EECS 545 Machine Learning - Sparse Kernel Density Estimates

Gowtham Bellala, Ganga Devadas, Bhavani Gopalakrishnan, Kumar Sricharan

University of Michigan, Ann Arbor

Introduction

- Density Estimation – Backbone of numerous Machine Learning problems.
- Standard Kernel Density Estimation (KDE) assigns equal weights for all the kernels.
- As the training data available becomes large, standard KDE becomes intractable for subsequent use.
- For *n* data samples, the order of complexity for computing KL divergence is $O(n^2)$
- We explicitly impose sparsity constraints on the objective function to induce sparse KDE.

Integrated Squared Error (ISE)

ISE is a measure of the quality of the estimate. The ISE between the true density and the estimated density is defined as:

$$\int (f(x) - \sum_{i=1}^{n} \alpha_i k_\sigma (x - x_i))^2 dx$$

The empirical estimate of the ISE can be reduced to:

$$\alpha^T Q \alpha - c^T \alpha$$

where α is the weight vector,

$$Q_{ij} = k_{\sqrt{2}\sigma}(x_i, x_j), \quad c_i = \frac{2}{m} \sum_{j=1}^{m} k_\sigma (y_j, x_i)$$

Girolami et. al. observed that the weights obtained by minimizing the ISE were sparse.

We extend this by imposing different penalties to increase the sparsity.

l_1 penalty

- Obvious choice to induce sparsity
- In the problem of KDE, the weights are subject to the constraint $\sum_{i=1}^{n} \alpha_i = 1$
- Therefore, l_1 penalty becomes redundant and cannot be used.

Weighted l_1 penalty

By increasing the contribution of the $c^T\alpha$ term we can increase the sparsity. The new objective function is:

$$(P_{l1}) \quad \alpha^T Q \alpha - \lambda c^T \alpha$$

As the above objective function remains convex, it can be solved using the Sequential Minimal Optimization (SMO) algorithm

Negative l_2 penalty

The objective function with a negative l_2 penalty imposed is :

$$(P_{l_2}) \min_{\alpha} \int (f(x) - \sum_{i=1}^{n} \alpha_i k_\sigma (x - x_i))^2 dx - \lambda \sum_{i=1}^{n} \alpha_i^2$$

which reduces to

$$\min_{\alpha} \alpha^T \hat{Q} \alpha - c^T \alpha \quad \text{where} \quad \hat{Q}_{ij} = k_{\sqrt{2}\sigma}(x_i, x_j) - \lambda \delta_{ij}$$

The above function is not convex for all values of λ

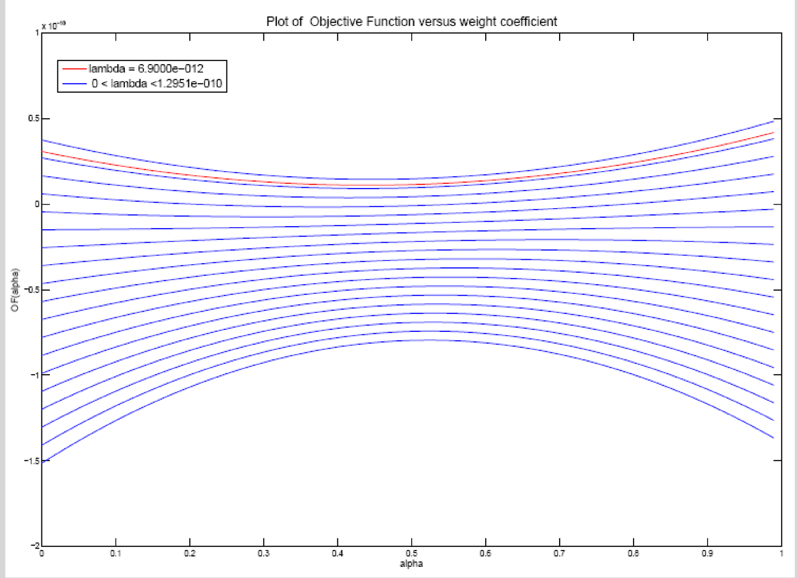
To solve this con-convex problem we use the following continuation search strategy

