
Generalized Spectral Bounds for Sparse LDA

Baback Moghaddam

Mitsubishi Electric Research Laboratories (MERL), Cambridge MA 02139 USA

BABACK@MERL.COM

Yair Weiss

The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

YWEISS@CS.HUJI.AC.IL

Shai Avidan

Mitsubishi Electric Research Laboratories (MERL), Cambridge MA 02139 USA

AVIDAN@MERL.COM

Abstract

We present a discrete spectral framework for the sparse or *cardinality-constrained* solution of a generalized Rayleigh quotient. This NP-hard combinatorial optimization problem is central to supervised learning tasks such as sparse LDA, feature selection and relevance ranking for classification. We derive a new generalized form of the *Inclusion Principle* for variational eigenvalue bounds, leading to exact and *optimal* sparse linear discriminants using branch-and-bound search. An efficient greedy (approximate) technique is also presented. The generalization performance of our sparse LDA algorithms is demonstrated with real-world UCI ML benchmarks and compared to a leading SVM-based gene selection algorithm for cancer classification.

1. Introduction

Feature selection for classification is quickly becoming an integral part of machine learning applications in both scientific and commercial domains. Examples range from text/information-retrieval to bioinformatics and sensor networks. Much of the recent attention has focused on the identification, categorization and evaluation of various selection algorithms (Blum & Langley, 1997; Guyon & Elisseeff, 2003). There are generally three types of feature selection methods: *filters*, *wrappers* and *embedded* techniques (Kohavi & John, 2003), depending on whether the core selection mechanism is independent and causally precedent to the classification stage (filter), is iteratively refined

based on classifier outputs (wrapper) or is an essential component of the classifier training itself (embedded).

Sparseness naturally constitutes one type of selection mechanism which is typically incorporated by means of continuous optimization with l_1 -norm penalty terms and/or “relevance priors.” Representative examples include sparse regression (Tibshirani, 1995) and sparse PCA (Zou et al., 2004), from both the supervised and unsupervised domains, respectively. A related class of *cardinality-constrained* optimization problems relying on Integer Programming (IP) are nowadays routine in operations research, leading to discrete and combinatorial search algorithms.

In this paper, we present a computational framework for a novel feature selection *filter*, using only the 2nd-order statistics (covariances), as needed for (Fisher) Linear Discriminant Analysis (LDA). We propose a *discrete* spectral formulation based on variational modes of the Courant-Fischer “Min-Max” theorem for eigenvalue maximization, as specifically adapted to *cardinality-constrained* subspaces (variable subsets). This methodology is the direct (supervised) extension of our previous framework for sparse PCA using variational eigenvalue bounds (Moghaddam et al., 2006) and thereby constitutes a more general formulation — *i.e.*, it subsumes sparse PCA as a special case of sparse LDA. As shown previously, a discrete formulation reveals a simple post-processing (renormalization) step for improving *any* approximate solution while providing bounds on its (sub)optimality. More importantly, the discrete approach leads to *exact* and provably *optimal* solutions using branch-and-bound search. We demonstrate the power of both greedy and exact sparse LDA algorithms with experiments on real-world datasets and also present summary findings from an extensive comparative study using Monte Carlo (MC) evaluation.

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

1.1. Background

In the supervised domain, the prototypical machine learning task is that of binary classification where given a *finite* number of input-output training pairs (x, y) from an unknown probability distribution, we try to learn (estimate) a function $f(x) : \mathcal{R}^n \rightarrow \{\pm 1\}$ from a preset function class F such that future (un-labeled) test data drawn from the same distribution are correctly classified. The simplest function class is a linear perceptron: $f(x) = \text{sign}(w^T x + b)$, for which *sparsity* corresponds to a weight vector w with many zero elements, thereby indicating that only few of the variables x_i actually participate in the decision rule $f(x)$. In the resulting *lower-dimensional* subspace, the variable subset selected forms a linear hyperplane which then discriminates between the two classes.

A representative optimization algorithm for (kernel) Fisher linear discriminant was given by (Mika et al., 2001) which is instructive to review as it is also a good example of an *embedded* variable selection technique. In very general terms, it is formulated as the following mathematical programming problem,

$$\begin{aligned} \min \quad & \|w\|_p^p + C \|\zeta\|_q^q \\ \text{subject to} \quad & y_i(w^T x_i + b) = 1 - \zeta_i \end{aligned} \quad (1)$$

where $p = q = 2$ is its *regularized* form and by setting $p = 1$ we obtain the *sparse* Fisher discriminant (SFD). Note the similarity to the formulation of SVMs where the inequality constraints are replaced by equalities and positivity is relaxed on the slack variables ζ_i . SVM training minimizes the L_2 -norm of w ($p = 2$) for a wide margin. Meanwhile, KKT complementarity leads to a sparse vector ζ which is typically penalized with an l_1 -norm ($q = 1$). One key difference from a classification standpoint is that SVMs maximize the *minimum* margin whereas LDA-based discriminants tend to maximize the *average* margin.

