International Journal of Research and Development in Engineering Sciences
ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/
Regular Issue, Vol. 8, Issue. 6 (November – December) , 2026 , Pages: 9 – 19
Received: 02 March 2026 ; *Accepted:* 31 October 2026 ; *Published:* 10 November 2026
https://doi.org/10.63328/IJRDES-V8RI6P2
https://www.youtube.com/watch?v=m4hvaL2qO1Q
https://archive.org/details/jsp-ijrdes-v8ri6p2
Original Paper : Classification Research – Problem-Solution Research Paper

Robust Threshold Selection for Cost-Sensitive Classification under Changing Prevalence and Limited Validation Data

Mahender Reddy Sheri [1] , Vijayalakshmi Kakulapati [2]

[1] Otto-Friedrich-Universität Bamberg , Bamberg, Frankfurt , Germany
[2] Department of Computer Science and Engineering , MVSR Engineering College, Hyderabad, Telangana, India
vldms@yahoo.com

* *Corresponding Author : Mahender Reddy Sheri ; mahendersheri@gmail.com*

Abstract: A classifier produces a score, but a decision needs a threshold, and the threshold depends on the cost of each kind of error, on how common the positive class is, and on how much labelled data is available to set it. We study these three dependences in a binormal model, where every quantity has a closed form. The optimal threshold is $\tau^* = d/2 - \ln \varrho/d$, with ϱ the cost-weighted prior odds. The excess cost of a threshold error $\Delta\tau$ is $\frac{1}{2}\, dA\, \Delta\tau^2$ to second order. The default rule that predicts the positive class when the posterior exceeds 0.5 is off by exactly $\ln r/d$ in score units, where r is the ratio of the two error costs, whatever the prevalence. A change of prevalence from π to π' moves the optimum by $-[\text{logit}\,\pi' - \text{logit}\,\pi]/d$. A threshold chosen by minimising the empirical cost on n labelled points has excess cost of order $n^{-2/3}$, against n^{-1} for a plug-in estimate that uses the Gaussian model. In simulation the excess-cost formulas matched within 0.3 cost units per 1,000 cases for the cost curve and 0.004 for the default rule, and fitted slopes were -0.69 and -1.01. When prevalence fell from 0.3 to 0.01, keeping the old threshold cost 5.4 times as much as always predicting the negative class. On the breast cancer and digits datasets, a shift-corrected Bayes threshold was best or near best in all four settings, and tuning a threshold on up to 160 labelled points did not beat it. Synthetic results are simulations; the real data are two small benchmark sets with resampled prevalence.

Keywords: Cost-Sensitive Learning, Decision Threshold, Class Imbalance, Prior Shift, Calibration.

1. Introduction

Most classifiers output a score. Turning that score into a decision needs one more number, the threshold, and the number is rarely chosen with care. The usual choice is the score at which the estimated posterior probability is one half. That choice is optimal only when both kinds of error cost the same and when the class frequencies at deployment match those at training [1, 2]. A fraud detector, a tumour screen and a spam filter all violate at least one of these.

Three questions arise. If the threshold is a little wrong, how much does that cost? When the prevalence of the positive class changes between training and deployment, how far should the threshold move, and what happens if it does not? And when the threshold is tuned on a labelled validation set, how many labelled points are needed? Each has a standard answer in words (use the cost ratio, correct for the prior [3], tune on validation data), but the sizes of the effects are seldom stated, and the sample-size question is complicated by the fact that the empirical cost of a threshold is a step function, whose minimiser converges at an unusual rate [4].

We answer these questions in the binormal model of signal detection [6], in which scores of the two classes are Gaussian with equal variance, and then test the answers on two real datasets. We make six contributions.

- The optimal threshold and a local expression for the excess cost of a threshold error, with a measured check on its range of validity (Section 4).
- The observation that the default threshold is off by exactly $\ln r/d$, independent of prevalence, and the resulting excess cost.
- The shift of the optimal threshold under a change of prevalence and the cost of ignoring it, including the

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P2

prevalence below which keeping the old threshold is worse than a trivial classifier.

- Measured convergence rates for threshold estimation from n labelled points: slopes -0.69 for the empirical minimiser and -1.01 for a plug-in rule, against −2/3 and −1 from theory, and the sample sizes that reach 5% relative excess cost (Section 6).
- The cost of assuming the wrong cost ratio, and a check that the equal-variance results survive unequal variances.
- A real-data comparison of five threshold rules on the breast cancer and digits datasets under natural and shifted prevalence.

All synthetic experiments are simulations from known models. The real-data experiment uses two small datasets with prevalence changed by resampling.

2. Background and Related Work

Cost-sensitive decisions. For known class-conditional densities, the cost-minimising rule is a likelihood-ratio test [2], which for a calibrated probability is a threshold of $c_{FP}/(c_{FP}+c_{FN})$ [1]. Provost and Fawcett study how to choose classifiers when costs and class distributions are uncertain [16], and Zadrozny and colleagues show how to learn cost-sensitively by reweighting examples [11]. Resampling and synthetic oversampling are common ways to handle imbalance [10, 12]; we do not use them, and instead keep the model fixed and move the threshold.

Prior shift and calibration. Saerens and colleagues showed how to adjust posteriors when class priors change [3], and Lipton and colleagues extended this to estimating the shift from unlabelled data [9]. Posteriors from many models are poorly calibrated [8, 17], which is usually repaired by a monotone map such as Platt scaling [7]. Threshold choice is related to ROC analysis, where an operating point is picked on the curve [5, 16].

What we add. The ingredients are known. We contribute closed forms for the binormal case that give sizes, not directions, of the effects: the quadratic excess cost, the prevalence-independent gap of the default rule, the cost of ignoring a prior shift, and measured rates and sample sizes for threshold estimation.

