

Nearest-neighbor estimation for ROC analysis under verification bias

Supplementary Material

S1. Simulation study: choice of distance and K

Results derived in Section 3 of the main paper make it quite clear that the use of the proposed estimators is not restricted to any special type of distance function, nor to any particular choice of the neighborhood size K . Nevertheless, to apply the proposed estimators, such choices have to be taken.

With the goal of exploring the impact of such choices, various simulation experiments, employing different distance measures and different choices of K , have been conducted within the scenario (i) of the main paper. Tables 1 to 6 show Monte Carlo means and standard deviations (in brackets) of the KNN estimators for the sensitivity and the specificity. As in the main paper, two different sample sizes are considered: $n = 50$ (Tables 1 to 3) and $n = 100$ (Tables 4 to 6), for three different values of α and of the cutpoint c . The number of replicates in each simulation experiment is 5000.

Results show that bias increases, in particular for sensitivity, on increasing the number of nearest neighbors, although such an effect tends to attenuate for increasing sample sizes. This can be explained the fact that the use of values of K which are large compared to the size of the verified sample fails to represent the local pattern of the measurement space, i.e., of the Y space. Overall, a choice of a small value of K (within the range 1 to 3) seems a good choice. In more general scenarios, however, the “optimal” value for K might depend upon the dimension of the feature space. If the number of features increases, it could be convenient to consider higher values for K , always taking into account the number of verified units in the sample.

Performance of the KNN estimators are quite comparable for different choices of the distance, with the Euclidean distance performing slightly better than competing distances. This might be due to the fact that, in our simulation setting, there is not a large disparity in the range of the data in each dimension. In practical situations, however, the selection among the distances here discussed or other distances is generally dictated by features of the data, by the sample size and the verification rate and by computational concerns, so that a general indication on an adequate choice is difficult to express.

Table 1: *Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity. Sample size = 50, $\alpha = 0.5$.*

Manhattan							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.787	(0.143)	0.593	(0.161)	0.377	(0.155)
	3NN	0.775	(0.136)	0.586	(0.155)	0.372	(0.149)
	5NN	0.768	(0.134)	0.579	(0.153)	0.367	(0.147)
	10NN	0.731	(0.137)	0.548	(0.152)	0.347	(0.143)
	20NN	0.594	(0.123)	0.434	(0.127)	0.272	(0.116)
Specificity	1NN	0.741	(0.076)	0.877	(0.057)	0.954	(0.036)
	3NN	0.740	(0.074)	0.876	(0.055)	0.954	(0.034)
	5NN	0.739	(0.074)	0.875	(0.054)	0.953	(0.034)
	10NN	0.735	(0.073)	0.872	(0.054)	0.951	(0.034)
	20NN	0.712	(0.073)	0.857	(0.057)	0.942	(0.037)
Euclidean							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.783	(0.143)	0.596	(0.166)	0.382	(0.153)
	3NN	0.775	(0.136)	0.587	(0.159)	0.376	(0.147)
	5NN	0.770	(0.133)	0.581	(0.155)	0.373	(0.148)
	10NN	0.733	(0.132)	0.545	(0.151)	0.353	(0.142)
	20NN	0.595	(0.121)	0.435	(0.126)	0.276	(0.114)
Specificity	1NN	0.742	(0.076)	0.877	(0.058)	0.954	(0.036)
	3NN	0.741	(0.074)	0.875	(0.057)	0.954	(0.035)
	5NN	0.738	(0.074)	0.874	(0.055)	0.952	(0.035)
	10NN	0.733	(0.073)	0.870	(0.055)	0.950	(0.035)
	20NN	0.710	(0.074)	0.855	(0.058)	0.942	(0.039)
Lagrange							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.783	(0.143)	0.593	(0.161)	0.377	(0.154)
	3NN	0.774	(0.136)	0.585	(0.155)	0.372	(0.149)
	5NN	0.764	(0.134)	0.576	(0.152)	0.366	(0.147)
	10NN	0.721	(0.131)	0.540	(0.146)	0.342	(0.140)
	20NN	0.587	(0.116)	0.429	(0.122)	0.269	(0.112)
Specificity	1NN	0.741	(0.075)	0.877	(0.057)	0.954	(0.036)
	3NN	0.740	(0.074)	0.876	(0.055)	0.953	(0.034)
	5NN	0.739	(0.073)	0.875	(0.054)	0.953	(0.034)
	10NN	0.732	(0.072)	0.870	(0.054)	0.950	(0.034)
	20NN	0.709	(0.073)	0.855	(0.057)	0.943	(0.037)
Mahalanobis							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.7777	(0.149)	0.5884	(0.168)	0.3784	(0.157)
	3NN	0.7578	(0.142)	0.5710	(0.160)	0.3662	(0.149)
	5NN	0.7383	(0.139)	0.5537	(0.156)	0.3541	(0.145)
	10NN	0.6814	(0.131)	0.5062	(0.145)	0.3225	(0.134)
	20NN	0.5707	(0.108)	0.4145	(0.115)	0.2618	(0.106)
Specificity	1NN	0.7392	(0.078)	0.8751	(0.059)	0.9529	(0.037)
	3NN	0.7373	(0.077)	0.8737	(0.057)	0.9520	(0.036)
	5NN	0.7352	(0.077)	0.8718	(0.056)	0.9507	(0.036)
	10NN	0.7272	(0.077)	0.8660	(0.057)	0.9475	(0.036)
	20NN	0.7057	(0.075)	0.8521	(0.059)	0.9408	(0.040)

