# FACTORIZED TOPIC MODELS

Cheng Zhang, Carl Henrik Ek, Andreas Damianou, Hedvig Kjellström

{chengz, chek, hedvig}@kth.se, andreas.damianou@sheffield.ac.uk

## ABSTRACT

We present a modification to a latent topic model, which makes the model exploit supervision to produce a factorized representation of the observed data. The structured parameterization separately encodes variance that is shared between classes from variance that is private to each class by the introduction of a new prior over the topic space. The approach allows for a more efficient inference and provides an intuitive interpretation of the data in terms of an informative signal together with structured noise. The factorized representation is shown to enhance inference performance for image, text, and video classification.
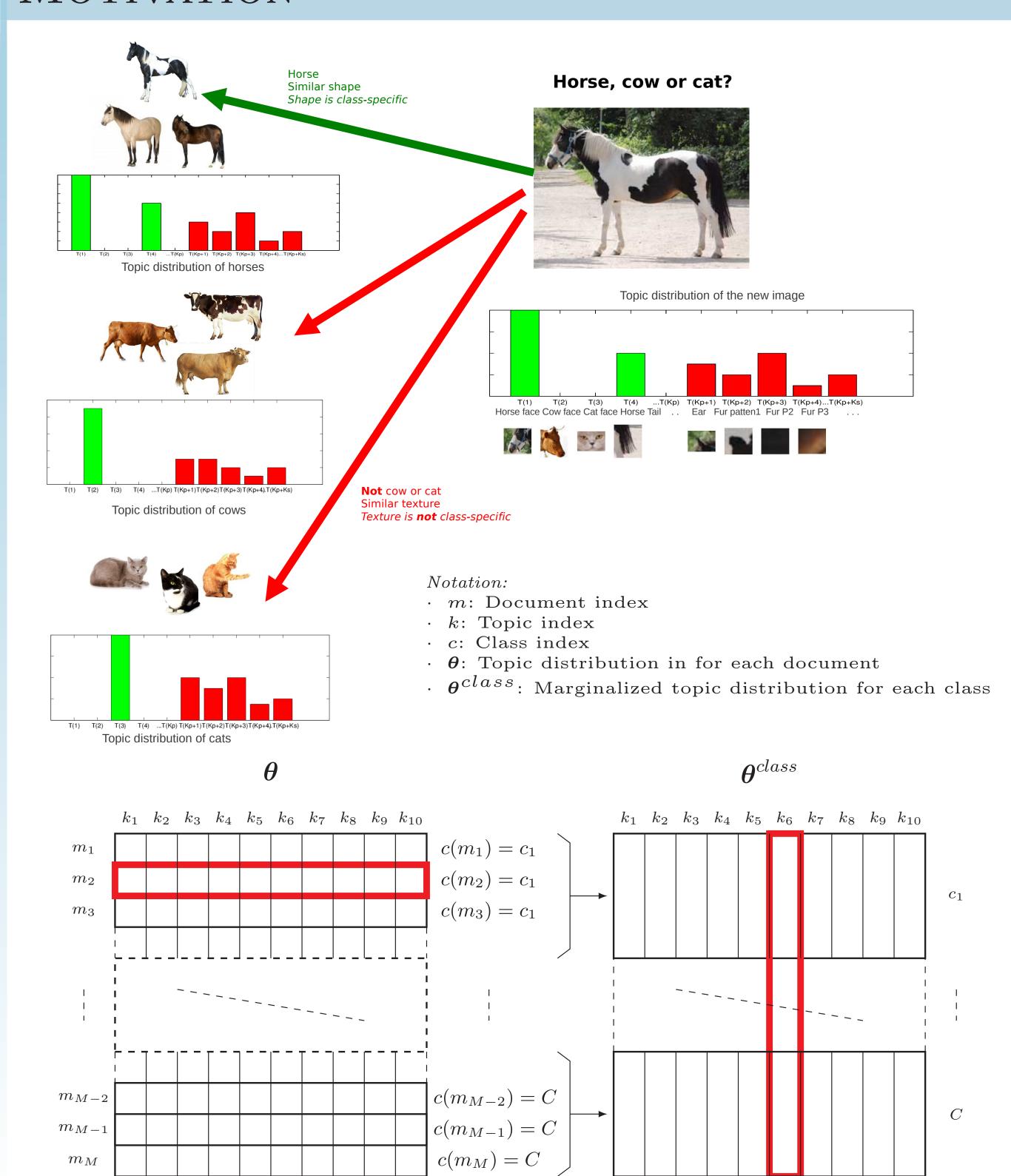
## MOTIVATION