3. Research Method and Gap

3.1. Score model and Cost

A score Z is observed for each case with label $Y \in \{0, 1\}$. In the binormal model, $Z \mid Y = 0 \sim N(0, 1)$ and $Z \mid Y = 1 \sim N(d, 1)$, where d is the separation. The positive class has prevalence π. A false positive costs $c_{FP} = 1$ and a false negative costs $c_{FN} = r$, so r is the cost ratio. Predicting positive when $Z > \tau$ gives expected cost per case

$$C(\tau) = \pi\, r\, \Phi(\tau - d) + (1 - \pi)[1 - \Phi(\tau)]. \qquad (1)$$

The trivial classifiers cost $\min(\pi r,\ 1 - \pi)$ per case. A classifier is useful only if its cost is below this. For a model output that is the log-odds of a calibrated posterior, the same analysis applies with score differences measured in log-odds units, and the default rule is $Z > 0$ at the training prevalence.

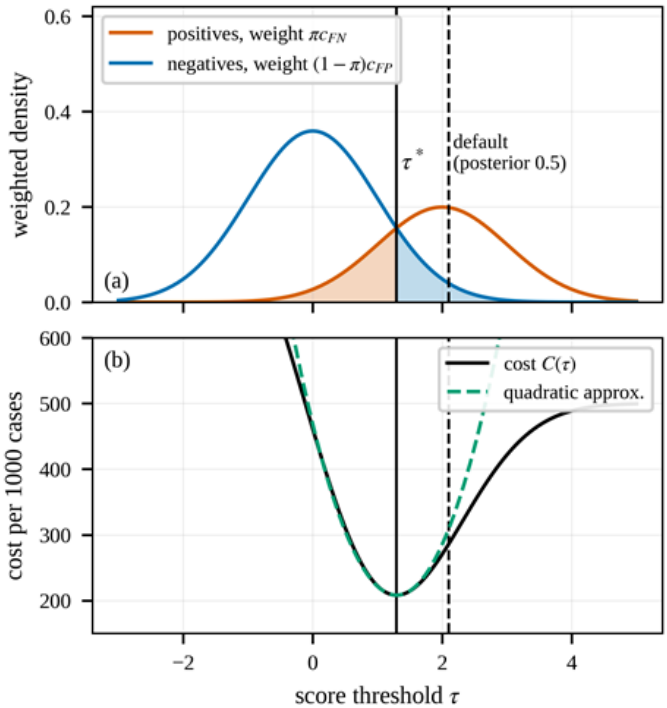


Figure 1: The binormal model with $d = 2$, $\pi = 0.1$, $r = 5$. (a) Cost-weighted class densities; the optimal threshold τ^* is where they cross, and the default threshold (posterior 0.5) lies to its right. (b) Expected cost $C(\tau)$ and its quadratic approximation at τ^*.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P2

3.2 Threshold rules

We compare five rules. **Default:** $z > 0$, that is posterior above one half. **Bayes:** $z > -\ln r$, the cost-sensitive threshold for a posterior calibrated at the training prevalence π_{tr}. **Shift-corrected Bayes:** $z > -\ln r - [\text{logit } \pi' - \text{logit } \pi_{tr}]$, which also corrects for deployment prevalence π' (Eq. 6 below). **Empirical minimiser:** the threshold that minimises the cost on n labelled validation points, with ties broken at the middle. **Platt plus Bayes:** logistic regression of the label on the score, using n validation points, then the Bayes threshold on the recalibrated posterior; this is the plug-in rule for real data. In the synthetic study the plug-in rule instead estimates the two means, the pooled variance and the prevalence, and inserts them in the optimal threshold of Eq. (2).

4. Proposed System

Proposition 1 (optimal threshold). *The cost $C(\tau)$ in Eq. (1) is minimised at*

$$\tau^* = d/2 - \ln \varrho / d, \quad \varrho = \pi r / (1 - \pi). \tag{2}$$

This follows by setting $C'(\tau) = 0$, which gives $\pi r \varphi(\tau - d) = (1 - \pi)\varphi(\tau)$, and taking logarithms [2]. The threshold sits at the midpoint of the two class means, moved towards the less costly class by $\ln \varrho / d$.

Proposition 2 (excess cost). *Let $A = \pi r \varphi(\tau^* - d) = (1 - \pi)\varphi(\tau^*)$ and $\Delta\tau = \tau - \tau^*$. Then*

$$C(\tau) - C(\tau^*) = \tfrac{1}{2}\, d\, A\, \Delta\tau^2 + O(|\Delta\tau|^3). \tag{3}$$

Proof. At τ^* the first derivative is zero. The second derivative is $-\pi r(\tau - d)\varphi(\tau - d) + (1 - \pi)\tau\varphi(\tau)$, which at τ^* equals $A[\tau^* - (\tau^* - d)] = dA$.

Corollary 1 (the default threshold). The default rule predicts positive when the posterior exceeds one half, that is when the cost-free odds exceed one. In score units its threshold is $\tau_0 = d/2 - \ln(\pi/(1-\pi))/d$. Hence $\tau^* - \tau_0 = -\ln r/d$, which does not depend on π, and its excess cost is

$$C(\tau_0) - C(\tau^*) \approx A\, (\ln r)^2 / (2d). \tag{4}$$

The default is optimal only when $r = 1$. For a fixed cost ratio and separation, the *threshold error* is the same at every Synthetic experiments use $d = 2$ and $r = 5$ unless stated. Excess costs in the exact columns come from the formulas. Simulated values use 4 million cases for Table 1, Table 2 and Table 3. The sample-size study draws $n \in \{50, \ldots, 10{,}000\}$ labelled points with prevalence $\pi \in \{0.5, 0.1, 0.02\}$ and uses 2,000 repetitions per setting. Validation sets with fewer than two cases of either class are discarded and their share is reported. The excess cost of each estimated threshold is evaluated exactly from Eq. (1). Slopes are fitted by least squares on log–log data for $n \geq 200$ ($n \geq 500$ for $\pi = 0.02$).

prevalence, though the cost it causes scales with A, which shrinks with π.

