1-Dimensional Splines as Building Blocks for Improving Accuracy of Risk Outcomes Models

David S. Vogel
MEDai, Inc.
University of Central Florida
Orlando, FL
dvogel1@cfl.rr.com

Morgan C. Wang
Department of Statistics & Actuarial Science
University of Central Florida
Orlando, FL
cwang@mail.ucf.edu

ABSTRACT

Transformation of both the response variable and the predictors is commonly used in fitting regression models. However, these transformation methods do not always provide the maximum linear correlation between the response variable and the predictors, especially when there are non-linear relationships between predictors and the response such as the medical data set used in this study. A spline based transformation method is proposed that is second order smooth, continuous, and minimizes the mean squared error between the response and each predictor. Since the computation time for generating this spline is O(n), the processing time is reasonable with massive data sets. In contrast to cubic smoothing splines, the resulting transformation equations also display a high level of efficiency for scoring. Data used for predicting health outcomes contains an abundance of non-linear relationships between predictors and the outcomes requiring an algorithm for modeling them accurately. Thus, a transformation that fits an adaptive cubic spline to each of a set of variables is proposed. These curves are used as a set of transformation functions on the predictors. A case study of how the transformed variables can be fed into a simple linear regression model to predict risk outcomes is presented. The results show significant improvement over the performance of the original variables in both linear and non-linear models.

Categories and Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning

Keywords

spline, variable transformation, linear model, data mining, prediction, adaptive, risk, outcomes.

1. INTRODUCTION

Transformation methods (Draper and Smith, 1981) are commonly used in regression analysis. Most of these transformation methods are very simple because it is not only of interest in the prediction, but also in observing the relationship between predictors and the response. Although there are transformation methods such as those proposed by Breiman and Friedman (1985) that can maximize the correlation between variables, this method is not suitable when there are piecewise relationships. A transformation that can maximize the linearity between predictors and the response is proposed. The proposed method works well even if there are piecewise relationships. In data mining context, this method is very useful because the focus is on maximizing the predictive value (Hastie et al., 2001) while limited in time spent on understanding this relationship between variables.

Many different non-linear modeling techniques claim to describe relationships in non-linear variables. However, due to the large number of non-linear patterns to investigate, even if sufficient computational power is available, the probability of finding false patterns is increased. Rather, analysis and transformation of one variable at a time to maximize its linear relationship to the dependent variable maintains low complexity of the solution, if done carefully. While this algorithm does not identify interactive relationships in the data, it does not take away from the ability to model interactions either. The gain in using a piecewise algorithm versus combinations of functions on the entire range of values is that the complexity of the model can be kept low, while effectively adapting to different behavior at different ranges within a predictor.

While many spline algorithms (Boucheta and Djeddi, 1994; Lindstorm, 1999; Prvan, 2000; Luo and Wahba, 1997; and Jupp, 1978) exist, a smooth series of third order polynomials was chosen. Because of the strange distributions of real-world data, we will show that polynomial curves cannot always fit the relationship between variables accurately. Cubic smoothing splines may be a good fit, but it is often the case that one part of the curve works well with one smoothing constant, and the other side of the curve works well with a different smoothing constant.

The data is an 115,000 member subset of a repository of various health plans stored at Medical Artificial Intelligence, Inc. (MEDai). The sample is constrained to members enrolled for the full period being evaluated and the variables available to be used as predictors are based on ETGs (Episode Treatment Groups). The data set is one used to project future annual cost based on previous charges within 21 ETG classes of drugs being

taken by those members. The classes of drugs include Arthritis, Asthma, Bronchitis, Breast Neoplasm, CHF (Congestive Heart Failure), CNS (Central Nervous System), CV medications, Degenerative Diseases, Dermatology, Diabetes, Fractures, GI, GU, Hypertension, Metabolic, Pneumonia, Psychosis, Renal Failure, Skin Inflammation, Tonsillitis and Trauma.

The relationship between these drug costs and future healthcare cost generally is extremely non-linear, and has a high number of outliers. These relationships will be modeled individually using the adaptive spline, transformed, and finally fed into a multivariate model for evaluation.

While multi-dimensional splines are theoretically obtainable, it has been found that these over-fit very quickly as the number of variables increases (Fan and Gijbels, 2000). In order to create a generalizable model, we chose to smooth the data in one dimension at a time, and feed the transformed variables into a global linear model.

