Information Theory (2021/22)
Homework 2: Estimating Differential Entropies

Tommaso Bergamasco (ID: 2052409)

Contents

1 Introduction

In this report I'm going to describe the procedures I used to estimate the differential entropy and other related quantities. Then I will show the results obtained by testing these procedure comparing them to their "true" theoretical values. I will do this both for the univariate and multivariate case.

2 Estimating the Differential Entropy for a vector x

The procedure is almost entirely based on the estimation of the continous pdf (exactly as we saw in the discrete case in HW1). Since this time the vector of realization x can assume continous (float) values, we need to gather the values of x which are reasonably near to each other.

We do this by using an histogram: every value associated to each bin of the histogram represents the probability for a certain sample of the vector x to be in that specific bin. Namely we reduced the problem to a discrete one once again.

At this point we can obtain the differential entropy $h(x)$ by simply using a modified version of the definition and by keeping in mind that the "approximated integral" we use must be "corrected" with the **width** of the bins we used for the histogram, namely:

$$h(x) = \int_{S_x} p_x(a) i_x(a) \approx \delta \sum_{p_\delta \neq 0} p_\delta \log_{1/2} p_\delta \tag{1}$$

where p_δ is the probability of each bin of the histogram and δ is the costant width used ∀ bin.

3 True Theoretical Values (Univariate case)

By using the definition we can simply obtain the differential entropy in the 3 required cases:

Uniform Distribution over $[0, A]$

$$h_{uniform}(x) = \log_2 A \text{ [bit]}$$

Gaussian Distribution with mean $m_x = 0$ **and variance** $\sigma_x^2 = p$

$$h_{normal}(x) = \frac{1}{2} \log_2(2\pi e \sigma^2) \text{ [bit]}$$

Exponential Distribution with parameter λ

$$h_{exp}(x) = 1 - \ln \lambda \text{ [nat]} = \frac{1 - \ln \lambda}{\log 2} \text{ [bit]}$$

4 Test on the Uniform

4.1 Varying A

As we can see in Figure 1 the estimation leads to a very small estimation error. The error increases with the parameter A exactly as we saw in the discrete case, since we "expand" the alphabet leading to more uncertainty on our estimation.

Note also how the plot is almost the same of the discrete case uniform (this was expected since with the estimation procedure I am basically forcing a discrete alphabet which approximates the "continous" one).

4.2 Varying the Number of Realizations L

In Figure 2 we simply note how the estimation tends better to the true value when L grows.

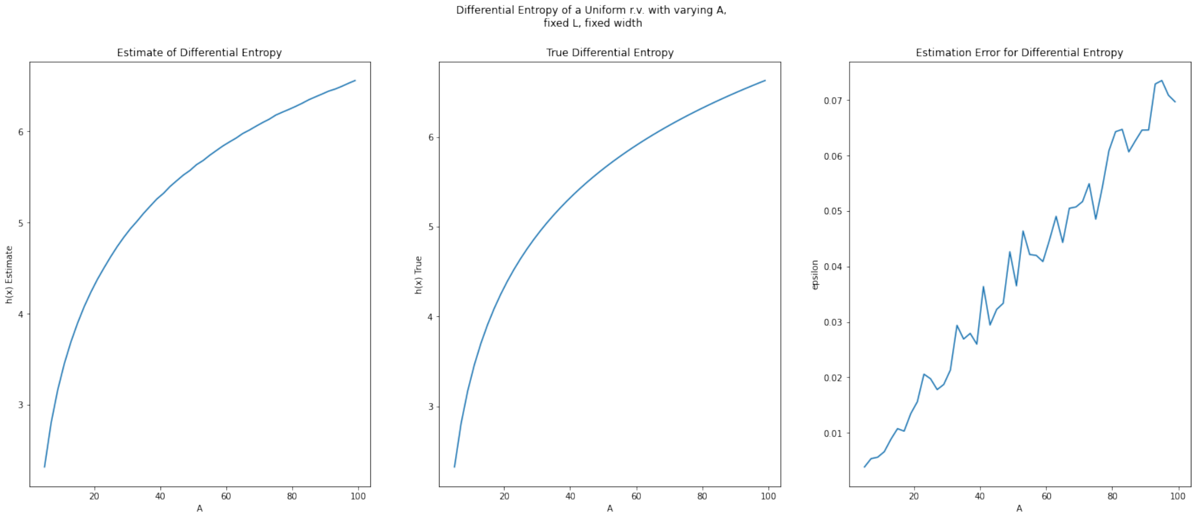


Figure 1: Estimation of the Differential Entropy for a Uniform r.v. with parameter A varying over [2,100], fixed length(x)=L, fixed bins' width.

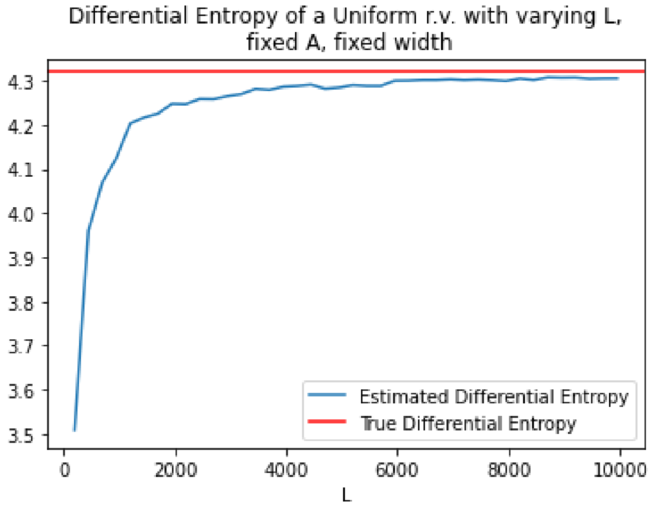


Figure 2: Diferential Entropy of a Uniform with varying L, fixed A and fixed width.

4.3 Varying the width of the bins

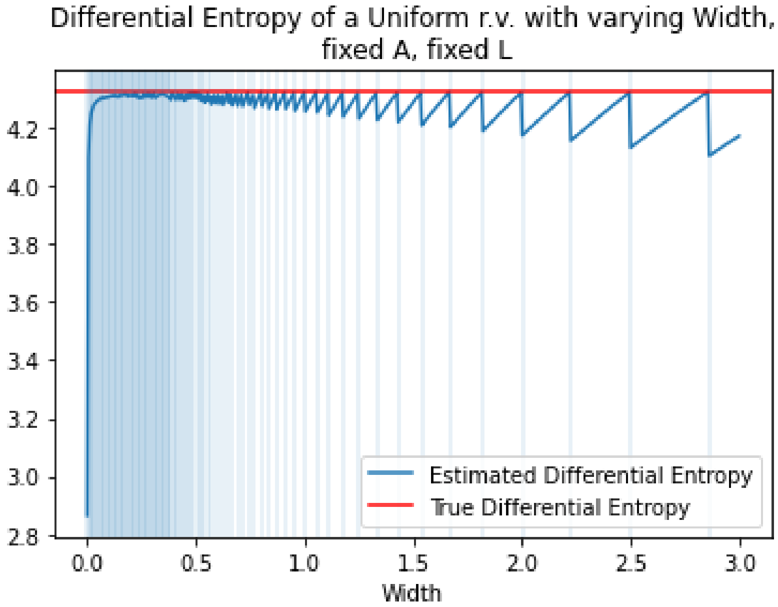


Figure 3: Differential Entropy of a Uniform with varying width, fixed A and fixed L. Every vertical line indicates when the increase of the width involves a decrease in the number of bins used in the histogram (of course the bigger the width, the fewer bins we need to "store" every value of x).

