Approximating unknown functions when fitting errors involve costs

Hans-Dieter Heike, Matei Demetrescu

Abstract. The problem of approximating unknown functions is discussed for the case where approximation errors are evaluated by means of a general cost-of-error (loss) function, not necessarily the squared-error one. To ensure minimal overall loss, approximation is carried out by fitting under the relevant loss function. Convergence results are provided in a general framework, allowing among others for Taylor approximations, approximations with Hermite polynomials or approximations with Neural Networks and for random measurement error.

M.S.C. 2000: 62G08, 65D10.
Key words: asymmetric loss, interpolation, sieve approximation.

1 Introduction

Function approximation is a common task in many branches of applied science, e.g. in control engineering or operations research. A specific response function is often desired, say $f(x)$ with $x \in D \subset \mathbb{R}$, but, in many cases, the desired function is not available in closed form or is not easily handled. Also, it may be that the function is only known at some points (nodes) x_t of its support, $t = 1, 2, \ldots, T$, and needs interpolation, resulting in the use of an approximating function, say $\widetilde{f}_n(x)$ where $\widetilde{f}_n$ denotes the approximation of n^{th} order in a family of approximating functions.

An important aspect of approximating a known or unknown function is the approximation error $f(x) - \widetilde{f}_n(x)$ (sometimes called bias). While uniform upper bounds for the approximation error in the domain D can be derived for many interpolation methods, these are only a general indication of the precision of the method in cause. In many practical applications, approximation errors can be identified as the source of costs, e.g. in Taguchi's approach to quality optimization. In time series analysis there is a long tradition of evaluating forecast errors according to the costs these incur, where the cost-of-error (loss) function is not necessarily the familiar squared-error loss, see Granger [4].

Proceedings of The 4-th International Colloquium "Mathematics in Engineering and Numerical Physics" October 6-8 , 2006, Bucharest, Romania, pp. 68-72.

Should the function be observed only at some points of its support, or observed with random measurement noise, fitting the desired function to the available, usually experimental, data is the only way to obtain an approximation. Then, following Weiss [7], this function should be fitted under the relevant loss function, i.e. by minimizing mean loss due to approximation error evaluated at each node. Even if f is completely known, and the parameters of an approximation can be computed analytically (e.g. for a Taylor expansion), these will lead to higher mean loss than fitting.

The main contribution of this note is to study the asymptotic behavior of fitting unknown functions under the relevant loss. We show that, if allowing the approximation order to grow to infinity (this is sometimes called sieve approximation), but at a slower rate that the number of nodes, the fitted parameters converge to their true values, allowing the desired function to be approximated with arbitrary accuracy. If keeping the approximation order constant, the mean cost due to approximation error, although higher than before, is minimized. This adds to a result due to White [8] for the nonlinear least squares fitting procedure.

2 The fitting problem

We extend the studied problem to the case where the true value of the function is only given with random additive noise. This allows us to use tools of mathematical statistics that simplify our task, but comes at the cost that all convergence results are stated in probability. However, this is not a serious restriction: the statistical framework reduces to the stated interpolation problem if letting the probability distribution of the noise degenerate to a constant. In this case, convergence in probability turns into standard convergence.

Assumption 1. *Let $y_t = f(x_t) + \varepsilon_t$, where ε_t is independent and identically distributed.*

The *iid* assumption is standard for controlled experiments. With observed data (e.g. in social sciences), this may not be fulfilled and leads to complications, see Christoffersen and Diebold [3].

Assumption 2. *Let $T \to \infty$ such that the empirical distribution function of the nodes x_t converges to a proper distribution function.*

This ensures, for instance, that x_t may be treated as stochastic *iid*, and, more importantly, that x_t is independent of ε_t, see Amemiya [1].

Denote now $\{\theta_i\}_{i\in\mathbb{N}}$ a sequence of parameters. To ease the exposition, let ${}_n\theta = (\theta_0, \theta_1, \theta_2, \dots, \theta_n)'$.

Assumption 3. *There is a sequence $\left\{\widetilde{f}_n\right\}_{n\in\mathbb{N}}$ of approximating functions, $\widetilde{f}_n : D \times \mathbb{R}^{n+1} \to \mathbb{R}$, such that a unique sequence of parameters $\left\{\widetilde{\theta}_i\right\}_{i\in\mathbb{N}}$ exists for any function $f(x) : D \to \mathbb{R}$ in a given set with $\widetilde{f}_n\left(x, {}_n\widetilde{\theta}\right) \to f(x)$ as $n \to \infty$ uniformly $\forall x \in D$. Further, let $\widetilde{f}_n\left(x, {}_n\widetilde{\theta}\right)$ be Lipschitz continuous w.r.t. ${}_n\widetilde{\theta}$.*

Taylor polynomials, Hermite polynomials, Tschebyscheff polynomials, Fourier series or Neural Networks can be checked to fulfill Assumption 3.

