1 Theoretical exercise: Local linear regression as a linear smoother

Setup. We observe i.i.d. data (X_i, Y_i) from

$$Y_i \;=\; m(X_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i] = 0, \quad i = 1, \dots, n,$$

with scalar covariate $X_i \in \mathbb{R}$. For a target point $x \in \mathbb{R}$, the local linear estimator $(\hat\beta_0(x), \hat\beta_1(x))$ is defined as the weighted least-squares solution

$$(\hat\beta_0(x), \hat\beta_1(x)) \;=\; \arg\min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^{n} \big(Y_i - \beta_0 - \beta_1(X_i - x)\big)^2 K\left(\frac{X_i - x}{h}\right), \tag{1}$$

where K is a kernel and $h > 0$ is a bandwidth. The fitted value is $\hat m(x) := \hat\beta_0(x)$.

1.1 Linearity in Y: $\hat m(x)$ is a weighted average of the Y_i

Write the problem in matrix form. Define

$$r_i := X_i - x, \qquad w_i := K\left(\frac{X_i - x}{h}\right) \;(\geq 0),$$

and the $n \times 2$ design matrix and diagonal weight matrix

$$X \;=\; \begin{bmatrix} 1 & r_1 \\ \vdots & \vdots \\ 1 & r_n \end{bmatrix}, \qquad W \;=\; \mathrm{diag}(w_1, \dots, w_n).$$

Let $Y = (Y_1, \dots, Y_n)^\top$ and $\beta = (\beta_0, \beta_1)^\top$. Then (1) is the weighted least squares (WLS)

$$\hat\beta(x) \;=\; \arg\min_{\beta \in \mathbb{R}^2} (Y - X\beta)^\top W (Y - X\beta).$$

Under the mild condition that $X^\top W X$ is invertible (e.g., at least two distinct X_i with nonzero weight), the WLS solution is

$$\hat\beta(x) \;=\; (X^\top W X)^{-1} X^\top W Y.$$

Since $\hat m(x) = \hat\beta_0(x) = e_1^\top \hat\beta(x)$ with $e_1 = (1, 0)^\top$, we have

$$\hat m(x) \;=\; e_1^\top (X^\top W X)^{-1} X^\top W\, Y \;=\; \sum_{i=1}^{n} \underbrace{\big[e_1^\top (X^\top W X)^{-1} X^\top W\, e_i\big]}_{=:\, w_{ni}(x)} Y_i, \tag{2}$$

where e_i is the ith standard basis vector in $\mathbb{R}^n$. Thus $\hat m(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i$, and the weights $w_{ni}(x)$ depend only on x, $\{X_i\}$, K, and h (through W and X), not on the Y_i's. This proves 1.1.

1.2 Explicit weights in terms of $S_{n,k}(x)$

Introduce the shorthand

$$S_0 := \sum_{i=1}^{n} w_i, \qquad S_1 := \sum_{i=1}^{n} w_i r_i, \qquad S_2 := \sum_{i=1}^{n} w_i r_i^2.$$

Then

$$X^\top W X \;=\; \begin{bmatrix} S_0 & S_1 \\ S_1 & S_2 \end{bmatrix}, \qquad (X^\top W X)^{-1} \;=\; \frac{1}{\Delta}\begin{bmatrix} S_2 & -S_1 \\ -S_1 & S_0 \end{bmatrix}, \quad \Delta := S_0 S_2 - S_1^2.$$

Also, the ith column of $X^\top W$ is $(w_i,\, w_i r_i)^\top$. Therefore,

$$w_{ni}(x) \;=\; e_1^\top (X^\top W X)^{-1}(X^\top W e_i) \;=\; \frac{1}{\Delta}\,[S_2,\, -S_1]\begin{bmatrix} w_i \\ w_i r_i \end{bmatrix} \;=\; \frac{w_i}{\Delta}\,\big(S_2 - S_1 r_i\big).$$

Next, rewrite everything using the normalized sums

$$S_{n,k}(x) \;:=\; \frac{1}{nh}\sum_{i=1}^{n} r_i^k\, K\Big(\frac{r_i}{h}\Big) \;=\; \frac{1}{nh}\sum_{i=1}^{n} r_i^k\, w_i, \qquad k = 0, 1, 2.$$

Because $S_k = \sum_{i=1}^n w_i r_i^k = nh\, S_{n,k}(x)$ for $k = 0, 1, 2$, we have

$$\Delta \;=\; S_0 S_2 - S_1^2 \;=\; (nh)^2 \big(S_{n,0} S_{n,2} - S_{n,1}^2\big).$$

Finally, $w_i = K\big(r_i/h\big)$ and

$$w_{ni}(x) \;=\; \frac{K\Big(\frac{X_i - x}{h}\Big)\,\big(S_2 - S_1 r_i\big)}{S_0 S_2 - S_1^2} \;=\; \frac{1}{nh}\,\frac{K\Big(\frac{X_i - x}{h}\Big)\,\Big(S_{n,2}(x) - (X_i - x)\, S_{n,1}(x)\Big)}{S_{n,0}(x)\, S_{n,2}(x) - \big(S_{n,1}(x)\big)^2}.$$

