Understanding and Evaluating Sparse Linear Discriminant Analysis: Supplementary File

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

Here we first review the details of the standard LDA classification rule. Later we present high-level proofs of all lemmas and theorems in our submission. In a few places technical details are omitted for brevity.

Classification Rule of LDA

Sparse variants of LDA only differ in how the projection matrix B is obtained. However, once this B is fixed, classification of a new data point proceeds in the same way based on Fisher's original analysis.

To begin, we first use B to project all the training data into a low-dimensional discriminant space. Let Σ_w^B and μ_k^B denote the within-class pooled covariance matrix and the mean of class k respectively for all the projected data. When given a new observation $x^\star$, we assume that it is draw from a normal distribution with mean given by some μ_k^B and covariance given by Σ_w^B. We then assign $x^\star$ to the class $\delta_B(x^\star)$ with maximum posterior probability in the discriminant space. Assuming that each class has the same prior distribution, $\delta_k(x^\star)$ can be computed using

$$\delta_B(x^\star) = \arg\min_k \left(B^\top x^\star - \mu_k^B\right)^\top \left(\Sigma_w^B\right)^{-1} \left(B^\top x^\star - \mu_k^B\right).$$

Proof of Lemma 1

Under the conditions stipulated by the lemma, we are concerned with finding a single discriminant vector β by solving

$$\begin{array}{ll} \max_\beta & \beta^\top \Sigma_b \beta - \lambda\phi(\beta) \\ \text{s.t.} & \beta^\top \Sigma_w \beta = 1. \end{array} \tag{26}$$

Based on properties of Σ_w, any feasible β can be decomposed as $\beta = \beta_1 + \beta_2$, where $X\beta_1 \in \text{span}[P_Y]$, $X\beta_2 \in \text{span}[I - P_Y]$, and $\beta_2^\top \Sigma_w \beta_2 = 1$. Now consider optimizing (26) over β_1 with β_2 fixed at any feasible value with finite entries. Since $\Sigma_b = \frac{1}{N} X^\top P_Y X$ by construction, the problem reduces to

$$\max_{\beta_1} \quad \beta_1^\top \Sigma_b \beta_1 - \lambda\phi(\beta_1 + \beta_2), \tag{27}$$

which is equivalent to

$$\max_{\beta_1} \quad \|P_Y X \beta_1\|_2^2 - \lambda\phi(\beta_1 + \beta_2). \tag{28}$$

Since $p > N$, we can always find some β_1' with nonzero entries such that $X\beta_1' \in \text{span}[P_Y]$. We may then consider any candidate solution $\alpha\beta_1'$, where $\alpha > 0$ is a scaler, and compute

$$\max_\alpha \quad \alpha^2 \|P_Y X \beta_1'\|_2^2 - \lambda\phi(\alpha\beta_1' + \beta_2). \tag{29}$$

Since ϕ is a concave, non-decreasing function, it can be strictly upper-bounded for all α using a linear, first-order Taylor series approximation, and therefore $\lambda\phi(\alpha\beta_1' + \beta_2) \leq O(\alpha)$. So clearly then (29) is unbounded from above as α becomes large since such a linear term will grow much slower than the $O(\alpha^2)$ term. Consequently, $\beta = \beta_1 + \beta_2$ will be non-sparse as well with unbounded coefficients.

Proof of Theorem 1

If $\phi(\beta)$ is a concave, non-decreasing function of $|\beta| \triangleq [|\beta_1|, |\beta_2|, \ldots]^\top$, then it can be expressed as $\phi(\beta) = \min_{\gamma \geq 0} \beta^\top \Gamma^{-1} \beta + z(\gamma)$ where $\Gamma = \text{diag}(\gamma)$ and $z(\gamma)$ is a concave and non-decreasing function of γ (Wipf et al., 2011). Hence, for any fixed $\gamma > 0$, (12) is equivalent to

$$\begin{array}{ll} \min_{\beta,\theta} & \frac{1}{N}\|Y\theta - X\beta\|_2^2 + \lambda\beta^\top \Gamma^{-1}\beta + z(\gamma) \\ \text{subject to} & \theta^\top Y^\top Y \theta = 1. \end{array} \tag{30}$$