Proposition 3 (prevalence shift). *Suppose the class-conditional score laws stay fixed and only π changes to π'. Then*

$$\tau^*(\pi') - \tau^*(\pi) = -[\text{logit } \pi' - \text{logit } \pi] / d. \tag{5}$$

Keeping the old threshold therefore costs about $A'(\Delta\text{logit})^2/(2d)$ per case, with A' evaluated at π'. In terms of a model whose log-odds output z is calibrated at the training prevalence π_{tr}, the Bayes threshold at deployment prevalence π' is

$$z^* = -\ln r - [\text{logit } \pi' - \text{logit } \pi_{tr}]. \tag{6}$$

This is the prior correction of Saerens and colleagues [3] combined with the cost threshold. The two corrections have opposite signs when $\pi' < \pi_{tr}$ and $r > 1$, so they partly cancel, which is why the default can look better than the uncorrected Bayes rule when positives become rarer.

Proposition 4 (estimating the threshold). Let $\hat{\tau}$ minimise the empirical cost on n points. The empirical cost is a step function with local fluctuations of order $n^{-1/2}$, and its minimiser satisfies $\hat{\tau} - \tau^* = O_P(n^{-1/3})$ [4]. By Eq. (3) the excess cost is of order $n^{-2/3}$. A plug-in estimate from a smooth model with estimated parameters has error $O_P(n^{-1/2})$ and excess cost of order n^{-1}.

We state these as known asymptotic rates [4] and check them by simulation; we do not prove them here.

Corollary 2 (wrong cost ratio). *If the threshold is set with an assumed cost ratio $\hat{r}$ instead of r, the threshold error is $\Delta\tau = \ln(r/\hat{r})/d$ and*

$$C(\tau) - C(\tau^*) \approx A\, [\ln(r/\hat{r})]^2 / (2d). \tag{7}$$

Corollary 1 is the case $\hat{r} = 1$. The excess cost depends on the ratio $r/\hat{r}$ only through its logarithm, so a cost ratio that is wrong by a factor of two is as harmful as one that is wrong by a factor of one half, and costs the same fraction of A/d whatever the prevalence.

5. Experimental Setup And Implementation

Real data are the Wisconsin breast cancer set (569 cases, 30 features; malignant is positive, natural prevalence 0.373) [13] and the optical digits set (1,797 cases, 64 features; digit 9 against the rest, prevalence 0.100) [14], both from scikit learn [15]. In each of 400 repetitions the data are split at random into 40% for fitting a logistic regression on standardised features, 30% as a validation pool and 30% as a test pool. Validation sets of size $n \in \{20, 40, 80, 160\}$ are drawn from the pool with the deployment prevalence, class by class with replacement. The test cost is measured on 4,000 draws from the test pool at the same prevalence, with $r = 5$. We study the natural prevalence and a shifted

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P2

one (0.05 for breast cancer, 0.01 for digits). The prior correction uses the true deployment prevalence. The reference "tuned on test" is the empirical minimiser on the test draws, an optimistic bound.

5.1. Cost of a threshold error

Table 1 compares Eq. (3) with the exact cost and with simulation for $d = 2$, $\pi = 0.1$, $r = 5$, where $\tau^* = 1.294$, $A = 0.155$ and the minimum cost is 208.1 per 1,000 cases. Simulation matches the exact value within 0.3 units per 1,000 cases, which is consistent with sampling error.

The quadratic approximation is within 15% for $|\Delta\tau| \le 0.5$ but overstates the cost by 42% at $\Delta\tau = 1$, because the cost curve is asymmetric (Figure 2). A threshold that is too low by one unit costs 159.9 units, and one that is too high by one unit costs 109.5, 46% less. Too low a threshold accepts many of the abundant negatives.

