Supplementary Materials for Tensor Analyzers

Yichuan Tang
Ruslan Salakhudinov
Geoffrey Hinton

Department of Computer Science
University of Toronto
Toronto, Ontario, Canada.
tang@cs.toronto.edu

1 Derivations of the M-step in EM learning of TAs

The objective function Q of the EM algorithm is the expected complete log-likelihood, taken over the posterior distribution of the latent factors, and summed over N training cases:

$$Q = E\Big[\log \prod_i^N (2\pi)^{-D/2} |\mathbf{\Psi}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}_i - \mathbf{e}_i)^\mathsf{T} \mathbf{\Psi}^{-1}(\mathbf{x}_i - \mathbf{e}_i)\}\Big] \tag{1}$$

$$= const - \frac{N}{2}\log|\mathbf{\Psi}| - \sum_i^N E\Big[\frac{1}{2}\mathbf{x}_i^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{x}_i - \mathbf{x}_i^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{e}_i + \frac{1}{2}\mathbf{e}_i^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{e}_i\Big] \tag{2}$$

$$= const - \frac{N}{2}\log|\mathbf{\Psi}| - \sum_i^N E\Big[\frac{1}{2}\mathbf{x}_i^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{x}_i - \mathbf{x}_i^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{y}_i - \mathbf{x}_i^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{T}_{(1)}\mathbf{u}_i$$
$$+ \mathbf{y}_i^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{T}_{(1)}\mathbf{u}_i + \frac{1}{2}\mathbf{y}_i^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{y}_i + \frac{1}{2}\mathbf{u}_i^\mathsf{T}\mathbf{T}_{(1)}^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{T}_{(1)}\mathbf{u}_i\Big] \tag{3}$$

$$= const - \frac{N}{2}\log|\mathbf{\Psi}| - \sum_i^N \Big(\frac{1}{2}\mathbf{x}_i^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{x}_i - \mathbf{x}_i^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{W}E[\mathbf{y}_i] - \mathbf{x}_i^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{T}_{(1)}E[\mathbf{u}_i]$$
$$+ E[\mathbf{u}_i^\mathsf{T}\mathbf{T}_{(1)}^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{y}_i] + \frac{1}{2}E[\mathbf{y}_i^\mathsf{T}\mathbf{W}^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{W}\mathbf{y}_i] + \frac{1}{2}E[\mathbf{u}_i^\mathsf{T}\mathbf{T}_{(1)}^\mathsf{T}\mathbf{\Psi}^{-1}\mathbf{T}_{(1)}\mathbf{u}_i]\Big) \tag{4}$$

The closed-form M-step update equations are:

$$\frac{\partial Q}{\partial \mathbf{W}} = -\sum_i^N \Big(-\mathbf{\Psi}^{-1}\mathbf{x}_i E[\mathbf{y}_i]^\mathsf{T} + \mathbf{\Psi}^{-1}\mathbf{T}_{(1)}E[\mathbf{u}_i\mathbf{y}_i^\mathsf{T}] + \mathbf{\Psi}^{-1}\mathbf{W}E[\mathbf{y}_i\mathbf{y}_i^\mathsf{T}]\Big) = 0 \tag{5}$$

$$\mathbf{W} = \Big(\sum_i^N \mathbf{x}_i E[\mathbf{y}_i^\mathsf{T}] - \mathbf{T}_{(1)}\sum_i^N E[\mathbf{u}_i\mathbf{y}_i^\mathsf{T}]\Big)\Big(\sum_i^N E[\mathbf{y}_i\mathbf{y}_i^\mathsf{T}]\Big)^{-1} \tag{6}$$

$$\frac{\partial Q}{\partial \mathbf{T}_{(1)}} = -\sum_i^N \Big(-\mathbf{\Psi}^{-1}\mathbf{x}_i E[\mathbf{u}_i^\mathsf{T}] + \mathbf{\Psi}^{-1}\mathbf{W}E[\mathbf{y}_i\mathbf{u}_i^\mathsf{T}] + \frac{1}{2}\mathbf{\Psi}^{-1}\mathbf{T}_{(1)}(2)E[\mathbf{u}_i\mathbf{u}_i^\mathsf{T}]\Big) = 0 \tag{7}$$