Thus,

$$\boxed{\;w_{ni}(x) \;=\; \frac{1}{nh}\,\frac{K\left(\dfrac{X_i - x}{h}\right)\,\Big(S_{n,2}(x) - (X_i - x)\, S_{n,1}(x)\Big)}{S_{n,0}(x)\, S_{n,2}(x) - \big(S_{n,1}(x)\big)^2}\;} \tag{3}$$

which yields the explicit linear-smoother representation

$$\hat{m}(x) \;=\; \sum_{i=1}^{n} w_{ni}(x)\, Y_i.$$

1.3 The weights sum to one: $\sum_{i=1}^n w_{ni}(x) = 1$

Starting from (3),

$$\sum_{i=1}^{n} w_{ni}(x) \;=\; \frac{1}{nh}\,\frac{S_{n,2}(x)\, \sum_{i=1}^n K\Big(\frac{X_i - x}{h}\Big) \;-\; S_{n,1}(x)\, \sum_{i=1}^n (X_i - x)\, K\Big(\frac{X_i - x}{h}\Big)}{S_{n,0}(x)\, S_{n,2}(x) - \big(S_{n,1}(x)\big)^2}.$$

Recognize the two sums in the numerator:

$$\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \;=\; S_0 \;=\; nh\, S_{n,0}(x), \qquad \sum_{i=1}^{n} (X_i - x)\, K\left(\frac{X_i - x}{h}\right) \;=\; S_1 \;=\; nh\, S_{n,1}(x).$$

Hence

$$\sum_{i=1}^{n} w_{ni}(x) \;=\; \frac{1}{nh}\,\frac{S_{n,2}(x)\, nh\, S_{n,0}(x) - S_{n,1}(x)\, nh\, S_{n,1}(x)}{S_{n,0}(x)\, S_{n,2}(x) - \big(S_{n,1}(x)\big)^2} \;=\; \frac{S_{n,0}(x)\, S_{n,2}(x) - \big(S_{n,1}(x)\big)^2}{S_{n,0}(x)\, S_{n,2}(x) - \big(S_{n,1}(x)\big)^2} \;=\; 1.$$

This proves $\sum_{i=1}^n w_{ni}(x) = 1$.

2 Practical exercise: Global bandwidth selection

2.1 Setup and goal

We observe i.i.d. (X_i, Y_i) from

$$Y_i = m(X_i) + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n,$$

with $X_i \in [0, 1]$ drawn from Beta(α, β) and

$$m(x) \;=\; \sin\left\{ \left(\tfrac{x}{3} + 0.1\right)^{-1} \right\}.$$

We consider the local linear estimator $\hat{m}$ and focus on selecting a global bandwidth h that minimizes AMISE under homoscedasticity and a quartic (biweight) kernel. Under these assumptions the optimal bandwidth admits the explicit form

$$h_{\text{AMISE}} \;=\; n^{-1/5} \left(\frac{35\,\sigma^2\,|\text{supp}(X)|}{\theta_{22}} \right)^{1/5}, \qquad \theta_{22} = \int \{m''(x)\}^2 f_X(x)\,dx, \tag{4}$$

We must estimate the unknowns σ^2 and θ_{22}.
Following the blocked quartic pilot strategy, we partition the sample into N blocks and fit a separate quartic:

$$Y_i = \beta_{0j} + \beta_{1j} X_i + \beta_{2j} X_i^2 + \beta_{3j} X_i^3 + \beta_{4j} X_i^4 + \varepsilon_i, \quad X_i \in \mathcal{X}_j,\ j = 1, \dots, N,$$

then define

$$\widehat{\theta}_{22}(N) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} \left\{ \hat{m}_j''(X_i) \right\}^2 \mathbf{1}\{X_i \in \mathcal{X}_j\}, \qquad \widehat{\sigma}^2(N) = \frac{1}{n - 5N} \sum_{i=1}^{n} \sum_{j=1}^{N} \{Y_i - \hat{m}_j(X_i)\}^2 \mathbf{1}\{X_i \in \mathcal{X}_j\}.$$

To choose N in a data-driven way we minimize Mallow's C_p over $N = 1, \dots, N_{\max}$,

$$C_p(N) \;=\; \frac{\text{RSS}(N)}{\text{RSS}(N_{\max})/(n - 5N_{\max})} - (n - 10N), \quad N_{\max} = \max\{\min(\lfloor n/20 \rfloor, 5), 1\}.$$

With $\widehat{\sigma}^2 = \widehat{\sigma}^2(N_{\text{opt}})$ and $\widehat{\theta}_{22} = \widehat{\theta}_{22}(N_{\text{opt}})$, we plug into (4) to obtain $\hat{h}_{\text{AMISE}}$.

2.2 Simulation design

We use:

- **Sample size** $n \in \{100, 1000, 10000, 100000\}$;
- **Design shape** $(\alpha, \beta) \in \{(2, 2), (2, 5), (5, 2), (0.5, 0.5)\}$ for $X \sim$ Beta(α, β);
- **Noise level** $\sigma^2 = 1$;
- **Replicates** $R = 100$ per setting.

Within each replicate we:

1. Simulate (X_i, Y_i) and compute $(\widehat{\theta}_{22}(N), \widehat{\sigma}^2(N), C_p(N))$ for all $N = 1, \dots, N_{\max}$;
2. Choose $N_{\text{opt}} = \arg\min_N C_p(N)$;
3. Compute $\hat{h}_{\text{AMISE}} = n^{-1/5} \left(35\,\widehat{\sigma}^2\,|\text{supp}(X)|/\widehat{\theta}_{22}\right)^{1/5}$.