Applying Proof A.6 in (Witten & Tibshirani, 2011), we can show that (30) is equivalent to $\min_\beta -h(\beta^\top \Sigma_b \beta)$ s.t. $\beta^\top \left(\Sigma_w + \lambda\Gamma^{-1}\right)\beta = 1$ by optimizing the objective over θ first. Therefore, (12) is equivalent to

$$\begin{array}{ll} \max_{\beta,\gamma\geq 0} & h(\beta^\top \Sigma_b \beta) - \lambda z(\gamma) \\ \text{subject to} & \beta^\top \left(\Sigma_w + \lambda\Gamma^{-1}\right)\beta = 1 \end{array} \tag{31}$$

for some concave and non-decreasing function $z(\gamma)$. Applying Lagrangian transformation to (31), this is equivalent to solving

$$\max_{\beta,\gamma\geq 0} h\left(\beta^\top \Sigma_b \beta\right) - \lambda z(\gamma) - \mu\lambda\beta^\top \Gamma^{-1}\beta - \mu\beta^\top \Sigma_w \beta \tag{32}$$

for some non-negative constant μ such that the constraint is satisfied at the optimal solution. Note that while the

value of $\beta^\top \left(\Sigma_w + \lambda\Gamma^{-1}\right)\beta$ is monotonically increasing as μ is decreasing, it is possible that in certain cases there will not exist a μ such that at the global optimum of (32) $\beta^\top \left(\Sigma_w + \lambda\Gamma^{-1}\right)\beta$ is equal to one exactly (i.e., this exact point will be skipped over as μ is varied continuously). In this circumstance we would then have to go back and modify slightly our optimal scoring starting point.

Now letting $\alpha_1 \triangleq \sqrt{\mu}$ we have that $-\phi(\alpha_1\beta) = \max_{\gamma\geq 0} -\mu\lambda\beta^\top\Gamma^{-1}\beta - z(\gamma)$. We may then convert $\mu\beta^\top\Sigma_w\beta$ to the constraint form such that $\beta^\top\Sigma_w\beta$ equals some constant α_2 where μ has been absorbed. Finally, the local minimum condition follows as a special case of Theorem 2 below.

Proof of Theorem 2

Similar to before, if $\phi(b)$ is a concave, non-decreasing function of $b = [\|\beta^1\|_2, \ldots, \|\beta^p\|_2]^\top$, then $\phi(b) = \min_{\gamma\geq 0} \text{trace}[B^\top\Gamma^{-1}B] + Lz(\gamma)$, with $z(\gamma)$ also a concave and non-decreasing function as above. Under these conditions, if we optimize over B and Θ, then the optimal scoring problem reduces to

$$\min_{\gamma\geq 0} \sum_{k=1}^{L} \theta_{k*}^\top Y^\top(\lambda I + X\Gamma X^\top)^{-1} Y\theta_{k*} + Lz(\gamma), \quad (33)$$

where Θ_* denotes the optimal value of Θ. Based on (Wipf et al., 2011) it can be shown that problems of this form have at most $N \cdot \text{rank}[Y\Theta_*]$ nonzero elements of γ at any local minimum. Let γ_* denote any minimizing solution (either local or global). Then the optimal B satisfies $B_* = \Gamma_* X^\top(\lambda I + X\Gamma_* X^\top)^{-1} Y\Theta_*$, and hence the number of nonzero rows will be bounded by the number of nonzero elements in γ_*. Additionally, for all λ sufficiently large, it can be shown that any minimizer of must have $\gamma = 0$ by taking derivatives (or subgradients) of (33) and checking first-order optimality conditions.