$$\mathbf{T}_{(1)} = \Big(\sum_i^N \mathbf{x}_i E[\mathbf{u}_i^\mathsf{T}] - \mathbf{W}\sum_i^N E[\mathbf{y}_i\mathbf{u}_i^\mathsf{T}]\Big)\Big(\sum_i^N E[\mathbf{u}_i\mathbf{u}_i^\mathsf{T}]\Big)^{-1} \tag{8}$$

$$\frac{\partial Q}{\partial \mathbf{\Psi}^{-1}} = \frac{N}{2}\mathbf{\Psi} - \sum_i^N \Big(\frac{1}{2}\mathbf{x}_i\mathbf{x}_i^\mathsf{T} - \mathbf{W}E[\mathbf{y}_i]\mathbf{x}_i^\mathsf{T} - \mathbf{T}_{(1)}E[\mathbf{u}_i]\mathbf{x}_i^\mathsf{T} + \mathbf{T}_{(1)}E[\mathbf{u}_i\mathbf{y}_i^\mathsf{T}]\mathbf{W}^\mathsf{T} + \frac{1}{2}\mathbf{W}E[\mathbf{y}_i\mathbf{y}_i^\mathsf{T}]\mathbf{W}^\mathsf{T} + \frac{1}{2}\mathbf{T}_{(1)}E[\mathbf{u}\mathbf{u}^\mathsf{T}]\mathbf{T}_{(1)}^\mathsf{T}\Big) = 0 \tag{9}$$

$$\mathbf{\Psi} = \frac{1}{N}diag\Bigg\{\sum_i^N \Big(\mathbf{x}_i\mathbf{x}_i^\mathsf{T} - 2\mathbf{W}E[\mathbf{y}_i]\mathbf{x}_i^\mathsf{T} - 2\mathbf{T}_{(1)}E[\mathbf{u}_i]\mathbf{x}_i^\mathsf{T} + 2\mathbf{T}_{(1)}E[\mathbf{u}_i\mathbf{y}_i^\mathsf{T}]\mathbf{W}^\mathsf{T} + \mathbf{W}E[\mathbf{y}_i\mathbf{y}_i^\mathsf{T}]\mathbf{W}^\mathsf{T} + \mathbf{T}_{(1)}E[\mathbf{u}\mathbf{u}^\mathsf{T}]\mathbf{T}_{(1)}^\mathsf{T}\Big)\Bigg\} \tag{10}$$

$$= \frac{1}{N}diag\Bigg\{\sum_i^N \Big(\mathbf{x}_i\mathbf{x}_i^\mathsf{T} - 2\mathbf{T}_{(1)}\big(E[\mathbf{u}_i]\mathbf{x}_i^\mathsf{T} - E[\mathbf{u}_i\mathbf{y}_i^\mathsf{T}]\mathbf{W}^\mathsf{T} - \frac{1}{2}E[\mathbf{u}\mathbf{u}^\mathsf{T}]\mathbf{T}_{(1)}^\mathsf{T}\big) - 2\mathbf{W}\big(E[\mathbf{y}_i]\mathbf{x}_i^\mathsf{T} - \frac{1}{2}E[\mathbf{y}_i\mathbf{y}_i^\mathsf{T}]\mathbf{W}^\mathsf{T}\big)\Big)\Bigg\} \tag{11}$$

2 Convergence of Gibbs Sampling

Posterior inference in TA using alternating Gibbs sampling is efficient. We present trace plots of the 6 random latent factors of two TAs in the figure below. Left panel is from a TA learned on 2D synthetic datasets of Sec. 4, while the right panel is from a TA modeling high dimensional face images under illumination variations. The plots demonstrate that samples mixes very quickly after around 20 Gibbs iterations.

