for values of t>0 given some negative s-coefficients. Having a basis inverse polynomial with a vertical asymptote is a bad idea, as the resultant basis matrix will most likely be ill-conditioned, due to extremes of magnitude that may occur within in matrix elements.

The second problem is due to the existence of a fixed origin for the inverse polynomials. Given a standard basis polynomial $p_k(x)$, the polynomial $p_k(x\text{-}x_0)$ for some constant x_0 is also a polynomial of order k, and thus can also be derived from the same (carefully chosen) set of basis polynomials. The same does not hold for inverse polynomials, and so they would be unsuitable for fitting anything other than a specific shape of curve. In fact, it was for curves of a specific shape – namely, those with a period of increase followed by a period of decrease – for which the original 1966 study was developed. That is not to say that the inverse polynomials could not be utilised in this study, rather, a careful modification of them would provide an adequate basis set, as will be demonstrated.

In order to avoid the first problem of vertical asymptotes, a set of basis inverse polynomials was required in which the underlying polynomials had no real roots. One such set of "root-less" polynomials are those for which the highest degree is even, and where the coefficients of the terms of even degree are not exceeded by those of their corresponding neighbour terms of odd degree. One way to achieve this was to simply define two coefficients; one for the terms of even degree, θ_1, and the other for the terms of odd degree, θ_2, where $|\theta_1| \geqslant |\theta_2|$. To further simplify, only coefficients of 1 or -1 were used. Because this underlying polynomial is essentially divided by x to achieve the form of the RHS of equation 18, the terms θ_1 and θ_2 will be coefficients for the terms of odd and even degree, respectively, in this modified inverse polynomial of odd order k with odd indices m = {1, 3, ..., k}:

$$\phi_k^{-1} = \frac{\theta_{k,1}}{x} + \sum_m \left(\theta_{k,2}x^{m-3} + \theta_{k,1}x^{m-2}\right) \qquad \theta_{k,1}, \theta_{k,2} \in \{-1, 1\} \tag{19}$$

For a particular basis inverse polynomial $\phi_k^{-1}(x)$ for order k, we now had four potential candidate functions, each corresponding to the choice of combined θ_k-coefficients. An example of them for an order of n = 3 can be seen in figure 2.1:

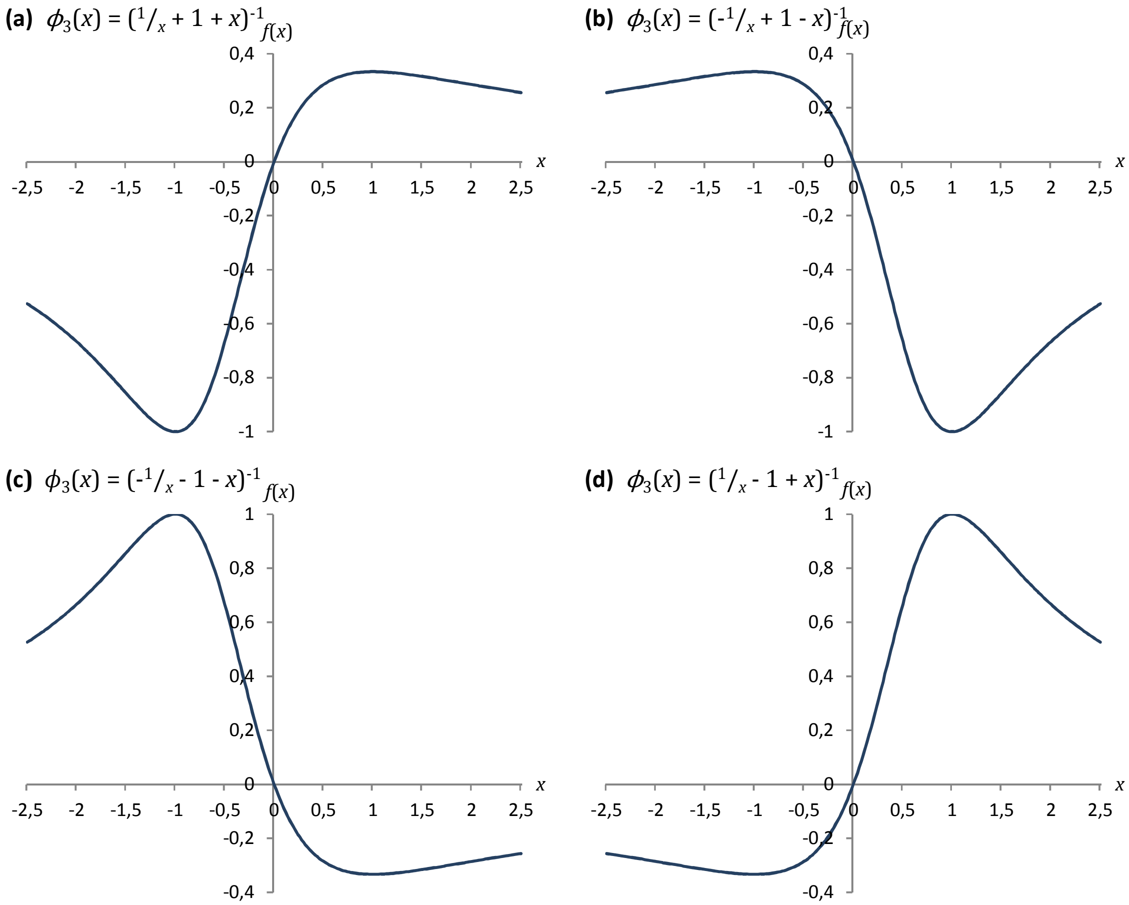


Figure 2.1 – Plots of inverses of four potential basis functions, $\phi_k(x)$ of order k=3 with varying θ_k coefficients, evaluated over the interval [-2.5, 2.5].

It is worth noting that whilst none of them are symmetrical, there is symmetry between all four of them. This symmetry can be summarised mathematically, where a function $\phi_k(x)$ using coefficients θ_1 and θ_2 is defined as $\phi_k(x, \theta_1, \theta_2)$:

$$\phi_k(x, 1, 1) = \phi_k(-x, -1, 1) = -\phi_k(-x, 1, -1) = -\phi_k(x, -1, -1) \qquad (20)$$

It is therefore only necessary to consider two of these functions as potential candidates for basis inverse polynomials, as the other two are merely scalar multiples of the first ones. The functions $\phi_k(x, 1, 1)$ and $\phi_k(x, 1, -1)$ were chosen in this case, thereby permanently setting the coefficients of the terms of odd degree to be 1.

Some problems would still remain at this stage. Namely:

(1) This would still leave us with twice as many basis functions for a given order of n than other models would otherwise provide, risking the problem of over-parameterisation for this model.
(2) The problem of requiring a defined origin per basis function still remained.
(3) In a related manner, the minima and maxima of the inverse polynomials are fixed within the interval [-1, 1] for any order k (with $|\theta_1|= |\theta_2| = 1$). Along with their fixed origin, this

creates a problem for accurately fitting the model to data for which (local) minima or maxima lie outside of this interval.

These issues can all be adequately accounted for however, by fixing the order of the potential set of basis inverse polynomials to an (odd) constant, and instead introducing a variable origin parameter.

Due to the limitations with the location(s) of minima and maxima, a pair of inverse polynomials employing the same linear transformation in their x-variable could be shifted/stretched in such a way that their minima and maxima points could be any two points on the x-axis. Naturally, choosing two points for which the data showed maximal variation and/or magnitude would make a good choice. By extension, another pair of inverse polynomials, this time with a different linear transformation applied to their x-variable, could attempt to capture the variation/magnitude of the next two points of greatest variation/magnitude in the data.

The question remained however, as to what choice of fixed order to use for the model. For the sake of generalisation, $\theta_{k,2}$ was chosen to be 1 for a series of $\phi_k(x)$ potential basis functions, with orders of k = {3, 5, 7, 9, 21}. These functions were determined (table 2.2), and plotted (figure 2.2).

Table 2.2 – Equations for potential inverse polynomial basis functions $\phi_k(x)$

k	$\phi_k^{-1}(x)$
3	$1/_x + 1 + x$
5	$1/_x + 1 + x + x^2 + x^3$
7	$1/_x + 1 + x + x^2 + x^3 + x^4 + x^5$
9	$1/_x + 1 + x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7$
$\vdots$	$\vdots$
21	$1/_x + 1 + x + \cdots + x^{18} + x^{19}$

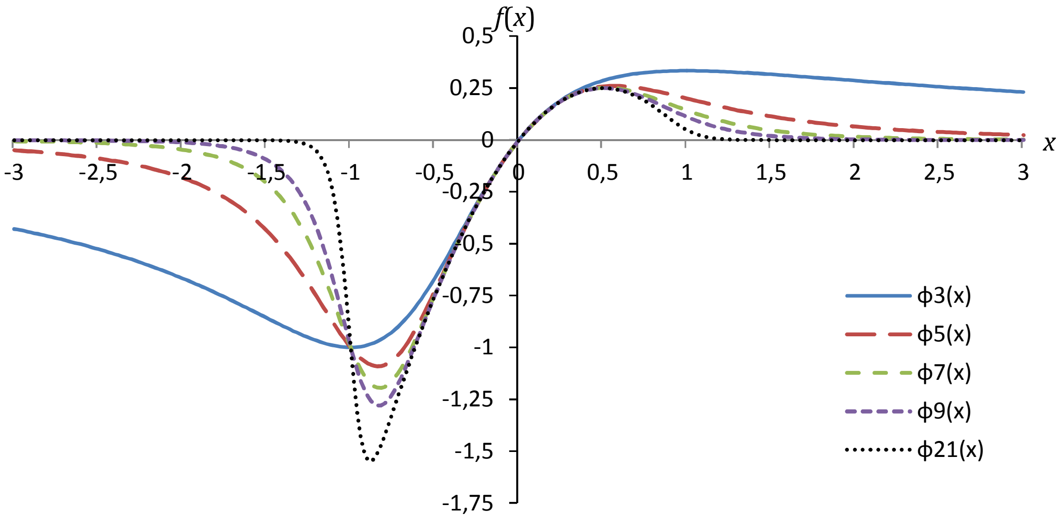


Figure 2.2 – Plots of inverses of potential basis functions $\phi_k(x)$ evaluated over the interval [-3, 3]

It could be seen that as the order of functions increased, the curve would approach an idealised curve of order k = ∞, within which the interval of [-1, 1] would contain the only values of significant magnitude. Because significant (relative) magnitude of values were desired across a broad range for any potential basis function, the function for which this extra-interval magnitude was best preserved was chosen; that is, the inverse polynomial function of order 3. This function also has the useful property that its minima and maxima occur exactly at the points x = {-1, 1}, which will simplify the final equations.

In order to shift/stretch a pair inverse polynomial basis functions to have their minima/maxima fall upon two certain points in x, a transformation in the x-variable was necessary for these two functions. Supposing that t_a and t_b were two points where high variation/magnitude existed in the data, then a pair of basis inverse polynomials of order n = 3 would be defined as:

$$\phi_{3,k}^{-1}(t) = \frac{1}{x_k(t)} + 1 + x_k(t) \tag{21}$$

$$\phi_{3,k+1}^{-1}(t) = \frac{1}{x_k(t)} - 1 + x_k(t) \tag{22}$$

Where:

$$x_k(t) = \frac{2\left(t - t_{a,k}\right)}{t_{b,k} - t_{a,k}} - 1 \tag{23}$$

For basis functions of odd index k and even index k+1, where $t_{a,k}$ and $t_{b,k}$ are two points specific for those two basis functions.

Here, the transformation function from t-space into x-space would scale the independent variable in such a way so as to ensure that the minima and maxima of the pair of inverse polynomial basis functions occurred at points t_a and t_b.

The final thing to consider was the number and choice of t_a and t_b variables. Many possible options were available, but three methods in particular were decided upon:

Type I Method: As mentioned earlier, for a given *odd* number n of inverse polynomial basis functions, the first pair of functions (of the forms in equations 21 & 22) would share the same t-transformation using the t_a and t_b variables of the top two points of greatest variation/magnitude in the data; a second pair of functions would use the second pair of t_a and t_b variables, and so forth.

Type II Method: For any given number n > 1 of inverse polynomial basis functions, with an list of t-variables ordered by variation/magnitude, the k^{th} basis function would utilise variables t_k and t_{k+1} of the list, except for the n^{th} basis function which would use variables t_1 and t_n. In this method, only a single form of basis function (i.e.: that from equation 21) would be necessary.

Type III Method: For any given number $n > 1$ of inverse polynomial basis functions, with n equi-spaced t-variables spanning the longitudinal range of the data (not necessarily the actual points in time of the observations), the k^{th} basis function ($k{\neq}n$) would utilise variables t_k and t_{k+1} defined by the equation:

$$t_k = t_1 + \left(k - \tfrac{1}{2}\right)\left(\frac{t_d - t_1}{n}\right) \tag{24}$$

The n^{th} basis function would utilise t-variables defined by t_1+½(t_d-t_1)/n and t_1+(n-½)(t_d-t_1)/n. In this way, the t-variables were chosen which were in the centre of each of the new n intervals which collectively spanned the range of the data. As in method II, only a single form of basis function (i.e.: that from equation 21) would be necessary.

The subsequent set of inverse polynomial basis functions would then be evaluated for the scaled x-variable on the data points, and the ensuing basis vectors would be used as regressors in the regression model.

2.4.3 Eigenvector model

The eigenvector model differs from the previous polynomial-based models in that it does not utilise a predefined set of basis functions, which by their definition, can be evaluated for any value of t within a continuous interval. Rather, it uses a set of basis vectors, and moreover, these vectors are not predefined, but defined by the data for which they are modelling.

The data, represented in matrix notation as matrix Y, with N rows corresponding to each record, and d columns corresponding to specific time points in which data observations were made for each record, would then have its corresponding covariance matrix calculated. This d×d covariance matrix, $\mathrm{cov}(\mathrm{Y})$, would then have the eigenvectors derived from it.

Given a d×d matrix M, an eigenvalue λ exists if there is a d-dimensional (non-empty) vector $\mathbf{u}$ for which $\mathrm{M}\mathbf{u} = \lambda\mathbf{u}$, where $\mathbf{u}$ is the eigenvector corresponding to λ. If M is a symmetrical matrix of real values, then there exist d real eigenvalues (not necessarily distinct), with d corresponding eigenvectors. The eigenvectors constitute an orthonormal basis of $\mathbb{R}^d$; the d-dimensional vector-space of real numbers.

Such a matrix M can be represented as a product of matrices comprised of its eigenvectors and eigenvalues as follows, in what is known as the spectral decomposition:

$$\mathrm{M} = \mathrm{Q}\Lambda\mathrm{Q}^T = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_d \end{bmatrix} \begin{bmatrix} \leftarrow & \mathbf{u}_1 & \rightarrow \\ \leftarrow & \mathbf{u}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{u}_d & \rightarrow \end{bmatrix} \tag{25}$$

Here Q is the matrix of eigenvectors $\mathbf{u}_i$, and Λ is the diagonal matrix of corresponding eigenvectors λ_i.

