Bayesian Penalized Splines in Semi-Parametric Modeling

Naomi S. Altman
Penn State University
and
Cornell University

naomi_altman@stat.psu.edu

Abstract

Penalized regression splines provide a useful tool for fitting complicated models with smooth components. Because they are sieve estimators, parametric tools such as likelihood and information criteria can be used for fitting. In this paper, I will demonstrate applications of Bayesian Penalized Splines to self-modeling regression and to varying coefficient models. We will also see that some statistics that are pivotal for the fixed knots (parametric) case do not appear to be pivotal for the sieve estimator. Also, knot placement, which has been shown to be of minimal importance for univariate smoothing, can have a large effect in more complicated settings.

Keywords: nonparametric regression; smoothing; varying coefficient models; self-modeling regression; longitudinal models; growth curves; p-spline;

1, Introduction

In this paper we discuss the use of (empirical) Bayesian penalized splines for semi-parametric modeling. Since penalized regression splines are parametric for a fixed set of knots, estimation in the semi-parametric setting is readily achieved via maximum likelihood or variants thereof. Both of the seminal papers in this area envisioned the use of penalized splines in complex problems involving nonparametric regression:

"[The penalized regression spline is] a flexible and easily implemented methodology for fitting complex nonparametric models"

(Ruppert and Carroll, 1996)

"it [is] natural to extend P-splines to semi-parametric models in which additional explanatory variables occur"

(Eilers and Marx, 1996)

On the other hand, since penalized regression splines are sieve estimators, for which the knot set must vary with sample size to achieve convergence in the general case, statistical inference cannot be based on the usual asymptotics for maximum likelihood. Randomization methods such as randomization tests or a semi-parametric bootstrap are more valid. Furthermore, knot location, which has been shown to be of little import for penalized regression splines for univariate nonparametric regression, proves to be a critical factor in the semi-parametric problems discussed here.

In this paper we consider the following two semi-parametric models:

Self-Modeling Regression: $y_i\,(t) = \alpha_{i0} + \alpha_{i1}\mu_0\,(\beta_{i0} + \beta_{i1}t) + \varepsilon_{it}$
(Lawton et al 1972)

Varying Coefficient Regression: $y(x_1,x_2,z) = \mu_0(z) + \mu_1(z)x_1 + \mu_2(z)x_2 + \varepsilon$
(Hastie and Tibshirani, 1993)

Section 2 of the paper is a very quick introduction to (empirical) Bayesian penalized splines. Section 3 considers the self-modeling regression problem in the context of growth curves for nestlings in a designed experiment with two treatments and a covariate. Section 4 considers one of the examples in Hastie and Tibshirani (1993) and shows the computational simplicity of the penalized spline for this type of problem. Remarks scattered throughout the paper are based on my experience with these and similar applications, as well as unpublished simulation results.

2. **The Bayesian Regression Spline**

Consider a regression problem:

$Y_i = \mu(t_i) + \varepsilon_i$ where $E(\varepsilon_i)=0$ and $\mu(t_i)$ is an unknown regression function for fixed or random predictor t_i. $\mu(t)$ is in the family of penalized regression splines of order p, with knot set $(\tau_1 \ldots \tau_K)$ and quadratic penalty C, if

$$\mu(\mathbf{t}) = \sum_{j=0}^{p} \lambda_j \mathbf{t}^j + \sum_{k=1}^{K} \lambda\, \gamma_k\, (\mathbf{t} - \tau_k)_+^{\mathbf{p}} \quad \text{with} \quad \Sigma\, \gamma_k^{\,2} \le \mathbf{C}.$$

