# Landmark-Dependent Hierarchical Beta Process for Robust Sparse Factor Analysis

**Mingyuan Zhou**[1] MZ1@EE.DUKE.EDU
**Hongxia Yang**[2] YANGHO@IBM.US.COM
**Guillermo Sapiro**[3] GUILLE@UMN.EDU
**David Dunson**[4] DUNSON@STAT.DUKE.EDU
**Lawrence Carin**[1] LCARIN@EE.DUKE.EDU

[1]Department of ECE, [4]Department of Statistical Science, Duke University, Durham, NC 27708, USA
[2]IBM Watson Research Center, Yorktown Heights, NY 10598, USA
[3]Department of ECE, University of Minnesota, Minneapolis, MN 55455, USA

## Abstract

A computationally efficient landmark-dependent hierarchical beta process is developed as a prior for data with associated covariates. The landmarks define local regions in the covariate space where feature usages are likely to be similar. The landmark locations are learned, to which the data are linked through normalized kernels. The proposed model is well suited for local latent feature discovery, and adding a robustness term, it successfully separates out non-local sparse spiky components, as demonstrated in image denoising and document analysis applications.

## 1. Introduction

In this paper we address dictionary learning for data that are endowed with an associated covariate. We explore the idea that data nearby in the covariate space ($e.g.$, nearby in terms of temporal, spatial or cosine distances) are likely to share similar sparseness properties. Hence we employ "landmarks" in the covariate space, whose positions are learned, to guide the usage probabilities of dictionary atoms. Links between the data and landmarks are established via normalized kernels. The proposed model may be viewed as a landmark-based extension of the recently proposed dependent hierarchical beta process (dHBP) (Zhou et al., 2011). We refer to the proposed model as Landmark-dHBP, which is considerably more efficient than the original dHBP for large scale learning and out-of-sample prediction.

A robustness term is employed to model sparse spiky noise or localized data anomalies, related to robust principal component analysis (RPCA) (Candès et al., 2009). However, our model differs from RPCA in that the low-rank assumption on data is replaced with the richer covariate-dependent union-of-subspace assumption, realized with a sparse factor analysis (SFA) model using the Landmark-dHBP as the prior. Therefore, our model might better handle data that violate the low-rank assumption, and the use of covariates to define a local region is important to distinguish non-local spiky or anomalous data components from local latent features.



## 2. Review of Motivating Models

Following (Thibaux & Jordan, 2007), a beta process (BP) $B \sim \text{BP}(c, B_0)$ is a positive random measure on a space $\Omega$, where $c$ is a positive function over $\Omega$, and $B_0$ is a fixed measure on $\Omega$, called the base measure; here we assume $c$ is a positive constant. A BP draw can be expressed as $B = \sum_{k=1}^{\infty} p_k \delta_{\omega_k}$, where $\omega_k \sim B_0/B_0(\Omega)$ is the $k$th atom, and each $p_k$ is drawn i.i.d. from a degenerate beta distribution with parameter $c$. We draw the atom usage vector for the $i$th sample, $i = 1, \cdots, N$, from a Bernoulli process as $X_i \sim \text{BeP}(B)$.

The BP construction above is exchangeable. Building on a hierarchical BP (HBP) construction (Thibaux & Jordan, 2007), a dependent HBP (dHBP) was recently proposed to introduce covariant dependence, demonstrating significant improvements over BP in image interpolation and denoising applications. The model can be express as

$$B_i = \sum_{j=1}^{N} a_{ij} B_j^*, B_j^* \sim \text{BP}(c_1, B), B \sim \text{BP}(c_0, B_0) \quad (1)$$

$$a_{ij} = \mathcal{K}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_j) / \sum_{j'=1}^{N} \mathcal{K}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_{j'}) \quad (2)$$

where $c_1$ is a constant, $a_{ij}$ is an element of the random-walk matrix $\mathbf{A} \in \Re^{N \times N}$, each row of which sums to one,

$\boldsymbol{\ell}_i \in \mathbb{R}^{\mathcal{L}}$ is a covariate associated with the $i$th sample, the kernel $\mathcal{K}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_j) \in [0,1]$ monotonically decreases towards 0 with increasing $\|\boldsymbol{\ell}_i - \boldsymbol{\ell}_{i'}\|$. As shown in (Zhou et al., 2011), for any measurable subset $S$, when $B_i$ and $B_{i'}$ are constituted as in the above model, these dependent random measures satisfy

$$\text{corr}\{B_i(S), B_{i'}(S)\} = \frac{< \boldsymbol{a}_i, \boldsymbol{a}_{i'} >}{\|\boldsymbol{a}_i\| \cdot \|\boldsymbol{a}_{i'}\|} \tag{3}$$

where $\boldsymbol{a}_i = [a_{i1}, \ldots, a_{iN}]^T$.