2.3 Key figures and tables

Below we include representative visualizations created by the script.

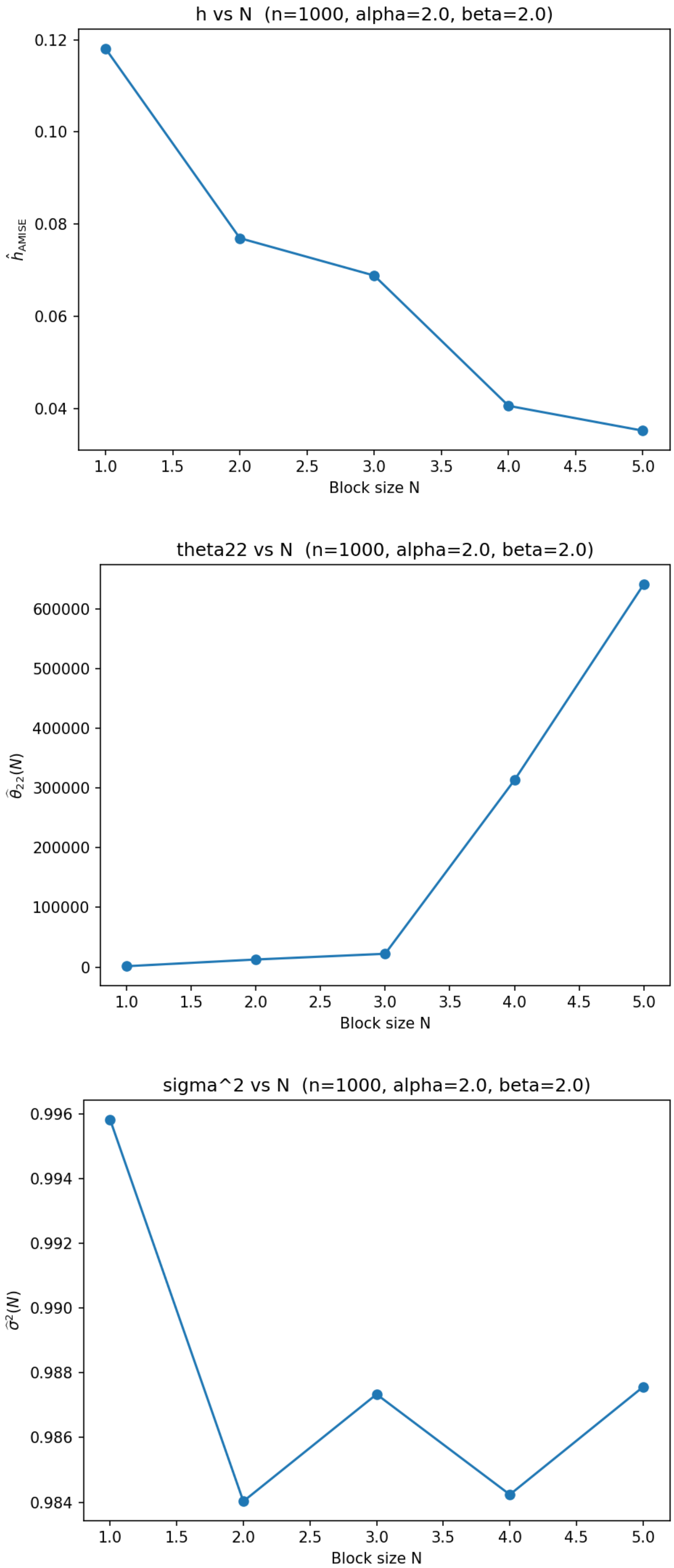

Figure 1: For a single dataset ($n = 200$, $\alpha = \beta = 2$): $\hat{h}_{\mathrm{AMISE}}$, $\widehat{\theta}_{22}$, and $\widehat{\sigma}^2$ versus the number of blocks N.

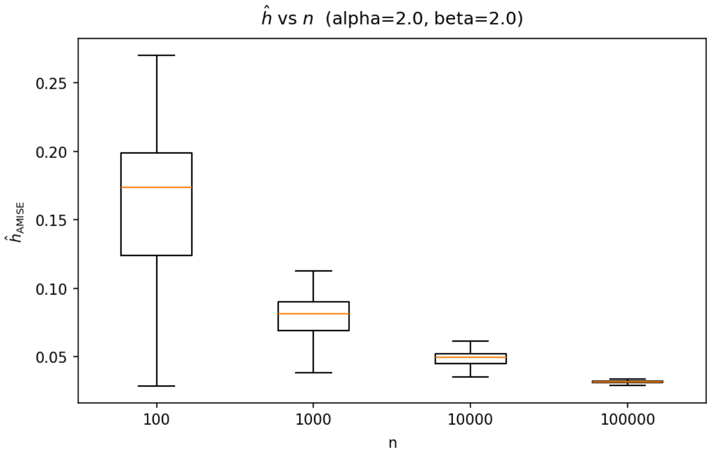

Figure 2: Distribution of $\hat{h}_{\text{AMISE}}$ vs n for $X \sim \text{Beta}(2, 2)$.

