Deep Clustering on a Hypersphere for High-Dimensional Healthcare Data

Abstract

We consider a task that arises often in exploratory analysis of healthcare data, as well as in other fields. The task is to produce a useful clustering of high-dimensional sparse count data, e.g., the number of times each drug, procedure, or billing code occurs, while taking into account important geometric constraints in the original space. In this work, we describe a general deep neural network framework to address this task. We propose a generic architecture that simultaneously (1) optimizes image recovery in an autoencoding framework, (2) creates a low-dimensional embedded representation of the high-dimensional space, (3) assigns examples to clusters in a soft clustering, and (4) optimizes the quality of this clustering. Crucially, image recovery is measured not just by ability to reconstruct data, but also by preservation of geometric relationships in this data. By integrating these components, our method is able to find a better clustering than under a simple two-step alternative where a representation is learned and then clustering is applied to the representation. By contrast, in our method, the representation can be informed not just by the data, but also by the clustering and constraints. The degree to which each component is prioritized can be adjusted by the user to enable enhanced exploratory analysis. Additionally, our framework can be used to efficiently compare clusterings with different numbers of clusters. We demonstrate our method's characteristics through through quantitative and qualitative analysis of real and simulated data, including in several real-world healthcare case studies.

Introduction

We consider a task that arises often in exploratory analysis and visualization of data from several fields. The task is to produce a useful clustering of high-dimensional sparse count data. For example, in the analysis of large-scale healthcare data arising from electronic health record databases, we might be interested in clustering patients or providers based on the number of time each billing code, procedure code, or drug prescription is observed. In this context, similarity between point is typically best measured through non-euclidean distances like cosine similarity.

Because straightforward clustering methods often perform poorly on complex high-dimensional data like that in our task, several authors have explored clustering in a lower-dimensional representation of the high-dimensional data.

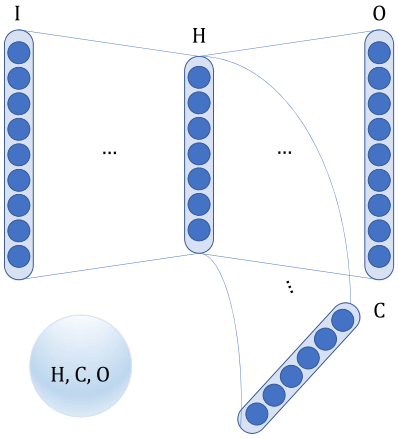


Figure 1: Autoencoder with clustering module. Nodes are Input (I), Output (O), Hidden (H), and Cluster (C). H and O are constrained by a user-specified batch distance measure; we use cosine similarity for sparse count data. H (h°) and C are constrained by batch cosine similarity that ties the angular representation to the cluster probability.

One way to approach to this problem is through a two-step process: (a) learn a low-dimensional representation, (b) apply clustering over the low-dimensional representation. For example, one popular pipeline is to use principal components analysis (PCA) to find a low-dimensional representation, followed by k-means to cluster points based on on this representation (Ding and He 2004).

Recently, deep learning learning methods have been used to find a low dimensional representation in the first step (Aljalbout et al. 2018). For example, Tian et al. (2014) use stacked autoencoders to learn a representation, followed by k-means. Similarly, Trigeorgis et al. (2014) define a deep non-negative matrix factorization and use the representation thus learned to cluster data with k-means.

A major drawback of this two-step process is that the representation learned in the first step is not necessarily well

suited to producing a clustering in the second step. For example, autoencoding is designed to minimize reconstruction loss, and a representation that minimizes reconstruction loss is not guaranteed – or even particularly likely – to map points to a meaningful clusterable space.

To address this issue, several methods have been defined that integrate both steps – reconstruction learning and clustering – in a single deep learning framework. This enables the representation to be informed by the clustering, not just by its ability to reconstruct data, and thus creates the possibility of finding both a more clusterable representation and a better clustering. For example, Xie et al. (2016) add a centroid-based soft clustering criterion to an autoencoder backbone and iteratively refine the autoencoder's parameters so as to improve the quality of the clustering. Similarly, Yang et al. (2017) define a method that integrates an autoencoding deep neural network and a k-means clustering loss, and fits this model with an alternating stochastic gradient descent algorithm.

While these approaches show excellent performance on certain types of data, particularly in imaging, they are not well-suited to sparse count data. This is primarily due to their dependence on autoencoding in high-dimensional euclidean space as the primary backbone of representation learning. Even with additional constraints to encourage the representation to be well-clusterable, the autoencoders disregard cosine similarity information and thus cannot take it into account in producing a clusterable representation. Given this limitation, use of centroid based clustering methods makes the problem worse, as identifying an appropriate centroid location is non-obvious.

Our contribution. In this paper, we propose a method that learns a low-dimensional embedding with geometric constraints that enables simultaneous representation learning and cluster determination. We propose a generic architecture that simultaneously (1) optimizes image recovery in an autoencoding framework, (2) creates a low-dimensional embedded representation of the high-dimensional space, (3) assigns examples to clusters in a soft clustering, and (4) optimizes the quality of this clustering. Crucially, image recovery is measured not just by ability to reconstruct data, but also by preservation of geometric relationships in this data. Our method also allows use of a side-constraint component, which encodes side constraints of interest, relating to general clustering characteristics or specific constraints based on prior information. The degree to which each component is prioritized can be adjusted by the user to enable enhanced exploratory analysis. Additionally, our framework can be used to efficiently compare clusterings with different numbers of clusters. We demonstrate our method's characteristics and its advantages over previous work through through quantitative analysis, and we illustrate the value of this approach through several case studies in healthcare analytics.

Method

Consider a sparse count matrix with N rows and K columns. Define M an $N \times K$ as a matrix with elements log transformed using the function $f(x) = \log(1 + x)$. We are interested in clustering over rows (providers) and columns (prescriptions and procedures). We will consider clustering across rows and columns separately. For each case, we construct an autoencoder consisting of an encoder Ψ_1 and a decoder Ψ_2. Define the hidden representation h of size H such that $\Psi_1(M) = h$ and $\hat{M}$ as the reconstruction given by $\Psi_2 \circ \Psi_1(M) = \hat{M}$. Define Ψ_3 as the function transformation for the clustering module that takes as input h and outputs cluster probabilities c: $\Psi_3(h) = c$. Fig. 1 illustrates the framework for the neural network.

