Infinite Sparse Factor Analysis and Infinite Independent Component Analysis

David Knowles and Zoubin Ghahramani

Paper Presented by

Mark Buller

Outline

- Motivation
- Background
  - Factor Analysis
  - ICA
- Extension to infinite models
  - A finite model
  - Indian buffet (One parameter)
  - Indian buffet + (Two parameter)
- Inference
  - Gibbs Sampling
  - Metropolis-Hastings
    - Sampling new features
    - Finding 2nd Indian Buffet Parameter
- Experimental Results

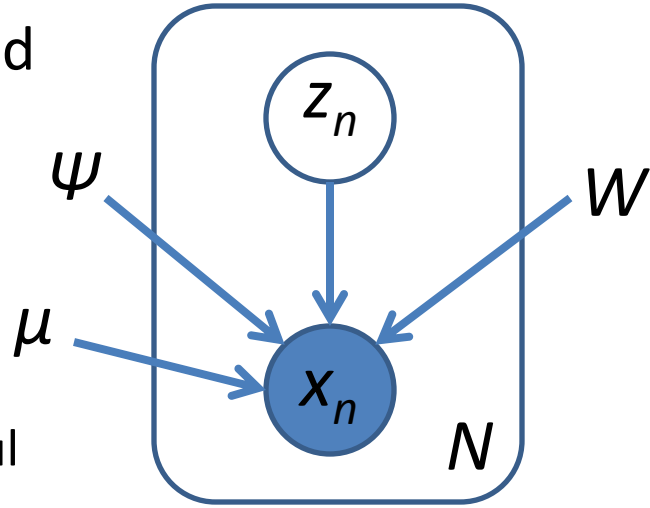
Motivation

- Cocktail Party Problem or Blind Source Separation
  - Given an observed signal (**y**) of a mixture of sources
  - y = Ax (y = observed, A = mixing matrix, x = source)
  - Want to recover the mixing coefficients and sources

Example from Aalto University, Department of Information and Computer Science:
http://research.ics.tkk.fi/ica/cocktail/cocktail_en.cgi

Background

- Factor Analysis
  - Linear Gaussian latent variable model
  - Similar to probabilistic PCA (Tipping and Bishop 1997, 1999; and Roweis 1998)
  - $P(x|z)=N(x|Wz + \mu, \Psi)$
  - Solved using ML computed from EM
  - Limitations:
    - Latent coordinate system is not meaningful
    - From y = Ax (y = observed, G = mixing matrix, x = source) We can recover the sources but not the mixing matrix.



Graphical Model from Bishop 2006.

Background

- Independent Component Analysis (ICA)
  - Linear non-Gaussian latent variable model
  - A non-Gaussian latent variable distribution allows the mixing matrix (G) to be estimated. (y = Gx)
  - Number of Sources (K) <= Observation Dimensions (D)

Example from Aalto University, Department of Information and Computer Science: http://research.ics.tkk.fi/ica/cocktail/cocktail_en.cgi

Contribution

- Model Extension:
  - Allows sources K> observed D
  - Can switch sources on and off
    - Provides a sparse model
  - Allows solutions using both Gaussian and non-Gaussian assumptions about the latent variable distribution

Infinite Model

Binary Mask

Gaussian Noise

$$\mathbf{y}_t = \mathbf{G}(\mathbf{z}_t \bullet \mathbf{x}_t) + \boldsymbol{\varepsilon}_t$$

Observed (dimension D)

Mixing Matrix

Element Wise Multiplication

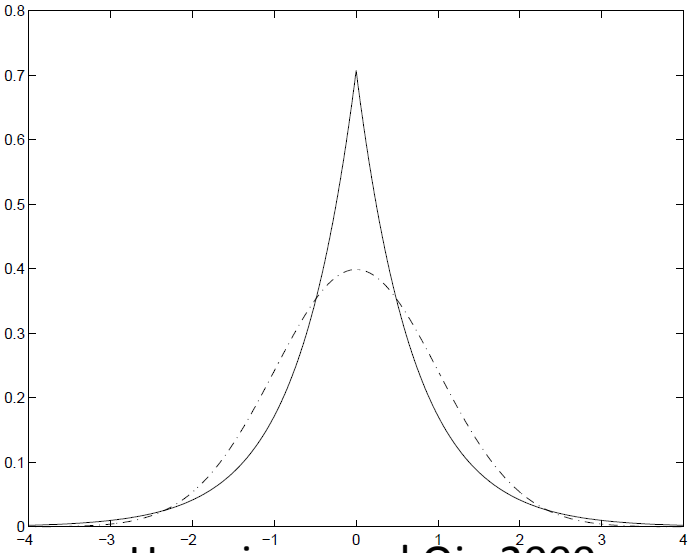
Latent (dimension K)

