Sparsity?
A Bayesian view

Zoubin Ghahramani
Department of Engineering
University of Cambridge

SPARS Conference
Cambridge, July 2015

Sparsity

Many people are interested in sparsity. Why?

Sparsity

Many people are interested in sparsity. Why?

- **Real world statistics** often have sparsity
  - Natural statistics of images, sounds, and other signals
  - Compressed sensing, independent components analysis
  - Feature/variable selection in e.g. gene expression data
  - The structure of many natural graphs is sparse.

Sparsity

Many people are interested in sparsity. Why?

- **Real world statistics** often have sparsity
  - Natural statistics of images, sounds, and other signals
  - Compressed sensing, independent components analysis
  - Feature/variable selection in e.g. gene expression data
  - The structure of many natural graphs is sparse.

- Sparsity assumptions can be a very good **regulariser** to avoid overfitting
  - Feature selection
  - SVMs
  - Data dependent generalisation bounds

Sparsity

Many people are interested in sparsity. Why?

- **Real world statistics** often have sparsity
  - Natural statistics of images, sounds, and other signals
  - Compressed sensing, independent components analysis
  - Feature/variable selection in e.g. gene expression data
  - The structure of many natural graphs is sparse.
- Sparsity assumptions can be a very good **regulariser** to avoid overfitting
  - Feature selection
  - SVMs
  - Data dependent generalisation bounds
- Sparsity can be exploited for **fast computation**
  - Matrix factorisation for recommender systems
  - Sparse solutions in kernel machines

Outline

- The Bayesian view
- Bayesian nonparametrics and sparsity
- Sparse factor models

Part I: The Bayesian view

Probabilistic Modelling

- A model describes data that one could observe from a system

- If we use the mathematics of probability theory to express all forms of uncertainty and noise associated with our model...

- ...then *inverse probability* (i.e. Bayes rule) allows us to infer unknown quantities, adapt our models, make predictions and learn from data.

Probabilistic Modelling

Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x,y)$

Product rule: $P(x,y) = P(x)P(y|x)$

$$P(\theta|\mathcal{D},m) = \frac{P(\mathcal{D}|\theta,m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(\mathcal{D}\|\theta,m)$	likelihood of parameters θ in model m
$P(\theta\|m)$	prior probability of θ
$P(\theta\|\mathcal{D},m)$	posterior of θ given data $\mathcal{D}$

Prediction:

$$P(x|\mathcal{D},m) = \int P(x|\theta,\mathcal{D},m)P(\theta|\mathcal{D},m)d\theta$$

Model Comparison:

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta,m)P(\theta|m)\,d\theta$$

Three Key Observations

1. Modelling and prediction require assumptions; the Bayesian approach uses probability theory to express uncertainty in all such assumptions

Three Key Observations

1. Modelling and prediction require assumptions; the Bayesian approach uses probability theory to express uncertainty in all such assumptions

2. Given these assumptions, the rest is applications of the sum and product rules, or approximations thereof

Three Key Observations

1. Modelling and prediction require assumptions; the Bayesian approach uses probability theory to express uncertainty in all such assumptions

2. Given these assumptions, the rest is applications of the sum and product rules, or approximations thereof

3. There is no "optimisation rule" in probability theory; optimisation is used either to approximate integration, or to make decisions under some loss

Part II: Bayesian nonparametrics and sparsity

Why Bayesian nonparametrics

- **Why Bayesian?**

  Simplicity (of the framework)

- **Why nonparametrics?**

  Complexity (of real world phenomena)

Parametric vs Nonparametric Models

- *Parametric models* assume some finite set of parameters θ. Given the parameters, future predictions, x, are independent of the observed data, $\mathcal{D}$:

  $$P(x|\theta, \mathcal{D}) = P(x|\theta)$$

  therefore θ capture everything there is to know about the data.

- So the complexity of the model is bounded even if the amount of data is unbounded. This makes them not very flexible.

- *Non-parametric models* assume that the data distribution cannot be defined in terms of such a finite set of parameters. But they can often be defined by assuming an *infinite dimensional* θ. Usually we think of θ as a *function*.

- The amount of information that θ can capture about the data $\mathcal{D}$ can grow as the amount of data grows. This makes them more flexible.

Overview of nonparametric models and uses

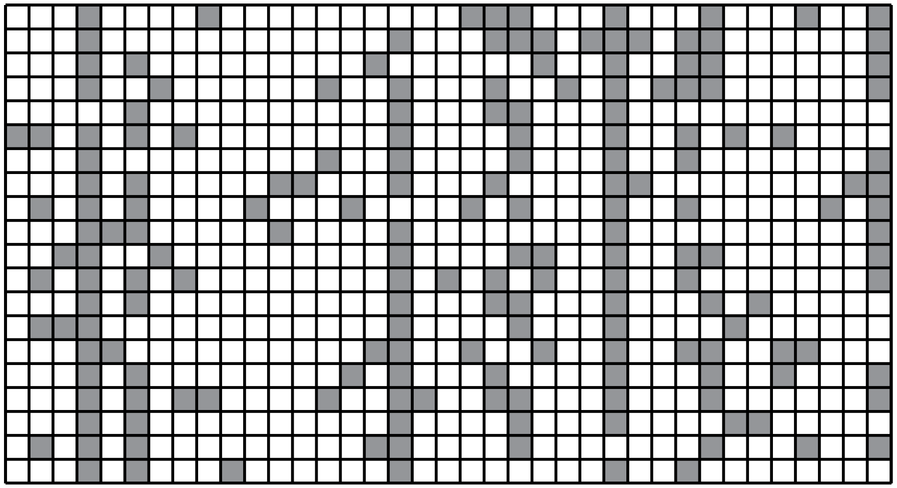
Bayesian nonparametrics has many uses.

Some modelling goals and *examples* of associated nonparametric Bayesian models:

Modelling goal	**Example process**
Distributions on functions	Gaussian process
Distributions on distributions	Dirichlet process
	Polya Tree
Clustering	Chinese restaurant process
	Pitman-Yor process
Hierarchical clustering	Dirichlet diffusion tree
	Kingman's coalescent
Sparse binary matrices	Indian buffet processes
Survival analysis	Beta processes
Distributions on measures	Completely random measures
...	...

Sparse Matrices

From finite to infinite sparse binary matrices



$z_{nk} = 1$ means object n has feature k:

$$z_{nk} \sim \text{Bernoulli}(\theta_k)$$

$$\theta_k \sim \text{Beta}(\alpha/K, 1)$$

- Note that $P(z_{nk} = 1|\alpha) = E(\theta_k) = \frac{\alpha/K}{\alpha/K+1}$, so as K grows larger the matrix gets sparser.

- So if $\mathbf{Z}$ is $N \times K$, the expected number of nonzero entries is $N\alpha/(1+\alpha/K) < N\alpha$.

- Even in the $K \to \infty$ limit, the matrix is expected to have a finite number of non-zero entries.

- $K \to \infty$ results in an Indian buffet process (IBP)[1]

[1] Naming inspired by analogy to "Chinese restaurant process" (CRP) from probability theory.