When the eigenvalues were defined such that $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d$, a subset of the corresponding eigenvectors $\mathbf{u}_1$, …, $\mathbf{u}_n$ would form the set of basis vectors for an eigenvector model of order $n{\leqslant}d$.

The transpose of this subset, $\mathrm{Q}_n{}^{\mathrm{T}}$, would thus function as the basis matrix Φ_n. From there, the set of basis vectors would go on to become the regressors in the regression model of the data.

In order to perform the regression for any time-points within the intervals bounded by the d time-points of the study, an extrapolation of the eigenvectors was required. Piece-wise functions were determined for each eigenvector; both a linear piece-wise function, and a natural cubic spline function. For each function of each eigenvector, values were determined from the evaluation of 100 equi-spaced time-points per interval, and in turn came to comprise the elements of extended eigenvectors of length $100(d\text{-}1)\text{+}1$.

2.5 Analysis of models

The effectiveness of each longitudinal model when applied within the fixed regression model would be evaluated by three different methods, for different groupings of the longitudinal data. The measures of goodness of fit could then be compared between models to assess their relative merit.

An F-test would also be applied to combinations of any two models to get a direct quantitative measure of the significance by which one model may perform over another.

2.5.1 Measure of accuracy

The measure of accuracy was a direct correlation between the actual and the predicted values, for any time-point t_j:

$$r_{y_j,\hat{y}_j} = \frac{\sum_{i=1}^{N}(y_{ij} - \bar{y}_j)(\hat{y}_{ij} - \bar{\hat{y}}_j)}{(N-1)\sigma_{y_j}\sigma_{\hat{y}_j}} \tag{26}$$

The accuracy of the model itself, over all d time-points, was simply the average of the correlations for each time-point:

$$r_{y,\hat{y}} = \frac{1}{d}\sum_{j=1}^{d} r_{y_j,\hat{y}_j} \tag{27}$$

2.5.2 Measure of variance

The calculation for measure of variance of the estimated values for the time-points with each model came directly from the regression equation:

$$\begin{aligned} var(\hat{y}_{ij}) &= var\left(\sum_{k=0}^{n} \phi_{jk}\beta_{ik}\right) \\ &= \sum_{a=1}^{n}\sum_{b=1}^{n} \Phi_{a,j}\Phi_{b,j} cov(\beta_{ia},\beta_{ib}) \\ &= \sum_{a=1}^{n}\sum_{b=1}^{n} \Phi_{a,j}\Phi_{b,j}\,[cov(\beta)]_{a,b} \\ &= var(\hat{y}_j) \end{aligned} \tag{28}$$

The covariance of the β-coefficients came from the elements of the covariance matrix of all β-coefficients, β, and were thus independent of i. Therefore the variance for any time-point t_j is the same for all i.

2.5.3 Percentage squared bias

The percentage squared bias (PSB) is a measure of the sums of squares of the residuals, relative to the sum of squared actual values:

$$PSB_j = 100 * \frac{\sum_{i=1}^{N}\left(y_{ij} - \hat{y}_{ij}\right)^2}{\sum_{i=1}^{N} {y_{ij}}^2} \tag{29}$$

The PSB of the model itself, over all d time-points, was simply an extension of equation 29, this time with both the residuals and the squared data summed over all individuals *and* time-points:

$$PSB = 100 * \frac{\sum_{i=1}^{N}\sum_{j=1}^{d}\left(y_{ij} - \hat{y}_{ij}\right)^2}{\sum_{i=1}^{N}\sum_{j=1}^{d} y_{ij}^2} \tag{30}$$

2.5.4 *F*-test

An F-test to compare two models could be used to either compare two types of models of the same order, n, or for comparing two models of the same type but with differing orders n. The resultant F-statistic $F_{1,2}$ would explain which model best fitted the data, whilst the p-value of the F-statistic would indicate the level of significance of the difference in models.

For models 1 and 2, the F-statistic was given by:

$$\text{for } n_1 = n_2, \quad F_{1,2} = \frac{RSS_1}{RSS_2} \tag{31}$$

$$\text{for } n_1 < n_2, \quad F_{1,2} = \frac{(RSS_1 - RSS_2)/(n_2 - n_1)}{RSS_2/(N - n_2)} \tag{32}$$

Where n_1 and n_2 are the orders of the respective models, the number of individuals is N = 1,000, and per model the RSS, the sum of squares of the residuals, is simply:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{33}$$

If $F_{1,2}$ > 1, then the second model was better than the first at fitting the data, and if $F_{1,2}$ < 1 the first model was better than the second. In the case of $F_{1,2}$ < 1, the two models should be reversed in the equation to yield the larger F-statistic.

By supplying the F-statistic to an F-distribution function, along with the two degrees of freedom of the models, n_1 and n_2, a p-value would be obtained. For a p-value < 0.05, it could be concluded that the difference in the two models was a *significant* one.

2.6 Software

All analysis was performed using the R statistical software. Additional R-packages used in the programming included the *MASS* package (Venables and Ripley, 2002) for use in matrix inversion, and the *mvtnorm* package (Genz and Bretz, 2009) for use in sampling from the multi-variate normal distribution when generating the simulated data.

3. Results

3.1 Basis sets

For each longitudinal model type, basis vectors were determined from the evaluation of the respective interpolating basis functions of the data. The results of these evaluations are presented in both tabular and graphical form below.

3.1.1 Legendre basis set

The basis vectors of each Legendre polynomial were determined from the evaluation of the t-variables transformed for the [-1, 1] interval. Table 3.1 shows these vectors from the evaluation of the d time-points of the study, while figure 3.1 shows them for both the d time-points of the study, and for the $100(d\text{-}1)\text{+}1$ extrapolated time-points.

Table 3.1 - Legendre basis polynomials evaluated for the 6 (transformed) t-variables

t	0	1	2	3	4	5
$\phi_1(t)$	1	1	1	1	1	1
$\phi_2(t)$	-1	-0.6	-0.2	0.2	0.6	1
$\phi_3(t)$	1	0.04	-0.44	-0.44	0.04	1
$\phi_4(t)$	-1	0.36	0.28	-0.28	-0.36	1
$\phi_5(t)$	1	-0.408	0.232	0.232	-0.408	1
$\phi_6(t)$	-1	0.15264	-0.30752	0.30752	-0.15264	1

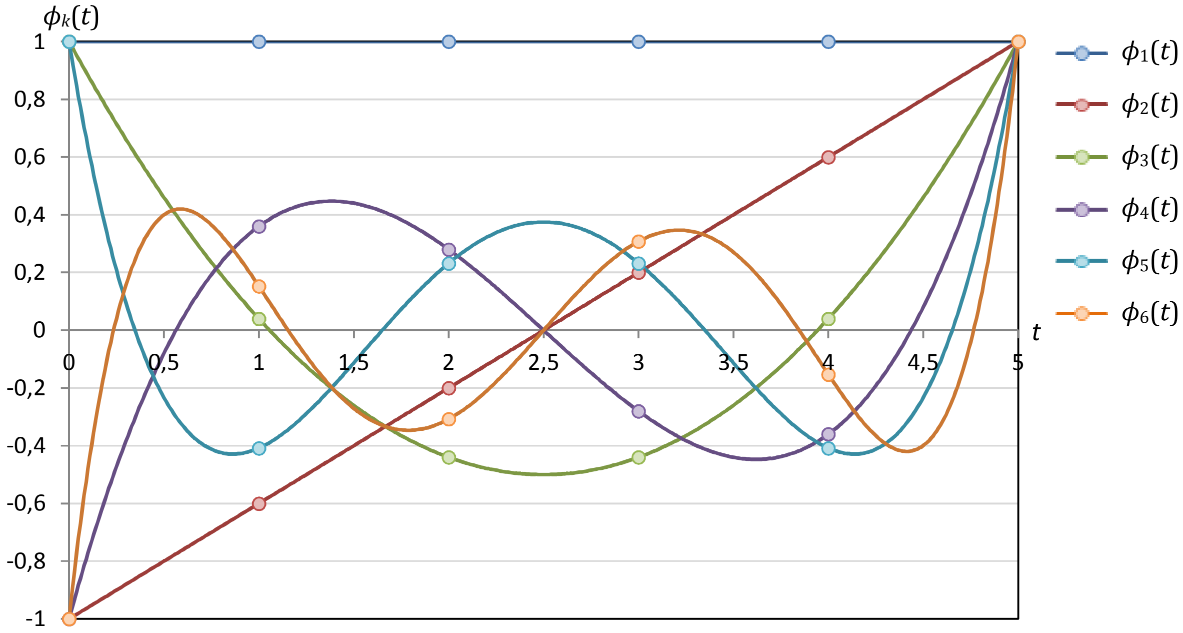


Figure 3.1 - Legendre basis polynomials evaluated for the (transformed) t-variables; the 6 time-points of the study (dots), and the extrapolated time-points over each interval (lines).

3.1.2 Inverse polynomial basis sets

For the inverse polynomials of types I and II, there was a single basis set each. For the type III inverse polynomials however, the basis set would depend upon the order of the set, so for the six different orders, there were six corresponding basis sets. This was due to the vector of n equi-spaced t_a and t_b time-points used in the set-up of the basis functions per basis set of order n.

Tables 3.2, 3.3 and 3.4 show the basis vectors from the evaluation of the d time-points of the study for inverse polynomial method type I, II and III respectively. Figures 3.2 to 3.9 show these vectors for both the d time-points of the study, and for the $100(d\text{-}1)\text{+}1$ extrapolated time-points.

Table 3.2 - Inverse basis polynomials of type I evaluated for the 6 (transformed) t-variables

t	0	1	2	3	4	5
$\phi_1(t)$	-0.4286	-1	0.3333	0.2308	0.1613	0.1228
$\phi_2(t)$	-0.2308	-0.3333	1	0.4286	0.2381	0.1628
$\phi_3(t)$	-1	-0.4286	0.2308	0.3333	0.3061	0.2658
$\phi_4(t)$	-0.3333	-0.2308	0.4286	1	0.7895	0.5676
$\phi_5(t)$	-0.1233	-0.1628	-0.2381	-0.4286	-1	0.3333
$\phi_6(t)$	-0.0989	-0.1228	-0.1613	-0.2308	-0.3333	1

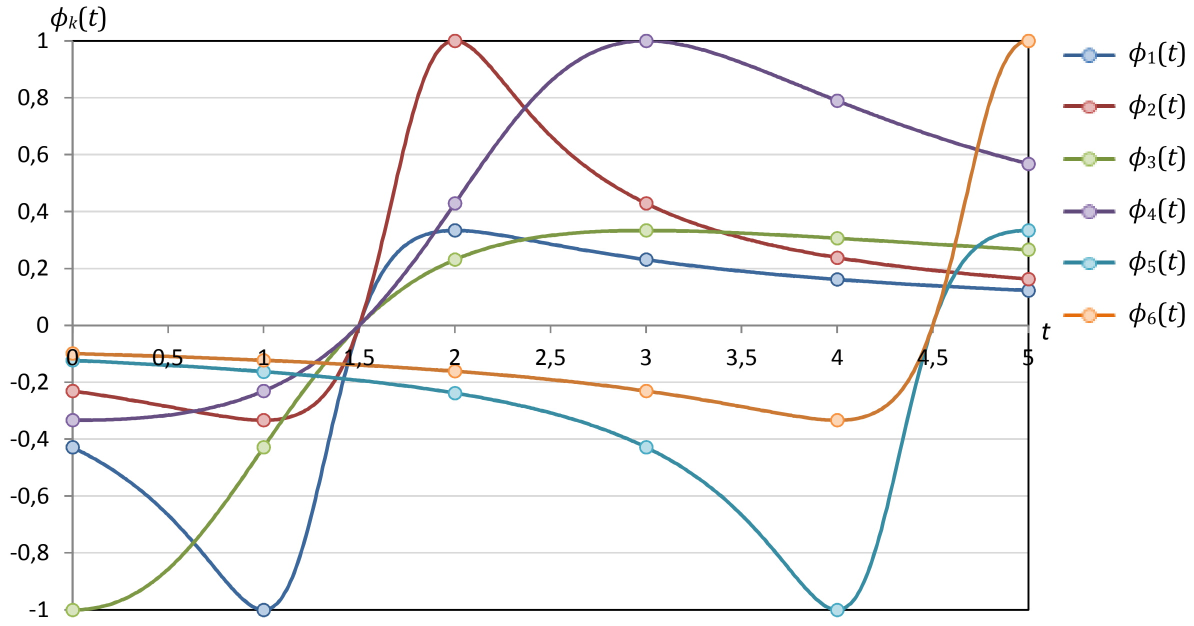


Figure 3.2 – Inverse polynomial type I basis polynomials evaluated for the (transformed) t-variables; the 6 time-points of the study (dots), and the extrapolated time-points over each interval (lines).

Table 3.3 - Inverse basis polynomials of type II evaluated for the 6 (transformed) t-variables

t	0	1	2	3	4	5
$\phi_1(t)$	-0.4286	-1	0.3333	0.2308	0.1613	0.1228
$\phi_2(t)$	0.3333	0	-1	-0.6667	-0.4286	-0.3077
$\phi_3(t)$	-1	-0.4286	0.2308	0.3333	0.3061	0.2658
$\phi_4(t)$	-0.1628	-0.2381	-0.4286	-1	0.3333	0.2308
$\phi_5(t)$	-0.1233	-0.1628	-0.2381	-0.4286	-1	0.3333
$\phi_6(t)$	0.3158	0.3333	0.2857	0	-0.6667	-1

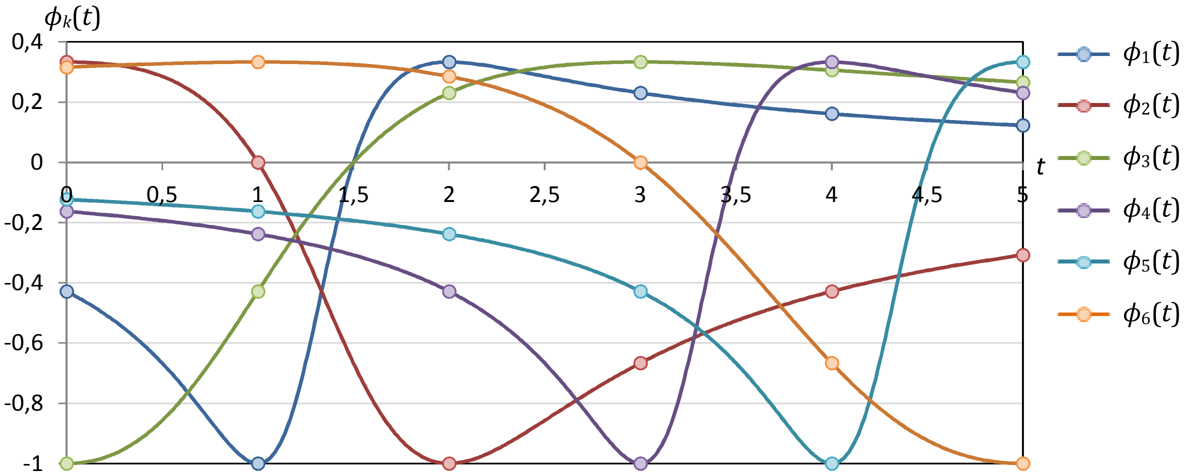


Figure 3.3 - Inverse polynomial type II basis polynomials evaluated for the (transformed) t-variables; the 6 time-points of the study (dots), and the extrapolated time-points over each interval (lines).