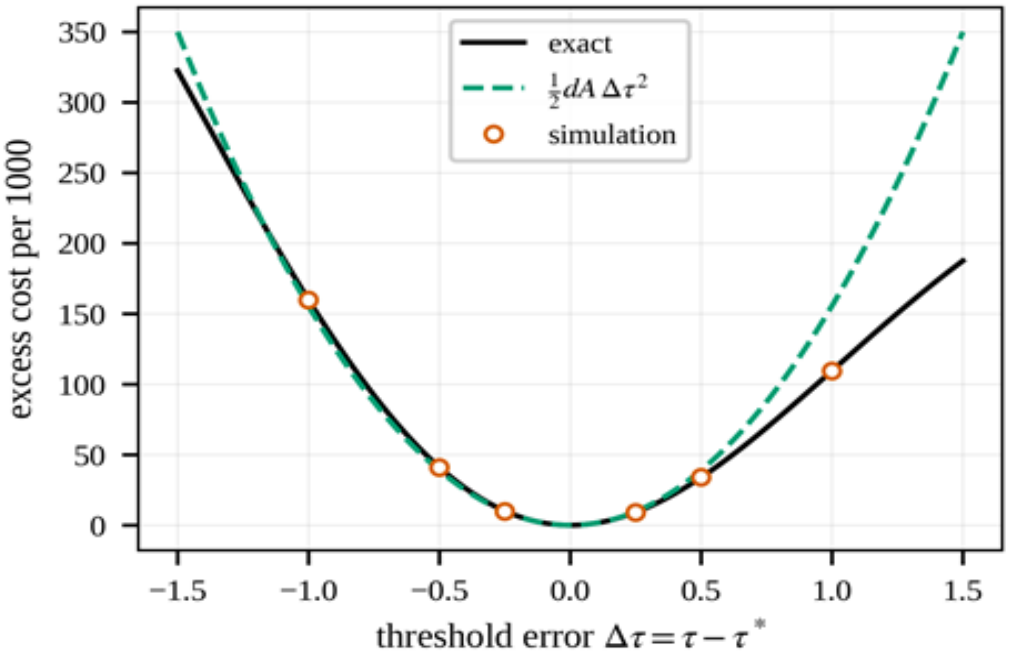


Figure 2: Excess cost against threshold error for $d = 2$, $\pi = 0.1$, $r = 5$. Black: exact. Dashed: quadratic approximation, Eq. (3). Circles: simulation with 4 million cases.

Table 1: Excess cost per 1,000 cases for a threshold error $\Delta\tau$ ($d = 2$, $\pi = 0.1$, $r = 5$).

$\Delta\tau$	Exact	Quadratic	Simulated	Quad./exact
-1.00	159.9	155.5	159.7	0.97
-0.50	41.1	38.9	41.1	0.95
-0.25	10.1	9.7	9.9	0.96
0.25	9.2	9.7	9.3	1.06
0.50	33.9	38.9	34.1	1.15
1.00	109.5	155.5	109.5	1.42

5.2. The default threshold

Table 2 checks Corollary 1. The gap between the optimal and default thresholds is $-\ln r/d$ in every row, and the relative excess cost of the default rule, from the exact formula, agrees with simulation within 0.004. At $d = 2$ and $r = 5$ the default costs 37% more than the optimum at $\pi = 0.1$ and 18% more at $\pi = 0.02$. For $d = 3$ and $r = 10$ it costs 73% more at $\pi = 0.05$. Figure 3 shows that the relative excess grows with the cost ratio and with prevalence.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P2

Table 2: Relative excess cost of the default threshold, theory against simulation (4 million cases). The gap is $\tau^* - \tau_0$.

d	r	π	Gap	Theory	Simulated
2.0	5.0	0.10	-0.80	0.373	0.377
2.0	5.0	0.02	-0.80	0.181	0.182
3.0	10.0	0.05	-0.77	0.733	0.733
1.5	2.0	0.30	-0.46	0.098	0.099

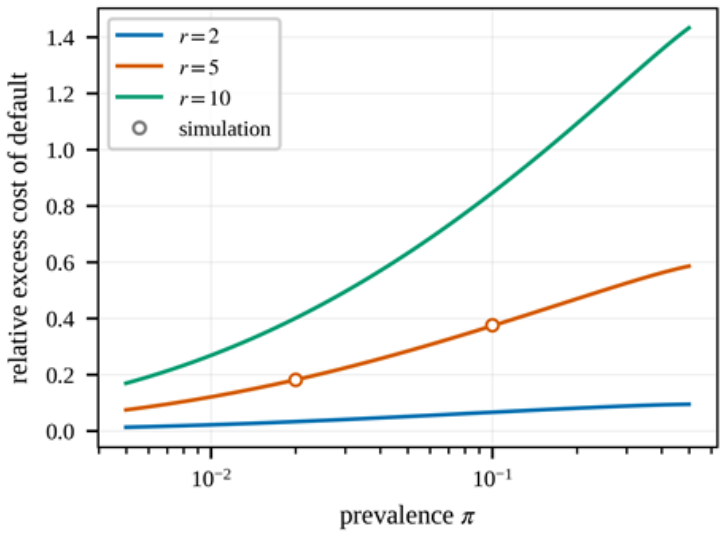


Figure 3: Relative excess cost of the default threshold (posterior 0.5) for $d = 2$ as prevalence and cost ratio vary. Lines: exact. Circles: simulation.

5.3. A Wrong Cost Ratio

Corollary 2 is checked in Figure 8. The relative excess cost is zero at $\hat{r} = r = 5$ and grows on both sides in the same way in log $\hat{r}$. Assuming equal costs ($\hat{r} = 1$) when $r = 5$ costs 37% more than the optimum at $d = 2$ and $\pi = 0.1$, and assuming $\hat{r} = 20$, a factor of four too high, costs 38% more. Eq. (7) predicts similar costs for errors of similar size in log $\hat{r}$, and a factor of five too low is slightly worse than a factor of four too high. Simulation agrees within 0.003. The practical message is that cost ratios need to be right in order of magnitude, not exactly: for $d = 3$, assuming $\hat{r} = 2$ (a factor of 2.5 too low) costs 13%

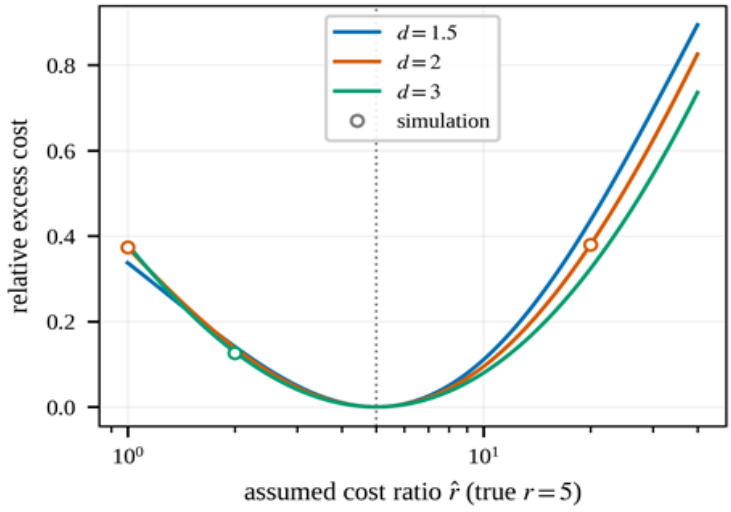


Figure 8: Relative excess cost when the cost ratio is wrongly assumed to be $\hat{r}$ (true $r = 5$, $\pi = 0.1$). Lines: exact. Circles: simulation. The minimum is at $\hat{r} = r$.