Modelling Data with Indian Buffet Processes

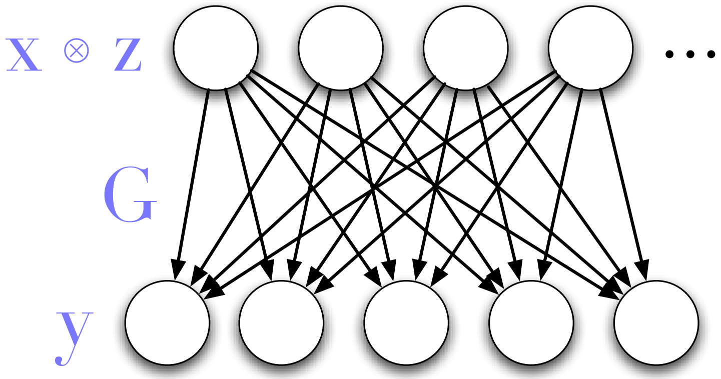
Latent variable model: let $\mathbf{X}$ be the $N \times D$ matrix of observed data, and $\mathbf{Z}$ be the $N \times K$ matrix of sparse binary latent features

$$P(\mathbf{X}, \mathbf{Z}|\alpha) = P(\mathbf{X}|\mathbf{Z})P(\mathbf{Z}|\alpha)$$

By combining the IBP with different likelihood functions we can get different kinds of models:

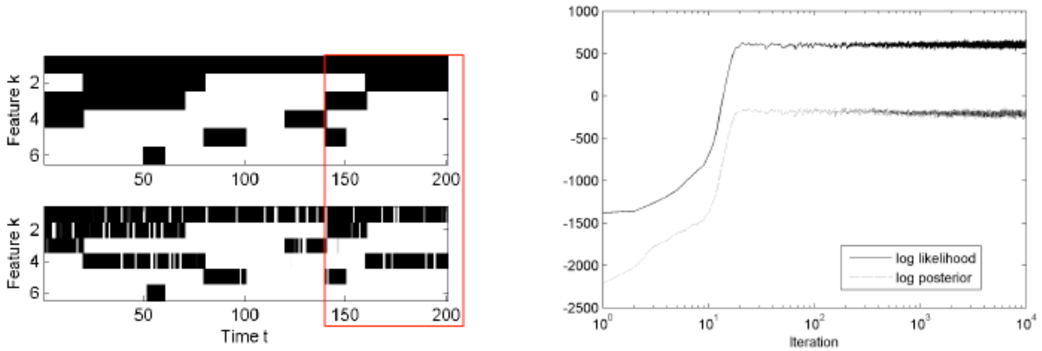
- Models for graph structures (w/ Wood, Griffiths, 2006; w/ Adams and Wallach, 2010)
- Models for protein complexes (w/ Chu, Wild, 2006)
- Models for choice behaviour (Görür & Rasmussen, 2006)
- Models for users in collaborative filtering (w/ Meeds, Roweis, Neal, 2007)
- Sparse latent trait, pPCA and ICA models (w/ Knowles, 2007, 2011)
- Models for overlapping clusters (w/ Heller, 2007)

Infinite Independent Components Analysis



Model: $\mathbf{Y} = \mathbf{G}(\mathbf{Z} \otimes \mathbf{X}) + \mathbf{E}$

where $\mathbf{Y}$ is the data matrix, $\mathbf{G}$ is the mixing matrix $\mathbf{Z} \sim \mathrm{IBP}(\alpha, \beta)$ is a mask matrix, $\mathbf{X}$ is heavy tailed sources and $\mathbf{E}$ is Gaussian noise.



(a) *Top:* True $\mathbf{Z}$. *Bottom:* Inferred $\mathbf{Z}$. Red box denotes test data.

(b) Plot of the log likelihood and posterior for the duration of the iICA$_2$ run.

Fig. 1. True and inferred $\mathbf{Z}$ and algorithm convergence.

(w/ David Knowles, 2007, 2011)

Infinite Sparse Factor Analysis

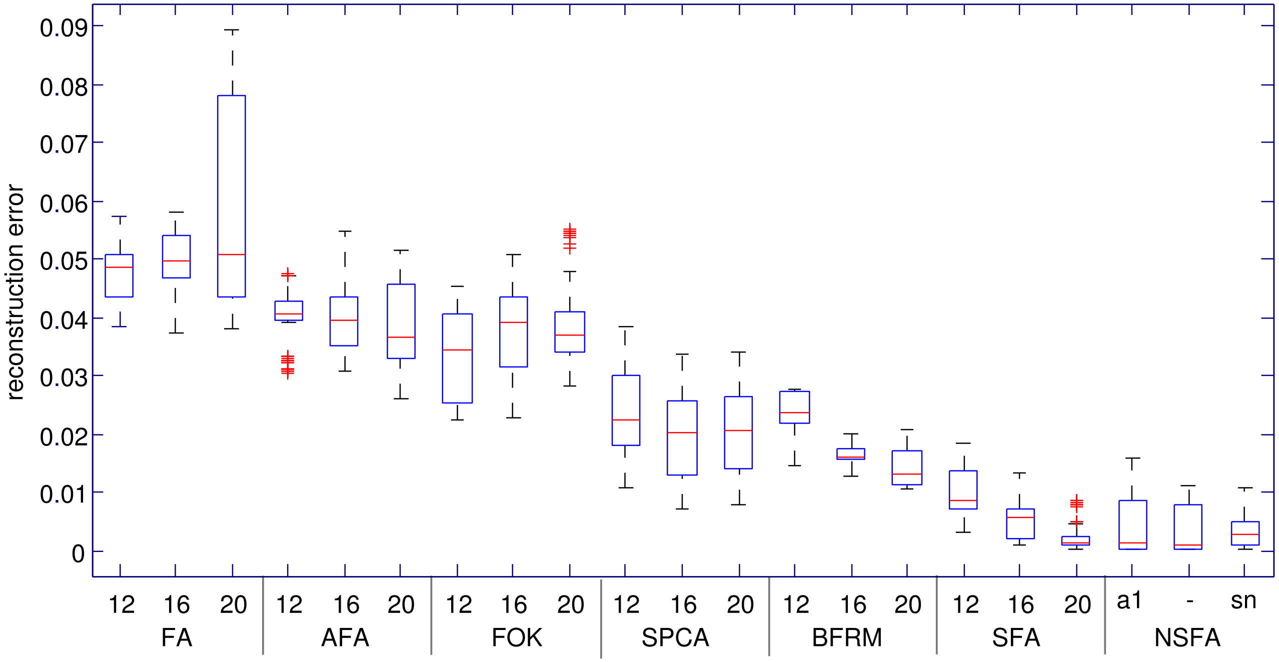


FIG. 4. *Boxplot of reconstruction errors for simulated data derived from the* E. Coli *connectivity matrix of Kao et al*. (2004). *Ten data sets were generated and the reconstruction error calculated for the last ten samples from each algorithm. Numbers refer to the number of latent factors used*, K. a1 *denotes fixing* $\alpha = 1$. sn *denotes sharing power between noise dimensions*.

- FA—Bayesian Factor Analysis; see, for example, Kaufman and Press (1973) or Rowe and Press (1998).
- AFA—Factor Analysis with ARD prior to determine active sources.
- FOK—The sparse Factor Analysis method of Fokoue (2004), Fevotte and Godsill (2006) and Archambeau and Bach (2009).
- SPCA—The Sparse PCA method of Zou, Hastie and Tibshirani (2004).
- BFRM—Bayesian Factor Regression Model of West et al. (2007).
- SFA—Sparse Factor Analysis, using the finite IBP.
- NSFA—The proposed model: Nonparametric Sparse Factor Analysis.

Part III: Sparse factor models

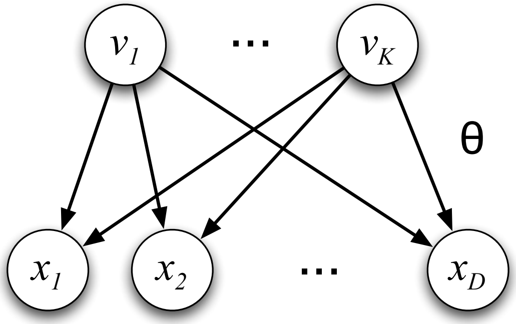
- Mohamed, S., Heller, K.A., and Ghahramani, Z. (2009) Bayesian Exponential Family PCA. In *Advances in Neural Information Processing Systems* **21**:1089–1096. Cambridge, MA: MIT Press.
- Mohamed, S., Heller, K.A., and Ghahramani, Z. (2012) Bayesian and L1 Approaches for Sparse Unsupervised Learning. ICML 2012.
- Mohamed, S., Heller, K. A. and Ghahramani, Z. (2014) Bayesian Approaches for Sparse Latent Variable Models: Reconsidering L_1 Sparsity. In Rish, I., Cecchi, G., Lozano, A. and Niculescu-Mizil (Eds.) *Practical Applications of Sparse Modeling*. MIT Press.

