Spectral Chinese Restaurant Processes: Nonparametric Clustering Based on Similarities

Richard Socher **Andrew Maas** **Christopher D. Manning**
Computer Science Department, Stanford University, Stanford, CA 94305

Abstract

We introduce a new nonparametric clustering model which combines the recently proposed distance-dependent Chinese restaurant process (dd-CRP) and non-linear, spectral methods for dimensionality reduction. Our model retains the ability of nonparametric methods to learn the number of clusters from data. At the same time it addresses two key limitations of nonparametric Bayesian methods: modeling data that are not exchangeable and have many correlated features. Spectral methods use the similarity between documents to map them into a low-dimensional spectral space where we then compare several clustering methods. Our experiments on handwritten digits and text documents show that nonparametric methods such as the CRP or dd-CRP can perform as well as or better than k-means and also recover the true number of clusters. We improve the performance of the dd-CRP in spectral space by incorporating the original similarity matrix in its prior. This simple modification results in better performance than all other methods we compared to. We offer a new formulation and first experimental evaluation of a general Gibbs sampler for mixture modeling with distance-dependent CRPs.

1 Introduction

Spectral clustering methods have the benefit of allowing us to include arbitrary features for representing data. They assume that the data lie on a low-dimensional manifold but are represented in a high-dimensional feature space. In order to recover the underlying cluster structure they perform the following steps. First, the features of each observation are used to compute a pairwise similarity matrix. This matrix is then used to map the observations from this implicit representation into a lower dimensional Euclidean space. In this space most methods apply k-means, fixing the number of clusters by hand [11]. In this paper, we argue that spectral methods can benefit from running nonparametric clustering methods in the reduced dimensional space.

Nonparametric Bayesian clustering methods such as the infinite Gaussian mixture model [17] or the hierarchical Dirichlet process [22] are appealing because they can infer the number of clusters from data. However, they have two fundamental restrictions. The first stems from the requirement to generate each observation from a well-defined distribution. For instance, in latent Dirichlet allocation [3], each word of a text document is sampled from a multinomial distribution of a corresponding topic. If we want to incorporate features such as the author of a document [18], then the model has to be changed and the inference procedure modified. The second restriction stems from the assumption that observations are exchangeable. Exchangeability refers to the invariance of a sequence of random variables to permutations of their indices. While exchangeability is often considered an advantageous property, much data in text, image and audio domains are not exchangeable. For example, image regions depicting the sky are not exchangeable since their position is important [19]. Topics change over time and future documents cannot influence past documents [23]. Lifting these restrictions is hard and often leads to models that are either specific to a certain modality or very complex.

We introduce a method to cluster non-exchangeable data that combines the advantages of nonparametric and spectral methods. Fig. 1 gives a high-level overview of our method. The input to our method can be a corpus of any modality such as text documents, handwritten digits or images. Similar to

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

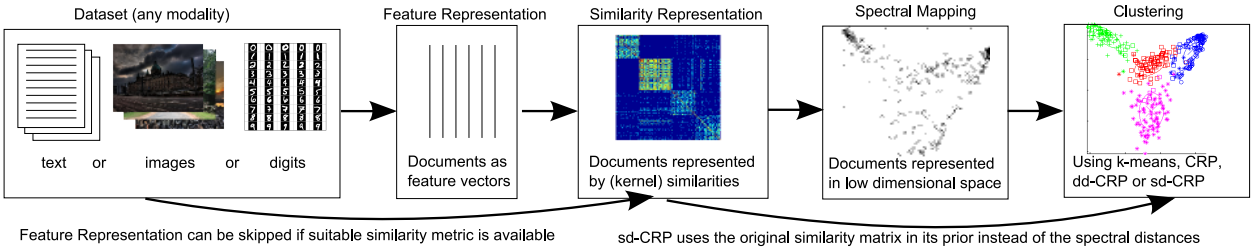

Figure 1: Overview of spectral-nonparametric clustering with the sd-CRP and other methods.

other spectral clustering methods, we first compute the similarity matrix from the features of the documents and then use this matrix to map documents into a low-dimensional Euclidean space. Instead of using k-means to cluster points based on their distances in this spectral space, we use nonparametric methods. This is also distinct from standard nonparametric clustering as we cluster in spectral space and not in the original data representation.

We first experiment with a Dirichlet Process mixture model [17]. To investigate other nonparametric clustering methods in spectral space, we implement the distance-dependent Chinese restaurant process (dd-CRP), a flexible class of distributions over partitions which was recently introduced by [2]. Previously, the dd-CRP had only been applied to temporal (sequential) data. We provide the combinatorial details of a *Gibbs sampling algorithm for mixture modeling with the dd-CRP in the case of arbitrary covariates*. In this setting, which we call the spatial or non-sequential case, cycles of seating assignments are possible. Lastly, we introduce the *similarity dependent CRP* (sd-CRP), which improves upon the dd-CRP by incorporating the original similarity matrix in the prior while still relying on the distances in spectral space for the likelihood. Our experiments show that the sd-CRP outperforms other methods on most criteria when clustering text documents and handwritten digits.