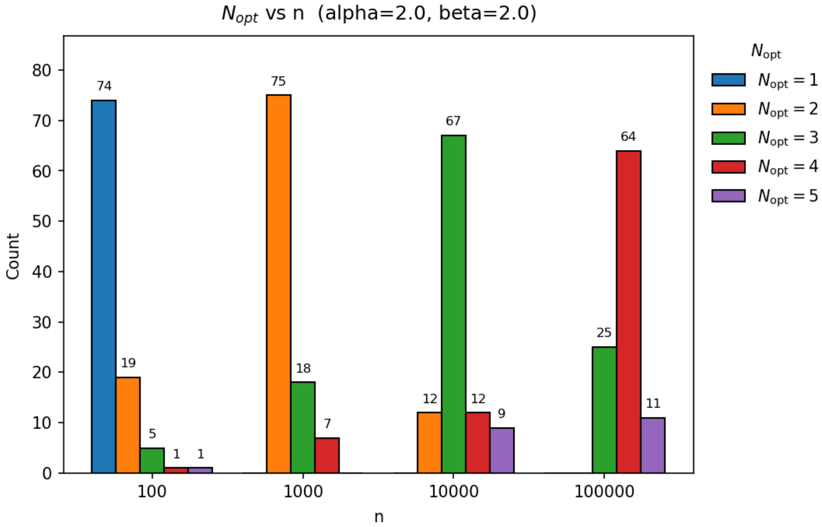

Figure 3: Distribution of N_{opt} (by C_p) vs n for $X \sim \text{Beta}(2, 2)$.

Table 1: Summary statistics of $\hat{h}_{\text{AMISE}}$ and N_{opt} across simulation scenarios.

α	β	n	Median($\hat{h}_{\text{AMISE}}$)	IQR($\hat{h}_{\text{AMISE}}$)	Mode(N_{opt})
0.5	0.5	100	0.083122	0.031411	2
		1000	0.053817	0.004222	2
		10000	0.035148	0.002759	3
		100000	0.025698	0.001491	4
2.0	2.0	100	0.173648	0.075198	1
		1000	0.081186	0.020890	2
		10000	0.049761	0.006789	3
		100000	0.031673	0.001402	4
2.0	5.0	100	0.137356	0.055404	1
		1000	0.059287	0.010956	2
		10000	0.038903	0.001932	2
		100000	0.024997	0.000778	3
5.0	2.0	100	0.181831	0.082602	1
		1000	0.163484	0.037461	1
		10000	0.112094	0.011551	1
		100000	0.067144	0.008476	2

2.4 Findings and interpretation

2.4.1 Overall scaling of the optimal bandwidth

Across all designs, the median $\hat{h}_{\text{AMISE}}$ decreases as n grows (Table 1; Figures 2, 7–9). This matches the AMISE formula

$$\hat{h}_{\text{AMISE}} \;=\; n^{-1/5}\left(\frac{35\,\hat{\sigma}^2\,|\text{supp}(X)|}{\hat{\theta}_{22}}\right)^{1/5},$$

whose leading $n^{-1/5}$ factor drives a systematic shrinkage with sample size. Empirically, the interquartile range (IQR) also tightens with n, reflecting reduced Monte Carlo variability as more information is available.

2.4.2 How $\hat{h}_{\text{AMISE}}$ behaves when N grows (single–dataset diagnostics)

From Figures 1 and 4–6, we observe a common pattern:

(a) $\widehat{\theta}_{22}(N)$ typically increases with N. More blocks yield quartic pilots that adapt more locally, revealing curvature and pushing the denominator of $\hat{h}_{\text{AMISE}}$ up.

(b) $\widehat{\sigma}^2(N)$ is comparatively stable around 1 for moderate N (the d.f. adjustment $n - 5N$ compensates the added flexibility).

(c) Because $\hat{h}_{\text{AMISE}} \propto \left(\widehat{\sigma}^2/\widehat{\theta}_{22}\right)^{1/5}$, $\hat{h}$ often decreases with N (driven by the growth of $\widehat{\theta}_{22}$) and flattens.

2.4.3 Should N depend on n?

Yes, and the data confirm it. As n increases, the C_p-selected N_{opt} slowly increases (Figures 3 and 10–12; Table 1). Intuitively, larger samples support more blocks without destabilizing the blockwise quartic fits. Concretely:

- **Beta(2,2):** N_{opt} moves from 1 (n=100) to 2 (n=1000), then to 3 and 4 (n=$10^4, 10^5$).
- **Beta(2,5):** $1 \to 2 \to 2 \to 3$ over the same n grid.
- **Beta(0.5,0.5):** $2 \to 2 \to 3 \to 4$.
- **Beta(5,2):** the most conservative: $1 \to 1 \to 1 \to 2$.

Growth remains modest because C_p imposes a linear complexity penalty in N and because the cap $N_{\max} = \min(\lfloor n/20 \rfloor, 5)$ prevents over-fragmentation.

2.4.4 Impact of the design density f_X (via Beta(α, β))

The curvature functional $\theta_{22} = \int \{m''(x)\}^2 f_X(x)\,dx$ weights curvature by f_X. This explains the systematic differences across designs:

- **Skewed right (Beta(5,2)):** Many points cluster near $x \approx 1$, with a sparser left tail. Extra blocks create very small, noisy blocks in sparse regions, so the RSS reduction rarely offsets the C_p penalty. Hence N_{opt} stays at 1 until very large n, and the median $\hat{h}$ is comparatively large (Table 1).
- **Skewed left (Beta(2,5)) and symmetric (Beta(2,2)):** The mass is more balanced where m'' is non-negligible, so $\widehat{\theta}_{22}$ increases more with N, making a second (and later a third) block worthwhile as n grows. Correspondingly, N_{opt} rises to 2 and then to 3–4 at large n, while $\hat{h}$ decreases more rapidly.