5.4. Prevalence Shift

Table 3 starts from a threshold tuned at $\pi = 0.3$ and moves prevalence to π'. The optimal threshold moves by $-\Delta\text{logit}/d$ (Eq. 5), from -0.42 units at $\pi' = 0.5$ to +1.87 at $\pi' = 0.01$. The regret of keeping the old threshold agrees with simulation

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P2

within 1.0 units per 1,000 cases. The relative regret is 36% at $\pi' = 0.1$ and 562% at $\pi' = 0.01$. At $\pi' = 0.01$ the old threshold costs 269.5 per 1,000 cases, 5.4 times the 50.0 of always predicting negative, while a re-set threshold costs 40.7 (Figure 4). A threshold that is not updated can therefore be worse than no classifier at all.

Table 3: Prevalence shift from $\pi = 0.3$ to π' ($d = 2$, $r = 5$). Shift of the optimal threshold, regret of keeping the old threshold (exact and simulated) and cost per 1,000 cases when kept, re-set, and for the trivial classifier.

π'	Shift	Regret exact	Regret sim.	Cost kept	Cost re-set	Trivial
0.5	-0.42	42.9	43.3	343.1	300.2	500.0
0.3	+0.00	0.0	0.0	313.0	313.0	700.0
0.1	+0.67	74.9	75.4	283.0	208.1	500.0
0.03	+1.31	175.7	176.7	272.5	96.8	150.0
0.01	+1.87	228.8	229.1	269.5	40.7	50.0

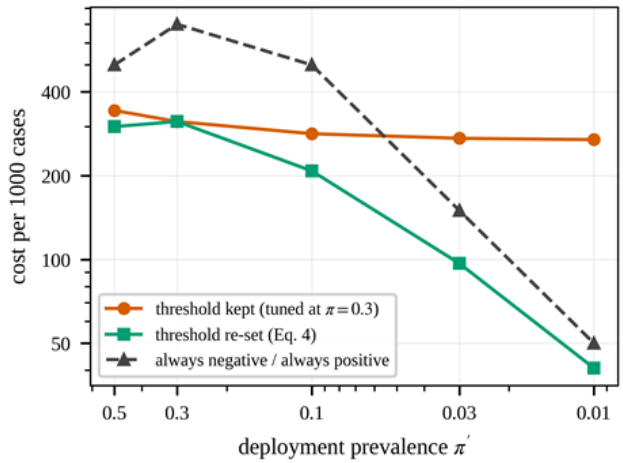


Figure 4: Cost per 1,000 cases after prevalence moves from 0.3 to π' ($d = 2$, $r = 5$). Keeping the old threshold becomes worse than the trivial classifier once prevalence falls enough.

5.5. Estimating The Threshold From Labelled Data

Figure 5 and Table 4 show how the excess cost of an estimated threshold falls with n. The fitted log–log slopes of the empirical minimiser are -0.69, -0.69 and -0.68 for $\pi = 0.5$, 0.1 and 0.02, close to the rate −2/3 of Proposition 4. For the plug-in rule they are -0.99, -1.01 and -1.08, close to −1. With $\pi = 0.1$ the empirical minimiser needs about 332 labelled points to be within 5% of the optimal cost, and the plug-in rule about 77. At $n = 1{,}000$ the relative excess cost is 2.4% for the empirical minimiser and 0.4% for the plug-in rule.

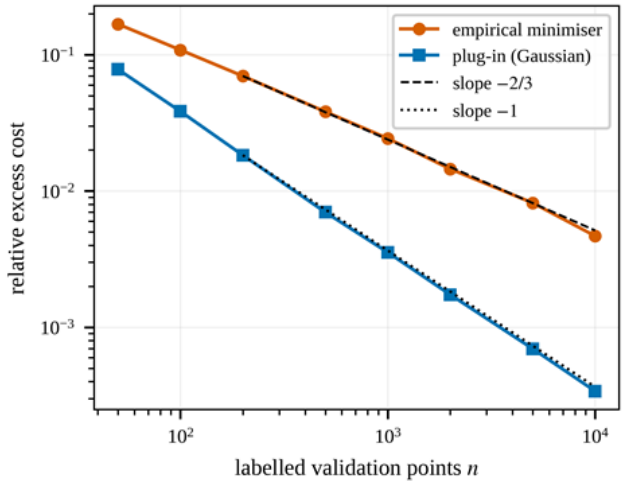


Figure 5: Relative excess cost of two ways to estimate the threshold from n labelled points ($\pi = 0.1$, $d = 2$, $r = 5$, 2,000 repetitions). Dashed lines show slopes −2/3 and −1.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P2

Rarity makes both worse. At $\pi = 0.02$ the empirical minimiser needs about 784 points to reach 5%, against 216 at $\pi = 0.5$ (Figure 6), and many small validation sets contain almost no positives: 73% of sets with $n = 50$ and 40% with $n = 100$ had fewer than two positive cases. The plug-in rule is better at every size here, but it relies on the Gaussian model being right; on real data we use a calibration-based version.

Table 4: Threshold estimation from n labelled points: fitted slopes, relative excess cost at $n = 1{,}000$ and n needed for 5% relative excess cost.