Table 3.4 - Inverse basis polynomials of type III for each order *n*, evaluated for the 6 (transformed) *t*-variables

n	k	$\phi_k(0)$	$\phi_k(1)$	$\phi_k(2)$	$\phi_k(3)$	$\phi_k(4)$	$\phi_k(5)$
1	1	0.2857	0.3297	0.2564	-0.5263	-0.9677	-0.6667
2	1	-0.6667	-0.9524	0.2564	0.3101	0.2405	0.1905
	2	0.3158	0.3321	0.2158	-0.3797	-0.989	-0.8571
3	1	-0.6667	-0.5263	0.3297	0.2405	0.1777	0.1395
	2	-0.3077	-0.5505	-0.9524	0.3279	0.262	0.1905
	3	0.3243	0.3279	0.1993	-0.3315	-0.9524	-0.9231
4	1	-0.6667	0	0.2857	0.1905	0.1395	0.1096
	2	-0.3077	-0.6667	0	0.2857	0.1905	0.1395
	3	-0.1935	-0.3077	-0.6667	0	0.2857	0.1905
	4	0.3279	0.3243	0.1905	-0.3077	-0.9231	-0.9524
5	1	-0.6667	0.2564	0.2405	0.1564	0.1145	0.0901
	2	-0.3077	-0.8163	0.3279	0.2216	0.1475	0.1096
	3	-0.1935	-0.3475	-0.9677	0.3297	0.205	0.1395
	4	-0.1404	-0.2093	-0.398	-0.9524	0.3101	0.1905
	5	0.3297	0.3217	0.185	-0.2935	-0.9018	-0.9677

6	1	-0.6667	0.3279	0.205	0.1323	0.097	0.0764
	2	-0.3077	-0.9677	0.3101	0.1777	0.1199	0.0901
	3	-0.1935	-0.398	-0.5263	0.262	0.1564	0.1096
	4	-0.1404	-0.2277	-0.5505	0.2564	0.2216	0.1395
	5	-0.1099	-0.1578	-0.2757	-0.8163	0.3297	0.1905
	6	0.3307	0.3196	0.1812	-0.2842	-0.8861	-0.9767

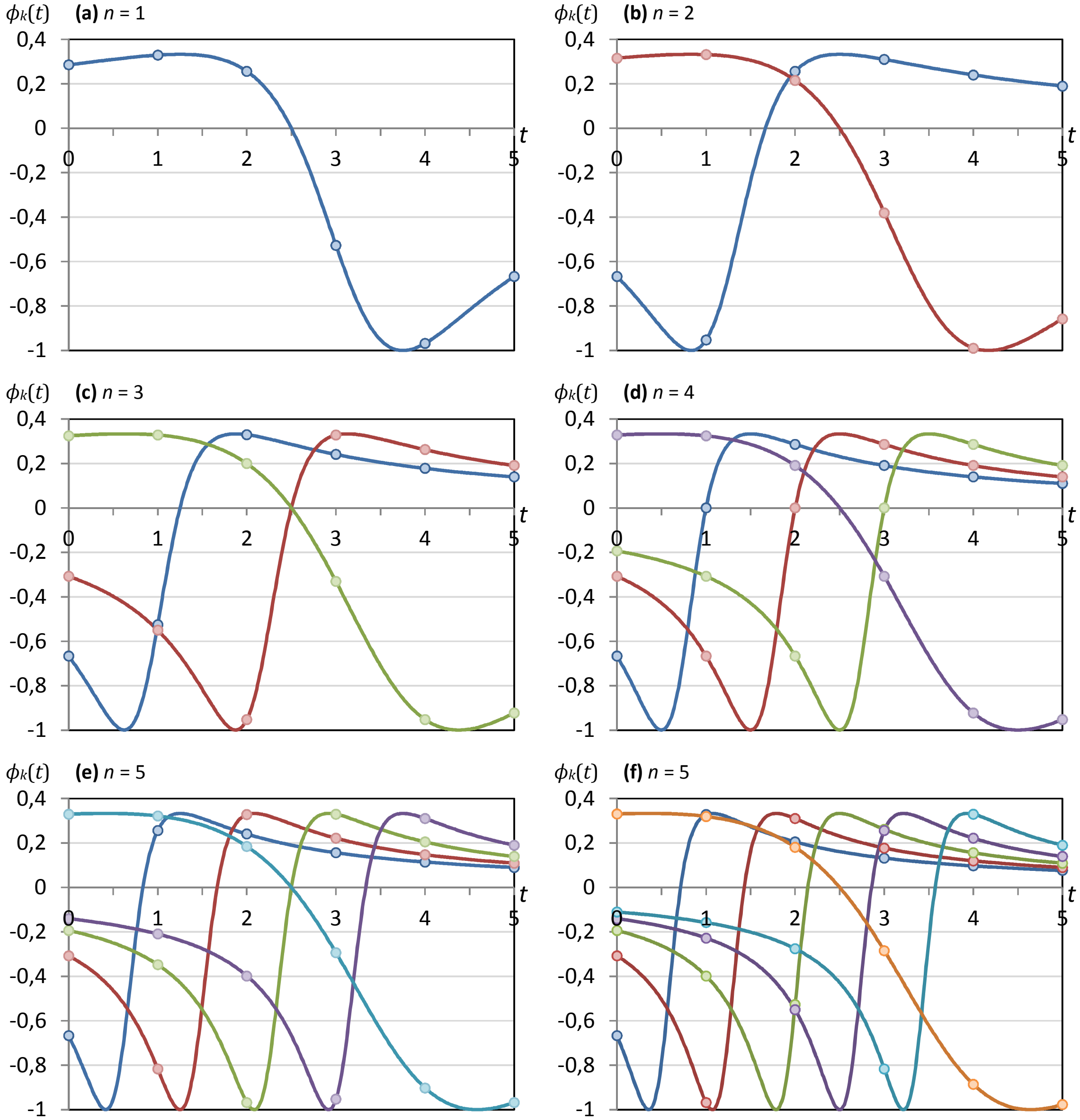


Figure 3.4 - Inverse polynomial type III basis polynomials for basis sets of order *n*, evaluated for the (transformed) t-variables; the 6 time-points of the study (dots), and the extrapolated time-points over each interval (lines).

3.1.3 Eigenvector basis sets

The basis vectors for the eigenvector sets were the eigenvectors obtained from the spectral decomposition of the covariance matrix of a dataset. These eigenvectors were ordered by their corresponding eigenvalues, from largest to smallest, to form the order of basis vectors.

Because the eigenvectors are not independent of the data like the other model types, there will be a basis set of eigenvectors for each specific dataset. For each dataset of a particular noise variance, tables of eigenvalues and eigenvectors, along with plots of both the piecewise linear and natural cubic spline interpolations of the eigenvectors, are presented below.

Table 3.5 - Eigenvalues and eigenvectors derived from the simulated dataset with noise variance of 0.

Eigenvalues:	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
	2.2724	0.3291	1.5×10^{-16}	4.2×10^{-17}	3.5×10^{-17}	2.0×10^{-17}
Eigenvectors:						
t	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$
0	-0.3938	0.2586	0.8821	0	0	0
1	-0.6706	0.3668	-0.4069	0.1173	-0.2941	-0.3872
2	-0.4927	-0.0175	-0.2148	-0.3392	0.3772	0.6734
3	-0.2668	-0.4417	0.0104	0.7018	0.4769	-0.117
4	-0.2166	-0.5797	0.0732	-0.5939	0.0916	-0.5006
5	-0.1854	-0.5169	0.0687	0.1611	-0.7317	0.3638

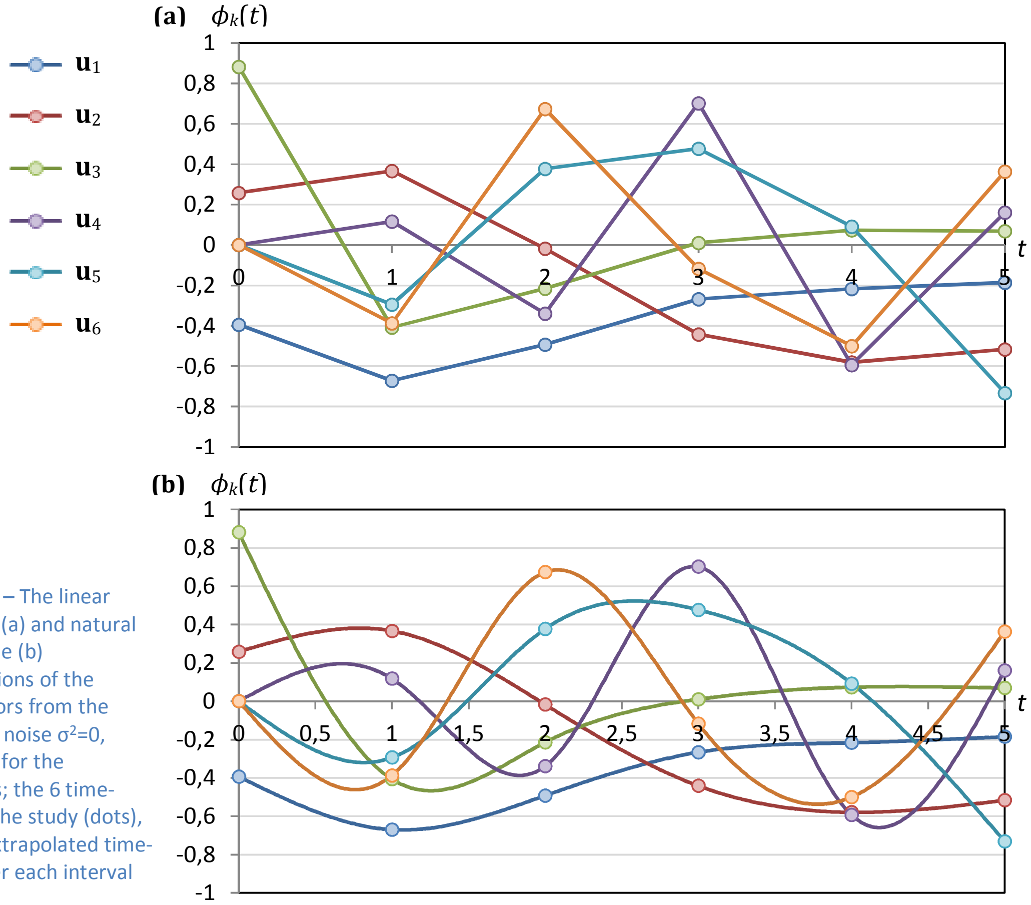


Figure 3.5 – The linear piecewise (a) and natural cubic spline (b) interpolations of the eigenvectors from the dataset of noise σ^2=0, evaluated for the t-variables; the 6 time-points of the study (dots), and the extrapolated time-points over each interval (lines).

It is worth paying particular attention to the eigenvalues in this data. When there is no added noise in the data, there are effectively only two eigenvectors of significance; the remaining four have eigenvalues of approximately zero. As the amount of added noise in the data increases, so do the value of the eigenvalues, and thus the significance of the corresponding eigenvectors.

Table 3.6 - Eigenvalues and eigenvectors derived from the simulated dataset with noise variance of 0.2

Eigenvalues:	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
	2.2921	0.3685	0.0418	0.0412	0.0405	0.0381
Eigenvectors:						
t	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$
0	-0.3922	0.2612	0.309	-0.2114	0.7322	0.3188
1	-0.6671	0.3732	-0.196	-0.2877	-0.2915	-0.4578
2	-0.4949	-0.0188	-0.1121	0.761	-0.1539	0.3734
3	-0.2691	-0.4366	0.6227	-0.3035	-0.4448	0.2434
4	-0.2199	-0.5748	-0.6486	-0.3412	0.156	0.2447
5	-0.1886	-0.5207	0.2124	0.2914	0.3648	-0.6559

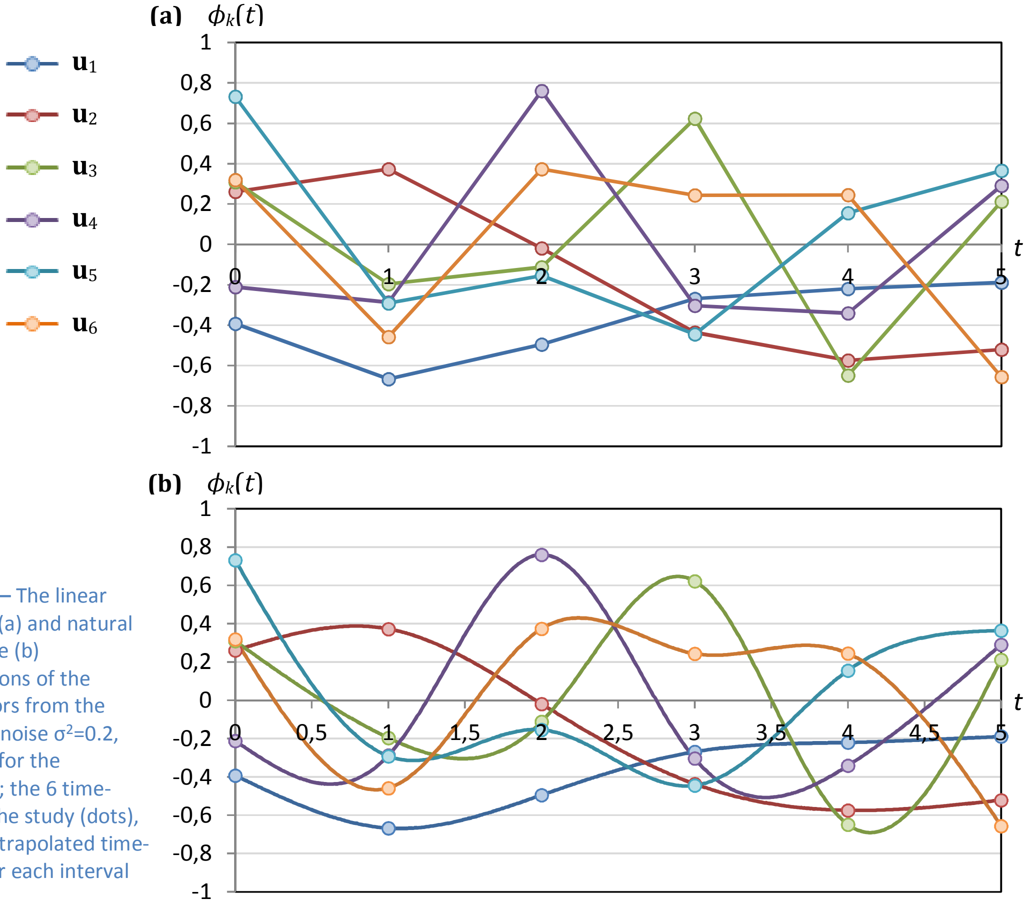


Figure 3.6 – The linear piecewise (a) and natural cubic spline (b) interpolations of the **e**igenvectors from the dataset of noise σ²=0.2, evaluated for the t-variables; the 6 time-points of the study (dots), and the extrapolated time-points over each interval (lines).