When constructing the random-walk matrix, (Zhou et al., 2011) used a neighborhood constraint on the kernel that $\mathcal{K}(\boldsymbol{\ell}_i, \boldsymbol{\ell}_j) \neq 0$ if and only if $j \in \mathcal{Q}_i$, where $\mathcal{Q}_i$ includes the indexes of the $L$ nearest covariates of $\boldsymbol{\ell}_i$ from $\{\boldsymbol{\ell}_j\}_{j=1,N}$. This construction directly links $B_j^*$ to $\{B_i\}_{\{i:\{j\in\mathcal{Q}_i\}\}}$. Since given $B_i$, $X_i$ is independently drawn as $X_i \sim \text{BeP}(B_i)$, this construction makes it feasible to infer the posterior distribution of $B_j^*$ based on $\sum_{i:\{j\in\mathcal{Q}_i\}} X_i$.

## 3. Landmarks and Kernels

The salutary characteristics of (3) are undermined by several additional attributes. First, the complexity of the model grows with the number of data samples $N$, even if many of them provide redundant information. Second, inference on a new sample $i'$ is complicated by the fact that one must retain all training data to compute the weights $\{a_{i'j}\}$ when subsequently using the model. We seek to remove these deficiencies of the dHBP model, while retaining its desirable characteristics, particulary the covariate-dependent removal of exchangeability. Toward this end, we introduce the concept of "landmarks" in the covariate space.

Consider the $J$ landmarks $\{\tilde{\boldsymbol{\ell}}_j\}_{j=1,J}$ to be learned, where each $\tilde{\boldsymbol{\ell}}_j \in \mathbb{R}^{\mathcal{L}}$. These landmarks are meant to capture the support of the set of covariates $\{\boldsymbol{\ell}_i\}_{i=1,N}$, serving as reference points to guide the covariate-dependent usage of atoms $\{\omega_k\}$. We are typically interested in $J \ll N$. For general covariate $\boldsymbol{\ell}$, the landmark-dependent probability measure for atom usage is expressed as

$$B_{\boldsymbol{\ell}} = \sum_{j=1}^{J} a_{\boldsymbol{\ell}j} B_j^*, B_j^* \sim \text{BP}(c_1, B), B \sim \text{BP}(c_0, B_0) \tag{4}$$

where the $\boldsymbol{\ell}$-dependent probabilities $\{a_{\boldsymbol{\ell}j}\}$ are defined as

$$a_{\boldsymbol{\ell}j} = \mathcal{K}(\boldsymbol{\ell}, \tilde{\boldsymbol{\ell}}_j) / \sum_{j'=1}^{J} \mathcal{K}(\boldsymbol{\ell}, \tilde{\boldsymbol{\ell}}_{j'}) \tag{5}$$

where $\mathcal{K}(\boldsymbol{\ell}_i, \tilde{\boldsymbol{\ell}}_j)$ is a kernel as discussed above. Note that $B_{\boldsymbol{\ell}}$ varies smoothly with the covariate $\boldsymbol{\ell}$.

To complete the model specification, the landmarks are assumed drawn from a probability measure $H$ in $\mathbb{R}^{\mathcal{L}}$, as $\tilde{\boldsymbol{\ell}}_j \sim H$. A simple way to do this is to densely select $\widetilde{N}$ positions $\{\boldsymbol{\ell}_i\}_{i=1,\widetilde{N}}$ in the covariate space as potential landmarks, and assume $H = \sum_{i=1}^{\widetilde{N}} \theta_i \delta_{\boldsymbol{\ell}_i}$ with $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_{\widetilde{N}}]^T$ drawn from a Dirichlet distribution. In this way the model selects a subset of the $\widetilde{N}$ covariates as landmarks.

