Received: 2 March 2020 | Accepted: 16 December 2021

DOI: 10.1111/rssb.12492

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY Series B Statistical Methodology

Supervised multivariate learning with simultaneous feature auto-grouping and dimension reduction

Yiyuan She[1] | **Jiahui Shen**[1] | **Chao Zhang**[2]

[1]Department of Statistics, Florida State University, Tallahassee, USA

[2]Center for Information Science, Peking University, Beijing, China

Correspondence
Yiyuan She, Department of Statistics, Florida State University, Tallahassee, USA.
Email: yshe@stat.fsu.edu

Abstract

Modern high-dimensional methods often adopt the 'bet on sparsity' principle, while in supervised multivariate learning statisticians may face 'dense' problems with a large number of nonzero coefficients. This paper proposes a novel clustered reduced-rank learning (CRL) framework that imposes two joint matrix regularizations to automatically group the features in constructing predictive factors. CRL is more interpretable than low-rank modelling and relaxes the stringent sparsity assumption in variable selection. In this paper, new information-theoretical limits are presented to reveal the intrinsic cost of seeking for clusters, as well as the blessing from dimensionality in multivariate learning. Moreover, an efficient optimization algorithm is developed, which performs subspace learning and clustering with guaranteed convergence. The obtained fixed-point estimators, although not necessarily globally optimal, enjoy the desired statistical accuracy beyond the standard likelihood setup under some regularity conditions. Moreover, a new kind of information criterion, as well as its scale-free form, is proposed for cluster and rank selection, and has a rigorous theoretical support without assuming an infinite sample size. Extensive simulations

and real-data experiments demonstrate the statistical accuracy and interpretability of the proposed method.

KEYWORDS

clustering, information criterion, low-rank matrix estimation, minimax lower bounds, nonasymptotic statistical analysis, nonconvex optimization

1 | INTRODUCTION

Modern statistical applications create an urgent need for analysing and interpreting high-dimensional data with low-dimensional structures. This paper works in a supervised multivariate setting with n samples for m responses and p features (or predictors): $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m] \in \mathbb{R}^{n\times m}$ and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}$. Given a loss l_0, not necessarily a negative log-likelihood function, one can solve the following optimization problem to model the set of responses of interest

$$\min_{\boldsymbol{B}\in\mathbb{R}^{p\times m}} l_0(\boldsymbol{XB}; \boldsymbol{Y}). \tag{1}$$

Here, the unknown coefficient matrix $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p]^T$ has pm unknowns, with $\boldsymbol{b}_j$ summarizing the contributions of the jth predictor to all the responses.

The modern-day challenge comes from large p and/or m. Statisticians often prefer selecting a small subset of features—for example, a group-ℓ_1 penalty $\lambda \sum \|\boldsymbol{b}_j\|_2$ (Yuan & Lin, 2006) can be added in the criterion to promote row-wise sparsity in $\boldsymbol{B}$, which results in a more interpretable model than using an ℓ_2-type penalty $\lambda\|\boldsymbol{B}\|_F^2$. However, with a large m, there may exist few features that are completely irrelevant to the whole set of responses. One may perform variable selection in a transformed space rather than the original space (Johnstone & Lu, 2009), but how to find a proper transformation to reveal sparsity is problem specific.

Perhaps a natural alternative is to make the coefficients form a relatively small number of groups, within each of which all coefficients are forced to be equal. This is referred to as 'equi-sparsity' in She (2010). In the general multivariate setup, instead of requiring a large number of zero rows in the true signal $\boldsymbol{B}^*$, we assume that it has relatively few distinct row patterns $\boldsymbol{b}_{(1)}^{*T}, \boldsymbol{b}_{(2)}^{*T}, \ldots, \boldsymbol{b}_{(q)}^{*T}$. Then, from

$$\boldsymbol{XB}^* = \boldsymbol{x}_1\boldsymbol{b}_1^{*T} + \cdots + \boldsymbol{x}_p\boldsymbol{b}_p^{*T} = \left(\sum_{j\in\mathcal{G}_1}\boldsymbol{x}_j\right)\boldsymbol{b}_{(1)}^{*T} + \cdots + \left(\sum_{j\in\mathcal{G}_q}\boldsymbol{x}_j\right)\boldsymbol{b}_{(q)}^{*T}, \tag{2}$$

the features sharing the same $\boldsymbol{b}_{(k)}^*$ $(1 \le k \le q)$ are automatically grouped, based on their contributions to $\boldsymbol{Y}$. The feature grouping is as interpretable as feature selection and can offer further parsimony, since the latter only targets the set of irrelevant features with $\boldsymbol{b}_j^* = \boldsymbol{0}$.

The problem of how to cluster the unknown coefficient matrix to achieve the best predictability falls into supervised learning, where both $\boldsymbol{Y}$ and $\boldsymbol{X}$ are available. This is in contrast to conventional clustering tasks for unsupervised learning that operate on a single data matrix. But it shares the same computational challenge in large dimensions. A nice but sometimes unnoticed

fact is that if p points form q clusters in a large m-dimensional vector space, then the clusters can be revealed in just a q-dimensional subspace, such as the one spanned by the cluster centroids. In real data analysis, it is not rare that the dimension of the cluster centroid space is much less than q (even as low as 2 or 3). This motivates us to perform simultaneous dimension reduction to ease the job of clustering.

Specifically, we propose to including an additional low-rank constraint, and the resulting jointly regularized form provides an extension of the celebrated reduced rank regression (RRR, Izenman, 1975). RRR assumes that the rank of the true $\boldsymbol{B}^*$ is no more than a small number r, or equivalently, $\boldsymbol{B}^* = \boldsymbol{B}_1\boldsymbol{B}_2^T$ with each $\boldsymbol{B}_i$ having r columns. Once locating a proper loading matrix $\boldsymbol{B}_1$, the final model amounts to fitting $\boldsymbol{Y}$ on r factors formed by $\boldsymbol{XB}_1$. Unfortunately, it is well known that the factor construction from a large number of features lacks interpretability. Our proposal of clustered rank reduction enforces row-wise equisparsity in $\boldsymbol{B}_1$ (or the overall coefficient matrix) so that in extracting r predictive factors, the original features can be automatically consolidated into q groups at the same time.

It is perhaps best to illustrate the idea on a real-world example. The yeast cell cycle data used in Chun and Keleş (2010) studies transcription factors (TFs) related to gene expression over time. In addition to the predictor matrix $\boldsymbol{X} \in \mathbb{R}^{542\times106}$ with 106 TFs collected on 542 genes, a response matrix $\boldsymbol{Y} \in \mathbb{R}^{542\times18}$ containing RNA levels measured on the same genes is available at 18 time points. A naive multivariate regression would have about 2k unknowns, and so we fit an RRR with $r = 2$ and plot the loadings of the 106 TFs in Figure 1. The fact that most of the loadings are apparently nonzero makes variable selection less effective in reducing the complexity of the model. Indeed, because of the high-quality experimental design by biologists, quite a few TFs seem to have effects on the gene expressions during the cell cycle process.

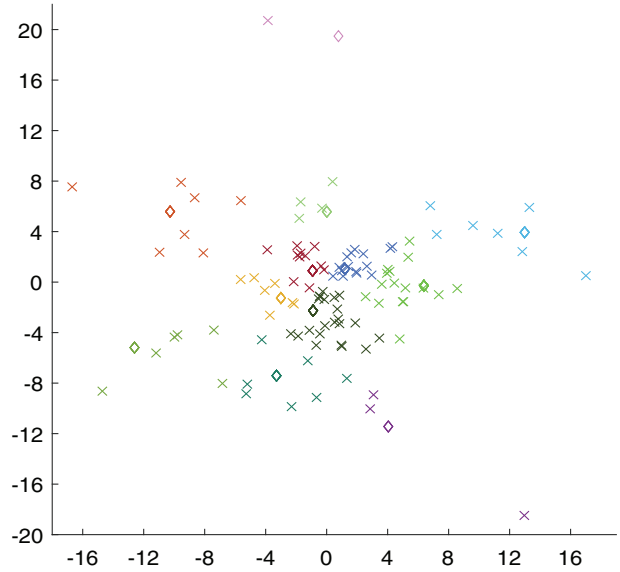

FIGURE 1 Yeast cell cycle data: the loading weights, denoted by '×', by fitting an RRR with $r = 2$. The many noticeable nonzeros imply that a number of transcription factors have effects on the gene expressions during the cell cycle process. Here, the 12 diamonds represent the consolidated loadings obtained by the proposed clustered reduced-rank learning method [Colour figure can be viewed at wileyonlinelibrary.com]

Intuitively, clustering the TFs' loadings would offer a significant reduction of the number of free parameters. Note that to enhance interpretability and avoid ad-hoc tuning, we force the estimates within a group to be equal. A stagewise procedure performing estimation and clustering in two distinct steps would be suboptimal; we aim to solve the problem *as a whole*. The diamonds in Figure 1 show the loadings obtained by the proposed method that simultaneously groups the features in performing dimension reduction.

Compared with the low-rank modelling, the new parsimonious model not only boosted the prediction accuracy by 23% (over 200 repeated training-test splits with 50% for training and 50% for test), but also offered some meaningful TF groups. For example, that ACE2, SWI5 and SOK2 fall into the same group and share the same set of large coefficients provides useful biological insights, as it is well known that ACE2 and SWI5 are paralogs, meaning that they are related to each other through a gene duplication event and are highly conserved in yeast cell cycle gene progression, and according to Pan and Heitman (2000), with regards to nitrogen limitation, SOK2, along with ACE2 and SWI5, is essential in the pseudohyphal growth of yeast cells. Our analysis also provides a cluster with three TFs, namely HIR1, STP2 and SWI4, all of which are chromatin-associated transcription factors involved in regulating the expression of multiple genes at distinct phases of yeast cells (Lambert et al., 2010).

This paper studies simultaneous feature auto-grouping and dimension reduction, and attempts to tackle some related challenges in methodology, theory and computation. Our main contributions are as follows.

- A novel clustered reduced-rank learning (CRL) framework is proposed, which imposes joint matrix regularizations through a convenient SV formulation. It relaxes the assumption of sparsity and offers improved interpretability compared with vanilla low-rank modelling. The concurrent dimension reduction substantially eases the task of clustering in high dimensions.
- Universal information-theoretical limits reveal the intrinsic cost of seeking for clusters, as well as the benefit of accumulating a large number of responses in multivariate learning, which seems to be largely unknown in the literature before.
- Tight error bounds are shown for CRL beyond the standard likelihood setup and justify its minimax optimality in some common scenarios. These nonasymptotic results are strikingly different from those for sparse learning and are the first of their kind. Our theoretical studies favour CRL over variable selection when the numbers of relevant features and irrelevant features are of the same order, or when the number of responses is greater than or equal to the number of features up to a multiplicative constant.
- An efficient optimization-based algorithm is developed, which performs simultaneous subspace pursuit and clustering with guaranteed convergence. The resulting fixed-point estimators, although not necessarily globally optimal, achieve the desired statistical accuracy under some regularity conditions.
- A predictive information criterion is proposed for joint cluster and/or rank selection. Its brand new model complexity notion differs from existing information criteria, but has a rigorous theoretical support in finite samples. A scale-free form is further proposed to bypass the noise scale estimation.

