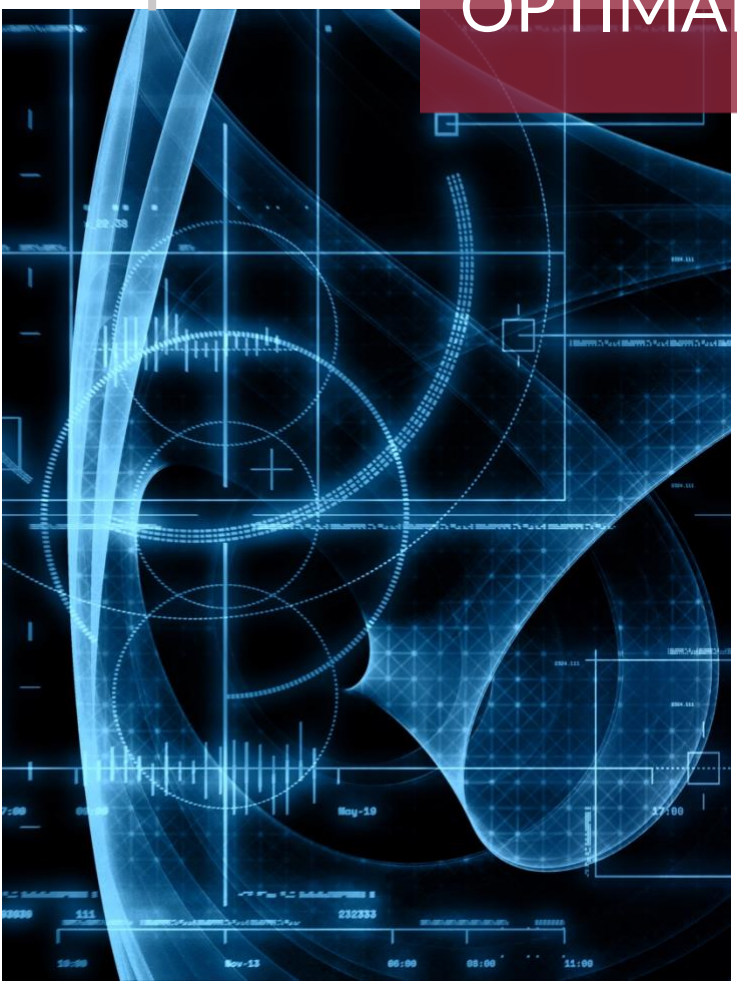
CIRANO
Knowledge into action

OPTIMAL BINARY CLASSIFICATION



RACHIDI KOTCHONI

CS
2025s-30
WORKING PAPER

CIRANO
Knowledge into action

The purpose of the **Working Papers** is to disseminate the results of research conducted by CIRANO research members in order to solicit exchanges and comments. These reports are written in the style of scientific publications. The ideas and opinions expressed in these documents are solely those of the authors.

Les cahiers de la série scientifique visent à rendre accessibles les résultats des recherches effectuées par des chercheurs membres du CIRANO afin de susciter échanges et commentaires. Ces cahiers sont rédigés dans le style des publications scientifiques et n'engagent que leurs auteurs.

CIRANO is a private non-profit organization incorporated under the Quebec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the government of Quebec, and grants and research mandates obtained by its research teams.

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du gouvernement du Québec, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO Partners – *Les partenaires du CIRANO*

Corporate Partners – Partenaires Corporatifs

Autorité des marchés financiers
Banque de développement du Canada
Banque du Canada
Banque Nationale du Canada
Bell Canada
BMO Groupe financier
Caisse de dépôt et placement du Québec
Énergir
Hydro-Québec
Intact Corporation Financière
Investissements PSP
Manuvie
Mouvement Desjardins
Power Corporation du Canada
Pratt & Whitney Canada
VIA Rail Canada

Governmental partners - Partenaires gouvernementaux

Ministère des Finances du Québec
Ministère de l'Économie, de l'Innovation et de l'Énergie
Innovation, Sciences et Développement Économique Canada
Ville de Montréal

University Partners – Partenaires universitaires

École de technologie supérieure
École nationale d'administration publique de Montréal
HEC Montreal
Institut national de la recherche scientifique
Polytechnique Montréal
Université Concordia
Université de Montréal
Université de Sherbrooke
Université du Québec
Université du Québec à Montréal
Université Laval
Université McGill

CIRANO collaborates with many centers and university research chairs; list available on its website. *Le CIRANO collabore avec de nombreux centres et chaires de recherche universitaires dont on peut consulter la liste sur son site web*.

The observations and viewpoints expressed in this publication are the sole responsibility of the authors; they do not represent the positions of CIRANO or its partners. *Les idées et les opinions émises dans cette publication sont sous l'unique responsabilité des auteurs et ne représentent pas les positions du CIRANO ou de ses partenaires.*

ISSN 2292-0838 (online version)

Optimal Binary Classification[*]

Rachidi Kotchoni[†]

Abstract/Résumé

It is shown that the Mean Integrated Square Error (MISE) of a binary classifier is a weighted average of its probabilities of type I (α) and type II errors (β). This provides a foundation for minimizing a linear cost function consisting of a weighted average of α and β to design an optimal classifier. Such a cost function is shown to have the interpretation of a MISE of the classifier under a subjective probability distribution. We derive the closed-form expression of the optimal α for the mean test, provide an equation that can be solved numerically to find the optimal cutoff of the Probit classifier, and illustrate the relevance of the results by simulation. In general, the optimal α for a significance test is different from the conventional 0.05 or 0.01 and the optimal cut-off for probabilistic classifiers deviates from 0.5.

Nous démontrons que l'erreur quadratique moyenne intégrée (EQMI) d'un classificateur binaire est une moyenne pondérée de ses probabilités d'erreurs de type I (α) et de type II (β). Ceci justifie la minimisation d'une fonction de coût linéaire, consistant en une moyenne pondérée de α et β, pour l'obtention d'un classificateur optimal. Une telle fonction de coût peut s'interpréter comme une EQMI du classificateur sous une distribution de probabilité subjective. Nous établissons l'expression analytique du α optimal pour le test de la moyenne, fournissons une équation résoluble numériquement pour la détermination du seuil optimal du classificateur Probit, et illustrons les résultats par simulation. En général, le α optimal pour un test de significativité est différent de 0,05 ou 0,01 utilisé conventionnellement, et le seuil optimal des classificateurs probabilistes est différent de 0,5.

Keywords/Mots-clés: Binary Classification – Hypothesis Testing – Mean Integrated Square Error – Probit – Subjective probabilities / Classification binaire – Tests d'hypothèses – Erreur quadratique moyenne intégrée – Probit – Probabilités subjectives

Pour citer ce document / To quote this document

Kotchoni, R. (2025). Optimal Binary Classification (2025s-30, Cahiers scientifiques, CIRANO.) https://doi.org/10.54932/TUZA3484

[*] Disclaimer: The views expressed in this paper are those of the author and should not be attributed to the institutions to which he is affiliated, nor to the Executive Board or Management of these institutions.

[†] Economist at the International Monetary Fund (IMF). Maitre de Conférences HDR (Associate Professor) at the Université Paris Nanterre (on leave). Associate Researcher at the Centre interuniversitaire de recherche en analyse des organisations (CIRANO). Email: rkotchoni@imf.org / rkotchoni@parisnanterre.fr / rachidi.kotchoni@cirano.qc.ca.

1. Introduction

It is customary for researchers in economics and social sciences to encounter binary choice problems where they must decide whether to act as if one of two statements is true given the available information. In statistical terms, this often boils down to telling whether an observed sample conforms to certain assumptions on the underlying data generating process (Fisher, 1925; Neyman and Pearson, 1933; Romano, 2005)[1]. In this context, the research question may be cast as the prediction of a binary variable *Y* that takes 1 when a given statement is true and 0 otherwise. However, the binary classification problem goes one step beyond the estimation of the conditional distribution of *Y* given the available information (*X*). It entails designing an oracle (or classifier, $\hat{Y}$) that delivers the "best guess" of *Y* based on the realization of *X*.

The cost of predicting $\hat{Y} = 1$ while $Y = 0$ can be quite different from that of predicting $\hat{Y} = 0$ while $Y = 1$. For that reason, the two types of misclassifications are rarely treated symmetrically by the investigator: one outcome is often considered the *default* assumption that is held to be true until "proven" wrong while the *alternative* outcome carries the burden of proof. For instance, a banker will most likely assume that a client is insolvent until the data suggest otherwise. In this example, the choice of default assumption is guided by prudence: the outcome against which misclassification is the costliest is erected as default assumption[2]. In general, the research design should reflect where the investigator wishes to place the burden of proof (Lavergne 2014, pp. 414-415).

With no loss of generality, let us assume that $Y = 0$ is the default assumption and $Y = 1$ the alternative. In a classical hypothesis testing, the default assumption is the null hypothesis. A misclassification against the default outcome (predicting $\hat{Y} = 1$ while Y=0) is an error of type I while a misclassification against the alternative outcome (predicting $\hat{Y} = 0$ while Y=1) is an error

[1] Concrete examples include (i) telling whether an economy is a recession or not (Boldin, 1994; Stock and Watson, 2010; Hamilton, 2011); (ii) predicting whether a firm will go bankrupt or not (Altman, 1968); (iii) enquiring whether a treatment has achieved the intended impact (Imbens and Rubin, 2015), etc.

[2] However, subjective preferences (as opposed to objective costs) may lead a person to consider a theory wrong until proven true (e.g., global warming is a danger to humanity; an accused is guilty until proven innocent).

of type II. The performance of a binary classifier is typically assessed via its probability of type I errors (denoted α) and its probability of type II errors (denoted β). A common approach to tackle binary classification problem when Y is deterministic is the classical Neyman and Pearson (1933)'s hypothesis testing framework. In this paradigm, the best classifier is the one with the lowest probability of type II errors among a set of classifiers with given probability of type I errors (typically, $\alpha = 0.05$ or 0.01). This approach does not directly control the type II errors rate and hence, its ability to detect the alternative hypothesis when it is true may remain undesirably low. Moreover, arbitrarily fixing α at 0.05 or 0.01 does not necessarily deliver a cost-minimizing classifier.

Another approach consists of fitting a model to $Pr(Y = 1|X)$ from which one deduces a classifier that assigns $\hat{Y} = 1$ if $Pr(\hat{Y} = 1|X) > p_0 \in (0,1)$ and $\hat{Y} = 0$ otherwise. This approach may be used when (Y, X) is random and a nontrivial sample of it can be observed naturally or generated via repeated experiments. The conditional probability $Pr(Y = 1|X)$ may be modeled as a Probit or Logit (Cox, 1958) or deduced via Bayes' posterior probability rule (Bishop, 1995). The value assigned to the cut-off p_0 determines the type I and type II errors rates of the classifier. It is customary in the applied literature to assign the value that maximizes the posterior probability $Pr(Y = k|X), k = 0,1$ to $\hat{Y}$, which is equivalent to fixing $p_0 = 0.5$. However, this arbitrary choice does not necessarily result is a cost-minimizing classifier[3].

In the current paper, a unified framework to design optimal binary classifiers is proposed. The proposed approach can be equally applied to hypothesis testing and probabilistic classifications. First, it is shown that the Mean Integrated Square Error (MISE) of a classifier is a weighted average of its type I and type II errors probabilities (α and β). This provides a justification for minimizing a penalty function consisting of the expected costs of misclassification to design of an optimal classifier. Interestingly, the configuration of the misclassification costs implies a subjective probability distribution on the possible outcomes, and the expected costs of misclassification is the

[3] Other popular approaches in the Machine Learning literature include the Support Vector Machines, the Multilayer Perceptron, Discriminant Analysis, etc. For an overview, see Bishop (1995) and Hastie and Tibshirani (1996).