Table 1: Deep network architecture

Layer	Index	Size	Act.
(Input)	–	$B \times K$	
Dense	0	$B \times H$	LReLU
Haar wavelet	–	$B \times H$	–
Expand, permute	–	$B \times H \times P$	–
Pool	1	$B \times H$	–
Sum (0) and (1)	–	$B \times H$	–
Batch norm	–	$B \times H$	–
Dense	–	$B \times H$	LReLU
Batch norm	–	$B \times H$	–
Dense	–	$B \times H$	LReLU
Batch norm	–	$B \times H$	–
Dense	–	$B \times H$	LReLU
Dense	–	$B \times H$	LReLU
(Hidden)	2	$B \times H$	
Dense	–	$B \times H$	LReLU
Dense	–	$B \times H$	LReLU
Dense	–	$B \times K$	LReLU
(Output)	–	$B \times K$	
Copy (2)	–	$B \times H$	–
Dense	–	$B \times H$	LReLU
Dense	–	$B \times H$	LReLU
Dense	–	$B \times C$	Softmax
(Cluster)	–	$B \times C$	

For log-transformed sparse count data, cosine similarity is a useful distance representation. However, computing the cosine similarity for all pairs may be problematic when N is large: $O(N^2)$. We propose to learn a hidden representation $h = \{h^\circ, h^{||\cdot||}\}$ given by angle h° and norm $h^{||\cdot||}$, such that the distance measure between examples in M is approximately preserved in h°. To achieve this, we define batch pairwise similarity ($O(B^2)$) and its loss $\mathcal{L}_{h^\circ,O}$ as a penalization to the autoencoder loss. For sparse count data, we use cosine similarity.

Define batch size B such that h_B and $\hat{M}_B$ are of size $B \times \{\cdot\}$ for H and M respectively. Let $\delta(x_i, x_j)$ be the pairwise distance measure. Define δ_B the batch distance measure, *i.e.* $\delta_B(\{x_1, \ldots, x_B\}) = [\delta(x_i, x_j)]_{ij}\ \forall i, j \in \{1, \ldots, B\}$. Then, $\mathcal{L}_{h^\circ,O} = \frac{1}{B}||\delta_B(h^\circ_B) - \delta_B(\hat{M}_B)||^2_2$.

While $\mathcal{L}_{h^\circ,O}$ encourages matching distances in h° and $\hat{M}$, two vectors in h could have the same angle but different embedding locations, *i.e.*, different $h^{||\cdot||}$. Note this could be problematic because the hidden vectors are used to learn cluster membership probabilities c. To encourage approxi-

mate injectivity, we introduce the loss $\mathcal{L}_{||\cdot||=1}$ that penalizes hidden representations away from the surface of the unit norm hypersphere. We could enforce this as a hard constraint, however, the ability to violate the constraint may facilitate alignment of the embedded representation angle with that of the output space.

While the hidden representation h° approximately preserves distances of the output space, its unit norm and approximate injectivity are useful for our clustering. Note that for a probability vector that sums to 1, its element-wise square root vector has unit norm and can be interpreted as an angle. Therefore, we can match the angular representation of $c^{\frac{1}{2}}$ to that of h° with batch cosine similarity.

Define $c_B = \Psi_3(h_B)$ the set of probability vectors indicating cluster membership. Note that the cosine similarity of ${c_B}^{\frac{1}{2}}$ is non-negative and we are not interested in incurring loss due to differences between 0 and negative cosine similarities. We define $\mathcal{L}_{c,h^\circ} = \frac{1}{B}||\delta_B({c_B}^{\frac{1}{2}}) - \max(\delta_B(h^\circ_B), 0)||_2^2$.

We consider additional loss terms to assist customizable representation learning. First, we introduce an entropy loss term to encourage cluster probabilities to be spread across more than one cluster. This encourages large clusters with overlap to spread across multiple clusters to reveal subgroup characteristics and enables injecting belief about characteristics of the clustering. For optimization, it encourages exploration in cluster membership and could help avoid local optima.

Second and optionally, the framework can use accessory information to inform the clustering. We investigate the following: our framework encodes a single hidden representation for clustering, but could use multiple settings of numbers of clusters C. In place of a single softmax, the terminal layer C can be a concatenation of possible numbers of clusters, *e.g.* 2, 3, ..., 100, where the softmax is applied to nodes corresponding to each of the possible numbers. Therefore, one can provide a cluster assignment of size C_{in} that informs the hidden representation and request a more or less granular clustering C_{out}. Unlike hierarchical clustering, this extension produces a multiarchy.

Thus, our overall objective function is $\sum_i \lambda_i \mathcal{L}_i$ for $\mathcal{L}_i \in \{\mathcal{L}_{M,\hat{M}}, \mathcal{L}_{c,h^\circ}, \mathcal{L}_{h^\circ,O}, \mathcal{L}_{||h^\circ||=1}, \mathcal{L}_{entropy}, \mathcal{L}_{C_{in}}\}$. Unless otherwise specified, we set λ_i respectively: $\lambda_i \in \{1, 1, 10^{-1}, 10^{-1}, 10^{-4}, 0\}$.

The autoencoder framework we adopt is shown in Table 1. The permute-pool layer copies the tensor P times, permutes the values, and performs max-pooling over the P dimension with size 3 and stride 2. For our experiments we set $B = 128$, $P = 4$, and $H = 128$. We set C to be twice the desired number of centroids, and in post-processing merge the clusters identified based on maximum pairwise cluster distances.

Results

We evaluate our method on real and simulated data. Our results on simulated data show (1) that our method produces superior clusterings, in this setting, than other recently proposed deep clustering methods, and (2) how different components of our method affect the results of the method. Our case study evaluations on three real-world healthcare data show (1) the quality of our method's clustering results, (2) the breadth of applicability of this method in a real-world context, and (3) the method's scalability on large real-world data.

Table 2: Quantitative evaluation on simulated data, relative to the true cluster labels. Data are generated as described in the text, with baseline parameters set at centroids=25, samples=1000, dims=1000, sod=0.01, explode=10000, and varied in each dataset as indicated. Data are clustered using our method (Ours), and the closest neural embedding clustering techniques DCN, SAE+KM, and DEC. The best score is in boldface.