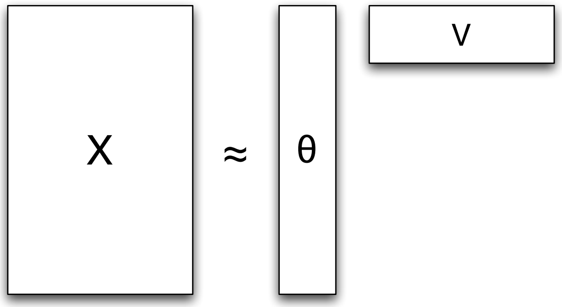
Factor Models and Matrix Factorization

Factor analysis and matrix factorization models have the following general form:

$$\mathbf{x}_n = \Theta \mathbf{v}_n + \mathbf{e}_n = \sum_k \boldsymbol{\theta}_k v_{nk} + \mathbf{e}_n$$

where $\mathbf{x}_n \in \mathbb{R}^D$ is a data vector, $\mathbf{v}_n \in \mathbb{R}^K$ is a vector of latent factors, $\Theta \in \mathbb{R}^{D \times K}$ is a matrix of parameters (factor loadings), and $\mathbf{e}_n$ is Gaussian noise.



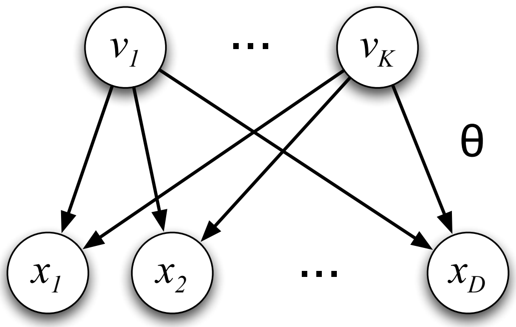


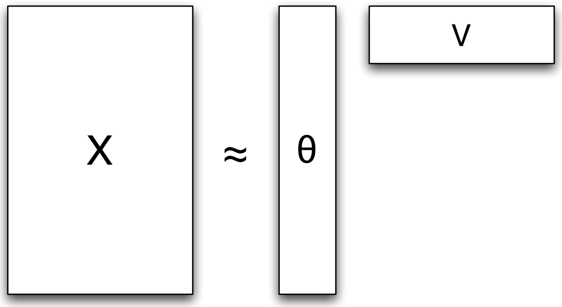
Factor Models and Matrix Factorization

Factor analysis and matrix factorization models have the following general form:

$$\mathbf{x}_n = \Theta \mathbf{v}_n + \mathbf{e}_n = \sum_k \boldsymbol{\theta}_k v_{nk} + \mathbf{e}_n$$

where $\mathbf{x}_n \in \mathbb{R}^D$ is a data vector, $\mathbf{v}_n \in \mathbb{R}^K$ is a vector of latent factors, $\Theta \in \mathbb{R}^{D \times K}$ is a matrix of parameters (factor loadings), and $\mathbf{e}_n$ is Gaussian noise.



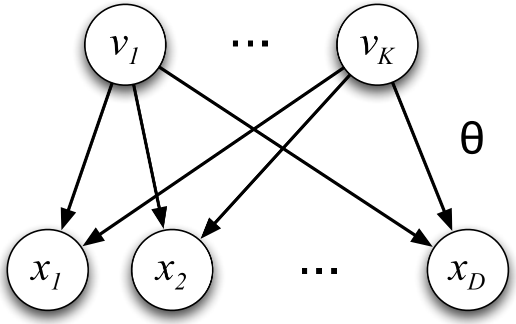


We can rewrite this in **matrix form** to more clearly see it as matrix factorization:

$$X_{D \times N} = \Theta_{D \times K} V_{K \times N} + E_{D \times N}$$

These models have been around for over 100 years (Spearman, 1904).
We are interested in *sparse* variants...

Sparse Factor Models



- Consider a **sparse factor model**:

$$\mathbf{x}_n = \Theta \mathbf{v}_n + \mathbf{e}_n = \sum_k \boldsymbol{\theta}_k v_{nk} + \mathbf{e}_n$$

where $\mathbf{x}_n$ is a data vector, $\mathbf{v}_n$ is a *sparse* vector of latent factors, Θ is a matrix of parameters,[2] and $\mathbf{e}_n$ is Gaussian noise.

- **Sparsity:** our "solution"[3] should have many of the elements of $v_{nk} = 0$

- Extension to general **exponential family distributions** for non-Gaussian $\mathbf{x}_n$:

$$\mathbf{x}_n \sim \mathrm{Expon}(\Theta \mathbf{v}_n)$$

This generalization can handle binary, count, discete, positive, and many other data types, and *combinations*! It is a sparse version of Exponential Family PCA.

[2] To keep things simple, we don't consider sparse Θ.

[3] There is a different notion of a solution under a Bayesian or optimization viewpoint.

Aside: Exponential Family

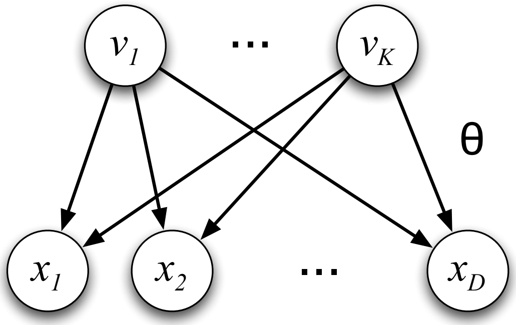
$p(x|\theta)$ in the **exponential family** if it can be written as:

$$p(x|\theta) = f(x)g(\theta)\exp\{\phi(\theta)^\top s(x)\}$$

ϕ	vector of *natural parameters*
$s(x)$	vector of *sufficient statistics*
f and g	positive functions of x and θ, respectively.

Examples include: Gaussian, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, categorical, Poisson, Wishart...

Sparse Factor Models: Optimization approach

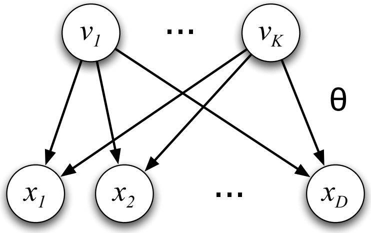


- The classical approach for inducing sparsity is to optimise a cost function/likelihood with *an L_1 regularizer on the elements of $\mathbf{v}_n$*.

$$\min_{V,\Theta} \sum_n \ell(\mathbf{x}_n, \Theta \mathbf{v}_n) + \alpha \|V\|_1 + \beta R(\Theta)$$

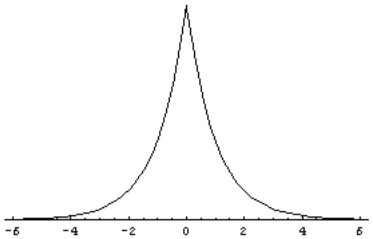
Is this a good idea?