With data $(Y_1, t_1) \ldots (Y_n, t_n)$ the regression function can be estimated using the penalized least squares formulation:

pick λ_j and γ_k to minimize $\left\| \mathbf{Y} - [\mathbf{T}\ \mathbf{Z}] \begin{bmatrix} \lambda \\ \gamma \end{bmatrix} \right\|^2 + \mathbf{f}\, \gamma'\gamma$

where T is the matrix of monomials in ($t_1 \ldots t_n$) of degrees 0 through p, Z is the matrix of truncated polynomials $(\mathbf{t} - \tau_k)_+^{\mathbf{p}}$ and **f** is the Lagrange Multiplier corresponding to the constraint. In the usual usage in nonparametric regression, the knots τ_k are prespecified with large K (generally equally spaced between the minimum and maximum of the observed values of t, or equal quantiles of the observed distribution of t). However, the penalty parameter C or f is chosen adaptively from the data, using a criterion such as generalized cross-validation. We find it convenient to use the GML criterion (Wahba,). For ε_i distributed i.i.d. Normal, the GML estimate of **f** is estimated using the normal mixture model:

$$\mathbf{Y} = \mathbf{T}\lambda + \mathbf{Z}\gamma + \varepsilon \quad \text{where} \quad \varepsilon \sim N(0, \sigma_\varepsilon^2 \mathbf{I})$$
$$\gamma_1 \ldots \gamma_K \sim \mathbf{N}(\mathbf{0}, \sigma_\gamma^2 \mathbf{I})$$

This yields the likelihood $\| \mathbf{Y} - [\mathbf{T}\ \mathbf{Z}] \begin{bmatrix} \lambda \\ \gamma \end{bmatrix} \|^2 + \mathbf{f}\, \gamma'\gamma$ where $\mathbf{f} = \sigma_\varepsilon^2 / \sigma_\gamma^2$. Substituting the maximum likelihood estimator for **f** yields the GML criterion.

Comments:

1. We should not take the variance component for the jumps (γ) seriously - rather this is the GML method for determining the smoothing parameter.

2. This formulation allows for diagonal variance matrix for the γ_i's, leading to variable smoothing parameters (but we do not take advantage of this).

3. The smoothing parameter has been shown (by simulation in the univariate case) to greatly reduce the dependence of the fit on the knot locations. The popular wisdom is that once there are “enough” knots, the fit is largely independent of the knot location.

4. Although estimation proceeds via maximum likelihood, this is a sieve method, where the sieve is defined by the knot set. Thus statistical inference based on this estimator of the regression function should be based on a sieve interpretation of the likelihood.

3. **Self-Modeling Regression Model for a Longitudinal Experiment in Bird Growth**

We now use the Bayesian penalized spline to model growth of nestlings in an experiment on nestling growth with and without a dietary supplement. Although parametric nonlinear growth models may be available, we prefer a nonparametric shape for the growth curve for a number of reasons:

a) There are several experiments with different species that may exhibit different growth patterns.
b) In each experiment, there are several response variables which exhibit differing growth patterns.
c) The “parametric” growth curves usually used have been selected for their empirical similarity to the observed growth patterns – not due to biology.

The main questions of interest within each study are similar:

a) Are there treatment effects?
b) Does the shape of the growth curve vary with response variable?
c) Do treatment effects for the response variables differ? - e.g. If a treatment delays growth of tarsus length, does it delay growth of head circumference.
d) Hatch date is an important determinant of nestling survival. Does it also affect growth parameters?

The analysis proceeds by fitting the self-modeling regression model:

$$y_{ij}(t) = \alpha_{ij0} + \exp(\alpha_{ij1})\mu_0(\exp(\beta_{ij1})t) + \varepsilon_{ijt}. \qquad (1)$$

Here $y_{ij}(t)$ represents a response on the j[th] nestling under treatment i at time t. The exponential forces some of the parameters to be positive, thus ensuring that growth is in the same direction for all of the experimental units. In the basic model, α_{ij0}, α_{ij1} and β_{ij1} are modeled as random effects, but I will continue to refer to them as parameters.