Table 2: *Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity. Sample size = 50, $\alpha = 1$.*

Manhattan							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.948	(0.079)	0.871	(0.115)	0.742	(0.145)
	3NN	0.942	(0.077)	0.863	(0.111)	0.734	(0.140)
	5NN	0.933	(0.081)	0.853	(0.114)	0.724	(0.141)
	10NN	0.883	(0.108)	0.802	(0.132)	0.680	(0.151)
	20NN	0.673	(0.119)	0.598	(0.127)	0.502	(0.131)
Specificity	1NN	0.746	(0.076)	0.856	(0.062)	0.932	(0.045)
	3NN	0.745	(0.075)	0.855	(0.061)	0.931	(0.044)
	5NN	0.745	(0.075)	0.854	(0.060)	0.930	(0.044)
	10NN	0.739	(0.075)	0.849	(0.060)	0.926	(0.044)
	20NN	0.706	(0.075)	0.825	(0.063)	0.911	(0.048)
Euclidean							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.950	(0.079)	0.872	(0.117)	0.744	(0.142)
	3NN	0.944	(0.077)	0.863	(0.117)	0.735	(0.142)
	5NN	0.932	(0.084)	0.853	(0.117)	0.722	(0.141)
	10NN	0.876	(0.112)	0.798	(0.135)	0.674	(0.152)
	20NN	0.670	(0.120)	0.597	(0.128)	0.499	(0.132)
Specificity	1NN	0.745	(0.075)	0.856	(0.062)	0.931	(0.044)
	3NN	0.744	(0.074)	0.855	(0.060)	0.930	(0.043)
	5NN	0.744	(0.073)	0.852	(0.060)	0.929	(0.042)
	10NN	0.734	(0.073)	0.847	(0.059)	0.926	(0.042)
	20NN	0.704	(0.074)	0.822	(0.063)	0.910	(0.048)
Lagrange							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.949	(0.079)	0.871	(0.114)	0.741	(0.144)
	3NN	0.941	(0.079)	0.861	(0.113)	0.732	(0.141)
	5NN	0.927	(0.086)	0.847	(0.117)	0.720	(0.143)
	10NN	0.865	(0.108)	0.785	(0.132)	0.665	(0.151)
	20NN	0.666	(0.114)	0.592	(0.122)	0.496	(0.127)
Specificity	1NN	0.746	(0.076)	0.856	(0.062)	0.931	(0.045)
	3NN	0.745	(0.075)	0.855	(0.060)	0.931	(0.044)
	5NN	0.744	(0.075)	0.854	(0.060)	0.929	(0.044)
	10NN	0.737	(0.075)	0.848	(0.061)	0.925	(0.045)
	20NN	0.705	(0.075)	0.824	(0.064)	0.910	(0.049)
Mahalanobis							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.934	(0.091)	0.856	(0.124)	0.725	(0.149)
	3NN	0.908	(0.094)	0.828	(0.124)	0.699	(0.145)
	5NN	0.877	(0.100)	0.796	(0.126)	0.671	(0.144)
	10NN	0.790	(0.109)	0.710	(0.127)	0.596	(0.141)
	20NN	0.637	(0.101)	0.563	(0.108)	0.469	(0.112)
Specificity	1NN	0.745	(0.076)	0.853	(0.062)	0.930	(0.045)
	3NN	0.742	(0.075)	0.851	(0.061)	0.928	(0.044)
	5NN	0.738	(0.074)	0.848	(0.061)	0.926	(0.044)
	10NN	0.729	(0.073)	0.839	(0.062)	0.920	(0.045)
	20NN	0.698	(0.075)	0.817	(0.065)	0.907	(0.049)

Table 3: *Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity. Sample size = 50, $\alpha = 1.5$.*

Manhattan							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.990	(0.036)	0.968	(0.062)	0.916	(0.094)
	3NN	0.985	(0.041)	0.962	(0.064)	0.909	(0.094)
	5NN	0.978	(0.052)	0.953	(0.073)	0.900	(0.099)
	10NN	0.923	(0.098)	0.895	(0.113)	0.842	(0.130)
	20NN	0.679	(0.122)	0.645	(0.128)	0.599	(0.133)
Specificity	1NN	0.730	(0.076)	0.821	(0.066)	0.897	(0.054)
	3NN	0.730	(0.075)	0.821	(0.065)	0.896	(0.052)
	5NN	0.729	(0.075)	0.820	(0.065)	0.895	(0.052)
	10NN	0.724	(0.075)	0.815	(0.065)	0.890	(0.052)
	20NN	0.684	(0.075)	0.783	(0.068)	0.868	(0.057)
Euclidean							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.990	(0.033)	0.970	(0.060)	0.918	(0.092)
	3NN	0.986	(0.038)	0.965	(0.061)	0.912	(0.091)
	5NN	0.979	(0.049)	0.954	(0.071)	0.902	(0.096)
	10NN	0.924	(0.097)	0.895	(0.111)	0.844	(0.128)
	20NN	0.678	(0.120)	0.643	(0.125)	0.598	(0.128)
Specificity	1NN	0.731	(0.077)	0.824	(0.067)	0.898	(0.054)
	3NN	0.731	(0.076)	0.824	(0.065)	0.897	(0.053)
	5NN	0.733	(0.073)	0.823	(0.063)	0.897	(0.051)
	10NN	0.728	(0.073)	0.818	(0.063)	0.893	(0.051)
	20NN	0.687	(0.073)	0.786	(0.067)	0.870	(0.057)
Lagrange							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.990	(0.036)	0.968	(0.062)	0.916	(0.094)
	3NN	0.984	(0.045)	0.961	(0.067)	0.908	(0.096)
	5NN	0.974	(0.058)	0.949	(0.079)	0.895	(0.103)
	10NN	0.911	(0.100)	0.881	(0.116)	0.830	(0.134)
	20NN	0.676	(0.118)	0.642	(0.125)	0.596	(0.129)
Specificity	1NN	0.730	(0.076)	0.821	(0.066)	0.896	(0.053)
	3NN	0.730	(0.075)	0.821	(0.065)	0.896	(0.052)
	5NN	0.729	(0.075)	0.820	(0.065)	0.895	(0.052)
	10NN	0.723	(0.075)	0.814	(0.066)	0.890	(0.053)
	20NN	0.684	(0.076)	0.783	(0.066)	0.868	(0.058)
Mahalanobis							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.976	(0.055)	0.950	(0.078)	0.899	(0.103)
	3NN	0.949	(0.066)	0.918	(0.084)	0.865	(0.107)
	5NN	0.913	(0.079)	0.878	(0.095)	0.826	(0.114)
	10NN	0.812	(0.102)	0.775	(0.111)	0.723	(0.122)
	20NN	0.643	(0.097)	0.606	(0.101)	0.561	(0.104)
Specificity	1NN	0.733	(0.075)	0.823	(0.066)	0.898	(0.053)
	3NN	0.732	(0.074)	0.821	(0.064)	0.895	(0.051)
	5NN	0.730	(0.073)	0.818	(0.064)	0.893	(0.052)
	10NN	0.717	(0.073)	0.808	(0.065)	0.884	(0.053)
	20NN	0.681	(0.075)	0.780	(0.069)	0.866	(0.059)

