

### Part 3b Asymptotic Results for MRR1 using PRESS

The PRESS statistic is a special type of cross validation procedure (see Allen (1971))

particular to the regression problem and is given by  $\sum_{i=1}^n (Y_i - \hat{Y}_{i,-i})^2$  where  $\hat{Y}_{i,-i}$  is

the estimate at the  $i$ th observation found by removing the  $i$ th data pair,  $(Y_i, \mathbf{x}_i)$  from

the data set. Here  $\text{PRESS}(\mathbf{I}) = \sum_{i=1}^n (Y_i - (\mathbf{I}\hat{f}^{(i)} + (1-\mathbf{I})\hat{g}^{(i)}))^2$ . One choice for  $\lambda$ , is to find

that value of  $\lambda$  that minimizes  $\text{PRESS}(\mathbf{I})$ . This is done by setting  $\frac{d}{d\mathbf{I}} \text{PRESS}(\mathbf{I}) =$

$\text{PRESS}'(\mathbf{I}) = 0$  and solving for  $\mathbf{I}$ . Then  $\text{PRESS}'(\mathbf{I}) =$

$$\begin{aligned} & \sum_{i=1}^n 2(Y_i - (\mathbf{I}\hat{f}^{(i)} + (1-\mathbf{I})\hat{g}^{(i)}))(\hat{g}^{(i)} - \hat{f}^{(i)}) \\ &= \sum_{i=1}^n (Y_i(\hat{g}^{(i)} - \hat{f}^{(i)}) - \hat{g}^{(i)2} - \hat{f}^{(i)}\hat{g}^{(i)} + \mathbf{I}(\hat{f}^{(i)} - \hat{g}^{(i)})^2) \\ &= \sum_{i=1}^n (\hat{g}^{(i)} - \hat{f}^{(i)})(Y_i - \hat{g}^{(i)}) + \mathbf{I}(\hat{f}^{(i)} - \hat{g}^{(i)})^2 = 0. \end{aligned}$$

Solving this equation results in  $\hat{\mathbf{I}}^P$  as

$$\hat{\mathbf{I}}^P = \frac{\sum_{i=1}^n (Y_i - \hat{g}^{(i)})(\hat{f}^{(i)} - \hat{g}^{(i)})}{\sum_{i=1}^n (\hat{f}^{(i)} - \hat{g}^{(i)})^2} = \frac{\langle (Y_i - \hat{g}^{(i)}), (\hat{f}^{(i)} - \hat{g}^{(i)}) \rangle}{\|\hat{f}^{(i)} - \hat{g}^{(i)}\|^2}.$$

Of course we must ensure that  $\text{PRESS}''(\mathbf{I}) > 0$ . This follows from

$$\text{PRESS}''(\mathbf{I}) = \frac{\sum_{i=1}^n (\hat{f}^{(i)} - \hat{g}^{(i)})^2}{n} > 0, \forall \mathbf{I}, n.$$

Thus  $\hat{\mathbf{I}}^P$  does, in fact, produce a global minimum. And it is this estimate that we will investigate asymptotically in the remainder of this section.

We will investigate asymptotically the difference between  $\mathbf{I}^*$  and  $\hat{\mathbf{I}}^P$ , which is given by

$$\mathbf{I}^* - \hat{\mathbf{I}}^P = \frac{\langle \hat{g} - \mathbf{q}, \hat{g} - \hat{f} \rangle}{\|\hat{f} - \hat{g}\|^2} - \frac{\langle Y_i - \hat{g}^{(i)}, \hat{f}^{(i)} - \hat{g}^{(i)} \rangle}{\|\hat{f}^{(i)} - \hat{g}^{(i)}\|^2} \quad (3.B.1)$$

$$= \frac{\sum(\hat{g} - \mathbf{q})(\hat{g} - \hat{f})}{n(\|\hat{f} - \hat{g}\|^2)} - \frac{\sum(Y_i - \hat{g}^{(i)})(\hat{f}^{(i)} - \hat{g}^{(i)})\sum(\hat{f} - \hat{g})^2}{n^2\|\hat{f}^{(i)} - \hat{g}^{(i)}\|^2\|\hat{f} - \hat{g}\|^2}. \quad (3.B.2)$$

