

Adaptive Dependence-Weighted Naïve Bayes for Robust and Calibrated Probabilistic Classification

Journal Title
XX(X):1–19
©The Author(s) 2026
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

SAGE

Isaac Touza^{1,2}, Didier Alain Njomen Njamen¹ and Kaladzavi Guidedi^{2,3}

Abstract

The naive Bayes classifier is valued for its simplicity, efficiency, and interpretability, but its conditional independence assumption causes redundant features to be counted repeatedly, leading to overconfident predictions and degraded accuracy. We propose *Adaptive Bayesian Dependence Naive Bayes* (ABD-NB), a method that preserves the standard naive Bayes factorisation while making the independence assumption adaptive through feature-specific weights $w_i \in (0, 1]$. These weights decrease according to each feature's estimated within-class dependence on the others. The proposed weighting scheme combines three components: (i) class-conditional dependence estimation using rank correlation (or alternative dependence measures) with asymptotic soft-thresholding for variance control; (ii) a harmonic graph-based weighting rule that provably assigns equal fractional weights to duplicated features, recovering the exact Bayes posterior under duplication; and (iii) spectral rescaling based on the eigenvalues of the dependence matrix to match the effective number of independent features. The correction parameter γ is selected by internal validation and includes $\gamma = 0$, ensuring graceful degradation to standard naive Bayes. We establish theoretical guarantees, including duplicate recovery, weight consistency, and provable reduction of naive Bayes overconfidence under correlated features. Experiments on 15 real and synthetic datasets against 10 baseline classifiers show that ABD-NB improves accuracy by up to 17.3 percentage points in highly redundant settings, significantly reduces log-loss (Wilcoxon $p = 0.008$), achieves a higher average accuracy rank than random forests and logistic regression, and never performs more than 0.1 accuracy points below naive Bayes. A dynamic extension further adapts to evolving dependence structures in data streams while maintaining superior predictive performance.

Keywords

naive Bayes, conditional independence, dependence-adaptive weighting, effective dimension, probability calibration, concept drift

Introduction

Few learning algorithms combine simplicity, speed, and empirical reliability as effectively as the naive Bayes (NB) classifier. Given a feature vector $X = (X_1, \dots, X_d)$ and a class variable Y , NB estimates the posterior under the *conditional independence* assumption

$$P(X_1, \dots, X_d | Y) = \prod_{i=1}^d P(X_i | Y), \quad (1)$$

so that classification reduces to d one-dimensional density estimates per class. This factorisation yields $O(nd)$ training, natural incremental updates, graceful behaviour in high dimensions and small samples, and transparent per-feature evidence contributions—properties that keep NB in production use across text categorisation, medical triage, spam filtering, and embedded systems (Lewis 1998; McCallum and Nigam 1998; Rish 2001; Hand and Yu 2001), while recent ontology-driven and hybrid approaches further demonstrate the continued relevance of combining statistical learning with semantic knowledge for automatic text classification (Touza et al. 2025).

Assumption (1) is, however, almost always false. Anthropometric variables such as weight, height, and body-mass index are deterministically entangled; neighbouring pixels

co-vary; engineered feature sets routinely contain near-duplicates. When (1) fails, NB *multiplies the same evidence several times*. The consequences are well documented (Domingos and Pazzani 1997; Hand and Yu 2001): posterior probabilities become grossly overconfident even when the decision remains correct, and in less benign configurations—typically when a large group of mutually redundant features competes with a smaller amount of independent evidence—the redundant group out-votes the reliable evidence and the *decisions themselves* become suboptimal, with an excess risk that does not vanish as $n \rightarrow \infty$.

Two broad families of remedies exist. *Structure-extending* methods—tree-augmented naive Bayes (TAN), k -dependence estimators, averaged one-dependence estimators

¹Department of Mathematics-Computer Science, Faculty of Science, University of Maroua, Maroua, Cameroon

²Laboratoire de Recherche en Informatique (LaRI), University of Maroua, Maroua, Cameroon

³Department of Computer Science and Telecommunications, National Advanced School of Engineering, University of Maroua, Cameroon

Corresponding author:

Isaac Touza, Department of Mathematics-Computer Science, Faculty of Science, University of Maroua, Maroua, Cameroon.

Email: isaac.touza@univ-maroua.cm

(AODE), and general Bayesian network classifiers—enrich the graphical structure to represent dependence explicitly (Friedman et al. 1997; Sahami 1996; Webb et al. 2005). They can be markedly more accurate, but they abandon the factorised architecture: training requires structure search (NP-hard in general, Chickering 1996), incremental updating becomes awkward, and per-feature interpretability is lost. *Feature-weighting* methods instead retain the NB architecture and attach a weight to each feature, but nearly all existing schemes weight by *relevance to the label*—mutual information with Y , decision-tree depth, gain ratio, or discriminatively optimised objectives (Hall 2007; Lee et al. 2011; Zaidi et al. 2013; Jiang et al. 2016, 2019). Relevance weighting addresses a different failure mode: it can silence noisy features, but it does not see, let alone correct, the double-counting of *mutually dependent* features, which is the actual violation of (1).

Our proposal

This paper develops *Adaptive Bayesian Dependence Naive Bayes* (ABD-NB), a classifier that keeps the naive factorisation and makes assumption (1) adaptive rather than binary. The starting point is the observation that independence violations are graded: some features are nearly conditionally independent of the rest, others are heavily entangled. ABD-NB estimates, for every feature (and optionally every class), *how much* the independence assumption is violated, and converts that estimate into an exponent on the feature’s likelihood contribution:

$$P(Y = c \mid x) \propto \pi_c \prod_{i=1}^d P(x_i \mid c)^{w_{ic}}, \quad w_{ic} \in (0, 1]. \quad (2)$$

In log-space the model is a dependence-weighted evidence sum, $\log \pi_c + \sum_i w_{ic} \log P(x_i \mid c)$, so training and prediction retain the cost and the incremental structure of NB.

The scientific content of the paper lies in *how* the weights are constructed and *what can be guaranteed* about the result. Our construction rests on four statistical design decisions, each of which we motivate, analyse, and validate in isolation:

1. **The dependence that matters is class-conditional.** Marginal association between two informative features is induced by the label itself and does not violate (1). We therefore estimate pairwise dependence *within* classes (per class, and pooled across classes with prior weights), a distinction that turns out to be essential in practice: weighting by marginal dependence systematically punishes exactly the informative features (Section).
2. **Redundancy accumulates on the shared-information scale.** Summing raw pairwise dependences over-counts redundancy: many moderate correlations do not carry the same information overlap as one near-duplicate. Guided by the Gaussian identity $MI = -\frac{1}{2} \log(1 - \rho^2) \approx \rho^2/2$, the redundancy degree of a feature is $\sum_{j \neq i} \hat{D}_{ij}^2$, and the harmonic rule $w_i = 1/(1 + \gamma \sum_{j \neq i} \hat{D}_{ij}^2)$ assigns weight exactly $1/r$ to each of r duplicated features. We prove that this choice recovers the exact Bayes posterior under duplication (Theorem 1).

3. **The total evidence mass is pinned to the effective dimension.** Down-weighting alone deflates the likelihood scale and lets the prior dominate. We rescale the weight vector so that $\sum_i w_i = M_{\text{eff}}$, the effective number of independent features computed from the eigenvalues of the dependence matrix in the spirit of the effective-number-of-tests literature (Cheverud 2001; Nyholt 2004; Li and Ji 2005). The rescaling factor is provably 1 under exact duplication, so Theorem 1 is preserved.
4. **The correction strength is selected, not imposed.** A single scalar $\gamma \geq 0$ interpolates between plain NB ($\gamma = 0$) and aggressive decorrelation. γ (and the choice between global and class-specific weights) is selected by an internal cross-validated criterion whose grid contains $\gamma = 0$; consequently ABD-NB inherits a *never-worse* safety property (Proposition 4) that we verify empirically on a dataset (Digits) where dependence weighting is genuinely harmful.

Contributions

The contributions of this paper are:

- **A framework.** A dependence-adaptive weighting framework for naive Bayes that is agnostic to the dependence measure (Spearman, Pearson, Kendall, mutual information, distance correlation, HSIC, Cramér’s V are implemented and compared), supports global, class-specific, and pairwise-graph weights, and includes a dynamic variant for non-stationary streams (Section).
- **Theory.** Degeneracy to NB at $\gamma = 0$ (Proposition 1); exact posterior recovery under feature duplication (Theorem 1); properties of the spectral effective dimension (Lemma 1); strong consistency and $\sqrt{n}$ asymptotic normality of the weight estimator with $O(1/n)$ mean-squared error (Theorem 2, Proposition 3); an explicit $1 + (d - 1)\rho$ overconfidence factor for NB under equicorrelation together with the correction achieved by ABD-NB (Proposition 2); and the asymptotic never-worse guarantee of the tuned classifier (Proposition 4). All results are verified numerically (Section).
- **Evidence.** A pre-registered-style evaluation on 15 datasets (4 real UCI benchmarks, 2 redundancy-augmented real datasets, 9 controlled synthetic families) against 10 baselines including relevance-weighted NB, logistic regression, SVM, and random forests, with Friedman, Nemenyi, and Holm-corrected Wilcoxon analyses (Demšar 2006): up to +17.3 accuracy points under adversarial redundancy, a 10/15-dataset log-loss improvement over NB ($p = 0.008$), an average accuracy rank above random forests and logistic regression, and verified robustness (worst regression versus NB across all 15 datasets: 0.05 accuracy points). A concept-drift study demonstrates that the dynamic variant re-estimates the dependence graph within ~ 500 samples after an abrupt structural change (Section).

- **Reproducibility.** A documented open-source implementation (scikit-learn compatible) with scripts that regenerate every number, table, and figure of the paper.

Organisation

Section reviews related work. Section formalises the double-counting pathology. Section develops the ABD-NB framework. Section states and proves the theoretical guarantees. Section describes the experimental protocol, Section reports the results, Section discusses scope and limitations, and Section concludes.

Related Work

The paradox of naive Bayes

The empirical success of NB despite its false premise has been analysed since the 1990s. Domingos and Pazzani (1997) showed that NB is optimal under zero-one loss for a strictly larger class of problems than the independence assumption suggests, because correct classification only requires the correct *ranking* of posteriors, not their values. Hand and Yu (2001) sharpened the picture: dependence damages NB precisely when it is *asymmetric*—when the amount of double-counted evidence differs across regions of the feature space or across feature groups—and they highlight redundant measurement clusters as the canonical failure case. Rish (2001) mapped the accuracy of NB against the entropy of the feature distribution and found the largest degradation for functionally dependent features. Our conflict-block construction (Section) is a distilled version of exactly this pathology, and our theory quantifies both faces of it: decision distortion (Theorem 1 and its corollary) and calibration distortion (Proposition 2).

Structure-extending Bayesian classifiers

Semi-naive methods relax (1) by enriching the dependency graph. Kononenko (1991) merged dependent attributes into joint nodes. TAN (Friedman et al. 1997) allows each feature one parent besides the class, learned via a maximum-spanning tree on conditional mutual information; k -dependence Bayesian classifiers (Sahami 1996) generalise to k parents; AODE (Webb et al. 2005) averages over all one-dependence estimators, and its descendants weight or select super-parents (Webb et al. 2012). These methods share three costs: parameter tables grow exponentially with the in-degree (requiring discretisation and larger samples), incremental maintenance of the structure is difficult, and the model no longer decomposes into per-feature evidence. ABD-NB takes the complementary route: the structure stays naive and only the *exponents* of the factors change; the price is that dependence is corrected on average rather than modelled, and the benefit is that every operational property of NB survives, including $O(d^2)$ -per-sample online updates (Section).

Feature-weighted naive Bayes

Attribute weighting has a long history in the NB literature; see Zaidi et al. (2013) for a survey. Filter approaches weight by label relevance: gain ratio, Kullback–Leibler divergence between class-conditional and marginal

distributions (Lee et al. 2011), depth in a decision tree (Hall 2007), or deep-feature statistics for text (Jiang et al. 2016). Wrapper approaches optimise the weights discriminatively—WANBIA (Zaidi et al. 2013) maximises conditional log-likelihood, and class-specific attribute weighting (Jiang et al. 2019) learns one weight per attribute–class pair. Locally weighted NB (?) reweights *instances* instead. Closest in spirit to us, correlation-based weight adjustment (Yu et al. 2020) multiplies a relevance term by a heuristic redundancy discount. Recent ontology-driven and hybrid approaches to text classification further illustrate the use of explicit semantic knowledge alongside statistical and transformer-based learning (Touza et al. 2025).

Our contribution differs from this literature in three ways. First, the weighting target is different and statistically explicit: ABD-NB estimates the violation of *conditional* independence—pairwise dependence measured within classes—rather than relevance to the label, and we show that the marginal/conditional distinction is not a nicety but a necessity (Section). Second, the functional form is derived from a recovery requirement (weight $1/r$ under r -fold duplication) instead of being an ad-hoc discount, and it comes with a spectral rescaling that preserves the total evidence mass; both ingredients are needed for the theory in Section . Third, the correction strength is selected with a safeguard that provably contains plain NB, which none of the cited schemes provides. Empirically, the relevance-weighted representative (MI-weighted NB) and ABD-NB fail and succeed on *different* datasets, confirming that they correct different pathologies (Section).

Dependence measures and the effective number of tests

We treat the dependence measure as a plug-in. Rank correlations (Spearman 1904; Kendall 1938; Kruskal 1958) are invariant to monotone transformations and robust to outliers; mutual information (Cover and Thomas 2006) and bias-corrected Cramér’s V (Cramér 1946; Bergsma 2013) capture general association on discretised data; distance correlation (Székely et al. 2007) and HSIC (Gretton et al. 2005) characterise independence against arbitrary (non-monotone) alternatives. Section compares all seven options inside ABD-NB. The spectral rescaling of Section adapts the effective-number-of-tests construction of Cheverud (2001), Nyholt (2004), and Li and Ji (2005)—developed for multiple-testing correction in genetics—to a new purpose: budgeting the total exponent mass of a weighted likelihood.

Calibration

Well-calibrated class probabilities are increasingly required of deployed classifiers (Guo et al. 2017; Gneiting and Raftery 2007). NB is a notorious offender: ? place it among the worst-calibrated standard models, pushing predicted probabilities towards 0 and 1, and post-hoc recalibration (Platt scaling, isotonic regression, binning) is the usual fix (Platt 1999; Zadrozny and Elkan 2002; Pakdaman Naeini et al. 2015). ABD-NB attacks the cause rather than the symptom: overconfidence arises because the likelihood exponent effectively counts d pieces of evidence when only $M_{\text{eff}} < d$ are independent, and the weights restore the correct

evidence budget *inside* the model, before any recalibration. The two approaches are complementary; unlike post-hoc maps, the internal correction also changes (and can improve) the decisions themselves, as the conflict experiments show.

Learning under concept drift

Streaming classifiers must track non-stationarity (Widmer and Kubat 1996; Gama et al. 2014; Bifet et al. 2010). NB is a popular base learner for streams because its sufficient statistics update in $O(d)$; but a stream’s *dependence structure* can drift even when the marginals are stable, silently changing how much evidence is double-counted. To our knowledge, ABD-NB is the first NB variant whose independence correction is itself dynamic: exponentially weighted dependence statistics let the weights track structural change (Section), which the drift experiment isolates.

