Balanced Population Stochastic Variational Inference

Cheng Zhang[*] **Stephan Mandt**[†] **Hedvig Kjellström**[*]

[*]RPL, KTH Royal Institute of Technology
Stockholm, Sweden
{chengz, hedvig}@kth.se

[†]Disney Research
Pittsburgh, PA, USA
stephan.mandt@disneyresearch.com

Abstract

Latent variable models are important tools to infer the underlying structure of a set of data. When we condition on observed data in Bayesian inference, we implicitly assume that the modeling assumptions are true and that the data can be considered a representative draw from the model. However, realistic data rarely agrees with these modeling assumptions. Especially when the observed data is highly imbalanced, inference will commonly result in redundant latent structures representing highly populated data clusters and miss the information contained in the scarce clusters. In this work, we propose a principled and scalable way to handle imbalanced data. We call our approach Balanced Population Stochastic Variational Inference (BP-SVI). Building on SVI, we use the Determinantal Point Process (DPP) to draw diversified mini-batches from the observed data set under the assumption that the observed data are an imbalanced realization of a true population. We show that this results in a more representative latent representation. From a theoretical side, our approach can be considered an instance of population Bayesian methods which were recently proposed for Bayesian inference on streams. While BP-SVI is applicable to a broad class of latent variable models, here we demonstrate how to use BP-SVI on Latent Dirichlet Allocation (LDA) as an example. Experiments are performed on both synthetic data and real-life news data showing clear improved performance on latent structure learning.

1 Introduction

Realistic data is commonly imbalanced [3]. For example, there are more images of cats and babies than other categories in social networks and there are more normal money transactions than fraudulent ones. Latent variable models that are used for tasks such as news/video summarization or fraud detection suffer from such imbalanced data. Aiming to maximize probability on the training data, they will infer redundant latent structures to capture the data clusters with high density but are not able to model the scarce data well that may contain possibly important information. For example, a latent space model may capture very fine differences between cats but might not even be able to model any feature of a reindeer. To extract more efficient latent representations of the underlying structure, we propose an approach based on biased subsampling of the original data set.

Nowadays with large amounts of data and complex probabilistic models, stochastic variational inference [4] with mini-batch subsampling is among the top choices for a broad range of applications. Sampling representative and balanced mini-batches is a desirable improvement and may result in more interpretable and less redundant latent structures, such as the topics in LDA [2].

Several related methods have been proposed to tackle the problem of model mismatch and imbalanced data from different perspectives, one example being diversity priors for latent variable models [8]. Yet, for large amounts of data, the effect of the prior will be drowned when not explicitly upweighting it relative to the likelihood. This, in turn, causes the new problem of balancing prior and data likelihood. Inspired by bootstrapping, population empirical Bayes [6] is proposed with sampling many bootstrapping populations and select the one with highest predictive likelihood. Local variational tempering or data re-weighting introduces per-datapoint temperatures or weights which are dynamically adjusted while training the model [9, 11], such that the resulting model can fit the

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

data better. This approach may lead to better topic representations for the highly populated clusters but has the opposite effect on the scarce data clusters which are further downweighted. Finally, we show that population posterior on streams [10] which introduces the population posterior helps us embed our inference scheme in a rigorous theoretical framework.

In this work, we propose a principled way to uncover the balanced unknown population for stochastic variational inference. The Determinantal Point Process (DPP) is used for mini-batch sampling (Section 2). We demonstrate the usage of our method using LDA with experiments using both synthetic data and real-world news data (Section 3).

2 Method

To tackle the inference problem with imbalanced data, we propose to sample the mini-batches with DPP. Thus, we firstly revisit DPP, then describe how it is employed for SVI with LDA as an example.

2.1 Determinantal Point Processes

The DPP [7] is a probabilistic model that models random subsets of a ground set with repulsive interactions between the elements. In this work, we use the DPP to create diversified mini-batches of the data that we then use for the SVI updates. For a good introduction of the DPP we recommend [7].

