IBM **Research**

Multi-task Multi-modal Models for Collective Anomaly Detection

Tsuyoshi Ide (“**Ide-san**”), Dzung T. Phan, J. Kalagnanam
PhD, Senior Technical Staff Member
IBM Thomas J. Watson Research Center

This slides are available at ide-research.net.

IEEE International Conference on Data Mining (ICDM 2017)

Outline

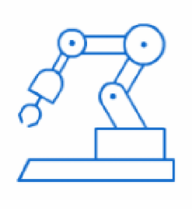
- Problem setting
- Modeling strategy
- Model inference approach
- Experimental results

Wish to build a collective monitoring solution

System 1
(in New Orleans)

⋮

System *s*



⋮

System *S*
(in New York)

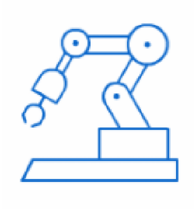
- You have many similar but not identical industrial assets
- You want to build an anomaly detection model for each of the assets
- Straightforward solutions have serious limitations
  - 1. Treat the systems separately. Create each model individually
    - ✓ Suffers from lack of fault examples
  - 2. Build one universal model by disregarding individuality
    - ✓ Model fit is not good

Practical requirements: Need to capture both commonality and individuality

System 1
(in New
Orleans)

⋮

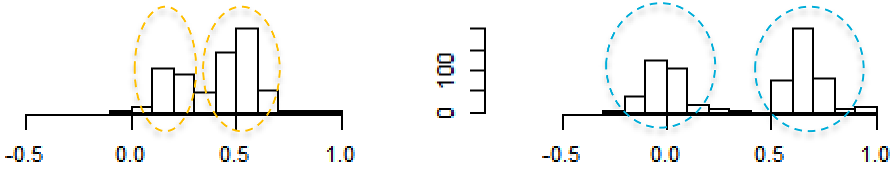
System *s*



⋮

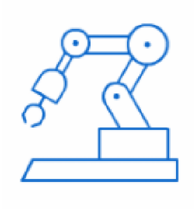
System *S*
(in New York)

- Capture both individuality and commonality
- Automatically capture multiple operational states
  - Real-world is not single-peaked (single-modal)



- Be robust to noise
- Be highly interpretable for diagnosis purposes

Formalizing the problem as multi-task density estimation for anomaly detection

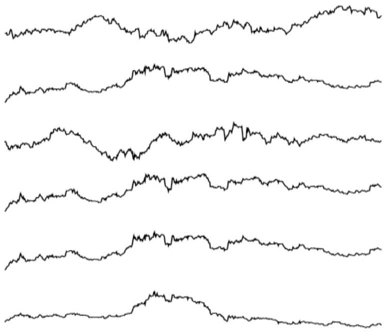
		Data		Prob. density	Anomaly score
System 1 (in New Orleans)		$\{\boldsymbol{x}^{1(n)} \in \mathbb{R}^M\}$	multi-task learning (MTL)	$p^1(\boldsymbol{x}^1 \mid \mathcal{D})$ ↑ all data	$a^1(\boldsymbol{x}^1)$
⋮					
System *s*		$\{\boldsymbol{x}^{s(n)} \in \mathbb{R}^M\}$		$p^s(\boldsymbol{x}^s \mid \mathcal{D})$	$a^s(\boldsymbol{x}^s)$ • overall • variable-wise
⋮					
System *S* (in New York)		$\{\boldsymbol{x}^{S(n)} \in \mathbb{R}^M\}$		$p^S(\boldsymbol{x}^S \mid \mathcal{D})$	$a^S(\boldsymbol{x}^S)$

Outline

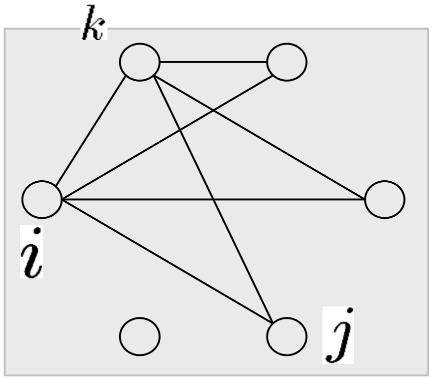
- Problem setting
- Modeling strategy
- Model inference approach
- Experimental results