Notation:
- $m$: Document index
- $k$: Topic index
- $c$: Class index
- $\boldsymbol{\theta}$: Topic distribution in for each document
- $\boldsymbol{\theta}^{class}$: Marginalized topic distribution for each class



Illustration of topic distribution $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{class}$. K=10 is used here as an example
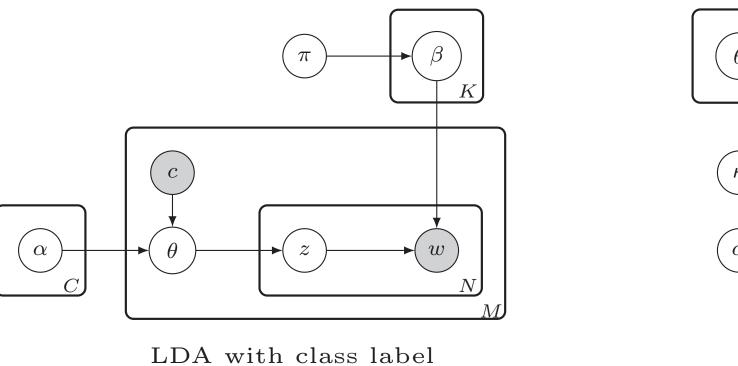
## ACKNOWLEDGEMENT

## REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. *Latent Dirichlet Allocation* Journal of Machine Learning Research, 3:993-1022, 2003

[2] L. Fei-Fei and P. Perona. *A Bayesian hierarchical model for learning natural scene categories* In CVPR, 2005.

[3] D. M. Blei and J. D. McAuliffe. *Supervised topic models* In NIPS, 2007.

## MODEL



LDA with class label        Our factorized LDA

We employ an entropy-like measure $H(k)$ over class for each topic $k$:

$$H(k) = \frac{1}{\log(1/C)} \sum_{c=1}^{C} \frac{\theta_{c,k}^{class}}{\sum_{\xi=1}^{C} \theta_{\xi,k}^{class}} \log(\frac{\theta_{c,k}^{class}}{\sum_{\xi=1}^{C} \theta_{\xi,k}^{class}})$$

- High $H(k) \rightarrow$ class-independent topics
- Low $H(k) \rightarrow$ class-depended topics

Hence we introduce the following function:

$$A(k) = H(k)^2 - H(k) + 1$$



The static function A

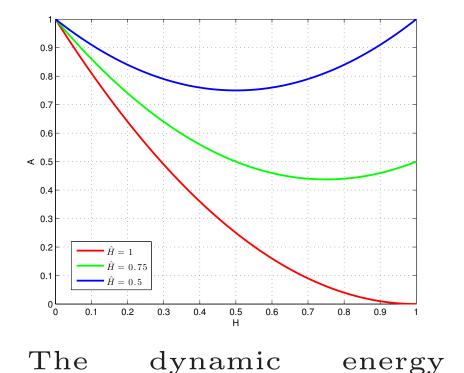We define the new prior as:

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} A(k) ,$$

which treats each column of $\boldsymbol{\theta}^{class}$ independently.