MISE of the classifier (up to a multiplicative constant) under this subjective distribution. We derive the closed-form expression of the optimal α for the mean test (both under the classical and model equivalence approach) and provide a numerical approximation of the optimal cutoff for a Probit classifier. It is found that in general, the optimal α is different from the conventional 0.05 and 0.01 and the optimal cut-off for a Probit classifier is different from 0.5.

The remainder of the paper is organized as follows. Section 2 shows that the MISE of a binary classifier is a weighted average of its type I and type II errors. Based on this result, Section 3 presents a general characterization of the optimal classifier. Sections 4, 5 and 6 specialize the previous result to cases of the classical mean test, the model equivalence mean test and the Probit classification. Each case is supported by a simulation experiment. Section 7 concludes, and an appendix collects the mathematical proofs.

2. The Mean Integrated Square Error of a Binary Classifier

Let us consider testing the null hypothesis $H_0: g(\theta) = 0$ against the alternative hypothesis $H_1: g(\theta) \neq 0$. The deterministic target associated with this test is $Y = I(|g(\theta)| > \eta)$, where $I(\quad)$ is the indicator function that equals 1 if the statement inside the parentheses is true and 0 otherwise; $\theta \in \mathbb{R}^d$ a finite-dimensional parameter of the distribution of X; and $g()$ a mapping from $\mathbb{R}^d$ into $\mathbb{R}$ and $\eta > 0$. The decision rule for this test is generally of the form:

$$\widehat{Y}_t = I\big(S(\theta, X_1, \dots, X_n) \notin (q_1, q_2)\big), \tag{1}$$

where $S(\theta, X_1, \dots, X_n)$ is a test statistic and q_1 and q_2 are chosen to satisfy $Pr\big(\widehat{Y}_t = 1 | H_0\big) = \alpha$. The corresponding probability of type II errors is $\beta(\alpha) = Pr\big(\widehat{Y}_t = 0 | H_1\big)$. Indeed, β is a decreasing function of α Jeffreys (1939).

Alternatively, one may specify the test as: $H_0: g(\theta) \neq 0$ against $H_1: g(\theta) = 0$. This set-up is most useful when the objective of the researcher is to provide evidence in favor of a theory stipulating that $g(\theta) = 0$. The target associated with this test is $Y = I(|g(\theta)| < \eta)$, where η is a small, positive violation margin below which the theory is considered valid (Romano, 2005; Lavergne, 2014). The decision rule becomes:

$$\widehat{Y}_t = I\big(S(\theta, X_1, \dots, X_n) \in (q_1, q_2)\big). \tag{2}$$

The probability of type II errors $\beta = Pr\big(\widehat{Y}_t = 0 | H_1\big)$ is more easily controlled for this test. The implied probability of type I errors is then deduced as $\alpha(\beta) = Pr\big(\widehat{Y}_t = 1 | H_0\big)$.

Hypothesis testing has raised severe criticisms in the literature, most of which revolve around the arbitrariness of the choice of α and the lack of clarity in the interpretation of the verdict of the test. On these grounds, Jeffreys (1939) proposes to minimize a weighted sum of α and β for the purpose of designing the optimal test. Romano (2005) and Lavergne (2014) put forward a model equivalence approach acknowledging that the hypothesis that the researcher wants to "*prove*" must carry the burden of proof. Johnson (2013) and Pericchi, Pereira and Perez (2014) suggest selecting the appropriate value of α based on connections between the Bayesian and frequentist approaches. Gelman and Robert (2014) observed that the appropriate value depends on the context. Other attempts to optimally select α include Miller and Ulrich (2019) and Maier and Lakens (2021). The current paper extends the discussion to probabilistic classifiers, provides a statistical justification for minimizing a weighted sum of the probabilities of type I and type II errors, and derives the closed-form expression of α in specific cases.

When the target Y_t is random and a sample $(Y_t, X_t), t = 1, \dots, T$ can be observed, one can estimate a parametric probabilistic model for $Pr(\mathrm{Y}_t = 1 | X_t)$. For instance:

$$Pr(Y_t = 1 | X_t) = p(\theta, X_t) \tag{3}$$

where $\theta \in \mathbb{R}^d$ is a finite-dimensional parameter. This model may be used to define a classifier that assigns $\widehat{Y}_t = 1$ if $p(\theta, X_t) > p_0$ and $\widehat{Y}_t = 0$ otherwise, for some cut-off $p_0 \in (0,1)$:

$$\widehat{Y}_t = I(p(\theta, X_t) > p_0) \tag{4}$$

It is tempting to assign $p_0 = 0.5$ based on the intuition that $\widehat{Y}_t$ should be the most likely realization of Y_t. While such a choice sounds reasonable, it entails abandoning any attempt to control misclassification rates, which are given by:

$$\alpha = Pr\big(\widehat{Y}_t = 1 | Y_t = 0\big) = E[I(p(\theta, X_t) > p_0) | Y_t = 0] \tag{5}$$

$$\beta = Pr\left(\hat{Y}_t = 0 | Y_t = 1\right) = E[I(p(\theta, X_t) < p_0) | Y_t = 1] \quad (6)$$

To see why the optimal calibration of p_0 is important, let us consider the situation of an epidemiologist trying to assess the prevalence of a disease in a population. The investigator would typically collect data $X_t, t = 1, \dots, T$ on several patients (e.g., from blood tests). Let us assume that the verdict of the test is $\hat{Y}_t = I(p(X_t) > 0)$, where $\hat{Y}_t = 1$ for a patient that is declared "*positive*" if $\hat{Y}_t = 0$ otherwise. Let π denote the probability of $Y = 1$:

$$Pr(Y_t = 1) = \pi \quad (7)$$

The probability of declaring a patient positive is given by:

$$\Pr\left(\hat{Y}_t = 1\right) = \alpha(1 - \pi) + \pi(1 - \beta) = \alpha + \pi[1 - \alpha - \beta]$$

If the disease of interest is rare so that π is very small (e.g. $\pi = 1/10000$), then probability of declaring a patient positive is approximately equal α. In this context, arbitrarily fixing α at 5% or 1% implies predicting $\hat{Y} = 1$ at a rate that is several times larger than the actual prevalence of the disease. This discussion remain valid for criminal trials, with $\hat{Y}_t = 1$ meaning a guilty judgement, $\hat{Y}_t = 0$ a non-guilty judgement and X_t the evidence available to the court.

Let us explore the possibility for selecting α to minimize the statistical precision of $\hat{Y}_t$ as a predictor of Y_t. When $Y = 0$, the Mean Square Error (MSE) of the classifier is given by:

$$MSE\left(\hat{Y}_t | Y_t = 0\right) = E\left[\left(\hat{Y}_t - Y_t\right)^2 | Y_t = 0\right] = (1 - \alpha)(0 - 0)^2 + \alpha(1 - 0)^2 = \alpha.$$

When $Y = 1$, the MSE is:

$$MSE\left(\hat{Y}_t | Y_t = 1\right) = E\left[\left(\hat{Y}_t - Y_t\right)^2 | Y_t = 1\right] = \beta(0 - 1)^2 + (1 - \beta)(1 - 1)^2 = \beta.$$

The Mean Integrated Square Error (MISE) of the classifier is the expected value of the MSE across all possible states. We have:

$$MISE\left(\hat{Y}_t\right) = (1 - \pi)\alpha + \pi\beta$$

This shows that the MISE of any binary classifier is the weighted average of its probability of type I and probability of type II errors. In the deterministic case, repeated observation of Y is not

possible and π is not well-defined. In this case, the MISE may be replaced by the concept of expected disutility associated with misclassification. Indeed, it is shown in the next section that the relative costs of misclassification have the interpretation of subjective probabilities. A classifier with small α tends to have a large β and vice versa. Therefore, the MISE of $\hat{Y}_t$ may be represented as a function of α only:

$$MISE\left(\alpha, \hat{Y}_t\right) = (1-\pi)\alpha + \pi\beta(\alpha) \tag{8}$$

Hence, the optimal classifier may be obtained by minimizing the MISE with respect to α. Alternatively, the MISE may be parameterized in terms of β, leading to:

$$MISE\left(\beta, \hat{Y}_t\right) = (1-\pi)\alpha(\beta) + \pi\beta \tag{9}$$

In this case, the optimal classifier may be obtained by minimizing the MISE with respect to β.

3. Optimal Binary Classifier: A General Result

The expected costs of misclassifications of the classifier described by Equation (9) is given by:

$$C\left(\alpha, \widehat{Y}\right) = (1-\pi)\alpha c_0 + \pi\beta(\alpha)c_1,$$

where $\Pr\left(\hat{Y}=1, Y=0\right) = (1-\pi)\alpha$ and $\Pr\left(\hat{Y}=0, Y=1\right) = \pi\beta$ are respectively the unconditional probabilities of misclassification in the states $Y=0$ and $Y=1$ and c_0 and c_1 are corresponding costs. In the context of bankruptcy prediction, c_0 is the cost of wrongly predicting that a creditworthy agent will default while c_1 is the cost of wrongly predicting that a financially distressed agent is creditworthy (see for example Hsieh, 1993).

This expected cost function can be rewritten as:

$$C\left(\alpha, \widehat{Y}\right) = [(1-\pi)c_0 + \pi c_1][(1-\pi^*)\alpha + \pi^*\beta(\alpha)],$$

where π^* is interpreted as the subjective probability of $Y=1$:

$$\pi^* = \frac{\pi c_1}{(1-\pi)c_0 + \pi c_1} \tag{10}$$

This shows that $C\left(\alpha, \hat{Y}\right)$ is the MISE of $\hat{Y}$ (up to a multiplicative constant) under a risk neutral

distribution. The interpretation of π^* as a subjective probability avoid us the need to resort to a Bayesian interpretation when Y is deterministic (e.g., hypothesis testing).

The multiplicative constant irrelevant for the purpose of determining the optimal classifier. It is therefore dropped so that expected cost function becomes:

$$C(\alpha, \widehat{Y}) = (1 - \pi^*)\alpha + \pi^*\beta(\alpha) \tag{11}$$

where π^* is a subjective probability that embodies the costs of misclassifications as well as the physical probabilities of the states of the world. Alternatively, we may consider parameterizing the cost function in β and write:

$$C(\beta, \widehat{Y}) = (1 - \pi^*)\alpha(\beta) + \pi^*\beta \tag{12}$$

The following result presents a general characterization of the optimal classifier.

Proposition 1. *When the MISE is parameterized in α, the optimal classifier satisfies:*

$$\alpha^* = \beta'^{-1}\left(-\frac{1-\pi^*}{\pi^*}\right), or \tag{13}$$

where β'^{-1} is the reciprocals of the first order derivative of $\beta(\alpha)$. Otherwise, the optimal classifier satisfies

$$\beta^* = \alpha'^{-1}\left(-\frac{\pi^*}{1-\pi^*}\right), \tag{14}$$

where α'^{-1} is the reciprocals of the first order derivative of $\alpha(\beta)$.

Proposition 1 follows immediately from the first order condition for the minimization of the MISE given by either Equations (11) or Equation (12), that is:

$$\beta'(\alpha^*) = -\frac{1 - \pi^*}{\pi^*} \text{ or } \alpha'(\beta^*) = -\frac{\pi^*}{1 - \pi^*}$$

For the solution (13) to be a minimum, the second order derivative of $C(\alpha, \hat{Y})$ with respect to α must be positive. Therefore, it must be the case that $\beta(\alpha)$ is a strictly convex function of α so that $\beta'(\alpha)$ is negative and increasing in α. This means that the smaller π^* is, the larger and negative $-\frac{1-\pi^*}{\pi^*}$ is and the smaller the optimal α^* becomes.