Use Gaussian graphical model (GGM)-based anomaly detection approach as the basic building block

Multi-variate data



Sparse graphical model



$$\max_{\Lambda} \{ \ln \det \Lambda - \mathrm{tr}(\Sigma\Lambda) - \rho ||\Lambda||_1 \}$$

sample covariance

training data

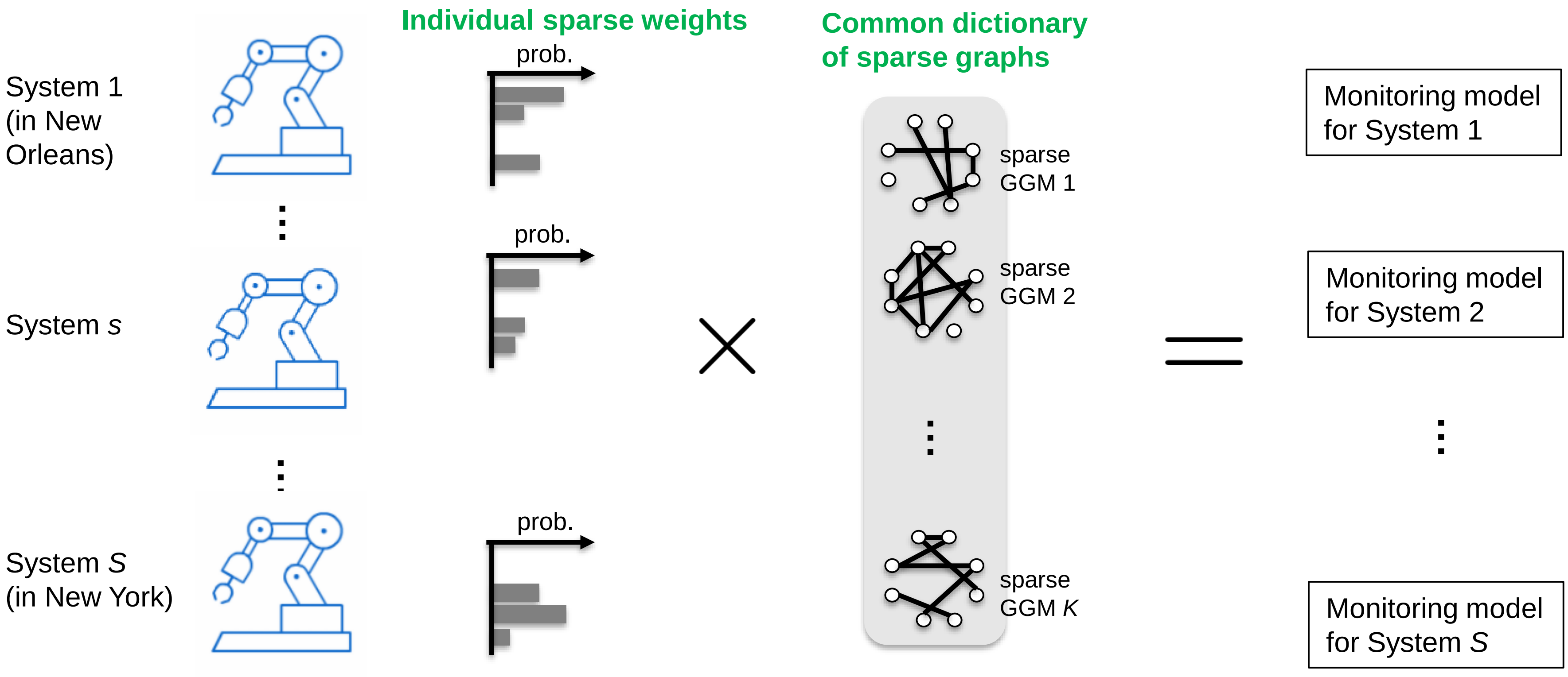
Anomaly score

$$a(\boldsymbol{x}) = \begin{cases} -\ln p(\boldsymbol{x} \mid \mathcal{D}) \\ \text{Overall score} \\ -\ln p(x_i \mid \boldsymbol{x}_{-i}, \mathcal{D}) \end{cases}$$

