Chapter 3 Asymptotic Results for MRR1

In this chapter we will develop asymptotic results for the MRR1 estimate using various schemes to select the mixing parameter λ. It is important to note the critical role the mixing parameter plays in the asymptotic quality of the final estimate. Burman and Chaudhuri (1992) presented an estimate similar to the MRR1 estimate of Einsporn and Birch (1987) along with some basic asymptotic results. Their paper is the subject of section 3a and is being utilized as a "benchmark" for the work in this dissertation. For the purpose of comparison we will utilize their notation throughout the next three chapters.

Overview of Conventions

Again, we will write the model as

$$Y_i = \boldsymbol{q}(\mathbf{x}_i) + \boldsymbol{e}_i \text{ for } i = 1 \text{ to } n.$$

We will require the $\boldsymbol{e}_i$'s to be i.i.d. with $E(\boldsymbol{e}_i) = 0$ and $V(\boldsymbol{e}_i) = \boldsymbol{s}^2$. We will assume that the $\mathbf{x}_i$'s are fixed uniformly on a compact set C in $\Re^{p+1}$. We will also assume that $\boldsymbol{q}(\mathbf{x}_i)$ is continuous on C. The ordered pairs $(\mathbf{x}_i, Y_i)$ form the observations from which two estimates are derived. Throughout the next three chapters we will be dealing with univariate regression, but will often write $\mathbf{x}_i$ in vector notation to indicate the p + 1 dimensional vector $(1, x_i, x_i^2, \ldots x_i^p)$. In this dissertation the term "asymptotic" means that the number of observations increases without bound, and the manifestation of these observations in the X space increases uniformly in C. Thus, we have a fixed effects picture, as $n \to \infty$, over a uniform design. As in chapter 2, the two estimates used in forming the MRR1 estimate are the parametric estimate and the nonparametric estimate, designated $\hat{f}, \hat{g}$ respectively.

We also assume that

$$\hat{g} = n^{-1}\sum_{i=1}^{n} W_2(.,x_i)Y_i$$

for some weight function W_2, where the weight function W_2 satisfies

$$n^{-1}\sum_{i=1}^{n} W_2(.,x_i) = 1.$$

These forms cover most all forms of nonparametric approximations including local polynomial estimates and spline estimates (see Hardle (1990), and Fan and Gijbels (1996)).

The parametric and nonparametric estimates are both based on the original data, and, after obtaining each separately, are combined to form

$$\lambda\hat{f}(\mathbf{x}_i) + (1-\lambda)\hat{g}(\mathbf{x}_i) = \hat{\theta}(\mathbf{x}_i)\text{ , for } i = 1 \text{ to } n$$

where λ is a constant chosen by various formulae.

We will be using cross validation estimates, and as such we will indicate the parametric and non-parametric estimates obtained by leaving out the ith observation, as $\hat{f}^{(i)}, \hat{g}^{(i)}$ respectively.

Also, in this dissertation, the bandwidth is designated by τ_n^{-1} and it is assumed $\tau_n \to \infty$, and $\lim_{n\to\infty}(\tau_n / n) \to 0$ with $\tau_n^{-1} \le O_P(n^{-.2})$ for p odd, and $\tau_n^{-1} \le O_P(n^{-\frac{1}{9}})$ for p even. We are not concerned with the speed with which the resulting bandwidth is obtained; only that the selection procedure satisfies the prerequisites above, and a couple of assumptions mentioned later.

Bandwidth selection methods for each n (see section 2b), and asymptotic performance are not addressed here. We will now turn our attention to the concept of distance measure, a critical issue when dealing with asymptotic rates of convergence.

For any two functions of $\mathbf{x}_i$, $\mathbf{h}_1$ and $\mathbf{h}_2$, we define the inner product and norm (respectively) as

$$\langle \mathbf{h}_1, \mathbf{h}_2 \rangle = n^{-1}\sum_{i=1}^{n} \mathbf{h}_1(\mathbf{x}_i)\mathbf{h}_2(\mathbf{x}_i) \text{ , and}$$

$$\|\mathbf{h}_1\|^2 = \langle \mathbf{h}_1, \mathbf{h}_1 \rangle, \text{ with } \|\mathbf{h}_1\| = (\langle \mathbf{h}_1, \mathbf{h}_1 \rangle)^{.5}.$$

This is the standard L_2 (Euclidean) norm. We also set forth a convention here for the entire discussion. We shall often write scalar type functions as arguments for the inner product and/or the norm. It is assumed, in that case, that the function represents the entire vector as defined over all n observations.