The kernel is constructed as

$$\mathcal{K}(\boldsymbol{\ell}_i, \tilde{\boldsymbol{\ell}}_j) = \delta\big(j \in \mathcal{Q}_i\big) \exp(-\|\boldsymbol{\ell}_i - \tilde{\boldsymbol{\ell}}_j\|_2/\sigma) \tag{6}$$

where $\sigma$ is a constant and $\mathcal{Q}_i$ includes the indexes of the $L$ nearest landmarks of $\boldsymbol{\ell}_i$ from $\{\tilde{\boldsymbol{\ell}}_j\}_{j=1,J}$. We now have a random walk matrix $\mathbf{A} \in \Re^{N\times J}$, each row of which $\boldsymbol{a}_i \in \Re^J$ has $L$ nonzero components and sums to one. As $J$ increases to $N$ and $\tilde{\boldsymbol{\ell}}_j = \boldsymbol{\ell}_j$, the Landmark-dHBP becomes the dHBP, and as $J$ decreases to one, the Landmark-dHBP becomes HBP. An appropriate size of $J$ is selected based on a compromise between computational complexity and the desired level of covariant dependence.

A property analogous to (3) holds for the Landmark-dHBP model as well, and it is is true for *any* fixed set of landmarks, $\{\tilde{\boldsymbol{\ell}}_j\}$, which may not necessarily well reflect the dependencies between data. When using the proposed prior to fit data, the likelihood function encourages the posterior distribution on landmark positions to be well matched to the data, as discussed below in specific applications.

## 4. Robust Sparse Factor Analysis

As shown in (Zhou et al., 2009), a beta-Bernoulli process sparse factor analysis model can be constructed as

$$\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i), \gamma_\epsilon^{-1}\mathbf{I}_P) \tag{7}$$

$$\boldsymbol{d}_k \sim \mathcal{N}(0, P^{-1}\mathbf{I}_P),\ \boldsymbol{s}_i \sim \mathcal{N}(0, \gamma_s^{-1}\mathbf{I}_K) \tag{8}$$

$$z_{ik} \sim \text{Bernoulli}(\pi_k),\ \pi_k \sim \text{Beta}(c\eta, c(1-\eta)) \tag{9}$$

where $\boldsymbol{x}_i$ is the $i$th sample, $\odot$ represents the Hadamard product, $\mathbf{D} = [\boldsymbol{d}_1, \cdots, \boldsymbol{d}_K] \in \mathbb{R}^{P\times K}$ is the dictionary, $\boldsymbol{s}_i = [s_{i1}, \cdots, s_{iK}]^T$, $\boldsymbol{z}_i = [z_{i1}, \cdots, z_{iK}]^T$, $s_{ik} \in \mathbb{R}$, $z_{ik} = X_i(\boldsymbol{d}_k) \in \{0,1\}$ indicates whether the $k$th atom is *active* within sample $i$, and $\pi_k = B(\boldsymbol{d}_k)$ is the probability for the $k$th atom to be selected. Gamma hyper-priors are placed on both $\gamma_\epsilon$ and $\gamma_s$.

Motivated by recent increased interest in RPCA (Candès et al., 2009), we assume that in addition to the Gaussian noise $\boldsymbol{\epsilon}_i$, there are sparse spiky data components as

$$\boldsymbol{x}_i = \mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i) + \boldsymbol{\epsilon}_i + \boldsymbol{v}_i \odot \boldsymbol{m}_i \tag{10}$$

where $\boldsymbol{v}_i = [v_1, \cdots, v_P]^T$ and $\boldsymbol{m}_i = [m_1, \cdots, m_P]^T$, $v_{ip} \in \mathbb{R}$ is the weight and $m_{ip}$ is the binary spiky indicator. A beta-Bernoulli prior is constituted on $\boldsymbol{m}_i$ as

$$m_{ip} \sim \text{Bernoulli}(\tilde{\pi}_{ip}), \quad \tilde{\pi}_{ip} \sim \text{Beta}(a_0, b_0). \tag{11}$$

We impose $\boldsymbol{v}_i \sim \mathcal{N}(0, \gamma_v^{-1}\mathbf{I}_P)$ with a gamma hyper-prior on $\gamma_v$ and we have Gaussian noise with precision $\gamma_\epsilon$. After performing analysis with this model, the noise and outliers free data is estimated as $\hat{\boldsymbol{x}}_i = \mathbf{D}(\boldsymbol{s}_i \odot \boldsymbol{z}_i)$.