- **U-shaped (Beta(0.5,0.5)):** Many points lie near the boundaries; if those regions also feature substantial curvature, $\widehat{\theta}_{22}$ is relatively large, yielding a smaller $\hat{h}$ already at moderate n. The design naturally supports at least two blocks (left/right), so N_{opt} starts at 2 and increases with n.

2.4.5 Practical takeaways

(i) The dominant effect is the theoretical $n^{-1/5}$ shrinkage of $\hat{h}_{\text{AMISE}}$; the data follow this trend across all designs.

(ii) The pilot block count N should increase slowly with n; C_p finds small values (typically 1–4 here), balancing curvature resolution against variance.

(iii) When f_X is highly uneven, a global h must compromise between dense and sparse regions. Designs concentrating mass where $|m''|$ is large (e.g., $\text{Beta}(0.5, 0.5)$) tend to produce smaller $\hat{h}$; designs with long sparse tails (e.g., $\text{Beta}(5, 2)$) tend to keep N_{opt} small and $\hat{h}$ larger.

2.5 Additional figures

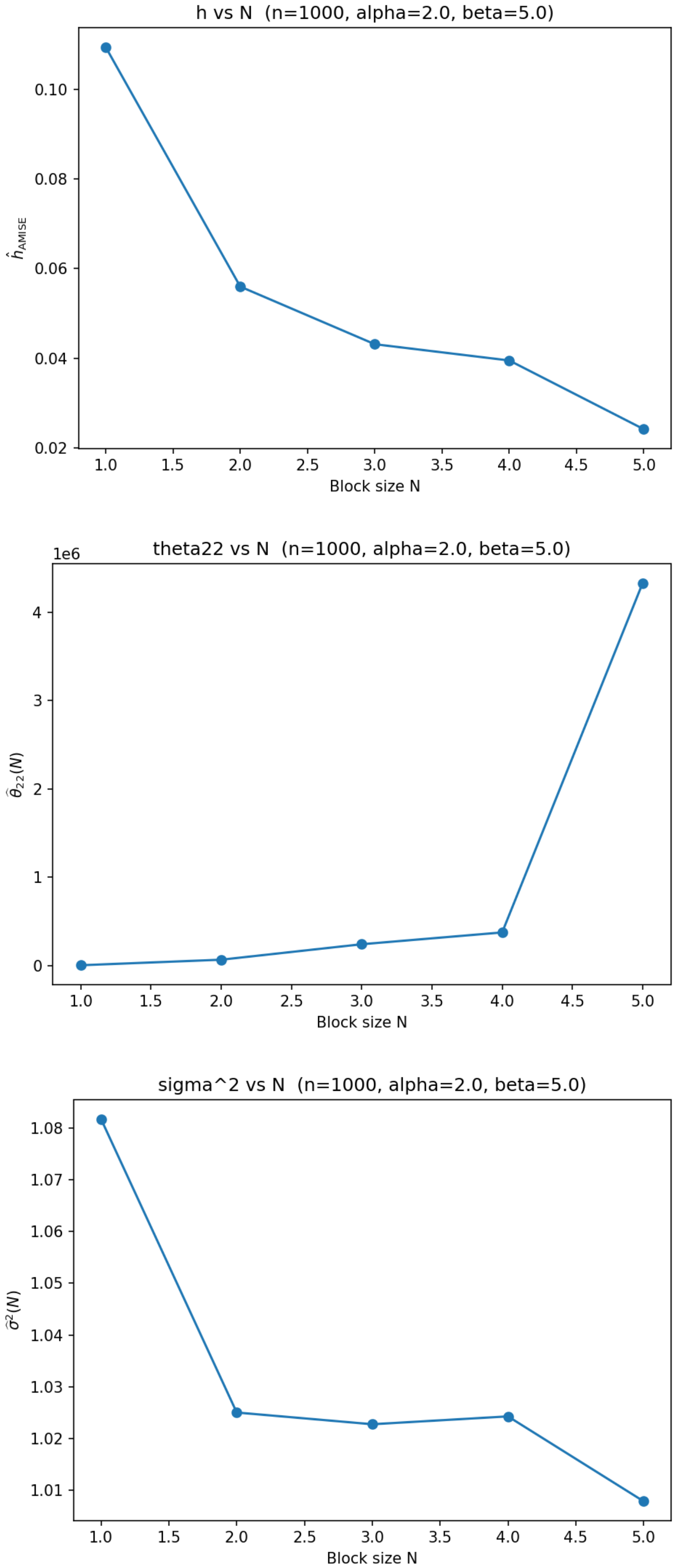

Figure 4: For a single dataset ($n = 1000$, $\alpha = 2$, $\beta = 5$): $\hat{h}_{\text{AMISE}}$, $\widehat{\theta}_{22}$, and $\widehat{\sigma}^2$ versus the number of blocks N.

