Reduced rank vector generalized linear models for feature extraction

Yiyuan She[*]

Supervised linear feature extraction can be achieved by fitting a reduced rank multivariate model. This paper studies rank penalized and rank constrained vector generalized linear models. From the perspective of thresholding rules, we build a framework for fitting singular value penalized models and use it for feature extraction. Through solving the rank constraint form of the problem, we propose progressive feature space reduction for fast computation in high dimensions with little performance loss. A novel projective cross-validation is proposed for parameter tuning in such nonconvex setups. Real data applications are given to show the power of the methodology in supervised dimension reduction and feature extraction.

1. BACKGROUND

Recently, high dimensional data analysis attracts a great deal of interest from statisticians. The availability of a large pool of variables (relative to the sample size) poses challenges in statistical modeling because in this high-dimensional setup, both estimation accuracy and model interpretability can be seriously hurt. Dimension reduction is a natural and effective means to reduce the number of unknowns. One can remove nuisance and/or redundant variables, referred to as variable/feature selection; alternatively, one can find low dimensional linear or nonlinear projections of the input data, referred to as feature extraction. In this paper, we focus on linear feature extraction for dimension reduction purposes.

The most popular approach for linear feature extraction is perhaps the principle component analysis (PCA). Given $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with n observations and p features, perform the Singular Value Decomposition (SVD) on the data $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$. Given any $1 \leq r \leq rank(\boldsymbol{X})$, denote by $\boldsymbol{V}_r$ the submatrix of $\boldsymbol{V}$ consisting of its first r columns. Then

$$\boldsymbol{Z}_r = \boldsymbol{X}\boldsymbol{V}_r$$

constructs r new features as linear combinations of the original features. The extraction is optimal in the sense that $\hat{\boldsymbol{B}} = \boldsymbol{X}\mathcal{P}_{\boldsymbol{V}_r} = \boldsymbol{X}\boldsymbol{V}_r\boldsymbol{V}_r^T$, with $\mathcal{P}_{\boldsymbol{V}_r}$ being the projection matrix onto the column space of $\boldsymbol{V}_r$, gives the best rank-r approximation to $\boldsymbol{X}$:

$$\hat{\boldsymbol{B}} = \arg \min_{\boldsymbol{B}:rank(\boldsymbol{B}) \leq r} \|\boldsymbol{X} - \boldsymbol{B}\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm. A by-product is that the gram matrix of $\boldsymbol{X}\boldsymbol{V}_r$ is diagonal, which means all new features are uncorrelated with each other.

On the other hand, PCA is unsupervised. In many statistical learning problems, we want to construct new features that best predict the responses. Suppose $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$ is the response matrix, n being the sample size and m being the number of response variables. Supervised feature extraction can be given by the *reduced rank regression* (RRR) (Anderson, 1951), with the RRR estimator $\hat{\boldsymbol{B}}$ defined by

$$\hat{\boldsymbol{B}} = \arg \min_{\boldsymbol{B}:rank(\boldsymbol{B}) \leq r} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2.$$

Assume $\boldsymbol{X}$ has full column rank and define $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Then $\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_{ols}\boldsymbol{V}_r\boldsymbol{V}_r^T$, where $\hat{\boldsymbol{B}}_{ols} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. $\boldsymbol{V}_r$ is formed by the first r columns of $\boldsymbol{V}$ from the spectral decomposition $\boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$. See Reinsel and Velu (1998) for more details. Therefore,

$$\boldsymbol{Z}_r = \boldsymbol{X}(\hat{\boldsymbol{B}}_{ols}\boldsymbol{V}_r)$$

constructs r new features that best approximate $\boldsymbol{Y}$ in Frobenius norm, and these new features are, again, uncorrelated. The RRR framework includes the PCA as a special case, by setting $\boldsymbol{Y} = \boldsymbol{X}$ (Izenman, 2008). The above RRR solves a *nonconvex* optimization problem in the classical setup $n > p$. Recently, Bunea, She and Wegkamp (2011) studied the problem under $p > n$ and developed finite-sample theories as well as a computational algorithm.

On the other hand, the squared error loss may not always be appropriate. For vector generalized linear models (GLMs), such as discrete responses arising in classification problems, deviance loss is much more reasonable.

Although there is a large body of literature on the RRR—Robinson (1974), Rao (1979), and Brillinger (1981), to name a few, to the best of our knowledge, there is very little work beyond the Gaussian model. Yee and Hastie (2003)

studied the reduced-rank vector GLM problem and used an iterative approximate estimation by fitting RRR repeatedly. Yet this only provides an approximation solution to the original problem and there is no guarantee of converge. Heinen and Rengifo (2008) resorted to a continuation technique to deal with discrete responses.

This paper aims to tackle the penalized and constrained vector GLMs

$$(1.1) \quad \min_{\boldsymbol{B}} - \text{log-likelihood}(\boldsymbol{B}; \boldsymbol{Y}, \boldsymbol{X}) + \frac{\lambda^2}{2} rank(\boldsymbol{B}), \quad \text{and}$$

$$(1.2) \quad \min_{\boldsymbol{B}} - \text{log-likelihood}(\boldsymbol{B}; \boldsymbol{Y}, \boldsymbol{X}) \quad \text{s.t. } rank(\boldsymbol{B}) \leq r.$$

The imposed reduced rank structure is based on the belief that the features' relevant directions, in response to $\boldsymbol{Y}$, define a lower dimensional subspace in $\mathbb{R}^p$. The rank of such an estimator determines the number of new features to construct. These two problems are *not* equivalent to each other due to their nonconvexity. In fact, the rank function is nonconvex and discrete (and thus nondifferentiable), thereby posing a challenge in optimization. Our algorithms boil down to an iterative version, which is not surprising in the GLM setup.

The rest of the paper is organized as follows. Section 2 starts by studying a matrix approximation problem, and then builds a framework for fitting singular-value penalized multivariate GLMs. Supervised feature extraction can be attained for non-Gaussian models not necessarily using the squared error loss. The framework covers a wide family of penalty functions. A new parameter tuning strategy is proposed. Section 3 tackles the rank constrained GLM problem and comes up with a feature space reduction technique. Through this, (1.1) and (1.2) can be combined to achieve better estimation accuracy and computational efficiency. Section 4 illustrates real applications of the proposed methodology. We conclude in Section 5. All technical details are left to the Appendix.

2. PENALIZED VECTOR GLMS FOR FEATURE EXTRACTION

In this section, we study the *penalized* form reduced rank GLMs (1.1). Our algorithm and analysis apply to $p > n$ situations and cover a large family of singular-value penalties, including nuclear norm, Frobenius norm, Schatten p-penalties ($0 < \text{p} < 1$), and rank penalty. To achieve such generality, we start by studying a simpler matrix approximation problem.

2.1 Singular-value penalized matrix approximation

We consider the problem of matrix approximation with a singular value penalty

$$(2.1) \quad \min_{\boldsymbol{B}} \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{B}\|_F^2 + \sum_i P(\sigma_i^{(\boldsymbol{B})}; \lambda),$$

where $\sigma_i^{(\boldsymbol{B})}$ denote the singular values of $\boldsymbol{B}$. The choice of the penalty function P is flexible. For example, $P(t; \lambda) = \lambda|t|$ gives a multiple of the sum of singular values corresponding to the trace norm or nuclear norm penalty. When $P(t; \lambda) = \lambda^2 1_{t \neq 0}/2$, we get the rank penalty which is discrete and nonconvex. For a general P, the closed-form solution to (2.1) is not known (to the best of our knowledge). We address the problem from the standpoint of threshold functions.

Definition 2.1 (Threshold function). A threshold function is a real valued function $\Theta(t; \lambda)$ defined for $-\infty < t < \infty$ and $0 \leq \lambda < \infty$ such that

1. $\Theta(-t; \lambda) = -\Theta(t; \lambda)$,
2. $\Theta(t; \lambda) \leq \Theta(t'; \lambda)$ for $t \leq t'$,
3. $\lim_{t \to \infty} \Theta(t; \lambda) = \infty$, and
4. $0 \leq \Theta(t; \lambda) \leq t$ for $0 \leq t < \infty$.

Remarks. (i) A vector version of Θ (still denoted by Θ) is defined componentwise if either t or λ is replaced by a vector. (ii) There may be some ambiguity in defining a threshold function. For example, the hard-thresholding can be defined as $\Theta_H(t; \lambda) = t1_{|t|>\lambda}$ or $\Theta_H(t; \lambda) = t1_{|t|\geq\lambda}$. Fortunately, commonly used thresholding rules have at most finitely many discontinuity points and such discontinuities rarely occur in real data. When applying Θ to a quantity (say t), we always make the implicit assumption that Θ is continuous at t. (iii) By definition, $\Theta^{-1}(u; \lambda) \triangleq \sup\{t : \Theta(t; \lambda) \leq u\}, \forall u > 0$ must be monotonically increasing on $(0, \infty)$ and bounded between the identity line and $u = 0$; its derivative is defined almost everywhere on $(0, \infty)$. We assume that

$$d\Theta^{-1}(u; \lambda)/du \geq 1 - \mathcal{L}_\Theta \quad \text{a.e. on } (0, \infty)$$

for some constant $\mathcal{L}_\Theta \in [0, 1]$ independent of λ. (In fact, for all *convex* penalties constructed through (2.5), $\mathcal{L}_\Theta$ can be set to be 0.)

Next we introduce the matrix thresholding.

Definition 2.2 (Matrix threshold function). Given any threshold function $\Theta(\cdot; \lambda)$, its matrix version Θ^σ is defined as follows

$$(2.2) \quad \Theta^\sigma(\boldsymbol{B}; \lambda) \triangleq \boldsymbol{U}\text{diag}\{\Theta(\sigma_i^{(\boldsymbol{B})}; \lambda)\}\boldsymbol{V}^T, \quad \forall \boldsymbol{B} \in \mathbb{R}^{n \times m},$$

where $\boldsymbol{U}$, $\boldsymbol{V}$, and $\text{diag}\{\sigma_i^{(\boldsymbol{B})}\}$ are obtained from the SVD of $\boldsymbol{B}$: $\boldsymbol{B} = \boldsymbol{U}^T \text{diag}\{\sigma_i^{(\boldsymbol{B})}\}\boldsymbol{V}$.

Note that $\Theta(0; \lambda) = 0$ by definition, and $\Theta^\sigma(\boldsymbol{B}; \lambda)$ is not affected by the ambiguity of the SVD form.