$$\boldsymbol{\epsilon}_t \sim \mathcal{N}\left(0, \sigma_\epsilon^2 \mathbf{I}\right) \qquad \sigma_\epsilon^2 \sim \mathcal{IG}\left(a, b\right)$$

$$\mathbf{g}_k \sim \mathcal{N}\left(0, \sigma_G^2\right) \qquad \sigma_G^2 \sim \mathcal{IG}\left(c, d\right)$$

$$\mathbf{Z} \sim \mathcal{IBP}(\alpha, \beta) \qquad \alpha \sim \mathcal{G}\left(e, f\right)$$

	$x_{kt} \sim \mathcal{N}(0,1)$	$x_{kt} \sim \mathcal{L}(1)$
$\beta = 1$	$isFA_1$	$iICA_1$
$\beta \sim \mathcal{G}(1,2)$	$isFA_2$	$iICA_2$



Hyvarinen and Oja 2000

Infinite Binary Matrix Distribution (Z)

- Approach:
  - Define a finite model with K sources
  - Take limit as $K \to \infty$
  - Demonstrate that the infinite case is a simple stochastic process
    - Follows: Ghahramani, Griffiths and Sollich: Bayesian nonparametric latent feature models (2007)

Finite Model

- Assumptions:
  - Sources independent
  - Probability of source being active: π_k

$$P(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{k=1}^{K}\prod_{t=1}^{N} P(z_{kt}|\pi_k) = \prod_{k=1}^{K} \pi_k^{m_k}(1-\pi_k)^{N-m_k}$$

  - $m_k = \sum_{t=1}^{N} z_{kt}$ number of data points for which source *k* is active
  - Define prior for π by assuming drawn from Beta distribution
    - Conjugate to Binomial Distribution
    - Chinese Restaurant Process made use of Dirichlet Process conjugacy to Multinomial Distribution
    - B(r,s)

Beta Distribution Prior

$$p(\pi_k) = \frac{\pi_k^{r-1}(1-\pi_k)^{s-1}}{B(r,s)}$$

$$B(r,s) = \int_0^1 \pi_k^{r-1}(1-\pi_k)^{s-1}\, d\pi_k$$

$$= \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}.$$

By taking $r = \frac{\alpha}{k}$ and s = 1:

$$B(\tfrac{\alpha}{K}, 1) = \frac{\Gamma(\frac{\alpha}{K})}{\Gamma(1+\frac{\alpha}{K})} = \frac{K}{\alpha}$$