2 Distance-Dependent Chinese Restaurant Processes

In this section, we briefly cover the basics of both the CRP and dd-CRP to underline their differences and to help understanding the proposed Gibbs sampler of Sec.3.3 which learns dd-CRP based infinite mixture models. For more information on the CRP and its connection to the Dirichlet process (DP), see [16, 6].

The Chinese restaurant process defines the following procedure. Imagine a restaurant with an infinite number of tables. The first customer sits down at a table. The ith customer sits down at a table with a probability that is proportional to the number of people already sitting at that table or she opens up a new table with a probability proportional to the hyperparameter α. Because of exchangeability, the order in which customers sit down is irrelevant and we can draw each customer's table assignment z_i by pretending they are the last person to sit down. Let K be the number of tables and let n_k be the number of people sitting at each table. For the ith customer, we define a multinomial distribution over *table assignments* conditioned on $\mathbf{z}_{-i}$, i.e. all other table assignments except the ith:

$$p(z_i = k|\mathbf{z}_{-i}, \alpha) \propto \begin{cases} n_k & \text{if } k \leq K \\ \alpha & \text{if } k = K+1. \end{cases} \quad (1)$$

Given the cluster assignment each data point is conditionally independent of the other ones. The exchangeability assumption in this process holds for some datasets but not in others. While several special models for spatial and temporal dependencies have been proposed, the distance-dependent CRP offers an elegant general method to modeling additional features and non-exchangeability.

The main difference between the dd-CRP and the standard CRP is that in the dd-CRP customers sit down with other customers instead of directly at tables. Connected groups of customers sit together at a table only implicitly. Using a similar culinary metaphor, imagine a restaurant full of people. The ith customer sits with some other customer j (denoted as $c_i = j$) with a probability proportional to a decreasing function of the distance between the two: $f(d_{ij})$, or by herself with a probability proportional to α. Hence, the larger your distance, the less likely you are to sit with somebody. This leads to the following multinomial over *customer assignments* conditioned on distances $D \in \mathbb{R}^{N\times N}$, where N is the number of customers and the decay function $f : \mathbb{R}^+ \to \mathbb{R}^+$ needs to be non-increasing and have $f(\infty) = 0$,

$$p(c_i = j|D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if } i \neq j \\ \alpha & \text{if } i = j. \end{cases} \quad (2)$$

The distance function is usually parameterized, e.g. for exponential decay, we have the parameter a: $f_a(d) = \exp(-d/a)$. Notice that this seating prior is

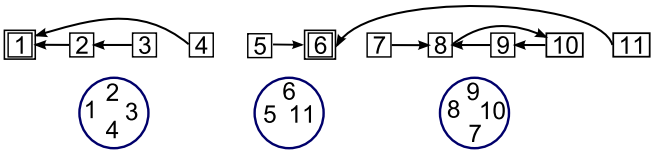

Figure 2: Illustration of the distance-dependent Chinese restaurant process. Customers sit behind other customers. Each group of linked customers implicitly sits at a table. Only the distance to a customer determines seating, not the order. Hence, 5 may choose to sit with 6 (and therefore at a table with 11). Cycles implicitly form tables. A customer who decides to sit with somebody in a cycle joins that table, like 7 who joined the cycle of 8,9,10.

not conditioned on $\mathbf{c}_{-i}$, the seating of other customers. Note also that customers may sit in cycles. Each connected component (of which cycles are a special case) forms its own table which may be joined by other customers who sit with a member of that component. Figure 2 illustrates a possible seating assignment.

Dd-CRPs are not *marginally invariant*, meaning that if we integrate out one observation we would not get the same probability distribution as if that observation was not part of the model. Furthermore, dd-CRPs are not exchangeable and therefore have the freedom to model spatial and temporal correlations. For instance, news articles are much more likely to cluster together if they appear nearby in time.

Distance dependent CRPs are also related to the recent effort to introduce more complex features into directed graphical models [13, 5, 21]. The dd-CRP can be embedded into graphical models and could provide an alternative to Dirichlet Process mixture models. Several other methods can use distances or feature similarities for clustering, e.g. kernel k-means [4].

3 Nonparametric Clustering in Spectral Space

The general spectral clustering procedure is as follows:
Inputs: Similarity matrix $S \in \mathbb{R}^{N \times N}$ and hyperparameters.
Outputs: Number of clusters K; their parameters and cluster assignments.

1. Map data into M-dimensional spectral space using similarity matrix S (Alg. 1).
2. Cluster data points in the M-dimensional space.

We now briefly describe step 1, which is the same for all the clustering methods. We then provide an overview of clustering alternatives including the dd-CRP, its modification and related mixture model.

