Nonparametric Reduced Rank Regression

Rina Foygel[†,*], **Michael Horrell**[†], **Mathias Drton**[†,‡], **John Lafferty**[†]

[*]Department of Statistics
Stanford University

[†]Department of Statistics
University of Chicago

[‡]Department of Statistics
University of Washington

Abstract

We propose an approach to multivariate nonparametric regression that generalizes reduced rank regression for linear models. An additive model is estimated for each dimension of a q-dimensional response, with a shared p-dimensional predictor variable. To control the complexity of the model, we employ a functional form of the Ky-Fan or nuclear norm, resulting in a set of function estimates that have low rank. Backfitting algorithms are derived and justified using a nonparametric form of the nuclear norm subdifferential. Oracle inequalities on excess risk are derived that exhibit the scaling behavior of the procedure in the high dimensional setting. The methods are illustrated on gene expression data.

1 Introduction

In the multivariate regression problem the objective is to estimate the conditional mean $\mathbb{E}(Y \mid X) = m(X) = (m^1(X), \ldots, m^q(X))^\top$ where Y is a q-dimensional response vector and X is a p-dimensional covariate vector. This is also referred to as multi-task learning in the machine learning literature. We are given a sample of n iid pairs $\{(X_i, Y_i)\}$ from the joint distribution of X and Y. Under a linear model, the mean is estimated as $m(X) = BX$ where $B \in \mathbb{R}^{q \times p}$ is a $q \times p$ matrix of regression coefficients. When the dimensions p and q are large relative to the sample size n, the coefficients of B cannot be reliably estimated, without further assumptions.

In reduced rank regression the matrix B is estimated under a rank constraint $r = \operatorname{rank}(B) \le C$, so that the rows or columns of B lie in an r-dimensional subspace of $\mathbb{R}^q$ or $\mathbb{R}^p$. Intuitively, this implies that the model is based on a smaller number of features than the ambient dimensionality p would suggest, or that the tasks representing the components Y^k of the response are closely related. In low dimensions, the constrained rank model can be computed as an orthogonal projection of the least squares solution; but in high dimensions this is not well defined.

Recent research has studied the use of the nuclear norm as a convex surrogate for the rank constraint. The nuclear norm $\|B\|_*$, also known as the trace or Ky-Fan norm, is the sum of the singular vectors of B. A rank constraint can be thought of as imposing sparsity, but in an unknown basis; the nuclear norm plays the role of the ℓ_1 norm in sparse estimation. Its use for low rank estimation problems was proposed by Fazel in [2]. More recently, nuclear norm regularization in multivariate linear regression has been studied by Yuan et al. [10], and by Neghaban and Wainwright [4], who analyzed the scaling properties of the procedure in high dimensions.

In this paper we study nonparametric parallels of reduced rank linear models. We focus our attention on additive models, so that the regression function $m(X) = (m^1(X), \ldots, m^q(X))^\top$ has each component $m^k(X) = \sum_{j=1}^p m_j^k(X_j)$ equal to a sum of p functions, one for each covariate. The objective is then to estimate the $q \times p$ matrix of functions $M(X) = \left[m_j^k(X_j) \right]$.

The first problem we address, in Section 2, is to determine a replacement for the regularization penalty $\|B\|_*$ in the linear model. Because we must estimate a matrix of functions, the analogue of the nuclear norm is not immediately apparent. We propose two related regularization penalties for

1

nonparametric low rank regression, and show how they specialize to the linear case. We then study, in Section 4, the (infinite dimensional) subdifferential of these penalties. In the population setting, this leads to stationary conditions for the minimizer of the regularized mean squared error. This subdifferential calculus then justifies penalized backfitting algorithms for carrying out the optimization for a finite sample. Constrained rank additive models (CRAM) for multivariate regression are analogous to sparse additive models (SPAM) for the case where the response is 1-dimensional [6] (studied also in the reproducing kernel Hilbert space setting by [5]), but with the goal of recovering a low-rank matrix rather than an entry-wise sparse vector. The backfitting algorithms we derive in Section 5 are analogous to the iterative smoothing and soft thresholding backfitting algorithms for SPAM proposed in [6]. A uniform bound on the excess risk of the estimator relative to an oracle is given Section 6. This shows the statistical scaling behavior of the methods for prediction. The analysis requires a concentration result for nonparametric covariance matrices in the spectral norm. Experiments with gene data are given in Section 7, which are used to illustrate different facets of the proposed nonparametric reduced rank regression techniques.

2   Nonparametric Nuclear Norm Penalization

We begin by presenting the penalty that we will use to induce nonparametric regression estimates to be low rank. To motivate our choice of penalty and provide some intuition, suppose that $f^1(x), \ldots, f^q(x)$ are q smooth one-dimensional functions with a common domain. What does it mean for this collection of functions to be low rank? Let $x_1, x_2, \ldots, x_n$ be a collection of points in the common domain of the functions. We require that the $n \times q$ matrix of function values $\mathbb{F}(x_{1:n}) = [f^k(x_i)]$ is low rank. This matrix is of rank at most $r < q$ for every set $\{x_i\}$ of arbitrary size n if and only if the functions $\{f^k\}$ are r-linearly independent—each function can be written as a linear combination of r of the other functions.

