Ordinal Discriminant Analysis: A new approach to the construction of optimal linear scores for ordinal category classification

Yizhar Toren*

Draft: October 13, 2010

Abstract

Most classification methods provide either a prediction of group membership or an assessment of class membership probability. In the case of two-group classification the predicted probability can be described as "risk" of belonging to a special group. When the required output is a set of ordinal risk categories, a discretization of the continuous risk prediction is achieved by two common methods: dividing the output of an "optimal" classification model into adjacent intervals that correspond to the desired risk categories, or constructing a set of models that describe the conditional risk function at specific points (quantile regression). By defining a new error measure for the distribution of risk onto intervals we are able identify lower bounds on the accuracy of these methods, showing sub-optimality both in their distribution of risk and in the efficiency of their resulting partition into intervals. Using the framework of discriminant analysis, we combine the two approaches by adding a new form of constraint to the existing discriminant analysis optimization problem and by introducing a penalty function to avoid degenerate solutions. Finally we show an example using linear discriminant analysis as a reference.

1 Introduction

The classical problem of discriminating between two groups of observations based on a given dataset has been widely discussed in the statistical literature. When only two groups are involved, the question of discrimination is reduced to whether or not a given observation is a member of a special group (where the other group is the default state, for example sick vs. healthy). Some classification methods, such as Fisher's linear discriminant analysis (LDA [4]), make decisive prediction of group membership while minimizing error in some sense (usually the miss-classification rate). Other methods, such as logistic regression, provide an estimate of the exact conditional probability of belonging to the

*This paper is based on a thesis submitted to the School of Mathematics, Tel-Aviv Univesity. Contact: yizhar.toren@math.tau.ac.il.

special class given a set of predictor variables. Throughout this paper we shall refer to this conditional probability as "*conditional risk*" or simply "*risk*", although sometimes belonging to the special group might actually have a very positive context (i.e success).

There are two ways to estimate the conditional risk function: parametric and non-parametric. Parametric methods include logit/probit models (Martin 1977 [14], Ohlsen 1980 [15]) and linear models (Amemiya 1981 [1], Maddala 1986 [10] and Amemiya 1985 [2]). Powell (1994 [16]) has a review of non-parametric estimators. For a comparison of these approaches and complete review see Elliott and Lieli (2006) [3] and more recently Green and Hensher (2010) [5].

The estimation of the exact structure of the conditional risk function comes in handy when we wish to make distinctions between observations that are finer than simply group membership. However, in realistic scenarios acting upon such estimations alone may prove to be difficult. An assessment on a scale of 1:100 (as percentages) or finer assessments have little practical use, primarily since the possible actions resulting from this information are usually few. For these cases a ordinal output is required.

It is important to note that our problem is not equivalent to multiple-group classification in two ways: first, our categories are ordinal by nature and relate to the underlying risk; second, the assignment into categories is not given a-priori and greatly depends on the selection of model, model parameters and the borders of the intervals assigned to each risk category.

There are two common approaches to creating an ordered set of categories that match such a finite set of escalating actions. The first approach is to divide post-hoc the continuous risk estimation of a known model into intervals; the second approach is to create multiple models describing the behaviour of the conditional risk function at specific points (a.k.a "quantile regression").

In the first approach, the continuous output of an existing optimal risk model (logit, linear or non-parametric) is divided into intervals, thus translating the prediction of risk (usually continuous in nature) into large "bins of risk" - i.e "low"/ "medium"/ "high" or "mild"/ "moderate"/ "severe" (depending on context). The final result of this discretization process is a set of ordinal categories based on the continuous prediction of conditional risk. The primary drawback of this approach is that the selection of the classification model and it's parameters is not performed in light of the final set of desired categories. Instead, an "optimal model" (in some sense) is constructed first, and the partition into discrete categories is performed post-hoc.

The second approach attempts to construct separate models that describe the behaviour of the conditional risk function at specific levels of risk. In linear models this approach is know as *quantile regression* (Koenker & Bassett 1978 [8]). Manksi ([11], [12], [13]) implemented this notion to binary response models (the equivalent of two-group classification) naming it "Maximum Score Estimation". The primary justification for using the second approach is methodological: it demands that we specify in advance the levels of risk that are of interest to us (a vector q of quantiles), and then constructs a series of models that describe conditional risk at these quantiles. However, as we shall demonstrate in section 3, using quantiles (or conditional probability over left-unbounded intervals) to describe the conditional behaviour of $Y \mid X$ over adjacent intervals has very limited relevance to our case and even the term "conditional quantiles" is in itself misleading.

The primary objective of this paper is to combine the idea of pre-set levels of risk over adjacent intervals (rather than risk quantiles) into the existing framework of discriminant analysis. Instead of constructing multiple models, we offer a process that optimizes a single classification model paired with a matching set of breakpoints that partition the model's output into ordinal risk categories. To that end we define a new measure of accuracy - *Interval Risk Deviation* (IRD) - which describes a model's ability to distribute risk correctly into intervals given a pre-set vector r of risk levels. We show how this new measure of error can be integrated into the framework of discriminant analysis by adding a constraint to the existing optimization problem. In addition, we address the more practical problem of effectively selecting breakpoints by introducing a penalty function to the modified optimization scheme.

The remainder of this paper is organized as follows. Section 2 defines risk categories, illustrates the problems of using the first approach by showing the existence of lower bounds on the accuracy of such models and suggests a new formulation to the optimization problem to provide accurate, optimal and non-degenerate solutions. Section 3 demonstrates the problem of using the second approach by showing that if our objective is to find a model that optimally assigns risk to adjacent intervals (rather than to left-unbounded overlapping intervals) we cannot use quantiles to construct optimal models. Finally, section 4 compares our method to Fisher's linear discriminant analysis by assuming Gaussian marginals and equal variance and showing an example.

2 Ordinal Discriminant Analysis

2.1 Definition

A *risk category* i ($i \in \{1, \dots T\}$) is defined by a single number $r_i \in [0, 1]$ which describes the conditional probability of an observation to be a member of the "special" class given that it belongs to the i'th category. Our predefined set of ordinal risk categories is uniquely characterized by a vector $r \in [0, 1]^T$ such that $0 \leq r_1 < r_2 < \dots < r_T \leq 1$.

Let $X = (X_1, \dots, X_P)$ be a continuous P-dimensional random vector and let $Y \in \{0, 1\}$ be a Bernoulli random variable representing group membership. An *Ordinal Risk Score* for a risk vector r is a couplet (Ψ, τ) where $\Psi : \mathbb{R}^P \to \mathbb{R}$ is a continuous (possibly not normalized) risk predictor and $\tau \in \mathbb{R}^{T-1}$ is a partition of $\mathbb{R}$ into T intervals ($-\infty = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_{T-1} < \tau_T = \infty$). Standard classification methods typically set $r = (0, 1)$ (and $T = 2$) meaning that we look for a single breakpoint $\tau \in \mathbb{R}$ and that the two open-ended intervals $(-\infty, \tau), (\tau, \infty)$ define an absolute prediction of class membership ($X \in$ Category $1 \Leftrightarrow \Psi(X) > \tau$).