π	Slope emp.	Slope plug-in	Rel. excess, n=1,000 emp. (%)	Rel. excess, n=1,000 plug-in (%)	n for 5% emp.	n for 5% plug-in
0.5	-0.69	-0.99	1.7	0.2	216	53
0.1	-0.69	-1.01	2.4	0.4	332	77
0.02	-0.68	-1.08	4.2	1.0	784	193

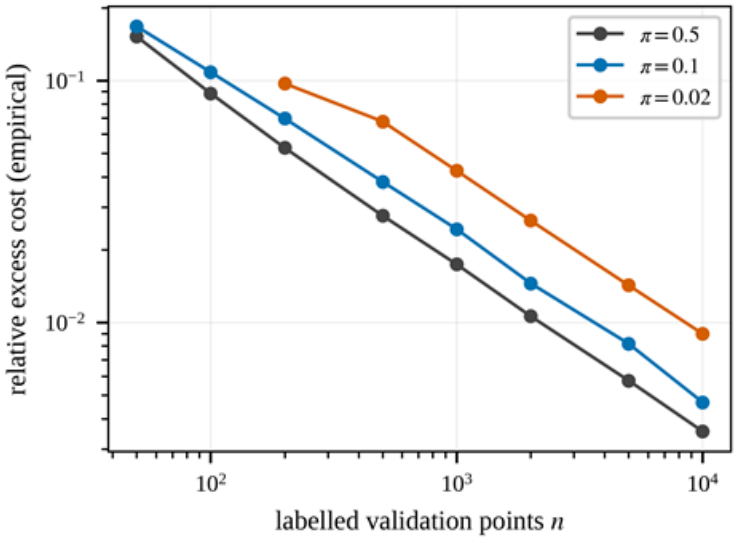


Figure 6: Relative excess cost of the empirical minimiser for three prevalences. Settings where more than 10% of validation sets lack two cases of a class are omitted.

5.6. Unequal Variances

The equal-variance model underlies Eqs. (2) to (5). To test how much depends on it, we let the positive-class scores have standard deviation $s_1 \neq 1$ ($d = 2$, $\pi = 0.1$, $r = 5$), find the exact optimal threshold numerically, and compare it with the equal-variance rule that uses the midpoint and pooled variance (Table 7). The rule is still close to optimal for moderate departures: the relative excess cost is 0.0% at $s_1 =$ 1.5 and 0.5% at $s_1 = 2$, and grows to 6.8% at $s_1 = 3$. The default rule remains far worse (44% to 73% excess), so the main message survives; but at large variance ratios the optimal rule is no longer a single threshold shifted by $\ln \varrho/d$, and the formulas of Section 4 should be taken as local guides.

Table 7: Unequal variances ($d = 2$, $\pi = 0.1$, $r = 5$): exact optimal threshold, equal-variance threshold, and relative excess cost (%) of the equal-variance rule and of the default rule. s_1 is the standard deviation of the positive-class scores.

s_1	Optimal τ	Equal-var. τ	Excess (%) eq.-var.	Excess (%) Default
1.00	1.29	1.29	0.0	37.3
1.25	1.37	1.38	0.0	39.4
1.50	1.46	1.48	0.0	44.0
2.00	1.61	1.73	0.5	55.8
3.00	1.84	2.47	6.8	73.4

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P2

5.7. Real Data

Table 5 gives the test cost at n = 160. At natural prevalence the Bayes threshold from the cost ratio, with no tuning, beats the default on both datasets (91.6 against 110.2 per 1,000 on breast cancer, 51.4 against 73.5 on digits). The two data-driven rules do not beat it, even with 160 labelled points: the empirical minimiser costs 100.7 and 63.5, and Platt plus Bayes 92.6 and 55.0. With 20 points the empirical minimiser is much worse (152.4 and 84.1).

Table 5: Real data, test cost per 1,000 cases at n = 160 validation points (r = 5; 400 repetitions). "Tuned on test" is an optimistic bound.

Setting	Default	Bayes	Shift-corr.	Empirical	Platt	Tuned on test
Breast, π'=0.37	110.2	91.6	92.1	100.7	92.6	59.9
Breast, π'=0.05	25.5	74.1	23.7	32.5	27.6	15.5
Digits, π'=0.10	73.5	51.4	51.2	63.5	55.0	39.8
Digits, π'=0.01	13.8	33.7	13.1	21.6	19.7	9.9