Transformation of the response variables are not considered (Draper and Smith, 1981) because of the large residuals that those transformations create in modeling these kinds of data sets with a high level of noise. In fact, previous tests with similar data sets where the response variable under-went a transformation yielded R-Squared values less than zero because the bias is so great.

2. SPLINE SELECTION

The proposed choice of splines is one that is flexible enough to adapt to practically any non-linear relationship, but one that is smooth and has a minimal number of degrees of freedom to prevent over-fitting. Given a set of knots (x-coordinates specified only), we construct the set of cubic curves beginning and ending at the knots that collectively minimize the Mean Squared Error. To keep the degrees of freedom at a minimum, we require 1st and 2nd order smoothness at the knots. No boundary conditions are imposed at the end-points, leaving the number of degrees of freedom to be N+3, where N is the number of segments used for constructing the curve.

2.1 Derivation of the Spline Coefficients

Given the set of x-coordinates

$$x_0, x_1 \ldots x_N$$

construct the N cubic equations with end-points at these x-coordinates minimizing the sum of the squares of the errors, while meeting the following 3N-3 smoothing conditions (matching end-points, first, and second derivatives at nodes):

For i = 1 to N-1,

$$a_{i+1}x_i^{\,3} + b_{i+1}x_i^{\,2} + c_{i+1}x_i + d_{i+1} = a_i x_i^{\,3} + b_i x_i^{\,2} + c_i x_i + d_i$$
$$3a_{i+1}x_i^{\,2} + 2b_{i+1}x_i + c_{i+1} = 3a_i x_i^{\,2} + 2b_i x_i + c_i$$
$$6a_{i+1}x_i + 2b_{i+1} = 6a_i x_i + 2b_i$$

Define the N equations, I = 1 to N:

$$\hat{y}_i = a_i x_i^{\,3} + b_i x_i^{\,2} + c_i x_i + d_i$$

Define the Error of the aggregate curve:

$$E = \sum (\hat{y} - y)^2 = \sum_{i=1}^{N} \sum_i (\hat{y} - y)^2$$

There are 4N coefficients that need to be solved for, but the 3N-3 smoothing conditions leave only N+3 degrees of freedom on which the optimization will be performed. If the following unknowns can be solved for

(1) $$\left[b_1, c_1, d_1, a_{1..N}\right]$$

the remaining 3N-3 coefficients can be written using the following algorithm:

For i = 2 to N:

(2)

$$b_i = 3x_{i-1}a_{i-1} + b_{i-1} - 3x_{i-1}a_i$$
$$c_i = 3x_{i-1}^{\,2}a_{i-1} + 2x_{i-1}b_{i-1} + c_{i-1} - 3x_{i-1}^{\,2}a_i - 2x_{i-1}b_i$$
$$d_i = x_{i-1}^{\,3}(a_{i-1} - a_i) + x_{i-1}^{\,2}(b_{i-1} - b_i) + x_{i-1}(c_{i-1} - c_i) + d_{i-1}$$

Now set the N+3 partial derivatives of the Error Function equal to zero. Also, we may refer to the variables in (1) as

$$[v] = \left[v_1, v_2 \ldots v_{N+3}\right]$$

Setting the partial derivatives to zero yields the following N+3 equations. For each i from 1 to N+3

$$0 = \frac{\partial E}{\partial v_i} = 2\sum_{j=1}^{N}\sum_j (\hat{y}_j - y)\frac{\partial \hat{y}_j}{\partial v_i}$$

(3)

$$\sum_{j=1}^{N}\sum_j \hat{y}_j \frac{\partial \hat{y}_j}{\partial v_i} = \sum_{j=1}^{N}\sum_j y \frac{\partial \hat{y}_j}{\partial v_i}$$

The partial derivatives of the N individual cubic pieces can be denoted by

$$\frac{\partial \hat{y}_j}{\partial v_i} = \frac{\partial a_j}{\partial v_i}x^3 + \frac{\partial b_j}{\partial v_i}x^2 + \frac{\partial c_j}{\partial v_i}x + \frac{\partial d_j}{\partial v_i}$$