Table 3.7 - Eigenvalues and eigenvectors derived from the simulated dataset with noise variance of 0.5

Eigenvalues:	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
	2.5343	0.5795	0.2636	0.2615	0.2340	0.2296
Eigenvectors:						
t	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$
0	-0.3883	0.2865	0.2229	0.6837	-0.4901	0.0991
1	-0.6632	0.3362	-0.2777	-0.1087	0.3453	-0.4888
2	-0.5022	-0.0038	0.4496	-0.5026	0.0086	0.5412
3	-0.276	-0.3799	-0.7102	-0.1268	-0.4586	0.2208
4	-0.2264	-0.5489	-0.011	0.4919	0.5887	0.2425
5	-0.1723	-0.5994	0.4081	-0.1011	-0.2891	-0.5923

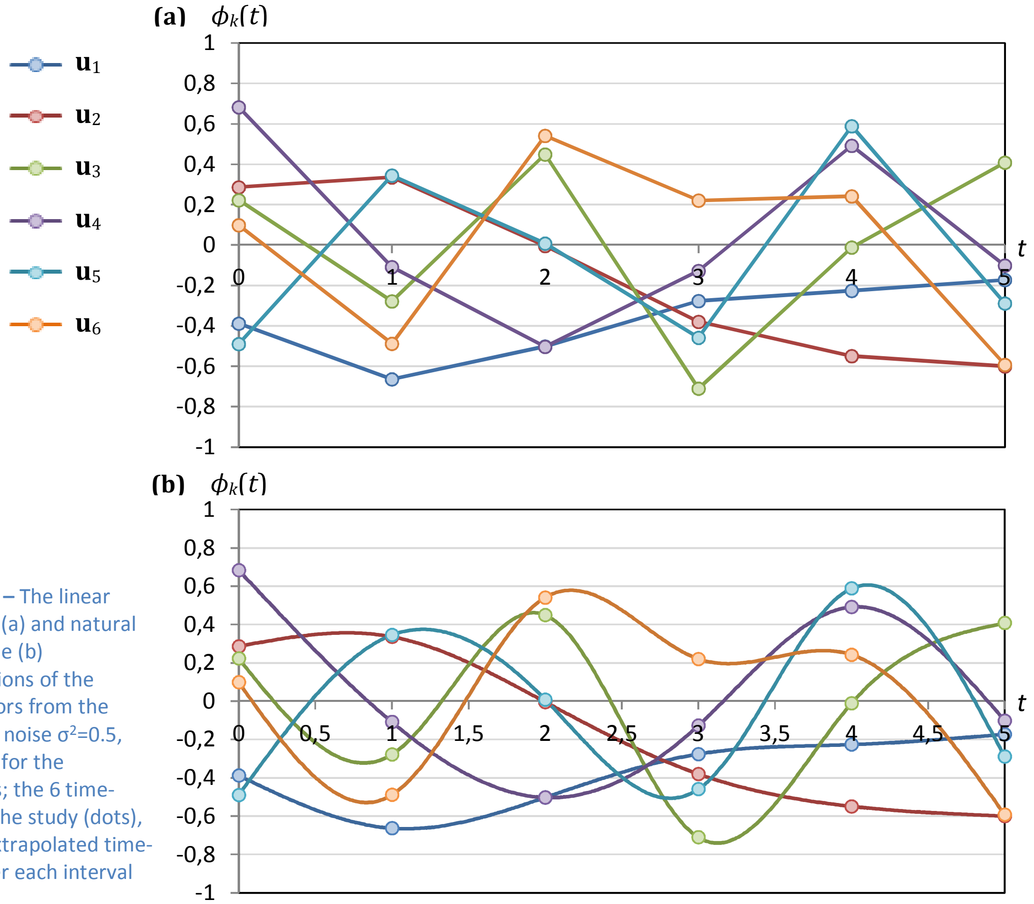


Figure 3.7 – The linear piecewise (a) and natural cubic spline (b) interpolations of the eigenvectors from the dataset of noise σ^2=0.5, evaluated for the t-variables; the 6 time-points of the study (dots), and the extrapolated time-points over each interval (lines).

Table 3.8 - Eigenvalues and eigenvectors derived from the simulated dataset with noise variance of 1.

Eigenvalues:	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
	3.4552	1.2801	1.0085	1.0026	0.9800	0.8934
Eigenvectors:						
t	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$
0	-0.4518	0.4299	-0.1074	-0.4376	-0.2244	0.5981
1	-0.6343	0.2714	-0.1814	0.2412	-0.2298	-0.6166
2	-0.4795	-0.0519	0.5377	0.3001	0.5862	0.2111
3	-0.2903	-0.4955	-0.5699	-0.3683	0.4534	-0.0648
4	-0.1954	-0.5113	-0.2045	0.5406	-0.4428	0.4126
5	-0.2031	-0.4815	0.5475	-0.4821	-0.3893	-0.2077

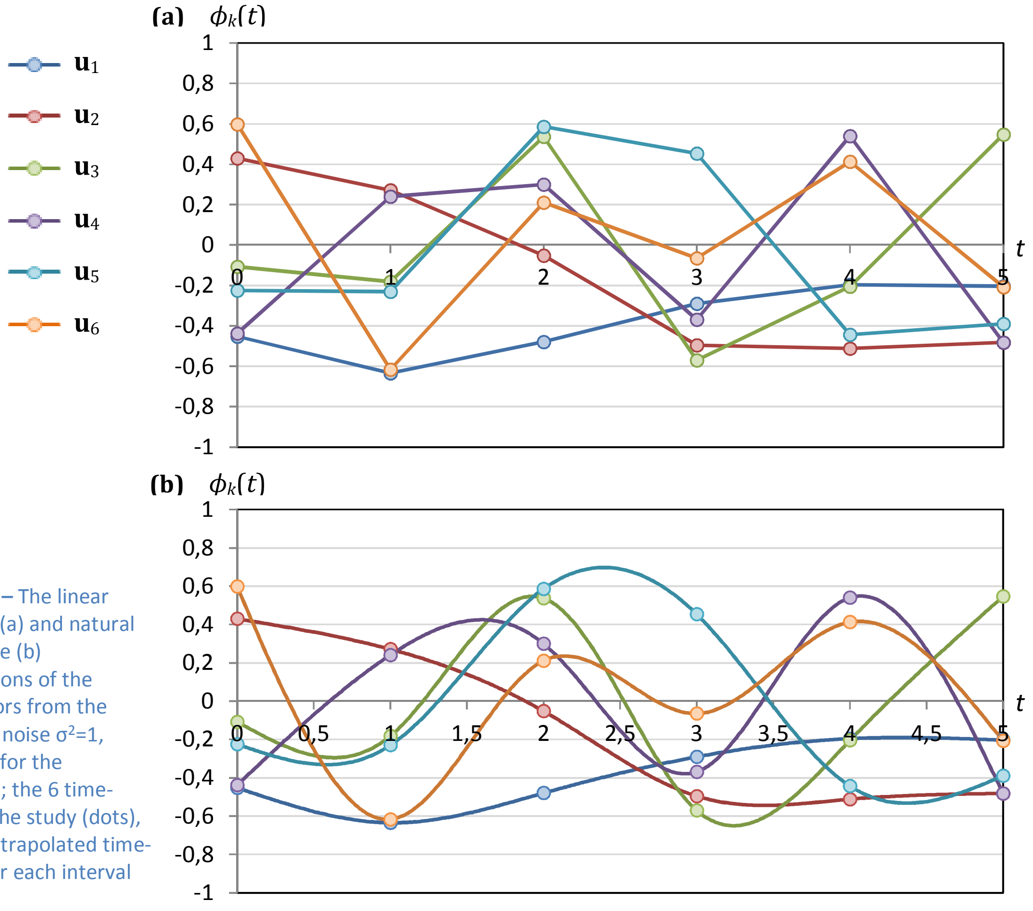


Figure 3.8 – The linear piecewise (a) and natural cubic spline (b) interpolations of the eigenvectors from the dataset of noise σ^2=1, evaluated for the t-variables; the 6 time-points of the study (dots), and the extrapolated time-points over each interval (lines).

3.2 Accuracy of models

The measures of accuracy of the models is primarily based upon the simulated dataset with an added noise variance of 1, for the various time-points of the data, although the accuracies when using other datasets are presented for average model accuracy.

3.2.1 Legendre model accuracy

The accuracies for the Legendre models of different orders for fitting the dataset with a noise variance of 1 are plotted as follows:

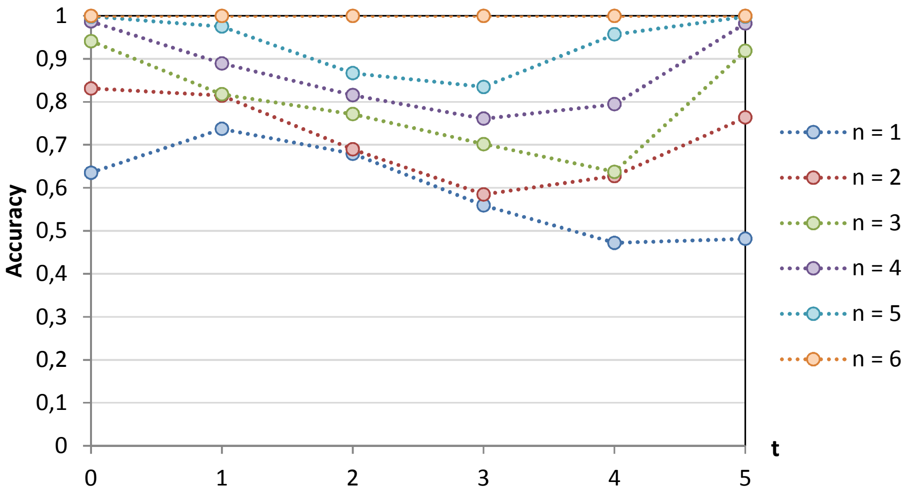


Figure 3.9 – Accuracies of orders *n* of the Legendre model, per time-point of the data, for dataset with noise $\sigma^2 = 1$.

The Legendre model appears to perform well at the edges of the range of the data for most orders, whilst the central time-points are not fitted quite as well.

Additionally, the average accuracy across all time-points was plotted for different orders of the Legendre model when fitting datasets of varying noise variances:

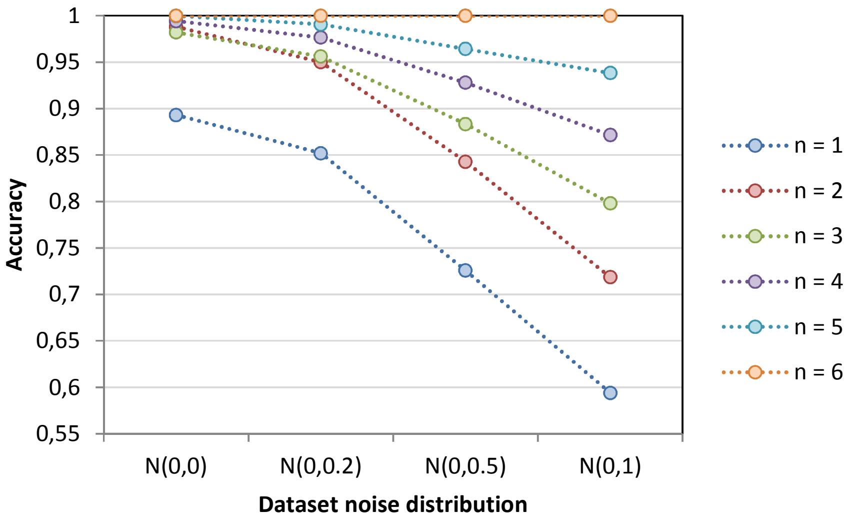
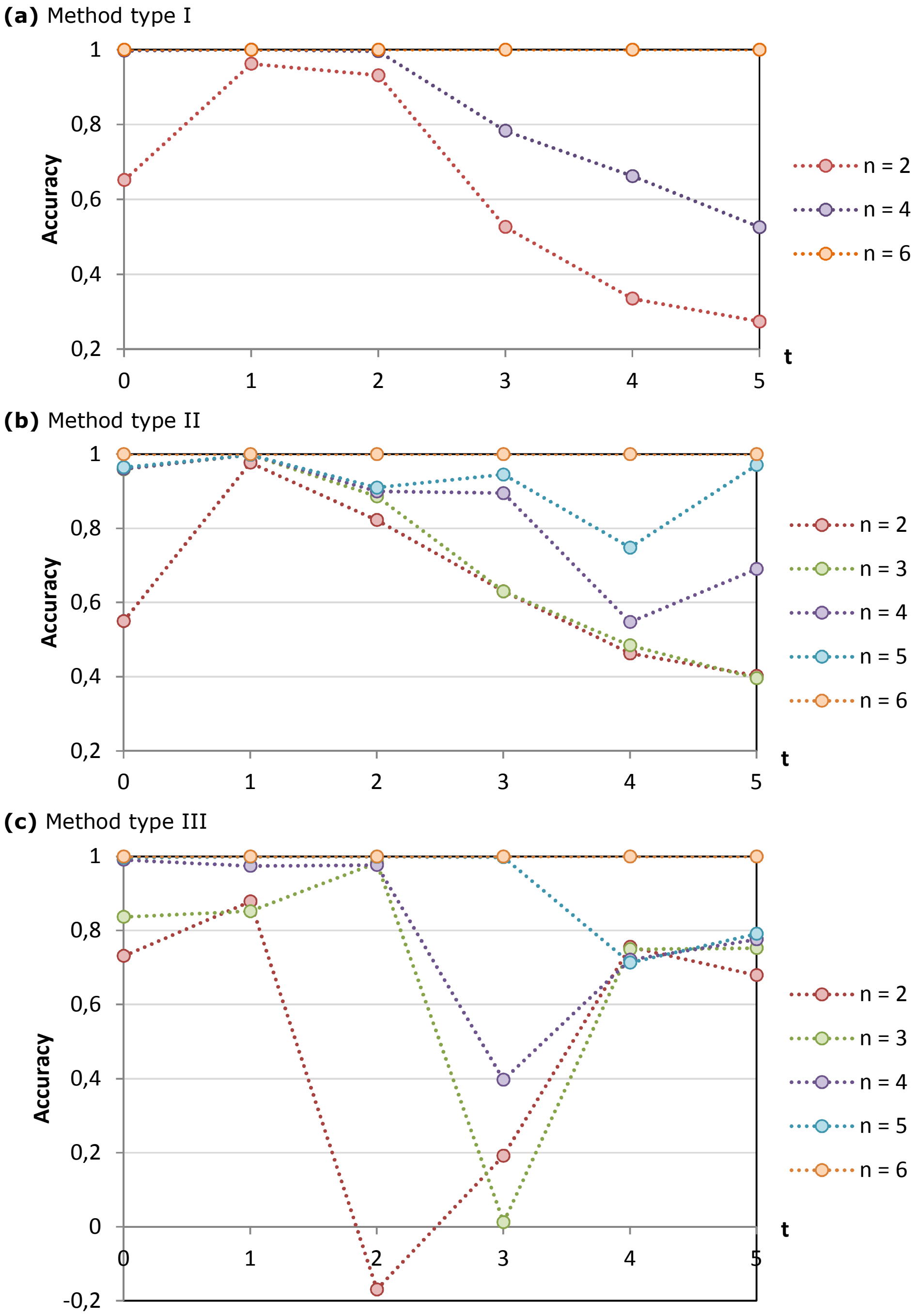


Figure 3.10 – Average accuracies for Legendre models of orders *n*, for datasets of differing noise variance.