Proposition 2.1. *Given an arbitrary thresholding rule Θ, let P be any function satisfying*

$$(2.3) \quad P(\theta; \lambda) - P(0; \lambda)$$

$$= \int_0^{|\theta|} (\sup\{s : \Theta(s; \lambda_k) \leq u\} - u) \, du + q(\theta; \lambda),$$

Table 1. Some basic singular-value penalties and their coupled thresholding functions

	Nuclear norm	Frobenius	Rank	Schatten-p, $p \in (0,1)$
Penalty function	$\lambda\|\boldsymbol{B}\|_* = \lambda \sum \sigma_i^{(\boldsymbol{B})}$	$\lambda\|\boldsymbol{B}\|_F^2$	$\frac{\lambda^2}{2} rank(B)$	$\lambda \sum (\sigma_i^{(\boldsymbol{B})})^p$
Thresholding rule	$(t - \mathrm{sgn}(t)\lambda)1_{\|t\|>\lambda}$ (soft)	$\frac{t}{1+\lambda}$ (ridge)	$t 1_{\|t\|>\lambda}$ (hard)	Ex 2.7 in She (2012)

where $q(\cdot; \lambda)$ is nonnegative and $q(\Theta(t; \lambda); \lambda) = 0$ for all $t \in \mathbb{R}$. Then, the singular-value penalized minimization

$$(2.4) \quad \min_{\boldsymbol{B}} F(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{B}\|_F^2/2 + \sum P(\sigma_i^{(\boldsymbol{B})}; \lambda)$$

has a unique optimal solution $\hat{\boldsymbol{B}} = \Theta^\sigma(\boldsymbol{Y}; \lambda)$ for every $\boldsymbol{Y}$, provided $\Theta(\cdot; \lambda)$ is continuous at any singular value of $\boldsymbol{Y}$.

See Appendix A for its proof.

The function q is often just zero, but can be nonzero in certain cases. In fact, we can use a nontrivial q to attain the exact rank penalty; see (2.7). The proposition implies that multiple (infinitely many, as a matter of fact) penalties can result in the same solution, which justifies our thresholding launching point (rather than a penalty one). Some examples of the penalty P and the coupled Θ are listed in Table 1.

We point out two special cases of Proposition 2.1 as follows.

A fusion between nuclear norm and Frobenius norm Define a continuous thresholding rule

$$(2.5) \quad \Theta_B(t; \lambda, M) = \begin{cases} 0, & \text{if } |t| \leq \lambda, \\ t - \lambda\,\mathrm{sgn}(t), & \text{if } \lambda < |t| < \lambda + M, \\ \frac{t}{1 + \frac{\lambda}{M}}, & \text{if } |t| > \lambda + M. \end{cases}$$

When $M \to \infty$, Θ_B becomes the soft-thresholding. When $M = 0$, Θ_B reduces to the ridge thresholding. The penalty constructed from (2.5) is given by

$$P(\theta; \lambda, M) = \begin{cases} \lambda|\theta|, & \text{if } |\theta| \leq M, \\ \lambda\frac{\theta^2 + M^2}{2M}, & \text{if } |\theta| > M, \end{cases}$$

which is exactly the "*Berhu*" penalty (Owen, 2007) whose composition reverses that of *Huber*'s robust loss function. The Berhu penalty on the singular values provides a *convex* fusion of the nuclear norm penalty and the Frobenius norm (squared) penalty in the problem of (2.4). Unlike the elastic net (Zou and Hastie, 2005), this fusion is nonlinear and fully preserves the nondifferentiable behavior (around zero) of the nuclear norm.

A fusion between rank and Frobenius norm A direct thresholding rule that fuses the hard-thresholding and the ridge-thresholding is the hard-ridge thresholding (She, 2009)

$$(2.6) \quad \Theta_{HR}(t; \lambda, \eta) = \begin{cases} 0, & \text{if } |t| < \lambda, \\ \frac{t}{1+\eta}, & \text{if } |t| \geq \lambda. \end{cases}$$

Setting $q \equiv 0$ in Proposition 2.1, we obtain one associated penalty

$$P(\theta; \lambda, \eta) = \begin{cases} -\frac{1}{2}\theta^2 + \lambda|\theta|, & \text{if } |\theta| < \frac{\lambda}{1+\eta}, \\ \frac{1}{2}\eta\theta^2 + \frac{1}{2}\frac{\lambda^2}{1+\eta}, & \text{if } |\theta| \geq \frac{\lambda}{1+\eta}. \end{cases}$$

Interestingly, noticing that Θ_{HR} is discontinuous at λ, we can choose

$$(2.7) \quad q(\theta; \lambda, \eta) = \begin{cases} \frac{(1+\eta)(\lambda-|\theta|)^2}{2}, & \text{if } 0 < |\theta| < \lambda, \\ 0, & \text{if } \theta = 0 \text{ or } |\theta| > \lambda, \end{cases}$$

which leads to $P(\theta) = \frac{1}{2}\eta\theta^2 + \frac{1}{2}\frac{\lambda^2}{1+\eta}1_{\theta\neq0}$. Therefore, $\Theta_{HR}(\cdot; \lambda, \eta)$ can deal with the following rank-Frobenius penalty in (2.4)

$$(2.8) \quad \frac{1}{2}\eta\|\boldsymbol{B}\|_F^2 + \frac{1}{2}\frac{\lambda^2}{1+\eta}\mathrm{rank}(\boldsymbol{B}).$$

This penalty may be of interest in statistical learning tasks that have joint concerns of accuracy and parsimony: the rank portion enforces high rank deficiency, while the ridge (Frobenius) portion shrinks $\boldsymbol{B}$ to compensate for large noise and decorrelates the input variables in large-p applications.

At the end of this subsection, we present a perturbation result which will be used to establish the main result in the next subsection.

Proposition 2.2. *Given* $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$, *let* $Q(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{B}\|_F^2/2 + \sum P_\Theta(\sigma_i^{(\boldsymbol{B})}; \lambda)$, *where* P_Θ *is the penalty obtained from* (2.5). *Denote by* $\hat{\boldsymbol{B}}$ *the minimizer of* $Q(\boldsymbol{B})$. *Then for any matrix* $\boldsymbol{\Delta} \in \mathbb{R}^{n \times m}$

$$Q(\hat{\boldsymbol{B}} + \boldsymbol{\Delta}) - Q(\hat{\boldsymbol{B}}) \geq \frac{C_1}{2}\|\boldsymbol{\Delta}\|_F^2,$$

where $C_1 = 1 - \mathcal{L}_\Theta \geq 0$.

See Appendix B for its proof.

2.2 Singular-value penalized vector GLMs

In this subsection, we generalize the results obtained for matrix approximation to vector GLMs.

Let $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m] \in \mathbb{R}^{n \times m}$ be the response matrix with m response variables and n samples for each. Assume $y_{i,k}$ are independent and each follows a distribution in the natural exponential family $f(y_{i,k}; \theta_{i,k}) = \exp(y_{i,k}\theta_{i,k} - b(\theta_{i,k}) + c(y_{i,k}))$, where $\theta_{i,k}$ is the natural parameter. Let $L_{i,k} = \log f(y_{i,k}, \theta_{i,k})$, $L = \sum_k \sum_i L_{i,k}$. The canonical link

function $g = (b')^{-1}$ is applied throughout the paper. Let the model matrix and the corresponding coefficient matrix be

(2.9)
$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T = [\tilde{\boldsymbol{x}}_0, \tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_p] \in \mathbb{R}^{n \times (p+1)} \quad \text{and}$$
$$\boldsymbol{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m] = [\tilde{\boldsymbol{\beta}}_0, \tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_p]^T \in \mathbb{R}^{(p+1) \times m},$$

respectively. If $\tilde{\boldsymbol{x}}_0 = \boldsymbol{1}$, $\tilde{\boldsymbol{\beta}}_0$ represents the intercept vector. For convenience, we use $\boldsymbol{B}^\circ = [\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_p]^T$ to denote the submatrix of $\boldsymbol{B}$ obtained by deleting the first row $\tilde{\boldsymbol{\beta}}_0^T$, and use $\boldsymbol{X}^\circ$ to denote the submatrix of $\boldsymbol{X}$ obtained by deleting the first column $\tilde{\boldsymbol{x}}_0$. Given any GLM with coefficients $\boldsymbol{\beta}$, we introduce

$$\mu(\boldsymbol{\beta}) \triangleq [g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})]_{n \times 1} \quad \text{and}$$
$$\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) \triangleq \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} = \boldsymbol{X}^T \text{diag} \left\{ b''(\boldsymbol{x}_i^T \boldsymbol{\beta}) \right\}_{i=1}^n \boldsymbol{X}$$

to denote the mean vector and the information matrix at $\boldsymbol{\beta}$. For the m-response vector GLM, the mean matrix $\boldsymbol{\mu}(\boldsymbol{B}) = [\mu_{i,k}]_{n \times m}$ is defined as $[\boldsymbol{\mu}(\boldsymbol{\beta}_1), \ldots, \boldsymbol{\mu}(\boldsymbol{\beta}_m)]$.

Remarks. (i) Having $\tilde{\boldsymbol{x}}_0$ and $\tilde{\boldsymbol{\beta}}_0$ is necessary. For non-Gaussian GLMs, one cannot center the response variables because this may violate the distributional assumption. (ii) For clarity, the above setup does not include any dispersion parameter. But all discussions in this subsection can be trivially extended to the exponential dispersion family $f(y_{i,k}; \theta_{i,k}, \phi) = \exp\{(y_{i,k}\theta_{i,k} - b(\theta_{i,k}))/a(\phi) + c(y_{i,k}, \phi)\}$ which covers the vector Gaussian regression.

Our goal is to minimize (1.1) or more generally, the following objective function

(2.10)
$$F(\boldsymbol{B}) \triangleq -\sum_{k=1}^m \sum_{i=1}^n L_{i,k}(\boldsymbol{\beta}_k; \boldsymbol{x}_i, y_{ik}) + \sum_{s=1}^{p \wedge m} P(\sigma_s^{(\boldsymbol{B}^\circ)}; \lambda)$$

for a large family of penalty functions (possibly nonconvex). The penalty is *not* imposed on $\tilde{\boldsymbol{\beta}}_0$.

We construct the following sequence of iterates for solving the problem: given $\boldsymbol{B}^{(j)}$, perform the update

(2.11)
$$\begin{cases} \boldsymbol{B}^{\circ(j+1)} &= \Theta^\sigma(\boldsymbol{B}^{\circ(j)} + \boldsymbol{X}^{\circ T}\boldsymbol{Y} - \boldsymbol{X}^{\circ T}\boldsymbol{\mu}(\boldsymbol{B}^{(j)}); \lambda), \\ \tilde{\boldsymbol{\beta}}_0^{(j+1)} &= \tilde{\boldsymbol{\beta}}_0^{(j)} + (\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{B}^{(j)}))^T \tilde{\boldsymbol{x}}_0. \end{cases}$$

Theorem 2.1. *Given an arbitrary thresholding rule Θ, let $P(\cdot)$ be any function satisfying (2.5). Starting with any $\boldsymbol{B}^{(0)} \in \mathbb{R}^{(p+1) \times m}$, run (2.11) to obtain a sequence $\{\boldsymbol{B}^{(j)}\}$. Denote by A_k the set of $\{t\boldsymbol{\beta}_k^{(j)} + (1-t)\boldsymbol{\beta}_k^{(j+1)} : t \in (0,1), j = 1, 2, \ldots\}$, $1 \le k \le K$, and define*

$$\rho = \max_{1 \le k \le m} \sup_{\boldsymbol{\xi}_k \in A_k} \|\boldsymbol{\mathcal{I}}(\boldsymbol{\xi}_k)\|_2.$$

Suppose $\rho < 2 - \mathcal{L}_\Theta$. Then $F(\boldsymbol{B}^{(j)})$ is decreasing and satisfies

(2.12)
$$F(\boldsymbol{B}^{(j)}) - F(\boldsymbol{B}^{(j+1)}) \ge C\|\boldsymbol{B}^{(j)} - \boldsymbol{B}^{(j+1)}\|_F^2/2,$$
$$j = 1, 2, \ldots$$

where $C = 2 - \mathcal{L}_\Theta - \rho$. Any limit point $\boldsymbol{B}^*$ of the sequence $\boldsymbol{B}^{(j)}$, referred to as a Θ^σ-estimator, is a solution to the following equation:

(2.13)
$$\begin{cases} \boldsymbol{B}^\circ = \Theta^\sigma(\boldsymbol{B}^\circ + \boldsymbol{X}^{\circ T}\boldsymbol{Y} - \boldsymbol{X}^{\circ T}\boldsymbol{\mu}(\boldsymbol{B}); \lambda), \\ \boldsymbol{0} = (\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{B}))^T \tilde{\boldsymbol{x}}_0, \end{cases}$$

under the assumption that Θ is continuous at all singular values of $\boldsymbol{B}^{\circ *} + \boldsymbol{X}^{\circ T}\boldsymbol{Y} - \boldsymbol{X}^{\circ T}\boldsymbol{\mu}(\boldsymbol{B}^*)$.

