On Forwards Prediction Error

Giles Hooker and Stephen P. Ellner

November 3, 2010

1 Introduction

This report examines the theoretical properties of Forwards Prediction Error (FPE) as proposed in two different settings: as a means of selecting smoothing parameters in nonparametric regression (Morton, Kang, and Henderson, 2009) and for selecting robustness parameters in the generalized profiling methodology for estimating differential equations proposed in Ramsay, Hooker, Campbell, and Cao (2007); Ellner (2007). FPE provides a criterion for selecting a smoothing parameter in nonparametric regression and uses the value of an estimated smooth $\hat{x}_\lambda(t)$ and its derivative $\hat{\dot{x}}_\lambda(t)$ to linearly extrapolate to an observation at time $t + h$ via $\hat{x}(t) + h\hat{\dot{x}}(t)$. Here λ represents a smoothing parameter to be chosen to give minimum predicted mean squared error between the observations and the extrapolated prediction. In the case of differential equation models, the equivalent criterion involves solving a differential equation forwards h time units from estimated initial conditions with estimated parameters in a manner made precise below.

This report studies the asymptotic properties of these methods. We demonstrate that in the case of local polynomial regression and other kernel smoothing methods, FPE results in an estimate of λ that is of the same order as h and that intuitive choices for h do not then lead to optimal convergence rates for λ, or even to consistent estimates $\hat{x}_\lambda(t)$. However, in the case of generalized profiling for differential equation models under infill asymptotics, choosing h to be constant leads to consistent estimates of parameters in the differential equation.

Two variations on the methods above are also considered. When local polynomial regression is employed and the global polynomial model is assumed correct, the regression function is estimated consistently. We also examine the use of FPE and the "gradient matching" approach suggested in Ellner, Seifu, and Smith (2002); Brunel (2008) and show that non-parametric rates of convergence are obtained.

2 FPE and Smoothing Parameters in Nonparametric Regression

We suppose that we have data pairs (y_i, t_i) for $i = 1, \dots, n$ with an assumed relationship

$$y_i = x(t_i) + \epsilon_i$$

where the ϵ_i are independent $N(0, \sigma^2)$ random variables. There are several non-parametric estimates of x; smoothing splines (Wahba, 1990) or local polynomial estimators (Fan and Gijbels, 1996) being among the most popular. These all have some smoothing parameter that controls the roughness of the resulting estimate, either governing the strength of the penalty for smoothing splines, or the width of a kernel for local polynomial estimators. Other estimates use the number of terms in a basis expansion such as splines, Fourier series or wavelets (see **?**). There are many potential means of estimating smoothing parameters: cross validation, various information criteria, ReML estimates and so forth, (see Gu, 2002), but these generally yield unsatisfactory results from the point of view of visual aesthetics.

Morton et al. (2009) suggested minimizing forwards prediction error as a means of obtaining smoothing parameters for spline estimates. Specifically, they considered an estimate $x_\lambda(t)$ for a curve along with its derivative $\dot{x}_\lambda(t)$ and measured:

$$FPE(\lambda, h) = \sum_{j=1}^{N_n} \frac{1}{n_j} \sum_{t_i \in [t_j^0 \;\, t_j^0 + h]} \left(y_i - x_\lambda(t_j^0) - (t_i - t_j^0)\dot{x}_\lambda(t_j) \right)^2 . \tag{1}$$

where the t_j^0 are a set of N_n starting points, typically taken as being the same as the data points t_i. The n_j are the number of observed data points falling in the interval $[t_j^0, \; t_j^0 + h]$.

It is worthwhile asking under what conditions minimizing $FPE(\lambda, h)$ will result in theoretically useful estimates. There will be a clear advance if a simple rule for selecting h is available. If consistency or optimal convergence rates require $h = O(n^\beta)$ for $\beta \notin \{0, 1\}$, there will be less utility in the approach.

A brief example will demonstrate that forwards prediction error does not always yield a useful estimate for λ. Consider the local polynomial regression estimate:

$$(\beta_{\lambda 0}(t), \beta_{\lambda 1}(t)) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum (y_i - \beta_0 - (t - t_i)\beta_1)^2 K\left(\frac{t - t_i}{\lambda}\right)$$

with the natural estimates $x_\lambda(t) = \beta_{\lambda 0}(t)$, $\dot{x}_\lambda(t) = \beta_{\lambda 1}(t)$ used in (1). A general examination of the size of λ will follow below. However, the particular choice of $K(t) = I(t \in [0 \; 1])$ yields an optimal value of $\lambda = h$ since for each t_j,

$$(\beta_0(t_j), \beta_1(t_j)) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{t_i \in [t_j \;\, t_j + h]} (y_i - \beta_0 - (t_i - t_j)\beta_1)^2 .$$

Thus, the conditions on h under which $\hat{x}_h(t)$ is estimated consistently or optimally are the same, and just as practically unhelpful, as the original conditions on λ.