With this in mind, we will denote by δ_n, the distance between the unknown regression function, θ, and the parametric family of continuous (on C) regression functions under consideration.

$$\delta_n = \inf\{\|\theta - f(\beta)\| : \beta \in \Re^p\}.$$

If the infimum is attained at a unique β, we will designate this as β^* and write

$$\delta_n = \|\theta - f(\beta^*)\|.$$

Note that the subscript n indicates that this quantity is based on a sum of n realizations of the function θ and the parametric function f. But, as $n \to \infty$, $\delta_n \to \left(\int_C (\theta(x) - f(x, \beta^*))^2 dx\right)^{.5}$.

It is important to note that the parametric family of continuous functions under consideration is set a priori by the user. That is, p, the dimension of the vector β, is a constant. And so, for instance, if the user is convinced that θ can be modeled with a second degree polynomial, when in fact θ is actually cubic, then $\lim_{n\to\infty} \delta_n \neq 0$, even though θ can be modeled parametrically.

Note also that we restrict the user to a parametric family of continuous (on C) functions. This, coupled with the fact that $\theta(\mathbf{x}_i)$ is continuous on C, implies that δ_n must be finite.

Similarly we set γ_n as a distance measure for the nonparametric estimate,

$$\gamma_n^2 = \mathrm{E}(\|\hat{g}(\mathbf{x}_i) - \theta\|^2).$$

Notice that this is precisely the AVEMSE $\left(\sum MSE(\hat{g}(\mathbf{x}_i))\right)$ from Mays (1995).

The terms δ_n, γ_n represent the convergence distances for the parametric and nonparametric estimates, respectively.

We also need to define what is meant by $O_P(b_n)$. Using the definitions from Bishop, Feinberg, and Holland (1975), a stochastic sequence $\{X_n\}$ is said to be $O_P(1)$ if for every $\eta > 0$, there exist constants $K(\eta)$ and $n(\eta)$, such that for $n \geq n(\eta)$,

$$P\{|X_n| \leq K(\eta)\} \geq 1 - \eta.$$

Second, a stochastic sequence $\{X_n\}$ is said to be $O_P(b_n)$ if $\{X_n/b_n\} = O_P(1)$. The authors go on to state that the stochastic order of a sequence is akin to the limiting standard deviation of the sequence. Hence, if $\{X_n\}$ is a stochastic sequence with $\mu_n = E(X_n)$ and $\sigma_n^2 = V(X_n) < \infty$, then

$$\{X_n - \mu_n\} = O_P(\sigma_n).$$

For example, by the Weak Law of Large Numbers, if a sample mean is constructed from repeated samples of increasing size n, from a population with any distribution having mean μ, we might write (using our developed notation).

$$\left|\{\overline{X}_n - \mu\}\right| = O_P(n^{-.5}) \text{ .[].}$$

Having the distance measures in hand, along with the definitions of stochastic convergence, we are now ready to apply these to the basic MRR1 estimate.

Part 3a Basic MRR1 Asymptotics

In finding the optimal data driven mixing parameter, it is necessary to determine exactly what is the ideal value for λ. In other words, what would the user select for λ if the true mean function were known.

By Burman and Chaudhuri (1992), it is easy to demonstrate by simple algebra that

$\| \lambda \hat{f}(\mathbf{x}_i) + (1-\lambda)\hat{g}(\mathbf{x}_i) - \theta(\mathbf{x}_i) \|$ attains a minimum at

$$\lambda^* = \frac{< \hat{f} - \hat{g}, \theta - \hat{g} >}{\left\| \hat{f} - \hat{g} \right\|^2} = \frac{< \hat{f} - \hat{g}, \theta > - < \hat{f} - \hat{g}, \hat{g} >}{\left\| \hat{f} - \hat{g} \right\|^2}.$$

Burman and Chaudhuri (1992) originally estimated the optimal mixing parameter by

$$\hat{\lambda}^* = \frac{< \hat{f}^{(i)} - \hat{g}^{(i)}, Y_i - \hat{f} > - < \hat{f} - \hat{g}, \hat{g} >}{\left\| \hat{f} - \hat{g} \right\|^2}.$$