We can describe the accuracy of the model (Ψ, τ) as it's ability to assign the intervals $[\tau_{i-1}, \tau_i]$ to the pre-defined set of conditional probabilities r_i by defining a measure of error for risk-category oriented models, which will be a parallel of *misclassification rate* in standard classification methods. We define $R(\Psi, \tau)_i$ as the actual conditional risk over the interval $[\tau_{i-1}, \tau_i]$, meaning that $R(\Psi, \tau)_i = P(Y = 1 \mid \Psi(X) \in [\tau_{i-1}, \tau_i])$. Using this notation can we describe the overall difference between the pre-defined risk vector r and

the actual conditional risk as *Interval Risk Deviation* IRD:

$$IRD_r(\Psi, \tau) = \|R(\Psi, \tau) - r\| \tag{1}$$

2.2 Lower bounds on Interval Risk Deviation

A common practice when building scores for risk categories is to build a model Ψ that is optimal in some sense (usually for accurate prediction of risk) and then partition the range of $\Psi(X)$ into adjacent intervals corresponding to the desired risk categories. In this section we demonstrate how, under relatively simple assumptions, using this approach with existing classification models is not optimal for more than two risk categories.

Our first observation is that if we set $r = (0, 1)$ (hence $T = 2$ and $\tau \in \mathbb{R}$) our description of the problem makes it essentially identical to the classical two-group classification. The fundamental difference becomes evident only when we require two or more breakpoints in our model ($T > 2$).

Using Bayes theorem we can represent R as:

$$R(\Psi, \tau)_i = P(Y = 1 \mid \Psi(X) \in [\tau_{i-1}, \tau_i]) = P(Y = 1)\frac{P(\Psi(X) \in [\tau_{i-1}, \tau_i] \mid Y = 1)}{P(\Psi(X) \in [\tau_{i-1}, \tau_i])}$$

By assuming that $R(\Psi, \tau)_i$ is finite, continuous and strictly increasing in τ_i we can describe the behaviour of infinitely short intervals ($\tau_i \to \tau_{i-1}$):

$$\begin{aligned} \lim_{\tau_i \to \tau_{i-1}} R(\Psi, \tau)_i &= \lim_{\tau_i \to \tau_{i-1}} \frac{P(Y = 1)\ P(\Psi(X) \in [\tau_{i-1}, \tau_i] \mid Y = 1)}{P(\Psi(X) \in [\tau_{i-1}, \tau_i])} = \\ &= P(Y = 1)\frac{\lim_{\tau_i \to \tau_{i-1}} \frac{P(\Psi(X) \in [\tau_{i-1}, \tau_i] | Y=1)}{\tau_i - \tau_{i-1}}}{\lim_{\tau_i \to \tau_{i-1}} \frac{P(\Psi(X) \in [\tau_{i-1}, \tau_i])}{\tau_i - \tau_{i-1}}} = P(Y = 1)\frac{f_{\Psi(X)|Y=1}(\tau_{i-1})}{f_{\Psi(X)}(\tau_{i-1})} \end{aligned} \tag{2}$$

where f is the appropriate density function and the limit is from the right-hand side. These equalities holds since by our assumptions on R all limits are well defined and finite.

To understand the implications of (2) on optimal model selection we look at the problem from a different perspective. Instead of a predefined set of probabilities r we fix a risk score (Ψ, τ) and assume that our model (Ψ, τ) has a perfect distribution of conditional risk up to the $(i-1)$ category, meaning that $R(\Psi, \tau)_j = r_j$ for all $j < i$. We can now show that under these assumptions not all values of r_i are exactly achievable without introducing some IRD, and in fact we can explicitly define a feasibility criteria:

$$r_i > P(Y = 1)\frac{f_{\Psi(X)|Y=1}(\tau^*_{i-1})}{f_{\Psi(X)}(\tau^*_{i-1})} \tag{3}$$

If r_i does not meet the feasibility criteria (3) then, by (2) and strict monotonicity of R, any selection of $\tau_i > \tau_{i-1}$ will have $R(\Psi, \tau)_i > r_i$, even if we set the interval $[\tau_{i-1}, \tau_i]$ to be arbitrarily small. Although we focus our discussion on the strictly monotone case, such bounds exist in the non monotone case as well, but in a more complicated form: we must have $r_i > min_{\tau_i}(R(\Psi, \tau)_i)$ in order to obtain and exact solution, and under reasonable assumptions we can expect $min_{\tau_i}(R(\Psi, \tau)_i) > 0$.

The inevitable conclusion is that the optimality of a model for a given r does not ensure it's optimality for other values of r or for models with a different number of breakpoints. The existing classification models (typically optimized for $r = (0, 1)$) are therefore not necessarily optimal for other choices of r.

Assuming some sense of continuity we can try to overcome this problem by considering augmented "local" solutions, whereby some algorithm lowers the breakpoint τ^*_{i-1} down to the point where the next value r_i is feasible, while balancing the increase in IRD on both sides of the breakpoint. Such local approaches are likely to produce very small or even degenerate intervals (where $|\tau_k - \tau_{k-1}| < \varepsilon$) which we would like to avoid if our models are to have meaningful application to real-world scenarios. Thus, we cannot "amend" the existing methods by focusing on a local neighbourhood of their optimal solutions and must optimize over a wider set of (Ψ, τ).

The existence of lower bounds on the IRD of some solutions is perhaps the most counter-intuitive result of this paper. The reason why these limitations have not been addressed before has to do with the fact that most discriminant analysis methods use only a single breakpoint between categories ($\tau \in \mathbb{R}$ and the issue of degenerate solutions or non-feasibility of r is avoided altogether). Once we define our objective as risk over multiple adjacent intervals we must recognize the existence of inherent limitations on IRD and as a result define new conditions for optimality.

2.3 Constrained optimization

Using simple algebraic manipulations we can represent (3) as a condition on the likelihood ratio Λ:

$$\Lambda_{X,Y,\Psi}(\tau_{i-1}) = \frac{f_{\Psi(X)|Y=1}(\tau_{i-1})}{f_{\Psi(X)|Y=0}(\tau_{i-1})} < \frac{1 - P(Y = 1)}{P(Y = 1)} \frac{r_i}{1 - r_i} \tag{4}$$

This formulation demonstrates how optimality in the sense of maximizing the likelihood ratio (as is the case for many classification methods, LDA in particular) will actually result in maximizing the lower bound on "feasible" values of r (with $IRD_r = 0$). On the other hand, minimizing the likelihood ratio Λ can ensure the existence of a solution with $IRD_r = 0$ (for example a degenerate solution $\Psi(X) = 0$). It becomes evident that rather than defining a single optimal solution the equation $IRD_r = 0$ defines a set of scores:

$$C_r(0) = \{(\Psi, \tau) \mid IRD_r(\Psi, \tau) = 0\} \tag{5}$$

If indeed $|C_r(0)| > 0$ it becomes possible to use IRD_r as a constraint in an optimization problem rather than as the definition of optimality itself. Since we are still in the domain of classification problems it would be reasonable to preserve the basic concepts of discriminant analysis, particularly the definition of optimality: We seek a model that on the one hand distributes risk optimally ($IRD_r(\Psi, \tau) = 0$), and on the other hand maximizes the discrimination between the two groups. In the case of discriminant analysis this notion is equivalent to maximizing the likelihood ratio under the constraint $IRD_r(\Psi, \tau) = 0$.