The proof details are given in Appendix C.

Recall that $\mathcal{L}_\Theta \le 1$. In implementation, we can scale the model matrix by $\boldsymbol{X}/k_0$ for any $k_0 \ge \sqrt{\rho}$ regardless of Θ, and then perform (2.11). The Θ^σ-estimate, obtained on the scaled data, can be scaled back to give an estimate on the original $\boldsymbol{X}$. The procedure has a theoretical guarantee of convergence and (2.12) yields a good stopping criterion based on the change in $\boldsymbol{B}^{(j)}$. Empirically, we always observe $\boldsymbol{B}^{(j)}$ has a unique limit point. Similar to She (2012), when it is possible to explicitly calculate the curvature parameter $\mathcal{L}_\Theta$, say for SCAD or soft-thresholding, we recommend using the smallest possible value of $k_0 = \sqrt{\rho/(2 - \mathcal{L}_\Theta)}$, which significantly speeds the convergence of the algorithm based on extensive experience. (For example, with a convex penalty we can set $k_0 = \sqrt{\rho/2}$.) We give two typical situations to find an upper bound for ρ in theory.

Example 2.1 (**Penalized Vector Gaussian GLM**). For Gaussian regression, we can ignore the intercept term (after centering both responses and predictors beforehand), and the objective function (2.10) becomes

(2.14)
$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2/2 + \sum_{s=1}^{p \wedge m} P(\sigma_s^{(\boldsymbol{B})}; \lambda).$$

(2.11) reduces to

(2.15)
$$\boldsymbol{B}^{(j+1)} = \Theta^\sigma(\boldsymbol{B}^{(j)} + \boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}; \lambda).$$

Here, $\boldsymbol{\mathcal{I}} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{X}$. According to the theorem, k_0 can be chosen to be $\|\boldsymbol{X}\|_2$ regardless of the thresholding rule and the penalty, where $\|\cdot\|_2$ denotes the spectral norm.

In the special case of imposing a direct rank penalty, where $\sum_{s=1}^{p \wedge m} P(\sigma_s^{(\boldsymbol{B})}; \lambda) = \frac{\lambda^2}{2} rank(B)$, another computational procedure based on the classical RRR algorithm can be used. In fact, RRR studies a relevant but different problem, with no penalty but subject to a low rank constraint. But we can adapt the procedure to minimizing $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2/2 + \lambda^2/2 \cdot \text{rank}(\boldsymbol{B})$ as follows (cf.

Bunea, She and Wegkamp (2011)). Suppose $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular and $\boldsymbol{H}$ is the hat matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. (i) Apply spectral decomposition to $\boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y}$: $\boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$ where $\boldsymbol{D} = \mathrm{diag}\{d_1,\ldots,d_m\}$ with $d_1 \geq d_2 \geq \cdots \geq d_m \geq 0$. (ii) Given any value of λ, define $r = \max\{i : d_i \geq \lambda\}$ and $\boldsymbol{V}_r = \boldsymbol{V}[\,,1{:}r]$, by taking the first r columns in $\boldsymbol{V}$. (iii) Then the (globally) optimal solution is given by

$$\hat{\boldsymbol{B}}(\lambda) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\mathcal{P}_{\boldsymbol{V}_r} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{V}_r\boldsymbol{V}_r^T,$$

where $\mathcal{P}_{\boldsymbol{V}_r}$ is the orthogonal projection onto the column space of $\boldsymbol{V}_r$. We can show the Θ^σ-estimate defined by (2.15) reduces to the RRR estimate in this case, the proof details given in Appendix D.

Proposition 2.3. *Suppose $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ ($n \geq p$) has full column rank and $\|\boldsymbol{X}\|_2 \leq 1$. Then the RRR estimate $\hat{\boldsymbol{B}}(\lambda)$ constructed above must satisfy the Θ^σ-equation $\hat{\boldsymbol{B}} = \Theta_H^\sigma(\hat{\boldsymbol{B}} + \boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{B}}; \lambda)$ for matrix hard-thresholding Θ_H^σ.*

Unlike RRR, our algorithm and convergent analysis do not require $\boldsymbol{X}$ to have full rank or $n > p$. In comparison with the large-p RSC (Bunea, She and Wegkamp, 2011), (2.11) applies to any Θ (and covers all vector GLMs).

Example 2.2 (**Penalized Vector Logistic GLM**). Assume a classification setup where y_{ik} are all binary. The singular-value penalized vector logistic regression minimizes

$$(2.16) \qquad -\sum_{k=1}^m \sum_{i=1}^n \left(y_{i,k}\boldsymbol{x}_i^T\boldsymbol{\beta}_k - \log(1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_k))\right)$$
$$+ \sum_{s=1}^{p\wedge m} P(\sigma_s^{(\boldsymbol{B}^\circ)}; \lambda).$$

The first iteration step in (2.11) becomes

$$(2.17)$$
$$\boldsymbol{B}^{\circ(j+1)} = \Theta^\sigma\left(\boldsymbol{B}^{\circ(j)} + \boldsymbol{X}^{\circ T}\boldsymbol{Y}\right.$$
$$\left. - \boldsymbol{X}^{\circ T}\left[1/(1 + \exp(-\boldsymbol{x}_i^T\boldsymbol{\beta}_k^{(j)}))\right]_{n\times m}; \lambda\right).$$

In R, the matrix $\boldsymbol{\mu}(\boldsymbol{B})$ can be simply constructed by `1/(1+exp(-X%*%B))`. Since $\boldsymbol{W}(\boldsymbol{\beta}) = \mathrm{diag}\{b''(\boldsymbol{x}_i^T\boldsymbol{\beta})\} = \mathrm{diag}\{\pi_i(1 - \pi_i)\} \preceq \boldsymbol{I}/4$ with $\pi_i = 1/(1 + \exp(-\boldsymbol{x}_i^T\boldsymbol{\beta}))$, a crude but general choice of the scaling constant is $k_0 \geq \|\boldsymbol{X}\|_2/2$, again, regardless of Θ and λ. Yet in applying a convex penalty such as the nuclear norm penalty, we can use $k_0 = \|\boldsymbol{X}\|_2/(2\sqrt{2})$ to speed the convergence.

Some related works There has been a surge of interest in nuclear norm penalization recently, in which case the penalty in (2.10) simplifies to a multiple of the sum of all singular values of $\boldsymbol{B}^\circ$ or $\lambda\|\boldsymbol{B}^\circ\|_*$. This gives a *convex* optimization problem. In the statistics community, Yuan et al.

(2007) seem to be the first to study the nuclear norm penalized least squares estimator. A popular equivalent formulation of the nuclear norm minimization in optimization is through semidefinite programming (SDP) Fazel (2002). See, e.g., Candès and Recht (2009), Candès and Tao (2010), and Ma et al. (2009) for some recent theoretical and computational achievements.

Although the nuclear norm provides a convex relaxation to the rank penalty, this approximation works only under certain regularity conditions (e.g., Candès and Plan (2011)). Bunea, She and Wegkamp (2011) show that direct rank penalization achieves the same oracle rate in a much less restrictive manner. Yet in addition to the reduced rank regression studies (see Example 2.1), there have been very few attempts to extend the rank penalization beyond the Gaussian framework. Two commonly cited works are Yee and Hastie (2003) and Heinen and Rengifo (2008). See Section 1 for their limitations. In comparison with these works, our matrix thresholding algorithm has a theoretical guarantee of convergence, is simple to implement, and covers a wide family of penalty functions as well as loss functions.

Finally, we point out a major difference between the thresholding-based iterative selection procedures (TISP) (She, 2009) and the proposed algorithm which can be referred to as *matrix-TISP*. TISP aims for variable selection in a single-response model, while here we discuss singular value regularization in vector GLMs. The singular-value sparsity or low rankness, different than coefficient sparsity, offers a new type of parsimony that can be used for supervised feature extraction. It brings a true multivariate flavor into our analysis.

Feature extraction In many high-dimensional problems, *feature extraction*, by transforming the input variables and creating a reduced set of new features, is a useful technique for dimension-reduction. For example, PCA considers linear projections of correlated variables to construct new orthogonal features ordered by decreasing variances. For singular-value penalized models, once a low-rank estimate $\hat{\boldsymbol{B}}$ is obtained, one can attain the same goal. Suppose the rank of $\hat{\boldsymbol{B}}$ is r. A direct way is to apply the reduced form SVD on $\hat{\boldsymbol{B}}$, getting $\hat{\boldsymbol{B}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ with $\boldsymbol{D}$ an $r \times r$ diagonal matrix. Next, construct a new model matrix

$$(2.18) \qquad \textit{Type-I:} \quad \boldsymbol{Z} \triangleq \boldsymbol{X}\boldsymbol{U} \text{ (or } \boldsymbol{X}\boldsymbol{U}\boldsymbol{D}),$$

which has only r new predictors. We refer to this as *Type-I extraction*. It can be used for parameter tuning later.

On the other hand, it may be preferred to work on $\boldsymbol{X}\hat{\boldsymbol{B}}$ in some situations. Perform the spectral decomposition $\hat{\boldsymbol{B}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$, where $\boldsymbol{V}$ is an $m \times r$ orthogonal matrix. It follows that $\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{X}\hat{\boldsymbol{B}}\boldsymbol{V}\boldsymbol{V}^T$. Therefore, for the new design matrix $\boldsymbol{Z}$ defined by

$$(2.19)$$
$$\textit{Type-II (Post-Decorrelation):} \quad \boldsymbol{Z} \triangleq \boldsymbol{X}(\hat{\boldsymbol{B}}\boldsymbol{V}) \in \mathbb{R}^{n\times r},$$

each column (z-predictor) can be represented as a linear combination of the columns of $\boldsymbol{X}$, and the r newly obtained z-predictors are uncorrelated with each other, i.e., $\boldsymbol{Z}^T\boldsymbol{Z}$ is diagonal. We refer to this as *Type-II extraction* or *post-decorrelation*. Type-I and Type-II are not equivalent in general (but coincide for the RRR estimate). The (linear) feature extraction is supervised and the corresponding dimension reduction can be dramatic when r is much smaller than p.