We prove the remainder of the theorem by showing that for any fixed $\gamma \geq 0$, the global optimum of

$$\begin{array}{ll} \min_{B,\Theta} & \sum_{k=1}^{L} \frac{1}{N}\|Y\theta_k - X\beta_k\|_2^2 + \lambda\beta_k^\top\Gamma^{-1}\beta_k \\ \text{subject to} & \Theta^\top Y^\top Y\Theta = I \end{array} \quad (34)$$

is equal to $\sum_{k=1}^{L} -h(a_k)$ where a_k is the kth eigenvalue of $\tilde{\Sigma}_b = \left(\Sigma_w + \lambda\Gamma^{-1}\right)^{-1/2}\Sigma_b\left(\Sigma_w + \lambda\Gamma^{-1}\right)^{-1/2}$. Note that (34) is a standard optimal scoring problem and therefore equivalent to (4) with $\Omega = \lambda\Gamma^{-1}$ and naturally the sequential version

$$\begin{array}{ll} \max_{\beta_k} & \beta_k^\top\Sigma_b\beta_k \\ \text{subject to} & \beta_k^\top(\Sigma_w + \lambda\Gamma^{-1})\beta_k = 1 \\ & \forall i < k,\ \beta_k^\top(\Sigma_w + \lambda\Gamma^{-1})\beta_i = 0. \end{array} \quad (35)$$

We claim that (35) is equivalent to

$$\begin{array}{ll} \min_{\theta_k,\beta_k} & \frac{1}{N}\|Y\theta_k - X\beta_k\|_2^2 + \lambda\beta_k^\top\Gamma^{-1}\beta_k \\ \text{subject to} & \theta_k^\top Y^\top Y\theta_k = 1 \\ & \forall i < k,\ \theta_k^\top Y^\top Y\theta_i = 0 \end{array} \quad (36)$$

and (36) takes its optimal objective value at $-h(a_k)$ where a_k is the k-th largest eigenvalue of $\tilde{\Sigma}_b$. If this claim holds, due to the equivalence of (36) and (34) which holds because the latter is separable as discussed in Section 4, then plugging the optimal objective value of (36) into (34) leads to the theorem.

We show this claim by induction, noting that the $k = 1$ case is derived in (Witten & Tibshirani, 2011), although not in the context of connecting sparse LDA models. To begin we define $\hat{\theta}_k$ by $\hat{\theta}_k = (Y^\top Y)^{1/2}\theta_k$. Optimizing (36) over $\hat{\theta}_k$, we derive the optimal $\hat{\theta}_k^\star$ by $\hat{\theta}_k^\star = c \cdot P_k^\perp\left((Y^\top Y)^{-1/2}Y^\top X\beta_i\right)$ where c is the normalization constant, and $P_k^\perp$ is an orthogonal projection matrix to the orthogonal space of $(Y^\top Y)^{-1/2}Y^\top X\beta_i$ for all $i < k$. We plug $\hat{\theta}_k^\star$ back to (36) and obtain

$$\min_{\beta_k} -\frac{2}{\sqrt{N}}\sqrt{\beta_k^\top\Sigma_b^k\beta_k} + \beta_k^\top\Sigma_b^k\beta_k + \beta_k^\top\left(\Sigma_w + \lambda\Gamma^{-1}\right)\beta_k, \quad (37)$$

where $\Sigma_b^k = \frac{1}{N}X^\top Y(Y^\top Y)^{-1/2}P_k^\perp(Y^\top Y)^{-1/2}Y^\top X$. Let $\tilde{\beta}_k = \left(\Sigma_w + \lambda\Gamma^{-1}\right)^{-1/2}\beta_k$. Then (37) becomes

$$\min_{\tilde{\beta}_k} -\frac{2}{\sqrt{N}}\sqrt{\tilde{\beta}_k^\top\tilde{\Sigma}_b^k\tilde{\beta}_k} + \tilde{\beta}_k^\top\left(\tilde{\Sigma}_b^k + I\right)\tilde{\beta}_k, \quad (38)$$