Data Set 1: 2 to 8 times per curve

The first example is nestling growth of tree sparrows with or without dietary supplementation. All of the eggs in the same nest hatch simultaneously, and have the same dietary supplement. So this is a nested design with covariate HatchDate on the nesting factor (which is nest). The data were provided by Matt Wasson, Cornell University.

The model fitted is:

$$y_{ij}(t) = \alpha_{ij0} + \exp(\alpha_{ij1})\mu_0(\exp(\beta_{ij1})t) + \varepsilon_{ijt}$$

$$\theta_{ij} = \gamma_0 + g(\text{HatchDate}_i) + \text{Treatment}_i + \theta^*_{ij} \qquad (2)$$

θ_{ij} is any of the parameters α_{ij0}, α_{ij1}, or β_{ij1}. θ^*_{ijk} is a random effect but each parameter is assumed to have fixed components due to treatment and a nonlinear regression on HatchDate. HatchDate and Treatment are time invariant and are applied to every bird in the nest. μ_0 is the shape of the growth curve and is assumed to be the same for all nestlings.

We did not model nest effects in this preliminary analysis, but these will be added before the analysis is complete. The errors were assumed to be i.i.d., but time series errors can also be modeled with suitable modification to the likelihood equations. For simplicity we look at a single response: tarsus length.

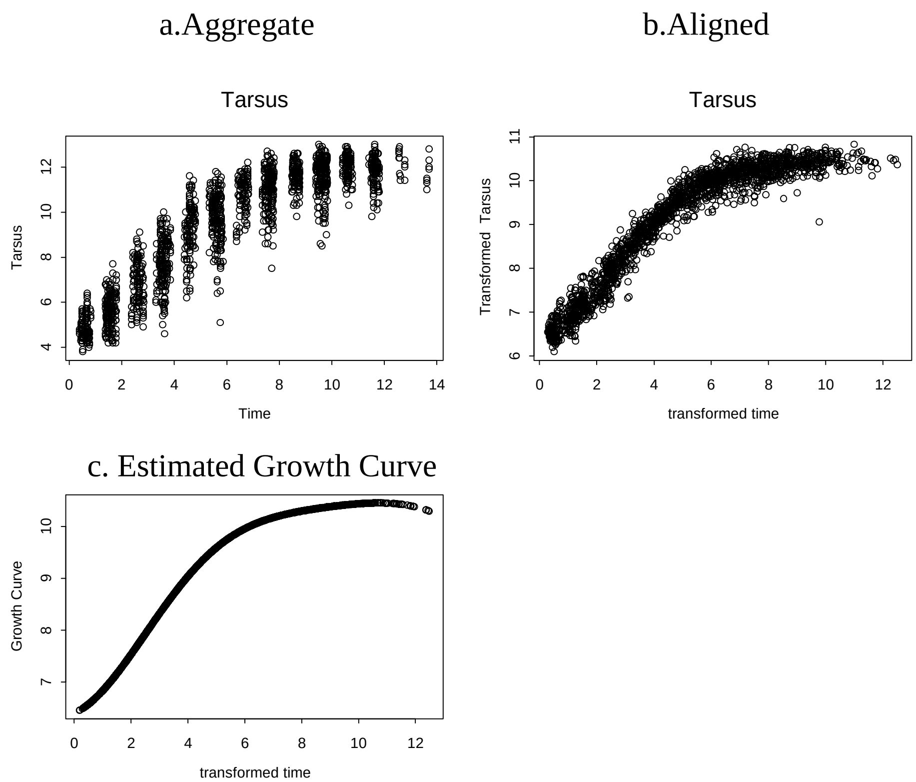


Figure 1: *Tarsus length as a function of time. Panel a) shows the raw data. Panel b) shows the* $(Y_{ij}-a_{ij0})/exp(a_{ij1})$ *plotted against* $exp(b_{ij1})t$*. where aij0,* a_{ij1} *and* b_{ij1} *are the estimated values. Panel c) shows the estimated common shape of the growth curve.*

Figure 1 shows the shape of the growth curve and plots of the raw data and data aligned in time and size. Figure 2 shows the fit of the model to data from some of the individuals in the study.