With the additional prior, the generative model becomes:

$$p(\mathbf{w}|\alpha, \pi, c) = \int p(\boldsymbol{\theta}|\alpha, c)p(\boldsymbol{\theta}) \left( \prod_{n=1}^{N} \left[ \sum_{z_n} p(z_n|\boldsymbol{\theta})p(w_n|z_n, \boldsymbol{\beta}) \right] p(\boldsymbol{\beta}|\pi) \right) d\boldsymbol{\theta} .$$



We introduce an "auto-annealing" procedure, where $A$ is changed to dynamic function, since $H$ is close to 1 during initialization.

$$A(k) = H(k)^2 - 2\hat{H}H(k) + 1 , \quad (1)$$

where the average $H$, $\hat{H} = \sum_{k=1}^{K} H(k)/K$, is used as an annealing parameter in the function.

The dynamic energy optimization function $A$

We consider topics with low $H$ as class-dependent while topics with high $H$ are considered as independent.

An additional factor $\mathcal{F}_{ck}$ is introduced in the update equation of the Gibbs Sampling,

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}, \alpha, \pi, \kappa) \propto \frac{n_{k,\neg i}^{(t)} + \pi_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \pi_t} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha_k}{[\sum_{k=1}^{K} n_m^{(k)} + \alpha_k] - 1} \cdot \mathcal{F}_{ck}.$$

In the updating procedure for $\mathcal{F}_{ck}$ a gradient step is carried out by measuring the possible change with respect to the prior $\Delta A(k) = A^{assume}(k) - A(k)$, where $A^{assume}(k)$ is the evaluation of Eq. 1 for the current sample. $\Delta A(k) \in (-1, 1)$. Thus the factor takes the following form,
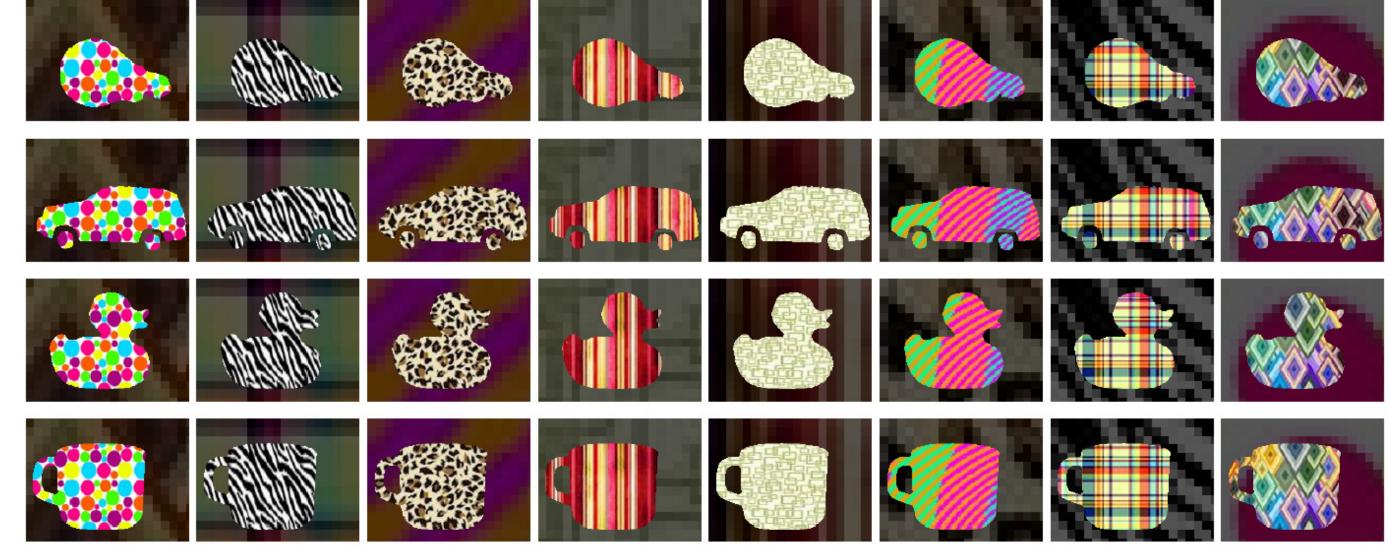
$$\mathcal{F}_k = 1 + \Delta A(k).$$

