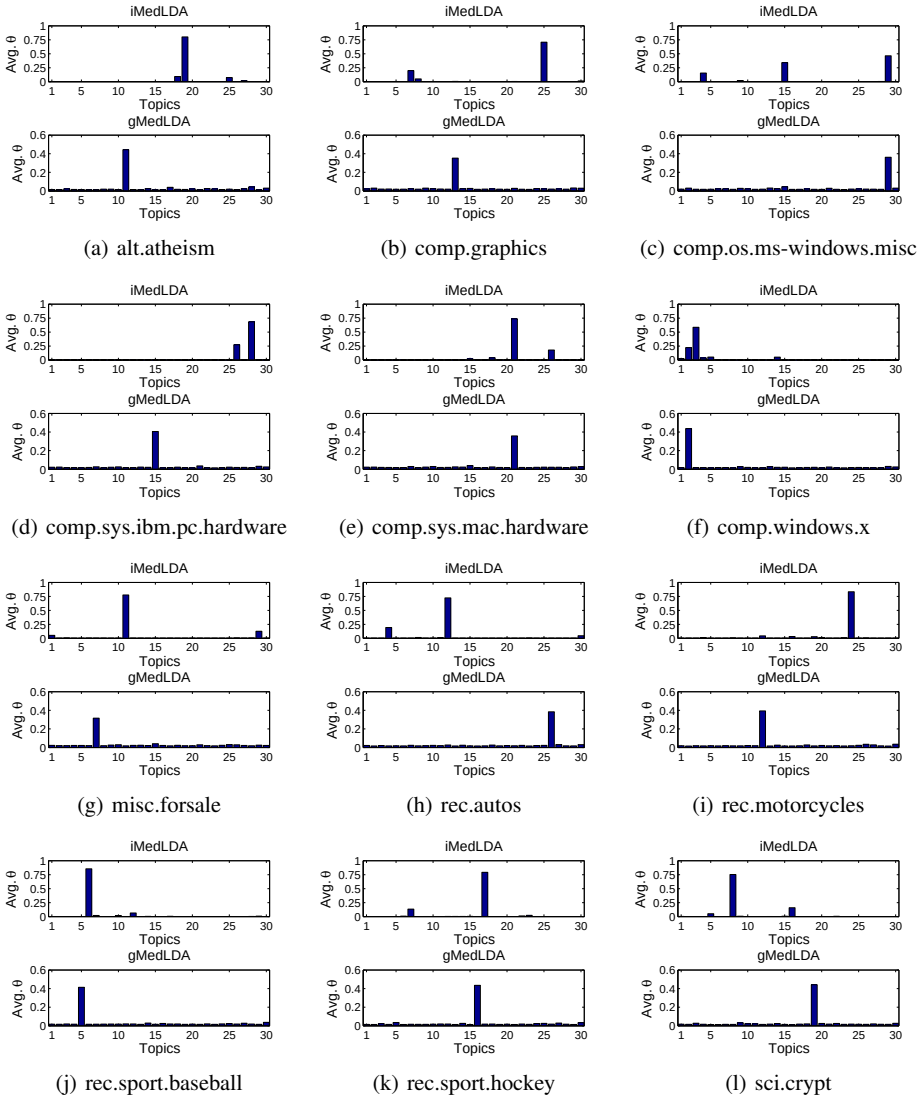
## Appendix A: Inferred latent topic space

To give an overview of the latent topic space discovered by our methods, we calculate the per-class average distribution over inferred topics for both iMedLDA and gMedLDA on the 20-Newsgroups data set. In this experiment, the topic number is set to be 30 for both models. The per-class distribution is computed by averaging the expected latent representations (i.e., $\theta$) of the documents in each class.

As shown in Figure 4 and Figure 5, both iMedLDA and gMedLDA can yield very sharp and sparse per-class distributions over topics. These sparse patterns are consistent with those reported in [16]. Moreover, for different categories, the per-class average topic representations are quite different, which suggests that the latent representations are good at distinguishing the documents from different categories.



