Robust Statistical Methods Using Student-t Distributions

Kalpesh Patil

130040019

Department of Electrical Engineering

IIT Bombay

Overview

Gaussian Mixture Model

Weighted sum of Gaussian distributions

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

Limitations of GMM

- Fails to model large outliers since decays very fast and assigns less density on points away from centres
- Need of heavy-tailed distribution to model large outliers

Student-t Distribution

Combination of infinitely many Gaussians with "scaled covariances" with scaling parameter (u) generated from Gamma distribution

$$f(x;\mu,\Sigma,\nu) = \int_0^\infty \mathcal{N}\left(x_j;\mu,\frac{\Sigma}{u}\right)\Gamma\left(u;\frac{\nu}{2},\frac{\nu}{2}\right)du$$

The integral results in

$$f(x_j;\mu,\Sigma,\nu) = \frac{\Gamma(\frac{\nu+p}{2})|\Sigma|^{-1/2}}{(\pi\nu)^{\frac{p}{2}}\Gamma(\frac{\nu}{2})(1+\delta(x_j;\mu,\Sigma)/\nu))^{\frac{\nu+p}{2}}}$$

$$\delta(x_j;\mu,\Sigma) = (x_j-\mu)^T\Sigma^{-1}(x_j-\mu)$$

Student-t Mixture Models

Modeling a point in SMM [McLachlan, 2016]

$$X_j \mid (u_j, z_{ij} = 1) \sim \mathcal{N}\left(\mu_i, \frac{\Sigma_i}{u_j}\right)$$

$$U_j \mid (z_{ij} = 1) \sim \Gamma\left(\frac{v_i}{2}, \frac{v_i}{2}\right)$$

Parameters are optimized using EM algorithm

Kernel Space

- Mapping input data to higher (possibly infinite) dimensional subspace (kernel space)
- Computation of explicit mapping is expensive/impossible
- Kernel tricks are employed to carry out task by using only inner produces in kernel spaces

$$x \mapsto \Phi(x)$$
$$G_{i,j} = \langle x_i, x_j \rangle_{\mathcal{H}}$$

Prior Arts

- Kernel Support Vector Machines [Boser, 1992]
  introduce kernel tricks for finding optimal separating hyperplane in kernel space
- Kernel Principal Component Analysis (KPCA)[Schölkopf, 1997]
  proved a relation of eigenvalues and eigenvectors of covariance matrix with Gram kernel matrix
- Kernel Gaussian Mixture Models (KGMM) [Wang, 2003]
  incorporated kernel tricks for Gaussian Mixture Models in kernel space

KSMM model

Define parameters $\{a_{li}\}$ and $\{b_{li}\}$

$$\mu_l = \sum_{i=1}^{N} a_{li}^2 \phi(x_i)$$

$$\Sigma_l = \sum_{i=1}^{N} b_{li}^2 (\phi(x_i) - \mu_l) \otimes (\phi(x_i) - \mu_l).$$

$$f(\phi(x_i); \theta) = \sum_{l=1}^{g} \pi_l f(\phi(x_i); a_l, b_l, \nu_l)$$

Parameters(θ): $\{a_1, a_2 \cdots a_g, b_1, b_2 \cdots b_g, \nu_1, \cdots \nu_g, \pi_1, \pi_2 \cdots \pi_g\}$
complete data(y_c): $\{y_1, y_2 \cdots y_n, z_1, z_2 \cdots z_n, u_1, u_2 \cdots u_n\}$.

