k-Means Clustering with Hölder divergences

Frank Nielsen[1,2], Ke Sun[3], and Stéphane Marchand-Maillet[4]

[1] École Polytechnique [2] Sony Computer Science Laboratories Inc.
[3] King Abdullah University of Science and Technology (KAUST)
[4] University of Geneva

Abstract. We introduced two novel classes of Hölder divergences and Hölder pseudo-divergences that are both invariant to rescaling, and that both encapsulate the Cauchy-Schwarz divergence and the skew Bhattacharyya divergences. We review the elementary concepts of those parametric divergences, and perform a clustering analysis on two synthetic datasets. It is shown experimentally that the symmetrized Hölder divergences consistently outperform significantly the Cauchy-Schwarz divergence in clustering tasks.

1 Introduction

To build dissimilarity measures between p and q in a common domain, one can use bi-parametric inequalities (Mitrinovic *et al.*, 2013) $\mathrm{lhs}(p,q) \leq \mathrm{rhs}(p,q)$, and measure the inequality tightness. When $\mathrm{lhs}(p,q) > 0$, a dissimilarity can be constructed by the log-ratio gap:

$$D(p:q) = -\log\left(\frac{\mathrm{lhs}(p,q)}{\mathrm{rhs}(p,q)}\right) = \log\left(\frac{\mathrm{rhs}(p,q)}{\mathrm{lhs}(p,q)}\right) \geq 0. \quad (1)$$

Notice that this divergence construction allows one to consider the equivalence class of scaled inequalities: $\lambda \times \mathrm{lhs}(p,q) \leq \lambda \times \mathrm{rhs}(p,q), \forall \lambda > 0$. Following this divergence construction principle, we defined Hölder divergences based on the Hölder's inequality, and presented the basic properties of this divergence family (Nielsen *et al.*, 2017). In this paper, we further extend the empirical clustering study with respect to Hölder divergences, and show that symmetrized Hölder divergences consistently outperform significantly the Cauchy-Schwarz divergence (Hasanbelliu *et al.*, 2014). We build Hölder divergences that are invariant by rescaling: These divergences D are called projective divergences and satisfy the property $D(\lambda p : \lambda' q) = D(p:q), \forall \lambda, \lambda' > 0$.

The term "Hölder divergence" was coined previously based on the definition of the Hölder score (Kanamori *et al.*, 2014; Kanamori, 2014): The score-induced Hölder divergence $D(p:q)$ is a proper gap divergence that yields a scale-invariant divergence. A key difference with our work is that this score-induced divergence is not projective and does not include the Cauchy-Schwarz (CS) divergence, while our definition is projective and includes the CS divergence.

This paper is organized as follows: Section 2 reviews the definition of Hölder pseudo divergence and Hölder proper divergence. Section 3 gives algorithms for

clustering based on Hölder divergences, and presents the experimental clustering results. Section 4 concludes this work.

2 Hölder Divergences: Definitions and properties

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measurable space where μ is the Lebesgue measure, and let $L^\gamma(\mathcal{X}, \mu)$ denote the space of functions with their γ-th power of absolute value Lebesgue integrable, for any $\gamma > 0$. When $\gamma \geq 1$, this is a Lebesgue space but we consider the wider scope of $\gamma > 0$ in this work. Hölder's inequality states that $\|pq\|_1 \leq \|p\|_\alpha \|q\|_\beta$ for conjugate exponents $\alpha > 0$ and $\beta > 0$ (satisfying $\frac{1}{\alpha} + \frac{1}{\beta} = 1$), $p \in L^\alpha(\mathcal{X}, \mu)$ and $q \in L^\beta(\mathcal{X}, \mu)$. Let $p(x) \in L^{\alpha\sigma}(\mathcal{X}, \mu)$ and $q(x) \in L^{\beta\tau}(\mathcal{X}, \mu)$ be positive measures where $\sigma > 0$ and $\tau > 0$ are prescribed parameters. We define (Nielsen *et al.*, 2017) a tri-parametric family of divergences as follows:

Definition 1 (Hölder pseudo-divergence). *The Hölder pseudo-divergence (HPD) between $p(x)$ and $q(x)$ is the log-ratio-gap:*

$$D^{\mathrm{H}}_{\alpha,\sigma,\tau}(p : q) := -\log \left(\frac{\int_{\mathcal{X}} p(x)^\sigma q(x)^\tau \mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^{\alpha\sigma} \mathrm{d}x\right)^{\frac{1}{\alpha}} \left(\int_{\mathcal{X}} q(x)^{\beta\tau} \mathrm{d}x\right)^{\frac{1}{\beta}}} \right).$$

The non-negativeness follows straightforwardly from Hölder's inequality (1889). However the symmetry, the triangle-inequality, and the law of indiscernibles (self distance equals to zero) are not satisfied for HPDs. Therefore these dissimilarity measures are said to belong to a broader class of "pseudo-divergences."