$$\pi_k \mid \alpha \sim \text{Beta}(\tfrac{\alpha}{K}, 1)$$

$$z_{kt} \mid \pi_k \sim \text{Bernoulli}(\pi_k)$$

Marginalize

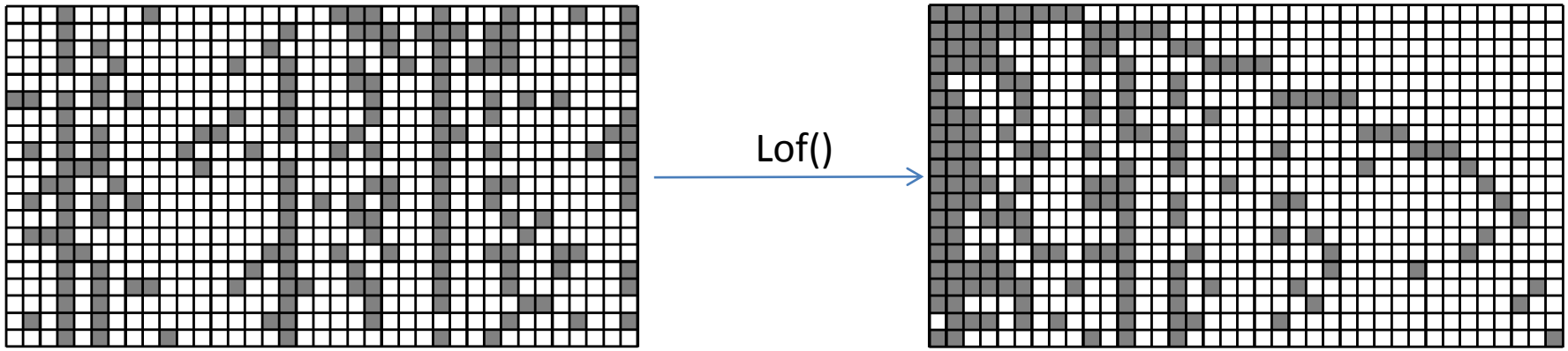
$$
\begin{aligned}
P(\mathbf{Z}) &= \prod_{k=1}^{K} \int \left(\prod_{i=1}^{N} P(z_{ik}|\pi_k) \right) p(\pi_k)\, d\pi_k \\
&= \prod_{k=1}^{K} \frac{B(m_k + \frac{\alpha}{K}, N - m_k + 1)}{B(\frac{\alpha}{K}, 1)} \\
&= \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}
\end{aligned}
$$

Conjugacy makes this marginalization possible

Distribution is exchangeable depending on m_k

Take the Limit

- OK they use a cool trick here…
  - Define an equivalence class of left ordered binary matrices
    - Analog of partitions for assignment vectors
  - Use a function to map binary matrix into left order (lof(●))
  - A distribution over collections of histories. I.e.
    - History for source k at observation t is defined as (Z_{1k}, … $Z_{(t-1)k}$)



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

$H_N = \sum_{j=1}^{N} \frac{1}{j}$ N-th Harmonic Number, K_h num. rows with binary num. h