In unsupervised learning, PCA (factor analysis) is an essential tool for modeling and representation of data. Despite its power and popularity, a key drawback is the lack of sparseness (*i.e.*, factor loadings are linear combinations of *all* the input variables). Yet sparse representations are generally desirable since they aid in human understanding (*e.g.*, with gene expression data), reduce computational costs and can even promote better generalization. In machine learning, input sparseness is closely related to variable selection and automatic relevance determination, problems of enduring interest to the learning community.

Recently, (Zou et al., 2004) proposed a sparse PCA algorithm (SPCA) using their “Elastic Net” framework for l_1 -penalized regression on regular PCs.

Subsequently, (d’Aspremont et al., 2004) relaxed the “hard” cardinality constraint with a simpler *convex* approximation using semi-definite programming (SDP) for a more “direct” formulation (called DSCPA). In contrast, an alternative *discrete* spectral framework was recently proposed by (Moghaddam et al., 2006), using variational eigenvalue bounds on the covariance “sub-spectrum” as defined by the *inclusion principle*, which yielded substantial performance gains using a simple *greedy* technique (GSPCA). In addition, an exact *optimal* algorithm (ESPCA) based on branch-and-bound search was given. We will now extend this framework to the *supervised* case of sparse LDA, cast as a *generalized* eigenvalue problem $Ax = \lambda Bx$, but in a *sparse* form. We will also draw some unifying connections between sparse PCA/LDA algorithms.

2. Sparse LDA as Generalized EVD

Classical Fisher or Linear Discriminant Analysis (LDA) can be formulated as a *generalized* eigenvalue decomposition (EVD), where given a pair of symmetric positive-semidefinite matrices $A, B \in \mathcal{S}_+^n$, corresponding to the *between-class* and *within-class* covariance matrices respectively, we seek to maximize a class-separability criterion defined by the *generalized* Rayleigh quotient: $R(x; A, B) = (x^T A x) / (x^T B x)$ where B is now assumed positive definite. Since this quotient is invariant to the magnitude of x , we can reformulate the problem in terms of a quadratically-constrained quadratic program (QCQP):

$$\begin{aligned} \max \quad & x^T A x \\ \text{subject to} \quad & x^T B x = 1 \end{aligned} \quad (2)$$

Fortunately, this problem has a closed-form solution obtained by differentiating the corresponding Lagrangian, yielding $Ax = \lambda Bx$ with the *determinantal* characteristic equation $\det(A - \lambda B) = 0$. Hence, the optimal x is the eigenvector corresponding to the largest root of the resulting characteristic polynomial in λ — or equivalently, the largest eigenvalue of $B^{-1}A$. Hereafter, we will denote eigenvalue rank in *increasing* order of magnitude, thus $\lambda_{\min} = \lambda_1$ and $\lambda_{\max} = \lambda_n$.

We can now define the *sparse* LDA optimization in terms of the following *cardinality-constrained* QCQP:

$$\begin{aligned} \text{Sparse LDA :} \quad \max \quad & x^T A x \\ \text{subject to} \quad & x^T B x = 1 \\ & \text{card}(x) = k \end{aligned} \quad (3)$$

The feasible set is all *sparse* $x \in \mathcal{R}^n$ with k non-zero elements and $\text{card}(x)$ as their l_0 -norm. Unfortunately, this optimization problem is *non-convex*, NP-hard and therefore generally intractable.

Note that the special case of $B = I$ defaults to a sparse maximal-variance QP which is equivalent to sparse PCA. Therefore, any strategy for the sparse LDA in Eq. (3) can also solve sparse PCA. To make this equivalence explicit, it is sufficient (and instructive) to view this generalized EVD as a standard eigenvalue problem in the non-singularly transformed space induced by the *bijection* $y = B^{1/2}x$

$$\begin{aligned} \max \quad & y^T C y \\ \text{subject to} \quad & y^T y = 1 \\ & \text{card}(B^{-1/2}y) = k \end{aligned} \quad (4)$$

where $C = B^{-1/2}AB^{-1/2}$. Except for the cardinality constraint, this is a standard Rayleigh quotient in terms of C which has the *same* eigenvalues as $B^{-1}A$ (but *not* the same eigenvectors). Without the cardinality constraint, this standard Rayleigh quotient obeys the analytic bounds $\lambda_{\min}(C) \leq y^T C y / y^T y \leq \lambda_{\max}(C)$, where unlike $B^{-1}A$, the new matrix C is *symmetric* by construction.

Despite the odd cardinality constraint on $B^{-1/2}y$, the above reformulation may provide a potentially useful method for adapting existing sparse PCA algorithms — *e.g.*, SPCA (Zou et al., 2004) or DSPCA (d’Aspremont et al., 2004) — to find sparse *discriminant* factors (to the best of our knowledge this reformulation has not been attempted). Another (perhaps simpler) alternative is to use the equivalence of Fisher linear discriminant to a least-squares regression (on suitably scaled output labels) and add an l_1 -norm penalty term as in Lasso for subset selection (Tibshirani, 1995).