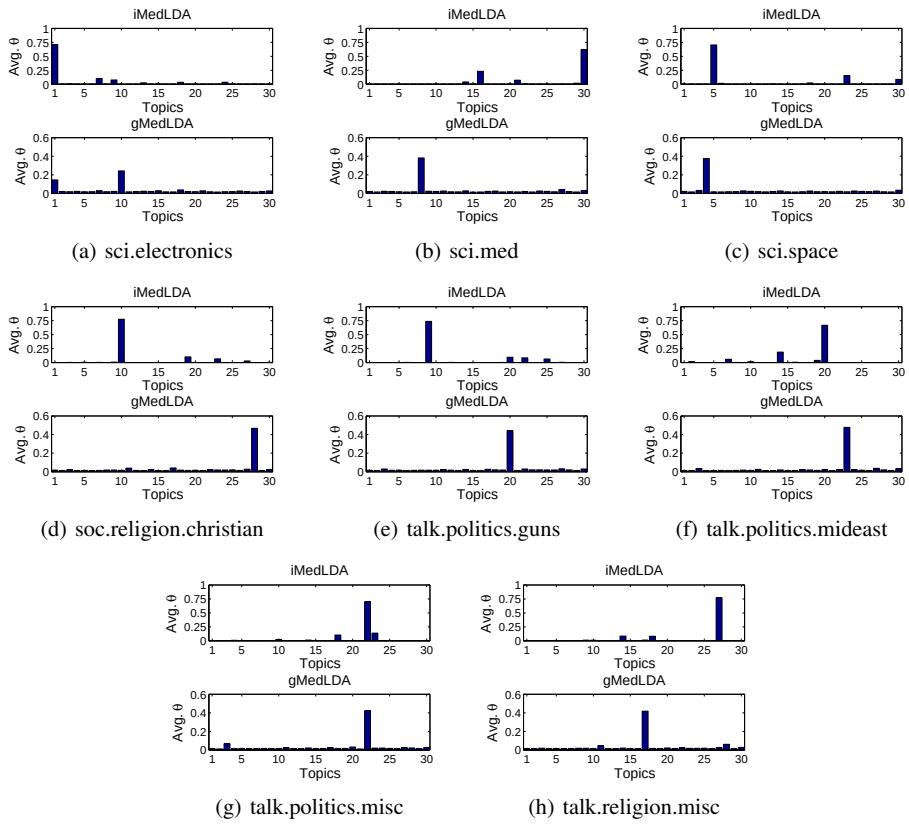

(a) alt.atheism (b) comp.graphics (c) comp.os.ms-windows.misc

(d) comp.sys.ibm.pc.hardware (e) comp.sys.mac.hardware (f) comp.windows.x

(g) misc.forsale (h) rec.autos (i) rec.motorcycles

(j) rec.sport.baseball (k) rec.sport.hockey (l) sci.crypt

Figure 4: Per-class distribution over topics for iMedLDA and gMedLDA methods on the 20-Newsgroups data set. (a)∼(l) the distribution of the 1st∼12th class respectively.




(a) sci.electronics (b) sci.med (c) sci.space

(d) soc.religion.christian (e) talk.politics.guns (f) talk.politics.mideast

(g) talk.politics.misc (h) talk.religion.misc

Figure 5: Per-class distribution over topics for the iMedLDA and gMedLDA methods on the 20-Newsgroups data set. (a)∼(h) the distribution of the 13th∼20th class respectively.

Finally, to illustrate the semantic meanings of the learned topics, we also report the ranked top-10 words in each of the 30 topics by both iMedLDA and gMedLDA in Table 1 and Table 2, respectively. Then, by examining Figure 4 and Figure 5 again, we can see the clear connections between class categories and the semantic meanings of the topics. For example, for the newsgroup "alt.atheism", iMedLDA uses the most salient topic "Topic 19" to describe the documents in that group, where "Topic 19" has the indicative top words "god", "religion" and "atheism", as shown in Table 1. For the same group, gMedLDA uses the most salient topic "Topic 11", which again has the similar indicative top words "god", "atheism", and "religion", as shown in Table 2. Note that due to the unidentifiability issue of topic models, we can't control the ordering of the topics learned by iMedLDA and gMedLDA.

## Appendix B: Binary classification

As in [16], binary classification is to distinguish the documents from the *alt.atheism* group and the documents from the *talk.religion.misc* group wit. We randomly sample 569 documents from such two groups as the test set and the rest 856 as the training set. All the parameters are set to be the same as in the multi-class classification experiments.