In addition, once we are looking for more than a single breakpoint the discrimination problem is no longer symmetric, meaning that relative placement of the groups on the scale becomes important. In most classification schemes the order of groups on the scale is

simply a question of interpreting a high scale result as "positive" or "negative". If our risk vector r is symmetric around $\frac{1}{2}$, meaning that $r = (r_1, \ldots, r_{T/2}, 1 - r_{T/2}, \ldots, 1 - r_1)$ then the problem remains symmetric. However if r is not symmetric around $\frac{1}{2}$ then reversing the score's orientation demands that we also reverse the order of the risk categories in order to maintain the original structure of our problem. Wishing conform with the original structure of our risk categories (meaning that the "special" group tends to be located to the left of the "default" group on the scale and the risk categories are ordered from left to right) we enforce another constraint on the optimization problem that maintains the relative order of the groups on the scale. The constrained optimization problem becomes:

$$\begin{aligned} \underset{\Psi,\tau}{\text{maximize}} \quad & \Lambda(\Psi) \\ \text{subject to} \quad & IRD_r(\Psi, \tau) = 0 \\ & \mathbb{E}(\Psi(X) \mid Y = 0) < \mathbb{E}(\Psi(X) \mid Y = 1) \end{aligned} \tag{6}$$

2.4 Penalty functions

While the optimization of Ψ is well defined, requiring $IRD_r = 0$ may lead to degenerate solutions of τ^* for certain values of r, as we shall show in the following example. Consider the case where $r = (0, 0.5, 1)$, $X \mid Y = 1 \sim N(\mu, \sigma)$, $X \mid Y = 0 \sim N(-\mu, \sigma)$ (a single real valued variable X with equal variance and symmetry around 0), $P(Y = 1) = \frac{1}{2}$ and our model Ψ is linear ($\Psi(X) = \beta X$ where $\beta \in \mathbb{R}$). In our terms $P = 1$ and $T = 3$ and since choosing Ψ is reduced to selecting a scale parameter β we may set $\beta = 1$ and optimize over τ alone.

The symmetry of the problem around $x = 0$ ensures that for any choice of $t \in \mathbb{R}$ we can minimize IRD for $i = 2$:

$$R(1, (-t, t))_2 = P(Y = 1 \mid X \in [-t, t]) = 0.5 = r_2$$

Symmetry also ensures that the risk prediction errors are equal on both sides:

$$R(1, (-t, t))_1 = 1 - R(1, (-t, t))_3$$

Using the Bayes theorem we have:

$$\begin{aligned} P(Y = 1 \mid X < -t) =& \frac{P(Y = 1)P(X < -t \mid Y = 1)}{P(Y = 1)P(X < -t \mid Y = 1) + P(Y = 0)P(X < -t \mid Y = 0)} \\ =& \frac{P(Y = 1)\Phi(-t - \mu)}{P(Y = 1)\Phi(-t - \mu) + P(Y = 0)\Phi(-t + \mu)} \\ =& \left(1 + \frac{P(Y = 0)}{P(Y = 1)} \frac{\Phi(-t + \mu)}{\Phi(-t - \mu)}\right)^{-1} \end{aligned}$$

where Φ is the CDF of the standard normal distribution. Using the known inequality:

$$\frac{\phi(x)}{x + 1/x} < \Phi(-x) < \frac{\phi(x)}{x} \quad \forall x > 0 \tag{7}$$

(where ϕ is the PDF of the standard normal distribution) we can show an upper bound:

$$\frac{\Phi(-t+\mu)}{\Phi(-t-\mu)} > \frac{(t+\mu)+\frac{1}{t+\mu}}{t-\mu}\frac{\phi(t-\mu)}{\phi(t+\mu)} = \frac{(t+\mu)+\frac{1}{t+\mu}}{t-\mu}\,e^{2\mu t} \quad \forall t > \mu$$

Therefore $\lim_{t\to\infty} P(Y=1 \mid X<-t) = 0$ and similarly $\lim_{t\to\infty} P(Y=1 \mid X>t) = 1$, and for any arbitrarily small $\varepsilon > 0$ we can find a sufficiently large t such that

$$P(Y=1 \mid X<-t) = R(1,(-t,t))_1 \le \varepsilon/2$$

The total IRD becomes:

$$IRD_{(0,0.5,1)}(1,(-t,t)) = \sqrt{\sum_{i=1}^{3}(R(1,(-t,t))_i - r_i)^2} \le \varepsilon$$

resulting in a degenerate optimal solution:

$$\lim_{t\to\infty} IRD_{(0,0.5,1)}(1,(-t,t)) = 0$$

There are several alternatives for dealing with this issue. First, we can decide that methodologically we do not allow setting $r_1 = 0$ or $r_T = 1$. This will ensure that the values of τ are finite but might still lead to very large or very small intervals, depending on the parameters of the model. Alternatively, we can introduce a *penalty function* $Pen : \mathbb{R}^P \times \mathbb{R}^{T-1} \to \mathbb{R}$, which will enable us to balance the properties of τ (minimal or maximal distance between breakpoints) with the discrimination properties of Ψ. We define the penalized optimization problem:

$$\begin{aligned} \underset{\Psi,\tau}{\text{maximize}} \quad & \Lambda(\Psi) + \gamma Pen(\tau) \\ \text{subject to} \quad & IRD_r(\Psi,\tau) = 0 \\ & \mathbb{E}(\Psi(X) \mid Y=0) < \mathbb{E}(\Psi(X) \mid Y=1) \end{aligned} \tag{8}$$

where the parameter γ is a tuning parameter that represents the degree of our aversion to degenerate solutions.

Possible candidates for $Pen(\tau)$ can be: $Pen(\tau) = max_{1\le i\le T-1}(\tau_i - \tau_{i-1})^{-1}$ which penalizes for the minimal distance between breakpoints and represents aversion to degenerate solutions with a pair $|\tau_k - \tau_{k-1}| < \varepsilon$ and ε arbitrarily small; $Pen(\tau) = \sum_{i=2}^{T-1}(\tau_i - \tau_{i-1}) = \tau_{T-1} - \tau_1$ which represents aversion to solutions where the overall span of τ too large (as in the above example); $Pen(\tau) = max_{1\le i\le T-1}(\tau_i - \tau_{i-1})/min_{1\le i\le T-1}(\tau_i - \tau_{i-1})$ which aims to control the ratio between the maximal and minimal distance between breakpoints (again aiming to avert degenerate solutions); any other combination of penalty functions, depending on the context of the problem and specific aversions.

The formulation of the optimization problem, using both constraints and penalty, reflects our understanding of the relative importance of IRD on the one hand and minimal distance between breakpoints on the other hand. Since IRD represents an absolute measure of the model's quality, we believe it must be tightly controlled (preferably having $IRD_r = 0$) and must be a constraint on any model we might consider. In contrast, the

requirement of "evenly spread" breakpoints is relatively subjective and should allow for some discretion as to the balance between the ability of the model to separate groups and the resulting minimal interval length. By choosing an appropriate penalty function and the aversion parameter γ we enable better fitting of the model according to the circumstances at hand, while introducing a relatively small number of additional parameters.