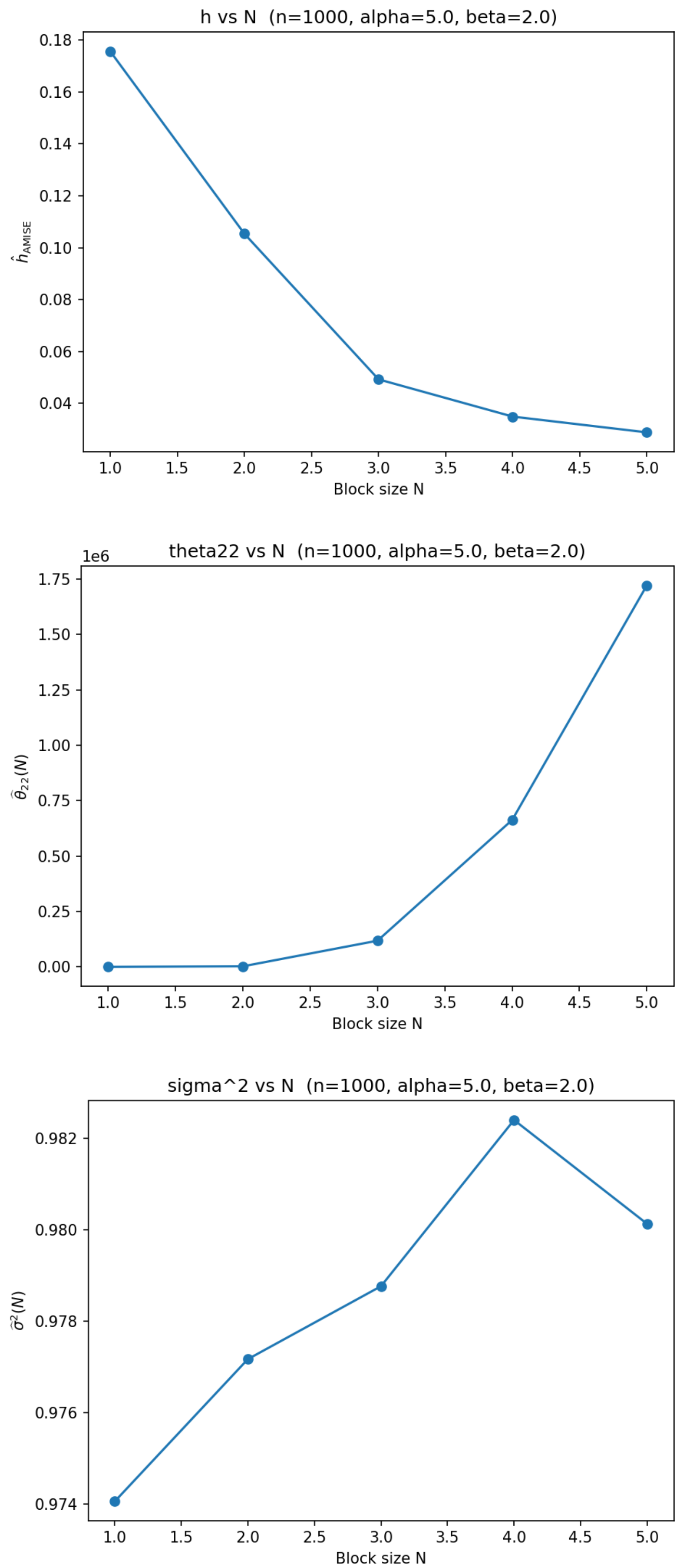

Figure 5: For a single dataset ($n = 1000$, $\alpha = 5$, $\beta = 2$): $\hat{h}_{\mathrm{AMISE}}$, $\widehat{\theta}_{22}$, and $\widehat{\sigma}^2$ versus the number of blocks N.

