BAYESIAN NONPARAMETRIC CLUSTERING WITH DIRICHLET PROCESS MIXTURE MODEL FOR MIXED-TYPE DATA

By

NURUL AFIQAH BINTI BURHANUDDIN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

January 2024

IPM 2024 6

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

BAYESIAN NONPARAMETRIC CLUSTERING WITH DIRICHLET PROCESS MIXTURE MODEL FOR MIXED-TYPE DATA

By

NURUL AFIQAH BINTI BURHANUDDIN

January 2024

Chair : Hani Syahida binti Zulkafli, PhD
Institute : Mathematical Research

Mixture models have been applied regularly by many researchers for clustering and density estimations. In particular, the Bayesian nonparametric mixture model involving the Dirichlet process prior has recently enjoyed popularity in clustering due to its flexibility, allowing the number of mixture components to grow infinitely. In this thesis, we aim to present some modifications of Bayesian nonparametric methods focusing on clustering mixed-type data, where the data comprises of continuous, ordinal, and nominal data.

Many studies have shown successful applications of the Dirichlet process mixture (DPM) model for clustering continuous data. However, the recent DPM model for clustering mixed-type data assumes a common covariance matrix across clusters, which is too restrictive in real practice. Accordingly, we develop a DPM model for clustering mixed-type data that allows for cluster-specific covariance matrices. To demonstrate the flexibility of our model, we compare it with the model with a common covariance matrix. Through this comparison, our model shows superior performance in terms of Normalized Mutual Information (NMI) in simulated

datasets with different cluster shapes and two real data applications. Our model also succeeds in estimating the true number of clusters in all cases as opposed to the model with a common covariance assumption that tends to overcluster the data.

When dealing with multivariate data, not all variables contribute towards cluster discrimination. To distinguish between relevant and irrelevant clustering variables, the DPM model for mixed-type data is further extended by specifying hierarchical shrinkage prior on the component means. This can be thought of as an implicit variable selection in clustering. The hierarchical shrinkage prior considered involves the normal-gamma prior for the continuous and ordinal data; while for nominal data, the grouped normal-gamma prior is used. The performances of the proposed model with shrinkage prior and without shrinkage prior are then compared. The comparison shows that the model with shrinkage prior achieves better clustering performance with higher NMI value, especially in simulated datasets with highly overlapping clusters and real datasets. Throughout the comparison, the model with shrinkage prior also produces a tighter clustering output measured in the form of silhouette width. Furthermore, the proposed model also successfully distinguishes relevant variables from noisy ones, as reflected by higher NMI value observed when the model is fitted with only the relevant variables.

The standard DPM model is introduced to address unsupervised learning problems where the data is analyzed without any background knowledge. To consider this extra knowledge in the clustering process, we develop a constrained DPM model that can incorporate labels as side information. These labels are considered in our formulation through a product partition prior that gives clusters of observations with similar labels a higher prior preference. The formulation is further extended to handle multiple side information. The empirical results on

several simulated and real datasets show that our model consistently improves its clustering performance in terms of NMI value as more labeled data become available. Even in the presence of noisy labels, the proposed model rarely performs worse than the standard unsupervised model, especially on continuous datasets. In multiple side information experiments, consistent increments in NMI value are also observed with access to more side information.

Keywords: Bayesian nonparametric, clustering, Dirichlet process, mixture model, model-based clustering

SDG: GOAL 9: Industry, Innovation and Infrastructure

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PENGELOMPOKAN BAYESIAN TAK PARAMETRIK DENGAN MODEL CAMPURAN PROSES DIRICHLET UNTUK DATA JENIS GABUNGAN

Oleh

NURUL AFIQAH BINTI BURHANUDDIN

Januari 2024

Pengerusi : Hani Syahida binti Zulkafli, PhD
Institut : Penyelidikan Matematik

Model campuran sering digunakan oleh para penyelidik dalam pengelompokan dan anggaran ketumpatan. Khususnya, model campuran Bayesian tak parametrik yang melibatkan prior proses Dirichlet sangat popular dalam pengelompokan kerana keanjalannya yang membolehkan bilangan komponen campuran bertambah tanpa had. Dalam tesis ini, kami bermatlamat untuk membentangkan beberapa pengubahsuaian kaedah Bayesian tak parametrik yang memfokuskan pada pengelompokan data data jenis gabungan. Di sini, data jenis gabungan merujuk kepada data yang terdiri daripada gabungan data selanjar, data ordinal, dan data nominal.