Initial point   When the problem (2.10) is convex, we can further show (based on Theorem 2.1) that any Θ^σ-estimate is a global minimum point. In this case, the choice of the initial estimate $\boldsymbol{B}^{(0)}$ is not essential and a pathwise algorithm with warm starts can be used in computing the solution path $\hat{\boldsymbol{B}}(\lambda)$ for a series of values of λ. However, for nonconvex problems, we do not have such global optimality given any initial point $\boldsymbol{B}^0$. Although one can try multiple random starts, we found that empirically, simply setting $\boldsymbol{B}^{(0)}$ to be the zero matrix leads to a solution with good statistical performance. Intuitively, such initialization looks for a local optimum that is close to zero. Of course, other initialization choices are possible.

Parameter tuning   The challenge still comes from nonconvexity. Take the rank penalty as an example. The solution path $\hat{\boldsymbol{B}}(\lambda)$ is *discontinuous*, while the optimal penalty parameter λ (as a surrogate for the Lagrange multiplier in convex programmings) is a function of both the data $(\boldsymbol{X}, \boldsymbol{Y})$ and the true coefficient $\boldsymbol{B}$. Therefore, plain cross-validation with respect to λ does not seem to be appropriate, as slightly perturbed data may result in serious regularization parameter mismatches. We propose to cross-validate the *range space* of the low rank estimator (as a function of λ) and refer to it as the projective cross-validation (**PCV**). In the following, we focus on the rank-Frobenius penalty (2.8) and the associated hard-ridge thresholding rule (2.6) to describe the idea. Let $\hat{\boldsymbol{B}}$ be a Θ^σ_{HR} estimator obtained from (2.11) and write it as $\begin{bmatrix}\hat{\boldsymbol{\beta}}_0^T\\\hat{\boldsymbol{B}}^\circ\end{bmatrix}$ following our previous notation (cf. (2.9)). For the penalized part $\hat{\boldsymbol{B}}^\circ$, denote its rank by r and its SVD by $\hat{\boldsymbol{B}}^\circ = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ with $\boldsymbol{D} \in \mathbb{R}^{r\times r}$. Construct r new features (Type-I) $\boldsymbol{Z}^\circ = \boldsymbol{X}^\circ\boldsymbol{U}$ and set $\boldsymbol{Z} = [\tilde{\boldsymbol{x}}_0, \boldsymbol{Z}^\circ] = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^T$. Let $\hat{\boldsymbol{C}}^\circ = \boldsymbol{D}\boldsymbol{V}^T$ and $\hat{\boldsymbol{C}} = [\hat{\tilde{\boldsymbol{\beta}}}_0, \hat{\boldsymbol{C}}^{\circ T}]^T = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m] \in \mathbb{R}^{(r+1)\times m}$.

Proposition 2.4. *Under the condition on ρ given in Theorem 2.1, for any Θ^σ_{HR}-estimator $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{C}}$ defined above is a Frobenius penalized estimator associated with new model matrix $\boldsymbol{Z}$, i.e.,*

(2.20)
$$\hat{\boldsymbol{C}} \in \arg\min_{\boldsymbol{C}\in\mathbb{R}^{(r+1)\times m}} -\sum_{k=1}^m\sum_{i=1}^n L_{i,k}(\boldsymbol{c}_k; \boldsymbol{z}_i, y_{ik}) + \frac{\eta}{2}\|\boldsymbol{C}^\circ\|_F^2.$$

See its proof in Appendix E. When $\eta > 0$ or $\boldsymbol{Z}$ has full column rank, the optimization problem (2.20) is strictly convex and so $\hat{\boldsymbol{C}}$ is the unique optimal solution.

The proposition implies that once $\boldsymbol{U}$ is extracted, we can simply use maximum likelihood estimation on the projected data to obtain the rank penalized estimator, or ridge penalized estimation to obtain the rank-Frobenius penalized estimator. There is no need to run the more expensive reduced rank fitting algorithms.

We now state the K-fold PCV procedure for tuning the rank penalty parameter in (1.1).

1. Run Algorithm (2.11) on the *whole* dataset for a grid of values for λ. The solution path is denoted by $\hat{\boldsymbol{B}}(\lambda_l)$, $l = 1, \ldots, L$.
2. Obtain L candidate models via (2.18), each with a new model matrix $\boldsymbol{Z}(l) = \boldsymbol{X}\boldsymbol{U}(l)$, $1 \le l \le L$.
3. Compute the cross-validation error for each model. Concretely, partition the sample index set into K (roughly) even subsets $\mathcal{T}_1, \ldots, \mathcal{T}_K$. Given $\boldsymbol{Z}(l)$, fit a vector GLM on the data without the subset indexed by $\mathcal{T}_k$, and evaluate its validation error (measured by deviance) on the left-out subset. In all, K maximum likelihood estimates are obtained and their validation errors are summed up to yield the CV error of the candidate model $\boldsymbol{Z}(l)$. Repeat this for all $l : 1 \le l \le L$.
4. Find the optimal model that minimizes the CV error.

In the pursuit of a parsimonious model with very low rank, a BIC penalty term can be added to the CV error (She, 2012). This is necessary in the large-p setup (Chen and Chen, 2008).

PCV is much more efficient than CV because the more involved reduced rank fitting algorithm runs only once beforehand, rather than K times in the CV trainings. The ML fitting in Step 3, justified by Proposition 2.4, involves very few predictors. Another benefit of PCV is that the parameter mismatch issue is eliminated and all K trainings are regarding the same model and feature space.

When there is an additional ridge parameter (cf. (2.8)), the procedure still applies, but a two-dimensional grid for (λ, η) has to be used. Fortunately, according to our experience, the statistical performance is not very sensitive to small changes in the ridge parameter and we can choose a sparse grid for it. Step 3 now fits a series of l_2-penalized GLMs. But again, this type of problems is smooth and convex; Newton-based algorithms are reasonably fast. Finally, we mention that PCV shares some similarities to the selective cross-validation (SCV) proposed for variable selection (She, 2012).

3. RANK CONSTRAINED VECTOR GLMS FOR FEATURE SPACE REDUCTION

In this section, we study the reduced rank GLMs in constraint form (cf. (1.2)). For any $r \ge 1$ and $\eta \ge 0$, the problem

of interest is

$$(3.1) \quad \min_{\boldsymbol{B} \in \mathbb{R}^{(p+1) \times m}} -\sum_{k=1}^{m} \sum_{i=1}^{n} L_{i,k}(\boldsymbol{\beta}_k; \boldsymbol{x}_i, y_{ik}) + \frac{\eta}{2} \|\boldsymbol{B}^{\circ}\|_F^2$$
$$\text{s.t.} \quad rank(\boldsymbol{B}^{\circ}) \leq r.$$

The additional Frobenius norm penalty is to handle collinearity. Again, neither the penalty nor the constraint is imposed on the first row of $\boldsymbol{B}$.

We introduce a *quantile thresholding rule* $\Theta^{\#}(\cdot; r, \eta)$ as a variant of the hard-ridge thresholding. Given $1 \leq r \leq p$ and $\eta \geq 0$, $\Theta^{\#}(\boldsymbol{a}; r, \eta) : \mathbb{R}^p \to \mathbb{R}^p$ is defined for any $\boldsymbol{a} \in \mathbb{R}^p$ such that the r largest components of $\boldsymbol{a}$ (in absolute value) are shrunk by a factor of $(1 + \eta)$ and the remaining components are all set to be zero. In the case of ties, a random tie breaking rule is used. The matrix version of $\Theta^{\#}$ is defined as

$$\Theta^{\#\sigma}(\boldsymbol{B}; \lambda) \triangleq \boldsymbol{U} \text{diag}\{\Theta^{\#}([\sigma_i^{(\boldsymbol{B})}]; r, \eta)\} \boldsymbol{V}^T, \quad \forall \boldsymbol{B} \in \mathbb{R}^{p \times m}$$

where $\boldsymbol{U}$, $\boldsymbol{V}$, and $\text{diag}\{\sigma_i^{(\boldsymbol{B})}\}$ are obtained from the SVD of $\boldsymbol{B}$: $\boldsymbol{B} = \boldsymbol{U}^T \text{diag}\{\sigma_i^{(\boldsymbol{B})}\} \boldsymbol{V}$.

Then, a simple procedure similar to (2.11) can be used to solve (3.1): given $\boldsymbol{B}^{(j)}$, perform the update

$$(3.2)$$
$$\begin{cases} \boldsymbol{B}^{\circ(j+1)} = \Theta^{\#\sigma}(\boldsymbol{B}^{\circ(j)} + \boldsymbol{X}^{\circ T} \boldsymbol{Y} - \boldsymbol{X}^{\circ T} \boldsymbol{\mu}(\boldsymbol{B}^{(j)}); r, \eta), \\ \tilde{\boldsymbol{\beta}}_0^{(j+1)} = \tilde{\boldsymbol{\beta}}_0^{(j)} + (\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{B}^{(j)}))^T \tilde{\boldsymbol{x}}_0. \end{cases}$$

Starting with any $\boldsymbol{B}^{(0)} \in \mathbb{R}^{(p+1) \times m}$, denote the sequence obtained from (3.2) by $\{\boldsymbol{B}^{(j)}\}$. Let F be the objective function (3.1). Define ρ as in Theorem 2.1.

Theorem 3.1. *If $\rho \leq 1$, $F(\boldsymbol{B}^{(j)})$ is decreasing and satisfies*

$$F(\boldsymbol{B}^{(j)}) - F(\boldsymbol{B}^{(j+1)}) \geq (1 - \rho) \|\boldsymbol{B}^{(j)} - \boldsymbol{B}^{(j+1)}\|_F^2 / 2,$$

and $rank(\boldsymbol{B}^{\circ(j)}) \leq r$, $\forall j \geq 1$.

See Appendix F for its proof. The preliminary scaling of $\boldsymbol{X}/k_0$ for any $k_0 \geq \sqrt{\rho}$ guarantees the convergence. Still, PCV can be used for parameter tuning and model selection although the obtained estimate may not be globally optimal due to nonconvexity.

Rank penalty vs. rank constraint We have developed algorithms (2.11) and (3.2) for solving (1.1) (or (2.10)) and (1.2) (or (3.1)), respectively. The obtained estimates are usually local optimizers of the corresponding objective functions. However, in non-Gaussian GLM setups, we found that the nonconvexity of either problem can be very strong. For instance, there may exist many local optima all having the same rank but spanning different subspaces in $\mathbb{R}^p$. In consideration of this, the penalized solution path $\hat{\boldsymbol{B}}(\lambda)$ ($0 \leq \lambda < +\infty$) may provide more candidate models of certain rank (if existing) than the constrained solution path $\hat{\boldsymbol{B}}(r)$ ($r = 1, 2, \ldots, p \wedge m$), which is advantageous in the

stage of parameter tuning. This phenomenon is often observed in datasets where p is comparable to or larger than n. Note that typically the direct rank penalized $\hat{\boldsymbol{B}}(\lambda)$ has no rank monotonicity.

On the other hand, computing the solution path for the penalty form is often more time-consuming in large-p applications. The path $\hat{\boldsymbol{B}}(\lambda)$ has jumps. Assuming no prior knowledge of the appropriate interval for λ, one has to specify a large search grid fine enough to cover a reasonable number of candidate models. By contrast, for the problem of constraint form, we can set a small upper bound for r in pursuing a low rank model (say, $r \leq 0.5n \wedge p \wedge m$ could be good), and the natural grid spacing is 1. With the grid focusing on small values of r (which amounts to applying large thresholds in the iteration steps), Algorithm (3.2) runs efficiently.