EM updates

EM algorithm is used to find iterative updates of the parameters. Let

$$\tau_{lj}^t = \frac{\pi_l^t f(y_j; \mu_l^t; \Sigma_l^t; \nu_l^t)}{f(y_j; \theta^t)} \qquad w_{lj}^t = \frac{\nu_l + p}{\nu_l + (x_j - \mu_l)^T \Sigma_l^{-1} (x_j - \mu_l)}$$

EM updates are as follows (derivation in the report):

$$\pi_l^{t+1} = \frac{1}{n} \sum_{j=1}^{n} \tau_{lj}^t \qquad a_{lj}^{t+1} = \sqrt{\frac{\tau_{lj}^t w_{lj}^t}{\sum_{j=1}^n \tau_{lj}^t w_{lj}^t}} \qquad b_{lj}^{t+1} = \sqrt{\frac{\tau_{lj}^t w_{lj}^t}{\sum_{j=1}^n \tau_{lj}^t}}$$

Eigenanalysis of Covariance Matrix

Let centered features $\tilde{\Phi}_l(x_j) = \Phi(x_j) - \mu_l$ and centered kernel matrix be $\tilde{\mathcal{K}}_l(i,j) = \langle b_{li}\tilde{\Phi}_l(x_i), b_{lj}\tilde{\Phi}_l(x_j)\rangle_{\mathcal{H}}$

Theorem

Eigenvalues of Σ_l are same as eigenvalues of $\tilde{\mathcal{K}}_l$ and there exists a relation between eigenfunction of Σ_l and $\tilde{\mathcal{K}}_l$ as follows

$$\Sigma_l v = \lambda v \iff \tilde{\mathcal{K}}_l \beta = \lambda \beta$$

$$v = \sum_{j=1}^{N} b_{lj}\beta_{lj}\tilde{\Phi}_l(x_j)$$

Computation of Probability

$$f(\phi(x_i);\mu_l,\Sigma_l,\nu_l)=\frac{\Gamma(\frac{\nu_l+p}{2})|\Sigma_l|^{-1/2}}{(\pi\nu_l)^{\frac{p}{2}}\Gamma(\frac{\nu_l}{2})(1+\frac{\delta_{li}}{\nu_l}))^{\frac{\nu_l+p}{2}}}$$

All components can be computed using kernel matrices only

- δ_{li}

$$\delta_{li}=\langle\tilde{\phi}_l(x_i),\sum_{m=1}^{n}\frac{v_{lm}\otimes v_{lm}}{\lambda}\tilde{\phi}_l(x_i)\rangle_{\mathcal{H}}=\sum_{m=1}^{n}\frac{\langle v_{lm},\tilde{\phi}_l(x_i)\rangle_{\mathcal{H}}^2}{\lambda_m}$$

$$\langle v_{lm},\tilde{\phi}_l(x_i)\rangle_{\mathcal{H}}=\sum_{j=1}^{n}\beta_{lmj}b_{lj}G(j,i)$$

- $|\Sigma_l|=\prod_{k=1}^{N}\lambda_{lk}$

KSMM Algorithm

Input: A set of points $\{x_n\}_{n=1}^{N}$, Gram matrix G such that $G(i,j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$. Number of clusters g, degree of freedom parameters $\{\nu_l\}$.

Initialization: Initialize τ_{li} s.t. $\sum_{l=1}^{g} \tau_{li} = 1 \; \forall \; i$ and w_{li} to 1

while *stopping criterion* $==$ *false* **do**

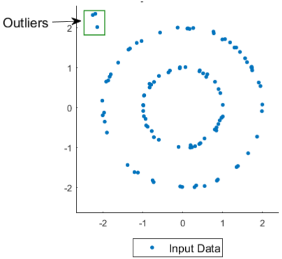
1. Compute $\{\pi_l\}$, $\{a_{li}\}$ and $\{b_{li}\}$ from the updates mentioned earlier using old values of $\{\tau_{li}\}$ and $\{w_{li}\}$.
2. Compute matrices $\{\tilde{\mathcal{K}}_l\}$ and their eigenvectors and eigenvalues
3. Compute $\{\delta_{li}\}$ using the the eigenvectors and eigenvalues of $\{\tilde{\mathcal{K}}_l\}$. to further calculate $Pr(\phi(x_i); \mu_l, \Sigma_l, \nu_l)$.
4. Update $\{\tau_{li}\}$ and $\{w_{li}\}$
5. Check stopping criterion for convergence i.e. either $t > t_{max}$ or $\sum_{i=1}^{n} \sum_{l=1}^{g} (\tau_{li}^{t} - \tau_{li}^{t-1})^2 < \epsilon$.