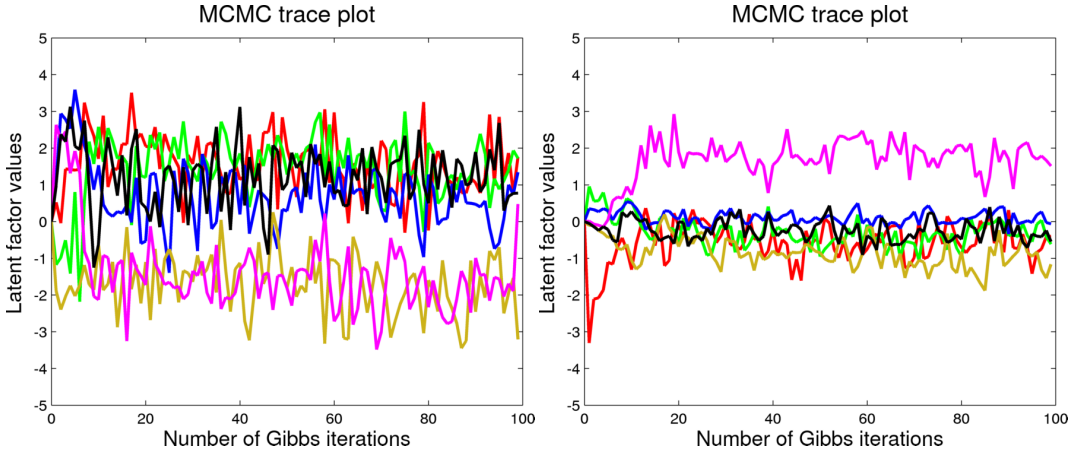

Figure 1: ***left**: MCMC trace plot for TA learning on synthetic data. **right**:MCMC trace plot for TA learning on high dimensional face images.*

We also looked at how posterior inference converges in a 200 component MTA trained on 8 x 8 natural image patches.

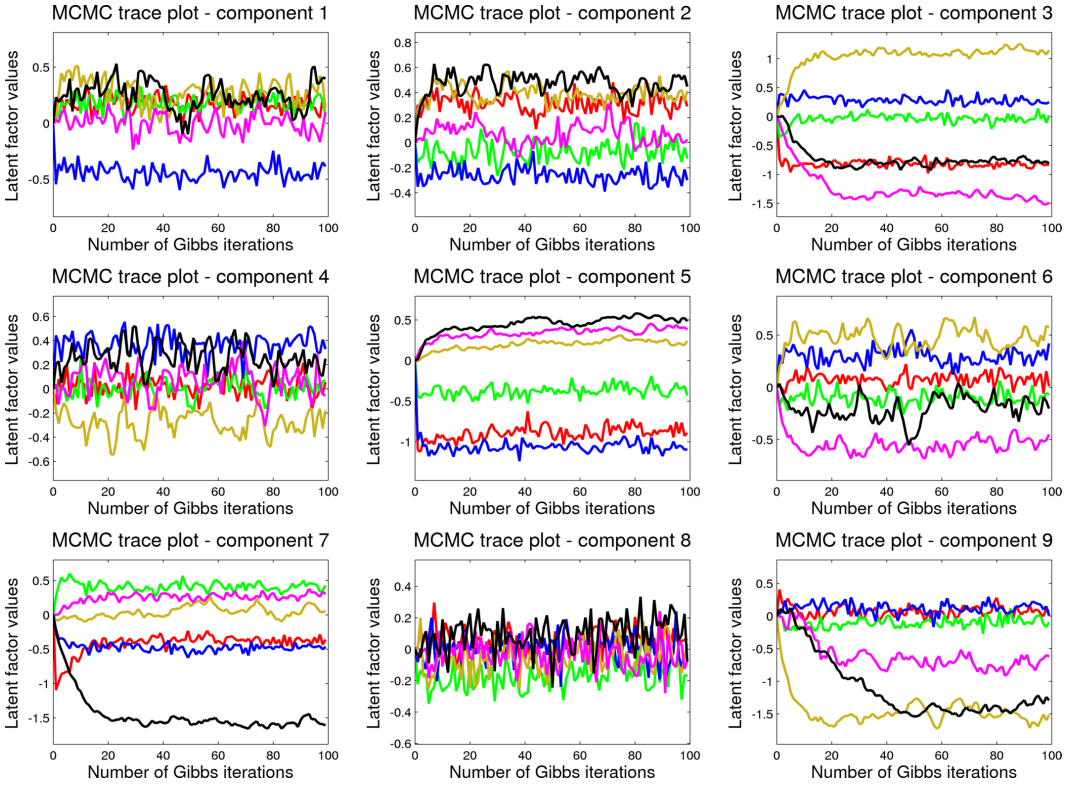

Figure 2: Gibbs sampling quickly converges around 20 iterations. 9 components were randomly picked from a MTA with 200 components trained on natural image patches. For each component, we randomly selected 6 latent factors (one for every color).