We propose the following correction which we shall call $\hat{\lambda}^{*C}$. $\hat{\lambda}^{*C}$ =

$$\hat{\lambda}^* - \frac{< \hat{f}, \hat{g} - \hat{f} >}{\left\| \hat{f} - \hat{g} \right\|^2} = \frac{< \hat{f}^{(i)} - \hat{g}^{(i)}, Y_i - \hat{f} > - < \hat{f} - \hat{g}, \hat{g} - \hat{f} >}{\left\| \hat{f} - \hat{g} \right\|^2} = \frac{< \hat{f}^{(i)} - \hat{g}^{(i)}, Y_i - \hat{f} > + \left\| \hat{f} - \hat{g} \right\|^2}{\left\| \hat{f} - \hat{g} \right\|^2}$$

$$= \frac{< \hat{f}^{(i)} - \hat{g}^{(i)}, Y_i - \hat{f} >}{\left\| \hat{f} - \hat{g} \right\|^2} + 1.$$

This correction to Burman and Chaudhuri (1992) was necessary to have successive parts of one of their proofs (Lemma 5.3) coincide mathematically, and was also noted by Rahman, Ghokhale, and Ullah (1997). Notice that the estimate is formed by essentially replacing θ_i with Y_i and, in some cases, replacing the component estimates with their cross validated counterparts.

One aspect of finding asymptotic convergence rates is finding the rate at which the data driven mixing parameter converges to the theoretically optimal mixing parameter. This rate is instrumental in obtaining other, more general rates for the estimate. The estimate $\hat{\lambda}^{*C}$ yields the following asymptotic results:

$$\lambda^* - \hat{\lambda}^{*C} =$$

$$\lambda^* - \hat{\lambda}^* + \frac{< \hat{f}, \hat{g} - \hat{f} >}{\left\| \hat{f} - \hat{g} \right\|^2} = \left[n^{-1} \sum_{i=1}^{n} \{Y_i - \hat{f}\} \{\hat{g}^{(i)}(\mathbf{x}_i) - \hat{f}^{(i)}(\mathbf{x}_i)\} - < \theta, \hat{g} - \hat{f} > + < \hat{f}, \hat{g} - \hat{f} > \right] \left\| \hat{f} - \hat{g} \right\|^{-2}$$

So that within the brackets, we have

$$n^{-1} \sum_{i=1}^{n} \varepsilon_i \{\hat{g}^{(i)}(\mathbf{x}_i) - \hat{f}^{(i)}(\mathbf{x}_i)\} + n^{-1} \sum_{i=1}^{n} \{\theta(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\} \{\hat{g}^{(i)}(\mathbf{x}_i) - \hat{g}(\mathbf{x}_i)\}$$

$$- n^{-1} \sum_{i=1}^{n} \{\theta(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\} \{\hat{f}^{(i)}(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\}$$

$$= J_1 + J_2 - J_3 (say) .$$

Burman and Chaudhuri (1992) go on to demonstrate that

$$J_1 = O_P(n^{-.5} \gamma_n) + O_P(n^{-.5} \delta_n) ,$$

$$J_2 = O_P(\delta_n \gamma_n^{\,2}) + O_P(n^{-.5} \gamma_n^{\,2}) ,$$

$$J_3 = O_P(n^{-1.5}) + O_P(n^{-1} \delta_n) .$$

These results are used to prove Lemma 3.a.3 and part of Lemma 3.a.4 which appear later in this section. The details of the proofs for results in this section (with the exception of Theorem 3.A.2 which is in Appendix 3a) are found in Burman and Chaudhuri (1992). We are primarily interested in results in this section, and so we will move on to the assumptions.

The following assumptions (and variations on them) will be used throughout the rest of the dissertation.

A1. There exists a function W_1 of two variables which is defined and bounded on C x C where C is the predictor space, and

$$\| \hat{f}(\hat{\beta}, .) - f(\beta^*, .) - n^{-1} \sum_{i=1}^{n} W_1(., x_i) e_i \| = O_P(n^{-1}) .$$

A2. $$\frac{\left\| \hat{g} - \theta \right\|^2 - E(\left\| \hat{g} - \theta \right\|^2)}{E(\left\| \hat{g} - \theta \right\|^2)} \xrightarrow{P} 0, \text{ as } n \to \infty .$$

A3. There exist $c_1 > 0$ and $c_2 > 0$, such that for all i, $c_1 \tau_n \le W_2(x_i, x_i) \le c_2 \tau_n$.

A4. There exists a constant $c > 0$, such that $n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_2^2(x_i, x_j) \le c \tau_n$.

A5. $\lim_{n \to \infty} \frac{1}{\gamma_n \sqrt{n}} = 0$.

