Supplementary material to
Double logistic regression approach to biased positive-unlabelled data

Konrad Furmańczyk[a], Jan Mielniczuk[b,d], Wojciech Rejchel[c] and Paweł Teisseyre[b,d;*]

ORCiD ID: Konrad Furmańczyk https://orcid.org/0000-0002-7683-4787,
Jan Mielniczuk https://orcid.org/0000-0003-2621-2303, Wojciech Rejchel https://orcid.org/0000-0003-1148-1439,
Paweł Teisseyre https://orcid.org/0000-0002-4296-9819

Appendix 1

Proof of Theorem 1

Assume that the alternative parametrisation exists. Then the assumption yields for $x := -x$ (with substitution $\beta_0^* := -\beta_0^*, \gamma_0^* := -\gamma_0^*$ and analogously for β_0, γ_0):

$$e^{\beta_0^* + \beta^{*T} x} + e^{\gamma_0^* + \gamma^{*T} x} + e^{\beta_0^* + \beta^{*T} x + \gamma_0^* + \gamma^{*T} x} = e^{\beta_0 + \beta^T x} + e^{\gamma_0 + \gamma^T x} + e^{\beta_0 + \beta^T x + \gamma_0 + \gamma^T x} \tag{1}$$

Assume that vectors β and γ have the same support and consider the case $\beta_i^* > 0, \gamma_i^* > 0$ for some fixed $i \leq p$, let $x_n = (0, \ldots, 0, n, 0, \ldots, 0)^T$ with n at i-th place, $\tilde{x}_n = -x_n$. Then $\beta_i > 0, \gamma_i > 0$ as for $\beta_i > 0, \gamma_i < 0$ the dominating term on the RHS of (1) is $e^{\gamma_0} e^{-\gamma_i n} \to \infty$ whereas the LHS$\to$ 0. The case $\beta_i < 0, \gamma_i < 0$ is excluded in the same way. For $\beta_i > 0, \gamma_i > 0$ and x_n the dominating term of the RHS is $e^{\beta_0 + \gamma_0} e^{(\beta_i + \gamma_i)n}$. Comparing rates and coefficients of dominating terms

$$\beta_i + \gamma_i = \beta_i^* + \gamma_i^*, \quad \beta_0 + \gamma_0 = \beta_0^* + \gamma_0^* \tag{2}$$

By the same token as before we have that if $\beta_i^* > \gamma_i^* > 0$ then either $\beta_i > \gamma_i > 0$ or $\gamma_i > \beta_i > 0$. For the first case comparing rates and coefficients for $\tilde{x}_n$ again we get $\gamma_i = \gamma_i^*$ and $\gamma_0 = \gamma_0^*$ and (2) implies that $\beta_i = \beta_i^*$ and $\beta_0 = \beta_0^*$. If $\gamma_i > \beta_i > 0$ then the parameters are interchanged i.e. $\gamma_i = \beta_i^*, \beta_i = \gamma_i^*$ and $\beta_0 = \gamma_0^*, \gamma_0 = \beta_0^*$. Note that since i is arbitrary, it follows from the proof that if the parameters are interchanged for a certain i there are interchanged *for any* i as this is determined by the fact whether intercepts are interchanged or not. All other cases are dealt with analogously, as well the case of arbitrary supports of β and γ. For example it is easy to see that if $\beta_i > 0$ and $\gamma_i = 0$ then either $\beta_i^* > 0$ and $\gamma_i^* = 0$ or $\gamma_i^* > 0$ and $\beta_i^* = 0$.

Proof of Lemma 1

Fix x. The conditional expected value

$$E_{S|X=x} \left[S \log s_{\tilde{\beta}, \tilde{\gamma}}(X) + (1 - S) \log(1 - s_{\tilde{\beta}, \tilde{\gamma}}(X)) \right]$$

is maximised by success probability $s_{\tilde{\beta}^*, \tilde{\gamma}^*}(x)$ by Information Inequality [1, Theorem 2.6.3]. If there exists another vector (a, b), which also maximizes $Q(\tilde{\beta}, \tilde{\gamma})$, then in view of Theorem 1 we obtain $(\tilde{\beta}^*, \tilde{\gamma}^*) = (a, b)$ or $(\tilde{\beta}^*, \tilde{\gamma}^*) = (b, a)$. So, uniqueness follows from the assumption $|\tilde{\beta}^*|_1 > |\tilde{\gamma}^*|_1$ as the interchange of parameters is impossible in such a case.