4. Testing the Mean of a Distribution: The Classical Approach

4.1. Optimal Test Design

Let us consider testing whether an unknown mean $\theta = E(X)$ equals a specified value θ_0 (the null hypothesis, H_0) against one of the following alternatives: (i) Unilateral on the left (LH_1): $\theta < \theta_0$; (ii) Unilateral on the right (RH_1): $\theta > \theta_0$; Bilateral (BH_1): $\theta \neq \theta_0$. Let Y be a binary outcome that equals 0 when the null hypothesis is true and 1 under the alternative.

To perform the test, one first constructs a statistic $S(\theta, X)$ whose distribution under the null hypothesis is known. The custom choice is:

$$S(\theta, X) = \frac{\sqrt{T}(\bar{X}-\theta)}{\sigma} \tag{15}$$

where $\bar{X}$ is the average of an independent and identically distributed sample $(X_1, \ldots, X_T)$ of X. Let us assume that X follows a normal distribution with variance σ^2 so that $S(\theta_0, X)$ follows a $N(0,1)$ distribution under the null hypothesis.

Next, one constructs an interval $[\delta_1(\alpha), \delta_2(\alpha)]$ such that:

$$Pr\{S(\theta_0, X) \in [\delta_1(\alpha), \delta_2(\alpha)]\} = 1 - \alpha \tag{16}$$

where $\alpha \in (0,1)$ is desirably small, typically below 10%. There are an infinite number of approaches to design the intervals $[\delta_1(\alpha), \delta_2(\alpha)]$. For instance, we could let:

$$Pr\{S(\theta_0, X) \in]-\infty, \delta_1(\alpha, \eta)]\} = \eta\alpha, \tag{17}$$

$$Pr\{S(\theta_0, X) \in [\delta_2(\alpha, \eta), +\infty[\} = (1-\eta)\alpha. \tag{18}$$

where the dependence of the bounds on $\eta \in (0,1)$ is made explicit.

In subsequent derivations of this section, we maintain that $\alpha < \frac{1}{2}$ so that:

$$\delta_1(\alpha, \eta) = \Phi^{-1}(\eta\alpha) < 0 \tag{19}$$

$$\delta_2(\alpha, \eta) = \Phi^{-1}(1-(1-\eta)\alpha) = -\Phi^{-1}\big((1-\eta)\alpha\big) > 0 \tag{20}$$

where Φ is the cumulative distribution function (CDF) of a standard normal random variable. The

family of binary classifier implied by this test is

$$\widehat{Y}(\eta) = 1(S(\theta_0, X) \notin [\delta_1(\alpha, \eta), \delta_1(\alpha, \eta)]) \tag{21}$$

By design, the probability of type I errors of these classifiers are all equal to α (selected beforehand):

$$Pr\big(\widehat{Y}(\eta) = 1 | Y = 0\big) = \eta\alpha + (1 - \eta)\alpha = \alpha.$$

The probability of type II errors depends on the choices of α and η as well as on the unknown true value of θ. Straightforward calculations show that:

$$\beta(\alpha, \eta, \Delta) = \Phi(\delta_2(\alpha, \eta) - \Delta) - \Phi(\delta_1(\alpha, \eta) - \Delta), \tag{22}$$

where $\delta_1(\alpha, \eta) = \Phi^{-1}(\eta\alpha)$, $\delta_2(\alpha, \eta) = -\Phi^{-1}\big((1 - \eta)\alpha\big)$ and $\Delta = \frac{\sqrt{T}(\theta - \theta_0)}{\sigma}$ is a signal-to-noise ratio measuring the distance between the null and the alternative hypotheses in units of standard deviation of the underlying test statistics.

The Uniformly Most Powerful (UMP) tests are obtained by letting $\eta = 1$ so that $\delta_1(\alpha, 1) = \Phi^{-1}(\alpha)$ and $\delta_2(\alpha, 1) = +\infty$ when the alternative hypothesis is unilateral of type LH_1; and $\eta = 0$ so that $\delta_1(\alpha, 0) = -\infty$ and $\delta_2(\alpha, 0) = \Phi^{-1}(1 - \alpha)$ when the alternative is unilateral of type RH_1 (Neyman and Pearson, 1933, pp. 302-303). When the alternative hypothesis is bilateral (BH_1), the most powerful test (but not uniformly over the rejection region) is obtained by letting $\eta = 0.5$ so that $\delta_1\left(\alpha, \frac{1}{2}\right) = -\delta_2\left(\alpha, \frac{1}{2}\right) = \Phi^{-1}\left(\frac{\alpha}{2}\right)$. The corresponding probabilities of type II errors are respectively given by:

$$\beta(\alpha, 1, \Delta) = 1 - \Phi(\Phi^{-1}(\alpha) - \Delta), \ \Delta < 0 \tag{23}$$

$$\beta(\alpha, 0, \Delta) = \Phi(-\Phi^{-1}(\alpha) - \Delta), \ \Delta > 0 \tag{24}$$

$$\beta\left(\alpha, \frac{1}{2}, \Delta\right) = \Phi\left(-\Phi^{-1}\left(\frac{\alpha}{2}\right) - \Delta\right) - \Phi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right) - \Delta\right), \Delta \in \mathbb{R}\backslash\{0\} \tag{25}$$

Finally, the MISE of the classifier is:

$$MISE\big(\widehat{Y}, \eta\big) = (1 - \pi^*)\alpha + \pi^*\beta(\alpha, \eta, \Delta)$$

where π^* is the subjective probability that is assigned to the outcome $Y = 1$ by the investigator and $\eta \in \left\{0; \frac{1}{2}; 1\right\}$. Let us consider designing the optimal test by minimizing the MISE above with respect to α. The classical test is consistent with the view that $Y = 0$ is the outcome with highest misclassification cost. The assumption $0 < \pi^* \leq 0.5$ is therefore maintained in the Proposition 2 below.

Proposition 2. *Let* $\Delta = \frac{\sqrt{T}(\theta - \theta_0)}{\sigma}$, $\pi^* \in]0,0.5]$. *The optimal* α *for the classical mean test satisfies the following:*

(i) When $\eta = 1$ *so that the alternative hypothesis is* LH_1*:*

$$\alpha^* = \Phi\left(\frac{1}{\Delta} ln\left(\frac{1-\pi^*}{\pi^*}\right) + \frac{\Delta}{2}\right) \tag{26}$$

(ii) When $\eta = 0$ *so that the alternative hypothesis is* RH_1*:*

$$\alpha^* = \Phi\left(-\frac{1}{\Delta} ln\left(\frac{1-\pi^*}{\pi^*}\right) - \frac{\Delta}{2}\right) \tag{27}$$

(iii) When $\eta = 1/2$ *so that the alternative hypothesis is* BH_1*:*

$$\alpha^* = 2\Phi\left(\frac{1}{\Delta} ln\left(\frac{1-\pi^*}{\pi^*} exp\left(\frac{\Delta^2}{2}\right) - sign(\Delta)\sqrt{\left(\frac{1-\pi^*}{\pi^*}\right)^2 exp(\Delta^2) - 1}\right)\right) \tag{28}$$

The probability of type II errors at the optimum is deducted using Equations (23)-(25).

In practice, σ may be unknown and replaced by $\hat{\sigma} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(X_t - \bar{X})^2}$, and $\hat{\Delta} = \frac{\sqrt{T}(\hat{\theta} - \theta_0)}{\hat{\sigma}}$ may be used as a guess for Δ. Accordingly, one should replace Φ by the CDF of Student's *t*-distribution with T-1 degrees of freedom.

Figure 1 shows the behavior of the optimal α^* and the corresponding $\beta(\alpha^*)$ for different values of the signal-to-noise ratio Δ and subjective probability π^*. Whether the test is bilateral or unilateral, α^* and $\beta(\alpha^*)$ decrease to zero fast as $|\Delta|$ converges to infinity. The optimal α falls below 5% for most values of π^* as soon as $|\Delta|$ exceeds 3.5. Indeed that $\alpha = 0.05$ is too high when

$|\Delta| > 3.5$ and $\alpha = 0.01$ becomes a better choice than $\alpha = 0.05$ when $|\Delta| > 5$.

Figure 1. Optimal Classical Mean Test: Analytical Formula

Figure 1.1. Unilateral Test on the Left: Optimal α

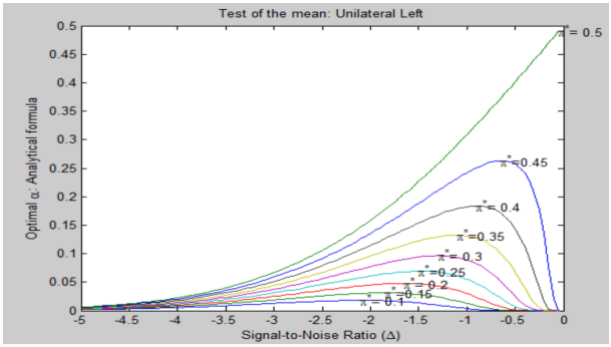


Figure 1.2. Unilateral Test on the Left: Optimal β

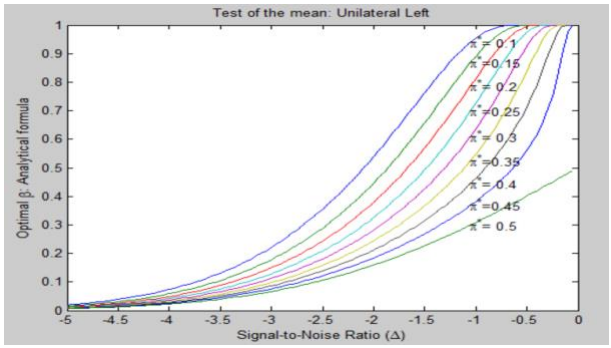


Figure 1.3. Unilateral Test on the Right: Optimal α

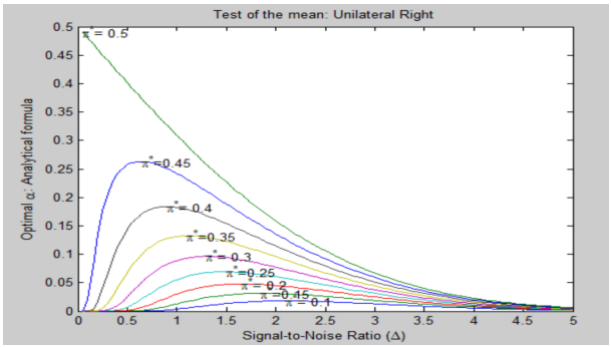


Figure 1.4. Unilateral Test on the Right: Optimal β

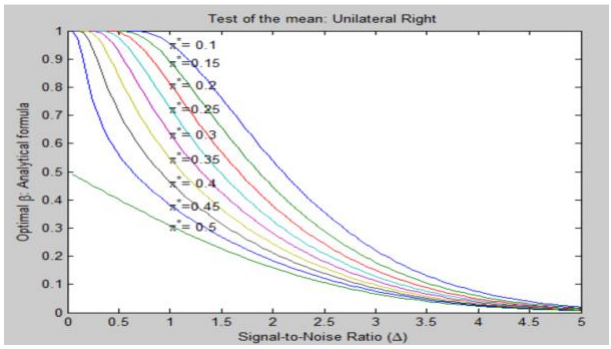


Figure 1.5. Bilateral Test: Optimal α

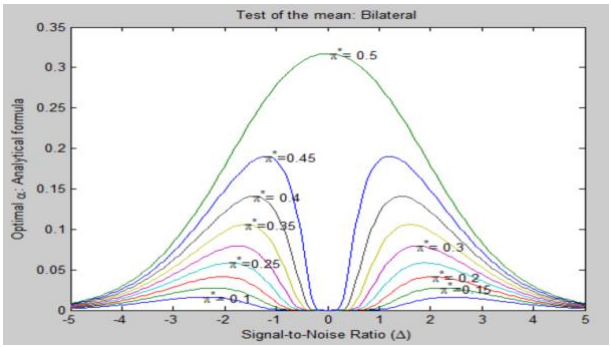
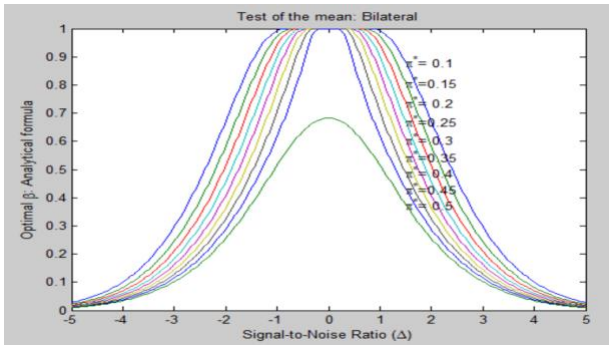


