Gamma Belief Networks

Mingyuan Zhou

IROM Department, McCombs School of Business
The University of Texas at Austin

Computational and Financial Econometrics (CFE 2015)
University of London, UK, December 12, 2015

Joint work with Yulai Cong and Bo Chen

Deep learning

- There is significant recent interest in deep learning due to its excellent performance in large-scale real applications, such as image classification and speech recognition.

- State-of-the-art results in supervised learning when the labeled data are abundant.

- Multilayer generative models for nonlinear distributed representations:
  - SBN, sigmoid belief network
  - DBN, deep belief network (a SBN whose last layer is replaced with a restricted Boltzmann machine that is undirected)
  - DBM, deep Boltzmann machine (a hierarchy of restricted Boltzmann machines)

Challenges in deep learning

- The hidden units are often restricted to be binary.
- Difficult to train a deep network in an unsupervised manner.
- A greedy layer-wise training strategy is often used due to the difficulty of jointly training all hidden layers.
- Lack of principled ways to determine the network structure, including the depth (number of layers) of the network and width (number of units) of each of its hidden layers.
- Commonly used deep learning models are not naturally designed for count data.

Our objectives

- Design a multilayer deep model that is well suited for extracting nonlinear distributed representations for high-dimensional sparse count, binary, and nonnegative real vectors.

- Construct the deep network using nonnegative real hidden units rather than using binary ones.

- Using nonparametric Bayesian priors to automatically infer the network structure from the data.

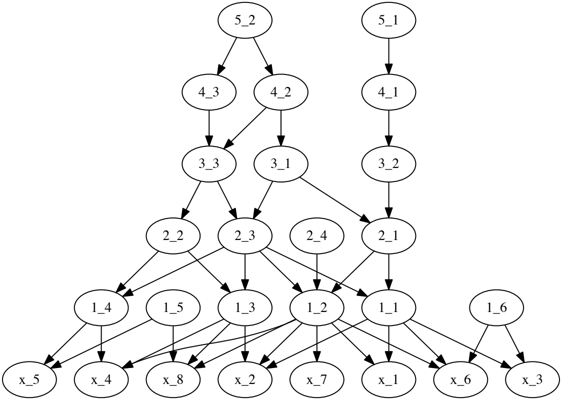
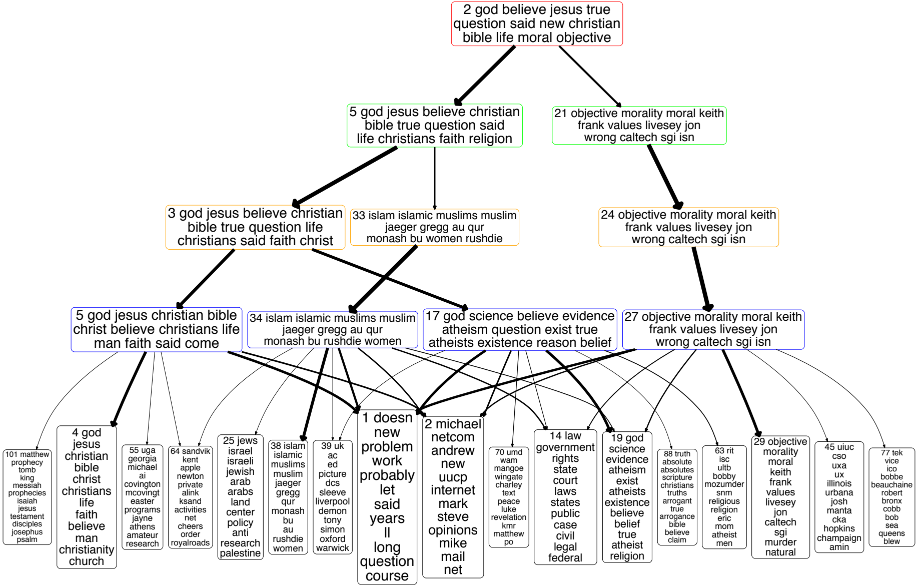


Figure: An example directed network of five hidden layers, with $K_0 = 8$ visible units, $[K_1, K_2, K_3, K_4, K_5] = [6, 4, 3, 3, 2]$, and sparse connections between the hidden units of adjacent layers.

$$
\begin{aligned}
P\left(\boldsymbol{x}_j^{(1)}, \{\boldsymbol{\theta}_j^{(t)}\}_t \,\middle|\, \{\boldsymbol{\Phi}^{(t)}\}_t\right) = {} & P\left(\boldsymbol{x}_j^{(1)} \,\middle|\, \boldsymbol{\Phi}^{(1)}, \boldsymbol{\theta}_j^{(1)}\right) \\
& \times \left[\prod_{t=1}^{T-1} P\left(\boldsymbol{\theta}_j^{(t)} \,\middle|\, \boldsymbol{\Phi}^{(t+1)}, \boldsymbol{\theta}_j^{(t+1)}\right)\right] P\left(\boldsymbol{\theta}_j^{(T)}\right).
\end{aligned}
$$



2 god believe jesus true question said new christian bible life moral objective

5 god jesus believe christian bible true question said life christians faith religion

21 objective morality moral keith frank values livesey jon wrong caltech sgi isn

3 god jesus believe christian bible true question life christians said faith christ

33 islam islamic muslims muslim jaeger gregg au qur monash bu women rushdie

24 objective morality moral keith frank values livesey jon wrong caltech sgi isn

5 god jesus christian bible christ believe christians life man faith said come

34 islam islamic muslims muslim jaeger gregg au qur monash bu rushdie women

17 god science believe evidence atheism question exist true atheists existence reason belief

27 objective morality moral keith frank values livesey jon wrong caltech sgi isn

101 matthew prophecy tomb king messiah prophecies isaiah jesus testament disciples josephus psalm

4 god jesus christian bible christ christians life faith believe man christianity church

55 uga georgia michael ai covington mcovingt easter programs jayne athens amateur research

64 sandvik kent apple newton private alink ksand activities net cheers order royalroads

25 jews israel israeli jewish arab arabs land center policy anti research palestine

38 islam islamic muslims muslim jaeger gregg qur monash bu au rushdie women

39 uk ac ed picture dcs sleeve liverpool demon tony simon oxford warwick

1 doesn new problem work probably let said years ll long question course

2 michael netcom andrew new uucp internet mark steve opinions mike mail net

70 umd wam mangoe wingate charley text peace luke revelation kmr matthew po

14 law government rights state court laws states public case civil legal federal

19 god science evidence atheism exist atheists existence believe belief true atheist religion

88 truth absolute absolutes scripture christians truths arrogant true arrogance bible believe claim

63 rit isc ultb bobby mozumder snm religious religion eric mom atheist men

29 objective morality moral keith frank values livesey jon caltech sgi murder natural

45 uiuc cso uxa ux illinois urbana josh manta cka hopkins champaign amin

77 tek vice ico bobbe beauchaine robert bronx cobb bob sea queens blew

Figure: A tree on “religion.”

The Poisson gamma belief network (PGBN)

- Assume the observations are multivariate count vectors $\boldsymbol{x}_j^{(1)} \in \mathbb{Z}^{K_0}$, where $\mathbb{Z} = \{0, 1, \ldots\}$.
- We construct the PGBN to infer a multilayer deep representation for $\{\boldsymbol{x}_j^{(1)}\}_j$.
- With $\boldsymbol{\Phi}^{(t)} \in \mathbb{R}_+^{K_{t-1} \times K_t}$, the generative model of the PGBN with T hidden layers, from the top to bottom, is expressed as

$$\boldsymbol{\theta}_j^{(T)} \sim \text{Gam}\left(\boldsymbol{r}, 1/c_j^{(T+1)}\right),$$
$$\cdots$$
$$\boldsymbol{\theta}_j^{(t)} \sim \text{Gam}\left(\boldsymbol{\Phi}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}, 1/c_j^{(t+1)}\right),$$
$$\cdots$$
$$\boldsymbol{x}_j^{(1)} \sim \text{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}\right), \quad \boldsymbol{\theta}_j^{(1)} \sim \text{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}, p_j^{(2)}/(1 - p_j^{(2)})\right).$$

- The PGBN factorizes the observed count vectors under the Poisson likelihood into the product of a factor loading matrix and the gamma distributed hidden units of layer one.