3 Annealed Importance Sampling for TA

We can treat the problem of estimating $\log p(\mathbf{x})$ as calculating the partition function of unnormalized posterior distribution $p^*(\mathbf{z}|\mathbf{x}) \triangleq p(\mathbf{x}, \mathbf{z})$, where $p^*(\cdot)$ denotes an unnormalized distribution. The basic Importance Sampling gives:

$$p(\mathbf{x}) = \int_{\mathbf{z}} d\mathbf{z}\, p(\mathbf{x}, \mathbf{z}) \tag{12}$$

$$p(\mathbf{x}) = \int_{\mathbf{z}} d\mathbf{z}\, \frac{p^*(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} q(\mathbf{z}) \tag{13}$$

$$p(\mathbf{x}) = \frac{Z_p}{1} \simeq \frac{1}{M} \sum_i^M w^{(i)}, \quad w^{(i)} = \frac{p^*(\mathbf{z}^{(i)}|\mathbf{x})}{q(\mathbf{z}^{(i)})}, \quad z^{(i)} \sim q(\mathbf{z}) \tag{14}$$

Annealed Importance Sampling specifies a set of intermediate distributions, where β varies from 0.0 to 1.0.

$$p_\beta(\mathbf{z}) \propto q(\mathbf{z})^{1-\beta} p^*(\mathbf{z}|\mathbf{x})^\beta \tag{15}$$

For TAs, the $\log$ of the tractable base distribution $q(\mathbf{z})$ is:

$$\log q(\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_J\}) = \sum_{j=1}^{J} \Big(-\frac{d_j}{2} \log(2\pi) - \frac{1}{2} \mathbf{z}_j^\top \mathbf{z}_j \Big), \tag{16}$$

which is simply the prior distribution over the latent factors.

Since we are using the prior as the base distribution and the unnormalized posterior distribution is the distribution of interest, we can write the intermediate distribution as:

$$p_\beta(\{\mathbf{z}_j\}) \propto q(\{\mathbf{z}_j\})^{1-\beta} p(\mathbf{x}, \{\mathbf{z}_j\})^\beta = p(\{\mathbf{z}_j\}) p^\beta(\mathbf{x}|\{\mathbf{z}_j\}) \tag{17}$$

Algorithm 1 AIS for TA

let $k = 1, 2, \ldots, K$, $\beta_{k=1} = 0$, $\beta_{k=K} = 1$.
for $i = 1$ to M **do**
  Draw $\mathcal{Z}_2$ from $q(\mathcal{Z})$
  $w^{(i)} = \frac{p_{\beta_2}(\mathcal{Z}_2)}{p_{\beta_1}(\mathcal{Z}_2)}$
  for $k = 2$ to $K - 1$ **do**
    Sample $\mathcal{Z}_{k+1}$ from $T_{\beta_k}(\mathcal{Z}' \leftarrow \mathcal{Z}_k)$
    $w^{(i)} = w^{(i)} \times \frac{p_{\beta_{k+1}}(\mathcal{Z}_{k+1})}{p_{\beta_k}(\mathcal{Z}_{k+1})}$
  end for
end for
$p(\mathbf{x}) \simeq \frac{1}{M} \sum_i^M w^{(i)}$

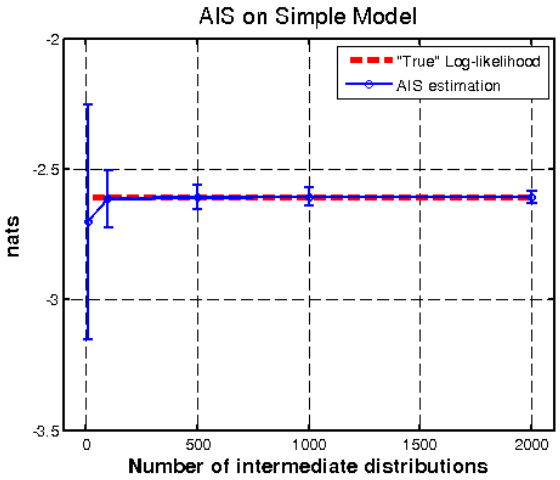

Figure 3: ***left**: AIS algorithm for TA. **right**: Experimental validation of AIS on a small model, where 100,000 Monte Carlo samples estimate the "true" log-likelihood.*