### 4.1. Landmark-dependent dictionary learning

Assuming we have data and associated covariates $\{\boldsymbol{x}_i, \boldsymbol{\ell}_i\}_{i=1,N}$, we wish to impose that if $\|\boldsymbol{\ell}_i - \boldsymbol{\ell}_{i'}\|$ is relatively small then $\boldsymbol{z}_i$ and $\boldsymbol{z}_{i'}$ will have similar sparseness

properties ($i.e.$, $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i'}$ will be constituted in terms of a similar set of the columns of $\mathbf{D}$). When employing the landmark-based construction in (4), (9) generalizes as

$$z_{ik} \sim \text{Bernoulli}(\pi_{ik}), \quad \pi_{ik} = \sum_{j=1}^{J} a_{ij}\pi^*_{jk} \tag{12}$$

$$\pi^*_{jk} \sim \text{Beta}(c_1\eta_k, c_1(1-\eta_k)) \tag{13}$$

$$\eta_k \sim \text{Beta}(c_0\eta_0, c_0(1-\eta_0)) \tag{14}$$

where $z_{ik} = X_{\boldsymbol{\ell}_i}(\boldsymbol{d}_k)$, $\pi_{ik} = B_{\boldsymbol{\ell}_i}(\boldsymbol{d}_k)$, $\pi^*_{jk} = B^*_{\tilde{\boldsymbol{\ell}}_j}(\boldsymbol{d}_k)$, $\eta_k = B(\boldsymbol{d}_k)$ and $a_{ij}$ is constructed via the landmarks as in (5). We use all the observed covariates as potential landmarks and learn the landmark locations with

$$\tilde{\boldsymbol{\ell}}_j = \boldsymbol{\ell}_{g_j}, \quad g_j \sim \sum_{i=1}^{N} \theta_i\delta_i \ , \ \boldsymbol{\theta} \sim \text{Dir}(\alpha_0, \cdots, \alpha_0). \tag{15}$$

In one application considered below, the objective is to denoise an image, potentially in the presence of non-Gaussian noise, the $\{\boldsymbol{x}_i\}$ correspond to pixels from image patches, and the associated covariates $\{\boldsymbol{\ell}_i\}$ locate patch positions within the image. For applications in which data reside on a manifold (or approximate manifold), we can define covariates as sample locations relative to each other. Specifically, we can define

$$\boldsymbol{\ell}_i = \boldsymbol{x}_i / \|\boldsymbol{x}_i\|_2 \tag{16}$$

and we have $\|\boldsymbol{\ell}_i - \boldsymbol{\ell}_{i'}\|_2 = \sqrt{2 - 2\cos(\boldsymbol{x}_i, \boldsymbol{x}_{i'})}$, where the cosine distance $\cos(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \frac{\boldsymbol{x}_i^T \boldsymbol{x}_{i'}}{\|\boldsymbol{x}_i\|_2 \|\boldsymbol{x}_{i'}\|_2}$ is widely used to measure similarity between vectors.

After the dictionary is learned on the training data, for the sparse coding of a new sample $\boldsymbol{x}_{i'}$, we can calculate its kernel distances to the $J$ landmarks and calculate $\boldsymbol{\pi}_{i'}$ as

$$a_{i'j} = \frac{\mathcal{K}(\boldsymbol{\ell}_{i'}, \tilde{\boldsymbol{\ell}}_j)}{\sum_{j'=1}^{J} \mathcal{K}(\boldsymbol{\ell}_{i'}, \tilde{\boldsymbol{\ell}}_{j'})}, \ \ \pi_{i'k} = \sum_{j=1}^{J} a_{i'j}\pi^*_{jk}. \tag{17}$$

With $\boldsymbol{\pi}_{i'}$ and the learned dictionary $\mathbf{D}$, we can infer the landmark-dependent sparse codes of $\boldsymbol{x}_{i'}$.

## 5. Example Results

The inference is performed using MCMC analysis. With the same robust term in (10) employed to model the sparse spiky noise, we apply BP, dHBP, and Landmark-dHBP as the prior for the robust sparse factor analysis (RSFA) model; our goal is to discover local latent features, and separate outliers from given data. Let $L_0$ denote the prior belief that a local region formed by a set of $L_0$ nearest samples share similar features, we set the kernel parameter as $L = \max\{\lfloor L_0 \frac{J}{N} \rceil, 2\}$, where $\lfloor \cdot \rceil$ represents the closest integer.