The rest of the paper is organized as follows. Section 2 describes in detail the clustered reduced-rank learning framework to automatically group the predictors in building a predictive low-rank model. Section 3 shows some universal minimax lower bounds and

tight upper bounds of CRL, from which one can conclude that CRL enjoys minimax optimality if the number of clusters is at most polynomially large in the rank. The obtained rates differ substantially from the standard results assuming sparsity, and interestingly, having a large number of responses seems to be a blessing. Section 4 develops an iterative and easy-to-implement algorithm by linearization and block coordinate descent (BCD), where Procrustes rotations and clusterings are performed repeatedly with guaranteed convergence. A new predictive information criterion, together with its scale-free form, is proposed for model selection in the context of clustered rank reduction. Section 5 shows some real data analysis. We conclude in Section 6. The appendices provide all technical details and more computer experiments.

Notation and symbols. The following notation and symbols will be used. Given a differentiable f, we use ∇f to denote its gradient, and f is called μ-strongly convex if $f(\boldsymbol{\eta}') \geq f(\boldsymbol{\eta}) + \langle \nabla f(\boldsymbol{\eta}), \boldsymbol{\eta}' - \boldsymbol{\eta} \rangle + \mu \|\boldsymbol{\eta}' - \boldsymbol{\eta}\|_2^2/2, \forall \boldsymbol{\eta}, \boldsymbol{\eta}'$, and L-strongly smooth if $f(\boldsymbol{\eta}') \leq f(\boldsymbol{\eta}) + \langle \nabla f(\boldsymbol{\eta}), \boldsymbol{\eta}' - \boldsymbol{\eta} \rangle + L\|\boldsymbol{\eta}' - \boldsymbol{\eta}\|_2^2/2, \ \forall \boldsymbol{\eta}, \boldsymbol{\eta}'$. In particular, $f(\boldsymbol{\eta}) = \|\boldsymbol{\eta} - \boldsymbol{y}\|_2^2/2$ is 1-strongly convex. For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times m}$, we denote by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle$ the inner product of $\boldsymbol{A}, \boldsymbol{B}$. Given $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, $\boldsymbol{A}^+$ denotes its Moore–Penrose inverse, $\text{rank}(\boldsymbol{A})$ denotes its rank, and when $n = m$, $\sigma_{\max}(\boldsymbol{A})$ denotes its maximal eigenvalue. We use $\boldsymbol{A}[i, j]$ to represent the (i, j)th element in $\boldsymbol{A}$ and $\boldsymbol{A}[i, :]$ (or $\boldsymbol{A}[:, i]$) to represent the ith row (or column) of $\boldsymbol{A}$. Some conventional matrix norms of $\boldsymbol{A}$ are as follows: $\|\boldsymbol{A}\|_F$ denotes the Frobenius norm, $\|\boldsymbol{A}\|_2$ the spectral norm, $\|\boldsymbol{A}\|_*$ the nuclear norm, and $\|\boldsymbol{A}\|_{2,\infty} = \max_{1 \leq j \leq p} \|\boldsymbol{a}_j\|_2$ for $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p]^T$. The constants denoted by C, c are not necessarily the same at each occurrence. Finally, $a \lesssim b$ means $a \leq cb$ up to a multiplicative positive constant c, and $a \asymp b$ means $a \lesssim b$ and $b \lesssim a$.

2 | CLUSTERED REDUCED RANK REGRESSION

This section focuses on the quadratic loss commonly used in multivariate regression,

$$l_0(\boldsymbol{XB}; \boldsymbol{Y}) = \|\boldsymbol{Y} - \boldsymbol{XB}\|_F^2/2.$$

The discussions in this important case will lay out a foundation for computation and theoretical analysis in later sections regarding a general loss.

Motivated by Section 1, rather than assuming that most features are irrelevant to the responses, we propose to enforce row-wise equisparsity in $\boldsymbol{B}$ so that we can group the features in modelling $\boldsymbol{Y}$. Sparsity is just a special case of equisparsity, and clustering the nonzero values can gain further parsimony. Meanwhile, we would like to regularize the multivariate model with low rank, making it possible to project the data into a much smaller subspace to reveal the row patterns of $\boldsymbol{B}$.

To mathematically formulate the problem, we use $\|\boldsymbol{b}\|_C$ to denote the number of distinct elements in vector $\boldsymbol{b}$, and $\|\boldsymbol{B}\|_{2,C}$ the number of distinct rows of $\boldsymbol{B}$. Then, our clustered reduced-rank learning (CRL) involves the minimization of the loss criterion with two constraints $\text{rank}(\boldsymbol{B}) \leq r, \|\boldsymbol{B}\|_{2,C} \leq q$. The joint regularization formulation poses significant challenges in both computation and theory.

A trick to decouple the two intertwined constraints is to write $\boldsymbol{B} = \boldsymbol{SV}^T$ with $\boldsymbol{V}$ an $m \times r$ column-orthogonal matrix. The 'SV' formulation will be used in optimization as well. Since $\boldsymbol{S} = \boldsymbol{BV}$, we get $\|\boldsymbol{S}\|_{2,C} = \|\boldsymbol{B}\|_{2,C}$ and thus an equivalent CRL problem with separate constraints on $\boldsymbol{S}$ and $\boldsymbol{V}$:

$$\min_{\boldsymbol{S}\in\mathbb{R}^{p\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}} \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{S}\boldsymbol{V}^T\|_F^2 \quad \text{s.t. } \boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I},\ \|\boldsymbol{S}\|_{2,C}\le q. \tag{3}$$

In Equation (3), q ($1\le q\le p$) controls the number of feature groups, and r ($r\le m$) is the dimension of the subspace to pursue coefficient clustering. The joint tuning of the regularization parameters q and r is an important task, too, and a data-adaptive solution with theoretical support will be given in Section 4.2. Let $\boldsymbol{V}=[\boldsymbol{v}_1,\ \dots\ ,\boldsymbol{v}_r]$ and $\boldsymbol{S}=[\boldsymbol{s}_1,\ \dots\ ,\boldsymbol{s}_r]$. We call $\boldsymbol{s}_i$ ($1\le i\le r$) the *clustering vectors*. An intercept term $\mathbf{1}\boldsymbol{\alpha}^T$ can be added to the loss to help control the scale of clustering vectors. Equivalently, for regression, one can simply centre $\boldsymbol{Y},\boldsymbol{X}$ columnwise in advance. We also suggest standardizing the predictors beforehand, as done in other regularization methods like LASSO and ridge regression, unless all the predictors are on the same scale.

We discuss a special case of Equation (3) to provide a more intuitive understanding on the mathematical problem, which can also be used to develop a sequential estimation procedure. Letting $r=1$, Equation (3) becomes $\min_{\boldsymbol{s}\in\mathbb{R}^p,\boldsymbol{v}\in\mathbb{R}^m}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{s}\boldsymbol{v}^T\|_F^2$ s.t. $\|\boldsymbol{v}\|_2^2=1$, $\|\boldsymbol{s}\|_C\le q$. Reparametrize $\boldsymbol{s}=d\boldsymbol{s}^\circ$ with $d\in\mathbb{R}$ and $\boldsymbol{s}^\circ$ satisfying $\|\boldsymbol{X}\boldsymbol{s}^\circ\|_2=1$. Simple algebra shows that the problem is minimized at $d=\langle\boldsymbol{X}\boldsymbol{s}^\circ\boldsymbol{v}^T,\boldsymbol{Y}\rangle$ and $\boldsymbol{v}=\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{s}^\circ/\|\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{s}^\circ\|_2$. Therefore, Equation (3) reduces to (4) when $r=1$:

$$\max_{\boldsymbol{s}^\circ\in\mathbb{R}^p} \boldsymbol{s}^{\circ T}(\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X})\boldsymbol{s}^\circ \text{ s.t. } \boldsymbol{s}^{\circ T}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{s}^\circ=1,\ \|\boldsymbol{s}^\circ\|_C\le q. \tag{4}$$

Without the last constraint, Equation (4) is a *generalized eigenvalue decomposition* problem. The regularization enforces equisparsity in estimating the generalized eigenvector. Although Equation (4) is intuitive, Equation (3) is much more amenable to optimization.

The regularization admits other variants via the SV formulation. For example, with $\boldsymbol{B}=\boldsymbol{S}\boldsymbol{V}^T=\boldsymbol{s}_1\boldsymbol{v}_1^T+\cdots+\boldsymbol{s}_r\boldsymbol{v}_r^T$, one can pursue equisparsity in each component $\boldsymbol{s}_i\boldsymbol{v}_i^T$ or $\boldsymbol{s}_i$:

$$\min_{\boldsymbol{S}\in\mathbb{R}^{p\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}} \left\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{S}\boldsymbol{V}^T\right\|_F^2 \quad \text{s.t. } \boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I},\ \|\boldsymbol{s}_i\|_C\le q^e,\ 1\le i\le r. \tag{5}$$

This rankwise CRL allows each feature to belong to more than one cluster as $r>1$. In comparison, the constraint in Equation (3) offers a uniform control. Unless otherwise mentioned, we will focus on the row-wise problem (3), but our algorithm applies to both.

Remark 1 *Alternative formulations of CRL*. Given a positive definite matrix $\boldsymbol{\Gamma}$ of size $m\times m$, consider a weighted criterion: $\min_{(\boldsymbol{S},\boldsymbol{V})\in\mathbb{R}^{p\times r}\times\mathbb{R}^{m\times r}}\mathrm{Tr}\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{S}\boldsymbol{V}^T)\boldsymbol{\Gamma}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{S}\boldsymbol{V}^T)^T\}$ s.t. $\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}$, $\|\boldsymbol{S}\|_{2,C}\le q$. Then, for $\boldsymbol{B}=\boldsymbol{S}\boldsymbol{V}^T\boldsymbol{\Gamma}^{1/2}$, we have $\mathrm{rank}(\boldsymbol{B})\le r$ and $\|\boldsymbol{B}\|_{2,C}=\|\boldsymbol{S}\|_{2,C}$. Applying the SV representation to $\boldsymbol{B}$ gives an equivalent problem (with $\boldsymbol{S}$, $\boldsymbol{V}$ redefined)

$$\min_{(\boldsymbol{S},\boldsymbol{V})\in\mathbb{R}^{p\times r}\times\mathbb{R}^{m\times r}} \frac{1}{2}\|\boldsymbol{Y}\boldsymbol{\Gamma}^{1/2}-\boldsymbol{X}\boldsymbol{S}\boldsymbol{V}^T\|_F^2 \text{ s.t. } \boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I},\|\boldsymbol{S}\|_{2,C}\le q. \tag{6}$$

Equation (6) is of the same form of Equation (3) with an adjusted response matrix.

