The Impact of Hyperparameter on Optimal Bandwidth

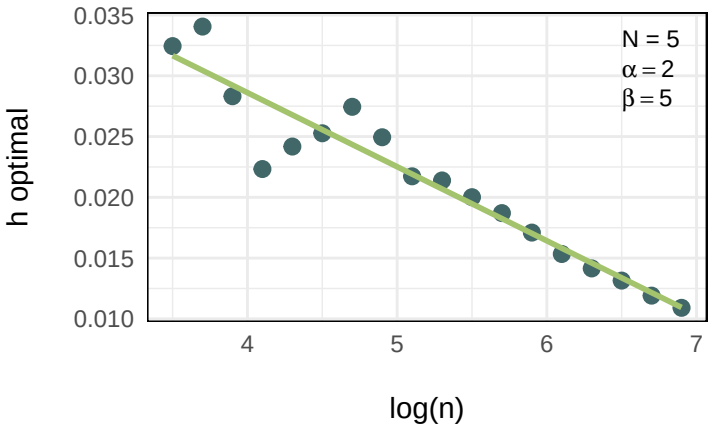
Alessandro Ferrera

The choice of bandwidth h is a crucial factor in the performance of non-parametric estimators, particularly in local linear regression. Bandwidth selection directly influences the estimator's ability to find the link function between the covariate X and the response variable Y. In our example, a well-chosen bandwidth can lead to precise estimates of the conditional expectation $m(x) = E[Y|X = x]$, while a poorly selected bandwidth may result in misleading inferences.

In this context, the asymptotic mean integrated squared error (AMISE) is a particularly valuable criterion for bandwidth selection. The AMISE considers both the bias and variance associated with the estimator, considering a fair trade-off between the two. A small bandwidth reduces bias by closely fitting the data but increases variance due to sensitivity to noise. Conversely, a larger bandwidth smooths out the noise, leading to lower variance but potentially higher bias as the estimator may miss finer details of the relationship. Thus, selecting the optimal bandwidth h_{AMISE} helps achieve an effective compromise.

Sample size

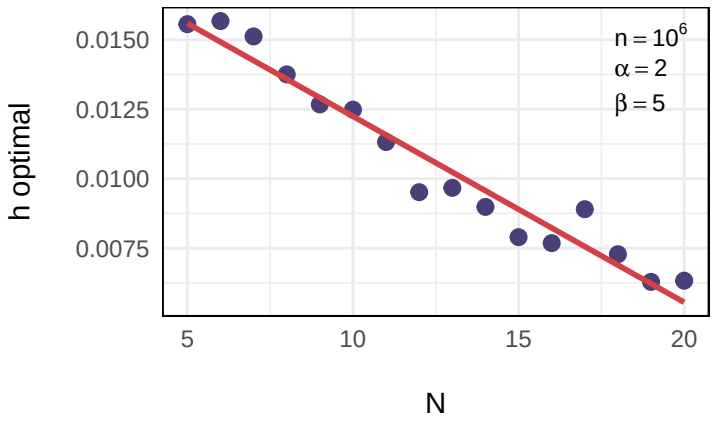
Turning our attention to the sample size n, we observed a decreasing trend of h_{AMISE} in n.



This decrease may be attributed to the large amount of data, that allows the model to reduce the bandwidth in order to minimize the bias, since the noise is already low. This trend underscores the importance of the amount of available data in the estimation process.

Block size

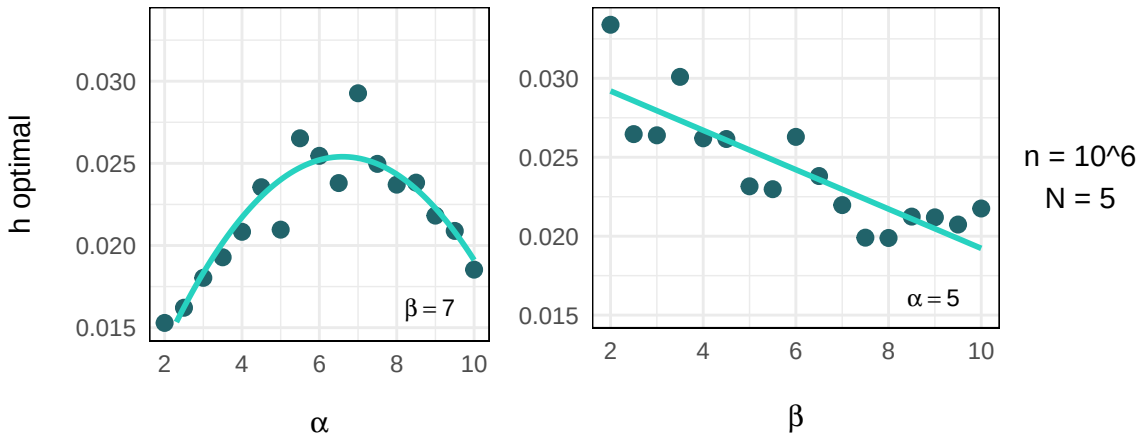
Our analysis revealed a clear trend: h_{AMISE} exhibits a decreasing pattern as the block size N increases. This observation can be rationalized by considering the statistical principles underlying the estimation process. With a larger number of blocks, we gain the advantage of obtaining more reliable estimates of the unknown quantities σ^2 and θ_{22} .



As N increases, the variance in our estimates is effectively reduced, thereby enhancing the precision of the estimated optimal bandwidth. However, it is important to note that beyond a certain threshold, further increases in N may cause problem in the model, if we keep n fixed. In such cases, having too few observations can undermine the reliability of linear regression. Therefore, while larger block can enhance estimation reliability, a careful balance between n and N must be maintained.

Alpha and Beta

When examining the parameters α and β of the Beta distribution of the covariate X, we observed distinct patterns in the behavior of h_{AMISE}. Specifically, for α we have a peek around 7, while for β we notice a consistently decreasing trend.



These behaviors can be understood in the context of the shape of the beta distribution.
We identified a trend indicating that h_{AMISE} reaches a maximum when the distribution is more symmetric. In this scenario, the aim is to minimize the variance, which makes a larger bandwidth.
Conversely, in cases where the distribution is asymmetric, characterized by values concentrated near 0 or 1, so values around 0 or 1, the variance is naturally lower. Here, the focus shifts to minimizing bias, indicating that a smaller bandwidth becomes more suitable.

Conclusion

In conclusion, our analysis revealed important connections between h_{AMISE} and the various parameters we examined. It's clear that these parameters don't just affect it in isolation; they are interconnected. For instance, the sample size n and block size N should vary together for the best results. Similarly, the parameters α and β of the Beta distribution also need to be considered in relation to one another. This understanding highlights the importance of looking at these parameters collectively to improve the accuracy of our estimates.