Banyak kajian berjaya menunjukkan aplikasi model campuran proses Dirichlet (DPM) bagi pengelompokan data selanjar. Walau bagaimanapun, model DPM terkini untuk pengelompokan data jenis gabungan mengandaikan matriks kovarians yang sama untuk semua kelompok, di mana andaian ini terlalu ketat dalam amalan aplikasi sebenar. Sehubungan dengan itu, kami membangunkan model DPM untuk data jenis gabungan yang membenarkan matriks kovarians

berbeza untuk setiap kelompok. Untuk menunjukkan keanjalan model kami, kami membandingkannya dengan model yang mengandaikan matriks kovarians sama. Melalui perbandingan ini, model kami menunjukkan prestasi unggul dari segi Maklumat Bersama Ternormal (NMI) dalam set data simulasi dengan bentuk kelompok yang berbeza dan dua aplikasi data sebenar. Model kami juga berjaya menganggar bilangan kluster sebenar dalam semua kes berbanding model dengan andaian kovarians sama yang cenderung mengelompok data secara berlebihan.

Apabila berurusan dengan data multivariat, tidak semua pembolehubah menyumbang ke arah diskriminasi kelompok. Untuk membezakan antara pembolehubah pengelompokan yang relevan dan tidak relevan, model DPM untuk data jenis gabungan dilanjutkan lagi dengan penggunaan prior kecutan hierarki pada min komponen. Ini boleh dianggap sebagai pemilihan pembolehubah tersirat dalam pengelompokan. Prior kecutan hierarki yang dipertimbangkan melibatkan prior normal-gamma untuk data selanjar dan ordinal; manakala untuk data ordinal, prior normal-gamma berkumpulan pula digunakan. Prestasi model yang dicadangkan dengan prior kecutan dan tanpa prior kecutan kemudian dibandingkan. Perbandingan menunjukkan bahawa model dengan prior kecutan mencapai prestasi pengelompokan yang lebih baik dengan nilai NMI lebih tinggi, terutamanya dalam set data simulasi dengan kelompok yang sangat bertindih dan juga set data sebenar. Sepanjang perbandingan, model dengan prior kecutan juga menghasilkan output pengelompokan yang lebih ketat diukur dalam bentuk lebar bayang. Tambahan pula, model yang dicadangkan juga berjaya membezakan pembolehubah yang relevan daripada yang hingar, seperti ditunjukkan oleh nilai NMI yang lebih tinggi diperhatikan apabila model ini disuai dengan hanya pembolehubah yang relevan tersebut.

Model piawai DPM diperkenalkan untuk menangani masalah pembelajaran tanpa pengawasan di mana data dianalisis tanpa pengetahuan latar belakang. Untuk mempertimbangkan pengetahuan tambahan ini dalam proses pengelompokan, kami membangunkan model DPM terkekang yang boleh menggunakan data label sebagai maklumat sampingan. Label ini dipertimbangkan dalam rumusan kami menerusi prior produk partisi yang memberi keutamaan yang lebih tinggi kepada kelompok dengan label yang serupa. Rumusan yang sama juga diperluaskan bagi mengendalikan maklumat sampingan berganda. Keputusan empirikal pada beberapa set data simulasi dan sebenar menunjukkan bahawa model kami secara konsisten meningkatkan prestasi pengelompokannya dari segi nilai NMI apabila lebih banyak data berlabel tersedia. Walaupun terdapat label yang hingar, model yang dicadangkan jarang menunjukkan prestasi yang lebih buruk daripada model piawai tanpa pengawasan, terutamanya pada set data selanjar. Dalam eksperimen maklumat sampingan berganda, kenaikan konsisten dalam nilai NMI juga diperhatikan dengan lebih banyak capaian kepada maklumat sampingan.

Kata Kunci: Bayesian tak parametrik, pengelompokan, proses Dirichlet, model campuran, pengelompokan berasaskan model

SDG: *GOAL 9: Industry, Innovation and Infrastructure*

ACKNOWLEDGEMENTS

First and foremost, all praise to Allah S.W.T. the Almighty for giving me the blessings, chance, strength, and endurance to complete this program. Nothing can be achieved without the Almighty's will. May Allah's blessing go to the final Prophet Muhammad (peace be upon him), his family, and his companions.