Variable-wise score

[Ide+ SDM09] [Ide+ ICDM16]

Basic modeling strategy: Combine common pattern dictionary with individual weights

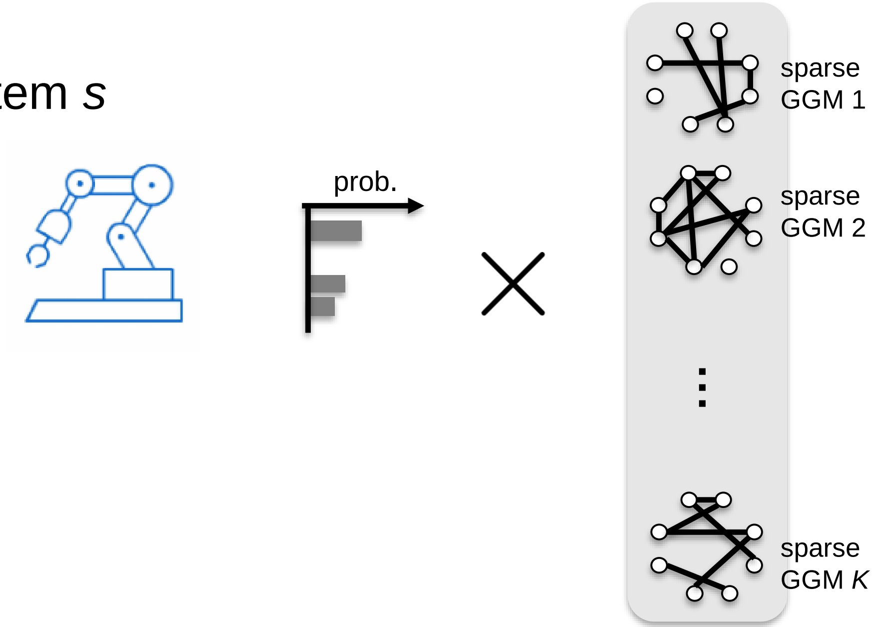


GGM=Gaussian Graphical Model

Basic modeling strategy: Resulting model will be a sparse mixture of sparse GGM

Monitoring model for System s

System *s*



GGM=Gaussian Graphical Model

Gaussian mixture

$$= \sum_{k=1}^{K} \pi_k^s \, \mathcal{N}(\boldsymbol{x}^s \mid \boldsymbol{\mu}^k, (\Lambda^k)^{-1})$$

<u>Sparse</u> mixture weights

(= automatic determination of the number of patterns)

Sparse Gaussian graphical model

Outline

- Problem setting
- Modeling strategy
- Model inference approach
- Experimental results

Employing a Bayesian model for multi-modal MTL

- Observation model (for the *s*-th task)
  - Gaussian mixture with task-dependent weight

$$\prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}^s \mid \boldsymbol{\mu}^k, (\Lambda^k)^{-1})^{z_k^s}$$

- Sparsity enforcing priors (non-conjugate)
  - Laplace prior for the precision matrix
  - Bernoulli prior for the mixture weights

$$p(\Lambda^k) = \left(\frac{\rho}{4}\right)^{M^2} \exp\left(-\frac{\rho}{2}\|\Lambda^k\|_1\right)$$

$$p(\boldsymbol{\pi}^s) = p_0^{\|\boldsymbol{\pi}^s\|_0}(1-p_0)^{G-\|\boldsymbol{\pi}^s\|_0}$$

- Conjugate prior on $\{\boldsymbol{\mu}^k\}$ and $\{\boldsymbol{z}^s\}$

Maximizing log likelihood using variational Bayes combined with point-estimation

- Log likelihood

$$L = \underbrace{\sum_{s=1}^{S}\sum_{n=1}^{N_s}\sum_{k=1}^{K} \ln \mathcal{N}(\boldsymbol{x}^{s(n)} \mid \boldsymbol{\mu}^k)^{z^{s(n)}}}_{\text{Likelihood by the obs. model}} + \underbrace{\sum_{k=1}^{K} \mathrm{Lap}(\Lambda^k \mid \rho) p(\boldsymbol{\mu}^k \mid \Lambda^k) + \sum_{s=1}^{S} z^{s(n)} \ln \pi_k^s + \sum_{s=1}^{S} \ln p(\boldsymbol{\pi}^s)}_{\text{Prior distributions}}$$