Table 4: *Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity. Sample size = 100, $\alpha = 0.5$.*

Manhattan							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.784	(0.100)	0.594	(0.113)	0.380	(0.106)
	3NN	0.780	(0.095)	0.591	(0.109)	0.377	(0.102)
	5NN	0.777	(0.094)	0.588	(0.108)	0.375	(0.101)
	10NN	0.767	(0.093)	0.580	(0.107)	0.369	(0.100)
	20NN	0.733	(0.095)	0.550	(0.106)	0.349	(0.097)
Specificity	1NN	0.742	(0.054)	0.878	(0.040)	0.954	(0.026)
	3NN	0.742	(0.053)	0.877	(0.039)	0.954	(0.024)
	5NN	0.741	(0.052)	0.877	(0.039)	0.954	(0.024)
	10NN	0.741	(0.052)	0.876	(0.039)	0.953	(0.024)
	20NN	0.736	(0.052)	0.873	(0.039)	0.951	(0.024)
Euclidean							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.783	(0.101)	0.594	(0.115)	0.378	(0.106)
	3NN	0.780	(0.096)	0.590	(0.110)	0.376	(0.104)
	5NN	0.774	(0.094)	0.584	(0.109)	0.373	(0.102)
	10NN	0.765	(0.093)	0.576	(0.107)	0.368	(0.101)
	20NN	0.730	(0.094)	0.546	(0.106)	0.348	(0.098)
Specificity	1NN	0.742	(0.054)	0.878	(0.040)	0.954	(0.026)
	3NN	0.741	(0.053)	0.877	(0.039)	0.954	(0.025)
	5NN	0.743	(0.052)	0.877	(0.039)	0.953	(0.024)
	10NN	0.742	(0.052)	0.876	(0.037)	0.953	(0.024)
	20NN	0.737	(0.052)	0.873	(0.037)	0.951	(0.024)
Lagrange							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.784	(0.101)	0.594	(0.113)	0.379	(0.105)
	3NN	0.780	(0.096)	0.591	(0.109)	0.377	(0.102)
	5NN	0.776	(0.094)	0.588	(0.108)	0.375	(0.101)
	10NN	0.764	(0.093)	0.577	(0.106)	0.368	(0.100)
	20NN	0.724	(0.092)	0.543	(0.103)	0.345	(0.096)
Specificity	1NN	0.743	(0.054)	0.878	(0.040)	0.954	(0.026)
	3NN	0.742	(0.053)	0.877	(0.039)	0.954	(0.024)
	5NN	0.741	(0.052)	0.877	(0.039)	0.954	(0.024)
	10NN	0.740	(0.052)	0.876	(0.039)	0.953	(0.024)
	20NN	0.734	(0.051)	0.872	(0.037)	0.951	(0.024)
Mahalanobis							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.776	(0.104)	0.586	(0.116)	0.376	(0.107)
	3NN	0.767	(0.100)	0.577	(0.111)	0.369	(0.104)
	5NN	0.758	(0.099)	0.570	(0.110)	0.364	(0.102)
	10NN	0.735	(0.097)	0.550	(0.108)	0.350	(0.099)
	20NN	0.679	(0.092)	0.503	(0.101)	0.318	(0.092)
Specificity	1NN	0.743	(0.054)	0.877	(0.040)	0.954	(0.026)
	3NN	0.742	(0.054)	0.876	(0.039)	0.953	(0.024)
	5NN	0.741	(0.054)	0.876	(0.039)	0.953	(0.024)
	10NN	0.739	(0.054)	0.874	(0.039)	0.951	(0.024)
	20NN	0.731	(0.053)	0.868	(0.039)	0.948	(0.024)

Table 5: *Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity. Sample size = 100, $\alpha = 1$.*