	Adjusted Rand Index (ARI)			
Dataset	Ours	SAE+KM	DEC	DCN
baseline	**1.00**	0.32	0.10	0.39
samples=100	**0.82**	0.36	0.26	0.36
dims=100000	**0.64**	0.03	0.00	0.01
explode=1000000	**1.00**	0.36	0.12	0.47
centroids=100	**1.00**	0.18	0.00	0.18
sd=0.1	**0.25**	0.02	0.00	0.02

Table 3: Quantitative evaluation of variants of our method. Each column corresponds to a different variant of our own method, as defined in the text. The best score(s) is in boldface.

	Adjusted Rand Index (ARI)				
Dataset	Ours	Headless	HKM	Clu:t-SNE	Monotonic
baseline	**1.00**	0.95	0.21	0.96	**1.00**
samples=100	0.82	0.43	0.09	**0.84**	0.75
dims=100000	**0.64**	0.05	0.01	0.13	0.16
explode=1000000	**1.00**	0.99	0.91	0.93	**1.00**
centroids=100	**1.00**	0.72	0.61	0.08	0.75
sd=0.1	0.25	0.02	0.01	0.24	**0.41**

Quantitative evaluation

First, on simulated data generated so as to be similar to our target healthcare applications, we show that our method typically finds a clustering that, under several measures, is more similar to the ground truth labels from the simulation than competing methods are.

We generate data as follows. Centroids are sampled from a multivariate normal $\mathcal{N}(\mathbf{0}, \mathbf{1})$ and normalized onto the unit ball. We generate an equal number of samples for each centroid by adding noise distributed according to $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Then we translate these points away from the origin by multiplying each point by an "explosion" factor κ drawn from a random uniform on $[1, \kappa]$. The purpose of this step is to make the data more similar to that in our healthcare applications, where patients and providers often have widely varying total encounter counts (e.g., due to length of tenure) unrelated to the clustering problem at hand.

Thus, the simulation is parameterized by the number of centroids, the number of dimensions, the explode factor, the

number of samples, and the standard deviation of the noise. We set each parameter to a baseline level and vary one at a time to test our algorithm.

Comparison to other methods. We cluster the data using our method ("Ours"), as well as three leading deep clustering frameworks that are similar to our method in combining autoencoders and clustering criteria: Stacked Autoencoder plus K-means (SAE+KM) (Tian et al. 2014), Deep Embedded Clustering (DEC) (Xie, Girshick, and Farhadi 2016), and Deep Clustering Network (DCN) (Yang et al. 2017). We evaluate results of these clustering methods using the Adjusted Rand Index (ARI) relative to the ground truth labels from the simulation.

Results are shown in Table 2. Our method produces better clusterings than the other methods on all six simulated datasets, as measured by ARI. Of note, the two-step SAE+KM method performs approximately as well as the integrated DCN and DEC methods on this data. This suggests that merely integrating clustering and representation learning into a single framework is not enough on its own to produce a good clustering; rather, it is also essential in this context for the method to utilize geometric data, as our method does.

Comparison to variants of our method. Next, we compare our method to variants of our method. These variants are created by omitting different components of the method one at a time, and thus, this evaluation demonstrates the importance of each component.

In the first variant ("Headless"), we remove the decoder from the framework. Doing so results in worse clustering performance on all six datasets. The possible limitation of decoder removal is that h no longer needs to preserve the information in the input because the network does not need to reconstruct the input.

In the second variant ("HKM"), we run k-means in the hidden, embedded space our method constructs, instead of using our method's clustering module. The embedded space is not designed to work well with k-means, and in fact, is constrained to have similar geometry to the original space, where k-means performs poorly. Likely for this reason, the HKM variant performs worse than our method on all datasets.

In the third variant ("Clu:t-SNE"), we substitute a t-SNE manifold constraint for $\mathcal{L}_{h^\circ,O}$, with the exception that to conform with the overall architecture of our method, instead of computing all $O(N^2)$ pairwise distances, we define the KL-loss based on $O(B^2)$ batch pairwise distances. This variant slightly outperforms our method when sample size is small ("samples=100"), but otherwise performs worse than our method.

The final variant ("Monotonic") adds a constraint to our method rather than removing a component. Specifically, motivated by autoregressive flow literature (Huang et al. 2018), we apply positive weight constraints and strictly monotonic activations (already present) to our network to get a autoregressive network. This variant outperforms ours in one case, but otherwise achieves equal or worse performance.

Single-representation multiple clustering. To illustrate single-representation multiple clustering, we run the simulation variant with 100 samples and ten clusters, with samples 0 through 9 in cluster 0, 10 through 19 in cluster 1, and so on. Figure 2 illustrates the predicted cluster assignments for different C along the x axis, with colors for a given C determining co-cluster membership and with colors for different C learned to be close to the previous cluster color but far from other cluster colors. The figure demonstrates recovery of the ground truth clusters and also provides clusterings identified when $C \neq 10$.

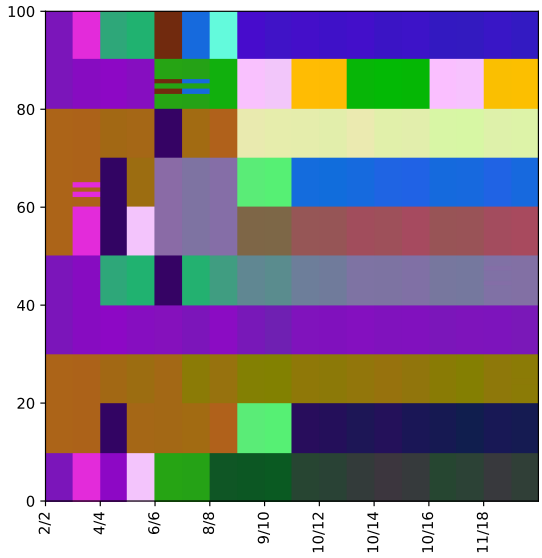


Figure 2: Single-representation multiple clustering with x-axis: number of clusters found divided by number available, and y-axis: ordered samples. Best viewed in color.

Case study: Biclustering healthcare providers, prescriptions, and procedures with CMS data

The quantitative results in the previous section show that our method performs quite well on simulated data that we constructed with an eye to its similarity to real-world high-dimensional sparse count healthcare data. In this section, we complement these quantitative results with an extensive case study to demonstrate that our method does indeed work quite well on real-world data for which it is designed. We (1) describe the case study's objectives and data, (2) compare our method's results to those obtained by k-means, and (3) use our method to successfully investigate three questions of interest to our physician co-authors.