Feature space reduction To combine the virtues of both approaches, we propose to solve the rank constrained problem to perform feature space reduction, and then run the rank penalized algorithm in the reduced feature space. This is very helpful in large-p applications. A crude sketch is as follows. First, we set $r = \alpha n \wedge p \wedge m$ with $\alpha < 1$ (e.g., $\alpha = 0.5$) and solve (1.2). Using the estimate $\hat{\boldsymbol{B}}(r)$, we execute Type-I feature extraction (2.18) to construct a new model matrix $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{U}_1(r)$ with only r factors (in addition to the intercept). Next, we turn to the penalized problem (2.10) on $(\boldsymbol{Y}, \boldsymbol{Z})$: Get the solution path from running Algorithm (2.11), and tune the regularization parameters to find the optimal estimate, denoted by $\hat{\boldsymbol{B}}'(\lambda_o)$. Our final coefficient matrix estimate is given by $\boldsymbol{U}_1(r)\hat{\boldsymbol{B}}'(\lambda_o)$. A small number of new predictive features can be constructed (and decorrelated) based on (2.19).

According to this scheme, the sample size of the reduced problem on $\boldsymbol{Z}$ is large relative to the reduced dimension. It is not difficult to show that for $n > p$, the update in (2.11) is essentially a contraction, and so Algorithm (2.11) converges fast.

A crucial assumption here is that the rank of the true model, denoted by r^*, is very small, compared with the sample size n. This makes it possible to choose a safe rank constraint value r in (1.2), which, though possibly much less than p, is still much larger than the true r^*. Hence the computational cost of obtaining a solution path according to Algorithm (2.11) can be effectively reduced with little performance loss. This idea shares similarity with the variable screening (Fan and Lv, 2008) proposed in the context of sparse variable selection. In the process of screening, all relevant variables should be kept, while in feature space reduction, only the necessary factors, being linear combinations of the original predictors and typically as few as a handful, are required to lie in reduced feature space we project $\boldsymbol{X}$ onto.

In implementation, we further adopt a path-following (annealing) idea to reduce computational load and avoid greedy reduction. Define a cooling schedule $r(t)$ ($0 \leq t \leq T$) with $r(0) = p$ and $r(T) = r$, where r is an upper bound of

Table 2. Rank constraint vs. rank penalty. Misclassification rates of the constrained and penalized reduced rank logistic regressions (RRL$^{(c)}$ and RRL$^{(p)}$) are shown for the `zipcode` (sub)dataset where $p = 257, n = 300$. The rank r controls the # of newly constructed features

r	1	2	3	4
RRL$^{(c)}$	66.52%	55.06%	38.47%	33.83%
RRL$^{(p)}$	66.52%	55.06%	38.32%, 38.37%	33.83%
RRL$^{(c)}$+SVM	59.24%	47.48%	33.58%	30.79%
RRL$^{(p)}$+SVM	59.24%	47.48%	33.63%, 33.58%	30.79%

5	6	7	8	9
24.86%	22.27%	21.33%	21.33%	20.43%
24.81%, 24.86%	22.42%	21.47%, 21.33%	21.33%	20.43%, 19.13%
23.02%	20.38%	20.43%	20.28%	18.53%
23.02%	21.08%	20.33%, 20.43%	20.33%, 20.28%	18.53%, 15.84%

the target rank. We conduct *progressive* feature space reduction as follows. (As aforementioned, $\boldsymbol{Z}^\circ$ refers to $\boldsymbol{Z}$ without the first column, $\boldsymbol{B}^\circ$ refers to $\boldsymbol{B}$ without the first row, and $\boldsymbol{U}^\circ$ refers to $\boldsymbol{U}$ without the first row and the first column.)

1. Let $t \leftarrow 0$, $\boldsymbol{Z} \leftarrow \boldsymbol{X}$, $\boldsymbol{U} \leftarrow \boldsymbol{I}$.
2. Iterate until $r(t) \leq r$:
   (a) Set the rank constraint value to be $r(t)$ and perform the update (3.2) on $(\boldsymbol{Y}, \boldsymbol{Z})$ for at most M times (with M pre-specified);
   (b) Obtain the left singular vectors of the current slope estimate $\boldsymbol{B}^\circ$, denoted by $\boldsymbol{U}_1(r(t))$;
   (c) Let $\boldsymbol{Z}^\circ \leftarrow \boldsymbol{Z}^\circ \boldsymbol{U}_1(r(t))$, $\boldsymbol{U}^\circ \leftarrow \boldsymbol{U}^\circ \boldsymbol{U}_1(r(t))$;
   (d) $t \leftarrow t + 1$.

At the end, $\boldsymbol{Z}$ is delivered as the new design, and the orthogonal matrix $\boldsymbol{U}$ gives the accumulated transformation matrix.

The previously described prototype reduction scheme corresponds to $r(t) = r$ for any t. With an annealing algorithm design, the dimensionality of the feature space keeps dropping; the $\boldsymbol{B}$ involved in (3.2) has only $r(t)$ columns. A slow cooling schedule with a small number of M is recommended. Based on our experiments, it is not too greedy and is usually computationally affordable for large-p problems.

4. DATA EXAMPLES

We use two real data examples to illustrate the proposed methodology for dimension reduction and supervised feature extraction.

Example 4.1. First, we make a practical comparison of the rank penalized estimators from solving (1.1) and the rank constrained estimators from solving (1.2) by use of a zipcode dataset. The whole dataset (available at the website of Hastie et al. (2001)) contains normalized handwritten digits in 16×16 grayscale images. The digits were originally scanned from envelopes by the U.S. Postal Service and have

been deslanted and size normalized. The space of pixel predictors is of dimension 256. We standardized all such predictors. The intercept term is included in the model and is always unpenalized. We introduced $m = 9$ indicator response variables for digits 0–8, using 9 as the reference class.

The training set is large in comparison with p and m (7,291 images). We chose a subset of $n = 300$ at random in this experiment to compare the penalized solution path and the constrained solution path. No additional Frobenius-norm penalty was enforced. The prediction results of the estimates are shown in Table 2, evaluated on 2,007 test observations.

From the table, at certain values of r, the rank penalty offered more candidate models along its solution path than the rank constraint. Note that these rank-r estimators may behave differently in prediction and feature extraction. For $p \sim n$ or $p > n$, this phenomenon is commonly seen. With an appropriate parameter tuning strategy, the penalty form gives better chances to achieve a low error rate.

Of course, this comes with a price in computation. In our experiment, the time for obtaining the RRL$^{(c)}$ path was less than one minute, while computing the RRL$^{(p)}$ path, with a 50-point grid for λ, took about four minutes.

There is no obligation to predict through the obtained estimator; perhaps more useful is the much lower dimensional feature space yielded from such an estimator. Fancier classifiers such as SVM can be applied with the new features automatically extracted and decorrelated via (2.19), and result in lower error rates as shown in the table.

Finally, we add a comment that in some situations there may exist no penalized solution at certain rank values. Yet with a large λ-grid chosen, the performance of the penalized estimator (after parameter tuning) does not seem to be worse than that of the constrained estimator.

Example 4.2. The *Computer Audition Lab 500* (`CAL500`) dataset is collected by Turnbull et al. (2008) and involves 502 Western popular songs by different artists selected from

the past 50 years. Digital audio files were played to students to annotate these songs with $m = 174$ words representing emotion, genre, instrument, vocals, etc. The concepts characterized by the words are not mutually exclusive and one song can be annotated with multiple labels. This is called *multi-label* data in machine learning. The predictors are MFCC-Delta audio features from analyzing a short-time segment of the audio signal. Turnbull et al. (2008) used 68 such feature vectors. To allow for interactions between these audio features and to make a more challenging problem, we consider a full quadratic model including all main effects, quadratic effects, and pairwise interactions. Hence $p = 68 + 68(69)/2 + 1 = 2,415$. We split the data into two halves and used $n = 251$ songs for training and the other 251 for testing.

For this small-sample-size-high-dimensional problem, the SVM using all 2,415 predictors gave a total misclassification rate of 21.2%, which is not all bad. On the other hand, the proposed reduced rank methodology can be applied to automatically construct new predictive audio features, possibly much fewer than 2,415. The supervised nature is important because only the audio features helpful in annotation (classification) are truly meaningful in this learning task.

First, we conducted the progressive feature space reduction introduced in Section 3, with the upper bound of the target rank set to be 20. Then we ran Algorithm (2.11) to fit a penalized reduced rank vector logistic regression with the 20 extracted features. The rank-Frobenius penalty was chosen due to serious collinearity arising from the high-dimensional quadratic model. The parameters were tuned by 5-fold PCV with BIC correction.

Surprisingly, our final estimate $\hat{\boldsymbol{B}}$ has $rank(\hat{\boldsymbol{B}}^\circ) = 2$, which gives a dramatic dimension reduction from 2,514 to 2. But the SVM trained based on just the two new features yielded an improved error rate of 14.13%. In fact, even using the vanilla reduced rank estimator, we can achieve an error rate of 14.36%. The per-word precision and recall (cf. Turnbull et al. (2008) for the detailed definitions) are, respectively, 35.6% and 8.7% on the test dataset, comparable to the rates of the two advocated approaches in Turnbull et al. (2008). But our model is more parsimonious and creates two concise audio summary indexes for semantic annotation.

5. CONCLUSION

Supervised linear feature extraction can be obtained from a reduced rank vector model. We studied rank penalized and rank constrained generalized linear models and discussed how to adapt them to feature extraction and feature space reduction. The latter technique helps to reduce the computational cost significantly in high dimensions. We also noticed the strong nonconvexity of such problems raises some serious issues in data-resampling based parameter tunings, but the proposed projective cross-validation works decently

in general and is efficient. Through reduced rank modeling, dimension reduction can be attained if the rank of the model is small relative to the number of predictors. The work can be viewed as a supervised and parametric generalization of the principle component analysis.

APPENDIX A. PROOF OF PROPOSITION 2.1

To prove Proposition 2.1, we first introduce two lemmas.

Lemma A.1 (von Neumann (1937)). *Let $\boldsymbol{A}, \boldsymbol{B}$ be two $n \times n$ matrices. Then*

$$(A.1) \qquad |Tr(\boldsymbol{AB})| \leq \sum \sigma_i(\boldsymbol{A})\sigma_i(\boldsymbol{B}),$$

where $\sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \cdots \geq \sigma_n(\boldsymbol{A})$ and $\sigma_1(\boldsymbol{B}) \geq \sigma_2(\boldsymbol{B}) \geq \cdots \geq \sigma_n(\boldsymbol{B})$ are ordered singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively.

We refer to Grigorieff (1991) for an elementary proof.

