
Supplementary Materials: Alpha-Beta Divergences Discover Micro and Macro Structures in Data

We first discuss details regarding continuity in Alpha-Beta SNE which allow us to make settings, e.g., $(\alpha, \alpha + \beta) = (1, 1)$ used in t-SNE, while bypassing issues arising from singularity. We then display a smattering of ABSNE plots across a large selection of datasets.

1. Continuity and Gradients for the Alpha-Beta Objective

The form of the original Alpha-Beta divergence used in the paper, defined $\mathcal{J}(\mathcal{E}; \alpha, \beta) = D_{AB}^{\alpha\beta}(\mathbf{P} \parallel \mathbf{Q})$, is computed as:

$$\frac{1}{\alpha\beta} \sum_{i \neq j} \left(-\mathbf{P}_{ij}^\alpha \mathbf{Q}_{ij}^\beta + \frac{\alpha}{\alpha + \beta} \mathbf{P}_{ij}^{\alpha+\beta} + \frac{\beta}{\alpha + \beta} \mathbf{Q}_{ij}^{\alpha+\beta} \right), \quad (1)$$

where $\alpha \in \mathbb{R} \setminus \{0\}$, $\beta \in \mathbb{R}$ are hyperparameters. According to (Cichocki et al., 2011), to account for cases such as $\beta = 0$ and $\alpha + \beta = 0$ in this objective, we can re-define:

$$D_{AB}^{\alpha\beta}(\mathbf{P} \parallel \mathbf{Q}) = \sum_{ij} d_{AB}^{(\alpha,\beta)}(\mathbf{P}_{ij}, \mathbf{Q}_{ij}), \quad (2)$$

where:

$$d_{AB}^{(\alpha,\beta)}(\mathbf{P}_{ij}, \mathbf{Q}_{ij}) = \begin{cases} -\frac{1}{\alpha\beta} \left(\mathbf{P}_{ij}^\alpha \mathbf{Q}_{ij}^\beta - \frac{\alpha}{\alpha + \beta} \mathbf{P}_{ij}^{\alpha+\beta} - \frac{\alpha}{\alpha + \beta} \mathbf{Q}_{ij}^{\alpha+\beta} \right), & \alpha, \beta, \alpha + \beta \neq 0 \\ \frac{1}{\alpha^2} \left(\mathbf{P}_{ij}^\alpha \ln \frac{\mathbf{P}_{ij}^\alpha}{\mathbf{Q}_{ij}^\alpha} - \mathbf{P}_{ij}^\alpha + \mathbf{Q}_{ij}^\alpha \right), & \alpha \neq 0, \beta = 0 \\ \frac{1}{\alpha^2} \left(\ln \frac{\mathbf{Q}_{ij}^\alpha}{\mathbf{P}_{ij}^\alpha} + \frac{\mathbf{P}_{ij}^\alpha}{\mathbf{Q}_{ij}^\alpha} - 1 \right), & \alpha = -\beta \neq 0 \\ \frac{1}{\beta^2} \left(\mathbf{Q}_{ij}^\beta \ln \frac{\mathbf{Q}_{ij}^\beta}{\mathbf{P}_{ij}^\beta} + \mathbf{P}_{ij}^\beta \right), & \alpha = 0, \beta \neq 0 \\ \frac{1}{2} (\ln \mathbf{P}_{ij} - \ln \mathbf{Q}_{ij})^2, & \alpha = \beta = 0 \end{cases} \quad (3)$$

As described in the paper, note that we obtain the KL-divergence by setting $\alpha = 1, \beta = 0$ and the Itakura-Saito divergence by setting $\alpha = 1, \beta = -1$, since \mathbf{P} and \mathbf{Q} are probability distributions. Given these various definitions, we would like to ensure that the gradient described in the paper is the same particularly for the first three cases (namely where $\alpha \neq 0$), as these are the ones that arise in our reductions. Specifically, we need to ensure that the gradients $\partial D_{AB}^{(\alpha,\beta)}(\mathbf{P} \parallel \mathbf{Q}) / \partial \mathbf{y}_i$ all match.