Granger [4] stated general conditions for loss functions. We require additional regularity conditions, namely continuity and convexity:

Assumption 4. *Let $\mathcal{L}$ be continuous and convex on $\mathbb{R}$, increasing on $\mathbb{R}_+$ and decreasing on $\mathbb{R}_-$ with $\mathcal{L}(0) = 0$.*

Convexity can be dropped at the cost of a more complicated proof, see Remark 1.

Assumption 5. *Let $E(\mathcal{L}(\varepsilon_t - b)) < \infty$, $\forall b \in \mathbb{R}$, and assume further that* $\arg\min_{b\in\mathbb{R}} E(\mathcal{L}(\varepsilon_t - b)) = 0$.

The parameters for the approximation are then estimated from

$$ {}_n\widehat{\theta} = \underset{{}_n\theta\in\mathbb{R}^{n+1}}{\arg\min} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}\left(y_t - \widetilde{f}_n\left(x_t, {}_n\theta\right)\right). $$

We address the question of convergence of ${}_n\widehat{\theta}$ to a limiting sequence $\{\theta_i^*\}_{i\in\mathbb{N}}$ in the following sense

$$ \left\| {}_n\widehat{\theta} - {}_n\theta^* \right\|_1 \xrightarrow{p} 0 \text{ as } n, T \to \infty, $$

where $\|\cdot\|_1$ denotes the L_1 vector norm. If $\{\theta_i^*\}_{i\in\mathbb{N}} \equiv \left\{\widetilde{\theta}_i\right\}_{i\in\mathbb{N}}$, this convergence (stronger than elementwise convergence), together with Assumption 3, implies uniform convergence in probability of $\widetilde{f}_n\left(x, {}_n\widehat{\theta}\right)$ to $f(x)$.

3 Convergence result

We examine first the case where $n \to \infty$, but slower than T, i.e. $n/T \to 0$. We show the estimated parameters to converge to the true values, $\left\{\widetilde{\theta}_i\right\}_{i\in\mathbb{N}}$.

Proposition 1. *Under Assumptions 1 through 5, it holds as $n, T \to \infty$ and $n/T \to 0$ that*

$$ \left\| {}_n\widehat{\theta} - {}_n\widetilde{\theta} \right\|_1 \xrightarrow{p} 0. $$

Proof. The result is derived in two steps. In the first step, the target function $T^{-1}\sum_{t=1}^{T} \mathcal{L}\left(y_t - \widetilde{f}_n\left(x_t, {}_n\theta\right)\right)$ is shown to converge pointwise in probability as $n, T \to \infty$ to a deterministic function minimized only at $\left\{\widetilde{\theta}_i\right\}_{i\in\mathbb{N}}$. In the second, pointwise convergence is shown to imply uniform convergence. Then, using continuity of the arg min operator w.r.t. $\|\cdot\|_1$ and the supremum norm, the desired result is established. To prove step 1, note that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(y_t-\widetilde{f}_n\left(x_t,\,{}_n\theta\right)\right) &= \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(f\left(x_t\right)+\varepsilon_t-\widetilde{f}_n\left(x_t,\,{}_n\theta\right)\right)\\ &= \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\varepsilon_t+f\left(x_t\right)-\widetilde{f}_n\left(x_t,\,{}_n\theta\right)\right).\end{aligned}$$

At ${}_n\theta = {}_n\widetilde{\theta}$, we have $f(x_t)-\widetilde{f}_n\left(x_t,\,{}_n\widetilde{\theta}\right)=o_p\left(1\right)$, with $o_p\left(\cdot\right)$ the correspondent of the Landau symbol in probabilistic terms, and thus

$$\mathcal{L}\left(\varepsilon_t+f(x_t)-\widetilde{f}_n\left(x_t,\,{}_n\theta\right)\right)=\mathcal{L}\left(\varepsilon_t\right)+o_p\left(1\right),$$

so

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(y_t-\widetilde{f}_n\left(x_t,\,{}_n\theta\right)\right)=\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\varepsilon_t\right)+o_p\left(1\right).$$

A Law of Large Numbers for *iid* variables thus delivers

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(y_t-\widetilde{f}_n\left(x_t,\,{}_n\theta\right)\right)\xrightarrow{p}E\left(\mathcal{L}\left(\varepsilon_t\right)\right).$$

