Table 1: Comparison of the PMI_q to the clustering comparison metrics in the systematic review by Gösgens et al. [9]. We consider a metric computationally tractable if its asymptotic complexity is linear in the number of data points N but not necessarily in the numbers of clusters k_A, k_B. The rationale is that in many cases, the number of clusters is much lower than the number of data points and metrics like the AMI_1 with $\mathcal{O}(N \max\{k_A, k_B\})$ are widely used in practice [27, 29]. The PMI_2 is the first metric to be Type II unbiased and monotonous and, while computationally demanding, has efficient approximations.

	NMI	$\mathrm{NMI}_{\max}$	Fair NMI	VI	FMeasure	BCubed	Jaccard	Wallace	Dice	Corr. Coeff.	Sokal&Sneath	Corr. Dist.	Rand Index	AMI_1	AMI_2	SMI_1	SMI_2	PMI_1	PMI_2
Type I unbiased	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓
Type II unbiased	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	○	○	✓	✓
Monotonicity	✓	✗	✗	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓
Comp. tractable	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	○	○

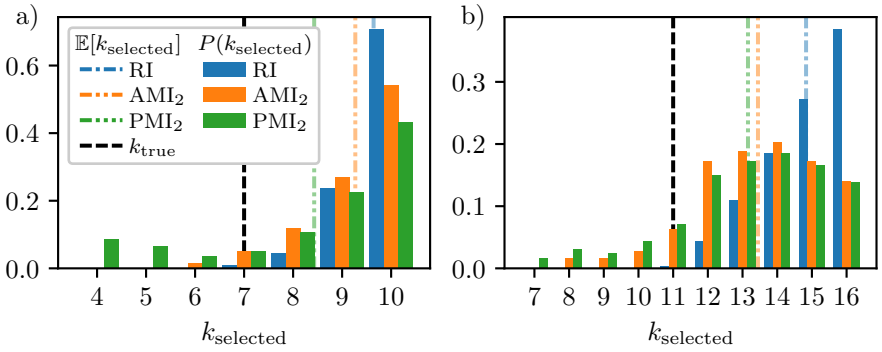

Figure 1: We apply spectral clustering to a) the UCI image segmentation dataset and b) a texture classification dataset. The number of clusters parameter k is set to eleven values between $k_{\text{true}}/2$ and $3k_{\text{true}}/2$. We compare the resulting clusterings with the ground truth via RI, AMI_2, and PMI_2 and select the best clustering k_{selected} according to each metric. We repeat the experiment with 1000 different random seeds and plot the selection probabilities of k_{selected} for each metric. The PMI_2 selects candidates where the number of clusters is closer to the true number of clusters k_{true} on average (dashed lines) compared to the RI and AMI_2.

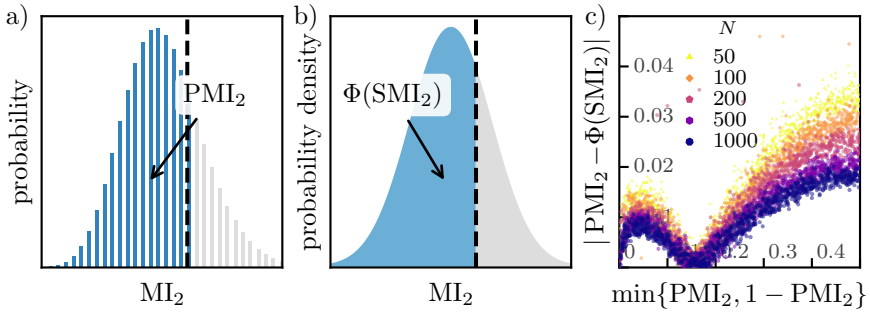

Figure 2: Following Reviewer i4ZB's suggestion, we changed the presentation of Figure 2c. We sampled 1000 pairs of clusterings uniformly at random for different numbers of elements N. We plot the absolute difference between Monte Carlo estimates of the PMI_2 and normalized SMI_2 values as a function of the two-sided p-value. The larger the dataset size N, the better $\Phi(\mathrm{SMI}_2)$ approximates the true PMI_2.