As expected, the less noise in the dataset, the better the model prediction. Whilst higher orders of the model make little difference for datasets of low noise, these higher orders become more important as the dataset noise increases.

3.2.2 Inverse polynomial models accuracy

The accuracies for the three different inverse polynomial model types, of different orders for fitting the dataset with a noise variance of 1, are plotted as follows:


(a) Method type I

(b) Method type II

(c) Method type III

Figure 3.11 – Accuracies of orders *n* of the inverse polynomials model types I, II & III, per time-point of the data, for dataset with noise σ^2 = 1.

These models did not perform as well as the Legendre models, with inverse polynomial method type III being particularly unhelpful. While the accuracies for some of the central points for the higher orders of model performed quite well compared to the Legendre model, the edges of the range of the data had poor accuracies, and moreover, the inconsistencies across the range were greater for all orders of the inverse polynomials than they were for the Legendre polynomials.

3.2.3 Eigenvector model accuracy

The accuracies for the Eigenvector models of different orders for fitting the dataset with a noise variance of 1 are plotted as follows:

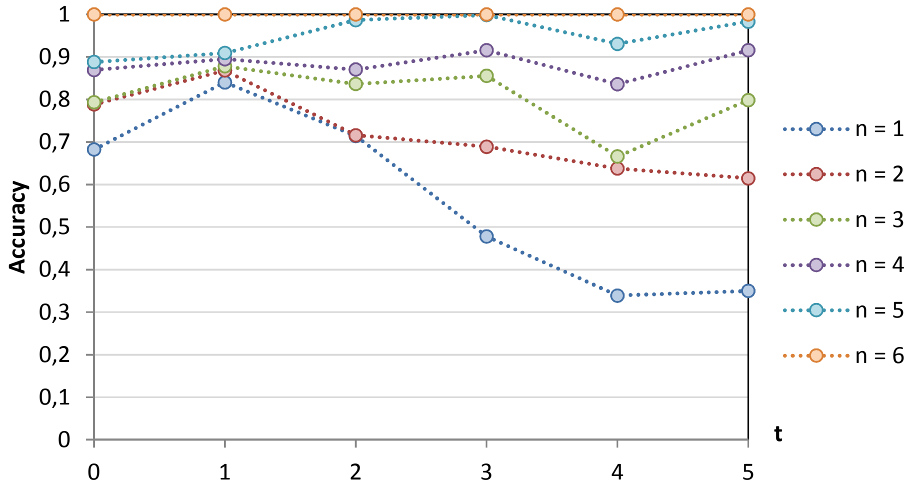


Figure 3.12 - Accuracies of orders *n* of the Eigenvector model, per time-point of the data, for dataset with noise σ^2 = 1.

Apart from the first order model, the eigenvector model appears to have a marginal advantage over the Legendre model, particularly for the mid-range time-points. However, the accuracies at the edges of the range of the data do not fare quite as well.

Additionally, the average accuracy across all time-points was plotted for different orders of the eigenvector model when fitting datasets of varying noise variances:

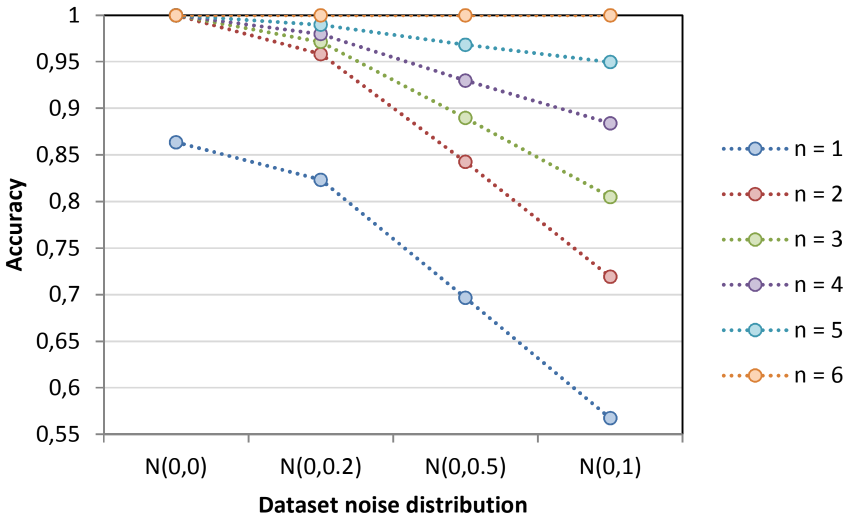


Figure 3.13 - Average accuracies for eigenvector models of orders *n*, for datasets of differing noise variance.

The average accuracy per model order for each of the datasets of differing noise variance is almost identical to that of the Legendre models. Apart from the order *n* = 1 models, the eigenvector models show a very slight advantage over the Legendre models, mainly for the datasets of low noise variance.

3.3 Variance of models

Similar to the previous section on model accuracy, the measures of variation of the models is primarily based upon the simulated dataset with an added noise variance of 1, for the various time-points of the data.

The actual variance of the data for the various datasets of differing noise variance added is provided for each time-point as follows:

Table 3.9 – Actual variance of the data per time-point, for each dataset with noise of differing variances

t	0	1	2	3	4	5
N(0,0)	0.3744	1.0662	0.5517	0.2259	0.2172	0.1660
N(0,0.2)	0.4091	1.0878	0.5921	0.2665	0.2582	0.2086
N(0,0.5)	0.6235	1.2864	0.8259	0.4742	0.4624	0.4301
N(0,1)	1.5145	1.9672	1.5562	1.2742	1.1460	1.1617

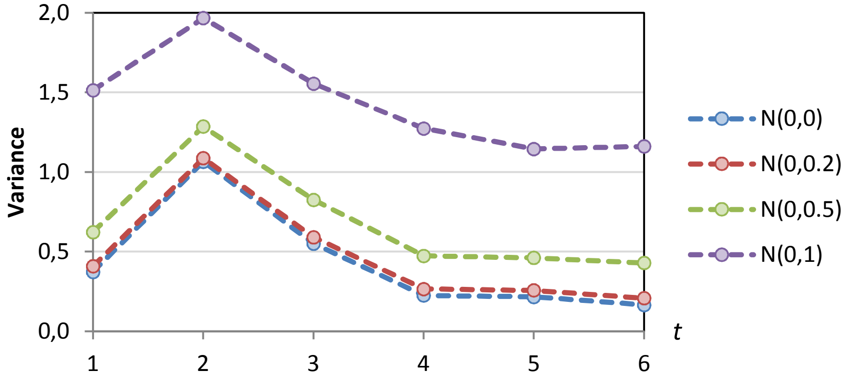


Figure 3.14 - Actual variance of the data per time-point, for each dataset with noise of differing variances

The variances given for each of the following models are presented as percentages of the actual variance of the data.

3.3.1 Legendre model variance

The variance over time (relative to the actual variance over time) for the Legendre models of different orders for fitting the dataset with a noise variance of 1 are plotted as follows:

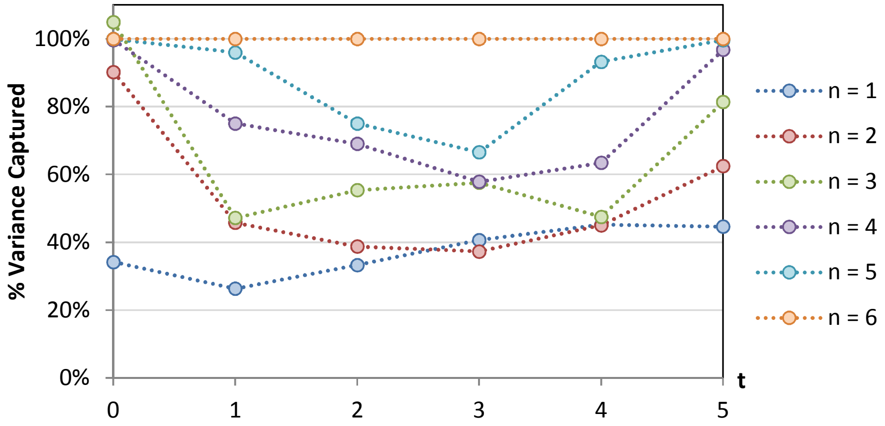


Figure 3.15 – Variance of the Legendre model per time-point for dataset with noise σ^2 = 1, as a percentage of actual variance

Again the Legendre model appears to do better at the edges of the range of the data than it does for the mid-range area. Interestingly, the model of order n=3 has over-captured the variance of time-point t_1. In fact, such over-capture is worse for datasets of lower noise. For example, the relative variance of the Legendre models on the dataset with additional noise from the N(0,0.2) distribution is plotted as follows:

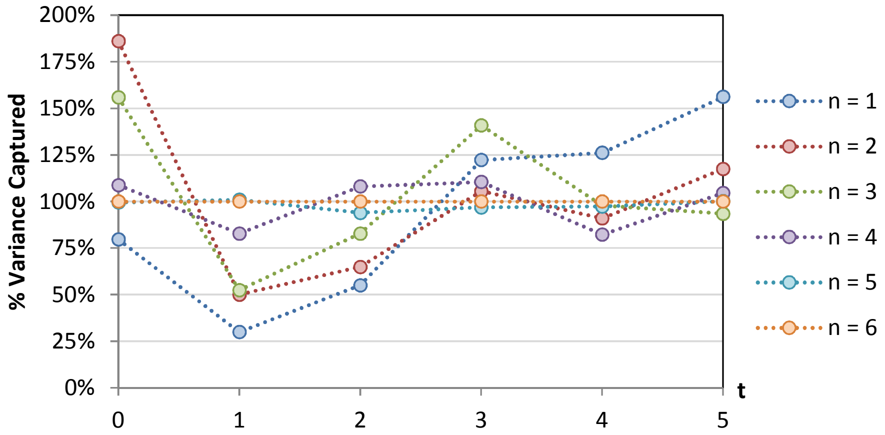
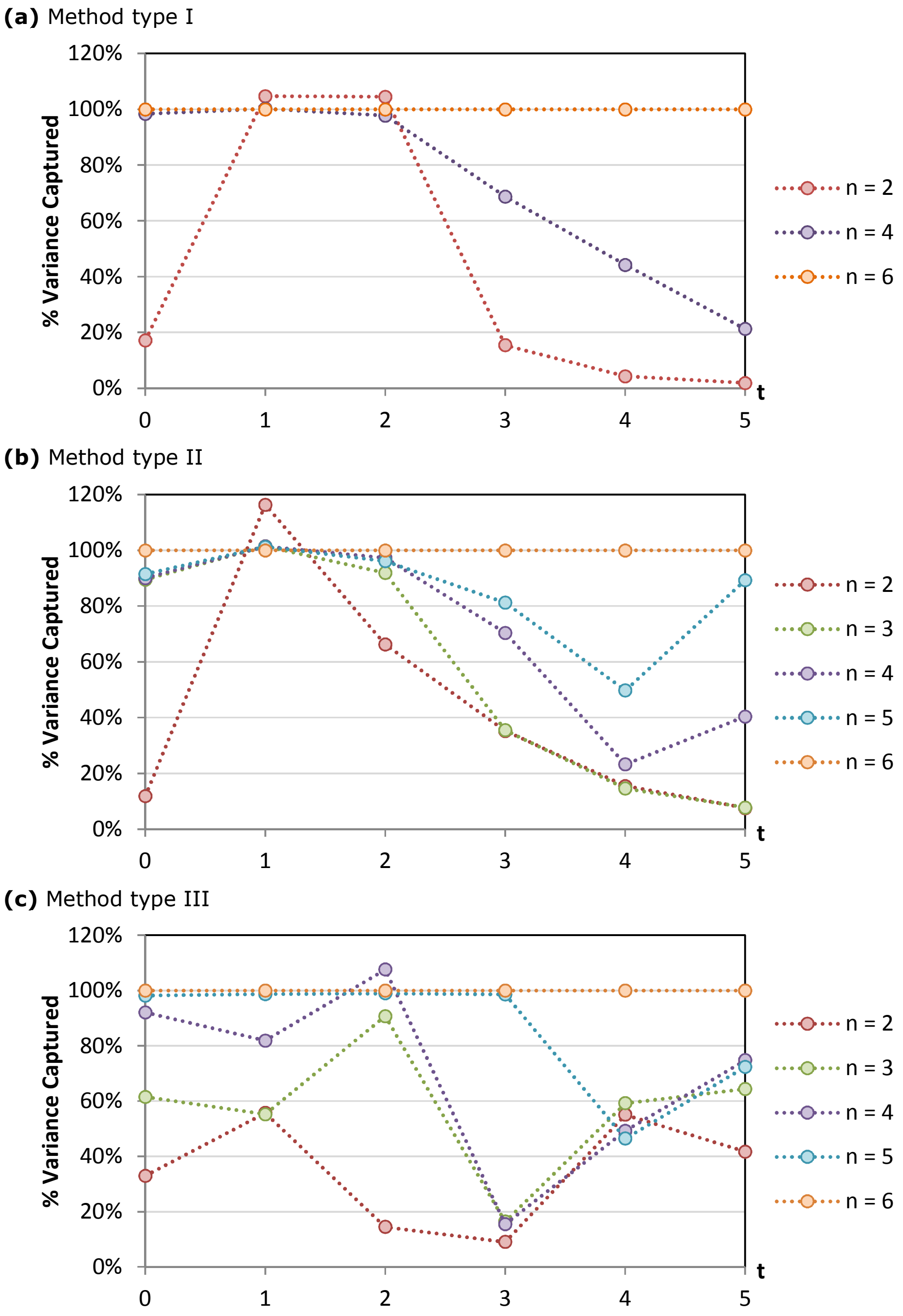


Figure 3.15 – Variance of the Legendre model per time-point for dataset with noise σ^2 = 0.2, as a percentage of actual variance

Clearly a model variance which is close to the actual variance does not translate to a measure of accuracy, particularly for lower orders of the model.

