Understanding and Evaluating Sparse Linear Discriminant Analysis

Yi Wu
EECS, UC Berkeley
jxwuyi@cs.berkeley.edu

David Wipf
Microsoft Research Asia
davidwipf@gmail.com

Jeong-Min Yun
Pohang University of Science and Techology
azida@postech.ac.kr

Abstract

Linear discriminant analysis (LDA) represents a simple yet powerful technique for partitioning a p-dimensional feature vector into one of K classes based on a linear projection learned from N labeled observations. However, it is well-established that in the high-dimensional setting ($p > N$) the underlying projection estimator degenerates. Moreover, any linear discriminate function involving a large number of features may be difficult to interpret. To ameliorate these issues, two general categories of sparse LDA modifications have been proposed, both to reduce the number of active features and to stabilize the resulting projections. The first, based on optimal scoring, is more straightforward to implement and analyze but has been heavily criticized for its ambiguous connection with the original LDA formulation. In contrast, a second strategy applies sparse penalty functions directly to the original LDA objective but requires additional heuristic trade-off parameters, has unknown global and local minima properties, and requires a greedy sequential optimization procedure. In all cases the choice of sparse regularizer can be important, but no rigorous guidelines have been provided regarding which penalty might be preferable. Against this backdrop, we winnow down the broad space of candidate sparse LDA algorithms and promote a specific selection based on optimal scoring coupled with a particular, complementary sparse regularizer. This overall process ultimately progresses our understanding of sparse LDA in general, while leading to targeted modifications of existing algorithms that produce superior results in practice on three high-dimensional gene data sets.

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

1 Introduction

Classical linear discriminant analysis (LDA) or Fisher's LDA addresses the classification problem of finding projections such that a high-dimensional data vector can be mapped into the most discriminative low-dimensional subspace (Fisher, 1936; Hastie et al., 2009). At the most basic level, this is accomplished by maximizing an estimate of the between-class variance subject to a constraint on the within-class variance in the projected space. We can formalize this procedure as follows: Assume we are given an $N \times p$ design matrix X where each row contains a p-dimensional sample belonging to one of K different classes, and each column denotes N observations of a given feature. Let x_i denote the ith row of X and μ_k the sample mean of class k. For convenience, we assume that each column of X is centered to have zero mean and unit ℓ_2 norm. Let Y be an $N \times K$ matrix of zeros and ones, where $Y_{i,k}$ is an indicator of whether sample x_i belongs to class k.

The standard between-class covariance Σ_b is defined by

$$\Sigma_b = \frac{1}{N}\sum_{k=1}^{K} n_k \mu_k \mu_k^\top = \frac{1}{N} X^\top P_Y X, \tag{1}$$

where n_k denotes the number of samples belonging to class k and the projection P_Y is defined by $P_Y = Y(Y^\top Y)^{-1}Y^\top$. The corresponding within-class covariance Σ_w is defined by

$$\Sigma_w = \frac{1}{N} X^\top (I - P_Y) X. \tag{2}$$

such that the total sample covariance satisfies $\frac{1}{N}X^\top X = \Sigma_b + \Sigma_w$. LDA then finds L discriminant vectors or a projection matrix B that solves

$$\max_B \operatorname{tr}(B^\top \Sigma_b B) \;\text{ s.t. } B^\top \Sigma_w B = I, \tag{3}$$

where $B = [\beta_1, \ldots, \beta_L] \in \mathbb{R}^{p\times L}$. We call β_k the kth discriminant vector and B the discriminant matrix. Given this B, new observations can be projected to a low-dimensional space and classified using a simple probabilistic decision rule (see supplementary file for details).

In general, there are at most $K-1$ non-trivial or linearly independent discriminant vectors since $\operatorname{rank}[\Sigma_b] \leq K-1$; hence we typically choose $L \leq K-1$.

When Σ_w is not full-rank, which will necessarily be the case in the high-dimensional setting where $p > N$, then the LDA problem is no longer well-posed. Consequently, several regularized versions of LDA have been proposed (Hastie et al., 1994). One such example is to modify the within-class covariance using some Ω such that $\Sigma_w + \Omega$ is positive definite and convert the canonical LDA cost to

$$\max_B \ \mathrm{tr}(B^\top \Sigma_b B) \ \ \text{s.t. } B^\top(\Sigma_w + \Omega)B = I, \quad (4)$$

which can be solved as an eigenvalue problem by converting to the reparameterized form

$$\max_B \ \mathrm{tr}(\tilde{B}^\top \widetilde{\Sigma}_b \tilde{B}) \ \ \text{s.t. } \tilde{B}^\top \tilde{B} = I, \quad (5)$$

where $\widetilde{\Sigma}_b = (\Sigma_w + \Omega)^{-1/2} \Sigma_b (\Sigma_w + \Omega)^{-1/2}$. The latter expression is optimized by setting $\tilde{B}$ to the eigenvectors of $\widetilde{\Sigma}_b$ associated with the L largest eigenvalues.

Alternatively, it has been shown in (Hastie et al., 1994) and (Hastie et al., 1995) that (4) is equivalent to solving

$$\begin{array}{ll} \min_{B,\Theta} & \frac{1}{N}\|Y\Theta - XB\|_F^2 + \mathrm{tr}(B^\top \Omega B) \\ \text{s.t.} & \Theta^\top Y^\top Y \Theta = I, \end{array} \quad (6)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Here we refer to $\Theta = [\theta_1, \ldots, \theta_L]$ as the scoring matrix and θ_k as the kth scoring vector. Consequently, LDA can be recast as a multi-output regression problem, which opens the door to more elaborate forms of regularization.

Both (4) and (6) solve the problem of a degenerate Σ_w, but require the heuristic selection of Ω. Additionally, even when an effective choice for Ω is somehow provided, the resulting high-dimensional discriminant vectors β_k will involve a combination of every feature and therefore will be difficult to interpret and possibly expensive to deploy in real-time environments. Consequently, sparse variants of LDA represent a highly desirable alternative.