Case 1: $\alpha, \beta \neq 0$. We have that:

$$\frac{\partial D_{AB}^{(\alpha, \beta)}(\mathbf{P} \parallel \mathbf{Q})}{\partial \mathbf{y}_i} = \frac{\partial}{\partial \mathbf{y}_i} \left[-\frac{1}{\alpha\beta} \left(\mathbf{P}_{ij}^\alpha \mathbf{Q}_{ij}^\beta - \frac{\alpha}{\alpha+\beta} \mathbf{P}_{ij}^{\alpha+\beta} - \frac{\beta}{\alpha+\beta} \mathbf{Q}_{ij}^{\alpha+\beta} \right) \right] \quad (4)$$

$$= \frac{\partial}{\partial \mathbf{y}_i} \left[-\frac{1}{\alpha\beta} \mathbf{P}_{ij}^\alpha \mathbf{Q}_{ij}^\beta + \frac{1}{\alpha(\alpha+\beta)} \mathbf{Q}_{ij}^{\alpha+\beta} \right] \quad (5)$$

$$= \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i} \cdot \left[-\frac{1}{\alpha\beta} \mathbf{P}_{ij}^\alpha \cdot \beta \cdot \mathbf{Q}_{ij}^{\beta-1} + \frac{1}{\alpha(\alpha+\beta)} \cdot (\alpha+\beta) \cdot \mathbf{Q}_{ij}^{\alpha+\beta-1} \right] \quad (6)$$

$$= \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i} \cdot \left[-\frac{1}{\alpha} \mathbf{P}_{ij}^\alpha \cdot \mathbf{Q}_{ij}^{\beta-1} + \frac{1}{\alpha} \cdot \mathbf{Q}_{ij}^{\alpha+\beta-1} \right] \quad (7)$$

$$= -\frac{1}{\alpha} \cdot \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i} \cdot \left[\mathbf{P}_{ij}^\alpha \mathbf{Q}_{ij}^{\beta-1} - \mathbf{Q}_{ij}^{\alpha+\beta-1} \right] \quad (8)$$

In the remainder of the cases, we set β to the appropriate value to demonstrate that the gradients match.

Case 2: $\alpha \neq 0, \beta = 0$. We have that:

$$\frac{\partial D_{AB}^{(\alpha, \beta)}(\mathbf{P} \parallel \mathbf{Q})}{\partial \mathbf{y}_i} = \frac{\partial}{\partial \mathbf{y}_i} \left[\frac{1}{\alpha^2} \left(\mathbf{P}_{ij}^\alpha \ln \frac{\mathbf{P}_{ij}^\alpha}{\mathbf{Q}_{ij}^\alpha} - \mathbf{P}_{ij}^\alpha + \mathbf{Q}_{ij}^\alpha \right) \right] \quad (9)$$

$$= \frac{\partial}{\partial \mathbf{y}_i} \left[-\frac{1}{\alpha^2} \mathbf{P}_{ij}^\alpha \ln \mathbf{Q}_{ij}^\alpha + \frac{1}{\alpha^2} \mathbf{Q}_{ij}^\alpha \right] \quad (10)$$

$$= \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i} \left[-\frac{1}{\alpha^2} \mathbf{P}_{ij}^\alpha \cdot \alpha \mathbf{Q}_{ij}^{\alpha-1} \cdot \frac{1}{\mathbf{Q}_{ij}^\alpha} + \frac{1}{\alpha} \mathbf{Q}_{ij}^{\alpha-1} \right] \quad (11)$$

$$= \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i} \left[-\frac{1}{\alpha} \frac{\mathbf{P}_{ij}^\alpha}{\mathbf{Q}_{ij}^\alpha} + \frac{1}{\alpha} \mathbf{Q}_{ij}^{\alpha-1} \right] \quad (12)$$

$$= -\frac{1}{\alpha} \cdot \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i} \cdot \left[\frac{\mathbf{P}_{ij}^\alpha}{\mathbf{Q}_{ij}^\alpha} - \mathbf{Q}_{ij}^{\alpha-1} \right] \quad (13)$$

Setting $\beta = 0$ in Equation (8), we observe that the gradients match, as desired.