In order to have a proper divergence with the law of the indiscernibles, we note that the equality $D^{\mathrm{H}}_{\alpha,\sigma,\tau}(p : q) = 0$ holds if and only if $p(x)^{\alpha\sigma} \propto q(x)^{\beta\tau}$ (almost everywhere). To make this equality condition to be $p(x) = q(x)$ (ae.) for probability distributions, we take $\gamma := \alpha\sigma = \beta\tau$.

Let $p(x)$ and $q(x)$ be positive measures in $L^\gamma(\mathcal{X}, \mu)$ for a prescribed scalar value $\gamma > 0$. Let $\alpha, \beta > 0$ be conjugate exponents. We define (Nielsen *et al.*, 2017) a bi-parametric divergence family, which is a sub-family of HPD that satisfies both non-negativeness and law of the indiscernibles as follows:

Definition 2 (Proper Hölder divergence). *The proper Hölder divergence (HD) between two densities $p(x)$ and $q(x)$ is:*

$$D^{\mathrm{H}}_{\alpha,\gamma}(p : q) = D^{\mathrm{H}}_{\alpha,\frac{\gamma}{\alpha},\frac{\gamma}{\beta}}(p : q) := -\log \left(\frac{\int_{\mathcal{X}} p(x)^{\gamma/\alpha} q(x)^{\gamma/\beta} \mathrm{d}x}{(\int_{\mathcal{X}} p(x)^\gamma \mathrm{d}x)^{1/\alpha} (\int_{\mathcal{X}} q(x)^\gamma \mathrm{d}x)^{1/\beta}} \right).$$

Notice that D^{H} is used to denote both HPD and HD. One has to check the number of subscripts to distinguish between these two pseudo and proper cases.

An important fact about Hölder divergences is that they encapsulate both the Cauchy-Schwarz divergence and the one-parameter family of skew Bhattacharyya divergences (Nielsen and Boltz, 2011). In the definition of HD, setting

$\alpha = \beta = \gamma = 2$ yields the CS divergence:

$$D^{\mathrm{H}}_{2,2}(p:q) = D^{\mathrm{H}}_{2,1,1}(p:q) = \mathrm{CS}(p:q) := -\log\left(\frac{\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^2\mathrm{d}x\right)^{\frac{1}{2}}\left(\int_{\mathcal{X}} q(x)^2\mathrm{d}x\right)^{\frac{1}{2}}}\right).$$

In the definition of HD, setting $\gamma = 1$ yields the skew Bhattacharyya divergences:

$$D^{\mathrm{H}}_{\alpha,1}(p:q) = D^{\mathrm{H}}_{\alpha,\frac{1}{\alpha},\frac{1}{\beta}}(p:q) = -\log\int_{\mathcal{X}} p(x)^{1/\alpha}q(x)^{1/\beta}\mathrm{d}x := B_{1/\alpha}(p:q).$$

It is easy to check from definition 1 that the HPD is a projective divergence which is invariant to scaling of its parameter densities:

$$D^{\mathrm{H}}_{\alpha,\sigma,\tau}(\lambda p:\lambda' q) = D^{\mathrm{H}}_{\alpha,\sigma,\tau}(p:q) \quad (\forall\lambda,\lambda' > 0).$$

Figure 1 illustrates the relationships between those divergence families.

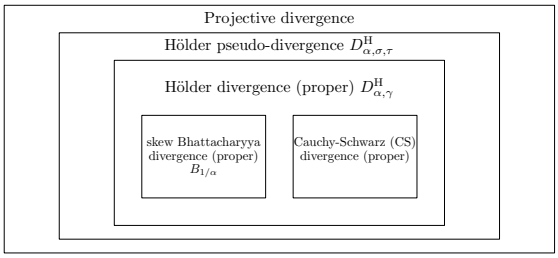

Fig. 1: Hölder proper divergence (bi-parametric) and Hölder improper pseudo-divergence (tri-parametric) encompass the Cauchy-Schwarz divergence and the skew Bhattacharyya divergences.