I will be ever grateful to Dr. Mohd. Bakri Adam, who, even though he is no longer with us, continues to inspire us with his dedication throughout his life. I would like to thank Prof. Dr. Kamarulzaman Ibrahim for standing by me during my time of loss. I must express my gratitude to Dr. Hani Syahida Zulkafli for guidance and support throughout this program. I am also grateful to Assoc. Prof. Datin Norwati Mustapha for her effort as a supervisory committee member.

Words cannot express my deep love and gratitude to Abah, Mak, and Andak for their patience and presence through all the ups and downs of my studies. This would not be a proper gratitude if I did not acknowledge my siblings, especially Zainul and Ain, for lending me their laptops, which I would not be able to finish if they did not.

Completing this study would have been more difficult without the support provided by the members of the Department of Mathematical Sciences at UKM and the staff of the Institute of Mathematical Research at UPM. I cannot thank them enough for their assistance. Those people, the postgraduate students at UPM, who gave me a very much-required form of escape from my studies, also deserve my gratitude for helping keep things in perspective.

Last but not least, I am especially grateful to the Universiti Kebangsaan Malaysia and the Ministry of Higher Education for providing the funding that allowed me to undertake this program.

I certify that a Thesis Examination Committee has met on 15 January 2024 to conduct the final examination of Nurul Afiqah binti Burhanuddin on her thesis entitled "Bayesian Nonparametric Clustering with Dirichlet Process Mixture Model for Mixed-Type Data" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Leong Wah June, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Shamarina binti Shohaimi, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Fadzilah binti Md Ali, PhD
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Nazrina binti Aziz, PhD
Associate Professor
College of Arts and Sciences
Universiti Utara Malaysia Sintok
Malaysia
(External Examiner)

RUSMAWATI BINTI SAID, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 30 April 2024

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Hani Syahida binti Zulkafli, PhD
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Norwati binti Mustapha, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Kamarulzaman bin Ibrahim, PhD
Professor
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
(Member)

ZALILAH MOHD SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 16 May 2024

Declaration by Members of the Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: _____
Name of Chairman of
Supervisory Committee: Hani Syahida binti Zulkafli

Signature: _____
Name of Member of
Supervisory Committee: Norwati binti Mustapha

Signature: _____
Name of Member of
Supervisory Committee: Kamarulzaman bin Ibrahim

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CKD	chronic kidney disease
CRP	Chinese restaurant process
CS	compound symmetry
DP	Dirichlet process
DPGM	Dirichlet process Gaussian mixture
DPM	Dirichlet process mixture
ec-GM	Gaussian mixture with equivalence constraints
EM	Expectation Maximization
GIG	generalized inverse Gaussian
GM	Gaussian mixture
lc-DPM	label-constrained Dirichlet process mixture
MCMC	Markov chain Monte Carlo
NMI	Normalized Mutual Information
non-CKD	non-chronic kidney disease
PCA	principal component analysis
pl-Kmeans	K-means with partition level side information
PP	product partition
SW	silhouette width

CHAPTER 1

INTRODUCTION

1.1 Background

We live in the era of big data. With the Internet, information can be obtained almost instantly. This has changed the dynamic of data acquisition; we expect to see, for example, a greater variety of topics as we read more news online, a greater variety of tags as we view more images online, and more debates as we examine individuals engaging in a social network. As more data are created and collected daily, the demand for effective data analysis tools continues to rise. Clustering is one of the common data analysis tools to identify natural grouping in a dataset. There are many different clustering methods available, such as hierarchical-based, centroid-based, density-based, and model-based. These methods differ in their underlying methodology in defining the target cluster. Hierarchical clustering builds a series of clustering outputs through a merger or division strategy based on some criteria function. Since each clustering output is irrevisable, the clustering results depend heavily on the choice of the criteria function (Murtagh, 1983). Centroid-based clustering, such as K-means (Hartigan and Wong, 1979) and K-medoid (Kaufman and Rousseeuw, 2009), groups the data based on their proximity to the cluster center called centroid. In contrast to hierarchical clustering, the centroid-based clustering method produces only one clustering output that is optimized iteratively. The density-based clustering, such as DBSCAN (Ester et al., 1996), characterizes clusters as dense regions in the sample space, separated by regions of lower density. Although density-based clustering is effective in identifying clusters of arbitrary shapes, interpretability suffers as a result. Overall, all these three clustering methods are purely heuristic without any underlying formal model.