Background and Problem Formulation

Setting and notation

Let $(X, Y) \sim P$ with $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ and $Y \in \{1, \dots, K\}$, class priors $\pi_c = \mathbb{P}(Y = c)$, and class-conditional densities $p(x | c)$. A training set $\mathcal{D} = \{(x^{(t)}, y^{(t)})\}_{t=1}^n$ is drawn i.i.d. from P . Gaussian naive Bayes (GNB) estimates $\hat{\pi}_c$, $\hat{\mu}_{ic}$, $\hat{\sigma}_{ic}^2$ and scores

$$S_c^{\text{NB}}(x) = \log \hat{\pi}_c + \sum_{i=1}^d \log \phi(x_i; \hat{\mu}_{ic}, \hat{\sigma}_{ic}^2), \quad (3)$$

where ϕ is the normal density; the predicted label is $\arg \max_c S_c^{\text{NB}}(x)$, and the posterior estimate follows by normalising $\exp S_c^{\text{NB}}$. Throughout, “dependence” refers to statistical association between *features conditionally on Y*; $\|\cdot\|$ is the Euclidean norm.

The double-counting pathology

Two elementary calculations show what dependence does to (3); both are formalised in Section .

Decision distortion. Suppose feature X_1 appears r times (exact copies $X_1 = X_{d+1} = \dots = X_{d+r-1}$, a limiting model of redundant measurements). Each copy contributes the identical term $\log p(x_1 | c)$ to (3), so the block’s evidence is multiplied by r . If the duplicated feature is weakly informative and competes with a strong independent feature, an arbitrarily large r makes the weak evidence dominate: the NB decision converges to the decision of the weak feature alone, and the excess risk over Bayes does not vanish with n . This is not a small-sample artefact but a bias of the population-level NB rule (Hand and Yu 2001).

Calibration distortion. Even when the decision survives (e.g. by symmetry), the posterior does not. Under class-conditional equicorrelation ρ the NB log-posterior-odds equal $1 + (d-1)\rho$ times the true log-odds (Proposition 2): with $d = 12$ and $\rho = 0.9$ the model reports odds on a scale *eleven times* too large. Downstream consumers of the probabilities—cost-sensitive decisions, triage thresholds, ensembling—inherit this distortion.

Design requirements

These two calculations dictate what a within-architecture correction must do. It must (R1) estimate the strength of conditional dependence per feature; (R2) discount groups of mutually dependent features so that a group of r duplicates counts once in total; (R3) preserve the overall evidence scale so the prior–likelihood balance is not destroyed; and (R4) never do damage when (1) happens to hold, or when the dependence pattern is one that helps NB. Sections – implement R1–R4 in turn.

The ABD-NB Framework

Figure 1 previews the pipeline’s raw material on the Breast Cancer dataset: the estimated pairwise dependence matrices, marginal and within-class, whose structure the weights will summarise. The framework has four stages: dependence estimation (Section), weight construction (Section), spectral rescaling (Section), and adaptive selection (Section); Section adds the dynamic variant and Section the complexity analysis.

Stage 1: class-conditional dependence estimation

Definition 1. Dependence matrix. A dependence measure is a functional D that maps a pair of random variables to $[0, 1]$, with $D = 0$ under independence and $D = 1$ under almost-sure functional dependence. For a sample $Z \in \mathbb{R}^{m \times d}$, $\hat{D}(Z) \in [0, 1]^{d \times d}$ denotes the matrix of pairwise estimates with zero diagonal.

Seven measures are implemented: absolute Pearson and Spearman correlation, Kendall’s τ_b , normalised mutual information and bias-corrected Cramér’s V on quantile-discretised data (Bergsma 2013), distance correlation (Székely et al. 2007), and normalised HSIC with Gaussian kernels and median-heuristic bandwidths (Gretton et al. 2005); the latter two are computed on subsamples for tractability. Spearman is the default: it is $O(d^2 n + dn \log n)$, invariant under monotone reparametrisations of the features (so the weights do not change if a feature is logged), robust to outliers, and its estimator is an asymptotically normal U-statistic, which Section exploits.

The estimand is conditional, not marginal. The pitfall in requirement R1 is subtle: assumption (1) concerns dependence *given Y*. Marginally, two features that both respond to the class are correlated through it—in Iris, the two petal dimensions have marginal $|\rho| \approx 0.96$ but within-class dependence of only ≈ 0.3 . Weighting by marginal dependence therefore penalises exactly the informative features and rewards noise. ABD-NB estimates one matrix per class,

$$\hat{D}^{(c)} = \hat{D}(\{x^{(t)} : y^{(t)} = c\}), \quad c = 1, \dots, K, \quad (4)$$

and the prior-weighted pooled matrix $\hat{D} = \sum_c \hat{\pi}_c \hat{D}^{(c)}$. Classes with fewer than a minimum count (8 by default) fall back to a matrix estimated on class-centred residuals, a shrinkage device for rare classes.

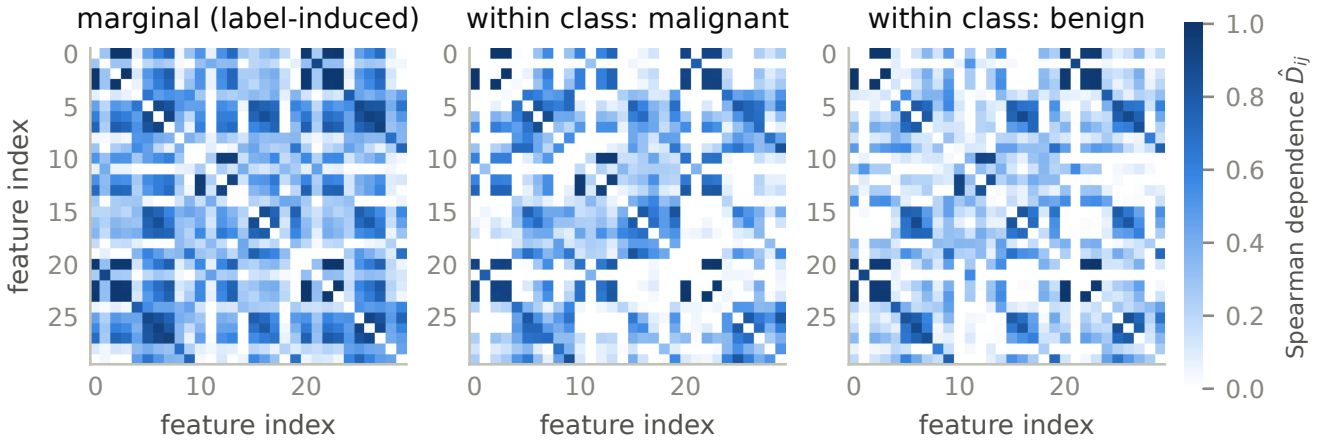


Figure 1. Pairwise Spearman dependence $\hat{D}_{ij}$ between the 30 features of the Breast Cancer Wisconsin dataset. Left: marginal dependence, inflated by label-induced association between informative features. Centre and right: within-class dependence for the malignant and benign classes—the estimand relevant to assumption (1). The visible block structure (radius/perimeter/area triplets, mean/worst duplicates) is what ABD-NB converts into weights; note also that the two classes do not share the same dependence pattern, motivating class-specific weights.

Variance control by soft-thresholding. With $d(d-1)/2$ entries estimated from n_c samples, small-sample noise in $\hat{D}$ propagates to the weights. We apply the asymptotic noise floor $t_n = z_{1-\alpha/2}/\sqrt{n-3}$ (the Fisher-scale critical value for zero correlation, Fisher 1915), soft-thresholding each entry:

$$\hat{D}_{ij} \leftarrow \max\left(0, \frac{\hat{D}_{ij} - t_n}{1 - t_n}\right). \quad (5)$$

Entries indistinguishable from zero at level α are removed and the rest are shrunk continuously, which reduces the variance of the plug-in weights at a $O(n^{-1/2})$ cost in bias that vanishes asymptotically (Theorem 2).

Stage 2: from dependence to weights

Given a dependence matrix D , define per-feature dependence degrees and weights. We study a family of four link functions (Figure 2); the first three collapse the row D_i into a scalar $\bar{D}_i$ (mean or max of the off-diagonal entries) and apply a decreasing link,

$$w_i^{\text{lin}} = 1 - \bar{D}_i, \quad w_i^{\text{exp}} = e^{-\gamma \bar{D}_i}, \quad w_i^{\text{inv}} = \frac{1}{1 + \gamma \bar{D}_i}, \quad (6)$$

while the flagship *harmonic graph rule* operates on the pairwise structure directly:

$$w_i^{\text{har}} = \frac{1}{1 + \gamma \sum_{j \neq i} D_{ij}^2}. \quad (7)$$

Two choices in (7) deserve justification.

Why squared dependence? Summing raw dependences treats twenty correlations of 0.3 as six duplicates’ worth of redundancy, which grossly over-counts: information overlap, not correlation, is what double-counting duplicates. For bivariate Gaussians the mutual information is $-\frac{1}{2} \log(1 - \rho^2) \approx \rho^2/2$ for moderate ρ , so the squared scale aggregates approximately shared *information*. The exponent leaves exact duplicates ($D_{ij} = 1$) untouched, preserving the recovery property below, while damping the contribution of moderate associations by an order of magnitude ($0.3^2 = 0.09$).

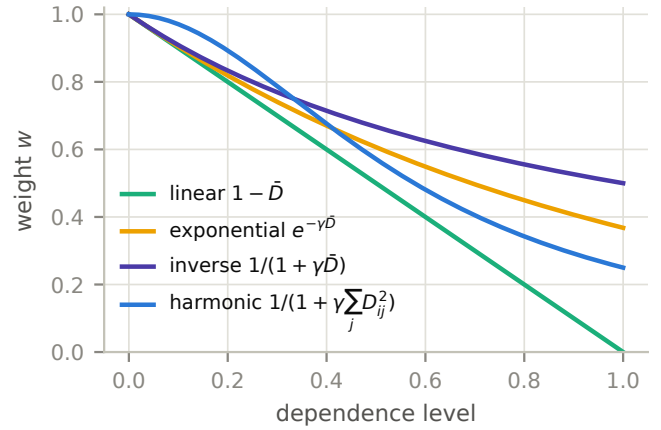


Figure 2. The four weight link functions at $\gamma = 1$. The scalar links (linear, exponential, inverse) act on the aggregated row dependence $\bar{D}_i$; the harmonic graph rule acts on the squared-dependence degree $\sum_{j \neq i} D_{ij}^2$ and is the only one with the exact $1/r$ duplicate-recovery property.

Why the harmonic form? Requirement R2 asks that a group of r interchangeable copies count once in total. Within such a group, $\sum_{j \neq i} D_{ij}^2 = r - 1$ for every member, so (7) with $\gamma = 1$ gives $w_i = 1/r$ for each copy—the unique symmetric allocation whose sum is 1. Theorem 1 shows this is not merely an appealing heuristic: the resulting classifier reproduces the *exact* Bayes posterior of the de-duplicated problem. The quantity γ acts as a correction strength: $\gamma = 0$ switches the correction off, $\gamma = 1$ is the canonical value at which the recovery property holds, and larger values decorrelate more aggressively.

Stage 3: spectral rescaling to the effective dimension

Down-weighting alone violates requirement R3: if all d weights shrink, the log-likelihood term of (2) deflates relative to the log-prior and, with class-specific weights, classes with smaller total weight are mechanically favoured because log-densities are negative. The correct total scale is dictated by

information content: the weighted evidence should be worth as much as M_{eff} independent features, where M_{eff} measures how many independent dimensions the feature set really spans.

Definition 2. Spectral effective dimension. Let $R = D + I$ (the dependence matrix with unit diagonal restored) with eigenvalues $\lambda_1 \geq \dots \geq \lambda_d \geq 0$. Following [Li and Ji \(2005\)](#),

$$M_{\text{eff}}(D) = \sum_{k=1}^d \left[\mathbf{1}\{\lambda_k \geq 1\} + (\lambda_k - \lfloor \lambda_k \rfloor) \right], \quad (8)$$

clipped to $[1, d]$.

The final weights are

$$w = \tilde{w} \cdot \frac{M_{\text{eff}}(D)}{\sum_i \tilde{w}_i}, \quad (9)$$

where $\tilde{w}$ is the output of Stage 2. Lemma 1 establishes that $M_{\text{eff}} = d$ under independence (the rescaling is then the identity for γ arbitrary, because $\tilde{w} \equiv 1$), that $M_{\text{eff}} = d - r + 1$ under r -fold duplication (the rescaling factor is then exactly 1, preserving Theorem 1), and that $1 \leq M_{\text{eff}} \leq d$ always. We write $d_{\text{eff}} = \sum_i w_i$ for the realised effective dimension; Figure 3 shows the resulting per-feature weights on Breast Cancer, where the 30 raw features collapse to $d_{\text{eff}} \approx 14.5$.

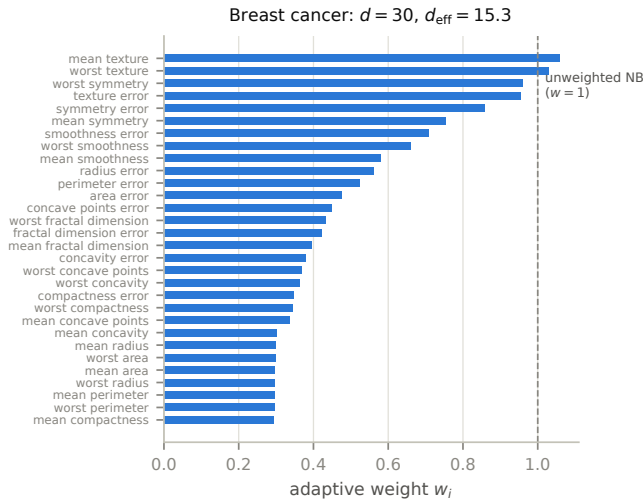


Figure 3. Adaptive feature weights on Breast Cancer. Redundant size features receive weights near 0.4, more independent shape features near or above 1, reducing 30 nominal features to $d_{\text{eff}} \approx 14.5$ effective ones.

Class-specific weights on a common evidence scale

Dependence structure can differ across classes (Figure 1), suggesting class-specific weights w_{ic} computed from $\hat{D}^{(c)}$. This flexibility carries a hidden trap that, to our knowledge, has not been documented in the attribute-weighting literature: because $\log p(x_i | c) \leq 0$ typically, a class whose weight row sums to less has its total penalty shrunk and is *mechanically* favoured, regardless of the data. ABD-NB therefore renormalises every class row to the prior-weighted common budget,

$$w_{\cdot c} \leftarrow w_{\cdot c} \frac{\sum_{c'} \hat{\pi}_{c'} \sum_i w_{ic'}}{\sum_i w_{ic}}, \quad (10)$$

so classes differ in the *shape* of their weight profile but not in total evidence mass. The ablation in Section shows what happens without safeguards: on a design with class-asymmetric dependence, naive class-specific weighting collapses to 40.7% accuracy while the safeguarded variants remain above 65%.