We first consider denoising an image corrupted by both sparse spiky noise and white Guassian noise. We use the same test procedure as that described in (Zhou et al., 2011) (omitted here for brevity). We find that as the landmark size increases from 1 to $N$ in Landmark-dHBP, the computational complexity increases and the performance generally improves until it saturates. Fig. 1 shows example results on denoising the House image.

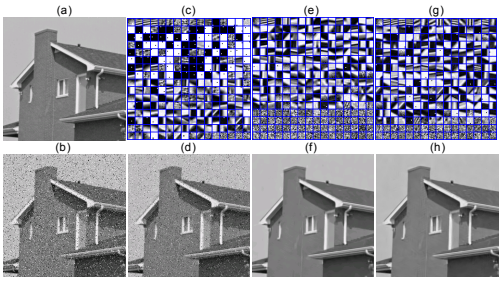


*Figure 1.* Denoisng results of BP, dHBP, Landmark-dHBP on "House" image, with 10% pixels corrupted by spiky noise situated uniformly at random. The spike amplitudes are uniformly distributed between -255 and 255. Zero-mean WGN with standard deviation 10 is also added to the image. (a) the original image; (b) the noisy image; (c), (e), and (g) show learned dictionaries; and (d), (f), and (h) the denoised images for BP (21.63 dB), dHBP (35.32 dB) and Landmark-dHBP (35.66 dB) with $J = \lfloor N/64 \rceil$, respectively.

The second example we consider is analyzing a document matrix $\mathbf{X} \in \mathbb{R}^{P\times N}$, whose $i$th column $\boldsymbol{x}_i$ encodes the relative number of times each term/word is manifested in the $i$th document. Word-count matrices of this form have usually been analyzed via latent semantic indexing (LSI) (Deerwester et al., 1990) or latent Dirichlet allocation (LDA) (Blei et al., 2003). Here we show that RSFA can infer not only easily interpretable topics sparsely used by documents, but also unique spiky keywords to each document, providing a novel way for document analysis. We use covariates defined in (16) for Landmark-dHBP.

We choose the widely studied NIPS dataset, available online at http://www.cs.nyu.edu/∼roweis/data.html, which includes 1740 conference papers from 1987 to 1999. We consider the most frequently used 2000 terms and use the *tf-idf* to construct $\mathbf{X} \in \mathbb{R}^{2000\times 1740}$. For better interpretation, we impose a non-negative constraint on both the dictionary and sparse codes with a truncated normal prior lower bounded by zero, and impose that only the top $N_s = 10$ most probable spiky terms can be captured. Each dictionary atom represents a topic, and the components of such a topic encode each term's relative importance in that topic. We set the landmark size $J = \lfloor N/2 \rceil = 870$ and kernel parameter $L = 16$, aiming to capture locally popular topics as well as distinct spiky keywords for each document. The dictionary size is truncated to be $K = 200$, and the atoms are initialized at random.

After learning, RSFA summarizes a document by a sparse set of topics and $N_s$ spiky keywords. RSFA with Landmark-dHBP discovers 104 topics, 80 of which have been used more than 10 times. For example, ranking the