Early on, it was discovered that the clustering method could also be built based on a statistical framework (Bock, 1996). Such clustering methods are usually based on a finite mixture model, referred to as the model-based clustering approach. In this approach, the data is assumed to arise from a mixture of distributions. Suppose we wish to partition $\mathbf{x} = (x_1, \ldots, x_N)$ into K clusters. This task can be formulated through the following model:

$$p(x_i; \boldsymbol{\theta}, \mathbf{w}) = \sum_{k=1}^{K} w_k p_k(x_i; \theta_k)$$

where $p_k(x_i; \theta_k)$ is the parametric density function of observation x_i from the kth component, θ_k is the parameter that characterized the density function, and w_ks are the component weights with $\sum_{k=1}^{K} w_k = 1$. This formulation requires the number of components K to be known in advance. Since each component is associated to one cluster, K is used interchangeably as the number of clusters. The use of the mixture model is gaining popularity in cluster analysis primarily due to the fact that it allows us to leverage standard statistical tools in assessing and advancing the clustering method. Moreover, some of the most widely used heuristic clustering methods have been proven to be approximate estimations of some statistical models. For instance, the standard K-means can be seen as a special case of the standard Gaussian mixture (GM) model with fixed mixing proportions and covariance matrices (Neal and Hinton, 1998).

Model-based clustering has been applied successfully in a wide area of applications, including population structure (Pritchard et al., 2000), genetics (McLachlan et al., 2002), computer vision (Lee, 2005), and econometrics (Frühwirth-Schnatter and Kaufmann, 2008), just to mention a few. However, despite the success of model-based clustering, quite a few practical issues should be considered. In particular, in the applications where there is little information available on the exact number of clusters, which in this case is the K; therefore, the unknown K has to be estimated from the data. From a fully Bayesian perspective, prob-

ably the most naive approach is to consider the K as an unknown parameter by specifying a prior distribution on it. Here, the term "fully" indicates that all the model parameters are assumed to be random, including the K. For this approach, several inference algorithms have been put forth, many of them are built upon the reversible jump Markov chain Monte Carlo (MCMC) proposed by Richardson and Green (1997). Reversible jump MCMC creates a Markov chain that jumps between mixture models of different K based on a proposal density. This proposal density can be challenging to construct, especially in a multivariate setting (Dellaportas and Papageorgiou, 2006). While there are advances in generalizing the construction of an effective proposal density for the Reversible jump MCMC (Brooks et al., 2003; Hastie and Green, 2012), these increase the complexity of the inference algorithms. Alternatively, Nobile (2004) proposed estimating the K based on the model marginal likelihood. Nonetheless, the marginal likelihood computation turns out to be demanding even for a moderate value of K (Frühwirth-Schnatter, 2004).

With the existence of a relatively simple MCMC (Escobar and West, 1995; MacEachern and Müller, 1998; Neal, 2000), it motivates us to turn to the Bayesian nonparametric approach in this thesis. The Bayesian nonparametric clustering allows the number K of the mixture model to grow infinitely, thus addressing the issue of an unknown number of clusters. This framework replaces the corresponding finite dimensional prior distribution of classical Bayesian analysis with infinite dimensional stochastic processes. Moreover, this model can be considered as an extension of the Bayesian finite mixture model with the advantage that it does not need any model selection to find the appropriate value of the K.

1.2 Scope of the Study

Motivated by such flexibility of the Bayesian nonparametric approach, this thesis focuses on using this approach to the mixture model in a clustering context. Much of our work builds upon a mixture model of the form:

$$\begin{aligned} x_i &\sim F = \int p(x_i|\theta_i) dG(\theta_i)\,, \\ \theta_i|G &\sim G\,, \\ G &\sim \mathcal{P}\,, \end{aligned}$$

where $p(x_i|\theta_i)$ denotes a probability density function parameterized by a random variable θ_i. This density function constitutes the kernel of the mixture model, and G acts as a mixing measure following some nonparametric prior $\mathcal{P}$. There have been many nonparametric priors available in the literature, such as the Dirichlet process, Gaussian process, Pólya tree process, and beta process. However, in this study, we restrict ourselves to the Dirichlet process (DP) prior, which plays a key role in the Bayesian nonparametric mixtures. The DP was introduced by Ferguson (1973) to serve as a prior over the space of discrete probability measures. The discreteness of the DP makes it suitable to be used as a mixing measure in mixture modeling. Basically, the role of the DP is to tie together the observations that share the same support, thus forming the different components of the mixture model. The mixture model that uses DP as the mixing measure is referred to as the Dirichlet process mixture (DPM). For Bayesian nonparametric models with other priors, refer to the following papers: Griffiths and Ghahramani (2011), Lijoi et al. (2005), Ishwaran and James (2001), and Pitman and Yor (1997).