Manhattan							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.951	(0.054)	0.874	(0.081)	0.744	(0.101)
	3NN	0.948	(0.052)	0.871	(0.077)	0.740	(0.097)
	5NN	0.945	(0.051)	0.867	(0.077)	0.737	(0.097)
	10NN	0.935	(0.056)	0.856	(0.079)	0.726	(0.098)
	20NN	0.888	(0.077)	0.808	(0.094)	0.684	(0.106)
Specificity	1NN	0.746	(0.053)	0.856	(0.044)	0.931	(0.032)
	3NN	0.746	(0.052)	0.854	(0.042)	0.931	(0.030)
	5NN	0.745	(0.052)	0.854	(0.042)	0.930	(0.030)
	10NN	0.745	(0.052)	0.853	(0.041)	0.929	(0.030)
	20NN	0.740	(0.052)	0.849	(0.042)	0.926	(0.030)
Euclidean							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.950	(0.056)	0.873	(0.083)	0.744	(0.103)
	3NN	0.947	(0.053)	0.870	(0.079)	0.741	(0.099)
	5NN	0.946	(0.052)	0.869	(0.077)	0.737	(0.097)
	10NN	0.936	(0.057)	0.858	(0.079)	0.726	(0.098)
	20NN	0.888	(0.075)	0.809	(0.092)	0.683	(0.106)
Specificity	1NN	0.745	(0.054)	0.855	(0.044)	0.931	(0.032)
	3NN	0.745	(0.053)	0.855	(0.043)	0.931	(0.031)
	5NN	0.744	(0.052)	0.853	(0.042)	0.930	(0.030)
	10NN	0.743	(0.052)	0.853	(0.042)	0.929	(0.030)
	20NN	0.739	(0.051)	0.848	(0.042)	0.926	(0.030)
Lagrange							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.952	(0.054)	0.875	(0.081)	0.745	(0.101)
	3NN	0.948	(0.052)	0.871	(0.077)	0.740	(0.097)
	5NN	0.945	(0.053)	0.867	(0.077)	0.736	(0.097)
	10NN	0.931	(0.058)	0.852	(0.081)	0.723	(0.098)
	20NN	0.869	(0.077)	0.791	(0.095)	0.669	(0.106)
Specificity	1NN	0.746	(0.053)	0.855	(0.044)	0.931	(0.032)
	3NN	0.746	(0.052)	0.854	(0.042)	0.931	(0.030)
	5NN	0.745	(0.052)	0.854	(0.042)	0.930	(0.030)
	10NN	0.744	(0.052)	0.852	(0.042)	0.929	(0.030)
	20NN	0.738	(0.052)	0.847	(0.042)	0.925	(0.030)
Mahalanobis							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.944	(0.061)	0.868	(0.084)	0.736	(0.104)
	3NN	0.933	(0.059)	0.855	(0.081)	0.723	(0.100)
	5NN	0.921	(0.061)	0.842	(0.082)	0.711	(0.099)
	10NN	0.886	(0.066)	0.805	(0.085)	0.677	(0.100)
	20NN	0.798	(0.075)	0.718	(0.088)	0.601	(0.098)
Specificity	1NN	0.744	(0.053)	0.854	(0.045)	0.931	(0.032)
	3NN	0.743	(0.052)	0.853	(0.044)	0.930	(0.030)
	5NN	0.743	(0.052)	0.852	(0.042)	0.929	(0.030)
	10NN	0.740	(0.052)	0.849	(0.042)	0.926	(0.030)
	20NN	0.730	(0.052)	0.840	(0.044)	0.920	(0.030)

Table 6: *Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity. Sample size = 100, $\alpha = 1.5$.*

Manhattan							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.991	(0.024)	0.968	(0.044)	0.917	(0.066)
	3NN	0.989	(0.024)	0.967	(0.041)	0.914	(0.063)
	5NN	0.988	(0.024)	0.964	(0.041)	0.911	(0.063)
	10NN	0.981	(0.032)	0.956	(0.047)	0.901	(0.066)
	20NN	0.932	(0.068)	0.903	(0.078)	0.848	(0.090)
Specificity	1NN	0.730	(0.053)	0.821	(0.046)	0.896	(0.037)
	3NN	0.731	(0.052)	0.821	(0.046)	0.896	(0.037)
	5NN	0.731	(0.052)	0.821	(0.045)	0.896	(0.036)
	10NN	0.731	(0.052)	0.821	(0.045)	0.895	(0.036)
	20NN	0.727	(0.052)	0.817	(0.045)	0.892	(0.036)
Euclidean							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.991	(0.024)	0.969	(0.043)	0.917	(0.066)
	3NN	0.990	(0.022)	0.967	(0.041)	0.914	(0.063)
	5NN	0.988	(0.024)	0.966	(0.042)	0.911	(0.064)
	10NN	0.982	(0.032)	0.958	(0.047)	0.902	(0.067)
	20NN	0.935	(0.065)	0.907	(0.077)	0.851	(0.091)
Specificity	1NN	0.731	(0.053)	0.824	(0.047)	0.898	(0.037)
	3NN	0.731	(0.053)	0.823	(0.046)	0.898	(0.037)
	5NN	0.731	(0.052)	0.821	(0.046)	0.897	(0.036)
	10NN	0.731	(0.052)	0.821	(0.046)	0.896	(0.036)
	20NN	0.727	(0.052)	0.817	(0.046)	0.893	(0.037)
Lagrange							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.991	(0.024)	0.968	(0.044)	0.917	(0.066)
	3NN	0.989	(0.024)	0.967	(0.041)	0.914	(0.063)
	5NN	0.988	(0.024)	0.964	(0.042)	0.911	(0.063)
	10NN	0.979	(0.035)	0.953	(0.050)	0.899	(0.069)
	20NN	0.920	(0.069)	0.890	(0.081)	0.835	(0.094)
Specificity	1NN	0.730	(0.053)	0.821	(0.046)	0.897	(0.037)
	3NN	0.731	(0.052)	0.821	(0.045)	0.896	(0.036)
	5NN	0.731	(0.052)	0.821	(0.045)	0.896	(0.036)
	10NN	0.730	(0.052)	0.820	(0.045)	0.895	(0.036)
	20NN	0.726	(0.052)	0.816	(0.046)	0.891	(0.037)
Mahalanobis							
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity	1NN	0.985	(0.032)	0.962	(0.050)	0.907	(0.071)
	3NN	0.975	(0.035)	0.950	(0.050)	0.893	(0.070)
	5NN	0.963	(0.039)	0.935	(0.054)	0.878	(0.071)
	10NN	0.923	(0.051)	0.891	(0.063)	0.833	(0.079)
	20NN	0.820	(0.072)	0.783	(0.079)	0.728	(0.088)
Specificity	1NN	0.731	(0.053)	0.821	(0.047)	0.897	(0.037)
	3NN	0.731	(0.053)	0.821	(0.046)	0.896	(0.037)
	5NN	0.731	(0.053)	0.820	(0.046)	0.895	(0.037)
	10NN	0.728	(0.053)	0.817	(0.046)	0.892	(0.037)
	20NN	0.717	(0.052)	0.807	(0.047)	0.884	(0.039)