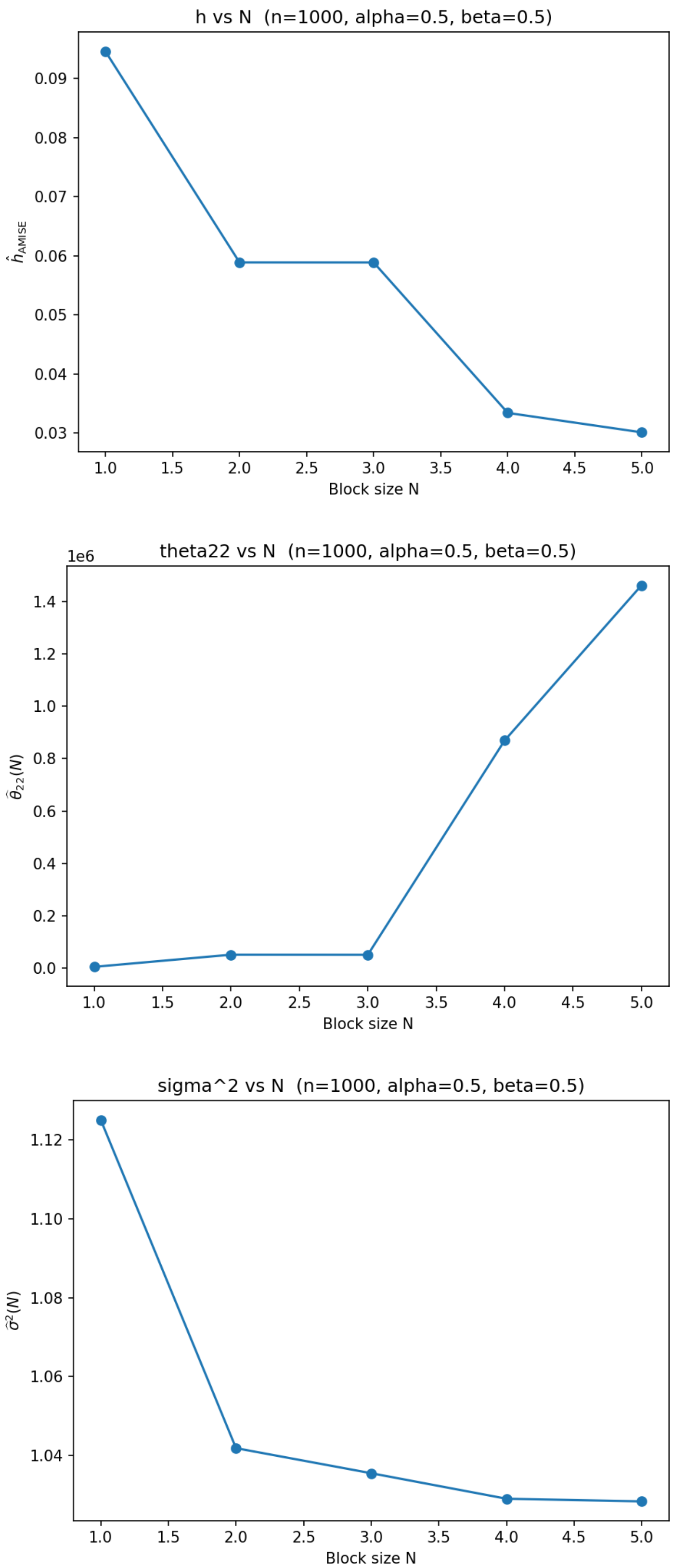

Figure 6: For a single dataset ($n = 1000$, $\alpha = 0.5$, $\beta = 0.5$): $\hat{h}_{\mathrm{AMISE}}$, $\widehat{\theta}_{22}$, and $\widehat{\sigma}^2$ versus the number of blocks N.

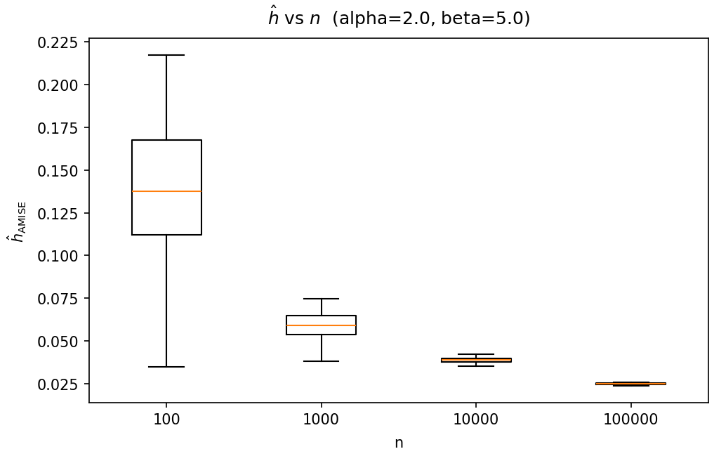

Figure 7: Distribution of $\hat{h}_{\text{AMISE}}$ vs n for $X \sim \text{Beta}(2, 5)$.

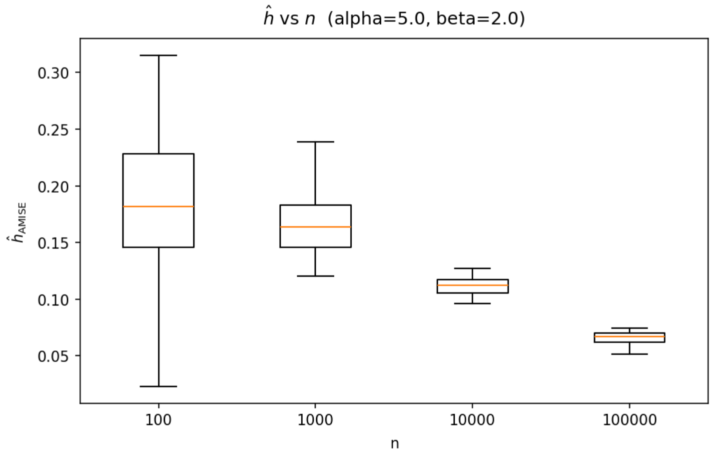

Figure 8: Distribution of $\hat{h}_{\text{AMISE}}$ vs n for $X \sim \text{Beta}(5, 2)$.

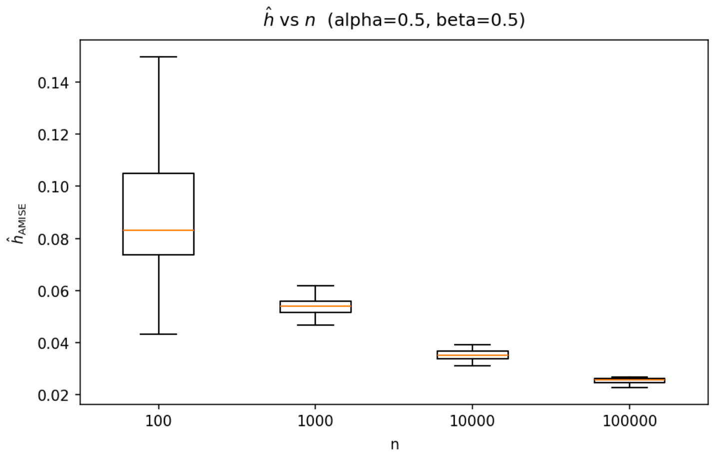

Figure 9: Distribution of $\hat{h}_{\text{AMISE}}$ vs n for $X \sim \text{Beta}(0.5, 0.5)$.

