1 Experimental Settings

All experiments were run with the same hyperparameters for MAF, only varying the number of max training epochs for MAF between table 1 and table 2. For all experiments the MAF consisted of $blocks = 5$ MADE layers stacked, with $H = 512$ hidden neurons each, as this proved to be a good setting for both datasets used. A fairly small learning rate of $lr = 5 * 10^{-5}$ was chosen, with weight decay for the Adam optimizer being set to $wd = 1 * 10^{-6}$. The batch size was $b = 128$. Training data was shuffled to prevent batches consisting only of examples from one class. A patience of 5 was employed to stop training early if no improvement on the validation set in terms of log likelihood occurred within 5 epochs. If noise is *True*, Gaussian noise with $\mu = 0$ and $\sigma = 0.025$ was added directly to the raw data. Whereas, augment indicates that augmented samples, i.e. samples which were jittered with Gaussian noise ($\mu = 0, \sigma = 0.05$) and each variable being scaled with a factor drawn from a normal distribution ($\mu = 1, \sigma = 0.1$), were added to the training samples. Results are reported for the test set, which is the seventh subject of the Carrots data set, and the pre-made test set of UCIHAR respectively.

Dataset	Window	W_size	W_step	Augment	Noise	MVN LL	MAF LL	MVN ACC	MAF ACC
CARROTS	False	0	0	False	False	63.0739	68.2259	0.6006	0.5384
CARROTS	False	0	0	False	True	56.7594	59.6865	0.5932	0.6037
CARROTS	False	0	0	True	False	52.1412	68.1716	0.5683	0.5624
CARROTS	False	0	0	True	True	45.6814	59.1802	0.5103	0.6166
CARROTS	True	26	13	False	False	2986.2914	1555.2181	0.3299	0.3290
CARROTS	True	26	13	False	True	2016.7640	1416.1207	0.4398	0.3861
CARROTS	True	26	13	True	False	1881.0891	1720.2389	0.4312	0.4121
CARROTS	True	26	13	True	True	1566.8524	1513.4050	0.4909	0.5351
CARROTS	True	8	4	True	False	530.5136	695.9487	0.3759	0.5341
CARROTS	True	64	32	True	False	1375.9667	3319.6480	0.1215	0.4030
UCIHAR	True	128	64	False	False	717.9488	351.7945	0.9444	0.6291

Table 1: Log likelihoods and prediction accuracies of MAF and a baseline multivariate normal distribution trained per class with unconstrained covariance matrix. The Settings of the experiments are reported in the first 6 columns, the results in the remaining four. MAF was trained for a maximum of 2 epochs with the settings described above.

Dataset	Window	W_size	W_step	Augment	Noise	MVN LL	MAF LL	MVN ACC	MAF ACC
CARROTS	False	0	0	False	False	63.0739	47.1715	0.6006	0.3938
CARROTS	False	0	0	False	True	56.7594	61.4120	0.5932	0.6083
CARROTS	False	0	0	True	False	52.1412	53.1310	0.5683	0.3289
CARROTS	False	0	0	True	True	45.6814	60.8103	0.5103	0.6276
CARROTS	True	26	13	False	False	2986.2914	2860.8159	0.3299	0.4026
CARROTS	True	26	13	False	True	2016.7640	1951.0628	0.4398	0.3377
CARROTS	True	26	13	True	False	1881.0891	2965.4284	0.4312	0.4779
CARROTS	True	26	13	True	True	1566.8524	1972.0382	0.4909	0.3203
CARROTS	True	8	4	True	False	530.5136	1032.2761	0.3759	0.5125
CARROTS	True	64	32	True	False	1375.9667	5491.9022	0.1215	0.4051
UCIHAR	True	128	64	False	False	717.9488	749.2362	0.9444	0.9552

Table 2: Log likelihoods and prediction accuracies of MAF and a baseline multivariate normal distribution trained per class with unconstrained covariance matrix. The Settings of the experiments are reported in the first 6 columns, the results in the remaining four. MAF was trained for a maximum of 100 epochs with the settings described above.