More generally, we can examine the relative rates of λ and h within a local linear setting. For the sake of mathematical simplicity, we assume the t_i occur on a circle in order to avoid special calculations for the edges of a data domain and we use the shorthand $|s|_c$ for $s\mathrm{mod}1$.

Theorem 2.1. *Let (y_i, t_i) $i = 1, \ldots, n$ be measurements of a function $x(t)$ with continuous fourth derivatives such such that*

$$y_i = x(t_i) + \epsilon_i$$

with $t_i = (i-1)/n$. Define an estimate $\hat{x}_\lambda(t)$ by a local smooth

$$\hat{x}_\lambda(t) = \frac{1}{n\lambda} \sum_{i=1}^{n} y_i K\left(\frac{|t - t_i|_c}{\lambda}\right)$$

with an estimated derivative

$$\hat{\dot{x}}_\lambda(t) = \frac{1}{n\lambda} \sum_{i=1}^{n} y_i \left(\frac{|t - t_i|_c}{\lambda^2}\right) K\left(\frac{|t - t_i|_c}{\lambda}\right).$$

where K is symmetric with continuous fourth derivatives and $\int K(u)du = \int u^2 K(u)du = 1$.

For each h, define a selection criterion for λ by

$$FPE(\lambda, h) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \hat{x}_\lambda(|t_i - h|_c) - h\hat{\dot{x}}_\lambda(|t_i - h|_c) \right]^2$$

and let

$$\hat{\lambda}_h = \underset{\lambda}{\operatorname{argmin}} \, FPE(\lambda, h)$$

then

$$\hat{\lambda}_h = O_p\left(\max(h, n^{-1/2})\right)$$

as $h \to 0$.

In this theorem our definitions of $\hat{x}_\lambda(t)$ and $\hat{\dot{x}}_\lambda(t)$ are nearly those for a local linear smooth for regularly spaced data, differing only by the absence of normalization. This will make no difference asymptotically, but serves to simplify our calculations.

Proof. We first observe that as in Härdle and Marron (1985), it is sufficient to examine the expectation of forwards prediction error. Expanding

$$\begin{aligned} E\left[FPE(\lambda,h)\right] = \frac{1}{n}\sum\left(x(t_i) - \frac{1}{n\lambda}\sum x(t_j)K\left(\frac{|t_i-h-t_j|_c}{\lambda}\right)\left[1+h\frac{|t_i-h-t_j|_c}{\lambda^2}\right]\right)^2 \\ + \frac{\sigma^2}{n^2}\sum\left(1-\frac{1}{n\lambda}K\left(\frac{-h}{\lambda}\right)\left[1-\frac{h^2}{\lambda^2}\right]\right)^2 \end{aligned}$$

From this, second-order Taylor series expansions give us that

$$\begin{aligned} \frac{1}{n\lambda}\sum x(t_j)K\left(\frac{|t_i-h-t_j|_c}{\lambda}\right) &= x(t_i-h) + o\left(\lambda^2\right) + o\left(\frac{1}{n}\right) \\ \frac{1}{n\lambda}\sum x(t_j)\frac{|t_i-h-t_j|_c}{\lambda^2}K\left(\frac{t_i-h-t_j}{\lambda}\right) &= \dot{x}(t_i-h) + o\left(\lambda^2\right) + o\left(\frac{1}{n}\right) \\ x(t_i) &= x(t_i-h) + h\dot{x}(t_i-h) + o\left(h^2\right) \end{aligned}$$

Thus we can characterize

$$E\left[FPE(\lambda,h)\right] = \left[o\left(\lambda^2(1+h)\right) + o\left(h^2\right) + o\left(\frac{1+h}{n}\right)\right]^2 + o\left(\frac{1}{n}\right).$$

From here, assuming $h \to 0$, we can match

$$\lambda = o\left(\sqrt{h^2+1/n}\right)$$

yielding in general

$$\lambda = o\left(\max\left(n^{-1/2}, h\right)\right).$$

□

We also note that the result above can be generalized to higher-order extrapolation, assuming greater regularity of $x(t)$. We further note

1. Because we have defined $FPE(\lambda)$ on a circle, the same results will apply to ordinary kernel smoothing with the derivative estimated directly:

   $$\hat{\dot{x}}_\lambda(t) = \frac{1}{n\lambda^2}\sum K'\left(\frac{t-t_i}{\lambda}\right)$$

   using the calculations as above.