2.5 Uniqueness of τ under monotonicity

Once we expand our optimization problem to include the set of breakpoints τ (in addition to the parameters of the model Ψ), it is easy to see that the problem becomes more complicated. Although the increase in the number of parameters should not be significant (in practical scenarios we would not expect $T > 10$), introducing τ into the optimization process would still mean a more complicated parameter space (for example we require $\tau_{i-1} < \tau_i$) and as a result longer running times for the optimization algorithms. It would therefore be useful to identify sufficient conditions for the uniqueness of τ:

Lemma 1. *If for a given Ψ the function $R(\Psi,\tau)_i = P(Y = 1 \mid \Psi(X) \in [\tau_{i-1}, \tau_i])$ is strictly increasing in τ_i, then if there exists τ_Ψ such that $(\Psi, \tau_\Psi) \in C_r(0)$ it is unique.*

Proof. By induction: Strict monotonicity of R_1 means that if exists τ_1 which satisfies $R(\Psi,\tau)_1 = r_1$ then it is unique. Fixing τ_{i-1}, if (3) holds (meaning that τ_i is "feasible") then again by strict monotonicity if exists τ_i that satisfies $R(\Psi,\tau)_i = r_i$ then it is unique. Therefore if (3) holds for all i then only a single τ satisfies $IRD(\Psi,\tau) = 0$. □

Corollary 1. *Under strict monotonicity of R we can denote $\tau = \tau^\Psi$ and rewrite our definitions using using Ψ alone:*

$$R(\Psi)_i = P(Y = 1 \mid \Psi(X) \in [\tau_{i-1}^\Psi, \tau_i^\Psi]), \quad IRD(\Psi) = \|R(\Psi) - r\| \tag{9}$$

The optimization problem becomes:

$$\begin{aligned} \underset{\Psi}{\text{maximize}} \quad & \Lambda(\Psi) + \gamma Pen(\Psi) \\ \text{subject to} \quad & IRD_r(\Psi) = 0 \\ & \mathbb{E}(\Psi(X) \mid Y = 0) < \mathbb{E}(\Psi(X) \mid Y = 1) \end{aligned} \tag{10}$$

where $Pen(\Psi) = Pen(\tau_\Psi)$. We can now solve a much simpler optimization problem with reduced dimensionality.

3 Quantile-oriented models (and why we can't use them)

When first presented with the problem of selecting an optimal model paired with a set of optimal breakpoints, our initial idea was to use quantile-oriented models. Such models have been extensively researched in econometrics, where they are commonly referred to

as "ordered choice models" ([17], [5]). The most relevant model in that group is Manski's *maximum score estimation* which defines the optimization problem using a set of probabilities over *left-unbounded overlapping* intervals (or *rays*) in contrast to our definition of the problem over *adjacent, non-overlapping* intervals. In order to better illustrate the differences between our approach and Manski's quantile-oriented approach we must first describe quantile oriented models in our terms. First we replace the vector of interval probabilities r with a vector of "conditional quantiles" q which are in fact conditional probabilities over left-unbounded and overlapping intervals. Using Manski's adaptation of quantile regression [11] we can build a different set of model parameters for each quantile q_i optimizing:

$$|P(Y = 1 \mid \Psi_i(X) \leq 0) - q_i| \longrightarrow \min_{\Psi_i}$$

It is easy to see how this approach can be slightly modified to match our original objective of finding a single model: By coercing the models Ψ_i to be parallel we can create a "master model" $\Psi(X)$ and derive the appropriate thresholds $\{\tau_i\}$ such that:

$$\Psi_i^*(X) < 0 \quad \Leftrightarrow \quad \Psi^*(X) < \tau_i$$

$$|P(Y = 1 \mid \Psi(X) \leq \tau_i) - q_i| \longrightarrow \min_{\Psi} \quad i \in \{1, \ldots T\}$$

Using the above notions we can easily define $Q(\Psi, \tau)_i = P(Y = 1 \mid \Psi(X) < \tau_i)$ and the equivalent *Quantile Risk Deviation* $QRD(\Psi, \tau) = \|Q(\Psi, \tau) - q\|$ and look for a model with $QRD = 0$.

However, while it is tempting to describe the vector q as a vector of "*conditional quantiles*", the term is in itself misleading and should be avoided. Figure 1 demonstrates how even under relatively simple assumptions (Gaussian marginals and unequal covariance matrices) the function $Q(\Psi, \tau)_i = P(Y = 1 \mid \Psi(X) < \tau_i)$ is not even monotone in τ_i.

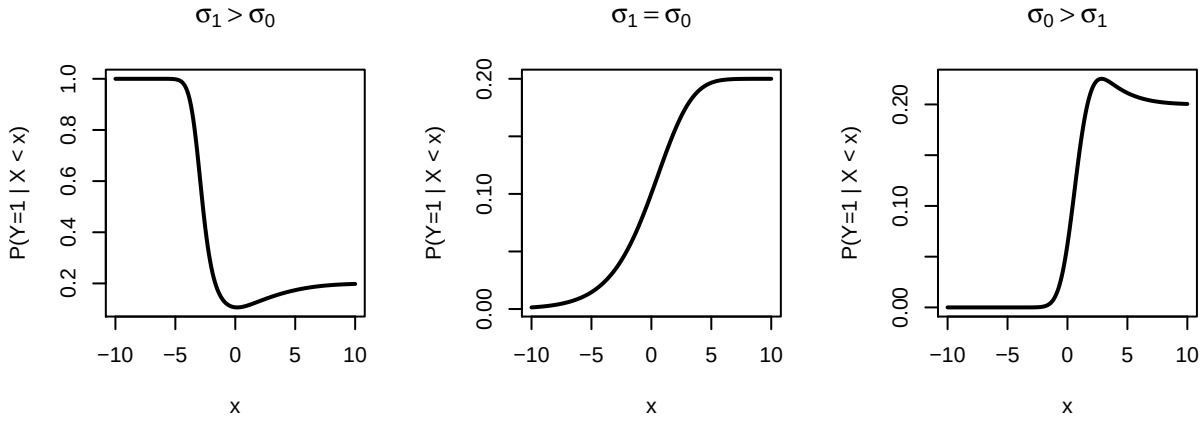

Figure 1: Different behaviour of conditional probability of left-unbounded intervals as a function of the left bound x in the case of gaussian marginals. The left panel is an example of $\sigma_1 > \sigma_0$, the middle panel is an example of the homoscedastic (strictly monotone) case and the right panel is an example of $\sigma_1 < \sigma_0$. Note that $P(Y = 1) = 0.2$.

Even if we assume strict monotonicity of $P(Y = 1 \mid \Psi(X) < x)$ (as is the case with homoscedastic Gaussian marginals, see appendix A) and can therefore use the term "conditional quantiles" in a meaningful sense, it would still be impossible to apply this approach to optimizing the distribution of risk over adjacent intervals.

If we could find a simple way to "translate" any given vector of conditional probabilities over intervals r to a vector of conditional probabilities over left unbounded intervals q then the two problems become equivalent. However it is easy to show that the relation between r and q depends on other parameters which make this simple translation impossible. Using the *law of total probability* in its conditional form we can calculate for any given R the equivalent Q_R (actual probabilities over left unbounded intervals):

$$\begin{aligned} Q_R(\Psi,\tau)_i =& P(Y=1 \mid \Psi(X) < \tau_i) \\ =& \sum_{j<i} P(Y=1 \mid \Psi(x) \in [\tau_{j-1},\tau_j]) P(\Psi(X) < \tau_j \mid \Psi(X) < \tau_i) \\ =& \frac{1}{P(\Psi(X) < \tau_i)} \sum_{j<i} R(\Psi,\tau)_j P(\Psi(X) < \tau_j) \end{aligned} \tag{11}$$

and the corresponding vector of target probabilities q_r:

$$(q_r)_i = \frac{1}{P(\Psi(X) < \tau_i)} \sum_{j<i} r_j P(\Psi(X) < \tau_j) \tag{12}$$

And as a result:

$$\begin{aligned} R(\Psi,\tau)_i = r_i \Leftrightarrow Q_R(\Psi,\tau) = (q_r)_i \\ IRD_r(\Psi,\tau) = 0 \Leftrightarrow QRD_{q_r}(\Psi,\tau) = 0 \end{aligned} \tag{13}$$

The primary problem of using quantiles to define our problems stems from the relation between r and the resulting q_r. Equation (12) demonstrates this paradox: in order to construct a maximum score estimation problem (with the appropriate q_r) for a given r we must first find the optimal model Ψ^* for this problem (which is what we are looking for in the first place). In other words, the translation $r \leftrightarrow q$ is possible only once we have an optimal solution to our problem, and if we begin with a given vector q we cannot know in advance what will be the conditional probabilities over adjacent intervals. Since by our own definitions the central aspect of the problem is the probability over adjacent intervals and not overlapping left-unbounded intervals, the optimization must be performed against a fixed parameter r and if there is no a-priori translation then we cannot use quantile-oriented methods to construct an optimal model.