S2. Simulation study: scenario (i)

The following tables report supplementary simulation results under scenario (i) of the main paper, for values of α equal to 0.1 and 0.25, giving rise respectively to true AUC values of about 0.59 and 0.71, and sample sizes equal to 50 and 100. The number of replicates in each simulation experiment is 5000.

Table 7: *Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity, when the models for $\rho(y)$ and $\pi(y)$ are correctly specified. “True” denotes the true parameter value. KNN estimators are based in the Euclidean distance. Sample size = 50.*

		$\alpha = 0.1$					
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity							
True		0.444		0.228		0.097	0.020
IPW		0.449	(0.169)	0.241	(0.137)	0.114	(0.094) 0.021 (0.047)
MSI		0.446	(0.154)	0.241	(0.125)	0.113	(0.033) 0.020 (0.043)
SPE		0.446	(0.159)	0.241	(0.129)	0.113	(0.084) 0.021 (0.044)
1NN		0.447	(0.163)	0.240	(0.131)	0.113	(0.088) 0.020 (0.043)
3NN		0.444	(0.154)	0.239	(0.125)	0.112	(0.084) 0.020 (0.042)
Specificity							
True		0.684		0.859		0.952	0.996
IPW		0.679	(0.107)	0.856	(0.079)	0.957	(0.045) 0.998 (0.010)
MSI		0.682	(0.079)	0.859	(0.059)	0.955	(0.035) 0.997 (0.009)
SPE		0.682	(0.080)	0.859	(0.060)	0.956	(0.036) 0.997 (0.009)
1NN		0.682	(0.080)	0.859	(0.060)	0.955	(0.036) 0.997 (0.010)
3NN		0.682	(0.079)	0.858	(0.059)	0.955	(0.035) 0.997 (0.009)
		$\alpha = 0.25$					
		$c = 0.2$		$c = 0.5$		$c = 0.8$	
Sensitivity							
True		0.588		0.361		0.173	0.049
IPW		0.598	(0.174)	0.369	(0.161)	0.190	(0.120) 0.051 (0.069)
MSI		0.593	(0.157)	0.367	(0.148)	0.189	(0.109) 0.051 (0.064)
SPE		0.594	(0.166)	0.364	(0.147)	0.189	(0.111) 0.051 (0.065)
1NN		0.594	(0.169)	0.368	(0.154)	0.189	(0.113) 0.051 (0.066)
3NN		0.589	(0.158)	0.364	(0.147)	0.186	(0.109) 0.050 (0.064)
Specificity							
True		0.717		0.874		0.956	0.995
IPW		0.713	(0.106)	0.870	(0.072)	0.956	(0.042) 0.997 (0.012)
MSI		0.717	(0.077)	0.873	(0.056)	0.957	(0.034) 0.996 (0.011)
SPE		0.717	(0.078)	0.873	(0.057)	0.957	(0.034) 0.996 (0.011)
1NN		0.717	(0.079)	0.872	(0.057)	0.957	(0.034) 0.996 (0.011)
3NN		0.716	(0.077)	0.872	(0.056)	0.956	(0.033) 0.996 (0.011)

Table 8: Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity, when the models for $\rho(y)$ and $\pi(y)$ are correctly specified. “True” denotes the true parameter value. KNN estimators are based in the Euclidean distance. Sample size = 100.

