**Title**
A Note on Adaptive Estimation



**Author**
Steigerwald, Douglas G

**Publication Date**
2006-04-24

A Note on Adaptive Estimation

Douglas G. Steigerwald

## Adaptive Estimation

An adaptive estimator is an efficient estimator for a model that is only partially specified.

For example, consider estimating a parameter that describes a sample of observations drawn from a distribution $F$. One natural question is: Is it possible that an estimator of the parameter constructed without knowledge of $F$ could be as efficient (asymptotically) as any well-behaved estimator that relies on knowledge of $F$? For some problems the answer is yes, and the estimator that is efficient is termed an adaptive estimator.

Consider the familiar scalar linear regression model

$$Y_t = \beta_0 + \beta_1 X_t + U_t,$$

where the regressor is exogenous and $\{U_t\}$ is a sequence of $n$ independent and identically distributed random variables with distribution $F$. The parameter vector $\beta = (\beta_0, \beta_1)'$ is often of interest rather than the distribution of the error, $F$. If we assume that $F$ is described by a parameter vector $\alpha$ (that is, we parameterize the distribution), then the resultant (maximum likelihood or ML) estimator of $\beta$ is parametric. If we assume only that $F$ belongs to a family of distributions, then the resultant estimator of $\beta$ is *semi*parametric. Because the OLS estimator does not require that we parameterize $F$, the OLS estimator is semiparametric. If the population error distribution is Gaussian, we know that the OLS estimator is equivalent to the ML estimator, and so is efficient. Although the OLS estimator is generally inefficient if $F$ is not Gaussian, it may be possible to construct an alternative (semiparametric) estimator that retains asymptotic efficiency if $F$ is not Gaussian. If we find that, for a family of distributions that includes the Gaussian this estimator is asymptotically equivalent to the ML estimator, then this estimator is adaptive for that family.

The question then is: How can we verify that an estimator is adaptive? As there will generally be an arbitrarily large number of distributions in the family, it is not feasible to algebraically verify asymptotic equivalence for each distribution. In a creative paper, Stein (1956) first proposed a solution to this problem. Let $\mathcal{F}_\alpha$ define a set of distributions parameterized by a value of $\alpha$ (each member of this family must satisfy certain technical conditions, such as differentiability with respect to $\beta$, which will not be explicitly defined). Although primary interest centers on $\beta$, the full set of parameters includes $\alpha$. The information matrix, evaluated at the population parameter values, is

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix},$$

where $\mathcal{I}_{11}$ corresponds to the elements of $\beta$. Estimators of $\beta$ (again, the estimators must satisfy technical conditions, such as $\sqrt{n}$ consistency, which are also not explicitly defined) will have covariance matrix that is at least as large as $\mathcal{I}^{11}$, which is the upper left component of $\mathcal{I}^{-1}$. If the partial derivative of the log-likelihood with respect to $\beta$ (the score for $\beta$) is orthogonal to the score for $\alpha$, then $\mathcal{I}_{12} = 0$ and $\mathcal{I}^{11} = \mathcal{I}_{11}^{-1}$. Because $\mathcal{I}_{11}$ corresponds only to the parameter $\beta$, the asymptotically efficient estimator of $\beta$ can be constructed without knowledge of $\alpha$. Stein argued that if the condition $\mathcal{I}_{12} = 0$ holds for all the sets $\mathcal{F}_\alpha$ that comprise a family of distributions $\mathcal{F}$, then $\beta$ is adaptively estimable within the family $\mathcal{F}$.

While Stein's condition holds intuitive appeal, it is not straightforward how to use the condition to define estimators that are adaptive. In an invited lecture, Bickel (1982) laid out a simpler condition that does yield a straightforward link to the construction of adaptive estimators. To understand the condition, let $E_F$ denote expectation with respect to the population error distribution and let $E_{\tilde{F}}$ denote expectation with respect to an arbitrary distribution $\tilde{F} \in \mathcal{F}$. Let $l$ be the log-likelihood for the regression model with data $z = (y, x)$ and let $\dot{l}(z, \beta, F)$ denote the score for $\beta$, constructed from the model in which $F$ is the error distribution. A familiar condition that arises in the context of likelihood estimation is that the expected population score $E_F\left[\dot{l}(z, \beta, F)\right]$ equal 0. Bickel's condition is simply that the population score must have expectation zero over the entire family $\mathcal{F}$, that is, for any $\tilde{F} \in \mathcal{F}$,

$$E_{\tilde{F}}\left[\dot{l}(z, \beta, F)\right] = 0.$$

The two conditions are linked: If $\mathcal{F}$ is a convex family, then Stein's condition is implied by Bickel's condition. For the linear regression model, an adaptive estimator of $\beta$ exists for the family $\mathcal{F}$ that consists of all distributions that are symmetric about the origin (and several other technical conditions). If interest centers on the slope coefficient alone, then one need not restrict attention to distributions that are symmetric about the origin, as an adaptive estimator of $\beta_1$ can exist even if $\beta_0$ is not identified.