To promote sparse discriminant vectors, a sparse penalty function ϕ can be added to either of the equivalent LDA expressions given above, although once such penalty is adopted the equivalence no longer holds as will be discussed more in the sequel. Perhaps the most common selection is $\phi(\beta) = \|\beta\|_1$ (which is associated with the Lasso sparse estimator in the context of regression (Tibshirani, 1996)), but later we will see that other possibilities can lead to considerable improvement.

In (Witten & Tibshirani, 2011) the canonical Fisher LDA objective (4) is supplemented with an ℓ_1-based penalty and a closely-related, specialized variant is proposed in (Zhang & Chu, 2013). Similar ideas are applied in (Wu et al., 2009) and (Moghaddam et al., 2006), where a sparsity regularizer is added instead to the constraints. Alternatively, (Clemmensen et al., 2011) and (Grosenick et al., 2008) apply a sparse penalty to the optimal scoring framework (6). Likewise, (Merchante et al., 2012) and (Leng, 2007) consider optimal scoring variants based upon a group Lasso penalty. Some other related work and applications can be found in (Dundar et al., 2005) and (Fung & Ng, 2007).

In general, applying a sparse penalty ϕ to (4) leads to problems of the form

$$\begin{array}{ll} \max_B & \mathrm{tr}(B^\top \Sigma_b B) - \lambda \sum_{k=1}^L \phi(\beta_k) \\ \text{s.t.} & B^\top(\Sigma_w + \Omega)B = I \end{array} \quad (7)$$

or related sequential versions. In contrast, with optimal scoring we must solve something akin to

$$\begin{array}{ll} \min_{B,\Theta} & \frac{1}{N}\|Y\Theta - XB\|_F^2 + \mathrm{tr}(B^\top \Omega B) + \lambda \sum_k \phi(\beta_k) \\ \text{s.t.} & \Theta^\top Y^\top Y \Theta = I, \end{array} \quad (8)$$

where Ω may or may not equal zero. However, one important modification of the sparse penalty produces a B matrix that is row sparse rather than element-wise sparse, the advantage being that a zero-valued row corresponds with a feature being completely pruned from the model. A specific successful instance of this comes from (Merchante et al., 2012), which proposes to solve the group Lasso penalized optimal scoring problem

$$\begin{array}{ll} \min_{B,\Theta} & \frac{1}{2}\|Y\Theta - XB\|_F^2 + \lambda \sum_i \|\beta^i\|_2 \\ \text{s.t.} & \Theta^\top Y^\top Y \Theta = I, \end{array} \quad (9)$$

where β^i denotes the ith row of B and $\Omega = 0$.

The remainder of the paper is organized as follows. In Section 2 we describe limitations in the current analysis and understanding of existing sparse LDA algorithms, all of which make choosing an optimal variant out of the numerous possibilities difficult. Next, Section 3 presents new analysis regarding the relationship between the two primary classes of sparse LDA techniques, demonstrating that the optimal scoring route is preferable in many respects despite pervasive conventional wisdom to the contrary. This perspective allows us to focus our attention on optimal scoring paradigms, with competing algorithms differentiated only by various families of sparsity penalties and associated update rules as discussed in Section 4. As a particular special case, we then argue that sparse regularizers that emerge implicitly from classical Bayesian learning algorithms, e.g., the relevance vector machine (Bishop & Tipping, 2000; Tipping, 2001), are natural candidates for retrofitting into a principled, optimal scoring LDA framework. Moreover, in a restricted setting we prove that this adaptation will produce maximally sparse projections unlike existing sparse LDA algorithms. Finally, Section 5 reveals that these insights can lead to state-of-the-art performance on real-world high-dimensional datasets. Note that our overall objective here is not to derive completely new algorithms per se; rather it is to better understand existing frameworks leading to specific design choices and targeted enhancements such that sparse LDA is optimally utilized to the extent possible.

2 Limitations of Existing Sparse LDA Algorithms and Analyses

As described above, there are basically two entry points for enforcing sparsity in the LDA framework: either directly to the original LDA cost function leading to (7) (e.g., (Witten & Tibshirani, 2011)), or indirectly via optimal scoring producing (8) and related row sparse variants (e.g., (Clemmensen et al., 2011; Merchante et al., 2012)). We now describe algorithmic and related analytical issues that have previously not been fully appreciated, and yet which nonetheless greatly impact the evaluation of ideal algorithm selections.

2.1 Algorithmic Issues

Applying sparse penalties directly to (7) comes with a substantial downside which is not addressed in existing sparse LDA studies. Simply put, sparse penalties themselves do not actually help solve the problem of a degenerate within-class covariance matrix Σ_w, which always occurs with $p > N$. To understand this, note that sparse penalties in a vector space are typically jointly concave, non-decreasing functions of the magnitudes of each element, otherwise individual coefficients have essentially no chance of being shrunk to zero (Rao et al., 2003). Now consider again the sparse LDA cost function from (7) but without the extra regularization to Σ_w provided by Ω, i.e., we are assuming $\Omega = 0$. This leads to the following result:

Lemma 1. *Let $\phi(z)$ be any concave, non-decreasing function of $|z| \triangleq [|z_1|, \ldots, |z_p|]^\top \in \mathbb{R}^p_+$ (e.g., $\phi(z) = \sum_i |z_i|^q$ with $q \leq 1$ or $\phi(z) = \sum_i \log(|z_i| + \epsilon)$ with $\epsilon \geq 0$) and assume that $K = 2$ (binary classification for simplicity). Then if $\Omega = 0$, $p > N$, and rank$[X] = N - 1$, the LDA cost function from (7) is unbounded from above for some β vector that has no zero-valued elements (non-sparse) and magnitudes tending towards infinity.*

Proofs of all results are deferred to the supplementary file; however, briefly in words Lemma 1 follows from the fact that *non-sparse* discriminate vectors with infinite magnitude and lying in the null-space of Σ_w will maximize the quadratic Σ_b term in (7) while swamping out the effects of the sparse regularizer which grows much more slowly as elements of β become large.