end

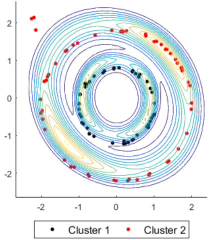
Output: A set of optimal parameters $\{\pi_l\}$, $\{a_{li}\}$ and $\{b_{li}\}$ representing the student-t mixture model in kernel space.

Results (synthetic dataset)

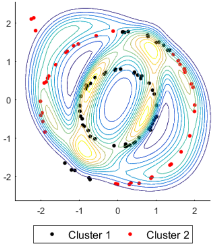
Unsupervised Clustering of 2D synthetic dataset



(a) Synthetic 2D data



(b) KSMM



(c) KGMM

Figure: KSMM vs KGMM on synthetic 2D dataset containing outliers

Results: Robust Outlier Detection (real datasets)

MNIST

- Inliers: images of digit 1
- Outliers: images of other digits
- Features: raw pixel values of images
- g = 4
- Kernel: RBF

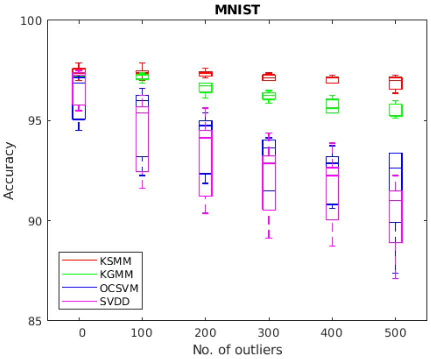


Figure: Accuracy with varying number of outliers for MNIST dataset

Results: Robust Outlier Detection (real datasets)

ORL faces

- Inliers: images of subject no. 1 to 30
- Outliers: images of remaining subjects
- Features: top 30 PCA components of images
- g = 4
- Kernel: RBF

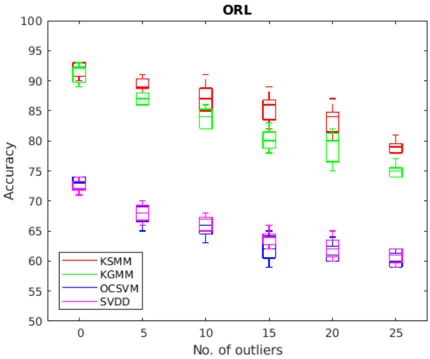


Figure: Accuracy with varying number of outliers for ORL dataset

Results: Robust Outlier Detection (real datasets)

Imagenet

- Inliers: images of class "flower"
- Outliers: images of other classes
- Features: 2 PCA components of features extracted from fc1 layer of VGG-16
- g = 4
- Kernel: RBF

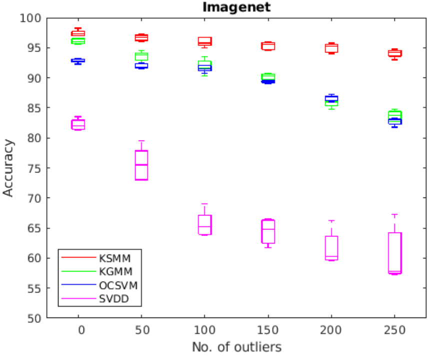


Figure: Accuracy with varying number of outliers for Imagenet dataset

Results: Robust Outlier Detection (real datasets)

Breast Cancer

- Inliers: benign tumors
- Outliers: malignant tumors
- Features: numerical features provided
- g = 2
- Kernel: RBF