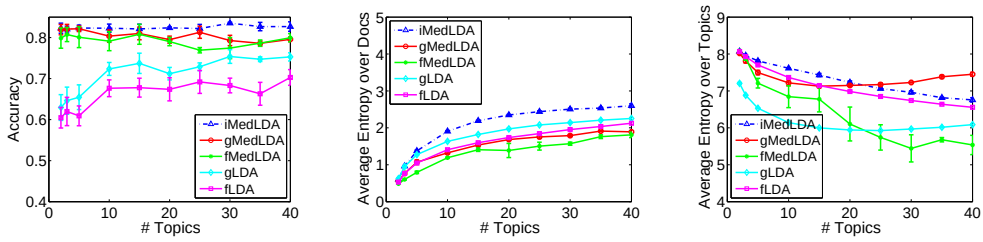
Fig.6(a) presents the binary classification accuracy of different models. As in the multi-class classification experiment, both MedLDA models using Monte Carlo approximation methods (i.e. iMedLDA and gMedLDA) can obtain the best classification accuracy, which owns to the fact that Monte Carlo methods for MedLDA impose weaker constrictions on the true posterior distributions than the variational methods.

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 |
|---|---|---|---|---|---|
| power | file | window | windows | space | year |
| good | window | server | don | nasa | team |
| don | program | file | car | launch | game |
| work | entry | motif | good | orbit | baseball |
| current | output | program | driver | gov | won |
| output | lib | widget | file | moon | don |
| circuit | widget | application | problem | earth | games |
| ground | number | mit | people | apr | runs |
| audio | line | sun | engine | shuttle | season |
| voltage | motif | display | cars | data | player |
| **Topic 7** | **Topic 8** | **Topic 9** | **Topic 10** | **Topic 11** | **Topic 12** |
| don | key | gun | god | sale | car |
| ca | encryption | people | jesus | offer | cars |
| time | chip | guns | people | shipping | don |
| apr | government | writes | church | mail | good |
| good | clipper | don | christ | price | engine |
| university | keys | article | christians | dos | apr |
| ve | system | weapons | christian | condition | time |
| center | writes | firearms | don | interested | year |
| points | security | fire | bible | sell | oil |
| ll | law | law | writes | email | speed |
| **Topic 13** | **Topic 14** | **Topic 15** | **Topic 16** | **Topic 17** | **Topic 18** |
| pub | people | db | key | team | people |
| file | don | windows | article | game | don |
| data | writes | file | information | hockey | time |
| ftp | didn | files | people | play | mr |
| anonymous | told | um | public | season | writes |
| contact | time | cs | don | ca | system |
| wire | ll | bh | writes | players | make |
| jpeg | work | di | privacy | nhl | article |
| archive | children | mov | time | writes | work |
| information | turkish | ei | number | games | ve |
| **Topic 19** | **Topic 20** | **Topic 21** | **Topic 22** | **Topic 23** | **Topic 24** |
| god | israel | mac | writes | people | bike |
| people | people | apple | people | space | dod |
| writes | turkish | writes | article | time | writes |
| don | israeli | problem | government | don | article |
| article | armenian | don | don | president | don |
| religion | jews | system | president | writes | ride |
| atheism | armenians | ve | mr | make | apr |
| evidence | writes | work | state | article | ca |
| atheists | government | drive | apr | government | good |
| time | article | lc | health | mr | motorcycle |
| **Topic 25** | **Topic 26** | **Topic 27** | **Topic 28** | **Topic 29** | **Topic 30** |
| graphics | drive | god | drive | windows | msg |
| image | scsi | jesus | scsi | dos | science |
| file | mb | people | mb | file | food |
| writes | card | writes | controller | card | time |
| files | drives | bible | card | pc | medical |
| software | memory | christian | bus | problem | years |
| bit | disk | don | system | system | disease |
| images | hard | life | ide | mail | patients |
| don | os | good | disk | program | good |
| color | system | christians | pc | mouse | health |

Table 1: The ten most probable words in the topics discovered by iMedLDA on the 20-Newsgroups data set.