Stage 4: adaptive selection of the correction strength

Requirement R4 is the crux. There exist datasets—Digits is our running example—where within-class dependence is strong *among the informative features*, so any pure redundancy discount transfers evidence mass towards uninformative ones and hurts accuracy; NB’s double-counting is, on such data, a beneficial bias. No fixed γ can be simultaneously optimal for Digits and for a redundancy-poisoned dataset. ABD-NB therefore treats $(\gamma, \text{structure})$ as data-determined: an internal V -fold stratified cross-validation (default $V = 3$) evaluates the grid $\gamma \in \{0, 0.25, 0.5, 1, 2, 4\} \times \{\text{global, class-specific}\}$ and selects with a *tolerance rule*: among all configurations whose validation accuracy is within one percentage point of the best, choose the one with the smallest validation log-loss; break remaining ties towards the smallest γ . The log-loss criterion is a strictly proper scoring rule ([Gneiting and Raftery 2007](#)), hence sensitive to the calibration gains that accuracy cannot see; the tolerance ensures those gains are never bought with a material accuracy loss; and the presence of $\gamma = 0$ in the grid makes plain NB an ever-available fallback (Proposition 4). The search is cheap because all configurations share the same Gaussian statistics and dependence matrices per internal fold: one pass computes them, and each configuration is then a reweighted scoring (Algorithm 1).

Algorithm 1 ABD-NB: training with adaptive selection

Require: data $\mathcal{D}$; measure D ; grid $\Gamma = \{0, \dots, 4\} \times \{\text{glob, cls}\}$; folds V ; tolerance $\epsilon = 0.01$
Ensure: fitted parameters $(\hat{\pi}, \hat{\mu}, \hat{\sigma}^2, W)$

- 1: **for** each internal fold $v = 1, \dots, V$ **do** $\triangleright$ selection phase
- 2: fit $\hat{\pi}, \hat{\mu}, \hat{\sigma}^2$ and $\hat{D}^{(c)}, \hat{D}$ on $\mathcal{D} \setminus \mathcal{D}_v$ $\triangleright$ once per fold
- 3: **for** each $(\gamma, s) \in \Gamma$ **do**
- 4: $W \leftarrow$ weights via (7)–(10)
- 5: accumulate accuracy and log-loss of (2) on $\mathcal{D}_v$
- 6: **end for**
- 7: **end for**
- 8: $A^* \leftarrow \max_{(\gamma, s)} \text{acc}(\gamma, s)$
- 9: $(\hat{\gamma}, \hat{s}) \leftarrow \arg \min \{\log \text{loss}(\gamma, s) : \text{acc}(\gamma, s) \geq A^* - \epsilon\}$, ties $\rightarrow$ smallest γ
- 10: refit $\hat{\pi}, \hat{\mu}, \hat{\sigma}^2, \hat{D}^{(c)}, \hat{D}$ on all of $\mathcal{D}$; $W \leftarrow$ weights at $(\hat{\gamma}, \hat{s})$
- 11: **return** $(\hat{\pi}, \hat{\mu}, \hat{\sigma}^2, W)$

Prediction is exactly (2): $\hat{y}(x) = \arg \max_c [\log \hat{\pi}_c + \sum_i w_{ic} \log \phi(x_i; \hat{\mu}_{ic}, \hat{\sigma}_{ic}^2)]$, with posteriors obtained by soft-max normalisation. Although this paper instantiates the likelihood with Gaussians, stages 1–4 only consume $\log p(x_i | c)$ and thus port verbatim to multinomial, Bernoulli, or kernel-density factors.

Dynamic variant for non-stationary streams

All quantities that ABD-NB needs are moments, so the model has a natural streaming form. *Online ABD-NB* maintains (i) per-class exponentially weighted means and variances with forgetting factor λ (effective memory $(1 - \lambda)^{-1}$ samples), and (ii) an exponentially weighted feature covariance, from which a correlation-based dependence matrix and harmonic weights (7)–(9) are refreshed every T_r samples (default $T_r = 25$). Each update costs $O(d^2)$; setting $\gamma = 0$ recovers online GNB exactly, which serves as the matched control in the drift experiment. When the stream’s dependence structure changes, the weights follow within the forgetting window—the model literally watches features become redundant and discounts them on the fly (Figure 19).

Computational complexity

With n samples, d features, K classes, grid size G , and V internal folds: Gaussian statistics cost $O(nd)$; the Spearman dependence stage costs $O(dn \log n + d^2n)$ (per class-conditional matrix, on class subsamples summing to n); M_{eff} costs one symmetric eigendecomposition, $O(d^3)$; the selection phase multiplies the validation scoring, $O(nKdG)$, but *not* the estimation. Total training is $O(V(d^2n + d^3 + nKdG))$, and prediction is identical to NB, $O(Kd)$ per instance. Empirically (Section), full ABD-NB including selection trains in 84 ms on average across our benchmark folds—an order of magnitude above GNB’s 1.5 ms but below logistic regression (108 ms) and far below random forests (418 ms).

Theoretical Analysis

This section establishes the formal properties. Throughout, D^* denotes the population dependence matrix (on the scale of the chosen measure; e.g. for Spearman on Gaussian pairs, $D_{ij}^* = \frac{6}{\pi} \arcsin(\rho_{ij}/2)$ in absolute value), $\tilde{w}(D)$ the Stage-2 weights, and $w(D)$ the rescaled weights (9).

Proposition 1. Degeneracy. *For $\gamma = 0$, $w_i = 1$ for all i and ABD-NB coincides with Gaussian naive Bayes—identical scores, decisions, and posteriors.*

Proof. With $\gamma = 0$, $\tilde{w} \equiv 1$ by (7) (the rescaling is skipped in this case by construction), so (2) reduces to (3).

Lemma 1. Properties of the spectral effective dimension. *Let $D \in [0, 1]^{d \times d}$ be symmetric with zero diagonal and $R = D + I$ positive semi-definite. Then (i) $1 \leq M_{\text{eff}}(D) \leq d$; (ii) if $D = 0$ (mutual independence) then $M_{\text{eff}} = d$; (iii) if D consists of one block of r exact duplicates ($D_{ij} = 1$ inside the block) and zeros elsewhere, then $M_{\text{eff}} = d - r + 1$.*

Proof. (ii) $R = I$ has all eigenvalues 1, each contributing $1\{1 \geq 1\} + 0 = 1$. (iii) R is block-diagonal: the $r \times r$ all-ones block has eigenvalues r (once) and 0 ($r-1$ times); the remaining $d-r$ coordinates contribute eigenvalue 1 each. Hence $M_{\text{eff}} = f(r) + (r-1)f(0) + (d-r)f(1)$ with $f(\lambda) = 1\{\lambda \geq 1\} + (\lambda - \lfloor \lambda \rfloor)$, i.e. $M_{\text{eff}} = 1 + 0 + (d-r) = d - r + 1$. (i) Each $f(\lambda_k) \in [0, 2)$ and $\sum_k \lambda_k = \text{tr } R = d$; the definition clips to $[1, d]$, and both endpoints are attained by (ii)–(iii) with $r = d$.

Theorem 1. Exact posterior recovery under duplication. *Let the “reduced” problem have features $Z = (Z_1, \dots, Z_m)$ that are conditionally independent given Y , and let the observed vector X consist of Z together with $r-1$ exact copies of Z_1 (so $d = m + r - 1$). Consider population-level ABD-NB with the harmonic rule at $\gamma = 1$ (any $\kappa \geq 1$ in D_{ij}^κ) and spectral rescaling. Then: (i) every copy of Z_1 (including Z_1) receives weight $1/r$ and every other feature weight 1; (ii) the rescaling factor in (9) equals 1; (iii) the ABD-NB score equals the Bayes score of the reduced problem, so ABD-NB attains the Bayes posterior and Bayes risk, whereas the NB score counts the Z_1 -evidence r times:*

$$S_c^{\text{NB}}(x) - S_c^{\text{ABD}}(x) = (r-1) \log p(z_1 | c).$$

Proof. (i) In the population matrix, $D_{ij}^* = 1$ within the duplicate block (any reasonable measure assigns dependence 1 to almost-surely identical variables) and $D_{ij}^* = 0$ between conditionally independent pairs. Hence for a copy i of Z_1 : $\sum_{j \neq i} (D_{ij}^*)^\kappa = r-1$ and $\tilde{w}_i = 1/(1 + (r-1)) = 1/r$; for any other feature the degree is 0 and $\tilde{w}_i = 1$. (ii) $\sum_i \tilde{w}_i = r \cdot \frac{1}{r} + (m-1) = m$, while $M_{\text{eff}} = d - r + 1 = m$ by Lemma 1(iii); the factor $M_{\text{eff}} / \sum \tilde{w}_i = 1$. (iii) The ABD-NB score is $\log \pi_c + \frac{1}{r} \sum_{\text{copies}} \log p(z_1 | c) + \sum_{i=2}^m \log p(z_i | c) = \log \pi_c + \sum_{i=1}^m \log p(z_i | c)$, which is the exact log-joint of the reduced problem by conditional independence of Z ; normalising yields the true posterior. The NB score replaces the first term’s coefficient $\frac{1}{r} \cdot r = 1$ by r , giving the stated gap.

Corollary 1. Non-vanishing NB excess risk. *In the setting of Theorem 1 with $m = 2$, let Z_1, Z_2 be Gaussian with class-mean separations $\delta_1 < \delta_2$ (weak duplicated evidence vs. strong independent evidence) and unit variances. As $r \rightarrow \infty$ the population NB decision converges to sign-classification by Z_1 alone, with risk $\Phi(-\delta_1/2) > \Phi(-\sqrt{(\delta_1^2 + \delta_2^2)}/2)$, the Bayes risk attained by ABD-NB. The gap persists for every sample size.*

Proof. The NB discriminant is $r\delta_1 z_1 + \delta_2 z_2$ (up to centring); dividing by r and letting $r \rightarrow \infty$ leaves $\delta_1 z_1$, whose risk is $\Phi(-\delta_1/2)$. ABD-NB’s discriminant is $\delta_1 z_1 + \delta_2 z_2$ by Theorem 1(iii), the Fisher rule for the reduced problem, with risk $\Phi(-\|\delta\|/2)$.

Corollary 1 is exactly the mechanism of the conflict experiments (Sections –), where empirical NB loses up to 17 accuracy points to ABD-NB at large r .

Proposition 2. Overconfidence under equicorrelation. *Let $X | Y=c \sim \mathcal{N}(\mu_c, \Sigma)$, $c \in \{0, 1\}$, with $\Sigma = (1 - \rho)I + \rho \mathbf{1}\mathbf{1}^\top$, $\mu_1 - \mu_0 = \frac{\delta}{\sqrt{d}} \mathbf{1}$, and equal priors. Let $\Lambda^*(x)$ denote the true log-posterior odds and $\Lambda^{\text{NB}}, \Lambda^{\text{ABD}}$ the population log-odds of NB and of Pearson-based ABD-NB with uniform weights $\bar{w} = M_{\text{eff}}/d$. Then*

$$\Lambda^{\text{NB}}(x) = (1 + (d-1)\rho) \Lambda^*(x),$$

$$\Lambda^{\text{ABD}}(x) = \bar{w} (1 + (d-1)\rho) \Lambda^*(x),$$

with $M_{\text{eff}} = d - \lfloor (d-1)\rho \rfloor$ for the equicorrelated matrix. The ABD-NB inflation factor $h(\rho) = \bar{w} (1 + (d-1)\rho)$

satisfies

$$h(\rho) \leq \frac{(d+2)^2}{4d} \quad \text{for all } \rho \in [0, 1),$$

$$\lim_{\rho \rightarrow 1^-} h(\rho) = 2,$$

whereas the NB factor increases monotonically to d as $\rho \rightarrow 1^-$. In the strong-dependence regime ABD-NB thus reduces the overconfidence from $\Theta(d)$ to $O(1)$; both classifiers keep the Bayes-optimal decision (the factors are positive), so the correction here acts purely on calibration.

Proof. $\mathbf{1}$ is an eigenvector of Σ with eigenvalue $1 + (d-1)\rho$. The true log-odds are $\Lambda^*(x) = (\mu_1 - \mu_0)^\top \Sigma^{-1}(x - \bar{\mu}) = \frac{\delta}{\sqrt{d}} \frac{\mathbf{1}^\top (x - \bar{\mu})}{1 + (d-1)\rho}$ with $\bar{\mu} = (\mu_0 + \mu_1)/2$. NB uses $\Sigma_{\text{NB}} = I$ (unit marginal variances), giving $\Lambda^{\text{NB}} = \frac{\delta}{\sqrt{d}} \mathbf{1}^\top (x - \bar{\mu})$, i.e. the first identity; the decisions agree because Λ^{NB} is a positive multiple of Λ^* —only the confidence is inflated. By the symmetry of the design all Stage-2 degrees coincide, so the rescaled weights are uniform with sum M_{eff} , i.e. $w_i \equiv \bar{w} = M_{\text{eff}}/d$; multiplying every term of the NB score by $\bar{w}$ gives the second identity. For the equicorrelated R the eigenvalues are $1 + (d-1)\rho$ (once) and $1 - \rho$ ($d-1$ times); with $f(\lambda) = \mathbf{1}\{\lambda \geq 1\} + \lambda - \lfloor \lambda \rfloor$ and $\rho \in (0, 1)$,

$$M_{\text{eff}} = 1 + \underbrace{(d-1)\rho - \lfloor 1 + (d-1)\rho \rfloor + 1}_{f(1+(d-1)\rho)} + (d-1)(1-\rho) = d - \lfloor (d-1)\rho \rfloor.$$

Writing $u = (d-1)\rho \in [0, d-1)$ and using $\lfloor u \rfloor \geq u - 1$,

$$h(\rho) = \frac{(d - \lfloor u \rfloor)(1 + u)}{d}$$

$$\leq \frac{(d + 1 - u)(1 + u)}{d} \leq \frac{(d + 2)^2}{4d},$$

the last step maximising the concave quadratic at $u = d/2$. As $\rho \rightarrow 1^-$, $u \rightarrow d-1$ and $\lfloor u \rfloor = d-2$ eventually, so $h(\rho) \rightarrow (d - (d-2)) \cdot d/d = 2$.

We now turn to the sampling behaviour of the weights. The next result requires a mild genericity assumption ruling out the measure-zero configurations where the Li-Ji functional (8) is non-smooth.

Assumption 1. The population matrix D^* satisfies: (i) the dependence estimator applied to i.i.d. data is strongly consistent for D^* entrywise with entrywise asymptotic normality at rate $\sqrt{n}$ (true for Pearson, Spearman, and Kendall, which are smooth functionals of U -statistics, [van der Vaart 1998](#); [Lehmann 1999](#)); (ii) no eigenvalue of $D^* + I$ is a positive integer, and the eigenvalues are simple where they exceed 1.