At ${}_n\theta\neq{}_n\widetilde{\theta}$, it follows that $\widetilde{f}_n\left(x,\,{}_n\theta\right)\rightarrow f^*\left(x\right)$, with $\left|f^*\left(x\right)-f(x)\right|>0\ \forall x\in D$. Hence,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(y_t-\widetilde{f}_n\left(x_t,\,{}_n\theta\right)\right) &= \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\varepsilon_t+f(x_t)-f^*\left(x_t\right)\right)+o_p\left(1\right)\\ &\xrightarrow{p} E\left(\mathcal{L}\left(\varepsilon_t+f(x_t)-f^*\left(x_t\right)\right)\right),\end{aligned}$$

where the expectation is taken w.r.t. the joint distribution of ε_t and x_t. It was shown by Heike and Demetrescu [5] that

$$E\left(\mathcal{L}\left(\varepsilon_t\right)\right)<E\left(\mathcal{L}\left(\varepsilon_t+u_t\right)\right),$$

if u_t is a random variable, independent of ε_t, and non-zero with probability 1, see their Lemma 1. Since, due to Assumption 2, $f(x_t)-f^*\left(x_t\right)$ fulfills these conditions, the limit of the target function is minimized at the true parameters alone. Step 2 is easily completed with the result of Andersen and Gill [2], who showed uniform convergence to follow from pointwise convergence and convexity of the loss function, see their Lemma II.1. The result follows. □

Remark 1. *Step 2 of the proof can also be completed by showing stochastic equicontinuity of the sequence of target functions. Uniform convergence then follows due to a result of Newey [6]. This, however, complicates the proof significantly.*

Let us now consider the case where n is fixed. The estimated parameters will not converge to the true values, $\left\{\widetilde{\theta}_i\right\}_{i\in\mathbb{N}}$, but to those parameter values ${}_n\theta$ that minimize the overall expected loss due to measurement noise *and* bias $b\left(x\right)=f\left(x\right)-\widetilde{f}_n\left(x,\,{}_n\theta\right)$, where the expectation is taken w.r.t. the distribution of x_t and ε_t jointly.

Proposition 2. *Under Assumptions 1 through 5, it holds as* $T \to \infty$ *and* $n < \infty$ *that*

$$ {}_n\widehat{\theta} \xrightarrow{p} \underset{{}_n\theta \in \mathbb{R}^{n+1}}{\arg\min}\, E\left(\mathcal{L}\left(\varepsilon_t + b\left(x_t\right)\right)\right). $$

Proof. By arguments similar to those in the proof of Proposition 1, the target function converges pointwise to $E\left(\mathcal{L}\left(\varepsilon_t + b\left(x_t\right)\right)\right)$. The result follows as before. □

Remark 2. *Should* $\varepsilon_t = 0$ *with probability 1, the fitted parameters converge to those parameters minimizing the expected bias.*

4 Concluding remarks

Asymptotic treatment of the problem of fitting a possibly unknown function under a general cost-of-error (loss) function is provided. Even if analytical approximations are available, fitting delivers better results in terms of overall loss due to approximation error.

References

[1] T. Amemiya, *Advanced Econometrics*, Harvard University Press, 1985.

[2] P.K. Andersen and R.D. Gill, *Cox's regression model for counting processes: a large sample study*, The Annals of Statistics 10 (1982), 1100-1120.

[3] P.F. Christoffersen and F.X. Diebold, *Optimal prediction under asymmetric loss*, Econometric Theory 13 (1997), 808-817.

[4] C.W.J. Granger, *Prediction with a generalized cost of error function*, Operational Research Quarterly 20 (1969), 451-468.

[5] H.-D. Heike and M. Demetrescu (2006), *Forecasting stationary processes under asymmetric loss*, mimeo.

[6] W.K. Newey, *Uniform convergence in probability and stochastic equicontinuity*, Econometrica 59 (1991), 1161-1167.

[7] A.A. Weiss, *Estimating time series models using the relevant cost function*, Journal of Applied Econometrics 11 (1996), 539-560.

[8] H. White, *Consequences and detection of misspecified nonlinear regression models*, Journal of the American Statistical Association 76 (1981), 419-433.

Authors' addresses:

Hans-Dieter Heike
Statistics and Econometrics, Technical University Darmstadt,
Residenzschloss, 64283 Darmstadt, Germany.
e-mail: heike@vwl.tu-darmstadt.de

Matei Demetrescu
Statistics and Econometric Methods, Goethe-University Frankfurt,
Gräfstr. 78 PF 76, 60054 Frankfurt, Germany.
e-mail: deme@wiwi.uni-frankfurt.de