Therefore,

$$\log p_\beta(\{\mathbf{z}_j\}) = \sum_j^J \Big(-\frac{d_j}{2}\log(2\pi) - \frac{1}{2}\mathbf{z}_j^\mathsf{T}\mathbf{z}_j \Big) - \frac{\beta D}{2}\log(2\pi) - \frac{\beta}{2}\log|\mathbf{\Psi}| - \frac{\beta}{2}(\mathbf{x}-\mathbf{e})^\mathsf{T}\mathbf{\Psi}^{-1}(\mathbf{x}-\mathbf{e}) + C(\beta) \tag{18}$$

To ensure that $p_\beta(\{\mathbf{z}_j\})$ sums to 1.0, we find that:

$$C(\beta) = \frac{\beta}{2}\log|\mathbf{\Psi}| + \frac{\beta D}{2}\log(2\pi) - \frac{1}{2}\log|\beta\mathbf{\Psi}| - \frac{D}{2}\log(2\pi) \tag{19}$$

Therefore,

$$\log p_\beta(\{\mathbf{z}_j\}) = \sum_j^J \Big(-\frac{d_j}{2}\log(2\pi) - \frac{1}{2}\mathbf{z}_j^\mathsf{T}\mathbf{z}_j \Big) - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\beta\mathbf{\Psi}| - \frac{1}{2}(\mathbf{x}-\mathbf{e})^\mathsf{T}(\beta\mathbf{\Psi}^{-1})(\mathbf{x}-\mathbf{e}) \tag{20}$$

For AIS, we also need a MCMC operator which leaves $p_\beta(\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_J\})$ invariant. We use a Gibbs sampler, which simply performs alternating Gibbs sampling of the TA's posterior where the original diagonal noise $\mathbf{\Psi}$ is modified to be $\beta\mathbf{\Psi}$. We denote this operator as $T_\beta(\mathbf{z}' \leftarrow \mathbf{z})$.

In Fig. 3 (left), We present the AIS algorithm. For clarity, we use $\mathcal{Z}_k$ to denote the set of all latent factors $\mathcal{Z} = \{\mathbf{z}_j\}$ of the k-th intermediate distribution, $k = 1, 2, \ldots, K$. M is the number of independent AIS chains.

On the right panel of Fig. 3, we experimented with the variance of the AIS estimator. Using a small model of TA{2,2,2} trained on a 2D dataset from Sec. 4.1 of the main paper, we ran AIS algorithm with varying number of intermediate distributions to estimate the average data log-likelihood. M is set to 10 in all experiments. In the plot, we can see that the variance of the estimator quickly shrinks to less than 0.1 nats as we use 500 or more intermediate distributions. The dashed line represent the estimated log-likelihood by sampling from the prior (Sec. 3.3 of main paper), using 100,000 samples. Since we are sampling from only 2 dimensions, this Monte Carlo estimator has very low variance and its value is taken to be the true data log-likelihood.

4 Experiments - Synthetic Data

We present both the training and test average log-likelihood on 4 2D synthetic datasets in Table 1. The first two is presented in the main paper. (M)TAs are better models for complicated density and just as good as MFAs on the simpler ones.