Theorem 2. Consistency and asymptotic normality of the weights. Under Assumption 1, with soft-threshold level $t_n = z_{1-\alpha/2}/\sqrt{n-3} \rightarrow 0$ and fixed (γ, κ) : (i) $\hat{w} \rightarrow w^* \equiv w(D^*)$ almost surely; (ii) $\sqrt{n}(\hat{w} - w^*) \rightsquigarrow \mathcal{N}(0, J \Sigma_D J^\top)$, where Σ_D is the asymptotic covariance of the half-vectorised $\hat{D}$ and J is the Jacobian of the map $D \mapsto w(D)$ at D^* ; (iii) consequently $\|\hat{w} - w^*\| = O_P(n^{-1/2})$.

Proof. (i) The map $D \mapsto \tilde{w}(D)$ is a composition of polynomials and the reciprocal on a domain bounded away from poles ($1 + \gamma \sum_j D_{ij}^{2\kappa/2} \geq 1$), hence continuous; $D \mapsto M_{\text{eff}}(D)$ is continuous at any D^* satisfying (ii) of Assumption 1, because eigenvalues are continuous functions of the matrix and the floor/indicator discontinuities of f sit exactly at integer eigenvalues, excluded by assumption. The soft-threshold (5) converges uniformly to the identity as $t_n \rightarrow 0$. The continuous-mapping theorem applied to $\hat{D} \rightarrow D^*$ (a.s.) yields $\hat{w} \rightarrow w^*$ a.s. (ii) On a neighbourhood of D^* excluded from the discontinuity set, $w(\cdot)$ is continuously differentiable: $\tilde{w}$ is rational with non-vanishing denominator; M_{eff} is differentiable because, near a matrix with no integer eigenvalues, $\lambda \mapsto f(\lambda)$ is locally the smooth map $\lambda \mapsto \lambda - c_k$ or $\lambda \mapsto 1 + \lambda - c_k$ with locally constant integers c_k , and simple eigenvalues are differentiable in the matrix. The delta method ([van der Vaart 1998](#), Ch. 3) applied to the entrywise CLT for $\hat{D}$ (the threshold contributes $O(t_n) = O(n^{-1/2})$ deterministic perturbation, absorbed in the limit) gives the stated limit. (iii) follows from (ii).

Proposition 3. Bias, variance, and MSE. Under the conditions of Theorem 2, each weight coordinate satisfies

$$\mathbb{E}[\hat{w}_i] - w_i^* = O(n^{-1/2}), \quad \text{Var}(\hat{w}_i) = O(n^{-1}),$$

$$\mathbb{E}[(\hat{w}_i - w_i^*)^2] = O(n^{-1}),$$

and the bias term is $O(n^{-1})$ when the threshold is disabled ($t_n \equiv 0$), by the second-order delta method.

Section verifies all three rates numerically: the observed $\sqrt{\text{MSE}}$ of the raw harmonic weight tracks the $n^{-1/2}$ reference over two decades of n , and the empirical bias sits an order of magnitude below the standard deviation (Figure 14).

Proposition 4. Asymptotic safety of the tuned classifier. Let $R(g)$ denote the risk (expected zero-one loss) of a classifier g , let Γ be the finite configuration grid containing $(\gamma, s) = (0, \text{glob})$, and let $g_n^{(\gamma, s)}$ be ABD-NB fitted at configuration (γ, s) on n samples. Assume each configuration's fitted classifier has a risk that converges in probability to a limit $R^{(\gamma, s)}$ (true under Assumption 1, since the Gaussian parameters and weights converge). Then the internally selected classifier g_n^{sel} satisfies

$$R(g_n^{\text{sel}}) \leq \min_{(\gamma, s) \in \Gamma} R^{(\gamma, s)} + o_P(1) \leq R^{\text{NB}} + o_P(1),$$

where $R^{\text{NB}} = R^{(0, \text{glob})}$ is the asymptotic risk of Gaussian naive Bayes. In particular ABD-NB is asymptotically never worse than NB, and strictly better whenever some grid point has strictly smaller limiting risk (e.g. the setting of Corollary 1).

Proof. The internal validation folds provide, for each of the $|\Gamma|$ configurations, an empirical risk estimate based on $\Theta(n)$ held-out points. By Hoeffding's inequality ([Hoeffding 1963](#)) and a union bound over the finite grid, all estimates converge uniformly in probability to the corresponding limiting risks. The selection rule picks a configuration whose estimated accuracy is within ϵ_n (the tolerance, which may be sent to 0 with n) of the best; standard arguments for selection by validation ([Kohavi 1995](#)) then give the first inequality. The second holds because $(0, \text{glob}) \in \Gamma$.

Remark 1. *Proposition 4 formalises requirement R4 and is visible in the benchmark: on Digits—adverse for dependence weighting—the selector returns $\gamma = 0$ in most folds and the flagship matches GNB to 0.05 accuracy points, while fixed- γ variants lose up to 23 points (Section).*

Experimental Setup

All experiments are fully scripted and reproducible; the implementation (Python 3.11, scikit-learn compatible, Pedregosa et al. 2011) and every experiment script are provided in the companion repository.

Datasets

Table 1 lists the 15 datasets. Four real benchmarks (Iris, Wine, Breast Cancer Wisconsin Diagnostic, Digits) come from the UCI repository (Dua and Graff 2019) as distributed with scikit-learn and span 4–64 features and 2–10 classes. Two *redundancy-augmented* versions of the real data (iris-red, wine-red) append two noisy duplicates of every feature (relative noise 10%), modelling repeated measurements. Nine controlled synthetic families make the dependence structure exactly known: equicorrelated Gaussians at $\rho \in \{0, 0.3, 0.6, 0.9\}$ (constant Bayes risk, growing redundancy); block-correlated features; the *conflict* design of Corollary 1 (one strong independent feature vs. a block of 12 near-duplicated weak features, $\rho = 0.95$); a class-asymmetric design where the correlated half of the features switches between classes; a non-monotone design ($X_2 = X_1^2 + \varepsilon$, $X_3 = |X_1| + \varepsilon$) where rank correlations are provably blind; and the scikit-learn `make_classification` generator with 10 redundant linear combinations.

Table 1. Benchmark datasets. n = samples, d = features, K = classes. “Dependence” summarises the dominant within-class dependence character.

Dataset	n	d	K	Dependence
iris	150	4	3	mild
wine	178	13	3	moderate
breast-cancer	569	30	2	strong blocks
digits	1797	64	10	spatial, informative
iris-red	150	12	3	duplicates ($\times 3$)
wine-red	178	39	3	duplicates ($\times 3$)
synth-rho0.0	600	12	2	none (control)
synth-rho0.3	600	12	2	equicorr. $\rho=0.3$
synth-rho0.6	600	12	2	equicorr. $\rho=0.6$
synth-rho0.9	600	12	2	equicorr. $\rho=0.9$
synth-blocks	600	20	2	4 blocks, $\rho=0.85$
synth-conflict	600	13	2	adversarial (Cor. 1)
synth-classdep	600	12	2	class-asymmetric
synth-nonlin	600	6	2	non-monotone
synth-sklearn	600	20	2	linear redundant

Methods compared

The NB family: **GNB** (Gaussian naive Bayes); **MI-WNB**, the relevance-weighted representative, with $w_i = \text{MI}(X_i; Y) / \max_j \text{MI}(X_j; Y)$ (Lee et al. 2011; Jiang et al. 2016); three fixed-configuration ABD-NB ablations, **ABD-NB-L** (linear link, $\gamma=1$, global), **ABD-NB-G** (harmonic, $\gamma=1$, global), **ABD-NB-C** (harmonic, $\gamma=1$, class-specific); and the flagship **ABD-NB** with internal selection (Algorithm 1). Non-Bayesian references: logistic regression (**LR**),

k -nearest neighbours (**kNN**, $k=5$), a CART decision tree, a 200-tree random forest (**RF**), and an RBF **SVM** with probability outputs; LR/kNN/SVM are standardised in-pipeline. Defaults are used throughout—no per-dataset tuning for any method except ABD-NB’s *internal* selection, which uses training data only.

Protocol and metrics

Every (dataset, model) pair is evaluated by $3\times$ repeated stratified 10-fold cross-validation (90 fits per pair; identical folds across models). We report accuracy, macro-F1, log-loss (the strictly proper primary calibration score, Gneiting and Raftery 2007), expected calibration error (ECE, 15 equal-width bins, Guo et al. 2017; Pakdaman Naeini et al. 2015), and wall-clock fit/predict times. Following Demšar (2006), cross-dataset comparisons use the Friedman test (Friedman 1937) with the Iman–Davenport correction, Nemenyi critical differences for the rank diagram, and Wilcoxon signed-rank tests (Wilcoxon 1945) of the flagship against every opponent with Holm correction (Holm 1979). Controlled studies (equicorrelation sweep, conflict sweep, duplication stress, learning curves, drift) use 5 to 30 replications with fresh generator seeds; figures show mean $\pm$ one standard deviation bands.

Results

Controlled dependence studies

Equicorrelation: calibration is the first casualty. Figure 4 sweeps the within-class correlation ρ at constant Bayes difficulty. Accuracy (left) is indistinguishable across GNB, ABD-NB, and LR at every ρ —as Proposition 2 predicts, equicorrelation preserves the NB decision. The probabilistic story is different: GNB’s log-loss grows linearly with ρ (0.35 $\rightarrow$ 1.97 at $\rho=0.95$) and its ECE more than quintuples (0.06 $\rightarrow$ 0.32), while ABD-NB caps the degradation (log-loss 1.10, ECE 0.19 at $\rho=0.95$)—the empirical signature of the $\Theta(d) \rightarrow O(1)$ overconfidence reduction. LR, which models the joint structure, bounds what any within-architecture correction can hope for.

Conflict blocks: decisions fail next. Figure 5 sweeps the size r of the redundant weak block competing with one strong feature. GNB behaves exactly as Corollary 1 predicts: accuracy falls monotonically from 86.4% ($r=1$) to 69.2% ($r=24$) as the duplicated weak evidence out-votes the strong feature, log-loss explodes to 3.31, and ECE reaches 0.29. ABD-NB is essentially flat (86.5% accuracy, log-loss 0.41, ECE 0.09 at $r=24$), statistically indistinguishable from LR. The mechanism is visible in the fitted weights: the block members receive $w \approx 1/r$ while the strong feature keeps $w \approx 1$.

Duplication stress on real data

The same mechanism operates on real data. For each of Iris, Wine, and Breast Cancer we append r noisy copies (5% relative noise) of one randomly chosen feature and track 5-fold CV accuracy over 10 random choices (Figure 6). GNB degrades steadily—on Wine from 97.2% to 80.9% at $r=32$; on Iris from 96.0% to 88.4%—because whichever feature

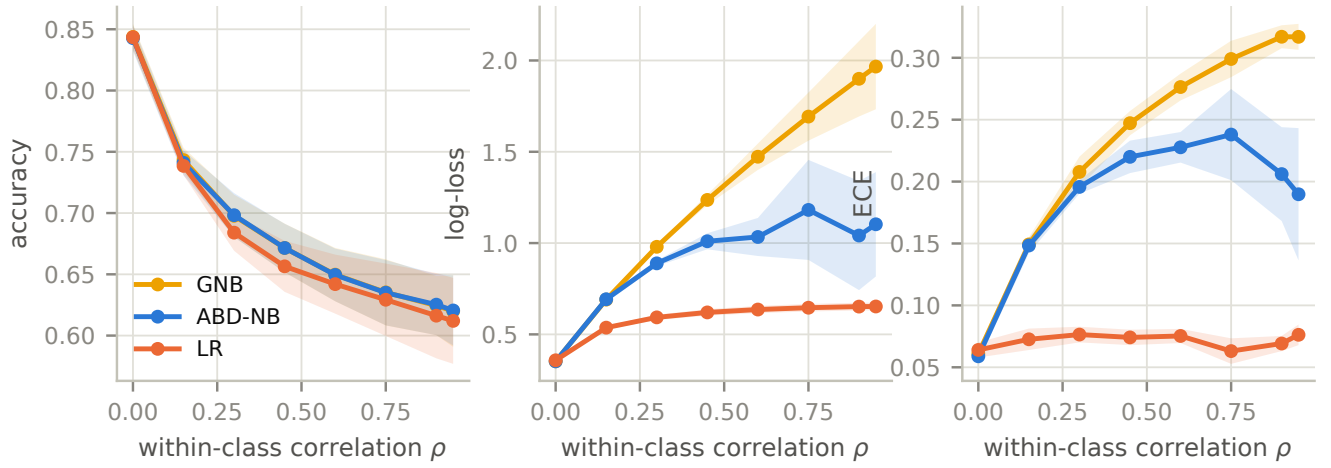


Figure 4. Equicorrelated Gaussians, $d = 12$, constant Bayes risk, 5 replications $\times$ 5-fold CV (bands: ± 1 sd). Left: accuracy is unaffected by ρ -equicorrelation for all methods (the NB decision is preserved; Proposition 2). Centre, right: GNB’s log-loss and ECE deteriorate linearly with ρ ; ABD-NB contains both, closing most of the gap towards the correctly specified logistic regression.

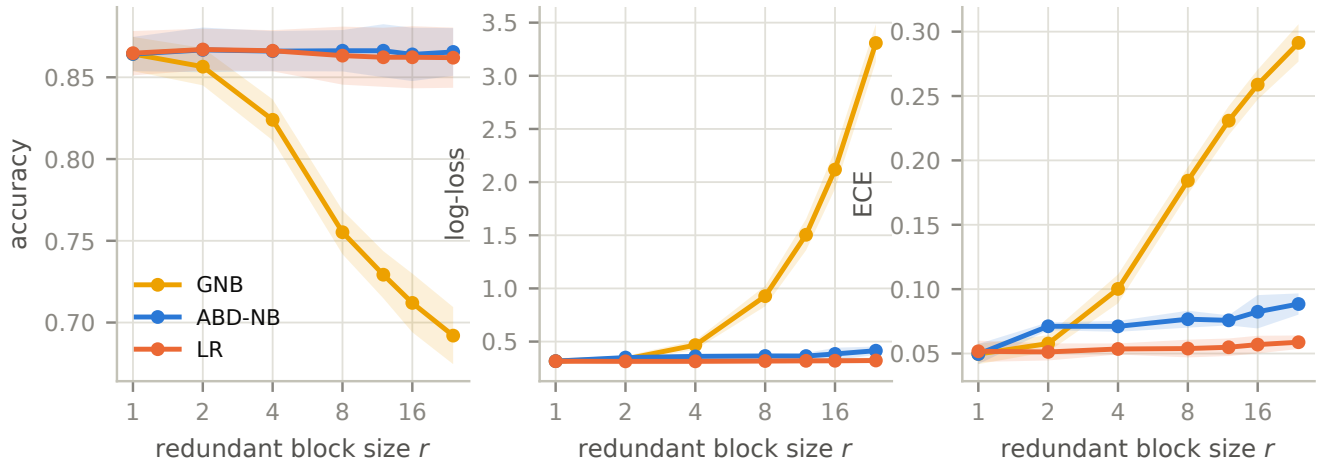


Figure 5. Conflict design (Corollary 1): one strong independent feature vs. r near-duplicated weak features ($\rho = 0.95$). GNB’s accuracy collapses with the block size while its log-loss grows super-linearly; ABD-NB matches logistic regression across the whole sweep. This is the double-counting pathology in its purest form, and the harmonic weights remove it.

happens to be duplicated commanders the posterior. ABD-NB is flat on all three datasets (Wine 97.2% $\rightarrow$ 97.2%; Breast Cancer even improves slightly as duplicates sharpen the dependence estimates), matching the invariance of LR. This experiment is a practical warning: a single accidentally repeated column in a feature pipeline can silently cost naive Bayes 16 accuracy points, and ABD-NB neutralises it without any user intervention.