- The PGBN factorizes the hidden units of each hidden layer into the product a connection weight matrix and the hidden units of the next layer under the gamma likelihood.

- The PGBN with a single hidden layer (i.e., $T = 1$) reduces to Poisson factor analysis as

$$\boldsymbol{x}_j^{(1)} \sim \text{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}\right), \quad \boldsymbol{\theta}_j^{(1)} \sim \text{Gam}\left(\boldsymbol{r}, p_j^{(2)}/(1 - p_j^{(2)})\right).$$

  The gamma-negative binomial process can be used to support potentially $K_1 = \infty$ number of factors.

Model likelihood

- The joint distribution of the observed counts and gamma hidden units given the network in the PGBN:

$$
\begin{aligned}
P\left(\boldsymbol{x}_j^{(1)}, \{\boldsymbol{\theta}_j^{(t)}\}_t \,\middle|\, \{\boldsymbol{\Phi}^{(t)}\}_t\right) = {} & P\left(\boldsymbol{x}_j^{(1)} \middle| \boldsymbol{\Phi}^{(1)}, \boldsymbol{\theta}_j^{(1)}\right) \\
& \times \left[\prod_{t=1}^{T-1} P\left(\boldsymbol{\theta}_j^{(t)} \middle| \boldsymbol{\Phi}^{(t+1)}, \boldsymbol{\theta}_j^{(t+1)}\right)\right] P\left(\boldsymbol{\theta}_j^{(T)}\right).
\end{aligned}
$$

$$
P\left(\theta_{vj}^{(t)} \,\middle|\, \phi_{v:}^{(t+1)}, \boldsymbol{\theta}_j^{(t+1)}, c_{j+1}^{(t+1)}\right) = \frac{\left(c_{j+1}^{(t+1)}\right)^{\phi_{v:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}}}{\Gamma\left(\phi_{v:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}\right)} \left(\theta_{vj}^{(t)}\right)^{\phi_{v:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}-1} e^{-c_{j+1}^{(t+1)}\theta_{vj}^{(t)}}
$$

- The joint distribution of the binary visible and hidden units given the network for the sigmoid belief network:

$$
P\left(\theta_{vj}^{(t)} = 1 \,|\, \phi_{v:}^{(t+1)}, \boldsymbol{\theta}_j^{(t+1)}, b_v^{(t+1)}\right) = \sigma\left(b_v^{(t+1)} + \phi_{v:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}\right).
$$

where $\sigma(x) = 1/(1+e^{-x})$ denotes the sigmoid function.

Interpreting the structure of the PGBN

- Using the law of total expectation, we have

$$\mathbb{E}\big[\boldsymbol{x}_j^{(1)} \,\big|\, \boldsymbol{\theta}_j^{(t)}, \{\boldsymbol{\Phi}^{(\ell)}, c_j^{(\ell)}\}_{1,t}\big] = \left[\prod_{\ell=1}^{t} \boldsymbol{\Phi}^{(\ell)}\right] \frac{\boldsymbol{\theta}_j^{(t)}}{\prod_{\ell=2}^{t} c_j^{(t)}}.$$

$$\mathbb{E}\big[\boldsymbol{\theta}_j^{(t)} \,\big|\, \{\boldsymbol{\Phi}^{(\ell)}, c_j^{(\ell)}\}_{t+1,T}, \boldsymbol{r}\big] = \left[\prod_{\ell=t+1}^{T} \boldsymbol{\Phi}^{(\ell)}\right] \frac{\boldsymbol{r}}{\prod_{\ell=t+1}^{T+1} c_j^{(\ell)}}.$$

- Consider $\prod_{\ell=1}^{t} \boldsymbol{\Phi}^{(\ell)}$ as the K_t topics/factors/nodes of layer $t \in \{1, \dots, T\}$ and use $\boldsymbol{r}^{(t)} := \left[\prod_{\ell=t+1}^{T} \boldsymbol{\Phi}^{(\ell)}\right] \boldsymbol{r}$ to rank them.
- Consider $\phi_{k'k}^{(t)} = \boldsymbol{\Phi}^{(t)}(k', k)$ as the weight that connects node k of layer t and node k' of layer $t-1$.
- Our intuition is that examining the topics/factors from the top to bottom layers will gradually reveal less general and more specific aspects of the data.

1 doesn new problem work probably let said years ll long question course

11 team game hockey season games nhl year league teams players play cup

21 disease doctor pain treatment medical patients medicine cancer help skin blood problems

31 key chip bit keys number serial encrypted bits clipper escrow algorithm des

41 henry toronto zoo spencer work utzoo zoology man umd kipling eng allen

51 ground wire wiring neutral outlets circuit cable electrical wires outlet connected house

61 callison uoknor james car ecn lights uokmax fake bird continental alarm slb

71 candida yeast noring steve symptoms jon vitamin dyer body quack patients sinus

81 sex child copy women children protection pregnancy depression abstinence sexual education sehari

2 michael netcom andrew new uucp internet mark steve opinions mike mail net

12 sale new shipping offer condition price asking sell cover cd interested best

22 space nasa earth orbit gov data shuttle mission lunar spacecraft solar jpl

32 israel israeli lebanese lebanon attacks peace arab israelis civilians killed palestinian soldiers

42 graphics ftp pub data ray risc contact sgi machines pc instruction mail

52 won lost san new kaldis rutgers york houston st astros louis reds

62 sweden fi finland germany canada players finnish german april swedish wc czech

72 gatech prism gt georgia atlanta technology institute internet uucp gilkey hydra lankford

82 hall dave fame smith eddie murray winfield kingman yount steve guys bsu

3 thanks mail email help fax advance looking hi info information phone send

13 power output input high signal low voltage chip radio battery circuit data

23 gun guns firearms crime control weapons self criminals defense handgun firearm violent

33 cramer men gay homosexual optilink sexual clayton sex homosexuals homosexuality male virginia

43 ohio state magnus acs drugs ncr atlantaga wilson ncratl legal war ryan

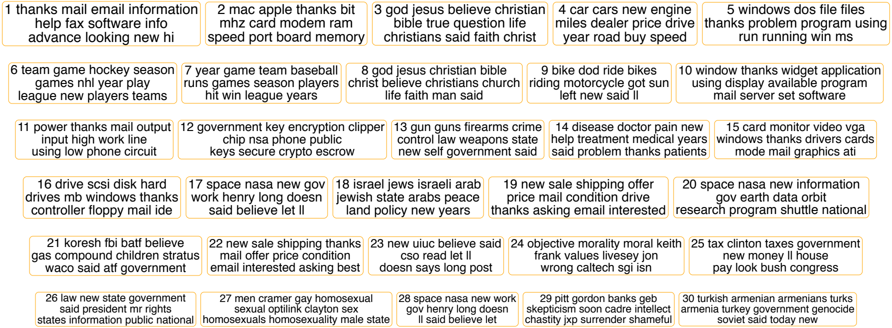
53 printer hp print laser ink printers deskjet postscript bj canon quality toner

63 rit isc ultb bobby mozumder snm religious religion eric mom atheist men

73 msg dyer spdcc glutamate food effects brain studies humans blood olney foods

83 radar detector detectors police car beam illegal antenna radio virginia law band

Figure: Example topics of layer one of the PGBN learned on the 20newsgroups corpus.