In contrast, we approached sparse LDA using the same discrete variational framework developed in (Moghaddam et al., 2006), motivated by the goal of finding *exact* and *optimal* discriminants — with optimality defined by the generalized Rayleigh quotient. We will next show how the *spectrum* of C (equivalently that of $B^{-1}A$) plays a key role in the design of exact and optimal sparse LDA algorithms.

2.1. Optimality Conditions

First let us consider what conditions *must* be true if the *oracle* revealed the optimal solution: a sparse vector $x \in \mathcal{R}^n$ with cardinality k yielding the *maximum* objective value R^* . This would necessarily imply that

$$R(x; A, B) = \frac{x^T A x}{x^T B x} = \frac{z^T A_k z}{z^T B_k z} \quad (5)$$

where $z \in \mathcal{R}^k$ contains the k non-zero elements in x and (A_k, B_k) are the $k \times k$ principal submatrices

of (A, B) obtained by deleting the rows and columns corresponding to the zero indices of x — equivalently, by extracting the rows and columns of non-zero indices. The k -dimensional quadratic form in z is equivalent to a standard *unconstrained* generalized Rayleigh quotient. Since this subproblem’s maximum objective value is $\lambda_{\max}(A_k, B_k)$, this therefore *must* be the optimal objective value R^* . We now summarize this key observation in the following proposition.

Proposition 1. The optimal value R^* of the sparse LDA optimization problem in Eq.(3) is equal to $\lambda_{\max}(C_k^*)$, where $C_k \stackrel{\text{def}}{=} B_k^{-1/2} A_k B_k^{-1/2}$ is $k \times k$ and C_k^* in particular is the one submatrix pair with *largest* maximal generalized eigenvalue. Moreover, the non-zero sub-vector z^* of the optimal x^* is equal to the inverse *bijection* of the principal eigenvector v_k of C_k^*

$$z^* = B_k^{-1/2} v_k, \quad v_k^T C_k^* v_k = \lambda_{\max}(C_k^*) \quad (6)$$

This therefore reveals the true combinatorial nature of sparse LDA (and equivalent cardinality-constrained optimization problems), wherein solving for the optimal solution is inherently a discrete search for the k indices which maximize λ_{\max} of the *subproblem* (A_k, B_k) . While such an exact definition of optimality is illuminating, it does not suggest an efficient method for actually *finding* the optimal subproblem, short of an exhaustive search which is impractical for $n > 30$ due to the exponential growth of candidate submatrices. Nevertheless, exhaustive search is a viable method for small n that *guarantees* optimality for “toy problems” and small real-world datasets, thus calibrating the *quality* of approximations (via the optimality gap). Moreover, it suggests a simple but effective “fix” for *improving* approximate factors obtained by other algorithms — *e.g.*, by SVMs.

Proposition 2. Let \tilde{x} be a candidate solution with (approximate) cardinality k found by any method. Let \tilde{z} be the non-zero subvector of \tilde{x} and v_k be the principal generalized eigenvector of (A_k, B_k) , as indexed by the non-zero indices of \tilde{x} . If $\tilde{z} \neq v_k(A_k, B_k)$, then \tilde{x} is *not* optimal. However, replacing \tilde{x} ’s nonzero elements with v_k in Eq.(6) will guarantee an increase in $R(\tilde{x}; A, B)$.

This *variational renormalization* suggests that continuous relaxations are only useful in providing a sparsity pattern with which to solve a smaller *unconstrained* subproblem (A_k, B_k) . In effect, their factor loadings are more sub-optimal than need be and should be renormalized. Indeed, the common *ad-hoc* technique of “simple thresholding” (ST) for sparse PCA (*i.e.*, setting the smallest absolute value loadings of the principal eigenvector to zero and renormalizing it to unit-norm) can be enhanced by applying this “fix.”

2.2. Variational Eigenvalue Bounds

We saw that the generalized eigenvalues of $Ax = \lambda Bx$ play a fundamental role in *defining* sparse LDA factors of a given cardinality k — as the generalized eigenvalues associated with the principal submatrices (A_k, B_k) . Not surprisingly, the two eigenvalue spectra can be related by the following result.

Theorem 1 *Generalized Inclusion Principle.* Let the pair (A, B) be $n \times n$ symmetric matrices with generalized spectrum $\lambda_i(A, B)$, with B positive definite. Let (A_k, B_k) be a corresponding pair of $k \times k$ principal submatrices with $1 \leq k \leq n$, with generalized eigenvalues $\lambda_i(A_k, B_k)$. Then, for all i , $1 \leq i \leq k$

$$\lambda_i(A, B) \leq \lambda_i(A_k, B_k) \leq \lambda_{i+n-k}(A, B) \quad (7)$$

Proof: Our proof (see Appendix) is an extension of a classic proof of the original (non-generalized) eigenvalue inclusion principle, derived by imposing a sparsity pattern of cardinality k as an additional subspace orthogonality constraint on the variational form of the Courant-Fischer “Min-Max” theorem.