In the multivariate regression setting, but still assuming the domain is one-dimensional for simplicity ($q > 1$ and $p = 1$), we have a random sample $X_1, \ldots, X_n$. Consider the $n \times q$ sample matrix $\mathbb{M} = [m^k(X_i)]$ associated with a vector $M = (m^1, \ldots, m^q)$ of q smooth (regression) functions, and suppose that $n > q$. We would like for this to be a low rank matrix. This suggests the penalty $\|\mathbb{M}\|_* = \sum_{s=1}^q \sigma_s(\mathbb{M}) = \sum_{s=1}^q \sqrt{\lambda_s(\mathbb{M}^\top \mathbb{M})}$, where $\{\lambda_s(A)\}$ denotes the eigenvalues of a symmetric matrix A and $\{\sigma_s(B)\}$ denotes the singular values of a matrix B. Now, assuming the columns of $\mathbb{M}$ are centered, and $\mathbb{E}[m^k(X)] = 0$ for each k, we recognize $\frac{1}{n}\mathbb{M}^\top \mathbb{M}$ as the sample covariance $\widehat{\Sigma}(M)$ of the population covariance $\Sigma(M) := \mathrm{Cov}(M(X)) = [\mathbb{E}(m^k(X)m^l(X))]$. This motivates the following sample and population penalties, where $A^{1/2}$ denotes the matrix square root:

$$\text{population penalty:} \quad \|\Sigma(M)^{1/2}\|_* = \|\mathrm{Cov}(M(X))^{1/2}\|_* \tag{2.1}$$

$$\text{sample penalty:} \quad \|\widehat{\Sigma}(M)^{1/2}\|_* = \frac{1}{\sqrt{n}}\|\mathbb{M}\|_*. \tag{2.2}$$

With $\mathbb{Y}$ denoting the $n \times q$ matrix of response values for the sample (X_i, Y_i), this leads to the following population and empirical regularized risk functionals for low rank nonparametric regression:

$$\text{population penalized risk:} \quad \frac{1}{2}\mathbb{E}\|Y - M(X)\|_2^2 + \lambda\|\Sigma(M)^{1/2}\|_* \tag{2.3}$$

$$\text{empirical penalized risk:} \quad \frac{1}{2n}\|\mathbb{Y} - \mathbb{M}\|_F^2 + \frac{\lambda}{\sqrt{n}}\|\mathbb{M}\|_*. \tag{2.4}$$

We recall that if $A \geq 0$ has spectral decomposition $A = UDU^\top$ then $A^{1/2} = UD^{1/2}U^\top$.

3   Constrained Rank Additive Models (CRAM)

We now consider the case where X is p-dimensional. Throughout the paper we use superscripts to denote indices of the q-dimensional response, and subscripts to denote indices of the p-dimensional covariate. We consider the family of additive models, with regression functions of the form $m(X) = (m^1(X), \ldots, m^q(X))^\top = \sum_{j=1}^p M_j(X_j)$, where each term $M_j(X_j) = (m_j^1(X_j), \ldots, m_j^q(X_j))^\top$ is a q-vector of functions evaluated at X_j.

2

In this setting we propose two different penalties. The first penalty, intuitively, encourages the vector $(m_j^1(X_j), \ldots, m_j^q(X_j))$ to be low rank, for each j. Assume that the functions $m_j^k(X_j)$ all have mean zero; this is required for identifiability in the additive model. As a shorthand, let $\Sigma_j = \Sigma(M_j) = \mathrm{Cov}(M_j(X_j))$ denote the covariance matrix of the j-th component functions, with sample version $\widehat{\Sigma}_j$. The population and sample versions of the first penalty are then given by

$$\left\| \Sigma_1^{1/2} \right\|_* + \left\| \Sigma_2^{1/2} \right\|_* + \cdots + \left\| \Sigma_p^{1/2} \right\|_* \tag{3.1}$$

$$\left\| \widehat{\Sigma}_1^{1/2} \right\|_* + \left\| \widehat{\Sigma}_2^{1/2} \right\|_* + \cdots + \left\| \widehat{\Sigma}_p^{1/2} \right\|_* = \frac{1}{\sqrt{n}} \sum_{j=1}^{p} \| \mathbb{M}_j \|_*. \tag{3.2}$$

The second penalty, intuitively, encourages the set of q vector-valued functions $(m_1^k, m_2^k, \ldots, m_p^k)^\top$ to be low rank. This penalty is given by

$$\left\| \left(\Sigma_1^{1/2} \cdots \Sigma_p^{1/2} \right) \right\|_* \tag{3.3}$$

$$\left\| \left(\widehat{\Sigma}_1^{1/2} \cdots \widehat{\Sigma}_p^{1/2} \right) \right\|_* = \frac{1}{\sqrt{n}} \| \mathbb{M}_{1:p} \|_* \tag{3.4}$$

where, for convenience of notation, $\mathbb{M}_{1:p} = \left(\mathbb{M}_1^\top \cdots \mathbb{M}_p^\top \right)^\top$ is an $np \times q$ matrix. The corresponding population and empirical risk functionals, for the first penalty, are then

$$\frac{1}{2} \mathbb{E} \left\| Y - \sum_{j=1}^{p} M_j(X) \right\|_2^2 + \lambda \sum_{j=1}^{p} \left\| \Sigma_j^{1/2} \right\|_* \tag{3.5}$$

$$\frac{1}{2n} \left\| \mathbb{Y} - \sum_{j=1}^{p} \mathbb{M}_j \right\|_F^2 + \frac{\lambda}{\sqrt{n}} \sum_{j=1}^{p} \| \mathbb{M}_j \|_* \tag{3.6}$$

and similarly for the second penalty.