Proof of Theorem 2

A set $C = \{(\tilde{\beta}, \tilde{\gamma}) : |\tilde{\beta}|_1 > |\tilde{\gamma}|_1\}$ is open in R^{2p+2}. Using the fact that $|\tilde{\beta}^*|_1 > |\tilde{\gamma}^*|_1$, we can choose a small enough number $r > 0$ such that $K := K((\tilde{\beta}^*, \tilde{\gamma}^*), r) \subset C$. Notice that with probability one $Q_n(\tilde{\beta}, \tilde{\gamma})$ converges uniformly to $Q(\tilde{\beta}, \tilde{\gamma})$ on K by [2, Theorem 16a].
Next, functions $Q_n(\tilde{\beta}, \tilde{\gamma})$ are continuous, so there exists a sequence $(\hat{\beta}_n, \hat{\gamma}_n) \in K$ such that $\inf_{(\tilde{\beta}, \tilde{\gamma}) \in K} Q_n(\tilde{\beta}, \tilde{\gamma}) = Q_n(\hat{\beta}_n, \hat{\gamma}_n)$. From the fact that $(\tilde{\beta}^*, \tilde{\gamma}^*)$ uniquely maximizes $Q(\tilde{\beta}, \tilde{\gamma})$, uniform convergence of $Q_n(\tilde{\beta}, \tilde{\gamma})$ and standard arguments (as e.g. in [4]) it follows that $(\hat{\beta}_n, \hat{\gamma}_n) \to_{a.s.} (\tilde{\beta}^*, \tilde{\gamma}^*)$. This implies that for n large enough $(\hat{\beta}_n, \hat{\gamma}_n)$ is in the interior of K, so it is a local maximiser of Q_n.

Proof of Theorem 3

The proofs of Theorem 3 and Theorem 4 refer to a simple consequence of Lemma 2 in [3]. For the sake of completeness we state it below as Lemma 3.

Lemma 3 *Suppose $A_n(s)$ is a sequence of concave random functions defined on an open convex set S of the Euclidean space, which converges in probability to $A(s)$ for each $s \in S$. If s_0 is the unique maximiser of A, then every maximiser of A_n converges to s_0 in probability.*

It is easily seen that $W_n(\tilde{\beta})$ can be written as

$$W_n(\tilde{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_1(S_i, X_i)(\beta_0 + \beta^T X_i) + \log(1 - y(X_i, \tilde{\beta})),$$

From concavity of $W_n(\tilde{\beta})$ and Lemma 3 we have only to prove pointwise convergence: $W_n(\tilde{\beta}) \to_P E_X W(X, \tilde{\beta})$ for each $\tilde{\beta}$. Notice that

$$E_X W(X, \tilde{\beta}) = E_{X,S}[w_1(S, X)(\beta_0 + \beta^T X) + \log(1 - y(X, \tilde{\beta}))].$$

* Corresponding Author. Email: teisseyrep@ipipan.waw.pl

Thus, the required convergence is a simple consequence of LLN and assumption of Theorem 3.

Proof of Theorem 4

Function $\hat{R}_n(\tilde{\gamma})$ is concave, thus again using Lemma 3 we have only to prove point-wise convergence: $\hat{R}_n(\tilde{\gamma}) \to_P R(\tilde{\gamma})$ for each $\tilde{\gamma}$. We can write

$$\hat{R}_n(\tilde{\gamma}) - R(\tilde{\gamma}) = \hat{R}_n(\tilde{\gamma}) - R_n(\tilde{\gamma}) + R_n(\tilde{\gamma}) - R(\tilde{\gamma}). \quad (3)$$

To finish the proof, we have to establish that the second difference on the right hand side of (3) goes to zero in probability. To this end we notice that $R_n(\tilde{\gamma}) = \frac{n}{n_1}\frac{1}{n}\sum_{i=1}^n Y_i K(S_i, X_i, \tilde{\gamma})$ and $R(\tilde{\gamma}) = EYK(S, X, \tilde{\gamma})/P(Y = 1)$ and apply LLN.

Generalization of Theorem 1

Theorem 5 *Assume that $s(x)$ is defined as*

$$s(x) = P(S = 1|Y = 1, x)P(Y = 1|x) = e(x)y(x).$$

and

$$y(x) = p(\beta_0^* + \beta^{*T}x), \qquad e(x) = p(\gamma_0^* + \gamma^{*T}x)$$

for a function p such that its logarithmic derivative $h(s) = p'(s)/p(s)$ satisfies the following conditions:
(i) h is a positive 1-1 function;
(ii) for all $a > 0, a > b, a_0, a_1 \in R$ $h(a_0 + as)/h(a_1 + bs) \to 0$ as $s \to \infty$;
(iii) for all $a_0 \neq b_0, a_0, b_0 \in R, b > 0$ $h(a_0 + bs)/h(b_0 + bs)$ tends to a finite limit different from 1 as $s \to \infty$;
(iv) for all $a_0, b_0 \in R, b < 0$ $h(a_0+bs)/h(b_0+bs) \to 1$ as $s \to \infty$.
Then parameters $\tilde{\beta}^*$ and $\tilde{\gamma}^*$ are uniquely defined up to interchange of $y(x)$ and $e(x)$ i.e. if for some $\tilde{\beta}$ and $\tilde{\gamma}$ $s(x) = p(\beta_0+\beta^T x)p(\gamma_0+\gamma^T x)$ for all $x \in R^p$, then $(\tilde{\beta}, \tilde{\gamma}) = (\tilde{\beta}^*, \tilde{\gamma}^*)$ or $(\tilde{\beta}, \tilde{\gamma}) = (\tilde{\gamma}^*, \tilde{\beta}^*)$