1 thanks mail email information help fax software info advance looking new hi

2 mac apple thanks bit mhz card modem ram speed port board memory

3 god jesus believe christian bible true question life christians said faith christ

4 car cars new engine miles dealer price drive year road buy speed

5 windows dos file files thanks problem program using run running win ms

6 team game hockey season games nhl year play league new players teams

7 year game team baseball runs games season players hit win league years

8 god jesus christian bible christ believe christians church life faith man said

9 bike dod ride bikes riding motorcycle got sun left new said ll

10 window thanks widget application using display available program mail server set software

11 power thanks mail output input high work line using low phone circuit

12 government key encryption clipper chip nsa phone public keys secure crypto escrow

13 gun guns firearms crime control law weapons state new self government said

14 disease doctor pain new help treatment medical years said problem thanks patients

15 card monitor video vga windows thanks drivers cards mode mail graphics ati

16 drive scsi disk hard drives mb windows thanks controller floppy mail ide

17 space nasa new gov work henry long doesn said believe let ll

18 israel jews israeli arab jewish state arabs peace land policy new years

19 new sale shipping offer price mail condition drive thanks asking email interested

20 space nasa new information gov earth data orbit research program shuttle national

21 koresh fbi batf believe gas compound children stratus waco said atf government

22 new sale shipping thanks mail offer price condition email interested asking best

23 new uiuc believe said cso read let ll doesn says long post

24 objective morality moral keith frank values livesey jon wrong caltech sgi isn

25 tax clinton taxes government new money ll house pay look bush congress

26 law new state government said president mr rights states information public national

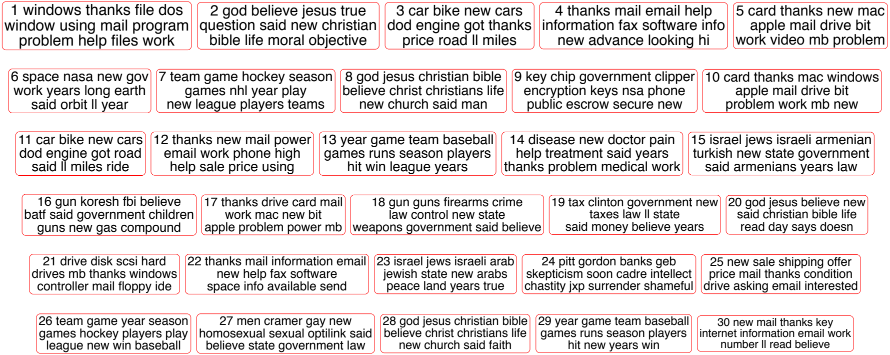
27 men cramer gay homosexual sexual optilink clayton sex homosexuals homosexuality male state

28 space nasa new work gov henry long doesn ll said believe let

29 pitt gordon banks geb skepticism soon cadre intellect chastity jxp surrender shameful

30 turkish armenian armenians turks armenia turkey government genocide soviet said today new

Figure: The top 30 topics of layer three of the PGBN learned on the 20newsgroups corpus.



1 windows thanks file dos window using mail program problem help files work

2 god believe jesus true question said new christian bible life moral objective

3 car bike new cars dod engine got thanks price road ll miles

4 thanks mail email help information fax software info new advance looking hi

5 card thanks new mac apple mail drive bit work video mb problem

6 space nasa new gov work years long earth said orbit ll year

7 team game hockey season games nhl year play new league players teams

8 god jesus christian bible believe christ christians life new church said man

9 key chip government clipper encryption keys nsa phone public escrow secure new

10 card thanks mac windows apple mail drive bit problem work mb new

11 car bike new cars dod engine got road said ll miles ride

12 thanks new mail power email work phone high help sale price using

13 year game team baseball games runs season players hit win league years

14 disease new doctor pain help treatment said years thanks problem medical work

15 israel jews israeli armenian turkish new state government said armenians years law

16 gun koresh fbi believe batf said government children guns new gas compound

17 thanks drive card mail work mac new bit apple problem power mb

18 gun guns firearms crime law control new state weapons government said believe

19 tax clinton government new taxes law ll state said money believe years

20 god jesus believe new said christian bible life read day says doesn

21 drive disk scsi hard drives mb thanks windows controller mail floppy ide

22 thanks mail information email new help fax software space info available send

23 israel jews israeli arab jewish state new arabs peace land years true

24 pitt gordon banks geb skepticism soon cadre intellect chastity jxp surrender shameful

25 new sale shipping offer price mail thanks condition drive asking email interested

26 team game year season games hockey players play league new win baseball

27 men cramer gay new homosexual sexual optilink said believe state government law

28 god jesus christian bible believe christ christians life new church said faith

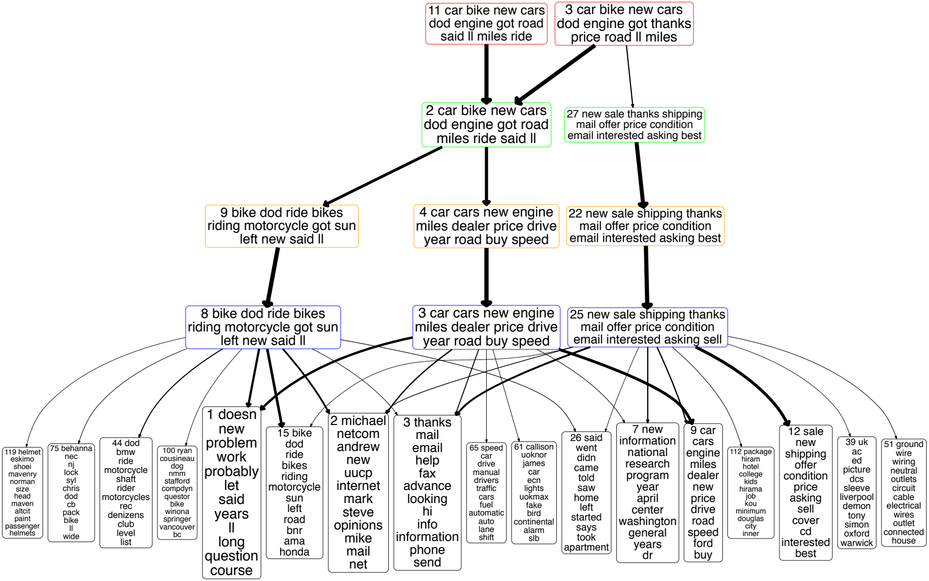
29 year game team baseball games runs season players hit new years win

30 new mail thanks key internet information email work number ll read believe

Figure: The top 30 topics of layer five of the PGBN learned on the 20newsgroups corpus.

Visualize the inferred deep network

- Visualize the whole network is challenging if the network is large.
- But we can easily visualize a tree or a subnetwork that consists of multiple trees.
- Construct a tree starting from a top-layer node: grow the tree downward by linking each leaf node of the tree to all the hidden units at the layer below that are connected to the leaf node with nonnegligible weights.



11 car bike new cars dod engine got road said ll miles ride

3 car bike new cars dod engine got thanks price road ll miles

2 car bike new cars dod engine got road miles ride said ll

27 new sale thanks shipping mail offer price condition email interested asking best

9 bike dod ride bikes riding motorcycle got sun left new said ll

4 car cars new engine miles dealer price drive year road buy speed

22 new sale shipping thanks mail offer price condition email interested asking best

8 bike dod ride bikes riding motorcycle got sun left new said ll

3 car cars new engine miles dealer price drive year road buy speed

25 new sale shipping thanks mail offer price condition email interested asking sell

119 helmet eskimo shoei mavenry norman size head maven altcit paint passenger helmets

75 behanna nec nj lock syl chris dod cb pack bike ll wide

44 dod bmw ride motorcycle shaft rider motorcycles rec denizens club level list

100 ryan cousineau dog nmm stafford compdyn questor bike winona springer vancouver bc

1 doesn new problem work probably let said years ll long question course

15 bike dod ride bikes riding motorcycle sun left road bnr ama honda

2 michael netcom andrew new uucp internet mark steve opinions mike mail net

3 thanks mail email help fax advance looking hi info information phone send

65 speed car drive manual drivers traffic cars fuel automatic auto lane shift

61 callison uoknor james car ecn lights uokmax fake bird continental alarm slb

26 said went didn came told saw home left started says took apartment

7 new information national research program year april center washington general years dr