By definition, the HPD is asymmetric and satisfies the reference duality:

$$D^{\mathrm{H}}_{\alpha,\sigma,\tau}(p:q) = D^{\mathrm{H}}_{\beta,\tau,\sigma}(q:p),$$

for conjugate exponents α and β. Similarly, the HD satisfies:

$$D^{\mathrm{H}}_{\alpha,\gamma}(p:q) = D^{\mathrm{H}}_{\beta,\gamma}(q:p).$$

The HPD and the HD admit closed-form formulas for exponential family distributions. For example, consider $p(x) = \exp(\theta_p^\top t(x) - F(\theta_p))$ and $q(x) = \exp(\theta_q^\top t(x) - F(\theta_q))$, where $t(x)$ is a vector of sufficient statistics, and $F(\theta)$ is the convex cumulant generating function. Then from straightforward derivations, the symmetrized Hölder divergence is:

$$\begin{aligned} S^{\mathrm{H}}_{\alpha,\gamma}(p:q) &:= \frac{1}{2}\left(D^{\mathrm{H}}_{\alpha,\gamma}(p:q) + D^{\mathrm{H}}_{\alpha,\gamma}(q:p)\right) \\ &= \frac{1}{2}\left[F(\gamma\theta_p) + F(\gamma\theta_q) - F\left(\frac{\gamma}{\alpha}\theta_p + \frac{\gamma}{\beta}\theta_q\right) - F\left(\frac{\gamma}{\beta}\theta_p + \frac{\gamma}{\alpha}\theta_q\right)\right]. \end{aligned}$$

This $S^{\mathrm{H}}_{\alpha,\gamma}$ has the key advantage that its centroid can be solved efficiently using the concave-convex procedure (CCCP) (Nielsen and Boltz, 2011). Consider a set of fixed densities $\{\theta_1, \cdots, \theta_n\}$ with positive weights $\{w_1, \cdots, w_n\}$ ($\sum_{i=1}^n w_i = 1$) of the same exponential family. The symmetrized HD centroid with respect to $\alpha, \gamma > 0$ is defined as:

$$\begin{aligned} O_{\alpha,\gamma} &:= \arg\min_{\theta} \sum_{i=1}^{n} w_i S^{\mathrm{H}}_{\alpha,\gamma}(\theta_i : \theta) \\ &= \arg\min_{\theta} \sum_{i=1}^{n} w_i \left[F(\gamma\theta) - F\left(\frac{\gamma}{\alpha}\theta_i + \frac{\gamma}{\beta}\theta\right) - F\left(\frac{\gamma}{\beta}\theta_i + \frac{\gamma}{\alpha}\theta\right)\right]. \end{aligned} \tag{2}$$

Because $F(\theta)$ is a strictly convex function, the energy function to be minimized in eq. (2) is the difference between two convex functions. Setting the derivatives to zero, we get the CCCP iterations given by:

$$O^{t+1}_{\alpha,\gamma} = \frac{1}{\gamma}(\nabla F)^{-1}\left[\sum_{i=1}^{n} w_i \left(\frac{1}{\beta}\nabla F\left(\frac{\gamma}{\alpha}\theta_i + \frac{\gamma}{\beta}O^t_{\alpha,\gamma}\right) + \frac{1}{\alpha}\nabla F\left(\frac{\gamma}{\beta}\theta_i + \frac{\gamma}{\alpha}O^t_{\alpha,\gamma}\right)\right)\right],$$

where ∇F and $(\nabla F)^{-1}$ are forward and reverse transformations between the natural parameters and the dual expectation parameters, respectively.

3 Clustering Based on Symmetric Hölder Divergences

Given a set of densities $\{p_1, \cdots, p_n\}$, we can perform variational k-means (Nielsen and Nock, 2015) clustering based on $S^{\mathrm{H}}_{\alpha,\gamma}$. The cost function is the Hölder information:

$$E := \sum_{i=1}^{n} S^{\mathrm{H}}_{\alpha,\gamma}(p_i : O_{l_i}), \tag{3}$$

where $O_1, \cdots, O_L$ are the cluster centers and $l_i \in \{1, \cdots, L\}$ is the cluster label of p_i. Algorithm 1 presents a revision of the clustering algorithm given in (Nielsen *et al.*, 2017) with k-means++ initialization (Arthur and Vassilvitskii, 2007).