2. If multiple values of h are used $h_1, \dots, h_k$, we obtain that

   $$\lambda = o\left(\max\left(\sqrt{n}, \sqrt{\sum h_i^2}\right)\right)$$

3 FPE and Robustness Parameters

While nonparametric estimates do not yield consistent results, the concept of forwards prediction error can be used to motivate a measure of robustness for parametric models. In particular, if we consider a model in which the nominal value of λ is ∞ and the smoothing process is used as a means of providing robustness to model miss-specification. We consider the two-level criterion described for the estimation of parameters in ordinary differential equations in Ramsay et al. (2007). A model is assumed of the form

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{f}(\mathbf{x}(t); t, \boldsymbol{\theta}) && (2) \\ \mathbf{x}(0) &= \mathbf{x}_0 \\ y_i &= \mathbf{x}_j(t_i) + \epsilon_i && (3) \end{aligned}$$

Here the ϵ_i are independent and identically distributed $N(0, \sigma^2)$ random variables. $\boldsymbol{\theta}$ represents an unknown set of parameters to be estimated; we denote the correct value by $\boldsymbol{\theta}_0$ below. The system is thought to be potentially inexact due to system disturbances and model miss-specification. Ramsay et al. (2007) proposed allowing extra flexibility to the model by allowing departures from (2). In particular, they proposed a nested optimization criterion. For each candidate value $\boldsymbol{\theta}$, the procedure estimates a smooth

$$\hat{\mathbf{x}}_{\lambda,\boldsymbol{\theta}}(t) = \underset{\mathbf{x} \in \otimes^k W^1}{\operatorname{argmin}} \sum (y_i - \mathbf{x}(t_i))^2 + \lambda \int \|\dot{\mathbf{x}}(t) - \mathbf{f}(\mathbf{x}(t); t, \boldsymbol{\theta})\|^2 dt.$$

Where in practise, $\hat{\mathbf{x}}_{\lambda,\boldsymbol{\theta}}(t)$ is represented via a basis expansion. Values for $\boldsymbol{\theta}$ were then chosen as

$$\hat{\boldsymbol{\theta}}(\lambda) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum (y_i - \hat{\mathbf{x}}_{\lambda,\boldsymbol{\theta}}(t_i))^2.$$

Hooker (2007) demonstrated that for $\lambda \to \infty$, the estimates for $\hat{\boldsymbol{\theta}}(\lambda)$ were identical to those that would be obtained by solving (2) for each value of $\boldsymbol{\theta}$ and estimating both $\boldsymbol{\theta}$ and $\mathbf{x}_0$ by minimizing squared error. Qi and Zhao (2010) further showed that any choice of $\lambda_n \to \infty$ as $n \to \infty$ resulted in consistent estimates if (2) is an exact model, but that efficiency required $\lambda_n = O(n^2)$.

In this context, we redefine forwards prediction error through solving (2) forwards. We let $\mathbf{x}(t, \boldsymbol{\theta}, \mathbf{x}_0)$ denote the solution, up to time t, of (2) at parameters $\boldsymbol{\theta}$ and initial conditions $\mathbf{x}_0$. Then we define

$$\mathrm{FPE}_2(\lambda, h) = \sum_{j=1}^{N_n} \frac{1}{n_j} \sum_{t_i \in [t_j^0 \;\; t_j^0 + h]} \left(y_i - x_i \left(t_i - t_j^0, \hat{\boldsymbol{\theta}}(\lambda), \mathbf{x}_{\lambda, \hat{\boldsymbol{\theta}}(\lambda)}(t_j^0) \right) \right)^2. \tag{4}$$

This measures the deviation of the data from the solution of (2) going forwards from the point $\mathbf{x}_{\lambda,\hat{\boldsymbol{\theta}}(\lambda)}(t_j^0)$ with estimated parameters $\hat{\boldsymbol{\theta}}(\lambda)$.

We can show that under the in-fill sampling studied for non-parametric estimators, with h fixed, choosing

$$\lambda_{FPE} = \operatorname{argmin} FPE_2(\lambda, h)$$

results in a consistent estimator of $\boldsymbol{\theta}$ regardless of the t_j^0 or N_n.

Theorem 3.1. *If (2-3) holds, $\boldsymbol{\theta}$ lies in a compact space Θ and is identifiable in the sense that*

$$\forall \epsilon > 0, \exists \delta > 0 \text{ such that } \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon \to \sup_{t \in [0 \;\; 1], \mathbf{x}_0} \int_0^h |\mathbf{x}(t, \boldsymbol{\theta}, \mathbf{x}_0) - \mathbf{x}(t, \boldsymbol{\theta}_0, \mathbf{x}_0)\|^2 \, dt > \delta$$

and $\max(t_{i+1} - t_i) = o(n^{-1})$*, then* $\boldsymbol{\theta}_{\lambda_{FPE}} \to \boldsymbol{\theta}_0$ *in probability.*