In addition, since we are focusing on the clustering problem, it is stressed that the primary purpose of all the empirical examples in this thesis is to find the clustering classification rather than to obtain precise point estimates of parameters as in

density estimation.

1.3 Problem Statement

Due to the abundance of available data, researchers have encountered increasingly complex data structures in an attempt to describe or explain some real world events. This raises many new and exciting challenges in cluster analysis.

The standard DPM model is limited to handling a single type of data at a time. In particular, the Gaussian distribution is usually adopted as the mixture kernel to handle continuous data, and the multinomial distribution to handle nominal data. In real applications, we frequently encountered mixed-type data. The mixed-type data is defined as a set of data that contains several different types of variables. For instance, we are often asked for gender and age when filling out a survey. These two pieces of information already constitute two different types of data. However, clustering approaches for handling mixed-type data are less studied in the literature, let alone its extension. Moreover, most clustering approaches and their extensions are exclusively constructed to handle only one type of variable; see Vouros and Vasilaki (2021), Sarkar et al. (2020), and Peralta et al. (2016) among others. Furthermore, the current method of DPM for clustering mixed-type data has a very restrictive assumption in terms of the cluster shape (Carmona et al., 2019). This restriction leads us to develop a more flexible model for clustering mixed-type data.

The next problem that needs to be addressed is that not all variables are relevant and contribute towards cluster discrimination. The inclusion of these irrelevant variables could obscure the true cluster structure. This is especially true in multivariate settings when there are so many variables involved. In many application domains, some variables that make up the clusters are perceived to bring about

more information than others. For example, suppose in a medical diagnosis, the practitioner is clustering tissue samples into multiple cancer types. At the same time, the practitioner is also interested in isolating which genes give rise to cancerous cell behavior. This process of identifying subsets of variables that are crucial in distinguishing a cluster structure is called the variable selection process. Having a clustering model that can also identify both the cluster structure available in the data as well as the variables that contribute to this particular structure is beneficial. Some works on variable selection in clustering include Fu et al. (2021), Prakash and Singh (2019), and Marbac and Sedki (2017). Nevertheless, these methods are only applicable to continuous data. This motivates us to explore variable selection in clustering mixed-type data.

In addition, the standard DPM is initially introduced to address an unsupervised learning problem where the data is analyzed with nothing known about the true cluster structure. However, in many practical applications, one often performs cluster analysis with a rough idea of how the cluster structure should be. For instance, a few hundred billion emails are estimated to be sent and received daily. Suppose we have access to the database of emails, with some already being classified as "spam" or "non-spam". In this case, clustering can be used to classify the emails such that one group is predominantly made up of "spam" and another group is primarily made up of "non-spam". Then, to fully leverage all the available information, the pre-classified emails can be used to guide the clustering algorithm to make a more accurate grouping. This grey area between having absolutely no knowledge and having some information has encouraged the researcher to extend the current clustering algorithm to a case called constrained clustering; see, for example, Covoes et al. (2013) and Basu et al. (2008). In this thesis, we are interested in extending the DPM model to a constrained case for continuous and mixed-type data by considering labeled data as side information.

1.4 Objectives

Indeed, the present challenges give us many opportunities to devise new tools for capturing the hidden patterns in the data. Therefore, to address these challenges, we aim to provide a few contributions to the current DPM model. More specifically, the objectives of our study are:

1. To develop a DPM model for clustering mixed-type data based on latent variables approach.
2. To develop a method for variable selection in clustering mixed-type data using shrinkage prior.
3. To develop a constrained DPM model for clustering mixed-type data that can incorporate labeled data as side information.

Model verification of all the proposed methods in this study is done by simulation studies. In addition, the applicability of each method is also illustrated through the usage of publicly available datasets.

1.5 Outline of the Thesis

We organize the rest of the thesis as follows.

In Chapter 2, we start with a brief introduction to cluster analysis, finite mixture model, and infinite mixture model. Then, we briefly review the technical concepts and discuss the necessary tools that lay a foundation for our study. This is followed by a literature review discussing important works related to this study.