4 Comparison with Linear Discriminant Analysis

Linear discriminant analysis (LDA) is one of the most studied classification methods in scientific literature and has been widely applied in statistics, scientific research and industry. Proposed by Fisher in his seminal paper from 1936 [4], it has been extensively studied and extended by numerous followers - see for example Hastie and Tibshirani [7] and recently [6] and a different extension from and to correspondence analysis by LeBart

et. al. [9]. Classical LDA provides a final prediction of class membership based on a linear scale $\beta_{LDA} \in \mathbb{R}^P$ and a single breakpoint between groups ($X\beta_{LDA} \gtrless \tau_{LDA}$).

In order to implement our notions on LDA we must make several assumptions: Gaussian marginals, linearity and homoscedasticity. LDA was originally developed for multivariate normal distributions and most of its elegant optimal properties are in fact closely tied to assumptions of normality. Linearity means our choice of risk predictor becomes a choice of $\beta \in \mathbb{R}^P$ ($\Psi(X) = X\beta$), and homoscedasticity means that the marginal covariance matrices are equal ($\Sigma_1 = \Sigma_0 = \Sigma$). Let $X = (X_1, \ldots, X_P)$ be a random vector from a P-dimensional multivariate normal distribution, and $\mu_0, \mu_1 \in \mathbb{R}^P$ and $\Sigma \in M_{PXP}$ represent two alternative distributions corresponding to the value of Y:

$$y \in \{0,1\} \quad X \mid Y = y \sim MVN_P(\mu_y, \Sigma_y), \quad X\beta \mid Y = y \sim N(\mu_y^T\beta, \beta^T\Sigma_y\beta)$$

$$P(X\beta \in [\tau_{i-1}, \tau_i] \mid Y = y) = \Phi\left(\frac{\tau_i - \mu_y^T\beta}{\sqrt{\beta^T\Sigma_y\beta}}\right) - \Phi\left(\frac{\tau_{i-1} - \mu_y^T\beta}{\sqrt{\beta^T\Sigma_y\beta}}\right) \tag{14}$$

where $i = 1, \ldots, T-1$, $y \in \{0,1\}$

Under these assumptions LDA can be said to maximizes the likelihood ratio Λ, or equivalently maximize the distance between the center of the transformed Gaussians:

$$\beta^*_{LDA} = \Sigma^{-1}(\mu_1 - \mu_0) = \underset{\|\beta\| \leq 1}{\operatorname{argmax}}\, \Lambda(\beta, \mu_0, \mu_1, \Sigma) \tag{15}$$

where the limitation on β is for scale normalization. By adding the constraints on IRD and the direction of the scale, we can modify the LDA optimization problem. We name this method Linear Ordinal Discriminant Analysis (LODA):

$$\begin{aligned} \underset{\beta,\tau}{\text{maximize}} \quad & \Lambda_{X,Y}(\beta) \\ \text{subject to} \quad & IRD_r(\beta, \tau) = 0 \\ & \beta^T\mu_0 < \beta^T\mu_1 \\ & \|\beta\| \leq 1 \end{aligned} \tag{16}$$

Equal variance under Gaussian assumptions ensures both the monotonicity of $P(Y = 1 \mid X < x)$ (see appendix A) and that the problem is symmetric around $(\mu_1^T\beta + \mu_0^T\beta)/2$. The result is that for any β we can find an optimal solution that minimizes IRD for $r = (0,1)$ by setting $\tau \in \mathbb{R}$ half way between the center of the projected Gaussians ($\tau^*(\beta) = (\mu_1^T\beta + \mu_0^T\beta)/2$). By lemmas 1 and 2, monotonicity also ensures that for any given error rate and our optimal solution is in fact unique. For the precise details about the strictly monotonicity of R under homoscedastic Gaussian marginals see appendix A.

This definition of the optimization problem is a natural extension of LDA: a single breakpoint between groups means that $T = 2$ and $r = (0,1)$. Since τ is a single real-valued number and therefore bounded by the values of β (and $\|\beta\| \leq 1$) we no longer need to add a penalty function to avoid degenerate or unbounded optimal solutions, and the optimization problem reverts to (15).

As we've seen in section 2.2, once we have more than a single breakpoint we introduce limits of feasibility into the maximization problem and may discover that β_{LDA} is no longer

feasible. However, if for our choice of r we have $\beta_{LDA} \in C_r(0)$ and the distances between the set of breakpoints τ_{LDA} are non-degenerate, then the global (unconstrained) optimality of β_{LDA} ensures that it is also the optimal solution of the bounded maximization problem. It may also be the optimal solution for the bounded and penalized optimization problem, but that will depend of the selection of the aversion parameter.

4.1 Example

In this section we shall demonstrate two cases: A case where the LDA solution meets the optimality constraints $IRD = 0$ and a case where a different optimal solution is required. As a homage to Fisher we shall use the same data set he used to originally describe LDA - the Iris dataset [4], which describes 3 species of flowers: Iris setosa, Iris virginica and Iris versicolor. The dataset contains 4 measurements of sepal length, sepal width, petal length and petal width performed on 150 flowers (50 from each species). Figure 2 shows the separation into classes using the canonical LDA solution.

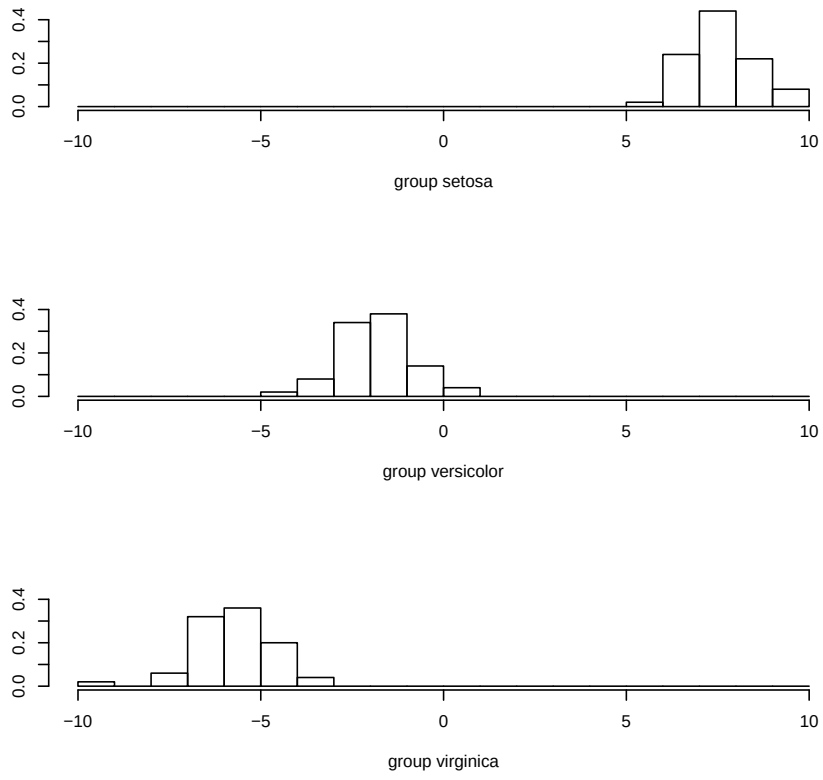

Figure 2: Iris data separation into groups using LDA

The code for this example was written using the R programming language (http://www.r-project.org) using the internally provided Iris dataset and the Augmented Lagrangian Adaptive BArrier Minimization Algorithm (ALABAMA) from the library *alabama* (http://cran.r-project.org/web/packages/alabama/).