We would like to form a common denominator, but in order to do so we need to know the asymptotic rates for  $\|\hat{f}^{(i)} - \hat{f}\|$  and  $\|\hat{g}^{(i)} - \hat{g}\|$ . Via some simple algebra combined with Burman and Chaudhuri (1992) results 6.21, and 6.20 respectively, we have that these norms are  $O_p(n^{-1})$ , and  $O_p(\mathbf{g}_i^2)$  respectively. Next, set  $\alpha =$

$$(\|\hat{f}^{(i)} - \hat{g}^{(i)}\|^2 - \|\hat{f} - \hat{g}\|^2) = \frac{\sum((\hat{f}^{(i)} - \hat{g}^{(i)})^2 - (\hat{f} - \hat{g})^2)}{n}.$$

Information on this difference is needed to deal with the related discrepancy in the denominators of 3.B.1 and 3.B.2. So we need the following lemma and its corollary.

**Lemma 3.b.1:** Assuming conditions A1-A6 hold...

$$\alpha = \begin{cases} O_p(\mathbf{g}_n^2), & \text{if } \lim_{n \rightarrow \infty} \mathbf{d}_n \neq 0 \\ O_p(\mathbf{g}_n^3), & \text{if } \mathbf{d}_n = 0 \end{cases}.$$

**Corollary 3.b.1:** Assuming conditions A1-A6...

$$\|\hat{f}^{(i)} - \hat{g}^{(i)}\| - \|\hat{f} - \hat{g}\| = \begin{cases} O_p(\mathbf{g}_n), & \text{if } \lim_{n \rightarrow \infty} \mathbf{d}_n \neq 0 \\ O_p(\mathbf{g}_n^{1.5}), & \text{if } \mathbf{d}_n = 0 \end{cases}$$

Lemma 3.b.1 gives us the rate of convergence for  $\alpha$ , while Corollary 3.b.1 translates the result into a usable form for subsequent results. An important artifact of this lemma is that  $\alpha$  converges to zero faster than  $\|\hat{f} - \hat{g}\|$ . Thus we can interchange this difference and the cross validated difference (particularly in denominators) without penalty. The proofs for both results are found in appendix 3b.

Rewriting the right hand term in 3.B.2 we have

$$\frac{\sum(Y_i - \hat{g}^{(i)})(\hat{f}^{(i)} - \hat{g}^{(i)})\sum(\hat{f} - \hat{g})^2}{n^2\|\hat{f} - \hat{g}\|^2(\|\hat{f} - \hat{g}\|^2 + (\|\hat{f}^{(i)} - \hat{g}^{(i)}\|^2 - \|\hat{f} - \hat{g}\|^2))}$$

$$\begin{aligned}
&= \frac{\sum (Y_i - \hat{g}^{(i)})(\hat{f}^{(i)} - \hat{g}^{(i)})}{n(\|\hat{f} - \hat{g}\|^2 + \mathbf{a})} \\
&= \frac{\sum (Y_i - \hat{g}^{(i)})(\hat{f}^{(i)} - \hat{g}^{(i)})}{n(\|\hat{f} - \hat{g}\|^2)} \left( 1 - \frac{\mathbf{a}}{\|\hat{f} - \hat{g}\|^2 + \mathbf{a}} \right) \\
&= \frac{\sum (Y_i - \hat{g}^{(i)})(\hat{f}^{(i)} - \hat{g}^{(i)})}{n^2(\|\hat{f} - \hat{g}\|^2)} - \frac{\sum (Y_i - \hat{g}^{(i)})(\hat{f}^{(i)} - \hat{g}^{(i)})}{n(\|\hat{f} - \hat{g}\|^2)} \left( \frac{\mathbf{a}}{\|\hat{f} - \hat{g}\|^2 + \mathbf{a}} \right).
\end{aligned}$$

The left part of the last term is exactly what we need to complete the problem. The right part, however, we need to investigate, and shall call it remainder term 1 ( $R1$ ). We have the following lemmas for  $R1$ .

**Lemma 3.b.2:** Assuming conditions A1-A6...

$$\frac{\sum (Y_i - \hat{g}^{(i)})(\hat{f}^{(i)} - \hat{g}^{(i)})}{n(\|\hat{f} - \hat{g}\|^2)} = O_p(1)$$

Lemma 3.b.2 provides an important result that can be used in the proof of Lemma 3.b.3.