The model was fitted to the basic self-modeling regression model (1) without the additional parametrization due to the experimental design (2). Figure 3 shows a smooth of the fitted parameters versus HatchDate. Basically, it shows that nestlings born in the middle of the breeding season are larger and grow faster than those born early or late in the season. This suggests that the regression of the parameters on HatchDate can be modeled as a quadratic.

Some individual growth curves

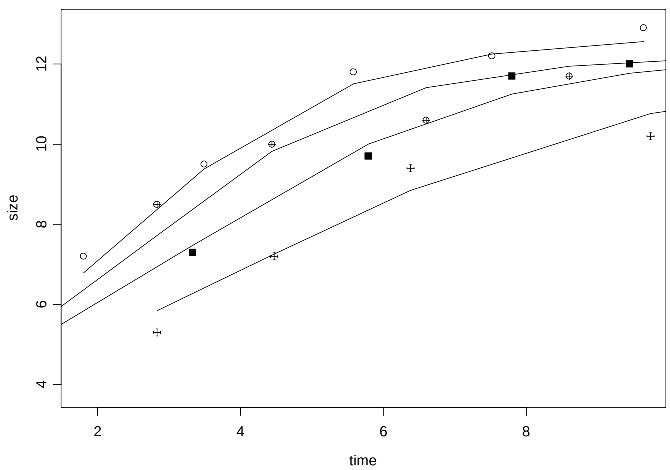


Figure 2: *The fit of the model to the raw data for 4 nestlings.*

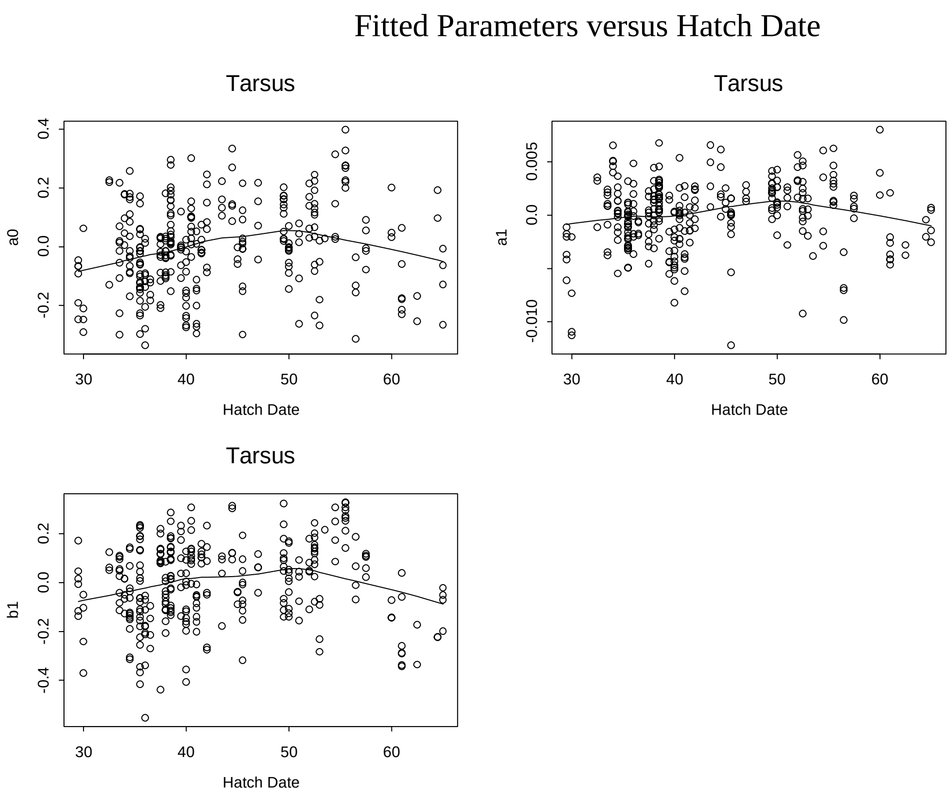
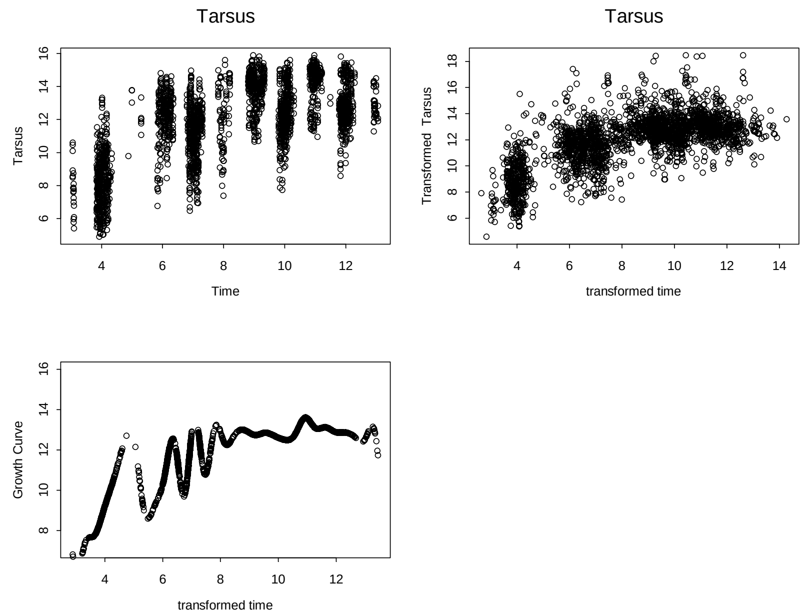


