Supplementary material for Maximum Mean Discrepancy for Class Ratio Estimation

1 Proofs and additional experiments

1.1 Detailed Proof of Lemma 1

With probability at least $1-\delta$,

$$\|\widehat{A}(n)\theta^* - \widehat{a}(n)\|^2 - \|\widehat{A}(n)\widehat{\theta}(n) - \widehat{a}(n)\|^2 \leq R^2 \left(\frac{c^2+2c+2}{n_u} + \sum_{y=0}^{c} \frac{2}{n_y}\right)\left(1+\sqrt{\log\frac{2}{\delta}}\right)^2$$

Proof. First note that $\|\widehat{A}(n)\theta^* - \widehat{a}(n)\|^2 - \|\widehat{A}(n)\widehat{\theta}(n) - \widehat{a}(n)\|^2 \leq \|\widehat{A}(n)\theta^* - \widehat{a}(n)\|^2$. Now, we upper-bound the RHS. Let $f(X_0,\ldots,X_c,X_u) \equiv \widehat{A}(n)\theta^* - \widehat{a}(n) = \sum_{y=0}^{c}\theta_y^*\widehat{\Phi}_y(n_y) - \widehat{\Phi}_U(n_u)$, where X_y, X_u denote independent samples of size n_y, n_u from $P_D(\mathbf{x}|y)$ and $P_U(\mathbf{x})$ respectively. We will proceed to prove this result in two steps:

1. Show that with probability $1-\delta$, $\|f\| \leq \mathbb{E}\|f\| + R\sqrt{\log\frac{2}{\delta}\left(\sum_{y=0}^{c}\frac{2}{n_y}+\frac{2}{n_u}\right)}$.
2. Show that $\mathbb{E}\|f\| \leq \left(R\sqrt{\frac{c^2+2c+2}{n_u}+\sum_{y=0}^{c}\frac{2}{n_y}}\right)$

Combining the results from the two steps gives us the final result. Next, we give proofs for the two steps.

Step 1: We would like to now first show that $\|f\|$ satisfies the bounded difference property. Take a point $\mathbf{x}_i$ from class $\tilde{y}$ and replace that point with $\mathbf{x}_i'$. Let the new class average of class $\tilde{y}$ be denoted as $\widehat{\Phi}'_{\tilde{y}}(n_{\tilde{y}})$. Note that the number of points in the class do not change. Let the new value of f be denoted as $f'_{\tilde{y}}$. Therefore,

$$\|f'_{\tilde{y}}\| = \left\|\sum_{\substack{y=0\\y\neq\tilde{y}}}^{c}\theta_y^*\widehat{\Phi}_y(n_y) + \theta_{\tilde{y}}^*\widehat{\Phi}'_{\tilde{y}}(n_{\tilde{y}}) - \widehat{\Phi}_U(n_u)\right\|$$

$$= \left\|\sum_{y=0}^{c}\theta_y^*\widehat{\Phi}_y(n_y) + \theta_{\tilde{y}}^*\widehat{\Phi}'_{\tilde{y}}(n_{\tilde{y}}) - \theta_{\tilde{y}}^*\widehat{\Phi}_{\tilde{y}}(n_{\tilde{y}}) - \widehat{\Phi}_U(n_u)\right\|$$

In the second step, we just added and subtracted $\theta_{\tilde{y}}^*\widehat{\Phi}_{\tilde{y}}(n_{\tilde{y}})$. Now using triangle inequality, we can see that,

$$\|f'_{\tilde{y}}\| \leq \|f\| + |\theta_{\tilde{y}}^*|\|\widehat{\Phi}'_{\tilde{y}}(n_{\tilde{y}}) - \widehat{\Phi}_{\tilde{y}}(n_{\tilde{y}})\|$$

Therefore,

$$\|f'_{\tilde{y}}\| - \|f\| \leq |\theta^*_{\tilde{y}}| \|\widehat{\Phi}'_{\tilde{y}}(n_{\tilde{y}}) - \widehat{\Phi}_{\tilde{y}}(n_{\tilde{y}})\| \leq \frac{1}{n_{\tilde{y}}} \|\Phi(\mathbf{x}'_i) - \Phi(\mathbf{x}_i)\| \leq \frac{2R}{n_{\tilde{y}}}$$