Bickel's score condition leads naturally to estimators that contain nonparametric estimators of the distribution, $\hat{F}$. In consequence, adaptive estimation requires a second condition: the nonparametric estimator of the score must converge in quadratic mean to the population score. The resulting estimators of $\beta$ are two-step estimators. The estimators require, as the first step, a $\sqrt{n}$-consistent estimator such as the OLS estimator. To understand the estimator's form, note that if the distribution were known, then the two-step (linearized likelihood) estimator is

$$B_{OLS} + n^{-1} \sum_{t=1}^{n} s\left(Z_t, B_{OLS}, F\right),$$

with $s\left(Z_t, B_{OLS}, F\right) = \mathcal{I}^{11}\left(B_{OLS}, F\right) \dot{l}\left(Z_t, B_{OLS}, F\right)$. The linearized likelihood estimator is asymptotically efficient. To form an adaptive estimator of $\beta$, we must replace $F$ with a nonparametric estimator $\hat{F}$. If $\hat{F}$ is constructed so that $s\left(Z_t, B_{OLS}, \hat{F}\right)$ converges in quadratic mean to $s\left(Z_t, B_{OLS}, F\right)$, then

$$B_{AD} = B_{OLS} + n^{-1} \sum_{t=1}^{n} s\left(Z_t, B_{OLS}, \hat{F}\right)$$

is an adaptive estimator of $\beta$ for the family $\mathcal{F}$.

For the linear regression model, as for numerous other models, nonparametric estimation of $F$ entails nonparametric estimation of the density $f$. One popular nonparametric density estimator is the kernel estimator, which is employed by Portnoy and Koenker (1989) in their proof that semiparametric quantile estimators are also adaptive for $\beta$. If $\left\{\hat{U}_t\right\}$ denotes the OLS residuals, then a kernel density estimator is defined for all $u$ in a small neighborhood of each value of $\hat{U}_t$ as

$$\hat{f}_t\left(u\right) = \left(n-1\right)^{-1} \sum_{\substack{s=1 \\ s \neq t}}^{n} \xi_\sigma\left(u - \hat{U}_s\right),$$

where $\xi_\sigma$ is a weight function that depends on the smoothing parameter $\sigma$. In Steigerwald (1992), $\xi_\sigma$ corresponds to a Gaussian density with mean 0 and variance $\sigma^2$. The variance controls the amount of smoothing, as $\sigma^2$ declines the weight given to residuals that lie some distance from $\hat{U}_t$ tends to zero. Of course, there are many other ways to form the nonparametric score estimator. Newey (1988) approximates the score by a series of moment conditions, which arise from exogeneity of the regressor and symmetry of $F$. Faraway (1992) uses a series of spline functions to approximate the score. Chicken and Cai (2005) use wavelets to form the basis for nonparametric estimation of $f$.

Recent results in adaptive estimation have focused on problems in which the error distribution is known, but other features are modeled nonparametrically. Some of the most intriguing results concern the type of stochastic differential equation often encountered in financial models. The price of an asset that is measured continuously over time, $P_t$, is often modeled as

$$dP_t = f_t dt + \upsilon_t dB_t.$$

The presence of standard Brownian motion, $B_t$, makes the model of price a stochastic differential equation. The function $f_t$ captures the deterministic movement or drift while $\upsilon_t$ is the potentially time-varying scale of the random component. Lepski and Spokoiny (1997) study the model in which $\upsilon_t$ is constant and $f_t$ is unknown. They establish that a nonparametric estimator of $f$ is pointwise adaptive. Yet an estimator that is pointwise adaptive, that is for a given point $t_0$ the nonparametric estimator of $f(t_0)$ is asymptotically efficient, may not perform well for all values within the range of the function $f$. Such an idea is intuitive, without knowledge of the smoothness of $f$, estimators designed to be optimal for one value of $t$ may be very different from optimal estimators for another value of $t$. Cai and Low (2005) study efficient estimation of $f$ over neighborhoods of $t_0$ and show that an estimator constructed from wavelets is adaptive. The restriction that the scale is constant is often difficult to support with financial data. A more realistic model, which Mercurio and Spokoiny (2004) study, models the asset return as a stochastic differential equation with drift 0 and $\nu_t$ varying over time. The time-varying scale is assumed to be constant over (short) intervals of time, but is otherwise unspecified. They construct a nonparametric estimator of the volatility from a kernel that performs local averaging and show that the resultant estimator is adaptive.

Douglas G. Steigerwald

## Bibliography


Bickel, P. 1982. On adaptive estimation. Annals of Statistics 10, 647-671.

Cai. T. and M. Low 2005. Nonparametric estimation over shrinking neighborhoods: superefficiency and adaptation. Annals of Statistics 33, 184-213.

Chicken, E. and T. Cai 2005. Block thresholding for density estimation: local and global adaptivity. Journal of Multivariate Analysis 95, 76-106.

Faraway, J. 1992. Smoothing in adaptive estimation. Annals of Statistics 20, 414-427.

Lepski, O. and V. Spokoiny 1997. Optimal pointwise adaptive methods in nonparametric estimation. Annals of Statistics 25, 2512-2546.

Mercurio, D. and V. Spokoiny 2004. Statistical inference for time-inhomogeneous volatility models. Annals of Statistics 32, 577-602.

Newey, W. 1988. Adaptive estimation of regression models via moment restrictions. Journal of Econometrics 38, 301-339.

Portnoy, S. and R. Koenker 1989. Adaptive L-estimation for linear models. Annals of Statistics 17, 362-381.

Steigerwald, D. 1992. On the finite sample behavior of adaptive estimators. Journal of Econometrics 54, 371-400.

Stein, C. 1956. Efficient nonparametric testing and estimation, in: J. Neyman ed. Proceedings of the third Berkeley symposium on mathematical statistics and probability (University of California Press, Berkeley, CA) 187-195.