Looking at Figure 3 we can note that:

- When width << we have a huge estimation error. This is due to the fact that the bins are so small they probably contain only 1 sample each, leading to a pdf which could be visually similar to a Dirichlet function (of course not representative of the uniform).
- The best estimation is in the interval $[0.1, 0.5]$.
- When width >> (and hence the number of bins becomes too small) we notice sort of a periodic pattern which tends to distance from the true value more and more as the width grows. To understand this behaviour we note that our estimate is on the red line only when $\frac{A-1}{width} \in \mathbb{N}$, otherwise we have that the bins edges do not perform a balanced partition of the domain of x.

5 Test on the Gaussian

5.1 Varying the Variance p

In Figure 4 note that our estimation is very bad when $p <<$ since in this case we have a lot values gathered in a small interval and (relatively) too many bins to store them (see first point of paragraph 4.3).

Otherwise our estimation seems good but becomes worse when the variance $p \uparrow$, namely when the differential entropy (uncertainty) about the r.v. increases.

5.2 Varying L

In Figure 5 note once again how our estimate becomes better when L increases.

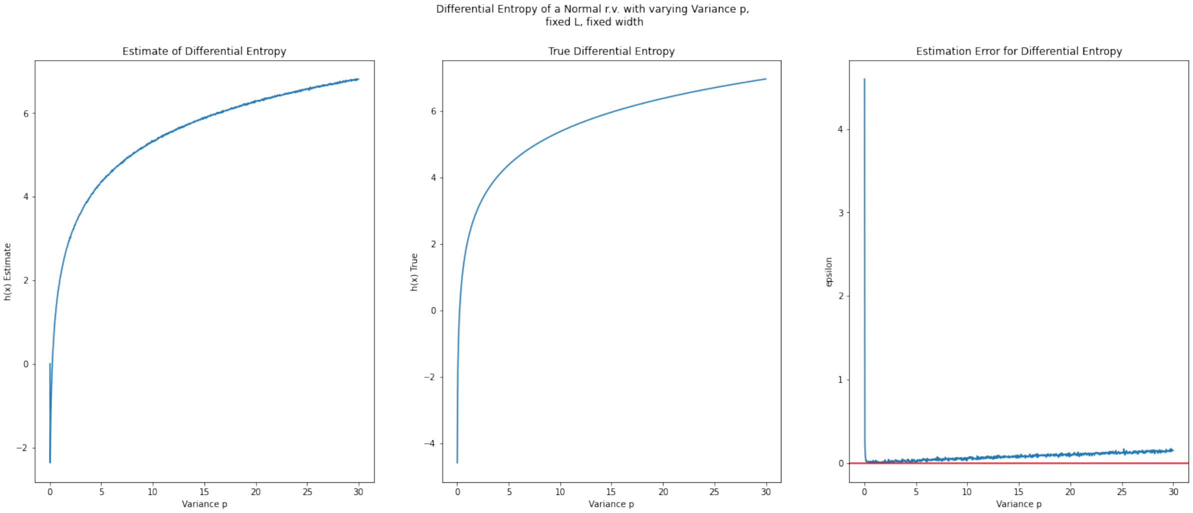


Figure 4: Estimation of the Differential Entropy for a Gaussian with varying variance p, fixed L, fixed width

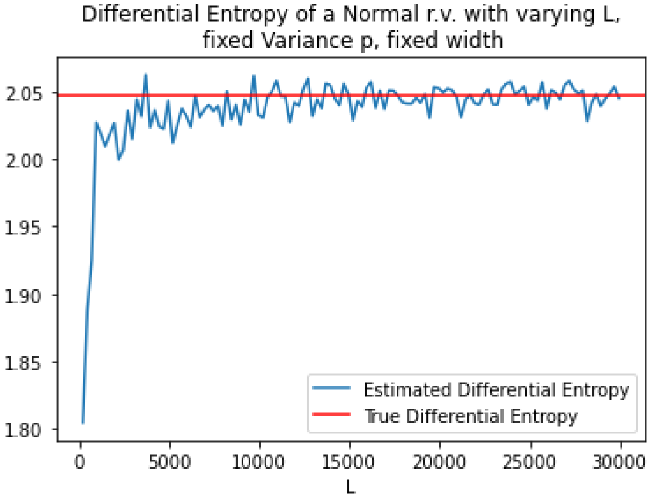


Figure 5: Differential Entropy of a Gaussian with varying L, fixed p, fixed width.

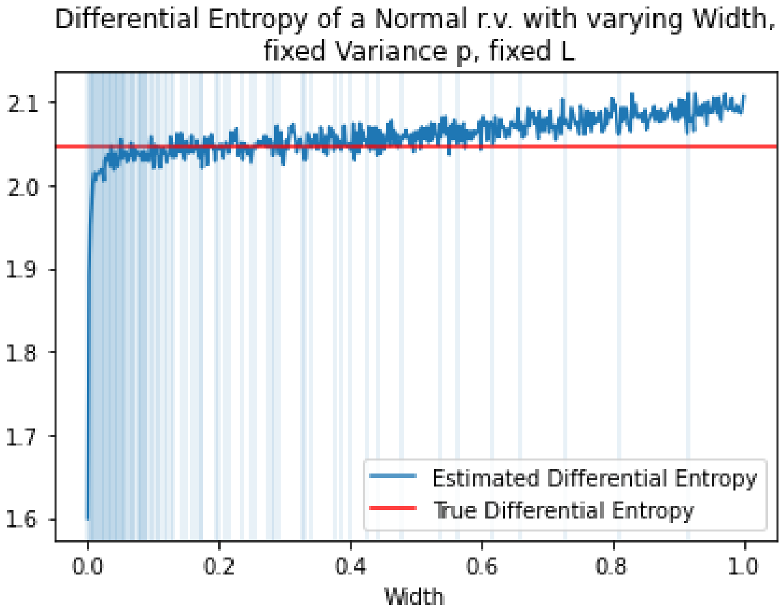


Figure 6: Differential Entropy of a Gaussian with varying width, fixed p, fixed L. The vertical lines have the same meaning described in Figure 3

5.3 Varying the Width

In Figure 6 note once again as in the uniform case (and for the same reason) that our estimate is bad for width <<. Note also that even in the best case (width $\in [0.1, 0.4]$) the estimation is a lot more noisy than the uniform case (this is due to the fact that in the continous case the Gaussian achieves maximum entropy and hence maximum uncertainty).

Finally note that for width >> (when the number of bins decreases) the estimate starts to diverge from the true value. Note that in this case we don't have a periodic pattern since for construction the extremes of the domain of x can vary a lot due to the low-probability-tails of the Gaussian distribution.

6 Test on the Exponential

6.1 Varying the parameter λ

In Figure 7 note how the entropy decreases when λ increases. This is due to the fact higher λ are associated with a faster-decreasing-exponential, leading to a distribution of x more concentrated near 0 (smaller alphabet).

Note how our estimation is good, but deeply affected by noise, this is due to the fact that unlikely events (big x_i) can significantly vary the bin edges (similarly at what we saw in the Gaussian case) modifying how the histogram works.

6.2 Varying L

In Figure 8 we can once again note how our estimates becomes better when L increases.