Lemma A.2. *Given a thresholding rule Θ, let P be any penalty satisfying condition (2.5) in Proposition 2.1. Then, the univariate minimization problem $\min_\theta (t-\theta)^2/2 + P(\theta; \lambda)$ has a unique optimal solution $\hat{\theta} = \Theta(t; \lambda)$ for every t at which $\Theta(\cdot; \lambda)$ is continuous.*

Proof. Apply Lemma 1 in She (2012). $\qquad\square$

Proof of the optimality part of Proposition 2.1. Let $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$ and assume $n \geq m$ without any loss of generality. Let $\boldsymbol{Y} = \boldsymbol{U}_0 \boldsymbol{D}_0 \boldsymbol{V}_0^T$ and $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ be the SVDs where $\boldsymbol{D}_0 = \text{diag}(d_{0,i})$ and $\boldsymbol{D} = \text{diag}(d_i)$ with $d_{0,1} \geq d_{0,2} \geq \cdots \geq d_{0,m}$ and $d_1 \geq d_2 \geq \cdots \geq d_m$. Clearly,

$$\|\boldsymbol{Y} - \boldsymbol{B}\|_F^2 = \|\boldsymbol{Y}\|_F^2 + \|\boldsymbol{B}\|_F^2 - 2Tr(\boldsymbol{Y}^T\boldsymbol{B})$$
$$= \|\boldsymbol{Y}\|_F^2 + \|\boldsymbol{B}\|_F^2 - 2Tr([\boldsymbol{Y} \quad \boldsymbol{0}]^T[\boldsymbol{B} \quad \boldsymbol{0}]),$$

where $[\boldsymbol{B} \quad \boldsymbol{0}] \in \mathbb{R}^{n \times n}$ and $[\boldsymbol{Y} \quad \boldsymbol{0}] \in \mathbb{R}^{n \times n}$. It follows from Lemma A.1 that $Tr(\boldsymbol{Y}^T\boldsymbol{B}) \leq \sum d_{0,i}d_i$. Hence

$$(A.2)$$
$$F(\boldsymbol{B}) \geq (\|\boldsymbol{D}_0\|_F^2 + \|\boldsymbol{D}\|_F^2 - 2Tr(\boldsymbol{D}_0\boldsymbol{D}))/2 + \sum P(d_i; \lambda)$$
$$= \sum (d_{0,i} - d_i)^2/2 + \sum P(d_i; \lambda).$$

Now the problem reduces to

$$\min_{d_i} \sum (d_{0,i} - d_i)^2/2 + \sum P(d_i; \lambda).$$

The optimal solution $\hat{\boldsymbol{B}}$ then follows from Lemma A.2. $\quad\square$

The argument above only implies the singular values of $\hat{\boldsymbol{B}}$ are unique (up to permutation). Although one can possibly argue the uniqueness of $\hat{\boldsymbol{B}}$ by studying the condition under which equality is achieved in (A.2), another formal proof of the uniqueness is deferred to Appendix B.

APPENDIX B. PROOF OF PROPOSITION 2.2

Let $\boldsymbol{B} = \hat{\boldsymbol{B}} + \boldsymbol{\Delta}$. Suppose $\boldsymbol{Y} = \boldsymbol{U}_0 \boldsymbol{D}_0 \boldsymbol{V}_0^T$, $\hat{\boldsymbol{B}} = \boldsymbol{U}_0 \hat{\boldsymbol{D}} \boldsymbol{V}_0^T$, and $\boldsymbol{B} = \boldsymbol{U} \boldsymbol{D} \boldsymbol{V}^T$ are the SVDs. We have

$$\|\boldsymbol{Y} - \boldsymbol{B}\|_F^2/2 - \|\boldsymbol{D}_0 - \boldsymbol{D}\|_F^2/2$$
$$= -Tr(\boldsymbol{B}^T \boldsymbol{Y}) + Tr(\boldsymbol{D}_0 \boldsymbol{D})$$
$$= -Tr(\boldsymbol{B}^T(\boldsymbol{Y} - \hat{\boldsymbol{B}})) + Tr((\boldsymbol{D}_0 - \hat{\boldsymbol{D}})\boldsymbol{D})$$
$$\quad + Tr(\boldsymbol{D}\hat{\boldsymbol{D}}) - Tr(\boldsymbol{B}^T \hat{\boldsymbol{B}})$$
$$= -Tr(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T \boldsymbol{U}_0(\boldsymbol{D}_0 - \hat{\boldsymbol{D}})\boldsymbol{V}_0^T) + Tr((\boldsymbol{D}_0 - \hat{\boldsymbol{D}})\boldsymbol{D})$$
$$\quad + Tr(\boldsymbol{D}\hat{\boldsymbol{D}}) - Tr(\boldsymbol{B}^T \hat{\boldsymbol{B}}).$$

By Proposition 2.1 $\hat{\boldsymbol{D}} \preceq \boldsymbol{D}_0$, i.e., $\boldsymbol{D}_0 - \hat{\boldsymbol{D}}$ is positive semi-definite. By augmenting $\boldsymbol{Y} - \boldsymbol{B}$ and $\boldsymbol{B}$ and applying Lemma A.1, we can prove

$$Tr(\boldsymbol{D}(\boldsymbol{D}_0 - \hat{\boldsymbol{D}})) \geq Tr(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T \boldsymbol{U}_0(\boldsymbol{D}_0 - \hat{\boldsymbol{D}})\boldsymbol{V}_0^T),$$

from which it follows that

$$\|\boldsymbol{Y} - \boldsymbol{B}\|_F^2/2 - \|\boldsymbol{D}_0 - \boldsymbol{D}\|_F^2/2$$
$$\geq Tr(\boldsymbol{D}\hat{\boldsymbol{D}}) - Tr(\boldsymbol{B}^T \hat{\boldsymbol{B}}) \geq C_1(Tr(\boldsymbol{D}\hat{\boldsymbol{D}}) - Tr(\boldsymbol{B}^T \hat{\boldsymbol{B}})).$$

Now we have

$$Q(\boldsymbol{B}) - Q(\hat{\boldsymbol{B}})$$
$$= \|\boldsymbol{Y} - \boldsymbol{B}\|_F^2/2 - \|\boldsymbol{Y} - \hat{\boldsymbol{B}}\|_F^2/2$$
$$\quad + \sum P_\Theta(d_i; \lambda) - \sum P_\Theta(\hat{d}_i; \lambda)$$
$$\geq \|\boldsymbol{D}_0 - \boldsymbol{D}\|_F^2/2 - \|\boldsymbol{D}_0 - \hat{\boldsymbol{D}}\|_F^2/2$$
$$\quad + \sum P_\Theta(d_i; \lambda) - \sum P_\Theta(\hat{d}_i; \lambda) + C_1 Tr(\boldsymbol{D}\hat{\boldsymbol{D}} - \boldsymbol{B}^T \hat{\boldsymbol{B}})$$
$$= \sum((d_{0,i} - d_i)^2/2 + P_\Theta(d_i; \lambda))$$
$$\quad - ((d_{0,i} - \hat{d}_i)^2/2 + P_\Theta(\hat{d}_i; \lambda)) + C_1 Tr(\boldsymbol{D}\hat{\boldsymbol{D}} - \boldsymbol{B}^T \hat{\boldsymbol{B}})$$
$$\geq C_1 \sum(d_i - \hat{d}_i)^2/2 + C_1 Tr(\boldsymbol{D}\hat{\boldsymbol{D}} - \boldsymbol{B}^T \hat{\boldsymbol{B}})$$
$$= C_1(\|\boldsymbol{D} - \hat{\boldsymbol{D}}\|_F^2/2 + Tr(\boldsymbol{D}\hat{\boldsymbol{D}}) - Tr(\boldsymbol{B}^T \hat{\boldsymbol{B}}))$$
$$= C_1(\|\boldsymbol{D}\|_F^2/2 + \|\hat{\boldsymbol{D}}\|_F^2/2 - Tr(\boldsymbol{B}^T \hat{\boldsymbol{B}}))$$
$$= C_1 \|\boldsymbol{B} - \hat{\boldsymbol{B}}\|_F^2/2.$$

The second inequality is due to Lemma 2 in She (2012).

Proof of the optimality part of Proposition 2.1. From the comment in Appendix A, any optimal solution $\boldsymbol{B}$ must have the same nonzero singular values (up to permutation) as $\hat{\boldsymbol{B}}$, i.e., $d_i = \hat{d}_i$, seen from the proof of Proposition 2.1. A more careful examination of the proof of Proposition 2.2 shows $Q(\boldsymbol{B}) - Q(\hat{\boldsymbol{B}}) \geq Tr(\boldsymbol{D}\hat{\boldsymbol{D}} - \boldsymbol{B}^T \hat{\boldsymbol{B}}) = \|\boldsymbol{B} - \hat{\boldsymbol{B}}\|_F^2/2$. Therefore, the globally optimal solution $\hat{\boldsymbol{B}}$ in Proposition 2.1 must be unique. $\square$

APPENDIX C. PROOF OF THEOREM 2.1

The proof is similar to that of Theorem 2.1 in She (2012). Define a surrogate function G for any $\boldsymbol{A} = [\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_m] = [\tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}}_1, \dots, \tilde{\boldsymbol{\alpha}}_p]^T$ and $\boldsymbol{B} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m] \in \mathbb{R}^{(p+1) \times m}$

$$G(\boldsymbol{B}, \boldsymbol{A})$$
$$= -\sum_{k=1}^m \sum_{i=1}^n L_{i,k}(\boldsymbol{\alpha}_k) + \sum_{s=1}^{p \wedge m} P(\sigma_s^{(\boldsymbol{A}^\circ)}; \lambda)$$
$$\quad + \frac{1}{2}\|\boldsymbol{A} - \boldsymbol{B}\|_F^2 - \sum_{k=1}^m \sum_{i=1}^n (b(\boldsymbol{x}_i^T \boldsymbol{\alpha}_k) - b(\boldsymbol{x}_i^T \boldsymbol{\beta}_k))$$
$$\quad + \sum_{k=1}^m \sum_{i=1}^n \mu_{i,k}(\boldsymbol{x}_i^T \boldsymbol{\alpha}_k - \boldsymbol{x}_i^T \boldsymbol{\beta}_k),$$

where $\mu_{i,k} = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}_k) = b'(\boldsymbol{x}_i^T \boldsymbol{\beta}_k)$. It can be shown that given $\boldsymbol{B}$, minimizing G over $\boldsymbol{A}$ is equivalent to

$$\arg\min_{\boldsymbol{A}} \frac{1}{2}\left\|\boldsymbol{A} - \left[\boldsymbol{B} + \boldsymbol{X}^T \boldsymbol{Y} - \boldsymbol{X}^T \boldsymbol{\mu}(\boldsymbol{B})\right]\right\|_F^2$$
$$+ \sum_{s=1}^{p \wedge m} P(\sigma_s^{(\boldsymbol{A}^\circ)}; \lambda).$$

By Proposition 2.1, $\boldsymbol{B}^{(j+1)}$ in (2.11) can be characterized by $\arg\min_{\boldsymbol{A}} G(\boldsymbol{B}^{(j)}, \boldsymbol{A})$. Furthermore, we have for any $\boldsymbol{\Delta} \in \mathbb{R}^{(p+1) \times m}$

(C.1)
$$G(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j+1)} + \boldsymbol{\Delta}) - G(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j+1)}) \geq \frac{C_1}{2}\|\boldsymbol{\Delta}\|_F^2$$

with $C_1 = \max(0, 1 - \mathcal{L}_\Theta)$, by applying Proposition 2.2 and Lemma 1 in She (2012), and noting that $q_s(\sigma^{(\boldsymbol{B}^{\circ(j+1)})}) = 0$, for $\boldsymbol{B}^{\circ(j+1)}$ obtained by Θ^σ-thresholding.