Case 3: $\alpha = -\beta \neq 0$. We have that:

$$\frac{\partial D_{AB}^{(\alpha, \beta)}(\mathbf{P} \parallel \mathbf{Q})}{\partial \mathbf{y}_i} = \frac{\partial}{\partial \mathbf{y}_i} \left[\frac{1}{\alpha^2} \left(\ln \frac{\mathbf{Q}_{ij}^\alpha}{\mathbf{P}_{ij}^\alpha} + \frac{\mathbf{P}_{ij}^\alpha}{\mathbf{Q}_{ij}^\alpha} - 1 \right) \right] \quad (14)$$

$$= \frac{\partial}{\partial \mathbf{y}_i} \left[\frac{1}{\alpha^2} \left(\ln \mathbf{Q}_{ij}^\alpha + \frac{\mathbf{P}_{ij}^\alpha}{\mathbf{Q}_{ij}^\alpha} \right) \right] \quad (15)$$

$$= \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i} \cdot \left[\frac{1}{\alpha^2} \cdot \alpha \mathbf{Q}_{ij}^{\alpha-1} \cdot \frac{1}{\mathbf{Q}_{ij}^\alpha} + \frac{1}{\alpha^2} \cdot \mathbf{P}_{ij}^\alpha \cdot -\alpha \cdot \frac{1}{\mathbf{Q}_{ij}^{\alpha+1}} \right] \quad (16)$$

$$= -\frac{1}{\alpha} \cdot \frac{\partial \mathbf{Q}_{ij}}{\partial \mathbf{y}_i} \cdot \left[\frac{\mathbf{P}_{ij}^\alpha}{\mathbf{Q}_{ij}^{\alpha+1}} - \frac{1}{\mathbf{Q}_{ij}^\alpha} \right] \quad (17)$$

Setting $\beta = -\alpha$ in Equation (8), we again observe that the gradients match, as desired. It follows that for all cases in the paper that we explore, we can simply use the gradient in Equation (8).

2. Embedding Plots of Various Datasets

In Figures 1, 2 and 3 we showcase visualizations of datasets on which we presented some results in the main paper, but did not include plots due to lack of space.

These plots aim to show the impact of α and β on qualitative properties of the embedding. As discussed in the paper: *changing λ should primarily affect global over local structure, where (i) $\lambda < 1$ should lead to greater cluster separation*

220	<i>while (ii) $\lambda > 1$ should lead to low separation. Further, ABSNE should tend to produce lots of small, fine-grained clusters</i>	275
221	<i>for $\alpha < 1$ with few global changes in visualization while $\alpha > 1$ should lead to fewer, larger clusters with more global</i>	276
222	<i>visualization changes.</i>	277
223		278
224	We directly use pixels for the ORL and COIL-20 vision datasets, while we compute fc7 features yielded by Caffe's ImageNet model(Jia et al., 2014) for Caltech256. We use raw features for the other datasets. For all datasets with over 100	279
225	dimensions, we first apply 100 dimensional PCA before computing neighborhood scores.	280
226		281
227		282
228	References	283
229		284
230	Cichocki, A., Cruces, S., and Amari, S. Generalized alpha-beta divergences and their application to robust nonnegative	285
231	matrix factorization. <i>Entropy</i> , 13:134–170, 2011.	286
232	Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolu-	287
233	tional architecture for fast feature embedding, 2014.	288
234		289
235		290
236		291
237		292
238		293
239		294
240		295
241		296
242		297
243		298
244		299
245		300
246		301
247		302
248		303
249		304
250		305
251		306
252		307
253		308
254		309
255		310
256		311
257		312
258		313
259		314
260		315
261		316
262		317
263		318
264		319
265		320
266		321
267		322
268		323
269		324
270		325
271		326
272		327
273		328
274		329