The notation is changed to accommodate the fact that the 4N coefficients are not known, but are linear functions of the N+3 unknowns. a_j, b_j, c_j, and d_j will now be notated as the numeric vectors $[a_j]$, $[b_j]$, $[c_j]$, and $[d_j]$ all with dimension N+3. This reduces the equation (3) to a system of linear equations [A][v]=[b]. Multiplying out the terms within the summations so that both sides of the equation are functions of x, we get:

[A]=

$$\sum_{j=1}^{N}\left\{\begin{array}{l}
[a_j]\left(\frac{\partial a_j}{\partial v_i}\sum_j x^6 + \frac{\partial b_j}{\partial v_i}\sum_j x^5 + \frac{\partial c_j}{\partial v_i}\sum_j x^4 + \frac{\partial d_j}{\partial v_i}\sum_j x^3\right) \\
+[b_j]\left(\frac{\partial a_j}{\partial v_i}\sum_j x^5 + \frac{\partial b_j}{\partial v_i}\sum_j x^4 + \frac{\partial c_j}{\partial v_i}\sum_j x^3 + \frac{\partial d_j}{\partial v_i}\sum_j x^2\right) \\
+[c_j]\left(\frac{\partial a_j}{\partial v_i}\sum_j x^4 + \frac{\partial b_j}{\partial v_i}\sum_j x^3 + \frac{\partial c_j}{\partial v_i}\sum_j x^2 + \frac{\partial d_j}{\partial v_i}\sum_j x\right) \\
+[d_j]\left(\frac{\partial a_j}{\partial v_i}\sum_j x^3 + \frac{\partial b_j}{\partial v_i}\sum_j x^2 + \frac{\partial c_j}{\partial v_i}\sum_j x + \frac{\partial d_j}{\partial v_i}N_j\right)
\end{array}\right\}$$

[b]=

$$\sum_{j=1}^{N}\left\{\frac{\partial a_j}{\partial v_i}\sum_j x^3 y + \frac{\partial b_j}{\partial v_i}\sum_j x^2 y + \frac{\partial c_j}{\partial v_i}\sum_j xy + \frac{\partial d_j}{\partial v_i}\sum_j y\right\}$$

Any matrix solver will yield the N+3 unknown values of [v], and the remaining 3N-3 coefficients can be calculated by substituting the N+3 values into equation (2). The 4N coefficients have now been calculated.

2.2 Resulting Spline Curves

For the result presented in this paper, an automated procedure is used, recursively finding the best location for an additional knot. A significance test is performed comparing the change in accuracy before and after the proposed addition of each knot. If the p-value is greater than 0.01, the knot is not added and the iterations are stopped. This procedure finds that it is best to concentrate most of the knots where the curve is most non-linear and dense with data points. With the drug cost variables, it is clearly the left side of the curve. On hypertension drug cost, knots were placed at x={0,2,5,15,30,50,100}. The knot at x=0 would normally be unnecessary because it is at the end-point, but this particular data set had 5-10 missing values per variable (out of over 300,000), and they were assigned a value of –1. Having a knot there prevented those few data points from having undue influence the rest of the curve.

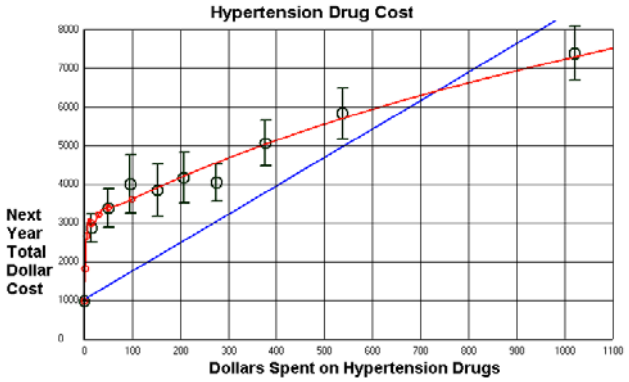

Figure 3: Modeling hypertension drugs versus next year dollar cost. Spline is shown in red. Regression line is shown in blue.

Similarly, knots were placed at x={0,2,5,10,20,60,200} for arthritis drug cost.