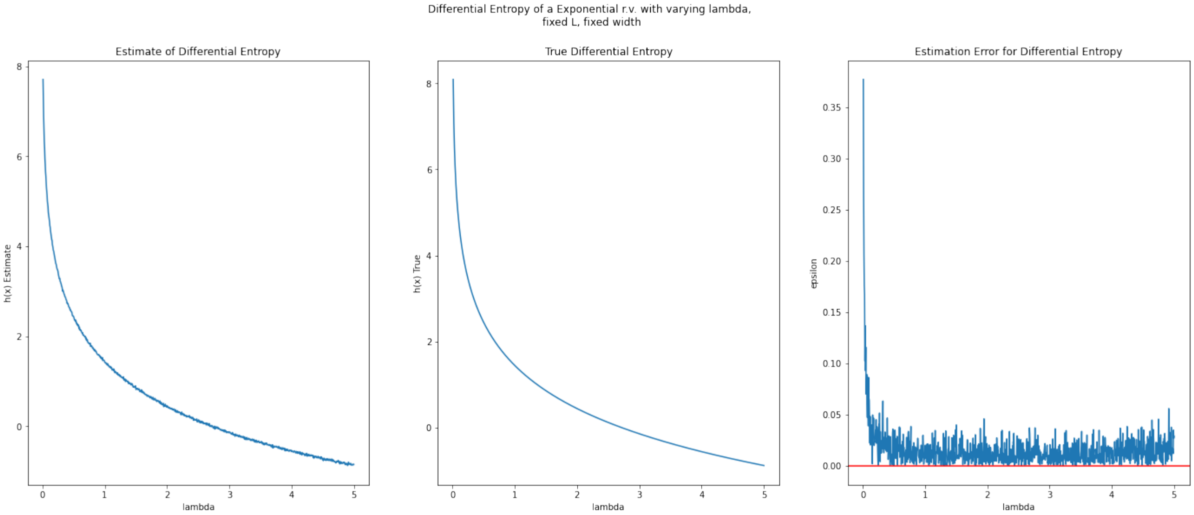


Figure 7: Differential Entropy of an Exponential with varying λ, fixed L, fixed width.

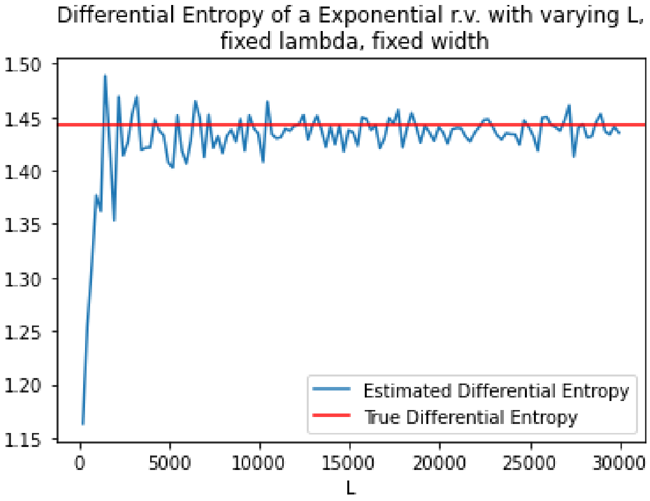


Figure 8: Differential Entropy an Exponential with varying L, fixed ,λ, fixed width.

6.3 Varying the width

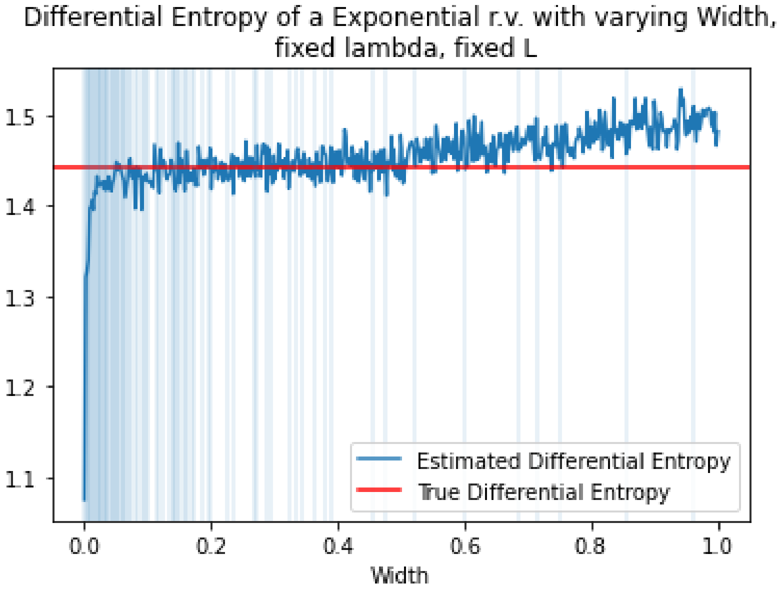


Figure 9: Differential Entropy of an Exponential with varying width, fixed L, fixed λ. The vertical lines have the same meaning described in the previous cases.

In Figure 9 note once again how our estimate is very bad for width << (same reason as previous cases), it becomes better for width $\in [0.1, 0.5]$ and then starts to diverge when the number of bins decreases (the same considerations of the 2 previous cases hold).

7 "Joint" Estimations for x and y

Joint Differential Entropy

In order to estimate it (as we already saw in the discrete case and in the univariate continous case) the problem is reduced to compute the joint pdf. The procedure is the 2-dimensional extension of what we have seen for a continous vector x.

In python we can use `numpy.histogram2d(x,y)` in order to obtain the multidimensional histogram and then we can use a discrete approximation of the (continous) definition as we saw in the univariate case, namely:

$$h(x,y) = \int_{S_{xy}} p_{xy}(a,b) i_{xy}(a,b) \approx \sum_{p_{xy}(a,b)\neq 0} p_{xy}(a,b) i_{xy}(a,b) \cdot \text{width}^2.$$

Where notice that this time the discrete summation is corrected with the term width^2 since we are dealing with a 2-dimensional case.

Conditional Differential Entropy $h(x|y)$

Since we have already computed $h(x,y)$ and and we are able to compute also univariate differential entropies, we can simply use the following formula:

$$h(x|y) = h(x,y) - h(y).$$

Relative Entropy $D(p_x||p_y)$

In this case it is sufficient to extract the singular pdfs and then to use the well known formula (same of discrete case) keeping in mind that it must be $S_x \subset S_y$:

$$D(p_x||p_y) = \mathbb{E}_x \left[\log_2 \frac{p_x(x)}{p_y(x)}\right].$$

Alternatively I could have used the following formula:

$$D(p_x||p_y) = h(p_x, p_y) - h(x).$$

But this way we avoid propagation of errors due to the estimating procedure of the joint entropy $h(p_x, p_y)$.

Mutual Information $I(x; y)$

Once again by exploiting the joint differential entropy and the single differential entropies of x and y, we obtain:

$$I(x; y) = h(x) + h(y) - h(x, y).$$

8 "True" Theoretical Values (Multivariate case)

In both the proposed experiments (Uniform and Gaussian distributions) it would be necessary to first compute the joint pdfs of the vectors x and y and after that to compute the related quantities by applying the definitions.

To avoid the computations to derive the general expressions of the joint pdfs, I will consider as reference values the ones obtained with the simple procedure described above performed with "optimal" parameters, namely with $L >>$ and a width=0.1 (which led to best performances in the univariate case).