Proof. We let $\mathbf{x}^*(t)$ represent the trajectory solving (2) that generates the data. Then as $n \to \infty$ with fixed h we have

$$\sum_{t_i \in [t_j^0 \;\; t_j^0 + h]} \left(y_i - x_i \left(t_i - t_j^0, \hat{\boldsymbol{\theta}}(\lambda), \mathbf{x}_{\lambda,\hat{\boldsymbol{\theta}}(\lambda)}(t_j^0) \right) \right)^2 \to \sigma^2 + \int_{t_j^0}^{t_j^0 + h} \left(x_i \left(s - t_j^0, \hat{\boldsymbol{\theta}}(\lambda), \mathbf{x}_{\lambda,\hat{\boldsymbol{\theta}}(\lambda)}(t_j^0) \right) - x_i^*(s) \right)^2 ds.$$

uniformly over $\boldsymbol{\theta} \in \Theta$. Under the identifiability conditions above, the last term on the right hand side is minimized at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ with $\mathbf{x}^*(t) = \mathbf{x}(t, \boldsymbol{\theta}_0, \mathbf{x}_0)$ which is achieved as $\lambda \to \infty$. □

4 Robustness in Parametric Regression Models and Gradient Matching

We note that similar results hold in the case of local-linear estimation, or spline smoothing when an appropriate null model is employed. In particular, we can demonstrate the following theorem:

Theorem 4.1. *Let $max(t_{i+1} - t_i) = o\left(n^{-1}\right)$ and*

$$y_i = \beta_0 + \beta_1 t_i + \epsilon_i$$

with $E\epsilon_i = 0$ and let a non-parametric estimator $x_\lambda(t)$ be such that

$$\lim_{\lambda \to \infty} x_\lambda(t) \to \hat{\beta}_0 + \hat{\beta}_1 t$$

for $\hat{\beta}_0$ and $\hat{\beta}_1$ the least-squares estimators of β_0 and β_1, then $x_{\lambda_{FPE}}(t) \to \beta_0 + \beta_1 t$ in probability.

The proof of this proceeds along the same lines as that of Theorem 3.1; essentially FPE can be minimized for $x_\lambda(t) = \beta_0 + \beta_1 t$ which occurs at $\lambda = \infty$.

We can also examine forwards prediction error in the context of gradient matching as described in Ellner et al. (2002). This is a two-step estimate, first obtaining a non-parametric smooth $\hat{\mathbf{x}}_\lambda(t)$ and derivative $\hat{\dot{\mathbf{x}}}_\lambda(t)$ and then choosing $\hat{\boldsymbol{\theta}}$ from

$$\operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \int \|\hat{\dot{\mathbf{x}}}_\lambda(t) - \mathbf{f}(\hat{\mathbf{x}}_\lambda(t); t, \boldsymbol{\theta})\|^2 dt.$$

In this context, using the minimizing values $\mathrm{FPE}_2(\lambda, h)$ for λ can also be shown to achieve consistency for $\hat{\boldsymbol{\theta}}$. In this case, the choice of $\lambda = \infty$ will not yield a parametric form. However, there is a sequence λ_n for which $\hat{\mathbf{x}}_{\lambda_n}(t)$ and $\hat{\dot{\mathbf{x}}}_{\lambda_n}(t)$ are consistent (Brunel, 2008) and choosing this sequence will yield, asymptotically, the minimizing values of $\mathrm{FPE}_2(\lambda, h)$.

References

Brunel, N. (2008). Parameter estimation of ODEs via nonparametric estimators. *Electronic Journal of Statistics 2*, 1242–1267.

Ellner, S. P. (2007). Commentary on "parameter estimation in differential equations: A generalized smoothing approach". *Journal of the Royal Statistical Society, Series B 16*, 741–796.

Ellner, S. P., Y. Seifu, and R. H. Smith (2002). Fitting population dynamic mo9dels to time-series data by gradient matching. *Ecology 83*, 2256–2270.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications.* New York: Chapman & Hall/CRC.

Gu, C. (2002). *Smoothing Spline ANOVA Models.* New York: Springer.

Härdle, W. and J. S. Marron (1985). Optimal bandiwth selection in nonparametric regression function estimation. *Annals of Statistics 13*, 1465–1481.

Hooker, G. (2007). Theorems and calculations for smoothing-based profiled estimation of differential equations. Technical Report BU-1671-M, Dept. Bio. Stat. and Comp. Bio., Cornell University.

Morton, R., E. L. Kang, and B. L. Henderson (2009). Smoothing splines for trend estimation and prediction in time series. *Environmetrics 20*, 249–259.

Qi, X. and H. Zhao (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *Annals of Statistics 38*(1), 435–481.

Ramsay, J. O., G. Hooker, D. Campbell, and J. Cao (2007). Parameter estimation in differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society, Series B 16*, 741–796.

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM CBMS-NSF Regional Conference Series in Applied Mathematics.