Figure 1.6. Bilateral Test: Optimal β



Second, the optimal α^* is monotonically decreasing in $|\Delta|$ when $\pi^* = 0.5$, and it is non-monotonic for all other $\pi^* \in (0,0.5)$ (Figures 1.1, 1.3 and 1.5). By contrast, $\beta(\alpha^*)$ is monotonically decreasing in $|\Delta|$ for all $\pi^* \in (0,0.5]$ (Figures 1.2, 1.4 and 1.6). On the one hand, the probability of type II errors increases fast to unity as $|\Delta|$ vanishes to zero, and its curve flattens as $|\Delta|$ falls below 0.5. On the other hand, the probability of type II errors decreases to zero as $|\Delta|$ increases to infinity, and its curve flattens as $|\Delta|$ exceeds 3.5. In these two regions, the sensitivity of the MISE to β is close to zero so that minimizing the MISE essentially boils down to minimizing α. However, this explanation holds only when α is given more weight than β in the MISE function.

Third, α^* is increasing in π^* while $\beta(\alpha^*)$ is decreasing in π^*. This result is quite intuitive: the more costly it is to wrongly reject the alternative hypothesis, the larger the optimal probability of type I error is. In the frequentist approach where there is no prior probabilities assigned to the hypotheses, one may consider using $\pi^* = \frac{c_1}{c_0+c_1}$ to express neutrality. In this case, using $\pi^* = 0.5$ is appropriate only if the researcher is further neutral about the costs of misclassification. In a Bayesian framework, $\pi^* = 0.5$ does not necessarily express neutrality as it only tells us that $\pi c_1 = (1-\pi)c_0$. When $\pi^* > 0.5$, we recommend implementing the model equivalent test discussed subsequently. Finally, the analytical expressions of α^* and $\beta(\alpha^*)$ only depend on Δ and π^* as well as on the normality assumption made for the distribution of the test statistics. In particular, the results of Proposition 2 hold for all hypothesis testing exercises where the (appropriately normalized) test statistic follows a pivotal $N(0,1)$ distribution under the null hypothesis (e.g., a regression slope coefficient).

4.2. Monte Carlo Simulation

For this simulation exercise, we draw M=25000 samples of size T=250 from each of the normal distributions with mean $\theta = 1 + \theta_k$ and variance $\sigma^2 = 2$, where $\theta_k = \frac{k\sigma}{10\sqrt{T}}, k = 0,1,\dots,50$. We use each sample to test for the null hypothesis $\theta = 1$ against the bilateral alternative. The simulated samples are of the form $X_t^{(k)} = 1 + \theta_k + \sigma\varepsilon_t, t = 1,\dots,T$ so that the replications use common random numbers across *k*. Hence, the test statistics take of the form:

$$S^{(k)}(\theta, X) = \Delta_k + \sqrt{T}\bar{\varepsilon}, k = 1, \dots, 50.$$

Where $\Delta_k = k/10$ and $\bar{\varepsilon}$ is the sample mean of $\varepsilon_t, t = 1, \dots, T$.

Figure 2. MISE and Optimal Classical Mean Test: Monte Carlo Simulations

Figure 2.1. MISE and Optimal α for $\pi^* = 0.1$

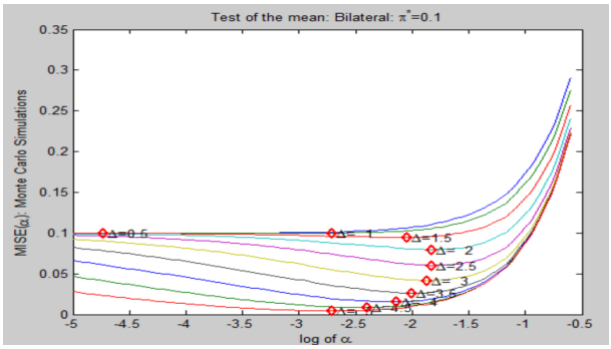


Figure 2.1. MISE and Optimal α for $\pi^* = 0.3$

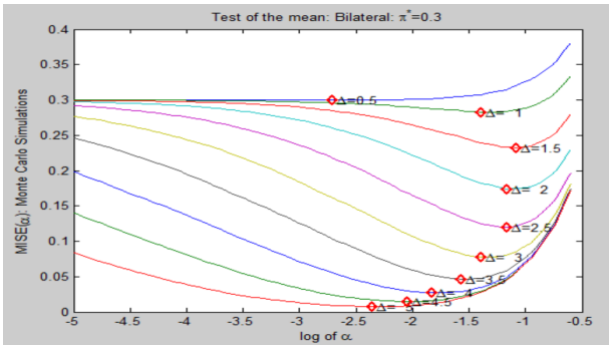


Figure 2.3. MISE and Optimal α for $\pi^* = 0.5$

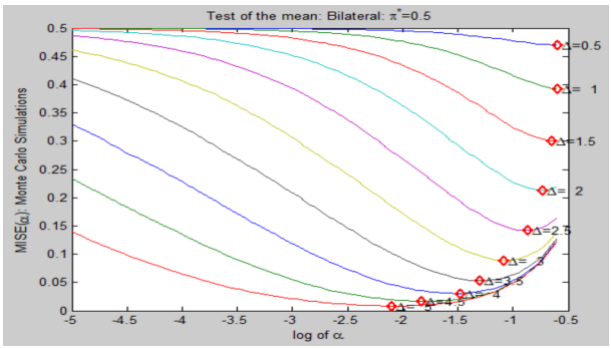
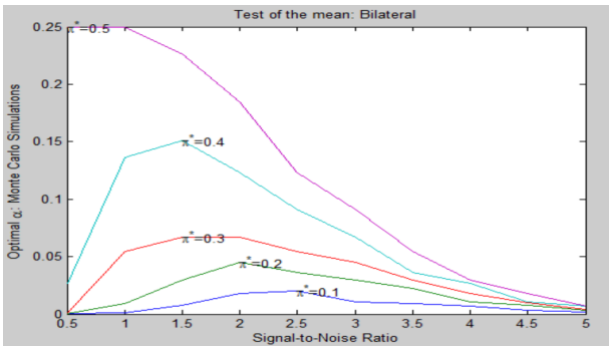


Figure 2.4. Optimal α for $\Delta > 0$



Note: The red diamonds dots mark the minimizer of the MISE.

Next, we calculate the rejection rates $\alpha(\Delta_k)$ over the M Monte Carlo replications. Noting that $\alpha(0)$ is the probability of type I errors while $\beta(\Delta_k) = 1 - \alpha(\Delta_k), k = 1, \dots, 50$ are probabilities of type II errors, we calculate the MISE as:

$$MISE(\Delta_k, \pi^*) = (1 - \pi^*)\alpha(0) + \pi^*\beta(\Delta_k),$$

for $k = 1, \dots, 50$ and $\pi^* = 0.1, 0.2, \dots, 0.5$. Figures 2.1, 2.2 and 2.3 show the MISE for different values of Δ_k and π^*. We see that the degree of convexity of the MISE is increasing in Δ_k.

The optimal $\alpha^* s$ (indicated by the red diamond dots) are collected and plotted against the Δ_k on Figure 2.4. This graph replicates the right quadrant of Figure 1.5, thereby providing a simulation-

based check of the formulas of Proposition 2.

5. Testing the Mean of a Distribution: The Model Equivalence Approach

5.1. Optimal Test Design

Let us now consider the situation where one wishes to test the approximate validity of a restriction on the mean $\theta = E(X)$ of a distribution, for instance:

- Null hypothesis (H_0): $|\theta - \theta_0| > 0$;
- Alternative hypothesis (H_1): $\theta - \theta_0 = 0$.

Let us consider rejecting the null hypothesis as soon as $|\bar{X} - \theta_0| < \eta$, where $\eta \geq 0$ is the tolerated violation margin of the assumption $\theta = \theta_0$. Lavergne (2014) proposes a framework to tackle this test in the general case where the null hypothesis consists of a set of possibly nonlinear restrictions on a finite-dimensional vector of parameters.

Note that $|\bar{X} - \theta_0| < \eta$ if and only if:

$$-(\theta - \theta_0) - \eta < \bar{X} - \theta < -(\theta - \theta_0) + \eta.$$

Therefore, the probability of rejecting the null is:

$$\alpha(\eta, \Delta) = \Phi\left(-\Delta + \frac{\sqrt{T}\eta}{\sigma}\right) - \Phi\left(-\Delta - \frac{\sqrt{T}\eta}{\sigma}\right).$$

The probability of type II errors for this test is:

$$\beta(\eta) = \Pr(|\bar{X} - \theta_0| > \eta | \theta = \theta_0) = 2\Phi\left(-\frac{\sqrt{T}\eta}{\sigma}\right) \quad (29)$$

Or equivalently, $\eta(\beta) = -\frac{\sigma}{\sqrt{T}}\Phi^{-1}\left(\frac{\beta}{2}\right)$. Hence, the MISE of the associated classifier is given by:

$$MISE\left(\widehat{Y}, \beta\right) = (1 - \pi^*)\alpha(\beta, \Delta) + \pi^*\beta$$

where

$$\alpha(\beta, \Delta) = \Phi\left(-\Delta - \Phi^{-1}\left(\frac{\beta}{2}\right)\right) - \Phi\left(-\Delta + \Phi^{-1}\left(\frac{\beta}{2}\right)\right) \quad (30)$$

We consider designing the optimal test by minimizing this MISE with respect to α. The model equivalence mean test is consistent with the view that $Y = 1$ is the outcome with highest misclassification cost. The assumption $0.5 \leq \pi^* < 1$ is therefore maintained in the Proposition 3 below.