A6. $\hat{g}^{(i)}(x_i) = \frac{\sum_{j:j \ne i} W_2(x_i, x_j) Y_j}{\sum_{j:j \ne i} W_2(x_i, x_j)}$.

Some comments on the assumptions follow. A1 gives a distance measure between the optimal parametric estimate and the given parametric estimate in terms of a weighting function that is independent of ε_i. It is trivially satisfied for the case where $\hat{f}(., \beta)$ is linear in β, and can be satisfied for non-linear cases via regularity conditions on $f(\beta, .)$. Also, in A1 we assume that β^* represents an optimal β (if not the optimal β). A2 guarantees that $\| \hat{g} - \theta \| = O_P(\gamma_n)$, while A5 implies that the nonparametric estimate converges at a rate slower than $n^{-.5}$. The fact that the parametric estimate (for the model presented in this chapter) converges at a rate of $n^{-.5}$ (provided $\delta_n = 0$), and the nonparametric estimate converges at a slower rate (but converges ($\lim_{n \to \infty} (\gamma_n) = 0$)) was shown to be the case by Stone (1982) and Olkin and Spiegelman (1987).

A3 and A4 assure us that the weighting function (based on the kernel) does not place too much or too little weight on any one observation, and that our bandwidth is not infinitely large, respectively. A6 is automatically satisfied for local polynomial and spline smoothing estimates, and is necessary (along with A4) for the proof of the aforementioned Lemmas 3.a.3 and 3.a.4 (which we shall see presently). Assumptions A1, A3, A4 are essential in the use of Whittle's Inequality (Whittle(1960)) in the proofs of the Lemmas on asymptotic convergence of the mixing parameter, both in this section and later chapters (particularly Chapter 5). Whittle's Inequality is presented in appendix 3a.

We now present the following technical results and theorems. The lemmas serve to construct proofs of the theorems that follow in this discussion. As such they deal with very basic parts of the estimate. We prove Theorem 3.A.2 in appendix 3a, since this proof contains an important equation which will be useful later. Throughout the dissertation we will be dealing with similar lemmas and theorems, hence we will provide comments here that will apply for each section.

The first three lemmas deal with the dichotomy that the parametric estimate is correct ($\delta_n = 0$) versus the parametric estimate is incorrect ($\lim_{n\to\infty} \delta_n \neq 0$). Lemma 3.a.4 addresses the situation in which the parametric model becomes correct as the sample size increases ($\lim_{n\to\infty} \delta_n = 0$).

Lemma 3.a.1: Assuming conditions A1, A2, and A5…

$$\left\| \hat{f} - \hat{g} \right\| = \begin{cases} O_P(1), & \text{if } \lim_{n\to\infty} \delta_n \neq 0 \\ O_P(\gamma_n), & \text{if } \delta_n = 0 \end{cases}$$

Lemma 3.a.1 deals with the proximity of the two estimates. As expected, if the parametric estimate is not correct, the two will not converge. If the parametric estimate is correct then the two converge at the rate of the slower nonparametric estimate. The proof is found in Burman and Chaudhuri (1992).

Lemma 3.a.2: Assuming conditions A1, A2, and A5…

$$\lambda^* = \begin{cases} O_P(\gamma_n), & \text{if } \lim_{n\to\infty} \delta_n \neq 0 \\ 1 + O_P(n^{-.5}\gamma_n^{-1}), & \text{if } \delta_n = 0 \end{cases}.$$

Lemma 3.a.2 gives asymptotic rates of convergence for the optimal mixing parameter.

Note that λ^* converges to 1 when the parametric model is correct because of conditions A2

and A5. The convergence of the mixing parameter to its optimum is obtained whether the model is correct or incorrect. The proof is found in Burman and Chaudhuri (1992).

Lemma 3.a.3: Assuming conditions A1-A6 hold…

$$\lambda^* - \hat{\lambda}^{*c} = \begin{cases} O_P(n^{-.5}) + O_P(\gamma_n^2), & \text{if } \lim_{n\to\infty} \delta_n \neq 0 \\ O_P(n^{-.5}\gamma_n^{-1}), & \text{if } \delta_n = 0 \end{cases}.$$

This is the all important mixing parameter lemma. To date, this provides the optimal dichotomous rates of convergence. Notice that in the first instance, two terms must be listed since we cannot tell which will be the slowest (i.e. $O_P(\gamma_n)$ is unknown to some extent). Lemma 3.a.3 is critical in the proofs of the theorems that follow. Its proof is found in Burman and Chaudhuri (1992).