9 Test on x Uniform and $y = x + z$ (z Uniform)

9.1 Varying A

In Figure 10 are shown the behaviours of the joint quantities, note the following:

- The **joint entropy** is proportional to A (same as discrete and/or univariate case). In facts the smaller the alphabet the smaller the uncertainty about x and therefore about also y.
- The **estimation error** about the **joint entropy** is proportional to the alphabet cardinality (this happens for the same reason of the point above: more uncertainty → bigger error).
- The **conditional entropy** $h(x|y) = h(x, y) - h(y)$ decreases when the alphabet cardinality increases. This is due to the fact that since y is a shifted version of x, when A increases then the window of values which x can take given the value of y becomes (relatively) smaller and smaller w.r.t. the entire extension of the domain [1,A].
- The **estimation error** about the **conditional entropy** increases with tha alphabet cardinality since it is obtained starting from $h(x, y)$ and $h(y)$ for which the estimation error has exactly the same behaviour.

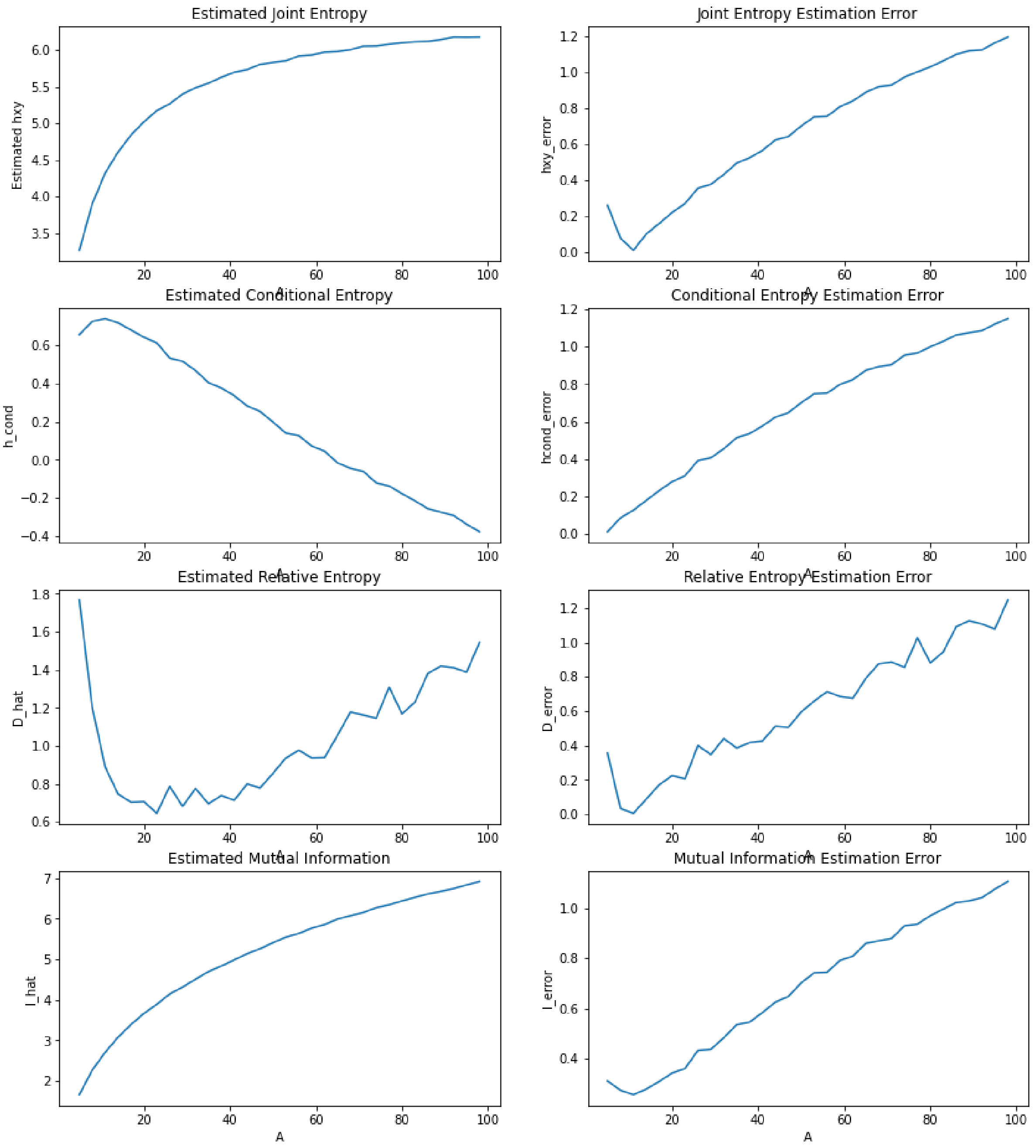


Figure 10: Estimates (left) and Estimation Errors (right) regarding the "joint" quantities of a vector $x \sim \mathcal{U}[1, A]$ and $y = x + z$ where $z \sim \mathcal{U}[-1, 1]$.

- The **relative entropy** is large for small A when the shift produced on y by z is not neglegibile (and hence when x and y have a significantly different alphabet). It then increases with the alphabet cardinality this sould not be the case since for $A \to \infty$, p_x and p_y should tend to the same distribution but it's probably due to the complexity of estimating the 2 single pdfs with "few" samples and an alphabet which becomes greater. (we already saw and commented the problem of estimating "small" probabilities in the discrete case). For the same reason its **estimation error** increases with the cardinality of A.
- The **mutual information** $I(x; y) = h(x) + h(y) - h(x, y)$ increases with the alphabet cardinality since $h(x, y)$ doesn't grow as much as the singular diferential entropies $h(x)$ and $h(y)$.

9.2 Varying L

In Figure 11 it's possible to see how the estimates become better when L increases.

9.3 Varying the width

From Figure 12 note how (analogously to the univariate case) the best performances are obtained in a restricted interval of the width (more or less width$\in [0.1, 0.3]$). This happens for the same reasons of the univariate case (when width << then too much "empty bins", when width >> we can't represent properly the distribution due to low precision).

Note also how we can't see any pattern relating the behaviour of the errors and the number of bins.