9 car cars engine miles dealer new price drive road speed ford buy

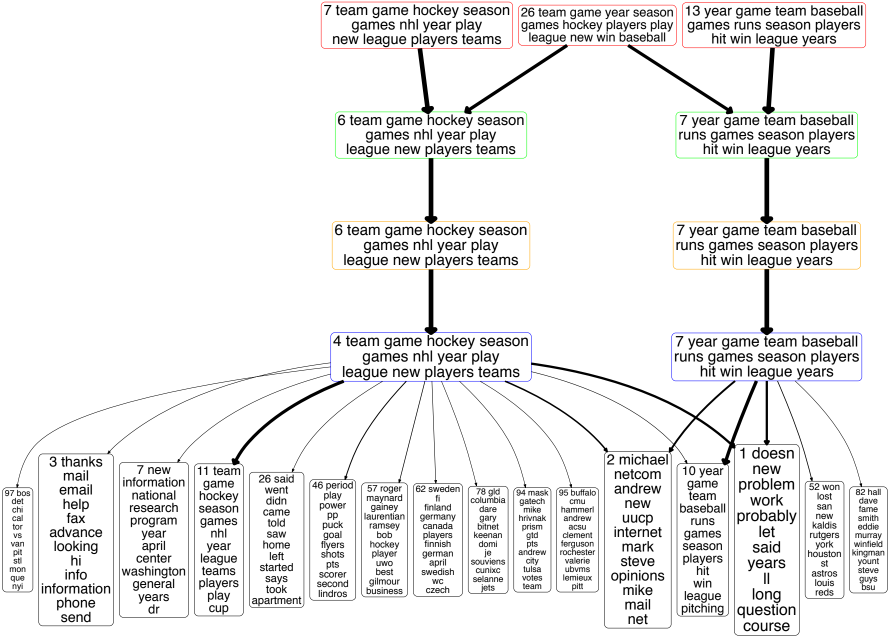
112 package hiram hotel college kids hirama job kou minimum douglas city inner

12 sale new shipping offer condition price asking sell cover cd interested best

39 uk ac ed picture dcs sleeve liverpool demon tony simon oxford warwick

51 ground wire wiring neutral outlets circuit cable electrical wires outlet connected house

Figure: A subnetwork on “car & bike.”



7 team game hockey season games nhl year play new league players teams

26 team game year season games hockey players play league new win baseball

13 year game team baseball games runs season players hit win league years

6 team game hockey season games nhl year play league new players teams

7 year game team baseball runs games season players hit win league years

6 team game hockey season games nhl year play league new players teams

7 year game team baseball runs games season players hit win league years

4 team game hockey season games nhl year play league new players teams

7 year game team baseball runs games season players hit win league years

97 bos det chi cal tor vs van pit stl mon que nyi

3 thanks mail email help fax advance looking hi info information phone send

7 new information national research program year april center washington general years dr

11 team game hockey season games nhl year league teams players play cup

26 said went didn came told saw home left started says took apartment

46 period play power pp puck goal flyers shots pts scorer second lindros

57 roger maynard gainey laurentian ramsey bob hockey player uwo best gilmour business

62 sweden fi finland germany canada players finnish german april swedish wc czech

78 gld columbia dare gary bitnet keenan domi je souviens cunixc selanne jets

94 mask gatech mike hrivnak prism gtd pts andrew city tulsa votes team

95 buffalo cmu hammerl andrew acsu clement ferguson rochester valerie ubvms lemieux pitt

2 michael netcom andrew new uucp internet mark steve opinions mike mail net

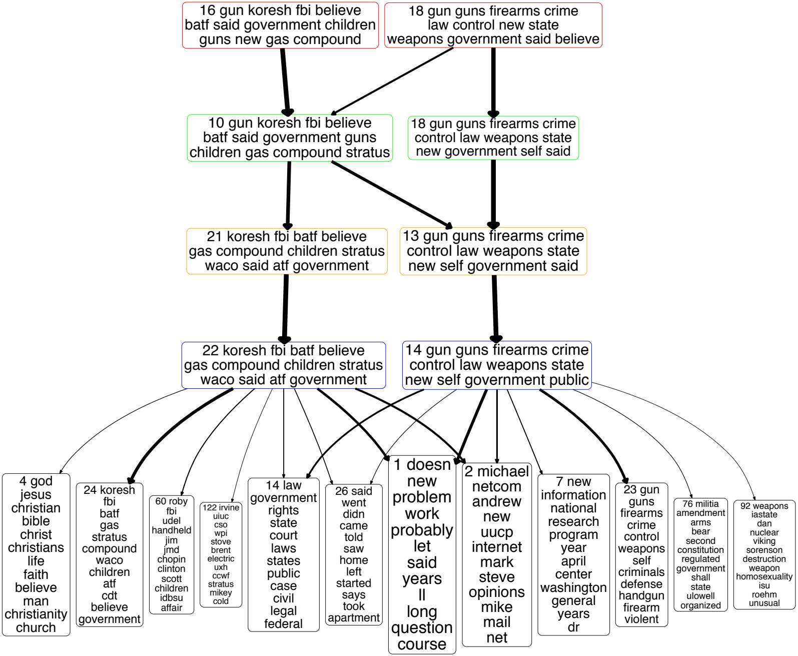
10 year game team baseball runs games season players hit win league pitching

1 doesn new problem work probably let said years ll long question course

52 won lost san new kaldis rutgers york houston st astros louis reds

82 hall dave fame smith eddie murray winfield kingman yount steve guys bsu

Figure: A subnetwork on “sports.”



16 gun koresh fbi believe batf said government children guns new gas compound

18 gun guns firearms crime law control new state weapons government said believe

10 gun koresh fbi believe batf said government guns children gas compound stratus

18 gun guns firearms crime control law weapons state new government self said

21 koresh fbi batf believe gas compound children stratus waco said atf government

13 gun guns firearms crime control law weapons state new self government said

22 koresh fbi batf believe gas compound children stratus waco said atf government

14 gun guns firearms crime control law weapons state new self government public

4 god jesus christian bible christ christians life faith believe man christianity church

24 koresh fbi batf gas stratus compound waco children atf cdt believe government

60 roby fbi udel handheld jim jmd chopin clinton scott children idbsu affair

122 irvine uiuc cso wpi stove brent electric uxh ccwf stratus mikey cold

14 law government rights state court laws states public case civil legal federal

26 said went didn came told saw home left started says took apartment

1 doesn new problem work probably let said years ll long question course

2 michael netcom andrew new uucp internet mark steve opinions mike mail net

7 new information national research program year april center washington general years dr

23 gun guns firearms crime control weapons self criminals defense handgun firearm violent

76 militia amendment arms bear second constitution regulated government shall state ulowell organized

92 weapons iastate dan nuclear viking sorenson destruction weapon homosexuality isu roehm unusual

Figure: A subnetwork on “gun.”