Lemma 3.a.4 addresses each situation involving the comparison of the parametric and nonparametric rates of convergence. Similar lemmas will be listed in subsequent sections. The proof of Lemma 3.a.4 is found in Burman and Chaudhuri (1992).

Lemma 3.a.4: Assume that $\lim_{n\to\infty} \delta_n = 0$.

a) Under assumptions A1, A2, and A5 we have

$$\left\| \hat{f} - \hat{g} \right\| = \begin{cases} O_P(\delta_n), & \text{if } \dfrac{\delta_n}{\gamma_n} > 1 \\ O_P(\gamma_n), & \text{if } \dfrac{\delta_n}{\gamma_n} < 1 \end{cases},$$

b) Under assumptions A1-A6, we have

$$\lambda^* - \hat{\lambda}^{*c} = \begin{cases} O_P(n^{-.5}\delta_n^{-1}) + O_P(\gamma_n^{\,2}\delta_n^{-1}), & \text{if } \dfrac{\delta_n}{\gamma_n} > 1 \\ O_P(n^{-.5}\gamma_n^{\,-1}) + O_P(\delta_n), & \text{if } \dfrac{\delta_n}{\gamma_n} < 1 \end{cases}.$$

The following theorems deal with the estimate convergence rates. The first two deal with the dichotomous scenario mentioned earlier while the last two deal with the instance in which the parametric model becomes correct for large n. The first and third concern the theoretically optimal mixing parameter. The second and fourth deal with the asymptotically optimal data driven mixing parameter. And each one deals with how quickly the MRR1 estimate converges to the true mean function in the L_2 norm as $n \to \infty$. The proof for Theorem 3.A.1 is found in Burman and Chaudhuri (1992). The proof for Theorem 3.A.2 is found there as well, although an expanded form can be found in appendix 3a.

Theorem 3.A.1: Assuming conditions A1-A6 hold...

$$\left\| \lambda^* \hat{f} + (1-\lambda^*)\hat{g} - \theta \right\| = \begin{cases} O_P(\gamma_n), & \text{if } \lim_{n\to\infty} \delta_n \neq 0 \\ O_P(n^{-.5}), & \text{if } \delta_n = 0 \end{cases}$$

The result of Theorem 3.A.1 is artificial in that it involves the theoretical asymptotically optimal mixing parameter. However, it does provide a benchmark for the MRR1 estimate. The question remains as to how fast an estimate with a data driven mixing parameter can attain its optimum. As indicated earlier, Lemma 3.a.3 plays a large role in answering that question in the form of the next theorem.

Theorem 3.A.2: Assuming conditions A1-A6 hold...

$$\left\| \hat{\lambda}^{*C} \hat{f} + (1-\hat{\lambda}^{*C})\hat{g} - \theta \right\| = \begin{cases} O_P(\gamma_n), & \text{if } \lim_{n\to\infty} \delta_n \neq 0 \\ O_P(n^{-.5}), & \text{if } \delta_n = 0 \end{cases}$$

Theorem 3.A.2 is the “Golden Result of Model Robust Regression”. It provides a result that is both useful and striking. The user obtains the best rate of convergence available by using the MRR1 estimate, regardless of whether or not the model is correctly specified.

We will demonstrate the convergence rates of MRR1 with an example. Suppose a user is estimating a function q by using MRR1 and attempting to model the function parametrically with an OLS quadratic regression and nonparametrically by a local linear regression (LLR) using the asymptotically optimal constant bandwidth, h_{OPT}, from p. 68 of Fan and Gijbels (1996) (in spite of the fact that the user would be unlikely to utilize such a bandwidth, we will present this here since it sets the stage for future examples. Hardle, Hall and Marron (1988) present evidence that, for at least the local constant regression, the plug in estimates for the optimal bandwidth have the same rate of convergence). We will also use the Epanechnikov kernel in the nonparametric estimate and $\hat{l}^{*C}$ for the mixing parameter. From Ruppert and Wand (1994) we have that at any given $\mathbf{x}$ in C, the convergence rate of the LQR estimate is given by

$$\left|\hat{g}(x)-\boldsymbol{q}(x)\right|^2 = O_P(h_{OPT}^4)+O_P(n^{-1}h_{OPT}^{-1})$$