Next, Taylor series expansion gives

$$F(\boldsymbol{B}^{(j+1)})$$
$$+ \sum_k \frac{1}{2}(\boldsymbol{\beta}_k^{(j+1)} - \boldsymbol{\beta}_k^{(j)})^T (\boldsymbol{I} - \boldsymbol{\mathcal{I}}(\boldsymbol{\xi}_k^{(j)}))(\boldsymbol{\beta}_k^{(j+1)} - \boldsymbol{\beta}_k^{(j)})$$
$$= G(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j+1)})$$
$$\leq G(\boldsymbol{B}^{(j)}, \boldsymbol{B}^{(j)})$$
$$\quad - \sum_k \frac{C_1}{2}(\boldsymbol{\beta}_k^{(j+1)} - \boldsymbol{\beta}_k^{(j)})^T (\boldsymbol{\beta}_k^{(j+1)} - \boldsymbol{\beta}_k^{(j)})$$
$$= F(\boldsymbol{B}^{(j)}) - \frac{C_1}{2}\|\boldsymbol{B}^{(j+1)} - \boldsymbol{B}^{(j)}\|_F^2$$

(2.12) can be obtained. In fact, this decreasing property holds for any $\rho \leq 2 - \mathcal{L}_\Theta$.

Let $\boldsymbol{B}^{(j_l)} \to \boldsymbol{B}^*$ as $l \to \infty$. Under the condition $\rho < 2 - \mathcal{L}_\Theta$, C is strictly positive and $\|\boldsymbol{B}^{(j_l+1)} - \boldsymbol{B}^{(j_l)}\|_F^2/2 \leq (F(\boldsymbol{B}^{(j_l)}) - F(\boldsymbol{B}^{(j_l+1)}))/C \leq (F(\boldsymbol{B}^{(j_l)}) - F(\boldsymbol{B}^{(j_{k+1})}))/C \to 0$. That is, $\Theta^\sigma(\boldsymbol{B}^{\circ(j_l)} + \boldsymbol{X}^{\circ T} \boldsymbol{Y} - \boldsymbol{X}^{\circ T} \boldsymbol{\mu}(\boldsymbol{B}^{(j_l)}); \lambda) - \boldsymbol{B}^{\circ(j_l)} \to 0$. Therefore, $\boldsymbol{B}^*$ is a solution to (2.13) due to the continuity assumption.

APPENDIX D. PROOF OF PROPOSITION 2.3

Let $\boldsymbol{M} = (\boldsymbol{X}^T\boldsymbol{X})^{-1/2}\boldsymbol{X}^T\boldsymbol{Y}$ and $r_0 = \text{rank}(\boldsymbol{M})$. Obviously, $r_0 \leq p \wedge m$. Note that $\boldsymbol{M}^T\boldsymbol{M} = \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y}$. Assume $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ is the SVD of $\boldsymbol{M}$ with $\boldsymbol{U} \in \mathbb{R}^{p \times r_0}$, $\boldsymbol{V} \in \mathbb{R}^{m \times r_0}$, and $\boldsymbol{D} \in \mathbb{R}^{r_0 \times r_0}$. Suppose all (positive) diagonal entries of $\boldsymbol{D}$ are arranged in decreasing order. Let $\boldsymbol{A} \triangleq \boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{B}}$. To prove $\hat{\boldsymbol{B}}$ obeys the Θ^σ-equation (2.13) for hard-thresholding, it suffices to show that (i) there exists a $p \times r_0$ orthogonal matrix $\boldsymbol{U}_*$ satisfying $\boldsymbol{U}_*^T\boldsymbol{U}_* = \boldsymbol{I}$ such that $\boldsymbol{U}_*^T(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T)\boldsymbol{U}_*$ and $\boldsymbol{U}_*^T(\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{U}_*$ are both diagonal; (ii) there exists an $m \times r_0$ orthogonal matrix $\boldsymbol{V}_*$ such that $\boldsymbol{V}_*^T(\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}})\boldsymbol{V}_*$ and $\boldsymbol{V}_*^T(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{V}_*$ are both diagonal; (iii) $Tr(\hat{\boldsymbol{B}}^T\boldsymbol{A}) = 0$; (iv) the singular values of $\boldsymbol{A}$ are all bounded by λ.

Recall that $r = \max\{i : d_i \geq \lambda\}$ and $\boldsymbol{V}_r = \boldsymbol{V}[\,,1{:}r]$. Introduce $\boldsymbol{V}_{-r} = \boldsymbol{V}[\,,(r+1){:}r_0]$, formed by deleting the first r columns in $\boldsymbol{V}$. Then we have

$$
\begin{aligned}
\text{(D.1)} \quad \boldsymbol{A} &= \boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{H}\boldsymbol{Y}\mathcal{P}_{\boldsymbol{V}_r} \\
&= \boldsymbol{X}^T\boldsymbol{Y}(\boldsymbol{I} - \mathcal{P}_{\boldsymbol{V}_r}) = \boldsymbol{X}^T\boldsymbol{Y}\mathcal{P}_{\boldsymbol{V}_{-r}} \\
&= \boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{V}_{-r}\boldsymbol{V}_{-r}^T = (\boldsymbol{X}^T\boldsymbol{X})^{1/2}\boldsymbol{M}\boldsymbol{V}_{-r}\boldsymbol{V}_{-r}^T.
\end{aligned}
$$

Obviously, $Tr(\hat{\boldsymbol{B}}^T\boldsymbol{A}) = 0$. (iii) is true. On the other hand, we can rewrite $\hat{\boldsymbol{B}}$ as

$$
\begin{aligned}
\text{(D.2)} \quad \hat{\boldsymbol{B}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1/2}\boldsymbol{M}\boldsymbol{V}_r\boldsymbol{V}_r^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1/2}\boldsymbol{M}\boldsymbol{V}\begin{bmatrix} \boldsymbol{I}_{r \times r} & \\ & \boldsymbol{0}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\boldsymbol{V}^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1/2}\boldsymbol{U}\boldsymbol{D}\begin{bmatrix} \boldsymbol{I}_{r \times r} & \\ & \boldsymbol{0}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\boldsymbol{V}^T \\
\text{(D.3)} \quad &= (\boldsymbol{X}^T\boldsymbol{X})^{-1/2}\Theta_H^\sigma(\boldsymbol{M};\lambda).
\end{aligned}
$$

Now we obtain

$$
\begin{aligned}
\text{(D.4)} \quad \hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}} &= \boldsymbol{V}\begin{bmatrix} \boldsymbol{I}_{r \times r} & \\ & \boldsymbol{0}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\boldsymbol{D}\boldsymbol{U}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{U}\boldsymbol{D} \\
&\cdot \begin{bmatrix} \boldsymbol{I}_{r \times r} & \\ & \boldsymbol{0}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\boldsymbol{V}^T,
\end{aligned}
$$

$$
\begin{aligned}
\text{(D.5)} \quad \boldsymbol{A}^T\boldsymbol{A} &= \boldsymbol{V}\begin{bmatrix} \boldsymbol{0}_{r \times r} & \\ & \boldsymbol{I}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\boldsymbol{D}\boldsymbol{U}^T(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{U}\boldsymbol{D} \\
&\cdot \begin{bmatrix} \boldsymbol{0}_{r \times r} & \\ & \boldsymbol{I}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\boldsymbol{V}^T.
\end{aligned}
$$

(iv) is straightforward from (D.5):

$$
\begin{aligned}
\|\boldsymbol{A}^T\boldsymbol{A}\|_2 &\leq \|\boldsymbol{X}\|_2^2 \cdot \left\| \boldsymbol{U}\boldsymbol{D}\begin{bmatrix} \boldsymbol{0}_{r \times r} & \\ & \boldsymbol{I}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\boldsymbol{V}^T \right\|_2^2 \\
&\leq 1 \cdot d_{r+1}^2 \leq \lambda^2.
\end{aligned}
$$

(D.4) + (D.5) also implies (ii). In fact, introducing $\boldsymbol{G} = \boldsymbol{D}\boldsymbol{U}^T(\boldsymbol{X}\boldsymbol{X})^{-1}\boldsymbol{U}\boldsymbol{D}$, $\boldsymbol{H} = \boldsymbol{D}\boldsymbol{U}^T(\boldsymbol{X}\boldsymbol{X})\boldsymbol{U}\boldsymbol{D}$, $\boldsymbol{G}_{11} = \boldsymbol{G}[1{:}r,1{:}r]$, $\boldsymbol{H}_{22} = \boldsymbol{H}[(r+1){:}r_0,(r+1){:}r_0]$, and assuming the spectral decompositions of the two submatrices are given by $\boldsymbol{G}_{11} = \boldsymbol{U}_{11}^G\boldsymbol{D}_{11}^G(\boldsymbol{U}_{11}^G)^T$ and $\boldsymbol{H}_{22} = \boldsymbol{U}_{22}^H\boldsymbol{D}_{22}^H(\boldsymbol{U}_{22}^H)^T$, respectively, then,

$$
\boldsymbol{V}_* = \boldsymbol{V}\begin{bmatrix} \boldsymbol{U}_{11}^G & \\ & \boldsymbol{U}_{22}^H \end{bmatrix}
$$

simultaneously diagonalizes $\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}}$ and $\boldsymbol{A}^T\boldsymbol{A}$ and satisfies $\boldsymbol{V}_*^T\boldsymbol{V}_* = \boldsymbol{I}$.

Finally, we construct $\boldsymbol{U}_*$ to prove (i). From (D.3) and (D.2),

$$
\begin{aligned}
\text{(D.6)} \quad \hat{\boldsymbol{B}}\boldsymbol{V}_* &= (\boldsymbol{X}^T\boldsymbol{X})^{-1/2}\boldsymbol{U}\boldsymbol{D} \\
&\cdot \begin{bmatrix} \boldsymbol{I}_{r \times r} & \\ & \boldsymbol{0}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\begin{bmatrix} \boldsymbol{U}_{11}^G & \\ & \boldsymbol{U}_{22}^H \end{bmatrix},
\end{aligned}
$$

$$
\begin{aligned}
\text{(D.7)} \quad \boldsymbol{A}\boldsymbol{V}_* &= (\boldsymbol{X}^T\boldsymbol{X})^{1/2}\boldsymbol{U}\boldsymbol{D} \\
&\cdot \begin{bmatrix} \boldsymbol{0}_{r \times r} & \\ & \boldsymbol{I}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\begin{bmatrix} \boldsymbol{U}_{11}^G & \\ & \boldsymbol{U}_{22}^H \end{bmatrix}.
\end{aligned}
$$