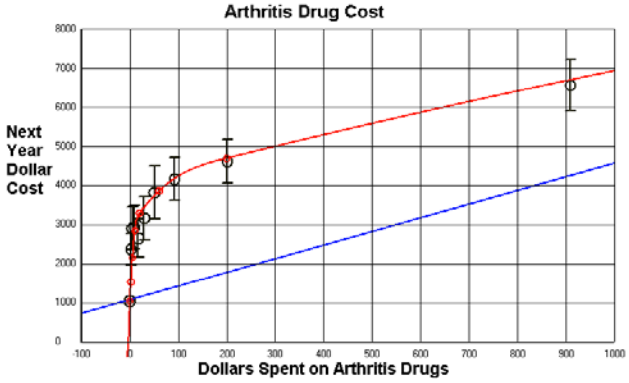

Figure 4: Modeling arthritis drugs versus next year dollar cost. Spline is shown in red. Regression line is shown in blue.

The curves appear visually to follow the non-linear trend of the relationship between the independent variables and the response variable. It is easier to see in the hypertension graph that the smoothness restrictions prevent the third order segments of the spline from over-fitting to some of the bobbles in the relationship that appear random. The 95% confidence intervals at those points show that the bobbles are not statistically significant, and that the spline correctly maintains the relationship of lowest complexity that truly follows the trend of the data.

Attempts were be made to fit these relationships with polynomial curves. The resulting curves are shown in figures 5 and 6.

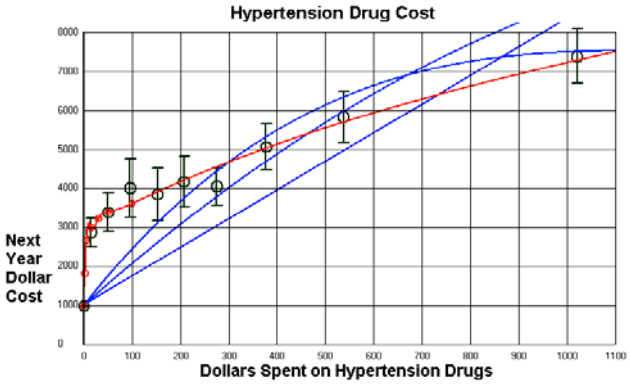

Figure 5: Modeling hypertension drugs versus next year dollar cost. Similar to figure 3 plus 2nd and 3rd order polynomial curves shown in blue.

On hypertension drugs, the 3rd order polynomial fits the data better than the 2nd order polynomial, which in turn fits the data better than the linear regression. However, it is apparent that the inherent relationship of the variable is not that of a polynomial. Of course a polynomial with a high enough degree will fit the data perfectly, but over-fitting becomes a concern with this.

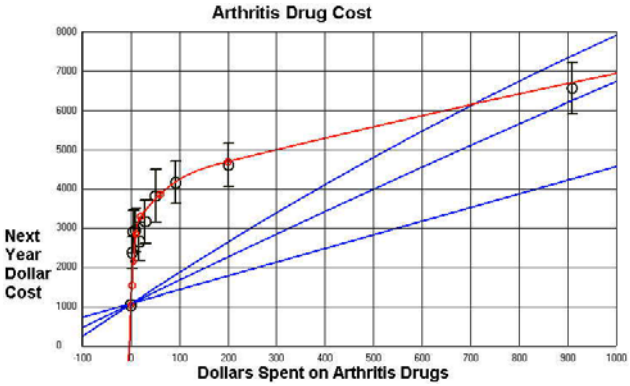

Figure 6: Modeling arthritis drugs versus next year dollar cost. Similar to figure 4 plus 2nd and 3rd order polynomial curves shown in blue.

On arthritis drug cost, the polynomial is even less of a decent fit for this non-linear relationship. The curves tend to accommodate the few outliers even more. More specifically, 2,952 members (2.6%) of the 115,000 take arthritis drugs. Out of those 2,952, only 319 have an arthritis drug cost of greater than $500. It is apparent from the graph that the polynomial focuses on less than 10% of the data and is completely inaccurate on the 90% of arthritis members with drug costs less than $500. The spline is much more flexible, and is given more knots at the lower costs so that all the members can have their risk assessed accurately.