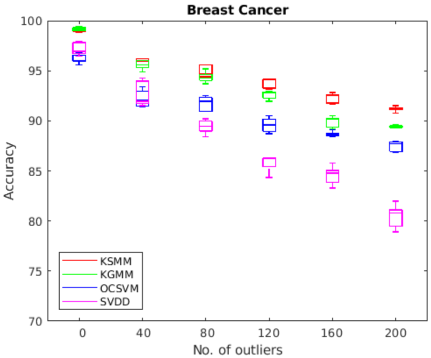


Figure: Accuracy with varying number of outliers for Breast Cancer dataset

Results: Robust Outlier Detection (real datasets)

Ionosphere

- Inliers: good radar data
- Outliers: bad radar data
- Features: numerical features provided
- $g = 2$
- Kernel: RBF

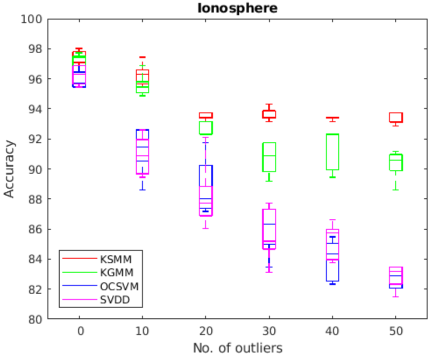


Figure: Accuracy with varying number of outliers for Ionosphere dataset

Sparse KSMM

- Sparsity is required to speed up computation
- constrain eigenfunctions of Σ_l to be sparse so that computing inner products with them is computationally inexpensive
- Naive way to do this is to keep only the largest k coefficients (by absolute value) of eigenfunction and make remaining coefficients zero after training is completed.
- Better way to do this is to encourage sparsity while training itself. We put a hard L1 norm constraint along with the sparsity constraint on eigenfunctions of Σ_l.

Sparse KSMM

Let v be an eigenfunction of Σ_l and w be any general vector in the kernel space. Let

$$v = \sum_{n=1}^{N} \alpha_n \tilde{\Phi}(x_n) \quad w = \sum_{n=1}^{N} \beta_n \tilde{\Phi}(x_n)$$

We intend to solve the following optimization problem.

$$\beta^* = \underset{\beta}{\operatorname{argmin}} ||\alpha - \beta||_2^2 \quad \text{s.t. } ||\beta||_0 \leq \kappa \quad \text{and} \quad ||\beta||_1 = \eta$$

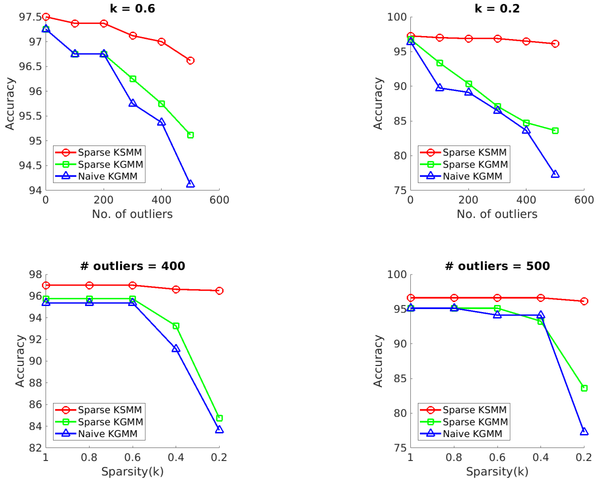
$$w^* = \sum_{n=1}^{N} \beta^* \tilde{\Phi}(x_n)$$

It is solved using the method mentioned in [Kyrillidis, 2013]

Results: MNIST

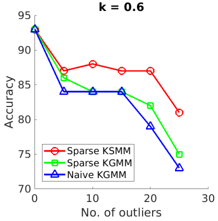
Varying no. of outliers for a given sparsity

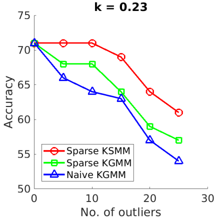
Varying sparsity for a given no. of outliers