In other words, the generalized eigenvalues of (A, B) form upper and lower bounds for the generalized eigenvalues of all their principal submatrices (A_k, B_k) . Therefore, the spectra of (A_m, B_m) and (A_{m+1}, B_{m+1}) interleave or *interlace* each other, with the eigenvalues of the larger matrix pair “bracketing” those of the smaller one.¹ For *positive-definite* symmetric matrices (covariances), augmenting A_m to A_{m+1} (adding a new variable) will always *expand* the spectral range: reducing λ_{\min} and increasing λ_{\max} . This *monotonicity* property has important theoretical as well as practical consequences for greedy and exact combinatorial algorithms, as we will see in the next section.

Since the solution of sparse LDA seeks to *maximize* the generalized Rayleigh quotient, the relevant inequality in Eq.(7) has $i = k$, thus yielding the inclusion bounds

$$\lambda_k(A, B) \leq \lambda_{\max}(A_k, B_k) \leq \lambda_n(A, B) \quad (8)$$

which shows that the k -th *smallest* generalized eigenvalue of (A, B) is a lower bound for the *class-separability* criterion of sparse LDA with cardinality k . The eigenvalue bound $\lambda_k(A, B)$ is also useful for speeding up branch-and-bound search with various predictive pruning techniques (Somol et al., 2004). We note that the right-hand inequality in Eq.(8) is a fixed (often loose) upper bound $\lambda_{\max}(A, B)$ for *all* k . However, branch-and-bound algorithms mostly work with *intermediate* subproblems (A_m, B_m) with

¹The well-known eigenvalue “interlacing” property comes from the basic inclusion principle with $k = n - 1$.

$k \leq m \leq n$, and will invariably encounter *smaller* submatrices with *tighter* bounds $\lambda_{\max}(A_m, B_m)$ which eventually *fathom* most branches of the search tree.

2.3. Combinatorial Optimization

In view of our discrete formulation and the generalized inclusion principle, binary Integer Programming (IP) techniques like branch-and-bound (Nemhauser & Wolsey, 1988) seem ideally suited for sparse LDA. Greedy techniques like *backward elimination* can also exploit the monotonic nature of successively *nested* submatrices and their “bracketing” eigenvalues: start with the full index set $I = \{1, 2, \dots, n\}$ and sequentially delete the variable j which yields the maximum $\lambda_{\max}(A_{\setminus j}, B_{\setminus j})$ until only k elements remain. For *small* cardinalities $k \ll n$, the computational cost of backward search can grow to near maximum complexity $\approx O(n^4)$. Hence its counterpart *forward selection* is often preferred: start with the null index set $I = \{\}$ and sequentially add the variable j which yields the maximum $\lambda_{\max}(A_{+j}, B_{+j})$ until k elements are selected. Forward search has *worst-case* complexity $< O(n^3)$. A powerful greedy strategy is a *bi-directional* search: run a forward pass (from 1 to n) plus a second (independent) backward pass (from n to 1) and pick the better solution at each k . We call this dual-pass algorithm *greedy* sparse LDA or **GSLDA**.

Despite the expediency of near-optimal greedy search, it is nevertheless worthwhile to invest in optimal solution strategies, especially if the sparse LDA problem is in a critical application domain like bioinformatics, where even a small optimality gap could lead to costly diagnostic failures. As with (Ko et al., 1995), our branch-and-bound relies on computationally efficient bounds, in our case the upper bound in Eq.(8) computable by the *power* method, for all active *subproblems* in a (FIFO) queue for *depth-first* search. The *lower* bound in Eq.(8) can be used to sort the queue for a more efficient *best-first* search. Our *exact* sparse LDA algorithm (called **ESLDA**) is *guaranteed* to terminate with the optimal discriminant. Naturally, the total search time depends on the *quality* of the starting candidate in the branch-and-bound initialization. The solutions found by our dual-pass greedy search (**GSLDA**) were ideal for initializing **ESLDA**, as their generalized Rayleigh quotient was typically near-optimal. However, we should note that even with good initialization, branch-and-bound search can still take a long time — *e.g.* ≈ 2 hours for $n = 40, k = 20$. In actual practice, preset thresholds based on generalized eigenvalue bounds can be used for early (premature) termination at the desired goal.

After extensive evaluation, we found that the most cost-effective strategy was to first run GSLDA (or at least the forward pass) and then either settle for its (near-optimal) discriminant or else use this to initialize ESLDA for a branch-and-bound search for the *optimal* discriminant. A full GSLDA run has the added benefit of giving near-optimal solutions for *all* cardinalities at once, with running times that are typically far less demanding than finding a single- k approximation with most continuous methods — *e.g.*, with SVMs.