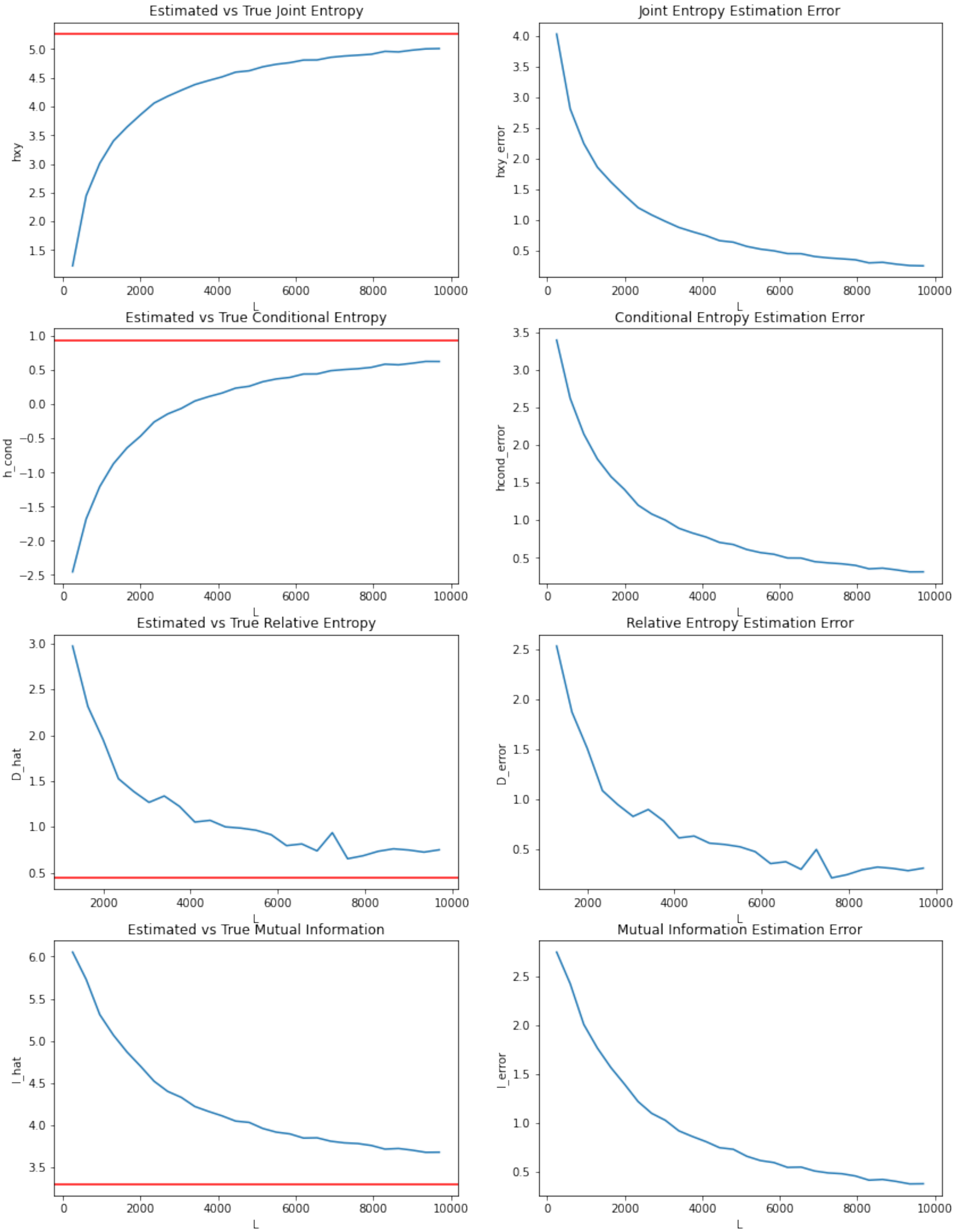


Figure 11: Estimates vs true (left) and estimation errors (right) for a varying L of the joint quantities regarding a uniform x and $y = x + z$ (where z uniform).

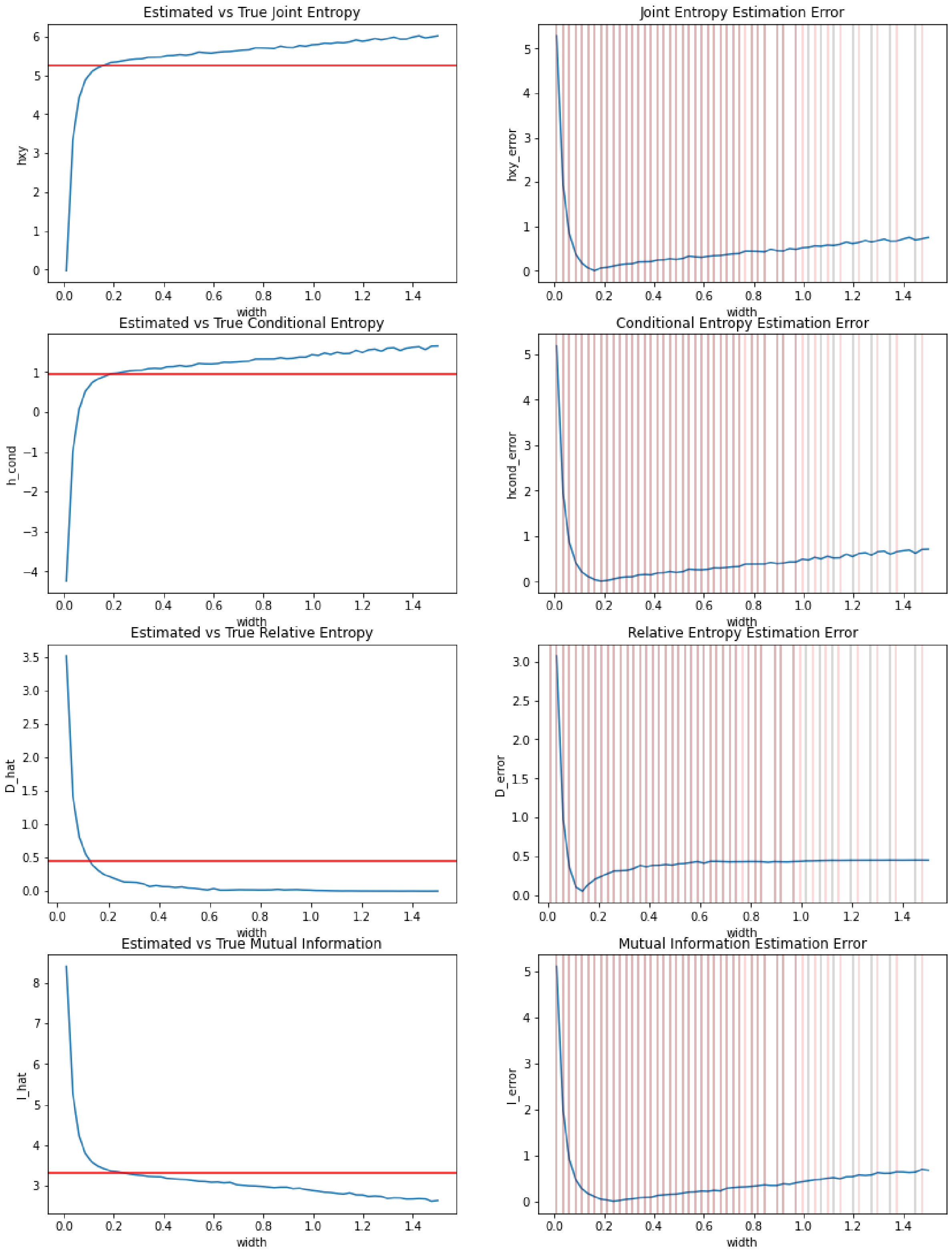


Figure 12: Estimates vs true (left) and estimation errors (right) when varying the width for a uniform x and $y = x + z$ where z uniform. The red and black vertical lines represents the coordinates for which a change (increase) in the width leads to a change (decrease) also in the number of bins (respectively for x and y).

10 Test on x Gaussian and $y = ax + bz$ (z Gaussian)

10.1 Varying m_x and σ_x^2

We now vary the parameter of x. Note that we don't need to vary also the parameters of z since the 2 distribution are "simmetrical".

In Figure 13 note how all the stimates are more or less invariant to m_x (we can see just small oscillations around a certain value). This was expected since the mean of the normal distribution doesn't change the statistical characteristics of the distribution.