- Use VB for $\{\boldsymbol{\mu}^k\}, \{\boldsymbol{z}^{s(n)}\}$

- Use point-estimate for $\{\Lambda^k\}, \{\boldsymbol{\pi}^s\}$
  - Results in two convex optimization problems

Maximizing log likelihood using variational Bayes combined with point-estimation

- Update sample weights
- Update cluster weights
- Update precision matrices
- Update other parameters

Use new semi-closed form solution

$$\max_{\boldsymbol{\pi}^s} \left\{ \sum_{k=1}^{K} c_k^s \ln \pi_k^s - \tau \|\boldsymbol{\pi}^s\|_0 \right\} \quad \text{s.t.} \ \|\boldsymbol{\pi}^s\|_1 = 1.$$

The ratio of samples assigned to the *k*-th cluster

Solved by graphical lasso [Friedman 08]

$$\max_{\Lambda^k} \left\{ \ln |\Lambda^k| - \mathrm{Tr}(\Lambda^k \mathsf{Q}^k) - \frac{\rho}{N_k} \|\Lambda^k\|_1 \right\}$$

total # of samples assigned to the *k*-th cluster

Solving the L0-regularized optimization problem for mixture weights

- What is the problem of the conventional VB approach?
  - Simply differentiate w.r.t. π_k^s
  - Claims to get a sparse solution [Corduneanu+ 01]
  - But mathematically π_k^s cannot be zero due to logarithm

$$\max_{\boldsymbol{\pi}^s} \left\{ \sum_{k=1}^{K} c_k^s \ln \pi_k^s \right\} \quad \text{s.t.} \quad \|\boldsymbol{\pi}^s\|_1 = 1.$$

- We re-formalized the problem as a convex mixed-integer programming
  - A semi-closed form solution can be derived (→ see paper)

$$\max_{\boldsymbol{\pi}^s, \boldsymbol{y}^s} \sum_{k=1}^{K} \{c_k^s \ln \pi_k^s - \tau y_k^s\} \quad \text{s.t.} \quad \sum_{k=1}^{K} \pi_k^s = 1,$$
$$y_k^s \geq \pi_k^s - \epsilon, \quad y_k^s \in \{0,1\} \quad \text{for} \quad k = 1, \ldots, K,$$

Comparison with possible alternatives

		Interpretability	Noise reduction	Fleet-readiness	Multi-modality
Our work [Ide et al. ICDM 17]		Yes	Yes	Yes	Yes
(single) sparse GGM	[Ide et al. SDM 2009, Ide et al. ICDM 2016]	Yes	Yes	No	No
Gaussian mixtures	[Yamanishi et al., 2000; Zhang and Fung, 2013; Gao et al., 2016]	Limited	Limited	No	Yes
Multi-task sparse GGM	[Varoquaux et al., 2010; Honorio and Samaras, 2010; Chiquet et al., 2011; Danaher et al., 2014; Gao et al., 2016; Peterson et al., 2015].	Yes	Yes	Yes	No
Multi-task learning anomaly detection	[Bahadori et al., 2011; He et al., 2014; Xiao et al., 2015]	No	(depends)	Yes	No