Algorithm 1: Continuation Search to solve (P_{l_2})
Step 1: Initialize $\lambda^{(0)} = 0$ and $\alpha_i^{(0)} = 1/n$, i=1,2,...n
Step 2: $\alpha_i^{(j+1)} = \mathrm{SMO}(\hat{Q}^{(j)}, c, \alpha_i^{(j)})$
Step 3: $\lambda^{(j+1)} = \lambda^{(j)} + \epsilon$, where ϵ is a small value Update $\hat{Q}^{(j+1)}$
Step 4: Compute KL Divergence between the standard KDE and the KDE with the weights $\alpha_i^{(j+1)}$
Step 5: Goto Step 2 if KL divergence less than threshold

Convexity of objective function

The figure illustrates how the objective function reduced to 2D changes from convex to concave as λ increases

We observed that for the final value of λ used by the above algorithm, the objective function reduced to 2D remained convex



l_0 penalty

Impose a penalty on the number of non-zero coefficients. The objective function is: $(P_{l_0}) \min_{\alpha} \alpha^T Q \alpha - c^T \alpha + \lambda \| \alpha_0 \|$

This is not convex. Wakin et. al. propose that the following function can be viewed as a relaxed version of the above function

$$(\hat{P}_{l_0}) \min_{\alpha} \alpha^T Q \alpha - c^T \alpha + \lambda w^T \alpha$$

which can be easily solved using the following iterative algorithm

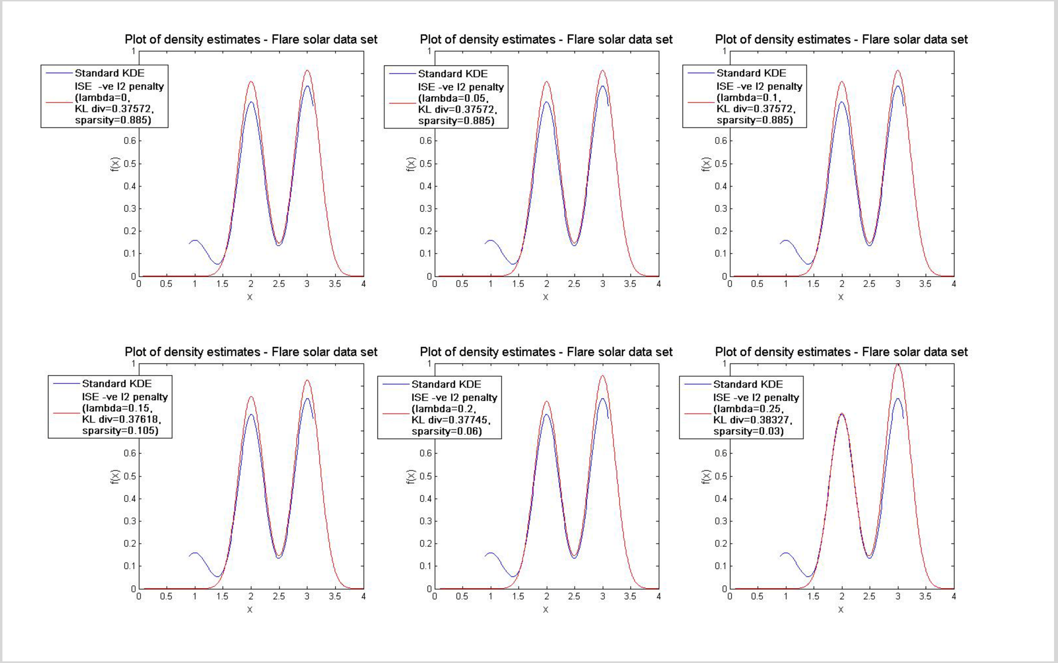
Algorithm 2 : Iterative algorithm to solve $(\hat{P}_{l_0})$
Step 1: Set iteration count l to zero and $w_i^{(0)} = 1$, i=1,2,...n
Step 2: Set $\alpha_i^{(l)} = \mathrm{argmin} \quad \alpha^T Q \alpha - c^T \alpha + \lambda w^{(l)T} \alpha$
Step 3: Update the weights for each i=1,2,..,n, $w_i^{(l+1)} = \frac{1}{\alpha_i^{(l)} + \epsilon}$
Step 4: Terminate after specified iterations l_{max}