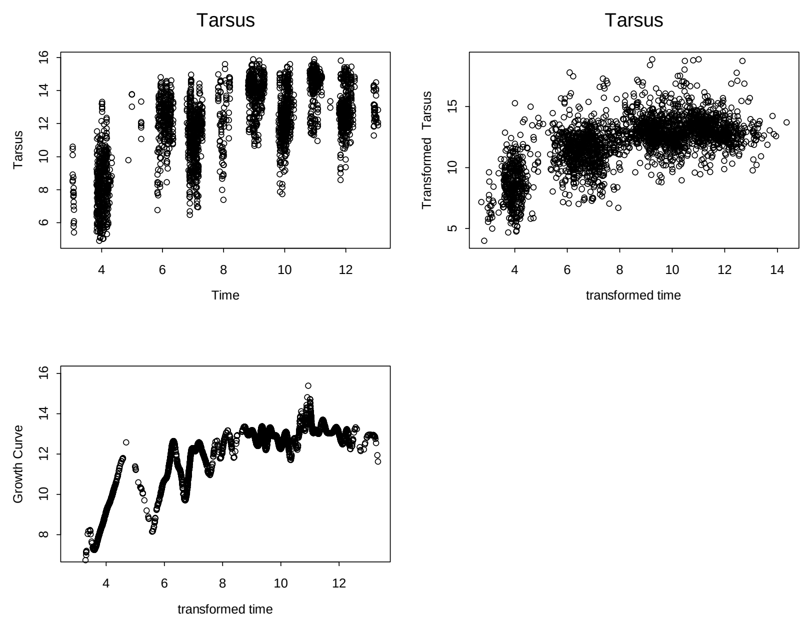
Figure 3: *A smooth of Hatch Date versus the values of the parameters for each nestling.*