In Figure 14 we can see instead how changing the variance of x, produces a variation in the results, note that:

- The **joint entropy** and the related **estimation error** increases when the variance increase since it means dealing with higher uncertainty random variables (same for conditional entropy and mutual information).
- The **conditional entropy** $h(x|y) = h(x, y) - h(y)$ decreases as the variance of x increases. In facts in such a case y becomes similar to a vertically shifted gaussian (since x tends to a uniform as $\sigma_{\text{x}} \uparrow$) and hence the residual uncertainty on x decreases.
- The **relative entropy** tends to decrease since y tends to become more similar (on average) to x, since it will be almost exactly x with a "bonus peak" (given by z). Note also that when σ_x^2 is around 2 (since I used $\sigma_z^2 = 1$) then it starts to occur that $S_x \supset S_y$ (of course theoretically both x and y have infinite domain) and therefore the relative entropy doesn't exist anymore.
- The **mutual information** grows with σ_x^2 for the same reason of the uniform case. Intuitively the more variable is x the less importance will have the peak due to the z contribution in y, making x and y (at the limit for $\sigma_x^2 \to \infty$) one function of the other.

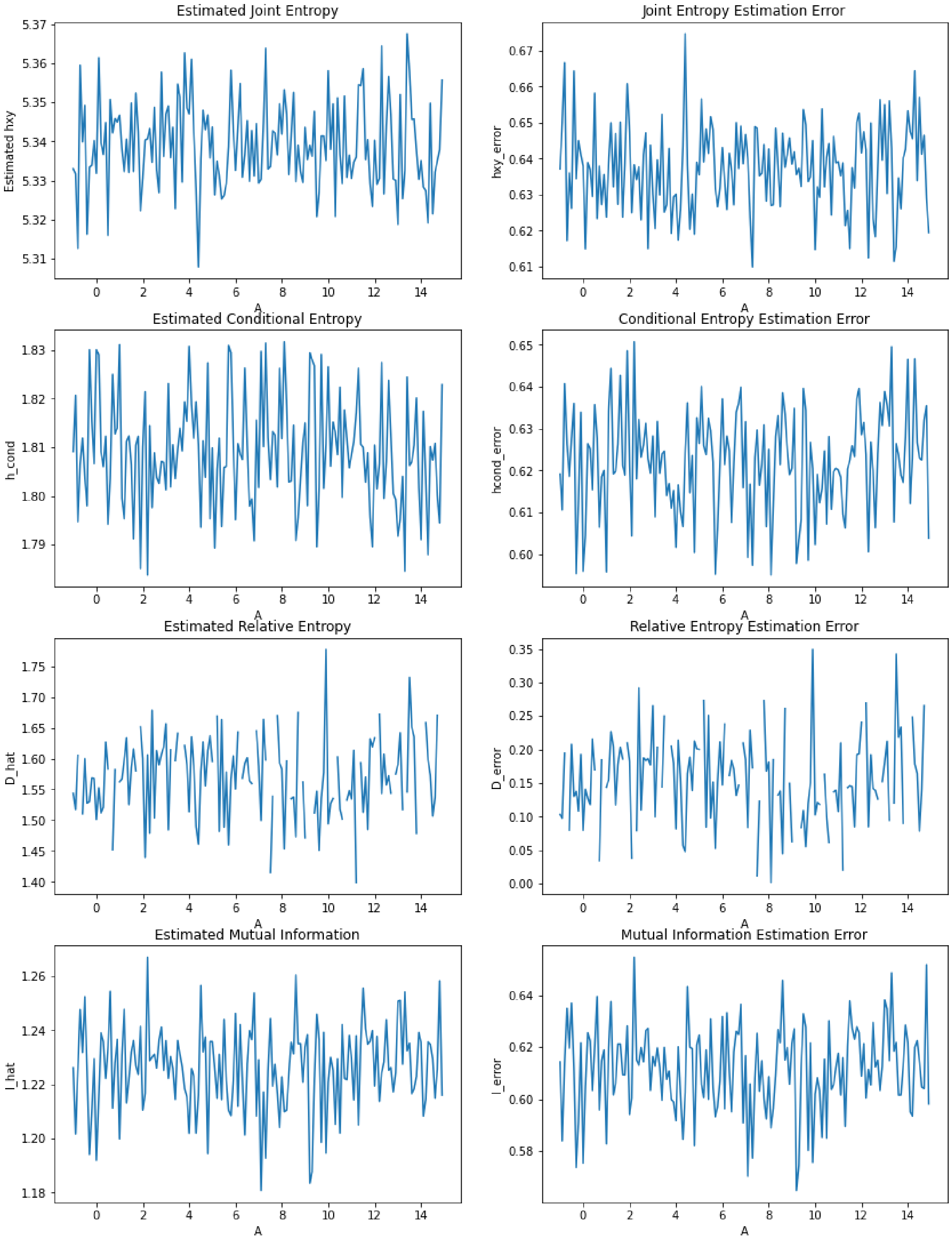
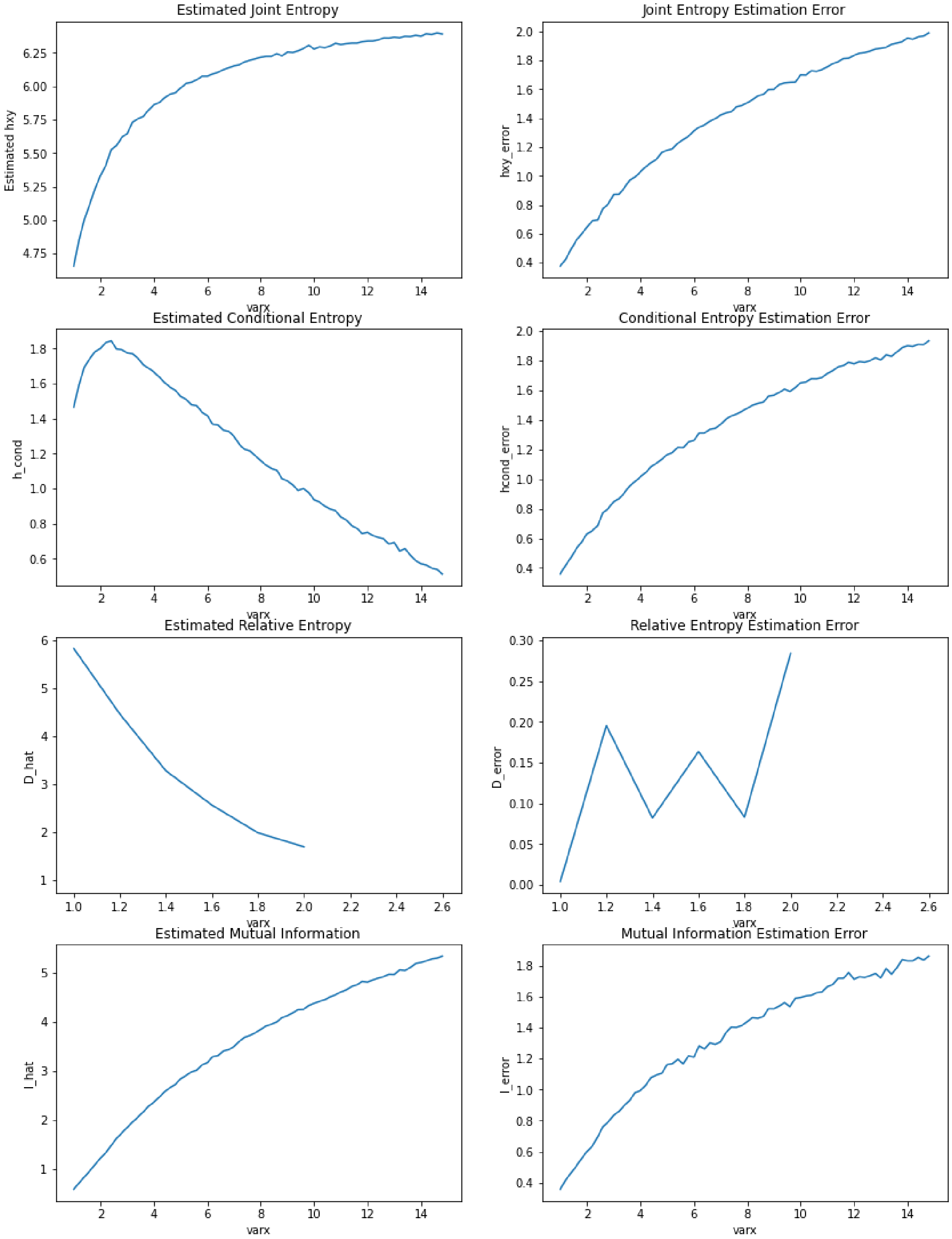