Objectives and Data. The purpose of this case study is to cluster healthcare providers, prescriptions, and procedures based on data obtained from the Centers for Medicare and Medicaid Services, with the ultimate objective to gain insights on providers' patterns of care based on this clustering, as well as on groupings of prescription and procedure use. To that end, we obtained Medicare Provider Utilization and Payment Data: Part D Prescriber Summary Table

CY2015 (Medicare 2017), which tabulates all prescriptions and procedures given under the Medicare Part D program in 2015 in the United States. This data consists of approximately 10^6 providers, 10^4 procedures, and 10^3 drugs. We used our method and k-means to separately cluster healthcare providers, prescriptions, and procedures based on this data.

The Medicare database assigns each provider a specialty code. We do not use these codes for clustering, but we do reference them below to further understand each clustering and for quality assessment. In the interest of space, we only show results for a clustering of providers based on the prescriptions they gave. Our physician authors assessed the quality of the resulting clusters.

Comparison to k-means. The clusterings formed by our method and by k-means, each with 20 clusters, are shown in Table 4. On examination of these results, our method produces qualitatively better clusters than k-means does. For example, our clustering consistently includes a larger fraction of specialists in specialist clusters, *e.g.*, Dentist (85k), Psychiatry (21k), Emergency Medicine (20k). Our clustering, unlike k-means, also identifies clean obstetrics and hematology oncology clusters. Although k-means identifies a cardiology and interventional cardiology cluster that our method does not initially identify, our clustering identified this cluster and merged it (Cardiology (18k), Nurse Prac (3k), Internal Medicine (2k)) into the internal medicine subgroup in post-processing.

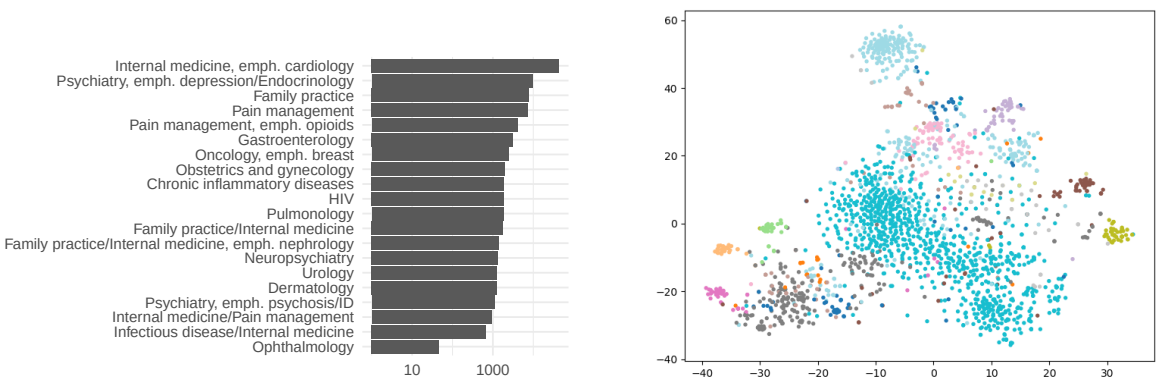
Our method also provides insight in regard to providers in ontology specialties that are not as common. For example, our method's urology cluster also includes a large fraction of the radiation oncologists in our dataset. On detailed assessment of this cluster, we found that urology medications tamsulosin and finasteride were most commonly prescribed in this cluster and that radiation oncologists most commonly prescribed tamsulosin, followed by hydrocodone/acetaminophen and dexamethasone, possibly for prevention and treatment of complications of radiation therapy. Our clustering identified radiation oncologists and urologists as being similar according to the drugs they commonly prescribe, a finding that would not be identified through the use of a standard ontology alone.

To investigate our method's biclustering (clustering on dimensions of provider and prescriptions), we investigated three questions: (1) does the clustering provide insight into how nurse practitioners provide care, (2) similarly, can we differentiate among family practice and internal medicine doctor subtypes, and (3) is the pain management cluster identifying commonality in opioid prescribing despite provider membership to many specialties?

Nurse practitioners. Nurse practitioners (NPs) tend to provide care in care units defined by medical specialties and subspecialties, but the CMS does not provide characteristics at that granularity. To characterize the type of care they provide, we can investigate their membership to identified clusters. Figure 3(a) shows the breakdown of membership across clusters. We pulled the top 10 prescriptions administered by NP cluster members, as defined by mean $\log(1+\cdot)$, and used our clinical expertise to characterize the care provided. Ev-

Table 4: Provider clustering by (a) our method, (b) k-means. Each row corresponds to a cluster. We display the top three specialties associated with providers in that cluster, and, in parentheses, the number of providers in the cluster with each of those specialties.

(a) Our Clustering (Top 3 Specialties)
Dentist (85k); Oral Surgery (3k); Podiatry (1k)
Internal Med (82k); Family Practice (82k); Nurse Prac (44k)
Psychiatry (21k); Nurse Prac (9k); Psychiatry & Neurology (6k)
Emergency Medicine (21k); Orthopedic Surgery (15k); Phys Asst (15k)
Optometry (18k); Ophthalmology (17k); Student (<1k)
Obstetrics/Gynecology (17k); Nurse Prac (2k); Phys Asst (<1k)
Gastroenterology (12k); Nurse Prac (3k); Internal Med (3k)
Dermatology (10k); Phys Asst (3k); Nurse Prac (1k)
Neurology (10k); Nurse Prac (1k); Psychiatry & Neurology (1k)
Urology (9k); Phys Asst (1k); Nurse Prac (1k)
Nurse Prac (8k); Phys Asst (6k); Emergency Medicine (6k)
Pulmonary Disease (7k); Allergy/Immunology (3k); Otolaryngology (2k)
Hematology/Oncology (6k); Nurse Prac (2k); Medical Oncology (2k)
Internal Med (5k); Emergency Medicine (3k); Nurse Prac (2k)
Phys Asst (4k); Nurse Prac (4k); Orthopedic Surgery (2k)
Dentist (3k); Emergency Medicine (2k); Phys Asst (2k)
Infectious Disease (3k); Obstetrics/Gynecology (2k); Nurse Prac (2k)
Pharmacist (2k); Nurse Prac (1k); Internal Med (1k)
Podiatry (2k); Nurse Prac (1k); Optometry (1k)
Physical Med/Rehab (2k); Podiatry (2k); Nurse Prac (2k)
(b) K-Means Clustering (Top 3 Specialties)
Dentist (52k); Nurse Prac (1k); Phys Asst (1k)
Nurse Prac (35k); Phys Asst (27k); Internal Med (27k)
Family Practice (23k); Internal Med (17k); Nurse Prac (10k)
Family Practice (22k); Internal Med (20k); Nurse Prac (4k)
Emergency Medicine (20k); Orthopedic Surgery (13k); Phys Asst (12k)
Dentist (19k); Oral Surgery (3k); Maxillofacial Surgery (1k)
Internal Med (16k); Nurse Prac (12k); Family Practice (9k)
Family Practice (16k); Nurse Prac (13k); Internal Med (11k)
Cardiology (14k); Nurse Prac (1k); Interventional Cardiology (1k)
Dentist (14k); Oral Surgery (<1k); Infectious Disease (<1k)
Optometry (12k); Ophthalmology (5k); Student (<1k)
Ophthalmology (11k); Optometry (2k); Student (<1k)
Internal Med (11k); Family Practice (10k); General Practice (1k)
Psychiatry (10k); Nurse Prac (3k); Psychiatry & Neurology (1k)
Psychiatry (8k); Psychiatry & Neurology (3k); Nurse Prac (3k)
Urology (8k); Phys Asst (1k); Nurse Prac (1k)
Neurology (7k); Nurse Prac (<1k); Phys Asst (<1k)
Pulmonary Disease (6k); Allergy/Immunology (2k); Otolaryngology (1k)
Neurology (3k); Nurse Prac (1k); Physical Med/Rehab (1k)
Rheumatology (2k); Physical Med/Rehab (2k); Nurse Prac (2k)