Data Set 2: 3 to 4 times per curve

Aggregate data - 32 knots



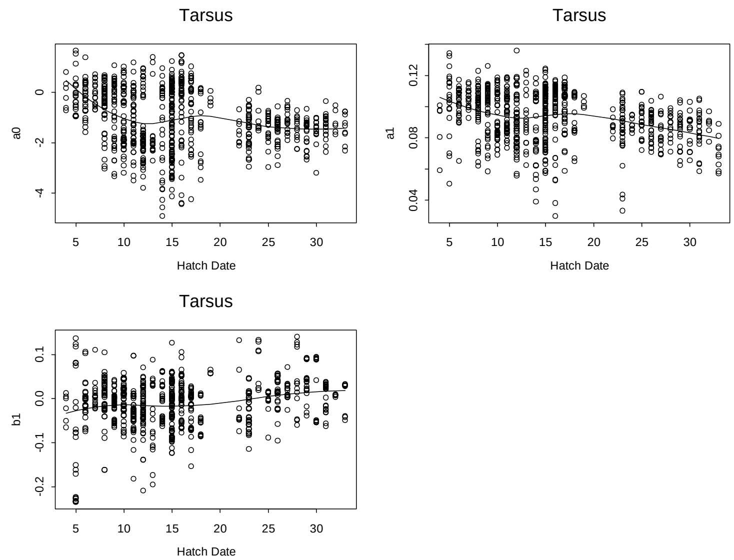
Aggregate data - 86 knots



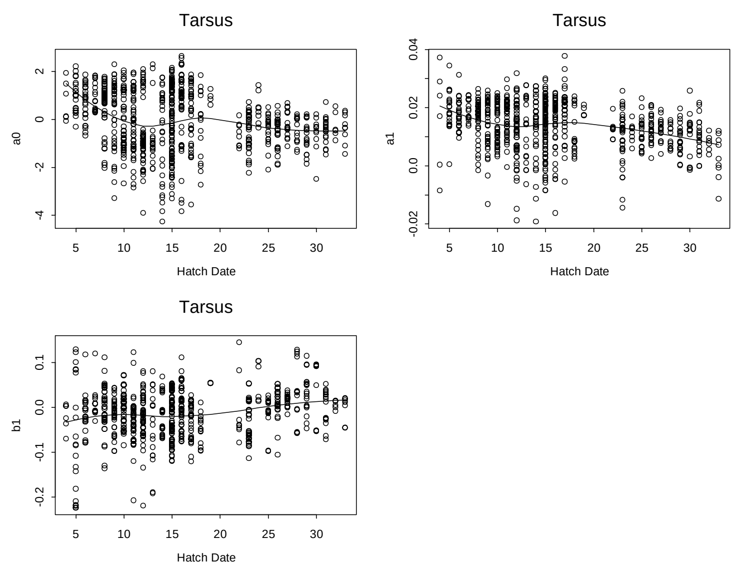
Moral: The insensitivity of the fit to the knot placement may not carry-over to complex models

On the other hand, the parametric part of the model is very insensitive to the knot placement, even for the sparse data

Hatch Date versus parameters - 32 knots



Hatch Date versus parameters - 86 knots



Varying Coefficient Model

$Y(X,Z) = \beta_0(Z) + \Sigma \beta_i(Z) X_i + \varepsilon$

Questions of interest:

a) Which X variables are important predictors?
b) Which X variables have constant coefficients?
c) Could a polynomial regression model do as well?

Modeling with a Bayesian Penalized Spline:

$$\beta_i(Z) = \Sigma\, \lambda_{ij}\, Z^j + \Sigma\, \gamma_{ik}\, (Z - Z_{ik})_+^p$$

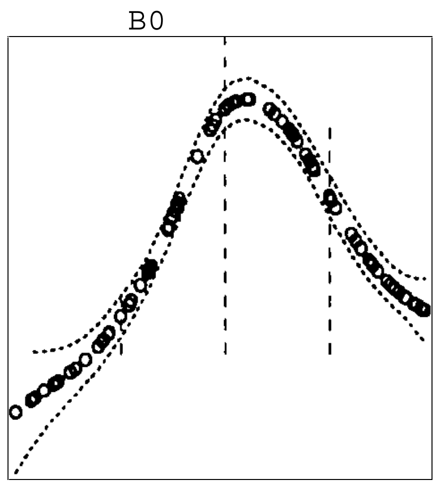
Here the variance component for γ_i is σ_i^2