Now suppose that each X_j is normalized so that $\mathbb{E}(X_j^2) = 1$. In the linear case we have $M_j(X_j) = X_j B_j$ where $B_j \in \mathbb{R}^q$. Let $B = (B_1 \cdots B_p) \in \mathbb{R}^{q \times p}$. Some straightforward calculation shows that the penalties reduce to $\sum_{j=1}^{p} \| \Sigma_j^{1/2} \|_* = \sum_{j=1}^{p} \| B_j \|_2$ for the first penalty, and $\| \Sigma_1^{1/2} \cdots \Sigma_p^{1/2} \|_* = \| B \|_*$ for the second. Thus, in the linear case the first penalty is encouraging B to be column-wise sparse, so that many of the B_js are zero, meaning that X_j doesn't appear in the fit. This is a version of the group lasso [11]. The second penalty reduces to the nuclear norm regularization $\| B \|_*$ used for high-dimensional reduced-rank regression.

4 Subdifferentials for Functional Matrix Norms

A key to deriving algorithms for functional low-rank regression is computation of the subdifferentials of the penalties. We are interested in $(q \times p)$-dimensional matrices of functions $F = [f_j^k]$. For each column index j and row index k, f_j^k is a function of a random variable X_j, and we will take expectations with respect to X_j implicitly. We write F_j to mean the jth column of F, which is a q-vector of functions of X_j. We define the inner product between two matrices of functions as

$$\langle\!\langle F, G \rangle\!\rangle := \sum_{j=1}^{p} \sum_{k=1}^{q} \mathbb{E}(f_j^k g_j^k) = \sum_{j=1}^{p} \mathbb{E}(F_j^\top G_j) = \mathrm{tr}\left(\mathbb{E}(FG^\top) \right), \tag{4.1}$$

and write $\| F \|_2 = \sqrt{\langle\!\langle F, F \rangle\!\rangle}$. Note that $\| F \|_2 = \left\| \sqrt{\mathbb{E}(FF^\top)} \right\|_F$ where $\mathbb{E}(FF^\top) = \sum_j \mathbb{E}(F_j F_j^\top) \succeq 0$ is a positive semidefinite $q \times q$ matrix.

We define two further norms on a matrix of functions F, namely,

$$\| F \|_{\mathrm{sp}} := \sqrt{\| \mathbb{E}(FF^\top) \|_{\mathrm{sp}}} = \left\| \sqrt{\mathbb{E}(FF^\top)} \right\|_{\mathrm{sp}} \quad \text{and} \quad \| F \|_* := \left\| \sqrt{\mathbb{E}(FF^\top)} \right\|_*,$$

where $\| A \|_{\mathrm{sp}}$ is the spectral norm (operator norm), the largest singular value of A, and it is convenient to write the matrix square root as $\sqrt{A} = A^{1/2}$. Each of the norms depends on F only through $\mathbb{E}(FF^\top)$. In fact, these two norms are dual—for any F,

$$\| F \|_* = \sup_{\| G \|_{\mathrm{sp}} \leq 1} \langle\!\langle G, F \rangle\!\rangle, \tag{4.2}$$

3

where the supremum is attained by setting $G = \left(\sqrt{\mathbb{E}(FF^\top)}\right)^{-1} F$, with A^{-1} denoting the matrix pseudo-inverse.

Proposition 4.1. *The subdifferential of* $\|F\|_*$ *is the set*

$$\mathcal{S}(F) := \left\{ \left(\sqrt{\mathbb{E}(FF^\top)}\right)^{-1} F + H \; : \; \|H\|_{\mathrm{sp}} \le 1, \; \mathbb{E}(FH^\top) = \mathbf{0}_{q\times q}, \; \mathbb{E}(FF^\top)H = \mathbf{0}_{q\times p} \; a.e. \right\}. \tag{4.3}$$

Proof. The fact that $\mathcal{S}(F)$ contains the subdifferential $\partial\|F\|_*$ can be proved by comparing our setting (matrices of functions) to the ordinary matrix case; see [9, 7]. Here, we show the reverse inclusion, $\mathcal{S}(F) \subseteq \partial\|F\|_*$. Let $D \in \mathcal{S}(F)$ and let G be any element of the function space. We need to show

$$\|F + G\|_* \ge \|F\|_* + \langle\!\langle G, D \rangle\!\rangle, \tag{4.4}$$

where $D = \left(\sqrt{\mathbb{E}(FF^\top)}\right)^{-1} F + H =: \widetilde{F} + H$ for some H satisfying the conditions in (4.3) above. Expanding the right-hand side of (4.4), we have

$$\|F\|_* + \langle\!\langle G, D \rangle\!\rangle = \|F\|_* + \langle\!\langle G, \widetilde{F} + H \rangle\!\rangle = \langle\!\langle F + G, \widetilde{F} + H \rangle\!\rangle \le \|F + G\|_* \|D\|_{\mathrm{sp}},$$

where the second equality follows from $\|F\|_* = \langle\!\langle F, \widetilde{F}\rangle\!\rangle$, and the fact that $\langle\!\langle F, H\rangle\!\rangle = \mathrm{tr}\left(\mathbb{E}(FH^\top)\right) = 0$. The inequality follows from the duality of the norms.