(a) (b)

Figure 3: (a) Histogram of nurse practitioner counts by cluster. Cluster names are given by clinical assessment of top 10 prescriptions. (b) t-SNE representation of 2000 randomly selected nurse practitioners. Cluster membership is indicated by color.

idently, the NPs do cluster based on prescription behavior alon lines of medical specialties. The top 10 medications for each cluster are provided in the Appendix. To empirically demonstrate the variation in the NP clustering, Figure 3(b) shows a t-SNE(cos) plot of a random subset of 2000 NPs to illustrate their similarity (manifold distance) alongside their cluster membership (color). This demonstrates that the method does provide cluster separation and could be used to select NPs based on approximate specialization for future investigation.

Family practice and internal medicine. Similar to nurse practitioners, the specialty titles "family practice" and "internal medicine" are underdifferentiated. To investigate these providers' prescription patterns, we inspected the FP/IM predominating clusters for differences. Figure 4(a) shows the breakdown of the providers across clusters, with labels corresponding to average care descriptions of the members in those groups. We selected three of the clusters that roughly corresponded to IM emphasis cardiology (IMC), IM emphasis infectious disease (IMID), and FP emphasis infectious disease (FPID). We compared the three clusters by listing the three prescriptions with the largest mean $\log(1 + \cdot)$ difference in prescribing.

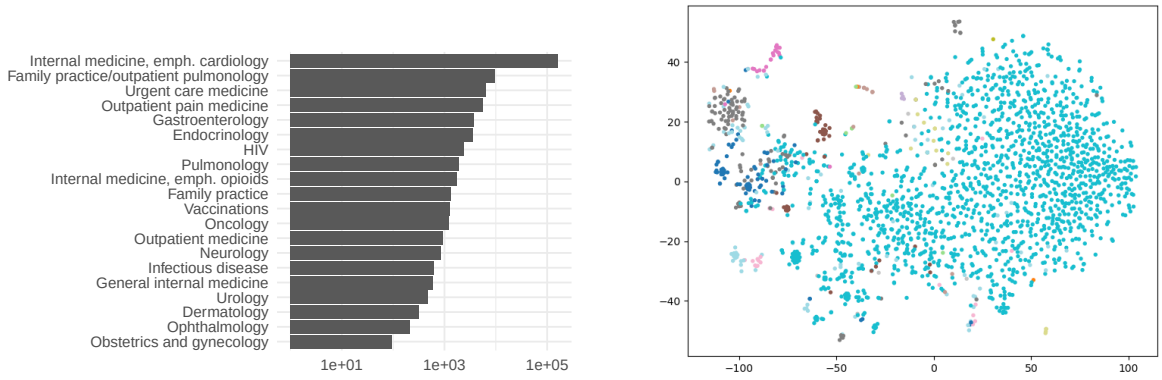
- Compared to FPID providers, IMC providers prescribed more general IM medication ["levothyroxine", "atorvastatin", "lisinopril"] and fewer antibiotics ["levofloxacin", "amoxicillin-clavulanate", "azithromycin"].
- Compared to IMID providers, IMC providers prescribed more general internal medicine medication ["atorvastatin", "levothyroxine", "lisinopril"] and fewer outpatient antibiotics ["clindamycin", "amoxicillin", "doxycycline"].
- Compared to FPID, IMID providers prescribed stronger community-acquired pneumonia antibiotics ["levofloxacin", "azithromycin", "amoxicillin-clavulanate"] and fewer older and narrow-spectrum antibiotics and community care medications ["clindamycin", "amoxicillin", "ranitidine"].

Table 5: Top three medications by specialty for providers in opioid-prescribing cluster.

Specialty	Prescriptions
Anesthesiology	oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen
Emergency medicine	oxycodone hcl, hydrocodone/acetaminophen, prednisone
Family practice	oxycodone hcl, hydrocodone/acetaminophen, morphine sulfate
General practice	oxycodone hcl, hydrocodone/acetaminophen, gabapentin
General surgery	oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen
Hematology/oncology	oxycodone hcl, morphine sulfate, ondansetron hcl
Hospice and palliative care	oxycodone hcl, morphine sulfate, fentanyl
Internal medicine	oxycodone hcl, morphine sulfate, hydrocodone/acetaminophen
Interventional pain management	hydrocodone/acetaminophen, oxycodone hcl, oxycodone hcl/acetaminophen
Medical oncology	oxycodone hcl, ondansetron hcl, dexamethasone
Neurology	oxycodone hcl, gabapentin, hydrocodone/acetaminophen
Neurosurgery	oxycodone hcl, hydrocodone/acetaminophen, gabapentin
Nurse practitioner	oxycodone hcl, hydrocodone/acetaminophen, morphine sulfate
Obstetrics/gynecology	oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen
Orthopaedic surgery	oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen
Orthopedic surgery	oxycodone hcl, hydrocodone/acetaminophen, tramadol hcl
Otolaryngology	oxycodone hcl, fluticasone propionate, oxycodone hcl/acetaminophen
Pain management	oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen
Physical medicine and rehabilitation	oxycodone hcl, hydrocodone/acetaminophen, gabapentin
Physician assistant	oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen
Plastic and reconstructive surgery	oxycodone hcl, hydrocodone/acetaminophen, cephalexin
Podiatry	oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen
Radiation oncology	oxycodone hcl, tamsulosin hcl, lidocaine hcl
Student	oxycodone hcl, hydrocodone/acetaminophen, oxycodone hcl/acetaminophen