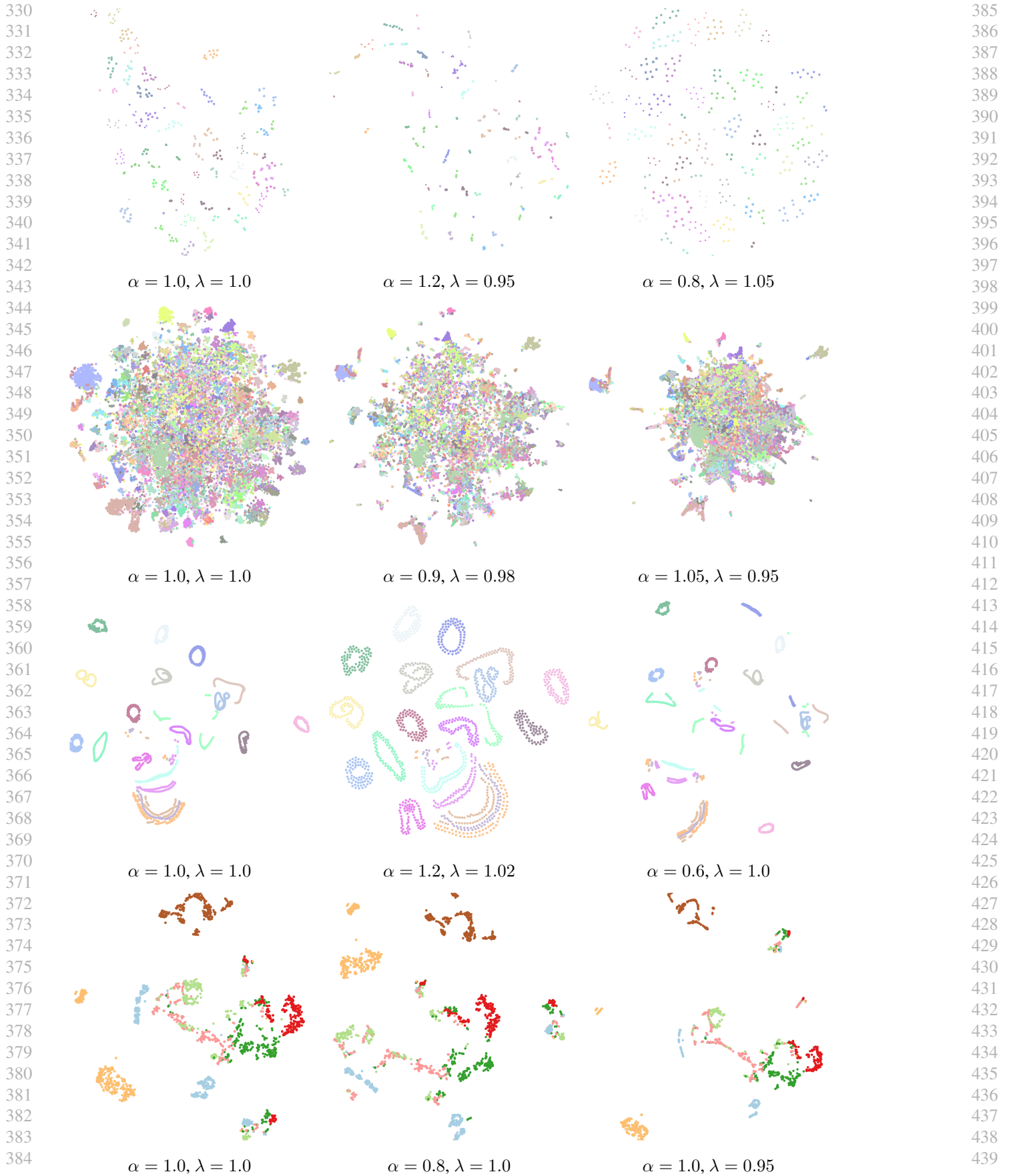


Figure 1. ABSNE visualizations for (rows from the top) ATT-Faces, Caltech256, COIL20, Segmentation Datasets. The left column corresponds to t-SNE. The center and right column contain visualizations with α and λ set manually to emphasize local or global clustering. These two parameters can be used by a data scientist for goal-driven exploratory data visualization.