3. TRANSFORMATION

Splines are generally good only for small dimensional problems as their complexity grows geometrically for each additional dimension. To avoid over-fitting in this high dimensional problem, we keep the complexity as low as possible by transforming each of the variables individually to its one-dimensional spline curve. The end goal is that the new variables will perform well in a linear regression. Therefore, if follows intuitively that the best transformation would likely create a new variable with a linear relationship to the dependent variable. Using the splines for hypertension drug cost and arthritis drug cost, we create new variables called "Hypertension Spline Estimate" and "Arthritis Spline Estimate."

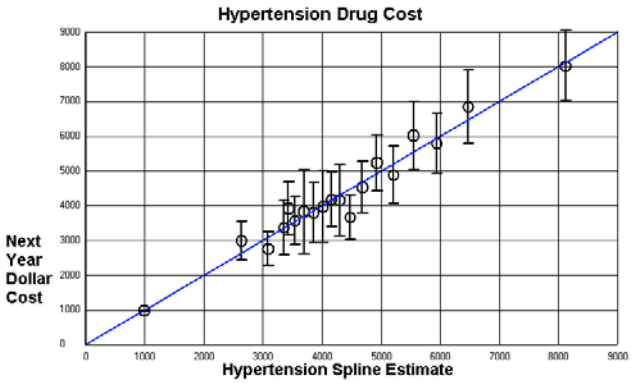

Figure 7: Transformed hypertension drug variable versus next year dollar cost. The new variable is linear with respect to the next year cost.

Graphing the transformed variable against the dependent variable, it is clear that the new relationship is now linear, as the grouped means match up very well with the regression line. The correlation of the original hypertension drug cost variable to the dependent variable is 0.1603. The correlation of the new variable to the dependent variable is 0.1758.

While this improvement is only 10%, a more substantial improvement can be observed in the arthritis drug cost transformation, as that variable is much more non-linear to begin with.

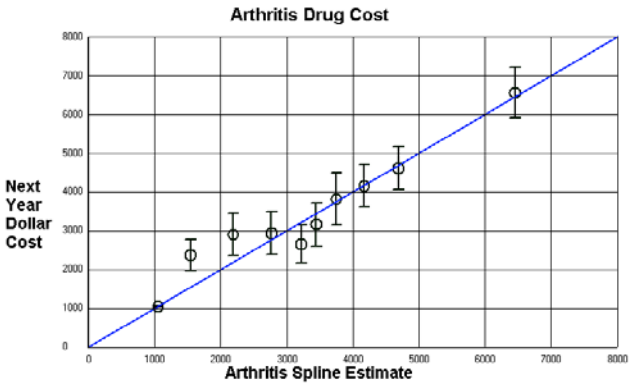

Figure 8: Transformed arthritis drug variable versus next year dollar cost. The new variable is linear with respect to the next year cost.

The spline fit is not perfect here, and could have benefited from more knot placement on the low end. Nevertheless, the fit is still "good" as the regression line passes through or is close to all the grouped means. The correlation of the original arthritis drug cost variable to the dependent variable is only 0.0847. The correlation of the new spline estimate variable to the dependent variable 0.1142, an increase of 35%.

The transformation is applied to all 21 independent variables, and the 7 variables most improved in correlation are recorded in the table 1.

Table 1: Improvements in correlation between original and transformed predictors to the dependent variable.

ETG	Original Correlation	Spline transformed correlation	Percent Increase
Arthritis	.0847	.1142	35%
CNS	.0658	.0881	34%
Degenerative	.1019	.1425	40%
Renal Failure	.1870	.2547	36%
Skin Inflammation	.0449	.0590	31%
Tonsillitis	.0166	.0217	31%
Trauma	.0440	.0616	40%

The average percent increase in correlation of all 21 variables is 22%. The range of improvement is from 5% to 40% as the most non-linear variables are going to be altered more drastically in the transformation than will the variables with a fairly linear relationship to begin with.

4. RESULTS

It is known that the improvement of correlation of individual variables does not necessarily improve the accuracy of the aggregate model, so it is necessary to combine the transformed variables into a model and evaluate the change in accuracy. Five models were calculated and evaluated for comparison:

1) Linear regression on the original variables.
2) 2nd degree polynomial regression on the original variables.
3) 3rd degree polynomial regression on the original variables.
4) Linear regression on binary versions of the original variables.
5) Linear regression on the new variables created from applying the spline.

The linear regression on the original variables is important to include, because it serves as the baseline. It allows us to see the results of a straight linear model with no transformations.