Outline

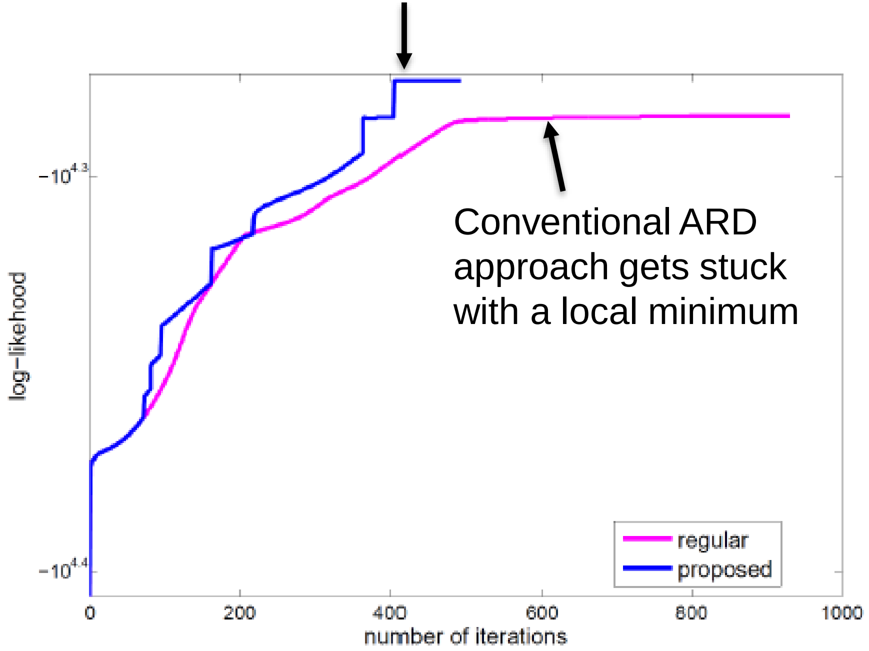
- Problem setting
- Modeling strategy
- Model inference approach
- Experimental results

Experiment (1): Learning sparse mixture weights

- Conventional ARD approach sometimes get stuck with local minima
  - ARD = automatic relevance determination
  - Often less sparser than the proposed convex L0 approach

- Typical result of log likelihood vs VB iteration count →

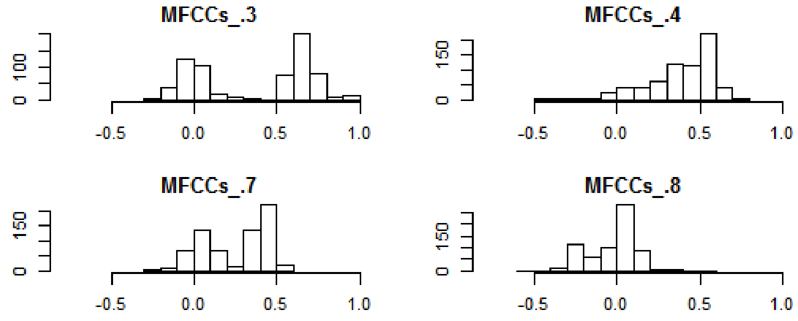
Proposed convex L0 approach gives better likelihood



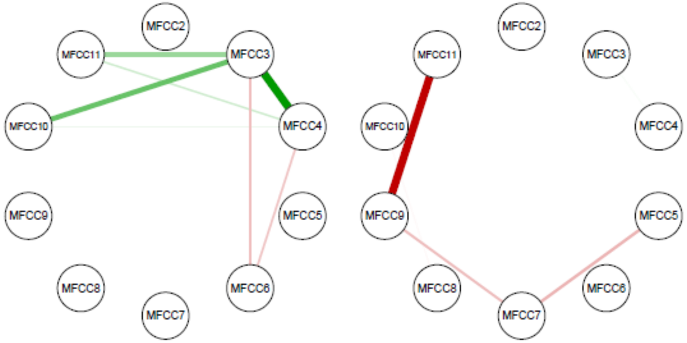
Experiment (2): Learning GGMs and detecting anomalies

- “Anuran Calls” (frog voice) data in UCI Archive
  - Multi-modal (multi-peaked)
  - Voice signal + attributes (species, etc.)
- Created 10-variate, 3-task dataset
  - Use the species of “Rhinellagranulosa” as the anomaly
- Results
  - Two non-empty GGMs are automatically detected starting from K=9
  - Clearly outperformed single-modal MTL alternative in anomaly detection
    - ✓ Group graphical lasso, fused graphical lasso