where similarly, $\tilde{\Sigma}_b^k$ is defined by $\left(\Sigma_w + \lambda\Gamma^{-1}\right)^{-1/2}\Sigma_b^k\left(\Sigma_w + \lambda\Gamma^{-1}\right)^{-1/2}$. Note that the optimal solution $\tilde{\beta}_k^\star$ for (38) must be such that $\forall i < k$, $\tilde{\beta}_k^{\star T}\tilde{\Sigma}_b\tilde{\beta}_i = 0$, otherwise by taking orthogonal projection, the objective value can be reduced because of the identity matrix term. This observation also implies that $\forall i < k$, $\tilde{\beta}_k^{\star T}\tilde{\beta}_i = 0$ since $\tilde{\beta}_i$ is eigenvector of $\tilde{\Sigma}_b$ for $i < k$. Consequently, (38) is equivalent to

$$\begin{array}{ll} \min_{\tilde{\beta}_k} & -\frac{2}{\sqrt{N}}\sqrt{\tilde{\beta}_k^\top\tilde{\Sigma}_b\tilde{\beta}_k} + \tilde{\beta}_k^\top(\tilde{\Sigma}_b + I)\tilde{\beta}_k \\ \text{subject to} & \forall i < k,\ \tilde{\beta}_k^\top\tilde{\beta}_i = 0. \end{array} \quad (39)$$

Differentiating over $\tilde{\beta}_k$ indicates

$$\tilde{\Sigma}_b\tilde{\beta}_k\left(1 - \frac{1}{\sqrt{N\tilde{\beta}_k^\top\tilde{\Sigma}_b\tilde{\beta}_k}}\right) + \tilde{\beta}_k = 0. \quad (40)$$

From (40) we deduce that $\tilde{\beta}_k$ is an eigenvector of $\tilde{\Sigma}_b$ with eigenvalue

$$\tau = \left(\frac{1}{\sqrt{N\tilde{\beta}_k^\top\tilde{\Sigma}_b\tilde{\beta}_k}} - 1\right)^{-1} = \frac{\sqrt{N\tau\omega}}{1 - \sqrt{N\tau\omega}}$$

where $\omega = \tilde{\beta}_k^\top \tilde{\beta}_k$. Subsequently, (39) evaluated at the optimal $\tilde{\beta}_k$ equals $-\frac{2}{\sqrt{N}}\sqrt{\tau\omega}+\tau\omega+\omega$. Plugging $\omega = \frac{\tau}{N(1+\tau)^2}$ into this expression leads to $-h(\tau)$. Since $\tilde{\beta}_k$ is orthogonal to $\tilde{\beta}_i$ for $i < k$, for minimizing $-h(\tau)$, the optimal solution is the kth eigenvector of $\bar{\Sigma}_b$ and the optimal objective value becomes $-h(a_k)$. Summing over all k then leads to the stated result.

Proof of Lemma 2

Because X is centered, any non-trivial optimal scoring matrix is orthogonal to $\mathbf{1}$ and any two scoring matrices are equivalent via an orthogonal projection. Right multiplying an orthogonal matrix to the scoring matrix Θ will not influence the cost function (15) since it can be compensated for by an equivalent inconsequential transformation of B; this occurs essentially because an ℓ_2-norm-based row-sparse penalty is invariant to rotations.

Proof of Theorem 3

For notational ease we first define $S \triangleq Y\Theta$. We then note that with $\alpha \to 0$, by extending the analysis in (Wipf et al., 2011), the underlying optimization problem and local minima profile is equivalent to minimizing

$$\mathcal{L}(\gamma) = \text{tr}\left[S^\top \Sigma_s^{-1} S\right] + L \log|\Sigma_s|, \tag{41}$$

over $\gamma \in \mathbb{R}_+^p$, where $\Sigma_s \triangleq X\Gamma X^\top$ and $\Gamma \triangleq \text{diag}[\gamma]$, and then computing $\hat{B} = \Gamma X^\top \Sigma_s^{-1} S$. Let $\tilde{B}$ denote the nonzero rows in a maximally row-sparse feasible solution to $S = XB$, and $\tilde{X}$ the corresponding columns, such that $S = \tilde{X}\tilde{B}$.