Proposition 3. *Let* $\Delta = \frac{\sqrt{T}(\theta-\theta_0)}{\sigma}$ *and* $\pi^* \in [0.5,1[$. *The optimal* α *for the mean test in the model equivalence approach satisfies:*

$$\alpha^* = \Phi\left(-\Delta - \Phi^{-1}\left(\frac{\beta^*}{2}\right)\right) - \Phi\left(-\Delta + \Phi^{-1}\left(\frac{\beta^*}{2}\right)\right) \tag{31}$$

where:

$$\beta^* = 2\Phi\left(\frac{1}{\Delta} ln\left(\frac{\pi^*}{1-\pi^*} exp\left(\frac{\Delta^2}{2}\right) - sign(\Delta)\sqrt{\left(\frac{\pi^*}{1-\pi^*}\right)^2 exp(\Delta^2) - 1}\right)\right) \tag{32}$$

Figure 3 shows the optimal β and the implies α^* for the model equivalence test for different values of Δ and π^*. The curves of β^* the coincide with the right quadrants of Figure 1.5 depicting the optimal probability of type I errors for the classical test. Likewise, the curves of α^* for the model equivalence test coincide with those of the optimal probability of type II errors for the classical test (Figure 1.6). Finally, the scenario based on π^* in the model equivalence approach coincide with the scenario based on $1 - \pi^*$ in the classical approach.

Figure 3. Optimal Model Equivalence Mean Test: Analytical Formula

Figure 3.1. Optimal β

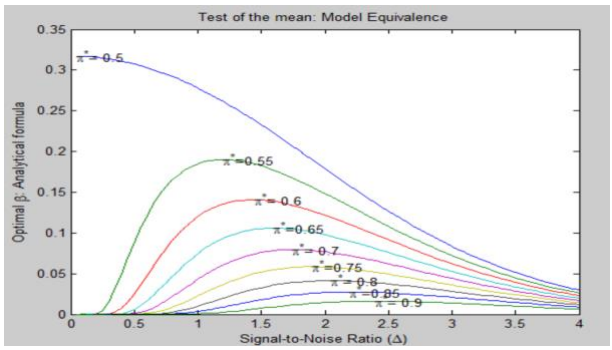


Figure 3.2. Optimal α^*

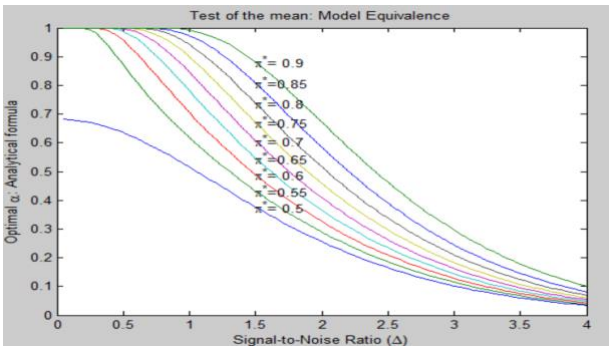


Figure 4. MISE and Optimal Model Equivalence Mean Test: Monte Carlo Simulations

Figure 4.1. MISE and Optimal α for $\pi^* = 0.9$

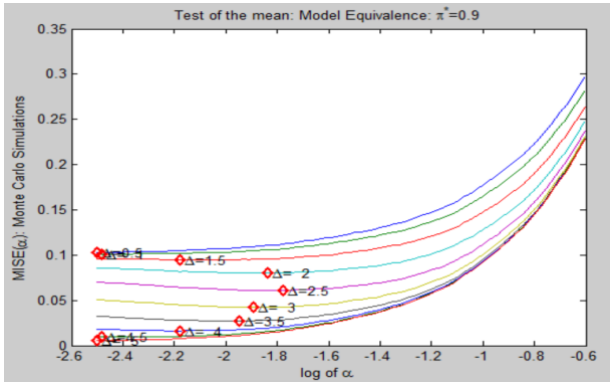


Figure 4.2. MISE and Optimal α for $\pi^* = 0.7$

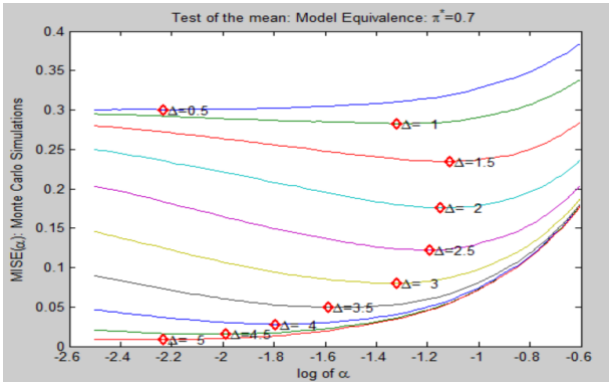


Figure 4.3. MISE and Optimal α for $\pi^* = 0.5$

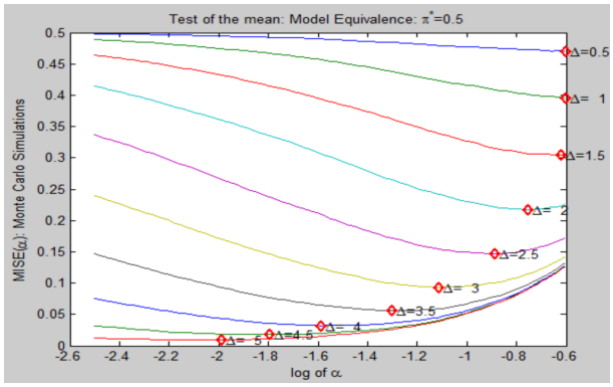
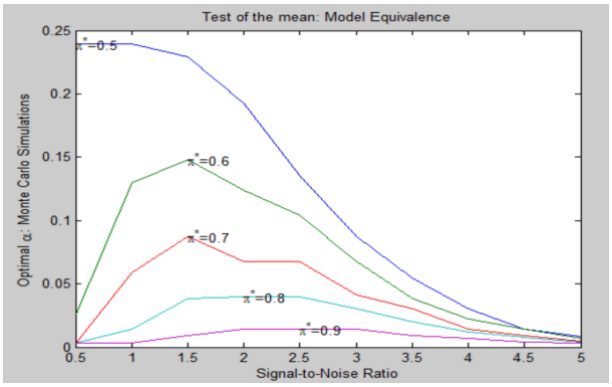


Figure 4.4. Optimal α



5.2. Monte Carlo Simulation

The simulation setup is the same as in the previous subsection except that here we test null hypothesis $\theta \neq 1$ against the alternative $\theta = 1$. We calculate the rejection rates $\alpha(\Delta_k)$ for each k over the M Monte Carlo replications. Given our simulation design, $\beta(1) = 1 - \alpha(1)$ is the probability of type II errors while $\alpha(\Delta_k), k = 1, \dots, 50$ are probabilities of type II errors. We calculate the MISE as:

$$MISE(\Delta_k, \pi^*) = (1 - \pi^*)\alpha(\Delta_k) + \pi^*\beta(1),$$

for $k = 1, \dots, 50$ and $\pi^* = 0.5, 0.2, \dots, 0.9$.

Figures 4.1, 4.2 and 4.3 show the MISE for different values of Δ_k and π^*. We see that the degree of convexity of the MISE is increasing in Δ_k. We see that the MISE curves have the same shape as on Figure 2. Furthermore, Figure 4.4 showing the curves of the optimal α^* replicates Figures 2.4, with the notable difference that here $\pi^* \geq 0.5$.

6. Probabilistic Classification

6.1. Optimal Algorithm Design

Let $p(\theta, X_t) = Pr(Y_t = 1|X_t)$ be a probabilistic model for the binary outcome $Y_t, t = 1, \dots, T$. A popular approach to specify $p(\theta, X_t)$ is:

$$p(\theta, X_t) = \frac{\pi f_1(X_t, \lambda)}{(1-\pi) f_0(X_t, \lambda) + \pi f_1(X_t, \lambda)}. \quad (33)$$

where

$$f_k(x, \lambda) \equiv f(x|Y_t = k), k = 0,1$$

is the conditional density of X_t and $\theta = (\pi, \lambda)$. The Naïve Bayes classifier is obtained by postulating that $f_k(x, \lambda)$ is the density a multivariate normal random variable with a diagonal covariance matrix.

Alternatively, one may consider the logistic regression model so that:

$$p(\theta, X_t) = \frac{1}{1+\exp(X_t\theta)} \tag{34}$$

This approach has a "reduced form" flavor as it avoids the estimation of the "structural" parameters governing the distribution of X_t. It coincides with the SoftMax specification that is very popular in the Machine Learning literature.

Another famous approach to perform binary classification is based on the Probit model, which assumes the existence of a latent variable Z_t such that

$$Z_t = X_t\theta + u_t \tag{35}$$

and $Y_t = 1 \Leftrightarrow Z_t > 0$, where $u_t \sim N(0,1)$. This model implies that

$$p(\theta, X_t) = Pr(X_t\theta + u_t > 0 | X_t) = \Phi(X_t\theta) \tag{36}$$

Upon observing a sample $(Y_t, X_t), t = 1, \dots, T$ and training any probabilistic model using an algorithm of our choice, we may consider defining a binary classifier as $\hat{Y}_t = I(p(\theta, X_t) > p_0)$ for some $p_0 \in (0,1)$. The implied misclassification errors are given by:

$$\alpha = E(I(p(\theta, X_t) > p_0) | Y_t = 0) \text{ and } \beta = E(I(p(\theta, X_t) < p_0) | Y_t = 1).$$

Fixing p_0 arbitrarily at 0.5 may result in undesirably large misclassification rates. Indeed, the formulas above indicate that α and β remain dependent on the distributional properties of X_t. This suggest that there are rooms left to fine-tune the Probit classifier to achieve the best trade-off between the two types of misclassification errors rates.

An empirical counterpart of the MISE may be computed as:

$$\widehat{MISE}\left(\hat{Y}, \pi^*\right) = (1 - \pi^*)\hat{\alpha} + \pi^*\hat{\beta} \tag{37}$$

where $\hat{\pi} = \frac{1}{T}\sum_{t=1}^{T} Y_t$, $\hat{\alpha} = \frac{\frac{1}{T}\sum_{t=1}^{T}(1-Y_t)\hat{Y}_t}{1-\hat{\pi}}$ and $\hat{\beta} = \frac{\frac{1}{T}\sum_{t=1}^{T} Y_t(1-\hat{Y}_t)}{\hat{\pi}}$.

This empirical MISE can be minimized numerically (by a grid search) to obtain the optimal p_0.

Let us consider analyzing the behavior of the optimal cut-off for the Probit model described by Equations (35)-(36). For that purpose, we first need to compute $F_0(z)$ and $F_1(z)$, the CDFs of $\bar{Z}_t =$

$X_t\theta$ conditional on $Y = 0$ and $Y = 1$ respectively. We have:

$$F_0(z) = F(z)\int_{-\infty}^{-z}\frac{\varphi(u)}{F(-u)}du + \Phi(z) \text{ and}$$

$$F_1(z) = F(z)\int_{-z}^{\infty}\frac{\varphi(u)}{1-F(-u)}du - \int_{-z}^{\infty}\frac{F(-u)\varphi(u)}{1-F(-u)}du$$

where $F(z)$ is the unconditional CDF of $\bar{Z}_t$. See proof of Proposition 4 in appendix.

The MISE of the Probit is therefore given by:

$$MISE(\delta, \pi^*) = (1-\pi^*)\alpha(\delta) + \pi^*\beta(\delta) \tag{38}$$

where $\delta = \Phi^{-1}(p_0)$,

$$\alpha(\delta) = 1 - F(\delta)\int_{-\infty}^{-\delta}\frac{\varphi(u)}{F(-u)}du - \Phi(\delta) \tag{39}$$

$$\beta(\delta) = F(\delta)\int_{-\delta}^{\infty}\frac{\varphi(u)}{1-F(-u)}du - \int_{-\delta}^{\infty}\frac{F(-u)\varphi(u)}{1-F(-u)}du \tag{40}$$

We have the following result.