Example of variable-wise distribution



Automatically learned GGMs

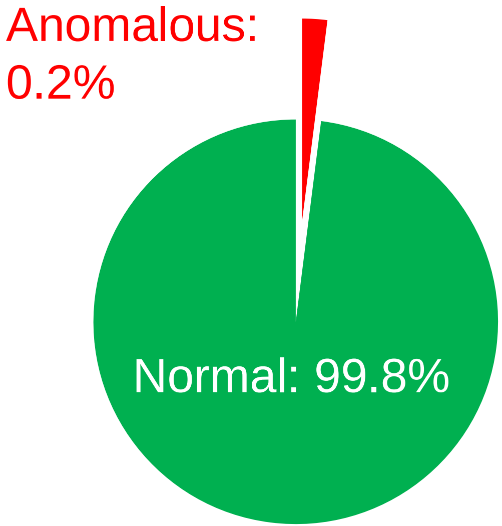


Conclusion

- Developed multi-task density estimation framework that can handle multi-modality
  - Featuring double sparsity: mixture weights, variable dependency
- Demonstrated the utility in the context of condition-based asset management

Thank you!

Integrated monitoring tool allows sharing rare anomaly data across different assets



- In condition-based monitoring, big data may not be really big
  - Anomalous samples account for less than 0.2% in a metal smelting process
- Coverage of anomalies and thus accuracy can be limited due to lack of data

Existing methods cannot handle multi-modality

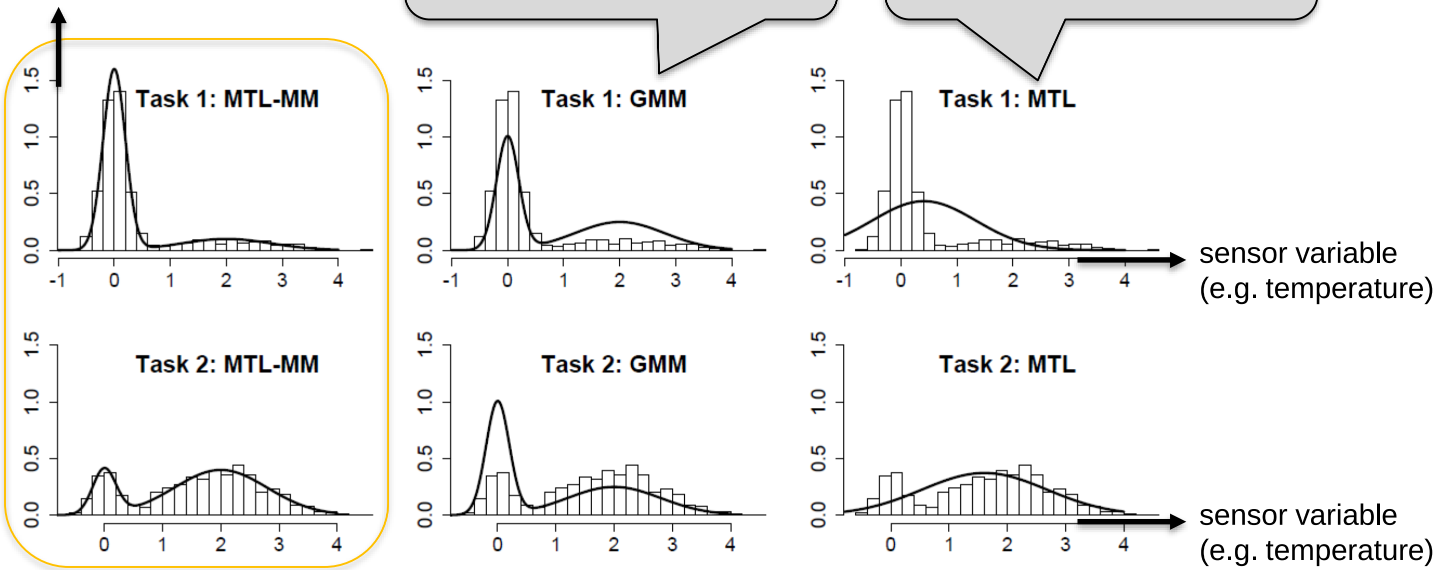
Can handle multi-modality but two systems must have the same model

Can treat different systems differently but cannot handle multi-modality

probability density

System (task) 1
(in New Orleans)

System (task) 2
(in New York)



Comparing the proposed multi-task multi-modal (MTL-MM) model with standard Gaussian mixture (GMM) and multi-task learning (MTL) models