The DPP relies on a similarity matrix of the data K. In the example of LDA, we use the kernel $K(x_i, y_j) = x_i^{\rho} x_j'^{\rho}$ with TF-IDF feature vectors (x) of the documents. We use $\rho = 1$, which is linear kernel when the vocabulary size is small for example in the synthetic data experiment in Section 3 and use $\rho \leq 1$ in the real-world data to amplify the similarities in high-dimensional (large vocabulary) scenario. Let A denote a subset of the ground set, and K_A the submatrix of K constrained on that subset. The probability of this subset under the DPP is

$$\mathscr{P}(A) \propto det(K_A), \tag{1}$$

For instance, if $A = \{i, j\}$ consists of only two elements, then $\mathscr{P}(A) \propto K_{ii}K_{jj} - K_{ij}K_{ji}$. Because K_{ij} and K_{ji} measure the similarity between elements i and j, being similar lowers the probability of co-occurence. The more different the data are, the bigger the determinant.

In our application we need need to condition the DPP on the size S of the subsets (mini-batch size). This is also called the k-DPP which is extensively discussed in [7] (see section 5 and alg. 8 therein). Sampling from this process is $O(NS^3)$ overall, assuming we already have an eigendecomposition of K. Since the computational bottleneck in SVI are the local updates, linear scaling in N does not spoil the efficiency of SGD.

2.2 Balanced Population SVI and its application in LDA

The theoretical framework of BP-SVI can be described by the population posterior [10]. We assume that there is a balanced unknown population of datasets, and the observed data is an imbalanced realization from that population. Let θ denote latent variables of a Bayesian model. Suppose we sequentially observe S data points from the underlying population distribution as specified by the DPP, $\mathbf{X}_S \sim \text{DPP}(\mathbf{X})$. This is the mini-batch. Every $\mathbf{X}_S$ induces a posterior $p(\theta|\mathbf{X}_S)$, which is a function of the random data. Our posterior of interest is the *population posterior*,

$$p(\theta|\text{DPP}(\mathbf{X})) = \mathbb{E}_{\mathbf{X}_S \sim \text{DPP}(\mathbf{X})}[p(z, \beta|\mathbf{X}_S)], \tag{2}$$

which is the posterior averaged over many realizations of random data. This was originally invented for streaming settings; it mixes Bayesian and frequentist paradigms. It was shown in [10] that in the end, stochastic inference in this setup simply results in the conventional SVI updates [4], but where the mini-batches come from the data generating mechanism (the DPP in our case).

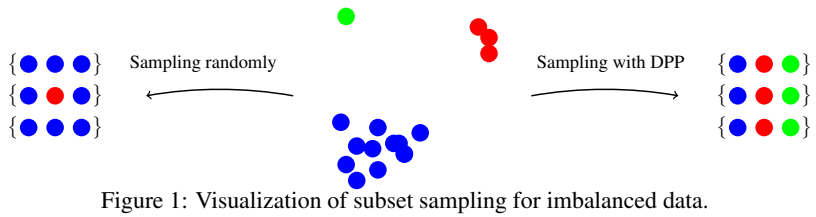


Figure 1: Visualization of subset sampling for imbalanced data.

During inference, we repetitively subsample mini-batches from the DPP to perform stochastic variational inference. Each time, we sample from the ground set which is in a similar manner as

bootstrapping, but which is biased towards diversification. This way, we effectively train our model on a more balanced population. This procedure is shown in Figure 1.

To demonstrate how to use BP-SVI for latent variable models, we use LDA [2] as an example which is directly comparable with SVI-LDA in [4]. Algorithms 1 shows the general usage of BP-SVI with LDA as an example. This is used in the experimental evaluation in Section 3. The detail of mini-batch sampling steps is shown in the appendix.