Primarily it allows us to work only with the term involving  $\alpha$  in  $R1$ . Notice that the expression in

Lemma 3.b.2 is almost exactly (with the exception of the denominator)  $\hat{\mathbf{I}}^P$ . Thus we would

expect the result that is presented. Lemma 3.b.3 gives the final result concerning  $R1$ . The

proofs of both lemmas are found in appendix 3b.

**Lemma 3.b.3:** Assuming conditions A1-A6...

$$R1 = \begin{cases} O_p(\mathbf{g}_n^2), & \text{if } \lim_{n \rightarrow \infty} \mathbf{d}_n \neq 0 \\ O_p(\mathbf{g}_n), & \text{if } \mathbf{d}_n = 0 \end{cases}.$$

Then 3.B.1 becomes

$$\begin{aligned}
\mathbf{I}^* - \hat{\mathbf{I}}^P &= \frac{\sum (\hat{g} - \mathbf{q})(\hat{g} - \hat{f}) - \sum (Y_i - \hat{g}^{(i)})(\hat{f}^{(i)} - \hat{g}^{(i)})}{n(\|\hat{f} - \hat{g}\|^2)} + R1 \\
&= \frac{\sum (\hat{g}^2 - \mathbf{q}\hat{g} + \mathbf{q}\hat{f} - \hat{f}\hat{g}) - \sum (Y_i \hat{f}^{(i)} - \hat{g}^{(i)} \hat{f}^{(i)} - Y_i \hat{g}^{(i)} + \hat{g}^{(i)2})}{n(\|\hat{f} - \hat{g}\|^2)} + R1 \\
&= \mathbf{I}^* - \hat{\mathbf{I}}^{*C} + \frac{\sum (\hat{g}^{(i)} \hat{f}^{(i)} - \hat{f}\hat{g}) - \sum (\hat{f}\hat{f}^{(i)} - \hat{f}\hat{g}^{(i)}) + \sum (\hat{g}^2 - \hat{g}^{(i)2}) + \sum (\hat{f}^2 - \hat{f}\hat{g})}{n(\|\hat{f} - \hat{g}\|^2)} + R1 \\
&= \mathbf{I}^* - \hat{\mathbf{I}}^{*C} + R1 + \\
&\quad \frac{\sum ((\hat{g}^{(i)} \hat{f}^{(i)} - \hat{f}\hat{g}) - (\hat{f}\hat{f}^{(i)} - \hat{f}\hat{g}^{(i)}) + (\hat{g} - \hat{g}^{(i)})(\hat{g} + \hat{g}^{(i)}) + (\hat{f}^2 - \hat{f}\hat{g}))}{n(\|\hat{f} - \hat{g}\|^2)} \tag{3.B.3}
\end{aligned}$$

Next note that  $(\hat{g}^{(i)} \hat{f}^{(i)} - \hat{f}\hat{g})$  can be rewritten as  $\hat{g}(\hat{f}^{(i)} - \hat{f}) + \hat{f}^{(i)}(\hat{g}^{(i)} - \hat{g})$ . Then 3.B.3

$$\begin{aligned}
&= \mathbf{I}^* - \hat{\mathbf{I}}^{*C} + R1 + \\
&\quad \frac{\sum (\hat{g}(\hat{f}^{(i)} - \hat{f}) + \hat{f}^{(i)}(\hat{g}^{(i)} - \hat{g}) - \hat{f}(\hat{f}^{(i)} - \hat{g}^{(i)}) + (\hat{g} - \hat{g}^{(i)})(\hat{g} + \hat{g}^{(i)}) + \hat{f}(\hat{f} - \hat{g}))}{n(\|\hat{f} - \hat{g}\|^2)} \\
&= \mathbf{I}^* - \hat{\mathbf{I}}^{*C} + R1 + T3(\text{say}).
\end{aligned}$$