Sparse Factor Models: Bayesian approach with Laplace priors



Exponential family likelihood as before:

$$\mathbf{x}_n \sim \mathrm{Expon}(\Theta \mathbf{v}_n)$$



Use conjugate prior for Θ, and Laplace prior on the elements of $\mathbf{v}_n$.

$$p(v_{nk}|\alpha) \propto \exp\{-\alpha|v_{nk}|\}$$

- **Maximum a posteriori (MAP)** in this model is equivalent to L_1 regularization

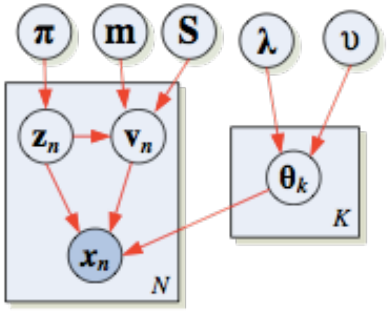
- We also explore doing full **Bayesian inference** by averaging (over V, Θ, etc).

- Other variants are non-negative v_{nk} etc...

Weak vs Strong Sparsity

- **Weak Sparsity**: L_1, or priors that have high density at 0
- **Strong Sparsity**: L_0, or priors that have probability *mass* at 0

Bayesian Sparse Factor Models: Spike and Slab Priors



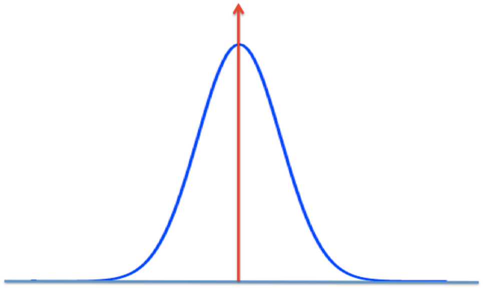
Exponential family likelihood as before:

$$\mathbf{x}_n \sim \mathrm{Expon}(\Theta \mathbf{v}_n)$$

Use conjugate prior for Θ, and spike and slab prior on the elements of $\mathbf{v}_n$:

$$v_{nk} = z_{nk}\, w_{nk} \qquad z_{nk} \sim \mathrm{Bern}(\pi_k) \qquad w_{nk} \sim \mathrm{Norm}(\mu_k, \sigma_k^2)$$

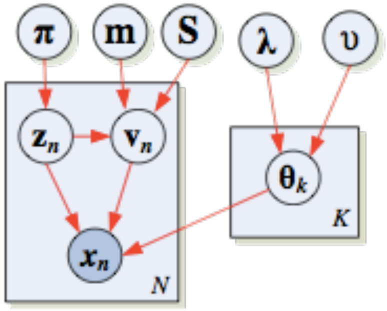
where z_{nk} is a binary indicator variable creating a *spike* (δ-function) at 0 with probability π_k, and w_{nk} is drawn from a *slab* distribution.



The spike and slab distribution encourages strong sparsity in the factors.

How does this compare to L_1 regularisation?

Inference and Learning

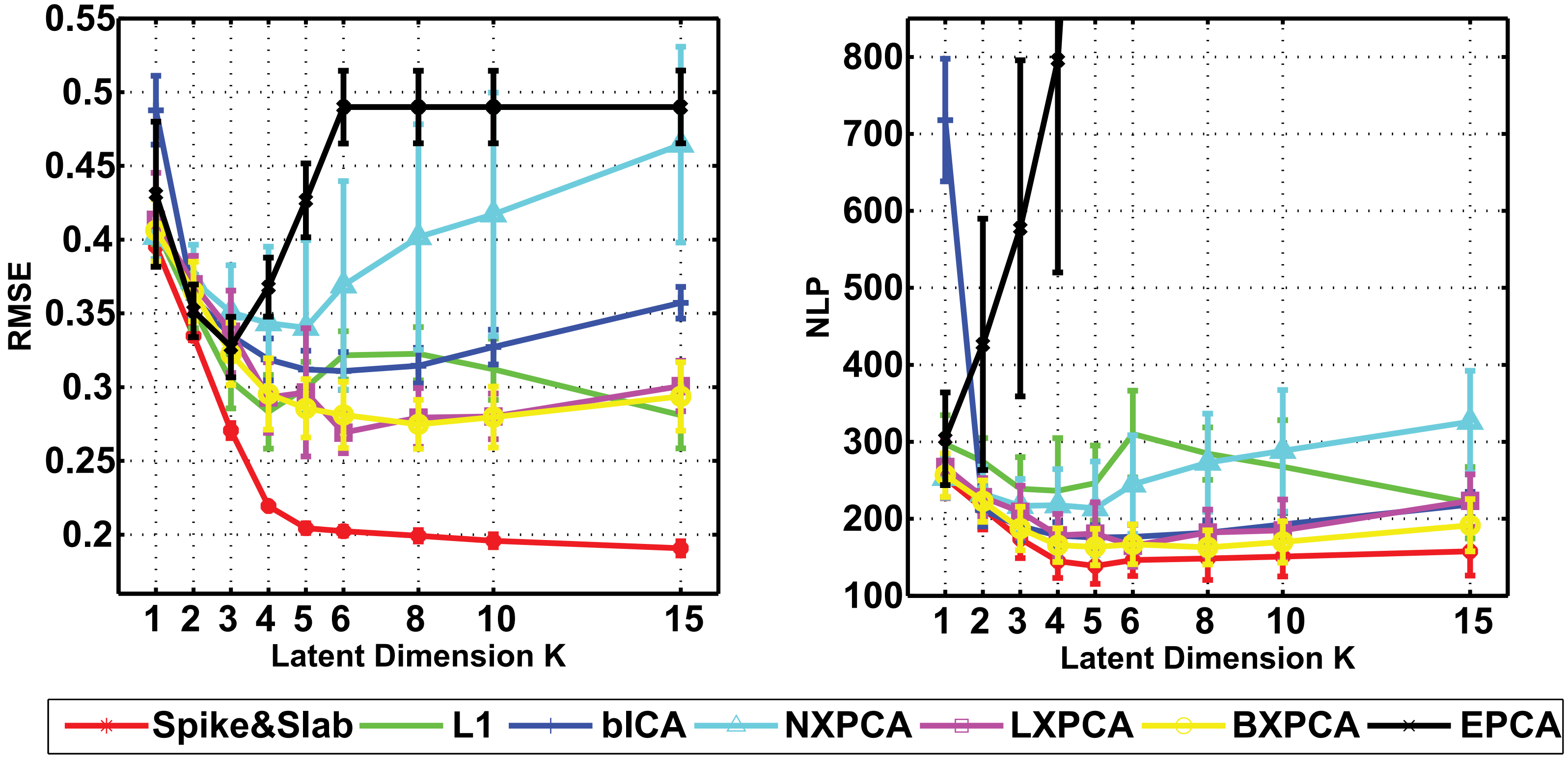


- **Strongly Sparse Bayesian Model** (`Spike&Slab`):
  Inference is done via MCMC, combining:
  - Slice sampling for Θ
  - Gibbs sampling for $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$
  - and Laplace marginalisation of the slab distribution to sample $\mathbf{Z}$

- **Weakly sparse Bayesian models** using Laplace (`LXPCA`) and Exponential non-negative (`NXPCA`) priors:
  - All variables are continuous so we use Hamiltonian Monte Carlo.

- **Regularised L_1 models** (`L1`):
  - cross validation to determine hyperparameters
  - fast L_1 projection method of Schmidt, Fung and Rosales (2007).