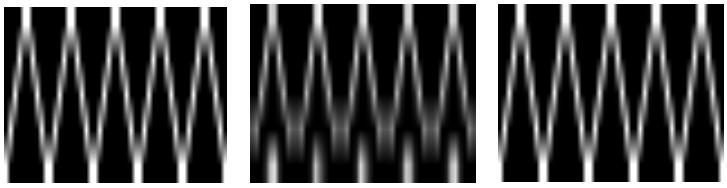
Algorithm 1 BP-SVI

Input: Mini-batch size S, eigendecomposition $\{(v_n, \lambda_n)\}_{n=1}^{N}$ of similarity matrix K.
for $t = 0$ *to* $MaxIter$ **do**
  Sampling mini-batch indices Y **with DPP**;
  Sampling S eigenvectors V with indices J using eigenvalues;
  Sampling S data points indexed by Y using V.
  Update variational parameters;
  Update local variational parameters (e.g. ϕ and λ for LDA) for mini-batch.
  Compute the intermediate global parameters as if the mini-batch is replicated $\frac{D}{S}$ times.
  (e.g. $\tilde{\lambda}_{kw} = \eta + \frac{D}{S}\sum_{s=1}^{S} n_{tw}\phi_{twk}$ for LDA)
  Update the current estimate of the global variational parameters with $\rho_t = (\tau_0 + t)^{-k}$.
  $\lambda = (1-\rho_t)\lambda + \rho_t\tilde{\lambda}$
end

3 Experiments

We first demonstrate BP-SVI on synthetic data with LDA. We show that by balancing our mini-batches, we find a much better recovery of the topics that were used to generate the data. Second, we use a real-world news dataset. We demonstrate that we can learn more diverse topics that are also better features for text classification tasks.

Synthetic data. We generate a synthetic dataset (shown in the appendix) following the generative process of LDA with a fixed global latent parameter (the graphical topics). We chose distinct patterns as shown in Figure 2 (a), where each row represents a topic and each column represents a word. To generate an imbalanced data set, we use different Dirichlet priors for the per document topic distribution θ. 300 documents are generated with prior (0.5 0.5 0.01 0.01 0.01); 50 with prior (0.01 0.5 0.5 0.5 0.01) and 10 with prior (0.01 0.01 0.01 0.5 0.5). Hence, the first two topics are used very often in the corpus. Topic 3 and 4 are shown a few times and topic 5 appears very rarely.

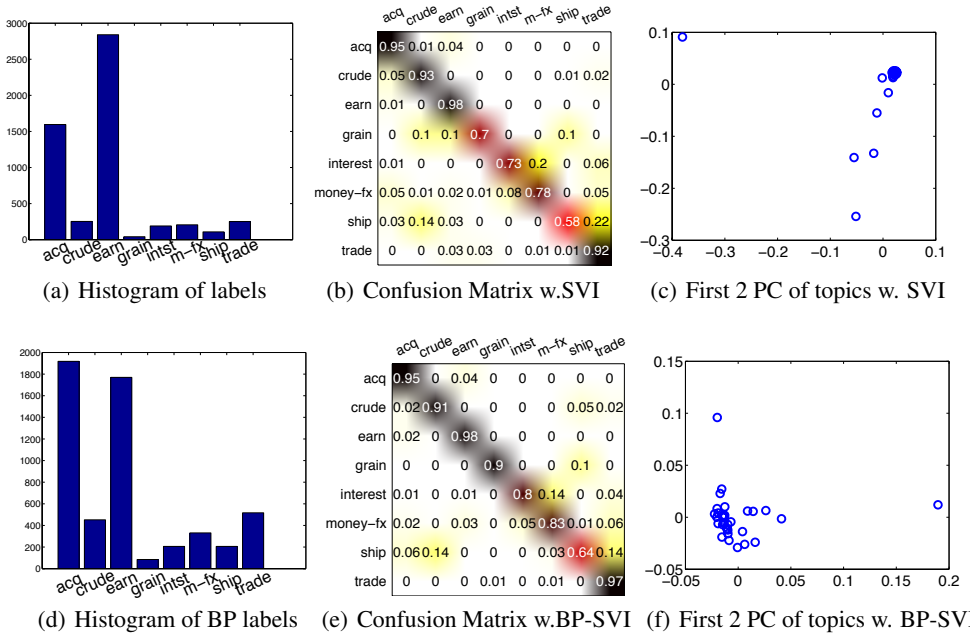


(a) Ground Truth (b) Est w. SVI LDA (c) Est w. BP-SVI

Figure 2: The above figures show the per topic word distribution for the synthetic data. Each row presents a topic and each column presents a word. (a) shows the ground truth with which the synthetic data is generated using LDA. (b) shows the estimation of this latent variable with LDA using traditional stochastic variational inference (SVI). (c) shows the estimation of this latent variable with BP-SVI