A view from Kernel Feature Space

- $x_1, x_2, \ldots, x_n \in R^d$ are i.i.d samples from multivariate Gaussian *f(x;θ)* with unknown mean θ
- ML estimate is $\hat{\theta} = 1/n \sum_{i=1}^{n} x_i$
- Kernel density estimate can be interpreted as
  $\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \langle \phi(x), \phi(x_i) \rangle = \left\langle \phi(x), \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \right\rangle$
- $\hat{\theta} = 1/n \sum_{i=1}^{n} \phi(x_i)$ can be interpreted as ML estimate of θ in
  $\max_{\theta} - \|\phi(x) - \theta\|^2 = \min_{\theta} \|\phi(x) - \theta\|^2$
- Reformulated as $\hat{\alpha} = \arg\min_{\alpha} \left\| \phi(x) - \sum_{i=1}^{n} \alpha_i \phi(x_i) \right\|$

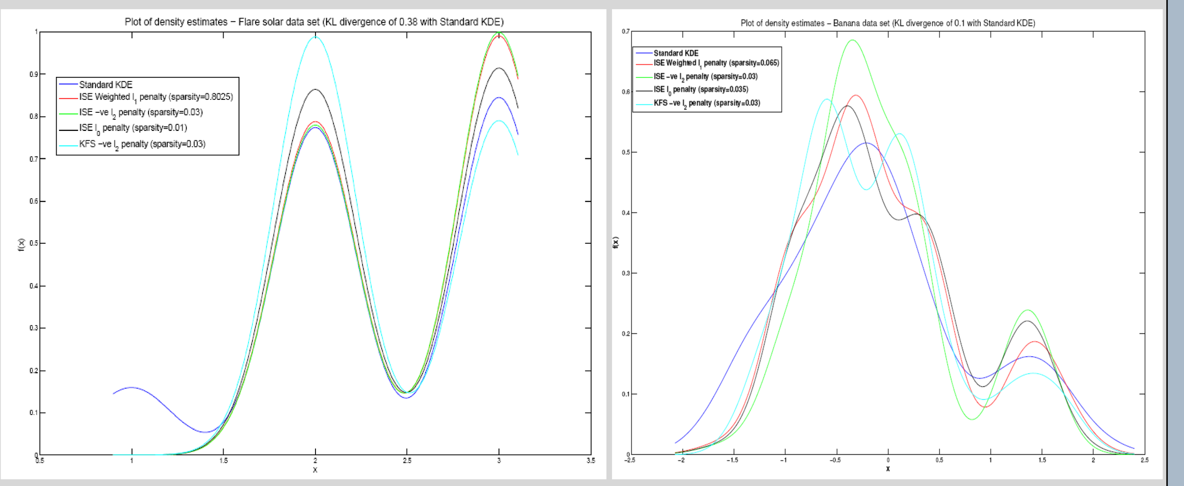
Results on Synthetic Data

The following figure compares the standard KDE and the sparse KDE obtained using negative l_2 penalty for different values of λ



Comparison of Methods

The following figure and table provide the sparsity induced by the different methods with and without penalty. The KL divergence for these values of sparsity are also specified



Method	Dataset	No Penalty		With Penalty	
		Sparsity	KL Div	Sparsity	KL Div
(P_{l_2})	Banana	14.5%	0.0518	2.5%	0.15
	Flare Solar	88.5%	0.3757	1%	0.7176
	Breast Cancer	84.5%	0.3083	23%	0.7271
(P_{Wl_1})	Banana	14.5%	0.0518	6%	0.1020
	Flare Solar	88.5%	0.3757	41%	0.7349
	Breast Cancer	84.5%	0.3083	8.5%	0.5929
(P_{l_0})	Banana	14.5%	0.0518	4%	0.0537
	Flare Solar	88.5%	0.3757	1%	0.5124
	Breast Cancer	84.5%	0.3083	2%	0.3146
KFS	Banana	79.5%	0.0147	4.5%	0.0364
	Flare Solar	95.5%	0.0068	3%	0.3811
	Breast Cancer	89.5%	0.1240	1.5%	1.6319