The rank condition implies that X is full-rank (after removing column means) and is merely included to rule out a high-dimensional problem with $p > N$ collapsing to an equivalent lower-dimensional problem with $p < N$. Likewise, while for simplicity Lemma 1 addresses the $K = 2$ case, the basic result can be extended to all $K > 2$ as well, implying that generally we cannot rely on a concave, sparsity-inducing regularizer to ameliorate the effects of a degenerate within-class covariance matrix. In fact, it naturally follows from the proof of Lemma 1 that $\Omega > 0$ is required to counteract this effect. But this then necessitates that both the λ and Ω weighting factors in (7) must be chosen via some potentially expensive cross-validation procedure. Moreover, it is entirely unclear to what extent actual sparse estimates for B even remain possible here since the larger we make Ω, the more the feasible region $B^\top(\Sigma_w + \Omega)B = I$ begins to disproportionately constrain sparse solutions (e.g., if $\Omega = \eta I$ for some constant $\eta > 0$, then the resulting quadratic factor in Lagrangian form can be viewed as a quadratic penalty well known to *disfavor* sparsity).

In contrast, the more indirect optimal scoring formulation does not suffer from these complications. Rather, a sparsity penalty applied as in (8) is sufficient to resolve the corresponding underdetermined system such that an additional Ω factor is not required; hence $\Omega = 0$ represents a simple robust choice. However, these methods are considered to be indirect in the sense that optimal scoring will no longer be equivalent to LDA once sparse penalties have been introduced. This is a perceived weakness as described in (Witten & Tibshirani, 2011), where it is argued that the direct form more accurately reflects the most natural, sparse extension of LDA.

To summarize then, the direct form of incorporating sparsity suffers from degenerate global optima without additional heuristics, making subsequent analysis of sparsity and efficient implementations more difficult and tuning-parameter-dependent. Meanwhile the indirect optimal scoring form maintains an ambiguous connection with the original LDA formulation but retains the structure of sparse regression problems more amenable to rigorous investigation and implementation. Moreover, as an additional algorithmic issue none of these methods come with a clear prescription for choosing an optimal penalty function ϕ. Note also that even if ϕ is convex, joint optimization over both B and Θ will generally not be.

2.2 Analytical Issues

Here we briefly unpack existing attempts to analytically characterize some of the differences between existing sparse LDA algorithms. First, in (Witten & Tibshirani, 2011) it is demonstrated that under certain special conditions (including that $K = 2$ and ϕ is an ℓ_1-norm penalty) optimal scoring (8) and direct sparse LDA (7) formulations may share a common stationary point. However, we are able to show (in part by extending the range of applicability of Lemma 1 to the case where $\Omega > 0$) that in general, the actual global minima may be very different depending upon the value of Ω. In other words, this common stationary point may be completely unrelated to the respective global optima. Consequently, it is likely to have little overall relevance regarding commonalities in the region of the respective cost functions that we most care about.

Secondly, in (Merchante et al., 2012) another initial attempt

is made to connect sparse optimal scoring with canonical sparse LDA. The core argument is that if $B_\star$ solves the particular sparse optimal scoring problem given by (9), again with a convex sparsity penalty, then an appropriate $\Omega = f(B_\star)$ can be chosen such that the LDA problem from (6) is optimized by some B proportional to $B_\star$. But this result is circular in that the optimal Ω depends directly on $B_\star$, and hence ultimately the resulting sparsity profile is completely independent of the original LDA model, significantly muting the overall relevance.

Finally, a related approach to handling sparse LDA involves selecting a particular choice of Ω such that (4) has a relatively high-dimensional subspace of global solutions. A sparsity penalty can then be directly applied to solutions in this subspace, resulting in a specialized case of (8). In particular, with $\Omega = \Sigma_b$ and $\phi(\beta_k) = \|\beta_k\|_1$, then the algorithm proposed in (Zhang & Chu, 2013), called SULDA (for sparse uncorrelated LDA) is effectively minimizing (8) in the limit as $\lambda \to 0$ (note that the limit must be taken outside of the maximization). Although originally presented in a much different way in (Zhang & Chu, 2013), we can use the equivalency between (4) and (6) to show that SULDA is tantamount to executing the following procedure. First, let B_* denote an optimum to

$$\begin{array}{cc} \min_{B,\Theta} & \|Y\Theta - XB\|_F^2 + \mathrm{tr}(B^\top X^\top P_Y X B) \\ \text{s.t.} & \Theta^\top Y^\top Y \Theta = I. \end{array} \tag{10}$$

Now clearly any additional $\bar{B} \in \mathrm{null}[X]$ added to B_* will not alter (10), and hence $B_* + \bar{B}$ will also represent a solution to (10). Therefore SULDA simply looks for new sparse solutions by solving L decoupled ℓ_1 linear programs of the form

$$\begin{array}{cc} \min_{\bar{\beta}_k} & \|\beta_{k*} + \bar{\beta}_k\|_1 \\ \text{s.t.} & X\bar{\beta}_k = 0. \end{array} \tag{11}$$

One appealing aspect of SULDA is that Ω and λ have been chosen for us, and moreover, for these specialized choices we obtain full equivalency between optimal scoring and canonical sparse LDA formulations, and there is no degenerate global solution for the latter.

But all of this comes with a price. For example, it is unclear why $\Omega = \Sigma_b$ might be the best regularizer for sparse LDA classification performance beyond algorithmic convenience. Additionally, the number of nonzero elements in any potential solution is bounded from *below* by NL (unlike algorithms we discuss in Section 3 which are bounded from *above* by this same amount, potentially leading to much greater sparsity and interpretability). Moreover, if we wish to relax the strict equality constraint in (11) to allow for greater sparsity, then any solution is no longer optimal to (10) and we lose any direct connection with the original Fisher LDA, which was the goal to begin with.

