Bayesian Rose Trees

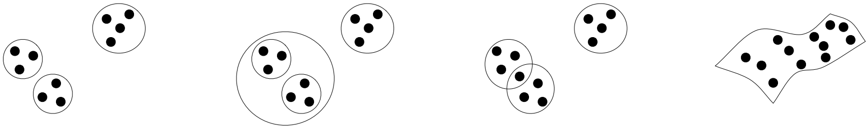
Charles Blundell and Katherine Heller
Yee Whye Teh

Gatsby Computational Neuroscience Unit
University College London

May 2010
CRiSM

Learning and Representational Structures

- Clustering.
- Hierarchical representations with trees.
- Overlapping clusters.
- Low dimensional embeddings.
- Distributed representations with multiple latent variables.

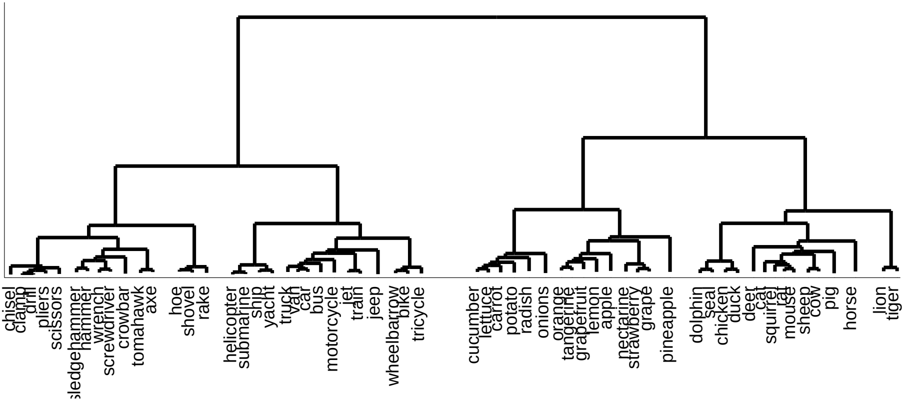


Psychological Objects and Features

apple	axe	bike	bus	car
carrot	cat	chicken	chisel	clamp
cow	crowbar	cucumber	deer	dolphin
drill	duck	grape	grapefruit	hammer
helicopter	hoe	horse	jeep	jet
lemon	lettuce	lion	motorcycle	mouse
nectarine	onions	orange	pig	pineapple
pliers	potato	radish	rake	rat
scissors	screwdriver	seal	sheep	ship
shovel	sledgehammer	squirrel	strawberry	submarine
tangerine	tiger	tomahawk	train	tricycle
truck	van	wheelbarrow	wrench	yacht

a fruit	a mammal	a tool	a vegetable	a vehicle
a weapon	an animal	beh - eats	beh - flies	beh - roars
beh - swims	eaten in salads	found in toolboxes	grows in Florida	grows in gardens
grows on trees	grows underground	has 2 wheels	has 4 legs	has 4 wheels
has a blade	has a handle	has a head	has a long handle	has a mane
has a metal head	has a tail	has a wooden handle	has an end	has an engine
has an inside	has doors	has eyes	has fur	has green leaves
has handles	has leaves	has legs	has peel	has propellers
has sections	has seeds	has skin	has teeth	has vitamin C
has wheels	has whiskers	has wings	hunted by people	is black
is brown	is citrus	is crunchy	is cute	is dangerous
is domestic	is edible	is fast	is ferocious	is green
is grey	is heavy	is juicy	is large	is long
is loud	is nutritious	is orange	is red	is round
is sharp	is small	is smooth	is white	is yellow
lives in wilderness	lives on farms	made of metal	made of wood	requires crews
requires drivers	requires gasoline	tastes good	tastes sour	tastes sweet
used by riding	used for cargo	used for carpentry	used for construction	used for cruising
used for digging	used for gardening	used for juice	used for loosening	used for passengers
used for pulling	used for tightening	used for transportation	used for turning	used on water