Proposition 4. *The optimal cutoff of the Probit model (35)-(36) for a one-off decision satisfies* $p_0 = \Phi(\delta^*)$*, where* δ^* *solves the following nonlinear equation in* δ*:*

$$\frac{\int_{-\infty}^{-\delta}\frac{\varphi(u)}{F(-u)}du}{\int_{-\infty}^{-\delta}\frac{\varphi(u)}{F(-u)}du + \int_{-\delta}^{\infty}\frac{\varphi(u)}{1-F(-u)}du} = \pi^* \tag{41}$$

In practice, θ must be estimated beforehand and used to compute $\bar{Z}_t$. Likewise, one may consider estimating the CDF of $F(z)$ using kernels (that is, Parzen Window):

$$\hat{F}(z) = \frac{1}{T}\sum_{t=1}^{T}\Phi\left(\frac{z-\hat{Z}_t}{h}\right) \tag{42}$$

Finaly, the quantities involved in Equation (41) can then be approximated by Monte Carlo:

$$\int_{-\infty}^{-\delta}\frac{\varphi(u)}{F(-u)}du \approx \frac{1}{K}\sum_{k=1}^{K}\frac{1(u_k \le -\delta)}{\hat{F}(-u_k)} \tag{43}$$

$$\int_{-\delta}^{\infty}\frac{\varphi(u)}{1-F(-u)}du \approx \frac{1}{K}\sum_{k=1}^{K}\frac{1(u_k \ge -\delta)}{1-\hat{F}(-u_k)} \tag{44}$$

where $(u_1, \dots, u_K)$ are generated from the standard normal distribution.

6.2. Monte Carlo Simulations

We generate T=250 observations from the following process:

$$\bar{Z}_t = \frac{\sigma}{\sqrt{1+\gamma^2}}\left(\sqrt{\frac{\nu-2}{\nu}}\varepsilon_{1,t} + \frac{\gamma}{\sqrt{2\lambda}}\left(\varepsilon_{1,t} - \lambda\right)\right), t = 1, \dots, T$$

where $\varepsilon_{1,t}$ follows Student's t-distribution with ν degree of freedom and $\varepsilon_{1,t}$ follows a Chi-square distribution with λ degrees of freedom. For this exercise, we use $\sigma^2 = 3, \nu = 7, \lambda = 3$ and $\gamma \in \{-1,0,1\}$. This ensures that the variance of $\bar{Z}_t$ is equal to σ^2 and that the distribution of $\bar{Z}_t$ is negatively skewed when $\gamma = -1$, symmetric when $\gamma = 0$ and positively skewed when $\gamma = 1$.

The process $\bar{Z}_t$ is assumed to be latent. The observed processes are generated as $Z_t = \bar{Z}_t + u_t$ and $Y_t = 1(Z_t > 0), t = 1, \dots, T$ where the $u_t s$ are independent and identically distributed (IID) draws from the $N(0,1)$ distribution. This design imply that $Pr(Y_t = 1|\bar{Z}_t) = \Phi(\bar{Z}_t)$. We consider the ideal setup where the latent process $\bar{Z}_t$ is estimated with no error by a Probit model. The MISE implied by this assumption is:

$$\widehat{MISE}(p_0, \pi^*) = (1 - \pi^*)\hat{\alpha} + \pi^*\hat{\beta}$$

where $\hat{\pi} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{Y}_t$, π^* is the subjective probability, $\hat{Y}_t = 1(\Phi(\bar{Z}_t) > p_0)$,

$$\hat{\alpha}(p_0) = \frac{\frac{1}{T}\sum_{t=1}^{T}(1 - \mathrm{Y}_t)\hat{Y}_t}{1 - \hat{\pi}} \text{ and } \hat{\beta}(p_0) = \frac{\frac{1}{T}\sum_{t=1}^{T}\mathrm{Y}_t\left(1 - \hat{Y}_t\right)}{\hat{\pi}}.$$

We simulate M=25000 trajectories of the processes described above. For each trajectory, we compute the empirical MISE of the Probit classifier above and identify the cutoff that minimizes it. The corresponding averages over the M replications are labeled "*Pure Monte Carlo*" subsequently. We also compute the theoretical MISE and optimal cutoff based on Proposition 4. For a particular trajectory, the results are conditional on the path $\bar{Z}_t, t = 1, \dots, T$. The corresponding results are labelled "*Analytical Formula and Monte Carlo*" subsequently.

Figure 5: True and Simulated Mean Integrated Square Error of the Probit

Figure 5.1. True MISE under Negative skewness

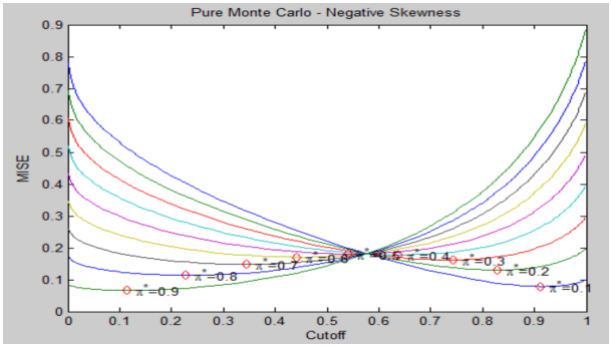


Figure 5.2. Simulated MISE under Negative skewness

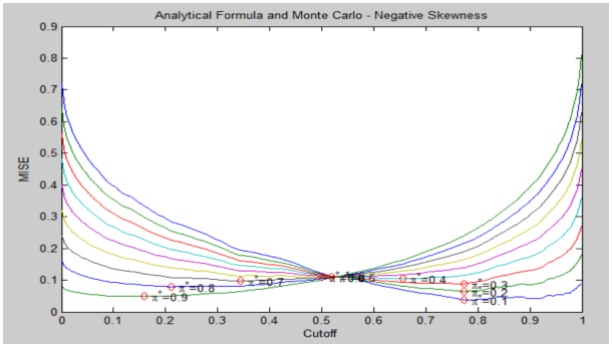


Figure 5.3. True MISE under Symmetric Distribution

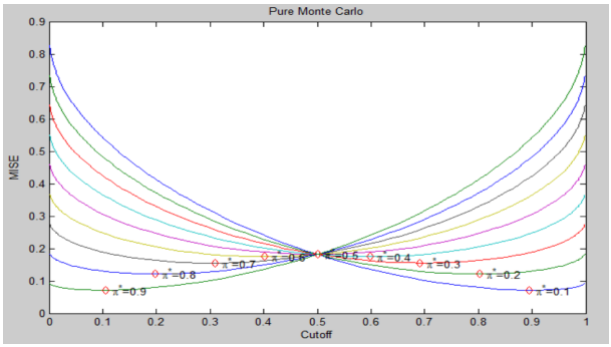


Figure 5.4. Simulated MISE under Symmetric Distribution

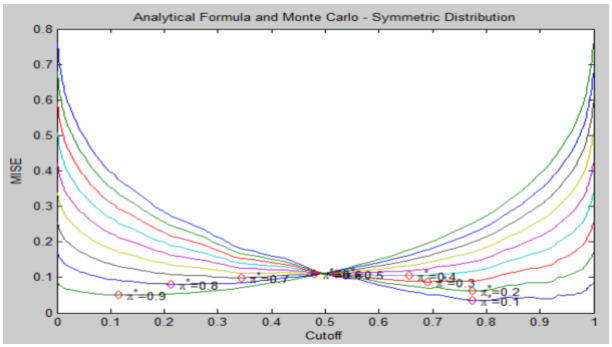


Figure 5.5. True MISE under Positive skewness

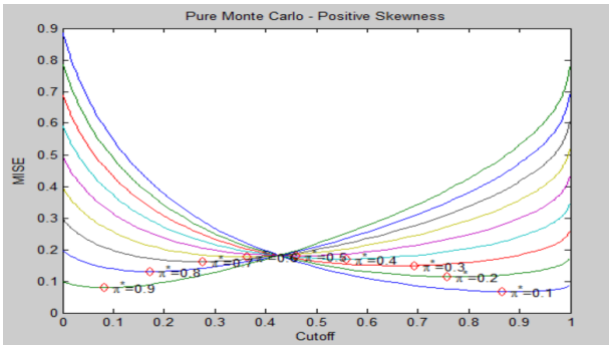
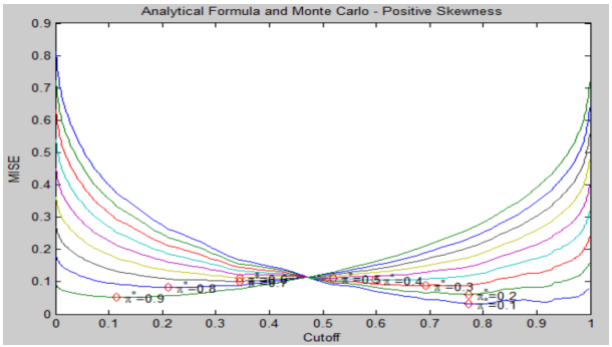


Figure 5.6. Simulated MISE under Positive skewness



Figures 5.1, 5.3 and 5.5 show the true MISE computed using Equation (38) while Figures 5.2, 5.4 and 5.6 show the simulated MISE. The similarity of the two MISEs confirms the correctness of the theoretical formula. We also note that the optimal cutoffs (marked by the red diamond dots) are uniformly tilted to the right (resp. to the left) when the distribution of $\bar{Z}_t$ is negatively skewed (resp. positively skewed) compared to the symmetric distribution. This is confirmed by Figure 6,

which shows the plots of the average optimal cutoffs against π^*. This result proves that the optimal cutoff $p_0^* = \Phi(\delta^*)$ is sensitive to the distributional properties of $\bar{Z}_t$. In fact, $p_0 = 0.5$ is optimal only in the very special case where the distribution of $\bar{Z}_t$ is symmetric and $\pi^* = 0.5$.

Figure 6 suggests that p_0^* is almost linearly decreasing in π^*, with a slope that is smaller than one is absolute value. As seen on Figure 5 above, negative skewness causes the curve to drift to the right while positive skewness induces a drift to the left. The slopes of the curves are not affected by the drift.

Finally, Figure 7 show the optimal α^* and $\beta(\alpha^*)$ as functions of the subjective probability π^*. Interestingly, the optimal error rates are less sensitive to the skewness of $\bar{Z}_t$ than are the optimal cutoffs. This result should not be surprising: the optimal cutoff adjust to the distributional properties of $\bar{Z}_t$ to deliver this optimal error rates. Note that α^* and $\beta(\alpha^*)$ are larger in the Pure Monte Carlo exercise due to the finiteness of the sample. This indicates that the empirical MISE given at Equation (37) should be preferred if one wishes to account for finite sample correction.

7. Conclusion

We show that the Mean Integrated Square Error (MISE) of a binary classifier is a weighted average of its probabilities of Type I and Type II error (α and β), where the weights are the unconditional probabilities ($1-\pi$ and π) of the outcomes. This provides a justification for minimizing a weighted average of the two error rates to design the optimum classifier. Any choice of weights imply a particular subjective probability distribution for the outcomes ($1-\pi^*$ and π^*), and the corresponding weighted average of α and β is proportional to the MISE under this subjective distribution. We derive closed-form expressions for the optimal α for standard significance tests on the mean of a distribution as well as for the Probit classifier. Simulation experiments confirm the relevance of optimally selecting α. The optimal α rarely coincide with 0.05 and the optimal cut-off of the Probit model is generally different from 0.5.