Benchmark across 15 datasets

Table 2 reports the NB-family accuracies; Figure 7 extends the comparison to all 11 methods as a rank-coloured matrix, and Figure 8 visualises the three-way comparison GNB vs. relevance weighting vs. ABD-NB. Four findings stand out.

First, *where dependence hurts NB, ABD-NB fixes it*: +12.8 points on synth-conflict (74.6 $\rightarrow$ 87.4), +0.4 on Breast Cancer, +0.4/+0.5 on the redundancy-augmented real datasets for the $\gamma=1$ variant, with the largest gains exactly where Theorem 1 operates. Second, *the safeguard works*: across all 15 datasets the flagship’s worst regression relative to GNB is 0.05 accuracy points (iris, wine: within fold noise), and on Digits—where fixed- γ dependence

weighting costs up to 23 points (ABD-NB-C: 61.1%)—the selector returns $\gamma=0$ and ties GNB. Third, *class-specific weights need their safeguards*: the synth-classdep row shows raw class-specific weighting collapsing to 40.7% (the mechanical bias of Section 4 compounds with estimation noise), while the common-scale global variants sit at 65.4%; the flagship avoids the trap automatically. Fourth, *relevance and redundancy weighting are orthogonal*: MI-WNB gains +4.8 points on Digits but loses on 8 of the remaining 14 datasets, whereas ABD-NB never materially loses—supporting our claim that the two families correct different failure modes and could in principle be composed.

Across all 11 methods the Friedman test rejects rank equality decisively ($\chi^2_{10} = 29.2$, $p = 1.2 \times 10^{-3}$; Iman–Davenport $F = 3.38$, $p = 5.5 \times 10^{-4}$). Figure 9 shows the Nemenyi diagram: ABD-NB-G (mean rank 4.73) and the flagship (5.13) sit in the top group with SVM (4.47) and LR (4.97), above random forests (5.53), and 1.1–1.5 ranks above GNB (6.20); CART is the only method separated beyond the critical difference (CD = 3.90 at $\alpha = 0.05$, $N = 15$, $k = 11$). Per-fold dispersion for four representative datasets is shown in Figure 10. Pairwise Wilcoxon tests

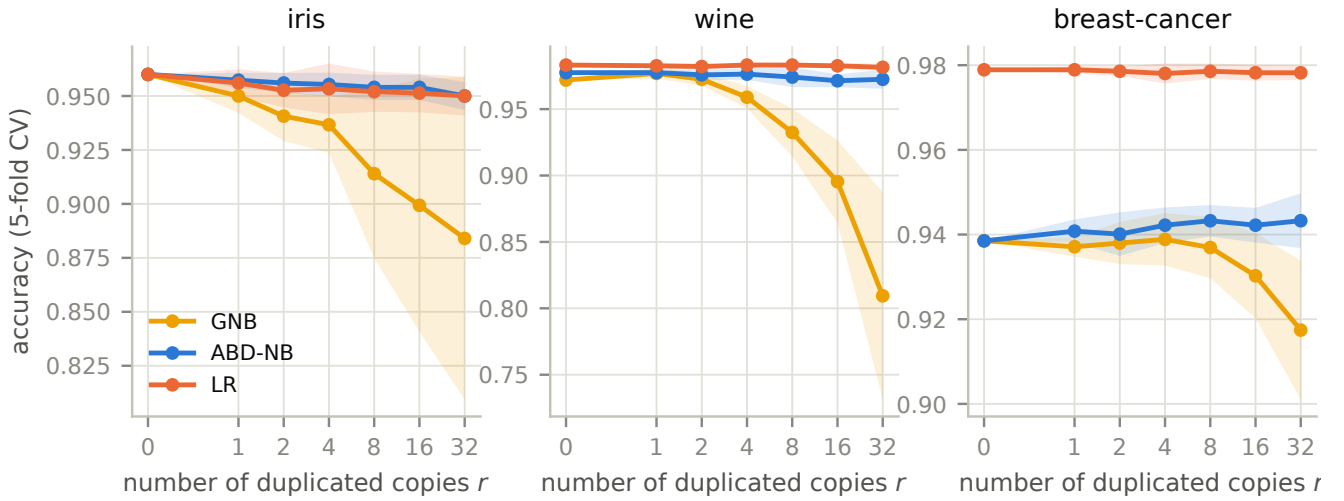


Figure 6. Duplication stress on real data: r noisy copies of one randomly chosen feature are appended (10 random choices, bands: ± 1 sd across choices). GNB loses up to 16 accuracy points; ABD-NB and LR are invariant. The flagship’s internal selector consistently chooses $\gamma \geq 1$ once duplicates appear.

Table 2. Accuracy (% , mean $\pm$ sd over 3×10 -fold CV) of the naive Bayes family. **Bold:** best NB-family result per dataset. ABD-NB-L/G/C are fixed-configuration ablations ($\gamma=1$); the flagship ABD-NB selects (γ , structure) internally (Algorithm 1).

Dataset	GNB	MI-WNB	ABD-NB-L	ABD-NB-G	ABD-NB-C	ABD-NB
iris	95.8 $\pm$ 5.7	96.0 $\pm$ 5.5	95.6 $\pm$ 5.6	95.6 $\pm$ 5.6	95.3 $\pm$ 5.9	95.3 $\pm$ 5.9
wine	97.7 $\pm$ 3.6	95.5 $\pm$ 4.9	97.7 $\pm$ 3.6	97.7 $\pm$ 3.6	97.7 $\pm$ 3.6	97.5 $\pm$ 3.6
breast-cancer	93.9 $\pm$ 3.3	94.0 $\pm$ 3.2	94.1 $\pm$ 3.5	94.2 $\pm$ 3.6	94.2 $\pm$ 3.7	94.3 $\pm$ 3.8
digits	84.1 $\pm$ 2.2	88.9 $\pm$ 2.3	83.3 $\pm$ 2.4	75.4 $\pm$ 2.7	61.1 $\pm$ 3.0	84.1 $\pm$ 2.2
iris-red	96.2 $\pm$ 4.9	96.4 $\pm$ 4.7	96.4 $\pm$ 4.7	96.7 $\pm$ 4.6	96.0 $\pm$ 4.9	96.4 $\pm$ 4.9
wine-red	97.5 $\pm$ 3.6	95.9 $\pm$ 4.4	97.7 $\pm$ 3.4	97.9 $\pm$ 3.3	97.5 $\pm$ 3.6	97.5 $\pm$ 3.6
synth-rho0.0	65.7 $\pm$ 5.3	62.4 $\pm$ 5.8	65.7 $\pm$ 5.3	65.7 $\pm$ 5.3	65.7 $\pm$ 5.3	65.7 $\pm$ 5.4
synth-rho0.3	59.1 $\pm$ 5.5	57.0 $\pm$ 6.2	59.4 $\pm$ 5.6	59.6 $\pm$ 5.5	58.9 $\pm$ 5.7	59.1 $\pm$ 5.6
synth-rho0.6	57.9 $\pm$ 4.3	57.4 $\pm$ 4.6	58.0 $\pm$ 4.4	58.1 $\pm$ 4.4	58.1 $\pm$ 4.5	58.2 $\pm$ 4.4
synth-rho0.9	57.7 $\pm$ 5.5	57.6 $\pm$ 5.6	58.2 $\pm$ 5.4	58.2 $\pm$ 5.4	58.2 $\pm$ 5.4	58.2 $\pm$ 5.4
synth-blocks	96.7 $\pm$ 2.2	96.4 $\pm$ 2.3	96.7 $\pm$ 2.2	96.9 $\pm$ 1.8	96.9 $\pm$ 1.8	96.9 $\pm$ 1.8
synth-conflict	74.6 $\pm$ 6.1	86.9 $\pm$ 5.3	87.2 $\pm$ 5.2	86.3 $\pm$ 5.3	86.3 $\pm$ 5.3	87.4 $\pm$ 5.2
synth-classdep	64.8 $\pm$ 6.1	63.9 $\pm$ 6.4	65.2 $\pm$ 6.2	65.4 $\pm$ 6.2	40.7 $\pm$ 6.9	65.1 $\pm$ 6.1
synth-nonlin	89.4 $\pm$ 4.0	89.5 $\pm$ 4.0	88.8 $\pm$ 4.1	88.9 $\pm$ 4.1	88.8 $\pm$ 4.1	89.4 $\pm$ 4.0
synth-sklearn	85.2 $\pm$ 3.8	84.7 $\pm$ 4.0	85.5 $\pm$ 3.8	85.2 $\pm$ 3.9	84.9 $\pm$ 4.0	84.9 $\pm$ 3.9
Avg. rank (all 11 models)	6.20	6.93	5.33	4.73	6.30	5.13

of the flagship (Table 3) show accuracy parity with the strong discriminative baselines and dominance over CART, class-specific ablation, and MI-WNB at the uncorrected level; on *log-loss* the flagship beats GNB on 10/15 datasets ($p = 0.008$, Holm-adjusted 0.059) and kNN and CART decisively, while conceding to LR/SVM—models with 10^2 – $10^3 \times$ higher training cost.

Calibration

Calibration is where dependence weighting pays off most systematically. On the Friedman analysis of *log-loss* the effect is far stronger than on accuracy ($\chi^2_{10} = 77.5$, $p = 1.6 \times 10^{-12}$): the flagship’s mean *log-loss* rank is 5.47 against GNB’s 7.67 and kNN’s 8.20. Table 4 details the two probabilistic metrics for the three NB-family protagonists. The reductions are largest exactly where the theory predicts: Breast Cancer $0.589 \rightarrow 0.216$ (−63%), synth-conflict $1.237 \rightarrow 0.341$ (−72%), synth-rho0.9 $1.471 \rightarrow 0.783$ (−47%), wine-red $0.218 \rightarrow 0.129$ (−41%); on the independence control (synth-rho0.0) the selector leaves NB

Table 3. Wilcoxon signed-rank tests of the flagship ABD-NB against each opponent across the 15 datasets (per-dataset mean metrics; W/T/L = wins/ties/losses of ABD-NB). p_H = Holm-adjusted.

Opponent	Accuracy			Log-loss		
	W/T/L	p	p_H	W/T/L	p	p_H
GNB	7/3/5	.230	1.0	10/0/5	.008	.059
MI-WNB	11/1/3	.050	.399	8/0/7	.847	.847
ABD-NB-L	6/2/7	.909	.909	7/0/8	.762	1.0
ABD-NB-G	5/2/8	.669	1.0	8/0/7	.679	1.0
ABD-NB-C	9/3/3	.032	.289	12/0/3	.026	.153
LR	8/0/7	.820	1.0	3/0/12	.004	.034
kNN	10/0/5	.804	1.0	12/0/3	.055	.221
CART	12/0/3	.026	.256	15/0/0	10^{-4}	.001
RF	7/0/8	.762	1.0	5/0/10	.048	.240
SVM	7/1/7	.478	1.0	3/0/12	.002	.018

untouched, as it should. Figure 11 gives the complete per-dataset picture, and Figure 12 shows reliability diagrams: GNB’s curve sags far below the diagonal (predicted confidence 0.9, empirical accuracy 0.56 on conflict), while ABD-NB hugs it.

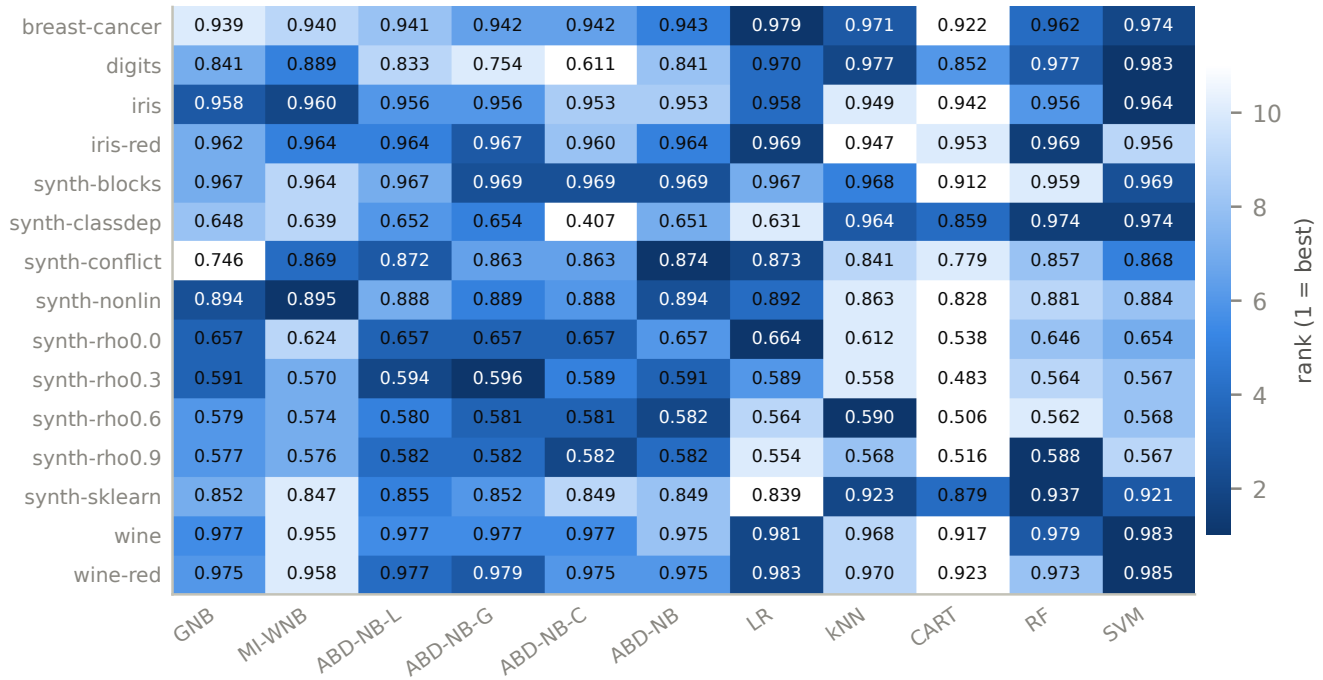


Figure 7. Mean CV accuracy of 11 methods on 15 datasets (darker = better). ABD-NB is the most consistent NB variant and automatically falls back to $\gamma = 0$ on Digits, matching GNB while fixed- γ variants fail.