Finally, we show that $\|D\|_{\mathrm{sp}} \le 1$. We have

$$\mathbb{E}(DD^\top) = \mathbb{E}(\widetilde{F}\widetilde{F}^\top) + \mathbb{E}(\widetilde{F}H^\top) + \mathbb{E}(H\widetilde{F}^\top) + \mathbb{E}(HH^\top) = \mathbb{E}(\widetilde{F}\widetilde{F}^\top) + \mathbb{E}(HH^\top),$$

where we use the fact that $\mathbb{E}(FH^\top) = \mathbf{0}_{q\times q}$, implying $\mathbb{E}(\widetilde{F}H^\top) = \mathbf{0}_{q\times q}$. Next, let $\mathbb{E}(FF^\top) = VDV^\top$ be a reduced singular value decomposition, where D is a positive diagonal matrix of size $q' \le q$. Then $\mathbb{E}(\widetilde{F}\widetilde{F}^\top) = VV^\top$, and we have

$$\mathbb{E}(FF^\top) \cdot H = \mathbf{0}_{q\times p} \text{ a.e. } \Leftrightarrow V^\top H = \mathbf{0}_{q'\times p} \text{ a.e. } \Leftrightarrow \mathbb{E}(\widetilde{F}\widetilde{F}^\top)H = \mathbf{0}_{q\times p} \text{ a.e. }.$$

This implies that $\mathbb{E}(\widetilde{F}\widetilde{F}^\top) \cdot \mathbb{E}(HH^\top) = \mathbf{0}_{q\times q}$ and so these two symmetric matrices have orthogonal row spans and orthogonal column spans. Therefore,

$$\left\|\mathbb{E}(DD^\top)\right\|_{\mathrm{sp}} = \left\|\mathbb{E}(\widetilde{F}\widetilde{F}^\top) + \mathbb{E}(HH^\top)\right\|_{\mathrm{sp}} = \max\left\{\left\|\mathbb{E}(\widetilde{F}\widetilde{F}^\top)\right\|_{\mathrm{sp}}, \left\|\mathbb{E}(HH^\top)\right\|_{\mathrm{sp}}\right\} \le 1,$$

where the last bound comes from the fact that $\|\widetilde{F}\|_{\mathrm{sp}}, \|H\|_{\mathrm{sp}} \le 1$. Therefore $\|D\|_{\mathrm{sp}} \le 1$. $\qquad\square$

This gives the subdifferential of penalty 2, defined in (3.3). We can view the first penalty update as just a special case of the second penalty update. For penalty 1 in (3.1), if we are updating F_j and fix all the other functions, we are now penalizing the norm

$$\|F_j\|_* = \left\|\sqrt{\mathbb{E}(F_j F_j^\top)}\right\|_*, \tag{4.5}$$

which is clearly just a special case of penalty 2 with a single q-vector of functions instead of p different q-vectors of functions. So, we have

$$\partial\|F_j\|_* = \left\{ \left(\sqrt{\mathbb{E}(F_j F_j^\top)}\right)^{-1} F_j + H_j \; : \; \|H_j\|_{\mathrm{sp}} \le 1, \; \mathbb{E}(F_j H_j^\top) = \mathbf{0}, \; \mathbb{E}(F_j F_j^\top)H_j = \mathbf{0} \; a.e. \right\}. \tag{4.6}$$

5 Stationary Conditions and Backfitting Algorithms

Returning to the base case of $p = 1$ covariate, consider the population regularized risk optimization

$$\min_M \left\{ \frac{1}{2}\mathbb{E}\|Y - M(X)\|_2^2 + \lambda\|M\|_* \right\}, \tag{5.1}$$

where M is a vector of q univariate functions. The stationary condition for this optimization is

$$\mathbb{E}(Y \mid X) = M(X) + \lambda V(X) \quad \text{a.e. for some } V \in \partial\|M\|_*. \tag{5.2}$$

Define $P(X) := \mathbb{E}(Y \mid X)$.

4

CRAM BACKFITTING ALGORITHM — FIRST PENALTY

Input: Data (X_i, Y_i), regularization parameter λ.

Initialize $\widehat{M}_j = 0$, for $j = 1, \dots, p$.