Note that the above assumptions are satisfied for $p(\cdot) = \sigma(\cdot)$ for logistic model or $p(\cdot) = \Phi(\cdot)$ for probit model and thus Theorem 5 generalises Theorem 1. Observe that (ii) is equivalent to (ii*): for all $b < 0, a > b, a_0, b_0 \in R$ $h(a_0 + as)/h(a_1 + bs) \to \infty$ as $s \to -\infty$; and (iii) is equivalent to (iii*): for all $a_0, b_0 \in R, b < 0$ $h(a_0 + bs)/h(b_0 + bs) \to K(a_0, b_0) \neq 1$ as $s \to -\infty$.

Proof Assume that the alternative parametrisation exists. Then for all $x \in R^p$

$$p(\beta_0^* + \beta^{*T}x)p(\gamma_0^* + \gamma^{*T}x) = p(\beta_0 + \beta^T x)p(\gamma_0 + \gamma^T x) \quad (4)$$

Taking the logs of both sides of (4) and then the derivative with respect to x, we have

$$\beta^* h(\beta_0^* + \beta^{*T}x) + \gamma^* h(\gamma_0^* + \gamma^{*T}x) = \beta h(\beta_0 + \beta^T x) + \gamma h(\gamma_0 + \gamma^T x) \quad (5)$$

Letting $x = 0$, we obtain

$$\beta^* h(\beta_0^*) + \gamma^* h(\gamma_0^*) = \beta h(\beta_0) + \gamma h(\gamma_0) \quad (6)$$

It is enough to show that for all $1 \leq i \leq p$: (A1) $\beta_i = \beta_i^*, \gamma_i = \gamma_i^*, \beta_0 = \beta_0^*, \gamma_0 = \gamma_0^*$ or $\beta_i = \gamma_i^*, \gamma_i = \beta_i^*, \beta_0 = \gamma_0^*, \gamma_0 = \beta_0^*$. Indeed, assume there exist $i \neq j$ such that $\beta_i = \beta_i^*, \gamma_i = \gamma_i^*, \beta_0^* = \beta_0, \gamma_0^* = \gamma_0$ and $\beta_j = \gamma_j^*, \gamma_j = \beta_j^*$. From (6), we have $\beta_j^* h(\beta_0^*) + \gamma_j^* h(\gamma_0^*) = \beta_j h(\beta_0) + \gamma_j h(\gamma_0) = \gamma_j^* h(\beta_0^*) + \beta_j^* h(\gamma_0^*)$. Then $h(\beta_0^*)(\beta_j^* - \gamma_j^*) = h(\gamma_0^*)(\beta_j^* - \gamma_j^*)$ and from (i) we obtain a contradiction if $\beta_j^* \neq \gamma_j^*$ and $\beta_0^* \neq \gamma_0^*$.

Now, we will check (A1). Fix $1 \leq i \leq p$, we deal with the following cases: (C1) $\beta_i^* > \gamma_i^* > 0$; (C2) $\gamma_i^* < \beta_i^* < 0$; (C3) $\gamma_i^* < 0, \beta_i^* > 0$. The remaining cases: (C1*) $\gamma_i^* > \beta_i^* > 0$, (C2*) $\beta_i^* < \gamma_i^* < 0$, (C3*) $0 > \beta_i^*, \gamma_i^* > 0$ follow in the same way. Consider i-th coordinate of (5)-(6). Setting $x = (0, \ldots, 0, t, 0, \ldots, 0)^T$ with t at i-th place, we have

$$\beta_i^* h(\beta_0^* + \beta_i^* t) + \gamma_i^* h(\gamma_0^* + \gamma_i^* t) = \beta_i h(\beta_0 + \beta_i t) + \gamma_i h(\gamma_0 + \gamma_i t) \quad (7)$$

and

$$\beta_i^* h(\beta_0^*) + \gamma_i^* h(\gamma_0^*) = \beta_i h(\beta_0) + \gamma_i h(\gamma_0) \quad (8)$$