K_+ Number of active features

Indian Buffet Process

- P(Z) relates to a simple stochastic process
- Buffet with infinite choices

Customer #

…

…

1 Samples Poisson (**α**)dishes

i Samples previously sampled dishes with probability $\frac{m_k}{i}$

And tries $\left(\frac{\alpha}{i}\right)$ new dishes

$$P(z_{kt} = 1 | \mathbf{z}_{-kt}) = \frac{m_{k,-t}}{N}$$

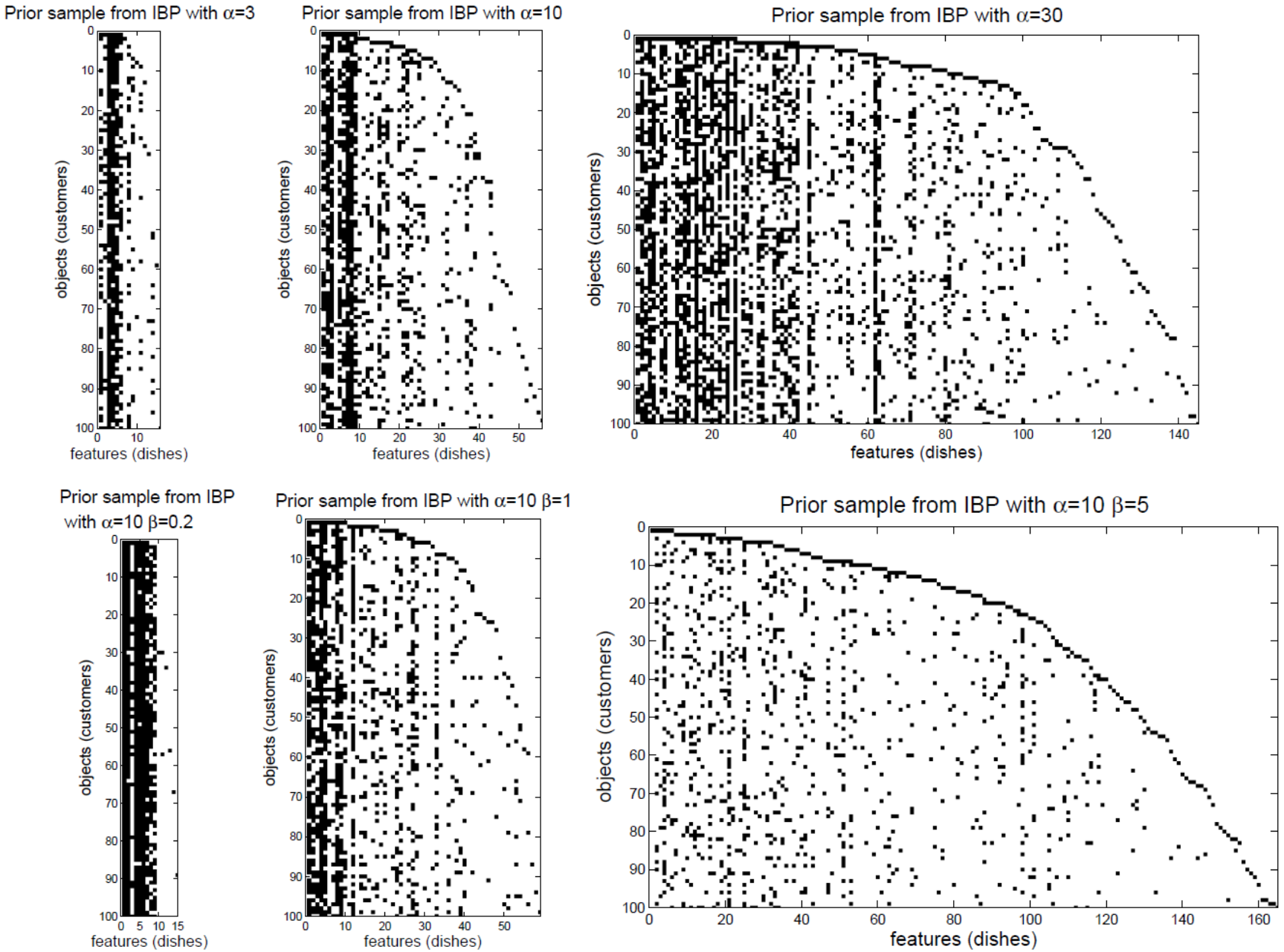
$$m_{k,-t} = \sum_{s \neq t} z_{ks}$$

Two Parameter Generalization

- "Distribution on the number of features per object and the total number of features are coupled through α"
- Add an additional parameter β "a measure of feature repulsion"
- i_{th} customer now samples dish *k* with:
  - *Probability* $\frac{m_k}{\beta+i-1}$
  - *Samples* Poisson $\left(\frac{\alpha\beta}{\beta+i-1}\right)$ new dishes
  - Marginal probability of **Z** becomes:

$$P(\mathbf{Z}|\alpha,\beta) = \frac{(\alpha\beta)^{K_+}}{\prod_{h>0} K_h!} \exp\{-\alpha H_N(\beta)\} \prod_{k=1}^{K_+} B(m_k, N-m_k+\beta)$$

Ghahramani, Griffiths, Sollich 2006


Prior sample from IBP with α=3
objects (customers)
features (dishes)
Prior sample from IBP with α=10
objects (customers)
features (dishes)
Prior sample from IBP with α=30
objects (customers)
features (dishes)
Prior sample from IBP with α=10 β=0.2
objects (customers)
features (dishes)
Prior sample from IBP with α=10 β=1
objects (customers)
features (dishes)
Prior sample from IBP with α=10 β=5
objects (customers)
features (dishes)

Inference

- Given observations **Y**

- Wish to infer:
  - hidden sources **X**

  - which sources are active **Z**
  - mixing matrix **G**
  - Hyper parameters
- Gibbs sampling
  - Metropolis-Hastings steps for:
    - β
    - New Features
  - Samples are drawn from marginal distribution of the model parameters by
    - Successively sample conditional distribution of each parameter in turn, given all other parameters

Hidden Sources

- Sample each element of X for which $z_{kt}=1$
- For:
  - isFA conditional distribution is a Gaussian

$$P(x_{kt}|\mathbf{G}, \mathbf{x}_{-kt}, \mathbf{y}_t, \mathbf{z}_t) = \mathcal{N}\left(x_{kt}; \frac{\mathbf{g}_k^T \boldsymbol{\epsilon}_{-kt}}{\sigma_\epsilon^2 + \mathbf{g}_k^T \mathbf{g}_k}, \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \mathbf{g}_k^T \mathbf{g}_k}\right)$$