Iterate until convergence:

    For each $j = 1, \dots, p$:

    (1) Compute the residual: $Z_j = Y - \sum_{k \neq j} \widehat{M}_k(X_k)$;

    (2) Estimate $P_j = \mathbb{E}[Z_j \,|\, X_j]$ by smoothing: $\widehat{P}_j = \mathcal{S}_j Z_j$;

    (3) Compute SVD: $\frac{1}{n} \widehat{P}_j \widehat{P}_j^\top = U \operatorname{diag}(\tau) U^\top$

    (4) Soft threshold: $\widehat{M}_j = U \operatorname{diag}([1 - \lambda/\sqrt{\tau}]_+) U^\top \widehat{P}_j$;

    (5) Center: $\widehat{M}_j \leftarrow \widehat{M}_j - \operatorname{mean}(\widehat{M}_j)$.

Output: Component functions $\widehat{M}_j$ and estimator $\widehat{M}(X_i) = \sum_j \widehat{M}_j(X_{ij})$.

Figure 1: The CRAM backfitting algorithm, using the first penalty, which penalizes each component.

Proposition 5.1. *Let* $\mathbb{E}(PP^\top) = U \operatorname{diag}(\tau) U^\top$ *be the singular value decomposition and define*

$$M = U \operatorname{diag}([1 - \lambda/\sqrt{\tau}]_+) U^\top P \tag{5.3}$$

where $[x]_+ = \max(x, 0)$. *Then* M *satisfies stationary condition* (5.2)*, and is a minimizer of* (5.1)*.*

Proof. Assume the singular values are sorted as $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_q$, and let r be the largest index such that $\sqrt{\tau_r} > \lambda$. Thus, M has rank r. Note that $\sqrt{\mathbb{E}(MM^\top)} = U \operatorname{diag}([\sqrt{\tau} - \lambda]_+) U^\top$, and therefore

$$\lambda \big(\sqrt{\mathbb{E}(MM^\top)} \big)^{-1} M = U \operatorname{diag}(\lambda/\sqrt{\tau_{1:r}}, \mathbf{0}_{q-r}) U^\top P \tag{5.4}$$

where $x_{1:k} = (x_1, \dots, x_k)$ and $c_k = (c, \dots, c)$. It follows that

$$M + \lambda \big(\sqrt{\mathbb{E}(MM^\top)} \big)^{-1} M = U \operatorname{diag}(\mathbf{1}_r, \mathbf{0}_{q-r}) U^\top P. \tag{5.5}$$

Now define

$$H = \frac{1}{\lambda} U \operatorname{diag}(\mathbf{0}_r, \mathbf{1}_{q-r}) U^\top P \tag{5.6}$$

and take $V = \big(\sqrt{\mathbb{E}(MM^\top)} \big)^{-1} M + H$. Then we have $M + \lambda V = P$.

It remains to show that H satisfies the conditions of the subdifferential in (4.3). Since $\sqrt{\mathbb{E}(HH^\top)} = U \operatorname{diag}(\mathbf{0}_r, \sqrt{\tau_{r+1}}/\lambda, \dots, \sqrt{\tau_q}/\lambda) U^\top$ we have $\|H\|_{\text{sp}} \leq 1$. Also, $\mathbb{E}(MH^\top) = \mathbf{0}_{q \times q}$ since

$$\operatorname{diag}(1 - \lambda/\sqrt{\tau_{1:r}}, \mathbf{0}_{q-r}) \operatorname{diag}(\mathbf{0}_r, \mathbf{1}_{q-r}/\lambda) = \mathbf{0}_{q \times q}. \tag{5.7}$$

Similarly, $\mathbb{E}(MM^\top) H = \mathbf{0}_{q \times q}$ since

$$\operatorname{diag}((\sqrt{\tau_{1:r}} - \lambda)^2, \mathbf{0}_{q-r}) \operatorname{diag}(\mathbf{0}_r, \mathbf{1}_{q-r}/\lambda) = \mathbf{0}_{q \times q}. \tag{5.8}$$

It follows that $V \in \partial \|M\|_{\text{sp}}$. $\qquad\square$

The analysis above justifies a backfitting algorithm for estimating a constrained rank additive model with the first penalty, where the objective is

$$\min_{M_j} \Big\{ \frac{1}{2} \mathbb{E} \Big\| Y - \sum_{j=1}^{p} M_j(X_j) \Big\|_2^2 + \lambda \sum_{j=1}^{p} \|M_j\|_* \Big\}. \tag{5.9}$$

For a given coordinate j, we form the residual $Z_j = Y - \sum_{k \neq j} M_k$, and then compute the projection $P_j = \mathbb{E}(Z_j \,|\, X_j)$, with singular value decomposition $\mathbb{E}(P_j P_j^\top) = U \operatorname{diag}(\tau) U^\top$. We then update

$$M_j = U \operatorname{diag}([1 - \lambda/\sqrt{\tau}]_+) U^\top P_j \tag{5.10}$$

5

and proceed to the next variable. This is a Gauss-Seidel procedure that parallels the population backfitting algorithm for SPAM [6]. In the sample version we replace the conditional expectation $P_j = \mathbb{E}(Z_j \,|\, X_j)$ by a nonparametric linear smoother, $\widehat{P}_j = \mathcal{S}_j Z_j$. The algorithm is given in Figure 1. Note that to predict at a point x not included in the training set, the smoother matrices are constructed using that point; that is, $\widehat{P}_j(x_j) = S_j(x_j)^\top Z_j$.