We apply LDA to recover the topics of the synthetic data using traditional SVI and our proposed BP-SVI respectively. The aim is to recover the ground truth global parameter which indicates that the model is able to capture is underlying structure of the data. Figure 2 (b) shows the estimated per topic words distribution with SVI and Figure 2 (c) shows the result with our proposed BP-SVI.

In Figure 2 (b), we see that the first three topics are recovered using traditional SVI. Topic four is roughly recovered but with information from topic two mixed in. The last topic were not recovered at all, instead, it is a repetition of the first topic. This shows the drawback of the tradition method that when the data is not balanced, the model create redundant topics to refine the likelihood of the dense data but ignore the scarce data even they carry important information. In Figure 2 (c), we see that all the topics are correctly recovered thanks to the balanced population.



(a) Histogram of labels (b) Confusion Matrix w.SVI (c) First 2 PC of topics w. SVI

(d) Histogram of BP labels (e) Confusion Matrix w.BP-SVI (f) First 2 PC of topics w. BP-SVI

Figure 3: The first row shows results from LDA with traditional SVI and the second row shows results with BP-SVI. The first column shows the histogram of class label of the original training data and with balanced population. The second column shows the confusion matrix when doing classification with SVI / BP-SVI features. With traditional SVI, the average performance over 8 classes is 82.11%, the total accuracy (num of correct classified docs over number of test documents) is 94.11%. With traditional SVI, the average performance over 8 classes is 87.24% and the total accuracy is 94.7%. The last column visualizes the per topic words distributions using the first 2 principal components from PCA.

News data. We also evaluated the effect of BP-SVI the Reuters news R8 dataset [1]. This dataset contains classes of news with an extremely imbalanced number of documents per class, as shown in Figure 3 (a). To measure similarities between documents, we defined an annealed linear kernel $K(x_i, y_j) = x_i^\rho y_j^\rho$ with inverse temperature $\rho = 0.1$ which is more sensitive to small feature overlap. We ran LDA with SVI and BP-SVI (proposed here) with one effective pass through the data, where we set the mini-batch size 80 and used 30 topics.

We first compared the frequencies at which documents with particular labels were subsampled. While Fig. (a) shows the actual frequency of these classes in the original data set, (d) shows the histogram of class labels over the balanced population (as subsampled using the DPP). We can see that the number of documents are more balanced among different classes. To demonstrate that BP-SVI leads to a more useful topic representation, we predicted the class labels for each document based on the learned topic proportions. We used the linear SVM. The resulting confusion matrices are shown in Figure 3 (b) and Figure 3 (e) using traditional SVI and BP-SVI respectively. The overall classification performance is improved using BP-SVI features, especially the performance on the classes with few documents (such as "grain" and "ship") is improved significantly.

We also visualized the first two principal components (PC) of the the global topics in Figure 3 (c) and (f) respectively. In traditional SVI, many topics are redundant and share large parts of their vocabulary, resulting in a single dense cluster. In contrast, we see that the topics in BP-SVI are more spread out. In this regard, BP-SVI achieves a similar effect as when using diversity priors as in [8] without the need to grow the prior with the data. The top words from each topics are shown in the appendix, where we present more evidence that the topics leaned by BP-SVI are more diverse.

4 Discussion

In this paper, we proposed a principled framework BP-SVI for imbalanced data with clear improvements in various experimental evaluations. In the future, we will explore the possibility to further improve the efficiency of the algorithm with data reweighing [9, 11] and tackle imbalance problems involving different modalities for supervised [12] and multi-modal [5] settings.