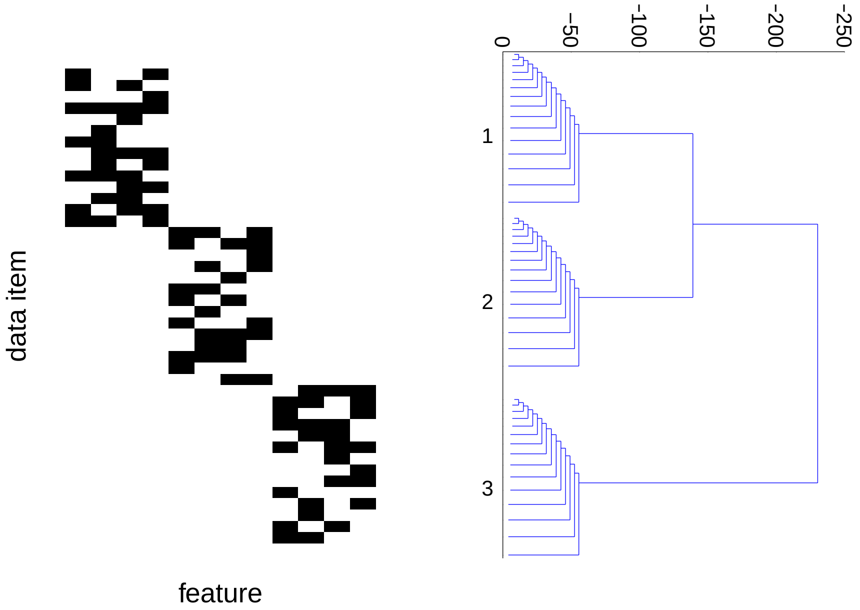
Psychological Objects and Features



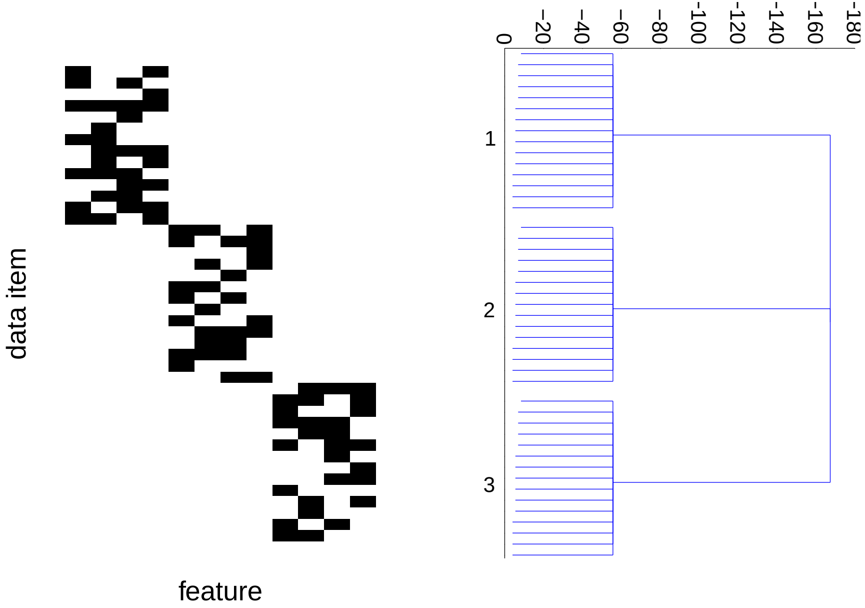
Hierarchical Clustering

- Linkage algorithms.
- Maximum likelihood, MAP, maximum parsimony [Vinokourov and Girolami 2000, Segal and Koller 2002, Friedman 2003].
- Bayesian hierarchical clustering (BHC) [Heller and Ghahramani 2005].
- Even more Bayesian models [Williams 2000, Neal 2003, Teh et al. 2008].
- Phylogenetics [Felsenstein 2003].

Non-binary Hierarchical Clusterings



Non-binary Hierarchical Clusterings

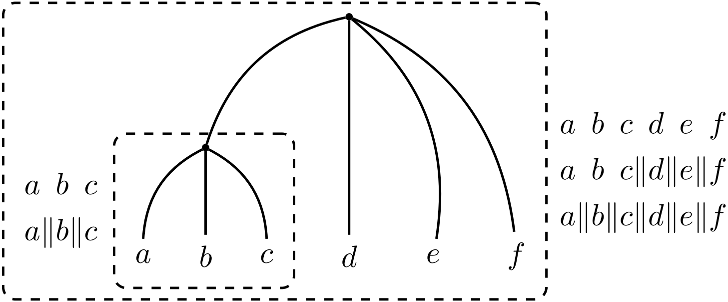


Bayesian Rose Trees

- Allows for non-binary trees if this is supported by data.
- Computational efficiency.

- Likelihood-based, probabilistic approach.
- most likely tree should offer a simple explanation of the data.

Tree-Consistent Partitions



An internal node means: Data at its leaves are more similar.

Each internal node denotes:

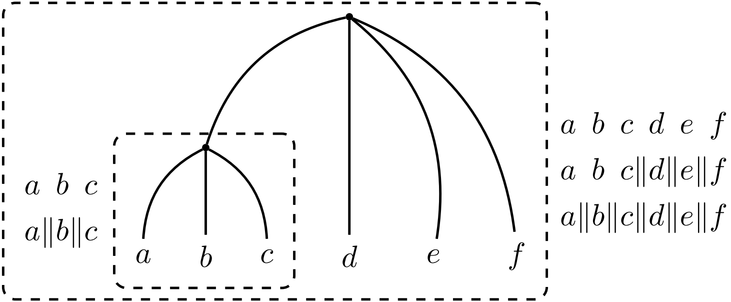
1. a cluster of its leaves
2. its children further partition the cluster into smaller subclusters.