		$\alpha = 0.1$							
		$c = 0.2$		$c = 0.5$		$c = 0.8$		$c = 1.2$	
Sensitivity									
True		0.444		0.228		0.097		0.020	
IPW		0.448	(0.119)	0.230	(0.095)	0.099	(0.062)	0.021	(0.029)
MSI		0.446	(0.109)	0.230	(0.087)	0.098	(0.056)	0.020	(0.026)
SPE		0.446	(0.113)	0.230	(0.090)	0.098	(0.058)	0.021	(0.027)
1NN		0.446	(0.117)	0.230	(0.093)	0.098	(0.059)	0.020	(0.027)
3NN		0.445	(0.112)	0.229	(0.089)	0.097	(0.056)	0.020	(0.026)
Specificity									
True		0.684		0.859		0.952		0.996	
IPW		0.682	(0.075)	0.857	(0.053)	0.953	(0.031)	0.997	(0.009)
MSI		0.683	(0.056)	0.859	(0.041)	0.952	(0.025)	0.996	(0.008)
SPE		0.684	(0.056)	0.859	(0.042)	0.952	(0.025)	0.996	(0.008)
1NN		0.683	(0.057)	0.859	(0.042)	0.952	(0.025)	0.996	(0.008)
3NN		0.683	(0.056)	0.859	(0.041)	0.952	(0.025)	0.996	(0.008)
		$\alpha = 0.25$							
		$c = 0.2$		$c = 0.5$		$c = 0.8$		$c = 1.2$	
Sensitivity									
True		0.588		0.361		0.173		0.049	
IPW		0.589	(0.107)	0.364	(0.111)	0.174	(0.083)	0.050	(0.045)
MSI		0.589	(0.107)	0.363	(0.102)	0.175	(0.077)	0.050	(0.045)
SPE		0.589	(0.111)	0.363	(0.105)	0.175	(0.079)	0.050	(0.043)
1NN		0.588	(0.115)	0.362	(0.107)	0.174	(0.080)	0.050	(0.043)
3NN		0.586	(0.110)	0.360	(0.103)	0.174	(0.078)	0.050	(0.042)
Specificity									
True		0.717		0.874		0.956		0.995	
IPW		0.715	(0.070)	0.873	(0.048)	0.955	(0.029)	0.995	(0.010)
MSI		0.717	(0.052)	0.874	(0.039)	0.956	(0.024)	0.995	(0.009)
SPE		0.717	(0.053)	0.874	(0.040)	0.956	(0.024)	0.995	(0.009)
1NN		0.716	(0.054)	0.873	(0.040)	0.956	(0.025)	0.995	(0.009)
3NN		0.716	(0.053)	0.873	(0.040)	0.956	(0.024)	0.994	(0.009)

S3. Simulation study: scenario (ii)

The following table reports supplementary simulation results under scenario (ii) of the main paper and shows a good and stable behavior of KNN estimators when the sample size is equal to 300. The choice of such sample size depends mostly on the value of the verification rate, which is quite small in this setting, especially for healthy subjects.

Table 9: *Mean estimated sensitivity, mean estimated specificity and standard deviations (Monte Carlo/estimated) from 5000 replications in scenario (ii) for 1NN and 3NN estimators based on the Euclidean distance. “Full” indicates the estimator based on complete data, with does not change with δ . Sample size = 300.*

δ	Sensitivity			Specificity		
	1NN	3NN	Full	1NN	3NN	Full
$c = 0.2$						
0.1	0.886 (0.084/0.075)	0.875 (0.076/0.072)		0.603 (0.045/0.044)	0.605 (0.040/0.044)	
0.3	0.886 (0.085/0.075)	0.874 (0.075/0.072)		0.603 (0.046/0.044)	0.604 (0.041/0.044)	
0.5	0.884 (0.088/0.077)	0.869 (0.079/0.074)	0.889 (0.037)	0.605 (0.045/0.045)	0.605 (0.041/0.044)	0.603 (0.033)
0.7	0.881 (0.094/0.080)	0.866 (0.082/0.076)		0.604 (0.047/0.045)	0.605 (0.041/0.045)	
0.9	0.876 (0.101/0.089)	0.858 (0.084/0.083)		0.604 (0.049/0.048)	0.606 (0.042/0.048)	
$c = 0.4$						
0.1	0.819 (0.097/0.086)	0.806 (0.088/0.082)		0.744 (0.042/0.041)	0.745 (0.036/0.041)	
0.3	0.819 (0.099/0.086)	0.804 (0.090/0.082)		0.746 (0.042/0.041)	0.746 (0.036/0.041)	
0.5	0.820 (0.103/0.087)	0.802 (0.092/0.083)	0.820 (0.045)	0.745 (0.042/0.041)	0.745 (0.041/0.041)	0.743 (0.029)
0.7	0.817 (0.109/0.091)	0.798 (0.098/0.086)		0.745 (0.043/0.042)	0.746 (0.038/0.042)	
0.9	0.813 (0.123/0.100)	0.791 (0.104/0.093)		0.746 (0.045/0.045)	0.748 (0.038/0.044)	
$c = 0.6$						
0.1	0.742 (0.104/0.093)	0.725 (0.095/0.087)		0.853 (0.034/0.035)	0.854 (0.029/0.035)	
0.3	0.738 (0.108/0.093)	0.720 (0.099/0.088)		0.854 (0.034/0.034)	0.855 (0.030/0.034)	
0.5	0.735 (0.113/0.094)	0.718 (0.101/0.089)	0.737 (0.051)	0.853 (0.034/0.035)	0.854 (0.030/0.034)	0.852 (0.024)
0.7	0.738 (0.119/0.097)	0.718 (0.107/0.092)		0.854 (0.034/0.035)	0.855 (0.030/0.035)	
0.9	0.736 (0.136/0.107)	0.710 (0.114/0.098)		0.855 (0.035/0.036)	0.857 (0.030/0.036)	