Since our method was developed for two-group classification we compare only two classes - Iris virginica vs. Iris setosa. The LDA solution for this pair of classes (normalized such that $\beta^T \Sigma \beta = 1$) is:

$$\beta_{LDA} = (0.1599, 0.1213, -0.3686, -0.3711)$$

In order to demonstrate the differences between canonical LDA and our methods we use two sets of r: A symmetric risk stratification $r^{(1)} = (0.25, 0.5, 0.75)$ where we expect the canonical LDA to be feasible; A "skewed" risk stratification $r^{(2)} = (0.5, 0.7, 0.9)$ where we would expect to encounter feasibility problems.

For the first set of risk levels $r^{(1)}$ we calculated a set of matching breakpoints $\tau^{(1)}_{LDA} = (-0.909, 0.119)$. The solution $(\beta_{LDA}, \tau^{(1)}_{LDA})$ meets the optimization constraints ($IRD_{r^{(1)}} = 7.648511\text{e-}06$) and has a reasonable distance between the breakpoints (τ_1, τ_2). Hence it is the optimal solution for this set of risk levels.

For the second set of risk levels $r^{(2)}$ we calculated a set of matching breakpoints $\tau^{(2)}_{LDA} = (0.6048, 1.5097)$. However this time the solution $(\beta_{LDA}, \tau^{(2)}_{LDA})$ did not meet the optimization constraints ($IRD_{r^{(2)}} = 0.0281 > 0$). We constructed a constrained optimization problem and solved it using the tools mentioned above. Since we used a relatively non-specific optimization algorithm we encountered cases where the ALABAMA algorithm converged to a local minima. In order to overcome this problem we randomly selected 100,000 different starting points for the algorithm and analysed the results. The optimal solution of Linear Ordinal Discriminant Analysis (LODA) is:

$$\beta^* = (0.0084, 0.0891, -0.3289, -0.3331), \quad \tau^* = (-0.6437, 0.2232)$$

This solution meets the feasibility criteria ($IRD = 1.64\text{E-}10$). Figure 3 shows the difference between our solution and the LDA solution $(\beta_{LDA}, \tau^{(2)}_{LDA})$.

As expected our solution (β^*, τ^*) slightly compromises the quality of separation in order to achieve feasibility: the distance between the group means is $d_{LODA} = \mu_1^T \beta^* - \mu_0^T \beta^* = 1.9312$ compared with $d_{LDA} = 1.941$ for LDA. On the other hand our solution meets the IRD criteria while the LDA solution does not. As we expected, our constraints ensure that the relative order of the groups and the direction of influence of the variables remained similar to the LDA solution (the signs of all of the coefficients remained identical). The most notable difference between the two solutions is the relative weight of the first variable - "sepal length" (0.1599 in LDA vs. 0.0084 in LODA).

5 Conclusions

The construction of exact assessments of the conditional risk function is an important part of practical and theoretical research. However the practical application of this information is very often in the form of a finite and small set of resulting actions. In this paper we showed that the practice of dividing the continuous estimation of conditional risk into intervals ignores the limitations introduced by the lower bounds on IRD (1) and may also produce degenerate solutions.

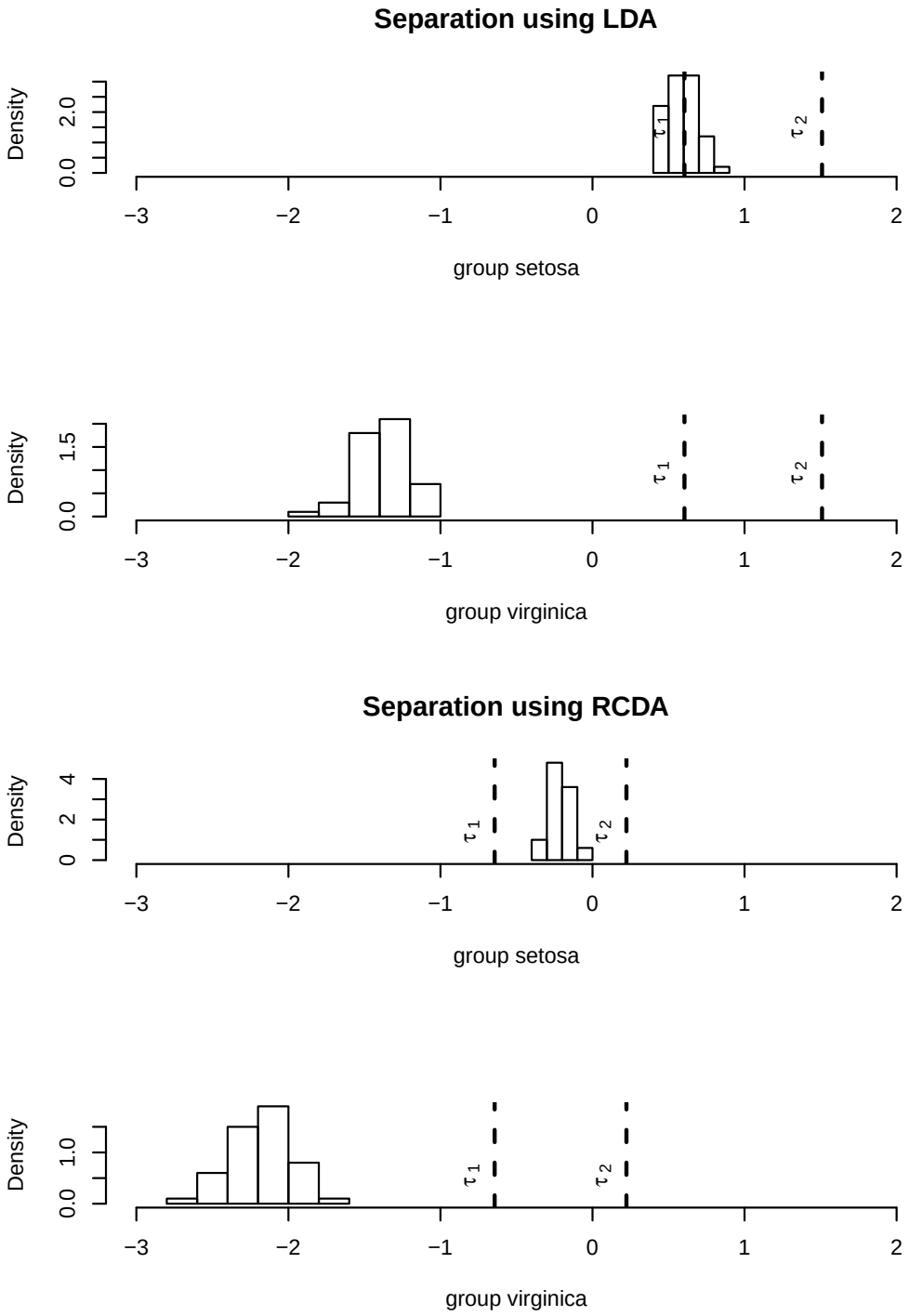

Figure 3: Separation of Iris virginica and Iris setosa data into groups using β_{LDA} (top graphs) and β^* (bottom graphs). Dotted lines mark the matching sets of breakpoints $\tau^{(2)}_{LDA}$ and τ^*.