We investigate two different datasets. The first (Nielsen *et al.*, 2017) consists of n random 2D Gaussians with two or three clusters. In the first cluster, the mean of each Gaussian $G(\mu, \Sigma)$ has the prior distribution $\mu \sim G((-2, 0), I)$; the covariance matrix is obtained by first generating $\sigma_1 \sim \Gamma(7, 0.01)$, $\sigma_2 \sim \Gamma(7, 0.003)$, where Γ means a gamma distribution with prescribed shape and scale, then rotating the covariance matrix $\mathrm{diag}(\sigma_1, \sigma_2)$ so that the resulting Gaussian has a "radial direction" with respect to the center $(-2, 0)$. The second and third clusters are similar to the first cluster with the only difference being that their μ's are centered around $(2, 0)$ and $(0, 2\sqrt{3})$, respectively.

The second dataset consists of multinomial distributions in Δ_9, the 9D probability simplex. The dataset presents two or three clusters. For each cluster, we first pick a random center $(c_0, \cdots, c_d)$ based on the uniform distribution in Δ_9. Then we randomly generate a distribution $(p_0, \cdots, p_d)$ based on $p_i =$

Algorithm 1: Hölder variational k-means.

Input: $p_1, \cdots, p_n$; number of clusters L; $1 < \alpha \leq 2$; $\gamma > 0$

Output: A clustering scheme assigning each p_i to a label in $\{1, \cdots, L\}$

Randomly pick one center $O_1 \in \{p_i\}_{i=1}^n$, then sequentially pick O_k $(2 \leq k \leq L)$ with probability proportional to $\min_{j=1}^{k-1} S^{\mathrm{H}}_{\alpha,\gamma}(p_i : O_j)$

while *not converged* **do**

  for $i = 1, \ldots, n$ **do**

    Assign $l_i = \arg\min_l S^{\mathrm{H}}_{\alpha,\gamma}(p_i : O_l)$

  for $l = 1, \ldots, L$ **do**

    `/* Carry CCCP iterations until the current center improves the former cluster Hölder information */`

    Compute the centroid $O_l = \arg\min_O \sum_{i:l_i=l} S^{\mathrm{H}}_{\alpha,\gamma}(p_i : O)$

return $\{l_i\}_{i=1}^n$

$\frac{\exp(\log c_i + \sigma\epsilon_i)}{\sum_{i=0}^d \exp(\log c_i + \sigma\epsilon_i)}$, where $\sigma > 0$ is a noise level parameter, and each ϵ^i follows independently a standard Gaussian distribution. We repeat generating random samples for each cluster center, and make sure that different clusters have almost the same number of samples.

Our clustering algorithm involves two additional hyper-parameters γ and α as compared with standard k-means clustering. Therefore it is meaningful to study how these two hyper-parameters affect the performance. We extend the experiments reported previously (Nielsen *et al.*, 2017) (where $\alpha = \gamma$ is applied for simplicity) with a grid of α and γ values. Notice that the reference duality gives $S^{\mathrm{H}}_{\alpha,\gamma} = S^{\mathrm{H}}_{\beta,\gamma}$ for conjugate exponents α and β. Therefore we assume $1 < \alpha \leq 2$ without loss of generality. If we choose $\alpha = \gamma = 2$, then $S^{\mathrm{H}}_{\alpha,\gamma}$ becomes the CS divergence, and Algorithm 1 reduces to traditional CS clustering.

We performed clustering experiments by setting the number of clusters $k \in \{2, 3\}$ and setting the sample size $n \in \{50, 100\}$. Tables 1 and 2 show the clustering accuracy measured by the Normalized Mutual Information (NMI). The large variance of the clustering accuracy is because different runs are based on different random datasets. We see that the symmetric Hölder divergence can give *strikingly better* clustering as compared to CS clustering. An empirical range of well-performed parameter values is given by $\gamma \in [0.5, 1.5]$ and $\alpha \in (1, 2]$. In practice, one has to setup a configuration grid and apply cross-validation to find the best α and γ values.

This hints that one should use instead the general Hölder divergence to replace CS in similar clustering applications (Hasanbelliu *et al.*, 2014; Rami *et al.*, 2016). Although one faces the problem of tuning the parameter α and γ, Hölder divergences can potentially give much better results.