S4. Simulation study: multidimensional X vector

In this section, we report some simulation results based on the introduction of a multidimensional vector X of observed covariates. In detail, $X = ({}_1X, {}_2X, {}_3X)^\top$. The simulation setting is a generalization of that in scenario (i) of the main paper. Starting from four independent random variables $Z_1 \sim N(0, 0.5)$ to $Z_4 \sim N(0, 0.5)$, the disease indicator D is specified as $D = I[Z_1 + Z_2 + Z_3 + Z_4 > r_1]$. The threshold r_1 determines a disease prevalence of about 0.30. The diagnostic test result T and the auxiliary covariates are generated as follows:

- $T = \alpha \sum_{i=1}^4 Z_i + \epsilon_1$;
- ${}_1X = \sum_{i=1}^4 Z_i + \epsilon_2$;
- ${}_2X = 0.5Z_1 + 2Z_2 + 1.5Z_3 + Z_4 + \epsilon_3$;
- ${}_3X = 2Z_1 - 0.5Z_2 + Z_3 + 0.5Z_4 + \epsilon_4$;

where ϵ_i , $i = 1, \dots, 4$, are independent $N(0, 0.25)$ random variables, independent also from Z_i , $i = 1, \dots, 4$. Finally, the verification probability π is set to be

$$\pi(T, X) = \frac{e^{\delta_0 + \delta_1 T + \delta_2^\top X}}{1 + e^{\delta_0 + \delta_1 T + \delta_2^\top X}},$$

with $\delta_0 = 0.05$, $\delta_1 = 0.9$, $\delta_2 = (0.7, 0.4, 0.2)^\top$. This choice corresponds to a verification rate of about 0.5. For the parametric estimators IPW, MSI, SPE, disease probabilities and verification probabilities are estimated using correctly specified models. For the KNN estimators, the Euclidean distance is employed. The number of replicates in each simulation experiment is 5000.

Results, given in Table 10 and 11, show, as expected, a certain efficiency loss of KNN estimators compared to parametric competitors. Among parametric estimators, the IPW shows, comparatively, poorer performances in some cases.

Table 10: Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity. “Full” indicates the estimator based on the complete data. Dimension of X is 3. KNN estimators are based in the Euclidean distance. Sample size = 100.

		$\alpha = 0.5$							
		$c = 0.2$		$c = 0.5$		$c = 0.8$		$c = 1.2$	
Sensitivity									
Full		0.834	(0.067)	0.674	(0.085)	0.481	(0.091)	0.246	(0.077)
IPW		0.833	(0.074)	0.674	(0.090)	0.482	(0.096)	0.247	(0.079)
MSI		0.832	(0.070)	0.673	(0.087)	0.481	(0.094)	0.246	(0.078)
SPE		0.833	(0.072)	0.673	(0.088)	0.481	(0.094)	0.246	(0.078)
1NN		0.821	(0.074)	0.663	(0.089)	0.474	(0.095)	0.242	(0.077)
3NN		0.812	(0.072)	0.654	(0.088)	0.466	(0.093)	0.239	(0.076)
Specificity									
Full		0.788	(0.048)	0.899	(0.037)	0.962	(0.023)	0.993	(0.010)
IPW		0.752	(0.106)	0.883	(0.066)	0.955	(0.036)	0.992	(0.014)
MSI		0.788	(0.049)	0.899	(0.037)	0.962	(0.023)	0.993	(0.010)
SPE		0.788	(0.049)	0.899	(0.038)	0.962	(0.024)	0.993	(0.010)
1NN		0.788	(0.050)	0.899	(0.038)	0.962	(0.023)	0.993	(0.010)
3NN		0.787	(0.049)	0.898	(0.037)	0.962	(0.023)	0.993	(0.010)
		$\alpha = 1$							
		$c = 0.2$		$c = 0.5$		$c = 0.8$		$c = 1.2$	
Sensitivity									
Full		0.967	(0.032)	0.912	(0.050)	0.815	(0.070)	0.638	(0.087)
IPW		0.967	(0.037)	0.912	(0.056)	0.816	(0.075)	0.639	(0.090)
MSI		0.966	(0.034)	0.911	(0.053)	0.814	(0.073)	0.637	(0.089)
SPE		0.966	(0.036)	0.912	(0.055)	0.814	(0.074)	0.637	(0.089)
1NN		0.956	(0.042)	0.899	(0.060)	0.802	(0.078)	0.627	(0.091)
3NN		0.945	(0.045)	0.887	(0.062)	0.790	(0.078)	0.618	(0.091)
Specificity									
Full		0.794	(0.049)	0.883	(0.039)	0.944	(0.028)	0.985	(0.015)
IPW		0.736	(0.110)	0.849	(0.075)	0.928	(0.046)	0.981	(0.022)
MSI		0.794	(0.050)	0.883	(0.040)	0.944	(0.029)	0.985	(0.015)
SPE		0.795	(0.050)	0.883	(0.040)	0.944	(0.029)	0.985	(0.015)
1NN		0.795	(0.051)	0.882	(0.041)	0.943	(0.029)	0.985	(0.015)
3NN		0.795	(0.050)	0.881	(0.040)	0.943	(0.029)	0.984	(0.015)
		$\alpha = 1.5$							
		$c = 0.2$		$c = 0.5$		$c = 0.8$		$c = 1.2$	
Sensitivity									
Full		0.994	(0.014)	0.980	(0.025)	0.945	(0.041)	0.851	(0.065)
IPW		0.994	(0.016)	0.981	(0.028)	0.945	(0.045)	0.851	(0.069)
MSI		0.993	(0.015)	0.979	(0.027)	0.944	(0.043)	0.849	(0.068)
SPE		0.994	(0.017)	0.980	(0.029)	0.944	(0.044)	0.849	(0.068)
1NN		0.987	(0.025)	0.971	(0.035)	0.934	(0.050)	0.839	(0.073)
3NN		0.978	(0.030)	0.961	(0.040)	0.923	(0.053)	0.829	(0.074)
Specificity									
Full		0.783	(0.050)	0.857	(0.043)	0.916	(0.034)	0.969	(0.021)
IPW		0.696	(0.127)	0.803	(0.092)	0.883	(0.065)	0.956	(0.035)
MSI		0.783	(0.051)	0.857	(0.044)	0.916	(0.035)	0.969	(0.022)
SPE		0.783	(0.051)	0.857	(0.044)	0.916	(0.035)	0.969	(0.022)
1NN		0.784	(0.051)	0.857	(0.044)	0.915	(0.035)	0.968	(0.022)
3NN		0.784	(0.051)	0.857	(0.044)	0.915	(0.035)	0.968	(0.022)