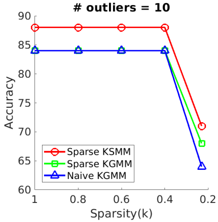
Results: ORL faces

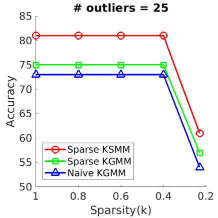
Varying no. of outliers for a given sparsity





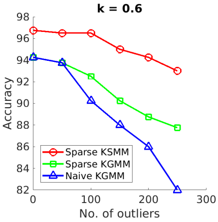
Varying sparsity for a given no. of outliers

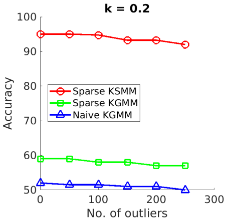




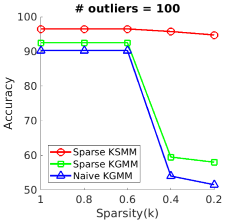
Results: Imagenet

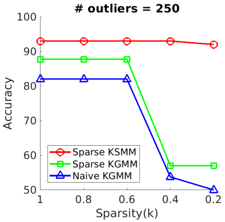
Varying no. of outliers for a given sparsity





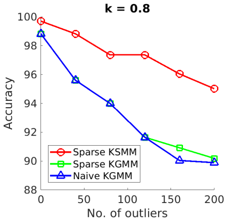
Varying sparsity for a given no. of outliers

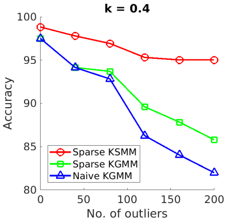




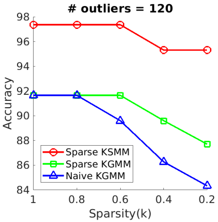
Results: Breast Cancer

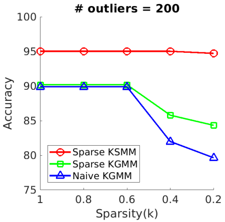
Varying no. of outliers for a given sparsity





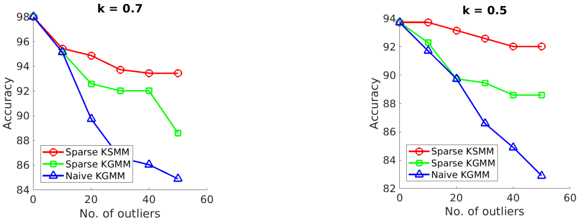
Varying sparsity for a given no. of outliers



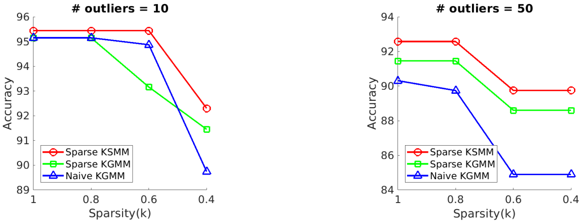


Results: Ionosphere

Varying no. of outliers for a given sparsity



Varying sparsity for a given no. of outliers



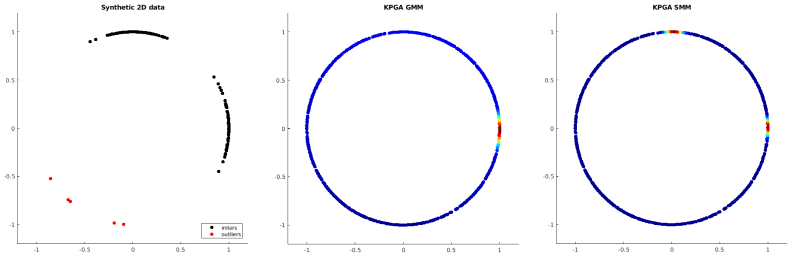
Kernel Principal Geodesic Analysis for SMM