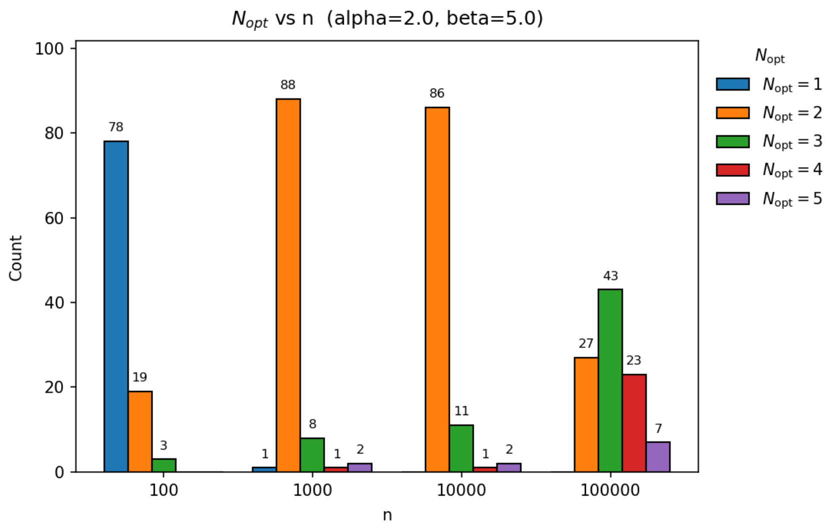

Figure 10: Distribution of N_{opt} (by C_p) vs n for $X \sim \mathrm{Beta}(2, 5)$.

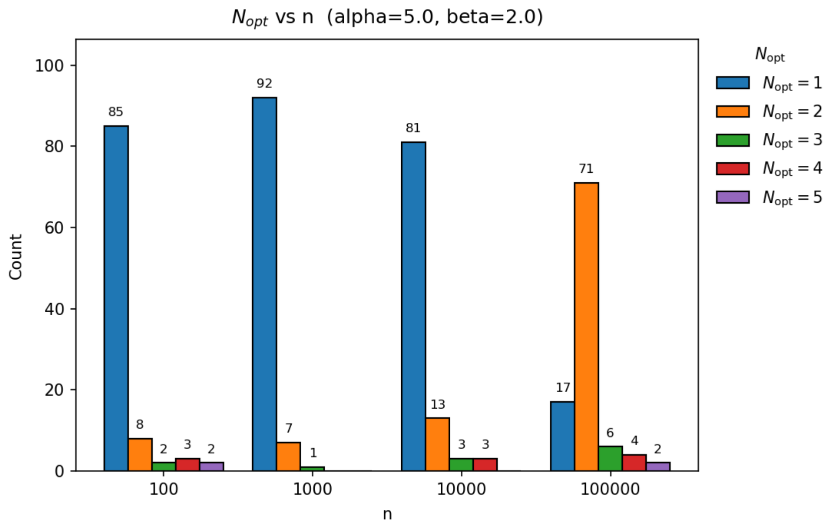

Figure 11: Distribution of N_{opt} (by C_p) vs n for $X \sim \mathrm{Beta}(5, 2)$.

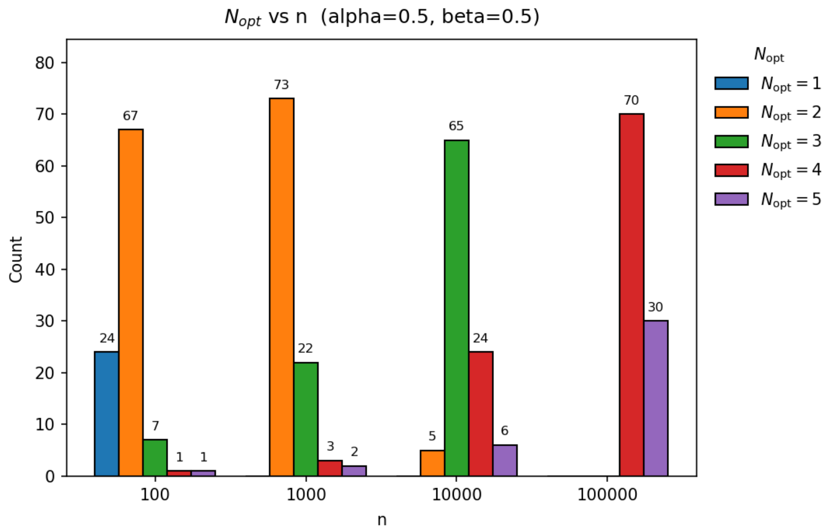

Figure 12: Distribution of N_{opt} (by C_p) vs n for $X \sim \mathrm{Beta}(0.5, 0.5)$.