Upward-downward Gibbs sampling

Lemma (Augment-and-conquer the gamma belief network)

With $p_j^{(1)} := 1 - e^{-1}$ *and*

$$p_j^{(t+1)} := -\ln(1 - p_j^{(t)}) \Big/ \left[c_j^{(t+1)} - \ln(1 - p_j^{(t)})\right]$$

for $t = 1, \ldots, T$*, one may connect the observed or latent counts* $\boldsymbol{x}_j^{(t)} \in \mathbb{Z}^{K_{t-1}}$ *to the product* $\boldsymbol{\Phi}^{(t)}\boldsymbol{\theta}_j^{(t)}$ *at layer* t *under the Poisson likelihood as*

$$\boldsymbol{x}_j^{(t)} \sim \mathit{Pois}\left[-\boldsymbol{\Phi}^{(t)}\boldsymbol{\theta}_j^{(t)} \ln\left(1 - p_j^{(t)}\right)\right].$$

Upward propagate latent counts

Corollary (Propagate the latent counts upward)

With $m_{kj}^{(t)(t+1)} := x_{\cdot jk}^{(t)} := \sum_{v=1}^{K_{t-1}} x_{vjk}^{(t)}$ *representing the number of times that factor* $k \in \{1, \ldots, K_t\}$ *of layer* t *appears in observation* j*, we can propagate the latent counts* $x_{vj}^{(t)}$ *of layer* t *upward to layer* $t+1$ *as*

$$\left\{ \left(x_{vj1}^{(t)}, \ldots, x_{vjK_t}^{(t)} \right) \middle| \, x_{vj}^{(t)}, \boldsymbol{\phi}_{v:}^{(t)}, \boldsymbol{\theta}_j^{(t)} \right\} \sim Mult\left(x_{vj}^{(t)}, \frac{\phi_{v1}^{(t)}\theta_{1j}^{(t)}}{\sum_{k=1}^{K_t}\phi_{vk}^{(t)}\theta_{kj}^{(t)}}, \ldots, \frac{\phi_{vK_t}^{(t)}\theta_{K_tj}^{(t)}}{\sum_{k=1}^{K_t}\phi_{vk}^{(t)}\theta_{kj}^{(t)}} \right),$$

$$(\boldsymbol{\phi}_k^{(t)} | -) \sim Dir\left(\eta_1^{(t)} + x_{1\cdot k}^{(t)}, \ldots, \eta_{K_{t-1}}^{(t)} + x_{K_{t-1}\cdot k}^{(t)} \right),$$

$$\left(x_{kj}^{(t+1)} \,\middle|\, m_{kj}^{(t)(t+1)}, \boldsymbol{\phi}_{k:}^{(t+1)}, \boldsymbol{\theta}_j^{(t+1)} \right) \sim CRT\left(m_{kj}^{(t)(t+1)}, \boldsymbol{\phi}_{k:}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)} \right).$$

Downward sample the hidden units

Using the latent counts propagated upward and the gamma-Poisson conjugacy, we downward sample the hidden units as

$$
\begin{aligned}
(r_k \,|\, -) &\sim \text{Gam}\left(\gamma_0/K_T + x_{k\cdot}^{(T+1)}, \left[c_0 - \textstyle\sum_j \ln\left(1 - p_j^{(T+1)}\right)\right]^{-1}\right), \\
(\boldsymbol{\theta}_j^{(T)} \,|\, -) &\sim \text{Gam}\left(\boldsymbol{r} + \boldsymbol{m}_j^{(T)(T+1)}, \left[c_j^{(T+1)} - \ln\left(1 - p_j^{(T)}\right)\right]^{-1}\right), \\
(\boldsymbol{\theta}_j^{(T-1)} \,|\, -) &\sim \text{Gam}\left(\boldsymbol{\Phi}^{(T)}\boldsymbol{\theta}_j^{(T)} + \boldsymbol{m}_j^{(T)(T+1)}, \left[c_j^{(T)} - \ln\left(1 - p_j^{(T-1)}\right)\right]^{-1}\right), \\
&\vdots \\
(\boldsymbol{\theta}_j^{(t)} \,|\, -) &\sim \text{Gam}\left(\boldsymbol{\Phi}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)} + \boldsymbol{m}_j^{(t)(t+1)}, \left[c_j^{(t+1)} - \ln\left(1 - p_j^{(t)}\right)\right]^{-1}\right), \\
&\vdots \\
(\boldsymbol{\theta}_j^{(1)} \,|\, -) &\sim \text{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)} + \boldsymbol{m}_j^{(1)(2)}, \left[c_j^{(2)} - \ln\left(1 - p_j^{(1)}\right)\right]^{-1}\right).
\end{aligned}
$$

Modeling overdispersion with distributed representation

Assuming $\boldsymbol{\Phi}^{(t)} = \mathbf{I}$ for all $t \in 3, \ldots, T$, we have

$$m_{kj}^{(1)(2)} \sim \mathrm{NB}(\theta_{kj}^{(2)}, p_j^{(2)}), \ldots, \theta_{kj}^{(t)} \sim \mathrm{Gam}(\theta_{kj}^{(t+1)}, 1/c_j^{(t+1)}),$$
$$\ldots, \theta_{kj}^{(T)} \sim \mathrm{Gam}(r_k, 1/c_j^{(T+1)}).$$

- In comparison to PFA with $m_{kj}^{(1)(2)} \sim \mathrm{NB}(r_k, p_j^{(2)})$, the PGBN increases $\mathrm{VMR}\left[m_{kj}^{(1)(2)} \mid r_k\right]$ by a factor of

  $$1 + p_j^{(2)} \sum_{t=3}^{T+1} \left[\prod_{\ell=3}^{t} \left(c_j^{(\ell)} \right)^{-1} \right],$$

  which is equal to

  $$1 + (T-1) p_j^{(2)}$$

  if we further assume $c_j^{(t)} = 1$ for all $t \geq 3$.
- Therefore, by increasing the depth of the network to distribute the variance into more layers, the multilayer structure could increase its capability to model data variability.

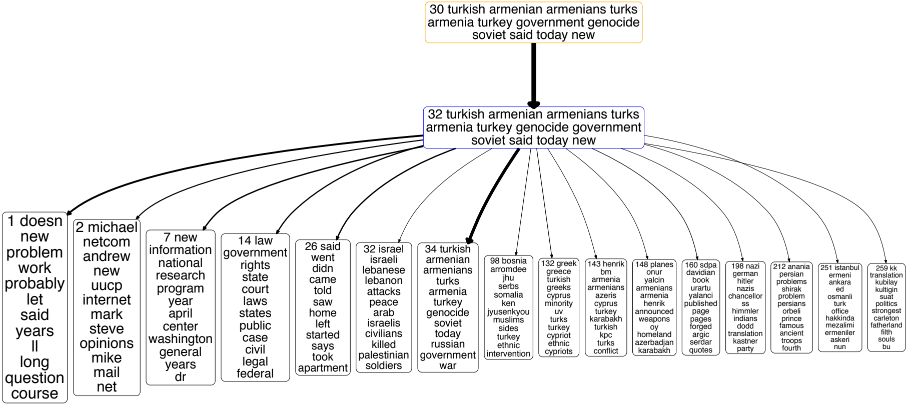
- For the GBN with $T = 1$, given the shared weight vector $\boldsymbol{r}$, we have

$$\mathbb{E}\big[\boldsymbol{x}_j^{(1)} \,\big|\, \boldsymbol{\Phi}^{(1)}, \boldsymbol{r}\big] = \boldsymbol{\Phi}^{(1)}\boldsymbol{r}/c_j^{(2)};$$

- For the GBN with $T \geq 2$, given the weight vector $\boldsymbol{\theta}_j^{(2)}$, we have

$$\mathbb{E}\big[\boldsymbol{x}_j^{(1)} \,\big|\, \boldsymbol{\Phi}^{(1)}, \boldsymbol{\Phi}^{(2)}, \boldsymbol{\theta}_j^{(2)}\big] = \boldsymbol{\Phi}^{(1)}\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}/c_j^{(2)}.$$

- Thus in the prior, the co-occurrence patterns of the columns of $\boldsymbol{\Phi}^{(1)}$ are captured in a single vector $\boldsymbol{r}$ when $T = 1$, and are captured in the columns of $\boldsymbol{\Phi}^{(2)}$ when $T \geq 2$.
- Similarly, in the prior, if $T \geq t + 1$, the co-occurrence patterns of the K_t columns of the projected topics $\prod_{\ell=1}^{t} \boldsymbol{\Phi}^{(\ell)}$ will be captured in the columns of the $K_t \times K_{t+1}$ matrix $\boldsymbol{\Phi}^{(t+1)}$.