3.3.2 Inverse polynomial models variance

The variance over time (relative to the actual variance over time) for the three different types of inverse polynomials models of different orders for fitting the dataset with a noise variance of 1 are plotted as follows:



(a) Method type I

(b) Method type II

(c) Method type III

Figure 3.16 – Variance of orders *n* of the inverse polynomials model types I, II & III, per time-point of the data, for dataset with noise σ^2 = 1, as a percentage of actual variance

Again the inverse polynomials do not seem to perform as well as the Legendre polynomials, particularly the one of method type III. For the higher orders of the method I and II types, the variance is looking quite good for the points of highest variance in the data (t_2 and t_3), but less so for the others. This is to be expected, given the way in which the basis functions were formed.

3.3.3 Eigenvector model variance

The variance over time (relative to the actual variance over time) for the eigenvector models of different orders for fitting the dataset with a noise variance of 1 are plotted as follows:

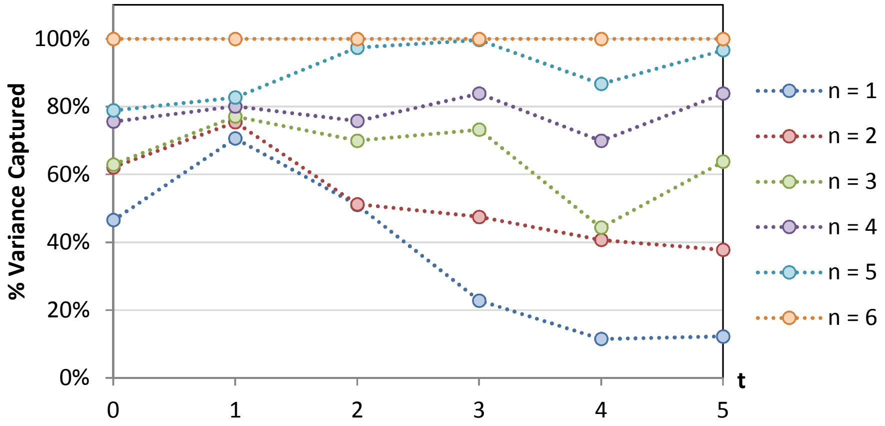


Figure 3.17 - Variance of the eigenvector model per time-point for dataset with noise σ2 = 1, as a percentage of actual variance

The higher orders of the eigenvector model seem to capture the variance reasonably well, particularly for the mid-range of the data. Furthermore, the eigenvector models of higher order would appear to be better suited to capturing the variance than the Legendre models of corresponding order, overall.

3.4 Percentage squared bias of models

The percentage squared bias (PSB) was calculated for the models in their fitting of the extended hundredfold dataset with noise from the N(0,1) normal distribution added.

3.4.1 Legendre model PSB

A simple plot of the PSB of Legendre models of different orders for the finite dataset time-points (with noise variance 1) gives a similar output to that of the (inverted) accuracy plot:

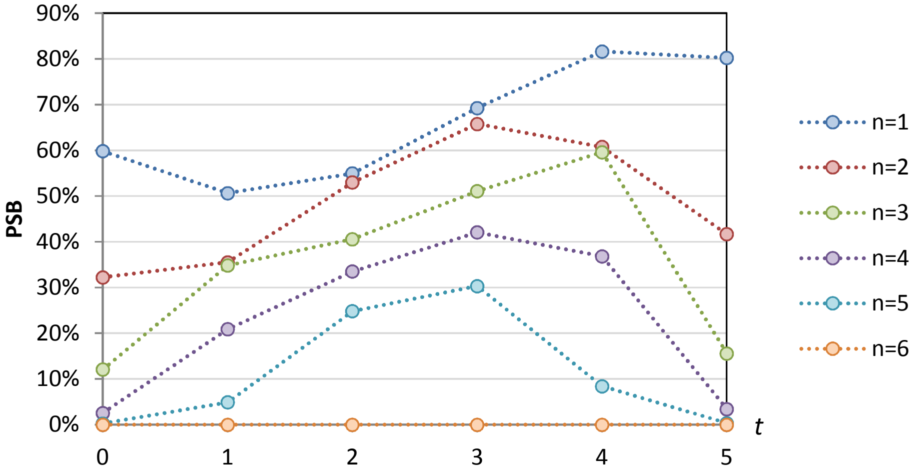


Figure 3.18 – PSB values of orders *n* of the Legendre model, per time-point of the data, for dataset with noise σ^2 = 1.

The same plot performed on the extended hundredfold dataset gives the following:

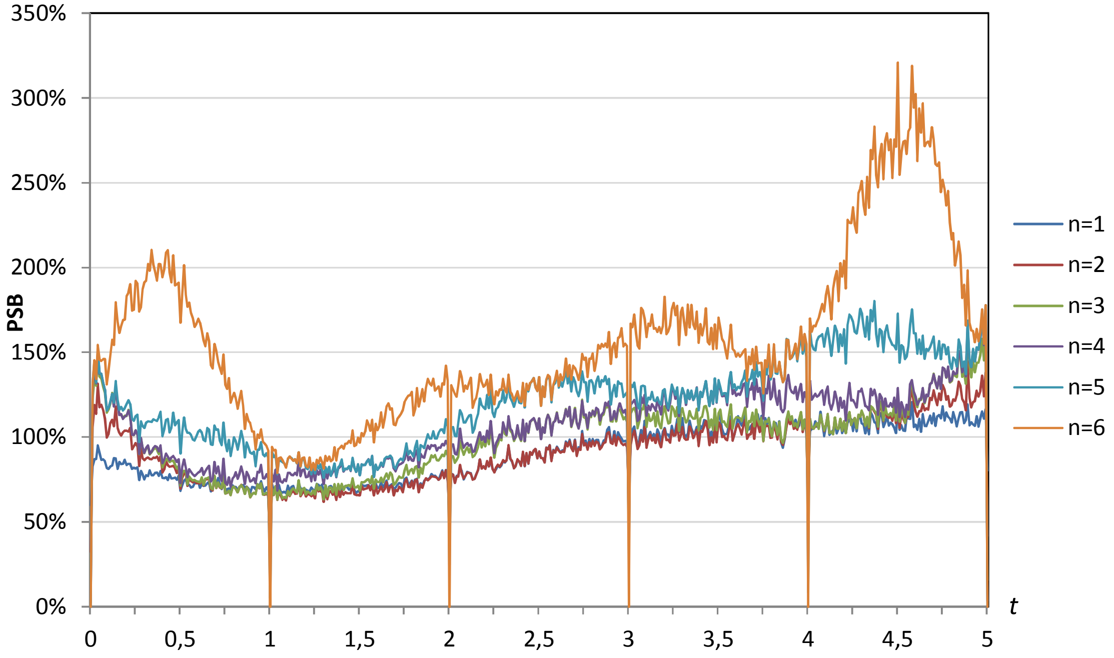


Figure 3.19 – PSB values of Legendre models of orders *n*, per time-point of the extended hundredfold dataset of noise σ^2=1

The plot indicates that while higher orders of the Legendre polynomials give better accuracies for the specific time-points of the finite dataset, the opposite is true for the intervals between those points.

The average PBS per order of Legendre polynomials for the four different extended hundredfold datasets of varying added noise was also plotted:

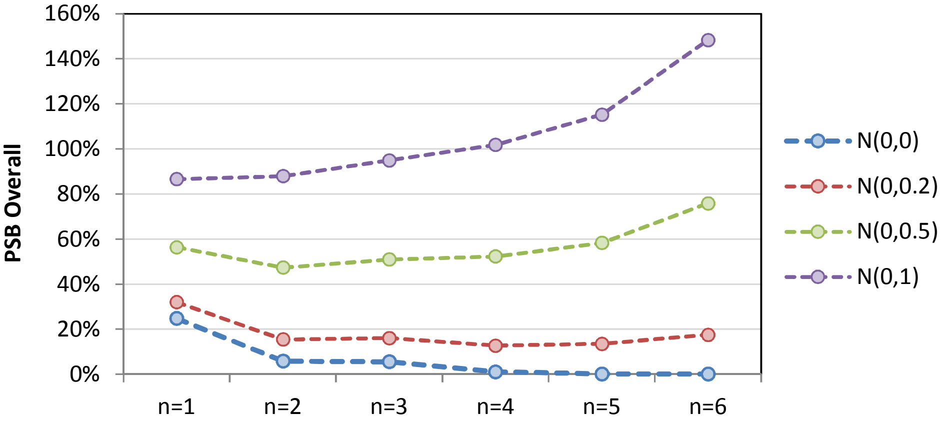
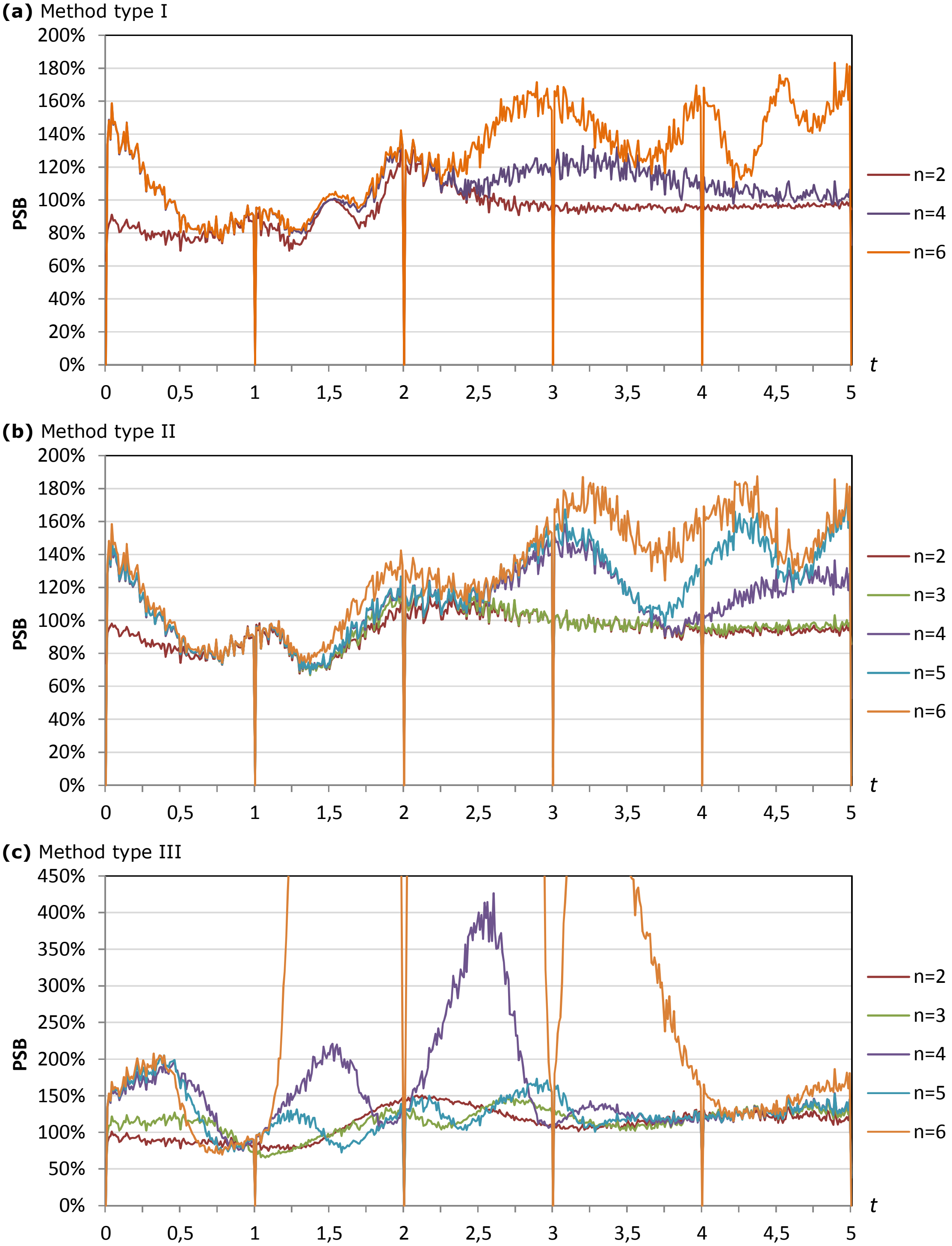


Figure 3.20 – Average PSB over the extended hundredfold datasets of differing noise variance, per order of the Legendre polynomials.

It can be seen that the overall PSB becomes pronounced for Legendre polynomial models of high order upon datasets containing a high noise component in their variance.