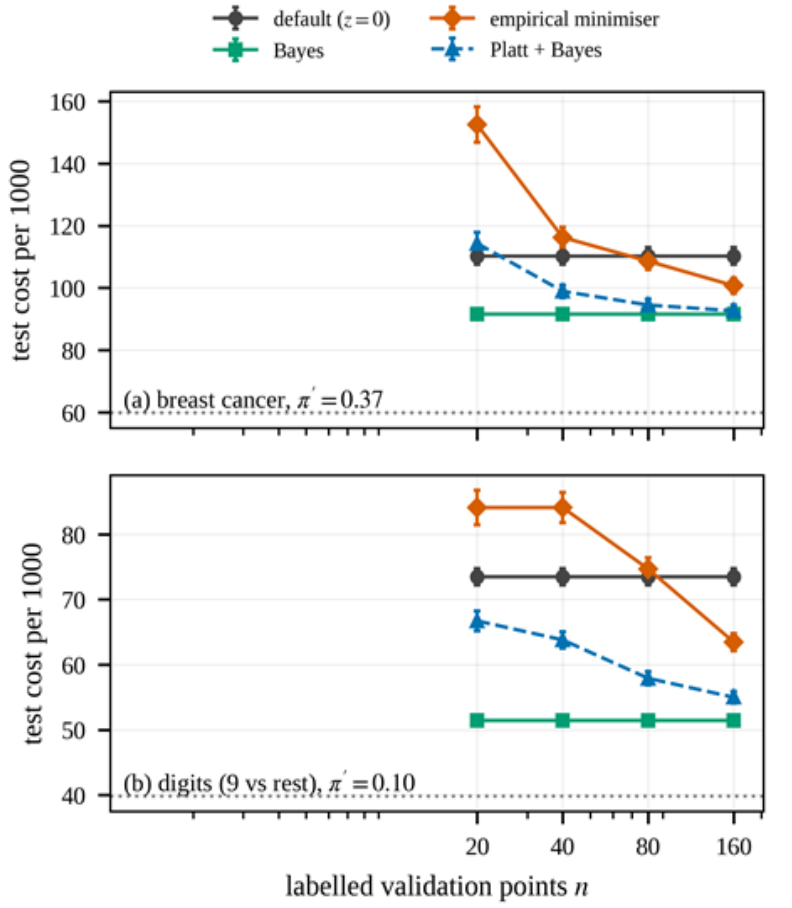


Figure 9: Real data at natural prevalence (400 repetitions, ±1 SE). Test cost per 1,000 cases against the number of validation points. Dotted: threshold tuned on the test draws (optimistic bound).

Under prevalence shift the picture follows Proposition 3. On breast cancer at π' = 0.05, the Bayes threshold without prior correction costs 74.1, 2.9 times the default (25.5); the prior-corrected rule costs 23.7. On digits at π' = 0.01 the numbers are 33.7, 13.8 and 13.1. The default does well here only by partial cancellation. The model was calibrated at prevalence 0.37, so the prior correction on breast cancer is -2.42 in log-odds and the cost correction is -1.61. The correct threshold is z^* = 0.81; the default z = 0 is 0.81 away from it, and the uncorrected Bayes threshold -1.61 is 2.42 away.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P2

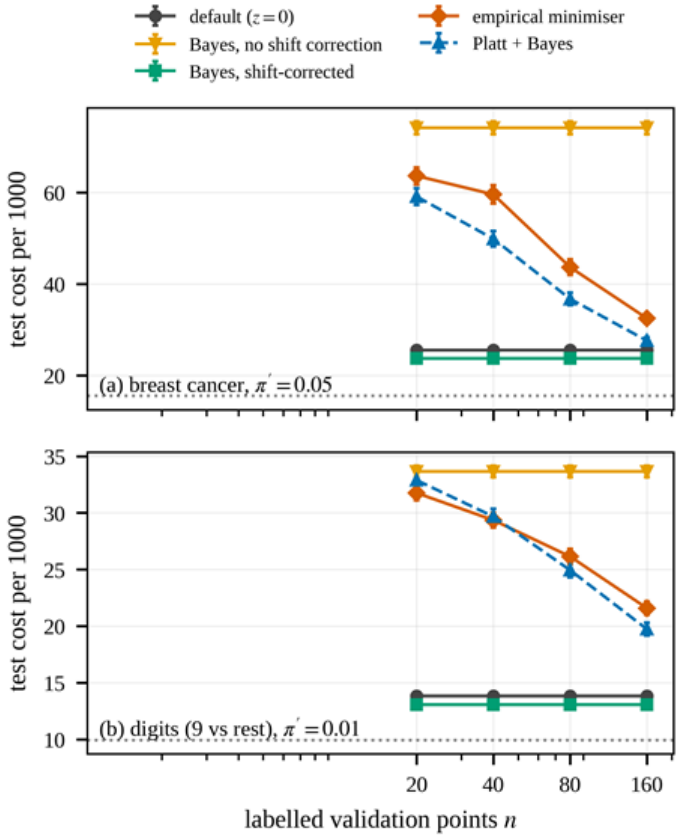


Figure 7: Real data under prevalence shift (400 repetitions, ±1 SE). Test cost per 1,000 cases against the number of validation points used by the data-driven methods. Dotted: threshold tuned on the test draws (optimistic bound).

The data-driven rules improve with n but remain above the corrected Bayes rule at n = 160: 32.5 (empirical) and 27.6 (Platt) against 23.7 on breast cancer, and 21.6 and 19.7 against 13.1 on digits. This agrees with Table 4: with a rare positive class, 160 validation points contain a handful of positives. The comparison favours the corrected rule partly because it is given the true prevalence and cost ratio.

6.8 A practical recipe

Table 6: The threshold rules compared. Numbers are from the real-data study at n = 160, breast cancer, π' = 0.05.

Rule	Needs	Fails by	Cost
Default	nothing	ignores r and π'	25.5
Bayes	cost ratio	ignores prior shift	74.1
Shift-corrected	cost ratio, π'	wrong π' or r	23.7
Tune on data	labelled set	noise at small n	32.5

The results suggest a short procedure. First, set the threshold from the costs on calibrated scores (Bayes rule). Second, correct it for the deployment prevalence with Eq. (6); this step changed the cost by a factor of 2.6 to 3.1 in the shifted real-data settings and is the one most often forgotten. Third, check whether the classifier beats the trivial one at deployment prevalence, using $\min(\pi' r, 1 - \pi')$. Fourth, tune on labelled data only if it is plentiful: Table 4 gives the order of magnitude, hundreds to thousands of points for a 5% loss, and more when positives are rare. If tuning, a smooth plug-in or calibration-based estimate needs fewer points than minimising the empirical cost. Fifth, remember that the quadratic cost of Eq. (3) is symmetric only locally; err towards a higher threshold when negatives are abundant.