Bayesian Sparse Factor Models: Test Prediction Results



Data: artificial block images

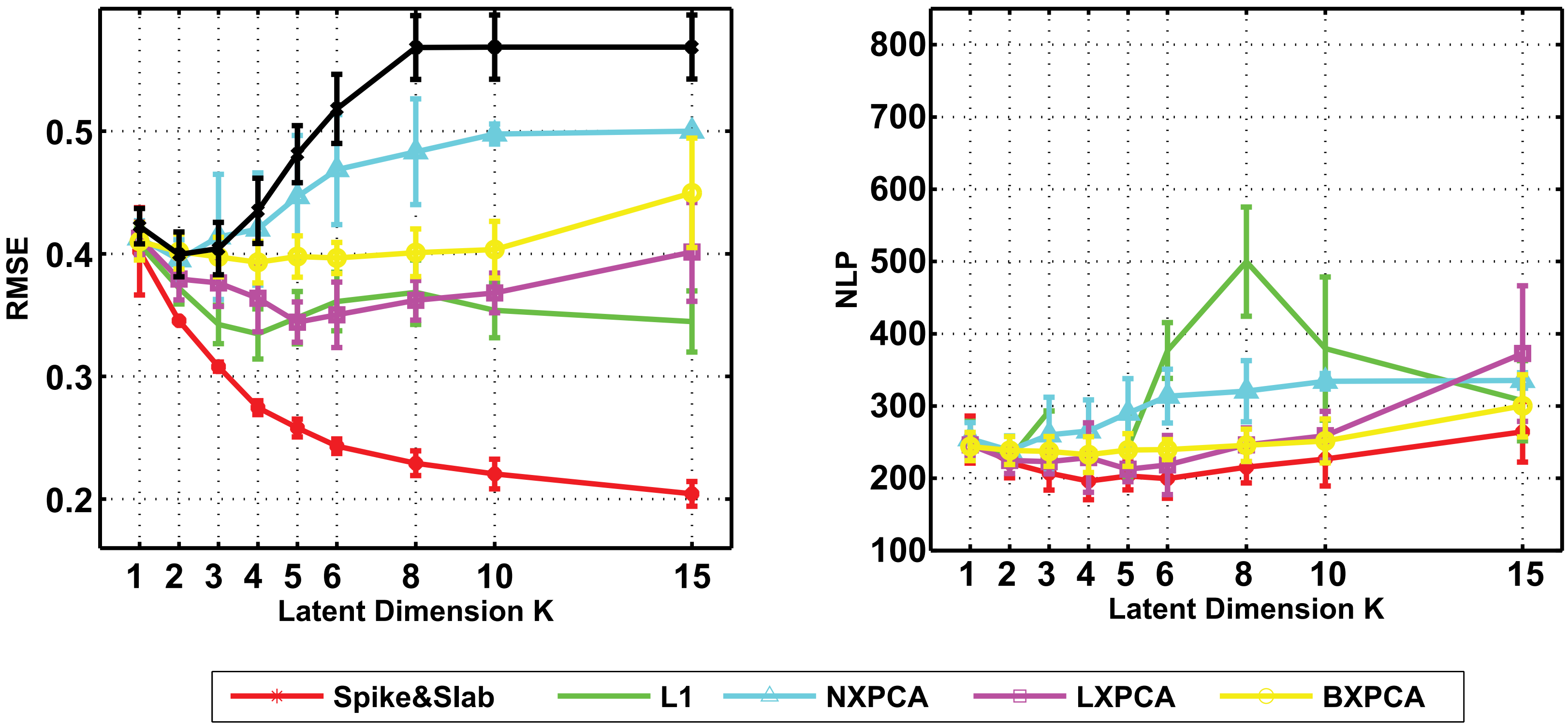
Spike&Slab:	Spike and slab with MCMC	* (This paper)
L1:	Optimization with cross validation	*
bICA:	Binary ICA	(Kaban and Bingham 2006)
NXPCA:	Non-negative exponential family PCA (MCMC)	*
LXPCA:	Laplace exponential family PCA (MCMC)	*
BXPCA:	Bayesian exponential family PCA (MCMC)	(Mohamed et al 2008)
EPCA:	Exponential family PCA (Opt)	(Collins, Dasgupta, Schapire, 2002)

Bayesian Sparse Factor Models: Test Prediction Results

Table 1. Summary of real data used.

#	Data	N	D	Type
1	Natural scenes	10,000	144	Real
2	Animal attributes	33	102	Binary
3	Newsgroups	100	200	Counts
4	Hapmap	100	200	Binary

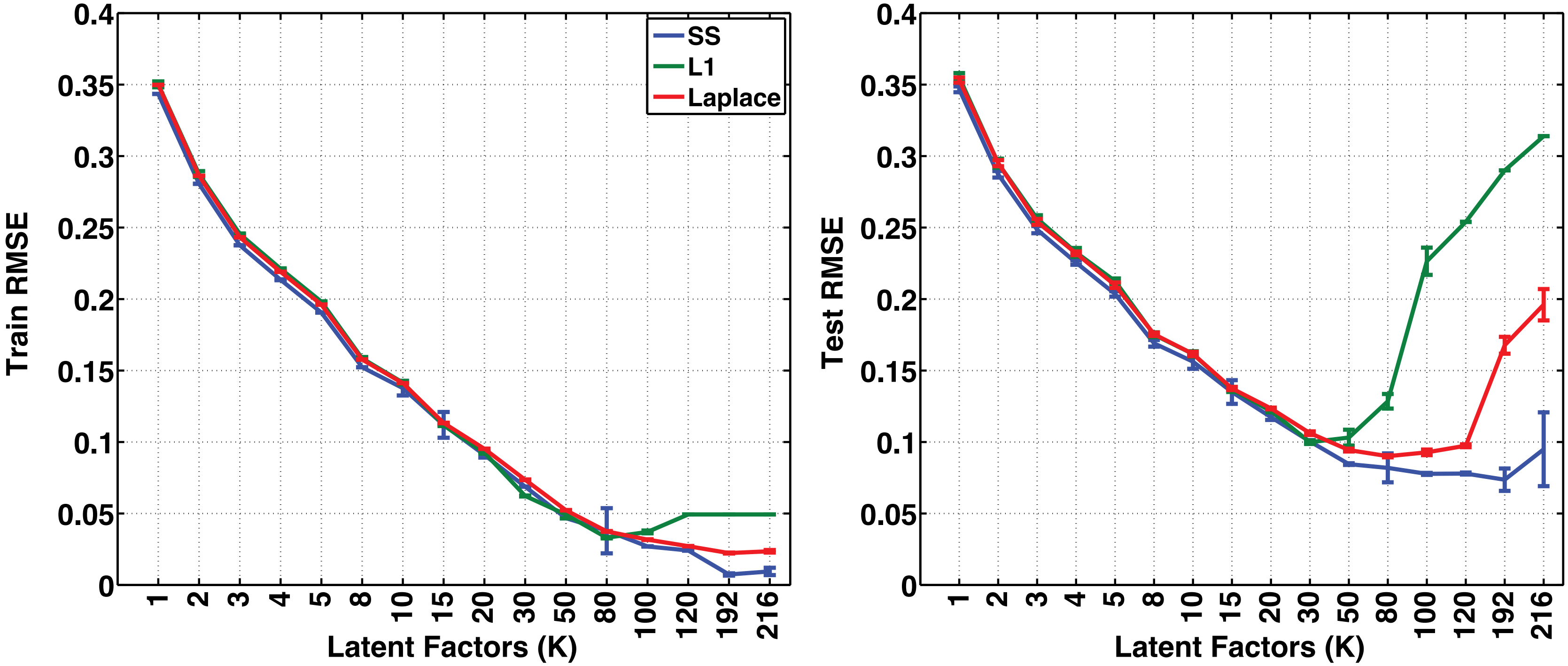
Bayesian Sparse Factor Models: Test Prediction Results