Pain management. Table 5 contains, by specialty, the top three medications prescribed by providers in the pain management cluster. The clustering demonstrates that, irrespective of the specialty, these providers prescribe large quantities of various opioids, including oxycodone, hydrocodone, morphine, fentanyl and others. We also compared the opioid-prescribing cluster against the other two clusters containing over 100 pain management specialty providers by listing the three prescriptions with the largest mean $\log(1 + \cdot)$ difference.

- Compared to the first cluster, the opioid-prescribing cluster prescribed more ["oxycodone", "morphine", "gabapentin"] and less ["celphalexin", "prednisone", "tramadol"].
- Compared to the second cluster, the opioid-prescribing cluster prescribed more ["oxycodone", "morphine",



(a) (b)

Figure 4: (a) Histogram of FP and IM provider counts by cluster. Cluster names given by clinical assessment of top 10 prescriptions. (b) t-SNE representation of 2000 randomly selected FP and IM providers. Cluster membership indicated by color.

"gabapentin"] and less ["ibuprofen", "acetaminophen with codeine", "azithromycin"].

Other Case Studies

In addition to our quantitative results on simulated data and our lengthy case study using CMS data, we also carried out two additional case studies. These case studies (1) demonstrate the ability of our method to work well on real-world datasets of large and small size, and (2) further demonstrate the ability of our method to provide clinically meaningful insights about high-dimensional healthcare data.

Identifying multiple myeloma subgroups. In this case study, we conducted clustering on patients diagnosed with multiple myeloma. The case study's cohort is derived from outpatient medical records of a regional care system from 2015 to 2018 (with services from outpatient clinic to tertiary care). The goal for this application was to identify subgroups of individuals based on the myeloma-relevant medications prescribed to these individuals. The regional care biclustering of patients with encounters for multiple myeloma and monogammopathy of unknown significance, according to ICD 9 and 10 billing codes, is shown in the Appendix. Our method successfully identifies characteristic treatment regimens and patients taking those regimens, such as the combination of lenalidomide and dexamethasone. It also identifies clusters of individuals with billing codes for myeloma but without any relevant prescription or prescription of immunologic therapy. These clusters could indicate individuals who were tested for but do not possess the disease, and individuals with the disease receiving care outside of network.

Characterizing radiology notes. In our final case study, we sought to organize a large collection of radiology notes based on the notes' contents, for individuals with critical care needs. We used de-identified radiology notes from MIMIC III v1.4 (Johnson et al. 2016), a natural language processing pipeline, and a bag of words representation to cluster radiology notes. The note descriptions provided are specific in some senses, *e.g.* anatomical: "X-ray of left foot 4th digit, two views", but nonspecific in other senses, *e.g.* indication: "CT chest w/o contrast" for cardiac, pulmonary, gastrointestinal disease, or other? After stop word removal, stemming, and lemmatization, the bag-of-word representation resulted in a sparse matrix of size: 522,279 notes by 275,263 unique word tokens. We performed clustering with 30 clusters. The result was a meaningful clustering, with for example, top members of distinct clusters including: kidney ultrasounds ["Renal US", "P Renal U.S. Port", "Renal transplant U.S."], infant radiographs ["Babygram (chest only)", "Neonatal head portable", "P babygram (chest only) portable"], and abdominal ultrasounds ["Liver or gallbladder US", "Abdomen US (complete study)", "US abd limit, single organ"]. The full cluster list is available in the Appendix.

Conclusion

In this paper, we defined a deep clustering method to address this task of clustering high-dimensional sparse count data, as often arises in analysis of healthcare data and other fields. To do so, we described a generic architecture that takes into account not just image recovery and clustering quality, but also geometric relationships in the data. We demonstrated that this approach works well through quantitative comparisons with other methods, as well as through detailed case studies on real healthcare data.

References

Aljalbout, E.; Golkov, V.; Siddiqui, Y.; and Cremers, D. 2018. Clustering with deep learning: Taxonomy and new methods. *CoRR* abs/1801.07648.

Ding, C., and He, X. 2004. K-means clustering via principal component analysis. *Proceedings of the Twenty-first International Conference on Machine Learning*.

Huang, C.-W.; Krueger, D.; Lacoste, A.; and Courville, A.

2018. Neural autoregressive flows. *Proceedings of Machine Learning Research: International Conference on Machine Learning*.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data* 3:160035.

Medicare. 2017. *Medicare Provider Utilization and Payment Data: Part D Prescriber Summary Table CY2015 (accessed April 30, 2018)*.

Tian, F.; Gao, B.; Cui, Q.; Chen, E.; and Liu, T.-Y. 2014. Learning deep representations for graph clustering. *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Trigeorgis, G.; Zafeiriou, K. B. S.; and Schuller, B. W. 2014. A deep semi-nmf model for learning hidden representations. *International Conference on Machine Learning*.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. *ICML*.

Yang, B.; Fu, X.; Sidiropoulos, N. D.; and Hong, M. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. *ICML*.