Another related *projected* form directly measures discrepancy in the projected space:

$$\min_{(\boldsymbol{S},\boldsymbol{A})\in\mathbb{R}^{p\times r}\times\mathbb{R}^{m\times r}} \frac{1}{2}\|\boldsymbol{Y}\boldsymbol{A}-\boldsymbol{X}\boldsymbol{S}\|_F^2 \text{ s.t. } (\boldsymbol{Y}\boldsymbol{A})^T\boldsymbol{Y}\boldsymbol{A}=n\boldsymbol{I},\|\boldsymbol{S}\|_{2,C}\le q, \tag{7}$$

where we assume $r\le\mathrm{rank}(\boldsymbol{Y})$. Let $\boldsymbol{\Sigma}_Y=\boldsymbol{Y}^T\boldsymbol{Y}/n=\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ with the diagonal matrix $\boldsymbol{D}$ containing $\mathrm{rank}(\boldsymbol{Y})$ nonzero eigenvalues, $\boldsymbol{W}=\boldsymbol{D}^{1/2}\boldsymbol{U}^T\boldsymbol{A}$, and $\boldsymbol{B}=\boldsymbol{S}\boldsymbol{W}^T\boldsymbol{D}^{1/2}\boldsymbol{U}^T$. Because

$\|\boldsymbol{YA}-\boldsymbol{XS}\|_F^2 = \mathrm{Tr}\{(\boldsymbol{Y}-\boldsymbol{XB})\boldsymbol{\Sigma}_Y^+(\boldsymbol{Y}-\boldsymbol{XB})^T\} + nr - n\,\mathrm{rank}(\boldsymbol{Y})$ (cf. Lemma A.8), $\mathrm{rank}(\boldsymbol{B}) \le r$, and $\boldsymbol{B}$, $\boldsymbol{S}$ have the same row patterns, we see that the projected form simply amounts to taking $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_Y^+$. Popular choices of $\boldsymbol{\Gamma}$ are based on the covariance matrix of $\boldsymbol{Y}$. See She et al. (2020) for a proposal to account for dependence in the case of a general loss. A further topic is to estimate the high-dimensional covariance matrix and mean matrix jointly, but it is beyond the scope of the current paper and we regard $\boldsymbol{\Gamma}$ as known. Then, based on Equation (6), one simply needs to 'whiten' $\boldsymbol{Y}$ by $\boldsymbol{Y\Gamma}^{\frac{1}{2}}$ beforehand.

Remark 2 *Pairwise-difference penalization*. An alternative idea, following She (2010) and Chi and Lange (2015), is to penalize the pairwise row-differences of $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p]^T$:

$$\sum_{1\le j<j'\le p} P(\|\boldsymbol{b}_j - \boldsymbol{b}_{j'}\|_2; \lambda),$$

where P is a sparsity-inducing function. This type of regularization is, however, not of our primary interest, due to its computational burden, suboptimal error rate and difficulties in parameter tuning. See Appendix A.5 and Theorem A.3 for more details.

Remark 3 *Unsupervised learning*. Supervised learning is the focus of our work, but when there is a single data matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$, we can set $\boldsymbol{X} = \boldsymbol{I}$ in Equation (3) to cluster its rows for unsupervised learning. (Substituting $\boldsymbol{Y}^T$ for $\boldsymbol{Y}$ offers clustered PCA as an alternative to sparse PCA.) Similar to the derivation in the rank-1 case, we can evaluate the optimal $\boldsymbol{V}$ to get $\min_{\boldsymbol{S}\in\mathbb{R}^{n\times r}} \|\boldsymbol{S}\|_F^2/2 - \|\boldsymbol{Y}^T\boldsymbol{S}\|_*$ s.t. $\|\boldsymbol{S}\|_{2,C} \le q$, or $\max_{\|\boldsymbol{S}^\circ\|_F=1} \|\boldsymbol{Y}^T\boldsymbol{S}^\circ\|_*^2/2$ s.t. $\|\boldsymbol{S}^\circ\|_{2,C} \le q$ via $\boldsymbol{S} = d\boldsymbol{S}^\circ$. (In the more general supervised setup, we can show that $\boldsymbol{S}$ solves $\min_{\boldsymbol{S}\in\mathbb{R}^{p\times r}} \|\boldsymbol{XS}\|_F^2/2 - \|\boldsymbol{Y}^T\boldsymbol{XS}\|_*$ s.t. $\|\boldsymbol{S}\|_{2,C} \le q$.) Because $\|\boldsymbol{Y}^T\boldsymbol{S}\|_* = \mathrm{Tr}\{(\boldsymbol{S}^T\boldsymbol{YY}^T\boldsymbol{S})^{1/2}\}$, CRL's clustering vectors depend on $\boldsymbol{Y}$ through its sample inner products only. One can then introduce a *kernel* CRL by substituting a positive semi-definite $\boldsymbol{K}$ for $\boldsymbol{YY}^T$. The desired clusters can still be obtained by solving the SV-form problem, with a suitable pseudo-response constructed from the kernel matrix. Let us consider two special cases to contrast the unsupervised CRL with some related methods. (a) No data projection, that is, $r = m$. Then we can show that K-means is an algorithm to solve the problem (cf. Section 4.1). Modern implementations of K-means make good use of seeding and can obtain a decent solution in low dimensions, which will assist the optimization of CRL, owing to its low-rank nature. Of course, as K-means operates in the input space, it can be ineffective for large m. (b) No equisparsity regularization. In this case, CRL reduces to spectral clustering (cf. Appendix B.1). Giving up the equisparsity regularization simplifies the computation significantly, but spectral clustering, as well as other similarity-motivated procedures, performs dimension reduction and clustering in two distinct steps. It would be less greedy to perform both steps simultaneously. In Section 4, we will see that the CRL algorithm can integrate clustering and subspace learning to solve the problem as a whole.

3 | NONASYMPTOTIC STATISTICAL ANALYSIS OF CLUSTERED REDUCED-RANK LEARNING

Rigorous theoretical guarantees must be provided to justify the proposed clustered rank reduction method. There is a big literature gap to fill in this regard. For example, how many samples are

needed for signal recovery by adopting the new notion of structural parsimony? In which situations will pursuing equisparsity be advantageous over performing variable selection? Is it always necessary to obtain a globally optimal solution in the nonconvex setup? The answers to these questions seem to be largely unknown.

In this section, we go beyond the regression setup and consider a loss $l_0(\boldsymbol{XB}; \boldsymbol{Y})$ that is defined on the systematic component $\boldsymbol{XB}$ with $\boldsymbol{Y}$ as parameters. Here, $\boldsymbol{B}$ is unknown and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are observed, and so we occasionally omit the dependence on $\boldsymbol{Y}$. Assume that l_0 is differentiable with respect to $\boldsymbol{XB}$. Our tool for tackling a general loss is the *generalized Bregman function* (She et al., 2021): given a differentiable function ψ,

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \psi(\boldsymbol{\alpha}) - \psi(\boldsymbol{\beta}) - \langle \nabla\psi(\boldsymbol{\beta}), \boldsymbol{\alpha} - \boldsymbol{\beta} \rangle. \tag{8}$$

If ψ is also strictly convex, $\boldsymbol{\Delta}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta})$ becomes the standard Bregman divergence denoted by $\mathbf{D}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta})$ (Bregman, 1967). A simple example is $\mathbf{D}_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2^2/2$, associated with $\psi = \|\cdot\|_2^2/2$, and its matrix version is $\mathbf{D}_2(\boldsymbol{A}, \boldsymbol{B}) = \|\mathrm{vec}(\boldsymbol{A}) - \mathrm{vec}(\boldsymbol{B})\|_2^2/2 = \|\boldsymbol{A} - \boldsymbol{B}\|_F^2/2$. In general, $\boldsymbol{\Delta}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta})$ may not be symmetric, and we define its symmetrized version by $\overline{\boldsymbol{\Delta}}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) := (\boldsymbol{\Delta}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}))/2$.

Introducing the notion of noise in the non-likelihood setup is another essential component, since l_0 may not correspond to a distribution function. We define the *effective noise* associated with the statistical truth $\boldsymbol{B}^*$ by

$$\boldsymbol{E} = -\nabla l_0(\boldsymbol{XB}^*; \boldsymbol{Y}). \tag{9}$$

So having a zero-mean noise means that the risk vanishes at the statistical truth, assuming we can exchange the gradient and expectation. In a canonical generalized linear model (GLM) with cumulant function $b(\cdot)$ and $g = (\nabla b)^{-1}$ as the canonical link (cf. Appendix A), the (unscaled) loss can be represented by $-\langle \boldsymbol{Y}, \boldsymbol{XB} \rangle + b(\boldsymbol{XB})$, and by matrix differentiation,

$$\boldsymbol{E} = \boldsymbol{Y} - g^{-1}(\boldsymbol{XB}^*) = \boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y}),$$

or $\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{XB}^*$ in regression. Unless otherwise specified, we assume that $\mathrm{vec}(\boldsymbol{E})$ is a *sub-Gaussian* random vector with mean zero and scale bounded by σ (namely, all marginals $\langle \mathrm{vec}(\boldsymbol{E}), \boldsymbol{\alpha} \rangle$ satisfy $\|\langle \mathrm{vec}(\boldsymbol{E}), \boldsymbol{\alpha} \rangle\|_{\psi_2} \le \sigma \|\boldsymbol{\alpha}\|_2$, $\forall \boldsymbol{\alpha} \in \mathbb{R}^p$, where $\|\cdot\|_{\psi_2} = \inf\{t > 0 : \mathbb{E} \exp[(\cdot/t)^2] \le 2\}$, cf. van der Vaart & Wellner, 1996). Note that the components of $\boldsymbol{E}$ need not be independent. Sub-Gaussian noises are typical in regression and classification problems, since Gaussian and bounded random variables are sub-Gaussian. Yet sub-Gaussianity is not critical for our analysis.

This section focuses on row-wise equisparsity $\|\boldsymbol{B}\|_{2,C}$. It turns out that the two measures $\|\cdot\|_{2,C}$ and $\mathrm{rank}(\cdot)$ can effectively bound the stochastic terms arising from CRL. Because it is be difficult in present-day applications to tell whether the sample size, relative to the problem dimensions, is large enough to apply asymptotics, all of our investigations will be nonasymptotic.

3.1 | Universal minimax lower bounds

The first question one must answer is how small the error could be under equisparsity with possibly low rank. We derive new minimax lower bounds to address the question. Let $I(\cdot)$ be an

arbitrary nondecreasing function with $I(0) = 0$, $I \not\equiv 0$; some particular examples are $I(t) = t$ and $I(t) = 1_{t \geq c}$.