3 Reexamining Sparse LDA

The previous section served to characterize the equivocal nature of present understanding of sparse LDA methods, making unequivocal design choices more challenging. In this section we intend to elucidate the precise connection between optimal scoring and direct sparse LDA, leading to arguments suggesting the superiority of the former. We first consider the $K = 2$ case (binary classification) which leads to cleaner analysis, although many of the conclusions carry over to more general multi-class situations. Later we address the $K > 2$ scenario explicitly.

3.1 Sparse LDA with $K = 2$

When $K = 2$, only one discriminant vector is required and so we simply ignore the vector subscript. In this restricted setting there exist two primary methods for introducing sparsity into the LDA model. The first, based on optimal scoring involves solving

$$\begin{array}{cc} \min_{\beta,\theta\geq 0} & \frac{1}{N}\|Y\theta - X\beta\|_2^2 + \lambda\phi(\beta) \\ \text{s.t.} & \theta^\top Y^\top Y \theta = 1, \end{array} \tag{12}$$

while the second is given by

$$\max_\beta \beta^\top \Sigma_b \beta - \lambda\phi(\beta) \;\; \text{s.t. } \beta^\top \Sigma_w \beta = 1. \tag{13}$$

In both cases we are assuming the heuristic Ω factor from (7) and (8) is set to zero. As we argued in Section 2, this poses no problem for solving (12), but produces a problematic degenerate global solution when solving (13). Hence some $\Omega > 0$ has traditionally been applied in a somewhat ad hoc fashion as a practical remedy. However, we will now show that (12) is actually equivalent to solving a problem very similar to (13), but with a principled modification that automatically removes this undesirable degeneracy. In doing so we demonstrate that, contrary to conventional wisdom (e.g., (Witten & Tibshirani, 2011)), the optimal scoring formulation can actually be viewed as a superior, more direct entry point for introducing sparsity into LDA models that does not require heuristic modifications to the within-class covariance Σ_w. In other words, we show that optimal scoring does in fact retain a close connection with the original LDA formulation but with problematic optima removed.

Theorem 1. *There exist non-negative constants α_1 and α_2 (dependent on λ) such that (12) is equivalent to solving*

$$\begin{array}{cc} \max_\beta & h\left(\beta^\top \Sigma_b \beta\right) - \lambda\phi(\alpha_1\beta) \\ \textit{s.t.} & \beta^\top \Sigma_w \beta = \alpha_2, \end{array} \tag{14}$$

where $h(v) \triangleq \frac{v}{N(1+v)}$. Moreover, every locally minimizing solution can be achieved at a solution with at most N nonzero elements (but typically there are fewer).

Clearly (14) possesses the same basic structure as (13), the primary difference being the monotonic transformation h

and some rescalings, although the overall scaling of LDA projections is irrelevant. In fact this squashing function h serves a very desirable purpose which is most pronounced when $p > N$ and so Σ_w is not full rank. Specifically, by rescaling $\beta^\top \Sigma_b \beta$ between zero and one, it prevents the type of degenerate solutions that occur when $\beta \in \text{null}[\Sigma_w]$ grows arbitrarily large. Instead, there is no significant advantage to large β coefficients. Consequently the sparse penalty $\phi(\beta)$ will not be overwhelmed and sparse bounded solutions will naturally be produced. Note also that when β is sparse, the effective subspace of Σ_w encountered by $\beta^\top \Sigma_w \beta$ will generally be full rank.

3.2 General Case for $K > 2$

There are at least two different ways to extend (12) to handle the general sparse LDA problem for $K > 2$. One direction is to solve (8) with $\Omega = 0$. Since each discriminant vector has an independent sparse penalty, this model will make each β_k sparse, which implies that the total number of non-zero entries in the resulting B will be small. In fact, using (Wipf et al., 2011) it is straightforward to show that every local minima can be achieved with at most NL nonzeros, while typically the actual number is far less. However, because the sparsity profile or support of each β_k will generally not be shared, it is unlikely that any given feature will be entirely pruned from the model when L is large, which would require that an entire *row* of B is equal to zero. Additionally while some of the basic intuitions from the previous section carry over, the explicit relationship between each discriminant vector and those produced by the direct sparse LDA formulation (7) is more difficult to quantify. Regardless, we will empirically demonstrate in Section 5 that minimizing (8) with $\Omega = 0$ can be very effective in practice.

As a second option, we may instead apply sparsity penalties which operate in a row-wise fashion, meaning that instead of setting individual elements of B to zero, we look to explicitly produce zero-valued rows. This leads directly to feature pruning (meaning greater model interpretability) and a somewhat more direct relationship with the analysis from Section 3.1. To accomplish this we need only apply a standard sparsity penalty ϕ to a vector of row norms, i.e., $b = [\|\beta^1\|_2, \ldots, \|\beta^p\|_2]^\top$, where β^i denotes the i-th row of B. We may then attempt to solve

$$\begin{array}{cc} \min_{B,\Theta} & \frac{1}{N}\|Y\Theta - XB\|_F^2 + \lambda\phi(b) \\ \text{s.t.} & \Theta^\top Y^\top Y \Theta = I. \end{array} \tag{15}$$

We will now demonstrate that this model has reasonable analytical properties related to row-sparsity and is strongly connected to the canonical LDA problem.