3. Experiments

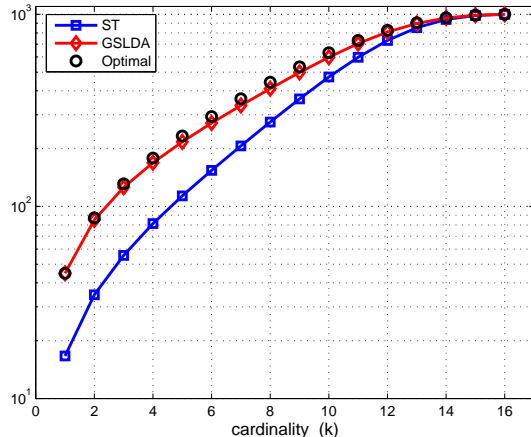
We evaluated GSLDA (and validated ESLDA) with various synthetic covariance matrices of size $10 \leq n \leq 40$, as well as real-world datasets from the UCI ML Repository with very encouraging results. We present a few representative examples in order to illustrate the advantages of using the discrete methodology advocated. In particular, we compare our performance with continuous approximation techniques like “simple thresholding” (ST) — *i.e.* thresholding the largest eigenvector of $C = B^{-1/2}AB^{-1/2}$ — but *with* the variational renormalization “fix” in post-processing.

We also compared both our discrete sparse LDA algorithms to more traditional feature selection (ranking) techniques such as the Pearson’s correlation coefficient between the individual variables x_i and their class labels $y_i \in \{\pm 1\}$, as typically used with DNA micro-arrays for gene selection and pruning.

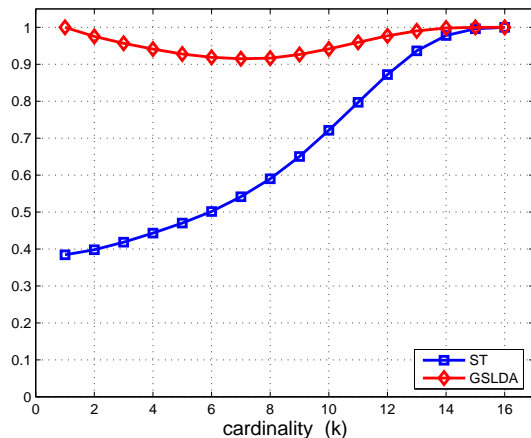
We first present experimental results for the synthetic datasets in our Monte Carlo evaluation, followed by two real-world datasets from the UCI ML Repository and then conclude this section with gene expression profiles from a DNA micro-array dataset for doing gene selection for diagnostic cancer classification.

3.1. Monte Carlo Evaluation

We give a representative summary of an extensive Monte Carlo (MC) evaluation of GSLDA against a simple continuous algorithm (ST). In order to show the typical or *average-case* performance, we present results with random covariance matrices from synthetic stochastic Brownian processes of various degrees of smoothness, ranging from sub-Gaussian to super-Gaussian. Every MC run consisted of 50,000 covariance matrices and the (normalized) generalized Rayleigh quotient $R(k)$ for each cardinality. For each sampled pair, ESLDA was used to find the *optimal* solution as “ground truth” for subsequent calibration and evaluation. Figure 1(a) shows the ensemble mean generalized Rayleigh quotient (for $n = 16$), which



(a)



(b)

Figure 1. Monte Carlo evaluation: (a) mean generalized Rayleigh score *vs.* k and (b) mean optimality ratio (captured-to-optimal Rayleigh quotients). Based on 50,000 random matrix pairs (A, B) . The algorithms shown are ST (blue □), GSLDA (red ◇) and optimal ESLDA (black ○).

demonstrates how close our greedy algorithm comes to achieving optimality. In part (b) we plot the ratio of captured-to-optimal Rayleigh quotient, where GSLDA is seen to capture *at least* 90% of the maximum LDA score across all cardinalities — its poorest performance (10% suboptimality) occurs at “half-sparsity” ($k = n/2$) corresponding (not surprisingly) to maximal combinatorial perplexity of candidates.

3.2. UCI ML Benchmarks

We next applied our discrete algorithms (GSLDA and ESLDA) to well-known and well-studied benchmark datasets from the UCI ML Repository; specifically, the medium-sized datasets Sonar ($n=60$) and Ionosphere

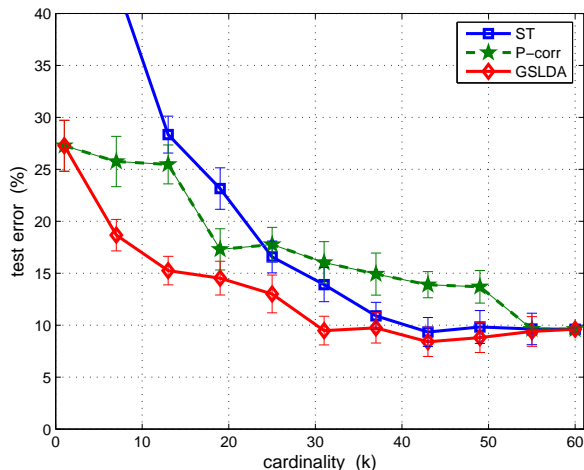


Figure 2. **Sonar** variable selection: cross-validation test error *vs.* cardinality (k). The 3 algorithms shown are ST (blue \square), P-correlation (green \star) and **GSLDA** (red \diamond).