The second and third models are polynomial regression models, which give an idea of how a simple non-linear model performs in comparison with the spline transformation model. It also gives an idea of how quickly over-fitting can occur, even with a data set this large.

The reason for including the fourth model is that it is common within the medical industry to use binary versions of ETGs for predictive models. For example, diabetes charge greater than zero indicates that the member has diabetes, and would be assigned a value of 1. A model with the binary variables gives an idea of what is the industry standard that we are comparing to.

In the evaluation of these 5 models, we use two statistics:

1) R-Squared
2) Sensitivity of the top 2% of predictions (where PPV = 2% as well). The "2%" evaluation is not chosen arbitrarily, as it seems to be the industry consensus on the number of members in a health plan that can be focused on for preventative measures.

Table 2: Training Results

	R-Squared	Sensitivity
Regression	0.1463	23.86
2nd Degree	0.1631	24.64
3rd Degree	0.1795	25.29
Binary	0.1498	23.01
Spline	0.1860	26.40

Table 3: Validation Results

	R-Squared	Sensitivity
Regression	0.1341	23.73
2nd Degree	0.1470	25.03
3rd Degree	0.1440	25.16
Binary	0.1500	21.63
Spline	0.1838	26.34

Table 4: Differential between training and validation results

	Out of Sample R-Squared Delta
Regression	-8%
2nd Degree	-10%
3rd Degree	-20%
Binary	0%
Spline	-1%

Table 5: Improvement made by using the spline variables over each of the other techniques (based on validation results).

Percent Improved	R-Squared	Sensitivity
Regression	37%	11%
2nd Degree	25%	5%
3rd Degree	28%	5%
Binary	23%	22%

The tables are based on a 67-33 split of the data set of 115,018 members. 76,679 members were placed in the training set and 38,339 members were placed in the validation set so that the training and validation results could be compared.

On the validation set, there was a small improvement from the linear model to the 2nd degree polynomial, but the R-Squared value started to decline when the 3rd degree polynomial was applied. It should also be noted that percent decline between the training and validation R-Squared increased dramatically for the 3rd degree polynomial, indicating that the complexity of the model was over-fitting on even this large data set. The accuracy on the linear regression dropped 8% from .1463 to .1341, a notable amount, but likely not enough to characterize the model as "over-fit". The accuracy on the 3rd degree polynomial dropped 20% from the training set to the validation set. The 2nd degree polynomial showed a 10% decline in R-Squared as it was applied on the validation sample, but yielded the best overall R-Squared value of the polynomial models.

The model using the spline transformed variables performed the best in every area. Although a simple linear regression was used as the final model, the variables had under-gone a low-complexity transformation that customized them to work well in a linear model.

5. ACKNOWLEDGMENTS

We acknowledge A.I. Insight, Inc. and MEDai, Inc. for the use of their predictive modeling technology, MITCH (Multiple Intelligent Tasking Computer Heuristics), as well as the use of their repository of medical data.

6. REFERENCES

[1] Boucheta, R., Djeddi, M. (1994). Smoothing spline: local adaptive smoothing parameter. ADV MODELL ANAL A GEN MATH COMPUT TOOLS, vol. 20, no. 2-4, pp. 55-64.

[2] Breiman, L. and Friedman, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation", Journal of the American Statistical Association, Vol. 80, No. 391, pp. 580-598.

[3] Draper, N. R. and Smith, H. (1981). Applied Regression Analysis, Second Edition. John Wiley & Sons. Pp. 221-241.

[4] Fan, J. and Gijbels, I. (2000). Local polynomial fitting , 228-275. In Smoothing and Regression.

[5] Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Sprienger, 127-133

[6] Jupp, D. (1978), "Approximation to data by splines with free knots", SIAM Journal of Numerical Analysis, 15, pp. 328-343.

[7] Lindstrom, M. (1999), "Penalized estimation of free-knot splines", Journal of Computational and Graphical Statistics, 8, 2, pp. 333-352

[8] Luo, Z., and Wahba, G. (1997), "Hybrid adaptive splines", Journal of the American Statistical Association, 92, pp. 107-115.

[9] Prvan, T. (2000), "Least squares splines with variable knots using a smooth spline basis", ANZIAM J. 42 (E) ppC1199-C1217.