In addition, although it is possible to define the distribution of risk in many ways, the definition of error on intervals as IRD (rather then using quantiles as QRD for example) reflects the desired outcome in the most straightforward manner. Although quantiles provide valuable information, we ultimately want to know the risk associated with non-overlapping intervals in order to create distinct ordinal categories, and therefore cannot use quantile regression for this purpose.

Our formulation of the optimization problem reflects our understanding that while the model's ability to separate the groups remains the key issue, we must introduce both a new constraint and a penalty function in order to achieve two additional objectives: an accurate risk distribution and a usable partition scheme. While IRD represents an absolute measure of the model's quality and must therefore be a constraint on the optimal solution, the "softer" requirement on minimal interval's length should allow for flexibility in application. We believe that a penalty function enables better control and adaptation through the choice of function and the aversion parameter.

Finally, we wish to emphasize the implications of the most counter intuitive result of this paper - the existence of limitations on certain values of r in the form of lower bounds on the error rate IRD (3). Although most of the examples we described are linear models with Gaussian marginals, the existence of lower bounds holds for any continuous risk estimator. A re-evaluation of the optimal properties of such estimators in the context of risk discretization is therefore required. We leave the specifics of applying these ideas to other classification methods to future studies.

References

[1] T. Amemiya. Qualitative response models: a survey. *Journal of economic literature*, 19(4):1483–1536, 1981.

[2] T. Amemiya. *Advanced econometrics*. Harvard Univ Pr, 1985.

[3] G. Elliott and R.P. Lieli. Predicting binary outcomes. *manuscript, UCSD*, 2006.

[4] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[5] W.H. Greene and D.A. Hensher. *Modelling Ordered Choices: A Primer*. Cambridge Univ Pr, 2010.

[6] Y. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86, 2007.

[7] Trevor Hastie, Andrea Buja, and Robert Tibshirani. Peanalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995.

[8] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, 46(1):33–50, 1978.

[9] Ludovic Lebart, Alain Morineau, and Kenneth M. Warwick. *Multivariate descriptive statistical analysis*. Wiley, 1984.

[10] G.S. Maddala. *Limited-dependent and qualitative variables in econometrics*. Cambridge Univ Pr, 1986.

[11] C.F. Manski. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3(3):205–228, 1975.

[12] C.F. Manski. Semiparametric analysis of discrete response:: Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27(3):313–333, 1985.

[13] C.F. Manski and T.S. Thompson. Operational characteristics of maximum score estimation* 1. *Journal of Econometrics*, 32(1):85–108, 1986.

[14] D. Martin. Early warning of bank failure: A logit regression approach. *Journal of Banking & Finance*, 1(3):249–276, 1977.

[15] J.A. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1):109–131, 1980.

[16] J. Powell. Estimation of Semiparametric Models. In R. Engle and D. McFadden, editors, *Handbook of Econometrics Vol. 4*. North Holland, 1994.

[17] K. Train. *Discrete choice methods with simulation*. Cambridge Univ Pr, 2003.

A Appendix: Monotonocity in the homoscedastic Gaussian case

Under homoscedastic Gaussian assumptions we can denote:

$$\Phi_y(x) = P(X\beta < x \mid Y = y) = \Phi\left(\frac{x - \mu_y^T\beta}{\beta^T\Sigma\beta}\right) \quad y = 0, 1 \tag{17}$$

and the conditional probability on $(-\infty, x]$ can be described as:

$$\begin{aligned} P(Y = 1 \mid X\beta < x) &= \frac{P(Y=1)\Phi_1(x)}{P(Y=1)\Phi_1(x) + P(Y=0)\Phi_0(x)} = \\ &= \left(1 + \frac{P(Y=0)}{P(Y=1)}\left(\frac{\Phi_1(x)}{\Phi_0(x)}\right)^{-1}\right)^{-1} \end{aligned} \tag{18}$$

Assuming that $\beta^T\mu_1 > \beta^T\mu_0$, showing strict monotonicity in x becomes equivalent to proving the following lemma:

Lemma 2. *The ratio* $\Psi(x) = \frac{\Phi_1(x)}{\Phi_0(x)}$ *is strictly increasing for all* $x \in \mathbb{R}$

Proof. For two smooth, positive and strictly increasing functions f, g the following equivalence holds:

$$\frac{f(x)}{g(x)} \text{ is strictly increasing } \Leftrightarrow \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)} > 0 \Leftrightarrow \frac{f(x)}{g(x)} < \frac{f'(x)}{g'(x)} \tag{19}$$

For convenience we denote $\mu_1 = \beta^T\mu_1$, $\mu_0 = \beta^T\mu_0$, $\sigma = \beta^T\Sigma\beta$ and ϕ_1, ϕ_0 as the appropriate derivatives. We prove that:

$$\frac{\Phi(\frac{x-\mu_1}{\sigma})}{\Phi(\frac{x-\mu_0}{\sigma})} < \frac{\phi(\frac{x-\mu_1}{\sigma})}{\phi(\frac{x-\mu_0}{\sigma})}, \quad \forall x \in \mathbb{R} \tag{20}$$

We distinguish 2 different cases: $x > \frac{\mu_1+\mu_0}{2}$, $x < \frac{\mu_1+\mu_0}{2}$

The first case is trivial: $x > \frac{\mu_1+\mu_0}{2} \Rightarrow \phi_1(x) > \phi_0(x)$, but for all $x \in \mathbb{R}$ we have $\Phi_1(x) < \Phi_0(x)$, meaning:

$$\frac{\Phi(\frac{x-\mu_1}{\sigma})}{\Phi(\frac{x-\mu_0}{\sigma})} < 1 < \frac{\phi(\frac{x-\mu_1}{\sigma})}{\phi(\frac{x-\mu_0}{\sigma})} \quad \forall x > \frac{\mu_1+\mu_0}{2} \tag{21}$$

For the second case we simplify our problem: Since Φ is invariant to scale and shift changes and we assume equal variance, we can prove strict monotonicity for a simpler case where $\mu_1 = \mu$, $\mu_0 = -\mu$, $\sigma = 1$.

$$\Psi(x) = \frac{\Phi_1(x)}{\Phi_0(x)} = \frac{\Phi(x-\mu)}{\Phi(x+\mu)} \tag{22}$$

From (21) we know that if $x_2 > x_1 > 0$ then:

$$\Psi(x_2) = \frac{\Phi(x_2-\mu)}{\Phi(x_2+\mu)} > \frac{\Phi(x_1-\mu)}{\Phi(x_1+\mu)} = \Psi(x_1) \tag{23}$$

For a negative number $-x$ (where $x > 0$):

$$\Psi(-x) = \frac{\Phi(-x-\mu)}{\Phi(-x+\mu)} = \frac{\Phi(-(x+\mu))}{\Phi(-(x-\mu))} = \frac{1-\Phi(x+\mu)}{1-\Phi(x-\mu)} = \frac{1-\Phi_0(x)}{1-\Phi_1(x)}$$

$$\Leftrightarrow \Psi(-x) = \frac{\frac{1}{\Phi_0(x)} - 1}{\frac{1}{\Phi_0(x)} - \Psi(x)} \tag{24}$$