The second inequality is obtained by noting the facts that 1] $\theta^*_{\tilde{y}} \leq 1$, 2] the number of points in class $\tilde{y}$ continue to be $n_{\tilde{y}}$ and 3] only one point has changed in the class and therefore the others will cancel out in the difference. Again using triangle inequality but starting from $\|f\|$ instead of $\|f'_{\tilde{y}}\|$, we can show that $\|f\| - \|f'_{\tilde{y}}\| \leq \frac{2R}{n_{\tilde{y}}}$. Similar bounds can be achieved for any point in any other class. For points in the unlabeled set, again using triangle inequality, we can show that the difference is bounded by $\frac{2R}{n_u}$. This shows that $\|f\|$ satisfies the bounded difference property. Therefore, we can apply McDiarmid's Inequality,

$$\Pr(|\|f\| - \mathbb{E}\|f\|| \geq \epsilon) \leq 2 \exp\left(\frac{-2\epsilon^2}{\sum_{y=0}^{c} \frac{4R^2}{n_y} + \frac{4R^2}{n_u}} \right)$$

By setting RHS to δ, we can say that with at least probability $1 - \delta$,

$$\|f\| \leq \mathbb{E}\|f\| + R \sqrt{\log \frac{2}{\delta} \left(\sum_{y=0}^{c} \frac{2}{n_y} + \frac{2}{n_u} \right)}$$

Step 2: Next we present the proof of the final step in Lemma 1 of the paper. We would like to show that, $\mathbb{E}\|f(X_0, \ldots, X_c, X_u)\| = \|\widehat{A}\theta^* - \widehat{a}\| \leq \left(R \sqrt{\frac{c^2+2c+2}{n_u} + \sum_{y=0}^{c} \frac{2}{n_y}} \right)$ where the symbols are as defined in Lemma 1 of the paper.

$$\begin{aligned}
f^2 = & \sum_{y=0}^{c} \frac{\theta^{*2}_y}{n_y^2} \sum_{i,j=1}^{n_y} k(\mathbf{x}_i, \mathbf{x}_j) \\
& + \sum_{y=0}^{c} \sum_{\substack{y'=0 \\ y \neq y'}}^{c} \frac{\theta^*_y \theta^*_{y'}}{n_y n_{y'}} \sum_{i=1}^{n_y} \sum_{j=1}^{n_{y'}} k(\mathbf{x}_i, \mathbf{x}_j) \\
& + \frac{1}{n_u^2} \sum_{i,j=1}^{n_u} k(\mathbf{x}_i, \mathbf{x}_j) \\
& - 2 \sum_{y=0}^{c} \frac{\theta^*_y}{n_u n_y} \sum_{i=1}^{n_u} \sum_{j=1}^{n_y} k(\mathbf{x}_i, \mathbf{x}_j)
\end{aligned}$$

Now, we can take expectation $\mathbb{E}f^2$. When we have two summations indexed by the same set X_k, we need to ensure that we handle the case $x_i, x_j \in X_k, i \neq j$ and $x_i, x_j \in X_k, i = j$ separately. Thus,

we get,

$$\mathbb{E}f^2 = \sum_{y=0}^{c} \theta_y^{*2} \left(1 - \frac{1}{n_y}\right) \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y)}[k(\mathbf{x},\tilde{\mathbf{x}})] + \frac{\theta_y^{*2}}{n_y} \mathbb{E}_{P_D(\mathbf{x}|y)}[k(\mathbf{x},\mathbf{x})]$$
$$+ \sum_{y=0}^{c} \sum_{\substack{y'=0 \\ y \neq y'}}^{c} \theta_y^* \theta_{y'}^* \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})]$$
$$+ \left(1 - \frac{1}{n_u}\right) \mathbb{E}_{P_U(\mathbf{x}),P_U(\tilde{\mathbf{x}})}[k(\mathbf{x},\tilde{\mathbf{x}})] + \frac{1}{n_u} \mathbb{E}_{P_U(\mathbf{x})}[k(\mathbf{x},\mathbf{x})]$$
$$-2 \sum_{y=0}^{c} \theta_y^* \mathbb{E}_{P_U(\mathbf{x}),P_D(\tilde{\mathbf{x}}|y)} k(\mathbf{x},\tilde{\mathbf{x}}) \tag{1}$$

We know that,

$$P_U(\mathbf{x}) = \sum_{y=0}^{c} P_U(y) P_U(\mathbf{x}|y) = \sum_{y=0}^{c} \theta_y^* P_D(\mathbf{x}|y)$$