Note that  $T3 =$

$$\begin{aligned}
&\frac{\sum (\hat{g}(\hat{f}^{(i)} - \hat{f}) + (\hat{g} - \hat{g}^{(i)})(\hat{g} + \hat{g}^{(i)} - \hat{f}^{(i)}) + \hat{f}((\hat{f} - \hat{g}) - (\hat{f}^{(i)} - \hat{g}^{(i)})))}{n(\|\hat{f} - \hat{g}\|^2)} \\
&= \frac{\sum (\hat{g}(\hat{f}^{(i)} - \hat{f}) + (\hat{g} - \hat{g}^{(i)})((\hat{g} - \hat{f}) + (\hat{g}^{(i)} - \hat{f}^{(i)})) + \hat{f}(\hat{f} - \hat{f}^{(i)}))}{n(\|\hat{f} - \hat{g}\|^2)} \\
&= \frac{\sum ((\hat{g} - \hat{g}^{(i)})(2(\hat{g} - \hat{f}) + (\hat{g}^{(i)} - \hat{f}^{(i)}) - (\hat{g} - \hat{f}))}{n(\|\hat{f} - \hat{g}\|^2)} + \frac{\langle \hat{f} - \hat{f}^{(i)}, \hat{f} - \hat{g} \rangle}{\|\hat{f} - \hat{g}\|^2}.
\end{aligned}$$

So that 3.B.3

$$\leq |\mathbf{I}^* - \hat{\mathbf{I}}^{*C}| + |R1| + \frac{2\|\hat{g} - \hat{g}^{(i)}\| \|\hat{g} - \hat{f}\| + \|\hat{g} - \hat{g}^{(i)}\| \langle (\hat{g}^{(i)} - \hat{g}) + (\hat{f} - \hat{f}^{(i)}) \rangle}{\|\hat{f} - \hat{g}\|^2} + \frac{\|\hat{f} - \hat{f}^{(i)}\|}{\|\hat{f} - \hat{g}\|}$$

(by the Cauchy-Schwartz and Triangle Inequalities)

$$\leq |\mathbf{I}^* - \hat{\mathbf{I}}^{*C}| + |R1| + \frac{2\|\hat{g} - \hat{g}^{(i)}\| \|\hat{g} - \hat{f}\| + \|\hat{g} - \hat{g}^{(i)}\| (\|\hat{g}^{(i)} - \hat{g}\| + \|\hat{f} - \hat{f}^{(i)}\|)}{\|\hat{f} - \hat{g}\|^2} + \frac{\|\hat{f} - \hat{f}^{(i)}\|}{\|\hat{f} - \hat{g}\|}$$

(by the Triangle Inequality)

$$= |\mathbf{I}^* - \hat{\mathbf{I}}^{*C}| + |R1| + \frac{O_p(n^{-1} \mathbf{g}_n^2) + O_p(\mathbf{g}_n^4)}{\|\hat{f} - \hat{g}\|^2} + \frac{O_p(n^{-1}) + O_p(\mathbf{g}_n^2)}{\|\hat{f} - \hat{g}\|} \quad (3.B.4)$$

by Burman and Chaudhuri (1992) results 6.20 and 6.21. We will use this result to develop the following general asymptotic results for the MRR1 estimate using the PRESS selected mixing parameter,  $\hat{\mathbf{I}}^P$ .

**Lemma 3.b.4:** Assuming conditions A1-A6...

$$\hat{\mathbf{I}}^P - \mathbf{I}^* = \begin{cases} O_p(n^{-5}) + O_p(\mathbf{g}_n^2), & \text{if } \lim_{n \rightarrow \infty} \mathbf{d}_n \neq 0 \\ O_p(n^{-5} \mathbf{g}_n^{-1}) + O_p(\mathbf{g}_n), & \text{if } \mathbf{d}_n = 0 \end{cases}$$

Lemma 3.b.4 gives the dichotomous rates of convergence for the PRESS selected mixing parameter (to the theoretical asymptotically optimal mixing parameter). Notice that it does not converge as quickly as  $\hat{\mathbf{I}}^{*C}$ . Also observe that this lemma corresponds to Lemma 3.a.3 in the previous section. The proof of Lemma 3.b.4 is found in appendix 3b.

Before we present the theorems, we will need the following result. Lemma 3.b.5 serves as a multi-purpose result for the case in which  $\lim_{n \rightarrow \infty} \mathbf{d}_n = 0$ . It is necessary in the proof of Theorem 3.B.4. The proof of Lemma 3.b.5 is found in appendix 3b.