We introduce a parameter $\kappa$ that controls the contribution of this term, hence

$$\begin{cases} \mathcal{F}_k = 1 + |\Delta A(k)|^{\kappa}, & if \ \Delta A(k) > 0 \\ \mathcal{F}_k = 1 - |\Delta A(k)|^{\kappa}, & if \ \Delta A(k) < 0. \end{cases}$$
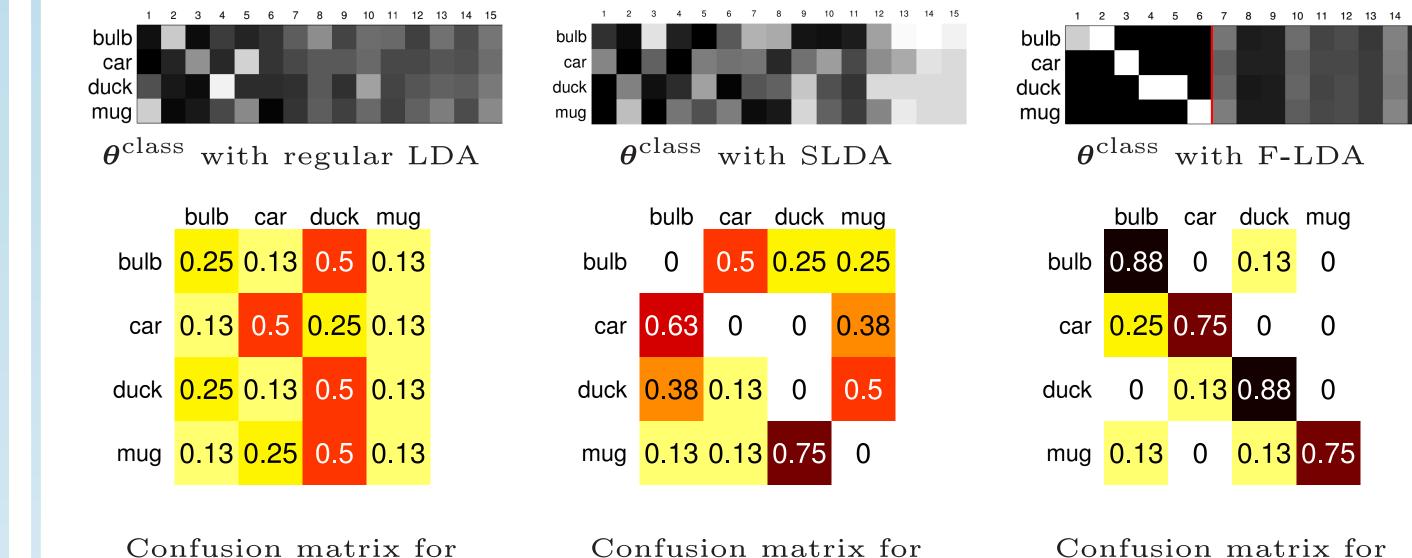
## EXPERIMENTS

We evaluated our model(F-LDA) against regular LDA model [2] and SLDA [3] on four different classification tasks with different data.
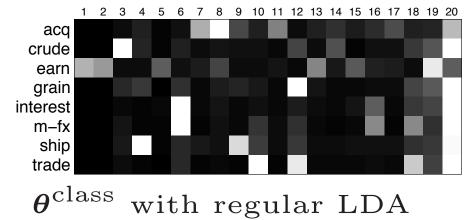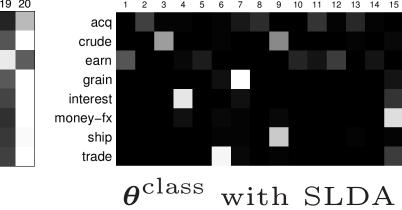
- **Object Classification**