Therefore,

$$\mathbb{E}_{P_U(\mathbf{x}),P_U(\tilde{\mathbf{x}})}[k(\mathbf{x},\tilde{\mathbf{x}})] = \sum_{y=0}^{c} \sum_{y'=0}^{c} \theta_y^* \theta_{y'}^* \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})] \tag{2}$$

and

$$\mathbb{E}_{P_U(\mathbf{x}),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})] = \sum_{y=0}^{c} \theta_y^* \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})] \tag{3}$$

Substituting Equation (2) and (3) in Equation (1), we get

$$\mathbb{E}f^2 = \sum_{y=0}^{c} \theta_y^{*2} \left(1 - \frac{1}{n_y}\right) \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y)}[k(\mathbf{x},\tilde{\mathbf{x}})] + \frac{\theta_y^{*2}}{n_y} \mathbb{E}_{P_D(\mathbf{x}|y)}[k(\mathbf{x},\mathbf{x})]$$
$$+ \sum_{y=0}^{c} \sum_{\substack{y'=0 \\ y \neq y'}}^{c} \theta_y^* \theta_{y'}^* \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})]$$
$$+ \sum_{y=0}^{c} \sum_{y'=0}^{c} \theta_y^* \theta_{y'}^* \left(1 - \frac{1}{n_u}\right) \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})] + \frac{1}{n_u} \mathbb{E}_{P_U(\mathbf{x})}[k(\mathbf{x},\mathbf{x})]$$
$$-2 \sum_{y=0}^{c} \sum_{y'=0}^{c} \theta_y^* \theta_{y'}^* \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})]$$

Rearranging we get,

$$\begin{aligned}
\mathbb{E}f^2 = \sum_{y=0}^{c} \theta_y^{*2} \left(1 - \frac{1}{n_y} + 1 - \frac{1}{n_u} - 2\right) \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y)}[k(\mathbf{x},\tilde{\mathbf{x}})] \\
+ \sum_{y=0}^{c} \frac{\theta_y^{*2}}{n_y} \mathbb{E}_{P_D(\mathbf{x}|y)}[k(\mathbf{x},\mathbf{x})] \\
+ \sum_{y=0}^{c} \sum_{\substack{y'=0 \\ y\neq y'}}^{c} \theta_y^* \theta_{y'}^* \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})] \\
+ \sum_{y=0}^{c} \sum_{\substack{y'=0 \\ y\neq y'}}^{c} \theta_y^* \theta_{y'}^* \left(1 - \frac{1}{n_u}\right) \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})] + \frac{1}{n_u} \mathbb{E}_{P_U(\mathbf{x})}[k(\mathbf{x},\mathbf{x})] \\
-2 \sum_{y=0}^{c} \sum_{\substack{y'=0 \\ y\neq y'}}^{c} \theta_y^* \theta_{y'}^* \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})]
\end{aligned}$$

Performing cancellation in third, fourth and the sixth term, and then we get,

$$\begin{aligned}
\mathbb{E}f^2 = -\sum_{y=0}^{c} \theta_y^{*2} \left(\frac{1}{n_y} + \frac{1}{n_u}\right) \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y)}[k(\mathbf{x},\tilde{\mathbf{x}})] \\
+ \sum_{y=0}^{c} \frac{\theta_y^{*2}}{n_y} \mathbb{E}_{P_D(\mathbf{x}|y)}[k(\mathbf{x},\mathbf{x})] \\
+ \sum_{y=0}^{c} \sum_{\substack{y'=0 \\ y\neq y'}}^{c} \theta_y^* \theta_{y'}^* \left(-\frac{1}{n_u}\right) \mathbb{E}_{P_D(\mathbf{x}|y),P_D(\tilde{\mathbf{x}}|y')}[k(\mathbf{x},\tilde{\mathbf{x}})] + \frac{1}{n_u} \mathbb{E}_{P_U(\mathbf{x})}[k(\mathbf{x},\mathbf{x})]
\end{aligned}$$