30 turkish armenian armenians turks armenia turkey government genocide soviet said today new

32 turkish armenian armenians turks armenia turkey genocide government soviet said today new

1 doesn new problem work probably let said years ll long question course

2 michael netcom andrew new uucp internet mark steve opinions mike mail net

7 new information national research program year april center washington general years dr

14 law government rights state court laws states public case civil legal federal

26 said went didn came told saw home left started says took apartment

32 israel israeli lebanese lebanon attacks peace arab israelis civilians killed palestinian soldiers

34 turkish armenian armenians turks armenia turkey genocide soviet today russian government war

98 bosnia arromdee jhu serbs somalia ken jyusenkyou muslims sides turkey ethnic intervention

132 greek greece turkish greeks cyprus minority uv turks turkey cypriot ethnic cypriots

143 henrik bm armenia armenians azeris cyprus turkey karabakh turkish kpc turks conflict

148 planes onur yalcin armenians armenia henrik announced weapons oy homeland azerbadjan karabakh

160 sdpa davidian book urartu yalanci published page pages forged argic serdar quotes

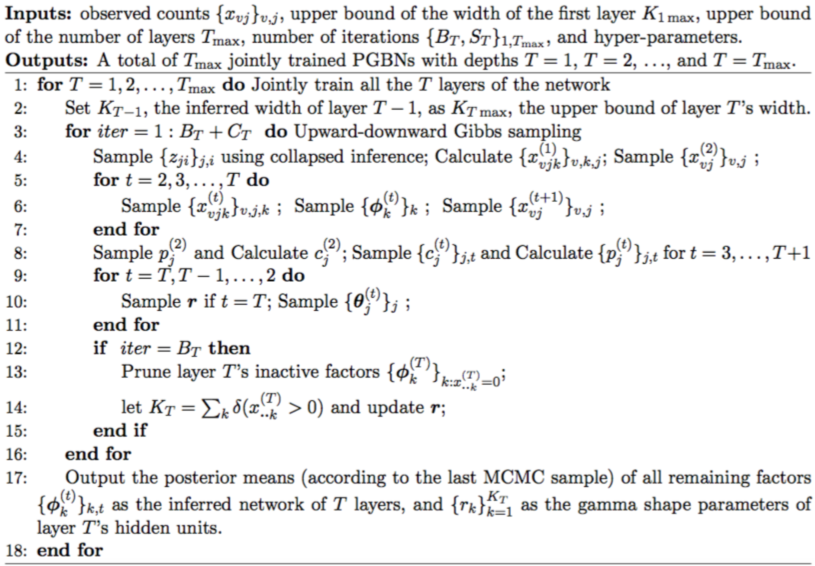
198 nazi german hitler nazis chancellor ss himmler indians dodd translation kastner party

212 anania persian problems shirak problem persians orbeli prince famous ancient troops fourth

251 istanbul ermeni ankara ed osmanli turk office hakkinda mezalimi ermeniler askeri nun

259 kk translation kubilay kultigin suat politics strongest carleton fatherland filth souls bu

Figure: The tree rooted at node 30 of layer three on “Turkey & Armenia.”



Inputs: observed counts $\{x_{vj}\}_{v,j}$, upper bound of the width of the first layer $K_{1\,\max}$, upper bound of the number of layers $T_{\max}$, number of iterations $\{B_T, S_T\}_{1,T_{\max}}$, and hyper-parameters.
Outputs: A total of $T_{\max}$ jointly trained PGBNs with depths $T = 1$, $T = 2$, …, and $T = T_{\max}$.

```
for $T = 1, 2, \ldots, T_{\max}$ do Jointly train all the $T$ layers of the network
    Set $K_{T-1}$, the inferred width of layer $T-1$, as $K_{T\,\max}$, the upper bound of layer $T$'s width.
    for $iter = 1 : B_T + C_T$ do Upward-downward Gibbs sampling
        Sample $\{z_{ji}\}_{j,i}$ using collapsed inference; Calculate $\{x_{vjk}^{(1)}\}_{v,k,j}$; Sample $\{x_{vj}^{(2)}\}_{v,j}$ ;
        for $t = 2, 3, \ldots, T$ do
            Sample $\{x_{vjk}^{(t)}\}_{v,j,k}$ ; Sample $\{\boldsymbol{\phi}_k^{(t)}\}_k$ ; Sample $\{x_{vj}^{(t+1)}\}_{v,j}$ ;
        end for
        Sample $p_j^{(2)}$ and Calculate $c_j^{(2)}$; Sample $\{c_j^{(t)}\}_{j,t}$ and Calculate $\{p_j^{(t)}\}_{j,t}$ for $t = 3, \ldots, T+1$
        for $t = T, T-1, \ldots, 2$ do
            Sample $\boldsymbol{r}$ if $t = T$; Sample $\{\boldsymbol{\theta}_j^{(t)}\}_j$ ;
        end for
        if $iter = B_T$ then
            Prune layer $T$'s inactive factors $\{\boldsymbol{\phi}_k^{(T)}\}_{k:x_{\cdot\cdot k}^{(T)}=0}$;
            let $K_T = \sum_k \delta(x_{\cdot\cdot k}^{(T)} > 0)$ and update $\boldsymbol{r}$;
        end if
    end for
    Output the posterior means (according to the last MCMC sample) of all remaining factors
  $\{\boldsymbol{\phi}_k^{(t)}\}_{k,t}$ as the inferred network of $T$ layers, and $\{r_k\}_{k=1}^{K_T}$ as the gamma shape parameters of
  layer $T$'s hidden units.
end for
```

Modeling binary and nonnegative real observations

- We link binary observations to the latent counts at layer one as

$$b_{vj}^{(1)} = \mathbf{1}\big(x_{vj}^{(1)} \geq 1\big).$$

  - For inference, we sample the latent counts at layer one from the truncated Poisson distribution as

$$\big(x_{vj}^{(1)} \mid -\big) \sim b_{vj}^{(1)} \cdot \mathrm{Pois}_{+}\left(\sum_{k=1}^{K_1} \phi_{vk}^{(1)} \theta_{kj}^{(1)}\right).$$

- We link nonnegative real observations to the latent counts at layer one using

$$y_{vj}^{(1)} \sim \mathrm{Gam}\big(x_{vj}^{(1)}, 1/a_j\big).$$

  - For inference, we let $x_{vj}^{(1)} = 0$ if $y_{vj}^{(1)} = 0$ and sample $x_{vj}^{(1)}$ from the truncated Bessel distribution as

$$\big(x_{vj}^{(1)} \mid -\big) \sim \mathrm{Bessel}_{-1}\left(2\sqrt{a_j y_{vj}^{(1)} \sum_{k=1}^{K_1} \phi_{vk}^{(1)} \theta_{kj}^{(1)}}\right)$$

  if $y_{vj}^{(1)} > 0$.

Multivariate count data

- Train on the 20 newsgroups dataset, each document of which is summarized as a word count vector over the vocabulary.
  - We use all 11,269 training documents to infer a five layer network.
  - After removing stopwords and terms that appear less than five times, we obtain a vocabulary with $K_0 = 33,420$.
  - With $\eta^{(t)} = 0.05$ for all t, the inferred network widths by the PGBN are $[K_1, K_2, K_3, K_4, K_5] =$
    - $[50, 50, 50, 50, 50]$ for $K_{1\,\max} = 50$
    - $[100, 99, 99, 94, 87]$ for $K_{1\,\max} = 100$
    - $[200, 161, 130, 94, 63]$ for $K_{1\,\max} = 200$
    - $[396, 109, 99, 82, 68]$ for $K_{1\,\max} = 400$
    - $[528, 129, 109, 98, 91]$ for $K_{1\,\max} = 600$
    - $[608, 100, 99, 96, 89]$ for $K_{1\,\max} = 800$
  - Test on the 7,505 documents in the testing set.

Feature learning for multi-class classification

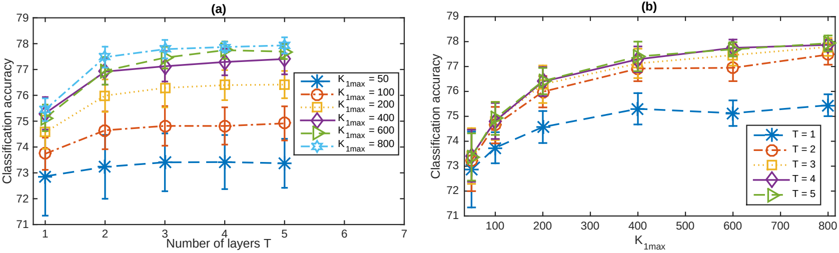


Figure: Classification accuracy (%) of the PGBNs with Algorithm 1 for 20newsgroups multi-class classification (a) as a function of the depth T with various $K_{1\,\mathrm{max}}$ and (b) as a function of $K_{1\,\mathrm{max}}$ with various depths, with $\eta^{(t)} = 0.05$ for all layers.