Now at any minimizing solution of (41), S must be an element of span$[X\Gamma^{1/2}]$ or the cost will be driven to infinity. In this regard for the time being we will assume that Σ_s is invertible. Near any candidate local minimum $\bar{\gamma}$, we may express (41) as

$$\begin{aligned}\mathcal{L}(a,b) &= L\log|a\bar{\Sigma}_s + b\tilde{X}\Lambda^2\tilde{X}^\top| \\ &+\text{tr}\left[S^\top(a\bar{\Sigma}_s + b\tilde{X}\Lambda^2\tilde{X}^\top)^{-1}S\right],\end{aligned} \tag{42}$$

where $\bar{\Sigma}_s = X\bar{\Gamma}X^\top$ and Λ is an arbitrary positive diagonal matrix.

If $\bar{\gamma}$ is a local minimum, it should satisfy

$$\left.\frac{\partial\mathcal{L}(a,b)}{\partial a}\right|_{a=1,b=0} = 0, \quad \left.\frac{\partial\mathcal{L}(a,b)}{\partial b}\right|_{a=1,b=0} \geq 0, \tag{43}$$

otherwise we could alter a (up or down) or increase b from zero to decrease (41). Let $Z = z(a,b) = a\bar{\Sigma}_s + b\tilde{X}\Lambda^2\tilde{X}^\top$. Then we have

$$\begin{aligned}\frac{\partial\mathcal{L}(a,b)}{\partial a} &= L\text{tr}\left[Z^{-1}\bar{\Sigma}_s\right] - \text{tr}\left[S^\top Z^{-1}\bar{\Sigma}_s Z^{-1}S\right] \\ \frac{\partial\mathcal{L}(a,b)}{\partial b} &= L\text{tr}\left[Z^{-1}\tilde{X}\Lambda^2\tilde{X}^\top\right] \\ &\quad -\text{tr}\left[S^\top Z^{-1}\tilde{X}\Lambda^2\tilde{X}^\top Z^{-1}S\right].\end{aligned} \tag{44}$$

Since $z(1,0) = \bar{\Sigma}_s$,

$$\begin{aligned}\left.\frac{\partial\mathcal{L}(a,b)}{\partial a}\right|_{a=1,b=0} &= L\text{tr}\left[I_N\right] - \text{tr}\left[S^\top\bar{\Sigma}_s^{-1}S\right], \\ \left.\frac{\partial\mathcal{L}(a,b)}{\partial b}\right|_{a=1,b=0} &= L\text{tr}\left[\bar{\Sigma}_s^{-1}\tilde{X}\Lambda^2\tilde{X}^\top\right] \\ &\quad -\text{tr}\left[S^\top\bar{\Sigma}_s^{-1}\tilde{X}\Lambda^2\tilde{X}^\top\bar{\Sigma}_s^{-1}S\right].\end{aligned} \tag{45}$$

Equating the first equation to zero gives tr$\left[S^\top\bar{\Sigma}_s^{-1}S\right] = LN$. For the second equation, let the singular value decomposition of $\tilde{X}$ be $\tilde{X} = U\Delta V^\top$. Then the righthand side of (45) becomes

$$\begin{aligned}&L\text{tr}\left[\Delta V^\top\Lambda^2 V\Delta U^\top\bar{\Sigma}_s^{-1}U\right] \\ &\quad -\text{tr}\left[\Delta V^\top\Lambda^2 V\Delta U^\top\bar{\Sigma}_s^{-1}SS^\top\bar{\Sigma}_s^{-1}U\right] \\ \leq\ & L\lambda_{max}(\Lambda\tilde{X}^\top\tilde{X}\Lambda)\text{tr}\left[U^\top\bar{\Sigma}_s^{-1}U\right] \\ &\quad -\lambda_{min}(\Lambda\tilde{X}^\top\tilde{X}\Lambda)\text{tr}\left[U^\top\bar{\Sigma}_s^{-1}SS^\top\bar{\Sigma}_s^{-1}U\right],\end{aligned} \tag{46}$$

where the inequality comes from the fact that $\lambda_{min}(X)\text{tr}(Y) \leq \text{tr}(XY) \leq \lambda_{max}(X)\text{tr}(Y)$ for any symmetric, positive semi-definite matrices X and Y. (Here $\lambda_{min}(X)$ and $\lambda_{max}(X)$ correspond with the smallest and largest eigenvalue of X respectively.)