- For certain kernels input points are mapped to a hyper-sphere in RKHS. This happens because for such kernels $\langle \Phi(x), \Phi(x) \rangle_{\mathcal{H}} = 1$ for kernels like radial basis function, exponential kernel
- KPGA [] projects (log map) points present on the Hilbert sphere to a hyperplane tangential to the sphere at the mean (μ) and analyze sample covariance of those projected points.

$$\mathrm{Log}_{\mu}(a) = \frac{x - \langle x, a \rangle_{\mathcal{H}} \mu}{||x - \langle x, a \rangle_{\mathcal{H}} \mu||_2} \arccos(\langle x, a \rangle_{\mathcal{H}})$$

- [] has proposed GMM model on a Hilbert sphere (KPGA-GMM). We expect KPGA-SMM to be more robust than KPGA-GMM.

KPGA-SMM Results



(a) Synthetic 2D data (b) KPGA-GMM (c) KPGA-SMM

Figure: KPGA-GMM vs KPGA-SMM on synthetic 2D dataset containing outliers

Robust SMM Prior

- FMRI: Data for 3D voxels collected at multiple time instants
- Time series are assumed to be generated from a mixture distribution (traditionally GMM is used)
- Prior learned by SMM is expected to be robust to large outliers present in data

Reconstruction

Reconstruction problem can be formulated as MAP estimation

$$Pr(X|Y) = \frac{Pr(Y|X)Pr(X)}{Pr(Y)}$$

$$\begin{aligned}\max_X \log(Pr(X|Y)) &\equiv \min_X - \log(Pr(X|Y)) \\ &\equiv \min_X \; - \log(Pr(Y|X) - \log(Pr(X))\end{aligned}$$

$$- \log(Pr(Y|X) = \text{Data Fidelity loss}$$

$$- \log(Pr(X)) = \alpha * \log \text{MRF misfit} + \beta * \log \text{mixture dist. misfit}$$

Prior is assumed be linear combination of MRF smoothness prior and mixture distribution prior in log space

Reconstruction Algorithm

- Step 1:
  Estimate parameters of mixture distributions (GMM or SMM) from data using EM algorithm. Note that this data may contain large outliers
- Step 2:
  Reconstruct fMRI data by Maximum Aposteriori Probability (MAP) estimation using gradient descent algorithm.

Experimental Setup

Creating data for prior estimation

- Image of Shepp Logan with sinusoidally varying intensity of each cluster
- Added Gaussian noise to simulate real life corruption of the image due to data collection process
- Added salt and pepper noise to simulate outliers
- Learned parameters of GMM and SMM from this corrupted data, which will be later used for reconstructions

Reconstruction

- Data Fidelity loss: Gaussian noise model for data corruption

$$|y_j - x_j|^2$$

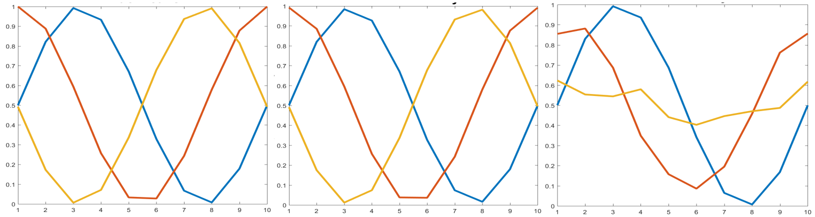
- Log MRF misfit: 4 neighborhood squared error

$$\sum_{p \in \{-1,1\}} \sum_{q \in \{-1,1\}} |x_{i,j} - x_{i+p,j+q}|^2$$

- Log Mixture Distribution misfit: Negative log probability under given mixture model