Figure 6: Optimal Cutoffs for the Probit Classifier

Figure 6.1. Average Optimal Cutoffs estimated by Pure Monte Carlo

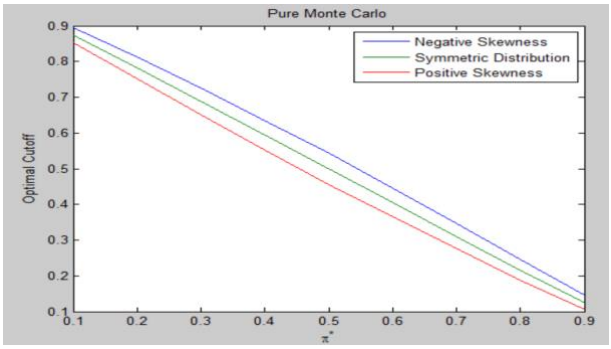


Figure 6.2. Average Optimal Cutoffs estimated by Analytical Formula and Monte Carlo

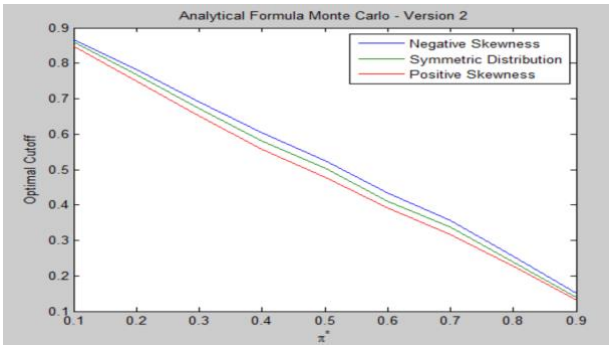


Figure 7

Figure 7.1. Optimal α Pure Monte Carlo

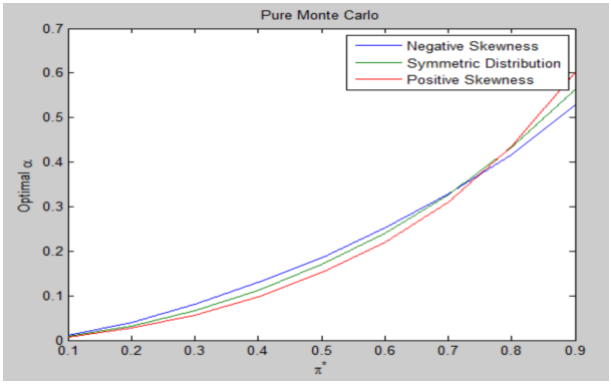


Figure 7.2. Optimal α Analytical Formula and Monte Carlo

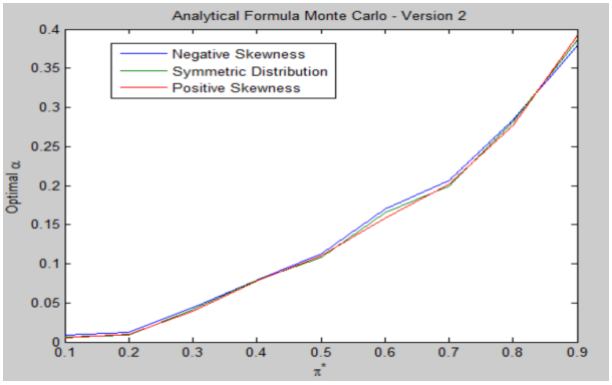


Figure 7.3. Optimal β Pure Monte Carlo

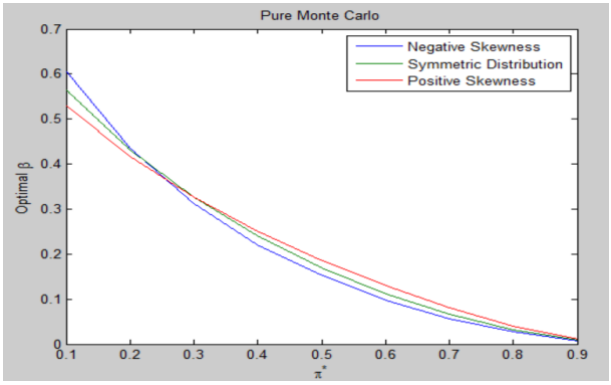
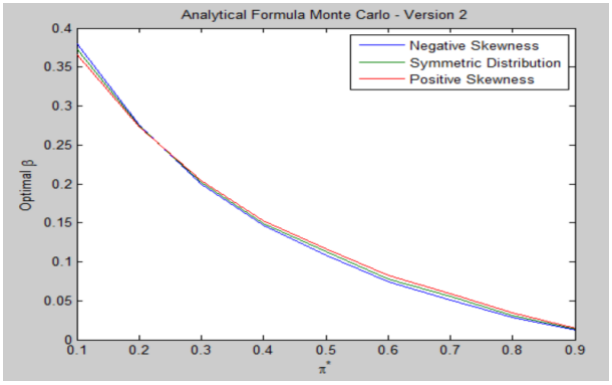


Figure 7.4. Optimal β Analytical Formula and Monte Carlo



References

Altman, E.I. 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance Vol XXIII (04) pp. 589-609.

Bishop, C. 1995. [book]. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Boldin, M.D. 1994. Dating Turning Points in the Business Cycle. *The Journal of Business*, Vol. 67, No. 1 (Jan 1994), pp. 97-131

Cox, D.R. 1958. The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 20, No. 2(1958), pp. 215-242

Fisher, R.A. 1925. Statistical Methods for Research Workers (10th Edition). Edinburgh: Olivier & Boyd.

Gelman, A. and C. P. Robert. 2014. Revised Evidence for Statistical Standards. Proceedings of the National Academy of Sciences of the United States of America, 111, E1933-E1933.

Hamilton, J.D. 2011. Calling recession in Real time. *International Journal of Forecasting*, Vol. 27 (4), pp. 1006-1026

Hastie, T. and R. Tibshirani. 1996. Discriminant Analysis by Gaussian Mixtures. *Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 58, No. 1(1996), pp. 155-176.

Hsieh, S-J. 1993. A note on the optimal cutoff point in bankruptcy prediction models. Journal of Business Finance and Accounting, 20(3), April 1993, 0306-686X.

Imbens, G.W. and D.B. Rubin. 2015. Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge University Press, New York.

Jeffreys, H. 1961. Theory of Probability (Oxford Univ Press, Oxford, UK).

Johnson V. E. 2013. Revised Standards for Statistical Evidence. Proceedings of the National Academy of Sciences of the United States of America, 110, pp. 19313-19317.

Lavergne, P. .2014. Model equivalence tests in a parametric framework. *Journal of Econometrics* 178 (2014) 414–425

Maier M. and D. Lakens. 2021. Justify Your Alpha: A Primer on Two Practical Approaches. https://doi.org/10.31234/osf.io/ts4r6.

Miller J. and R. Ulrich. 2019. The quest for an optimal alpha. PLoS ONE 14(1): e0208631. https://doi.org/10.1371/journal.pone.0208631.

Neyman, J. and E.S. Pearson. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philosophical Transactions of the Royal Society of London. Serie A, Containing Papers of a Mathematical or Physical Caracter, Vol 231 (1933), pp. 289-337

Pericchi L., Pereira, C. A. and M.-E. Perez. 2014. Adaptive Revised Standards for Statistical Evidence. Proceedings of the National Academy of Sciences of the United States of America, 111, E1935-E1935.

Stock J.H. and M.W. Watson. 2010. Indicators for Dating Business Cycles: Cross-History Selection and Comparisons. *American Economic Review*: Papers & Proceedings 100 (May 2010): pp. 16–19

Appendix: Mathematical Proof.

Proof of Proposition 2. The expression of $\beta(\alpha, \eta, \theta_0, \theta)$ is:

$$\beta(\alpha, \eta, \Delta) = \Phi\big(-\Phi^{-1}\big((1-\eta)\alpha\big) - \Delta\big) - \Phi(\Phi^{-1}(\eta\alpha) - \Delta),$$

where $\Delta = \frac{\sqrt{T}(\theta-\theta_0)}{\sigma}$. The derivative of $\beta(\alpha, \eta, \theta_0, \theta)$ with respect to α is given by:

$$\beta'(\alpha, \eta, \Delta) = -(1-\eta)\frac{\varphi\big(-\Phi^{-1}\big((1-\eta)\alpha\big) - \Delta\big)}{\varphi\left(\Phi^{-1}\big((1-\eta)\alpha\big)\right)} - \eta\frac{\varphi(\Phi^{-1}(\eta\alpha) - \Delta)}{\varphi\big(\Phi^{-1}(\eta\alpha)\big)},$$

where $\varphi(z) = \frac{1}{2}\exp(-z^2/2)$.

Case $\boldsymbol{\eta = 1}$: The alternative hypothesis is unilateral on the left so that $\Delta \leq 0$. The optimal α solves:

$$\frac{\varphi(\Phi^{-1}(\alpha) - \Delta)}{\varphi\big(\Phi^{-1}(\alpha)\big)} = \frac{1-\pi^*}{\pi^*} \Leftrightarrow \exp\left[-\frac{1}{2}(\delta - \Delta)^2 + \frac{1}{2}\delta^2\right] = \frac{1-\pi^*}{\pi^*},$$

where $\delta = \Phi^{-1}(\alpha)$. Hence:

$$\delta\Delta - \frac{1}{2}\Delta^2 = \ln\left(\frac{1-\pi^*}{\pi^*}\right) \Leftrightarrow \delta^* = \frac{1}{\Delta}\ln\left(\frac{1-\pi^*}{\pi^*}\right) + \frac{\Delta}{2},$$

$$\Leftrightarrow \alpha^* = \Phi\left(\frac{1}{\Delta}\ln\left(\frac{1-\pi^*}{\pi^*}\right) + \frac{\Delta}{2}\right).$$

Case $\boldsymbol{\eta = 0}$: Alternative hypothesis is unilateral on the right so that $\Delta \geq 0$. The optimal α solves:

$$-\frac{\varphi(-\delta - \Delta)}{\varphi(\delta)} = -\frac{1-\pi^*}{\pi^*} \Leftrightarrow \exp\left[-\frac{1}{2}(-\delta - \Delta)^2 + \frac{1}{2}\delta^2\right] = \frac{1-\pi^*}{\pi^*}$$

$$\Leftrightarrow -\delta\Delta - \frac{1}{2}\Delta^2 = \ln\left(\frac{1-\pi^*}{\pi^*}\right) \Leftrightarrow \delta^* = -\frac{1}{\Delta}\ln\left(\frac{1-\pi^*}{\pi^*}\right) - \frac{\Delta}{2}$$

$$\Leftrightarrow \alpha^* = \Phi\left(-\frac{1}{\Delta}\ln\left(\frac{1-\pi^*}{\pi^*}\right) - \frac{\Delta}{2}\right).$$

Case $\boldsymbol{\eta = 1/2}$: Alternative hypothesis is bilateral. Let $\delta = \Phi^{-1}\left(\frac{\alpha}{2}\right)$. The optimal α solves:

$$\frac{\varphi(-\delta-\Delta)+\varphi(\delta-\Delta)}{2\varphi(\delta)}=\frac{1-\pi^*}{\pi^*},$$

where $\varphi(x)=\frac{1}{2}\exp(-x^2/2)$. Hence:

$$\exp\left[-\frac{1}{2}(-\delta-\Delta)^2+\frac{1}{2}\delta^2\right]+\exp\left[-\frac{1}{2}(\delta-\Delta)+\frac{1}{2}\delta^2\right]=\frac{2(1-\pi^*)}{\pi^*}$$

$$\Leftrightarrow \exp(2\delta\Delta)-\frac{2(1-\pi^*)}{\pi^*}\exp\left(\frac{\Delta^2}{2}\right)\exp(\delta\Delta)+1=0.$$

Let $\tilde{\delta}=\exp(\delta\Delta)$, so that:

$$\tilde{\delta}^2-\frac{2(1-\pi^*)}{\pi^*}\exp\left(\frac{\Delta^2}{2}\right)\tilde{\delta}+1=0.$$

The (modified) determinant of this quadratic equation is:

$$d=\left(\frac{1-\pi^*}{\pi^*}\right)^2\exp(\Delta^2)-1.$$

To move forward, I need to verify that this determinant is positive:

$$\left(\frac{1-\pi^*}{\pi^*}\right)^2\exp(\Delta^2)>1.$$

This inequality is trivial because $\pi^*<1/2$.