The algorithm for penalty 2 is similar. In step (3) of the algorithm in Figure 1 we compute the SVD of $\frac{1}{n}\widehat{P}_{1:p}\widehat{P}_{1:p}^\top$. Then, in step (4) we soft threshold according to $\widehat{M}_{1:p} = U \operatorname{diag}\big(\big[1 - \lambda/\sqrt{\tau}\,\big]_+\big)U^\top \widehat{P}_{1:p}$. Both algorithms can be viewed as functional projected gradient descent procedures.

6 Excess Risk Bounds

The population risk of a $q \times p$ regression matrix $M(X) = [M_1(X_1)\cdots M_p(X_p)]$ is

$$R(M) = \mathbb{E}\|Y - M(X)\mathbf{1}_p\|_2^2,$$

with sample version denoted $\widehat{R}(M)$. Consider all models that can be written as

$$M(X) = U \cdot D \cdot V(X)^\top$$

where U is an orthogonal $q \times r$ matrix, D is a positive diagonal matrix, and $V(X) = [v_{js}(X_j)]$ satisfies $\mathbb{E}(V^\top V) = I_r$. The population risk can be reexpressed as

$$R(M) = \operatorname{tr}\left\{\begin{pmatrix}-I_q \\ DU^\top\end{pmatrix}^\top \mathbb{E}\left[\begin{pmatrix}Y \\ V(X)^\top\end{pmatrix}\begin{pmatrix}Y \\ V(X)^\top\end{pmatrix}^\top\right]\begin{pmatrix}-I_q \\ DU^\top\end{pmatrix}\right\}$$

$$= \operatorname{tr}\left\{\begin{pmatrix}-I_q \\ DU^\top\end{pmatrix}^\top \begin{pmatrix}\Sigma_{YY} & \Sigma_{YV} \\ \Sigma_{YV}^\top & \Sigma_{VV}\end{pmatrix}\begin{pmatrix}-I_q \\ DU^\top\end{pmatrix}\right\}$$

and similarly for the sample risk, with $\widehat{\Sigma}_n(V)$ replacing $\Sigma(V) := \operatorname{Cov}((Y, V(X)^\top))$ above. The "uncontrollable" contribution to the risk, which does not depend on M, is $R_u = \operatorname{tr}\{\Sigma_{YY}\}$. We can express the remaining "controllable" risk as

$$R_c(M) = R(M) - R_u = \operatorname{tr}\left\{\begin{pmatrix}-2I_q \\ DU^\top\end{pmatrix}^\top \Sigma(V) \begin{pmatrix}\mathbf{0}_q \\ DU^\top\end{pmatrix}\right\}.$$

Using the von Neumann trace inequality, $\operatorname{tr}(AB) \leq \|A\|_p \|B\|_{p'}$ where $1/p + 1/p' = 1$,

$$R_c(M) - \widehat{R}_c(M) \leq \left\|\begin{pmatrix}-2I_q \\ DU^\top\end{pmatrix}^\top (\Sigma(V) - \widehat{\Sigma}_n(V))\right\|_{\mathrm{sp}} \left\|\begin{pmatrix}\mathbf{0}_q \\ DU^\top\end{pmatrix}\right\|_*$$

$$\leq \left\|\begin{pmatrix}-2I_q \\ DU^\top\end{pmatrix}^\top\right\|_{\mathrm{sp}} \left\|\Sigma(V) - \widehat{\Sigma}_n(V)\right\|_{\mathrm{sp}} \|D\|_*$$

$$\leq C \max(2, \|D\|_{\mathrm{sp}}) \left\|\Sigma(V) - \widehat{\Sigma}_n(V)\right\|_{\mathrm{sp}} \|D\|_*$$

$$\leq C \max\{2, \|D\|_*^2\} \left\|\Sigma(V) - \widehat{\Sigma}_n(V)\right\|_{\mathrm{sp}} \tag{6.1}$$

where here and in the following C is a generic constant. For the last factor in (6.1), it holds that

$$\sup_V \left\|\Sigma(V) - \widehat{\Sigma}_n(V)\right\|_{\mathrm{sp}} \leq C \sup_V \sup_{w \in \mathcal{N}} w^\top \big(\Sigma(V) - \widehat{\Sigma}_n(V)\big) w$$

where $\mathcal{N}$ is a $1/2$-covering of the unit $(q+r)$-sphere, which has size $|\mathcal{N}| \leq 6^{q+r} \leq 36^q$; see [8]. We now assume that the functions $v_{sj}(x_j)$ are uniformly bounded from a Sobolev space of order two. Specifically, let $\{\psi_{jk} : k = 0, 1, \ldots\}$ denote a uniformly bounded, orthonormal basis with respect to $L^2[0,1]$, and assume that $v_{sj} \in \mathcal{H}_j$ where

$$\mathcal{H}_j = \left\{f_j : \; f_j(x_j) = \sum_{k=0}^\infty a_{jk}\psi_{jk}(x_j), \;\; \sum_{k=0}^\infty a_{jk}^2 k^4 \leq K^2\right\}$$

for some $0 < K < \infty$. The L_∞-covering number of $\mathcal{H}_j$ satisfies $\log \mathcal{N}(\mathcal{H}_j, \epsilon) \leq K/\sqrt{\epsilon}$.