**Lemma 3.b.5:** Assuming conditions A1-A6 hold, and that  $\lim_{n \rightarrow \infty} \mathbf{d}_n = 0 \dots$

$$\begin{aligned}
 \text{a) } \mathbf{a} &= \begin{cases} O_p(\mathbf{g}_n^2 \mathbf{d}_n), & \text{if } \frac{\mathbf{d}_n}{\mathbf{g}_n} > 1 \\ O_p(\mathbf{g}_n^3), & \text{if } \frac{\mathbf{d}_n}{\mathbf{g}_n} < 1 \end{cases} \\
 \text{b) } R1 &= \begin{cases} O_p(\mathbf{g}_n^2 \mathbf{d}_n^{-1}), & \text{if } \frac{\mathbf{d}_n}{\mathbf{g}_n} > 1 \\ O_p(\mathbf{g}_n), & \text{if } \frac{\mathbf{d}_n}{\mathbf{g}_n} < 1 \end{cases} \\
 \text{c) } \hat{\mathbf{I}}^P - \mathbf{I}^* &= \begin{cases} O_p(n^{-.5} \mathbf{d}_n^{-1}) + O_p(\mathbf{g}_n^2 \mathbf{d}_n^{-1}), & \text{if } \frac{\mathbf{d}_n}{\mathbf{g}_n} > 1 \\ O_p(n^{-.5} \mathbf{g}_n^{-1}) + O_p(\mathbf{g}_n), & \text{if } \frac{\mathbf{d}_n}{\mathbf{g}_n} < 1 \end{cases} .
 \end{aligned}$$

The first two parts of Lemma 3.b.5 deal with the asymptotics of the extraneous terms in 3.B.2, and are necessary in obtaining the asymptotic convergence rates for the mixing parameters in the situation where  $\lim_{n \rightarrow \infty} \mathbf{d}_n = 0$ , given in the third part. Again, the following theorems give the most important results; the asymptotic convergence rates for the MRR1 estimate (using the mixing parameter selected via PRESS) to the true function  $\mathbf{q}$

**Theorem 3.B.2:** Assuming conditions A1-A6 hold...

$$\|\hat{\mathbf{I}}^P \hat{f} + (1 - \hat{\mathbf{I}}^P) \hat{g} - \mathbf{q}\| = \begin{cases} O_p(\mathbf{g}_n), & \text{if } \lim_{n \rightarrow \infty} \mathbf{d}_n \neq 0 \\ O_p(\mathbf{g}_n n^{-.25}) + O_p(n^{-.5}) + O_p(\mathbf{g}_n^2), & \text{if } \mathbf{d}_n = 0 \end{cases}$$

Observe that Theorem 3.B.2 gives a strong result, but one that is not quite as good as that of Theorem 3.A.2. In the case where the model is properly specified the user's asymptotic results are still dependent to a large extent on the nonparametric rate of convergence. This is an artifact of using cross validation in the selection of the mixing parameter. The proof of Theorem 3.B.2 is found in appendix 3b.

We will again demonstrate the convergence rates of the MRR1 estimate with an example.

Suppose a user is estimating a function  $\mathbf{q}$  by using MRR1 and attempting to model the function parametrically with an OLS cubic regression and nonparametrically by local linear regression (LLR). We will use the nearest neighbor bandwidth,  $h_{kn}$ , from p. 151 of Fan and Gijbels (1996) (in this case we will take  $k_n$  to be  $[n^{.5}]_{\text{gif}}$ ; that is, we will utilize  $[n^{.5}]_{\text{gif}}$  observations (as possible) above and below the  $\mathbf{x}$  of interest). We will also use the Epanechnikov Kernel in the nonparametric estimate and the PRESS selected  $\hat{\mathbf{I}}^P$  for the mixing parameter. From Ruppert and Wand (1994) we have that at any given  $\mathbf{x}$  in C, the convergence rate of the LLR estimate is given by

$$|\hat{g}(x) - \mathbf{q}(x)|^2 = O_p(h_{kn}^4) + O_p(n^{-1}h_{kn}^{-1})$$

where

$$h_{kn} = o_p\left(n^{-\frac{1}{2}}\right),$$

so that

$$|\hat{g}(x) - \mathbf{q}(x)|^2 = O_p(n^{-5}).$$

Next, we extend this result to the  $n$  dimensional nonparametric vector estimate. For a rigorous presentation of this extension see the proof of Lemma 5.a.1 in appendix 5a. This extension results in