Given that $R = \max_{\mathbf{x}\in\mathcal{X}} \|\Phi(\mathbf{x})\|$,

$$\begin{aligned}
\mathbb{E}f^2 &\leq \sum_{y=0}^{c} \left(\frac{R^2}{n_y} + \frac{R^2}{n_u}\right) + \sum_{y=0}^{c} \frac{R^2}{n_y} + \sum_{y=0}^{c} \sum_{\substack{y'=0 \\ y\neq y'}}^{c} \left(\frac{R^2}{n_u}\right) + \frac{R^2}{n_u} \\
&= R^2 \left(\frac{c^2+2c+2}{n_u} + \sum_{y=0}^{c} \frac{2}{n_y}\right)
\end{aligned}$$

Since f is a norm, $\mathbb{E}f \leq \sqrt{\mathbb{E}f^2} \leq R\sqrt{\frac{c^2+2c+2}{n_u} + \sum_{y=0}^{c} \frac{2}{n_y}}$.

□

1.2 Proof required for Lemma 2.2

Lemma 2.2 uses the following claim that we prove here:

Lemma 1. *mineig*$(\bar{A}^\top \bar{A})) - $ *mineig*$(\widehat{A}(n)^\top \widehat{A}(n)) \leq \|\bar{A}^\top \bar{A} - \widehat{A}(n)^\top \widehat{A}(n)\|_{\mathcal{F}}$

Proof. Let $\mathbf{e} = \text{argmin}_{\mathbf{y}:\|\mathbf{y}\|=1}\mathbf{y}^\top[\widehat{A}(n)^\top\widehat{A}(n)]\mathbf{y}$. That is, $\mathbf{e}$ is the minimum eigen vector of $\widehat{A}(n)^\top\widehat{A}(n)$.

$$\begin{aligned}
\text{mineig}(\bar{A}^\top\bar{A}) &- \text{mineig}(\widehat{A}(n)^\top\widehat{A}(n)) \\
&= \text{argmin}_{\mathbf{y}:\|\mathbf{y}\|=1}\mathbf{y}^\top[\bar{A}^\top\bar{A}]\mathbf{y} - \mathbf{e}^\top[\widehat{A}(n)^\top\widehat{A}(n)]\mathbf{e} \\
&\leq \mathbf{e}^\top[\bar{A}^\top\bar{A} - \widehat{A}(n)^\top\widehat{A}(n)]\mathbf{e} \\
&\leq \max_{\mathbf{y}:\|\mathbf{y}\|=1}\mathbf{y}^\top[\bar{A}^\top\bar{A} - \widehat{A}(n)^\top\widehat{A}(n)]\mathbf{y} \\
&= \max_{\mathbf{y}:\|\mathbf{y}\|=1} < \bar{A}^\top\bar{A} - \widehat{A}(n)^\top\widehat{A}(n), \mathbf{y}\mathbf{y}^T > \\
&\leq \|\bar{A}^\top\bar{A} - \widehat{A}(n)^\top\widehat{A}(n)\|_{\mathcal{F}} \max_{\mathbf{y}:\|\mathbf{y}\|=1}\|\mathbf{y}\mathbf{y}^T\|_{\mathcal{F}} \quad \text{(Cauchy-Schwartz)} \\
&\leq \|\bar{A}^\top\bar{A} - \widehat{A}(n)^\top\widehat{A}(n)\|_{\mathcal{F}}
\end{aligned}$$

□

1.3 Sub-Gradient expression needed in Section 3

In this section, we derive the sub-gradient expression for the second term of the objective function given in Section 3.2.

Let $f(\mathbf{w}) \equiv \text{maxeig}\left(\sum_{j=1}^{n_k} -w_j\widehat{A}_j^\top\widehat{A}_j\right)$. We can write f as,

$$\begin{aligned}
f(\mathbf{w}) &= \max_{\|\mathbf{x}\|=1}\mathbf{x}^\top\left(\sum_{j=1}^{n_k} -w_j\widehat{A}_j^\top\widehat{A}_j\right)\mathbf{x} \\
&= \max_{\mathbf{y}\in S}\mathbf{w}^\top\mathbf{y}
\end{aligned}$$

where $\mathbf{y} = \begin{bmatrix} -\mathbf{x}^\top\widehat{A}_1^\top\widehat{A}_1\mathbf{x} \\ \vdots \\ -\mathbf{x}^\top\widehat{A}_{n_k}^\top\widehat{A}_{n_k}\mathbf{x}\end{bmatrix}$ and $S = \{\begin{bmatrix} -\mathbf{x}^\top\widehat{A}_1^\top\widehat{A}_1\mathbf{x} \\ \vdots \\ -\mathbf{x}^\top\widehat{A}_{n_k}^\top\widehat{A}_{n_k}\mathbf{x}\end{bmatrix} : \|\mathbf{x}\| = 1\}$.