The roots of the quadratic equation are therefore given by:

$$\tilde{\delta}=\frac{1-\pi^*}{\pi^*}\exp\left(\frac{\Delta^2}{2}\right)\pm\sqrt{\left(\frac{1-\pi^*}{\pi^*}\right)^2\exp(\Delta^2)-1}.$$

These two roots are both positive. Hence:

$$\exp(\delta\Delta)=\frac{1-\pi^*}{\pi^*}\exp\left(\frac{\Delta^2}{2}\right)\pm\sqrt{\left(\frac{1-\pi^*}{\pi^*}\right)^2\exp(\Delta^2)-1}.$$

It is easily shown that these two roots are inverses of each other. One root is greater than 1 and the other one lies between 0 and 1.

$$\frac{1-\pi^*}{\pi^*}\exp\left(\frac{\Delta^2}{2}\right)+\sqrt{\left(\frac{1-\pi^*}{\pi^*}\right)^2\exp(\Delta^2)-1}>1,$$

$$0<\frac{1-\pi^*}{\pi^*}\exp\left(\frac{\Delta^2}{2}\right)-\sqrt{\left(\frac{1-\pi^*}{\pi^*}\right)^2\exp(\Delta^2)-1}<1.$$

As $\delta^*=\Phi^{-1}\left(\frac{\alpha^*}{2}\right)<0$ by definition, $\exp(\delta^*\Delta)>1$ if $\Delta<0$, and $\exp(\delta^*\Delta)<1$ otherwise. Therefore, we have:

$$\exp(\delta^*\Delta)=\frac{1-\pi^*}{\pi^*}exp\left(\frac{\Delta^2}{2}\right)-\text{sign}(\Delta)\sqrt{\left(\frac{1-\pi^*}{\pi^*}\right)^2\exp(\Delta^2)-1}$$

$$\Leftrightarrow\delta^*=\frac{1}{\Delta}\ln\left(\frac{1-\pi^*}{\pi^*}\exp\left(\frac{\Delta^2}{2}\right)-\text{sign}(\Delta)\sqrt{\left(\frac{1-\pi^*}{\pi^*}\right)^2\exp(\Delta^2)-1}\right)$$

$$\Leftrightarrow\alpha^*=2\Phi\left(\frac{1}{\Delta}\ln\left(\frac{1-\pi^*}{\pi^*}\exp\left(\frac{\Delta^2}{2}\right)-\text{sign}(\Delta)\sqrt{\left(\frac{1-\pi^*}{\pi^*}\right)^2\exp(\Delta^2)-1}\right)\right).$$

QED.■

Proof of Proposition 3. The MISE of the equivalence test is given by:

$$MISE(\widehat{Y},\beta)=(1-\pi^*)\left(\Phi\left(-\Delta-\Phi^{-1}\left(\frac{\beta}{2}\right)\right)-\Phi\left(-\Delta+\Phi^{-1}\left(\frac{\beta}{2}\right)\right)\right)+\pi^*\beta.$$

Taking the derivative with respect to β and equating to zero yields:

$$-\frac{\varphi(-\Delta-\delta)}{\varphi(\delta)}-\frac{\varphi(-\Delta-\delta)}{\varphi(\delta)}=\frac{2\pi^*}{1-\pi^*},$$

where $\delta=\Phi^{-1}\left(\frac{\beta}{2}\right)$ and $\varphi(x)=\frac{1}{2}\exp(-x^2/2)$. Thus:

$$\exp\left(\frac{\delta^2}{2}-\frac{(-\Delta-\delta)^2}{2}\right)+\exp\left(\frac{\delta^2}{2}-\frac{(-\Delta+\delta)^2}{2}\right)=\frac{2\pi^*}{1-\pi^*}$$

$$\Leftrightarrow \exp(2\delta\Delta) - \frac{2\pi^*}{1-\pi^*}\exp\left(\frac{\Delta^2}{2}\right)\exp(\delta\Delta) + 1 = 0.$$

Let $\tilde{\delta} = \exp(\delta\Delta)$, so that:

$$\tilde{\delta}^2 - \frac{2\pi^*}{1-\pi^*}\exp\left(\frac{\Delta^2}{2}\right)\tilde{\delta} + 1 = 0.$$

The (modified) determinant of this quadratic equation is:

$$d = \left(\frac{\pi^*}{1-\pi^*}\right)^2 \exp(\Delta^2) - 1.$$

To move forward, I need to verify that this determinant is positive:

$$\left(\frac{\pi^*}{1-\pi^*}\right)^2 \exp(\Delta^2) \geq 1 \Leftrightarrow \frac{1}{1+\exp\left(\frac{\Delta^2}{2}\right)} < 1/2 \leq \pi^*.$$

This is basically saying that π^* must not be too small for the equivalent test to be justified. The roots of the quadratic equation are given by:

$$\tilde{\delta} = \frac{\pi^*}{1-\pi^*}\exp\left(\frac{\Delta^2}{2}\right) \pm \sqrt{\left(\frac{\pi^*}{1-\pi^*}\right)^2 \exp(\Delta^2) - 1}.$$

These two roots are both positive. Hence:

$$\exp(\delta\Delta) = \frac{\pi^*}{1-\pi^*}\exp\left(\frac{\Delta^2}{2}\right) \pm \sqrt{\left(\frac{\pi^*}{1-\pi^*}\right)^2 \exp(\Delta^2) - 1}.$$

It is easily shown that these two roots are inverses of each other. As $\delta^* = \Phi^{-1}\left(\frac{\beta^*}{2}\right) < 0$, $\exp(\delta^*\Delta) > 1$ if $\Delta < 0$, and $\exp(\delta^*\Delta) < 1$ otherwise. Therefore, we have:

$$\exp(\delta^*\Delta) = \frac{\pi^*}{1-\pi^*}\exp\left(\frac{\Delta^2}{2}\right) - \text{sign}(\Delta)\sqrt{\left(\frac{\pi^*}{1-\pi^*}\right)^2 \exp(\Delta^2) - 1}$$

$$\Leftrightarrow \delta^* = \frac{1}{\Delta}\ln\left(\frac{\pi^*}{1-\pi^*}\exp\left(\frac{\Delta^2}{2}\right) - \text{sign}(\Delta)\sqrt{\left(\frac{\pi^*}{1-\pi^*}\right)^2\exp(\Delta^2) - 1}\right)$$

$$\Leftrightarrow \beta^* = 2\Phi\left(\frac{1}{\Delta}\ln\left(\frac{\pi^*}{1-\pi^*}\exp\left(\frac{\Delta^2}{2}\right) - \text{sign}(\Delta)\sqrt{\left(\frac{\pi^*}{1-\pi^*}\right)^2\exp(\Delta^2) - 1}\right)\right).$$

QED. ■

Proof of Proposition 4. Let us first find the CDF $F_1(z)$ of $\bar{Z}_t = X_t\theta$ conditional on $Y_t = 1$. We have:

$$Pr(\bar{Z}_t < z|\bar{Z}_t > -u_t, u_t) = \frac{Pr(-u_t < \bar{Z}_t < z|u_t)}{Pr(\bar{Z}_t > -u_t|u_t)}$$

$$= \begin{cases} 0 & \text{if } z < -u_t \\ \dfrac{F(z) - F(-u_t)}{1 - F(-u_t)} & \text{otherwise} \end{cases}$$

$$\Rightarrow F_1(z) = Pr(\bar{Z}_t < z|\bar{Z}_t > -u_t) = \int_{-z}^{\infty}\frac{F(z) - F(-u)}{1 - F(-u)}\varphi(u)du$$

$$= F(z)\int_{-z}^{\infty}\frac{\varphi(u)}{1 - F(-u)}du - \int_{-z}^{\infty}\frac{F(-u)\varphi(u)}{1 - F(-u)}du.$$

The corresponding PDF is:

$$f_1(z) = f(z)\int_{-z}^{\infty}\frac{\varphi(u)}{1 - F(-u)}du.$$

The probability of type II errors for the Probit classifier is given by:

$$\beta = Pr(\bar{Z}_t < \delta|\bar{Z}_t > -u_t) = F_1(\delta)$$

$$= F(\delta)\int_{-\delta}^{\infty}\frac{\varphi(u)}{1 - F(-u)}du - \int_{-\delta}^{\infty}\frac{F(-u)\varphi(u)}{1 - F(-u)}du,$$

where $\delta = \Phi^{-1}(p_0)$.

Next, we derive the CDF $F_0(z)$ of $\hat{Z}_t$ conditional on $Y_t = 0$.

$$Pr(\bar{Z}_t < z|\bar{Z}_t < -u_t, u_t) = \begin{cases} \dfrac{Pr(\bar{Z}_t < z|u_t)}{Pr(\bar{Z}_t < -u_t|u_t)} & \text{if } z < -u_t \\ 1 & \text{otherwise} \end{cases}$$

Hence,

$$F_0(z) = Pr(\bar{Z}_t < z|\bar{Z}_t < -u_t) = \int_{-\infty}^{-z} \frac{F(z)}{F(-u)}\varphi(u)du + \int_{-z}^{\infty} \varphi(u)du.$$

$$\Rightarrow F_0(z) = F(z)\int_{-\infty}^{-z} \frac{\varphi(u)}{F(-u)}du + \Phi(z)$$

The corresponding PDF is:

$$f_0(z) = f(z)\int_{-\infty}^{-z} \frac{\varphi(u)}{F(-u)}du.$$

The probability of type I errors for the Probit classifier is given by:

$$\alpha = Pr(\bar{Z}_t > \delta|\bar{Z}_t < -u_t) = 1 - F_0(\delta)$$

$$= 1 - F(\delta)\int_{-\infty}^{-\delta} \frac{\varphi(u)}{F(-u)}du - \Phi(\delta).$$

The MISE is given by:

$$MISE(\hat{Y}, \delta) = (1 - \pi^*)\left(1 - F(\delta)\int_{-\infty}^{-\delta} \frac{\varphi(u)}{F(-u)}du - \Phi(\delta)\right)$$

$$+ \pi^*\left(F(\delta)\int_{-\delta}^{\infty} \frac{\varphi(u)}{1 - F(-u)}du - \int_{-\delta}^{\infty} \frac{F(-u)\varphi(u)}{1 - F(-u)}du\right).$$

Taking the derivative with respect to δ and equating to zero yields:

$$-(1 - \pi^*)f(\delta)\int_{-\infty}^{-\delta} \frac{\varphi(u)}{F(-u)}du + \pi^* f(\delta)\int_{-\delta}^{\infty} \frac{\varphi(u)}{1 - F(-u)}du = 0$$

$$\Leftrightarrow \frac{\int_{-\infty}^{-\delta} \frac{\varphi(u)}{F(-u)}du}{\int_{-\infty}^{-\delta} \frac{\varphi(u)}{F(-u)}du + \int_{-\delta}^{\infty} \frac{\varphi(u)}{1 - F(-u)}du} = \pi^* \blacksquare$$