  - ilCA piecewise Gaussian

$$P(x_{kt}|\mathbf{G}, \mathbf{x}_{-kt}, \mathbf{y}_t, \mathbf{z}_t) = \begin{cases} \mathcal{N}\left(x_{kt}; \mu_-, \sigma^2\right) & x_{kt} > 0 \\ \mathcal{N}\left(x_{kt}; \mu_+, \sigma^2\right) & x_{kt} < 0 \end{cases}$$

$$\text{where } \mu_\pm = \frac{\mathbf{g}_k^T \boldsymbol{\epsilon}_{-kt} \pm \sigma_\epsilon^2}{\mathbf{g}_k^T \mathbf{g}_k} \text{ and } \sigma^2 = \frac{\sigma_\epsilon^2}{\mathbf{g}_k^T \mathbf{g}_k}$$

Active Sources

- Sample Z define a ration of conditionals, r so that:

$$P(z_{kt} = 1 | \mathbf{G}, \mathbf{X}_{-kt}, \mathbf{Y}, \mathbf{Z}_{-kt}) = \frac{r}{r+1}$$

$$r = \underbrace{\frac{P(\mathbf{y}_t | \mathbf{G}, \mathbf{x}_{-kt}, \mathbf{z}_{-kt}, z_{kt} = 1, \sigma_\epsilon^2)}{P(\mathbf{y}_t | \mathbf{G}, \mathbf{x}_{-kt}, \mathbf{z}_{-kt}, z_{kt} = 0, \sigma_\epsilon^2)}}_{r_l} \underbrace{\frac{P(z_{kt} = 1 | \mathbf{z}_{-kt})}{P(z_{kt} = 0 | \mathbf{z}_{-kt})}}_{r_p}$$

$$r_p = \frac{m_{k,-t}}{\beta + N - 1 - m_{k,-t}}$$

isFA $$r_l = \sigma \exp\left\{\frac{\mu^2}{2\sigma^2}\right\}$$

iICA $$r_l = \sigma\sqrt{\frac{\pi}{2}}\left[F(0;\mu_+,\sigma)\exp\left\{\frac{\mu_+^2}{2\sigma^2}\right\} + (1 - F(0;\mu_-,\sigma))\exp\left\{\frac{\mu_-^2}{2\sigma^2}\right\}\right]$$

Creating New Features

- For a given time point the number of active features k_n are sampled with a Metropolis Hastings move $\xi \rightarrow \xi^*$
- Move Accepted: $\min\left(1, r_{\xi \rightarrow \xi^*}\right)$

- isFA $r_{\xi \rightarrow \xi^*} = |\boldsymbol{\Lambda}|^{-\frac{1}{2}} \exp\left(\frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\Lambda}\boldsymbol{\mu}\right)$

$$\boldsymbol{\Lambda} = \mathbf{I} + \frac{\mathbf{G}^{*T}\mathbf{G}^*}{\sigma_\epsilon^2} \qquad \boldsymbol{\Lambda}\boldsymbol{\mu} = \frac{1}{\sigma_\epsilon^2}\mathbf{G}^{*T}\boldsymbol{\epsilon}_t$$

- iICA

$$r_{\xi \rightarrow \xi^*} = \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\mathbf{x}_t'^T\mathbf{G}^{*T}(\mathbf{G}^*\mathbf{x}_t' - 2\boldsymbol{\epsilon}_t)\right\}$$

Mixture Weights

- Sample columns g_k of G

$$P(\mathbf{g}_k|\mathbf{G}_{-k},\mathbf{X},\mathbf{Y},\mathbf{Z},\sigma_\epsilon^2,\sigma_G^2) \propto P(\mathbf{Y}|\mathbf{G},\mathbf{X},\mathbf{Z},\sigma_\epsilon^2)P(\mathbf{g}_k|\sigma_G^2)$$