| TOPIC 1 | TOPIC 2 | TOPIC 3 | TOPIC 4 | TOPIC 5 | TOPIC 6 | TOPIC 7 | TOPIC 8 | TOPIC 9 | TOPIC 10 | TOPIC 11 | TOPIC 12 | TOPIC 13 | TOPIC 14 | TOPIC 15 | TOPIC 16 | TOPIC 17 | TOPIC 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| units | training | neuron | image | mixture | estimate | cells | gradient | theorem | dynamics | bayesian | classifier | weight | synaptic | stimulus | analog | cortex | memory |
| unit | hidden | neurons | images | likelihood | variance | cell | descent | proof | dynamical | posterior | classifiers | weights | synapses | stimuli | vlsi | cortical | associative |
| hidden | layer | network | pixel | mixtures | error | inhibitory | learning | lemma | equations | prior | classification | error | synapse | response | chip | visual | stored |
| network | network | synapses | pixels | gaussian | sample | response | convergence | prove | fixed | monte | class | decay | presynaptic | responses | circuit | cells | memories |
| TOPIC 19 | TOPIC 20 | TOPIC 21 | TOPIC 22 | TOPIC 23 | TOPIC 24 | TOPIC 25 | TOPIC 26 | TOPIC 27 | TOPIC 28 | TOPIC 29 | TOPIC 30 | TOPIC 31 | TOPIC 32 | TOPIC 33 | TOPIC 34 | TOPIC 35 | TOPIC 36 |
| markov | bound | object | recognition | receptive | matrix | frequency | energy | speech | map | examples | firing | control | reinforcement | signal | convergence | tree | search |
| state | bounds | objects | character | field | eigenvalues | frequencies | hopfield | phoneme | mapping | learning | spike | controller | sutton | signals | state | trees | algorithms |
| states | upper | recognition | handwritten | fields | eigenvectors | filter | solution | recognition | kohonen | class | spikes | state | barto | noise | programming | node | algorithm |
| transition | vapnik | visual | recognize | cells | eigenvalue | auditory | constraint | speaker | maps | positive | rate | system | learning | filter | optimal | nodes | population |
| TOPIC 37 | TOPIC 38 | TOPIC 39 | TOPIC 40 | TOPIC 41 | TOPIC 42 | TOPIC 43 | TOPIC 44 | TOPIC 45 | TOPIC 46 | TOPIC 47 | TOPIC 48 | TOPIC 49 | TOPIC 50 | TOPIC 51 | TOPIC 52 | TOPIC 53 | TOPIC 54 |
| connectionist | nearest | belief | circuit | word | images | bit | functions | policy | decision | eq | generalization | clustering | phase | state | eye | orientation | threshold |
| cognitive | neighbor | graphical | voltage | words | principal | processor | approximation | action | table | fig | phys | clusters | oscillatory | recurrent | position | spatial | boolean |
| representation | class | variables | transistor | recognition | image | bits | basis | state | classification | matrix | student | cluster | oscillations | symbol | movements | tuning | polynomial |
| representations | neighbors | conditional | current | speech | recognition | hardware | approximate | actions | accuracy | learning | training | unsupervised | coupling | finite | movement | visual | gates |
| TOPIC 55 | TOPIC 56 | TOPIC 57 | TOPIC 58 | TOPIC 59 | TOPIC 60 | TOPIC 61 | TOPIC 62 | TOPIC 63 | TOPIC 64 | TOPIC 65 | TOPIC 66 | TOPIC 67 | TOPIC 68 | TOPIC 69 | TOPIC 70 | TOPIC 71 | TOPIC 72 |
| kernel | intensity | trajectories | blind | hmm | membrane | movement | matching | edge | natural | motor | support | kalman | ocular | perception | tangent | cues | algorithm |
| kernels | koch | trajectory | separation | speech | channels | movements | graph | regions | gaussian | animal | vector | filter | dominance | stage | distance | cue | algorithms |
| support | visual | position | sources | recognition | conductance | arm | matches | surface | scale | behavioral | machines | nonlinear | eyes | white | simard | spatial | prediction |
| margin | motion | velocity | source | training | koch | control | match | region | statistics | nervous | vapnik | prediction | development | visual | euclidean | environment | instance |

*Figure 2.* The top 72 NIPS topics discovered by Landmark-dHBP (most-probable four words in each topic).

| ID | Title, Authors | Topics | Top 10 Spiky Keywords |
|---|---|---|---|
| 716 | Supervised Learning from Incomplete Data via an EM Approach, Ghahramani_Z, Jordan_M | 5 6 | missing density em mixture supervised expectation gaussians jacobs valued estimator |
| 1383 | Stacked Density Estimation, Smyth_P, Wolpert_D | 5 6 11 | density kernel kernels combining estimation estimators partition vowel weighting sample |
| 1521 | Batch and On-Line Parameter Estimation of Gaussian Mixtures Based on the Joint Entropy, Singer_Y, Warmuth_M | 5 8 20 24 | em update loss eq entropy estimation updates distance parameter setting |
| 1524 | SMEM Algorithm for Mixture Models, Ghahramani_Z, Hinton_G, Nakano_R, Ueda_N | 5 22 38 | split em operations criteria manifold usual mixture fig posterior algorithm |
| 1509 | Maximum Conditional Likelihood via Bound Maximization and the CEM Algorithm, Jebara_T, Pentland_A | 5 20 | conditional em gate bound experts likelihood maximization gates joint update |
| 1525 | Learning Mixture Hierarchies, Lippman_A, Vasconcelos_N | 5 6 4 22 62 21 | em hierarchy mixture sample hierarchical object kernel level bottom smoothing |
| 1654 | Variational Inference for Bayesian Mixtures of Factor Analysers, Beal_M, Ghahramani_Z | 5 11 24 | variational bayesian factor predictive posterior evidence overfitting intrinsic exact inference |
| 924 | ESTIMATING CONDITIONAL PROBABILITY DENSroES FOR PERIODIC VARIABLES, Bishop_C, Legleye_C | 5 44 6 | periodic conditional kernel density kernels bishop densities estimating euclidean angle |
| 717 | Training Neural Networks with Deficient Data, Ahmad_S, Neuneier_R, Tresp_V | 5 2 6 44 11 | missing uncertainty integral unknown closed noise complete convolution solutions network |