Theorem 1 *Assume $\boldsymbol{Y}|\boldsymbol{XB}^*$ follows a distribution in the regular exponential family with dispersion σ^2 with l_0 the associated negative log-likelihood function (cf. Appendix A for details). Define a signal class by*

$$\boldsymbol{B}^* \in \mathcal{S}(q, r) = \{\boldsymbol{B} \in \mathbb{R}^{p \times m} : \|\boldsymbol{B}\|_{2,C} \leq q, \ \operatorname{rank}(\boldsymbol{B}) \leq r\}, \tag{10}$$

where $p \geq q \geq r \geq 2$, $r(q \wedge \operatorname{rank}(\boldsymbol{X}) + m - r) \geq 4$. Let b, ζ be any integers satisfying

$$\sum_{i=0}^{\zeta} \binom{r}{i} (b-1)^i \geq q, \tag{11}$$

with $b \geq 2$, $1 \leq \zeta \leq r$, and define a complexity function

$$P(q, r) = (q + m)r + p\{\log(er) - \log\log q\}. \tag{12}$$

Assume for some $\kappa > 0$

$$\boldsymbol{\Delta}_{l_0}(\boldsymbol{0}, \boldsymbol{XB})\sigma^2 \leq \kappa \|\boldsymbol{B}\|_F^2/2, \quad \forall \boldsymbol{B} \in \mathcal{S}(q, r). \tag{13}$$

Then there exist positive constants c, c', depending on $I(\cdot)$ only, such that

$$\inf_{\hat{\boldsymbol{B}}} \sup_{\boldsymbol{B}^* \in \mathcal{S}(q,r)} \mathbb{E}\left\{ I\left(\|\boldsymbol{B}^* - \hat{\boldsymbol{B}}\|_F^2 \Big/ \left[c\sigma^2 \left\{ (q+m)r + \frac{p \log q}{b^2 \zeta} \right\} \Big/ \kappa \right] \right) \right\} \geq c' > 0, \tag{14}$$

where $\hat{\boldsymbol{B}}$ denotes an arbitrary estimator. In particular, under $8 \leq q \leq \exp(r)$,

$$\inf_{\hat{\boldsymbol{B}}} \sup_{\boldsymbol{B}^* \in \mathcal{S}(q,r)} \mathbb{E}\left[I\left(\|\boldsymbol{B}^* - \hat{\boldsymbol{B}}\|_F^2 / \{c\sigma^2 P(q, r)/\kappa\} \right) \right] \geq c' > 0. \tag{15}$$

A more complete theorem including minimax lower bounds for both the estimation error $\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F$ and prediction error $\|\boldsymbol{X}\hat{\boldsymbol{B}} - \boldsymbol{XB}^*\|_F$ is presented in Appendix A.1, from which one can see that the size of $\phi := q^{1/r}$ plays a vital role in determining the final error rate. These results are the first of their kind and provide useful guidance in the context of equisparsity. Our proof is nontrivial and makes use of q-ary codes in information theory (Pellikaan et al., 2017), as well as some useful facts of the generalized Bregman functions for GLMs.

The regularity condition (13) is not restrictive. For regression and logistic regression, the condition is implied by $\boldsymbol{\Delta}_{l_0}(\boldsymbol{0}, \boldsymbol{XB})\sigma^2 \leq \|\boldsymbol{XB}\|_F^2/2$ and $\boldsymbol{\Delta}_{l_0}(\boldsymbol{0}, \boldsymbol{XB}) \leq \|\boldsymbol{XB}\|_F^2/8$ respectively. The bound in Equation (14) is general, while Equation (15) is perhaps more illustrative: when $q \leq \exp(r)$ and $\kappa \leq c\, n$,

$$\mathbb{E}\left[\|\boldsymbol{B}^* - \hat{\boldsymbol{B}}\|_F^2 \right] \geq c\sigma^2 P(q, r)/n, \ \text{ and } \ \mathbb{P}\left[\|\boldsymbol{B}^* - \hat{\boldsymbol{B}}\|_F^2 \geq c\sigma^2 P(q, r)/n \right] > c_0 > 0,$$

by setting $I(t) = t$ and $I(t) = 1_{t \geq c}$ respectively. Therefore, in the scenario of q being polynomially large in r, that is, $q \leq r^c$ for some constant c, no estimator can beat the error rate

$P(q, r) \asymp (q + m)r + p \log q$ in a minimax sense. Interestingly, when m is a constant, the rate reduction compared to pm is not significant, whereas having a large number of response variables is (perhaps surprisingly) a blessing for pursuing equisparsity.

3.2 | Upper error bounds of CRL

Can we approach the optimal error rate using a particular estimator? This part shows that CRL is a legitimate method, and more importantly, pursuing its globally optimal solutions is unnecessary in many cases. Rather, finding a *fixed point* of CRL, defined by Equations (16) and (17) below, would suffice for regular problems.

Let $r^* = \text{rank}(\boldsymbol{B}^*)$ and $q^* = \|\boldsymbol{B}^*\|_{2,C}$. Given a differentiable l_0, define

$$G_\rho(\boldsymbol{B}; \boldsymbol{B}^-) = l_0(\boldsymbol{XB}; \boldsymbol{Y}) - \boldsymbol{\Delta}_{l_0}(\boldsymbol{XB}, \boldsymbol{XB}^-) + \rho \mathbf{D}_2(\boldsymbol{B}, \boldsymbol{B}^-), \tag{16}$$

where ρ, representing the inverse stepsize, is an algorithm parameter to be chosen. Then, for all fixed points defined by

$$\hat{\boldsymbol{B}} \in \underset{\boldsymbol{B}:\, \|\boldsymbol{B}\|_{2,C} \le q, \text{rank}(\boldsymbol{B}) \le r}{\text{argmin}} G_\rho(\boldsymbol{B}; \boldsymbol{B}^-)|_{\boldsymbol{B}^- = \hat{\boldsymbol{B}}}, \tag{17}$$

a nonasymptotic error bound can be derived by calculating the metric entropy of the associated manifolds and using the Stirling numbers of the second kind.

Theorem 2 *Let $r \ge r^*$, $q \ge q^*$ and $\hat{\boldsymbol{B}}$ be any fixed point satisfying Equation (17) for some $\rho > 0$. Define*

$$P_o(q, r) = \{q \wedge \text{rank}(\boldsymbol{X}) + m\}r + (p - q) \log q. \tag{18}$$

Assume $\rho > 0$ is chosen so that

$$\rho \mathbf{D}_2(\boldsymbol{B}_1, \boldsymbol{B}_2) \le (2\overline{\boldsymbol{\Delta}}_{l_0} - \delta \mathbf{D}_2)(\boldsymbol{XB}_1, \boldsymbol{XB}_2) + K\sigma^2 P_o(q, r), \ \forall \boldsymbol{B}_i\colon \text{rank}(\boldsymbol{B}_i) \le r, \|\boldsymbol{B}_i\|_{2,C} \le q \tag{19}$$

for some $\delta > 0$ and sufficiently large $K \ge 0$. Then, $\hat{\boldsymbol{B}}$ satisfies

$$\mathbb{E}\big[\|\boldsymbol{X}\hat{\boldsymbol{B}} - \boldsymbol{XB}^*\|_F^2\big] \lesssim \frac{K\delta \vee 1}{\delta^2}\{\sigma^2(q \wedge \text{rank}(\boldsymbol{X}) + m)r + \sigma^2(p - q) \log q + \sigma^2\}. \tag{20}$$

It is not difficult to see that when l_0 is μ-strongly convex, the following matrix restricted eigenvalue condition implies Equation (19) with $K = 0$:

$$\rho\|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_F^2 \le (2\mu - \delta)\|\boldsymbol{X}(\boldsymbol{B}_1 - \boldsymbol{B}_2)\|_F^2, \quad \forall \boldsymbol{B}_i\colon \text{rank}(\boldsymbol{B}_i) \le r, \|\boldsymbol{B}_i\|_{2,C} \le q. \tag{21}$$

When q is small, Equation (21) is applicable to large-p designs. Similar regularity conditions are widely used in compressed sensing, variable selection and low rank estimation (Bickel et al., 2009; Candès & Plan, 2011; Candès & Tao, 2007).

Let $q = \vartheta q^*$, $r = \vartheta r^*$ with $\vartheta \ge 1$. When ϑ, δ and K are treated as constants and $\sigma = 1$, from Equation (20), the prediction error bound is of the order

$$\{q^* \wedge \text{rank}(\boldsymbol{X}) + m\}r^* + (p - q^*) \log q^*, \tag{22}$$

ignoring all trivial multiplicative/additive terms. The rate distinguishes CRL from various sparse learning methods in the literature.

Remark 4 *Computational feasibility*. The fact that the fixed-point solutions, although not necessarily globally or even locally optimal, can have provable guarantees offers a feasible computation of the nonconvex CRL optimization problem in regular cases.

Specifically, regardless of the choice of the loss, G_ρ always has a simple quadratic form in terms of $\boldsymbol{B}$, which gives rise to an iterative update of the coefficient matrix. Similar results can be shown for $(\boldsymbol{S}, \boldsymbol{V})$ obtained by alternative optimization; see Theorem A.2 in Remark A.2. Section 4.1 designs an efficient algorithm on the basis of linearization and BCD.

Remark 5 *Error rate comparison*. To clarify the theoretical meaning of Equation (22), we make an error-rate comparison between CRL and some commonly used estimators in regression with $\sigma = 1$ (assuming all regularity conditions are met). First, assuming that $\boldsymbol{X}$ has full column rank, the ordinary least squares has $\mathbb{E}\|\boldsymbol{X}\hat{\boldsymbol{B}} - \boldsymbol{X}\boldsymbol{B}^*\|_F^2 = mp$. With no rank reduction ($r^* = q^*$), Equation (22) gives $(q^* + m)q^* + (p - q^*)\log q^*$, which is $\lesssim mp$ when the number of responses is larger than the number of feature groups. Of course, if $r^* < q^*$, CRL can achieve a much lower error rate. Comparing Equation (22) with $(p + m)r^*$ by low-rank matrix estimation (Bunea et al., 2011), we see that CRL does a substantially better job if the number of clusters does not grow exponentially with the rank, namely, $q^* \ll \exp(r^*)$.

Variable selection gives another important means of regularization. If $\boldsymbol{B}^* = [\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_p^*]^T$ is row-wise sparse with $s^* = |\{j : \boldsymbol{b}_j^* \neq \boldsymbol{0}\}|$, the prediction error by means of variable selection is of the order (Lounici et al., 2011)

$$s^* m + s^* \log p. \tag{23}$$

The comparison between Equations (23) and (22) shows no clear winner: for the degrees-of-freedom terms, $(\mathrm{rank}(\boldsymbol{X}) \wedge q^* + m)r^* \leq q^* r^* + ms^* \lesssim s^* m$, while for the 'inflation' terms, $(p - q^*)\log q^*$ is typically larger than $s^* \log p$. But two scenarios draw our particular attention:

$$\text{(i) ``many responses'': } m \geq cp \quad \text{(ii) ``linear sparsity'': } s^* = cp$$

where c is a positive constant. In either situation, CRL is advantageous over variable selection.

Concretely, in case (i), $s^* m + s^* \log p \asymp s^* m$, while from $s^* \geq q^* \geq r^*$, we have $s^* m \geq r^* m \geq r^* q^*$ and $s^* m \gg (\log q^*)(p - q^*)$, and so $s^* m + s^* \log p \gtrsim \{q^* \wedge \mathrm{rank}(\boldsymbol{X}) + m\}r^* + (p - q^*)\log q^*$. In case (ii), $s^* \log p \asymp p \log p \geq (p - q^*)\log q^*$ and the same conclusion holds. In other words, when the number of responses is greater than the number of features up to a multiplicative constant, or when the number of relevant features and the number of irrelevant features are of the same order, CRL has a lower error rate with rigorous theoretical support.

Remark 6 *Minimax optimality*. One may be curious if the upper error bound of CRL could match the universal minimax lower bound. Notably, under the mild conditions $q^* \leq \mathrm{rank}(\boldsymbol{X})$ and $q^* \ll \exp(r^*)$ (and so $q^* \log q^* \ll q^* r^*$), Equation (22) becomes

$$(q^* + m)r^* + p \log q^*,$$

which is exactly the rate shown at the end of in Section 3.1. So at least for canonical GLMs with q^* polynomially large in r^*, CRL does enjoy minimax rate optimality.

Of course, the previous discussions assume that q and r are specified so that they are not too large relative to q^* and r^*, respectively. In general, the data-adaptive tuning to be introduced in Section 4.2 still ensures Equation (22).