A Bayesian rose tree represents a set of partitions of the data.

$$\mathrm{part}(T) = \{\mathrm{leaves}(T)\} \cup \{e_1\|e_2\|e_3\| \cdots : T_k \in \mathrm{ch}(T), e_k \in \mathrm{part}(T_k)\}$$

[Heller and Ghahramani 2005]

Tree-Consistent Partitions



An internal node means: Data at its leaves are more similar.
Each internal node denotes:

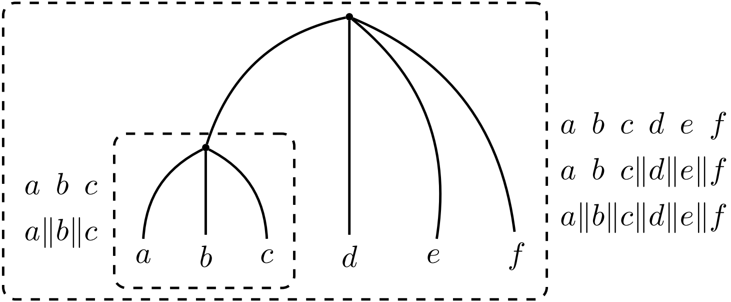
1. a cluster of its leaves
2. its children further partition the cluster into smaller subclusters.

A Bayesian rose tree represents a set of partitions of the data.

$$\text{part}(T) = \{\text{leaves}(T)\} \cup \{e_1 \| e_2 \| e_3 \| \cdots : T_k \in \text{ch}(T), e_k \in \text{part}(T_k)\}$$

[Heller and Ghahramani 2005]

Tree-Consistent Partitions



An internal node means: Data at its leaves are more similar.
Each internal node denotes:

1. a cluster of its leaves
2. its children further partition the cluster into smaller subclusters.

A Bayesian rose tree represents a set of partitions of the data.

$$\mathrm{part}(T) = \{\mathrm{leaves}(T)\} \cup \{e_1 \| e_2 \| e_3 \| \cdots : T_k \in \mathrm{ch}(T), e_k \in \mathrm{part}(T_k)\}$$

[Heller and Ghahramani 2005]

Likelihood of Clusters, Partitions and Trees

Cluster: *a b c*||*d*||*e*||*f*

A cluster is a set of data items. We use an exponential family distribution to model the cluster:

$$p(\mathcal{D}|\theta) = \exp\left(\theta^\top \sum_{x \in \mathcal{D}} s(x) - |\mathcal{D}|A(\theta)\right)$$

Using a conjugate prior for θ, we can marginalize out θ:

$$q(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$$

Example: Product of Beta-Bernoulli's:

$$q(\mathcal{D}) = \prod_{i=1}^{d} p(\mathcal{D}_i|\alpha_i, \beta_i) = \prod_{i=1}^{d} \frac{\mathrm{Beta}(\alpha_i + n_i^{\mathcal{D}}, \beta_i + N^{\mathcal{D}} - n_i^{\mathcal{D}})}{\mathrm{Beta}(\alpha_i, \beta_i)}$$

Likelihood of Clusters, Partitions and Trees

Partition: $a\ b\ c \| d \| e \| f$

A partition is a separation of data set into clusters. We model each cluster independently, so the likelihood of a partition is:

$$r(\{\mathcal{D}_1 \| \mathcal{D}_2 \| \ldots\}) = \prod_j q(\mathcal{D}_j)$$

Example:

$$r(a\ b\ c \| d \| e \| f) = q(a\ b\ c) q(d) q(e) q(f)$$

Likelihood of Clusters, Partitions and Trees

A tree is treated as a mixture of partitions. The likelihood of a tree will be a convex combination of partition likelihoods:

$$s(T) = \sum_{P \in \text{part}(T)} m_T(P) r(P)$$

Example:

$$\begin{aligned} s(T) = & m_T(a\ b\ c\ d\ e\ f) r(a\ b\ c\ d\ e\ f) + \\ & m_T(a\ b\ c \| d \| e \| f) r(a\ b\ c \| d \| e \| f) + \\ & m_T(a \| b \| c \| d \| e \| f) r(a \| b \| c \| d \| e \| f) \end{aligned}$$

Likelihood of Clusters, Partitions and Trees

Tree: $\{a\ b\ c\ d\ \ e\ f, a\ b\ c \| d \| e \| f, a \| b \| c \| d \| e \| f\}$

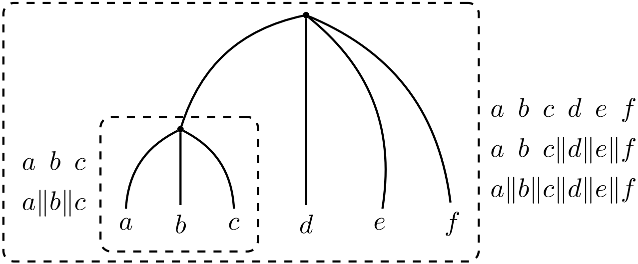
To make computations tractable, we will define the tree likelihood in a recursive fashion:

$$
\begin{aligned}
s(T) &= \sum_{P \in \mathrm{part}(T)} m_T(P) r(P) \\
&= \pi_T \underbrace{q(\mathrm{leaves}(T))}_{\text{cluster of leaves}} + (1 - \pi_T) \underbrace{\prod_{T_i \in \mathrm{ch}(T)} s(T_i)}_{\text{partitions of children}}
\end{aligned}
$$