- Likelihood function exponent:

$$-\frac{1}{2\sigma_\epsilon^2}\operatorname{tr}(\mathbf{E}^T\mathbf{E}) = -\frac{1}{2\sigma_\epsilon^2}((\mathbf{x}_k'^T\mathbf{x}_k')(\mathbf{g}_k^T\mathbf{g}_k) - 2\mathbf{g}_k^T\mathbf{E}|_{\mathbf{g}_k=0}) + \text{const}$$

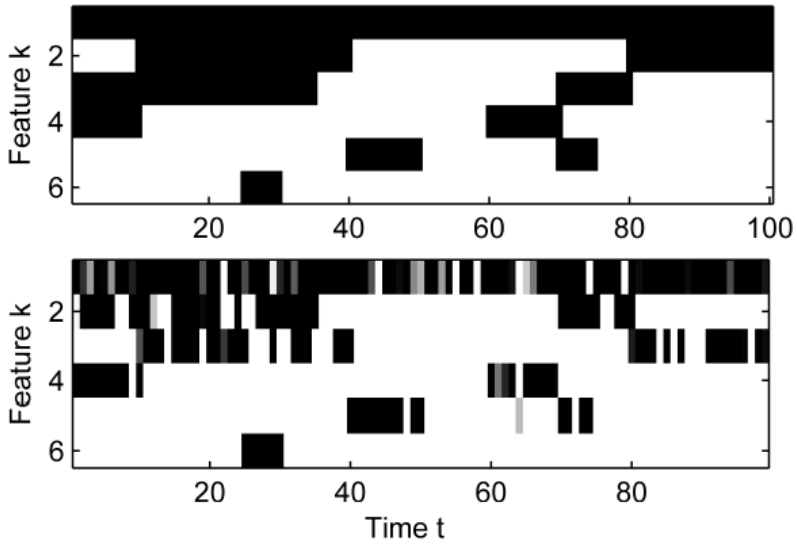
- Conditional of g_k is N(μ,Λ) where:

$$\boldsymbol{\mu} = \frac{\sigma_G^2}{\mathbf{x}_k'^T\mathbf{x}_k'\sigma_G^2+\sigma_\epsilon^2}\mathbf{E}|_{\mathbf{g}_k=0}\mathbf{x}_k'$$

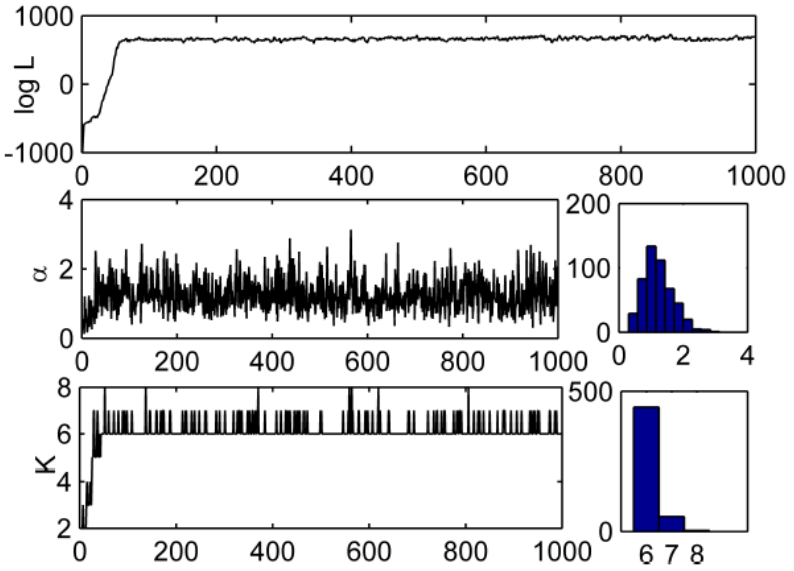
$$\boldsymbol{\Lambda} = \left(\frac{\mathbf{x}_k'^T\mathbf{x}_k'}{\sigma_\epsilon^2} + \frac{1}{\sigma_G^2}\right)\mathbf{I}_{D\times D}$$

Results

- Synthetic Data
  - 30 sets randomly generated data (D=7, K=6, N=200), using both Gaussian and Laplacian distributions