For (C1) using (ii), we have $h(\beta_0^* + \beta_i^* t)/h(\gamma_0^* + \gamma_i^* t) \to 0$ as $t \to \infty$. Then the dominating term on the LHS of (7) is $\gamma_i^* h(\gamma_0^* + \gamma_i^* t)$. For $\beta_i > \gamma_i$ from (i) and (8), we conclude $\beta_i > 0$. Using (ii), we have $h(\beta_0 + \beta_i t)/h(\gamma_0 + \gamma_i t) \to 0$ as $t \to \infty$ and the dominating term on the RHS of (7) is $\gamma_i h(\gamma_0 + \gamma_i t)$. Therefore comparing the dominating terms of both sides of (7) $\frac{\gamma_i^* h(\gamma_0^* + \gamma_i^* t)}{\gamma_i h(\gamma_0 + \gamma_i t)} \to 1$ as $t \to \infty$, and consequently from (iii) we have $\gamma_i^* = \gamma_i, \gamma_0^* = \gamma_0$. Similarly comparing the remaining terms of both sides of (7), we get $\beta_i^* = \beta_i, \beta_0^* = \beta_0$. For $\beta_i < \gamma$ from (i) and (8), we have $\gamma_i > 0$, and the dominating term on the RHS of (7) is $\beta_i h(\beta_0 + \beta_i t)$. By a similar reasoning, we have $\frac{\gamma_i^* h(\gamma_0^* + \gamma_i^* t)}{\beta_i h(\beta_0 + \beta_i t)} \to 1$ as $t \to \infty$. Hence $\gamma_i^* = \beta_i, \gamma_0^* = \beta_0$ and consequently $\beta_i^* = \gamma_i$. In (C2) we use (ii*), (iii*) analogously to (ii), (iii) in (C1). For (C3), $\beta_i^* > \gamma_i^*$ and, we have the following possibilities: (P1) $\gamma_i < 0, \beta_i > 0$; (P2) $\gamma_i > 0, \beta_i < 0$; (P3) $\gamma_i > 0, \beta_i > 0$; (P4) $\gamma_i < 0, \beta_i < 0$. We only discuss (P3)-(P4) as (P1)-(P2) are obtained analogously as in (C1). In (P3) for $t \to \infty$ the dominating term on the LHS of (7) is $\gamma_i^* h(\gamma_0^* + \gamma_i^* t)$ and on the RHS of (7) is $\gamma_i h(\gamma_0 + \gamma_i t)$ or $\beta_i h(\beta_0 + \beta_i t)$. Comparing the dominating terms of the (7) from (ii), we have $\frac{\gamma_i h(\gamma_0 + \gamma_i t)}{\gamma_i^* h(\gamma_0^* + \gamma_i^* t)} \to 0 \neq 1$ as $t \to \infty$. Thus, we have a contradiction and similarly, $\frac{\beta_i h(\beta_0 + \beta_i t)}{\gamma_i^* h(\gamma_0^* + \gamma_i^* t)} \to 0 \neq 1$. For (P4), we consider $t \to -\infty$. The dominating term on the LHS of (7) is $\beta_i^* h(\beta_0^* + \beta_i^* t)$ and on the RHS is $\beta_i h(\beta_0 + \beta_i t)$ or $\gamma_i h(\gamma_0 + \gamma_i t)$. Therefore, from (ii*) we have a contradiction $\frac{\beta_i^* h(\beta_0^* + \beta_i^* t)}{\beta_i h(\beta_0 + \beta_i t)} \to -\infty \neq 1$ and $\frac{\beta_i^* h(\beta_0^* + \beta_i^* t)}{\gamma_i h(\gamma_0 + \gamma_i t)} \to -\infty \neq 1$ as $t \to -\infty$.

Appendix 2

In this section we present additional results of numerical studies, which are mentioned in the main paper.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience, 2006.

[2] T. S. Ferguson, *A Course in Large Sample Theory*, Chapman and Hall, London, 1996.

[3] N. Hjort and D. Pollard. Asymptotics for minimisers of convex processes. http://www.stat.yale.edu/~pollard/Papers/convex.pdf, 1993.

[4] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge, 1998.

Table 1. Summary statistics of the datasets.