Algorithm 1 Spectral Dimensionality Reduction

Input: Similarity matrix S
Output: Points $(x_1, \ldots, x_N)$
Compute diagonal degree matrix:
$D_{ii} = \sum_{j=1}^{N} S_{ij}$
Compute unnormalized graph Laplacian:
$L = D - S$
Compute normalized Laplacian:
$L_{sym} = D^{-1/2} L D^{-1/2}$
Compute its first M eigenvectors: $u_1, \ldots, u_M$
Define: $U \in \mathbb{R}^{N \times M}$
Normalize rows of U to norm 1
Define rows of U as new observations:
$x_i = U(i, :)$

3.1 Spectral Dimensionality Reduction

Spectral dimensionality reduction methods try to preserve local distances when mapping data to a low dimensional space. They are often based on the eigendecomposition of the transformed similarity matrix S which can be interpreted as a weighted adjacency matrix of a graph. They reduce the dimensions of the observations and their features to an M-dimensional space. M is manually set to a low number of about 2 to 20, often close to the number of assumed clusters, we provide sensitivity analysis to this choice in the experiment section. An advantage of manifold learning methods is that they can be used even if no vectorial representation is available. They can purely rely on a similarity function between observations. The algorithm we use follows [15] except that we do not use k-means clustering at the end. It is given for completeness in Alg. 1.

To the best of our knowledge this is the first work which combines sophisticated non-linear dimensionality reduction methods with nonparametric clustering. Wood et al. [24] used PCA and therefore rely on the data lying on a linear manifold in a higher dimensional space. However, if the data lies on a lower dimensional manifold with a highly nonlinear structure, linear methods such as PCA fail to to correctly map the data into a lower dimensional space. For an in-depth tutorial on non-linear dimensionality reduction methods see [11].

3.2 Learning the Number and Parameters of Clusters in Spectral Space

We compare several alternatives to k-means that do not fix the number of clusters.

Model based clustering learns several Gaussian mixture models with varying geometric properties and numbers of clusters [7]. The models are selected using the Bayesian Information Criterion.

The infinite Gaussian mixture model (IGMM) assumes a Gaussian distribution for each cluster and -instead of manually setting the number of clusters - automatically determines the number of clusters and their parameters from the data. IGMMs may be based on Dirichlet Process mixture models [17].

The distance-dependent Chinese restaurant process (dd-CRP) provides another nonparametric alternative to the IGMM for learning infinite mixture models [2]. In Sec. 2 we described the dd-CRP prior over partitions using the analogy of seating assignments of customers with other customers. Similar to the Dirichlet process mixture model, we can define a mixture model with a base distribution G_0 and use the dd-CRP as a prior over cluster assignments. The decay function f is applied to the pairwise distances in matrix D. In our case, D represents the distances of points in the reduced-dimensional spectral space. Given the scaling parameter α, we get the following generative process for observation $x_i \in \mathbb{R}^M$:

1. For each observation $i \in [1, N]$ draw seating assignment $c_i \sim \text{dd-CRP}(\alpha, f, D)$.
2. For each cluster $k \in [1, K]$, induced by separate seating groups, draw parameters $\theta_k \sim G_0$
3. For each observation $i \in [1, N]$, draw $x_i \sim F(\theta_{k(i)})$,

where the function $k(i)$ returns the cluster number of the ith customer. $F(\theta_{k(i)})$ may for instance be a Gaussian distribution and $\theta = (\mu, \Sigma)$. Similar to IGMMs based on the CRP, the dd-CRP clusters the points based on their spatial layout in the reduced M-dimensional spectral space.

Since the spectral dimensionality reduction method is purely unsupervised it may throw away low variance variations which might be important for finding the correct groups [1]. Furthermore, one usually chooses a low number of reduced dimensions and therefore approximates the true spatial layout.[1] Hence, allowing the original similarity matrix to influence the clustering in spectral space might improve performance.

The similarity-dependent CRP (sd-CRP) is a modified version of the dd-CRP for clustering in spectral space. Unlike the authors in [2] who assume that the distances are 'innate properties of the customers', we reformulate the decayed distances between customers as similarities and as such, a modeling choice. This allows us to define arbitrary similarities instead of having to focus only on the spatial layout or time stamps of observations.

We modify the dd-CRP by computing the prior with the original similarity matrix S which was used to compute the spectral mapping instead of the decayed distances in spectral space. The combination of spectral dimensionality reduction and dd-CRP could be seen as a simple pipeline approach with a sophisticated pre-processing step. In contrast, the sd-CRP goes beyond such a pipeline by re-using the similarity information and essentially introducing a dual-view of observations. Each observation's prior holds the detailed, local similarities to all other observation while its likelihood is computed from the layout of the points in the spectral, reduced dimensional space. This space takes into consideration the global layout of points on a possibly non-linear manifold.