Likelihood of Clusters, Partitions and Trees

Tree: $\{a\ b\ c\ d\ e\ f, a\ b\ c\|d\|e\|f, a\|b\|c\|d\|e\|f\}$

Example:

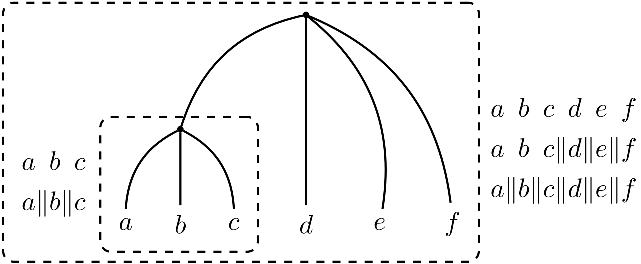


$$
\begin{aligned}
s(T_{abc}) &= \pi_{abc} q(\mathcal{D}_{abc}) + (1-\pi_{abc}) s(T_a) s(T_b) s(T_c) \\
&= \pi_{abc} q(\mathcal{D}_{abc}) + (1-\pi_{abc}) q(x_a) q(x_b) q(x_c) \\
s(T_{abcdef}) &= \pi_{abcdef} q(\mathcal{D}_{abcdef}) + (1-\pi_{abcdef}) s(T_{abc}) q(x_d) q(x_e) q(x_f) \\
&= \pi_{abcdef} q(\mathcal{D}_{abcdef}) + \\
&\quad (1-\pi_{abcdef}) \pi_{abc} q(\mathcal{D}_{abc}) q(x_d) q(x_e) q(x_f) + \\
&\quad (1-\pi_{abcdef})(1-\pi_{abc}) q(x_a) q(x_b) q(x_c) q(x_d) q(x_e) q(x_f)
\end{aligned}
$$

Likelihood of Clusters, Partitions and Trees

Tree: $\{a\ b\ c\ d\ e\ f, a\ b\ c\|d\|e\|f, a\|b\|c\|d\|e\|f\}$

Example:



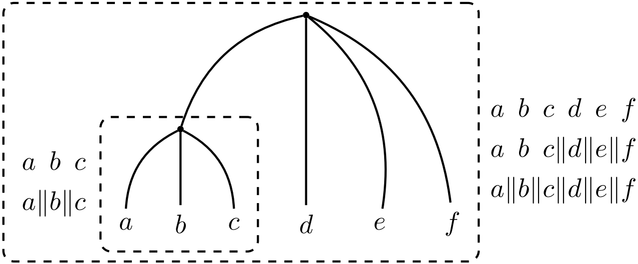
$$
\begin{aligned}
s(T_{abc}) =& \pi_{abc} q(\mathcal{D}_{abc}) + (1 - \pi_{abc}) s(T_a) s(T_b) s(T_c) \\
=& \pi_{abc} q(\mathcal{D}_{abc}) + (1 - \pi_{abc}) q(x_a) q(x_b) q(x_c)
\end{aligned}
$$

$$
\begin{aligned}
s(T_{abcdef}) =& \pi_{abcdef} q(\mathcal{D}_{abcdef}) + (1 - \pi_{abcdef}) s(T_{abc}) q(x_d) q(x_e) q(x_f) \\
=& \pi_{abcdef} q(\mathcal{D}_{abcdef}) + \\
& (1 - \pi_{abcdef}) \pi_{abc} q(\mathcal{D}_{abc}) q(x_d) q(x_e) q(x_f) + \\
& (1 - \pi_{abcdef})(1 - \pi_{abc}) q(x_a) q(x_b) q(x_c) q(x_d) q(x_e) q(x_f)
\end{aligned}
$$

Likelihood of Clusters, Partitions and Trees

Tree: $\{a\ b\ c\ d\ e\ f, a\ b\ c \| d \| e \| f, a \| b \| c \| d \| e \| f\}$

Example:

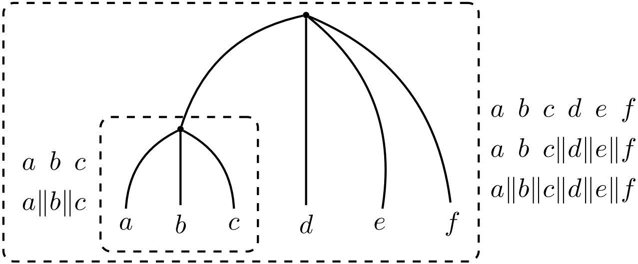


$$
\begin{aligned}
s(T_{abc}) =& \pi_{abc} q(\mathcal{D}_{abc}) + (1 - \pi_{abc}) s(T_a) s(T_b) s(T_c) \\
=& \pi_{abc} q(\mathcal{D}_{abc}) + (1 - \pi_{abc}) q(x_a) q(x_b) q(x_c) \\
s(T_{abcdef}) =& \pi_{abcdef} q(\mathcal{D}_{abcdef}) + (1 - \pi_{abcdef}) s(T_{abc}) q(x_d) q(x_e) q(x_f) \\
=& \pi_{abcdef} q(\mathcal{D}_{abcdef}) + \\
& (1 - \pi_{abcdef}) \pi_{abc} q(\mathcal{D}_{abc}) q(x_d) q(x_e) q(x_f) + \\
& (1 - \pi_{abcdef})(1 - \pi_{abc}) q(x_a) q(x_b) q(x_c) q(x_d) q(x_e) q(x_f)
\end{aligned}
$$