References

[1] R8 dataset. http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[4] M D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[5] A. Joulin, L. van der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, 2016.

[6] A. Kucukelbir and D. M. Blei. Population empirical bayes. In *Conference on Uncertainty in Artificial Intelligence*, 2015.

[7] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

[8] J. T. Kwok and R. P. Adams. Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems*, pages 2996–3004, 2012.

[9] S. Mandt, J. McInerney, F. Abrol, R. Ranganath, and Blei. Variational Tempering. In *International Conference on Artificial Intelligence and Statistics*, pages 704–712, 2016.

[10] J. McInerney, R. Ranganath, and D. M. Blei. The population posterior and bayesian modeling on streams. In *Advances in Neural Information Processing Systems*, pages 1153–1161, 2015.

[11] Y. X. Wang, A. Kucukelbir, and D. M. Blei. Reweighted Data for Robust Probabilistic Models. *arXiv preprint arXiv:1606.03860*, 2016.

[12] C. Zhang and H. Kjellström. How to Supervise Topic Models. In *European Conference on Computer Visionworkshop on Graphical Models in Computer Vision*, 2014.

A Appendix

Algorithm 2 shows the details of how to sample a mini-batch using DPP which is used for the BP-SVI algorithm in the paper.

Algorithm 2 Mini-batch Sampling

Input: Minibatch size S, eigendecomposition $\{(v_n, \lambda_n)\}_{n=1}^{N}$ of similarity matrix K.
Compute the elementary symmetric polynomials
$e_0^n \leftarrow 1 \forall n \in \{0, 1, 2, ..., N\}$
$e_0^l \leftarrow 1 \forall nl \in \{1, 2, ..., S\}$
for $l = 1, 2, ..., S$ **do**
  for $n = 1, 2, ..., N$ **do**
    $e_l^n \leftarrow e_l^{n-1} + \lambda_n e_{l-1}^{n-1}$
  end
end
Sampling S **eigenvectors** V **with indices** J
$J \leftarrow \emptyset$
$l \leftarrow S$
for $n = N, ..., 2, 1$ **do**
  if $l = 0$ **then**
    break;
  end
  if $u \sim U[0,1] \leq \lambda_n \frac{e_{l-1}^{n-1}}{e_l^n}$ **then**
    $J \leftarrow J \cup \{n\}$
    $l \leftarrow l - 1$
  end
end
$V \leftarrow \{v_i\}_{i \in J}$
$Y \leftarrow \emptyset$
while $|V| > 0$ **do**
  Select i with $Pr(i) = \frac{1}{|V|} \sum_{v \in V} (v^T e_i)^2$
  $Y \leftarrow Y \cup i$ $V \leftarrow V_\perp$, an orthonormal basis for the subspace of V orthogonal to e_i
end
Output: Y

Figure 4 shows the synthetic data that are used in the experiment. Each row represents a document and each column represents a word.

Table and show the top words using $K = 30$ for LDA using traditional SVI and our proposed BP-SVI respectively. We can see that the topics that learned by BP-SVI are more diverse and rare topic such as grain is captured with BP-SVI for LDA.

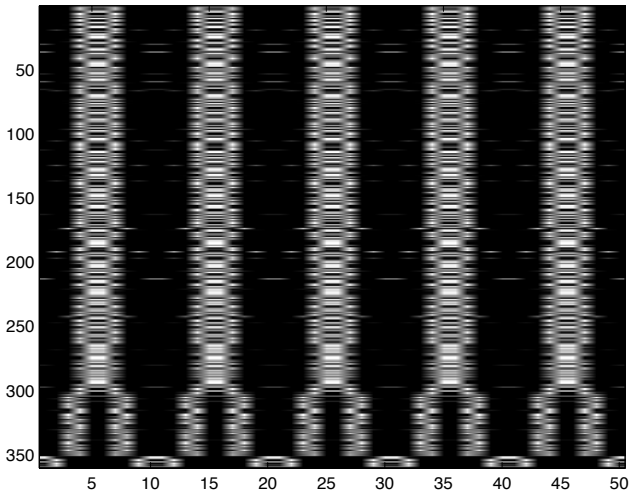


Figure 4: The synthetic data. Each row present a document and each column represent a word.