Generally, for a strictly decreasing function h ($h'(x) < 0$) and a constant $c \in (0,1)$, the function $(h(x)-1)/(h(x)-c)$ is strictly decreasing:

$$\frac{d}{dx}\left(\frac{h(x)-1}{h(x)-c}\right) = \frac{h'(x)(h(x)-c) - h'(x)(h(x)-1)}{(h(x)-c)^2} = \frac{h'(x)(1-c)}{(h(x)-c)^2} < 0 \tag{25}$$

We can replace $h(x)$ with the strictly decreasing function $\frac{1}{\Phi_0(x)}$ and replace c with $\Psi(x_1)$ to show that for $x_2 > x_1 \Leftrightarrow -x_2 < -x_1$:

$$\Psi(-x_1) = \frac{\frac{1}{\Phi_0(x_1)} - 1}{\frac{1}{\Phi_0(x_1)} - \Psi(x_1)} > \frac{\frac{1}{\Phi_0(x_2)} - 1}{\frac{1}{\Phi_0(x_2)} - \Psi(x_1)} > \frac{\frac{1}{\Phi_0(x_2)} - 1}{\frac{1}{\Phi_0(x_2)} - \Psi(x_2)} = \Psi(-x_2) \tag{26}$$

The last inequality holds by (23) and the fact that $\frac{1}{\Phi_0(x)} - 1 > 0$ □

Showing that $R(\beta,\tau)_i = P(Y = 1 \mid X\beta \in [\tau_{i-1}, \tau_i])$ is strictly increasing in τ_i is more complicated. Again we prove an equivalent argument about the strict monotonicity of $\Psi(c,x) = (\Phi_1(x) - \Phi_1(c))/(\Phi_0(x) - \Phi_0(c))$ in x. However the approximations such as (7) are now of little use since it is difficult to analytically manipulate equations with sums of exponents.

We bring instead an outline of a proof which shows that as a function of x the ratio $\Psi(c,x)$ has a single saddle point or equivalently that $d^2\Psi/dx^2$ has a single root. Solving equations with Φ and its derivatives is usually achievable only by using numerical methods, but a reasonable alternative would be to simulate a wide selection of parameters and to show numerically that for any choice we make the function has only one root. Alternatively we present some intuitive insight into the problem by graphically demonstrating that μ_1, μ_0, Σ, c are shift and scale parameters of the second derivative, and therefore it is sufficient to show numerically that the function has a single root for one set of parameters. Figure 4 demonstrates the shift and scale changes to $d^2\Psi/dx^2$ as a function of these parameters.

Having two horizontal asymptotes at $\pm\infty$ and a single saddle point s means that the function $\Psi(c,x)$ is first convex and then concave: It rises to the horizontal asymptote $(1 - \Phi_1(a))/(1 - \Phi_0(a))$ at $+\infty$ and is therefore concave and strictly increasing on the right side of the saddle point. The existence of $x_2 > x_1 > s$ such that $R(c,x_2) \leq \Psi(c,x_1)$ together with the existance of a horizontal asymptote at $+\infty$ means that exists $x_3 > x_2 > x_1 > s$ such that $\Psi(c,x_3) > R(c,x_1) \geq \Psi(c,x_2)$ and the function is not strictly concave (a contradiction to the uniqueness of the saddle point). For the same considerations $\Psi(c,x)$ descends to the asymptote $\Phi_1(a)/\Phi_0(a)$ at $-\infty$ and is therefore convex and strictly increasing on the left side of the saddle point.

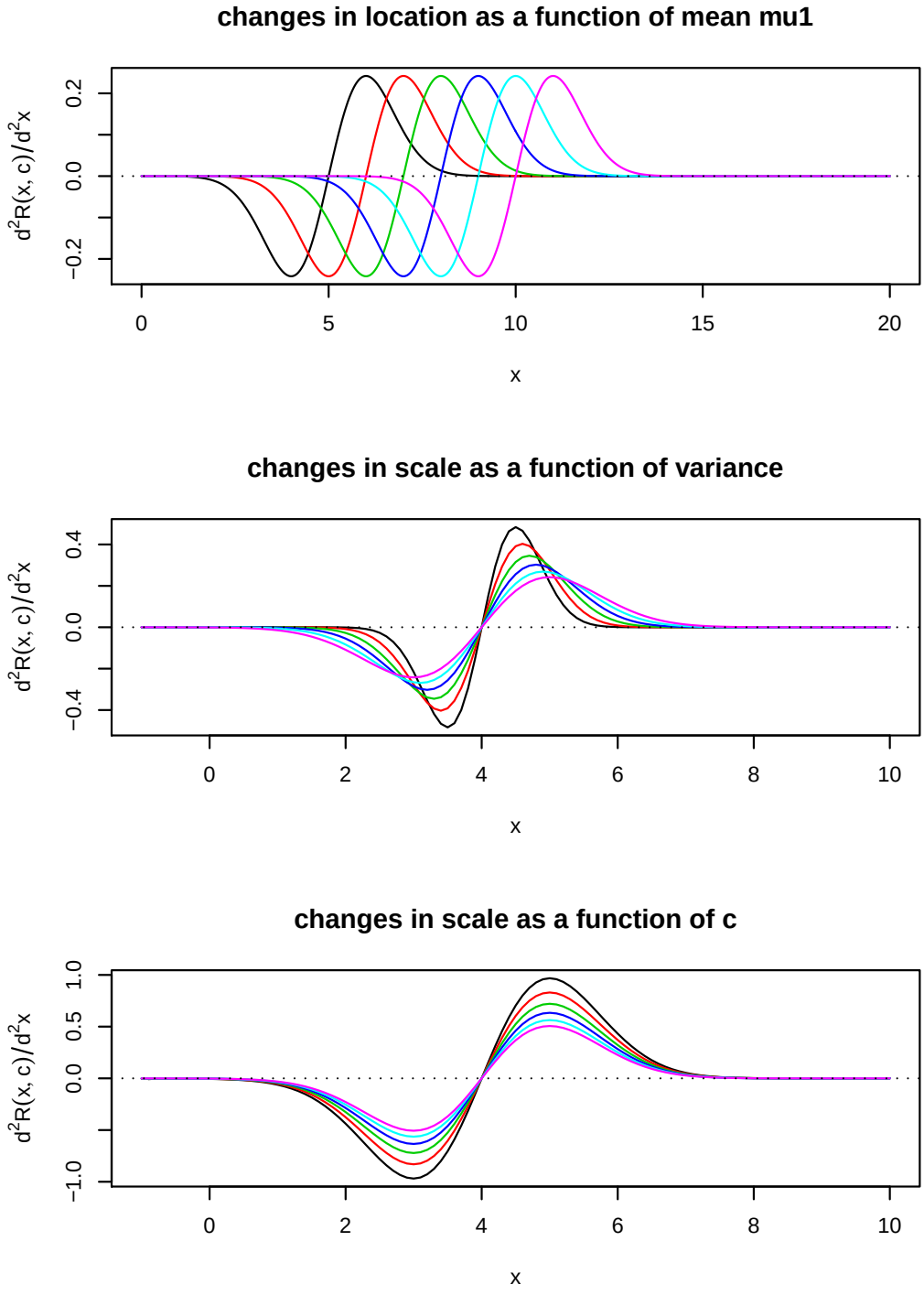

Figure 4: Changes in the parameters of the second derivative $d^2\Psi(x,c)/d^2x$ are scale and shift changes. The top graph shows shift changes for various values of μ_1 (μ_0 remains fixed). The center graphs shows scale and shift changes for various values of the covariance matrix Σ and. The bottom graph shows scale changes for various values of c. All functions have a single root and therefore a single saddle point.