Likelihood of Clusters, Partitions and Trees

Tree: $\{a\ b\ c\ d\ e\ f, a\ b\ c \| d \| e \| f, a \| b \| c \| d \| e \| f\}$

Example:



$$
\begin{aligned}
s(T_{abc}) =& \pi_{abc} q(\mathcal{D}_{abc}) + (1-\pi_{abc}) s(T_a) s(T_b) s(T_c) \\
=& \pi_{abc} q(\mathcal{D}_{abc}) + (1-\pi_{abc}) q(x_a) q(x_b) q(x_c) \\
s(T_{abcdef}) =& \pi_{abcdef} q(\mathcal{D}_{abcdef}) + (1-\pi_{abcdef}) s(T_{abc}) q(x_d) q(x_e) q(x_f) \\
=& \pi_{abcdef} q(\mathcal{D}_{abcdef}) + \\
& (1-\pi_{abcdef}) \pi_{abc} q(\mathcal{D}_{abc}) q(x_d) q(x_e) q(x_f) + \\
& (1-\pi_{abcdef})(1-\pi_{abc}) q(x_a) q(x_b) q(x_c) q(x_d) q(x_e) q(x_f)
\end{aligned}
$$

An End to Needless Cascades

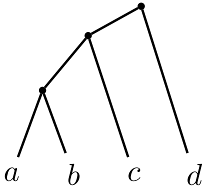
Define mixing proportions with parameter $0 < \gamma < 1$:

$$\pi_T = 1 - (1-\gamma)^{|\mathrm{ch}(T)|-1}$$

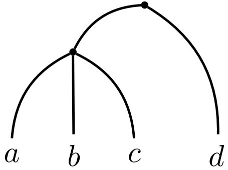
Suppose $r(a\ b\ c\|d) > r(a\ b\|c\|d)$ [other partitions of a, b, c as well].

$m(S,T)$	partition S
γ	$a\ b\ c\ d$
$(1-\gamma)\gamma$	$a\ b\ c\|d$
$(1-\gamma)(1-\gamma)\gamma$	$a\ b\|c\|d$
$(1-\gamma)(1-\gamma)(1-\gamma)$	$a\|b\|c\|d$

Cascading binary tree



Collapsed rose tree



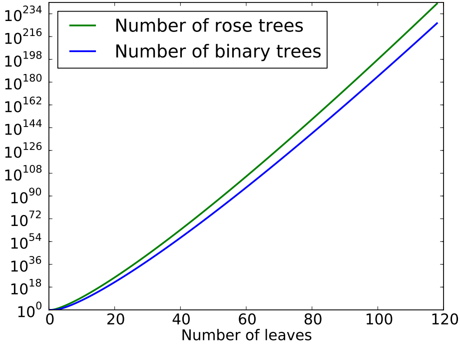
$m(S,T)$	partition S
γ	$a\ b\ c\ d$
$(1-\gamma)\left(1-(1-\gamma)^2\right)$	$a\ b\ c\|d$
$(1-\gamma)^3$	$a\|b\|c\|d$

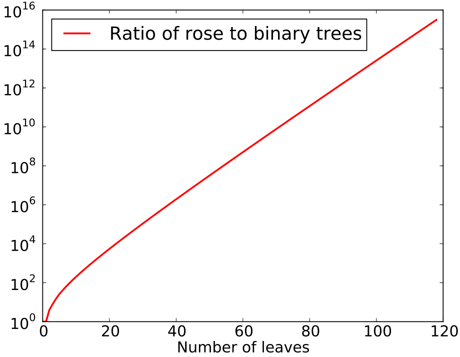
Complexity of Maximising $s(\mathcal{D}|T)$

There are too many rose trees T for an exhausitive search for the highest $s(T)$.

With L leaves there are:

Binary trees	$2^{O(L \log L)}$
Rose trees	$2^{O(L \log L + L)}$





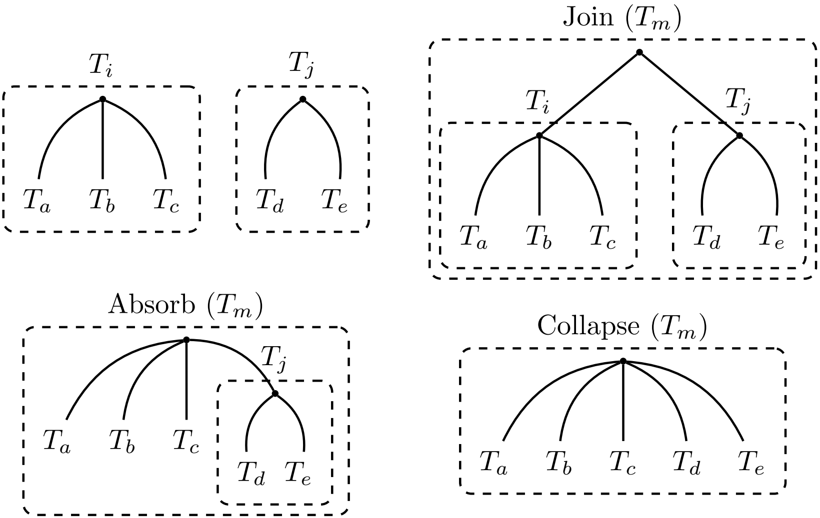
Construction by Greedy Model Selection

1. Let $T_i = \{x_i\} \, \forall i$.
2. For every ordered pair of trees (T_i, T_j) and possible merge operation producing tree T_m, pick the T_m with the largest Bayes factor:

$$\log \frac{s(T_m)}{s(T_i)s(T_j)}$$

3. Merge T_i, T_j into T_m.
4. Repeat 2 and 3 until one tree remains.

Merging Operations



Bayesian Hierarchical Clustering

Relationship between BRT and BHC:

- BHC produces binary trees; BRT can produce non-binary trees.
- BRT and one version of BHC interpret trees as mixtures over partitions.
- In other version, BHC interpreted as approximate inference in a DP mixture:
  - Uses a different π_T related to DP clustering prior.
  - BHC includes many partitions in its model as this encourages a tighter bound on the marginal probability under the DP mixture.
  - Unfortunately this leads to overly complicated models with many more partitions than necessary.
  - We found that this tends to produce trees with inferior likelihoods.

[Heller and Ghahramani 2005]

Bayesian Hierarchical Clustering

Relationship between BRT and BHC:

- BHC produces binary trees; BRT can produce non-binary trees.
- BRT and one version of BHC interpret trees as mixtures over partitions.
- In other version, BHC interpreted as approximate inference in a DP mixture:
  - Uses a different π_T related to DP clustering prior.
  - BHC includes many partitions in its model as this encourages a tighter bound on the marginal probability under the DP mixture.
  - Unfortunately this leads to overly complicated models with many more partitions than necessary.
  - We found that this tends to produce trees with inferior likelihoods.

[Heller and Ghahramani 2005]

Bayesian Hierarchical Clustering

Relationship between BRT and BHC:

- BHC produces binary trees; BRT can produce non-binary trees.
- BRT and one version of BHC interpret trees as mixtures over partitions.
- In other version, BHC interpreted as approximate inference in a DP mixture:
  - Uses a different π_T related to DP clustering prior.
  - BHC includes many partitions in its model as this encourages a tighter bound on the marginal probability under the DP mixture.
  - Unfortunately this leads to overly complicated models with many more partitions than necessary.
  - We found that this tends to produce trees with inferior likelihoods.

[Heller and Ghahramani 2005]

Bayesian Hierarchical Clustering

Relationship between BRT and BHC:

- BHC produces binary trees; BRT can produce non-binary trees.
- BRT and one version of BHC interpret trees as mixtures over partitions.
- In other version, BHC interpreted as approximate inference in a DP mixture:
  - Uses a different π_T related to DP clustering prior.
  - BHC includes many partitions in its model as this encourages a tighter bound on the marginal probability under the DP mixture.
  - Unfortunately this leads to overly complicated models with many more partitions than necessary.
  - We found that this tends to produce trees with inferior likelihoods.

[Heller and Ghahramani 2005]

Bayesian Hierarchical Clustering

Relationship between BRT and BHC:

- BHC produces binary trees; BRT can produce non-binary trees.
- BRT and one version of BHC interpret trees as mixtures over partitions.
- In other version, BHC interpreted as approximate inference in a DP mixture:
  - Uses a different π_T related to DP clustering prior.
  - BHC includes many partitions in its model as this encourages a tighter bound on the marginal probability under the DP mixture.
  - Unfortunately this leads to overly complicated models with many more partitions than necessary.
  - We found that this tends to produce trees with inferior likelihoods.

[Heller and Ghahramani 2005]

Results (anecdotal)