While this constitutes a redundant use of these similarities, it helps to exploit high variance variations in the prior and still use the spatial layout of points in the spectral space for the likelihood computation. It also removes another parameter for the distance function. The difference to the above procedure is: $c_i \sim \text{dd-CRP}(\alpha, I, S)$, where I is simply the identity function (and can therefore be ignored).

As we will see in the experiments section, this simple change results in an improved performance on several metrics. It could be argued that it is not surprising that more information improves clustering performance. However, such re-use has not been attempted before in the spectral clustering literatureWe note that the generative nature of the model is - in its current formulation - not preserved.

3.3 Posterior Inference via Gibbs Sampling

The goal of inference in dd-CRP and sd-CRP based models is to compute the posterior distribution over partitions given the observations in the lower dimensional space. As in most other mixture models this is intractable and we resort to sampling techniques. The sampling part of the outlined procedure is the same for both models, only the input to the prior changes.

Blei and Frazier [2] provide a general Gibbs sampler for the dd-CRP together with the algorithmic details for language modeling and the sequential case where customers can only sit with past customers. This results in a seating assignment in the form of a DAG. We introduce the details of a Gibbs sampler for the general case of mixture modeling where cycles are possible (see Fig. 2 for an illustration). While in principle it is the same sampler, one has to pay special attention to cycles which can start new tables even in cases where customers do not sit by themselves.

[1] We will explore the dependence of our method on the number of dimensions in the reduced spectral space in the experiments section.

Let us first introduce some notation. The function sitBehind(i), returns the set of all customers that sit behind i, including i itself.[2] This function is recursive. As examples based on the seating of Fig. 2: sitBehind(1) = $\{1,2,3,4\}$, sitBehind(2) = $\{2,3\}$, sitBehind(3) = $\{3\}$. Note that in a cycle, everybody sits behind each other: sitBehind(10) = sitBehind(8) = $\{7,8,9,10\}$.

During sampling, we want to compute the probability of each customer i to sit with any customer j: $p(c_i = j|\mathbf{c}_{-i}, \mathbf{x}_{1:N}, S, \alpha, \Lambda_0, \Theta_{1:K})$, where $\Theta_{1:K} = (\mu_k, \Sigma_k)_{k=1:K}$ are the table/cluster parameters, $\mathbf{x}_{1:N}$ are the observations and $\mathbf{c}_{-i}$ are other customers' seating assignments. The two hyperparameters are α (see Eq. 2) and Λ_0, the prior on the covariance of the Gaussian components. In general, we have that

$$p(c_i = j|\mathbf{c}_{-i}, \mathbf{x}_{1:N}, \cdot) \propto p(c_i = j|\cdot)p(\mathbf{x}_{1:N}|c_i = j, \mathbf{c}_{-i}) \tag{3}$$

For the dd-CRP in spectral space, we have $p(c_i = j|\alpha, f, D)$ and for sd-CRP, we have $p(c_i = j|\alpha, S)$, both are defined by Eq. 2. It is left to show the conditional likelihoods that can arise from the seating choice c_i. There are two main cases: either customer i implicitly creates a new table with her seating choice, or she connects to an existing table.

New tables can be created in two ways: Either the customer sits by herself or she sits with somebody behind her (and has not previously done so). The latter case creates a cycle. This is captured by the boolean predicate newTable:

$$\begin{aligned}&\text{newTable}(c_i) \text{ is true iff} \\ &\left(c_i \in \text{sitBehind}(i) \wedge c_i^{(old)} \notin \text{sitBehind}(i)\right).\end{aligned} \tag{4}$$

The likelihood computation has some resemblance to Gibbs sampling for Dirichlet process mixture models [14]. The difference is that we compute the likelihood for all the customers that sit behind i, denoted $X_i = \mathbf{x}_{\text{sitBehind}(i)}$. This can be seen as a sort of blocked sample. Note that we can ignore all other customers as their likelihood is not affected. In the case of a new table, we integrate over the normal-inverse Wishart base distribution $G_0 = \mathcal{NW}^{-1}$:

$$\begin{aligned}&\text{If newTable}(c_i), p(\mathbf{x}_{1:N}|c_i = j, \mathbf{c}_{-i}) \propto \\ &\int \mathcal{N}(X_i|\mu, \Sigma)\mathcal{NW}^{-1}(\mu, \Sigma|\nu_0, \mu_0, \Lambda_0)d(\mu, \Sigma).\end{aligned} \tag{5}$$

Since, we use a conjugate prior, the above integral has a closed form solution in a form of a multivariate Student-t distribution. We approximate this distribution by a moment-matched Gaussian [20]. We sample a new cluster covariance matrix from the inverse-Wishart prior (with hyperparameters Λ_0 which is fixed and $\nu_0 = M$) and a mean which depends on $\mu_0 = 0$ and this covariance matrix as described in [20],

$$\Sigma_{K+1} \sim \mathcal{W}^{-1}(\nu_0, \Lambda_0),\ \mu_{K+1} \sim \mathcal{N}(\mu_0, \Sigma_{K+1}) \tag{6}$$

and then computing the likelihood for all $l \in$ sitBehind(i) given this Gaussian distribution. Alternatively, one could just work directly with the t-distribution using a scale mixture.