Starting from Chapter 3, we begin with our main contributions. In this chapter, we propose a DPM model that simultaneously handles mixed continuous, ordinal, and nominal data. Using the latent variables approach, we describe how the ordinal and nominal data are incorporated into the Gaussian kernel. The restriction on the choice of prior is also discussed.

Chapter 4 explores the use of shrinkage prior on the component means in a mixed-type data setting., which allows for implicit variable selection. The comparison between DPM with and without shrinkage prior is also illustrated through simulated and real datasets. In addition, the validity of the variables selected is also investigated.

In Chapter 5, we propose a constrained version of the DPM model that takes into account the availability of labeled information. To be specific, the labeled data is integrated via a product partition prior, where the relationship between the product partition model and DPM is also described. The formulation is also extended to accommodate multiple side information. In the first part of the chapter, we focus on the conjugate Gaussian kernel case. We then also present a variation of the model in a mixed-type data setting.

Finally, Chapter 6 provides the concluding remarks and the significant contributions of this study. Suggestions for future works related to this study are also presented in this chapter.

REFERENCES

Abin, A. A. (2016). Clustering with side information: Further efforts to improve efficiency. *Pattern Recognition Letters*, 84:252–258.

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281. Akademiai Kiado.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Almasoud, M. and Ward, T. E. (2019). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Soft Computing and Its Applications*, 10(8):89–96.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821.

Basu, S., Davidson, I., and Wagstaff, K. (2008). *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press.

Bilenko, M., Basu, S., and Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pages 11–18. Association for Computing Machinery.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355.

Bock, H. H. (1996). Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis*, 23(1):5–28.

Brooks, S. P., Giudici, P., and Roberts, G. O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):3–39.

Browne, R. P. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference*, 142(11):2976–2984.

Brusco, M. J. and Cradit, J. D. (2001). A variable-selection heuristic for K-means clustering. *Psychometrika*, 66(2):249–270.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83(2):275–285.

Byar, D. P. and Green, S. B. (1980). The choice of treatment for cancer patients based on covariate information. *Bulletin du Cancer*, 67(4):477–490.

Carmona, C., Nieto-Barajas, L., and Canale, A. (2019). Model-based approach for household clustering with mixed scale variables. *Advances in Data Analysis and Classification*, 13(2):559–583.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793.

Charikar, M., Guruswami, V., and Wirth, A. (2005). Clustering with qualitative information. *Journal of Computer and System Sciences*, 71(3):360–383.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

Covoes, T. F., Hruschka, E. R., and Ghosh, J. (2013). A study of K-means-based algorithms for constrained clustering. *Intelligent Data Analysis*, 17(3):485–505.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441):294–302.

Davidson, I. (2012). Two approaches to understanding when constraints help clustering. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1312–1320.

Davidson, I. and Ravi, S. (2005). Clustering with constraints: Feasibility issues and the K-means algorithm. In *Proceedings of the Fifth SIAM International Conference on Data Mining*, pages 138–149. Society for Industrial and Applied Mathematics.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Dellaportas, P. and Papageorgiou, I. (2006). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing*, 16(1):57–68.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–22.

DeYoreo, M. and Kottas, A. (2018). Bayesian nonparametric modeling for multivariate ordinal regression. *Journal of Computational and Graphical Statistics*, 27(1):71–84.

DeYoreo, M., Reiter, J. P., and Hillygus, D. S. (2017). Bayesian mixture models with focused clustering for mixed ordinal and nominal data. *Bayesian Analysis*, 12(3):679–703.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2):363–375.

Dua, D. and Graff, C. (2019). UCI machine learning repository.

Dubes, R. C. and Jain, A. K. (1993). Random field models in image analysis. *Journal of Applied Statistics*, 20(5-6):121–154.

Dy, J. G. and Brodley, C. E. (2004). Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5(8):845–889.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press.

Evans, M., Guttman, I., and Olkin, I. (1992). Numerical aspects in estimating the parameters of a mixture of normal distributions. *Journal of Computational and Graphical Statistics*, 1(4):351–365.

Everitt, B. S. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, 6(5):305–309.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, pages 287–302. Elsevier.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, 7(1):143–167.

Frühwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26(1):78–89.

Fu, Y., Liu, X., Sarkar, S., and Wu, T. (2021). Gaussian mixture model with feature selection: An embedded approach. *Computers & Industrial Engineering*, 152:107000.