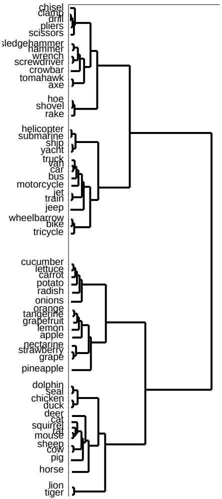
apple	axe	bike	bus	car
carrot	cat	chicken	chisel	clamp
cow	crowbar	cucumber	deer	dolphin
drill	duck	grape	grapefruit	hammer
helicopter	hoe	horse	jeep	jet
lemon	lettuce	lion	motorcycle	mouse
nectarine	onions	orange	pig	pineapple
pliers	potato	radish	rake	rat
scissors	screwdriver	seal	sheep	ship
shovel	sledgehammer	squirrel	strawberry	submarine
tangerine	tiger	tomahawk	train	tricycle
truck	van	wheelbarrow	wrench	yacht
a fruit	a mammal	a tool	a vegetable	a vehicle
a weapon	an animal	beh - eats	beh - flies	beh - roars
beh - swims	eaten in salads	found in toolboxes	grows in Florida	grows in gardens
grows on trees	grows underground	has 2 wheels	has 4 legs	has 4 wheels
has a blade	has a handle	has a head	has a long handle	has a mane
has a metal head	has a tail	has a wooden handle	has an end	has an engine
has an inside	has doors	has eyes	has fur	has green leaves
has handles	has leaves	has legs	has peel	has propellers
has sections	has seeds	has skin	has teeth	has vitamin C
has wheels	has whiskers	has wings	hunted by people	is black
is brown	is citrus	is crunchy	is cute	is dangerous
is domestic	is edible	is fast	is ferocious	is green
is grey	is heavy	is juicy	is large	is long
is loud	is nutritious	is orange	is red	is round
is sharp	is small	is smooth	is white	is yellow
lives in wilderness	lives on farms	made of metal	made of wood	requires crews
requires drivers	requires gasoline	tastes good	tastes sour	tastes sweet
used by riding	used for cargo	used for carpentry	used for construction	used for cruising
used for digging	used for gardening	used for juice	used for loosening	used for passengers
used for pulling	used for tightening	used for transportation	used for turning	used on water

[Cree and McRae 2003]

Results (anecdotal)

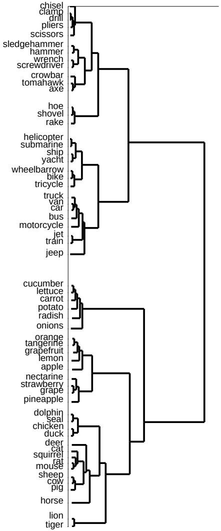
BHC (DP)

log likelihood -1418
468,980,051 partitions



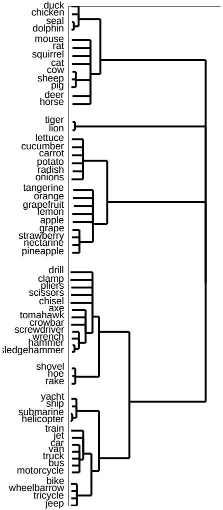
BHC (fixed)

log likelihood -1266
908,188,506 partitions

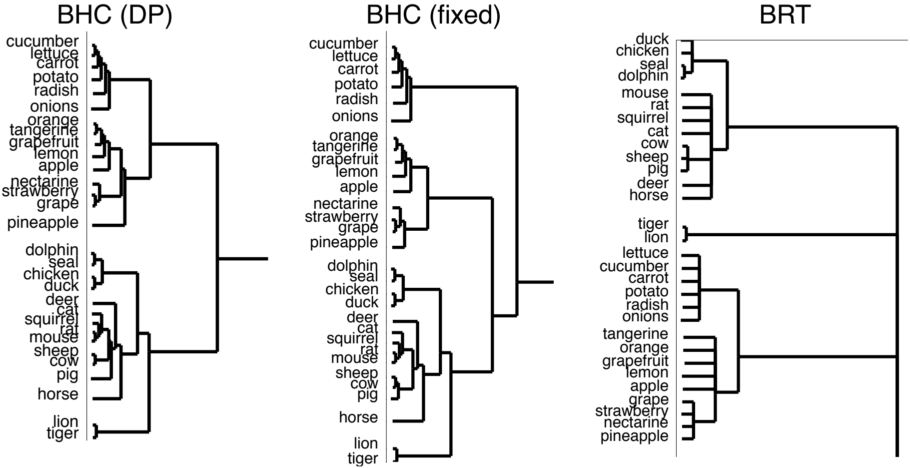


BRT

log likelihood -1258
1,441 partitions

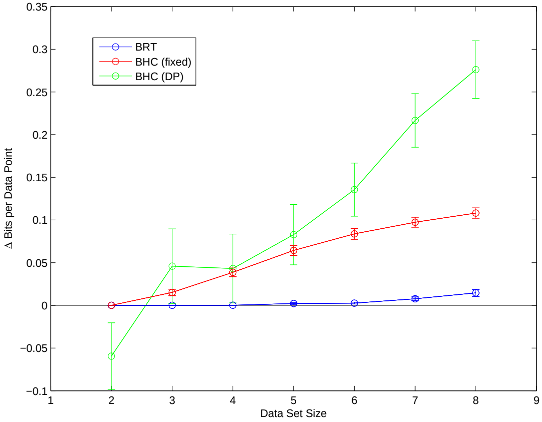


Results (anecdotal)



Results (quantitative)

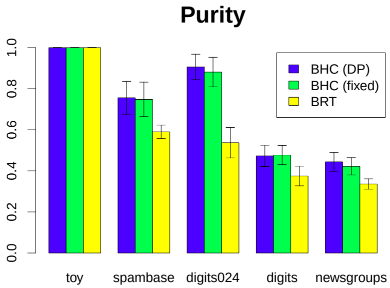
Does greedy search find the best tree?