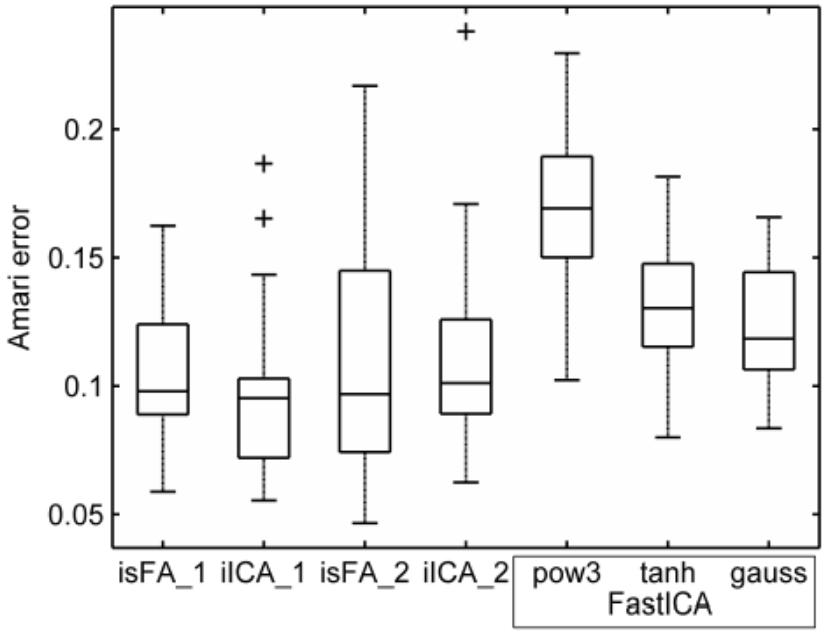


(a) *Top:* True $\mathbf{Z}$. *Bottom:* Inferred $\mathbf{Z}$.

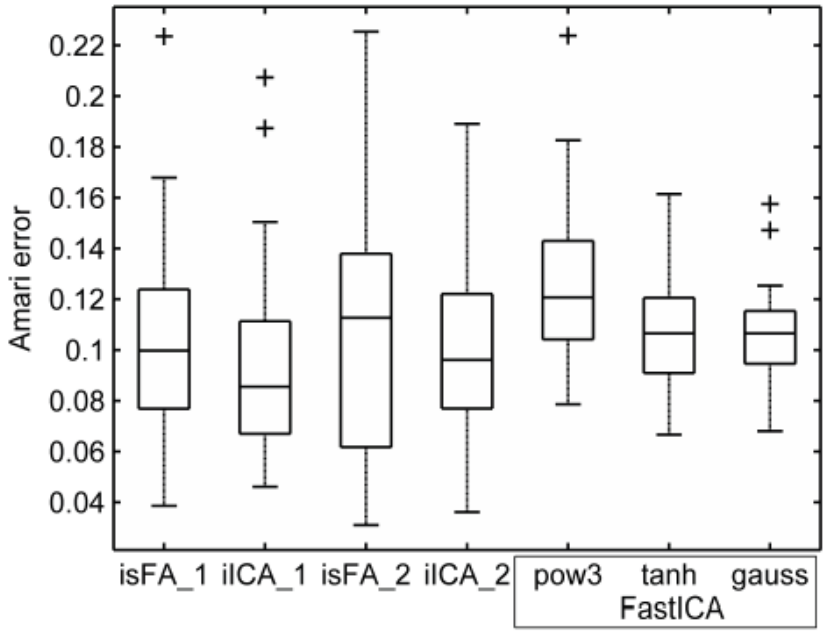


(b) Log likelihood, α and K^+ for duration of 1000 iteration run.

Model Comparison



(a) Gaussian sources.



(b) Laplacian sources.