Because S is orthogonal by virtue of the assumed LDA optimal scoring constraint, rank$[S] = L$. Moreover, because of the spark assumption, rank$[\tilde{X}] = D \leq L$. However, this restricts any feasible $\tilde{B}$ to satisfy rank$[\tilde{B}] \geq$ rank$[S]$, and therefore $\tilde{B}$ must be invertible. It then follows that span(U) = span$(\tilde{X})$ = span$(\tilde{X}\tilde{B})$ = span(S). To further simplify the righthand side of (46), we use the fact that for any matrices with orthonormal columns X and Y, if span(X) = span(Y), then tr$(X^\top WX)$ = tr$(Y^\top WY)$ and tr$(X^\top WXX^\top WX)$ = tr$(X^\top WYY^\top WX)$ for an arbitrary symmetric matrix W.

Now given $\lambda_1, \cdots, \lambda_D$ are the eigenvalues of $S^\top\bar{\Sigma}_s^{-1}S$, then tr$\left[S^\top\bar{\Sigma}_s^{-1}S\right] = \sum_{i=1}^D \lambda_i = LN$. And with $A \triangleq \Lambda\tilde{X}\tilde{X}^\top\Lambda$, the righthand side of (46) reduces to

$$L\lambda_{max}(A)\sum_{i=1}^D \lambda_i - \lambda_{min}(A)\sum_{i=1}^D \lambda_i^2 \tag{47}$$

$$\leq L^2 N\|A\|_2 - \frac{(LN)^2}{D\|A^{-1}\|_2}. \tag{48}$$

Consequently, in total we may conclude that, for a local

minima to occur, it must be that

$$L^2 N\|A\|_2 - \frac{(LN)^2}{D\|A^{-1}\|_2} \geq \left.\frac{\partial \mathcal{L}(a,b)}{\partial b}\right|_{a=1,b=0} \geq 0. \quad (49)$$

However, if $\|A\|_2\|A^{-1}\|_2 < \frac{N}{D}$, then $L^2 N\|A\|_2 - \frac{(LN)^2}{D\|A^{-1}\|_2} < 0$, which means that $\bar{\gamma}$ cannot be a local minimum. Since Λ can be an arbitrary positive diagonal matrix, we choose $\Lambda = \arg\min_{\Lambda}\|A\|_2\|A^{-1}\|_2$ to form the strongest bound. This rules out as a local minima any $\bar{\gamma}$ such that the corresponding $\bar{\Sigma}_s$ is full rank. Similar arguments apply to a general stationary point that may not be a local minima.

We now only need consider the rare γ values such that both S is an element of span$[X\Gamma^{1/2}]$ and Σ_s is *not* full rank. Technically, if Σ_s is not full rank, the cost function (41) is not defined. It is here that careful consideration of the limit of $\alpha \to 0$, where the limit is take outside of the minimization, ameliorates the problem. With this in mind, it is then straightforward to demonstrate that only some γ^* with sparsity profile matching the maximally row-sparse feasible solution is eligible to be a stationary point. However, for brevity in a short conference paper we defer rigorous treatment of these details, which are ultimately straightforward to handle, to a subsequent journal publication. All of this then guarantees that the associated $\hat{B}$ estimator will be maximally row sparse.

Finally, strategic counter-examples can be used to show that no problem of the form

$$\min_B \sum_{i=1}^{p} f(\|\beta^i\|_2) \quad \text{s.t. } S = XB \quad (50)$$

can satisfy the same result. For all functions f, there can exist cases which fulfill the stipulations of the theorem and yet have one or more local minima that are not maximally row sparse. Again, we defer details to a subsequent journal publication.

References

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.