Prediction of heldout words

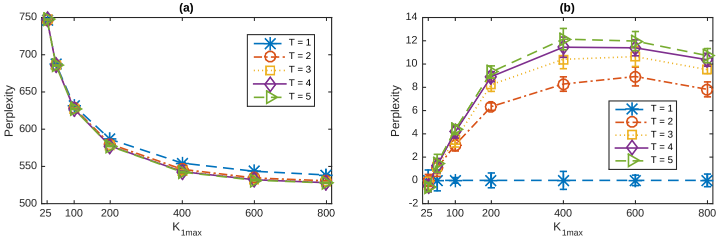


Figure: (a) per-heldout-word perplexity (the lower the better) for the NIPS12 corpus (using the 2000 most frequent terms) as a function of the upper bound of the first layer width $K_{1\,\max}$ and network depth T, with 30% of the word tokens in each document used for training and $\eta^{(t)} = 0.05$ for all t. (b) for visualization, each curve in (a) is reproduced by subtracting its values from the average perplexity of the single-layer network.

Simulate documents

- Draw $\boldsymbol{\theta}^{(T)}, \boldsymbol{\theta}^{(T-1)}, \ldots, \boldsymbol{\theta}^{1}$ using

$$
\begin{gathered}
\boldsymbol{\theta}_{j'}^{(T)} \sim \operatorname{Gam}\left(\boldsymbol{r}, \left[c_{j'}^{(T+1)}\right]^{-1}\right), \\
\boldsymbol{\theta}_{j'}^{(T-1)} \sim \operatorname{Gam}\left(\boldsymbol{\Phi}^{(T)}\boldsymbol{\theta}_{j'}^{(T)}, \left[c_{j'}^{(T)}\right]^{-1}\right), \\
\vdots \\
\boldsymbol{\theta}_{j'}^{(t)} \sim \operatorname{Gam}\left(\boldsymbol{\Phi}^{(t+1)}\boldsymbol{\theta}_{j'}^{(t+1)}, \left[c_{j'}^{(t+1)}\right]^{-1}\right), \\
\vdots \\
\boldsymbol{\theta}_{j'}^{(1)} \sim \operatorname{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_{j'}^{(2)}, \left[c_{j'}^{(2)}\right]^{-1}\right).
\end{gathered}
$$

- Calculate $E[\boldsymbol{x}_{j'}^{(1)}] = \boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_{j'}^{(1)}$
- Display the top 100 words

Simulate documents

- mac apple bit mhz ram simms mb like memory just don cpu people chip chips think color board ibm speed does know se video time machines motherboard hardware lc cache meg ns simm need upgrade built vram good quadra want centris price dx run way processor card clock slots make fpu internal did macs cards ve pin power really machine say faster said software intel macintosh right week writes slot going sx performance things edu years nubus possible thing monitor work point expansion rom iisi ll add dram better little slow let sure pc ii didn ethernet lciii case kind

Simulate documents

- image jpeg gif file color files images format bit display convert quality formats colors programs program tiff picture viewer graphics bmp bits xv screen pixel read compression conversion zip shareware scale view jpg original save quicktime jfif free version best pcx viewing bitmap gifs simtel viewers don mac usenet resolution animation menu scanner pixels sites gray quantization displays better try msdos tga want current black faq converting white setting mirror xloadimage section ppm fractal amiga write algorithm mpeg pict targa arithmetic export scodal archive converted grasp lossless let space human grey directory pictures rgb demo scanned old choice grayscale compress

Simulate documents

- medical health disease doctor pain patients treatment medicine cancer edu hiv blood use years patient writes cause skin don like just aids symptoms number article help diseases drug com effects information doctors infection physician normal chronic think taking care volume condition drugs page says cure people tobacco hicnet know newsletter effective therapy problem common time women prevent surgery children center immune research called april control effect weeks low syndrome hospital physicians states clinical diagnosed day med age good make caused severe reported public safety child said cdc usually diet national studies tissue months way cases causing migraine smokeless infections does

Simulate documents

- men homosexual sex gay sexual homosexuality male don people partners promiscuous number just bi like study homosexuals percent cramer heterosexual think did dramatically numbers straight church reform population know report pythagorean life man good accept time said considered kinsey posted general optilink irrational social gays behavior way children make published johnson survey table new activity showing million statistics american sexuality shows want right women ve article ago exclusively eating virginia masters repent really say purpose member clayton apparent kolodny writes going press society evil function engaged relationships ryan evolutionary different does compared person join figure community edu chose interesting things

Simulate documents

- nissan electronics wagon altima delcoelect kocrsv station gm subaru sumax delco spiros hughes wax pathfinder legacy kokomo wagons smorris scott toyota seattleu don just like strong silver software luxury derek proof stanza seattle cisco morris cymbal triantafyllopoulos sportscar think people know near fool ugly proud claims flat statistics lincoln sedans bullet karl lee perth puzzled miata sentra maxima acura infiniti corolla mgb untruth verbatim good time consider way based make stand guys writes noticed want ve heavy suggestion eat steven horrible uunet studies armor fisher lust designs study definately lexus remove conversion embodied aesthetic elvis attached honey stole designing wd

Modeling high-dimensional sparse binary vectors

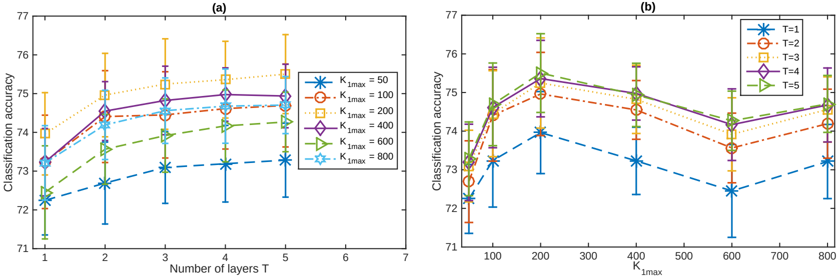


Figure: Results of the BerPo-GBNs on the binarized 20newsgroups term-document count matrix. The widths of the hidden layers are automatically inferred. In a random trial with Algorithm 2, the inferred network widths $[K_1, \ldots, K_5]$ for $K_{1\,\max} = 50, 100, 200, 400, 600$, and 800 are $[50, 50, 50, 50, 50]$, $[100, 97, 95, 90, 82]$, $[178, 145, 122, 97, 72]$, $[184, 139, 119, 101, 75]$, $[172, 165, 158, 138, 110]$, and $[156, 151, 147, 134, 117]$, respectively.

Modeling high-dimensional sparse nonnegative real vectors

- Train on the 60,000 MNIST digits in the training set, each digit of which is represented as a 784 dimensional nonnegative real vector.
    - We use all 60,000 to infer a five layer network.
    - With $\eta^{(t)} = 0.05$ for all t, the inferred network widths by the gamma-PGBN are $[K_1, K_2, K_3, K_4, K_5] =$
        - $[50, 50, 50, 50, 50]$ for $K_{1\,\max} = 50$
        - $[100, 100, 100, 100, 100]$ for $K_{1\,\max} = 100$
        - $[200, 200, 200, 200, 200]$ for $K_{1\,\max} = 200$
        - $[400, 400, 399, 385, 321]$ for $K_{1\,\max} = 400$
    - Test on the 10,000 MNIST digits in the test set.