Topic 1	pct shares stake and group investment securities stock commission firm
Topic 2	year pct and for last lower growth debt profits company
Topic 3	and merger for will approval companies corp acquire into letter
Topic 4	and for canadian company management pacific bid southern court units
Topic 5	baker official and that treasury western policy administration study budget
Topic 6	and president for executive chief shares plc company chairman cyclops
Topic 7	bank pct banks rate rates money interest and reuter today
Topic 8	and unit inc sale sell reuter company systems corp terms
Topic 9	mln stg and reuter months year for plc market pretax
Topic 10	and national loan federal savings reuter association insurance estate real
Topic 11	trade and for bill not united imports that surplus south
Topic 12	and february for china january gulf issue month that last
Topic 13	market dollar that had and will exchange system currency west
Topic 14	dlrs quarter share for company earnings year per and fiscal
Topic 15	billion mln tax year profit credit marks francs net pct
Topic 16	usair inc twa reuter trust air department chemical diluted piedmont
Topic 17	and will union spokesman not two that reuter security port
Topic 18	offer share tender shares that general and gencorp dlrs not
Topic 19	and company for that board proposal group made directors proposed
Topic 20	that japan japanese and world industry government for told officials
Topic 21	american analysts and that analyst chrysler shearson express stock not
Topic 22	loss profit mln reuter cts net shr dlrs qtr year
Topic 23	mln dlrs and assets for dlr operations year charge reuter
Topic 24	mln net cts shr revs dlrs qtr year oper reuter
Topic 25	cts april reuter div pay prior record qtly march sets
Topic 26	dividend stock split for two reuter march payable record april
Topic 27	oil and prices crude for energy opec petroleum production bpd
Topic 28	agreement for development and years program technology reuter conditions agreed
Topic 29	and foreign that talks for international industrial exchange not since
Topic 30	corp inc acquisition will company common shares reuter stock purchase

Table 1: Top 10 words from topics learned from LDA with traditional SVI.

Topic 1	oil and that prices for petroleum dlrs energy crude field
Topic 2	pct and that rate market banks term rates this will
Topic 3	billion and pct mln group marks sales year capital rose
Topic 4	and saudi oil gulf that arabia december minister prices for
Topic 5	and dlrs debt for brazil southern mln will medical had
Topic 6	and grain that will futures for program farm certificates agriculture
Topic 7	bank banks rate and pct interest rates for foreign banking
Topic 8	and union for national seamen california port security that strike
Topic 9	and trade that for dollar deficit gatt not exports economic
Topic 10	and financial for sale inc services reuter systems agreement assets
Topic 11	dollar and for yen mark march that dealers sterling market
Topic 12	and for south unit equipment reuter two will state corp
Topic 13	and firm stock company will for pct not share that
Topic 14	and world that talks economic official for countries system monetary
Topic 15	and gencorp for offer general company partners that dlrs share
Topic 16	mln canada canadian stg and pct will air that royal
Topic 17	usair and twa that analysts not for pct analyst piedmont
Topic 18	and that for companies not years study this areas overseas
Topic 19	trade and bill for house that reagan foreign states committee
Topic 20	company dlrs offer stock and for corp share shares mln
Topic 21	dlrs year and quarter company for earnings will tax share
Topic 22	mln cts net loss dlrs profit reuter shr year qtr
Topic 23	exchange paris and rates that treasury baker allied for western
Topic 24	and shares inc for group dlrs pct offer reuter share
Topic 25	merger and that pacific texas hughes baker commerce for company
Topic 26	and american company subsidiary china french reuter pct for owned
Topic 27	japan japanese and that trade officials for government industry pact
Topic 28	oil opec mln bpd prices production ecuador and output crude
Topic 29	and that had shares block for mln government not san
Topic 30	mln pct and profits dlrs year for billion company will

Table 2: Top 10 words from topics learned from LDA with BP-SVI.