6

Suppose that $Y - \mathbb{E}(Y\,|\,X) = W$ is Gaussian and the true regression function $\mathbb{E}(Y\,|\,X)$ is bounded. Then the family of random variables $Z_{(V,w)} \coloneqq \sqrt{n} \cdot w^\top (\Sigma(V) - \widehat{\Sigma}_n(V))w$ is sub-Gaussian and sample continuous. It follows from a result of Cesa-Bianchi and Lugosi [1] that

$$\mathbb{E}\left(\sup_V \sup_{w \in \mathcal{N}} w^\top (\Sigma(V) - \widehat{\Sigma}_n(V))w\right) \le \frac{C}{\sqrt{n}} \int_0^B \sqrt{q \log(36) + \log(pq) + \frac{K}{\sqrt{\epsilon}}}\; d\epsilon$$

for some constant B. Thus, by Markov's inequality we conclude that

$$\sup_V \left\|\Sigma(V) - \widehat{\Sigma}_n(V)\right\|_{\mathrm{sp}} = O_P\left(\sqrt{\frac{q + \log(pq)}{n}}\right). \tag{6.2}$$

If $\|M\|_* = \|D\|_* = o\left(n/(q + \log(pq))\right)^{1/4}$, then returning to (6.1), this gives us a bound on $R_c(M) - \widehat{R}_c(M)$ that is $o_P(1)$. More precisely, we define a class of matrices of functions:

$$\mathcal{M}_n = \left\{M : M(X) = UDV(X)^\top,\; \text{with } \mathbb{E}(V^\top V) = I,\; v_{sj} \in \mathcal{H}_j,\; \|D\|_* = o\left(\frac{n}{q + \log(pq)}\right)^{1/4}\right\}.$$

Then, for a fitted matrix $\widehat{M}$ chosen from $\mathcal{M}_n$, writing $M_* = \arg\min_{M \in \mathcal{M}_n} R(M)$, we have

$$R(\widehat{M}) - \inf_{M \in \mathcal{M}_n} R(M) = R(\widehat{M}) - \widehat{R}(\widehat{M}) - (R(M_*) - \widehat{R}(M_*)) + (\widehat{R}(\widehat{M}) - \widehat{R}(M_*))$$

$$\le \left[R(\widehat{M}) - \widehat{R}(\widehat{M})\right] - \left[R(M_*) - \widehat{R}(M_*)\right].$$

Subtracting $R_u - \widehat{R}_u$ from each of the bracketed differences, we obtain that

$$R(\widehat{M}) - \inf_{M \in \mathcal{M}_n} R(M) \le \left[R_c(\widehat{M}) - \widehat{R}_c(\widehat{M})\right] - \left[R_c(M_*) - \widehat{R}_c(M_*)\right]$$

$$\le 2 \sup_{M \in \mathcal{M}_n} \left\{R_c(M) - \widehat{R}_c(M)\right\}$$

$$\overset{\text{by (6.1)}}{\le} O\left(\|D\|_*^2 \left\|\Sigma(V) - \widehat{\Sigma}_n(V)\right\|_{\mathrm{sp}}\right) \overset{\text{by (6.2)}}{=} o_P(1).$$

This proves the following result.

Proposition 6.1. *Let $\widehat{M}$ minimize the empirical risk $\frac{1}{n}\sum_i \|Y_i - \sum_j M_j(X_{ij})\|_2^2$ over the class $\mathcal{M}_n$. Then*

$$R(\widehat{M}) - \inf_{M \in \mathcal{M}_n} R(M) \xrightarrow{P} 0.$$

7 Application to Gene Expression Data

To illustrate the proposed nonparametric reduced rank regression techniques, we consider data on gene expression in *E. coli* from the "DREAM 5 Network Inference Challenge"[1] [3]. In this challenge genes were classified as transcription factors (TFs) or target genes (TGs). Transcription factors regulate the target genes, as well as other TFs.

We focus on predicting the expression levels Y for a particular set of $q = 27$ TGs, using the expression levels X for $p = 6$ TFs. Our motivation for analyzing these 33 genes is that, according to the gold standard gene regulatory network used for the DREAM 5 challenge, the 6 TFs form the parent set common to two additional TFs, which have the 27 TGs as their child nodes. In fact, the two intermediate nodes d-separate the 6 TFs and the 27 TGs in a Bayesian network interpretation of this gold standard. This means that if we treat the gold standard as a causal network, then up to noise, the functional relationship between X and Y is given by the composition of a map $g : \mathbb{R}^6 \to \mathbb{R}^2$ and a map $h : \mathbb{R}^2 \to \mathbb{R}^{27}$. If g and h are both linear, their composition $h \circ g$ is a linear map of rank no more than 2. As observed in Section 2, such a reduced rank linear model is a special case of an additive model with reduced rank in the sense of penalty 2. More generally, if g is an additive function and h is linear, then $h \circ g$ has rank at most 2 in the sense of penalty 2. Higher rank can in principle occur