Data: Binary human judgements of different animals

Spike&Slab:	Spike and slab with MCMC	* (This paper)
L1:	Optimization with cross validation	*
NXPCA:	Non-negative exponential family PCA (MCMC)	*
LXPCA:	Laplace exponential family PCA (MCMC)	*
BXPCA:	Bayesian exponential family PCA (MCMC)	(Mohamed et al 2008)
EPCA:	Exponential family PCA (Opt)	(Collins, Dasgupta, Schapire, 2002)

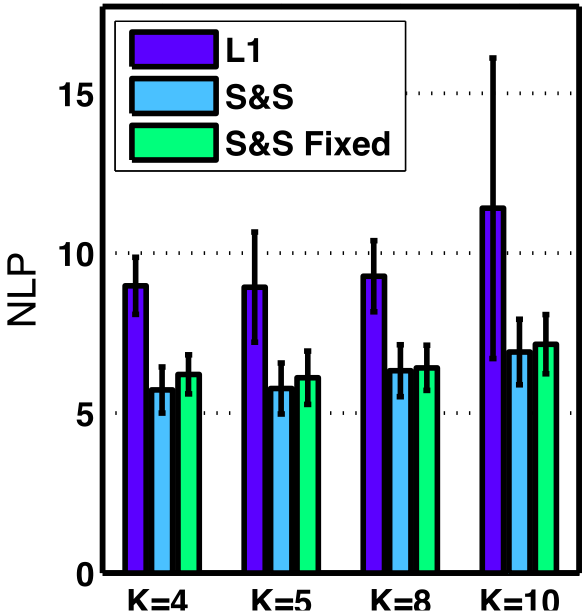
Bayesian Sparse Factor Models: Test Prediction Results

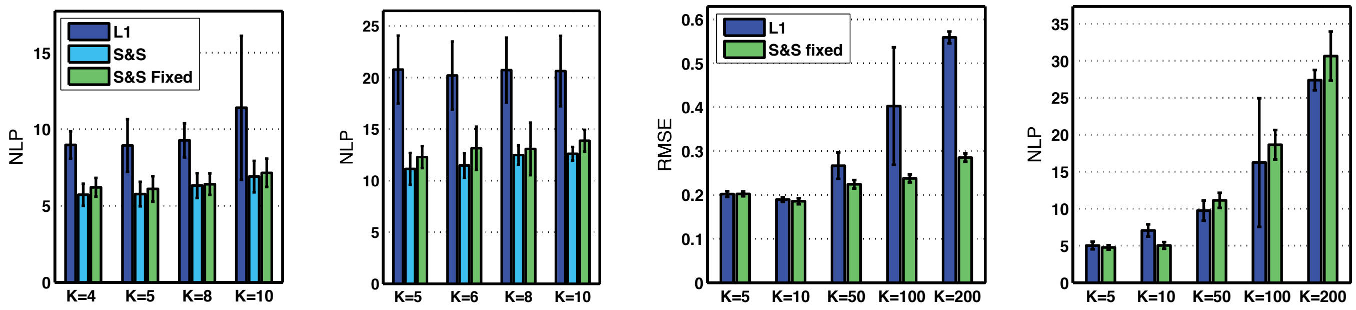


Data: natural scenes dataset (real-valued)

SS:	Spike and slab with MCMC	* (This paper)
L1:	Optimization with cross validation	*
Laplace:	Laplace exponential family PCA (MCMC)	*

Bayesian Sparse Factor Models: Timing Results





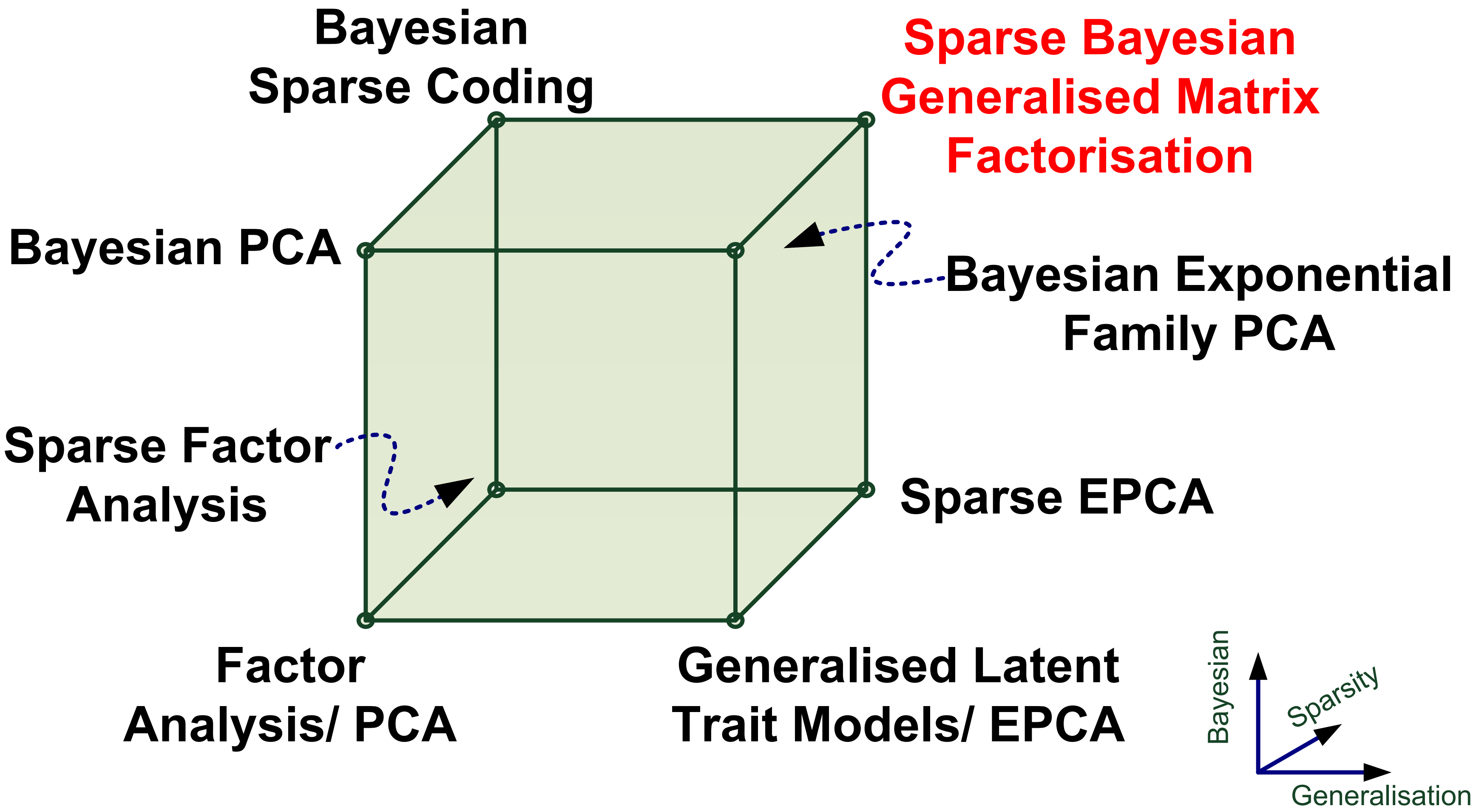
K	L_1	Spike-Slab
5	475 ±36	1446 ±24
6	483±57	1418± 29
8	592±207	1400±18
10	934±440	1367±32

(a) Animal Attr. (b) Newsgroups (c) Hapmap Data (d) Newsgroups sparsity

Figure 4. (a) - (c) Comparison of predictive probabilities (NLP). 'S&S fixed' is the time-matched spike-and-slab performance (elaborated upon in sect. 7). (d) Num. of non-zeros in newsgroups reconstruction - the true number is 1436.

Note: optimization times include cross-validation for setting regularizers.