$$-\log \left(\sum_{l=1}^{3} \pi_l Pr(x_j|\theta_l) \right)$$

Prior Parameters Results



(a) True means (b) SMM prior means (c) GMM prior means

Figure: Estimated mean time series by GMM and SMM in presence of outliers

Prior Parameters Results

Quantitative Results

Quantity	Value	SMM	GMM
Avg. Euclidean distance between cluster means	$\frac{1}{3}\sum_{l=1}^{3} \|\|\mu_l^{est} - \mu_l^{orig}\|\|_2$	0.0074	0.3651
Avg. Frobenius distance between cluster covariance matrices	$\frac{1}{3}\sum_{l=1}^{3} \|\|\Sigma_l^{est} - \Sigma_l^{orig}\|\|_{frobenius}$	0.0209	0.3029
Avg. absolute difference between eigenvalues of covariance matrices	$\frac{1}{3}\sum_{l=1}^{3}\sum_{t=1}^{10} \|\lambda_{t,l}^{est} - \lambda_{t,l}^{orig}\|$	0.0631	0.6468

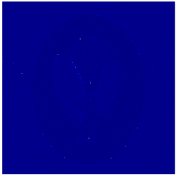
Table: GMM vs SMM quantitative results for parameter estimation

SMM Prior Reconstruction Results (frame #3)

Figure: Original

SMM-MRF

(a) reconstructed



(b) absolute diff.

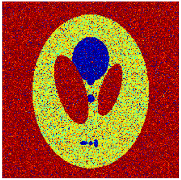
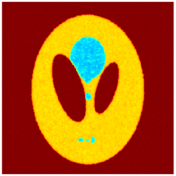
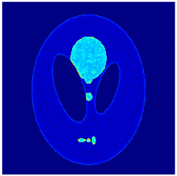


Figure: Corrupted

GMM-MRF



(a) reconstructed



(b) absolute diff.

SMM Prior Reconstruction Results (frame #8)

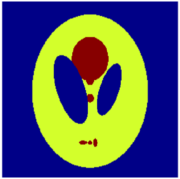
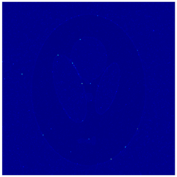


Figure: Original

SMM-MRF

(a) reconstructed



(b) absolute diff.

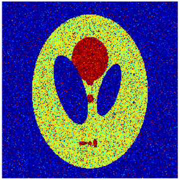
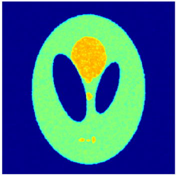
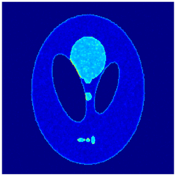


Figure: Corrupted

GMM-MRF



(a) reconstructed



(b) absolute diff.

Conclusion and Future Work

- Showed that robustness can be instilled in kernel spaces using our formulation of KSMM, which can also incorporate sparsity.
- Updates for degree of freedom parameter $(\{\nu_l\})$ of KSMM needs to be found using kernel tricks in future
- Show results of KPGA-SMM on real datasets in future
- Demonstrated how SMM is a better prior than GMM.
- Results of this robust prior needs to be shown on real FMRI.
- Data fidelity loss and smoothness prior loss can be enhanced using wavelet domain and edge preserving smoothness prior like Huber respectively.

References

- Boser, Bernhard E and Guyon, Isabelle M and Vapnik, Vladimir N (1992)
  A training algorithm for optimal margin classifiers
- Schölkopf, Bernhard and Smola, Alexander and Müller, Klaus-Robert (1997)
  Kernel principal component analysis
- Wang, Jingdong and Lee, Jianguo and Zhang, Changshui
  Kernel trick embedded Gaussian mixture model
- McLachlan, Geoffrey J and Ng, Shu-Kay and Bean, Richard
  Robust cluster analysis via mixture models
- Kyrillidis, Anastasios and Becker, Stephen and Cevher, Volkan and Koch, Christoph
  Sparse projections onto the simplex
- Awate, Suyash P and Yu, Yen-Yun and Whitaker, Ross T
  Kernel principal geodesic analysis

Thank You