| **Topic 1** | **Topic 2** | **Topic 3** | **Topic 4** | **Topic 5** | **Topic 6** |
|---|---|---|---|---|---|
| db | window | mr | space | year | ma |
| cs | file | president | nasa | game | pa |
| writes | program | people | launch | team | um |
| don | server | states | gov | baseball | em |
| si | motif | money | earth | games | ei |
| water | entry | stephanopoulos | moon | runs | el |
| article | sun | work | orbit | hit | di |
| mov | output | time | satellite | won | mu |
| work | widget | years | shuttle | players | mi |
| bh | set | american | data | season | de |
| **Topic 7** | **Topic 8** | **Topic 9** | **Topic 10** | **Topic 11** | **Topic 12** |
| price | msg | information | power | people | dod |
| sale | health | mail | current | god | bike |
| offer | food | list | company | writes | writes |
| mail | disease | send | radio | don | article |
| shipping | medical | internet | high | evidence | ride |
| dos | patients | faq | line | argument | motorcycle |
| interested | science | anonymous | phone | system | back |
| sell | people | email | audio | atheism | dog |
| condition | doctor | group | low | exist | riding |
| original | pitt | ftp | input | religion | bmw |
| **Topic 13** | **Topic 14** | **Topic 15** | **Topic 16** | **Topic 17** | **Topic 18** |
| image | don | drive | team | people | ground |
| graphics | good | scsi | game | god | point |
| software | ve | mb | hockey | jesus | time |
| images | make | card | play | writes | case |
| color | doesn | disk | season | christian | wire |
| file | real | system | games | world | work |
| article | current | security | control | mb | case |
| don | subject | public | fire | speed | care |
| word | run | law | state | hardware | free |
| program | difference | hard | la | bible | make |
| **Topic 19** | **Topic 20** | **Topic 21** | **Topic 22** | **Topic 23** | **Topic 24** |
| key | gun | mac | people | israel | university |
| encryption | people | apple | article | turkish | april |
| chip | guns | problem | writes | armenian | national |
| government | weapons | bit | government | jews | center |
| clipper | law | drive | state | people | research |
| keys | firearms | system | drugs | israeli | washington |
| armenians | san | data | ll | bus | nhl |
| war | number | version | give | pc | period |
| turkey | dr | files | de | controller | players |
| system | government | computer | don | turks | institute |
| **Topic 25** | **Topic 26** | **Topic 27** | **Topic 28** | **Topic 29** | **Topic 30** |
| mark | car | people | god | windows | writes |
| man | cars | didn | jesus | file | article |
| andrew | engine | time | people | files | apr |
| st | writes | don | church | dos | ca |
| thing | article | back | christians | win | cs |
| book | speed | told | bible | program | uiuc |
| appears | good | left | faith | driver | uk |
| day | oil | started | christian | mouse | org |
| black | driving | things | christ | card | news |
| cmu | dealer | home | truth | version | cc |

Table 2: The ten most probable words in the topics discovered by gMedLDA on the 20-Newsgroups data set.




(a) (b) (c)

Figure 6: Performance of binary classification of different topic models with different topic numbers (from 1 to 40) on *alt.atheism* group and *talk.religion.misc*: (a) classification accuracy, (b) the average entropy of $\boldsymbol{\Theta}$ over test documents, and (c) The average entropy of topic distributions $\boldsymbol{\Phi}$.

Fig.6(b) shows the average entropy of latent topic representations $\boldsymbol{\Theta}$ over test documents. We can see that fMedLDA yields the smallest entropy than all the other models, which is because fully-factorized variational methods tend to obtain too compact results. iMedLDA's entropy is the largest.

Fig.6(c) reports the average entropy of inferred topic distributions $\boldsymbol{\Phi}$. As the sampling method for LDA (i.e. gLDA) yields larger entropy than the variational method for LDA (i.e. fLDA), both MedLDA models' entropy using Monte Carlo sampling methods is larger than the variational MedLDA's.

## Appendix C: Distribution of training time

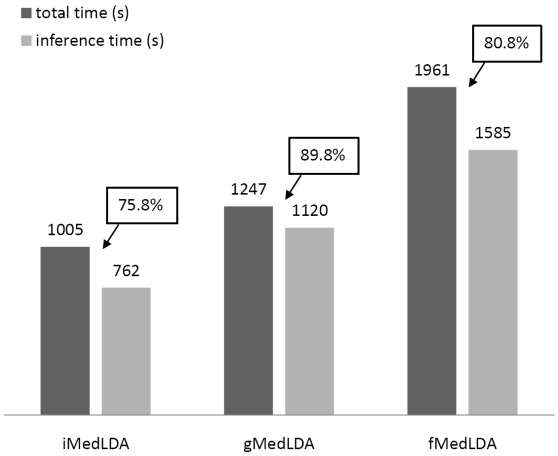



Figure 7: The total training time and the amount of time spent on the inference phase for different methods on the 20-Newsgroups data set when the topic number is 30.

We have presented the total training time in Figure 3, where the training includes two phases – inferring the latent topic representations and training SVMs. Now, we present a closer examination. Specifically, Figure 7 presents the total training time and the time (as well as the proportion) taken by posterior inference. Here, we have adopted the equivalent 1-slack formulation (i.e., with only one constraint and one slack variable) of the multi-class SVM as in Eq. (12), which is more efficient to solve than the original $n$-slack formulation as in Eq. (12). From the results, we can see that for all the MedLDA methods, most of the training time is spent on the inference.