In the case of i sitting with customer j and at its table $k(j)$, we compute the likelihood of all the customers behind i, given that table's parameters.

$$\text{If } \neg\text{newTable}(c_i), p(\mathbf{x}_{1:N}|c_i = j, \mathbf{c}_{-i}) \propto \mathcal{N}(X_i|\mu_{k(j)}, \Sigma_{k(j)}) \tag{7}$$

As we noted above, the dd-CRP needs to take into account the current and all connected customers since it is not marginally invariant. While this results in a higher computational cost for computing each step, it allows for larger steps through the state space[3] and therefore faster convergence than Gibbs sampling in a CRP mixture. Note also that unless seating assignments actually change, we can use cached likelihood computations of previous iterations.

After each iteration of sampling the seating assignments, we need to update the table parameters given the new table members. Since our $\mathcal{NW}$-prior is conjugate, we can sample the new table parameters from the posterior density in the same family [8]. Let $x_1, \ldots, x_n$ be the customers at a table, $\bar{x}$ the sample mean and we define $Q = \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T$; then the posterior is $\mathcal{NW}(\mu_n, \kappa_n, \nu_n, \Lambda_n)$ with the following parameters ($\kappa_n = \kappa_0 + n, \nu_n = \nu_0 + n$):

$$\begin{aligned}\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{x}, \\ \Lambda_n &= \Lambda_0 + Q + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T.\end{aligned} \tag{8}$$

Algorithm 2 summarizes the complete inference procedure for dd-CRPs and sd-CRPs in spectral space.

Typically, in such conjugate models, one could simply integrate out the parameters and only sample the assignments of observations to clusters. In our experiments this worked very well on synthetic data that

[2] For notational convenience that will become apparent soon, each customer sits *behind* herself. Intuitively, this is the set of customers that point to i, including i but excluding c_i, (the customer that i sits with). If i and c_i are in a cycle they are both in each other's sitBehind set.

[3] Larger steps through the state space are a result of the seating assignments. When customer i moves to a different table, all customers who recursively sit behind her also move to that table. As a special case, if one customer of a cycle moves, all of them move.

Algorithm 2 Inference for dd-&sd-CRP

Input: Similarity matrix S, hyperparameters (α, Λ_0)
Output: Clustering and cluster parameters.
SpectralMapping$(S) = (x_1, \ldots, x_n)$ (Alg.1)
Initialize random clustering of points
for iterations **do**
  Sample random permutation τ of $1, \ldots, N$
  for $i \in (\tau(1), \ldots, \tau(N))$ **do**
    Remove c_i, the outgoing link of customer i
    Sample $c_i \sim p(\cdot|\mathbf{c}_{-i}, \cdot)$ (Eqs. 3)
    if newTable(c_i) (Eq. 4) **then**
      K = K+1
      Sample new parameters (μ_K, Σ_K) (Eq. 6)
      Add them to $\Theta_{1:K}$
    end if
    Move sitBehind(i) to c_i's table.
  end for
  Re-sample table parameters (Eq. 8)
  Sample α using a Metropolis step
end for

Method	mutI	randI	VoI	K
Oracle	1.38	1	0	(4)
$\mathcal{N}$-Oracle	0.98	0.90	1.13	(4)
k-means	0.93	0.88	1.29	(4)
MBC	0.96	0.86	1.50	5
CRP	0.72	0.82	1.48	**4.2**
dd-CRP	**0.98**	0.86	1.65	6.0
sd-CRP	**0.98**	**0.89**	**1.22**	4.4

Table 1: **Digits 1-4.** Comparison of k-means, model based clustering, CRP, dd-CRP and sd-CRP on a 4 digits subset of MNIST. The $\mathcal{N}$-Oracle uses the ground truth to compute the Gaussian mixture and then clusters with these Gaussians, providing an upper bound of what methods based on Gaussians can achieve. Metrics are mutual information (mutI, higher is better), rand Index (randI, higher is better) and variation of information (VoI, lower is better). K is the average number of found clusters.

we sampled from true Gaussian mixture distributions. However, in experiments on real data we achieved higher performance when explicitly sampling the cluster parameters.