	DATASET	n	p	R^2	$P(Y=1)$
1	Artif1	2000	50	0.28	0.49
2	Artif2	2000	50	0.23	0.51
3	diabetes	768	8	0.27	0.35
4	BreastCancer	683	9	0.88	0.35
5	heart-c	303	19	0.55	0.46
6	credit-a	690	38	0.56	0.44
7	adult	32561	57	0.41	0.24
8	vote	435	32	0.85	0.39
9	wdbc	569	31	0.93	0.37
10	spambase	4601	57	0.62	0.39

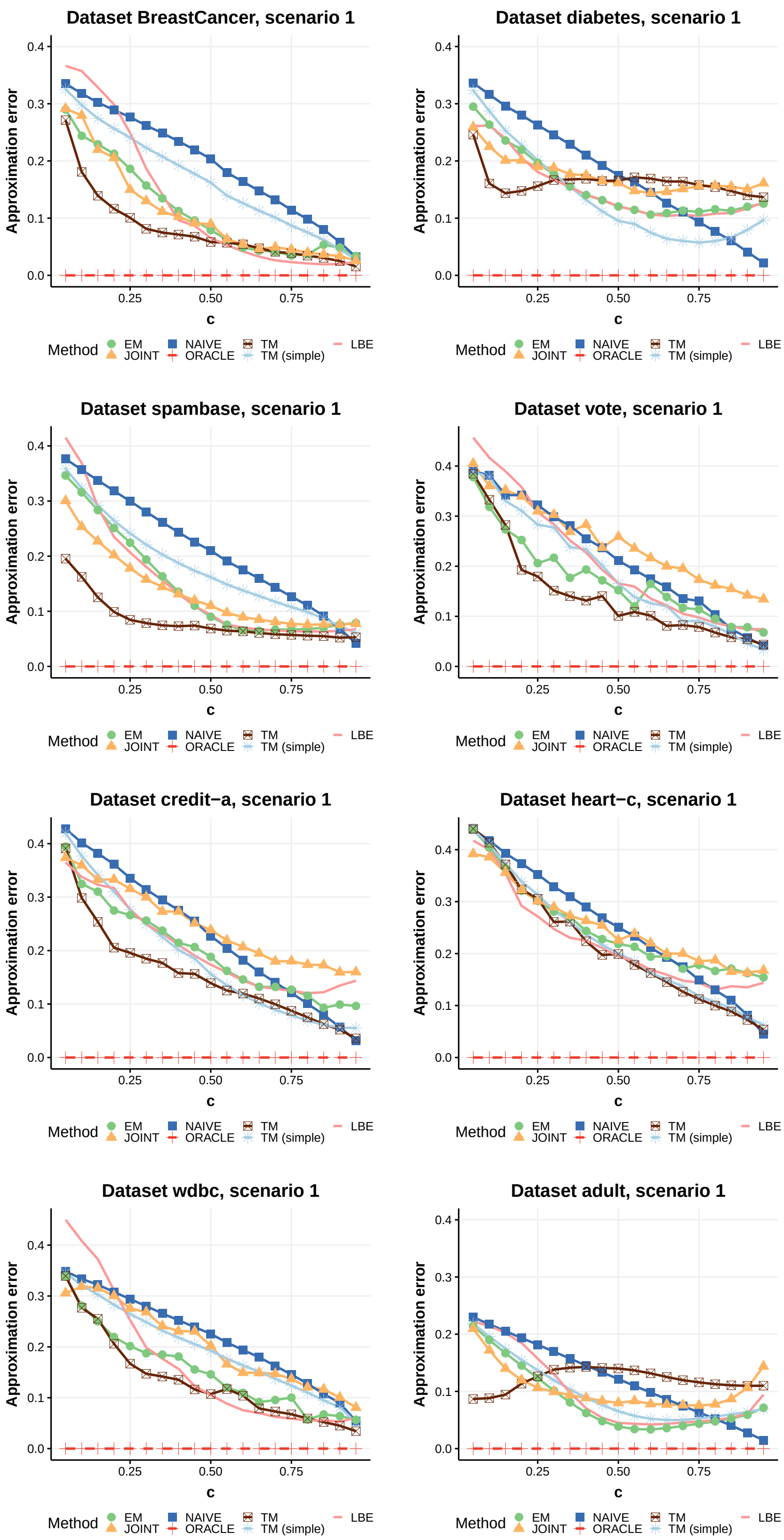

Figure 1. Approximation errors for benchmark datasets for scenario 1 and different values of c.

Table 2. Accuracy ($\pm$ standard deviation of the mean) for labelling scenario 1. The values are averaged over different $c = 0.05, 0.1, \ldots, 0.95$. The last column is p-value of the t-test. The t-test is used to verify whether the difference between the winner method (in bold) and the second best is significant.