*Figure 3.* Topics and spiky keywords for the first 8 retrieved papers for Paper 716.

topics and the words in a topic based on their usage probabilities, the top 8 words of Landmark-dHBP Topic 73 are "segmentation frame category speaker segment classifier boundaries classification," and shown in Fig. 2 are the top 72 topics in the corpora, which are all very easily interpretable. The average number of topics used per document is about 4.80. The unique number of landmark locations is 259. The results presented are based on the last MCMC sample of 2500 iterations, for simple interpretation.

We also tested using the BP. We found that BP only discovers 67 topics, 65 of which have been used more than 6 times. Examining the correlation coefficients between the top 65 BP topics and the top 80 Landmark-dHBP topics, as shown in Fig. 4 (a), we find that most of the BP topics are highly correlated with one of the Landmark-dHBP topics. Examining the not well correlated ones, we find that they are either not obviously interpretable, or evenly correlated with multiple Landmark-dHBP topics. For example, shown in Figs. 4 (b)-(d) are the correlation coefficients between BP Topics 10, 56 and 62 and the top 80 Landmark-dHBP topics. Comparing these three BP topics with their correlated Landmark-dHBP topics, we can find that a BP topic of this kind is the superposition of several related but distinct Landmark-dHBP topics. This suggests that BP tends to model some distinct but related topics as a single topic. This is not surprising, since BP is trying to model *globally* popular features while Landmark-dHBP is able to discover *locally* popular topics as defined by regions near landmarks. This also suggests that the covariate dependence plays an important role in revealing locally popular features, and hence to better model the corpora.

We pick a widely cited NIPS paper, and retrieve the 8 closest papers based on the cosine distance between their topic

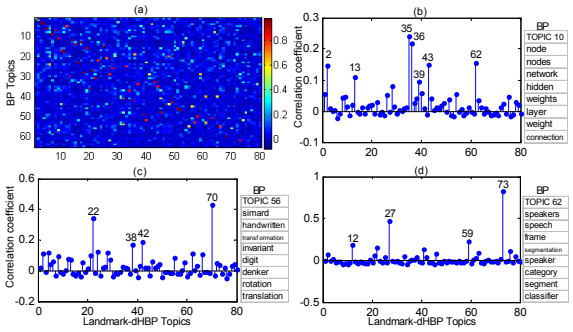


*Figure 4.* Correlation coefficients between the top 65 BP topics (dictionary atoms) and the top 80 Landmark-dHBP topics.

usage weights (sparse codes). Shown in Fig. 3 are the topics and 10 spiky keywords discovered by RSFA for these papers. We find that the captured spiky keywords well correspond to a paper's unique aspects compared to its neighbors, and the underlying topics are shared often, but can also be different.

## References


Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 2003.

Candès, E.J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *to apear in Journal of the ACM*, 2009.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. Indexing by latent semantic analysis. *J. Amer. Soc. Inf. Sci.*, 1990.

Thibaux, R. and Jordan, M. I. Hierarchical beta processes and the Indian buffet process. In *AISTATS*, 2007.

Zhou, M., Chen, H., Paisley, J., Ren, L., Sapiro, G., and Carin, L. Non-parametric bayesian dictionary learning for sparse image representations. In *NIPS*, 2009.

Zhou, M., Yang, H., Sapiro, G., Dunson, D.B., and Carin, L. Dependent hierarchical beta process for image interpolation and denoising. In *AISTATS*, 2011.