$$\mathbf{g}_n^2 = O_p(n^{-5}),$$

so that asymptotically the user has an estimate such that

$$\|\hat{\mathbf{I}}^P \hat{f} + (1 - \hat{\mathbf{I}}^P) \hat{g} - \mathbf{q}\| = \begin{cases} O_p(n^{-2.5}), & \text{if } \lim_{n \rightarrow \infty} \mathbf{d}_n \neq 0 \\ O_p(n^{-5}), & \text{if } \mathbf{d}_n = 0 \end{cases},$$

and will thus converge to the true mean function at a rate no slower than  $O_p(n^{-2.5})$  if the model is misspecified, and as fast as  $O_p(n^{-5})$  if  $\mathbf{q}(\mathbf{x})$  is truly a cubic function on  $\mathbb{C}$ .[].

**Theorem 3.B.4:** Assuming conditions A1-A6 hold, and that  $\lim_{n \rightarrow \infty} \mathbf{d}_n = 0$ .

$$\|\hat{\mathbf{I}}^P \hat{f} + (1 - \hat{\mathbf{I}}^P) \hat{g} - \mathbf{q}\| = \begin{cases} O_p(\mathbf{g}_n), & \text{if } \frac{\mathbf{d}_n}{\mathbf{g}_n} > 1 \\ O_p(\mathbf{g}_n^2) + O_p(\mathbf{g}_n \mathbf{d}_n^{.5}) + O_p(\mathbf{d}_n), & \text{if } n^{-.5} \mathbf{g}_n^{-1} < \frac{\mathbf{d}_n}{\mathbf{g}_n} < 1. \\ O_p(\mathbf{g}_n^2) + O_p(\mathbf{g}_n n^{-.25}) + O_p(n^{-5}), & \text{if } \frac{\mathbf{d}_n}{\mathbf{g}_n} < n^{-.5} \mathbf{g}_n^{-1} \end{cases}.$$

Once more, the final theorem gives the asymptotic results for the MRR1 estimate using the PRESS selected mixing parameter in the case where  $\lim_{n \rightarrow \infty} \mathbf{d}_n = 0$ . The proof makes use of

Lemma 3.b.5 and is found in appendix 3b.

### Comments

The MRR1 estimate using the PRESS statistic possesses appealing asymptotic qualities, since it



does converge to  $\mathbf{q}$  in each case. The results indicate that this estimate, though consistent, is not as efficient as the MRR1 estimate with the data driven mixing parameter presented in part 3a.

The mixing parameter  $\hat{\mathbf{I}}^P$  tends to react somewhat slower to the situations where either  $\lim_{n \rightarrow \infty} \mathbf{d}_n = 0$ , or if  $\mathbf{d}_n = 0$ . In these situations, the parametric model is suitable asymptotically or for small samples respectively. Also, in both situations the user wants  $\hat{\mathbf{I}}^P$  to be as close to 1 as possible. Perhaps the problem lies in the denominator,  $\|\hat{f}^{(i)} - \hat{g}^{(i)}\|^2$  which is seemingly larger than that of the theoretically optimal mixing parameter estimate  $\hat{\mathbf{I}}^{*C}$ . At any rate, we feel that mathematically the real problem lies with the term  $\|\hat{g}^{(i)} - \hat{g}\|$ , the difference between the cross validated nonparametric estimate and its non cross validated counterpart. Clearly, this difference may not be resolved very quickly depending on the convergence rate  $\mathbf{g}$ . This probably stems from the fact that cross validation in the nonparametric case involves removing the most influential observation (at least in the case of the non-uniform kernel function) in the “realm” of the bandwidth. We are dealing with an ever increasing number of observations, but we must remember that even though the bandwidth is going to zero at a rate slower than  $n^{-1}$ , the fact that we are removing an influential observation in a decreasing neighborhood will still cause the rate of convergence of  $\|\hat{g}^{(i)} - \hat{g}\|$  to be relatively slow. According to personal communication with George Terrell, this is a problem in practice as well. However, we are comparing the results to those of Theorem 3.A.2, and in light of the strengths of that estimate, the PRESS version of MRR1 does quite well.

We turn our attention to the asymptotic investigation of  $\hat{\mathbf{I}}^{*P}$ , the mixing parameter chosen via a method similar to PRESS\*, and its performance in the MRR1 estimate.