Let $\tilde{\boldsymbol{G}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1/2}\boldsymbol{U}\boldsymbol{D}$ and $\tilde{\boldsymbol{H}} = (\boldsymbol{X}^T\boldsymbol{X})^{1/2}\boldsymbol{U}\boldsymbol{D}$. Then $\tilde{\boldsymbol{G}}^T\tilde{\boldsymbol{G}} = \boldsymbol{G}$, $\tilde{\boldsymbol{H}}^T\tilde{\boldsymbol{H}} = \boldsymbol{H}$. By construction, $\boldsymbol{U}_{11}^G$ and $\boldsymbol{U}_{22}^H$ must be the right-singular vectors of $\tilde{\boldsymbol{G}}_1 = \tilde{\boldsymbol{G}}[\,,1{:}r]$ and $\tilde{\boldsymbol{H}}_2 = \tilde{\boldsymbol{H}}[\,,(r+1){:}r_0]$ respectively. Denoting by $\boldsymbol{U}_1^{\tilde{G}}$ and $\boldsymbol{U}_2^{\tilde{H}}$ their associated left-singular vectors respectively, we get

$$
\boldsymbol{U}_* = \begin{bmatrix} \boldsymbol{U}_1^{\tilde{G}} & \boldsymbol{U}_2^{\tilde{H}} \end{bmatrix},
$$

which makes both $\boldsymbol{U}_*^T\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T\boldsymbol{U}_*$ and $\boldsymbol{U}_*^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{U}_*$ diagonal. To prove $\boldsymbol{U}_*$ is the desired matrix in (i), it remains to show the orthogonality of $\boldsymbol{U}_*$. It follows from (D.6) and (D.7) that

$$
\begin{aligned}
&(\tilde{\boldsymbol{G}}_1\boldsymbol{U}_{11}^G)^T\tilde{\boldsymbol{H}}_2\boldsymbol{U}_{22}^H \\
&= (\boldsymbol{U}_{11}^G)^T\left((\boldsymbol{X}^T\boldsymbol{X})^{-1/2}\boldsymbol{U}\boldsymbol{D}\begin{bmatrix} \boldsymbol{I}_{r \times r} \\ \boldsymbol{0} \end{bmatrix}\right)^T \\
&\quad \cdot (\boldsymbol{X}^T\boldsymbol{X})^{1/2}\boldsymbol{U}\boldsymbol{D}\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\boldsymbol{U}_{22}^H \\
&= (\boldsymbol{U}_{11}^G)^T\begin{bmatrix} \boldsymbol{I}_{r \times r} & \boldsymbol{0} \end{bmatrix}\boldsymbol{D}\boldsymbol{U}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1/2}(\boldsymbol{X}^T\boldsymbol{X})^{1/2} \\
&\quad \cdot \boldsymbol{U}\boldsymbol{D}\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\boldsymbol{U}_{22}^H \\
&= (\boldsymbol{U}_{11}^G)^T\begin{bmatrix} \boldsymbol{I}_{r \times r} & \boldsymbol{0} \end{bmatrix}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_{(r_0-r) \times (r_0-r)} \end{bmatrix}\boldsymbol{U}_{22}^H
\end{aligned}
$$

$$= (\boldsymbol{U}_{11}^G)^T \begin{bmatrix} \boldsymbol{I}_{r \times r} & \boldsymbol{0} \end{bmatrix} \boldsymbol{D}^2 \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_{(r_0 - r) \times (r_0 - r)} \end{bmatrix} \boldsymbol{U}_{22}^H$$

$$= \boldsymbol{0}.$$

Since $\boldsymbol{G}_{11}$ and $\boldsymbol{H}_{22}$ are positive definite (noting that $\boldsymbol{D} \in \mathbb{R}^{r_0 \times r_0}$ is nonsingular), we further obtain $(\boldsymbol{U}_1^{\tilde{G}})^T \boldsymbol{U}_2^{\bar{H}} = \boldsymbol{0}$. Hence $\boldsymbol{U}_*^T \boldsymbol{U}_* = \boldsymbol{I}$. The proof is now complete.

APPENDIX E. PROOF OF PROPOSITION 2.4

From Theorem 2.1, $\hat{\boldsymbol{B}}$ satisfies

(E.1) $\begin{cases} \hat{\boldsymbol{B}}^\circ = \Theta_{HR}^\sigma(\hat{\boldsymbol{B}}^\circ + \boldsymbol{X}^{\circ T}\boldsymbol{Y} - \boldsymbol{X}^{\circ T}\boldsymbol{\mu}(\hat{\boldsymbol{B}}; \boldsymbol{X}); \lambda, \eta), \\ \boldsymbol{0} = (\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{B}}; \boldsymbol{X}))^T \tilde{\boldsymbol{x}}_0. \end{cases}$

Here, we write $\boldsymbol{\mu}(\hat{\boldsymbol{B}}; \boldsymbol{X})$ to emphasize the dependence of the mean matrix on the design. In this proof, we use the same submatrix notation as in Appendix D.

Given the SVD $\boldsymbol{B}^\circ = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, by Definition 2.2, there exist orthogonal matrices $\bar{\boldsymbol{U}}$ and $\bar{\boldsymbol{V}}$, as augmented versions of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively, i.e., $\boldsymbol{U} = \bar{\boldsymbol{U}}[I,]$ and $\boldsymbol{V} = \bar{\boldsymbol{V}}[I,]$ for some index set I, such that $\hat{\boldsymbol{B}}^\circ = \bar{\boldsymbol{U}}\boldsymbol{\Sigma}\bar{\boldsymbol{V}}^T$ and $\boldsymbol{X}^{\circ T}\boldsymbol{Y} - \boldsymbol{X}^{\circ T}\boldsymbol{\mu}(\hat{\boldsymbol{B}}; \boldsymbol{X}) = \bar{\boldsymbol{U}}\boldsymbol{W}\bar{\boldsymbol{V}}^T$ are both the SVDs. Clearly, $\boldsymbol{\Sigma}[I, I] = \boldsymbol{D}$, $\boldsymbol{\Sigma}[I^c, I^c] = \boldsymbol{0}$. Using the hard-ridge thresholding (2.6), we rewrite the first equation in (E.1) as

(E.2)
$$(1 + \eta)\boldsymbol{B}^\circ + \lambda(1 + \eta)\bar{\boldsymbol{U}}\boldsymbol{S}\bar{\boldsymbol{V}}^T = \boldsymbol{B}^\circ + \boldsymbol{X}^{\circ T}(\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{B}}; \boldsymbol{X})),$$

where $\boldsymbol{S}$ is diagonal and satisfies $\boldsymbol{S}[I, I] = \boldsymbol{0}$ and $\boldsymbol{S}[i, i] \le 1$ for any $i \in I^c$. Left-multiplying both sides of (E.2) by $\bar{\boldsymbol{U}}^T$ yields

$$\eta \boldsymbol{D}\boldsymbol{V}^T = \boldsymbol{U}^T \boldsymbol{X}^{\circ T}(\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{B}}; \boldsymbol{X})).$$

On the other hand, from the construction of $\hat{\boldsymbol{C}}$ and $\boldsymbol{Z}$, it is easy to verify $\boldsymbol{x}_i^T \hat{\boldsymbol{B}} = \boldsymbol{z}_i^T \hat{\boldsymbol{C}}$, from which it follows that $\boldsymbol{\mu}(\hat{\boldsymbol{B}}; \boldsymbol{X}) = \boldsymbol{\mu}(\hat{\boldsymbol{C}}; \boldsymbol{Z})$. Therefore, $\hat{\boldsymbol{C}}$ satisfies

(E.3) $\begin{cases} \eta \boldsymbol{C}^\circ & = \boldsymbol{Z}^{\circ T}(\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{C}}; \boldsymbol{Z})), \\ \boldsymbol{0} & = (\boldsymbol{Y} - \boldsymbol{\mu}(\hat{\boldsymbol{C}}; \boldsymbol{Z}))^T \tilde{\boldsymbol{z}}_0. \end{cases}$

Noticing that the optimization problem in (2.20) is convex and (E.3) gives its KKT equation, the conclusion follows.

APPENDIX F. PROOF OF THEOREM 3.1

Lemma F.1. *Given any* $\boldsymbol{Y} \in \mathbb{R}^{n \times}$, $\hat{\boldsymbol{B}} = \Theta^{\#\sigma}(\boldsymbol{Y}; r, \eta)$ *is a globally optimal solution to*

(F.1) $\quad \min_{\boldsymbol{B}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{B}\|_F^2 + \frac{\eta}{2}\|\boldsymbol{B}\|_F^2 \quad s.t. \quad rank(B) \le r.$

Proof. The problem is equivalent to minimizing

$$\frac{1 + \eta}{2}\sum(\sigma_i^{(\boldsymbol{B})})^2 - \langle \boldsymbol{Y}, \boldsymbol{B} \rangle$$

subject to $rank(\boldsymbol{B}) \le r$. Applying Lemma A.1 yields the result. $\square$

The remainder of the proof follows the same lines to the proof of Theorem 2.1. See Appendix C for details.

ACKNOWLEDGEMENTS

REFERENCES

ANDERSON, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics* **22** 327–351. MR0042664

BRILLINGER, D. R. (1981). *Time Series: Data Analysis and Theory*, expanded edn, San Francisco, CA: Holden-Day. MR0595684

BUNEA, F., SHE, Y. and WEGKAMP, M. (2011) Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics* **39**(2) 1282–1309. MR2816355

CANDÈS, E. J. and PLAN, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory* **57**(4) 2342–2359. MR2809094

CANDÈS, E. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* **9** 717–772. MR2565240

CANDÈS, E. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56**(5) 2053–2080. MR2723472

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika* **95** 759–771. MR2443189

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B* **70**(5) 849–911. MR2530322

FAZEL, M. (2002). Matrix rank minimization with applications, PhD thesis, Stanford University.

GRIGORIEFF, R. D. (1991). A note on von Neumann's trace inequality. *Math. Nachr.* **151** 327–328.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*, New York: Springer-Verlag. MR1121214

HEINEN, A. and RENGIFO, E. (2008). Multivariate reduced rank regression in non-Gaussian contexts, using copulas. *Computational Statistics & Data Analysis* **52**(6) 2931–2944. MR2424769

IZENMAN, A. (2008). *Modern Multivariate. Statistical Techniques: Regression, Classification and Manifold Learning*, New York: Springer.

MA, S., GOLDFARB, D. and CHEN, L. (2009). Fixed point and Bregman iterative methods for matrix rank minimization. arXiv:0905.1643 [math.OC]. MR2445017

OWEN, A. B. (2007). A robust hybrid of lasso and ridge regression. In: *Prediction and Discovery*, Contemporary Mathematics, vol. 443, pp. 59–71. MR2433285

RAO, C. R. (1979). Separation theorems for singular values of matrices and their applications in multivariate analysis. *Journal of Multivariate Analysis* **9**(3) 362–377. MR0548787

REINSEL, G. and VELU, R. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*, New York: Springer. MR1719704

ROBINSON, P. M. (1974). Identification, estimation and large sample theory for regression containing unobservable variables. *International Economic Review* **15** 680–692. MR0356376

SHE, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics* **3** 384–415. MR2501318

SHE, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis* **56** 2976–2990.

TURNBULL, D., BARRINGTON, L., TORRES, D. and LANCKRIET, G. (2008). Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech and Language Processing* **16**(2) 467–476.

VON NEUMANN, J. (1937). Some matrix-inequalities and metrization of matrix-space. *Tomskii Univ. Rev.* **1** 286–300.

YEE, T. W. and HASTIE, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling* **3** 15–41.

YUAN, M., EKICI, A., LU, Z. and MONTEIRO, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B* **69**(3) 329–346. MR1977163

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *JRSSB* **67**(2) 301–320. MR2137327

Yiyuan She
Department of Statistics
Florida State University
Tallahassee, FL 32306-4330
USA
E-mail address: yshe@stat.fsu.edu