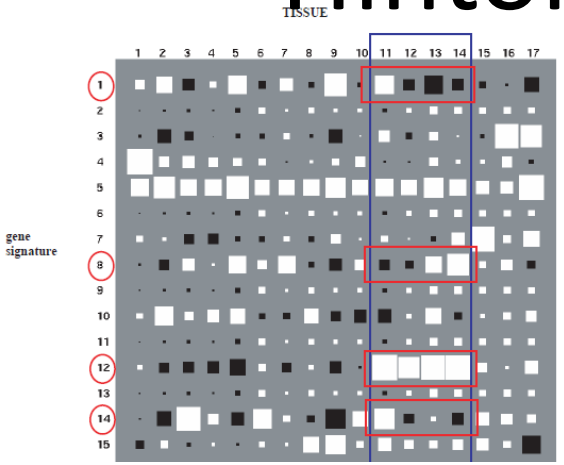
2 Parameter IBP results show more variance

FastICA is affected more by source distribution

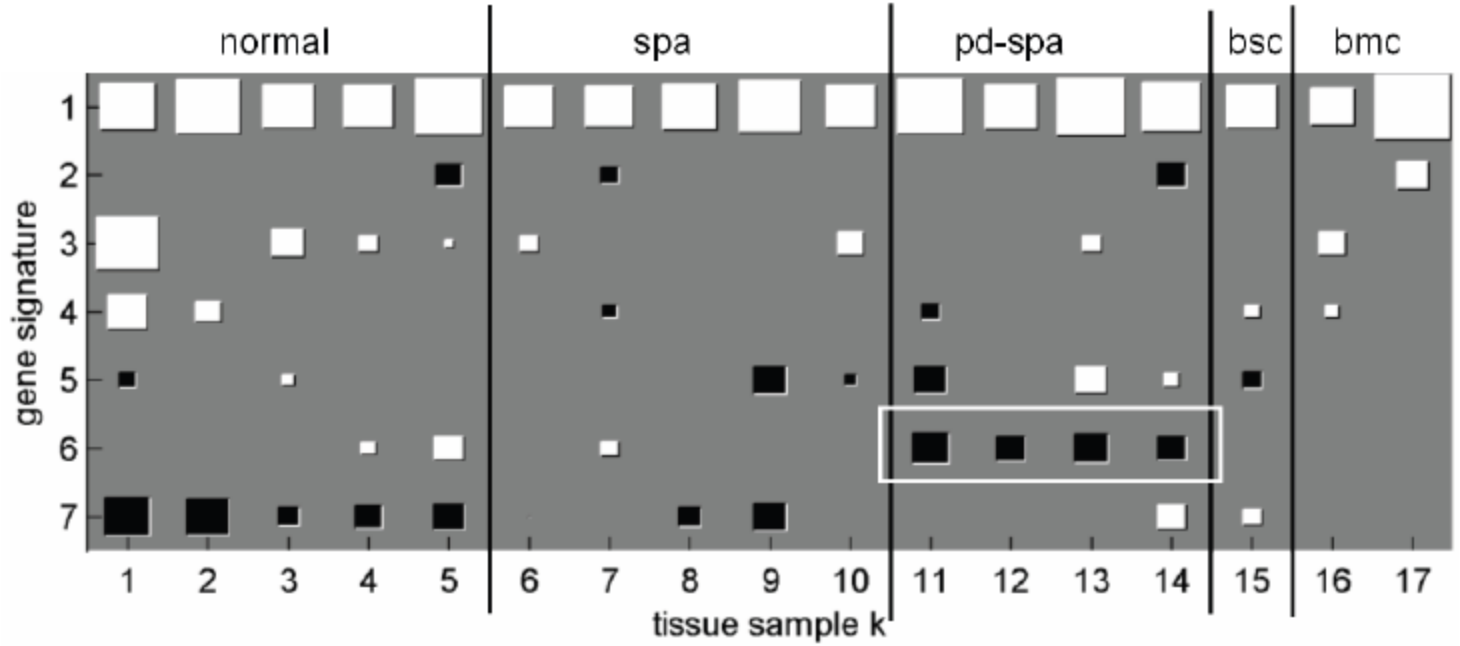
Gene Expression Study

- Ovarian Cancer Study
- N=172 genes (data points)
- D=17 tissue samples
- Tissues grouped into 5 tissue types:
  - 1 healthy
  - 4 diseased
- Some gene signatures are expressed across all samples, others are silent
  - The sparse model ICA model is applicable here

Hinton Diagram of G



ICA Technique, Martoglio, Miskin, Smith and MacKay 2002



Questions

- Need to switch sources on and off? Can't a source be 0 in the mixing matrix?
- Are the results that much better than FastICA?
- No speed comparisons with FastICA?
- Does having the 2 parameter IBP help?