4 Experiments

The goal of our experiments is to demonstrate the main advantages of the sd-CRP. It can (i) use the same model for different tasks and modalities, (ii) it can exploit powerful features that represent documents; and (iii) it more robustly recovers the clusters than most other methods we try on most metrics. We compare the sd-CRP to the other clustering algorithms mentioned in the previous section on handwritten digits [10] and newsgroup articles.[4] *All methods we compare cluster the data in spectral space.*

The 20 newsgroups dataset consists of 20 classes, divided into six main topics. The approximate main topics and number of subgroups are *computers*(5), *sports*(4), *science*(4), *sale*(1), *politics*(3) and *religions*(3). The main topics can be quite diverse, for instance *auto* and *hockey* are both in the sports category. Hence, we only evaluate on the 20 subclasses. Matlab code for the complete inference procedure can be downloaded at `http://uponAcceptance`.

We use the *same hyperparameter* Λ_0 for all experiments on digits 1-4, digits 0-9 and main topics of newsgroups and expect to recover the main clusters. This is possible because the spectral dimensionality reduction method maps all data into a similar range, allowing us to set the covariance prior to $\Lambda_0 = 0.005 \cdot \text{diag}(1)$ for all CRPs. We set an initial $\alpha = 10^{-6}$ for *all* experiments. The dd-CRP uses the exponential decay function with $a = 0.01$: $f(d) = \exp(-d/a)$. Unless otherwise stated the number of dimensions in the embedding space is set to the number of clusters. We explore the importance of the dimensionality of the embedding space in the last experiment.

In all our experiments, k-means has an unfair advantage since it is given the true number of clusters. Model based clustering and the CRP-based methods recover the true number of clusters and still outperform the ubiquitous k-means algorithm in several settings.

4.1 Digits 1-4 and Clustering Visualization

We first show results on a subset of the MNIST digits dataset. We cluster 400 digits of the classes 1 to 4. This setup allows us to use only two eigenvectors in the lower dimensional space and to visualize some of the differences between clustering methods.

Since one of our goals is to show that we can exploit powerful features we use deep features [9] which are among the best performing features for handwritten digit classification and clustering. In all following digit experiments we first compute a linear kernel of deep learned features. The features are the top layer activations of a deep belief network trained in a completely unsupervised fashion as described in [9] and simply using their accompanied code. Using the similarity matrix we map the digits to a lower dimensional space and all cluster methods are compared in that space.

In table 1, we compare the sd-CRP to k-means, model based clustering [7], the original CRPand the dd-CRP (which uses the Euclidean distance in spectral space). As clustering metrics we use mutual information, rand index and variation of information [12]. The first two are higher if the clustering is better whereas the last

[4] `http://kdd.ics.uci.edu/databases/20newsgroups/`

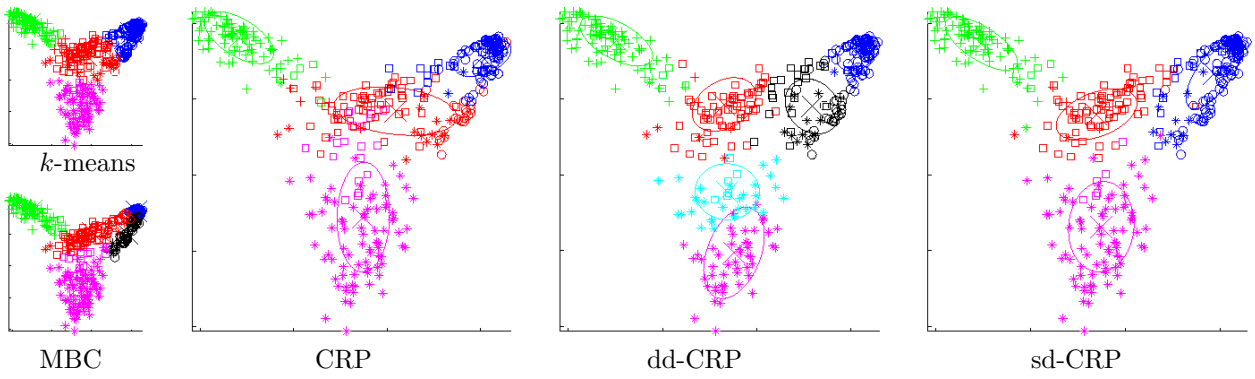

Figure 3: Clustering of 100 randomly sampled digits from 1 to 4 in spectral space. Model based clustering is abbreviated as MBC. Markers are based on the ground truth, colors are based on the method's clustering. The sd-CRP uses the original similarity matrix when computing the seating prior. This allows points far away from the mean to still join a cluster and influence its parameters. Hence, the sd-CRP correctly recovers the elongated structure of the blue cluster.

one is better if it's close to zero. Table 1 shows the results. All CRPs are averaged over 5 runs and model based clustering was run for $K = 1, \ldots, 20$ clusters.

Fig. 3 shows the clustering result of these five methods. Table 1 demonstrates that the dd-CRP with a prior based on Euclidean distances in spectral space does not perform as well as the sd-CRP which uses the original similarity matrix S for the seating prior. This is apparent in the blue cluster of Fig. 3 whose elongated structure of nearby points is only captured by the sd-CRP.