To the best of our knowledge, Theorem 3 is the first nonasymptotic statistical analysis of the set of CRL's fixed points in nonconvex optimization. One might ask whether the error rate can be further improved by pursuing a global CRL estimator. The following theorem shows that this is not the case, but the regularity condition (19) gets relaxed to some extent.

Theorem 3 *Let $\hat{\boldsymbol{B}}$ be an optimal CRL solution with $r \geq r^*$ and $q \geq q^*$. Assume that there exists some $\delta > 0$ such that*

$$\boldsymbol{\Delta}_{l_0}(\boldsymbol{XB}_1, \boldsymbol{XB}_2) \geq \delta \mathbf{D}_2(\boldsymbol{XB}_1, \boldsymbol{XB}_2), \quad \forall \boldsymbol{B}_i\colon \|\boldsymbol{B}_i\|_{2,C} \leq q, \ \operatorname{rank}(\boldsymbol{B}_i) \leq r. \tag{24}$$

Then $\mathbb{E}[\|\boldsymbol{X}\hat{\boldsymbol{B}} - \boldsymbol{XB}^*\|_F^2] \lesssim \frac{1}{\delta^2}\{\sigma^2(q \wedge \operatorname{rank}(\boldsymbol{X}) + m)r + \sigma^2(p-q)\log q + \sigma^2\}$.

Unlike Equation (19) in Theorem 2, Equation (24) uses $\boldsymbol{\Delta}_{l_0}$ (in place of twice of its symmetrized version) and does not involve ρ. The conclusion of Theorem 3 can be extended to an oracle inequality (Donoho & Johnstone, 1994), and these ℓ_2-recovery results can be used to give some estimation error bounds under proper regularity conditions. Corollary 1 gives an illustration.

Corollary 1 *Let l_0 be μ-strongly convex. Then for any $\boldsymbol{B}$: $\operatorname{rank}(\boldsymbol{B}) \leq r, \|\boldsymbol{B}\|_{2,C} \leq q$,*

$$\mathbb{E}\mathbf{D}_{l_0}(\boldsymbol{X}\hat{\boldsymbol{B}}, \boldsymbol{XB}^*)\|_F^2 \lesssim \mathbb{E}\mathbf{D}_{l_0}(\boldsymbol{XB}, \boldsymbol{XB}^*) + \frac{\sigma^2}{\mu}\{(q \wedge \operatorname{rank}(\boldsymbol{X}) + m)r + (p-q)\log q\} + \frac{\sigma^2}{\mu}. \tag{25}$$

Furthermore, assume $\|\boldsymbol{XB}\|_F^2/n \geq \delta\|\boldsymbol{B}\|_{2,\infty}^2, \forall \boldsymbol{B}$: $\operatorname{rank}(\boldsymbol{B}) \leq (1+\vartheta)r^*, \|\boldsymbol{B}\|_{2,C} \leq \vartheta q^{*2}$ for some $\delta > 0$, and $r = \vartheta r^*, q = \vartheta q^*$, $\vartheta \geq 1$, $q^* > 1$. Then with probability at least $1 - C\exp\{-c(m + \operatorname{rank}(\boldsymbol{X}))\}$,

$$\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,\infty}^2 \leq \frac{c_0\vartheta^2\sigma^2}{n\delta\mu^2}\{(q^* \wedge \operatorname{rank}(\boldsymbol{X}) + m)r^* + (p-q^*)\log q^*\} \tag{26}$$

for some constants $c_0, c, C > 0$.

The RHS of Equation (25) offers a bias-variance tradeoff and as a result, CRL applies to $\boldsymbol{B}^*$ with just *approximate* equisparsity and/or low rank. The $(2, \infty)$-norm error bound Equation (26) implies faithful cluster recovery with high probability, if the signal-to-noise ratio is properly large: $\min_{b_{j\cdot}^* \neq b_{k\cdot}^*} |\operatorname{avg}_l b_{jl}^{*2} - \operatorname{avg}_l b_{kl}^{*2}|/\sigma^2 > 2\zeta$. Here, $\operatorname{avg}_l b_{jl}^{*2}$ is the average of $b_{j1}^{*2}, \ldots, b_{jm}^{*2}$ and $\zeta = \frac{c_0\vartheta^2}{\delta\mu^2}\{\frac{(p-q^*)\log q^*}{nm} + \frac{r^*(q^* \wedge \operatorname{rank}(\boldsymbol{X}))}{nm} + \frac{r^*}{n}\}$. Then, from the bound on $\|\hat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,\infty}^2$, we know that for $q = q^*$, $\hat{\boldsymbol{B}}$ exhibits the same row clusters as does $\boldsymbol{B}^*$, and for $q > q^*$, $\hat{\boldsymbol{B}}$ refines the clustering structure of $\boldsymbol{B}^*$.

4 | COMPUTATION AND TUNING

In this section, we develop an efficient optimization-based CRL algorithm with implementation ease and guaranteed convergence. We also propose a novel model comparison criterion and provide its theoretical justification from a predictive learning perspective, without assuming any infinite sample-size or large signal-to-noise ratio conditions.

4.1 | Algorithm design

In this part, we discuss how to solve the CRL problem with q and r fixed. CRL poses some intriguing challenges in optimization: the problem is highly nonconvex, l_0 is not restricted to the quadratic loss or a negative log-likelihood function, and the equisparsity and low-rank constraints are of discrete nature. These obstacles render standard algorithms inapplicable. In addition, as p and m may be large in real applications, the coefficient matrix $\boldsymbol{B} \in \mathbb{R}^{p\times m}$ can easily contain an overwhelming number of unknowns. Then, how to make use of its low-rank nature to reduce the computational cost at each iteration, while maintaining the convergence of the overall procedure, is crucial in large-scale computation.

Before describing the algorithm design in full detail, we provide below a simplified version of the algorithm.

Algorithm 1 Clustered Reduced-rank Learning Algorithm

Require: $(\boldsymbol{Y}, \boldsymbol{X}) \in \mathbb{R}^{n\times m} \times \mathbb{R}^{n\times p}$, l_0: a loss function with ∇l_0 L-Lipschitz continuous; q: desired number of clusters; ϵ: error tolerance; M: maximum number of iterations; a feasible starting point: $\boldsymbol{F}^0, \boldsymbol{\mu}^0, \boldsymbol{V}^{[0]}, \boldsymbol{\alpha}^{[0]}$; ρ: inverse stepsize (e.g., $L\|\boldsymbol{X}\|_2^2$)

$k \leftarrow 0$, $\boldsymbol{S}^{[0]} = \boldsymbol{F}^0\boldsymbol{\mu}^0$;
while $k < M$ **and** $\epsilon < \|\boldsymbol{B}^{[k]} - \boldsymbol{B}^{[k-1]}\|$ (if existing) **do**
  $k \leftarrow k + 1$;
  $\boldsymbol{B}^{[k-1]} = \boldsymbol{S}^{[k-1]}(\boldsymbol{V}^{[k-1]})^T$, $\widetilde{\boldsymbol{Y}} \leftarrow \boldsymbol{B}^{[k-1]} - \boldsymbol{X}^T\nabla l_0(\mathbf{1}\boldsymbol{\alpha}^{[k-1]} + \boldsymbol{X}\boldsymbol{B}^{[k-1]})/\rho$;
  $\boldsymbol{\alpha}^{[k]} \leftarrow \boldsymbol{\alpha}^{[k-1]} - [\nabla l_0(\mathbf{1}\boldsymbol{\alpha}^{[k-1]} + \boldsymbol{X}\boldsymbol{B}^{[k-1]})]^T\mathbf{1}/\rho$;
  $t \leftarrow 0, \boldsymbol{S}^{(0)} \leftarrow \boldsymbol{S}^{[k-1]}, \boldsymbol{V}^{(0)} \leftarrow \boldsymbol{V}^{[k-1]}, \boldsymbol{B}^{(0)} \leftarrow \boldsymbol{S}^{(0)}\boldsymbol{V}^{(0)T}$;
  while not converged **do**
    $t \leftarrow t + 1$;
    $\boldsymbol{W} \leftarrow \widetilde{\boldsymbol{Y}}^T\boldsymbol{S}^{(t-1)}$;
    $\boldsymbol{V}^{(t)} \leftarrow \boldsymbol{U}_w\boldsymbol{V}_w^T$ with $\boldsymbol{U}_w$, $\boldsymbol{V}_w$ from the SVD $\boldsymbol{W} = \boldsymbol{U}_w\boldsymbol{D}_w\boldsymbol{V}_w^T$;
    $\boldsymbol{L} \leftarrow \widetilde{\boldsymbol{Y}}\boldsymbol{V}^{(t)}$;
    With $\boldsymbol{\mu}^0$ as the initial cluster centroid matrix, call K-means on $\boldsymbol{L}$ to update $\boldsymbol{F}^l$ and $\boldsymbol{\mu}^l$ ($l \geq 1$) alternatively till convergence;
    $\boldsymbol{S}^{(t)} \leftarrow \boldsymbol{F}^l\boldsymbol{\mu}^l$, $\boldsymbol{B}^{(t)} = \boldsymbol{S}^{(t)}(\boldsymbol{V}^{(t)})^T$, $\boldsymbol{\mu}^0 \leftarrow \boldsymbol{\mu}^l$;
  end while
  $\boldsymbol{S}^{[k]} \leftarrow \boldsymbol{S}^{(t)}, \boldsymbol{V}^{[k]} \leftarrow \boldsymbol{V}^{(t)}, \boldsymbol{B}^{[k]} \leftarrow \boldsymbol{S}^{[k]}(\boldsymbol{V}^{[k]})^T$;
end while
return $\boldsymbol{S} = \boldsymbol{S}^{[k]}, \boldsymbol{V} = \boldsymbol{V}^{[k]}, \boldsymbol{F} = \boldsymbol{F}^l$.