Gelman, A. and King, G. (1990). Estimating the electoral consequences of legislative redistricting. *Journal of the American Statistical Association*, 85(410):274–282.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient metropolis jumping rules. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 5*, pages 599–608. Oxford University Press.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2):339–373.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

Gondek, D. and Hofmann, T. (2007). Non-redundant data clustering. *Knowledge and Information Systems*, 12(1):1–24.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Griffiths, T. L. and Ghahramani, Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12(4):1185–1224.

Gupta, E., Kunjal, R., and Cury, J. D. (2015). Severe hyponatremia due to valproic acid toxicity. *Journal of Clinical Medicine Research*, 7(9):717–719.

Hart, H., Chantiluke, K., Cubillo, A. I., Smith, A. B., Simmons, A., Brammer, M. J., Marquand, A. F., and Rubia, K. (2014). Pattern classification of response inhibition in ADHD: Toward the development of neurobiological markers for ADHD. *Human Brain Mapping*, 35(7):3083–3094.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108.

Hastie, D. I. and Green, P. J. (2012). Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*, 66(3):309–338.

Hunt, L. and Jorgensen, M. (1999). Theory & methods: Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, 41(2):154–171.

Hunt, L. and Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information. *Computational Statistics & Data Analysis*, 41(3-4):429–440.

Hunt, L. and Jorgensen, M. (2011). Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):352–361.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323.

Jerlin Rubini, L. and Perumal, E. (2020). Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm. *International Journal of Imaging Systems and Technology*, 30(3):660–673.

Johnson, V. E. and Albert, J. H. (2006). *Ordinal data modeling*. Springer Science & Business Media.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4):877–893.

Kottas, A., Müller, P., and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, 14(3):610–625.

Krzanowski, W. (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification*, 10(1):25–49.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In Stouffer, S. A., Guttman, L., Suchman, E. A., and Lazarsfeld, P. F., editors, *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*, pages 362–412. Princeton University Press.

Lee, D.-S. (2005). Effective Gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 7(5):827–832.

Lee, H. and Li, J. (2012). Variable selection for clustering by separability based on ridgelines. *Journal of Computational and Graphical Statistics*, 21(2):315–336.

Lelis, L. and Sander, J. (2009). Semi-supervised density-based clustering. In *2009 Ninth IEEE International Conference on Data Mining*, pages 842–847.

Li, C., Rana, S., Phung, D., and Venkatesh, S. (2016). Dirichlet process mixture models with pairwise constraints for data clustering. *Annals of Data Science*, 3(2):205–223.

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291.

Liu, H., Tao, Z., and Fu, Y. (2018). Partition level constrained clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2469–2483.

Lu, Z. and Leen, T. K. (2005). Semi-supervised learning with penalized probabilistic clustering. In *Advances in Neural Information Processing Systems 17*, pages 849–856. MIT Press.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26:303–324.

Marbac, M. and Sedki, M. (2017). Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*, 27:1049–1063.

Marbac, M., Sedki, M., and Patin, T. (2020). Variable selection for mixed data clustering: Application in human population genomics. *Journal of Classification*, 37:124–142.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009). Variable selection for clustering with Gaussian mixture models. *Biometrics*, 65(3):701–709.

McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1):207–240.

McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173–193.

McLachlan, G. J., Bean, R., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422.

McParland, D. and Gormley, I. C. (2016). Model based clustering for mixed data: clustMD. *Advances in Data Analysis and Classification*, 10(2):155–169.

McParland, D., Gormley, I. C., McCormick, T. H., Clark, S. J., Kabudula, C. W., and Collinson, M. A. (2014). Clustering South African households based on their asset status using latent variable models. *The Annals of Applied Statistics*, 8(2):747–776.

Morlini, I. (2012). A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model. *Advances in Data Analysis and Classification*, 6(1):5–28.

Moustaki, I. and Papageorgiou, I. (2005). Latent class models for mixed variables with applications in Archaeometry. *Computational Statistics & Data Analysis*, 48(3):659–675.

Müller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278.

Murray, J. S. and Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical statistics*, 9(2):249–265.

Neal, R. M. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Springer.

Nestler, S. and Hall, A. (2019). The variance gamma distribution. *Significance*, 16(5):10–11.

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *The Annals of Statistics*, 32(5):2044–2073.