For this purpose, let γ denote a non-negative vector of auxiliary variables and define $\Gamma = \text{diag}(\gamma)$. Then we consider the optimization problem

$$\max_{\gamma \geq 0} \quad \sum_{k=1}^{L} h\left(\nu_k(\gamma)\right) - L\lambda z(\gamma), \tag{16}$$

where $\nu_k(\gamma)$ is the kth largest eigenvalue of $\left(\Sigma_w + \lambda\Gamma^{-1}\right)^{-1/2} \Sigma_b \left(\Sigma_w + \lambda\Gamma^{-1}\right)^{-1/2}$ and z is some penalty function on γ. This construction leads to the following:

Theorem 2. *For any concave non-decreasing function ϕ, (15) is equivalent to solving (16) with some function z that is also concave and non-decreasing (sparsity-inducing), and then choosing each $\beta_k = \left(\Sigma_w + \lambda\Gamma^{-1}\right)^{-1/2} \tilde{\beta}_k$, where $\tilde{\beta}_k$ is the eigenvector associated with the optimal eigenvalue $\nu_k(\gamma)$. Moreover, any locally minimizing solution B can be achieved with at most NL nonzero rows (regardless of λ) and there exists some finite $\bar{\lambda}$ such that for all $\lambda > \bar{\lambda}$, any locally minimizing solution has no nonzero rows, i.e., $B = 0$.*

This theorem suggests the the optimal scoring problem is equivalent to a regularized Fisher LDA problem where Ω is set equal to the inverse of a diagonal factor penalized with a sparse regularizer. Returning to the original regularized LDA problems from (4) and (6), it suggests that we may equivalently inject sparsity by learning the diagonal elements of Ω in the presence of an appropriate additional regularization term. Simply put, if (4) is viewed as the ideal starting point for applying additional sparse regularizers, there exist two candidate mechanisms for injecting sparsity: we can either apply a sparse penalty to B or alternatively, to the inverse of Ω, since if elements of Ω^{-1} go to zero, the constraint can only be satisfied with corresponding rows of B being driven to zero. Both strategies are equally plausible; however, the second variant provides a close connection to optimal scoring via Theorem 2, which also then leads to efficient implementations (see Section 4). And importantly, by adding a sparsity penalty to Ω^{-1} instead of B, there is no longer any issue with degenerate, non-sparse global solutions as occurs in (7) without a judiciously chosen Ω and λ combination. Here we only need to select a single scalar parameter λ.

Additionally, unlike (9) in (Merchante et al., 2012), Theorem 2 quantifies a general equivalence without any circular dependency on the value of the optimizer of (15). Moreover, it allows us to quantify a minimal amount of row-sparsity that is to be expected from such a model. It is worth mentioning here that efficient algorithms and certain attendant analysis are possible when defining the problem either in terms of the original sparse function ϕ in B space or in terms of the auxiliary penalty z in γ space. This duality has been explored previously in the context of compressive sensing-type models (Wipf et al., 2011).

4 Robust Algorithms for Sparse LDA

We began with a variety of sparse LDA algorithms that could largely be partitioned into those based on either (7) or (8) and row-sparse extensions to (15). In searching for the

most principled adaptation for handling sparsity, the previous section then served to narrow the field by rigorously motivating the latter optimal scoring variants, an added benefit being that we no longer need any heuristic selection criteria for Ω, which winnows the possibilities even further. However, for practical application we must still select some regularization function ϕ for enforcing sparsity, and a particular strategy for minimizing the resulting penalized cost function, either (8) or (15).

For this purpose we will first consider solving (8) where $\Omega = 0$ and the sparse penalty ϕ is an arbitrary concave, non-decreasing function of coefficient magnitudes; given that each discriminant vector is penalized separately, we may expect to minimize the overall number of nonzero elements. We refer to this as an *entry-wise sparse model*. Secondly, we describe algorithms for solving (15) with arbitrary row-sparsity penalties, which we refer to as a *row sparse model*. In both cases we will first review simple iterative reweighted algorithms that scale linearly in p, a desirable property for the high-dimensional setting where $p \gg N$. From this pool we will then motivate a specific choice for ϕ, inspired by the popular relevance vector machine (RVM) from (Bishop & Tipping, 2000) and (Tipping, 2001), that is particularly well-suited for incorporation into the optimal scoring sparse LDA pipeline.

4.1 General Iterative Reweighted Algorithms

Both iterative reweighted ℓ_1 and ℓ_2 style algorithms (Wipf & Nagarajan, 2010) can be applied to either the entry-wise or row sparse models for general ϕ. Here for simplicity we consider the former for the entry-wise case and the latter for the row-spare model, in part because this sometimes leads to the simplest implementations.

Entry-wise Sparse Model: In order to solve (8) with $\Omega = 0$, we iteratively update B while holding Θ fixed and then find the optimal Θ holding B fixed. When B is fixed, the optimal Θ can be computed in closed form using the SVD decomposition of $(Y^\top Y)^{-1/2} Y^\top X B$. Next, when Θ is fixed the optimization problem for each discriminant vector β_k conveniently decouples and hence each can be computed independently. For β_k, we rewrite the relevant objective as

$$\min_{\beta_k} \frac{1}{N} \|Y\theta_k - X\beta_k\|_2^2 + \lambda\phi(\beta_k). \tag{17}$$

Solving (17) can be accomplished whenever ϕ is concave (in coefficient magnitudes) using the general iterative ℓ_1 reweighing technique summarized in (Wipf & Nagarajan, 2010). This produces the updates

$$\begin{aligned} \beta_k^{(t+1)} &\leftarrow \arg\min_{\beta} \frac{1}{N} \|Y\theta_k - X\beta\|_2^2 + \lambda \sum_i w_{ki} |x_i| \\ w_{ki}^{(t+1)} &\leftarrow \left. \frac{\partial \phi(\mu)}{\partial |\mu_i|} \right|_{\mu = \beta_k^{(t+1)}}. \end{aligned} \tag{18}$$

Note that solving for β_k here is a classical weighted Lasso problem, which can be efficiently computed using fast convex solvers, hence this procedure is computationally efficient. Although the connection to the canonical LDA problem may be somewhat more ambiguous, not surprisingly the entry-wise model can nonetheless achieve the overall sparsest discriminant matrix (see Section 5)

Row Sparse Model: In contrast to the entry-wise sparse model, we only need compute Θ once and then solve for the resulting B a single time. This observation is based directly on the following result:

Lemma 2. *When $L = K - 1$, any Θ in the subspace orthogonal to $\mathbf{1}$ and satisfying $\Theta^\top Y^\top Y \Theta = I$ leads to the same optimal B for solving (15) up to an inconsequential rotation as long as b is computed using an ℓ_2 norm.*