440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494

495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549

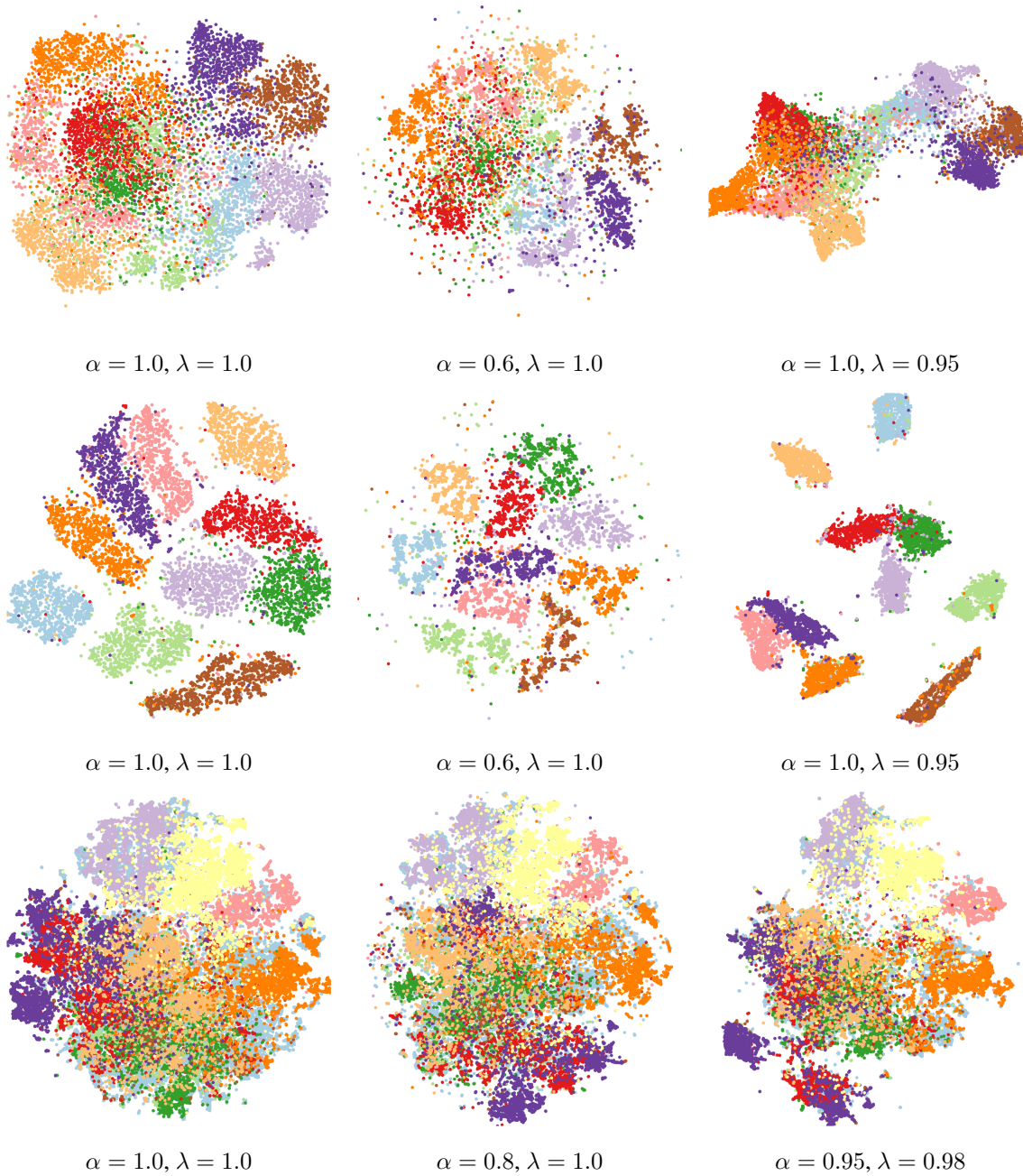


Figure 2. ABSNE visualizations for (rows from the top) CIFAR10, MNIST, ILSVRC2012 (validation) Datasets. The left column corresponds to t-SNE. The center and right column contain visualizations with α and λ set manually to emphasize local or global clustering. These two parameters can be used by a data scientist for goal-driven exploratory data visualization.

550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604

605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659



Figure 3. BSNE visualizations for (rows from the top) WDBC, WINE, IRIS Datasets. The left column corresponds to t-SNE. The center and right column contain visualizations with α and λ set manually to emphasize local or global clustering. WINE and IRIS are very small datasets with clear clusters, and ABSNE plots do not differ widely from t-SNE