4 Conclusion

We experimentally confirmed the usefulness of the novel parametric Hölder classes of statistical divergences and pseudo-divergences. In general one should use Hölder clustering instead of Cauchy-Schwarz clustering to get much better results. These new concepts can open up new insights and applications in statistics and information sciences.

Reproducible source code is available online at:

`https://www.lix.polytechnique.fr/~nielsen/HPD/` .

Bibliography

Mitrinovic, D.S.; Pecaric, J.; Fink, A.M. *Classical and New Inequalities in Analysis*; Springer Science & Business Media: New York, NY, USA, 2013; Volume 61.

Kanamori, T.; Fujisawa, H. Affine invariant divergences associated with proper composite scoring rules and their applications. *Bernoulli* **2014**, *20*, 2278–2304.

Kanamori, T. Scale-invariant divergences for density functions. *Entropy* **2014**, *16*, 2611–2628.

Arthur, D.; Vassilvitskii, S. k-means++: The advantages of careful seeding ACM-SIAM symposium on Discrete algorithms, 2007; pp. 1027–1035

Nielsen, F.; Sun, K.; Marchand-Maillet, S. On Hölder projective divergences. *Entropy* **2017**, *19*, 122.

Holder, O.L. Über einen Mittelwertssatz. *Nachr. Akad. Wiss. Gottingen Math. Phys. Kl.* **1889**, vol. 44, 38–47.

Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466.

Nielsen, F.; Nock, R. Total Jensen divergences: Definition, properties and clustering. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Queensland, Australia, 19–24 April 2015; pp. 2016–2020.

Hasanbelliu, E.; Giraldo, L.S.; Principe, J.C. Information theoretic shape matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2436–2451.

Rami, H.; Belmerhnia, L.; Drissi El Maliani, A.; El Hassouni, M. Texture Retrieval Using Mixtures of Generalized Gaussian Distribution and Cauchy-Schwarz Divergence in Wavelet Domain. *Image Commun.* **2016**, *42*, 45–58.

Table 1: Performance in NMI (mean±std) when clustering 2D Gaussians based on 1000 independent runs for each configuration. Bold numbers indicate the best obtained performance. The boxed numbers are given by the Cauchy-Schwarz (CS) clustering.

(a) $k = 2$; $n = 50$

	$\alpha = 1.01$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$\gamma = 0.25$	0.91 ± 0.10	0.89 ± 0.13	0.86 ± 0.14	0.85 ± 0.14	0.85 ± 0.15	0.85 ± 0.16
$\gamma = 0.5$	$\mathbf{0.92 \pm 0.09}$	0.86 ± 0.16	0.84 ± 0.17	0.84 ± 0.17	0.82 ± 0.19	0.82 ± 0.17
$\gamma = 0.75$	$\mathbf{0.92 \pm 0.10}$	0.85 ± 0.16	0.83 ± 0.17	0.82 ± 0.18	0.82 ± 0.19	0.81 ± 0.18
$\gamma = 1$	$\mathbf{0.92 \pm 0.10}$	0.84 ± 0.18	0.81 ± 0.20	0.82 ± 0.18	0.82 ± 0.20	0.81 ± 0.20
$\gamma = 1.5$	$\mathbf{0.92 \pm 0.10}$	0.82 ± 0.18	0.80 ± 0.20	0.81 ± 0.19	0.81 ± 0.19	0.80 ± 0.21
$\gamma = 2$	$\mathbf{0.92 \pm 0.10}$	0.81 ± 0.20	0.82 ± 0.19	0.80 ± 0.21	0.80 ± 0.20	$\boxed{0.81 \pm 0.20}$

(b) $k = 2$; $n = 100$

	$\alpha = 1.01$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$\gamma = 0.25$	0.91 ± 0.07	0.88 ± 0.08	0.87 ± 0.09	0.86 ± 0.10	0.86 ± 0.10	0.86 ± 0.10
$\gamma = 0.5$	0.91 ± 0.07	0.87 ± 0.12	0.85 ± 0.11	0.85 ± 0.13	0.84 ± 0.14	0.84 ± 0.12
$\gamma = 0.75$	0.91 ± 0.07	0.86 ± 0.11	0.84 ± 0.13	0.84 ± 0.13	0.84 ± 0.14	0.84 ± 0.14
$\gamma = 1$	$\mathbf{0.92 \pm 0.07}$	0.86 ± 0.12	0.83 ± 0.15	0.83 ± 0.13	0.83 ± 0.14	0.84 ± 0.12
$\gamma = 1.5$	$\mathbf{0.92 \pm 0.07}$	0.84 ± 0.14	0.83 ± 0.14	0.83 ± 0.15	0.83 ± 0.14	0.83 ± 0.13
$\gamma = 2$	0.91 ± 0.08	0.84 ± 0.14	0.82 ± 0.15	0.83 ± 0.14	0.83 ± 0.14	$\boxed{0.83 \pm 0.14}$