Define $\iota_{\boldsymbol{V}}(\boldsymbol{V}) = 0$ if $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ and $+\infty$ otherwise. Similarly, $\iota_{2,C}(\boldsymbol{S}) = 0$ if $\|\boldsymbol{S}\|_{2,C} \leq q$ and $+\infty$ otherwise, and $\iota_C(\boldsymbol{S}) = 0$ if $\|\boldsymbol{s}_i\|_C \leq q^e$ for all $1 \leq i \leq r$ and $+\infty$ otherwise. We use $\iota(\boldsymbol{S})$ to denote $\iota_{2,C}(\boldsymbol{S})$ in the row-equisparsity case and $\iota_C(\boldsymbol{S})$ in the rankwise case. The loss $l_0(\boldsymbol{XB}; \boldsymbol{Y})$ is also written as $l_0(\boldsymbol{XSV}^T; \boldsymbol{Y})$ or $l(\boldsymbol{S}, \boldsymbol{V}; \boldsymbol{X}, \boldsymbol{Y})$, often abbreviated as $l_0(\boldsymbol{XSV}^T)$ or $l(\boldsymbol{S}, \boldsymbol{V})$ for convenience. The general CRL optimization problem can be stated as

$$\min_{\boldsymbol{S}\in\mathbb{R}^{p\times r}, \boldsymbol{V}\in\mathbb{R}^{m\times r}} f(\boldsymbol{S}, \boldsymbol{V}) := l(\boldsymbol{S}, \boldsymbol{V}; \boldsymbol{X}, \boldsymbol{Y}) + \iota(\boldsymbol{S}) + \iota_{\boldsymbol{V}}(\boldsymbol{V}). \tag{27}$$

For simplicity, assume the gradient of l_0 is L-Lipschitz continuous for some $L > 0$:

$$\|\nabla l_0(\boldsymbol{\Theta}_1) - \nabla l_0(\boldsymbol{\Theta}_2)\|_F \leq L\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|_F, \ \forall \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2. \tag{28}$$

One idea might be to apply alternating optimization directly, but it would encounter difficulties when l_0 is non-quadratic. Motivated by Theorem 2, we use a surrogate function to design an iterative algorithm. Given $(\boldsymbol{S}^-, \boldsymbol{V}^-)$, define

$$G_\rho(\boldsymbol{S}, \boldsymbol{V}; \boldsymbol{S}^-, \boldsymbol{V}^-) = l(\boldsymbol{S}^-, \boldsymbol{V}^-) + \langle \nabla l_0(\boldsymbol{X}\boldsymbol{B}^-), \boldsymbol{X}(\boldsymbol{B} - \boldsymbol{B}^-)\rangle + \frac{\rho}{2}\|\boldsymbol{B} - \boldsymbol{B}^-\|_F^2 + \iota(\boldsymbol{S}) + \iota_{\boldsymbol{V}}(\boldsymbol{V}), \tag{29}$$

where $\boldsymbol{B}^- = \boldsymbol{S}^-(\boldsymbol{V}^-)^T$, $\boldsymbol{B} = \boldsymbol{S}\boldsymbol{V}^T$. The dependence of G on ρ is often dropped for notational simplicity. Equation (29) applies linearization on $\boldsymbol{S}\boldsymbol{V}^T$ as a whole to construct the surrogate, but not on $\boldsymbol{S}$ or $\boldsymbol{V}$ individually.

Given any $\left(\boldsymbol{S}^{[0]}, \boldsymbol{V}^{[0]}\right)$, let $\left(\boldsymbol{S}^{[k]}, \boldsymbol{V}^{[k]}\right)$ $(k \geq 1)$ satisfy

$$\left(\boldsymbol{S}^{[k]}, \boldsymbol{V}^{[k]}\right) \in \underset{(\boldsymbol{S},\boldsymbol{V})}{\operatorname{argmin}} G\left(\boldsymbol{S}, \boldsymbol{V}; \boldsymbol{S}^{[k-1]}, \boldsymbol{V}^{[k-1]}\right), \tag{30}$$

or just

$$G\left(\boldsymbol{S}^{[k]}, \boldsymbol{V}^{[k]}; \boldsymbol{S}^{[k-1]}, \boldsymbol{V}^{[k-1]}\right) \leq G\left(\boldsymbol{S}^{[k-1]}, \boldsymbol{V}^{[k-1]}; \boldsymbol{S}^{[k-1]}, \boldsymbol{V}^{[k-1]}\right). \tag{31}$$

We can show that whenever ρ is chosen large enough,

$$f\left(\boldsymbol{S}^{[k]}, \boldsymbol{V}^{[k]}\right) \leq G\left(\boldsymbol{S}^{[k]}, \boldsymbol{V}^{[k]}; \boldsymbol{S}^{[k-1]}, \boldsymbol{V}^{[k-1]}\right), \tag{32}$$

from which it follows that $f\left(\boldsymbol{S}^{[k]}, \boldsymbol{V}^{[k]}\right) \leq G\left(\boldsymbol{S}^{[k-1]}, \boldsymbol{V}^{[k-1]}; \boldsymbol{S}^{[k-1]}, \boldsymbol{V}^{[k-1]}\right) = f\left(\boldsymbol{S}^{[k-1]}, \boldsymbol{V}^{[k-1]}\right)$. A conservative choice is $\rho = L\|\boldsymbol{X}\|_2^2$ (cf. Appendix B.1), but the structural parsimony in $\boldsymbol{B}^{[k]}$ makes it possible to pick a much smaller ρ, which is beneficial from Theorem 2. Let $\bar{\kappa}_2(q, r)$ satisfy $\mathbf{D}_2(\boldsymbol{X}\boldsymbol{B}_1, \boldsymbol{X}\boldsymbol{B}_2) \leq \bar{\kappa}_2(q, r)\mathbf{D}_2(\boldsymbol{B}_1, \boldsymbol{B}_2)$, for $\boldsymbol{B}_i$: $\|\boldsymbol{B}_i\|_{2,C} \leq q$, $\operatorname{rank}(\boldsymbol{B}_i) \leq r$. Summarizing the above derivations gives the following computational convergence.

Theorem 4 *Given any feasible initial point* $(\boldsymbol{S}^{[0]}, \boldsymbol{V}^{[0]})$, *the sequence of iterates* $(\boldsymbol{S}^{[k]}, \boldsymbol{V}^{[k]})$ *generated from Equation* (30) or (31) *satisfies*

$$f\left(\boldsymbol{S}^{[k-1]}, \boldsymbol{V}^{[k-1]}\right) - f\left(\boldsymbol{S}^{[k]}, \boldsymbol{V}^{[k]}\right) \geq \frac{\rho - L\bar{\kappa}_2(q, r)}{2}\|\boldsymbol{B}^{[k]} - \boldsymbol{B}^{[k-1]}\|_F^2$$

for any $k \geq 1$, *where* $\boldsymbol{B}^{[k]} = \boldsymbol{S}^{[k]}\left(\boldsymbol{V}^{[k]}\right)^T$. *Therefore, if* $\rho > L\bar{\kappa}_2(q, r)$, $f\left(\boldsymbol{S}^{[k]}, \boldsymbol{V}^{[k]}\right)$ *is monotonically decreasing,* $\boldsymbol{B}^{[k]} - \boldsymbol{B}^{[k-1]} \to 0$ *as* $k \to +\infty$ *and* $\min_{1\leq k\leq K}\|\boldsymbol{B}^{[k]} - \boldsymbol{B}^{[k-1]}\|_F^2 \leq \frac{1}{K} \cdot \frac{2f(\boldsymbol{S}^{[0]}, \boldsymbol{V}^{[0]})}{\rho - L\bar{\kappa}_2(q,r)}$, $\forall K \geq 1$.

When the value of L is unknown, Equation (32) can be used for line search to get a proper ρ in implementation. After some simple algebra, the optimization problem in Equation (30) is

$$\min_{\boldsymbol{S}\in\mathbb{R}^{p\times r}, \boldsymbol{V}\in\mathbb{R}^{m\times r}} \frac{1}{2}\left\|\boldsymbol{B}^{[k-1]} - \frac{\boldsymbol{X}^T\nabla l_0(\boldsymbol{X}\boldsymbol{B}^{[k-1]})}{\rho} - \boldsymbol{S}\boldsymbol{V}^T\right\|_F^2 + \iota(\boldsymbol{S}) \quad \text{s.t. } \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}. \tag{33}$$

Equation (33) is the unsupervised CRL problem. There is no need to solve the problem in depth though; we use BCD to get some $\left(\boldsymbol{S}^{[k]}, \boldsymbol{V}^{[k]}\right)$ that satisfies (31). Let

$$\tilde{\boldsymbol{Y}} = \boldsymbol{B}^{[k-1]} - \boldsymbol{X}^T\nabla l_0(\boldsymbol{X}\boldsymbol{B}^{[k-1]})/\rho, \quad \boldsymbol{W} = \tilde{\boldsymbol{Y}}^T\boldsymbol{S}, \ \boldsymbol{L} = \tilde{\boldsymbol{Y}}\boldsymbol{V}. \tag{34}$$

First, with $\boldsymbol{S}$ held fixed, a globally optimal $\boldsymbol{V}$ can be obtained by Procrustes rotation: $\boldsymbol{V} = \boldsymbol{U}_w \boldsymbol{V}_w^T$, where $\boldsymbol{U}_w$ and $\boldsymbol{V}_w$ are from the SVD of $\boldsymbol{W}$. Equivalently, $\boldsymbol{V} = \{(\boldsymbol{W}\boldsymbol{W}^T)^+\}^{1/2}\boldsymbol{W}$. The Procrustes rotation simplifies to a normalization operation $\boldsymbol{W}/\|\boldsymbol{W}\|_2$ when $r = 1$.

Next, we solve for $\boldsymbol{S}$ given $\boldsymbol{V}$. Using the orthogonal decomposition $\|\tilde{\boldsymbol{Y}} - \boldsymbol{S}\boldsymbol{V}^T\|_F^2 = \|\boldsymbol{L} - \boldsymbol{S}\|_F^2 + \|\tilde{\boldsymbol{Y}}\boldsymbol{V}_\perp\|_F^2$, the problem reduces to a low-dimensional one:

$$\min_{\boldsymbol{S}\in\mathbb{R}^{p\times r}} \|\boldsymbol{L} - \boldsymbol{S}\|_F^2 + \iota(\boldsymbol{S}).$$

Consider $\|\boldsymbol{S}\|_{2,C} \le q$ first. Let $\boldsymbol{S} = \boldsymbol{F}\boldsymbol{\mu}$, where $\boldsymbol{\mu} \in \mathbb{R}^{q\times r}$ stores the q cluster centroids, and $\boldsymbol{F} \in \mathbb{R}^{p\times q}$ is the associated binary membership matrix, with $\boldsymbol{F}[j, k] = 1$ indicating that the j-row of $\boldsymbol{L}$ falls into the kth cluster, that is, $\boldsymbol{F} \in \mathcal{F}^{p\times q} := \{\boldsymbol{F} \in \mathbb{R}^{p\times q} : \boldsymbol{F} \ge 0, \boldsymbol{F}\mathbf{1} = \mathbf{1}, \boldsymbol{F}^T\boldsymbol{F} \text{ is diagonal}\}$. BCD can be used to update $\boldsymbol{F}$ and $\boldsymbol{\mu}$ alternatively: given $\boldsymbol{F}$, the optimal $\boldsymbol{\mu}$ is $(\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{L}$, while given $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_q]^T$, it suffices to solve $\min_{\boldsymbol{f}\in\{0,1\}^q, \mathbf{1}^T\boldsymbol{f}=1} \|\boldsymbol{L}[j, :] - \boldsymbol{f}^T\boldsymbol{\mu}\|_2^2$, $1 \le j \le p$, from which it follows that $\boldsymbol{F}[j, c_0] = 1$ for $c_0 = \operatorname{argmin}_{1\le c\le q} \|\boldsymbol{L}[j, :] - \boldsymbol{\mu}_c\|_2^2$, and $\boldsymbol{F}[j, c] = 0$ for any $c \ne c_0$. The algorithm turns out to be K-means. Similarly, for the rank-wise constraint $\|\boldsymbol{s}_i\|_C \le q^e$ $(1 \le i \le r)$, we just need to run K-means on each column of $\boldsymbol{L}$ to get $\boldsymbol{S}$. State-of-the-art implementations of K-means skilfully use initialization strategies and usually give a high-quality or even globally optimal solution in low dimensions (Bachem et al., 2016; Zhang & Xia, 2009). Of course, other unsupervised clustering criteria and algorithms can be seamlessly integrated into the framework.

To sum up, the CRL algorithm performs simultaneous dimension reduction and clustering and has guaranteed convergence. Regarding the per-iteration complexity, apart from some fundamental matrix operations, the algorithm involves the SVD of $\boldsymbol{W}$ and the clustering on $\boldsymbol{L}$. Neither is costly in computation, since $\boldsymbol{W}$ and $\boldsymbol{L}$ have only r columns.