Flow Cytometry Data

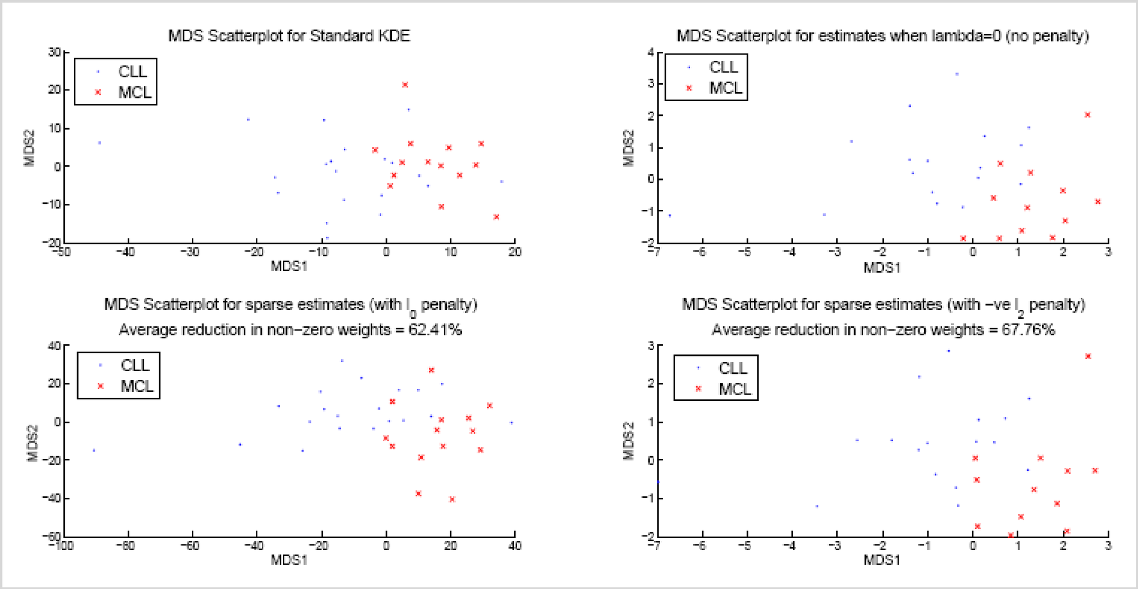
The sparsity induced by the different methods with similar quality of the estimate

Method	Dataset	No Penalty		With Penalty	
		Sparsity	KL Div	Sparsity	KL Div
(P_{l_2})	CLL1	31%	1	10.5%	1.985
	CLL2	24.5%	1.158	10.5%	1.9942
	MLL1	14.5%	1.431	4%	1.999
	MLL2	37%	0.8411	13.5%	1.6805
(P_{Wl_1})	CLL1	31%	1	19.5%	1.9293
	CLL2	24.5%	1.158	14.5%	1.9960
	MLL1	14.5%	1.431	8%	1.9979
	MLL2	37%	0.8411	23.5%	1.6659
(P_{l_0})	CLL1	30%	0.8902	13.5%	1.7712
	CLL2	30.5%	1.0486	9%	1.9962
	MLL1	14%	1.4478	4.5%	1.9838
	MLL2	45%	0.7659	24%	1.4259

Method	No Penalty		With Penalty	
	Average Sparsity	Average KL Div	Average Sparsity	Average KL Div
(P_{l_2})	25.14%	1.1543	8.607%	1.9181
(P_{Wl_1})	25.2%	1.115	14.7%	1.917
(P_{l_0})	17.1%	0.8797	8.47%	1.1429

Method	Increase in Sparsity(%)	Increase in KL divergence
(P_{l_2})	67.76	1.7654
(P_{Wl_1})	62.41	1.7546
(P_{l_0})	42.97	1.6218

Low dimensional representation



Conclusions

- The penalty methods allow for a user defined trade off between the sparsity and the quality of the estimates.
- Of the different methods proposed, the performances of the negative l_2 and l_0 penalties were better compared to the weighted l_1 penalty.
- Performance of the ISE and KFS objective functions with negative l_2 penalty were quite similar.
- The sparsity induced for 1D data is much more than the sparsity induced for the higher dimensional data.
- Extensions : Other choices of objective functions – KL divergence
  Other forms of penalties – l_p and entropy
- Acknowledgements : We would like to thank Professor Clayton Scott and Ami Wiesel for their valuable suggestions.

References :

[1] M.Girolami and C.He, "Probability density estimation from optimally condensed data samples," *IEEE Trans. on pattern analysis and machine learning*, 2003.
[2] J.Kim and C.Scott, "Robust Kernel Density Estimation," to appear in ICASSP,2008
[3] E.Candes, M.Wakin and S.Boyd, "Enhancing sparsity by reweighted l1 minimization," *Preprint*
[4] M.Carter, R.Raich and A.Hero, "Learning on manifolds for clustering and visualization," *Proc of Allerton Conference,* 2007.