3.4.2 Inverse polynomial models PSB

The plots of the PSB values over the range of the extended hundredfold dataset of noise variance 1, for the inverse polynomials of types I, II and III of various orders, are given below:



(a) Method type I

(b) Method type II

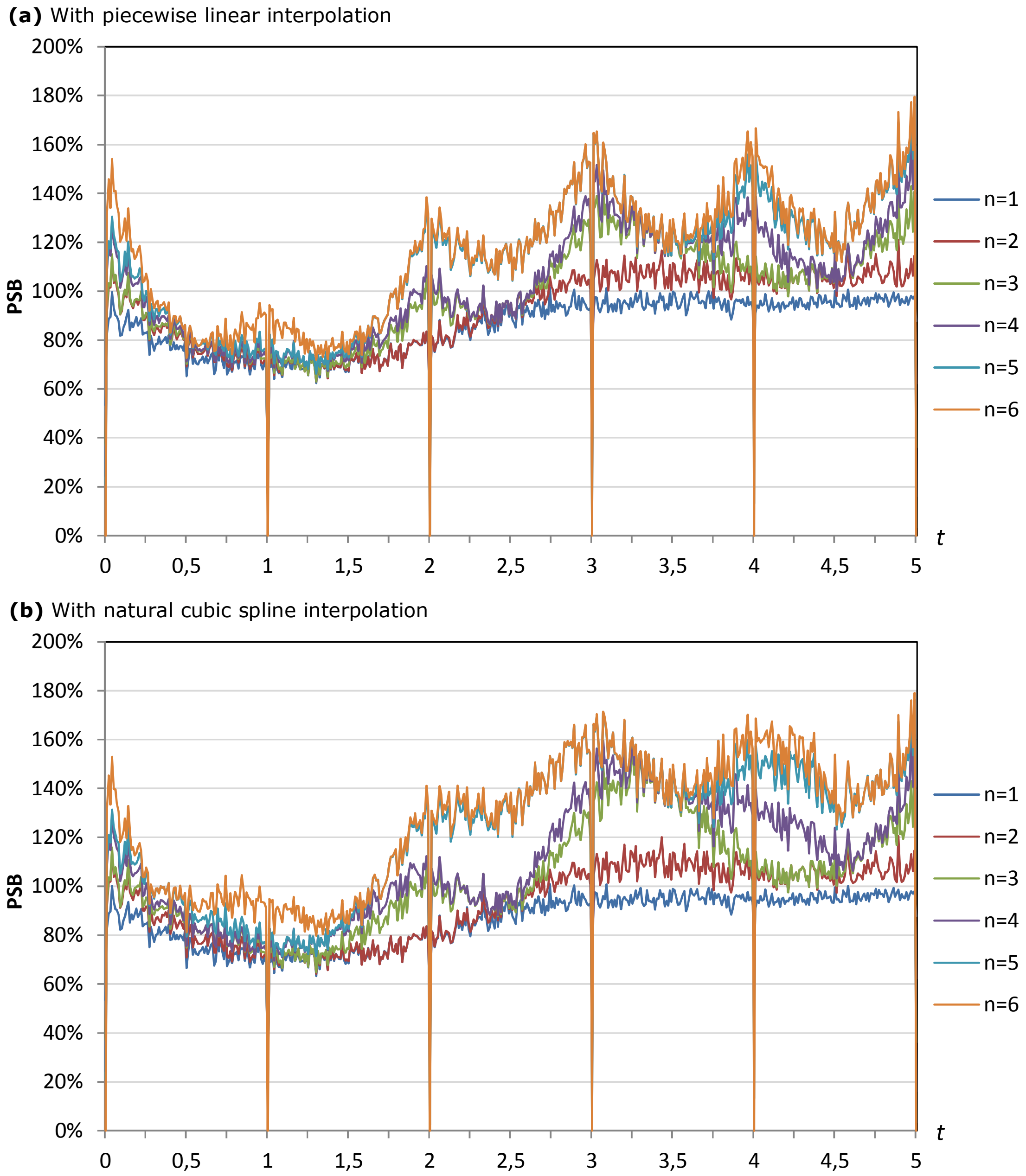
(c) Method type III

Figure 3.21 - PSB values of Inverse polynomial models type I, II and II, of orders *n*, per time-point of the extended hundredfold dataset of noise σ^2=1

Despite the poorer accuracies for the time-points of the finite dataset, the inverse polynomials of types I and II appear to be better behaved for higher orders than the Legendre polynomials are. The inverse polynomial of type III proves to be an entirely unsuitable method here, as it's PSB values for higher orders have increased massively.

3.4.3 Eigenvector model PSB

The plot of the PSB values over the range of the extended hundredfold dataset of noise variance 1 for the eigenvector model of various orders is given below. The evaluation of the non-dataset points was made by both a piecewise linear interpolation (a), and a natural cubic spline interpolation (b).


(a) With piecewise linear interpolation

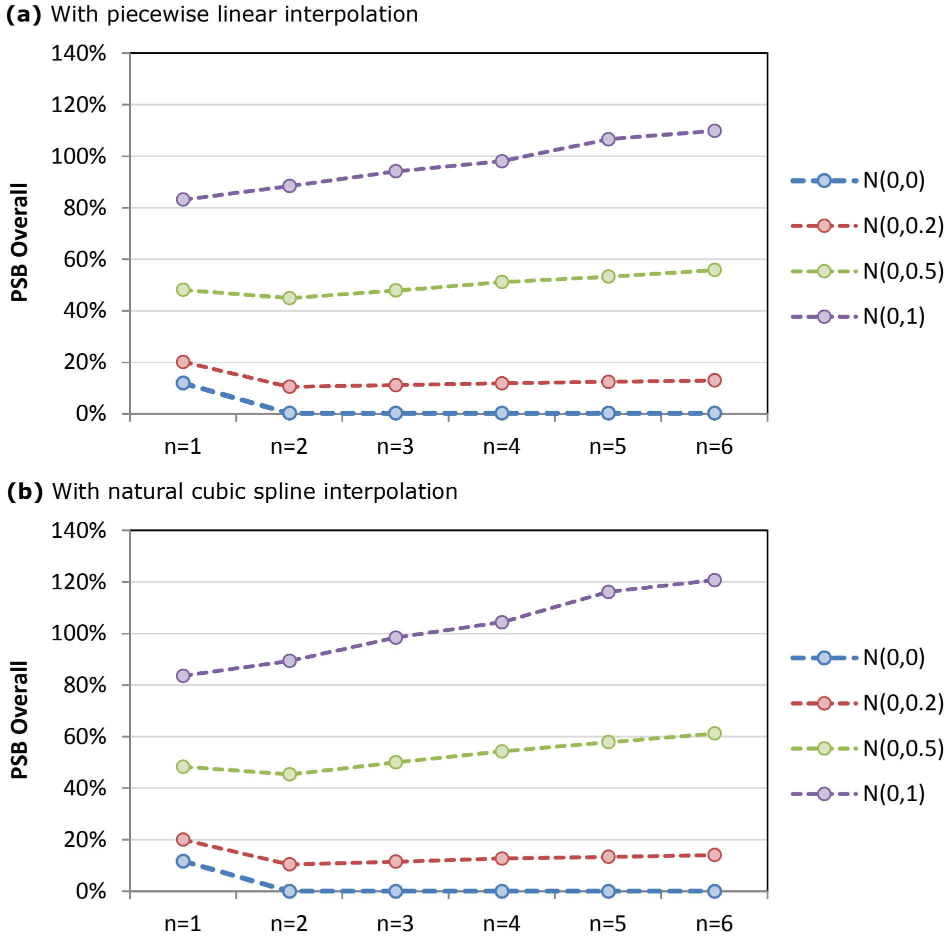
(b) With natural cubic spline interpolation

Figure 3.22 - PSB values of eigenvectors extended by (a) piecewise linear and (b) cubic spline interpolations, for various orders *n*, per time-point of the extended hundredfold dataset of noise σ^2=1

It appears that the eigenvector set extended by piecewise linear interpolation produced less biased estimates that the eigenvector set extended by natural cubic spline interpolation, particularly for the higher order models.

A striking comparison can be made with the Legendre PSB evaluated over the extended hundredfold dataset, for the higher order models. It can be seen that while the PSB generally increases with greater distance from a data-point of the finite dataset for the Legendre models, the PSB for the eigenvector model does the opposite, generally decreasing with greater distance from a data-point of the finite dataset.

The average PBS per order of eigenvector model for the four different extended hundredfold datasets of varying added noise was also plotted, for both eigenvectors extended with linear piecewise interpolation, and eigenvectors extended with natural cubic spline interpolation:



(a) With piecewise linear interpolation

(b) With natural cubic spline interpolation

Figure 3.23 - Average PSB over the extended hundredfold datasets of differing noise variance, per order of the eigenvector models, for eigenvectors extended by either the (a) piecewise linear, or (b) cubic spline interpolations.

For higher order models, the eigenvectors extended with piecewise linear interpolation fit the data with less bias than the eigenvectors extended with natural cubic splines. Moreover, while the PSB does increase for the higher order models, it does not do so in the way that the Legendre polynomials do.

3.5 F-tests

F-tests were performed upon the Legendre and eigenvector models, for data with an added noise variance of 1 only. Within a certain type of model, each pair of models of consecutive orders of n were compared via an F-test for two models of differing order (equation 32). Between the Legendre and eigenvector model types, for each order n model, the two model types were compared via an F-test for two models of equal order (equation 31). These F-tests were performed per time-point of the finite study data, and for the overall values across all those time-points; and per time-point of the extended hundredfold dataset, and for the overall values across all those time-points.

For the F-tests performed on each pair of models of consecutive orders of n for either the Legendre or eigenvector model types, all results were highly significant with p < 0.001, whether it was per each of the d = 6 finite time-points, or for all of these finite time-points together. This showed that each model of consecutively higher order gave significantly better estimates than its lower-order counterpart, for the finite dataset. When performing the same F-tests on the continuous data, almost all of the non-study data-points had significant p-values of p > 0.95. This showed that each model of consecutively higher order gave significantly worse estimates than its lower-order counterpart, for the continuous dataset.

For the F-tests performed on both the Legendre and eigenvector model types of equal order n, for the d = 6 finite time-points, the resultant F-statistics and p-values are given below in table 3.10. The p-values have been colour-coded, where the results showing the eigenvector model to be significantly better than the Legendre model (p < 0.05) are blue; whereas the results showing the Legendre model to be significantly better than the eigenvector model (p > 0.95) are red. The results for models of order n = 6 have been excluded, as the RSS of those models were almost infinitesimally small, and so the comparing the two model types for this order would have been meaningless.

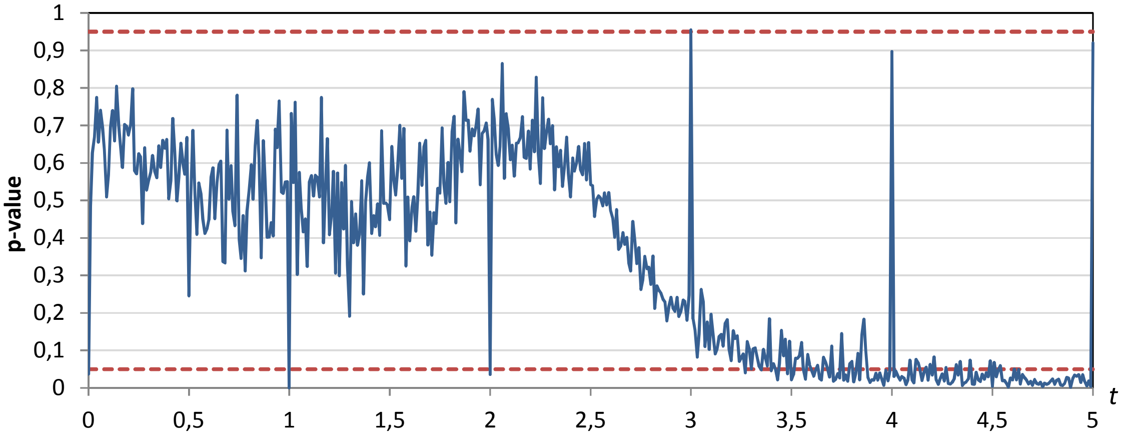
Table 3.10 – F-test results for the comparison of Legendre and eigenvector models of different orders

F-statistics:					
t	*n*=1	*n*=2	*n*=3	*n*=4	*n*=5
1	1.120	0.852	0.326	0.106	0.012
2	1.726	1.447	1.523	1.050	0.284
3	1.121	1.085	1.352	1.385	9.704
4	0.898	1.254	1.910	2.611	102.851
5	0.923	1.025	1.073	1.224	0.635
6	0.915	0.671	0.431	0.213	0.100
Overall	1.067	1.037	1.048	1.051	1.090
p-values:					
t	*n*=1	*n*=2	*n*=3	*n*=4	*n*=5
1	0.0362	0.9942	1	1	1
2	5.27×10^{-18}	2.91×10^{-9}	1.91×10^{-11}	0.2185	0.9999
3	0.0359	0.0985	1.02×10^{-6}	1.45×10^{-7}	2.09×10^{-236}
4	0.9548	0.0002	2.15×10^{-24}	2.73×10^{-50}	0
5	0.8971	0.3461	0.1333	0.0007	1
6	0.9200	1	1	1	1
Overall	0.0063	0.0782	0.0352	0.0273	0.0004

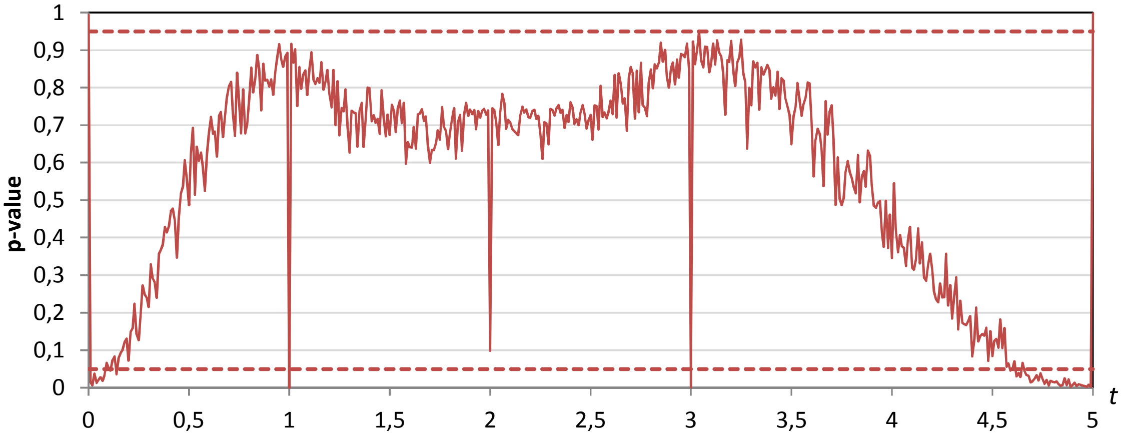
Apart from the models of order n = 1, these results confirm that the eigenvector model performs significantly better than the Legendre model within the mid-range of the data, whereas the Legendre model performs significantly better than the eigenvector model at the boundary of the data range, for fitting of the values of the d = 6 finite time-points.