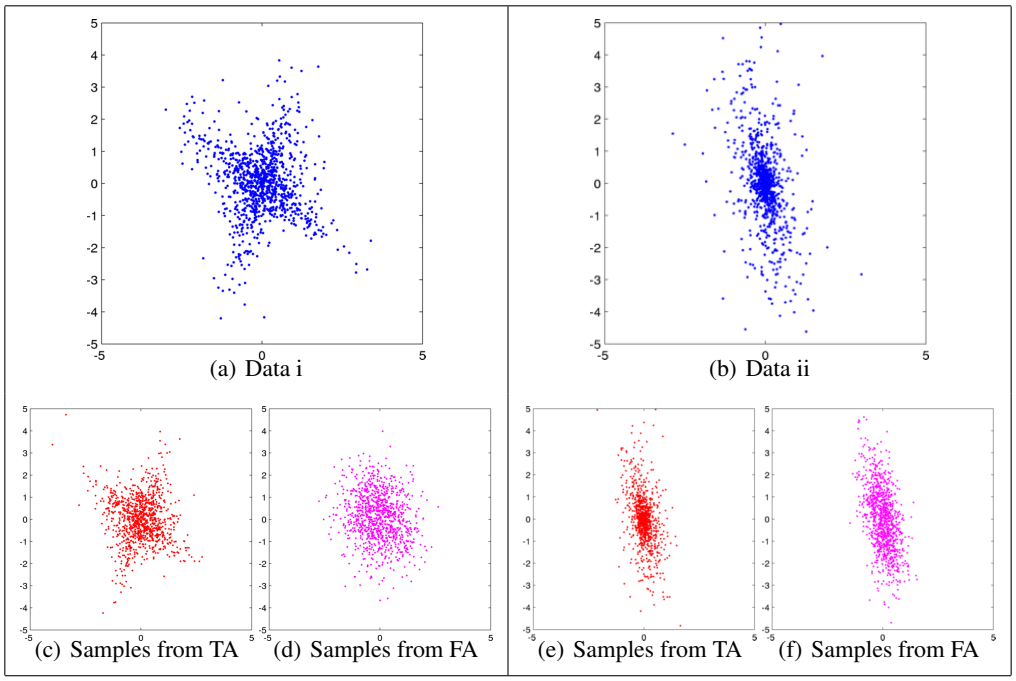

Figure 4: *TA vs. FA on 2D synthetic data.*

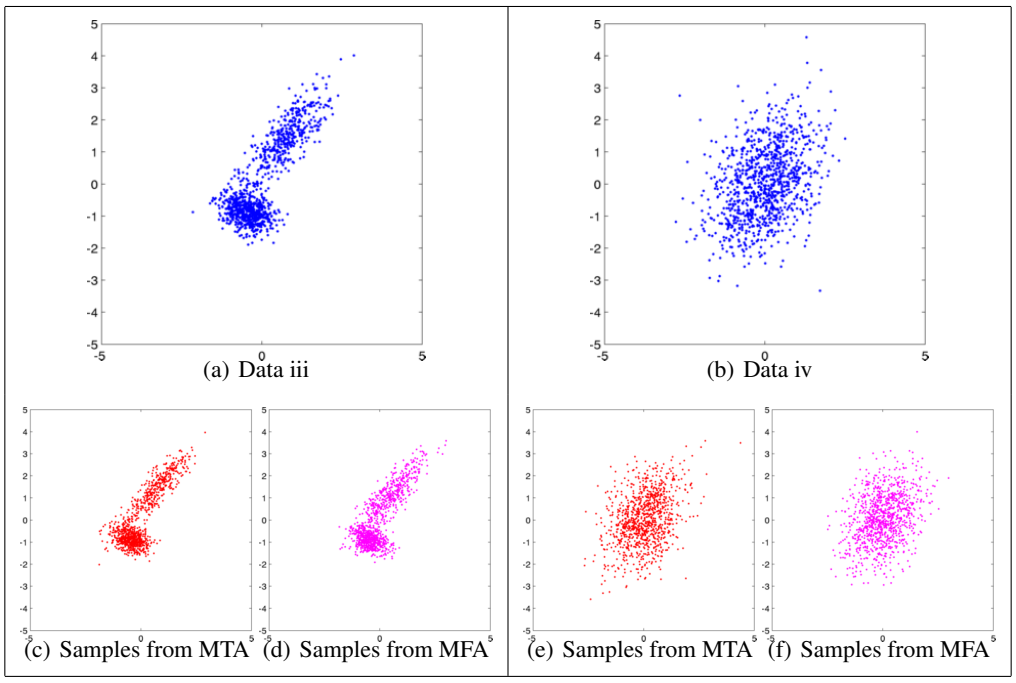

Figure 5: *MTA vs. MFA on 2D synthetic data. A mixture of two components are used in both MTA and MFA.*

Data i and ii are generated by a randomly initialized TA model, while Data iii and iv are generated by a randomly initialized MFA model.

Dataset	i	ii	iii	iv
MTA	-2.62 (-2.58)	-2.04 (-1.90)	-1.81 (-1.75)	-2.76 (-2.74)
MFA	-2.85 (-2.78)	-2.48 (-2.37)	-1.82 (-1.76)	-2.75 (-2.73)

Table 1: *Average test and (training) log-likelihood in nats of Mixture of Tensor Analyzers and Mixture of Factor Analyzers on the 2D synthetic datasets*.