Bayesian Sparse Factor Models: The Big Picture



Bayesian Sparse Coding

Sparse Bayesian Generalised Matrix Factorisation

Bayesian PCA

Bayesian Exponential Family PCA

Sparse Factor Analysis

Sparse EPCA

Factor Analysis/ PCA

Generalised Latent Trait Models/ EPCA

Discussion

Modelling contribution

- A new general latent factor model for strongly sparse unsupervised learning based on spike-and-slab priors and exponential family likelihoods

Algorithmic contribution

- An MCMC inference method for this model

Discussion

Modelling contribution

- A new general latent factor model for strongly sparse unsupervised learning based on spike-and-slab priors and exponential family likelihoods

Algorithmic contribution

- An MCMC inference method for this model

Experimental contribution

Some potentially controversial conclusions of this work:

- **Strong sparsity** is useful in unsupervised learning; it may better approximate the goal of L0 optimisation

Discussion

Modelling contribution

- A new general latent factor model for strongly sparse unsupervised learning based on spike-and-slab priors and exponential family likelihoods

Algorithmic contribution

- An MCMC inference method for this model

Experimental contribution

Some potentially controversial conclusions of this work:

- **Strong sparsity** is useful in unsupervised learning; it may better approximate the goal of L0 optimisation

- The **Bayesian** sparse model has much better test performance than optimization/cross-validation L1 approach

Discussion

Modelling contribution

- A new general latent factor model for strongly sparse unsupervised learning based on spike-and-slab priors and exponential family likelihoods

Algorithmic contribution

- An MCMC inference method for this model

Experimental contribution

Some potentially controversial conclusions of this work:

- **Strong sparsity** is useful in unsupervised learning; it may better approximate the goal of L0 optimisation

- The **Bayesian** sparse model has much better test performance than optimization/cross-validation L1 approach

- **MCMC** can be faster than optimisation (i.e. can get better predictive performance given the same compute-time budget)

Thanks.

Acknowledgements

Shakir Mohamed
Google DeepMind

Katherine A. Heller
Duke U

David A. Knowles
Stanford U

References

- Ghahramani, Z. (2013) Bayesian nonparametrics and the probabilistic approach to modelling. *Phil. Trans. R. Soc. A* 371: 20110553.
- Ghahramani, Z. (2015) Probabilistic machine learning and artificial intelligence. *Nature* **521**:452–459. `http://www.nature.com/nature/journal/v521/n7553/full/nature14541.html`
- Griffiths, T.L., and Ghahramani, Z. (2011) The Indian buffet process: An introduction and review. *Journal of Machine Learning Research* **12**(Apr):1185–1224.
- Knowles, D.A. and Ghahramani, Z. (2007) Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. In *7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007)*. Lecture Notes in Computer Science Series (LNCS) **4666**:381–388.
- Knowles, D.A. and Ghahramani, Z. (2011) Nonparametric Bayesian Sparse Factor Models with application to Gene Expression modelling. *Annals of Applied Statistics* **5**(2B):1534-1552.
- Mohamed, S., Heller, K.A., and Ghahramani, Z. (2009) Bayesian Exponential Family PCA. In *Advances in Neural Information Processing Systems* **21**:1089–1096. Cambridge, MA: MIT Press.
- Mohamed, S., Heller, K.A., and Ghahramani, Z. (2012) Bayesian and L1 Approaches for Sparse Unsupervised Learning. ICML 2012.
- Mohamed, S., Heller, K. A. and Ghahramani, Z. (2014) Bayesian Approaches for Sparse Latent Variable Models: Reconsidering L_1 Sparsity. In Rish, I., Cecchi, G., Lozano, A. and Niculescu-Mizil (Eds.) *Practical Applications of Sparse Modeling*. MIT Press.

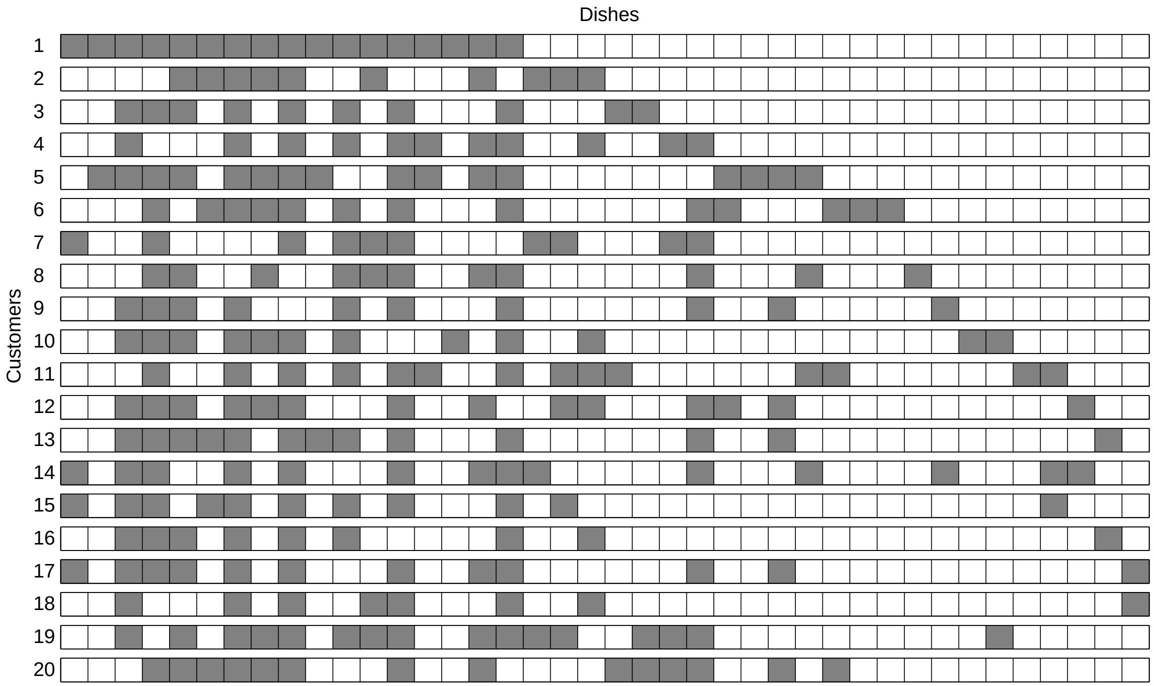
Other Related Work

Lots!

- **Spike and Slab Priors:** (Mitchell and Beauchamp, 1988; Ishwaran and Rao, 2005)
- **Feature selection, compressed sensing and regression using L_1 norm:** e.g. Tibshirani (1996); dAspremont et al. (2005); Candes (2006); Lee et al. (2006).
- **Bayesian sparse regression:** Seeger. et al. (2007); Carvalho et al. (2010); OHara and Sillanpaä (2009).
- **Sparse PCA:** (Zou et al., 2004; dAspremont et al., 2005; Rattray et al., 2009).
- **Matrix factorisation:** lots of papers!
- **Sparse deep belief networks:** Courville et al. (2010)

Appendix

Indian buffet process



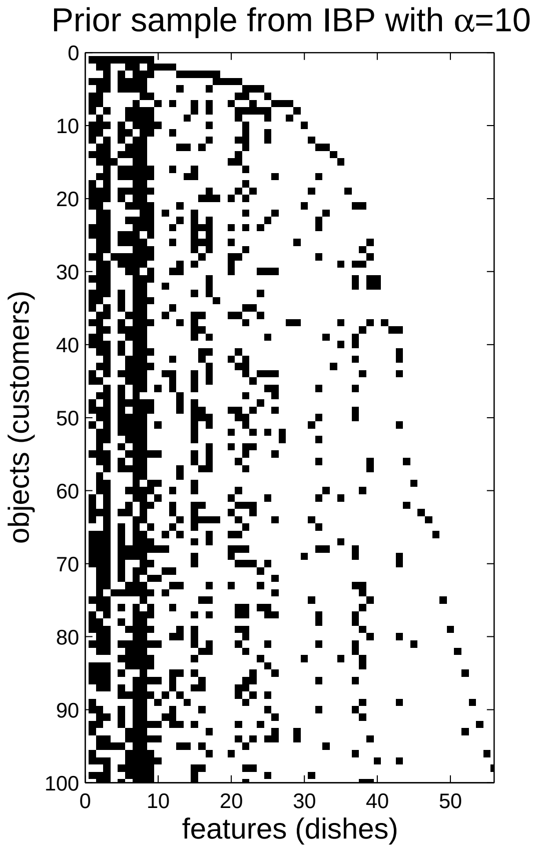
"Many Indian restaurants in London offer lunchtime buffets with an apparently infinite number of dishes"