Therefore,

$$f(\mathbf{w}) = \max_{\mathbf{y}\in\text{dom}(f^*)}\mathbf{w}^\top\mathbf{y} - f^*(\mathbf{y}) \tag{4}$$

where $f^*(\mathbf{y}) = \begin{cases} 0, & \text{if } \mathbf{y}\in\text{dom}(f^*) \\ \infty, & \text{otherwise}\end{cases}$ and $\text{dom}(f^*) = S$.

With (4), we have that f, f^* are conjugates and hence we obtain that:

$$\nabla f(\mathbf{w}) = \begin{bmatrix} -\mathbf{x}^\top\widehat{A}_1^\top\widehat{A}_1\mathbf{x} \\ \vdots \\ -\mathbf{x}^\top\widehat{A}_{n_k}^\top\widehat{A}_{n_k}\mathbf{x}\end{bmatrix}$$

where $\mathbf{x}$ is any eigenvector corresponding to the maximum eigenvalue of $\left(\sum_{j=1}^{n_k} -w_j\widehat{A}_j^\top\widehat{A}_j\right)$.

1.4 Results for SVM-CR

We created a second baseline SVM-CR as a representative of a classifier that is trained to handle shifted class ratios. Here, we modified the training objective of a SVM classifier to minimize errors in class ratio estimation on various test dataset sampled as for the MMD-MKL method below. We trained using the structured learning framework of [1] but estimated class ratios as in SMO-MKL.

In Figure 1, we present results for SVM-CR. The results of this model are presented on the five UCI datasets. We can observe that SVM-CR is worse than SMO-MKL on all datasets except one.

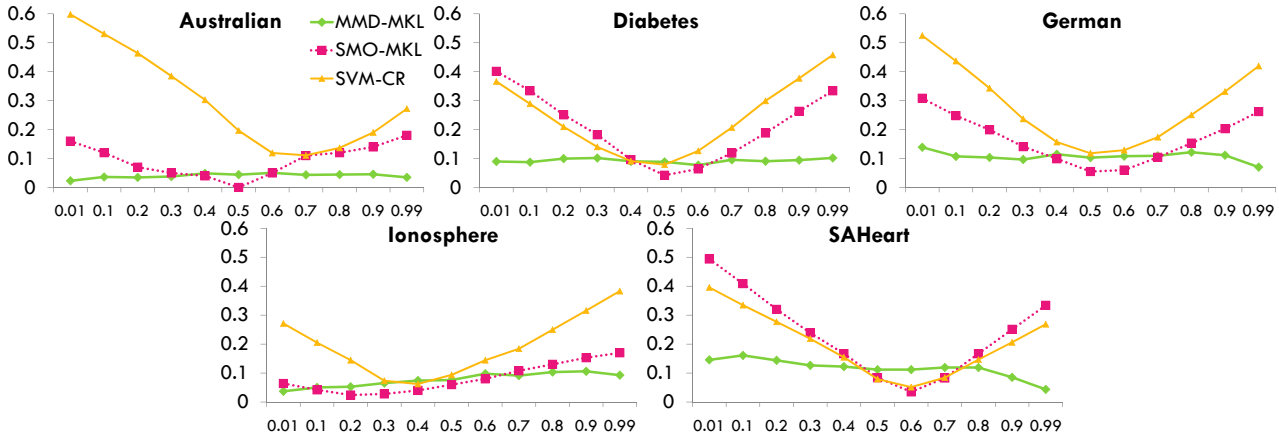

Figure 1: Class ratio estimation error ($|\theta_0^* - \widehat{\theta}_0|$) on Y-axis against varying true fractions (θ_0^*) for five binary datasets including the SVM-CR method. The methods compared are same in all datasets; the legend is present in only one of them to reduce clutter.