(c) $k = 3$; $n = 50$

	$\alpha = 1.01$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$\gamma = 0.25$	0.88 ± 0.12	0.83 ± 0.14	0.81 ± 0.15	0.80 ± 0.15	0.79 ± 0.14	0.80 ± 0.15
$\gamma = 0.5$	0.88 ± 0.12	0.80 ± 0.15	0.77 ± 0.16	0.77 ± 0.15	0.77 ± 0.15	0.76 ± 0.16
$\gamma = 0.75$	$\mathbf{0.89 \pm 0.12}$	0.80 ± 0.14	0.77 ± 0.15	0.76 ± 0.16	0.75 ± 0.15	0.76 ± 0.16
$\gamma = 1$	0.88 ± 0.12	0.78 ± 0.15	0.76 ± 0.16	0.75 ± 0.16	0.75 ± 0.16	0.76 ± 0.15
$\gamma = 1.5$	0.88 ± 0.13	0.76 ± 0.16	0.76 ± 0.16	0.76 ± 0.15	0.76 ± 0.16	0.76 ± 0.16
$\gamma = 2$	0.88 ± 0.12	0.76 ± 0.16	0.75 ± 0.16	0.74 ± 0.16	0.75 ± 0.16	$\boxed{0.76 \pm 0.16}$

(d) $k = 3$; $n = 100$

	$\alpha = 1.01$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$\gamma = 0.25$	$\mathbf{0.89 \pm 0.08}$	0.84 ± 0.11	0.82 ± 0.12	0.82 ± 0.11	0.82 ± 0.11	0.82 ± 0.12
$\gamma = 0.5$	$\mathbf{0.89 \pm 0.08}$	0.83 ± 0.11	0.81 ± 0.12	0.79 ± 0.12	0.78 ± 0.14	0.79 ± 0.13
$\gamma = 0.75$	$\mathbf{0.89 \pm 0.09}$	0.81 ± 0.12	0.79 ± 0.13	0.78 ± 0.13	0.77 ± 0.14	0.78 ± 0.14
$\gamma = 1$	0.88 ± 0.10	0.80 ± 0.12	0.78 ± 0.14	0.78 ± 0.14	0.78 ± 0.13	0.78 ± 0.13
$\gamma = 1.5$	$\mathbf{0.89 \pm 0.09}$	0.78 ± 0.13	0.77 ± 0.14	0.77 ± 0.14	0.76 ± 0.14	0.77 ± 0.14
$\gamma = 2$	$\mathbf{0.89 \pm 0.09}$	0.78 ± 0.13	0.77 ± 0.13	0.77 ± 0.14	0.77 ± 0.13	$\boxed{0.78 \pm 0.13}$

Table 2: Performance in NMI (mean±std) when clustering multinomial distributions in Δ_9 based on 1000 independent runs for each configuration. Bold numbers indicate the best obtained performance. The boxed numbers are given by the Cauchy-Schwarz (CS) clustering.

(a) $k = 2$; $n = 50$

	$\alpha = 1.01$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$\gamma = 0.25$	$\mathbf{0.93 \pm 0.14}$	$\mathbf{0.93 \pm 0.15}$	$\mathbf{0.93 \pm 0.13}$	$\mathbf{0.93 \pm 0.15}$	0.92 ± 0.16	$\mathbf{0.93 \pm 0.13}$
$\gamma = 0.5$	0.91 ± 0.16	0.92 ± 0.15	0.90 ± 0.18	0.91 ± 0.17	0.91 ± 0.16	0.91 ± 0.16
$\gamma = 0.75$	0.87 ± 0.20	0.86 ± 0.21	0.87 ± 0.20	0.87 ± 0.20	0.88 ± 0.19	0.88 ± 0.19
$\gamma = 1$	0.83 ± 0.23	0.83 ± 0.23	0.83 ± 0.23	0.82 ± 0.24	0.81 ± 0.23	0.82 ± 0.23
$\gamma = 1.5$	0.75 ± 0.26	0.71 ± 0.28	0.72 ± 0.27	0.70 ± 0.28	0.71 ± 0.28	0.71 ± 0.28
$\gamma = 2$	0.68 ± 0.28	0.65 ± 0.29	0.64 ± 0.29	0.62 ± 0.29	0.62 ± 0.30	$\boxed{0.61 \pm 0.30}$