where

$$h_{OPT} = o_P(n^{-.2}).$$

Note that $o_P(b_n)$ is a stronger form of stochastic convergence than $O_P(b_n)$ (for further information on this see Bishop, Feinberg, and Holland (1975)). We then obtain

$$\left|\hat{g}(x)-\boldsymbol{q}(x)\right|^2 = O_P(n^{-.2}).$$

Next, we extend this result to the n dimensional nonparametric vector estimate. This is a simple extension which we shall not formalize here (for a rigorous presentation see the proof of Lemma 5.a.1 in appendix 5a). The extension results in

$$g_n^2 = O_P(n^{-.4}).$$

Asymptotically, then, the user has an estimate such that

$$\left\| \hat{l}^{*C} \hat{f} + (1 - \hat{l}^{*C})\hat{g} - \boldsymbol{q} \right\| = \begin{cases} O_P(n^{-.2}), & \text{if } \lim_{n\to\infty} \boldsymbol{d}_n \neq 0 \\ O_P(n^{-.5}), & \text{if } \boldsymbol{d}_n = 0 \end{cases},$$

and will thus converge to the true mean function at a rate no slower than $O_P(n^{-.2})$ if the model is misspecified, and as fast as $O_P(n^{-.5})$ if $\boldsymbol{q}(\mathbf{x})$ is truly a quadratic function on C.[].

Notice in the next two theorems that there is now a third condition based on whether or not the optimal parametric function converges to the true mean function faster than $n^{-.5}$. As before, the first deals with the artificial situation in which the theoretical asymptotically optimal mixing parameter is known, while the second deals with the asymptotically optimal data driven mixing parameter. Once again the proofs for both theorems are found in Burman and Chaudhuri (1992).

Theorem 3.A.3: Assuming conditions A1-A6 hold, and that $\lim_{n\to\infty} \boldsymbol{d}_n = 0$ ...

$$\left\| l^{*} \hat{f} + (1 - l^{*})\hat{g} - \boldsymbol{q} \right\| = \begin{cases} O_P(\boldsymbol{g}_n), & \text{if } \dfrac{\boldsymbol{d}_n}{\boldsymbol{g}_n} > 1 \\ O_P(\boldsymbol{d}_n), & \text{if } \dfrac{n^{-.5}}{\boldsymbol{g}_n} < \dfrac{\boldsymbol{d}_n}{\boldsymbol{g}_n} < 1 \\ O_P(n^{-.5}), & \text{if } \dfrac{\boldsymbol{d}_n}{\boldsymbol{g}_n} < \dfrac{n^{-.5}}{\boldsymbol{g}_n} \end{cases}$$

Theorem 3.A.4: Assuming conditions A1-A6 hold, and that $\lim_{n\to\infty} \boldsymbol{d}_n = 0$ ...

$$\left\| \hat{l}^{*C} \hat{f} + (1 - \hat{l}^{*C})\hat{g} - \boldsymbol{q} \right\| = \begin{cases} O_P(\boldsymbol{g}_n), & \text{if } \dfrac{\boldsymbol{d}_n}{\boldsymbol{g}_n} > 1 \\ O_P(\boldsymbol{d}_n), & \text{if } \dfrac{n^{-.5}}{\boldsymbol{g}_n} < \dfrac{\boldsymbol{d}_n}{\boldsymbol{g}_n} < 1 \\ O_P(n^{-.5}), & \text{if } \dfrac{\boldsymbol{d}_n}{\boldsymbol{g}_n} < \dfrac{n^{-.5}}{\boldsymbol{g}_n} \end{cases}$$

Note that the data driven mixing parameter performs as well as the theoretical asymptotically optimal mixing parameter. Also note that the convergence rate for the estimate is dependent upon how fast the optimal parametric model converges to q, and that d_n is the convergence rate in the middle instance.

Concluding Remarks

These results demonstrate one of the main advantages of MRR1. That is, if the user's model is incorrect, MRR1 can achieve a consistent estimate at the asymptotic convergence rate of the nonparametric estimate. On the other hand, if the model is correct, MRR1 achieves consistency at the parametric rate, faster than a purely nonparametric estimate. It is also interesting to note that the estimate with the asymptotically optimal data driven mixing parameter $\hat{l}^{*C}$ performs just as well as the estimate with the asymptotically optimal theoretical mixing parameter l^*. This phenomenon occurs as long as we do not invoke too much cross validation in the computation. We now turn our attention to the performance of the MRR1 estimate utilizing $\hat{l}^P$, the data driven mixing parameter chosen via the PRESS statistic.