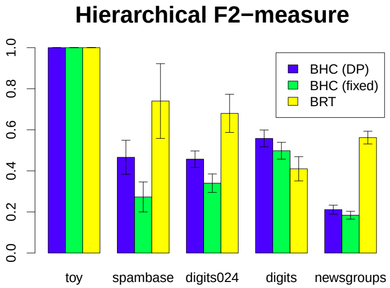


Results (quantitative)

Log likelihood:

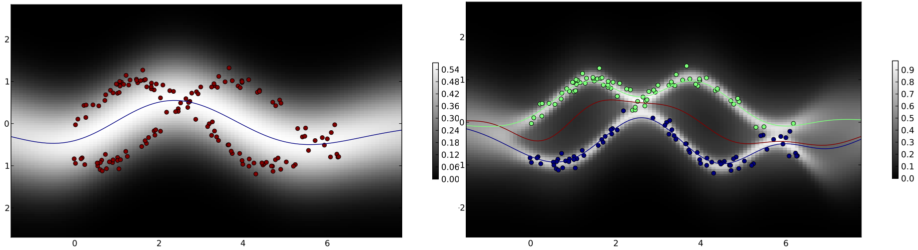
Data set	BHC (DP)	BHC (fixed)	BRT
toy	-230 ± 0	-169.4 ± 0	$-\mathbf{167} \pm 0$
spambase	-2354 ± 4.7	-2000 ± 4.5	$-\mathbf{1991} \pm 4.5$
digits024	-4154 ± 5.2	-3759 ± 4.6	$-\mathbf{3748} \pm 4.6$
digits	-4429 ± 3.3	-3966 ± 3.1	$-\mathbf{3954} \pm 3.1$
newsgroups	-11602 ± 104	-10833 ± 106	-10827 ± 105





Mixtures of Gaussian Process Experts

Mixtures of GPs are simple ways to construct nonparametric density regression models. A type of dependent Dirichlet process mixtures. MCMC inference can be very time consuming.



[MacEachern 1999, Rasmussen and Ghahramani 2002, Müller et al. 2010]

Discussion

A hierarchical clustering model that:

- allows arbitary branching structure.
- uses this flexibility to find simpler models better explaining data.
- Finding good trees in $O(L^2 \log L)$ time (same as BHC).

To explore more computationally efficient algorithms.
There are other (unexplored wrt hierarchical clustering) models of non-binary trees such as Λ-coalescents and Gibbs fragmentation trees.
[Pitman 1999, McCullagh et al. 2008]

Thanks

References I

Cree, G. S. and McRae, K. (2003).
Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese and cello (and many other such conc rete nouns).
Journal of Experimental Psychology: General, 132(2):163–201.

Felsenstein, J. (2003).
Inferring Phylogenies.
Sinauer Associates.

Friedman, N. (2003).
Pcluster: Probabilistic agglomerative clustering of gene expression profiles.
Technical Report Technical Report 2003-80, Hebrew University.

Heller, K. A. and Ghahramani, Z. (2005).
Bayesian hierarchical clustering.
In *Proceedings of the International Conference on Machine Learning*, volume 22.

MacEachern, S. (1999).
Dependent nonparametric processes.
In *Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association.

McCullagh, P., Pitman, J., and Winkel, M. (2008).
Gibbs fragmentation trees.
Bernoulli, 14(4):988–1002.

Müller, P., Quintana, F. A., and Rosner, G. L. (2010).
A product partition model with regression on covariates.
http://www.mat.puc.cl/ quintana/publications/publications.html.

Neal, R. M. (2003).
Density modeling and clustering using Dirichlet diffusion trees.
In *Bayesian Statistics*, volume 7, pages 619–629.

Pitman, J. (1999).
Coalescents with multiple collisions.
Annals of Probability, 27:1870–1902.

References II

Rasmussen, C. E. and Ghahramani, Z. (2002).
Infinite mixtures of Gaussian process experts.
In *Advances in Neural Information Processing Systems*, volume 14.

Segal, E. and Koller, D. (2002).
Probabilistic hierarchical clustering for biological data.
In *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*, pages 273–280, New York, NY, USA. ACM.

Teh, Y. W., Daume III, H., and Roy, D. M. (2008).
Bayesian agglomerative clustering with coalescents.
In *Advances in Neural Information Processing Systems*, volume 20, pages 1473–1480.

Vinokourov, A. and Girolami, M. (2000).
A probabilistic hierarchical clustering method for organizing collections of text documents.
Pattern Recognition, International Conference on, 2:2182.

Williams, C. K. I. (2000).
A MCMC approach to hierarchical mixture modelling.
In *Advances in Neural Information Processing Systems*, volume 12.

Animal Features

in tiger/lion?	**not in tiger/lion?**	**maybe in both?**	**in both?**
is fast	beh - flies	lives in wilderness	has teeth
has a mane	has wings	hunted by people	has eyes
roars	swims		has fur
is ferocious	is domestic		has a tail
is dangerous	is edible		has 4 legs
	lives on farms		eats
	is cute		an animal
	taste good		a mammal
			has whiskers
			has skin