Appendix C

Calculation of minimum value of test-MSE

In the present simulation study, average test-MSE is theoretically equal to $E_T\left[E_{X,Y}\left(Y-\hat{Y}\right)^2\right]$, where the outer expectation refers to the variability due to using different training sets (denoted by T) to build the models, while the inner expectation refers to the variability of the predictor variable X, as well as the variability of the "dependent" variable Y. Some derivations give:

$$E_T\left[E_{X,Y}\left(Y-\hat{Y}\right)^2\right]=E_T\left[E_{X,\varepsilon}\left(f(X)+\varepsilon-\hat{f}(X)\right)^2\right]=E_T\left[E_X\left(E_\varepsilon\left(f(x_0)+\varepsilon-\hat{f}(x_0)\,|\,X=x_0\right)^2\right)\right]=$$
$$E_T\left[E_X\left(E_\varepsilon\left(f(x_0)+\varepsilon-\hat{f}(x_0)\,|\,X=x_0\right)^2\right)\right]=E_T\left[E_X\left[\left(f(X)-\hat{f}(X)\right)^2+E_\varepsilon(\varepsilon^2)\right]\right]=$$
$$E_T\left[E_X\left[\left(f(X)-\hat{f}(X)\right)^2\right]\right]+E_X\left[E_\varepsilon(\varepsilon^2)\right]=E_T\left[E_X\left[\left(f(X)-\hat{f}(X)\right)^2\right]\right]+E_X\left[Var(\varepsilon)\right]$$

The first term is reducible and depends on the accuracy of the model $\hat{f}(X)$ to approximate the true mean $f(X)$ of the outcome conditional on the predictor X. The second term $E_X\left\lfloor Var(\varepsilon)\right\rfloor$ is independent of the performance of the model and it depends only on the variability of the data.

Under DGM=1, the simulated utility score is given by U = -0.3+1.3*Y, with Y ~ beta(2+16*X, 2), where X ~ Uniform(0,1). Therefore, we have that $Var(\varepsilon)=Var(U)=1.3^2Var(Y)=1.69\dfrac{(2+16X)\cdot 2}{(4+16X)^2(4+16X+1)}$, based on the variance of the beta distribution.

Using simple Monte-Carlo simulation, we find that $E_X\left\lfloor Var(\varepsilon)\right\rfloor\approx 0.0260$.

Under DGM=2, a close form expression is much more complex to describe. We therefore utilize a Monte-Carlo simulation for estimating the irreducible error, where we simulate data and use as “predictions” for a given X the true expected values, given by p + (1-p)*E(U|X), where p and E(U|X) are known, given the assumptions described in Section 2.2. The simulation was repeated over 1000 iterations and the results were averaged, given a value of 0.0314.