Discussion and Limitations

What we think carries over. Three points seem general. The cost of a threshold error is quadratic near the optimum, so small errors are cheap and large ones are not. The default threshold embodies two assumptions, equal costs and unchanged prevalence, and the two errors partly cancel in rare-positive settings, so a default that looks fine can be wrong for the right reasons. And a step-function objective makes threshold tuning data-hungry: the cube-root rate means that the first thousand labels buy much

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P2

less than the next thousand would suggest.

Synthetic setting. The binormal model with equal variance is a convenient case, not a typical one. Real score distributions are skewed, have unequal variances and may be multimodal; Corollary 1 and Eq. (5) depend on the equal-variance form, which gives log-odds that are linear in the score. The rates in Proposition 4 are asymptotic and we did not prove them. Costs were constant and known.

Real data. We used two small benchmark datasets and a logistic regression. Validation and test draws were taken with replacement from pools of a few hundred cases, so repeated cases appear, and the "tuned on test" bound is optimistic. Prevalence shift was created by resampling classes, which satisfies the prior-shift assumption by construction. The corrected rule used the true prevalence and cost ratio; estimated values would add error. We did not compare smoothed or cross-validated threshold estimators, threshold-free metrics, or models trained with cost weights.

6. Conclusion

A threshold is a decision about costs, prevalence and data. In the binormal model the optimum has a closed form, the excess cost of an error is quadratic, the default rule is off by $\ln r/d$ whatever the prevalence, and a change of prevalence moves the optimum by the change in log-odds divided by d. In our experiments ignoring a prevalence shift multiplied cost by up to 6.6 in simulation and by up to 3.1 on real data, while tuning on up to 160 labelled points did not beat a corrected rule from known costs. Next steps are to test these results on skewed and unequal-variance scores, to estimate the prevalence from unlabelled data, and to combine a corrected Bayes threshold with a small amount of tuning.

References

[1]. C. Elkan. The foundations of cost-sensitive learning. In *IJCAI*, 2001.

[2]. H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. Wiley, 1968.

[3]. M. Saerens, P. Latinne, C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002.

[4]. J. Kim, D. Pollard. Cube root asymptotics. *Ann. Statist.*, 18(1):191–219, 1990.

[5]. T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[6]. D. M. Green, J. A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, 1966.

[7]. J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, MIT Press, 1999.

[8]. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[9]. Z. C. Lipton, Y.-X. Wang, A. Smola. Detecting and correcting for label shift with black box predictors. In *ICML*, 2018.

[10]. H. He, E. A. Garcia. Learning from imbalanced data. *IEEE Trans. Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[11]. B. Zadrozny, J. Langford, N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *ICDM*, 2003.

[12]. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.

[13]. W. N. Street, W. H. Wolberg, O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *SPIE*, vol. 1905, 1993.

[14]. E. Alpaydin, C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998.

[15]. F. Pedregosa et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[16]. F. Provost, T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.

[17]. A. Niculescu-Mizil, R. Caruana. Predicting good probabilities with supervised learning. In *ICML*, 2005.

How To Cite:

Mahender Reddy Sheri , Vijayalakshmi Kakulapati, " Robust Threshold Selection for Cost-Sensitive Classification under Changing Prevalence and Limited Validation Data ", International Journal of Research and Development in Engineering Sciences(IJRDES), Volume 8 , Issue 6 , pp 9-19, November 2026.

CrossRef DOI: https://doi.org/10.63328/ijrdes-v8ri6p2

A. Derivations

Eq. (2). $C'(\tau) = \pi r\varphi(\tau - d) - (1 - \pi)\varphi(\tau) = 0$ gives $\varphi(\tau - d)/\varphi(\tau) = (1 - \pi)/(\pi r)$. The left side is $\exp(d\tau - d^2/2)$, so $d\tau - d^2/2 = -\ln \varrho$, which is Eq. (2).

Eq. (4). Substituting $\Delta\tau = -\ln r/d$ into Eq. (3) gives $\tfrac{1}{2}\, dA\, (\ln r)^2/d^2 = A\, (\ln r)^2/(2d)$.

Eq. (5). Differentiating Eq. (2) with respect to logit π, with $\ln \varrho = \text{logit}\, \pi + \ln r$, gives $-1/d$.

Eq. (6). A posterior calibrated at π_{tr} has log-odds $z = d(s - d/2) + \text{logit}\, \pi_{tr}$ for score s in the binormal model. The threshold $-\ln r$ at the deployment prior adds the term $-[\text{logit}\, \pi' - \text{logit}\, \pi_{tr}]$, by the prior correction of [3].

Eq. (7). The threshold with assumed ratio $\hat{r}$ is τ^* with $\ln \varrho$ replaced by $\ln(\pi\hat{r}/(1-\pi))$, so it differs from τ^* by $\ln(r/\hat{r})/d$; Eq. (3) then gives Eq. (7).

B. Parameters

Synthetic: $d = 2$, $r = 5$, $c_{FP} = 1$; $\pi = 0.1$ for Table 1, Tables 2 and 3 as stated; 4 million cases per simulated cost; sample-size study with $n \in \{50, 100, 200, 500, 1{,}000, 2{,}000, 5{,}000,$

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P2

10,000}, $\pi \in \{0.5, 0.1, 0.02\}$, 2,000 repetitions; plug-in uses sample means, pooled variance and sample prevalence. Real data: logistic regression ($C = 1$) on standardised features, split 40/30/30, 400 repetitions, validation sizes 20 to 160, 4,000 test draws; Platt step is a logistic regression ($C = 1$) of the label on the score, falling back to the Bayes threshold when the slope is below 0.05 or one class is absent; empirical minimiser falls back to the Bayes threshold when one class is absent. All figures were produced by code at 600 dpi and are printed at their stored width. Code, results and figures are in the supplementary archive.

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com