Appendix for: Deep Clustering on a Hypersphere for High-Dimensional Healthcare Data

Table 1: Top 10 drugs prescribed by nurse practitioners, by cluster

Cluster5	Cluster6	Cluster7	Cluster9	Cluster11
GABAPENTIN	PREDNISONE	PREDNISOLONE ACETATE	IBUPROFEN	TRIAMCINOLONE ACETONIDE
TIZANIDINE HCL	LEVOFLOXACIN	LATANOPROST	ACETAMINOPHEN WITH CODEINE	CLOBETASOL PROPIONATE
HYDROCODONE/ACETAMINOPHEN	AZITHROMYCIN	OFLOXACIN	HYDROCODONE/ACETAMINOPHEN	KETOCONAZOLE
CYCLOBENZAPRINE HCL	HYDROCODONE/ACETAMINOPHEN	KETOROLAC TROMETHAMINE	PREDNISONE	MUPIROCIN
BACLOFEN	CIPROFLOXACIN HCL	GATIFLOXACIN	TRAMADOL HCL	DOXYCYCLINE HYCLATE
TRAMADOL HCL	SULFAMETHOXAZOLE/TRIMETHOPRIM	NEPAFENAC	CEPHALEXIN	METRONIDAZOLE
MELOXICAM	ALBUTEROL SULFATE	TIMOLOL MALEATE	AZITHROMYCIN	FLUOCINONIDE
PREGABALIN	FUROSEMIDE	ERYTHROMYCIN BASE	SULFAMETHOXAZOLE/TRIMETHOPRIM	FLUOROURACIL
DULOXETINE HCL	PANTOPRAZOLE SODIUM	DORZOLAMIDE HCL/TIMOLOL MALEAT	CYCLOBENZAPRINE HCL	CLINDAMYCIN PHOSPHATE
OXYCODONE HCL/ACETAMINOPHEN	METHOTREXATE SODIUM	TRAVOPROST	NAPROXEN	DESONIDE

Cluster15	Cluster21	Cluster22	Cluster24	Cluster25
TAMSULOSIN HCL	ANASTROZOLE	OMEPRAZOLE	LEVETIRACETAM	CLINDAMYCIN HCL
CIPROFLOXACIN HCL	WARFARIN SODIUM	PANTOPRAZOLE SODIUM	GABAPENTIN	AMOXICILLIN
FINASTERIDE	PROCHLORPERAZINE MALEATE	RANITIDINE HCL	TOPIRAMATE	PREDNISONE
OXYBUTYNIN CHLORIDE	LETROZOLE	ESOMEPRAZOLE MAGNESIUM	DONEPEZIL HCL	HYDROCODONE/ACETAMINOPHEN
SOLIFENACIN SUCCINATE	ONDANSETRON HCL	DICYCLOMINE HCL	CARBIDOPA/LEVODOPA	AZITHROMYCIN
MIRABEGRON	DEXAMETHASONE	POLYETHYLENE GLYCOL 3350	LAMOTRIGINE	CEPHALEXIN
SULFAMETHOXAZOLE/TRIMETHOPRIM	TAMOXIFEN CITRATE	DEXLANSOPRAZOLE	MEMANTINE HCL	SULFAMETHOXAZOLE/TRIMETHOPRIM
NITROFURANTOIN MONOHYD/M-CRYST	OXYCODONE HCL	LINACLOTIDE	DIVALPROEX SODIUM	AMOXICILLIN/POTASSIUM CLAV
CEPHALEXIN	HYDROCODONE/ACETAMINOPHEN	SUCRALFATE	BACLOFEN	TRAMADOL HCL
NITROFURANTOIN MACROCRYSTAL	GABAPENTIN	MESALAMINE	CLONAZEPAM	CIPROFLOXACIN HCL

Cluster26	Cluster28	Cluster29	Cluster30	Cluster32
FLUTICASONE PROPIONATE	OXYCODONE HCL	AZITHROMYCIN	RITONAVIR	SIMVASTATIN
ALBUTEROL SULFATE	HYDROCODONE/ACETAMINOPHEN	AMOXICILLIN/POTASSIUM CLAV	EMTRICITABINE/TENOFOVIR	SEVELAMER CARBONATE
MONTELUKAST SODIUM	MORPHINE SULFATE	PREDNISONE	DARUNAVIR ETHANOLATE	NORGESTIMATE-ETHINYL ESTRADIOL
FLUTICASONE/SALMETEROL	OXYCODONE HCL/ACETAMINOPHEN	CIPROFLOXACIN HCL	SULFAMETHOXAZOLE/TRIMETHOPRIM	PEG 3350/NA SULF,BICARB,CL/KCL
TIOTROPIUM BROMIDE	GABAPENTIN	FLUTICASONE PROPIONATE	RALTEGRAVIR POTASSIUM	MEDROXYPROGESTERONE ACETATE
PREDNISONE	FENTANYL	ALBUTEROL SULFATE	EFAVIRENZ/EMTRICITAB/TENOFOVIR	LAMOTRIGINE
BUDESONIDE/FORMOTEROL FUMARATE	TRAMADOL HCL	LEVOFLOXACIN	ATAZANAVIR SULFATE	LEDIPASVIR/SOFOSBUVIR
AZITHROMYCIN	PREGABALIN	SULFAMETHOXAZOLE/TRIMETHOPRIM	FLUCONAZOLE	CITALOPRAM HYDROBROMIDE
OMEPRAZOLE	TIZANIDINE HCL	METHYLPREDNISOLONE	ABACAVIR SULFATE/LAMIVUDINE	CINACALCET HCL
MOMETASONE/FORMOTEROL	METHADONE HCL	CEPHALEXIN	LISINOPRIL	FLUOXETINE HCL

Cluster33	Cluster35	Cluster36	Cluster38	Cluster39
ESTRADIOL	COLLAGENASE CLOSTRIDIUM HIST.	LISINOPRIL	TRAZODONE HCL	HYDROCODONE/ACETAMINOPHEN
ESTROGENS, CONJUGATED	RISPERIDONE	AMLODIPINE BESYLATE	CLONAZEPAM	TRAMADOL HCL
FLUCONAZOLE	DIVALPROEX SODIUM	LEVOTHYROXINE SODIUM	QUETIAPINE FUMARATE	OXYCODONE HCL/ACETAMINOPHEN
MEDROXYPROGESTERONE ACETATE	BENZTROPINE MESYLATE	ATORVASTATIN CALCIUM	SERTRALINE HCL	CEPHALEXIN
ALENDRONATE SODIUM	OLANZAPINE	OMEPRAZOLE	LORAZEPAM	PREDNISONE
VALACYCLOVIR HCL	ENOXAPARIN SODIUM	METFORMIN HCL	RISPERIDONE	MELOXICAM
PROGESTERONE,MICRONIZED	CLOZAPINE	SIMVASTATIN	ALPRAZOLAM	GABAPENTIN
OXYBUTYNIN CHLORIDE	URSODIOL	FUROSEMIDE	BUPROPION HCL	CYCLOBENZAPRINE HCL
CLOBETASOL PROPIONATE	GENTAMICIN SULFATE	METOPROLOL TARTRATE	FLUOXETINE HCL	OXYCODONE HCL
METRONIDAZOLE	LITHIUM CARBONATE	HYDROCHLOROTHIAZIDE	DULOXETINE HCL	METHYLPREDNISOLONE

Table 2: Top three radiology descriptions per cluster with counts.