Noh, H., Jo, Y., and Lee, S. (2015). Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Systems with Applications*, 42(9):4348–4360.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(5):1145–1164.

Papageorgiou, G., Richardson, S., and Best, N. (2015). Bayesian non-parametric models for spatially indexed data of mixed type. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(5):973–999.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, 185:71–110.

Pedrycz, W. (1985). Algorithms of fuzzy clustering with partial supervision. *Pattern Recognition Letters*, 3(1):13–20.

Pedrycz, W. and Waletzky, J. (1997). Fuzzy clustering with partial supervision. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 27(5):787–795.

Pelleg, D. and Baras, D. (2007). K-means with large and noisy constraint sets. In *Proceedings of the 18th European Conference on Machine Learning*, ECML '07, pages 674–682. Springer-Verlag.

Peralta, B., Caro, A., and Soto, A. (2016). A proposal for supervised clustering with Dirichlet process using labels. *Pattern Recognition Letters*, 80:52–57.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.

Prakash, J. and Singh, P. K. (2019). Gravitational search algorithm and K-means for simultaneous feature selection and data clustering: A multi-objective approach. *Soft Computing*, 23:2083–2100.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):557–574.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5(2):121–125.

Roshtkhari, M. J. and Levine, M. D. (2013). An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer Vision and Image Understanding*, 117(10):1436–1452.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sarkar, J. P., Saha, I., Chakraborty, S., and Maulik, U. (2020). Machine learning integrated credibilistic semi supervised clustering for categorical data. *Applied Soft Computing*, 86:105871.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650.

Shental, N., Bar-Hillel, A., Hertz, T., and Weinshall, D. (2003). Computing Gaussian mixture models with EM using equivalence constraints. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 465–472. MIT Press.

Storlie, C. B., Myers, S. M., Katusic, S. K., Weaver, A. L., Voigt, R. G., Croarkin, P. E., Stoeckel, R. E., and Port, J. D. (2018). Clustering and variable selection in the presence of mixed variable types and missing data. *Statistics in Medicine*, 37(19):2884–2899.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(12):583–617.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Tran, H. (2005). Extreme hyperkalemia. *Southern Medical Journal*, 98(7):729–733.

Van Hattum, P. and Hoijtink, H. (2009). Market segmentation using brand strategy research: Bayesian inference with respect to mixtures of log-linear models. *Journal of Classification*, 26(3):297–328.

Verdinelli, I. and Wasserman, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler. *Statistics and Computing*, 1(2):105–117.

Vlachos, A., Korhonen, A., and Ghahramani, Z. (2009). Unsupervised and constrained Dirichlet process mixture models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 74–82. Association for Computational Linguistics.

Vouros, A. and Vasilaki, E. (2021). A semi-supervised sparse K-means algorithm. *Pattern Recognition Letters*, 142:65–71.

Wade, S. (2023). Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2247):20220149.

Wade, S. and Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*, 13(2):559–626.

Wagstaff, K. and Cardie, C. (2000). Clustering with instance-level constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 1103–1110. Morgan Kaufmann Publishers Inc.

Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. (2001). Constrained K-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 577–584. Morgan Kaufmann Publishers Inc.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics*, 64(2):440–448.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

West, M. and Escobar, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *A tribute to D. V. Lindley, eds.* New York: Wiley.

Willse, A. and Boik, R. J. (1999). Identifiable finite mixtures of location models for clustering mixed-mode data. *Statistics and Computing*, 9(2):111–121.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5(3):329–350.

Xie, B., Pan, W., and Shen, X. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics*, 2:168–212.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.

Yau, C. and Holmes, C. (2011). Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Analysis*, 6(2):329–351.

Yi, J., Jin, R., Jain, A., and Jain, S. (2012). Crowdclustering with sparse pairwise labels: A matrix completion approach. In *The 4th AAAI Workshop on Human Computation*, AAAI Workshops. AAAI Press.

Zhang, X., Boscardin, W. J., and Belin, T. R. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, 15(4):880–896.

Zhang, X., Boscardin, W. J., Belin, T. R., Wan, X., He, Y., and Zhang, K. (2015). A Bayesian method for analyzing combinations of continuous, ordinal, and nominal categorical data with missing values. *Journal of Multivariate Analysis*, 135:43–58.

Zhou, D., Platt, J. C., Basu, S., and Mao, Y. (2012). Learning from the wisdom of crowds by minimax entropy. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, pages 2195–2203. Curran Associates Inc.