For the F-tests performed on both the Legendre and eigenvector model types of equal order n, for the extended hundredfold dataset, the resultant p-values are given in the following series of figures:

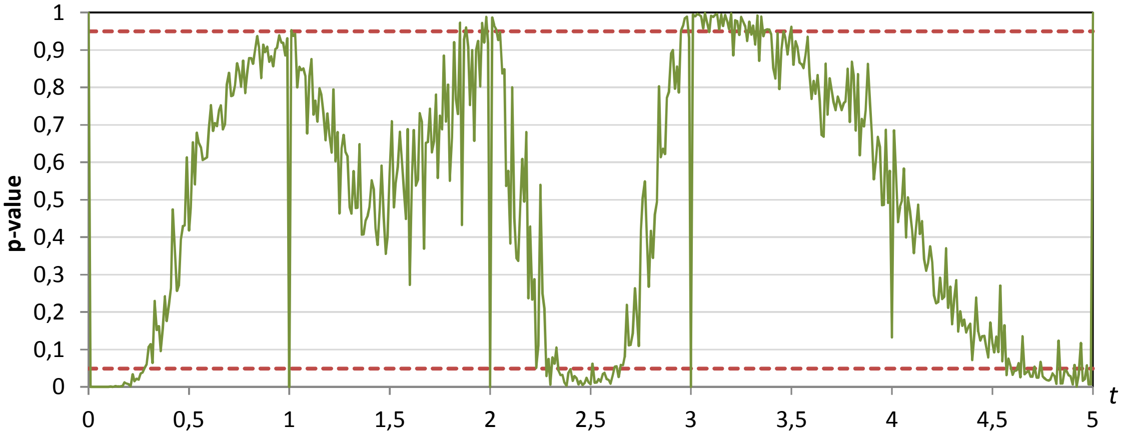
(a) *n* = 1



(b) *n* = 2



(c) *n* = 3



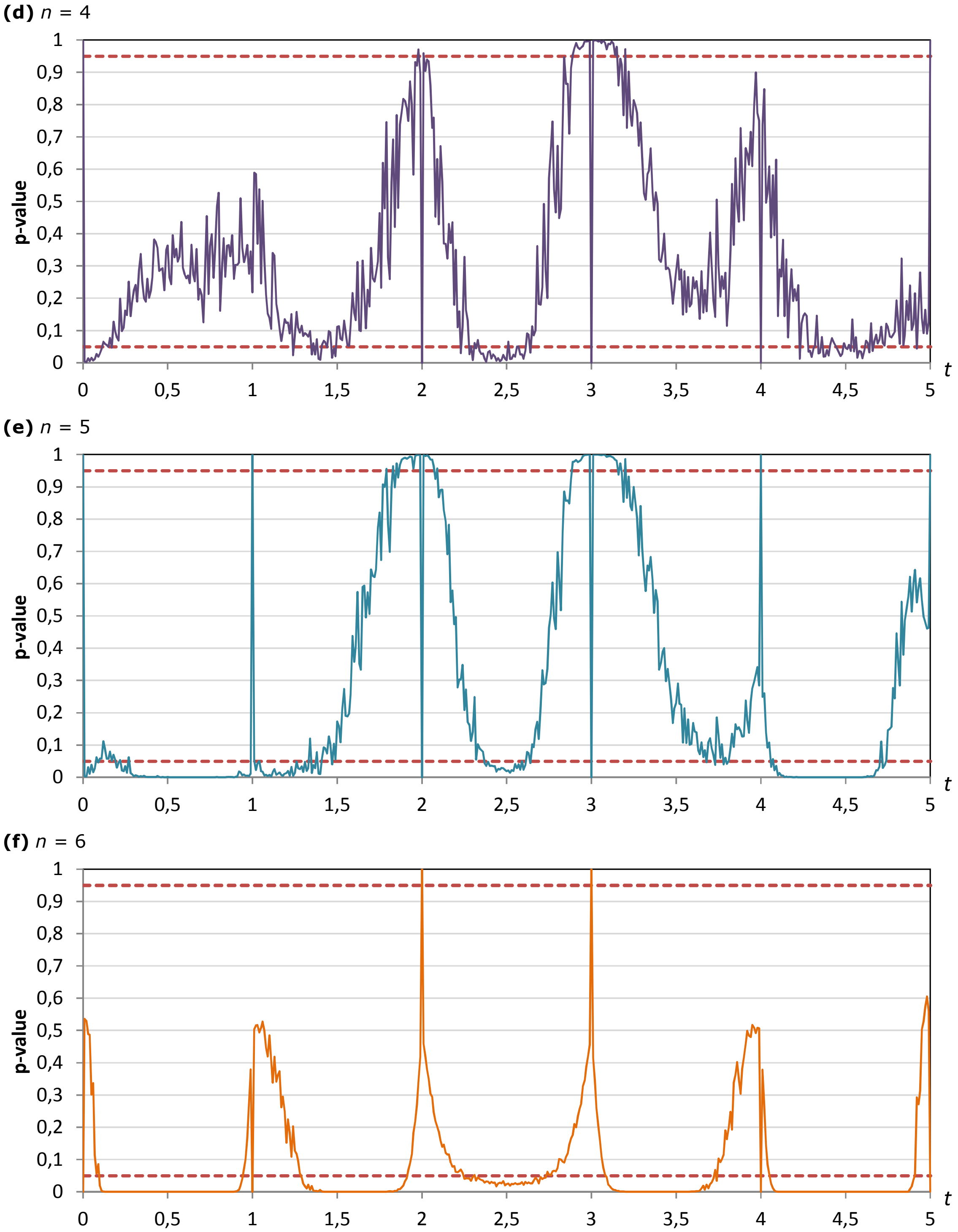


Figure 3.24 – p-values of the F-test of the Legendre and eigenvector models, for each of the six orders *n* of model, over the extended hundredfold dataset of noise variance 1.

Here, p-values of $p > 0.95$ indicate that the Legendre model was significantly better than the eigenvector model, whilst p-values of $p < 0.05$ indicate that the eigenvector model was significantly better than the Legendre model.

The results show that neither model has a consistent advantage over the other across the whole range of the continuous dataset. Additionally, while one model is generally significantly better than the other at a specific one of the d = 6 finite time-points, in the local region around such a time-point, the other model gives significantly better fitting than the former.

For higher orders of the models, the eigenvector model is generally better suited to fitting the data than the Legendre model, particularly for the mid-ranges of the intervals bounded by the consecutive d = 6 finite time-points.

4. Discussion

As this study was focussed on the development of a novel method for modelling longitudinal data, there is relatively little to mention in the way of current literature concerning it. Nonetheless, comparative studies on other longitudinal methods have been conducted, and can be considered insofar as they pertain to the methods compared in this study.

It was not only the use of eigenvectors themselves in modelling longitudinal data which was novel in nature, but their utilisation within a linear regression model also. A comparative study by Ali & Schaefer (1967) investigated the use of, amongst others, a regression model in modelling lactation yield data, and found that it out-performed the other two models for which they were investigating, namely, a gamma function and an inverse quadratic polynomial function. By combining a regression model with a longitudinal model in this study, the resultant model might be “greater than the sum of its parts”, and improve estimation of data further.

The primary objective in this study of assessing the new eigenvector model, particularly in comparison to other methods, would be primarily achieved in comparing it with a Legendre polynomial model. The Legendre polynomials are a well-established basis set for the purposes of data interpolation, having properties which make them a well-suited and popular tool to this end, and thus a worthy candidate for comparison with the eigenvector model.

A secondary objective of this study came to be the development of a modified inverse polynomial basis set, suitable for applying to general datasets. Although the results showed that it did not perform as well as the others, it nonetheless demonstrates some important aspects in the choice of basis sets, and in the type of dataset being fitted.

4.1 Obtaining basis sets

One of the first comparisons to be made is to do with the practical nature in which basis sets may be obtained for the purposes of regression. Generally there are two steps involved in this: determining the parameters of the basis functions to be used; and evaluating these functions for the time-points of the dataset, in order to obtain vectors for the regression.

The Legendre polynomial basis set, like any polynomial basis set, is relatively straight-forward to obtain. The polynomial coefficients can be derived from a recursion equation, obtaining the elements of a coefficient matrix. Similarly, the polynomial terms $\{1, t, t^2, \ldots\}$ can be evaluated for all data time-points, obtaining the elements of a data-specific term-matrix, although for the Legendre polynomials in particular, it is the transformed terms in t which will need to be evaluated by their various degrees. These two matrices can then be multiplied to form the basis matrix of vectors to be regressed.

The obtaining of the basis matrix for the modified inverse polynomials is more complicated, and somewhat less intuitive. Unlike the single transformation of the independent variable required by the Legendre polynomials, the modified inverse polynomials require a specific transformation per function, and the choice of specific t-variables upon which to base each transformation further complicates things. Fortunately there are only three terms to consider, the coefficients of which do not change.

The eigenvector basis matrix, on the other hand, is much simpler to obtain than the basis matrices of the other methods. The covariance matrix of the data, and the eigendecomposition of that covariance matrix, are both very straight-forward procedures, which have readily available pre-defined functions in standard statistical or mathematical software packages. Additionally, the lack of a need to evaluate functions for the time-points further simplifies the process.

4.2 Goodness of fit

The goodness of fit of the models can be divided into two categories: the goodness of fit on the finite set of time-points for which the dataset provides values (also referred to as the study time-points, or the interpolation points), and the goodness of fit on the continuous intervals between the aforementioned finite data-points (also referred to as the extended/hundredfold dataset).

4.2.1 The finite set of data-points

When considering models of mid- to high-order, the Legendre polynomial model had a clear advantage over the others at the boundaries of the range of the data. This could be seen in the results of the accuracies of the models, and of the variance of the models. Apart from the models of high order, neither the modified inverse polynomial model nor eigenvector model had comparatively good predictions for those boundary areas (although the predictions were not too bad either).

On the other hand, the Legendre polynomial model did not perform so well in the mid-range of the data, when compared with the eigenvector model at least.

When it came to consistency across the range of the data, the modified inverse polynomial models of all three types performed the worst. This can be expected due to the nature in which their basis functions were established; each function was specified to explain the variance at different areas of the data range, and so they would have been “competing” with each other in the regression, which led to the oscillatory nature of the goodness of fit of the predictions of these models.

The Legendre polynomial models were much more consistent, although for the higher orders they were not able to match for the mid-range of the data the relative goodness of fit obtained at the boundaries of the data range.

The eigenvectors fared better overall when it came to consistency of goodness of fit across the data range, but only for the higher model orders. The F-test data confirms this, as although the Legendre models outperformed the eigenvector models at the boundaries of the data range, the eigenvector models had significantly better predictions overall.

4.2.2 The continuous intervals between data-points

A different story was revealed for the intervals in-between the data-points of the finite study dataset. The presence of high percentage squared bias (PSB) of the range of approximately 60%-160% (for the data with known noise variance of 1) indicates that the noise variance in the data is having a significant effect on the predictions. In particular, it is likely that the regression model has treated the data as being *too* representative of the true “noiseless” data underlying the dataset. By fitting specific data (which contains noise) too closely, the resulting predictions are biased in favour of the noise present in the fitted data.

This can be seen in the higher order models, which, while they fit the actual set of finite data-points better with increasing order of model, the PSB becomes higher with increasing order of model also, for the prediction of data immediately beyond the finite study time-points. This can be seen in the F-test results for comparing two models over the extended dataset also: when one of the models gives significantly better estimates for the actual set of finite study data-points (that is, it is fitting that data, with its particular noise component, better), it is generally giving significantly worse estimates for the local surrounding "continuous" data.

Despite this issue with the noise component, the continuous data does give some further insight into the nature of the models. The Runge phenomenon (Runge, 1901) becomes increasingly apparent for Legendre polynomials of higher order, where the predictions oscillate between the interpolation points for polynomials of high order, and in particular for those intervals near the edges of the range of the data.

The eigenvector model, on the other hand, tends to correct for its "over-prediction" on the set of study data-points, when considering the central regions of the intervals bounded by those data-points.

Despite the eigenvector's capacity to only evaluate for a finite number of data-points (as opposed to the basis functions, which are continuous on a data range and can therefore be used to evaluate any point in time), the interpolation of the eigenvectors, and the piecewise linear interpolation in particular, has demonstrated that the eigenvector model when extended with this interpolation, not only can *compete* with the continuous functions in estimation of values in-between points of the dataset, but can indeed *outperform* these continuous functions.

It is worth considering if basis polynomial sets, such as the Legendre basis set, would likewise benefit from using a piecewise linear interpolation of a finite set of time-points evaluated by such a basis set. Certainly, a problem like the Runge phenomenon could be negated by such a method.

4.3 Dataset effect

This study dealt with quite a specific sort of simulated dataset. While some longitudinal models can be generally applied to any sort of data, others are tailored for data of a specific kind. For example, in the modelling of cattle lactation data, the lactation curve has a specific form of increase and decrease which some functions are well-suited to model.

Thus the shape of the curve underlying the data being modelled can influence the judgement of the model(s) under consideration. In this study, while the Legendre and eigenvector models have the capacity to be applied to a broad range of dataset-types, the modified inverse polynomial is probably more suited to a specific type of data.

Indeed, the inverse polynomial was chosen due in part to its horizontal asymptote, which it was hoped would be a useful feature of a basis set derived from it, when fitting data with flattened boundary values, such as one based upon a normal density distribution, as in our simulated dataset.

The trade-off between generality and specificity is something worth considering when assessing a model.

4.4 Potential improvements

A major issue in this study was that of an over-accounting of noise in the data. This was likely due to the choice of regression model, and in particular, of having a fixed regression coefficient rather than a random one. By using a random regression model (Schaffer and Dekkers, 1994), the calculation of regression coefficients could adequately allow for random variation in the particular combination of basis functions per individual, thereby taking the random noise of the data into account.

4.5 Further lines of inquiry

It has already been mentioned that piecewise linear interpolation of evaluated Legendre basis polynomials could help eliminate the Runge phenomenon, and that the use of a random regression model could better take the data noise into account.

It could be worthwhile to perform a heritability study on the regression coefficients of an eigenvector model, the cases where the "trait-space" has been significantly reduced in dimensionality to that of the new "beta-space" of regression coefficients, to see how the heritabilities of the new β-parameters compare to those of the original traits, and thus if an eigenvector model is capable of producing a meaningful re-parameterisation of traits in such a way as to be of use in breeding selection.

Although the modified inverse polynomial model proved to be ineffective for the fitting of data in this study, perhaps being better suited to data of a more specific type, the method(s) for constructing its basis functions may yet be used for other fixed-origin functions, such as the gamma function or Gaussian (bell-curve) function.

Finally, the use of a non-simulated dataset would be good to see how well an eigenvector model can handle real data.

5. Conclusion

The use of an eigenvector basis set within a regression model for the interpolation and fitting of longitudinal data has good potential to be a competitive model with other well-established models currently in use, such as the Legendre polynomial regression model.

Additionally, a piecewise linear interpolation of the eigenvectors will allow for the eigenvector model to be applied to longitudinal points outside of the finite set of points from which it is generated, again giving it a competitive opportunity against models based upon continuous functions.

Further investigation of the eigenvector model is yet required, preferably within a random regression model, in order to better establish its validity in comparison to other models.

6. References

Ali, T. E. and L. R. Schaeffer (1987). *Accounting for covariances among test day milk yields in dairy cows*. Can. J. Anim. Sci. 67: 637-644.

Genz, Alan and Frank Bretz (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics, Vol. 195., Springer-Verlage, Heidelberg. ISBN 978-3-642-01688-2

Interbull (2000). *National Genetic Evaluation Programmes for Dairy Production Traits Practised in Interbull Member Countries 1999-2000*. Department of Animal Breeding and Genetics, Uppsala, Sweden, Bulletin 24.

Macciotta, N.P.P., D. Vicario, and A. Cappio-Borlino (2005). *Detection of Different Shapes of Lactation Curve for Milk Yield in Dairy Cattle by Empirical Mathematical Models*. Journal of Dairy Science, Volume 88, Issue 3, March 2005, Pages 1178–1191

Nelder , J. A. (1966). *Inverse Polynomials, a Useful Group of Multi-Factor Response Functions.* Biometrics, Vol. 22, No. 1 (Mar., 1966), pp. 128-141

Runge, Carl (1901). *Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten*. Zeitschrift für Mathematik und Physik 46: 224–243.

Schaeffer, L.R. and J.C.M. Dekkers (1994). *Random regression in animal models for test-day production in dairy cattle*. Proc. 5th World Congr. Genet. Appl. Livest. Prod., Guelph, ON, XVIII (1994), pp. 443–446

Venables, W. N. & B.D. Ripley (2002). *Modern Applied Statistics with S. Fourth Edition*. Springer, New York. ISBN 0-387-95457-0

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no