Table 1. Mean training time in seconds and mean classification accuracy when the number of dimensions is chosen by cross-validation. Bold font highlights methods that were statistically the best or not statistically different from the best with 95% confidence. Symbol legend: "-" method did not converge in under three hours per training/test split, "nc" not computable.

	Orig	Train	LDG		PCA		FDA		LFDA		ITML		NCA	
	Dim	Ex	acc	time	acc	time	acc	time	acc	time	acc	time	acc	time
Diabetes	8	538	**71.3**	1	67.9	<1	**72.7**	<1	69.2	<1	69.7	40	**70.7**	135
Wine	13	125	**97.7**	<1	95.8	<1	**98.5**	<1	**98.5**	<1	**97.2**	54	**97.9**	6
Image Seg	19	1617	**96.5**	4	92.5	2	**96.2**	1	95.0	3	95.6	138	95.4	4641
German	20	700	**71.1**	<1	68.9	<1	**71.4**	<1	**71.9**	<1	69.5	15	**71.1**	1015
Ringnorm	20	3000	**86.9**	9	85.8	5	71.9	1	85.8	6	80.6	23	85.7	9119
Derm	33	256	**89.2**	<1	92.5	<1	91.5	<1	92.5	<1	**93.4**	139	**95.5**	381
Ion	34	246	**86.2**	<1	85.2	<1	83.2	<1	83.1	<1	84.3	10	**89.1**	222
Statlog	36	3000	**90.1**	10	**90.1**	5	86.6	3	88.3	6	88.0	232	-	-
USPS	256	3000	**93.5**	24	**92.0**	10	90.9	5	92.6	7	90.8	4886	-	-
Isolet	617	3000	87.9	94	73.1	15	**88.9**	18	**90.2**	21	-	-	-	-
MNIST	784	3000	**89.1**	55	**87.5**	13	76.2	10	32.7	34	-	-	-	-
Gisette	5000	3000	**95.5**	466	**96.7**	63	51.5	1983	49.8	2313	-	-	-	-
Mutants	5408	200	**90.0**	2908	64.8	2	52.0	1946	48.0	2003	-	-	-	-
Arcene	10K	70	**76.0**	578	61.3	15	53.7	39	nc	nc	-	-	-	-
Dexter	20K	210	**84.0**	4365	58.1	2	nc	nc	nc	nc	-	-	-	-

imizes the k-NN leave-one-out cross-validation error at dimensionality equal to the number of classes plus five. We have found that, in general, a few more dimensions than the number of classes present in the data is a good dimensionality at which to choose γ. In the case of ties, we select the largest value of γ. MRMR requires that we discretize the ITML features for feature selection, and we do so by thresholding at the mean, as recommended in the authors' code.

We perform experiments on fifteen datasets, and for each we average the accuracy over ten random 70/30 splits of the training and test data (up to a maximum of 3000 training samples). The datasets that we use can be found either at the UCI Machine Learning Repository or the Machine Learning Dataset Repository. The P53-Mutants dataset contained a large degree of class asymmetry. Therefore, we randomly sampled 143 of the *inactive* class samples and discarded the rest in order to make a 50/50 split between *inactive* and *active* class data (as opposed to the 1% vs 99% split in the original dataset).

Figure 1 and Table 1 show that for small datasets, LDG is comparable to other state-of-the-art methods. However, LDG provides a clear advantage on the datasets with the largest feature dimensionality.

NCA and ITML failed to converge in under three hours per training/test split on a standard 2.8 GHz PC for the datasets marked with "-" in Table 1, and results for these datasets are not plotted in Figure 1. Figure 1 also shows that ITML has difficulty with the Ringnorm dataset which has some features that are only noise.

LDG also outperforms FDA and LFDA on some of the datasets. FDA can provide dimensionality only up to one fewer than the number of classes, which limits its performance on the Ionoshpere and Ringnorm datasets. Furthermore, FDA and LFDA exhibit numerical instability in some of the datasets with large feature dimensionality due to the fact that the within-class covariance matrix is underdetermined. Thus, the generalized eigenvalue decomposition that these algorithms solve fails to find discriminative dimensions. LFDA returns complex eigenvalues for the Arcene and Dexter datasets, and FDA does the same on the Dexter dataset; thus, the LFDA and FDA results are not computable for these datasets.

In Table 1, we show the average classification accuracy when the dimensionality is chosen by leave-one-out cross-validation. We do this by increasing the dimensionality until the cross-validation accuracy decreases by adding another dimension. The run-time numbers measure the mean time it takes, in seconds, for the method to produce the dimensions shown in Figure 1 and to select the best dimensionality.

5. LDG for Transfer Learning

In this section, we apply LDG dimensionality reduction to transfer learning. In transfer learning, we wish