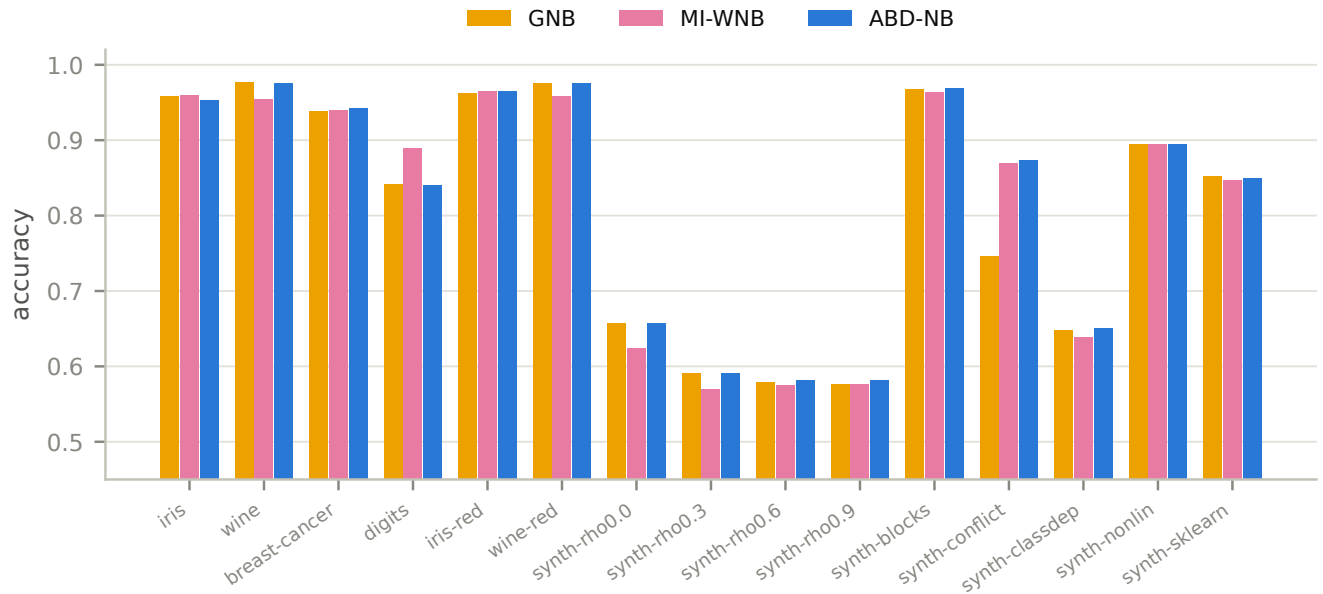


Figure 8. GNB, MI-WNB, and ABD-NB across benchmark datasets. MI-WNB favors relevance-dominated data, whereas ABD-NB is superior on redundancy-dominated datasets and avoids the performance losses observed with MI-WNB.

Empirical verification of the theory

Duplicate recovery. With r exact copies planted among 6 base features ($n = 3000$), the estimated per-copy weight tracks $1/r$ within estimation noise for every $r \in \{1, \dots, 10\}$ (e.g. 0.492 vs. 0.5 at $r=2$; 0.110 vs. 0.1 at $r=10$), and the realised effective dimension equals the theoretical $d^* = 6$ exactly in all runs (Figure 13)—Theorem 1 and Lemma 1 in action.

Convergence rates. Figure 14 tracks the weight estimation error on the equicorrelated design as n grows from 50 to 6400 (30 replications). The ℓ_2 error of the full weight vector decays parallel to the $n^{-1/2}$ reference over two decades; the

bias of the raw harmonic weight is an order of magnitude smaller than its standard deviation at every n , so the MSE is variance-dominated—precisely the regime described by Theorem 2 and Proposition 3.

Sample efficiency and the persistence of NB's excess risk.

Figure 15 shows learning curves on the conflict design with a fixed test set of 4000 points. ABD-NB reaches 84.5% with only 60 training samples—weight estimation is not a large-sample luxury—and converges to 85.4%, near the Bayes rate. GNB plateaus at 72.1%: more data does not help, because its bias is structural (Corollary 1). The 13-point asymptotic gap

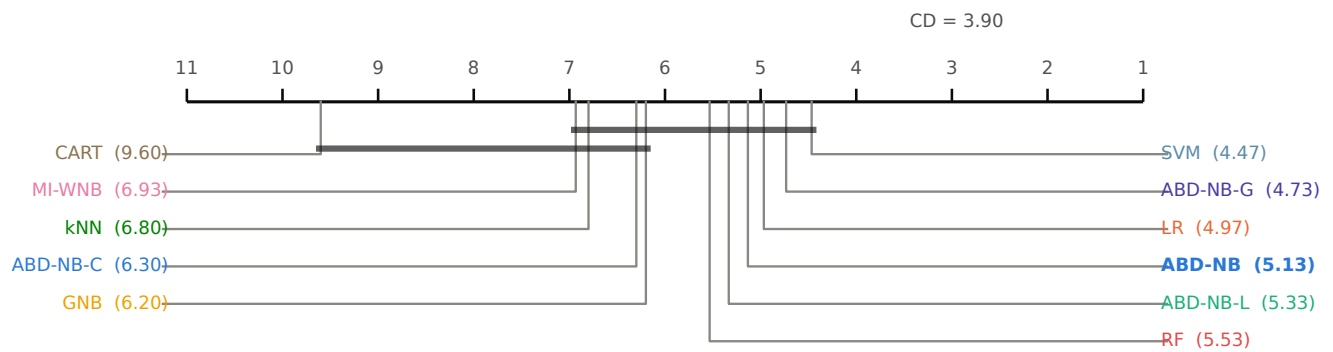


Figure 9. Nemenyi critical-difference diagram on mean accuracy ranks (15 datasets, 11 methods, $\alpha = 0.05$). Lower rank is better. Both ABD-NB variants with harmonic weights rank above random forests and GNB; methods joined by a thick segment are not distinguishable at the critical difference.

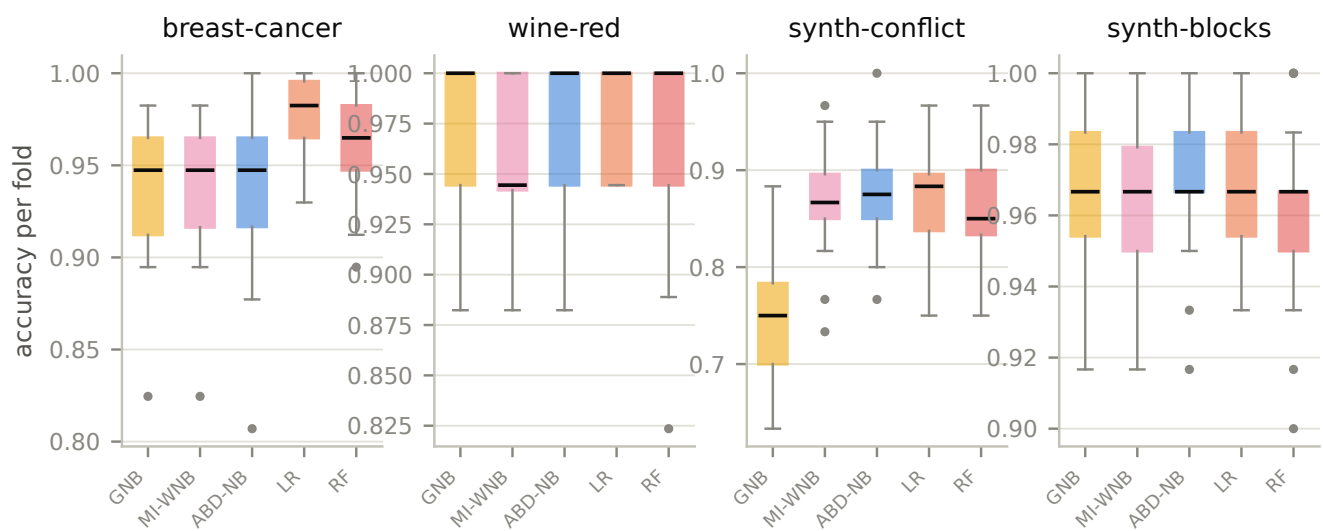


Figure 10. Per-fold accuracy distributions (90 folds) on four representative datasets. On synth-conflict the entire GNB distribution sits below the ABD-NB distribution; on breast-cancer and wine-red the median shifts up with smaller spread; MI-WNB's relevance weighting does not help on wine-red.

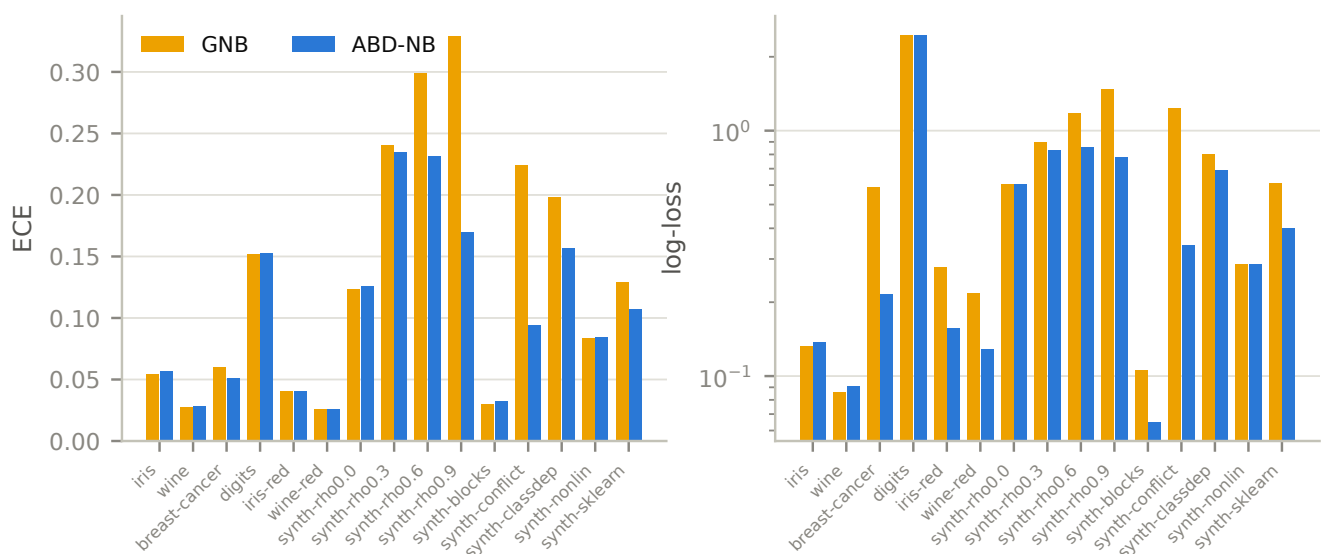


Figure 11. Per-dataset calibration comparison, GNB vs. ABD-NB: ECE (left) and log-loss (right, log scale). The improvements concentrate on the dependence-heavy datasets; on the independence control and on Digits the internal selector leaves NB unchanged rather than degrading it.

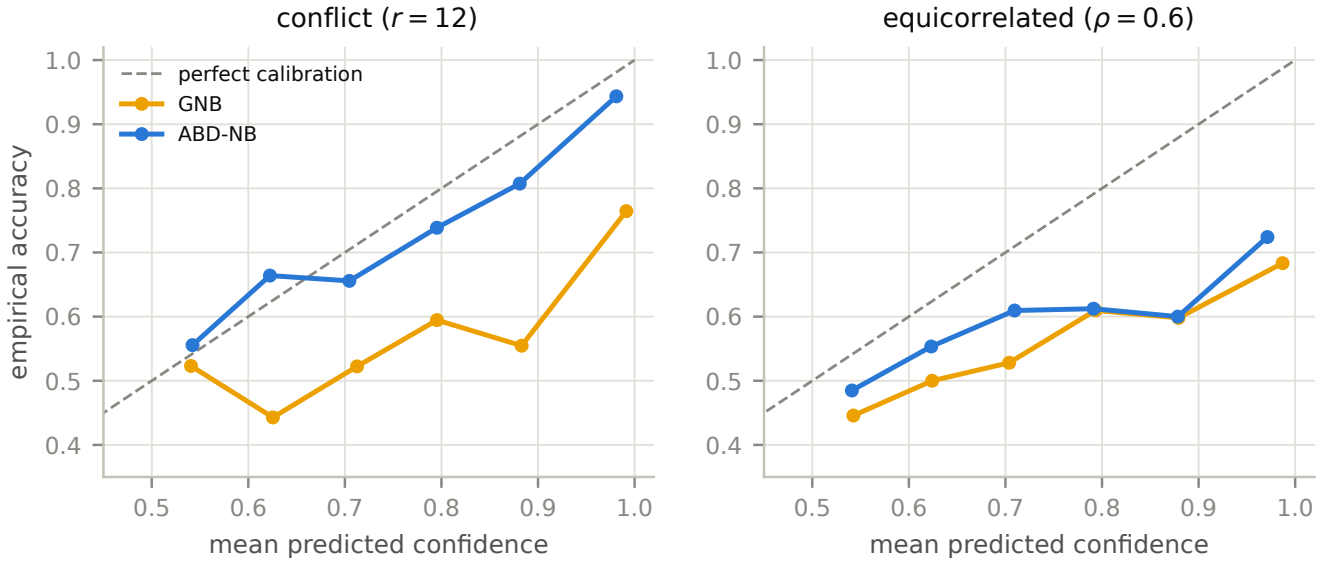


Figure 12. Reliability diagrams (12 confidence bins, held-out halves of $n = 4000$ samples). GNB is grossly overconfident—on the conflict design, predictions made with 90% confidence are correct 56% of the time; ABD-NB tracks the diagonal without any post-hoc recalibration.

Table 4. Calibration metrics (mean over 3×10 -fold CV). LL = log-loss, ECE = expected calibration error. Lower is better; **bold** marks the best of the three.

Dataset	GNB		MI-WNB		ABD-NB	
	LL	ECE	LL	ECE	LL	ECE
iris	.132	.054	.096	.046	.138	.056
wine	.086	.027	.073	.039	.091	.028
breast-cancer	.589	.059	.480	.062	.216	.051
digits	2.441	.152	1.504	.102	2.445	.153
iris-red	.279	.040	.187	.036	.157	.041
wine-red	.218	.026	.169	.040	.129	.026
synth-rho0.0	.607	.124	.649	.106	.607	.125
synth-rho0.3	.895	.240	.691	.136	.837	.235
synth-rho0.6	1.181	.299	.739	.181	.854	.232
synth-rho0.9	1.471	.329	.830	.215	.783	.170
synth-blocks	.106	.029	.125	.033	.065	.032
synth-conflict	1.237	.224	.332	.094	.341	.094
synth-classdep	.801	.198	.668	.137	.692	.156
synth-nonlin	.286	.083	.289	.089	.286	.084
synth-sklearn	.611	.129	.485	.119	.402	.107

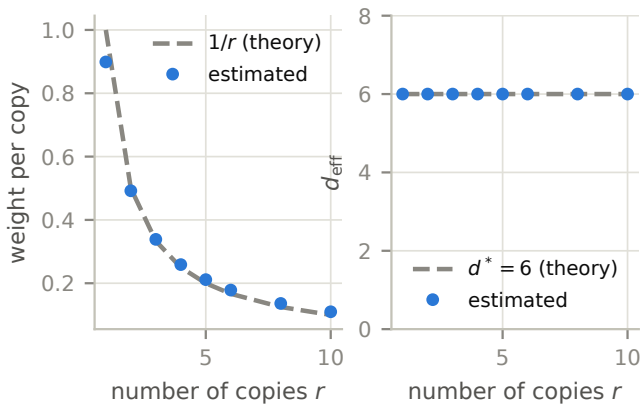


Figure 13. Verification of Theorem 1: with r exact copies of one feature, each copy’s estimated weight (left) matches $1/r$ and the effective dimension (right) stays pinned at the de-duplicated value $d^* = 6$.

is the excess risk that no amount of data can remove from within the unweighted architecture.