Figure 13: Estimations (left) and estimation errors (right) when varying m_x for x gaussian and $y = ax + bz$ (z gaussian).


Estimated Joint Entropy

Joint Entropy Estimation Error

Estimated Conditional Entropy

Conditional Entropy Estimation Error

Estimated Relative Entropy

Relative Entropy Estimation Error

Estimated Mutual Information

Mutual Information Estimation Error

Figure 14: Estimations (left) and estimation errors (right) when varying σ_x^2 for x gaussian and $y = ax + bz$ (z gaussian).

10.2 Varying the coefficient a in $y = ax + bz$ (x, z gaussian)

In Figure 15 note that:

- The **joint entropy** doesn't vary with the value of a. This is due to the fact that a changes only the alphabet but not the values of the joint pdf $p_{xy}(a, b)$.
- The **conditional entropy** $h(x|y) = h(x, y) - h(y)$ decreases with a this is due to the fact that if we multiply by a constant a r.v. we increase its entropy, namely if we suppose $y = cx$ then:

$$h(y) = h(x) + \log_2(|c|).$$

- The **relative entropy** decreases as a increases since when $a >>$ we have that $y = ax + bz$ is more and more similar to x.
- The **mutual information** (given by $I(x; y) = h(x) + h(y) - h(x, y)$) increases for exactly the same reason of the conditional entropy described in the point above.

10.3 Varying L

Note from Figure 16 how the estimates get better when L increases.

10.4 Varying the width

In Figure 17 note how, onece again the optimal width is in the interval [0.2,0.5] (the same observation on the quality of the estimates hold in this case as in the previous ones).

Note also how we cannot see any pattern relating the change in the number of bins and the behaviour of the estimation errors.

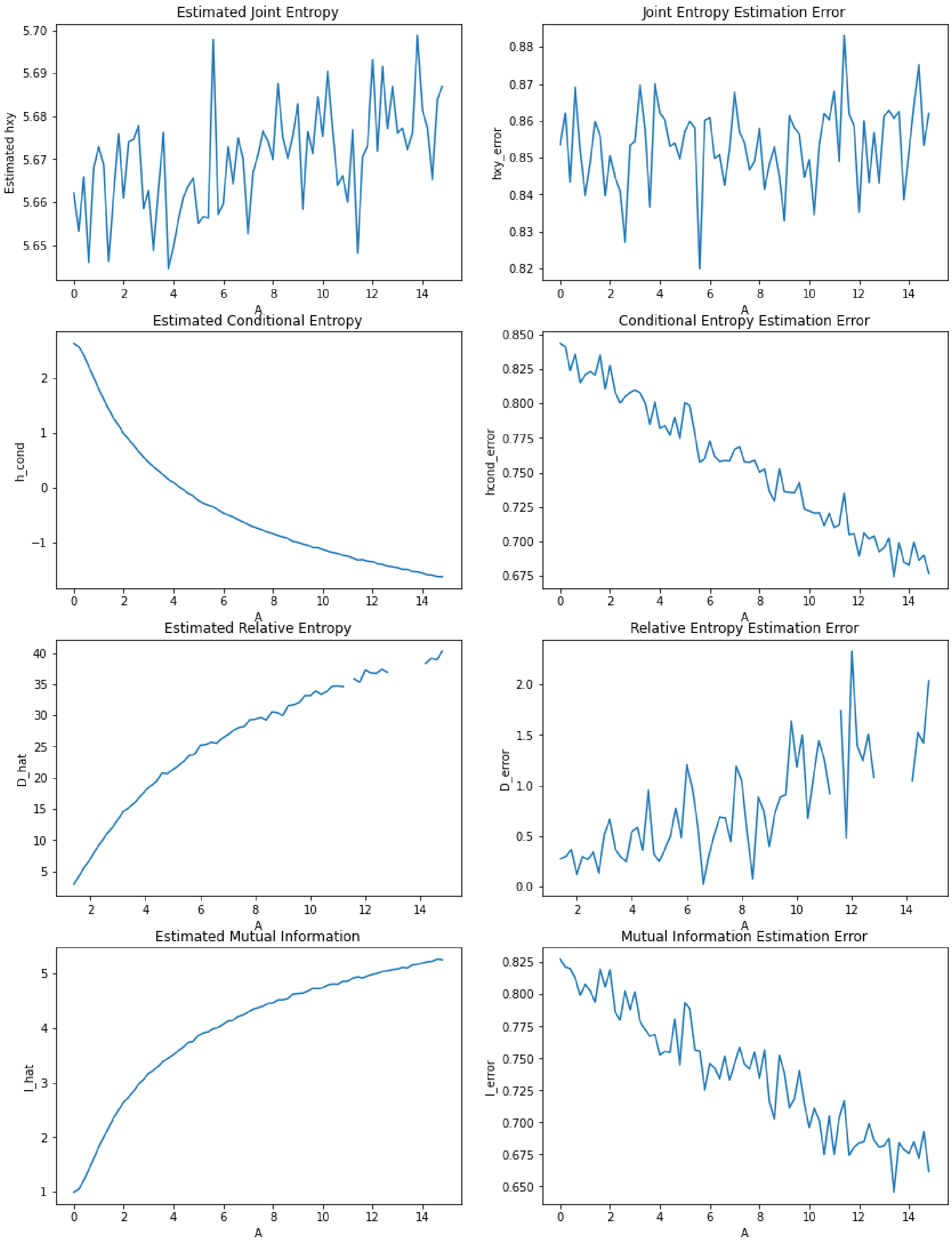


Figure 15: Estimates (left) and estimation errors (right) when varying the coefficient a for x gaussian and $y = x + z$ (z gaussian).