By Lemma 2, we can arbitrarily pick a valid Θ by a simple eigenvalue decomposition and then compute once

$$\min_B \frac{1}{N} \|Y\Theta - XB\|_F^2 + \lambda\phi(b). \tag{19}$$

This can be iteratively solved analogous to the previous section using an iterative reweighted convex program. However, because the discriminant vectors are now coupled, a single second-order cone program must replace the set of independent ℓ_1 problems from before. While there exist many efficient methods for solving such a problem, here we present a simple alternative based upon iterative ℓ_2 reweighting. An advantage is that often simple updates are available in closed form, although the convergence rate can be slower (Wipf & Nagarajan, 2010). For any concave, non-decreasing ϕ, the required iterations are

$$\begin{aligned} B^{(t+1)} &\leftarrow \widetilde{W}^{(t)} X^\top \left(\lambda N I + X \widetilde{W}^{(t)} X^\top \right)^{-1} Y\Theta \\ w_i^{(t+1)} &\leftarrow \left. \frac{\partial \phi(\mu)}{\partial \|\mu^i\|_2^2} \right|_{\mu = B^{(t+1)}} \end{aligned} \tag{20}$$

where $\widetilde{W} = \operatorname{diag}\left(w_1^{-1}, \ldots, w_p^{-1}\right)$.

4.2 Special Case: Using the RVM for Sparse LDA

The RVM represents a popular probabilistic model that has been shown to be effective for solving sparse regression problems (Bishop & Tipping, 2000; Tipping, 2001). Here we derive two simple sparse LDA variants that rely on a penalty function ϕ inspired by the RVM. These adaptations display a unique mechanism for smoothing away bad local minima that is specific to this merger with LDA, all at a computational cost equivalent to only a single RVM regression problem (at least for the row sparse LDA model). Later in Section 5 we will empirically demonstrate that these variants outperform existing sparse LDA algorithms on three important benchmark datasets. Therefore, we would argue that the RVM is notably well-suited for integrating with LDA.

For an observed vector y, the basic RVM builds from the likelihood function $p(y|\beta) = \mathcal{N}(y; X\beta, \lambda I)$ and coefficient prior $p(\beta;\gamma) = \mathcal{N}(\beta; 0, \text{diag}(\gamma))$ where $\gamma \in \mathbb{R}^p_+$. If γ were somehow known, then the posterior distribution over β is Gaussian with closed-form mean and covariance available as standard formulae. However, because it is typically not known, we can obtain an estimate via Type II maximum likelihood, which entails solving $\max_\gamma \int p(y|\beta)p(\beta;\gamma)d\beta$. Once an optimal γ is computed, β can then be estimated via the corresponding posterior mean. While not directly amenable to LDA adaptation, it is shown in (Wipf et al., 2011) that the optimal solution β_{RVM} resulting from this procedure satisfies

$$\beta_{RVM} = \arg\min_\beta \|y - X\beta\|_2^2 + \lambda g_\alpha(\beta), \quad \text{with}$$

$$g_\alpha(\beta) = \min_{\gamma \geq 0} \beta^\top \Gamma^{-1}\beta + \log|\alpha I + X\Gamma X^\top|, \; \alpha > 0. \quad (21)$$

Note that this implicit penalty (which is concave, non-decreasing, and hence favors sparsity) can be extended to the multi-column case via

$$g_\alpha(B) = \min_{\gamma \geq 0} \sum_{i=1}^{L} \beta_i^\top \Gamma^{-1}\beta_i + L\log|\alpha I + X\Gamma X^\top|. \quad (22)$$

When viewed as abstract penalty functions, it is then straightforward to derive RVM-based sparse LDA algorithms with either reweighted ℓ_1 or ℓ_2 updates. It follows that the general ℓ_1 updating rule (18) for entry-wise sparse model becomes

$$w_k{}_i^{(t+1)} \leftarrow \left[X_i^\top \left(\alpha I + X\widetilde{W}_k^{(t)} \widetilde{B}_k^{(t+1)} X^\top \right)^{-1} X_i \right]^{1/2}, \quad (23)$$

where $\widetilde{B}_k^{(t)} = \text{diag}\left(|\beta_k{}_1^{(t)}|, \ldots, |\beta_k{}_p^{(t)}| \right)$. Likewise, the general ℓ_2 updating rule (20) for row sparsity becomes

$$w_i^{(t+1)} \leftarrow \Big[\frac{1}{L} \sum_{j=1}^{L} \left(B_{j,i}{}^{(t+1)} \right)^2 + {w_i^{(t)}}^{-1} - {w_i{}^{(t)}}^{-2} X_i^\top \left(\alpha I + X\widetilde{W}^{(t)} X^\top \right)^{-1} X_i \Big]^{-1}. \quad (24)$$

We will use the entry-wise sparse model with reweighing rule (23) and row sparse model with reweighing rule (24) to handle real world datasets in the following section. While both models produce extremely sparse solutions, in specialized situations the RVM row sparse model can be proven to be a particularly sensible choice. This occurs in part because of the orthogonal structure of $Y\Theta$.

We first define spark$[X]$ as the smallest number of linearly dependent columns in a matrix X. Now consider the optimization problem

$$\min_B g_\alpha(B), \quad \text{s.t. } Y\Theta = XB, \; \alpha \to 0, \quad (25)$$

This problem can be viewed as the noiseless version of (19) with $\phi(b)$ replaced by the RVM row sparse penalty from (22). And finally, define $\tilde{X}$ as the D columns of X associated with nonzero rows in a maximally row sparse feasible solution to $Y\Theta = XB$ (i.e., the solution with the fewest number of nonzero rows, meaning D is minimal). We then have the following:

Theorem 3. *If $D \leq L < N$, spark$[X] = N + 1$, and*

$$\min_\Lambda \|A\|_2 \|A^{-1}\|_2 < \frac{N}{D} \quad \textit{s.t.} \quad \begin{array}{l} A = \Lambda\tilde{X}\tilde{X}^\top\Lambda, \\ \Lambda \textit{ a positive diagonal}, \end{array}$$

then (25) has a single stationary point that is guaranteed to be a maximally row sparse feasible solution. No row-sparse penalty of the separable form $\phi(b) = \sum_i f(\|\beta^i\|_2)$ can satisfy this result.