($n=36$), which represent typical application domains in geospatial sensing where sparsity is required due to limited budgets and the high cost of (permanent) installation and maintenance of equipment. A remote sensing or environmental monitoring application will typically consist of multiple sites making homogeneous measurements of a possibly redundant nature — *e.g.*, data of the same modality at different locations or spatial/frequency channels.

Figure 2 is a summary of our experiments on **GSLDA** for greedy variable selection on the **Sonar** dataset, showing the generalization error (and error-bars) from 100 randomized trials of 5-fold cross-validation for each k . In each CV run, 80% of the samples were used for covariance estimation and subsequent variable selection — by directly solving for the sparse LDA factor in Eq.(6) — and 20% for subsequent testing. Throughout our experiments with real datasets (especially DNA micro-arrays in the next section) all rank-deficient within-class covariance matrices B were automatically regularized: $B \leftarrow B + \alpha I$ with α set smaller than $\lambda_{\min}(B)$ by $O(10^3)$. This not only constitutes sound numerical computing practice but also provides the necessary *shrinkage* needed to avoid over-fitting small samples. Figure 2 shows that **GSLDA** yields the best generalization performance, especially in the truly sparse regime of $k < n/2$. In fact, unlike correlation-based ranking (P-corr) and simple thresholding (ST) whose test error begins to increase as soon as we discard more than just a few variables, **GSLDA** maintains essentially the *same* level of performance with only half the number of features ($k = 30$). This was especially fortuitous as the

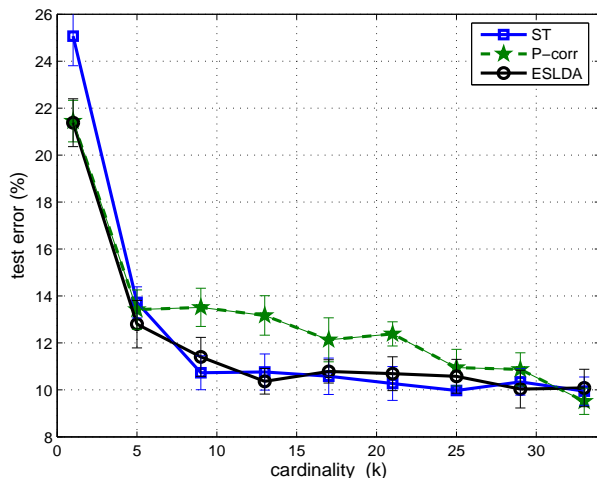


Figure 3. **Ionosphere** variable selection: cross-validation test error *vs.* cardinality (k). Algorithms are ST (blue \square), P-correlation (green \star) and the *optimal* **ESLDA** (black \circ).

majority of the 30 variables eliminated corresponded to the high-frequency band of the sensor. The “half-sparsity” test error of **GSLDA** (9% at $k = 30$) is as good as any reported on this dataset *without* sparsity (*i.e.*, with all the variables).

Figure 3 is the summary of feature selection experiments on the **Ionosphere** dataset, which consists of “interesting” scattering phenomena measured by bouncing radio waves off of electrons in the ionosphere with 16 high-frequency antennae as part of a phased array radar system with a fixed power budget. This is a good example of the type of remote sensing application where sparsity translates to power (money) saved, if not increased discrimination accuracy. Indeed, this does appear to be the case here since even though the majority of variables can be eliminated, there is no discernible gain in classification performance. This example is also instructive in that it shows that *optimality* in the generalized Rayleigh quotient sense (here obtained with **ESLDA**) does not necessarily translate to improved generalization performance, as the “fixed” sub-optimal **ST** method appears to yield essentially the same accuracy. The error rate at “half-sparsity” ($\approx 11\%$ at $k = 16$) is competitive with past benchmarks established with neural networks and SVMs using all variables. Note that the correlation-based filter is once again found to be inferior to our spectral algorithm.

3.3. Gene Classification

We applied the forward pass of our **GSLDA** algorithm to the colon cancer dataset of (Alon et al., 1999),

with a total of 62 tissue samples (22 normal and 40 cancerous) with the expression profiles of $n = 2000$ genes (this dataset was an obvious candidate for regularization of B). Following the methodology used in (Guyon et al., 2002), we first ran our variable selection — (A, B) covariance estimation and subsequent GSLDA — on the entire dataset of 62 samples and then computed the *leave-one-out* error rates for various values of k (number of genes selected) and compared them to other techniques: simple thresholding (ST), correlation ranking and the SVM-embedded *recursive feature elimination* (SVM-RFE) of (Guyon et al., 2002). The resulting error curves are shown in Figure 4 where the RFE-SVM results are taken from (Guyon et al., 2002) (see their Fig.4). Note that GSLDA is very competitive with RFE-SVM, especially at high levels of sparsity ($k \leq 10$). The difference was most pronounced (15%) at the extreme case of a single discriminant gene ($k = 1$).