	NAIVE	**TM simple**	**EM**	**TM**	**JOINT**	**LBE**	p-value
Artif1	0.676 ± 0.004	0.738 ± 0.003	0.784 ± 0.004	$\mathbf{0.823 \pm 0.003}$	0.786 ± 0.005	0.785 ± 0.005	<0.001
Artif2	0.656 ± 0.003	0.710 ± 0.003	0.746 ± 0.004	$\mathbf{0.785 \pm 0.004}$	0.737 ± 0.006	0.741 ± 0.005	<0.001
diabetes	0.701 ± 0.004	0.721 ± 0.004	$\mathbf{0.724 \pm 0.005}$	0.716 ± 0.005	0.706 ± 0.007	0.719 ± 0.005	0.013
BreastCancer	0.795 ± 0.006	0.840 ± 0.005	0.875 ± 0.007	$\mathbf{0.907 \pm 0.005}$	0.891 ± 0.007	0.881 ± 0.006	<0.001
heart-c	0.638 ± 0.018	0.645 ± 0.019	0.624 ± 0.019	0.654 ± 0.018	0.642 ± 0.017	$\mathbf{0.693 \pm 0.011}$	<0.001
credit-a	0.700 ± 0.012	0.743 ± 0.012	0.711 ± 0.013	$\mathbf{0.753 \pm 0.013}$	0.699 ± 0.011	0.734 ± 0.007	0.003
adult	0.790 ± 0	0.806 ± 0.001	$\mathbf{0.819 \pm 0.001}$	0.797 ± 0.001	0.805 ± 0.002	0.813 ± 0.001	<0.001
vote	0.770 ± 0.023	0.782 ± 0.023	0.787 ± 0.023	$\mathbf{0.821 \pm 0.019}$	0.762 ± 0.016	0.804 ± 0.011	<0.001
wdbc	0.773 ± 0.007	0.812 ± 0.006	0.825 ± 0.011	0.839 ± 0.01	0.810 ± 0.011	$\mathbf{0.860 \pm 0.008}$	<0.001
spambase	0.725 ± 0.007	0.770 ± 0.008	0.814 ± 0.007	$\mathbf{0.847 \pm 0.012}$	0.833 ± 0.003	0.818 ± 0.005	<0.001
avg. rank	6.7	4.8	4.2	**2.9**	5.0	3.4	

Table 3. Approximation error ($\pm$ standard deviation of the mean) for labelling scenario 1. The values are averaged over different $c = 0.05, 0.1, \ldots, 0.95$. The last column is p-value of the t-test. The t-test is used to verify whether the difference between the winner method (in bold) and the second best is significant.

	NAIVE	**TM simple**	**EM**	**TM**	**JOINT**	**LBE**	p-value
Artif1	0.287 ± 0.002	0.242 ± 0.002	0.187 ± 0.004	$\mathbf{0.145 \pm 0.003}$	0.186 ± 0.005	0.183 ± 0.004	<0.001
Artif2	0.298 ± 0.002	0.261 ± 0.002	0.208 ± 0.004	$\mathbf{0.163 \pm 0.003}$	0.218 ± 0.006	0.211 ± 0.005	<0.001
diabetes	0.178 ± 0.003	$\mathbf{0.137 \pm 0.003}$	0.156 ± 0.005	0.163 ± 0.006	0.174 ± 0.008	0.150 ± 0.005	0.0145
BreastCancer	0.194 ± 0.003	0.165 ± 0.003	0.113 ± 0.007	$\mathbf{0.079 \pm 0.004}$	0.109 ± 0.007	0.128 ± 0.007	<0.001
heart-c	0.250 ± 0.006	0.219 ± 0.005	0.246 ± 0.009	$\mathbf{0.212 \pm 0.008}$	0.252 ± 0.012	0.221 ± 0.01	0.4188
credit-a	0.229 ± 0.003	0.184 ± 0.004	0.199 ± 0.006	$\mathbf{0.154 \pm 0.006}$	0.247 ± 0.009	0.204 ± 0.006	<0.001
adult	0.122 ± 0	0.097 ± 0	$\mathbf{0.084 \pm 0.001}$	0.120 ± 0.001	0.105 ± 0.003	0.098 ± 0.001	<0.001
vote	0.217 ± 0.009	0.189 ± 0.008	0.174 ± 0.017	$\mathbf{0.143 \pm 0.012}$	0.249 ± 0.013	0.210 ± 0.009	<0.001
wdbc	0.218 ± 0.003	0.195 ± 0.003	0.153 ± 0.01	$\mathbf{0.133 \pm 0.008}$	0.203 ± 0.01	0.165 ± 0.007	<0.001
spambase	0.211 ± 0.001	0.178 ± 0.001	0.145 ± 0.002	$\mathbf{0.082 \pm 0.002}$	0.135 ± 0.002	0.147 ± 0.004	<0.001
avg. rank	6.7	4.4	3.7	**2.7**	5.3	4.2	

Table 4. Accuracy (± standard deviation of the mean) for labelling scenario 2. The values are averaged over different $g = 0.1, 0.2, \ldots, 1$. The last column is p-value of the t-test. The t-test is used to verify whether the difference between the winner method (in bold) and the second best is significant.