Here $\|\cdot\|_2$ represents the spectral norm of a matrix. Hence, as long as the columns of $\tilde{X}$ display a sufficient degree of spread, $\|A\|_2\|A^{-1}\|_2$ will be sufficiently small given the optimal diagonal weighting factor Λ. Interestingly, unlike typical sparse recovery results that rely on convex penalty functions and RIP conditions, Theorem 3 is independent of any correlation structure in the remaining columns of X, i.e. $X\backslash\tilde{X}$. In fact, the other columns only play a weak role by virtue of the spark condition, but this is extremely minor given that any design matrix X with even an infinitesimally small continuously random component will satisfy it. Therefore in this regard, the RVM maintains a substantial advantage. And while still obviously an idealized result, which relies on $Y\Theta$ being orthogonal, more standard separable penalties (convex or not) such as in (Merchante et al., 2012) cannot satisfy something similar (details to follow in a subsequent publication). At the very least, Theorem 3 motivates applying the RVM adaptation as a viable candidate for sparse LDA, that can promote maximal sparsity without incurring too many suboptimal minima.

We close this section by mentioning that, although the original RVM is very straightforward to implement for regression, it requires an additional heuristic Laplace approximation for classification and cannot be extended to multi-class problems without incurring an unacceptable complexity cost with order-K^3 dependency on the number of classes K (Tipping, 2001). Note that, although a more efficient, greedy version of the RVM has already been developed, this algorithm has not be adapted for problems with $K > 2$ nor row sparse models. Consequently, an appealing byproduct of the developments herein is an extremely efficient, general-purpose multi-class extension of the RVM.

5 Experimental Results

This section presents experimental results comparing sparse LDA models built upon the RVM (both entry-wise denoted *RVM-ent* and row-wise denoted *RVM-row*) to several existing state-of-the-art LDA methods: the GLOSS

algorithm (Merchante et al., 2012), ℓ_1-PLDA (Witten & Tibshirani, 2011), and SDA (Clemmensen et al., 2011). Because of limited space and for ease of direct comparison in the high-dimensional setting, we select the same three gene datasets used in (Merchante et al., 2012; Witten & Tibshirani, 2011). These are three gene datasets. The *Ramaswamy* dataset (Ramaswamy et al., 2001) contains 198 samples of 16063 gene expression measurements from 14 distinct cancer subtypes. The *Nakayama* dataset (Nakayama et al., 2007) contains contains 195 samples of 22283 gene expression measurements from 10 types of soft tissue tumours. Consistent with previous experiments using ℓ_1-PLDA and GLOSS, we only consider the 5 main types of 86 samples. Finally, the *Sun* data (Sun et al., 2006) contains 180 samples of 54613 gene expression measurements from 4 classes of tumors.

We follow the training and testing protocol of (Merchante et al., 2012) and (Witten & Tibshirani, 2011). Each dataset was split into a training portion containing 75% of the samples and a testing set containing the remaining 25%. This process is repeated 10 times with random choice of the split. The tuning parameters for the entry-wise sparse RVM model are obtained by 10-fold cross validation (using only the 75%). However, for the row sparse RVM model the tuning parameter was simply assigned to a constant value prior to the random splits and all the splits used the same parameter for simplicity. This constant value was determined based on grid-search over another random split.

The test error rates and corresponding sparsity metrics are presented in Table 1 along with standard deviations. For ℓ_1-PLDA, SDA, and GLOSS, the results were obtained directly from (Merchante et al., 2012) and (Witten & Tibshirani, 2011); for both RVM models we use our own simple implementation. Additionally, there are two criteria for evaluating the sparsity, row- or equivalently feature-wise sparsity (#FEATURE) and entry-wise sparsity (#ENTRY). SDA fails to return a solution on the *Ramaswamy* data because of numerical instabilities (Merchante et al., 2012). In all cases, an RVM model achieves better accuracy with dramatically fewer nonzeros. Additionally, the ℓ_1-PLDA algorithm performs far worse than all of the other algorithms which are based on optimal scoring, consistent with our analysis presented previously and the requirement that a heuristic modification of Σ_w, e.g., by adding an $\Omega > 0$, is required to avoid degeneracy, which can counteract the effects of sparse penalties.

While the RVM performs well, there are of course other non-convex penalty functions that could potentially boost sparsity beyond existing convex surrogates in the context of LDA. However, care must be taken such that local minima do not disrupt performance. In this context, we incorporated a non-convex Gaussian entropy penalty function (Figueiredo, 2001) commonly used for sparse estimation; however, the performance was substantially worse than the RVM because of local minima. We also performed tests with the recent SULDA algorithm from (Zhang & Chu, 2013) using code provided by the authors, but the results were not competitive on these high-dimensional datasets. The accuracy is considerably worse than the RVM on the Ramaswamy and Nakayama data and extremely poor for the Sun data. Consequently, for simplicity we do not include either Gaussian entropy or SULDA results in Table 1. Finally, (Cai & Liu, 2011) recently derived a slightly different convex sparse LDA variant; however, this algorithm only applies when $K = 2$ and does not scale well with p (reported results require $p \leq 3000$).

Table 1: Comparisons between different models on three gene datasets. ERR% denotes the error rate for test split. #FEATURE denotes the number of non-zero rows in discriminant matrix B (i.e., number of selected features). #ENTRY is the number of non-zero entries in B. Standard deviations are shown in parentheses. The best performance for each criteria is in bold.