0	CT ABDOMEN W/CONTRAST	6393
0	CT CHEST W/CONTRAST	5350
0	CT ABDOMEN W/O CONTRAST	3892
1	CHEST (PORTABLE AP)	13010
1	CHEST (PA & LAT)	611
1	CHEST (SINGLE VIEW)	108
2	CHEST (PORTABLE AP)	11938
2	CHEST (PA & LAT)	1693
2	CT CHEST W/O CONTRAST	857
3	CHEST (PORTABLE AP)	14512
3	CHEST PORT. LINE PLACEMENT	1725
3	TRAUMA #3 (PORT CHEST ONLY)	217
4	BABYGRAM (CHEST ONLY)	3117
4	NEONATAL HEAD PORTABLE	2782
4	P BABYGRAM (CHEST ONLY) PORT	1609
5	LIVER OR GALLBLADDER US (SINGLE ORGAN)	4522
5	ABDOMEN U.S. (COMPLETE STUDY)	2102
6	CT C-SPINE W/O CONTRAST	4632
6	T-SPINE	1091
6	L-SPINE (AP & LAT)	937
7	CHEST (PORTABLE AP)	21394
7	CHEST (PA & LAT)	1968
7	CT HEAD W/O CONTRAST	604

8	CHEST (PORTABLE AP)	4785
8	CHEST (PA & LAT)	877
8	CHEST (SINGLE VIEW)	442
9	CHEST (PORTABLE AP)	8740
9	CT CHEST W/O CONTRAST	2726
9	CHEST (PA & LAT)	2031
10	CHEST (PORTABLE AP)	9343
10	CHEST (PA & LAT)	3046
10	VIDEO OROPHARYNGEAL SWALLOW	2246
11	CT HEAD W/O CONTRAST	28423
11	CT HEAD W/ & W/O CONTRAST	589
11	CTA HEAD W&W/O C & RECONS	564
12	CHEST (PORTABLE AP)	13728
12	CHEST (PA & LAT)	791
12	CHEST PORT. LINE PLACEMENT	298
13	MR HEAD W & W/O CONTRAST	6623
13	MR HEAD W/O CONTRAST	2824
13	MR C-SPINE W& W/O CONTRAST	638
14	PORTABLE ABDOMEN	5587
14	ABDOMEN (SUPINE & ERECT)	4139
14	ABDOMEN (SUPINE ONLY)	706
15	CHEST (PORTABLE AP)	6515
15	CT HEAD W/O CONTRAST	3829
15	CHEST (PA & LAT)	1389
16	PELVIS (AP ONLY)	824
16	L HIP UNILAT MIN 2 VIEWS LEFT	793
17	CHEST (PORTABLE AP)	11202
17	CHEST (PA & LAT)	2767
17	CHEST PORT. LINE PLACEMENT	264
18	CAROTID SERIES COMPLETE	3287
18	CTA HEAD W&W/O C & RECONS	1783
19	RENAL U.S.	2668
19	P RENAL U.S. PORT	911
19	RENAL TRANSPLANT U.S.	536
20	CHEST (PORTABLE AP)	4092
20	CHEST (PA & LAT)	778
20	CHEST PORT. LINE PLACEMENT	430
21	CHEST (PORTABLE AP)	6725
21	CHEST (PA & LAT)	266
21	CHEST (SINGLE VIEW)	186
22	CHEST (PORTABLE AP)	7657
22	CHEST (PA & LAT)	827
22	CHEST PORT. LINE PLACEMENT	495
23	CHEST (PORTABLE AP)	7086
23	PORTABLE ABDOMEN	1473
23	NASO-INTESTINAL TUBE PLACEMENT (W/FLUORO)	1347
24	CHEST (PORTABLE AP)	19050
24	CHEST PORT. LINE PLACEMENT	4408
24	BY DIFFERENT PHYSICIAN	2198
25	CHEST PORT. LINE PLACEMENT	11057
25	CHEST (PORTABLE AP)	7495
25	CHEST (PA & LAT)	519
26	BILAT LOWER EXT VEINS	3050
26	P BILAT LOWER EXT VEINS PORT	1203
26	VEN DUP EXTEXT BIL (MAP/DVT)	1154
27	FDG TUMOR IMAGING (PET-CT)	1186
27	BONE SCAN	755

27	PERSANTINE MIBI	752
28	PICC W/O PORT	2442
28	PARACENTESIS DIAG. OR THERAPEUTIC	1391
28	TUNNELED W/O PORT	736
29	CHEST (PA & LAT)	24695
29	CHEST (PRE-OP PA & LAT)	6460
29	CHEST (LAT DECUB ONLY)	68

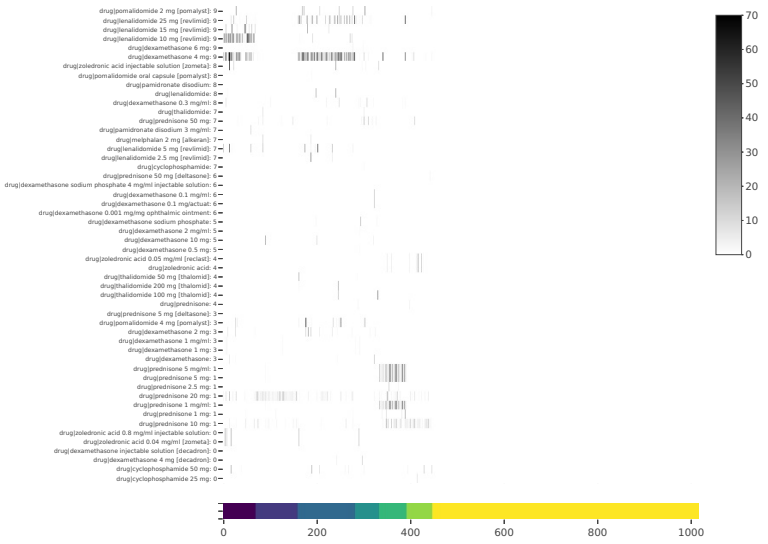


Figure 1: Bicluster graph of patients with multiple myeloma or MGUS diagnoses, based on prescriptions. Patients' cluster membership is indicated by color (x-axis), and prescriptions' cluster membership is indicated by number (y-axis).