	NAIVE	**TM simple**	**EM**	**TM**	**JOINT**	**LBE**	p-value
Artif1	0.740 ± 0.004	0.829 ± 0.003	0.842 ± 0.004	**0.858 ± 0.003**	0.822 ± 0.005	0.844 ± 0.004	<0.001
Artif2	0.709 ± 0.004	0.794 ± 0.004	0.796 ± 0.004	**0.817 ± 0.003**	0.772 ± 0.005	0.797 ± 0.004	<0.001
diabetes	0.722 ± 0.003	**0.751 ± 0.004**	0.745 ± 0.004	0.729 ± 0.004	0.721 ± 0.008	0.745 ± 0.004	<0.001
BreastCancer	0.915 ± 0.004	0.942 ± 0.003	0.935 ± 0.008	0.940 ± 0.003	0.944 ± 0.003	**0.959 ± 0.002**	<0.001
heart-c	0.644 ± 0.02	0.657 ± 0.022	0.630 ± 0.02	0.662 ± 0.022	0.647 ± 0.02	**0.710 ± 0.012**	<0.001
credit-a	0.731 ± 0.01	**0.801 ± 0.01**	0.755 ± 0.012	0.795 ± 0.009	0.738 ± 0.011	0.773 ± 0.006	0.0177
adult	0.806 ± 0.001	0.832 ± 0	**0.841 ± 0.001**	0.803 ± 0.001	0.815 ± 0.003	0.840 ± 0.001	<0.001
vote	0.789 ± 0.025	0.808 ± 0.024	0.779 ± 0.026	0.855 ± 0.02	0.782 ± 0.016	**0.864 ± 0.013**	0.044
wdbc	0.891 ± 0.006	0.914 ± 0.005	0.886 ± 0.011	0.905 ± 0.006	0.892 ± 0.007	**0.945 ± 0.004**	<0.001
spambase	0.772 ± 0.008	0.836 ± 0.01	0.869 ± 0.011	0.868 ± 0.011	0.865 ± 0.002	**0.881 ± 0.002**	<0.001
avg. rank	**6.4**	3.9	4.8	3.8	5.4	2.5	

Table 5. Approximation errors (± standard deviation of the mean) for labelling scenario 2. The values are averaged over different $g = 0.1, 0.2, \ldots, 1$. The last column is p-value of the t-test. The t-test is used to verify whether the difference between the winner method (in bold) and the second best is significant.

	NAIVE	**TM simple**	**EM**	**TM**	**JOINT**	**LBE**	p-value
Artif1	0.230 ± 0.002	0.174 ± 0.002	0.128 ± 0.003	**0.114 ± 0.002**	0.154 ± 0.004	0.155 ± 0.003	<0.001
Artif2	0.244 ± 0.002	0.190 ± 0.002	0.153 ± 0.004	**0.126 ± 0.002**	0.183 ± 0.005	0.175 ± 0.004	<0.001
diabetes	0.149 ± 0.003	**0.078 ± 0.002**	0.120 ± 0.004	0.144 ± 0.005	0.154 ± 0.009	0.117 ± 0.004	0.0066
BreastCancer	0.091 ± 0.003	0.067 ± 0.003	0.053 ± 0.008	0.048 ± 0.003	0.055 ± 0.005	**0.036 ± 0.003**	<0.001
heart-c	0.243 ± 0.006	0.196 ± 0.006	0.228 ± 0.011	**0.194 ± 0.008**	0.255 ± 0.015	0.198 ± 0.012	<0.001
credit-a	0.194 ± 0.004	0.132 ± 0.006	0.154 ± 0.005	**0.131 ± 0.004**	0.207 ± 0.01	0.162 ± 0.005	<0.001
adult	0.103 ± 0	0.049 ± 0	**0.038 ± 0.001**	0.114 ± 0.001	0.083 ± 0.004	0.046 ± 0.001	<0.001
vote	0.209 ± 0.014	0.160 ± 0.01	0.137 ± 0.015	**0.100 ± 0.011**	0.234 ± 0.014	0.165 ± 0.011	<0.001
wdbc	0.106 ± 0.003	0.090 ± 0.003	0.088 ± 0.011	0.073 ± 0.003	0.125 ± 0.007	**0.069 ± 0.003**	<0.001
spambase	0.161 ± 0.001	0.107 ± 0.001	**0.079 ± 0.002**	0.082 ± 0.002	0.107 ± 0.002	0.084 ± 0.002	<0.001
avg. rank	6.4	4.4	3.4	**3.1**	6.0	3.7	

Table 6. Accuracy ($\pm$ standard deviation of the mean) for scenario 3, $k = 5$, $p^- = 0.2$ and $p^+ = 0.6$. The last column is p-value of the t-test. The t-test is used to verify whether the difference between the winner method (in bold) and the second best is significant.