Table 11: Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity. “Full” indicates the estimator based on the complete data. Dimension of X is 3. KNN estimators are based in the Euclidean distance. Sample size = 200.

		$\alpha = 0.5$							
		$c = 0.2$		$c = 0.5$		$c = 0.8$		$c = 1.2$	
Sensitivity									
Full		0.834	(0.047)	0.673	(0.060)	0.480	(0.063)	0.247	(0.054)
IPW		0.834	(0.051)	0.674	(0.064)	0.480	(0.065)	0.248	(0.055)
MSI		0.834	(0.049)	0.674	(0.062)	0.480	(0.064)	0.247	(0.055)
SPE		0.834	(0.050)	0.674	(0.062)	0.480	(0.064)	0.247	(0.055)
1NN		0.827	(0.051)	0.666	(0.064)	0.475	(0.065)	0.244	(0.055)
3NN		0.822	(0.050)	0.662	(0.062)	0.471	(0.064)	0.243	(0.054)
Specificity									
Full		0.788	(0.036)	0.900	(0.026)	0.962	(0.017)	0.993	(0.007)
IPW		0.767	(0.075)	0.888	(0.045)	0.958	(0.024)	0.992	(0.009)
MSI		0.788	(0.036)	0.900	(0.026)	0.962	(0.017)	0.993	(0.007)
SPE		0.788	(0.037)	0.900	(0.027)	0.962	(0.017)	0.993	(0.007)
1NN		0.788	(0.037)	0.899	(0.027)	0.962	(0.017)	0.993	(0.007)
3NN		0.788	(0.036)	0.899	(0.026)	0.962	(0.017)	0.993	(0.007)
		$\alpha = 1$							
		$c = 0.2$		$c = 0.5$		$c = 0.8$		$c = 1.2$	
Sensitivity									
Full		0.967	(0.023)	0.912	(0.036)	0.816	(0.050)	0.635	(0.062)
IPW		0.967	(0.026)	0.912	(0.040)	0.816	(0.053)	0.636	(0.064)
MSI		0.967	(0.023)	0.912	(0.038)	0.816	(0.052)	0.635	(0.063)
SPE		0.967	(0.025)	0.912	(0.039)	0.816	(0.052)	0.635	(0.063)
1NN		0.960	(0.028)	0.903	(0.042)	0.808	(0.054)	0.629	(0.064)
3NN		0.955	(0.027)	0.898	(0.041)	0.802	(0.054)	0.624	(0.064)
Specificity									
Full		0.794	(0.034)	0.883	(0.027)	0.944	(0.020)	0.986	(0.010)
IPW		0.756	(0.079)	0.861	(0.052)	0.935	(0.030)	0.983	(0.014)
MSI		0.794	(0.035)	0.883	(0.028)	0.944	(0.020)	0.986	(0.010)
SPE		0.794	(0.035)	0.883	(0.028)	0.944	(0.020)	0.986	(0.011)
1NN		0.794	(0.035)	0.882	(0.028)	0.944	(0.020)	0.985	(0.011)
3NN		0.794	(0.035)	0.882	(0.028)	0.943	(0.020)	0.985	(0.010)
		$\alpha = 1.5$							
		$c = 0.2$		$c = 0.5$		$c = 0.8$		$c = 1.2$	
Sensitivity									
Full		0.995	(0.009)	0.980	(0.017)	0.945	(0.029)	0.852	(0.045)
IPW		0.994	(0.011)	0.980	(0.020)	0.945	(0.032)	0.852	(0.047)
MSI		0.994	(0.009)	0.980	(0.018)	0.945	(0.031)	0.852	(0.046)
SPE		0.994	(0.011)	0.980	(0.019)	0.946	(0.031)	0.852	(0.046)
1NN		0.991	(0.014)	0.974	(0.023)	0.938	(0.035)	0.845	(0.049)
3NN		0.987	(0.014)	0.970	(0.022)	0.933	(0.034)	0.841	(0.048)
Specificity									
Full		0.783	(0.035)	0.856	(0.030)	0.917	(0.023)	0.969	(0.015)
IPW		0.728	(0.090)	0.820	(0.066)	0.896	(0.043)	0.962	(0.023)
MSI		0.783	(0.036)	0.856	(0.031)	0.917	(0.024)	0.970	(0.015)
SPE		0.783	(0.036)	0.856	(0.031)	0.917	(0.024)	0.969	(0.015)
1NN		0.784	(0.036)	0.856	(0.031)	0.916	(0.024)	0.969	(0.015)
3NN		0.785	(0.036)	0.856	(0.031)	0.916	(0.024)	0.969	(0.015)