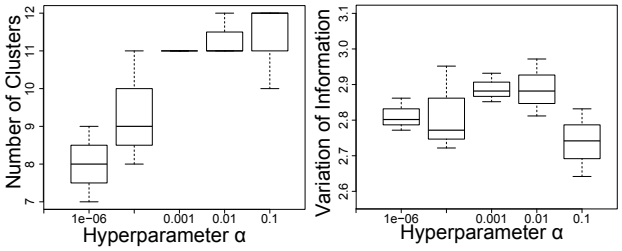

Figure 4: **Sensitivity Analysis of sd-CRP** on the number of clusters and variation of information (lower is better) given different initial values of the hyperparameter α. Even if α changes several orders of magnitude, the number of clusters stays around the ground truth of 10.

4.2 Digits 0-9 and Hyperparameter Sensitivity Analysis

In this experiments we cluster 1000 digits from all classes $0, 1, \ldots, 9$. Table 2 (left) lists the results. CRP and sd-CRP show a similar performance on the variation of information criterion, while the sd-CRP is on par with model based clustering on mutual information and rand index. Overall, the sd-CRP performs best across all three metrics and is closest to the true number of clusters.

The importance of the initial prior over seating assignments α is reduced since we sample it after each iteration using a Metropolis Hastings step (in both the CRP and sd-CRP). A sensitivity analysis shows that the number of clusters can be learned robustly even when the initial α varies several orders of magnitude. Fig. 4 shows a sensitivity analysis of the variation of information metric and the number of clusters given different values of the hyperparameter α. For values of $\alpha = 10^{-6}, \ldots, 0.1$, the number of clusters robustly varies between 8 to 12, i.e., around the true number 10.

4.3 Newsgroups

We sample 1000 newsgroup articles uniformly from the full set of 18,846. The similarity matrix is again computed with a linear kernel of deep features of the newsgroup articles similar to the procedure of [9]. When using the same hyperparameters as in the digit experiments, sd-CRP finds 8 clusters (compared to 11 in MBC and 10.6 for CRP). Since the newsgroup dataset is hierarchical (with 6 main classes and 20 subclasses), no clustering algorithm can automatically determine the preferred granularity of the clusters. We show that by using a slightly smaller value for the covariance prior, $\Lambda_0 = 0.001 \cdot \text{diag}(1)$, the sd-CRP can recover the 20 classes more accurately than other methods. Table 2 shows that sd-CRP achieves the best score on the variation of information metric and is most accurate in its estimate of the number of clusters. Only k-means, which was given the true number of clusters outperforms it by 0.01 on the other 2 metrics.

	Digits 0-9				Newsgroups			
Method	mutI	randI	VoI	K	mutI	randI	VoI	K
Oracle	2.30	1.00	0.00	(10)	2.98	1.00	0.00	(20)
$\mathcal{N}$-Oracle	1.58	0.95	2.05	(10)	1.41	0.91	4.29	(20)
k-means	1.23	0.88	2.98	(10)	**1.06**	**0.88**	5.12	(20)
MBC	**1.27**	**0.89**	3.22	13	0.93	0.87	4.96	11
CRP	1.11±0.09	0.85±0.02	2.73±0.09	8.0±1.15	1.01±0.07	**0.88**±0.01	4.89±0.07	15.0±1.9
dd-CRP	1.09±0.04	0.85±0.01	2.98±0.10	8.0±0.78	0.98±0.05	0.82±0.03	5.15±0.06	22.8±1.5
sd-CRP	**1.27**±0.01	**0.89**±0.00	**2.72**±0.08	**9.3**±0.96	1.05±0.04	0.87±0.01	**4.78**±0.05	**17.8**±1.1

Table 2: **Left: Digits 0-9.** Comparison on a 10 digits subset of MNIST, including standard deviation. sd-CRP in on par with the best methods on 2 metrics and outperforms others on VoI. It also most closely recovers the true number of clusters. **Right: Newsgroups.** Comparison on the newsgroup dataset with 20 classes as ground truth. The sd-CRP outperforms other methods on the variation of information criterion and is only 0.01 below k-means in the other two metrics. However, k-means has an *unfair advantage* since it was given the true number of clusters.

4.4 Influence of Dimensionality Reduction

The idea of combining spectral dimensionality reduction methods and nonparametric clustering has only been explored briefly by Wood et al. [24] who use PCA. They then discard completely the original data (like all previous spectral clustering methods) and cluster using only the first 2 eigenvectors. As we will see, this throws away a lot of valuable information and hurts clustering performance badly.