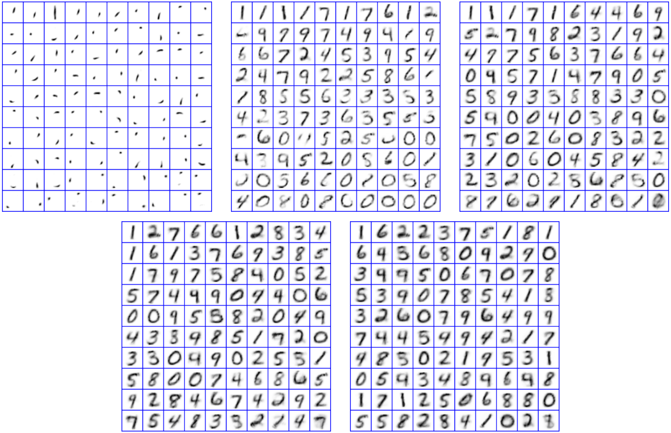
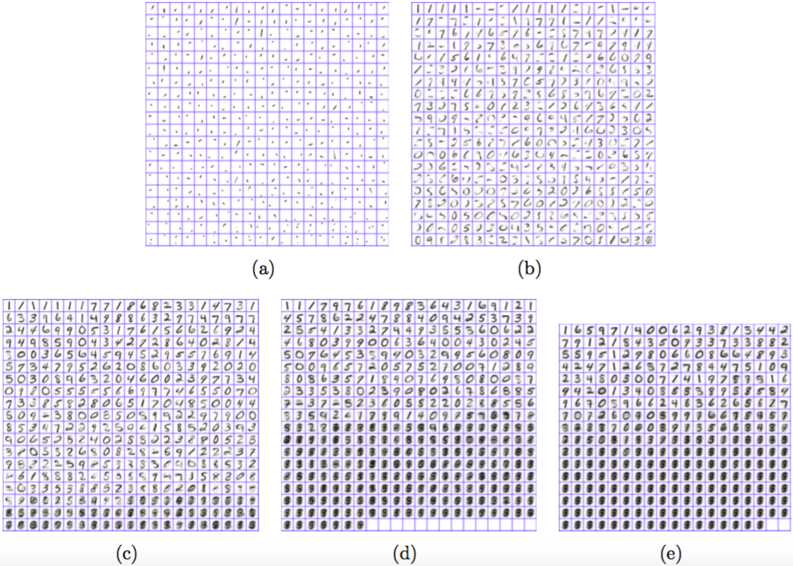


Figure: Visualization of the inferred $\{\Phi^{(1)}, \cdots, \Phi^{(T)}\}$ with $K_{1max} = 100$. All Φ's are projected to the first layer.
First row from left to right: $\Phi^{(1)}$, $\Phi^{(1)}\Phi^{(2)}$, $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}$
Second row from left to right: $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}\Phi^{(4)}$, $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}\Phi^{(4)}\Phi^{(5)}$.



(a)

(b)

(c)

(d)

(e)

Figure: Visualization of the inferred $\{\Phi^{(1)}, \cdots, \Phi^{(T)}\}$ with $K_{1max} = 400$. All Φ's are projected to the first layer. From (a) to (e): $\Phi^{(1)}$, $\Phi^{(1)}\Phi^{(2)}$, $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}$, $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}\Phi^{(4)}$, $\Phi^{(1)}\Phi^{(2)}\Phi^{(3)}\Phi^{(4)}\Phi^{(5)}$.

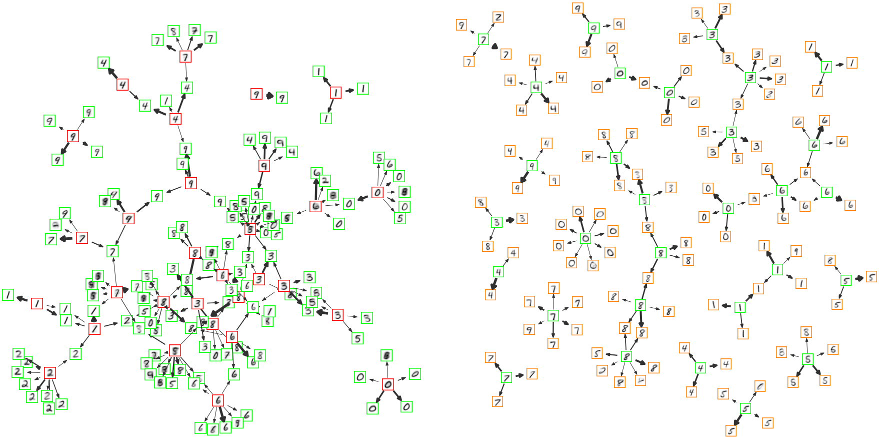


Figure: Left: layers 5 to 4; Right: layers 4 to 3

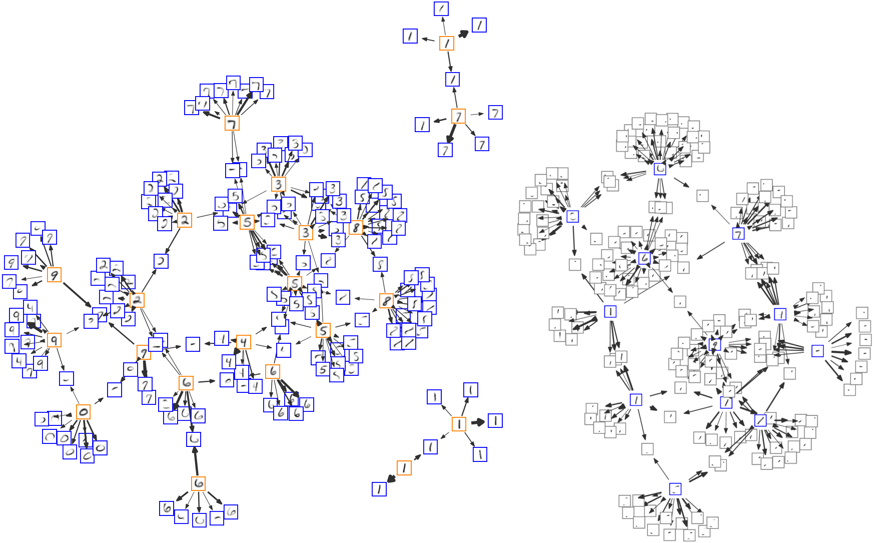


Figure: Left: layers 3 to 2; Right: layers 2 to 1

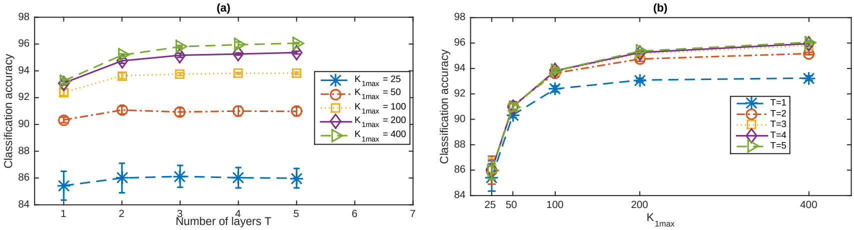


Figure: Left: Classification accuracy (%) of the PRG-GBNs with various K_{1max} as a function of $T_{1\,\max}$. Right: Classification accuracy (%) of the PRG-GBNs with various depths as a function of K_{1max}. ($T_{max} \in \{1, \cdots, 5\}$).

Conclusions

- The Poisson gamma belief network is proposed to extract a multilayer representation for high-dimensional count vectors.
- An upward-downward Gibbs sampler is used to jointly train all layers.
- A layer-wise training strategy is used to automatically infer the network structure.
- Extension to modeling binary and nonnegative real data.
- Understanding the data by examining the features of different layers and their relationships using the structure of the network.
- For big data problems, in practice one may rarely has a sufficient budget to allow the first-layer width to grow without bound, thus it is natural to consider a deep network that can use a multilayer deep representation to better allocate its resource and increase its representation power with limited computational power.
- Our algorithm provides a natural solution to achieve a good compromise between the width of each layer and the depth of the network.