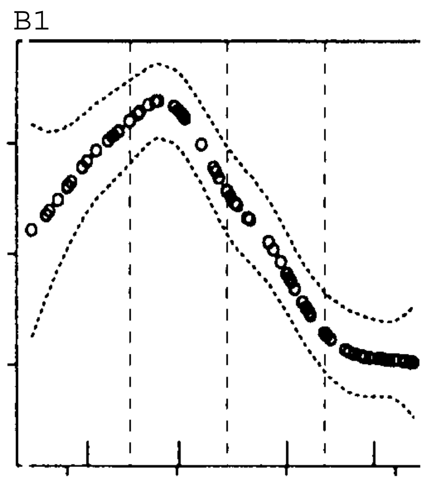
The varying coefficient model is now just a LME with Z*X multiplicative terms.

This facilitates model selection and determining parametric alternatives.

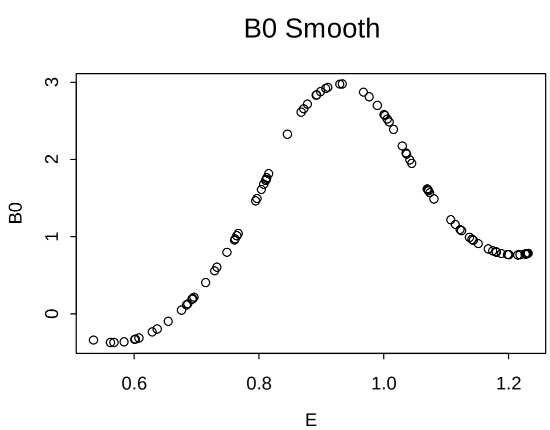
Example: Cleveland et al, 1991
Hastie and Tibshirani, 1993

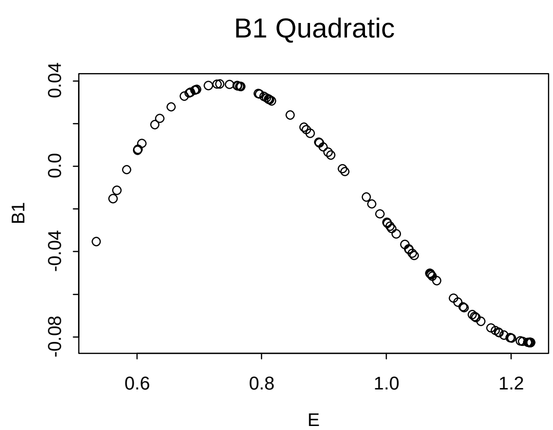
NOx in car exhaust as a function of
X=compression ratio
Z=equivalence ratio





Penalized Spline





Model Selection

Hastie and Tibshirani: approximate F-tests based on equivalent d.f.

Note that the Bayesian Penalized Spline method is a sieve nonparametric estimator.

I used AIC, but probably better to use the Fan semiparametric likelihood ratio test.

The Next Project

Soil compaction as a function of pressure and water content

- nonlinear curves all with the same "shape"
- time and compaction axes depend nonlinearly on soil properties

$\Rightarrow$

self-modeling regression with time-varying coefficients (come back next year)

Conclusions

Bayesian Penalized splines are convenient for fitting complex semi-parametric models.

Fitting is (relatively) rapid, and can be done with off-the-shelf software.

"Folklore" developed from experience with simple (univariate) smoothing, may not apply to complex situations.

There is (still) no free lunch. Even though the fitted models are "conditionally parametric" correct statistical inference requires use of nonparametric inference methods.