MODEL	ERR%	#FEATURE	#ENTRY
Ramaswamy: $N = 198$, $p = 16063$, $K = 14$			
ℓ_1-PLDA	38.36(6.0)	14874(720)	14874(720)
SDA	-	-	-
GLOSS	20.61(6.9)	372.4(122.1)	4841.2(1457)
RVM-ENT	17.76(6.2)	1940.5(31.9)	**1940.5(31.9)**
RVM-ROW	**16.73(6.1)**	**218.7(7.3)**	2843.1(94.9)
Nakayama: $N = 86$, $p = 22283$, $K = 5$			
ℓ_1-PLDA	20.95(1.3)	10479(2116)	10479(2116)
SDA	25.71(1.7)	252.5(3.1)	252.5(3.1)
GLOSS	20.48(1.4)	129.0(18.6)	516.0(74.4)
RVM-ENT	**19.52(4.2)**	74.3(7.7)	**74.3(7.7)**
RVM-ROW	20.00(1.2)	**61.5(7.9)**	246.0(31.6)
Sun: $N = 180$, $p = 54613$, $K = 4$			
ℓ_1-PLDA	33.78(5.9)	21635(7443)	21635(7443)
SDA	36.22(6.5)	384.4(16.5)	384.4(16.5)
GLOSS	31.77(4.5)	93.0(93.6)	279.0(280.8)
RVM-ENT	**30.00(4.9)**	86.0(4.7)	**86.0(4.7)**
RVM-ROW	30.68(4.3)	**36.1(3.6)**	108.3(10.8)

6 Conclusion

Our starting point was the fragmented understanding of various sparse LDA algorithms, which can generally be partitioned into two broad categories, optimal scoring-based and canonical Fisher LDA-based. Contrary to conventional wisdom, we demonstrated that the former actually maintains an intrinsic advantage over the latter when sparse penalties are introduced. We further motivated a very specific penalty function inspired by the RVM that assimilates particularly well with optimal scoring, leading to state-of-the-art performance with theoretical support. A natural byproduct of this process is an extremely simple, scalable multi-class extension of the RVM that relies on no additional approximation steps.

References

Bishop, Christopher M and Tipping, Michael E. Variational relevance vector machines. In *Proceedings of the 16th conference on Uncertainty in artificial intelligence*, pp. 46–53, 2000.

Cai, Tony and Liu, Weidong. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496), 2011.

Clemmensen, Line, Hastie, Trevor, Witten, Daniela, and Ersbøll, Bjarne. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

Dundar, Murat, Fung, Glenn, Bi, Jinbo, Sandilya, Sathyakama, and Rao, R Bharat. Sparse fisher discriminant analysis for computer aided detection. In *SDM*. SIAM, 2005.

Figueiredo, Mário. Adaptive sparseness using jeffreys prior. In *Advances in neural information processing systems*, pp. 697–704, 2001.

Fisher, Ronald A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Fung, Eric S and Ng, Michael K. On sparse fisher discriminant method for microarray data analysis. *Bioinformation*, 2(5), 2007.

Grosenick, Logan, Greer, Stephanie, and Knutson, Brian. Interpretable classifiers for fmri improve prediction of purchases. *Neural Systems and Rehabilitation Engineering, IEEE Trans.*, 16(6):539–548, 2008.

Hastie, Trevor, Tibshirani, Robert, and Buja, Andreas. Flexible discriminant analysis by optimal scoring. *Journal of the American statistical association*, 89(428): 1255–1270, 1994.

Hastie, Trevor, Buja, Andreas, and Tibshirani, Robert. Penalized discriminant analysis. *The Annals of Statistics*, pp. 73–102, 1995.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, Hastie, T, Friedman, J, and Tibshirani, R. *The elements of statistical learning*, volume 2. Springer, 2009.

Leng, Chenlei. Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection q1 using microarray data. 2007.

Merchante, Luis Francisco Sanchez, Grandvalet, Yves, and Govaert, Gerrad. An efficient approach to sparse linear discriminant analysis. In Langford, John and Pineau, Joelle (eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pp. 1167–1174, New York, NY, USA, July 2012. Omnipress. ISBN 978-1-4503-1285-1.

Moghaddam, Baback, Weiss, Yair, and Avidan, Shai. Generalized spectral bounds for sparse lda. In *Proceedings of the 23rd international conference on Machine learning*, pp. 641–648. ACM, 2006.

Nakayama, Robert, Nemoto, Takeshi, Takahashi, Hiro, Ohta, Tsutomu, Kawai, Akira, Seki, Kunihiko, Yoshida, Teruhiko, Toyama, Yoshiaki, Ichikawa, Hitoshi, and Hasegawa, Tadashi. Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma. *Modern pathology*, 20 (7):749–759, 2007.

Ramaswamy, Sridhar, Tamayo, Pablo, Rifkin, Ryan, Mukherjee, Sayan, Yeang, Chen-Hsiang, Angelo, Michael, Ladd, Christine, Reich, Michael, Latulippe, Eva, Mesirov, Jill P, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

Rao, Bhaskar D, Engan, Kjersti, Cotter, Shane F, Palmer, Jason, and Kreutz-Delgado, Kenneth. Subset selection in noise based on diversity measure minimization. *Signal Processing, IEEE Trans.*, 51(3):760–770, 2003.

Sun, Lixin, Hui, Ai-Min, Su, Qin, Vortmeyer, Alexander, Kotliarov, Yuri, Pastorino, Sandra, Passaniti, Antonino, Menon, Jayant, Walling, Jennifer, Bailey, Rolando, et al. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer cell*, 9(4):287–300, 2006.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Tipping, Michael E. Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211–244, 2001.

Wipf, David and Nagarajan, Srikantan. Iterative reweighted l1 and l2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 4(2): 317–329, 2010.

Wipf, David P, Rao, Bhaskar D, and Nagarajan, Srikantan. Latent variable bayesian models for promoting sparsity. *Information Theory, IEEE Trans.*, 57(9):6236–6255, 2011.

Witten, Daniela M and Tibshirani, Robert. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011.

Wu, Michael C, Zhang, Lingsong, Wang, Zhaoxi, Christiani, David C, and Lin, Xihong. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009.

Zhang, Xiaowei and Chu, Delin. Sparse uncorrelated linear discriminant analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 45–52, 2013.