We should note a certain caveat here in regards to the optimistic *bias* introduced by training the feature selection algorithm on the *entire* dataset before doing cross-validation. Clearly, this is not the purist’s view of what cross-validation was meant to achieve. In fact, the authors of the RFE-SVM study (Guyon et al., 2002) have since published a retraction regarding this “flawed” methodology (with discussion). However, the true skeptic would discount the *zero* error rates of RFE-SVM and GSLDA in Figure 4 *anyway*, fully expecting to see poor(er) accuracy on the next batch of (truly held-out) tissue samples. Nevertheless, this protocol is still valid for judgments of *relative* merit when comparing different algorithms, which is indeed our sole intention with Figure 4.

4. Discussion

We presented an *exact* variational framework for sparse LDA, complete with requisite eigenvalue bounds and two discrete algorithms: fast and effective greedy search (GSLDA) and a less efficient but *optimal* method (ESLDA). In addition, we gave a simple renormalization “fix” for *any* continuous approximation (relaxation). Indeed, the “straw-man” of simple thresholding (ST) was seen to be adequate (when *fixed*, naturally) but *not* always reliable.

Note that since binary classification results in a rank-1 A matrix, it is mostly the eigen-structure of the within-class B matrix that governs the performance of continuous approximations (discrete methods are not affected as much as long as a small regularization term is added for numerical stability). Of course, sparse LDA is not restricted to binary classification.

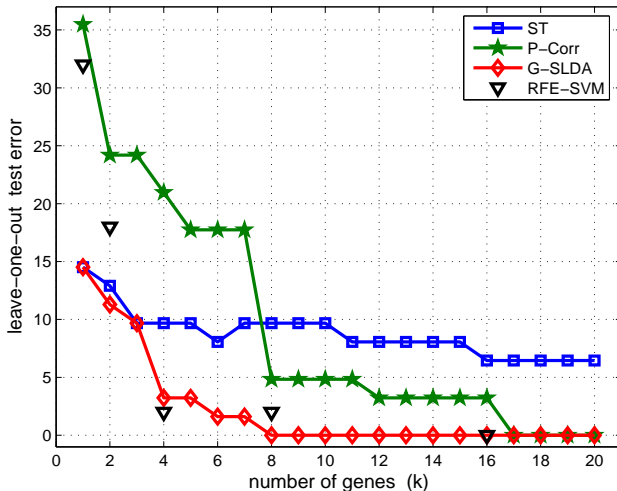


Figure 4. Colon Cancer DNA Micro-array gene expression profiles for cancer diagnosis: *leave-one-out* cross-validation test error vs. cardinality (k). Algorithms are ST (blue □), P-correlation (green ★), GSLDA (red ◇) and RFE-SVM (black ▽). RFE-SVM results are from (Guyon et al., 2002).

The *multi-factor* form of the generalized Rayleigh quotient (with a matrix of factors X) can lead, for example, to a *trace* criterion which as an eigenvalue sum can also be bounded using the generalized inclusion principle. In fact, *any* objective that can be formulated with eigenvalues alone (*e.g.* a log-determinant for entropy-based criteria) can be solved in discrete form using essentially the same algorithms.

The remarkable effectiveness of GSLDA is supported by empirical observations in combinatorial optimization, wherein greedy search with (sub)modular and *monotonic* cost functions very often produces excellent results (Nemhauser & Wolsey, 1988). In our experiments, GSLDA consistently out-performed continuous algorithms like simple thresholding (ST) and variable ranking by correlation. Although our computational burden is greater than such simple techniques, our method compares favorably to more powerful continuous algorithms like SVMs. Nevertheless, processing very high-dimensional datasets, with $n = O(10^4)$, is generally beyond the reach of matrix-based algorithms without specialized numerical computing techniques.

We close by reflecting on the modularity of our discrete algorithms and the ease of transition from the supervised domain (sparse LDA) to the unsupervised domain (sparse PCA) — the default case of $B = I$. Indeed, there is (almost) no modification required in the derivations or implementation. Consequently, our discrete algorithms for sparse LDA automatically *subsume* the unsupervised case of sparse PCA.

In the future, we plan to investigate applications of generalized inclusion bounds to the related problems of clustering and regression. In particular, the ubiquitous “kernel trick” for the *dual* problem of *basis* selection — as in *reduced-set* methods for SVMs, sparse Gaussian processes or sparse kernel machines in general.

APPENDIX

The basic proof of the Inclusion Principle is derived by adding subspace (cardinality) constraints to the variational form of the Courant-Fischer “Min-Max” theorem (Horn & Johnson, 1985). We now extend this to the *generalized* Rayleigh quotient ($x^T A x / x^T B x$).