There are several factors playing into the choice of dimensionality for the reduced space. The lower it is, the more the spectral method reduces the noise and the faster the computations. However, choosing too few eigenvectors also results in a loss of information. Fig.5 shows results on the mutual information criterion for the 1-4 digit dataset under different dimensionalities of the embedding space. The bottom two horizontal lines, k-means(O) and the CRP(O), are results in the original feature space (the dd-CRP(O) performed worse than .85 and is not show). While the results vary, most methods perform best when using 5 or 6 eigenvectors and the sd-CRP achieves the highest performance among all methods with only 5 eigenvectors. Furthermore, the sd-CRP has the largest area under the curve and is robust to the number of dimensions. This experiment also shows that the dimensionality reduction improves clustering in most settings.

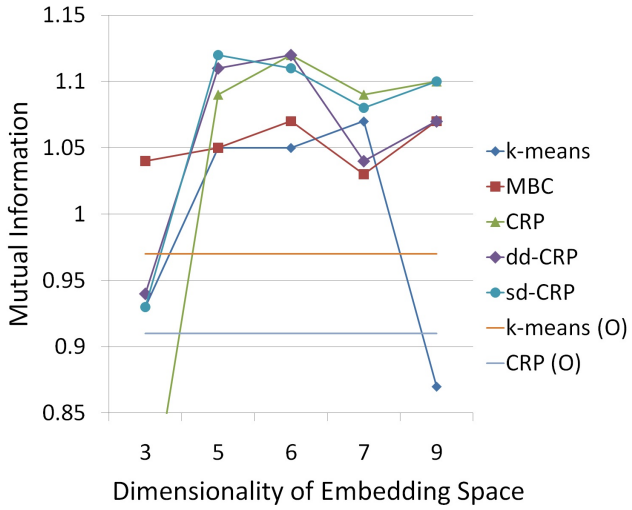

Figure 5: Dependence of the clustering performance (mutI) on the dimensionality of the embedding space of the 4 digits dataset. The sd-CRP achieves the highest performance with only 5 dimensions underlining its ability to exploit both the local information from its similarity prior and the globally consistent likelihood computed in spectral space.

5 Conclusion

We introduced a new nonparametric clustering technique based on the distance-dependent Chinese restaurant process and spectral dimensionality reduction. Our method combines the advantages of both of these methods and we hope it opens the door for further research into such combined models. We showed that nonparametric methods are a reasonable alternative to the widely used k-means clustering in spectral space. With the sd-CRP we introduce a simple but powerful modification to clustering with the dd-CRP in spectral space. The sd-CRP does not simply use the spectral dimensionality reduction as a pre-processing step. It incorporates the original similarity matrix in its prior instead of purely relying on the spatial layout in the reduced dimensional space. We showed that the sd-CRP is robust with respect to its hyperparameters and outperforms other clustering methods on handwritten digits and text documents. Our Gibbs sampler, provides the necessary combinatorial details needed when using the dd-CRP in a non-sequential setting. A possible direction of future research could be to jointly model and infer dimensionality reduction and clustering in one generative model.

Acknowledgments

We thank Neal Parikh, Adam Vogel and the anonymous reviewers for helpful comments on the draft.

References

[1] Y. Bengio, O. Delalleau, N. L. Roux, J.-F. Paiement, P. Vincent, and M. Ouimet. Spectral Dimensionality Reduction. Cirano working papers, CIRANO, 2004.

[2] D. M. Blei and P. I. Frazier. Distance dependent Chinese restaurant processes. In *ICML*, 2010.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. In *KDD 2004*, pages 551–556. ACM, 2004.

[5] L. Du, L. Ren, D. Dunson, and L. Carin. A Bayesian model for simultaneous image clustering, annotation and object segmentation. *NIPS 22*, 2009.

[6] T. S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[7] C. Fraley and A. E. Raftery. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 41:578–588, 1998.

[8] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, 2 edition, July 2003.

[9] G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, July 2006.

[10] Y. Lecun and C. Cortes. The MNIST database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`.

[11] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[12] M. Meila. Comparing Clusterings by the Variation of Information. In *Learning Theory and Kernel Machines*, pages 173–187, 2003.

[13] D. Mimno, H. Wallach, and A. McCallum. Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors. In *NIPS Workshop on Analyzing Graphs, 2008*, 2008.

[14] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[15] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS 14*. MIT Press, 2001.

[16] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, 2006.

[17] C. E. Rasmussen. The infinite Gaussian mixture model. In *NIPS 12*, volume 12, pages 554–560, 2000.

[18] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI 2004*, pages 487–494, 2004.

[19] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing Visual Scenes Using Transformed Objects and Parts. *IJCV*, 77(1):291–330, May 2008.

[20] E. B. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. Ph.D. thesis, MIT, Cambridge, MA, 2006.

[21] E. B. Sudderth and M. I. Jordan. Shared Segmentation of Natural Scenes Using Dependent Pitman-yor Processes. In *NIPS*, 2008.

[22] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[23] X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD 2006*, pages 424–433. ACM, 2006.

[24] F. Wood and M. Black. A nonparametric Bayesian alternative to spike sorting. *Journal of Neuroscience Methods*, 173, 2008.