- First customer starts at the left of the buffet, and takes a serving from each dish, stopping after a Poisson(α) number of dishes as his plate becomes overburdened.
- The n^{th} customer moves along the buffet, sampling dishes in proportion to their popularity, serving himself dish k with probability m_k/n, and trying a Poisson(α/n) number of new dishes.
- The customer-dish matrix, $\mathbf{Z}$, is a draw from the IBP.

(w/ Tom Griffiths 2006; 2011)

Properties of the Indian buffet process

$$P([\mathbf{Z}]|\alpha) = \exp\left\{ - \alpha H_N \right\} \frac{\alpha^{K_+}}{\prod_{h>0} K_h!} \prod_{k \leq K_+} \frac{(N - m_k)!(m_k - 1)!}{N!}$$



Shown in (Griffiths and Ghahramani 2006, 2011):

- It is infinitely exchangeable.
- The number of ones in each row is Poisson(α)
- The expected total number of ones is αN.
- The number of nonzero columns grows as $O(\alpha \log N)$.

Additional properties:

- Has a stick-breaking representation (Teh, et al 2007)
- Has as its de Finetti mixing distribution the Beta process (Thibaux and Jordan 2007)
- More flexible two and three parameter versions exist (w/ Griffiths & Sollich 2007; Teh and Görür 2010)

Posterior Inference in IBPs

$$P(\mathbf{Z}, \alpha|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{Z})P(\mathbf{Z}|\alpha)P(\alpha)$$

Gibbs sampling: $P(z_{nk} = 1|\mathbf{Z}_{-(nk)}, \mathbf{X}, \alpha) \propto P(z_{nk} = 1|\mathbf{Z}_{-(nk)}, \alpha)P(\mathbf{X}|\mathbf{Z})$

- If $m_{-n,k} > 0$, $P(z_{nk} = 1|\mathbf{z}_{-n,k}) = \frac{m_{-n,k}}{N}$
- For infinitely many k such that $m_{-n,k} = 0$: Metropolis steps with truncation* to sample from the number of new features for each object.
- If α has a Gamma prior then the posterior is also Gamma → Gibbs sample.

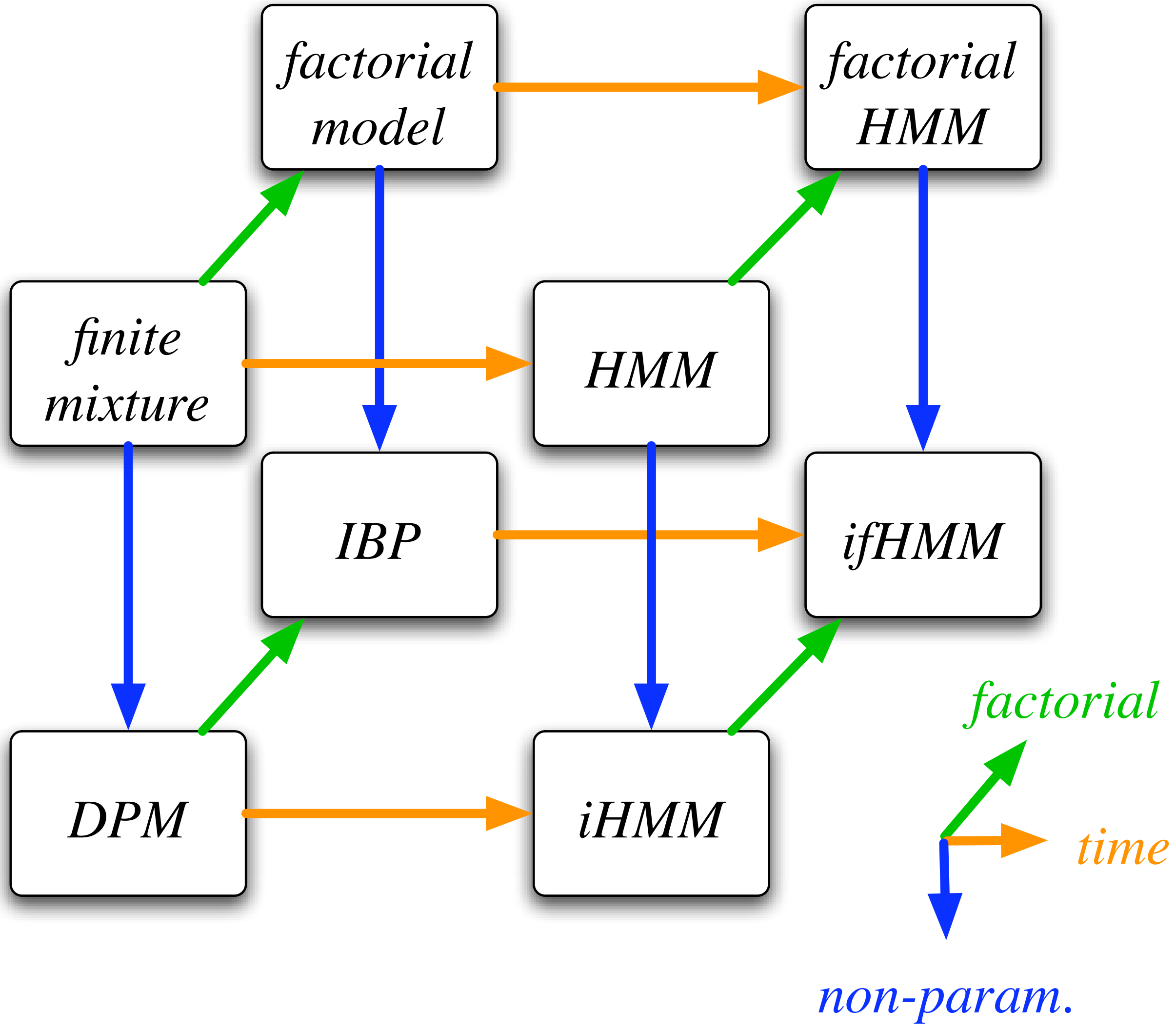
Conjugate sampler: assumes that $P(\mathbf{X}|\mathbf{Z})$ can be computed.

Non-conjugate sampler: $P(\mathbf{X}|\mathbf{Z}) = \int P(\mathbf{X}|\mathbf{Z}, \theta)P(\theta)d\theta$ cannot be computed, requires sampling latent θ as well (e.g. approximate samplers based on (Neal 2000) non-conjugate DPM samplers).

Slice sampler: works for non-conjugate case, is not approximate, and has an *adaptive truncation level* using an IBP stick-breaking construction (Teh, et al 2007) see also (Adams et al 2010).

Deterministic Inference: variational inference (Doshi et al 2009a) parallel inference (Doshi et al 2009b), beam-search MAP (Rai and Daume 2011), power-EP (Ding et al 2010)

The Big Picture:
Relations between some models



factorial model
factorial HMM
finite mixture
HMM
IBP
ifHMM
DPM
iHMM
factorial
time
non-param.