4.2 | Parameter tuning

CRL has two regularization parameters q and r; once they are given, CRL can determine the model structure that fits best to the data, including the cluster sizes and the projection subspace. In many applications, we find it possible to directly specify these bounds based on domain knowledge, and they are not very sensitive parameters. But for the sake of cluster and/or rank selection, one must carefully tune the regularization parameters in a data-adaptive manner. The goal of this subsection is to design a proper model comparison criterion assuming a series of candidate models have been obtained (rather than developing a numerical optimization algorithm).

It is well known that parameter tuning is quite challenging in the context of clustering. AIC, BIC and many other known information criteria do not seem to work well, and what makes a sound complexity penalty term is a notable open problem. Fortunately, the statistical studies in Section 3 shed some light on the topic. We advocate a new model penalty $P_o(\cdot)$ as follows

$$P_o(\boldsymbol{B}) = \{\|\boldsymbol{B}\|_{2,C} \wedge \operatorname{rank}(\boldsymbol{X}) + m\} \operatorname{rank}(\boldsymbol{B}) + \{p - \|\boldsymbol{B}\|_{2,C}\} \log \|\boldsymbol{B}\|_{2,C}. \tag{35}$$

Theorem 5 *Given any differentiable loss l_0, assume that vec$(\boldsymbol{E})$ (cf. Equation 9) is sub-Gaussian with mean zero and scale bounded by σ, $\boldsymbol{B}^* \ne \mathbf{0}$ and there exist constants $\delta > 0$ and $C \ge 0$ such that $\boldsymbol{\Delta}_{l_0}(\boldsymbol{X}\boldsymbol{B}_1, \boldsymbol{X}\boldsymbol{B}_2) + C\sigma^2 P_o(\boldsymbol{B}_1) + C\sigma^2 P_o(\boldsymbol{B}_2) \ge \delta \mathbf{D}_2(\boldsymbol{X}\boldsymbol{B}_1, \boldsymbol{X}\boldsymbol{B}_2)$ for all $\boldsymbol{B}_i$. Then, any $\hat{\boldsymbol{B}}$ that minimizes the following criterion*

$$l_0(\boldsymbol{X}\boldsymbol{B}; \boldsymbol{Y}) + A\sigma^2[\{\|\boldsymbol{B}\|_{2,C} \wedge \operatorname{rank}(\boldsymbol{X}) + m\} \operatorname{rank}(\boldsymbol{B}) + \{p - \|\boldsymbol{B}\|_{2,C}\} \log \|\boldsymbol{B}\|_{2,C}], \tag{36}$$

where A is a sufficiently large constant, must satisfy

$$\mathbb{E}[\|\boldsymbol{X}\hat{\boldsymbol{B}} - \boldsymbol{X}\boldsymbol{B}^*\|_F^2 \vee P_o(\hat{\boldsymbol{B}})] \lesssim \sigma^2\{(q^* \wedge \text{rank}(\boldsymbol{X}) + m)r^* + (p - q^*)\log q^*\}.$$

Compared with the results in Section 3.2, Theorem 5 offers the same desired order of statistical accuracy, but involves no regularization parameters. We refer to the new information criterion defined by Equation (36) as the predictive information criterion (**PIC**). Unlike most information criteria, PIC has a nonasymptotic justification, and does not need $n \to +\infty$ or any growth conditions on p or m. The new criterion aims to achieve the best prediction accuracy, and applies regardless of the signal-to-noise ratio.

If the effective noise has a constant scale parameter like in classification, Equation (36) can be directly used. But some problems have an unknown σ. For example, $\boldsymbol{Y}|\boldsymbol{X}\boldsymbol{B}^*$ may belong to the exponential dispersion family with a density

$$\exp[\{\langle \cdot, \boldsymbol{X}\boldsymbol{B}^*\rangle - b(\boldsymbol{X}\boldsymbol{B}^*)\}/\phi]$$

with respect to some base measure. For such models with dispersion, the standard practice is to substitute a preliminary estimate $\hat{\sigma}^2$ for σ^2 in Equation (36), but a fascinating fact is that in some scenarios like regression ($b = \|\cdot\|_F^2/2$), the estimation of σ^2 can be totally bypassed with a scale-free form of PIC.

Recall the Orlicz ψ_α-norm (van der Vaart & Wellner, 1996) defined for a random variable Y: $\|Y\|_{\psi_\alpha} = \inf\{t > 0 : \mathbb{E}\exp[(Y/t)^\alpha] \le 2\}$. Sub-Gaussian random variables have finite ψ_2-norm, and sub-exponential random variables (like Poisson and χ^2) have finite ψ_1-norm. As $\alpha < 1$, random variables with even heavier tails are included (Götze et al., 2021).

Theorem 6 *Let the loss be*

$$l_0(\boldsymbol{X}\boldsymbol{B}; \boldsymbol{Y}) = -\langle \boldsymbol{Y}, \boldsymbol{X}\boldsymbol{B}\rangle + b(\boldsymbol{X}\boldsymbol{B}), \tag{37}$$

where b is differentiable, μ-strongly convex and μ'-strongly smooth with $\kappa = \mu'/\mu$, the domain $\Omega = \{\boldsymbol{\eta} \in \mathbb{R}^{n\times m} : b(\boldsymbol{\eta}) < \infty\}$ is open, and $\boldsymbol{Y}$ takes values in the closure of $\{\nabla b(\boldsymbol{\eta}): \boldsymbol{\eta} \in \Omega\}$. Assume the effective noise $\boldsymbol{E} = \boldsymbol{Y} - \nabla b(\boldsymbol{X}\boldsymbol{B}^*)$ has independent, zero-mean entries e_{ik} that satisfy $\|e_{ik}\|_{\psi_\alpha} \le \sigma$ for some $\alpha \in (0, 2]$ and are nondegenerate in the sense that $\text{var}(e_{ik}) \asymp \sigma^2$, where σ is an unknown parameter. Suppose that the true model is not over-complex in the sense that $\kappa P_o(\boldsymbol{B}^*) < mn/A_0$ for some constant $A_0 > 0$. Let $\delta(\boldsymbol{B}) = A\{P_o(\boldsymbol{B})/(mn/\kappa)\}$ for some constant $A : A < A_0$, and so $\delta(\boldsymbol{B}^*) < 1$. Consider the following criterion

$$\frac{l_0(\boldsymbol{X}\boldsymbol{B}; \boldsymbol{Y}) + b^*(\boldsymbol{Y})}{1 - \delta(\boldsymbol{B})}, \tag{38}$$

where $b^*(\cdot) = \sup_{\boldsymbol{\eta}} \langle \cdot, \boldsymbol{\eta}\rangle - b(\boldsymbol{\eta})$ is the Fenchel conjugate of $b(\cdot)$.

Then, for sufficiently large values of A_0, A, any $\hat{\boldsymbol{B}}$ that minimizes (38) subject to $\delta(\boldsymbol{B}) < 1$ satisfies

$$\|\boldsymbol{X}\hat{\boldsymbol{B}} - \boldsymbol{X}\boldsymbol{B}^*\|_F^2 \vee \frac{P_o(\hat{\boldsymbol{B}})}{\mu^2} \lesssim \frac{\kappa\sigma^2}{\mu^2}\{(q^* \wedge \text{rank}(\boldsymbol{X}) + m)r^* + (p - q^*)\log q^*\}$$

with probability at least $1 - C\exp\{-c(m + P_o(\boldsymbol{B}^*))\} - C\exp\{-c(mn)^{\alpha/2}\}$, or $1 - C\exp\{-c(m + \text{rank}(\boldsymbol{X}))\} - C\exp\{-c(mn)^{\alpha/2}\}$ as $q^* > 1$, for some constants $C, c > 0$.

The theorem needs no restricted eigenvalue or signal strength assumptions. Some other scale-free forms can be used based on the techniques in She and Tran (2019); see Remark A.3.

5 | EXPERIMENTS

We performed extensive simulation studies which, due to limited space, are presented in the appendices. The results show the benefits of CRL: the desired structural parsimony can be successfully captured by simultaneous clustering and dimension reduction, and the removal of nuisance dimensions leads to improved statistical performance and reduced computational cost. We also tested kernel CRL on a variety of benchmark datasets (cf. Figure B.1 and Tables B.1–B.3) and performed experiments in network community detection (cf. Figure B.2 and Table B.4). In addition, simulation studies were conducted to test the performance of CRL when model misspecification occurs, compared with LASSO, group LASSO, reduced rank regression and fused LASSO (Tibshirani et al., 2005) (cf. Table B.5). Interested readers may refer to Appendix B for details. Here, we use two real datasets to demonstrate the performance of CRL in supervised learning. Our code can be found in the supplementary material.[1]

5.1 | Horseshoe crab data

This part performs 'model segmentation' on a horseshoe crab dataset in Agresti (2012), to showcase an application of CRL. The response variable is the number of male crabs residing near a female crab's nest, denoted by satellites, ranging from 0 to 15. The dataset records the number of satellites for 173 female horseshoe crabs in vector $\boldsymbol{y}$, as well as some covariates in $\boldsymbol{X}$, such as width, colour, weight and the intercept. Here, width refers to the carapace width of a female crab, measured in centimetres; colour has several categories from light to dark, and darker female crabs tend to be older than lighter-coloured ones. Following Agresti, we removed some redundant and irrelevant predictors and used width and a dummy variable dark to model satellites. Fitting a simple regression model to the overall data gives $-10+0.5\cdot$ width $-0.4\cdot$ dark.

An interesting question in statistical modelling is to study the possible existence of latent 'sub-populations', across which predictors have different coefficients. To this end, we re-characterize the problem using a trace regression (Koltchinskii et al., 2011):

$$y_i \sim \langle \boldsymbol{X}_i, \boldsymbol{B} \rangle, \ \ i = 1, \ \ldots \ , n \tag{39}$$

where $\boldsymbol{B} \in \mathbb{R}^{n\times p}$ is a matrix of unknowns, and $\boldsymbol{X}_i \in \mathbb{R}^{n\times p}$ has all rows zero except the ith row, which is equal to $\boldsymbol{X}[i, :]$. For the horseshoe crab data, the 173 rows of matrix $\boldsymbol{B}$ give sample-specific coefficient vectors, and the model is clearly overparameterized. CRL helps to estimate the coefficient matrix and identify a small number of sub-models. After running the optimization algorithm and parameter tuning, the whole sample is split to two sub-groups ($q = 2$). The model on the first subset (117 observations) is

$$\text{model 1:} \ \ -9.6 + 0.4 \cdot \text{width} - 0.5 \cdot \text{dark}, \tag{40}$$

[1]Also available at https://ani.stat.fsu.edu/~yshe/code/CRL.zip

while on the second subset (56 observations), we get

$$\text{model 2:} \quad -12 + 0.7 \cdot \text{width} + 2.1 \cdot \text{dark}. \tag{41}$$

The two resulting models are quite different. For example, for every 1-cm increase in width, Equation (40) predicts an increase of 0.4 in the number of satellites, while Equation (41) predicts an increase of 0.7, and the *p*-values associated with the two slopes are both low (<3e-4). Also, notice the positive sign of the coefficient estimate for dark in Equation (41).