Sensitivity and ablation studies

Correction strength γ . Figure 16 sweeps γ over $[0, 8]$ on five representative datasets. The response surfaces explain why no fixed γ suffices and why selection is safe: on synth-conflict, accuracy rises steeply from 72.6% ($\gamma=0$) to a broad plateau around $\gamma \in [1, 2]$ (86.4%); on wine and breast-cancer the curves are nearly flat with a mild optimum near $\gamma \in [1, 4]$; on synth-nonlin accuracy *decreases* slowly in γ (rank measures misread the non-monotone structure), and on the equicorrelated control it is exactly flat. Log-loss (right panel) is uniformly non-increasing in γ up to its optimum on the dependence-heavy datasets. The internal selector needs only to distinguish these coarse regimes, a much easier task than locating a sharp optimum—which is why $V=3$ internal folds suffice.

Choice of weight function. Figure 17 compares the four links of Section 4 at $\gamma=1$. All four behave identically on benign data, but on synth-conflict the harmonic graph rule (85.9%) and the linear link (85.4%) clearly beat the exponential (79.5%) and inverse (77.6%) scalar links, and the harmonic rule attains the best log-loss on the dependence-heavy datasets—empirical support for deriving the functional form from the recovery requirement rather than choosing it by convenience.

Choice of dependence measure. Figure 18 swaps the seven dependence estimators inside an otherwise identical ABD-NB. The differences are small (within 1–2 points on all datasets except a 5-point dip for discretised MI on synth-conflict), confirming that the framework is robust to the plug-in: the heavy lifting is done by the weighting scheme, not by exotic dependence estimation. Spearman’s rank measure is an excellent default—as accurate as distance correlation and HSIC at a fraction of their cost—while MI is preferable

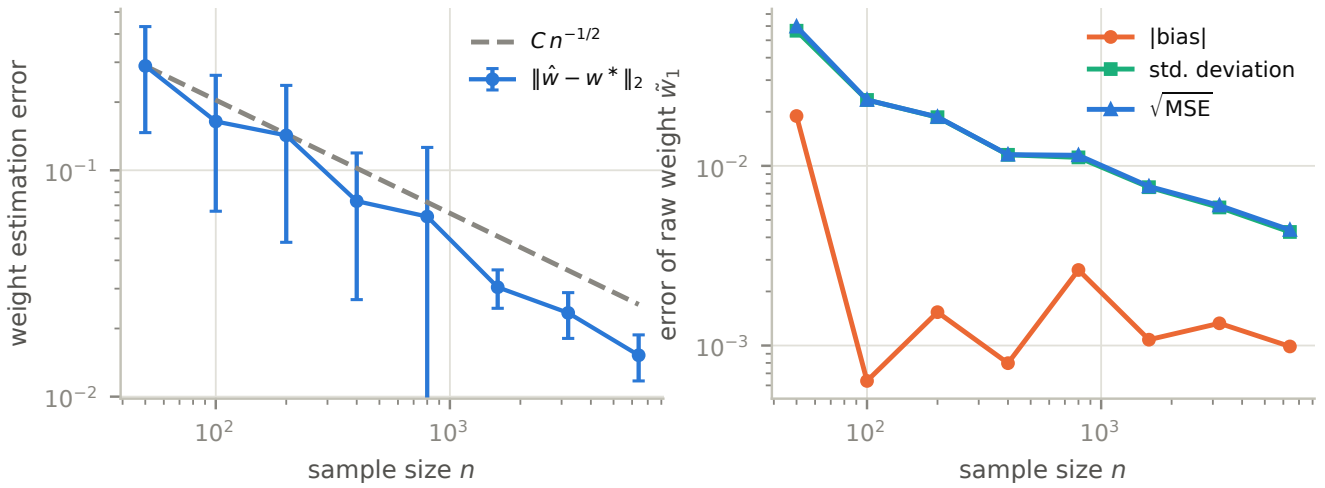


Figure 14. Weight estimator asymptotics on the equicorrelated design ($\rho = 0.6$, $d = 12$, 30 replications). Left: $\|\hat{w} - w^*\|_2$ decays at the parametric $n^{-1/2}$ rate (dashed reference). Right: bias–variance decomposition of the raw harmonic weight $\tilde{w}_1$; the squared bias is negligible against the variance at every sample size.

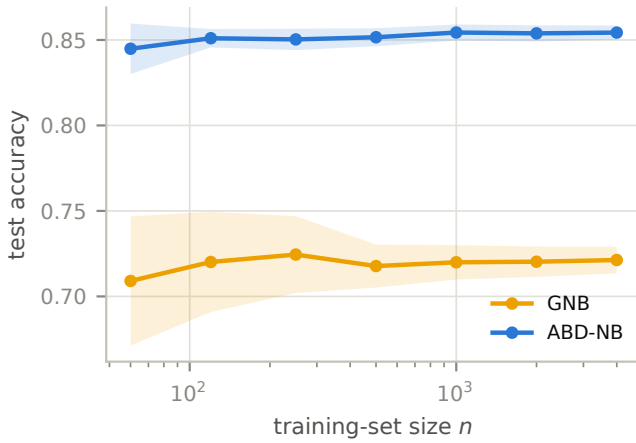


Figure 15. Learning curves on the conflict design (20 replications, test $n = 4000$). GNB’s plateau is the non-vanishing excess risk of Corollary 1; ABD-NB approaches the Bayes rate and is already effective at $n = 60$.

when non-monotone dependence is suspected (synth-nonline: MI 86.4% vs. Spearman 85.0%).

Tracking dependence drift in streams

Figure 19 reports the concept-drift experiment: a stream of 8000 samples from a conflict-type design in which, at $t = 4000$, the redundant block abruptly swaps identity with an independent block of equal marginal behaviour—a pure *dependence drift* that leaves every marginal distribution unchanged and is therefore invisible to standard drift detectors monitoring feature means or error rates. Online ABD-NB (forgetting factor $\lambda = 0.995$) re-estimates the weights within roughly 500 samples: block A’s weights rise from 0.22 to 1.05 and block B’s fall symmetrically, while the strong feature’s weight stays flat (left panel). Prequential accuracy (right panel) tells the operational story: online ABD-NB runs ~ 6 points above online GNB before the drift (90.5% vs. 84.2%), dips by 4 points at the switch, and fully recovers within the forgetting window, restoring its advantage (90.1% vs. 84.3% post-drift). The matched control

($\gamma=0$, identical code path) confirms that the entire gap is attributable to the dynamic dependence correction.

Computational cost

Figure 20 summarises wall-clock costs over all benchmark folds. Fixed-configuration ABD-NB trains in 17 ms on average ($11\times$ GNB’s 1.5 ms, dominated by the dependence matrices); the flagship including the full internal selection takes 84 ms—still below logistic regression (108 ms) and $5\times$ below random forests (418 ms). Prediction cost is identical to GNB (0.4 ms per fold) since scoring (2) multiplies precomputed log-densities by constants. The dynamic variant updates in $O(d^2)$ per sample ($d = 13$: $\sim 40 \mu\text{s}$). These costs position ABD-NB in the same deployment envelope as NB: cheap enough for embedded and streaming use, with none of the structure-search overhead of TAN-style models.

Discussion

What the evidence supports

Assembling the results: (i) when the independence violation is of the redundancy type, ABD-NB converts NB’s structural bias into near-Bayes behaviour, with accuracy gains up to 17 points that persist at every sample size; (ii) when dependence merely inflates confidence, ABD-NB leaves decisions untouched and repairs the probabilities, cutting log-loss by 40–70% on the affected datasets; (iii) when dependence weighting is counterproductive, the selection layer detects it and disengages, at a measured worst-case cost of 0.05 accuracy points; and (iv) the correction itself is estimable at the parametric rate, cheap, and portable to streams. We regard (iii) as as important as (i)–(ii): an adaptive correction that can silently fail on unfavourable data would be of little practical value, and our Digits result quantifies precisely this failure mode for fixed-strength weighting.

Limitations

Four limitations bound the claims. *Redundancy is not relevance*: ABD-NB corrects double-counting but cannot silence

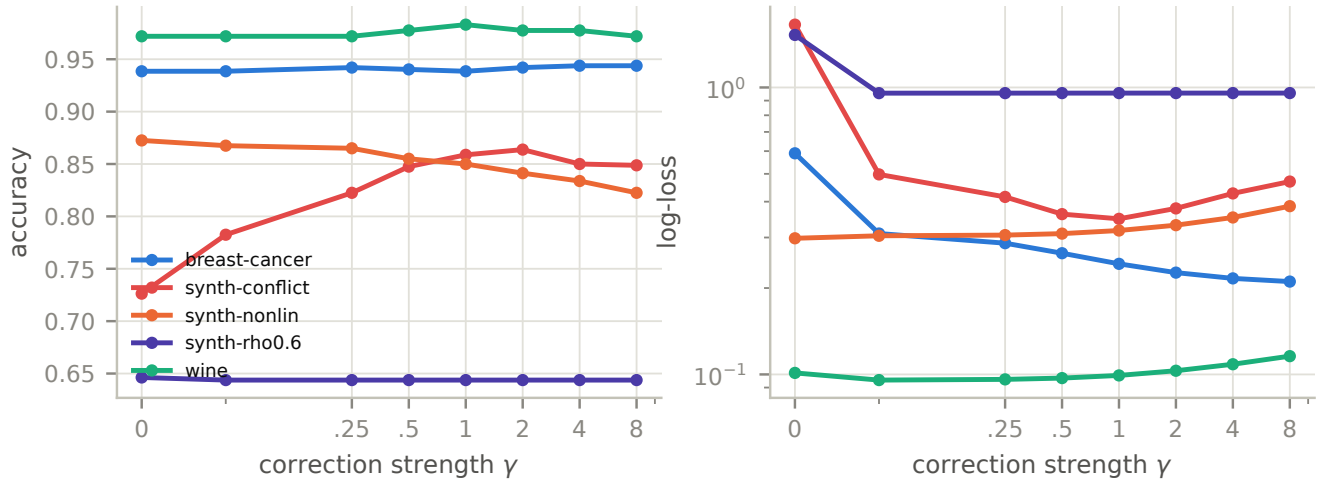


Figure 16. Sensitivity to the correction strength γ (harmonic rule, global weights, 5-fold CV). $\gamma = 0$ is exactly GNB. The dependence-poisoned dataset needs $\gamma \geq 1$; benign datasets are flat; the non-monotone design mildly prefers $\gamma = 0$. The broad plateaus make the internal selection problem easy.

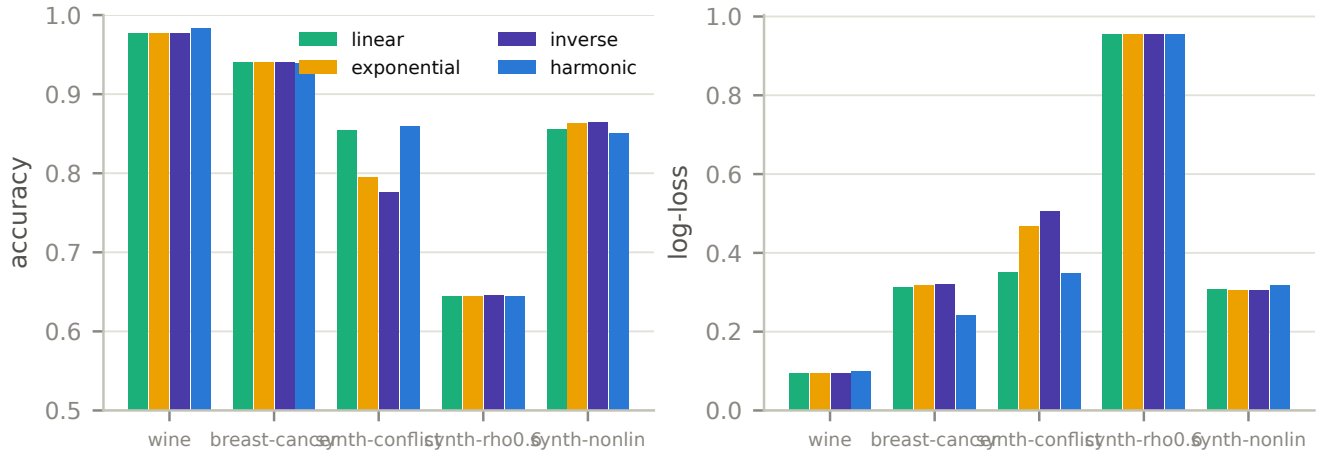


Figure 17. Ablation of the weight link function ($\gamma = 1$, global weights). The graph-based harmonic rule dominates or ties the scalar links on both accuracy (left) and log-loss (right); the gap opens exactly on the adversarial-redundancy design that Theorem 1 addresses.

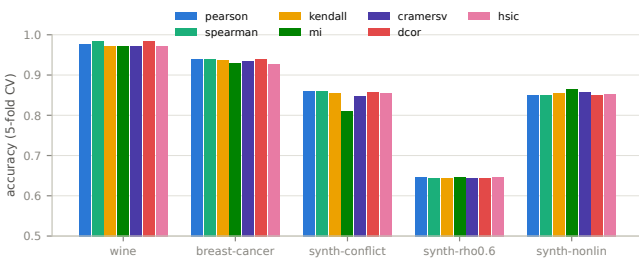


Figure 18. Ablation of the dependence measure ($\gamma = 1$, global weights, 5-fold CV). The framework is largely insensitive to the plug-in estimator; rank correlation (Spearman) matches the kernel and distance measures at far lower cost, and only strongly non-monotone dependence (synth-nonlin) mildly favours mutual information.

irrelevant features—Digits shows relevance weighting (MI-WNB, +4.8 points) succeeding where redundancy weighting has nothing to offer. The two mechanisms are orthogonal and composable; a combined relevance-redundancy weighting with the same safety layer is a natural next step. *Pairwise*

estimation: the dependence graph sees only bivariate structure; higher-order dependence with pairwise-independent margins (e.g. XOR-type interactions) is invisible to every measure we plug in. *Class-specific weights are fragile:* even with the common-scale safeguard (10), per-class matrices estimated from n_c samples are noisy, and the benchmark shows the class-specific ablation losing to the global one on 9 of 15 datasets; we recommend the class-specific option only through the selection layer, never hard-wired. *Gaussian factors:* our experiments instantiate continuous likelihoods; while stages 1–4 are factor-agnostic, the empirical behaviour with multinomial or Bernoulli factors (text, categorical data) remains to be mapped, and rank-based dependence requires adaptation for sparse counts.