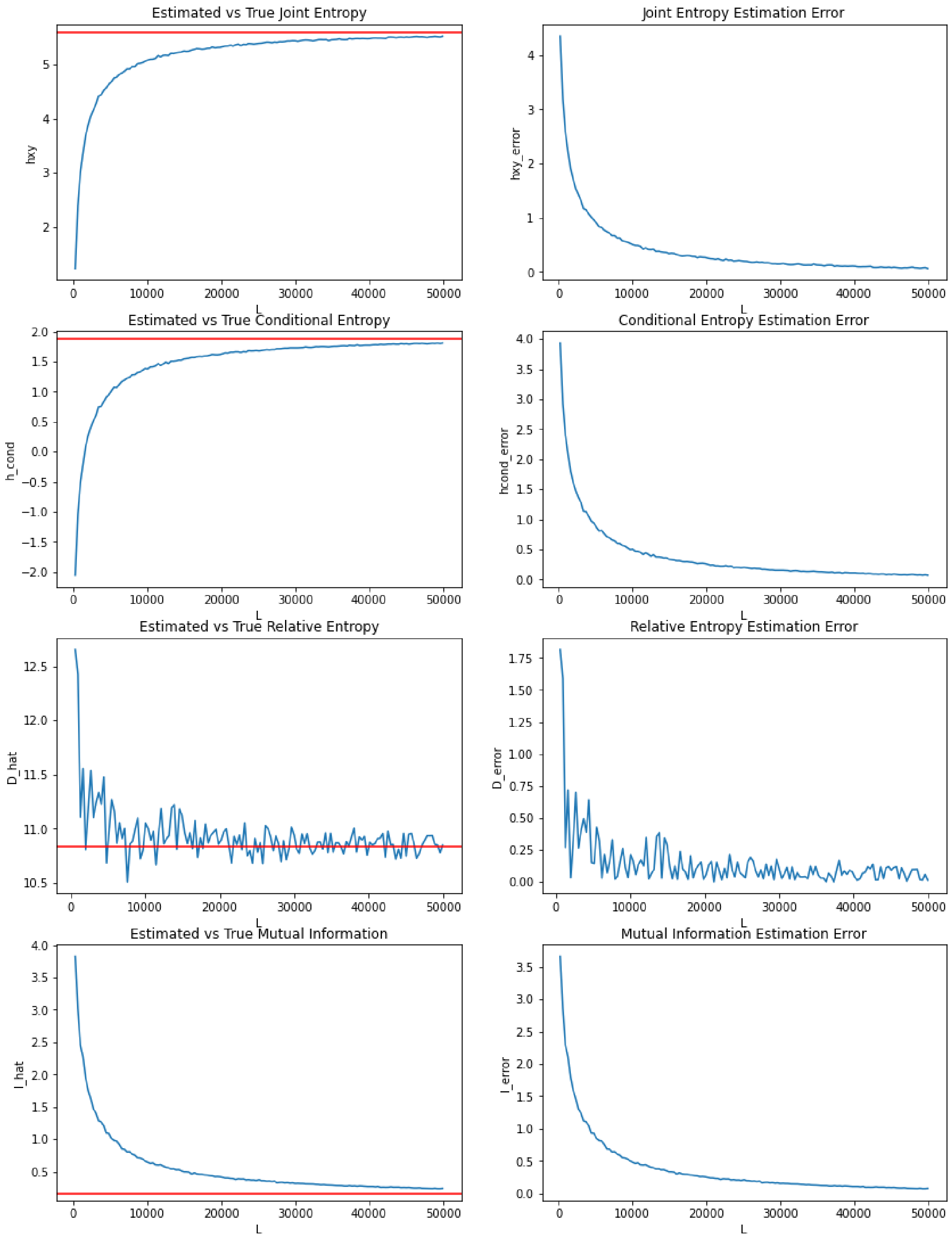


Figure 16: Estimates vs True (left) and estimation errors (right) when varying L, for a gaussian x and $y = ax + bz$ (z gaussian).

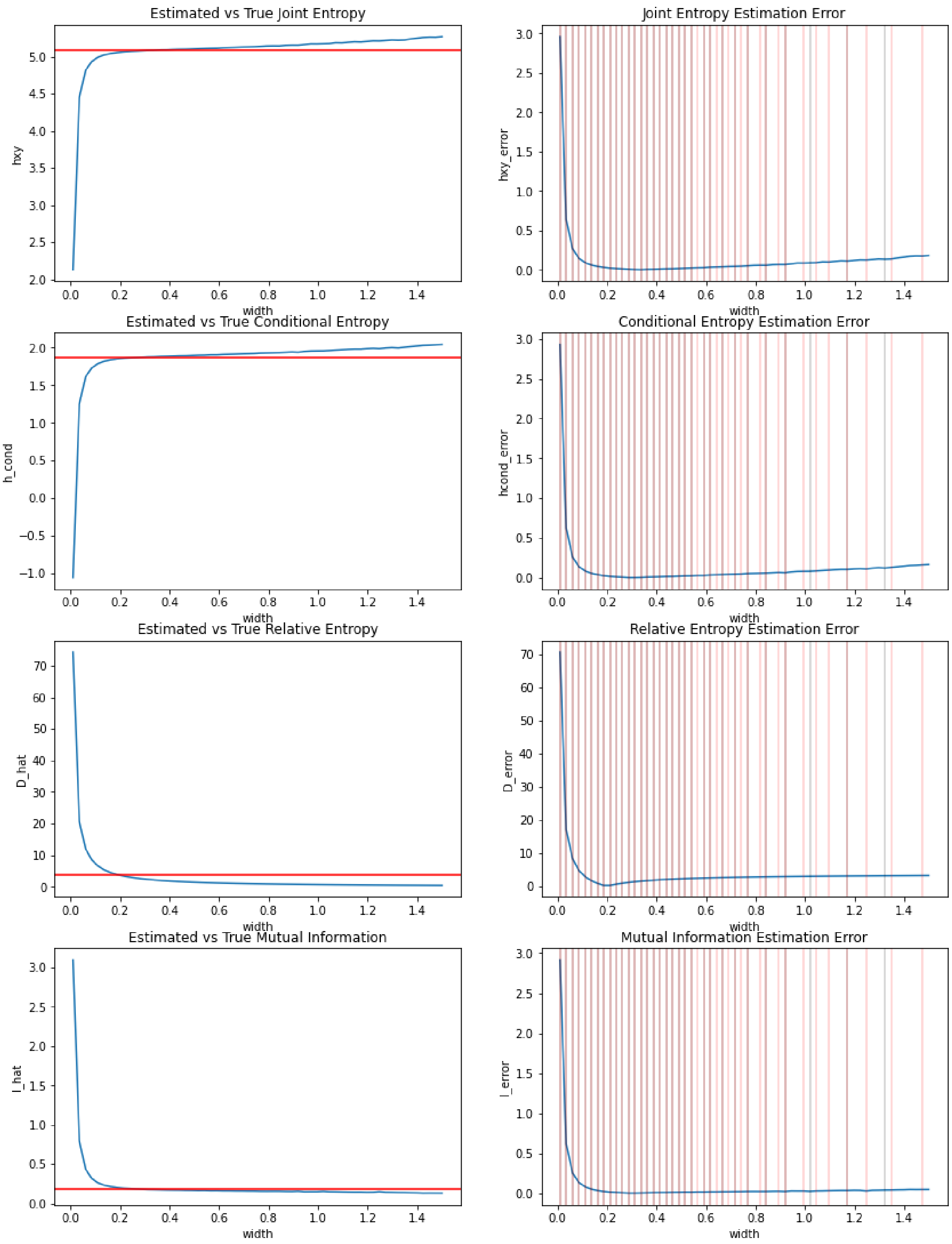


Figure 17: Estimates vs true (left) and estimation errors (right) when varying the width for x gaussian and $y = ax + bz$ (z gaussian). The vertical red and black lines have the same meaning described above (see e.g. Figure 12).