To get more intuition of the two detected sub-populations, we built a decision tree using CART (Breiman et al., 1984), which has a pretty simple structure: the prediction outcome is the second sub-population if

$$\text{(a) satellites} \geq 4 \quad \text{and} \quad \text{(b) width} < 28.7,$$

and the first if either condition is violated. Therefore, for the group of female crabs that have at least four satellites but do not yet have an extremely large carapace in width, Equation (41) states that being dark is actually a beneficial factor in attracting more satellites.

5.2 | Newsgroup data

The 20 newsgroup dataset, available at http://ftp.ics.uci.edu, contains about 18k documents falling into 20 binary categories which we treat as responses. The feature matrix records the occurrence information of a large dictionary of words. We chose $p = 200$ words at random and used $n = 2000$ documents for training and the remaining for test. On this dataset, CRL produced $q = 50$ word groups and constructed $r = 16$ factors. A prediction error comparison can be made using the test data. The classification accuracy of an SVM trained on the original 200 words is 40.8%. Using only the 16 CRL factors improves the rate to 45.6%, while a LASSO model with 16 selected words only reaches an accuracy rate of 31.30%.

Next, we study the interpretability of the CRL model. Figure 2 plots the coefficients for three clusters for illustration purposes. Note that we did not use any available word groups from the literature, which may or may not be useful for modelling the responses here.

First, cluster 14, composed of words bitmap, format, graphics and image, shows a single large coefficient in response to category 2. This is sensible, as the documentation shows that the category corresponds to computer graphics.

Cluster 37 contains two words only, hockey and nhl. This group has two big coefficients in magnitude, +2.25 and −1.05 for the categories of hockey and baseball respectively. So the occurrence of these two words seems helpful for differentiating the two related sport categories.

Finally, let us turn to cluster 11 which consists of 42 words. All its coefficients are pretty small, varying between −0.02 and 0.01 for different responses. A careful examination of its composition explains the mild effects: almost all are the so-called 'stop words', such as the, very and yours, and removing this cluster gave almost identical results. CRL was able to capture these essentially irrelevant features and group them together. To sum up, CRL contributes as a beneficial complement to conventional variable selection.

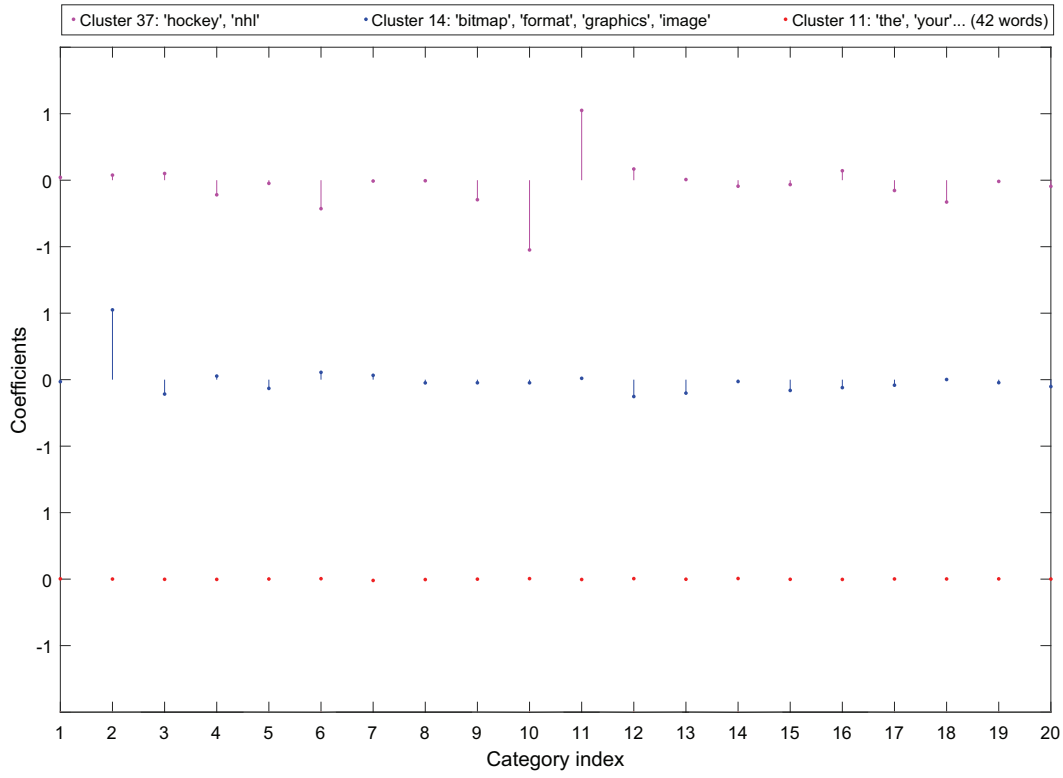

FIGURE 2 Newsgroup data: the coefficients of some word clusters obtained by clustered reduced-rank learning in response to the 20 categories [Colour figure can be viewed at wileyonlinelibrary.com]

6 | CONCLUSIONS

Many high-dimensional methods adopt the 'bet on sparsity' principle (Hastie et al., 2009), but in real multi-response applications, statisticians often face 'dense' problems with such a large number of relevant features that variable selection may be ineffective. This paper proposed a clustered reduced-rank learning framework to build a predictive and interpretable model through feature auto-grouping and dimension reduction.

The joint matrix regularization formulation poses intriguing challenges in both theory and computation. We provided universal information-theoretical limits to reveal the intrinsic cost of seeking for clusters, as well as the benefit of accumulating a large number of response variables in multivariate learning. The obtained error rates are strikingly different from those assuming sparsity. Moreover, we proved that CRL, unlike the class of methods based on pairwise-difference penalization, achieves the minimax optimal rate in some common scenarios. The remarkable fact that the CRL estimators need not be global minimizers but just fixed points in some regular problems paved the way for the design of an efficient optimization algorithm in the nonconvex setup. Furthermore, a new information criterion, along with its scale-free form, was proposed to address cluster and rank selection. Overall, our new method is as interpretable as variable selection, and is advantageous when the numbers of relevant features and irrelevant features are of the same order, or when the number of responses is greater than the number of features up to a multiplicative constant.

CRL can be extended to tensors, and one possible application is model segmentation in a multi-task setting. For example, given $\boldsymbol{Y} = [\tilde{\boldsymbol{y}}_1, \ldots \tilde{\boldsymbol{y}}_n]^T, \boldsymbol{X} = [\tilde{\boldsymbol{x}}_1, \ldots \tilde{\boldsymbol{x}}_n]^T$ and an unknown order-3 tensor $\boldsymbol{B} \in \mathbb{R}^{p\times m\times n}$, we can fit a model $\tilde{\boldsymbol{y}}_i^T \sim \tilde{\boldsymbol{x}}_i^T \boldsymbol{B}_{::i}\, (1 \le i \le n)$ by enforcing low rank in $\boldsymbol{B}$ and equi-sparsity along its third dimension. This can be applied to heterogeneous populations. Moreover, our algorithm often shows a linear convergence rate for small q and r, which deserves further study. Finally, to reduce the search cost (thereby the overall error rate), one possible way is to limit the min/max cluster size; a new form of regularization (convex or nonconvex) that guarantees both interpretability and efficiency is an interesting topic that merits future research.

ACKNOWLEDGEMENTS

The authors thank the editor, associated editor and three anonymous referees for suggestions that significantly improved the paper. We also thank Eric Chi, Fangyun Wei, Zhisheng Zhong and Pranay Tarafdar for helpful discussions on some related topics and assistance in some real data analysis. The first author is particularly grateful to Art Owen for his valuable comments and encouragement. The work is partially supported by the National Science Foundation.

ORCID

Yiyuan She https://orcid.org/0000-0002-5110-3179

REFERENCES

Agresti, A. (2012) *Categorical data analysis*. Wiley series in probability and statiscs. Hoboken: Wiley.

Bachem, O., Lucic, M., Hassani, S.H. & Krause, A. (2016) Fast and provably good seedings for K-means. In: *Proceedings of the 30th international conference on neural information processing systems*, NIPS'16. Curran Associates Inc., pp. 55–63.

Bickel, P.J., Ritov, Y. & Tsybakov, A.B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37, 1705–1732.

Bregman, L. (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3), 200–217.

Breiman, L., Friedman, J., Stone, C. & Olshen, R. (1984) *Classification and regression trees*. Monterey, CA: Taylor & Francis.

Bunea, F., She, Y. & Wegkamp, M. (2011) Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39, 1282–1309.

Candès, E.J. & Plan, Y. (2011) Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Transactions on Information Theory*, 57(4), 2342–2359.

Candès, E. & Tao, T. (2007) The Dantzig selector: satistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6), 2313–2351.

Chi, E.C. & Lange, K. (2015) Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4), 994–1013.

Chun, H. & Keleş, S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 3–25.

Donoho, D.L. & Johnstone, J.M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3), 425–455.

Götze, F., Sambale, H. & Sinulis, A. (2021) Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26, 1–22.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The elements of statistical learning*, 2nd edn. New York: Springer-Verlag.

Izenman, A.J. (1975) Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2), 248–264.

Johnstone, I.M. & Lu, A.Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 682–693.

Koltchinskii, V., Lounici, K. & Tsybakov, A.B. (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5), 2302–2329.

Lambert, J.-P., Fillingham, J., Siahbazi, M., Greenblatt, J., Baetz, K. & Figeys, D. (2010) Defining the budding yeast chromatin-associated interactome. *Molecular Systems Biology*, 6(1), 448.

Lounici, K., Pontil, M., Tsybakov, A.B. & van de Geer, S. (2011) Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39, 2164–2204.

Pan, X. & Heitman, J. (2000) Sok2 regulates yeast pseudohyphal differentiation via a transcription factor cascade that regulates cell-cell adhesion. *Molecular and Cellular Biology*, 20(22), 8364–8372.

Pellikaan, R., Wu, X.-W., Bulygin, S. & Jurrius, R. (2017) *Codes, cryptology and curves with computer algebra*, 1st edn. Cambridge: Cambridge University Press.

She, Y. (2010) Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4, 1055–1096.

She, Y. & Tran, H. (2019) On cross-validation for sparse reduced rank regression. *Journal of the Royal Statistical Society: Series B*, 81, 145–161.

She, Y., Tang, S. & Zhang, Q. (2020) Indirect Gaussian graph learning beyond Gaussianity. *IEEE Transactions on Network Science and Engineering*, 7, 918–929.

She, Y., Wang, Z. & Jin, J. (2021) Analysis of generalized Bregman surrogate algorithms for nonsmooth nonconvex statistical learning. *The Annals of Statistics*, 49(6), 3434–3459.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005) Sparsity and smoothness via the fused LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1), 91–108.

van der Vaart, A. & Wellner, J. (1996) *Weak convergence and empirical processes: with applications to statistics*. Berlin: Springer.

Yuan, M. & Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

Zhang, C. & Xia, S. (2009) K-means clustering algorithm with improved initial center. In: *Proceedings of the 2009 second international workshop on knowledge discovery and data mining*. IEEE, pp. 790–792.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: She, Y., Shen, J. & Zhang, C. (2022) Supervised multivariate learning with simultaneous feature auto-grouping and dimension reduction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3), 912–932. Available from: https://doi.org/10.1111/rssb.12492