---

[1] http://wiki.c2b2.columbia.edu/dream/index.php/D5c4

7

Penalty 1, $\lambda = 20$ Penalty 2, $\lambda = 5$
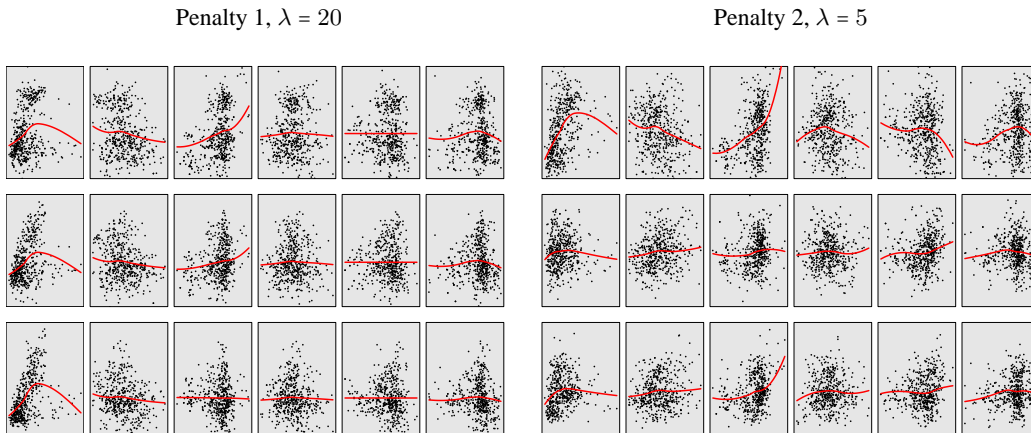
Figure 2: Left: Penalty 1 with large tuning parameter. Right: Penalty 2 with tuning parameter obtained through 10-fold cross-validation. Plotted points are residuals holding out the given predictor.

under functional composition, since even a univariate additive map $h : \mathbb{R} \to \mathbb{R}^q$ may have rank up to q under our penalties (recall that penalty 1 and 2 coincide for univariate maps).

The backfitting algorithm of Figure 1 with penalty 1 and a rather aggressive choice of the tuning parameter λ produces the estimates shown in Figure 2, for which we have selected three of the 27 TGs. Under such strong regularization, the 5th column of functions is rank zero and, thus, identically zero. The remaining columns have rank one; the estimated fitted values are scalar multiples of one another. We also see that scalings can be different for different columns. The third plot in the third row shows a slightly negative slope, indicating a negative scaling for this particular estimate. The remaining functions in this row are oriented similarly to the other rows, indicating the same, positive scaling. This property characterizes the difference between penalties 1 and 2; in an application of penalty 2, the scalings would have been the same across all functions in a given row.

Next, we illustrate a higher-rank solution for penalty 2. Choosing the regularization parameter λ by ten-fold cross-validation gives a fit of rank 5, considerably lower than 27, the maximum possible rank. Figure 2 shows a selection of three coordinates of the fitted functions. Under rank five, each row of functions is a linear combination of up to five other, linearly independent rows. We remark that the use of cross-validation generally produces somewhat more complex models than is necessary to capture an underlying low-rank data-generating mechanism. Hence, if the causal relationships for these data were indeed additive and low rank, then the true low rank might well be smaller than five.

8 Summary

This paper introduced two penalties that induce reduced rank fits in multivariate additive nonparametric regression. Under linearity, the penalties specialize to group lasso and nuclear norm penalties for classical reduced rank regression. Examining the subdifferentials of each of these penalties, we developed backfitting algorithms for the two resulting optimization problems that are based on soft-thresholding of singular values of smoothed residual matrices. The algorithms were demonstrated on a gene expression data set constructed to have a naturally low-rank structure. We also provided a persistence analysis that shows error tending to zero under a scaling assumption on the sample size n and the dimensions q and p of the regression problem.

Acknowledgements

Research supported in part by NSF grants IIS-1116730, DMS-0746265, and DMS-1203762, AFOSR grant FA9550-09-1-0373, ONR grant N000141210762, and an Alfred P. Sloan Fellowship.

8

References

[1] Nicolò Cesa-Bianchi and Gábor Lugosi. On prediction of individual sequences. *The Annals of Statistics*, 27(6):1865–1894, 1999.

[2] Maryam Fazel. Matrix rank minimization with applications. Technical report, Stanford University, 2002. Doctoral Dissertation, Electrical Engineering Department.

[3] D. Marbach, J. C. Costello, R. Küffner, N. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, the DREAM5 Consortium, M. Kellis, J. J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012.

[4] Sahan Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39:1069–1097, 2011.

[5] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. arxiv:1008.3654, 2010.

[6] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society, Series B, Methodological*, 71(5):1009–1030, 2009.

[7] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[8] Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? arxiv:1004.3484, 2010.

[9] G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and Applications*, 170:1039–1053, 1992.

[10] Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coeffcient estimation in multivariate linear regression. *J. R. Statist. Soc. B*, 69(3):329–346, 2007.

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

9