Practical guidance

For practitioners we distil: use the flagship configuration (Spearman, harmonic $\kappa=2$, spectral rescaling, internal selection) as shipped; expect material gains whenever the feature pipeline contains engineered ratios, repeated measurements, or sensor clusters; inspect the fitted weights and d_{eff} —a large gap between d and d_{eff} is itself a

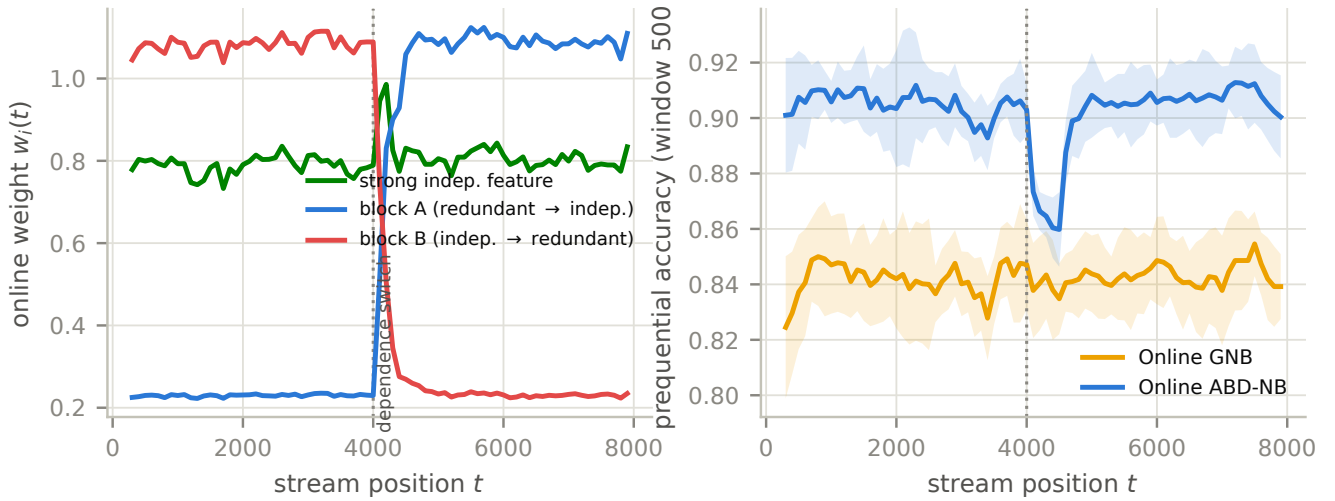


Figure 19. Dependence drift (10 seeds; bands: ± 1 sd). After the switch at $t = 4000$, Online ABD-NB adapts within ~ 500 samples and consistently outperforms online GNB by about 6 prequential accuracy points.

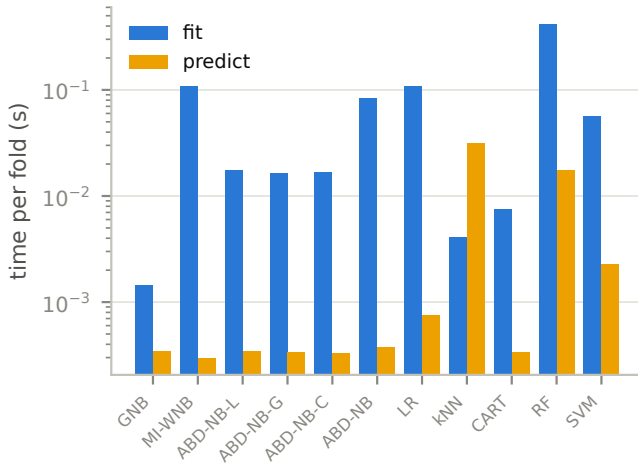


Figure 20. Mean wall-clock fit and predict times per benchmark fold (log scale). The full adaptive ABD-NB trains faster than logistic regression and random forests; prediction cost is indistinguishable from GNB.

useful diagnostic of pipeline redundancy (Breast Cancer: 30 features, $d_{\text{eff}} \approx 14.5$); and if calibrated probabilities feed downstream decisions, prefer ABD-NB over post-hoc recalibration of NB, since it also repairs the decisions where they are distorted.

Conclusion

We introduced Adaptive Bayesian Dependence Naive Bayes, a classifier that treats the conditional-independence assumption of naive Bayes as a matter of degree: class-conditional pairwise dependence is estimated, converted into per-feature likelihood exponents by a harmonic graph rule on the shared-information scale, rescaled to the spectral effective dimension, and applied with a strength selected under a never-worse safeguard. The construction is backed by guarantees—exact posterior recovery under duplication, parametric-rate weight consistency, an explicit overconfidence factor that the weights provably shrink, and asymptotic safety of the tuned classifier—each verified numerically. On 15 datasets against 10 baselines, the method

delivers large accuracy gains under adversarial redundancy, systematic calibration improvements (log-loss down on 10 of 15 datasets, $p = 0.008$), an average accuracy rank above random forests and logistic regression, robust behaviour where dependence weighting cannot help, and a dynamic variant that tracks structural drift in streams—all while preserving the $O(Kd)$ prediction cost, incremental updates, and per-feature interpretability that keep naive Bayes in service.

Beyond the composition with relevance weighting and the extension to discrete factors noted above, we see three research directions. First, *richer graph functionals*: the weight of a feature currently depends on its dependence degree; spectral or community structure of the dependence graph (which our M_{eff} already summarises globally) could allocate weights within blocks by informativeness rather than symmetrically. Second, *finite-sample theory for the selection layer*: Proposition 4 is asymptotic, and oracle-type inequalities quantifying the finite- n cost of adaptivity would sharpen the safety claim. Third, *drift-aware forgetting*: the dynamic variant uses a fixed forgetting factor; coupling it to a change detector on the dependence statistics themselves would shorten the recovery window after abrupt structural change. We hope the framework—keep the architecture, adapt the assumption—proves useful beyond naive Bayes, wherever a simple model’s known misspecification can be measured and dosed rather than merely tolerated.

Data Availability Statement

All datasets used in this study are publicly available: the real datasets ship with scikit-learn (Pedregosa et al. 2011) and originate from the UCI repository (Dua and Graff 2019); the synthetic generators are part of the released code. The complete implementation, experiment scripts, result files, and figure-generation code are available in the ABD-NB repository: <https://github.com/2zalab/ABD-NB-Public>.

Disclosure of Funding

Funding. The authors declare that this research received no external funding. The work presented in this manuscript was conducted as part of the authors' academic research activities without financial support from governmental agencies, private organizations, foundations, or commercial entities.

Competing Interests. The authors declare that they have no competing financial interests or personal relationships that could have influenced the work reported in this manuscript.

References

- Bergsma W (2013) A bias-correction for Cramér's V and Tschuprow's T. *Journal of the Korean Statistical Society* 42(3): 323–328. DOI:10.1016/j.jkss.2012.10.002.
- Bifet A, Holmes G, Kirkby R and Pfahringer B (2010) Moa: Massive online analysis. *Journal of Machine Learning Research* 11(52): 1601–1604. URL <https://jmlr.org/papers/v11/bifet10a.html>.
- Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* 87(1): 52–58. DOI:10.1046/j.1365-2540.2001.00901.x.
- Chickering DM (1996) Learning Bayesian networks is NP-complete. In: Fisher D and Lenz HJ (eds.) *Learning from Data: Artificial Intelligence and Statistics V, Lecture Notes in Statistics*, volume 112. New York: Springer, pp. 121–130. DOI: 10.1007/978-1-4612-2404-4_12.
- Cover TM and Thomas JA (2006) *Elements of Information Theory*. 2 edition. Hoboken, NJ: Wiley-Interscience. ISBN 9780471241959. DOI:10.1002/047174882X.
- Cramér H (1946) *Mathematical Methods of Statistics, Princeton Mathematical Series*, volume 9. Princeton, NJ: Princeton University Press.
- Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7: 1–30. DOI:10.5555/1248547.1248548. URL <https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.html>.
- Domingos P and Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29(2–3): 103–130. DOI:10.1023/A:1007413511361.
- Dua D and Graff C (2019) UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. URL <https://archive.ics.uci.edu>.
- Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10(4): 507–521. DOI:10.1093/biomet/10.4.507. URL <https://doi.org/10.1093/biomet/10.4.507>.
- Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32(200): 675–701. DOI: 10.1080/01621459.1937.10503522.
- Friedman N, Geiger D and Goldszmidt M (1997) Bayesian network classifiers. *Machine Learning* 29(2–3): 131–163. DOI:10.1023/A:1007465528199.
- Gama J, Žliobaitė I, Bifet A, Pechenizkiy M and Bouchachia A (2014) A survey on concept drift adaptation. *ACM Computing Surveys* 46(4): 44:1–44:37. DOI:10.1145/2523813.
- Gneiting T and Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477): 359–378. DOI:10.1198/016214506000001437.
- Gretton A, Bousquet O, Smola AJ and Schölkopf B (2005) Measuring statistical dependence with Hilbert–Schmidt norms. In: Jain S, Simon HU and Tomita E (eds.) *Algorithmic Learning Theory: 16th International Conference, ALT 2005, Proceedings, Lecture Notes in Computer Science*, volume 3734. Springer, pp. 63–77. DOI:10.1007/11564089_7.
- Guo C, Pleiss G, Sun Y and Weinberger KQ (2017) On calibration of modern neural networks. In: Precup D and Teh YW (eds.) *Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research*, volume 70. PMLR, pp. 1321–1330. URL <https://dl.acm.org/doi/10.5555/3305381.3305518>.
- Hall MA (2007) A decision tree-based attribute weighting filter for naive Bayes. *Knowledge-Based Systems* 20(2): 120–126. DOI: 10.1016/j.knosys.2006.11.008.
- Hand DJ and Yu K (2001) Idiot's Bayes—not so stupid after all? *International Statistical Review* 69(3): 385–398. DOI: 10.1111/j.1751-5823.2001.tb00465.x.
- Hoeffding W (1963) Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58(301): 13–30. DOI:10.1080/01621459.1963.10500830.
- Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2): 65–70. DOI:10.2307/4615733.
- Jiang L, Li C, Wang S and Zhang L (2016) Deep feature weighting for naive Bayes and its application to text classification. *Engineering Applications of Artificial Intelligence* 52: 26–39. DOI:10.1016/j.engappai.2016.02.002.
- Jiang L, Zhang L, Yu L and Wang D (2019) Class-specific attribute weighted naive Bayes. *Pattern Recognition* 88: 321–330. DOI: 10.1016/j.patcog.2018.11.032.
- Kendall MG (1938) A new measure of rank correlation. *Biometrika* 30(1–2): 81–93. DOI:10.1093/biomet/30.1-2.81.
- Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 2. Morgan Kaufmann, pp. 1137–1145. URL <https://doi.org/10.5281/zenodo.19712699>.
- Kononenko I (1991) Semi-naive Bayesian classifier. In: Kodratoff Y (ed.) *Machine Learning – EWSL-91, Lecture Notes in Computer Science*, volume 482. Berlin, Heidelberg: Springer, pp. 206–219. DOI:10.1007/BFb0017015.
- Kruskal WH (1958) Ordinal measures of association. *Journal of the American Statistical Association* 53(284): 814–861. DOI: 10.1080/01621459.1958.10501481.
- Lee CH, Gutierrez F and Dou D (2011) Calculating feature weights in naive Bayes with Kullback–Leibler measure. In: *2011 IEEE 11th International Conference on Data Mining*. IEEE, pp. 1146–1151. DOI:10.1109/ICDM.2011.29.
- Lehmann EL (1999) *Elements of Large-Sample Theory*. Springer Texts in Statistics. New York, NY: Springer. ISBN 9780387985954. DOI:10.1007/b98855.
- Lewis DD (1998) Naive (Bayes) at forty: The independence assumption in information retrieval. In: Nédellec C and

- Rouveirol C (eds.) *Machine Learning: ECML-98, Lecture Notes in Computer Science*, volume 1398. Berlin, Heidelberg: Springer, pp. 4–15. DOI:10.1007/BFb0026666.
- Li J and Ji L (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95(3): 221–227. DOI:10.1038/sj.hdy.6800717.
- McCallum A and Nigam K (1998) A comparison of event models for naive Bayes text classification. In: *AAAI-98 Workshop on Learning for Text Categorization*. AAAI Press, pp. 41–48. URL <https://cdn.aaai.org/Workshops/1998/WS-98-05/WS98-05-007.pdf>.
- Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics* 74(4): 765–769. DOI:10.1086/383251.
- Pakdaman Naeini M, Cooper GF and Hauskrecht M (2015) Obtaining well calibrated probabilities using Bayesian binning. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, volume 29. AAAI Press, pp. 2901–2907. DOI:10.1609/aaai.v29i1.9602.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay É (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12(85): 2825–2830. URL <https://www.jmlr.org/papers/v12/pedregosa11a.html>.
- Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Bartlett PL, Schölkopf B and Schuurmans D (eds.) *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, pp. 61–74.
- Rish I (2001) An empirical study of the naive Bayes classifier. In: *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*. pp. 41–46. URL <https://api.semanticscholar.org/CorpusID:61568722>.
- Sahami M (1996) Learning limited dependence Bayesian classifiers. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*. pp. 335–338. URL <https://cdn.aaai.org/KDD/1996/KDD96-061.pdf>.
- Spearman C (1904) The proof and measurement of association between two things. *The American Journal of Psychology* 15(1): 72–101. DOI:10.2307/1412159.
- Székely GJ, Rizzo ML and Bakirov NK (2007) Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35(6): 2769–2794. DOI:10.1214/009053607000000505.
- Touza I, Lazarre W, Balama G, Guidedi K and Kolyang (2025) SHADO: A semantics-preserving hybrid framework for automatic text classification using domain ontology. *Ingénierie des Systèmes d'Information* 30(11): 2961–2976. DOI:10.18280/isi.301114.
- van der Vaart AW (1998) *Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics*, volume 3. Cambridge: Cambridge University Press. ISBN 9780521496032. DOI:10.1017/CBO9780511802256.
- Webb GI, Boughton JR and Wang Z (2005) Not so naive Bayes: Aggregating one-dependence estimators. *Machine Learning* 58(1): 5–24. DOI:10.1007/s10994-005-4258-6.
- Webb GI, Boughton JR, Zheng F, Ting KM and Salem H (2012) Learning by extrapolation from marginal to full-multivariate probability distributions: Decreasingly naive Bayesian classification. *Machine Learning* 86(2): 233–272. DOI:10.1007/s10994-011-5263-6.
- Widmer G and Kubat M (1996) Learning in the presence of concept drift and hidden contexts. *Machine Learning* 23(1): 69–101. DOI:10.1007/BF00116900.
- Wilcoxon F (1945) Individual comparisons by ranking methods. *Biometrics Bulletin* 1(6): 80–83. DOI:10.2307/3001968.
- Yu L, Gan S, Chen Y and He M (2020) Correlation-based weight adjusted naive Bayes. *IEEE Access* 8: 51377–51387. DOI: 10.1109/ACCESS.2020.2973331.
- Zadrozny B and Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, Canada: ACM. ISBN 1-58113-567-X, pp. 694–699. DOI: 10.1145/775047.775151.
- Zaidi NA, Cerquides J, Carman MJ and Webb GI (2013) Alleviating naive Bayes attribute independence assumption by attribute weighting. *Journal of Machine Learning Research* 14(24): 1947–1988. URL <https://jmlr.org/papers/volume14/zaidi13a/zaidi13a.pdf>.