(b) $k = 2$; $n = 100$

	$\alpha = 1.01$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$\gamma = 0.25$	0.93 ± 0.12	0.93 ± 0.12	0.93 ± 0.11	0.93 ± 0.12	$\mathbf{0.94 \pm 0.11}$	0.93 ± 0.12
$\gamma = 0.5$	0.92 ± 0.14	0.91 ± 0.14	0.92 ± 0.13	0.91 ± 0.15	0.92 ± 0.14	0.91 ± 0.14
$\gamma = 0.75$	0.89 ± 0.16	0.88 ± 0.16	0.89 ± 0.17	0.89 ± 0.16	0.88 ± 0.16	0.89 ± 0.15
$\gamma = 1$	0.83 ± 0.20	0.84 ± 0.19	0.84 ± 0.19	0.83 ± 0.19	0.84 ± 0.19	0.84 ± 0.19
$\gamma = 1.5$	0.77 ± 0.24	0.74 ± 0.25	0.74 ± 0.23	0.73 ± 0.24	0.74 ± 0.23	0.74 ± 0.24
$\gamma = 2$	0.70 ± 0.26	0.67 ± 0.26	0.65 ± 0.27	0.64 ± 0.27	0.63 ± 0.27	$\boxed{0.63 \pm 0.27}$

(c) $k = 3$; $n = 50$

	$\alpha = 1.01$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$\gamma = 0.25$	$\mathbf{0.87 \pm 0.16}$	$\mathbf{0.87 \pm 0.16}$	$\mathbf{0.87 \pm 0.16}$	$\mathbf{0.87 \pm 0.15}$	$\mathbf{0.87 \pm 0.16}$	0.86 ± 0.16
$\gamma = 0.5$	0.84 ± 0.17	0.84 ± 0.17	0.84 ± 0.17	0.83 ± 0.17	0.84 ± 0.17	0.84 ± 0.18
$\gamma = 0.75$	0.80 ± 0.18	0.79 ± 0.18	0.79 ± 0.18	0.78 ± 0.19	0.79 ± 0.18	0.78 ± 0.19
$\gamma = 1$	0.73 ± 0.20	0.72 ± 0.20	0.73 ± 0.20	0.73 ± 0.20	0.72 ± 0.20	0.71 ± 0.21
$\gamma = 1.5$	0.65 ± 0.21	0.63 ± 0.20	0.61 ± 0.19	0.59 ± 0.20	0.59 ± 0.20	0.60 ± 0.20
$\gamma = 2$	0.57 ± 0.20	0.55 ± 0.20	0.53 ± 0.19	0.51 ± 0.18	0.52 ± 0.18	$\boxed{0.51 \pm 0.18}$

(d) $k = 3$; $n = 100$

	$\alpha = 1.01$	$\alpha = 1.2$	$\alpha = 1.4$	$\alpha = 1.6$	$\alpha = 1.8$	$\alpha = 2$
$\gamma = 0.25$	$\mathbf{0.90 \pm 0.13}$	0.88 ± 0.14	0.88 ± 0.14	0.88 ± 0.13	0.89 ± 0.13	0.89 ± 0.12
$\gamma = 0.5$	0.87 ± 0.14	0.86 ± 0.14	0.86 ± 0.15	0.86 ± 0.14	0.86 ± 0.14	0.86 ± 0.14
$\gamma = 0.75$	0.82 ± 0.16	0.82 ± 0.17	0.83 ± 0.15	0.82 ± 0.16	0.82 ± 0.16	0.82 ± 0.16
$\gamma = 1$	0.77 ± 0.18	0.77 ± 0.17	0.77 ± 0.18	0.75 ± 0.18	0.76 ± 0.18	0.76 ± 0.18
$\gamma = 1.5$	0.66 ± 0.19	0.63 ± 0.19	0.64 ± 0.19	0.63 ± 0.19	0.63 ± 0.19	0.63 ± 0.19
$\gamma = 2$	0.57 ± 0.18	0.56 ± 0.18	0.54 ± 0.18	0.53 ± 0.18	0.53 ± 0.19	$\boxed{0.53 \pm 0.18}$