Given a pair of symmetric matrices $A, B \in \mathcal{S}_+^n$, let $\lambda_j(A, B)$ for $j = 1, \dots, n$ be their generalized eigenvalues ranked in *increasing* order. The main result establishes the following inequalities:

$$\lambda_j(A, B) \leq \lambda_j(A_k, B_k) \leq \lambda_{j+n-k}(A, B) \quad (9)$$

where $\lambda_j(A_k, B_k)$ are the generalized eigenvalues of corresponding $k \times k$ principal submatrices of (A, B) .

By the variational form of the “Min-Max” theorem, the generalized eigenvalues of (A, B) satisfy

$$\lambda_j(A, B) = \min_{\mathcal{S}_n^j} \max_{x \in \mathcal{S}_n^j} \frac{x^T A x}{x^T B x} \quad (10)$$

where \mathcal{S}_n^j denotes an arbitrary j -dimensional subspace of \mathcal{R}^n . The same variational form holds independently for the generalized eigenvalues of (A_k, B_k)

$$\lambda_j(A_k, B_k) = \min_{\mathcal{S}_k^j} \max_{z \in \mathcal{S}_k^j} \frac{z^T A_k z}{z^T B_k z} \quad (11)$$

where \mathcal{S}_k^j is an arbitrary j -dimensional subspace of \mathcal{R}^k . Next we define a “sparse” j -dimensional subspace \mathcal{S}_0^j formed by the direct sum $\mathcal{R}^k \oplus \mathbf{0}$, which by definition consists of all vectors $x \in \mathcal{R}^n$ given by

$$x = \begin{bmatrix} z \\ 0 \end{bmatrix}, \quad \text{where } z \in \mathcal{R}^k \quad (12)$$

We now derive the *l.h.s.* inequality in Eq.(9) — the lower bound for the eigenvalues of (A_k, B_k) — starting from the variational equality in Eq.(10)

$$\begin{aligned} \lambda_j(A, B) &= \min_{\mathcal{S}_n^j} \max_{x \in \mathcal{S}_n^j} \frac{x^T A x}{x^T B x} \\ &\leq \min_{\mathcal{S}_0^j} \max_{x \in \mathcal{S}_0^j} \frac{x^T A x}{x^T B x} \\ &= \min_{\mathcal{S}_0^j} \max_{x \in \mathcal{S}_0^j} \frac{z^T A_k z}{z^T B_k z} \\ &= \lambda_j(A_k, B_k) \end{aligned} \quad (13)$$

where in the 2nd line the subspace $x \in \mathcal{S}_n^j$ is restricted to $x \in \mathcal{S}_n^j \cap \mathcal{S}_0^j$ and since *adding* constraints can not

further *decrease* the minimized expression we obtain the inequality. The 3rd line follows by definition of z as the leading k -dimensional subvector of \mathcal{S}_0^j and the last line follows from Eq.(11). The upper bound on $\lambda_j(A_k, B_k)$ — *r.h.s.* of Eq.(9) — is found by using the same derivation on the *negation* of the Rayleigh quotient. The proof is completed by noting that eigenvalues are invariant to permutation of the indices, hence the derived bounds hold true for *any* principal submatrix of (A, B) not just the leading one.

References

- Alon et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues *Cell Biology*, 96, 6745–6750.
- Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97, 245–271.
- d’Aspremont, A., Ghaoui, L. E., Jordan, M. I., & Laffont, G. R. G. (2004). A direct formulation for sparse PCA using semidefinite programming. *Advances in Neural Information Processing Systems 17* (pp. 803–809). MIT Press.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46, 389–422.
- Horn, R. A., & Johnson, C. R. (1985). *Matrix Analysis*. Cambridge, England: Cambridge Press.
- Ko, C., Lee, J., & Queyranne, M. (1995). An exact algorithm for maximum entropy sampling. *Operations Research*, 43, 684–691.
- Kohavi, R., & John, G. (2003). Wrappers for feature subset selection. *Artificial Intelligence Journal*, 3, 1157–1182.
- Mika, S., Rätsch, G., & Müller, K.-R. (2001). A mathematical programming approach to the kernel fisher algorithm. *Advances in Neural Information Processing Systems 13* (pp. 591–597). MIT Press.
- Moghaddam, B., Weiss, Y., & Avidan, S. (2006). Spectral Bounds for Sparse PCA: Exact & Greedy Algorithms. *Advances in Neural Information Processing Systems 18* (pp. 915–922). MIT Press.
- Nemhauser, G. L., & Wolsey, L. A. (1988). *Integer and Combinatorial Optimization*. New York: John Wiley.
- Somol, P., Pudil, P., & Kittler, J. (2004). Fast branch & bound algorithms for optimal feature selection. *IEEE Pattern Analysis and Machine Intelligence*, 26, 900–912.
- Tibshirani, R. (1995). Regression shrinkage and selection via Lasso. *Journal of the Royal Statistical Society B*, 58, 267–288.
- Zou, H., Hastie, T., & Tibshirani, R. (2004). *Sparse principal component analysis* (Technical Report). Statistics Department, Stanford University.