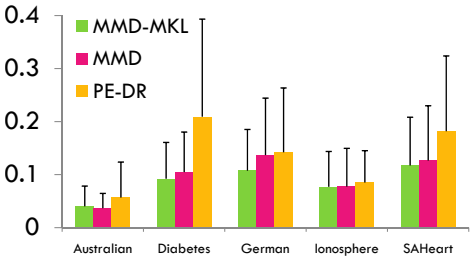

Figure 2: Comparing the methods PE-DR, MMD and MMD-MKL on the 5 UCI datasets and their corresponding variance in their estimation errors on these datasets

1.5 Error-bars in estimation for PE-DR, MMD and MMD-MKL

In Figure 2, we plot the average error of PE-DR, MMD and MMD-MKL on the 5 UCI datasets with the corresponding variance in the estimation error. PE-DR shows high variance in its estimates whereas MMD and MMD-MKL show lower variance in their estimates. MMD-MKL has overall better accuracy than both MMD and PE-DR.

1.6 Median Bandwidth vs Selection via Cross Validation

In our experiments, we selected the bandwidth for MMD via cross validation over the set of bandwidths given in Section 4 of the paper. In Figure 3, we present results where we compare accuracy on the 5 UCI datasets for MMD whose bandwidth is selected via cross validation from our set (MMD-CV) versus accuracies given by MMD whose bandwidth is simply selected as the median bandwidth (MMD-MED). We observe that bandwidth selected as per our method performs better than simply selecting the median bandwidth.

References

[1] Thorsten Joachims. A support vector method for multivariate performance measures. In *ICML*, pages 377–384, 2005.

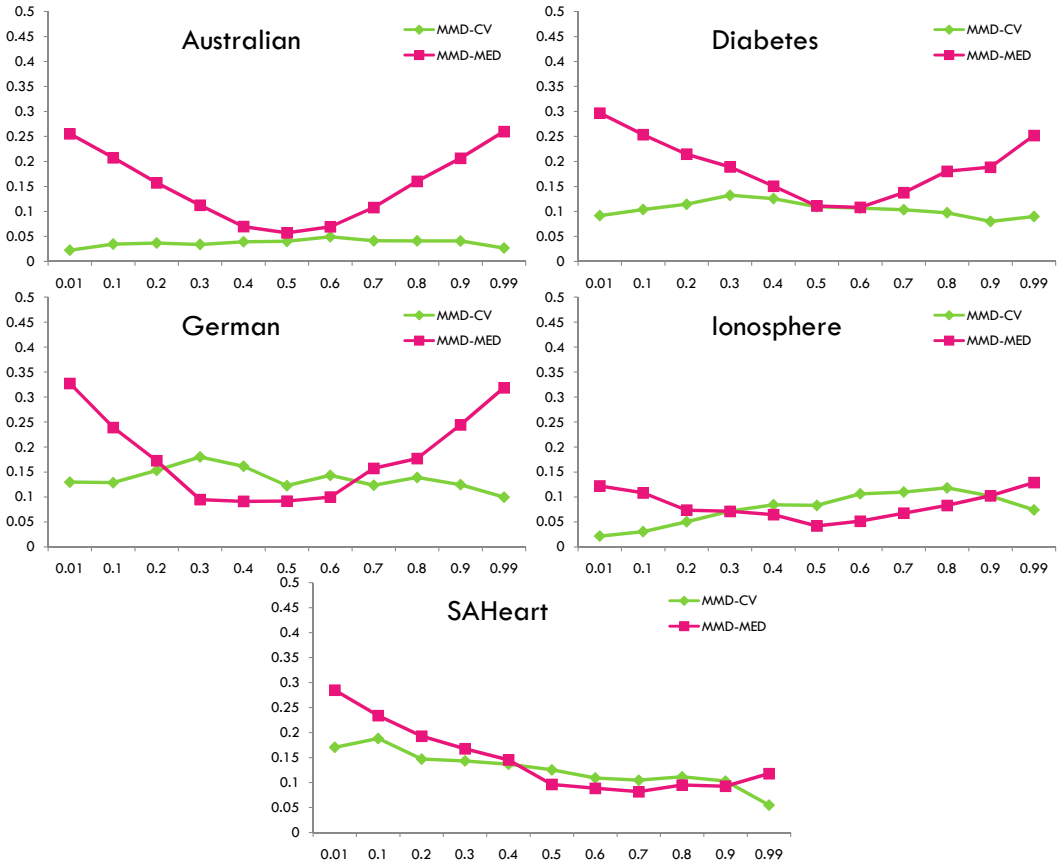

Figure 3: Absolute difference (on the Y axis) between estimated and true negative fraction for different estimation algorithms against different true negative fraction of the test set (on the X axis). The five graphs correspond to the five different datasets as shown in the graph's title.