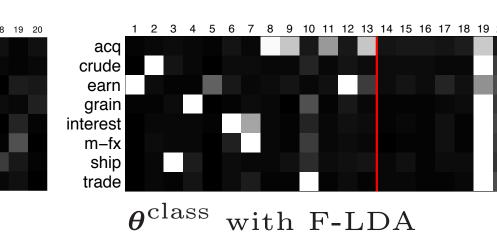The toy dataset used in the object classification



$\boldsymbol{\theta}^{class}$ with regular LDA    $\boldsymbol{\theta}^{class}$ with SLDA    $\boldsymbol{\theta}^{class}$ with F-LDA

| | bulb | car | duck | mug |
|---|---|---|---|---|
| bulb | 0.25 | 0.13 | 0.5 | 0.13 |
| car | 0.13 | 0.5 | 0.25 | 0.13 |
| duck | 0.25 | 0.13 | 0.5 | 0.13 |
| mug | 0.13 | 0.25 | 0.5 | 0 |

| | bulb | car | duck | mug |
|---|---|---|---|---|
| bulb | 0 | 0.5 | 0.25 | 0.25 |
| car | 0.63 | 0 | 0 | 0.38 |
| duck | 0.38 | 0.13 | 0 | 0.5 |
| mug | 0.13 | 0.13 | 0.75 | 0 |

| | bulb | car | duck | mug |
|---|---|---|---|---|
| bulb | 0.88 | 0 | 0.13 | 0 |
| car | 0.25 | 0.75 | 0 | 0 |
| duck | 0 | 0.13 | 0.88 | 0 |
| mug | 0.13 | 0 | 0.13 | 0.75 |

Confusion matrix for regular LDA: 34.38%    Confusion matrix for SLDA: 0%    Confusion matrix for F-LDA: 81.25%

- **Text Classification**



$\boldsymbol{\theta}^{class}$ with regular LDA    $\boldsymbol{\theta}^{class}$ with SLDA    $\boldsymbol{\theta}^{class}$ with F-LDA

- **Scene Classification**
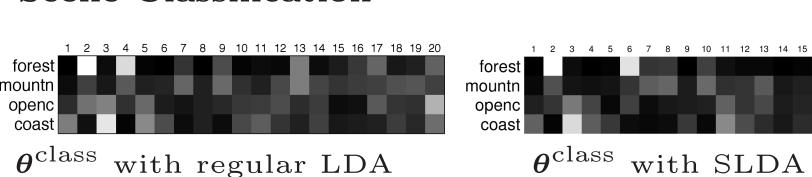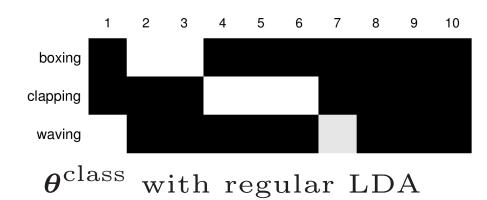


$\boldsymbol{\theta}^{class}$ with regular LDA    $\boldsymbol{\theta}^{class}$ with SLDA    $\boldsymbol{\theta}^{class}$ with F-LDA

- **Action Classification**



$\boldsymbol{\theta}^{class}$ with regular LDA    $\boldsymbol{\theta}^{class}$ with SLDA    $\boldsymbol{\theta}^{class}$ with F-LDA

| Performance | Regular LDA | SLDA | Factorized LDA |
|---|---|---|---|
| Text Classification | 74.63% | 63.75% | 83.91% |
| Scene Classification | 80.50% | 84.00% | 84.50% |
| Action Classification | 38% | 52.33% | 65.22% |