	NAIVE	**TM simple**	**EM**	**TM**	**JOINT**	**LBE**	p-value
Artif1	0.664 ± 0.005	0.777 ± 0.005	0.830 ± 0.005	$\mathbf{0.857 \pm 0.003}$	0.816 ± 0.005	0.831 ± 0.005	<0.001
Artif2	0.643 ± 0.005	0.743 ± 0.003	0.780 ± 0.005	$\mathbf{0.805 \pm 0.004}$	0.762 ± 0.007	0.773 ± 0.005	<0.001
diabetes	0.682 ± 0.004	0.726 ± 0.005	0.732 ± 0.005	0.714 ± 0.005	0.710 ± 0.007	$\mathbf{0.733 \pm 0.006}$	0.486
BreastCancer	0.814 ± 0.008	0.877 ± 0.007	0.903 ± 0.005	0.908 ± 0.005	0.905 ± 0.006	$\mathbf{0.913 \pm 0.007}$	0.0024
heart-c	0.636 ± 0.014	0.654 ± 0.013	0.636 ± 0.017	0.679 ± 0.014	0.622 ± 0.02	$\mathbf{0.693 \pm 0.01}$	<0.001
credit-a	0.624 ± 0.01	0.688 ± 0.014	0.691 ± 0.012	$\mathbf{0.765 \pm 0.011}$	0.676 ± 0.011	0.722 ± 0.006	<0.001
adult	0.767 ± 0.001	0.795 ± 0.001	$\mathbf{0.828 \pm 0.001}$	0.778 ± 0.001	0.809 ± 0.002	0.827 ± 0.001	<0.001
vote	0.735 ± 0.024	0.750 ± 0.026	0.757 ± 0.028	$\mathbf{0.837 \pm 0.018}$	0.733 ± 0.019	0.817 ± 0.02	<0.001
wdbc	0.766 ± 0.009	0.825 ± 0.009	0.838 ± 0.008	0.852 ± 0.01	0.806 ± 0.011	$\mathbf{0.895 \pm 0.01}$	<0.001
spambase	0.633 ± 0.009	0.648 ± 0.009	0.690 ± 0.011	$\mathbf{0.810 \pm 0.02}$	0.775 ± 0.005	0.738 ± 0.008	<0.001
avg. rank	6.8	5.2	4.0	3.0	5.3	**2.8**	

Table 7. Approximation error ($\pm$ standard deviation of the mean) for scenario 3, $k = 5$, $p^- = 0.2$ and $p^+ = 0.6$. The last column is p-value of the t-test. The t-test is used to verify whether the difference between the winner method (in bold) and the second best is significant.

	NAIVE	**TM simple**	**EM**	**TM**	**JOINT**	**LBE**	p-value
Artif1	0.292 ± 0.003	0.227 ± 0.002	0.147 ± 0.004	$\mathbf{0.118 \pm 0.002}$	0.159 ± 0.005	0.166 ± 0.004	<0.001
Artif2	0.300 ± 0.002	0.243 ± 0.002	0.177 ± 0.005	$\mathbf{0.139 \pm 0.003}$	0.194 ± 0.007	0.194 ± 0.005	<0.001
diabetes	0.206 ± 0.003	$\mathbf{0.128 \pm 0.004}$	0.144 ± 0.004	0.162 ± 0.004	0.171 ± 0.007	0.140 ± 0.004	<0.001
BreastCancer	0.180 ± 0.003	0.139 ± 0.002	0.093 ± 0.005	$\mathbf{0.083 \pm 0.004}$	0.098 ± 0.007	0.094 ± 0.008	<0.001
heart-c	0.270 ± 0.007	0.222 ± 0.006	0.249 ± 0.01	$\mathbf{0.219 \pm 0.009}$	0.262 ± 0.014	0.223 ± 0.01	0.0077
credit-a	0.287 ± 0.004	0.218 ± 0.005	0.229 ± 0.007	$\mathbf{0.173 \pm 0.005}$	0.283 ± 0.008	0.218 ± 0.005	0.0185
adult	0.150 ± 0	0.093 ± 0.001	$\mathbf{0.071 \pm 0.001}$	0.134 ± 0.001	0.091 ± 0.003	0.079 ± 0.001	<0.001
vote	0.272 ± 0.011	0.228 ± 0.01	0.197 ± 0.021	$\mathbf{0.148 \pm 0.017}$	0.275 ± 0.014	0.192 ± 0.012	0.3601
wdbc	0.219 ± 0.004	0.186 ± 0.004	0.147 ± 0.006	$\mathbf{0.117 \pm 0.007}$	0.197 ± 0.008	0.127 ± 0.008	<0.001
spambase	0.308 ± 0.001	0.250 ± 0.001	0.237 ± 0.002	$\mathbf{0.088 \pm 0.002}$	0.194 ± 0.003	0.223 ± 0.005	<0.001
avg. rank	6.9	4.8	3.8	**2.7**	5.2	3.7	